Do flight delays in the US follow a seasonal pattern across the year?

Abstract:

Our assignment investigates the well - established belief that flight delays (arrival and departure) in the United States demonstrates a predictable seasonal trend. In certain months, particularly during summer and holiday periods, delays slightly increase due to several factors, such as increased travel demand, weather conditions, and operational congestion. To explore this trend, we analyzed a large dataset of US domestic flights from 2020, with over 7 million records. Key attributes such as month, departure delay, and arrival delay were extracted from this dataset. The missing values in the dataset were addressed by imputing the median to ensure consistency across the full dataset.

With the aid of aggregated monthly averages, we produced a set of visualisations comparing departures and arrival delay trends all throughout the year. These visualisations reveal clear seasonal patterns, such as the highest number of delays were observed in the months June-August and again in December, while the least number of delays was observed in the early months of the year, January to March. These visual trends show that the flight performance is heavily influenced by seasonal factors, confirming the initial hypothesis.

Overall, this project demonstrates that U.S flight delays were not random but show strong month-to-month variation, caused by various factors, highlighting the importance of this data when analysing national aviation performance.

1. Datasets:

We used the "**US National Flight Data 2015 - 2020**" dataset from kaggle to perform analysis in this project.

Dataset link: https://www.kaggle.com/datasets/bingecode/us-national-flight-data-2015-2020

This dataset contains U.S. domestic flight records collected over six years, from 2015 to 2020, with a combined size of tens of millions of rows spread across multiple CSV files.

Although the full collection spans several years, we selected only the 2020 dataset for our analysis. This decision was made because 2020 is the most recent year, and provides a significant volume of data.

The 2020 dataset has 327 MB file size, and contains approximately 6.2 million rows and more than 20 columns, making it sufficiently large to demonstrate big-data characteristics while focusing on monthly analysis as well.

Within the 2020 subset, we extracted only the fields relevant to our research question and those attributes allowed us to compute monthly averages and investigate whether flight delays follow a seasonal pattern. The choice to focus on 2020 improves clarity, avoids

multi-year confounding effects, and ensures the results reflect recent operational conditions while still showcasing the large volume of flight-level data.

2. Data Exploration, Processing, Cleaning and Integration:

In our initial investigation, we examined how the data looks by sampling a few rows from the dataframe and obtaining summary statistics to look for null values/ outliers/ shape of data etc.

The raw dataset is very huge over 30 million records for multiple years, including more than 6 million in 2020 alone. Given the size of our data, rapid loading the whole file into memory at once caused overloading of our machines. To address this issue, we used only the 2020 file. Not only did we filter the dataset to a year, but also decided to limit it columns-wise to what was strictly needed for our analysis which helped in reducing Memory and pushed forward the process. Its popularity is comparable of **Chunk - Loading** or **Column filtering** used in real-world large-scale data processing and it was to make sure that the dataset can be processed smoothly by our system.

Data Cleaning and Processing:

During cleaning, we found that some flights had missing delay values, which is common in aviation data because of incomplete reporting, cancellations, or diversions. To deal with this, we used **median** imputation on both delay and arrival columns.

We selected median because it provided a stable replacement that isn't skewed by extremes outliers like:

- ➔ very long delays caused by storms
- ➔ mechanical issues
- ➔ ATC ground holds
- ➔ crew shortages
- ➔ COVID - 19 disruptions

After cleaning, we grouped the dataset by month to calculate average departure and average arrival delays for each month of 2020. This step converted millions of rows into a simple and meaningful monthly summary. Finally, we validated the results to ensure the imputation did not introduce unrealistic patterns. These cleaned, aggregated values formed the basis of the visualisations used to analyse seasonal trends in U.S. flight delays.

Next, we extracted only the three columns relevant to our visualisation:

1. month
2. arr_delay
3. dep_delay

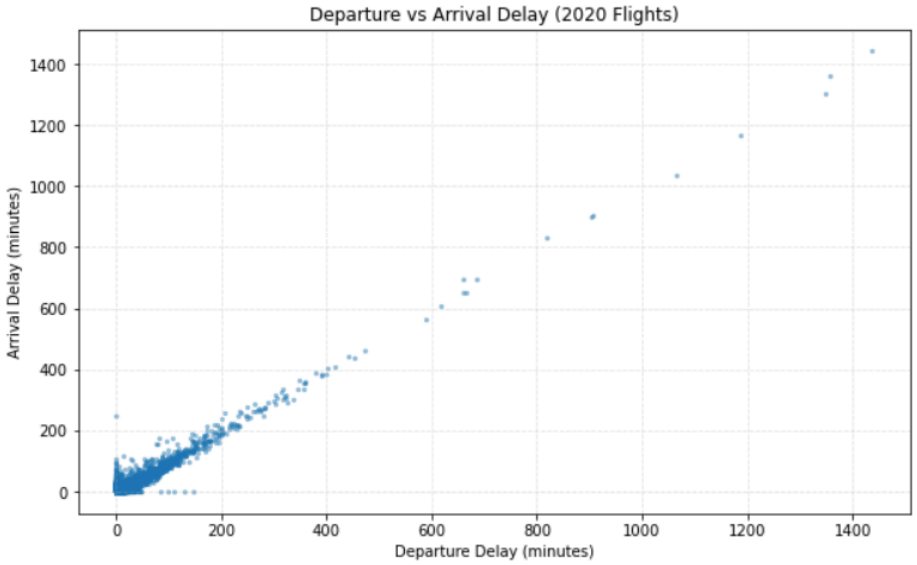
This helped reduce memory usage and made our further analysis more efficient. We then grouped the cleaned dataset by MONTH and calculated the average departure and average

arrival delay for each of the twelve months. This aggregation step transformed millions of individual flight rows into a manageable structure that could be used directly for visualisation.

Data Exploration:

During exploration, we compared delay trends month-by-month to confirm that the cleaned values made sense. As expected, early months like January and February showed lower average delays, while summer and holiday periods - particularly July and December showed noticeably higher delay values. This early insight supported our main hypothesis and helped guide the design of the final visualisation.

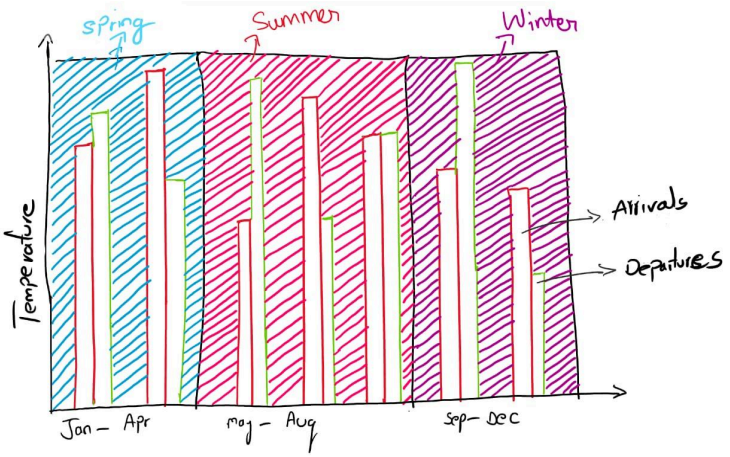
Before creating the final visualisation, we generated a scatter plot as part of our exploratory data analysis (EDA). The aim of this graph was to check whether departure delays had a direct impact on arrival delays and to see if the dataset contained unusual values. The plot showed a clear upward trend, meaning that flights with long departure delays almost always arrived late as well. We also noticed several extreme outliers, such as delays over 1,000 minutes, which confirmed that using the median for imputation was the right choice. This scatter plot helped us understand the overall behaviour of delays and reassured us that the two delay variables moved together consistently. Although the scatter plot was not suitable as a final visualisation, it played an important role in confirming the structure of the data and guiding us toward a month-based comparison chart for answering our seasonal delay question.



3. Visualisation:

Sketching the Idea:

We made a simple sketch to plan the structure of our visualisation: a bar graph showing seasonal trend along with arrival and departure delay



The idea was to create a graph that:

- Shows the two delay types clearly (dep vs arr),
- Compares them month by month, and
- Highlight seasonal changes across the year.

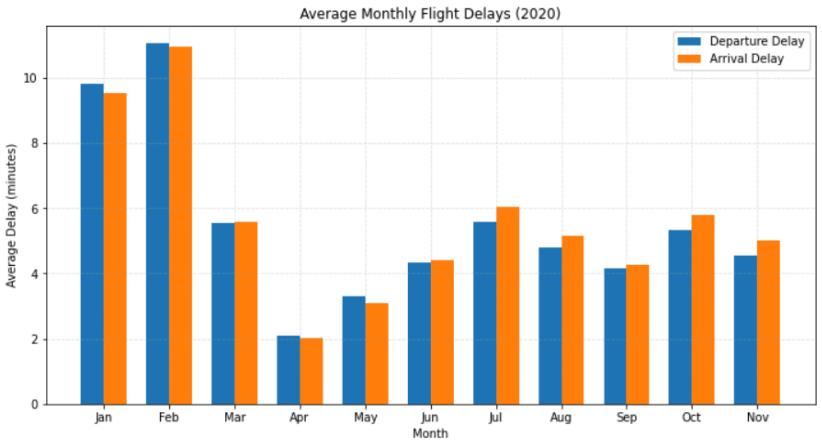
To reach this final design, we created a few trial graphs to understand the data and figure out which visualisation communicated our findings most effectively.

Graph 1 - Vertical Bar Chart:

We created a vertical grouped bar chart like our sketch, and it gave us a clear comparison, showed exact differences, but it was

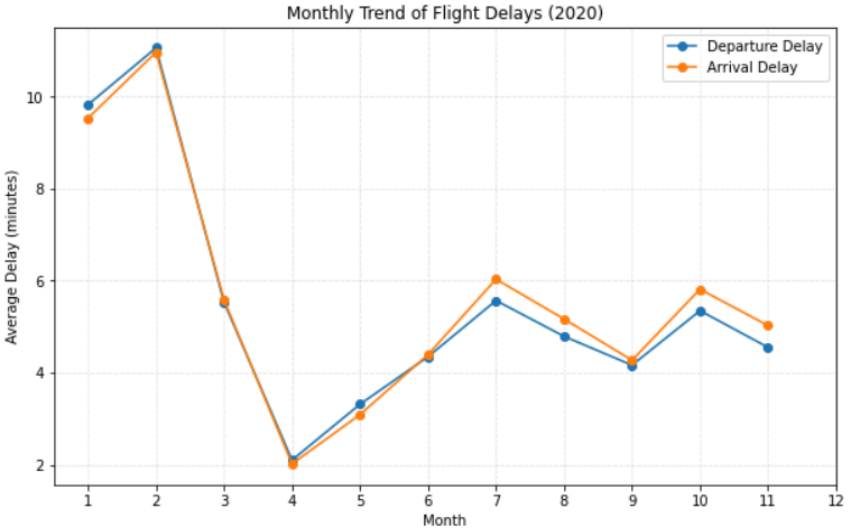
- Hard to read when months run left-to-right
- Labels on the x-axis look cluttered
- Bars look squeezed together
- Seasons were not clearly visible

This chart was better, but still didn't feel close to our goal, as our goal was to show seasonality clearly and visually. The vertical orientation was not the best layout.

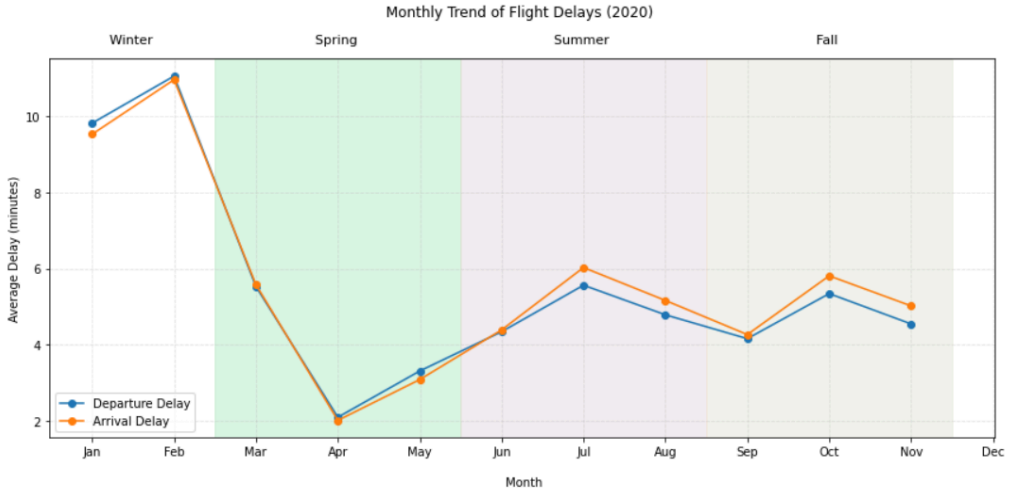


Graph 2 - Line Chart

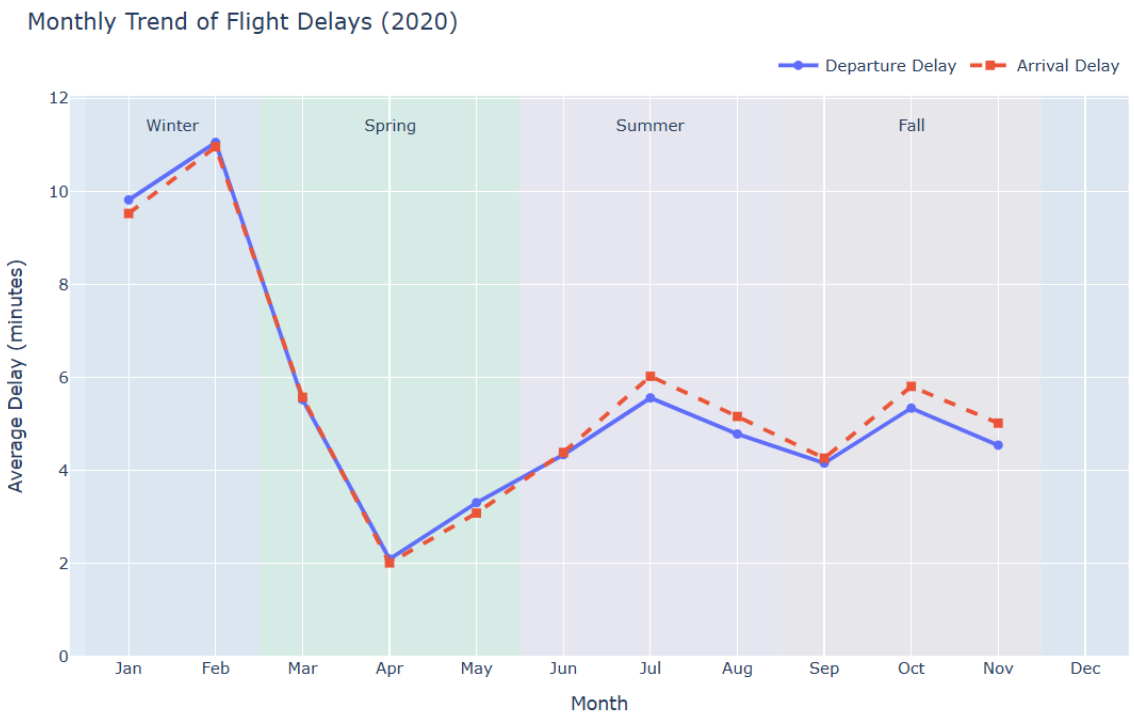
“How do average flight delays vary across the months?”



- Since the vertical bar chart did not give us exact visualisation, we tried a basic **line chart** to explore how average departure and arrival delays change month-to-month.
- This helped us quickly see that **both delay types move together**, suggesting a consistent seasonal trend.
- From the line chart we noticed an early pattern: **higher delays in winter/summer months** and a clear dip around April.
- We then added **seasonal background bands (Winter, Spring, Summer, Fall)** to make the trend easier to read in real-world terms, not just month numbers.
- We also kept the two lines in contrasting colours with markers, so the overlap and gaps between departure vs arrival delays are easy to spot at a glance in every season.



Our final graph:



Finally, we converted this chart into an **interactive Plotly horizontal grouped line graph**. By hovering over any point, the exact departure and arrival delay values appear, which makes month-to-month comparison much clearer.

The **seasonal background bands** remain visible in the interactive view, so the viewer can directly link each rise or dip to Winter, Spring, Summer, or Fall.

We also made the final chart **colour-blind friendly** by using two clearly contrasting line colours plus distinct markers and a legend, so the two delay types are still easy to tell apart even if someone can't rely on colour alone.

Overall, this interactive, accessible version communicates the seasonal pattern more clearly than our earlier static graphs while still letting the audience read precise values when needed.

4. Conclusion:

Tools and libraries used:

We used Jupyter Notebook to explore the dataset and filter out only the columns needed for our analysis. Because the original dataset was extremely large, we relied heavily on pandas for reading, cleaning, and grouping the data. Google Colab was then used so both group members could work together on the cleaning process and testing different visualisations. Finally, Plotly was used to create our interactive horizontal bar chart, which became our final visualisation.

Analysis of the outcome of the visualisation:

One of the biggest challenges we faced was deciding which type of chart best communicated the seasonal pattern in flight delays. We tried multiple options, including scatter plots, line graphs, and vertical bar charts, but each of them had limitations; either they were too cluttered, misleading, or didn't highlight month-to-month differences clearly. After testing all these possibilities, the horizontal grouped bar chart in Plotly proved to be the clearest and easiest to understand.

What could be improved?

A limitation of our dataset is that December had missing or incomplete records, which affected the monthly average. Ideally, analysing several years together (instead of only 2020) would give a more reliable long-term pattern. Another improvement could be adding airline specific or state-level insights, but this was outside the domain of our chosen research question.

What does the chart tell us?

Our final visualisation clearly shows that flight delays do follow seasonal trends. We observed higher delay averages in January, February, July, and October, and much lower averages in April and May. Departure and arrival delays also moved together almost identically, meaning the same seasonal factors affect both sides of flight operations.

Collaboration:

The work was shared fairly evenly. One member handled the initial data download, filtering, and the memory-efficient loading of the dataset. Both members collaborated on cleaning and exploratory charts in Google Colab. The final Plotly graph and colour styling were decided together after reviewing sketches and earlier draft visuals.

References:

Dataset Source:

- Bingecode (2021). *US National Flight Data 2015-2020*. Kaggle. https://www.kaggle.com/datasets/bingecode/us-national-flight-data-2015-2020

Python Libraries & Documentation:

- pandas documentation: https://pandas.pydata.org/
- Matplotlib documentation: https://matplotlib.org/
- Plotly Express documentation: https://plotly.com/python/plotly-express/

Additional Resources (used for guidance on visualisation concepts):

- Evergreen Data - "Choosing the Right Chart": https://stephanieevergreen.com/
- Data-to-Viz: https://www.data-to-viz.com/
- Kaggle community discussions on handling missing values and outliers.