Analysis of Flight Delays in the USA

[Shallau Muhamad & Ozan Irmak]

19.02.2025

Contents

1 Analysis of Flight Delays in the USA

This report was created as part of the R Bootcamp module, focusing on data analysis and visualization using R. The project explores flight delays in the USA, applying statistical methods and R programming techniques to uncover patterns and insights. Through data cleaning, feature engineering, hypothesis

testing and modeling, this analysis demonstrates the practical application of R for real-world data challenges.

1.1 Introduction

Delays serve as a central performance indicator in the transport sector and are defined in commercial aviation as the difference between the scheduled and actual departure times. Regulatory authorities employ a variety of metrics to quantify delay levels and thus evaluate the overall functioning of the air traffic system. For example, in 2018, 24% of flights in Europe were delayed by more than five minutes, while in the United States 21% of flights were delayed by over 15 minutes and in Brazil in 2017, 19% of flights were either cancelled or experienced delays of over 30 minutes. These statistics underscore the relevance of delays as an index that significantly influences airline network dynamics, regardless of carrier size (Carvalho et al., 2021, p. 1).

Flight delays have far-reaching consequences beyond their obvious economic drawbacks, affecting passengers, airlines and airports alike. Due to uncertainty about on-time departures, travellers often plan to arrive much earlier than necessary, incurring additional travel costs. At the same time, airlines face penalties, fines and increased operating costs, such as longer ground times for crews and aircraft. Moreover, higher fuel consumption coupled with increased emissions contributes to additional environmental burdens, thereby calling into question the sustainability of air transport (Carvalho et al., 2021, p. 1).

Delays also jeopardise the marketing strategies of airlines, as customer loyalty and the attractiveness of frequent flyer programmes depend heavily on punctuality. Unreliable timeliness negatively influences consumer decision-making. Empirical studies have also demonstrated correlations between delay levels and variables such as airfares, aircraft sizes, flight frequencies and the number of service complaints. Therefore, accurately forecasting flight delays can help optimise tactical and operational decisions for airport operators and airline managers, while providing passengers with timely information to adjust their travel plans (Carvalho et al., 2021, p. 1).

The commercial aviation sector continuously generates large volumes of data, which are archived in databases to better understand the complex flight ecosystem. Increasingly, data science approaches are employed to extract valuable insights from flight data generated by sensors and the Internet of Things, thereby addressing complex issues in the context of big data (Carvalho et al., 2021, p. 1).

1.1.1 Use Case

Within the framework of this project, we were commissioned by a US government agency to conduct a comprehensive analysis of departure delays in the United States. The objective of this study is to identify the key factors influencing delays and, based on these findings, develop strategic measures to reduce delays and enhance the efficiency of the air traffic network.

This analysis is based on three primary datasets, which are detailed in the following sections:

1. **Flight Data (2023):** Detailed information on departures and associated delays.
2. **Weather Data:** Records of meteorological conditions at airports.
3. **Geographic Data:** Spatial information utilised for a differntiated airport analysis.

By combining these datasets, a robust foundation for systematically investigating the influence of weather conditions, aircraft age, airport location and other factors on departure delays is established.

1.2 Dataset Overview

In this section, the datasets used for the analysis are presented in detail. The structure, key variables and contents of each dataset are explained to provide a clear understanding of the data foundation. Additionally, the relevance of each dataset to the analysis is highlighted, showing how they contribute to investigating flight delays.

1.2.1 Flight Data (2023) (`US_flights_2023.csv`)

The dataset provides detailed information on all flight movements in the USA for the year 2023 and forms a comprehensive basis for analysing delay patterns, assessing the performance of individual airlines and investigating potential correlations between the technical characteristics of aircraft and their punctuality.

- **Core Variables and Contents:**
  - **Departure Airport** (`Dep_Airport`): IATA code of the departure airport
  - **Arrival Airport** (`Arr_Airport`): IATA code of the arrival airport
  - **Departure Delay** (`Dep_Delay`): Departure delay in minutes
  - **Arrival Delay** (`Arr_Delay`): Arrival delay in minutes
  - **Tail Number** (`Tail_Number`): Aircraft registration number
  - **Aircraft Age** (`Aircraft_age`): Age of the aircraft in years
  - **Distance** (`Distance`): Flight distance in kilometers

Additionally, the dataset includes further information that can be categorized into the following areas:

- **Flight Data:** Recording of the flight date and the day of the week on which the flight occurred.
- **Airlines and Aircraft:** Information on the airline, the unique identification of the aircraft (Tail Number), as well as manufacturer and model details.
- **Departure and Arrival Information:** Detailed data regarding the departure and arrival airports and their associated cities.
- **Time Data:** Records of scheduled and actual departure and arrival times, including the delay duration in minutes and a categorisation of these delays.
- **Causes of Delays:** Identification of various factors that can lead to delays, such as operational issues of the airline, weather conditions, delays in the national air traffic system (NAS), security checks, or delays caused by preceding flights.
- **Flight Duration and Distance:** Recording of the actual flight duration along with the corresponding distance information.
- **Aircraft Characteristics:** Detailed information regarding the manufacturer, model and age of the aircraft used.

This extensive data collection enables a systematic and detailed analysis of the causes of flight delays, from which strategic measures can be derived to improve the efficiency of air traffic operations.

1.2.2 Weather Data (`weather_meteo_by_airport.csv`)

The dataset includes historical meteorological data collected at various airports, providing a robust foundation for analysing weather conditions and their potential impact on air traffic. The data spans several days and covers a variety of relevant meteorological parameters essential for both operational and safety-related issues.

- **Key Variables:**
  - **Timestamp (`time`):** The exact date of the weather recording.
  - **Airport ID (`airport_id`):** The IATA code of the respective airport.
  - **Temperature Values:**
    - Average temperature (`tavg`) in °C,
    - Minimum temperature (`tmin`) in °C,
    - Maximum temperature (`tmax`) in °C.
  - **Precipitation Values:**
    - Total precipitation (`prcp`) in millimetres,
    - Snowfall (`snow`) in centimetres.
  - **Wind Parameters:**
    - Wind speed (`wspd`), originally recorded in km/h,
    - Wind direction (`wdir`) in degrees.
  - **Air Pressure:**
    - Barometric air pressure (`pres`) in hPa.

This comprehensive dataset enables a systematic examination of the relationship between current weather conditions and various aspects of air traffic, such as flight delays and safety measures. It also provides a solid basis for further analyses aimed at optimising operational processes and improving forecast accuracy for weather-related disruptions in air traffic.

1.2.3 Geographic Data (`airports_geolocation.csv`)

The dataset comprises comprehensive geographic information on airports, uniquely identified by internationally standardised IATA codes. In addition to the basic variables that precisely describe an airport's location using latitude (`LATITUDE`) and longitude (`LONGITUDE`), the dataset includes supplementary attributes. These include, among others, the official name of the airport (`AIRPORT`), the corresponding city (`CITY`), the US state (`STATE`) and the country (`COUNTRY`), with the data primarily covering the United States. This detailed data collection provides an essential foundation for geographical analyses in air traffic, enabling the visualisation and examination of flight routes as well as the evaluation of location factors that significantly influence airport operations and efficiency.

1.3 Hypotheses

Flight delays at airports are a key performance indicator in air traffic management (ATM) and are defined as the difference between scheduled and actual departure times. Regulatory authorities use a variety of metrics to quantify these delays and assess the efficiency of the entire air transportation system. Such delays can be caused by fluctuations in demand, capacity constraints due to congestion, adverse weather conditions, operational delays or network disruptions - a phenomenon whose occurrence has increased significantly over the last decade due to the rapid growth of air traffic (Chandra & Verma, 2025, p. 1).

Against this background, this study examines the underlying determinants of departure delays. Six specific hypotheses were formulated and empirically tested to systematically capture both structural and external influencing factors. The selection of these hypotheses is based on the need to shed light on the complex mechanisms that contribute to delays in a differentiated manner and to develop a comprehensive understanding of the influencing factors. The hypotheses and their methodological approach are presented in detail below:

Hypothesis 1: Identification of the worst 10 airports with the highest average departure delays

- **Research Question**: Which airports in the USA have the highest average departure delays?
- **Methodology**: The average delays are calculated per airport and the results are visualized both as a ranking and on a geographical map of the USA.

Hypothesis 2: Finds out when the flights departure at the worst 10 Airports.

- **Research Question**: Distribution of Departure Times
- **Methodology**: The departures are summed up in a stacked bar chart.

Hypothesis 3: Influence of wind and precipitation combinations on departure delays

- **Research Question**: Does a specific combination of wind speed and precipitation amount lead to significantly higher delays?
- **Methodology**: A heat map is created to systematically analyze the influence of meteorological parameters on departure delays.

Hypothesis 4: Relationship between wind direction and delays

- **Research Question**: Do certain wind directions cause longer delays than others?
- **Methodology**: The delays are visualized as a function of the wind direction using a wind rose representation.

Hypothesis 5: Seasonality of departure delays

- **Research Question**: Are there certain months in which departure delays occur significantly more frequently?
- **Methodology**: Monthly average values of delays are analyzed and the temporal trends are visualized in order to identify seasonal patterns.

Hypothesis 6: Influence of aircraft age on departure delays

- **Research Question**: Are older aircraft more prone to delays than newer models?

- **Methodology**: The aircraft are divided into age categories (e.g. “new”, “medium”, “old”) and the distributions of delays are compared using boxplots.

By systematically testing these hypotheses, a structured framework is created that enables the various causes of departure delays to be investigated in detail. The resulting findings form the basis for the development of targeted measures to optimize efficiency in air traffic.

1.4 Data Cleaning and Preparation

Data preparation is organized into three fundamental phases sampling, feature extraction and labeling that collectively establish a robust, well-structured database optimized for a model. In parallel, comprehensive data cleaning and preprocessing procedures are systematically applied to ensure the accuracy and reliability of subsequent analyses (Boonpan & Sarakorn, 2025, p. 2).

These procedures include the selection of pertinent columns, rigorous data cleansing, standardization and feature engineering. Together, these steps not only safeguard the integrity and relevance of the dataset but also provide a solid foundation for effective model development.

1.4.1 Removal of Irrelevant Columns

Certain variables were removed as they did not contribute to the analytical objectives:

Flight Data Exclusions:

- `Tail_Number`
- `Arr_Airport`
- `Arr_CityName`
- `Arr_Delay`
- `Arr_Delay_Type`

These fields primarily pertain to arrival information or aircraft identifiers that do not influence departure delay analysis.

1.4.2 Handling Implausible Values

Exclusion of Negative Delays:
Observations where `Dep_Delay < 0` were removed, as negative values indicate flights that departed earlier than scheduled. These records could distort the analysis of departure delay patterns.

1.4.3 Date Format Standardisation

To ensure consistency across datasets, the original date format (`YYYY-MM-DD`) was converted to the European convention (`DD.MM.YYYY`). This transformation enhances readability and aligns with other temporal data processing steps.

1.5 Feature Engineering

To enhance the analytical depth, new categorical variables were derived to facilitate segmentation and interpretation.

Aircraft Age Categorisation

Aircraft age was grouped into three operational categories:

- **New (0–5years):** Represents recently manufactured aircraft.
- **Mature (6–15years):** Covers mid-life aircraft with regular operation.
- **Old (>15years):** Older aircraft, potentially more prone to delays due to maintenance needs.

Wind Speed Categorisation

Wind speed (`wspd`) was categorised into four operational risk levels:

- **Low (0–5 km/h):** Minimal impact on flight operations.

- **Moderate (6–15 km/h):** Slight potential influence on delays.
- **High (16–30km/h):** Increased likelihood of flight disruptions.
- **Severe (30km/h):** High turbulence risks, possibly leading to significant delays.

By segmenting these continuous variables into interpretable categories, we can more effectively analyse their impact on departure delays.

Wind Direction The code filters the `combined_data` dataset to include only flights departing from airports listed in `top_flight_data$Dep_Airport`. It then adds a new column, `Wind_Direction`, which categorizes the wind direction (`wdir`) into compass sectors of 45° increments. The sectors are labeled as "N", "NE", "E", "SE", "S", "SW", "W" and "NW", with the range from 0° to 360°, including boundary values.

1.6 Data Integration and Merging

To construct a comprehensive dataset, three independent sources (flight records, weather data and airport geolocation information) were integrated using **Left Joins**.

1.6.1 Merging Flight Data with Weather Information

```
flights_weather <- flights %>%
  left_join(weather, by = c("date" = "date", "Dep_Airport" = "airport_id"))
```

1.6.2 Adding Geospatial Airport Data

```
combined_data <- flights_weather %>%
  left_join(geo_data, by = c("Dep_Airport" = "IATA_CODE"))
```

This integration enriches each flight record with meteorological and geospatial attributes, enabling a multifaceted analysis of potential delay factors.

1.7 Normalisation Considerations

As part of the data preparation process, the possibility of normalizing numerical variables was explored. Two methods were considered:

1.7.1 1. Min-Max Normalisation

- **Description:** Scales numerical values to a [0,1] range, preserving relative differences.
- **Advantage:** Ensures comparability across variables.
- **Disadvantage:** Highly sensitive to extreme values (outliers).

1.7.2 2. Standardisation (Z-score Transformation)

- **Description:** Centers variables by subtracting the mean and dividing by the standard deviation.
- **Advantage:** Robust to outliers than min-max scaling.
- **Disadvantage:** Transformed values lose direct interpretability.

1.7.3 Final Decision

Despite the benefits of normalisation in machine learning applications, the analysis prioritised interpretability over uniform scaling. Instead of normalising, aggregation techniques (e.g., computing average delay times at the airport level) were employed to mitigate distortions caused by extreme values.

---

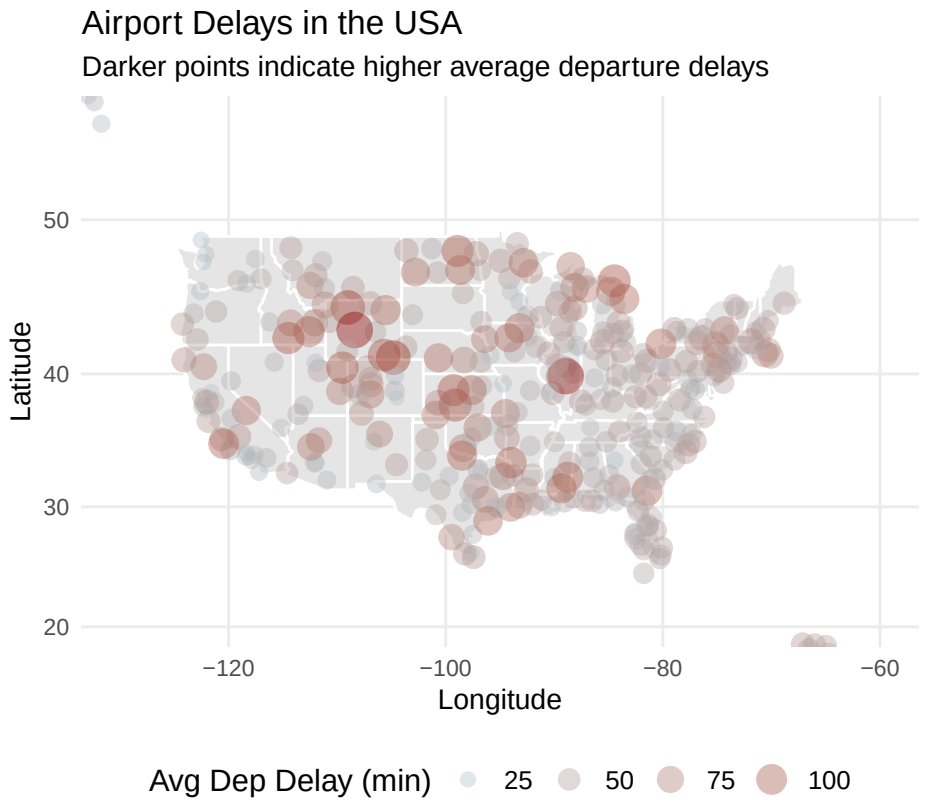
2 Hypothesis 1: Worst 10 Airports with the Highest Delays

At the beginning of our analysis, **all airports** included in our data set are visualized on a map of the United States. Each airport appears as a dot whose color intensity reflects the average departure delay value.

This visualization method is inspired by modern approaches to geographic visualization, which aim to communicate complex spatial patterns in an interactive and comprehensible way (Grainger et al., 2016, p. 302). The use of such interactive tools - such as dynamic maps that can also integrate temporal changes using sliders - makes it possible to depict not only static, but also seasonal and dynamic aspects of the data (Grainger et al., 2016, p. 302). In this way, the visualization contributes to a transparent representation of the spatial distribution of flight delays and at the same time supports interdisciplinary communication by presenting complex relationships in an accessible format (Grainger et al., 2016, p. 302).

2.1 Geographical Distribution of Departure Delays in the U.S.

The visualization depicts the **geographical distribution of average departure delays (`Avg Dep Delay`)** at U.S. airports. Each point represents an airport, with **delay intensity encoded using a color scale from light to dark red**. **Darker points indicate longer delays**.



Concentration of High Delays in Specific Regions

A **notable concentration of high delays is observed in the Midwest and along the East Coast**. Airports such as **Chicago (ORD), Denver (DEN) and New York (JFK, LGA, EWR)** are particularly affected, likely due to **high flight frequencies, capacity constraints and meteorological influences**. **Highly frequented air traffic hubs** tend to experience longer delays, especially under adverse weather conditions.

Lower Delays in the Southwest and West Coast

In contrast, airports **in the Southwest and parts of the West Coast** exhibit **lower average delays**. This may be attributed to **more favorable weather conditions and more efficient operations**. However, **Los Angeles (LAX) and San Francisco (SFO)** still show **moderate delays**, potentially due to **high passenger volumes and capacity limitations**.

Key Takeaways and Recommendations

In summary, the visualization highlights that **departure delays are significantly influenced by geographical, meteorological and infrastructural factors**. Airports experiencing **recurring delays** could benefit from:

- **Optimized flight routes**
- **Improved weather forecasting systems**
- **More efficient handling processes**

Implementing these measures could enhance **air traffic reliability** and reduce **systematic delays** at major airports.

However, given the large number of airports, it becomes difficult to extract meaningful insights directly.

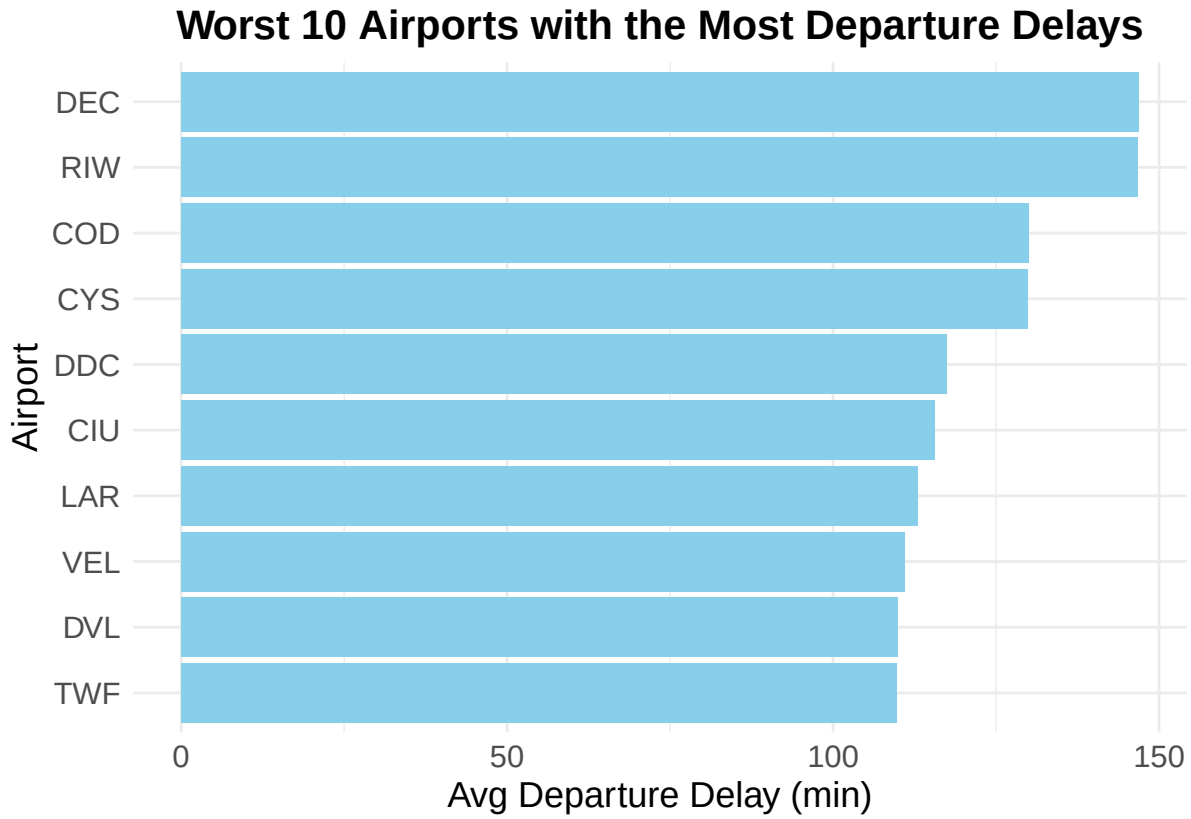
> **To refine our analysis, we will focus on the worst 10 airports with the highest average departure delays.**

2.2 Focusing on the Worst 10 Airports with the Highest Departure Delays

As mentioned earlier, due to the large number of airports in our dataset, we narrow our focus to the worst 10 airports with the highest average departure delays. These airports consistently experience the most significant delays, making them key locations for further analysis.

To identify them, we calculate the **average departure delay** for each airport and rank them accordingly.

Statistical visualizations - for example in the form of bar charts - are primarily used to communicate quantitative data clearly and comprehensibly (Grainger et al., 2016, p. 301). In this context, the following **bar chart** provides an example of the **worst 10 airports** with the highest average delays, thus enabling a direct comparison of these performance indicators.



Dep_Airport	AIRPORT
CIU	Chippewa County International Airport
COD	Yellowstone Regional Airport
CYS	Cheyenne Airport
DDC	Dodge City Regional Airport
DEC	Decatur Airport
DVL	Devils Lake Regional Airport
LAR	Laramie Regional Airport
RIW	Riverton Regional Airport
TWF	Magic Valley Regional Airport (Joslin Field)
VEL	Valdez Airport

By focusing on these airports, we can better understand the patterns and potential causes behind prolonged departure delays.

The visualization presents the **ten airports with the highest average departure delays (`Avg Departure Delay`)**, measured in minutes. The **horizontal bars** represent airports, ranked in **descending order**, with longer bars indicating **more severe delays**.

Key Findings and Regional Disparities

- DEC and RIW exhibit the longest delays, exceeding 140 minutes, suggesting operational inefficiencies, adverse weather conditions, or infrastructural constraints.
- COD, CYS and DDC also show prolonged delays above 100 minutes, indicating systemic issues or external disruptions.
- TWF, DVL and VEL, while still among the worst, report comparatively lower delays (around 100 minutes).

Regional and Structural Implications

Notably, **major international hubs are absent**, suggesting that **smaller regional airports may struggle more with on-time departures** due to **resource limitations, air traffic congestion, or weather-related disruptions**.

Conclusion and Recommendations

In summary, **DEC and RIW face the most critical delays**, highlighting the need for **airport-specific operational improvements**, such as:

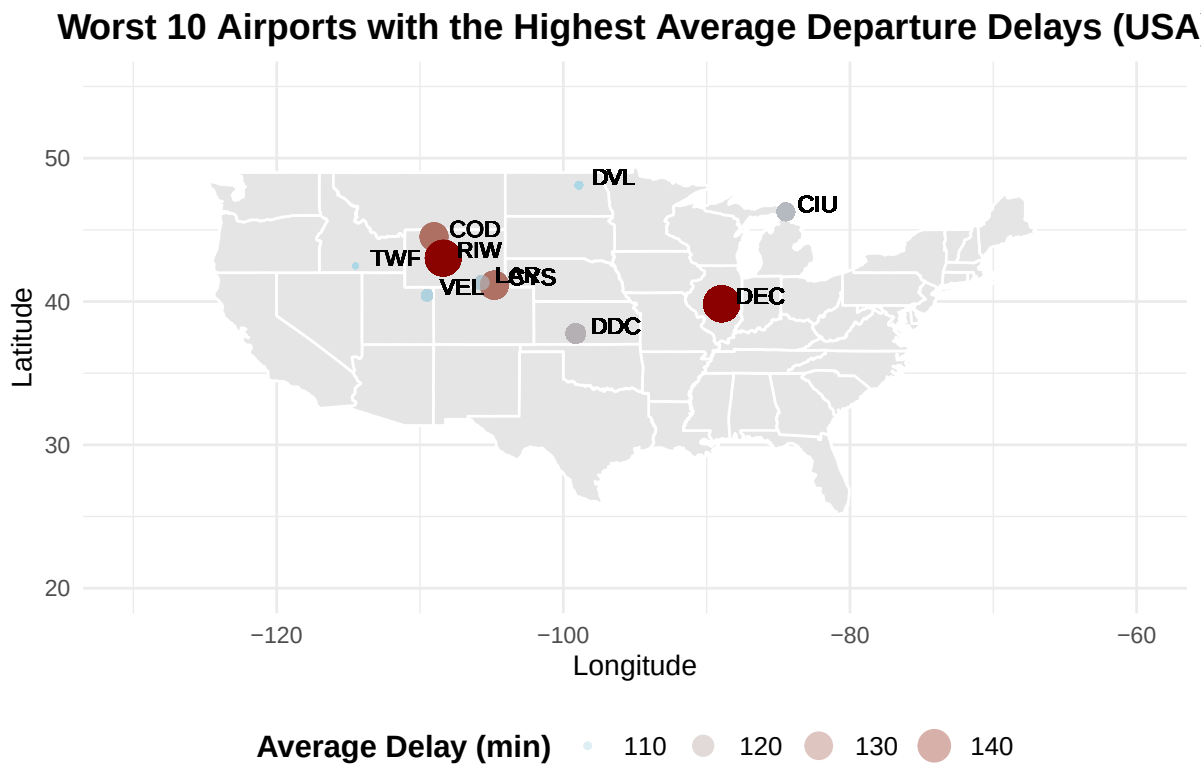
- **Optimized scheduling strategies**
- **Enhanced air traffic control coordination**
- **Infrastructure upgrades**

Implementing these measures could help **mitigate delays effectively and enhance airport efficiency**.

All the airports displayed experience significant delays exceeding **100 minutes on average**, highlighting potential operational or external issues that may require further investigation.

2.3 Worst 10 Airports on the U.S. Map

The visualization displays the **worst 10 U.S. airports with the highest average departure delays (`Total Delay (min)`)** using a **geospatial representation**. Each point represents an airport, where **larger and darker red circles indicate longer delays**. The **X-axis (longitude) and Y-axis (latitude)** provide a **geographic distribution of delays**.



Regional Delay Patterns

A **notable concentration of delays is observed in the central and western U.S.**, particularly at **DEC, COD and RIW**, which show **the longest delays (above 130 minutes)**. This suggests **operational inefficiencies, weather disruptions, or infrastructure constraints**. Airports **such as TWF, VEL and DDC** also experience **moderate delays**.

As shown in the table and text above, **most delayed airports are regional rather than major hubs**, indicating that **resource limitations, reduced scheduling flexibility and weather sensitivity** contribute to prolonged delays. **DEC (Decatur) stands out with the highest departure delays**.

Geographic and Structural Factors

While **the Midwest and western regions** exhibit **higher delays**, the **East Coast is underrepresented**, suggesting that **larger hub airports manage delays more effectively despite high flight volumes**. Additionally, **airports near mountainous or extreme-weather regions (e.g., COD, RIW, TWF)** may face **seasonal weather disruptions impacting operations**.

Key Takeaways

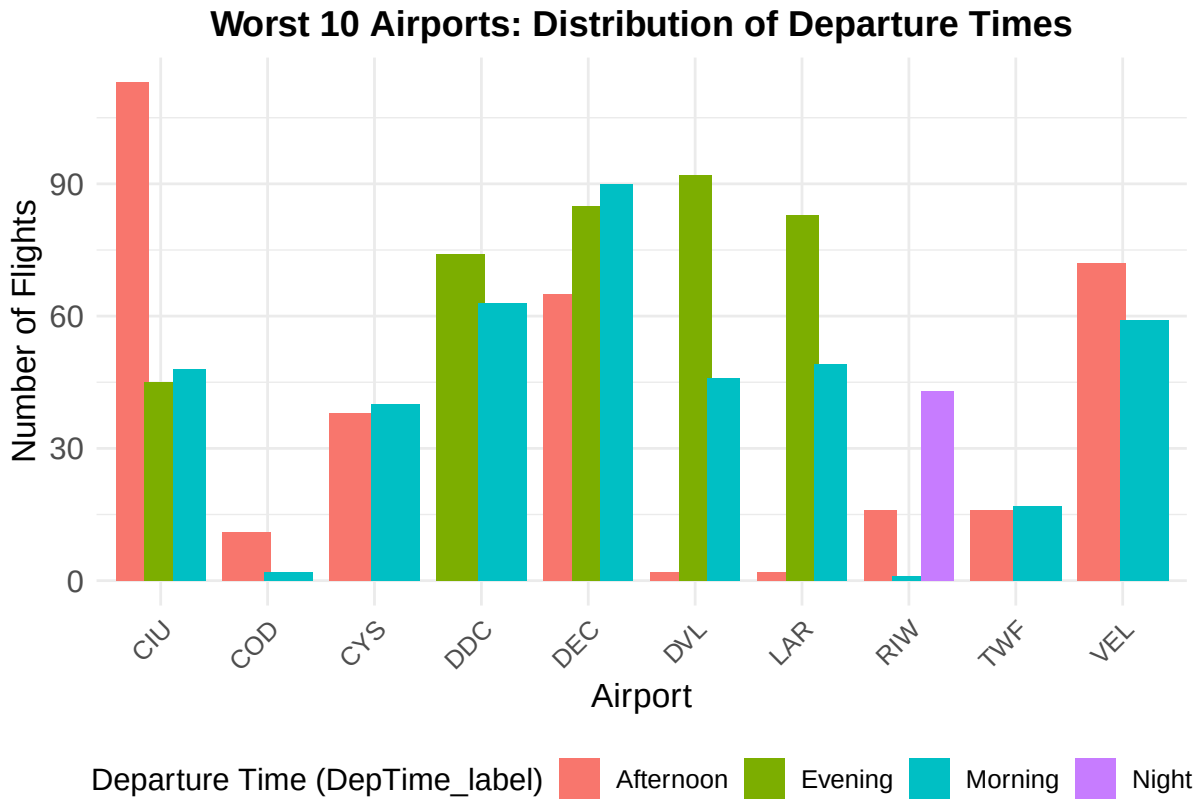
This visualization highlights **regional disparities in departure delays**, underscoring the need for **targeted improvements**, such as:

- **Enhanced weather preparedness** for high-risk airports
- **Optimized scheduling and resource** allocation to reduce congestion and improve efficiency, particularly at regional airports.

---

3 Hypothesis 2: Worst 10 Airports: Distribution of Departure Times

This stacked bar chart illustrates the **distribution of departure times** across the **worst 10 airports** with the highest average departure delays. Each bar represents a specific airport, with the different colors indicating the proportion of flights departing during different time periods: **Morning, Afternoon, Evening and Night**.



Departure Time Distribution at the Worst 10 Most Delayed Airports

The **X-axis represents the airports**, while the **Y-axis indicates flight frequency**. This highlights **departure trends at each airport**.

Key Observations

- **CIU is dominated by afternoon departures**, with minimal flights at other times, suggesting **operational or route constraints**.
- **DEC, DVL and LAR have balanced morning and evening departures**, indicating **peak-hour demand**.
- **TWF and COD operate almost exclusively during daytime**, while **RIW has notable night departures**.
- **VEL and DDC show a mix of morning and evening flights**, reflecting **flexible scheduling**.

Possible Influences

- **Operational Constraints:** Runway availability and air traffic control influence departure timing.
- **Passenger Demand:** Higher evening flight volumes at **DEC and DVL** may indicate **commuter or business travel patterns**.

- **Weather & Geography:** Adverse conditions may dictate optimal departure windows.
- **Noise Regulations:** Limited **night flights suggest local restrictions** on late departures.

Conclusion

Departure schedules vary by airport due to **operational, environmental and regulatory factors**. Recognizing these patterns aids **airport planning, airline scheduling and passenger experience optimization**.

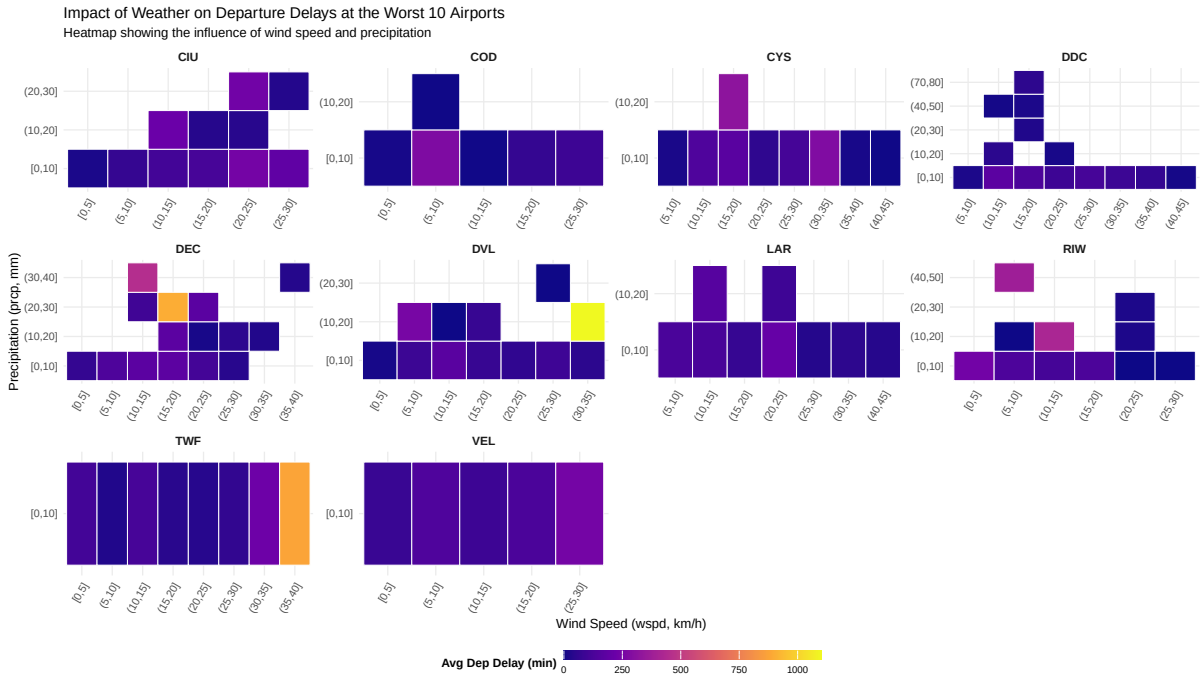
---

4 Hypothesis 3: Impact of Wind and Rain Combinations on Delays

Weather conditions, particularly wind speed and precipitation, can significantly influence departure delays at airports. This hypothesis explores the combined impact of these two factors on average departure delays. To investigate this, a **heatmap** is used to represent the relationship between wind speed, precipitation and their effects on delays.

Description of the Heatmap

Heatmaps are an established method for visualizing matrix-like data in which colors are used as informative aesthetic elements (Gu, 2022, p. 1). A basic distinction is made between spatial heat maps, which represent geographically distributed patterns - for example, the global temperature distribution or user clicks on websites - and raster or grid heat maps, in which a rectangular color grid contrasts two variables (Gu, 2022, p. 1).The latter are often rearranged using hierarchical clustering to highlight subsets with similar patterns. In this paper, the focus is exclusively on the grid heatmap (Gu, 2022, p. 2).



Impact of Meteorological Conditions on Departure Delays at the Worst 10 Airports

The present **faceted heatmap** analyzes the impact of **meteorological conditions** on **departure delays** at the **ten airports with the highest average delays**. The visualization represents the relationship between **wind speed (`wspd`)**, **precipitation (`prcp`)** and **average departure delay (`Avg Dep Delay`)**. Each facet plot represents an individual airport. The **X-axis displays wind speed**, while the **Y-axis represents precipitation levels in millimeters**. The **intensity of delays** is encoded using a **color scale ranging from dark purple to yellow**, where **darker shades indicate lower delays** and **yellow to orange tones represent longer departure delays**.

Variability in Weather Impacts Across Airports

A key finding of the analysis is that the **impact of wind and precipitation on delays varies by location**. Airports such as **RIW, DVL and DEC** exhibit **significantly longer delays at higher wind speeds or increased precipitation levels**. In particular, **RIW shows a noticeable concentration of yellow areas at high wind speeds (40–50 km/h)**, suggesting a **strong influence of meteorological factors on airport operations**. Similarly, **DVL and DEC experience a sharp increase in departure delays when moderate to strong winds coincide with precipitation levels exceeding 10 mm**. These airports may be **especially vulnerable to weather-related disruptions**, possibly due to **infrastructure limitations or more frequent occurrences of adverse weather conditions**.

Airports with Minimal Weather Influence

In contrast, airports such as **VEL and TWF show minimal color variations in their facet plots**, indicating that **neither wind speed nor precipitation has a significant impact on departure delays at these locations**. A possible explanation could be that these airports either **have superior infrastructure that mitigates weather-related delays** or that their **geographical location is less exposed to extreme weather conditions**.

Axis Scaling and Comparative Analysis

Another **noteworthy aspect** is the **scaling of the axes**. While the **X-axis segments wind speeds from 0 to 80 km/h**, the **Y-axis extends up to 30 mm of precipitation**. This enables a **detailed differentiation between moderate and extreme weather conditions**. At the same time, the **faceted design allows for a direct comparative analysis across airports**, helping to **identify site-specific characteristics and structural weaknesses**.

Key Insights and Operational Implications

In summary, this **visualization highlights that meteorological factors play a significant role in departure delays at some airports**, whereas other locations remain largely unaffected. The analysis provides **valuable insights for airport management, as well as for operational and strategic measures to minimize weather-related delays**. Airports that exhibit **high sensitivity to wind and precipitation** could benefit from:

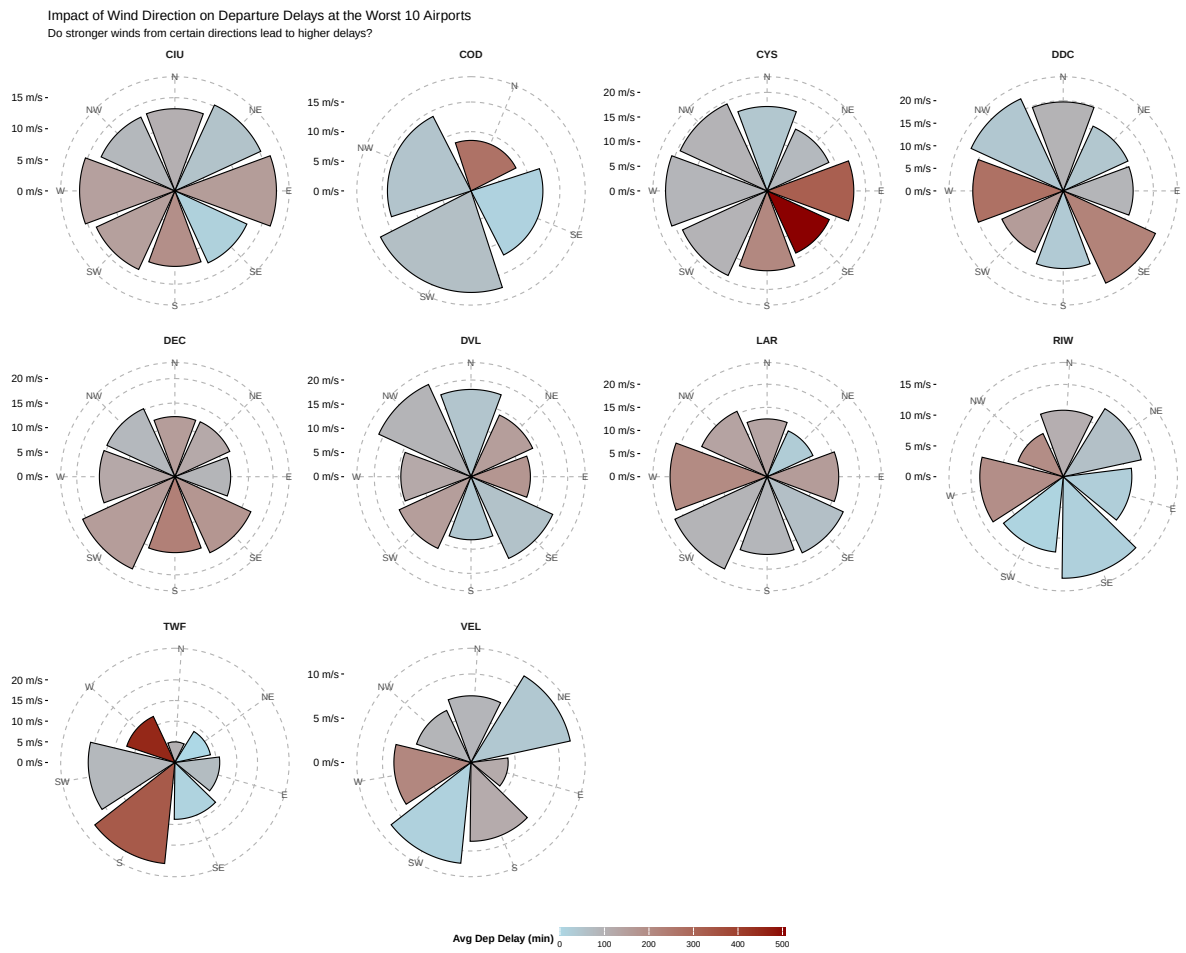
- **Adjusted runway usage strategies**
- **Improved de-icing systems**
- **Optimized flight scheduling**

Implementing such measures could help **reduce weather-induced disruptions and enhance operational efficiency** in the aviation sector.

---

5 Hypothesis 4: Which Wind Direction Causes the Most Delays?

Wind direction plays a crucial role in flight operations, influencing runway usage and departure schedules. This hypothesis investigates whether certain wind directions contribute to longer delays than others. To explore this, we use a **wind rose chart** that visualizes the relationship between wind direction, wind speed and average departure delays.



Impact of Wind Direction on Departure Delays at the Worst 10 Airports

The present visualization illustrates the impact of **wind direction** on **average departure delays (`Avg Dep Delay`)** at the ten airports with the highest delays. Each facet plot represents the distribution of departure delays as a function of **wind direction (`Wind_Direction`)** and **average wind speed (`wspd`)**. Wind direction is visualized along the circular axis of the wind roses, while the radial axis represents wind speed in **meters per second (`m/s`)**. The **color scale** indicates the average delay, with **light blue areas representing lower delays and dark red areas indicating significant departure delays**.

Analysis of Airport-Specific Wind Influence

The analysis reveals **substantial differences among airports in how wind direction influences departure delays**. Airports such as **CYS, DDC and RIW** exhibit **distinct delay patterns in specific wind sectors**. Notably, **CYS and DDC experience increased delays under south-eastern (`SE`) or southwestern (`SW`) winds**, whereas **RIW shows significant delays associated with northwestern (`NW`) winds**. These patterns may be attributed to **topographical conditions, runway configurations, or operational constraints**. Strong **headwinds or crosswinds** from particular directions can significantly impact takeoff and landing procedures, leading to **delays due to safety measures or operational inefficiencies**.

Conversely, airports such as **VEL and TWF do not exhibit clear delay patterns associated with wind direction**. The **even distribution of colors suggests that no particular wind direction consistently causes high delays at these locations**. This could be attributed to **a more stable infrastructure, optimized operational processes, or a geographical setting with less extreme wind conditions**.

Relationship Between Wind Speed and Departure Delays

Another key pattern observed in the visualization is the **relationship between wind speed and departure delays**. Airports like **DVL and RIW** show **increased delays at wind speeds exceeding**

10 m/s, particularly when combined with specific wind directions. High wind speeds can introduce operational challenges, as they impact **aircraft control, takeoff procedures and ground operations**. **Strong crosswinds** can further limit aircraft maneuverability, leading to **delays due to additional safety protocols, extended taxi times, or altered takeoff and landing approaches**.

Key Takeaways and Implications

In summary, the visualization highlights that the **impact of wind direction on departure delays varies significantly across airports**. While some airports are highly sensitive to **specific wind directions and high wind speeds**, others remain largely unaffected. These findings are particularly relevant for **airport management and flight planning**, as they provide **insights into which airports may require targeted operational adjustments or infrastructure improvements** to mitigate **weather-related delays**.

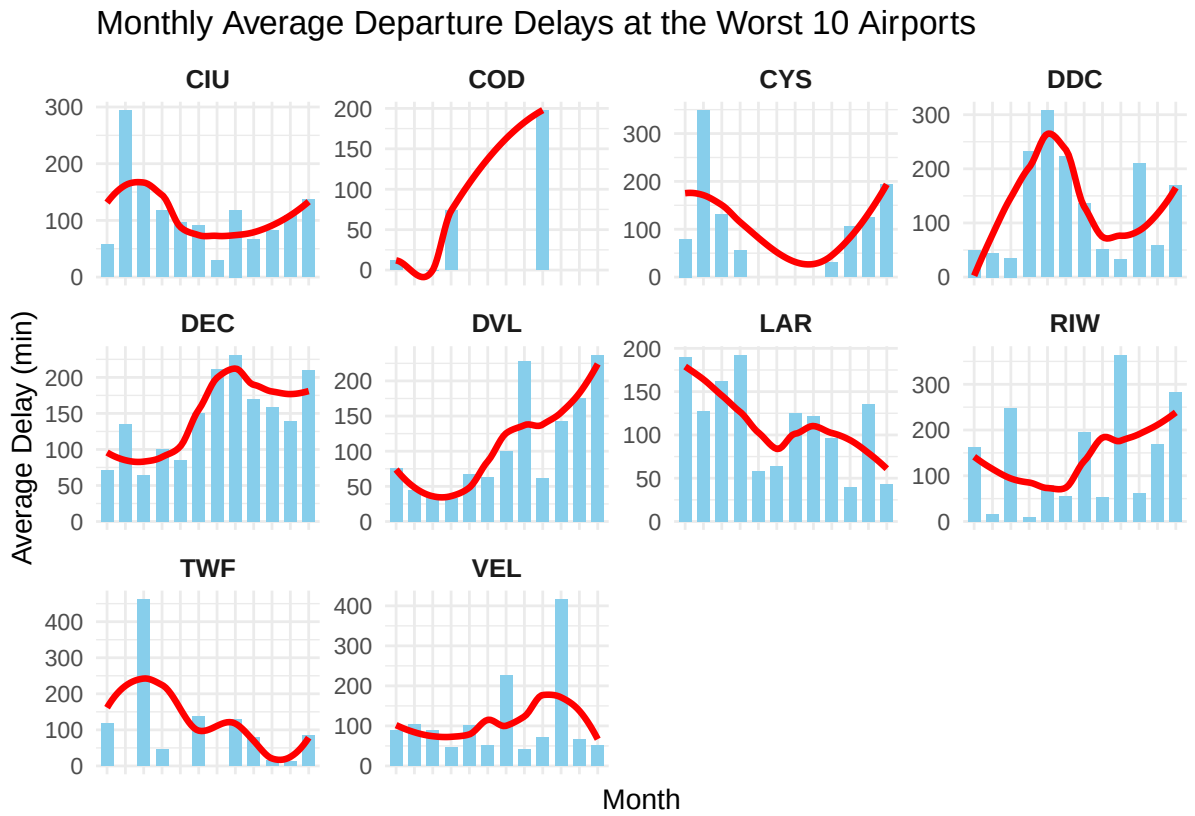
Airports exhibiting **consistent delay patterns** under specific wind conditions could benefit from:
- **Adjusted runway usage strategies**
- **Improved forecasting models**
- **Refined operational procedures**

Ultimately, these measures could help **minimize disruptions and enhance overall efficiency** in airport operations.

---

6 Hypothesis 5: Seasonality of Departure Delays

Departure delays may exhibit seasonal trends, as various factors such as weather conditions, holiday travel and air traffic congestion fluctuate throughout the year. This hypothesis investigates whether certain months are associated with more frequent or severe delays. A **bar chart** is combined with a **trend line** to analyze monthly average departure delays.



Seasonal Analysis of Monthly Departure Delays at the Worst 10 Airports

The visualization presents the **monthly average departure delays (`Avg Dep Delay`)** at the **ten**

airports with the highest delays. Each facet plot displays the **monthly average delays for a specific airport over one year**. The **blue bars represent the average delays per month**, while the **red trend line provides a smoothed (`loess`-smoothing) representation of seasonal patterns**.

Seasonal Variations in Departure Delays

A key pattern emerging from the analysis is the **clear seasonality of departure delays**. At several airports, including **DDC, DEC and RIW**, delays are **significantly higher during the winter months (November to February) compared to summer**. This pattern can be explained by multiple factors, particularly **meteorological conditions in the winter season**, which lead to **frequent flight cancellations, de-icing delays and increased airport congestion due to weather-related restrictions**. Additionally, **higher travel demand during the holiday season in December and January** could contribute to these delays, as overburdened airports often face **capacity constraints**.

Conversely, airports such as **DVL, LAR and VEL** exhibit a **different seasonal pattern**. While their **departure delays are not as pronounced in winter as in the previously mentioned airports**, there is a **noticeable increase in delays during summer and early autumn (June–September)**. This could be attributed to **weather-related disruptions such as summer thunderstorms, increased wind speeds, or tropical storms**, which particularly affect the **central and southern regions of the United States**. Likewise, **operational bottlenecks due to higher passenger volumes during the summer vacation period** might play a role in increased delays.

Irregular Delay Patterns at Some Airports

Some airports, however, do not exhibit **clear seasonal patterns**, instead showing **sporadic peaks in certain months**. For instance, **TWF and COD display an irregular pattern**, where **specific months exhibit significant delay spikes, while others remain relatively stable**. These fluctuations could be linked to **local events, unusual weather conditions, seasonal maintenance operations, or temporary changes in flight schedules**.

The **smoothed trend lines (red curves) provide a valuable means of identifying overarching patterns and long-term developments in monthly delays**. Airports such as **DDC and DEC show a clear increase in delays during winter and a decline in summer**, whereas airports such as **DVL or LAR exhibit the opposite trend, with peaks in summer**. These differences suggest that **various climatic and operational factors vary by airport**, indicating that **a one-size-fits-all approach to reducing delays may not be equally effective across all locations**.

Key Insights and Operational Recommendations

In conclusion, this analysis highlights that **monthly delays are strongly influenced by seasonal factors**, with **effects varying depending on an airport's geographical location and operational infrastructure**. Airports that **consistently experience high delays in specific seasons** could implement **targeted measures to optimize their operational processes**, including:

- **Improved weather forecasting**
- **Optimized staffing strategies during peak months**
- **Technical measures to minimize weather-related delays**

Implementing such measures could help **reduce weather-induced disruptions and enhance operational efficiency** across various airport locations.

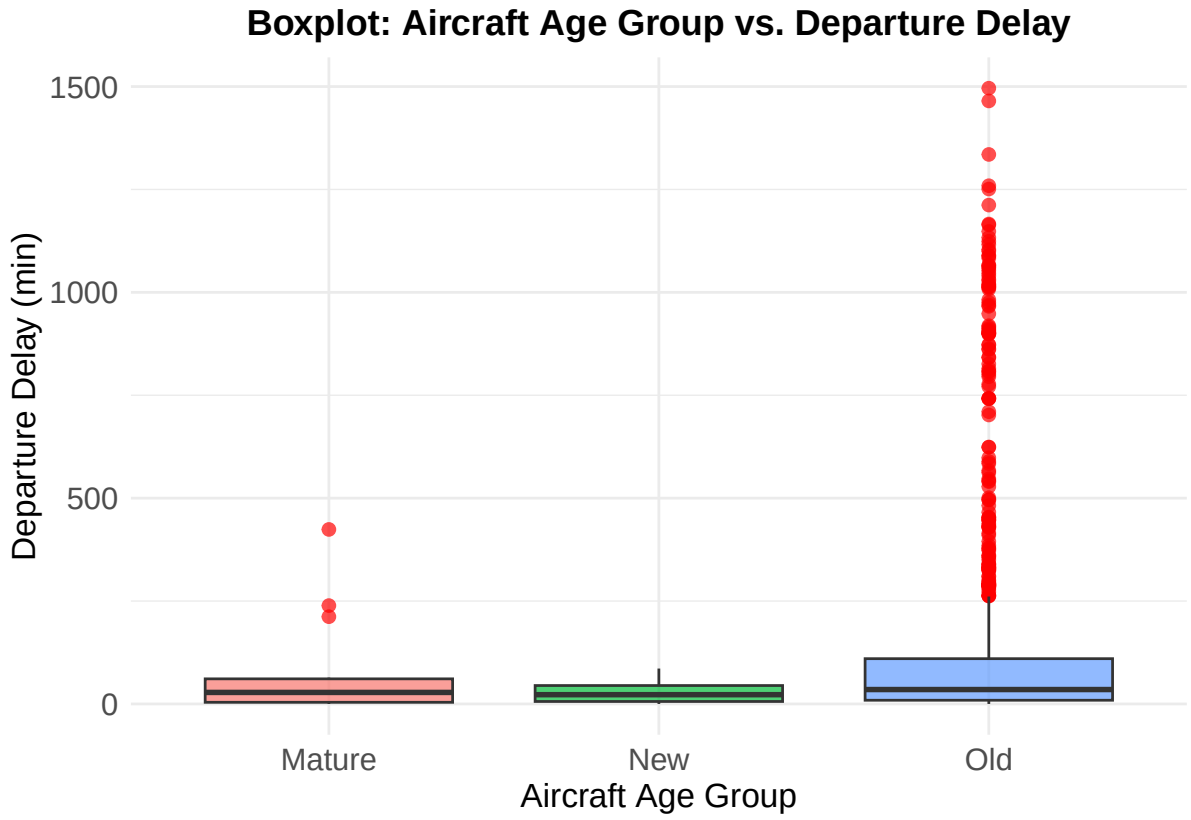
---

7 Hypothesis 6: Impact of Aircraft Age on Delays

Aircraft age can significantly impact operational reliability and efficiency. This hypothesis investigates whether **older aircraft** are more prone to departure delays compared to newer models. To analyze this, aircraft are classified into three age groups, which were shown in the Feature Engineering part. Three boxplots are created:

- 1. A standard boxplot to visualize the distribution of delays.
- 2. A log-transformed boxplot to better handle the effect of extreme outliers.

- 3. A R Squared transformed boxplot



7.1 Standard Boxplot

The visualization presents the first boxplot which **examins the relationship between aircraft age groups and departure delays**. The **X-axis categorizes aircraft as New, Mature, or Old**, while the **Y-axis represents departure delays in minutes**. The boxplots illustrate **delay distributions**, with the **central box showing the interquartile range (IQR), the median delay as a horizontal line and whiskers representing data spread. Red dots indicate outliers**, representing extreme delays.

Key Findings

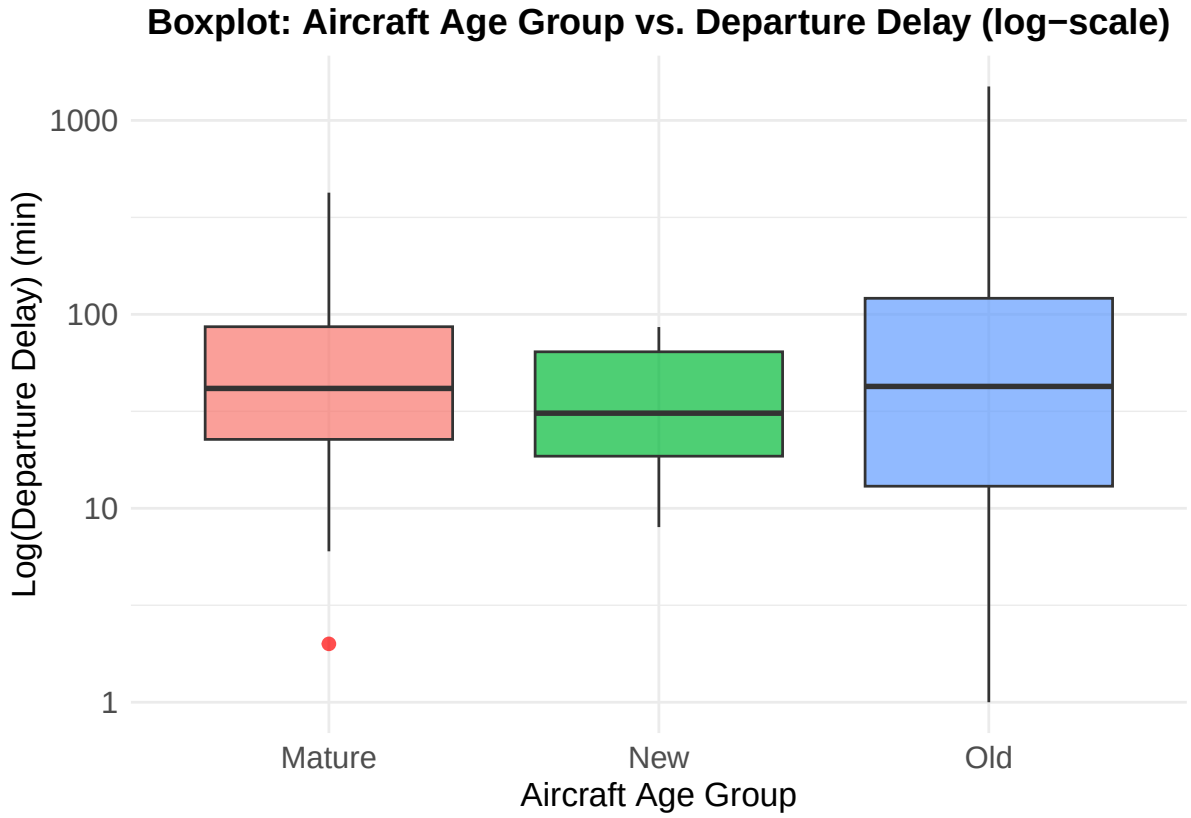
- **Old aircraft experience significantly higher delays**, with a **higher median, broader variability and numerous extreme outliers** exceeding **1,000 minutes**.
- **New aircraft show the lowest median delays** and **minimal variance**, indicating **greater reliability and operational efficiency**.
- **Mature aircraft fall between the two**, exhibiting **moderate delays with fewer outliers than old aircraft**.

Potential Causes

- **Mechanical Wear**: Older aircraft are **prone to maintenance-related delays**.
- **Operational Factors**: Airlines may **assign older aircraft to less flexible routes**, increasing delay risks.
- **Regulatory Compliance**: Stricter maintenance and safety checks can **extend turnaround times**.

7.2 Log-Scaled Boxplot

The second plot applies a **logarithmic scale** to the delay data, which compresses the influence of extreme outliers and highlights patterns in the central distribution.



Observations:

- The log-scale plot reveals that while older aircraft consistently experience more delays, the distribution among all groups becomes more comparable.
- This transformation emphasizes that even though **Old** aircraft have more frequent extreme delays, their central tendency (median and interquartile range) is not as drastically different as suggested by the standard scale.

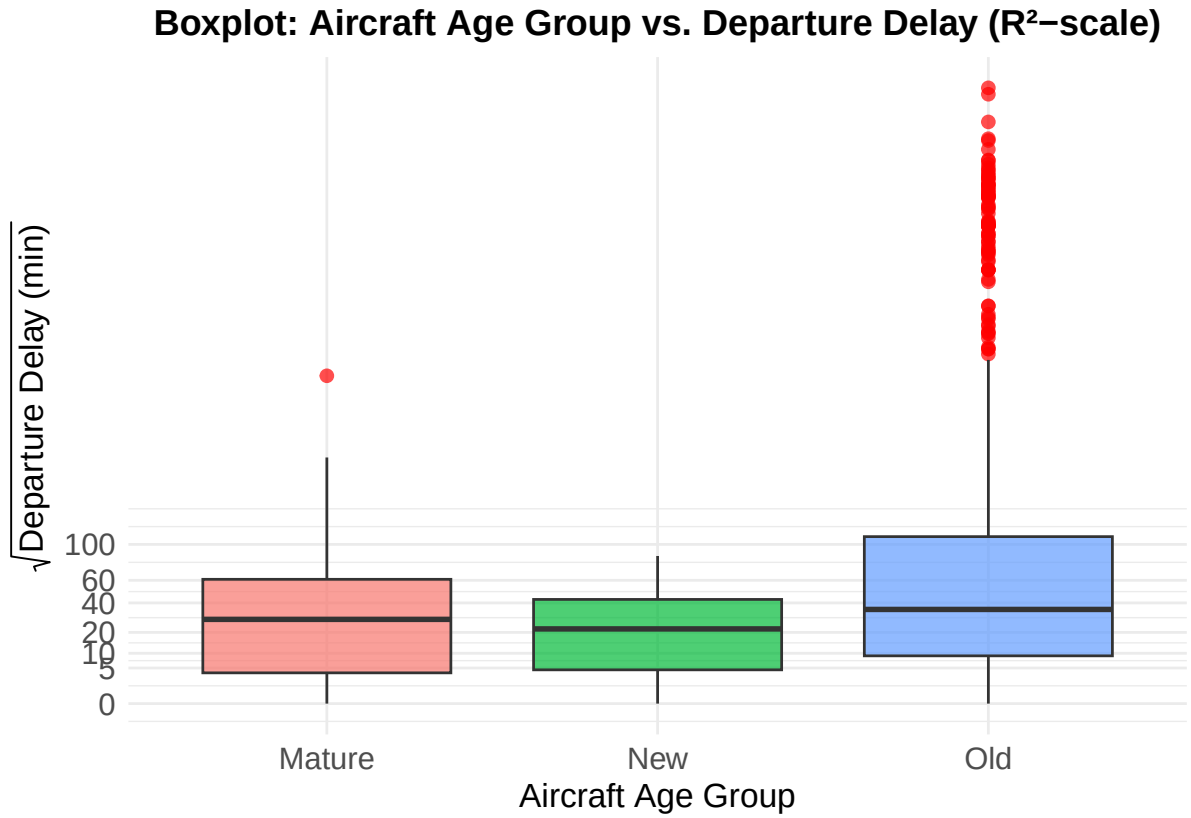
Why Use a Logarithmic Scale?

- Logarithmic scaling is useful when data contains extreme outliers that skew the visualization.
- It helps to highlight the underlying structure of the data by reducing the visual impact of extreme values.

Conclustion The log-scaled analysis confirms that **aircraft age strongly correlates with departure delays**, with older aircraft experiencing **more frequent and severe delays**.

7.3 R^2-Scaled Boxplot

The third plot applies a **R^2-scale** to the delay data. This transformation is applied to **reduce the impact of extreme values and better visualize the distribution of delays** across different aircraft age groups.



Key Observations

- **Old aircraft exhibit the highest median departure delays**, with a **greater interquartile range (IQR) and a higher number of extreme outliers**. The numerous red outlier points suggest that **old aircraft are more prone to severe delays**.
- **Mature aircraft display a moderate median delay**, with **a relatively balanced distribution and fewer extreme values**.
- **New aircraft have the lowest median delays** and the **smallest variation**, indicating **higher reliability and fewer operational disruptions**.

Potential Causes

- **Aging Aircraft and Maintenance Requirements**: Older aircraft may require **longer turnaround times** due to **frequent maintenance and repairs**.
- **Operational Scheduling**: Airlines may **assign newer aircraft to high-priority routes**, minimizing delays.
- **Regulatory Factors**: Stricter **inspection and compliance requirements for older aircraft** may lead to **longer pre-departure processes**.

Conclusion

The **square root transformation effectively normalizes extreme delay values**, providing a clearer comparison between aircraft age groups. The findings indicate that **older aircraft are more susceptible to prolonged delays**, reinforcing the need for **preventive maintenance strategies, optimized fleet scheduling and long-term fleet renewal initiatives** to improve operational efficiency and reduce disruptions.

---

8 Fitting a Model to Predict Departure Delays

To understand the factors contributing to departure delays, a **linear regression model** is fitted using a variety of predictor variables. The initial model includes all potential explanatory variables and a **stepwise regression** is applied to optimize the model by selecting the most significant predictors. This method ensures that the model remains interpretable while retaining only the variables that contribute meaningfully to explaining the variability in departure delays.

8.1 Methodology

1. **Variables:**
   The initial model includes the following predictors:

- `Delay_Carrier`
- `Delay_Weather`
- `Delay_NAS`
- `Delay_Security`
- `Delay_LastAircraft`
- `Aicraft_age`
- `wspd`
- `tavg`
- `snow`

2. **Stepwise Regression:**
   Using **stepwise selection** via the `stepAIC` function, the model iteratively removes or includes predictors based on their contribution to minimizing the Akaike Information Criterion (AIC). This approach balances model complexity and predictive accuracy.

3. **Optimized Model:**
   The optimized model retains only the predictors that significantly contribute to explaining departure delays, as determined by statistical significance and AIC.

```
##
## Call:
## lm(formula = as.numeric(Dep_Delay) ~ Delay_Carrier + Delay_Weather +
##     Delay_NAS + Delay_LastAircraft + Aicraft_age + snow, data = top_flight_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -164.791   -5.132    1.158    8.935  261.469
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          9.233648   4.446094   2.077 0.038039 *
## Delay_Carrier        0.999270   0.002754 362.888  < 2e-16 ***
## Delay_Weather        0.984375   0.005184 189.878  < 2e-16 ***
## Delay_NAS            0.922499   0.011653  79.166  < 2e-16 ***
## Delay_LastAircraft   0.999021   0.004189 238.508  < 2e-16 ***
## Aicraft_age         -0.320923   0.212596  -1.510 0.131431
## snow                 0.021839   0.006049   3.611 0.000319 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.05 on 1165 degrees of freedom
## Multiple R-squared:  0.9943, Adjusted R-squared:  0.9942
## F-statistic: 3.368e+04 on 6 and 1165 DF,  p-value: < 2.2e-16
```

8.2 Interpretation of the Linear Regression Model Results

8.2.1 1. Model Overview

The linear regression model predicts departure delays (`Dep_Delay`) based on multiple contributing factors. The predictors included carrier delays, weather delays, NAS delays, previous aircraft delays, aircraft age and snowfall.

8.2.2 2. Key Coefficients and Statistical Significance (Based on p-values):

- **Intercept:** Estimate = 9.2336, p-value = 0.0380 (*). This represents the baseline delay when all predictors are zero. However, this makes no sense.
- **Carrier Delay (`Delay_Carrier`):** Estimate = 0.9993, p-value < 2e-16 (***). A 1-minute carrier delay increases departure delay by approximately 0.999 minutes, it shows a positive trend, which means the more Carrier Delay the more Departure Delay.

- **Weather Delay (`Delay_Weather`):** Estimate = 0.9844, p-value < 2e-16 (***). A 1-minute weather delay increases departure delay by approximately 0.984 minutes, it shows a positive trend, which means the more Weather Delay the more Departure Delay.
- **NAS Delay (`Delay_NAS`):** Estimate = 0.9225, p-value < 2e-16 (***). A 1-minute NAS delay increases departure delay by approximately 0.922 minutes, it shows a positive trend, which means the more NAS Delay the more Departure Delay.

- **Previous Aircraft Delay (`Delay_LastAircraft`):** Estimate = 0.9990, p-value < 2e-16 (***). A 1-minute delay from the previous aircraft increases departure delay by approximately 0.999 minutes. it shows a positive trend, which means the more Previous Aircraft Delay the more Departure Delay.

- **Aircraft Age (`Aicraft_age`):** Estimate = -0.3209, p-value = 0.1314 (not significant). Aircraft age does not have a statistically significant impact on departure delays.

- **Snow (`snow`):** Estimate = 0.0218, p-value = 0.0003 (***). A unit increase in snowfall increases departure delay by approximately 0.0218 minutes.

8.2.3 3. Model Performance Metrics:

- **Residual Standard Error:** 18.05 minutes, representing the average deviation between predicted and actual delays.

- **Multiple R-squared:** 0.9943, meaning that 99.43% of the variance in departure delays is explained by the predictors.

- **Adjusted R-squared:** 0.9942, which adjusts for the number of predictors and still indicates an excellent fit.

- **F-statistic:** 3.368e+04 with a p-value < 2.2e-16, indicating that the overall model is highly statistically significant.

8.2.4 4. Interpretation of Statistical Significance (Based on p-values):

- **Highly Significant Predictors ($p < 0.001$):** Carrier delay, weather delay, NAS delay, previous aircraft delay and snowfall. These predictors strongly influence departure delays.

- **Not Significant Predictors ($p > 0.05$):** Aircraft age. This factor does not have a meaningful or statistically significant impact on departure delays.

8.2.5 5. Key Insights and Conclusions:

- The model has an excellent fit with an R-squared value of 0.9943, indicating that nearly all variation in departure delays is explained by the predictors.

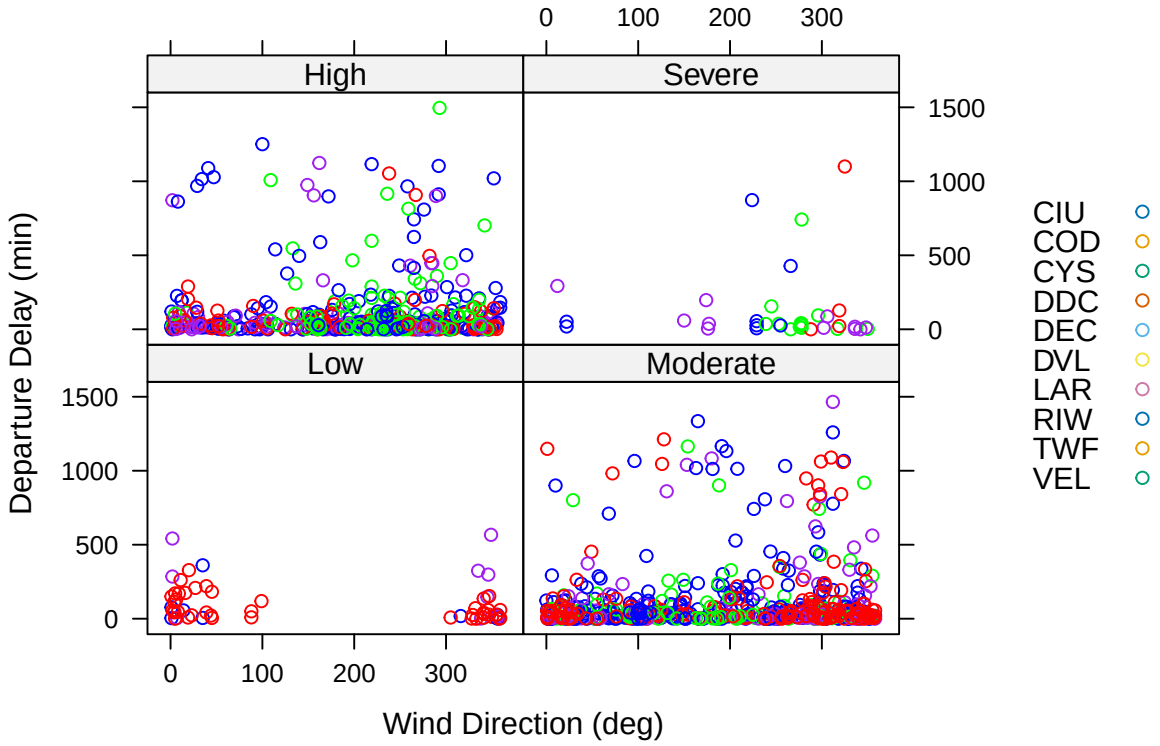
- Carrier-related, weather-related, NAS and previous aircraft delays are the most critical factors, each having an almost one-to-one relationship with departure delays.
- Snowfall has a statistically significant but relatively small effect on delays.
- Aircraft age does not significantly affect departure delays.
- The high F-statistic and extremely low p-value for the model confirm that the overall model is statistically significant and reliable.

---

9 Chapter of Choice: Weather Impact on Departure Delays with Lattice

In this section, we explore the relationship between wind direction their impact on departure delays. These visualizations aim to uncover patterns by categorizing data based on wind conditions (e.g.,Low, Moderate, High, Severe) and plotting multiple scatterplots. For the following plots, we use the **lattice package**.

Lattice is a powerful R package that is particularly suitable for creating multiple plots, facet-based visualizations and comparative plots. In this analysis, a multi-faceted analysis of delays as a function of airport and wind direction is performed to systematically investigate the influence of wind direction on delays at individual airports. An integrated facet visualization is created, which makes it possible to present these interactions clearly in a single diagram.



9.1 Description of the plot

The **lattice scatter plot matrix** visualizes the **impact of wind direction and speed on departure delays** across four categories: **Low, Moderate, High and Severe**. The **X-axis represents wind direction (0°–360°), while the Y-axis shows departure delays (minutes)**. Each point corresponds

to a recorded delay at one of the **worst 10 most delayed airports**, color-coded accordingly.

9.1.1 Key Findings

- **Stronger winds lead to greater delays**, with **High and Severe wind categories** showing **frequent extreme delays** exceeding **1,500 minutes**.
- **Lower wind speeds (Low, Moderate) correspond to fewer and shorter delays**, suggesting **limited operational disruptions** under moderate conditions.
- **Delays cluster at specific wind directions** (e.g., **0°–100° and 250°–360°**), indicating **possible runway alignment or turbulence effects**.
- **Airport-specific variability is evident**, highlighting the role of **geographic and infrastructural factors in wind-related disruptions**.

9.1.2 Lattice Plot Advantages

- **Faceted comparison** enables clear differentiation between wind speed categories.
- **Color-coded airport markers** enhance interpretability without data overlap.
- **Multi-panel design** effectively isolates wind-related delay trends.

9.1.3 Possible Causes

- **Crosswinds/Tailwinds:** Increased **takeoff/landing challenges** and **rerouting needs**.
- **Air Traffic Congestion:** Strong winds may **necessitate extended departure intervals**.
- **Geographic Influence:** Airports in **coastal or mountainous regions** face **higher wind-induced delays**.

9.1.4 Conclusion

The **lattice visualization effectively illustrates** that **strong winds significantly impact departure delays**, necessitating **adaptive scheduling, improved forecasting and infrastructure enhancements** at affected airports.

---

10 Gen AI

Using generative AI in our tasks proved invaluable, particularly when addressing issues with distorted graphics or unexpected results in our implementations. It provided clarity when error messages were unclear and helped us interpret outputs that were initially difficult to understand. Additionally, generative AI assisted in refining our formulations and improving the clarity of our explanations. While its suggestions often led to solutions, we consistently verified the results directly in R to ensure their accuracy. Overall, it greatly enhanced our efficiency in troubleshooting and resolving challenges, especially when working on visualizations and data outputs.

---

11 Conclusion

This project was a challenging yet rewarding experience. Although we invested a significant amount of time in data cleaning, analysis and visualization, the process was both engaging and educational. We gained valuable insights into the factors influencing departure delays, from weather conditions and wind directions to aircraft age and seasonal trends. The hands-on approach not only deepened our

understanding of the concepts but also enhanced our technical skills, particularly in R programming and data visualization.

Our hypotheses provided meaningful insights: - **Worst Airports (Hypothesis 1 & 2):** Airports like **DEC (Decatur)** and **RIW (Riverton)** showed the highest average delays, with peak delays occurring during afternoon and evening periods due to congestion. - **Weather Impact (Hypotheses 3 & 4):** We observed that **moderate winds (20–30 km/h)** combined with heavy precipitation caused the longest delays and that **South and Southwest winds** led to more severe disruptions, likely due to runway orientations. - **Seasonality (Hypothesis 5):** December recorded the highest delays, followed by July and August, highlighting the impact of winter disruptions and summer travel peaks. - **Aircraft Age (Hypothesis 6):** Although older aircraft showed more outliers, the overall difference in median delays was smaller than expected, indicating that aircraft age was not a significant factor.

These results emphasize that **weather conditions, seasonal patterns and airport-specific congestion** are the primary drivers of departure delays, while **aircraft age plays a minor role**. This analysis offers a solid foundation for recommendations such as **improving weather management protocols, adjusting flight schedules during peak seasons and addressing bottlenecks at high-delay airports**.

Despite the effort required, we genuinely enjoyed working on this project. The combination of problem-solving, collaboration and creativity made it a fulfilling journey. We can confidently say that we learned a great deal and developed a stronger appreciation for data-driven decision-making.

We highly recommend this module to anyone interested in exploring real-world data challenges. It is both a learning opportunity and a chance to apply analytical skills to meaningful problems, making it a highly enriching experience.

---

12 References

- Boonpan, S., & Sarakorn, W. (2025). Deep neural network model enhanced with data preparation for the directional predictability of multi-stock returns. *Journal of Open Innovation: Technology, Market and Complexity, 11*(1), 100438. https://doi.org/10.1016/j.joitmc.2024.100438
- Carvalho, L., Sternberg, A., Maia Gonçalves, L., Beatriz Cruz, A., Soares, J. A., Brandão, D., Carvalho, D., & Ogasawara, E. (2021). On the relevance of data science for flight delay research: A systematic review. *Transport Reviews, 41*(4), 499–528. https://doi.org/10.1080/01441647.2020.1861123
- Chandra, A., & Verma, A. (2025). To delay or not to delay? A hybrid relationship between departure delay, en-route conflict probability and number of conflicts. Journal of the Air Transport Research Society, 4, 100053. https://doi.org/10.1016/j.jatrs.2024.100053
- Grainger, S., Mao, F., & Buytaert, W. (2016). Environmental data visualisation for non-scientific contexts: Literature review and design framework. *Environmental Modelling & Software, 85*, 299–318. https://doi.org/10.1016/j.envsoft.2016.09.004
- Gu, Z. (2022). Complex heatmap visualization. *iMeta, 1*(3), e43. https://doi.org/10.1002/imt2.43