2017 NHTS Data Competition

Improving Commute Time with Data Analysis

Authors: Walter Yu, P.E. and Bonny Nyaga, P.E.

Abstract

Travel behavior by US citizens has a great impact on the country's economy, productivity, infrastructure and environment. Different modes of travel for all purposes, looked at individually or in a combination, have a direct relationship on how travel impacts these factors. In this study we evaluate the National Household Travel Survey (NHTS) national transportation dataset[1] for 3 major metro areas specifically focusing on the impact at the average commute times has a large impact on the economy, productivity, infrastructure and environment. Longer commute times cause lost wages for workers with longer commute times, additional wearing of highway infrastructure and environmental impacts. Specifically, this study evaluates commute patterns with the NHTS dataset[1] and whether public transportation or additional transportation planning could reduce commute times based on data analysis.

Introduction

This study outlines the data, methods and results used to identify commute patterns. Specifically, it seeks to answer the following questions:

1. Which are the census divisions with the most trips per household[2]
2. What are the average commute distance and time within those divisions?
3. Could public transportation or transportation planning reduce commute times?
4. What are some recommendations for improving commute times based on demographic data[3]?

NHTS Data Analysis

This study analyzes the households, trips and vehicles tables of the NHTS dataset to evaluate commute trends by census division. Specifically, the tables were analyzed to evaluate average commute distance and time.

Tools and Process

The tools and process listed below were used to analyze data and provide recommendations:

1. Jupyter Notebook - Exploratory data analysis and visualization were completed using this notebook.
2. Python Modules - The modules listed below will need to be installed in order to run this notebook:

Pandas - Data Analysis
NumPy and SciPy - Scientific Calculations
Matplotlib and Seaborn - Data Visualization

[1] NHTS Datasets Website
[2] Section 5.1, 2017 NHTS User Guide
[3] City-Data.com Website

Data Cleaning and Preparation

Data cleaning was completed prior to analysis; as a result, the datasets were cleaned to minimize the impact of outlier, missing or repeated values as follows:

1. Replaced empty and missing values since they may cause errors during analysis.
2. Removed negative values since they may skew summary statistics and results.
3. Removed outlier values by removing values greater or less than 3 standard deviations from the mean since they will skew results. Specifically, outlier values will skew summary statistics such as the mean, median and standard deviation.

After data cleaning, the household, trip and vehicle tables were sorted by census district to begin the analysis.

Household Count by Division

The weighted household count was calculated to identify the following:

1. Which divisions have the most households?
2. How do household count differ between divisions with and without subway systems?
3. Are there any other noticeable trends based on chart plot?

As a result, household count is shown Appendix A (Figure A.1) with the following observations:

1. Total Count: The Pacific and Atlantic divisions had the highest count while the Central divisions; this trend is intuitive given that many coastal states have higher populations than those in the midwest.
2. Subway Systems: The Pacific and Atlantic had more households with access to a subway system then those which did not have access; however, the Central divisions had more households without access to a subway system than those which did have access.
3. Subway Access: The trend observed above implies that the Pacific and Atlantic regions may have more urban areas with more households centered around transportation hubs instead of more rural or equal distribution of the population as in the Central divisions.
4. In general, household count is highest in the Pacific and Atlantic divisions with a higher percentage of households having access to a subway system whereas household count is lower the Central divisions with lower percentage of households having access to a subway system.

Average Weighted Trips per Household by Census Divisions

The average trips per household by division were calculated to identify commute trends; as a result, the household table was sorted for census divisions as listed below and are shown in Appendix A (Figure A.2):

1. Selected census divisions with the 5 highest total of weighted households per the NHTS codebook.
2. Divisions were further sorted by ones with and without subway system.
3. The households identified in these divisions were then matched with trips.
4. The weighted trip values within each division were totaled, then divided by total households.
5. Result was average weighted trips per household by division.

The divisions identified with highest total weight households were as follows:

1. Mid-Atlantic > 1M with Subway
2. Mid-Atlantic > 1M w/o Subway
3. East North Central > 1M with Subway
4. East North Central > 1M w/o Subway
5. South Atlantic > 1M with Subway
6. South Atlantic > 1M w/o Subway
7. East South Central > 1M with Subway
8. East South Central > 1M w/o Subway
9. Pacific > 1M with Subway
10. Pacific > 1M w/o Subway

The results were plotted in the chart below with the following observations:

1. Trip Count: Total trip count was higher in all divisions with access to a subway system than those without access. This trend implies that households with access to a subway system tend to be more densely populated and result in higher trip count.
2. Trip Count Distribution: Trip count were distributed evenly between divisions despite household count differences which may be due to weight ranking or household formation (larger or smaller size).

Annual Miles per Household by Census Divisions

The average miles per household by division were calculated to identify commute trends; as a result, the vehicle table was sorted as follows:

1. The households within each selected division were matched with vehicle values.
2. The annual miles for each joined value were totaled, then divided by total households.
3. Result was average annual miles per household by division.

Results are shown in Appendix A (Figure A.3) with the following observations:

1. Total Miles: Total miles traveled were lower in divisions with access to a subway system than those without access to one (except for the South-Atlantic which had slightly higher miles traveled). This implies that more residents may be taking the subway and reducing the number of miles traveled.
2. Pacific/Atlantic Divisions: These divisions showed the largest decrease in miles traveled with access to a subway system than those without access. This implies that subway systems in more urban/densely populated divisions may reduce the total number of miles traveled and minimize impacts to the environment and infrastructure.

Commute Time per Household by Census Divisions

The average commute time per household by division were calculated to identify commute trends; as a result, the trips table was sorted as follows:

1. The households within each selected division were matched with trip values.
2. The commute time for each joined value were totaled, then divided by total households.
3. Result was average commute time per household by division.

Results are shown in Appendix A (Figure A.4) with the following observations:

1. Average Commute: Higher commute times were observed in all divisions with access to a subway system which may be a result of them being located in urban areas with higher traffic congestion. In addition, this observation implies that subway systems are having an impact in reducing commute time within areas that already have high traffic congestion.
2. Commute Time Distribution: Average commute time was slightly higher in the Pacific/Atlantic divisions and lower in the Central regions. This observation implies that the Pacific/Atlantic divisions have more urban areas with shorter distances, whereas the Central divisions have more rural areas with longer distances.

Comparison with Demographic Data

Demographic data visualizations were analyzed for major metropolitan areas with subway access to verify commuting trends; as a result, commuting trends were identified which verify observations made in this study as follows:

1. San Francisco Bay Area (Appendix B, Figure B.1): Average commute time is lower within city limits; however, commute time increase in the North, South and East Bay regions located outside of San Francisco.
2. Southern California, Los Angeles (Appendix B, Figure B.2): Average commute time is lower within city limits; however, commute time increased in the North and Eastern regions located outside of San Francisco.
3. Tri-State Area, New York (Appendix B, Figure B.3): Average commute time is low for most of the tri-state area which validates the conjecture that commute times are typically lower due to high population density and subway access.

The data visualizations validate the conjecture that commute times are typically lower due to high population density and subway access. However, the analysis only establish an associations between commute time, population density and subway access; it does establish whether subway access reduces commute time.

Recommendations

Recommendations to improve commuting trends are based on the initial data analysis and comparison with demographic data and as follows:

1. Increase Population Density: Implement good urban planning practices by increasing population density to create positive effects such as lower annual miles traveled and commute times.
2. Provide Subway Systems in Urban Areas: Since most subway systems are developed based on usage and ridership, then continue to develop them in areas that have sufficient urban density.
3. Spoke and Hub Model: Some metropolitan areas have established other cities, such as Oakland for San Francisco, which can serve as transportation spokes to the hub. As a result, this model may be applied to urban planning in developing areas.

Next Steps

This study included a limited scope to briefly clean, analyze and visualize the dataset; however given more time, additional analysis should be completed as follows:

1. Additional Data Attributes: Analyze additional factors from the NHTS dataset; the NHTS codebook contains many more factors which are work additional analysis.
2. Establish Correlation: Compare different factors and establish numeric correlation values; although it would not prove causation, the analysis would formally establish correlation.
3. Predictive Analysis: Use machine learning to develop a predictive model to identify future commuting trends based on current NHTS data.

Conclusion

The study provided exploratory data analysis, visualization and recommendations based on the NHTS dataset; in general, subway access are typically associated with higher population areas and lower annual travel miles. As a result of subway systems being located within more densely populated areas, these areas also indicated longer commute times and higher number of annual trips.

The primary recommendation is to continue good urban planning practices by increasing population density to create positive effects such as lower annual miles traveled and commute times. Next steps beyond this study are analyze other data attributes within the NHTS dataset, establish correlation between these attributes and develop a predictive model with machine learning.

NHTS Data Challenge - Appendix A

Figure A.1 - Weighted Household Count by Census Division and Subway System Access

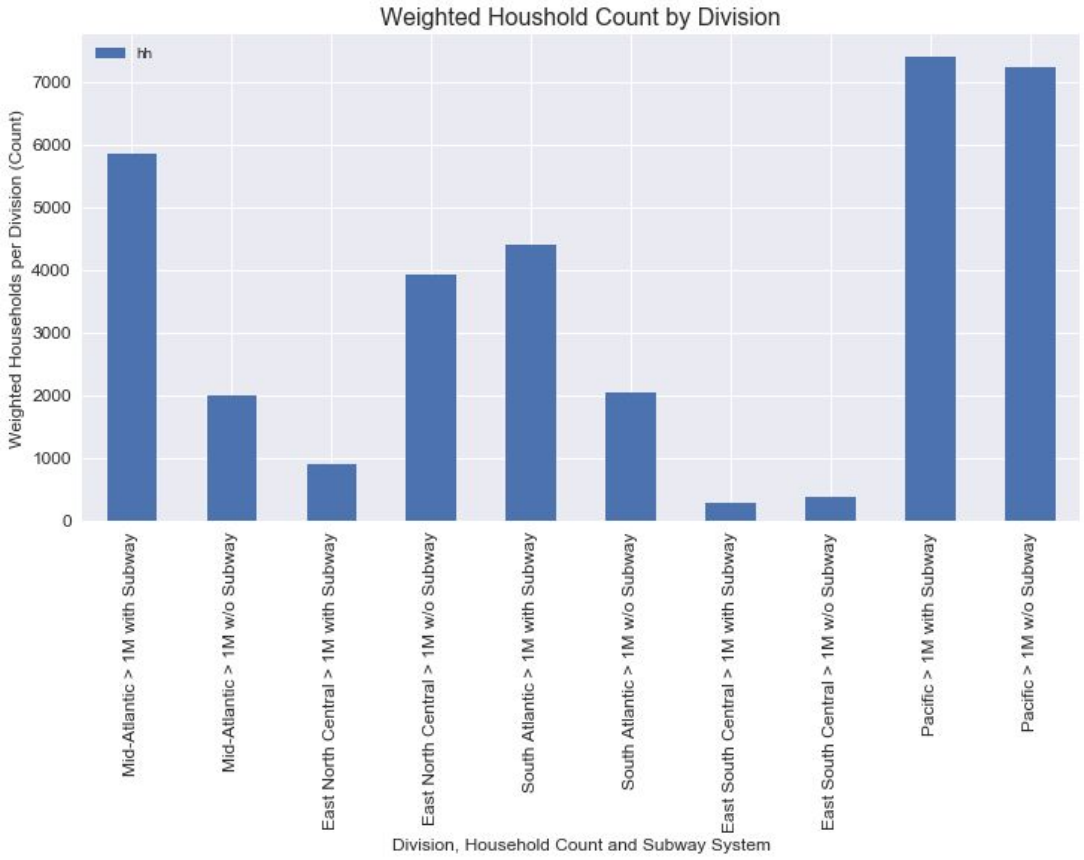


Figure A.2 - Annual Weighted Trip Count by Census Division and Subway System Access

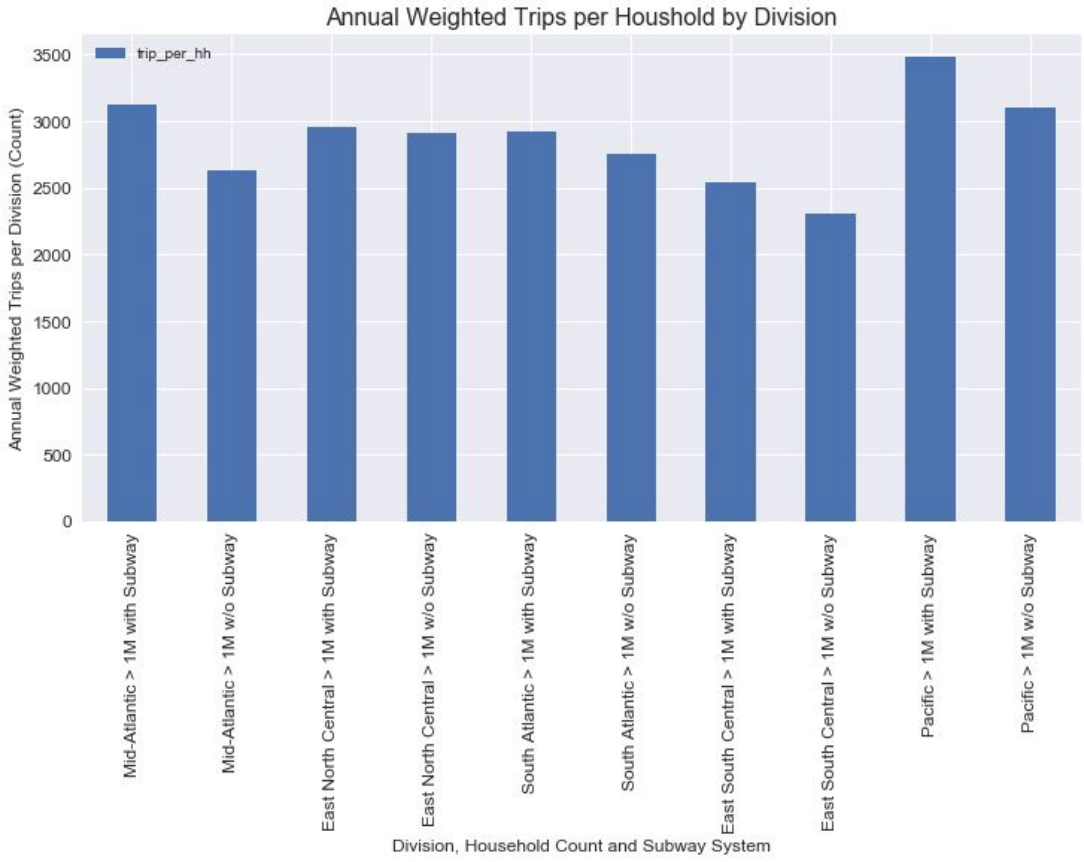


Figure A.3 - Average Commute Time by Census Division and Subway System Access

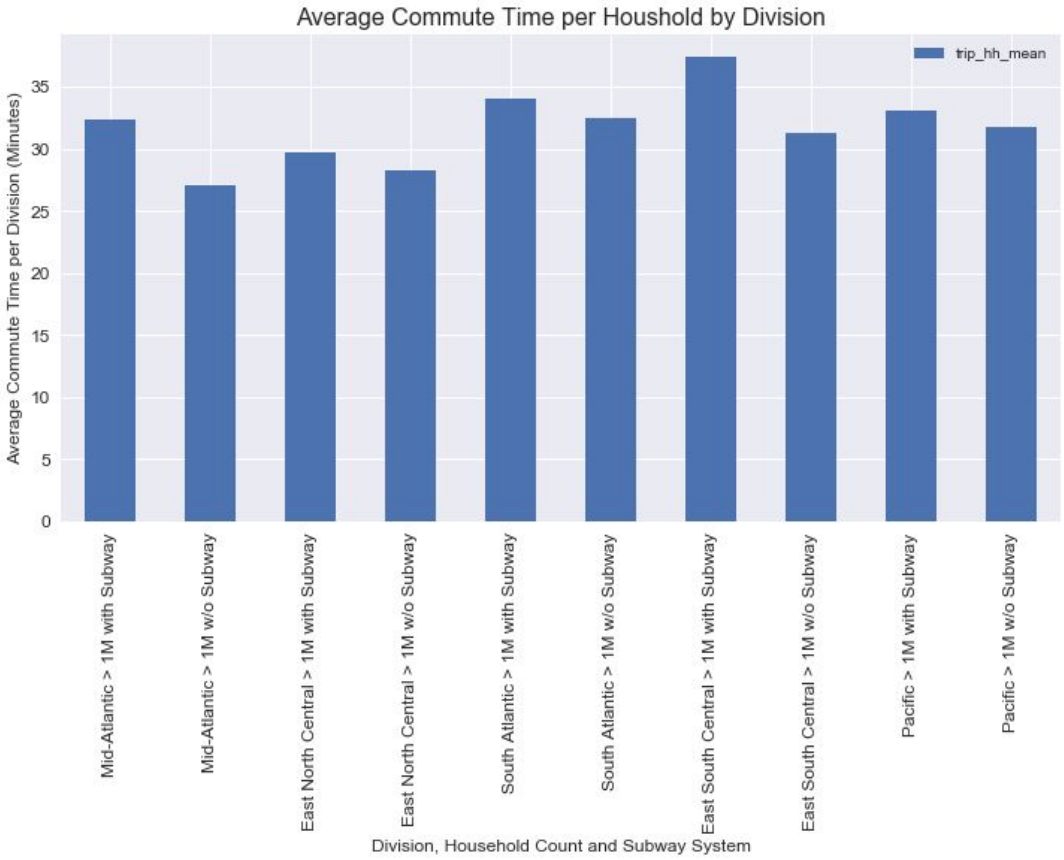
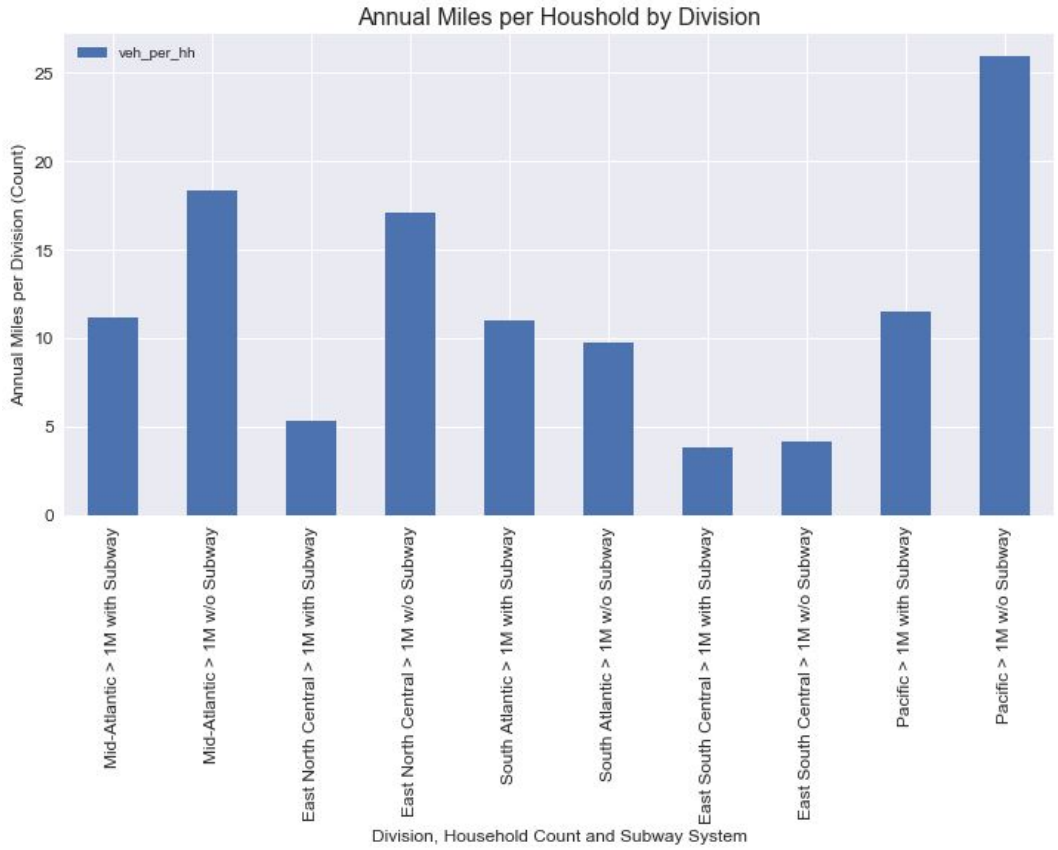
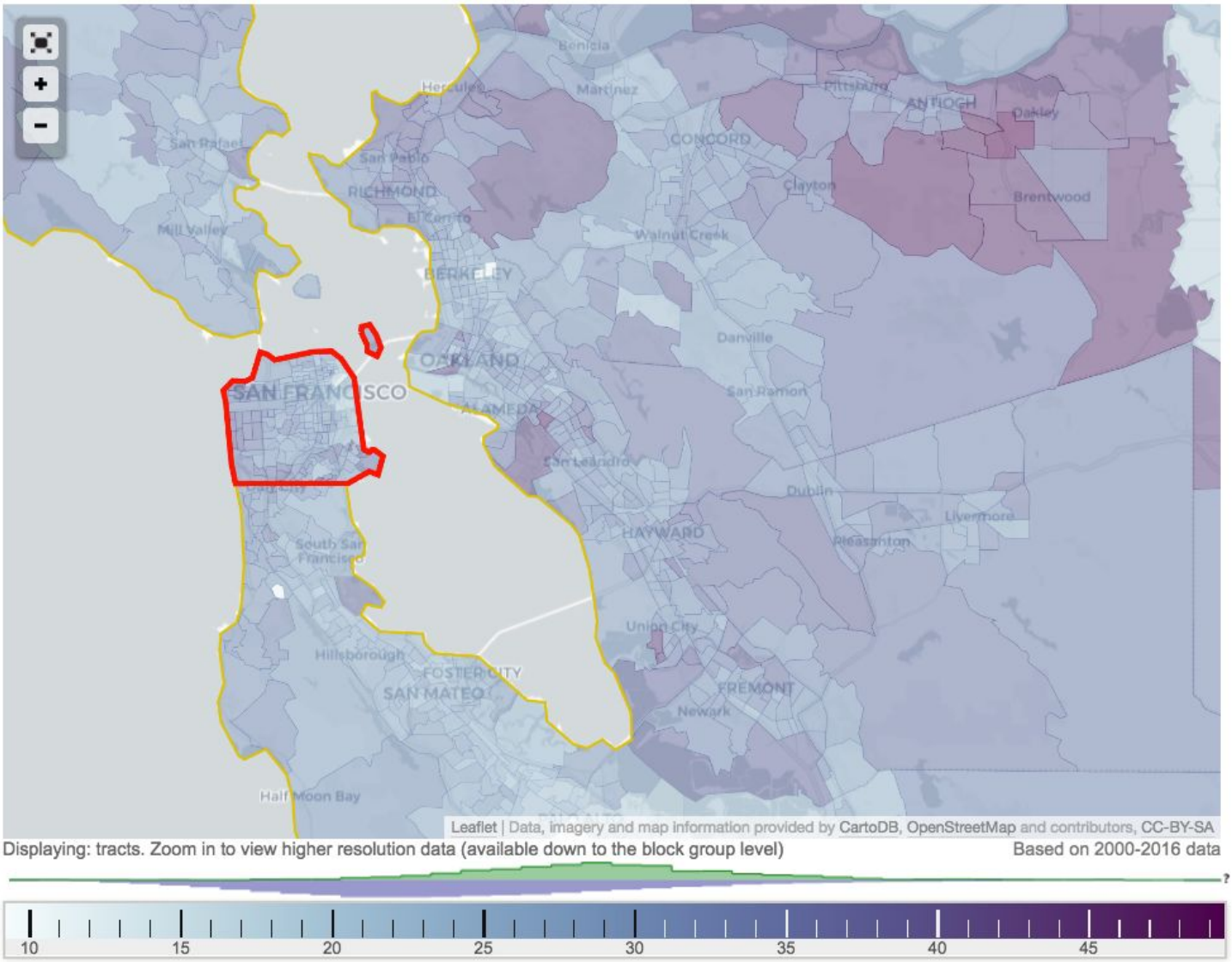


Figure A.4 - Annual Miles by Census Division and Subway System Access



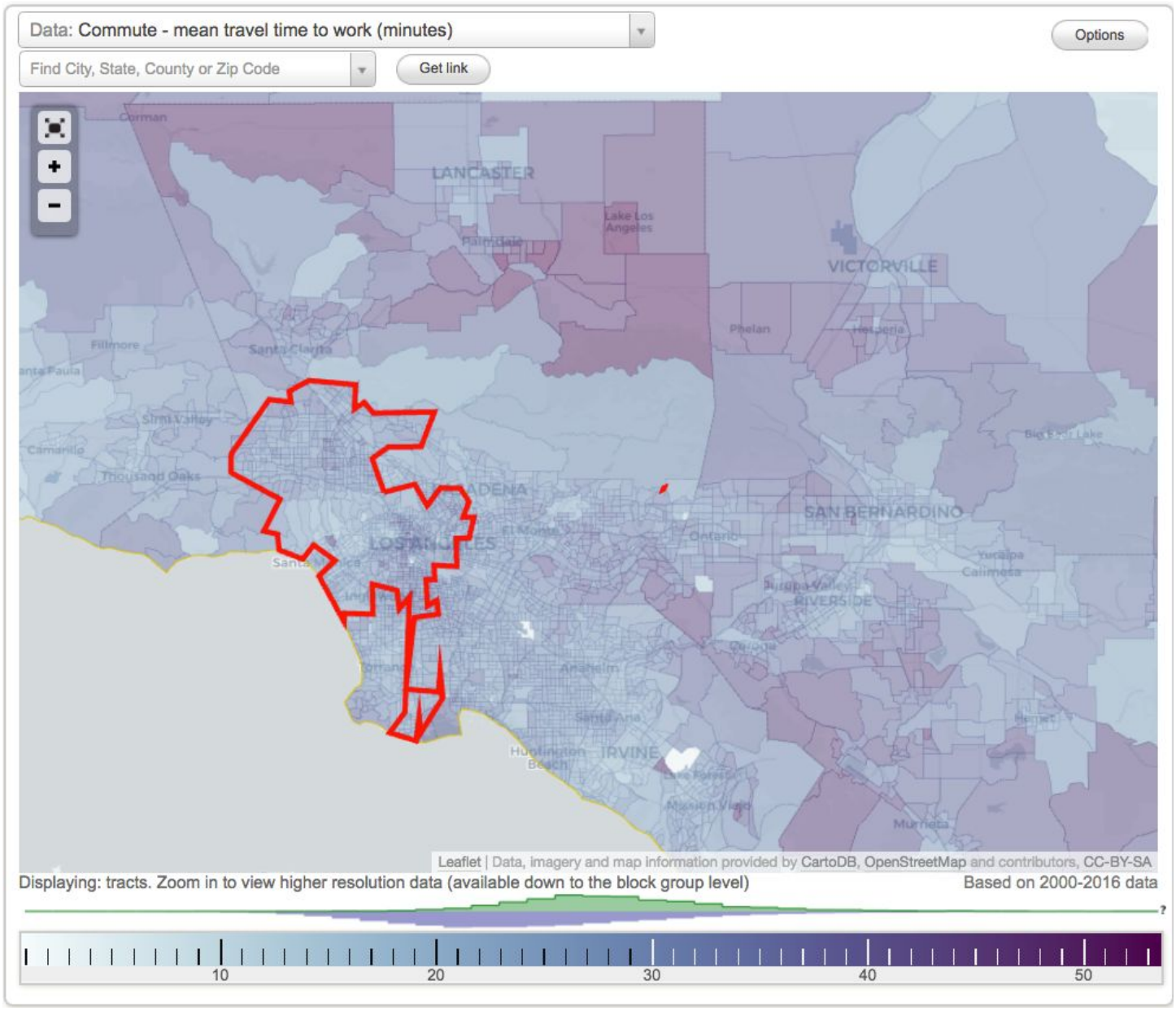
NHTS Data Challenge - Appendix B

Figure B.1 - Commute Time - San Francisco, CA
(White = Short Commute & Purple = Longer Commute)



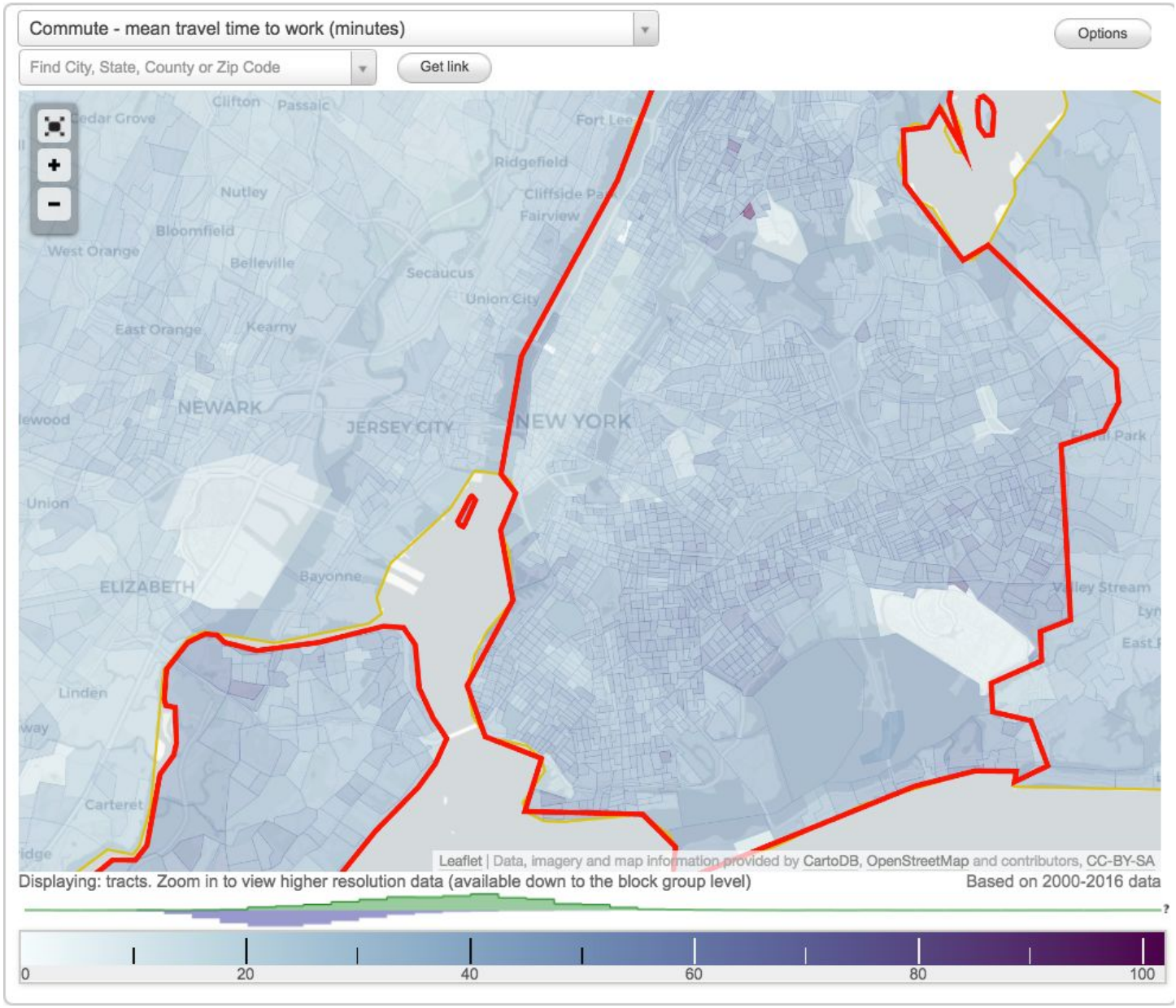
Reference: City-Data.com

Figure B.2 - Commute Time - Los Angeles, CA
(White = Short Commute & Purple = Longer Commute)



Reference: [City-Data.com]

Figure B.3 - Commute Time - New York, NY
(White = Short Commute & Purple = Longer Commute)



Reference: [City-Data.com]