ExploreCensus: *A Statistical Exploration of the 2015 United States Census Dataset*

Ishaan Prasad, Abhishek Malani, and Massimo Aufiero

Dataset:

Our dataset (available here) provides demographic information for each census tract in the US (i.e., the 50 states, Washington D.C., and Puerto Rico) from 2015. The dataset included the following features: *Census Tract*, *County*, *State*, *Total Population*, *Gender Split*, *Racial Split*, *Income*, *Poverty*, *Employment Type*, *Commuting Style*, *Commuting Time*, and *Employment Status*. This dataset is both a wide and long one, offering a lot of opportunities for analysis. The *census tract*, *county*, and *state* features are all methods of identification for the location of the data. The *total population*, *gender makeup*, and *racial makeup* offer insights into the size and types of people living in each census tract. One of the most used variables listed was the median income of people in the census tract. Finally, the last several columns list out various other demographic information collected from the tract (i.e., from information on poverty level to percentage breakdowns of transportation type for work commutes).

Analysis:

Upon cleaning up the dataset, we began our analysis by visualizing the median household incomes using a histogram. Overlaying the probability density graph for the normal distribution over the histogram, it became clear that median household incomes *do not* fall within this distribution. From here, we explored income with respect to the populations present within the census tracts. In particular, we used scatterplots to visualize the percentage of a specific race in a tract with respect to the median income of the tract overall. Ultimately, we found that the slope for the Asian population was the most positive of all the scatterplots, reflecting broader trends amongst one of the fastest growing, highest-educated minority groups in the nation.[1]

In addition, we developed several hypotheses about public transportation that we hoped to explore. Given that many people compare Chicago and Boston as cities with great public transportation, we wanted to see if one of the cities had a higher percentage of people using public transportation. To start, we conducted a permutation test on the states of Illinois and Massachusetts to see if there was any significant differences in public transportation usage. We found that the difference in public transportation usage was not significant at a state-wide level. However, when we specified our search to just Cook County and Suffolk County (i.e., the countie the contain the city of Chicago and Boston, respectively), we could see that with a two-sided test, the percent of Boston's population using public transportation is significantly larger than the percent of Chicago's population! Given our initial conclusions from our first test, we had not originally expected a significantly different percentage of people using public transport between the two regions.

Similarly, we developed hypotheses with respect to the different forms of transportation that populations use to commute to work. In particular, we hypothesized that more people would prefer to walk to work in Los Angeles than in Chicago because the weather is much nicer year-round in the former. To determine the validity and potential statistical significance of this hypothesis, we conducted a permutation test. Ultimately, we found that the percentage of people walking to work in Los Angeles was not statistically greater than Chicago; in fact, and to our surprise, we actually found that the number of people walking to work in Chicago was significantly greater than Los Angeles!

Following this analysis on transportation, we turned back to exploring incomes, using machine learning to see how well we could predict the incomes of census tracts. We segmented the data so that 70% of the data was used for training and the remaining 30% was used to test the out-of-sample accuracy. This separation was done through random sampling. We started by doing an Ordinary Least Squares (OLS) regression with the provided features; our model had a 71% R^2 value indicating that 71% of the variation in income is predicted by changes in one of the factors in our regression. From this regression we found that *Poverty*, *ChildPoverty*, *MeanCommute*, *Employed*, *Private Work*, and *Unemployment* were the only statistically significant correlates to income at the 5% level. Ultimately, with the inclusion of more features (e.g., health behaviors, child count, single parent households, etc.) in the dataset, we hypothesize that the OLS regression would have been even more accurate. In addition to our OLS regression, we made a Decision Tree over the same variables using 5-fold cross validation to prevent overfitting. Interestingly, the only factors that were used were by the tree were *poverty* and *professional*, with poverty rate being the most predictive split at 9.35%. We were skeptical of the ability of just these two variables to predict incomes, but found it to be quite an accurate model for predicting incomes in the out-of-sample data. From there, we created a 500-tree random forest model to model incomes across census tracts. Testing all three models against the out-of-sample data, the random forest had the lowest mean error followed by the OLS regression. Finally, we wrote a new file called predictions.csv to display the income followed by the predictions from the three models and their respective errors.

Upon completing our machine learning tasks, we used a number of other columns (e.g., race, profession, poverty) to predict income. After seeing via bar plot that average income for tracts with predominantly Asian populations was the highest, we tested to see if it was significantly greater than the others, and found that in each case it was. We also used the percentage of people working in a "professional" industry (management, business, science, and arts) could be used as a predictor of being "wealthy" (median income over $50,000) in Massachusetts using a logistic regression curve. The curve demonstrated a pretty strong positive correlation, and the model gave fairly accurate predictions of the data. Finally, we ended our analysis by exploring the skewness and kurtosis of *drive* (i.e., the percentage of people commuting alone in a car, van, or truck).

---

[1] The Rise of Asian Americans. (2018, September 27). Retrieved from https://www.pewsocialtrends.org/2012/06/19/the-rise-of-asian-americans/