Predicting the Concentration of Starbucks Locations using Socioeconomic Data

Created By: Trevor Stull

Created: February 2020

Introduction

- Being able to predict the number of Starbucks locations in an area based on socioeconomic metrics could be valuable to the company.
- This model could answer the questions:
  - Does the Starbucks company need any additional locations based on the predictions from the model?
  - Are the current Starbucks locations being built in areas to target the sector of the market the company is trying to target?

Data Collection and Processing

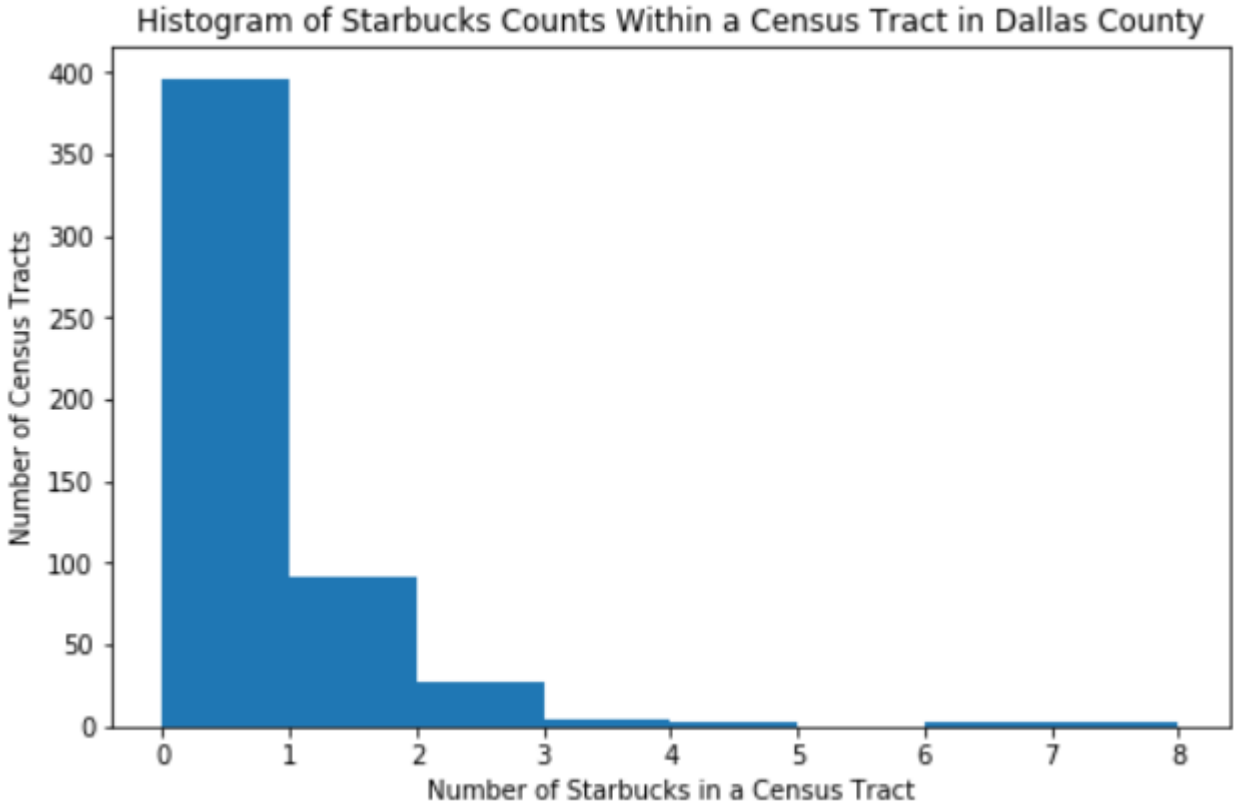
- Foursquare API calls were used to find Starbucks locations in Dallas and Tarrant county.
- US Census data was downloaded from IMPUS NHGIS website.
- GIS was used to process some of the geographic information collected.
- The different data sets were combined and paired down to a single data set containing 19 columns.
- Census tracts with no total population were excluded from the analysis (This happened to be DFW and Dallas Love Field Airports)

Model Development

- Dallas county specific data was investigated and used to develop a model to predict the number of Starbucks locations in a census tract
- Once the model was created, it was tested against Tarrant county data to determine if the model of location specific

Dallas County Data

- Most of the census tracts contain 2 or fewer Starbucks locations
- This could lead to prediction inaccuracies for census tracts with 3 or more Starbucks



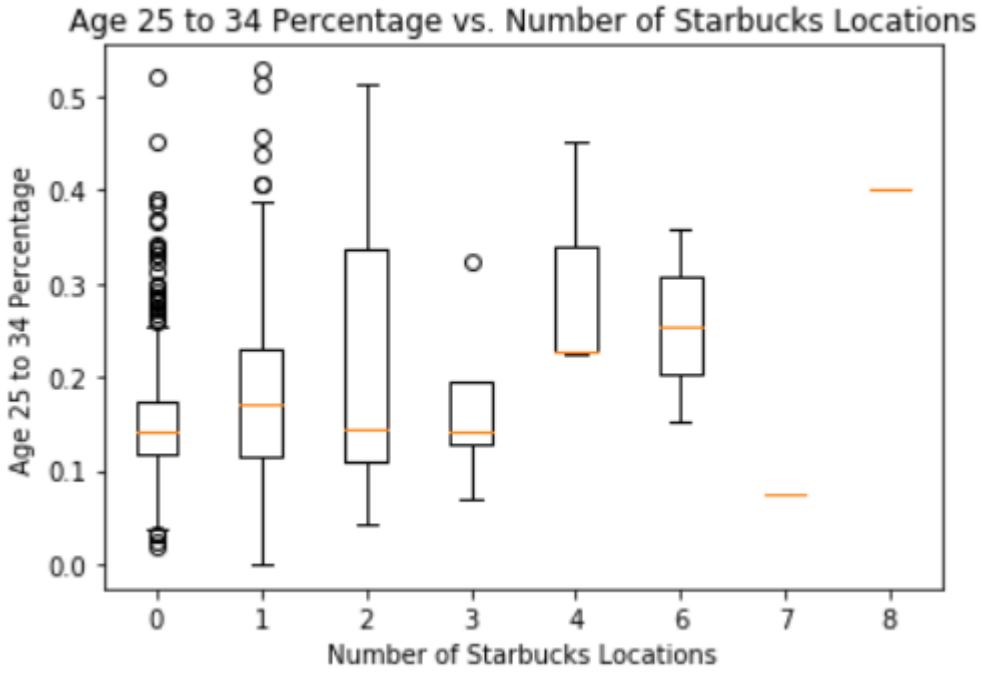
Dallas County Data Correlation

- All US Census data metrics collected do not strongly correlate to the number of Starbucks locations in a census tract
- All data metrics with an absolute Pearson correlation greater than 0.1 were selected for model development

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

Dallas County Data Correlation

- The five metrics that were used for model development were investigated further
- This graph shows an example of how each metric was investigated

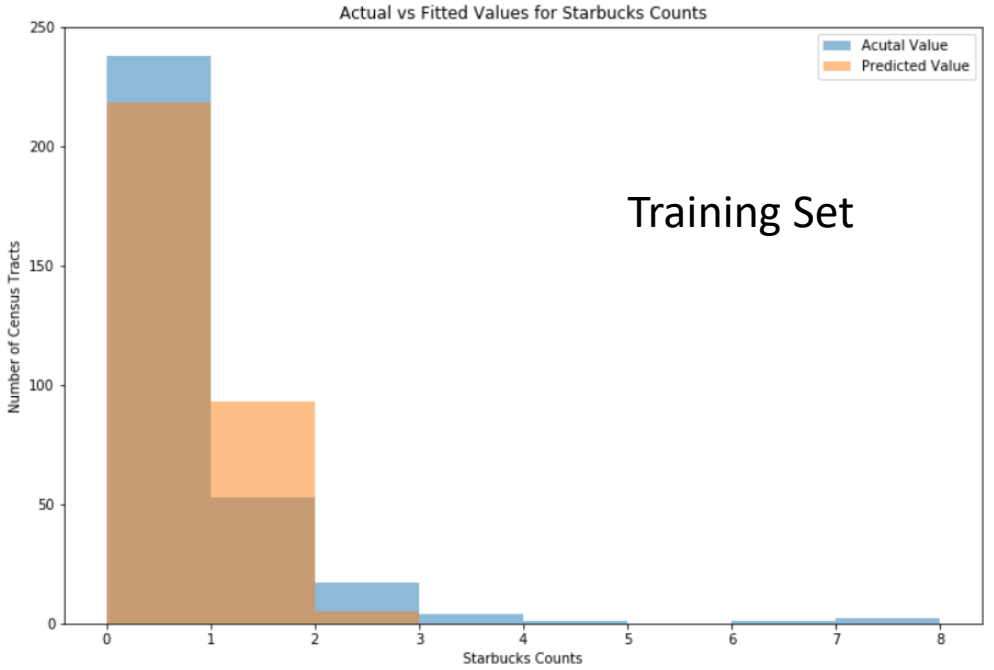


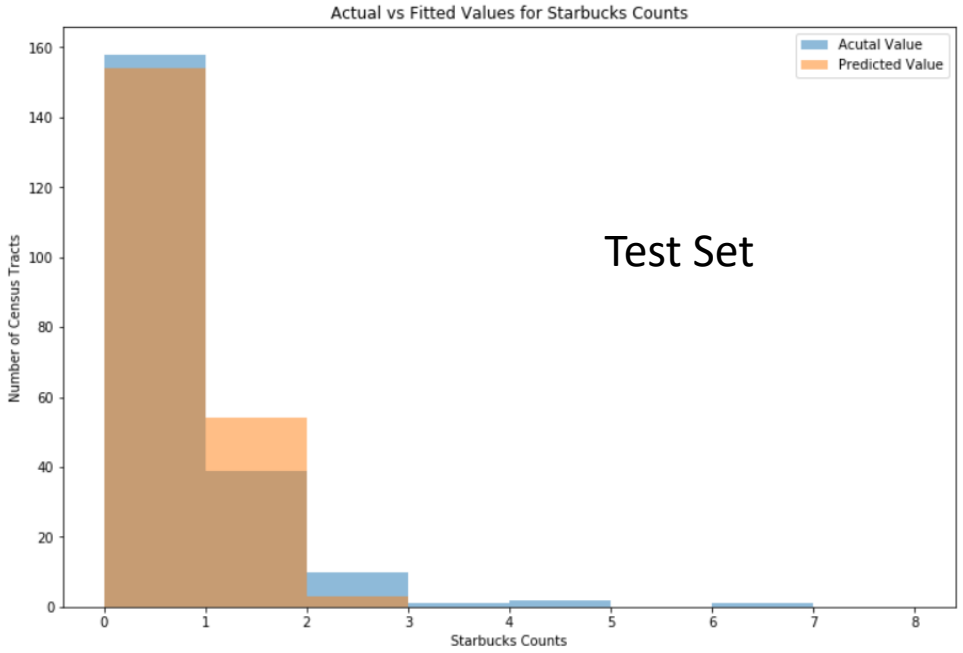
Dallas County Data Correlation - Findings

- All median values generally increase or decrease the data metrics investigated if is has a positive or negative Pearson 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.
- Percentage of people aged 25 to 34 and Asian Percentage trend in median values deviate widely from the overall trend for census tracts with 6 or more Starbucks locations
- Each data metric has 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.

Multiple Linear Regression Performance

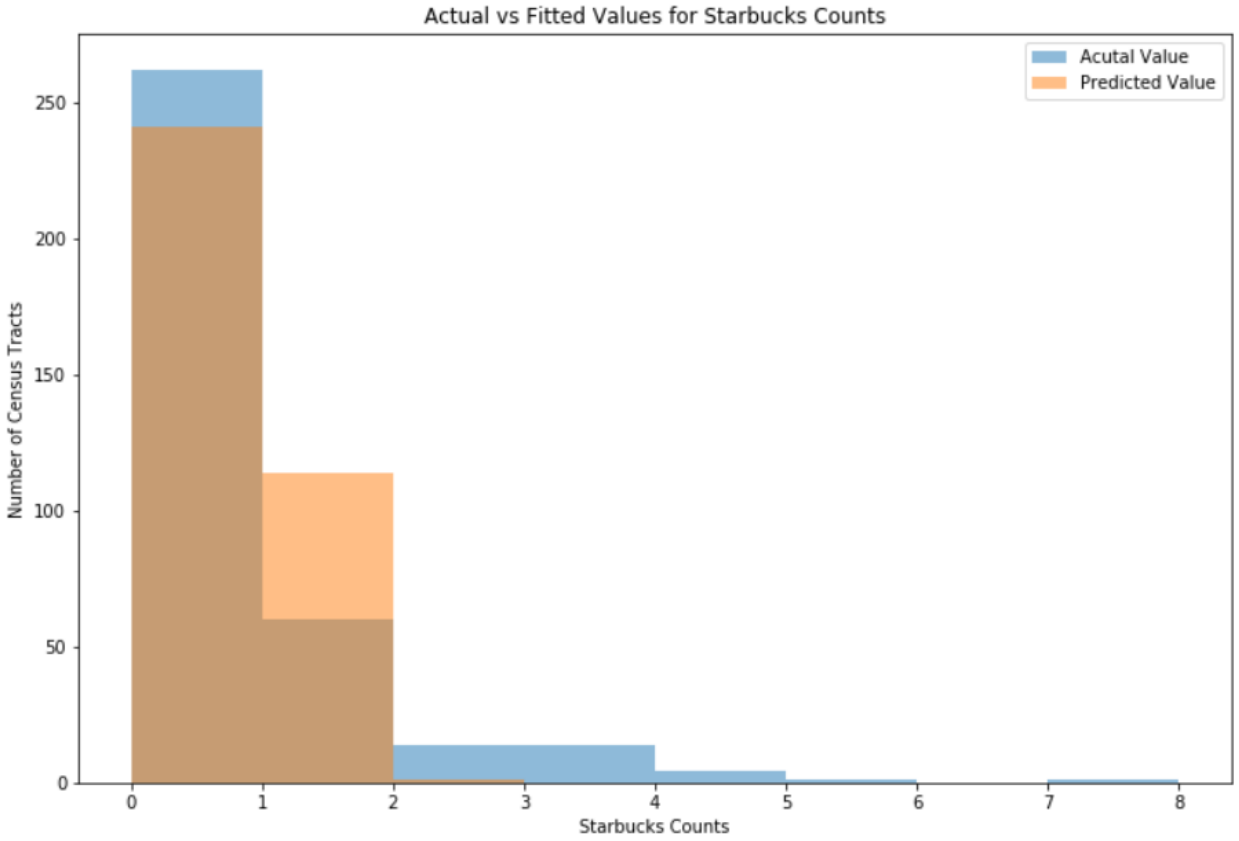
- A multiple linear regression was used since appears multiple factors play a role in predicting the number of Starbucks
- The model is not able to predict the number of Starbucks locations in census tract accurately if it has 2 or more locations





Multiple Linear Regression Performance – Tarrant County

- Same problems seen in Dallas county model predictions



Overall Model Performance

- Models are able to predict the number of Starbucks locations correctly a majority of the time.
- Model under performs for census tracts with more than 2 Starbucks locations.
- Right the model would be able to very accurately predict if a census tract should contain a Starbucks or not

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

Conclusion

- This model shows promise in being able to predict the number of Starbucks locations in a census tract.
- More data metrics could be included outside of the ones gathered for this project to improve the model related to labor statistics.
- In the Dallas/Fort Worth area, it appears Starbucks locations are being targeted to predominant young, white, and affluent individuals.