MIS41040 – Business Decision Support System

Professor: Dr Peter Keenan

Tableau Assignment – Group 34

Richard Roy Kattiparmbil	22200144
Siddhant Bajpai	22200242
Aravind Renjan	22200890

Date - 16/04/2023

Declaration of Authorship

I declare that all material in this assessment is my own work except where there is clear acknowledgement and appropriate reference to the work of others.

Richard Roy Kattiparmbil Date: 16/04/2023

Siddhant Bajpai

Aravind Renjan

Tableau Public Link

https://public.tableau.com/views/MIS41040_Team-34/FlightDelay2018?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link

Problem Statement

Provide an interactive decision support system (DSS) to better understand the issue of flight delays in the airline industry. The dashboard should be able to help mine insights as to what are the reasons for these delays and how they might be interconnected causing a cascading effect throughout the industry. The information can be further used to find solutions to the flight delay issue at hand. The data to develop the dashboard used is flight details for multiple airlines in the US for the year of 2018.

Raw Data Exploration and Cleaning

Flight information Dataset:

The raw "Flights" data is divided into 12 files for each month of the year 2018 which were combined to form a single dataset.

The dataset contains 5689512 rows and 120 columns out of which not all are necessary for our analysis. Filtering out the required dataset for the purpose of creating the dashboard for our problem statement. List of chosen columns:

'FlightDate', 'Origin', 'Cancelled', Diverted', 'Dest', 'CRSDepTime', 'DepTime', 'DepDelayMinutes', 'DepDelay', 'ArrTime', 'ArrDelayMinutes', 'AirTime', 'Year', 'Quarter', 'Month', 'DayofMonth', 'DayOfWeek', 'Marketing_Airline_Network', 'IATA_Code_Operating_Airline', 'Tail_Number', 'OriginAirportID', 'OriginCityName', 'OriginState', 'OriginStateName', 'DestAirportID', 'DestCityName', 'DestState', 'DestStateName', 'DepDel15', 'DepartureDelayGroups', 'DepTimeBlk', 'TaxiOut', 'WheelsOff', 'WheelsOn', 'TaxiIn', 'CRSArrTime', 'ArrDelay', 'ArrDel15', 'ArrivalDelayGroups', 'ArrTimeBlk', 'DivAirportLandings', 'Operating_Airline', 'NASDelay', 'WeatherDelay', 'CarrierDelay', 'SecurityDelay', 'LateAircraftDelay'.

Which is 5689512 rows and 47 columns.

In the primary exploration phase of the dataset, it can be identified that the fields "Cancelled" and "Diverted" does not have additional supporting fields to provide insights into the reasons for cancellation or where the flight was diverted to.

Also checking the total number of NULL values in the dataset we can see that the maximum number of NULL values are associated to the supporting fields for these two fields. Hence dropping the "Cancelled" and "Diverted" columns from the dataset to focus on the delay aspect in the airline industry for this project (Agarwal, 2023).

The highest number of NULL values are in the fields 'NASDelay', 'WeatherDelay', 'CarrierDelay', 'SecurityDelay' and 'LateAircraftDelay' which is 4601173 for each field. **One interesting pattern that was identified during the data exploration phase is that the records where there are NULL values for these fields are for flights that did not have a**

delay. This can be confirmed as no record is available in the dataset where all the five delays have the values "0" meaning no delays and no value in the arrival and departure delay. So, it is considerably safe to assume that the records with NULL value can be replaced with "0".

Next the maximum number of NULL values are in the column "AirTime". After further investigation of the dataset, it can be confirmed that the columns were there are NULL values for "AirTime" are for the flights that are cancelled or diverted. Also, other columns with high NULL values like "ArrDelay", "DepDelay" and "Tail_Number" are also for the records for which flights were either cancelled or diverted. Compared to the whole dataset these records are only 1.9% therefore dropping these records (Agarwal, 2023).

Python data cleaning step code snippets:

```
#Columns with max null values:
AirTime                     109271
ArrDelayMinutes             102893
ArrDelay                    102893
ArrDel15                    102893
ArrivalDelayGroups          102893
#WheelsOn                    97549
#TaxiOut                     94694
#TaxiIn                      97559
```

```
#Column with highest number of NULL values is #AirTime

#Number of columns where AirTime and ArrDelay are NULL
flights_data_subset[flights_data_subset['AirTime'].isna()&flights_data_subset['ArrDelay'].isna()] #Count-102345

#Number of columns where AirTime and DepDelay are NULL
flights_data_subset[flights_data_subset['AirTime'].isna()&flights_data_subset['DepDelay'].isna()] #Count-85470

#Number of columns where AirTime and DepDelay are NULL
flights_data_subset[flights_data_subset['AirTime'].isna()&flights_data_subset['Tail_Number'].isna()] #Count-20548

#This shows that most columns have NULL values for which the AirTime is NULL meaning the flight was either cancelled or diverted
```

After removing these records, the remaining NULL values are less than 1% of the whole dataset hence dropping those records as well to have a dataset with no NULL values which is now 5578618 rows and 45 columns.

Check if columns that are supposed to have standard values have non-standard data. For example, "DepTimeBlk" and "ArrTimeBlk". Both have 19 intervals in hourly basis except for 0001-0559 midnight to early morning intervals that are consolidated into a single block. Confirmed that these values have no issue.

```
#Check if columns that are supposed to have standard values have non-standard data
delint = flights_data_subset['DepTimeBlk'].value_counts()
arrint = flights_data_subset['ArrTimeBlk'].value_counts()
print('Departure Intervals:', delint)
print('Arrival Intervals:', arrint)
#Both have 19 intervals in hourly basis except for 0001-0559 midnight to early morning intervals that are consolidated
#So no issues
```

```
Departure Intervals: 0600-0659    407886
0700-0759    385375
1700-1759    380583
0800-0859    374031
1200-1259    354433
1100-1159    350151
1000-1059    346695
1400-1459    342155
1500-1559    340458
0900-0959    335335
1800-1859    320661
1900-1959    320494
1600-1659    318870
1300-1359    314184
2000-2059    262828
2100-2159    181657
0001-0559    164576
2200-2259    147502
2300-2359     41638
Name: DepTimeBlk, dtype: int64
```

Airlines Information Dataset:

In the "Airlines" dataset there was only 1 NULL value out of the 1571 rows and 2 columns for the record.

```
airlines_data[airlines_data['Code'].isna()]
```

	Code	Description
929	NaN	North American Airlines

After investigation it was found the same Airlines is mentioned twice for "North American Airlines Inc" as "North American Airlines" without a code and hence was removed.

In the raw Flights dataset, there are only 29 unique airlines that are listed with flight information hence dropping the rest of the records in the Airlines dataset.

Airport Information Dataset:

In the "Airport" dataset there were only three records that had NULL values for "Latitude" and "Longitude" fields which were manually inserted since these are constant values that can be collected from open data sources (Maps.ie, 2023).

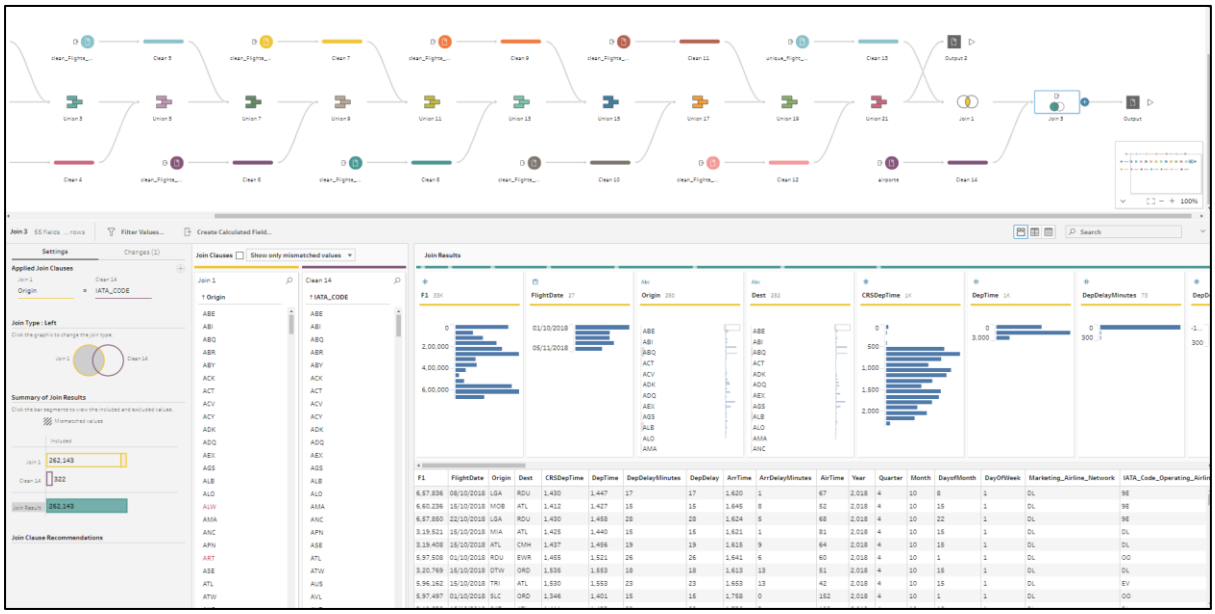
Python data cleaning step code snippets:

```
IATA_CODE    0
AIRPORT      0
CITY         0
STATE        0
COUNTRY      0
LATITUDE     3
LONGITUDE    3
dtype: int64
```

	IATA_CODE	AIRPORT	CITY	STATE	COUNTRY	LATITUDE	LONGITUDE
96	ECP	Northwest Florida Beaches International Airport	Panama City	FL	USA	NaN	NaN
234	PBG	Plattsburgh International Airport	Plattsburgh	NY	USA	NaN	NaN
313	UST	Northeast Florida Regional Airport (St. August...	St. Augustine	FL	USA	NaN	NaN

After cleaning all the three datasets they were joined together in Tableau Data Prep.

Joining data files in Tableau Prep Builder



Details of the most used fields from the dataset (extracted from data dictionary attached to the dataset):

- **Operating_Airline:** Unique Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2).
- **DepDelay:** Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.
- **DepDelayMinutes:** Difference in minutes between scheduled and actual departure time. Early departures set to 0.
- **DepDel15:** Departure Delay Indicator, 15 Minutes or More (1=Yes)
- **DepartureDelayGroups:** Departure Delay intervals, every (15 minutes from <-15 to >**180)**
- **ArrDelay:** Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.
- **ArrDelayMinutes:** Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0.
- **ArrDel15:** Arrival Delay Indicator, 15 Minutes or More (1=Yes)
- **ArrivalDelayGroups:** Arrival Delay intervals, every (15-minutes from <-15 to >180)
- **CarrierDelay:** Carrier Delay, in Minutes
- **WeatherDelay:** Weather Delay, in Minutes

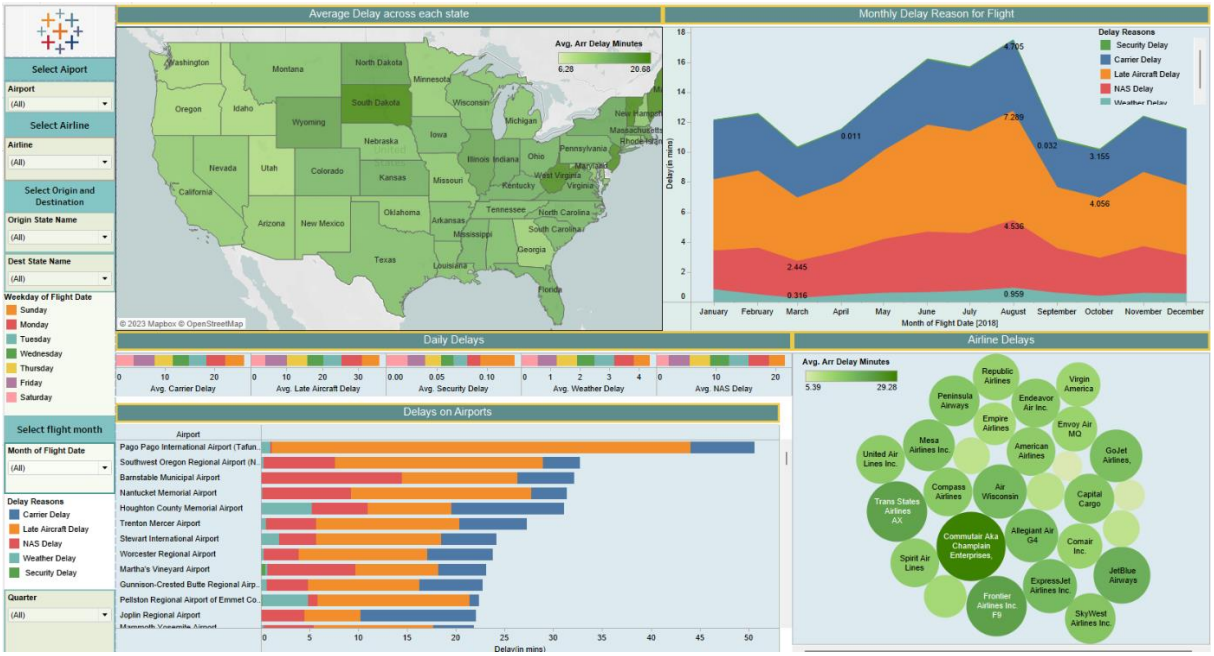
- **NASDelay:** National Air System Delay, in Minutes
- **SecurityDelay:** Security Delay, in Minutes
- **LateAircraftDelay:** Late Aircraft Delay, in Minutes

Decision Support System Dashboard

The dashboard is divided into three parts focusing on the overview of all the information mined from the dataset and also individual focus points of delay as well as the airlines.

Dashboard View – Delay Information

In the DSS dashboard the default window that opens up first has the delay information which is the main focus of our solution.



This dashboard is divided into 5 main parts which are:

- Average Delay across each state
  - In this map view of the US hovering over each state would give further details of average delay for all airlines for the year by default.
- Monthly Delay Reasons for flight
  - This area chart shows monthly average value specific to each type of reason that caused the delay by default.
- Daily Delays
  - This bar chart shows daily average value specific to each type of reason that caused the delay by default.
- Airline Delays
  - The bubble chart shows the average delay for each airline for the whole year by default when hovering over the bubbles.
- Delays on Airport
  - This bar chart shows yearly average value specific to each type of reason that caused the delay for each airport by default.

The filers that can be utilized to further drill down for specific information are:

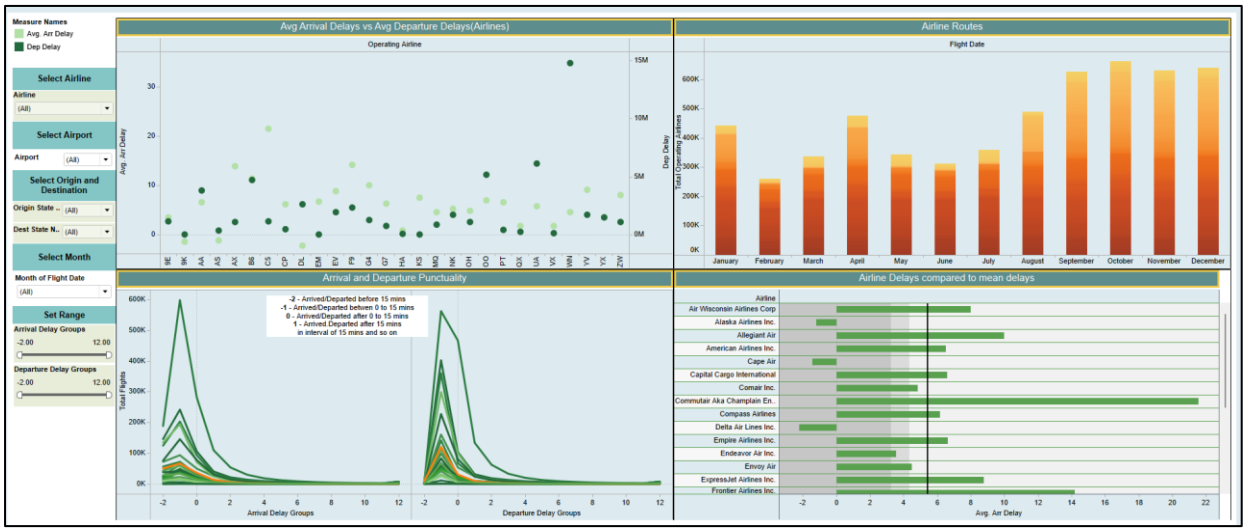
- Airport
- Airline
- Origin State Name
- Destination State Name
- Month of the Flight Date
- Quarter

All filters are applied to above mentioned 5 parts of the dashboard and can be used to find out detailed and specific information in various combinations using these filters. Some examples include:

- How many flights were delayed in a certain month?
- Which state had maximum number of delays for a certain month?
- How many flights were delayed in a certain airport for a certain month for a certain airline?
- What is the average daily/monthly/quarterly delay for airport/airlines, etc?

Dashboard View – Airline Information

In the next dashboard the window displays information that focuses on the various KPIs related to an airline to further drill down to the airline industry stats and performance.



This dashboard is divided into 4 main parts which are:

- Average Arrival Delay vs Average Departure delay
  - In the scatter plot the average arrival delay is plotted along with the average departure delay for each airline against each airport.
- Airline Routes
  - This bar graph shows how many flights were flown monthly from source to destination city.
- Arrival and Departure Punctuality

  - This line chart shows the Arrival and Departure Punctuality KPI for the airlines in a 15 min interval from -15 to 180 minutes.
- Airline Delays compared to mean delays.
  - This comparative bar graph shows the delay stats for each airline with respect to the average mean delay for the entire year.

The filers that can be utilized to further drill down for specific information are:

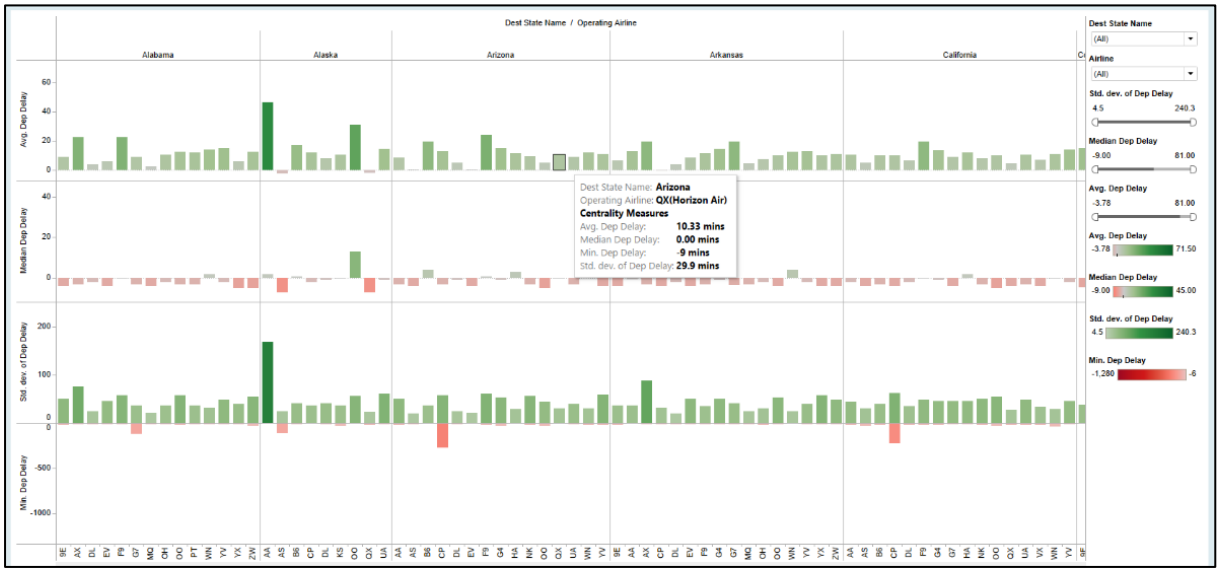
- Airline
- Airport
- Origin State Name
- Destination State Name
- Set Range for Arrival Delay Groups
- Set Range for Departure Delay Groups

All filters are applied to above mentioned 5 parts of the dashboard and can be used to find out detailed and specific information in various combinations using these filters. Some examples include:

- What is the arrival and departure delay relationship for each airline?
- KPI metrics for each airline such as Arrival and Departure Punctuality
- Comparative average delay metrics for each airline
- Monthly flight traffic for each airline from one/all airports, etc.

Dashboard View – Centrality Measures Information

In this dashboard the centrality measures of the airlines are displayed with respect to each destination airport.



- This dashboard is sectioned into 5 parts which are:
- Average Departure Delay
- Median Departure Delay
- Standard Deviation Departure Delay
- Minimum Departure Delay
- Maximum Departure Delay

The filers that can be utilized to further drill down for specific information are:

- Airline
- Destination State Name
- Value Range sliders for Median, Average, Standard Deviation

All filters are applied to above mentioned 5 parts of the dashboard and can be used to find out detailed and specific information in various combinations using these filters. Some examples include:

- What is the average departure delay for an airline?
- What is the minimum departure delay for an airline?
- What is the maximum departure delay for an airline, etc.

Dashboard View – Overview Information

In this dashboard an overview of the entire flights over the year of 2018 in the US.

This dashboard is divided into 6 main parts which are:

- Airport Traffic
  - The number of flights that departed from each airport for the year 2018.
- Total Flights per month

    - Information regarding the number of flights per month in the US country as a whole.
- Flight Status
    - Information regarding the punctuality of flights with regards to airports and airlines.
- Daily Flights
    - Information regarding daily flights with regards to airports and airlines.
- Types of Delay
    - Overall delay percentage analysis
- Number of Flights per airline
    - Information about number of flights per airline

The filers that can be utilized to further drill down for specific information are:

- Airline
- Airport
- City
- Month

All filters are applied to above mentioned 6 parts of the dashboard and can be used to find out detailed and specific information in various combinations using these filters. Some examples include:

- What is number of flights that departed from a city in certain month?
- Which airline had maximum number of flights scheduled to depart from a city/airport?
- What is the airport traffic statistics for each airline, etc.?

Reference List

Agarwal, M. (2023) *Pythonic data cleaning with pandas and NumPy*. Real Python. Available at: https://realpython.com/python-data-cleaning-numpy-pandas/ (Accessed 13 April 2023).

Ravikiran, S. (2023) *Learn how to create Tableau Dashboard [a step by step guide]*, *Simplilearn.com*. Simplilearn. Available at: https://www.simplilearn.com/tutorials/tableau-tutorial/tableau-dashboard (Accessed 15 April 2023).

Maps.ie *Find GPS coordinates on Google Maps, map of Ireland*. Available at: https://www.maps.ie/coordinates.html (Accessed 15 April 2023).