Capstone Project - U.S. Census Bureau, Center for Adaptive Design

Title of the Project: Combining Traditional and New Data Sources in the Production of Official Statistics

Timing: Spring 2016

Client:
The Center for Adaptive Design at the U.S. Census Bureau

Issue Definition:
The U.S. Census Bureau is the largest statistical agency in the U.S. Government. In addition to the once-a-decade population count, the Census Bureau collects data on the U.S. economy, healthcare, poverty, education, crime, and business innovation, among others. The predominant method used for collecting data by the Census Bureau is the traditional sample survey administered to households, businesses and government entities. Response rates to sample surveys, however, are declining rapidly while survey costs are climbing at an unsustainable rate. Further, the Census Bureau is highly interested in reducing public burden (the effort and time spent responding to surveys). This combination of challenges is driving the Census Bureau to explore alternative data collection methods, including passive approaches using Big Data techniques. The overall issue this project will help address is how to combine the data from new sources with data collected using the traditional approaches. The goal is to increase efficiency and timeliness, reduce cost and burden, and maintain data quality.

Scope of Work:
A particular line of survey questions that have proven to be problematic are those surrounding the topic of people's travel and commuting habits. With the availability of traffic and commuting data from public and private sources, as well as crowd-sourced applications (such as Waze) the Census Bureau should be able to access data that may enable a reduction in travel and commuting questions in our sample surveys (such as the American Community Survey). This project would research possible data sources to infer selected travel and commuting statistics in the American Community Survey, suggest methods and tactics for accessing and interpreting those sources, and explore approaches for verifying the quality of these sources with the goal over time of incorporating them into official statistics. Possible data sources include roadway traffic counts, travel time, parking, public transportation, etc. Work also involves devising methods to test new data source quality against established survey data benchmarks, as well as suggesting a viable long-term plan for transitioning to new data sources.

Expected Deliverables:

- Research and recommendations on possible new data sources to complement/replace targeted questions on existing sample surveys such as the American Community Survey

- Literature review of existing work on incorporating new data sources into travel survey
- Documentation of methods, technology, cost-benefits, and tactics for accessing and interpreting new data sources
- Suggested approaches for verifying the quality of new data sources
- Documentation of a possible long-term plan for transitioning to new data sources

Skills Required:
Data Analytics/Data Science
Survey Methodology
Information Technology/System Development

Advisory Board:
Michael Thieme, Chief of the Center for Adaptive Design, U.S. Census Bureau
Benjamin Reist, Assistant Center Chief, U.S. Census Bureau, Center for Adaptive Design, U.S. Census Bureau
Stephanie Coffey, Mathematical Statistician, Center for Adaptive Design, U.S. Census Bureau

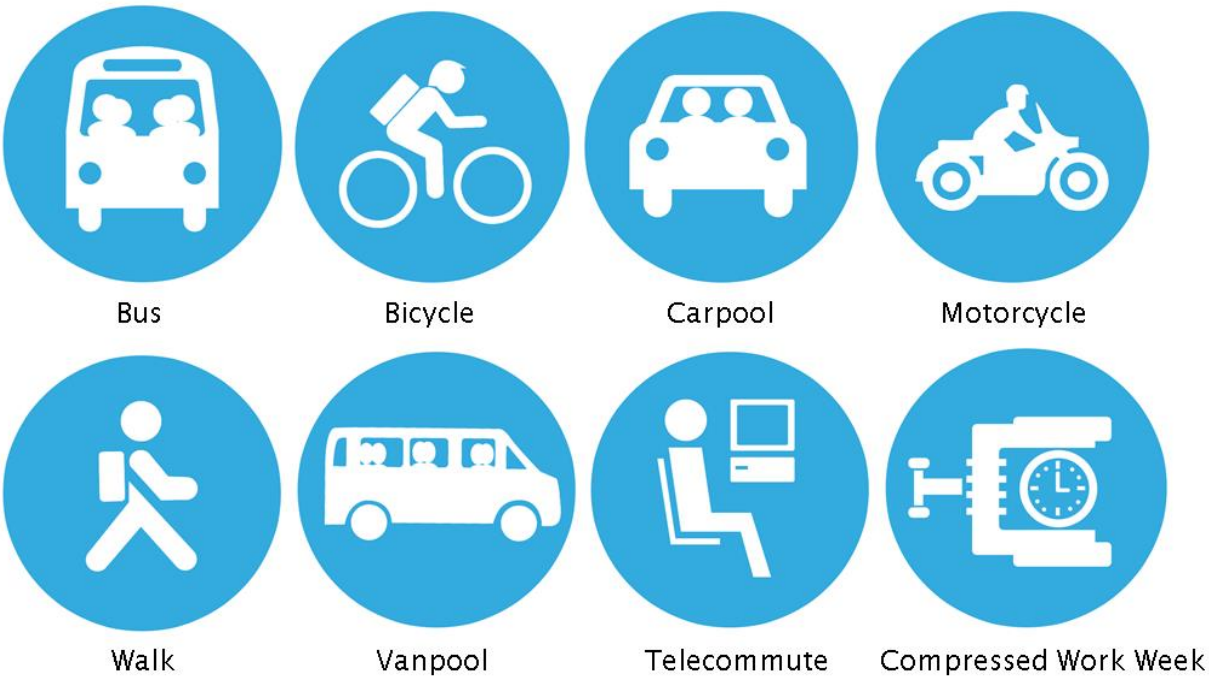
Deadline: 12/5/2015
Faculty Advisor: Sean Qian
Proposer contact info:
Name, email, phone #
Client point of contact:
Michael Thieme
michael.t.thieme@census.gov
(301) 763-9062

Carnegie Mellon University
Heinz College

United States Census Bureau

Capstone Report

Spring 2016



Advisor

Sean Qian

Team

Manikandan Palani | Nisha Rao | Sahil Aggarwal| Yifei Jiang | Yun Fu

Carnegie Mellon University Heinz College

United States Census Bureau

Table of Contents

Executive Summary

This report is one of the deliverables associated with the Heinz Capstone Project program at Carnegie Mellon University. In this program, students are grouped in teams of five and assigned a client for whom they will undertake an initiative over the course of a semester.

Our client for this project is the Centre for Adaptive Design of the United States Census Bureau. The U.S Census Bureau is the principle U.S Federal Statistical system in the U.S.

Our goal was to help reduce the cost of conducting the American Community Survey. This report is divided into several sections that detail the steps taken by the team to develop a solution that helps achieve the goal. Firstly, it provides a brief background on the team' objective. Secondly, it details the model built by CMU Capstone team to estimate response to two of the five work-commute related questions in the American Community Survey (ACS) using historical ACS data and data from external sources – INRIX, Pittsburgh Parking. The team compared the accuracy of the results of the model using a 20%, 40%, 60% and 80% sampling plan to observe the utility of the model in scenarios where only 20% or 40% of the existing respondents were to be contacted. The model could help the Bureau to reduce the number of people surveyed yet obtain reliably accurate responses for the rest of the intended target population and subsequently reduce survey costs. Thirdly, the report details the future work of including factors such as congestion and weather that could help further refine the model.

Introduction

Background

US Census Bureau conducts the American Community Survey (ACS) on an ongoing basis to collect information about the American people and their community. This data is used by planners and entrepreneurs to assess the past and plan the future. It also helps determine the distribution of more than 400$ billion in federal and state funds each year.

The survey is administered through several ways to one in thirty-eight U.S households per year. The form can be completed online or on a paper form. U.S Census Bureau has a follow-up procedure in case a respondent does not respond within a stipulated time frame. The non-respondent either receives a call or a personal visit from the Census staff to help complete the survey.

The cost of conducting the survey is increasing while the survey response rate is declining. ACS currently has seventy-seven questions. To reduce respondent burden, Centre for Adaptive Design of US Census Bureau is researching ways to reduce survey cost.

Of its seventy-seven questions, ACS has five questions that capture information about work-commute of the respondents. These questions have proved to be very problematic when it comes to eliciting responses.

Goal

Under the guidance of Sean Qian, CMU Professor, our goal of this project was to access ways to infer the travel and commute related questions in ACS.

Project Methodology

Of the five work-commute related questions in ACS, the team concentrated on two questions - “How did this person usually get to work last week?” and “How many minutes did it usually take this person to get from home to work last week?”. The objective was to build models to estimate responses to these questions.

So, the project methodology has been detailed with respect to the two tasks.

Task 1: ACS question - “How did this person usually get to work last week?”

Summary: The team through its initial research (Appendix I: NHTS Data Analysis) identified the methodology to develop a model to predict the mode of transport as public or private. The explanatory variables for the model were initially identified from both NHTS and ACS data sources.

NHTS data is detailed and can answer many questions directly but its latest data is from year 2009. NHTS is not conducted as frequently as ACS. Due to this fact, the utility of the model was limited. More details are provided in "Step I: Predictive model using National Household Travel Survey data"

The team then identified equivalent of NHTS variables in ACS data. The team members developed a model to predict the mode of transport as public or private using solely the variables from ACS data. They refined the model further to be able to identify the type of public or private mode of transport taken by a respondent. The utility of this model was tested using 20%, 40%, 60% and 80% random sampling plan. More details are provided in "Step II: Predictive model using PUMS data of American Community Survey"

Task 2: ACS question - “How many minutes did it usually take this person to get from home to work last week?”

Summary: Using the Commuting in America III report (Appendix II: Factors affecting travel time in US), the team identified intuitively the variables that could affect the work-commute travel time. The team members also researched (Appendix III: Bus Transit Time and Appendix IV: INRIX Traffic Data) on how to make use of the external data sources such as Bus transit data and INRIX traffic speed data in the model to predict travel-time.

The team faced challenges with respect to data granularity and overcame them by using data cleaning methods and aggregating the data at the appropriate level. The team also had to make quite a few assumptions in order to make use of the external data sets. The members developed a method to calculate mean travel time for bus transit at block level. More details are provided in "Step I: Mean travel time for Public transportation - Bus in a given block ".

The details of the travel-time predicting model are provided in "Step II: Predictive model using historical data from ACS along with inputs from external data set".

The team also did some research on calculating the congestion level in urban areas (Appendix V: Urban Congestion Level). As per the "Commuting in America III report", travel time is an attribute of commuters whereas congestion is an attribute of facilities. Therefore, measures of travel times and measures of congestion do not necessarily converge. Hence the congestion factor has not been delved into detail and currently not been incorporated in the travel-time model.

The details of each step in each task are described below.

Task 1: ACS question "How did this person usually get to work last week?"

Response options in ACS:

- Car, truck, or van
- Bus or trolley bus
- Streetcar or trolley car
- Subway or elevated
- Railroad
- Ferryboat
- Taxicab
- Motorcycle
- Bicycle
- Walked
- Worked at home
- Other method

Step I: Predictive model using National Household Travel Survey data

Summary: This model was built to initially classify the population in terms of whether an individual used a public or a private mode of transport to commute to work, and once that model showed promise, to dig deeper into the prediction by being able to also identify which particular mode of transport was used (for e.g. car, bicycle etc. in private vehicles and bus, train etc. in public modes of transport).

Tool: R Studio

Input Data:

Source: National Household Travel Survey

Details:

Year	Level	# of variables	# of records
2001	National	142	70k
2009	National	150	75k

Data Cleaning Step: Some of the input variables were non-binary categorical variables. The logistic regression model cannot accept non-binary categorical variables. These variables were converted to relevant dummy variables for use in the predictive process. For e.g. The variable for 'Race' can take up to 9 different values. So at least 8 dummy variables were needed for each value that the variable could assume.

After taking the dummy variables into account, we had a total of around 160 variables. The variables that did not realistically hold any promise were weeded out either due to their inherent definition or due to their infrequency in the actual data. After removing the variables that had almost zero frequency of occurrence, the shortlisted variables totaled in the 60-80 range.

Explanatory Variables:

- Age
- Sex
- Race
- Region
- Urban housing status
- Block housing status
- MSA size
- Education level
- Hispanic indicator population density
- Total income
- Household income
- Renter percentage
- Worker status
- Number of vehicles
- Homeownership status
- Place of birth

Selection logic: Details are provided in Appendix I: NHTS Data Analysis

Predicted Variable: Mode of transport – Public or Private

Model: GLM based logistic classifier

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_m x_{m,i}$$

$$\text{where } \ln\left(\frac{p_i}{1-p_i}\right) = \text{log odds of private mode of transport used}$$

$x_{1,i}$ = predictor variable i

Code: "Details in Appendix V under section "Code used to predict mode of transport: Pittsburgh_Mode_Choice.R"

Output:

- 93% prediction accuracy for 2001 NHTS data at 0.5 cutoff as a decision rule
- 95% prediction accuracy for 2009 NHTS data at 0.5 cutoff as a decision rule

Validation: 10-fold cross validation method was used to validate the results. It had an average accuracy rate of 93%.

Challenge:

- Highly skewed nature of data - 92% people use private mode of transport
- Trade-off between accuracy and the number of false positives for tweaking the decision rule and increasing it beyond 0.5

The overall rate of public transport use in the NHTS sample was around 7% which was roughly our model's misclassification rate. This happened due to the inherently skewed nature of the data, where ~93% of the population used a private form of transport to commute to work whereas only 7% used a public mode of transport. Furthermore, our model's ROC curve showed 26% greater recall than a flipping-of-coin based model.

Conclusion: The most important and most correlated variables (with the mode of transport used to commute to work) are Age, Sex, Education level and Total Income- all of which made perfect intuitive sense.

The National Household Travel Survey data can be used to estimate cannot be used as an alternative source.

Step II: Predictive model using PUMS data of American Community Survey

Summary of the model: This model was built to classify the population in terms of the particular mode of transport (for e.g. car, bicycle, train, etc.) used to commute to work.

Tool: R Studio, Amazon Web Service

Input Data: American Community Survey's historical data

The American Community Survey (ACS) Public Use Microdata Sample (PUMS) files are a set of untabulated records about individual people or housing units. The Census Bureau produces the PUMS files so that data users can create custom tables that are not available through pretabulated (or summary) ACS data products.

Data Cleaning Step:

- The team extracted the Pittsburgh City data from the whole national wide dataset based on the PUMA code of Pittsburgh City (01701 and 01702).
- For the missing values, we replaced them with 0 or 01 according to certain circumstances.

Explanatory Variables:

There are 512 variables in the dataset, the team then selected the variables that most influence which transport mode the person is most likely to use for both Pittsburgh City and U.S. national wide.

Variables for Pittsburgh City:

Name	Explanation
JWMNP	Travel time to work
FMRGIP	First mortgage payment includes fire, hazard, flood insurance allocation flag
MIG	Mobility status
PINCP	Total person's income
COW	Class of worker
JWRIP	Vehicle occupancy
INTP	Interest, dividends, and net rental income past 12 months
FLANXP	Language other than English allocation flag
PWGTP	Person's weight
SPORDER	Person Number
PERNP	Total person's earnings
HINS2	Insurance purchased directly from an insurance company

TYPE	Type of unit
FS	Yearly food stamp recipiency
RACASN	Asian recode (Asian alone or in combination with one or more other races)
RNTM	Meals included in rent
JWTR(Target)	Means of transportation to work

Variables for U.S. National Wide:

Name	Explanation
JWMNP	Travel time to work
POWSP	Place of work - State or foreign country recode
HINS7	Indian Health Service
Wgtp	Housing Weight replicate 3
REFR	Refrigerator
RACBLK	Black or African American recode (Black alone or in combination with one or more other races)
MHP	Mobile home costs (yearly amount)
SSP	Social Security income past 12 months
PUMA	Public use microdata area code (PUMA)
OCCP	Occupation recode
POWPUMA	Place of work PUMA
FINDP	Industry allocation flag
SMOCP	Selected monthly owner costs
HINCP	Household income (past 12 months)
OCPIP	Selected monthly owner costs as a percentage of household income during the past 12 months
JWTR	Means of transportation to work

Selection Logic: Same as the one mentioned in the previous section.

Predicted Variable: Mode of transport – Car, Bus, Streetcar, Subway, Railroad, Ferryboat, Taxicab, Motorcycle, Bicycle, Walked, Work at home or other method

Model:

Multinomial Logistic Regression is the linear regression analysis to conduct when the dependent variable is nominal with more than two levels. Thus it is an extension of logistic regression, which analyzes dichotomous (binary) dependents.

$$\Pr(Y_i = c) = \frac{e^{\boldsymbol{\beta}_c \cdot \mathbf{X}_i}}{\sum_{k=1}^{K} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}}$$

Also, in order to prevent the over fitting problem, we conducted 10-folder cross validation. Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (testing dataset). The goal of cross validation is to define a dataset to "test" the model in the training phase (i.e., the validation dataset), in order to limit problems, like over fitting, give an insight on how the model will generalize to an independent dataset.

Code:

Details in Appendix V under section “Code used to predict mode of transport: Pittsburgh_Mode_Choice.R”

Output:

Pittsburgh City

Accuracy / Percentage	Car	Bus	Streetcar	Subway	Railroad	Ferryboat	Taxicab	Motorcycle	Bicycle	Walked	Worked.at.home	Other.method
20%	85.444%	59.288%	0.200%	0.700%	0.000%	0.000%	0.000%	0.000%	3.123%	46.172%	17.648%	1.665%
40%	90.418%	59.762%	0.650%	1.692%	0.000%	0.000%	0.000%	0.000%	7.494%	48.328%	37.454%	3.340%
60%	93.415%	61.997%	0.400%	3.075%	0.000%	0.000%	0.000%	0.000%	8.775%	51.798%	51.333%	6.263%
80%	94.800%	64.093%	0.150%	4.517%	0.000%	0.000%	0.000%	0.000%	10.639%	54.222%	63.457%	10.173%
100%	99.032%	73.411%	0.000%	0.160%	0.000%	0.000%	0.000%	0.000%	12.883%	63.552%	86.961%	27.586%

U.S. National Wide

Accuracy / Percentage	Car	Bus	Streetcar	Subway	Railroad	Ferryboat	Taxicab	Motorcycle	Bicycle	Walked	Worked.at.home	Other.method
10%	80.284%	39.628%	0.000%	45.524%	35.015%	0.000%	2.817%	0.000%	5.226%	21.684%	81.075%	9.019%
20%	76.700%	34.175%	7.692%	36.015%	35.252%	0.000%	3.030%	0.000%	1.829%	22.154%	74.696%	1.569%
30%	79.294%	35.880%	0.763%	52.198%	32.941%	1.695%	1.031%	0.315%	3.828%	18.568%	88.898%	5.550%
40%	81.880%	33.657%	2.367%	47.558%	30.679%	1.266%	1.667%	0.605%	3.434%	21.922%	81.998%	5.762%
50%	83.723%	31.566%	0.885%	49.272%	33.236%	0.980%	1.017%	0.191%	3.487%	22.405%	80.013%	7.215%

Validation:

Due to the highly skewed original dataset (overall 95% of the households use car to commute), the team has to rebalance the dataset. We specifically extracted half of the data as commuters who use car and the rest half as commuters who don't use car.

For Pittsburgh City, the team then randomly subset the dataset into samples of 20%, 40%, 60% and 80% and ran simulation 1000 times to compare the accuracy. However, for U.S. National wide, even if we deployed AWS and launched the largest instance (x10.Large) on it, we still couldn't run the whole national dataset. So alternatively, we randomly subset the dataset into samples of 10%, 20%, 30%, 40% and 50%, and ran simulation only 1 times to compare the accuracy.

Challenge: Large volume of data

Conclusion:

For Pittsburgh City data, the average accuracy after 1000 times simulation runs shows that partial sampling can lead to roughly lesser accuracy. However, for U.S. National Wide data, we cannot see a clear pattern. We believe that the reason behind of this phenomenon is the lack of simulation times. Given the budget and time restriction, we realized that there is a limit of computer power.

Task 2: ACS question "How many minutes did it usually take this person to get from home to work last week?"

Step I: Mean travel time for Public transportation - Bus in a given block

Summary: **Use the arithmetic mean formula to calculate the mean travel time**

Tool: **Excel and Python**

Input: **APC AVL (Automatic passenger counter and automatic vehicle location) data**

Year	Level	Location
September 2013	City	Pittsburgh

Data Cleaning Step: **Eliminate missing values by filtering out those records from the input.**

Input variables:

- People getting on the bus at a stop
- People getting off the bus at a stop
- Current load in the bus
- Stop ID
- Block ID or FIPS code
- Time taken between the stops
- Dwell times

Selection Logic: Major variables were used directly or indirectly for calculating mean travel time. There was no variable selection methods used. This is direct calculation.

Predicted Variable: mean travel time per person in a given tract (FIPS code)

Model: Formula for calculating the mean travel time is given below.

$$MTT = \frac{\sum_1^n t}{n}$$

MTT: Mean travel time of people who use public transport in a particular block

n: Number of people who get on the bus

t: Difference in time taken between getting on the bus and getting of the bus.

Code: Python code files: Code is in Appendix.

- getStopFIPS.py: this is to get the FIPS code of stop from the Census API
- filter.py: Filtered mean travel time from tract level to county level
- calculateMeanTT.py: This is to calculate the mean time

Assumptions:

- Location at which a passenger boards is assumed to be that passenger's home/ work location in that particular tract.
- This may necessarily not be true. It is important to realize that a person can get in and get down at random stops according to his wish. Though there are patterns in the data, there is no way to determine the source or the destination of a passenger from this data. This affects the average travel time critically. There is no data to identify the person in question or for what purpose the person travels in the public transport. Mean travel time cannot be computed at an individual level.
- FIFO assumption visualized

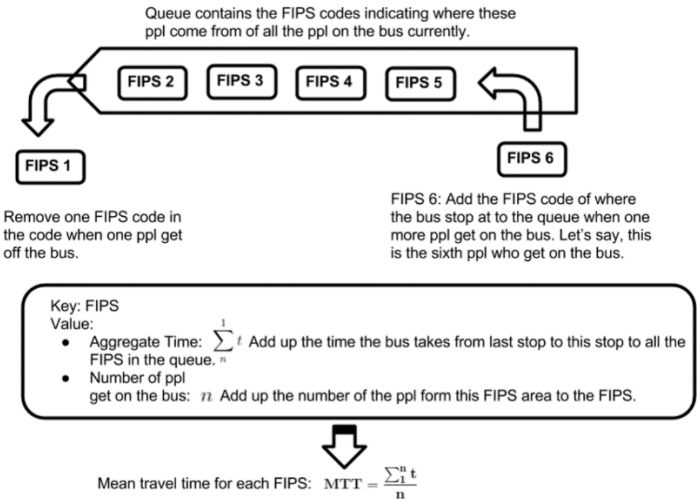


Figure 8 section 2

Output:

Mean travel time visualized

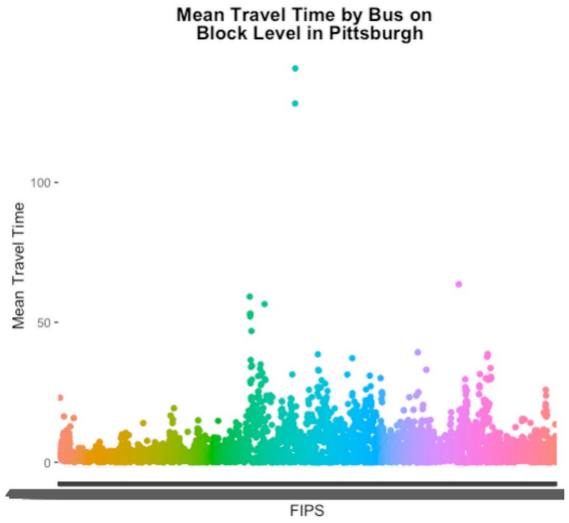
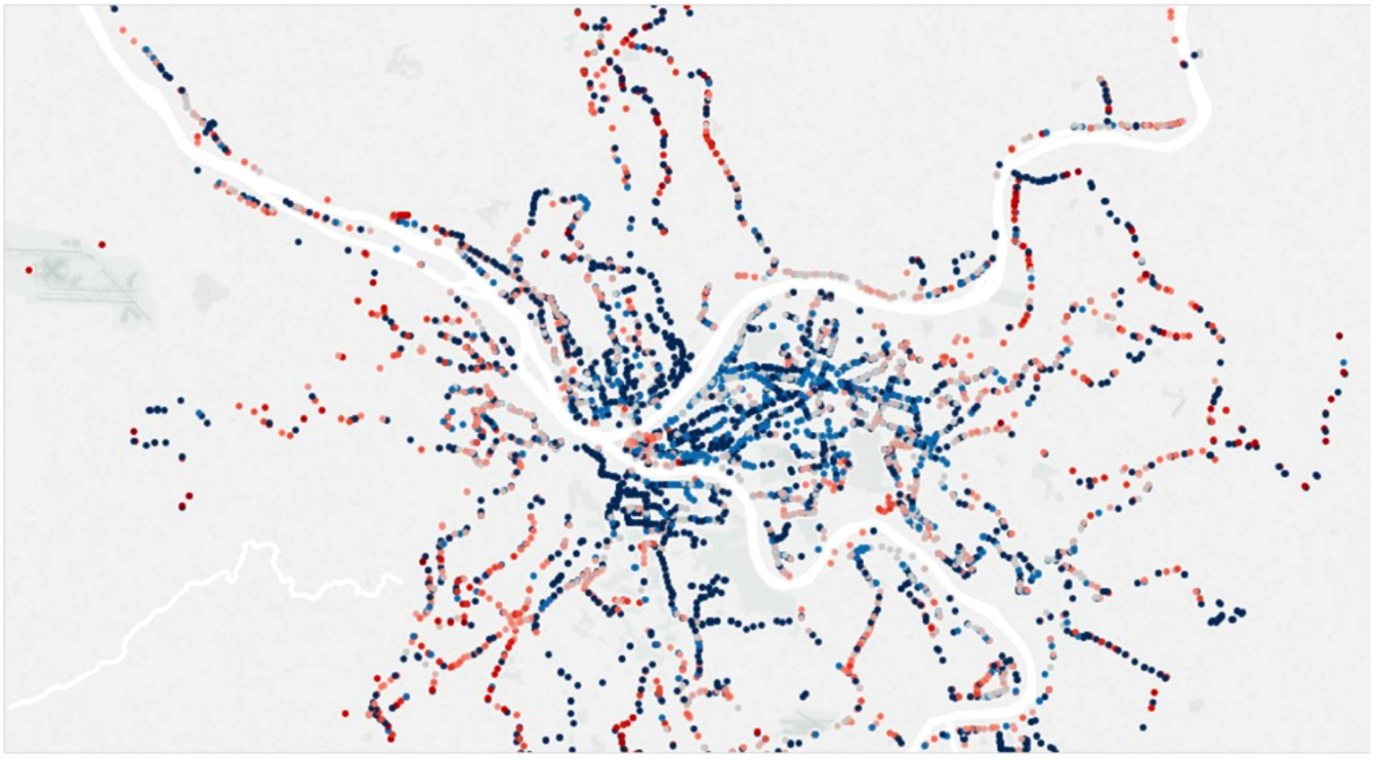


Figure 9 Section 2



Map based on Longitude and Latitude. Color shows sum of Mean Time. Details are shown for Fips.

Figure 10 Section 2 – The visualization of mean travel for bus transport at block level for Pittsburgh

Challenge:

- The bus transit data only has stop ID numbers and stop name with no coordinates.
- Finding the coordinates for the stop ID was a difficult task. The team searched several APIs including google maps and finally found the values in Pittsburgh city GIS data.
- Joined the Pittsburgh city GIS data and the bus transit data using stop ID as primary key and obtained the coordinates for the stop.
- Once the coordinates were obtained, the coordinates were used to get the FIPS code, which identifies the block to which the stop belongs.
- All the bus stops were then assigned to their respective block using FIPS code.
- The stops were grouped according to FIPS code.
- Once this grouping was done, average travel time for each FIPS code was calculated using the python code: calculateMeanTT.py

Conclusion:

Though, we make several assumptions the relative mean travel time values for different blocks can still give us a good idea about the trends in travel time across blocks.

For instance, Mean travel time is stops in downtown and quite long for stops in rural areas. Below figure shows, the visualization Red indicates more time. Blue indicates higher time.

Step II: Predictive model using historical data from ACS along with inputs from external data set

Summary: The team tried to answer these questions by building a linear regression model by combining variables from ACS data and the External Data Sources. The main motive was to find the effect of adding External Data Source variable on the ACS travel time for all tracts within Pittsburgh.

Tool: R Studio, WEKA

Input:

Data Source	Time Period	Level	Location	# of records	# of variables
ACS	5year estimates	Census Tract	Pittsburgh	22	33
Pittsburgh Parking Terminals	2013 and 2014	City	Pittsburgh		1
APC AVL	2013 and 2014	City	Pittsburgh		1

Data Cleaning Step:

Input variable: Details about calculation of Mean travel time for public transportation is mentioned in the previous section.

Variable name	Description (All variables are in tract level)
TRACTCE10	TRACT ID
Public MTT	Pittsburgh public transport mean travel time calculated from external data set
Parking Transactions	Number of parking transaction in Pittsburgh
Population Total	Total population
Income Per Capita	Per capita income
Means of Transport Total	Total number of vehicles
Car truck van	Number of cars, trucks or van
Drove Alone	Number of people who drove alone
Carpooled	Number of People who car pooled
Public transport	Number of People who used public transport
Bicycle	Number of People who used bicycle
Walked	Number of people who waked
Other means	Number of people who used other modes of transport
Worked at home	Number of people who worked from home
Less than 5 minutes	Number of people who took less than 5 minutes to travel to work
5 to 9 minutes	Number of people who took between 5 to 9 minutes to travel to work
10 to 14 minutes	Number of people who took between 10 to 14 minutes to travel to work
15 to 19 minutes	Number of people who took between 15 to 19 minutes to travel to work

20 to 24 minutes	Number of people who took between 20 to 24 minutes to travel to work
25 to 29 minutes	Number of people who took between 25 to 29 minutes to travel to work
30 to 34 minutes	Number of people who took between 30 to 34 minutes to travel to work
35 to 39 minutes	Number of people who took between 35 to 39 minutes to travel to work
40 to 44 minutes	Number of people who took between 40 to 44 minutes to travel to work
45 to 59 minutes	Number of people who took between 45 to 49 minutes to travel to work
60 to 89 minutes	Number of people who took between 60 to 89 minutes to travel to work
90 or more minutes	Number of people who took between 90 minutes to travel to work
White	Race identifier
African American	Race identifier
Asian	Race identifier
Other	Race identifier
Male	Number of males
Female	Number of females
Avg TT	Average travel time

Selection Logic: This is detailed under the section "Factors affecting travel time in U.S" in "Appendix I: Literature Review".

Predicted variable: Travel time at tract level

Model: Linear regression

Code: Lm (formula = X$Avg_TT ~X$Public_MTT + X$Parking_Transactions + X$Population_Total + X$Means_of_Transport_Total, data = X)

Output:

Below is the result of the regression model without excluding the highly correlated variables.

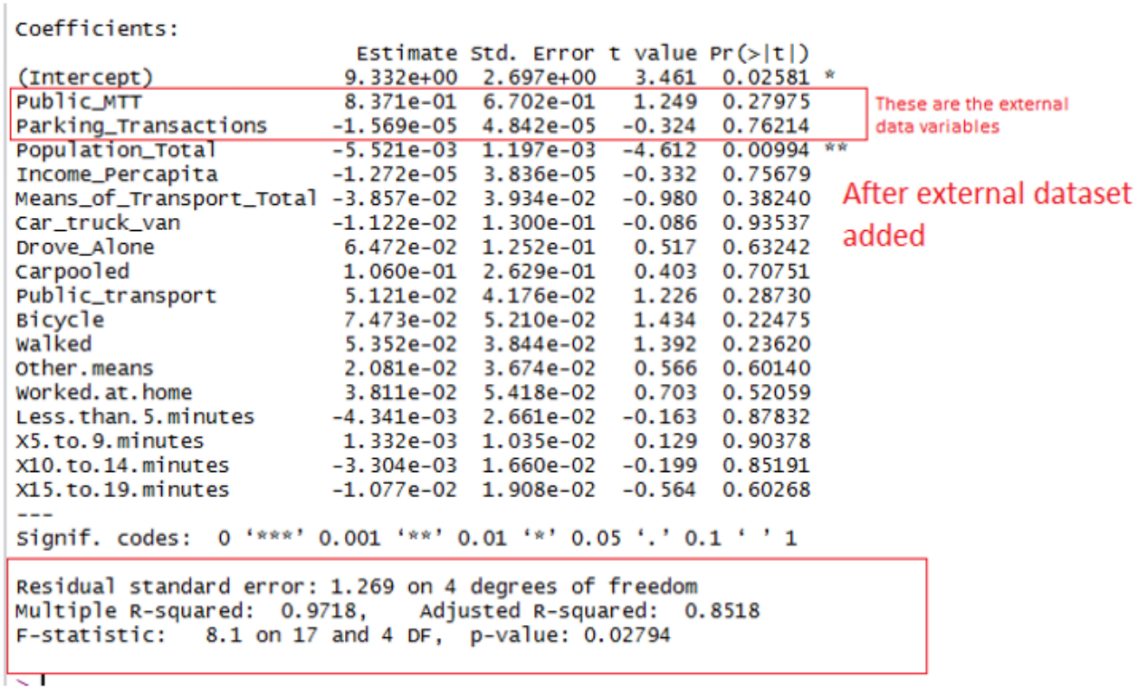


Figure 11 Section 2

It can be clearly seen that R square is high. R square explains the variability in the model and a High R square value (close to 1) implies there are many variables that are explaining the variation in the model. P value which denotes the significance is also lesser than 0.05. Therefore, we decided to exclude the highly correlated variables and run the regression model only with the essential variables. The table below shows the highly correlated variables that were excluded.

Variable	Response(Y)	Coorelation	P value
TRACTCE10	Avg_TT	0.66996247	0.00065
Public_MTT	Avg_TT	0.04346859	0.84768
Parking_Transactions	Avg_TT	-0.15737116	0.48429
Population_Total	Avg_TT	-0.29064602	0.18945
Income_Percapita	Avg_TT	0.45658502	0.03268
Means_of_Transport_Total	Avg_TT	0.31166765	0.15796
Car_truck_van	Avg_TT	0.49907997	0.01805
Drove_Alone	Avg_TT	0.46218032	0.03034
Carpooled	Avg_TT	0.50182629	0.01733
Public_transport	Avg_TT	0.49013776	0.02057
Bicycle	Avg_TT	0.42754686	0.04717
Walked	Avg_TT	-0.18694438	0.40482
Other.means	Avg_TT	-0.15612894	0.48778
Worked.at.home	Avg_TT	-0.14554225	0.51811
Less.than.5.minutes	Avg_TT	-0.0155714	0.94517

Figure 12 Section 2

Output of Regression runs after exclusion:

```
Call:
lm(formula = X$Avg_TT ~ X$Public_MTT + X$Parking_Transactions +
    X$Population_Total + X$Means_of_Transport_Total, data = X)

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                9.301e+00  7.183e-01  12.950 3.11e-10 ***
X$Public_MTT               7.227e-01  2.826e-01   2.557   0.0204 *
X$Parking_Transactions    -4.102e-05  1.912e-05  -2.145   0.0467 *
X$Population_Total        -4.064e-03  3.614e-04 -11.245 2.70e-09 ***
X$Means_of_Transport_Total 7.996e-03  6.955e-04  11.496 1.93e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.155 on 17 degrees of freedom
Multiple R-squared:  0.9006,    Adjusted R-squared:  0.8772
F-statistic: 38.49 on 4 and 17 DF,  p-value: 2.607e-08
```

Figure 13 Section 2

Goodness of Fit:

Consider any linear regression model, which looks like the following

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Are all the assumptions - Normality of residuals or errors from the model, constant residual variance throughout the range – true? It is mostly unlikely. It is only plausible that the assumptions are close enough. Goodness of fit defines how closely the assumptions for the model are useful in practice.

There are several measures for goodness of fit they are:

- Examining residues
- A global measure of variance explained by R^2
- A global measure of variance adjusted for number of parameters in the model called adjusted R^2

Residuals can be used descriptively, by looking at either histograms or scatter plots of residuals. Consider the following model -

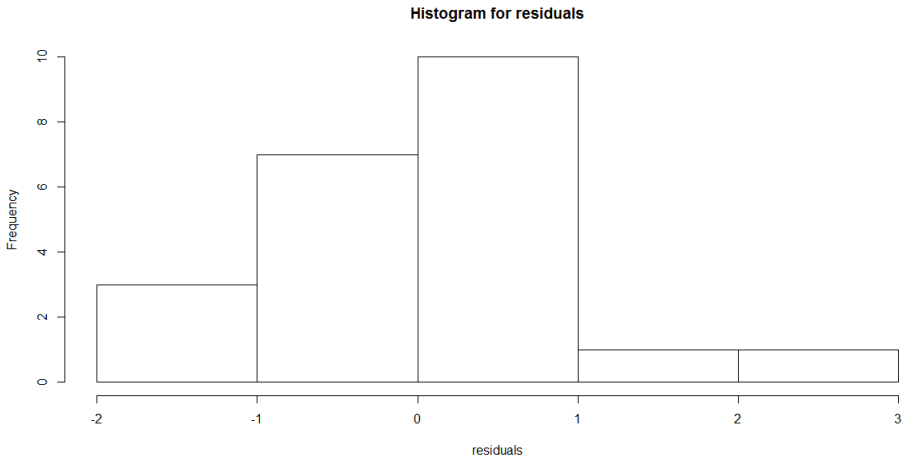
$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

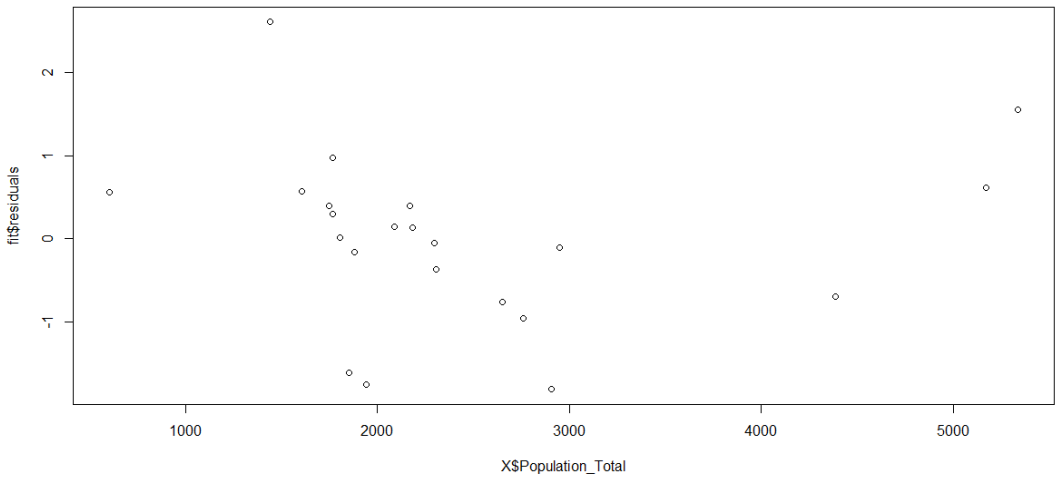
The i^{th} residual for the i^{th} observation is given by

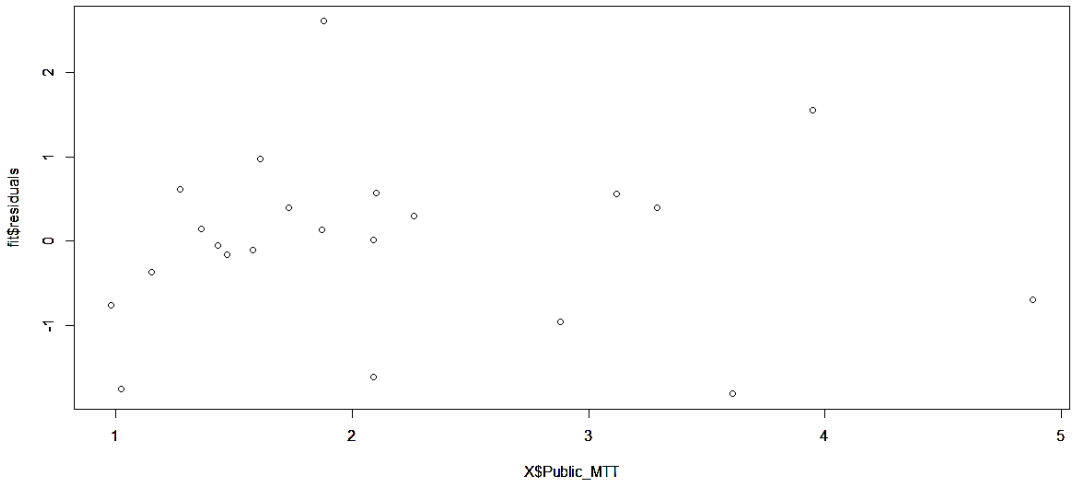
$$Y_i - \hat{Y}_i = Y_i - \left(\hat{\alpha} + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \ldots + \hat{\beta}_p X_{ip}\right)$$

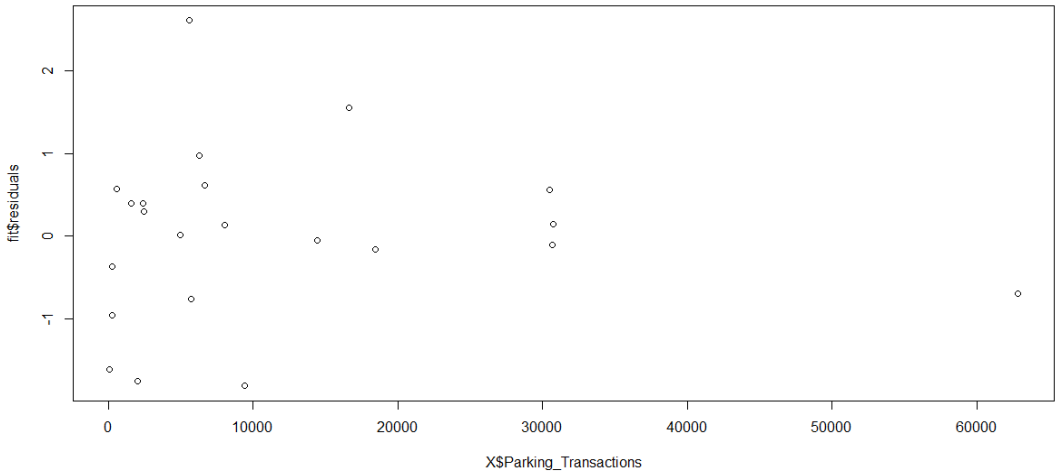
Where Yi is the observed dependent variable and Xij is the observed covariates for the i^{th} observation. Now let us plot the graph and see the outcomes.

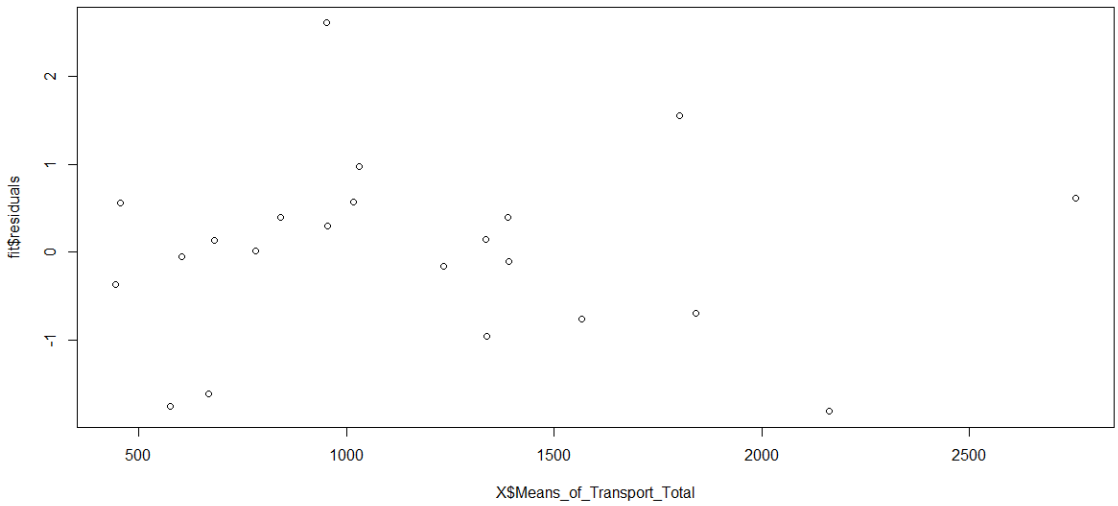
```
Fit = lm(X$Avg_TT~
X$Public_MTT+X$Parking_Transactions+X$Population_Total+X$Means_of_Transport_Total)
> hist(fit$residuals, main = "Histogram for residuals", xlab = "residuals")
> plot(X$Population_Total,fit$residuals)
> plot(X$Public_MTT,fit$residuals)
> plot(X$Parking_Transactions,fit$residuals)
> plot(X$Means_of_Transport_Total,fit$residuals)
```






fit$residuals
2
1
0
-1
1
2
3
4
5
X$Public_MTT


fit$residuals
2
1
0
-1
0
10000
20000
30000
40000
50000
60000
X$Parking_Transactions



The histogram is not strongly normal in shape, but, then again, there are just 22 Observations. Scatter plots look reasonable, i.e., no obvious departures from Constant variance or linearity.

*R*2: A global measure of "variance explained"

R2 value is repeatedly used in regression to explain the fit. We will now see exactly how to interpret this measure. Consider first simple linear regression. We are usually interested in whether the independent variable is worth having, so really we are comparing the model
$Y = \alpha + \beta X$ to the simpler model $Y = \alpha$, using the intercept only.

We define:
*R*2 = 1 –*{(*sum of squared residuals from model with α and β/ (sum of squared residuals from model with α only)}

Or

*R*2 = 1 – *{SS* (*res*)/*SS* (*total*)}

where SS(res) (often also referred to as SSE or "sum of squares of error") is defined as the sum of residual distances from model with X, and SS(total) is the same, but from the intercept only model.

By definition of least squares regression, SS (res) ≤ SS (total), because if the best regression line was really using α only, then SS (res) = SS (total), and in all other cases, adding β improves SS (res).

So, $0 \leq R^2 \leq 1$. If SS (res) = SS (total), then $R^2 = 0$, and model is not useful. If SS (res) = zero, then R2 = one, and model fits all points perfectly. Almost all models will be between these extremes.

Therefore, SS (res) shows how much closer the points get to the line when β is used, compared to a flat line using α only (which is always Y = α = Y).

Because of this, we can call R2 the "proportion of variance explained by adding the variable X".

Essentially the same thing happens when there is more than one independent variable, except residuals are from the model with all X variables for the numerator in the definition of $R^{2.}$ Thus, R^2 gives the "proportion of variance explained by adding the variables X1, X2. . . Xp, if there are p independent variables in the model.

How large does R^2 need to be to be considered as "good"? This depends on the context; there is no absolute answer here. For hard to predict Y variables, smaller values may be "good". Overall, R2 provides a useful measure of how well a model fits, in terms of (squared) distance from points to the best fitting line. However, as one adds more regression coefficients, R2 never goes down, even if the additional X variable is not useful. In other words, there is no adjustment for the number of parameters in the model.

Adjusted *R*2

In a simple linear regression, p is the number of independent variables.
It p = 1 then Adjusted $R^2 = R^2$. As the number of parameters increases, Adjusted $R^2 \leq R^2$.

With this definition:
*R*2 = 1 − {(*n* − 1) × sum of squared residuals from model with *α* and *β*/ (*n* − *p*) × sum of squared residuals from model with *α* only}

Therefore, there is some attempt to adjust for the number of parameters. Let us see this for our model below.

```
Call:
lm(formula = X$Avg_TT ~ X$Public_MTT + X$Parking_Transactions +
   X$Population_Total + X$Means_of_Transport_Total)

Residuals:
    Min      1Q  Median      3Q     Max
-1.80948 -0.61435  0.07758  0.52286  2.60368

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              9.301e+00  7.183e-01  12.950 3.11e-10 ***
X$Public_MTT             7.227e-01  2.826e-01   2.557   0.0204 *
X$Parking_Transactions  -4.102e-05  1.912e-05  -2.145   0.0467 *
X$Population_Total      -4.064e-03  3.614e-04 -11.245 2.70e-09 ***
X$Means_of_Transport_Total  7.996e-03  6.955e-04  11.496 1.93e-09 ***
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.155 on 17 degrees of freedom
Multiple R-squared: 0.9006, Adjusted R-squared: 0.8772
F-statistic: 38.49 on 4 and 17 DF, p-value: 2.607e-08

R report: R^2 = 0.9006 and Adjusted R^2 = 0.8772, so, in either case, about 90% of the total variance is explained by the variables used, which is very high. At least by these measures, the model fits well.

Challenge:

We initially collected the tract level data for the 33 selected variables and combined them into one file. 138 tracts fall under Allegany- Pittsburgh borders.

Once this data was collated, we used the TRACT ID and TRACT name as key to combine the External Data Source variables to the ACS dataset.

Parking data and APCAVL data are collected by different authorities. They do not have a common ground for combining. When the data sets are brought to a common level for example, at tract level it results in missing data.

The main reason for this is parking data is not available for all tracts in Pittsburgh i.e. parking spots are not spread across all areas in a tract in Pittsburgh. Similarly, bus lines travel across particular routes and data for some tracts are not available. When combined this naturally results in data not being available for some tracts.

Once the datasets were combined it resulted in data only for 22 out of 138 Pittsburgh tracts.
This sample size of 22 is extremely small to run a regression model. However, taking into account that this is only a proof of concept we built a regression model using R.
The regression model was initially run with all the intuitively selected variables for ACS along with the external data set variables.

However, this model had large number of variables that were highly correlated and effects of the external dataset variables could not be realized clearly.

So we excluded the highly correlated variables and ran a regression model only with 4 variables. 2 from external data set. Public mean travel time and number of parking transactions per tract and the ACS variables. Total tract population and number of vehicles in a tract. The results are discussed in the next page.

Conclusion:

- The first R square value in the regression results after excluding the significant variables is very high. It means the existing variables in the model explain the variability in the

external dataset very well. The p-value is also less than 0.05 and it means the variables are significant.

- Mode of transport for a commuter can be predicted with just a fraction of the current sample available to a good degree of accuracy.
- Mean travel time for people using bus as public transportation has been calculated for block level. Other calculations, involving data from external data source, include parking transactions for all tracts within Pittsburgh and average speed data and average travel time.
- Using external and existing ACS data mean travel time for Pittsburgh tract level is well predicted. External data sets are not uniform. So, data needs to be cleaned and normalized. Normalization may result in a very small sample. As a result, this can only be used as a proof of concept, further research is encouraged and results need to be scrutinized.

Conclusion

- Mode of transport for a commuter can be predicted well, even with a fraction of the current sample.
- Mean travel time at Pittsburgh tract level is well predicted using a combination of APC AVL and PA parking with the ACS data (87% of the variability explained)

Lessons Learned

Data

Real world data is extremely different and multifaceted. Data sets are complex as well as large in volume. Useful information must be extracted from raw data. Even after extracting useful information missing values and normalization issues can occur. Large data sets further create problems when it comes to computing. They require huge processing power and distributed computing. Laptops and mini computers cannot handle such data hence they need to be moved to the cloud. This results in extra costs. So better and more efficient code needs to be developed.

Assumptions

While building models using data mining several assumptions like FIFO, Queue are used these assumptions can give models that are very good for creating proof of concepts but not entirely accurate models. For building, more accurate models, one needs to develop or use customized machine learning algorithms based on the dataset.

Suggestions for Future Work

Travel mode model can be expanded to national level using the nation wide data with more computing power and distributed computing.

Expand linear regression model to predict travel time for

- All tracts of Pittsburgh.
- State wide and nationwide using other external datasets.

Travel mode model can be further refined by including the congestion and weather as independent variables.

Acknowledgement

This capstone project was undertaken by the team for Centre for Adaptive Design, US Census Bureau, with support from Carnegie Mellon University. We thank our advisor Professor Sean Qian for his continued support and guidance. We also thank our client for their patience and support.

This project allowed some of us to hone our analytical skills while others gained an understanding of the analytics field. We thank our Heinz college, Carnegie Mellon University, for providing us this learning opportunity.

Appendix I: Literature Review

National Highway Travel Survey Data Analysis

Source: The Impacts of Socio-Economic and Demographic Shifts in Transit Served Neighborhoods On Mode Choice And Equity by Steven Apell

Highlight the impacts of socio-economic and demographic changes on TOD. TOD stands for Transit-oriented development. It is a mixed-use residential and commercial area designed to maximize access to public transport, and often incorporates features to encourage transit ridership. The paper also evaluates the overall effectiveness of TOD policy in supporting a sustainable community.

Answered following four questions:

- How has the socio-economic and demographic character of communities changed in block groups, which are within 0.5 mile and 0.5-1.0 mile of TOD and non-TOD stations?
- Is the presence of TOD interrelated with the occurrence of gentrification in block groups within 0.5 and 0.5-1.0 mile of transit stations?
- How have changes in socio-economic and demographic factors influenced mode choice in block groups within 0.5 and 0.5-1.0 mile radius of TOD and non-TOD stations?
- What are the differences between socio-demographic characteristics of communities living in TOD compared with communities in non-TOD; further, what differences exist between communities within 0.5 miles of TOD compared with communities living within 0.5-1.0 mile radius?

Helped identify the following factors that affect the mode of transport chosen by a commuter to travel to work:

- Percentage of households with no car or more than one car
- Median and per capita income
- Percentage of college graduates
- Median housing value
- Median contract rent
- Percentage of owner and renter occupied housing
- Racial categories (Percentages for Black and White)
- Percentage of college graduates
- Employment Type
- Length of Residency

Factors affecting travel time in US

Source: Commuting in America Report III - The Third National Report on Commuting Patterns and Trends by Alan E. Pisarski

This report examines commuting patterns in terms of longstanding trends and emerging factors that affect commuting every day. It examines the data trend identified through the responses to each travel question of ACS in detail.

Travel time is a function of both speed and distance. The increase in work trip distances when matched by the increase in work travel speeds indicates an actual improvement in work travel speeds. Travel times increasing faster than the travel distances might indicate the effect of congestion in the travel time. Travel distances growing faster than travel speeds might indicate that the commuter prefers staying in a neighborhood with reduced housing cost and this neighborhood might be at a considerable distance from his work location.

It has also been observed in past that women make frequent stops while travelling and as such, their travel time will be more. Areas smaller in size will tend to send at least some portion of commuters to other metropolitan areas for work. Families who have toddlers travel less and a lot.

It helped us to intuitively identify the factors that could affect travel-time of a commuter. The variables identified were as follows –

- Age
- Gender
- Marital Status
- Presence and age of own children
- Gave birth to child within past 12 months
- Frequent residence shifters
- Citizenship status - categorical
- Education
- Income
- Number of vehicles in the household
- Monthly rent
- Number of bread earners in household

Bus Transit Time

Source: Impact of Traffic Congestion on Bus Travel Time in Northern New Jersey by Claire E. McKnight, Herbert S. Levinson, Kaan Ozbay, Camille Kamga, and Robert E. Paaswell

This paper details the impact of congestion on the bus travel time rate. It also provides details about a regression model developed to estimate travel time rate (in minutes per mile) of a bus as a function of car traffic time rate, number of passengers hoarding per mile, and the number of bus stops per mile.

It details the methodology in the following way -

- Give a brief overview of the previous studies conducted
- Describe the data collection and initial analysis of data
- Define the model
- State the conclusions

Predict Bus travel time for a particular route, particular bus and particular time slot.

A route segment is defined as a section of route between two adjacent time points, with a time point (TP) being the location at which the schedule had a recorded time.

Bus speed variation between stops can indicate congestion. A.m peak – 7 am to 10 am; Mid day – 10 am to 4 pm; P.m peak – 4 pm to 7 pm; Post p.m peak – after 7 pm.

Factors	Characteristics
Bus	➢ # of stops ➢ Stop spacing ➢ Dwell times at stops

	➢ # of passengers boarding ➢ # of passengers alighting
Route	➢ Segment length ➢ # of traffic signals ➢ # of left turns in route
Traffic	➢ Car travel time rate = traffic level ➢ # of vehicles in queue waiting to make left turn ➢ # of taxis making sudden stops or turns to pick up/ drop off passengers ➢ # of vehicles on the road
Parking	➢ # of double or triple parked cars

INRIX Traffic Data

Source: INRIX Interface Guide December 2014

This report provides detailed information about the INRIX data. It also acts as a guide to help understand how to make use of INRIX data.

It contains the following information -

- Interpreting TMC codes
- Understanding XD segments, sub-segments and INRIX-managed set files
- Integrating graphical traffic data in user applications
- Generating speed buckets

The data dictionary included in this report helped us to understand the variables in the input data file containing the INRIX traffic speed data.

Urban Congestion Level

Source: 2015 Urban Mobility Scorecard

It is published jointly by the Texas A&M Transportation Institute and INRIX. This report details the methodology used to estimate the congestion level at an urban area-wide level. This allows for comparison of congestion level in a consistent way across urban areas.

It uses a combination of traffic volume data along with the traffic speed data to compute the congestion level measures. All the measures and many of the input variables for each year and every city are provided in a spreadsheet that can be downloaded at http://mobility.tamu.edu/ums/congestion-data/. The roadway inventory data source for most of the calculations is the Highway Performance Monitoring System from the Federal Highway Administration. Traffic volume data is sourced from INRIX.

The following steps were used to calculate the congestion performance measures for each urban roadway section.

- Obtain HPMS traffic volume data by road section
- Match the HPMS road network sections with the INRIX traffic speed dataset road sections

- Estimate traffic volumes for each hour time interval from the daily volume data
- Calculate average travel speed and total delay for each hour interval
- Establish free-flow (i.e., low volume) travel speed
- Calculate congestion performance measures
- Additional steps when volume data had no speed data match

The mobility measures require four data inputs:

- Actual travel speed
- Free-flow travel speed
- Vehicle volume
- Vehicle occupancy (persons per vehicle) to calculate person-hours of travel delay

This paper could be referred in future to calculate the congestion factor to be included as another independent variable in the travel-time predicting model.

Appendix II: Information on External Data Sources

NHTS: National Household travel survey

Data: NHTS 2001 and 2009

NHTS is a travel survey that collects information about people, especially an individual's travel behavior over a period through questionnaire. These surveys collect information about demographics of the individual. Especially information about socio-economic status, household (size and structure), vehicle data, travel mode, vehicles owned, vehicles used, purpose of the journey, starting point and ending point of the journey and number of people who the person travelled with.

The National Household Travel Survey (NHTS) provides information to assist transportation planners and policy makers who need comprehensive data on travel and transportation patterns in the United States. The 2009 NHTS updates information gathered in the 2001 NHTS and in prior Nationwide Personal Transportation Surveys (NPTS) conducted in 1969, 1977, 1983, 1990, and 1995.

Data collected:

The NHTS/NPTS acts as nation's inventory of travel data. The data is collected on daily trips taken in a 24-hour period, and includes:

- Purpose of the trip (work, shopping, etc.)
- Means of transportation used (car, bus, subway, walk, etc.)
- Travel time
- Time of day when the trip took place
- Day of week when the trip took place

For a private vehicle trip, it also captures

- number of people in the vehicle, i.e., vehicle occupancy
- Driver characteristics (age, sex, worker status, education level, etc.)
- Vehicle attributes (make, model, model year, amount of miles driven in a year)

Scope – What the NHTS Includes

- Household data on members, education, income
- Housing characteristics, information on each vehicle, including year, make, model, and estimates of annual miles traveled
- Information on travel as part of work; Data about one-way trips taken during a designated 24-hour period
- Information to describe characteristics of geographical area

Scope — What Is Not Included

- Costs of travel
- Specific travel routes or types of roads used

- Travel of the sampled household changes over time
- Information that would identify the exact household or workplace location

APC AVL - Automatic passenger counter and automatic vehicle location data

Location: Pittsburgh
Time: 2012, 2013, 2014
Data used: 2013 / 2014

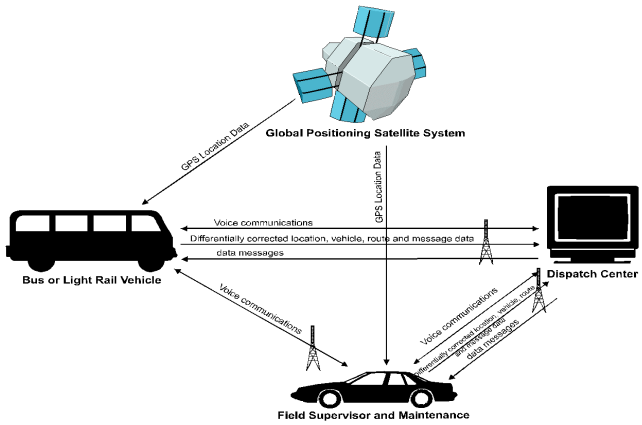
AVL Automatic Vehicle Location (AVL) Systems, is a part of ITS (intelligent transportation system), which has been adopted by many transit agencies to track their transit vehicles in real time. It is essentially a vehicle tracking system that combines the use of automatic vehicle location gathered by GPS attached to individual vehicles with software that collects a fleet of data for a comprehensive picture. The data collected includes vehicle locations, how long they travel, where they stop, how long it takes to move from one stop to another and the distance between the stops. Urban public transit authorities are an increasingly common user of vehicle tracking systems, particularly in large cities like Pittsburgh.

Automated Passenger Counter (APC) on the other hand is a device installed on transit vehicles, which accurately records boarding and alighting data. This device accurately tracks transit ridership when compared to the traditional methods of accounting by drivers or estimation through surveying. These devices are becoming more common among American transit operators seeking to improve the accuracy of reporting patronage as well as analyzing transit use patterns by linking boarding and alighting data with stop location.

A schematic diagram for AVL APC system is given below.

AUTOMATIC VEHICLE LOCATION

References: ntl.bts.gov



AUTOMATIC PASSENGER COUNTER

References: Pinterest

Figure 2 Section 1

A sample of how APC AVL system collects data is given below along with the variable description.

DOW	dir	ROUTE	TRIPA	BLOCKA	VEHNOA	daymoyr	STOPA	QSTOPA	ANAME	HR	MIN	SEC	DHR	DMIN
1	0		10615	001005	5505	09-Sep-2013	14	N17300	EAST OHIO ST AT GRANT	6	42	28	6	42
1	0		10615	001005	5505	09-Sep-2013	15	N12510	BUTLER ST AT #188	6	45	42	6	45
1	0		10615	001005	5505	09-Sep-2013	16	N12440	BUTLER ST AT ISABELLA	6	45	59	6	45
1	0		10615	001005	5505	09-Sep-2013	17	N12380	BUTLER ST AT CENTER	6	46	33	6	46
1	0		10615	001005	5505	09-Sep-2013	18	N12520	BUTLER ST AT FREEPORT	6	47	17	6	47
1	0		10615	001005	5505	09-Sep-2013	19	N20750	FREEPORT ST AT BRIDGE	6	47	46	6	47
1	0		10615	001005	5505	09-Sep-2013	20	N31500	MAIN ST AT 62ND ST BRIDGE FS	6	48	39	6	48
1	0		10615	001005	5505	09-Sep-2013	21	N31600	MAIN ST AT PRAEGER ST	6	49	1	6	49
1	0		10615	001005	5505	09-Sep-2013	22	N31520	MAIN ST AT 6TH ST	6	49	23	6	49
1	0		10615	001005	5505	09-Sep-2013	23	N31540	MAIN ST AT ALI ST	6	50	2	6	50
1	0		10615	001005	5505	09-Sep-2013	24	N31550	MAIN ST OPP 9TH ST	6	50	19	6	50
1	0		10615	001005	5505	09-Sep-2013	25	N31770	MAIN ST OPP CANAL ST	6	50	35	6	50
1	0		10615	001005	5505	09-Sep-2013	26	N31620	MAIN ST AT #1113	6	50	52	6	50
1	0		10615	001005	5505	09-Sep-2013	27	N31660	MAIN ST OPP 13TH ST	6	51	0	6	51
1	0		10615	001005	5505	09-Sep-2013	28	N31670	MAIN ST OPP 14TH ST	6	51	16	6	51
1	0		10615	001005	5505	09-Sep-2013	29	N31680	MAIN ST OPP 15TH ST	6	51	29	6	51
1	0		10615	001005	5505	09-Sep-2013	30	N31690	MAIN ST OPP 16TH ST	6	51	51	6	51
1	0		10615	001005	5505	09-Sep-2013	31	N31700	MAIN ST OPP 17TH ST	6	52	5	6	52
1	0		10615	001005	5505	09-Sep-2013	32	N31710	MAIN ST OPP 18TH ST	6	52	18	6	52
1	0		10615	001005	5505	09-Sep-2013	33	N31780	MAIN ST AT #1841	6	52	32	6	52
1	0		10615	001005	5505	09-Sep-2013	34	N31720	MAIN ST OPP 19TH ST	6	52	46	6	52
1	0		10615	001005	5505	09-Sep-2013	35	N31730	MAIN ST OPP 21ST ST	6	53	7	6	53
1	0		10615	001005	5505	09-Sep-2013	36	N31750	MAIN ST OPP 23RD ST	6	53	32	6	53
1	0		10615	001005	5505	09-Sep-2013	37	N22810	FREEPORT RD OPP WESTERN A	6	54	32	6	54
1	0		10615	001005	5505	09-Sep-2013	38	N22260	FREEPORT RD OPP CENTER AVE	6	54	53	6	54
1	0		10615	001005	5505	09-Sep-2013	39	N22320	FREEPORT RD OPP EASTERN AV	6	55	6	6	55
1	0		10615	001005	5505	09-Sep-2013	40	N22210	FREEPORT RD OPP BRILLIANT AV	6	55	28	6	55

MIN	SEC	DHR	DMIN	DSEC	ON	OFF	LOAD	DLMILES	DLMIN	DLPMLS	DWTIME	DELTA	SCHTIM	SCHDEV	SRTIME	ARTIME
42	28	6	42	36	2	0	23	2.21	8.2	46.396	0.13	140	636	6.47	99.90	99.90
45	42	6	45	45	2	0	25	1.41	2.7	32.503	0.05	132	9,999	99.00	99.90	99.90
45	59	6	45	59	0	0	25	0.33	0.8	8.335	0.00	15	9,999	99.00	99.90	99.90
46	33	6	46	47	0	2	23	0.05	0.2	1.355	0.23	111	9,999	99.00	99.90	99.90
47	17	6	47	18	1	0	24	0.15	1.0	3.563	0.02	52	640	7.28	4.00	4.68
47	46	6	47	46	0	0	24	0.11	0.6	2.576	0.00	64	9,999	99.00	99.90	99.90
48	39	6	48	40	0	1	23	0.12	0.3	2.779	0.02	93	9,999	99.00	99.90	99.90
49	1	6	49	1	0	0	23	0.19	0.9	4.265	0.00	24	9,999	99.00	99.90	99.90
49	23	6	49	32	3	0	26	0.08	0.4	1.830	0.15	29	9,999	99.00	99.90	99.90
50	2	6	50	3	0	1	25	0.08	0.6	1.973	0.02	146	9,999	99.00	99.90	99.90
50	19	6	50	19	0	0	25	0.06	0.3	1.460	0.00	20	9,999	99.00	99.90	99.90
50	35	6	50	35	0	0	25	0.08	0.3	2.106	0.00	53	644	6.58	4.00	3.28
50	52	6	50	52	0	0	25	0.09	0.3	2.295	0.00	11	9,999	99.00	99.90	99.90
51	0	6	51	0	0	0	25	0.05	0.1	1.350	0.00	21	9,999	99.00	99.90	99.90
51	16	6	51	16	0	0	25	0.09	0.2	2.370	0.00	11	9,999	99.00	99.90	99.90
51	29	6	51	31	1	0	26	0.07	0.2	1.678	0.03	15	9,999	99.00	99.90	99.90
51	51	6	51	51	0	0	26	0.07	0.4	1.910	0.00	13	9,999	99.00	99.90	99.90
52	5	6	52	5	0	0	26	0.09	0.2	2.466	0.00	42	9,999	99.00	99.90	99.90
52	18	6	52	18	0	0	26	0.09	0.2	2.225	0.00	17	9,999	99.00	99.90	99.90
52	32	6	52	32	0	0	26	0.09	0.2	2.463	0.00	18	9,999	99.00	99.90	99.90
52	46	6	52	46	0	0	26	0.11	0.3	2.990	0.00	25	9,999	99.00	99.90	99.90
53	7	6	53	8	1	0	27	0.11	0.3	2.814	0.02	35	9,999	99.00	99.90	99.90
53	32	6	53	32	0	0	27	0.11	0.5	2.858	0.00	4	9,999	99.00	99.90	99.90

Figure 3 section 1

Variable description

Field Name	Short Descriptoin	Field name	Short Descriptoin
DOW	Day of week code	SEC	Arrival Sec
dir	Direction of trip along route	DHR	Departure Hour
ROUTE	Route Code	DMIN	Departure Min
TRIPA	Trip Number? Need to double check	DSEC	Departure Sec
BLOCKA	I think these are for scheduling bus drivers. Need to double check	ON	Observed Number of Passengers Boarding
VEHNOA	Vehicle Number	OFF	Observed Number of Passengers Alighting
daymoyr	Day/Month/Year of run	LOAD	Number of Passengers on Bus
STOPA	stop sequential number	DLMILES	Miles travelled from last stop
QSTOPA	PAAC stop alpha numeric ID number	DLMIN	Minutes travelled from last stop
ANAME	Stop Name	DLPMLS	Change in passenger miles from last stop
HR	Arrival Hour	DWTIME	Can't find any documentation on this. Need to ask
Min	Arrival Min	DELTA	Distance in feet from observed GPS coordinates of the record to GPS coordinates for the stop

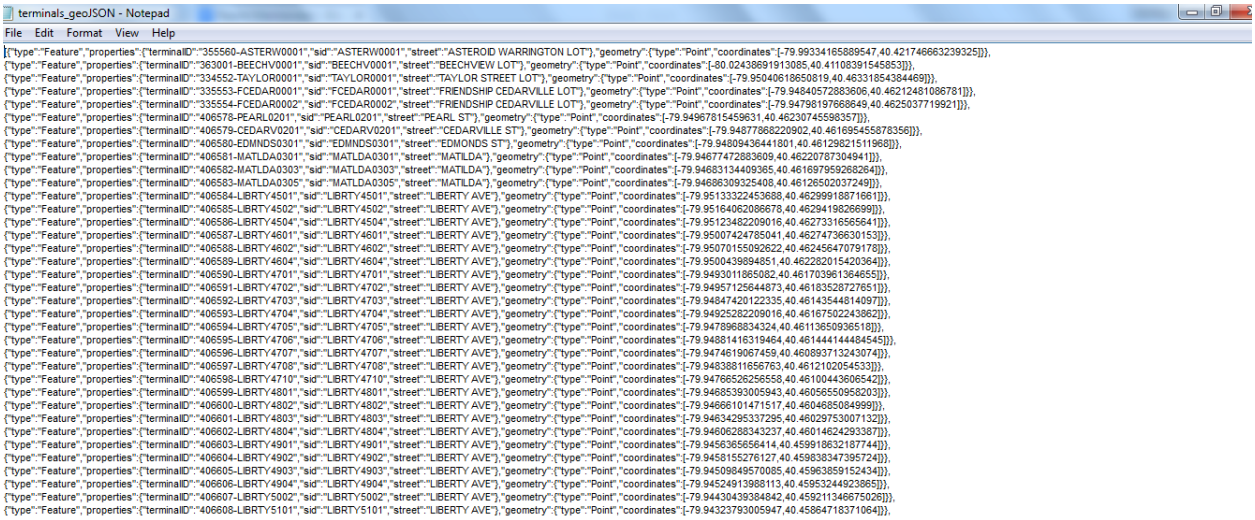
Figure 4 section 1

Pittsburgh Parking Terminals

Location: Pittsburgh
Time: 2013 and 2014

The data (shown in the snapshot below) is for Pittsburgh downtown parking meter occupancy contains all Downtown Pittsburgh parking terminals identified by "terminal ID". The data file given to the team was in geoJSON format and PPA_terminals contains parking occupancy data for entire Pittsburgh

Pittsburgh parking terminals data file in geoJSON format


terminals_geoJSON - Notepad

File Edit Format View Help

{"type":"Feature","properties":{"terminalID":"355560-ASTERW0001","sid":"ASTERW0001","street":"ASTEROID WARRINGTON LOT"},"geometry":{"type":"Point","coordinates":[-79.99334165889547,40.4217466663239325]}},
{"type":"Feature","properties":{"terminalID":"363001-BEECHV0001","sid":"BEECHV0001","street":"BEECHVIEW LOT"},"geometry":{"type":"Point","coordinates":[-80.02438691913085,40.41108391545853]}},
{"type":"Feature","properties":{"terminalID":"334552-TAYLOR0001","sid":"TAYLOR0001","street":"TAYLOR STREET LOT"},"geometry":{"type":"Point","coordinates":[-79.95040618650819,40.46331854384469]}},
{"type":"Feature","properties":{"terminalID":"335553-FCEDAR0001","sid":"FCEDAR0001","street":"FRIENDSHIP CEDARVILLE LOT"},"geometry":{"type":"Point","coordinates":[-79.94840572883606,40.46212481086781]}},
{"type":"Feature","properties":{"terminalID":"335554-FCEDAR0002","sid":"FCEDAR0002","street":"FRIENDSHIP CEDARVILLE LOT"},"geometry":{"type":"Point","coordinates":[-79.94798197668649,40.4625037719921]}},
{"type":"Feature","properties":{"terminalID":"406578-PEARL0201","sid":"PEARL0201","street":"PEARL ST"},"geometry":{"type":"Point","coordinates":[-79.94967815459631,40.462307455983 57]}},
{"type":"Feature","properties":{"terminalID":"406579-CEDARV0201","sid":"CEDARV0201","street":"CEDARVILLE ST"},"geometry":{"type":"Point","coordinates":[-79.94877868220902,40.46169455878356]}},
{"type":"Feature","properties":{"terminalID":"406580-EDMNDS0301","sid":"EDMNDS0301","street":"EDMONDS ST"},"geometry":{"type":"Point","coordinates":[-79.94809436441801,40.46129821511968]}},
{"type":"Feature","properties":{"terminalID":"406581-MATLDA0301","sid":"MATLDA0301","street":"MATILDA"},"geometry":{"type":"Point","coordinates":[-79.94677472883609,40.46220787304941]}},
{"type":"Feature","properties":{"terminalID":"406582-MATLDA0303","sid":"MATLDA0303","street":"MATILDA"},"geometry":{"type":"Point","coordinates":[-79.94683134409365,40.46169759268264]}},
{"type":"Feature","properties":{"terminalID":"406583-MATLDA0305","sid":"MATLDA0305","street":"MATILDA"},"geometry":{"type":"Point","coordinates":[-79.94686309325408,40.461265020372491]}},
{"type":"Feature","properties":{"terminalID":"406584-LIBRTY4501","sid":"LIBRTY4501","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.95133322453688,40.46299918871661]}},
{"type":"Feature","properties":{"terminalID":"406585-LIBRTY4502","sid":"LIBRTY4502","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.95164062086678,40.462941982699]}},
{"type":"Feature","properties":{"terminalID":"406586-LIBRTY4504","sid":"LIBRTY4504","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.95123482209016,40.462733165656 41]}},
{"type":"Feature","properties":{"terminalID":"406587-LIBRTY4601","sid":"LIBRTY4601","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.95007424785041,40.462747366301 53]}},
{"type":"Feature","properties":{"terminalID":"406588-LIBRTY4602","sid":"LIBRTY4602","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.95070155092622,40.462456470791 78]}},
{"type":"Feature","properties":{"terminalID":"406589-LIBRTY4604","sid":"LIBRTY4604","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.9500439894851,40.462282015420364]}},
{"type":"Feature","properties":{"terminalID":"406590-LIBRTY4701","sid":"LIBRTY4701","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.9493011865082,40.461703961364655]}},
{"type":"Feature","properties":{"terminalID":"406591-LIBRTY4702","sid":"LIBRTY4702","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.94957125644873,40.461835287276 51]}},
{"type":"Feature","properties":{"terminalID":"406592-LIBRTY4703","sid":"LIBRTY4703","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.94847420122335,40.461435448140 97]}},
{"type":"Feature","properties":{"terminalID":"406593-LIBRTY4704","sid":"LIBRTY4704","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.94925282209016,40.461675022438 62]}},
{"type":"Feature","properties":{"terminalID":"406594-LIBRTY4705","sid":"LIBRTY4705","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.9478968834324,40.46113650936518]}},
{"type":"Feature","properties":{"terminalID":"406595-LIBRTY4706","sid":"LIBRTY4706","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.94881416319464,40.461444144484545]}},
{"type":"Feature","properties":{"terminalID":"406596-LIBRTY4707","sid":"LIBRTY4707","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.9474619067459,40.460893713243074]}},
{"type":"Feature","properties":{"terminalID":"406597-LIBRTY4708","sid":"LIBRTY4708","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.94838811656763,40.46121020545 33]}},
{"type":"Feature","properties":{"terminalID":"406598-LIBRTY4710","sid":"LIBRTY4710","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.94766526256558,40.461004436065 42]}},
{"type":"Feature","properties":{"terminalID":"406599-LIBRTY4801","sid":"LIBRTY4801","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.94685393005943,40.460565509582 03]}},
{"type":"Feature","properties":{"terminalID":"406600-LIBRTY4802","sid":"LIBRTY4802","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.94666101471517,40.460468508499 9]}},
{"type":"Feature","properties":{"terminalID":"406601-LIBRTY4803","sid":"LIBRTY4803","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.94634295337295,40.460297530071 32]}},
{"type":"Feature","properties":{"terminalID":"406602-LIBRTY4804","sid":"LIBRTY4804","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.94606288343237,40.460146242933 87]}},
{"type":"Feature","properties":{"terminalID":"406603-LIBRTY4901","sid":"LIBRTY4901","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.94563656564414,40.459918632187744]}},
{"type":"Feature","properties":{"terminalID":"406604-LIBRTY4902","sid":"LIBRTY4902","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.94581552761 27,40.459838347395724]}},
{"type":"Feature","properties":{"terminalID":"406605-LIBRTY4903","sid":"LIBRTY4903","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.94509849570085,40.459638591524 34]}},
{"type":"Feature","properties":{"terminalID":"406606-LIBRTY4904","sid":"LIBRTY4904","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.94524913988113,40.459532449233 85]}},
{"type":"Feature","properties":{"terminalID":"406607-LIBRTY5002","sid":"LIBRTY5002","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.94430439384842,40.459211346675026]}},
{"type":"Feature","properties":{"terminalID":"406608-LIBRTY5101","sid":"LIBRTY5101","street":"LIBERTY AVE"},"geometry":{"type":"Point","coordinates":[-79.94323793005947,40.458647183710 64]}},

Figure 5 section 1

Pittsburgh Parking terminal converted to table format

Purchase Date Local	Terminal - Terminal ID	Pay Unit - Name	Ticket Nu	Amount	Units	Masked PAN	Pay Interval Start Local	Pay Interval End Local	Tariff Pacl	Article Name	Node	Tariff Package - ID
9/30/2013 23:14	342336-ECARSN0001	Card	22473	1	48	443060*5695	9/30/2013 23:13	10/1/2013 0:01	Regular	Article 1	SouthSide	0
9/30/2013 23:08	402391-3RDAVE0202	Card	12158	3	60	443040*3435	9/30/2013 23:08	10/1/2013 0:08	Pgm12	Article 1	Downtown	12
9/30/2013 23:05	415054-S12TH-0001	Card	6293	1	56	517545*6804	9/30/2013 23:05	10/1/2013 0:01	Regular	Article 1	SouthSide	0
9/30/2013 23:03	415027-ECARSN1703	Card	13955	1	58	517545*2673	9/30/2013 23:03	10/1/2013 0:01	Regular	Article 1	SouthSide	0
9/30/2013 22:59	407113-FORBES3701	Card	20469	2	60	443040*9335	9/30/2013 22:59	9/30/2013 23:59	Regular	Article 1	Oakland	0
9/30/2013 22:58	405008-BUTLER4301	Card	800	2	120	443040*9390	9/30/2013 22:58	10/1/2013 0:58	Pgm1	Article 1	Lawrenceville	1
9/30/2013 22:47	402394-4THAVE0201	Card	8965	4	80	443040*3090	9/30/2013 22:47	10/1/2013 0:07	Pgm12	Article 1	Downtown	12
9/30/2013 22:18	407133-LTHROP0203	Card	9108	3.47	103	449449*5905	9/30/2013 22:18	10/1/2013 0:01	Regular	Article 1	Oakland	0
9/30/2013 22:13	328547-IVYBEL0003	Coin	38952	0.5	30		9/30/2013 22:13	9/30/2013 22:43	Regular	Article 1	ShadySide	0
9/30/2013 22:07	415089-SIDNEY2601	Card	19994	1.9	114	442791*4742	9/30/2013 22:07	10/1/2013 0:01	Regular	Article 1	SouthSide	0
9/30/2013 21:56	402426-COMMON0101	Coin	8064	0.5	10		9/30/2013 21:56	9/30/2013 22:06	Pgm12	Article 1	Downtown	12
9/30/2013 21:55	401359-9THST-0101	Card	7718	6	120	482853*4029	9/30/2013 21:55	9/30/2013 23:55	Pgm28	Article 1	Downtown	28
9/30/2013 21:53	401359-9THST-0101	Card	7717	6	120	517545*6837	9/30/2013 21:53	9/30/2013 23:53	Pgm28	Article 1	Downtown	28

Figure 6 section 1

Table 1: Summary of the Original Dataset[1]

Variable	Description
Purchase Date Local	Purchase date stored in YYYY-MM-DD h:mm:ss AM/PM format
Created in Cwo	[…][2]
Terminal - Terminal ID*	Transaction terminal information stored as the combination of 6 digit terminal identifiers unique to each terminal and alphanumeric name of the terminal [i.e. 410412-FREWST4905]
Pay Unit - Name	Payment method information [Card or Coin]
Ticket Number	Ticket number stored
Amount	Amount paid in dollars
Units	[…]
Masked PAN	[…]
Code	[…]
Pay Interval Start Local*	Pay interval start time stored in YYYY-MM-DD h:mm:ss AM/PM format
Pay Interval End Local*	Pay interval end time stored in YYYY-MM-DD h:mm:ss AM/PM format

INRIX

INRIX is a global SaaS and DaaS company, which provides a variety of Internet services and mobile applications pertaining to road traffic and driver services. INRIX provides historical, real-time traffic information, traffic forecasts, travel times, travel time polygons and traffic count. (References: Inrix website and Wiki)

We had access to information for Pittsburgh and all fields for INRIX data for Northbound, southbound , eastbound, westbound, clockwise and counter clock wise roads in Allegheny county, PA (1930 tmcs). The variable description and snap shot is provided in next page.

Variable name Description

Tmc _code:	The segment of road for which data is being reported.
Measurement timestamp:	Time stamp
Speed:	Traffic speed on this segment
Average_speed:	Historical average speed on the segment
Reference_speed:	Typical speed on this segment under free flow conditions.
Travel_time_minutes:	The time in minutes required to traverse the segment.
Confidence _score:	Real time, historic or referenced speed reported
Cvalue:	A measure of the confidence in the real-time data. (Higher = better

Snapshot of INRIX data

tmc_code	measurement_tstamp	speed	average_speed	reference_speed	travel_time_minutes	confidence_score	cvalue
104N0938	9/3/2013 0:00	36	31	31	0.12	30	100
104+09362	9/3/2013 0:00	28	28	28	0.72	10	0
104-08602	9/3/2013 0:00	31	31	31	1.68	10	0
104P0865	9/3/2013 0:00	63	63	63	0.35	10	0
104+04493	9/3/2013 0:00	36.92	44	44	0.04	26	34.4
104-09187	9/3/2013 0:00	28	28	28	1.02	10	0
104-09451	9/3/2013 0:00	37	37	37	4.05	10	0
104N0927	9/3/2013 0:00	34	34	34	0.16	10	0
104+09203	9/3/2013 0:00	32	32	32	3.76	10	0
104N0959	9/3/2013 0:00	24	24	24	0.06	10	0
104+06630	9/3/2013 0:00	43	43	43	3.01	10	0
104+06237	9/3/2013 0:00	41	41	41	1.19	10	0
104P0945(	9/3/2013 0:00	28	28	28	0.09	10	0
104P0451	9/3/2013 0:00	42	44	44	0.38	30	85.6
104P0861	9/3/2013 0:00	32	32	32	0.05	10	0
104P0955(	9/3/2013 0:00	47	47	47	0.05	10	0
104N0455	9/3/2013 0:00	58	58	58	0.15	10	0
104-04555	9/3/2013 0:00	50	50	50	0.57	10	0
104+04554	9/3/2013 0:00	50	50	50	0.15	10	0
104+09444	9/3/2013 0:00	31	31	31	2.11	10	0
104N0908	9/3/2013 0:00	31	31	31	0.17	10	0
104-09539	9/3/2013 0:00	29	29	29	0.75	10	0
104+09593	9/3/2013 0:00	22	22	22	0.22	10	0

Figure 7 section 1

Appendix III: Data Cleaning Steps

TOOLS USED:

- R Studio
- Weka
- Excel

APC AVL: Automatic passenger counter and automatic vehicle location data

Available data: September 2012, 2013 and 2014. (Refer to Figure 3 and 4 in Section 1). The data set clearly contains many variables. However, looking at it closely the team realized that it does not contain the most important variables that identify the location of the bus stop, which is the geographic coordinates.

Geographic coordinates act as the missing link between finding the tracts through which the bus transits. In order to find this data, we used the Pittsburgh GIS data from the Pittsburgh City website. http://pittsburghpa.gov/dcp/gis/gis-data

Key used: Bus Stop ID

The following files were used to obtain the missing information and combine the datasets

- Census Tracts 2010: Contains all details about census tracts in Pittsburgh
- Neighborhoods: Contains details about all neighborhoods in Pittsburgh
- Census Blocks 2010: Contains details about all blocks and block groups in Pittsburgh
- Port Authority Bus Stops: Contains details about all bus stops and their respective stop IDs in Pittsburgh

Pittsburgh Downtown parking terminals

Available data: September 2013 and 2014. Refer to figure 6 section 1. Missing information about neighborhood and tracts. Data was at a very granular level (Terminal ID level that is a geographic point). Therefore, it needed to be scaled up to tract level. We used the terminal IDs coordinates which acted as the missing link between finding the tracts. In order to find this data, we used the Pittsburgh GIS data from the Pittsburgh city website. Key used to join the data set is coordinates.

http://pittsburghpa.gov/dcp/gis/gis-data

The following files were used to obtain the missing information and combine the datasets

- Census Tracts 2010: Contains all details about census tracts in Pittsburgh
- Neighborhoods: Contains details about all neighborhoods in Pittsburgh
- Census Blocks 2010: Contains details about all blocks and block groups in Pittsburgh

After obtaining the data, we identified 138 major tracts in Pittsburgh, isolated the parking data for all these 138 tracts, and aggregated the number of parking transactions in each tract. This data was very relevant in creating the regression model that will be discussed later.

INRIX

Available data set: September 2011 and 2013. Refer (figure 7 section 1). This data set consists of vehicle speeding data for different TMC road segments and its respective coordinates. Again, the missing data is tract FIPS code. We used the coordinates which acted as the missing link between finding the tracts. In Order to find this data we used the Pittsburgh GIS data from the Pittsburgh city website. Key used to join is coordinates.

http://pittsburghpa.gov/dcp/gis/gis-data

The following files were used to obtain the missing information and combine the datasets

- Census Tracts 2010: Contains all details about census tracts in Pittsburgh
- Neighborhoods: Contains details about all neighborhoods in Pittsburgh
- Census Blocks 2010: Contains details about all blocks and block groups in Pittsburgh

Appendix IV: Models

Logistic Regression

It is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

Applied in NHTS model to predict mode of transport taken by commuter to reach work location.

Multinomial Logistic Regression

It is the linear regression analysis to conduct when the dependent variable is nominal with more than two levels. Thus, it is an extension of logistic regression, which analyzes dichotomous (binary) dependents.

Applied in ACS model to predict mode of transport taken by commuter to reach work location.

Arithmetic Mean

It is simply "mean") of a sample $x_1, x_2, \ldots, x_n$, usually denoted by $\bar{x}$, is the sum of the sampled values divided by the number of items in the sample:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Simple Linear Regression

It is the least squares estimator of a linear regression model with a single explanatory variable, fits a straight line through the set of *n* points in such a way that makes the sum of squared residuals of the model (that is, vertical distances between the points of the data set and the fitted line) as small as possible.

The outcome variable is related to a single predictor. The slope of the fitted line is equal to the correlation between *y* and *x* corrected by the ratio of standard deviations of these variables. The intercept of the fitted line is such that it passes through the center of mass (*x*, *y*) of the data points.

Applied in model used to predict time taken by a commuter to reach his work location.

Appendix V: Code

Code used to predict mode of transport: Pittsburgh_Mode_Choice.R

(Applicable for both the steps of Task 1)

```
library("nnet")
#Function for import csv file
import.csv <- function(filename) {
  return(read.csv(filename, sep = ",", header = TRUE))
}
#Import dataset
pa <- import.csv('mergedPUMS2013.csv')
#Extract only Pittsburgh City data by PUMA code
pit <-  subset(pa, PUMA==01701 | PUMA ==01702)

#Create an empty data frame for later use
final <- data.frame()
#Random Sampling
for(i in 1:1000){
  #Shuffle
  pit <- pit[sample(nrow(pit)),]
  #20%
  #pit.20 <- pit[1:189,]
  #40%
  #pit.40 <- pit[1:377,]
  #60%
  #pit.60 <- pit[1:566,]
  #80%
  #pit.80 <- pit[1:754,]
  #100%
```

```
  pit.100 <- pit
  #Combine the results
  final <- rbind(final,results(pit.100))
}
#Replace missing values
final[is.na(final)] <- 0
#Calulate the mean accuracy as final results
onethousandtimes.hundredpercent <- data.frame(Car=mean(final$Car),
                                              Bus=mean(final$Bus),
                                              StreetCar=mean(final$Streetcar),
                                              Subway=mean(final$Subway),
                                              Railroad=mean(final$Railroad),
                                              Ferryboat=mean(final$Ferryboat),
                                              Taxicab=mean(final$Taxicab),
                                              Motorcycle = mean(final$Motorcycle),
                                              Bicycle = mean(final$Bicycle),
                                              Walked = mean(final$Walked),
                                              Worked.at.home = mean(final$Worked.at.home),
                                              Other.method = mean(final$Other.method)
                                              )


#Function for the model
results <- function(df) {
  #Variables
  variables <- c("JWMNP","FMRGIP","MIG","PINCP","COW","JWRIP","INTP","FLANXP","PWGTP","SPORDER","PERNP","HINS2","TYPE","FS","RACASN","RNTM","JWTR")
  #Subset the dataset by provided variables
  df <- df[variables]
  #Data processing
```

```
df <- subset(df, JWTR != "")
df[is.na(df)] <- 0
df$FMRGIP <- as.factor(df$FMRGIP)
df$MIG <- as.factor(df$MIG)
df$COW <- as.factor(df$COW)
df$JWRIP <- as.factor(df$JWRIP)
df$FLANXP <- as.factor(df$FLANXP)
df$HINS2 <- as.factor(df$HINS2)
df$TYPE <- as.factor(df$TYPE)
df$FS <- as.factor(df$FS)
df$RACASN <- as.factor(df$RACASN)
df$RNTM <- as.factor(df$RNTM)
df$JWTR <- as.factor(df$JWTR)
private <- subset(df,JWTR == "1")
public  <- subset(df,JWTR != "1")

#Function for 10 folders cross validation
smoothCV <- function(x, y, K = 10) {
  result <- data.frame()
  x <- c(x)
  y <- c(y)
  df <- data.frame(x,y)
  df <- df[sample(nrow(df)),]
  folds = cut(1:nrow(df),breaks=K,labels=FALSE)
  for(j in 1:K){
    train <- df[folds!=j,]
    test <- df[folds==j,]
    fit <- multinom(JWTR ~ ., data = train,MaxNWts=8268)
    predicted2 <- predict(fit,test,type="class")
```

```
    result <- rbind(result,data.frame(pred = predicted2, true = test$JWTR))
  }
  return(result)
}
#Run cross validation
x <- df[,names(df) != "JWTR"]
y <- data.frame(df$JWTR)
colnames(y) <- c("JWTR")
result <- smoothCV(x, y,10)
#Create variables for later use
Car <- 0
Bus <- 0
Streetcar <- 0
Subway<- 0
Railroad<- 0
Ferryboat<- 0
Taxicab<- 0
Motorcycle<- 0
Bicycle<- 0
Walked<- 0
Worked.at.home<- 0
Other.method<- 0
#Accumulate the occurrences of each transportation tool
for (i in 1:nrow(result)){
  if(result$true[i] == "1" && result$pred[i] == "1"){
    Car <- Car+1
  }else if (result$true[i] == "2" && result$pred[i] == "2"){
    Bus <- Bus+1
  }else if (result$true[i] == "3" && result$pred[i] == "3"){
```

```
    Streetcar <- Streetcar+1
  }else if (result$true[i] == "4" && result$pred[i] == "4"){
    Subway <- Subway+1
  }else if (result$true[i] == "5" && result$pred[i] == "5"){
    Railroad <- Railroad+1
  }else if (result$true[i] == "6" && result$pred[i] == "6"){
    Ferryboat <- Ferryboat+1
  }else if (result$true[i] == "7" && result$pred[i] == "7"){
    Taxicab <- Taxicab+1
  }else if (result$true[i] == "8" && result$pred[i] == "8"){
    Motorcycle <- Motorcycle+1
  }else if (result$true[i] == "9" && result$pred[i] == "9"){
    Bicycle <- Bicycle+1
  }else if (result$true[i] == "10" && result$pred[i] == "10"){
    Walked <- Walked+1
  }else if (result$true[i] == "11" && result$pred[i] == "11"){
    Worked.at.home <- Worked.at.home+1
  }else if (result$true[i] == "12" && result$pred[i] == "12"){
    Other.method <- Other.method+1
  }
}
#Calculate the accuracy
car <- subset(df, JWTR == "1")
car.accurary <- Car/nrow(car)
bus <- subset(df, JWTR == "2")
bus.accurary <- Bus/nrow(bus)
streetcar <- subset(df, JWTR == "3")
streetcar.accurary <- Streetcar/nrow(streetcar)
subway <- subset(df, JWTR == "4")
```

```
subway.accurary <- Subway/nrow(subway)
railroad <- subset(df, JWTR == "5")
railroad.accurary <- Railroad/nrow(railroad)
ferryboat <- subset(df, JWTR == "6")
ferryboat.accurary <- Ferryboat/nrow(ferryboat)
taxicab <- subset(df, JWTR == "7")
taxicab.accurary <- Taxicab/nrow(taxicab)
motorcycle <- subset(df, JWTR == "8")
motorcycle.accurary <- Motorcycle/nrow(motorcycle)
bicycle<- subset(df, JWTR == "9")
bicycle.accurary <- Bicycle/nrow(bicycle)
walked<- subset(df, JWTR == "10")
walked.accurary <- Walked/nrow(walked)
worked.at.home<- subset(df, JWTR == "11")
worked.at.home.accurary <- Worked.at.home/nrow(worked.at.home)
other.method<- subset(df, JWTR == "12")
other.method.accurary <- Other.method/nrow(other.method)
accu <- data.frame(Car = car.accurary,
                   Bus =bus.accurary,
                   Streetcar =streetcar.accurary,
                   Subway =subway.accurary,
                   Railroad =railroad.accurary,
                   Ferryboat =ferryboat.accurary,
                   Taxicab =taxicab.accurary,
                   Motorcycle =motorcycle.accurary,
                   Bicycle =bicycle.accurary,
                   Walked =walked.accurary,
                   Worked.at.home =worked.at.home.accurary,
                   Other.method =other.method.accurary
```

```
 )
 return(accu)
}
```

Code used to get the FIPS code of stop from the Census API: getStopFIPS.py

```
import re
import requests
from bs4 import BeautifulSoup, Tag
import csv

def main():
	a = 1
	map = {}
	with open('stopIDFIPS', 'w', newline='') as fp:
		a = csv.writer(fp, delimiter=',')
		title = ["stopID","FIPS"]
		a.writerow(title)
		with open('PAAC_Stops_1603_Public2.csv','r') as csvfile:
			readCSV = csv.reader(csvfile, delimiter=',')
			for row in readCSV:
				if a==1:
					a=2
					continue
				all = row[4]+"+"+row[5]
				value = map.get(all)
				if value == None:

	url="http://data.fcc.gov/api/block/find?format=json&latitude="+row[4]+"&longitude="+row[5]+"
&showall=true"
					html = requests.get(url).content
					htmltxt = BeautifulSoup(html, 'html.parser')

	matchobj=re.findall(r'{"Block":{"FIPS":"(\d+)"},"County":',str(htmltxt))
					if len(matchobj)>0:
						map[all] = matchobj[0]
						value = matchobj[0]
				if value!=None:
					data=[row[0],value]
					a.writerow(data)

main()
```

Code used to filter mean travel time from tract level to county level: filter.py

```
import re
import requests
from bs4 import BeautifulSoup, Tag
import csv
def main():
    a = 1
    begin = set()
    map = {}
    PUMAMap = {}
    with open('filteredMeanTT.csv', 'w', newline='') as fp:
        a = csv.writer(fp, delimiter=',')
        title = ["PUMAtract","meanTT"]
        a.writerow(title)
        with open('NeighborhoodPuma.csv','r') as csvfile:
            readCSV = csv.reader(csvfile, delimiter=',')
            for row in readCSV:
                if a==1:
                    a=2
                    continue
                puma = row[0]
                tract = row[2][0:11]
                begin.add(tract)
                PUMAMap[tract] = puma
        with open('FIPSMeanTT.csv','r') as input:
            readInput = csv.reader(input,delimiter=',')
            for row in readInput:
                try:
                    FIPS = row[0]
                    time = float(row[1])
                    FIPSstart = FIPS[0:11]
                except:
                    continue
                if FIPSstart in begin:
                    pumaValue = PUMAMap.get(FIPSstart)
                    value = map.get(pumaValue)
                    if value==None:
                        map[pumaValue]=[time,1]
                    else:
                        map[pumaValue] = [value[0]+time,value[1]+1]
        for key, value in map.items():
            meanTT = value[0]/value[1]
            data = [key,meanTT]
            a.writerow(data)
```

```
main()
```

Code used to calculate the mean time: calculateMeanTT.py

```
import re
import requests
import csv
from os import listdir
from os.path import isfile, join
from collections import deque
def main():
    countmiss = 0
    countsucc = 0
    FIPSMap = {}
    trackMap = {}
    d = deque()
    countMissOff = 0
    with open('FIPSMeanTT.csv', 'w', newline='') as fp:
        a = csv.writer(fp, delimiter=',')
        title = ["FIPS","meanTT"]
        a.writerow(title)
        with open('stopIDFIPS.csv', 'r', newline='') as csvfile1:
            readCSV1 = csv.reader(csvfile1,delimiter=',')
            for row in readCSV1:
                FIPSMap[row[0]]=row[1]
        mypath = '/Users/Yun/Documents/capstone/findTrackPublicMeanTT/StopInfo'
        onlyfiles = [f for f in listdir(mypath) if isfile(join(mypath, f))]
        for file in onlyfiles:
            if file.endswith(".csv"):
                with open(mypath+"/"+file,'r') as csvfile:
                    readCSV = csv.reader(csvfile,delimiter=',')
                    for row in readCSV:
                        if len(row) >20:
                            if row[0]!='DOW':
                                try:
                                    stopID = row[8].strip()
                                    on = int(row[16])
                                    off = int(row[17])
                                    min = float(row[20])
                                except Exception:
                                    continue
                                FIPS = FIPSMap.get(stopID)
```

```
                if FIPS == None:
                    countmiss = countmiss + 1
                    continue
                else:
                    countsucc = countsucc + 1
                    for i in d:
                        trackInfo = trackMap.get(i)
                        if trackInfo == None:
                            trackMap[i] = [min,0]
                        else:
                            if len(trackInfo)==2:
                                trackMap[i] = [trackInfo[0]+min,trackInfo[1]]
                    trackInfoCur = trackMap.get(FIPS)
                    if trackInfoCur == None:
                        trackMap[FIPS] = [0,0]
                        trackInfoCur = [0,0]
                    trackMap[FIPS] = [trackInfoCur[0],trackInfoCur[1]+on]
                    while on>0:
                        d.append(FIPS)
                        on = on - 1
                    while off>0:
                        if len(d)==0:
                            countMissOff = countMissOff+off
                            break
                        else:
                            d.popleft()
                            off = off - 1
                            while countMissOff>0:
                                if len(d)>0:
                                    d.popleft()
                                    countMissOff = countMissOff-1
                                else:
                                    break
        for key, value in trackMap.items():
            if value[1] == 0:
                continue
            else:
                meanTT = value[0]/value[1]
                print(key,value)
                data = [key,meanTT]
                a.writerow(data)
        print(countmiss,countsucc)

main()
```

Code used to predict the travel-time: travel_time.R

```
Lm (formula = X$Avg_TT ~X$Public_MTT + X$Parking_Transactions + X$Population_Total + X$Means_of_Transport_Total, data = X)
```

References

- National Household Travel Survey Data Analysis:
- U.S. Census Bureau projections to 2060 methodology: http://www.census.gov/population/projections/data/national/2014.html
- U.S. Census Bureau projections to 2060 summary tables: http://www.census.gov/population/projections/data/national/2014/summarytables.html
- Summary report about the U.S. Census Bureau projections to 2060: http://www.census.gov/content/dam/Census/library/publications/2015/demo/p25-1143.pdf
- ACS Longitudinal Employer-Household Dynamics: Job-to-Job Flows (J2J) Data (Beta) Data Files: http://lehd.ces.census.gov/data/j2j_beta.html
- ACS Longitudinal Employer-Household Dynamics: Job-to-Job Flows (J2J) Data (Beta) Data dictionary: http://lehd.ces.census.gov/data/schema/V4.1c-draft/lehd_public_use_schema.html
- Measuring Commuting in the American Time Use Survey: http://bae.uncg.edu/econ/files/2015/02/Kimbrough-working-paper-15-02.pdf
- Ohio TIMES: transportation information mapping system http://gis.dot.state.oh.us/tims/
- 2011 Oregon Household Activity Survey: http://www.oregon.gov/ODOT/TD/TP/Pages/Data.aspx
- The Impacts of Socio-Economic and Demographic Shifts in Transit Served Neighborhoods On Mode Choice And Equity by Steven Apell
- Demographic Trends The Effects of Demographic Change on Selected Transportation Services and Demand by Murdock SH, Cline ME, Zey M, Perez D,& Jeanty PW
- Mode Choices of Millennials: How Different? How Enduring? by Robert B. Case, PE, PhD, Seth Schipinski
- Travel Time Use Over Five Decades by Chen Song & Chao Wei
- Northeast Travel Choice Survey (NTCS) University of Vermont's Transportation Research Center (data not public)
- 2015 Urban Mobility Scorecard published jointly by the Texas A&M Transportation Institute and INRIX
- INRIX Interface Guide December 2014 published by INRIX
- Impact of Traffic Congestion on Bus Travel Time in Northern New Jersey by Claire E. McKnight, Herbert S. Levinson, Kaan Ozbay, Camille Kamga, and Robert E. Paaswell
- Commuting in America Report III - The Third National Report on Commuting Patterns and Trends by Alan E. Pisarski
- R Studio Tool: https://www.rstudio.com/products/rstudio/download/
- Weka Tool: https://weka.waikato.ac.nz/explorer
- National Household Travel Survey Data:
- INRIX data: Heinz College, Carnegie Mellon University Research Centre
- Pittsburgh Downtown Parking Terminals: Heinz College, Carnegie Mellon University Research Centre
- Automatic passenger counter and automatic vehicle location data: Heinz College, Carnegie Mellon University Research Centre

- Pittsburgh GIS information: http://pittsburghpa.gov/dcp/gis/gis-data
- Description of Statistical Models: Wikipedia and www.statisticssolutions.com
- Goodness of Fit explanation for statistical model: http://www.medicine.mcgill.ca/epidemiology/joseph/courses/EPIB-621/fit