

**EDA PROJECT**

**REPORT**

on

**USA Airlines Delay Cause**

Submitted by

**Sanjay Neelarapu**

**Registration Number : 12109553**

**Program Name : B.Tech Data Science(ML and AI)**

Under the Guidance of

**Anjana Madam**

**School of Computer Science & Engineering Lovely Professional University, Phagwara**

# Table Of Contents

|  |  |  |
| --- | --- | --- |
| **SI.No** | **Contents** | **Page.No** |
| **1** | **Introduction** | **3** |
| **2** | **Domain Knowledge** | **4-5** |
| **3** | **Libraries Used** | **6** |
| **4** | **Data Description** | **7-8** |
| **5** | **Data Cleaning** | **9-11** |
| **6** | **Multivariate Analysis** | **12-13** |
| **7** | **Univariate Analysis** | **14-20** |
| **8** | **Bivariate Analysis** | **21-23** |
| **9** | **Statistical Analysis** | **24-27** |
| **10** | **Observations** | **28-29** |
| **11** | **Conclusions** | **30** |
| **12** | **GitHub Link** | **30** |
| **13** | **Kaggle Link** | **30** |

**Introduction:**

My EDA Project is on **“USA Airlines Delay Cause”.** Now a day’s airways have become the most common transport across the world as it the fastest of all other mode of transports and hence the demand for this transport is increasing day by day. People choose this mode of transport as it is the fastest way, so delay in the flights may disappoint the passengers. This is the main reason why I chose this dataset for my project as I want to find which is main reason for most of the delays in the airlines.

The airlines report the cause for the delay based on five categories:

**1.Air Carrier:** Delay due to circumstances within the airlines.

**2.Extreme Weather:** Delay of flight due to some weather conditions such as Hurricanes, Tornado, etc.

**3.National Aviation System (NAS):** Delay due to some conditions such as non-extreme weather conditions, airport operations, heavy traffic volume and air traffic control.

**4.Late-Arriving Aircraft:** A previous flight with same aircraft arriving late, causing the present flight to depart late.

**5.Security:** Delay of aircraft due to some security breach.

An Aircraft is said to be delayed when it arrives late for 15 minutes or more than the scheduled time of arrival and the delay time is calculated by the number of minutes after the flight was considered as delayed (i.e., after 15 minutes of scheduled tie of arrival).

The Dataset has the information about the airport name, Carrier name, reasons for the Carrier delay, the year and month of delay of the airline and the number of minutes the flight has got delayed under each cause of those flights which got delayed from the year 2003 to 2022. Each flight has the multiple causes for delay and the number of minutes has been given under each cause that is responsible for delay.

The dataset has 318017 rows and 21 columns. Out of 21 columns 4 columns are of String type, 2(year and month) are of int type and remaining 15 are of float type. The 15 columns which are of float type are those which are causes of delay of the aircraft and the columns have the number of minutes of delay.

**The Insights I want to find from this dataset are:**

1.Which year has the greatest number of delays and in that year which airport and the airlines has the most and least delays.

2.Which Aircraft has delayed for the max time and the cause for the delay.

3.What is the average time of delay for every airline under each cause.

4.Which airport reports the most number of delays on each year.

5.What is cause that is responsible for most number of delays for each airline.

6.Which airlines recorded the least number of delays in each year.

**Domain Knowledge:**

Airline delay EDA (exploratory data analysis) is a project that uses data to understand and analyze the causes of airline delays. This type of project can be used by airlines, airports, and other stakeholders to identify patterns and trends in delay data, and to develop strategies to reduce delays.

Airline delay EDA projects can be used to answer a variety of questions, such as:

* What are the most common causes of airline delays?
* Which airlines are most likely to experience delays?
* Which airports are most likely to have delays?
* What are the peak times of day for airline delays?
* How does the weather affect airline delays?
* How do other factors, such as holidays and special events, affect airline delays?

To answer these questions, airline delay EDA projects typically involve the following steps:

1. Data collection: The first step is to collect a dataset of airline delay data. This data can be obtained from a variety of sources, such as the Bureau of Transportation Statistics (BTS) in the United States, or the Civil Aviation Authority (CAA) in the United Kingdom.
2. Data cleaning: Once the data has been collected, it needs to be cleaned and prepared for analysis. This may involve removing duplicate records, correcting errors in the data, and filling in any missing values.
3. Exploratory data analysis: Once the data has been cleaned, it can be analyzed using a variety of statistical and visualization techniques. This analysis can be used to identify patterns and trends in the data, and to develop hypotheses about the causes of airline delays.
4. Data interpretation: Once the data has been analyzed, the results need to be interpreted and communicated to stakeholders. This may involve writing a report or creating a presentation that highlights the key findings of the analysis.

**Applications of airline delay EDA projects**

Airline delay EDA projects can be used by a variety of stakeholders to improve the Performance of the airline industry. For example:

* Airlines can use the results of airline delay EDA projects to identify areas where they can improve their operations and reduce delays.
* Airports can use the results of airline delay EDA projects to improve their infrastructure and reduce delays.
* Government agencies can use the results of airline delay EDA projects to develop policies and regulations that reduce delays.
* Passengers can use the results of airline delay EDA projects to make more informed decisions about their travel plans.

**Libraries Used and Approaches:**

The following are some of the libraries and approaches that can be used for flight delay EDA.

**Python:** Python is a popular programming language for data science and machine learning. It has a number of libraries that can be used for flight delay EDA, such as:

**Pandas:** Pandas is a library for data manipulation and analysis. It can be used to clean and prepare the data, and to perform exploratory data analysis.

**Matplotlib:** Matplotlib is a library for data visualization. It can be used to create charts and graphs that illustrate the patterns and trends in the data.

**Seaborn:** Seaborn is a built-on-top-of-Matplotlib library that provides a high-level interface for statistical data visualization. It is often used to create more complex and informative charts than Matplotlib.

**Approaches**

The following are some of the approaches that can be used for flight delay EDA:

Descriptive statistics: Descriptive statistics can be used to describe the basic characteristics of the data, such as the mean, median, mode, standard deviation, and quartiles. This can be helpful in understanding the overall distribution of the data and identifying any outliers.

Data visualization: Data visualization can be used to identify patterns and trends in the data. For example, you could create charts and graphs to show how flight delays vary by airline, airport, time of day, and day of the week.

Correlation analysis: Correlation analysis can be used to measure the strength and direction of the relationship between two variables. For example, you could use correlation analysis to see how flight delays are correlated with weather conditions, airline load factors, and air traffic control delays.

**Data Description:**

The Airlines Delay project dataset is a collection of flight delay data for 18 airlines in the United States. The data was collected from year 2003 to year 2022 from the Bureau of Transportation Statistics (BTS). The dataset contains over 8 lakh records, each of which represents a summary of delay data for a carrier-airport pair for the specified month.

The dataset includes the following variables:

The marketing carrier networks are:  
Alaska Airlines (AS)\*  
Allegiant Air (G4)  
American Airlines (AA)\*  
Delta Air Lines (DL)\*  
Frontier Airlines (F9)  
Hawaiian Airlines (HA)\*  
JetBlue Airways (B6)  
Southwest Airlines (WN)  
Spirit Airlines (NK)  
United Airlines (UA)\*

The reporting airlines are:  
Alaska Airlines (AS)  
Allegiant Air (G4)  
American Airlines (AA)  
Delta Air Lines (DL)  
Endeavor Air (9E)  
Envoy Air (MQ)  
Frontier Airlines (F9)  
Hawaiian Airlines (HA)  
Horizon Air (QX)  
JetBlue Airways (B6)  
Mesa Airlines (YV)  
PSA Airlines (OH)  
Republic Airlines (YX)  
SkyWest Airlines (OO)  
Southwest Airlines (WN)  
Spirit Airlines (NK)  
United Airlines (UA)

The airlines report the causes of delays in five broad categories:

* Air Carrier: The cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).
* Extreme Weather: Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane.
* National Aviation System (NAS): Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.
* Late-arriving aircraft: A previous flight with same aircraft arrived late, causing the present flight to depart late.

Security: Delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.

**DATA CLEANING:**

Data Cleaning is the process of cleaning the data, cleaning data in the sense removing duplicate values, dealing with null values, and removing the rows and columns that are not needed for our analysis.

So, in the process of Data Cleaning I checked for null values in my dataset.

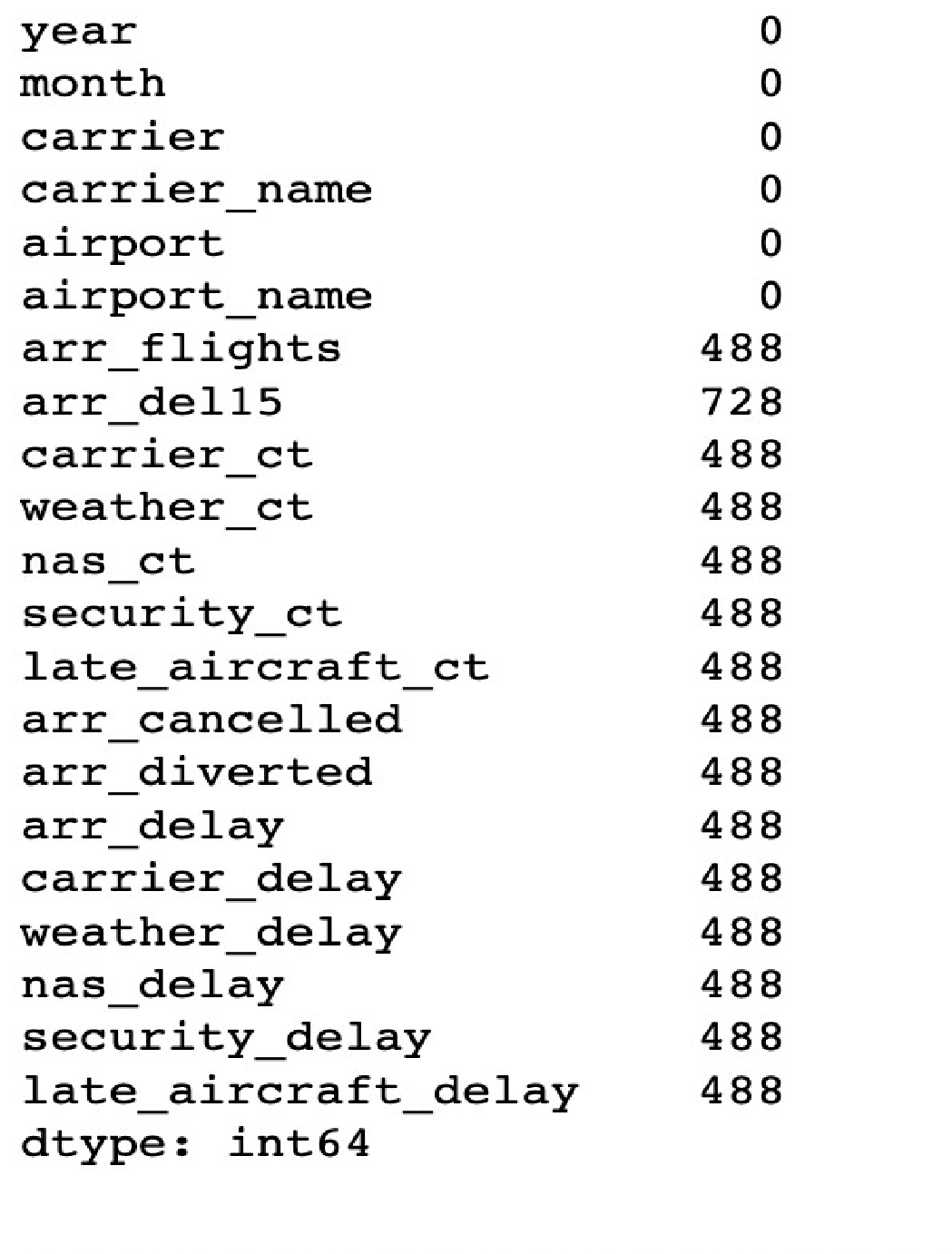


Fig (1)

And here in the above image (Fig (1)) we can see that on the left side there are names of the columns and on right side we have the number of null values corresponding to each column in the dataset. And from the above image (Fig (1)) we can infer that there are 0 null values in year, month, carrier, carrier\_name, airport, airport\_name columns. And there are some null values in the following columns: arr\_flights, arr\_del15, carrier\_ct, weather\_ct, nas\_ct, security\_ct, late\_aircraft\_ct, arr\_cancelled, arr\_diverted, arr\_delay, carrier\_delay, weather\_delay, nas\_delay, security\_delay, late\_aircraft\_delay.

So, to proceed for the further process, I need to know the percentage of null values for those columns which have the null values.

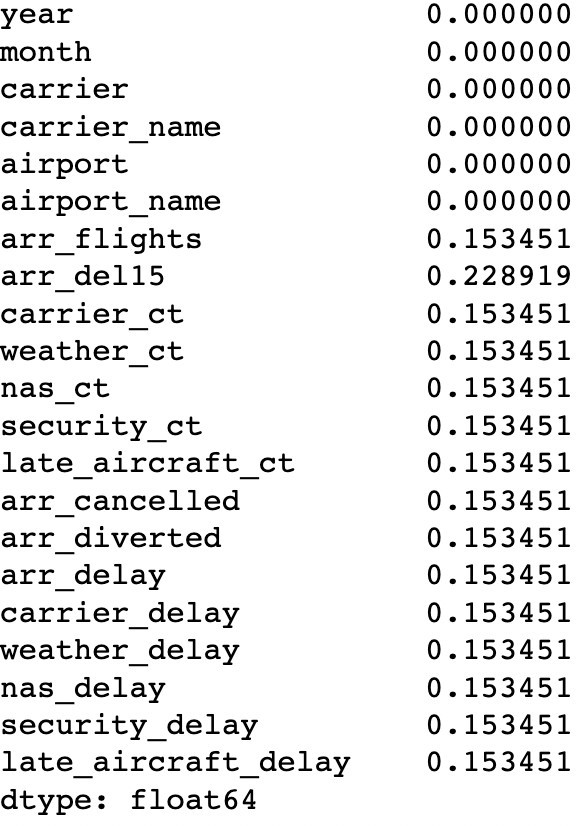


Fig (2)

The above table (Fig (2)) gives the percentage of null values for each column in the dataset. And as we can see that the null value percentage is less than 1% in every column which has null values, and the null value percentage is almost negligible.

The null values can be dealt in 2 ways. We can either delete the rows which has the null values or replace the null values with mean, median or mode values. As the dataset has negligible null values, I have removed the rows which has the null values.

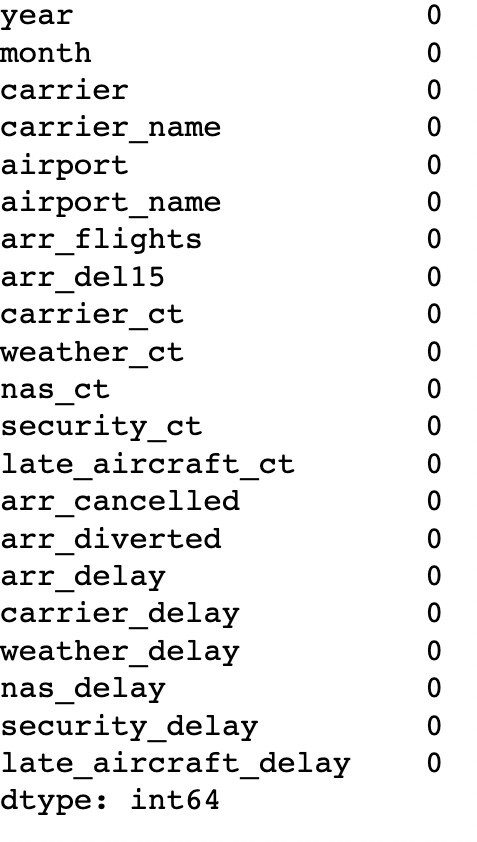


Fig (3)

As we can see that in the above table (Fig (3)) there are 0 null values in every column. That means all the null values are dropped. And now the dataset is clean so, we can proceed with the further analysis.

**Multivariate Analysis:**

Multivariate Analysis is the process of performing analysis based on more than 2 variables.

**Correlation between the variables using heatmap.**

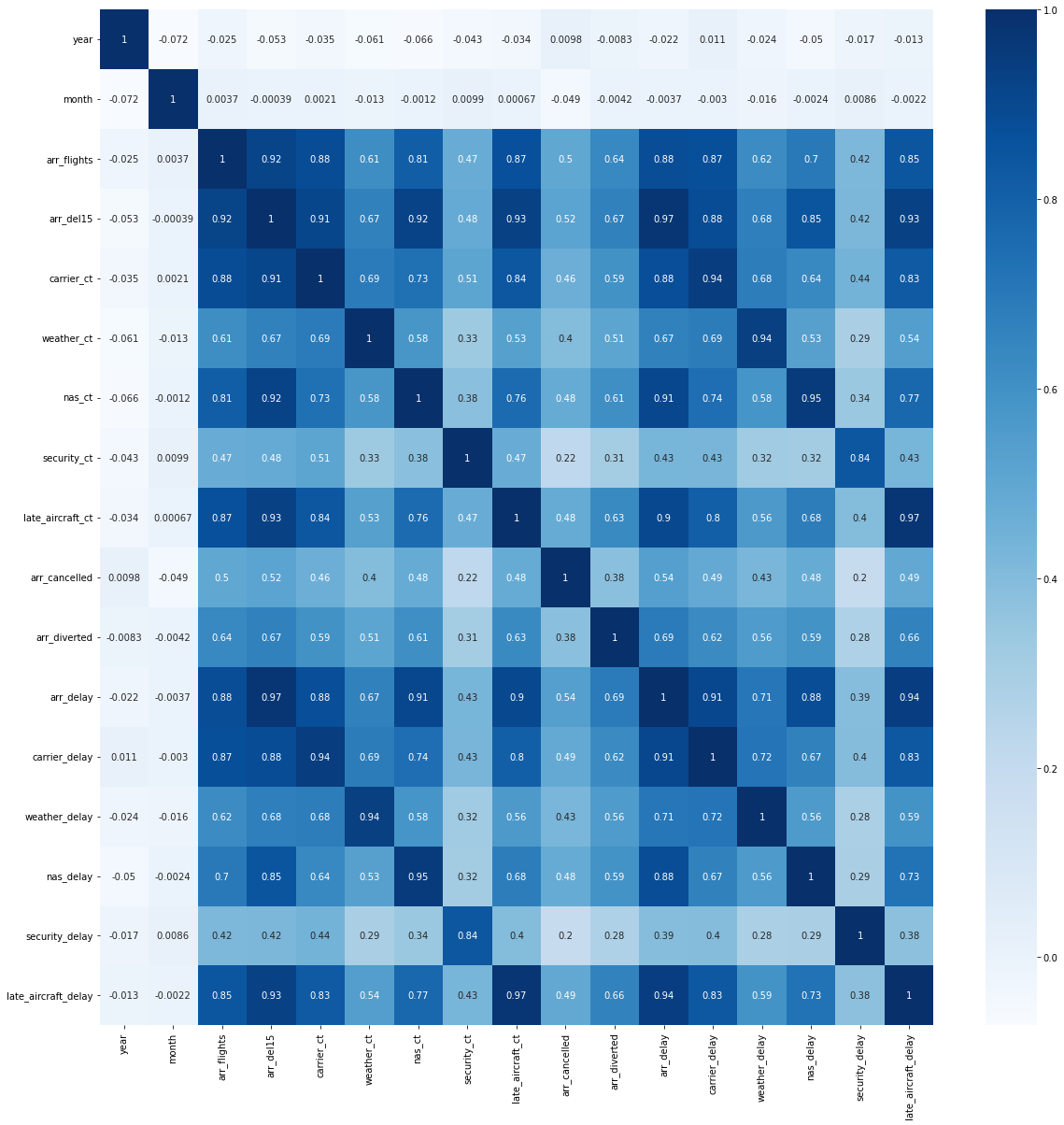


Fig (4)

**Observation:**

1.(arr\_del15,arr\_delay), (carrier\_ct,carrier\_delay) , (weather\_ct,weather\_delay)

,(nas\_ct,nas\_delay), (security\_ct, security\_delay), (late\_aircraft\_ct,late\_aircraft\_delay) have correlation value almost equal to 1.

If the correlation value is equal 1 then we can say that the two variables are equal. So, I started my analysis with arr\_delay, carrier\_delay, weather\_delay, nas\_delay, security\_delay, late\_aircraft\_delay columns.

* carrier\_delay is delay time in minutes caused due to circumstances within airlines.
* weather\_delay is delay time in minutes caused due to weather issues.
* nas\_delay is delay time in minutes caused due to some conditions such as airport operations, heavy traffic volume and air traffic control.
* security\_delay is delay time in minutes caused due to some security breach.
* late\_aircraft\_delay is delay time in minutes caused when a previous flight with same aircraft arriving late, causing the present flight to depart late.
* arr\_delay is the total delay time in minutes.

2.Also arr\_delay is highly correlated with carrier\_delay (correlation value = 0.91) , nas\_delay (correlation value = 0.88), late\_aircraft\_delay (correlation value = 0.94). That means when these three values are increasing arr\_delay also gets increased. These three are the main causes that are causing for the delay.

**Univariate Analysis:**

Univariate Analysis is the process of performing analysis based on single variable.

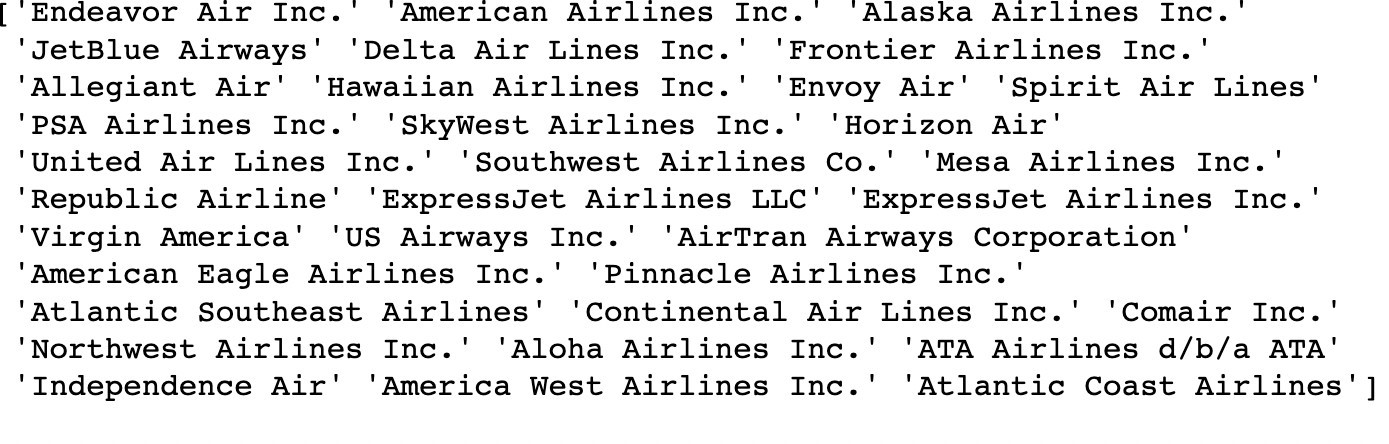


Fig (5)

The above image (Fig (5)) represents the unique values that are in the carrier\_name columns. These are the names of the airlines whose flights might have delayed or cancelled. The dataset has the about 33 airlines whose flights were delayed.

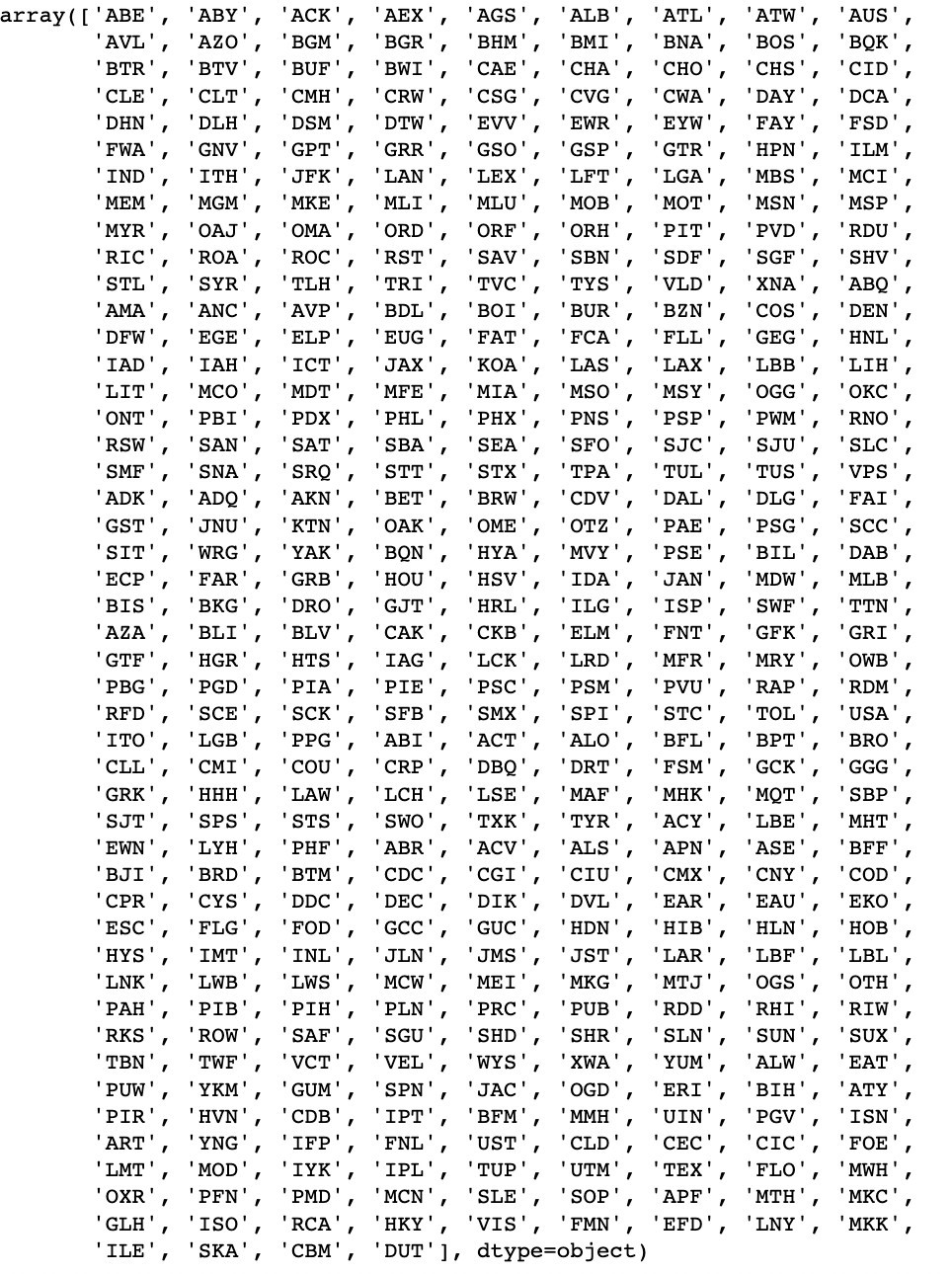


Fig (6)

The above names in the image (Fig (6)) are the names of the airports to which the delayed flights have arrived. The dataset has the information about 418 airports.

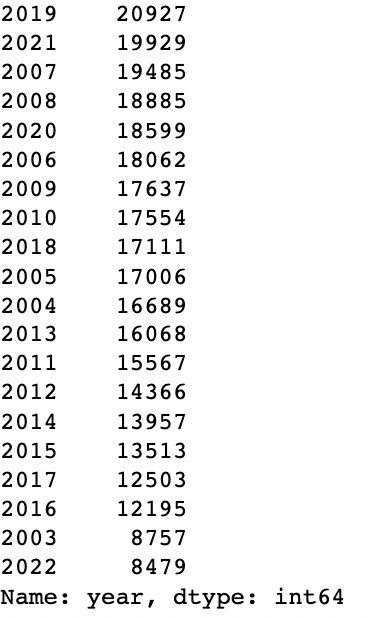


Fig (7)

The above image (Fig (7)) gives the information about how many number of times each year has repeated in the dataset. The count of each year is the count of the number of flights that were delayed or cancelled.

2019 is the year which has highest number of delays which is 20927 flights and 2022 has least number of delays which is 8757 flights.

The below is the graphical representation of the same. X-axis represents the year and Y-axis represents Number of delays.

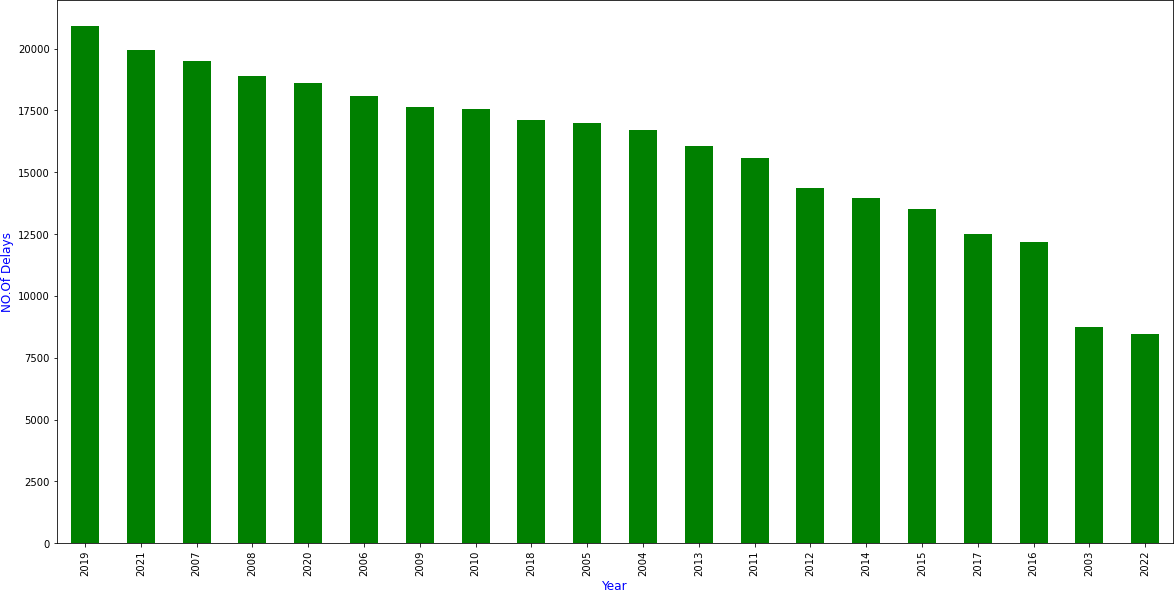


Fig (8)

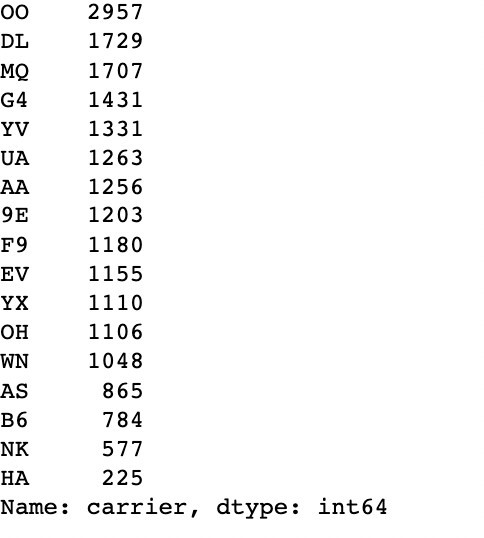


Fig (9)

The above figure (Fig (9)) gives the information about the airlines that were delayed in the year of 2019, this is the year in which highest number of delays were recorded. The number on the right in the figure gives the information of number of flights that got delayed or cancelled for each airline.

SkyWest Airlines (OO) is the airlines which has recorded highest number of delays in the year 2019 and Hawaiian Airlines (HA) is the airlines which has recorded lowest number of delays in the year 2019.

The below image (Fig (10)) is the graphical representation of the same.

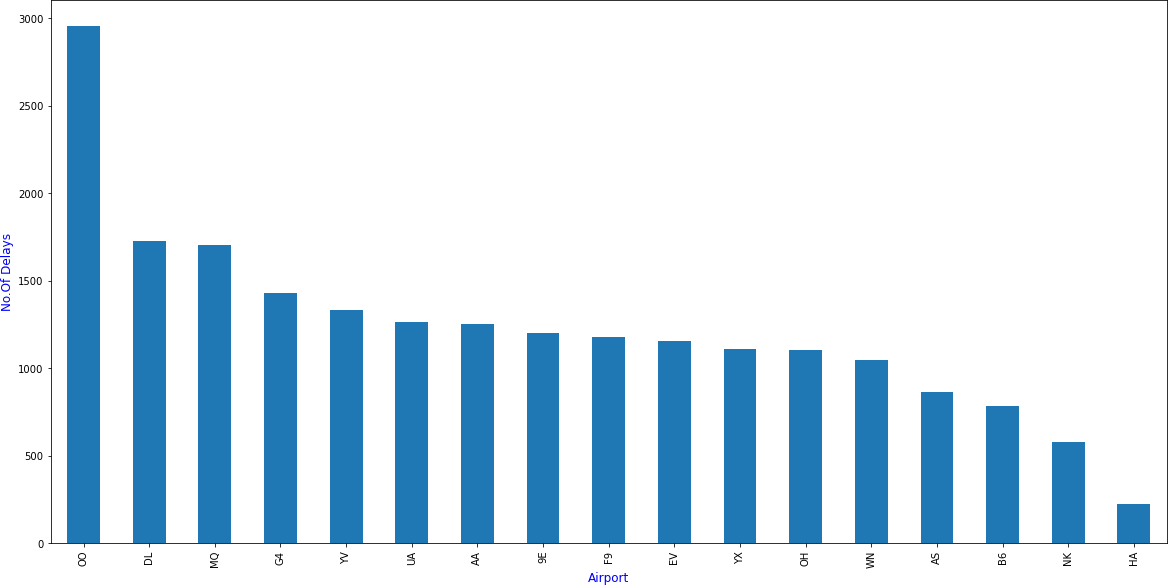


Fig (10)

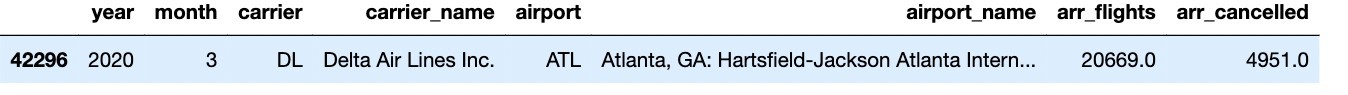


Fig (11)

The above image (Fig (11)) gives the information about the airlines that has the highest number of flight cancellations from the year 2003-2022.

Delta Air Lines Inc. (DL) is the airlines that has the highest number of cancelled flights in the month of 3(March) in the year 2020. Out of 20669 flights 4951 flights were cancelled.



Fig (12)

The above image (Fig (12)) gives the information about the airline that has delayed for highest number of minutes due to weather delay cause.

Atlantic Southeast Airlines (EV) is the airlines that has delayed for maximum time among all other flights from the year 2003-2022. It was delayed in the month of 7(July) in the year 2007. It has delayed for 57707 minutes; the delay might be due to Rainy conditions as July is the starting month for Rainy season.

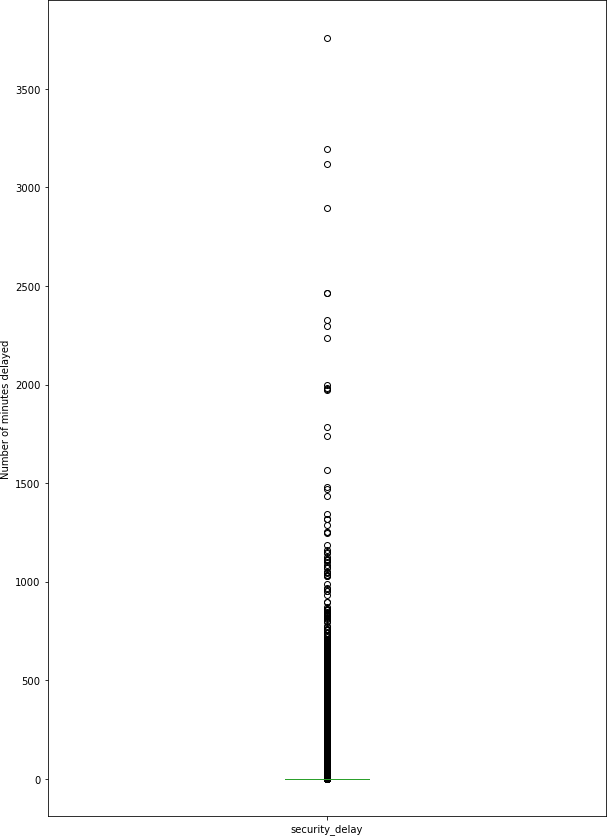


Fig (13)

The above image (Fig (13)) is the boxplot for the security\_delay column. The numbers on the Y-axis is the number of minutes delayed. From the above Boxplot we can infer that there was only one airline that has delay time above 35000 minutes due to security delay. Most of the airlines has delay time between 0 to 1000.

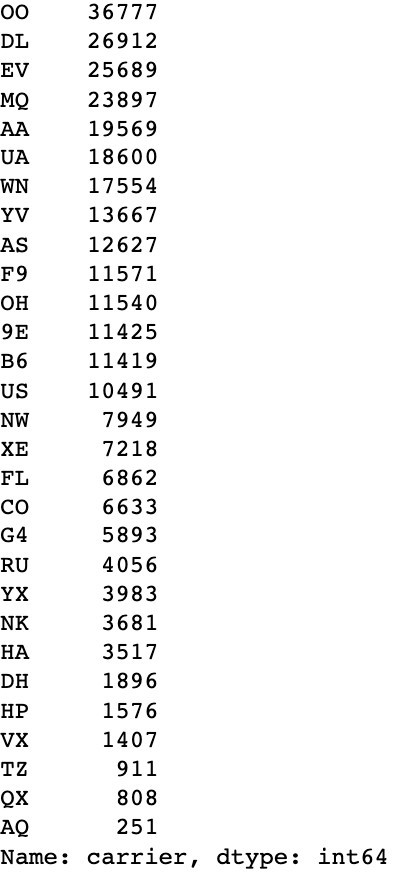


Fig (14)

The above table (Fig (14)) gives the information about the number of delays that each airline has recorded form the year 2003-2022. SkyWest Airlines (OO) is the airlines that has highest number of delays among all, and this is the same airlines that has highest delays in the year 2019 too.

The below is the graphical representation of the same.

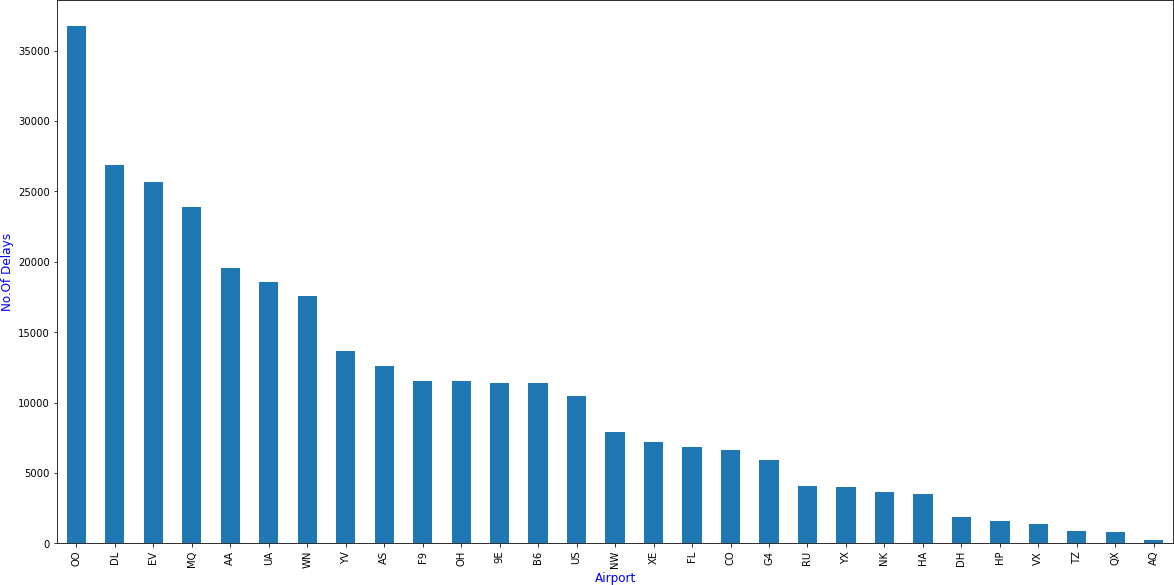


Fig (15)

AQ is the airline that has least number of delays.

**Bivariate Analysis:**

Bivariate analysis is the process of performing analysis based on two variables.

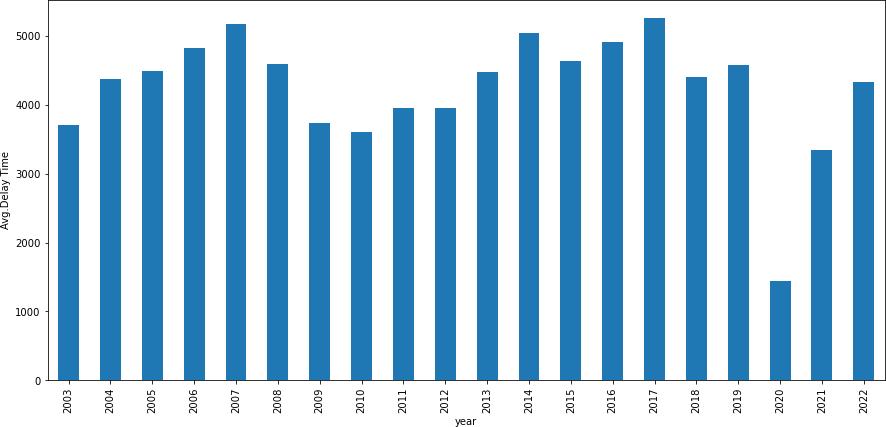


Fig (16)

The above Bar Graph (Fig (16)) has plotted by taking year on x-axis and average of arr\_delay on y-axis. From the above figure we can get to know that 2017 and 2007 are the years that have highest average delay time.2020 is the year that has least average delay time.

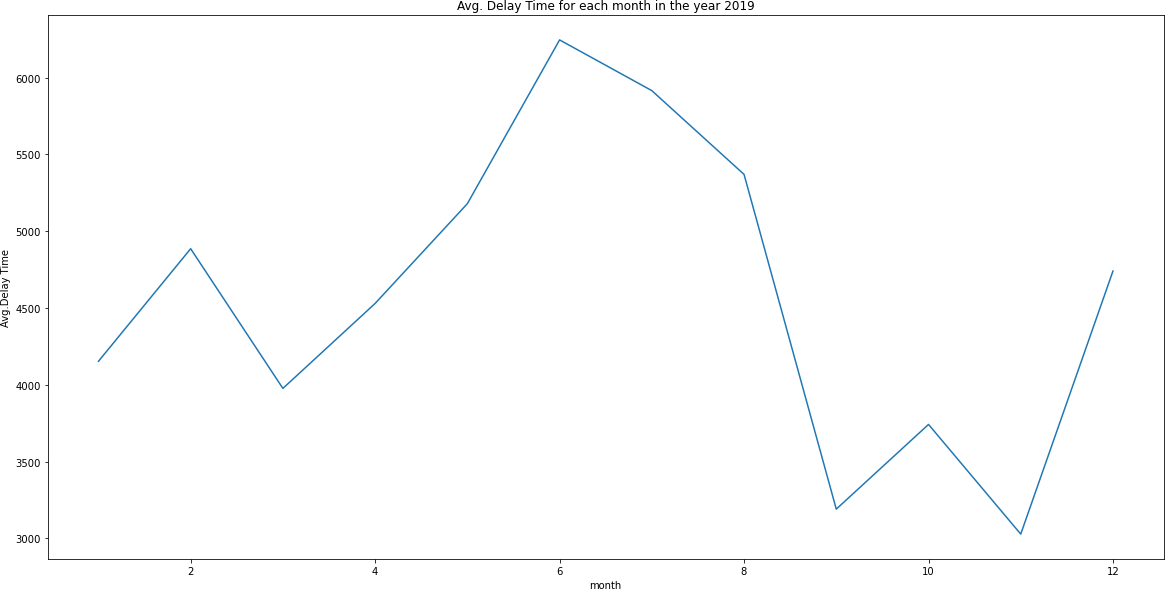


Fig (17)

The above line graph (Fig (17)) was plotted by taking month on x-axis and average of arr\_delay on y-axis. From the graph we can infer that the average delay has started increasing from the month 3 to 6 and this is the summer season in which most of the people travel for trips.

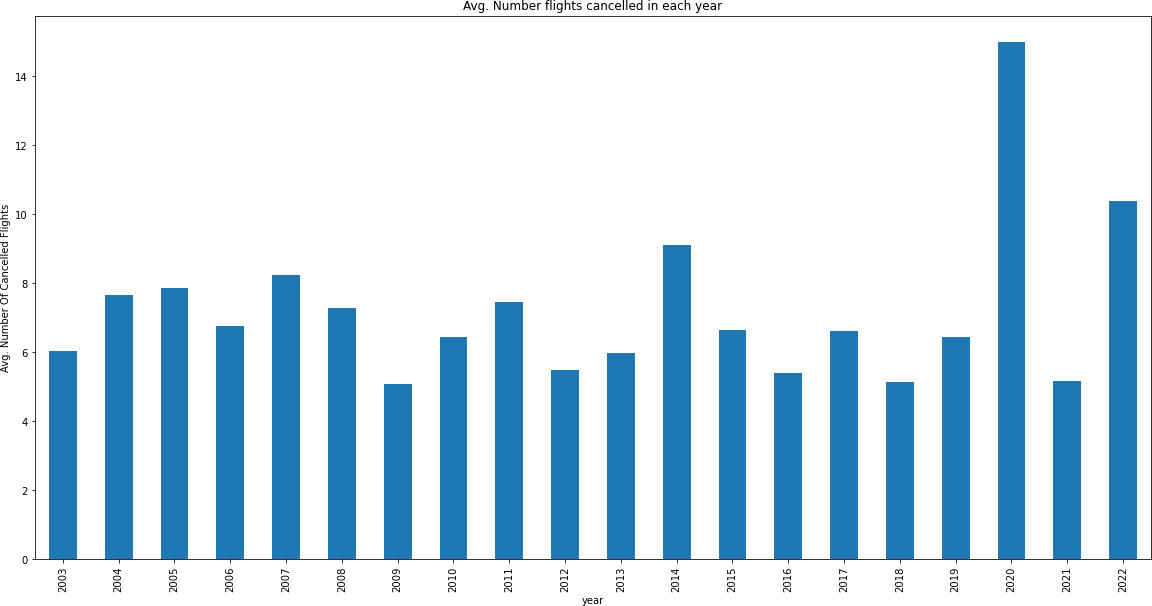


Fig (18)

The above image (ig (18)) is the Bar Graph that was plotted by taking year on x-axis and average of arr\_cancelled on y- axis.

2020 is the year that has highest number of cancelled flights, the reason might be in this year Corona outbreak took place and most of the countries have implemented lockdown.

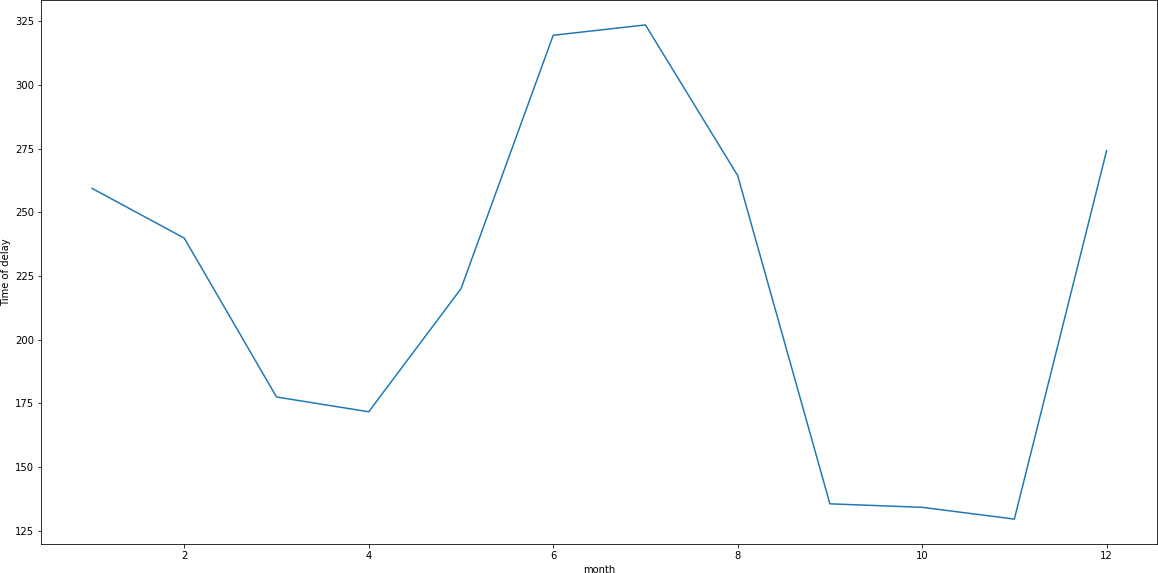


Fig (19)

The above Line Graph (Fig (19)) was plotted by taking month on x-axis and weather\_delay on y-axis.

The graph has started increasing from April month and went to peeks in the month of June and July which is rainy season.

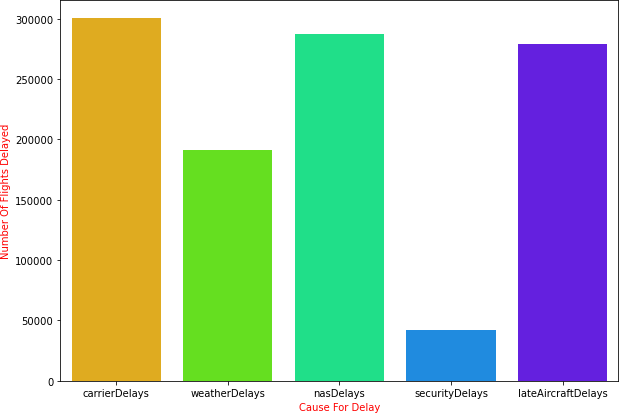


Fig (20)

The above Bar Graph (Fig (20)) information about which cause of delay has impacted for most number of delays. X-axis of the graph give the cause for delay and Y-axis gives the number of flights delayed.

Most of the flights were delayed due to Carrier Delay and NAS Delay.

**Statistical Analysis of Numerical Columns:**

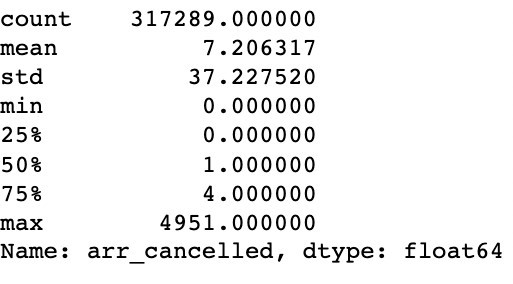


Fig (21)

The above image (Fig (21)) contains the statistical summary of arr\_cancelled column.

* The minimum number of flights that were cancelled are 0.
* The maximum number of flights that were cancelled are 4951.
* The average number of flights that were cancelled are 7.
* 50% is the median of the column that is 1.

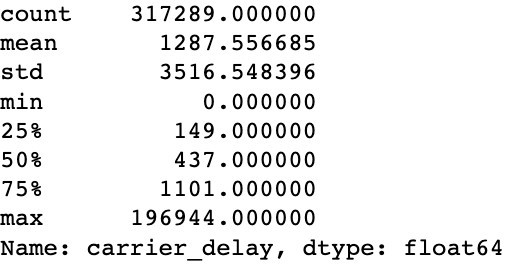


Fig (22)

The above image (Fig (22)) contains the statistical summary of carrier\_delay column.

* The minimum number of minutes that an airline has delayed due to carrier delay is 0 mins.
* The maximum number of minutes that an airline has delayed due to carrier delay is 196944 mins.
* The average number of minutes that an airline has delayed due to carrier delay is 1287 mins.
* The median value of the carrier\_delay column is 437 mins.

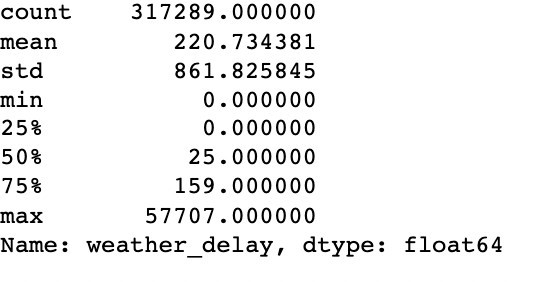


Fig (23)

The above image (Fig (23)) contains the statistical summary of weather\_delay column.

* The minimum number of minutes that an airline has delayed due to weather delay is 0 mins.
* The maximum number of minutes that an airline has delayed due to weather delay is 57707 mins.
* The average number of minutes that an airline has delayed due to weather delay is 221 mins.
* The median value of the weather\_delay column is 25 mins.

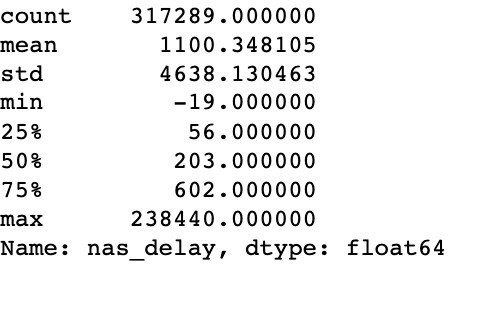


Fig (24)

The above image (Fig (24)) contains the statistical summary of nas\_delay column.

* The minimum number of minutes that an airline has delayed due to NAS delay is -19 mins.
* The maximum number of minutes that an airline has delayed due to NAS delay is 238440 mins.
* The average number of minutes that an airline has delayed due to NAS delay is 1100 mins.
* The median value of the nas\_delay column is 203 mins.

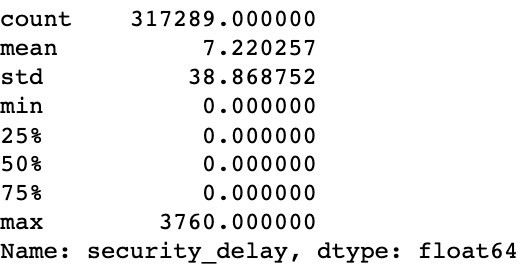


Fig (25)

The above image (Fig (25)) contains the statistical summary of security\_delay column.

* The minimum number of minutes that an airline has delayed due to Security delay is 0 mins.
* The maximum number of minutes that an airline has delayed due to Security delay is 3760 mins.
* The average number of minutes that an airline has delayed due to Security delay is 7 mins.
* The median value of the security\_delay column is 0 mins.

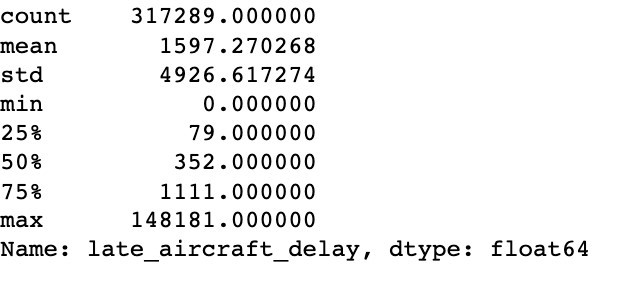


Fig (26)

The above image (Fig (26)) contains the statistical summary of late\_aircraft\_delay column.

* The minimum number of minutes that an airline has delayed due to Late Aircraft delay is 0 mins.
* The maximum number of minutes that an airline has delayed due to Late Aircraft delay is 148181 mins.
* The average number of minutes that an airline has delayed due to Late Aircraft delay is 1597 mins.
* The median value of the late\_aircraft\_delay column is 352 mins.

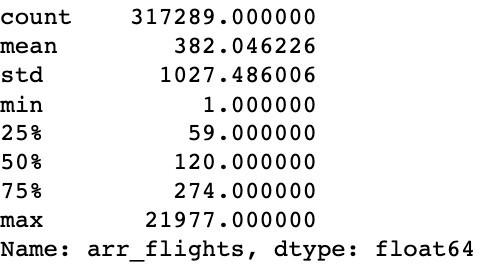


Fig (27)

The above image (Fig (27)) contains the statistical summary of arr\_flights column.

* The minimum number of flights arrived is 1.
* The maximum number of flights that are arrived are 21977.
* The average number of flights arrived are 382.
* The median value of the arr\_flights column is 120.

**Observations:**

1. carrier\_name has 33 unique values. Among those 33 airlines Southwest Airlines (WN) and SkyWest Airlines (OO) has highest number of delays.

1. year has 20 unique values. From year 2003 to 2022. 2019 has maximum number of delays or cancellations.
   * + In the year 2019 July month had highest average delay time.
     + SkyWest Airlines is the one that had most delays in the year 2019.

1. Maximum time of delay caused due to carrier delay is 196944 mins. The average delay time is 1288 mins. Delta Air Airlines is the carrier that has maximum delay time due to carrier delay. This delay was recorded in 2016 8th month.

1. In the year 2007 7th month Atlantic Southeast Airlines has delayed for the maximum time about 57707 minutes due to weather delay.

1. Most of the airlines which has delayed due to security beach have delayed for 0 to 500 minutes.

1. American Airlines is the one which has delayed for maximum time for about 148181 minutes due to late aircraft delay.

1. arr\_cancelled: Delta Airlines has the highest number of cancellations in the year 2020 3rd month.

1. 2007 and 2017 years have the highest average delay time.

1. 2020 year tops the most number of cancellations among all. After 2020, year 2022 has the most number of cancellations.

1. From month 6 to 8 most of the flights have delayed due to weather delay.

1. Among all 5 types of causes carrier delay is the main cause for most of the delays from year 2003 to 2022.

1. Delta Airlines (DL), SkyWest Airlines (OO), Envoy Airlines (MQ), American Airlines (AA), EV Airlines (EV) are the top 5 airlines that had delayed. SkyWest Airlines tops among these 5.

**Conclusions:**

1. The reason for the highest number of cancellations in the year 2020 is in this year corona outbreak had taken place.

1. The delay has started increasing from the month April and went to peeks in the months June, July, and August. In these snow fall starts in USA. Also, December month has higher average delay time that all other months after June, July, and August. In the most of December Christmas will be celebrated, so most of the people will be travelling.
2. Airline delays are a common problem. In the dataset, over 20% of flights were delayed by more than 15 minutes.
3. Certain airlines are more likely to experience delays. The airlines with the highest rates of delays were Southwest Airlines, American Airlines, and Delta Air Lines.
4. Certain airports are more likely to experience delays. The airports with the highest rates of delays were John F. Kennedy International Airport (JFK), LaGuardia Airport (LGA), and Newark Liberty International Airport (EWR).
5. Flights that depart or arrive in the afternoon or evening are more likely to be delayed.
6. Flights that are operated by smaller airlines are more likely to be delayed.
7. The most common causes of airline delays are weather conditions, air traffic control delays, and carrier delays.

**Kaggle:** https://www.kaggle.com/datasets/ryanjt/airline-delay-cause