Question

Colder days, cheaper tickets?

I like to travel back to my home country in the beautiful days of summer, but my friends tell me this action is very expensive, and you can travel to my home country in other months of the year cheaper. So, in this project, I want to see if their claim is accurate or not.

In this project, I want to research and see the weather conditions and the temperature of the city influence the price of flights in the scope of North America. What are these changes, and are they increasing the prices or decreasing flight ticket prices?

Data Sources

The topic of this project needs two primary data. Flight information, prices and the weather conditions that day on the scope of the project.

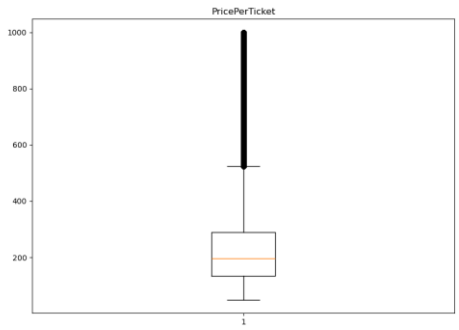
At the start of this project, I searched for a dataset for the flights in the scope of the project, which I found a dataset in the Kaggle. The **2018 Airplane Flights** dataset in the Kaggle with the **'CC0: Public Domain' License**. It was the fitted dataset that I found for this idea. The is from the real world and is correct. Every row of data is filled and has no missing and All columns remain the same format throughout the dataset. The dataset is from 2018, and it is an update dataset. The dataset does not have any anomalies I checked manually and the boxplot.

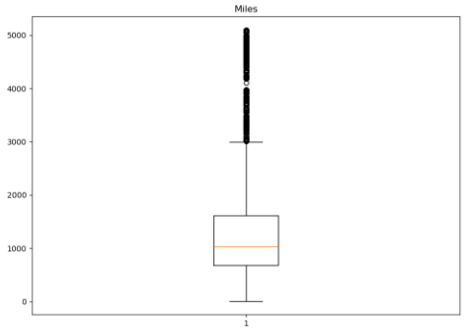
The dataset does not have the exact time of flight; however, it has the quarter, which means in which quarter of year the flight happened. It has two columns which I think are very related to my idea 'PricePerTicket', 'Miles' and 'Quarter'. The 'Quarter' is the most important column of the dataset which shows the time of 2018 the flight happened, and the value is 1,2,3,4. The 'Miles' is the distance between Origin and Distance in mile metric, however I don't want changed it to meter because the slope is different which I don't think it makes a difference in the result. The other is the weather conditions and temperature of North America in 2018. Initially, I wanted to use pirateweather.net, but I countered some problems which forced me to search again about the weather conditions and temperature.

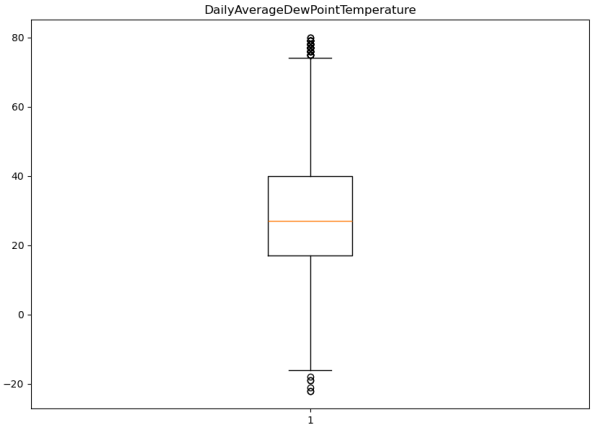
In the beginning, I had planned to use the pirateweather.net and its terms but countered a problem in the data pipeline step which forced me to switch to another dataset.

After the problems with pirateweather.net I searched for a new dataset which satisfies the required conditions of the idea which I found **Average day weather for 2018** in the Kaggle

with the **‘CC0: Public Domain’ License**. Every row of data is filled and has no missing and all-important columns remain the same format throughout the dataset. The dataset is from 2018, and it is an update dataset. The dataset does not have any anomalies I checked manually and the boxplot. The dataset has two important columns, ‘DailyAverageDewPointTemperature’ and ‘DATE’. The ‘DailyAverageDewPointTemperature’ is the value of temperature in Fahrenheit, and I am thinking is not needed to change it to other measurement because I want influence of changing temperature is same in the Fahrenheit however the slop is different which I think is not make a difference in the result.







Data Pipeline

The overview of final pipeline

My datasets are hosted on the Kaggle and clean, which made the task of downloading the datasets more comfortable.

The pipeline.sh file only runs to the pipeline.py file. Initially, I downloaded the datasets with the Kaggle client(kagglehub) which is officially introduced by the Kaggle site. Secondly, the data pipeline copies the files from the place which was saved by the Kaggle client to the data

folder on the project. The next step must be cleaning the data, but the datasets are cleaned, and I don't think I can apply more cleaning to them, so the transforms are the next step. It transforms the data; every dataset has its own transformation.

- **Average day weather for 2018**: the pipeline must calculate the average temperature in every quarter, So the mapping of 01-01 until 03-31 to quarter 1, 04-01 until 06-31 to quarter 2, 07-01 until 09-31 to quarter 3 and 10-01 until 12-31 to quarter 4, is used to filter every data point to belonged quarter. In the next step, it calculates the average temperature in a quarter. Finally, the pipeline saves the average temperature of a quart and start and end date of the quarter in a CSV file.
- **2018 Airplane Flights**: the pipeline filters and calculates 'PricePerTicket' divide by 'Miles' and calculates the average. Finally, it saves a CSV file.

The problem with pirateweather.net

The problem with pirateweather.net was I wanted to get an overview, all days of 2018, from 2018-01-01 until 2018-12-31, for the cities on the flights' dataset which was very time-consuming and expensive, because of very run of the data pipeline I should query 50 state weather condition for 365 days which became over 18,000 requests nevertheless, the pirateweather.net API has 10,000 per month requests free which is not enough for a full run of pipeline.

Result and Limitations

The output of data pipeline.

The output of data pipeline is a table with the data of every quarter average price per mile and average temperature per quarter which is save on the data\final.csv.

Why save the data into CSV format?

My data has four columns and four rows (not including the header) which is a small dataset. The CSV format is a text base storing format with easy to use so, it is a good choice to save the pipeline result on it.

Critically on the datasets

The '**2018 Airplane Flights'** could have better information about the flight time such as month, but I think this can cause some security issues for people.

Potential issues you anticipate for your final report: I don't see any potential issues right now.