11253444

Abstract

This report analyses an airline dataset containing 98619 rows and 15 columns. It aims to apply big data manipulation and analytic techniques to the data. Additionally, the report will provide data visualisation and address ethical issues, along with strategies to mitigate them.

Big data  
604IT

Using big data manipulation, mining and analytics techniques to extract meaningful information from large-scale study data set



Contents

[INTRODUCTION 4](#_Toc158905579)

[DATA EXTRACTION AND MANIPULATION 5](#_Toc158905580)

[Accurately identify and source an appropriate large-scale dataset from reliable online resources, relevant to the module's learning outcomes 5](#_Toc158905581)

[Apply data manipulation techniques to clean, preprocess, and transform the selected dataset for analysis 5](#_Toc158905582)

[Demonstrate proficiency in handling big data, ensuring efficiency and scalability of the data manipulation process 11](#_Toc158905583)

[BIG DATA ANALYTICS 13](#_Toc158905584)

[Critically evaluate a range of algorithms and big data analytics techniques suitable for analysing the complex dataset 13](#_Toc158905585)

[Select and justify the most appropriate methods for extracting meaningful information from the large-scale data 13](#_Toc158905586)

[Implement and apply the selected algorithms using appropriate tools or programming languages 13](#_Toc158905587)

[DATA VISUALISATION FOR BUSINESS DECISION MAKING 17](#_Toc158905588)

[Apply data visualisation techniques to present the extracted insights in a clear, informative, and visually appealing manner 17](#_Toc158905589)

[Create appropriate visualisations that cater to the needs of the company executive board and support effective decision-making 17](#_Toc158905590)

[Interpret the visualisations, identify key trends, patterns, and actionable insights, and effectively communicate them to a non-technical audience 17](#_Toc158905591)

[Task 1: What is the total proportion of passengers according to gender? 17](#_Toc158905592)

[Task 2: What is the distribution of passengers per month? Which has the highest and lowest number of passengers? 0](#_Toc158905593)

[Task 3: Which nationality has the highest number of travellers by gender? 0](#_Toc158905594)

[Task 4: What is the percentage of **On time flights** compared to **Cancelled** and **Delayed**? 2](#_Toc158905595)

[Task 5: What is the distribution of passengers across various countries 0](#_Toc158905596)

[Task 6: What is the age trend of passengers? Which country do most young adults travel to? 0](#_Toc158905597)

[Task 7: What airport do certain young adults most frequent? 1](#_Toc158905598)

[Task 8: What is the minimum, median and maximum age of passenger per continent? 0](#_Toc158905599)

[Task 9: Which month has the most delayed flights? 1](#_Toc158905600)

[Task 10 Using chi-square is there any corelation between passenger gender and continent? 2](#_Toc158905601)

[Task 11: Is there a corelation between age and country? 3](#_Toc158905602)

[ETHICAL AND LEGAL CONSIDERATION 0](#_Toc158905603)

[Address and discuss ethical and legal issues related to the handling of big data and the presentation of findings 0](#_Toc158905604)

[Ensure compliance with data privacy regulations and other relevant legal considerations 0](#_Toc158905605)

[Propose strategies to mitigate potential ethical challenges in the use of data for business decision-making purposes 0](#_Toc158905606)

[REFERENCES 1](#_Toc158905607)

[Appendix 2](#_Toc158905608)

[Using WordCloud 2](#_Toc158905609)

[Using KDE plot 4](#_Toc158905610)

[Using Map visuaisation 5](#_Toc158905611)

[Using swarm plot to visualise the top 10 most used airports 0](#_Toc158905612)

[Using stacked bar chart to find out which continent has the highest number of passengers 1](#_Toc158905613)

[Figure 1 Analytical questions 4](#_Toc158905534)

[Figure 2 Airline Dataset 5](#_Toc158905535)

[Figure 3 Counting rows and columns 5](#_Toc158905536)

[Figure 4 Names of columns 6](#_Toc158905537)

[Figure 5 Dropping columns 6](#_Toc158905538)

[Figure 6 Locating NULL values 7](#_Toc158905539)

[Figure 7 Drop duplicate 7](#_Toc158905540)

[Figure 8 Date format 8](#_Toc158905541)

[Figure 9 Changing '/' to '-' 8](#_Toc158905542)

[Figure 10 Changing date format 8](#_Toc158905543)

[Figure 11 Checking for unique values 9](#_Toc158905544)

[Figure 12 Cleaned dataset 10](#_Toc158905545)

[Figure 13 Handling Null values 10](#_Toc158905546)

[Figure 14 Max and Min function 11](#_Toc158905547)

[Figure 15 Grouping by month 11](#_Toc158905548)

[Figure 16 Resetting the index 12](#_Toc158905549)

[Figure 17 Using plotly.express to make a Bar chart 12](#_Toc158905550)

[Figure 18 Importing libraries 13](#_Toc158905551)

[Figure 19 Converting CSV to Data Frame 13](#_Toc158905552)

[Figure 20 Count unique values in Gender 14](#_Toc158905553)

[Figure 21 Using group by and aggregate function 14](#_Toc158905554)

[Figure 22 Sorting values in descending order 15](#_Toc158905555)

[Figure 23 Filter data frame and show Ages 16 to 30 16](#_Toc158905556)

[Figure 24 Pie chart of Male and Female passengers 17](#_Toc158905557)

[Figure 25 Horizontal bar graph distribution of passengers per month 0](#_Toc158905558)

[Figure 26 Line graph of distribution of passengers per month 1](#_Toc158905559)

[Figure 27 Code for sorting the highest number of passengers by Gender and Nationality 0](#_Toc158905560)

[Figure 28 Heatmap code 0](#_Toc158905561)

[Figure 29 Heatmap by Nationality and Gender 1](#_Toc158905562)

[Figure 30 Percentage of Cancelled, On time and Delayed flights 2](#_Toc158905563)

[Figure 31 Pie chart distribution of top 20 passengers from various countries 0](#_Toc158905564)

[Figure 32 Histogram of Age frequency among passengers 0](#_Toc158905565)

[Figure 33 Bar graph of the Top 10 countries visited by young adults 16-30 0](#_Toc158905566)

[Figure 34 Bar graph of the top 10 airports used by young adults 16 - 30 1](#_Toc158905567)

[Figure 35 Box and whiskers chart for Age distribution among Continents 0](#_Toc158905568)

[Figure 36 Horizontal bar chart of delayed flights 1](#_Toc158905569)

[Figure 37 Chi-square computation and Gender to continent heatmap 2](#_Toc158905570)

[Figure 38 Scatter plot of Age to Country 3](#_Toc158905571)

[Figure 39 WordCloud code 2](#_Toc158905572)

[Figure 40 WordCloud 3](#_Toc158905573)

[Figure 41 KDE plot of passenger age 4](#_Toc158905574)

[Figure 42 Geopy code for Map display 5](#_Toc158905575)

[Figure 43 Map visualisation of top 10 most used airports 6](#_Toc158905576)

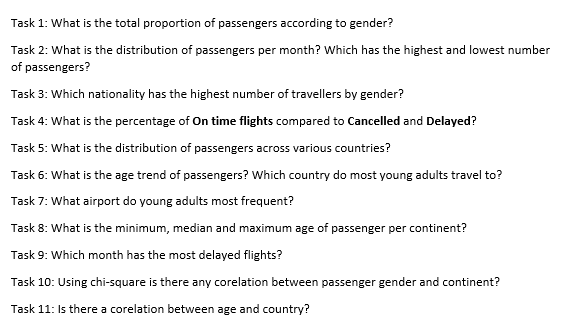
[Figure 44 Swarm plot of the top 10 most used airports 0](#_Toc158905577)

[Figure 45 Horizontal bar chart of continents and their passenger count 1](#_Toc158905578)

# INTRODUCTION

This report aims to utilise data mining, manipulation and analytics technique on an Airline data set (13MB) obtained from Kaggle to extract meaningful information. Furthermore, by using data visualisation it will solve the following analytic questions:

Figure 1 Analytical questions



# DATA EXTRACTION AND MANIPULATION

## Accurately identify and source an appropriate large-scale dataset from reliable online resources, relevant to the module's learning outcomes

Airline dataset is a 13MB downloadable file from Kaggle (see Figure 1). It provides information about passengers, demographics, airports, and flights status. It is a dataset used by aviation industries in identifying trends and can be used in constructing strategies to improve growth in the industry (Airline Dataset, 2023).

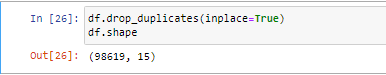
Figure 2 Airline Dataset



## Apply data manipulation techniques to clean, preprocess, and transform the selected dataset for analysis

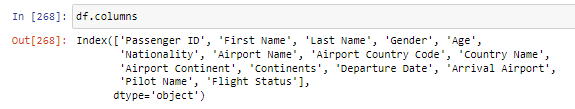
To clean the data set I first identify the number of rows and columns (Figure 3).

Figure 3 Counting rows and columns



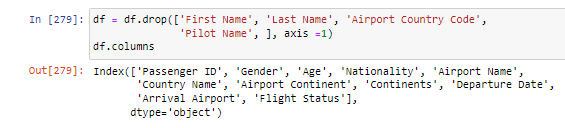
Then identify the different columns (Figure 4)

Figure 4 Names of columns



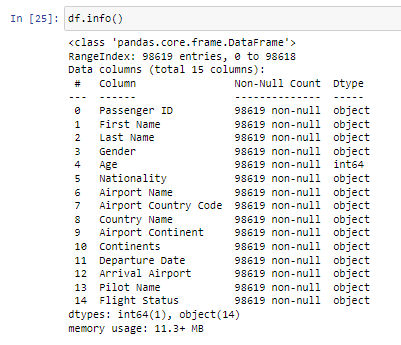
I have identified that the first name, last name, airport country code, and pilot names will not be important. Therefore, I dropped these columns using the drop function (Figure 5).

Figure 5 Dropping columns



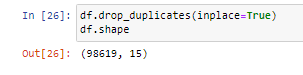
I would then use **.info()** function to check the integrity of the dataset and check the data types. Figure 6 shows that all Non-Null values are 98619, indicating there are no missing information.

Figure 6 Locating NULL values



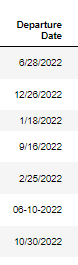
Use **drop\_duplicates()** function to drop any duplicates in the data set. Figure 7 illustrates that there are no duplicates as the number of rows are still 98619.

Figure 7 Drop duplicate



After running **df.head()** again, I have noticed some discrepancy in the data format. Some were separated by ‘**/**’ and some were ‘**-**’ (Figure 8).

Figure 8 Date format



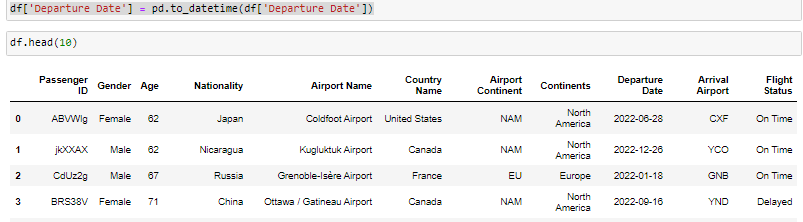
To standardise the column, I modified all ‘**/**’ into ‘**-**’ (Figure 9).

Figure 9 Changing '/' to '-'



Change the date format using the **pd.to\_datetime()** function (Figure 9).

Figure 10 Changing date format



Use the **.unique()** function to check each column if they contain unique outliers in the dataset (Figure 11).

Figure 11 Checking for unique values

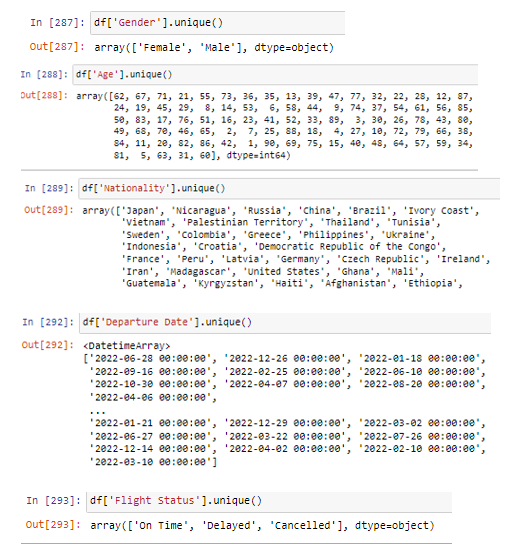
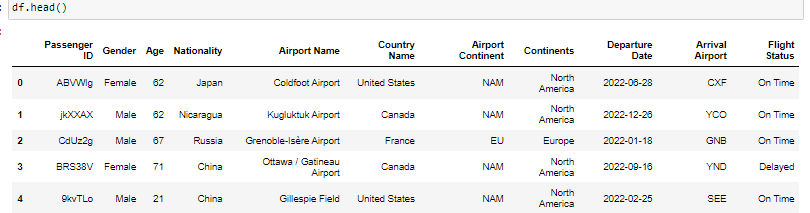


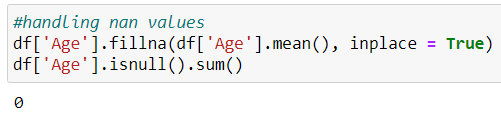
Figure 12 illustrates a cleaned data set.

Figure 12 Cleaned dataset



To handle null values, I impute the null values with the mean (Figure 13).

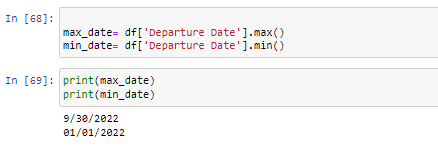
Figure 13 Handling Null values



## Demonstrate proficiency in handling big data, ensuring efficiency and scalability of the data manipulation process

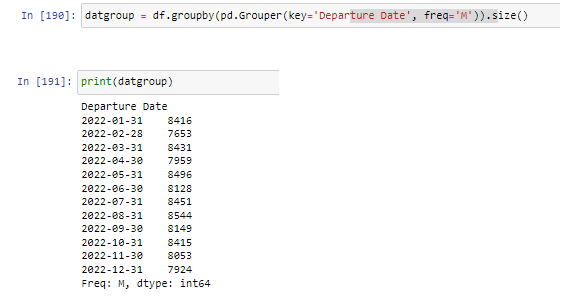
Find out the range of the departure date using **.max()** and **.min()** function (Figure 14).

Figure 14 Max and Min function



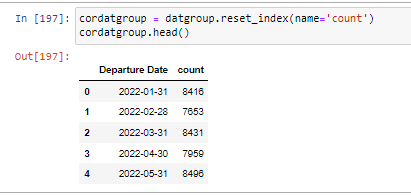
Group by month to see trends (Figure 15).

Figure 15 Grouping by month



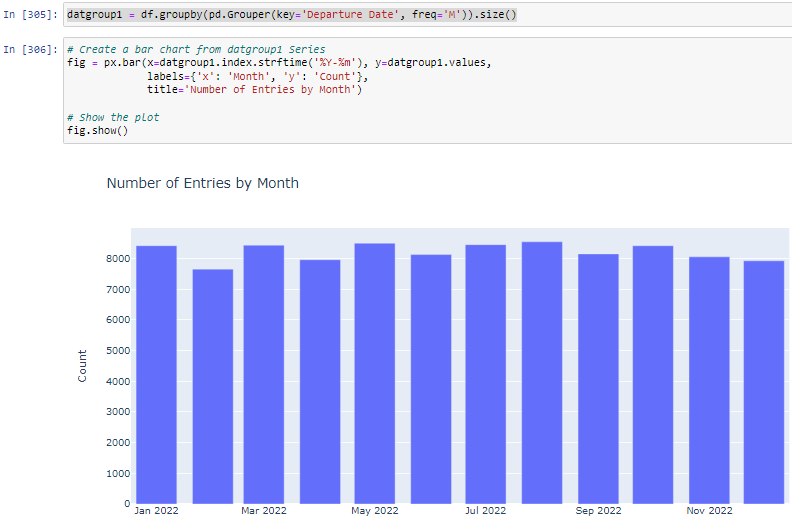
Use **reset\_index()** to set it back to the default integer index (Figure 16).

Figure 16 Resetting the index



I can also omit the **reset\_index()** function, use the Departure date as index and count as values, and create a bar chart and observe the count per month using plotly.express (Figure 17).

Figure 17 Using plotly.express to make a Bar chart



# BIG DATA ANALYTICS

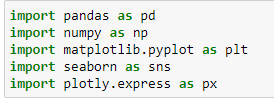
## Critically evaluate a range of algorithms and big data analytics techniques suitable for analysing the complex dataset

## Select and justify the most appropriate methods for extracting meaningful information from the large-scale data

## Implement and apply the selected algorithms using appropriate tools or programming languages

I have used the following libraries to manipulate my dataset (Figure 18)

Figure 18 Importing libraries



Pandas was used to convert the CSV file into data frames (Figure 19), which structured the data in a tabular format, making it easier to manipulate. This enabled me to perform powerful data manipulation techniques and analysis, including filtering, grouping and aggregating. Furthermore, it integrates well with matplotlib, seaborn and plotly.express, which are used in data visualisation. Additionally, Numpy is a library used in numerical computing.

Figure 19 Converting CSV to Data Frame



Using SQL queries, I have used **.groupby()**, **.value\_counts()**, **.sort\_values()** and **.agg()** to sort and index the values based on order. The sorting algorithm used will depend on the type DBMS system. Additionally, the most common would be a variation of merge sort and quick sort (Figure 20 – 23).

Figure 20 illustrates the total unique values in Gender which are then separated to Male and Female. It is useful in identifying the distribution of each value.

Figure 20 Count unique values in Gender

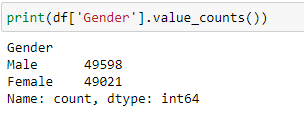


Figure 21 uses the .**groupby()** and **.agg()** function to handle 2 or more columns. It allows us to group values according to specification and counts the unique values under it.

Figure 21 Using group by and aggregate function

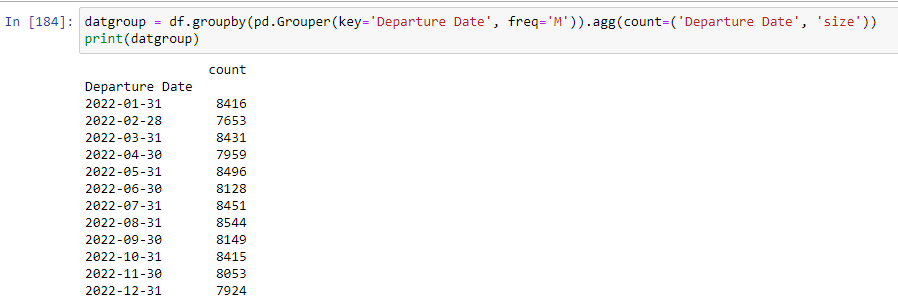


Figure 22 uses **.sort\_values()** to sort data in ascending or descending order. It helps visualise the values from highest or lowest.

Figure 22 Sorting values in descending order

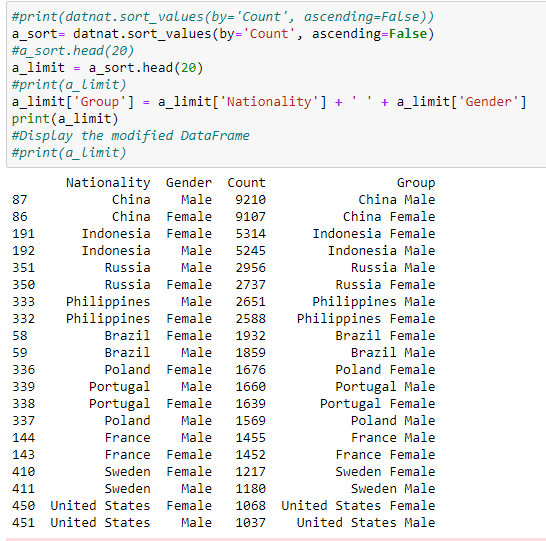
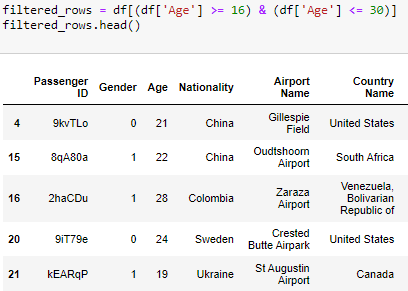


Figure 23 uses filtering to improve visualisation of a targeted subset of data.

Figure 23 Filter data frame and show Ages 16 to 30



# DATA VISUALISATION FOR BUSINESS DECISION MAKING

## Apply data visualisation techniques to present the extracted insights in a clear, informative, and visually appealing manner

## Create appropriate visualisations that cater to the needs of the company executive board and support effective decision-making

## Interpret the visualisations, identify key trends, patterns, and actionable insights, and effectively communicate them to a non-technical audience

### Task 1: What is the total proportion of passengers according to gender?

Figure 24 depicts a pie chart with the distribution of male and female passengers as whole. It can be seen that there is a slightly higher proportion of male travellers compared to the female.

Figure 24 Pie chart of Male and Female passengers



### Task 2: What is the distribution of passengers per month? Which has the highest and lowest number of passengers?

Figure 25 – 26 shows slight fluctuations in the number of passengers per month, with minimal differences. August has the highest count while February has the lowest.

Figure 25 Horizontal bar graph distribution of passengers per month

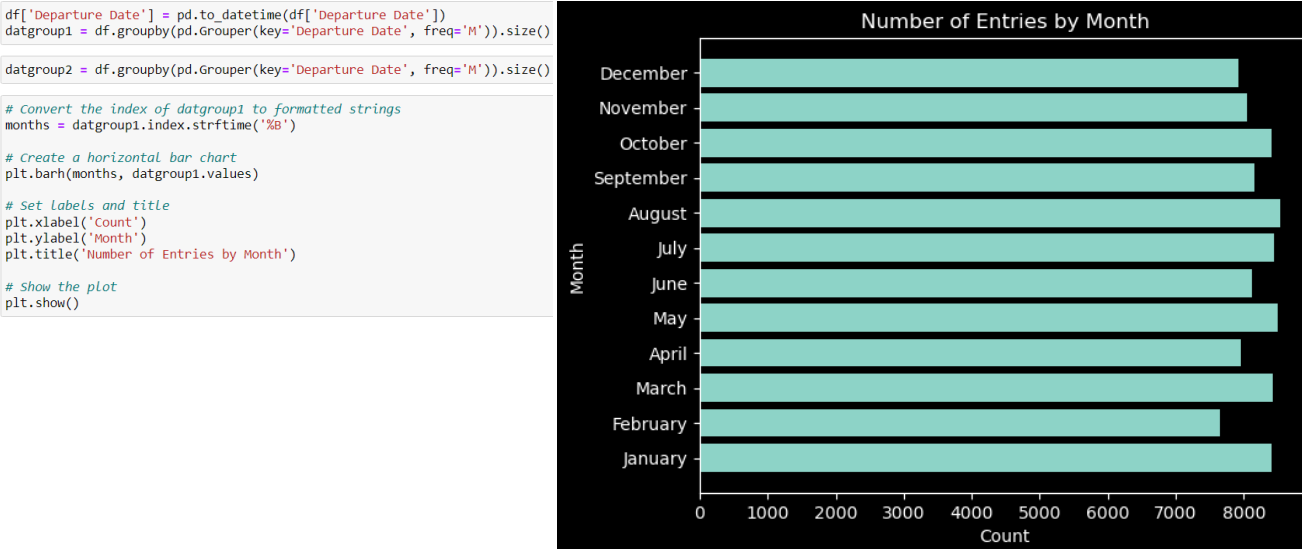
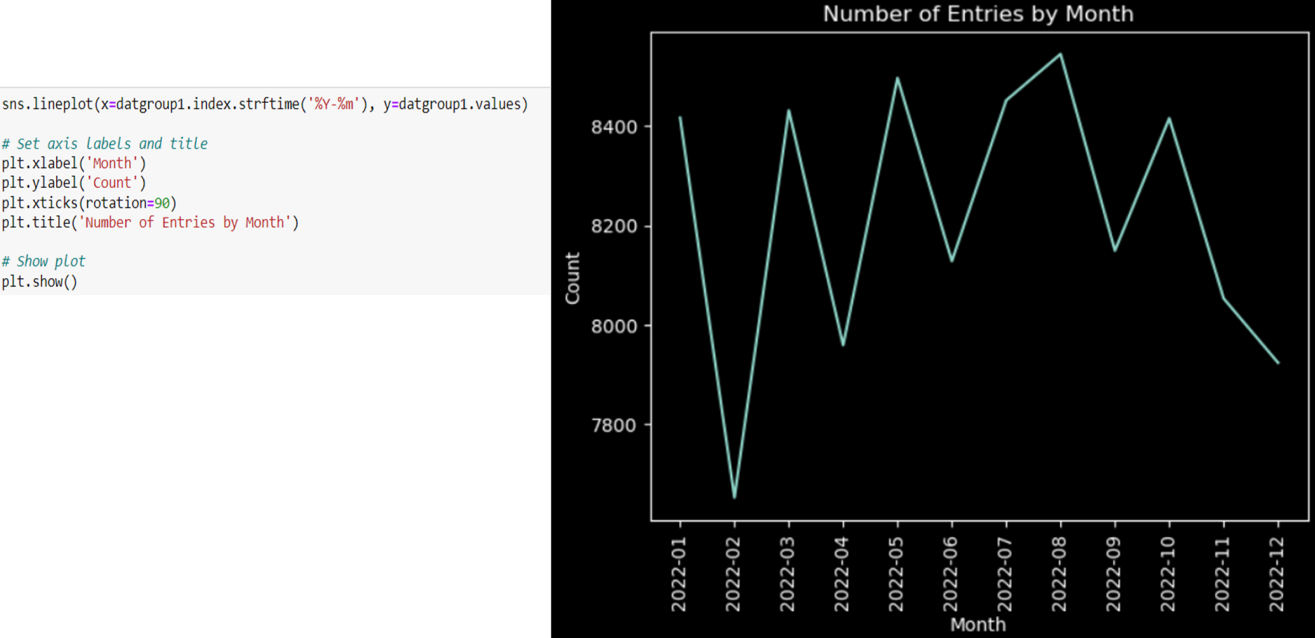


Figure 26 Line graph of distribution of passengers per month



### Task 3: Which nationality has the highest number of travellers by gender?

Figure 27 – 29 shows that among the top 10 nationalities, Chinese men and women travel the most.

Figure 27 Code for sorting the highest number of passengers by Gender and Nationality

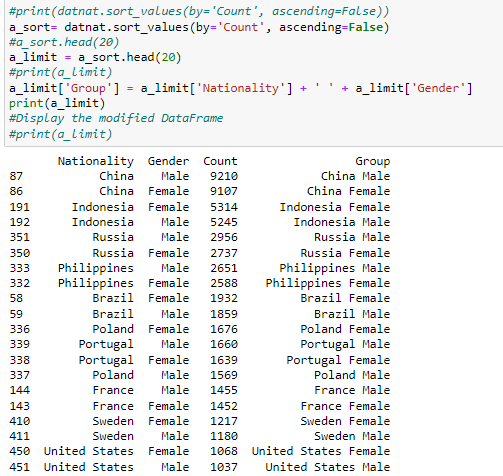


Figure 28 Heatmap code

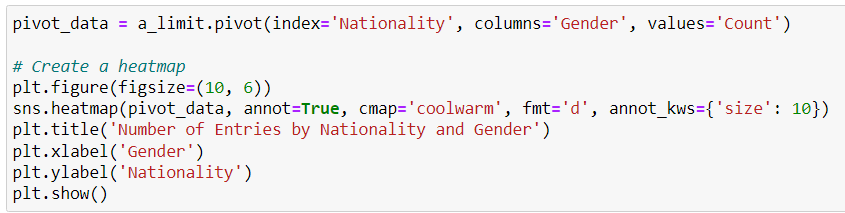


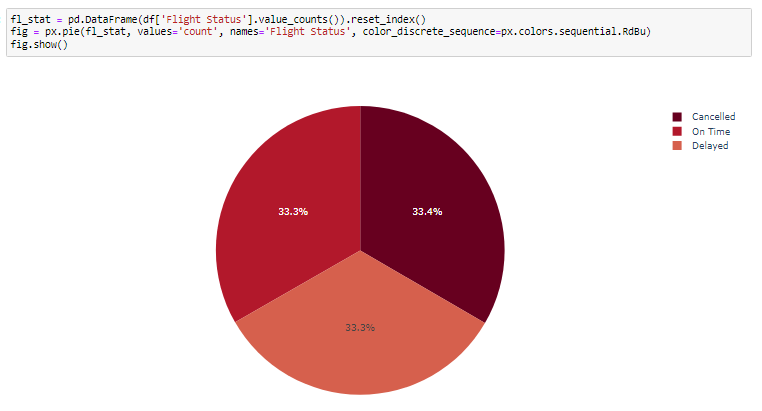
Figure 29 Heatmap by Nationality and Gender



### Task 4: What is the percentage of **On time flights** compared to **Cancelled** and **Delayed**?

There is an equal distribution of flights that are on time, delayed and cancelled (Figure 30).

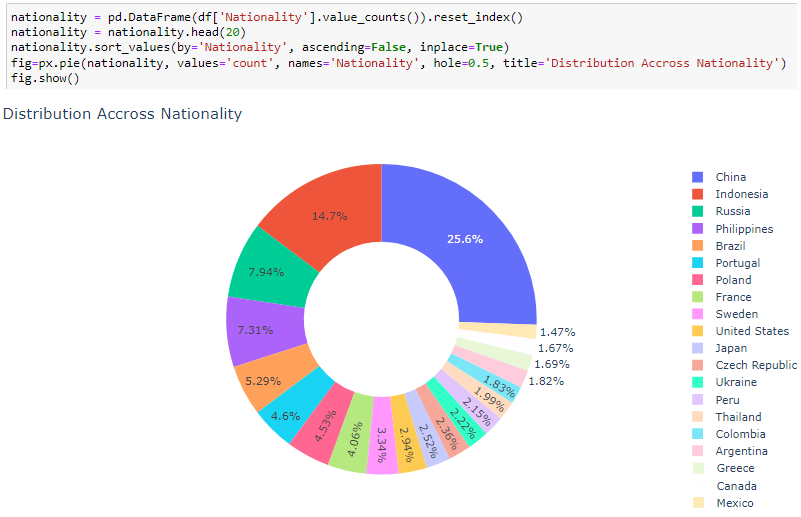
Figure 30 Percentage of Cancelled, On time and Delayed flights



### Task 5: What is the distribution of passengers across various countries

The pie chart shows the distribution of top 20 passengers per Nationality. Figure 31 shows, that the Chinese account for 25.6% of the total passengers, followed by Indians at 14.7%. Conversely, Mexicans represents the lowest with 1.47%.

Figure 31 Pie chart distribution of top 20 passengers from various countries



### Task 6: What is the age trend of passengers? Which country do most young adults travel to?

Figure 32 shows a significant drop of passengers at the ages of 30 and 60, in contrast to other age groups which have minimal differences amongst themselves.

Figure 32 Histogram of Age frequency among passengers

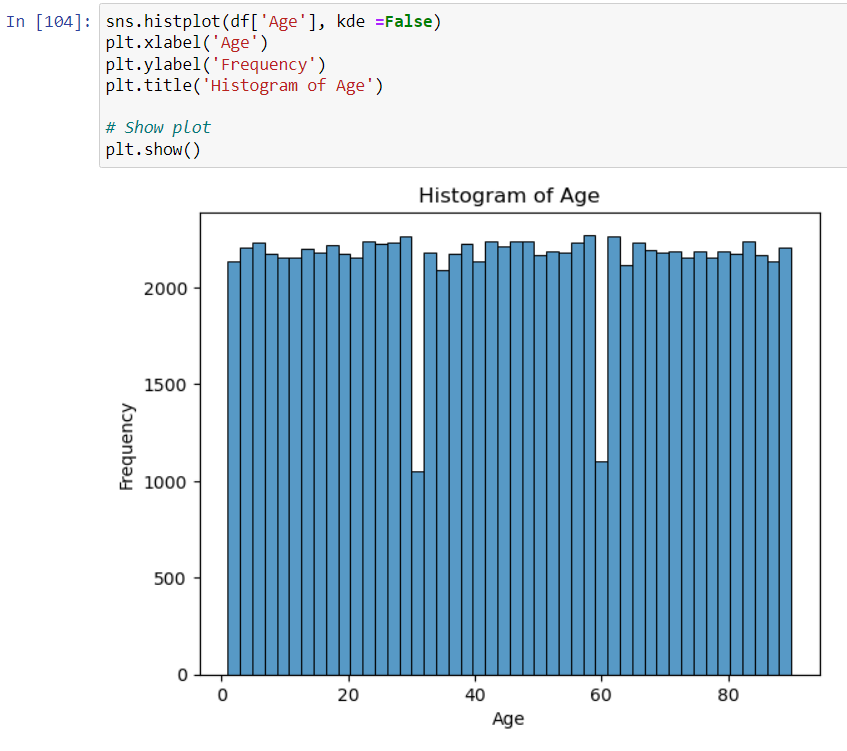
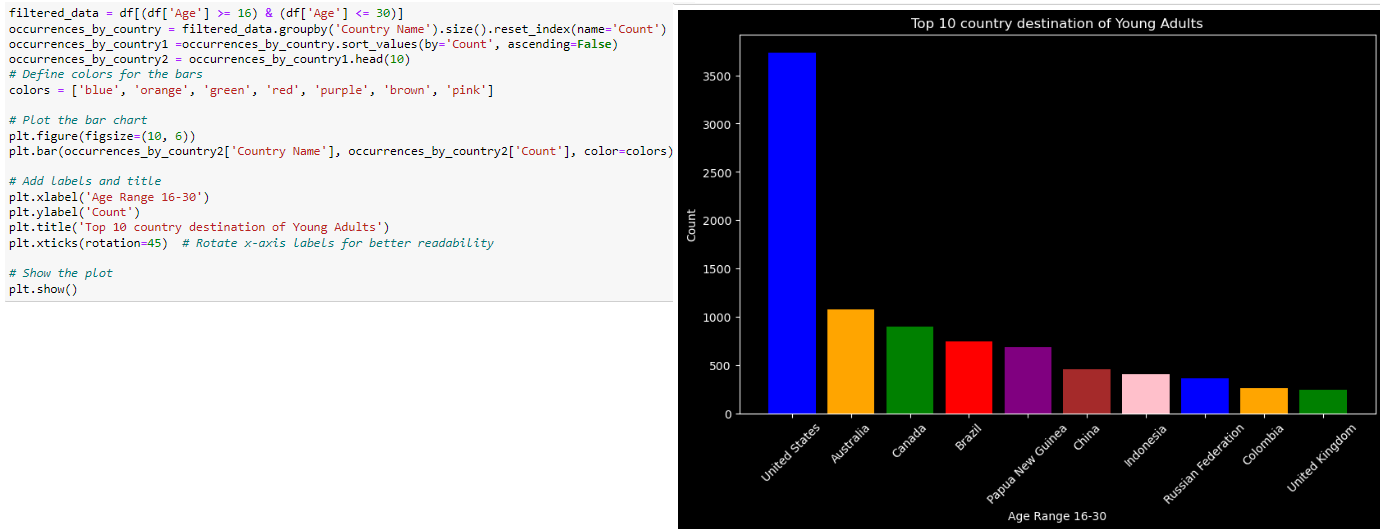


Figure 33 depicts that USA has the highest number of visits by young adults compared to other countries.

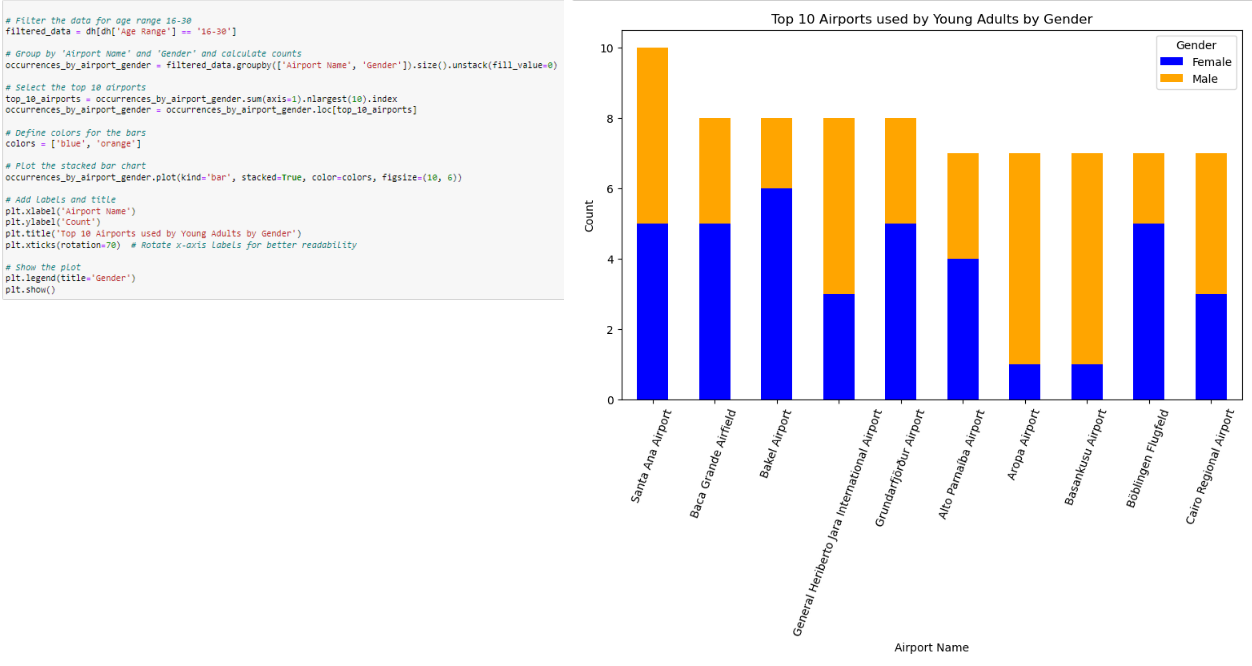
Figure 33 Bar graph of the Top 10 countries visited by young adults 16-30



### Task 7: What airport do certain young adults most frequent?

Figure 34 shows that Santa Ana airport is the most frequently used by young adults in the top 10 category. Furthermore, Basankusu and Aropa airports have more male young adults using it.

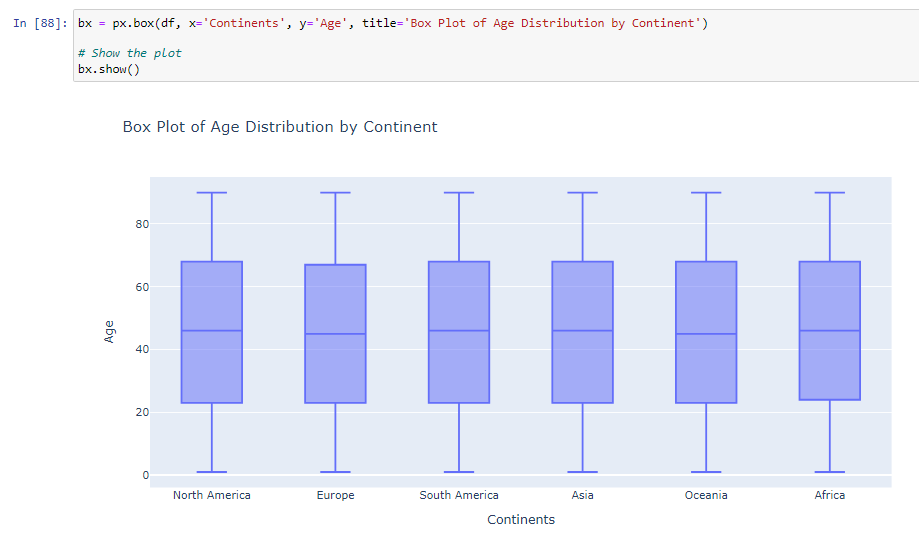
Figure 34 Bar graph of the top 10 airports used by young adults 16 - 30



### Task 8: What is the minimum, median and maximum age of passenger per continent?

Figure 35 reveals that the age of passengers across continents are closely comparable, with the mean age ranging from 45 to 46.

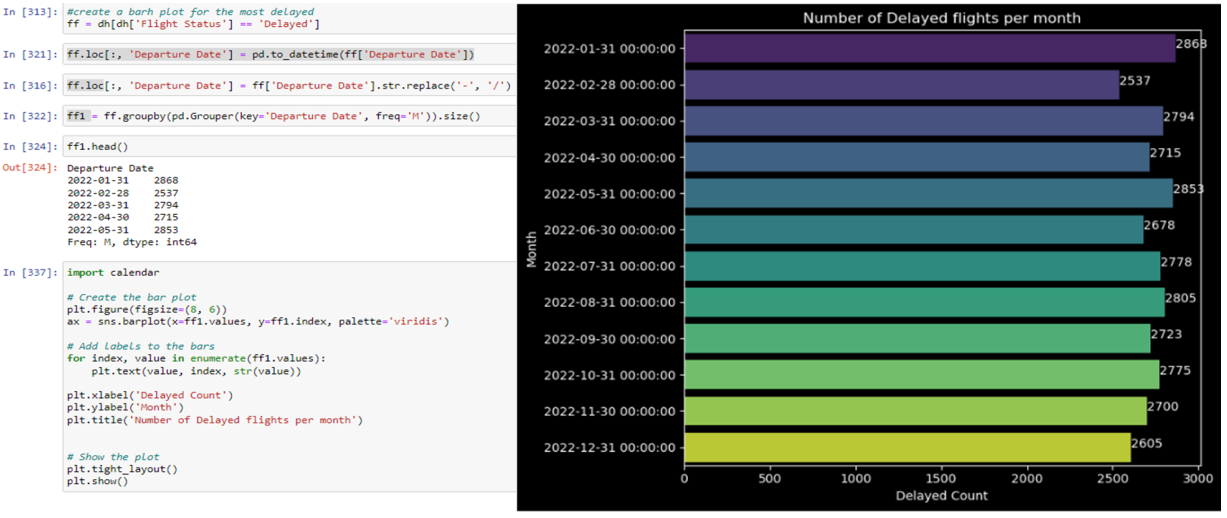
Figure 35 Box and whiskers chart for Age distribution among Continents



### Task 9: Which month has the most delayed flights?

Figure 36 displays that January has the highest number of delayed flights, closely followed by May. In contrast, February has the least number.

Figure 36 Horizontal bar chart of delayed flights



### Task 10 Using chi-square is there any corelation between passenger gender and continent?

Figure 37 illustrates the application of chi-square statistics. With a P-value is 0.1187, we reject the null hypotheses that there is no corelation between gender and origin continent for departure. A heatmap is then produced showing that North America has the highest number of flights for both male and female, as origin continent.

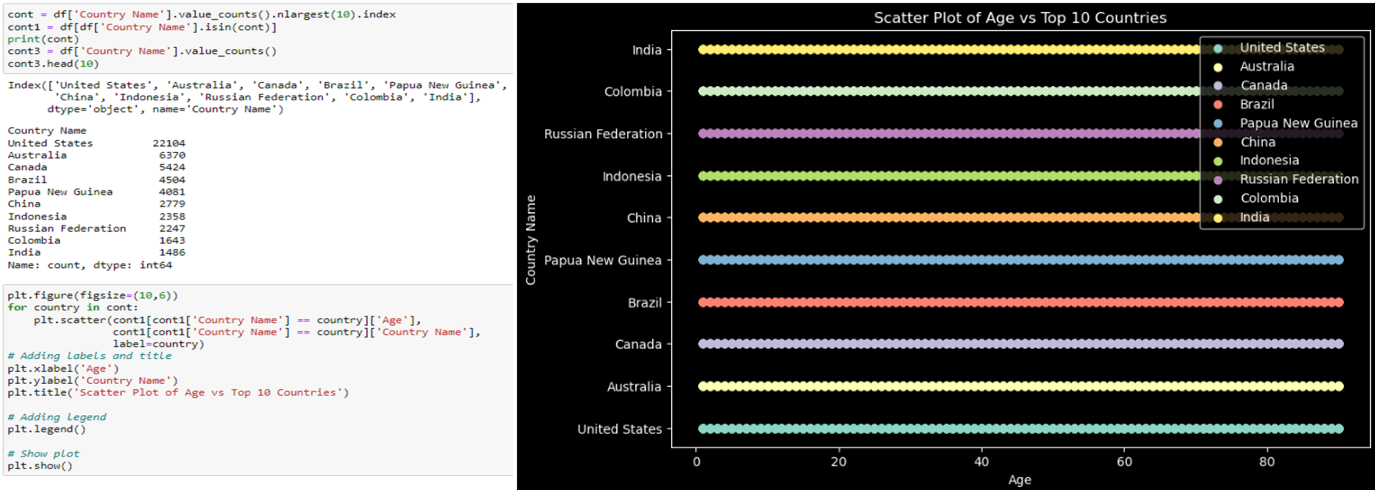
Figure 37 Chi-square computation and Gender to continent heatmap



### Task 11: Is there a corelation between age and country?

The scatter plot shows a uniform distribution of passengers across all age categories for each country, suggesting no corelation (Figure 38).

Figure 38 Scatter plot of Age to Country



# ETHICAL AND LEGAL CONSIDERATION

## Address and discuss ethical and legal issues related to the handling of big data and the presentation of findings

Big data and its analytics raise several ethical concerns including, consent, identity, trust and security. Privacy and data protection are among the most prominent worries. Individuals should be able to control what organisations can collect and have access to in their aggregated data as these data can reveal sensitive information about them (Stahl & Wright, 2018). Moreover, poor security can lead to data leak of personal information. This data can also be edited or deleted by third-party traffickers, which was the case with an airline company in collecting data (Chang, 2019)

.

Additionally, discrimination and inequality can also be an issues as algorithmic bias can cause privileging of one arbitrary group over the other.

In addition, despite multiple authors providing a standard for data visualisation, misinformation from bad visualisation exists. Studies have shown that size, colour and shape are the most common pit fall, with the pie and bar chart most commonly being misrepresented (Nguyen et al., 2021).

## Ensure compliance with data privacy regulations and other relevant legal considerations

To maintain legal compliance, it is important to adhere to the Data protection act 2018. It is the implementation of the General Data Protection Regulation (GDPR) in the UK which regulates how organisations and businesses can use personal data. (Government Digital Service, 2015). Ensuring compliance includes being transparent about data processing, secure user’s privacy rights, and implement data security measures.

## Propose strategies to mitigate potential ethical challenges in the use of data for business decision-making purposes

To mitigate ethical issues, compliance to the data protection act should be ensured. This entails respecting personal data, allowing transparency in data processing and allowing individuals to correct or delete wrong information about themselves. Additionally, algorithms in decision-making should be monitored on how they impact people as they can be viable to bias.

# REFERENCES

*Airline Dataset*. (2023, September 26). Kaggle. [https://www.kaggle.com/datasets/iamsouravbanerjee/airline- dataset/code?datasetId=3682609&sortBy=voteCount](https://www.kaggle.com/datasets/iamsouravbanerjee/airline-%09dataset/code?datasetId=3682609&sortBy=voteCount)

Chang, V. (2019, May 2). *Privacy and ethical issues of big data in the airline industry*. Teesside University’s Research Portal. <https://research.tees.ac.uk/en/publications/privacy-and-> ethical-issues-of-big-data-in-the-airline-industry

Government Digital Service. (2015, September 16). *Data protection*. GOV.UK. https://www.gov.uk/data-protection

Nguyen, V. T., Jung, K., & Gupta, V. (2021). Examining data visualization pitfalls in scientific publications. *Visual Computing for Industry, Biomedicine, and Art*, *4*(1). https://doi.org/10.1186/s42492-021-00092-y

Stahl, B. C., & Wright, D. (2018). Ethics and Privacy in AI and big Data: Implementing responsible research and innovation. *IEEE Security & Privacy*, *16*(3), 26– 33. <https://doi.org/10.1109/msp.2018.2701164>

# Appendix

## Using WordCloud

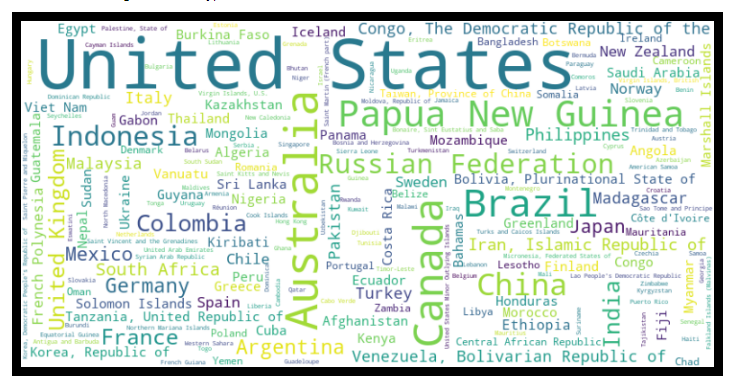
Top country of departure using WordCloud.

Figure 39 WordCloud code



Figure 40 shows that the United States is the top country of departure for passengers in the data set.

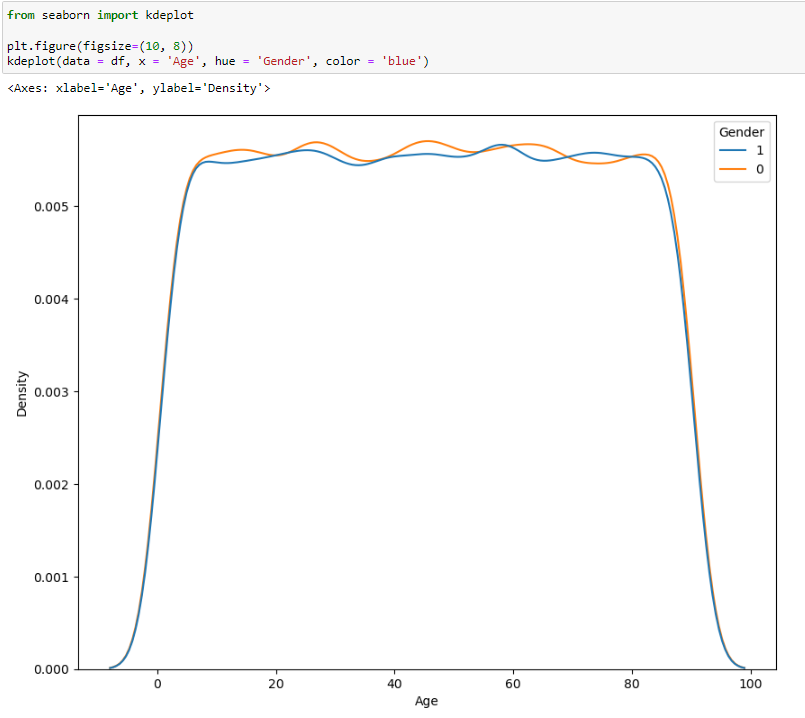
Figure 40 WordCloud



## Using KDE plot

The KDE plot reveals that there is minimal variance in age and that the data is relatively evenly distributed (Figure 41).

Figure 41 KDE plot of passenger age



## Using Map visuaisation

Figure 42 and 43 shows the code and map visualisation of the top 10 airports used by all passengers.

Figure 42 Geopy code for Map display

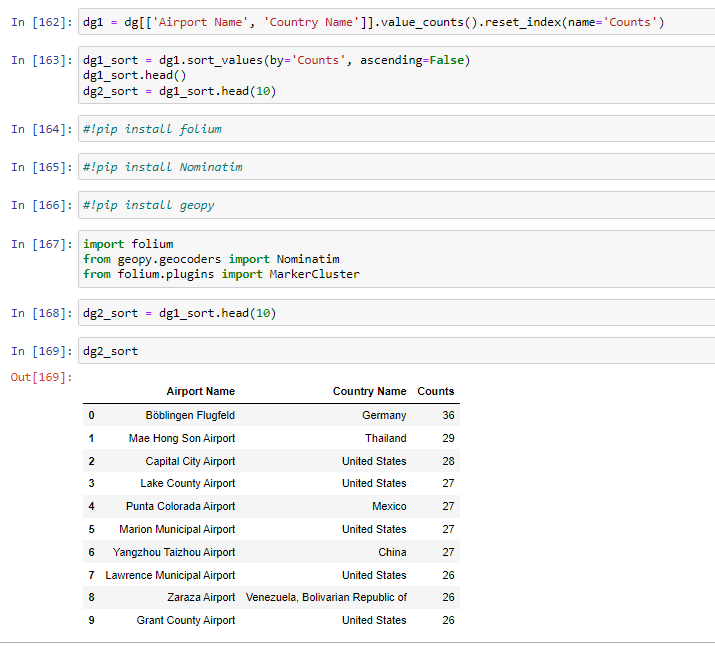
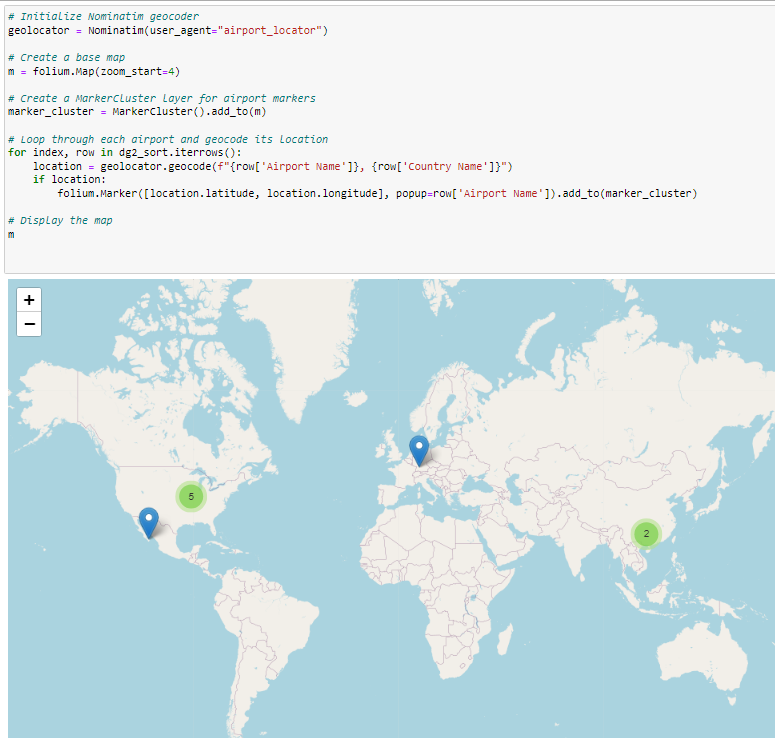


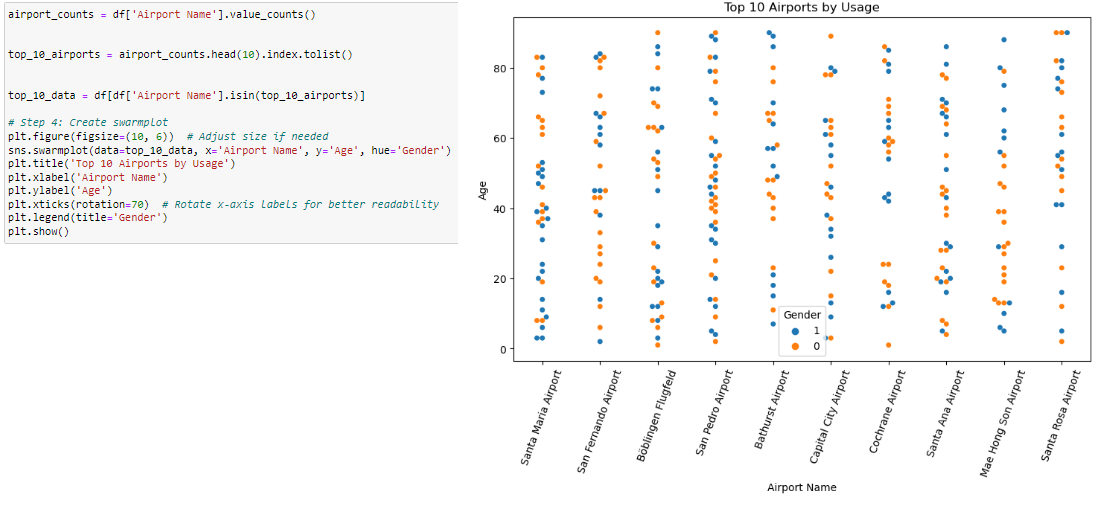
Figure 43 Map visualisation of top 10 most used airports



## Using swarm plot to visualise the top 10 most used airports

Figure 44 shows that there is a high number of clusters in San Pedro and San Fernando Airport. Additionally, the clusters are within the 40 to 50 age range,

Figure 44 Swarm plot of the top 10 most used airports



## Using stacked bar chart to find out which continent has the highest number of passengers

Figure 45 illustrates that North America has the highest number of passengers reaching over 30000, in contrast with South America with approximately 10000 passengers recorded. Furthermore, the distribution of passengers appears relatively even across the months.

Figure 45 Horizontal bar chart of continents and their passenger count

