Analyzing Means of Commute to work in the United States: US Census Data Exploration

Sagar D. Goswami

Data Analytics Engineering, George Mason University

AIT 580: Analytics: Big Data to Information

Dr. Harry J. Foxwell

December 05, 2021

Analyzing Means of Commute to work in the United States: US Census Data Exploration

This paper is submitted for the fulfillment of the Data Analytics Research Project for the course "AIT-580 Analytics: Big Data to Information."

Abstract

The United States Census Bureau conducts the Annual 1-Year American Census Survey (ACS) and maintains an open Data Repository for Public use. This paper discusses the ACS Census Data to study means of transportation to commute for all Counties and States in the US. Tools like Python, SQL, and R was used for calling ACS Data API (Application Programming Interface), Data Ingestion, Data Cleaning, Data Manipulation, Data Wrangling, Data Exploration, Visualization, and Data Analysis. The insights from this paper can enable stakeholders in the transportation domain and city planners to identify the potential scope of improvement for Public Transportation for a given county.

Introduction

The amount of research content published online that uses "Means of Transportation to Work" – ACS Census Data is significantly less. Thus, there is a broad scope for potential research and untapped insights from analyzing such data. Moreover, with many recent changes in how US Census Bureau publishes the Survey Data, it is becoming possible to analyze updated data more frequently and with more scope for analysis. One of the published literature concerning the Means of Transportation data is Dafeng Xu, which uses repeated cross-sectional survey data to analyze and discuss the relation between US Immigrant Length of Stay and Public Transit Use (Xu, 2018). Apart from it, the Transportation Research Board, US Census Bureau published a guideline which explains using the ACS Survey Data for transportation planning (Transportation Research Board., 2007). Another article on transportation analysis, planning, and implementation published in the

Journal of Economic and Social Measurement discusses using Census Information from the transportation community's perspective (Kristena & Jeffa, 2006).

This paper distinguishes from the other publications by exploring Census Data to generate metrics and thus setting a framework for future work, which will focus on comparing the Public Transportation Usage across various regions, states, counties, and even census tracts. It is achieved by training a linear model between No of people using Public Transportation and the population data. The research is carried out in the following steps:

1. Data Ingestion (Python)
2. Data Cleaning (Python)
3. Data Manipulation / Wrangling (Python)
4. Data Exploration (SQL/R)
5. Data Analysis (R)

Research Methodology

Data Ingestion

The US Census Bureau provides Census Data API (Application Programming Interface) that allows the public entities to access raw Census Data in aggregate and microdata formats (US Census Bureau, n.d.).

> An Application Programming Interface is a set of functions that allows applications to access data and interact with external software components, operating systems, or microservices (bigcommerce, n.d.).

For the purpose of this research, ACS Subject data [2011-2019] was accessed using Python by following the below API Link:

https://api.census.gov/data/2011/acs/acs1?get=group(B08301)&for=county:*

The above link can be interpreted as the 2011 American Census Survey 1-Yr Subject Data for (B08301) group of variables with geography data by county. The B08301 group of variables contains 84 metrics of census data related to "Means of Transportation to Work." The names and outline of the 84 metrics returned from the above API is as follows:

Figure 1 group(B08301) 84 variables "Means of Transportation to Work"

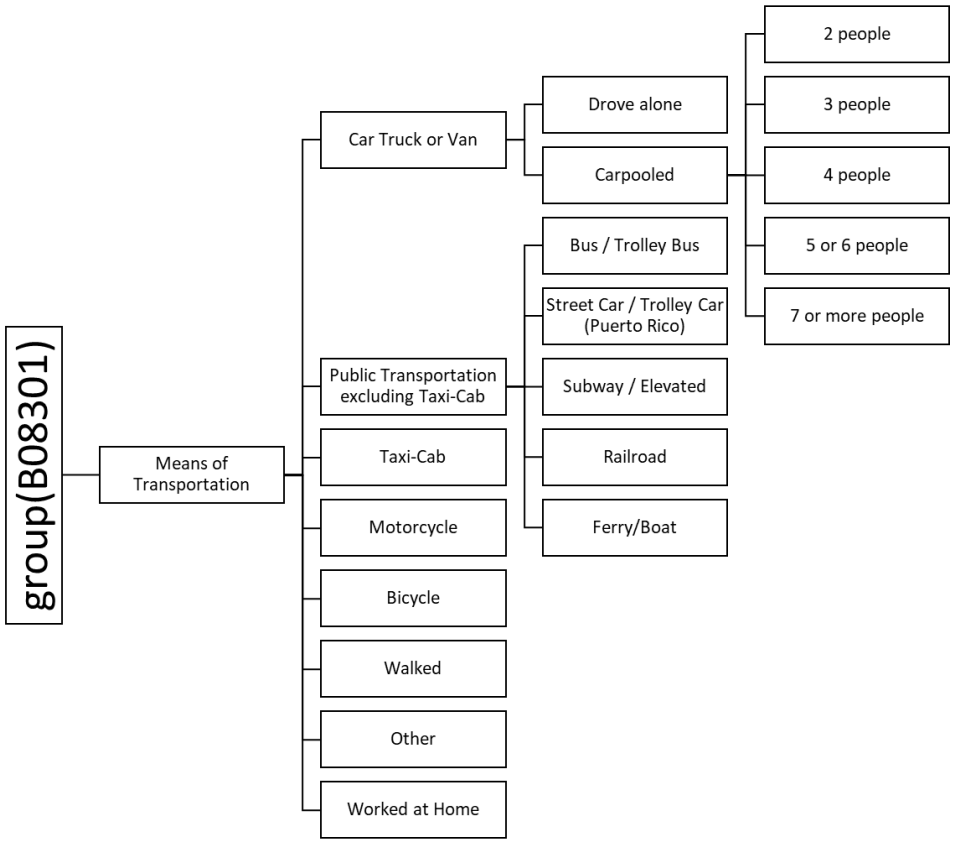


Figure 2 Layout of the Variables in the imported JSON File

The data was obtained in JSON format by calling the above API, then imported as Python Pandas Dataframe. The 84 variables returned have 21 Estimate values for various metrics, 21 pairs of respective Margin of Error, and respective Annotation values for Estimate and Margin of Error Values.

Data Cleaning

This data was then undergone through various cleaning processes using Python, where excess variables were removed, and the column names were renamed. Web-Scraping was carried out to extract the variables' names and replace the codes in the column names in raw data.

Data Manipulation and Wrangling

The above step was carried out for multiple ACS/ACS-1 datasets spanning 2011-2019 and then merged into a single Python Pandas Dataframe. The columns were rearranged, and the "NAME" column was split into two columns, namely 'StateName' and 'CountyName.' The counties for which no data was available were also removed from the Dataframe. Once the Dataframe was ready for analysis, it was separated into two new dataframes for separating State and County Data and exported into '.csv' format for importing it in SQL and R to carry out further analysis.

Data Exploration

The above files were then loaded into SQL for identifying essential features of the dataset, which are given below:

1. Los Angeles County, California, has the Highest Number of people (3.64 Million people) using Car/Truck/Van and driving alone to commute to work. The top 10 counties with the highest number of people commuting using Car/Truck/Van and driving alone are:

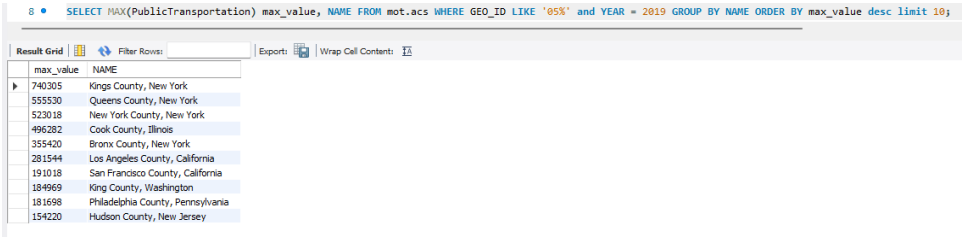
```
7 •  SELECT MAX(CarTruckVanDroveAlone) max_value, NAME FROM mot.acs WHERE GEO_ID LIKE '05%' and YEAR = 2019 GROUP BY NAME ORDER BY max_value desc limit 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

max_value	NAME
3643466	Los Angeles County, California
1815605	Harris County, Texas
1631675	Maricopa County, Arizona
1540712	Cook County, Illinois
1265110	San Diego County, California
1232474	Orange County, California
1034470	Miami-Dade County, Florida
1032499	Dallas County, Texas
847176	Tarrant County, Texas
840598	Clark County, Nevada

Figure 3 Top 10 Counties with the highest number of people commuting by driving alone

2. Kings County, New York, has the highest number of people (740k people) using Public Transport (excluding Taxis/Cab) to commute to work. The top 10 counties with the highest number of people commuting using Public Transport are:



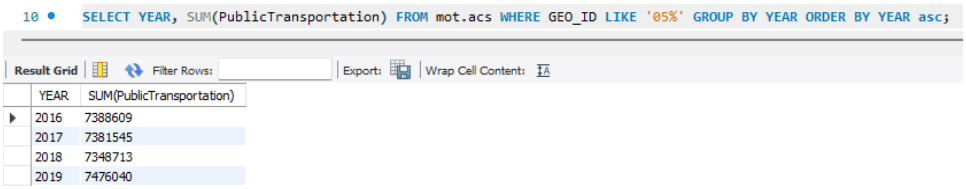
```
8 •  SELECT MAX(PublicTransportation) max_value, NAME FROM mot.acs WHERE GEO_ID LIKE '05%' and YEAR = 2019 GROUP BY NAME ORDER BY max_value desc limit 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

max_value	NAME
740305	Kings County, New York
555530	Queens County, New York
523018	New York County, New York
496282	Cook County, Illinois
355420	Bronx County, New York
281544	Los Angeles County, California
191018	San Francisco County, California
184969	King County, Washington
181698	Philadelphia County, Pennsylvania
154220	Hudson County, New Jersey

Figure 4 Top 10 Counties with the highest number of people using Public Transportation

3. In 2016, around 7.38 million people used Public Transportation systems like Buses, Subway/Metro Rails, Road Trains, Ferry, etc., to commute to work, which rose to 7.47 million people in 2019. The change in the number of people using Public Transportation is as follow:



```
10 •  SELECT YEAR, SUM(PublicTransportation) FROM mot.acs WHERE GEO_ID LIKE '05%' GROUP BY YEAR ORDER BY YEAR asc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

YEAR	SUM(PublicTransportation)
2016	7388609
2017	7381545
2018	7348713
2019	7476040

Figure 5 Trend of No of People using Public Transportation over the years

4. Whereas the number of people using Car/Truck/Van and driving alone to commute was as high as 79.4 million people in 2016, and it rose to 80.4 million in 2019, which is more than 11 times as much as people using Public Transportation. Its change over the years can be noted as follows:

```
11 •  SELECT YEAR, SUM(CarTruckVanDroveAlone) FROM mot.acs WHERE GEO_ID LIKE '05%' GROUP BY YEAR ORDER BY YEAR asc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

YEAR	SUM(CarTruckVanDroveAlone)
2016	79481855
2017	80453846
2018	80746699
2019	80471443

Figure 6 Trend of No of People using Car/Truck/Van and Driving Alone, over the year

5. The counties with the most significant mean of people commuting by Public Transport are as follows. Most of these counties are concentrated in New York, California, Massachusetts, Pennsylvania, and New Jersey.

```
14 •  SELECT NAME, AVG(PublicTransportation) AVG_PUBLIC FROM mot.acs WHERE GEO_ID LIKE '05%' GROUP BY NAME ORDER BY AVG_PUBLIC desc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

NAME	AVG_PUBLIC
Kings County, New York	741501.7500
Queens County, New York	570382.2500
New York County, New York	527075.0000
Cook County, Illinois	484238.5000
Bronx County, New York	351147.7500
Los Angeles County, California	281136.0000
San Francisco County, California	180094.7500
Philadelphia County, Pennsylvania	170370.7500
King County, Washington	168205.5000
Hudson County, New Jersey	156649.0000
Suffolk County, Massachusetts	140553.7500
District of Columbia, District of Co...	128774.0000
Alameda County, California	125488.7500
Nassau County, New York	115179.5000
Westchester County, New York	110314.0000
Middlesex County, Massachusetts	108206.7500
Montgomery County, Maryland	80796.7500
Essex County, New Jersey	79750.2500
Bergen County, New Jersey	76229.7500
Prince George's County, Maryland	70341.0000
Richmond County, New York	64045.5000
Miami-Dade County, Florida	60480.2500
Contra Costa County, California	59697.2500
Allegheny County, Pennsylvania	59695.0000

Figure 7 Mean no of people commuting by Public Transport for given Counties

Data Visualization

R Programming and popular R libraries (tidyverse, choropleth, maps, ggplot2) were used to create a spatial representation of the data instead of (tidycensus and tigris) libraries for simplicity and ease use. The ACS Dataset was imported into an R Dataframe from the '.csv' format. The rows with state data were filtered and converted into a lower case. Maps dataset was used to import spatial data for states of the United States. The imported ACS Dataset was then joined with the spatial map data by 'left_join().' 'ggplot' was used to create a map of the United States to visualize the following ACS Census Data:

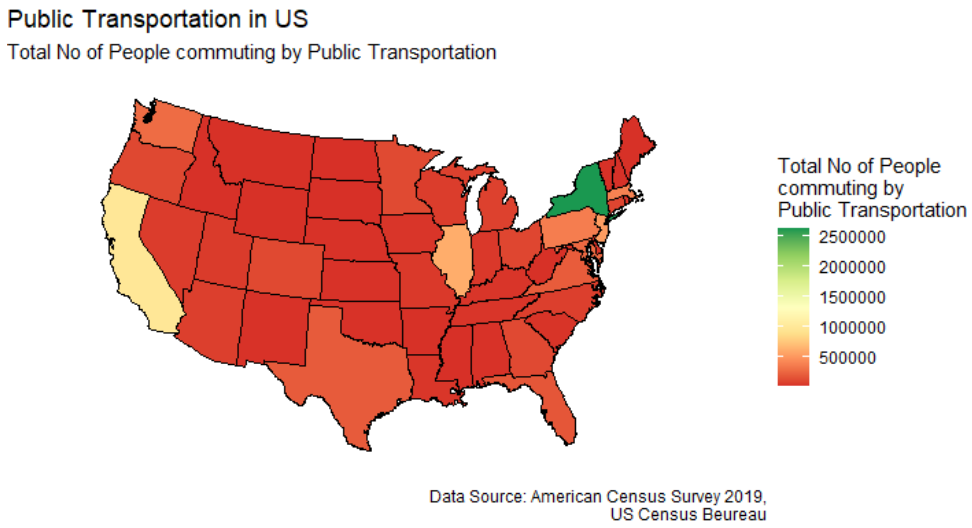


Figure 8 Map(States): No of People using Public Transportation

From the above map, it can be observed that the highest no of people using Public Transportation is in New York, followed by California and Illinois.

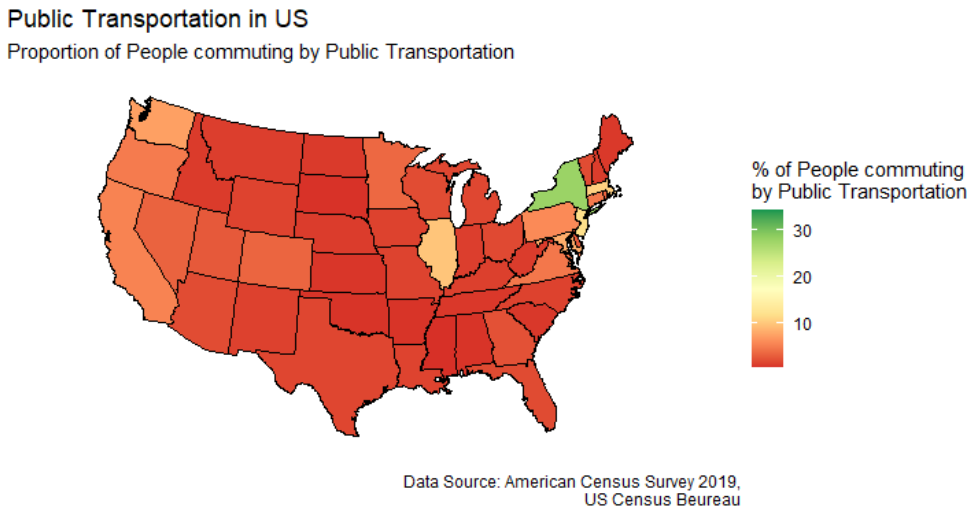


Figure 9 % of people using Public Transportation

From the above maps, it can be observed that the highest % of people using Public Transporation is again New York, followed by Illinois and Washington this time

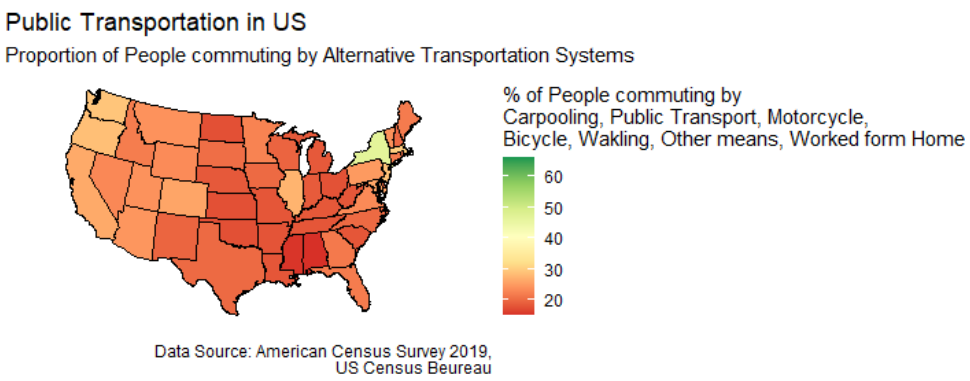


Figure 10 % of people using Alternative Transportation Systems

From the above map, it can be observed that the state with the highest no of people using the Alternative Transportation System is again New York, with Illinois, Washington, and Oregon in the following.

Data Analysis

A linear model was trained in R based on the Population and Public Transportation Usage Statistics from the ACS Census Dataset to analyze the Public Transportation across the United States. The relationship between Population and Public Transportation usage can be explained as follows:

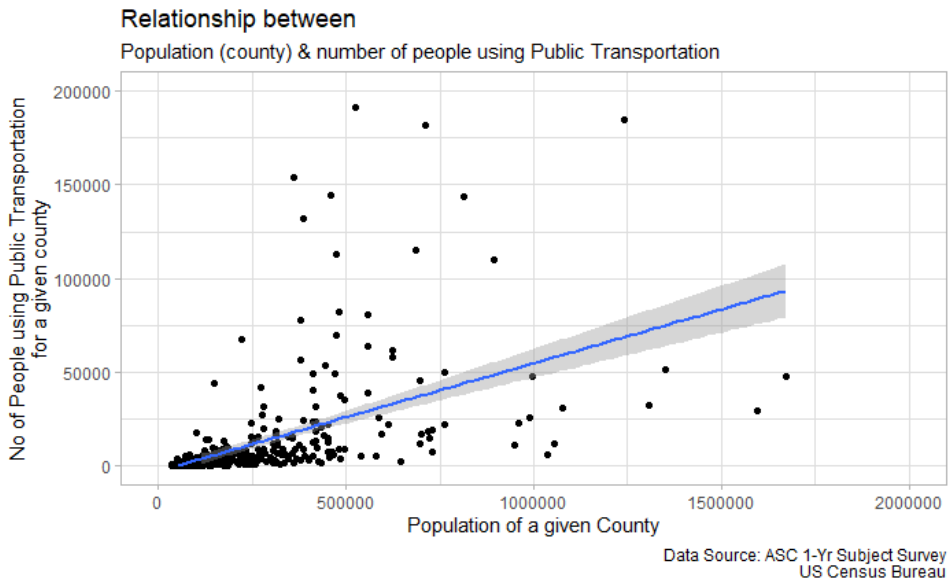


Figure 11 Relationship between Population and No of people using Public Transportation for a county

A positive correlation can be noted between the two entities, and the model can be summarized as follows:

```{r}
lm_model <- lm(formula = PubTran ~ Population,
               data = df2County)
summary(lm_model)
```

```
Call:
lm(formula = PubTran ~ Population, data = df2County)

Residuals:
    Min      1Q  Median      3Q     Max
-157945  -11002   -3564     855  637148

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.614e+03  3.858e+03  -1.714   0.0873 .
Population   9.046e-02  7.782e-03  11.624   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 59190 on 366 degrees of freedom
Multiple R-squared:  0.2696,    Adjusted R-squared:  0.2676
F-statistic: 135.1 on 1 and 366 DF,  p-value: < 2.2e-16
```

Figure 12 Summary of the Linear Model

As seen in the above model, the variables seem to be highly correlated with p-value statistics as low as 2.2×10^{-16}. Using this model, the expected value of Public Transportation Usage can be calculated and later compared with the actual Estimates by the ACS Census Survey. Thus, we can calculate the PT Score (Public Transportation usage Score) for all counties by following simple formula:

Equation 1 PTScore Calculation

$$PT\ Score = \frac{PubTran - E(PubTran)}{PubTran}$$

$$where, \qquad PubTran = \begin{pmatrix} \#\ of\ people\ using\ Public\ Transportation \\ for\ a\ given\ county \end{pmatrix}$$

$$and, \qquad E(PubTran)\ is\ the\ expected\ value, based\ on\ the\ linear\ model.$$

The PT Score determines whether a particular county is doing better or worse compared to all the counties across the United States. (NOTE: The PT Score determines the relative performance of all the counties. It is not an absolute value but can be treated as a ratio-metric for comparing performance metrics of counties.)

The top 10 counties with the best PT Score are:

Table 1 Counties with best PTScores

Name <chr>	PTScore <dbl>
Harnett County, North Caro...	77.283975
Crawford County, Pennsylv...	28.901848
Mercer County, Pennsylvania	24.083743
Rice County, Minnesota	20.684770
Cochise County, Arizona	17.992724
Robeson County, North Car...	11.307564
Wagoner County, Oklahoma	11.207785
Wayne County, Ohio	9.278144
Eau Claire County, Wisconsin	9.217732
Geauga County, Ohio	8.605395

1-10 of 10 rows

The ten counties with the worst PT Scores are:

Table 2 Counties with worst PTScore

Name <chr>	PTScore <dbl>
Brevard County, Florida	-17.23931
Shelby County, Tennessee	-17.89191
Oakland County, Michigan	-19.66476
Pinal County, Arizona	-20.10343
Benton County, Arkansas	-23.81197
Johnson County, Kansas	-24.38626
Madison County, Alabama	-27.82825
Chesterfield County, Virginia	-35.44284
Cleveland County, Oklahoma	-37.66521
Tulare County, California	-40.47746

1-10 of 10 rows

Results

The analysis performed in the previous section can be summarized as follows:

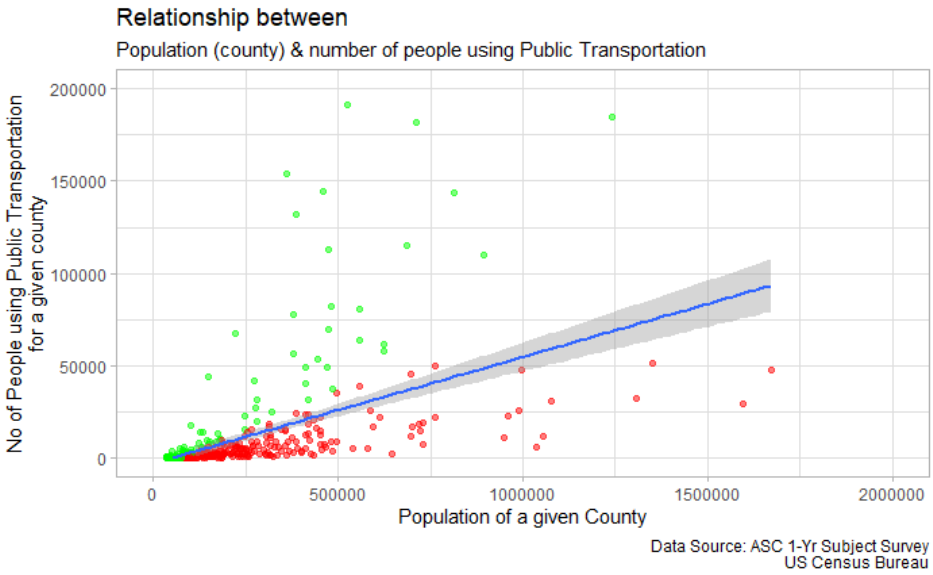


Figure 13 Interpreting Results of the Model

Here, the green dots represent the counties, representing a better Public Transportation Score than other counties in the US. At the same time, the red dots represent the opposite case, with the worst scores compared to other counties in the US. The dots on or closer to the regression line represent the scores reflecting average scores for the entire country.

Table 1 shows the list of counties with the best PTScores for the United States, representing the green dots in Figure 12. Whereas Table 2 shows the list of counties with the worst PTScores and represents the red dots on Figure 12. The results of Table 1 and Table 2 can be further analyzed by performing a validation check with real-life case studies/reports. For example, the Shelby County public transport under the Memphis Area Transport Authority (MATA) scored a very low PTScore of -17.8. Upon further investigation, there were additional insights into why this must have occurred. According to this article, the Memphis Area Transport has been undercut in fundings and improper planning and execution; between 2005 and 2015, the ridership fell by 28%, and MATA cut their services by 22% (Jarrett Walker & Associates, 2019).

Another example of the Brevard County public transport, which comes under the Space Coast Area Transit (SCAT), scored a low-performance Score of -17.2. According to the Space Coast Transport Planning Organization (SCTPO) report, the Brevard County transit ridership decreased by 14.9% from 2018 to 2019 (Kittelson & Associates, Inc., 2019). Contrary to our findings, although the public transport system in Crawford County, Pennsylvania, received a positive performance score of 28.9, according to the Crawford Area Transportation Authority (CATA), the passenger ridership decreased from 13.80% in 2012 to 10.96% in 2018.

According to the presented model, the expected scores of counties are dependent only on the Population Data. The better way to predict expected scores in future works would be to carry out multiple regression models by considering more socio-economic variables like Population Density, Income, Gross Domestic Product, Topological/Geological factors, et Cetra. Also, by using the 5-Yr ACS Data, individual scores can be calculated for individual Census Tracts, thus providing more refined approximations for the population when compared to County Data. It will provide better insight for identifying potential zones for improvement and the strategies that can be implemented for the given value of the predictors.

Conclusion

The ACS Census Survey Data provides invaluable insights on 'Means of Transportation to Work' segregated by counties. By performing Data Analysis on the finest level of data, it is possible to obtain metrics on the adoption of Public Transportation for every census tract and thus identify zones with a high scope of improvement. More variables can be added to the model in future work to obtain further precision in predictions.

References

bigcommerce. (n.d.). *What is an API?* Retrieved from Bigcommerce: https://www.bigcommerce.com/blog/what-is-an-api/#what-is-an-api

Dave Jensen, U.S. Census Bureau. (2021, 03 11). Virginia Means of Transportation to Work by Vehicles Available by Census Tract (ACS 5-Year). *Virginia Open Data Portal*. Virginia, Unites States of America.

Jarrett Walker & Associates. (2019). *Memphis 3.0 Transit Vision Recommended Network.* Memphis.

K. Shvachko, H. K. (2010). "The Hadoop Distributed File System". *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST).* IEEE. doi:10.1109/MSST.2010.5496972

Keller, S. A. (2020, 2 21). Doing Data Science: A Framework and Case Study. *Harvard Data Science Review*(2.1). doi:10.1162/99608f92.2d83f7f5

Kittelson & Associates, Inc. (2019). *2019 State of the System Report : Space Coast Transportation Planning Organization.* Brevard County, Florida.

Kristena, R., & Jeffa, T. (2006, 12 27). Census data for transportation planning, analysis, and implementation. (C. G. Renfro, Ed.) *Journal of Economic and Social Measurement, 31*(3-4).

Transportation Research Board. (2007). *A guidebook for using American Community Survey data for transportation planning.* US Census Bureau.

U.S. Census Bureau. (2021). *An assessment of the COVID-19 pandemic's impact on the 2020 ACS 1-year data : final report.* United States Department of Commerce. Washington, DC: United States Department of Commerce, U.S. Census Bureau.

US Census Bureau. (n.d.). *About: Census Bureau API*. Retrieved from Census.gov: https://www.census.gov/data/developers/about.html

Xu, D. (2018). Transportation assimilation revisited: New evidence from repeated cross-sectional survey data. *PLoS One*.

Table of Figures