Capstone Project 1: Final 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 a molecular biology appeal and a big enough depth for data analysis was the most challenging aspect of this part of the project.

In the end, I settled on a survey dataset by the CDC on older American adults (years 50 or older) about their health issues. The dataset contains 38 columns, with many being useless such as: Data\_Value\_Footnote\_Symbol, TopicID etc. In the end only 3 columns were useful for downstream data analysis: LocationDesc (where was the person located), Question (what kind of question was asked) and Data\_Value (percentage points).

Initially I hoped to do a comparative analysis, where I compare whether the health of older American men significantly differs from the health of older American women for most if not all the questions. Essentially running T-test of the various questions and later, during the machine learning aspect, using clustering to predict whether a given health percentage value belong to a man or a woman. However, in hindsight the survey dataset turned out to be terrible for analysis and even worse from machine learning. Here is why: as a survey, it wasn’t designed like an experiment where a control group is compared to experimental groups (this makes doing statistical analysis very easy because it was designed to have statistical analysis done on it!). The sampling was mixed for all but one question: *Percentage of older adult men/women who are up to date with select clinical preventive services.* This made it difficult to perform any statistical analysis aside from these two questions, because for all the other questions, not only was the men and women were mixed together, there was also few outliers/data point of interest. As such, I had to bring in outside information to complement my analysis – using data such as a state’s population, poverty rate and even economy to try to explain some of the differences in the data.

When it came to machine learning, usually in biology something akin to a regression analysis is the preferred method where for a given treatment X, there exists the response variable Y and the analysis will focus on whether there correlation or a causation between X and Y (the actual order of the polynomial equation can vary). As such, in the beginning I planned on writing a machine learning algorithm akin to the one shown below:

# Import necessary modules

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_squared\_error

from sklearn.model\_selection import train\_test\_split

# Assign X column and y columns into their own variables

X = df['x column']

y = df['y column']

# Create training and test sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size = 0.3, random\_state=42)

# Create the regressor: reg\_all

reg\_all = LinearRegression() #this part can also be adjusted for higher order polynomial equation, the goal is to have an equation that fits the data here

# If wanting to do a higher order polynomial instead of a straight line, do the below: (higher order can be achieved by adjusting the degree)

polynomial\_features= PolynomialFeatures(degree=2)

x\_poly = polynomial\_features.fit\_transform(x)

model = LinearRegression()

model.fit(x\_poly, y)

y\_poly\_pred = model.predict(x\_poly)

# Fit the regressor to the training data

reg\_all.fit(X\_train, y\_train)

# Predict on the test data: y\_pred

y\_pred = reg\_all.predict(X\_test)

# Compute and print R^2 and RMSE

print("R^2: {}".format(reg\_all.score(X\_test, y\_test)))

rmse = np.sqrt(mean\_squared\_error(y\_test, y\_pred))

print("Root Mean Squared Error: {}".format(rmse))

However, the problem with this survey dataset is that there is no correlation or causation between X or Y. As such the machine learning algorithm shown above will not work, as the data here is US territories and a value associated with them (for example, the percentage of older adults who are experiencing frequent mental distress across all US territories). As such, a regression algorithm will not work here, and a clustering algorithm may be more appropriate.

For a clustering algorithm the idea is to be given a value, and then predict what state the value are most likely to come from. For this I tried a K nearest neighbors classifier, as it is the algorithm that I am the most familiar with. For each question I wanted to write an algorithm similar to the one below:

# Import KNeighborsClassifier from sklearn.neighbors

from sklearn.neighbors import KNeighborsClassifier

# Create arrays for the features and the response variable

y = df['Percentage of older adults who are experiencing frequent mental distress'].values

X = df.drop('Percentage of older adults who are experiencing frequent mental distress', axis=1).values

# Create a k-NN classifier with 6 neighbors: knn

knn = KNeighborsClassifier(n\_neighbors=6)

# Fit the classifier to the data

knn.fit(X,y)

# Predict the labels for the training data X

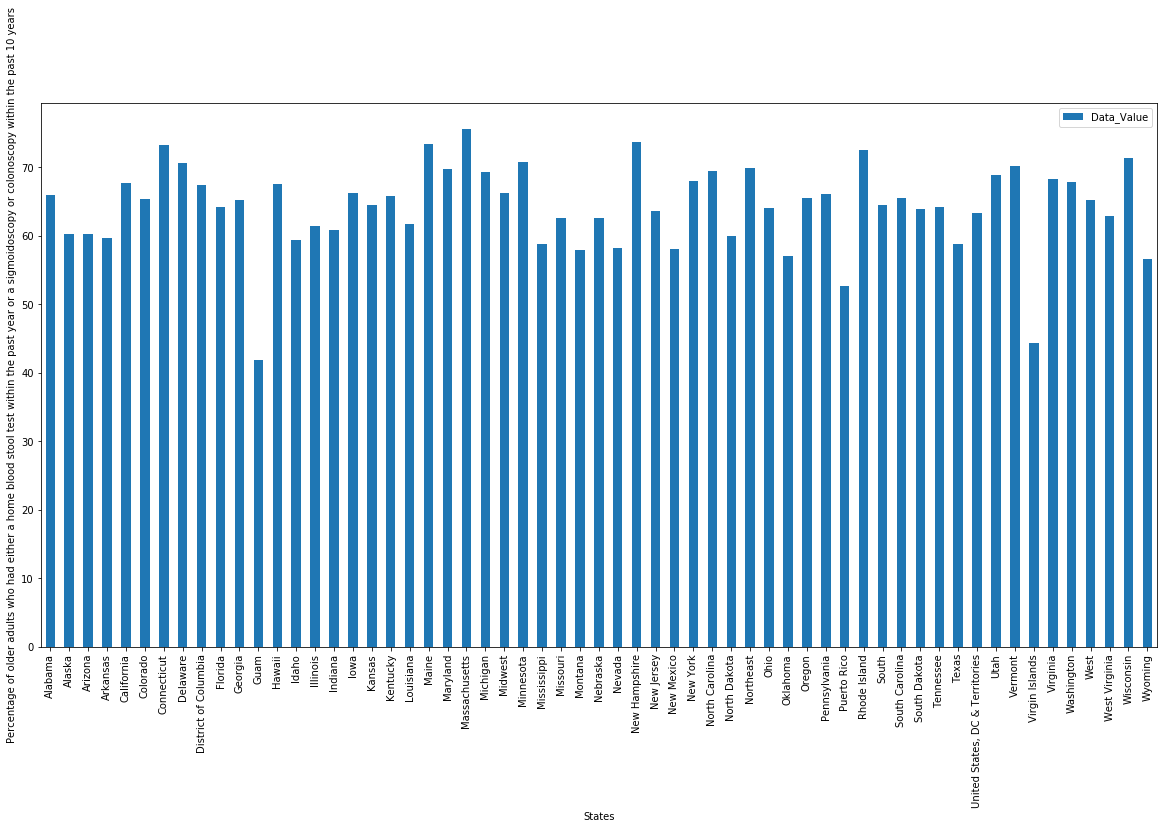
y\_pred = knn.predict(X)

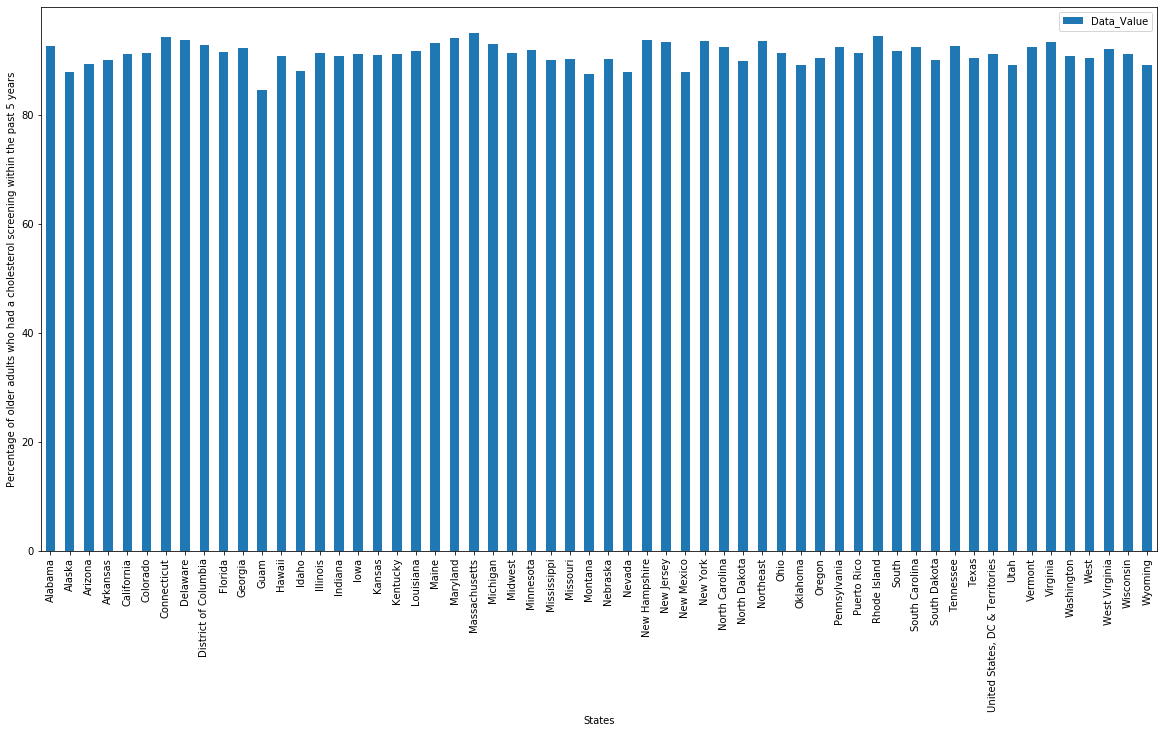
# Predict and print the label for the new data point X\_new

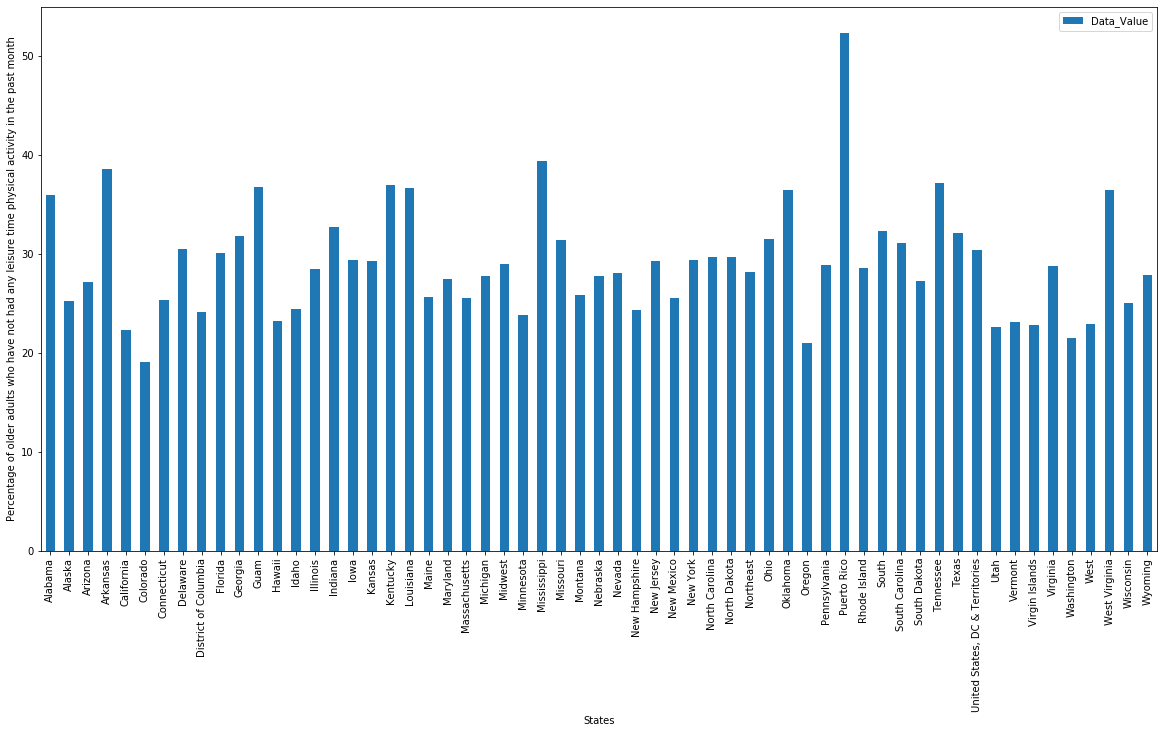
new\_prediction = knn.predict(X\_new)

print("Prediction: {}".format(new\_prediction))

However, this time another problem surfaced, in all of the questions, save for 3 US territories: Guam, US Virgin Islands and Puerto Rico, the rest of the locations are all relatively the same or irregular enough that there is no discernable pattern:







This meant that for both linear regression and classification it is basically impossible to write an algorithm that will work for any of the questions in this dataset. Proving that, once again, this survey data has extremely limited analysis value.

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. This is because for a survey, unlike a carefully designed experiment, it generally doesn’t have a specific focus, i.e. it is not asking one set of questions (does X(s) cause Y or can we predict, given a set of Xs, the corresponding value of Y?). Instead, the survey is focuses on a wide range of shallow questions (have you suffered from disease Z in the past W months?) that makes formulating a question hard going in (unlike, say, a dataset for the login locations for Pokemon Go, where questions like “does the login times/location vary by season?”, “how does the login location vary across different cities?” can easily be asked and analyzed). 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.

In this dataset, 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 question 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.

In hindsight, I most likely would have picked a different dataset, even one that is not wholly biologically related – that’s how hard this dataset was to work with.

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 or useless information. The dataset is comprised of different questions about various health topics and a numerical answer, in percentages. The dataset can be segregated by questions asked, where a distribution of a percentage value across most of US territories can be obtained.

Data Wrangling:

Initially I sorted the dataset by State location and removed some columns that contained unnecessary information (blank cells, for example). However, I soon found that sorting by the type of questions asked makes much more sense, as stated in the section above. 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. In the end a pretty lean set of data for such a big dataset.

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? (This, as stated in the Overview section, was also what I wanted, but could not do for the entire dataset).

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. This is actually really interesting, because generally speaking, men are much more risk prone than women and that men tend to take health issues less seriously than women, so I expected the opposite to be true, that women are significantly more up to date with select clinical preventive services than men.

Here several hidden factors could come into play, one possibility is the type of select clinical preventive services defined by CDC (there is no definition in the questionnaire) where it may be more biased towards men in some way (either the services are shorter or easier to preform on men, or men have an easier time accessing them). Another more realistic possibility could be the earning power difference between men and women. Generally, men tend to earn more than women, this is true across many industries as well as locations in the US. Higher earning power could easily mean more access to clinical preventive services, especially in the US, where the medical costs can be extremely high for something as basic as an X-ray. Also, men generally need less costly clinical preventive services compared to women, where events like childbirth can cause women to have the need, but not necessarily the financial means to certain clinical preventive services.

A counter to this argument, however, could be that if a man is earning a lot, his spouse should be able to “hitchhike” on his health benefits, therefore allowing her to be up to be as up to date with select clinical preventive services as her husband. Therefore, for every male that is doing well financially and therefore has the means to be up to date with select clinical preventive services, there should exist one female as well. This could be especially true given that the people surveyed are older adults who are more likely to be married than not. However, without knowing the number of single men/women surveyed, it is hard to make a definitive conclusion. As such, factors like these prevents an in-depth analysis of the survey results, and any conclusions drawn from the resulting statistical analysis are murky, at best.

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, due to the lack of depth for many of these questions, it was impossible for me to perform statistical analysis and machine learning on them. For some I managed to generate an interesting enough story, but for many others the information they present are extremely minimal.

Conclusion:

Conclusion here.