Introduction

Colder days, cheaper tickets?

I like to travel back to my home country on beautiful days in summer, but my friends tell me this action is expensive, and you can travel to your home country in colder 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, and are these changes alike? For example, if I buy a ticket on freezing days in winter, I should pay less than buying the same ticket in summer.

Used Data

Data Source

The project has two data sources, flight information, prices and the weather conditions that day on the scope of the project.

- **2018 Airplane Flights**, The flight of 2018 in North America. Important columns are 'PricePerTicket', 'Miles' and 'Quarter'. The 'Quarter' shows the time of 2018 when the flight happened, and the value is 1,2,3,4. The 'Miles' is the distance between Origin and Distance in mile metric. 'PricePerTicket' is the price for a ticket which the user paid.
  - License: CC0: Public Domain.
  - No missing data and clean dataset.
  - Data type: CSV
  - Metadata link: link
  - Source: Kaggle
- **Average day weather for 2018**, the average day weather of 2018 in North America. Important columns, 'DailyAverageDewPointTemperature' and 'DATE'. The 'DailyAverageDewPointTemperature' is the value of temperature in Fahrenheit, 'DATE' is the date of recorded weather:
  - License: CC0: Public Domain.
  - No missing data and clean dataset
  - Data type: CSV
  - Metadata link: link
  - Source: Kaggle

The license of both datasets is CC0: Public Domain, which is a well-known data source licensed. The license is free as possible of any restriction and people can use the data with this license for their research.

Pipeline

The pipeline which is written in python 3 download datasets from Kaggle. The datasets are clean, which was very helpful. The 'DATE' of The **Average day weather for 2018** dataset maps to the 'Quarter' value of 2018 Airplane Flights dataset which the map range:

Start	End	Quarter
2018-01-01	2018-03-31	1
2018-04-01	2018-06-31	2
2018-07-01	2018-09-30	3
2018-10-01	2018-12-31	4

Moreover, calculate the average weather in that quarter and the average of Price of the tickets per a mile in the quarter. The output of the pipeline is a CSV file which each row contains 'AvgTemperature' which is the average temperature in float, 'PricePerMiles' which is the average of Price of the tickets per a mile in float and the quarter number. The measurement of the temperature and distance are in Fahrenheit and miles which does not convert it to other measurement system because they don't have influence on the analysis part. The CSV file selected to store the output of pipeline because it produces a few rows as output. At the end, remove unnecessary files.

Analysis

The pipeline produces 'AvgTemperature' and 'PricePerMiles' features for each quarter. The question is wanting to see the weather is changes are the same way of changing the price or contrast. This need wants to calculate the correlation between these two columns. One way is the Pearson correlation coefficient (PCC) which was developed by Karl Pearson in the 1880s. It measures linear correlation between two sets of data. The result of PCC is between -1 and 1. The 1 shows storage correlation between two sets of data and -1 shows the opposite correlation between two sets of data, the exact -1 and 1 shows unrealistic perfect correlation. Also, 0 means the sets of data do not have any correlation. The value of PCC is normalized because it is the ration of two the covariance of two variables and the product of their standard deviation.

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 . \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Conclusions

Result

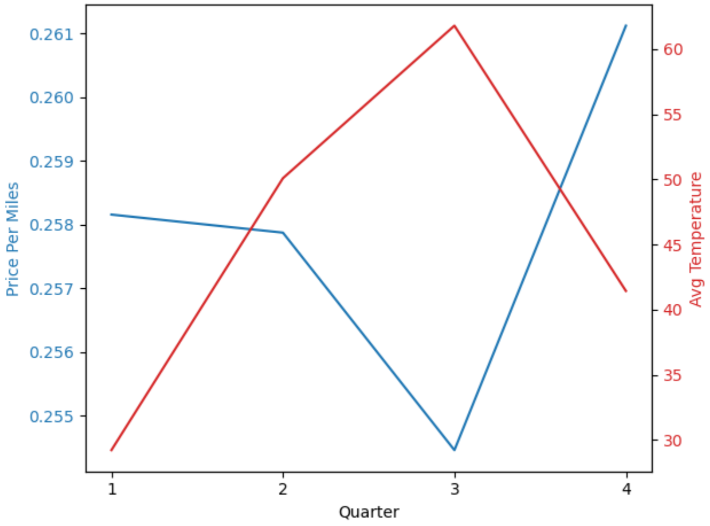
In previous part explains the analysis and how calculated PCC. The result of PCC is -0.65 which shows these two features' correlation are inverse which means with decrease of temperature, increase the price of ticket per mile. So, the answer to the posed question "**Colder days, cheaper tickets?"** is **No**, the price of tickets increase in the cold temperatures.

Some reasons for this phenomenon could be:

- Winter Holidays: the winter holiday which is the Christmas public holiday. Many people can travel to new places; however, the summer holiday does not have public holidays, and many people cannot travel freely.
- Tourism: In winter, many people travel from cold cities to warm cities such as Florida. However, people travel from warm cities to cold cities such as Colorado for winter sports such as skiing.

- Fighting in winter can face some weather challenges such as deicing, which cause the flight company to increase the price of tickets.

Also, this result is shown in the chart of two data.



The code of the result is in the file ` cal-ressult.ipynb`.

Limitation

As you read in the last part, the price can be influenced by public holidays, this project does not consider the public and school holidays. Moreover, the price changes in months but needs a more accurate flights dataset.

Future works

To remove the limitation of this result and be more accurate, there are suggestions to work on these topics:

- The result can be more accurate and shown in months of year, but the flight data source needs to have more accurate information about the flight time.
- Consider public holidays and school holidays as a feature which has influenced the price of tickets. A suggestion for these features is to be counted apart.
- The price of Operational Costs is an influence feature which this project does not investigate.