MGT 6203 Group Progress Report

TEAM # 104:

1. **Caroline Schmitt** (GTID: cschmitt33)
2. **Emy Ng** (GTID: ehang6)
3. **Matthew Kim** (GTID: mkim629)
4. **Mike Genovese** (GTID: mgenovese6)
5. **Osman Yardimci** (GTID: oyardimci3)

INTRODUCTION

For this project we are examining how gas prices affect public transit usage in the United States. The U.S. is broadly car-dependent, but as gas prices remain volatile, alternative methods of travel such as public transportation, carpooling, and bicycling are viable options to avoid spending more at the pump.

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. Furthermore, it may be interesting to explore these choices against a quantifiable measure such as air quality. Since not all forms of transportation are created equal (i.e., an automobile with a single rider will create much more pollution per capita than a subway), a statistically significant observation could be obtained by comparing all of these datasets.

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 provide a model for transportation organizations based on monthly US gas prices. To study this, we focus on the question: how does the average US gas price over time affect the ridership numbers of alternative transportation methods? While the relationship between gas price and ridership is our primary focus, we're also pursuing the following questions:

1. Which areas show the greatest increase in transit ridership in response to increased gas prices?
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?

This report outlines the progress we've made so far and what we've learned from our data. Once this research and analysis is complete, our findings may help alternative transportation organization or company model demand if gas prices are correlated to ridership. Organizations won't miss out on potential profits by having insufficient resources to meet demand and won't have a surplus of resources when ridership is lower.

DATA

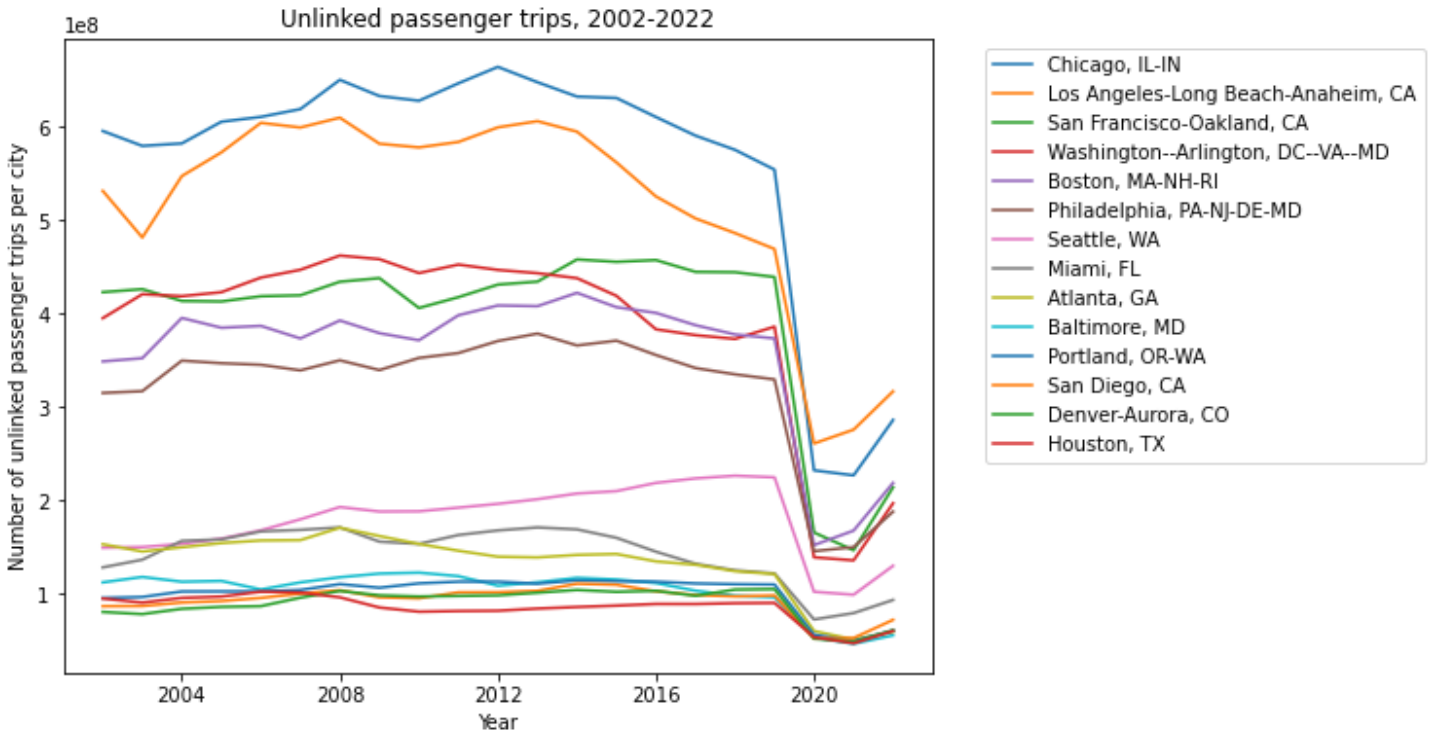
Complete monthly ridership dataset: Of the three primary sources our group is working with, the National Transit Database (NTD) monthly ridership dataset is the most complicated and messy. The four forms of counting in the ridership dataset are unlinked passenger trips (UPT), vehicle revenue miles (VRM), 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 presents 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. This makes it challenging to join the UPT data with the air quality index (AQI) data because the AQI data is encoded with state and city names rather than urbanized area designations. This makes joining datasets an ongoing challenge, though solvable via manual queries and troubleshooting.

For necessary transformations, we have successfully aggregated UPTs per UZA. 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.

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. When cleaning the data, the summary rows at the end of the Excel sheet were excluded. The data is available until April 2023. Because we only only have data for the first four months of the year, to get the top 15 cities, we plotted data only through 2022; the 2023 data collected so far cannot be compared to the annual sums in the rest of the dataset.

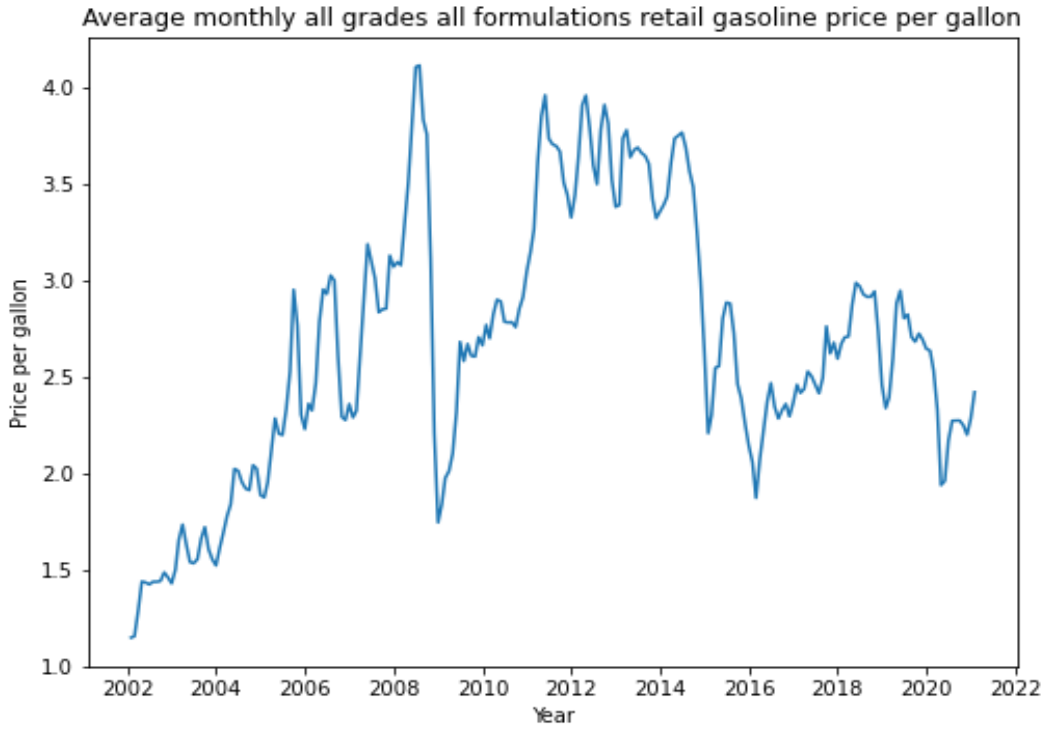


U.S. Gasoline and Diesel Retail Prices: Our original data source for gasoline and diesel rates was Kaggle. We have replaced it with the dataset from the U.S. Energy Information Administration, the source of the original Kaggle upload; the EIA source has been continuously updated and thus has nearly two years of additional data. We are additionally looking at 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 monthly price by the consumer price index, CPI.) As we continue modeling, we may use the U.S. Bureau of Labor Statistics CPI dataset for our own adjustments.

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. As we continue modeling, we'll be keeping a close eye on these dates for the relevant datasets. As for missing observations, there are gaps in coverage early in the dataset. However, starting in 2002, when the monthly ridership data begin, there are no missing values.

For EDA and preliminary modeling, we have performed time series resampling to work with monthly and annual data. Below is an exploratory line plot of average monthly gasoline prices, across all grades and formulations. This plot is not adjusted for inflation: it shows the nominal average price per gallon over time, from 2002 to 2022, for all grades all formulations retail gasoline.



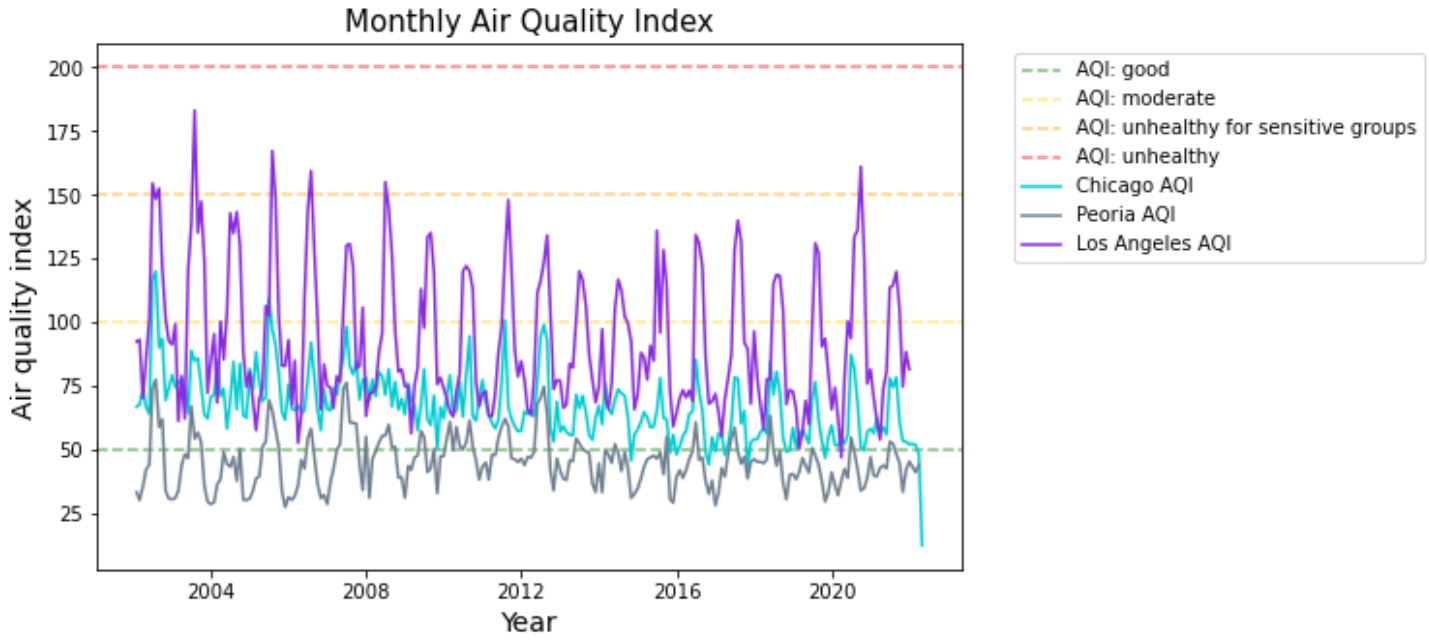
There are three major drawdowns in the graph:

- The 2008 financial crisis and the Great Recession.
- Decline in oil prices in 2014–2016 due to an oversupply of petroleum compared to demand.
- COVID-2019.

These events may have impacts on ridership unrelated to gasoline price.

Air quality 1980–present: The primary challenge presented by the air quality index (AQI) dataset is its size: there are well over half a million rows of data, for over fifty years and across 671 unique "counties or core based statistical areas" (CBSA.) Core based statistical areas are similar to urbanized areas, but the AQI dataset CBSA codes unfortunately do not match against the ridership dataset UZA codes. However, state and city are clearly given, making "manual" matches possible, as discussed above.

At over 5 million rows, the air quality index dataset is too large to upload to GitHub natively. To be able to analyze and load the data, the country-level data was chopped up into state-level files. The AQI dataset has a *state ID* column with no missing or null values, and there are 52 different state ID codes, one for each of the 50 states, Puerto Rico, and the District of Columbia. See below an exploratory plot of the air quality index (AQI) in Chicago, IL, Peoria, IL, and Los Angeles, CA.



Note substantial variation at the state level between Chicago (urban) and Peoria (rural), though similar seasonal or periodic patterns across all cities.

MODELING APPROACH/METHODOLOGY

We are currently most focused on modeling. We hope to work with a variety of multivariate models, including vector autoregressive (VAR) models; generalized autoregressive conditional heteroskedasticity (GARCH) models; decomposition models, and others. However, due to the nature of the datasets we are working with, we will start with simple multivariate time series regression, i.e. predicting future observations using lagged observations. Because transit usage and gas prices have a strong degree of autocorrelation, we expect that these simple preliminary models will be successful, though not necessarily useful for long-term forecasting or inference, which are our ultimate goals. However, starting with these more simple models will help us continue to better understand our datasets and the trends therein.

While working with more advanced models, we will look at different ways to optimize hyperparameters, including exploratory analysis like looking at ACF and PACF plots to figure out the VAR order and computational strategies like grid searching and auto-ARIMA.

For this project, inference is more important than prediction. When evaluating and comparing models, we'll be concerned with model accuracy, measured by, e.g., MAE, MAPE, and RMSE as evaluated on a holdout set, but goodness-of-fit measures will also be critical. For inference, model interpretability matters. Thus, we may prefer lower-performing but interpretable models with testable assumptions over higher-performing but black box-style models such as neural nets. In particular, decomposition models may allow us to pick apart the complicated trends in our data. Seasonality and cyclicality are observable in the line plots for transit use and air quality. Understanding these better will be useful for us and any transportation stakeholders.

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 a positive relationship between transit ridership and air quality will exist. We hypothesize this relationship will likely be marginal. Although public transportation is more efficient than automobiles per capita, the difference in ridership may not be enough to significantly offset emissions. 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, air quality, and transit ridership. In particular, we expect rises in transit ridership when gas prices are especially high. But gas prices, air quality, 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.

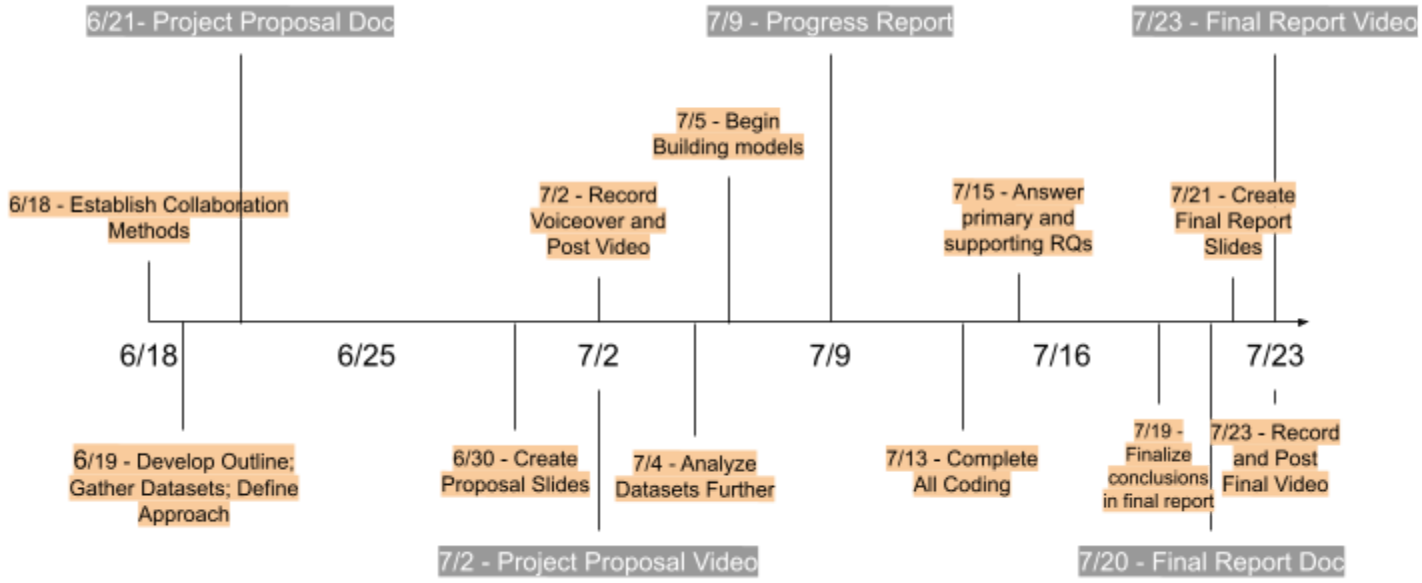
Ridership forecasting is useful for transit agencies as well as companies like Lyft or Uber that compete with public transit. Understanding regional ridership elasticity may also help regional and national policymakers increase public transit use.

Furthermore, air quality is an important measure for public health. Being able to more accurately forecast AQI in relation to readily available data could prove useful. Although benefits would be marginal for the individual, it could have a significant impact in aggregate.

PROJECT TIMELINE/PLANNING

In the chart below you can see a timeline of the project milestones as well as goals for tasks that we wish to complete in a given timeframe to stay on track and on schedule to complete the project by the 7/23 deadline.

Currently we have met all goals so far on our timeline to the current date. Currently, we are in the initial stages of developing our models to help answer our research questions but most work on this will be completed the week of 7/10 as we aim to complete our code by our next goal of 7/13 . We believe that we are on track for this goal but we did plan a decent buffer to the final report deadline in case we need to take some time away from writing our report to develop our code further.



LITERATURE SURVEY/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