[1]

A stormy flight: Investigating the correlation between weather patterns and delays on the JFK-LAX route

Project Group - 2

Members: Nynke Leppink, Alexander du Marchie Sarvaas, Samuel du Marchie Sarvaas, Sanne van Proosdij, Emma Toet

Student numbers: 5070988, 5547156, 4802829, 5326990, 4849892

Contribution Statement

General contribution: Together we contributed to the structure of the report, determining the research question and sub questions and reviewing eachothers work, cleaning of the code and organising

- Sanne: Coding and data analysis weather data and correlation
- Samuel: Coding and data analysis flight data and sub question 5
- Nynke: Coding data analysis weather data and correlation
- Emma: Coding data analysis flight data and introduction
- Alexander: Coding data analysis flight data, conclusion

Introduction

Flight delays are a common source of irritation for many travellers. Factors such as weather and technical issues frequently cause these disruptions. Additionally, the growing impact of climate change is leading to more extreme weather patterns, which could further contribute to disruptions in air travel. For frequent flyers, dealing with delays has become an unfortunate

norm. This report will explore flight delays on the route between John F. Kennedy International Airport (JFK) and Los Angeles International Airport (LAX), which is one of the busiest air routes in the world, with over 150 flights per week. LAX ranks among the top ten busiest airports globally [2]. The report will examine how weather conditions at JFK and LAX impact flight delays on this route.

Research objective

For this research the following question will be answered: ***"How do weather conditions at JFK (New York) and LAX (Los Angeles) affect flight delays on this flight route?"***

To answer this question a dataset of flight information from the year 2013 will be used [3]. This dataset contains flight information of different flight routes in the United States. This research focuses on the flight route between John F. Kennedy International Airport (JFK) in New York and Los Angeles International Airport (LAX). The information in this dataset can be used to analyse the departure and arrival delays of these airports. For the weather conditions at JFK and LAX the weather information of Iowa Environmental Mesonet will be used [4]. On this website different weather variables are available and will be combined in one dataset.

Sub questions

1. Which specific weather conditions most significantly impact flight delays on the JFK-LAX route?
2. Which specific flight and weather data is relevant for our research? (actual delay etc).
3. What is the relationship between the chosen (severe) weather patterns and the duration and occurence of flight delays? (arrival)
4. What is the relationship between the chosen (severe) weather patterns and the duration and occurence of flight delays? (departure)
5. How can we combine the different correlations of the weather data and look at the total effect on flight delays?

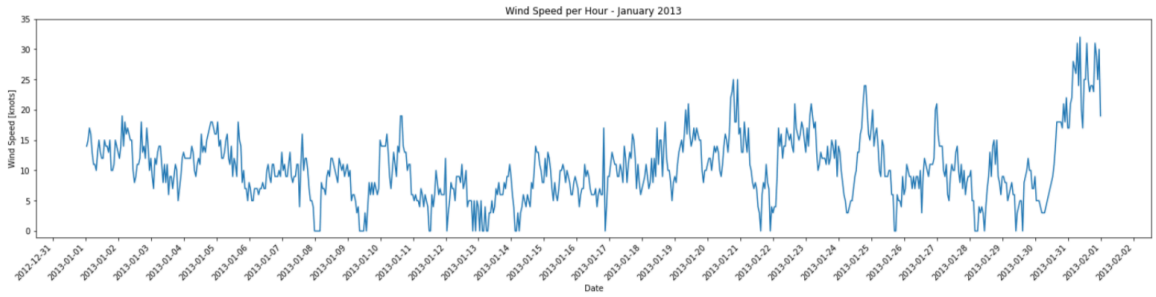
Conditions and restrains

In this study, there are several constraints that influence the analysis and findings. The weather conditions include a variety of elements that are important to understand flight operations. The study will consider factors such as visibility, wind speed and wind gust, sky level coverage, precipitation and pressure altimeter.

The to-be-used datasat has five airlines operating this route: **United Airlines Inc., Virgin America, JetBlue Airways, American Airlines Inc., and Delta Air Lines Inc.** By investigating the performance of these airlines, the aim is to identify trends and correlations between weather conditions and flight operations. This will enable a better understanding of how various factors impact delays throughout the year. The temporal scope of the study covers the year 2013, which provides a big dataset for analysis. The dataset (consisting of more than 11.000 flights) has flights at an hourly level, monitoring approximately 30 flights each day. Geographically, the analysis will concentrate on the route between two major international airports: John F. Kennedy International Airport (JFK) in New York and Los Angeles International Airport (LAX). By focusing on these key locations, the weather conditions specific to each airport will be explored.

Sub Question 1

In this analysis, we focused on key weather parameters—wind gust, wind speed, precipitation, visibility, and sky coverage—at both JFK and LAX airports. After cleaning the dataset, handling missing values, and converting sky coverage into six distinct categories, we visualised the behaviour of each variable for all months. All the graphs can be found in the submap of this subquestion. An example of the weather data is shown in the figure below: the wind speed per hour at JFK in january.



The analysis of weather patterns reveals noticeable differences between the two airports, reflecting the distinct climates of Los Angeles and New York. At LAX, the weather data showed minimal variability in visibility and precipitation, which remained consistently clear. However, an extreme outlier for wind speed (800 km/h) was identified, likely due to a data error. In contrast, JFK showed more frequent wind gusts and slightly more cloud cover, which can indicate more turbulent weather conditions compared to LAX.

These differences in weather patterns between the two cities will now be used to assess their relationship with flight delays. By comparing weather conditions to delay data, we aim to uncover how the weather conditions contribute to delays at each airport, which will inform strategies to mitigate these impacts on flight schedules.

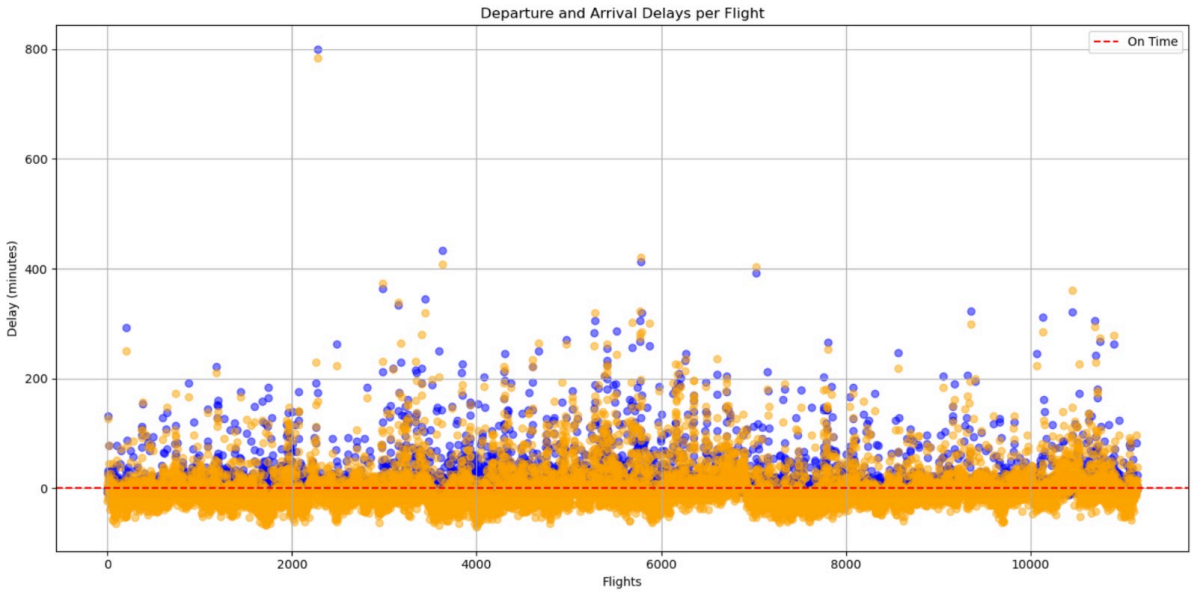
Sub Question 2

The flight data for this analysis comes from two datasets: 'arrivals.csv' and 'departures.csv', which contain detailed records of flights within the United States from January 1, 2013, to September 30, 2013. The original dataset consists of around 320,000 flights. To focus on the specific research question—examining flight delays at New York's John F. Kennedy International Airport (JFK) and Los Angeles International Airport (LAX)—we filtered the data to include only flights between these two airports. This resulted in a more manageable dataset of around 11,000 flights, divided into two sets: arrivals and departures. After filtering for flights specific to JFK and LAX, the relevant data was exported to a new CSV file that would serve as the primary source for the analysis. The dataset includes several key columns necessary for understanding the timing and nature of flight delays:

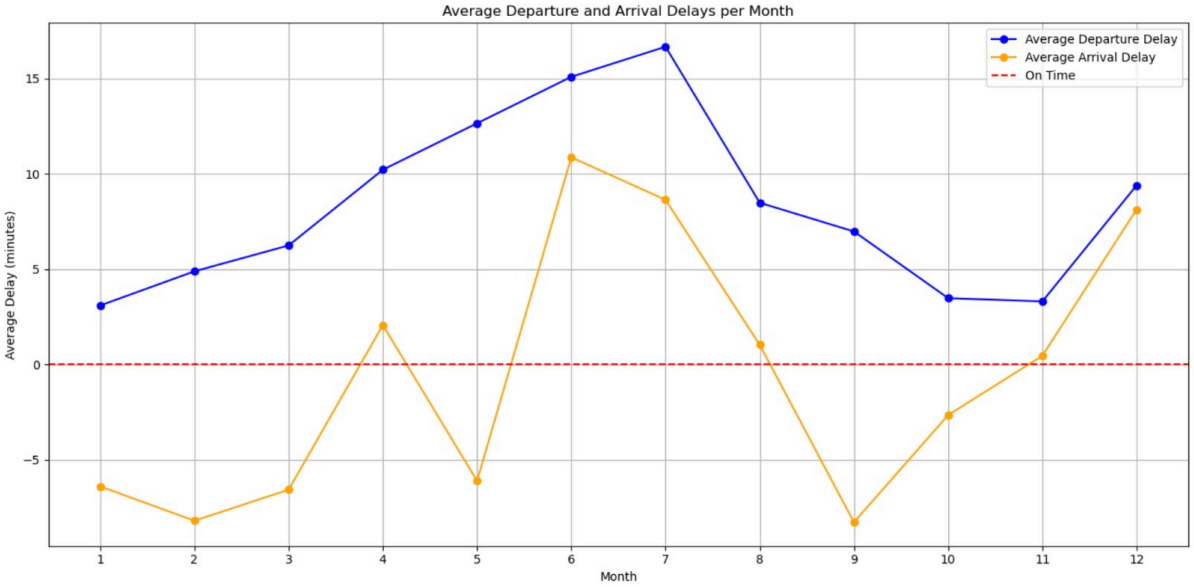
- date: The flight date.
- dep_time: Actual departure time.
- sched_dep_time: Scheduled departure time.
- sched_dep_time (modified): This column was created as an additional step to link departure times more accurately with weather events.
- dep_delay: The departure delay in minutes.

- arr_time: Actual arrival time.
- sched_arr_time: Scheduled arrival time.
- sched_arr_time (modified): This column, like the departure time, was modified to better align with weather event timings.
- arr_delay: The arrival delay in minutes.
- tailnum: The unique identifier of the aircraft.
- air_time: The duration of the flight in minutes.
- name: The name of the airline.

For the actual analysis, some columns were excluded to streamline the focus on the most relevant data points, such as delay times, departure and arrival information, and weather-related modifications. These cleaned and tailored datasets enabled a more precise investigation of how weather conditions impact flight delays at these two major airports.



This scatter plot compares departure and arrival delays for flights between JFK and LAX. Blue dots represent departure delays and orange dots show arrival delays, with the red dashed line at 0 marking on-time flights. Dots above the line indicate delays, while those below would show early departures or arrivals. The plot helps visualise delay patterns and distributions, highlighting any clusters of significant delays or areas where delays are frequent. Alpha transparency ensures overlapping points remain visible for clearer insights.



This chart visualises the average departure and arrival delays per month for flights between JFK and LAX. It first extracts the month and year from the flight dates, then calculates the average delays for each month. The results are plotted as a line chart, with blue representing average departure delays and orange representing average arrival delays. A red dashed line at 0 minutes serves as a reference for on-time flights. The x-axis shows the months (1 to 12), and the chart highlights how delays fluctuate over the year.

Sub Question 3 and 4

This chapter addresses the third and fourth sub-questions: *What is the relationship between the chosen (severe) weather patterns and the duration and occurence of flight delays? (arrival)* and *What is the relationship between the chosen (severe) weather patterns and the duration and occurence of flight delays? (departure)*. This sub-question is answered by plotting the correlation between flight delays and individual weather variables to identify any patterns. These graphs are then analyzed by examining the data points and plotting a regression line to identify trends. This analysis is conducted for departure delays at JFK Airport in New York using weather data from the same airport, and for arrival delays at LAX Airport in Los Angeles using corresponding weather data from LAX.

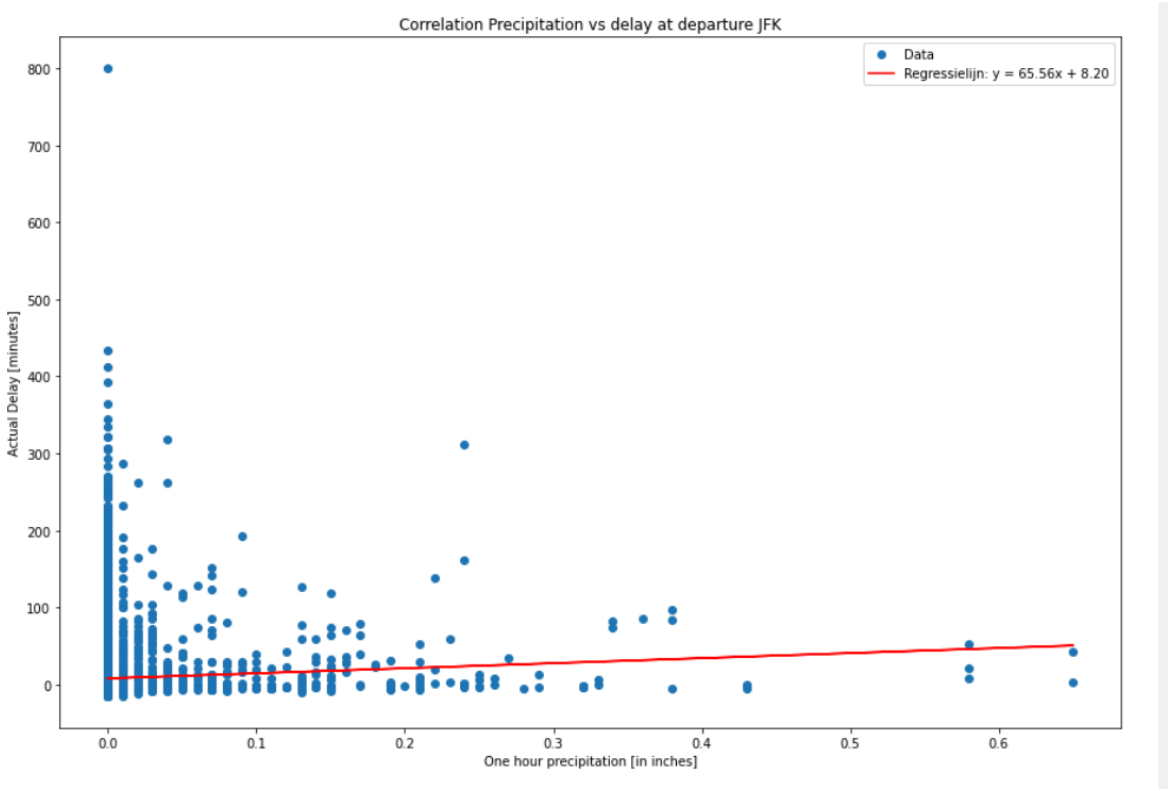
First, the datasets are merged, linking weather variables to the respective flights. The departure and arrival times of the flights are adjusted to align with the hourly intervals of the weather data. This allows the correct weather variables for a specific hour to be matched with each flight. The merged datasets are represented in the CSV files *Arrival_weather_combined.csv* and *Departures_weather_combined.csv*.

Additionally, based on the insights from sub-question 1, the arrival delay has been adjusted to reflect the actual arrival delay, by subtracting the departure delay from the initial reported arrival delay.

This report examines the relationship between severe weather patterns and the duration and occurrence of flight delays. Notably, flights without delays are not excluded from the analysis. This is important because some flights may experience severe weather but still arrive or departure on time, and excluding them could cause biassed results. By including all flights,

regardless of delay status, the analysis aims to determine whether a correlation exists between weather conditions and delays.

For example, as shown in the figure titled "Correlation between Precipitation and Departure Delays at JFK," extreme precipitation does not consistently result in significant delays. Several other plots display similar relationships, with the y-axis representing flight delays and the x-axis showing various weather conditions.



Despite plotting these relationships, the figures indicate no clear correlation between weather patterns and flight delays. A regression line has been included in the plots to highlight any potential trends, but it confirms the lack of a strong relationship. As a result, the next step in the analysis will explore whether combining different weather variables might reveal any trends, which could provide more insight into how weather impacts flight delays.

Sub Question 5

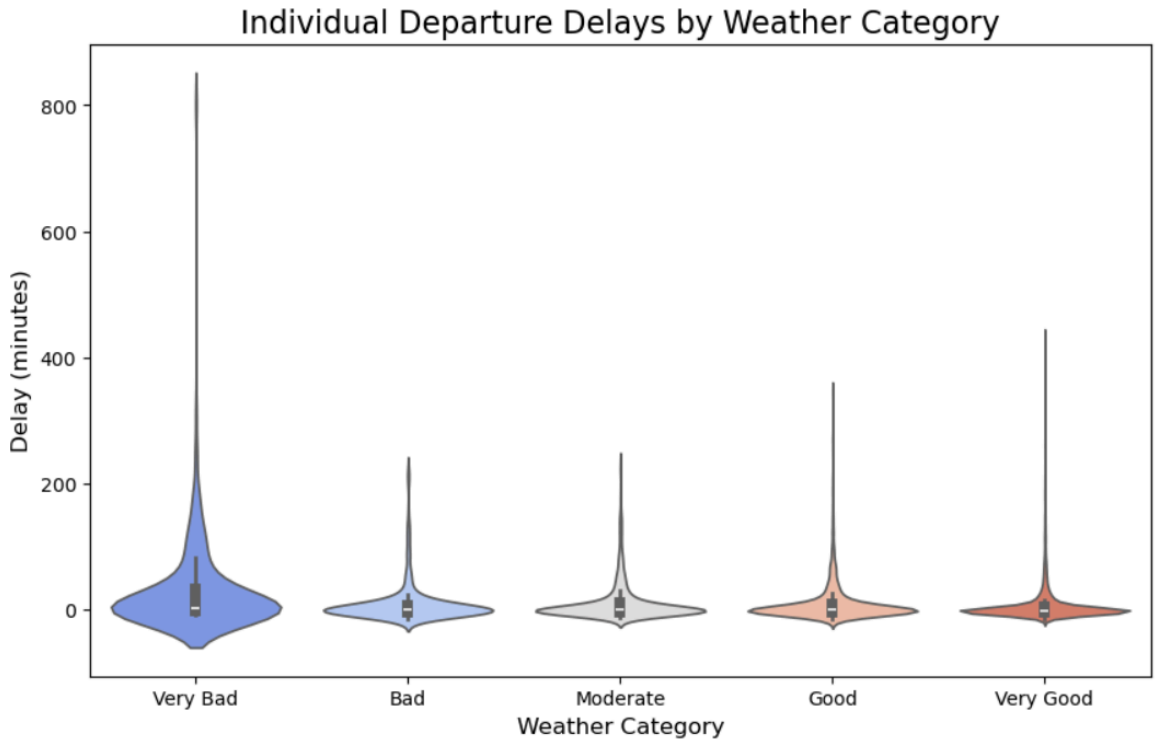
Subquestion 5 explores the effect of combined weather variables on the flight delays. First, the available weather variables have been normalized and each variable has been assigned a weight. This weight is then used to define a score which defines its category. There have been 5 weather categories defined, ranging from bad to good weather.

The normalisation is done using min-max normalisation. As every variable is in a different unit, the normalisation gives each of them a score between 0 and 1. The weight assignment is done using information based on literary research. (Álvaro Rodríguez-Sanz et al, 2021) state that wind conditions have the highest impact on flight delays than others. Furthermore, visibility is also noted as an important factor for flight delays, especially at arrivals. Using this information, the following weights have been assigned to the weather variables:

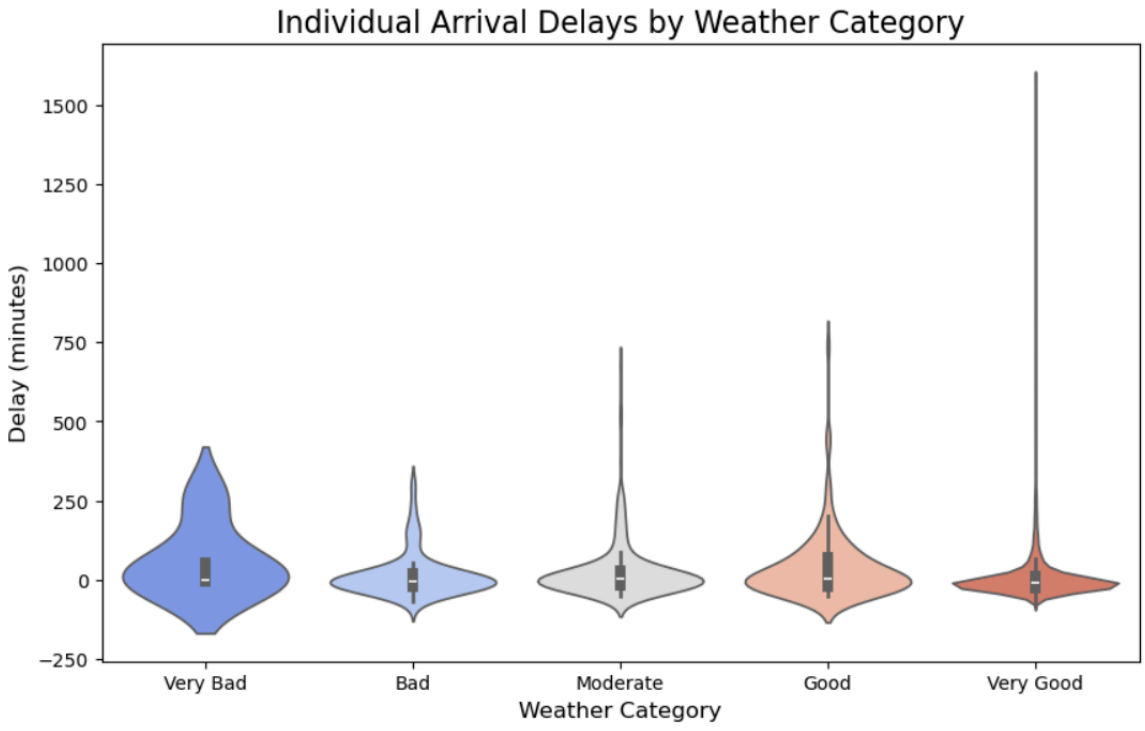
Wind speed: 0.3, wind gust: 0.3, pressure altimeter: 0.1, visibility: 0.2, sky coverage: 0.1

This will give each individual flight data point a weather score between 0 and 1, whereafter a weather category will be assigned.

Next, it is possible to plot all flights in each category into violin plots, as seen in the following images:



This has also been done for the arrivals.



Furthermore, the average delays per weather category have been calculated. For the departure delays these came out to be the following:

Very bad: 29.3 minutes, Bad: 9.4 minutes, Moderate: 10.8 minutes, Good: 9.9 minutes, Very Good: 7.28 minutes

And for the arrival delays the following:

Very bad: 60.2 minutes, Bad: 15.8 minutes, Moderate: 28 minutes, Good: 38.6 minutes, Very Good: 7.13 minutes

As seen from these results, there is an obvious decline in delay time when weather is better.

Conclusion

This analysis aimed to find and examine the relationship between weather conditions and flight delays at JFK and LAX airports, focusing on how distinct weather-related factors at each location could affect flight schedules. Key weather parameters such as wind speed, precipitation, visibility, and sky coverage were analysed and visualised for their potential impact on delays.

For subquestion 1, based on practical data, significant differences in weather patterns between JFK and LAX were found. While LAX experienced stable, clear weather with minimal precipitation, with one extreme outlier for wind speed (800 km/h) (likely due to a data error), JFK showed more variability, including frequent wind gusts and cloud cover. These findings provide a foundational understanding of the weather differences that could affect flight delays at each airport.

For subquestion 2, the flight data analysis focused on delays at JFK and LAX airports, using a filtered dataset from the original 320,000 U.S. flights (original dataset) to around 11,000 flights specific to these two airports. Key columns such as departure and arrival times, delays, and weather-modified schedules were included to better understand the timing and impact of delays. By refining the data and removing unnecessary columns, the analysis was able to more accurately investigate the relationship between weather conditions and flight delays, providing a better foundation for exploring the factors influencing delays at these airports. Next to preparing the data for later subquestions, some initial charts and plots were created for an outline of the data.

Sub questions 3 and 4 further explored correlations between specific weather variables and delays, but again, no clear patterns emerged. Even under severe weather conditions, such as heavy precipitation or high winds, delays did not show a consistent increase. This suggests that factors beyond just weather, such as operational or logistical issues, may be equally influential.

Sub question 5 sought out to find out the total effect of combined weather variables. The model described and combined the weather variables using normalisation weight assignment to give each flight entry a weather score. The weather scores describe how 'good' the weather was at each flight time, relative to each other. What is interesting to see, is that the method used has a preference for good weather, especially for the arrivals data set, which is in Los Angeles. This can be considered obvious, as the weather is generally better on the west coast.

The discrepancies in these categories only confirm the hypothesis. For both the departure and arrivals set, the total number of flights in the "very bad weather" category is much lower than the other categories, but the average delays are much higher.

The methods used for subquestion 5 can be up for discussion. The category assignment, scoring system, variable selection, all are changeable elements. In further research it could be possible to refine these.

Literature study

[1] Transport free icon https://www.flaticon.com/free-icon/transport_15226809?term=airplane&page=1&position=2&origin=search&related_id=15226809

[2] The top ten busiest airports in the world https://www.airport-technology.com/features/the-top-10-busiest-airports-in-the-world/

[3] Wu, Y., Mei, G., & Shao, K. (2022). Revealing influence of meteorological conditions and flight factors on delays Using XGBoost. Journal Of Computational Mathematics And Data Science, 3, 100030. https://doi.org/10.1016/j.jcmds.2022.100030

[4] Kulesa, G. (2003). WEATHER AND AVIATION: HOW DOES WEATHER AFFECT THE SAFETY AND OPERATIONS OF AIRPORTS AND AVIATION, AND HOW DOES FAA WORK TO MANAGE WEATHER-RELATED EFFECTS? http://www.climate.dot.gov/documents/workshop1002/kulesa.pdf

[5] Rodríguez-Sanz, Á., Cano, J., & Fernández, B. R. (2021). Impact of weather conditions on airport arrival delay and throughput. Aircraft Engineering And Aerospace Technology, 94(1), 60–78. https://doi.org/10.1108/aeat-12-2020-0318

[6] Álvaro Rodríguez-Sanz et al (2021). Impact of Weather Conditions on Airport Arrival Delay and Throughput. IOP Conf. Ser.: Mater. Sci. Eng. 1024 012107. https://iopscience.iop.org/article/10.1088/1757-899X/1024/1/012107

Data Used

[3] Iowa State University Iowa Environmental Mesonet https://www.kaggle.com/datasets/mahoora00135/flights?resource=download

[4] Flights: A report to analyze the performance of airlines in 2013 https://mesonet.agron.iastate.edu/request/download.phtml?network=NY_ASOS#

In [ ]: