Li Guan:

In this project, my main task is to collect data from the real world. Our project is airline delay prediction and I need to find the airline schedule from the airports API. The API I choose is Airlabs API, the tool I query the API is python with the request package.

What I learned through this project is to make trade-offs. Since I cannot find all the data we need for a machine learning model to predict, I have to narrow down the amount of information that is needed by our team and find an optimized solution to finish the project.

Hengchuan: I learned a lot from deploying my data visualization tools to the AWS and also learned a lot from using the matplotlib library. The data visualizations and analysis part is interesting. I learned to use advanced data statistic tools to manipulate my data such as normal distribution.

Robin Sah:

In our project, I undertook a meticulous data processing journey, initially dissecting the dataset to understand the various data types and structures within the 5.8 million records, which spanned multiple features pertinent to flight logistics and delays. Addressing missing values was paramount, necessitating strategic imputation techniques to preserve our model's integrity. Correlation analysis was pivotal, aiding in the distillation of the most impactful features for predicting.

My classification model analysis encompassed a suite of five distinct machine learning algorithms, each rigorously trained on select predictive variables to forecast arrival delays. The Random Forest model exhibited superior performance, capturing 95.27% of variance in delays, signaling its robustness. While the Neural Network, with its nuanced handling of nonlinear relationships, incurred slightly higher error rates, its potential was evident, albeit with concerns of overfitting.

The real-world applicability of our models was tested through integration with the AIRLAB API, which presented a reduced feature set for real-time prediction, leading to diminished accuracy. Despite this, my focus remained steadfast on refining our models to leverage real-time data, ensuring they were fed with well-structured and relevant inputs. The transformation of intricate timestamp data and categorical strings was a notable part of this endeavor, highlighting the agility of our processing pipeline.