Easyjet Assignment

Kevin Grainger

July 2025

Task 1

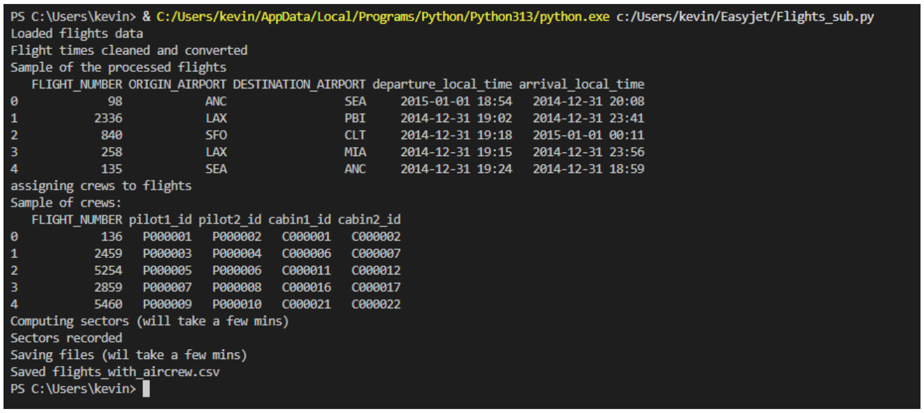
See linked python file containing code used for data generation.
Linked is the excel table, with the executed VBA code.
Possible Improvements to VBA:

- Addition of error handling, in the case of missing data or incorrect formatting.

  ```
  If IsNumeric(age) And town <> "" Then        (1)
  ```

- There could be issue with the value 'k', integers in VBA can overflow. Better to use 'long'.

Sample Output from flights.csv analysis:



```
PS C:\Users\kevin> & C:/Users/kevin/AppData/Local/Programs/Python/Python313/python.exe c:/Users/kevin/Easyjet/Flights_sub.py
Loaded flights data
Flight times cleaned and converted
Sample of the processed flights
   FLIGHT_NUMBER ORIGIN_AIRPORT DESTINATION_AIRPORT departure_local_time arrival_local_time
0             98            ANC                 SEA     2015-01-01 18:54   2014-12-31 20:08
1           2336            LAX                 PBI     2014-12-31 19:02   2014-12-31 23:41
2            840            SFO                 CLT     2014-12-31 19:18   2015-01-01 00:11
3            258            LAX                 MIA     2014-12-31 19:15   2014-12-31 23:56
4            135            SEA                 ANC     2014-12-31 19:24   2014-12-31 18:59
assigning crews to flights
Sample of crews:
   FLIGHT_NUMBER pilot1_id pilot2_id cabin1_id cabin2_id
0            136   P000001   P000002   C000001   C000002
1           2459   P000003   P000004   C000006   C000007
2           5254   P000005   P000006   C000011   C000012
3           2859   P000007   P000008   C000016   C000017
4           5460   P000009   P000010   C000021   C000022
Computing sectors (will take a few mins)
Sectors recorded
Saving files (wil take a few mins)
Saved flights_with_aircrew.csv
PS C:\Users\kevin> 
```

Figure 1: Output in Terminal

Task 2

Suppose that the data tables above were imported from the server, what kind of obstacles would you encounter? Explain how would you tackle with them?

1. Large data load will decide could overwhelm Pandas, you need to look out for bottle-necks and size restrictions (e.g. data dimension). Using cloud processing like Pyspark or Dask would help.

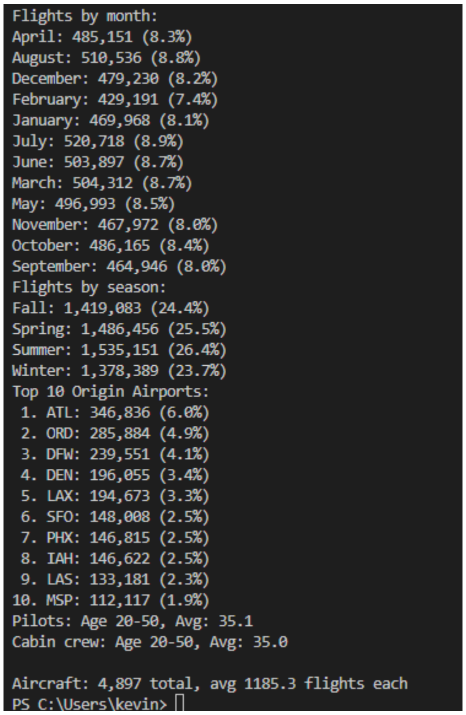
2. Data cleaning could be an issue (empty spaces, incorrect character types, etc.), the format could also be an issue. A code that can detect data formats, and process un-cleaned data would be an approach.

Check "flights with aircrew.csv" data and check what type of cleanings are required?

- We need to check the IATA code is in the correct form.
- Make sure the time format is consistent.
- Trim unwanted spaces.
- Account for all missing entries and duplicates.

Task 3

I did not get to complete this section but got a good chunk of the analysis done:



```
Flights by month:
April: 485,151 (8.3%)
August: 510,536 (8.8%)
December: 479,230 (8.2%)
February: 429,191 (7.4%)
January: 469,968 (8.1%)
July: 520,718 (8.9%)
June: 503,897 (8.7%)
March: 504,312 (8.7%)
May: 496,993 (8.5%)
November: 467,972 (8.0%)
October: 486,165 (8.4%)
September: 464,946 (8.0%)
Flights by season:
Fall: 1,419,083 (24.4%)
Spring: 1,486,456 (25.5%)
Summer: 1,535,151 (26.4%)
Winter: 1,378,389 (23.7%)
Top 10 Origin Airports:
 1. ATL: 346,836 (6.0%)
 2. ORD: 285,884 (4.9%)
 3. DFW: 239,551 (4.1%)
 4. DEN: 196,055 (3.4%)
 5. LAX: 194,673 (3.3%)
 6. SFO: 148,008 (2.5%)
 7. PHX: 146,815 (2.5%)
 8. IAH: 146,622 (2.5%)
 9. LAS: 133,181 (2.3%)
10. MSP: 112,117 (1.9%)
Pilots: Age 20-50, Avg: 35.1
Cabin crew: Age 20-50, Avg: 35.0

Aircraft: 4,897 total, avg 1185.3 flights each
PS C:\Users\kevin> 
```

Figure 2: Terminal Output

Task 4

Consider that you have a larger dataset in a cloud storage, what technologies would you utilise for processing a given task?

- Use Dask or PySpark for heavy processing.
- Databricks or similar large data tool, can be used for dashboards also.
- SQL for data queries.

Consider that a crew detail is wrong in the crews table, which is stored in a data lakehouse, how would you approach to resolve this problem?

1. Use SQL to find the incorrect entry.

   ```
   SELECT * FROM crews WHERE CrewID = ...;        (2)
   ```

2. Correct entry:

   ```
   UPDATE crews SET ...  = ...  WHERE CrewID = ...;    (3)
   ```

3. Cloud use Pyspark with SQL also.