Income and Expenditure in US Households

A Multivariate Analysis of Consumer Expenditure fmli161 Dataset

Mingzhao Hu[1]
[1]Department of Statistics, University of Wisconsin, 1300 University Avenue, Madison, WI 53706

Abstract

In recent decades, data-driven approaches have been developed to analyze demographic and economic surveys on a large scale. The goal of this report is to utilize multivariate analysis techniques to gain insight on relationship between income and expenditure of American households. The fmli161 dataset from PUMD (Public-Use Micro Dataset) is based on the Consumer Expenditure survey conducted by the Bureau of Labor Statistics, with initially 35 variables from three categories: demographics, income and expenditure. Missing values and categorical variables are the first to be handled in preliminary analysis. On the mathematical side, I propose to evaluate the data and the results for stability and reproducibility, based on several multivariate procedures. Principal component analysis is first adopted to directly visualize the data structure and gain intuitive understanding. Furthermore, efforts have been made to group the more significant income and expenditure variables through cluster of variables. Hierarchical clustering provides important solutions to problems later encountered in canonical correlation analysis revealing correlation between groups of income and expenditures. However, this extensive dataset is not only limited to economic interpretations, as demonstrated through the section of correspondence analysis discussing educational inequality due to racial factors. In conclusion, sparse PCA suggests FINCBTXM, FSALARYM, TOTEXPCQ, FOODCQ and HOUSCQ as the five most important variables of the selected, while cluster analysis gives more options depending on the number of clusters needed. CCA revealed high correlation between income and expenditure for middle class Americans, while correspondence analysis does not fully support suggestions of rebalancing higher educational rights based on race.

Keywords: Consumer Expenditure survey, PCA, Correspondence analysis, Cluster of variables, Canonical correlation analysis.

1. Introduction

1.1 Background

Statistics has become an integral part in the making of modern economics policy, which aims at balancing the allocation of resources within our society. At the level of daily life, the immediate concern is for every citizen to have a balanced income and expenditure, such that the income supports the necessities of the person and his dependents, while enough expenditures occur to ensure the healthy flow of resources within our society.

Statistics comes in as the tool for determining quantitatively how the overall situation is and providing the interpretations crucial for making a calculated decision. There are three main responsibilities of statistics in economic studies: data collection, analysis and presentation, inference and prediction. In particular, multivariate analysis is powerful in revealing the connections and relative significance between factors, which in reality can be both numerous and complex.

As the modern society develops and we enter the digital age, the datasets statisticians have to deal with are increasing rapidly, even if we only consider the simple measure in the dimensionality of the data. For example, at the beginning of modern statistics, the datasets in biological experiments were at most with sample size in the hundreds and number of variables in the dozens. Today, datasets from clinical trials in industries and research can easily have sample size in the millions and just as many variables. It is no wonder that high-dimensional data analysis is an integral part of current statistical research. However, high-dimensional data analysis have its origin and foundations in multivariate analysis, and methods in multivariate analysis prove to be, more often than not, direct and elegant solutions to most of the large datasets researchers encounter. In fact, since the 20th century, with the development of computer science and related industries such as biology, pharmacy, multivariate analysis has significantly developed to include an array of tools including Principal Component Analysis, Cluster Analysis, Canonical Correlation Analysis and so on. In this project the goal is to utilize a few appropriate techniques to reveal the economic knowledge that can be acquired via simple yet useful multivariate analysis methods

Modern economics research, no matter theoretical or applied, emphasize the importance of empirical analysis and simulation, due to a stronger need to provide quantitative support for claims and connect to reality situations. However, a significant trend in economics research is the application of linear regression models. While regression is undoubtedly a powerful and appropriate statistical tool, a lot could be gained by going beyond the boundary of only a good linear fit and going into the details of the interactions and comparisons between the variables.

1.2 Source

The Bureau of Labor Statistics (BLS) under the United States Department of Labor has been tasked with surveying, computing and reporting one key indicator of macroeconomic wellbeing: the Consumer Price Index(CPI). In light of the need of researchers from various disciplines, BLS has been releasing their collected survey dataset for computing weights for the up-to-date CPI in the Public-Use Micro Datasets, which includes two main files, the first of which is the Interview files. Compared to the Diary files in the second category which records the detailed daily expenses and incomes of the survey participants, the Interview files are based on quarterly interviews. During an interview, the CU is asked to report expenditures for a reference period of three months. Therefore one can hope to encounter less redundant information but sacrifice some level of accuracy. The Interview files are divided into five categories, each with different focus and updated every quarter. For this analysis, the interest is in FMLI 161, the first category of files covering the most recent published period of 2016's first quarter. All FMLI files contain Consumer Unit characteristics, income, and summary level expenditures. The BLS has conducted extensive research based on these datasets, including imputations for missing data. For beginners to the dataset, it is a good practice to start with the pumd_novice_guide and get a general understanding of the organization. However, in order to utilize the 807 variables in the FMLI 161 file, a very useful file to help the

understanding and selection is the data dictionary, which includes explanations for all the variables categorized in characteristics, income, and summary level expenditures.

The source of the original data is the website of Bureau of Labor Statistics [1], and is open to the entire public for free download. Formats are in SAS, SPSS, STAT and CSV.

1.3 Objectives

At the most general level, the goals of this report are:

i) Provide the readers with a strategic sense of the FMLI 161 dataset via subset selection and detailed discussions of the variable in forming the experimental dataset

ii) Indicate arguments for treating missing values in preparation for creating experiment dataset

iii) Adopt multivariate analysis techniques for a quantitative understanding of the relationship between the selected variables, both in pairs of special interest and as groups of similarity

iv) Interpret based on analysis results to shed light on real life implications of socioeconomic significance

1.4 Approach

The best subset was selected completely judgmentally. The variables selected are most representative and as quantitative as possible in the three categories of Characteristics, Income and Expenditures. Principal component analysis was adopted to verify these results and gain a deeper understanding of the data structure, directly from graphs. Also utilized Cluster Analysis to compare clusters with intuitive grouping based on meaning of variables, and further conducted Canonical Correlation Analysis to investigate the relationships between the groups. Correspondence analysis was done on a particular pair of important categorical variables.

2. Preliminary Analysis

2.1 Description of Variables

2.1.1 Initial survey dataset selected:

After reading the data, a total of 35 variables were selected. However, due to missing values a number were dropped in later steps. The justification for selection of variables are provided in the following section. Table I gives the corresponding detailed descriptions.

Table I: Descriptions of Selected Variables

NEWID	Consumer Unit* (CU) identification number. Values of NEWID contain a leading zero.
HH_CU_Q	Count of CUs in household

AS_COMP1	Number of males age 16 and over in CU
AS_COMP2	Number of females age 16 and over in CU
BLS_URBN	CODED 1 Urban 2 Rural
CUTENURE	Housing tenure CODED 1 Owned with mortgage 2 Owned without mortgage 3 Owned mortgage not reported 4 Rented 5 Occupied without payment of cash rent 6 Student housing
FAM_SIZE	Number of members in CU
NO_EARNR	Number of Earners*
PERSLT18	Number of children less than 18 in CU
PERSOT64	Number of persons over 64 in CU
VEHQ	Number of owned vehicles
AGE_REF	Age of reference person
EDUC_REF	Education of reference person* CODED 00 Never attended school 10 First through eighth grade 11 Ninth through twelfth grade (no H.S. diploma) 12 High school graduate 13 Some college, less than college graduate 14 Associate's degree (occupational/vocational or academic) 15 Bachelor's degree 16 Master's degree, (professional/Doctorate degree)
REF_RACE	Race of reference person: CODED 1 White 2 Black 3 Native American

	4 Asian 5 Pacific Islander 6 Multi-race
SEX_REF	Sex of reference person CODED 1 Male 2 Female
INC_HRS1	Number of hours usually worked per week by reference person
OCCUCOD1	The job in which reference person received the most earnings during the past 12 months best fits the following category. CODED by professions 1 - 18
FINCBTXM	Amount of CU income after taxes in the past 12 months (FINCBTAX – TOTTXPDX)
FSALARYM	Amount of wage and salary income, before deductions, received by all CU members in past 12 months
FRRETIRM*	Amount of Social Security and Railroad Retirement income, prior to deductions for medical insurance and Medicare*
FSMPFRXM	Total amount of income received from self-employment income
WELFAREM	During the past 12 months, the total amount of income from public assistance or welfare
INTRDVXM	Amount of income received from interest and dividends: savings or bonds; stocks; trust funds.
NETRENTM	Amount of income received from net rental income or loss
RETSURVX	The amount received in retirement, survivor, or disability pensions during the past 12 months.
TOTEXPCQ*	Total expenditures this quarter

FOODCQ	Total food this quarter
FDHOMECQ	Food at home this quarter
FDAWAYCQ	Food away from home this quarter
HOUSCQ	Housing this quarter
APPARCQ	Apparel and services this quarter
TRANSCQ	Transportation this quarter
HEALTHCQ	Health care this quarter
ENTERTCQ	Entertainment this quarter
EDUCACQ	Education this quarter
READCQ	Reading this quarter

*Consumer Unit**: For ease of understanding, can be simply considered as an individual consumer in real life

*Earners**: A consumer unit member, 14 years of age or older, who reported having worked at least 1 week during the 12 months prior to the interview date.

*FRRETIRM**: received by all CU members in the past 12 months (sum SOCRRXM from MEMB file for all CU members)

*TOTEXPCQ**: Food + Housing + Apparel and Services + Transportation + Healthcare + Entertainment + Other Expenditures

In total there are 6 categorical variables and 29 continuous variables. At this stage, the selected dataset has 6426 samples (Consumer Units) of 35 variables, but includes a considerable proportion of missing values read in as NA. Missing values is a major issue with this survey, and given the effort to avoid it in the survey stage, it relies on good statistical reasoning to deal with it in the analysis, shown in section II. B.

The four categorical variables are BLS_URBN (location urban or not), CUTENURE (housing tenure), EDUC_REF (education), REF_RACE (race), SEX_REF (sex), and OCCUCOD1 (occupation). These categorical variables exists naturally, as it is intuitive to categorize characteristics such as educational level and sex. However, since many multivariate analysis methods require continuous variables in input datasets, such as PCA, after discussions on missing values and truncating the dataset to its final version we will expand these categorical variables to become continuous before the analysis.

2.1.2 Justifications for variable selection:

Since the target of the Consumer Expenditure survey is to produce the Consumer Price Index, the 807 variables in the FMLI 161 files fall in three major categories: Consumer Unit characteristics, income, and summary level expenditures. This report will investigate

the relationship between variables in different categories and reveal or confirm claims and assumptions about the interactions. For example, some questions of interests include:

i) Do people with higher education necessarily have higher income? (Characteristics vs. expenditure)

ii) Do male earn more than females? Do they spend more? What is the significance of gender? (Characteristics vs. expenditure)

iii) Do people who spend more eating out necessarily have higher overall expenditures? (Expenditures vs. expenditure)

iv) Do people whose income relies on salaries have similar spending behaviors as those relying on pensions? (Income vs. income)

Therefore, not all 807 variables are necessary. In practice, choosing only relevant variables is a useful strategy in terms of avoiding missing values, lowering analysis difficulties, and reducing cost. However, that is based on good justification. For this report, NEW_ID is the row names of the dataset, and each CU is uniquely identified by NEWID. There are 14 variables in the characteristics category, 10 variables in the income category, and 11 variables in the expenditures. The criterion for selection is that for each subcategory within a category, such as income due to rent and expenditure on entertainment, only one variable is selected and the preference is laid on the BLS derived variables, which is the common setting for most of the information included in the PUMD files. Note that the decision to take the BLS derived variables is based on the fact that BLS treated the original data with care, using multiple imputations and records from previous-years in the estimation of missing components. The derivation also includes introducing anonymity to protect respondent identity. Details of the methods introduced in the csxguide file. Although sacrificed 773 variables, the remaining variables listed in the previous part do include the most useful information and make sense to the general readers. The dataset produced by the 35 variables selected offers an acceptable coverage of key characteristics, and sufficiently large number of sample size that even allows for further reduction and grouping. This dataset is recorded as **fmli**, and provides a good basis for the next step: missing values.

2.2 Missing Values

To deal with the missing values in the dataset, a strong tool is attained by building a dataset of variables in the original FMLI 161 dataset that correspond to the selected variables in the fmli dataset. For example, for HH_CU_Q the corresponding variable, known as a flag variable, is HH_CU_Q_. BLS has named these flag variables, which were solely derived for the purpose of recording the state of the data values, as shown below: \

Table II: Definitions of Missing Flag Variables

A	Valid non-response: a response is not anticipated
C	"don't know", refusal or any other type of non-response
D	Valid data value
T	Topcoding applied to value

When reading in the FMLI 161 dataset, note that the approach was to regard all " . " and "NA" values in the dataset as missing. However, in many expense columns where the input values were just 0, although the corresponding flag variables do not exist for these expense columns such as "FOODCQ", it can be known that the zero means the respondent replied instead of omitting the answer, therefore will not be considered as missing. It is necessary to note that in later analysis these will cause further interpretation. Based on the dataset created, named **flag**, one third of all the expenditures in TOTEXPCQ are missing. To be specific, out of the 62426 samples 2171 have missing values in TOTEXPCQ. I first tried to split the dataset according to if TOTEXPCQ is empty. This is valid because according to the corresponding **flag** dataset recordings the missing values in selected columns are all Missing Completely As Random, labeled as "A".

```
> library("mice")
> delete = which(fmli$TOTEXPCQ == 0)
> fmlifull = fmli[-delete,]
```

A list-wise deletion for the rows with missing values was performed. An important observation of the **fmli** output file is that of the 2171 respondents reporting 0 for the total expenditure, they have also reported 0 for the other detailed expenses. Of the ones who have reported a certain amount of total expenditure, all have managed to also report sums spent on "food at home", "eating out" and "housing". Admittedly, there are some who still reported 0 for transportation, healthcare and entertainment. But a more significant issue is that most people reported zero expenditure for reading and education.

The next step would be to check that once we ensured no missing values exist in one of the most important column, TOTEXPCQ, how many are missing in the other columns. As a general rule of thumb, we consider missing values below 10% of the total information in the variable to be acceptable.

First, consider the detailed expense columns which do not have corresponding flag variables yet contain a lot of zero entries. Since we have already split the dataset according to total expenditure missing or not, we consider these to be "valid data". It does not mean the respondents intentionally ignored answering questions such as "How much you spent on food you ate at home during the past three months", because they have accurately reported the total expenditure and expense on eating out. The best explanation to offer would be the amount spent is negligible. Therefore these zeros would affect conclusions and we should consider them as meaning literally spending $0 on each expense. While

trying to provide a quantitative justification for this argument, we have discovered a surprising fact: We need to check the missing percentages and decide the measure. The results are in Table III.

Table III: Missing Values Percentages

Expenditure Variable	Percentage of Missing
FOODCQ	0.61%
FDHOMECQ	1.01%
FDAWAYCQ	19.42%
HOUSCQ	0.50%
APPARCQ	47.92%
TRANSCQ	5.92%
HEALTHCQ	20.81%
ENTERTCQ	14.34%
EDUCACQ	84.06%
READCQ	80.22%

It is clear that in the survey reading and educational expenditures mostly 0. This is an interesting and somewhat disappointing feature as it reflects most individuals and household in America does not invest enough in the educations.

Eating out, healthcare and entertainment expenditures still has considerable 0s even after throwing away the part where TOTEXP is 0. This is expected, as most of the middle class or lower income families are not likely to frequently spend money on not so necessary pleasures such as these. However, it will be interesting in later analysis to group separately the respondents who have no spending on education/ reading and the respondents who have no spending on these recreations. Then compare their income levels, we can verify the effect of income status on spending emphasis. A more elaborate plan would be to investigate the cross relationship between educational level, income level, and expenditure on reading and education, to see if more educated people earn more and care more about education.

We also observe about half of people reported 0 for the expenditure on apparels. This echoes with the lack of expenditures on other recreational expenses. It might be a good idea to simply use these recreational expenses as indicators for economic well-being and consumer confidence in the market. While not related to this project, a time series analysis on the recreational expenses compared to other, more traditional factors such as unemployment rate since 2008 might be an interesting project. For our purposes, we will include all 11 expense data columns.

Finally, we will look at missing variables labeled as NA in income columns. For the characteristics columns we observe no missing values. This is easy to understand as it is unlikely people will withhold basic information such as gender or how many cars are owned, especially after splitting the dataset and reducing the sample size. After confirming that indeed no values are missing in the characteristics, the final step is dealing with missing values in the income variables. For "working hours per week" and "primary type of occupation", both missing 1453 values, 1453/4524 = 32.11%, which is very significant. Thus these two variables are dropped since they are not essential to defining the income situation, and not of essential concern for questions we are interested in. This is a regular approach in reducing complexity of dataset and avoiding missing values, but the sacrifice is in the robustness of results since we have fewer dimensions and explanatory variables. Four other income variables: Welfare income, Interest/Dividends income, Retirement income, and Rental income are missing 3519, 3645, 4048, 4209 values respectively, which are significant. While we can use multiple imputation, the original dataset has so many columns of variables that many are not that essential and even after section we have a large number of variables that are either not informative on the general population (e.g., Retirement income) or overlaps in information provided (e.g., Retirement and Welfare income). On the other hand, while multiple imputation is a useful tool when missing values cannot be avoided, it has its drawbacks, especially when we are dealing with complex reality surveys such as this Consumer Expenditure (CE) survey. More extensive and rigorous arguments and findings discussing the different approaches in managing missing values on real economic surveys datasets, in particularly the CE survey, can be found in Loh, Eltinge, Cho & Li (2016) [2]. However, for the purposes of this project we will only use the simplest approach. We name the resulting dataset **fmlifull**. Since we have 14 characteristics variables, 4 income variables and 11 expenditure variables, most of which contains no missing values, we can remove the rows with missing values. There are two reasons for this. First, based on the flag variables all are MCAR. Second, we have a large sample size that will not be significantly affected by a small number of loss when deleting missing values. However, using md.pattern() on the resulting dataset **fmlifull** has shown that in the remaining dataset there are no more missing values, so we need not make any deletion.

2.3 Correlation Matrix

Correlation matrix reflects the correlation between every pair of the variables. It gives a direct illustration of the significance of dependence. It is also the basis for R in computing the Principal Component Analysis, which is a major section of this report. It also makes sense in the later section of canonical correlation analysis which strives to solve the influence significance between income and expenditure variables to first check the pair-wise correlation between the variables before. However, as the number of variables increases, the dimension of the matrix increases, making it harder to interpret the correlations. In dealing with correlation matrix, it is important to remember that it just reflects dependency, i.e., similarity in change, not necessarily reflecting the causal relationship. In our case, with 29 variables, the dimensions are still acceptable. A very useful trick is to concatenate the digits before presenting.

```
> round(cor(fmlifull),digits = 2)
```

However, there are four categorical variables among the 29 mostly continuous variables: BLS_URBN, EDUC_REF, REF_RACE, SEX_REF. In order to incorporate them in later analysis, we need to avoid the problems caused by ignoring the difference between categorical variables and continuous variables, and simply consider the numbers representing different categories as continuous values. The best approach, is to expand each categorical variables according to their levels in to as many new columns of variables correspondingly. The new columns, which just comprises of 0's and 1's, should not cause further trouble. In total, we have changed from four categorical variables to 18 continuous variables. By sacrificing simplicity in the matrix we have achieved uniformity in the columns. After column-wise deleting the four categorical variables, we add the 18 new columns to the dataset **fmlifull**, making it now a 4254 by 43. Then it makes sense to calculate the correlation matrix for the 43 variables. Bear in mind that given the size of the correlation matrix, it is important to focus on portions and draw conclusions in general.

The correlation matrix is given in Appendix A. There are two main results to be gained from this. The first one is that even though 43 is still a relatively small amount of variables included for the data-driven world today, especially when compared to clinical trials and biological experiments where the number of variables involved easily exceed 10,000, the correlation matrix is already very difficult to read and is no longer an effective measure of reporting. The second is that a major observation from the correlation matrix is most of the correlations are below 0.5. In fact, using R can show that out of the 903 pairs of different variables, only 79 has correlation coefficient larger than 0.5. It makes sense since when dealing with real-life datasets, we often consider the influence of other factors on a pair-wise relationship. For example, there is 0.55 correlation between the variables FINCBTXM (income before tax in last 12 months) and TOTEXPCQ (total expenditure of this quarter). We would expect to see something higher, given that intuitively the amount of income should be reflected in the amount spent. Part of the explanation for this could be that economically speaking the first quarter is generally not the most consumption-intensive quarter when compared to others, and different income groups tend to have different strategy with spending money. On the other hand, if we consider the relative magnitude of the correlation to others, we would gain more information. The 0.55 correlation is pretty high among all, and correlations between other pairs of significantly related variables, such as FOODCQ and FDHOMECQ also reflect this trend. Therefore the correlation matrix is still providing robust results, just not in a very efficient manner.

3. Principal Component Analysis

3.1 Objectives

The quarterly interview of the CE survey was intentionally conducted with three sections: respondent characteristics, income summary and expenditure summary. This is reflected in the variables selected to compose this experimental dataset. While the CE survey was primarily for the purpose of calculating the Consumer Price Index, a lot more detailed information can be attained from this extensive and carefully conducted survey. For

example, one of the main questions this report would like to answer is whether the response of the survey supports the intuitive grouping of variables in the design, if so, do the response give more detail? If not, what are the explanations? Furthermore, what further relationship the response to the variables reflect on grouping and the influence between groups? To answer these questions, we will discuss one at a time using different multivariate techniques. This section is dedicated to adopting Principal Component Analysis to gain understanding of relative significance of variables. Grouping of the dataset by a single factor will also be demonstrated. This section is a prelude for the cluster analysis in the following section, and through the rich resources of PCA presents abundant graphical illustrations of the dataset. PCA also helps in standardizing the data, decrease dimensionality and demonstrating the more important variables. Here it is worth pointing out that since 43 variables is more than easy for any reader to process, selecting the more important ones and attempting dimensionality reduction is a main theme and challenge of this project, as important as endeavors on dealing with missing values.

3.2 Principal Component Analysis

After list-wise deleting all the missing values, we used the PCA generic function princomp() in software R to find the principal components of dataset on its correlation matrix.

```
> fmlifull.pca = princomp(x=fmlifull, cor = TRUE)
```

As an interesting point, a lot of users of R have been confused as to when to use princomp() instead of the other function, prcomp(), to compute principle components. The two gives slightly different eigenvalues, but one is always a multiple of the other. One explanation is that princomp() uses the correlation matrix and calculates the variance with the divisor N while prcomp() uses the covariance matrix and calculates variance with divisor N-1. For the purposes of this project, both are fine, but princomp() is more familiar to the author.

Using the summary of outputs, we get 43 principle components, each a linear combination of the 43 variables, and explain a fraction of the total variance. The entire output is included in Appendix B. We include the first 6 PCs.

```
##                  Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## HH_CU_Q            0.04  -0.07  -0.03   0.09   0.16   0.04
## AS_COMP1          -0.18   0.03   0.31  -0.17  -0.25  -0.02
## AS_COMP2          -0.11  -0.13  -0.30  -0.30  -0.15   0.07
## CUTENURE           0.15  -0.20   0.00   0.14   0.07  -0.13
## FAM_SIZE          -0.22  -0.21   0.00  -0.38  -0.24   0.00
## NO_EARNR          -0.25  -0.23   0.12  -0.13  -0.06   0.11
## PERSLT18          -0.15  -0.26  -0.02  -0.28  -0.12  -0.04
## PERSOT64           0.12   0.39  -0.17  -0.13  -0.24  -0.02
## VEHQ              -0.17   0.11   0.09  -0.18  -0.05   0.13
## AGE_REF            0.11   0.40  -0.15  -0.08  -0.19   0.01
## FINCBTXM          -0.32   0.06   0.01   0.07   0.06   0.10
## FSALARYM          -0.31  -0.05   0.06   0.07   0.08   0.09
## FRRETIRM           0.12   0.39  -0.16  -0.13  -0.22  -0.03
## FSMPFRXM          -0.11   0.04  -0.02   0.04   0.08   0.05
## TOTEXPCQ          -0.31   0.16  -0.11   0.10   0.12  -0.10
## FOODCQ            -0.31   0.10  -0.11   0.01  -0.04  -0.17
## FDHOMECQ          -0.25   0.05  -0.10  -0.10  -0.10  -0.15
## FDAWAYCQ          -0.25   0.11  -0.07   0.12   0.05  -0.12
## HOUSCQ            -0.27   0.09  -0.10   0.11   0.08  -0.08
## APPARCQ           -0.12   0.02  -0.08   0.08   0.10  -0.09
## TRANSCQ           -0.12   0.10  -0.07   0.01   0.06  -0.08
## HEALTHCQ          -0.13   0.21  -0.08   0.00  -0.01   0.00
## ENTERTCQ          -0.15   0.12  -0.08   0.07   0.08  -0.02
## EDUCACQ           -0.06   0.00  -0.01   0.09   0.12  -0.04
## READCQ            -0.07   0.12  -0.06   0.06   0.08   0.00
## BLS_URBN_URBAN    -0.06  -0.08  -0.11   0.36  -0.40   0.40
## BLS_URBN_RURAL     0.06   0.08   0.11  -0.36   0.40  -0.40
## EDUC_REF_NEVER     0.01  -0.01  -0.01  -0.01  -0.03   0.03
## EDUC_REF_FIRST     0.02  -0.03   0.00  -0.12  -0.05  -0.01
## EDUC_REF_SECOND    0.07  -0.04  -0.02  -0.04  -0.11  -0.09
## EDUC_REF_HIGH      0.07   0.03   0.05  -0.16  -0.06  -0.08
## EDUC_REF_ASSC      0.00  -0.04  -0.02  -0.05   0.00   0.10
## EDUC_REF_UDGRD    -0.09   0.00   0.01   0.15   0.10   0.06

##                  Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## EDUC_REF_GRD      -0.10   0.09  -0.05   0.12   0.07  -0.02
## REF_RACE_WH       -0.01   0.20   0.15  -0.22   0.29   0.49
## REF_RACE_BLK       0.06  -0.17  -0.12   0.17  -0.22  -0.37
## REF_RACE_NA       -0.01  -0.02   0.00   0.00  -0.04  -0.09
## REF_RACE_AS       -0.06  -0.08  -0.07   0.12  -0.13  -0.24
## REF_RACE_PI       -0.01  -0.02   0.00   0.02  -0.04  -0.04
## REF_RACE_MULTI    -0.02  -0.04  -0.03   0.04  -0.07  -0.09
## SEX_REF_MALE      -0.05   0.17   0.53   0.12  -0.18  -0.12
## SEX_REF_FEMALE     0.05  -0.17  -0.53  -0.12   0.18   0.12
```

Figure 1: First 6 principle components from regular PCA

The corresponding variance:

```
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6
Standard deviation     2.4575533 1.75764897 1.57076826 1.53409571 1.36991131 1.34170408
Proportion of Variance 0.1404551 0.07184488 0.05737937 0.05473139 0.04364319 0.04186441
Cumulative Proportion  0.1404551 0.21229996 0.26967933 0.32441072 0.36805390 0.40991832
```

Figure 2: Corresponding variances for the First 6 PCs

The command used is:

```
> round(t(summary(fmlifull.pca, loadings = TRUE)$loadings[,1:6]), digits = 2)
```

Instead of directly printing from the Summary, because most of the coefficients are small so it does not make sense to omit coefficients of value 0.06 in the output but including coefficient at 0.15.

The attempt at directly interpreting the coefficients of each PC has failed to offer useful insights into the grouping or relative significance of the variables. The hope was that each principle component, as a linear combination of all 43 variables, would have coefficients that can be interpreted such that one PC might be representing the contrast between income and expenditure as two groups, another representing contrast between income after tax and retirement income, etc. This failed because most PCs, as demonstrated above, have nonzero coefficients for each variable, with no pattern that can be easily distinguished. However, this is a natural result of having 43 variables, and we will solve this in Part C of this section. Next we look at the scree plot.

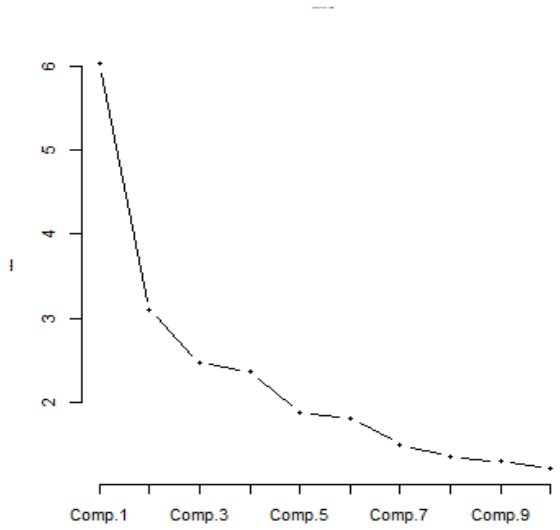


Figure 3: Scree plot

The scree plot is a direct and visual demonstration of the importance of the PCs by the amount of variance each have explained. In general, we look for the "elbow" to determine the best amount of principle components that we consider to be important. In this case, we observe the most significant "elbow" at 3, another at 5. With three PCs we have 26.96% of the total variance explained, which is not sufficient for showing the majority of the variance. For 5 PCs we have 36.80% of the variance explained, not ideal either. This, and the related fact that the percentage of variance explained by each PC is about the same from PC2 to PC6, goes on to show that the normal PCA is not ideal in showing relative significance among variables in this situation. However, the advantage of PCA is that it is robust, requires no sparsity constraint and does not misidentify the important variables.

These are the reasons why we should thoroughly investigate the results of PCA before adopting the sparse PCA in part C.

It would naturally follow that we plot the variables according to their importance to the PCs, when we just consider the PCs as linear combinations of original variables that are always orthogonal to each other. This will offer us a new perspective to the dataset, especially considering that with 43 variables it is difficult to have a direct graphic presentation in 2D. In particular it would be most helpful to see their dispersion in the plot of PC1 vs. PC2, the two most important PCs. It is not very informative to plot the ID number of 4524 sample points, so we choose to plot the 43 variables instead by their coefficients in the first two PCs.

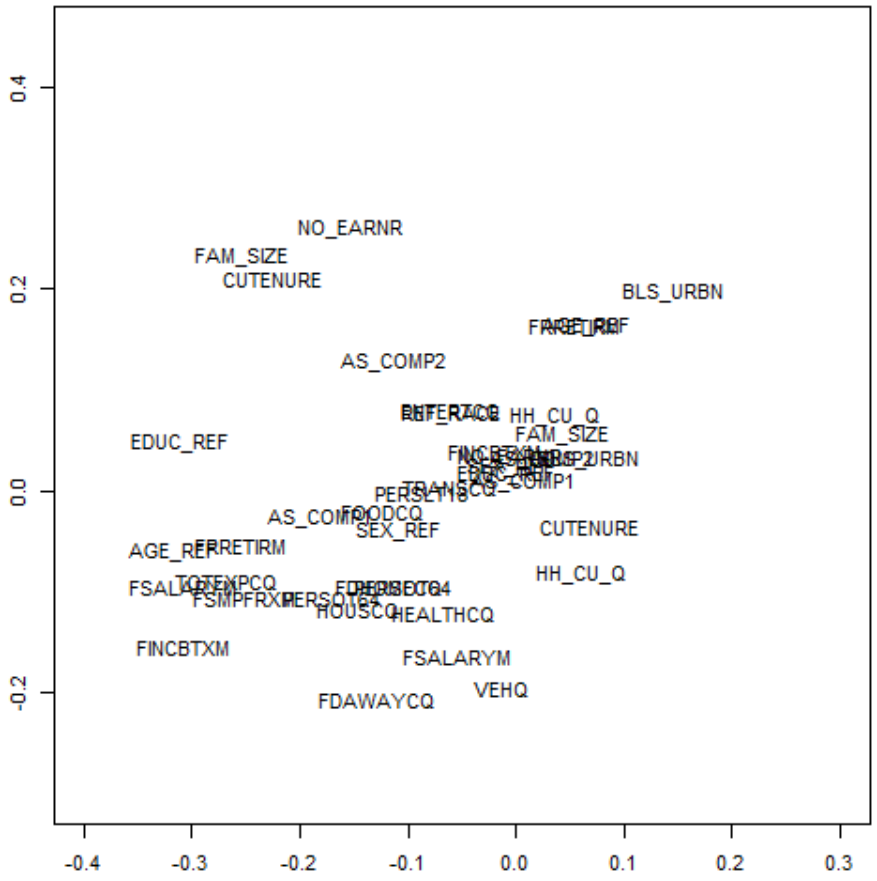
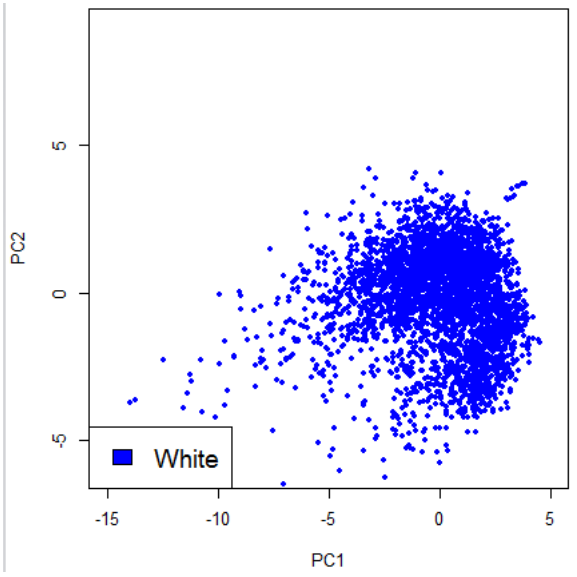


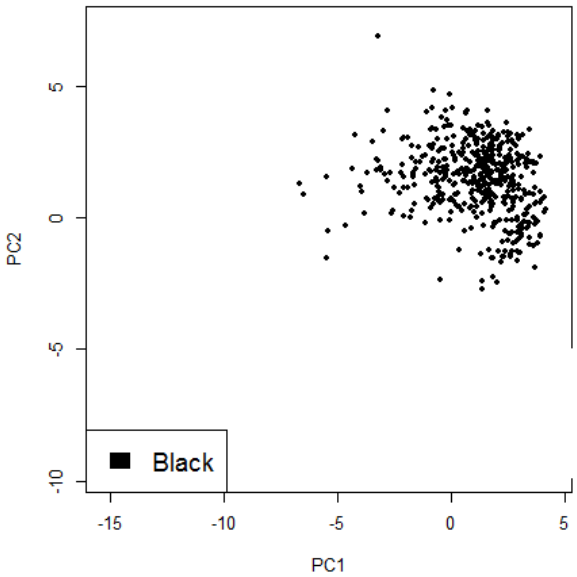
Figure 4: PC1 vs. PC2

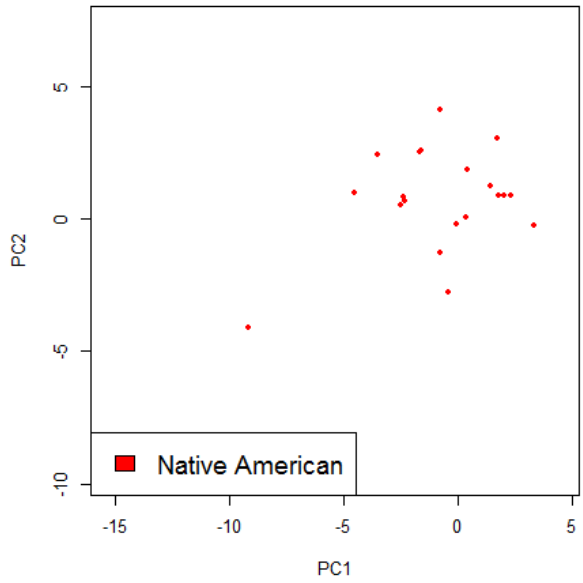
Basically we can observe no real outlying variables from this plot. The fact that PC1, PC2 only explains 21.22% of the total variance, and the rest of the PCs each explain small portions means that the whole dataset is probably arranged more similar to a sphere in dimension of 43, since there is no sharp variation along any direction. Then when we look at the plot above, we can see that variation along PC1 is about (-0.4, 0.2), along PC2 is about (-0.2, 0.3), not very significant difference. Another observation is that some

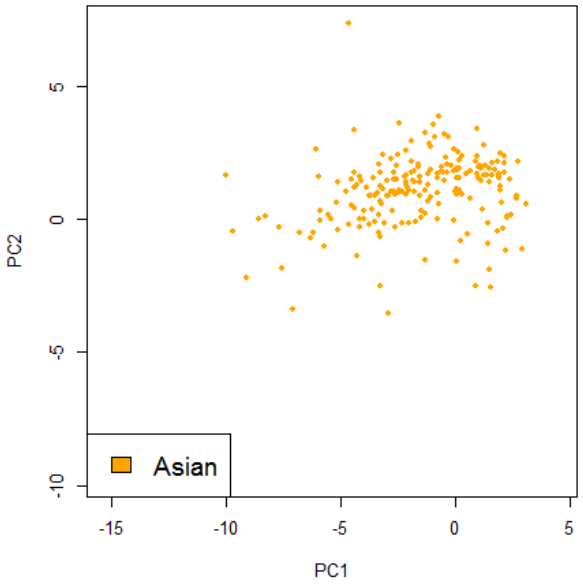
clustering seem to be around a subset of the characteristic variables, FAM_SIZE, BLS_URBN, PERSLT_18, etc. This is not true, as PC1 vs. PC2 plot is only in 2 dimensions, so being close in this plots may well mean being far apart when the 41 remaining dimensions have been taken into account.

Based on these results, more in-depth analysis of the survey can be conducted. For example, asking whether there is significant grouping effect in the PC1-PC2 plane by racial factors is of huge social and political significance. A lot of discussion on racial and social equity is based on survey analysis such as this. To avoid dense clustering of colored points covering up others, we decided to display the plots separately according to races.






PC2
5
0
-5
-10
-15
-10
-5
0
5
PC1
Native American


PC2
5
0
-5
-10
-15
-10
-5
0
5
PC1
Asian

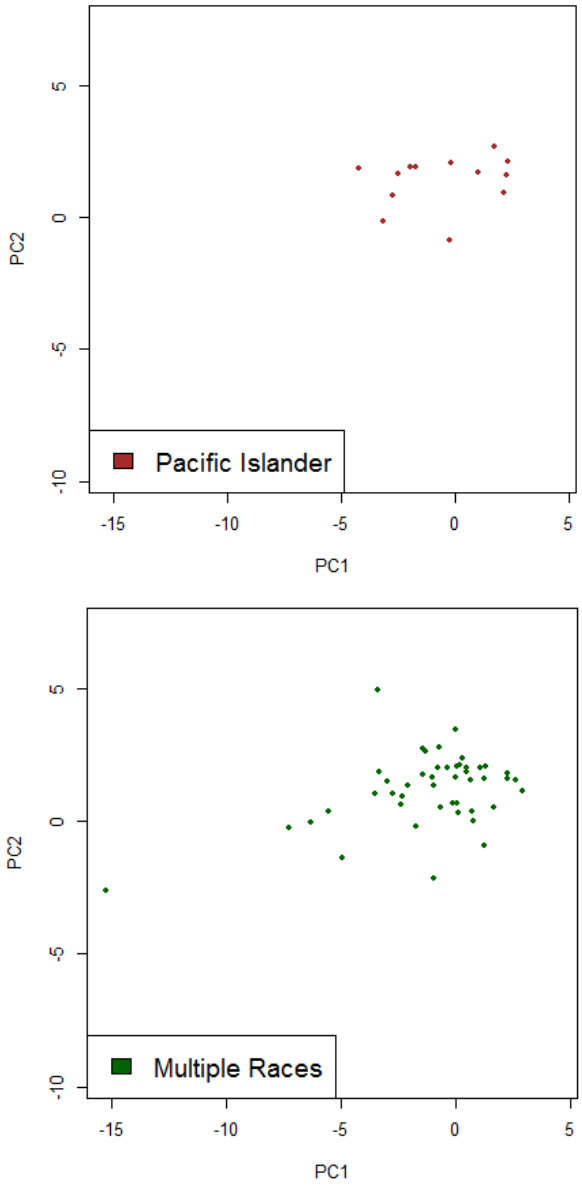


Figure 5: PC1 vs. PC2 according to races

The immediate visual perception is that white respondents take up the majority, while there are very few Pacific Islanders or Native Americans. Also, we can observe that for each plot the basic spread of points mimics that of the first. For Pacific Islanders, Native Americans and Multiple Races, the sapmle points are too few to make any conclusive arguments, but for Asians we can observe small dismillarities from the plot of Whites. This means that in the PC1-PC2 plane the Asian community's projections so have differences from the White community, while all variables have been taken into account. It is also worth noting that after dividing into groups the restriction of sample size made it difficult to reach robust conclusions.

The most useful graphical respresentation from PCA is probably the biplot. Many important information is included in one plot: Observations are points and variables are vectors, inter-point distances between points are Mahalanobis distances, length of a vector is standard deviation, dot product between two vectors is covariance, cosine of angle between two vectors is correlation, and length of difference of two vectors is standard deviations of difference of the two variables.

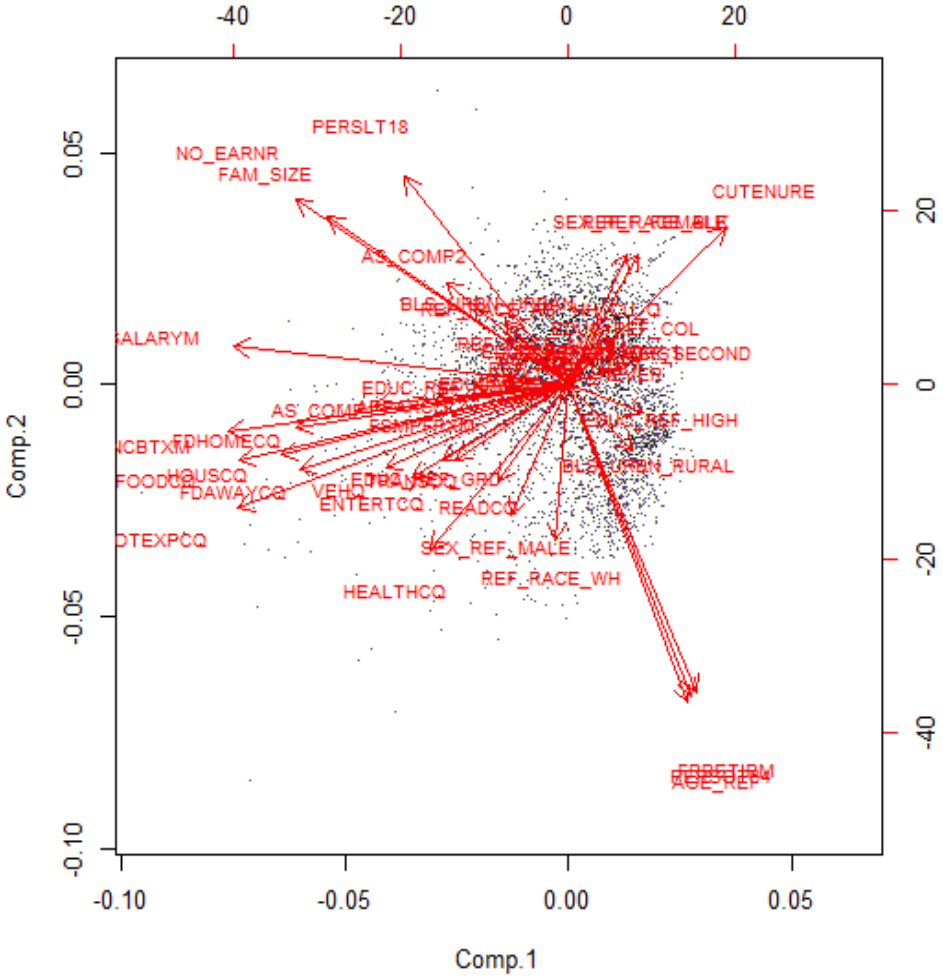
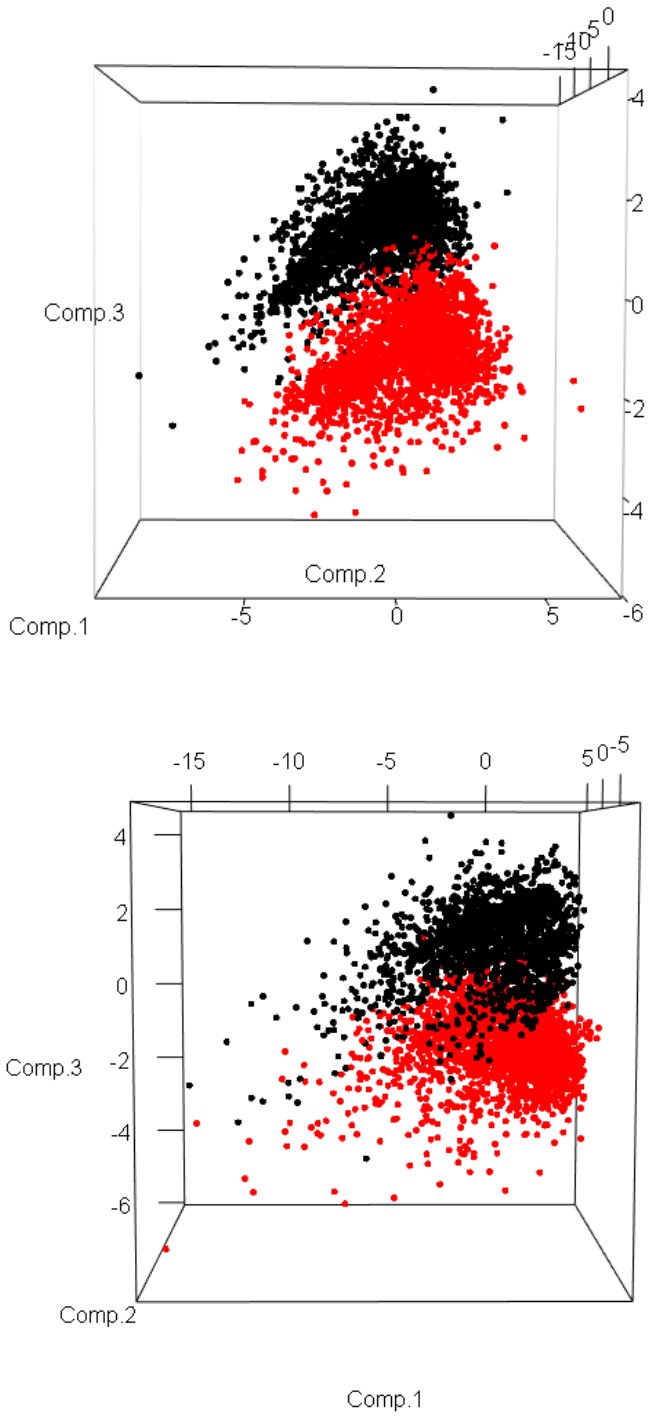


Figure 6: Biplot

To avoid the data points completely obscuring the axis, we chose the representation to minimize its presence, as the inter-observational distances are not of interest. This is still in the PC1-PC2 plane, but offers a more powerful demonstration of the grouping of variables. While careful to avoid that in 43 dimensions the same results hold, we can conclude that retirement income and age are close, number of earners and family size are close, and expenditure variables including total, eating out, housing, eating at home are in the same direction as income before tax. Also, the pattern of stretching from top-right to lower-left observed before in the grouping by races can now be explained as variation due to income and expenditure variables.

Lastly, we consider a 3D representation of the dataset. While we have taken a close look at the racial factor, it would be equally important to discuss gender equality in our society.



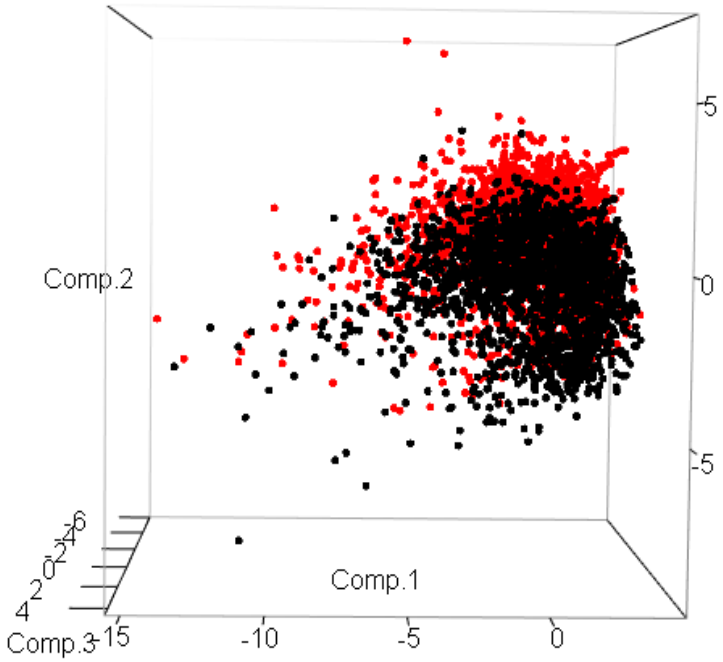


Figure 7: 3D plot, PC1 vs. PC2, PC3. Separated by gender,

Using the ggplot and rgl package in R, rotational 3D representation of PCA is attainable. This offers the researcher additional dimension of presenting the dataset projections. The red represents the females, while the black represent the males. A segregation based on gender is clearly observed in all of these plots. An interesting observation comes from the lowest plot, showing the PC1-PC2 projection again. We see that the male and females basically overlap each other, and the patter of stretch is also preserved. Going back to Table 2, gender is indeed of less importance to PC1. This is good news, as it shows that gender gap is not very easy to observe from this survey.

In conclusion, PCA has reinforced some intuitive grouping of the variables through the biplot and 3D projection, and revealed details via grouping of races. While not performing as well as expected, it laid the foundation for pursuing the more appropriate sparse PCA approach.

3.3 Sparse Principal Component Analysis

The usual PCA approach has provided useful insight into variable interactions, as well as grouping by race and gender. It is easy to implement and explain. However, given the 43 variables in the dataset, one deficiency of the usual PCA is the difficulty in interpreting the PCs and ineffectiveness in dimension reduction. The hope is that not every variable will have a nonzero coefficient in the PCs. To achieve this we need to enforce a sparsity constraint, proposed by Zou, Hastie, & Tibshirani (2006) **[3]**. The R package nsprcomp contains the function nsprcomp for conducting sparse pca. However, the user must set seed before to ensure the results are reproducible.

```
> set.seed(0)
```

> fmlifull.pca.sparse <- nsprcomp(fmlifull, ncomp = 3, center=T, scale.=T, k=c(5,5,5), nneg = FALSE)

For simplicity of interpretation and presentation we have chosen the number of PCs to be 3. However, to attain more robust results more careful determination is required, as discussed in Zou et. al. The outputs of PCs obtained on the same dataset is given below in two parts.

```
Importance of components:
                          PC1    PC2    PC3
Standard deviation     1.795 1.6842 1.5625
Proportion of Variance 0.379 0.3337 0.2873
Cumulative Proportion  0.379 0.7127 1.0000
```

Figure 8: PC components for sparse PCA

We can observe that there is now significant differences in the amount of variances explained by the 3 PCs. The mechanism involves forcing the constraint on the dataset, therefore forcing most coefficients to be 0, leaving the specified number of nonzero coefficients for each PC. The scree plot is peculiar, with no elbow. However, we no longer need the scree plot to tell us the optimal number of PCs. As a relatively new concept, sparse PCA is still developing and many questions are left unanswered. But for our case, sparse PCA has given 3 PCs as specified. PC1 represents the five most important factors of income and expenditure, FINCBTXM, FSALARYM, TOTEXPCQ, FOODCQ and HOUSCQ. PC2 is more concerned with the age structure, a contrast between NO_EARNR, PERSLT18 and PERSLT64, AGE_REF, FRRETIRM. PC3 is about the gender, male contrasting against female, AS_COMP1, VEHQ, SEX_REF_MALE against AS_COMP2, SEX_REF_FEMALE. It is interesting that number of vehicles owned is counted as an index of masculinity. The ease for interpretation has increased significantly with sparse PCA.

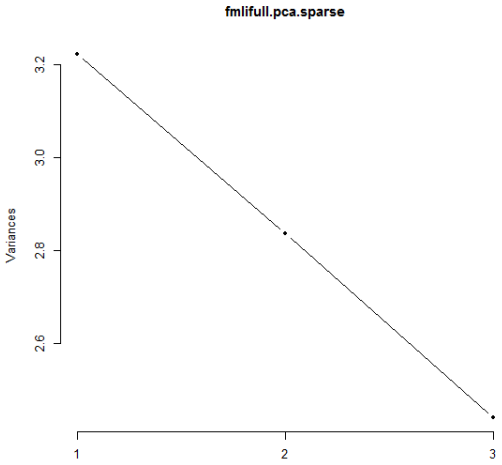


Figure 9: Scree plot for sparse PCA

	PC1	PC2	PC3
HH_CU_Q	0.00	0.00	0.00
AS_COMP1	0.00	0.00	-0.37
AS_COMP2	0.00	0.00	0.30
CUTENURE	0.00	0.00	0.00
FAM_SIZE	0.00	0.00	0.00
NO_EARNR	0.00	0.38	0.00
PERSLT18	0.00	0.28	0.00
PERSOT64	0.00	-0.52	0.00
VEHQ	0.00	0.00	-0.14
AGE_REF	0.00	-0.50	0.00
FINCBTXM	0.48	0.00	0.00
FSALARYM	0.46	0.00	0.00
FRRETIRM	0.00	-0.51	0.00
FSMPFRXM	0.00	0.00	0.00
TOTEXPCQ	0.46	0.00	0.00
FOODCQ	0.40	0.00	0.00
FDHOMECQ	0.00	0.00	0.00
FDAWAYCQ	0.00	0.00	0.00
HOUSCQ	0.44	0.00	0.00
APPARCQ	0.00	0.00	0.00
TRANSCQ	0.00	0.00	0.00
HEALTHCQ	0.00	0.00	0.00
ENTERTCQ	0.00	0.00	0.00
EDUCACQ	0.00	0.00	0.00
READCQ	0.00	0.00	0.00
BLS_URBN_URBAN	0.00	0.00	0.00
BLS_URBN_RURAL	0.00	0.00	0.00
EDUC_REF_NEVER	0.00	0.00	0.00
EDUC_REF_FIRST	0.00	0.00	0.00
EDUC_REF_SECOND	0.00	0.00	0.00
EDUC_REF_HIGH	0.00	0.00	0.00
EDUC_REF_COL	0.00	0.00	0.00
EDUC_REF_ASSC	0.00	0.00	0.00
EDUC_REF_UDGRD	0.00	0.00	0.00
EDUC_REF_GRD	0.00	0.00	0.00
REF_RACE_WH	0.00	0.00	0.00
REF_RACE_BLK	0.00	0.00	0.00
REF_RACE_NA	0.00	0.00	0.00
REF_RACE_AS	0.00	0.00	0.00
REF_RACE_PI	0.00	0.00	0.00
REF_RACE_MULTI	0.00	0.00	0.00
SEX_REF_MALE	0.00	0.00	-0.61
SEX_REF_FEMALE	0.00	0.00	0.61

Figure 10: First 3 PCs for sparse PCA

Surprisingly, sparse PCA does not give good biplots. Unsure of whether there is yet a good R package to be developed or there exists mathematical explanation, we move on to the more useful plot of PC1 vs. PC2, since now they account for 71.27% of total variance and has reality interpretations. The plot is demonstrated below, with points representing the data samples.

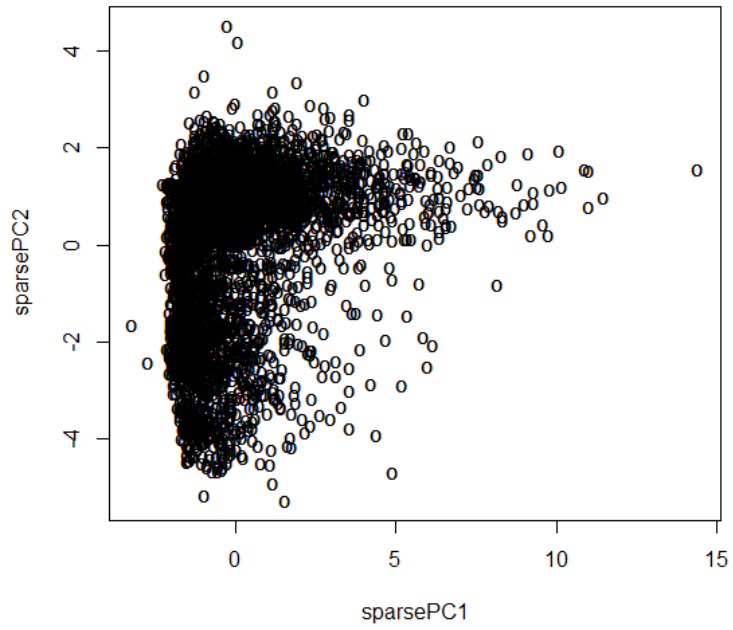
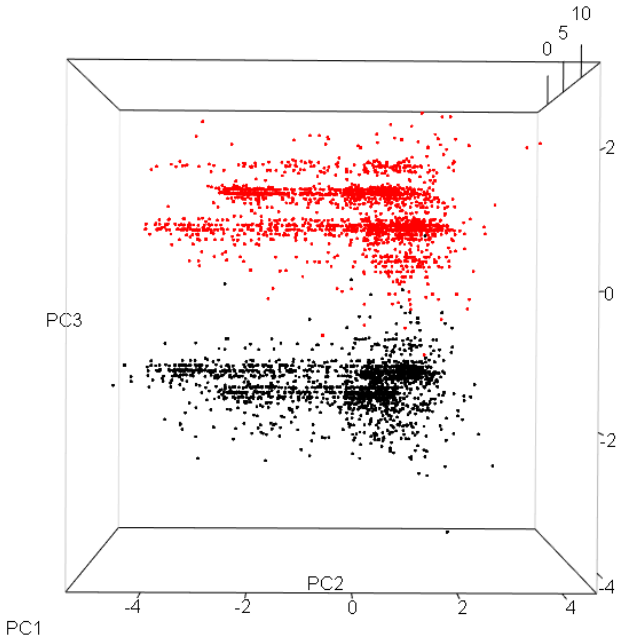


Figure 11: PC1 vs. PC2 for sparse PCA

It is difficult to produce a plot of variables as before, since the outputs of nsprcomp are based on that of prcomp. The plot shows different dispersion pattern from usual PCA, with variance along PC1 significantly larger. It also seems that PC1 has not been centered.



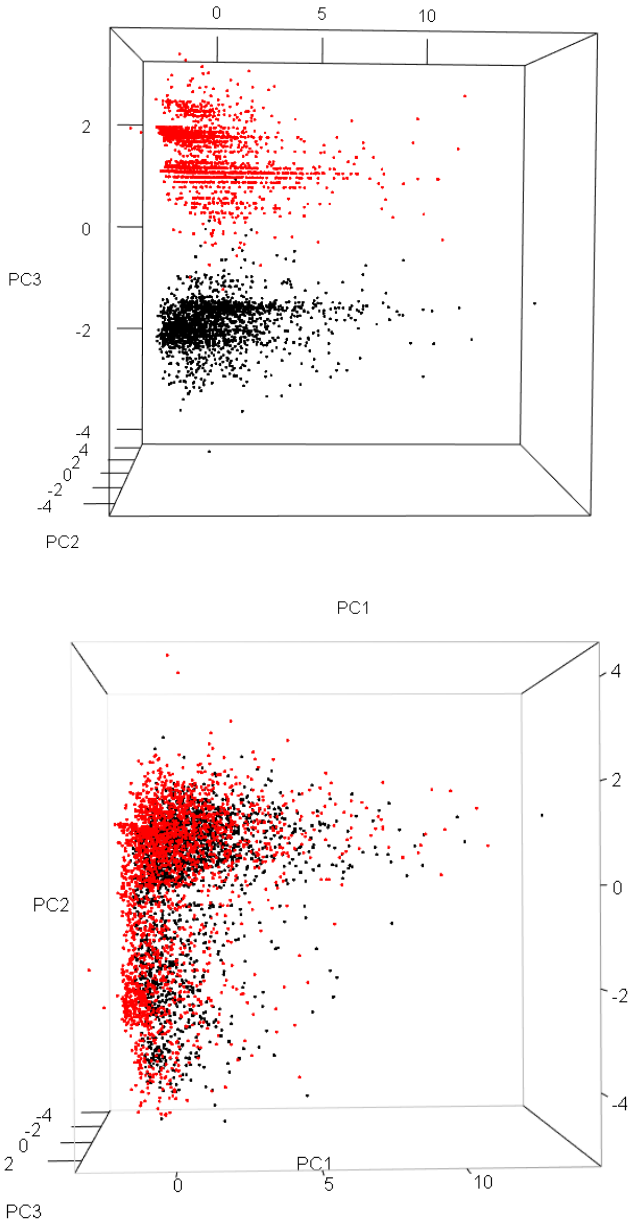


Figure 12: 3D representation of PC1, PC2 and PC3 in sparse PCA

Given above are the 3D PC1 – PC2 – PC3 projections of the dataset with black indicating male and red indicating female. We can observe similar characteristics as the outputs of usual PCA, namely overlap of male and female in PC1-PC2 plane, and clear segregation along PC3, which represents the contrast. More interestingly there seem to be some stratification of the sample points along PC3. This is possibly due to the influence of number of vehicles owned.

In conclusion, sparse PCA made interpretation a lot easier and revealed details PCA did not uncover. The three PCs produced by sparse PCA indicates related variables that are more significant, providing us with the first criterion for further selection and grouping.

4. Cluster Analysis

4.1 Objectives

After thorough discussion in section IV on the specific pair of Education vs. Race, we return to the first main question of variable grouping, significance and interaction. Based on the results from section III, PCA, we know that the intuitive grouping of variables in to characteristics, income and expenditure still holds. Furthermore, sparse PCA has revealed some important variables that coincides with my intuition, which is supported by findings from the usual PCA. It seems that further details can be explored by choosing the more important and informative variables and grouping variables of the sample points accordingly to reveal implicit interactions. Therefore, we arrive at the multivariate technique of cluster analysis. We hope to find primarily clusters of the variables based on distances such that it either reinforces or contradicts my intuitive classification of variables. Moreover, the process may well indicate the relative significance of some variables. Since cluster analysis is a useful and well-developed technique, we will compare several different approaches and provide comments. Clustering of variables is an essential step before advancing to dependencies between variables

4.2 Cluster of Variables

In general, cluster analysis is for the clustering of sample points in the dataset instead of the variables. The distances which provide the criterion for clustering are usually measured between data points. However, for this survey there are two problems. First of all, the data points are just individual Consumer Units marked by an ID number, so grouping them do not have huge significance. Since the interest is not in results for areas such as marketing, where the characteristics of special subgroups of respondents are important, there is less need for clustering the data points. Secondly, there are 4524 data points. While offering a sufficient sample size, it makes drawing conclusions when looking at the sample points and their distances difficult. In fact, an attempt was made at clustering the sample points. The original goal was to disregard the fact that the dataset does not perfectly satisfy the multivariate normality assumption for classification maximum likelihood methods and use the Mclust function in R to find an appropriate number of clusters for other method such as K-means. However, the result was one cluster for the whole dataset. The multivariate normality assumption should not be disregarded. On the other hand, using K-means method directly and selecting 4 clusters based on the scree plot of 9 clusters, we succeeded in clustering all the data points into the most appropriate four clusters. However, it was very difficult to find common traits in each cluster. Also, given that many variables, such as FOODCQ, FDHOMECQ and FDAWAYCQ have significant correlations (as revealed in section II), we have the effect of noise variables in computing the distances between points. On the other hand, variable clustering is used for finding collinearity, redundancy, and separating variables into clusters that can be scored as a single variable. It is also helpful as a method for data reduction. Considering these arguments, we will proceed to discussions on clustering of variables.

In R, there are several packages that aim at providing robust cluster of variables. One popular package is ClustOfVar **[4]**, which includes a function hclustvar that provides ascendant hierarchical clustering of a set of variables based on the decrease in homogeneity for the cluster being merged. This method centers on the first principle component produced by PCAmix, which was developed to incorporate a mixture of categorical and continuous variables. While the current dataset has been reorganized to be completely continuous, this is a very powerful tool, especially since we have verified in section III that PCA produces reasonable results. Below is the cluster dendogram.

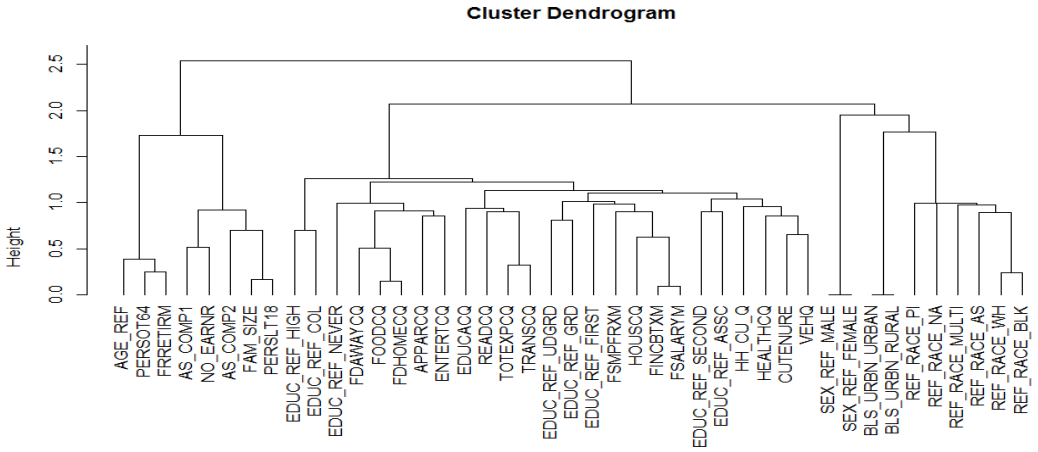
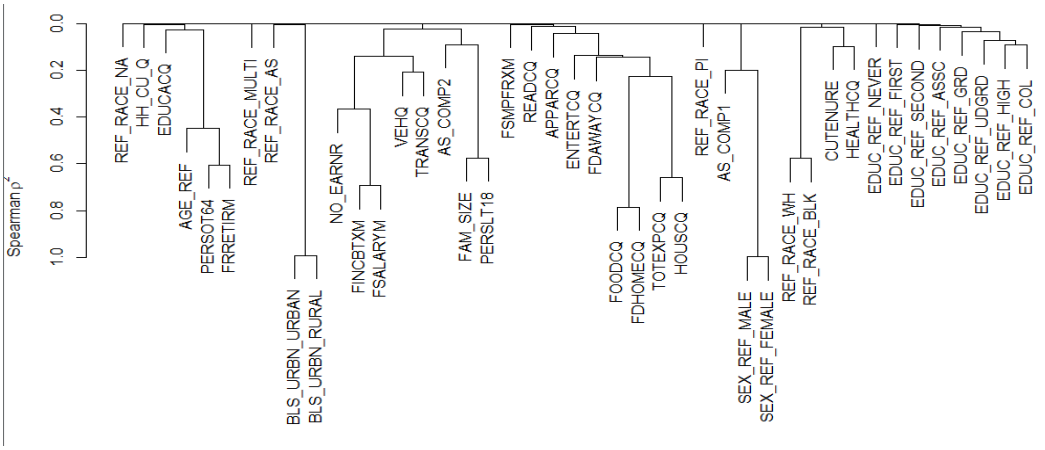


Figure 13: Dendogram for cluster analysis using *ClustOfVar*

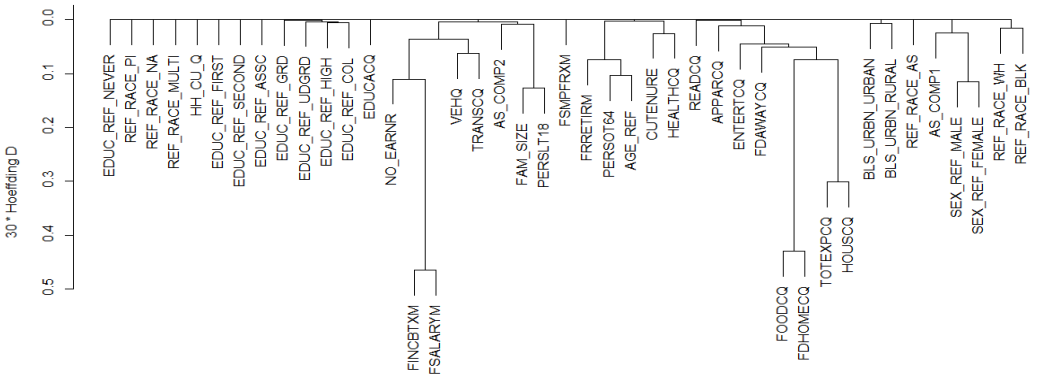
The results are easy to interpret and matched intuition. Age, number of seniors, and retirement income are in a cluster, echoing the second PCs of the sparse PCA in section III. Food related expenditures form a small cluster, as do education, reading, transportation and the total expenditure. Not surprisingly, references to having an undergraduate or graduate degree is close to income, housing expenditure and salary. After all, the level of education received is a crucial factor in determining the lifestyle and incomes of an individual. Note also that the SEX_REF_MALE and SEX_REF_FEMALE are in the same cluster right from the start. This makes sense since they are the opposite of each other. Same explains the urban references. On the more general level, observe how there are in general three main clusters. From left to right, each cluster at height 2.0 represents age and family composition, education, income and expenditures, and demographical categories of gender, race and geography. This is different from the intuitive grouping into demographical characters, income summary and expenditure summary. The explanation is that clustering of variables by hclustvar is based on homogeneity for the clusters when merging in a hierarchical fashion. Therefore, similarities between the variables become the deciding factor, and the pattern of response is compared between the variables. The criterion is no longer the intuitive sociological classifications. The clustering performed is useful for setting the Canonical Correlation Analysis later investigations of the dependencies between groups of variables. Note that Cluster Analysis and Canonical Correlation Analysis are not competitive techniques, but complementary methods focusing separately on similarity and dependency.

Although ClustOfVar is powerful, it is wise to implement other packages for variable clustering and compare the results. Another extensive package for multivariate analysis in

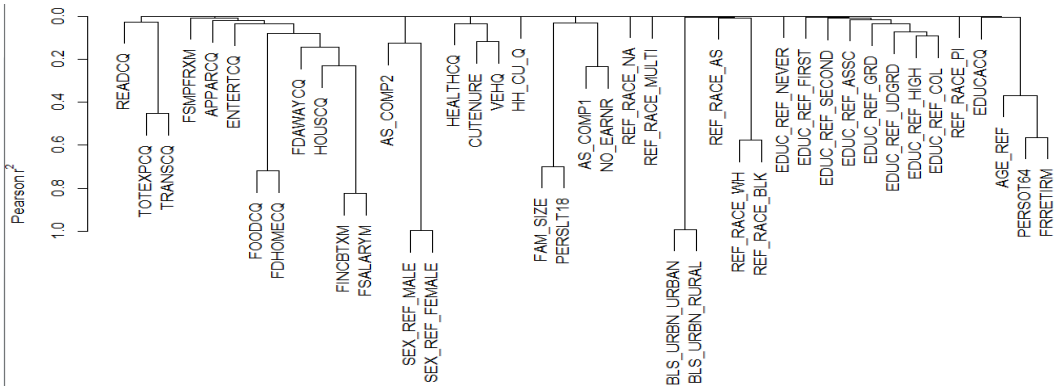
general is the Hmisc **[5]**. Hmisc includes the function varclus, which conducts hierarchical cluster analysis on variables, using the Hoeffding D statistic, squared Pearson or Spearman rho correlations. It is important to note that the function conducts pair-wise deletion of NA values implicitly, and the default aggregation criterion is default is squared Spearman correlation coefficients. The Spearman rho correlation coefficients are used for detecting monotonic but nonlinear relationships, while Hoeffding's D statistic is sensitive to many types of dependence, including highly non-monotonic relationships. Note that the difference between Spearman's correlation and Pearson's correlation is that one is more suitable for ordinal scales and the other for interval scales. Spearman is better for detecting monotonic relationships, while Pearson is better for linear relationships. However, the actual mathematical justification is beyond these explanations. The best approach is to compare all three methods.



varclus(as.matrix(fmlifull), similarity = "spearman")



varclus(as.matrix(fmlifull), similarity = "hoeffding")



varclus(as.matrix(fmlifull), similarity = "pearson")

Figure 14: Dendogram for using Hmisc cluster analysis with Hoeffding D statistic, squared Pearson or Spearman rho correlations.

In general, although still hierarchical clusters, we can no longer observe the clear levels of aggregation according to height as in ClustOfVar. All three have the tendency of grouping all variables into one cluster in total. In particular, this is apparent with Hoeffding's D. However, at the more detailed level, these methods have been successful in selecting the significant variables and clustering on a smaller scale. For example, FOODCQ and FDHOMECQ, AGE_REF, PERSOT64 and FRRETIRM, FAM_SIZE and PERSLT18, FINCBTXM and FSALARYM, as well as TOTEXPCQ, HOUSCQ and TRANSCQ have each been grouped in all three. This reaffirms the results from ClustOfVar. This provides a strong basis for solving a challenge presented in CCA due to multi-collinearity between these similar variables.

The limitations imposed by mathematical implications of each criterion is a challenge. However, the goals of this section are to verify to the best of my abilities grouping of variables, reveal implicit connections, and select suitable clustering for the following section of CCA. Then the most useful approach would be simply compare the results given. This is a rule of thumb in data analysis: Keep the objective in mind and compare different results to verify.

5. Canonical Correlation Analysis

5.1 Objectives

The last section is dedicated to revealing relationships between groups of variables. Much effort has been made in previous sections through PCA and Cluster Analysis to obtain meaningful grouping of variables. The natural question to follow would be investigating how groups of variables influence each other as groups instead of individual variables. Canonical Correlation Analysis is adopted to find the linear combination within each group that achieves the maximum possible correlation between two groups. This section

also utilizes previous preparations made in sections III and V on significant variables and reasonable grouping. This shed light on important relationships such as interaction between income and expenditure as groups with all variables contributing to the analysis. The results have significant macroeconomics meaning in promoting healthy consumption within the society and stimulating the national economy.

5.2 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) requires the user to provide two groups of variables, then provides the linear combination within each group that will give the largest correlation between the two linear combinations. Therefore, an important task is to find the appropriate groups. This should be based on the research question, in this case being: Is there dependency between the income variables and expenditure variables in the **fmlifull** dataset? We would like to include as much variables as possible to avoid loss of information and negligence of implicit influence between and within each group. Therefore, in the first attempt after standardizing the variables we include all the 4 income variables and 11 expenditure variables.

```
> round(sqrt(eigen(e1)$values), digits = 2)
[1] 0.68 0.25 0.14 0.09
> round((eigen(e1)$vectors), digits = 2)
      [,1]  [,2]  [,3]  [,4]
[1,]  0.95  0.71 -0.66 -0.06
[2,]  0.30 -0.67  0.67 -0.34
[3,] -0.01  0.12  0.24 -0.25
[4,]  0.07 -0.16  0.24  0.90
```

```
> sqrt(eigen(e2)$values)
 [1] 6.885289e-01+0.000000e+00i 2.520059e-01+0.000000e+00i
 [3] 1.428474e-01+0.000000e+00i 9.292323e-02+0.000000e+00i
 [5] 0.000000e+00+2.879308e-03i 1.946009e-06+0.000000e+00i
 [7] 1.160737e-06+0.000000e+00i 1.297464e-07+0.000000e+00i
 [9] 1.275201e-08+0.000000e+00i 7.107810e-09+8.250445e-09i
[11] 7.107810e-09-8.250445e-09i
> round((eigen(e2)$vectors), digits = 2)
          [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]
 [1,]  0.00+0i  0.00+0i  0.00+0i  0.00+0i  0.00+0i -0.12+0i -0.16+0i
 [2,] -0.76+0i -0.76+0i  0.76+0i -0.76+0i -0.76+0i -0.16+0i -0.11+0i
 [3,]  0.50+0i  0.50+0i -0.50+0i  0.50+0i  0.50+0i  0.86+0i  0.86+0i
 [4,]  0.43+0i  0.43+0i -0.43+0i  0.43+0i  0.43+0i -0.46+0i -0.46+0i
 [5,]  0.00+0i  0.00+0i  0.00+0i  0.00+0i  0.00+0i  0.09+0i  0.09+0i
 [6,]  0.00+0i  0.00+0i  0.00+0i  0.00+0i  0.00+0i  0.01+0i  0.01+0i
 [7,]  0.00+0i  0.00+0i  0.00+0i  0.00+0i  0.00+0i  0.06+0i  0.06+0i
 [8,]  0.00+0i  0.00+0i  0.00+0i  0.00+0i  0.00+0i -0.04+0i -0.03+0i
 [9,]  0.00+0i  0.00+0i  0.00+0i  0.00+0i  0.00+0i -0.01+0i -0.01+0i
[10,]  0.00+0i  0.00+0i  0.00+0i  0.00+0i  0.00+0i  0.02+0i  0.03+0i
[11,]  0.00+0i  0.00+0i  0.00+0i  0.00+0i  0.00+0i  0.05+0i  0.05+0i
          [,8]     [,9]       [,10]       [,11]
 [1,]  0.18+0i -0.19+0i -0.15+0.01i -0.15-0.01i
 [2,]  0.09+0i -0.07+0i -0.07+0.04i -0.07-0.04i
 [3,] -0.85+0i  0.87+0i  0.76+0.00i  0.76+0.00i
 [4,]  0.47+0i -0.34+0i -0.42+0.06i -0.42-0.06i
 [5,] -0.10+0i -0.05+0i -0.02-0.12i -0.02+0.12i
 [6,] -0.01+0i  0.05+0i  0.03-0.06i  0.03+0.06i
 [7,] -0.07+0i  0.27+0i  0.33+0.15i  0.33-0.15i
 [8,]  0.04+0i -0.08+0i -0.05-0.01i -0.05+0.01i
 [9,] -0.02+0i  0.01+0i  0.00-0.02i  0.00+0.02i
[10,] -0.04+0i -0.02+0i  0.17+0.15i  0.17-0.15i
[11,] -0.05+0i  0.00+0i  0.00+0.00i  0.00+0.00i
```

It is surprising at first to observe the eigenvalues, as the square of the canonical correlations, are complex. However, this is a good example of multiple mistakes in utilizing CCA. First and foremost, there should be the same amount of variables in both

groups. Note that the first four eigenvalues of eigen1 and eigen2 are the same, because they are the square of the canonical correlations. However, the rest of the eigenvalues of eigen2 is not. It does not make sense to use 4 variables to explain 11. Of course, the complex values are not only caused by the above, and there are other issues at hand. The variables in the income group are based on income in the last twelve months, while the expenditure group variables are based on the last three months. They measure different periods, so even though that should not cause computation problems it must be carefully interpreted. More importantly, the correlation matrix in section II has revealed that high correlation exist within the income group and expenditure group. For example, there is a 0.85 correlation between FOODCQ and FDHOMECQ. This leads to multi-collinearity, which influences the performance of CCA.

Therefore, improvements must be made to improve the performance of CCA. This is where previous grouping results are useful. Instead of 11 expenditure variables we choose 4 most significant ones that have fewer multi-collinearity concerns and form a group, then do CCA against the 4 income variables. The selection criterion is primarily from Cluster Analysis of Variables using ClustOfVar. See below for the enlarged section of interest.

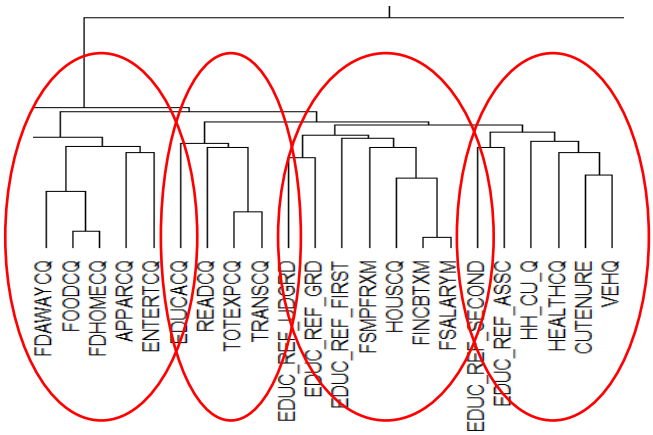


Figure 15: Enlarged section of cluster analysis using *ClustOfVar*

We can observe 4 distinct clusters from this part of the dendogram which contains all of the 11 expenditure variables. Within a cluster the variables have higher similarity. Therefore it makes sense to choose one variable from each cluster. Then in order to determine which varaible is most likely to be the most significant within each cluster, we look back to the PCs in section III PCA, in particular sparse PCA. We choose the ones present in the first three PCs, which give FOODCQ, TOTEXPCQ, HOUSCQ. The last cluster contains only one expenditure variable, HEALTHCQ. This forms the group of 4 expenditure variables needed for CCA. Results of the second attempt are given below.

```
> round(sqrt(eigen(e1)$values), digits = 2)
[1] 0.59 0.19 0.04 0.00
> round((eigen(e1)$vectors), digits = 2)
      [,1]  [,2]  [,3]  [,4]
[1,]  0.99  0.57 -0.57 -0.68
[2,]  0.13 -0.58  0.73  0.66
[3,]  0.05  0.57  0.33  0.14
[4,] -0.03  0.11 -0.19  0.28
```

```
> round(sqrt(eigen(e2)$values), digits = 2)
[1] 0.59 0.19 0.04 0.00
> round((eigen(e2)$vectors), digits = 2)
     [,1]  [,2]  [,3]  [,4]
[1,] 0.79  0.02  0.61  0.60
[2,] 0.31 -0.16  0.18 -0.79
[3,] 0.53 -0.31 -0.75  0.06
[4,] 0.05  0.94 -0.18 -0.09
> round(cor(u,x),digits = 2)
      [,1]  [,2]  [,3]  [,4]
[1,]  1.00  0.92 -0.14  0.33
[2,]  0.01 -0.33  0.90  0.32
[3,] -0.06  0.18  0.41 -0.80
[4,]  0.00  0.09  0.02  0.39
> round(cor(v,y),digits = 2)
     [,1]  [,2]  [,3]  [,4]
[1,] 0.93  0.73  0.86  0.38
[2,] 0.08 -0.05 -0.17  0.91
[3,] 0.26  0.17 -0.49 -0.14
[4,] 0.23 -0.66  0.05 -0.10
```

This time the outputs have avoided complex eigenvalues. We observe the 0.59 correlation between the first linear combination:

U1 = 0.99FINCBTXM + 0.57FSALARYM – 0.57FRRETIRM – 0.68FSMPFRXM.

And the second linear combination:

V1 = 0.79 TOTEXPCQ + 0.02FOODCQ + 0.61 HOUSCQ + 0.60 HEALTHCQ

This correlation is quite high, and the result is consistent with my intuitive understanding that expenditure variables should have dependence on the income variables. More information could be gained if we look into the underlying meaning of U1 nad V1. However, it is important to note that a common pitfall for many inexperienced users would be to mistakenly take the linear combination coefficients of U1 and V1 and interpret them as the coefficients to reflect the reationship with the original variables. As shown above, that is not the case. In fact, using correlation coefficients, we have U1 as a measure of high income before tax and salary, but low retirement and self-employment income. V1 is a combination of total expenditure and expenditure on housing and health. Food seem not important. This makes sense, as people with high income and salary have the lower percentage of total spending on necessities such as food and more on housing and health. This is where CCA verifies a socioeconomic phenomenon. CCA have also verified that previous work on clustering makes sense in applications to CCA. However, CCA does have significant shortcomings. Trying to compare groups of different numbers of variables seem not to work, and other implicit influences not included in either of the two groups are not accounted for. The basic requirement of having appropriate, meaningful groups prior to conducting CCA may present difficulty in real life, especially for large datasets such as in genetics or clinical trails where many factors are biomarkers and do not have distinct, easy-to-interpret meanings for grouping intuitively. Most importantly, CCA uses linear combinations of original variables to show the between group correlation, but the relationship between the linear combinations and the original variables is best reflected in the correlations instead of the linear coefficients. This adds complexity to the interpretation and leaves much space for the search of a more elegant procedure of reflecting group-to-group dependencies in multivariate analysis.

In conclusion, this section on CCA has utilized the results from previous PCA and Cluster Analysis, proving them to be reasonable with sensible results produced. Interesting challenges of complex eigenvalues in CCA served as a good example of some common pitfalls and shortcomings of CCA. The relatively high correlation between U1: higher income stable employment and V1: bulk expenditure is consistent with economic intuition. It is reasonable to argue that U1 and V1 reflect the consumption and income of the U.S. middle class. It would also be productive if this correlation, which reflect the dependency of expenditure based on income, could act as a factor of consumer confidence. If the correlation is low, that may suggest either a frenzy or a crisis. While this report does not make comparison with data from previous periods to establish the connection, this is something worth further discussion.

6. Correspondence Analysis

6.1 Objectives

Next we briefly suspend the discussion on the first main question of variable grouping, significance and interaction. In this section we will answer the second main objective investigating pairs of important variables and their socioeconomics significance. In particular, is the educational level related to race?

Racial equality is part of the foundation of modern American society. However, in recent years a heated debate erupted on whether protection of minority rights should be based on sacrifice of other racial groups, from the white majority to other minorities such as Asian Americans. This has led to two major events: Protest against the California legislation of SCA 5 and unexpected results in 2016 presidential election. In particular, the SCA 5 event advocated using race and nationality as admission criteria in California's higher educational systems, reducing admissions to Asian Americans in favor of Latino Americans and Native Americans. The cornerstone of the supporters' argument is Asian Americans have admission percentages disproportional to their population percentage. Although not the original objective of the survey, this section aims to utilize the **fmlifull** dataset to support or discredit this claim.

Since both EDUC_REF and REF_RACE are categorical variables, natural questions that arise include: What is the dependence between the levels of EDUC_REF and REF_RACE? What is the similarity between two variables? What about between different levels of the same variable? To answer these questions, the best approach is correspondence analysis.

6.2 Contingency table and Mosaic plot

Before conducting correspondence analysis, we need to visually verify the dependence between two variables and the significance of each level. The first step is to make a contingency table. This is straight forward, as the two categorical variables have already been expanded into 7 and 6 columns of '0's and '1's in section II. This creates the following Table IV.

Table IV: Contingency table

	WHITE	BLK	NAT AM	ASN	PAC IS	MULTI
NEVER	7	1	0	1	0	0
FIRST	113	7	0	9	0	0
SECOND	264	74	1	9	0	5
HIGH	860	115	7	33	4	9
COL	777	117	5	24	4	17
ASSC	332	45	3	11	1	4
UNDGRD	701	80	4	70	3	4
GRD	430	36	0	46	1	7

Also the mosaic plot which provides visual demonstration of the contingency table, but also shows significance based on Pearson residuals in colors:

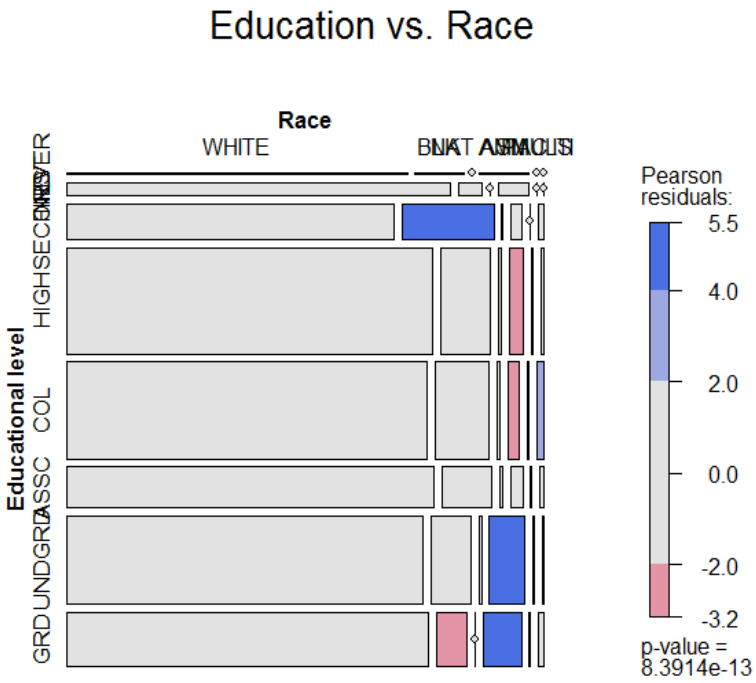


Figure 16: Mosaic plot based on contingency table

The categories, "Never" (no education) and "First" (only elementary education) in Education and "Native Americans", "Pacific Islanders" and "Multi-race" in Races. In the plot, blue means there are more sample points in that cell than would be expected under the assumed independence between variables. Red means there are fewer samples than would have been expected. It seems that in graduate education African Americans are underprivileged, while Asian Americans tend to b more likely to have a bachelor or graduate degree. Native Americans also tend to have a higher concentration in the category of only have a high school education. However, we will next cluster the insignificant categories to attain a more elegant plot.

We substitute "Native Americans" with the new category "Other", which is the sum of "Native American", "Pacific Islanders" and "Multirace". Also combine "Never" and "First". This is the new contingency table and Mosaic Plot.

Table V: Contingency table after truncation

	WHITE	BLK	OTHER	ASN
NEVER or FIRST	120	8	0	10
SECOND	264	74	6	9
HIGH	860	115	20	33
COL	777	117	26	24
ASSC	332	45	8	11
UNDGRD	701	80	11	70
GRD	430	36	8	46

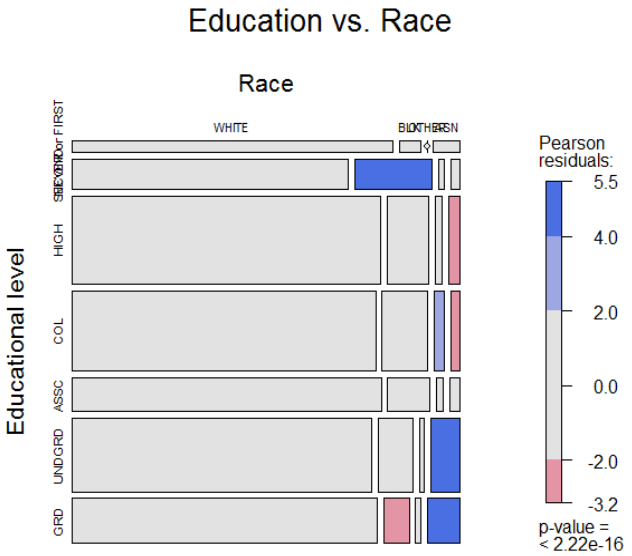


Figure 17: Mosaic plot based on contingency table after truncation

Apparently, while African Americans are fewer than expected on average in the higher education category, there is no indication that Native Americans, Pacific Islanders and Multi-race as a group together share the situation. This may be caused by the survey being conducted nation-wide instead of just in California, also not specifying Latino/Hispanic as a category. However, the results do show Asian Americans having a lead in obtaining higher education degrees. But should simply that be the reason for abandoning the ban on racial consideration factors? This analysis does not answer.

6.3 Correspondence Analysis

After thoroughly investigating the contingency table and Mosaic plot, we feel confident in using correspondence analysis based on the most appropriate table to reveal the relationship between Education categories, between Race categories and the interaction between Education and Race.

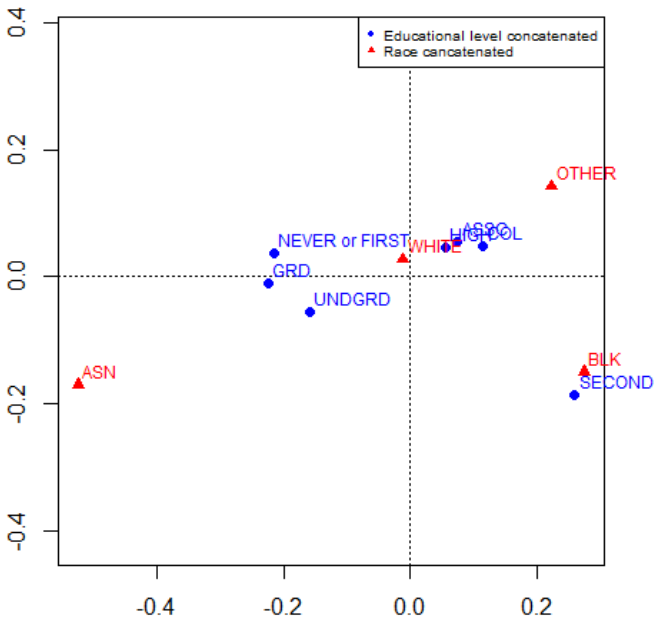


Figure 18: Correspondence plot, education vs. race

First we look at the similarity between Race profiles. Apparently, "Asian Americans" is far from other categories. While not exactly a cluster, "White", "Black" and "Other" are close, meaning more similarity. As for the categories of Education, "Associate", "Some College" and "High School" are very close, indicating very similar profiles. Surprisingly, "Graduate" and "Never or First" are closer than "Graduate" and "Undergraduate". This may be due to both "Graduate" and "Never or First" have very few sample points, and in reality few people choose a graduate education after obtaining a Bachelor's degree. Finally, the most important interaction. "Black" and "Second" are a group, which echoes the observation from the mosaic plot that African Americans have more sample points in the category of secondary education only than independence assumption expects. Similarly, "White" may occur more frequently with an associate degree, high school degree than a Bachelor's degree. This reinforces the recent trend in US that majority of young people are satisfied without a full college education for a Bachelor's degree.

7. Conclusion

This project is an endeavor at applying appropriate multivariate techniques to a real life economics survey dataset. Undoubtedly it was challenging, not only in selecting the correct methods, but also on finding the right questions and "taming" the original dataset before even beginning analysis. To begin with, selecting more important variables from 807 and gaining a general idea of each variable in the process was a huge task for anyone unfamiliar to the dataset. In a more rigorous discussion, the missing values would also become a much more significant problem. Moreover, methods such as PCA and CCA have encountered unexpected problems with this dataset, so adjustments or different methods have been adopted and results are compared. That being said, each section has

resulted in informative interpretations. Reasonable grouping of variables and relationship between groups have been discussed. The special section on the pair, Education and Race, have also provided interesting insight on recent political discussions.

Acknowledgements

The author would like to show his gratitude for the Bureau of Labor Statistics for providing the datasets on which this research was based. Furthermore, guidance from Prof. Weiyin Loh of the department of statistics, University of Wisconsin Madison has been much appreciated. Prof. Loh, who has done extensive research on the PUMD dataset, first introduced the author to this topic. His advice on decisions involving missing values and incorporating sparse PCA, among other suggestions, were essential to the improvement of this research project.

References

http://www.bls.gov/cex/pumd.htm

Loh, W., Eltinge, J., Cho, M., & Li, Y. (2016). Classification and regression trees methods for incomplete data from sample surveys. Web. https://arxiv.org/pdf/1603.01631v1.pdf. doi: 2016arXiv160301631L

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics, 15* (2), p. 265–p. 286. DOI: 10.1198/106186006X113430

Marie Chavent, Vanessa Kuentz, Benoit Liquet and Jerome Saracco (2013). ClustOfVar: Clustering of variables. R package version 0.8. https://CRAN.R-project.org/package=ClustOfVar

Frank E Harrell Jr, with contributions from Charles Dupont and many others. (2016). Hmisc: Harrell, Miscellaneous. R package version 4.0-0. https://CRAN.R-project.org/package=Hmisc

Appendix A: Correlation matrix of fmlifull

	FSMPFRXM	TOTEXPCQ	FOODCQ	FDHOMECQ	FDAWAYCQ	HOUSCQ	APPARCQ	TRANSCQ	HEALTHCQ	ENTERTCQ	EDUCACQ	READCQ
HH_CU_Q	-0.02	-0.03	-0.05	-0.06	-0.02	-0.05	0.02	0.00	-0.04	-0.03	0.04	-0.01
AS_COMP1	0.06	0.19	0.26	0.26	0.16	0.14	0.03	0.08	0.08	0.06	0.03	0.00
AS_COMP2	0.03	0.11	0.18	0.23	0.04	0.09	0.05	0.06	0.05	0.03	0.01	-0.01
CUTENURE	-0.08	-0.24	-0.21	-0.18	-0.15	-0.19	-0.04	-0.09	-0.19	-0.14	0.01	-0.04
FAM_SIZE	0.05	0.20	0.33	0.38	0.14	0.19	0.08	0.08	0.05	0.05	0.00	-0.03
NO_EARNR	0.13	0.27	0.28	0.25	0.20	0.23	0.10	0.09	0.04	0.08	0.05	0.01
PERSLT18	0.03	0.10	0.22	0.27	0.08	0.14	0.07	0.03	-0.01	0.02	-0.01	-0.04
PERSOT64	-0.06	-0.08	-0.07	-0.04	-0.08	-0.10	-0.06	0.00	0.13	-0.01	-0.07	0.05
VEHQ	0.09	0.28	0.20	0.17	0.15	0.15	0.07	0.17	0.18	0.16	0.03	0.08
AGE_REF	-0.01	-0.06	-0.09	-0.07	-0.09	-0.08	-0.08	-0.01	0.13	0.01	-0.07	0.06
FINCBTXM	0.38	0.55	0.43	0.31	0.41	0.50	0.20	0.14	0.23	0.23	0.09	0.14
FSALARYM	0.09	0.50	0.40	0.28	0.38	0.48	0.17	0.12	0.15	0.19	0.09	0.08
FRRETIRM	-0.08	-0.07	-0.07	-0.05	-0.07	-0.09	-0.06	0.03	0.13	0.02	-0.06	0.05
FSMPFRXM	1.00	0.20	0.15	0.11	0.13	0.14	0.09	0.04	0.13	0.11	0.03	0.05
TOTEXPCQ	0.20	1.00	0.57	0.45	0.49	0.67	0.28	0.67	0.37	0.37	0.30	0.22
FOODCQ	0.15	0.57	1.00	0.85	0.79	0.51	0.24	0.14	0.27	0.29	0.06	0.11
FDHOMECQ	0.11	0.45	0.85	1.00	0.34	0.41	0.17	0.11	0.24	0.21	0.04	0.08
FDAWAYCQ	0.13	0.49	0.79	0.34	1.00	0.44	0.23	0.13	0.21	0.26	0.07	0.10
HOUSCQ	0.14	0.67	0.51	0.41	0.44	1.00	0.23	0.15	0.23	0.28	0.10	0.11
APPARCQ	0.09	0.28	0.24	0.17	0.23	0.23	1.00	0.06	0.08	0.14	0.09	0.09
TRANSCQ	0.04	0.67	0.14	0.11	0.13	0.15	0.06	1.00	0.06	0.08	0.02	0.16
HEALTHCQ	0.13	0.37	0.27	0.24	0.21	0.23	0.08	0.06	1.00	0.14	0.01	0.10
ENTERTCQ	0.11	0.37	0.29	0.21	0.26	0.28	0.14	0.08	0.14	1.00	0.02	0.09
EDUCACQ	0.03	0.30	0.06	0.04	0.07	0.10	0.09	0.02	0.01	0.02	1.00	0.00
READCQ	0.05	0.22	0.11	0.08	0.10	0.11	0.09	0.16	0.10	0.09	0.00	1.00
BLS_URBN_URBAN	0.02	0.07	0.07	0.04	0.08	0.09	0.03	0.01	0.02	0.04	0.01	0.02
BLS_URBN_RURAL	-0.02	-0.07	-0.07	-0.04	-0.08	-0.09	-0.03	-0.01	-0.02	-0.04	-0.01	-0.02
EDUC_REF_NEVER	-0.01	-0.02	-0.02	-0.01	-0.02	-0.02	0.00	0.00	-0.01	-0.02	-0.01	-0.01
EDUC_REF_FIRST	-0.01	-0.06	-0.03	0.00	-0.06	-0.06	-0.02	-0.01	-0.05	-0.05	-0.02	-0.03
EDUC_REF_SECOND	-0.04	-0.11	-0.09	-0.04	-0.11	-0.12	-0.04	-0.01	-0.06	-0.05	-0.03	-0.03
EDUC_REF_HIGH	-0.06	-0.10	-0.09	-0.05	-0.11	-0.12	-0.07	0.02	-0.05	-0.08	-0.06	-0.04
EDUC_REF_COL	-0.03	-0.07	-0.06	-0.06	-0.05	-0.05	-0.01	-0.03	-0.04	-0.04	0.00	-0.04
EDUC_REF_ASSC	0.00	-0.02	-0.01	0.00	-0.03	-0.03	-0.02	-0.01	0.02	0.02	-0.02	-0.01
EDUC_REF_UDGRD	0.09	0.14	0.11	0.06	0.12	0.14	0.05	0.01	0.06	0.11	0.06	0.03
EDUC_REF_GRD	0.05	0.20	0.17	0.09	0.20	0.22	0.10	0.02	0.10	0.09	0.06	0.10
REF_RACE_WH	0.03	0.02	0.00	0.00	0.00	-0.01	-0.03	0.01	0.08	0.04	-0.01	0.05
REF_RACE_BLK	-0.04	-0.09	-0.09	-0.08	-0.07	-0.07	-0.01	-0.02	-0.11	-0.07	-0.03	-0.05
REF_RACE_NA	0.00	0.01	0.01	0.01	0.01	0.02	-0.01	0.00	0.04	0.01	0.00	0.02
REF_RACE_AS	0.01	0.08	0.11	0.09	0.10	0.10	0.04	0.02	0.01	0.02	0.06	-0.01
REF_RACE_PI	-0.01	0.00	0.00	0.01	0.00	0.02	-0.01	-0.01	0.00	0.01	-0.01	0.00
REF_RACE_MULTI	0.01	0.03	0.03	0.03	0.02	0.02	0.03	0.00	0.00	0.02	0.00	0.01
SEX_REF_MALE	0.01	0.06	0.07	0.03	0.08	0.03	0.00	0.01	0.05	0.01	0.01	0.01
SEX_REF_FEMALE	-0.01	-0.06	-0.07	-0.03	-0.08	-0.03	0.00	-0.01	-0.05	-0.01	-0.01	-0.01

	BLS_URBN_URBAN	BLS_URBN_RURAL	EDUC_REF_NEVER	EDUC_REF_FIRST	EDUC_REF_SECOND	EDUC_REF_HIGH	EDUC_REF_COL
HH_CU_Q	0.02	-0.02	0.00	-0.01	-0.02	-0.04	0.11

AS_COMP1	0.00	0.00	0.01	0.04	-0.01	0.02	-0.04
AS_COMP2	0.03	-0.03	0.01	0.06	0.02	-0.01	-0.02
CUTENURE	0.02	-0.02	0.00	0.05	0.10	0.02	0.09
FAM_SIZE	0.01	-0.01	0.00	0.09	0.00	-0.01	-0.05
NO_EARNR	0.09	-0.09	0.02	-0.01	-0.08	-0.05	-0.03
PERSLT18	-0.01	0.00	-0.01	0.08	0.01	-0.02	-0.02
PERSOT64	-0.06	0.06	0.01	0.02	0.03	0.07	-0.01
VEHQ	-0.02	0.03	-0.01	-0.06	-0.09	0.00	0.00
AGE_REF	-0.05	0.05	0.02	0.06	0.07	0.09	-0.08
FINCBTXM	0.11	-0.11	-0.02	-0.09	-0.15	-0.18	-0.10
FSALARYM	0.11	-0.11	-0.01	-0.08	-0.14	-0.16	-0.10
FRRETIRM	-0.07	0.07	-0.01	-0.01	0.05	0.07	0.01
FSMPFRXM	0.02	-0.02	-0.01	-0.01	-0.04	-0.06	-0.03
TOTEXPCQ	0.07	-0.07	-0.02	-0.06	-0.11	-0.10	-0.07
FOODCQ	0.07	-0.07	-0.02	-0.03	-0.09	-0.09	-0.06
FDHOMECQ	0.04	-0.04	-0.01	0.00	-0.04	-0.05	-0.06
FDAWAYCQ	0.08	-0.08	-0.02	-0.06	-0.11	-0.11	-0.05
HOUSCQ	0.09	-0.09	-0.02	-0.06	-0.12	-0.12	-0.05
APPARCQ	0.03	-0.03	0.00	-0.02	-0.04	-0.07	-0.01
TRANSCQ	0.01	-0.01	0.00	-0.01	-0.01	0.02	-0.03
HEALTHCQ	0.02	-0.02	-0.01	-0.05	-0.06	-0.05	-0.04
ENTERTCQ	0.04	-0.04	-0.02	-0.05	-0.05	-0.08	-0.04
EDUCACQ	0.01	-0.01	-0.01	-0.02	-0.03	-0.06	0.00
READCQ	0.02	-0.02	-0.01	-0.03	-0.03	-0.04	-0.04
BLS_URBN_URBAN	1.00	-1.00	0.01	-0.04	-0.03	-0.10	0.04
BLS_URBN_RURAL	-1.00	1.00	-0.01	0.04	0.03	0.10	-0.04
EDUC_REF_NEVER	0.01	-0.01	1.00	-0.01	-0.01	-0.03	-0.02
EDUC_REF_FIRST	-0.04	0.04	-0.01	1.00	-0.05	-0.10	-0.09
EDUC_REF_SECOND	-0.03	0.03	-0.01	-0.05	1.00	-0.17	-0.16
EDUC_REF_HIGH	-0.10	0.10	-0.03	-0.10	-0.17	1.00	-0.30
EDUC_REF_COL	0.04	-0.04	-0.02	-0.09	-0.16	-0.30	1.00
EDUC_REF_ASSC	0.01	-0.01	-0.01	-0.06	-0.10	-0.18	-0.17
EDUC_REF_UDGRD	0.06	-0.06	-0.02	-0.09	-0.15	-0.28	-0.27
EDUC_REF_GRD	0.04	-0.04	-0.02	-0.07	-0.11	-0.21	-0.20
REF_RACE_WH	-0.09	0.09	-0.01	0.02	-0.06	0.02	0.00

REF_RACE_BLK	0.06	-0.06	0.00	-0.03	0.09	0.00	0.02
REF_RACE_NA	-0.01	0.01	0.00	-0.01	-0.01	0.02	0.00
REF_RACE_AS	0.06	-0.06	0.01	0.02	-0.03	-0.04	-0.06
REF_RACE_PI	0.02	-0.02	0.00	-0.01	-0.02	0.01	0.01
REF_RACE_MULTI	0.03	-0.03	0.00	-0.02	0.01	-0.01	0.04
SEX_REF_MALE	-0.01	0.01	-0.01	-0.01	-0.02	0.01	-0.02
SEX_REF_FEMALE	0.01	-0.01	0.01	0.01	0.02	-0.01	0.02

	EDUC_REF_ASSC	EDUC_REF_UDGRD	EDUC_REF_GRD	REF_RACE_WH	REF_RACE_BLK	REF_RACE_NA	REF_RACE_AS	REF_RACE_PI	REF_RACE_MULTI	SEX_REF_MALE	SEX_REF_FEMALE
HH_CU_Q	-0.03	-0.02	-0.02	0.02	-0.02	-0.01	0.01	-0.01	0.01	-0.07	0.07
AS_COMP1	-0.01	0.01	0.01	0.07	-0.12	0.01	0.03	0.01	0.03	0.40	-0.40
AS_COMP2	0.03	-0.01	-0.02	-0.06	0.01	0.01	0.07	0.01	0.02	-0.36	0.36
CUTENURE	-0.05	-0.07	-0.12	-0.13	0.13	0.01	0.02	0.02	0.02	-0.07	0.07
FAM_SIZE	0.03	-0.01	0.01	-0.01	-0.05	0.02	0.07	0.01	0.02	-0.03	0.03
NO_EARNR	0.04	0.09	0.04	-0.01	-0.05	0.02	0.07	0.02	0.05	0.05	-0.05
PERSLT18	0.03	-0.02	0.01	-0.03	0.00	0.03	0.05	-0.01	0.01	-0.09	0.08
PERSOT64	-0.05	-0.08	0.02	0.08	-0.06	-0.01	-0.04	-0.02	-0.02	0.00	0.00
VEHQ	0.06	0.03	0.03	0.15	-0.16	0.01	-0.04	0.01	0.01	0.12	-0.12
AGE_REF	-0.03	-0.09	0.02	0.08	-0.04	-0.01	-0.07	-0.02	-0.03	-0.01	0.01
FINCBTXM	-0.01	0.22	0.27	0.04	-0.12	0.00	0.09	0.00	0.04	0.10	-0.10
FSALARYM	0.00	0.21	0.24	0.01	-0.10	0.00	0.11	0.01	0.03	0.10	-0.10
FRRETIRM	-0.03	-0.09	-0.02	0.09	-0.05	-0.02	-0.07	-0.02	-0.03	0.00	0.00
FSMPFRXM	0.00	0.09	0.05	0.03	-0.04	0.00	0.01	-0.01	0.01	0.01	-0.01
TOTEXPCQ	-0.02	0.14	0.20	0.02	-0.09	0.01	0.08	0.00	0.03	0.06	-0.06
FOODCQ	-0.01	0.11	0.17	0.00	-0.09	0.01	0.11	0.00	0.03	0.07	-0.07
FDHOMECQ	0.00	0.06	0.09	0.00	-0.08	0.01	0.09	0.01	0.03	0.03	-0.03
FDAWAYCQ	-0.03	0.12	0.20	0.00	-0.07	0.01	0.10	0.00	0.02	0.08	-0.08
HOUSCQ	-0.03	0.14	0.22	-0.01	-0.07	0.02	0.10	0.02	0.02	0.03	-0.03
APPARCQ	-0.02	0.05	0.10	-0.03	-0.01	-0.01	0.04	-0.01	0.03	0.00	0.00
TRANSCQ	-0.01	0.01	0.02	0.01	-0.02	0.00	0.02	-0.01	0.00	0.01	-0.01
HEALTHCQ	0.02	0.06	0.10	0.08	-0.11	0.04	0.01	0.00	0.00	0.05	-0.05
ENTERTCQ	0.02	0.11	0.09	0.04	-0.07	0.01	0.02	0.01	0.02	0.01	-0.01
EDUCACQ	-0.02	0.06	0.06	-0.01	-0.03	0.00	0.06	-0.01	0.00	0.01	-0.01
READCQ	-0.01	0.03	0.10	0.05	-0.05	0.02	-0.01	0.00	0.01	0.01	-0.01
BLS_URBN_URBAN	0.01	0.06	0.04	-0.09	0.06	-0.01	0.06	0.02	0.03	-0.01	0.01
BLS_URBN_RURAL	-0.01	-0.06	-0.04	0.09	-0.06	0.01	-0.06	-0.02	-0.03	0.01	-0.01

EDUC_REF_NEVER	-0.01	-0.02	-0.02	-0.01	0.00	0.00	0.01	0.00	0.00	-0.01	0.01
EDUC_REF_FIRST	-0.06	-0.09	-0.07	0.02	-0.03	-0.01	0.02	-0.01	-0.02	-0.01	0.01
EDUC_REF_SECOND	-0.10	-0.15	-0.11	-0.06	0.09	-0.01	-0.03	-0.02	0.01	-0.02	0.02
EDUC_REF_HIGH	-0.18	-0.28	-0.21	0.02	0.00	0.02	-0.04	0.01	-0.01	0.01	-0.01
EDUC_REF_COL	-0.17	-0.27	-0.20	0.00	0.02	0.00	-0.06	0.01	0.04	-0.02	0.02
EDUC_REF_ASSC	1.00	-0.16	-0.12	0.02	0.00	0.01	-0.03	0.00	0.00	-0.05	0.05
EDUC_REF_UDGRD	-0.16	1.00	-0.19	-0.01	-0.03	0.00	0.08	0.00	-0.03	0.03	-0.03
EDUC_REF_GRD	-0.12	-0.19	1.00	0.01	-0.05	-0.03	0.07	-0.01	0.01	0.03	-0.03
REF_RACE_WH	0.02	-0.01	0.01	1.00	-0.76	-0.15	-0.48	-0.12	-0.22	0.06	-0.06
REF_RACE_BLK	0.00	-0.03	-0.05	-0.76	1.00	-0.02	-0.08	-0.02	-0.04	-0.07	0.07
REF_RACE_NA	0.01	0.00	-0.03	-0.15	-0.02	1.00	-0.02	0.00	-0.01	0.02	-0.02
REF_RACE_AS	-0.03	0.08	0.07	-0.48	-0.08	-0.02	1.00	-0.01	-0.02	0.00	0.00
REF_RACE_PI	0.00	0.00	-0.01	-0.12	-0.02	0.00	-0.01	1.00	-0.01	0.01	-0.01
REF_RACE_MULTI	0.00	-0.03	0.01	-0.22	-0.04	-0.01	-0.02	-0.01	1.00	-0.01	0.01
SEX_REF_MALE	-0.05	0.03	0.03	0.06	-0.07	0.02	0.00	0.01	-0.01	1.00	-1.00
SEX_REF_FEMALE	0.05	-0.03	-0.03	-0.06	0.07	-0.02	0.00	-0.01	0.01	-1.00	1.00

Appendix B: Principle components

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12
Standard deviation	2.4575533	1.75764897	1.57076826	1.53409571	1.36991131	1.34170408	1.21957921	1.1585920	1.13443845	1.10081479	1.09097981	1.08399095
Proportion of Variance	0.1404551	0.07184488	0.05737937	0.05473139	0.04364319	0.04186441	0.03459008	0.0312171	0.02992908	0.02818124	0.02767993	0.02732643
Cumulative Proportion	0.1404551	0.21229996	0.26967933	0.32441072	0.36805390	0.40991832	0.44450840	0.4757255	0.50565458	0.53383582	0.56151575	0.58884218

	Comp.13	Comp.14	Comp.15	Comp.16	Comp.17	Comp.18	Comp.19	Comp.20	Comp.21	Comp.22	Comp.23	Comp.24
Standard deviation	1.04545522	1.01367535	1.0120255	1.00446306	1.00291073	1.00128939	0.98928973	0.97386978	0.95998226	0.94137364	0.93559635	0.91725285
Proportion of Variance	0.02541806	0.02389623	0.0238185	0.02346386	0.02339139	0.02331582	0.02276033	0.02205633	0.02143177	0.02060894	0.02035676	0.01956634
Cumulative Proportion	0.61426024	0.63815646	0.6619750	0.68543883	0.70883022	0.73214605	0.75490638	0.77696271	0.79839448	0.81900341	0.83936017	0.85892651

	Comp.25	Comp.26	Comp.27	Comp.28	Comp.29	Comp.30	Comp.31	Comp.32	Comp.33	Comp.34	Comp.35
Standard deviation	0.89749103	0.89338913	0.84284234	0.82926509	0.77979580	0.76425066	0.73646642	0.68875545	0.582040251	0.479358427	0.449628221
Proportion of Variance	0.01873233	0.01856149	0.01652054	0.01599257	0.01414143	0.01358323	0.01261355	0.01103219	0.007878392	0.005343826	0.004701524
Cumulative Proportion	0.87765884	0.89622033	0.91274087	0.92873345	0.94287488	0.95645811	0.96907166	0.98010385	0.987982244	0.993326069	0.998027594

	Comp.36	Comp.37	Comp.38	Comp.39	Comp.40	Comp.41	Comp.42	Comp.43
Standard deviation	0.1804696732	0.155522619	0.1293241285	0.0759640120	5.520174e-02	4.496607e-02	2.219122e-02	5.188468e-08
Proportion of Variance	0.0007574257	0.000562495	0.0003889472	0.0001341984	7.086587e-05	4.702203e-05	1.145233e-05	6.260511e-17
Cumulative Proportion	0.9987850192	0.999347514	0.9997364614	0.9998706598	9.999415e-01	9.999885e-01	1.000000e+00	1.000000e+00

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15	Comp.16	Comp.17	Comp.18
HH_CU_Q				0.160			0.218		0.122			-0.200			-0.216			
AS_COMP1	-0.181		0.313	-0.174	-0.250													
AS_COMP2	-0.114	0.129	-0.298	-0.299	-0.149							-0.116						
CUTENURE	0.146	0.197		0.134		-0.122	0.293		-0.101					-0.111				
FAM_SIZE	-0.224	0.211		-0.377	-0.241													
NO_EARNR	-0.254	0.233	0.115	-0.130		0.104	-0.147	0.101										
PERSLT18	-0.153	0.261		-0.281	-0.120		0.167				0.112				0.101			
PERSOT64	0.115	-0.391	-0.166	-0.125	-0.239							-0.117						
VEHQ	-0.170	-0.105		-0.179		0.137	-0.157	0.241	0.147		-0.120							
AGE_REF	0.111	-0.400	-0.154		-0.188		-0.141											
FINCBTXM	-0.317						-0.302		0.122				-0.109					
FSALARYM	-0.312						-0.278			0.106								
FRRETIRM	0.119	-0.386	-0.159	-0.122	-0.215													
FSMPFRXM	-0.108						-0.146						-0.161	-0.379		-0.203	-0.168	0.142
TOTEXPCQ	-0.308	-0.155	-0.113	0.103	0.122			0.327		-0.113	0.113							
FOODCQ	-0.307		-0.106			-0.156	0.309	-0.196			-0.158							
FDHOMECQ	-0.254		-0.105			-0.138	0.301	-0.162			-0.133							
FDAWAYCQ	-0.250	-0.106		0.125		-0.116	0.198	-0.158			-0.125							
HOUSCQ	-0.268		-0.104	0.115														
APPARCQ	-0.124						0.120								-0.109	-0.104	-0.111	
TRANSCQ	-0.117							0.606	-0.151	-0.195	0.211				0.111			
HEALTHCQ	-0.128	-0.206									-0.142						0.111	
ENTERTCQ	-0.146	-0.117									-0.142	0.135						
EDUCACQ					0.123			0.126		-0.137	0.226	-0.188	0.181	0.370	-0.242	-0.251		-0.203
READCQ		-0.121						0.203			0.148	0.107		-0.370		0.160	0.153	
BLS_URBN_URBAN			-0.108	0.362	-0.395	0.398			-0.135									
BLS_URBN_RURAL			0.109	-0.362	0.395	-0.396			0.135									
EDUC_REF_NEVER														0.143	-0.141	0.238	0.347	0.837
EDUC_REF_FIRST				-0.118				-0.149		-0.150	0.312	-0.103	0.313	-0.449	0.247		-0.314	0.126
EDUC_REF_SECOND					-0.112					-0.197	0.379	0.403	-0.385	0.102	-0.284	0.144	0.218	-0.150
EDUC_REF_HIGH				-0.167				0.261	-0.539	0.327	-0.291	-0.256	-0.158			-0.102		
EDUC_REF_COL							0.332	0.184	0.648		-0.154	-0.194			0.197			
EDUC_REF_ASSC											-0.169	0.567	0.618	0.125	-0.133		-0.132	

EDUC_REF_UDGRD				0.147	0.100		-0.250	-0.242	-0.153	-0.621	-0.170	-0.131						
EDUC_REF_GRD	-0.105			0.112				-0.164		0.530	0.442						0.126	
REF_RACE_WH		-0.195	0.146	-0.214	0.296	0.491	0.212		-0.132									
REF_RACE_BLK		0.164	-0.125	0.165	-0.230	-0.372	-0.170	0.140	0.104			0.219	-0.180	0.114	0.297	-0.175	-0.132	0.137
REF_RACE_NA											-0.211		0.157	-0.387		-0.308	0.689	-0.185
REF_RACE_AS				0.114	-0.133	-0.249	-0.115	-0.144			0.115	-0.387	0.384		-0.171			
REF_RACE_PI											-0.168				0.220	0.744		-0.315
REF_RACE_MULTI									0.195	0.134			-0.105	-0.326	-0.634	0.217	-0.310	
SEX_REF_MALE		-0.164	0.529	0.120	-0.179	-0.117												
SEX_REF_FEMALE		0.164	-0.529	-0.119	0.179	0.118												

	Comp.19	Comp.20	Comp.21	Comp.22	Comp.23	Comp.24	Comp.25	Comp.26	Comp.27	Comp.28	Comp.29	Comp.30	Comp.31	Comp.32	Comp.33	Comp.34	Comp.35
HH_CU_Q		0.378	0.226	0.682	-0.118	0.198				-0.199							
AS_COMP1							-0.112			0.209	0.448	-0.411	-0.262			0.170	-0.361
AS_COMP2				0.112		-0.123				0.227	-0.550	0.243		-0.185		0.187	-0.340
CUTENURE									-0.250	0.643	0.105		0.386	0.264	0.146		
FAM_SIZE			0.136			0.201											-0.104
NO_EARNR				0.123		-0.241				0.276		-0.111	-0.200	-0.156		-0.240	0.690
PERSLT18			0.138	-0.201	-0.130	0.430				-0.235	0.162	0.199			0.154		0.213
PERSOT64			0.108					0.142		0.123						-0.784	-0.133
VEHQ					0.208	-0.128		-0.233	0.421		0.167	0.115	0.649				
AGE_REF									-0.107						0.761	0.300	0.153
FINCBTXM								0.151	-0.268					0.331	-0.123		
FSALARYM				0.115					-0.336					0.363			
FRRETIRM			0.117			0.109		0.123		0.111	0.138				-0.551	0.394	0.409
FSMPFRXM	0.459	0.514		-0.195			-0.154	0.173	0.171								
TOTEXPCQ			-0.119			0.107											
FOODCQ						-0.192			0.144					0.104			
FDHOMECQ					-0.141	-0.132				-0.162	-0.305	-0.444	0.157	0.288			
FDAWAYCQ				0.105		-0.187		0.154	0.165	0.122	0.294	0.575	-0.192	-0.152			
HOUSCQ		-0.101				0.122			-0.372			-0.204	0.374	-0.653			
APPARCQ	0.129	-0.111	0.620		0.368	-0.128	0.520	0.165									
TRANSCQ	-0.129		-0.224			0.144	0.120	0.256	0.178				-0.135	0.120			
HEALTHCQ	0.102	0.177				-0.261		0.494	-0.646		0.155			-0.174			
ENTERTCQ						0.570	0.397	-0.438	-0.337	-0.129				-0.167	0.114		
EDUCACQ	0.476	-0.237			-0.133	-0.183	-0.126	-0.258	-0.165		-0.135						
READCQ	-0.269			0.449	-0.241	-0.257	-0.347	-0.300	-0.178		-0.181						
BLS_URBN_URBAN																	
BLS_URBN_RURAL																	
EDUC_REF_NEVER	0.183																
EDUC_REF_FIRST	0.176	-0.199	-0.158	0.307	0.147	-0.166		-0.157								-0.106	
EDUC_REF_SECOND		0.158	-0.172		0.208	-0.167			-0.104	-0.175							
EDUC_REF_HIGH									-0.165								
EDUC_REF_COL				-0.224													
EDUC_REF_ASSC		0.103	0.107					0.186	-0.136								
EDUC_REF_UDGRD	-0.155	-0.149	0.142		-0.199	0.114			0.181	0.114							
EDUC_REF_GRD						0.198			0.359	0.249	-0.102	-0.133					
REF_RACE_WH																	
REF_RACE_BLK			0.126	0.203	-0.121			-0.116									
REF_RACE_NA	0.120	-0.239		0.132	0.109			0.158									
REF_RACE_AS	-0.248	0.352		-0.290	0.245												
REF_RACE_PI	0.459																
REF_RACE_MULTI	-0.103	-0.374	-0.107		-0.182	0.117											
SEX_REF_MALE											-0.273	0.155					
SEX_REF_FEMALE											0.273	-0.156					

	Comp.36	Comp.37	Comp.38	Comp.39	Comp.40	Comp.41	Comp.42	Comp.43
HH_CU_Q								
AS_COMP1			-0.288					
AS_COMP2			-0.281					
CUTENURE								
FAM_SIZE			0.757					
NO_EARNR								
PERSLT18			-0.509					
PERSOT64								
VEHQ								
AGE_REF								
FINCBTXM	-0.112	0.689						
FSALARYM		-0.674						

FRRETIRM		-0.108					
FSMPFRXM		-0.216					
TOTEXPCQ	0.798						
FOODCQ							-0.756
FDHOMECQ							0.497
FDAWAYCQ							0.426
HOUSCQ	-0.269						
APPARCQ							
TRANSCQ	-0.454						
HEALTHCQ	-0.132						
ENTERTCQ	-0.101						
EDUCACQ	-0.185						
READCQ							
BLS_URBN_URBAN			0.414	0.430	-0.356	0.133	
BLS_URBN_RURAL			0.414	0.430	-0.356	0.132	
EDUC_REF_NEVER							
EDUC_REF_FIRST					0.161		
EDUC_REF_SECOND				0.108	0.259		
EDUC_REF_HIGH			0.125	0.168	0.401	0.139	
EDUC_REF_COL			0.122	0.164	0.389	0.135	
EDUC_REF_ASSC				0.115	0.272		
EDUC_REF_UDGRD			0.117	0.159	0.377	0.130	
EDUC_REF_GRD				0.129	0.307	0.106	
REF_RACE_WH			0.501	-0.469			
REF_RACE_BLK			0.413	-0.386			
REF_RACE_NA							
REF_RACE_AS			0.279	-0.261			
REF_RACE_PI							
REF_RACE_MULTI			0.135	-0.127			
SEX_REF_MALE			0.169	0.140	0.118	-0.662	
SEX_REF_FEMALE			0.169	0.138	0.118	-0.662	

Appendix C: Project R codes

```
[1] remove(list=ls())
[2]
[3] fmli161 <- read.csv("fmli161.csv", header = TRUE,na.string = c(".", "NA"), row.names = 1)
[4]
[5] name <- c("HH_CU_Q", "AS_COMP1", "AS_COMP2", "BLS_URBN", "CUTENURE",
[6]       "FAM_SIZE", "NO_EARNR", "PERSLT18", "PERSOT64", "VEHQ", "AGE_REF",
[7]       "EDUC_REF", "REF_RACE", "SEX_REF", "INC_HRS1", "OCCUCOD1", "FINCBTXM",
[8]       "FSALARYM", "FRRETIRM","FSMPFRXM", "WELFAREM", "INTRDVXM",
[9]       "NETRENTM","RETSURVM", "TOTEXPCQ","FOODCQ",
[10]      "FDHOMECQ", "FDAWAYCQ","HOUSCQ", "APPARCQ","TRANSCQ", "HEALTHCQ",
[11]      "ENTERTCQ", "EDUCACQ","READCQ")
[12]
[13] name2 <- c("HH_CU_Q_", "AS_C_MP1", "AS_C_MP2", "BLS_URBN", "CUTE_URE",
[14]       "FAM__IZE", "NO_E_RNR", "PERS_T18", "PERS_T64", "VEHQ_", "AGE_REF_",
[15]       "EDUC0REF", "REF__ACE", "SEX_REF_", "INC__RS1", "OCCU_OD1", "FINCB_XM",
[16]       "FSAL_RYM", "FRRE_IRM","FSMP_RXM", "WELF_REM", "INTR_VXM",
[17]       "NETR_NTM","RETS_RVM", "TOTEXPCQ","FOODCQ",
[18]       "FDHOMECQ", "FDAWAYCQ","HOUSCQ", "APPARCQ","TRANSCQ", "HEALTHCQ",
[19]       "ENTERTCQ", "EDUCACQ","READCQ")
[20] #Total of 34 variables selected
[21] fmli <- subset(fmli161, select = name)
[22] flag <- subset(fmli161,select = c(name2 ))
[23]
[24] library("mice")
[25]
[26] #Try to split the dataset according to if TOTEXPCQ is empty
[27] delete = which(fmli$TOTEXPCQ == 0)
[28] str(which(fmli$TOTEXPCQ == 0))#total of 2171 missing tot expenditure
[29] fmlifull = fmli[-delete,]
[30] str(which(fmlifull$FOODCQ == 0)) #28/4524 = 0.61%
[31] str(which(fmlifull$FDHOMECQ == 0)) #46/4524 = 1.01%
[32] str(which(fmlifull$FDAWAYCQ == 0)) #879/4524 = 19.42%
```

```
[33] str(which(fmlifull$HOUSCQ == 0)) #23/4524 = 0.50%
[34] str(which(fmlifull$APPARCQ == 0)) #2168/4524 = 47.92%
[35] str(which(fmlifull$TRANSCQ == 0)) #268/4524 = 5.92%
[36] str(which(fmlifull$HEALTHCQ == 0)) #941/4524 = 20.81%
[37] str(which(fmlifull$ENTERTCQ == 0)) #649/4524 = 14.34%
[38] str(which(fmlifull$EDUCACQ == 0)) #3803/4524 = 84.06%
[39] str(which(fmlifull$READCQ == 0)) #3629/4524 = 80.22%
[40]
[41] md.pattern(fmlifull)
[42] #drop X,Y,Z,AA,S:5000 - 6000+ NAs, not usable
[43] fmlifull = fmlifull[,-c(21:24)]
[44] #for INC-HRS1 and OCCUCOD1 missing 2191, 2191/4524 = 48.43%, dropped
[45]
[46] identical(which(is.na(fmli[,"OCCUCOD1"])), which(is.na(fmli[,"INC_HRS1"])))
[47]
[48] fmlifull = fmlifull[,-c(15:16)]
[49] md.pattern(fmlifull)
[50] name3 <- c("HH_CU_Q", "AS_COMP1", "AS_COMP2", "BLS_URBN",
     "CUTENURE",
[51]       "FAM_SIZE", "NO_EARNR", "PERSLT18", "PERSOT64", "VEHQ",
     "AGE_REF",
[52]       "EDUC_REF", "REF_RACE", "SEX_REF", "FINCBTXM",
[53]       "FSALARYM", "FRRETIRM","FSMPFRXM", "TOTEXPCQ","FOODCQ",
[54]       "FDHOMECQ", "FDAWAYCQ","HOUSCQ", "APPARCQ","TRANSCQ",
     "HEALTHCQ",
[55]       "ENTERTCQ")
[56]
[57] name4 <- c("HH_CU_Q_", "AS_C_MP1", "AS_C_MP2", "BLS_URBN",
     "CUTE_URE",
[58]       "FAM__IZE", "NO_E_RNR", "PERS_T18", "PERS_T64", "VEHQ_",
     "AGE_REF_",
[59]       "EDUC0REF", "REF__ACE", "SEX_REF_", "FINCB_XM",
[60]       "FSAL_RYM", "FRRE_IRM","FSMP_RXM", "TOTEXPCQ","FOODCQ",
[61]       "FDHOMECQ", "FDAWAYCQ","HOUSCQ", "APPARCQ","TRANSCQ",
     "HEALTHCQ", "ENTERTCQ")
[62] flagfull <- subset(fmli161,select = c(name4))[-delete,]
[63]
[64] str(fmlifull)
[65]
[66] blank = rep(c(0),4254)
[67] urb_ind <- which(fmlifull$BLS_URBN == 1)
[68] for(i in (0:(length(urb_ind)-1))){
[69]  blank[urb_ind[i]] = 1
```

```
[70] }
[71] BLS_URBN_URBAN = blank
[72]
[73] blank = rep(c(0),4254)
[74] rural_ind <- which(fmlifull$BLS_URBN == 2)
[75] for(i in (0:(length(rural_ind)-1))){
[76]   blank[rural_ind[i]] = 1
[77] }
[78] BLS_URBN_RURAL = blank
[79]
[80] blank = rep(c(0),4254)
[81] never_ind <- which(fmlifull$EDUC_REF == 0)
[82] for(i in (0:(length(never_ind)-1))){
[83]   blank[never_ind[i]] = 1
[84] }
[85] EDUC_REF_NEVER = blank
[86]
[87] blank = rep(c(0),4254)
[88] first_ind <- which(fmlifull$EDUC_REF == 10)
[89] for(i in (0:(length(first_ind)-1))){
[90]   blank[first_ind[i]] = 1
[91] }
[92] EDUC_REF_FIRST = blank
[93]
[94] blank = rep(c(0),4254)
[95] second_ind <- which(fmlifull$EDUC_REF == 11)
[96] for(i in (0:(length(second_ind)-1))){
[97]   blank[second_ind[i]] = 1
[98] }
[99] EDUC_REF_SECOND = blank
[100]
[101]blank = rep(c(0),4254)
[102]high_ind <- which(fmlifull$EDUC_REF == 12)
[103]for(i in (0:(length(high_ind)-1))){
[104]  blank[high_ind[i]] = 1
[105]}
[106]EDUC_REF_HIGH = blank
[107]
[108]blank = rep(c(0),4254)
[109]col_ind <- which(fmlifull$EDUC_REF == 13)
[110]for(i in (0:(length(col_ind)-1))){
[111]  blank[col_ind[i]] = 1
```

```
[112]}
[113]EDUC_REF_COL = blank
[114]
[115]blank = rep(c(0),4254)
[116]assc_ind <- which(fmlifull$EDUC_REF == 14)
[117]for(i in (0:(length(assc_ind)-1))){
[118]  blank[assc_ind[i]] = 1
[119]}
[120]EDUC_REF_ASSC = blank
[121]
[122]
[123]blank = rep(c(0),4254)
[124]udgrd_ind <- which(fmlifull$EDUC_REF == 15)
[125]for(i in (0:(length(udgrd_ind)-1))){
[126]  blank[udgrd_ind[i]] = 1
[127]}
[128]EDUC_REF_UDGRD = blank
[129]
[130]blank = rep(c(0),4254)
[131]grd_ind <- which(fmlifull$EDUC_REF == 16)
[132]for(i in (0:(length(grd_ind)-1))){
[133]  blank[grd_ind[i]] = 1
[134]}
[135]EDUC_REF_GRD = blank
[136]#
[137]#REF_RACE
[138]blank = rep(c(0),4254)
[139]wh_ind <- which(fmlifull$REF_RACE == 1)
[140]for(i in (0:(length(wh_ind)-1))){
[141]  blank[wh_ind[i]] = 1
[142]}
[143]REF_RACE_WH = blank
[144]
[145]blank = rep(c(0),4254)
[146]blk_ind <- which(fmlifull$REF_RACE == 2)
[147]for(i in (0:(length(blk_ind)-1))){
[148]  blank[blk_ind[i]] = 1
[149]}
[150]REF_RACE_BLK = blank
[151]
[152]blank = rep(c(0),4254)
[153]na_ind <- which(fmlifull$REF_RACE == 3)
```

```
[154]for(i in (0:(length(na_ind)-1))){
[155]  blank[na_ind[i]] = 1
[156]}
[157]REF_RACE_NA = blank
[158]
[159]blank = rep(c(0),4254)
[160]as_ind <- which(fmlifull$REF_RACE == 4)
[161]for(i in (0:(length(as_ind)-1))){
[162]  blank[as_ind[i]] = 1
[163]}
[164]REF_RACE_AS = blank
[165]
[166]blank = rep(c(0),4254)
[167]pi_ind <- which(fmlifull$REF_RACE == 5)
[168]for(i in (0:(length(pi_ind)-1))){
[169]  blank[pi_ind[i]] = 1
[170]}
[171]REF_RACE_PI = blank
[172]
[173]blank = rep(c(0),4254)
[174]multi_ind <- which(fmlifull$REF_RACE == 6)
[175]for(i in (0:(length(multi_ind)-1))){
[176]  blank[multi_ind[i]] = 1
[177]}
[178]REF_RACE_MULTI = blank
[179]#
[180]#SEX_REF
[181]blank = rep(c(0),4254)
[182]male_ind <- which(fmlifull$SEX_REF == 1)
[183]for(i in (0:(length(male_ind)-1))){
[184]  blank[male_ind[i]] = 1
[185]}
[186]SEX_REF_MALE = blank
[187]
[188]blank = rep(c(0),4254)
[189]female_ind <- which(fmlifull$SEX_REF == 2)
[190]for(i in (0:(length(female_ind)-1))){
[191]  blank[female_ind[i]] = 1
[192]}
[193]SEX_REF_FEMALE = blank
[194]
[195]fmlifull = fmlifull[,-(12:14)]
```

```
fmlifull = fmlifull[,-4]
fmlifull = cbind(fmlifull, BLS_URBN_URBAN, BLS_URBN_RURAL, EDUC_REF_NEVER, EDUC_REF_FIRST, EDUC_REF_SECOND,
EDUC_REF_HIGH, EDUC_REF_ASSC, EDUC_REF_UDGRD, EDUC_REF_GRD, REF_RACE_WH, REF_RACE_BLK,
REF_RACE_NA, REF_RACE_AS, REF_RACE_PI, REF_RACE_MULTI, SEX_REF_MALE, SEX_REF_FEMALE)
write.csv(fmlifull,"fmlifull.csv")

#Correlation
round(cor(fmlifull),digits = 2)
large = function(m, value){
  if(m >= value){
    return(TRUE)
  }
}
#PCA
fmlifull.pca = princomp(x=fmlifull, cor = TRUE)
summary(fmlifull.pca, loadings = TRUE)
round(summary(fmlifull.pca, loadings = TRUE)$loadings[,1:6], digits = 2)
print(fmlifull.pca)
plot(fmlifull.pca,type="lines")

#plot the variables in PC1-PC2
par(cex = 0.5)
plot(fmlifull.pca$loadings[,1],fmlifull.pca$loadings[,2],
     xlab="PC1",ylab="PC2",type="n", xlim = c(-0.4, 0.3), ylim = c(-0.3,0.45))
text(fmlifull.pca$loadings[,1],fmlifull.pca$loadings[,2],labels=name3)

#Plot the data points in PC1-PC2 grouping by race
m = matrix(data=c(1:6), nrow=3, ncol=2, byrow=TRUE)
layout(m)

par(cex=0.5)
z1 = fmlifull.pca$scores[,1]
z2 = fmlifull.pca$scores[,2]
colors <- c("blue","black","red","orange", "brown", "green")
leg.txt <- c("White","Black","Native American","Asian", "Pacific Islander", "Multiple Races")

plot(z1,z2,xlab="PC1",ylab="PC2",type="n", xlim = c(-15, 5), ylim = c(-6,9))
gp <- fmlifull$REF_RACE_WH == 1 #1 is white
points(z1[gp],z2[gp],col="blue")
```

```
[235]legend("bottomleft",legend=leg.txt[1],fill=colors[1])
[236]
[237]plot(z1,z2,xlab="PC1",ylab="PC2",type="n")
[238]gp <- fmlifull$REF_RACE_BLK == 1#2 is black
[239]points(z1[gp],z2[gp],col="black")
[240]legend("bottomleft",legend=leg.txt[2],fill=colors[2])
[241]
[242]plot(z1,z2,xlab="PC1",ylab="PC2",type="n")
[243]gp <- fmlifull$REF_RACE_NA  == 1#3 is native american
[244]points(z1[gp],z2[gp],col="red")
[245]legend("bottomleft",legend=leg.txt[3],fill=colors[3])
[246]
[247]plot(z1,z2,xlab="PC1",ylab="PC2",type="n")
[248]gp <- fmlifull$REF_RACE_AS  == 1#4 is asian
[249]points(z1[gp],z2[gp],col="orange")
[250]legend("bottomleft",legend=leg.txt[4],fill=colors[4])
[251]
[252]plot(z1,z2,xlab="PC1",ylab="PC2",type="n")
[253]gp <- fmlifull$REF_RACE_PI  == 1#5 is pacific islander
[254]points(z1[gp],z2[gp],col="brown")
[255]legend("bottomleft",legend=leg.txt[5],fill=colors[5])
[256]
[257]plot(z1,z2,xlab="PC1",ylab="PC2",type="n")
[258]gp <- fmlifull$REF_RACE_MULTI  == 1#6 is multiple
[259]points(z1[gp],z2[gp],col="darkgreen")
[260]legend("bottomleft",legend=leg.txt[6],fill=colors[6])
[261]
[262]layout(matrix(data=1, nrow=1, ncol=1))
[263]
[264]
[265]#Two ways to make biplots, not sure which is better for reading
[266]library(devtools)
[267]#install_github("ggbiplot", "vqv")
[268]
[269]library(ggbiplot)
[270]g <- ggbiplot(fmlifull.pca, obs.scale = 1, var.scale = 1,
[271]              ellipse = TRUE,
[272]              circle = TRUE)
[273]#g <- g + scale_color_discrete(name = '')
[274]g <- g + theme(legend.direction = 'horizontal',
[275]               legend.position = 'top')
[276]g$layers <- c(g$layers, g$layers[[2]])
```

```
[277]print(g)
[278]
[279]require(graphics)
[280]par( pch = 20)
[281]par(cex = 0.8)
[282]biplot(princomp(fmlifull, cor = TRUE), xlabs=rep("·", nrow(fmlifull)))
[283]
[284]#male/female plotted against pca, pc2, pc3
[285]library(rgl)
[286]fmlifull.pca = princomp(x=fmlifull, cor=TRUE)
[287]plot3d(fmlifull.pca$scores[,1:3], col=fmlifull$SEX_REF_FEMALE+1, size = 5)
[288]
[289]#Sparse PCA
[290]library(nsprcomp)
[291]set.seed(0)
[292]fmlifull.pca.sparse <- nsprcomp(fmlifull, ncomp = 3, center=T, scale.=T, k=c(5,5,5),
     nneg = FALSE)
[293]summary(fmlifull.pca.sparse, loadings = TRUE)
[294]print(fmlifull.pca.sparse)
[295]plot(fmlifull.pca.sparse,type="lines")
[296]
[297]#plot the variables in PC1-PC2
[298]par(cex = 0.5)
[299]plot(fmlifull.pca.sparse$x[,1],fmlifull.pca.sparse$x[,2],
[300]     xlab="sparsePC1",ylab="sparsePC2",type="n")
[301]text(fmlifull.pca.sparse$x[,1],fmlifull.pca.sparse$x[,2],labels="o")
[302]
[303]library(rgl)
[304]plot3d(fmlifull.pca.sparse$x, col=fmlifull$BLS_URBN_RURAL+1)
[305]plot3d(fmlifull.pca.sparse$x, col=fmlifull$SEX_REF_FEMALE+1)
[306]
[307]
[308]#Correspondence Analysis:
[309]#Question: is educational level related to race:
[310]#Admittedly that was not what this survey was designed for, but it would be nice to
     use this for this
[311]#purpose given that the survey is extensive and carefully conducted.
[312]library(vcd)
[313]
[314]x <- matrix(c(length(intersect(which(fmlifull$EDUC_REF_NEVER ==
     1),which(fmlifull$REF_RACE_WH == 1))),
[315]          length(intersect(which(fmlifull$EDUC_REF_NEVER == 1),
     which(fmlifull$REF_RACE_BLK == 1))),
```

```
[316]          length(intersect(which(fmlifull$EDUC_REF_NEVER                    ==
    1),which(fmlifull$REF_RACE_NA == 1))),
[317]          length(intersect(which(fmlifull$EDUC_REF_NEVER      ==      1)      ,
    which(fmlifull$REF_RACE_AS == 1))),
[318]          length(intersect(which(fmlifull$EDUC_REF_NEVER      ==      1)      ,
    which(fmlifull$REF_RACE_PI == 1))),
[319]          length(intersect(which(fmlifull$EDUC_REF_NEVER      ==      1      )      ,
    which( fmlifull$REF_RACE_MULTI == 1))),
[320]          length(intersect(which(fmlifull$EDUC_REF_FIRST      ==      1      )      ,
    which( fmlifull$REF_RACE_WH == 1))),
[321]          length(intersect(which(fmlifull$EDUC_REF_FIRST      ==      1      )      ,
    which( fmlifull$REF_RACE_BLK == 1))),
[322]          length(intersect(which(fmlifull$EDUC_REF_FIRST      ==      1      )      ,
    which( fmlifull$REF_RACE_NA == 1))),
[323]          length(intersect(which(fmlifull$EDUC_REF_FIRST      ==      1      )      ,
    which( fmlifull$REF_RACE_AS == 1))),
[324]          length(intersect(which(fmlifull$EDUC_REF_FIRST==          1          )          ,
    which( fmlifull$REF_RACE_PI == 1))),
[325]          length(intersect(which(fmlifull$EDUC_REF_FIRST          ==          1          ),
    which( fmlifull$REF_RACE_MULTI == 1))),
[326]          length(intersect(which(fmlifull$EDUC_REF_SECOND      ==      1      )      ,
    which( fmlifull$REF_RACE_WH == 1))),
[327]          length(intersect(which(fmlifull$EDUC_REF_SECOND      ==      1      ),
    which( fmlifull$REF_RACE_BLK == 1))),
[328]          length(intersect(which(fmlifull$EDUC_REF_SECOND==          1          ),
    which( fmlifull$REF_RACE_NA == 1))),
[329]          length(intersect(which(fmlifull$EDUC_REF_SECOND      ==      1      ),
    which( fmlifull$REF_RACE_AS == 1))),
[330]          length(intersect(which(fmlifull$EDUC_REF_SECOND      ==      1      ),
    which( fmlifull$REF_RACE_PI == 1))),
[331]          length(intersect(which(fmlifull$EDUC_REF_SECOND      ==      1      ),
    which( fmlifull$REF_RACE_MULTI == 1))),
[332]          length(intersect(which(fmlifull$EDUC_REF_HIGH          ==          1          ),
    which( fmlifull$REF_RACE_WH == 1))),
[333]          length(intersect(which(fmlifull$EDUC_REF_HIGH==          1          )          ,
    which( fmlifull$REF_RACE_BLK == 1))),
[334]          length(intersect(which(fmlifull$EDUC_REF_HIGH          ==          1          ),
    which( fmlifull$REF_RACE_NA == 1))),
[335]          length(intersect(which(fmlifull$EDUC_REF_HIGH          ==          1          ),
    which( fmlifull$REF_RACE_AS == 1))),
[336]          length(intersect(which(fmlifull$EDUC_REF_HIGH          ==          1          ),
    which( fmlifull$REF_RACE_PI == 1))),
[337]          length(intersect(which(fmlifull$EDUC_REF_HIGH      ==      1      )      ,
    which( fmlifull$REF_RACE_MULTI == 1))),
[338]          length(intersect(which(fmlifull$EDUC_REF_COL      ==      1      )      ,
    which(fmlifull$REF_RACE_WH == 1))),
```

```
[339]  length(intersect(which(fmlifull$EDUC_REF_COL == 1 ) , which( fmlifull$REF_RACE_BLK == 1))),
[340]  length(intersect(which(fmlifull$EDUC_REF_COL == 1 ) , which( fmlifull$REF_RACE_NA == 1))),
[341]  length(intersect(which(fmlifull$EDUC_REF_COL == 1 ) , which( fmlifull$REF_RACE_AS == 1))),
[342]  length(intersect(which(fmlifull$EDUC_REF_COL == 1 ) , which( fmlifull$REF_RACE_PI == 1))),
[343]  length(intersect(which(fmlifull$EDUC_REF_COL == 1 ), which( fmlifull$REF_RACE_MULTI == 1))),
[344]  length(intersect(which(fmlifull$EDUC_REF_ASSC == 1 ), which( fmlifull$REF_RACE_WH == 1))),
[345]  length(intersect(which(fmlifull$EDUC_REF_ASSC == 1 ) , which( fmlifull$REF_RACE_BLK == 1))),
[346]  length(intersect(which(fmlifull$EDUC_REF_ASSC == 1 ) ,which( fmlifull$REF_RACE_NA == 1))),
[347]  length(intersect(which(fmlifull$EDUC_REF_ASSC == 1 ) , which( fmlifull$REF_RACE_AS == 1))),
[348]  length(intersect(which(fmlifull$EDUC_REF_ASSC == 1 ) , which( fmlifull$REF_RACE_PI == 1))),
[349]  length(intersect(which(fmlifull$EDUC_REF_ASSC == 1 ) , which( fmlifull$REF_RACE_MULTI == 1))),
[350]  length(intersect(which(fmlifull$EDUC_REF_UDGRD == 1 ) , which( fmlifull$REF_RACE_WH == 1))),
[351]  length(intersect(which(fmlifull$EDUC_REF_UDGRD == 1 ) , which( fmlifull$REF_RACE_BLK == 1))),
[352]  length(intersect(which(fmlifull$EDUC_REF_UDGRD == 1) , which( fmlifull$REF_RACE_NA == 1))),
[353]  length(intersect(which(fmlifull$EDUC_REF_UDGRD == 1 ) , which( fmlifull$REF_RACE_AS == 1))),
[354]  length(intersect(which(fmlifull$EDUC_REF_UDGRD == 1 ), which(fmlifull$REF_RACE_PI == 1))),
[355]  length(intersect(which(fmlifull$EDUC_REF_UDGRD == 1 ) , which( fmlifull$REF_RACE_MULTI == 1))),
[356]  length(intersect(which(fmlifull$EDUC_REF_GRD == 1) , which( fmlifull$REF_RACE_WH == 1))),
[357]  length(intersect(which(fmlifull$EDUC_REF_GRD == 1 ), which( fmlifull$REF_RACE_BLK == 1))),
[358]  length(intersect(which(fmlifull$EDUC_REF_GRD == 1 ) ,which( fmlifull$REF_RACE_NA == 1))),
[359]  length(intersect(which(fmlifull$EDUC_REF_GRD == 1 ) , which( fmlifull$REF_RACE_AS == 1))),
[360]  length(intersect(which(fmlifull$EDUC_REF_GRD == 1 ), which( fmlifull$REF_RACE_PI == 1))),
[361]  length(intersect(which(fmlifull$EDUC_REF_GRD == 1 ), which( fmlifull$REF_RACE_MULTI == 1)))),
```

```
[362]          nrow = 8,byrow=TRUE,ncol=6)
[363]rownames <- c("NEVER","FIRST","SECOND","HIGH","COL","ASSC","UNDGRD","GRD")
[364]colnames <- c("WHITE","BLK","NAT AM","ASN","PAC IS", "MULTI")
[365]dimnames(x) <- list(rownames,colnames)
[366]names(dimnames(x)) <- c("Educational level","Race")
[367]par(cex.lab=0.1,cex.main=0.1)
[368]mosaic(x,main="Education vs. Race",shade=TRUE,color=TRUE)
[369]
[370]#Since the mosaic shows that some categories are so insignificant, drop those categories then redraw mosaic, or
[371]#concatenate those races together into one
[372]x1 <- x
[373]x1[,3] = x1[,3]+x1[,5]+x1[,6] # Substitute Native Americans with Other, the sum of
[374]#Native AMerican, Pacific Islander and multirace
[375]x1[1,] = x1[1,] + x1[2,]
[376]x1 = x1[-2,]
[377]x1 = x1[,-(5:6)]
[378]#Substitue NEVER with NEVER OR FIRST
[379]rownames <- c("NEVER or FIRST","SECOND","HIGH","COL","ASSC","UNDGRD","GRD")
[380]colnames <- c("WHITE","BLK","OTHER","ASN")
[381]dimnames(x1) <- list(rownames,colnames)
[382]names(dimnames(x1)) <- c("Educational level","Race")
[383]par(cex.lab=0.1,cex.main=0.1)
[384]mosaic(x1,main="Education vs. Race",shade=TRUE,color=TRUE,
[385]      labeling_args = list(gp_labels = gpar(fontsize = 8),
[386]  gp_varnames = gpar(fontsize = 16)
[387]))
[388]
[389]library(ca)
[390]model <- ca(x)
[391]par(pty="s",cex=1)
[392]plot(model,labels=2)
[393]legend("topright",legend=c("Educational level","Race"),
[394]      col=c("blue","red"),pch=c(16,17), cex=o.6)
[395]
[396]model <- ca(x1)
[397]par(pty="s",cex=1)
[398]plot(model,labels=2)
[399]legend("topright",legend=c("Educational level concatenated","Race cancatenated"),
[400]      col=c("blue","red"),pch=c(16,17), cex=0.6)
[401]
```

```
[402]
[403]#Cluster analysis
[404]# library(mclust)
[405]# mclus <- Mclust(fmlifull) # fits up to 9 clusters by default
[406]# mclass <- mclus$classification
[407]# k <- mclus$G # number of clusters
[408]# for(i in 1:k){ print(paste("Cluster",i))
[409]#   print(c.names[mclass == i])
[410]# }
[411]#Not useful, since we already know that the datset is not normal so should not use
     this, and returns only one cluster
[412]
[413]n <- dim(fmlifull)[1]
[414]k <- 9
[415]wss <- rep(0,k)
[416]fmlifull.m <- apply(fmlifull,2,mean)
[417]for(i in 1:n){
[418]  wss[1] <- wss[1]+sum((fmlifull[i,]-fmlifull.m)^2)
[419]}
[420]for(i in 2:k){
[421]  model <- kmeans(fmlifull,i)
[422]  wss[i] <- sum(model$withinss)
[423]}
[424]par(cex=0.8)
[425]plot(1:k,wss,type="b",xlab="Number of clusters",
[426]     ylab="Within cluster sum of squares", main="Scree plot")
[427]
[428]
[429]
[430]#try cluster of variables
[431]library(ClustOfVar)
[432]clustofvar <- hclustvar(fmlifull)
[433]par(cex = 0.7)
[434]dev.new(width = 10,height = 5)
[435]plot(clustofvar)
[436]
[437]#try Hmisc to cluster variables and compare to ClusofVar
[438]library(Hmisc)
[439]#using Spearman's rho
[440]Hclust_rho <- varclus(as.matrix(fmlifull), similarity = "spearman")
[441]Hclust_rho
[442]par(cex = 0.7)
[443]dev.new(width = 10,height = 5)
```

```
[444]plot(Hclust_rho)
[445]
[446]Hclust_D <- varclus(as.matrix(fmlifull), similarity = "hoeffding")#very slow, results
     not so good
[447]Hclust_D
[448]par(cex = 0.7)
[449]dev.new(width = 10,height = 5)
[450]plot(Hclust_D)
[451]
[452]
[453]Hclust_Pearson <- varclus(as.matrix(fmlifull), similarity = "pearson")
[454]Hclust_Pearson
[455]par(cex = 0.7)
[456]dev.new(width = 10,height = 5)
[457]plot(Hclust_Pearson)
[458]
[459]#Canonical correlation analysis, compare with
[460]#we are interested in comparing the 4 income variables:FINCBTXM, FSALARYM,
     FRRETIRM, FSMPFRXM
[461]#to the 11 expenditure variables:
[462]str(fmlifull)
[463]round(cor(fmlifull),digits = 2)
[464]fincbtm <- scale(fmlifull$FINCBTXM)
[465]fsalarym <- scale(fmlifull$FSALARYM)
[466]frretirm <- scale(fmlifull$FRRETIRM)
[467]fsmpfrxm <- scale(fmlifull$FSMPFRXM)
[468]totexpcq <- scale(fmlifull$TOTEXPCQ)
[469]foodcq <- scale(fmlifull$FOODCQ)
[470]fdhomecq <- scale(fmlifull$FDHOMECQ)
[471]fdawaycq <- scale(fmlifull$FDAWAYCQ)
[472]houscq <- scale(fmlifull$HOUSCQ)
[473]apparcq <- scale(fmlifull$APPARCQ)
[474]transcq <- scale(fmlifull$TRANSCQ)
[475]healthcq <- scale(fmlifull$HEALTHCQ)
[476]entertcq <- scale(fmlifull$ENTERTCQ)
[477]educacq <- scale(fmlifull$EDUCACQ)
[478]readcq <- scale(fmlifull$READCQ)
[479]cca.cor <- cor(cbind(fincbtm, fsalarym, frretirm, fsmpfrxm, totexpcq, foodcq,
     fdhomecq, fdawaycq, houscq,
[480]                    apparcq, transcq, healthcq, entertcq, educacq, readcq))
[481]r11 <- cca.cor[1:4,1:4]
[482]r22 <- cca.cor[5:15,5:15]
[483]r12 <- cca.cor[1:4,5:15]
```

```
[484]r21 <- t(r12)
[485]e1 <- solve(r11) %*% r12 %*% solve(r22) %*% r21
[486]e2 <- solve(r22) %*% r21 %*% solve(r11) %*% r12
[487]eigen(e1)
[488]round(sqrt(eigen(e1)$values), digits = 2)
[489]round((eigen(e1)$vectors), digits = 2)
[490]eigen(e2)
[491]round(sqrt(eigen(e2)$values), digits = 2)
[492]round((eigen(e2)$vectors), digits = 2)
[493]x <- cbind(fincbtm, fsalarym, frretirm, fsmpfrxm)
[494]y <- cbind(totexpcq, foodcq, fdhomecq, fdawaycq, houscq, apparcq, transcq, healthcq,
     entertcq, educacq, readcq)
[495]u <- x %*% eigen(e1)$vectors
[496]v <- y %*% eigen(e2)$vectors
[497]round(cor(u,x), digits = 2)
[498]round(cor(v,y),digits = 2)
[499]
[500]#
[501]#Second approach: pick subgroups of expenditure of size four and then do cca with
[502]#4 income variables
[503]#Selection criterion:based on sparse pca, cluster and intuition
[504]
[505]
[506]#Selection criterion:based on cluster of variables analysis
[507]#FOODCQ,TOTEXPCQ,HOUSCQ,HEALTHCQ, each from as far away a
     subcluster as possible
[508]fincbtm <- scale(fmlifull$FINCBTXM)
[509]fsalarym <- scale(fmlifull$FSALARYM)
[510]frretirm <- scale(fmlifull$FRRETIRM)
[511]fsmpfrxm <- scale(fmlifull$FSMPFRXM)
[512]totexpcq <- scale(fmlifull$TOTEXPCQ)
[513]foodcq <- scale(fmlifull$FOODCQ)
[514]houscq <- scale(fmlifull$HOUSCQ)
[515]healthcq <- scale(fmlifull$HEALTHCQ)
[516]
[517]cca.cor <- cor(cbind(fincbtm, fsalarym, frretirm, fsmpfrxm, totexpcq, foodcq, houscq,
     healthcq))
[518]r11 <- cca.cor[1:4,1:4]
[519]r22 <- cca.cor[5:8,5:8]
[520]r12 <- cca.cor[1:4,5:8]
[521]r21 <- t(r12)
[522]e1 <- solve(r11) %*% r12 %*% solve(r22) %*% r21
[523]e2 <- solve(r22) %*% r21 %*% solve(r11) %*% r12
```

```
[524]eigen(e1)
[525]round(sqrt(eigen(e1)$values), digits = 2)
[526]round((eigen(e1)$vectors), digits = 2)
[527]eigen(e2)
[528]round(sqrt(eigen(e2)$values), digits = 2)
[529]round((eigen(e2)$vectors), digits = 2)
[530]x <- cbind(fincbtm, fsalarym, frretirm, fsmpfrxm)
[531]y <- cbind(totexpcq, foodcq, houscq, healthcq)
[532]u <- x %*% eigen(e1)$vectors
[533]v <- y %*% eigen(e2)$vectors
[534]round(cor(u,x),digits = 2)
round(cor(v,y),digits = 2)
```