**Technical Report**

**ETL Project: Flight Delays and Cancellations**

Group members: Wei Zhu , Irene Okada

**Data Source:**

Our data focus was on flight delays and cancellations. Specifically, we were interested in the potential to project causes of flight delays and to predict likelihood of delay based on calendar, location, airline etc. We sourced data from Kaggle.com primarily in csv form. Although multiple years of data were available, we focused on the year 2018, the most recent full set of data. In addition, we extracted data regarding airports including codes and location. To close some gaps in data we extracted text data from a government aviation website.

**Sources included:**

<https://www.aviationweather.gov/docs/metar/stations.tx>

t<https://www.kaggle.com/gauravmehta13/airline-flight-delays?select=flights.csv>

<https://www.kaggle.com/sherrytp/airline-delay-analysis>

https://en.wikipedia.org/wiki/List\_of\_airline\_codes

* Airports: includes the IATA airport code, full name, address, long. and lat. of airports
* Flight data: includes flight delay information with delay time, airline short code and date details
* Cancellation data: includes the code and cause of cancellation, original origin/destination, potential flight length, month/day/year of flight among other potentially useful information

Workflow was as follows:

### **Extract/Data Utilized**

Data files (csv and txt) were read into Jupyter notebooks using pandas.

**Transformation**:

The files were cleaned, and fields prioritized based on hypothetical questions we would want to answer. Some columns were dropped, some renamed, some split and the datatype was changed on some to facilitate ultimate creation of a database in SQL. Transforming the data accommodated ease of reading and manipulation of the files and allows us to save space by creating multiple tables that can be referenced or joined depending on the need and question to be answered.

**Loading:**

We then created a connection between Jupyter Notebooks and Mongo dB to export Data Frames and create a database with multiple collections, ready for use in queries and analysis.

**Considerations:** Due to limitations on repository size it was necessary to use a sample of data from larger file