Exploring Flight Patterns & Delays in NYC

Project Overview

This project analyzes flight delays and patterns at New York City airports using the `nycflights13` dataset. The goal is to explore factors affecting flight delays, airline performance, and seasonal trends.

Datasets Used

We will use `flights` from the `nycflights13` package. It contains over 300,000 observations with details such as departure/arrival times, airline information, and delay times.

Learning Objectives

1. **Understand how real-world datasets are structured** (data types, missing values, distributions).
2. **Apply key dplyr functions** (filtering, grouping, summarizing, ranking, and feature engineering).
3. **Frame data-driven questions and extract insights** relevant to flight delays and performance.

Data Analysis Questions / Hypotheses

🔷 Level 1: Understanding the Dataset (Basic Exploration)

1. What are the column names and data types in the dataset?
2. Are there any missing values? If yes, in which columns?
3. What is the average departure and arrival delay for all flights?

🔷 Level 2: Data Extraction & Filtering

4. Find the top 10 airlines with the most flights.
5. Extract all flights from **JFK** airport that were delayed by more than 2 hours.
6. Identify flights that arrived at **LAX (Los Angeles)** with a delay of more than 60 minutes.

🔷 Level 3: Grouping & Summarization

7. Compute the **average departure delay per airline** to identify which airline has the worst delays.
8. Determine the **busiest month** by counting the number of flights.
9. Find the **airport with the most delays** by grouping data by origin.

🔷 Level 4: Sorting & Ranking Data

10. Rank **airlines based on average arrival delay** (worst to best).
11. Find the top 5 **most delayed flights** (highest dep_delay).
12. Identify the **day of the week** with the highest flight delays.

🔷 Level 5: Feature Engineering

13. Create a new column **"Delay_Category"** based on departure delay:

- **"On Time"** (dep_delay ≤ 0)
- **"Minor Delay"** (1-30 min)
- **"Major Delay"** (31-120 min)
- **"Severe Delay"** (>120 min)

14. Create a new column **"Total Delay"** as the sum of **departure delay + arrival delay**.
15. Compute the **percentage of delayed flights per airline**.

🔷 Level 6: Hypothesis Testing & Advanced Insights

16. Do longer flights (higher distance) tend to have more delays?
17. Are flights scheduled in the **morning (before 12 PM) less delayed than evening flights**?
18. Which months have the worst delays? Is there a seasonal trend?
19. Are delays worse for **certain airlines** compared to others?
20. Does the origin airport significantly affect delay times?

Hypothesis 1: Do longer flights (higher distance) tend to have more delays?

This calculates the **correlation coefficient** between flight distance and departure delay.

- If the value is **close to 1**, it means **longer flights tend to have more delays**.
- If it's **close to 0**, there is **no relationship** between flight distance and delays.
- If it's **negative**, it means longer flights **experience fewer delays**.

Interpretation

If the correlation is **low or negative**, it means that **flight duration does not strongly impact delays**, and other factors (e.g., airport congestion, weather) are more influential.

Hypothesis 2: Are morning flights less delayed than evening flights?

This groups flights into **morning (before 12 PM) and evening (after 12 PM)** and calculates the average departure delay.

Interpretation

- If **morning flights have significantly lower delays**, it suggests that delays **accumulate throughout the day** due to airport congestion and scheduling issues.
- If delays are **evenly distributed**, it means that **time of day does not impact flight delays significantly**.
- If **morning delays are higher**, it might indicate **early morning weather disruptions** or scheduling bottlenecks.

Hypothesis 3: Are delays worse in certain months? Is there a seasonal trend?

This calculates **the average delay per month** to identify seasonal trends.

Interpretation

- If **delays peak in winter**, it suggests that **snowstorms and bad weather** impact flights.
- If **summer has higher delays**, it may be due to **increased air traffic** and vacation travelers.
- A **steady pattern across all months** means seasonality **does not significantly affect delays**.

Hypothesis 4: Which airline has the worst delays?

This identifies the airline with the **highest average departure delay**.

Interpretation

- If **one airline has significantly higher delays**, it may indicate **poor scheduling, frequent technical issues, or mismanagement**.
- If delays are **evenly distributed**, it suggests that **delays are more dependent on external factors like airport congestion or weather** rather than airline policies.

Hypothesis 5: Does the origin airport impact flight delays?

This ranks airports based on **average departure delay** to see if certain airports experience more delays.

Interpretation

- If **one airport has much higher delays**, it might be due to **high congestion, runway constraints, or weather conditions**.
- If delays are **similar across airports**, then factors like **airline scheduling, staffing, and aircraft readiness** might be more responsible for delays.

Conclusion: Practical Insights from These Analyses

1. **Distance vs. Delay:** If longer flights aren't delayed more, **distance isn't a major factor**, and delays are due to **airport or airline-specific** issues.
2. **Morning vs. Evening Flights:** If evening flights are more delayed, travelers should **prefer morning flights** for on-time travel.
3. **Seasonal Delays:** If winter has worse delays, airlines should **schedule buffer time during bad weather months**.
4. **Worst Airlines for Delays:** If a specific airline is the worst, passengers might **avoid it** when booking flights.
5. **Worst Airports for Delays:** If a particular airport has high delays, travelers should **choose alternative airports when possible**.