Name: Refiloe Kekana

August 18, 2023

Student number: u12087752
Date: 18 August 2023

MIT 805 ASSIGNMENT 1: Exploring Flight Price Dataset

This academic paper delves into the technical aspects of the flight price dataset, designated as itineraries.csv. The dataset's attributes, source, and relevance are examined in light of the V's of big data: Variety, Volume, Velocity, and Veracity. The dataset, comprising flight details from various airports in the United States, presents insights into the aviation industry and flight pricing dynamics.

0.1 Technical Aspects of the Dataset:

The dataset encompasses the following technical attributes:

- Data Set Size: Approximately 6 GB.
- File Name: itineraries.csv
- Data Structure: The data is structured in a CSV file format and organized as a tabular structure.
- Data Types: It contains a mix of alpha-numeric entries, including 16 string columns, 3 integer columns, and 3 Boolean columns. The latter encompass features such as isBasicEconomy, isRefundable, and isnonstop.
- Temporal Scope: Data was collected from 2022 April 16 to 2022 October 05, resulting in a dataset just under a year old. The dataset consists of 27 columns and a significant 5,999,739 individual data entries. Data collection involved employing web scraping methodologies, and capturing information across diverse dates and times.

0.2 Purpose and Characteristics of Data:

The dataset encapsulates flight details for 9 notable US airports: ATL, DFW, DEN, ORD, LAX, CLT, MIA, JFK, EWR, SFO, DTW, BOS, PHL, LGA, IAD, and OAK (Wong, 2022). The rationale behind data collection remains undisclosed; however, a plausible purpose is to facilitate flight price comparison among different airports and days. This supposition aligns with the dataset's focus on individual clients' flight ticket data. The dataset was scraped from prominent travel booking websites like Expedia. Features like departureAirport, arrivalAirport, and totalFare are pivotal for querying flight data APIs, primarily used for extracting airfare prices. Notably, Python was leveraged for web scraping, automating URL requests to extract pertinent details.

The data was diligently collected by [1]. Central to the dataset's integrity is the leg-ID, which furnishes a unique identifier for each flight. Additionally, flightDate serves as a consistent timestamp, reflecting the data's uniformity in terms of year, month, and day.

Features such as startingAirport and destinationAirport serve the purpose of geo-locating flights, depicting flight paths. The utilization of fare basis codes enables insight into passengers' payment methodologies and highlights whether a client's ticket is part of a holiday package. This harmonizes with the data's origin from travel booking websites (Academic-accelerator.com).

Furthermore, the segmentsEquipmentDescription facet characterizes the aircraft types boarded by clients or to be boarded. The segmentsDepartureTimeRaw and segmentsArrivalTimeRaw features aid in discerning departure and arrival times, pivotal for determining flight duration.

0.3 Predictions and Expected Correlations:

- Predicted Positive Correlations: Totalfare is anticipated to correlate positively with isBasicEconomy, travel duration, and distance traveled, reflecting longer flights' higher costs.
- Additional Expected Correlations: It is expected that the relationship between totalfare and whether the flight ticket is for an economy seat (isBasicEconomy) will be positive. Furthermore, a positive correlation between travel duration, distance traveled (depicted by startingAirport and destinationAirport), and totalfare can be inferred. Longer flights may entail increased costs due to extended distances traversed. Conversely, a negative correlation between fare price and search date is plausible, as flight prices are likely to surge during specific seasons and holidays.

0.3.1 Implications Based on the V's of Big Data:

- Variety: The dataset's amalgamation of string, Boolean, and numeric data contributes to its variety. It encompasses multiple dimensions, including travel duration, departure and arrival locations, and industry-specific information denoted by fare basis codes (Academic-accelerator.com).

- Volume: The dataset's sheer volume is underscored by its 5,999,739 unique data entries, providing substantial coverage. Captured over approximately 6 months, spanning seasons of elevated travel, the dataset allows monitoring of ticket price fluctuations over time. As the dataset is updated daily, its expansion continues with growing flight bookings. The dataset explores a considerable number of attributes, boasting 27 distinct characteristics.

- Velocity: A daily recording frequency underscores the dataset's velocity. Data collection efficiency was enhanced using Cron Jobs, a scheduling tool, and Python's multiprocessing library (Python-crontab) (multiprocessing — Process-based parallelism).

- Veracity: The dataset's veracity is assured by its origin from Expedia, a reputable travel booking platform. The leg-ID acts as a verifiable, unique identifier, upholding data integrity by enabling traceability.

0.3.2 Conclusion:

In conclusion, the flight price dataset, meticulously collected and organized, imparts valuable insights into the aviation domain's intricacies. By assessing the dataset through the lens of big data's V's, this paper has explored its significance, applications, and future potential. This dataset serves as a valuable resource for investigating trends in flight pricing, thereby benefiting both travelers and aviation industry stakeholders.

Bibliography

[1] Wong Dillon. “Flight Prices”. In: (2022).