Milestone Report

1) The Issue

This data set pertains to crime and socioeconomic and law enforcement data for thousands of cities and towns sample across the United States. Knowing the crime rates for a certain city during a certain year is nice when trying to decide where to live, establish a business, invest more money into local infrastructure, etc. These are all great reasons to come up with the best regression algorithm as a method of predicting the amount of violet and nonviolent crimes based on the area. Particularly, I wanted to frame this issue on what communities should be focused on when dividing up resources. The ultimate goal is to be able to reduce crimes across the nation, which leads to an explanation of potential clients.

2) The Client

The best potential client for these findings would most likely be national non-profits whose goals it is to provide money to in need communities. Having money to bolster law enforcement, providing assistance to those most at risk to commit crime in the future, and otherwise increasing the overall standard of living will decrease crime as a whole in each area. The best approach to figuring out how funds should be distributed is to begin to predict which cities would have the highest crime rate for that time period then make a way down the list. Thus, for current data, I have decided to present the areas that need most work based on the available data set. However, a potential client would either have even more updated and inclusive data that what I was able to find, or also be interested in how to potentially gather the most useful data in the future. Therefore, for the client, I will also provide insights on how to maximize data collection in conjunction with the algorithm.

3) The Dataset

Information on the Dataset:

The dataset that I used for this project was compiled by the UCI Machine Learning Repository, and I decided to use the Communities and Crime Unnormalized Data Set. The dataset combines:

- data from the 1990 census
- law enforcement data from the 1990 Law Enforcement Management and Administrative Statistics Survey, originally investigated by the US Department of Justice and compiled by University of Michigan’s ICPSR, and
- data from the 1995 FBI Uniform Crime Report

The dataset mostly contains independent variables, hundreds in fact, most of which had to be eliminated or cleaned up in steps detailed shortly. There were also some categorical data that could not be included during regression but were important for reincorporating back into a user-friendly output. Thankfully, the dataset was already organized in such a way that the independent and dependent variables were easy to separate. However, despite this data being collected and organized well, the data contained within the dataset was far from perfect.

Cleaning and Wrangling:

- Merging

The format of column names needed to be manipulated first before being able to merged. I grabbed the list, placed it in excel, then separated out the data using Excels “Text to Columns” function, which allowed me to pull just the column names, and separate out other extraneous characters including the definitions of said columns, which I know that I’ll need when I do the final data analysis and wrap up, but is not necessary for data analysis right now, and therefore was excluded. With a little bit of massaging, I was able to create a CSV of column names, which I then uploaded to the Python notebook and saved as the original dataframe’s column names.

- Dealing with Missing Values

I noticed when I was first importing the data that there were missing values in the data which seemed to be denoted by “?”s. After in fact checking that they were strings of question marks, I then wrote a function to count the number of ?s(missing data in this case) to count the number of columns with missing data. I noted after running this function that there were a few data sets that jumped out at me, with large amounts of missing data. One of these was a per capita column, where my intention is to fill that in with a possibly weighted average of other per capitas in that community area using the average of other races in the area to weight them. One other staggering thing I noticed was that of the 2215 rows of information I had, 1872 of them were missing police data, a whopping 84.5% of total police data was denoted as “?”. With that much missing data, I did not feel comfortable wanting to keep that in my data set, and instead eliminated the columns entirely. Trying to make predictions based on such a small pool would most likely lead to inaccurate results. I decided to drop the 22 columns that had the vast majority of the data missing, most of which had to do with police census type data, and decided to just predict with the other 100+ columns. I used a manual drop, though I could have used a function to find where it said police or polic, since the spelling used was inconsistent, I decided to do it manually as to not accidentally keep or drop wrong columns.

- Dealing with Irrelevant Data

Although this was already denoted when I first retrieved the data set, it wasn’t actually clear what the irrelevant data was going to be until I successfully put the column names back into their associated columns. At that point, it was then I saw that there were going to be identifying variables, which would help in the end stages, but most likely hinder early data analysis. I knew that they would not need to be included in the analysis, but I did not believe they would need to be removed entirely. These columns included the community name, county code, and community code.

In addition, it was noted in the data gathering portion that the last 18 columns/variables would be the ones to be predicted, but I did see that they already had data in them. This could be useful to check how accurate the algorithm is at the end, but serve no purpose in the early data analysis, nor in the machine learning portion. While I plan on leaving them in the dataframe for the time being, they will not be used in the final product, other than to compare accuracy.

- Checking Data Values

I began doing this by first grabbing the columns that would be the easiest to check. For me, this was all the columns containing a percentage of some sort, which all began with “pct” which made it very easy pull all relevant columns into a new dataframe. At this point I wanted to make sure that the values were all between 0-100. Anything outside of those boundaries would be incorrect data and would need to be omitted. I wrote a np.where to check where the percentage would've been outside of these boundaries and an empty array was returned to me, so all data within that percentage dataframe was within acceptable boundaries. Will be doing similar things with the other columns.

In addition, I used a bit of graphical analysis to try to find outliers, and it was difficult to figure out down all columns so I will be trying to narrow down, and using boxplots and histograms to find outliers that would severely skew the data and eliminating them from the dataset.

4) Additional Datasets

The dataset itself notes that using this data alone to evaluate communities is a bit over simplistic since a lot of data is missing that has a lasting effect on crime rate, such as more visitors in a city than one doesn't have many visitors come in since the crim rate is based on per capita population and visitors would not count towards the resident population but could count towards the amount of crime that occurs. Therefore, to truly have a more thorough look at assessing more at-risk communities, we would need additional corroborating data. It would be impossible to determine all of the variables that would affect the ultimate data, but certainly additional community data would be useful, and entirely beneficial as it would help make the prediction algorithm even stronger.

I understand that the dataset is relatively old at this point, and of course the best-case scenario is to have the most updated data possible, so a massive improvement would to be to incorporate all new data as well. The most current data would be to be able to gather the newest census data at the time, and gather crime data by the year from the FBI, but law enforcement data would be much harder to gather. Testing would show whether or not the law enforcement data is important enough to absolutely necessitate when data gathering. However, my goal with using the existing data was to determine the best algorithm to use. Although rates and data within the communities may change over time, the contributing factors should not, thus all that would need to be changed is running the updated years' data into the working algorithm.

5) Initial Findings

Most of what was done to this dataset prior to working with the algorithm was exploring the dataset as well as some rudimentary statistical analysis. The exploration was done graphically to see whether I could pinpoint any variables before actually working deeply with the data that would significantly affect the end goal, which was crime rate. I looked at the homeless/shelter population, how urban vs rural communities looked, the poverty rate, unemployment, and median income which I guessed might be important factors in crime rate, but would not be able to confirm until testing was done.

During statistical analysis, I realized that the amount of inferential stats to be done with this data was limited at best, since showing how closely related two independent variables were was not an end goal in this project, and there were far too many independent to dependent variable combinations to test for significance of each one with t tests, so inferential statistics for the project is as follows: checking the crime rates, both violent and nonviolent are statistically significantly different from known national averages, seeing if there is a statistically significant difference between the rates of violent and nonviolent crime rates themselves, and in addition, looking at single correlation levels of variables that would be used later on in the machine learning portions.