**Flight Delay Prediction for aviation Industry using Machine learing**

# INTRODUCTION

# In the present world, the major components of any transportation system include passenger airline, cargo airline, and air traffic

# control system. With the passage of time, nations around the world have tried to evolve numerous techniques of improving the airline

# transportation system. This has brought drastic change in the airline operations. Flight delays occasionally cause inconvenience to the

# modern passengers . Every year approximately 20% of airline flights are canceled or delayed, costing passengers more than 20 billion

# dollars in money and their time.

# overview

* Over the last twenty years, air travel has been increasingly preferred among travelers, mainly because of its speed

and in some cases comfort. This has led to phenomenal growth in air traffic and on the ground. An increase in air

traffic growth has also resulted in massive levels of aircraft delays on the ground and in the air. These delays are

responsible for large economic and environmental losses.

* According to, taxi-out operations are responsible for 4,000 tons of hydrocarbons, 8,000 tons of nitrogen oxides and

45,000 tons of carbon monoxide emissions in the United States in 2007. Moreover, the economic impact of flight

delays for domestic flights in the US is estimated to be more than $19 Billion per year to the airlines and over $41

Billion per year to the national economy In response to growing concerns of fuel emissions and their negative impact

on health, there is active research in the aviation industry for finding techniques to predict flight delays accurately in

order to optimize flight operations and minimize delays.

**Data Analysis**

* In order to investigate what factor could have a significant influence on delays (potential correlations or a lack

thereof) we needed to analyse the data. Although the set had 31 pieces of data per sample, not all of them have

been relevant to delay time. For example, flight number hardly could influence any delay. Also, since features such

as taxi time or wheels-off time are directly related to departure time (it’s a standard set of procedures with more

or less fixed durations) these features should not be of any influence to our data set. For the data analysis, we

picked several features that could potentially have some correlations with our target and could enhance our

predictions.

**Purpose**

* Using a machine learning model, we can predict flight arrival delays. The input to our algorithm is rows of feature

vector like departure date, departure delay, distance between the two airports, scheduled arrival time etc. We then

use decision tree classifier to predict if the flight arrival will be delayed or not. A flight is delayed when difference

between scheduled and actual arrival times is greater than 15 minutes. Furthermore, we compare decision tree

classifier with logistic regression and a simple neural network for various figures of merit.

* Throughout the year 2015, there has been over 5,4 million domestic flights within the US. All of their metadata

are recorded and saved in the Department of Transportation's (DOT) Bureau of Transportation Statistics. Flight

delays cause significant financial and other losses to airlines, airports, and passengers. Their prediction is crucial

during the decision-making process for all players of American aviation industry. Therefore, predicting the

likelihood of delay based on flights' features bridges an important information asymmetry between airlines and

passengers. The primary use case of the algorithm will be: predicting a potential delay, on a given day, for a given

airport and airline.

**Data processing**

* Since we worked on a large data file, we had to prepare several datasets for different purposes. We had to fix some technical issues

with the set, primarily in the airport data - airport codes for October consisted of 5 numbers instead of two capital letter. After fixing

the dataset, we created the first subset - with only delayed flights that had a known reason for the delay.

* For the sake of simplicity, we also limited the number of airlines and airports to only include the major ones in the analysis.

However, we decided not to do that, as it would require us to arbitrarily determine limits of our dataset. For the data analysis part, we

created a set of 12,000 flights (trimmed down from 5.4m) by looping through each row and only taking each multiple of 480 since

5.4m/480≈12k 5.4m/480≈12k. Later on, we created several other files, one with 72,000 flights and one with 148,000. These files let

us compare the influence of data size on the accuracy of our prediction algorithm.

* It was especially important since our data has an asymmetry of delayed versus non-delayed flights, which causes bias against delays

in our predictions. It is for this reason that, for our final models, we used a total dataset of 406,559 flights, of which 204,760 delayed

and 201.799 non-delayed.

* The features we used for final implementation were: - Month (numeric) - Day of the week (numeric) - Departure time (numeric,

converted to full hours) - Airline (converted to dummy variables) - Departure Airport (converted to dummy variables) - Arrival

Airport (converted to dummy variables) The target we ended up using was: arrival time (discrete). Where class 0 is defined as “Any

delay smaller than or equal to 10 minutes”, and class 1 is defined as “Any delay larger than 10 minutes” Some other features we used

were: - Airtime (numeric) - Distance (numeric) - Scheduled arrival time (numeric).

* Some of the features, such as airport code or airline, were saved as categorical variables (strings), which cannot be an input to the

sklearn prediction functions. Therefore they had to be converted into so-called dummy variables. Every unique string in the feature

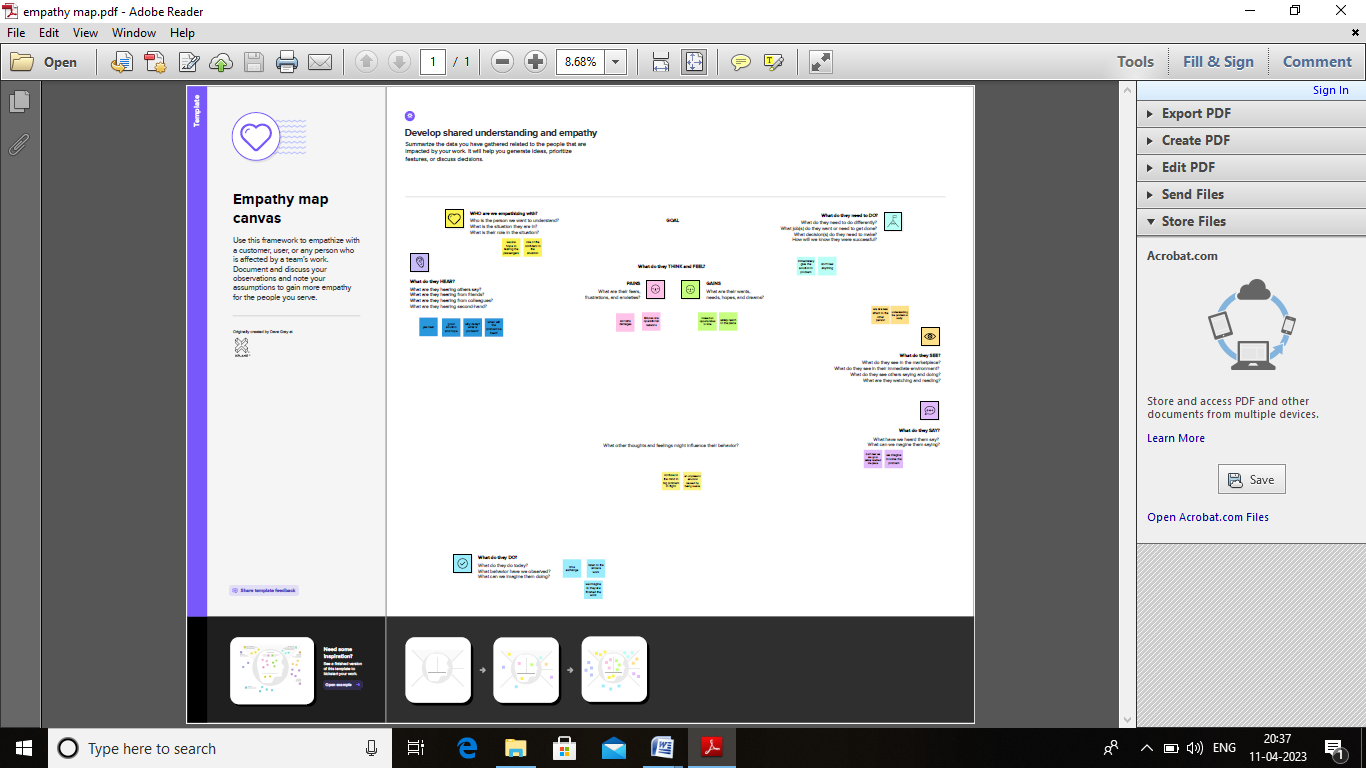
had to be converted into a separate feature, with a value of 1 for every data sample where that specific string shows up in the feature.

* We used the get\_dummies function from Pandas to do this. Aside from that, due to relatively large differences between numbers

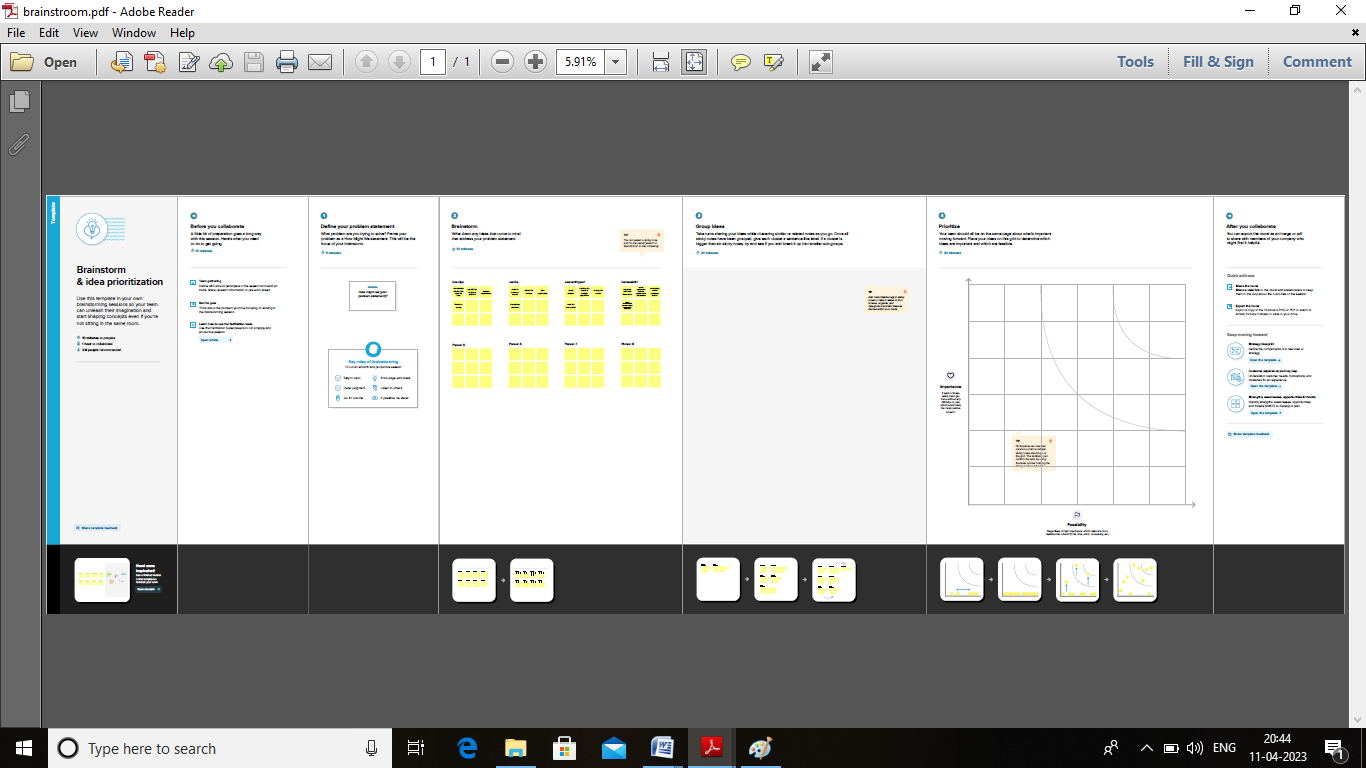
(hours vs day of the week for example) we had to scale our features, using the StandardScaler from sklearn.

**Problem Definition & Design Thinking**

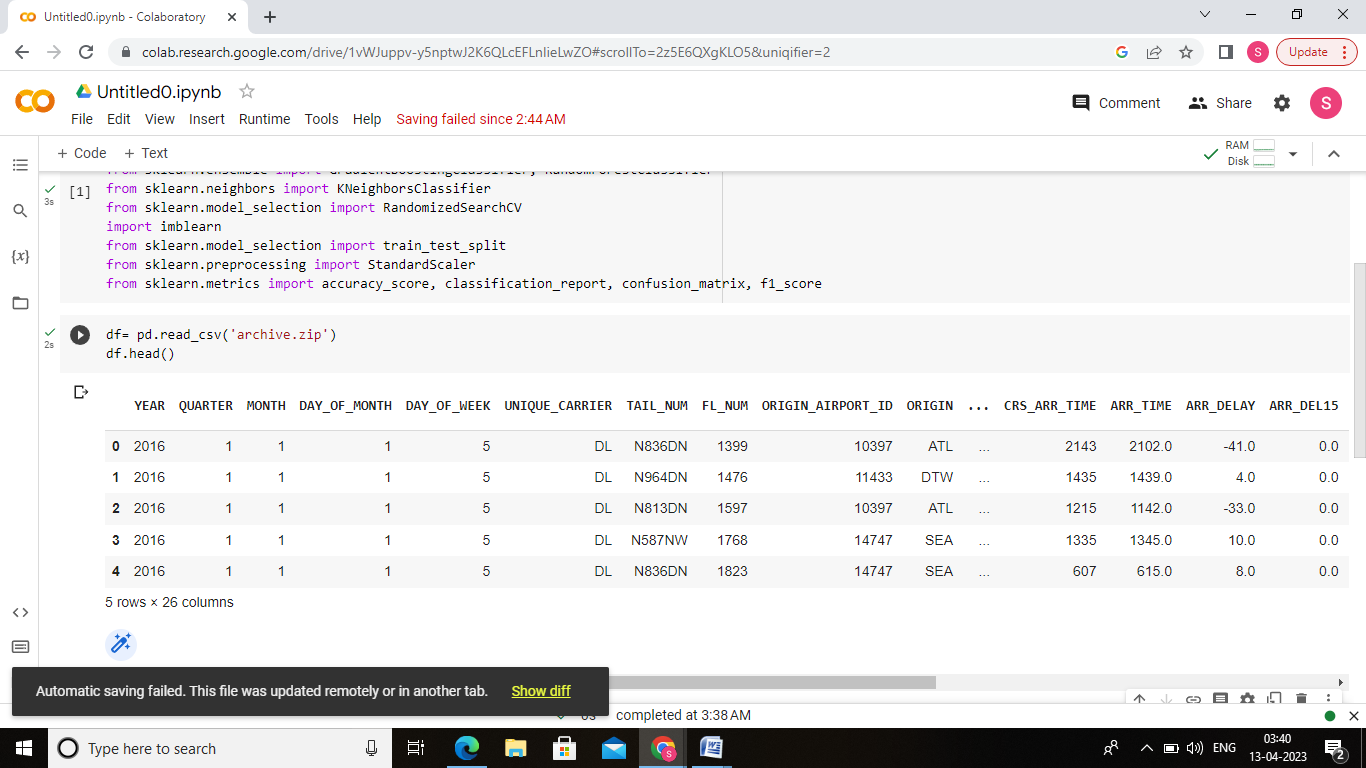
**Empathy Map**

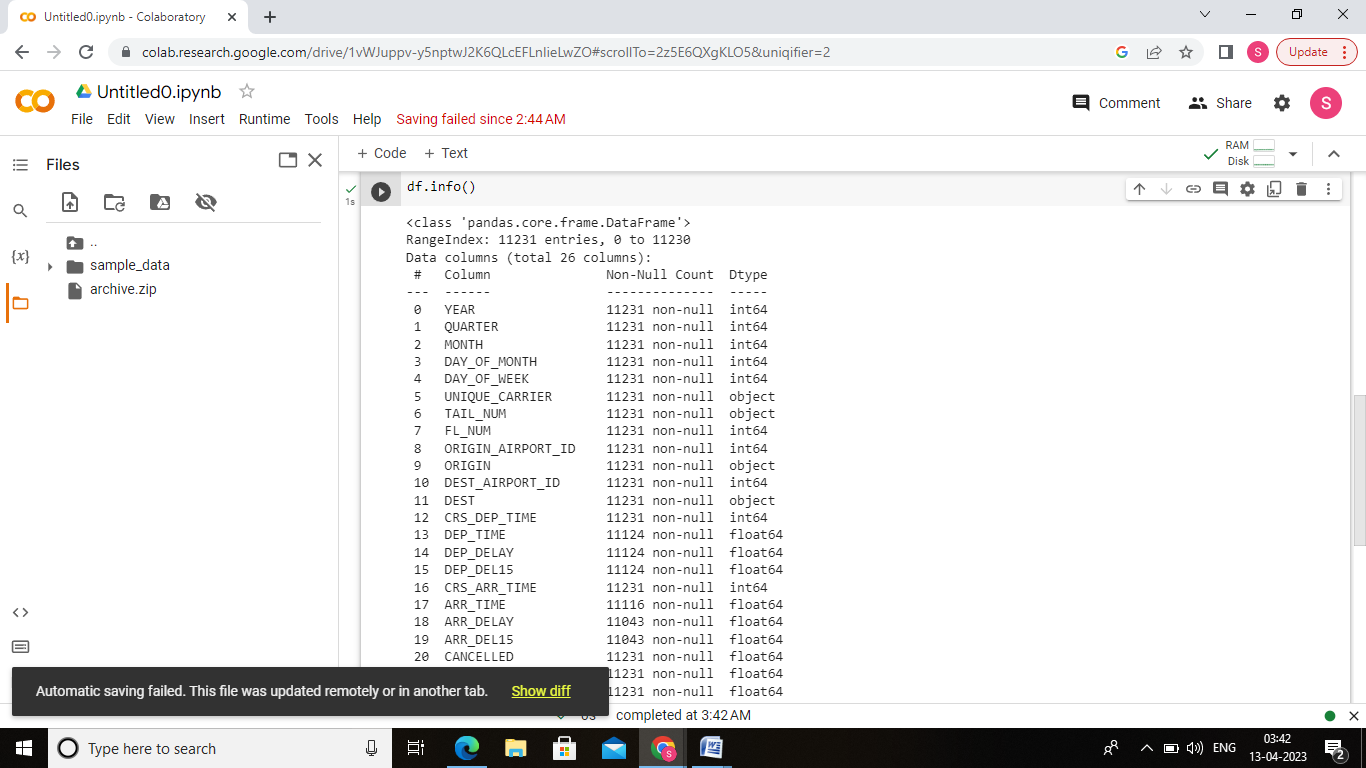


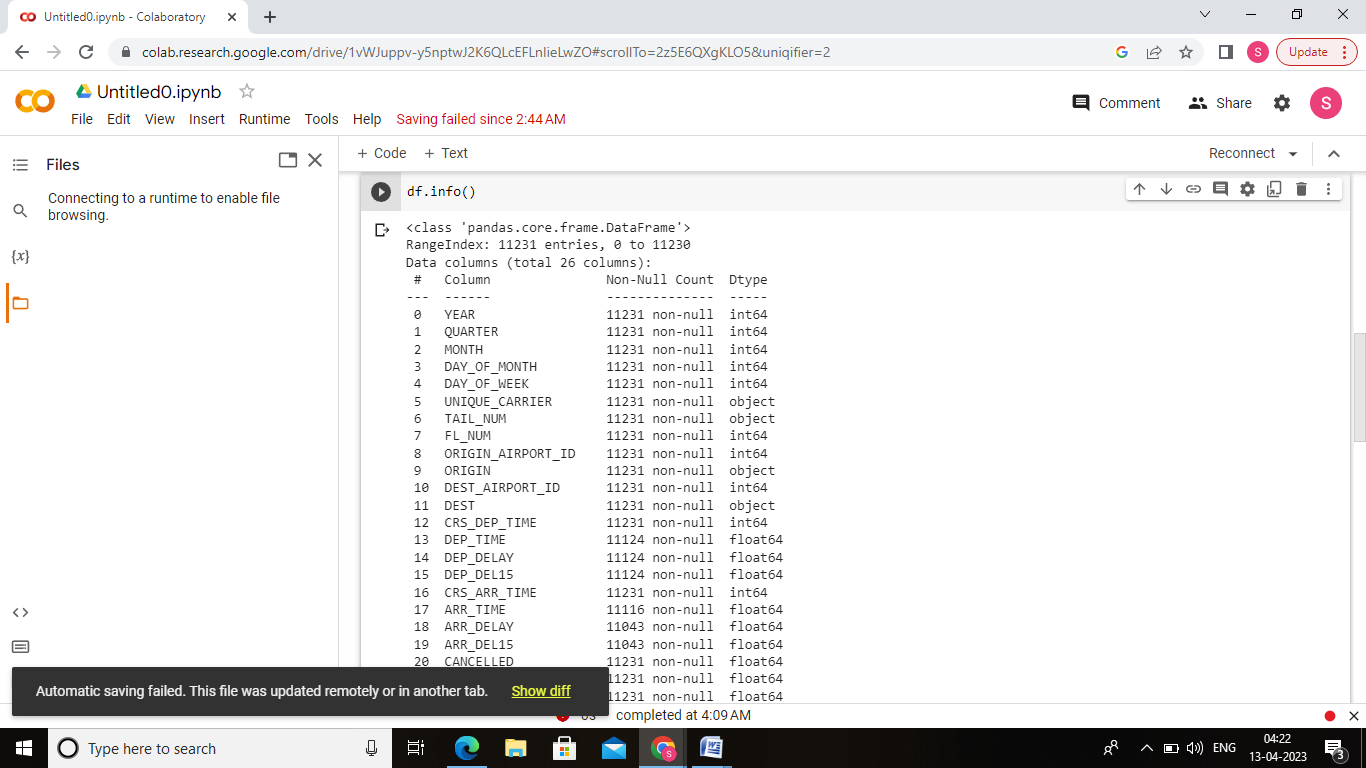
**Paste the Ideation & brainstorming map screenshot**

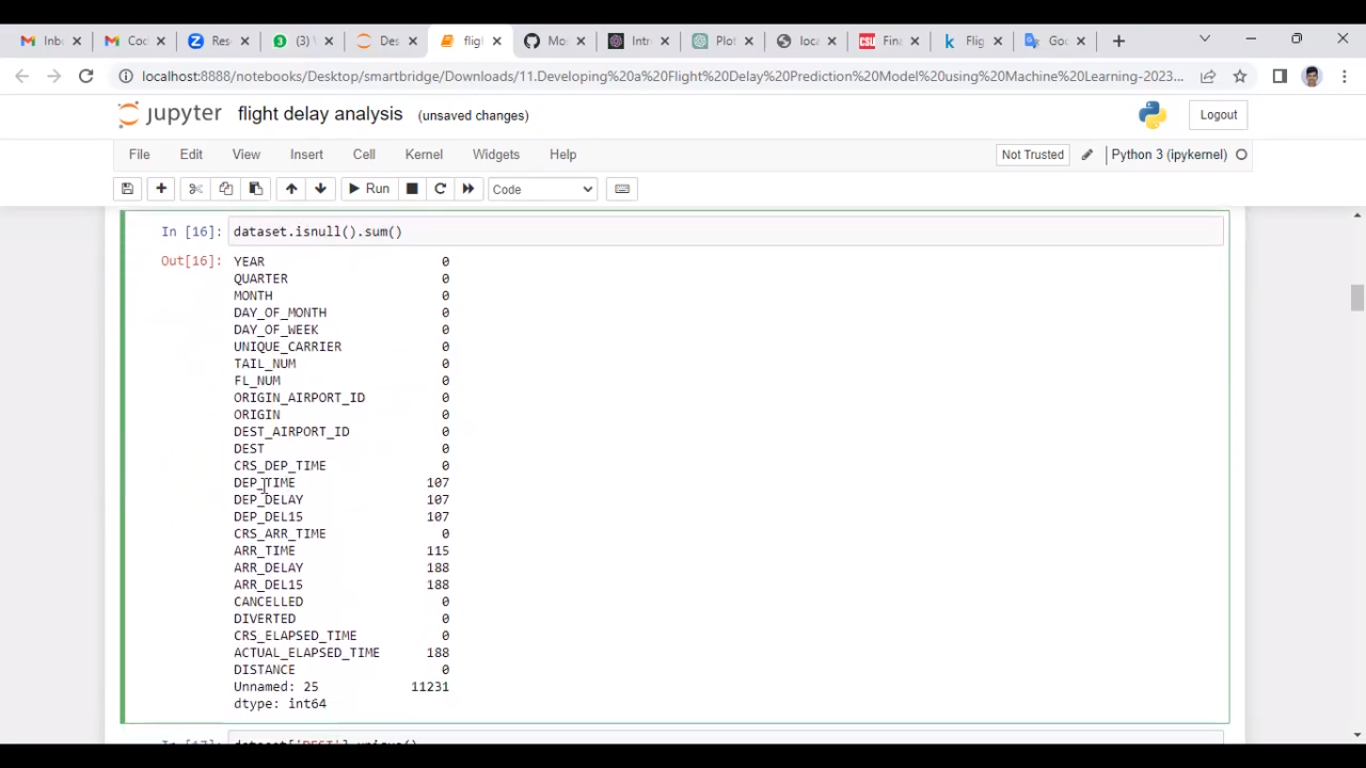
****

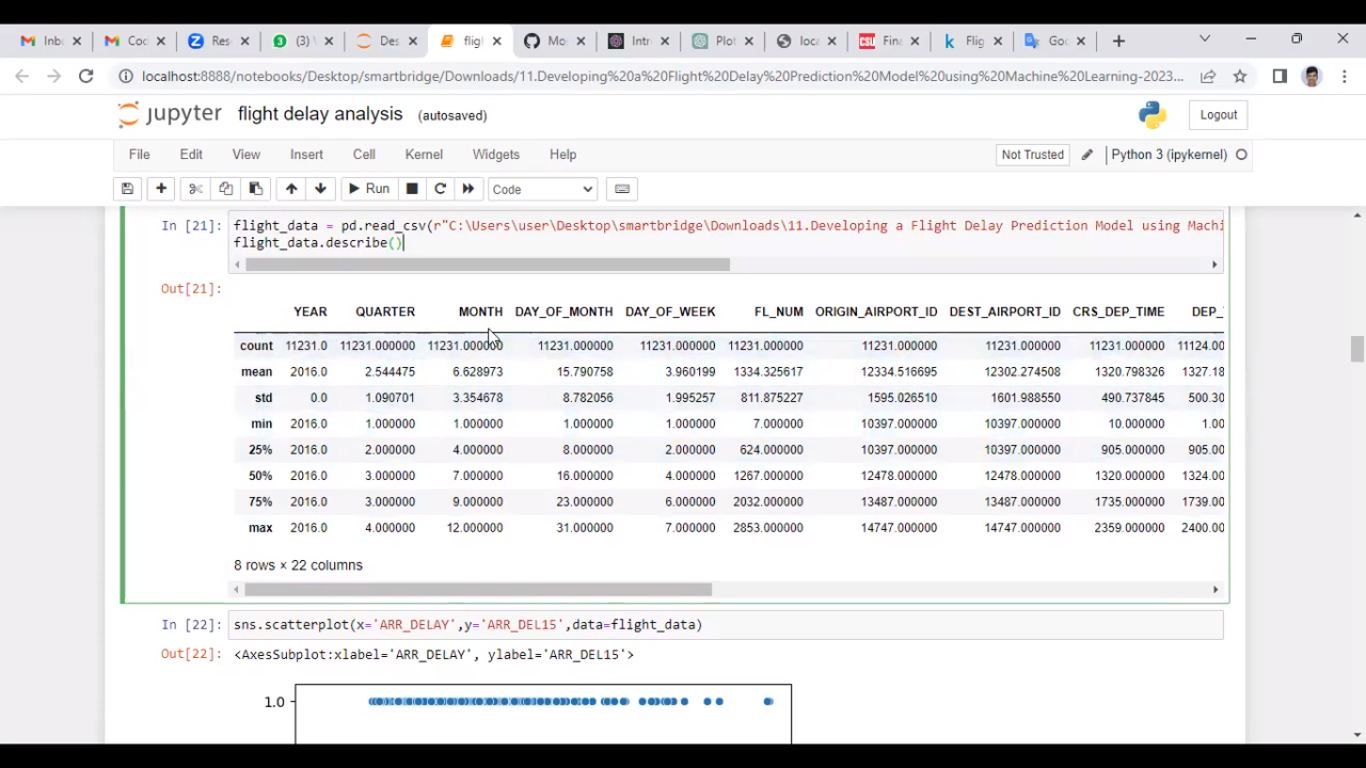
# RESULT

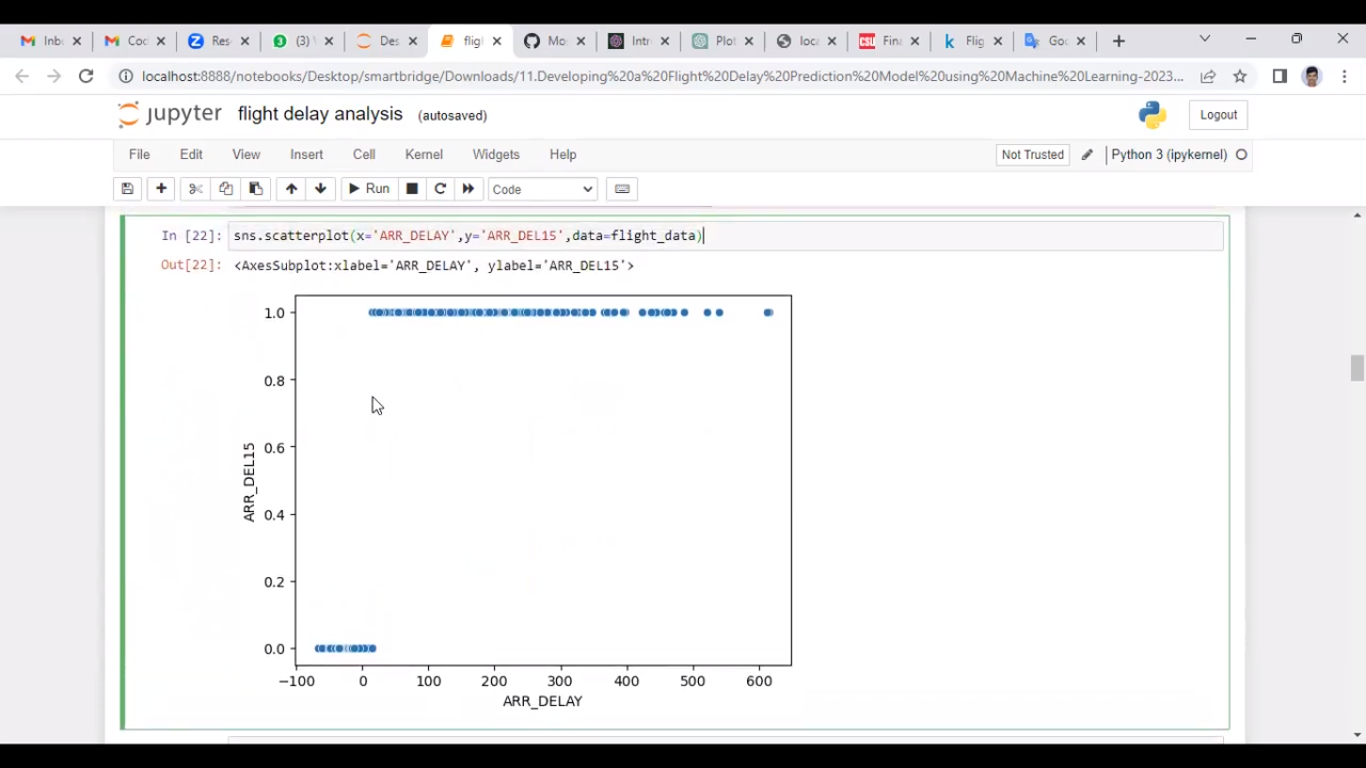


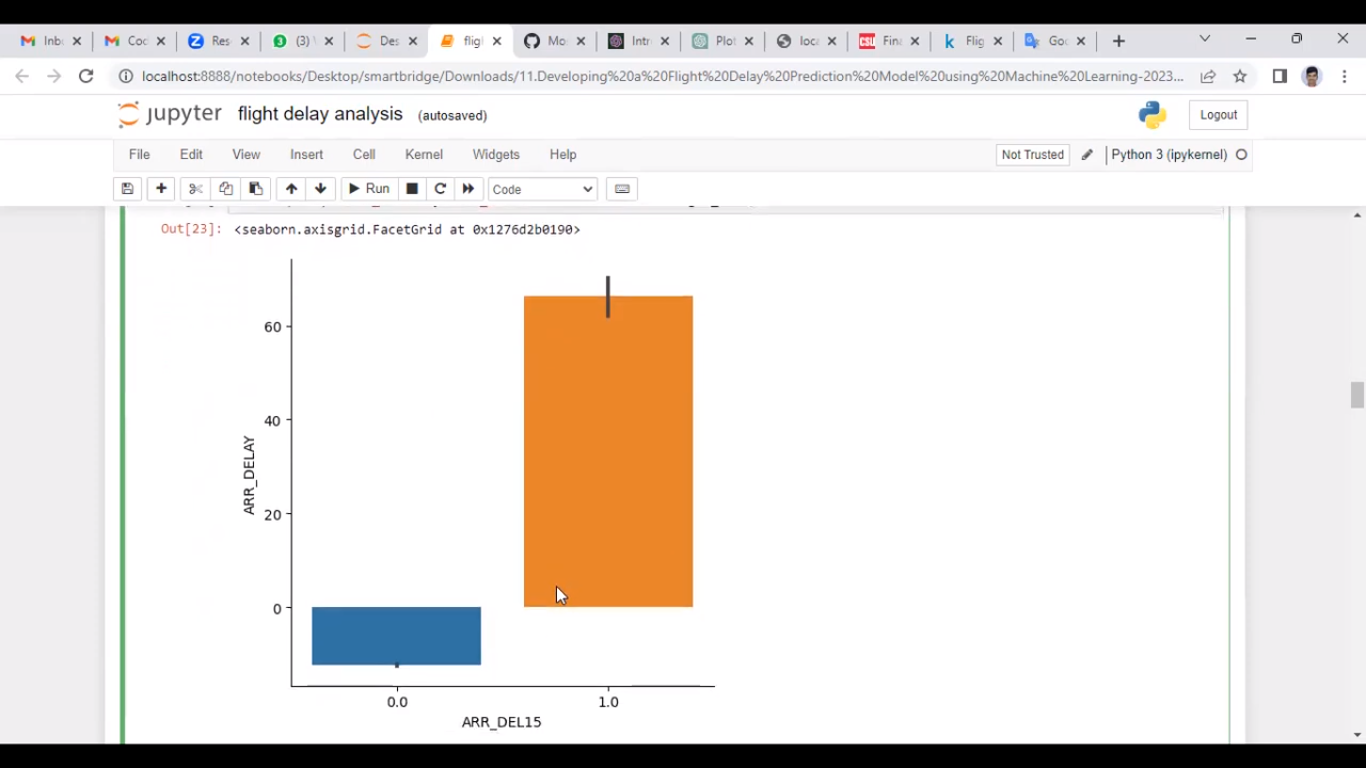












# 

# 

# 

# 

# ADVANTAGES & DISADVANTAGES

# Advantages

* Even if it is at the airport, you can enjoy your holiday destination for a little longer. Enjoy the local food, write and send a holiday card from the airport or sit back and relax and enjoy the vibrant life at the airport.
* The wait that comes with a [flight delay](https://www.euclaim.com/flight-delay-compensation) can last long. Luckily, in case of a delay of two hours or more, or with a delay due to exceptional circumstances, you are entitled to care. Food and drinks should be provided by the airline, usually by vouchers. If not, make sure to save your receipts to claim the extra costs from the airline. In case of a delay that forces you to take an overnight stay, you are also entitled to a hotel, including the transport to the accommodation.
* Reality teaches us that once you are home, you often do not take time to sort out your holiday photos let alone make a photo album. Are you delayed then sort them out to create that photo album back home that you are so looking forward.

# Disadvantages

# A delayed flight can be particularly costly to business travellers by causing them to miss scheduled appointments and interfering with

# other commitments. Furthermore, delayed passengers may suffer anger, frustration, and even air rage.

* Flight delays not only irritate air passengers and disrupt their schedules but also cause a decrease in efficiency, an increase in capital

costs, reallocation of flight crews and aircraft, and additional crew expenses

# Going Through Security.

# Cramped Economy Flights.

# Expensive Airports.

# Inconsiderate and Noisy Neighbors.

# Delays, Cancellations, and Lost Baggage.

# Passports, Bureaucracy, and Luggage Collection.

# Traveling To and From the Airport.

# APPLICATIONS

# There are several conclusions that can be drawn from the data: 1. We have two seasonal spikes in delayed flights.

# The first one is during the winter season - possibly because of the weather conditions and winter holidays. The second one is

# during the summer, especially in June - when many people travel for summer vacation.

# 2. For days of the week, Friday and Monday have the highest ratio of delayed to non-delayed flights, probably because the

# traffic is significantly higher right before and after the weekend.

# 3. Evening and late night flights are usually most likely to be delayed, either because of the increased traffic or other factors,

# such as low visibility, difficult weather conditions or workforce fatigue.

# 4. A majority of airports with a high number of delayed flights are regional, low-scale airports. It might be useful to check for

# inverse correlation between airport traffic and ratio of delayed to non-delayed flights.

# 5. Low-fare airlines tend to have a significantly greater amount of delays. The top three: Spirit Airlines (NK), Frontier (F9) and

# JetBlue (B6) are all low-cost operators that are more likely to compromise on customer experience to cut down maintenance

# costs. The biggest 3 airlines in the US - American, United and Delta are all less likely to delay, with Delta having the best score in this category.

**CONCLUSION**

# In this project, we use flight data, weather, and demand data to predict flight departure delay. Our result shows that the

# Random Forest method yields the best performance compared to the SVM model. Somehow the SVM model is very time

# consuming and does not necessarily produce better results. In the end, our model correctly predicts 91% of the non-delayed

# flights.

# However, the delayed flights are only correctly predicted 41% of time. As a result, there can be additional features

# related to the causes of flight delay that are not yet discovered using our existing data sources. In the second part of the

# project, we can see that it is possible to predict flight delay patterns from just the volume of concurrently published tweets,

# and their sentiment and objectivity. This is not unreasonable; people tend to post about airport delays on Twitter;

# 

# Reason that these posts would become more frequent, and more profoundly emotional, as the delays get worse. Without more

# data, we cannot make a robust model and find out the role of related factors and chance on these results. However, as a proof

# of concept, there is potential for these results. It may be possible to routinely use tweets to ascertain an understanding of

# concurrent airline delays and traffic patterns, which could be useful in a variety of circumstances.

# FUTURE SCOPE

# This project is based on data analysis from year 2008. A large dataset is available from 1987-2008 but handling a bigger

# dataset requires a great amount of preprocessing and cleaning of the data. Therefore, the future work of this project includes

# incorporating a larger dataset. There are many different ways to preprocess a larger dataset like running a Spark cluster over a

# server or using a cloud-based services like AWS and Azure to process the data. With the new advancement in the field of deep

# learning, we can use Neural Networks algorithm on the flight and weather data. Neural Network works on the pattern

# matching methodology. It is divided into three basic parts for data modelling that includes feed forward networks, feedback

# networks, and selforganization network. Feed-forward and feedback networks are generally used in the areas of prediction,

# pattern recognition, associative memory, and optimization calculation, whereas self-organization networks are generally used

# in cluster analysis. Neural Network offers distributed computer architecture with important learning abilities to represent

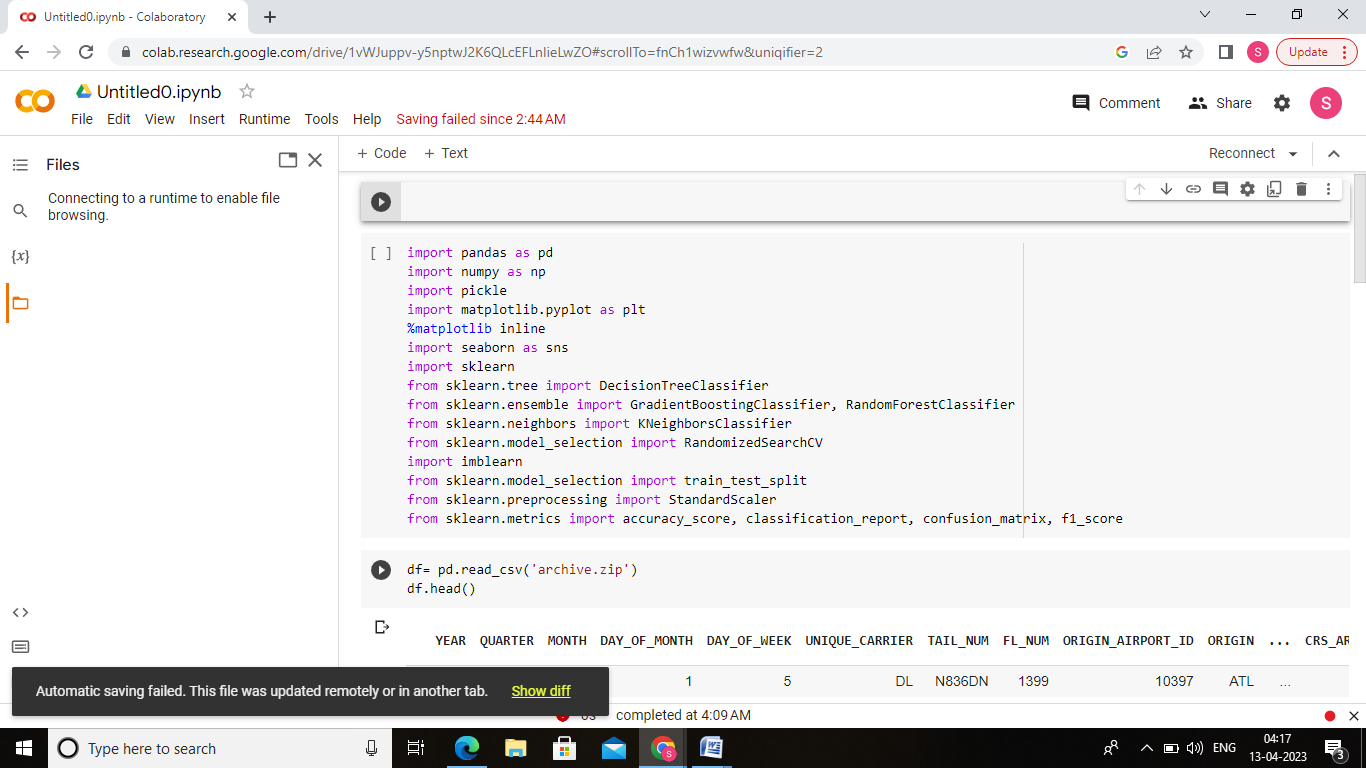
# nonlinear relationships. Also, the scope of this project is very much confined to flight and weather data of United States, but

# we can include more countries like China, India, and Russia. Expanding the scope of this project, we can also add the flight

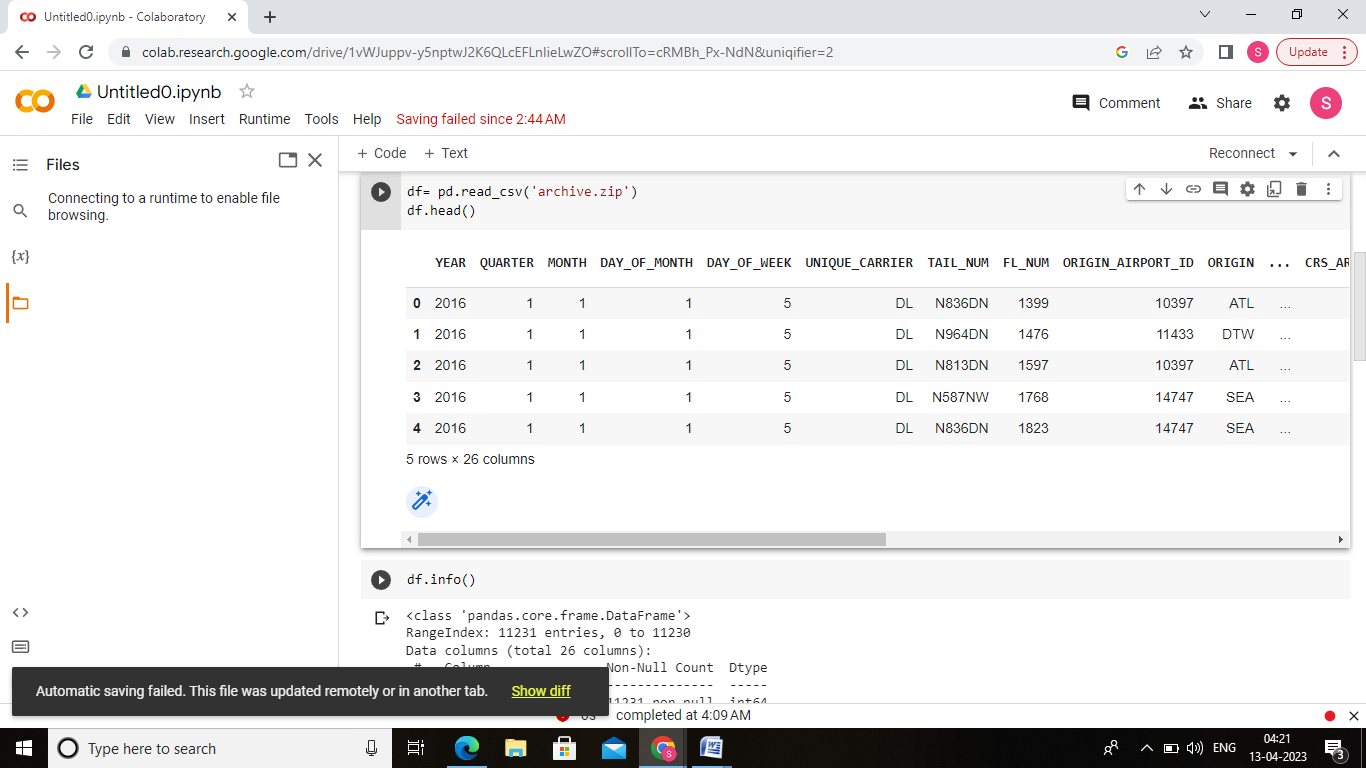
# data from international flights and not just restrict our self to the domestic flights

# APPENDIX

# Activity 1.1: Importing the libraries



**Activity 1.2: Read the dataset**



**Activity 2.1:Handling missing values**

