1. For our final project, the app that Jasmine and I chose to develop is a Delay Estimator, or a delay prediction model that forecasts rail delay on the New Jersey transit system. New Jersey rail serves as commuter rail between Philadelphia, New Jersey, and New York.

2.Amtrak is also commuter rail for the Northeast, so we originally planned to develop a website add-on for people buying their train tickets in advance. However, the Amtrak data does not include schedule times, which left us only with NJ Transit data. New Jersey Transit also has a different system than Amtrak: it only sells tickets on its mobile app, or in person.

3. Additionally, its ticket prices do not fluctuate like Amtrak tickets, which vary based on date of reservation. Because of this, predicting delay far in advance is not as important for our project. So, we decided to conceptualize a tool for the New Jersey Transit App, and in particular, its Trip Planner feature. This feature allows riders to look at future trips for a specific date and time.

4. In the resulting "Trip Options" screen, we aim to add an indicator for possible delay for each schedule option available, that refers to the total predicted time delay for each trip. Yellow indicates 10-19 minutes combined predicted delay, orange for 20-29 minutes, and red for over 30 minutes. There is no indicator for under 10 minutes, as we consider that "acceptable."

5. In the Detailed Trip View, we would also add the same icons and categorization color scheme for each leg of a trip. For an entire trip, these would all be summed up and displayed accordingly.

6. We are mainly using a dataset from Kaggle, which scrapes its data from the NJ Transit DepartureVision Real Time Train Status service. Our sample months are January and Feburary 2021. The dataset is made up of basic identification, spatial, and temporal data. We have also added weather data through the RIEM R package.

7. We first needed to actually determine if delay is severe enough often enough to warrant developing a model. As you can see, most trains in our sample set are delayed under ten minutes, with frequency reaching very low levels for delay over 40 minutes, and even less for over an hour.

8. To further examine delay, we used two metrics from the UK Southern Railroad; the first, "on time," considers the percentage of trains that arrive at each stop on schedule. For New Jersey Transit, the on-time rate of trains is usually low, with most trains only reaching each stop at its scheduled time 10-20% of the time.

9. The second indicator is the "Public Performance Measure," which refers to the percentage of trains that arrive at their stops within 5 minutes. Even using different brackets for PPM, it is clear that PPM varies spatially (and is especially low on the Philly-Atlantic City line). So, we concluded that there are sufficient delays within the system to warrant modeling.

10. For our models, we train on the first five weeks and test on the last three. So far, we have developed four different models with increasing levels of detail: time, space, weather, and time lag data. From testing these predictions, we can see that our most detailed model predicts general trends and small delays well, but misses periods of long delay.

11. It is clear that Mean Absolute Error decreases with the models that have additional independent variables. A couple of our ideas on improving accuracy are to expand our sample size to allow us to train on more data, and to improve on the current model by adding more variables or finding other relevant independent data such as weather.

12. Our current idea for the Delay Estimator is fairly simple, but has great potential to be expanded on. With more data, especially data on Amtrak and SEPTA, as they both are used as connectors or replacements for NJ transit, this estimator could be used for the benefit of more customers. With enough planning power, agencies could use this tool to predict and mitigate delays before they happen.