LOOKING TO THE **PAST** TO PREDICT THE **FUTURE**

Today, information is **abundant** yet often **complex**. Statisticians play a **pivotal role** in harnessing data to unravel patterns, trends, and insights in numerous domains to **extract meaningful predictions**.

Here, in celebration of Mathematics and Statistics Awareness Month, **American Statistical Association members** with expertise in climate science, sports analytics, epidemiology, and economics **answer the question:**

What data set would you suggest for high-school students and their teachers if they want to learn about how predictions are made **in your area of expertise?**

BO LI, UNIVERSITY OF ILLINOIS

You can easily download the local data specific to your area from the **National Oceanic and Atmospheric Administration Climate Data Online** (*www.ncei.noaa.gov/cdo-web*). I would recommend starting with **daily temperature data from a selection of weather stations**.

RON YURKO, CARNEGIE MELLON UNIVERSITY

The SCORE Network maintains a data repository (*data.scorenetwork.org*) with data sets and prediction problems across a variety of sports, such as forecasting tennis **match outcomes**. It is also easy to access data using various packages in the R programming language listed in the **CRAN Task View: Sports Analytics** (*cran.r-project.org/web/views/SportsAnalytics.html*).

MICHAEL JACKSON, RICE UNIVERSITY

I think **macroeconomic data** would be a great place for high-school students to start with and **potentially S&P 500 data** (*bit.ly/SandP500Data*) so they can look at the market as a whole. And potentially **fundamental data from the top 10 stocks** on the S&P 500. I think **Wharton Research Data Services** (*wrds-www.wharton.upenn.edu*) has a good data set that can be used for academic purposes.

ALEXANDRA HANLON AND TANNER BARBOUR, VIRGINIA TECH

Given that we are often asked to create clinical prediction models for cancer **patient outcomes as a function of baseline demographic and clinical factors**, a suggestion would be the **SEER database** (*seer.cancer.gov*). I also recommend the **UK Biobank** (*www.ukbiobank.ac.uk*) as a resource for **building phenotypic profiles** of health outcomes.

GET THE **TOOLS** OF THE **TRADE**

These **ASA members** also weighed in on the best tools for beginners to use for **creating visualizations** with data from sets like those mentioned above. There was a unanimous answer: **R**, along with several associated packages such as **ggplot2, plotly,** and **ggvis**.