

# Hop on the Bus Gus

Explanatory Factors and Recommendations for Increased Public Transit Ridership

# Authors

Robert Eason Jr, Michael Flynn, Mohamad Nayal, Seshu Miriyala, Bibhuti Timalsina,

Ben Reeve

# Abstract

Access to transportation is vital for the personal and economic prosperity of people and to provide this access is part of the fundamental mission of city and municipal government. The fulfilment of that mission has struggled in much of the United States and since the 2020 Covid-19 pandemic, public transportation ridership has fallen even in places where it was once strong. We wish to help public transit administrations find effective avenues for investment that will help them increase their ridership to pre-pandemic level and beyond. This paper presents an exploration of different factors that influence public transportation ridership and recommendations of where to invest public funds. In it, we find that to increase ridership the best value improvements are amenities that make public transportation stops comfortable and safe to use. Ensuring that stations have proper lighting, seating, and protection from the elements are cheap yet effective means of improving usage.

# Specifications

Problem: The main problem specified in this project is about what features of a city’s public transportation system can encourage an increase in the number of riders.

Hypotheses: The more expensive comforts are included in the stations, there is an increase in the numbers of riders.

Data: The source of our data is Infrastructure Database from American Public Transportation Association which provides a report on major transit infrastructure in the US and Canada that includes data on stations, stops and parking for all modes of transport. This database also includes information on rail systems and systems used by public transportation agencies for real- time information. The other source of data is the Monthly Ridership Report from National Transit Database. This database records the financial, operating and asset conditions of American transit systems. This dataset includes information about agencies, types of service and mode of transportation which contains ridership numbers by agency.

<https://www.apta.com/research-technical-resources/transit-statistics/ridership-report/>

<https://www.apta.com/research-technical-resources/transit-statistics/transit-infrastructure-database/>

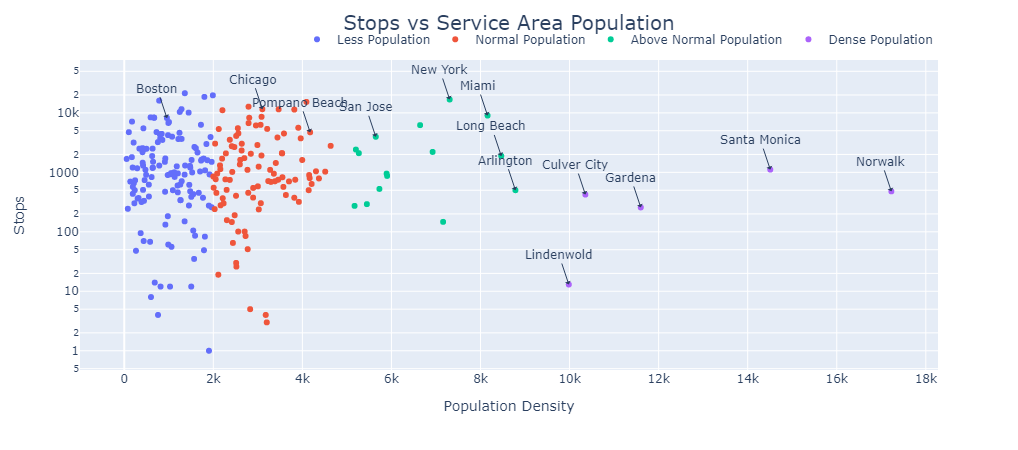
<https://www.transit.dot.gov/ntd/data-product/monthly-module-raw-data-release/>

# Observations:

## Clustering

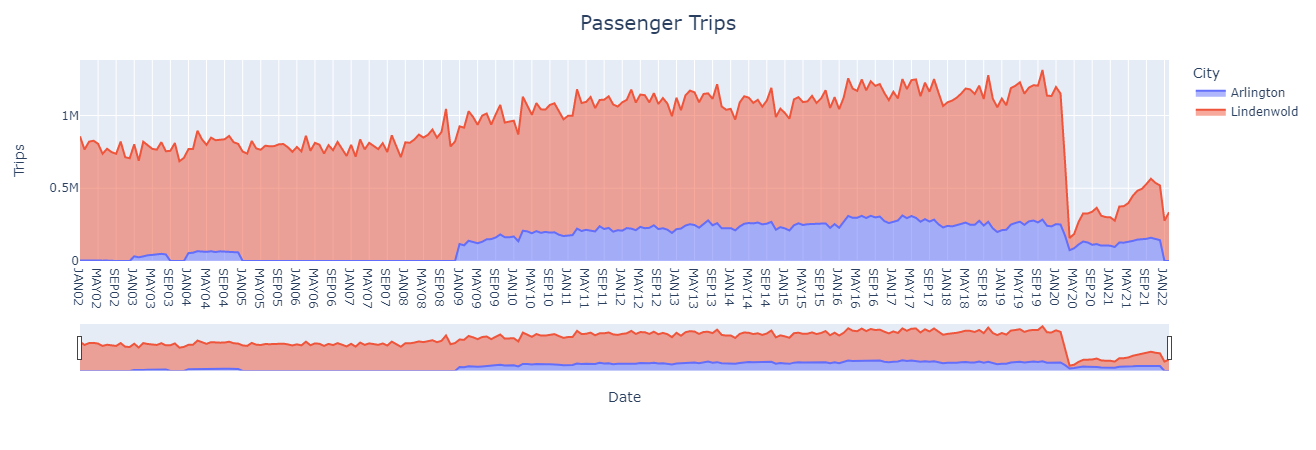
When we divide the cities based on the population density of the agencies that run the transportation system in those cities and the number of stops in those cities into 4 categories by using the K-means algorithm.

We get 4 clusters mainly centered on the population density. So, we named them as Less, Normal, Above Normal and Densely Populated Cities.



The reason we choose the number of Stops is because number of Stops represents how well connected the transportation infrastructure is within that City.

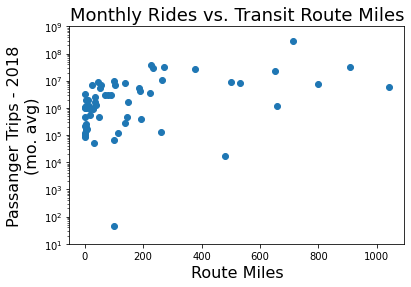
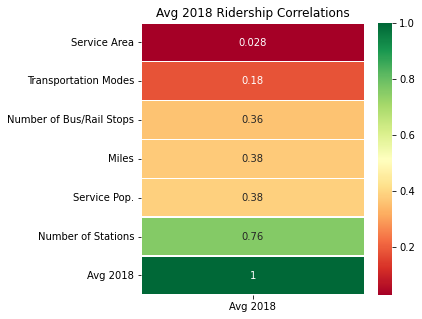
From the chart above, we see that some cities, even though they fall into a similar Population Density, have huge variations in number of Stops (namely Arlington and Lindenwold).

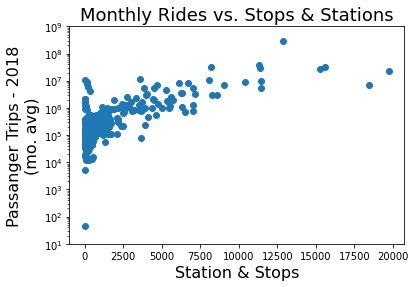
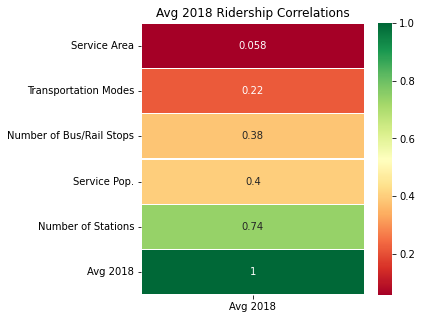


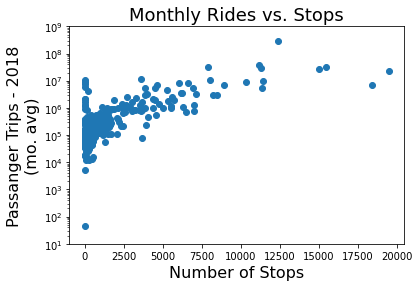
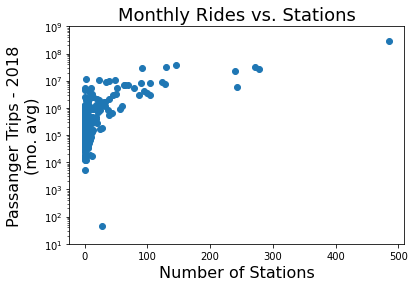
When we plot the Passenger Trips between the two Cities (Arlington and Lindenwold), we see a huge gap. This shows that number of Stops (or Connectivity within the City) places a crucial role in increasing the Passenger Trips.

## Connectivity

Information regarding the breadth of a transit network was separated out from two data sources, “February 2022 Raw Database” and “2018-APTA-Infrastructure-Database," two excel spread sheets. Applicable information included the dependent variable, average 2018 monthly ridership, and several predictor variables; station count, stop count, service area, service population, modes of transportation, route miles and max. speed.

A correlation matrix was built and the correlations to average monthly ridership were singled out. The number of stations being as high as it was fell in line with expectations for this data, but one surprising observation was that the number of route miles was relatively unimportant. Furthermore, when looking at the scatter plot for monthly rides and route miles, our ridership basically plateaus above 200 route miles except for New York which had the highest ridership of any transit system. Route miles would ultimately not be selected for modeling due to the incompleteness of the data available and the remaining correlations were as follows.





Stations and stops were looked at as a combined measure, but because of the drastic difference in magnitude they were kept separate for modeling.

## Passenger Comforts

Infrastructure data from the APTA database was combined with ridership time series data with the goal of predicting ridership based on conveniences or comforts that might improve the overall experience of using public transportation from each agency. These comforts included expected attributes such as the presence of a shelter or bench at a bus stop, the number of restrooms at a train station, or more modern comforts such as Wi-Fi and electric vehicle charging stations in parking lots, and real time updating video displays. The number of unlinked passenger trips per month for each transportation agency for every month in 2018 was averaged to decrease seasonal effects on correlation with comforts. We were concerned using a winter month might inflate the importance of shelters when there might be more inclement weather stand using a summer month might inflate the correlation between ridership and bike racks when more people would tend to ride their bikes for part of their commute. A correlation plot was then created to show average rides per month against attributes related to passenger comforts.

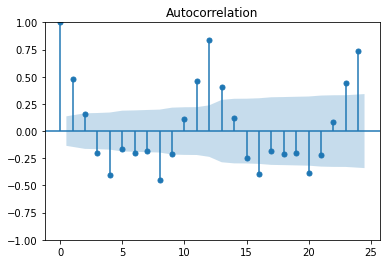
To ensure that our model would not simply recommend every aspect of the New York City / Los Angeles / Chicago transit systems which have far more unlinked passenger trips than smaller service areas the number of average monthly passenger trips was normalized in one of three different ways. Passenger trips were divided either by the population of the service area, the area in square miles of the service area, the population density of the service area per square mile. The intent of this method was to highlight some of the actions of smaller service areas that helped improve their ridership. In a similar attempt to ensure more representation of smaller transit agencies, each comfort related attribute was divided by the number of locations where it could be available. An example of this was dividing the number of elevators a transportation agency had by the number of stations. By doing so we are ensuring that it is the presence of an elevator that we are correlating with ridership, otherwise a large system such as NYC which may have more elevators due to the size of the NYC transit area but fewer stations with elevators might make it appear in our model that adding elevators would increase ridership, when the inputs are only showing that a large transit system is associated with higher ridership. Heatmaps were created for each correlation plot showing ridership normalized either by population, service area, or population density. In the plots below the darkest green and darkest red represent the highest correlation or highest degree of negative correlation.

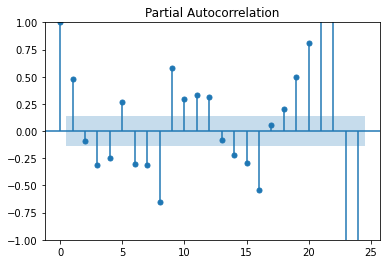
|  |  |
| --- | --- |
|  |  |
|  |  |

Analysis

## Time Series

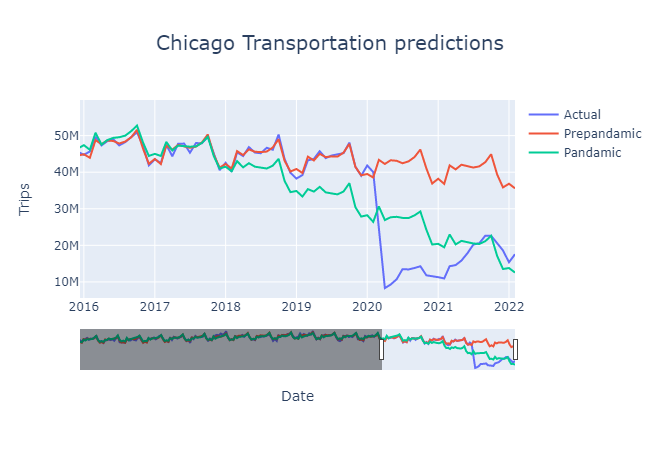
Let us pick a city (like Chicago) and start the time series modeling. The Root Mean Squared Error (RMSE) of ARIMA with (p, d, q) as (9,1,9) from the below ACF and PACF charts is 1231234.1.



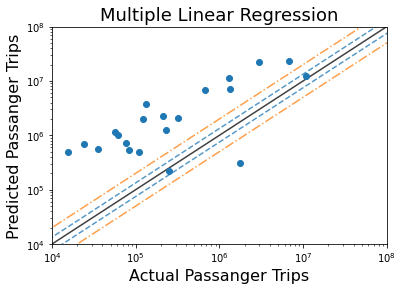


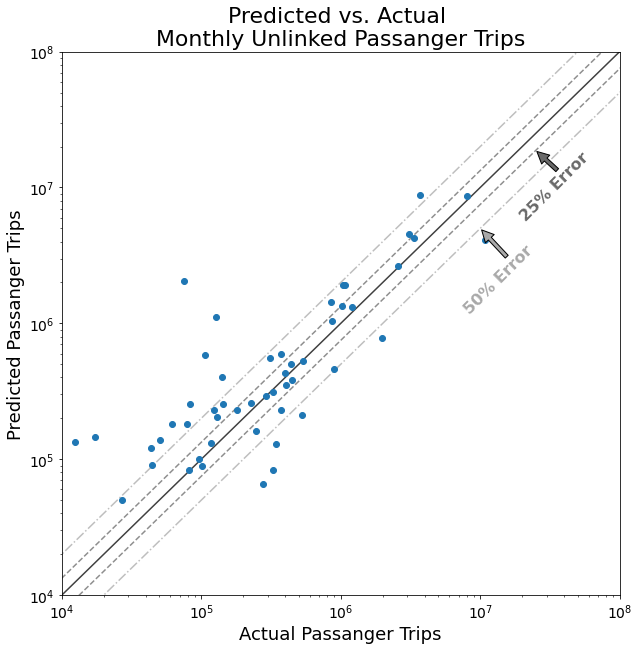
But prophet model gave us better model with a RMSE of 897311.6. So, we used the prophet model to make some predictions.

We used prophet to create a model with the pre-pandemic data (I.e., data before 2020) and with post-pandemic data. Then we used the models to make predictions. The chart below shows the model performance with pandemic data and without pandemic data.



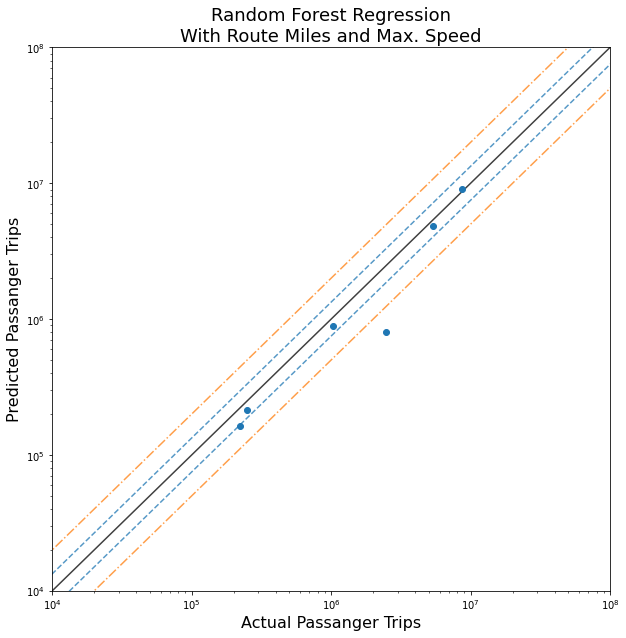
## Connectivity

Three different model types were used in predicting ridership from connectivity measures, linear regression, random forest regression, and a neural net. Using the sklearn linear model package a baseline was created for predictive accuracy based on 5 input measures. With a root mean squared error of nearly 8 million on the test set, and a consistent overestimate, the linear regression model set the bar low. The orange line shows a 50% error in prediction and the blue line shows a 25 % error, across the 51 test samples the predictions were only within the rather large error bounds on two occasions. Next a random forest would be used to see how additional modeling techniques would compare. Using the same base inputs, the sklearn ensemble RandomForestRegressor package showed significant predictive improvement, but still had a rather large deviation and a RMSE of over two million (rides per month).

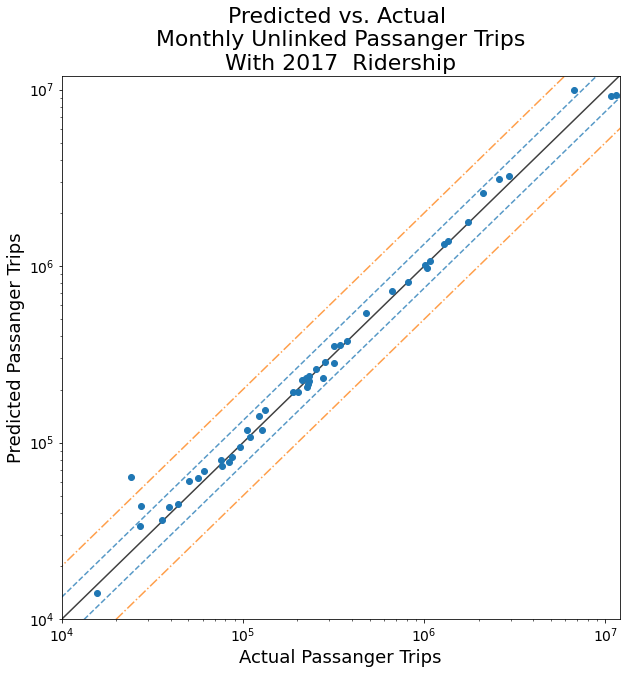


Lastly using Keras from TensorFlow, a neural net was attempted to see if once again accuracy could be improved with a different model. The baseline neural net had a RMSE of roughly four million, putting it in between the two previous models. The creation of the neural net however took significantly more time than the linear regression or the random forest. Because of the mixture of accuracy and low compute time, the random forest was chosen for some additional modeling.

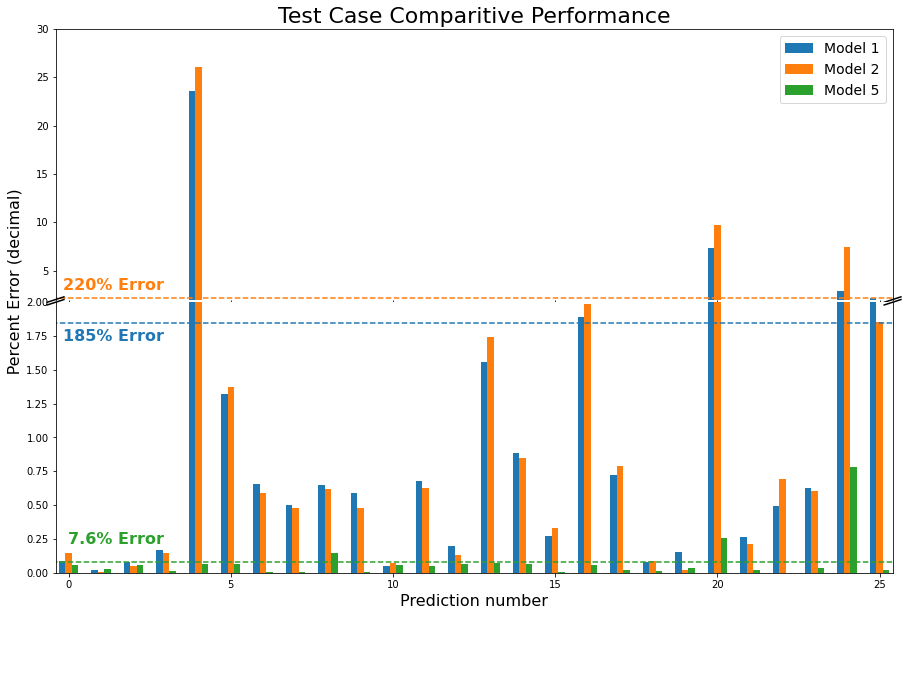
Different random forest models were attempted, reducing the sample size through the inclusion of route miles, removal of top and bottom outliers in ridership, using the log transformations.



The addition of route miles did seem beneficial, but the population was reduced too greatly to make an accurate comparison. Rather than 203 training samples and 51 test samples, this model used 50 training samples and only 6 test samples, so the RMSE of 735,000 monthly trips is not as comparable as a model using the full data set. There was however additional data from these data sets that could aid in the prediction of monthly ridership for the 2018 calendar year, that was the average 2017 monthly ridership. Another random forest was constructed incorporating this data as well and as expected, the predictions were much closer to the actual ridership.



Using the additional predictor variable, the overall RMSE was down to just under 600,000 monthly rides. While that number still seems sufficiently large, some of our monthly ride numbers were on the order of 10M in our test set. Looking at these results another way, based on percent error we could see the drastic improvement adding the 2017 numbers and how close they were.

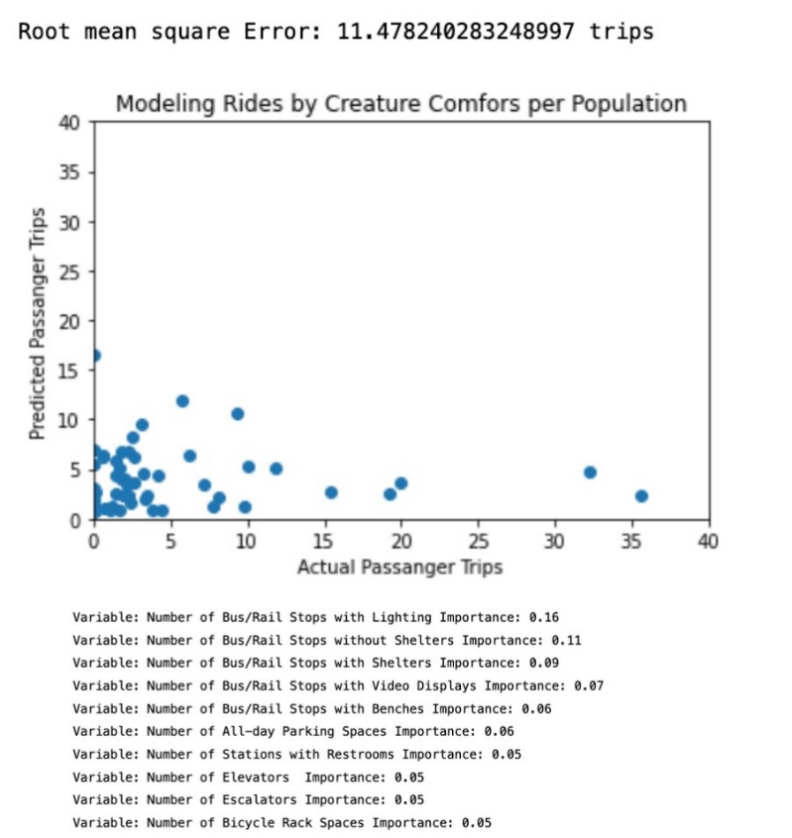


This chart looks at the baseline random forest, Model 1, the model created with the removal of the outliers, Model 2, and the model created with the 2017 ridership, Model 5. The reduction in average error rate for predictions to 7.6% was a huge improvement from any of the other random forest models.

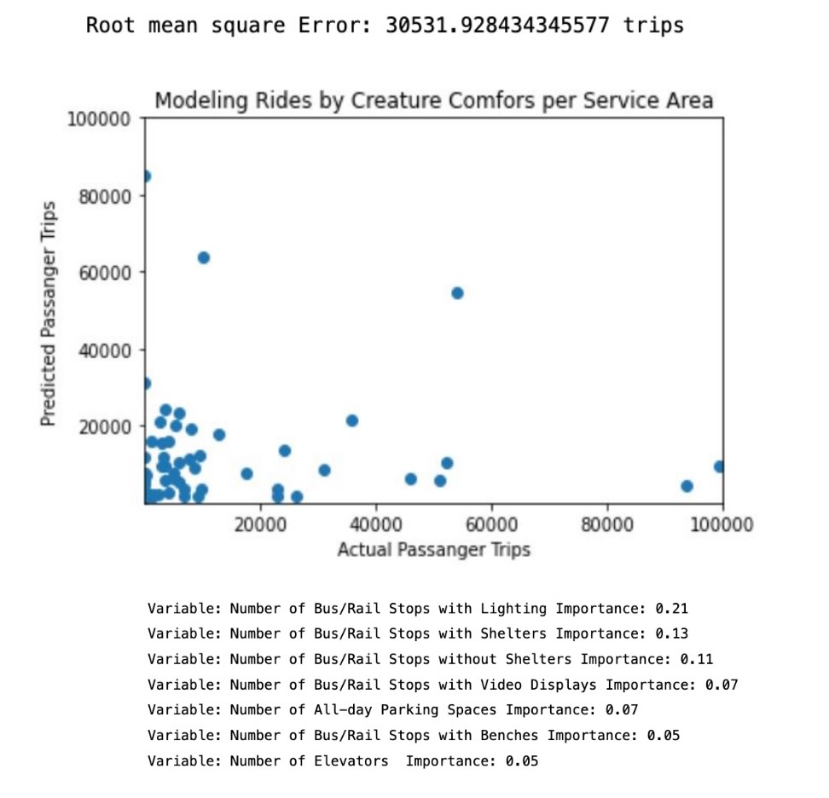
The additional takeaway using the random forest models was the feature importance. In all models built the most important input in generating predictions was the number of stations, followed by the number of stops. This information can help someone in an organization looking to increase ridership. Even more important than your service area population is the number of stations and stops, enabling passengers to get to many places may increase their likeliness to use your transit system. These are costly improvements to make, so to look at some lower budget improvements various creature comforts were explored.

## Comforts

The passenger comforts data frame was used to create a random forest model to predict average monthly passenger trips per person in the service area from all the creature comfort attributes. A similar model was created to predict average monthly passenger trips per person per square mile of the service area. When running this model for average monthly passenger trips normalized to the population density of the service area there were some roadblocks in dividing by zero and predicting attributes with infinity as a value. Excluding these agencies resulted in excluding too many instances and poor model accuracy. A plot of actual monthly passenger trips per person against predicted monthly passenger trips shows that the model is under predicting passenger trips and other attributes are involved. The model had a root mean square error of 11.47. The top contributing factor for the random forest model is the percentage of bus / rail stops with lighting, followed by the percentage of bus stops with shelters, and then percentage of bus stops with video screens.

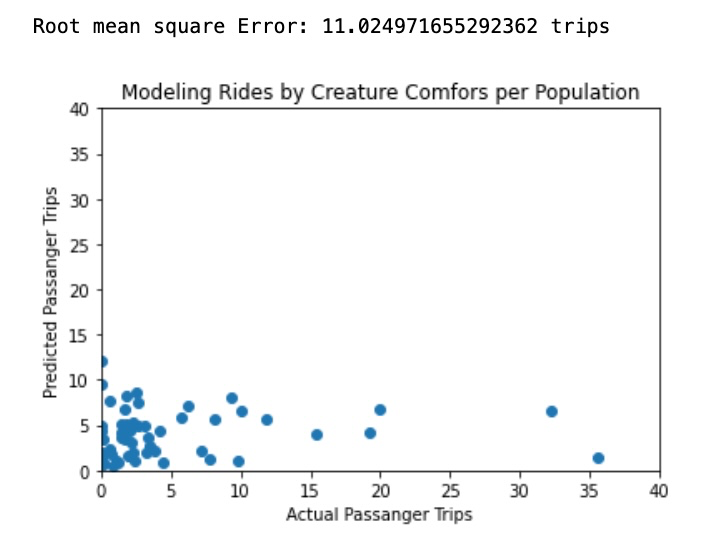


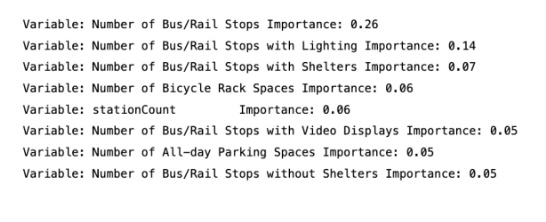
The model for prediction of the number of passenger trips per service area square mile had a much higher mean square error. Future iterations of the model focused on normalizing passenger trips by population.



## Combined Connectivity and Creature Comforts

The unalterable version of our model combined some attributes of the connectivity model with the passenger comfort attributes. This version of the model combined the passenger comforts data frame with a count of the number of stations and the number of bus and rail stops that a transportation Agency had as well as the type of facility including bus, light rail, commuter rail, or ferry. For this model, the historical data from the year 2017 was not added to the model, because the focus of this model was to help provide a list of actionable improvements a transportation agency could make to increase ridership. A random forest was run with these attributes and a plot of prediction average monthly passenger trips per person vs actual average monthly trips is plotted below. Unfortunately, the root means square error did not decrease significantly from combining these two modeling fronts. Root means square only dropped from 11.48 to 11.02. However, the variable impact analysis from this model showed that the number of Bus/ rail stops is the most important attribute for predicting ridership, followed by stops with lighting and then stops with shelters.





# Recommendations

Expanding the transportation grid can be expensive, require extended time for construction and can disrupt current services. Our models here show that some of the strongest influences on ridership come from the availability of cheap upgrades to the current infrastructure. In this project we have shown that the greatest increase in ridership from the lowest investment cost would come from ensuring passengers have lighting in their waiting area, have a bench to sit on, and are partially protected from the elements by a bus stop shelter.

Our models show that the number of bus stops and stations do have an impact on ridership. For cities with the resources to expand their service area, our models predict they will have an increase in their monthly ridership.

# References

1. American Public Transportation Association: Annual Ridership Report

* <https://www.apta.com/research-technical-resources/transit-statistics/ridership-report/>

1. American Public Transportation Association: Station and Infrastructure Report

* <https://www.apta.com/research-technical-resources/transit-statistics/transit-infrastructure-database/>

1. Department of Transportation’s National Transit Database: Monthly Usage Report

* <https://www.transit.dot.gov/ntd/data-product/monthly-module-raw-data-release/>