# What Makes Them Ride?

An Analysis of Public Transit Ridership in 9 American Cities

Raza Lamb

11/29/2021

# Summary

This analysis investigates the various factors that influenced public transportation ridership in 9 major U.S. cities between 2002 and 2019. The goal of the project is to attempt to explain the recent decrease in public transportation ridership seen in many cities. The results of the analysis suggest that seasonal public transportation ridership trends vary greatly by city and that an increasing trend in ridership is largely reversed following the introduction of Uber, even after controlling for other factors.

# Introduction

Public transportation is one of the most important aspects of modern infrastructure. In addition to being a critical driver of economic growth and productivity, it dramatically increases quality of life within major cities by providing access to food, education and entertainment, especially for those with lower socioeconomic status. Investment in public transportation is thus a key public policy issue, especially in the context of the Infrastructure Investment and Jobs Act (P.L. 117-58), which allocates nearly $20 Billion USD in new government spending on transportation.

In fact, public transportation ridership in the U.S. had been decreasing even prior to the COVID-19 pandemic. A 2018 study found that New York City was the only major American city to experience an increase in ridership. There are many potential reasons cited for this trend, including urban sprawl, population decreases, and the introduction of ride-sharing services such as Uber. Regardless, urban planners and other policy makers need to understand how various factors affect transit ridership in order to make appropriate funding and investment decisions. This analysis examines public transportation ridership by month in 9 major American cities in order to determine what factors influence rail ridership. Specifically, the questions of interest are: is there a trend over time after accounting for other factors, does the introduction of Uber in a city affect the ridership trend, and what potentially controllable factors influence ridership the most?

# Data

## Data Preparation

The data required for this analysis was derived from several different sources. The information on public transit systems comes from the National Transit Dataset (NTD), which is a public dataset tracking transportation statistics for all transit agencies in the United States. This is the ideal dataset for this analysis, because all of the statistics are standardized across systems. From this dataset, rail ridership totals (tram, light rail, and heavy rail — commuter rail was not included) per month were extracted for 9 cities: Atlanta, Boston, Chicago, Los Angeles, New York City, Philadelphia, Portland, San Francisco, and Washington, D.C. Also obtained from the NTD were yearly statistics by transit agency including: average fare revenue by mode of service, number of vehicles in operation at maximum service, and directional route miles (length of track). Other city-specific statistics were obtained through various sources. Yearly population and population density estimates were extracted from the United States Census, while labor data (unemployment and labor force participation) was obtained from the Bureau of Labor Statistics (BLS). Daily weather data, inlcuing temperature, precipitation, and snowfall, was scraped from the National Oceanic and Atmospheric Administration (NOAA). Overall monthly average gas price (urban) was obtained from the Federal Reserve Economic Data (FRED) website, along with the monthly consumer price index (CPI). Finally, Uber introduction into a city market was measured by a binary indicator variable, determined by sourcing local news articles and Uber's website.

After obtaining the data, significant work was required to appropriately join all data. First of all the response variable, originally total monthly rail ridership, was divided by the number of days in each month to obtain a more interpretable response: average daily ridership. Subsequently, the yearly transit statistics were joined with the response variable using a many-to-one join. The other yearly statistics, population and population density, were also joined in this manner. The monthly data, including gas price, labor data, and the CPI, were joined using a one-to-one join. The weather data was first converted to monthly averages, and then joined in a similar fashion. The Uber indicator variable was added separately for each city using subsetting. Additionally, a seasonal categorical variable was added based on the month. Finally, the monthly gas prices and yearly average fare per trip were inflation adjusted using the CPI (base year 2010).

## Data Exploration

During this project, exploratory data analysis was made difficult on account of the vast differences between cities. Average daily ridership ranges from approximately 100,000 in Portland to more than 6.5 million in New York City. Notably, New York has ridership far above any other city—New York's average ridership is more than 9 times the next highest city, Washington, D.C. As a result, initial data analysis also explored the log of average ridership. While not normally distributed, this scale is much more reasonable. However, visualization of trends was still very difficult, and typically had to be done by city for meaningful insights. Interestingly, none of the continuous predictors appeared to have a linear relationship with the response (or log response). Instead, all of the trends appeared to be quadratic. One initial hypothesis for this occurrence was the confluence of overlapping trends, such as seasonality, gas prices, population, and other trends unaccounted for by this set of predictors.

Despite challenges with the data, there were still useful insights gathered from initial data exploration. First of all, seasonal trends appeared to vary by city. For example, Washington, D.C. and Chicago visibly had lower ridership in the winter, while Philadelphia experienced the lowest ridership in the summer, and Los Angeles experienced very little differences by season. Labor force participation rate and gas prices had the most consistent strong positive relationship with ridership, across nearly every city. Finally, in examining the trend of ridership over time, the motivation for this project was confirmed. While the trends are different, every city in this analysis experienced a decrease in ridership before the end of 2019. Below are visualized the trends across time for different cities, displayed as LOESS lines. New York is displayed separately for easier visualiztion. This consistent trend is particularly striking, especially given the differences in population trends, expansion of transit services and climate across cities.

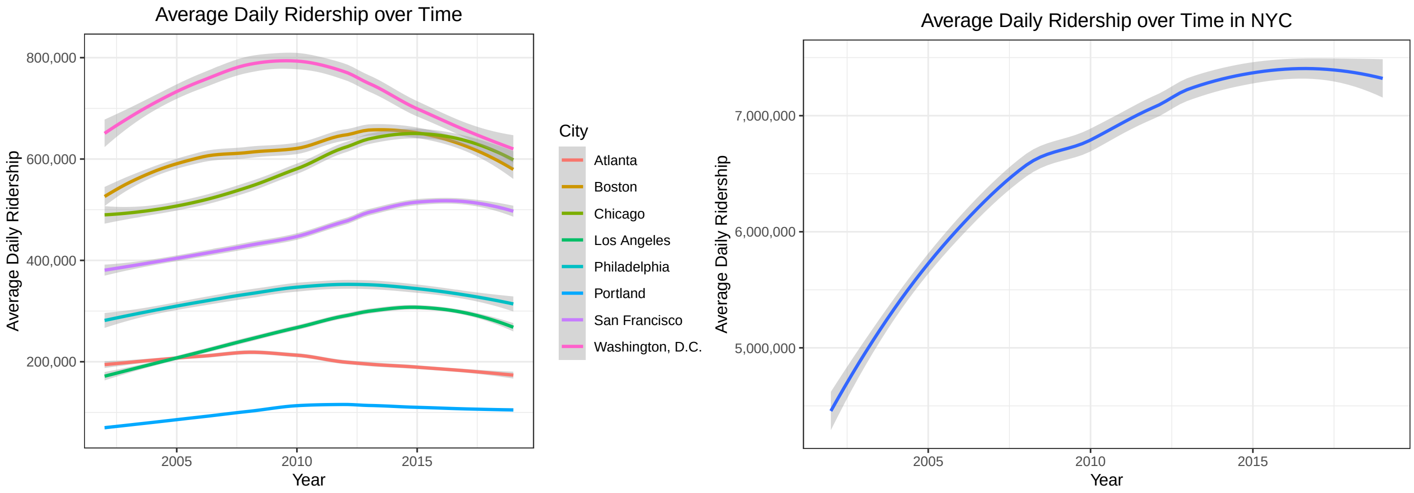


Figure 1: Average Daily Ridership Over Time by City

# Model

During the planning stage of the analysis, the goal was to fit a hierarchical model with cities as random intercepts, especially given the normal distribution of the response variable within cities. Unfortunately, during modeling, a hierarchical model failed to converge when including the interaction between year and the Uber indicator variable as a random slope. Because that interaction is one of the main outcomes of interest to this analysis, an alternative model was developed. The final model selected was a linear model, with the response as the logarithm of average daily ridership. The general formula for this model is included below. Here, each $x_{pi}$ represents either a continuous predictor, a dummy variable representing a categorical variable, or an interaction between two or more predictors.

- Predictors in the final model: city, year, population, vehicles, route miles, minimum temperature, precipitation, Uber, season, average fare, labor force participation
- Interactions in the final model: Uber:year:city, Uber:average fare, Uber:gas price, and city:season

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + ... + \beta_p x_{pi} + \epsilon_i, \ \epsilon_i \sim N(0, \sigma^2), \ i = 1, ..., n$$

## Model Selection

Before fitting any models, variables were centered, and rescaled given their massively different scales. Population was transformed into 100,000s of people, the number of vehicles was converted to 10s of vehicles, and the same transformation was done to miles. Initially, the full model selected included all predictors: city, year (as a continuous predictor), population, population density, labor force participation, unemployment rate, CPI adjusted gas price, max number of vehicles in service, number of directional route miles, CPI adjusted average fare, average minimum temperature, average precipitation, average snowfall, average snow depth, Uber indicator variable, and season. Also included were several interactions based on visualizations from preliminary data exploration. These included a three-way interaction between city, the Uber indicator variable, and year; an interaction between season and city; an interaction between Uber and average fare, and interaction between Uber and gas price. Also based on the findings from EDA, the response variable in this first model was the logarithm of average daily ridership per month.

After fitting this initial model, model selection was performed using step-wise selection with AIC as the selection criteria. This removed several predictors, including population density, unemployment rate, precipitation, snowfall, and snow depth. All of the chosen interactions were included in the model.

## Model Assessment

Subsequently, the model chosen with stepwise selection was assessed to verify that it meets the four assumptions required for linear regression: normality, linearity, independence, and constant variance. As discussed above, the logarithm of average daily ridership was selected as the response for the model, but this assumption was double checked here. Below is included the QQ plots for the same model, both without and with the log transformation of the response variable, on the left and right, respectively. Visibly, the log transformation is needed to met the assumption of normality.

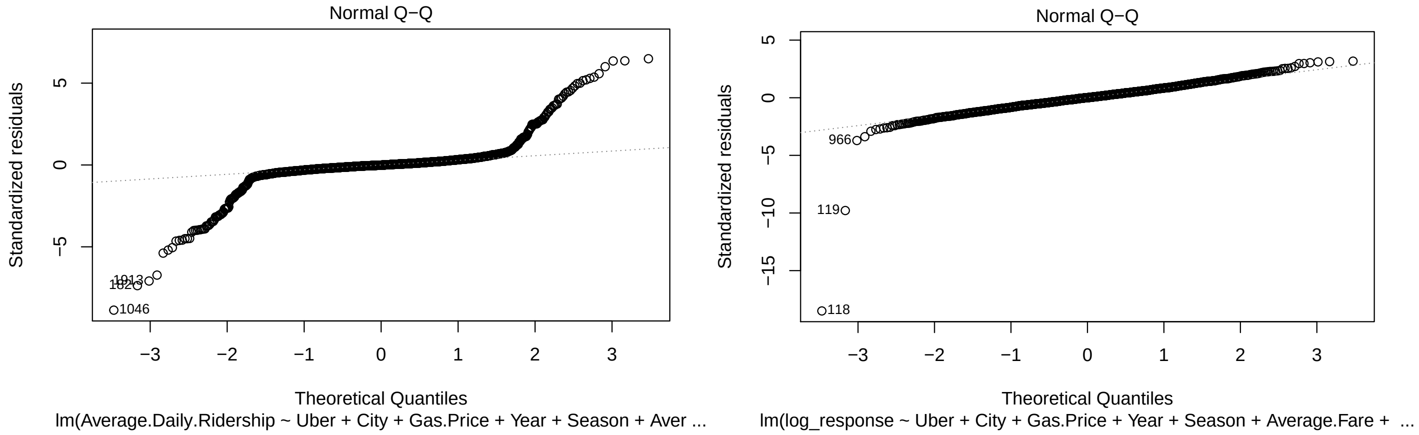


Figure 2: Residual QQ-Plot Without (Left) and With (Right) Log Transformation

Also visible in the above QQ-plot after log transformation are two large outliers. These outliers are also visible in the plot of residuals vs. fitted values and the plot of residuals vs. leverage, both displayed below. While the model clearly satisfies the independence and equal variance assumption, the magnitude of these outliers are concerning. Investigation of these outliers reveals that they are consecutive months (October and Novrmber) of data from Los Angeles in 2003. News articles from that period, including one from the LA Times reveal that there was a transit worker strike during this period, explaining the low ridership. Due to this, these observations were removed from the data and the model was refit. Stepwise selection was conducted again, and this time precipitation was kept as a predictor.

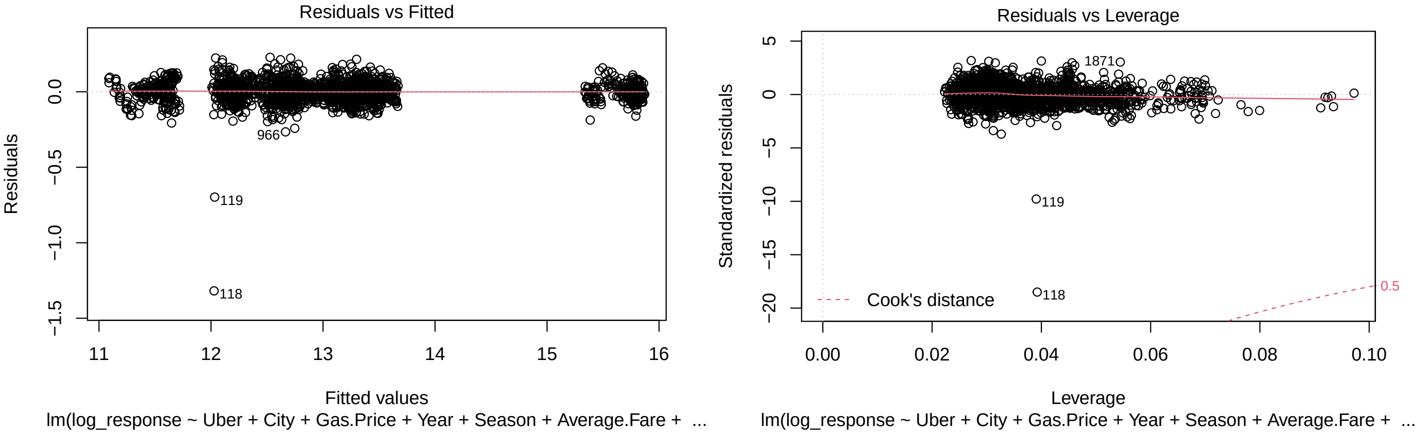


Figure 3: Residuals vs. Fitted Values

After removing these points, there were some other high leverage points, but none with high Cook's distance. Next, the assumptions for linearity were checked by plotting the residuals vs. each individual continuous predictor. There was no visible trend in these diagnostic plots, so this model was selected as the final model. All diagnostic plots for this model are included in Appendix 1. The last model assessment undertaken was checking for multicollinearity. There were several terms with very high VIF and GVIF values, which is expected given the number of interaction terms. Despite this, the model was not changed, because several of the terms with high VIF were critical to the inferential questions of this analysis.

## Model Interpretation

Due to space constraints, the output of the final model is included in Appendix 2. To interpret the effect of certain variables, namely City, Uber, Year, and Season, visualizations are required, due to the intricate nature of the interactions included in the final model. Figure 4 demonstrates the estimated percent change in ridership year over year by city, before and after the introduction of ridesharing service—under the assumption that all other variables in the model are held constant (i.e. population, weather, season, etc.). This figure shows that every city experienced a decrease in the year over year trend in ridership after Uber was introduced, and all cities except New York and San Francisco experienced a reversal of trends. Washington D.C. and Los Angeles are the most affected, while Chicago and San Francisco are the least affected.

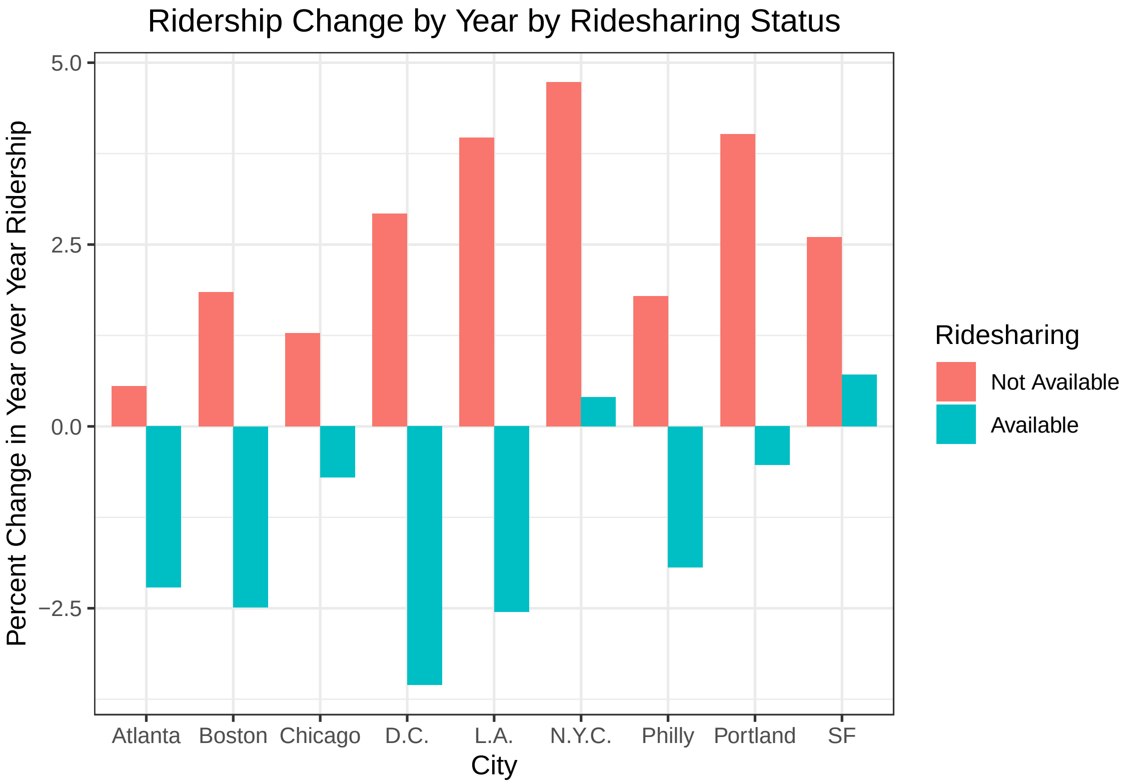


Figure 4: Projected Year over Year Change in Ridership Before and After Introduction of Ridersharing Services

Figure 5 displays how ridership changes by season and by city, again controlling for all other factors. This graph contains some very interesting trends. Overall, spring and fall appear to be high ridership months, while winter is generally a lower ridership month, and summer is closer to average. Notably, there are some deviations from this trend: Philadelphia and New York have decreased ridership in the summer. In fact, Philadelphia's summers show the greatest deviation from average than any other city's seasons. Los Angeles shows very little differences between seasons, which makes sense given the consistent climate.

Continually, we can interpret the remaining coefficients that were significant at the 95% confidence level. The baseline of this model is the city of Atlanta in the year 2010, in the fall, before Uber was introduced, with all continuous predictors fixed at their mean. Thus the intercept of the model, 11.7, can be interpreted as the expected log of average ridership given these statistics, or 120,571, with a 95% confidence interval of (81,010, 179502). In the context of this interpretation, it is important to note that the continuous predictors are fixed at their overall mean, not the city-specific mean. There were several continuous predictors that were significant. For each of the following interpretations, the coefficients are exponentiated, and all other variables are assumed to be held constant. With these assumptions, the model indicates that a $1 increase in the inflation adjusted gas price increases ridership by approximately 3.1%. However, after the introduction of Uber, the effect of gas prices flips, and a $1 increase in the gas price decreases ridership by 1.6%. Similarly, an increase of 1% in labor force participation increases average daily ridership by 0.8%, while an increase of 10 degrees in the minimum

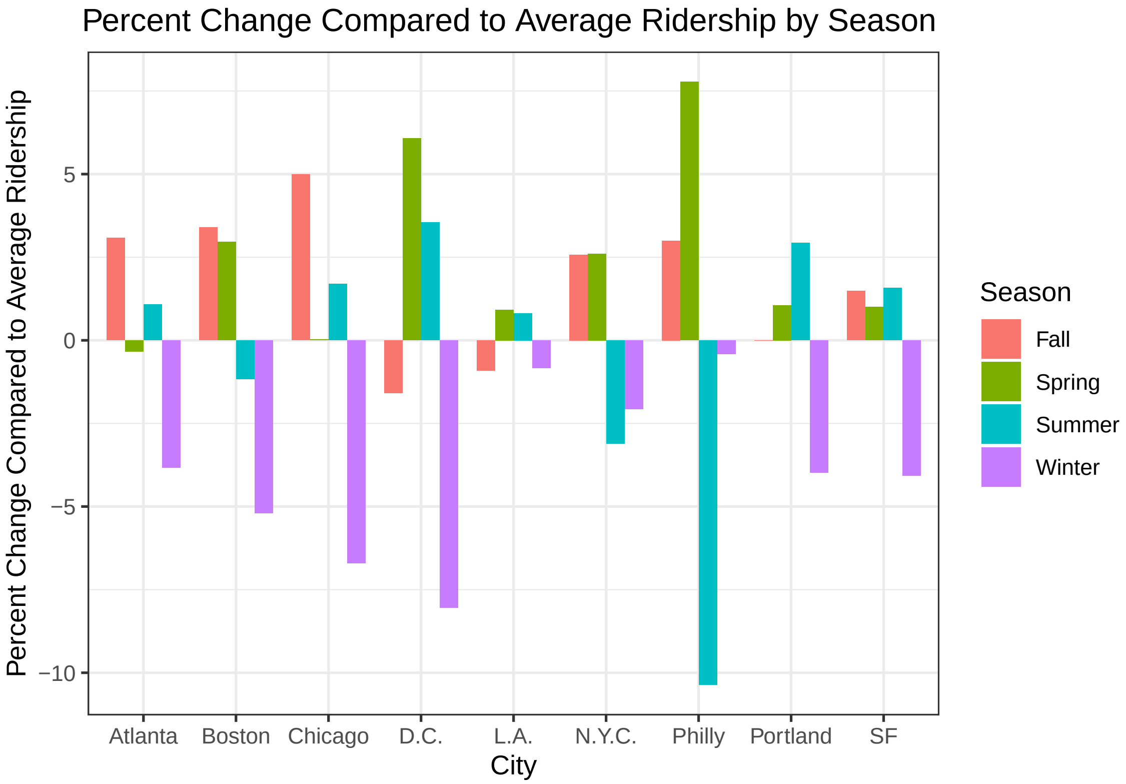


Figure 5: Percent Above Average Ridership by City by Season

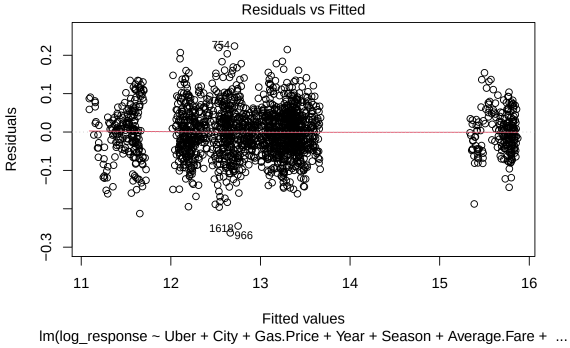
average temperature increases average daily ridership by 0.7% (accounting for season). Considering city-specific factors, an increase in a transit system of 10 miles increases ridership by 1.3%, and an increase in the population of 100,000 actually decreases ridership by 2.6%. Average fair is another variable that has a separate value before and after the introduction of Uber in a given city. Before Uber is available, an increase of 10 cents in the average fare decreases ridership by 2.6%, while after Uber is available the same increase increases ridership by 1.3%.

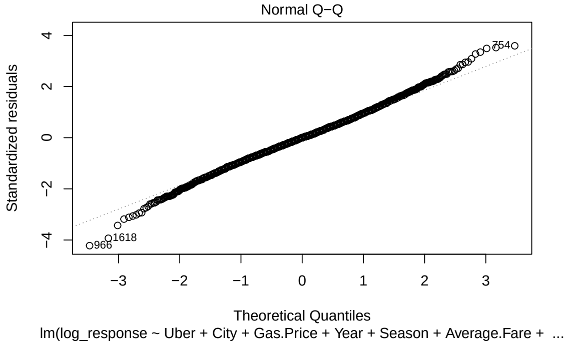
# Conclusion

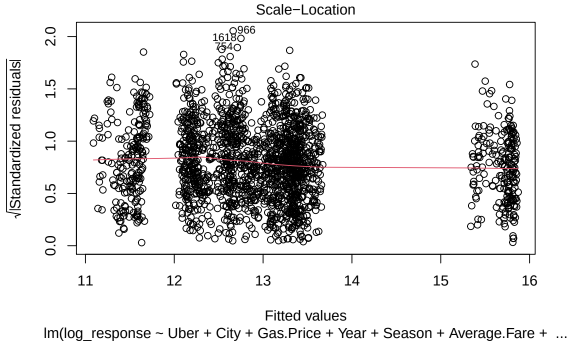
The final model utilized in this analysis answers the original questions of interest well. This model suggests that there is a significant difference in public transportation utilization before and after ridesharing services are introduced, after accounting for potential other explanatory factors. Every city in this analysis experienced a decrease in trend, with all but 2 cities seeing a significant reversal of trend. In the same vein, the introduction of ridesharing services also appears to have affected how gas prices and average fare are related to transit ridership. In addition to Uber-specific findings, this analysis also shed light on how other factors influence public transportation, ranging from controllable factors such as the length of the public transportation system to latent factors (i.e. weather).

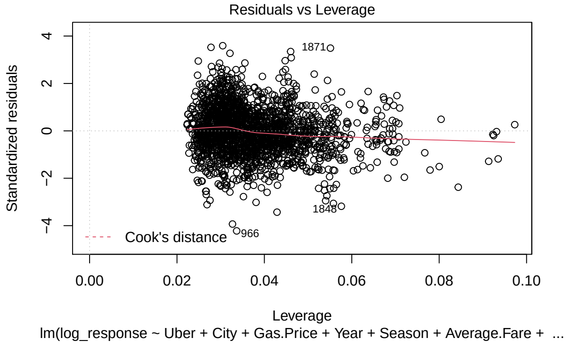
There are important limitations to consider in the context of interpretation, however. This analysis focuses only on rail ridership, and does not factor in bus ridership in these cities, which is often equal to or exceeding the number of unlinked passenger trips taken by rail. This is an important factor, because it's possible that rail losses could be accounted for by bus gains, especially with a shift towards bus-rapid transit (BRT) in the United States. Furthermore, another weakness of this analysis is the method by which ridesharing services are represented. Simply measuring whether Uber is available in a city is somewhat naive, given that many potential other factors could impact public transportation ridership: the number of Uber drives, the number of Uber trips, the cost of an Uber, and the availability of other ride-sharing services (including Lyft). However, most of this data is not available to the public. Finally, it is also important to recognize that there are definitely other factors that this analysis does not account for, including crime/safety, change in telecommuting prevelance, and more complicated population dynamics (such as population density near train stations). There is room for improvement in future work, including investigating bus ridership, including better population measures, and accounting for competition in the ridesharing market.

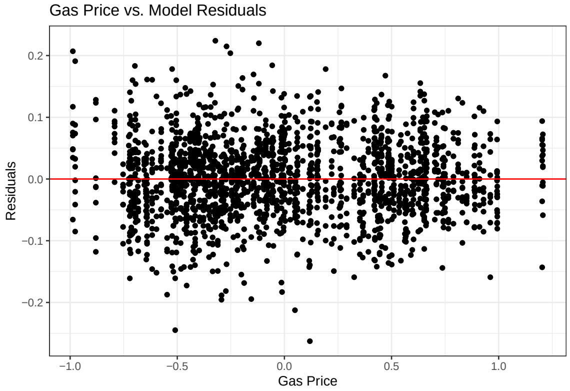
# Appendix 1: Final Model Diagnostics

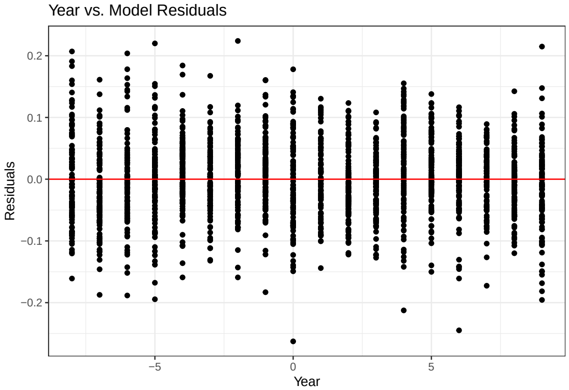


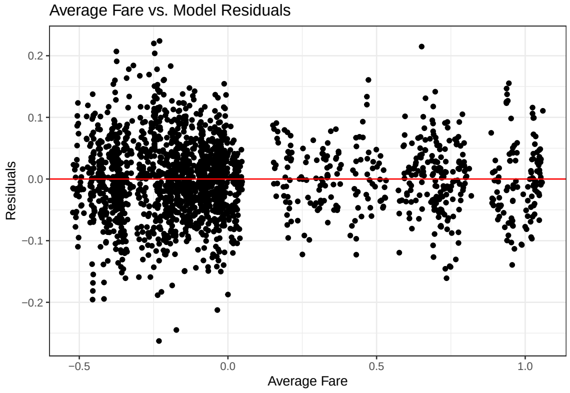


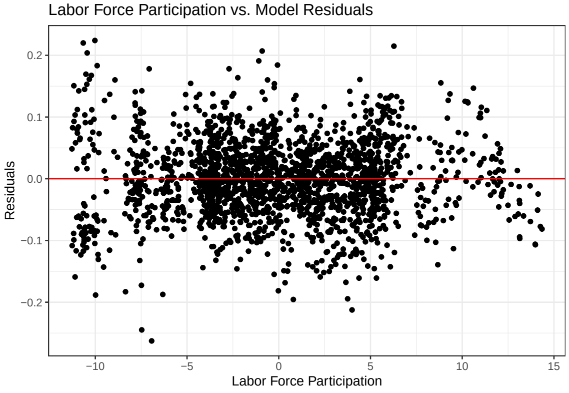


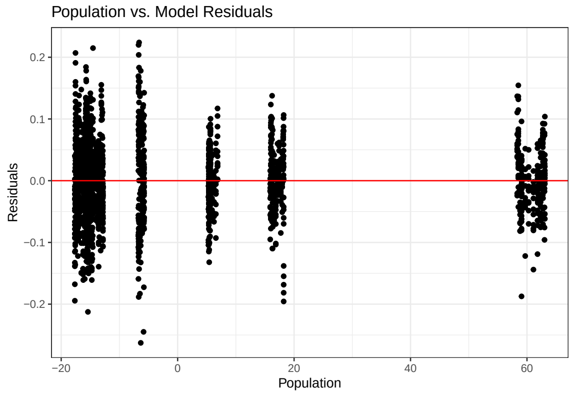


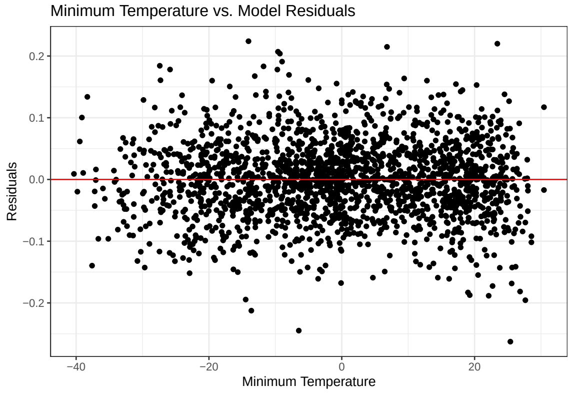


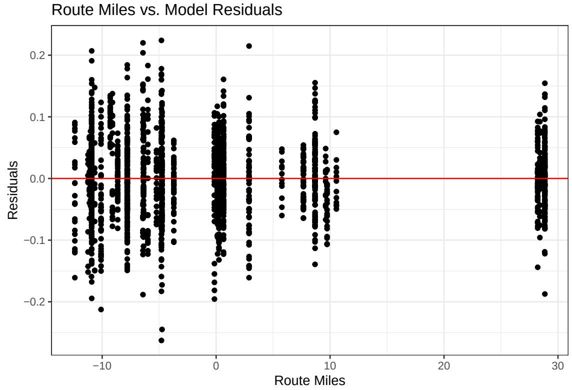


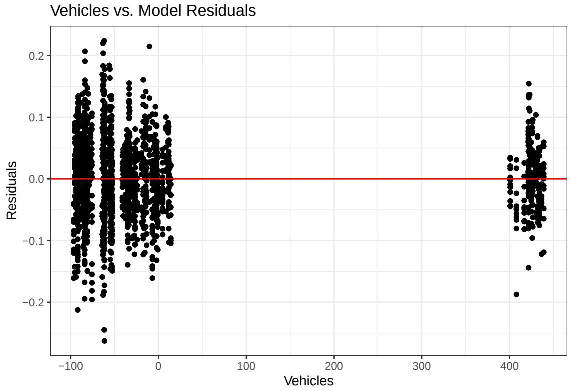


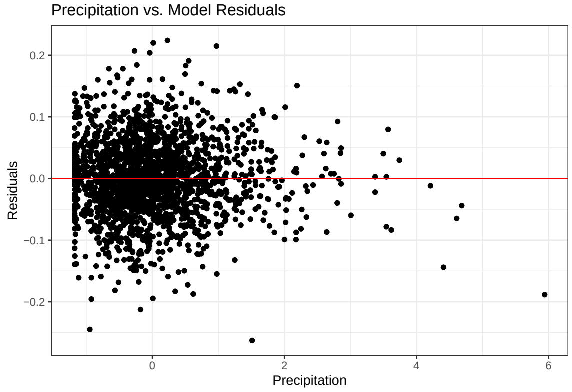



Population vs. Model Residuals


Minimum Temperature vs. Model Residuals


Route Miles vs. Model Residuals


Vehicles vs. Model Residuals


Precipitation vs. Model Residuals

# Appendix 2: Final Model Output

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---:|---:|---:|---:|---:|
| (Intercept) | 11.7001 | 0.2028 | 57.68 | 0.0000 |
| Uber1 | 0.1755 | 0.0274 | 6.40 | 0.0000 |
| CityBoston | 1.2305 | 0.0441 | 27.90 | 0.0000 |
| CityChicago | 1.7390 | 0.2436 | 7.14 | 0.0000 |
| CityLos Angeles | 1.0224 | 0.2886 | 3.54 | 0.0004 |
| CityNew York City | 5.9843 | 1.0211 | 5.86 | 0.0000 |
| CityPhiladelphia | 0.8723 | 0.1093 | 7.98 | 0.0000 |
| CityPortland | -0.6674 | 0.0199 | -33.57 | 0.0000 |
| CitySan Francisco | 0.9677 | 0.1036 | 9.34 | 0.0000 |
| CityWashington, D.C. | 1.4939 | 0.0839 | 17.80 | 0.0000 |
| Gas.Price | 0.0305 | 0.0059 | 5.17 | 0.0000 |
| Year | 0.0055 | 0.0021 | 2.68 | 0.0075 |
| SeasonSpring | -0.0337 | 0.0123 | -2.75 | 0.0061 |
| SeasonSummer | -0.0196 | 0.0126 | -1.55 | 0.1218 |
| SeasonWinter | -0.0693 | 0.0129 | -5.40 | 0.0000 |
| Average.Fare | -0.2684 | 0.0287 | -9.36 | 0.0000 |
| Labor__perc | 0.0082 | 0.0014 | 5.98 | 0.0000 |
| Minimum.Temperature | 0.0007 | 0.0002 | 3.34 | 0.0008 |
| Route.Miles | 0.0130 | 0.0039 | 3.35 | 0.0008 |
| Population | -0.0268 | 0.0086 | -3.11 | 0.0019 |
| Vehicles | -0.0016 | 0.0010 | -1.57 | 0.1155 |
| Precipitation | -0.0028 | 0.0020 | -1.40 | 0.1622 |
| Uber1:CityBoston | 0.0283 | 0.0296 | 0.96 | 0.3391 |
| Uber1:CityChicago | 0.0119 | 0.0319 | 0.37 | 0.7094 |
| Uber1:CityLos Angeles | 0.2684 | 0.0343 | 7.83 | 0.0000 |
| Uber1:CityNew York City | -0.0248 | 0.0345 | -0.72 | 0.4720 |
| Uber1:CityPhiladelphia | 0.0035 | 0.0312 | 0.11 | 0.9098 |
| Uber1:CityPortland | -0.0611 | 0.0527 | -1.16 | 0.2464 |
| Uber1:CitySan Francisco | -0.4842 | 0.0927 | -5.22 | 0.0000 |
| Uber1:CityWashington, D.C. | -0.3311 | 0.0746 | -4.44 | 0.0000 |
| CityBoston:Year | 0.0128 | 0.0030 | 4.27 | 0.0000 |
| CityChicago:Year | 0.0072 | 0.0035 | 2.10 | 0.0361 |
| CityLos Angeles:Year | 0.0334 | 0.0031 | 10.85 | 0.0000 |
| CityNew York City:Year | 0.0407 | 0.0051 | 8.01 | 0.0000 |
| CityPhiladelphia:Year | 0.0122 | 0.0027 | 4.52 | 0.0000 |
| CityPortland:Year | 0.0339 | 0.0024 | 13.87 | 0.0000 |
| CitySan Francisco:Year | 0.0202 | 0.0036 | 5.55 | 0.0000 |
| CityWashington, D.C.:Year | 0.0233 | 0.0036 | 6.52 | 0.0000 |
| CityBoston:SeasonSpring | 0.0295 | 0.0173 | 1.70 | 0.0885 |
| CityChicago:SeasonSpring | -0.0147 | 0.0173 | -0.85 | 0.3945 |
| CityLos Angeles:SeasonSpring | 0.0520 | 0.0174 | 2.99 | 0.0028 |
| CityNew York City:SeasonSpring | 0.0340 | 0.0174 | 1.96 | 0.0502 |
| CityPhiladelphia:SeasonSpring | 0.0791 | 0.0174 | 4.56 | 0.0000 |
| CityPortland:SeasonSpring | 0.0443 | 0.0173 | 2.56 | 0.0105 |
| CitySan Francisco:SeasonSpring | 0.0290 | 0.0173 | 1.67 | 0.0943 |
| CityWashington, D.C.:SeasonSpring | 0.1086 | 0.0173 | 6.29 | 0.0000 |
| CityBoston:SeasonSummer | -0.0256 | 0.0173 | -1.48 | 0.1403 |
| CityChicago:SeasonSummer | -0.0123 | 0.0174 | -0.71 | 0.4780 |
| CityLos Angeles:SeasonSummer | 0.0368 | 0.0174 | 2.12 | 0.0342 |
| CityNew York City:SeasonSummer | -0.0375 | 0.0173 | -2.17 | 0.0303 |
| CityPhiladelphia:SeasonSummer | -0.1194 | 0.0174 | -6.88 | 0.0000 |
| CityPortland:SeasonSummer | 0.0485 | 0.0174 | 2.78 | 0.0055 |
| CitySan Francisco:SeasonSummer | 0.0205 | 0.0177 | 1.16 | 0.2456 |
| CityWashington, D.C.:SeasonSummer | 0.0704 | 0.0173 | 4.07 | 0.0000 |
| CityBoston:SeasonWinter | -0.0175 | 0.0173 | -1.01 | 0.3129 |
| CityChicago:SeasonWinter | -0.0488 | 0.0175 | -2.79 | 0.0053 |
| CityLos Angeles:SeasonWinter | 0.0701 | 0.0173 | 4.04 | 0.0001 |
| CityNew York City:SeasonWinter | 0.0230 | 0.0173 | 1.32 | 0.1859 |
| CityPhiladelphia:SeasonWinter | 0.0357 | 0.0174 | 2.05 | 0.0402 |
| CityPortland:SeasonWinter | 0.0287 | 0.0173 | 1.66 | 0.0968 |
| CitySan Francisco:SeasonWinter | 0.0131 | 0.0174 | 0.75 | 0.4527 |
| CityWashington, D.C.:SeasonWinter | 0.0015 | 0.0173 | 0.09 | 0.9321 |
| Uber1:Year | -0.0279 | 0.0039 | -7.07 | 0.0000 |
| Uber1:Average.Fare | 0.4004 | 0.0799 | 5.01 | 0.0000 |
| Uber1:Gas.Price | -0.0465 | 0.0096 | -4.84 | 0.0000 |
| Uber1:CityBoston:Year | -0.0156 | 0.0054 | -2.87 | 0.0041 |
| Uber1:CityChicago:Year | 0.0081 | 0.0055 | 1.47 | 0.1420 |
| Uber1:CityLos Angeles:Year | -0.0369 | 0.0054 | -6.82 | 0.0000 |
| Uber1:CityNew York City:Year | -0.0143 | 0.0058 | -2.46 | 0.0141 |
| Uber1:CityPhiladelphia:Year | -0.0094 | 0.0054 | -1.76 | 0.0790 |
| Uber1:CityPortland:Year | -0.0168 | 0.0075 | -2.26 | 0.0242 |
| Uber1:CitySan Francisco:Year | 0.0093 | 0.0053 | 1.76 | 0.0783 |
| Uber1:CityWashington, D.C.:Year | -0.0371 | 0.0057 | -6.56 | 0.0000 |

Table 1: Final Regression Model