Data Science Course Exercise Workbook

Final Project Brief – Airline Flight Delays Analysis

Course Material

Stuart Brown, June 2017

Data Science – Final Project

Stuart Brown, June 2017

08

**Fall**

Contents

[Summary 3](#_Toc485218204)

[The Data: 3](#_Toc485218205)

[The Task: 4](#_Toc485218206)

[What you need to deliver? 4](#_Toc485218207)

# Summary

This document serves to provide some guidance for the trainees for their final project for the Data Science pathway. Trainees will use the following datasets.

1. 2019-01.csv
2. 2019-02.csv
3. 2019-03.csv
4. airports.csv
5. carriers.csv
6. plane-data.csv

# The Data:

The data for this project can be categorized into 4 sections.

Flight details are captured on files **2019-01.csv**, **2019-02.csv**, **2019-03.csv**.

Airport data are in the **airports.csv**,

carrier information in the **carriers.csv**

and plane data in the **plane-data.csv**.

Attributes shared by the flight details include:

* Year - 2019
* Month - 1-12
* DayofMonth - 1-31
* DayOfWeek - 1 (Monday) - 7 (Sunday)
* DepTime - actual departure time (local, hhmm)
* CRSDepTime - scheduled departure time (local, hhmm)
* ArrTime - actual arrival time (local, hhmm)
* CRSArrTime - scheduled arrival time (local, hhmm)
* UniqueCarrier - unique carrier code
* FlightNum - flight number
* TailNum - plane tail number
* ActualElapsedTime - in minutes
* CRSElapsedTime - in minutes
* AirTime - in minutes
* ArrDelay - arrival delay, in minutes
* DepDelay - departure delay, in minutes
* Origin - origin IATA airport code
* Dest - destination IATA airport code
* Distance - in miles
* TaxiIn - taxi in time, in minutes
* TaxiOut - taxi out time in minutes
* Cancelled - was the flight cancelled?
* CancellationCode - reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
* Diverted - 1 = yes, 0 = no
* CarrierDelay - in minutes
* WeatherDelay - in minutes
* NASDelay - in minutes
* SecurityDelay - in minutes
* LateAircraftDelay - in minutes

The airports.csv contains the following attributes: (for map)

* iata - the international airport abbreviation code
* airport - name of the airport
* city - name of the city the airport is located
* state - the state code for the airports location
* country - country in which airport is located.
* lat - latitude co ordinate
* long - longitude co ordinate

The carriers.csv contains the following attributes:

* code - unique carrier code
* description - carrier full name

The plane-data.csv contains the following attributes:

* tailnum - unique tail number of the aircraft
* type - the type ownership of the aircraft
* manufacturer - the manufacturer of the aircraft
* issue\_date - date aircraft was issued
* model - the model number of the aircraft
* status - status indicator
* aircraft\_type - description of aircraft
* engine\_type - engine type used
* year - year of manufacturing

# The Task:

The aim of the project is to provide a graphical summary of important features of the data set, combined with your own exploration.

You **must** use **HiveQL and Spark** to process the data. It’s up to you whether you decide to use all or just one. The only requirement is that you can justify your choices. Python can be used to visualise your results.

There should be a focus in your analysis to detect occurrences and cause of flight delays. A small example of this could be to look into:

* When is the best time of day/day of week/time of year to fly to minimise delays?
* Do older planes suffer more delays?
* How does the number of people flying between different locations change over time?
* Can you detect cascading failures as delays in one airport create delays in others? Are there critical links in the system?
* Which carrier experiences lowest/highest number of delays?

Aside from this requirement, you are free to choose which other feature(s) to explore, how far to take the analysis and the methods you used to do it. Perhaps you could look to the operational efficiency of the larger airports.

# What do you need to deliver?

Create a presentation, in the time allocated to you, that analyses the dataset and be prepared to present to and engage with the class by the end of the week.

To Include:

* Introduction to your presentation
* Method (What do you want to find out? How you went about doing it?)
* Findings (With accompanying evidence, any code & data visuals)
* Any challenges/ issues? And how you resolved them?
* Conclusion

**Notes:**

**Take care to take legible screenshots of the important parts of your code.**

**Have the application open for practical demonstration.**

**You will need to make clear use of at least 4 of the datasets.**

This task is to be completed individually. It will be graded on insight derived, technical content and professional practice.

Marks are available for clear data visualisation, the quality of your code used and on your presentation about your interpretation and insight into the data.