MSE 226 Project Part 1

Zhaoyu (Joe) Lou and Carina Zhang

Dataset

The dataset we are using is a combination of two data files, one on county level demographic information and one on election outcomes. The demographic information focuses on a few major aspects, including racial composition, educational attainment, age, housing, income, business ownership, and economic status. The election outcomes contain voting results from the 2016 presidential election, keeping track of the party affiliations of candidates and the proportion of the county population for each candidate. These files contain information on each of the 3007 counties in the US, and there are a total of 53 covariates. The election data was collected from the 2016 election, whereas the majority of the demographic information came from the 2014 US Census. For some covariates, the dataset also looks at differences between the 2014 and 2010 census data. We will need to do some preprocessing work to join the datasets together and format them to use a common standard naming convention, but that should be manageable. Given that it is published by the federal government from the results of a census, we believe that it is probably as accurate as possible for any demographic dataset we might be able to find. One potential concern about the data is that there is a two year time delay between the demographic data collection and the election data. While this will likely introduce a little bit of error into our analysis, we believe that given the slow changing nature of population demographics a time difference of one or two years will likely not change the data significantly.
One continuous variable we could look into is per capita income - based on the home ownership, economy, housing situation, racial composition, and educational attainment, can we predict the average per capita income of the county? The per capita income is certainly an important variable to be able to model, since it is a good measure of the economic prosperity of the region. Looking into associated covariates might allow us to examine the factors determining economic prosperity, or look into how economic prosperity changes with demographic composition - very important questions in economic policymaking. One binary variable we could explore is election outcomes - given the same aforementioned demographic data, can we see what demographic qualities are associated with Republican or Democratic counties? This is clearly an exceptionally important question for campaigners on both sides of the political spectrum so they can either target the concerns of their supporters or target the voters they might be able to turn from the other side. Perhaps even more interestingly, though, would be to look at what problems those demographics were facing to see what drove them to vote one way or another. With an increasingly politically divided nation, it is of paramount importance to examine the reasons why people align with different political sides, and hopefully with that information to facilitate more cooperation on issues that both sides agree on without knowing it.

Directions for Exploration

There are a wide range of questions that we could explore with this dataset, given its breadth of demographic information. As mentioned above, one natural family of questions posed by the election database is to look at predictors for election outcomes - what makes a county lean towards one end or another of the political outcome, and why might this be? This is a critical question for political parties to answer and reveals insights into the problems and frustrations driving people towards certain candidates, the proper analysis of which could lead to informed, targeted policy change. Another family of questions that we can look into is the associations of race with various other demographic factors relating to living conditions; is there a systematic bias towards certain races having better or worse living conditions and are certain races associated with greater or worse educational attainment? This is interesting from a fairness perspective - if there is such systematic bias, it's worth investigating the source of that bias and trying to take measures to combat it. It is important in this case, however, to discern the actual causal relationships between race and the associated covariates - this modelling might provide a starting point for causal inference study, but should not be considered conclusive evidence of racial bias. We could do a similar analysis based on gender; more generally, it'd be interesting to see how many different demographic factors are associated quality of life

(perhaps measured by income, or population below poverty level), the answers to which might allow us to identify vulnerable or underserved subpopulations in need of greater support.

Data Cleaning

Our dataset is complete, so we do not have to deal missing values. The columns we have now give sufficient information about each county's demographics, and no column seems to be completely irrelevant. However, since the data set contains some state-level and country-level data rows, I filtered out a data set with only county-level data and stored as a data frame,"dataclean". We take out 80% of the county-level data as our training set.

Data Exploration

Then, we use R to plot pairwise scatterplots and explore patterns within our data set.

```
## Loading required package: ggplot2
```

```
##         Var1      Var2      Freq
## 55 PST040210 PST045214 0.9996582
## 57 POP010210 PST045214 0.9996578
## 75 VET605213 PST045214 0.9245172
## 77 HSG010214 PST045214 0.9934552
## 81 HSD410213 PST045214 0.9960061
```

```
## [1] 178
```

From the coefficient matrix, we found that a lot of pairs in our data set seems to be highly correlated. There are 178 pairs of data that have correlation larger than 0.8. We will only examine three, pair for the sake of time, which are VET605213 and PST045214, HSG010214 and PST045214, and see whether VET605213 and HSG010214 are highly correlated as well. It seems that a lot of our covariates are highly correlated; however, that is because some columns have arithmetic relationship with others (for example, total population and the breakdown of population). We will need to do some work to remove such collinear columns. Note that these three variables also fall into three categories (population, veterans, and housing). From the scatterplots, we see that these three variables all follow a linear pattern pairwise. This warns us about the colinearity exist in our data set, which further suggests that some columns in our data might be repetitive. However, since there are plenty of county-level data on other potential covariates, such as water usage, environment quality, and etc, we can fix this problem by joining our data set with other county-level data sets that we deem significant. Since our current data is at county-level, is it possible for us to even extend our data. For example, the percentage of immigrants who do not currently hold a green card/U.S. citizenship will be an intersting data to explore. There three variables we plotted are: PST045214:"Population, 2014 estimate"; VET605213:"Veterans, 2009-2013"; HSG010214:"Housing units, 2014"