Flight Prices Analysis Using Hive and Hadoop

Authors: Ragi Dave, An Mach, Ankita Hasmukhbhai Savaliya, Bhumika Suvagia
Department of Information Systems, California State University, Los Angeles
CIS 5200 – System Analysis and Design
rdave4@calstatela.edu, amach3@calstatela.edu, asavali2@calstatela.edu, bsuvagi@calstatela.edu

Abstract: This paper explains an analysis of flight prices purchased on Expedia between April to November 2022. The paper outlines the methods and processes used for data manipulation and further analysis. Our primary objective of the project was to provide a clear understanding of handling large data files and data cleaning processes using Hadoop File System and Hive. Besides, data interpretation and analysis were conducted using Excel and Tableau where the visualizations such as bar graphs, heat maps, 3D maps, dual-axis charts, and pie charts were created to facilitate flight prices analysis.

1. Introduction

The airline industry is one that requires sophisticated database systems as there are many elements that airline companies need to keep track of, including but not limits to: number of available seats, types of tickets, flight schedules, airport terminals. Airfare is definitely one of the most crucial items of interest amongst the decision makers because it is directly related to a company's revenue and profit. While customers seek to get the lowest ticket prices possible to reduce their travel expense, airlines try to keep their overall revenue as high as possible to stay in business.[1] Hence, this motivates us to look into flight prices dataset.

We have chosen this dataset on Kaggle because it has both dimensions and measures, along with time and location data that can be used for tempo-spatial analysis as required by this project. Through analyzing this data set, we will gain insights into pricing patterns and trends for routes and travel periods during the specified timeframe. Moreover, as a popular travel booking website, Expedia can offer valuable information about consumer purchasing behavior and preferences. Overall, this dataset can be beneficial for the airline industry and travelers in terms of understanding pricing strategies, identifying cost-saving opportunities, and making informed travel decisions.

2. Related Work

Expedia is a popular online travel agency that offers services such as airline ticket booking and hotel reservations. There are quite a few related works that are publicly available based on flight ticket prices data. It offers customers relevant information about flight schedules, airline amenities, and other details to assist them in making informed decisions about their travel plans. These studies can include academic research papers that cover information such as air ticket prices prediction for consumer and profit margin in the airline industry. One of the works that is based on flight prices data is available in the form of research paper from International Research Journal of Modernization in Engineering Technology and Science (May 2021).[2] This study focuses on the use of machine learning to predict airline ticket prices. The outcome of the study is the development of a model that assists passengers in making informed decisions about when to purchase tickets for the best deal. To create the model, the researchers collected specific airline data, including features such as travel time, arrival time, and airways, and used machine learning techniques to extract relevant features and predict prices.

Another related work was done by Iraqi Journal of Computers, Communications, Control & Systems Engineering (IJCCCE) (September 2022).[3] This study explores the use of machine learning algorithms for predicting flight prices in an Indian airline network. It offers a promising approach for airlines to improve their revenue management and provide more accurate pricing for customers, leading to a better travel experience overall.

Another study done by Hakan Yilmazkuday focuses on estimating profit margins in the U.S. airline industry at the domestic route level using a dynamic estimation methodology (December 2021).[4] All in all, our work is similar to this study in the terms of deliverables but the tools that we used for this are different, since we are focusing on interactive visuals and depth of information, giving more insights on flight prices and weekdays-based analysis.

3. Specifications

The Flight Prices dataset comprises of itinerary related data such as flight date, airfare, starting airports, destination airports, and so on. Each record is a purchased ticket found on Expedia to and from 16 airports, including ATL, DFW, DEN, ORD, LAX, CLT, MIA, JFK, EWR, SFO, DTW, BOS, PHL, LGA, IAD, and OAK. The dataset is of the size 31.09 GB and covers several months in 2022 (April 2022 – November 2022) in the itineraries file. There are 27 columns in the original dataset, but we only utilized 17 of them: FlightID, SearchDate, FlightDate, StartingAirport, DestinationAirport, FareBasisCode, TravelDuration, ElapsedDays, IsBasicEconomy, IsRefundable, IsNonStop, BaseFare, TotalFare, SeatsRemaining, TotalTravelDistance, SegmentsAirlineName, and SegmentsEquipmentDescription.

Since the Flight Prices dataset only contains three-letter airport codes without other geographical data, an additional Airport Codes dataset that contains airports' latitude and longitude We will join these two tables using the airport code column to conduct the tempo-spatial analysis.
Data set URL:
https://www.kaggle.com/datasets/dilwong/flightprices
https://www.kaggle.com/datasets/mike90/airport-codes

Table 1 shows files and size of the files from dataset.

Table 1 Data Specification

Data Set	Size (Total 31.09 GB)
Itineraries	31.09 GB
Airports	580.47 kB

The below table shows the specification for Hadoop cluster specification for our project.

Table 2 H/W Specification

Cluster version	3.1.2
Number of nodes	5
Number of CPU cores	8 cores * 5 nodes = 40 cores
CPU speed	1995.312 MHz
Total Memory Size	390.71 GB

4. Implementation Flowchart

Initially, the raw dataset, which comprises the detail of flight itineraries and airports, was downloaded from Kaggle and unzipped on a local Windows laptop. The whole process of date manipulation is shown in the below flowchart (Figure 1). Two files in csv format were secure copied to Linux server, then uploaded to the Hadoop File System. After that, HiveQL is used as querying language to create the tables' schema, clean data, create queries and export the results. Once the output file has been downloaded and opened in Excel format, we used Excel's 3D map and Tableau to obtain the visualizations.

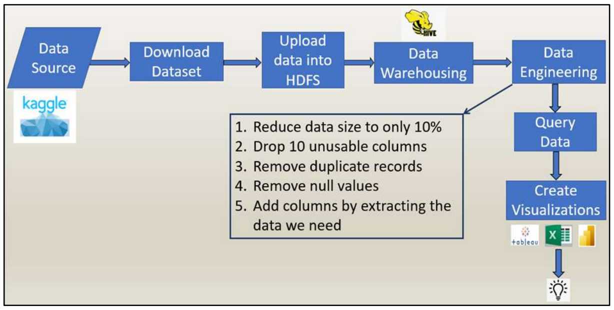

Figure 1 – Architecture Workflow

5. Data Cleaning

Raw files were uploaded and stored in HDFS and then loaded into tables using Beeline Client. We created two tables, FlightData and Airport, each point to the respective directory in HDFS. Data cleaning was conducted using different techniques such as random sampling, concatenation, null removals, and so on.

The FlightData table has 82,138,753 records. Since the big data size could negatively affect other classmates and labs, we were instructed to reduce our data to 10% of the original size by randomly sampling it. Below is the sampling code:

```
select * from FlightData
where rand() <= 0.1
distribute by rand()
sort by rand()
limit 8000000;
```

The original FlightData table has 27 columns, but we were not going to use all of them for our analysis and visualizations, so we dropped the unusable columns with the RELACE COLUMNS action using HiveQL. The resulting table has 17 columns as listed in the Specifications section. We further removed any duplicate records. Since our dataset is huge (almost 8 million rows of data), we chose to remove any row that contain null values. The resulting table has 7,406,035 rows.

We added two additional columns that we needed for our analysis. The first one is FlightMonth column which was populated with the month from FlightDate column. The second is FlightRoute column that concatenate startingAirport with destinationAirport columns.

6. Analysis and Visualization

After data cleaning and preparation for further analysis, we created our visualizations using Tableau, Power BI, and Excel 3D map. We chose Tableau because it is user-friendly and offers great visualizations with just a few drags and drops. We created a variety of charts in Tableau for our analysis, including Bar charts, Heat map, Pie chart, Dual axis chart, and Line chart. We also learned how to use 3D Map in Excel via our lab exercise, which is a great tool for temporal-spatial analysis in this project.

6.1 Bar Graphs in Tableau

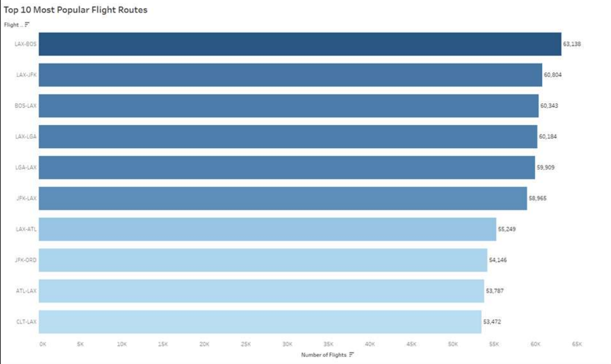

Figure 2 – Top 10 Popular Flight Routes

The horizontal bar graph depicts the top 10 most popular flight routes in the United States between April and November 2022 where the X and Y axis represent the number of flights and flight routes respectively. Most of the routes involve either LAX or one of the East Coast airports – BOS (Boston), JFK (Queens, NY), LGA (LaGuardia, NY), or CLT (Charlotte). Specifically, the route between LAX and BOS is the most popular, with over 63,000 flights recorded. After creating this chart, it can be observed that Los Angeles (LAX) was a popular destination or connection point for the travelers as it is involved in nine out of the top ten flight routes, except for the JFK-ORD route. According to the article "Busiest Airline Routes in the U.S.", New York to Los Angeles and New York's LaGuardia to Chicago are in the third and fourth place.[5] Comparing that to Figure 1, the JKF-LAX route is at the sixth place with 58,965 flights on Expedia; the JKF-ORD route is at the eighth place with 54,146 flights. All in all, this information can be useful for airlines to optimize their operations and tailor their services to meet the needs of their customers.

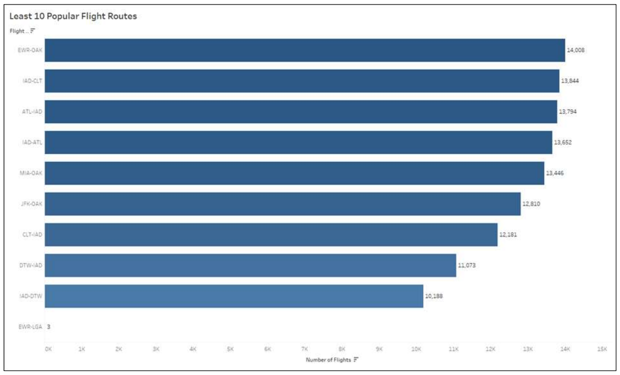

Figure 3 – Least 10 Popular Flight Routes

Based on the number of flights between April and November 2022, Figure 3 illustrates the least 10 popular flights in the US. It indicates the routes with the lowest demand by showing the airport names and the number of flights for each route. The route between EWR and LGA had the lowest number of flights, which accounted for only 3 flights of the cleaned dataset. This is most likely due to the short distance between the two airports. According to Air Miles Calculator website, "The driving distance from Newark (EWR) to New York (LGA) is 23 miles / 37 kilometers, and travel time by car is about 41 minutes… The estimated flight time from New York Newark Liberty International Airport to New York LaGuardia Airport is 31 minutes".[6] It is more convenient to drive between EWR and LGA airports unless they are part of a connected flight.

The EWR to OAK route topped the chart with only 14,008 flights. Moreover, based on this analysis, airlines can gain insights into the least popular flight routes and make data-driven decisions to optimize their route planning and capacity allocation. By understanding which routes have lower demand, airlines can adjust their pricing strategies and marketing efforts to attract more customers and increase revenue.

6.2 Heat Map in Tableau

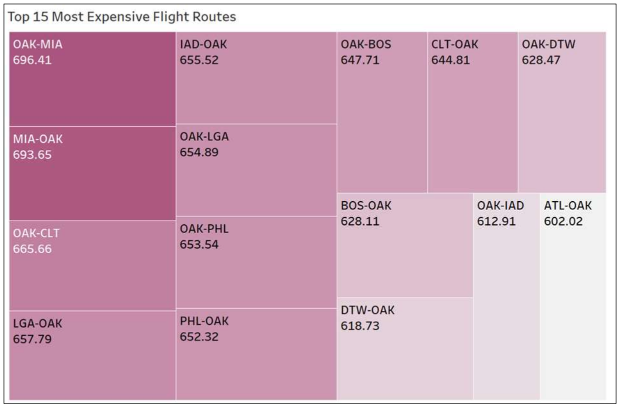

Figure 4 – Top 15 Most Expensive Flight Routes

For our analysis, we have chosen to focus on the top 15 flight routes that are the most expensive. The heat map shows that the route from Oakland (OAK) to Miami (MIA) ranks at the top of the chart with the average airfare of $695. Interestingly, all of the top fifteen most expensive flight routes involve OAK as either the departing or arriving airport. On average, any flight related to OAK has a cost greater than $600, while the least expensive route from Figure 4 is from Atlanta (ATL) to Oakland (OAK) at $602.

6.3. Dual Axis Chart in Tableau

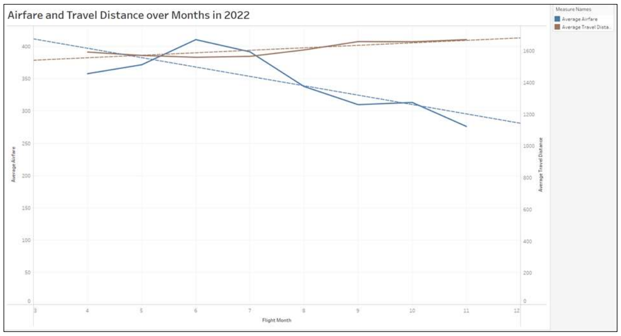

Figure 5 – Airfare and Travel Distance over Months in 2022

By analyzing the dual-axis chart displayed in Figure 5, we can gain insights into the trends of Average Airfare and Travel Distance between the months of April and November in 2022. It can be inferred that in the month of June, the airfare was considerably higher (exceeding $400), but the travel distance was comparatively shorter (less than 1600 miles). On the other hand, in the month of November, the airfare was more affordable (ranging between $250 to $300), but the travel distance was longer (more than 1600 miles).

We have included a dotted trend line to help visualize patterns more clearly. The trend line indicates that as airfare (represented by the blue dotted line) decreases, there is an increase in travel distance (represented by the brown dotted line).We assume that in order to attract customers, airlines could have implemented effective strategies such as loyalty programs, exceptional customer service, and targeted social media advertising. By providing incentives for customers to continue flying with their airline, offering exceptional service that goes above and beyond expectations, and utilizing social media platforms to reach potential customers, airlines can create a strong brand image and gain a competitive edge in the industry.

6.4. Bar Graph in Power BI

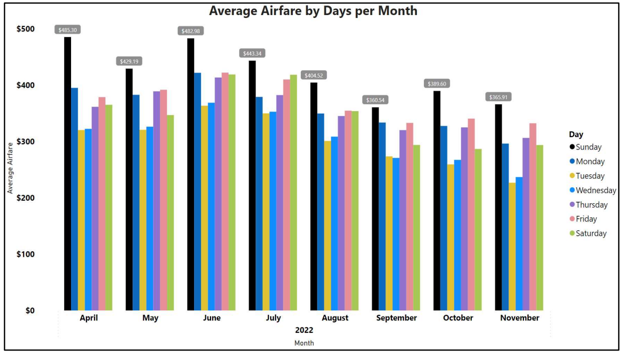

Figure 6 – Average Airfare by Days per Month

Figure 6 shows the average airfare from April 2022 to November 2022, categorized by days of the week. The graph provides insights into the fluctuation of airfare prices throughout these months. According to the data depicted in Figure 6, it can be observed that Sunday flights consistently had higher airfare compared to other days of the week during this period. The airfare for Sunday flights ranged from $360 to $450 across all the months. This indicates that if you were to book a flight on a Sunday, you would generally expect to pay a higher price than on other days.

On the other hand, Tuesday flights tended to have the cheapest airfare prices in most months, followed closely by Wednesday, with the exception of September. This implies that if you were looking for the most affordable airfare, booking a flight to travel on a Tuesday or Wednesday would be a favorable option, as it was consistently cheaper than other weekdays. According to Johnson in "What's the cheapest day to book flights?", "flying domestically on a Wednesday may help you save 15 percent off airfare" (February 2023).[7]

To gain further insights from the chart, users can interact with it by hovering over the individual bars. By doing so, they can access additional information such as the corresponding month, day, and the average airfare for that particular period. This interactive feature enables users to analyze and compare airfare prices for different months, providing a comprehensive view of the pricing trends over time.

6.4. 3D Map in Excel

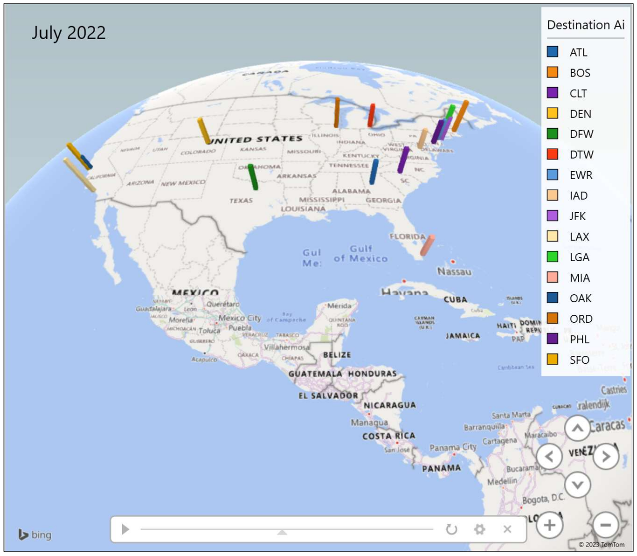

Figure 7 – Popular Destination Airports Over Month

Figure 7 is an animated 3D map in Excel that shows flight data from April to December 2022 utilizing a timeline across different months. To represent popular destinations, we used a bar chart that displays the total number of flights for each month. The time attribute is used to visualize which destinations experienced more flights during specific months. In the time column, month-year format has been chosen. This visualization is available in video format clearly displaying the size of each bar increasing during August and September, then decreasing until November. This format effectively illustrates the changes in flight patterns over time.

6.5. Pie Chart in Tableau

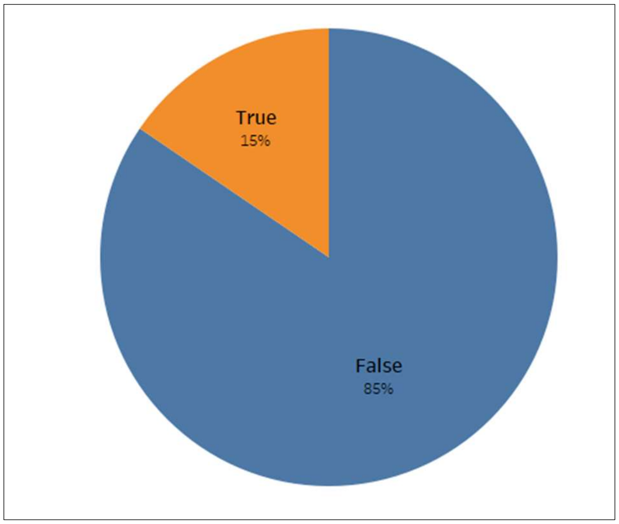

Figure 8 – Distribution of Basic Economy Tickets

It can be interpreted from Figure 8 about the distribution of Basic Economy Tickets. The chart is divided into two slices: one labeled "True" and the other labeled "False". The "True" slice is represented in orange and indicates the proportion of Basic Economy Tickets, which accounts for only 15% of the total dataset. Meanwhile, the "False" slice represents the percentage of Non-Basic Economy Tickets, which comprises 85% of the dataset. The size of the "True" slice is much smaller than the "False" slice, highlighting that Basic Economy Tickets are a relatively small proportion of the overall Flight Prices dataset.

7. Conclusion

In summary, this report presents an analysis of flight prices from April to November 2022. The results of this analysis will assist travelers in finding the best flight deals, while also enabling the airline industry to meet consumer demand, increase revenue, and provide a positive customer experience.

Based on the data analysis, the following points can be summarized:

i. The report highlights the top and least 10 flight routes based on the number of tickets sold.

ii. The report also identifies the top 15 most expensive flight routes, based on the ticket price.

iii. In June, airfare was found to be higher in comparison to the traveled distance.

iv. The cost of airfare for Sunday flights was higher, while Tuesday flights generally had lower prices, except for the month of September.

v. In August and September, the majority of destination airports experienced higher traffic. LAX experienced heavy traffic as compared to all the destination airports in August and September.

vi. It was observed that airlines offer only a small proportion of basic economy tickets.

However, it should be noted that this dataset only contains information on flights to and from 16 airports. Further research could be conducted by analyzing data from additional airports to gain a more comprehensive understanding of the trends in the airline industry.

For the complete project document repository (term-paper, PowerPoint slides, and code), visit the following GitHub link: https://github.com/amach3/Flight-Price-Analysis

References

[1] Abdella, J.A., Zaki, N.M., Shuaib, K., & Khan, F. (2021). Airline ticket price and demand prediction: A survey. Journal of King Saud University - Computer and Information Sciences, 33(4), 375-391. https://doi.org/10.1016/j.jksuci.2019.02.001.

[2] Uddin, M.S., Khan, M.I., Qadeer, M., & Mohammed, S. (2021). Airfare Intelligence Based on Machine Learning. International Research Journal of Modernization in Engineering Technology and Science, 3(5), e-ISSN: 2582-5208. Retrieved from http://www.irjmets.com/uploadedfiles/paper/volume3/issue_5_may_2021/9706/1628083389.pdf

[3] Fadhil, H., Abdullah, M., & Younis, M. (2022). A Framework for Predicting Airfare Prices Using Machine Learning. IRAQI JOURNAL OF COMPUTERS, COMMUNICATIONS, CONTROL AND SYSTEMS ENGINEERING, 22(3), 81-96. doi: https://doi.org/10.33103/uot.ijccce.22.3.8

[4] Yilmazkuday, H. (2021). Profit margins in U.S. domestic airline routes. Transport Policy, 114, 245-251. https://doi.org/10.1016/j.tranpol.2021.10.010

[5] Greenberg, P. (2022, November 6). Busiest airline routes in the U.S. https://petergreenberg.com/2022/11/04/busiest-airline-routes-in-the-u-s/

[6] Distance from Newark to New York (EWR – LGA). Air Miles Calculator. (n.d.). https://www.airmilescalculator.com/distance/ewr-to-lga/

[7] Johnson, H. D. (2023, February 28). What's the cheapest day to book flights. https://www.bankrate.com/finance/credit-cards/cheapest-day-to-book-flights/