Abstract

Flight delay

According to Federal Aviation Administration, the cost of flight delays in the US has been increasing for the past decade and it reached $ 33 billion in 2019. This project aims to conduct exploratory data analysis to analyze US airlines delay data set published by The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS) to have a high level understanding on the delay causes.

Design

This project is one of the T5-Batch 3 Data Science BootCamp requirements. Data provided by Kaggle and originally released by U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS) has been used in this project. The data sets contain up to 1.936.758 different internal flights in the US for 2008 and their causes for delays. Analyzing the causes at macro level will help in improving the delays.

Data

The dataset is provided in .csv format. It contains 1,936758 flight , each flight has 30 features. The most relevant feature to this project is the Flights Delay. It is obtained from DataExpo**. The date** has 29 features, as instance: (Year ,Month, Day of month, Day of week, , Carrier name ,Arrival Delay , Departure Delay ,Origin ,Destination ,Carrier Delay ,Weather Delay , NAS Delay, Security Delay ,Late Aircraft Delay, …).

Algorithms

- **Cleaning all data feature by removing all non-values, removing duplicated rows.**
- **Analyzing relationship between different data features.**
- **Filtering, aggregating delays to obtain overall delay time**
- **Conducting various data analysis on different causes to drive conclusions**

Tools

- Technologies: Python using Jupyter Notebook
- Libraries : Pandas , Matplotlib and Seaborn

Communication

The project code is available Here and the The slides are provided Here.