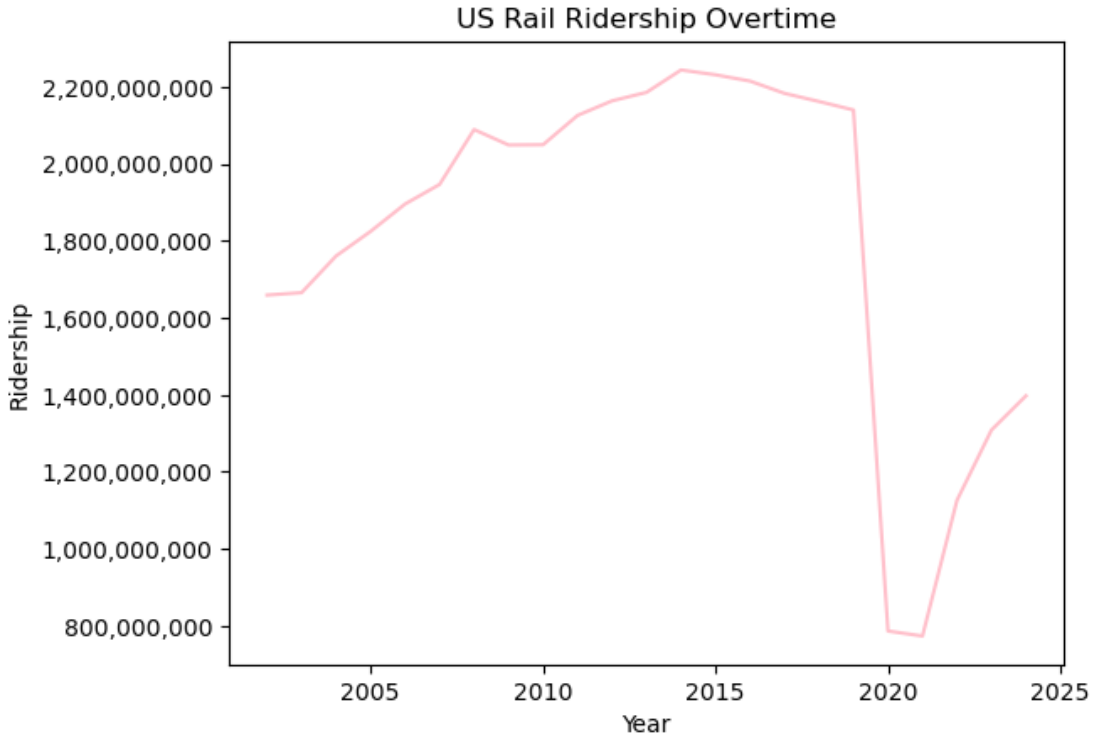
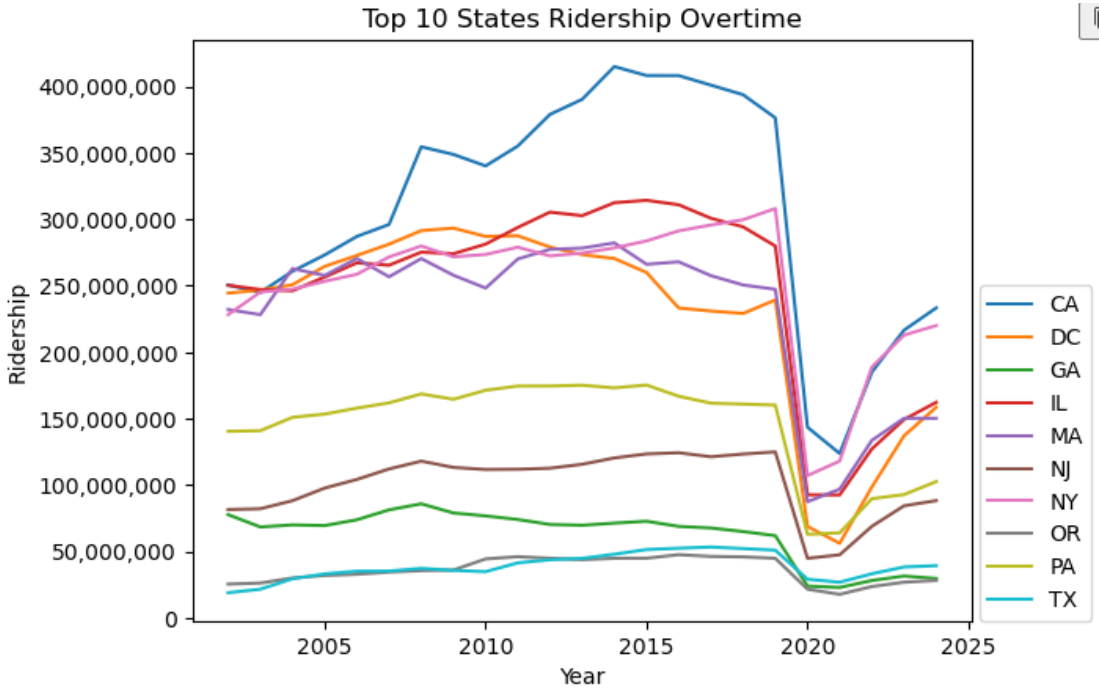
US Rail Analysis

Felix Menges & Jet Chanchom

How has ridership changed over time?

- Time Series plots of US ridership (UPT) & Top 10 States
  - Peak ridership around 2014-2015
  - Half of ridership since Covid
  - There was a decline before Covid
  - California having the highest ridership despite also being a high-unwalkable state.
- Questions & challenges
  - Forecast ridership
  - Why did it peak during 2015?
  - Why hasn't ridership returned to pre-Covid levels yet?





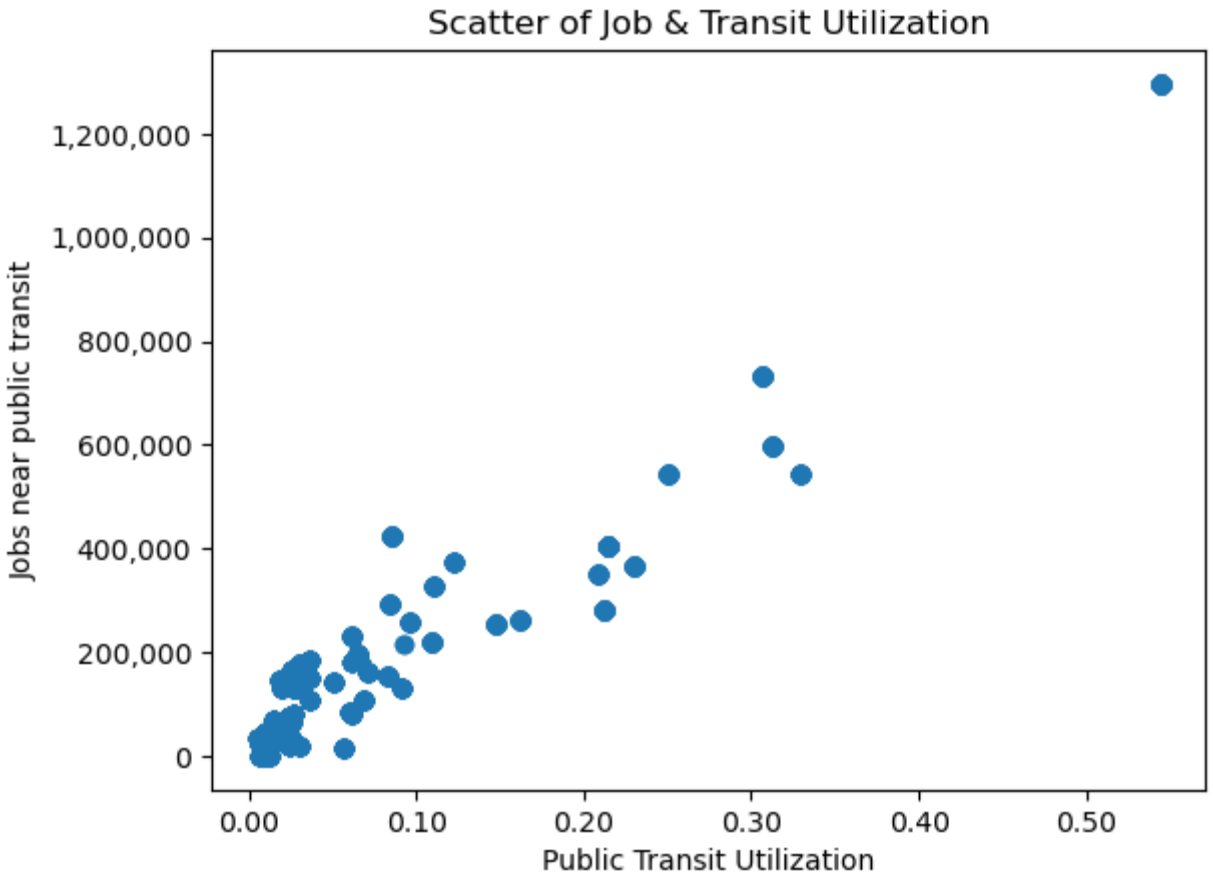
Agencies with the highest ridership

- Summary statistics via aggregation
  - The US' most densely populated cities expectedly has the highest ridership between 2002-Present
  - Californian cities occupies 5th and 8th despite being the state with the most ridership.
  - New York City has 3 agencies in the Top 10
- Questions/Challenges
  - Population density could be a factor of total ridership.

Agency	city	UPT
Massachusetts Bay Transportation Authority	Boston	5,302,250,865
Washington Metropolitan Area Transit Authority	Washington	5,249,483,547
Chicago Transit Authority	Chicago	4,299,462,519
Southeastern Pennsylvania Transportation Autho...	Philadelphia	3,164,852,196
San Francisco Bay Area Rapid Transit District	Oakland	2,315,514,732
MTA Long Island Rail Road	New York	2,143,060,547
New Jersey Transit Corporation	Newark	2,119,341,935
Los Angeles County Metropolitan Transportation...	Los Angeles	1,980,659,087
Metro-North Commuter Railroad Company, dba: MT...	New York	1,699,575,885
Port Authority Trans-Hudson Corporation	New York	1,656,219,248

How does access to jobs impact transit ridership?

- Scatter plot of jobs near public transit & transit utilization
- Access to jobs is near 1:1 proportional with the utilization of transit
- Questions/Challenges
  - Could delve deeper, unsure what metrics would be used, however.
  - Possibly look towards European or East Asian countries where rail transportation could less proportional (i.e. more jobs per %)



Which mode of rail transportation is most popular?

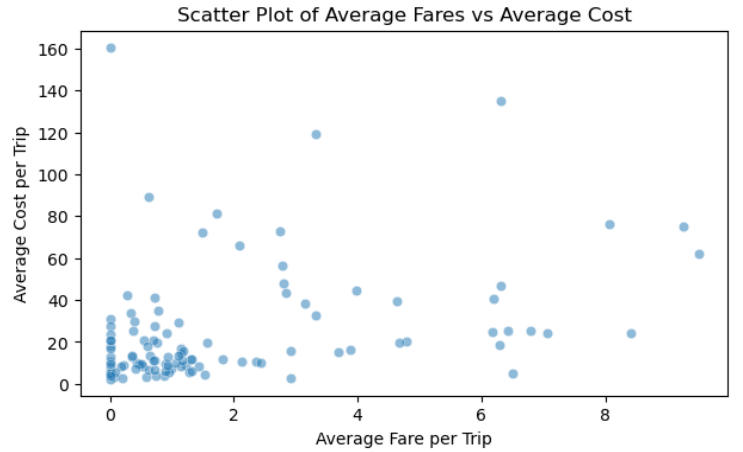
- Summary Statistics achieved via aggregation
  - Created City Class bins based of population size
- Light Rail & Commuter Rail are the most popular choices by far.
- Light Rail are popular in densely populated areas, save for the outlier of a Large Metropolis
- Questions & Challenges
  - Why are other forms of rail not as popular
  - Why is do large metropolises prefer commuter rail over light rail

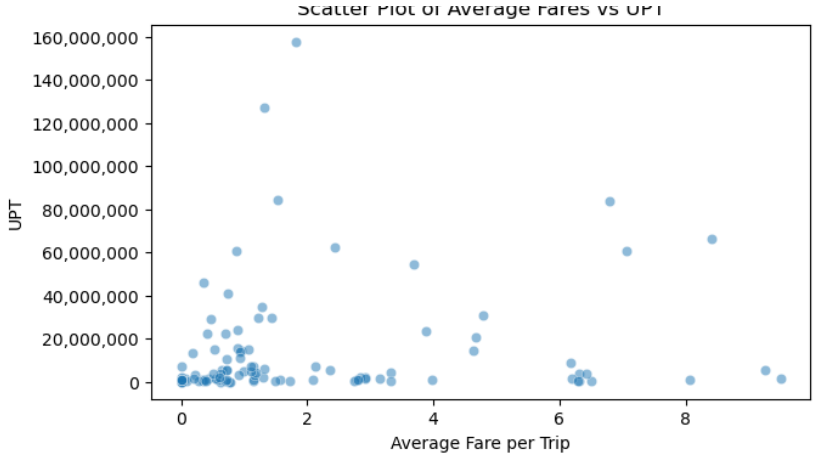
Mode	
LR	579
CR	551
HR	303
SR	254
MG	116

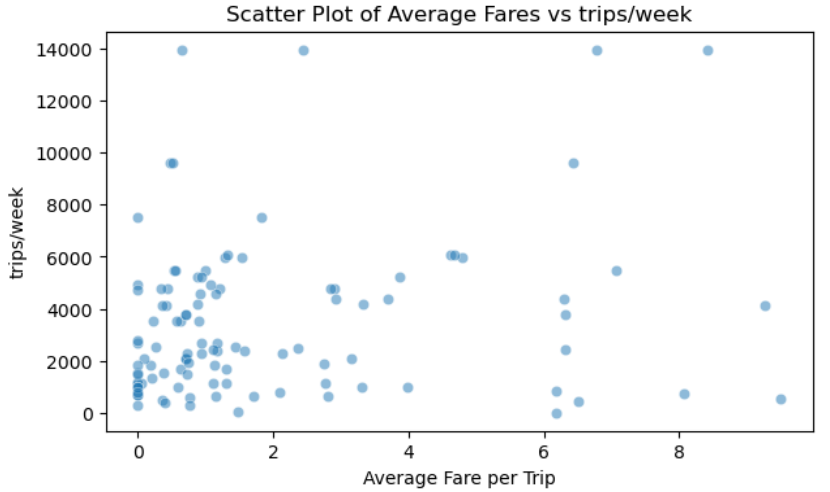
City Class	Mode	
Metropolis	LR	266
City	LR	201
Large Metropolis	CR	120
Small City	CR	100
Town	CR	99

How does fare price impact ridership? Does cost, population, and service area impact fare price?

- Filtered data to only include data from 2024 because fare data only accounts for the most recent year
- Average fare price per trip ranges from $0-$160.25, with a mean of $3.89 and median of $0.99. Most fares in our dataset are less than $3
- When comparing scatter plots of average fare per trip with UPT, trips/week, and average cost per trip, there doesn't seem to be a strong relationship
    - As fares goes up there is a trend of increased UPT, trips/week, and average cost, but most data points are clustered around the area where fares are less that $2
- Did hypothesis test using ANOVA for Avg_Fares_Per_Trip_FY & City_Class
    - Created categorical column for population using bins using City Class (Town, Small City, City, Metropolis, Large Metropolis) criteria from earlier question
    - P-value (.57) > alpha (.05), so fail to reject H0 – no significant difference in average fare price across different population sizes
- Questions/challenges:
    - Was Anova hypothesis test done correctly? When looking at mean fare price across city classes we get these price that seem significantly different across categories
        - Town: $3.88
        - Small City: $6.46
        - City: $7.41
        - Metropolis: $1.27
        - Large Metropolis: $2.85
    - What would be the best way to bin service area to create a categorical variable to use for ANOVA hypothesis testing?







Which factors can be used to predict how much ridership a city will have in the future?

- Created linear regression model to predict ridership
  - X: Mode, TOS, Organization Type, VRM, Avg Cost Per Trip FY, Avg Fares Per Trip FY, score, jobs, trips/week, routes, transit shed (mi2), %transit, & population
  - y: UPT
- Created dummy variables for our categorical variable (29 columns total for X)
- We got a R-squared score of .549 for our model, which means our model is not doing a good job of predicting UPT using our current variables

- Questions/challenges:
  - Model doesn't perform well, can experiment using different variables to see if we can get a better performing model.
  - Didn't check correlation of variables, which may have impacted model performance
  - What parameters can be used to improve the performance of our model?

```
R-squared score: 0.5486020212205438
Intercept: -2397072.293190535
Coefficients:
 [ 1.24780909e+00  1.99143956e+03 -6.56628822e+03 -8.53728731e+05
 -3.48344833e+01  3.40656520e+03  2.30200616e+02  5.14897939e+04
  8.08554135e+06  2.01547989e+00  3.27220736e+06 -5.39150715e+06
 -4.92130068e+06  7.94349349e+06 -1.96324237e+06  6.39895540e+06
 -1.01235462e+06  3.18819695e+05 -9.88906152e+05 -3.65616497e+06
 -7.28545463e+05  7.28545463e+05  1.92637035e+06  5.80491200e+06
 -4.69111233e+05  2.96287153e+06  2.63363947e+05 -1.61554671e+07
  5.66706049e+06]
```

Road Map for Finishing

- Finish answering our remaining analysis questions
- Do additional exploratory analysis for questions we've already answered
  - Try additional metrics (i.e., compare ridership by Urbanized Area (UZA) instead of city)
  - Improve linear regression model by changing parameters and variables
  - Create a forecast for total US rail ridership
- Clean up code and use markdown cells/comments to add documentation
- Begin final analysis and writing final report

Data Dictionary

Rank	Numeric	Ranking amongst other regions
Name	Text	Name of City & State
Score	Numeric	Overall Transit Score weighted by TCI, # of Jobs, and average Trips per Week. Higher score means better transit service Scaled [0.0:10.0]
Transit Connectivity Index (TCI)	Numeric	A normalized ranking of the sum of weekly bus & train traffic per region. Higher ranking means denser transit connectivity Scaled [0:100]
Jobs	Numeric	Jobs within 30-minute access of public transport
Trips/Week	Numeric	Transit Trips per Week within ½ Mile
Routes	Numeric	Total number of Transit Routes within ½ Mile
Transit Shed	Numeric	Size of geographic area accessible within 30 minutes by public transportation in square miles
%Transit	Numeric	% of commuters who use transit
Population	Numeric	Population of Region

Field	Type	Description
Agency	Text	Name of service provider agency
Mode	Text	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)
Type of Service (TOS)	Text	How services are provided: • Directly operated (DO) • Purchased transportation (PT)
Headquarters (HQ) City	Text	City of Agency's Headquarters
HQ State	Text	State of Agency's Headquarters
Urbanized Area (UZA) SQ Miles	Numeric	Urbanized Area (UZA): Region containing at least 50,000 people The size of the UZA
UZA Population	Numeric	The population of the UZA
Service Area Population	Numeric	Service Area: A ¾ mile radius surrounding a rail station The population within a service area
Service Area SQ Miles	Numeric	Total area covered by the service area
Unlinked Passenger Trips (UPT)	Numeric	Total number of times a person has boarded the railway
Avg Trip Length	Numeric	Average length of trip
Fares	Numeric	Total revenue made from fares
Operating Expenses	Numeric	Total expenses from operating the railway
Avg Cost per Trip	Numeric	Expenses divided by total number of trips
Avg Fares per Trip	Numeric	Total fares divided by total number of trips
Year	Numeric	Year of data collected
Vehicle Revenue Miles (VRM)	Numeric	Actual & scheduled miles during revenue service (excluding maintenance & training)