Analyzing the Impact of Fare Prices and Other Factors on Public Transportation Ridership

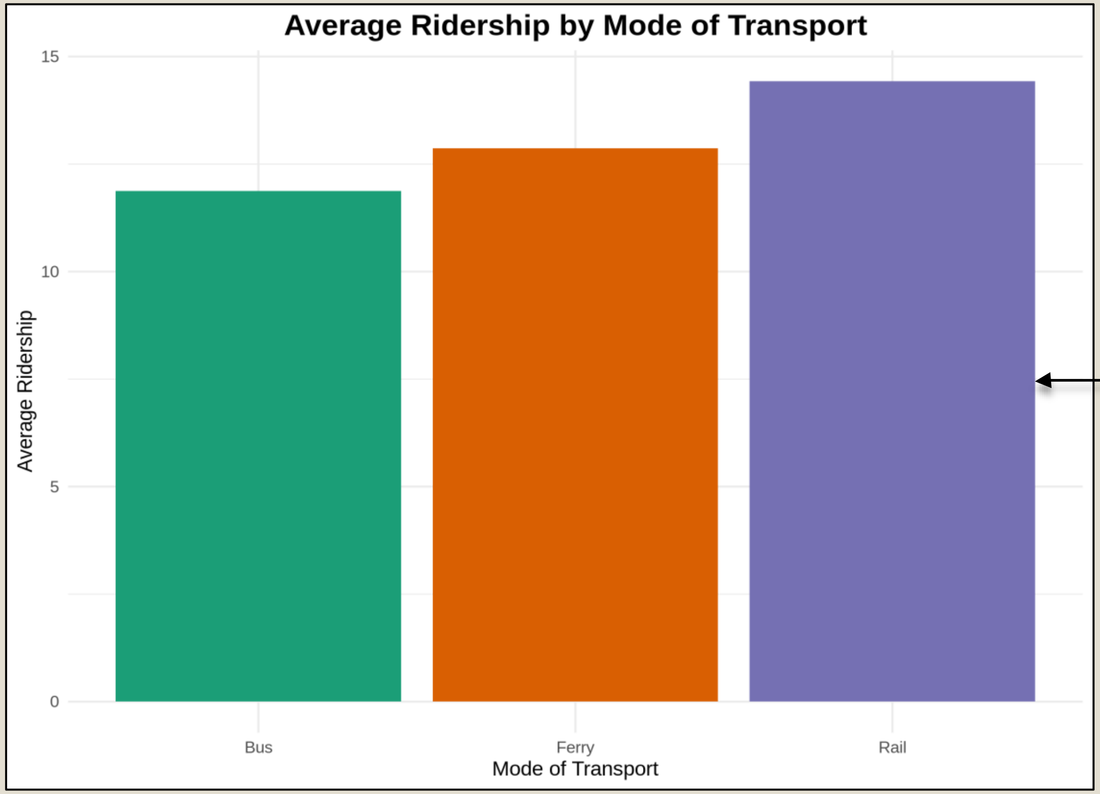
Introduction

Public transportation is a crucial service, that provides much-needed mobility in densely populated urban areas. Even with its necessity, several factors influence ridership patterns, with many assuming fare prices to be the sole factor with the most impact. This assumption ended up creating a common myth, causing many to believe that increase in fare prices lends to lower usage of public transportation. But of course, this is far from the truth. Various factors such as the type of service, the size of the service area, and even economic conditions all play a role in shaping travel behavior.

This analysis aims to explore fare prices along with other influencing factors, such as operating expenses, trip length, service area population, etc., to gain a proper understanding of what truly drives public transportation ridership. By examining these variables, I hope to challenge the existing beliefs and provide valuable insight that can improve transportation services.

Who is the intended audience ?

This analysis could potentially help professionals working in transportation infrastructure and economics by providing a better understanding of the complex factors that influence public transportation ridership. With the intention of demonstrating that fare prices alone do not dictate public transportation usage, the analysis offers a unique perspective aimed at challenging oversimplified assumptions and supporting data-driven strategies to improve transit systems and ensure sustainability in transportation planning.



From the plot above, we can see that Rail is the most popular among the 3 modes of transportation.
Why is that? Well, it's pretty simple. Rails are much more efficient and are not affected by traffic. They also operate on dedicated tracks and routes while also being well-integrated with the urban landscape, providing ease of access to the public. Overall, they are considered the most reliable.

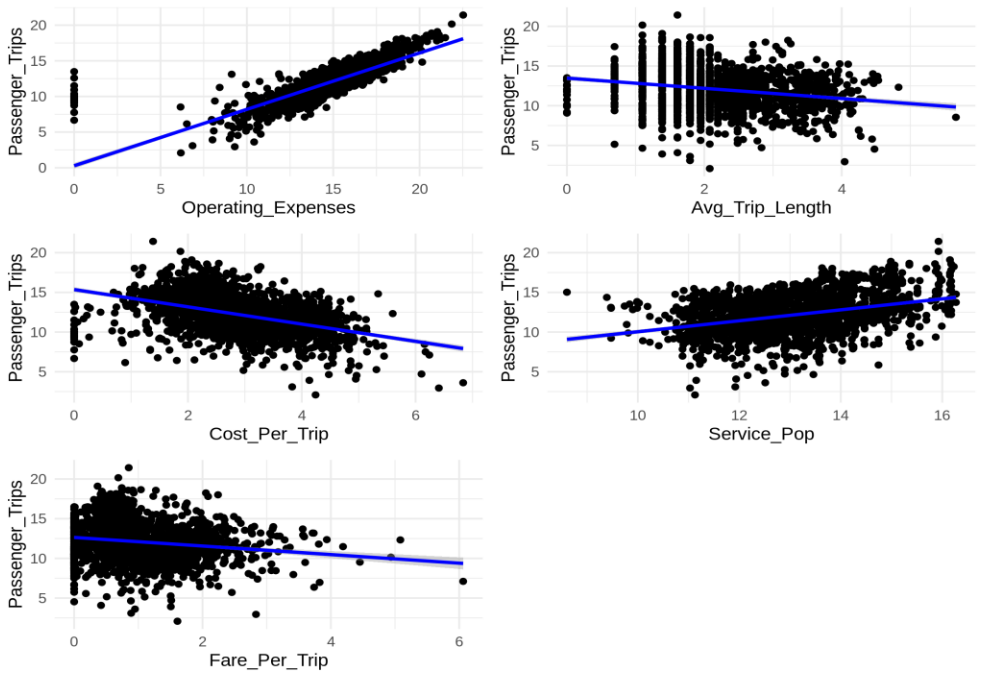
Hypothesis

While it is generally accepted that fare prices can influence ridership, my goal is to investigate the relationship between ridership and the combination of fare prices along with other significant factors. Therefore, my hypothesis is as such :

Null Hypothesis (H_0): Ridership is not significantly affected by fare prices and other factors.

Alternative Hypothesis (H_1): Ridership is affected by fare prices along with other factors.

Given the nature of the hypothesis, I will first identify the important key factors using Principal Component Analysis (PCA), Factor Analysis (FA), and Feature Importance techniques. Once the factors have been determined, I shall test their relationship with ridership using a Linear Regression model to assess whether the Null Hypothesis can be rejected.



Why Linear Regression for Hypothesis Testing?

From the graph, it can be understood that a few variables are linearly weak. While there is some variability, there is no strong indication of non-linearity.

As such, I shall use a Linear regression model for hypothesis testing instead of a Polynomial Regression Model.

	Service_Pop	Service_Area_SQMI	Passenger_Miles	Passenger_Trips	Avg_Trip_Length	Fares	Operating_Expenses	Cost_Per_Trip	Fare_Per_Trip	City_Freq	Agency_Freq
Agency_Freq											
City_Freq											0.66
Fare_Per_Trip										0.027	-0.04
Cost_Per_Trip									0.241	0.029	0.039
Operating_Expenses								0.086	-0.032	0.165	0.1
Fares							0.53	-0.057	0.496	0.118	0.099
Avg_Trip_Length						0.126	-0.164	0.132	0.488	0.056	0.038
Passenger_Trips					-0.216	0.5	0.809	-0.44	-0.166	0.159	0.1
Passenger_Miles				0.884	0.199	0.529	0.726	-0.366	0.016	0.175	0.108
Service_Area_SQMI			0.258	0.143	0.334	0.203	0.165	0.055	0.168	0.336	0.345
Service_Pop		0.687	0.424	0.363	0.201	0.302	0.343	-0.025	0.112	0.499	0.45
UZA_Pop	0.606	0.231	0.311	0.301	0.048	0.215	0.27	-0.045	0.063	0.266	0.164

1.0
0.5
0.0
-0.5
-1.0

A key observation is that the correlation between Passenger Trips/ Ridership and Fare Per Trip is -0.166.

This suggests that, as the fare price for each trip increases, ridership tends to decrease, which aligns with the established understanding.

However, since the strength of the correlation is weak, we can conclude that fare price is one of the factors that influences ridership, not the sole determinant. Other factors are likely to play a more significant role.

The key Predictors are as follows:

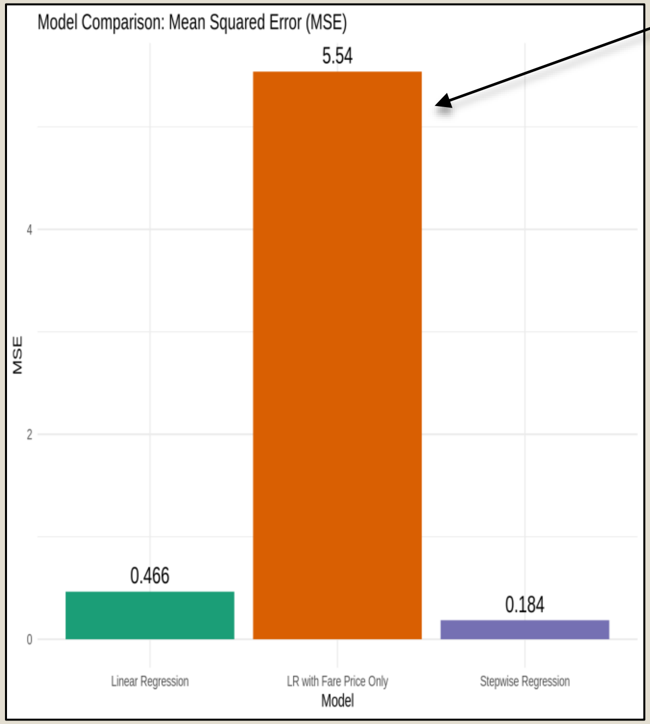
- **Operating_Expenses**
- **Avg_Trip_Length**
- **Cost_Per_Trip**
- **Service_Pop**
- **Mode**
- **Fare_Per_Trip**

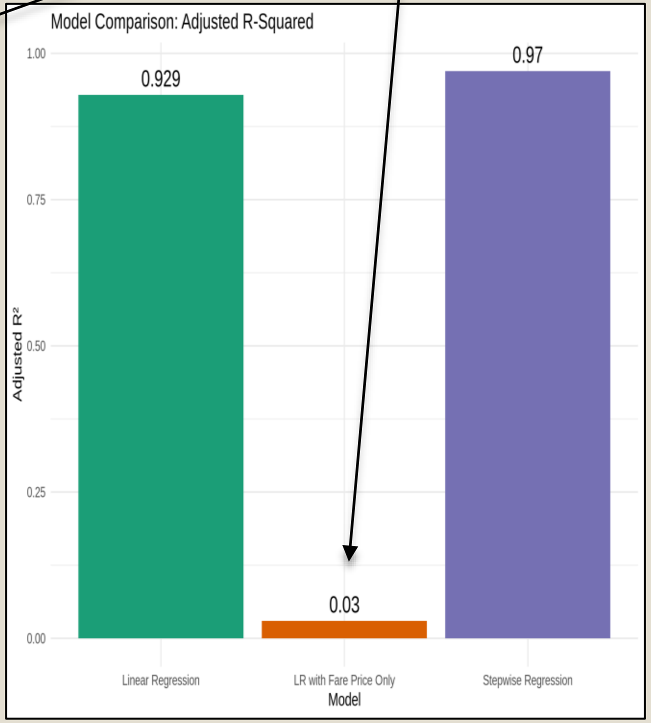
Results:

The P-values for the variables are below 0.05, indicating statistical significance and suggesting that ridership is influenced by fare prices along with other factors.
An F-test was also conducted on the model. Since the model's p-value was smaller than 0.5, it confirms that at least one predictor significantly contributes to explaining ridership, resulting in the rejection of the null hypothesis.

Model evaluation:

The model performs poorly when using only fare price as the predictor. It has a high **Mean Squared Error (MSE)** of 5.54 and explains 3% of the variance in ridership, as indicated by the low **Adjusted R² value**. This, of course, suggests that fare price alone is not a strong determinant of ridership, reinforcing the need to consider additional factors for a more accurate prediction.





The stepwise model selected a very similar set of factors to those chosen manually (Under the model name: **Linear Regression**), reinforcing the robustness of the selection process.

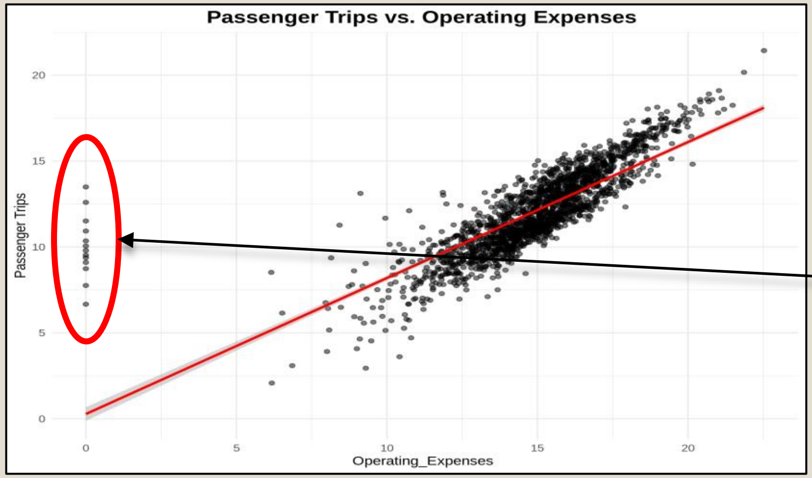
The difference in Adjusted R² value and MSE between the two models is also relatively small, indicating comparable performance.

Stepwise Model Variables: TOS, Mode, Service_Pop, Passenger_Miles, Avg_Trip_Length, Fares, Operating_Expenses, Cost_Per_Trip, Region

Manually Selected Variables (via Dimension Reduction): Operating_Expenses, Avg_Trip_Length, Cost_Per_Trip, Service_Pop, Fare_Per_Trip, Mode

Since stepwise regression selects factors that are the most impactful predictors, the strong overlap between both selections supports the fact that the **Manually Selected Variables** are not only influential but also reliable predictors of ridership.

Key Observations:



Something that surprised me was how much of an influence Operating_Expenses had over ridership. However, it does make sense that higher operation expenses often translate to better infrastructure, increased reliability, and improved frequency. In a way, a well-maintained and efficient transit system is always going to appear attractive to passengers, resulting in higher ridership.

That being said, there are some outliers where lower operating costs can lead to higher ridership. It could be representing a part of the population using older, well-established transit systems that require low operating costs while serving millions of people, such as the NYC subway or maybe buses with shorter routes and high passenger turnover.

Service_Population also has a decent influence over ridership. As the service population increases, the number of passenger trips also increases.

There are some noticeable outliers:

- This outlier could represent college towns, where students may not own personal vehicles and heavily rely on public transport.
- This outlier could represent dense urban cities, where transit access is high, but ridership is lower than expected due to the availability of alternative transportation options (walking, biking, or personal vehicles).

Another observation is that it does appear that small trips are more frequent than large trips. However, their considerable overlap indicates that trip length alone is not a strong predictor. Both types of trips exist across different populations.