Text

Description automatically generated

Logo

Description automatically generated

**Group Alpha**

ALY6110- Big Data and Data Management, 80443, Spring 2022.

Module 6 Signature Assignment: A comprehensive Airlines Analysis

**Group members:**

[Vishnu Vamsi Koye](https://northeastern.instructure.com/groups/101181/users/160956)

[Hari Babu Muga](https://northeastern.instructure.com/groups/101181/users/153492)

[Rambhupal Payyavula](mailto:Rambhupal Payyavula) ([Payyavula.r@northeastern.edu](mailto:Payyavula.r@northeastern.edu))

[Charan Thota](https://northeastern.instructure.com/groups/101181/users/150764)

[Saipavanvenkatanarayana Abhise Vegiraju](https://northeastern.instructure.com/groups/101181/users/95968)

**SUMMARY**

**CONTENT**

We comprehensively analyze the airline data from 1987 to 2008. As the data is enormous, around 118.9 million records, it best suits the big data tools to deal with it. So, we used the Apache spark framework to process and analyze the data.

**Importing the required libraries:**

Let’s start with importing the required libraries to analyze the airline's data. We will use the glob, pandas, NumPy, matplotlib, seaborn, and PySpark to perform the project.

Graphical user interface, text, application, email

Description automatically generated

We have imported glob to get the list of files with the same pattern, pandas to manipulate the data, seaborn and matplotlib to visualize the data, and PySpark to process and manipulate a large amount of data.

We need to create a spark session to utilize the spark framework and API in this project productively, so we have imported a SparkSession function from the PySpark library and started a spark session. And below is the code we executed to create a spark session.

Text

Description automatically generated

**Extracting the Data:**

We have around 26 CSV files; we have used the glob faction to combine all the CSV files with the same pattern into one large file named csv\_files.

Graphical user interface, application, Word

Description automatically generated

Then we sorted all the extracted files having a file name with the length of 4 characters, and we named that new variable a flight\_csv, which is shown in the below screenshot.

Graphical user interface, text, application

Description automatically generated

Then we have created a spark data frame “df “with the files listed in the flight\_csv. Below is the code we executed to generate the spark in-memory data frame.Graphical user interface, text, application

Description automatically generated

The below screenshot shows that the df spark data frame comprises 118.9 million records and 29 variables.

A picture containing background pattern

Description automatically generated

**Data exploration:**

**Data cleaning:**

Since we mainly focus on the airport traffic and canceled flights and have 29 columns, we don’t use most of the columns in our analysis. So, we dropped the unused 17 columns from the data frame, resulting in 12 columns.

Text

Description automatically generated with medium confidence

The above screenshot shows the list of fields we have chosen to drop.

**Data analysis:**

**Analyzing the flight cancelations:**

It would be beneficial for consumers to know the chance of their flight being canceled while planning vacations, mainly when booking flights. Customers will be able to plan their journeys better using this information.

The proportion of flights canceled to the flights not canceled gives us a better idea of the frequency of the cancelations. So, we have created a bar graph to analyze the flight cancelations among the different airlines.

Chart

Description automatically generated

It can be seen from the above graph that United Airlines (UA), American Airlines (US), Delta Air Lines, and US Airways have a significant proportion of cancelations to others.

However, the proportions are not giving much clarity on the flight cancelations. So, we have created a pie chart to understand the canceled flights comprehensively from the different airlines.

The pie graph below shows that US Airlines (US), American Airlines (UA), and united airlines (UA) individually contribute 13 percent of canceled flights. Together, Northwest Airlines (NW) and Delta Airlines (DL) contribute 21 percent of canceled flights. Southeast Airlines (WN) and Envoy airlines (MQ) contribute seven percent of canceled flights individually. Continental Airlines have five percent of flights canceled. And all the other airlines contribute less than or equal to 3 percent of canceled flights.

After researching, we realized that all these airlines attributed the cancelations to the bad weather and staff shortage.

Chart, pie chart

Description automatically generated

We have created a bar plot visualization to get more clarity on the yearly flight cancelations of all the registered airlines.

A picture containing bar chart

Description automatically generated

It is clear from the graph that in 2001 and 2000, a significant proportion of flights were canceled compared to the other years.

We wanted to investigate the canceled flights across the different airlines from 1987 to 2008 to know which year was evident the more canceled flights and compare canceled flights based on the year—bar chart best suited for the comparison. So, we have created a bar graph to compare.

Chart, bar chart

Description automatically generated

 It is clear from the above graph that 10.4 percent of flights were canceled in 2001. 8.4 percent of flights were canceled in 2000, and 7.2 percent of flights were canceled in 2007. And in all years, only less than 7 percent of flights were canceled.

We are very enthusiastic to know why most flights were canceled in 2001. After researching, it is no surprise that hijacked flights crashed into the world trade center on September 11, 2001. As a result, all the flights got abruptly stopped. Moreover, we also know that Storms, increased aviation traffic, and a lack of new airport runways contributed to making 2000 the Year of the Flight Delay, resulting in more flight cancelations.

In 2007, around 7.2 percent of flights were canceled due to the increased demand for aviation and, in other words, increased traffic, which motivated us to investigate the airline's traffic from the available data.

**Busiest Flights bookings:**

We had previously seen reports of flight cancellations by various airlines from 1987 to 2008. Now, it’s time to analyze the busiest airlines on the short-listed routes from 1987 to 2008. We used analysis in this process to answer the following five questions. We can advise customers on the best time to book flights by answering the following five questions.

1. Which Airline had the most traffic?
2. What was the busiest Year for flights?
3. What was the busiest Month for flights?
4. What was the Busiest Day of the Month for flights?
5. What was the Busiest Day of the Week for flights?

We have many airline journeys in the dataset. We are interested in the top twelve high-traffic airline routes mainly focused on. So, we have summarized the data of the top 12 high-traffic airports.

Graphical user interface, text, application, email

Description automatically generated

To get a detailed overview of the top-12 high traffic trips among the airports, we need to compare the trips from one airport to another. A bar chart visualization best suits to compare the results. So, we have created a bar graph visualization to compare the trips.

Chart, bar chart

Description automatically generated

The above bar graph shows the number of trips on the y-axis and trip names on the x-axis. It is evident from the chart that the most significant number of trips of around 329K happened between San Francisco International Airport to Los Angeles International Airport and vice versa.

### Which Airline had the most traffic from 1987 to 2007? Considering the use case, we should advise the customer to choose better airways for low-cost travel; to do so, we need to know which airlines passengers prefer. I used a bar plot to visualize the number of passengers traveling by airline to figure this out.

Chart, bar chart

Description automatically generated

We can see from the above plot that "DL (Delta)" airlines are the busiest, with a booking percentage of 13.7 percent, and "WN (Southwest)" airlines have the second-highest booking percentage of 12.9 percent. Let's look at why Delta Airlines was the busiest of all the airlines. There is only one reason for this: Delta Airlines has never failed to keep its promises to its customers. Delta Airlines was less expensive than other airlines, and they provided in-flight amenities such as food and Wi-Fi. And they are now offering the best baggage policy to their customers. Delta Airlines was also interlinked with many international airlines such as KLM, Air France, and others. Similarly, Southwest offers the best amenities and baggage policies on low-cost domestic flights.

### What was the busiest Year for flights?

Before we get into the above question, let's look at the total number of passengers traveling each year, regardless of which airline they choose.

Graphical user interface, application

Description automatically generated with medium confidence

It's difficult to tell which year was busiest from the above table, so we used a bar plot to present the graph for each year. And we had indicated with percentages how much of a share of bookings each year compared to previous years.

Chart, bar chart

Description automatically generated

We can see from the above graph that 2007 was the busiest year, accounting for 6.2% of year bookings. And we can't conclude that 1987 and 2008 were the least busy years because we don't have enough data for these years. The busiest year was 2007, which was due to two factors: one, tourism, and two, a significant part of the IT world was changing rapidly at the time, necessitating the conduct of most business meetings and client meetings. And according to the international tourist market, there were 56 million more tourists in 2006 than in 2007, with 19.6 million more in America. Compared to previous years, this represented a 6% increase in tourist numbers.

### What was the busiest Month for flights?

We can't give proper input to customers until we know which year was the busiest. So, to recommend the best month to book a flight, I created the bar graph below to show the most active months from 1987 to 2007.

Chart, bar chart

Description automatically generated

According to the above bar plot, June was the busiest month for flights, accounting for 8.8% of total passengers traveled across all months. June is the busiest month due to tourism. As we all know, summer vacation begins in June, and during this time, most people try to visit new places. We can also see that the busiest months are January and August due to a new academic semester. During these three months, primarily international students will fly.

### What was the Busiest Day of the Month for flights?

We can't give customers advice based on the busiest month because some days of the month are most active, and some months are not so busy. A bar plot was presented below to visualize these:

Chart, bar chart

Description automatically generated

According to the graph above, the average passenger traveled for the entire month. However, we can see that the end of the month was not as busy because only a few months have 31st, and February only has up to 29th. Apart from that, there is no difference between the days of the month. However, according to "CheapAir" analysis, rates will rise during the holiday season.

### What was the Busiest Day of the Week for flights?

We can give precise advice to customers if we know the busiest year, month, and day of the week. By considering the day of the week, we can advise customers on the best day to book a flight. I used a bar plot to visualize the data from weekdays:Chart, bar chart

Description automatically generated

According to the above graph, Friday was the busiest day. As we all know, most offices and schools are closed on Saturday and Sunday. Most offices and schools were closed after 5:00 p.m. People taking extended vacations prefer to travel on weekends to maximize their vacation time. Because some airports are busier than ever on Friday nights and Saturday mornings, Office Needle recommends flying before 4 p.m. on Fridays. Weekend flights are in higher demand and thus more expensive.

**COMMENTS**

Combine the results of the above analysis and add descriptions here

**CONCLUSIONS**

Add descriptions here

**REFERENCES**