Where do I fly next? A Data Science Mini Project

Umar Faruk ABDULLAHI

Google Colab Notebook: https://colab.research.google.com/drive/18GH7o9C1f-1vI_wCAxTh-Whwev4_vW7v#scrollTo=4UYmbsXRV2lJ&unigifier=1

Introduction

This project addresses the challenge faced by users in selecting the most optimal flight from a plethora of options available on various airline websites. The abundance of flight information can overwhelm users, leading to suboptimal choices and time inefficiency. The primary objective is to streamline this process by integrating data from flight websites, conducting comprehensive data processing, and employing exploratory data analysis techniques.

The culmination of this project will empower users with a tool that facilitates informed decision-making during the flight selection process. By offering a tailored user interface and leveraging data-driven insights, the project aims to save time and enhance the overall efficiency of flight searches.

To solve the problem, the project is structured into distinct four steps as follows:

1. **Data Collection:**
    - Extracting flight data from three prominent airline websites.
    - Ensuring data uniformity and reliability through rigorous validation processes.
2. **Data Processing:**
    - Implementing advanced data processing techniques to harmonize and cleanse the collected website data.
    - Standardizing data formats e.g., date, duration for seamless analysis.
3. **Exploratory Data Analysis (EDA):**
    - Utilizing comprehensive data visualization methods to uncover patterns, trends, and correlations within the flight data.
4. **User Interface Development:**
    - Designing an intuitive and user-friendly interface for flight criteria selection.
    - Integrating user preferences to display only flights that fit the selected criteria.

1.Data Collection (Web Scraping)

In this procedural phase, imperative flight data was procured from three distinct flight websites, adhering to predefined search parameters as follows:

- **Departure City:** Helsinki
- **Destination:** London (All airports)
- **Date:** 30th October 2023
- **Travel Type:** One way

The chosen flight websites for this comprehensive data retrieval process include the following:

1. **Kayak:** https://kayak.com
2. **Momondo:** https://momondo.com
3. **Skiplagged:** https://skiplagged.com

To execute the data extraction seamlessly, a **webdriver** agent was implemented utilizing the Selenium framework for each respective flight website. The resultant data were methodically stored in local **.html** files, ensuring fast accessibility and utilization in subsequent stages of the project. The choice

of **.html** as the archival format allows for the preservation of the structural representation of the web pages.

2.Data Selection and Processing

In this phase, the amassed website data, stored in the **.html** format, underwent systematic processing leveraging the BeautifulSoup library. The selected features were chosen based on their direct relevance to the project's objectives - ensuring alignment with the overarching aim of refining the flight searches. Simultaneously, features that promised substantive insights into the intricacies of the flight data were also considered, enriching the subsequent data analysis phase.

The extracted features can be grouped into two categories:

1. Quantitative Variables:
    - Flight Duration: The duration of the entire flight.
    - Price: The monetary cost associated with the flight.
    - Departure Time: The initiation of the flight.
    - Arrival Time: The culmination of the flight.
    - Number of Stops: Pertinent for non-direct flights, providing a crucial metric for traveller consideration.

2. Categorical Variables:
    - Airline Operator: The designated airline facilitating the flight.
    - Source of Data: The specific flight website from which the data originated.
    - Destination Airport: Recognizing that the destination city, London, encompasses multiple airports, this categorical variable captures the nuanced diversity in flight destinations.

For each of the data sources (flight websites), a provisional Pandas dataframe was created to facilitate data processing by utilizing the library's functions. The necessity for such a preliminary measure arose from the disparities in the data representation formats employed by each website. To mitigate the intricacies introduced by these divergent data formats, a meticulously defined final data structure was initialized within a designated final dataframe. Conforming to the pre-established criteria, each individual temporary dataset underwent systematic processing, thereby ensuring uniformity and coherence within the broader dataset. The amalgamation process involved merging all individual dataframes into a singular comprehensive dataframe. This final step holds paramount significance as it not only streamlines the dataset for the forthcoming analytical phase but also ensures the harmonious integration of diverse datasets, thus mitigating the potential discrepancies in the representation of flight information.

3.Exploratory Data Analysis (EDA)

In this phase of the project, the dataset underwent analysis by employing varying data visualization techniques to uncover the distributions, relationships and correlation between the numerous features in the dataset. The collection of data from different sources, each associated with a distinct flight operator, introduced notable variance in key quantitative attributes, notably in price, flight hours, and layover durations. This divergence was further exacerbated by the presence of outliers within these features. Figure 1 below shows a succinct summary of the quantitative data:

	Price	No of Stops	Flight Hours	Layover Duration
count	711.000000	711.000000	711.000000	711.000000
mean	189.616906	1.371308	11.306610	6.189873
std	151.930279	0.593361	7.127329	7.012601
min	26.320000	0.000000	3.000000	0.000000
25%	109.980000	1.000000	7.000000	2.000000
50%	139.120000	1.000000	9.000000	4.000000
75%	184.710000	2.000000	13.000000	7.000000
max	998.280000	3.000000	41.000000	37.000000

Figure 1: Summary of quantitative data

Additionally, interesting patterns were observed within the temporal dynamics of flight schedules. A noteworthy observation revealed a substantial proportion of flights departing during the evening hours and a substantial proportion arriving at their destinations in the morning. This suggests a strategic consideration by airline operators, possibly tailored to accommodate and align with passengers' preferred schedules.

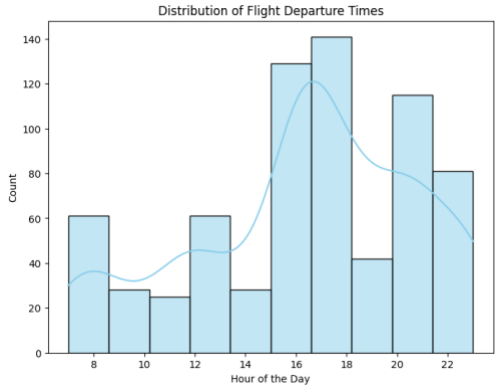

Figure 2: Departure Times

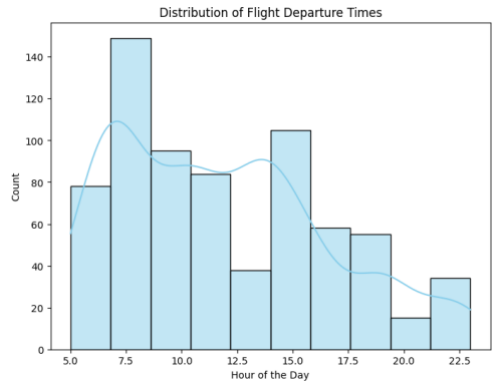

Figure 3: Arrival Times

As depicted in Figures 4 and 5 below, an intriguing revelation emerged—contrary to common expectations, no discernible correlation was observed between the duration of a flight and its corresponding price. This absence of a linear relationship underscores the complexity of pricing determinants, urging a nuanced examination of variables that contribute to the economic dynamics of air travel. However, noticeable correlation between the airline operators and price of flights was observed with certain operators consistently having higher fares than their counterparts.

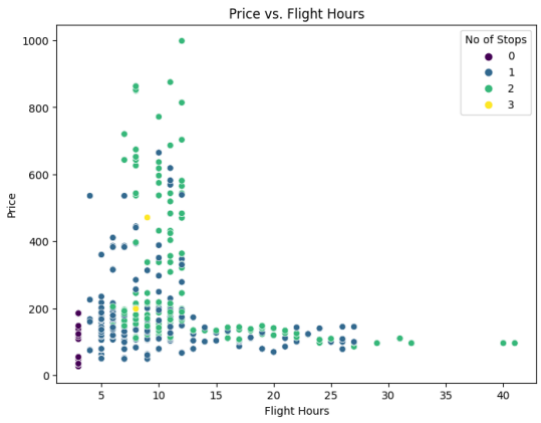

Figure 4: Flight Hours vs Price

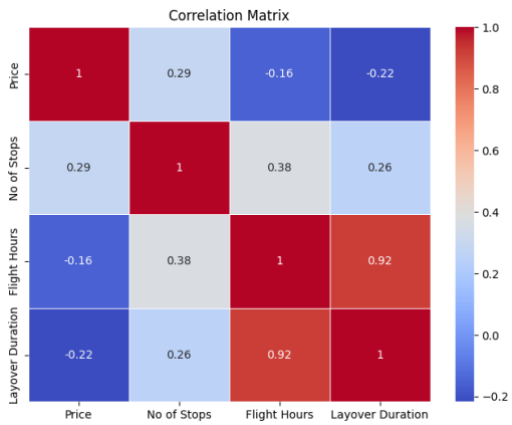

Figure 5: Correlation Matrix

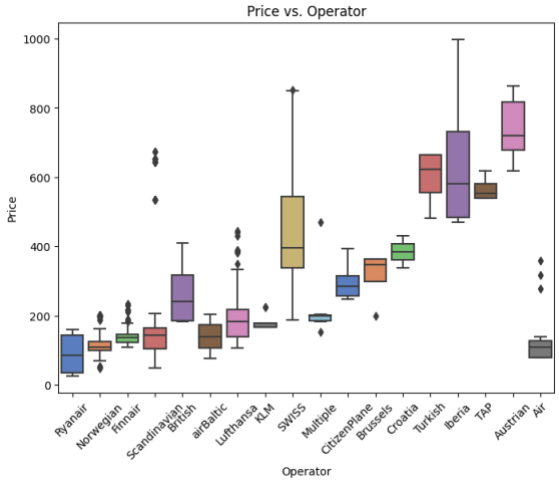

Figure 6: Price vs Operators

Upon scrutinizing the distribution of airports within the destination city, several noteworthy insights surfaced. A substantial number of airline operators predominantly directed their flights to a specific airport within the city. Furthermore, as shown in figures 7 and 8, discernible variations in flight prices were evident based on the geographical location of airports, with flights to city centre airports commanding notably higher fares compared to those serving the outskirts. This disparity underscores the premium associated with the convenience of proximity to the city centre, shedding light on the cost considerations that users must bear for accessibility to the nearest airports.

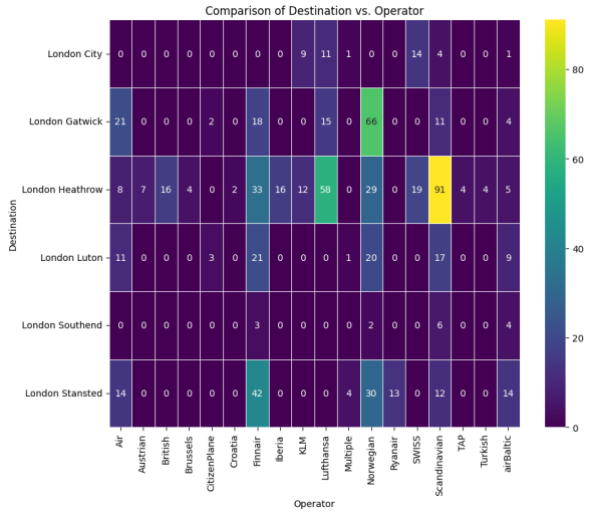

Figure 7: Destination airport vs airlines operators

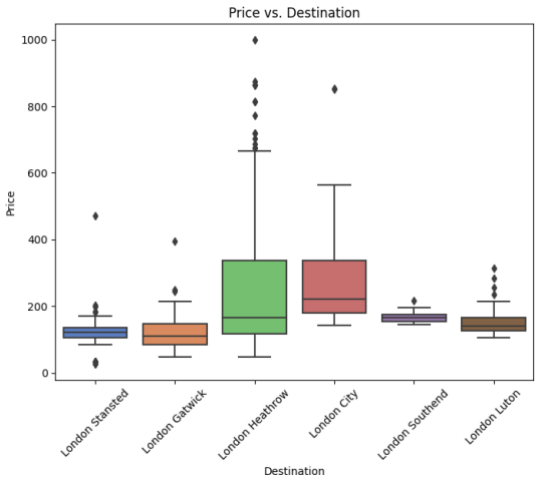

Figure 8: Airfare against the destination airport

Overall, the data analysis brought to the forefront a multitude of factors that intricately influence airline operations. The examination elucidated the intricate interplay of these factors in shaping the pricing and duration dynamics of flights. Consequently, users are urged to conscientiously consider the impact of these multifaceted determinants when making informed decisions about selecting the most suitable flight for their travel needs.

4.User Interaction

In the final step of the project, a simple user interface was created to take in user flight criteria and filter all flights based on that specified criterion. Users could specify their preferences, including the desired price range, flight duration, maximum number of stops, preference for direct flights, and a designated airline of choice. Beyond these fundamental criteria, users were also afforded the option to input additional criteria, contributing to a more refined and personalized flight selection process. The image below provides a visual representation of the cheapest and fastest flight options, as determined by the following sample criteria:

- **Price range:** 20 – 100 euros
- **Trip duration:** 3 – 5 hours
- **Direct flight only:** No
- **Maximum stops:** 2
- **Preferred airline:** None

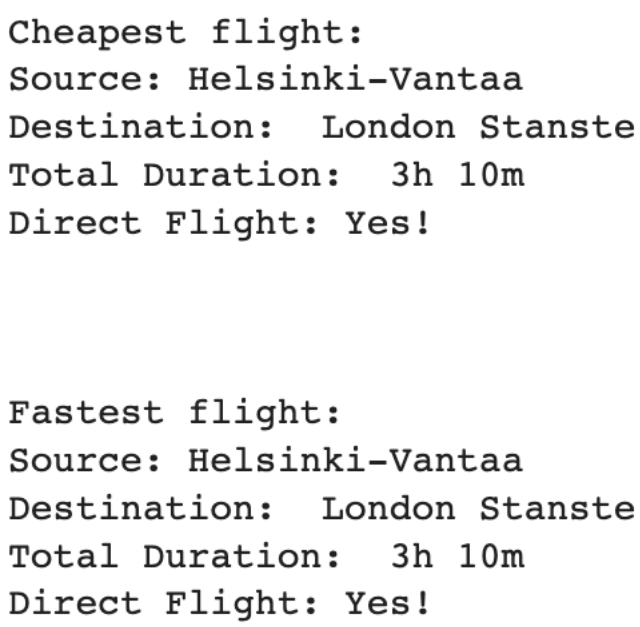

Figure 9: Cheapest and fastest flights
(In this case, it's the same flight)

Conclusion

Through this project traversed the intricate landscape of air travel, from the tedious collection of data across diverse flight websites to the comprehensive analysis of factors influencing airline operations. The consolidation of relevant quantitative and categorical data, harmonized through systematic data processing, laid the foundation for a nuanced exploratory analysis.

There have also been a number of challenges encountered while working on the project. The most challenging part of the process was in the data collection phase. Due to the dynamic nature of flight websites, a webdriver agent – which is a tool for automation and testing, had to be employed to facilitate the dynamic content generation of the websites. Websites also place a lot of emphasis in protecting their content from bots and other unidentified users. To bypass this, a stealth library had to deployed to spoof the connection to the websites. This also highlights a crucial point in carefully considering and adhering to the data policies set by external sources while collecting data.

In the data processing stage, the varying methods of data representation from the distinct websites was a bottleneck for harmonized standardization of data. To circumvent this, the data from every website was pre-processed separately according to a specified guideline before merging all data to a cumulative data frame.

In conclusion, the project unravelled patterns in flight schedules, highlighted the nuanced pricing dynamics influenced by airport locations, and unveiled surprising disconnections between flight duration and pricing. The creation of a user interface represents the translation of these analytical findings into a practical tool, affording users the ability to tailor their flight selections based on diverse criteria.