

1. Introduction

Travel has become an integral part of our lives in today's fast-paced world, providing opportunities for exploration, recreation, and business. In our role as travelers, we often find ourselves confronted with the daunting task of choosing a flight that aligns with our preferences, whether it is budget considerations, travel duration, or airline selection. This project aims to provide travelers with a comprehensive analysis of flight data in order to help them navigate an enormous array of flight options efficiently.

2. Data collection

2.1 Destination and Date Selection

The data collection process began with the selection of the city and date for the study. The destination city in this case was Dhaka, the capital city of Bangladesh, and the data collection date was March 15, 2024. As a result of the project's requirement for a dataset with a minimum of 50 flights to be able to conduct robust analysis, Dhaka was selected as the destination because it is relevant to the project and is expected to have a sufficient number of flights. Furthermore, this is my country of origin.

2.2 Web Scraping Sources

For the purposes of acquiring flight data, three reputable online travel booking websites were selected: Expedia.com, ebookers.fi, and booking.com. The sources selected were deemed appropriate due to their established presence in the online travel booking domain, their diverse range of flight options, and their mix of global and regional platforms. A comprehensive representation of flight information was achieved by combining these sources.

2.3 Data Attributes

The data collection process focused on retrieving specific attributes that were deemed essential for subsequent analysis and user interaction. The following attributes were targeted:

1. **Departure time:** The time at which a flight is scheduled to depart from Helsinki.
2. **Arrival time:** The time at which the same flight is scheduled to arrive in Dhaka.
3. **Number of stops:** An indication of whether the flight is direct or involves one or more stops.
4. **Total layover time:** The cumulative duration of all layovers in the itinerary.
5. **Individual layover information:** Detailed information about each layover, including the city and duration.
6. **Airlines:** The airline carrier(s) offering flights on the specified route.
7. **Total trip duration:** The total duration of the flight from departure to arrival, encompassing both in-flight and layover times.

3. Web scraping process

3.1 Initial configuration

- **Selenium in headless mode:** The web scraping process commenced with the configuration of Selenium, a powerful web automation tool, to facilitate data retrieval from online travel booking websites. To ensure efficient and unobtrusive data collection, Selenium was employed in headless mode. In this mode, a web browser interface is not displayed, allowing for automated interaction with web pages while minimizing resource consumption.
- **Chromium and chromedriver:** Chromium, an open-source web browser project, served as the underlying browser engine for Selenium. To establish communication and control with Chromium, the Chromedriver executable was utilized. Chromedriver acts as an intermediary, facilitating Selenium's commands to manipulate and navigate web pages within the Chromium environment.

3.2 Interaction with booking website

3.2.1 Interaction with the first page

In configuring the web scraping script for data collection, several key decisions and actions were taken to meet the project's requirements effectively:

- **One-way direction selection:** The script was set to collect data exclusively for one-way flights, specifically from Helsinki to Dhaka. This choice narrowed the scope of data collection to the specified route.
- **Departure and arrival places:** The script dynamically filled in the departure and arrival locations as Helsinki and Dhaka, respectively, ensuring that the collected data pertained to the intended travel route.
- **Date selection:** The script automated the selection of the travel date as March 15, 2024, ensuring that the data gathered was for flights available on that specific day.
- **Action button click:** The script was configured to programmatically trigger the "Search" or "Submit" button on the booking website, initiating the flight search process.

These actions were essential in streamlining the data collection process to align with the project's objectives. All fields are shown in the fig-1.

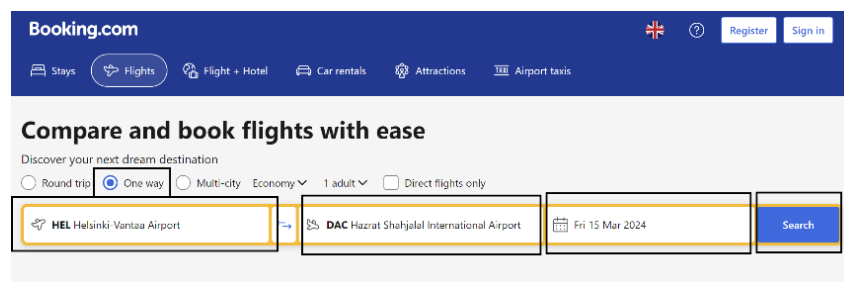


Fig-1: Interaction with the first page

3.2.2 Interaction with the flight information page

- **Arrival and departure times:** Precise arrival and departure times were directly extracted from the websites.
- **Airlines:** The names of the operating airlines were systematically identified.
- **Total trip duration:** The overall journey duration, encompassing in-flight and layover times, was collected.
- **Total stops:** The number of stops or layovers for each flight option was obtained.

In Figure 2, all data locations are indicated by square boxes

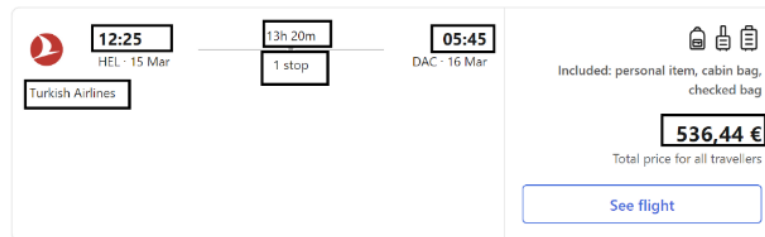


Fig – 2: Flight information page

3.2.3 Differential data extraction for booking.com

Booking.com required an additional step to extract detailed layover information.

- **Layover information on booking.com:** Selenium navigated to the "Flight Details" section for each flight on Booking.com and collected individual layover data, including layover city and duration. This step, specific to Booking.com, provided comprehensive layover insights but extended the scraping time.

In Figure 3, all layover data locations are indicated by square boxes.

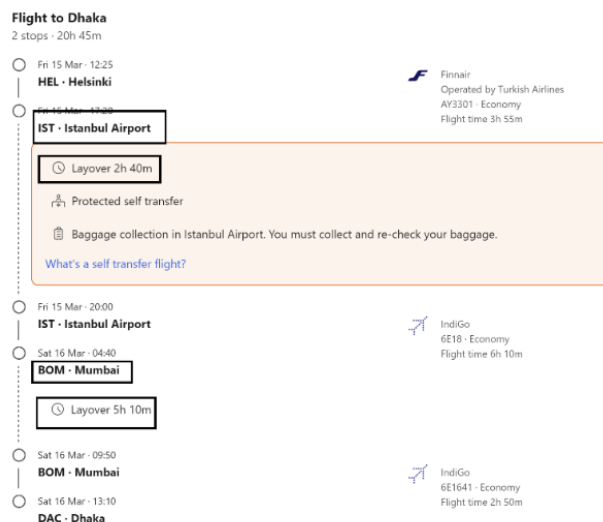


Fig – 3: Layover information page

4. Data processing

During the data preprocessing stage, the collected flight data underwent several transformations to enhance its usability and consistency for subsequent analysis. The key data preprocessing steps are summarized below:

Time format standardization

- **Arrival and departure Times:** The arrival and departure times were reformatted from a combined "HH:mm-HH:mm" format to individual HH:mm format for clarity and ease of computation. Furthermore, time values were converted into a unified representation in hours, with minutes represented as decimal fractions (e.g., 12 hours and 30 minutes as 12.5 hours).

Currency and price formatting

- **Currency sign removal:** Any currency symbols (e.g., Euro or Dollar signs) present in the price data were removed. This ensured uniformity in the representation of flight prices.
- **Space and comma removal:** Extraneous spaces and commas in the price data were eliminated. The result was a clean, numeric representation suitable for further analysis in float data type.

Layover time aggregation

- **Total layover time:** Raw layover information, such as "10h 20m Doha" and "5h 40m Istanbul," was processed to calculate the total layover time. The individual layover times were aggregated, resulting in a single value representing the cumulative layover duration. For instance, "10h 20m Doha, 5h 40m Istanbul" was consolidated into a total layover time of 16.0 hours.
- **Layover place extraction:** In addition to total layover time, the layover places (e.g., Doha and Istanbul) were extracted. This information provided insight into the cities where layovers occurred.

5. Data Storage

After completing the web scraping and data processing phases, the collected flight data was carefully organized and cleaned to ensure its readiness for analysis and user interaction. Following the meticulous preprocessing steps, the structured dataset was saved in a CSV (Comma-Separated Values) file format. This strategic decision to store the data in a CSV format serves as a pivotal juncture in the data pipeline, offering a portable and easily accessible representation of the flight information. The CSV file format facilitates seamless integration with a variety of data analysis tools and frameworks, providing the foundation for in-depth exploratory data analysis, user interaction, and the fulfillment of the project's objectives. Fig 4 is showing the data table.

	Booking website	airlines	departure_time	arrival_time	price(euro)	total_duration(hour)	total_stops_int	total_layover_time(hour)	layover_info
0	https://www.ebookers.fi/en/	Turkish Airlines	05:05	12:55	558.00	13.17	1	1.83	1h 50m in Istanbul (IST)
1	https://www.ebookers.fi/en/	Turkish Airlines	05:05	19:40	558.00	30.42	1	19.17	19h 10m in Istanbul (IST)
2	https://www.ebookers.fi/en/	Multiple airlines	08:40	08:00	629.00	21.67	2	6.75	5h 20m in London (LHR) 1h 25m in Dubai (DXB)
3	https://www.ebookers.fi/en/	Multiple airlines	17:20	16:00	629.00	22.33	2	7.58	4h 40m in London (LHR) 2h 55m in Dubai (DXB)
4	https://www.ebookers.fi/en/	Multiple airlines	17:20	14:05	629.00	24.25	2	9.50	6h 35m in London (LHR) 2h 55m in Dubai (DXB)
...
522	https://flights.booking.com/	Singapore Airlines	06:40	22:40	931.19	37.00	2	18.00	4h 15m - FRA Å Frankfurt International Apt 13...
523	https://flights.booking.com/	Qatar Airways	16:05	18:15	777.14	47.17	2	17.92	13h 45m - FRA Å Frankfurt International Apt 4...
524	https://flights.booking.com/	Qatar Airways	19:50	02:05	780.53	51.25	2	37.58	19h 25m - OSL Å Oslo Airport 18h 10m - DOH Å...
525	https://flights.booking.com/	Qatar Airways	19:50	03:35	780.53	52.75	2	39.08	19h 25m - OSL Å Oslo Airport 19h 40m - DOH Å...
526	https://flights.booking.com/	Qatar Airways	16:05	02:05	780.53	55.00	2	41.33	23h 10m - OSL Å Oslo Airport 18h 10m - DOH Å...

527 rows x 9 columns

Fig-4. Data table

6. Data Visualization

6.1 Data Visualization tools

- **Matplotlib:** Matplotlib is a widely used library for creating static, animated, and interactive visualizations in Python. It provides a flexible and comprehensive set of plotting functions and is highly customizable. You can create a wide range of plots, including line plots, bar charts, scatter plots, histograms, and more, using Matplotlib.
- **Seaborn:** Seaborn is built on top of Matplotlib and provides a higher-level interface for creating attractive and informative statistical graphics. It simplifies the process of creating complex visualizations by providing easy-to-use functions for tasks like creating heatmaps, violin plots, pair plots, and more. Seaborn also offers a range of built-in themes and color palettes to make your visualizations visually appealing.

6.2 Finding data insights

- **Airlines distribution (Fig-5):** The provided figure illustrates the distribution of airlines, presenting various unique airlines along with their respective counts. Additionally, the percentage representation of each airline is thoughtfully annotated within the visualization. *Insights: Finnair stands out as the airline with the largest number of flights, boasting a total of 107 flights (Out of 527), constituting 20.3% of the entire flight dataset.*

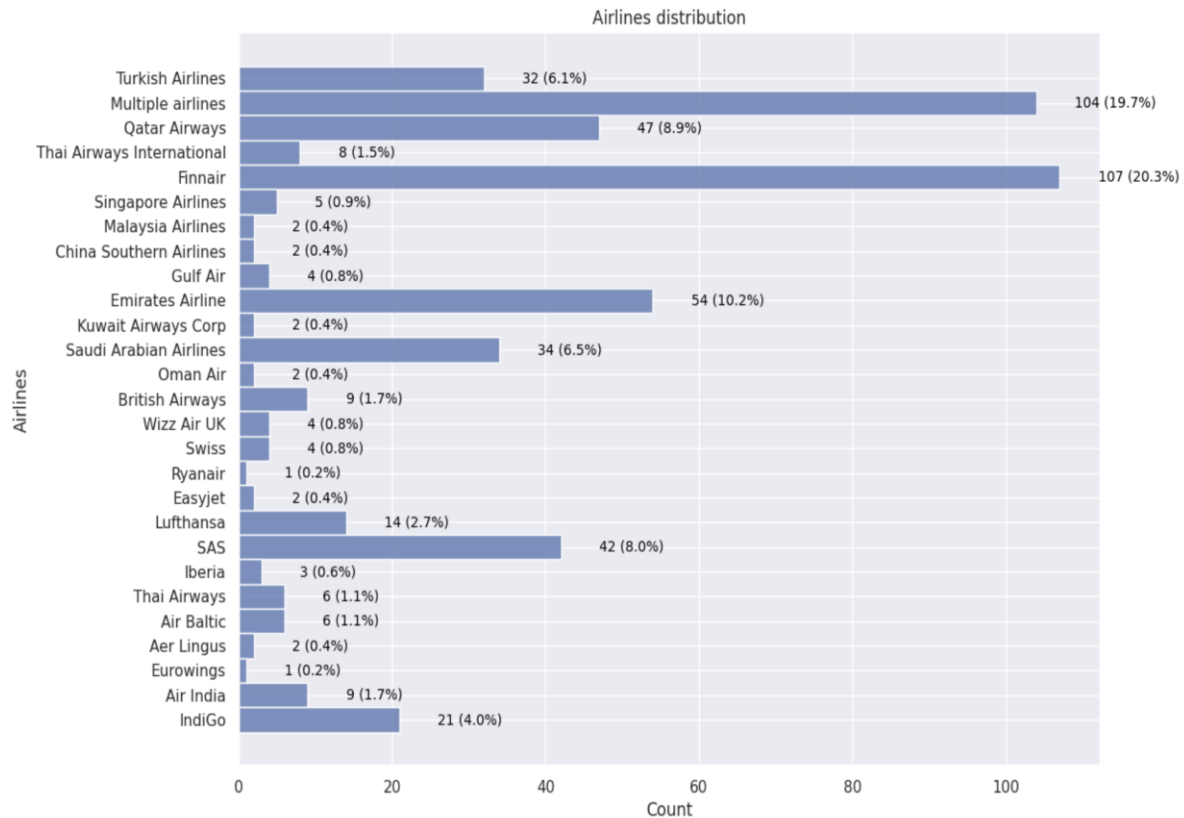


Fig - 5: Airlines distribution

- **Booking website distribution (Fig-6):** The below figure illustrates the distribution of booking websites, presenting flight availability in different websites.

Insights: Booking.com leads the pack by offering the highest number of flight options, comprising an impressive 65.1% of the total, while two other websites contribute equally, each accounting for 17.5% of the available flight choices.

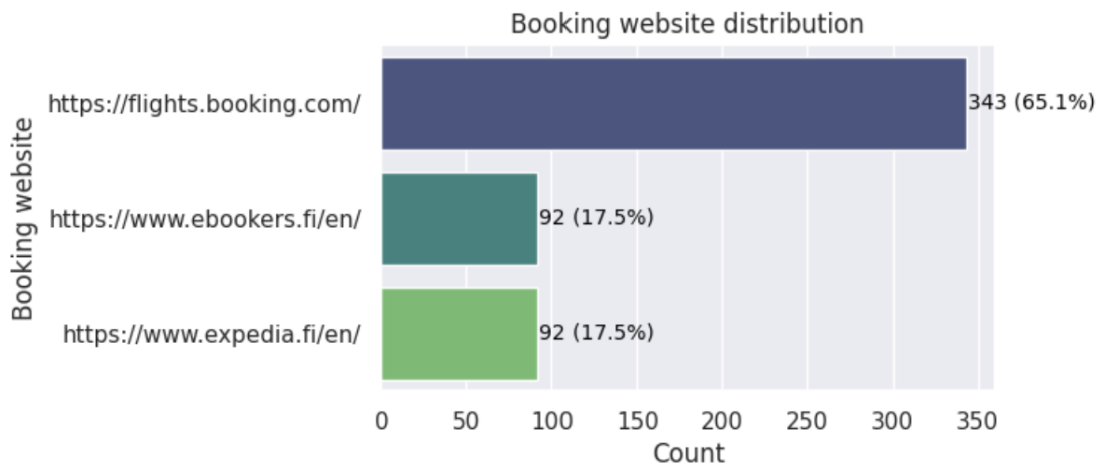


Fig - 6: Booking website distribution.

- **Departure time distribution (Fig-7):** The departure time distribution plot is like a chart that helps us understand when most flights take off on a given day. However, instead of looking at the exact time down to the minutes, we simplify things by only considering the hour when flights depart. For example, if one flight leaves at 12:00 PM and another at 12:15 PM, we treat both as if they left at 12:00 PM. This simplification helps us see the bigger picture of when flights are most likely to take off during the day. It's a way of looking at trends in departure times without getting lost in the small details of specific minutes.

***Insights:** The departure pattern analysis indicates that a significant proportion of flights exhibit a preference for morning departures, occurring between 6:00 AM and 10:00 AM, or alternatively, afternoon departures within the time frame of 4:00 PM to 5:00 PM. Of particular interest is the hour of 7:00 AM, which emerges as the peak departure time. Notably, the dataset encompasses more than 60 flights scheduled to depart from Helsinki during the specified day.*

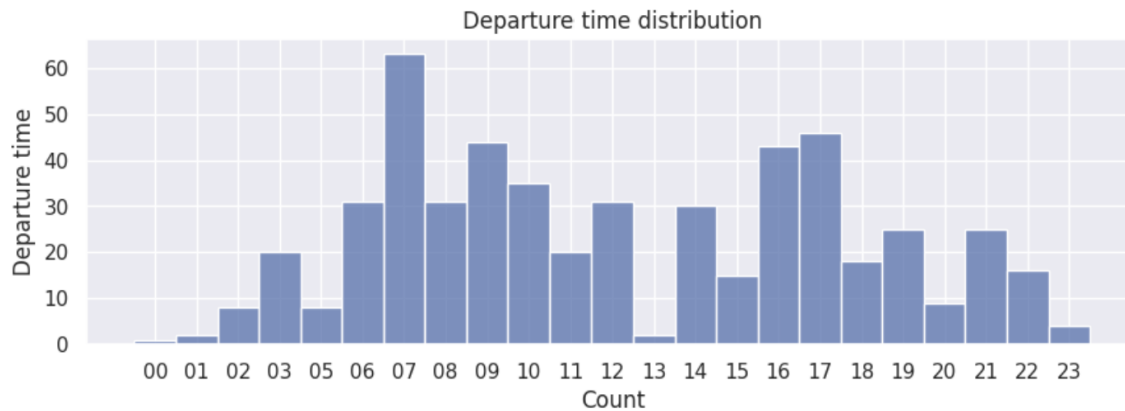


Fig - 7: Departure time distribution

- **Arrival time distribution (Fig-8):** The arrival time distribution plot is like a chart that helps us understand when most flights reach the destination. For normalizing the graph, I used same method as departure time.

***Insights:** The analysis reveals that a substantial portion of flights tend to arrive at their destination during the morning hours, specifically between 7:00 AM and 9:00 AM, as well as in the evening, spanning the time frame of 5:00 PM to 6:00 PM.*

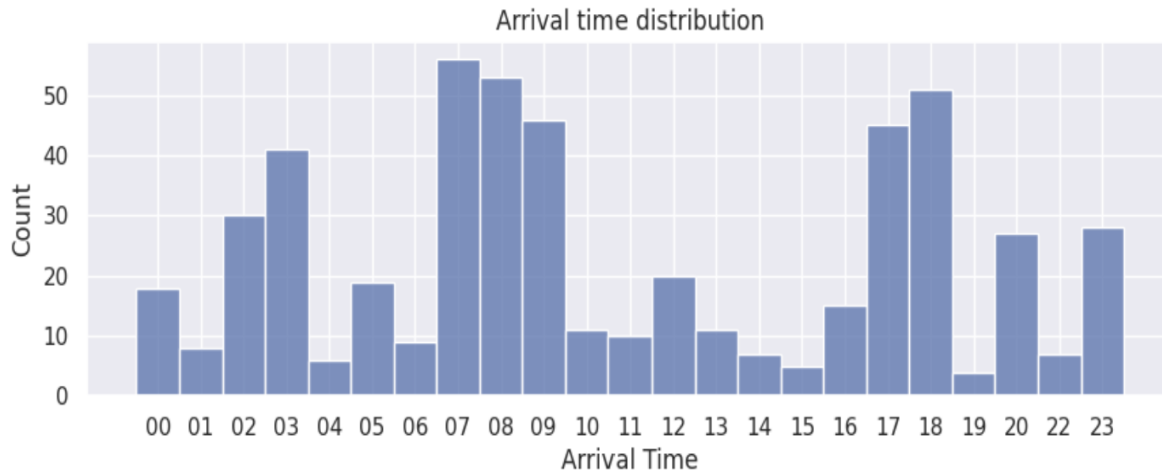


Fig - 8: Arrival time distribution

- **Price distribution (Fig-9):** The price distribution visualization serves as a valuable tool for gaining insights into the predominant price range observed among flights.

Insights: A notable observation within the dataset is that the majority of flight prices cluster within the range of 600 to 900 euros. Additionally, it is noteworthy that in proximity to the 600-euro mark, the dataset exhibits the highest concentration of flight offerings.



Fig - 9: Price distribution

- **Total trip duration distribution (Fig-10):** Total trip duration distribution graph help us to understand on an average how much time will be needed to reach the destination.

Insights: On average, the total trip duration falls within the range of 15 to 30 hours. Notably, the graphical representation illustrates a discernible peak in flight durations

slightly exceeding 21 hours, with approximately 104 flights falling within this specific timeframe. This observation underscores the distribution of trip durations within the dataset.

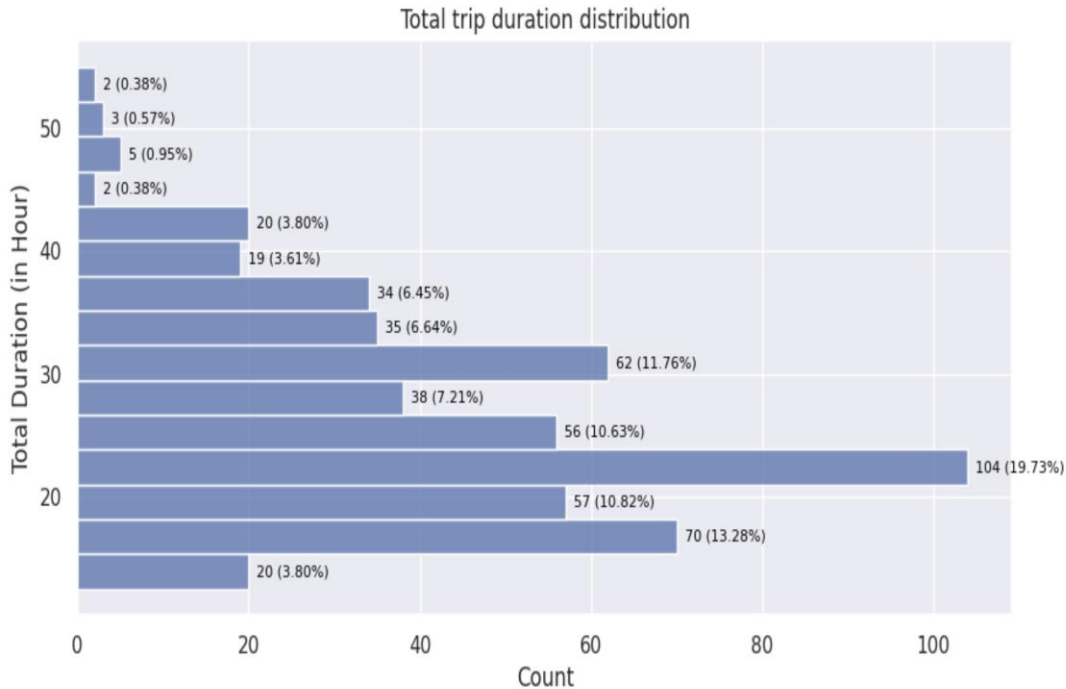


Fig - 10: Total trip duration distribution

- **No of stops distribution (Fig-11):** No of stops distribution graph help us to understand on an average how many stops will be needed to reach the destination.

***Insights:** Approximately 70% of the total flights, specifically 365 out of the 527 flights analyzed, necessitate two stops along their respective routes. It is worth noting that there are no direct flights available on this route.*

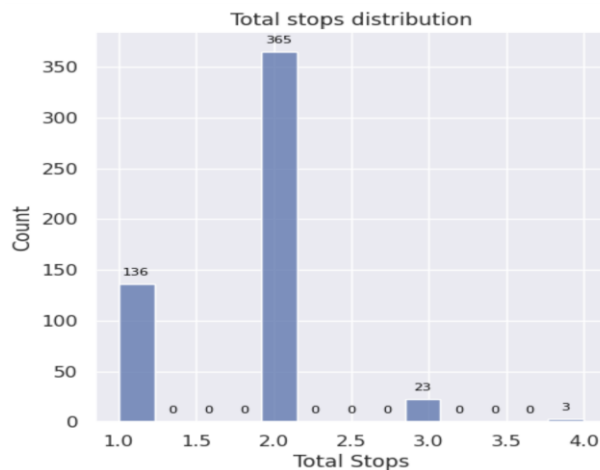


Fig - 11: Total no of stops distribution.

- **Total layover time distribution (Fig-12):** Total layover time distribution give us a comprehensive understanding of layover time in the trip.

***Insights:** In aggregate, roughly 50% of the flights in question require layover times ranging from 5 to 10 hours, representing a significant portion of the dataset. Furthermore, a minority fraction of flights exhibits layover times exceeding 30 hours, indicating the presence of extended layover durations within the dataset.*

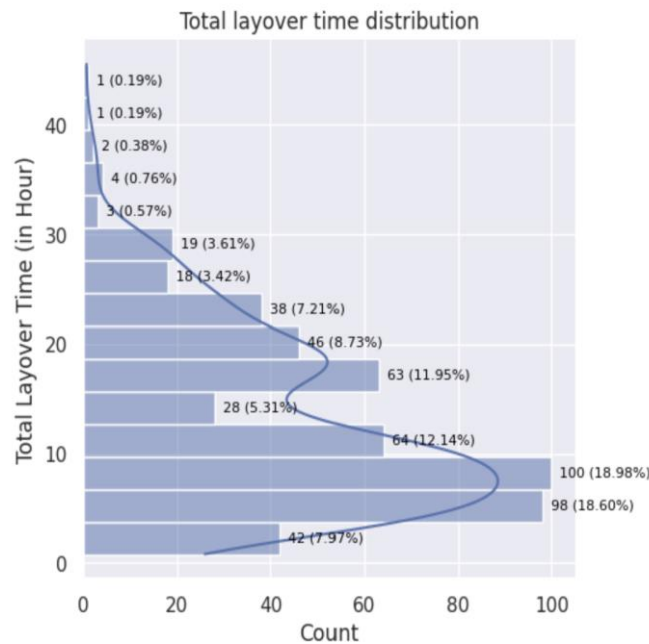


Fig - 12: Total layover time distribution

- **Heatmap (Fig-13):** Using the heatmap, we can determine whether there is any correlation among the different variables in the dataset.

***Insights:** A very good correlation exists only between the total trip duration and the total layover time, which is obvious given that layover time constitutes a subset of the total trip duration.*

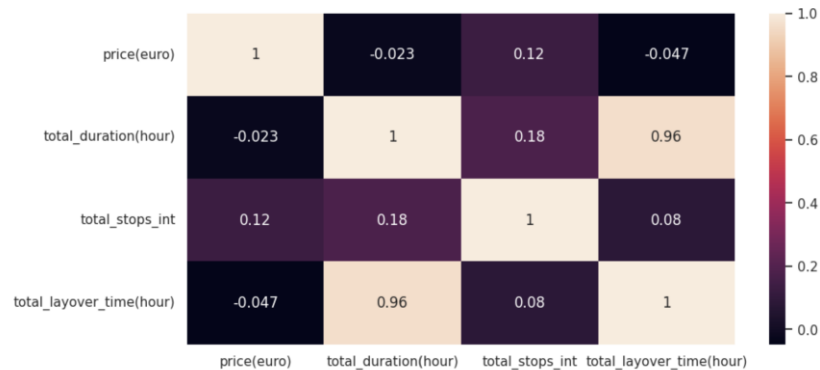


Fig - 13: Heatmap

- **Scatterplot 1 (Fig-14):** I am visualizing a scatterplot between trip duration and layover duration based on the heat map we have seen.

Insights: It is likely that the layover duration will be longer if the trip duration is longer.

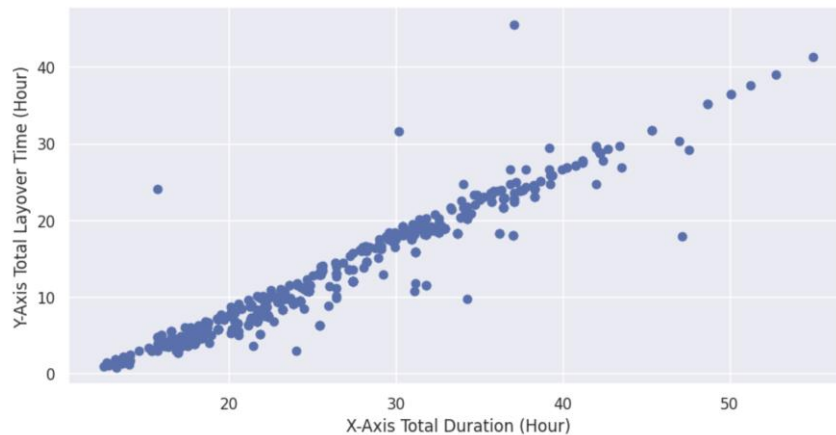


Fig - 14: Scatterplot 1

- **Scatterplot 2 (Fig-15):** I am visualizing a scatterplot between total stops and total duration, and total stops and layover duration to see the trend.

Insights: There exists a limited probability that an increase in trip duration and layover time may correspond to a higher number of stops. However, it is essential to note that there is no substantial or definitive correlation observed between trip duration, layover time, and the number of stops in flights.

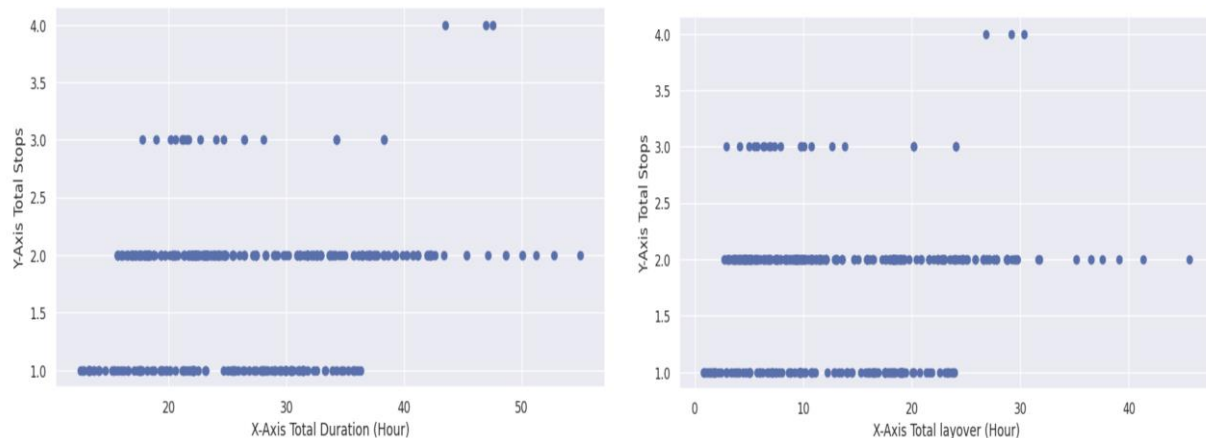


Fig - 15: Scatterplot 2

- **Bar graph between airlines and price (Fig-16):** This bar graph is employed for the purpose of elucidating the average pricing trends associated with different airlines. The annotations within the visualization serve the purpose of providing both the count and percentage representation of flights associated with various airlines.

Insights: The graphical representation yields several noteworthy observations regarding airline pricing. Air India emerges as a provider of competitively priced flights, with an average fare of 471.03 Euros. However, it is important to note that Air India represents a relatively modest share of the flight options, accounting for only 1.71% of the total. In contrast, China Southern Airlines commands the highest average price, with fares averaging 3741.00 Euros, albeit offering only two flights on the given day. Remarkably, Finnair boasts the highest share of flight options on this route, encompassing approximately 20.30% of all available flights. These flights are associated with an average price of 817.06 Euros.

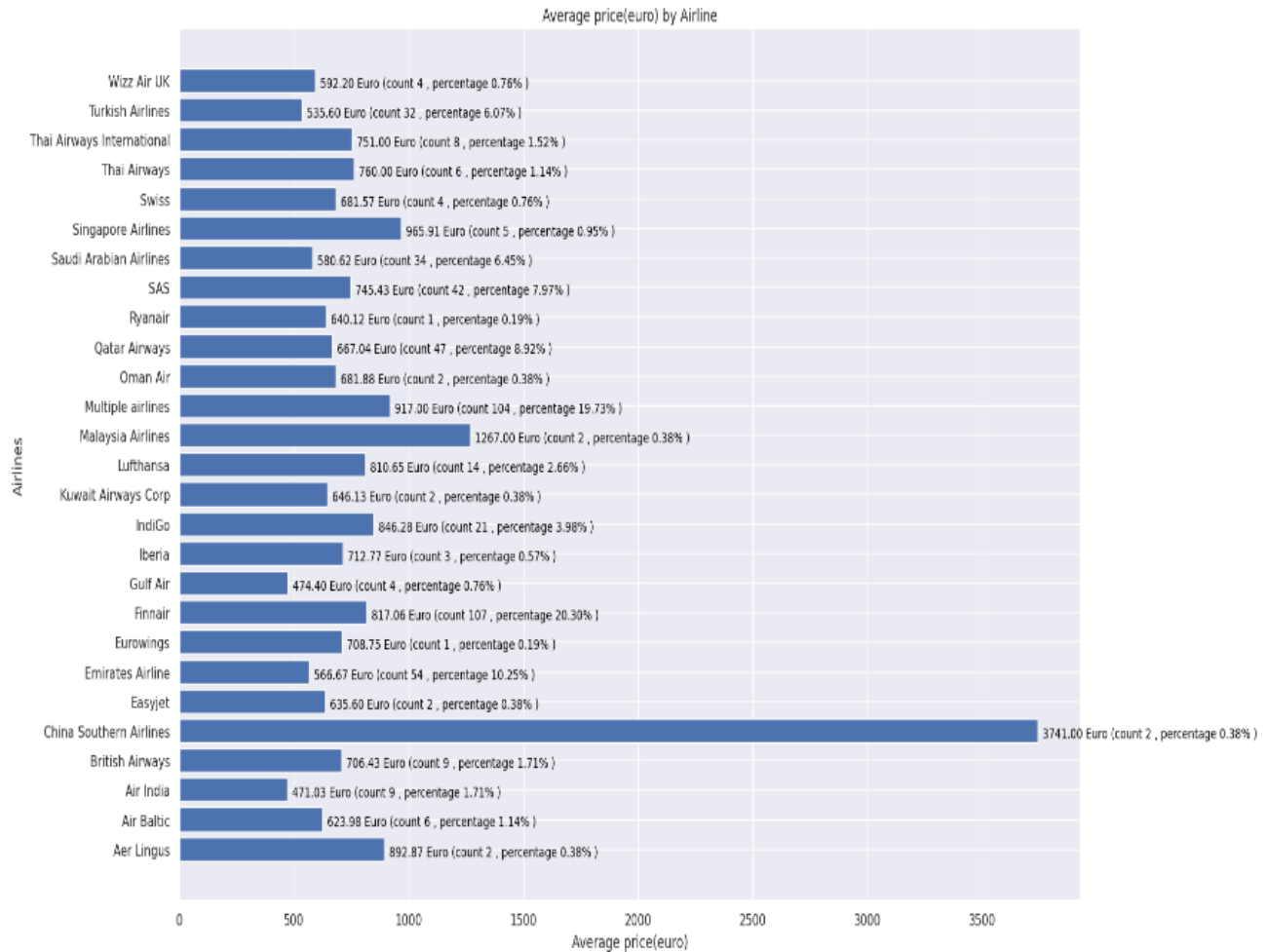


Fig - 16: Bar graph between airlines and price

- **Bar graph between airlines and trip duration (Fig-17):** This bar graph is employed for the purpose of elucidating the average trip duration associated with different airlines. The annotations within the visualization serve the purpose of providing both the count and percentage representation of flights associated with various airlines.

Insights: Air Lingus stands out with the shortest average total trip duration, requiring only 18.00 hours for its flights. However, it's important to note that Air Lingus offers a limited selection, with only two flight options available on that particular day. Conversely, Air India exhibits the longest average total trip duration, with an average of 42.28 hours, despite concurrently offering the most economically priced flights.

Finnair, with the highest frequency of flights on this route, possesses an average total trip duration of 28.54 hours for its most frequently operated flights. These findings offer valuable insights into the comparative temporal aspects of flight offerings by these airlines.

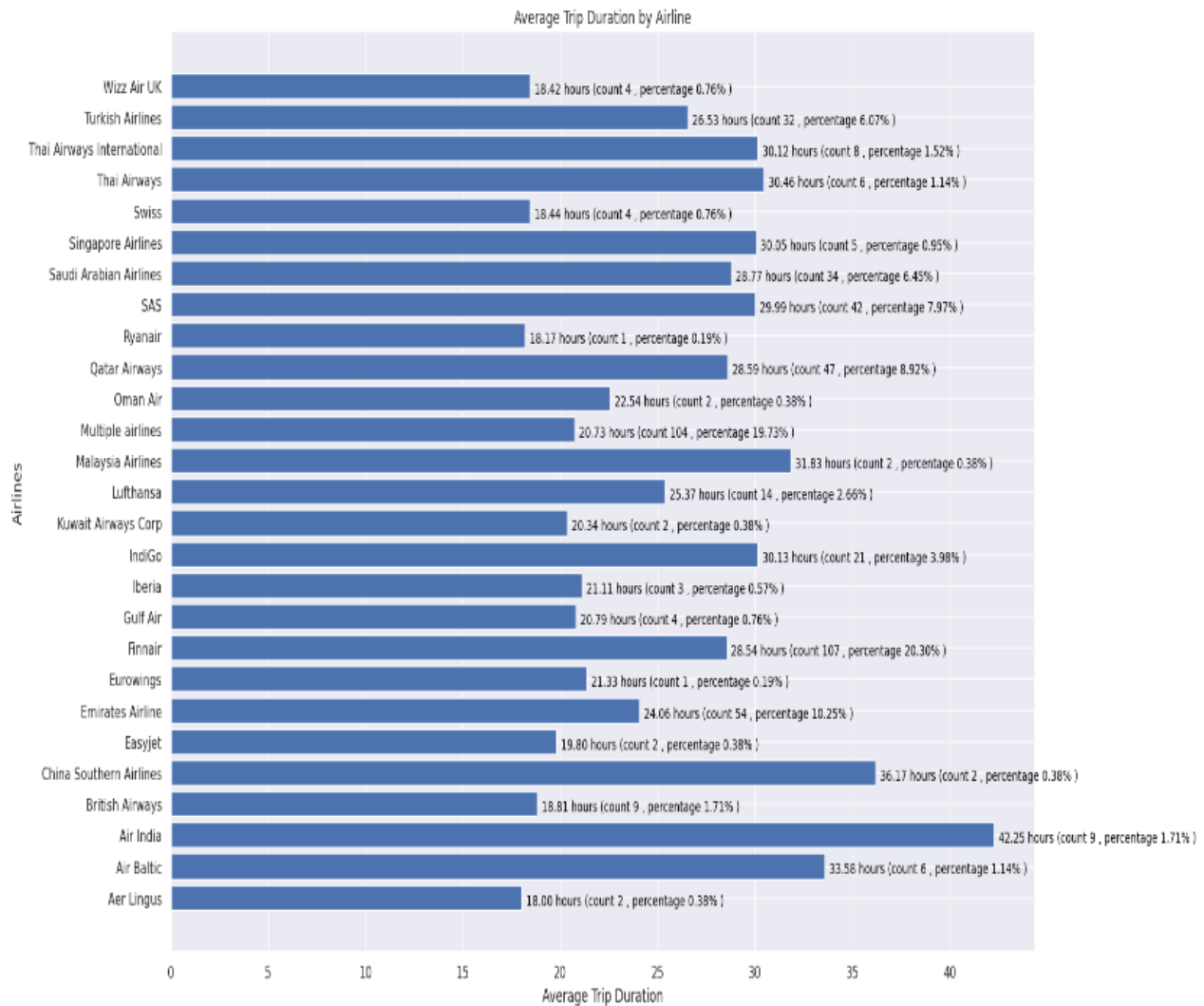


Fig - 17: Bar graph between airlines and trip duration

- **Bar graph between airlines and layover duration (Fig-18):** This bar graph illustrates the average layover duration associated with various airlines. In the visualization, annotations provide both a count and percentage representation of flights associated with different airlines.

Insights: Like before, Air Lingus has the shortest average total layover duration, requiring only 4.38 hours for its flights. Air India, however, exhibits the longest average total layover duration, with an average of 28.79 hours, despite simultaneously offering the most economically priced routes. The most frequently operated flights by Finnair occupy a total layover duration of 28.54 hours, with the highest frequency of flights on this route.

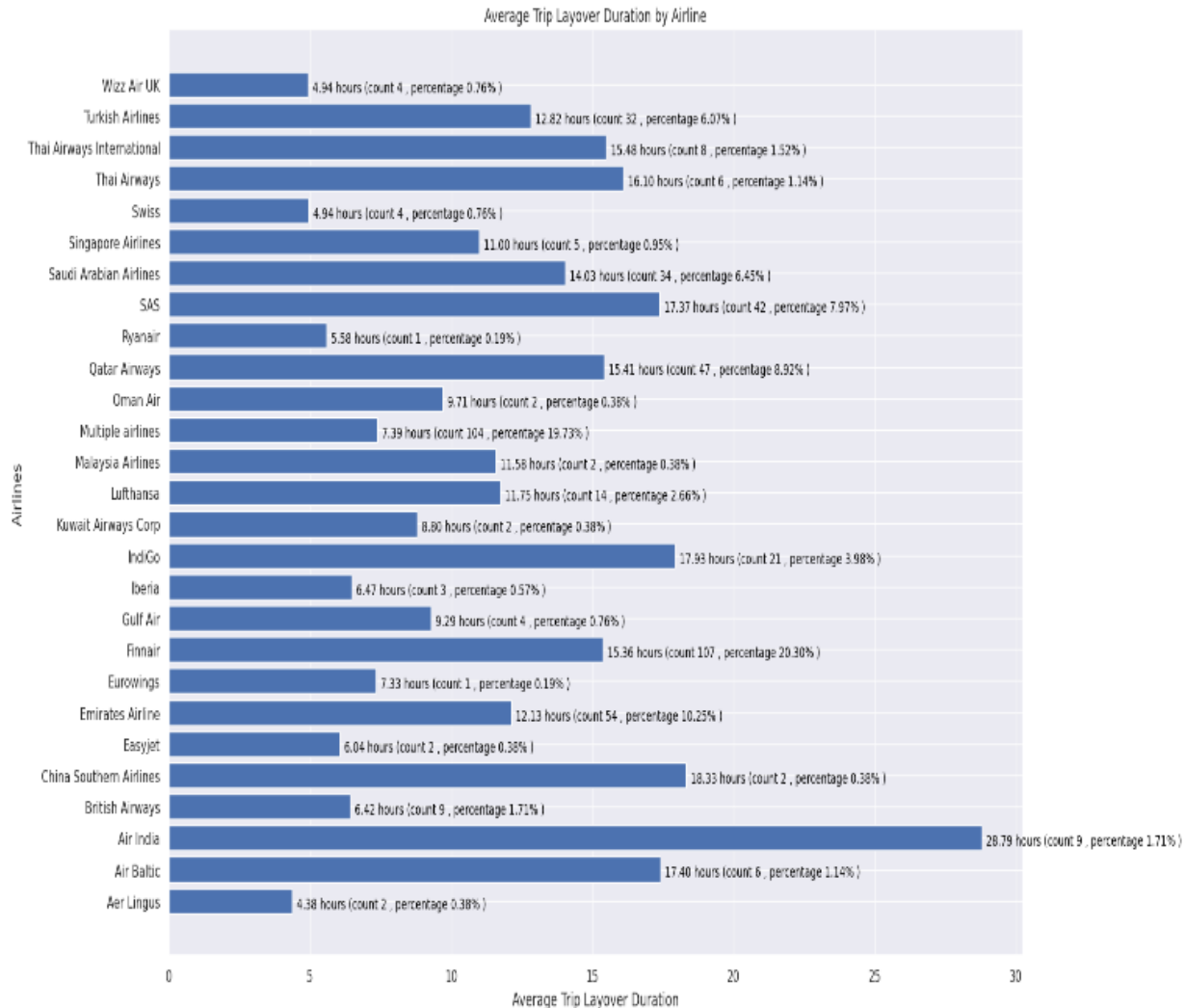


Fig - 18: Bar graph between airlines and layover duration

- **Bar graph between airlines and no of stops (Fig-19):** According to this bar graph, the average number of stops for various airlines can be seen. It is possible to view the statistics for flights associated with different airlines using annotations that present both a count and a percentage.

Insights: The graphical representation affords an observation that several airlines, including Omar Air, Kuwait Airways Crop, Gulf Air, and British Airways, exhibit the lowest average number of stops, which is 1. Conversely, Air Baltic stands out with the highest average number of stops, totaling 3.5. Furthermore, Finnair, the airline with the highest

flight frequency, necessitates an average of 1.94 stops for its flights. These insights provide valuable information concerning the distribution of stopover patterns among various airlines within the dataset.

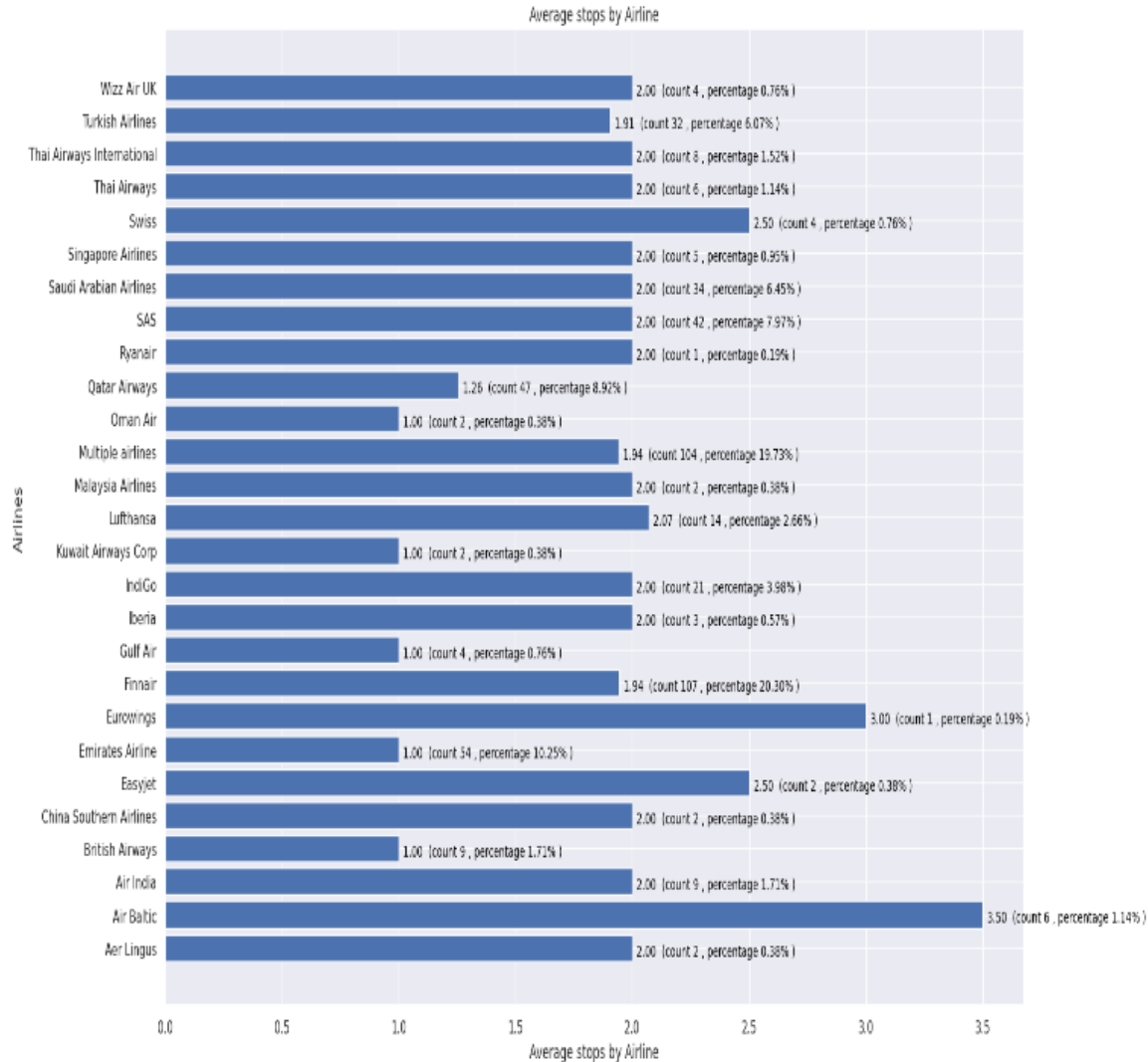


Fig - 19: Bar graph between airlines and No of Stops

7. Conclusion:

In the course of executing this project, I encountered several challenges and obstacles that needed to be addressed in order to meet the project's requirements. Below, I outline these challenges and provide insights into the strategies employed to overcome them successfully.

Dynamic Website Navigation: Initially, the web scraping process encountered difficulties with BeautifulSoup, primarily due to the limitations of parsing static HTML content. To overcome this challenge, Selenium, a more versatile tool for web automation, was adopted. Selenium allowed for

the dynamic interaction required to access and extract data from websites with dynamically generated content, such as flight information displayed through JavaScript.

XPath Precision: While XPath expressions are typically reliable for locating and extracting elements on a web page, there were instances where "unavailable element" errors occurred despite the use of seemingly correct XPath expressions. This challenge stemmed from variations in the website's structure or the loading speed of elements. The solution involved a meticulous review and adjustment of XPath expressions to ensure their accuracy and consistency.

Unpredictable Refresh Requests: The booking.com platform intermittently presented refresh requests during the scraping process, potentially disrupting data collection. These requests were triggered by the website's real-time data updates or changes in flight availability. To address this challenge, try-except constructs were implemented to detect and handle these unexpected refresh requests gracefully, allowing the scraping process to resume without data loss or interruption.

In spite of these challenges, the persistent efforts in web scraping and subsequent exploratory data analysis have culminated in a dataset that empowers individuals to make well-informed decisions when selecting their preferred flights. These challenges definitely increased my critical thinking and adaptability skills with new tools.