# GROUP 6 PROJECT 1 REPORT

# CHANGES IN AIR TRAFFIC PATTERNS DURING COVID-19 AND AFTER

Ava Lee, Silas Phillips, Jed Murphy, Christoph Guenther

April 22, 2024

## INTRODUCTION

As the Covid-19 virus spread globally in early 2020, the routines of daily life changed drastically for many people around the world. As the U.S. went into lockdown to slow the spread of the virus, travel, business as well as leisure, changed drastically. It is well known that the U.S. domestic airline industry, among many others, experienced an unprecedented crisis during that time [1].

Using United States Department of Transportation commercial air travel data, we compare U.S. air passenger traffic patterns before, during, and after the Covid-19 pandemic. The purpose is to obtain a more detailed and quantitative understanding of how air travel patterns look different during the pandemic from before and after. We also aim to understand how air traffic patterns might still be different after Covid-19 as compared to before Covid-19.

## CHALLENGES AND PROBLEMS:

1. Storage issues with GitHub.
   After many hours spent trying to fix what appeared to be corrupted datafiles, we found out that GitHub has a 1GB storage limitation. The size of our datafiles is about 20GB in total far exceeding GitHub's limit. All the tech help resources available to us could not provide a solution to the issue. We were finally able to resolve it by buying more storage.
2. Issues downloading large amount of data
   Due to their size, some of us were unable to download the data files to their local machines. We tried to resolve the issue by using a virtual CoLab machine. However, processing the large data files required almost 6GB of RAM too much for an instance of a free CoLab machine.
3. Not all airline datasets are available in sufficient granularity unless we wanted to pay for the data. Therefore, airfare information is only available by quarter. All other information is available at least by month.
4. Pandas was unable to read some of our data files. They had to be converted to ".csv" format manually.
5. One of the files in the flight data set could not be read as is. However, we were able to download the same information in a different file. That file had to be manually pre-processed before it could be read in automatically.
6. On Windows PC, not all values of the correlation heatmaps were displaying. We had to downgrade the "Matplotlib" library to version 3.7.3 for it to work.

## GOALS AND QUESTIONS

[1] Leslie Josephs (August 16, 2020). *A flood of job losses looms as airline industry struggles in pandemic*, CNBC website, https://www.cnbc.com/2020/08/16/a-flood-of-job-losses-looms-as-airlines-industry-struggle-in-coronavirus-pandemic.html as accessed on 4/22/2024.

Our goal is to illustrate and quantify some of the changes in commercial U.S. air passenger traffic before, during, and after the Covid-19 pandemic. We aim to answer the following questions:

1. How did domestic U.S. air passenger traffic change during Covid-19?=
2. How long did it take for passenger volume to normalize to pre-pandemic levels after Covid 19?
3. How did the average airfare change during Covid?
    a. If it changed, when did it change as compared to when passenger volume changed?
    b. If it changed, when did it normalize as compared to when passenger volume normalized?
4. Can we discern any changes between air passenger traffic before and after Covid-19?
   What were the impacts on?
    a. Airfare,
    b. Arrival delays,
    c. Passengers per flight.

## DATA COLLECTION AND SOURCES:

All of our data was obtained from the Bureau of Transportation Statistics (BTS) of the United States Department of Transportation, BTS. We used three different BTS data sources:

1. Flight data
   From https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGK&QO_fu146_anzr=b0-gvzr as accessed between 4/8/2024 and 4/17/2024.
   We downloaded a separate pre-zipped ".csv" file for each month between January 2018 and December 2023. Each file contains a list of all US domestic commercial flights in that month. We used these files to retrieve the number of flights, the flight delays, and the number of cancelled flights.
   We were able to follow this process for all files in this data set except one which had a format that could not be read by Pandas. We overcame this issue by downloading the data we needed not through a pre-zipped file. However, this required some manual processing before this file could be read with all the other files in this data set.
2. Passenger Volume
   From https://www.transtats.bts.gov/Data_Elements.aspx?Data=2 as accessed between 4/8/2024 and 4/17/2024.
   There is a separate ".csv" file for each of the major U.S. airports. Each file contains the number of passengers that enplaned from that airport each month starting in October of 2002 until December of 2023.
3. Average Airfares
   From https://www.transtats.bts.gov/AverageFare/ as accessed between 4/8/2024 and 4/17/2024.
   A file for each quarter from Q1 of 2018 to Q3 of 2023 was downloaded separately. These files were only available in a "pseudo" Excel format, so Pandas was unable to read them. This required us to convert the files one by one by opening them in the Microsoft Excel application and saving them as ".csv" files.

## METHODS AND TECHNIQUES USED FOR EXPLORATORY DATA ANALYSIS.

### STEP 1: DATA IMPORT AND VALIDATION

As mentioned in the "Data Collection and Sources" section above, the data is grouped into three different data sets.

- Flight Data
- Passenger Volume Data
- Airfare Data

After converting the Airfare data set files from "pseudo' Excel to ".csv" format manually, all data sets can be read and converted to Pandas DataFrames using Pandas' 'pd.read_csv()' function.

We process each data set separately by

1. Using the Python 'glob' library to retrieve a list of files,
2. Using a 'for' loop to read each file into a Pandas DataFrame,
3. Reducing or adding to the rows and columns of the DataFrame to extract only the data we are interested in,
4. Removing duplicate rows,
5. Concatenating all DataFrames pertaining to the same data set into one DataFrame.

At the end of this process, we end up with three DataFrames, each one corresponding to one of the data sets for flights, passenger volume and airfare.

For each of the three DataFrames, we handle invalid values ('NaN' values) appropriately (see code comments for details).

We make sure that the number of rows for each column is the same in each DataFrame (but different from a DataFrame corresponding to a different data set).

Finally, we convert some data to the appropriate data types.

### STEP 2: BASIC CONSISTENCY CHECK

Our first consistency check consists of summing the flight data by month, year, and destination airport and ensuring that the sum of sums is equal to the total number of rows in the unaggregated flight DataFrame.

The second consistency check consists of making sure that the number of rows in our aggregated dataset is what we expect and equal to the number of rows in the passengers DataFrame.

## STEP 3: DATA PREPARATION

We process the data from the separate DataFrames into a summary DataFrame that is used for subsequent data exploration and visualization. Please refer to the code for details on this process.

The summary DataFrame has the following columns:

- Year
- Quarter
- Month
- Year-Month (YYYY-MM)
- (Destination) Airport
- City (of the airport)
- State (of the airport)
- No. of Flights
- Passengers
- Average Delay (Minutes)
- Cancelled Flights
- Cancelled %
- Average Fare ($)
- Inflation Adjusted Average Fare ($)
- Passengers per Flight

Since the airfare data is only available by quarter but all our other data is either available by month or aggregated by month, we converted the airfare data into "monthly" data by taking the quarterly values and assigning them to each month in the quarter.

We added a "Year-Month" column to make it very easy to slice the data into different time periods.

Although, we do not use the "Airport", "City", and "State" columns here, we could use them in further analysis to determine whether air traffic patterns vary across airports, cities, or states (for example because of different travel restriction in different locales).

## STEP 4: DATA EXPLORATION AND VISUALIZATION

Our data exploration led us to define the following time periods with corresponding Frames to answer the questions posed in "Goals and Questions" section above.

1. Before Covid-19 (pre-pandemic): January 2018 – February 2020
2. During Covid-19: March 2020 – July 2021
3. After Covid-19 (post-pandemic): August 2021 – December 2023

As we will explain in more detail below, we used the passenger volume to define these time periods.

To answer our questions, we created plots of the following metrics by month for each time period.

- No. of Flights
- Passenger Volume
- Average Delay (Minutes)
- Cancelled Flights
- Cancelled %
- Average Fare ($)
- Inflation Adjusted Average Fare ($)
- Passengers per Flight

We then selected the most appropriate plots to answer and substantiate the answers to our questions.

We used the Pandas 'describe' function to understand how basic statistical measures were different across the different time periods.

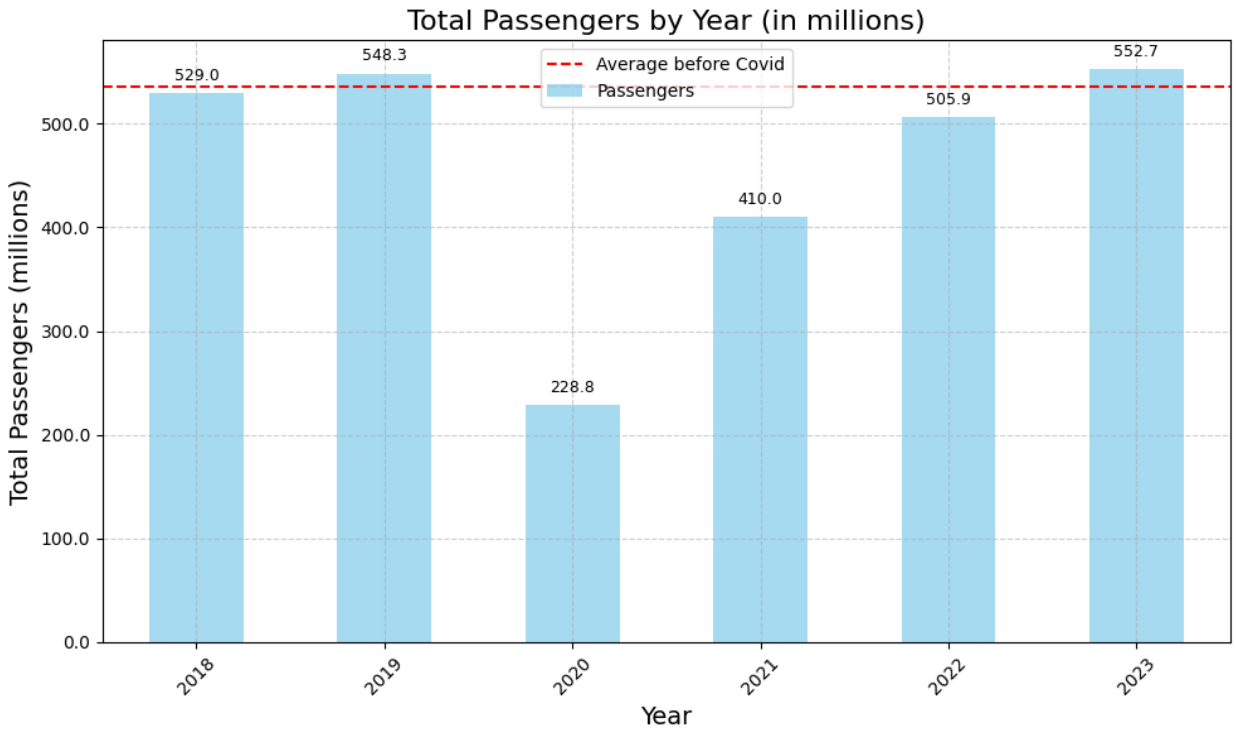
Finally, we used the Pandas 'corr' function to understand whether correlations between the different metrics listed above had shifted.

## RESULTS:

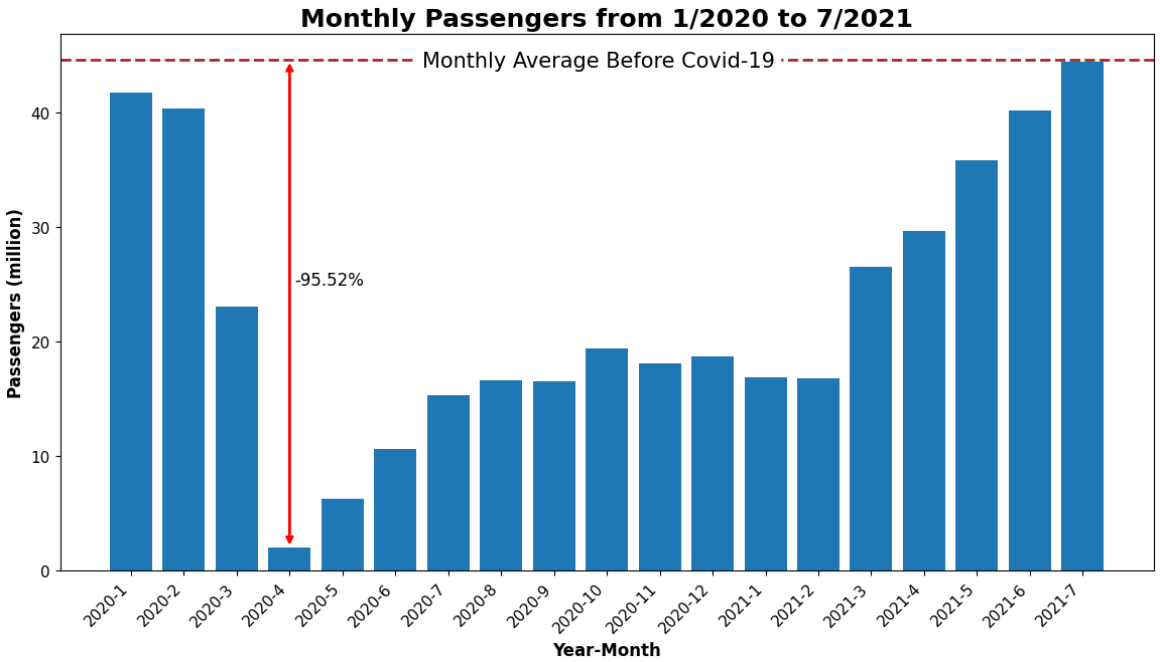
We will address each question in turn.

### QUESTION 1: HOW DID U.S. AIR PASSENGER TRAFFIC CHANGE DURING COVID-19?

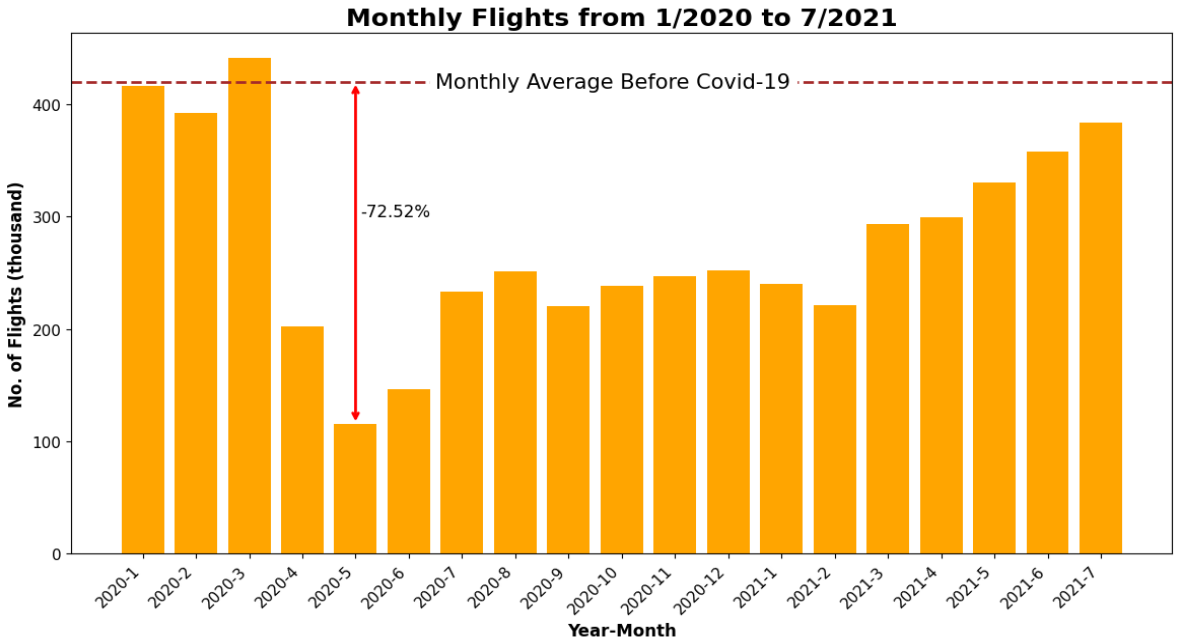
One of the most striking impacts of the COVID-19 pandemic on the U.S. domestic airline industry was the precipitous drop in passenger volume. The following plot shows the year over year decline in passenger volume.




Zooming in on the Covid-19 period, we see that passenger volume in April 2020 was 95.52% lower than the monthly average before Covid-19.



This is consistent with the drop in the number of flights as shown in the following plot.

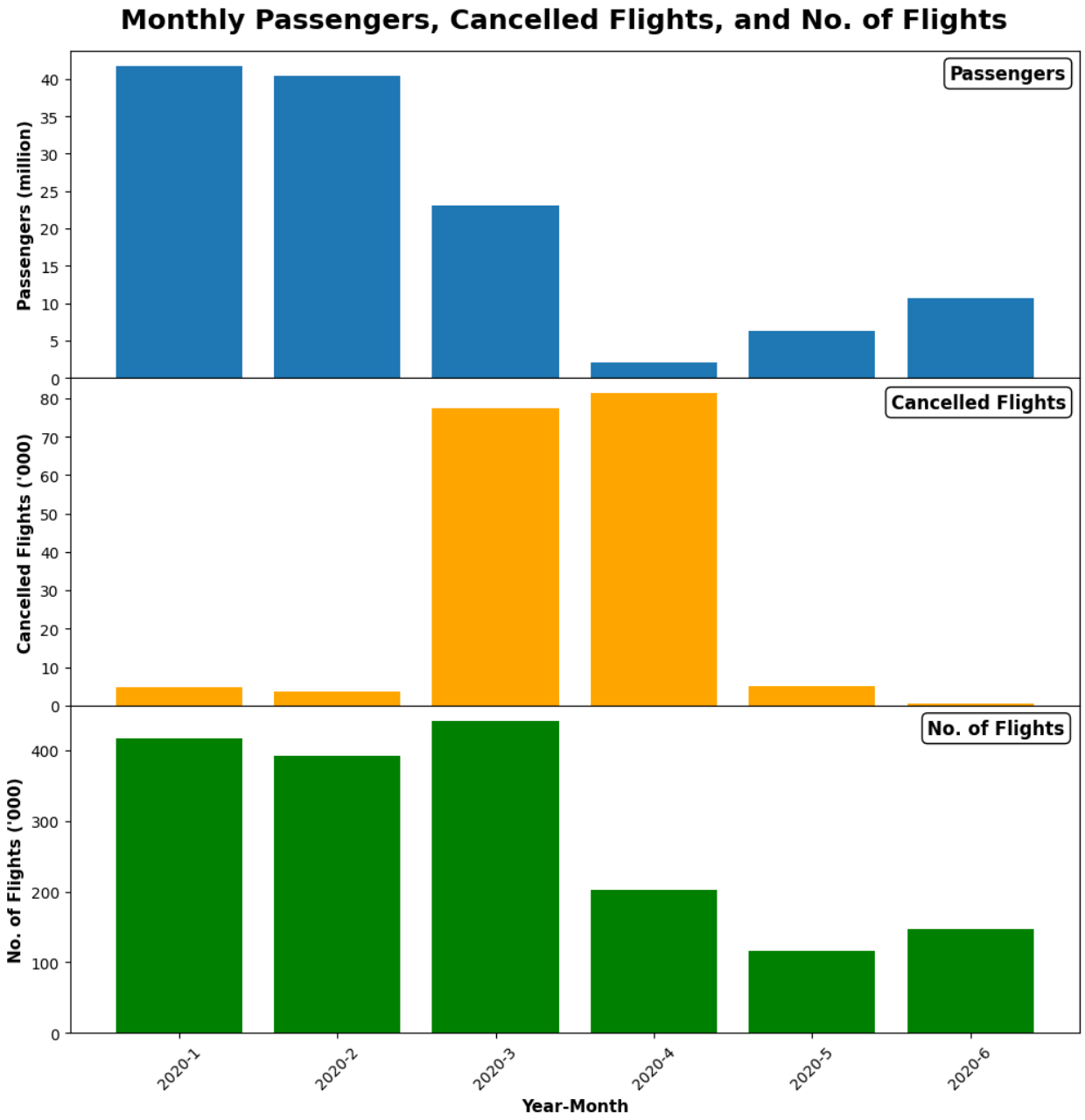



As the plot shows, the number of flights dropped by 75.52% as compared to its average before Covid. This decline is not quite as steep as the decline in passenger volume, and it would be interesting to understand the possible reasons for that. Unfortunately, we were not afforded more time to follow up on that question.

Note that the number of flights had not reached its pre-pandemic average in July 2021 when passenger volume had reached its pre-pandemic average.

Finally note that the minimum of the number of flights occurred one month after the passenger volume reached its minimum. We will do a little more detailed analysis later to see what the reason for this might be.

Comparing plots of the passenger volume, the number of cancelled flights, and the number of flights, se the next plot, what appears to be happening is that as the passenger volume dropped, the airlines cancelled many flights until they were able to adjust the number of flights to the new normal.




## QUESTION 2: HOW LONG DID IT TAKE FOR PASSENGER VOLUME TO NORMALIZE TO PRE-PANDEMIC LEVELS AFTER COVID-19

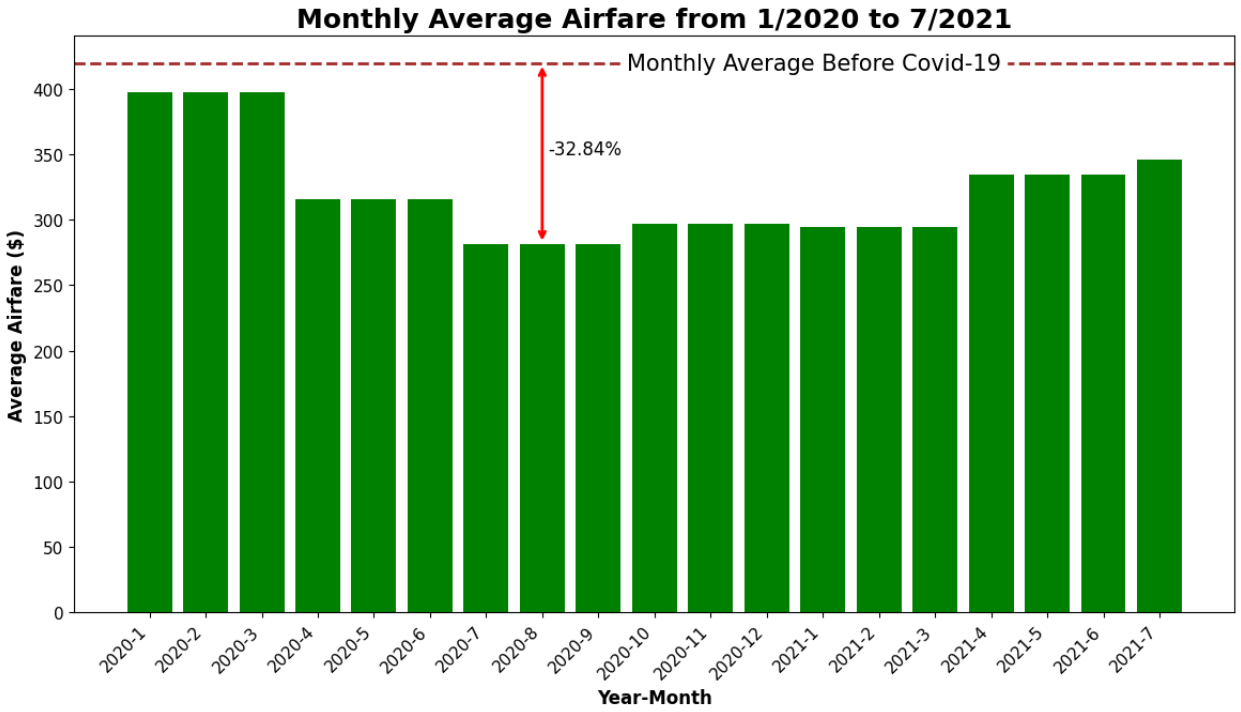
As is apparent from the "Monthly Passengers from 1/2020 to 7/2021" plot, passenger volume gradually increased from its April 2020 low, but it took until July of 2021 for it to reach its average before Covid-19.

We used this plot to define the Covid-19 period as starting in March 2020, the first month with a sizeable decline in passenger traffic until July 2021 when passenger volume had recovered to the monthly average before Covid-19.

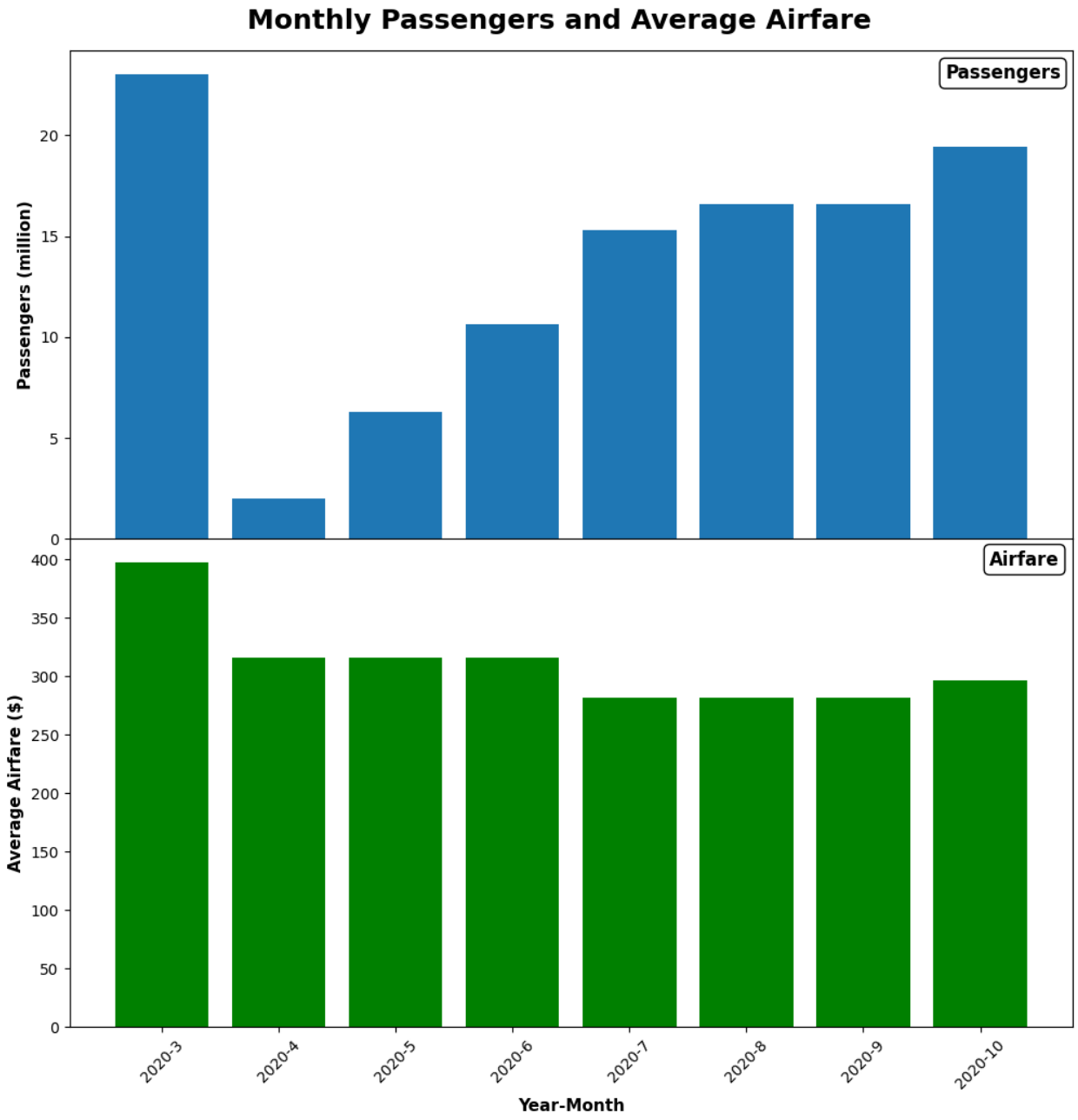
A more thorough analysis of when passenger volume recovered should take seasonality into account. Unfortunately, we did not have time to address this here.

## QUESTION 3: HOW DID THE AVERAGE AIRFARE CHANGE DURING COVID?

The following plot shows the change in airfare (inflation adjusted to quarter 3 of 2023) during the Covid-19 period. As is evident, it declined by 32.84% as compared to its average before Covid-19.




Note that the airfare reached its minimum about three months after the passenger volume reached its lowest point. This can be seen more clearly in this plot.




Unfortunately, we did not have monthly data for the airfare, making it impossible to pinpoint the exact month and percentage decline of the airfare.

One possible explanation for the delay between when the passenger volume reached its minimum and when the airfare reached its minimum might be that it took the airlines some time to determine the right pricing level given the new circumstances.

Also note that the average airfare did not drop nearly as much as the passenger volume or the number of flights. Considering the balance between demand and supply, this seems odd. Further investigation would need to be done to find possible reasons for that. Unfortunately, we do not have the time to address this question.

A final observation is that average airfare was still considerably lower than its pre-pandemic average at the end of the Covid-19 period in July 2021.

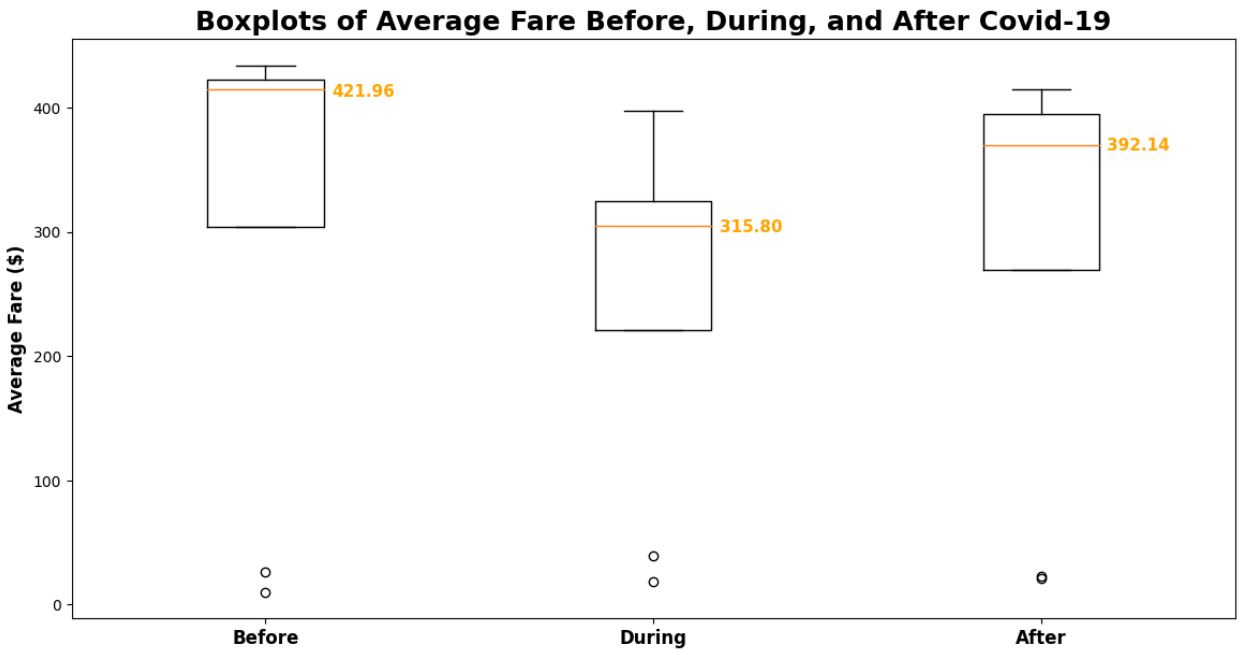
## QUESTION 4: CAN WE DISCERN ANY CHANGES BETWEEN AIR PASSENGER TRAFFIC BEFORE AND AFTER COVID-19?

What were the impacts on?

a. Airfare,
b. Arrival delays,
c. Passengers per flight.

We created the following boxplots to answer this question. Each boxplots depicts statistics for the periods before, during, and after Covid-19 for the airfare, delay, and passengers per flight metrics respectively.

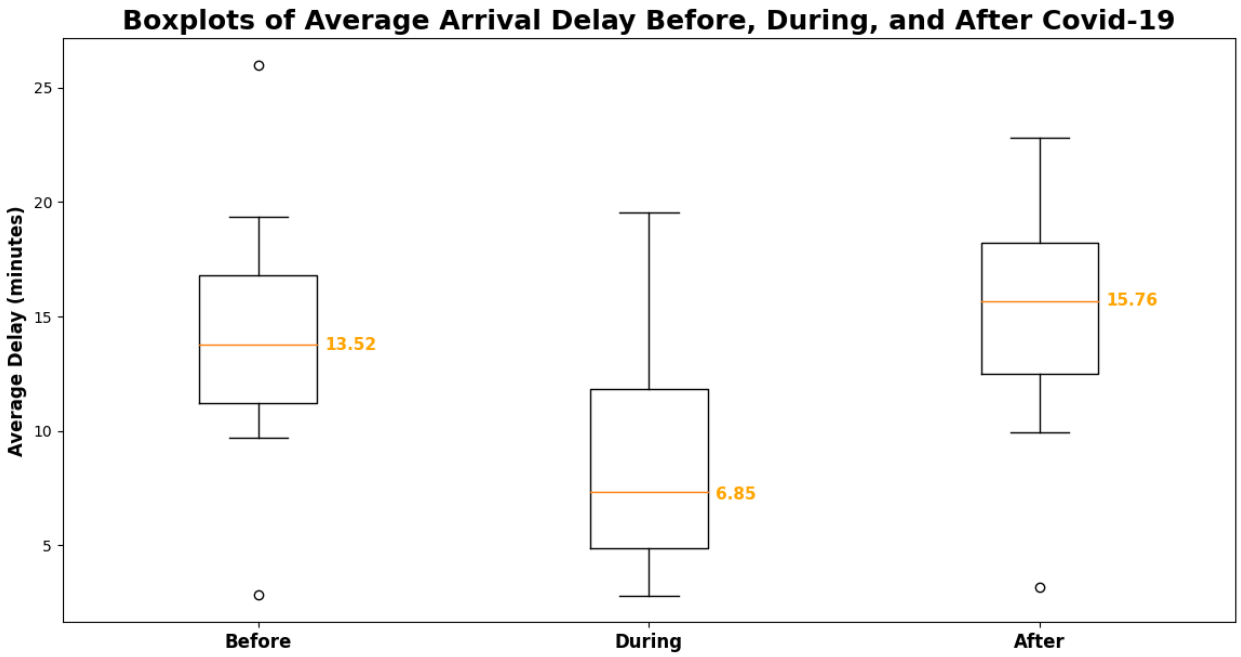
### AIRFARE



As can be seen from this plot the median airfare (inflation adjusted to quarter 3 2023 levels) for the period after Covid (defined from August of 2021 to September of 2023) is still 7% lower than the median before Covid-19 with an expected drop during Covid-19.

One might also notice that the percentage drop of the median values is smaller than the drop between the mean before Covid and the actual minimum during Covid. These differences can be attributed to the different statistical metrics used.

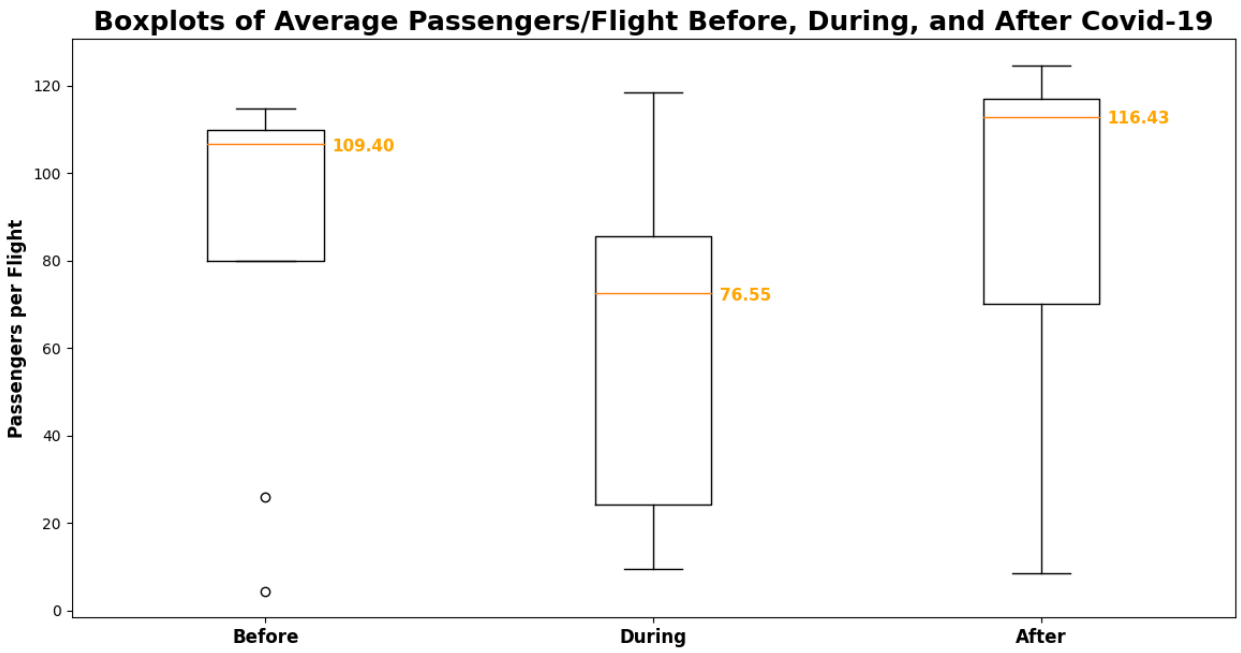
## ARRIVAL DELAYS




The median arrival delay dropped considerably during Covid-19 but as of the end of 2023 was over two minutes larger than its pre-pandemic level.

The reason for the drop in arrival delays during Covid-19 might be due to the reduced passenger volume allowing airline personnel more time to make sure flights were on time.

## PASSENGERS PER FLIGHT



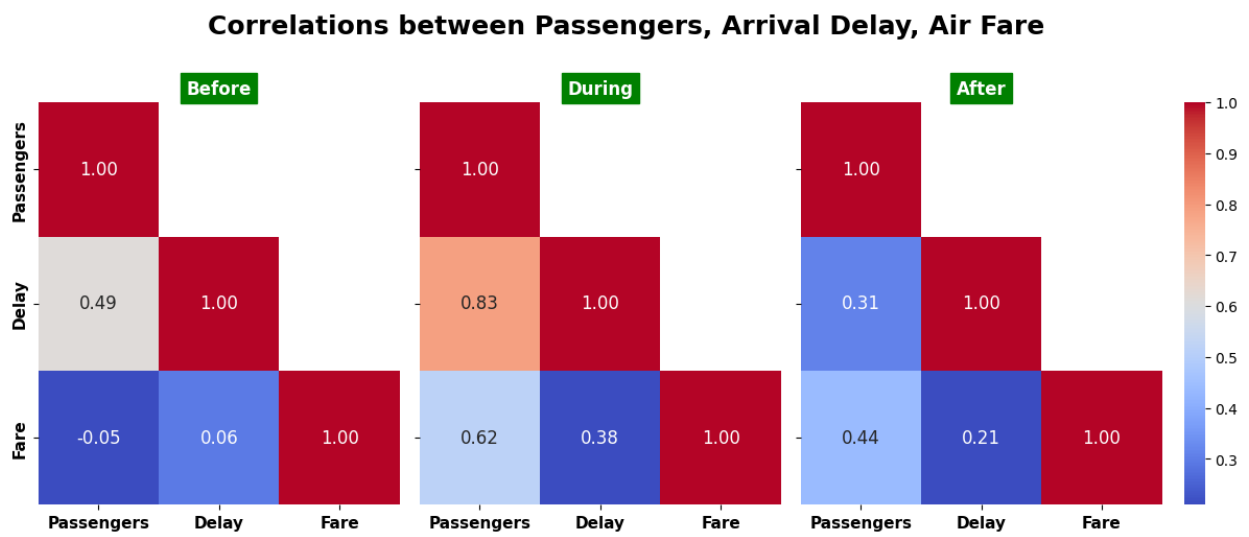

As expected, the median for the number of passengers per flight dropped during Covid-19.

The median for the number of passengers per flight after Covid-19 is considerably higher than before Covid-19. That makes sense since passenger volume seems to have recovered (close to) pre-pandemic levels, however, the number of flights is still lower than it was pre-pandemic.

This might also explain the increase in arrival delays since it might take more time to process more passengers per flight increasing the likelihood of delays.

## OTHER RESULTS

We also compared correlations between passenger volume, airfare, and arrival delays for the periods before Covid-19, during Covid-19, and after Covid-19, see the following plot.




One of the findings is that before Covid, there was practically no correlation between passenger volume and airfare. That changed to a moderately strong positive correlation during Covid-19. This might make sense because as passenger volumes gradually increased to pre-pandemic levels, the airlines were able to raise their prices as well. Note, however, that this moderately strong positive correlation has persisted after Covid-19, although it is less strong. Further investigation is needed to understand this observation.

A further observation is that the positive correlation between passenger volume and delay increased considerably during Covid-19, although the size of the average delay decreased considerably. Again, further analysis is necessary to better understand this observation.

## SUMMARY AND CONCLUSIONS

We have seen that air traffic patterns changed significantly during Covid-19. At its lowest points Passenger volume was down by 95.5% as compared to its average before Covid-19, the number of flights decreased by 75.5% from its pre-pandemic average, and airfare decreased by 32.8%. We have seen that the number of cancelled flights increased significantly in the first months of Covid-19.

Our analysis further shows that the minimum in passenger volume and the number of flights was reached in the first few months of the Covid-19 pandemic. It also revealed that there is a lag between when passenger volume reached its minimum and when the number of flights was at its lowest point. At the same time, the number of cancelled flights increased significantly in the first months of Covid-19.

We concluded that in the first few months of the pandemic, the airline industry reacted by cancelling flights before they reduced the number of flights.

We also observed that airfares reached its minimum about three months after the passenger volume reached its minimum and that the decline in airfares was much less severe than the decline in passenger volume or the number of flights. Further investigation is needed to better understand this pattern.

Recovering from the pandemic, we saw that passenger volume had recovered to its average before Covid-19 by July of 2021. However, other metrics, notably the number of flights and the airfare, were still depressed as compared to their pre-pandemic averages. In fact, the median airfare has still not reached the same level it had before Covid-19.

We also saw that the average arrival delay after the pandemic has increased compared to before Covid-19. This was accompanied by an increase in the number of passengers per flight after Covid-19. This makes sense since passenger volume is close to its pre-pandemic levels, but the number of flights is still smaller than pre-pandemic levels, which should lead to an increase in the number of passengers per flight.

The increased number of passengers per flight might also explain the observed increase in delay since more passengers per airplane might cause delays because it might take more time to process these passengers.

In addition, the airlines might have reduced staff during the pandemic and are still short on personnel. Therefore, fewer personnel must process a similar number of passengers, leading to delays. Further investigation is needed to confirm the correlations supporting this hypothesis.

Finally, we observed that some of the patterns and metrics have still not recovered to their patterns and levels before Covid-19. As mentioned above, the median airfare is still lower, arrival delays are higher, the number of passengers per flight is higher, and the correlation between passenger volume and airfare is still moderately strong. Will these lingering effects prevail? Further investigation and the passage of time is needed to answer this question.

## TOPICS FOR FURTHER INVESTIGATION

1. What explains the difference in decline between passenger volumes, number of flights, and average airfare?
2. What is the exact month when the airfare reached its minimum?
3. Why did it take three months for airfare to reach its minimum as compared to passenger volume?
4. Why did it take longer for the number of flights to reach pre-pandemic levels as compared to the passenger volume?
5. Why is the median average airfare still lower than before Covid-19 (in inflation adjusted dollars)?
6. Why has the correlation between passenger volume and airfare we observed during Covid-19 persisted after Covid-19?
7. Why did the correlation between passenger volume and arrival delay increase during Covid, while at the same time the average arrival delay decreased?
8. Did Covid-19 have any effect on seasonality?
9. How did air traffic patterns vary by airport, city, and state? Was there any correlation between Covid-19 restrictions and air traffic patterns?

## SOURCES

1. The data used for the report falls under the Freedom of Information act.

United States. Department of Transportation. (2022). MANAGING RIGHTS. https://doi.org/10.21949/1520564

2. Source for Flight Data
   Dataset Title: Marketing Carrier On-Time Performance (Beginning January 2018)
   U.S. Department of Transportation, Department of Transportation Statistics
   URL: https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGK&QO_fu146_anzr=b0-gvzr
   Dates Accessed: 4/8/2024 - 4/17/2024.
   Other Details: Data retrieved from January 2024 to December 2023 as monthly '.csv' files.
3. Source for Passenger Data
   Dataset Title: Passengers All Carriers – All Airports
   U.S. Department of Transportation, Department of Transportation Statistics
   URL: https://www.transtats.bts.gov/Data_Elements.aspx?Data=2
   Dates Accessed: 4/8/2024 - 4/17/2024.
   Other Details: Select "U.S. Carriers" and each major airport from the dropdowns and retrieve a '.csv' file for every major airport. Data from October 2002 through December of 2023.
4. Source for Air Fare Data by Quarter
   Dataset Title: Passengers All Carriers – All Airports
   U.S. Department of Transportation, Department of Transportation Statistics
   URL: https://www.transtats.bts.gov/AverageFare/
   Dates Accessed: 4/10/2024 - 4/13/2024- 4/18/2024
   Other Details: Airfare price data by quarter, each quarter is a separate file

**Additional resources used for the project:**

**Data processing:**

Python glob function:

https://www.geeksforgeeks.org/how-to-use-glob-function-to-find-files-recursively-in-python/

Replicating Pandas DataFrame columns

https://www.statology.org/pandas-replicate-rows/

**Visualizations:**

https://matplotlib.org/stable/gallery/lines_bars_and_markers/barchart.html#sphx-glr-gallery-lines-bars-and-markers-barchart-py

https://matplotlib.org/stable/api/axes_api.html#module-matplotlib.axes

To fix correlation matrix:

Seaborn heat map doesn't display annotations for all rows #14363,

https://github.com/microsoft/vscode-jupyter/issues/14363, accessed on 4/17/2024.

https://matplotlib.org/stable/gallery/lines_bars_and_markers/barchart.html#sphx-glr-gallery-lines-bars-and-markers-barchart-py

https://matplotlib.org/stable/api/axes_api.html#module-matplotlib.axes

https://medium.com/

Stackoverflow.com

www.python.org