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

Chase L Munson

Western Governors University

Table of Contents

A. Proposal Overview

A.1 Research Question or Organizational Need

This proposal 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 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.

A.2 Context and Background

This project was chosen because weather patterns in the United States are changing. We seem to be seeing more frequent severe weather events in every region of the United States. Due to this, the research question of whether or not flight delays are causing more frequent delays has been proposed. This is an important topic that must be studied, as there is an innate concern that delays may continue to become more frequent, and determining if weather is contributing would be helpful for airlines to decide on whether to study this topic further.

This project aims to determine whether or not weather changes are starting to impact the domestic airline market. This will be done by analyzing a dataset from the public platform Kaggle. The chosen dataset, Flight Delay, has a considerable amount of information about various delays, including weather, which will be this project's scope. Other relevant information in the dataset includes “FlightDate”, “Year”, “OriginCity”, “DestCity”, and “WeatherDelay”. The last column, “WeatherDelay”, contains a value in the form of minutes of delay time that will

be used extensively in this project. The “FlightDate” and “Year” columns will help track if delays increase as time passes. This project is being conducted to help airlines determine if they should be factoring weather delays into their decision-making process.

A.3 and A3A Summary of Published Works and Their Relation to the Project

Review of Work 1

Recently, in March of this year, Suzanne R. Kelleher’s Article on Forbes.com, Midwest Blizzard: Over 4,400 Flights Disrupted—From Kansas City To D.C. was put online. In this article, Kelleher (2025) spoke of several significant impacts of a massive winter storm on flights across a large swath of the United States (Kelleher, 2025). In the article, Kelleher (2025) states that: “Airports reporting double-digit flight cancellations include major hubs such as Charlotte/Douglas in N.C., Reagan National in Washington, D.C., Chicago O’Hare, Kansas City, Dallas and Philadelphia” (Kelleher, 2025). The author also states, “69%. That’s the percentage of flight delays caused by weather in the U.S” (Kelleher, 2025).

This article speaks to the importance of the research question presented, as significant weather events such as the one above can significantly impact a large swath of the United States. This article speaks to how one major weather event can dramatically affect the entire aviation industry and the importance of seeing if there is a trend toward more of these events impacting the aviation market in the United States.

Review of Work 2

In the journal article: Meteorological Impacts on Commercial Aviation Delays and Cancellations in the Continental United States, which focuses on weather impacts at a more granular level, Goodman & Griswold (2019) report that: “Weather creates numerous operational and safety hazards within the National Airspace System (NAS). In 2014, extreme weather events attributed 4.3% to the total number of delay minutes recorded by the Bureau of Transportation Statistics” (Goodman & Griswold, 2019). They also state, “When factoring weather’s impact on the NAS delays and aircraft arriving late delays, weather was responsible for 32.6% of the total number of delay minutes recorded” (Goodman & Griswold, 2019).

This journal article is essential to the question posed as it helps confirm that weather delays significantly affect the airline market. This enforces the need to determine whether or not these types of delays are increasing and need to be more thoroughly researched as we move forward with more and more weather pattern shifts over the next decade. While the research question does not focus as granularly on the data as this journal article does, this article helps to reinforce that this information needs to be examined from multiple angles.

Review of Work 3

In September 2024, Porter Fox wrote the article: The Dawn of Superstorms. In this article, Fox (2024) reports the increase in severe hurricanes and extreme weather (Fox, 2024). In one section of the article, Fox (2024) reports:

> More Category 4 and 5 hurricanes hit the U.S. mainland from 2017 to 2021 than from 1963 to 2016. Hurricanes today also last longer than they once did and move slower, multiplying the damage. Rapid intensification used to spin up once a century, but studies show that in the future, it could occur more frequently—especially in waters bordering

> the East Coast—putting cities like New Orleans, Houston, Tampa, and Charleston, S.C., at higher risk (Fox, 2024).

The author also states, “A recent study by Brooklyn’s First Street Foundation also shows how hurricanes will penetrate farther inland in decades to come, affecting U.S. states as far west as New Mexico, Kansas, and Wisconsin” (Fox, 2024).

This article points to the prevalence of increasing extreme weather in typically storm-prone areas and areas that generally have less severe weather. This is important to the research question as it shows that weather events, specifically extreme ones, are increasing. This article supports the fact that these events are growing and that analysis of these events compared to weather delays could derive meaningful insights that need further investigation.

A.4 Summary of Data Analytics Solution

The solution for this project is to use a downloadable flight delay dataset from the website Kaggle.com with information from 2018-2024. This dataset contains 29 columns and 30132672 rows of data. The columns contain various information concerning Dates, Time, Airline, Origin, Destination, Arrival, Departure, Delay Information, and Distance. For our use case, the relevant fields are the Date and Delay columns. This dataset is downloadable as a .zip file, the chosen method to acquire the data. The dataset will be worked on in a Jupyter Notebook to track and document the process easily. Cleaning, analyzing, and visualizing will be done in the Jupyter Notebook. Python with the libraries Pandas, Matplotlib, Numpy, Scipy, and Seaborn will be used to perform operations on the dataset and garner insights. Python will be used as it supports a variety of data analysis libraries for our use case. Pandas and Numpy will be used to clean and analyze the dataset. Scipy will be used for statistical analysis using the Pearson test.

The Pearson correlation test will be performed on the flight date, and the weather delay columns to attempt to prove the statistical significance of increasing weather delays over time (SciPy, n.d.). Finally, Matplotlib and Seaborn will be used to visualize the data insights from the dataset.

A.5 Benefits and Support of Decision-Making Process

The benefits that analyzing this dataset will provide are an attempt to garner information on whether domestic airlines should be placing a greater emphasis on weather delays at this time or whether they are currently focusing enough on the problem. This insight will help them determine whether they are doing so. If not, this will allow the airlines to decide if they want to look at weather delays more granularly and decide whether to prioritize the issue more. The insights garnered from the data analysis will be used to determine whether or not airlines may want to place a greater emphasis on weather delays when scheduling or adjusting flight schedules.

B. Data Analytics Project Plan

B.1 Goals, Objectives, and Deliverables

- Goal 1: Determine if there is a correlation between Weather and Domestic Airline Delays
    - Objective 1.1: Collect and Clean the Flight Delay Dataset for analysis.
        - Deliverable 1.1.1: Cleaned dataset ready for analysis.
        - Deliverable 1.1.2: Documented cleaning process steps in the Jupyter Notebook.

- Objective 1.2: Analyze the dataset to identify any correlation between Weather and Domestic Airline Delays.
    - Deliverable 1.2.1: Statistical analysis results concerning the proposed question.
    - Deliverable 1.2.2: Documented steps of the analysis in the Jupyter notebook.
- Objective 1.3: Visualise the results of the dataset analysis for more straightforward communication of results.
    - Deliverable 1.3.1: Visualization of results in the Jupyter Notebook.

B.2 Scope of Project

B.2.A Included in Project Scope

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

B.2.B Not included in Project Scope

- Data outside the years 2018-2024
- Real-time Data Tracking

B.3 Standard Methodology

CRISP-DM - Cross-Industry Process for Data Mining

- **Business Understanding**: Develop a research question for analysis, in this case, if there is a correlation between Weather and Domestic Airline Delays (Schröer, Kruse, & Gómez, 2021).
- **Data Understanding**: Determine and collect data to be used, explore and check the quality of data selected for analysis (Schröer, Kruse, & Gómez, 2021).
- **Data Preparation**: Cleaning of the dataset will occur. Duplicate rows, outliers, missing information, and transformation will occur (Schröer, Kruse, & Gómez, 2021).
- **Modeling**: Statistical analysis will be performed with the Pearson Correlation test (SciPy, n.d.).
- **Evaluation**: Check the statistical results and determine whether or not to reject the null hypothesis or if statistical significance exists (Schröer, Kruse, & Gómez, 2021).
- **Deployment**: Create visualizations and final report (Schröer, Kruse, & Gómez, 2021).

B.4 Timeline and Milestones

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*

B.5 Resources and Costs

- **Hardware**
    - Windows 11 Desktop Computer: No Cost
    - MacBook Laptop: No Cost
- **Software**
    - Anaconda: No cost
    - Jupyter Notebook: No Cost
    - Python & Python Libraries: No cost
- **Work**
    - 7 Days: No Cost

B.6 Criteria for Success

Several criteria will be used to evaluate the success of this project. Firstly, statistical significance will have to be proven to disprove the null hypothesis. This will be done by obtaining a p-value using the Pearson test to attempt to prove statistical significance (SciPy, n.d.). Secondly, visual representations of the findings need to be made to show the findings to any respective person or organization easily and understandably. Finally, a final report has to be made to distribute these findings to individuals who may use this information to guide their decisions in the future.

C. Design of Data Analytics Solution

C.1 Hypothesis

This project intends to determine if weather is increasingly impacting flights over time. This will be measured by comparing weather delays to dates to see if there is a trend toward increasing delays or significance.

Null Hypothesis: There is no correlation between weather and domestic flight delays.

Alternate Hypothesis: Weather will positively correlate with domestic airline delays.

C.2 and C.2.A Analytical Method

The method that will be used is the Pearson correlation test, as both of the data columns I will use are quantitative (SciPy, n.d.). The first column, "CleanDates," will be created from a date field into a format the model can read. The second column that will be used is called "WeatherDelay" and reports how long each delay is in minutes. These two columns will be compared to see if there is any correlation and statistical significance to the research question.

The Pearson correlation has been chosen as it is well suited to comparing two quantitative variables and will result in a statistical score and r value that can be used to determine correlation and statistical significance, respectively (SciPy, n.d.).

C.3 Tools and Environments

The programming language chosen for this project is Python, with Pandas, Matplotlib, NumPy, SciPy, and Seaborn libraries. These libraries will help clean, analyze, test, and visualize the data set. All of the operations will be documented in a Jupyter Notebook.

C.4 and C.4.A Methods and Metrics to Evaluate Statistical Significance

The Pearson correlation test is the chosen method to determine statistical significance and correlation. This test will provide an r-value and p-value, which will be used to determine how strong the correlation is and whether the null hypothesis can be rejected (SciPy, n.d.). The null hypothesis in this case is that there is no correlation between flight delays and weather. If the p-value is less than 0.05, we will reject the null hypothesis and assume that we have statistical significance. Otherwise, we accept the null hypothesis. In the case of this research question, any r-value greater than 0.0 will be considered a success, and conversely, any r-value less than 0.0 will be regarded as unsuccessful.

The Pearson correlation test was chosen because two quantitative variables are used (SciPy, n.d.). After considering all options, I found the simplicity of the test and the fact that it is tailor-made to compare these types of variables more than enough to make it the chosen test. I have chosen the metrics above as studying our research question seems novel, and we are just starting to understand how weather patterns are shifting. Even if the r-value is minimally positive, it will show that weather impacts flight delays more. In this case, the results being statistically significant and rejecting the null hypothesis alone would be considered a success.

C.5 Practical Significance

The practical significance of this project is that it may help airlines determine if they need to start 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.

C.6 Visual Communication

Visualisations that will be used for communication purposes include bar charts to show the theorized overall increase in delays by year. Histograms will be used to show the changes over time at a more granular level. Bar charts will help to visualize the research question as they will show the increase in delays year over year, if they exist, in a way that any prospective person or group will understand. Furthermore, bar charts will help communicate this data as the dataset is quite large, making it more straightforward to visualize the changes. Seaborn will be used to create the bar charts for the visualizations and will show each year and the respective impact of weather delays that year. For the bar charts, the year will be put on the x-axis, and weather delays on the y-axis.

Histograms will be used to show how long delays are over a larger timeframe, such as monthly, to communicate if the length of delays visually changes over time. Histograms will help explain the research question as they will show the possible increase in delays, or lack thereof, at a more granular level, making it easy to see the trends. These will also be produced using Seaborn and will likely show monthly information for each year to garner information on

specific times of year that are more delay-prone. Similar to the bar charts, on the histograms, the months will be put on the x-axis, and weather delays on the y-axis.

D. **Description of Dataset**

D.1 Source of Data

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 dataset can be found at this link: https://www.kaggle.com/datasets/arvindnagaonkar/flight-delay/data. This dataset can be downloaded directly from Kaggle as a .zip file.

D.2 Appropriateness of Dataset

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 the severity of 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.

This dataset contains several columns relating to: Dates, Time, Airline, Origin, Destination, Arrival, Departure, Delay Information, and Distance. In this particular circumstance, we will only look at the columns relating to dates and delays to compare and

determine trends. The delay columns contain delays in minutes, which will be appropriate for seeing if delays are increasing. This will be compared to the date column to determine the proposed research question. For this use case, the dataset in question is more than sufficient.

D.3 Data Collection Methods

This 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 comes as a parquet file called “Flight_Delay.parquet” and will be read into the Jupyter Notebook as a data frame with “pd.read_parquet”.

D.4 Observations on Quality and Completeness of Data

The initial assessment of data quality and completeness was good. The column names were noted as being well-named and consistent. Furthermore, each column seems to have consistent data quality with no outliers or duplicate records. The only noted data cleaning and quality steps will likely be removing columns that are 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.

D.5 and D.5.A Data Governance, Privacy, Security, Ethical, Legal, and Regulatory Compliances

- **Data governance**: This dataset uses publicly accessible flight information containing various columns related to flight delays. For transparency, the dataset will be preserved and not altered beyond the documented steps in the Jupyter Notebook.
- **Privacy**: This dataset uses only publicly available flight information and contains no personally identifiable information. Privacy concerns are minimal.
- **Security**: This dataset uses publicly available flight information and no information that could be considered compromising. Due to this, security concerns are minimal.
- **Ethical, legal, and regulatory compliance considerations**: All of the data used is publicly available information and does not contain any information that would violate privacy and/or security. Furthermore, the code, libraries, and platforms are open-source and publicly accessible.

References

Fox, P. (2024, September 30). The Dawn of Superstorms. *TIME Magazine, 204*(9/10), 27-28.

Goodman, C. J., & Griswold, J. D. (2019). Meteorological Impacts on Commercial Aviation Delays and Cancellations in the Continental United States. *Journal of Applied Meteorology & Climatology, 58*(3), 479-494. doi:https://doi.org/10.1175/JAMC-D-17-0277.1

Kelleher, S. R. (2025, March 5). Midwest Blizzard: Over 4,400 Flights Disrupted—From Kansas City To D.C. *Forbes.com*. Retrieved from Forbes.com: https://www.forbes.com/sites/suzannerowankelleher/2025/03/05/midwest-blizzard-flights-disrupted-from-kansas-city-to-dc/

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