Overview of the notebook(main.ipynb):

The focus of the notebook is analyzing flight delays for the year of 2015 and coming up with a model that can predict the delays.

The concepts covered are:

1. Data Cleaning
2. Data Exploration and Analysis
3. Data Visualization
4. Machine Learning:
    4.1 Prediction
    4.2 Classification

For the curious fellow python users, I have used pandas and numpy for data manipulation and matplotlib and seaborn for visualizations.

In the context of machine learning, I have tried to explore the ideas of randomizing the input data and cross-validation.

About the dataset:

The dataset that I have used is hosted on kaggle and is provided by the US transport department. It comprises of 3 different csv files: *airport, airlines and flights* containing information regarding domestic flights of USA in the year 2015. The dataset contains lot of features and hence, the possibilities of explorations are abundant! Which is why, I have chosen to perform some analysis on the delays. But again, there could have been many more explorations possible with such a huge dataset. For example, we are provided with airports.csv containing the cities, locations, etc and we also have flights.csv. From these, we can analyse which city has the most traffic,which has the least, which is a hub for tourism etc.

Note: flights.csv is an extremely large file. Hence, please use the below mentioned link to download the file:

https://www.dropbox.com/s/uce656ijxu8an66/flights.csv.zip?dl=0