Analysis of US Flight Data

Portfolio Project 1

Done By: Felix Liang

About The Project

This project aims to conduct an exploratory data analysis of US flight data to uncover insights about flight delays and route popularity through data visualization.

- Analysis 1 – What is the best time to travel to minimize delays?
- Analysis 2 – Do older planes suffer more delays?
- Analysis 3 – How does the popularity of flight routes change over time?
- Analysis 4 – How do the top diverted flight routes compare to the average diversion rate of all flights?
- Analysis 5 – Do flight delays cause a ripple effect across subsequent flights?

Methodology

- Given the large sample size, the data is stored in a database (PostgreSQL). Initial data cleaning and preprocessing are performed using SQL queries.
- Aggregated tables are created to summarize the data. These tables consolidate information and speeds up the analysis process.
- We then connect to the aggregated tables using Tableau to visualize the data.
- For analysis 5, regression analysis is conducted using python to explore the ripple effect of flight delays.

About the Dataset

The data consisted of flight arrival and departure details for all commercial flights on major carriers within the USA from October 1987 to April 2008. It contains nearly 120 million records and 48 attributes in total.

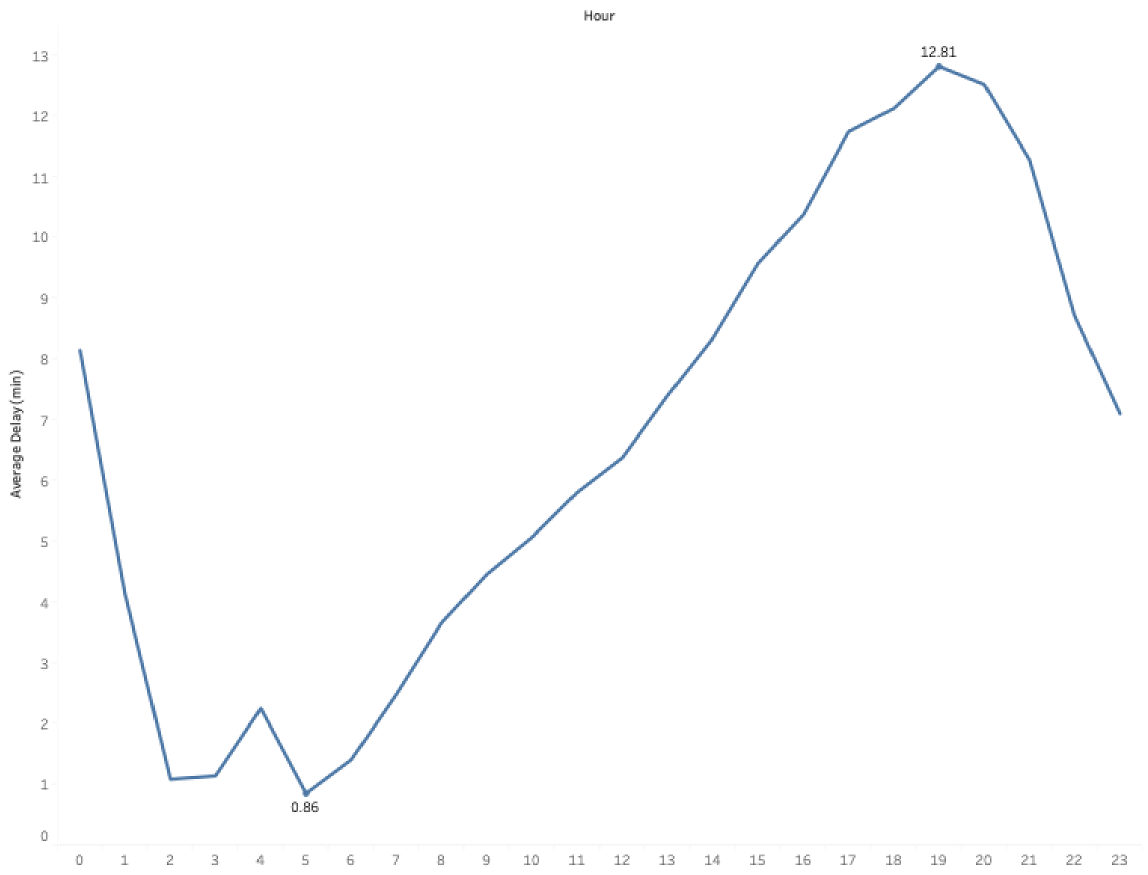
Analysis 1 – What is the best time to travel to minimize delays?

What is the best time to travel to minimize delays?

- This analysis aims to identify the time that exhibits the lowest average delay.
- A flight will be classified as delayed if both its departure and arrival time exceed 15 minutes.
- Both diverted and cancelled flights are not considered.
- Extreme values (above 1440 minutes, equivalent to one day) are excluded since they are not genuine and meaningful and can potentially skew our analysis.

Average Delay Hourly Trend

- The line chart highlights that 5am boasts the lowest average delay (0.86 min).
- Delays gradually increase and peak around 7pm. This pattern aligns with the phenomenon where delays from earlier flights can cascade and affect subsequent flights, leading to increased delays as the day progresses.

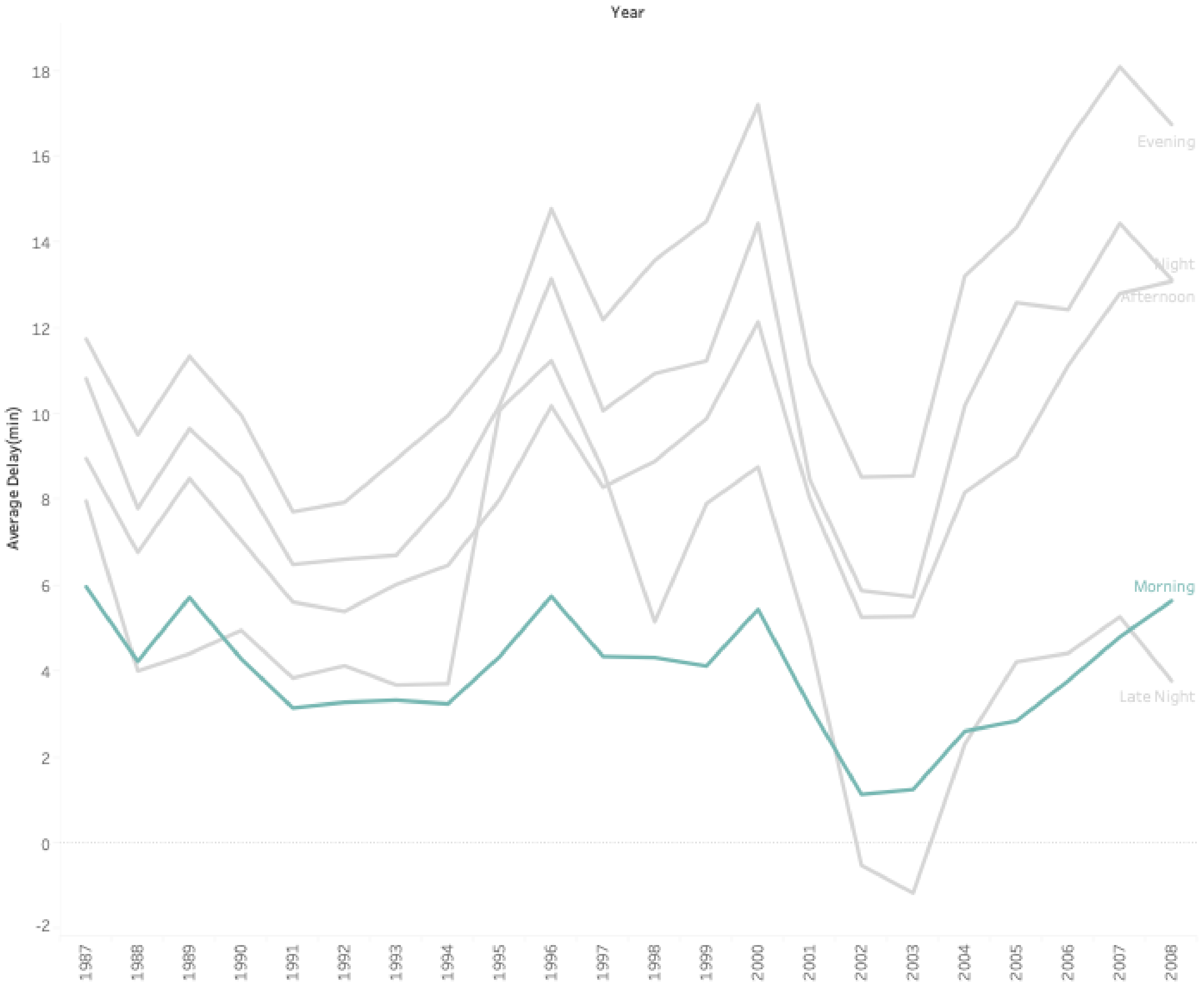


Is flying in the morning consistently the best option over the years?

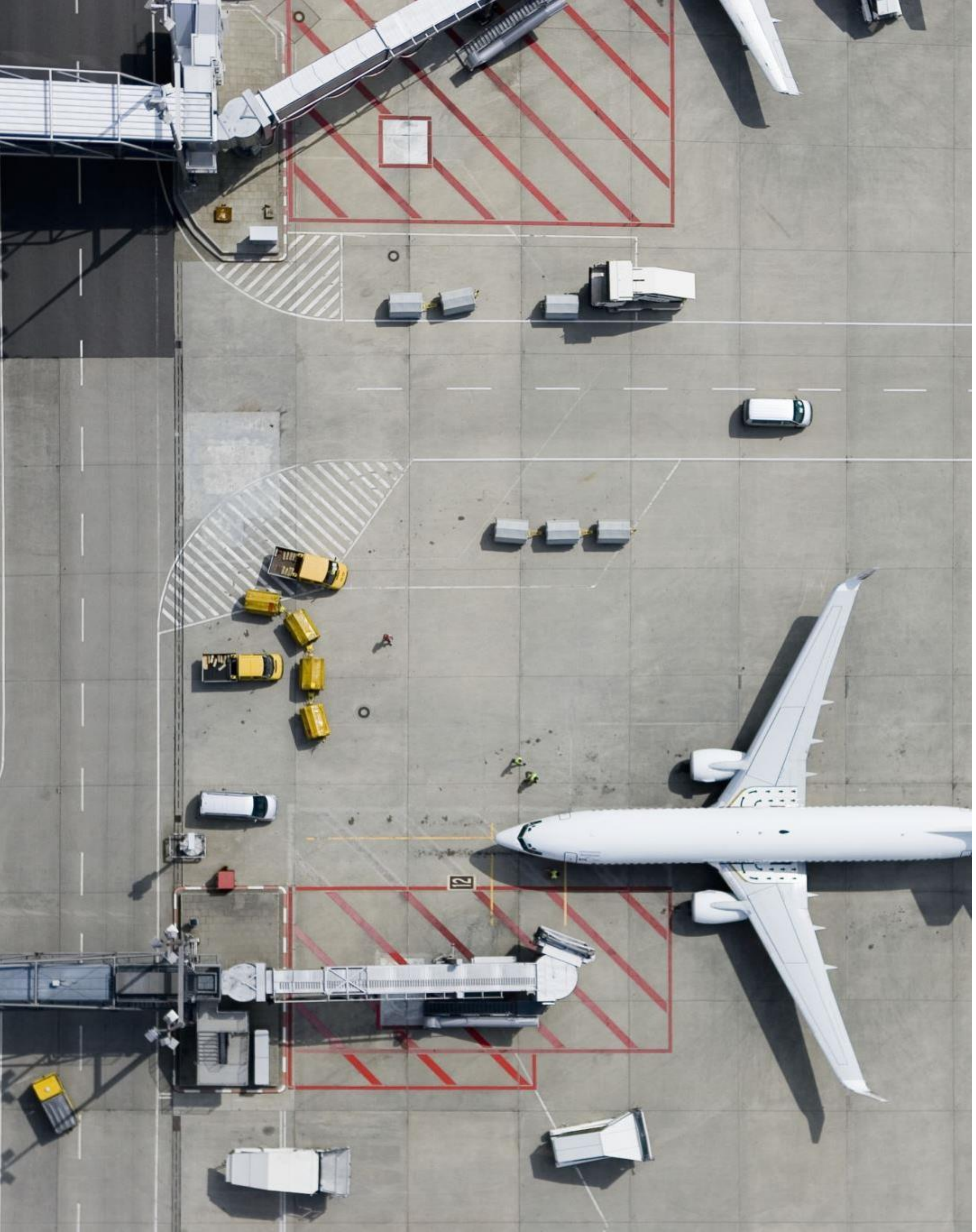
- The schedule departure hour is grouped into 5 intervals for enhanced interpretation.
  - Morning: 5am – 11:59am
  - Afternoon: 12pm – 4:59pm
  - Evening: 5pm – 8:59pm
  - Night: 9pm – 11:59pm
  - Late Night: 12am – 4:59am

Trends in Average Delay Over the Years by Time of Day

- With the exception of 1988 and the period from 2002 to 2004, when ‘Late Night’ had the lowest average delay, ‘Morning’ consistently showed the lowest average delay.



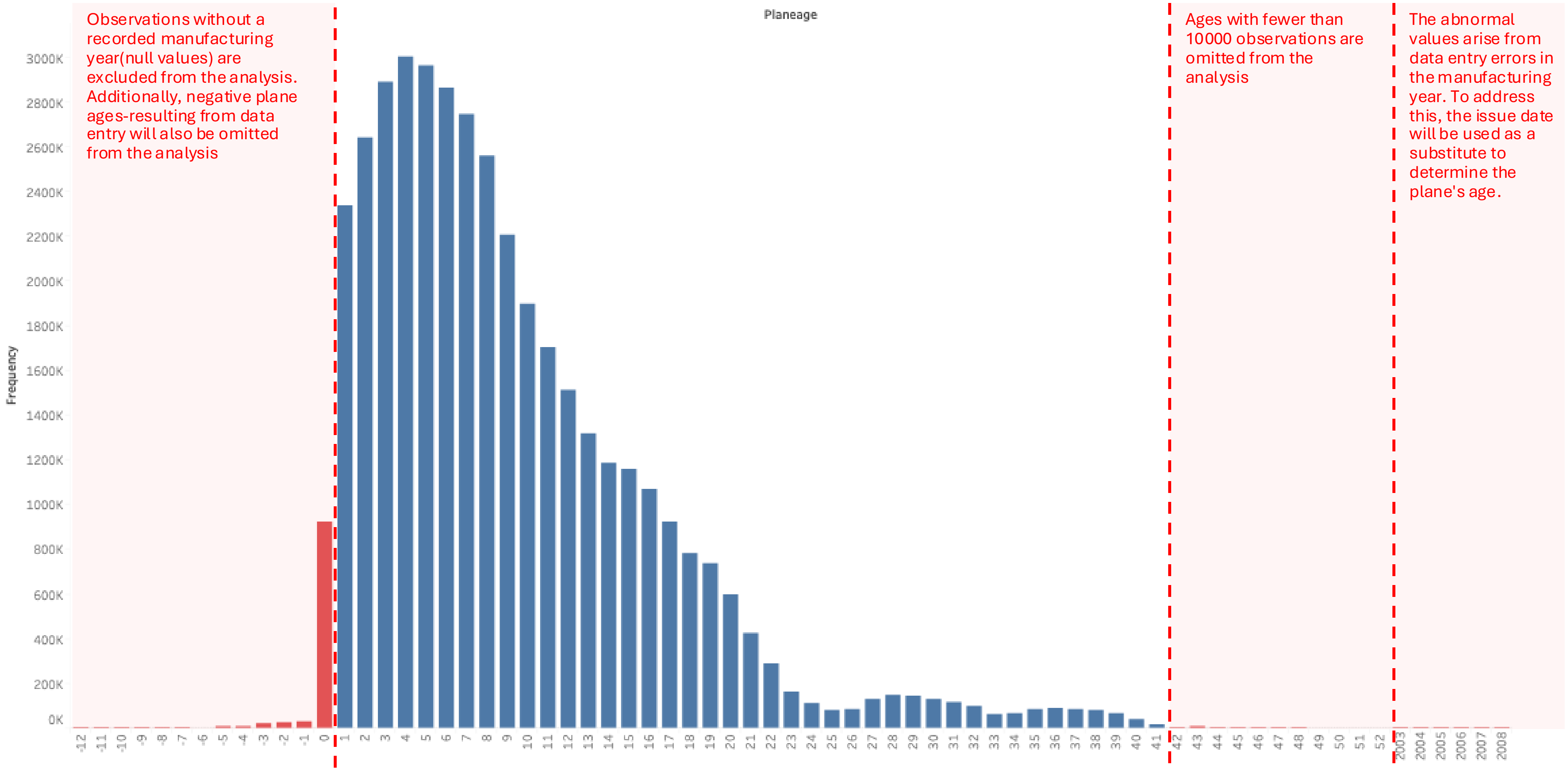
Analysis 2 – Do older planes suffer more delays?



Do older planes suffer more delays?

- To address this question, we will be utilizing a bar plot to visualize the mean delay across different plane ages.
- Plane age is calculated as the difference between the flight year and the plane’s manufacturing or issued year.
- A histogram is also plotted to examine the distribution of plane ages.

Plane-age distribution



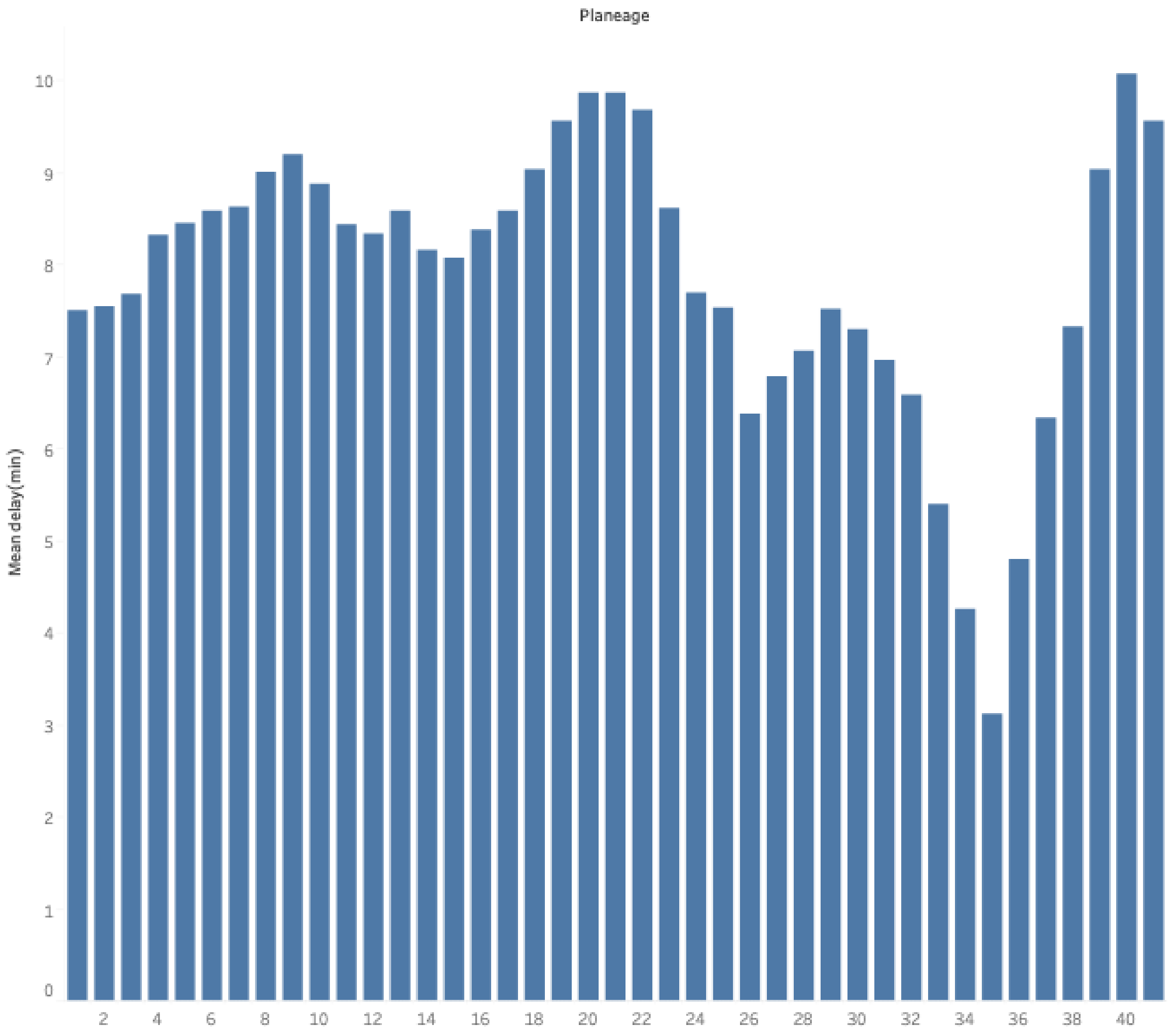
Observations without a recorded manufacturing year(null values) are excluded from the analysis. Additionally, negative plane ages-resulting from data entry will also be omitted from the analysis

Ages with fewer than 10000 observations are omitted from the analysis

The abnormal values arise from data entry errors in the manufacturing year. To address this, the issue date will be used as a substitute to determine the plane's age.

Mean Delay by Plane-age

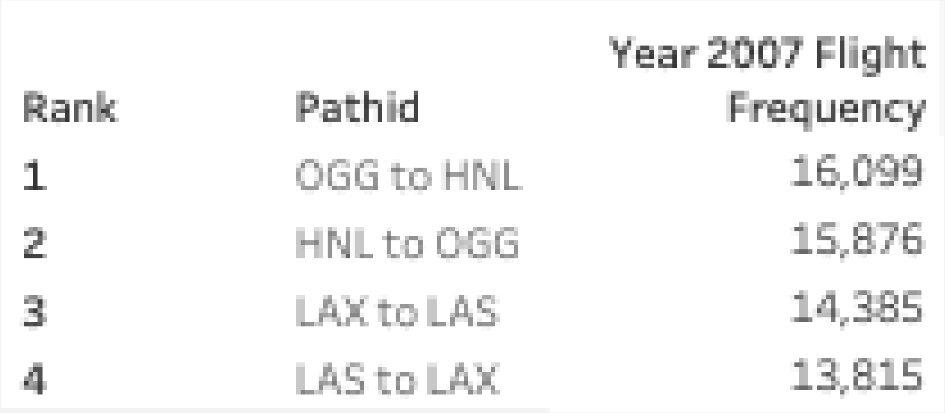
- The bar plot does not reveal any clear patterns indicating that older aircraft result in higher mean delay. Further analysis is required to draw conclusive insights



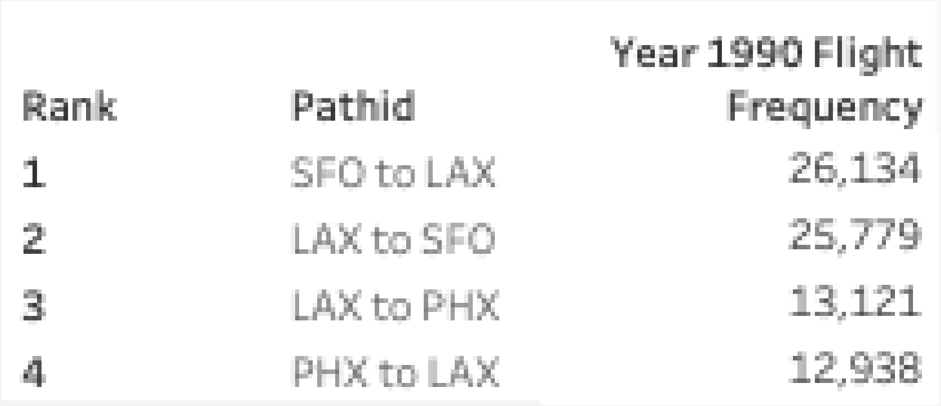
Analysis 3 – How does the popularity of flight routes change over time?

Top 4 flight routes for Year 1990 and 2007

- For this analysis, ‘popularity is defined by the frequency of flights on each route. Tables comparing the most popular routes in 1990 and 2007 show that the top 4 flight routes are different between both years.

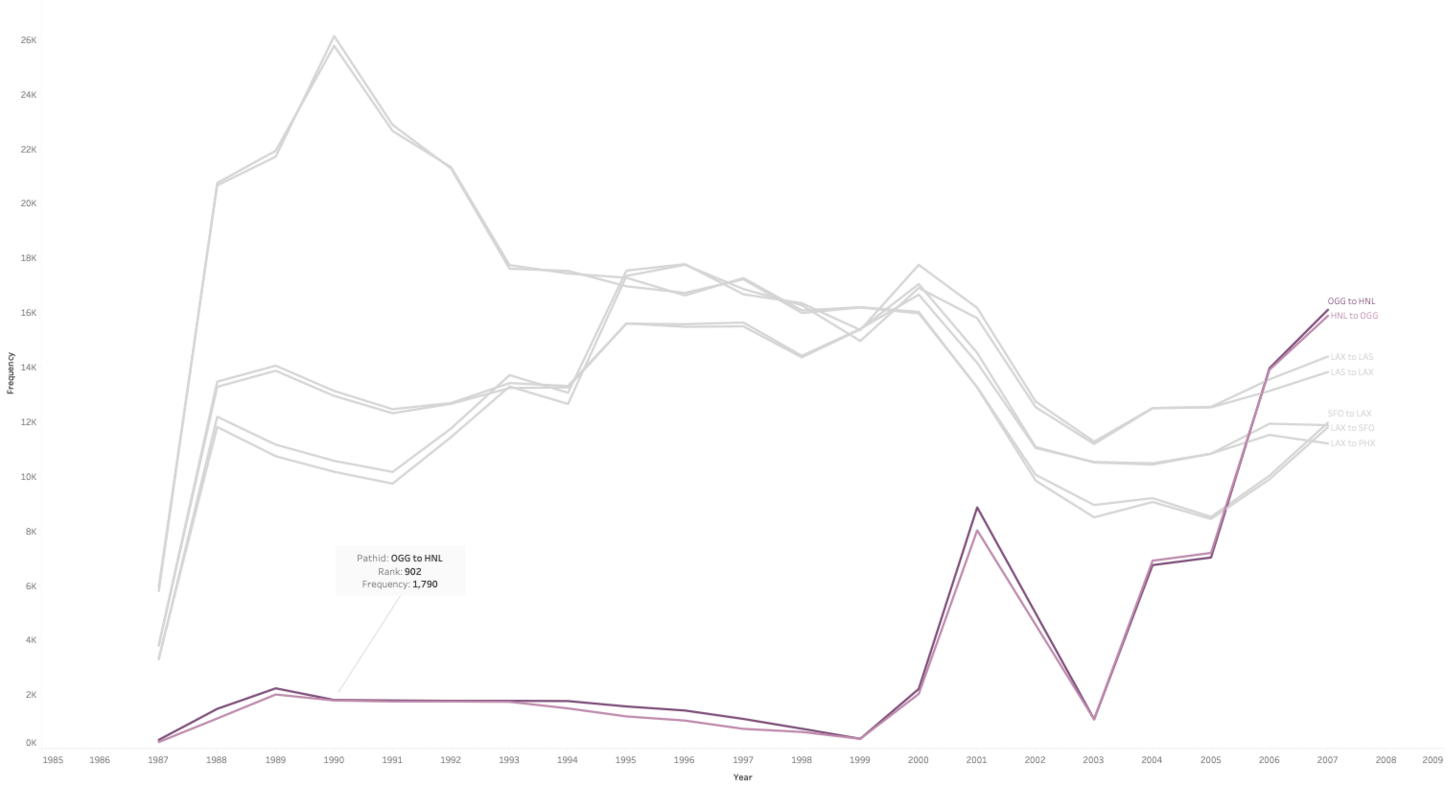


Rank	Pathid	Year 2007 Flight Frequency
1	OGG to HNL	16,099
2	HNL to OGG	15,876
3	LAX to LAS	14,385
4	LAS to LAX	13,815



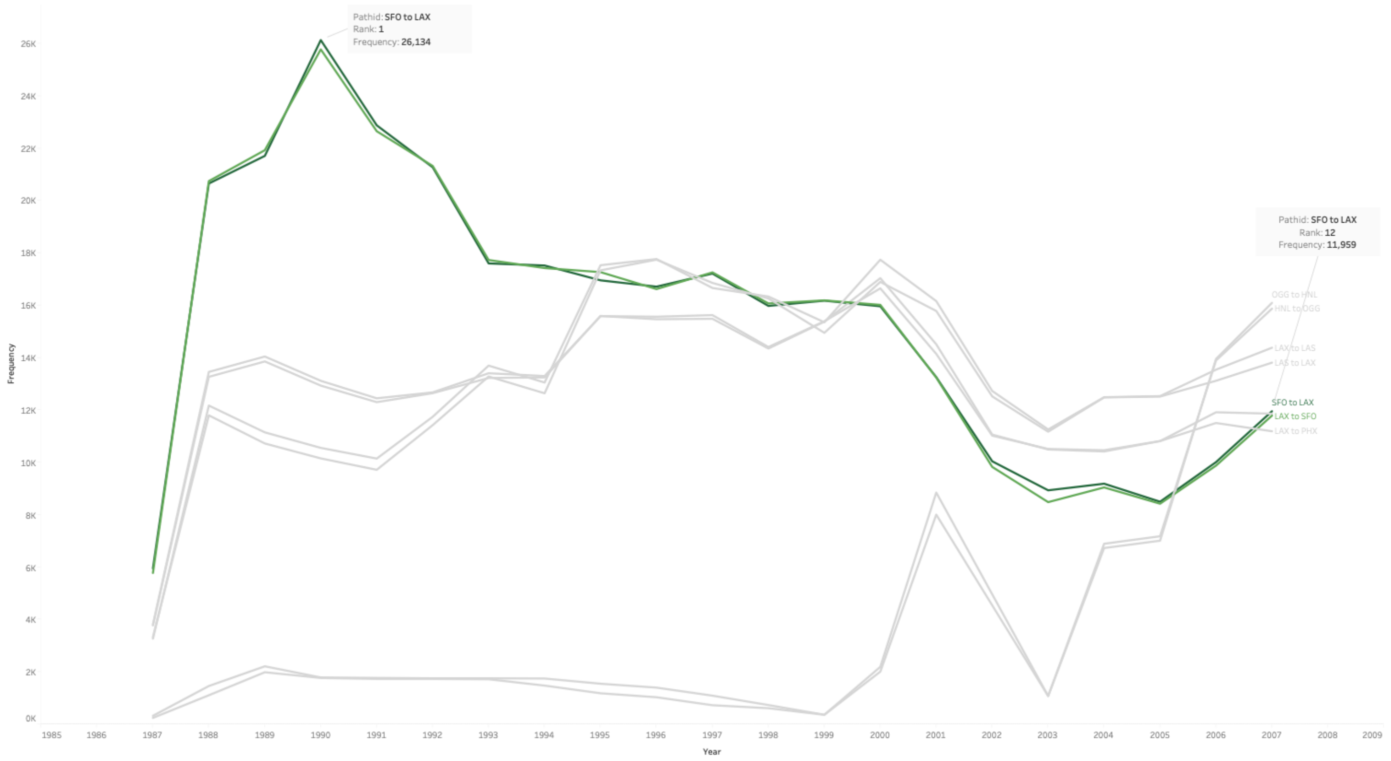
Rank	Pathid	Year 1990 Flight Frequency
1	SFO to LAX	26,134
2	LAX to SFO	25,779
3	LAX to PHX	13,121
4	PHX to LAX	12,938

Frequency Over Time of Top 4 Flight Routes in 1990 and 2007



- The flight route from OGG to HNL, which was ranked 902 in 1990, surpassed the SFO to LAX route in 2006 and reached rank 1 by 2007.

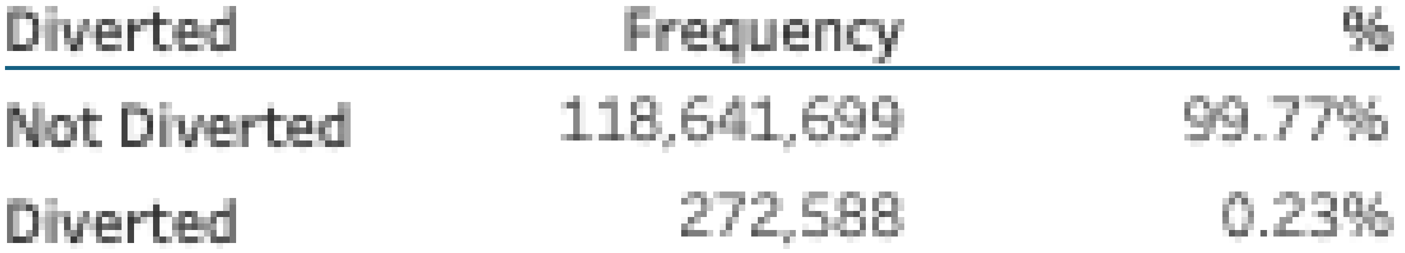
Frequency Over Time of Top 4 Flight Routes in 1990 and 2007



- SFO to LAX route, which ranked 1st with 26,134 flights in 1990, dropped to the 12th place by 2007.

Analysis 4 – How do the top diverted flight routes compare to the average diversion rate of all flights?

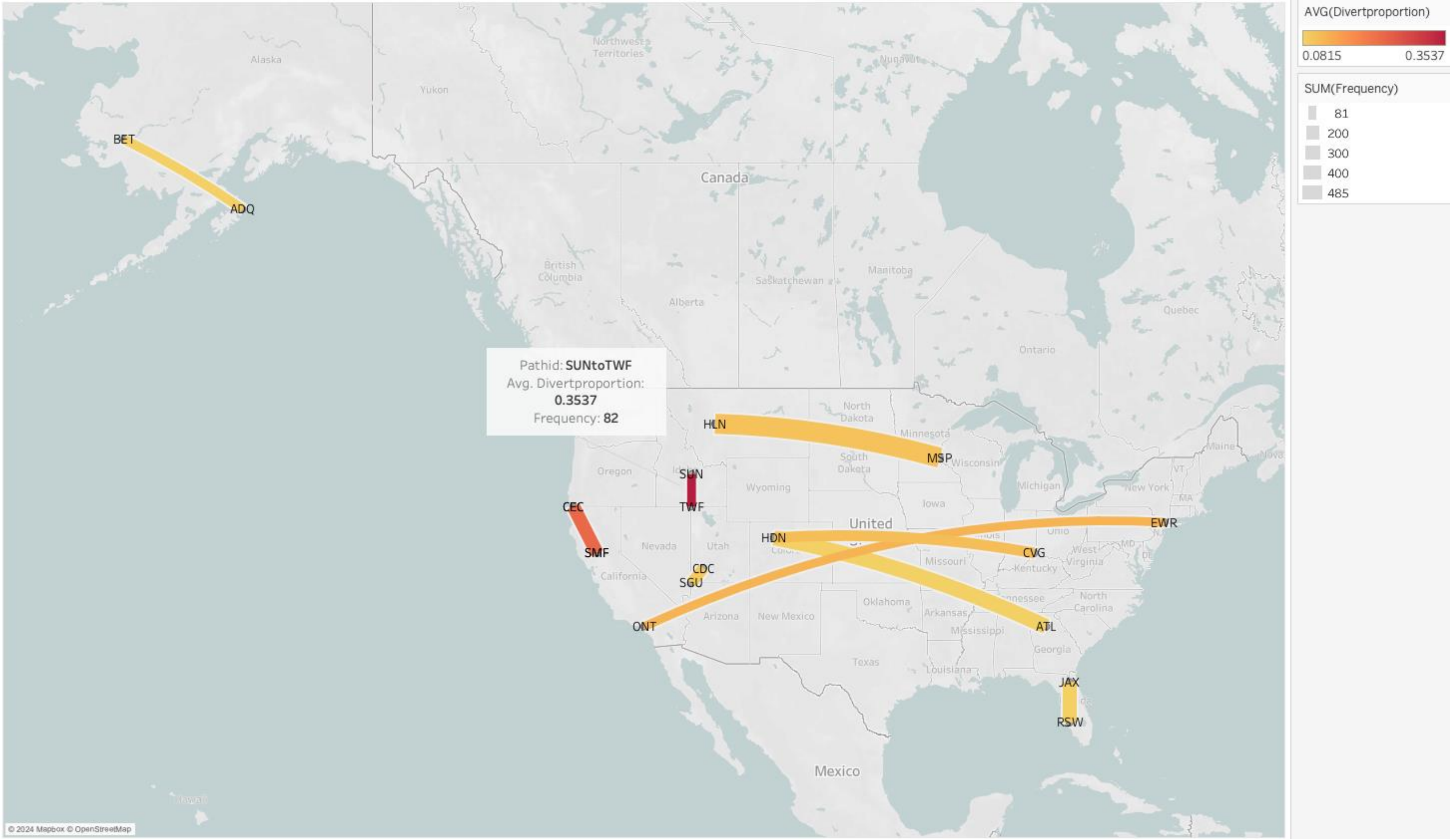
Diverted VS Non-Diverted Flights



Diverted	Frequency	%
Not Diverted	118,641,699	99.77%
Diverted	272,588	0.23%

Top 10 Flight routes with the largest flight diverted proportion

- Flight routes fewer than 50 flights were excluded from the analysis to ensure statistical reliability. Small number of flights can lead to disproportionate diversion rates.
- After exclusion, SUN to TWF routes has the highest diversion rate at 35.37%, which is significantly higher than the overall U.S commercial flight diversion rate of 0.23%.



Analysis 5 – Do flight delays cause a ripple effect across subsequent flights?

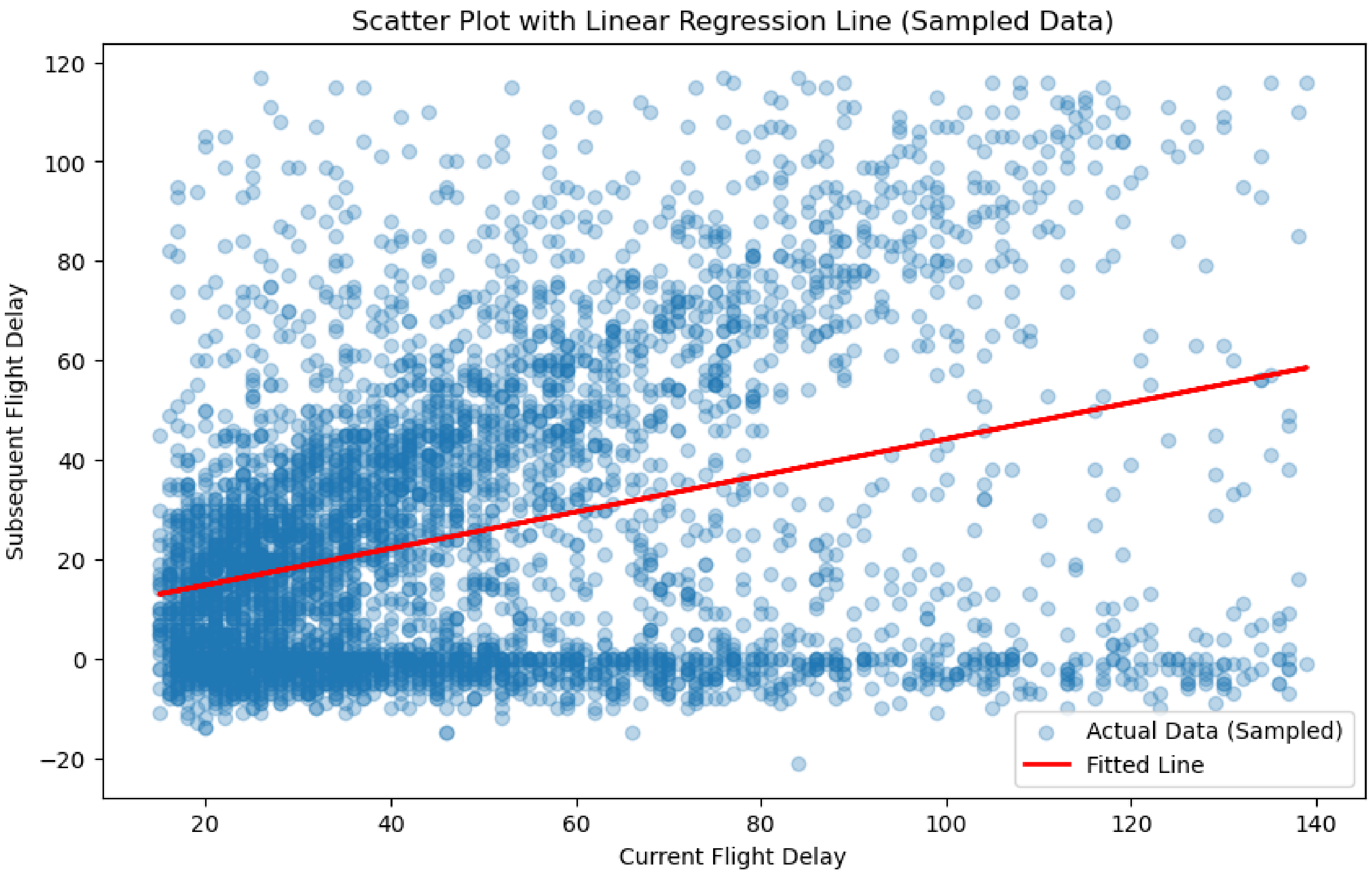
Do flight delays cause a ripple effect across subsequent flights?

- To address this, the data is partitioned by plane tail number and sorted by flight date and time. The data is then merged so that each observation includes the delay for two consecutive flight segments on the same route. For example, for a route like A-B-C, the observation will contain the delay from A to B and the departure delay for the subsequent flight from B to C.
- A regression analysis will then be conducted to determine if there is a meaningful relationship between the current flight delay and the subsequent flight delay.

Scatterplot – Current Flight Delay VS Subsequent Flight Delay

- For enhanced visualization, a scatter plot is created using a sample of 5,000 data points.

- The plot reveals a positive relationship, indicating that a delay in the current flight tends to lead to an increase in the delay of the subsequent flight.



Regression Results

Coefficient: 0.367

The coefficient is 0.367 which means that for every 1 minute in crease in the current flight delay, the subsequent flight delay increases by an average of 0.367min.

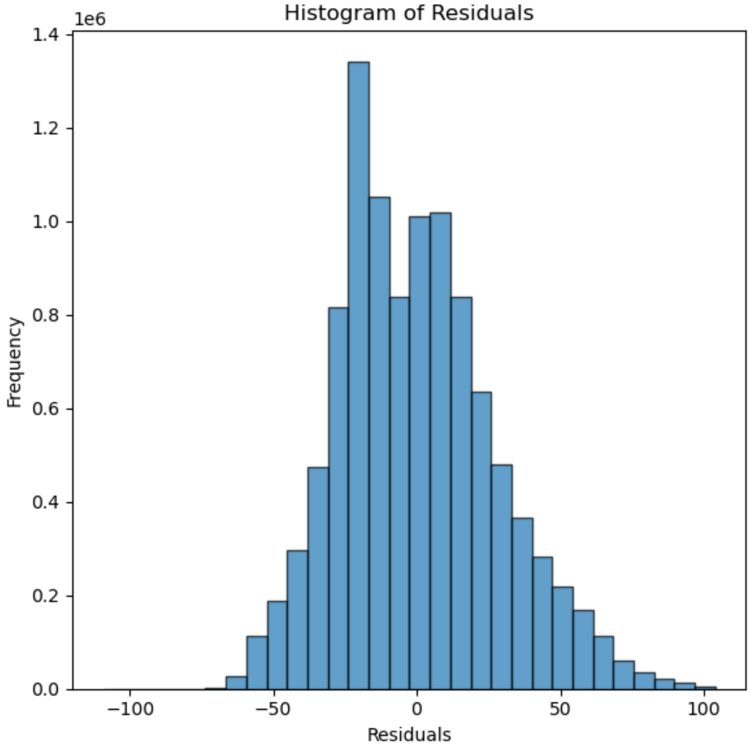
R-squared: 0.12

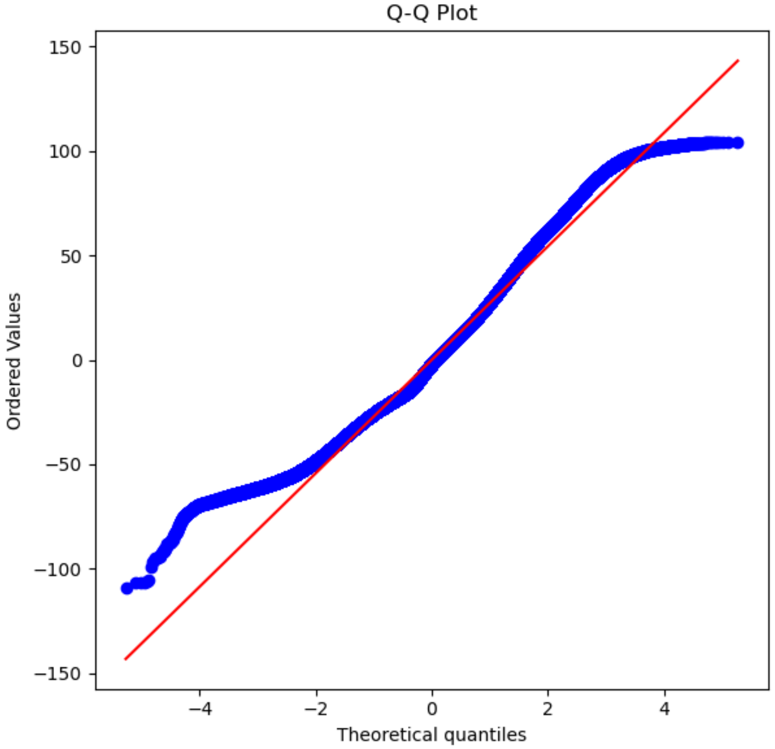
The r-squared value of 0.12 indicates that 12% of the variance in the subsequent flight delay is explained by the current flight delay.

While this shows a relationship between the two variables, it also means that 88% of the variation in the subsequent flight delay is explained by other factors not included in the model.

The End - Thank You

Technical Appendix – Regression Assumptions (analysis 5)





The residuals are normally distributed with mean 0.00 and standard deviation of 27.36. We ignore normality assumption as our sample size is large.