Rail Transit Scores & Statistics - Jet Chanchom & Felix Menges

Introduction:

Almost half of the United States has no access to public transportation despite the many benefits public transportation provides for communities. According to the American Public Transportation Association, "every $1 invested in public transportation generates $5 in economic returns." Public transportation can save households more than $13,000 annually if public transit replaces one car.[1] The Infrastructure Investment and Jobs Act (IIJA) was signed into law by President Joe Biden in 2021 to provide new funding for infrastructure projects across the United States to pursue these benefits. The IIJA was the biggest investment in public transportation in US history, providing $108 billion for public transit, with $193 million of that investment going towards the Public Transportation Innovation Program.[2]

For our project, we decided to limit our scope to rail transportation due to the vast variety of non-rail modes of transportation across the United States. According to the Federal Transit Administration, some non-rail forms of transportation include buses, ferries, and taxis. Even within the category of buses, there is a lot of variety in service provided, from high-frequency bus rapid transit systems in urban areas to on-demand bus or paratransit services in rural areas. Though rail transport also has many forms—street cars, light rail, and heavy rail, to name a few—all forms of rail share one distinct characteristic of traveling across fixed rails.[3] Unlike buses, rail routes cannot use alternate routes. Because of this characteristic, rail transportation has less variation across different modes, making it easier to compare the rail systems across the United States for our analysis.

Another reason we are limiting our scope to rail is due to the advantages rail has over bus transportation. Unlike buses, rail transportation is generally more reliable and frequent due to having a right of way. Additionally, rail is perceived as safer and more comfortable, which can help convince car users to switch to public transportation.[4] Though rail may not be the best solution for every city, the benefits of rail travel are worth exploring to see if implementing or improving existing rail services will be a good fit for a community. Our project aims to analyze

[1] https://www.apta.com/news-publications/public-transportation-facts/
[2] https://www.transit.dot.gov/IIJA
[3] https://www.transit.dot.gov/ntd/national-transit-database-ntd-glossary
[4] https://www.sciencedirect.com/science/article/pii/S0967070X02000094?via%3Dihub

rail public transit across the US and the factors impacting ridership to provide insights that public transit agencies can use to improve their service with the Public Transportation Innovation Program.

Data:

For this project, we used the Federal Transit Administration's Complete Monthly Ridership (with adjustments and estimates) data from the National Transit Database and the Center for Neighborhood Technology's AllTransit Rankings for cities with a population of over 10,000.

The Federal Transit Administration (FTA) data comes in an Excel spreadsheet downloaded from the FTA's website.[5] Specifically, we used the Master, Calendar Year UPT, and Calendar Year VRM sheets filtered to only show rows with rail modes. We created separate DataFrames for each sheet and merged them using an inner join on the *NTD ID, Agency, Mode/Type of Service Status, UZA Name, Mode, TOS,* and *Year* columns. For the Calendar Year UPT and Calendar Year VRM DataFrames, we transformed the data from wide to long data before merging it with the Master sheet using all the columns listed previously, except *Year*, to create a merged DataFrame.

The AllTransit Ranking compiles transit connectivity, access, and frequency data to determine the best cities for public transportation.[6] To get the AllTransit Rankings, we needed to scrape data from the rankings table on the webpage. By default, the webpage shows the top 10 cities with a population over 250,000, so when we web scrape, we will need to include the click() command to select the correct filters, all cities with at least a population of 10,000, from the webpage.

Both data sets contained city and state information, allowing the data to join horizontally using an inner join; However, data cleaning was needed before merging. In the AllTransit dataset, the city and state information was in one column, so we had to split the column into separate columns. We also had to rename the city and state columns in our previously merged DataFrame and change the data to title case to match the format of the city and state columns in our AllTransit DataFrame. When we initially merged the two DataFrames, we noticed that some

[5] https://www.transit.dot.gov/ntd/data-product/monthly-module-adjusted-data-release
[6] https://alltransit.cnt.org/rankings/

cities were excluded due to discrepancies in how their names were written. We manually changed their names to match since only three cities were affected.

The final merged DataFrame had missing values to be dealt with. Most of our missing values were in the *UPT* and *VRM* columns. We realized these null values were caused by the service being inactive or unbuilt. We could drop these rows, taking care of most of our missing values, but for our remaining missing values, we imputed the median value between the year before the gaps in reporting and the first year reporting resumed. By doing this, we assumed that ridership for the missing years will be a value between those two. Once we removed all of our missing values, we only needed to drop some unnecessary columns that wouldn't be used in our analysis. Our final cleaned dataset contained 2010 rows and 16 columns.

Table 1: Data Dictionary

Field	**Type**	**Source**	**Description**
Agency	Text	DOT	Name of service provider agency
Mode	Text	DOT	Mode of transportation: • Alaska Railroad (AR) • Cable car (CC) • Commuter rail (CR) • Heavy rail (HR) • Hybrid rail (YR) • Inclined plane (IP) • Light rail (LR) • Monorail/Automated guideway transit (MG) • Streetcar (SR)
Service Area SQ Miles	Numeric	DOT	Total area covered by the service area
Unlinked Passenger Trips (UPT) [Ridership]	Numeric	DOT	Total number of times a person has boarded the railway
Avg Cost per Trip	Numeric	DOT	Expenses divided by total number of trips
Avg Fares per Trip	Numeric	DOT	Total fares divided by total number of trips
Year	Numeric	DOT	Year of data collected
Vehicle Revenue Miles (VRM)	Numeric	DOT	Actual & scheduled miles during revenue service (excluding maintenance & training)
City	Text	All Transit	Name of City

State	Text	All Transit	Name of State
Transit Connectivity Index (TCI)	Numeric	All Transit	A normalized ranking of the sum of weekly bus & train traffic per region. Higher ranking means denser transit connectivity Scaled [0:100]
Jobs	Numeric	All Transit	Quantity of jobs within 30-minute access of public transport
Trips/Week	Numeric	All Transit	Transit Trips per Week within ½ Mile
Routes	Numeric	All Transit	Total number of Transit Routes within ½ Mile
%Transit	Numeric	All Transit	% of commuters who use transit
Population	Numeric	All Transit	Population of Region

Analysis:

1. **What's the agency with the highest ridership per state?**

By aggregating *Agency* & the summation of their *UPT* throughout the years, we could find the transit agencies with the most ridership from 2002 to 2024. In Table 2, we found that the Massachusetts Bay Transportation Authority, Washington Metropolitan Area Transit Authority, and Chicago Transit Authority are the top 3 agencies across the United States, indicating that rail transit is less defined by geographical region. We expected that New York would have made this list due to how famous New York City's subway system is, but unlike other cities, New York City splits its rail services across different agencies, causing it not to have an agency make the top 5.

Table 2: Unlinked Passenger Trips by Agency per State

state	Agency	UPT
MA	Massachusetts Bay Transportation Authority	5,302,250,865
DC	Washington Metropolitan Area Transit Authority	5,249,483,547
IL	Chicago Transit Authority	4,299,462,519
PA	Southeastern Pennsylvania Transportation Authority	3,164,852,196
CA	San Francisco Bay Area Rapid Transit District	2,315,514,732

2. How has ridership changed over time since 2002 across the US as a whole and by city?

Figure 1 was derived from aggregated data of *year* and the yearly sum of the *UPT*, displaying historical data in blue and forecasted data in red. The forecasted data was predicted with exponential smoothing with a multiplicative trend. It was found that transit peaked during 2015, declined till 2019, with the COVID-19 shutdown causing nearly a 2/3rds drop of ridership. A similar case is found in Figure 2, which breaks down the ridership of the top 10 states with the most ridership. Current ridership has only recovered to half of pre-pandemic ridership, with a trend forecasted to return to normal levels by 2030. It is unknown why rail ridership has not yet returned to pre-pandemic levels, but some assumptions could be made. The rise of on-demand personal transport businesses (i.e., Uber and Lyft) may have potentially demoted the usage of rail or public transport in general. Furthermore, the shutdowns may have caused pressure in some aspects, such as vehicle ownership and remote work, all reducing ridership.

Figure 1 US Ridership Overtime Trendline (2002-2024) & Forecast (2025-2030)

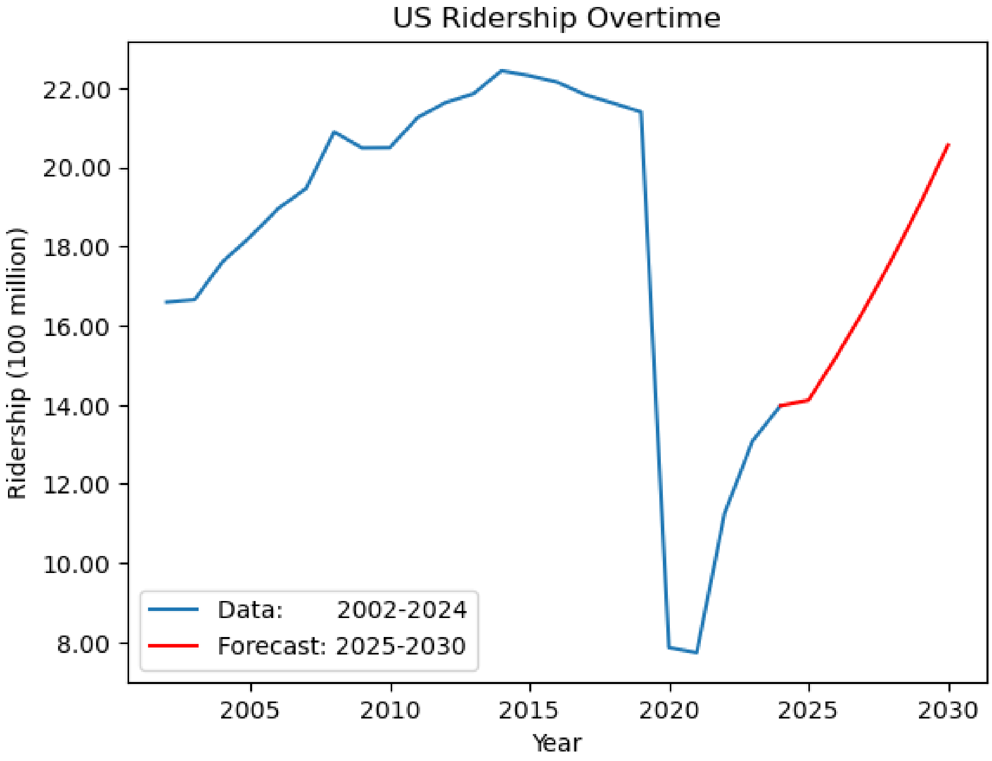
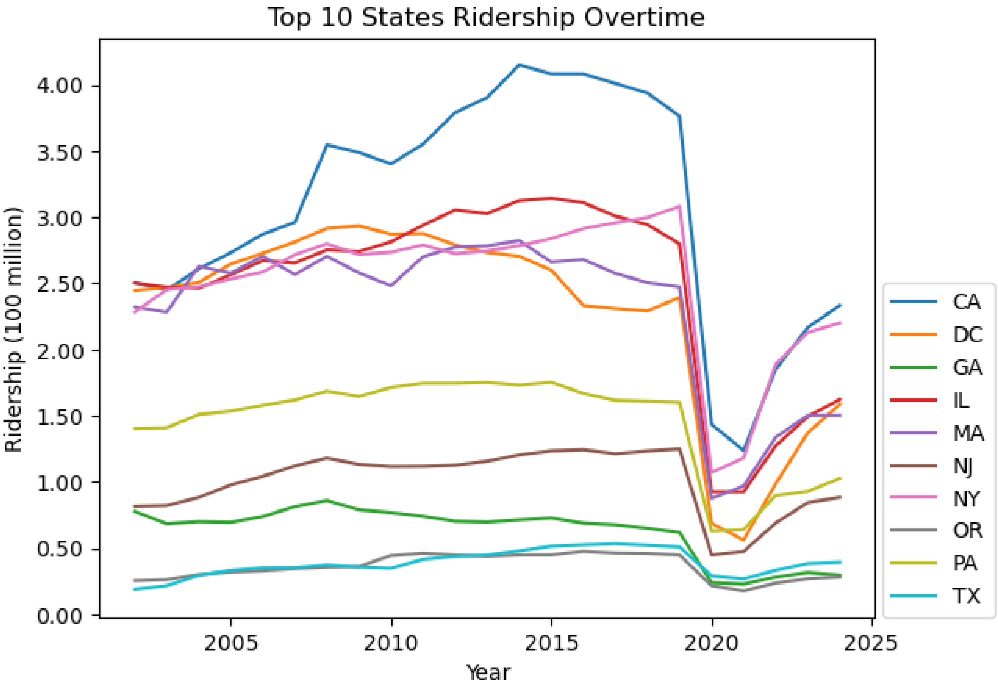


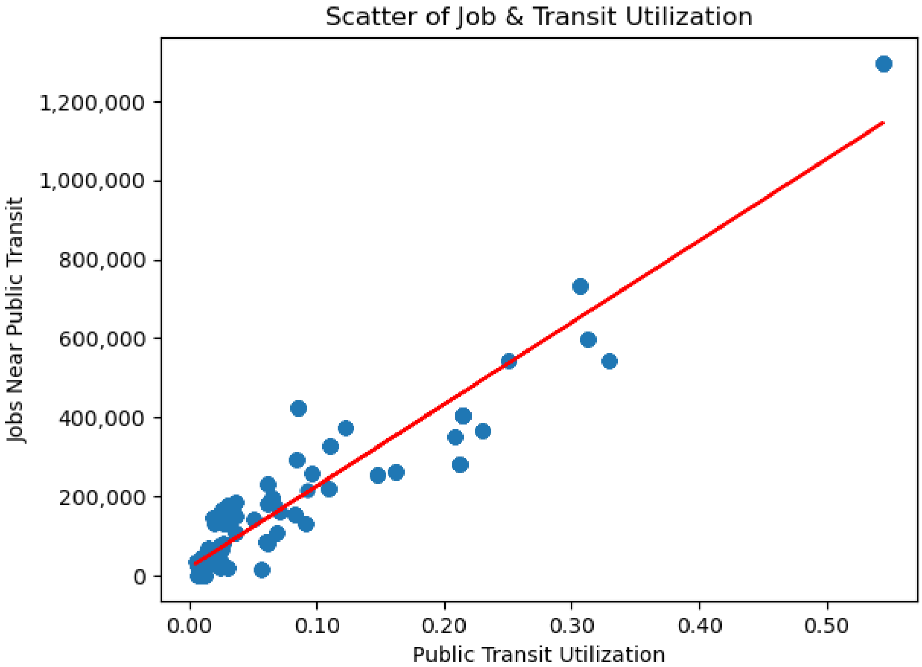
Figure 2 Top 10 States With the Most Ridership (2002-2024)



3. How does access to jobs impact transit ridership?

To determine if cities with more jobs accessible to public transit have higher ridership, we created a scatter chart comparing *%transit* with *jobs*. These 2 metrics give us an idea of how many jobs are available near transit and how many employees use public transit to get to work. When we initially plotted our scatter chart, we noticed a strong trend of the number of *jobs* near public transit increasing as public transit utilization increases. To test this observation, we added a trend line using poly fit and saw that our data matched up closely with our trend line, as seen in Figure 3 below. Additionally, we found a Pearson Correlation of .947 between *%transit* and *jobs* and a p-value less than our alpha value of 0.05 when we performed a Pearson Correlation Test, meaning there is a significant, strong linear correlation between jobs and transit utilization. Based on these tests, we can conclude that access to jobs via public transit leads to higher ridership.

Figure 3 Scatter Plot of Job and Transit Utilization with Trend Line



4. Which mode of rail transportation is most popular? Does it vary by population of a city?

Figure 4, grouping by *Mode* and counting by *Mode*, shows the most popular forms of rail transportation. Light rail and commuter rail make up the majority of rail transport within the United States. Table 3, made by binning *population* and then grouping by population for the count of rail *Mode*, shows the most common mode by city classification. Commuter rail makes up both the lower and upper ends of city sizes, while light rail makes up the majority of the cities and metropolises.

Figure 4 Scatter Plot of Job and Transit Utilization with Trend Line

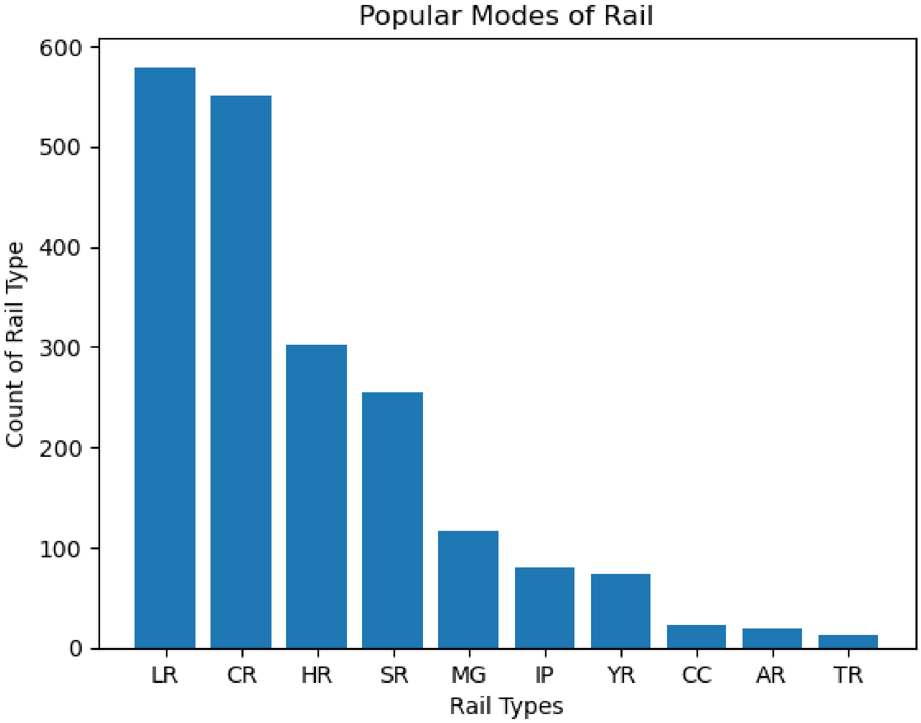


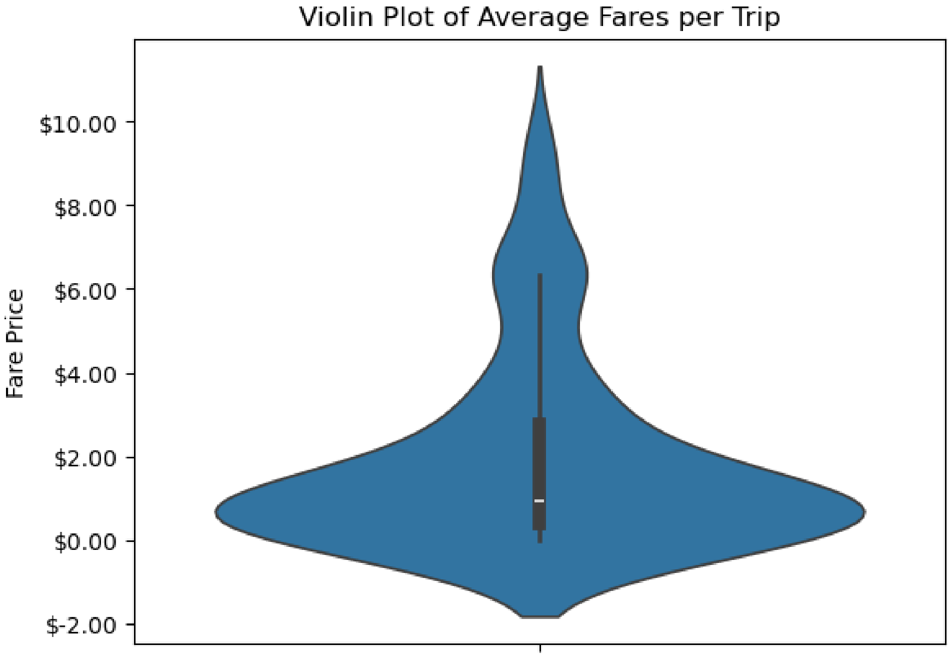
Table 3: Most Numerous Mode of Rail by City Class

City Class	Mode	Mode
Metropolis	LR	266
City	LR	201
Large Metropolis	CR	120
Small City	CR	100
Town	CR	99

5. Which factors impact fare price?

To answer this question, we used *Avg Fares Per Trip FY* as our metric for fare price because it represented the average fare a passenger would pay for a trip. Our data only included information for the most recently completed year, so we had to filter our data to show results for the year 2024, which left us with 109 rows of data to work with. Before we analyzed which factors impact fare price, we explored how fare price differed across different systems. We found that fare prices ranged from $0 to $160.25, with an average fare of $3.89 and a median of $0.99. When looking at the spread of our data, we saw that 75% of fares were under $2.85, suggesting that public transportation agencies keep fares low regardless of different factors. To help confirm these findings, we plotted a violin plot shown in Figure 5 below. When looking at the violin plot, we can see a high frequency of fares under $2, with the number of fares tapering off as fare price increases, with a small surge around $6.

Figure 5 Violin Plot of Average Fares per Trip



Once we finished analyzing fare price, we decided to see if *UPT*, *trips per week*, *population*, *service area*, *average cost per trip*, and *mode* impacted *fare price*. Since the *average fare price* and all our factors except *mode* were continuous variables, we plotted scatter charts to see if there was a linear relationship between our variables.

Figure 6-9 Scatter Plots of features with no or weak linear relationship with fare price

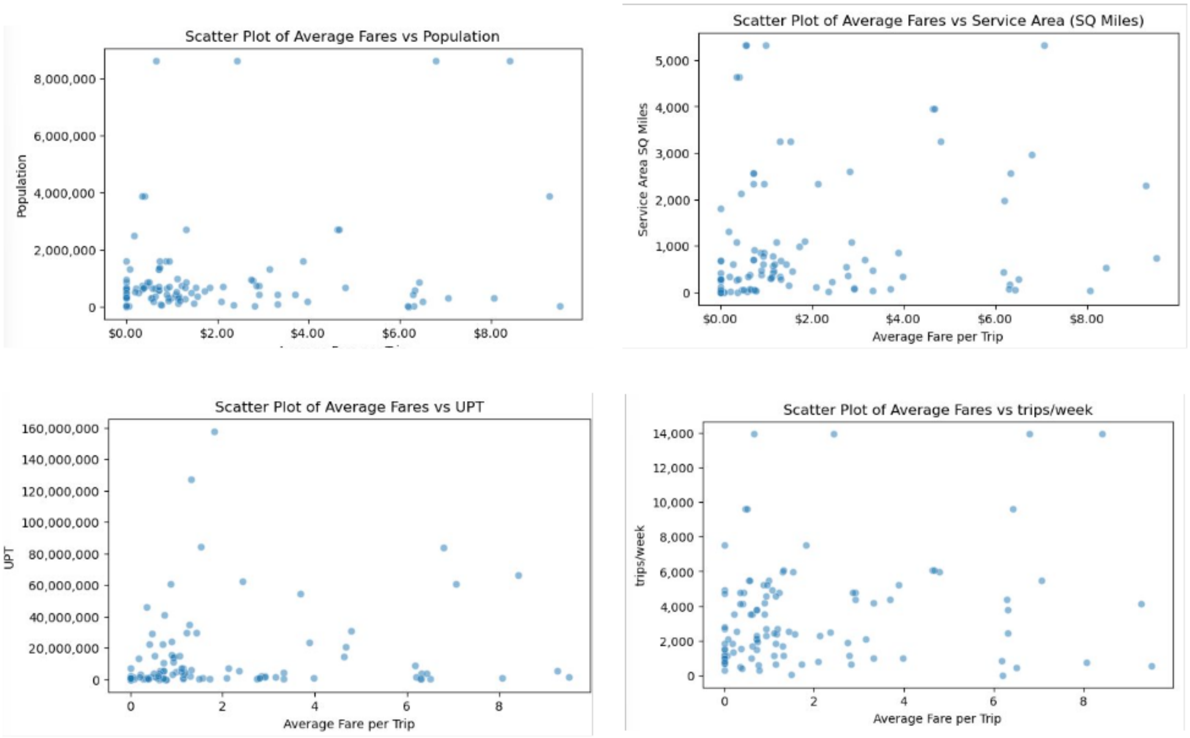
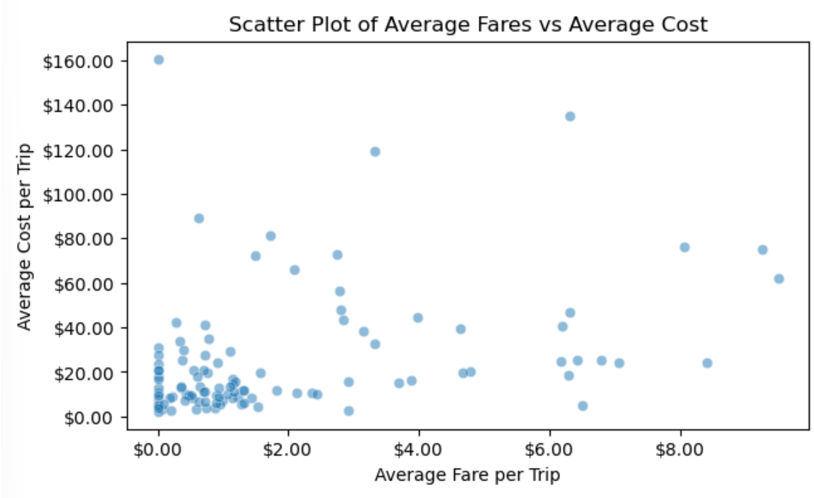


Figure 10 Scatter Plot of Average Fares by Average Cost

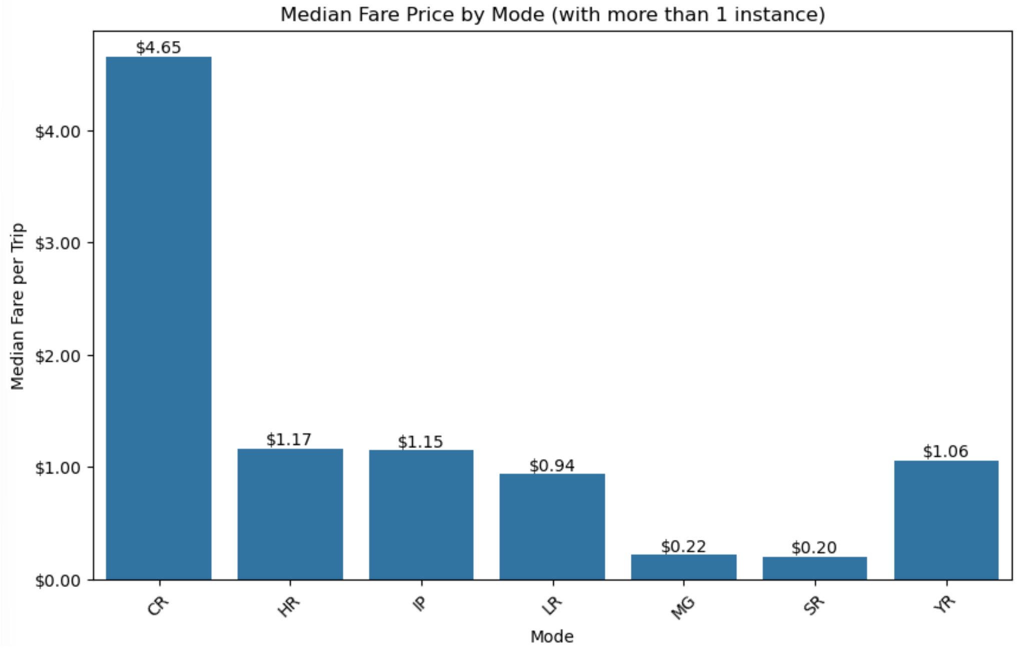


As Figures 6-9 above show, our scatter plots indicated that only *average cost per trip* had a linear relationship with *fare price*. To confirm our observations, we found the Pearson Correlation Coefficient and performed the Pearson Correlation Test to determine if a statistically significant linear relationship existed between the two variables. Only *average cost per trip* gave us a Pearson Correlation greater than 0.3 with a correlation value of .6827, suggesting a moderate linear correlation between average fare and cost per trip. Additionally, when performing the Pearson Correlation Test, only *average cost per trip's* p-value was less than the alpha value of 0.05, meaning that a significant correlation exists between *average fare price* and *cost per trip*. This observation makes sense when you consider that more expensive systems to run might need to charge higher fares to make up for operating costs. Surprisingly, there was no relationship between ridership and fare price—we thought high fares would decrease ridership, but fares are likely subsidized since most fares in our data are under $3.

The last factor we analyzed was rail *mode*, our only categorical variable. Initially, we plotted a bar graph to see how *fare price* is impacted by *mode*, when we noticed that Alaska Railroad was an outlier in our data with a median fare price of $160.26, nearly 25 times higher than the next highest median fare. Unlike other rail systems in our data, Alaska Railroad serves the entire state of Alaska and is comparable to a service like Amtrak, which is not in our data.[7] After noticing this, we filtered out Alaska Railroad and all other modes with only one instance. Once we filtered our data, we performed a Kruskal-Wallis test and found that the median fare prices were significantly different across different modes, suggesting that *mode* impacts *fare price*. Figure 10 below shows an updated bar chart and presents a wide gap between our most expensive *mode*, Commuter Rail, and our cheapest *mode*, Streetcar. Earlier, we found that *operating cost per trip* impacted *fare price*, so it makes sense that modes like Commuter Rail, which travel long distances, would have higher fare prices than streetcars, which usually operate on a smaller scale.

[7] https://www.alaskarailroad.com/ride-a-train

Figure 11 Bar Chart of Median Fare Price by Modes with more than 1 instance



6. Which factors can be used to predict how much ridership a city will have?

For this question, we created four regression models using *UPT* as our target variable to predict ridership. We chose *Mode, VRM, Avg Fares Per Trip FY, jobs, routes,* and *tci* as our features, and split the data using an 80/20 training-test split. *UPT* is a continuous variable, so we had to use regression models, and in particular, we used scikit-learn's linear regression, Huber regressor, decision tree regressor, and gradient boosting regressor models. Given the R-squared scores of our model, we found that the decision tree regressor performed the best with a score of .915, with the gradient boosting regressor coming in a close second with a score of .908. Decision tree regressor and gradient boosting regressor are both tree-based models, so they likely performed well because they are more equipped to deal with more complex, non-linear relationships like those found in our data.

Conclusion:

Our project explored public rail transit in the United States, analyzing rail ridership and factors related to ridership, such as jobs and fares. We found that the Massachusetts Bay Transportation Authority, Washington Metropolitan Area Transit Authority, and Chicago Transit Authority are the top 3 agencies by ridership from 2002 to 2024. All these agencies serve large metro regions, so we expected higher ridership. Additionally, when we look at ridership by state, we see that more populous states tend to have higher ridership, with the top 3 states—California, New York, and Illinois—all falling in the top 10 states by population. When we look at ridership across the United States, we see that ridership dropped nearly 2/3rds during the pandemic and has yet to recover. It is unknown why rail ridership has yet to recover, but recovery will be slow, as we forecast that ridership won't recover to pre-pandemic levels until 2030. We found that access to jobs via public transportation has a strong linear relationship with ridership, while fare price did not. In terms of fare price, only operating cost per trip and mode had a relationship with fare price due to most fares being subsidized to keep prices low for passengers.

One limitation of our project was that some variables, such as population, fare price, and all the features from the AllTransit Ranking, only accounted for 2024. Our data included UPT and VRM data from 2002 to 2024, but we couldn't use those data points before 2024 with many of our variables. Future work on this project could involve finding data for population, fare, and other features before 2024, so we could perform more in-depth analysis on how those features changed over time and how that impacted ridership. The AllTransit Ranking website had some data inaccuracies, such as the population of the much smaller cities. Additionally, future analysis would benefit from analyzing ridership based on the metro area instead of the headquarters city. Some modes, like commuter rail, serve multiple cities and towns in a metro area, so analyzing based on the metro area would provide a more comprehensive picture of ridership. Analysis could also be more limited in scope, such as purely metro, commuter, or streetcar data, as some modes were not as well tracked (i.e., Alaskan Rail), to give us data that is more granular and accurate. However, this comes at the cost of being less representative of the greater US.