Predicting the Concentration of Starbucks Locations using Socioeconomic Data

---

Prepared by:

Trevor Stull

February 2020

Coursera Capstone Project

Table of Contents

List of Figures

List of Tables

1.0 Introduction/Background

1.1 Background

The Starbucks Coffee Company was founded in 1971 with one coffee shop in Seattle, Washington. Since those humble beginnings, the company has grown to an international company with coffee shops open across the world. Like with any company, knowing what customer the product is trying to reach is important to Starbucks' success and growth. It is also paramount for a company to maximize its resources and ensure the company is efficiently using them. Keeping this in mind, it would be worthwhile to predict the number of Starbucks stores needed in an area based on its population's socioeconomic makeup. This would allow the company to use its resources more efficiently by either adding store locations in areas predicted to need more stores or eliminate store locations in areas that are predicted not to need the current number of stores in the location.

1.2 Problem Statement

Data that might contribute to the concentration of Starbucks in an area include the area's population, average age, per capita income, and race/ethnicity. This project will aim to try and quantify these relations and develop an expression to predict the number of Starbucks in an area.

1.3 Interest

This information could be used by the Starbucks Coffee Company to determine the optimum number of Starbucks stores needed in a location or could be leveraged to determine the typical Starbucks customer in America.

2.0 Data Acquisition and Cleaning

2.1 Data Sources

Foursquare data was leveraged to gather Starbucks location information throughout Dallas and Tarrant counties in North Texas. Socioeconomic factors such as population counts, average age, race/ethnicity, and per capita income will be obtained from the US Census data American Fact Finder tool. The Integrated Public Use Microdata Series (IPUMS) National Historical Geographic Information System (NHGIS) service provides easy to use processed data from the American Fact Finder tool and was leveraged for obtaining US Census data throughout this analysis.

2.2 Foursquare Data

Foursquare data was obtained leveraging Foursquare API calls along with Foursquare's search function to gather Starbucks locations in Dallas and Tarrant counties. Both Dallas and Tarrant county were divided into 256 evenly sized squares using GIS. Each square was searched for Starbucks locations and all unique Starbucks locations were recorded. A unique Starbucks location was defined to be any Starbucks location found with a latitude and/or longitude value that had not been recorded before. The unique Starbucks locations were counted within each census tract using GIS. Figure 2-1 and Figure 2-2 show the Starbucks locations identified in Dallas and Tarrant county.

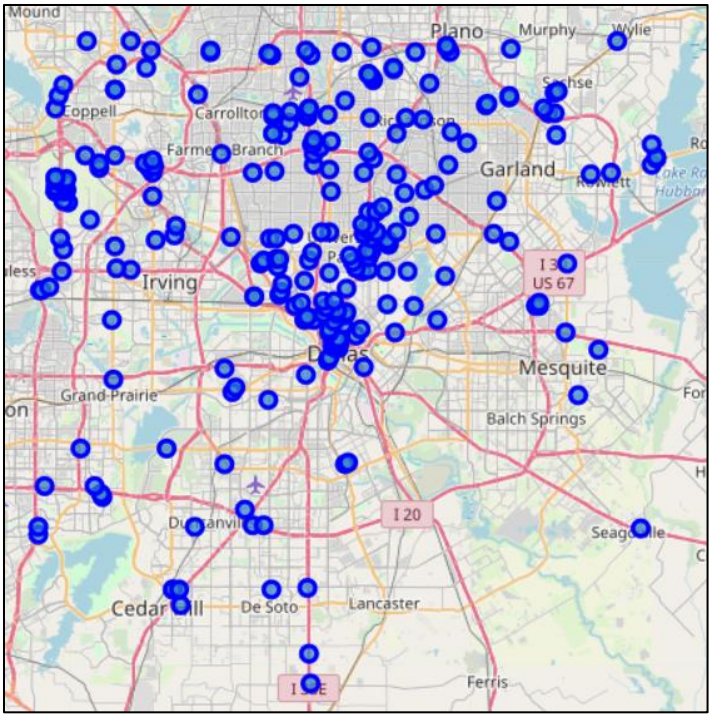

Figure 2-1: Starbucks Locations in Dallas County

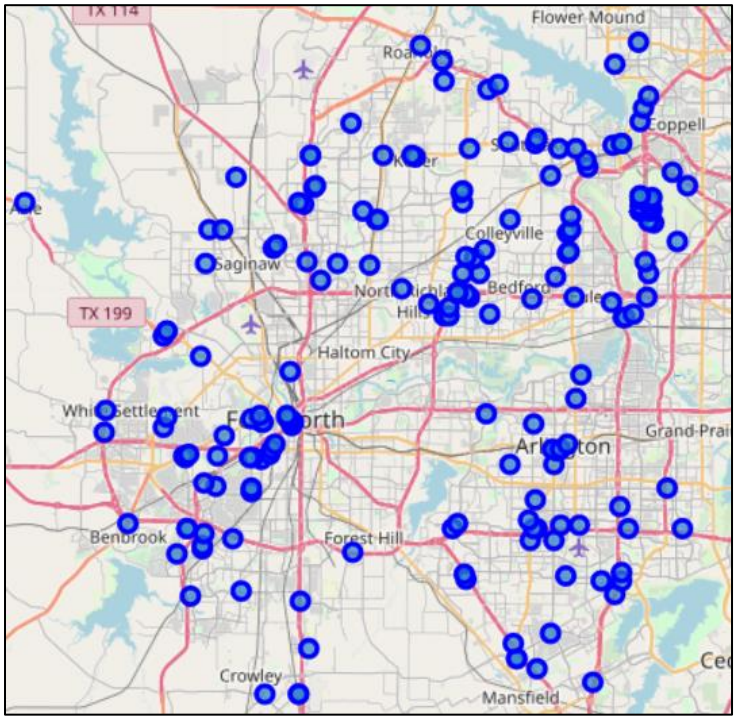

Figure 2-2: Starbucks Locations in Tarrant County

2.3 US Census Data

US census tracts were the geographic level used to collect US Census data. Census tracts are relatively small geographic boundaries that are relatively statically permanent in a county. Census tracts typically have an average population 4,000 people and can range from a population of 1,200 to 8,000 people. The following data was collected for the census tracts in both Dallas and Tarrant County. This data was from the 2013-2017 estimates developed by the American Community Survey. Census tracts with no total population were excluded from this analysis.

- Total population – This is the estimated total population within the census tract from 2013-2017.
- Per capita income – This is the estimated average income for an individual within the census tract from 2013-2017 and is inflation adjusted to 2017.
- White only percentage – This is the percentage of people in a census tract who are white only and were not Hispanic from 2013-2017.
- Hispanic only percentage – This is the percentage of people in a census tract who are white only and were Hispanic from 2013-2017.
- Black/African American only percentage – This is the percentage of people in a census tract who are black/African American from 2013-2017.
- American Indian only percentage – This is the percentage of people in a census tract who are American Indian/Native American from 2013-2017.
- Asian only percentage – This is the percentage of people in a census tract who are Asian from 2013-2017.
- Native Hawaiian percentage – This is the percentage of people in a census tract who are Native Hawaiian/Pacific Islander from 2013-2017.
- Other race percentage – This is the percentage of people in a census tract who are another race that listed above from 2013-2017.
- Two or more races percentage – This is the percentage of people in a census tract who are two or more races from 2013-2017.
- Male Percentage – This is the percentage of males in a census tract from 2013-2017.
- Female Percentage – This is the percentage of females in a census tract from 2013-2017.
- Percentage of people aged 18 to 24
- Percentage of people aged 25 to 34
- Percentage of people aged 35 to 44
- Percentage of people aged 45 to 54
- Percentage of people aged 55 to 64
- Percentage of people aged 65 and over

3.0 Methodology

Only Dallas county US Census data and Starbucks count data was used to develop a model in order to test the model versus both the test set and the Tarrant county data. All data summarized within this section is for Dallas county only.

3.1 Starbucks Count Distribution

Figure 3-1 summarizes the distribution of the number of Starbucks locations within a census tract in Dallas county. Dallas county contains a total of 527 census tracts (excluding census tracts with no total

population). It is clear from this figure that a large majority of the census tracts in Dallas county contain 2 or fewer Starbucks locations, with most of the census tracts containing no Starbucks locations. This does not provide a large sample size for census tracts with 3 or more Starbucks locations, which could make model prediction of these types of census tracts difficult. Also, the large number of census tracts with 2 or fewer Starbucks locations could lead to a wide variation in US Census data for census tracts in this category. This could cause the model to overpredict these types of census tracts because the wide variation in data could make it hard to identify differences in the census tract groups.

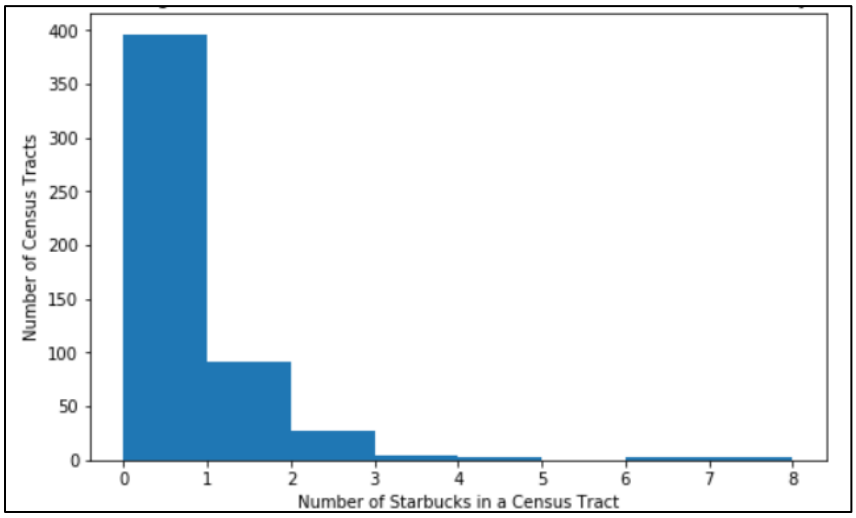

Figure 3-1: Histogram of the Number of Starbucks Within a Census Tract in Dallas County

3.2 US Census Data Correlation to the Number Starbucks Locations

The Pearson correlation between each US Census data metric obtained and the number of Starbucks locations within a census tract was calculated to evaluate how strongly each metric was correlated to the number of Starbucks locations in a census tract. Table 3-1 summarizes the correlation of each US Census metric and the number of Starbucks locations in a census tract. If a metric had a positive correlation value, then higher values of the metric correlated to a higher number of Starbucks locations in a census tract. Conversely, a negative correlation meant higher values of the data metric correlated to a lower number of Starbucks locations in a census tract.

Table 3-1: US Census Data Correlations in Dallas County

US Census Data Metric	Pearson Correlation to the Number of Starbucks Location in a Census Tract
Total Population	0.031
Per Capita Income	0.252
Male Percentage	0.076
Female Percentage	-0.076
Percentage of People Age 18 to 24	0.029
Percentage of People Age 25 to 34	0.231
Percentage of People Age 35 to 44	0.083
Percentage of People Age 45 to 54	0.007

Percentage of People Age 55 to 64	0.076
Percentage of People Age 65 and Over	0.015
White Percentage	0.246
Hispanic Percentage	-0.238
Black Percentage	-0.093
American Indian Percentage	0.013
Asian Percentage	0.151
Native Hawaiian Percentage	0.013
Other Percentage	-0.081
Two or More Race Percentage	0.010

All data metrics do not show a strong correlation to the number of Starbucks locations in a census tract. Per capita income showed the highest absolute correlation value to the number of Starbucks locations in a census tract at 0.252. A correlation value near 1 or -1 generally means the metric correlates strongly to the metric you are trying to predict. Since no single US census metric correlates very strongly to the number for Starbucks locations in a census tract, it could mean that multiple census data metrics play a role in determining how many Starbucks locations get built in an area. This makes sense, because typically companies try to cater a product to target a specific group in order to provide the most effective service to that group.

Furthermore, since there are few census tracts with multiple Starbucks locations, the correlation of each data metric could be easily impacted due to the variability of these low sample sizes. For example, the data metric could generally trend down as Starbucks locations increase from 0 to 5 but have a correlation close to zero if the handful of census tracts with 6 or more Starbucks locations all have abnormally high data metric values compared to the rest of the census tracts.

3.3 Machine Learning Model Development

To account for the likelihood multiple US Census data metrics play in a role in determining the number of Starbucks locations, a multiple linear regression machine learning algorithm was used to develop a model to predict the number of Starbucks locations in a census tract. US Census data metrics with an absolute correlation value less than 0.1 were not used to develop the machine learning algorithm. This was done to eliminate the potential increased variability caused by data metrics with very low correlations to Starbucks locations. A simplified model was also used to combat the potential of overfitting with increased model complexity. The US Census data metrics used to develop the linear regression model are listed below.

- Per Capita Income (correlation of 0.252)
- White Percentage (correlation of 0.246)
- Hispanic Percentage (correlation of -0.238)
- Percentage of People Age 25 to 34 (correlation of 0.231)
- Asian Percentage (correlation of 0.151)

To further evaluate the correlation of these five data metrics to the number of Starbucks locations in a census tract, box and whisker plots were developed for each metric for each group of census tracts with the same number of Starbucks locations. Figure 3-2 through Figure 3-6 show the box and whisker plots developed for each data metric. The following observations can be drawn from these figures.

- All median values generally increase if the data metric has a positive correlation to the number of Starbucks locations or decrease if the data metric has a negative correlation to the number of Starbucks locations.
- Generally, the census tracts with lower total number of Starbucks locations have a larger range in the data metric values. This is due to the large number of census tracts in these categories, which naturally leads to increased variability in the data.
- Figure 3-5 and Figure 3-6 show cases where the medians follow a similar pattern for census tracts with 6 or fewer Starbucks locations, but census tracts with 7 or more Starbucks locations deviate wildly from the overall pattern seen for census tracts with 6 or fewer Starbucks locations.
- For each data metric, there is a large amount of overlap between the interquartile ranges of the box and whisker plots for each group of census tracts with the same number of Starbucks locations. This could potentially lead to incorrect predictions because the model cannot distinguish how the metric differs between the census tracks with a different number of Starbucks locations.

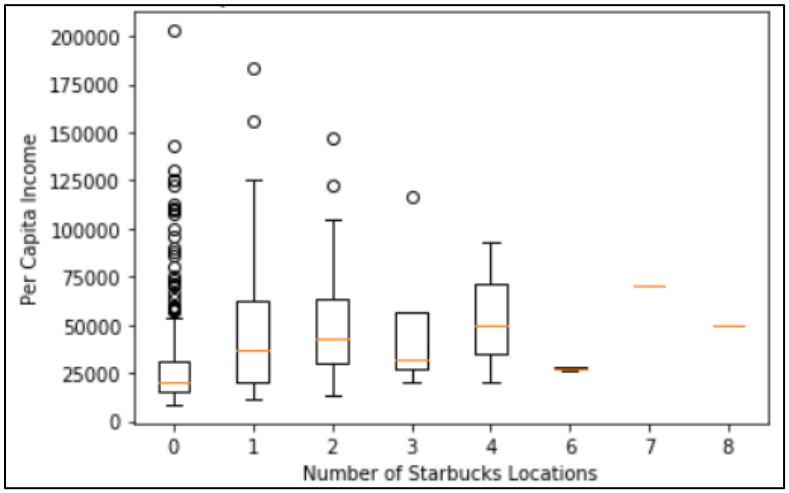

Figure 3-2: Per Capita Income vs. Number of Starbucks Locations

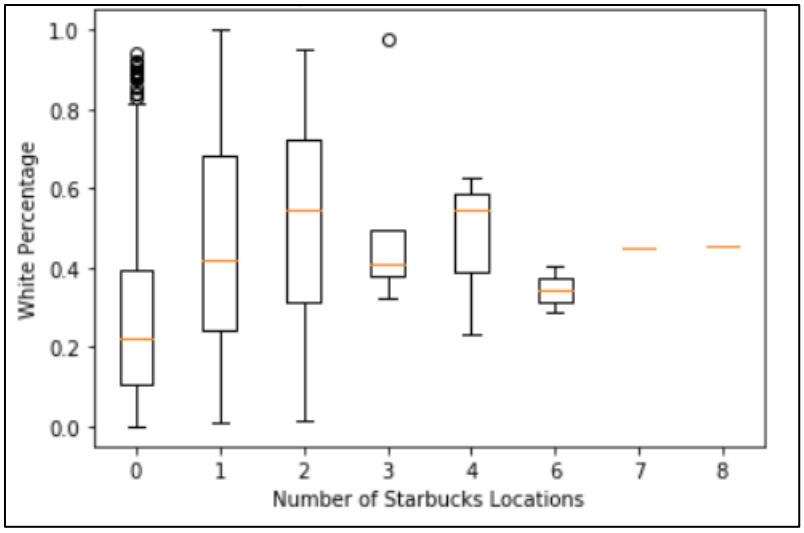

Figure 3-3: White Percentage vs. Number of Starbucks Locations

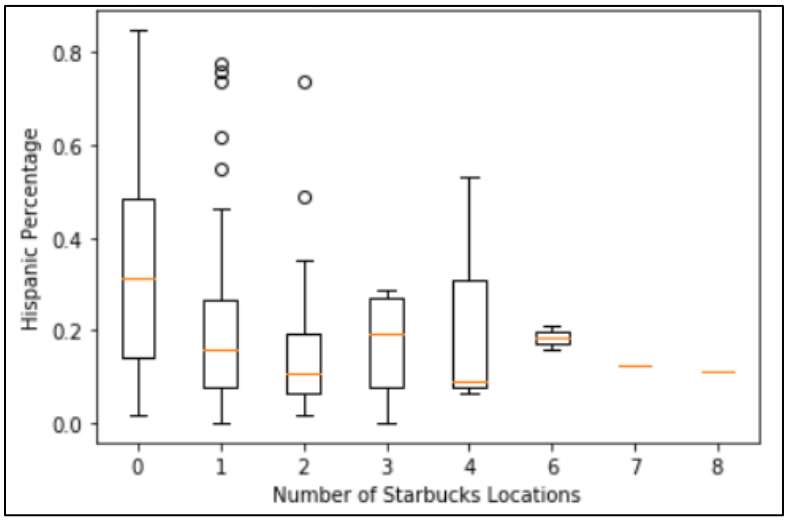

Figure 3-4: Hispanic Percentage vs. Number of Starbucks Locations

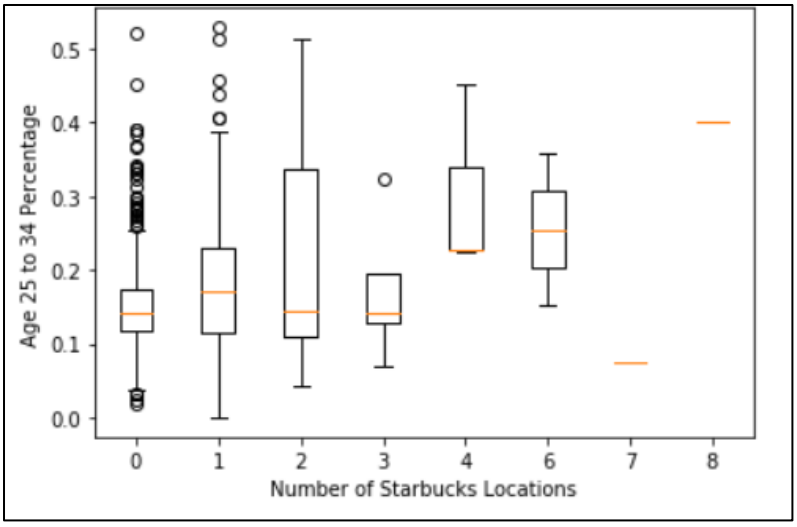

Figure 3-5: Percentage of People Age 25 to 34 vs. Number of Starbucks Locations

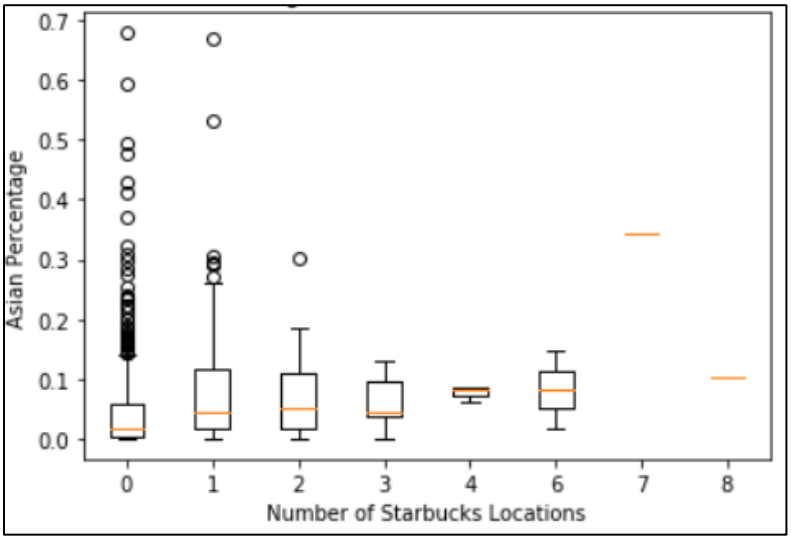

Figure 3-6: Asian Percentage vs. Number of Starbucks Locations

A range of test and training sets were used to determine the optimal test and training set sizes for this multiple linear regression model. Test set sizes from 5% to 50% of the data points were evaluated. Data was normalized using scikit learn library standard scalar normalization. To evaluate the model performance for each test and training set combination, the predicted number of Starbucks locations was rounded to the nearest whole number and compared to the actual number of Starbucks locations in the census tract. Figure 3-7 summarizes the accuracy of the model developed for each test set size to predict values in both the training set and test set. A test set size of 40% was used for the final model development, because the test set accuracy had a local maximum at this test set size.

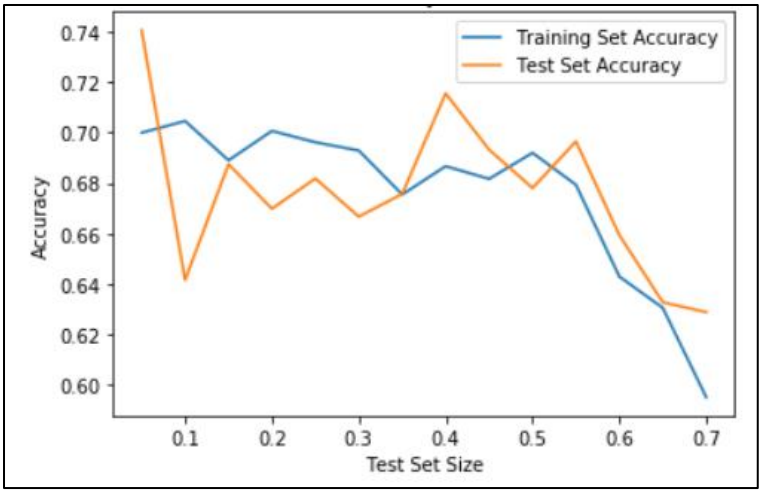

Figure 3-7: Accuracies of Predicted Values for Various Test Set Sizes

The final model developed with normalized data and a test set size of 40% is summarized below.

$$\begin{aligned} Number\ of\ Starbucks \\ &= 0.117 * (Per\ Capita\ Income) + 0.206 * (Percentage\ of\ People\ Age\ 25\ to\ 34) \\ &+ 0.041 * (White\ Percentage) - 0.081 * (Hispanic\ Percentage) \\ &+ 0.086 * (Asian\ Percentage) + 0.390 \end{aligned}$$

4.0 Results

The model's performance was tested for both the training set and test set for Dallas county data and the entire Tarrant county data set. The model summarized in Section 3.3 was used for the following evaluations and performance measures outlined in this section. Predicted values for the number of Starbucks location in a census tract were rounded to the nearest whole number to be able to compare to actual Starbucks locations values. Table 4-1 summarizes the accuracy the model was able to predict the number of Starbucks locations in each census tract based on the five US Census data metrics used. Overall, the model can predict the number of Starbucks locations within a census tract most of the time.

Table 4-1: Model Accuracy in Predicting Starbucks Locations

Census Tract Data Set	Model Accuracy (%)
Dallas County – Training Set	68.2
Dallas County – Test Set	69.3
Tarrant County	61.0

Figure 4-1 through Figure 4-3 summarize the distribution of the census tracts in each data set based on the number of Starbucks locations. From these comparisons, it can be seen that the multiple linear regression model developed does not accurately predict the number of Starbucks in a census tract if it has two of more Starbucks. Also, the model is predicting census tracts to have one Starbucks location significantly more often than the actual number of census tracts with one Starbucks location. Conversely,

the model is predicting census tracts to have no Starbucks location significantly less often than the actual number of census tracts with no Starbucks location.

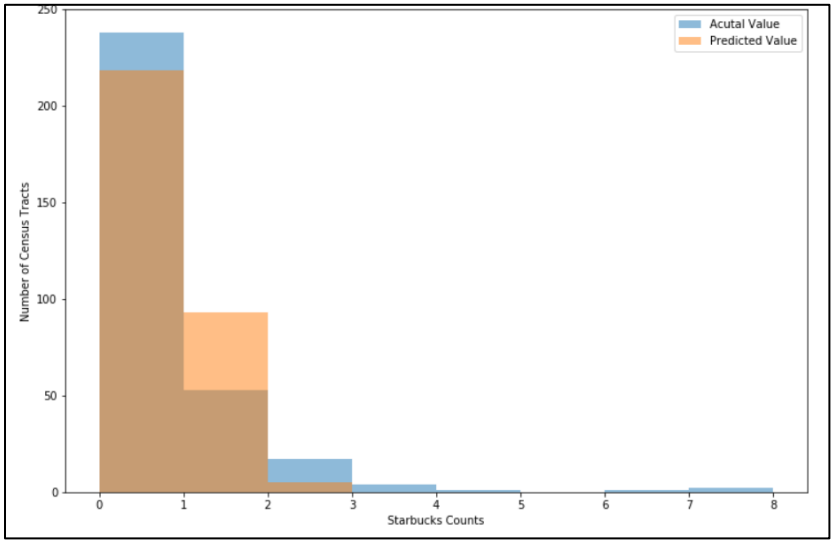

Figure 4-1: Actual vs. Predicted Values for Dallas County – Training Set

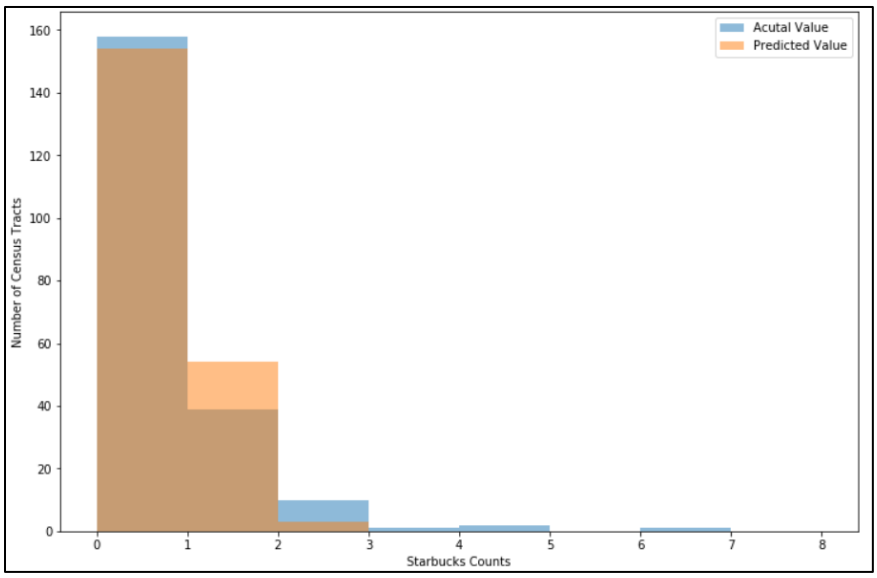

Figure 4-2: Actual vs. Predicted Values for Dallas County – Test Set

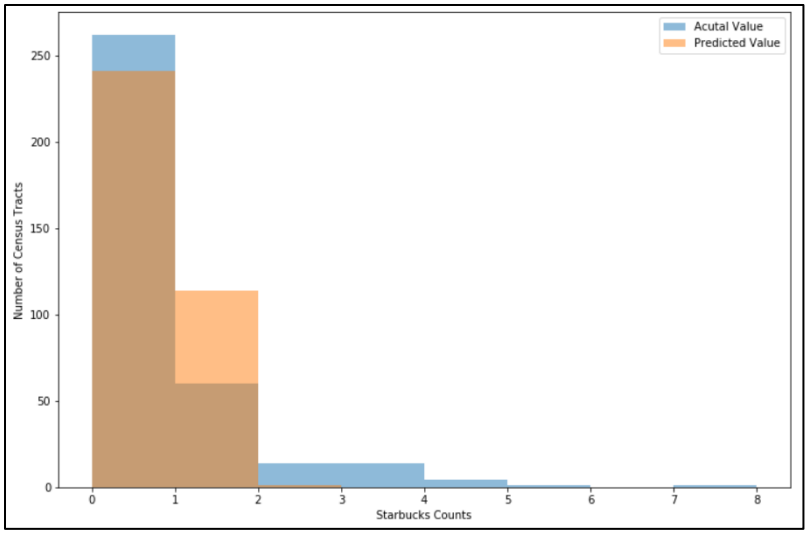

Figure 4-3: Actual vs. Predicted Values for Tarrant County

5.0 Discussion

Overall, the model developed predicts the number of Starbucks locations in a census tract correctly most of the time but does not accurately predict census tracts with 2 or more Starbucks locations. From Figure 2-1 and Figure 2-2, it can be seen that a large number of Starbucks are located in both Dallas and Fort Worth downtowns. These areas are generally where most of the Dallas/Fort Worth metroplex works, so it makes since large concentrations of Starbucks are in these areas. The model could likely be improved by incorporating data related to labor statistics. This will probably allow the model to perform better when predicting Starbucks locations numbers for areas with 2 or more Starbucks locations.

Tarrant county data was briefly investigated to determine why the model preformed slightly worse for Tarrant county compared to Dallas county. The Pearson correlation between each US Census data metric gathered and the number of Starbucks locations within a census tract was calculated. The following US Census metrics had an absolute correlation value greater than 0.1. Bolded metrics were also identified to have an absolute correlation value greater than 0.1 in the Dallas county data.

- Total Population (correlation of 0.198)
- **Per Capita Income** (correlation of 0.233)
- Percentage of People Age 18 to 24 (correlation of 0.190)
- **White Percentage** (correlation of 0.184)
- **Hispanic Percentage** (correlation of -0.193)
- **Asian Percentage** (correlation of 0.153)
- Other Percentage (correlation of -0.157)

While many of the same metrics showed relatively strong correlations in both data sets, some metrics showed stronger correlations in different data sets. This could mean that the model may have to be location specific or include more metrics.

Based on the similarities identified in some of the US Census metrics between the Dallas and Tarrant county data, a general customer target group could be identified for the Dallas/Fort Worth area. It appears Starbucks tries to target individuals who are younger (less than 35 years old), white, and wealthy.

The model developed excluded census tracts that contained the Dallas/Fort Worth and Dallas Love Field airports. These airport census tracts contained many Starbucks locations. For the Dallas/Fort Worth airport, approximately 18 Starbucks are within its census tract. In order to predict the number of Starbucks needed in an airport, a separate model may need to be developed because airports are unique situations. The number of Starbucks needed at an airport would probably be more related to the number of flights/travelers handled by the airport than US Census and labor data.

6.0 Conclusion

This project shows that a model could be developed to predict the number of Starbucks locations in a census tract based on US Census data. Including more US Census metrics as well as labor statistics could improve the model's performance for census tracts with many Starbucks locations. Furthermore, this model development and data evaluation identified one group of people Starbucks tries to target. It appears Starbucks tries to target individuals who are younger (less than 35 years old), white, and wealthy.