Analysis of Flight Delays Compared to Weather in the Domestic Flight Market

Chase L Munson

Western Governors University

Table of Contents

A. Project Highlights

This project aims to determine whether or not there is a correlation between domestic flight delays and weather in the domestic airline market. Specifically, the intent is to determine whether globally changing weather patterns are causing more or increased delays. This research question was decided upon because if weather is starting to affect the airline market more dramatically, then this information could be used to help inform the market on whether or not they need to factor this into their flight scheduling.

The scope of this project is as follows:

- Collection of Data
- Cleaning of Dataset
- Statistical Analysis of Research Question
- Visualization of Findings
- Final Documentation of Project

The methodology used for this project was the CRISP-DM - Cross-Industry Process for Data Mining (Schröer, Kruse, & Gómez, 2021). This methodology was chosen because it is tailor-made for data mining and was well suited for this endeavor. The programming language used for this project is Python, with Pandas, Matplotlib, NumPy, SciPy, and Seaborn libraries. These libraries help to clean, analyze, test, and visualize the data set. All of the operations were documented in a Jupyter Notebook.

B. Project Execution

This project aims to determine if there is a correlation between weather and domestic airline delays. This was done by first collecting and cleaning the dataset for analysis and documenting those findings in a Jupyter Notebook. This part of the project did not vary from the original proposal. The dataset was cleaned successfully, and minimal adjustments were needed for analysis. Furthermore, all processes were successfully documented in the Jupyter Notebook. The next part of the project consisted of analyzing the dataset to identify any correlations between weather and domestic airline delays. This part of the project did not vary from the initial proposal; statistical results were garnered, and all steps were documented in the Jupyter Notebook. The last objective was to visualize the results in the Jupyter Notebook. This was also done successfully in the form of bar charts and histograms.

The methodology used for this project was the Cross-Industry Process for Data Mining (Schröer, Kruse, & Gómez, 2021). Business understanding consisted of developing a research question for analysis, which, in this case, was whether there is a correlation between weather and domestic airline delays (Schröer, Kruse, & Gómez, 2021). Data understanding consists of determining and collecting the data to be used and checking the quality of that data (Schröer, Kruse, & Gómez, 2021). Data preparation involved cleaning the dataset by removing duplicate rows, outliers, and missing information and doing required data transformations (Schröer, Kruse, & Gómez, 2021). Modeling consisted of performing statistical analysis on the dataset using the Pearson Correlation Test (SciPy, n.d.). The evaluation involved checking the statistical analysis results and determining if there was a correlation, whether or not to reject the null hypothesis, and checking for statistical significance (Schröer, Kruse, & Gómez, 2021). The last step of this methodology, deployment, consisted of creating visualizations for the final report (Schröer, Kruse, & Gómez, 2021).

There was no variance from the original methodology plan for this project. The research question was created, the dataset to be used was checked for quality for the proposed research question and deemed to be acceptable, all the data cleaning steps were performed as needed, statistical analysis was performed with the Pearson Correlation Test (SciPy, n.d.), the statistical results were checked, and a determination was made on the null hypothesis and statistical significance. Visualizations for the final report were made.

Project Timeline

Milestone or deliverable	Duration (hours or days)	Projected start date	Anticipated end date
Data Collection	1 Day	*4/13/2025*	*4/13/2025*
Data Cleaning	1 Day	*4/14/2025*	*4/14/2025*
Data Analysis	2 Days	*4/15/2025*	*4/16/2025*
Data Visualization	1 Day	*4/17/2025*	*4/17/2025*
Report Creation	2 Days	*4/18/2025*	*4/19/2025*

The project timeline did have some variance as data collection and cleaning took less time than expected. Data analysis proceeded as planned. Data visualization took approximately one more day than expected, and report creation was on schedule.

C. Data Collection Process

The dataset chosen to work with is named Flight Delay. This data is hosted on a platform called Kaggle. This dataset has information ranging from 2018 to 2024. This dataset has 29 columns and 30132672 rows. This dataset was put on Kaggle by a user named Arvind Nagaonkar and is licensed under Public Domain usage. The proposed dataset has been chosen as it seems to fit the proposed research question well. It is a good fit as it has information about many delay types, which can be compared to the passage of time to determine the proposed research question. Furthermore, it tracks delays in minutes, which means that if further research on delays is warranted, the dataset could be used in the future. Finally, while this dataset is sourced from a third-party website, the data is from an official government source, which lends it credibility.

The dataset was downloaded from Kaggle as a .zip file, unzipped, and put into a folder tracked by a private GitHub repository called “d502” to keep a backup of progress on this project. The dataset initially came as a parquet file called “Flight_Delay.parquet” and was read into the Jupyter Notebook as a data frame with “pd.read_parquet”. The data collection method did not differ from the original proposal; the .zip file was successfully downloaded and put into the repository, and the parquet file was read into a data frame in the Jupyter Notebook.

The initial assessment of data quality and completeness was good when collecting the data. The column names were noted as being well-named and consistent. Furthermore, each column did have consistent data quality with no outliers or duplicate records. The only noted data cleaning and quality steps were removing columns that were not relevant to the proposed question, splitting two columns related to departure and arrival city and states, and converting the date column into a format readable by the statistical test. This did not vary from the initial assessment, and no issues were encountered when collecting the data.

Regarding data governance issues, the dataset consisted of publicly accessible flight information containing various columns related to flight delays. For added transparency, the dataset was preserved and not altered beyond the documented steps in the Jupyter Notebook. This project had no unexpected data governance issues, as the dataset used contained only publicly accessible information with no types of privileged information.

C.1 Advantages and Limitations of Data Set

The main advantage of this dataset is that it is based on an official government dataset and has various forms of information relating to delays that can be analyzed. This makes it very useful for the proposed research question.

The main disadvantage is that the dataset only has information from 2018-2024. This makes it more challenging to develop a more accurate trend line regarding the research question. Furthermore, 2020 and the COVID-19 pandemic may also impact the results.

D. Data Extraction and Preparation

Data extraction consisted of downloading from Kaggle as a .zip file, unzipping it, and putting it into a folder tracked by a private GitHub repository called “d502” to keep a backup of progress on this project. The dataset initially came as a parquet file called “Flight_Delay.parquet” and was read into the Jupyter Notebook as a data frame with “pd.read_parquet”. There were no issues with this process step, and it proceeded as planned.

The preparation section of the process also went well. Preparation consisted of exploring the dataset for outliers, duplicates, and missing values and performing some fundamental transformations for the analysis. This was done with functions such as the head() for an initial view of the dataset, info() to check on data types and entry counts, isnull(), sum(), to check for null records, duplicated().sum() to check for duplicate records, .max() .min() functions on the WeatherDelay column to check for outliers, and the .unique() function on the OriginCityName, DestCityName, and Year columns to check for unique items. Finally, an initial scatterplot was made to give a general idea of possible information to be garnered. No duplicates, missing values, or outliers were noted.

Transformation also proceeded as expected. Un-needed columns relating to dates, specific airlines, departure times, arrival times, taxi times, take-off and landing times, elapsed time, air time, and distance were dropped with .drop(). Furthermore, a new column called “CleanDates” was created for statistical analysis, as the provided date column was unsuitable as it was in object format. This was done using mdates.date2num(). Another transformation step that was performed was the splitting of the Columns “OriginCityName and “DestCityName” into “OriginState”, “OriginCity”, and “DestState”, “DestCity” for possible future analysis. This was performed using .split(). Finally, all the remaining columns of type float64 were converted to int64 for consistency with .astype().

E.1 Data Analysis Methods

Both exploratory and descriptive data analyses were used to answer the research question. The chosen null hypothesis for this project was that there is no correlation between weather and domestic flight delays. The alternative hypothesis was selected to be that weather would positively correlate with domestic airline delays. To try to disprove the null hypothesis, a new dataset called "df_weather_delays" was made, consisting of only rows of data with a delay. The Pearson Correlation test was then performed on the "CleanDates" column that was created against the column "WeatherDelay" (SciPy, n.d.). The Pearson Test was chosen as both of the variables used were quantitative, and it resulted in a statistical score and an p-value that could be used to determine correlation and statistical significance (SciPy, n.d.).

After determining statistical significance and correlation, an exploratory analysis was performed. Bar charts showing average weather delays by year were made to visually indicate whether or not there was an increase in the average delay time year over year. Secondary bar charts were made to show the average monthly weather delay for all dataset years for further visual representation. Histograms were also created to show the count of all weather delays for the entire dataset in 30-minute increments in the range of 30-300 minutes. A secondary set of histograms showing the same 30-minute increments in the range of 30-300 minutes was created for each year of the dataset for further breakdown. These charts were made to visually represent the impact of weather delays in a simple manner that prospective viewers can understand.

E.2 Advantages and Limitations of Tools and Techniques

The programming language chosen for this project was Python, with Pandas, Matplotlib, NumPy, SciPy, and Seaborn libraries. These libraries helped to clean, analyze, test, and visualize the data set. All of the operations were documented in a Jupyter Notebook. Python was chosen for its many advantages, mainly the various libraries that are tailor-made for data analysis. Disadvantage-wise, the main one discovered is the learning curve required to learn basic Python language and the need to type out commands manually.

Pandas main advantage is that it is a very versatile library for Python that supports many operations used to clean and transform a dataset. The disadvantage of this tool is similar to Python in that it can be challenging to learn and use. NumPy and SciPy are advantageous because they help simplify the process of statistical testing (SciPy, n.d.). The disadvantages of these two libraries are related to the learning curve required to use these tools.

Matplotlib and Seaborn have the advantage of being powerful visualization tools that can be used to visualize data findings. However, they are also more complicated than a tool such as Tableau. Jupyter Notebooks' main advantage is that it supports the organization of the whole data analysis process in notebooks. The main disadvantage is the difficulty of installing this tool before use.

E.3 Application of Analytical Methods

- Importation of dataset into Notebook using pd.read_parquet()
- Data Exploration:
  - head() for an initial check of the dataset layout
  - info() to check for column types, entry count
  - isnull.sum() to check for null values in columns
  - duplicated.sum() to check for duplicate records
  - .min() and .max() on “WeatherDelay” to check for outliers
  - .unique() on “OriginCityName”, “DestCityName” and “Year” to check values
  - plt.scatter() with “Year” and “WeatherDelay for initial visualization before cleaning
- Data Cleaning:
  - .drop() to remove columns relating to dates, specific airlines, departure times, arrival times, taxi times, take-off and landing times, elapsed time, air time, and distance
  - .split() to split “OriginCityName”, and “DestCityName” into “OriginCity”, “OriginState”, “DestCity”, and “DestState” for possible future more granular analysis
  - .astype() to convert remaining columns from float64 to int64 for consistent values
  - mdates.date2num() to create “CleanDates” from the “FlightDate” column for data analysis

    - .to_csv() to create "df_clean" dataset for analysis
- Data analysis:
    - Creation of "df_weather_delays" from df_clean for a data frame with only rows that have a weather delay
    - stats.pearsonr() on "CleanDates and "WeatherDelay" for an r-value score and a statistical significance score (SciPy, n.d.)
    - sns.barplot() to create a barplot with the columns "Year" and "WeatherDelay" to show the average weather delay in minutes by year.
    - sns.FacetGrid() and sns.barplot() to show average weather delay in minutes monthly, broken down by year
    - sns.displot() to create a histogram showing counts of delays for the entire dataset in increments of 30 minutes in the range of 30-300
    - sns.displot() to create histograms showing counts of delays by year, in increments of 30 minutes in the range of 30-300

F Data Analysis Results

F.1 Statistical Significance

To determine statistical significance and correlation, The Pearson Correlation Test was used on the quantitative variables in the columns "CleanDates" and "WeatherDelay" on a copy of the dataset with only rows that had a weather delay present (SciPy, n.d.). This was done to focus on the central question of whether weather correlates with flight delays. The null hypothesis was chosen to be that there is no correlation between weather and domestic flight delays. The alternative hypothesis was selected to be that weather positively correlates with domestic airline delays. This test provided a statistical result that can be used to determine correlation strength and a p-value that can be used to determine statistical significance (SciPy, n.d.). The chosen alpha value for this project is 0.05. This value will be used to either reject or accept the null hypothesis. A correlation value of over 0.0 for this research question will be considered successful.

The statistical scores that were generated are as follows:

Statistic	P-Value
0.04109091354365808	7.378745366859716e-144

Due to the P-value being less than 0.05, there is enough evidence to reject the null hypothesis and support the alternative hypothesis that weather correlates with domestic flight delays. While the statistical score is minimal, it still suggests a positive correlation.

F.2 Practical Significance

The practical significance of this project is that airlines could use it to determine if they need to emphasize weather more. This information can be used to, for example, determine if flight schedules may have to be adjusted as weather patterns continue to shift. Furthermore, airlines can use this information to inform them to start making machine-learning algorithms to judge whether or not to proactively cancel flights and offer flight changes to consumers to cut down on overall delays.

F.3 Overall Success

Based on the above information, this project has been successful. This project intended to do an initial analysis of the possible, increasing impact of weather on flight delays in the domestic market. This was done to help airlines determine whether they want to place more emphasis on researching this particular topic and decide if they want to attempt a more granular analysis or develop new tools to help forecast these events. Due to the null hypothesis being rejected and a positive correlation, this topic garners further research.

G. Conclusion

G.1 Summary of Conclusions

This project intended to determine if there was a correlation between domestic flight delays and weather. This was done with the intent that airlines may use it to help them decide whether or not to study this topic more in the future. I did so using Python with Pandas, Matplotlib, NumPy, SciPy, and Seaborn. These libraries helped to clean, analyze, test, and visualize the data set. All of the operations were documented in a Jupyter Notebook. I then performed a Pearson Correlation Test on two columns containing quantitative data, “CleanDates” and “WeatherDelay,” to determine the correlation and statistical significance (SciPy, n.d.). This resulted in a statistical score that rejected the null hypothesis and accepted the alternative hypothesis that weather correlates with flight delays.

Furthermore, a positive correlation, though minimal was garnered, showing that delays are increasing. Finally, I created bar charts to show the average delays by year, and monthly, broken down by year. I also created histograms showing total delay counts in 30-minute increments and delay counts by year in 30-minute increments.

G.2 Effective Storytelling

I chose a mix of bar charts and histograms for this project to show the selected variables from multiple angles. I chose bar charts to show the average weather delay in minutes and visualize the overall delay length increase year over year. I then made a second set of bar charts showing the same information broken down by year and month for further visualized information. Bar charts with an average of delays helped represent the findings because they very concisely show an increase in average delays in a simple way that most prospective viewers will understand.

The second set of visuals that I made was two sets of histograms. I made these to show the count of delays in 30-minute increments up to 300 minutes. The first histogram is based on all of the delays, so a viewer can get an overall view of how many delay events there were, and then the second set is broken down by year so that you can see the difference year over year. These histograms helped to represent the findings as they give a clearer picture of the amount of delays, their severity, and how they impact the averages in the bar charts.

G.3 Recommended Courses of Action

The first recommendation based on the information garnered by this analysis is for airlines to emphasize researching this topic further and starting to mitigate the impact of the results before they become worse. While the correlation is small, it still shows that the subject should be researched further to determine how much this will impact the flight market and possibly profit margins.

My second recommendation is to implement machine learning on these results to attempt to predict future weather delays. This, once again, is related to attempting to mitigate the issue before it worsens, but more importantly, it may help determine how much worse it can get. This would also be valuable information for the airlines as they can start the process of preparing to mitigate now instead of later.

H Panopto Presentation

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=1942db05-c7c7-443d-8b5a-b2c30160e07c

References

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science, 181*, 526-534. doi:https://doi.org/10.1016/j.procs.2021.01.199

SciPy. (n.d.). *pearsonr*. Retrieved April 11, 2025, from SciPy: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html

Appendix A

Additional Supporting Files

- Flight Data Dataset: https://www.kaggle.com/datasets/arvindnagaonkar/flight-delay/data
- Jupyter Notebook File
- Jupyter Notebook in .html format