IST 652 Scripting for Data Analysis Final Project Report

by

Luigi Penaloza, Alan Nguyen, Kartheek Sunkara

Professor Ben Nichols

Data Acquisition

Originally, we were using a combination of Expedia unstructured data obtained via web scraping in JSON format, and csv data from Expedia. This proved difficult given the Expedia API from which we intended to obtain the data was hard to scrape, and only available to vendors doing business with Expedia. The project was slightly modified to use a combination of three different csv files containing airline, airport and flight information.

The structured data was compiled by the Department of Transportation Bureau of Transportation Statistics, which tracks performance of domestic flights operated by large air carriers. Summary of delayed, on time and cancelled flights is included in the report and data used for the project.

The acquired data from Kaggle for the year 2015 will serve to predict which airline is best to travel with, as initially intended. We will be using the data to advise on the best domestic flights and airlines to use when traveling via air.

Example of data used:

Airports csv

	IATA_CODE	AIRPORT	CITY	STATE	COUNTRY	LATITUDE	LONGITUDE
0	ABE	Lehigh Valley International Airport	Allentown	PA	USA	40.65236	-75.44040
1	ABI	Abilene Regional Airport	Abilene	TX	USA	32.41132	-99.68190
2	ABQ	Albuquerque International Sunport	Albuquerque	NM	USA	35.04022	-106.60919
3	ABR	Aberdeen Regional Airport	Aberdeen	SD	USA	45.44906	-98.42183
4	ABY	Southwest Georgia Regional Airport	Albany	GA	USA	31.53552	-84.19447

Airlines csv

	IATA_CODE	AIRLINE
0	UA	United Air Lines Inc.
1	AA	American Airlines Inc.
2	US	US Airways Inc.
3	F9	Frontier Airlines Inc.
4	B6	JetBlue Airways

Flights csv

	YEAR	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	TAIL_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPOR
1	2015	1	1	4	AA	2336	N3KUAA	LAX	PBI
2	2015	1	1	4	US	840	N171US	SFO	CLT
3	2015	1	1	4	AA	258	N3HYAA	LAX	MIA
4	2015	1	1	4	AS	135	N527AS	SEA	ANC
5	2015	1	1	4	DL	806	N3730B	SFO	MSP

The flights csv contains 33 columns related for flights information. Using properties such as flight delation and cancellation we will help people choose the best airline domestically in the U.S.

We will be using SQL alchemy, scikit-learn, seaborn, pandas and as our primary Python libraries to predict the best airline to use. SQL Alchemy will be used to query the data that we will later use. The Sickit library will be used to run our algorithm models for the prediction analysis. Seaborn is the library used for visualization purposes on our project.

Data Preprocessing

After looking at the data, the main thing that we needed to take care of was formatting the time and dates data. We also noticed a significant amount of NA values, however these were kept as we created three different tables for each csv and moved them into a MariaDB. This made NA's non consequential for our analysis. Ultimately the transfer of data into the MariaDB allows us to run our models for analysis with greater efficiency due to the ease of handling larger amounts of data.

Dates and Time:

First we concatenated date columns and gave day of the week. Then we changed the times in the dataframe so they are formatted as HH:MM, 23:54 instead of 235. We also abbreviated the airlines names and gave numeric values to them for use in support vector machines model.

When transferring our tables into the MariaDB we noticed that we had some binary values with 0 being equal to No and 1 to Yes. This was not plausible for machine learning analysis, and we had to match category names with values in a dictionary instead.

Methods of Analysis [2 point] This includes: o The questions that are to be answered o Which fields in the data are used o How the results of the analysis are collated •

The main question to be answered by our analysis is:

1) What is the best airline to take for domestic flights based on cancellations and delays?

The fields or variables in the data that were most useful from the collation of the csv files to answer this question is the days of week cancellations and delays, originally from the airlines csv.

For better use of the data, we chunked or collated the data as shown below:

```
0    00:05:00
1    00:10:00
2    00:20:00
3    00:20:00
4    00:25:00
Name: SCHEDULED_DEPARTURE, dtype: object
0    23:54:00
1    00:02:00
2    00:18:00
3    00:15:00
4    00:24:00
Name: DEPARTURE_TIME, dtype: object
0    00:15:00
1    00:14:00
2    00:34:00
3    00:30:00
4    00:35:00
Name: WHEELS_OFF, dtype: object
0    04:04:00
1    07:37:00
2    08:00:00
3    07:48:00
4    02:54:00
Name: WHEELS_ON, dtype: object
0    04:30:00
1    07:50:00
2    08:06:00
3    08:05:00
4    03:20:00
Name: SCHEDULED_ARRIVAL, dtype: object
```

Chunking simply refers to grouping the data into chunks, which are easier to commit to memory than a longer uninterrupted string of data.

For storage purposes as briefly stated before, we created a table to store flight data in a MariaDB with three different tables based on the three different csv files. Below is an example of queried data into a data frame:

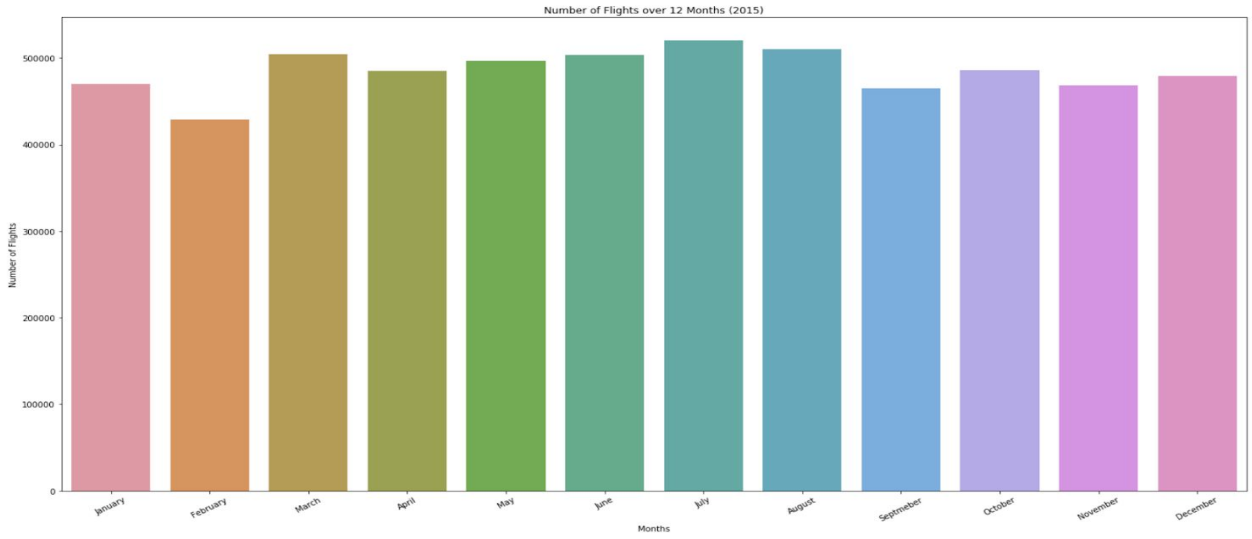
```
Successfully read from database
```

	DATE	count(*)
0	2015-01-01	469968
1	2015-02-01	429191
2	2015-03-01	504312
3	2015-04-01	485151
4	2015-05-01	496993
5	2015-06-01	503897
6	2015-07-01	520718
7	2015-08-01	510536
8	2015-09-01	464946
9	2015-10-01	486165
10	2015-11-01	467972
11	2015-12-01	479230

To finalize our analysis we ran machine learning models based on queried data.

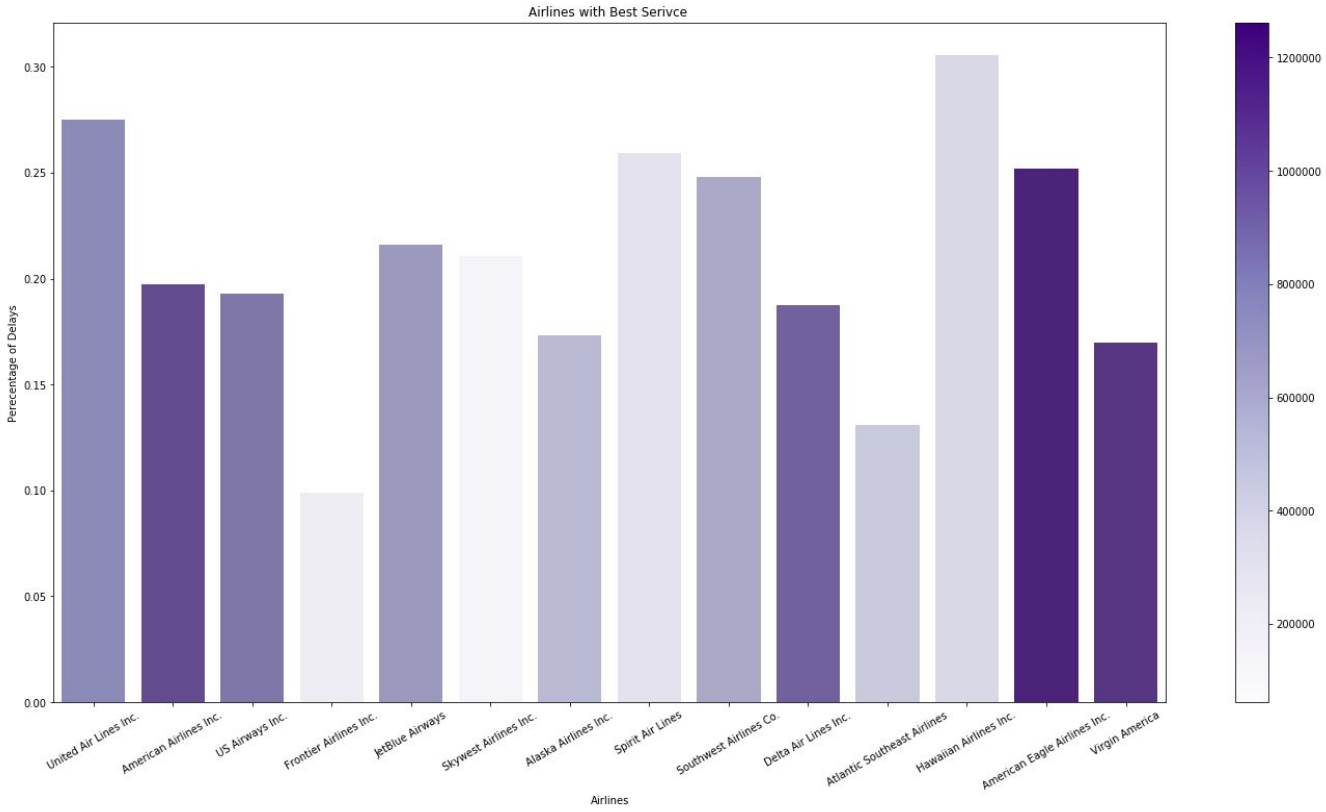
Data Analysis

The bar graph belows shows number of flights per month:



For the data analysis, we started with some descriptive statistics to get acquainted with the data. After we created machine learning model of support vector machines (SVM) to predict airline probability of having a delay with SciKit Learn values. SVM is mostly used for classification problems, making it ideal to answer our question. It works by segregating classes based on coordinates of individual observation on a hyperplane.

The bar graph below shows delays for each airline in percentages and the number of flights that each airline had over the entire year of 2015.



SVM Airline classification report, confusion matrix and accuracy score:

```
[[   0    0    0    0    0    0    0    0 1764    0    0    0   11    0]
 [   0    0    0    0    0    0    0    0 1864    0    0    0   42    0]
 [   0    0    0    0    0    0    0    0 1364    0    0    0   12    0]
 [   0    0    0    0    0    0    0    0  318    0    0    0    9    0]
 [   0    0    0    0    0    0    0    0  890    0    0    0    8    0]
 [   0    0    0    0    0    0    0    0 1993    0    0    0   74    0]
 [   0    0    0    0    0    0    0    0  525    0    0    0    0    0]
 [   0    0    0    0    0    0    0    0  353    0    0    0    5    0]
 [   0    0    0    0    0    0    0    0 4199    0    0    0   48    0]
 [   0    0    0    0    0    0    0    0 2627    0    0    0    3    0]
 [   0    0    0    0    0    0    0    0 2132    0    0    0   95    0]
 [   0    0    0    0    0    0    0    0  239    0    0    0    0    0]
 [   0    0    0    0    0    0    0    0 1051    0    0    0  164    0]
 [   0    0    0    0    0    0    0    0  210    0    0    0    0    0]]
              precision    recall  f1-score   support

           0       0.00      0.00      0.00      1775
           1       0.00      0.00      0.00      1906
           2       0.00      0.00      0.00      1376
           3       0.00      0.00      0.00       327
           4       0.00      0.00      0.00       898
           5       0.00      0.00      0.00      2067
           6       0.00      0.00      0.00       525
           7       0.00      0.00      0.00       358
           8       0.22      0.99      0.35      4247
           9       0.00      0.00      0.00      2630
          10       0.00      0.00      0.00      2227
          11       0.00      0.00      0.00       239
          12       0.35      0.13      0.19      1215
          13       0.00      0.00      0.00       210

   micro avg       0.22      0.22      0.22     20000
   macro avg       0.04      0.08      0.04     20000
weighted avg       0.07      0.22      0.09     20000

0.21815
```

SVM Day of Week Classification Report and Confusion Matrix:

```
[[   0    0    0   83 3219    0    0]
 [   0    0    0   80 2977    0    0]
 [   0    0    0   19 1131    0    0]
 [   0    0    0  103 2718    0    0]
 [   0    0    0   47 3278    0    0]
 [   0    0    0   68 3066    0    0]
 [   0    0    0   72 3139    0    0]]
              precision    recall  f1-score   support

           1       0.00      0.00      0.00      3302
           2       0.00      0.00      0.00      3057
           3       0.00      0.00      0.00      1150
           4       0.22      0.04      0.06      2821
           5       0.17      0.99      0.29      3325
           6       0.00      0.00      0.00      3134
           7       0.00      0.00      0.00      3211

   micro avg       0.17      0.17      0.17     20000
   macro avg       0.06      0.15      0.05     20000
weighted avg       0.06      0.17      0.06     20000

0.16905
```

Project Code Github Link:
https://github.com/ChillOutAlan/652-Final-Project/blob/master/Cleansing.ipynb