Module 6 Signature Assignment: Leveraging Big Data Analytics in Aviation Industry

College of Professional Studies,

Northeastern University

Ashutosh Singh (singh.ashut@northeastern.edu)

ALY 6110: Data Management & Big Data

CRN: 80442

Prof. Valeriy Shevchenko

June 26, 2022

ABSTRACT

In this report, we will propose a solution to a real-world problem that can be handled with the Big Data. We selected the Airlines dataset because there are many opportunities for data analysis and Big data in the aviation industry for a variety of factors. First up are the risk factors related to the airline industry, which may be further divided into two categories: risk factors affecting the whole airline industry and risk factors affecting a specific airline. As part of this project, we examined risk factors that might have an impact on the whole airline industry, such as terrorist attacks, geopolitical unrest, or security events, lengthy delays or service outages at crucial airports, fuel price risk, etc. Because airlines suffered losses for six straight years as a direct result of the terrorist events that took place in the United States in September 2001, we found that 2007 is the busiest airline year in this report's analysis and the most flight got cancelled in 2001. This demonstrates the enormous potential for predictive analytics and risk management in the aviation industry. Aside from that, the second reason is that optimizing the airline industry is crucial because each Boeing 787 aircraft generates half a terabyte of data, which, when combined with weather forecasts, customer service data, ticketing information, and airport communications, offers a wealth of business insights. We believe that the aviation business has a significant potential for data science and big data, which is also the reason that, according to a research from the website stratascratch, Delta Airlines is one of the "11 Best Companies to Work for as a Data Scientist." Big data may help airlines increase operational effectiveness while simultaneously enhancing consumer experience. As a result, this initiative is primarily focused on improving customer happiness, making predictions, risk managemnet and solving the problem of Air traffic. The most intriguing portion of the information we gathered from a website reads, "NASA has cooperated with United Airlines, Delta Airlines, British Airways, Southwest Airlines, Qatar, EasyJet, and Southwest Airlines to constantly enhance airline safety." Because of this, it's critical to examine the causes of the majority of flight cancellations and the years and months in which they occur so that we may use our study to solve the air traffic problem. Since we will be investigating how airlines may utilize technology and data to improve their operational performance, this project promises to be highly exciting. For our study, we utilized the libraries glob, pandas, NumPy, matplotlib, seaborn, and PySpark. We cocluded our study after performing data science techniques, which allowed us to derive some insightful conclusions from the data. We wish to thank Professor Valeriy Shevchenko for his guidance and help during this process.

SUMMARY

Using the big data analytics tool, we performed an analysis on the metadata of the airline flights that occurred in the United States between the years 1987 and 2008. To begin, it is necessary for us to investigate the past of aviation in the United States. Before the 1920s, people in the United States first began traveling by airplane. This was the decade in which air mail transportation in the United States became the most common, and it was during this time period that the airline transported its first load of mail along with passengers on a single flight. A new airline company would become accessible to the public and available for commercial use once every ten years. In recent decades, airlines have increased the economic benefits that they provide to their customers. These benefits include the provision of meals while in flight, access to electronic devices, in-flight entertainment, the internet, and other similar amenities. Another significant development that took place in the twenty-first century was the introduction of online travel booking. More than fifty percent of all reservations pertaining to travel were made online for the very first time in 2009. Because of this, more pressure was put on airlines to compete and offer the best value for their customers' money.

Flight cancellations and increased airline traffic continue to be a problem for airlines, despite the fact that the companies have increased all economic perks for customers. Because of this, the traveler decided to book a flight. However, this makes their schedule more difficult to follow. Because of this, we are going to delve even further into the subject during the course of this project. While we were working through some of the questions, we also provided recommendations for the most convenient flights that did not involve any complications.

To begin, we developed two primary use cases, one for flight cancellations and another for the busiest flight bookings. A number of insights have been developed with regard to both use cases. In the use case of flight cancellation, we are going to conduct an analysis of flight cancellation across various airlines, the percentage of canceled flights across all airlines, flight cancellation in previous years, and which year had the most cancellations. Second, in order to determine which airlines have the most traffic, the most active year for flights, the busiest month for aviation, the busiest day of the month for flights, and the busiest day of the week for flights will all be taken into consideration. We will provide the passenger who has experienced a more

minor cancellation on their flight with suggestions for airlines, as well as suggestions for days and weeks without heavy traffic.

That data is also very important for an analysis of how the use cases are important and how they are related. As a result, we are in possession of the metadata that was retrieved from this Bureau of Transportation Statistics of the United States Department of Transportation. The dataset includes information on nearly 200 million separate domestic flight movements that occurred in the United States between the years 1987 and 2008. It takes into account all of the 29 different variables, out of which we chose a few to focus on, including the Year, Month, Day of Month, Day of the Week, Unique Carrier, Cancelled, trips, and Airlines.

We chose Apache Spark as our big data analytics tool within the Hadoop ecosystem because spark is an open-source framework designed for ease of use, speed, and sophisticated analytics. In order to conduct our analysis, we relied on PySpark, which is essentially just the Python interface to Apache Spark.

CONTENT

We conducted an in-depth study of the data pertaining to airlines from 1987 to 2008. Because there are approximately 118.9 million records in the data set, the big data tools are the ones that are best suited to manage it. Therefore, in order to process and analyze the data, we utilized the Apache spark framework.

If we want these libraries to be installed in the Jupiter Notebook, running the commands that were just described will accomplish that installation. If we are going to perform a direct installation of the operating system, we need to take the (!) symbol out of the command.

Importing the required libraries:

Importing the necessary libraries is the first step in performing an analysis of the airline's data, so let's get to it. In order to complete the project, we made use of the glob, pandas, NumPy, matplotlib, seaborn, and PySpark libraries.

```
import os
from glob import glob
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
os.getcwd()
```

```
'/Users/lokaraju/Downloads'
```

```
import pyspark
import pyspark.sql.functions as f
from pyspark.sql import SQLContext
from pyspark.rdd import RDD
from pyspark.sql import DataFrame
from pyspark.sql import functions
from pyspark.sql import Row
from pyspark.sql import SparkSession
from pyspark.sql.functions import lit, desc, col, size, array_contains\
, isnan, udf, hour, array_min, array_max, countDistinct
from pyspark.sql.types import *

from pyspark.ml  import Pipeline
from pyspark.sql.functions import regexp_extract, mean,col,split, col, when, lit, count
```

We have imported glob so that we can get a list of files that match a specific pattern, pandas so that we can manipulate the data, seaborn and matplotlib so that we can visualize the data, and PySpark so that we can process and use a large amount of data.

Because we need to start a spark session in order to make effective use of the spark framework and API in this project, we imported a function called SparkSession from the PySpark library and then started a spark session. And below you will find the code that we executed in order to create a sparksession.

```
from pyspark.sql import SparkSession
from pyspark.sql.types import *

#create session in order to be capable of accessing all Spark API
spark = SparkSession \
    .builder \
    .appName("Purchase") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()
```

Extracting the Data:

We have about 26 CSV files, and we have used the glob faction to combine all of the CSV files that have the same pattern into a single massive file that we have called csv files.

```
csv = glob("./*.csv")
csv
```

```
['./1990.csv',
 './1991.csv',
 './1987.csv',
 './1993.csv',
 './1992.csv',
 './1996.csv',
 './1997.csv',
 './1995.csv',
 './1994.csv',
 './2008.csv',
 './2001.csv',
 './2000.csv',
 './2002.csv',
 './2003.csv',
 './2007.csv',
 './2006.csv',
 './2004.csv',
 './2005.csv',
 './1999.csv',
 './1998.csv',
 './1988.csv',
 './1989.csv']
```

```
os.getcwd()
```

```
'/Users/lokaraju/Downloads'
```

We then used a file name length of four characters to sort all of the extracted files containing airline annual data, and we named this new variable flight csv, as shown in the screenshot below.

```
Flightsdata = sorted([i for i in csv if len(i.split("/")[1].rsplit(".",1)[0])==4])
Flightsdata
```

```
['./1987.csv',
 './1988.csv',
 './1989.csv',
 './1990.csv',
 './1991.csv',
 './1992.csv',
 './1993.csv',
 './1994.csv',
 './1995.csv',
 './1996.csv',
 './1997.csv',
 './1998.csv',
 './1999.csv',
 './2000.csv',
 './2001.csv',
 './2002.csv',
 './2003.csv',
 './2004.csv',
 './2005.csv',
 './2006.csv',
 './2007.csv',
 './2008.csv']
```

The files listed in the flight csv were then copied into a spark data frame called "df." Spark's in-memory data frame was generated using the following code.

```
c=0
for file in Flightsdata:
    if c==0:
        df = spark.read.csv(file, header = "true")
    else:
        df_temp = spark.read.csv(file, header = "true")
        df = df.unionByName(df_temp)
        print((file, df.count(), len(df.columns)))
    c+=1
```

```
('./1988.csv', 6513922, 29)
('./1989.csv', 11555122, 29)
('./1990.csv', 16826015, 29)
('./1991.csv', 21902940, 29)
('./1992.csv', 26995097, 29)
('./1993.csv', 32065598, 29)
('./1994.csv', 37245646, 29)
('./1995.csv', 42573081, 29)
('./1996.csv', 47925064, 29)
('./1997.csv', 53336907, 29)
('./1998.csv', 58721628, 29)
('./1999.csv', 64249512, 29)
('./2000.csv', 69932559, 29)
('./2001.csv', 75900339, 29)
('./2002.csv', 81171698, 29)
('./2003.csv', 87660238, 29)
('./2004.csv', 94789508, 29)
('./2005.csv', 101930104, 29)
('./2006.csv', 109072026, 29)
('./2007.csv', 116525241, 29)
('./2008.csv', 118914458, 29)
```

Data exploration:

According to the screenshot that can be found below, the df spark data frame contains 118.9 million records and 29 different variables.

```
print((df.count(), len(df.columns)))
```

```
(118914458, 29)
```

Data cleaning:

Because our primary focus was on airport traffic and canceled flights, and because our analysis contains 29 columns, we do not make extensive use of the majority of the columns. The amount of time it takes to complete a command will increase if we include those columns in our dataset that are not being used. Therefore, we reduced the number of columns in the data frame from 17 to 12, which was more manageable.

```
removecolumns=['Unnamed: 0','CRSDepTime','ActualElapsedTime','AirTime','Origin',
               'Dest', 'CRSArrTime', 'TaxiIn', 'TaxiOut','CRSElapsedTime','CancellationCode','CarrierDelay',
               'WeatherDelay', 'NASDelay', 'SecurityDelay', 'Distance','LateAircraftDelay','Diverted']
df = df.drop(*removecolumns)
len(df.columns)
```

```
12
```

The list of fields that we have decided to remove can be seen in the screenshot that was just displayed.

Data analysis:

By providing customers with answers to the following questions, we are able to provide guidance on the optimal time to book flights.

1. Which airline has the greatest amount of cancelled flights? Why do these Airlines have the highest amount of cancellations, and what can those causes be?
2. Which year appears to have had the most flights cancelled overall, and what may the potential causes be? Is it wise to make decisions only based on data?
3. Which domestic airline routes in the US had the largest flight traffic, and between which two stations?
4. Which airline is the world's biggest public airline companies by number of passengers carried between 1987 and 2008?
5. Which airlines are the main rivals of the largest publicly traded airline corporations in the world in terms of passenger volume?
6. Which year sees the most flight bookings and the aviation industry's boom?
7. Which month of the year do individuals most frequently travel?
8. Which day of the month saw the highest number of passengers taking flights?
9. Which day of the week saw the highest number of passengers taking flights

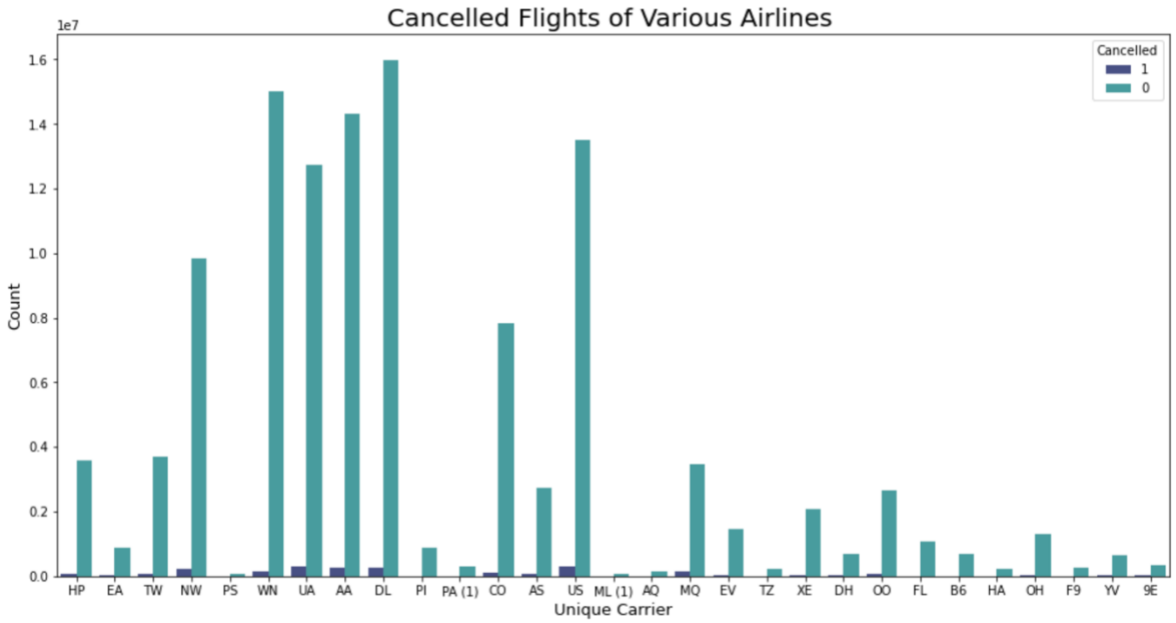
Analyzing the flight cancellations:

Airlines proportions:

When consumers are planning their vacations, especially when booking flights, it would be helpful for them to have information regarding the likelihood that their flight will be canceled. Customers will have an easier time arranging their travel plans as a result of having access to this information.

The frequency of flight cancellations can be more accurately estimated by calculating the percentage of total flights that were canceled compared to the total number of flights that were not canceled. As a result, we devised a bar graph in order to compare and contrast the various airlines' cancellation policies regarding flights.

Q1. Which airline has the greatest amount of cancelled flights? Why do these Airlines have the highest amount of cancellations, and what can those causes be?



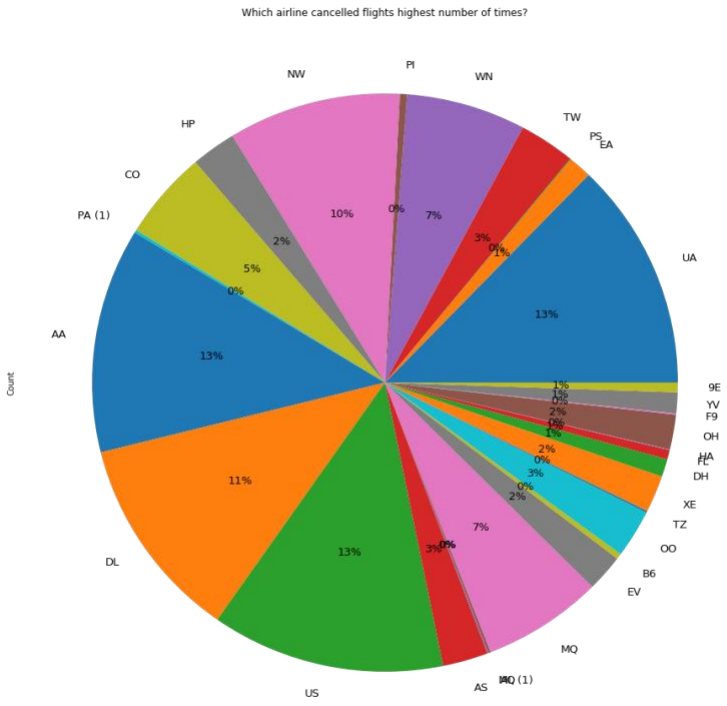
According to the data presented in the preceding graph, United Airlines (UA), American Airlines (US), Delta Air Lines, and US Airways have a significantly higher number of flight cancellations compared to other airlines.

On the other hand, the proportions do not provide a great deal of clarity regarding the flight cancellations. As a result, we devised a pie chart in order to gain a comprehensive understanding of the canceled flights that were offered by the various airlines.

Canceled airlines percentage:

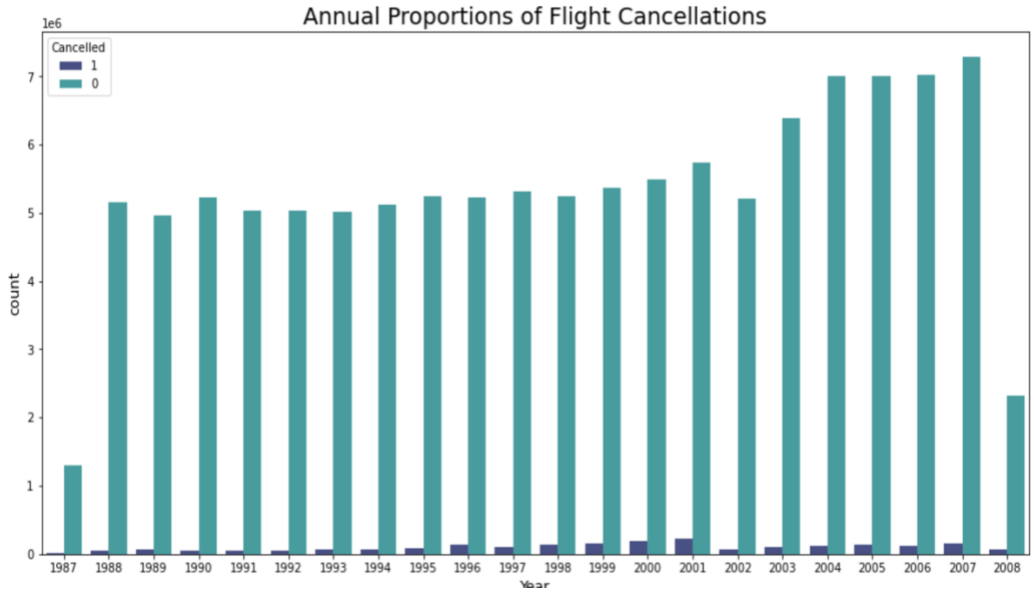
According to the diagram below, US Airlines (US), American Airlines (UA), and united airlines (UA) are each responsible for 13 percent of the total number of canceled flights. Northwest Airlines (NW) and Delta Airlines (DL), when combined, are responsible for 21 percent of all canceled flights. Both Southeast Airlines (WN) and Envoy Airlines (MQ) are responsible for seven percent of the total number of canceled flights. There have been cancellations of five percent of flights for Continental Airlines. And the combined total of all other airlines accounts for a percentage of canceled flights that is less than or equal to 3%.

After doing some research, we found out that all of these airlines blamed the flight cancellations on the inclement weather and the lack of available staff.



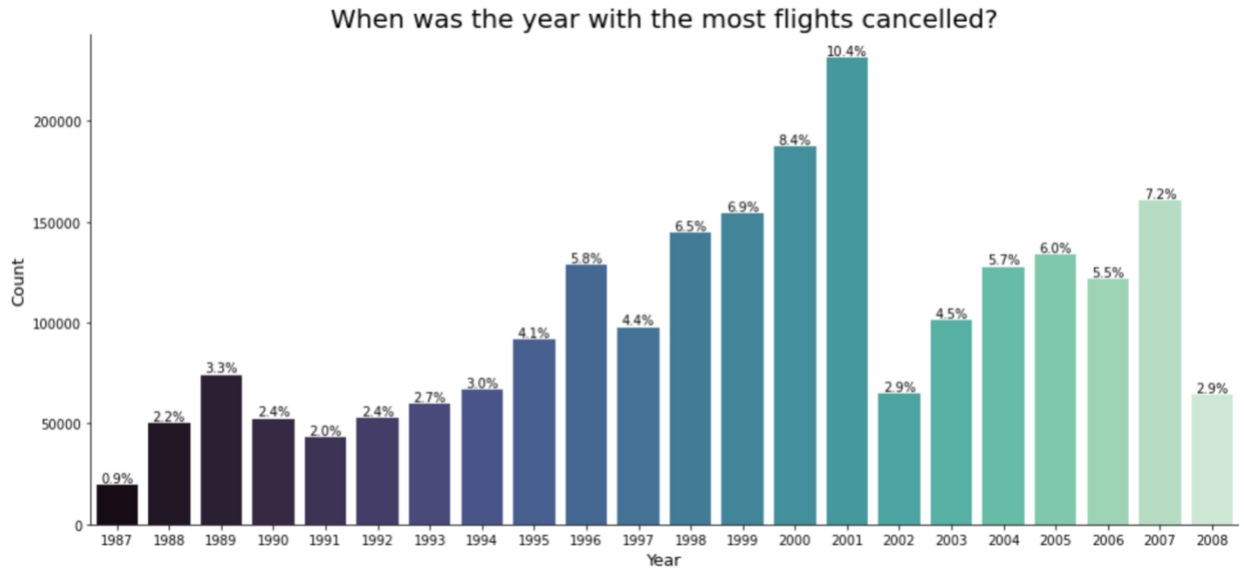
Annual Proportions of Flight Cancellations

To gain a better understanding of the annual flight cancellations that are experienced by all of the registered airlines, we have developed a bar plot visualization.



The graph makes it abundantly clear that, in comparison to the other years, 2001 and 2000 experienced a significantly higher rate of flight cancellations than the other years.

Q2. Which year appears to have had the most events postponed overall, and what may the potential causes be? Is it wise to make decisions only based on data?



We wanted to investigate the different airlines' canceled flights from 1987 to 2008 so that we could determine which year had the clearest trend of having the most canceled flights and

compare the number of flights that were canceled according to the year. A bar chart is the most effective visual representation for this comparison. As a result, a bar graph has been constructed for the purpose of comparison.

According to the graph that was just presented, there was a total cancellation rate of 10.4 percent in the year 2001. In the year 2000, 8.4 percent of flights were canceled, while in the year 2007 only 7.2 percent of flights were canceled. And even considering all of the years combined, fewer than 7 percent of flights were ever scrapped.

Our eagerness to learn the reason behind the cancellation of the vast majority of flights in 2001 cannot be overstated. After doing some research, the fact that hijacked planes hit the World Trade Center on September 11, 2001 should not come as a surprise to anyone. As a direct consequence of this, each and every flight was unexpectedly terminated. In addition, we are aware that Storms, increased aviation traffic, and a lack of new airport runways all contributed to making the year 2000 the Year of the Flight Delay, which in turn led to an increase in the number of flights that were canceled.

Because of the increased demand for aviation—or, to put it another way, the increased traffic—in 2007, approximately 7.2% of flights had to be canceled. This prompted us to investigate the airline's traffic based on the data that was readily available.

Busiest Flights bookings:

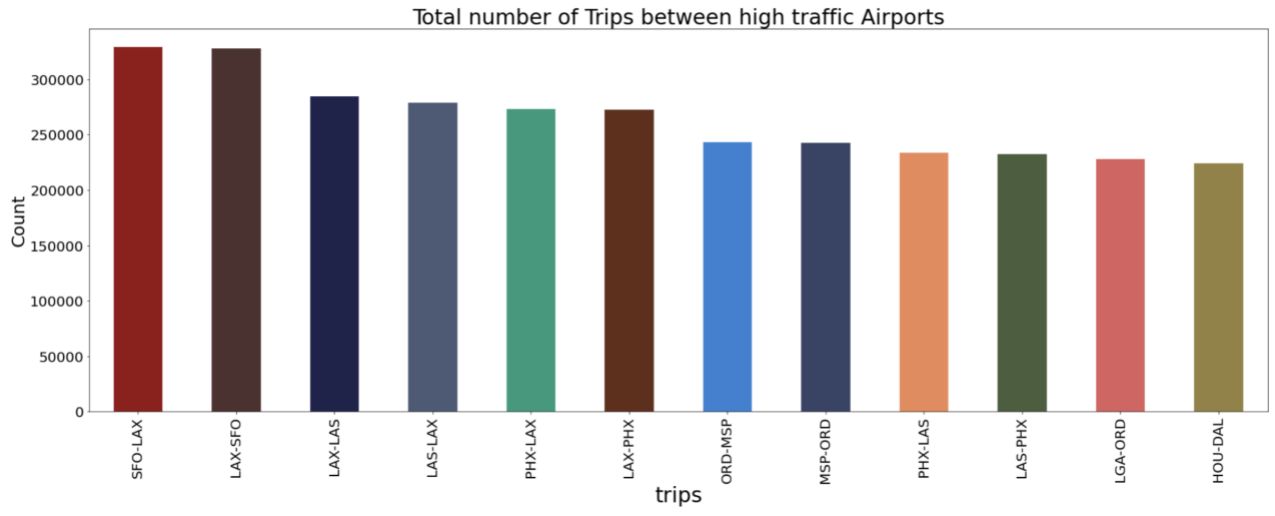
Previous to that, we had read reports of flights being canceled by a variety of airlines between the years 1987 and 2008. It is now time to conduct an investigation into the busiest airlines that operated on the routes that were selected between the years 1987 and 2008. During this step of the process, we answered five questions by using analysis. Within the dataset, we have a significant number of flights taken on airlines. The top twelve high-traffic airline routes that are the primary focus of attention are of particular interest to us. The following is a rundown of the top 12 airports in terms of passenger volume that we have compiled for your convenience.

```
df_journies_sorted = df_journies.sort_values(by=['count'], ascending=False)[:12]
df_journies_sorted.head(10)
```

	Origin	Dest	count
6283	SFO	LAX	329370
4246	LAX	SFO	328105
7245	LAX	LAS	284494
5287	LAS	LAX	278653
5340	PHX	LAX	273286
3151	LAX	PHX	272681
2101	ORD	MSP	243470
7474	MSP	ORD	242933
2196	PHX	LAS	233977
2090	LAS	PHX	232467

It is necessary to compare the trips from one airport to another in order to obtain a comprehensive view of the top twelve high-traffic trips that are shared by the airports. When comparing the results, a bar chart is the most appropriate visual representation. Therefore, in order to visualize the comparison of the trips, we devised a bar graph.

Q 3. Which domestic airline routes in the US had the largest flight traffic, and between which two stations?



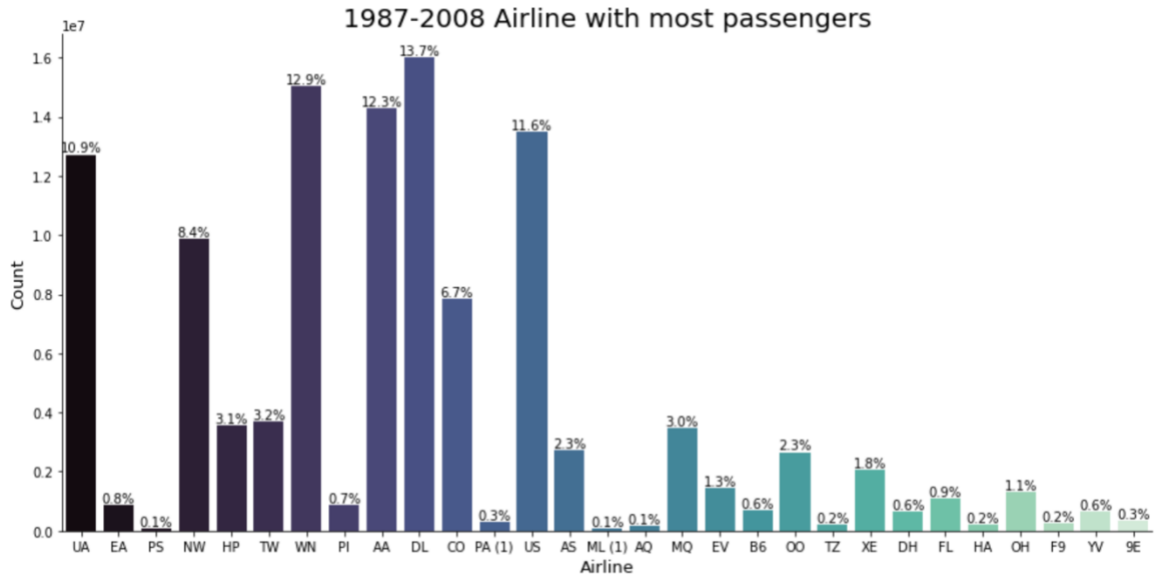
The bar graph that can be seen above displays the names of various trips along the x-axis, while the number of trips is displayed along the y-axis.

It is clear from the chart that the greatest number of flights, approximately 329K, took place between San Francisco International Airport and Los Angeles International Airport, as well as the other way around

Q 4. Which airline is the world's biggest public airline companies by number of passengers carried between 1987 and 2008?

Q 5. Which airlines are the main rivals of the largest publicly traded airline corporations in the world in terms of passenger volume?

In light of the use case, we ought to counsel the customer to select superior airlines for economical travel; however, in order to do so, we require information regarding the airlines that are most popular among passengers. In order to figure this out, I looked at data in the form of a bar plot, which allowed me to see the total number of passengers traveling by airline.



According to the chart that was just presented, "DL (Delta)" airlines have the highest booking percentage, which comes in at 13.7 percent, while "WN (Southwest)" airlines have the second-highest booking percentage, which comes in at 12.9 percent. Let's take a look at the reasons why Delta Airlines had the most passengers of any airline. There is only one explanation for this, and that is because Delta Airlines has a perfect track record of keeping its commitments to its passengers. Delta Airlines provided in-flight amenities such as food and Wi-Fi connectivity, and their fares were significantly lower than those of competing airlines. In addition to this, they have recently implemented the most customer-friendly baggage policy. Delta Airlines was also interconnected with a large number of international airlines, including KLM, Air France, and a number of others. In a similar vein, Southwest is the low-cost domestic carrier that provides the best amenities and the most generous policies regarding baggage.

Q 6. Which year sees the most flight bookings and the aviation industry's boom?

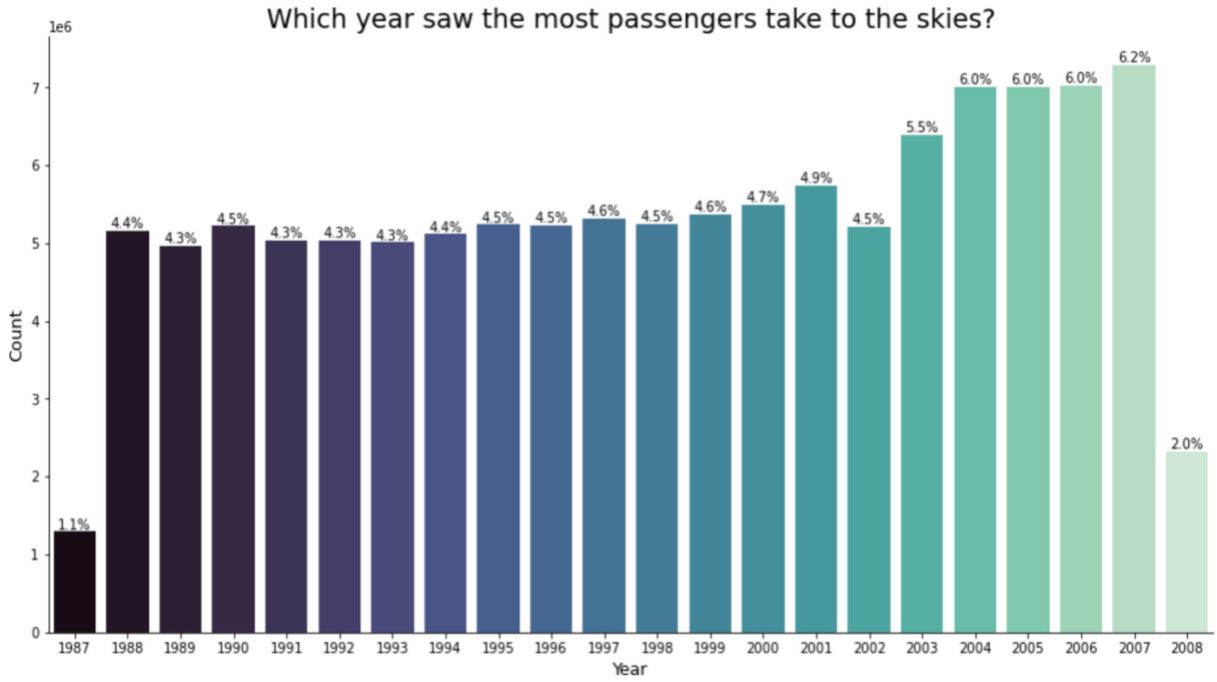
Let's take a look at the total number of passengers who travel each year before we address the question that was just posed, and this applies regardless of the airline that they choose.

```
Year_counts
```

```
Year
1987    1292141
1988    5151933
1989    4967035
1990    5218435
1991    5033420
1992    5039321
1993    5010656
1994    5113308
1995    5235530
1996    5223447
1997    5314080
1998    5240212
1999    5373573
2000    5495557
2001    5736582
2002    5206216
2003    6387071
2004    7001513
2005    7006866
2006    7019988
2007    7292467
2008    2324775
Name: Count, dtype: int64
```

Because it is difficult to determine which year had the most activity based on the information presented in the table above, we decided to present the graph for each year using a bar plot. And

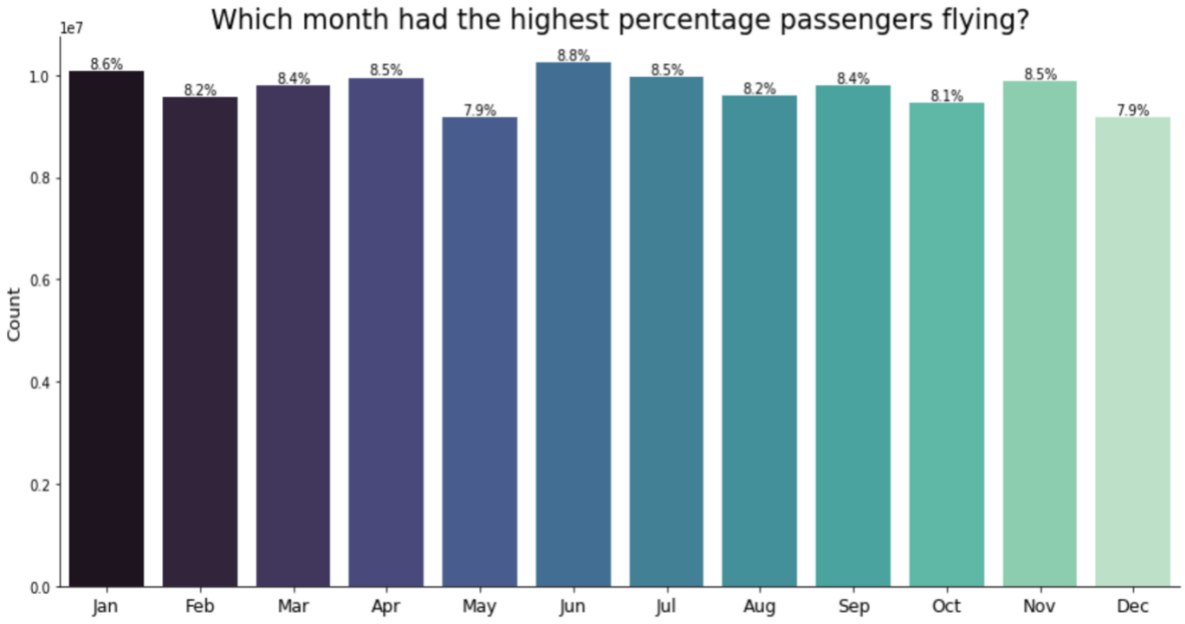
we had provided a percentage breakdown of how much of a share of bookings each year represented in comparison to the years before it.



The preceding graph demonstrates that 2007 was the busiest year, with year bookings accounting for 6.2 percent of the total for the year. Because we do not have sufficient data for the years 1987 and 2008, we are unable to draw the conclusion that these were the years with the lowest volume of activity. 2007 was the busiest year, which was caused by two different things: first, an increase in tourism, and second, the fact that a significant portion of the information technology world was changing rapidly at the time, which required the conduct of the majority of business meetings and client meetings. In addition, the international tourist market reports that there were 56 million more visitors in 2006 than there were in 2007, with 19.6 million of those visitors coming to the United States. This represented a 6 percent increase in the number of tourists that visited the area in comparison to the previous year.

Q 7. Which month of the year do individuals most frequently travel?

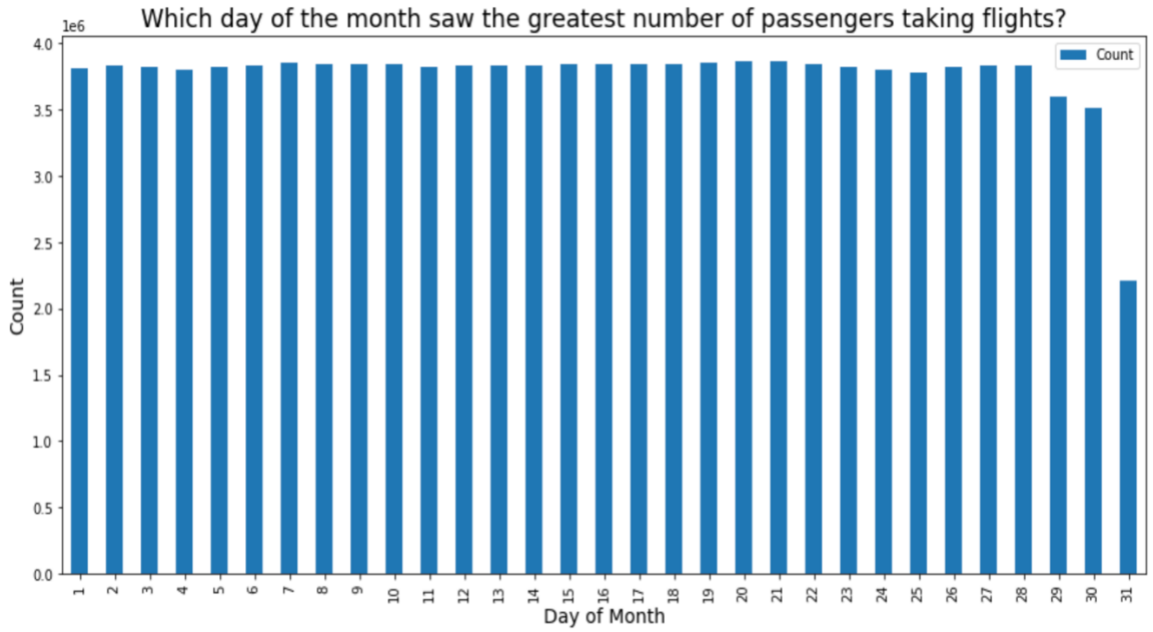
Before we can provide useful feedback to our customers, we need to determine which year was the busiest. As a result, in order to make a recommendation regarding which month is ideal for booking a flight, I made the bar graph shown below, which compares the busiest months from 1987 to 2007.



The data presented in the preceding bar plot indicates that June was the busiest month for flights, with an 8.8 percent share of the total number of passengers who traveled during any given month. Because of tourists, June is the busiest month of the year. June marks the beginning of summer vacation, and during this time of year, the vast majority of people make an effort to travel to unfamiliar locations. We can also see that January and August are the busiest months because a new academic semester begins during those months. During these three months, the majority of passengers on these flights will be students from other countries.

Q 8. Which day of the month saw the highest number of passengers taking flights?

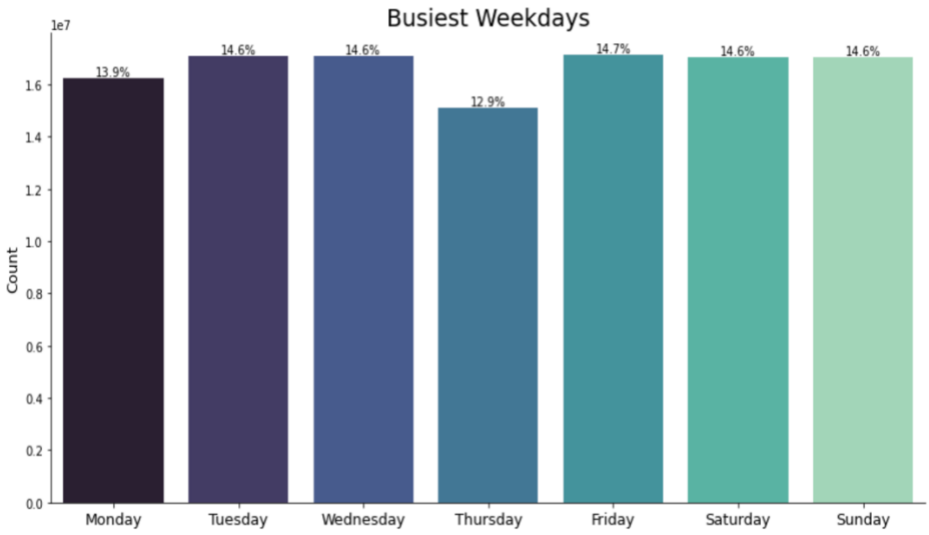
Because some days of the month are busier than others, and some months are not as busy as others, we are unable to provide our customers with guidance that is based on which month is the busiest. Below, a bar plot was presented in order to better illustrate these



According to the data presented in the preceding graph, the typical passenger traveled for the entirety of the month. However, we can see that the end of the month was not nearly as busy because February only goes up to the 29th, which is one of the few months that does not have a 31st. There is no other discernible difference between the days of the month other than that one. Nevertheless, research from "CheapAir" predicts that fares will be more expensive during the holiday season..

Q 9. Which day of the week saw the highest number of passengers taking flights?

If we are aware of the busiest year, month, and day of the week, we are better equipped to provide accurate guidance to our clients. We are able to provide our customers with guidance regarding the day of the week that is most advantageous for them to book a flight on. I visualized the data from weekdays by using a bar plot, which looks like this:



The graph that was just above this one shows that Friday is the busiest day. We are all aware that on Saturdays and Sundays, the majority of businesses and educational institutions are closed. After five o'clock in the evening, most offices and schools were already closed. People who are taking extended vacations typically travel during the weekends so that they can make the most of their time off. Office Needle suggests that passengers schedule their flights for Fridays before 4 p.m., as some airports see more passengers than usual on Friday evenings and Saturday mornings. The demand for flights over the weekends is higher, which results in higher prices..

COMMENTS

PySpark is an application programming interface (API) for Python that was developed by the Apache Spark team in order to integrate Python and Spark. PySpark makes it easy to integrate RDDs into Python programs and to manipulate them. In its role as a framework for the management of massive datasets, PySpark performs an important function. PySpark has been an invaluable tool for us as we work with and compute on large datasets.

Python is a tool that can help you maximize the potential of your data skills. Python is an appealing choice due to its many desirable characteristics. This includes making things easier to understand while simultaneously streamlining the syntax and increasing readability. The fact that Python can be used for both object-oriented and functional programming is perhaps its most appealing quality.

Pros of PySpark:

• PySpark will assist you in achieving faster disk performance at your organization. In most cases, it is ten times faster. In-memory processing is also a hundred times faster than before.

• PySpark facilitates the utilization of RDDs as a fault-tolerant data structure.

• A significant number of essential algorithms are already incorporated into Spark

There are 80 high-level operators available in PySpark's Natural Dynamics module. They will offer assistance in the creation of a parallel application for you to use.

• The framework has a lot of capability when it comes to synchronization points and mistakes.

Cons of PySpark:

• PySpark is frequently considered to be difficult to explain.

• When performing tasks that require a lot of processing power, Python will consequently be slower than Scala.

• When compared to other programs and models, its efficiency is lower than that of the others.

• When it comes to the overall performance of Spark Jobs, Python is typically much slower than Scala. approximately ten times more slowly.

CONCLUSIONS

We were able to gain a better understanding of the statistics and characteristics of the airline of time data as a result of the detailed data analysis that we performed. Based on the analysis of this entire dataset, we have determined that 2007 was the busiest year for flights. After suffering six consecutive years of losses as a direct result of the terrorist attacks that occurred in the United States in September 2001, airline companies saw enormous growth in their profits in 2007. Since 2001, the continuous efforts to cut costs have resulted in a 16 percent reduction in expenses that are not related to fuel. The month of June was the busiest one for flights, making it the most active month overall. This is due to the fact that the first half of the month of June is a relatively cheap time to fly in comparison to the prices that kick in from July, which marks beginning of the summer's peak season. It's common knowledge that Friday is the busiest day of the week for airline travel.

On the other hand, traveling close to or during the weekend is typically more expected and results in a higher volume of passengers. According to the findings of the analysis, the most traveled route between high traffic airports is from SFO to LAX. Overall, Delta Airlines experienced the highest proportion of flight cancellations compared to other airlines. 2001 was the year with the highest number of flight cancellations, with the FAA (Federal Aviation Administration) citing inclement weather as the cause of almost 69 percent of those cancellations. In addition, the FAA reported that thunderstorms caused unusually significant disruptions to air traffic during the spring and summer months..According to our findings, the Delta airlines airline is the one with the most experience in dealing with traffic. Delta is recognized for excellence in a variety of categories, ranging from passenger experience & customer service to operational performance & workplace culture. Free flight changes also offered.

FUTURE WORK: We hope to use the data to learn more about, for example, the best day of the week and time of year to fly in order to avoid delays and the general trend in the number of people traveling between various locations at various times. We would like to monitor the pattern of the number of people traveling between various locations at various times.We are interested in picking up some recent data of Airlines and validate that what changes has been done by the Airlines in recent years. We would also like to analyze the impact of covid-19 on aviation industry.

REFERENCES:

1. *Data Expo 2009: Airline on time data*. (2008, October 6). Harvard Dataverse; dataverse.harvard.edu. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7

2. Feliu, C. (n.d.). *Big Data case study: 5 relevant examples from the airline industry*. Big Data Case Study: 5 Relevant Examples from the Airline Industry; blog.datumize.com. Retrieved June 13, 2022, from https://blog.datumize.com/5-relevant-examples-of-a-big-data-case-study-from-the-airline-industry

3. *5 Real-World Problems Big Data Can Solve*. (n.d.). Techopedia.Com; www.techopedia.com. Retrieved June 13, 2022, from https://www.techopedia.com/2/29524/technology-trends/big-data/5-real-world-problems-big-data-can-solve

4. *Digging Deep, Flying High: The Airline Of The Future Will Run On Big Data | GE News*. (2019, February 18). Digging Deep, Flying High: The Airline Of The Future Will Run On Big Data | GE News; www.ge.com. https://www.ge.com/news/reports/digging-deep-flying-high-airline-future-will-run-big-data

5. Helsen, S. (2020, August 14). *The 3 Big Challenges with Big Data in Aviation - Sensing Change Blog*. Sensing Change Blog; blog.hexagongeospatial.com. https://blog.hexagongeospatial.com/the-3-big-challenges-with-big-data-in-aviation/

6. Zamiatina, A. (n.d.). *9 incredible ways data analytics is transforming airlines*. 9 Incredible Ways Data Analytics Is Transforming Airlines; blog.datumize.com. Retrieved June 13, 2022, from https://blog.datumize.com/9-incredible-ways-data-analytics-is-transforming-airlines

7. *Big Data Processing with Apache Spark – Part 1: Introduction*. (2022, June 21). InfoQ; www.infoq.com. https://www.infoq.com/articles/apache-spark-introduction%20/

8. Rane, Z. (n.d.). *11 Best Companies to Work for as a Data Scientist*. www.stratascratch.com. Retrieved June 27, 2022, from https://www.stratascratch.com/blog/11-best-companies-to-work-for-as-a-data-scientist/

9. Russom, P. (2011). Big data analytics. TDWI best practices report, fourth quarter, 19(4), 1-34.

10. Van Rossum, G. (2007, June). Python Programming language. In USENIX annual technical conference (Vol. 41, No. 1, pp. 1-36).