Predicting US Public Transit Ridership

Daniel Pickett
March 2023

Section 1: Problem Statement

For decades, the state of public transit in the United States has been suboptimal. Transit lines are generally underfunded and underutilized. Both of these qualities act cyclically: lack of funding decreases the quality of the lines, which subsequently decreases people's desire to use them, and which justifies even more decreased funding. In most places across the US, cars and car infrastructure are the least inconvenient, and often the only feasible, method of transportation, despite the many negative long-term effects that massive car usage has on city design, local culture, and the environment. The recent COVID pandemic has only made this issue worse. When it began in early 2020, many lines either shut down temporarily or restricted service, and even for the ones that didn't, ridership decreased drastically as people emphasized distancing from others. Several years later, some lines have recovered their ridership levels, but many are still in progress or are struggling to recover.

In this project, we wish to determine how many of these lines truly have potential to recover their ridership to pre-COVID levels. Given monthly data on the ridership of local and regional public transit lines across the US between 2002 and 2022, can we predict the ridership of these lines for 2023 (and perhaps later)?

Section 2: Data Wrangling

The raw dataset contained 2,237 records, where each record was a different local or regional public transit line in the United States. For each record, the following fields were provided:

- 12 descriptive characteristics, such as where the line is located in the US, and other identifying information. Each line was a unique combination of three of these characteristics: Agency, the transit property in the area that owns the line; Mode, the type of transit of the line (bus, light rail, etc.); and Type of Service, or TOS, an indicator of whether the line is directly owned and operated by its agency, or by a third party through a contract.
- 14 summary statistics, numbers about the line either in general (such as service area square mileage) or covering the last fiscal year (such as fares collected). The majority of these did not factor into the final analysis, but they did serve as gauges on whether to even consider analyzing certain lines.
- 250 months of ridership numbers, from January 2002 to October 2022. These were more officially referred to as unlinked passenger trips, or UPT, defined as the number of passengers who boarded the transit vehicle, regardless of how many total vehicles they used to reach their final destination. For this analysis, it was assumed that the months before January 2020 are considered pre-COVID months, and the ones from January 2020 and on are considered post-COVID (or during COVID) months.

With such a large dataset, the main task before the final modeling would be to narrow it down to a few lines most worth spending time on.

As an immediate filter, one of the descriptive fields, Active Status, indicated that 683 lines were currently inactive. The timing of these lines' closures varied from within the last few years to near the beginning of the 20-year provided history. Since many of the lines were clearly closed for the indefinite future, and the dataset provided no indication of when any might ever reopen, all of these lines were removed from the data.

Of the remaining lines, 421 were missing some variety of their descriptive and summary values. Many of these lines were also missing a good portion of their UPT history, and without the descriptive and summary values, it would be particularly hard to extrapolate or assume anything about these lines during modeling. Since checking these lines for fixes on an individual basis would be an inefficient use of time, these lines were simply removed instead. Afterward, a further 169 lines were missing at least 10% of their UPT history. It seemed less meaningful to

predict results on what would have to be largely interpolated data, especially when accounting for variations in the UPT due to seasonality, so these were also removed.

Superficial checks on the remaining data revealed little wrong with it. All columns contained the appropriate type for their data, and there were no duplicate records. A smell test also showed that fields were generally providing values within reasonable ranges and not displaying any immediately suspicious trends.

The wrangled dataset contained 964 records and all original 276 fields of data.

Section 3: Exploratory Data Analysis

In this step, lines were eliminated from final consideration based not on having erroneous data but on whether their data had the optimal qualities for this analysis. First, each line was checked for the relative usefulness of the area it served. For each agency, therefore, was calculated the percentage of the last fiscal year's UPT that each of its modes comprised (where UPT over the last fiscal year was provided as one of the fields of summary statistics). This determined that most areas were serviced primarily by one or two major lines, regardless of how many total lines serviced the area. More specifically – and more importantly – 422 lines comprised less than 10% of their agency's UPT, suggesting that their usefulness in the area's public transportation system was highly limited. These records were removed from the data.

The modes of the lines were then more deeply considered. Intuitively, transit ridership is driven by demand – people will ride the line if they find it efficient and safe (COVID-wise, in this case) to do so. In other words, people will decide to ride the line if it is the best *option* to take them to their destination. Certain modes did not fit the spirit of this type of line:

- Modes that are the only way to reach a destination, such as a ferry to an island. These are not optional modes – ridership is driven not by how desirable the mode itself is, but by how much people want to be at that destination.
- Modes that are used primarily as tourist attractions, such as San Francisco's cable car system. Ridership is not driven by the mode's efficiency but rather its entertainment value. Arriving at a particular destination efficiently and safely is not the main purpose of using the line.
- Modes that don't behave as "traditional" mass transit, such as vanpools. These are essentially on-demand cars, akin to taxis, which don't alleviate any of the problems that "traditional" mass transit does.

43 lines were of a mode that fit one of these descriptions. These records were also removed from the data.

The remaining records were grouped by different characteristics, and the groups' UPTs graphed together in order to visually determine any surface-level similarities between lines. The first attempt was to group lines by similar UPT numbers over the last fiscal year (Figure 1). However, this made obvious that lines in the same group often had different seasonal patterns, making them difficult to extract any generalities from. Worse, since UPT over the last fiscal year expressed the effects of COVID, it could often drastically differ from how the line operated pre-COVID, resulting in both pre-COVID and post-COVID comparisons between different lines *and* among the same line having little in common.

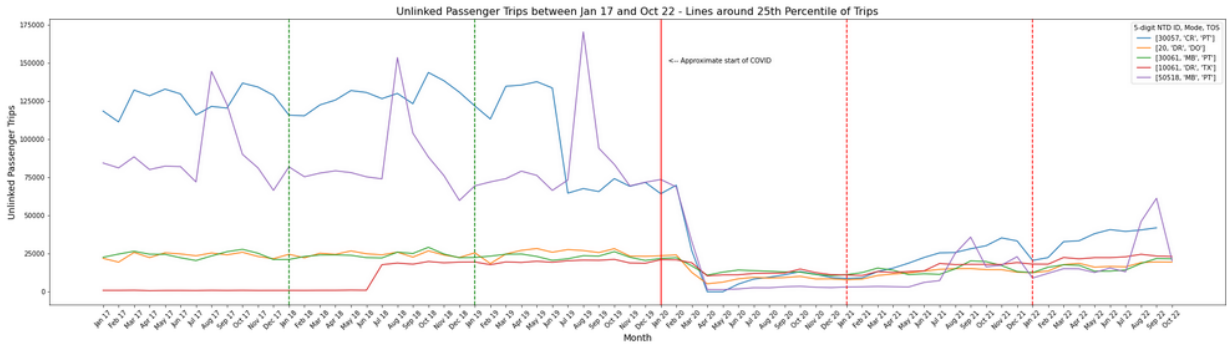

Figure 1: Graph of UPT data of 5 lines with similar UPT numbers over the last fiscal year – in this case, around the 25th percentile of UPT. These lines generally do not express similar pre-COVID or post-COVID patterns.

The second attempt was to group lines by their service area populations over the last fiscal year (another field provided as a summary statistic), defined as the number of people living within three-quarters of a mile of a stop within the line (Figure 2). Although flat UPT numbers for these lines tended to be more similar, this grouping still did not account for different seasonal patterns between the lines.

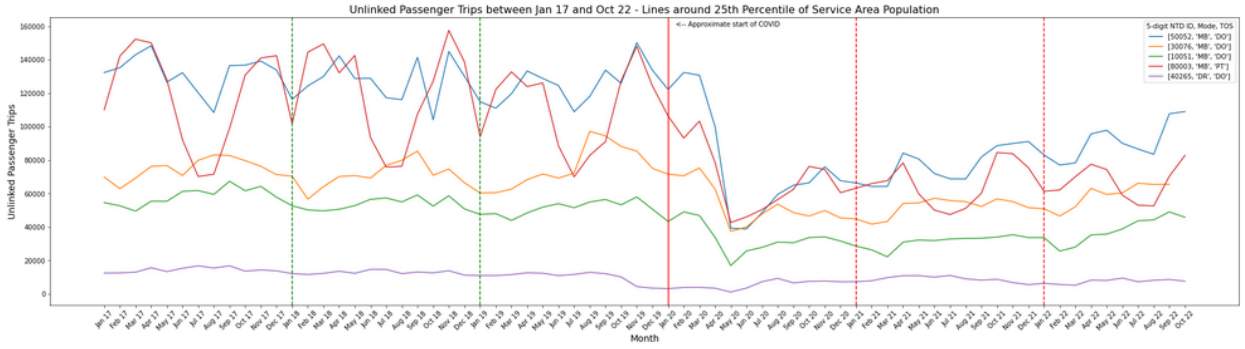

Figure 2: Graph of UPT data of 5 lines with similar service area populations over the last fiscal year – in this case, around the 25th percentile of population. These lines have more similar UPT numbers but still don't have similar seasonal patterns.

The third attempt was to group lines by both their mode and type of service (Figure 3). Additionally, lines' UPTs were standardized by their recent pre-COVID maximum amount (from between January 2017 and December 2019) in an attempt to display common seasonal patterns between lines more obviously. This was the best grouping of the lines yet – lines with similar modes indeed had similar seasonal patterns, with the standardization making this especially clear.

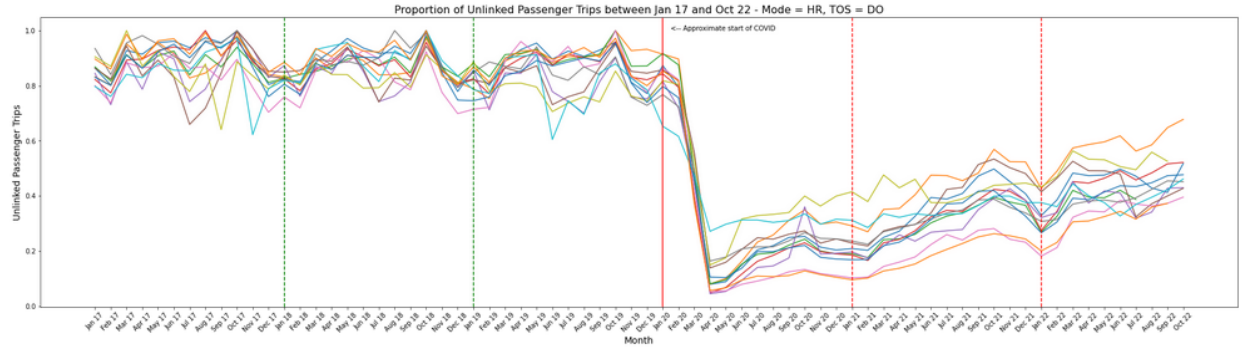

Figure 3: Graph of UPT data of all lines with the same mode and TOS – in this case, all 12 lines of the heavy rail mode and directly operated TOS. These lines have express similar seasonal patterns despite servicing different areas of the US – for example, their pre-COVID October UPT numbers all spiked upward, and their January 2022 UPT numbers each spiked downward.

Grouping the lines in this manner revealed the following general takeaways:

- The UPTs of lines with the same modes expressed similar seasonal patterns, even if the magnitudes of their seasonal changes differed – in other words, UPT rose and fell at the same times of year.
- Nearly all lines achieved their historical maximum UPT during their pre-COVID history – that is, UPT post-COVID usually did not ever surpass the maximum UPT achieved pre-COVID.
- At the beginning of COVID, between January 2020 and April 2020, all lines' UPT dropped drastically and to varying degrees, to between 0% and 40% of their pre-COVID maximum. In the following three years, their UPT slowly climbed back toward pre-COVID levels. Several lines managed to achieve or even surpass the pre-COVID maximum, but by October 2022, most lines achieved only between 25% and 75% of the maximum.

- Interestingly, many lines' post-COVID UPT patterns were not necessarily the same as their pre-COVID UPT patterns. This makes sense, as it is likely due to changes that the agencies were forced to make in their service schedules as a results of COVID, which may or may not have ever reverted.

After viewing the lines in this manner, it made sense to consider narrowing down the lines simply by how likely they appeared to be able to recover their UPT in the near future. After some experimenting, a threshold of historical success was established: the line must have had at least 12 post-COVID months that achieved at least 75% of the pre-COVID UPT maximum. Although this didn't account for seasonal patterns, such as how off-season months would almost certainly not meet the percentage threshold, this theoretically allowed for months near peak season, which are usually close in numbers, to help the line meet the duration threshold. This also was a fairly strict threshold that would eliminate all but the most likely successful lines, yet still allow for some diversity in line mode. Ultimately, 475 lines did not meet the threshold, and their records were removed from the data.

The explored dataset contained 24 records and all original 276 fields of data. Of these records, 13 were direct response lines, 9 were local bus lines, and 2 were commuter bus lines. These may be useful to refer to later to determine if they are at all correlated with the final predictions.

Section 4: Pre-Processing and Training Data Development

The data now needed to be arranged in a suitable manner for modeling. Because standardizing the data proved so useful for comparing different lines, this became permanently applied. The data was then split into training and test sets. Both sets needed to contain some post-COVID months in order to capture the most recent trends in UPT – however, there were only 2 years and 10 months available of this information. Therefore, the test sets consisted of 10 of these months, from January 2022 to October 2022; and the training sets consisted of the remaining 2 post-COVID years, from January 2020 to December 2021, plus the prior 3 years of pre-COVID data, from January 2017 to December 2019.

Section 5: Modeling

The datasets may have been cleaned and arranged for modeling, but one final check needed to be made on them as time series data. In order for them to return sensible time series forecasts, they also needed to express *stationarity*. Generally, this means that the statistical properties of the dataset do not change over time; specifically, this means that the dataset over time expresses three traits:

- Constant mean;
- Constant variance; and
- Insignificant autocorrelation. This is defined as the correlation between two datapoints in the data of a given fixed distance in time, or lag – for example, between all datapoints exactly two months apart from each other. For a dataset, there is therefore a different autocorrelation value for each lag.

There are two common signs of non-stationarity: trend, in which the data values increase or decrease generally throughout the entire time period; and seasonality, in which the data values increase and decrease in a cycle over a repeating unit of time during the period. Additionally, trend can also affect the dataset's trend itself, resulting in an exponential rather than linear change in value over time, or its seasonality, resulting in stronger fluctuations in value over later time periods.

To superficially visualize these signs, each of the 24 datasets was graphed individually. Each appeared to express some level of either trend or seasonality, either pre-COVID, post-COVID, or both (Figure 4 shows an example of one of these graphs). This was expected, as transit lines often have different levels of ridership in different months (seasonality) while generally increasing or decreasing in ridership due to local demand or population changes (trend).

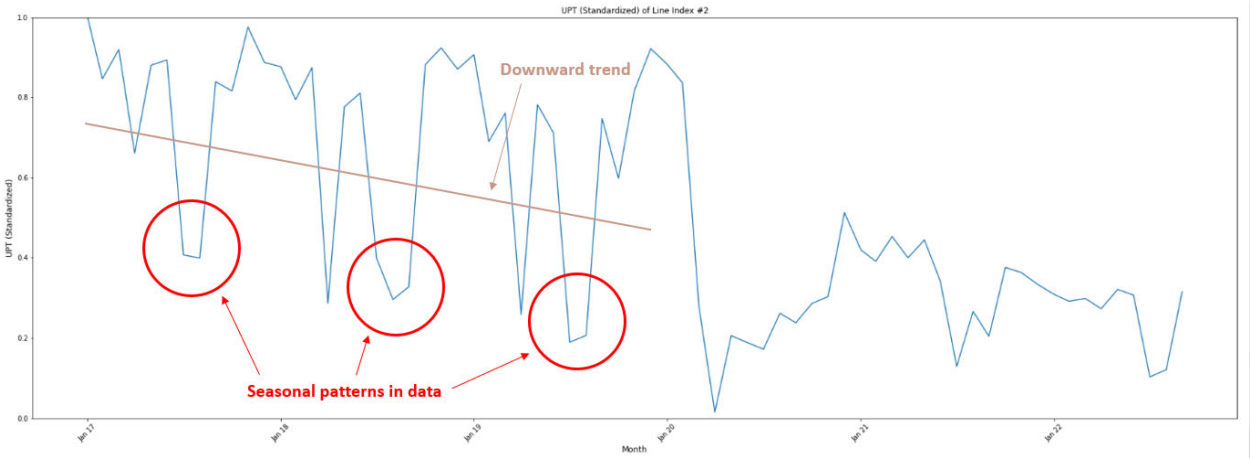

Figure 4: Graph of UPT data of one of the remaining 24 lines. The pre-COVID portion of the data expresses trend because it gradually decreases in value over each year. It also expresses seasonality because it has regular, similarly sized dips in value during early fall months (as well as early spring months).

Additionally, each dataset had its autocorrelation function (ACF) and partial autocorrelation function (PACF) graphed. Both the ACF and PACF are graphs of autocorrelation values at each lag from the current value in the data. However, the ACF accounts for how a given lag might affect later values that in turn affect the current value, while the PACF only accounts for how the given lag directly affects the current value. The former therefore indicates generally how far in the past that previous values have a significant effect on the current value, while the latter pinpoints individual months that have a significant effect. The ACF and PACF of each dataset showed significant autocorrelation between numerous lags (Figures 5 and 6 are examples of a dataset's ACF and PACF), lending further credence to how the datasets were currently non-stationary.

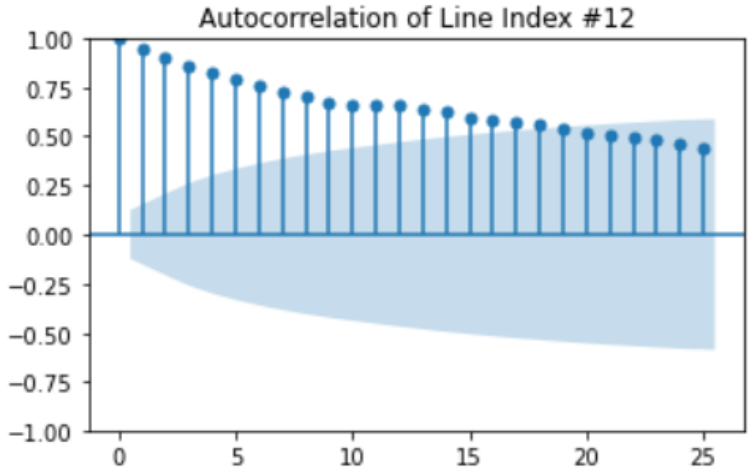

Figure 5: Graph of ACF of one of the lines. Values outside of the blue shaded region are considered significant. For this line, the graph indicates that the current value is significantly affected, directly or indirectly, by the last 18 months of lagged data.

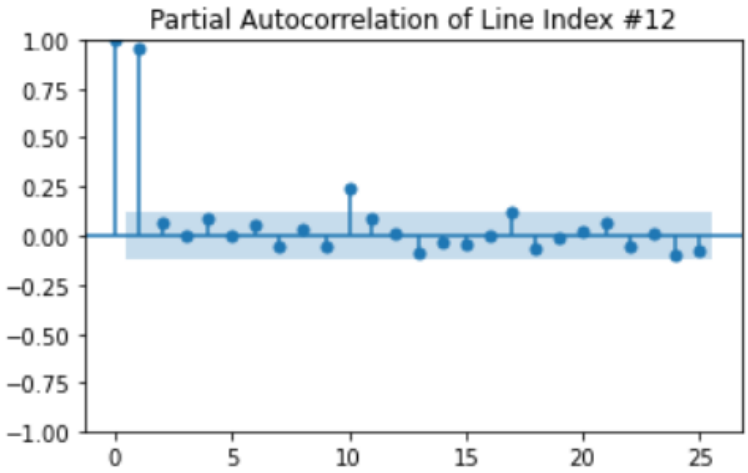

Figure 6: Graph of PACF of one of the lines. For this line, the graph indicates that the current value is directly significantly affected by the 1st, 10th, and perhaps the 17th months of lagged data.

The issue of non-stationarity is often addressed by transforming the data, uniformly shifting it in ways that remove any trends over time. Common transformations include differencing the data between a single unit of time (here, by a month), useful for adjusting linear trend; differencing the data between longer periods of time (such as by a year), useful for adjusting seasonality; and taking the logarithm of the data, useful for adjusting exponential trend. Which transformations are best for the dataset can be determined via the Augmented Dickey-Fuller (ADF) test, which (without delving too far into specifics) determines whether the overall formula for the current value of a dataset is significantly affected by previous values. A returned test statistic that is below a certain threshold (often 0.05) indicates an insignificant affect by the previous values, and that the dataset is likely stationary. Each dataset was transformed by each of the three individual transformations, as well as combinations of the three, and the ADF test subsequently applied. It was determined that stationarity could be achieved for all 24 lines by differencing by both month and year, and for 13 of the lines by additionally taking the logarithm of their data. These transformations were therefore permanently applied to the appropriate datasets. Figure 7 below shows the graphed dataset in Figure 4 after being transformed, which no longer displays its old annual trend or seasonality.

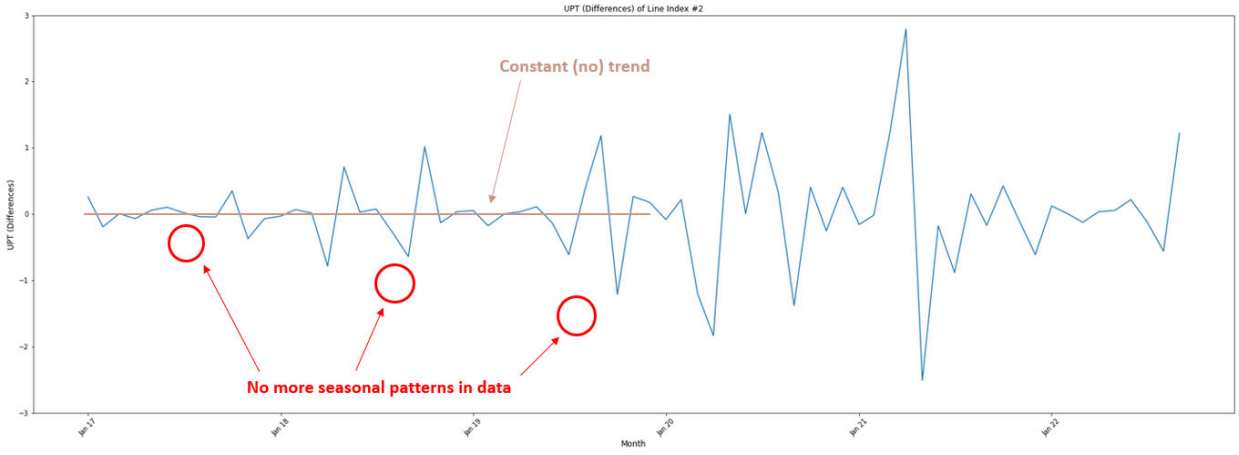

Figure 7: Graph of UPT data of the line graphed in Figure 4, post-transformation. The trend and seasonality that it previously expressed is now gone.

The appropriate time series models for the datasets could now be determined. The types and characteristics of the generally available models are as follows:

- Autoregressive, or AR(p). The first of the two basic models. In this model, the current value is a function of the *p* past values, plus some degree of white noise:

$$y_t = a_1 y_{t-1} + \cdots + a_{\mathrm{p}} y_{t-p} + \varepsilon_t$$

- Moving average, or MA(q). The second of the two basic models. In this model, the current value is a function of the *q* past error terms, plus some degree of (current) white noise:

$$y_t = m_1 \varepsilon_{t-1} + \cdots + m_{\mathrm{q}} \varepsilon_{t-q} + \varepsilon_t$$

- ARMA(p, q). This model is simply an additive combination of an AR(p) and an MA(q) model – the current value is a function of both past values and error terms:

$$y_t = a_1 y_{t-1} + \cdots + a_{\mathrm{p}} y_{t-p} + m_1 \varepsilon_{t-1} + \cdots + m_{\mathrm{q}} \varepsilon_{t-q} + \varepsilon_t$$

- SARIMA(p, d, q)(P, D, Q, s). This model is an extension of the ARMA(p, q) model, containing functionality to address, if present/required, seasonality (the "S") or differencing (the "I", short for "integrated"). It continues to be a function of the *p* past values and *q* past error terms, but is now also a function of the *P* past seasons of values and *Q* past seasons of error terms:

$$y_t = a_1 y_{t-1}^{[d-1]} + \cdots + a_{\mathrm{p}} y_{t-p}^{[d-1]} + m_1 \varepsilon_{t-1}^{[d-1]} + \cdots + m_{\mathrm{q}} \varepsilon_{t-q}^{[d-1]} + \varepsilon_t$$
$$+ n_1 y_{t-1}^{[D-1]} + \cdots + n_{\mathrm{sP}} y_{t-sP}^{[D-1]} + r_1 \varepsilon_{t-1}^{[D-1]} + \cdots + r_{\mathrm{sQ}} \varepsilon_{t-sQ}^{[D-1]}$$

- For all models:
  - t = current time;
  - y_{t-x} = data value x units of time in the past;
  - ε_{t-x} = error term x units of time in the past; and
  - p and q are known as the **orders** of the models.
- For the SARIMA model:
  - s = number of units of time that make up a season; and
  - $[d]$ = number of units of time that the data should be differenced by.

On each of the 24 datasets' training sets, models were fitted containing each combination of *p*, *q*, *P*, and *Q* from 0 to 6, and the best fit combination selected as the dataset's final model. Despite the vast number of combinations that needed to be tested as a result, a few simple lines of code were enough to run through each combination and determine the best for each model on its own; but as this code still required a significant amount of time to run, the coder left to have dinner in the meanwhile.

Once the final models were determined, few steps remained to achieve the final forecasts: the models predicted their respective lines' transformed ridership from November 2022 to December 2027 – approximately the next five years – and each dataset's transformations were backed out to obtain the final flat ridership numbers. It was immediately noticeable that most predictions for years after 2023 appeared to mimic exactly whatever seasonality was predicted for 2023, and their trend increased or decreased identically through all five years, leading to nonsensical results in later years. This is likely attributable to how each line's most recent data, which is post-COVID, both differed heavily from its pre-COVID data and had many fewer months available than the pre-COVID data, which limited complexity in the patterns of the predictions. The results were therefore truncated after December 2023, leaving approximately the next one year.

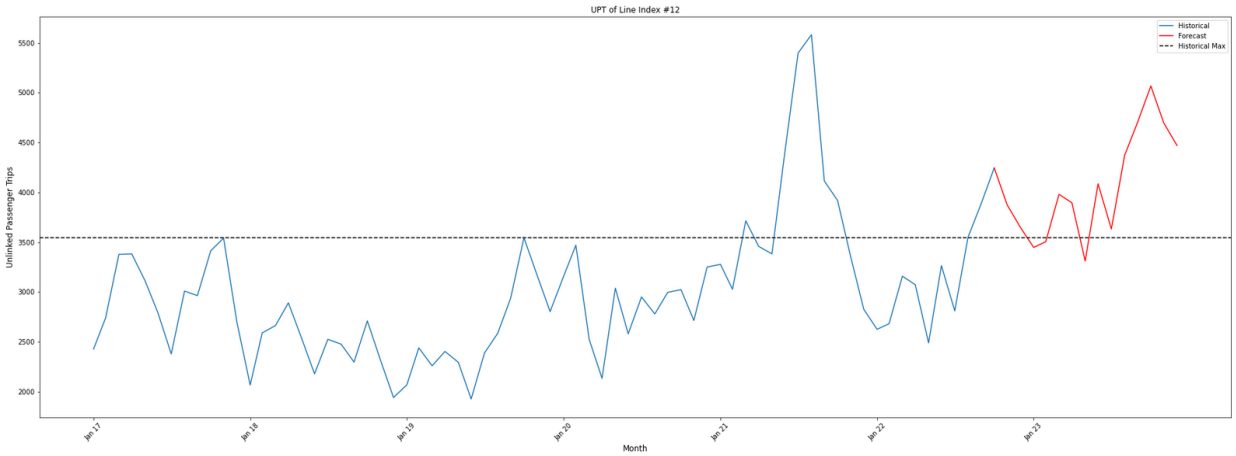

Figure 8: Graph of forecasted UPT data from November 2022 to December 2023 of one of the lines. Blue represents the historical values, red represents the forecasted values, and the black dotted line indicates the pre-COVID maximum UPT value for reference.

The breakdown of the results was as follows:

- 12 lines would achieve their maximum pre-COVID ridership during at least one month by the end of 2023, and were projected to be on an upward trend.
- 3 lines would achieve their maximum pre-COVID ridership during at least one month by the end of 2023, but were projected to be on a downward trend and therefore unable to maintain it.
- 2 lines would not achieve their maximum pre-COVID ridership during at least one month by the end of 2023, but based on their upward trend could potentially achieve it within the next 2-3 years.
- The remaining 5 lines would not achieve their maximum pre-COVID ridership during at least one month by the end of 2023, and based on their downward trend were not likely to achieve it within the foreseeable future.

Unfortunately, there was no discernable correlation between a line's success and its mode. There was also no immediate pattern between a line's TOS, service area population, or any other provided summary or descriptive value and its success or trend.

Section 6: Conclusion

Out of the 24 regional transit lines in the US that were analyzed for post-COVID ridership, 14 were projected to largely recover their ridership to pre-COVID levels within the next few years. At a 58.3% success rate, this is a potentially concerning prediction for the fate of US public transport, as the remaining 41.7% of lines could potentially face further budget or service cuts, further worsening its state. Additionally, these 24 lines were ones predetermined likely to succeed – the rate of success for most lines is therefore likely even lower.

Notes for future research:

- Some exogenous variables not included in the datasets could be useful to integrate into the analysis so that predictions beyond just the next year might make more sense. For example, incorporating trends in population growth of the area could allow for projected trends in ridership that fits them, instead of simply projecting that ridership increases linearly for all time.
- With more time, analysis could perhaps determine whether any summary or descriptive values are correlated with the predictions.
- Choosing and analyzing even only 24 of the lines out of the original 2,237 was a massive undertaking for this type of project, especially with regards to data cleanup and stationarity testing. It may make more sense to spend more time figuring out the single most representative line to perform this analysis on instead.