**Summary:**

For this project our team was asked to look at voter registration, and participation across the state of North Carolina. To do this we used a series of tools that had been taught in class, including critical thinking, exploratory data analysis, logistic regression, outside research, and hierarchical modeling techniques. Our analysis was targeted at answers the following research questions.

* How did demographic subgroups vote in 2016? For example, how did the turnout for males compare to the turnout for females after controlling for other potential predictors?
* Did the overall probability or odds of voting differ by county in 2016?
* How did the turnout rates differ between females and males for the different party affiliations?

Our analysis ultimately led to the following model:

####

**Intro:**

The North Carolina State Board of Elections (NCSBE) is administration and the election process for the State. They are also required to keep extensive records to ensure electoral compliance, as part of their duties, they also keep information on likely voters and registered voters.

For this project we were given a dataset that contained information about the 2016 election, and we were asked to use two separate datasets, one which contained registered voters and the other which had likely voters.

From that we needed to clean, import, explore, and ultimately model these two datasets in R to answer the above research questions.

**Data:**

One of the more challenging aspects of this project was getting the two separate dataframes (registered voters and voters) merged into one dataframe that could be analyzed. Initial attempts to merge resulted in R crashing because of missing values, additionally my first successful join deleted all of the registered voters data because it had the same column name as the total voters.

In order to successfully merge the two dataframses the team went through a series of steps to ensure a successful merge. This was first done by removing the null values in both dataframes, which still left us with over a million observations between the two dataframes. To validate the success of this operation, we then took the total voters over the registered voters and confirmed it matched roughly matched the voter turnout percentage in 2016 (which it did).

We also

**Data Processing:**

We were then able to merge the two dataframes and begin working towards what we would use for our final analysis (named VotedDataMerged). From here we could check and ensure all variables were the right type (for example county names and race as factors instead of strings).

From this we could then aggregate into the VoterStatDF. The team then validated that we did not have any unusual values/percentages, things like the ensuring the number of voters exceeded (which only happened in 0.84% of cases, or that the voter turnout was close to the average across the state, etc).

Based upon this and other analysis we then dropped some of the unnecessary or unclear data because we had a large enough sample size to ensure accuracy of the results. In total between the two datasets we had over a million observations, less than 2% of which were dropped for various reasons.

**Exploratory Data Analysis:**

From here, the team began graphically exploring variables within the dataset this starting with the total voters, then exploring the voter participation rate within each (voters vs registered voters). After that we selected a random sample of 20 different counties (varying in size and demographics) across the state.

With that sample we then began graphing how much the various ethic and gender groups participated. First looking at gender, then each of the ethnic groups. From our initial plots we could see that voting in the US was very split along ethnic, gender, age and party lines.

Additionally, we also did multiple plots that overlaid various factors (age and race etc). From this we could see that each group sub group had influence on voting patterns. Because of this we started modeling with a Hierarchal Model.

**Model Selection and Assessment:**

**Final Model:**

**­**

**Conclusion:**

* How did demographic subgroups vote in 2016? For example, how did the turnout for males compare to the turnout for females after controlling for other potential predictors?
* Did the overall probability or odds of voting differ by county in 2016?
* How did the turnout rates differ between females and males for the different party affiliations?