Final Project

Reducing Commute Time with Machine Learning and Graph Analysis

Yu, Walter

CSCI E-63 Big Data Analytics

Harvard University Extension School

Prof. Zoran B. Djordjević

Introduction

- Analyze commuter data (NHTS) to identify trends
- Goal to reduce U.S. commute time
- Machine learning (ML) and graph analysis
- Github: https://github.com/walteryu/e63-final

U.S. Department of Transportation
Federal Highway Administration

Introduction

- National Household Transportation Survey (NHTS) dataset: https://nhts.ornl.gov/
- Based on 2017 NHTS Data Challenge contest entry
- Final project extends analysis with ML and graph analysis
- National survey with ~1M records; ~700MB total

Data Analysis

- Data cleaning, exploration and visualization
- Calculate summary statistics
- Analyze with ML and graph analysis
- Document results in Jupyter Notebook

CSCI E-63 Big Data Analytics - Final Project (Fall 2018)

Reducing Commute Time with Machine Learning and Graph Analysis

Author: Walter Yu, Graduate Degree Candidate

Abstract

The average commute time within each U.S. census division has a large impact on its economy, productivity, infrastructure and environment. Longer commute times can lead to lost wages for workers, additional wearing of highway infrastructure and environmental impacts. As a result, this study evaluates U.S. commuter patterns with the National Household Transportation Survey (NHTS) dataset[1] provided by the Federal Highway Administration (FHWA) and whether public transportation or additional transportation planning could reduce commute times based on data analysis.

ML and Graph Analysis

- Evaluate ML algorithm performance
- Identify key factors with feature importance
- Develop graph and calculate out-degree relationships

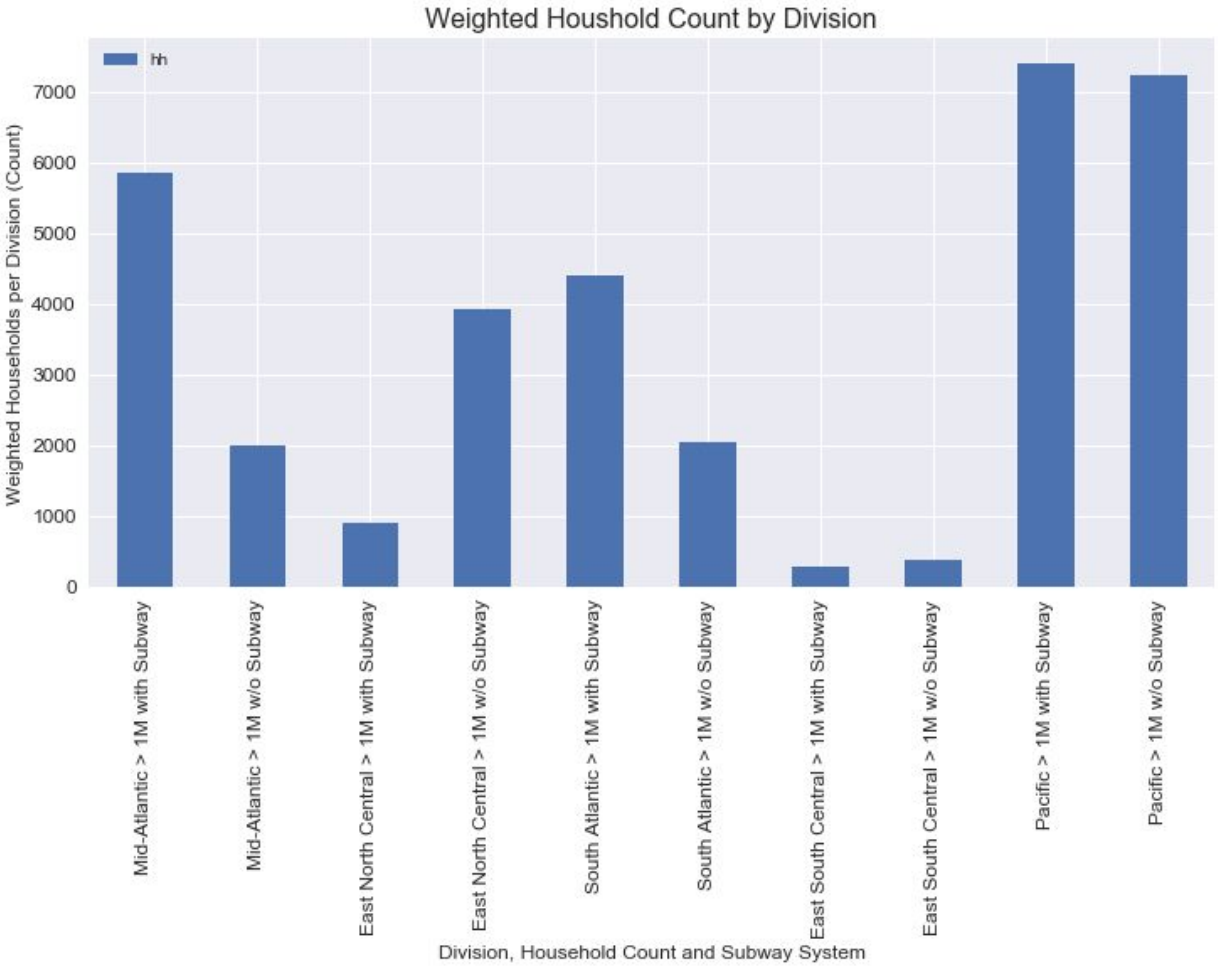
Example ML Output Feature Importance:

prediction	BIKE	BUS	CAR	PARA	PLACE	PRICE	PTRANS	features
1.8759342084160502	5	4	1	5	2	1	4	[5.0,4.0,1.0,5.0,...
2.1851742331474937	5	5	1	5	3	5	5	[5.0,5.0,1.0,5.0,...
2.271102893790943	4	4	1	5	5	5	5	[4.0,4.0,1.0,5.0,...
2.078192432037876	4	5	2	5	3	5	3	[4.0,5.0,2.0,5.0,...
2.348142330357143	2	5	1	5	3	4	5	[2.0,5.0,1.0,5.0,...
2.400746222594453	4	5	1	5	4	3	5	[4.0,5.0,1.0,5.0,...
2.1127218257500227	5	5	1	5	4	4	4	[5.0,5.0,1.0,5.0,...
0.3279613358204647	-9	1	-1	-9	1	1	1	[-9.0,1.0,-1.0,-9...
2.3522649979636734	3	5	1	5	3	2	4	[3.0,5.0,1.0,5.0,...
0.38146184482344175	5	1	4	5	3	5	1	[5.0,1.0,4.0,5.0,...

Example Graph Analysis Output:

```
Total Number of Households: 129696
Total Number of Relationships in Graph: 923572
Total Number of Relationships in Original Data: 923572
```

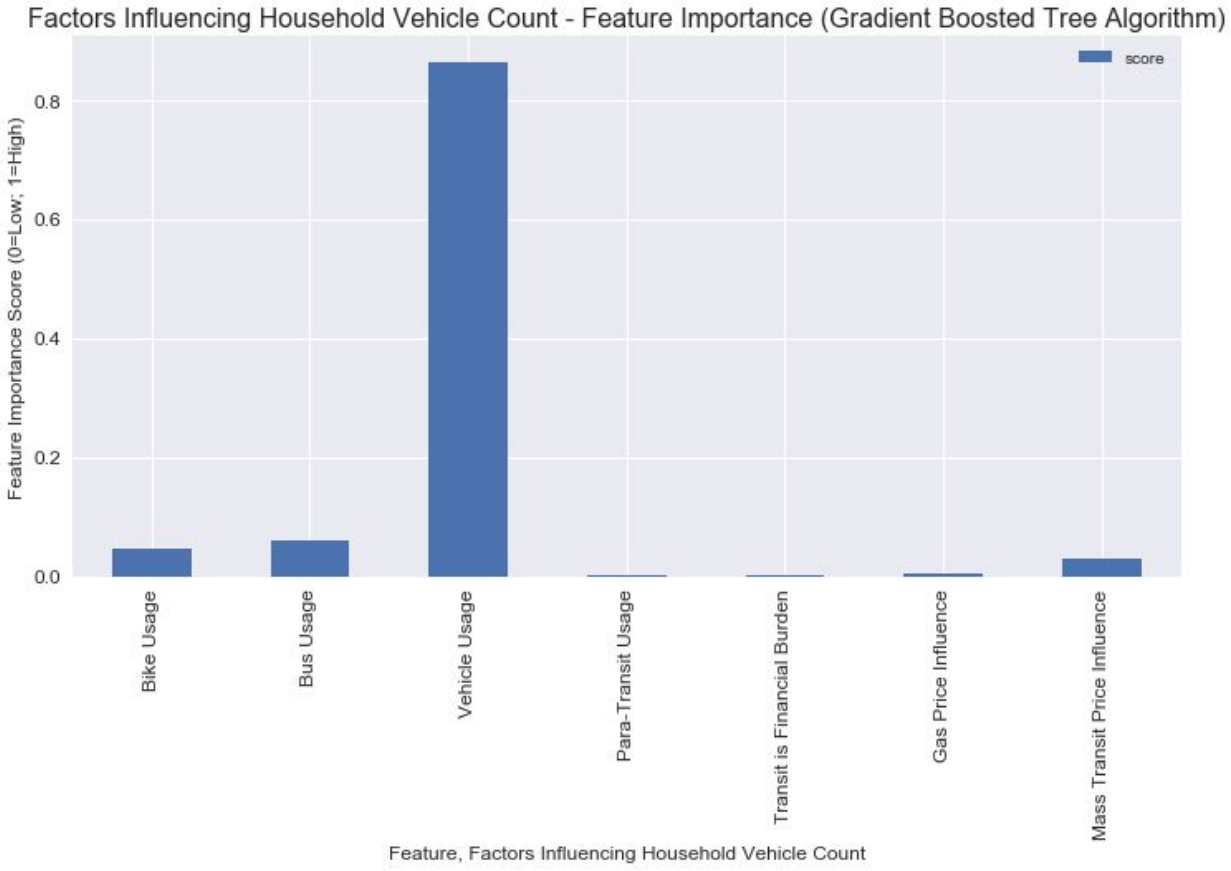
Household Data - Summary Statistics



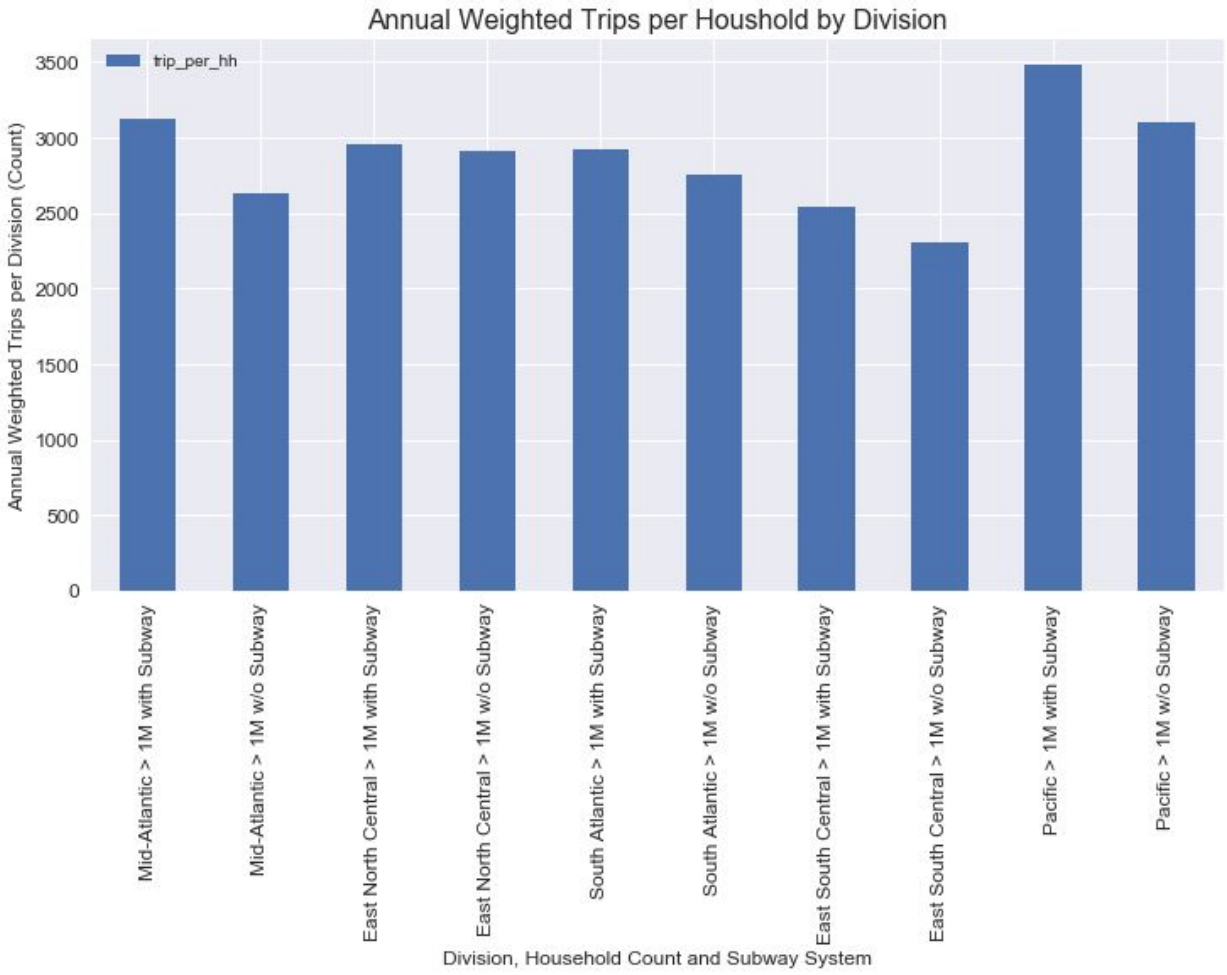
- Calculate household count by division
- Separate by access to mass transit
- Areas with mass transit typically have higher household count

Household Data - Feature Importance

- Vehicle usage most impacts total count
- Other features have low significance
- Mass transit feature has low significance with vehicle usage



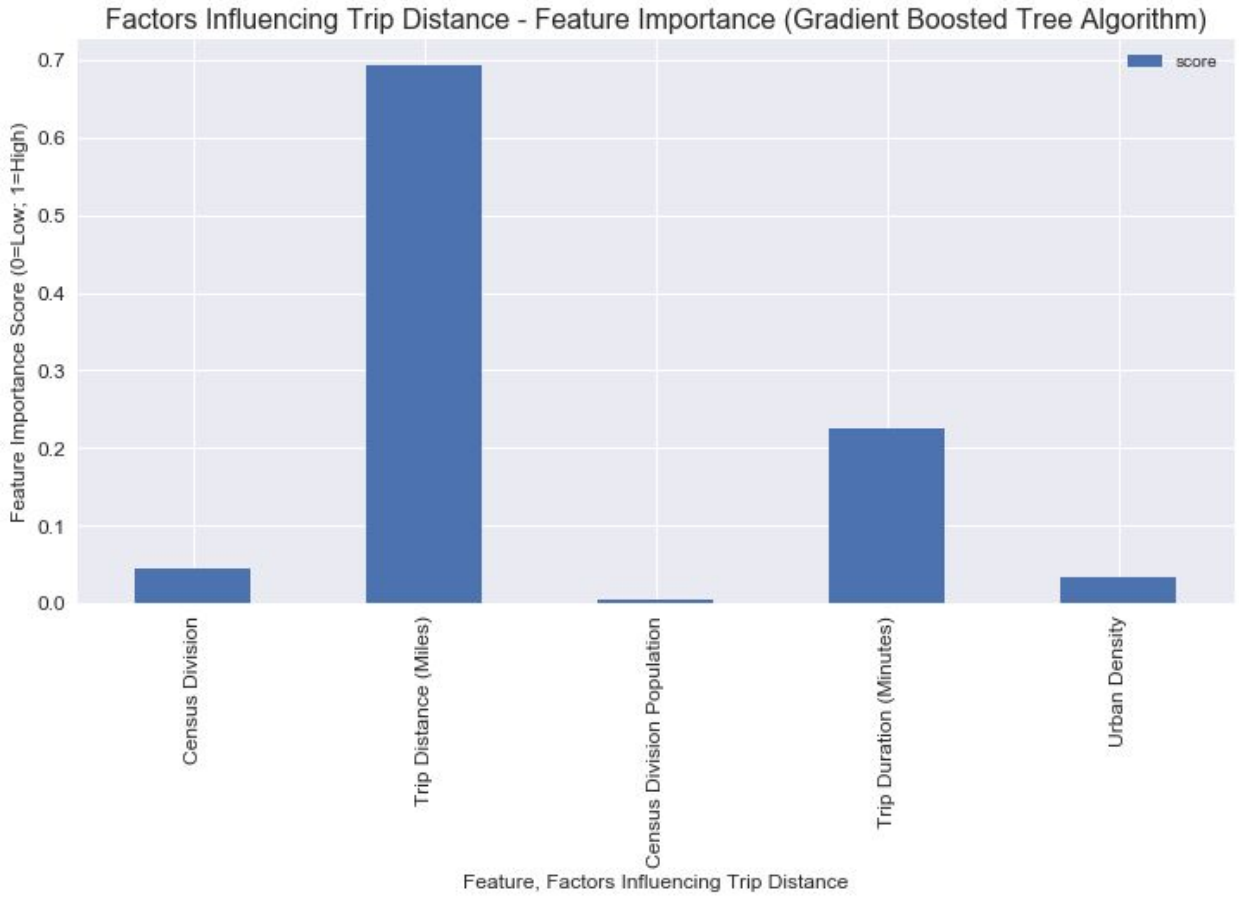
Trip Data - Summary Statistics



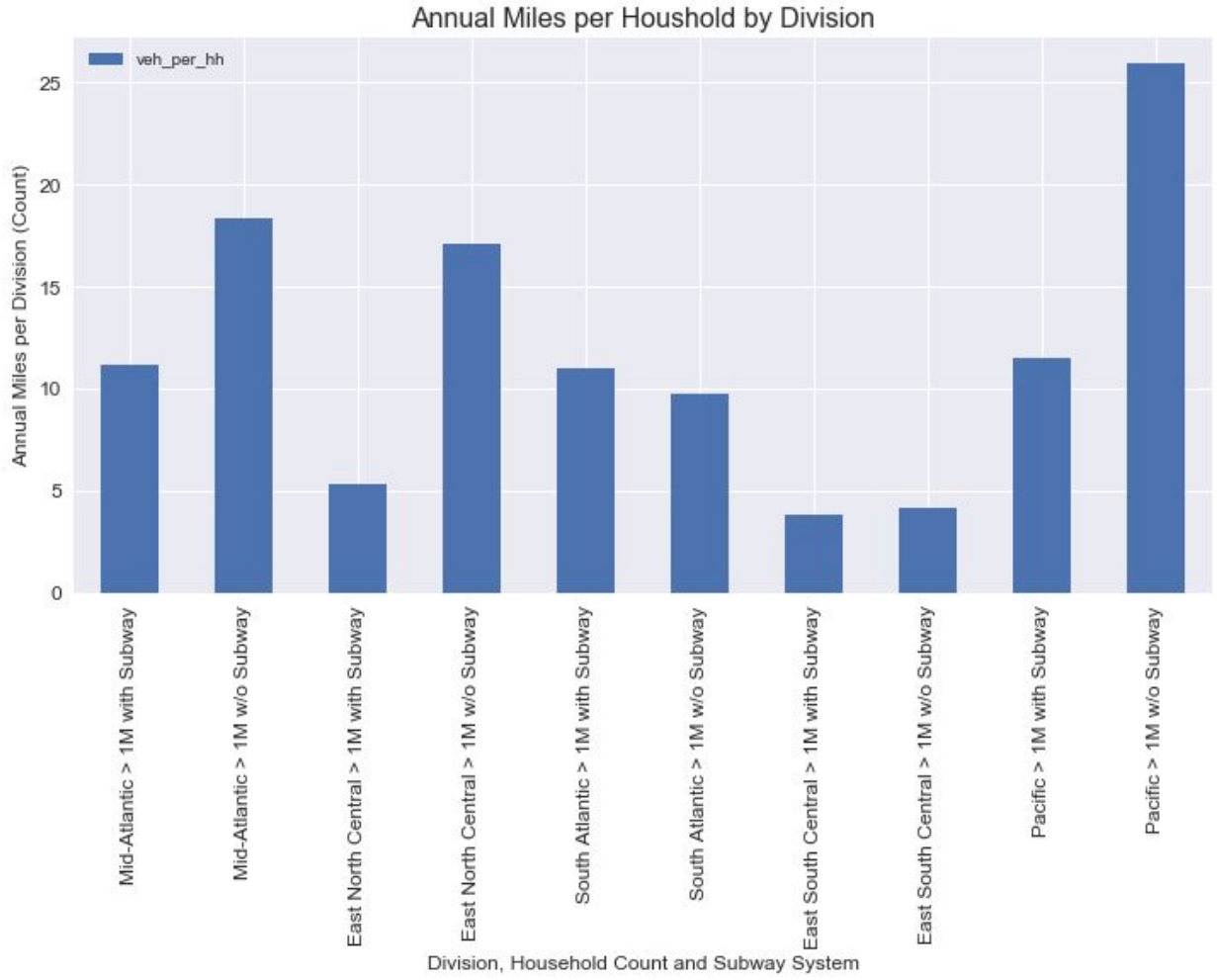
- Calculate trip count by division
- Separate by access to mass transit
- Areas with mass transit typically have higher trip count

Trip Data - Feature Importance

- Trip distance most impacts vehicle miles per trip
- Trip duration is also significant feature
- Other features have low significance

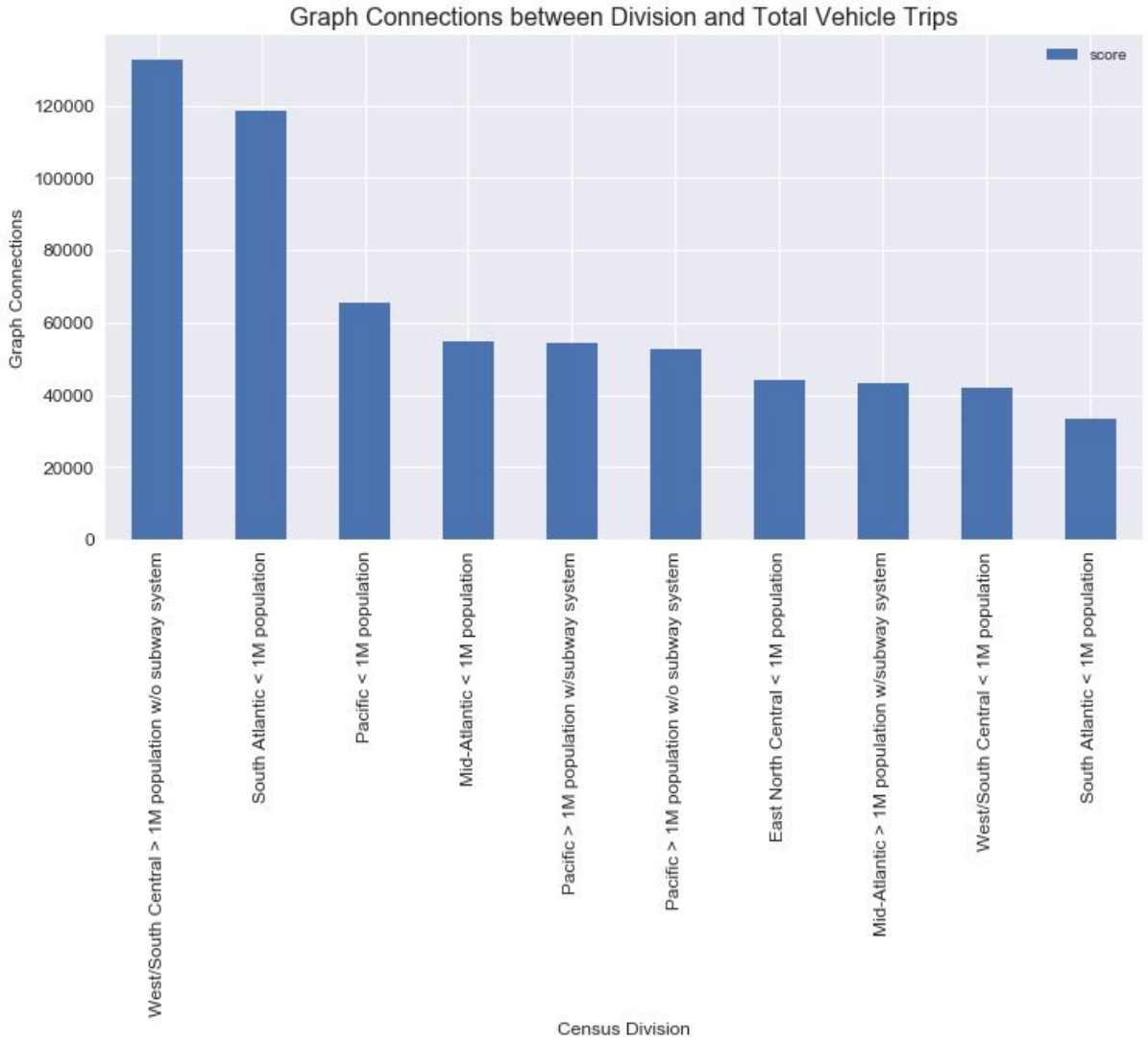


Annual Miles Driven - Summary Statistics



- Calculate annual miles driven by division
- Separate by access to mass transit
- Areas without mass transit typically have higher miles driven

Graph Analysis - Out-Degree Relationship Count



- Calculate out-degree relationships between division and total trips
- Areas without mass transit and lower population

Observations

Photo Reference: http://bit.ly/2DZ2avA

- Mass transit does not appear to directly influence driving behavior; however, trip distance/length has high significance
- Urban areas have higher population so appear to have higher total trips but shorter distances; rural areas have lower population so have lower total trips but longer distances
- Households with higher income appear to have higher vehicle count and usage; trip destination also has significance

Recommendations

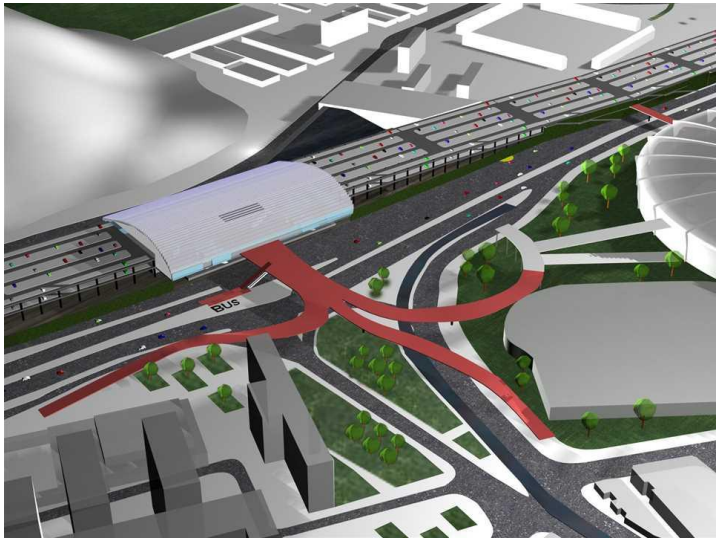


Photo Reference: http://bit.ly/2E2N7Rh

- Public education is recommended to raise awareness
- Results may be useful to urban planners
- Encourage households to take action by living near urban areas instead of relying on urban planning to solve problem
- Mass transit alone is unlikely to reduce vehicle usage; actions to change driving behavior are recommended

YouTube URLs, Last Page

- Two minute (short):
- 15 minutes (long):