**Capstone Project 1 Data Wrangling Report**

**Introduction**

The purpose of this project is to analyze data and develop a model to predict to best flight optimization by various methods. This project includes modeling procedures such as training and testing, cross-validation, linear and polynomial regression. This project focuses on real airport transportation problems including a number of flights in short time slots (Ex. every 15 min) that causes frequent delays and possible conflicts, route and taxi time and more.

In fact, a reliable and efficient air transportation system provides substantial benefits to society and the companies in terms of global, national finance, and making air transportation convenient for people.

**Data Collection and Cleaning**

The data was obtained from Kaggle and it is consist of more than 320 airports and about 6M rows that contains all flights details took place in the USA in 2015 from Jan/1/2015 00:00 am to Dec/31/2015 11:59 pm.

Additional data sets are used to build more reliable prediction or optimization models such as the USA holidays that were expected to change the number of flights remarkably, and weather data of the USA for each airport which was obtained from DarkskyAPI, also expected to have interesting results. Unnecessary columns are dropped and statistical analysis is run to become familiar with the dataset.

The first part of the data wrangling involved managing an uploading large data to jupyter notebook. Some columns had to be uploaded in string format (astype(str)) to prevent loss of data in the progress. This also saved time for extra computations for converting the integer to date.time object.

Second, since this data set is consist of about 6M rows, 10 most common airports are extracted from the datasets to be used as the original data set for this capstone 1 project. This effort reduced the size of the data set to about 2M rows.

Third, holiday data is added to the dataset in binary format. The holiday is labeled as one and non-holiday is labeled as zero.

In addition, weather data is obtained from Darksky using API response method. Darksky has provided hourly weather data for each airport. Darksky allows Latitude, Longitude, and timestamp as parameters to get the desired weather data for the whole year. This process was repeated for each ten most common airports.

Finally, in order to calculate the frequent flight for certain times, a number of flights in 15 minutes were analyzed and inserted into dataset. This will help to design and test optimization and prediction models.

**Data Sources**

Holiday Data Source: <https://www.opm.gov/policy-data-oversight/snow-dismissal-procedures/federal-holidays/#url=2015>

Flight Data Source: <https://www.kaggle.com/usdot/flight-delays#flights.csv>

Airport Data Source: <https://www.kaggle.com/usdot/flight-delays#airports.csv>

Weather Data Source: <https://darksky.net/dev/docs#time-machine-request>