GREAT LAKES
INSTITUTE OF MANAGEMENT, CHENNAI

greatlearning
Learning for Life

FLIGHT DELAY PREDICTION

Post Graduate Program in Data Science Engineering

Location: **Gurgaon** Batch: **Jan’22**

Submitted by

Ankit Chauhan

Ankush Dujania

Anukriti Singh Chauhan

Aryan Singh Bhadauria

Shivansh Verma

Srikar Yalamarthy

Mentored by

Mr. Muppidi Srikar

Table of Contents

1. ABSTRACT

Aviation plays a crucial role in our day-to-day mode of transportation and the costs of running incurred by these companies are significantly high due to which cost per delayed take-offs sets back the industry from growing and leaves customers with bad experiences.

The aim of this project is to explore ways to understand the cause of delay based on previous history of the take off occurring in all USA airports as per Federal Aviation Administration (FAA) maximum delays are caused due to variation in temperatures, weather and age of the airplane in service.

A machine learning model will be deployed to predict the future delays of the flight and inferences will be drawn from the predictions of the model which in turn will help the companies to deploy effective ways to reduce delay time and are equipped with better understanding on how to come up with better strategies to tackle such situations.

Some statistical numbers prove that:

- Delays were 33.4% of total delays in 2020.
- 5,400 average flights take off during a day
- 5,082 Public Airports, 14,551 Private Airports
- 2,900,000 Passengers fly every day in and out of U.S.A

2. INTRODUCTION

Commercial airlines are a backbone of the worldwide transportation system, bringing a good-sized socio-economic utility via means of permitting cheaper and less complex long-distance travel. After more than half a century of mainstream adoption (especially in the US), airline operations have seen primary optimizations and these days function with terrific reliability even withinside the face of exhausting engineering challenges. Still, the present-day passenger is once in a while inconvenienced through aircraft delays, disrupting an otherwise exacting system and causing tremendous inefficiencies at scale in 2007, 23% of US flights had been more than 15 minutes past due to depart (federal definition of flight delay), levying an aggregate price of $32.9bn on the US economy. Suboptimal weather situations had been the direct reason for ~17% of those delays, suggesting better expertise of aircraft unfriendly weather should enhance airline scheduling and notably reduce delays. The Federal Aviation Administration (FAA) considers a flight to be delayed when it is 15 minutes later than its scheduled time. Compensation for the delay is provided when the flight is delayed by more than 3 hours, the cost of which is borne by the aviation company. In some cases, the delay in flights is also caused by the age of the plane which suggests a plethora of factors involved in flight delays.

2.1 Domain and Feature Review

Nowadays, the aviation industry plays a crucial role in the world's transportation sector, and many businesses rely on various airlines to connect them with other parts of the world. But extreme weather conditions may directly affect the airline services causing flight delays.

To solve this issue, accurately predicting these flight delays allows passengers to be well prepared for the deterrent caused to their journey and enables airlines to respond to the potential causes of the flight delays in advance to diminish the negative impact.

We will be examining features like Distance group, The segment that this tail number is on for the day, Concurrent flights leaving from the airport in the same departure block, Number of seats on the aircraft, Airline, Average Airport Flights per Month, Average Airline Flights per Month, Average Flights per month for Airline and Airport, Average Passengers for the departing airport for the month, Average Passengers for the airline for the month, Flight attendants per passenger for airline, Age of departing aircraft, Departing Airport, Previous airport that aircraft departed from, and whether conditions like Inches of precipitation for the day, Inches of snowfall for the day, Inches of snow on the ground for the day, Maximum temperature for the day, Max wind speed for the day.

2.2 Dataset Information

The dataset is provided by machinehack.com as part of the Data Engineering Championship. The dataset consists of Domestic flights in the USA for January, 2020.

Link to the dataset: https://machinehack.com/hackathons/data_engineering_championship/data

2.3 Problem Statement

How can we predict the future delays of departure of flights based on the previous delay data?

Our project focuses upon the data of flights from various parts of the US. It gives us various features such as snow and wind to help in predicting if there will be a delay in the flights by 15 minutes. Our target variable is DEP_DEL15 which has binary values; 0 is for flight not delayed and 1 is for flight delayed.

2.4 Variable Categorization with description

Independent variables in the data are given below:

- AIRPLANE_ID

- YEAR

- MONTH

- DAY_OF_WEEK

- DEP_TIME_BLK

- DISTANCE_GROUP

- SEGMENT_NUMBER
- CONCURRENT_FLIGHTS
- NUMBER_OF_SEATS
- CARRIER_NAME
- AIRPORT_FLIGHTS_MONTH
- AIRLINE_FLIGHTS_MONTH
- AIRLINE_AIRPORT_FLIGHTS_MONTH
- AVG_MONTHLY_PASS_AIRPORT
- AVG_MONTHLY_PASS_AIRLINE
- FLT_ATTENDANTS_PER_PASS
- GROUND_SERV_PER_PASS
- PLANE_AGE
- DEPARTING_AIRPORT
- LATITUDE
- LONGITUDE
- PREVIOUS_AIRPORT
- PRCP
- SNOW
- SNWD
- TMAX

- AWND

Target Variable from the data is: DEP_DEL15

The variable information is as follows: -

- MONTH: Month, of which the data is collected
- YEAR: Year, of which the data is collected
- DAY_OF_WEEK: Day of Week
- DEP_DEL15: (Target) Binary of a departure delay over 15 minutes (1 is yes)
- DISTANCE_GROUP: Distance group to be flown by departing aircraft
- DEP_BLOCK: Departure block of the aircraft
- SEGMENT_NUMBER: The segment that this tail number is on for the day
- CONCURRENT_FLIGHTS: Concurrent flights leaving from the airport in the same departure block
- NUMBER_OF_SEATS: Number of seats on the aircraft
- CARRIER_NAME: Name of the carrier
- AIRPORT_FLIGHTS_MONTH: Average Airport Flights per Month
- AIRLINE_FLIGHTS_MONTH: Average Airline Flights per Month
- AIRLINE_AIRPORT_FLIGHTS_MONTH: Average Flights per month for Airline AND Airport
- AVG_MONTHLY_PASS_AIRPORT: Average Passengers for the departing airport for the month
- AVG_MONTHLY_PASS_AIRLINE: Average Passengers for the airline for the month

- FLT_ATTENDANTS_PER_PASS: Flight attendants per passenger for airline

- GROUND_SERV_PER_PASS: Ground service employees (service desk) per passenger for airline

- PLANE_AGE: Age of departing aircraft

- DEPARTING_AIRPORT: Departing Airport

- LATITUDE: Latitude of departing airport

- LONGITUDE: Longitude of departing airport

- PREVIOUS_AIRPORT: Previous airport that aircraft departed from

- PRCP: Inches of precipitation for the day

- SNOW: Inches of snowfall for the day

- SNWD: Inches of snow on the ground for the day

- TMAX: Max temperature for the day

- AWND: Max wind speed for the day

- AIRPLANE_ID: Airplane ID

2.4.1 Numerical

The following are the numerical variables in the data.

CONCURRENT_FLIGHTS

NUMBER_OF_SEATS

AIRPORT_FLIGHTS_MONTH

AIRLINE_FLIGHTS_MONTH

AIRLINE_AIRPORT_FLIGHTS_MONTH

AVG_MONTHLY_PASS_AIRPORT

AVG_MONTHLY_PASS_AIRLINE

FLT_ATTENDANTS_PER_PASS

GROUND_SERV_PER_PASS

PLANE_AGE

PRCP

SNOW

SNWD

TMAX

AWND

2.4.2 Categorical

The following are the categorical variables in the data.

CARRIER_NAME
DEPARTING_AIRPORT
PREVIOUS_AIRPORT_REGION
DEP_DEL15(Target)
DISTANCE_GROUP
SEGMENT_NUMBER
CARRIER_NAME

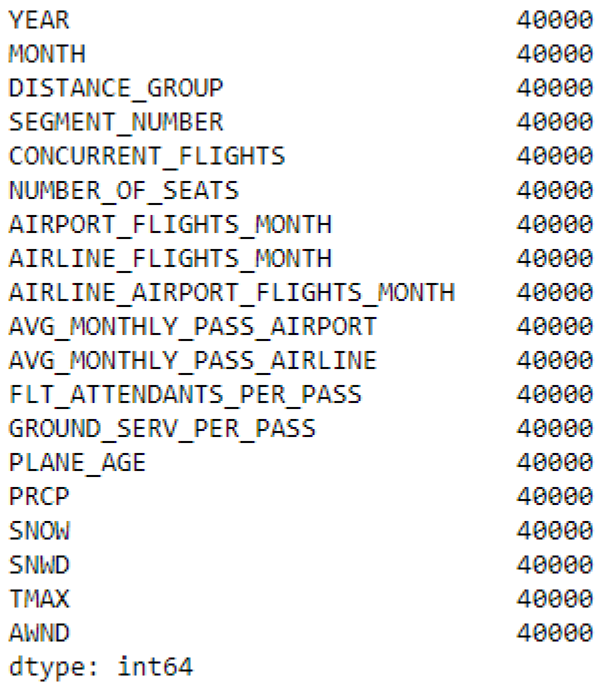
2.5 Target Variable

Our target variable is DEP_DEL15. This will give the outputs in binary, 0 and 1. 0 shows the flights that were not affected and delayed, whereas 1 refers to the flights that were delayed due to others significant variables in the data.

3. DATA PREPROCESSING

3.1 Missing Values

Following are the number of null values per column:

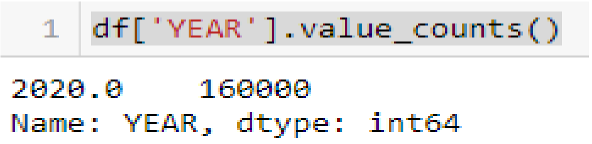


```
YEAR                            40000
MONTH                           40000
DISTANCE_GROUP                  40000
SEGMENT_NUMBER                  40000
CONCURRENT_FLIGHTS              40000
NUMBER_OF_SEATS                 40000
AIRPORT_FLIGHTS_MONTH           40000
AIRLINE_FLIGHTS_MONTH           40000
AIRLINE_AIRPORT_FLIGHTS_MONTH   40000
AVG_MONTHLY_PASS_AIRPORT        40000
AVG_MONTHLY_PASS_AIRLINE        40000
FLT_ATTENDANTS_PER_PASS         40000
GROUND_SERV_PER_PASS            40000
PLANE_AGE                       40000
PRCP                            40000
SNOW                            40000
SNWD                            40000
TMAX                            40000
AWND                            40000
dtype: int64
```

Out of 28 Columns there are 19 Columns having null values.

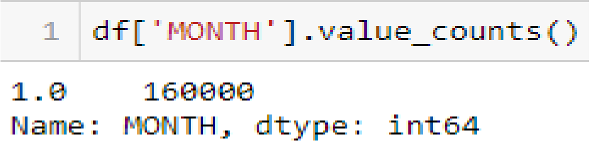
<u>PATTERN 1:</u>

Taking into account that this column has just one value in all the rows accounting for 1,60,000 single values in all the rows.



```
df['YEAR'].value_counts()
```

```
2020.0    160000
Name: YEAR, dtype: int64
```



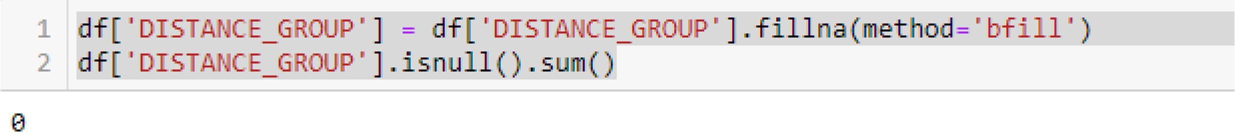
```
df['MONTH'].value_counts()
```

```
1.0    160000
Name: MONTH, dtype: int64
```

So, we imputed all the NaN values with the mode (highest number of occurrences).

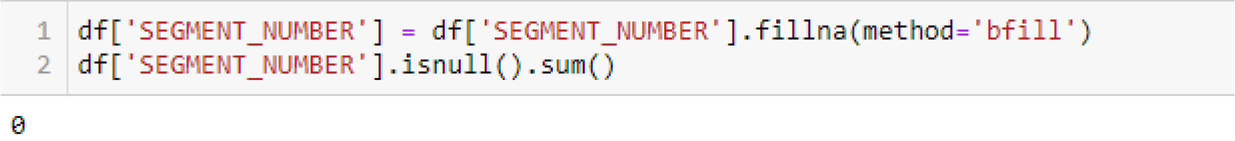
<u>PATTERN 2:</u>

On the same line after filling with mode we checked for other variables and they were not compatible for filling with mode as it changes the distribution hence we tried for bfill and ffill



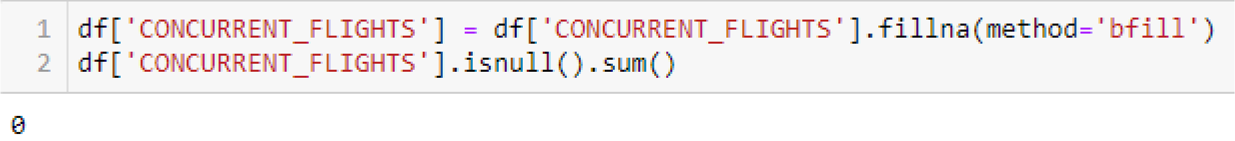
```
df['DISTANCE_GROUP'] = df['DISTANCE_GROUP'].fillna(method='bfill')
df['DISTANCE_GROUP'].isnull().sum()
```

```
0
```



```
df['SEGMENT_NUMBER'] = df['SEGMENT_NUMBER'].fillna(method='bfill')
df['SEGMENT_NUMBER'].isnull().sum()
```

```
0
```



```
df['CONCURRENT_FLIGHTS'] = df['CONCURRENT_FLIGHTS'].fillna(method='bfill')
df['CONCURRENT_FLIGHTS'].isnull().sum()
```

```
0
```

So, we imputed the missing values in [DISTANCE_GROUP, SEGMENT_NUMBER, CONCURRENT_FLIGHTS, AIRPORT_FLIGHTS_MONTH, AIRLINE_AIRPORT_FLIGHTS_MONTH, AVG_MONTHLY_PASS_AIRLINE, FLT_ATTENDANTS_PER_PASS, GROUND_SER_PER_PASS, SNWD, TMAX] with the bfill and ffill and checking for the distribution after imputing we found the data to be in same distribution before filling the NaN values.

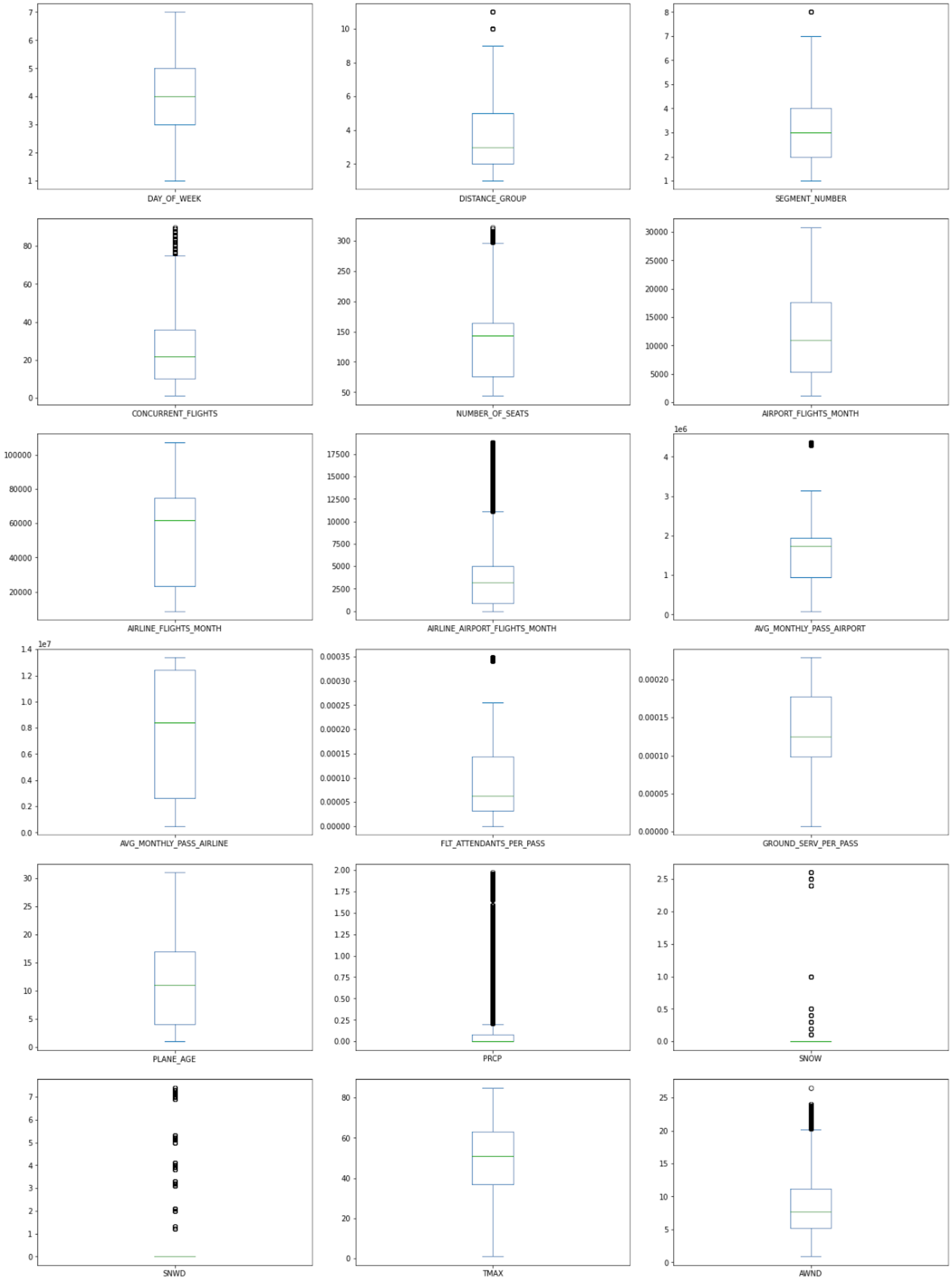
PATTERN 3:

After filling with bfill and fill and checking for the distribution, it remained same but for some variables the distribution was not same before and after filling of the values

PATTERN 4:

After filling maximum of the variables, we are left with some of variables

3.2 Check for Outliers



As we can see from the above graph, there are outliers in DISTANCE_GROUP, SEGMENT_NUMBER, CONCURRENT_FLIGHTS, NUMBER_OF_SEATS, AIRLINE_AIRPORT_FLIGHTS_MONTH, AVG_MONTHLY_PASS_AIRPORT, FLT_ATTENDANTS_PER_PASS, PRCP, SNOW, SNWD and AWND.

We checked for the specific values wherever we spotted outliers through above graph, and we saw that feature like SNOW and SNWD have very high outliers values. When we removed them, the overall rows of our data were declining by larger margin, so we preferred to keep such variables and values in our data. Also, we saw that dropping such columns would not be logically right as these can have high effect on a delay of a flight.

3.3 Transformation Techniques

Feature Engineering: The process of taking raw data and extracting or creating new features that allow a machine learning model to learn a mapping between these features and the target. This might mean taking transformations of variables, such as we do with the log and square root, or one-hot encoding categorical variables so they can be used in a model. Generally, feature engineering as additional features derived from the raw data. Now that we have explored the trends and relationships within the data, we worked on engineering a set of features for our models. In particular, we learned the following from EDA which can help us in engineering/selecting features:

One way to address this issue is to transform the distribution of values in a dataset using one of the three transformations:

1. Log Transformation: Transform the response variable from y to log(y).

2. Square Root Transformation: Transform the response variable from y to √y.

3. Cube Root Transformation: Transform the response variable from y to y1/3.

*We will split the data into train and test using train_test_split and then apply the appropriate data transformation technique (TBD).

3.4 Dropping of Columns

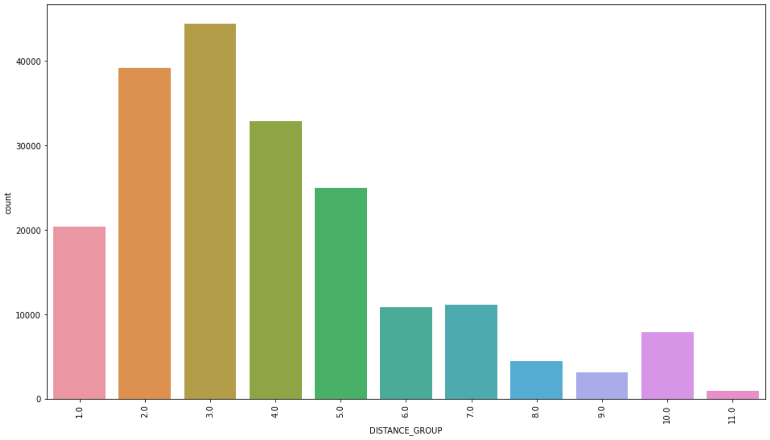
After the above steps, we found the following as the insignificant variables in accordance with our target variable.

Airplane ID
Latitude
Longitude
Year
Month

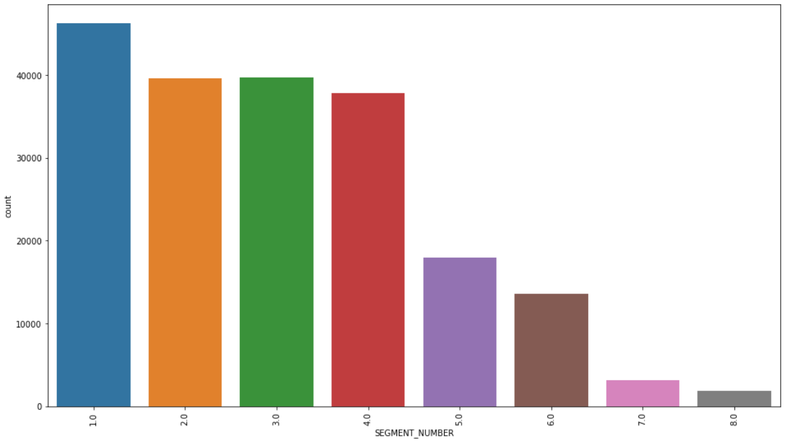
4. EXPLORATORY DATA ANALYSIS

4.1 Univariate Analysis

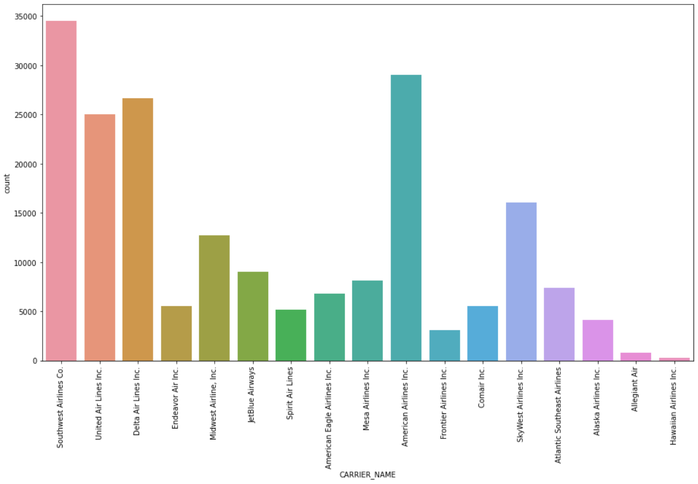
Analysis of Categorical Variables



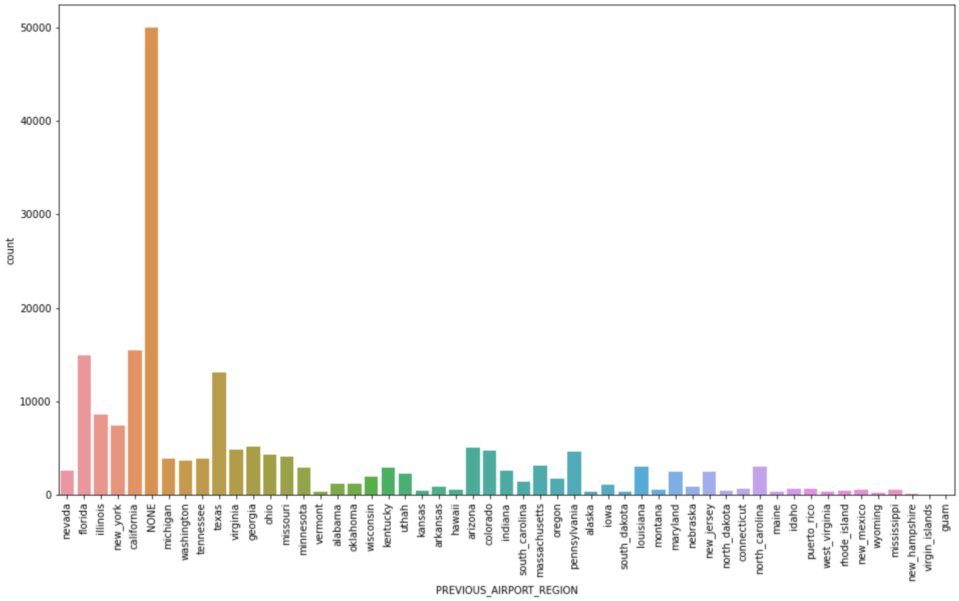
DISTANCE_GROUP : Group 3 has the highest spread of data, followed by Group 2 and 4.



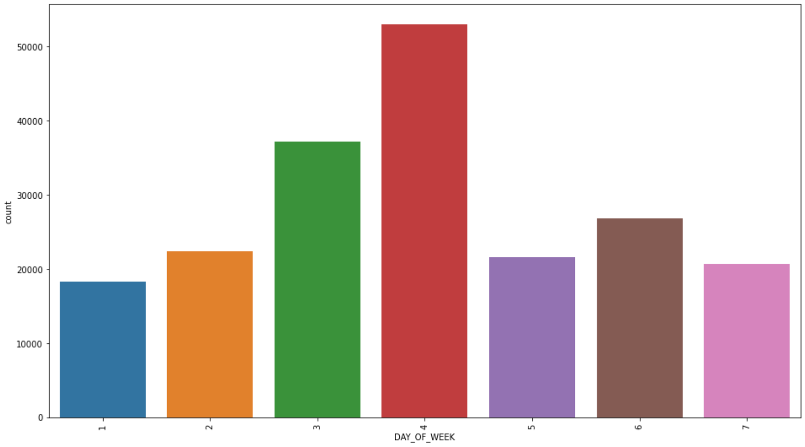
SEGMENT_NUMBER: Segment number 1 is highly spread as compared to the rest.



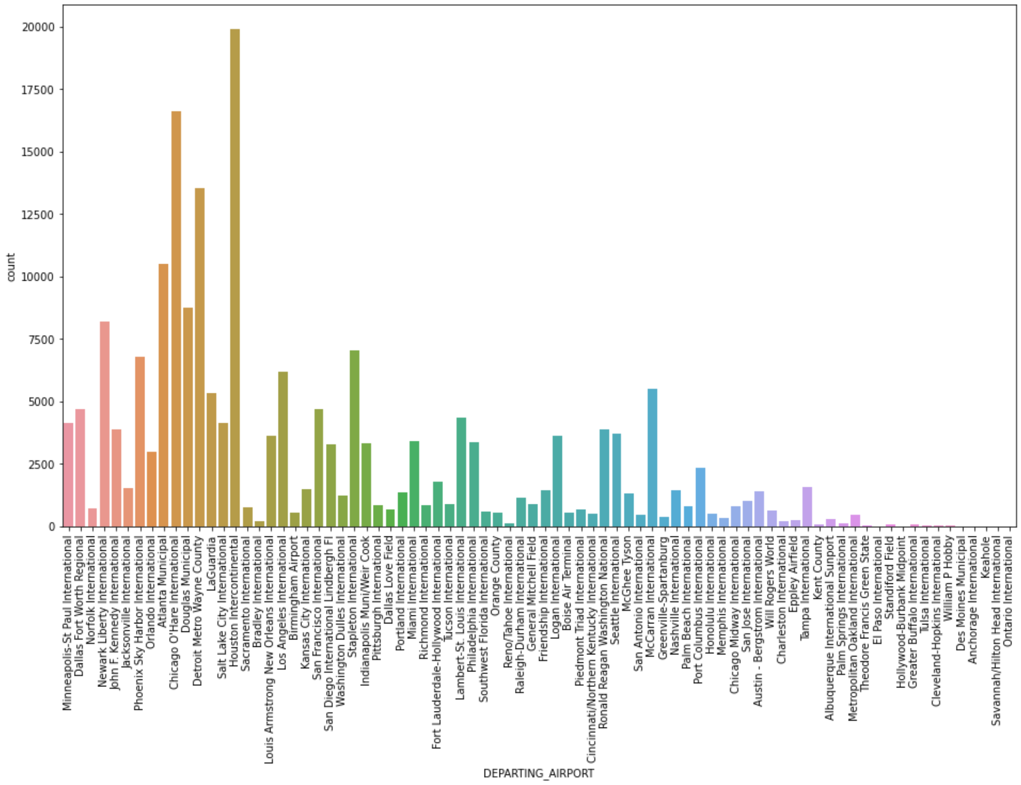
CARRIER_NAME: Most of the flights are of Southwest Airlines Co., American airlines inc. and Delta Airlines inc.



PREVIOUS_AIRPORT_REGION: Most of the flights have no previous departure data, shown under attribute names NONE.

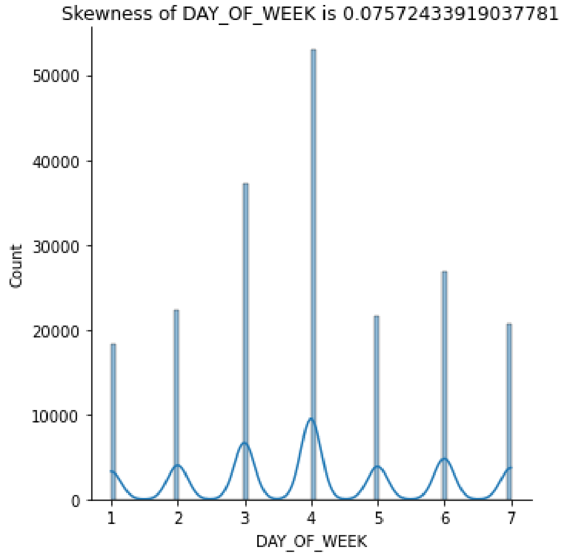


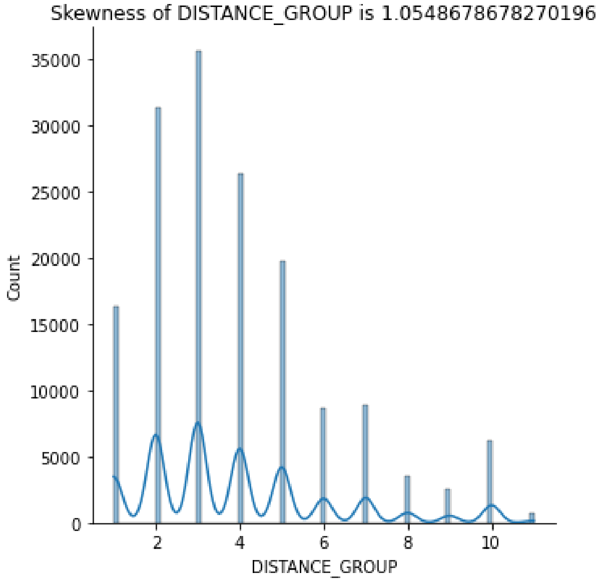
DAY_OF_WEEK: Normally distributed data

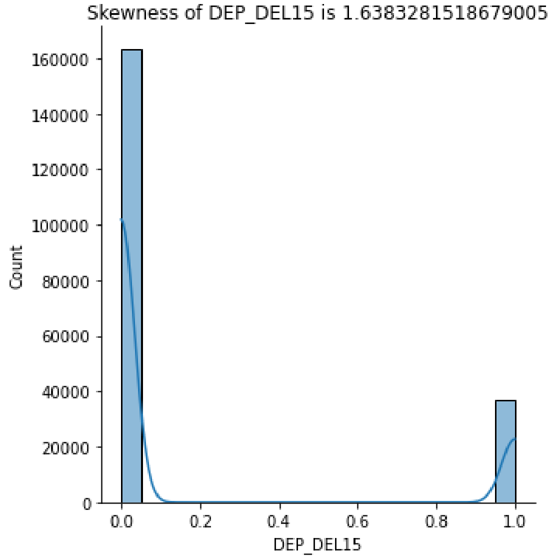


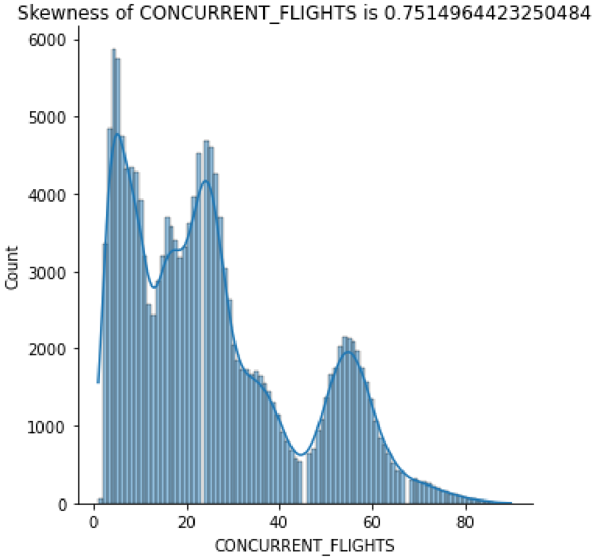
DEPARTING_AIRPORT: The above graph shows the previous airports from which the flights have taken off.

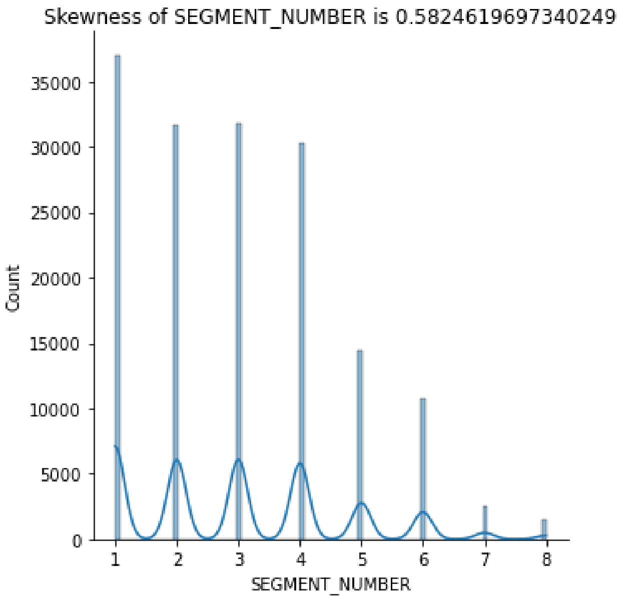
Analysis of Numerical variables

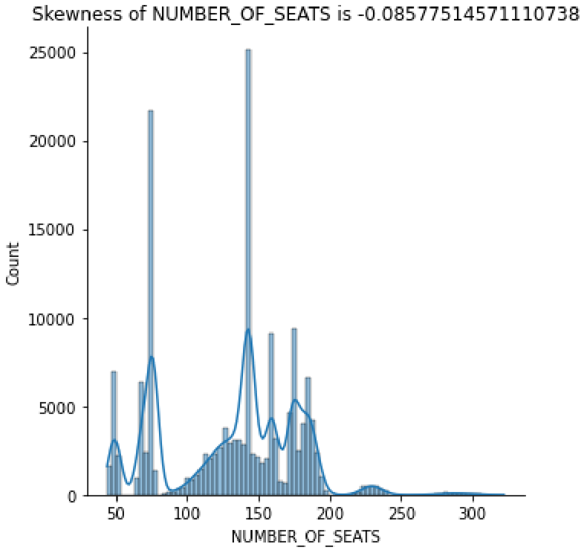


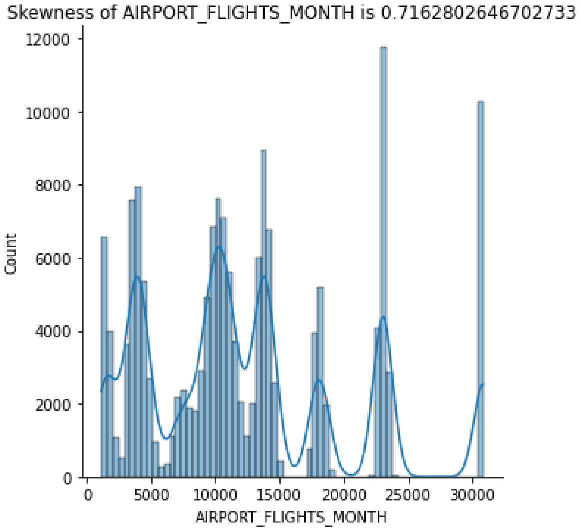


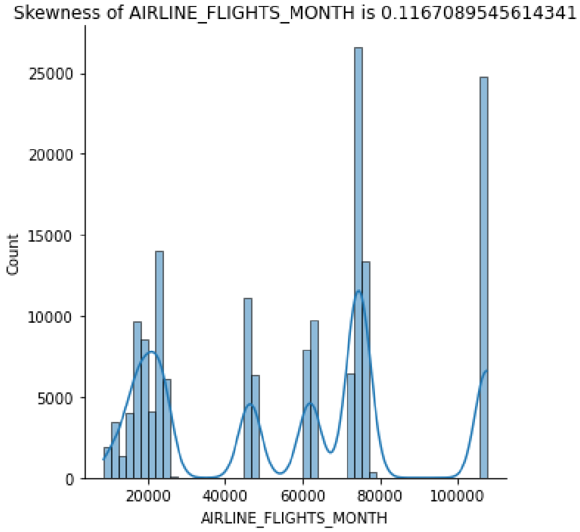


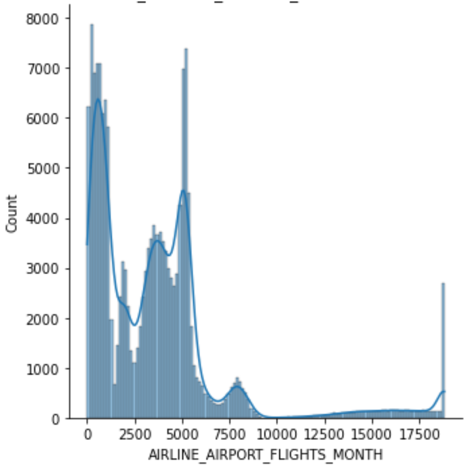



Skewness of SEGMENT_NUMBER is 0.5824619697340249
Count
35000
30000
25000
20000
15000
10000
5000
0
1
2
3
4
5
6
7
8
SEGMENT_NUMBER

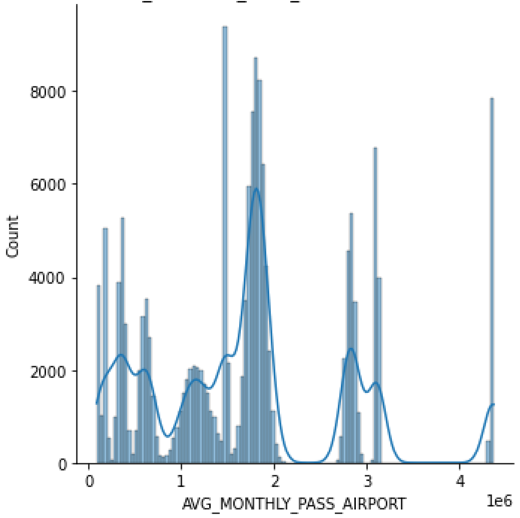

Skewness of NUMBER_OF_SEATS is -0.08577514571110738
Count
25000
20000
15000
10000
5000
0
50
100
150
200
250
300
NUMBER_OF_SEATS


Skewness of AIRPORT_FLIGHTS_MONTH is 0.7162802646702733
Count
12000
10000
8000
6000
4000
2000
0
0
5000
10000
15000
20000
25000
30000
AIRPORT_FLIGHTS_MONTH

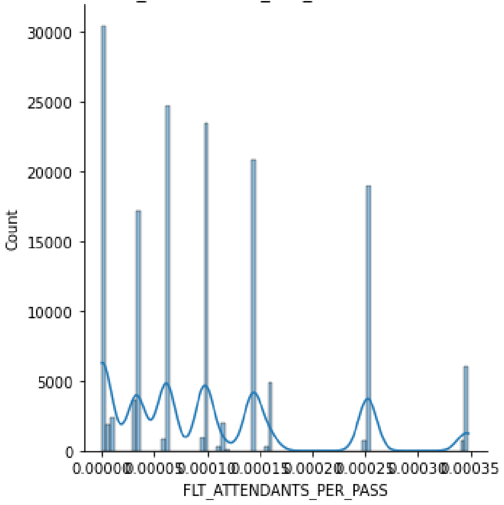

Skewness of AIRLINE_FLIGHTS_MONTH is 0.1167089545614341
Count
25000
20000
15000
10000
5000
0
20000
40000
60000
80000
100000
AIRLINE_FLIGHTS_MONTH



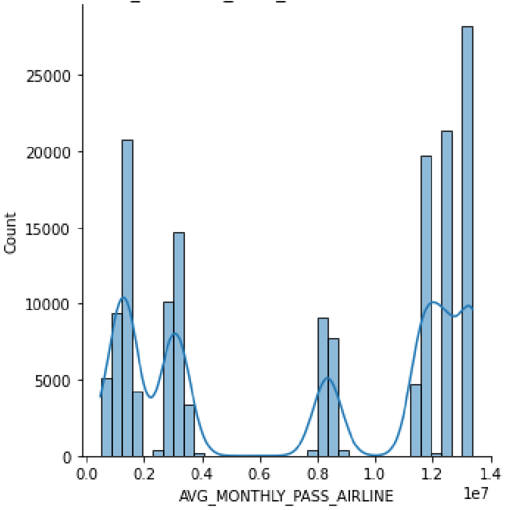
Skewness of AIRLINE_AIRPORT_FLIGHTS_MONTH is 2.2901834315565615



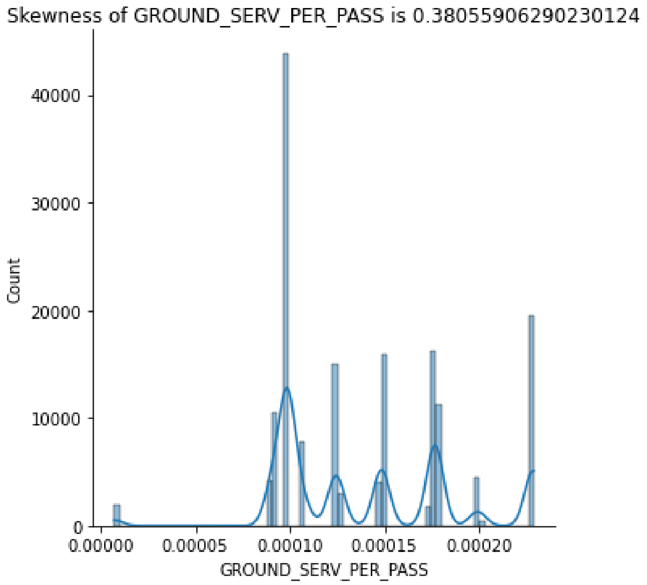
Skewness of AVG_MONTHLY_PASS_AIRPORT is 0.6509401456621381

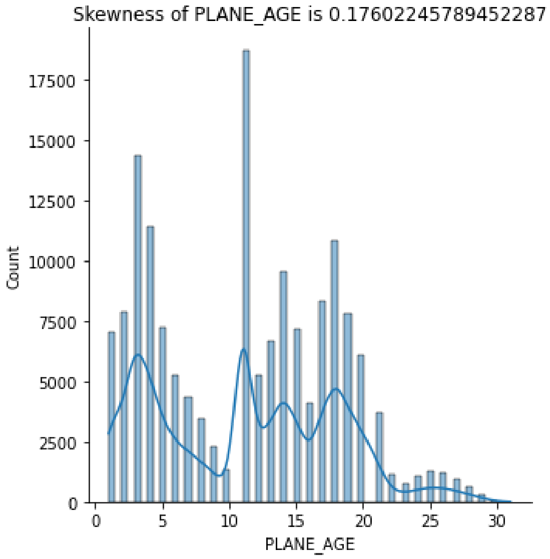


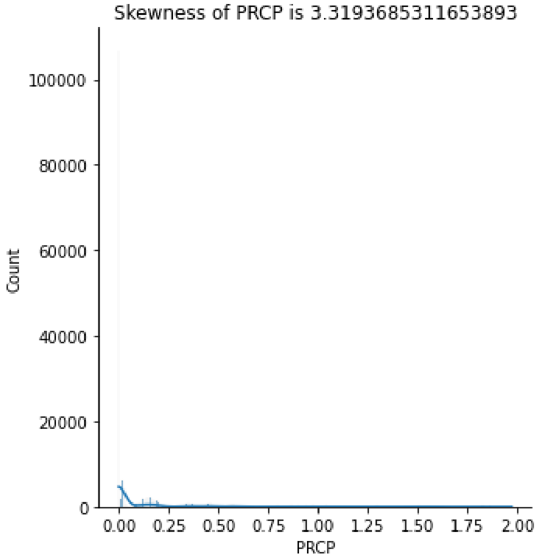
Skewness of FLT_ATTENDANTS_PER_PASS is 1.0027451144350805

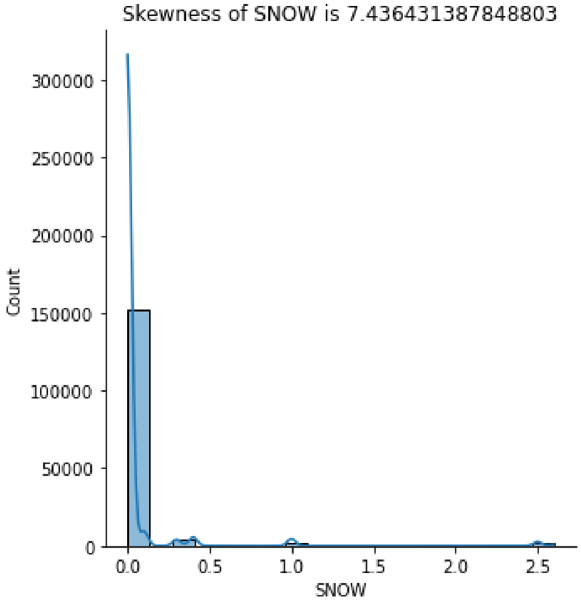


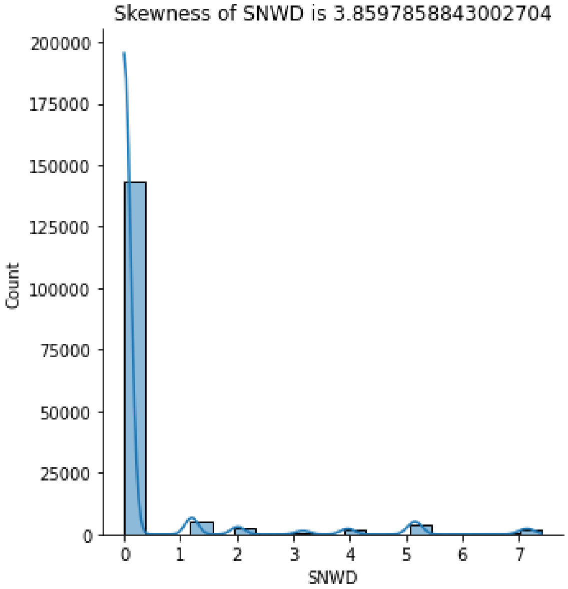
Skewness of AVG_MONTHLY_PASS_AIRLINE is -0.16204623718473274

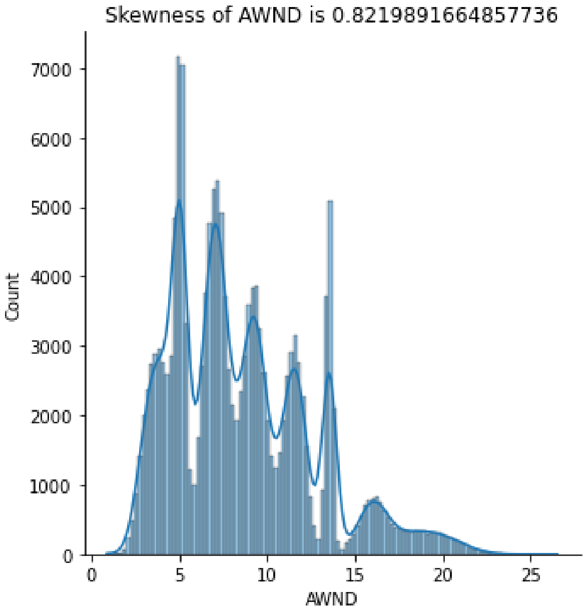

Skewness of GROUND_SERV_PER_PASS is 0.38055906290230124

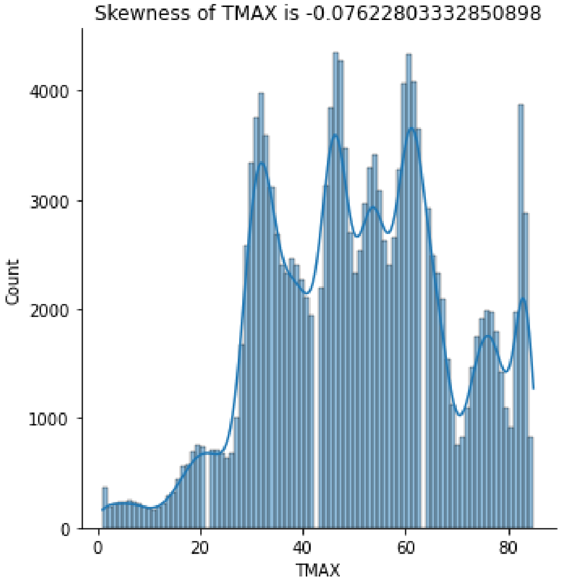

Skewness of PLANE_AGE is 0.17602245789452287


Skewness of PRCP is 3.3193685311653893


Skewness of SNOW is 7.436431387848803

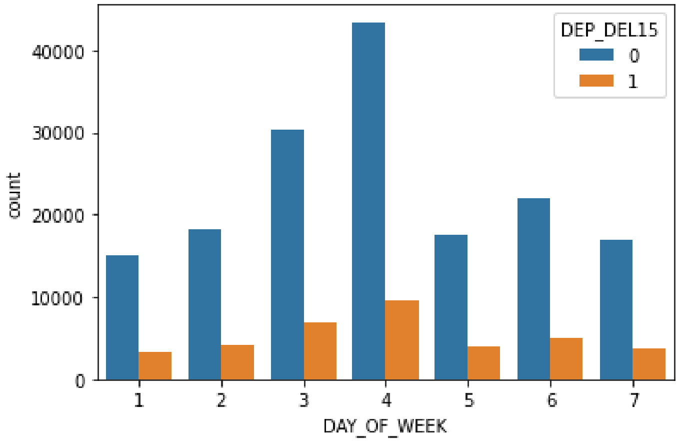




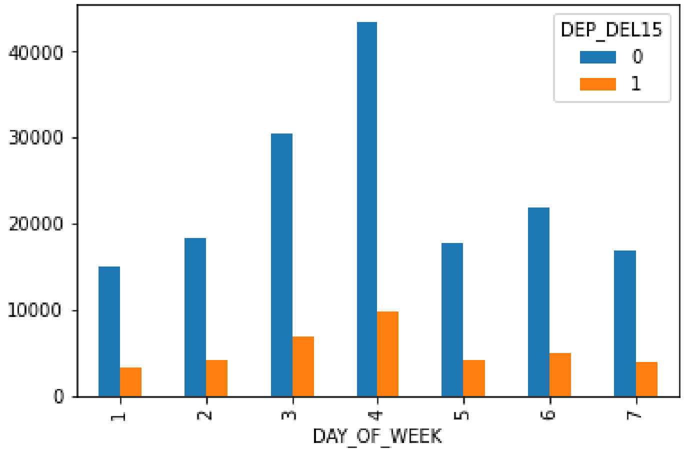


4.2 Bivariate and Multivariate Analysis

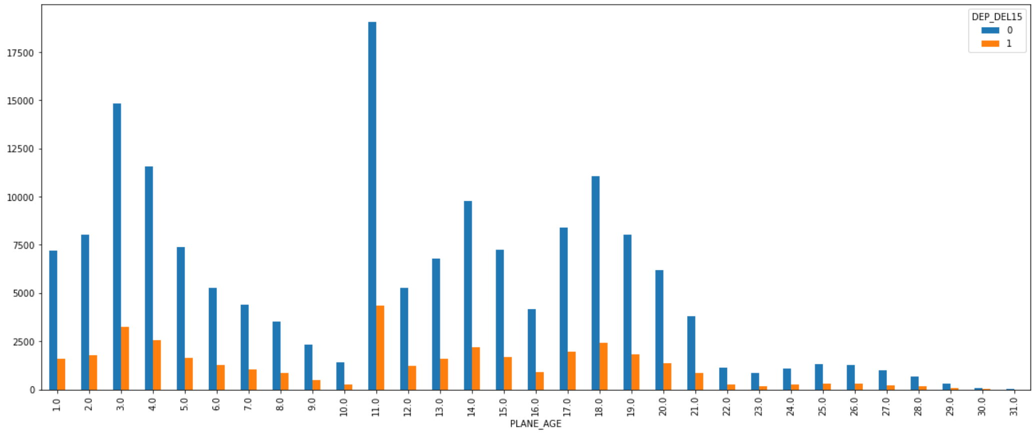
DAY_OF_WEEK: is contributing significantly to a **'DEP_DEL15,** but intuitively it can be one of the important features. This might be due to specificity in age numbers. So, we decided to bin DAY_OF_WEEK into different DAY_OF_WEEK categories to generalize the data to get more insights into the data.

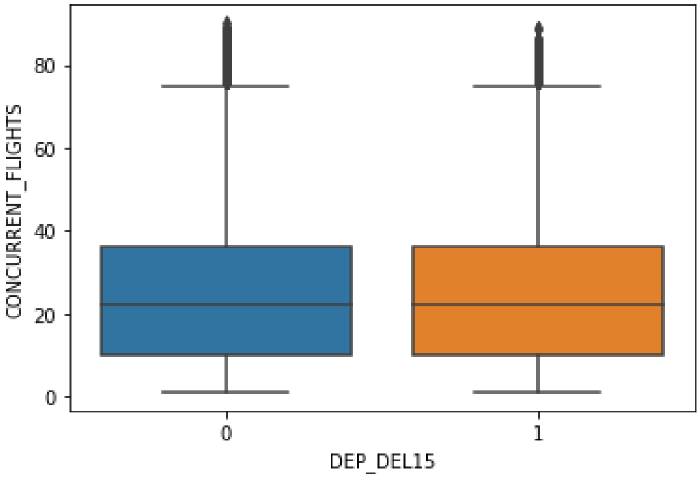


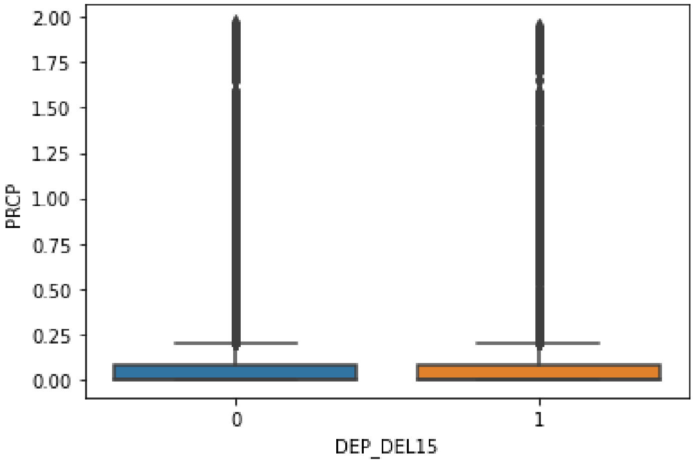
DAY_OF_WEEK: is contributing significantly to a **'DEP_DEL15,** but intuitively it can be one of the important features. This might be due to specificity on day 4th, which we can see in the graph as there is a high number of not delayed flights and parallelly high number of delayed flights but less in comparison to not delayed numbers. So, we decided to bin DAY_OF_WEEK into different DAY_OF_WEEK categories to generalize the data to get more insights into the data.

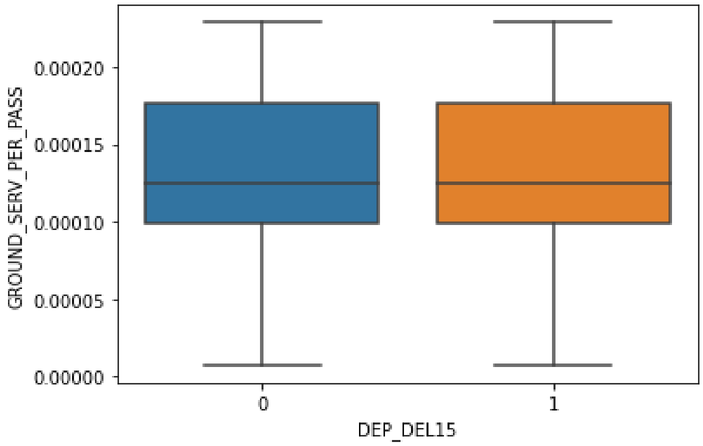


'PLANE_AGE': is contributing significantly to a **'DEP_DEL15,** but intuitively it can be one of the important features. This might be due to specificity on day 4th, which we can see in the graph as there is a high number of not delayed flights and parallelly high number of delayed flights but less in comparison to not delayed numbers. So, we decided to bin DAY_OF_WEEK into different DAY_OF_WEEK categories to generalize the data to get more insights into the data.


DEP_DEL15
0
1
17500
15000
12500
10000
7500
5000
2500
0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
14.0
15.0
16.0
17.0
18.0
19.0
20.0
21.0
22.0
23.0
24.0
25.0
26.0
27.0
28.0
29.0
30.0
31.0
PLANE_AGE


80
60
40
20
0
CONCURRENT_FLIGHTS
0
1
DEP_DEL15


2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
PRCP
0
1
DEP_DEL15



Inferences for

CONCURRENT_FLIGHTS VS DEP_DEL15

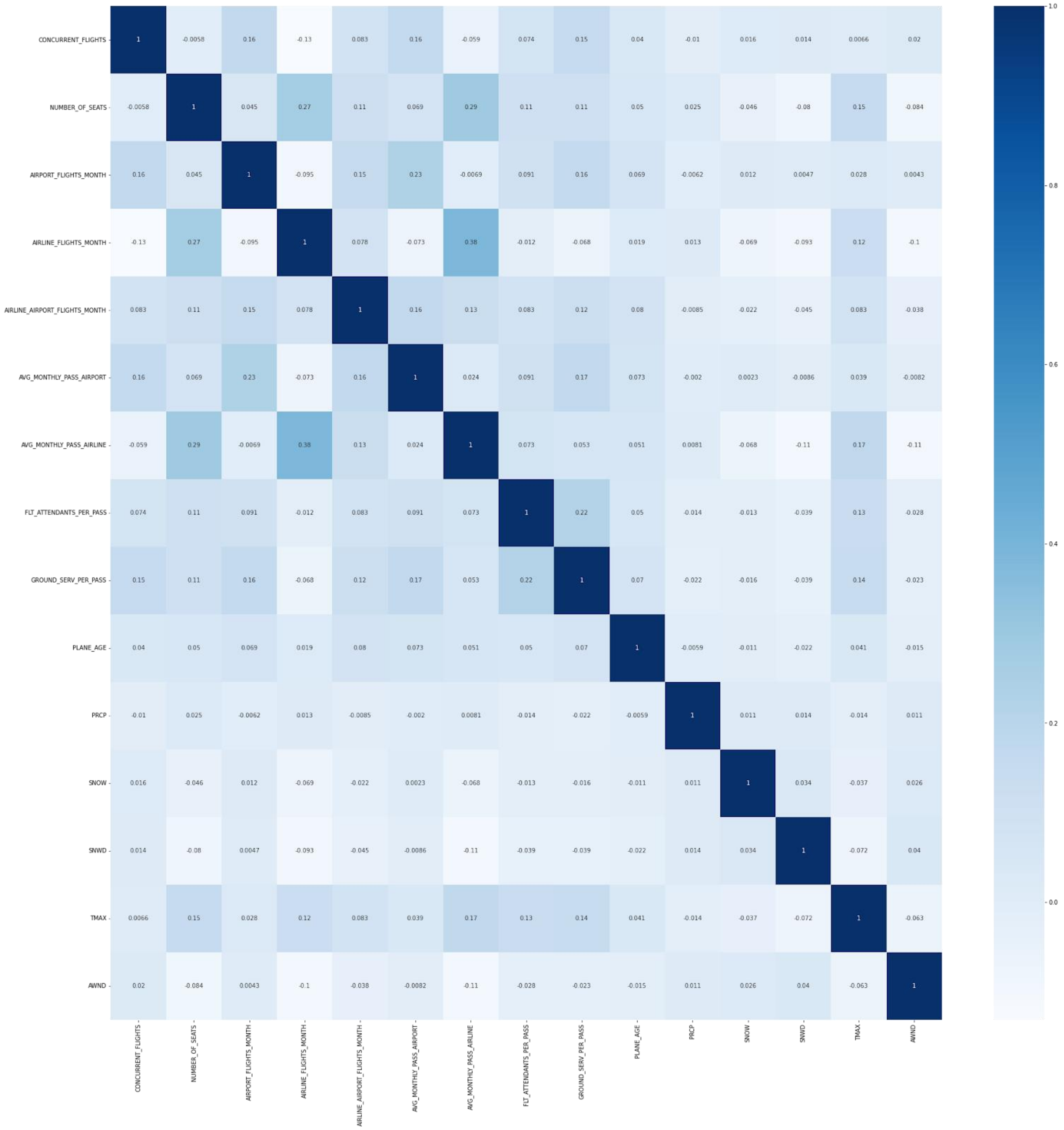
PRCP VS DEP_DEL15

GROUND_SERV_PER_PASS VS DEP_DEL15

All variables contributing significantly to a **'DEP_DEL15,** each variable has outliers even flights are delayed or not.

2. The min and max value of both variables are same for flights have delayed or not delayed.

A heat map is a graphical representation of data that uses a system of color-coding to represent different values. We mostly use heat map to visualize correlations between (continuous) features. Following is a heat map showing correlations between continuous features in the data:


CONCURRENT_FLIGHTS 1 -0.0058 0.16 -0.13 0.083 0.16 -0.059 0.074 0.15 0.04 -0.01 0.016 0.014 0.0066 0.02
NUMBER_OF_SEATS -0.0058 1 0.045 0.27 0.11 0.069 0.29 0.11 0.11 0.05 0.025 -0.046 -0.08 0.15 -0.084
AIRPORT_FLIGHTS_MONTH 0.16 0.045 1 -0.095 0.15 0.23 -0.0069 0.091 0.16 0.069 -0.0062 0.012 0.0047 0.028 0.0043
AIRLINE_FLIGHTS_MONTH -0.13 0.27 -0.095 1 0.078 -0.073 0.38 -0.012 -0.068 0.019 0.013 -0.069 -0.093 0.12 -0.1
AIRLINE_AIRPORT_FLIGHTS_MONTH 0.083 0.11 0.15 0.078 1 0.16 0.13 0.083 0.12 0.08 -0.0085 -0.022 -0.045 0.083 -0.038
AVG_MONTHLY_PASS_AIRPORT 0.16 0.069 0.23 -0.073 0.16 1 0.024 0.091 0.17 0.073 -0.002 0.0023 -0.0086 0.039 -0.0082
AVG_MONTHLY_PASS_AIRLINE -0.059 0.29 -0.0069 0.38 0.13 0.024 1 0.073 0.053 0.051 0.0081 -0.068 -0.11 0.17 -0.11
FLT_ATTENDANTS_PER_PASS 0.074 0.11 0.091 -0.012 0.083 0.091 0.073 1 0.22 0.05 -0.014 -0.013 -0.039 0.13 -0.028
GROUND_SERV_PER_PASS 0.15 0.11 0.16 -0.068 0.12 0.17 0.053 0.22 1 0.07 -0.022 -0.016 -0.039 0.14 -0.023
PLANE_AGE 0.04 0.05 0.069 0.019 0.08 0.073 0.051 0.05 0.07 1 -0.0059 -0.011 -0.022 0.041 -0.015
PRCP -0.01 0.025 -0.0062 0.013 -0.0085 -0.002 0.0081 -0.014 -0.022 -0.0059 1 0.011 0.014 -0.014 0.011
SNOW 0.016 -0.046 0.012 -0.069 -0.022 0.0023 -0.068 -0.013 -0.016 -0.011 0.011 1 0.034 -0.037 0.026
SNWD 0.014 -0.08 0.0047 -0.093 -0.045 -0.0086 -0.11 -0.039 -0.039 -0.022 0.014 0.034 1 -0.072 0.04
TMAX 0.0066 0.15 0.028 0.12 0.083 0.039 0.17 0.13 0.14 0.041 -0.014 -0.037 -0.072 1 -0.063
AWND 0.02 -0.084 0.0043 -0.1 -0.038 -0.0082 -0.11 -0.028 -0.023 -0.015 0.011 0.026 0.04 -0.063 1
CONCURRENT_FLIGHTS
NUMBER_OF_SEATS
AIRPORT_FLIGHTS_MONTH
AIRLINE_FLIGHTS_MONTH
AIRLINE_AIRPORT_FLIGHTS_MONTH
AVG_MONTHLY_PASS_AIRPORT
AVG_MONTHLY_PASS_AIRLINE
FLT_ATTENDANTS_PER_PASS
GROUND_SERV_PER_PASS
PLANE_AGE
PRCP
SNOW
SNWD
TMAX
AWND
1.0
0.8
0.6
0.4
0.2
0.0

4.3 Insights from EDA

- **AIRLINE_AIRPORT_FLIGHTS_MONTH and** FLT**_ATTENDANTS_PER_PASS** is correlated with **CONCURRENT_FLIGHTS.**

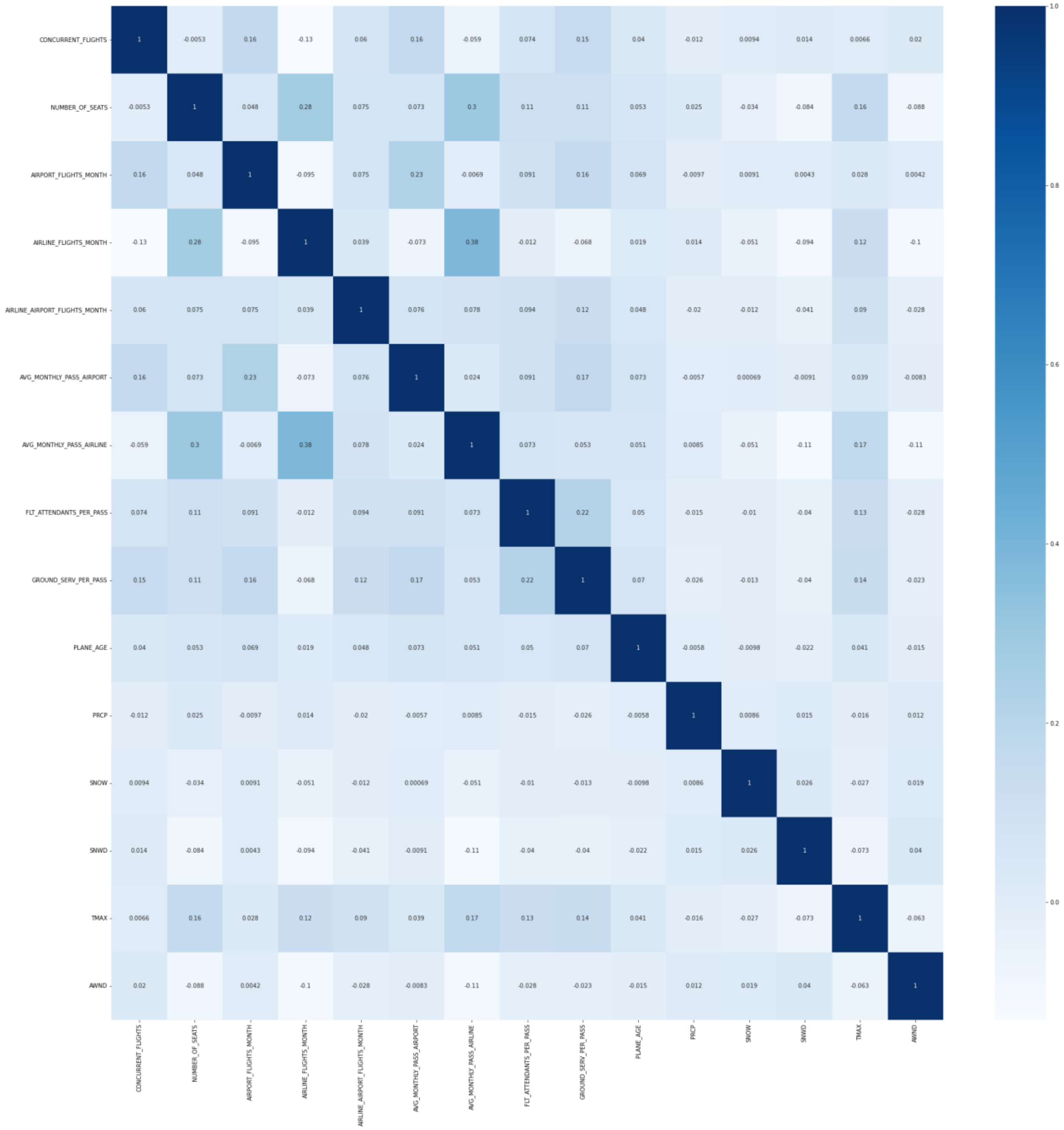
- **AIRLINE_FLIGHTS_MONTH** and **AIRLINE_AIRPORT_FLIGHTS_MONTH., TMAX, FLT_ATTENDANTS_PER_PASS, AIRLINE_FLIGHTS_MONTH, CONCURRENT_FLIGHTS** are correlated with **AIRLINE_AIRPORT_FLIGHTS_MONTH.**

- **NUMBER_OF_SEATS** are correlated with **AVG_MONTHLY_PASS_AIRPORT**

- Insignificant variables that are to be dropped are**: YEAR, MONTH, LATITUDE, LONGITUDE, AIRPLANE_ID.**

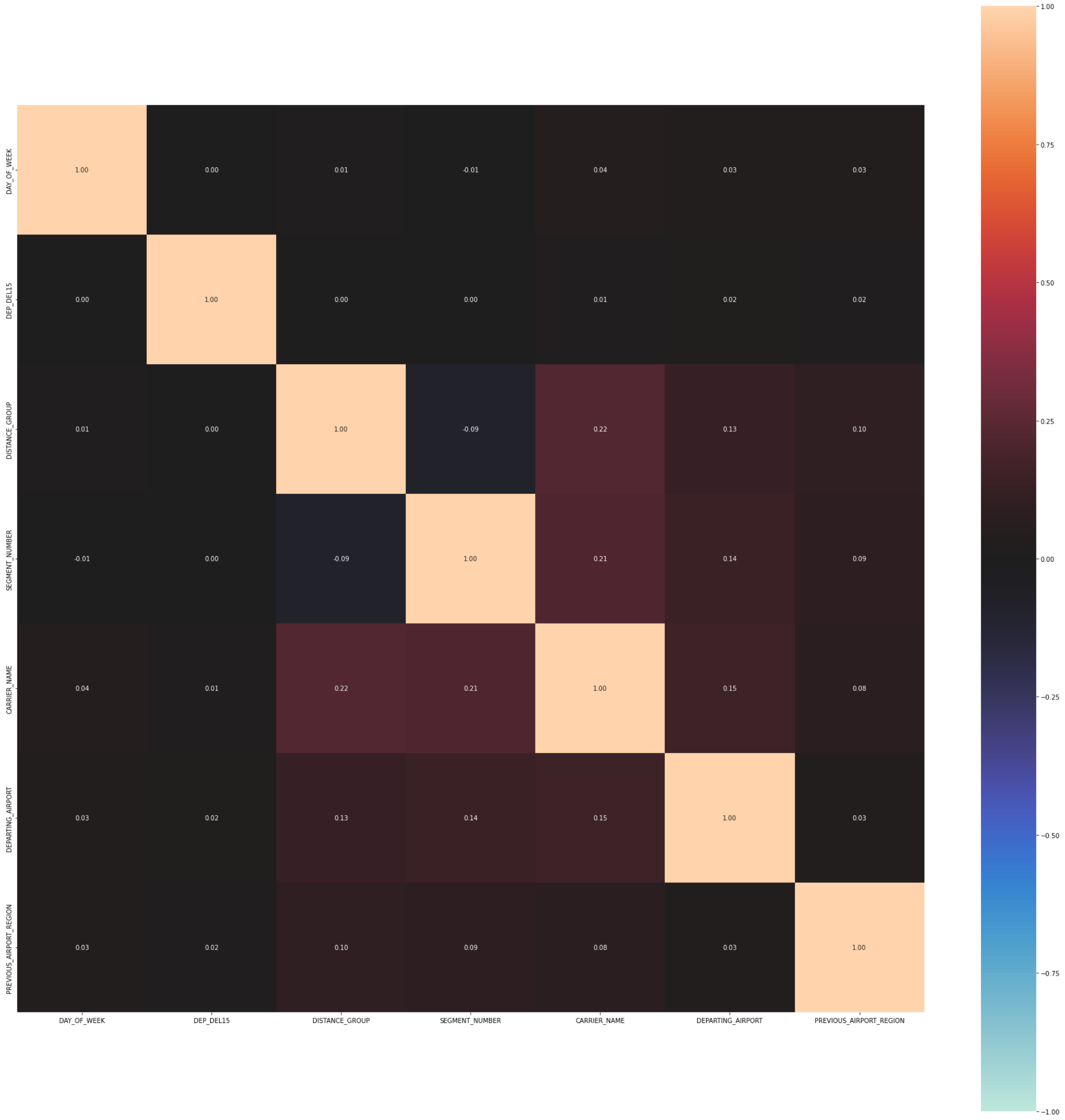
  - We only have data of January 2020 so there is no need to work with these variables.
  - Latitude and Longitude can be dropped as we have other attributes which gives us satisfactory information about the region.
  - Each plane has its own unique ID so AIRPLANE_ID as an attribute is not contributing much to the further steps in analysis the delay
- There are outliers present in various columns. When different techniques were applied to treat those outliers, we saw data leakage, hence, we decided to keep the outliers for further steps and analysis it better in order to have better precision model.
- Imputation technique was applied, and capping was done using upper whisker for outliers.
- Out of 28 columns, there were null values in 19 columns, each having 20%.
- Null Values were treated using various techniques according to the data patterns.
- Grouping was done and added to a new variable name, PREVIOUS_AIRPORT_REGION, which contains region- wise airports.
- There no two variables with high correlation.

4.2 Numerical feature Correlation



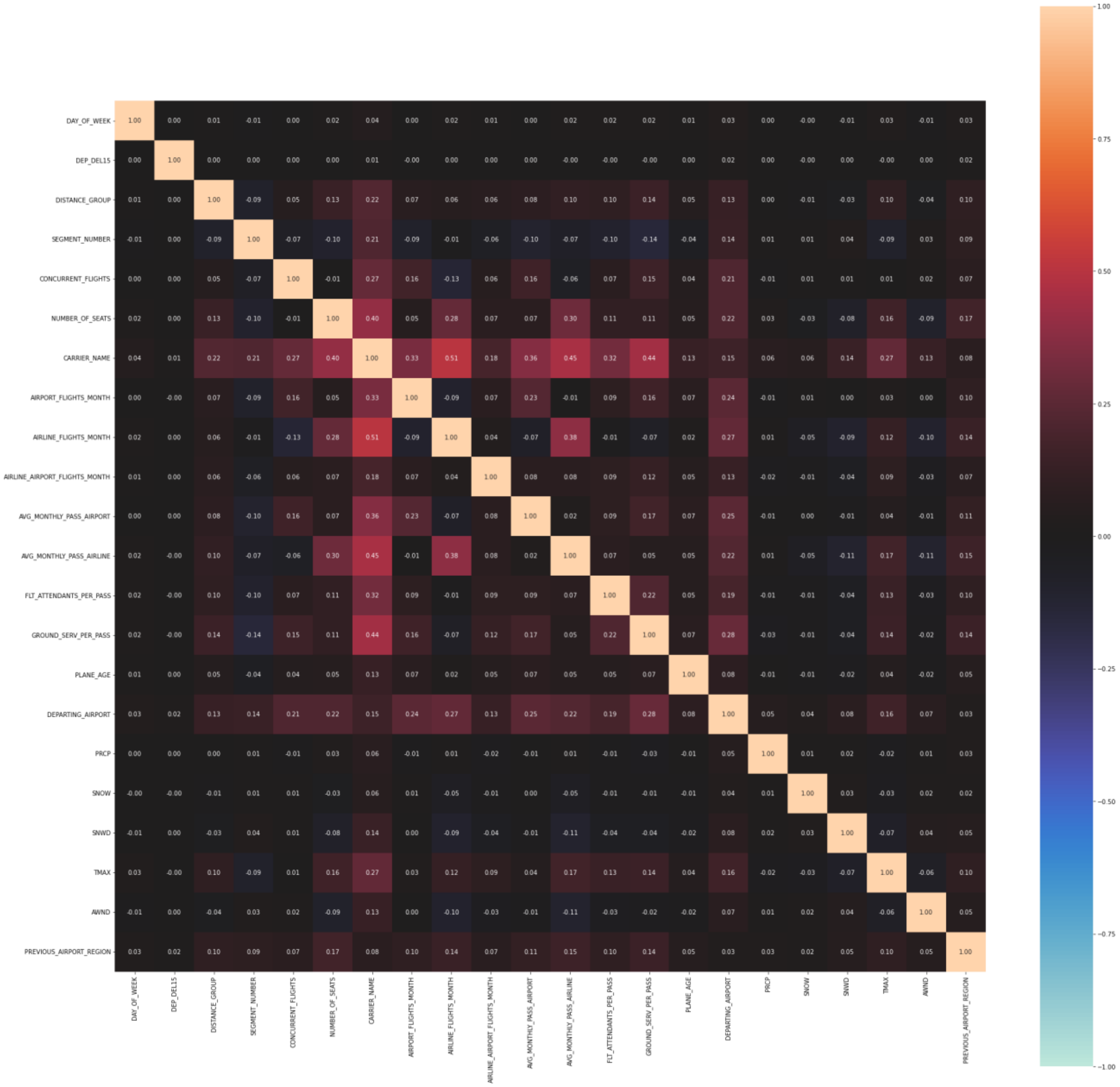
As we can see from the above correlation heatmap, there is no major correlation between the numerical variables.

4.3 Categorical feature Correlation



As we can see in the above graph there is no major correlation between categorical variables as well.

4.4 Association within features



	DAY_OF_WEEK	DEP_DEL15	DISTANCE_GROUP	SEGMENT_NUMBER	CONCURRENT_FLIGHTS	NUMBER_OF_SEATS	CARRIER_NAME	AIRPORT_FLIGHTS_MONTH	AIRLINE_FLIGHTS_MONTH	AIRLINE_AIRPORT_FLIGHTS_MONTH	AVG_MONTHLY_PASS_AIRPORT	AVG_MONTHLY_PASS_AIRLINE	FLT_ATTENDANTS_PER_PASS	GROUND_SERV_PER_PASS	PLANE_AGE	DEPARTING_AIRPORT	PRCP	SNOW	SNWD	TMAX	AWND	PREVIOUS_AIRPORT_REGION
DAY_OF_WEEK	1.00	0.00	0.01	-0.01	0.00	0.02	0.04	0.00	0.02	0.01	0.00	0.02	0.02	0.02	0.01	0.03	0.00	-0.00	-0.01	0.03	-0.01	0.03
DEP_DEL15	0.00	1.00	0.00	0.00	0.00	0.00	0.01	-0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	0.02	0.00	-0.00	0.00	-0.00	0.00	0.02
DISTANCE_GROUP	0.01	0.00	1.00	-0.09	0.05	0.13	0.22	0.07	0.06	0.06	0.08	0.10	0.10	0.14	0.05	0.13	0.00	-0.01	-0.03	0.10	-0.04	0.10
SEGMENT_NUMBER	-0.01	0.00	-0.09	1.00	-0.07	-0.10	0.21	-0.09	-0.01	-0.06	-0.10	-0.07	-0.10	-0.14	-0.04	0.14	0.01	0.01	0.04	-0.09	0.03	0.09
CONCURRENT_FLIGHTS	0.00	0.00	0.05	-0.07	1.00	-0.01	0.27	0.16	-0.13	0.06	0.16	-0.06	0.07	0.15	0.04	0.21	-0.01	0.01	0.01	0.01	0.02	0.07
NUMBER_OF_SEATS	0.02	0.00	0.13	-0.10	-0.01	1.00	0.40	0.05	0.28	0.07	0.07	0.30	0.11	0.11	0.05	0.22	0.03	-0.03	-0.08	0.16	-0.09	0.17
CARRIER_NAME	0.04	0.01	0.22	0.21	0.27	0.40	1.00	0.33	0.51	0.18	0.36	0.45	0.32	0.44	0.13	0.15	0.06	0.06	0.14	0.27	0.13	0.08
AIRPORT_FLIGHTS_MONTH	0.00	-0.00	0.07	-0.09	0.16	0.05	0.33	1.00	-0.09	0.07	0.23	-0.01	0.09	0.16	0.07	0.24	-0.01	0.01	0.00	0.03	0.00	0.10
AIRLINE_FLIGHTS_MONTH	0.02	0.00	0.06	-0.01	-0.13	0.28	0.51	-0.09	1.00	0.04	-0.07	0.38	-0.01	-0.07	0.02	0.27	0.01	-0.05	-0.09	0.12	-0.10	0.14
AIRLINE_AIRPORT_FLIGHTS_MONTH	0.01	0.00	0.06	-0.06	0.06	0.07	0.18	0.07	0.04	1.00	0.08	0.08	0.09	0.12	0.05	0.13	-0.02	-0.01	-0.04	0.09	-0.03	0.07
AVG_MONTHLY_PASS_AIRPORT	0.00	0.00	0.08	-0.10	0.16	0.07	0.36	0.23	-0.07	0.08	1.00	0.02	0.09	0.17	0.07	0.25	-0.01	0.00	-0.01	0.04	-0.01	0.11
AVG_MONTHLY_PASS_AIRLINE	0.02	-0.00	0.10	-0.07	-0.06	0.30	0.45	-0.01	0.38	0.08	0.02	1.00	0.07	0.05	0.05	0.22	0.01	-0.05	-0.11	0.17	-0.11	0.15
FLT_ATTENDANTS_PER_PASS	0.02	-0.00	0.10	-0.10	0.07	0.11	0.32	0.09	-0.01	0.09	0.09	0.07	1.00	0.22	0.05	0.19	-0.01	-0.01	-0.04	0.13	-0.03	0.10
GROUND_SERV_PER_PASS	0.02	-0.00	0.14	-0.14	0.15	0.11	0.44	0.16	-0.07	0.12	0.17	0.05	0.22	1.00	0.07	0.28	-0.03	-0.01	-0.04	0.14	-0.02	0.14
PLANE_AGE	0.01	0.00	0.05	-0.04	0.04	0.05	0.13	0.07	0.02	0.05	0.07	0.05	0.05	0.07	1.00	0.08	-0.01	-0.01	-0.02	0.04	-0.02	0.05
DEPARTING_AIRPORT	0.03	0.02	0.13	0.14	0.21	0.22	0.15	0.24	0.27	0.13	0.25	0.22	0.19	0.28	0.08	1.00	0.05	0.04	0.08	0.16	0.07	0.03
PRCP	0.00	0.00	0.00	0.01	-0.01	0.03	0.06	-0.01	0.01	-0.02	-0.01	0.01	-0.01	-0.03	-0.01	0.05	1.00	0.01	0.02	-0.02	0.01	0.03
SNOW	-0.00	-0.00	-0.01	0.01	0.01	-0.03	0.06	0.01	-0.05	-0.01	0.00	-0.05	-0.01	-0.01	-0.01	0.04	0.01	1.00	0.03	-0.03	0.02	0.02
SNWD	-0.01	0.00	-0.03	0.04	0.01	-0.08	0.14	0.00	-0.09	-0.04	-0.01	-0.11	-0.04	-0.04	-0.02	0.08	0.02	0.03	1.00	-0.07	0.04	0.05
TMAX	0.03	-0.00	0.10	-0.09	0.01	0.16	0.27	0.03	0.12	0.09	0.04	0.17	0.13	0.14	0.04	0.16	-0.02	-0.03	-0.07	1.00	-0.06	0.10
AWND	-0.01	0.00	-0.04	0.03	0.02	-0.09	0.13	0.00	-0.10	-0.03	-0.01	-0.11	-0.03	-0.02	-0.02	0.07	0.01	0.02	0.04	-0.06	1.00	0.05
PREVIOUS_AIRPORT_REGION	0.03	0.02	0.10	0.09	0.07	0.17	0.08	0.10	0.14	0.07	0.11	0.15	0.10	0.14	0.05	0.03	0.03	0.02	0.05	0.10	0.05	1.00

As we can observe from the above heatmap, significant associations can be found between following pairs:

- CARRIER_NAME and NUMBER_OF_SEATS: 0.4 association coefficient
- CARRIER_NAME and AIRLINE_FLIGHTS_MONTH: 0.51 association coefficient
- CARRIER_NAME and AVG_MONTHLY_PASS_AIRLINE: 0.45 association coefficient
- CARRIER_NAME and GROUND_SERV_PER_PASS: 0.44 association coefficient

5. FUTURE WORK

- After the preliminary data analysis, we found our insignificant variables. This will later be proven with the help of statistical tests.
- After the split done in the data, appropriate data transforming step will be taken to get to the prediction more precisely.
- Considering the above two steps, important attributes will be taken forward to build a base model for the future predictions.