Analysis of Factors Influencing US Public Transit Ridership Using the National Transit Database*

Diego Olaya†
Department of Physics
University of Colorado Boulder
Boulder CO, USA
diego.olaya@colorado.edu

Yashwant Gandham
Department of Data Science
University of Colorado Boulder
Boulder CO, USA
yaga6775@colorado.edu

Sasidhar Reddy
Department of Data Science
University of Colorado Boulder
Boulder CO, USA
saga3727@colorado.edu

ABSTRACT

Access to reliable transportation is essential to modern life. In most developed countries, public transit coexists with private vehicles as a common choice for the public. Places like Japan and Western Europe have robust and highly utilized public transit infrastructure in their major cities as well as developed intercity rail. In the United States, by contrast, private vehicles are overwhelmingly the most popular mode of transport, and even when public transit is available it is not always used. This study uses data from the National Transit Database to investigate the factors that contribute to public transit usage in the United States. Using a combination of linear regression and other regression learning techniques, our work finds that service frequency and low fares are the most important factors in promoting transit usage.

KEYWORDS

Public transit, United States, infrastructure, passenger rail.

1 Introduction

How a person gets around affects almost every aspect of their life. Transportation is a critical component of modern living: the capacity to get where you want to go is critical no matter where a person lives. Access to transportation affects where someone can work and socialize, what services they have access to, and their ability to access food and supplies. Cut off someone's access to transport, and you cut off their ties to their communities and society. Thus, promoting general access to transportation should be a priority for governments at all levels. How to go about doing this is a complicated policy question, but one common solution, especially for cities, is to invest in public transit systems alongside other infrastructure like roads, sidewalks, and bike paths.

In places like Western Europe and Japan, public transit is ubiquitous, and its presence is integrated in the broader society. In some cases, public transit becomes so emblematic of the cities it operates in that it becomes part of the cultural image of the place itself. Think of the London Underground, for example, whose station signs have become icons themselves. Japan's Shinkansen trains are a source of national pride, as well as a source of no little envy from transit advocates elsewhere. The New York City subway is a rare example of this in North America. The New York City subway is the exception, however. In many cases, even when public transportation is available in North America, people do not use it. The reasons for this disappointing trend are a subject of great interest to urban planners and researchers alike [4, 11].

This paper is by no means the first work to think about public transit's relationship to good urban design, nor is it the first to consider the question of how to encourage transit use more broadly. In the United States and Canada, advocacy groups have been working for years to encourage municipalities to develop alternatives to car travel by investing in changes to road infrastructure and urban design. They often focus on encouraging cycling and public transit, both to improve mobility for the public and to alleviate congestion. In addition, groups like these have been consistent advocates for investment in public transit solutions that work for the communities in which they are based. One positive about public transit design is that, because the service is designed to serve the public, the public often has a good idea of what they want in a good transit system. These are things like, safety, accessibility, frequency, ease of use, and cost. Academic studies support the importance of these factors, pointing out the importance of frequent, reliable service in attracting riders. Safety, accessible stops, and intuitive ticketing are also important in attracting riders [1, 3, 9].

There is also anecdotal evidence to suggest that transit in North America can work well given the chance. Places like Ottawa, whose public transit system has been plagued with problems, nonetheless have systems that perform significantly better than driving and walking in getting to major destinations within the city [2]. Other cities like Montreal are investing heavily in improvements, but there is still a long way to go [5].

Given the recent interest in investing in public transit, it is important to make sure that whatever money is being channeled to public transportation is well spent. This study uses data on transit agencies in the United States to draw conclusions about what factors are most important to keep in mind when planning improvements or additions to transit.

2 Methods

2.1 Data Collection and Processing

The data used in this work is taken from the National Transit Database (NTD). The database is managed by the Federal Transit Administration and is the primary data source for American transit data, as US transit agencies are required by law to report information about their operations to the NTD. The database contains information including capital expenditures, agency revenue, safety incidents, and ridership metrics [10]. It thus provides a good snapshot of the operations of transit agencies in the United States.

This work is intended to identify factors that contribute to public transit usage in the United States, so our sample of interest is the existing agencies in the United States. The NTD provides information about transit agencies by data category, and provides a search interface for researchers to find the specific data they are interested in. This work examines data from 2016 to 2022, taking information on agency capital expenditures, agency revenue, safety incidents, and ridership measures. Each of these categories are separate data products in the NTD and must be combined manually in the data processing phase.

Regarding issues of data bias, this is mitigated by the fact that transit agencies are required to report to the NTD by law, meaning that the database contains information for all agencies covered by the legal reporting obligations. The agency does allow more limited reporting obligations for the most remote and smallest agencies, but most are designated as full reporters, and even among full reporters, the agencies give a good cross-section of the United States. Agencies classified as full reporters include both the Washington Metropolitan Area Transit Agency (WMATA) in Washington, DC and the Regional Transportation District (RTD) in Colorado, which serve very different areas. For this reason, we are reasonably confident that any findings we make in this paper will have general applicability to transit behavior in the US.

2.2 Data Cleaning and Transformation

Processing the data downloaded from the NTD primarily required reconciling the different layouts that the Federal Transit Administration used in different years. As an example, data on capital expenditures can be found in the NTD for all years in our period of interest. For the 2022 data, the database has a separate webpage containing definitions of all the columns in the database as well as the format of the data. For the period of 2016 to 2021, the NTD includes the column definitions in the Excel file that they provide for download. In addition, the file formats change between XLSX and XLSM depending on the year, as do the included columns. In order to combine all of the data for our time period of interest, we need to reconcile the differences between the data tables before combining them into a master dataset. We did this by removing legacy identifiers that were not needed in our work as well as other additional information that was not collected in the earlier part of our period of interest.

The legal reporting obligations for agencies in the US is also beneficial when assessing the data for accuracy and cleaning up nonsensical values or missing data. The Federal Transit Administration (FTA) does a preliminary review of the data the agencies report and flags entries that the agency considers questionable, allowing for these dubious numbers to be excluded from analysis. In addition, the FTA will sometimes publish revisions to the data to correct estimates or improper counts, most notably with respect to the measurements that are used to estimate ridership. The work that the FTA does in doing initial cleaning of their published data allows us to proceed with the assumption that the numbers we are using are accurate.

Data cleaning, transformation, and preprocessing was done in Python using the Pandas package [6]. Details about the exact methods and the code used can be found in the project git repository, which is hosted on GitHub.

2.3 Model and Feature Selection

After downloading and processing the NTD data related to the categories of interest, we obtain a data set that has several columns focusing on highly specific features as well as more general measures that capture information about a more general concept. For example, the difference between total cyclists injured in safety incidents is a highly specific measurement of how safe the transit agency's services are, while the total number of injuries in much more general but still provides information about safety. When selecting the features to use to train the predictive models, we preferred broader measures that captured information about a key concept without being excessively narrow.

In addition, some of the features we wanted to include had to be derived or substituted for a proxy measurement. For example, the NTD does not contain direct information about the fares on US transit agencies, only the amount that each agency makes annually in revenue from fares. It also includes the total number of unlinked passenger trips (UPT), so an estimate of the fare can be made by dividing the revenue by the number of trips. This is generally reliable, but occasionally produces results that had to be discarded because they were unreasonably large. Regarding proxies, the NTD also does not directly contain information about an agency's service frequency, so this quantity is estimated by using the number of Vehicles Operated at Maximum Service (VOMS).

We list the features selected for use in model training below, along with a short explanation of the rationale behind each one.

1. **Vehicle Revenue Hours**: This quantity measures the amount of time that passenger vehicles are available to the public to ride. This is a measurement of availability, the thought being that higher revenue hours means more time during which the public can use the system.
2. **Passenger Miles Traveled:** This is a ridership measure that can be used to estimate the length of the trips taken

by passengers. A higher total indicates that riders are traveling for longer distances.

3. **Operating Expenses**: This is the total cost associated with running the transit system, including labor, maintenance, fuel, and administrative costs. This can also be used a measure of organizational efficiency, where lower costs per trip or per mile indicate leaner organizations.
4. **Fares (Total, Paid by Organizations/Passengers):** Total revenue from fares, with optional divisions on fares paid by passengers and those subsidized by organizations.
5. **Population:** Accounting for the population of the area served by a transit agency avoids the mistake of assuming that agencies in big cities are more successful when in reality ridership may only be higher because there are more people around to use the system.
6. **Vehicles Operated in Maximum Service:** The maximum number of passenger vehicles operated by the agency during peak service times. This measurement as discussed above, serves as a proxy for service frequency, but it also provides information about service capacity.
7. **Total Capital Spending:** The total amount of money spent by an agency on capital expenses, which include vehicles, guideway, buildings, and other non-labor expenses.
8. **Safety Statistics:** These included data on minor, major, and total injuries; total fatalities; and property damage caused by accidents. This is the factual record of each agency's safety.

To generate predictions regarding the relative importance of these features, we focus on models that generate continuous numerical predictions as their outputs. This means that we primarily select regressors. We focused on training a random forest regressor to avoid accidentally fixating on an extraneous variable, since the random forest generates many simple models and weights them to generate a prediction. While this technique can give us information regarding the relative importance of the features after the model is trained, it does not allow for statistical inference. For that, we use linear regression, which lets us draw conclusions regarding the significance of each of the individual features on the independent variable, the unlinked passenger trips. Other models are occasionally used to see if the choice of model changes the relative parameter weightings. We divided the analysis to separate rail systems and bus systems, as these have significantly different physical characteristics that makes it reasonable to assume the possibility that they would have different key features. Models were trained using scikit-learn and stats models [7, 8].

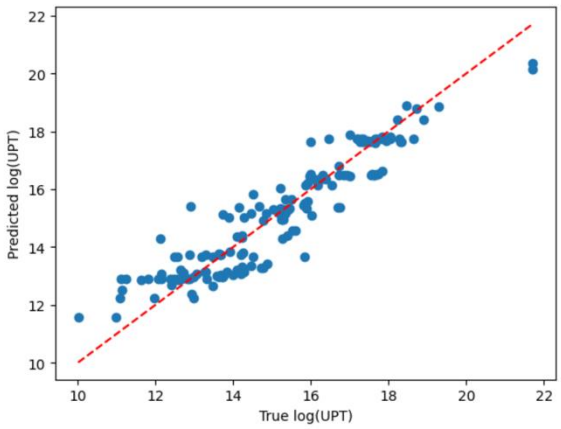
3 Evaluation and Results

3.1 Rail

Due to the different fixed costs for rail infrastructure as compared to buses, we chose to analyze rail and bus systems separately. The NTD database indicates the type of transit mode for each entry and separates rail and bus modes for agencies that operate both. Thus, it is rather simple to select only rail transit from our processed data and fit the models to this subset.

The first model fit to the data was a random forest regressor, setting the maximum depth of each tree to three levels and training 100 trees. The R^2 value of the trained model was 0.91, which suggested a highly performant model, but closer examination revealed an artificial prediction floor that rendered it impossible for the model to accurately predict the UPT for smaller agencies. This problem was caused by a significant disparity in the ranges of the numerical values used to train the models. The values for UPT were often in the tens of millions, as were values like area population and capital expenditures, while values like fares and VOMS were in the tens or at most the thousands. This mismatch in orders of magnitude was what resulted in the prediction floor.

In order to address the problem, values with large numbers were transformed using the logarithm, so the UPT value was replaced with its logarithm. This transformation was also applied to population, VOMS, and total property damage, replacing zero values with one where necessary to avoid undefined values. The random forest and all subsequent models were then trained on this dataset. Random forest regression on the scaled data generates an R^2 of 0.86 but avoids the artificial prediction floor of the previous model, as shown in the figure below.



Closer examination of the plot reveals good performance for the mid-sized agencies but suggests a possible systematic underprediction for large agencies and overprediction for small agencies. With the model trained, we list the relative feature importances in the table below.

Feature	Importance Score
Mode VOMS	0.884647
Total Minor Injuries	0.036011
Total Injuries	0.030670

Primary UZA Population	0.022279
Fare	0.016620
Total Spending	0.004799
Total Fatalities	0.002813
Total Serious Injuries	0.002161
Property Damage	0.000000

Most significant is VOMS, followed by some safety metrics. While this analysis works well as a black-box predictive model, it is limited by the lack of statistical significance, for which we turn to linear regression. After eliminating statistically insignificant coefficients and small effects, we obtain the minimal model outlined below, using the fewest number of features to predict overall ridership.

GLS Regression Results

Dep. Variable:	UPT	R-squared:	0.845
Model:	GLS	Adj. R-squared:	0.845
Method:	Least Squares	F-statistic:	1121.
Date:	Mon, 18 Nov 2024	Prob (F-statistic):	9.56e-249
Time:	18:11:27	Log-Likelihood:	-782.49
No. Observations:	619	AIC:	1573.
Df Residuals:	615	BIC:	1591.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025 0.975]
const	7.9791	0.505	15.792	0.000	6.987 8.971
Primary UZA Population	0.2549	0.037	6.952	0.000	0.183 0.327
Mode VOMS	1.0300	0.024	43.030	0.000	0.983 1.077
Fare	-0.1403	0.012	-11.981	0.000	-0.163 -0.117

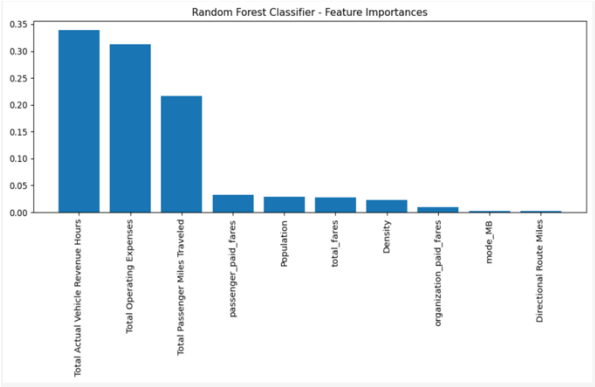
Omnibus:	38.770	Durbin-Watson:	0.795
Prob(Omnibus):	0.000	Jarque-Bera (JB):	45.521
Skew:	-0.595	Prob(JB):	1.30e-10
Kurtosis:	3.590	Cond. No.	228.

The regression model above shows us that the most important features are the area population, the VOMS, and the fare, with the fare having a negative effect on predicted ridership. We also see a large positive intercept term, which suggests that transit is generally popular and that there may be effects not captured by the NTD features selected that nonetheless contribute to rail ridership.

3.2 Bus

Road-based transportation systems, unlike rail, are not dependent on fixed guideway and other similarly expensive infrastructure, instead using existing roadways to move people around. Their success is highly dependent on how efficiently they operate. The NTD sorts of buses by subtype, but all road transit can be extracted from the data with relative ease by filtering for specific modes, including motor bus (MB) and paratransit bus (PB). As mentioned in the above section, the key ridership metric is Unlinked Passenger Trips (UPT). We expect bus ridership to be influenced by several factors, including operational capacity, population served, fares, and safety metrics.

To test this, we trained a **random forest classifier** on the data, which did quite well. "Total Actual Vehicle Revenue Hours" stands out as the most significant, which makes intuitive sense; the more time the buses are running, the more people are going to take them. “Total operating expenses” also contributed heavily. Factors such as the area population and the length of the routes were not as significant. The suggestion, based on the features, suggests that focusing on making transit more available, spending wisely, and ensuring it is accessible to people are important contributors to ridership.



Passenger Miles Traveled (PMT) was used as one of the features to classify ridership levels into categories: **Low, Medium, and High.** These categories were created from **Unlinked Passenger Trips (UPT)** using a method called quantile-based discretization. While **UPT** measures how often passengers are aboard a vehicle, **PMT** shows how far passengers travel.

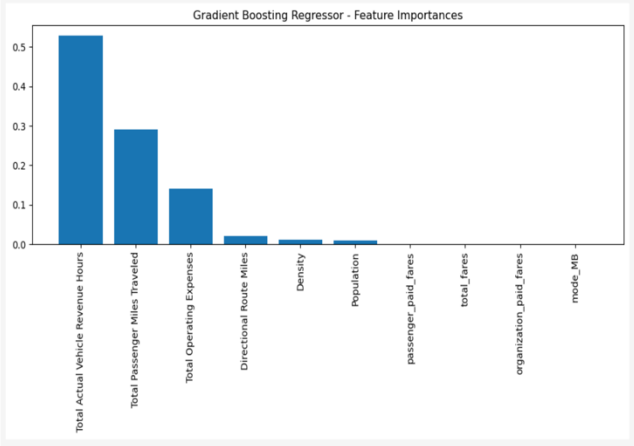
For Example:

1. Low UPT and Low PMT: A bus route with few passengers and short trips.
2. Low UPT and High PMT: A bus route with low passengers but those passengers are traveling long distances.

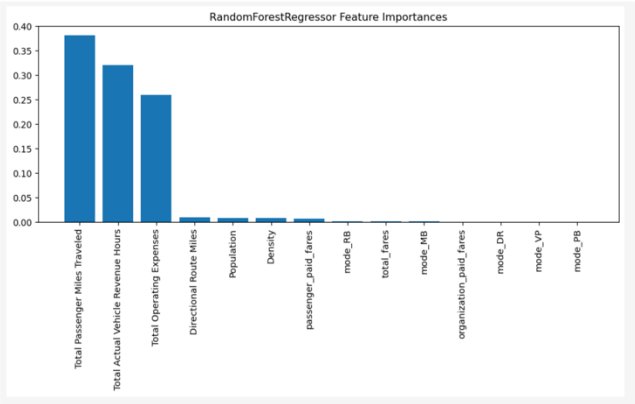
The random forest classifier stood out in our analysis, as it had accuracy of 94.86 accuracy and f1 score of 94.8%. The "Total Actual Vehicle Revenue Hours", basically the more hours buses and the transit systems in operation can drive ridership. "Operating Expenses" happened to be an important factor, as spending more on operations, the more people happened to use them. It makes sense hypothetically, because the developed a transit system is, it motivates people tousle it as their choice of commute. Features like population and Route miles didn't matter as much, it could be because having a bigger population and route miles doesn't intend to increase ridership.

Gradient Boosting was an obvious choice for this analysis because of its performance with non-linear patterns, which are common in real-world scenarios. The model delivered good

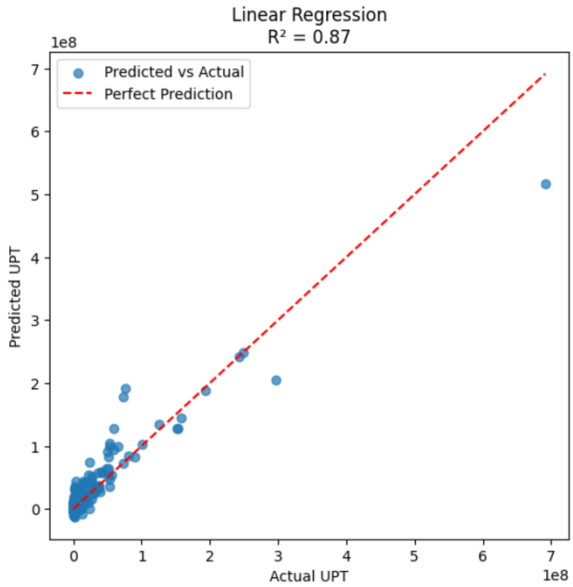
performance by achieving an **R^2 value of 0.99.** In simple terms this means it explained almost 100% of the variance in ridership. The **Mean Absolute Error (MAE)** was around **148,321 trips**. While that might sound like a lot, but looking at the scale of the dataset, it's actually really small. One of the Standout findings was the **Total Actual Vehicle Revenue Hours** as the most important predictor of ridership. This makes intuitive sense that the longer the buses are operational and accessible, the more the people are likely to use them. Features like **Directional Route Miles** and **Population** had a minimal impact, shows that expanding the routes or considering population size does not guarantee that it will boost the ridership.



In this analysis, the Random Forest model highlighted **Passenger Miles Traveled** and **Vehicle Revenue Hours** as the most significant predictors of ridership. These features make sense that the longer distances passengers travel and the more hours transit vehicles are available, the more likely people are going to use it and will rely on public transit.



Linear Regression was applied to predict **Unlinked Passenger Trips (UPT),** an important measure of transit ridership. This model was chosen to analyze how operational, demographic, and safety related features have contributed to ridership. The model achieved an **R^2 value of 0.87,** meaning it explained 87% of the variance in UPT. This certainly indicates that the selected features have a strong influence on ridership. Though we could see that there is still room for improvement.



The Predicted vs Actual UPT plot shows a strong linear alignment, for mid-range values. However, the model underperformed slightly at the extreme ends. Overestimating for higher ridership agencies and underestimating for low ridership, this shows that linear regression may not fully capture the non-linear relationship.

Feature	Importance Score
Mode VOMS	0.421195
Total Minor Injuries	0.409121
Total Injuries	0.135111
Total Spending	0.019043
Property Damage	0.007934
Primary UZA Population	0.005259
Total Fatalities	0.002337
Total Serious Injuries	0.000000

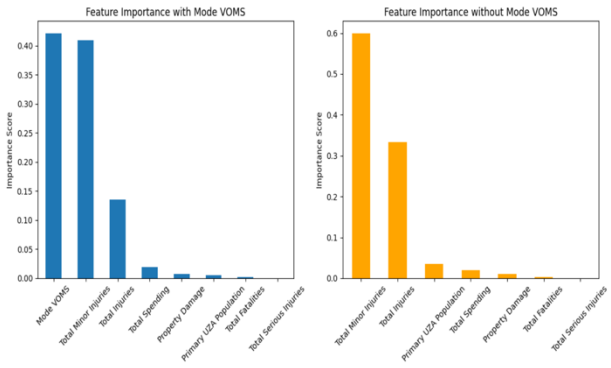
Mode **VOMS (Vehicle Operated Maximum Service)** emerged as the strongest predictor, explaining **42%** of the variance in UPT. This metric reflects a system's operational capacity—higher numbers of vehicles in service correlate with better coverage, convenience, and accessibility, directly boosting ridership. For example, a transit system with more buses running during peak hours offers shorter wait times and greater convenience, making it a more attractive option for commuters.

Significance of Safety Metrics:

Total Minor Injuries and Total Injuries together accounted for **54.4%** of the feature importance. These metrics are strongly tied to public perception of safety. A transit system perceived as unsafe is less likely to attract riders, regardless of its operational efficiency or affordability. Safety issues, such as frequent minor accidents or injuries, can deter passengers and lead to long-term reputational damage. **Total Spending** had a relatively small impact on ridership, suggesting that merely increasing financial resources does not guarantee better outcomes. The findings indicate that investments need to be strategically targeted—such as improving service frequency, safety, and reliability—to directly influence ridership.

Feature Importance:

Mode VOMS remained the most critical predictor, reaffirming its direct link to ridership. It highlighted the importance of operational efficiency in attracting passengers. When **Mode VOMS** was removed from the model, safety-related metrics—**Total Minor Injuries** and **Total Injuries**—became dominant, accounting for **93.2%** of feature importance. This suggests that in the absence of operational data, safety perceptions strongly influence ridership decisions.



4 Conclusions

Our analysis highlights various factors that influence rail and bus transit ridership, providing actionable insights for agencies.

For rail ridership, population size naturally correlates with higher usage, but the most influential factors were fare, and the number of vehicles operated. Lower fares and increased availability of vehicles tend to directly boost ridership. Interestingly, metrics like safety incidents were not significant predictors of ridership. However, this does not discount the potential impact of public perception on safety, which is not fully captured in the data. Additionally, the large value of the constant fitting parameter suggests two possible conclusions. First, that there are other effects not captured by the data in the NTD that could help explain more ridership trends, and second, that base ridership for rail systems is generally high independent of other factors. Some combination of these two could also be true as well.

This analysis underscores the importance of operational and safety metrics in predicting and boosting bus transit ridership. Investments in operational capacity (e.g., increasing Mode VOMS) and safety improvements should take precedence, while spending strategies need to be reevaluated for effectiveness. By adopting data-driven, targeted interventions, transit agencies can create systems that not only meet the needs of current users but also attract new riders, ultimately enhancing the sustainability of public transportation.

Finally, further research is needed to tease out effects related to more complicated or subtle effects, like distance-based pricing, the effects of station placement on ridership, or the effect of the overall system map on rider behavior. The necessity for public transit is not going away, and as more kinds of data become more available, more complex analyses are necessary to inform policymakers looking for the best way to bring effective public transportation to their communities.

4.1 Recommendations for Agencies

Based on our analysis and the conclusions above, we conclude the paper by collecting our set of recommendations for transit agencies looking to boost ridership. First, increase service availability and frequency, seeking to expand fleet capacity and optimize vehicle deployment during peak hours to improve service coverage, reduce wait times, and enhance convenience. Even outside of peak times, agencies should focus on making sure that service is as frequent as possible.

Second, agencies should foster a perception of safety on their systems. While the data suggests that transit agencies in the US are already quite safe, our analysis did find that buses are sensitive to safety numbers. Even though this is not the case for rail, that does not mean that passengers' subjective feelings about their safety do not affect their decisions to take transit. Agencies should make sure that their systems both are and feel safe to use.

Third, while spending does correlate with ridership, money should be directed towards improving the factors that have the most impact on riders. That means directing funds to expenditures that will increase frequency, improve safety, or reduce costs for riders.

Finally, agencies need to balance these investments against the necessity to keep their systems affordable for the public. Riders are price sensitive, which means that keeping fares low is important. Strategies like discounted fares during off-peak hours can help keep prices affordable while still raising enough money to keep the system running well.

SUPPLEMENTAL MATERIAL

A public-facing project site can be found at https://sites.google.com/view/transit-ridership-insights/home, and the project Git repository, containing full code and datasets, is accessible at https://github.com/Yash-Yashwant/transit-system-analysis.

REFERENCES

[1] Aboah, A. et al. 2021. Identifying the Factors that Influence Urban Public Transit Demand. arXiv.

[2] Answer in Progress 2023. why america is addicted to cars - YouTube.

[3] Chunyan Tang and Ying-En Ge 2017. Optimizing fare and operational strategies for an urban bus corridor using elastic demand. *2017 International Conference on Service Systems and Service Management* (Dalian, China, Jun. 2017), 1–5.

[4] Kim, G. and Rim, J. 2000. Seoul's Urban Transportation Policy and Rail Transit Plan—Present and Future. *Japan Railway & Transport Review*. 25, October (Oct. 2000).

[5] NotJustBikes 2023. I Visited the Best* City in North America - YouTube.

[6] pandas - Python Data Analysis Library: *https://pandas.pydata.org/#*. Accessed: 2024-12-06.

[7] Pedregosa, F. et al. 2012. Scikit-learn: Machine Learning in Python. (2012). DOI:https://doi.org/10.48550/ARXIV.1201.0490.

[8] Seabold, S. and Perktold, J. 2010. Statsmodels: Econometric and Statistical Modeling with Python. (Austin, Texas, 2010), 92–96.

[9] Taylor, B.D. and Fink, C.N.Y. 2013. Explaining transit ridership: What has the evidence shown? *Transportation Letters*. 5, 1 (Jan. 2013), 15–26. DOI:https://doi.org/10.1179/1942786712Z.0000000003.

[10] The National Transit Database (NTD) | FTA: 2024. *https://www.transit.dot.gov/ntd*. Accessed: 2024-12-06.

[11] 2024. Technological Innovation Drives The High-Quality Development Of The Shenzhen Metro. *Tunnels and Tunneling*.