EXPLORING GAS STATION EXPANSION WITH ANALYTICS

V. M
F. S
C. P
Benjamin McGriff

STAT 4220
Professor Ferrara
March 20, 2019

Table of Contents

Executive Summary

Empire State Gas & Fuel is attempting to determine which part of New York City would serve as the best location for a new gas station with the goal of capturing a large share of the market. To determine the best location in New York City, we want to maximize the percentage of commuters who drive alone or carpool to work, average commute time, and household income. Accordingly, we used multiple linear regression modelling to predict driver/carpooler percentage and commute time as well as binary logistic regression modelling to predict above or below average household income. We found that, based on our analyses of the aforementioned indicators of a strong location, Queens, Brooklyn, and Staten Island are likely the most financially rewarding options. We also determined that areas with a high percentage of people employed in service jobs and low percentages of people self-employed or employed in private industry are likely regions with a greater demand for a gas station. Going forward, we recommend using a) clustering to identify our target market in New York City and b) quantile regression modelling to focus on extremely high commute times and driver percentages. We also believe collecting more data such as education level, real estate costs, and the current degree of gas station market saturation would help improve our analysis and allow us to more accurately pinpoint a successful location. If Empire State Gas & Fuel consider the following analysis and conclusions, we believe they will have the best statistical chance of achieving their long-term strategic and financial goals.

Introduction

Company Overview

The following analysis will be conducted to help our company, Empire State Gas & Fuel, an American chain of gas stations located in the state of New York, determine a strategy for future expansion. After opening our first filling station near our headquarters in Yonkers, New York 20 years ago, we currently operate dozens of stations throughout the entire state. The majority of our gas stations are located in Westchester County as well as in several cities in Upstate New York like Buffalo and Rochester. However, we have yet to open any filling stations in New York City — the most populous metropolitan area in the state and country. As the company hopes to become one of the largest gas station chains in the state, we plan to explore new expansion opportunities to accelerate our growth.

Business Objectives

For the near future, our company’s primary goals are to establish a new source of revenue and gain a larger share of the New York gas station market. Accordingly, we have begun investigating the idea of opening a station in New York City whose high population density makes it the ideal target market. We specifically want to operate in high-traffic areas which offer our company the opportunity to serve a large amount of drivers without having to open many stations. As New York City is nearly three times more densely populated than Yonkers, the City would allow us to immediately capitalize on a readily available market and customer base. Compared to our other locations, a gas station in New York City could have hundreds of thousands of more customers, generating millions in additional revenue.

While expanding into New York City may be financially rewarding, there are a number of unique challenges that may derail successful expansion. For one, the monetary costs of opening a filling station in the City are enormous. The exorbitant cost of real estate in New York City means that we would have to spend a significant amount of money up-front just to secure the property. Furthermore, we would have to perpetually pay for labor, maintenance and gasoline on a weekly, or even daily, basis. The major problem our company faces is the financial and logistical risk of the project. If the marginal revenue earned from the new gas station isn't able to cover the initial set-up, fixed and variable costs, the expansion will actually hurt the company's bottom line. Additionally, opening an unsuccessful filling station wastes valuable resources like money, time and effort that could have been used for a more profitable opportunity.

Selecting the perfect location for our first gas station in New York City is critical as choosing a bad location with limited visibility and traffic will damage the company's financial health. Accordingly, we posed the following research question: "In which area of New York City should we establish a new gas station so that we can reach the most customers?" To successfully answer this question, we have identified high household income, commute length, and concentration of drivers as indicators of areas with a high demand for a gas station. Individuals who are high-income earners are more likely to both own and use a vehicle, while a population with a large percentage of driving commuters and longer commute time have a greater need for gasoline. The analysis conducted for this report will statistically determine which part of New York City would be the best location for our next gas station based on the aforementioned variables; the results of the analysis will help the company deploy its resources successfully, considering the company's strategic and financial interests.

Methods

Data Understanding

The data being used for this analysis is information collected through the 2015 American Community Survey (ACS), an ongoing survey conducted by the U.S. Census Bureau (Kaggle, 2017). Specifically, the dataset consists of key socioeconomic, demographic, and population data for each census tract in New York City based on 60 months of data collected from 2011 to 2015. For the survey, the Census Bureau used monthly independent samples of housing units extracted from the Bureau's own inventory of known living quarters in the country. Once selected, participants are given the option of completing the survey either through the Internet, a mail-in questionnaire, telephone, or an in-person visit.

Each observation in the data set represents one of the 2167 census tracts in New York City. Census tracts are small subdivisions used by the Census Bureau for data collection; in New York City, tracts typically have a population of about 3000-4000 and a land area of about 90 acres. In total, the data set measures 35 different qualitative and quantitative variables for each census tract. First, the borough (e.g. Bronx, Queens, etc.) and county (e.g. Kings, Richmond, etc.) of each tract are included. In addition, a number of relevant continuous variables are also given including total population, percentage of people commuting alone, percentage of people carpooling, percentage of people commuting on public transportation, percentage of people walking to work, mean commute time, and median household income (USD). After conducting basic exploratory analyses, there seem to be no clear outliers or influential points. While extreme cases exist regarding many of the variables, there is no obvious evidence that these observations deviate from the overall pattern. Moreover, 72 observations in the data set were found to have missing values for one or more of the important variables; these observations had to be addressed before any analysis could be conducted.

Data Preparation

While the data set chosen is relatively comprehensive and complete, a number of modifications were made to the data to ensure it was appropriate for our analysis. Data preprocessing steps were taken as the 72 incomplete observations found were removed. The missing values are likely due to low participation in lowly populated tracts as there are no signs of clear bias and every borough is well-represented. As these tracts represent an extremely small percentage of the total sample, we do not expect removing the observations to be a significant issue. This was the only data cleaning that was necessary.

Additionally, two important measures were created and added to the data set to facilitate the most beneficial analyses for our company. First, we calculated a new categorical variable — household income level — which describes whether or not a tract’s average household income is above the New York City average of $57,782. We decided to use this binary measure instead of the continuous income variable because of its practicality for our purposes. The company is not looking to specifically target one arbitrary income value or range, but simply prefers areas with generally higher household incomes. Secondly, we calculated another variable that combines the percentage who commute by driving alone and the percentage who commute via carpool. This new variable therefore represents the percentage of the population of a census tract that routinely uses some private vehicle to travel to work. We created this new measure as relying on only one of the variables used in the aforementioned calculation would not cover some sizable amount of people who contribute to vehicle usage in New York City (either solo drivers or carpoolers).

When assessing the data quality of the ACS, there are a number of issues and limitations to consider. As the data was pooled from samples, some sampling error will naturally be present. This problem is exacerbated by the fact that the Census Bureau does not increase the sample size to proportionally represent the growing population of the United States. In addition, since the research methodology was a survey, response error will also be present to an extent. Many individuals the Census Bureau reaches out to neglect to respond, producing imperfect sample data. This nonresponse bias is not consistent as some counties have less than 10% participation, while others have close to 80%. Finally, the ACS does not have population controls for census tract data. Population controls supplement the latest Decennial Census population count with birth, death, and net migration data to produce accurate and up-to-date population estimates. The absence of population controls increases the size of the standard errors by 15 to 20 percent, which means the sample used for the ACS is less representative of the overall population.

Modelling

Before any analysis was conducted, several data visualizations were created in order to observe any relationships that existed between our key variables. First, we compared the percentage of commuters who drive alone or carpool in a census tract with the tract's average household income (Appendix C). In general, there appears to be a positive relationship, suggesting that as household income increases, so too does the percentage of people who drive or carpool to work. However, the faceted graph indicates that the pattern differs depending on the borough in question. While Staten Island, Queens, and the Bronx display a clear positive correlation between the factors, a large portion of Brooklyn's observations do not follow the pattern. In Manhattan, on the other hand, there appears to be no correlation at all between driver percentage and household income; as household income increases there, the percentage of commuters who drive alone or carpool remains generally constant. In a second graph, the number of census tracts with "Below Average" or "Above Average" household income in each borough is displayed (Appendix D). Here, once again, a stark contrast exists as the Bronx and Brooklyn have more "Below Average" tracts than "Above Average", while the vice versa is true for the other 3 boroughs. Specifically, Brooklyn has the most "Below Average" observations and Queens has the most "Above Average" ones. A final visualization displays four map charts of New York City, measuring different continuous variables across the five geographically-marked boroughs (Appendix E). The charts clearly show Manhattan is the most densely populated borough. Additionally, Staten Island has the highest percentage of commuters who either drive alone or carpool and Brooklyn has the highest mean commute time. Manhattan has the lowest percentage of driving/carpooling commuters and average commute time, but has the highest average household income.

Two important factors that will affect our decision about our new location are the variables of mean commute time (MeanCommute) and percentage of commuters who drive alone or carpool (DriveOrCarpool). A location that has high values for these two variables will be a better overall investment for the company. Therefore, we made two multiple linear regression models that predict the values of these two variables based upon the other attributes in the dataset.

First, we checked the assumptions for multiple linear regression. One assumption is that the independent variables are not highly correlated with each other. To test this, we created a correlation matrix to determine the correlation between each of the independent variables. Based on this matrix, we found the highly correlated variables to be total population, men, women, income per capita, poverty, and percentage of people in professional jobs, percentage of people employed, and percentage of people in public work. Another assumption is that each independent variable has a linear relationship with the dependent variable. We created scatterplots plotting each independent variable against the dependent variable MeanCommute and against the dependent variable DriveOrCarpool. We found that all of the relationships appeared to be linear, although several had weak correlation, but decided to keep all of the variables and later run feature selection. The other two assumptions are that the standardized residuals follow a normal distribution and homoscedasticity, which we checked after we developed our model.

We separated our data into training and testing sets with a 70/30 split and built our model based on our training data. We then conducted stepwise selection on our linear models for MeanCommute and DriveOrCarpool to select the most relevant predictors from our dataset. From this, we removed the variables Borough, Hispanic, Native, ChildPoverty, Service, and FamilyWork from our DriveOrCarpool model. We also removed the variables Borough, Native, Income, IncomeErr, Office, WorkAtHome, and FamilyWork from our MeanCommute based on the results of the stepwise regression. These removals were based on the analysis of the deviance table generated by the stepAIC function. Finally, we developed our final linear models for predicting MeanCommuteTime and DriveOrCarpool, a summary of which can be found in Appendix K and L respectively. We tested our models using several different methods. We looked at the adjusted R squared value to analyze the proportion of the variation in DriveOrCarpool and MeanCommuteTime explained by our independent variables. Since we separated our dataset into a testing and training set, we also tested our model on our test set by comparing the actual and predicted values and calculating both the correlation accuracy and the mean absolute percentage error of our predictions.

Next, we modeled a binary logistic regression to predict whether the household incomes of relevant tract demographics will be above or below the average. Doing so provides us with the assurance (or uncertainty) of our decisions given projected changes in the values of our independent variables. To ensure that we obtain both a good fit of our model and statistical predictiveness, we did a preprocess of cleaning and formatting our data so that it only contains data of variables that we believed were most relevant to household income factors. Since we are primarily concerned with those who earn an income that is above the average, we created our fitting process to reflect just that by making the dependent variable a boolean factor that identifies incomes as being either above the household average or not.

To avoid overfitting our model, we partitioned our data so that 70% of it went to a training subset while the remainder 30% went to a testing subset as per standard practice. The demographics we observed, serving as our independent variables, are borough (Borough), mean commute time (MeanCommute), public work (PublicWork), private work (Private Work), service, office, construction, production, professional, drive/carpool preferences (DriveOrCarpool), and the racial groups of white, black, Hispanic, and Asian.

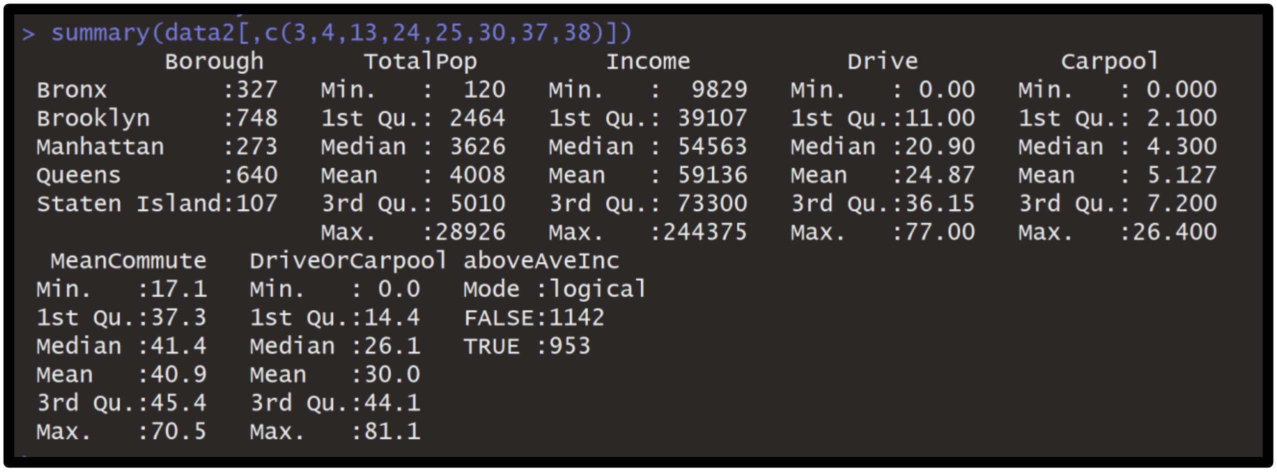
To assess the predictability of our model on any given new set of data, we created a model that evaluated the percentage accuracy of our model given a chosen decision boundary value of 0.5. This arrangement makes it so that if, in mathematical terms, $P(y=1|X) > 0.5$ then $y = 1$; otherwise y=0. An ANOVA table (Appendix G) was also devised to help give us a better picture of how impactful our selection of independent variables were to the trends of our model.

Finally, we plotted a ROC curve (Appendix H) as it is a typical performance measurement for a binary classifier such as ours. From this we then calculated the area under the curve to obtain a percentage value that denotes how well our model predicts above average incomes between the boundaries of 1 (being ideal) and 0.5 (our chosen boundary).

Results and Evaluation

Data Summary

A summary statistics table consisting of descriptive data for each of our most relevant variables is given below:



```
> summary(data2[,c(3,4,13,24,25,30,37,38)])
          Borough       TotalPop         Income            Drive          Carpool     
 Bronx        :327   Min.   :  120   Min.   :  9829   Min.   : 0.00   Min.   : 0.000  
 Brooklyn     :748   1st Qu.: 2464   1st Qu.: 39107   1st Qu.:11.00   1st Qu.: 2.100  
 Manhattan    :273   Median : 3626   Median : 54563   Median :20.90   Median : 4.300  
 Queens       :640   Mean   : 4008   Mean   : 59136   Mean   :24.87   Mean   : 5.127  
 Staten Island:107   3rd Qu.: 5010   3rd Qu.: 73300   3rd Qu.:36.15   3rd Qu.: 7.200  
                     Max.   :28926   Max.   :244375   Max.   :77.00   Max.   :26.400  
  MeanCommute   DriveOrCarpool aboveAveInc    
 Min.   :17.1   Min.   : 0.0   Mode :logical  
 1st Qu.:37.3   1st Qu.:14.4   FALSE:1142     
 Median :41.4   Median :26.1   TRUE :953      
 Mean   :40.9   Mean   :30.0                  
 3rd Qu.:45.4   3rd Qu.:44.1                  
 Max.   :70.5   Max.   :81.1                  
```

Multiple Linear Regression

We obtained two linear regression equations from our model, the coefficients of which can be viewed in Appendix K and L for Mean Commute Time and DriveOrCarpool respectively. We tested the validity of these models by looking at the adjusted R squared values, the correlation accuracy of our models' predictions, the min-max accuracy of our models' predictions, and the mean absolute percentage error of our models' predictions. The adjusted R squared for our Mean Commute Time model was 0.5643, the correlation accuracy was 0.7636, the min-max accuracy was 0.9173, and the mean absolute percentage error was 0.0917. For our DriveOrCarpool model, the adjusted R squared was 0.7112, the correlation accuracy was 0.8541, the min-max accuracy was 0.7094, and the mean absolute percentage error was 0.4921.

Because the adjusted R squared value for our mean commute time model is lower than expected, that indicates that there is a significant amount of variation in the mean commute time that is not explained by the dependent variables. This suggests that there is significant variability in our dataset. However, our mean commute time model has a high correlation accuracy, indicating that our model does relatively well at capturing the relationship between the independent and dependent variables (i.e. as the actual values go up, so do our predicted values). The min-max accuracy is high and the mean absolute percentage error is relatively low, indicating that our model predicts close to the actual values for mean commute time.

The equation of our mean commute time model is as follows:
MeanCommute = B0 + B1factor (County) + B2Hispanic + B3White + B4Black + B5Asian + B6Citizen + B7IncomePerCapErr + B8ChildPoverty + B9Service + B10Construction + B11Production + B12PrivateWork + B13SelfEmployed + B14Unemployment + B15factor (aboveAveInc).

The coefficients can be found in Appendix K.

For the model that predicts the percentage of commuters who drive or carpool, the adjusted R squared value is relatively high, indicating that most of the variation in the percentage of drivers and carpoolers is explained by our dependent variables. The correlation accuracy is also relatively high, indicating that our model is relatively accurate at predicting the relationship between the independent and dependent variables. However, the mean absolute percentage error is high and the min-max-accuracy is lower than ideal, indicating that while our model predicts the relationship between the variables well, it falls short when predicting values that are close to the actual values of DriveOrCarpool.

The equation of our percentage of drivers/carpoolers model is as follows:

DriveOrCarpool = B0 + B1factor (County) + B2White + B3Black + B4Asian + B5Citizen + B6Income + B7IncomeErr + B8IncomePerCapErr + B9ChildPoverty + B10Office + B11Construction + B12Production + B13WorkAtHome + B14PrivateWork + B15SelfEmployed + B16factor (aboveAveInc).

The coefficients can be found in Appendix L.

Based on these results, the mean commute model appears to be a reliable and accurate model. It indicates that the most significant variables that are positive drivers for mean commute time include Kings County and the percentage of the population that is in the service industry. On the other hand, New York County, the percentage of the population that is self-employed, and the percentage of the population that is in the private sector negatively affect mean commute time.

The drive and carpool model is more complicated to analyze, because it does have a high error rate when it comes to predicting the actual values, but since it has a high correlation accuracy, it accurately captures the relationship between the independent and dependent variables, and therefore can still be considered when analyzing which factors negatively and positively affect mean commute time. The percentage of the population that are white, the percentage that are asian, and the percentage that work in an office all positively affect the percentage of drivers/carpoolers. In addition, Richmond County is the only county that positively affects driver percentage.

Logistic Regression

From our model we observed that the sociodemographic independent variables of race (particularly asian), car use (DriveOrCarpool), borough (particularly that of Queens), and mean time spent driving (MeanCommute) to be statistically significant to our study. This means that the relationship between these variables and our dependent variable, above average income, is likely caused by something other than random chance. All of the other variables, while insignificant, still provide us with information that gives us a holistic picture of the general population we are focusing; we know that these variables in particular are likely arbitrary and thus allow us to better define our research scope. All of the conclusions made here can be viewed under Appendix F.

The overall accuracy of our model was calculated to be 85%, which is a very good accuracy percentage considering the manual data split we did to prevent the model from overfitting. The program function we used utilizes the general formula for logistic regression (identified as logit):

$$\text{logit}(p) = \log(p/(1-p) = B_0 + B_1*X_1 + \ldots + B_f*X*f$$

Where ‘p’ is the observed probability of an event, ‘B_0’ is the intercept, ‘B_0,...,B_f’ are the coefficients, represented in Appendix F, and ‘X_1,...,X_f’ are the variables relevant to the study. The list of all of the variables we considered in our model can be viewed again in Appendix F. Contextually, this means that if we were to look at a new set of data then our model would be able to predict the resulting above average incomes with 85% accuracy. The programmed calculation can be viewed under Appendix H.

We then calculated of classification accuracy of our model to determine whether or not the modeling process of our logistic model can be relied upon. In non-technical terms, this accuracy value represents how well our model distributes observations into categories/variables of the same type. This was calculated by finding the area under our Receiver Operating Characteristic (ROC) curve, which is a diagnostic curve standard to the field of statistical analysis. This curve allows us to ultimately test the accuracy of our results through a comparative study of both the aforementioned positive (or sensitive) and negative (or specific) rates of our results. From taking the area of this curve, represented under Appendix J, we found our model to be 93% accurate at classifying, giving us confidence to uphold our model’s predictive ability.

<u>Discussion</u>

Interpretations

From our multiple linear regression models, it appears that a good location would be somewhere in either Richmond County or Kings County and is near many service industry buildings or office buildings because the percentage of service workers and office workers positively affect mean commute time and driver percentage respectively. However, one important consideration for our model is that mean commute time doesn't necessarily mean high driving time. Commute also includes other forms of transportation such as the subway, walking, etc. Therefore, the mean commute time model should only be evaluated along with the driver percentage model to provide a more complete picture of the factors that will allow us to pick the best location to maximize profit. When taking both of these models into account as well as the limitation of the mean commute time model, it would appear that a location that is near many office buildings, possibly in either Richmond County, Kings County, or a location in Queens based upon our logistic model, would be an ideal location. In the future, if data were gathered on mean commute time for commuters who drive or carpool, that would be valuable to be able to further and more accurately research the locations that have maximum car usage. Another factor to consider is that these models do not consider the number of competitor gas companies that are also already established in the area, so data on the locations of current gas companies in New York City would also be valuable in conducting further research.

From our binary logistic regression model we were pleased to observe a positive relation between household income and borough, driving, and mean commuting time, as they provide validation to our business objectives. Since the conclusions of our logistic regression are dealt in log odds, we can simply interpret the relationships between our dependent and independent variables as thus: the odds of there being individuals earning above the average household income are highest in the borough of Queens than in the general population of regular drivers by a ratio of about 15:1. The odds of this for the borough of Queens and the overall population of those who have a high mean commute time are even more extreme by a ratio of 29:1. From this, then, it is expected for Queens to be of particular interest to us as a potentially rewarding location for business as it was shown to be the most significant of all of the other variables considered in our model. Our competitors are likely to have caught to this information as well, so we will be extra vigilant to ensure that we secure a competitive advantage over established competitors within the area.

Intuitively, the findings from our logistic model make complete sense in the context of our situation. Since we want to maximize profits, it is only natural that we would want to do business in an area with drivers who are not only regular commuters but also have the financial means of reliably sustaining such driving activities.

Recommendations

If our company were to pursue additional analyses in the future, we would suggest a number of changes to yield more beneficial and meaningful results. For one, we conducted multiple linear regression which assumes that the observations in the data set are independent. In practice, especially when working with geographic units that experience a similar environment, ensuring complete independence is virtually impossible. Accordingly, we suggest conducting a cluster analysis that identifies groups of census tracts based on some variable like commute time. The company could then specifically narrow down their target market to tracts with higher commute times. Additionally, linear regression looks at the relationship between the mean of the dependent variable and various independent variables. However, in this situation, it would be useful to look at the extremes of the outcome variable as, for instance, we would prefer a model that could predict an extraordinarily high percentage of driving/carpooling commuters. To do this, we suggest using quantile regression to model the 90th percentile of driver percentage.

Regarding our logistic regression model for income, this type of modelling is only truly useful when all of the relevant predictor variables are identified. While the data set we used includes a lot of useful information from the U.S. Census Bureau, many factors that clearly influence household income such as education level and industry type are not given. For example, if we knew the percentage of a census tract's population that has earned a Bachelor's or Master's degree, we could likely build a more accurate model. In a similar vein, before our company decides on a specific location in New York City, we believe there is some critical data that we have yet to find or analyze. First, looking at how real estate costs vary across the city would be very useful so that we can secure property while minimizing the amount of capital expended. Also, we suggest considering where competitive gas stations are already located in New York City. If a particular borough is already saturated with dozens of filling stations, we would likely want to avoid that borough. Since the Census Bureau generally does not collect information regarding property values and gas station establishments, we would have to access other databases.

By combining the results from the analyses we presented in this report and those we gain from the aforementioned modelling techniques we could use in the future, we could paint a clearer and more holistic picture of New York City's gas station market. Armed with such comprehensive statistical evidence, we are confident Empire State Gas & Fuel can successfully launch a new gas station and immediately attract customers.

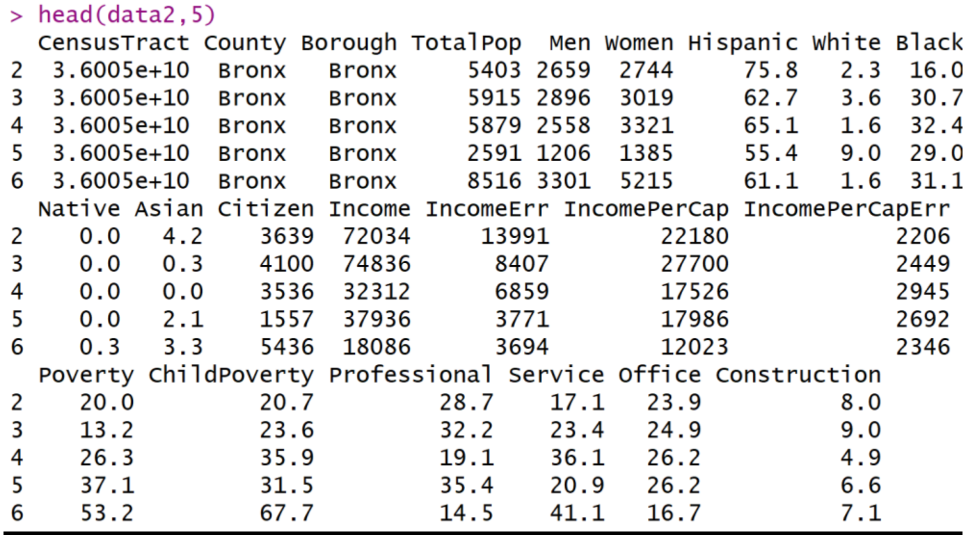
Appendix

Appendix A

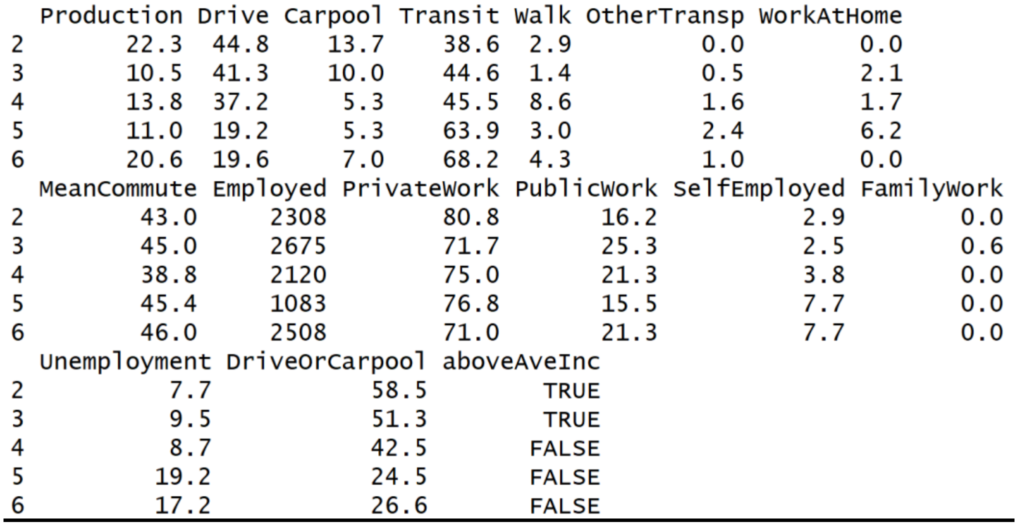
Variable Label	Variable Type	Levels of Variable
County	Factor w/ 5 levels "Bronx","Kings",..:	1 1 1 1 1 1 1 1 1 1 ...
Borough	Factor w/ 5 levels "Bronx","Brooklyn",..:	1 1 1 1 1 1 1 1 1 1 ..
TotalPop	int	7703 5403 5915 5879 2591 8516 4774 150 5355 3016 ...
Drive	num	NA 44.8 41.3 37.2 19.2 19.6 5.9 0 12.6 14 ...
Carpool	num	NA 13.7 10 5.3 5.3 7 0 0 2.8 1.7 ...
Transit	num	NA 38.6 44.6 45.5 63.9 68.2 74.5 100 62.5 64.7 ...
Walk	num	NA 2.9 1.4 8.6 3 4.3 14 0 17.7 18 ...
Income	num	NA 72034 74836 32312 37936 ...

*This table presents the structure of the original dataset we used, indicating the type of variable and the first few levels for each of the main variables for our analysis.

Appendix B



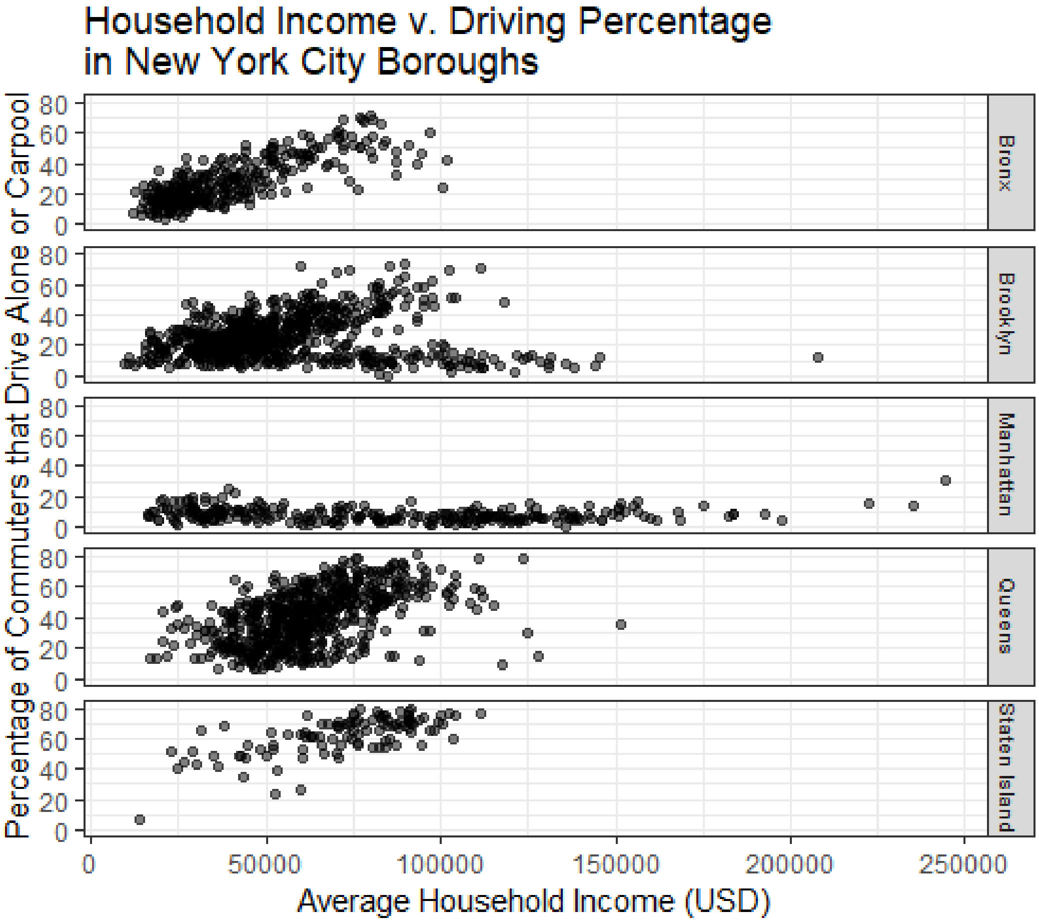
```
> head(data2,5)
  CensusTract County Borough TotalPop  Men Women Hispanic White Black
2  3.6005e+10  Bronx   Bronx     5403 2659  2744     75.8   2.3  16.0
3  3.6005e+10  Bronx   Bronx     5915 2896  3019     62.7   3.6  30.7
4  3.6005e+10  Bronx   Bronx     5879 2558  3321     65.1   1.6  32.4
5  3.6005e+10  Bronx   Bronx     2591 1206  1385     55.4   9.0  29.0
6  3.6005e+10  Bronx   Bronx     8516 3301  5215     61.1   1.6  31.1
  Native Asian Citizen Income IncomeErr IncomePerCap IncomePerCapErr
2    0.0   4.2    3639  72034     13991        22180            2206
3    0.0   0.3    4100  74836      8407        27700            2449
4    0.0   0.0    3536  32312      6859        17526            2945
5    0.0   2.1    1557  37936      3771        17986            2692
6    0.3   3.3    5436  18086      3694        12023            2346
  Poverty ChildPoverty Professional Service Office Construction
2    20.0         20.7         28.7    17.1   23.9          8.0
3    13.2         23.6         32.2    23.4   24.9          9.0
4    26.3         35.9         19.1    36.1   26.2          4.9
5    37.1         31.5         35.4    20.9   26.2          6.6
6    53.2         67.7         14.5    41.1   16.7          7.1
```



```
  Production Drive Carpool Transit Walk OtherTransp WorkAtHome
2       22.3  44.8    13.7    38.6  2.9         0.0        0.0
3       10.5  41.3    10.0    44.6  1.4         0.5        2.1
4       13.8  37.2     5.3    45.5  8.6         1.6        1.7
5       11.0  19.2     5.3    63.9  3.0         2.4        6.2
6       20.6  19.6     7.0    68.2  4.3         1.0        0.0
  MeanCommute Employed PrivateWork PublicWork SelfEmployed FamilyWork
2        43.0     2308        80.8       16.2          2.9        0.0
3        45.0     2675        71.7       25.3          2.5        0.6
4        38.8     2120        75.0       21.3          3.8        0.0
5        45.4     1083        76.8       15.5          7.7        0.0
6        46.0     2508        71.0       21.3          7.7        0.0
  Unemployment DriveOrCarpool aboveAveInc
2          7.7           58.5        TRUE
3          9.5           51.3        TRUE
4          8.7           42.5       FALSE
5         19.2           24.5       FALSE
6         17.2           26.6       FALSE
```

*This table prints out the first 5 observations for each column in the modified dataset.

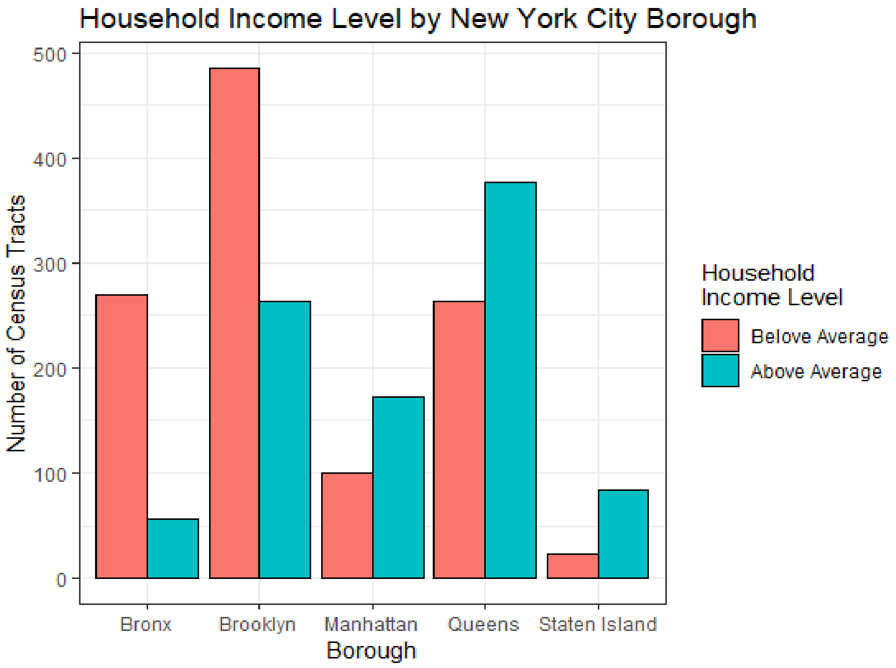
Appendix C



Household Income v. Driving Percentage in New York City Boroughs

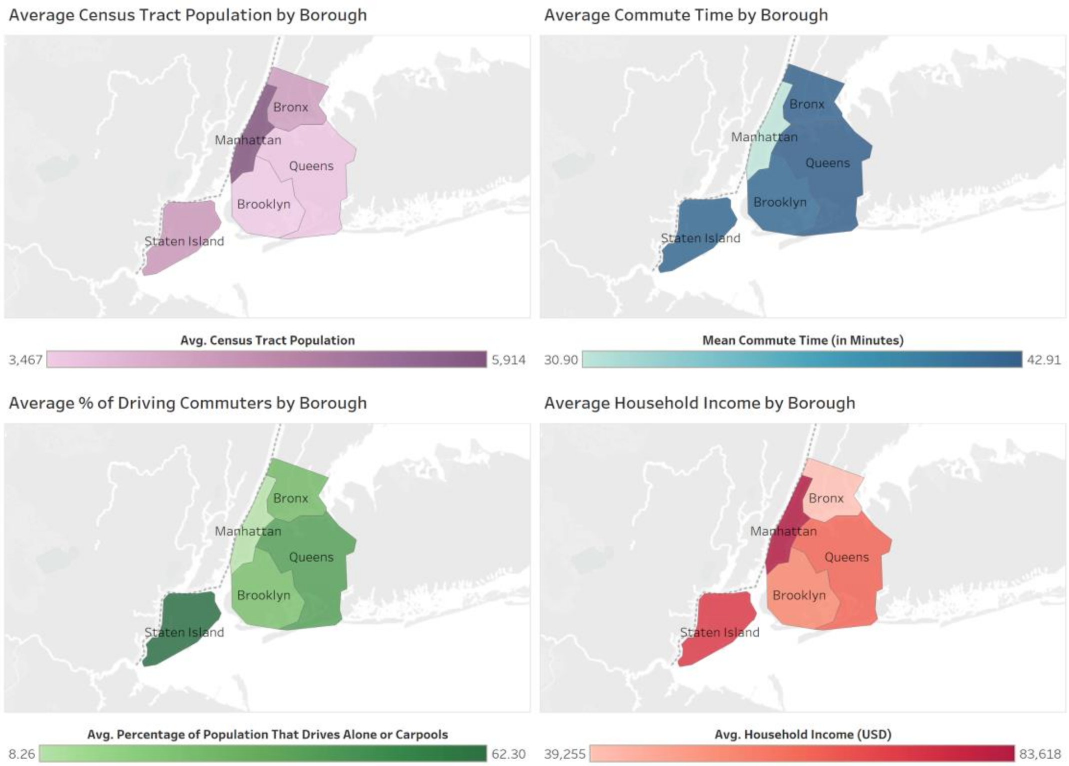
*The above scatter plot compares median household income ($) and percentage of people commuting alone or carpooling in a vehicle. The graph is also facetted by borough on the y axis.

Appendix D



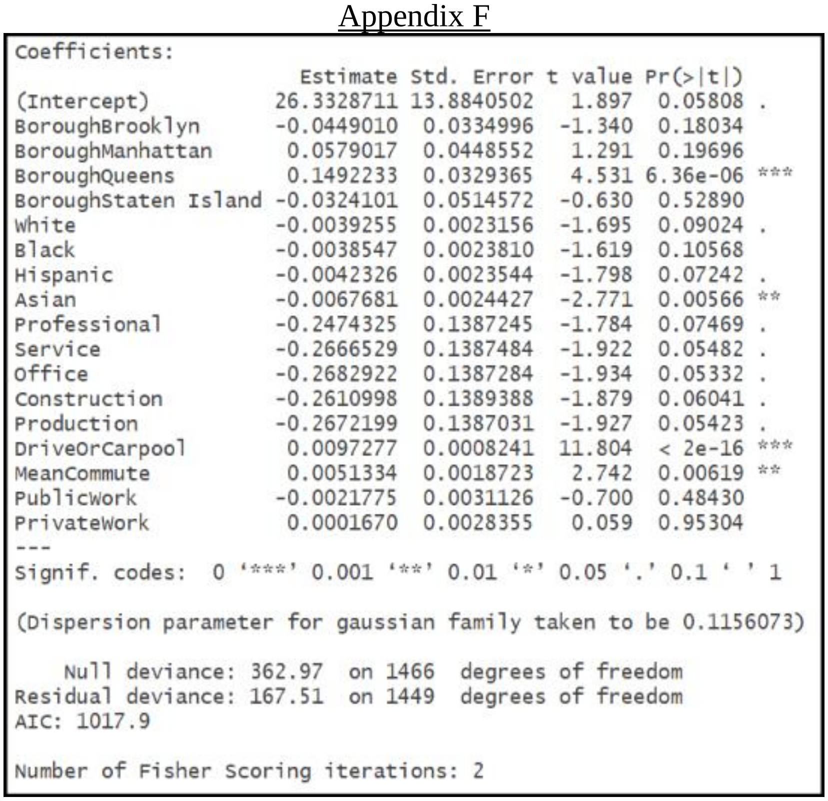
*The bar chart displays how many tracts with “Below Average” and “Above Average” household income are in each borough. The median household income in New York City is $57,782.

Appendix E


Average Census Tract Population by Borough
Bronx
Manhattan
Queens
Brooklyn
Staten Island
Avg. Census Tract Population
3,467
5,914
Average Commute Time by Borough
Bronx
Manhattan
Queens
Brooklyn
Staten Island
Mean Commute Time (in Minutes)
30.90
42.91
Average % of Driving Commuters by Borough
Bronx
Manhattan
Queens
Brooklyn
Staten Island
Avg. Percentage of Population That Drives Alone or Carpools
8.26
62.30
Average Household Income by Borough
Bronx
Manhattan
Queens
Brooklyn
Staten Island
Avg. Household Income (USD)
39,255
83,618

*The Tableau dashboard consists of four map charts displaying the five New York City boroughs with a different continuous variable mapped to color in each visualization.

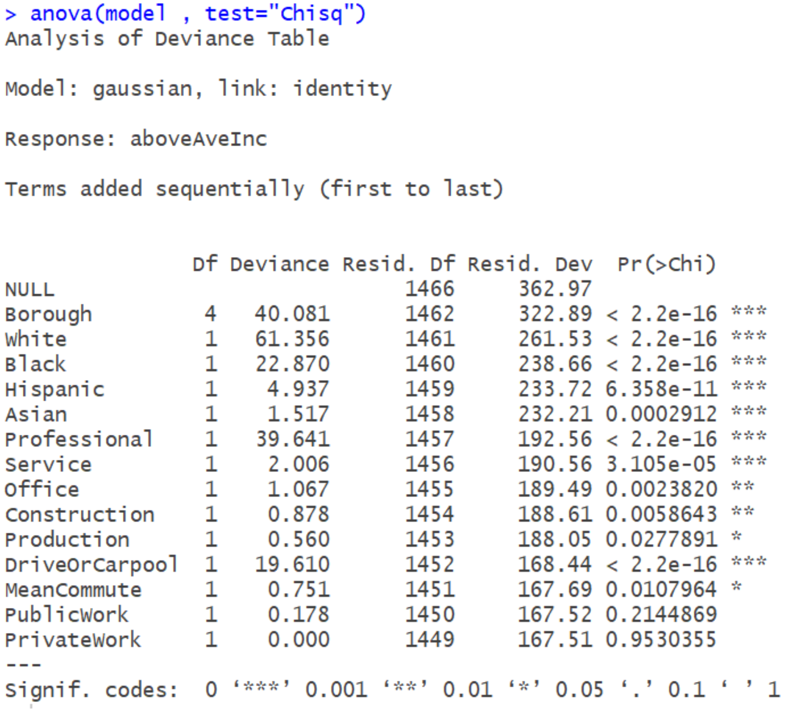
Appendix F



```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         26.3328711 13.8840502   1.897  0.05808 .  
BoroughBrooklyn     -0.0449010  0.0334996  -1.340  0.18034    
BoroughManhattan     0.0579017  0.0448552   1.291  0.19696    
BoroughQueens        0.1492233  0.0329365   4.531 6.36e-06 ***
BoroughStaten Island -0.0324101  0.0514572  -0.630  0.52890    
White               -0.0039255  0.0023156  -1.695  0.09024 .  
Black               -0.0038547  0.0023810  -1.619  0.10568    
Hispanic            -0.0042326  0.0023544  -1.798  0.07242 .  
Asian               -0.0067681  0.0024427  -2.771  0.00566 ** 
Professional        -0.2474325  0.1387245  -1.784  0.07469 .  
Service             -0.2666529  0.1387484  -1.922  0.05482 .  
Office              -0.2682922  0.1387284  -1.934  0.05332 .  
Construction        -0.2610998  0.1389388  -1.879  0.06041 .  
Production          -0.2672199  0.1387031  -1.927  0.05423 .  
DriveOrCarpool       0.0097277  0.0008241  11.804  < 2e-16 ***
MeanCommute          0.0051334  0.0018723   2.742  0.00619 ** 
PublicWork          -0.0021775  0.0031126  -0.700  0.48430    
PrivateWork          0.0001670  0.0028355   0.059  0.95304    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1156073)

    Null deviance: 362.97  on 1466  degrees of freedom
Residual deviance: 167.51  on 1449  degrees of freedom
AIC: 1017.9

Number of Fisher Scoring iterations: 2
```

*This table prints about the estimated log odds, standard errors, t-values, and p-values for each independent variable used in the logistic regression model.

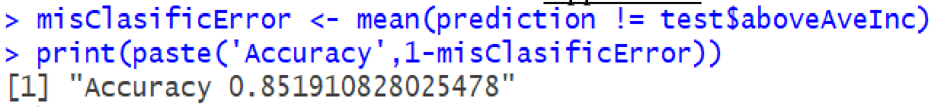
Appendix G



```
> anova(model , test="Chisq")
Analysis of Deviance Table

Model: gaussian, link: identity

Response: aboveAveInc

Terms added sequentially (first to last)


              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                           1466     362.97              
Borough        4   40.081      1462     322.89 < 2.2e-16 ***
White          1   61.356      1461     261.53 < 2.2e-16 ***
Black          1   22.870      1460     238.66 < 2.2e-16 ***
Hispanic       1    4.937      1459     233.72 6.358e-11 ***
Asian          1    1.517      1458     232.21 0.0002912 ***
Professional   1   39.641      1457     192.56 < 2.2e-16 ***
Service        1    2.006      1456     190.56 3.105e-05 ***
Office         1    1.067      1455     189.49 0.0023820 ** 
Construction   1    0.878      1454     188.61 0.0058643 ** 
Production     1    0.560      1453     188.05 0.0277891 *  
DriveOrCarpool 1   19.610      1452     168.44 < 2.2e-16 ***
MeanCommute    1    0.751      1451     167.69 0.0107964 *  
PublicWork     1    0.178      1450     167.52 0.2144869    
PrivateWork    1    0.000      1449     167.51 0.9530355    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*The above table is the ANOVA table for our logistic regression

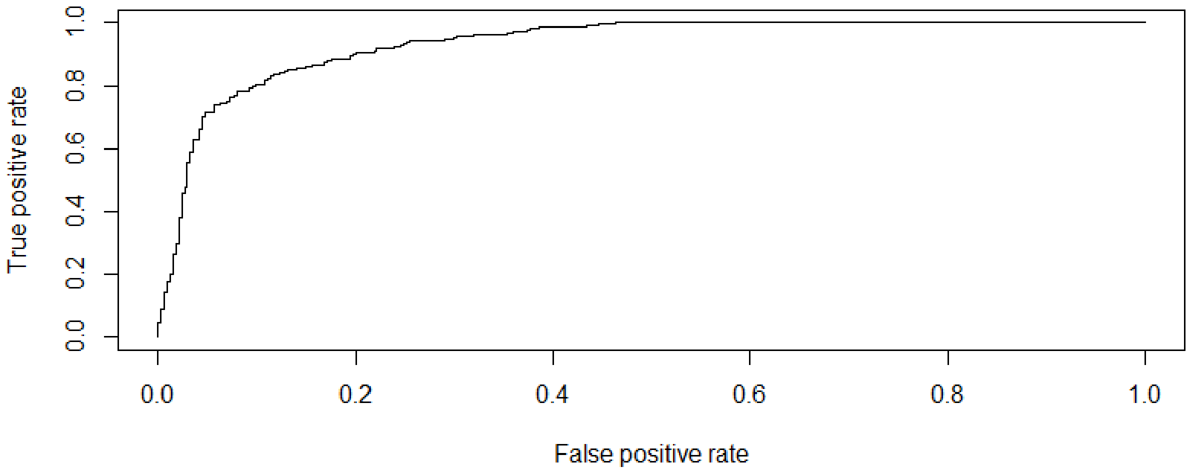
Appendix H



```
> misClasificError <- mean(prediction != test$aboveAveInc)
> print(paste('Accuracy',1-misClasificError))
[1] "Accuracy 0.851910828025478"
```

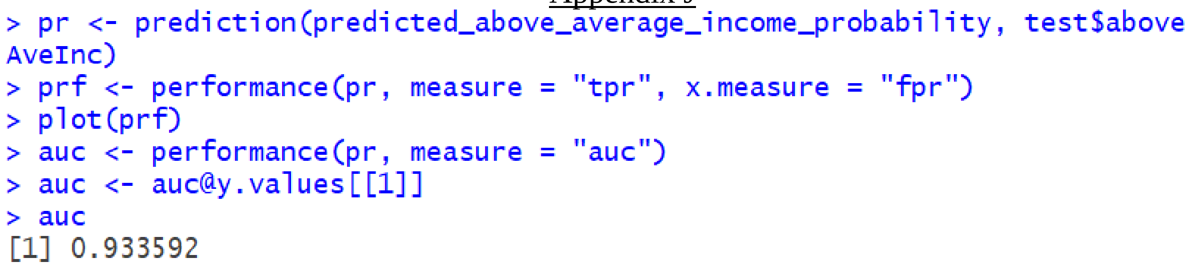
*Calculated predictability accuracy of our test set.

Appendix I



*Our ROC with the true positive rate (TPR) plotted against the false positive rate (FPR). The area under the curve resulted in a 0.93 .A model with strong predictability should have an area closer to 1 (1 being ideal) than to 0.5.

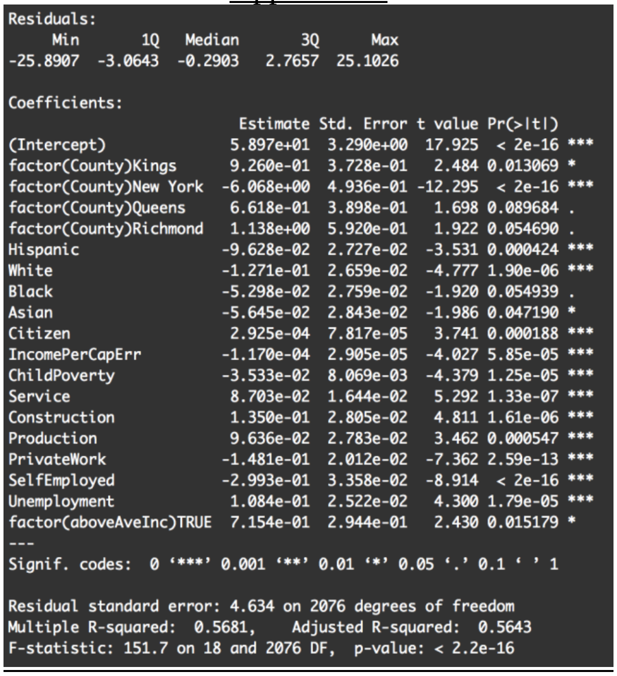
Appendix J



```
> pr <- prediction(predicted_above_average_income_probability, test$above
AveInc)
> prf <- performance(pr, measure = "tpr", x.measure = "fpr")
> plot(prf)
> auc <- performance(pr, measure = "auc")
> auc <- auc@y.values[[1]]
> auc
[1] 0.933592
```

*ROC curve plot (ending with 'plot(prf)' and calculation for the area under the ROC curve (starting with the first 'auc…') .

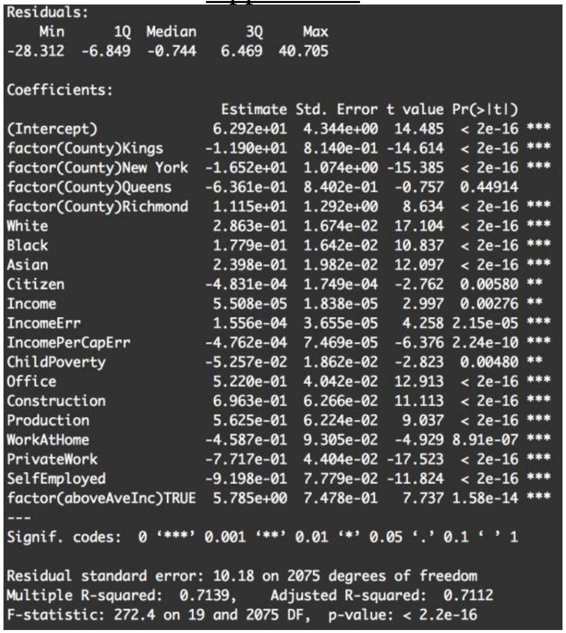
Appendix K



```
Residuals:
     Min       1Q   Median       3Q      Max 
-25.8907  -3.0643  -0.2903   2.7657  25.1026 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              5.897e+01  3.290e+00  17.925  < 2e-16 ***
factor(County)Kings      9.260e-01  3.728e-01   2.484 0.013069 *  
factor(County)New York  -6.068e+00  4.936e-01 -12.295  < 2e-16 ***
factor(County)Queens     6.618e-01  3.898e-01   1.698 0.089684 .  
factor(County)Richmond   1.138e+00  5.920e-01   1.922 0.054690 .  
Hispanic                -9.628e-02  2.727e-02  -3.531 0.000424 ***
White                   -1.271e-01  2.659e-02  -4.777 1.90e-06 ***
Black                   -5.298e-02  2.759e-02  -1.920 0.054939 .  
Asian                   -5.645e-02  2.843e-02  -1.986 0.047190 *  
Citizen                  2.925e-04  7.817e-05   3.741 0.000188 ***
IncomePerCapErr         -1.170e-04  2.905e-05  -4.027 5.85e-05 ***
ChildPoverty            -3.533e-02  8.069e-03  -4.379 1.25e-05 ***
Service                  8.703e-02  1.644e-02   5.292 1.33e-07 ***
Construction             1.350e-01  2.805e-02   4.811 1.61e-06 ***
Production               9.636e-02  2.783e-02   3.462 0.000547 ***
PrivateWork             -1.481e-01  2.012e-02  -7.362 2.59e-13 ***
SelfEmployed            -2.993e-01  3.358e-02  -8.914  < 2e-16 ***
Unemployment             1.084e-01  2.522e-02   4.300 1.79e-05 ***
factor(aboveAveInc)TRUE  7.154e-01  2.944e-01   2.430 0.015179 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.634 on 2076 degrees of freedom
Multiple R-squared:  0.5681,	Adjusted R-squared:  0.5643 
F-statistic: 151.7 on 18 and 2076 DF,  p-value: < 2.2e-16
```

*This table prints the value of the coefficients for each of the independent variables in the linear model that predicts the mean commute time

Appendix L



```
Residuals:
    Min      1Q  Median      3Q     Max 
-28.312  -6.849  -0.744   6.469  40.705 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            6.292e+01  4.344e+00  14.485  < 2e-16 ***
factor(County)Kings   -1.190e+01  8.140e-01 -14.614  < 2e-16 ***
factor(County)New York -1.652e+01  1.074e+00 -15.385  < 2e-16 ***
factor(County)Queens  -6.361e-01  8.402e-01  -0.757  0.44914    
factor(County)Richmond  1.115e+01  1.292e+00   8.634  < 2e-16 ***
White                  2.863e-01  1.674e-02  17.104  < 2e-16 ***
Black                  1.779e-01  1.642e-02  10.837  < 2e-16 ***
Asian                  2.398e-01  1.982e-02  12.097  < 2e-16 ***
Citizen               -4.831e-04  1.749e-04  -2.762  0.00580 ** 
Income                 5.508e-05  1.838e-05   2.997  0.00276 ** 
IncomeErr              1.556e-04  3.655e-05   4.258 2.15e-05 ***
IncomePerCapErr       -4.762e-04  7.469e-05  -6.376 2.24e-10 ***
ChildPoverty          -5.257e-02  1.862e-02  -2.823  0.00480 ** 
Office                 5.220e-01  4.042e-02  12.913  < 2e-16 ***
Construction           6.963e-01  6.266e-02  11.113  < 2e-16 ***
Production             5.625e-01  6.224e-02   9.037  < 2e-16 ***
WorkAtHome            -4.587e-01  9.305e-02  -4.929 8.91e-07 ***
PrivateWork           -7.717e-01  4.404e-02 -17.523  < 2e-16 ***
SelfEmployed          -9.198e-01  7.779e-02 -11.824  < 2e-16 ***
factor(aboveAveInc)TRUE  5.785e+00  7.478e-01   7.737 1.58e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.18 on 2075 degrees of freedom
Multiple R-squared:  0.7139,	Adjusted R-squared:  0.7112 
F-statistic: 272.4 on 19 and 2075 DF,  p-value: < 2.2e-16
```

*This table prints the value of the coefficients for each of the independent variables in the linear model that predicts the percentage of commuters who drive alone or carpool

Works Cited

1. MuonNeutrino from Kaggle (2017). *New York City Census Data: Demographic, Economic, and Location Data for Census Tracts in NYC*. Retrieved from https://www.kaggle.com/muonneutrino/new-york-city-census-data#nyc_census_tracts.cs

2. Population Studies Center of the University of Michigan. *Data Quality Issues with the American Community Survey (ACS).* Retrieved from https://www.psc.isr.umich.edu/dis/acs/aggregator/

3. The City of New York (n/a). *New York City Census FactFinder (NYC CFF)*. Retrieved from https://www1.nyc.gov/assets/planning/download/pdf/data-maps/maps-geography/census-factfinder/cff-faq.pdf

4. U.S. Census Bureau (2010-2017). *QuickFacts: New York City, New York.* Retrieved from https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork/PST040217

5. U.S. Census Bureau (2014). *American Community Survey (ACS): Design and Methodology Report.* Retrieved from https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html