**Data Science Project Protocol -**

**Flight Delays**

*Author(s): Michael Shnayderman*

# Introduction

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights is published in DOT's monthly Air Travel Consumer Report and in this dataset of 2015 flight delays and cancellations.

Flight delays is a wide problem that all the airliners deal with. The delays occur due to a large variety of reasons and we will inspect most of them in this work. Flight delays are one of the most critical things to the airline companies, since they cause large financial losses.

In this work,I want to check the reasons for flight delays and predict flight delays in the United states. This will help airliners to reduce to a minimum the main causes of the delays.

In addition, we will analyze what affects the most on the delays - is it specific airports, specific routes, days of the week or etc.

We will check what is the company with the lowest amount of delays and what is the company that has the highest amount of delays.

# Methodology (Project design)

## The data for this project was published by the U.S Department of Transportation’s Bureau of Transportation statistics. All the information is downloaded from Kaggle and has all domestic flight data in the USA for the year 2015. Data includes more than 5 million records of domestic flights in the USA.

## Data

All the data was downloaded from Kaggle, no external data was added. Possible data source that could enrich my study is the weather forecast for the year of 2015 - to see if it has an effect on flight delays. I could not find a reliable source for this so I didn’t add anything.

Before processing the actual data, I took all separate files and made one big data frame that includes all data on each flight : Airline, Arrival and Departure time, IATA codes, origin and destination - country and airport and etc.

These are the final columns I have on my data set (dataset name is All\_Flights) :

['DESTINATION\_AIRPORT\_NAME', 'DESTINATION\_CITY', 'DESTINATION\_STATE',

'DESTINATION\_LAT', 'DESTINATION\_LONG', 'YEAR', 'MONTH', 'DAY',

'DAY\_OF\_WEEK', 'FLIGHT\_NUMBER', 'TAIL\_NUMBER', 'ORIGIN\_AIRPORT',

'DESTINATION\_AIRPORT', 'SCHEDULED\_DEPARTURE', 'DEPARTURE\_TIME',

'DEPARTURE\_DELAY', 'TAXI\_OUT', 'WHEELS\_OFF', 'SCHEDULED\_TIME',

'ELAPSED\_TIME', 'AIR\_TIME', 'DISTANCE', 'WHEELS\_ON', 'TAXI\_IN',

'SCHEDULED\_ARRIVAL', 'ARRIVAL\_TIME', 'ARRIVAL\_DELAY', 'DIVERTED',

'CANCELLED', 'CANCELLATION\_REASON', 'AIR\_SYSTEM\_DELAY',

'SECURITY\_DELAY', 'AIRLINE\_DELAY', 'LATE\_AIRCRAFT\_DELAY',

'WEATHER\_DELAY', 'IATA\_CODE', 'AIRLINE', 'ORIGIN\_AIRPORT\_NAME',

'ORIGIN\_CITY', 'ORIGIN\_STATE', 'COUNTRY', 'ORIGIN\_LAT', 'ORIGIN\_LONG']

My outcome variable will be departure delay time and I will try to predict it with few different techniques.

Regarding the missings - there were less than 10 percent missings in all the data so I dropped all the missings.

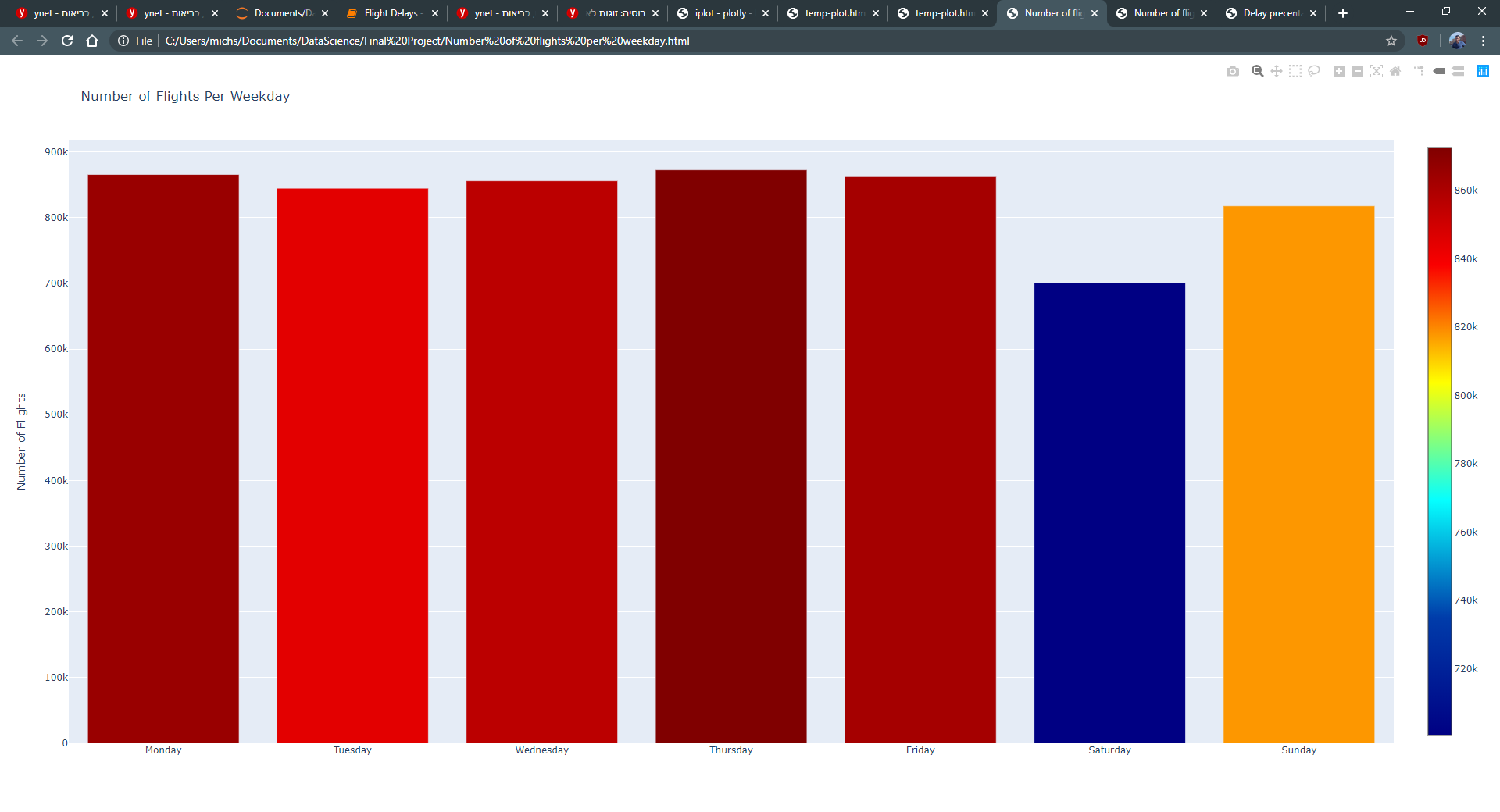
For exploring the data , I showed graphs that deployed delays by airline companies. In addition, I compared arrival delay with departure delay for each company - in which we can see the the arrival and departure delays are almost equal.

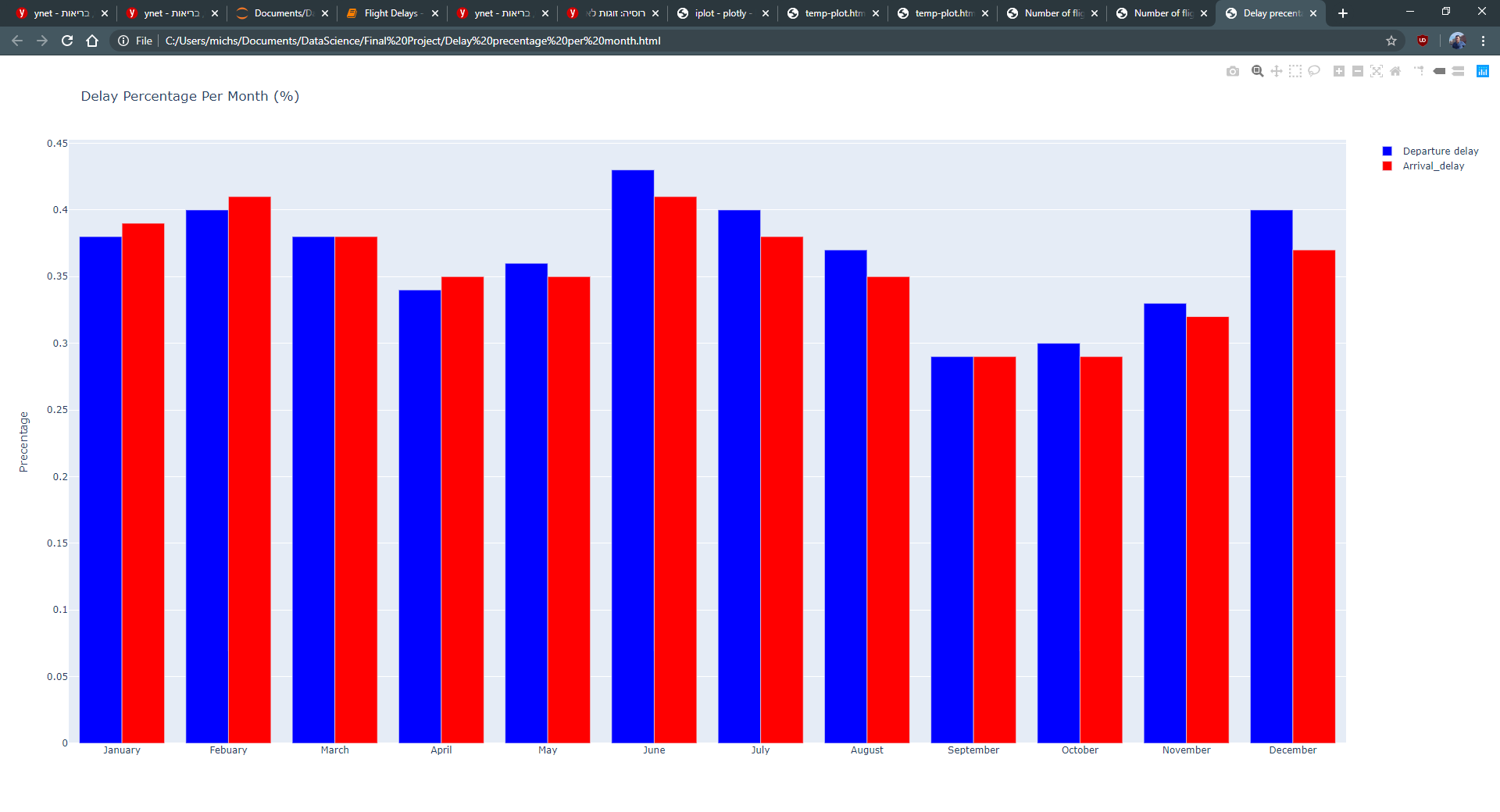
I showed the delay percentage (arrival and departure) for each month.

You will also find graphs that show that number of flights per each weekday and each month.

From the graphs above we can see that in June and July we get the highest percentage of flight delays as well as the highest number of flights.

As for the weekdays, the number of flights each day is almost equal except for Friday and Saturday (weekend). See too of those graphs below.





I also tried to find strong correlation between some of our variables to the delays. We can see that there is no strong correlation between states and days of the week regarding delays.

In the shown correlation matrix, we can see that there is a strong correlation between the arrival delay and the departure delay.

In addition, there is a correlation between the departure delay with the airline delay and the late aircraft delay, same for arrival delay.

## Models

In this work I want to predict flight delays (in minutes) using the gathered data.

Due to this statement, I will use regression techniques - In this case XGboost regressor.

I will train the model only with numeric variables. I will train the model twice - one with the variable arrival delay and the second without arrival delay.

As said before - The target variable for prediction is “arrival delay”.

First, I will prepare the data into train and test, the test size will be 15 percent.

Second, we have to train the model with x\_train and y\_train.

Third - we will “test” our model and predict for x\_test, comparing our result with y\_test.

When we are pleased with the results we can go further to the next stage.

The last stage is predicting an actual Y using our trained model. I didn’t have additional data for prediction so I don’t have this stage.

\*\* Normally I would try other models such as ADAboost or decision trees, but my data was too big (more than 5 Million rows) and it took hours to run each model.

## Deployment of your model

The model will be published and can be used by each airline that wants to reduce delays.

The final user is an airline company and not a customer so it can be given as a side program that can predict delays using specific parameters.

The analysts of each airline company will have to check the manipulations made on the data are useful for their company (For example - different states, routes and etc.

The airline company will subscript a responsible person for predicting delays and using the models. In addition he will be responsible for updating the model. I think that there is no need to update the model more than once in a year - the data is yearly periodic.

# Conclusion

First, The topic was very interesting but much more complicated than I thought.

There are so many variables that affect flight delays that it is very hard to add them all.

During the project I found it hard to add additional data from different sources. I am not sure why. Maybe I just don’t know the common sources, maybe the data I searched for is too old (2015).

It was interesting to see what variables reflect more on flight delays. I learned a lot of new definitions used by airline companies such as wheels on/off.

Also I learned that the airline companies use categories to classify different delays.

I think the model can work, but I think I could do a much better job using a much stronger computer to echeck some other models and find the one that fits the best.