**Global Aviation Network Analysis Report**

**Introduction**

This report summarises the steps it took to clean and analyse data from three CSV files. The report also contains findings from these CSV files.

Tools utilised include:

Python and libraries

VS Code, along with its Jupyter Notebook Feature

Final Analysis and results were displayed by building a web application. The web application script was saved under dashboard.py, and the technical documentation was saved under dashboard.txt.

**Dataset Description and Cleaning Steps**

Datasets were saved as three CSV files, namely airports.csv, airlines.csv and routes.csv.

* The Airports Dataset contains metadata on global airports, including name, location, and coordinates.
* The Airlines Dataset lists airline names, IATA codes, countries of operation, and other identifiers.
* The Routes Dataset captures source-destination route information for each airline.

Datasets were first loaded with correct headers. Then missing values were handled and replaced. The Imputation strategy used was the Constant Value Imputation method. The rationale is simple. If a column has too many missing values, this column would be deleted because it provides little relevant information. Thus, more memory is freed up to make computations and programming faster and efficient. However, if a column had only a few missing values, the missing values would be replaced with 'UNKNOWN'.

Airport codes in the routes file were verified against the airport codes in the airports file. Unverified airport codes were removed from the routes file. Inactive airlines and unknown airports were filtered out.

Duplicate entries were also removed from each of the files. Then the geographic coordinates were validated. Finally, the data type for the entries in each dataset was checked to ensure that each data type was consistent with its entry. Changes were made where necessary.

**Key findings from Exploratory Data and Analysis**

Exploratory Data Analysis was made possible by generating descriptive statistics for numerical variables in each of the datasets. Then, frequency tables were generated for the categorical variables. This allowed for the identification of:

The top countries by airport count were the United States, Australia, Brazil and Canada (see Appendix A and B).

Busiest airports by number of connections, which were found to be Atlanta (ATL), Chicago O’Hare (ORD), and Los Angeles (LAX) (see Appendix E and F).

The country with the most airlines is the United States (see Appendix H).

**Key Insights from Merged Data**

A master dataset was generated and saved by merging the route, airline, and airport datasets. Thus, the identification of top airlines by routes was possible. Top Airlines by number of routes, which were found to be FR, AA and UA (see Appendix C and D). With the help of the route distribution table, it was discovered that most airlines focus solely on domestic markets, while others provide flight service for the international market (see Appendix J). Also, geographic analysis showed that major airlines serve hundreds of destinations and show strong network centrality.

**Network Analysis Results**

NetworkX was used to model the air traffic network as a graph and saved it as a GML file. Nodes represented airports while edges represented routes. Graph has 5959 nodes and 18668 edges. Due to the high number of nodes and edges, visualisation in this document is not possible. However, the GML file will be submitted (Also see Appendix G for a sample of a graph). Top hub airports include ATL, ORD and LAX. The network is highly centralised with a few major nodes dominating connectivity. An airport is said to be centralised when it is connected to many other airports through direct routes. A high degree of centrality for an airport indicates that it is directly connected to other airports.

Using the Python Streamlit library, a dashboard with multiple tabs was built to display other analysis results under the geography tab. This includes a map of global airport locations, a map of routes taken by a particular airline, among others.

From the network analysis, airline activity in countries was analysed. It was discovered that only 1253 were active, with 4906 being inactive (see Appendix I).

Dashboard Specifics

The dashboard was created using Python Streamlit. It highlights the information obtained from this exercise under five tabs.

The Overview Tab: Holds general information regarding the total number of airports, the total number of routes, the top hub, etc. It displays the category of routes based on the distances of these routes. It also displays a global interactive map of the top 100 airports by connectivity, which means the airports with the most routes.

The Network Tab: displays information about the top airports with a high degree of centrality. Also displays the top airports that are close to other airports based on route distance. It also shows the most important airports based on the connections these airports have. An Airport that has a high degree of centrality is categorised as a major airport, and airports that are connected to the most major airports are categorised as having a high page rank; thus, they are important.

The Geography Tab: displays global interactive maps of airports, heatmaps and airline-route maps (see Appendix

The Business Tab: displays airline activity, top airlines based on routes, competitive routes, etc. It also allows for the comparison of airline network types.

The Search Tab is used to find routes based on airlines. This can assist in picking the best airline based on route and comparing airline routes.

**Challenges Faced and Solutions Applied**

Some key challenges included:

Conflicting data and records when merging the datasets caused issues like duplicate rows and errors during computation. This was fixed by removing duplicate rows and merging datasets along common columns, removing irrelevant columns and standardising.

**APPENDIX**

Appendix A

A graph with green bars

AI-generated content may be incorrect.

Appendix B

Top 10 Countries by Airport Count

|  |  |  |
| --- | --- | --- |
|  | Country | Number of Airports |
| 0 | United States | 1503 |
| 1 | Canada | 428 |
| 2 | Australia | 328 |
| 3 | Brazil | 260 |
| 4 | Russia | 251 |
| 5 | Germany | 248 |
| 6 | France | 217 |
| 7 | China | 207 |
| 8 | United Kingdom | 165 |
| 9 | India | 145 |

Appendix C

A graph with blue bars

AI-generated content may be incorrect.

Appendix D

Airline and Routes Distribution

|  |  |  |
| --- | --- | --- |
|  | airline | route\_count |
| 199 | FR | 2484 |
| 96 | AA | 2354 |
| 479 | UA | 2180 |
| 160 | DL | 1981 |
| 490 | US | 1960 |
| 150 | CZ | 1454 |
| 332 | MU | 1263 |
| 140 | CA | 1260 |
| 522 | WN | 1146 |
| 476 | U2 | 1130 |

Appendix E

A graph with blue bars

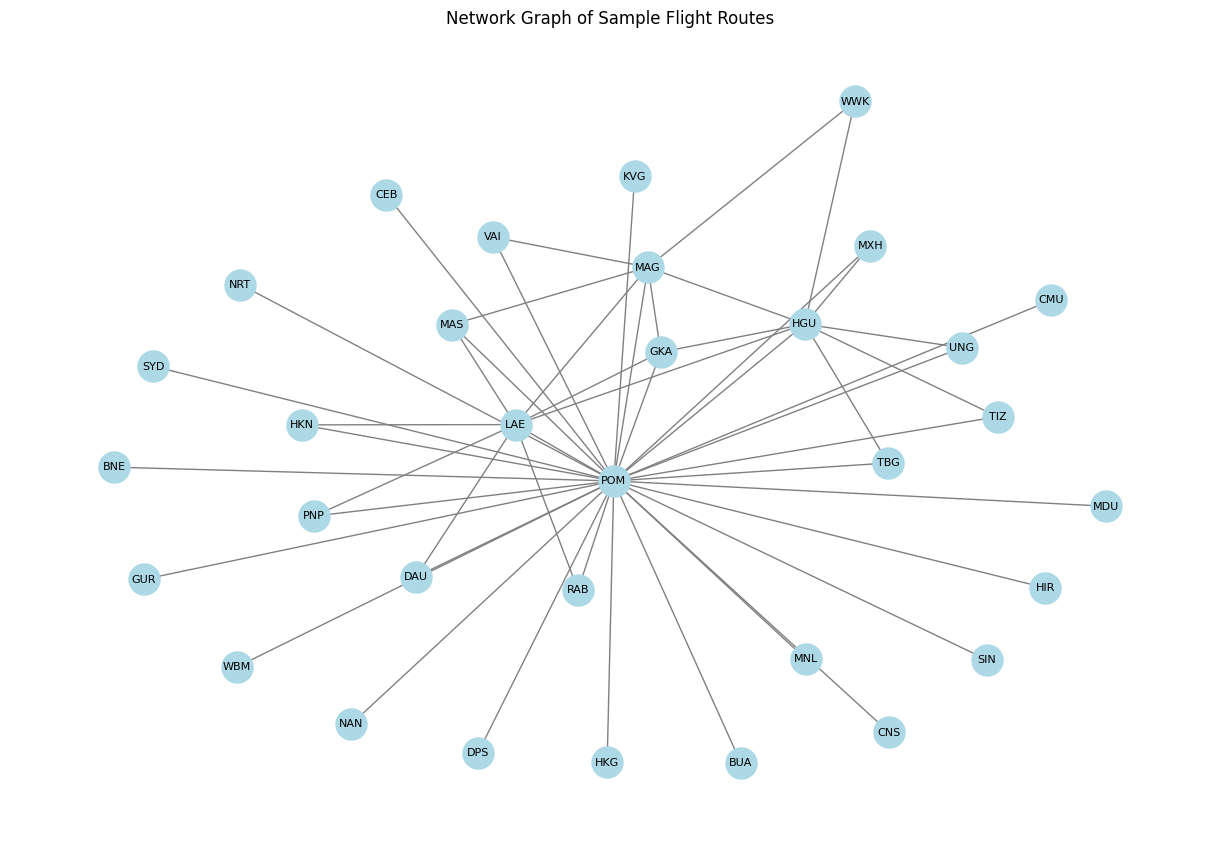
AI-generated content may be incorrect.

Appendix F

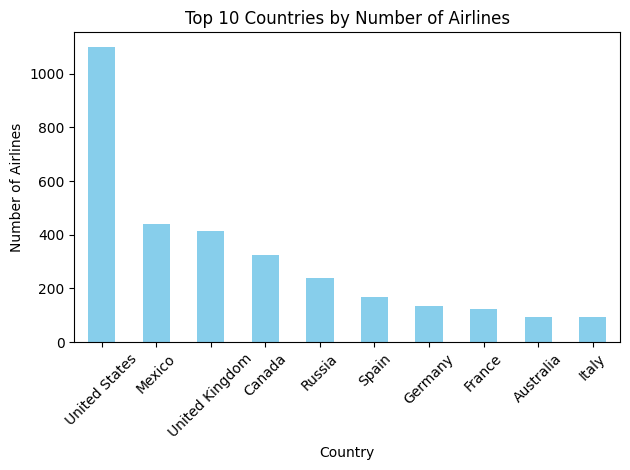
Top Ten Busiest Airports by Connections

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | IATA Code | Airport Name | Country | Connections |
| 165 | ATL | Hartsfield Jackson Atlanta International Airport | United States | 1826 |
| 2128 | ORD | Chicago O'Hare International Airport | United States | 1108 |
| 2200 | PEK | Beijing Capital International Airport | China | 1069 |
| 1626 | LHR | London Heathrow Airport | United Kingdom | 1051 |
| 463 | CDG | Charles de Gaulle International Airport | France | 1041 |
| 890 | FRA | Frankfurt am Main Airport | Germany | 990 |
| 1573 | LAX | Los Angeles International Airport | United States | 990 |
| 667 | DFW | Dallas Fort Worth International Airport | United States | 936 |
| 1293 | JFK | John F Kennedy International Airport | United States | 911 |
| 110 | AMS | Amsterdam Airport Schiphol | Netherlands | 903 |

Appendix G



Appendix H



Appendix I

A graph with a red and green rectangle

AI-generated content may be incorrect.

Appendix J

A green bar graph with white text

AI-generated content may be incorrect.

Appendix K (Picture of interactive Map)

