R Notebook

Post-Stratification Data

Data Source and The American Community Survey (ACS)

The data used to construct the post-stratification dataset was a selected subset of the data collected through the American Community Surveys (@cite_ACS). The data was retrieved from IPUMS with a density of 0.2 due to the full size of the dataset being quite large. Data from 2018 was chosen since the goal of the model is to predict 2020 election results, as a result, it was ideal to choose the most recent dataset that was available in IPUMS.

The ACS run by the Census Bureau is performed by sending a survey to a sample of the population in the 50 states every month of every year. The population of this survey would be the American population, and the frame would be the 140 million addresses of American citizens available to the Census Bureau. Respondents are chosen by their address to ensure geographical coverage, and no respondent is to be selected more than once within 5 years. According to the Census Bureau, the survey is mailed to approximately 295,000 respondents a month. Respondents are allowed to respond to the survey in multiple ways such as by mail, telephone, internet, or in-person interviews. Furthermore, it can be noted that the addresses selected for the ACS are required by law to accurately respond to all questions. The multitude of methods that respondents can use to respond to the survey, as well as the legal component, allow for the Census Bureau to collect data on most of the 295,000 chosen addresses that make up the sample.

In addition, the guide released by the Census Bureau states that in person interviews are conducted for the addresses if there had been a non-response for 3 months. Moreover, the Census Bureau conducted telephone follow-ups for the questionnaires that were returned incomplete or needed clarification.

The main strength of the survey is that it is conducted very often and the survey collects a large amount and variety of data on the respondents. As a result, the ACS can provide an accurate and current reflection of American communities. Moreover, the survey is conducted by an official government organization, which has access to a large amount of resources to facilitate the survey distribution and data collection. Furthermore, respondents are allowed to respond to the survey in multiple ways such as by mail, telephone, internet, or in-person interviews.

Dataset

The variables included in the post stratification dataset were the variables that were determined to have a larger impact in the model. As a result, the variables included were age, gender, income, census regions, state, and race. The age variable represents the age range which corresponds to the age of respondent. Originally the age variable was just the age of the respondent when extracted from the IPUMS data. However, it was changed to be a range for the development of a model since it is easier for the model to determine the impact of a range of ages in comparison to exact age. Similarly, income groups were built from household income for the same reason.

In addition, state was chose to represent geographical location over census region in the final model due to there being 50 states in the US. The large number of states could help draw more specific conclusions.

(Figure @ref(fig:varpie)) shows the distribution of our predictor variables

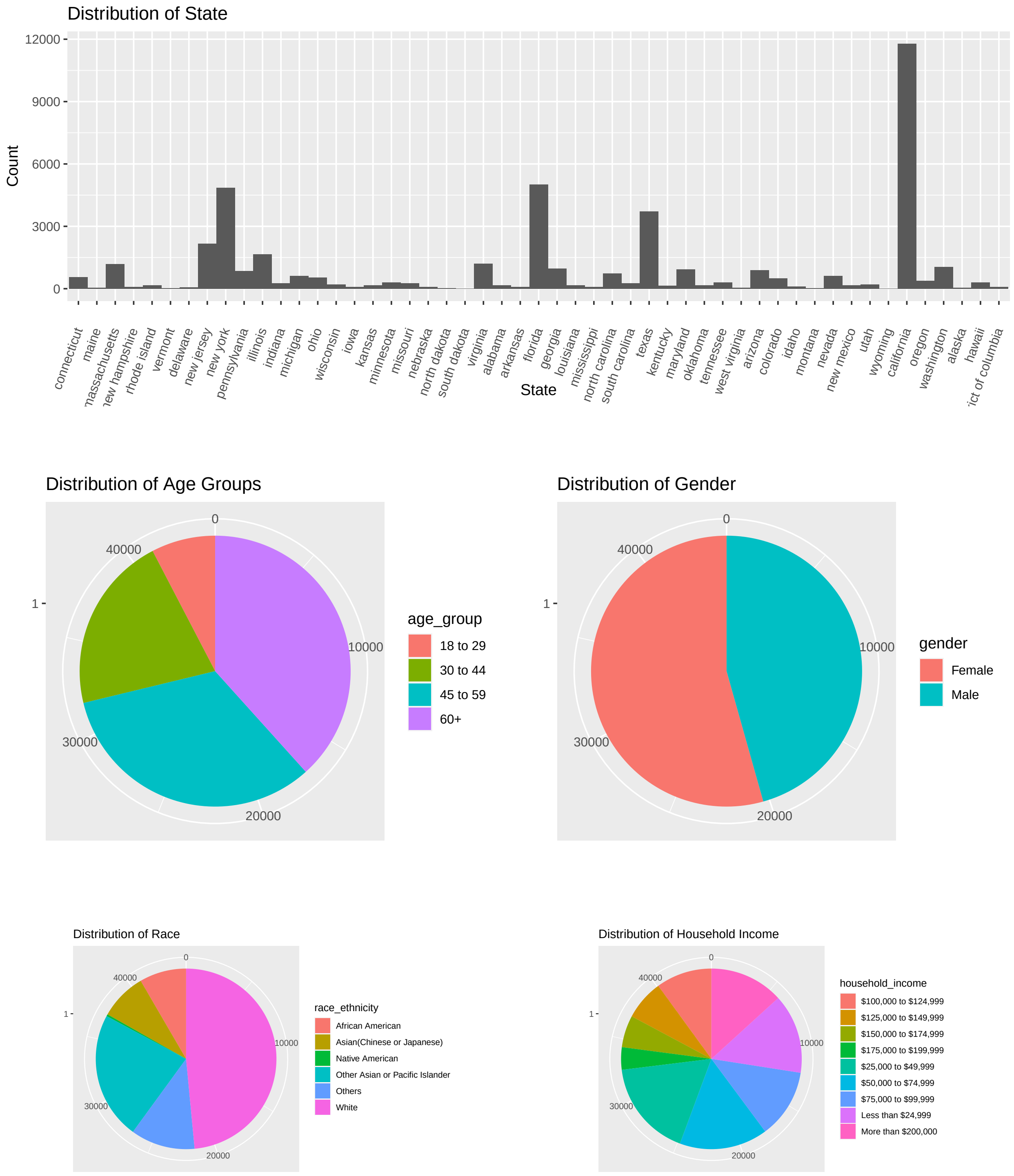

Figure 1: Distribution of Predictor Variables in Pie table

Figures 1.... Demonstrate the distribution of the variables used for the development of the model. We can notice that the distribution of the variables matches the distribution obtained by the survey mentioned in the previous section....... It can be noted that a lot of respondents are from California, Texas, and Florida, which is representative of the distribution of the american population among the 50 states.

Overview of MRP

Multilevel modelling with post stratification (MRP) is often used to help bridge the difference between non-representative survey samples and the actual demographics of the population. MRP is often done by first training a model from the survey data where the individual response is estimated. The estimation of the effect of individual predictors is typically improved through the use of multilevel modeling. Then the model is applied onto another dataset whose percentages of each type of sub population is more accurate to re-weight the estimates. (stat.columbia.edu) A benefit of this method is that it permits the use of broad surveys for analysis with regards to subgroups. However, it should be noted that there will be a larger uncertainty for the estimates. A caveat of MRP is that it requires the assumption that we know the composition of our population. Although this point is not as applicable to the point of research in this paper, it could cause an issue when trying to post-stratify a variable whose distribution is not clearly known. (https://statmodeling.stat.columbia.edu/2019/08/22/multilevel-structured-regression-and-post-stratification/)