Capstone Project 1: Milestone Report

Overview:

With this first project my goal was to find a dataset that is molecular biologically relevant while not being a niche dataset or something that requires specific knowledge to analyze (such as bioinformatics datasets where DNA/RNA/protein sequences are analyzed, in order to analyze these datasets, the person needs field-specific knowledge i.e. what sequences to look for, what sequences not to look for, what sequences are junk and needs to be removed etc.). This quickly became much harder than I initially imagined because not only is bioinformatics dataset the most prevalent biological dataset out there (with the second being ecological datasets, which does not serve me well), finding a molecular biology dataset with sufficient amount of data was extremely hard. Most of the molecular biology datasets has only a few hundred data points, which are enough for a molecular biology paper, but it is nowhere near enough for this project. Ultimately, finding a dataset that has both an molecular biology appeal and a big enough depth for data analysis was the most challenging aspect of this part of the project.

The Problem:

The one thing that differentiates survey biological datasets from other datasets is that, sometimes there is no question going in; that sometimes, the question comes about while exploring the data itself. In addition, with survey data the problem is naturally that, it's a survey data, there are too many random variables to account for, for example: truthfulness of the response, correlative nature of the survey itself (no causation, at best correlation) and how general the data is generated (from a person's response, not by a scientific instrument or device, as such the results are variable to a considerable degree, as different people can give difference responses to the same question). However, given that this is the only dataset of value

that I can find, it is the dataset I decided to use: Alzheimer's Disease and Healthy Aging Data. The problem initially is unknown, because the dataset itself contains the health data of older adults in US lands (50 years or older). However, after some initial data analysis, there quickly emerged a pattern: only 2 out of the 20+ questions included a comparison between male and female older adults, all other questions are simply centered on information of the various aspect of the health of older adults across the US.

At this juncture I encountered a problem: how can I generate an interesting problem when the dataset itself is very broad and general? Obviously I cannot go back and spend couple more weeks to find another dataset, I already spend 3 weeks on this and it was a very exhaustive search (both in terms of the scope and the level of my exhaustiveness). As such the only way forward is to examine the dataset very thoroughly and using additional external data (such as a State's population, economy, crime rate etc.) to generate at least some interesting stories about the dataset.

The problem I am trying to find in this dataset ironically is the dataset itself, but now that I have a path forward it is time to work on the dataset.

Description of Data:

The dataset itself is a .csv file with over 35 columns and over 46,676 rows of data. Some of the columns contains duplicate information.

Data Wrangling:

Initially I sorted the dataset by State location and removed some columns that contained unnecessary information. However, I soon found that sorting by the type of Questions asked makes much more sense. I removed some more columns and I also removed all rows that did not contain any values for the columns Data Value, Low Confidence Limit and High Confidence Limit. After this a list of Questions asked was generated and data analysis began.

Data Analysis:

The most interesting of the data revolves around the question “percentage of older adult men/women who are up to date with select clinical preventive services”. This is because it allows for the comparison between two groups over a specific topic: overall is older men or women in the US are more up to date with select clinical preventive services?

The distribution of the two datasets are strikingly similar, both are roughly normally distributed as such no transformations are needed before applying any statistical tests. From the plot here I noticed that overall men are more up to date with select clinical preventive services, with a range between 15 – 38 whereas for women the range is 9-32 approximately. For my analysis I employed a T-test as I want to calculate whether the means of the two datasets are significantly different from each other. I calculated all of the necessary components of the T-test and my final p value is highly significant: 1.87e-10 with a T-statistic of 6.98. This tells me that overall, men are significantly more up to date with select clinical preventive services than women.

For the rest of the questions, I employed real world statistics of US states along with the data in the graphs themselves to generate a story for each of the graphs. For some I managed to generate an interesting enough story, but for some others the information they present are extremely minimal.