Survey data analysis
Week 1 Class Exercise
Peter Lugtig p.lugtig@uu.nl

For this exercise we will use a database of the U.S. 2016 presidential election between Trump and Clinton. Hundreds of polls were conducted between 2015 and November 2016 when the election took place.

A database of all the polls, along with some characteristics of the polls have been assembled by the people behind www.fivethirtyeight.com They have made the data available as a .csv after which Thom Volker (2nd year MSBBSS student in 2020) created a Shiny App using R. The Shiney app is available through https://utrecht-university.shinyapps.io/SDA_shinyelectionbias/

The shiny app allows you to plot all the election polling data in different ways. It will always show you the raw polling percentages (in orange) and the adjusted percentages (after some correction for nonresponse/selection bias). The black dotted line gives you the true value (the vote in November 2016).

You can adjust several setting to adjust what polls are being visualized: You can choose to vary:

- State. Polls showing the entire USA or one of the individual states
- Trump and/or Clinton or the difference between the two (as in who will win).
- An independent variable (plotted on X-axis). Here you can select to either the date the survey was done, the sample size or the population of the survey (likely voters (people who say they will vote) and registered voters (in the USA you need to register yourself before you can vote. Finally, the grade of the survey. This is a subjective rating by the people of Fivethirtyeight of the quality of the polls. See https://projects.fivethirtyeight.com/pollster-ratings/

Take home exercise: Try to answer the following questions. Write your answers down just for yourself to share with the class. The easiest dependent variable to work with is the Difference (Trump-Clinton).

1. Was Trump underestimated by the polls?
2. … in the swing states? (Michigan, Wisconsin, Pensyllvania)
3. Did the quality of the pollster matter?
4. Was there a difference between sampling likely voters and registered voters?
5. Are larger polls better?
6. Is there a difference between raw and modeled polls?