Name: Alexa Canaan
Thesis Proposal: Predicting Public Transportation Inefficiencies Using LASSO

Research Question:
Where is public transportation being used inefficiently? For my research, I would like to identify areas where public transportation is not available, but would be utilized if it were made available and areas where public transportation is available, but it is not being used. These are areas that I would define as "inefficient" for public transportation. Efficient uses of public transportation are areas where public transportation is available and it is used, as well as areas where public transportation is not available and it would not be used if it was available.

Method:
I will use machine learning to predict inefficient uses of public transportation. Since I will be using the American Housing Survey, which has 4114 different explanatory variables to work with, I will use a LASSO method to select explanatory variables for my model.

Data Set:
I will be using the American Housing Survey data, a longitudinal housing unit survey, to develop my model. The American Housing Survey is sponsored by the Department of Housing and Urban Development and conducted by the U.S. Census Bureau covering all 50 states and the District of Columbia. The study is conducted biennially in odd-number years, spanning from 1973 to 2017 between May and September. The same housing units are surveyed every other year until new samples are drawn, allowing for analysis of households over time. The goal of the data set is to provide timely information on the quality and cost of housing in the United States and American metropolitan areas and the participating housing units were chosen to represent all housing units in the United States. It is used by policymakers to make decisions about housing for all demographics in America.

Each observation in the data set is a "housing unit" or any house, townhouse, apartment, mobile home or trailer, single room, group of rooms, or other location that is occupied as separate living quarters, or if vacant, is intended for occupancy as separate living quarters. The survey is conducted using computer-assisted personal interviewing using laptops. National data as well as metropolitan data are collected. As I continue to refine my research, I will decide whether or not I will focus on national data, metropolitan data, or analyze both. It is important to note though that while national data is always collected biennially with adjusted samples, no more than 30 metropolitan areas are sampled in one survey year due to budgetary constraints.

This particular data set is applicable to my research question because it has been used by policy makers to plan community development such as infrastructure. I will specifically focus on data from 2013 due to its large range of questions covering public transportation. There were approximately 84,400 sampled housing units with a supplemental sample of 15,533 housing units in the Chicago, Detroit, New York City, Northern New Jersey, and Philadelphia metropolitan areas. Out of the 84,400 sampled housing units, 2,715 were ineligible because the unit no longer existed or because it did not meet the definition of a housing unit. Further, 10,000 units had no response after repeat visits or refused to be interviewed. This led to an overall response rate of 86%. The AHS-National weight is 2,148.

The data does come with its own set of limitations in terms of incomplete data, wrong answers, and sampling variability. For example, incomplete data are adjusted by assuming that the respondents are similar to those not answering and the size of these errors is estimated. The data is also not adjusted for wrong answers and does not estimate the size of these errors. The same applies to sampling variability.

Data Sources:
https://www2.census.gov/programs-surveys/ahs/2013/2013%20AHS%20National%20Sample%20Design%20and%20Weightin
https://www2.census.gov/programs-surveys/ahs/2013/2013%20AHS%20National%20Errors.pdf?
https://www.census.gov/programs-surveys/ahs/tech-documentation.html

Summary Statistics:
To understand the data that we're working with, it is important to look at the summary statistics to get an idea of the general spread of data. For now, I will only examine the national data, although metropolitan data will be useful in providing more observations for my model.

```
> #read AHS data
> #national data
> dat.n<-read.csv("~/Desktop/ahs2013n.csv")
> #metropolitan data
> #dat.m<-read.csv("~/Desktop/ahs2013m.csv")
```

Since the data is supposed to be representative of the United States, we can look at the descriptive statistics for the demographics of the data set to confirm that they are reasonable. To get a basic idea of the data, I will be looking at the spread of sex, age, marital status, race, educational level, and household income.

It is important to note that this data has been cleaned very minimally. However, as I proceed in my data analysis process I hope to fully convert all variables I plan to use to their proper form, whether that be a factor, numeric, or character values.

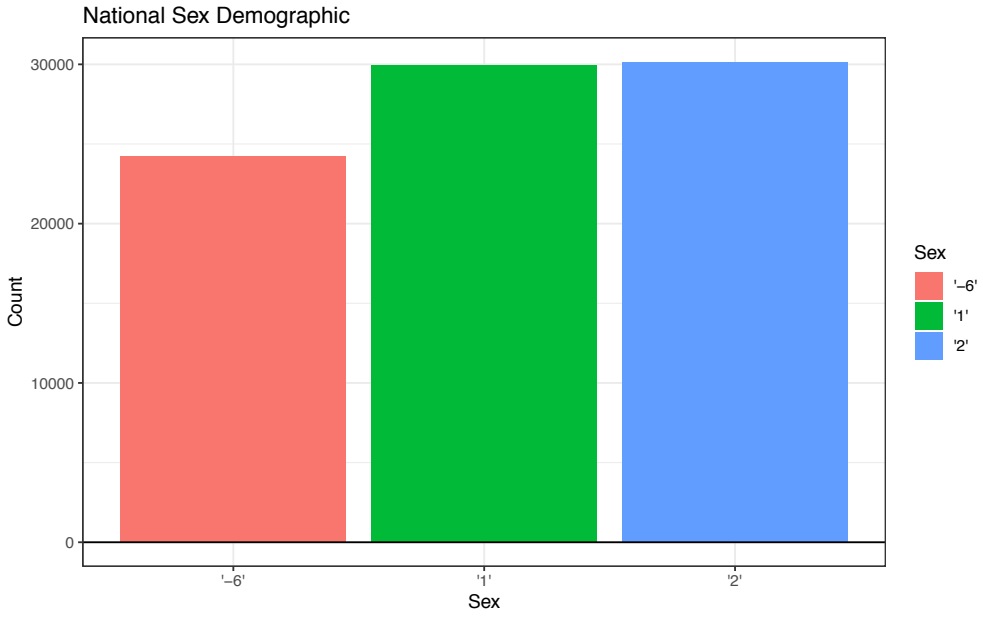

Figure 1: As we can see from the graph on sex, there is an almost even split between male(1) and female(2) household respondents. Missing values do make up a large fraction of the sex observations though and those will be values that will have to be cleaned for accurate analysis later on.

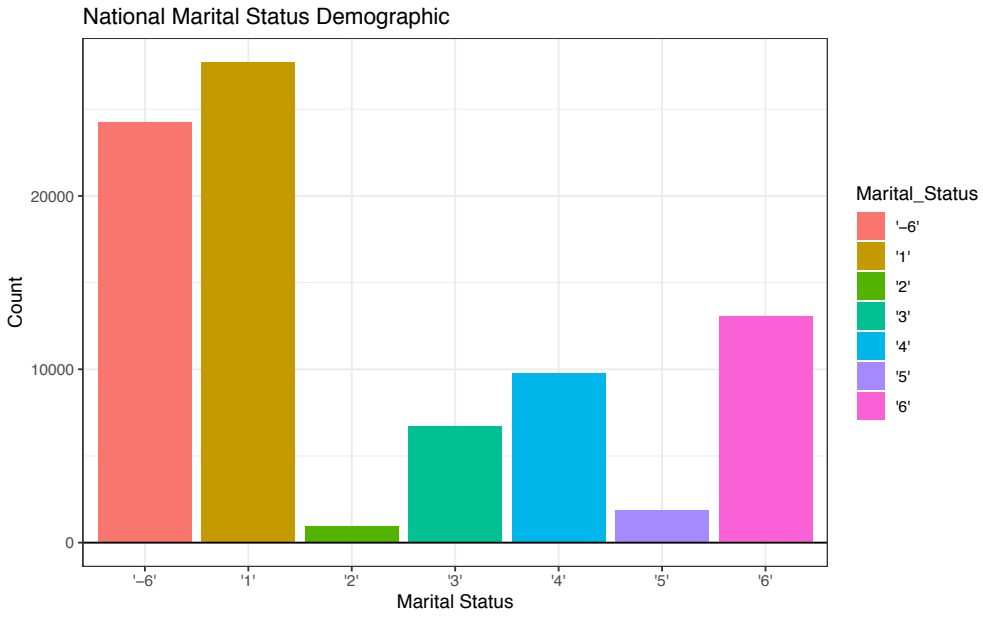

Figure 3: The graph on marital status has 7 different categories. Going from left to right, -6 corresponds to NA values, 1 corresponds to Married, spouse present, 2 represents Married, spouse absent, 3 represents widows, 4 represents divorced, 5 represents separated, and 6 represents never married. The majority of observations represent people who are married with their spouse present followed by those who were never married.

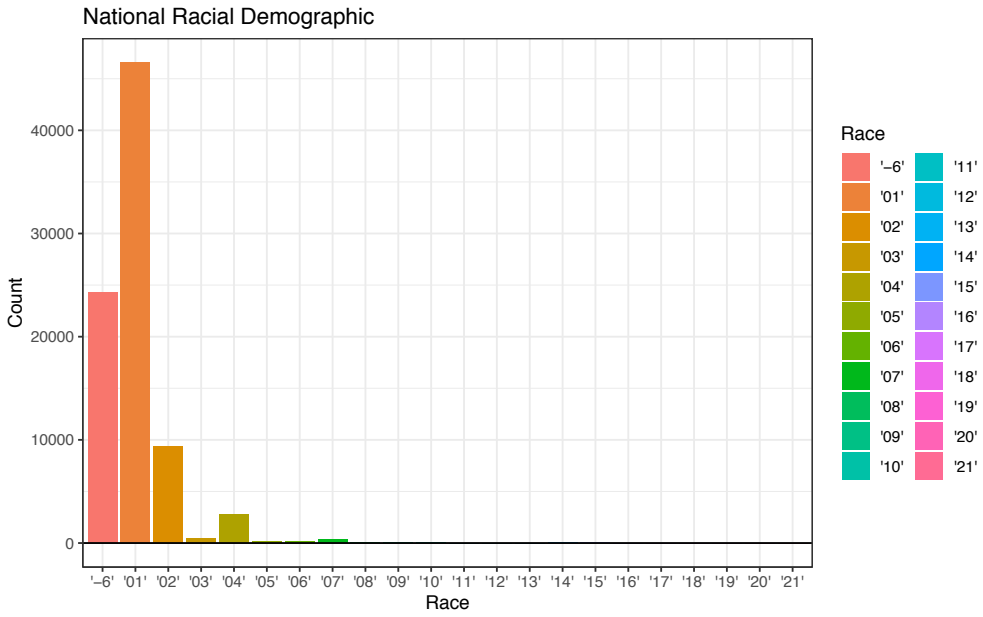

Figure 4: While the graph on race does not have many observations shown for the over 20 options available, we can gather from the graph that white people make up the overwhelming majority of the sample, followed by black people and asian people. Other races and combinations of races are included, but are far from the majority.

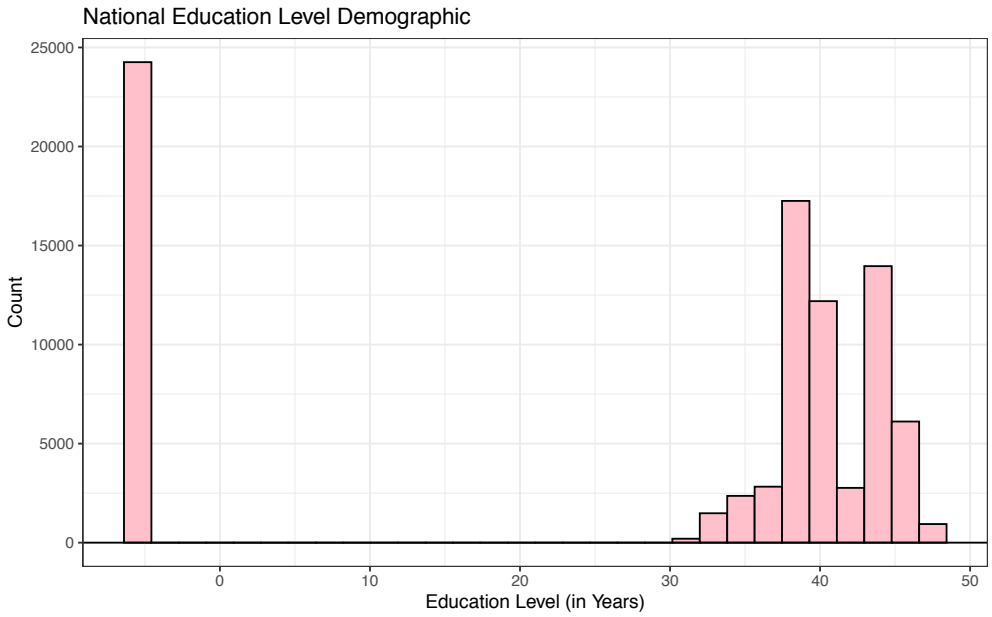

Figure 5: Education level data will need to be adjusted, but essentially 31 represents less than a 1st grade education, and then the amount of years one spent in school increases incrementally up to 47, which represents a doctoral degree. 31 to 34 represent elementary and middle school years, 35 to 39 represent high school years, and 40 to 47 ranges from some college to a doctorate. The majority of observations are concentrated around a high school diploma and a bachelor's degree.

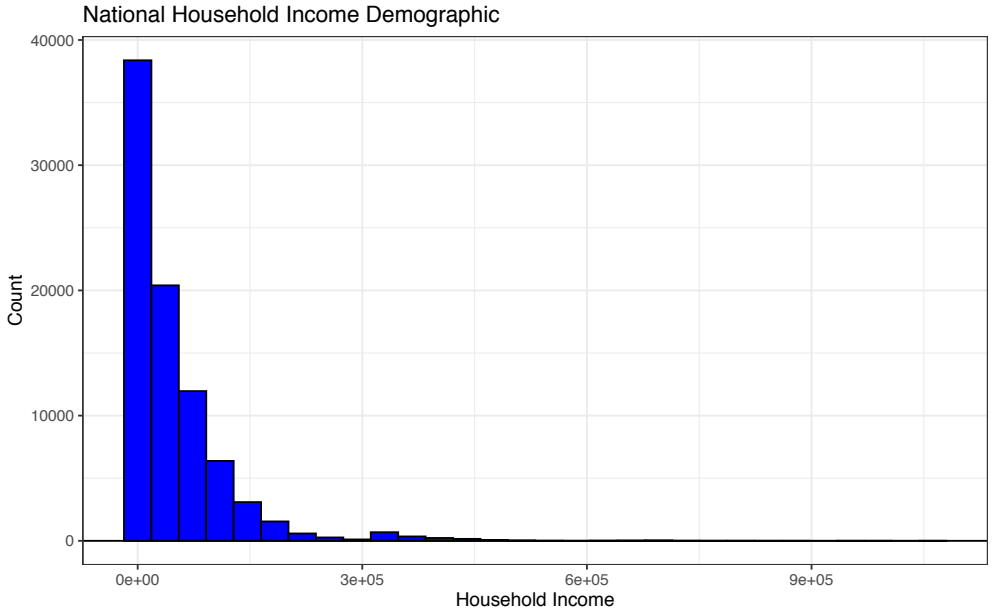

Figure 6: Household income is right-skewed with most observations being concentrated between 0 and 150,000. This is in line with what is generally known to be true about average household income.

Looking at all of the graphs, we see that there are many missing values for all variable types. This will be important to consider as I build out the model and decide which machine learning technique to use to predict public transportation usage. It may be hard to test the out-of-sample prediction accuracy since the

more observations there are, the more accurate the model will be.

Plots of Data:
Now we can look at variables related to our research question. While there are 60 variables in the data set covering public transportation, the following variables provide a holistic view of public transportation demand.

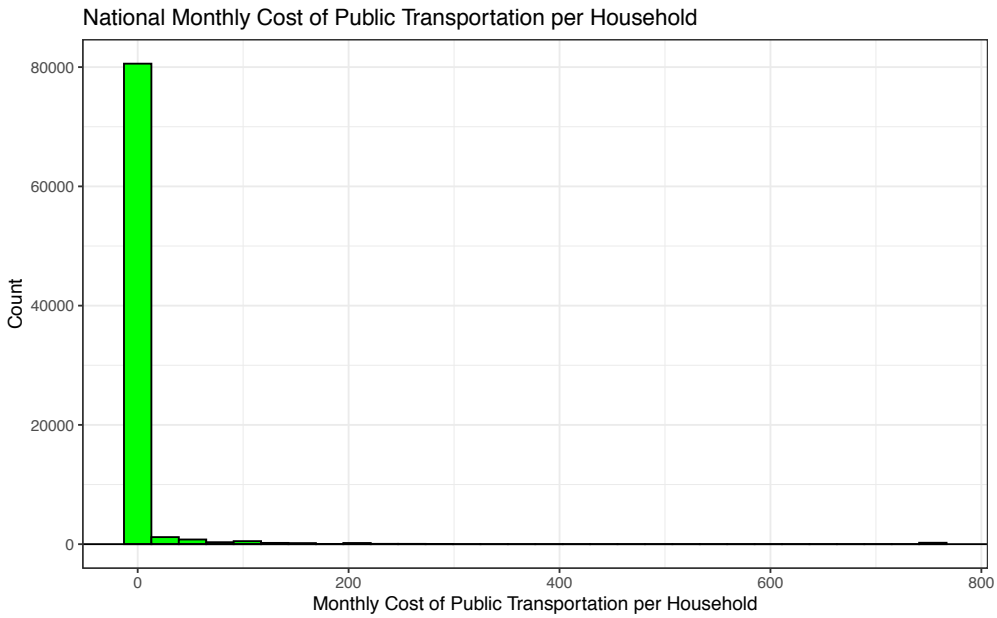

Figure 7: The vast majority of observations for the monthly cost of public transportation ranges from 0 to 200 dollars. A summary statistics table will be able to give more insight into this. This does highlight though that public transportation is an affordable option if it is available.

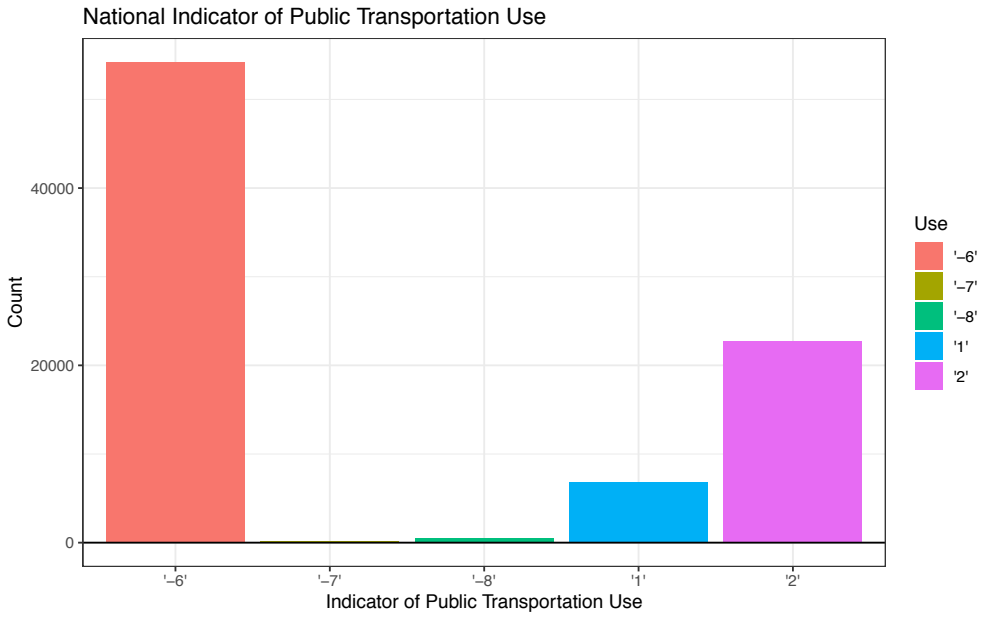

Figure 8: Out of those who responded to the question indicating use of public transportation, more respondents did not use public transportation (2) than those that did (1). Values -6 through -8 represent represent missing answers, NAs, don't knows, or those who refused to answer the question. It is important to distinguish between the observations where people do and don't use public transportation because we are interested in determining why these decisions are made and if it is due to public transportation inefficiencies.

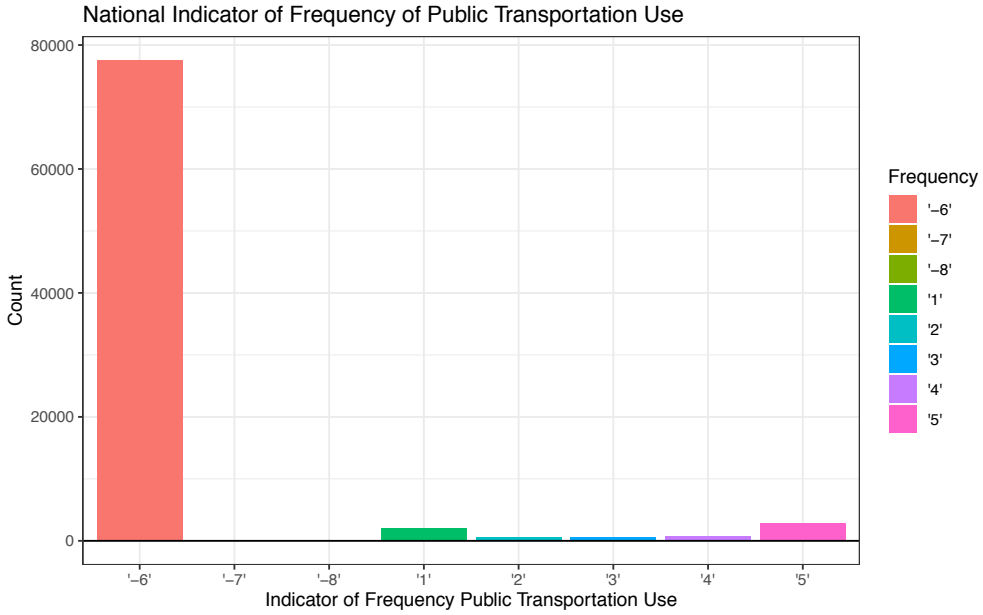

Figure 9: The last graph illustrates different frequencies of public transportation use. -6 through -8 represent values that indicate NA, don't know, or refused to answer the question. 1 represents those that always use public transportation, 2 is most of the time, 3 is sometimes, 4 is once in a while, and 5 is never. While this is an important question to ask, the possible answers are relative. This will require further research.

Description of Key Literature:

While I am still in the elementary stages of my research, there are important articles that have given me significant direction thus far. "Predicting travel mode of individuals by machine learning" by Hichem Omrani gave me some baseline knowledge of how to proceed with a machine learning analysis in an economic context. "Machine Learning Forecasts of Public Transport Demand" by Sebastian M. Palacio aided me in understanding what variables I will include in my regression and machine learning analysis and what patterns to consider in public transportation usage, such as weather and working days. "Predicting the Use of Public Transportation: A Case Study from Putrajaya, Malaysia" by Borhna et al also provided insight into some seemingly unmeasurable factors to consider, like service quality and attitude, that may not be readily available from the American Housing Survey alone. I hope that further research will help me determine what other machine learning approaches I should take in my analyses and refine my research question if necessary.