Mid term project EDA Notes

This file contains our hypothesis during EDA

## Passengers DB

Notes

* What is departure scheduled / departures performed?Is the payload only passengers or total? (is it freight + mail + available seats?)
* How to join this to the flights DB?

Numerical data (continuous):

'departures\_scheduled', 'departures\_performed', 'payload', 'seats',

'passengers', 'freight', 'mail', 'distance', 'ramp\_to\_ramp', 'air\_time',

'airline\_id', 'origin\_airport\_id',

'origin\_city\_market\_id', 'dest\_airport\_id', 'dest\_city\_market\_id',

'aircraft\_group', 'aircraft\_type', 'aircraft\_config',

Numerical data (categorical):

'Carrier\_group' (4 values), 'Carrier\_group\_new' (4 values)

'aircraft\_config', (4 values)

'year', 'month', 'Day'

String data

'unique\_carrier', 'unique\_carrier\_name' (377 values)

'Region', (6 values)

'Carrier' (388 values), 'carrier\_name', (396 values)

'origin', 'origin\_city\_name', (2514 values)

'origin\_country', 'origin\_country\_name', 154 values)

'dest', 'dest\_city\_name',

'dest\_country',

'dest\_country\_name',

'distance\_group',

'Class' (4 values)

Redundant features removed one of the two

carrier and unique\_carrier (99% similar)

unique\_carrier\_name and carrier\_name (94% similar)

carrier\_group and carrier\_group\_new (94% similar)

Highly correlated features remove one of the two?

origin and origin\_city\_name (70% correlated)

origin country and origin\_country\_name (91% correlated)

dest and destination\_city\_name (70% correlated)

dest\_country country and dest\_country\_name (91% correlated)

Unique\_carrier and unique\_carrier name (50% correlated - keeping those two)

## Fuels DB

Notes:

This DB is structured like a report showing fuel consumption (in gallons) per month & per airline.

Confirm the unique ID of this DB.

1. Scheduled

Total scheduled domestic

International atlantic

International pacific

International latin america

International scheduled subtotal

Total scheduled

1. Real

Total domestic

Total international

Grand total real

Numerical data (continuous):

'sdomt\_gallons',

'satl\_gallons', 'spac\_gallons', 'slat\_gallons', 'sint\_gallons',

'ts\_gallons', 'tdomt\_gallons', 'tint\_gallons', 'total\_gallons',

'sdomt\_cost', 'satl\_cost', 'spac\_cost', 'slat\_cost', 'sint\_cost',

'ts\_cost', 'tdomt\_cost', 'tint\_cost', 'total\_cost']

Numerical data (categorical):

'Month’, 'year', 'airline\_id', 'carrier\_group\_new'

String data

['unique\_carrier', 'carrier', 'carrier\_name']

Redundant features removed one of the two

carrier and unique\_carrier (99% similar)

**Task 1 - Maggie**

* Starting here to start small (with outliers it is not normally distributed)
* Using IQR to determine outliers you can take anything outside the range of -33 to 33 minutes as an outlier (this accounts for 12.6% of all observations) Hesitant to make such a narrow window
* LOF analysis in progress
  + 25 neighbors -- -67 to 286 minutes
  + 15 neighbors -68 to 302 minutes
  + 35 neighbors -58 to 266 minutes

**Task 2 - Maggie**

* Seems to be more “outliers” in june/july overall the median is lower but you can expect more longer delays in those months
* Less impact due to winter months than anticipated

**Task 3 - Isa**

Notes from technical documentation

*Below is a list of examples of causes for delays and cancellations that we believe are the result of weather. This list should be used as a guide for the type of occurrences that should be reported as an air carrier delay and/or cancellation. It should not be considered a complete list, and we welcome comments on additions or deletions. WEATHER*

*Below minimum conditions*

*Clear ice inspection*

*Deicing aircraft*

*Earthquake*

*Extreme high or low temperatures*

*Hail Damage*

*Holding at gate for enroute weather*

*Hurricane*

*Lightning damage*

*Pre-planned cancellations that result from predicted weather*

*Snow Storm*

*Thunder Storm T*

*ornado*

API research

Meteomatics

Registered for a free trial (valid until November 02)

**Task 4 - Maggie**

* Strategy will be binning
* Midnight to 3 am == 0
* 3am to 6am == 1
* 6am to 9am ==2
* 9am to noon =3
* Noon to 3pm =4
* 3pm to 6pm =5
* 6pm to 9pm = 6
* 9pm to midnight = 7

**Task 5 - Maggie**

* This is interesting to look at and good information about how delayed departure impacts
* If the departure is delayed ~75% you will be delayed in the end
* If the arrival is delayed you almost definitely we delayed on your departures
* Task 7 ties into this so I am going to look there next

**Task 6 -**

**Task 7 - Maggie**

* Ties into task 5
* use air time then distance over time
* Compare delayed departure to no delayed depature
* May also try to play with the direction in here
* With big sample you have strong evidence to support that planes fly faster if they have a delayed depature

**Task 8 - Maggie**

* Linking to Task 4 and going to regroup that one to hourly and then will be able to bin the distances

**Task 9**

**Task 10**

Additional Questions/Thoughts/Hypotheses

* Direction of travel and speed -- like is a flight more likely to be delayed if its traveling E-W than W-E or vice-versa
* Why do we have an origin airport ID and destination airport ID and origin city market ID and destination city market ID in the passengers DB?