Projects

There are seven projects described in this section. Each project will be assigned to at most two groups. Discuss with your group and email me an ordering of the projects in order from most preferred to least preferred. I will try to assign to each group one of their top choices.

1. US Census Data
2. Google Transparency Report
3. Course Evaluations
4. Diamond Prices
5. Mammal Sleep Patterns
6. BRFSS
7. SAT data

US Census County Data

Contains information on all US counties based on the 2010 census. This is the counties dataset that we have been using. You should focus on some part of the data that interests you. For instance you can focus on an in-depth comparison between a few states. Or conversely you can continue looking at all the states but focus on a small set of columns/variables.

A description of some of the variables: notes/countyExplanations[1]

Google Transparency Report

Contains information about requests for account information that Google has received from various countries. Records variables related to the country's demographics, politics and freedom of the press, as well as the number of requests received and complied with.

Brief description: https://www.openintro.org/stat/data/?data=goog

Data: http://www.openintro.org/stat/data/goog.csv

Course Evaluation Data

Data from the Hanover College course evaluation data for the year 2016-2017. This is a subset of the actual answers provided by the students, but with identifiable information removed.

Brief description: studentEvalsDescriptions.md[2]

Data: Ask me for access to the data if you want to do this project. This data is not to be shared with anyone else.

[1] countyExplanations.html
[2] studentEvalsDescriptions.html

Diamond Prices

A dataset called "diamonds" is included with the ggplot2 package and is available via our hanoverbase package. It contains prices, quality, color and size information for over 54000 diamonds. You can discuss various aspects of this dataset, but your focus should be how the "four Cs" variables (carat, cut, clarity, color) affect the price of diamonds. In particular, you should provide an explanation of why it appears that diamonds of "Ideal cut" appear to have relatively lower prices than diamonds of "Premium cut", which is a lower-quality cut. You will need to investigate how those 5 variables relate in various ways.

Use data(diamonds) to load the data, and help(diamonds) to bring up the documentation.

Mammal Sleep Patterns

A dataset called "msleep" is included with the ggplot2 package and is available via our hanoverbase package. It contains information about the various species of mammals related to their sleep cycles. This is a fairly open-ended dataset.

Use data(msleep) to load the data, and help(msleep) to bring up the documentation.

BRFSS

The brfss dataset that we have been using in our assignments contains many extra variables that we have not considered. You can choose to focus your project on some subset of these variables.

SAT Data

The dataset SAT, which is part of the mosaicData dataset and included with our hanoverbase package, contains state-by-state information related to the SAT exam. You are free to explore the dataset in various ways, but at the very least you should explore the relation between verbal and math scores, as well as the relation between the frac and sat variables.

Use data(SAT) to load the data, and help(SAT) to bring up the documentation.