## **AIRLINES DATASET TO PREDICT A FLIGHT DELAY.**

## **Business Understanding**

### **Business Overview**

With the instantaneous development of today’s air transport industry, airline travel has in the recent past become a very popular mode of travel. There are so many benefits of traveling by air as compared to other forms of travel. Most important of these benefits being safety and speed. Flying has also become quite affordable in the past decade and it is the most preferred by people traveling both far and short distances.

Despite these benefits there are still some disadvantages to flying, among them being delays in flights. Delay in flight is an inevitable phenomenon which has significant negative economic effects on airports, passengers and travel agencies.

With this study, we aim at predicting if a flight will be delayed or not depending on the factors that we will analyze. Predicting flight delays will help improve airline operations and attain passenger satisfaction, which will result in a positive economic impact on passengers, airlines, airports and travel agencies.

### **Business Objective**

The main objective of this report is to build a model that predicts whether a given flight will be delayed given the information in the scheduled departure.

**Business Success Criteria**

To do a thorough analysis and build a model to predict whether a given flight will be delayed given the information in the scheduled departure.

**Assessing the Situation**

1. **Resource Inventory**
2. Datasets:

* <https://www.kaggle.com/datasets/jimschacko/airlines-dataset-to-predict-a-delay>

1. Software used ( Github, Google Collaboratory, Google Docs, IBM® SPSS® Modeler)
2. Resources available for reference is found ([here](https://www.kaggle.com/datasets/jimschacko/airlines-dataset-to-predict-a-delay))

**2. Assumptions**

* The data provided is correct and up to date.
* All my data is relevant for analysis.
* There are no existing typos in my dataset.

**3. Constraints**

There are no constraints

### **Data Mining Goals**

Our data mining goals for this project are as follows:

* Determine which airline has the most delays.
* Determine from which airport there were most delays.
* Determine the average delay time for each airline.
* Determine which day of the week had the most delays.
* Determine if there is a relationship/ correlation between the airport and airline which may lead to delays in flights.

**Data Mining Success Criteria**

Our success criteria will be measured by the following criteria;

* Create a clear picture of how close the company is to its business goals and strategies to ensure overall business success and growth.
* Easy application of the analytics results to gain substantial and sustainable benefits for the company.
* Customer satisfaction, brand awareness or customer engagement to help evaluate the efficiency of the use and consumption of the product(Airlines).

## **Data Understanding**

### **Data Understanding Overview**

For this project, we are using the availed dataset. Data was extracted from Kaggle, where the availability information was available in real-time. The information of the scheduled departures is highlighted below.

**Different Feature Names**

1.Airline

2.Flight

3.Airport From

4.Airport To

5.DayOfWeek

6.Time

**Abbreviations for different Airlines**

Alaska Airlines AS / ASA

American Airlines AA/AAL

Air Canada AC/ACA

Aeromexico AM / AMX

Continental Airlines CO / COA

Delta Airlines DL / DAL

FedEx FX / FDX

Hawaiian Airlines HA / HAL

Northwest Airlines NW / NWA

Polar Air Cargo PO / PAC

Southwest Airlines SW / SWA

United Airlines UA / UAL

United Parcel (UPS) 5X / UPS

Virgin Atlantic VS / VIR

VivaAerobús VB / VIV

WestJet WS / WJ

ATL - Hartsfield-Jackson Atlanta International Airport - Georgia

AUS - Austin-Bergstrom International Airport - Texas

BNA - Nashville International Airport - Tennessee

BOS - Boston Logan International Airport - Massachusetts

BWI - Baltimore-Washington International Thurgood Marshall Airport - Washington

CLT - Charlotte Douglas International Airport - North Carolina

DAL - Dallas Love Field - Texas

DCA - Ronald Reagan Washington National Airport - Arlington, Virginia

DEN - Denver International Airport - Colorado

DFW - Dallas/Fort Worth International Airport - Texas

DTW - Detroit Metropolitan Airport - Michigan

EWR - Newark Liberty International Airport - New Jersey

FLL - Fort Lauderdale–Hollywood International Airport - Florida

HNL - Daniel K. Inouye International Airport - Honolulu, Hawaii

HOU - William P. Hobby Airport - Houston, Texas

IAD - Dulles International Airport - Virginia

IAH - George Bush Intercontinental Airport - Houston, Texas

JFK - John F. Kennedy International Airport - Queens, New York

LAS - McCarran International Airport - Las Vegas, Nevada

LAX - Los Angeles International Airport - California

LGA - LaGuardia Airport - Queens, New York

MCO - Orlando International Airport - Florida

MDW - Chicago Midway International Airport - Illinois

MIA - Miami International Airport - Florida

MSP - Minneapolis–Saint Paul International Airport - Minnesota

MSY - Louis Armstrong New Orleans International Airport - Louisiana

OAK - Oakland International Airport - California

ORD - O'Hare International Airport - Chicago, Illinois

PDX - Portland International Airport - Oregon

PHL - Philadelphia International Airport - Pennsylvania

PHX - Phoenix Sky Harbor International Airport - Arizona

RDU - Raleigh-Durham International Airport - North Carolina

SAN - San Diego International Airport - California

SEA - Seattle–Tacoma International Airport - Washington

SFO - San Francisco International Airport - California

SJC - Norman Y. Mineta San Jose International Airport - California

SLC - Salt Lake City International Airport - Utah

SMF - Sacramento International Airport - California

STL - St. Louis Lambert International Airport - Missouri

TPA - Tampa International Airport - Florida

### **Data Description**

We have one dataset available for this project. A detailed description of the datasets is provided as follows:

* This dataset consists of 539383 rows instances and 8 columns. The data contains the following fields:

1. Airline (different types of commercial airlines)
2. Flight (type of aircraft)
3. Airport From (source airport)
4. Airport To (destination airport)
5. Day of the week
6. Time (time of day/ time the flight should have been taken)
7. Length (how long the flight was delayed)
8. Delay (whether the flight is delayed or not)

### **Verifying Data Quality**

There were no missing values in our dataset.

See below visualization on the sum of all null values.



Our dataset had no duplicates as well.

## **Data Preparation**

These are the steps followed in preparing the data

#### **Loading Data**

Loading the dataset from the CSV file and then creating a dataframe to be used.

1. **Cleaning Data**

* **Validity check**

Checking our columns;

**Index(['id', 'Airline', 'Flight', 'AirportFrom', 'AirportTo', 'DayOfWeek',**

**'Time', 'Length', 'Delay'],**

**dtype='object')**

* **Accuracy check**

**<class 'pandas.core.frame.DataFrame'>**

**RangeIndex: 539383 entries, 0 to 539382**

**Data columns (total 9 columns):**

**# Column Non-Null Count Dtype**

**--- ------ -------------- -----**

**0 id 539383 non-null int64**

**1 Airline 539383 non-null object**

**2 Flight 539383 non-null int64**

**3 AirportFrom 539383 non-null object**

**4 AirportTo 539383 non-null object**

**5 DayOfWeek 539383 non-null int64**

**6 Time 539383 non-null int64**

**7 Length 539383 non-null int64**

**8 Delay 539383 non-null int64**

* **Completeness check**

There are no null values.

* **Consistency check**

Using the *df.duplicated().sum()* to identify duplicates in my data. No duplicates were found.

* **Uniformity check**

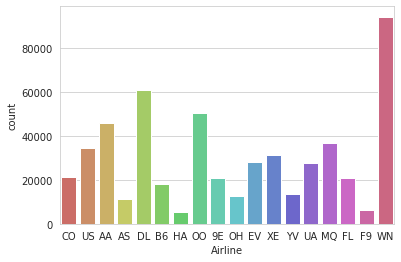
ID column was dropped during the cleaning process. All entries were relevant for analysis. The outliers will be retained. We have both delay and non delays having outliers due to various causes of delays like; Security reasons, Weather, Technical reasons- airplane parts etc, Flight crew delays.

1. **Analysis**

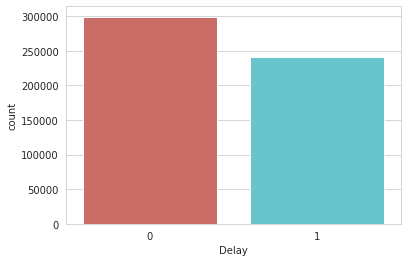
During our analysis, the following was covered;

1. Univariate
2. Bivariate
3. Multivariate

**Univariate**

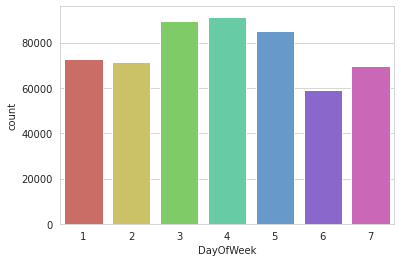
****

WN(Westjet Airline) airline had the most trips made while HA(Hawaiian Airlines) had the least.

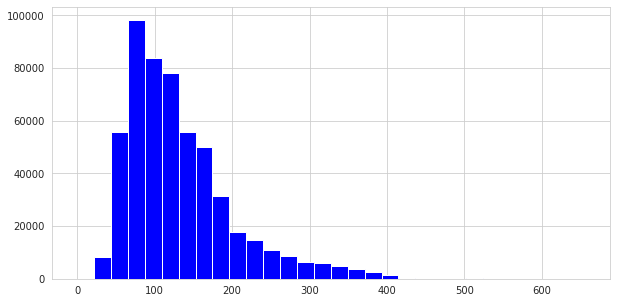
****

This also implies that this will be a binary classification problem where a “0” means that the flight arrives on time, and “1” means that the flight will be delayed.

We can also see more flights were on time compared to those delayed.

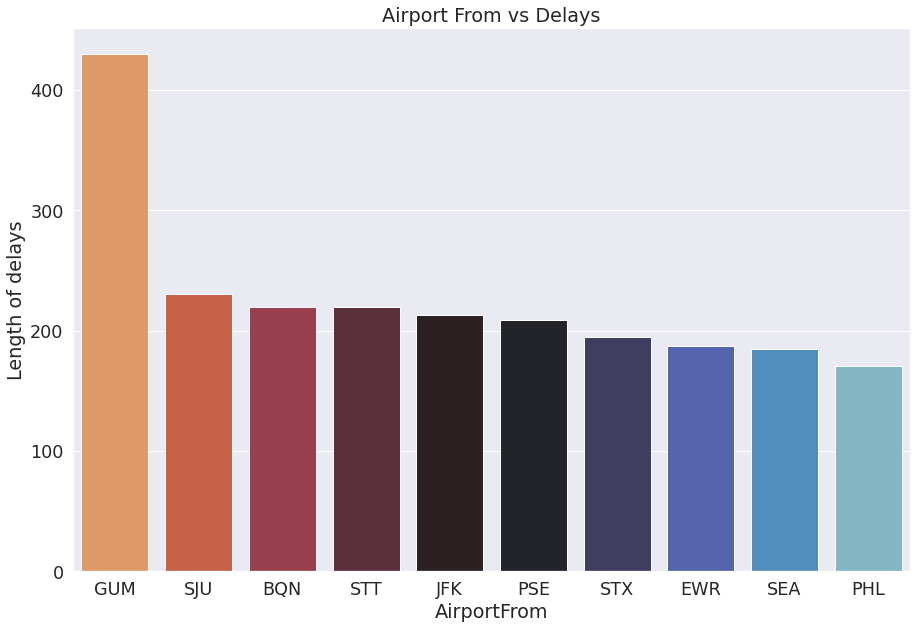
****

There were more flights on the fourth day of the week and least on the sixth day of the week



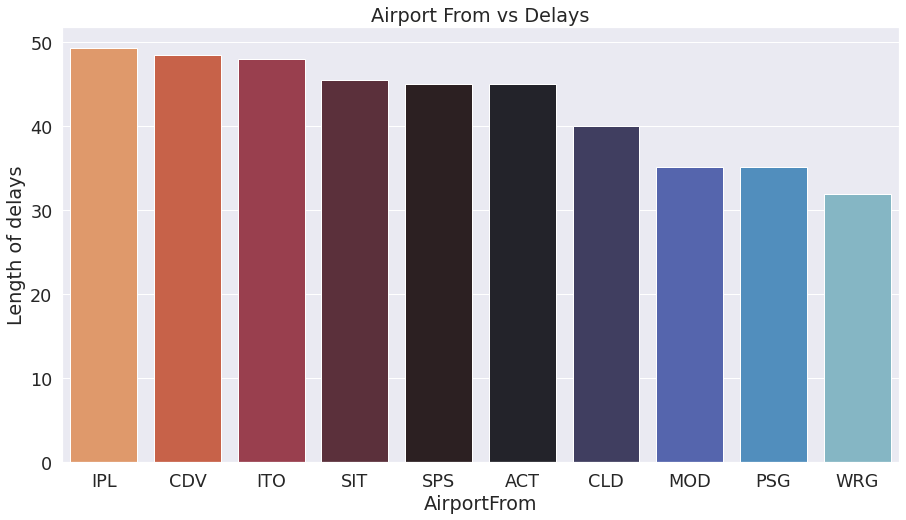
There were more flights from ATL (Hartsfield-Jackson Atlanta International Airport - Georgia ) airport.

**Bivariate**

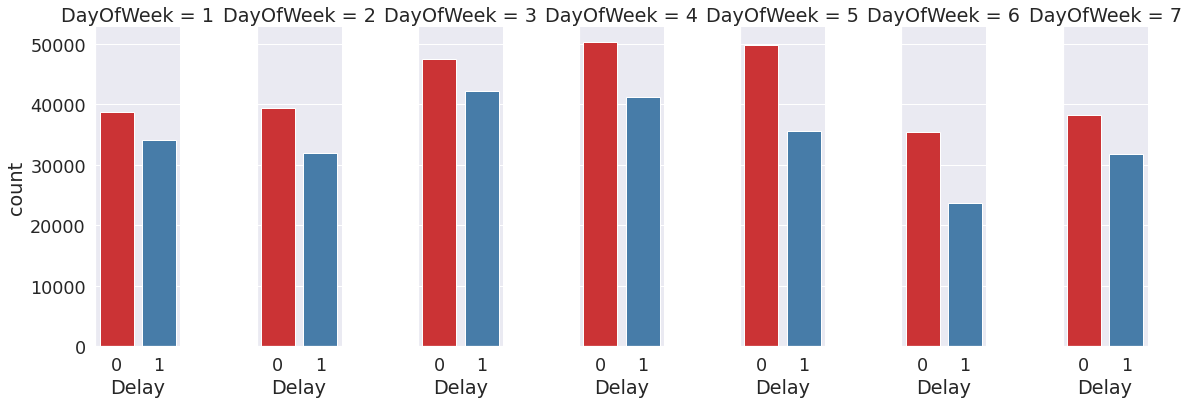


GUM Airport had the highest delays, being a US territory island in Micronesia it experiences some of the adverse weather such as crosswinds of 30-35 kts or more which are deemed unsafe for take\_off or landing.It also geographically lies in the path of typhoons ( a mature tropical cyclone that develops between 180° and 100°E in the Northern Hemisphere) where it has been hit some of the severe ones like Typhoon of 1900, Karen (1962), Pamela (1976), Paka (1997), and Pongsona (2002).

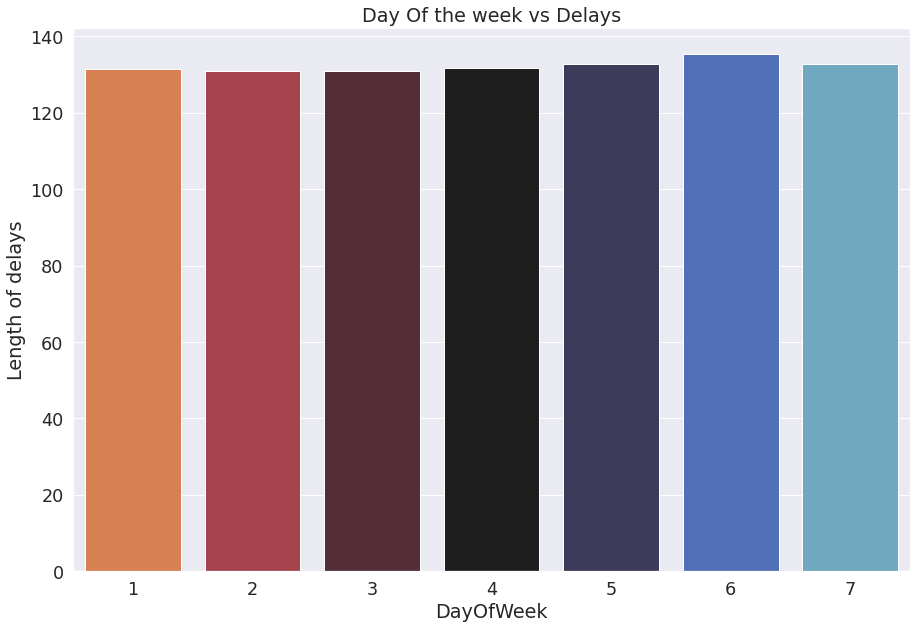
It also the home away from home of US military base who occupy 29% of the island

****

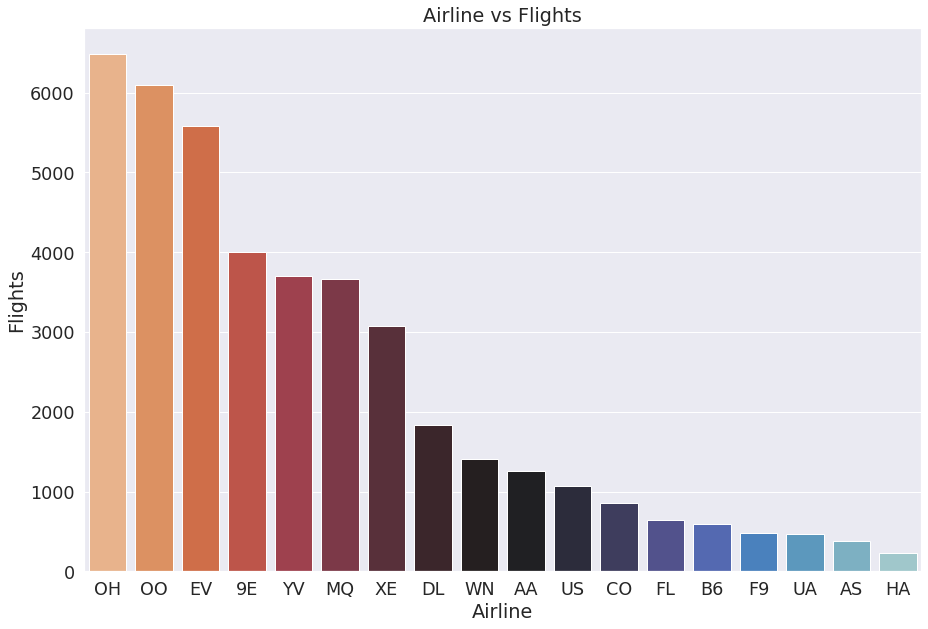
From our data set WRG (Wrangell Airport) had the least delay time. This is expected since previously it was among the airport with the least flights(traffic) and it also had least delay time compared to other with the least flights.



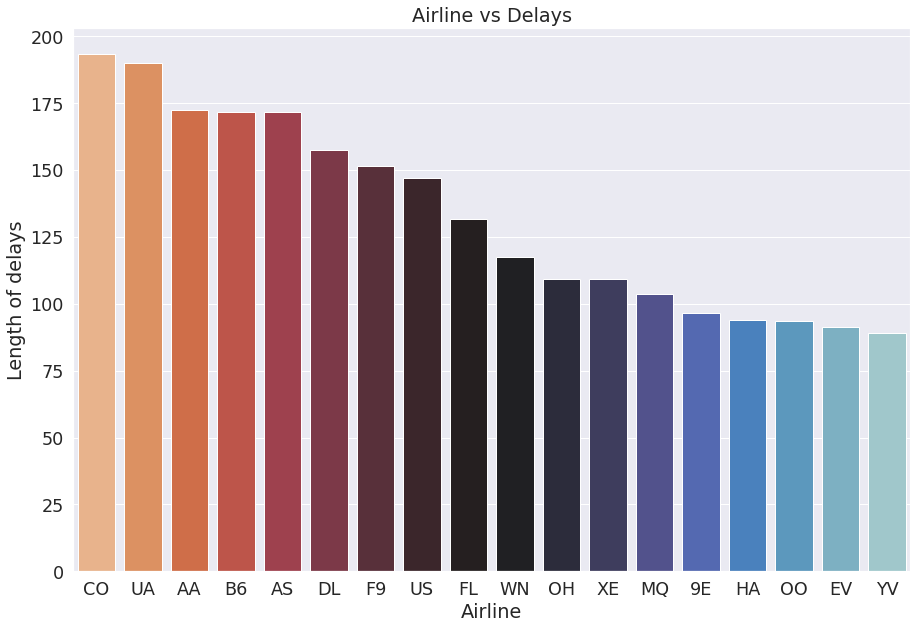
From a count plot view the day with most delays (1) was Day three and Day four.



From the Delay length point of view Day six had the maximum length of delay.

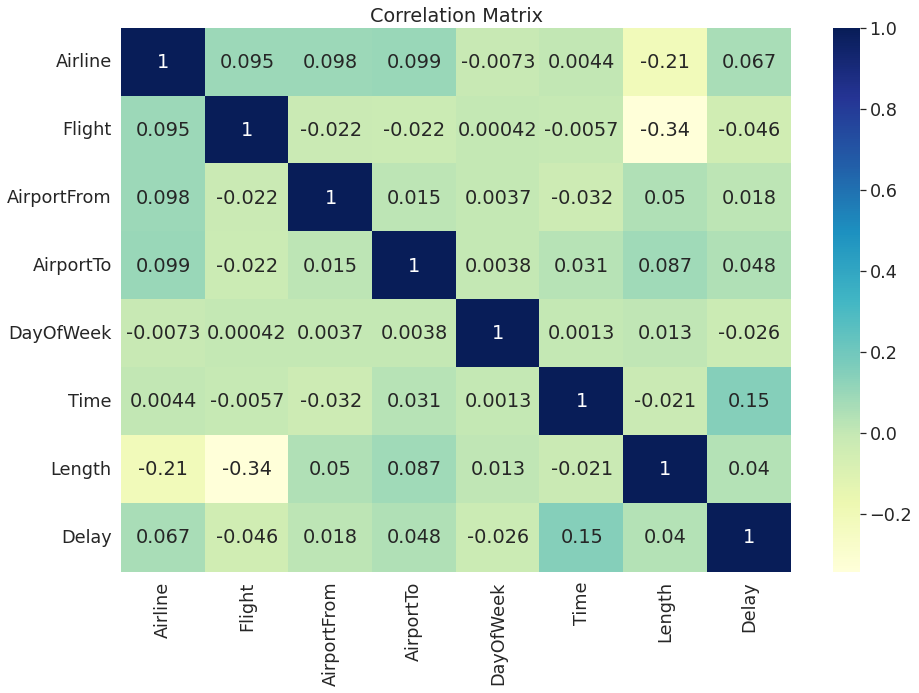


OH (Pacific Southwest) Airlines has the highest number of flights.Some of the contributing factors to highest trips is it was the first large discount airline in the United States. It also has its headquarters in Vandalia, Ohio, United States. It not only makes local and state trips, it also operates internationally.

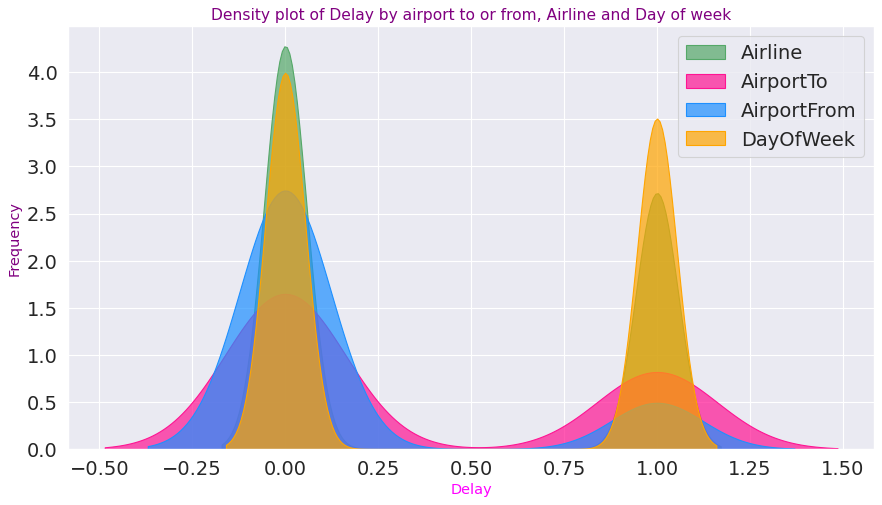


CO (Continental) airlines comes first with the maximum delay length in time. With its headquarter in Houston, Texas; an area affected by some adverse weather conditions like tonados are some of its contributing factors to delays.

**Multivariate**

****

There is no strong correlation between delay and features that might be leading to delay or not.



THe above graph tries to show the relationship between delay and various features. We can see that Airline and Dayof the week are some of the key factor in delay or no delay. This can be interpreted as some factor leading to delay such as weather. There could be storms on Tuesday but not on Wednesday leading to more delays on Tuesday unlike wednesday. On other factors such as Technical reasons, Flight crew delays are directly influenced by the airline.

The above analysis was done using Google Collab. The full analysis can be found here.[[Link](https://colab.research.google.com/drive/1rkEC2-ONOhLhkhuAENZOJso49E34KWn7?usp=sharing)]

1. **Modeling**

| **Model** | **Accuracy** | **After Hyperparameter Tuning** |
| --- | --- | --- |
| **Decision Trees (RandomForest)** | Our model is giving us an f1 score of 66% for non delayed class and 57% for the delayed class. | After tuning our parameters our model has improved by giving us an f1 score of 74% for no delayed class and 40% for delayed class. This means our model classifier works best for non delayed class. |
| **KNN** | Without tuning any parameter ou KNN model is giving us an f1 score of 69% for class 0 and 58% for class 1 | Our measure of success, which is f1 score, has decreased to 66% and 55% after tuning parameters. could mean this is not the best model for this dataset |
| **Naive Bayes****(Bernoulli)** | Using bernoulli we get a metric score of 55% with no parameters tuned | After tuning the model does a bit better giving us a metric score of 58% |
| **Naive Bayes****(Gaussian)** | On using Gaussian our model gives a better score with no parameters tuned of 59% | There is no change in metric score after parameter tuning. 59% |
| **Neural Networks** | Using the Multilayer perceptron our f1 score is 72% and 47% for our classes respectively. This is just a baseline; there is no activation or hyperparameter tuning done. | After activation and parameter tuning our model was able to get a better f1 score for second class 55% and 71% for first class |
| **XGBoost Classifier** | XGBoost Classifier gives a f1 score of 73% and 51% | Introducing early stopping as a form of parameter tuning. Our model hasn't improved it still gives us an f1 score of 73% and 51% |

## **Conclusion**

The best model is Neural network (MLp) followed by Random forest, however on challenging the given model with XGBoost, XGBoost gives the highest f1 score which is what we are looking for in our metric of success