# Incorporating Open Portal Data into Courses in Statistics

Roberto Rivera* Mario Marazzi† Pedro Torres‡

## 1. Abstract

The 2016 Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report emphasizes six recommendations to teach introductory courses in statistics. Among them: use of real data with context and purpose. Many educators have created databases consisting of multiple datasets for use in class; sometimes making hundreds of datasets available. Yet 'the context and purpose' component of the data may remain elusive if just a generic database is made available.

We describe the use of open portal data in introductory courses. Countries and cities continue to share data through these portals. Hence, educators can find regional data that engages their students more effectively. We present excerpts from case studies that show the application of statistical methods to data on: crime, bankruptcy, rainfall, tourist travel, and others. Data clean up and discussion of results are recognized as important case study components. Thus the open portal data based case studies attend all GAISE College Report recommendations. Reproducible R code is made available for each case study. Example uses of open portal data in more advanced courses in statistics are also presented.

## 2. Background

In 2016, the eleven year old GAISE College Report was revised. Two reasons for the revision were the increase in available data, and the emergence of Data Science as a discipline (GAISE 2016). The GAISE College report recommends the use of real data with context and purpose. The report even mentions the New York City data open portal data as a reference. However, almost all data sets used as example or referred to in the report can not be considered regional for most course students. Furthermore, most data sets referred to in the report are below a few thousand observations.

Two important obstacles to effective statistical education for non-statistics majors that receive little attention from statisticians is statistics anxiety and attitude toward statistics. Statistics anxiety has been defined as the feelings of anxiety encountered when taking a statistics course or doing statistical analysis (Cruise 1985). Attitude toward statistics is an individuals disposition to respond either favorably or unfavorably to statistics or statistics learning (Chew and Dillon 2015). It has been found that negative attitude towards statistics results in statistics anxiety (Chew and Dillon 2014). Several studies have found that there is a negative association between statistics anxiety and achievement in statistics courses (Chew and Dillon 2014), although some studies have suggested that some statistics anxiety (but not too much) may be beneficial to students (see Onwuegbuzie and Wilson (2003);

*College of Business, University of Puerto Rico, Mayaguez
†Puerto Rico Institute of Statistics
‡Math Department, University of Puerto Rico, Mayaguez

Keeley et al. (2008)). Statistics anxiety also affects non-statistics major graduate students (Williams 2010). The influence of statistics anxiety on student achievement has lead to several recommendations, among them using real data (Neumann et al. 2013), and the reduction of mathematical emphasis (Chew and Dillon 2014).

In real life students deal with data that is far more complex than the small, neat, data sets traditionally seen in introductory statistics (Baumer 2015; Grimshaw 2015). The curriculum must prepare students to engage in the entire data analysis process including data wrangling (Horton and Hardin 2015). With the emergence of big data, data science and data analytics, more emphasis in large data is needed in courses. Ridway (2016) suggested to devote course time to open data. In this paper we present a series of uses of open portal data in class. We argue that in terms of exposure for students, open portal datasets are hard to compete with because: data is current and localized allowing to incorporate current hot topics into the course, data modification is often necessary, real data can be messy and occasionally very large. Challenges of bringing open portal data into the classroom are also discussed.

## 3. Case Studies

A set of case studies revolving around open portal data are covered in this section. Usually students are presented with the problem addressed in the case study and asked what they think the solution or answer should be before seeing the results of the statistical procedure. Later students are introduced to the data. Even if the data is downloaded and made accessible to the students, we encourage instructors to briefly show the students the data from the website so that they can appreciate the authenticity of the observations. Any issues with the data are either resolved together with the class or pointed out as being fixed. One final remark is that for presentation and reproducibility purposes, all statistical procedures in this paper were performed through R (R Core Team 2016). Codes are available by emailing the primary author but some can be found online[1]. Statistics majors can use the codes to replicate the results or modify it for a different purpose. However, considering the influence of statistical anxiety on the performance of non-statistics majors, we recommend students rely on a more user friendly software for any coursework.

### 3.1 Obtaining a metric of violence in Puerto Rico

Residents in Puerto Rico are aware that violence is an issue in the United States territory, in part because traditionally, on New Years eve local media will report on the number of murders for the year and how it compares to previous years. But, how can we compare violence in one place to violence in another? Students tend to quickly realize that such a comparison should not be based on the number of murders and that the demographics of the places to be compared play a role. In this case study, Puerto Rico murder rates per 100,000 people are computed from open portal data and compared among several regions. As a first look at data science, students are told that the murder rates were obtained by processing two data sets in two open data portals (see Figure 1). Crime data used includes dates, times and location of 9 different types of crimes. As of writing, the raw crime data set had over 250,000 rows[2]. Only data from 2012 to 2015 were considered since 2016 data was incomplete. Furthermore, even if data would have been available until the

[1] `http://github.com/EstadisticasPR/Examples-for-Data.PR.gov`

[2] `https://data.pr.gov/Seguridad-P%C3%BAblica/Mapa-del-Crimen-Crime-Map/bkiv-k4gu`

end of 2016, many current cases were still under investigation. Vintage 2016 U.S. Census annual population estimates for the period of interest were also retrieved [3]. The quality of the crime data was assessed via checking for duplicated entries and comparison of annual number of murders found in other sources.

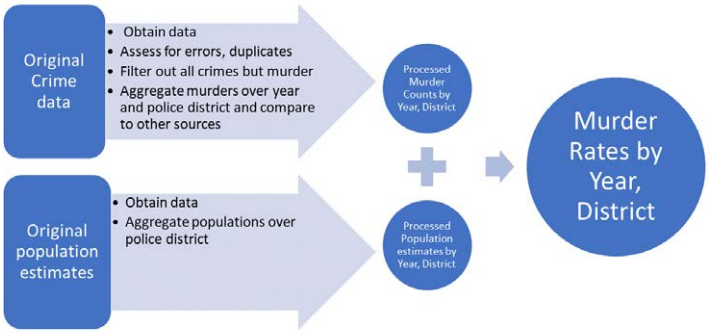


**Figure 1**: Steps required to calculate murder rate data for Puerto Rico by year and police district.

Initially a comparison of yearly murder rates in 4 police districts (San Juan, Fajardo, Ponce and Mayaguez) in Puerto Rico is performed. It is found that San Juan has the highest murder rate among the 4 police districts with results of 49.6, 42.2, 43.1, and 33.2 murders per 100,000 people for 2012, 2013 2014, and 2015 respectively. As in the other 3 regions, the murder rate in San Juan has been decreasing. Without prior knowledge about murder rates it is still hard to grasp the meaning of the San Juan murder rates. Thus, San Juan murder rates are compared next to cities of at least 250,000 people in the United States. To get this stage of the case study going, students are asked which U.S. cities have the highest murder rates. New York city, Los Angeles and Chicago are frequently mentioned. Next, the audience is presented with Table 1, which lists the cities with the five highest 2015 murder according to the 2015 FBI violent crime report[4], San Juan would place fifth highest among all large cities.

| St. Louis | Baltimore | Detroit | New Orleans | San Juan |
|---|---|---|---|---|
| 59.2 | 55.37 | 43.82 | 41.68 | 33.2 |

**Table 1**: Top five 2015 murder rates (per 100,000 people) in U.S. cities and San Juan, Puerto Rico.

Given the status of Puerto Rico as a commonwealth one can compare the overall island murder rate with U.S. states and with other countries as well. Table 2 summarizes the murder rates for Puerto Rico, New York, Florida and California. The statistics for the island are always over 3 times higher than these benchmarks. Moreover, the 2014 murder rate in Puerto Rico was almost twice as high as that of Louisiana, which was the state with the highest murder rate in the U.S. back then.

[3] `https://www.indicadores.pr/Demographics/Estimados-anuales-poblacionales-por-municipio-y-Pu/8ey7-awsw`

[4] Source: `https://ucr.fbi.gov/crime-in-the-u.s/2015/crime-in-the-u.s.-2015/tables/table-6`

| Year | Puerto Rico | Nueva York | Florida | California |
|---|---|---|---|---|
| **2012** | 26.9 | 3.5 | 5.2 | 5.0 |
| **2013** | 25.3 | 3.3 | 5.0 | 4.6 |
| **2014** | 19.3 | 3.1 | 5.8 | 4.4 |
| **2015** | 16.9 | – | – | – |

**Table 2**: Murder rate per 100,000 residents in some of U.S. states and Puerto Rico:

The United States overall 2015 murder rate was about 4.9 murders per 100,000 people, up 10% from the previous year, yet still far below the 16.9 murders per 100,000 people in Puerto Rico. Now, it is of general knowledge that the United States has a violence problem of its own. To put things in perspective, many European countries have murder rates below 1 murder per 100,000 people while Honduras had 84.6 murders per 100,000 people in 2014. In summary, the good news is that murder rates indicate a decrease in violence in Puerto Rico. But much work is needed. The case study is finalized with a discussion of the results, including the limitations of the procedure. Among the limitations: no expert in crime was involved in the study, murder is not the only measure of violence, and variables such as population density and economic hardship play a role in violence but were not included in the presentation.

### 3.2 Do police searches during police stops in San Diego depend on driver's race?

The students are presented with a case study aiming to visually assess the question that is the title of this section. Students are allowed to argue why they think police searches depend or do not depend on race before using the data. Police stops in San Diego from their open portal data is shown to the students[5]. Whether gender has any impact in the results is also assessed. Observations with missing values in race, search or gender were removed from the data set.

The statistical concept to be taught is data visualization, specifically bar plots summarizing multiple categorical variables. However, this case study can very well be applied in the context of probability, or inference on two categorical variables. The original data consists of over 100,000 incidents of traffic stops. After exploring the data, one of the things of notice is that the race variable is rather specific (Korean, Japanese, Indian, etc.). To simplify our task race is reclassified into 5 categories: Blacks, Hispanics, Whites, Asians and Others.

Figure 2 summarizes police vehicle stops in San Diego in 2016 by race and whether the driver was searched. Overall, White drivers were stopped the most but it is explained to students that this does not mean that White drivers are more likely to be stopped by police than drivers from other races. Assessing likeliness of being stopped by race requires using other data such as time of day of police stops. Drivers who are Black or Hispanic appear to be searched more often than drivers from other races. Specifically, for Black drivers the red bar covers more of the entire race bar than for White drivers. The same can be said for Hispanic drivers.

[5] `https://data.sandiego.gov/datasets/police-vehicle-stops/`

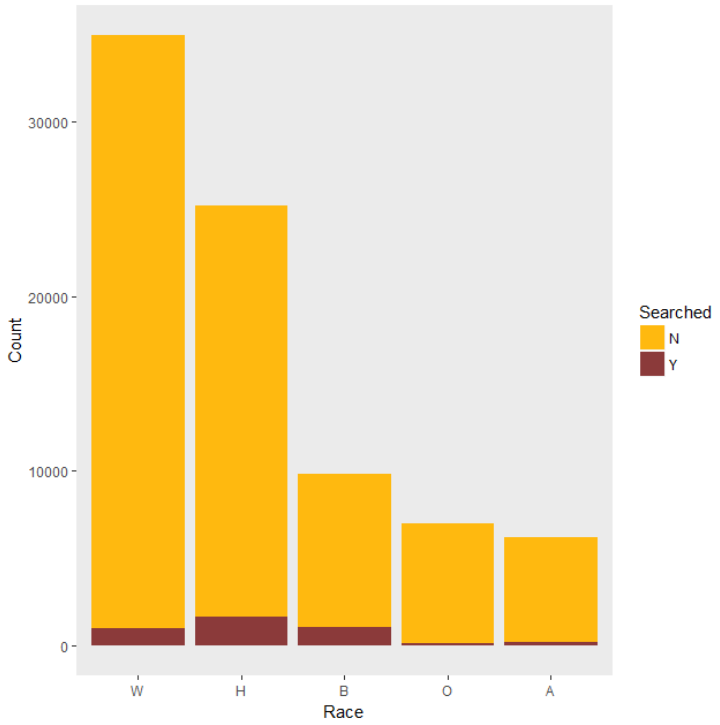


**Figure 2**: Number of police stops and searches by race and gender in San Diego in 2016. Race categories are: A - Asian, B - Black, H - Hispanic, O - Other, and W-White. Searches are: Y - Yes, N - No.

Figure 3 summarizes police vehicle stops in San Diego in 2016 by race, gender and whether the driver was searched. The class is asked to interpret the stacked bar chart, at first comparing genders and then just focusing on men. In summary:

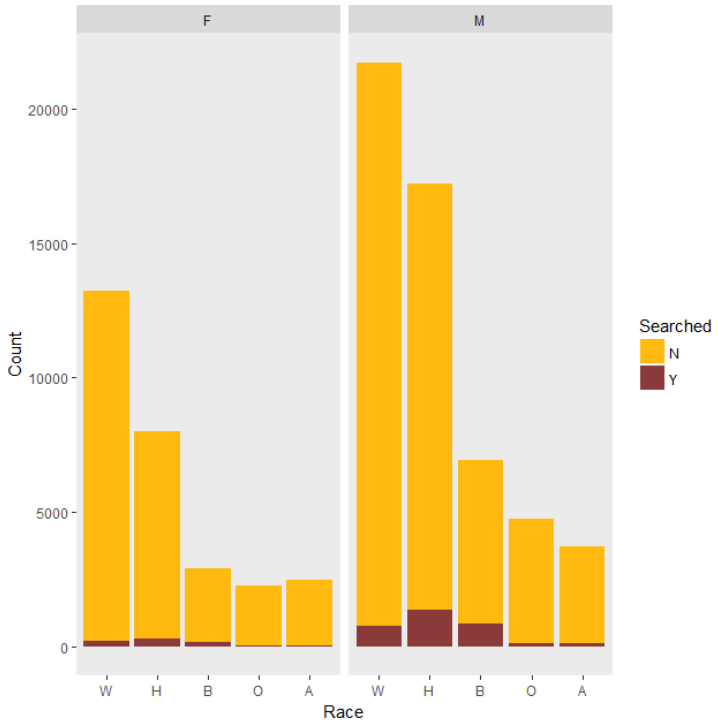


**Figure 3**: Number of police stops and searches by race and gender in San Diego in 2016. Race categories are: A - Asian, B - Black, H - Hispanic, O - Other, and W-White. Genders are F - Female, M - Male. Searches are: Y - Yes, N - No.

- Figure 2 indicates that searches occur more frequently when the driver is Black or Hispanic.
- Overall, men are generally stopped more frequently than women with White men being stopped the most (this is a perfect opportunity to joke about the misconception of women being reckless drivers.).
- Drivers of other races do not tend to be stopped nor searched too often.

To have a better idea of what is suggested by these bar charts, numbers are needed. The percent of searched drivers according to whether they were Black, Hispanic or White was 12.4%, 7.84%, 3.57% respectively.

Limitations of the completed statistical procedure must be pointed out. The assessment does not answer why there is such a discrepancy in searches of drivers by race. For one, the summary does not consider the cause for the traffic stop, or location in San Diego. Also, the data does not include the race of the officer which may (or may not) be associated to the chance that a traffic stop involves a search. Furthermore, statistical inference would be needed to draw conclusions from the current data. The presentation is followed by a discussion.

### 3.3 Other Open Portal Data Based Ideas

Many other case studies can be created using open portal datasets. An introduction to time series analysis could be presented using hotel registration data (Figure 4). Data visualization has received a lot of attention lately (Hullman et al. 2015; Nolan

and Perrett 2016). Hans Rosling showed how effective data visualization can be in engaging the audience. More frequently traditional graphics are being extended by adding additional dimensions: more variables (Figure 5). This has made displaying complicated information more aesthetically appealing (Figure 6).

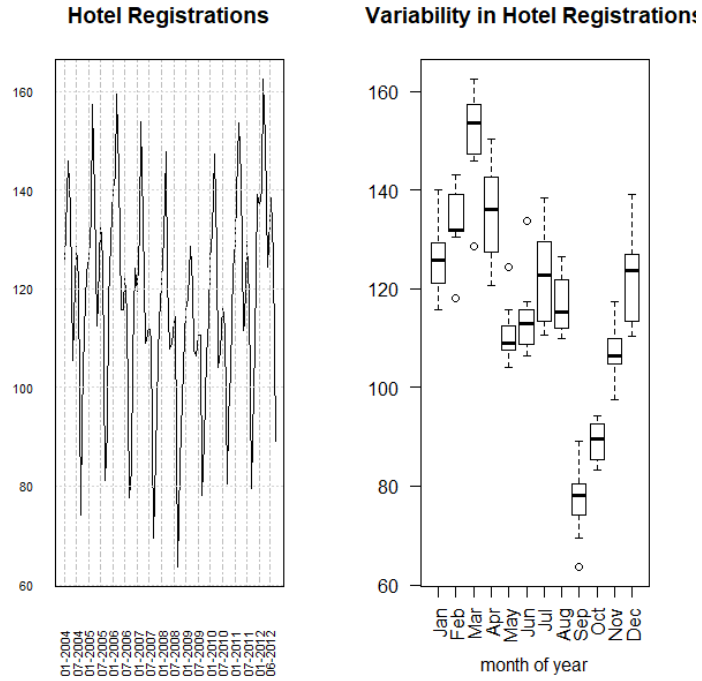


**Figure 4**: Left panel displays the Time series plot of nonresident hotel registrations in Puerto Rico (in thousands). Right panel shows boxplots of nonresident hotel registrations (in thousands) by month of year.

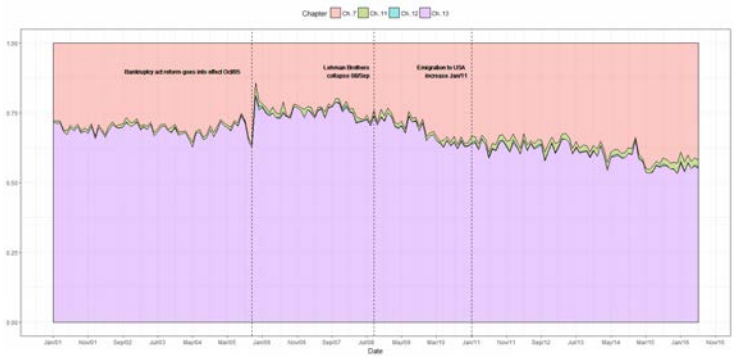

**Figure 5**: Distribution of Four Bankruptcy Chapters in Puerto Rico from 2001 to 2016.

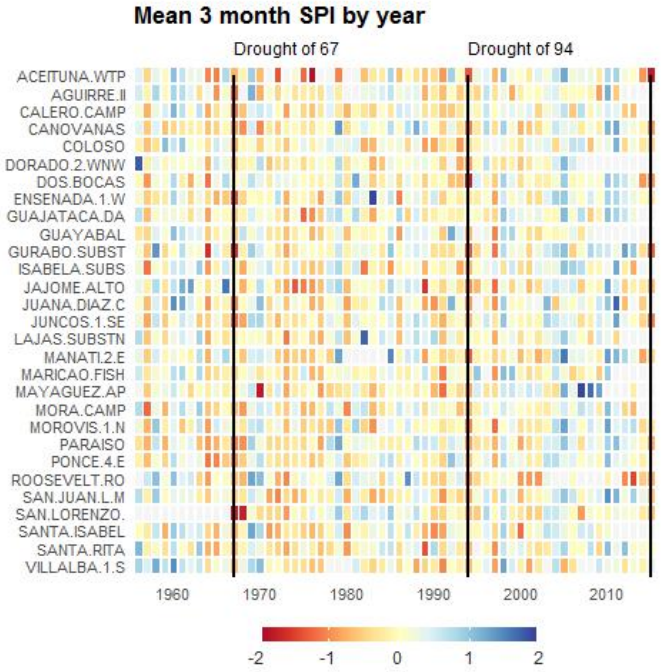


**Figure 6**: Annual average of Standard Precipitation Index (SPI) for accumulation of rain every 3 month across many weather stations in Puerto Rico. Each line represents a station and each column indicates a year. Period covered is 1956 thru 2015. Negative SPI indicates below average rain at the weather station while positive SPI indicates above average rain.

Interactive plots have generated a lot of buzz. For one case study, a motion chart was created by combining several different datasets[6]. Users can select variables to compare and track how the measure of a location changes in time. While aesthetic appeal is important and certainly catches the attention of students during class, one of the key roles of visualization techniques in data analysis is to present information effectively. On occasion, attractive plots (Figure 7) may not be as effective in providing information as more basic visualization techniques of the same data (Figure 8).

[6] `http://motionchartpr.blogspot.com/2017/03/jsdata-function-gvisdatamotionchartid19_28.html`

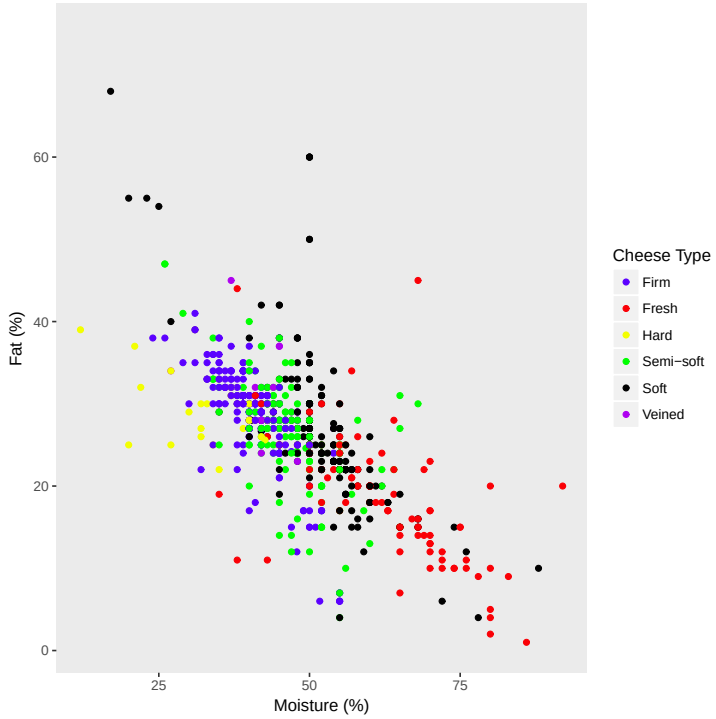


**Figure 7**: Scatterplot of chesse fat percentage and moisture percentage color coded by cheese type using data from the Canada open data portal. It is hard to tell in this chart if cheese type has any influence in the association between fat percentage and moisture percentage.

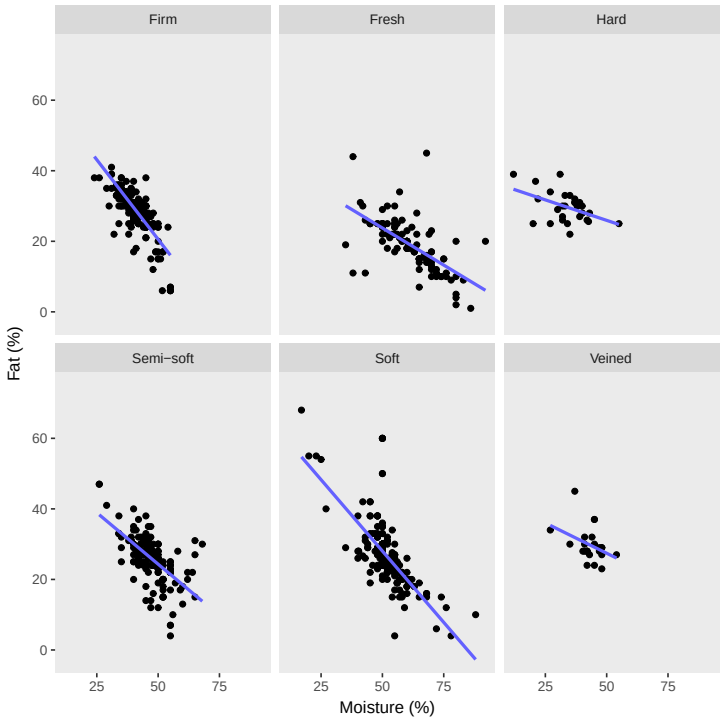


**Figure 8**: Scatterplots of chesse fat percentage and moisture percentage per cheese type using data from the Canada open data portal. It is easier to tell in this chart that cheese type has some influence in the association between fat percentage and moisture percentage.

The sample case studies presented in this article only scratch the surface of the possible ways open portal data can be incorporated into courses. The San Diego police stops data presented in section 3.2 could be used while discussing empirical probabilities or hypothesis testing on the dependence of two categorical variables. The Canadian cheese data could serve as a platform to teach multiple linear regression with interactions. Table 3 presents a few more ideas of statistical methods that can be taught using open portal data. The New York city emergency response time data set can be processed to perform ANOVA, initially to study the effect time of day treatments has in cardiac arrest response times. The 5 million plus raw data would be filtered to just below 60,000 observations, serving as an opportunity to discuss the impact of large sample sizes on hypothesis testing. Moreover, residual diagnostics would show non-normality and the data set can be reused when teaching the Kruskal-Wallis test. In advanced courses, a more sophisticated model can be fit to the response times for the purpose of flexibility and accounting for more predictors available in the data set.

| **Data** | **Location** | **Topic** | **Comments** |
|---|---|---|---|
| Lead in water | Toronto | Quartiles | Data is censored. |
| emergency response time for cardiac arrest | New York City | ANOVA | Large data makes it easy to reject null. Skewed distribution indicates nonparametric tests are best. |
| fatal shootings | Philadelphia | two proportion inference | confidence interval or testing |
| socioeconomic data | Chicago | multiple linear regression | Associations may be nonlinear |
| meteorite mass | World | shape of distribution | Choose specific classification |
| tourism employment | Canada | time series | Seasonal component and global trend present. |

**Table 3**: A few more ideas for incorporating open portal data in introductory stats.

Case Studies are not the only pedagogical tool that can be developed from open portal data. Class projects are also possible. Open portal data can also have a role in homeworks, quizzes or exams. For example, the class can be assigned a data set to perform a statistical procedure. Assignments can be individualized by asking pupils to take a sample of size $n$ based on, say the last two digits of the students identification number. For example, if a student's last two digits is 14 and $n = 50$ then the student must perform the assignment using observations from row 14 to 63. This way, the instructor can create R code to efficiently reproduce each student sample and grade the assignments.

## 4. Discussion

Open portals provide an excellent opportunity to integrate real data with context and purpose into introductory courses, as recommended by the 2016 GAISE College Report. The more localized the data, the more interesting students will find its use in class. Specifically, there is crime data from Los Angeles, New York and other cities with which a case study analogous to section 3.1 can be built. Similarly, police stops data is available for Austin Texas and other locations. R code for some of the case studies presented can be found online[7]. For other codes please contact the first author. Another benefit of open portal data is that it is often raw and very large. Thus, students get a first look into data science and big data with such data sets. For example, the New York city emergency response time raw data has over 5 million records, needs to be filtered, and period of day factors must be created. Although the emphasis in this article is introductory statistics courses, the applications proposed can be modified accordingly for more specialized courses.

There are some challenges to be aware of when incorporating open portal data into the classroom. Preprocessing is often needed and this can be time consuming. Specifically, the data may have errors or duplicates. It might be required to aggregate measurements, filter, create new variables, or combine data sets as seen in section 3.1. Challenges also arise when it is not clear where the data comes from, which puts the reliability of the data in question (Rivera 2016); or when there is no variable dictionary available with the data set. Depending on the data set and context of a case study, parameters can be obtained not statistics (e.g. mean number of inmates in New Orleans during October 2015.). Some data sets are very large (e.g. New York city emergency management system data) and can not be opened on a regular computer, and some open data websites do not work properly on some web browsers. Since a few data sets are routinely updated in open portals, for certain tasks it is wise to ensure everyone is using the same version of the data by providing a downloaded version. Furthermore, the data may hold hidden surprises. For example, lead samples of tap water from Toronto are available online. A potential application is regression by zip code, or ANOVA. However, the lead measurements are censored, requiring more advanced procedures for inference than seen in introductory stat courses. Instructors are advised to carefully explore the data before using it in class. Part of the idea is that the data is realistically complicated, but not too complicated for academic use.

The premise of making data available through open portals is that it helps the economy. Use of open portal datasets for statistical analysis and to teach statistics courses can encourage open portal managers to share data in a more efficient way. In principle, enough academic interest in open portal data could lead to improvements

[7] `http://github.com/EstadisticasPR/Examples-for-Data.PR.gov`

in open portal platforms: more data, better accessibility, and faster updating of data.

## References

Baumer, B. (2015), "A Data Science Course for Undergraduates: Thinking With Data," *The American Statistician*, 69, 334–342.

Chew, P. K. and Dillon, D. B. (2014), "Statistics Anxiety Update: Refining the Construct and Recommendations for a New Research Agenda," *Perspectives on Psychological Science*, 9, 196–208.

— (2015), "Statistics anxiety and attitudes toward statistics," in *D. Chhabra (Ed.), Proceedings of the 4th Annual International Conference on Cognitive and Behavioral Psychology (CBP 2015). Singapore.*

Cruise, R. J., C. R. W. . B. D. L. (1985), "Development and validation of an instrument to measure statistical anxiety," *Paper presented at the annual meeting of the Statistical Education Section, Chicago, IL.*, 92.

GAISE (2016), "Guidelines for Assessment and Instruction in Statistics Education College Report," Tech. rep., ASA Revision Committee.

Grimshaw, S. (2015), "A Framework for Infusing Authentic Data Experiences Within Statistics Courses," *The American Statistician*, 69.

Horton, N. J. and Hardin, J. S. (2015), "Teaching the Next Generation of Statistics Students to Think With Data: Special Issue on Statistics and the Undergraduate Curriculum," *The American Statistician*, 69, 259–265.

Hullman, J., Resnick, P., and Adar, E. (2015), "Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences About Reliability of Variable Ordering," *PLOS ONE*, 10.

Keeley, J., Zayac, R., and Correia, C. (2008), "Curvilinear relationships between statistics anxiety and performance among undergraduate students: Evidence for optimal anxiety," *Statistics Education Research Journal*, 7, 4–15.

Neumann, D. L., M., H., and Neumann, M. M. (2013), "Using Real-Life Data when Teaching Statistics: Student Perceptions of this Strategy in an Introductory Statistics Course," *Statistics Education Research Journal*, 12, 59–70.

Nolan, P. and Perrett, J. (2016), "Teaching and Learning Data Visualization: Ideas and Assignments," *The American Statistician*, 70, 260–269.

Onwuegbuzie, A. J. and Wilson, V. A. (2003), "Statistics anxiety: Nature, etiology, antecedents, effects, and treatmentsA comprehensive review of the literature," *Teaching in Higher Education*, 8, 195–209.

R Core Team (2016), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Ridway, J. (2016), "Implications of the Data Revolution for Statistics Education," *International Statistics Review*, 84, 528–549.

Rivera, R. (2016), "A dynamic linear model to forecast hotel registrations in Puerto Rico using Google Trends data," *Tourism Management*, 57, 12–20.

Williams, A. S. (2010), "Statistics Anxiety and Instructor Immediacy," *Journal of Statistics Education*, 18, 1–18.