The Effect of US Gas Prices on Alternative Methods of Travel

MGT6203: Data Analytics in Business
Summer 2023

Team 104
Caroline Schmitt
Emy Ng
Matthew Kim
Mike Genovese
Osman Yardimci

Abstract

As gas prices fluctuate in the US, people tend to shift to alternative methods of travel such as buses, light rail, heavy rail, commuter rail, carpooling, and bicycling. This brings attention to the travel behavior of Americans, historically known as a gasoline-powered car dependent society. These methods of transportation are substitutable from each other. An individual can choose to travel to a destination by car, by rail, or by bike, but may only select one mode at a time. Therefore, it may be useful to examine how the economics of refueling automobiles can affect consumer behavior, and ultimately what alternates they may (or may not) choose.

This research examines the relationship between volatility in gasoline prices and transit ridership on public transportation in the past years, which is a relatively unexamined topic. The data will be analyzed using correlation, causality test, and time series regression.

The results indicate a moderate positive correlation between gasoline prices and transit ridership that could suggest transit agencies to prepare for higher transit ridership in their planning strategies.

Objective/Problem Statement

Alternative methods of transportation like public transit and bike share programs are affordable options for people to avoid spending on gasoline for personal automobiles, but ridership trends for these alternative methods aren't publicly modeled. The purpose of our analysis will be to dig into the relationship between alternative transport ridership numbers and how the price of gasoline in the US affects these numbers.

Business Justification/Impact

There are multiple resources readily available that try to simulate and forecast future gasoline prices in the United States. Our findings will confirm if alternative transportation organizations and companies can utilize these forecasts when planning their business operations. If US gas prices do indeed affect alternative ridership numbers, organizations can plan ahead to have a closer estimation on the number of resources required to meet the demand of riders. This will profit maximum profits by not having a surplus of resources while demand is lower than expected or alternatively not having not enough resources available to meet demand.

Ridership forecasting is useful for transit agencies as well as companies like Lyft or Uber that compete with public transit. Understanding regional ridership elasticity may also help regional and national policymakers increase public transit use.

Research Questions

How does the average US gas price over time affect ridership numbers of alternative transportation methods?

1. Which areas show the closest increase/decrease in transit ridership in response to gas price changes?
2. Do all grades of gasoline fluctuate in price at roughly the same rate? And if not, does one affect alternative transportation ridership more than others?
3. Are there any particular significant events that explain any sudden spikes in alternative transportation ridership numbers?

Hypothesis

We expect to see a positive relationship between gas prices and transit ridership. However, we would assume this relationship is only one of many factors that drive transit ridership. We expect that the most significant independent variable will be regular-grade gas price, but that we will see different patterns in different cities due to different regional availability and the existing use of public transit.

We're expecting some relationship between gas prices and transit ridership. In particular, we expect rises in transit ridership when gas prices are especially high. But gas prices and transit ridership are all complicated and affected by many exogenous variables not explicitly included in our dataset. We don't expect to be able to explain any of our variables of interest completely.

Methodology/Approach

We will be working with different tools and calculations to analyze if gasoline prices and alternative transportation ridership numbers are correlated. We will be utilizing Spearman Rank Correlation tests to help solve our research questions related to correlation and the effects of gas prices in the United States. We will also be working with Granger Causality Tests. This test will add another layer to our correlation tests if they show there is a positive correlation between the two variables. The Granger Test will show if alternative ridership numbers are caused directly by gas prices at a significant level.

Data Overview, Cleaning and Preparation

Data Overview

Complete monthly ridership dataset: Of the three primary sources our group worked with, the National Transit Database (NTD) monthly ridership dataset is the most complicated and messy. The four forms of counting in the ridership dataset are unlinked passenger trips (UPT), vehicle revenue miles (VRM), vehicle revenue hours (VRH), and vehicles operated in maximum service (VOMS). The NTD glossary provides full explanations of these measurements. For our purposes, unlinked passenger trips are most relevant. The NTD defines unlinked passenger trips as: "The number of passengers who board public transportation vehicles. Passengers are counted each time they board vehicles, no matter how many vehicles they use to travel from their origin to their destination." We cannot interpret UPT counts as passenger counts, and the UPT count will necessarily be larger than the number of actual passenger trips.

The UPT count data presented a few challenges. There are multiple transit authorities *and* modalities per what NTD defines as an "urbanized area" (UZA), which is "an incorporated area with a population of 50,000 or more that is designated as such by the U.S. Department of Commerce, Bureau of the Census." For instance, there are three public ferries associated with the UZA name "New York--Jersey City--Newark, NY--NJ." They are operated by "Metro-North Commuter Railroad Company, dba: MTA Metro-North Railroad," the NYC Department of transportation, and the Port Authority Trans-Hudson Corporation. Consequently, we aggregated data across UZAs, thus working with summed unlinked passenger trips across all owners and modalities.

U.S. Gasoline and Diesel Retail Prices:
We are utilizing a dataset from the U.S. Energy Information Administration and additionally the EIA's real prices of gasoline and diesel fuel. Because gasoline and diesel price datasets stretch back decades, prices cannot be directly compared without adjusting for inflation. (The EIA performs this adjustment by dividing the monthly price by the consumer price index, CPI.)

Though formatted for Excel, the EIA datasets are clean once multi-row row and column labels are accounted for. There are some quirks and missing observations to be aware of. First, the EIA notes that twice they have changed their surveying methodologies: in 2018 to more accurately track gasoline prices and in 2022 to more accurately track on-highway diesel fuel prices.

Data Cleaning

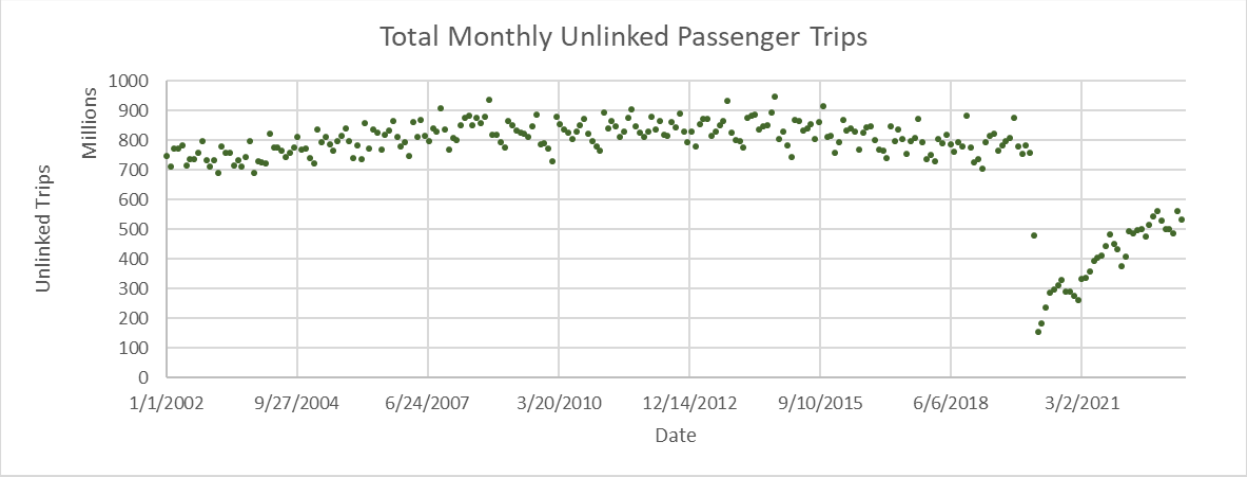
In order to perform our correlation and causality tests, the unlinked passenger transit data set and the US gasoline retail price data set needed to be cleaned and combined. The transit data was grouped by month and ridership values were summed across the months. This provided sufficient data to start our analysis. The data was further dissected to test different scenarios which will be walked through in the modeling section.

We have also successfully aggregated UPTs (unlinked passenger) per UZA (city area). While it may be of interest to study which modalities have the most elasticity, for now we're most interested in total UPTs across a subset of UZAs. There are 398 UZAs in the dataset; we will not model all 398 individually. This data was also combined with the US gasoline retail price data set to calculate correlations in specific cities to help answer our second research question. Some cities unfortunately had months where ridership data was not recorded. These cities had to be removed from the data set as there would be an impact on correlation scores.

Exploratory Analysis Conclusions

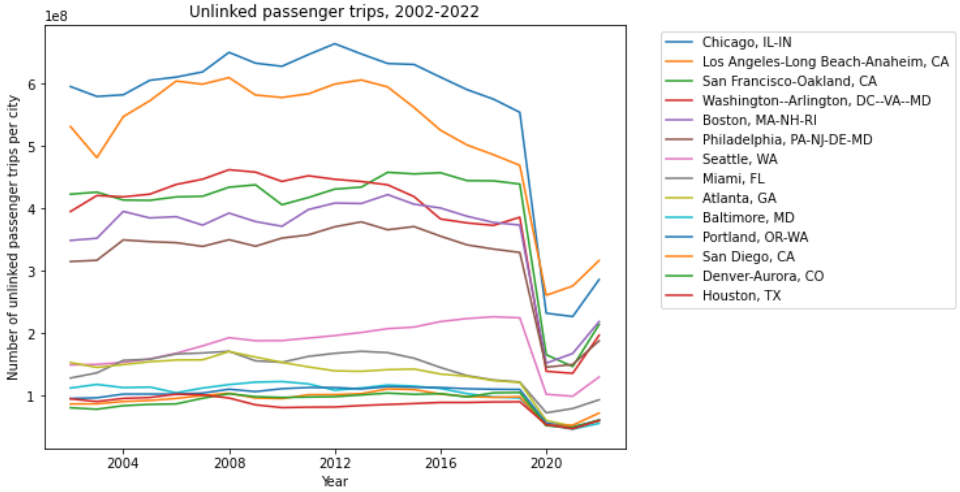
The exploratory analysis shows that Covid-19 had a heavy impact on the number of travelers using public transportation. This could be due to factors like transit services shutting down or decreasing operations, the increase in work-from-home jobs, and decrease in workforce participation (Appx. Fig. 16) mixed with an increase in unemployment (Appx. Fig. 17). This could affect the data modeling stage so we need to be aware and alter our models if needed.

Figure 1. Total Transit Ridership from Jan 2002 to Apr 2023



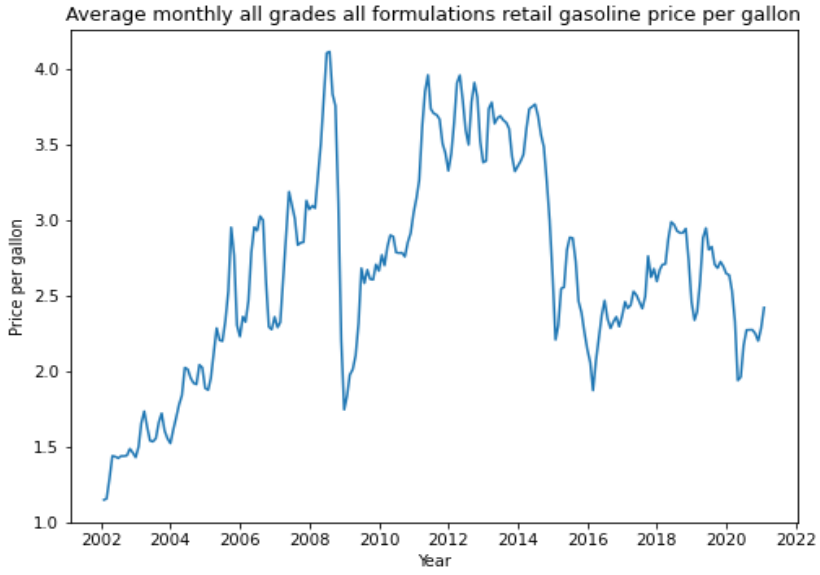
Below is an exploratory line plot of annual unlinked passenger trips counts from 15 of the highest-trip cities excluding New York, which is an extreme outlier.

Figure 2. Unlinked Passenger Trips - Top 15 Cities



When looking at inflation adjusted gasoline prices in the United States over time we see that there are a few major dips in prices. These can be attributed to the 2008 financial crisis, an oversupply of petroleum in the years 2014 to 2016, and Covid-19 in 2020.

Figure 3. US Average Monthly Price of Gasoline (All Grades)



Data Modeling

Model Assumptions

A major assumption that we are taking on is that our gas prices are a national average. We are using this as a proxy price of gas in city markets but the actual price of gas could be slightly different between cities. We hope that there aren't any cities that have significantly different fluctuations in gas prices and a monthly average won't impact data too greatly.

Another assumption that was briefly mentioned is the effect of weather on ridership numbers. We are assuming consistent weather for the most part but we understand that in reality cold months and months with heavy rainfall could affect the ridership values. Particularly hot

summers could also decrease ridership in the summer which typically has higher historical ridership numbers.

Lastly, our last assumption is that when testing correlation, larger cities will be assumed to not have a significant effect on ridership totals. For example, New York City has the largest transit ridership numbers as we saw from our exploratory analysis. Ideally all 398 cities are weighted by their population sizes but due to the complexity of the UZAs as mentioned before, it's unrealistic to factor these sizes into our calculations. Therefore we must assume that all cities have a near equal impact on our analysis.

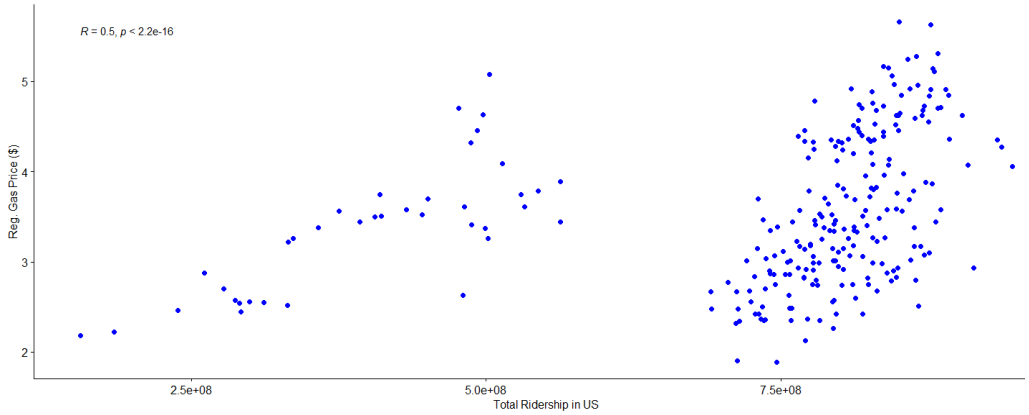
Correlation Between Regular Grade Gas Price and Transit Ridership

The first step to our data modeling is checking if there is a strong correlation present between gasoline prices and alternative transit ridership numbers. In order to test this, we utilized Spearman Rank Correlation calculations. The Spearman Rank Correlation is calculated in our instance by ranking all monthly average prices of gasoline from lowest to highest and ranking all monthly transit ridership numbers from lowest to highest. The delta of these ranks are taken and squared and plugged into the formula below where d is the delta and n is the number of rows of the dataset.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Multiple correlation tests were performed to account for different factors that affect the ridership data. First, was the correlation between regular grade gas and ridership without any adjustments. This yielded a value of 0.5 which is greater than the value of 0.123 which is the limit of a n=256 at a 95% significance interval. Therefore, we can conclude that there is a correlation between the two variables and there is a moderate relationship.

Figure 4. Reg. Gasoline vs Total Ridership in US Correlation (Jan. 2002 - Apr. 2023)



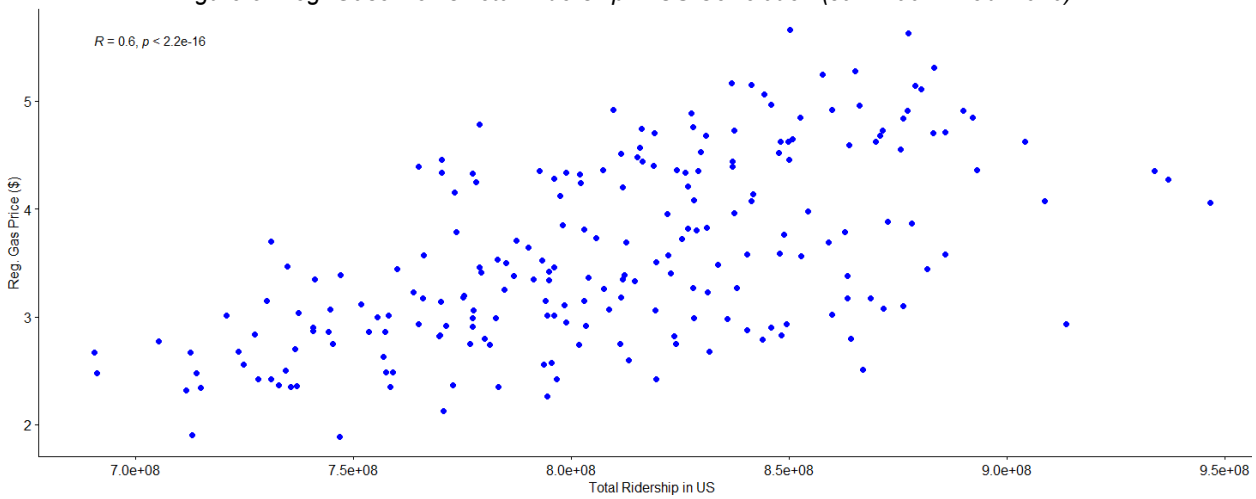
Based on our data exploration and the total correlation chart we see that there are 38 data points where ridership is exceptionally low. All of these data points are from March 2020 to the end of the dataset in April 2023. We will remove these points to see if there's a stronger correlation before Covid-19 changed the landscape of work and travel.

Date	Reg.Gas Price	Total Ridership
4/1/2020	2.18	155,607,108

5/1/2020	2.22	184,199,864
…	…	…
10/1/2022	3.89	563,264,404

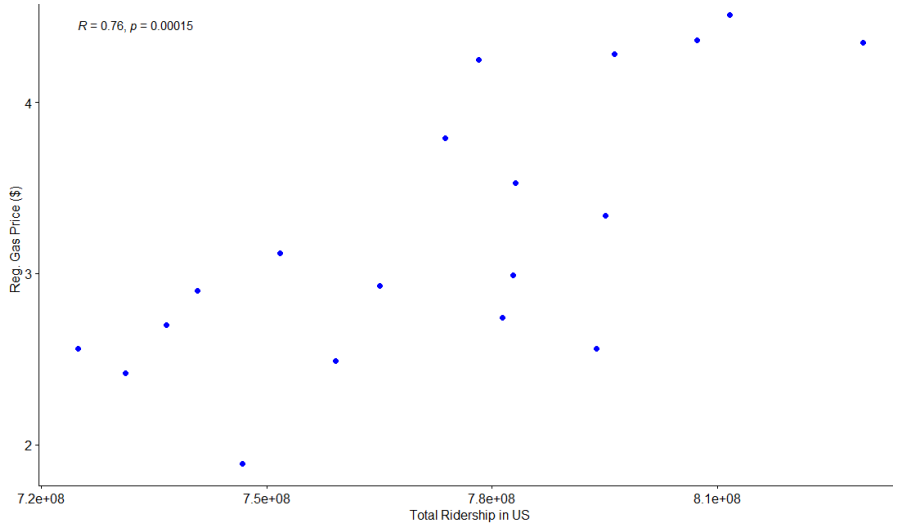
Our correlation after removing the post-covid months actually weakened the correlation now at a value of 0.6. Our Spearman rank correlation moved slightly closer to a perfect positive correlation.

Figure 5. Reg. Gasoline vs Total Ridership in US Correlation (Jan. 2002 - Feb. 2020)



Another factor that could be affecting our correlation is temperatures in the US. People may be less likely to ride public transit in the cold winter versus a warmer month. To adjust for this, the correlation between regular gas prices and ridership specifically just in January year-over-year was tested. The correlation was stronger than our past tests. With a calculated Spearman Rank Correlation value of 0.76, this is still greater than the 95% significance interval value of 0.46 for an n=19 dataset and the correlation value is even greater than our previous tests. A value greater than 0.7 is considered to be a strong relationship

Figure 6. Reg. Gasoline vs Total Ridership in US Correlation (Jan. YoY 2002 - 2020)



While other factors like weather may also play a role in ridership numbers, it appears that gas prices do prove they play a positive role in transit ridership numbers at a significant level.

Granger Causality Test

Going one step further we will test if there is causation between regular gas prices and alternative transit ridership using the Granger Causality Test. Using the same three sets of data used previously in the correlation tests, we calculated if there significant causality. From our results in the table below we found the pre-covid dataset showed significant signs of causality and could be valuable in predicting future values of transit ridership. The January year-over-year data also showed a level of significance but could have been affected by the small sample size.

Dataset	F Test Statistic	p-value
All Dates	1.8285	0.1775
Pre-Covid	42.753	4.487e-10 ***
January Year-Over-Year	5.1957	0.03769 *

Other Models

We did not implement multivariate ARCH/GARCH models because our exploratory analysis showed that the ridership data does not have the level of volatility which indicates use of ARCH/GARCH. All city data showed patterns in the ACF and PACF plots.

Seasonal decomposition models were useful for better understanding region-level data. With seasons accounted for, it's easier to identify regional trends -- for example, the Los Angeles-Long Beach-Anaheim area experienced an overall decrease in public transit use compared to other major cities. Decomposition analysis is discussed in more detail below. Due to the volatility of gas prices, however, decomposition analysis was of little use for forecasting with gas prices.

We implemented vector-autoregressive models for the five largest markets: NYC, Chicago, Los Angeles, San Francisco, and Washington, D.C. Data required scaling because the VAR model requires data to be of similar scales. Order was picked based on analysis of ACF and PACF plots as is indicated in the table below. The data used were weekly estimated unlinked passenger trips and adjusted gasoline prices.

Market	**Training MAE**	**Testing MAE**
NYC (order=2)	0.212	1.61
CHI (order=1)	0.180	1.49
LA (order=2)	0.180	0.585
SF (order=2)	0.183	1.54
DC (order=1)	0.229	1.25

The metric used for evaluation is the mean absolute error, abbreviated as MAE. This metric tells us about the distance between the actual data and the model's predictions. Thus, smaller

MAE scores are better. The data for each city were scaled, and thus the MAE scores are not directly comparable between city models. The MAE is consistently better for the training data than the testing data. This is expected; models perform better on data they have seen before. Additionally, autoregression models degrade as their forecast windows get longer. For our business case, this is acceptable; we're interested in analyzing the short-term spikes in transit demand that may be associated with spikes in gas prices. However, for evaluating model performance across a 70-30 train-test split, it is also unsurprising that the test MAE scores are consistently much worse.

Plots of the model predictions show that the VAR models generally capture the market level trends and some increases and decreases in demand, but they're not able to completely capture the trends and patterns present. Further modeling work is needed.

Discussion

RQ1: Which cities show the closest increase/decrease in transit ridership in response to gas price changes?

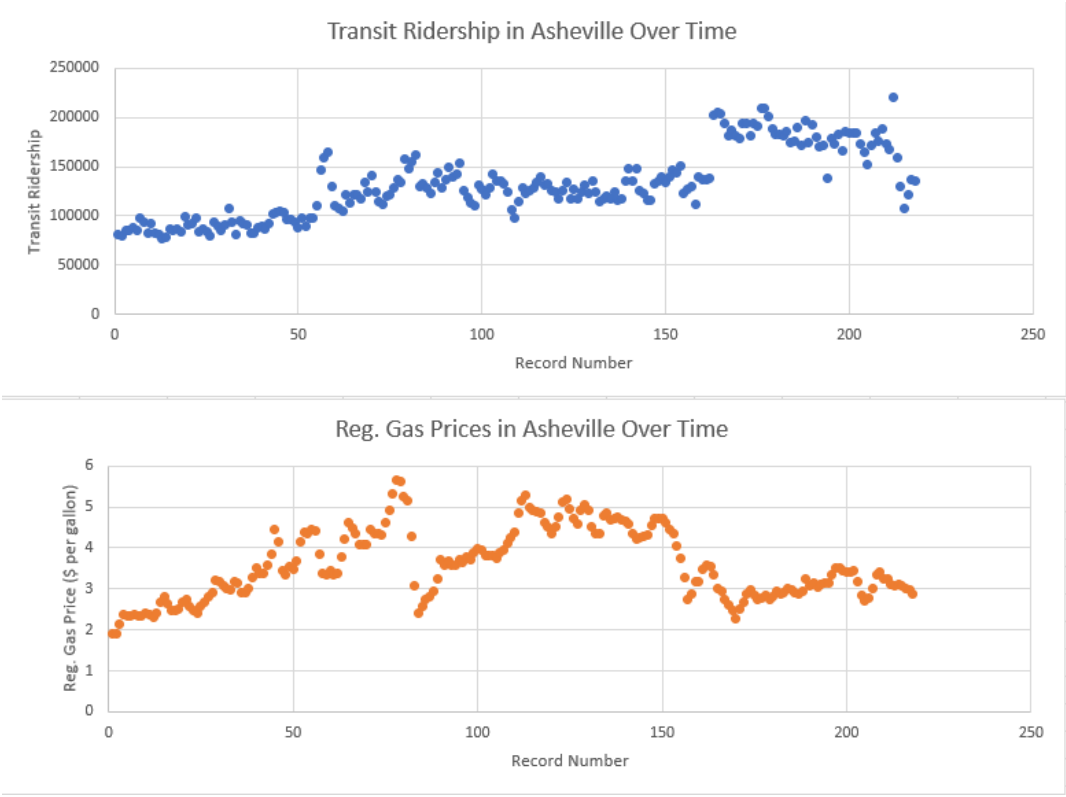
Using the same Spearman Rank Correlation values calculated to see if the gasoline prices were correlated to transit ridership numbers in general, we can also use it to find which city/area in the United States has the closest dependency on the price of gasoline when it comes to their public transit numbers.

After running spearman rank across all areas that didn't contain any zeros in their monthly ridership numbers we found that the most correlated area was Williamsburg, Virginia. Williamsburg scored a 0.3 in its correlation value which is lower than we expected but it's still a moderate positive correlation. There could have also been a more closely correlated city in our dataset however about half the cities contained zeros that would have altered their correlation values.

Figures 7 & 8. Williamsburg, VA Ridership and Reg. Gas Prices Over Time

The least correlated area was also calculated. This was Asheville, North Carolina that had a correlation value of just 0.03. This reflects in the graphs below when gas prices dip, there is either little to no movement or an increase in transit ridership.

Figures 9 & 10. Asheville, NC Ridership and Reg. Gas Prices Over Time

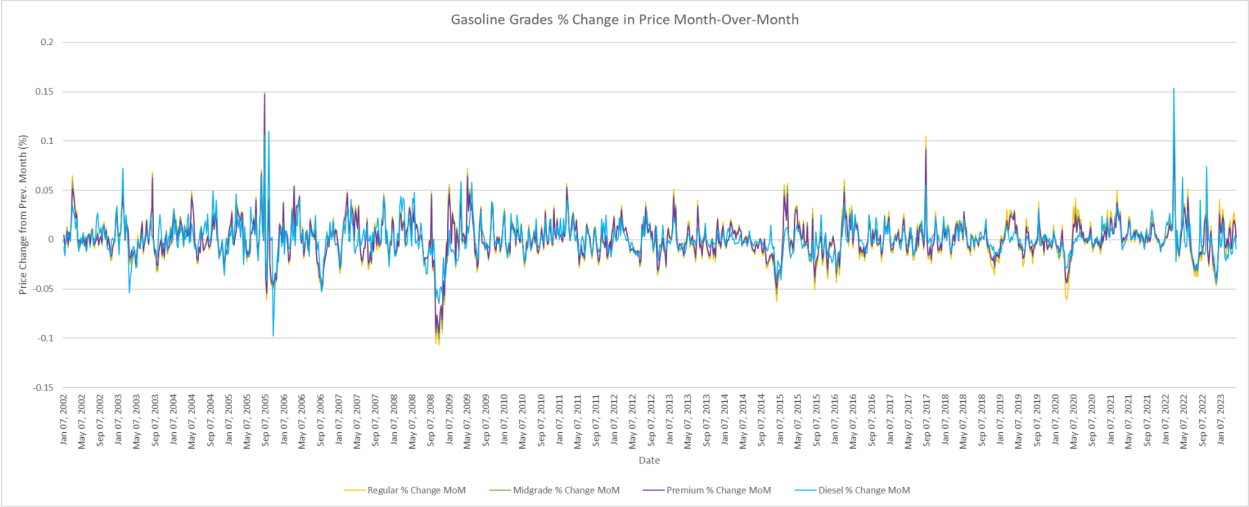


RQ2: Do all gasoline grades fluctuate at the same rate?
Yes, generally the gasoline grades do fluctuate at the same rates. The average price fluctuation for the different gas grades was within a 0.02% range of each other. After taking the highest and the lowest changes in price for each month, the average difference between the two was just 1.1%.

Grade Gasoline	Average Price Fluctuation Month-Over-Month
Regular (All formulations)	0.08%
Midgrade (All formulations)	0.09%
Premium (All formulations)	0.09%
Diesel	0.10%

In the graph below there are some noticeable points where one grade may have increased or decreased in price more than other grades but based on our data these points smooth out over the course of our roughly 11 year time period.

Figure 11. Percent Change in Gasoline Price from Previous Month



RQ3: Are there any particular events that explain sudden spikes in alternative transportation ridership numbers?

Yes, but to a much smaller amount than originally anticipated. To smooth the data, we decomposed both bus and commuter rail (the two most actively used methods of alternative transportation) and removed the seasonal component to clearly visualize the trend. As can be seen below, both the bus and commuter rail data both had a strong seasonal component.

Figure 12. Seasonal Component of Bus Commuting Data

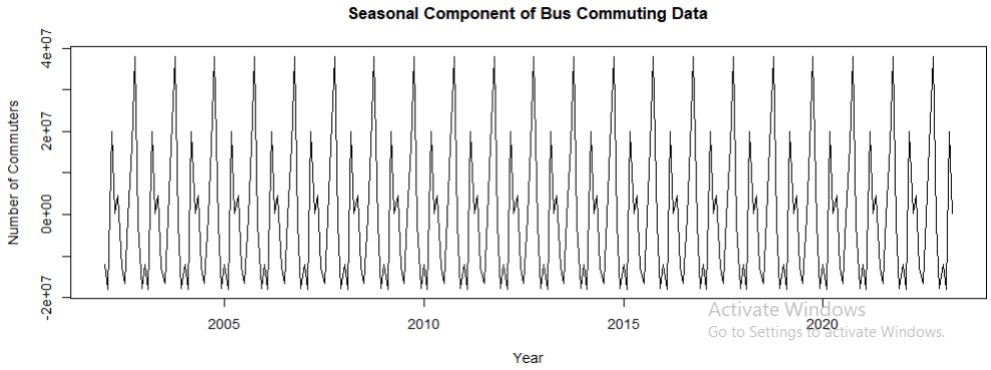


Figure 13. Seasonal Component of Commuting Rail Commuting Data

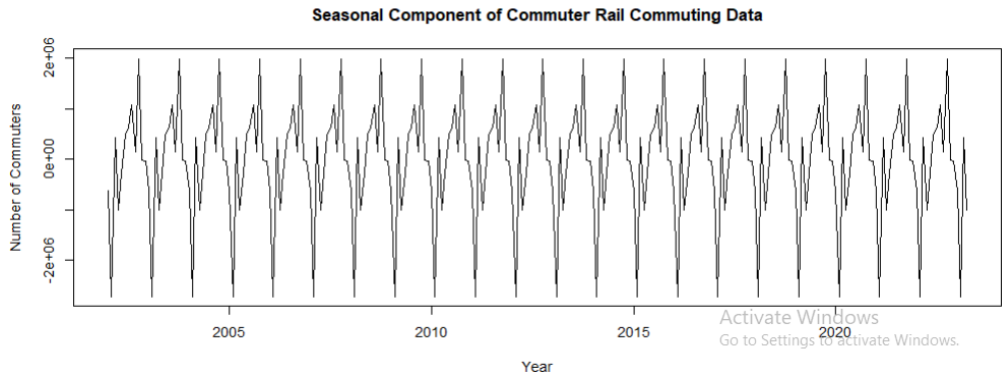


Figure 14. Bus Commuters by Year, Seasonally Adjusted

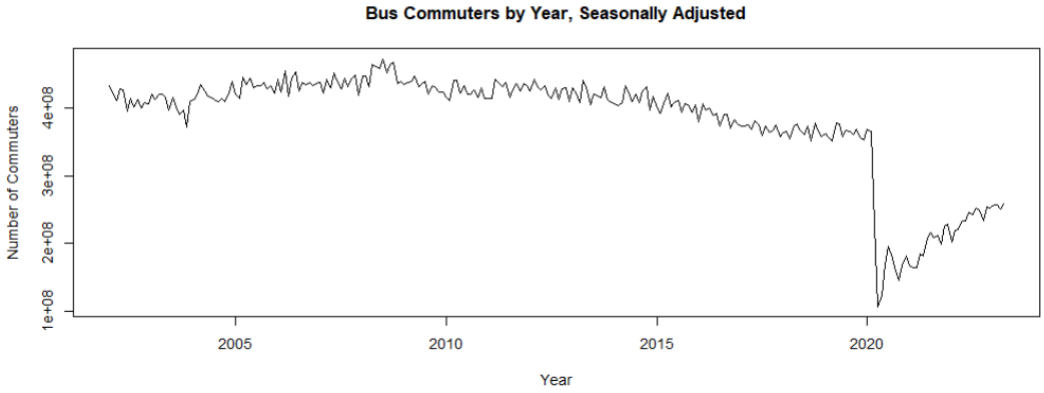
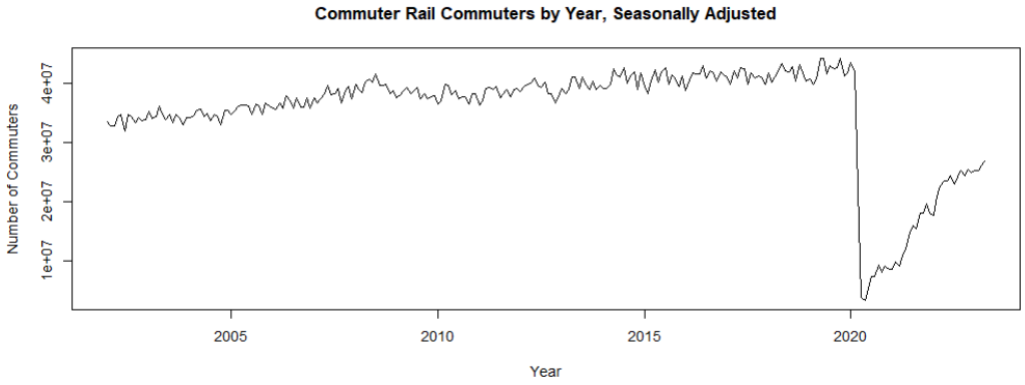
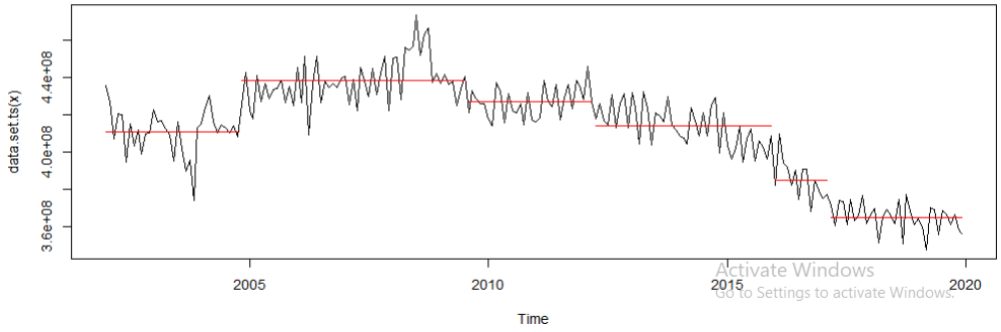


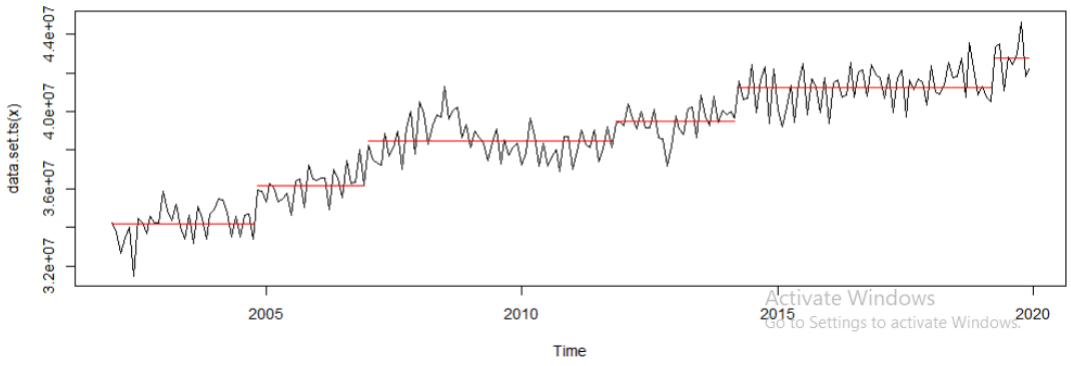
Figure 15. Commuter Rail Commuters by Year, Seasonally Adjusted



Most notably (and obviously), COVID-19 had such a large magnitude of change that any other trends are undetectable from a glance. To remedy this, we removed all data from 2020 onward to see the data without the outliers.

After cleaning the data, we used R's changepoint package and binary segmentation to identify five points where the residual sum of squares is highest (between each red line). Graphs for bus and commuter rail can be seen below:





Bus Data	Commuter Rail Data
October 2004	October 2004
July 2009	December 2006
March 2012	October 2011
December 2015	March 2014
February 2017	March 2019

Some commonalities were found in late 2004 (where regular-grade gas prices had increased by 23.73% YoY in October) and late 2011 (where regular-grade gas prices had increased 19.07% YoY in October as well). Although there are no notable discrete events that could intuitively affect these figures, historical context shows that 2004 was a particularly strong year for the US economy, while 2011 was more lackluster.

Curiously, the 2008 recession defied expectations. Although on both graphs a maximum exists followed by a steep decline that occurs post-financial crisis, accompanied by a substantial drop in gas prices, these points are not identified by changepoint without providing a higher Q value. These impacts, while significant, may be of smaller importance to alternative transportation than originally anticipated.

Conclusion

From the results of our analysis, it appears that our initial hypothesis was correct that there is a positive relationship between the price of gasoline and alternative transportation ridership numbers in the United States. After cleaning our data to attempt to decrease our limitations, we saw that there was a moderate to strong correlation between gas prices and transit ridership using Spearman Rank Correlation. We also concluded that gas prices had a significant level of causality by using a Granger Causality Test.

Although we know our analysis was limited to the data we collected and we understand that there are many other factors that play into the number of riders on alternative transportation, we are pleased that correlation and causation did show through the data like we thought it would.

We believe that our analysis acts as an initial proving point that further analysis could prove to be of great benefit to public transportation agencies across the United States. Modeling estimated transportation ridership while factoring in gas prices could be used in helping utilize resources more efficiently.

Further Research

There are many different directions that we could go to further our research and there are still many questions looming. To start, branching off to see if gas prices and/or alternative transportation methods have any impact on the quality of air in the United States would be an interesting topic to follow up on. This information could be utilized by green initiatives and lobbyists to support more maintaining and building infrastructure for alternative travel methods in our cities.

Adding on to our current analysis, we could break down the transit data set even further into specific modes of transportation provided in the data set. Getting down to a detailed level would help transportation agencies even greater by seeing how gas prices will affect ridership on their buses, trains, ferries, taxis, etc.

Lastly, the next big step would be to add predictive modeling to try to forecast ridership numbers based on gasoline price projections and estimates. We could try using historical data to predict gas prices ourselves but there are so many factors like sudden political changes and policies in the United States that we would rather have expert sources provide the forecasts to build our models on. This research would most likely be the greatest help to transportation agencies as they can plan their resources even further ahead into the future.

References/Works Cited

U.S. Energy Information Administration,
Independent Statistics and Analysis for Petroleum & Other Liquids
https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=RWTC&f=M
Available data by year and month of price FOB (Dollars per Barrel)

The 2014 plunge in import petroleum prices: What happened?
by Dave Mead and Porscha Stiger
https://www.bls.gov/opub/btn/volume-4/pdf/the-2014-plunge-in-import-petroleum-prices-what-happened.pdf

Cross-checking automated passenger counts for ridership analysis,
by Simon J. Berrebi, Sanskruti Joshi, Kari E. Watkins
https://digitalcommons.usf.edu/cgi/viewcontent.cgi?article=1900&context=jpt

Will transit recover? A retrospective study of nationwide ridership in the United States during the COVID-19 pandemic
Abubakr Ziedan, Candace Brakewood, and Kari Watkinsb
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10290420/

United States Labor Force Participation Rate
U.S. Bureau of Labor Statistics
https://www.bls.gov/charts/employment-situation/civilian-labor-force-participation-rate.htm

United States Civilian Unemployment
U.S. Bureau of Labor Statistics
https://www.bls.gov/charts/employment-situation/civilian-unemployment.htm

Spearman's Correlation
Stats Tutor
https://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf

Appendix

Figure 16. Participation Rate in US Workforce Over Time

Figure 17. US Civilian Unemployment Over Time

Figure 18. Average Weekly Gasoline Prices in US by Grade

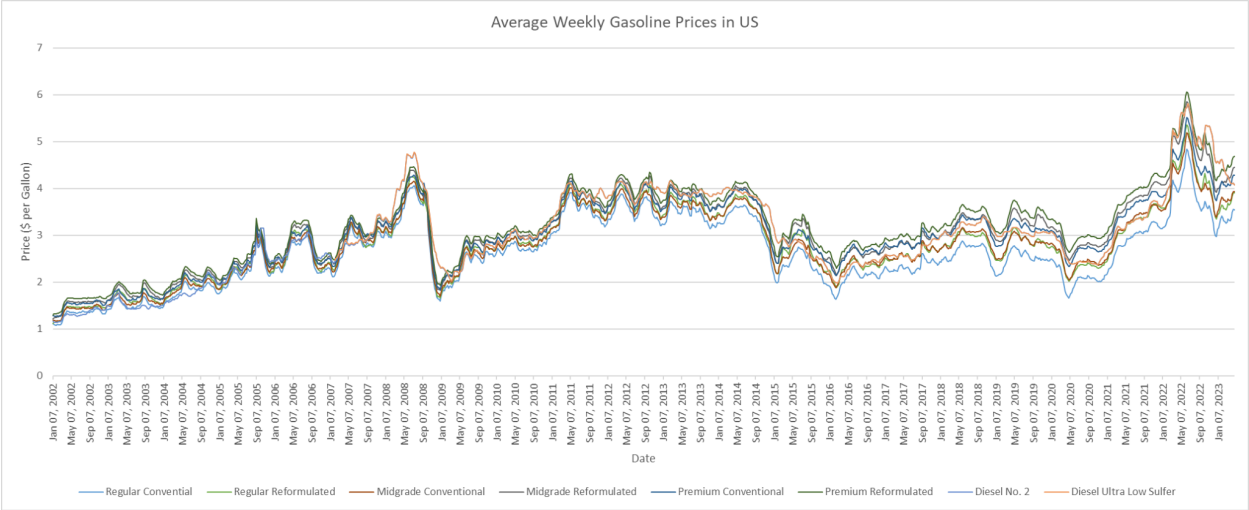


Figure 19. Average Weekly Gasoline Prices in US by Formulation Averages