SAS Analysis of Factors Affecting Flight Delays and Flight Times

Saransh Rakshak

Due: 12/11/2024

Introduction

Air travel is an essential mode of transportation, with millions of passengers flying every day across the globe. However, flight delays remain a persistent issue, causing inconvenience to passengers and airlines alike. Understanding the factors that contribute to these delays and how different attributes like weather conditions and airplane characteristics, affect flight times can help mitigate some of these challenges.

This R project data analysis explores the variables that influence flight delays and flight times for departures from New York City's three major airports: John F. Kennedy (*JFK*), LaGuardia (*LGA*), and Newark (*EWR*). The dataset being used for this analysis is from the *nycflights13* library, which contains detailed information on all flights departing from these airports in 2013 (**flights**), along with corresponding weather conditions (**weather**) and airplane details (**planes**). The data is provided by the R *TidyVerse* package, and raw data can be found here: https://github.com/tidyverse/nycflights13.

By integrating data on weather (temperature, wind speed, precipitation) and plane characteristics (year of manufacture, model), we aim to uncover patterns and relationships that contribute to delays and extended flight times.

Data Wrangling

To begin our analysis, we will examine the variables contained within the dataset and explore their impact on key metrics such as departure delays, arrival delays, and overall flight time.

Our **flights_data** frame has been cleaned using R by first removing all rows containing 'NA' values for *dep_delay*, *arr_delay*, and *air_time*. This data is again filtered to include only those that have had departure or arrival delays, as our analysis is focused on factors affecting delay time. Column *was_delayed* is added as a boolean column with values of TRUE for flights that had an arrival delay and FALSE for those that arrived on time, regardless of departure delay.

Now, we will join additional data to **flights_data** to add more information for analysis. First, **airlines** is merged on *carrier* to add information about airline names. Next, **weather** data was joined to our **flights_data** dataframe. Prior to this join, **weather**'s column *time_hour* is renamed to *time_hour_weather* to avoid conflicts with **flights** dataframe column *time_hour*. **Weather** is left joined to **flights_data** to preserve data from **flights**, and is joined by columns *year*, *month*, *day*, *hour*, and *origin*. Finally, data regarding details about **planes** for each corresponding flight is joined to our **flights_data**. Prior to this join, **planes** column *year* is renamed to *year_manufactured* to avoid conflicts with **flights** dataframe column *year*. **Planes** is left joined to **flights_data** to preserve data from **flights**, and is joined by *tailnum*.

We will import our semi-cleaned dataset from R to SAS as follows:

```{R}
# Load required libraries
library(nycflights13)
library(dplyr)  # For mutate() and across()
library(tidyr)  # For replace_na()
library(readr)  # For write_csv()

# Create a copy of the flights dataset
flights_processed <- flights
flights_processed <- flights_processed %>%
  mutate(across(where(is.character), ~ tidyr::replace_na(., ""))) %>%
  mutate(across(where(is.numeric), ~ tidyr::replace_na(., as.numeric(NA))))


airlines_cleaned <- airlines %>%
  mutate(across(where(is.character), ~ replace_na(., ""))) %>%
  mutate(across(where(is.numeric), ~ replace_na(., NA_real_)))
# Rename column 'name' in airlines to 'carrier_name'
colnames(airlines_cleaned)[colnames(airlines_cleaned) == 'name'] <- 'carrier_name'
# Join airline data to flights
flights_processed <- flights_processed %>%
  left_join(airlines_cleaned, by = "carrier")


planes_cleaned <- planes %>%
  mutate(across(where(is.character), ~ replace_na(., ""))) %>%
  mutate(across(where(is.numeric), ~ replace_na(., NA_real_))) %>%
  mutate(
    year = ifelse(is.na(year) | year == "NA", NA_integer_, as.integer(year))
  )
# Rename column 'year' in planes to 'year_manufactured'
colnames(planes_cleaned)[colnames(planes_cleaned) == 'year'] <- 'year_manufactured'
# Rename column 'engines' in planes to 'num_engines'
colnames(planes_cleaned)[colnames(planes_cleaned) == 'engines'] <- 'num_engines'
# Join plane data to flights
flights_processed <- flights_processed %>%
  left_join(planes_cleaned, by = "tailnum")


weather_cleaned <- weather %>%
  mutate(across(where(is.character), ~ tidyr::replace_na(., ""))) %>%
  mutate(across(where(is.numeric), ~ tidyr::replace_na(., 0))) %>%
  mutate(
    wind_dir = ifelse(wind_dir %in% c("NA", ""), NA_real_, as.numeric(wind_dir))
  )
# Rename column 'time_hour' in weather to 'time_hour_weather'
colnames(weather_cleaned)[colnames(weather_cleaned) == 'time_hour'] <- 'time_hour_weather'
# Join flights with weather data
flights_processed <- flights_processed %>%
  left_join(weather_cleaned, by = c("year", "month", "day", "hour", "origin"))

flights_processed

write_csv(flights_processed, "flights_processed.csv", na = ".")
```

And to load the saved ".csv" file containing the dataset into our SAS environment, we will do the following:

```
3   libname flights '/home/sraksha/final_project/data';
4   /* Importing Data */
5   proc import datafile="/home/sraksha/final_project/data/flights_processed.csv"
6       OUT=flights_data
7       DBMS=CSV
8       REPLACE;
9       GETNAMES=YES;
10  RUN;
11  /* Adding was_delayed boolean column */
12  DATA flights_data;
13      SET flights_data;
14      was_delayed = (arr_delay > 0 or dep_delay > 0);
15  RUN;
```

Our initial analysis shows that our dataset contains 327,346 rows and 39 columns. Of these columns, we will focus our analysis on the following variables.

- *dep_delay*, *arr_delay* : Departure and arrival delays, in minutes. Negative times represent early departures/arrivals.
- *was_delayed* : Boolean column, equals True if there was a delay in arrival, else False.
- *air_time* : Amount of time spent in the air, in minutes.
- *origin* : The airport in NYC from which the flight departed (*JFK*, *LGA*, or *EWR*).
- *dest* : The destination airport.
- *carrier* : Two letter carrier abbreviation for the airline responsible for the flight.
- *carrier_name* : Full name for the airline responsible for the flight.
- *temp*, *precip*, *visib* : Weather conditions at time of departure, including temperature (°F), precipitation (inches), and visibility (miles).
- *year_manufactured*, *model*, *num_engines*, *speed*, *engine* : Details regarding plane that took the journey, such as its model, year of manufacture, number of engines, average cruising speed (MPH), and type of engine.

Upon reviewing our dataset, we can see that it is semi-cleaned. All rows contain values for our column variables *dep_delay*, *arr_delay*, and *air_time*. For our merged columns from **airlines**, **weather** and **planes**, there are instances where certain extra column values contain missing data (represented by 'NA' in our system). When conducting analysis on these columns, we ensure that only data points from samples which do not have any missing information are considered, and those with 'NA' are filtered out. This approach is taken because there is currently no new data available to fill in the gaps left by these missing values.

Flights Analysis

We will be being our analysis by looking at **flights** data regarding carrier and origin.

1. Carrier

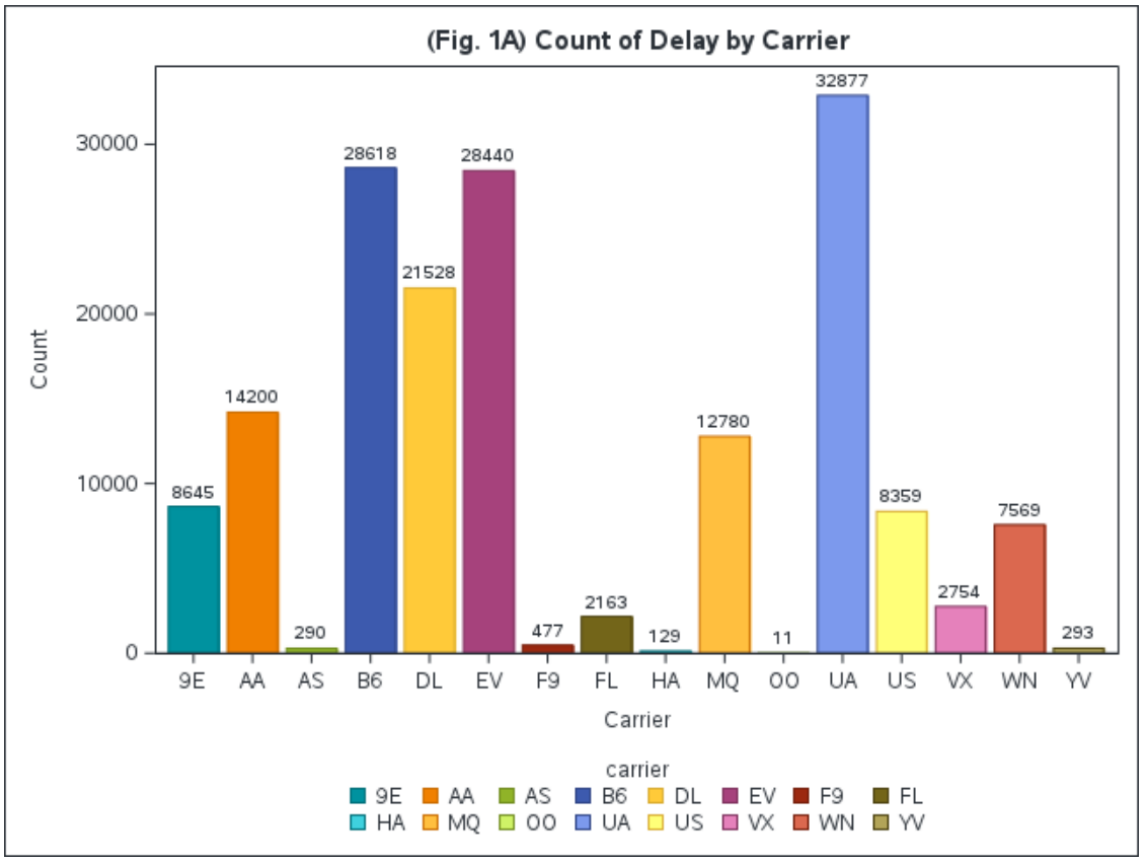
From Fig. 1A, it is clear that some carriers have significantly higher counts of delays than others. Carriers like United Airlines and American Airlines have the highest number of delays. This could be attributed to the number of flights they

operate or operational challenges unique to these carriers. The data reveals that carriers with higher traffic volumes have a larger share of delays, since their flight schedules may be more complex and susceptible to disruptions.

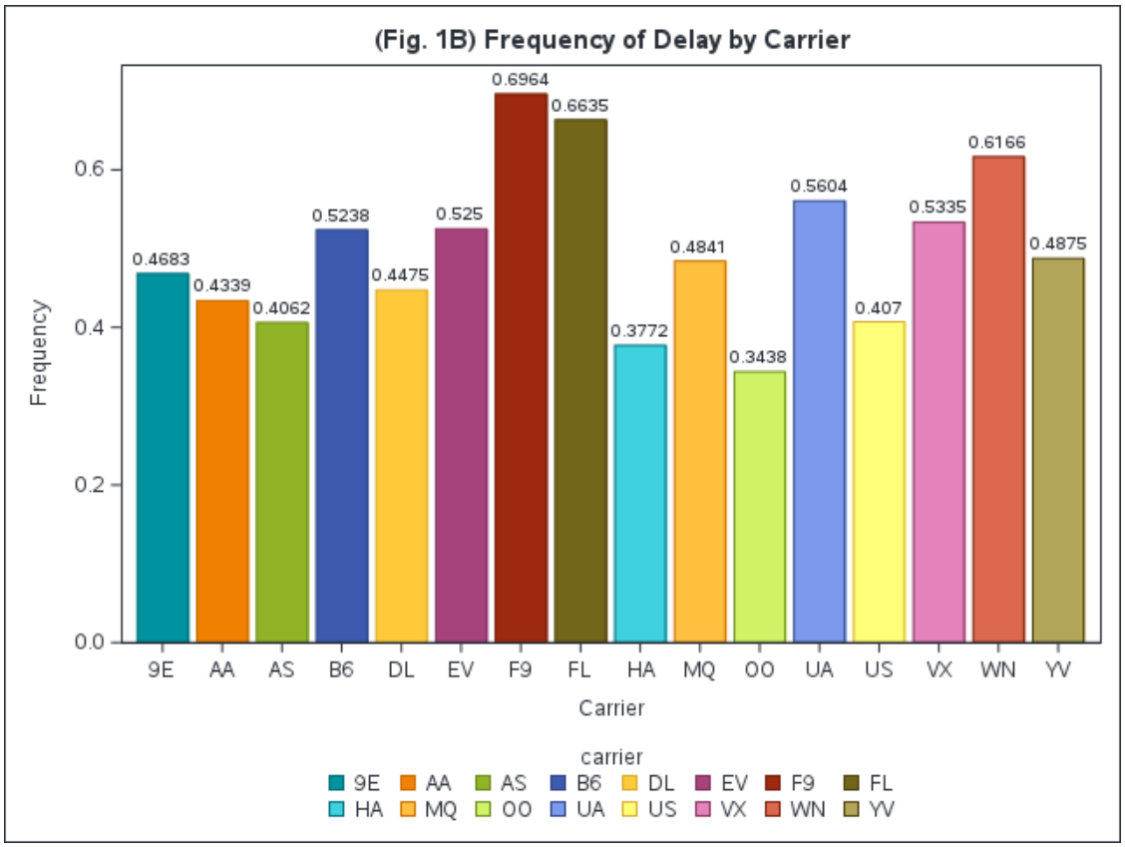
While the count provides raw numbers, Fig. 1B presents a normalized view by showing the frequency of delays relative to the total number of flights operated by each carrier. It offers a fairer comparison, as it accounts for the differing sizes of the airlines. Interestingly, some carriers with fewer total delays have higher delay frequencies, indicating they experience delays more often relative to their size. Low-cost carriers or smaller regional airlines sometimes exhibit higher delay frequencies, suggesting that operational inefficiencies may be more common for these airlines despite handling fewer flights.

Together, these two figures provide insights into both the absolute and proportional performance of airlines in terms of delays, highlighting operational challenges faced by larger carriers and frequent delays within smaller airlines.

```
/* 1A. Carrier Delay Count */
PROC SQL;
    CREATE TABLE flights_carrier_count AS
    SELECT carrier, COUNT(*) AS count
    FROM flights_data
    WHERE was_delayed = 1
    GROUP BY carrier
    ORDER BY count DESC;
QUIT;
PROC PRINT DATA=flights_carrier_count; RUN;
PROC SGPLOT DATA=flights_carrier_count;
    VBAR carrier / RESPONSE=count DATALABEL GROUP=carrier STAT=SUM;
    TITLE "(Fig. 1A) Count of Delay by Carrier";
    XAXIS LABEL="Carrier";
    YAXIS LABEL="Count";
RUN;
```



```sql
/* 1B. Carrier Delay Frequency */
PROC SQL;
  CREATE TABLE flights_delay_freq AS
  SELECT carrier,
         SUM(was_delayed) AS delayed,
         SUM(CASE WHEN was_delayed = 0 THEN 1 ELSE 0 END) AS notdelayed,
         SUM(was_delayed) / (SUM(was_delayed) + SUM(CASE WHEN was_delayed
  FROM flights_data
  GROUP BY carrier;
QUIT;
PROC PRINT DATA=flights_delay_freq; RUN;
PROC SGPLOT DATA=flights_delay_freq;
    VBAR carrier / RESPONSE=delay_freq DATALABEL GROUP=carrier STAT=SUM;
    TITLE "(Fig. 1B) Frequency of Delay by Carrier";
    XAXIS LABEL="Carrier";
    YAXIS LABEL="Frequency";
RUN;
```

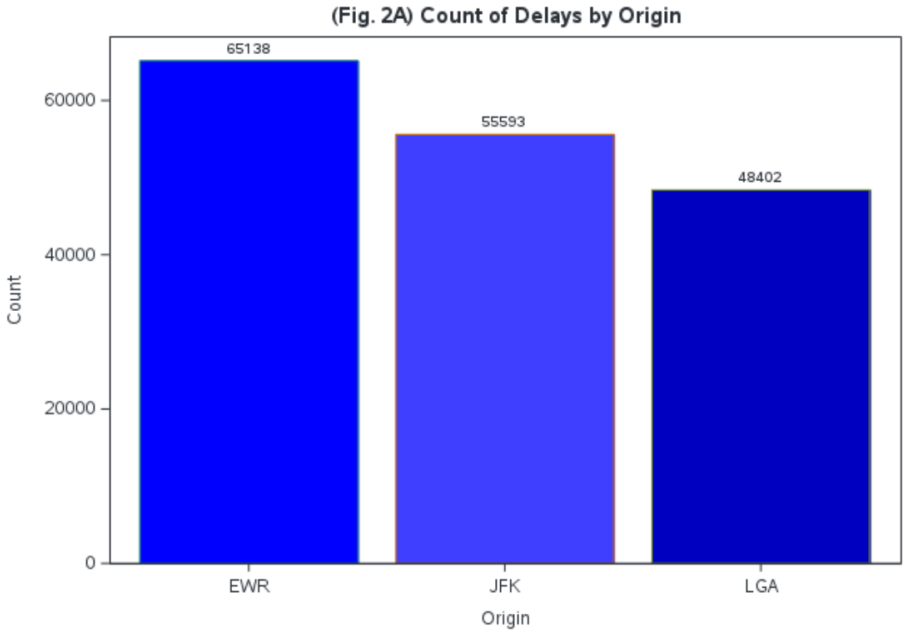


2. Origin

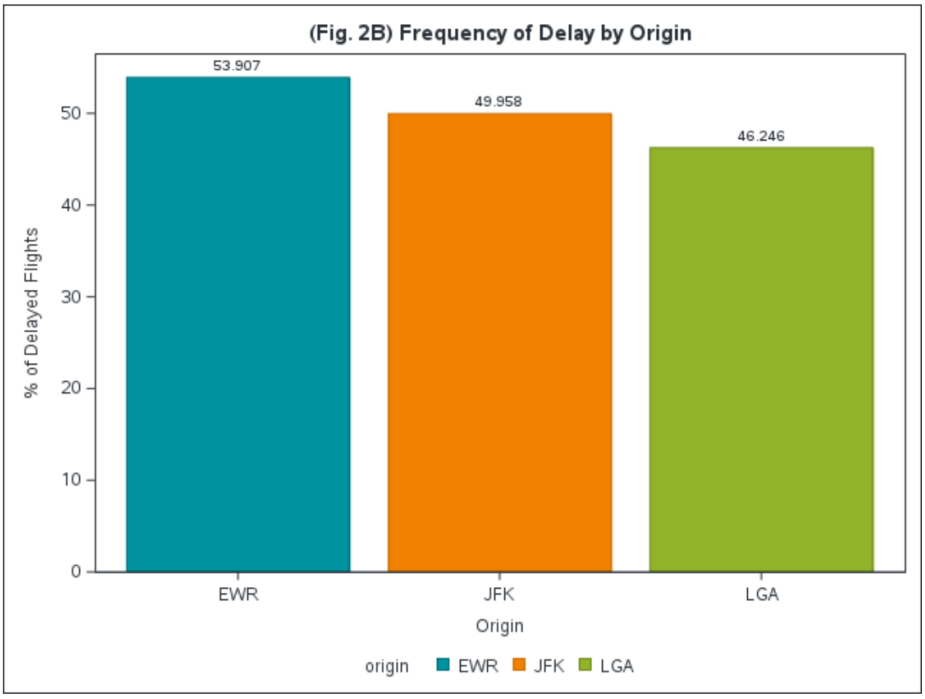
Fig. 2A reveals that *EWR* (Newark Liberty International Airport) has the highest total number of delayed flights. The volume of delays at *EWR* surpasses those at both *JFK* and *LGA.* Newark's high delay count may stem from its larger volume of traffic and greater complexity in managing both domestic and international flights which could significantly contribute to more frequent operational delays. Similarly, *EWR* also leads in the frequency of delays (Fig. 2B), indicating that a significant proportion of flights departing from Newark experience delays.

These finding highlights that not only does *EWR* handle a large number of delayed flights, but the percentage of flights delayed relative to the total number of flights is also the highest among the three airports. This suggests that Newark faces chronic issues affecting its punctuality.

```sas
/* 2A. Count of delays grouped by origin */
PROC SQL;
  CREATE TABLE flights_origin_count AS
  SELECT origin,
         COUNT(*) AS count
  FROM flights_data
  WHERE was_delayed = 1
  GROUP BY origin
  ORDER BY count DESC;
QUIT;
PROC SGPLOT DATA=flights_origin_count;
  VBAR origin / RESPONSE=count GROUP=origin DATALABEL;
  TITLE "(Fig. 2A) Count of Delays by Origin";
  XAXIS LABEL="Origin";
  YAXIS LABEL="Count";
  STYLEATTRS DATACOLORS=(blue);
RUN;
```



```sql
/* 2B. Origin Delay Frequency */
PROC SQL;
  CREATE TABLE flights_origin_freq AS
  SELECT origin,
         SUM(was_delayed) AS delayed,
         SUM(CASE WHEN was_delayed = 0 THEN 1 ELSE 0 END) AS notDelayed,
         100 * SUM(was_delayed) /
               (SUM(was_delayed) + SUM(CASE WHEN was_delayed = 0 THEN 1 ELSE 0 END)) AS delay_freq
  FROM flights_data
  GROUP BY origin;
QUIT;
PROC SGPLOT DATA=flights_origin_freq;
  VBAR origin / RESPONSE=delay_freq GROUP=origin DATALABEL;
  TITLE "(Fig. 2B) Frequency of Delay by Origin";
  XAXIS LABEL="Origin";
  YAXIS LABEL="% of Delayed Flights";
RUN;
```



Weather Analysis

We will now incorporate **weather** data in our delay analysis.

3. Temperature

Our dataset **flights_data** has some rows with 'NA' values for *temp*. For this analysis, those rows will be filtered out.

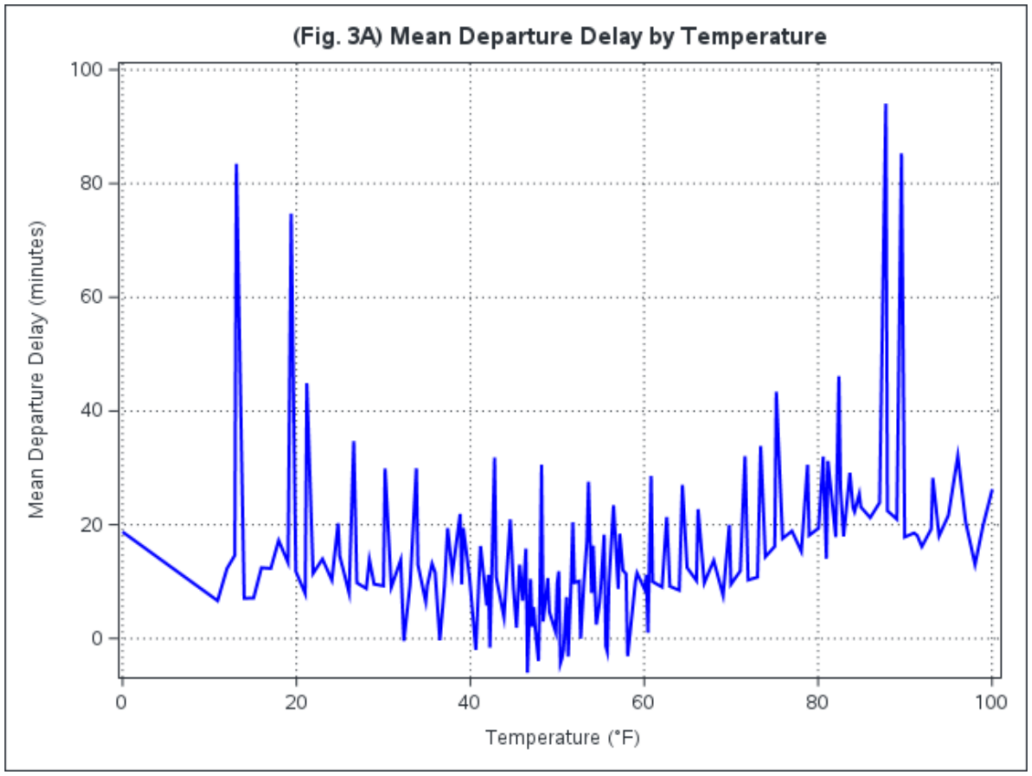
Fig. 3A reveals that flight delays generally increase in extreme weather conditions. Delays tend to be higher when temperatures drop below 30°F and when they rise above 85°F. These findings suggest that both very cold and very hot temperatures can contribute to disruptions in flight operations, likely due to issues such as de-icing, reduced aircraft performance, or weather-related complications like thunderstorms in high heat.

When examining how different airlines are affected by temperature changes, the results show distinct patterns among various carriers (Fig. 3B). Airlines such as Endeavor (*9E*), American Airlines (*AA*), JetBlue (*B6*), Frontier (*F9*), Hawaiian Airlines (*HA*), and US Airways (*US*) exhibit particularly high delays during cold weather (below 30°F). In contrast, carriers like AirTran (*FL*) and SkyWest Airlines (*OO*) tend to experience higher delays in hot temperatures (above 80°F). This suggests that certain airlines may have operational vulnerabilities in extreme weather, possibly due to factors like fleet composition, hub location, or preparedness for adverse weather conditions.

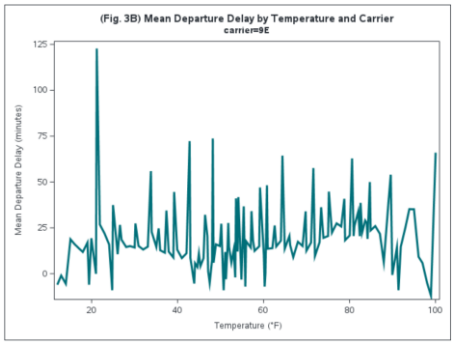
The analysis by airport of origin (Fig. 3C) highlights significant differences in how temperature impacts delays. Newark (*EWR*) stands out as the airport most affected by extreme temperatures, showing the greatest change in delay times in both very cold and very hot weather. This may be due to *EWR*'s operational characteristics, weather patterns in its vicinity, or infrastructure challenges. *JFK* and *LGA* also show increased delays in extreme conditions, but the effects are less pronounced than at *EWR*.

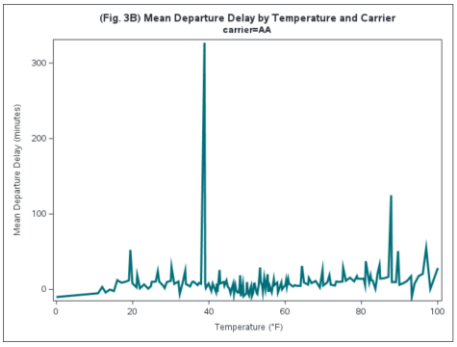
Our findings suggest that flight delays increase significantly in extreme temperatures. The most affected airlines in cold weather (*9E*, *AA*, *B6*, *F9*, *HA*, and *US*) may need to improve their winter operational strategies, while airlines like *FL* and *OO* could benefit from addressing challenges in hot weather conditions. Newark (*EWR*) is the most sensitive to temperature changes, indicating potential for focused improvements in weather-resilient infrastructure or procedures at this airport.

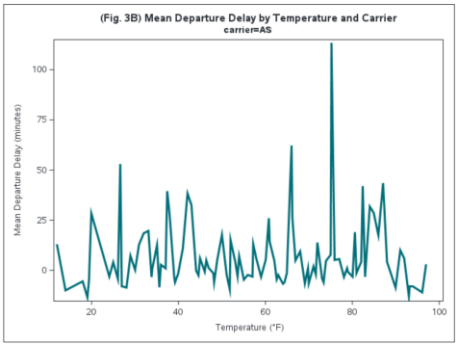
```
/* 3A. Mean Departure Delay by Temperature */
PROC SQL;
  CREATE TABLE flights_temp_mean AS
  SELECT temp,
         MEAN(dep_delay) AS mean_dep_delay
  FROM flights_data
  WHERE temp IS NOT NULL
  GROUP BY temp;
QUIT;
PROC SGPLOT DATA=flights_temp_mean;
  SERIES X=temp Y=mean_dep_delay / LINEATTRS=(THICKNESS=2 COLOR=blue);
  TITLE "(Fig. 3A) Mean Departure Delay by Temperature";
  XAXIS LABEL="Temperature (°F)" GRID;
  YAXIS LABEL="Mean Departure Delay (minutes)" GRID;
RUN;
```

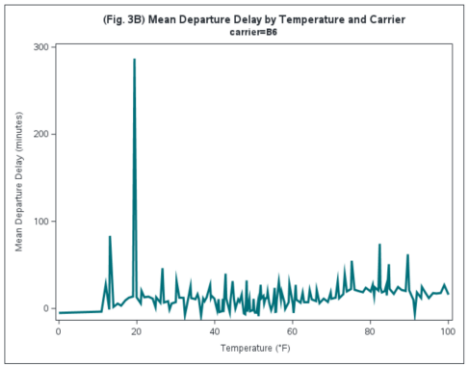


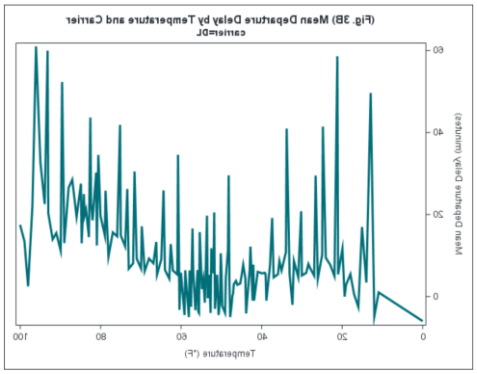
```sql
/* 3B. Mean Departure Delay by Temperature and Carrier */
PROC SQL;
    CREATE TABLE flights_temp_carrier_mean AS
    SELECT carrier,
           temp,
           MEAN(dep_delay) AS mean_dep_delay
    FROM flights_data
    WHERE temp IS NOT NULL
    GROUP BY carrier, temp;
QUIT;
PROC SORT DATA=flights_temp_carrier_mean;
    BY carrier;
RUN;
PROC SGPLOT DATA=flights_temp_carrier_mean;
    SERIES X=temp Y=mean_dep_delay;
    TITLE "(Fig. 3B) Mean Departure Delay by Temperature and Carrier";
    XAXIS LABEL="Temperature (°F)";
    YAXIS LABEL="Mean Departure Delay (minutes)";
    BY carrier;
RUN;
```

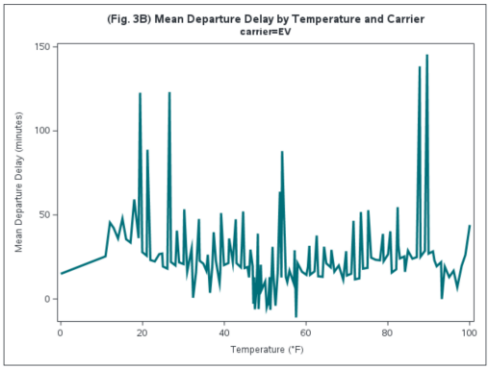


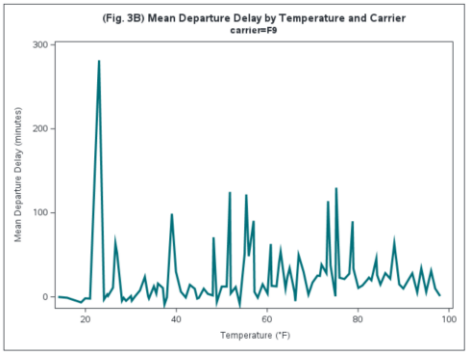


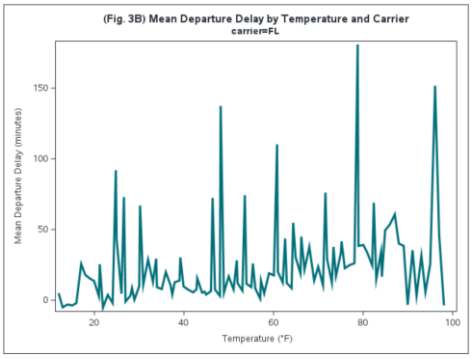


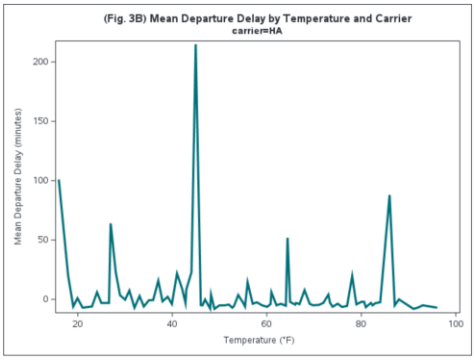


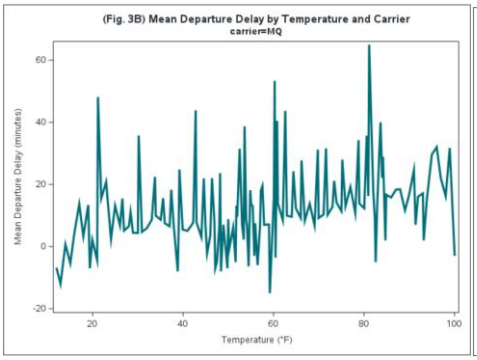


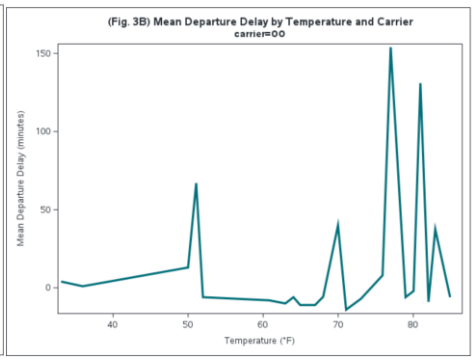


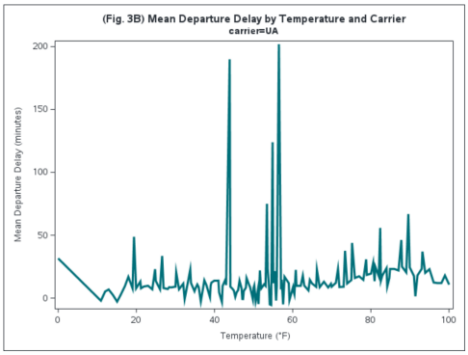

(Fig. 3B) Mean Departure Delay by Temperature and Carrier
carrier=F9
Mean Departure Delay (minutes)
Temperature (°F)

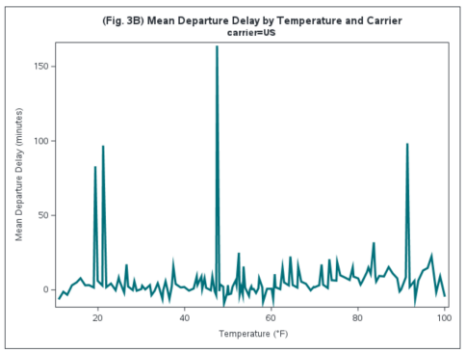

(Fig. 3B) Mean Departure Delay by Temperature and Carrier
carrier=FL
Mean Departure Delay (minutes)
Temperature (°F)

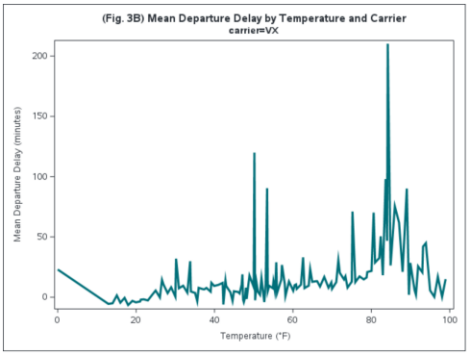

(Fig. 3B) Mean Departure Delay by Temperature and Carrier
carrier=HA
Mean Departure Delay (minutes)
Temperature (°F)

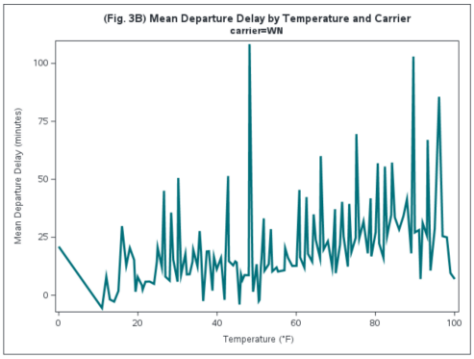

(Fig. 3B) Mean Departure Delay by Temperature and Carrier
carrier=MQ
Mean Departure Delay (minutes)
Temperature (°F)

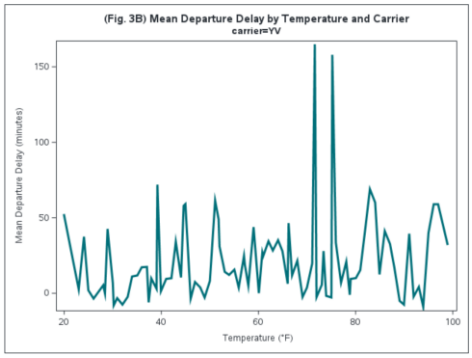

(Fig. 3B) Mean Departure Delay by Temperature and Carrier
carrier=OO
Mean Departure Delay (minutes)
Temperature (°F)


(Fig. 3B) Mean Departure Delay by Temperature and Carrier
carrier=UA
Mean Departure Delay (minutes)
Temperature (°F)

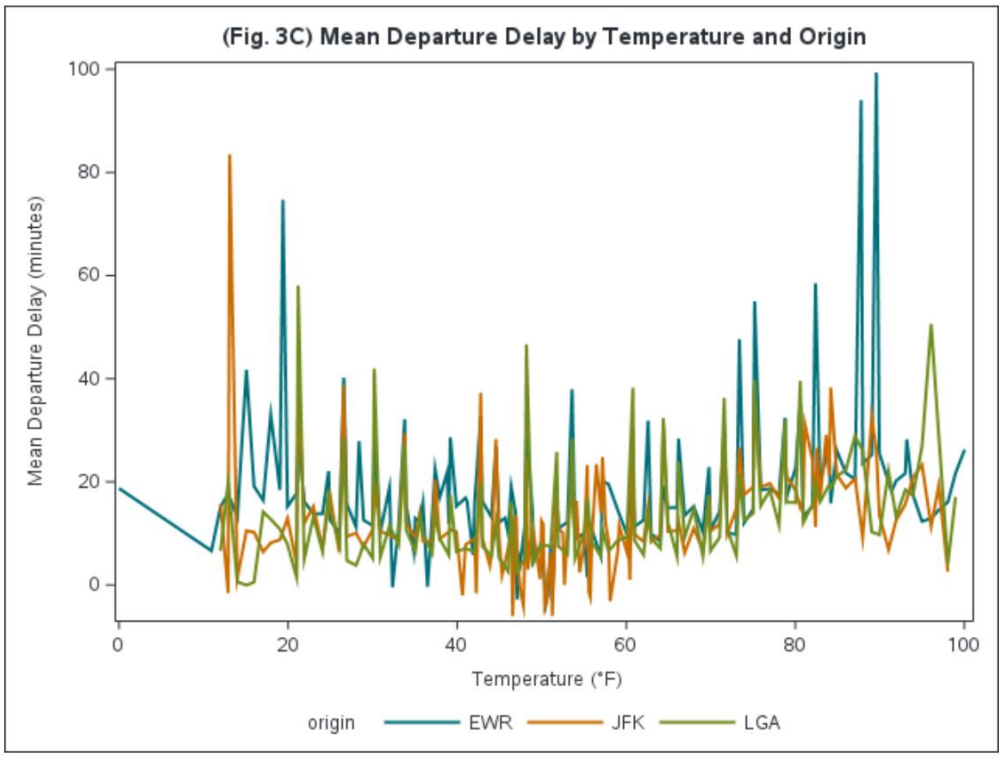

(Fig. 3B) Mean Departure Delay by Temperature and Carrier
carrier=US
Mean Departure Delay (minutes)
Temperature (°F)


(Fig. 3B) Mean Departure Delay by Temperature and Carrier
carrier=VX
Mean Departure Delay (minutes)
Temperature (°F)


(Fig. 3B) Mean Departure Delay by Temperature and Carrier
carrier=WN
Mean Departure Delay (minutes)
Temperature (°F)


(Fig. 3B) Mean Departure Delay by Temperature and Carrier
carrier=YV
Mean Departure Delay (minutes)
Temperature (°F)

```sql
/* 3C. Mean Departure Delay by Temperature and Origin */
PROC SQL;
    CREATE TABLE flights_temp_origin_mean AS
    SELECT origin, temp,
           MEAN(dep_delay) AS mean_dep_delay
    FROM flights_data
    WHERE temp IS NOT NULL
    GROUP BY origin, temp;
QUIT;
PROC SGPLOT DATA=flights_temp_origin_mean;
    SERIES X=temp Y=mean_dep_delay / GROUP=origin LINEATTRS=(THICKNESS=2);
    XAXIS LABEL="Temperature (°F)";
    YAXIS LABEL="Mean Departure Delay (minutes)";
    TITLE "(Fig. 3C) Mean Departure Delay by Temperature and Origin";
RUN;
```



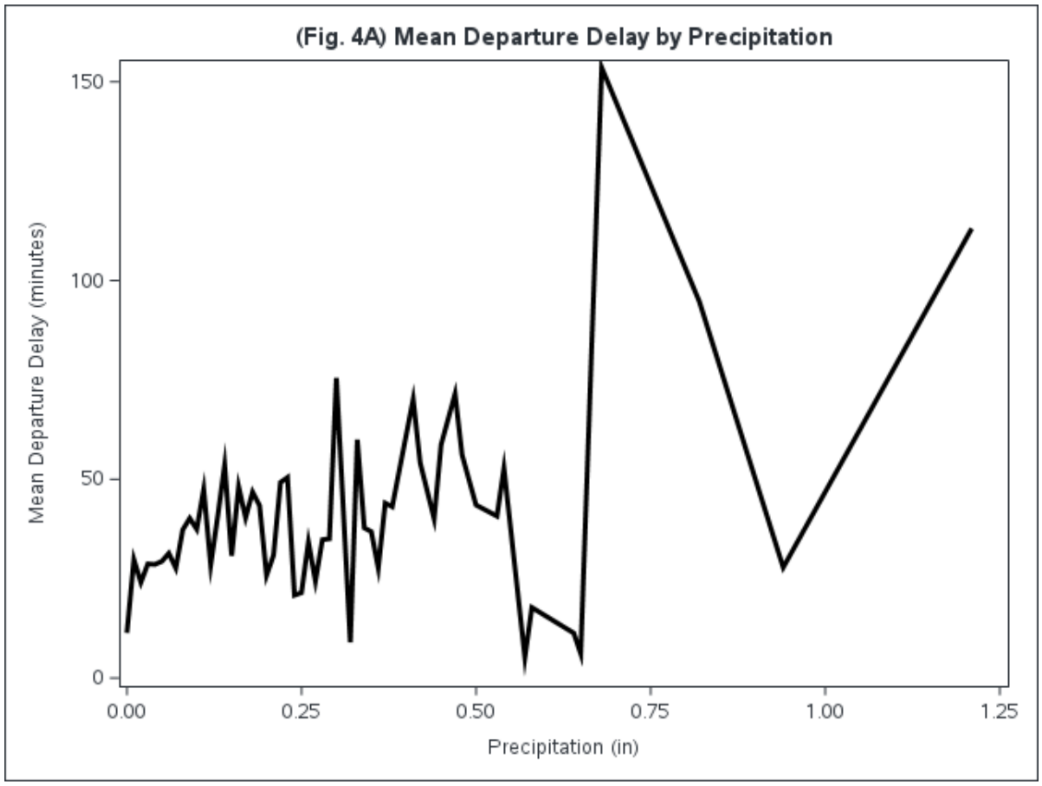
4. Precipitation

Our dataset **flights_data** has some rows with 'NA' values for *precip*. For this analysis, those rows will be filtered out.

Fig. 4A shows a clear trend: as precipitation increases, so do flight departure delays. This is expected since rain, snow, and other forms of precipitation can disrupt airport operations, slow down runway traffic, and affect aircraft performance. For instance, wet or icy runways necessitate longer takeoff and landing times, and heavy precipitation often leads to delays in both air traffic control and ground services.

However, it's important to note that the dataset contains relatively few data points for higher precipitation levels. The total number of observations for this analysis is only 55 (55 levels of precipitation observed), with much fewer samples in cases of heavy precipitation. This limited data makes it difficult to draw strong conclusions about the exact impact of severe weather on delays, as the trends observed may not fully represent all conditions. The small sample size suggests that while there is a general relationship between precipitation and delays, further data collection would be necessary to make more reliable and detailed observations about the effects of heavy rain or snow on flight departures.

```sql
/* 4A. Mean Departure Delay by Precipitation */
PROC SQL;
    CREATE TABLE flights_precip_mean AS
    SELECT precip,
           MEAN(dep_delay) AS mean_dep_delay
    FROM flights_data
    WHERE precip IS NOT NULL
    GROUP BY precip;
QUIT;
PROC SQL;
    SELECT COUNT(*) INTO :row_count
    FROM flights_precip_mean;
QUIT;
%PUT Number of rows in flights_precip_mean: &row_count;
PROC SGPLOT DATA=flights_precip_mean;
    SERIES X=precip Y=mean_dep_delay / LINEATTRS=(COLOR=BLACK);
    TITLE "(Fig. 4A) Mean Departure Delay by Precipitation";
    XAXIS LABEL="Precipitation (in)";
    YAXIS LABEL="Mean Departure Delay (minutes)";
RUN;
```



5. Visibility

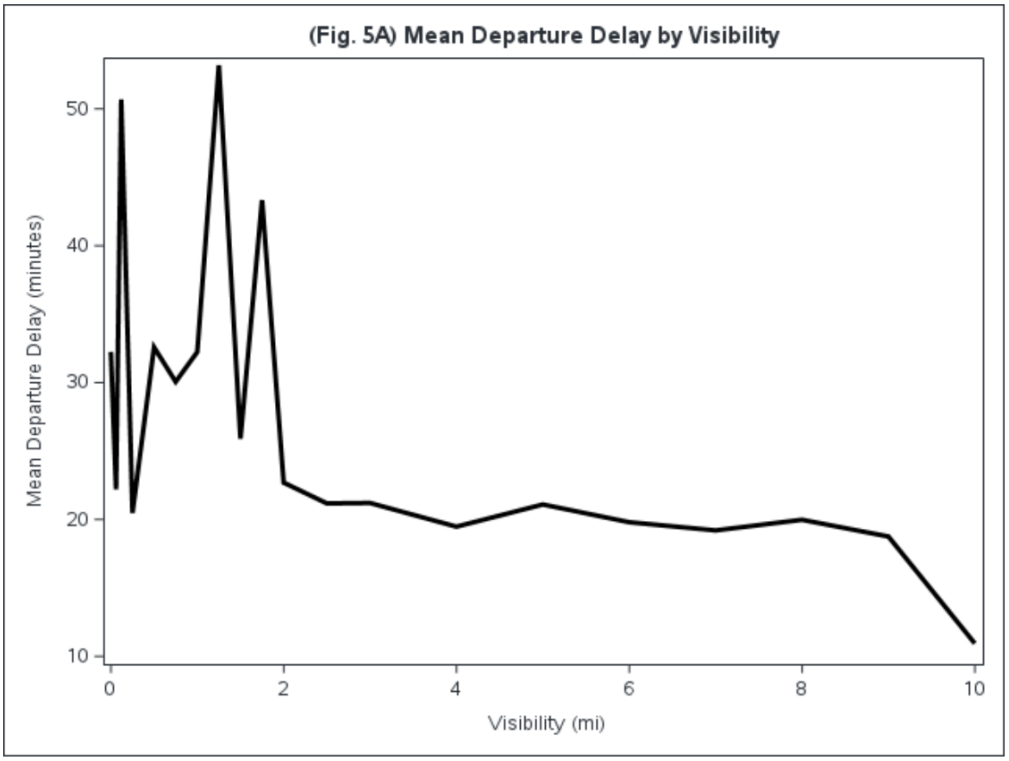
Our dataset **flights_data** has some rows with 'NA' values for *visib*. For this analysis, those rows will be filtered out.

Fig. 5A (Mean Departure Delay by Visibility) highlights how visibility levels affect flight departure delays. From the graph, we observe a general trend where reduced visibility correlates with longer departure delays. This pattern aligns with expectations, as lower visibility conditions—such as fog or heavy precipitation—typically disrupt airport operations, requiring more caution during takeoff and landing procedures, which in turn leads to delays. Similarly, higher visibility conditions allow for more on-time departures.

However, it is critical to address the limitations of the dataset used for this analysis. The dataset includes only 20 rows of data (20 observed visibilities), which significantly limits the ability to make strong generalizations or conclusive statements about the impact of visibility on flight delays. A small sample size increases the possibility of statistical noise,

meaning that any trends observed might not hold true across a larger, more comprehensive dataset. Therefore, while the current data provides some insight, it lacks the depth and breadth needed for robust conclusions. Expanding the dataset would provide more reliable and statistically significant results.

```
169  /* 5A. Mean Departure Delay by Visibility */
170  PROC SQL;
171      CREATE TABLE flights_visib_mean AS
172      SELECT visib, MEAN(dep_delay) AS mean_dep_delay
173      FROM flights_data
174      WHERE visib IS NOT NULL
175      GROUP BY visib;
176  QUIT;
177  PROC SGPLOT DATA=flights_visib_mean;
178      SERIES X=visib Y=mean_dep_delay / LINEATTRS=(COLOR=BLACK);
179      XAXIS LABEL="Visibility (mi)";
180      YAXIS LABEL="Mean Departure Delay (minutes)";
181      TITLE "(Fig. 5A) Mean Departure Delay by Visibility";
182  RUN;
```



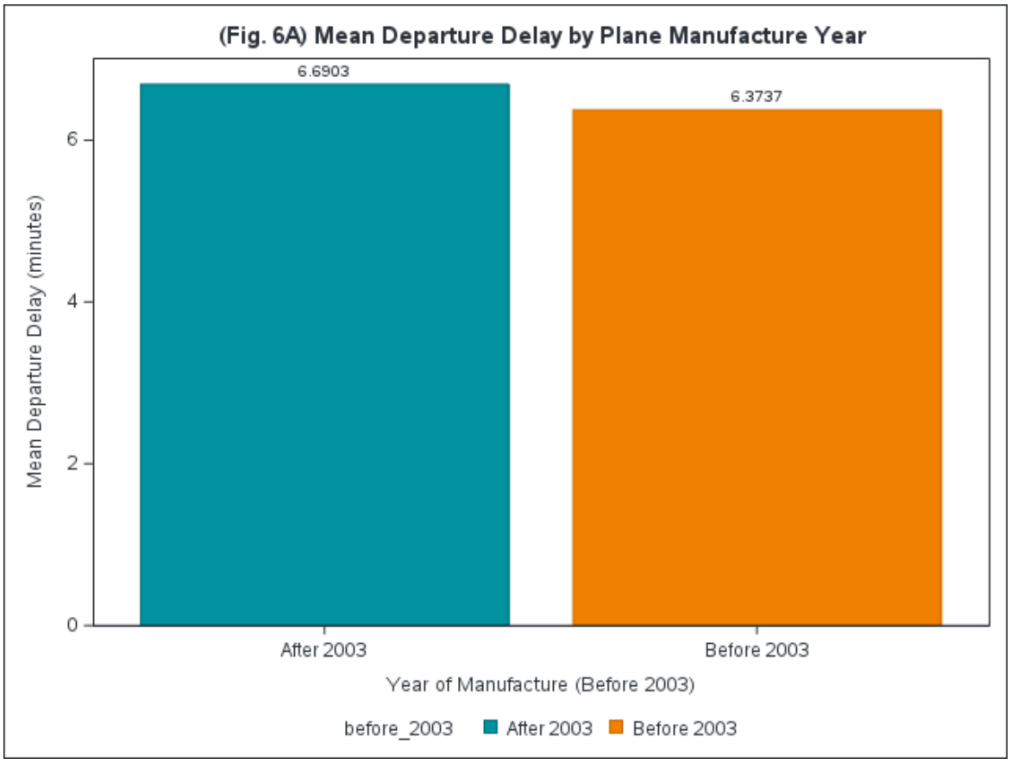
Plane Analysis

We will now incorporate **planes** data in our delay analysis. We will first see if planes manufactured more than a decade ago have greater frequency of delays.

6. Year Manufactured

Figure 6A compares the mean departure delay for planes manufactured before and after 2003. Contrary to what one might expect, the analysis reveals that planes manufactured after 2003 experience slightly higher average departure delays compared to those manufactured before 2003. Planes manufactured before 2003 have a mean departure delay of 12.88 minutes, indicating relatively shorter delays while planes manufactured after 2003 show a slightly higher mean departure delay of 13.43 minutes. The difference in the two can be considered negligible, and we can conclude that a planes ‘age’ has no effect on its expected delay time.

```
185  /* 6A. Mean Departure Delay by Plane Manufacture Year */
186  PROC SQL;
187    CREATE TABLE flights_year_mean AS
188    SELECT
189      (CASE WHEN year_manufactured <= 2003 THEN 'Before 2003' ELSE 'After 2003' END) AS before_2003,
190      MEAN(dep_delay, 'na.rm'='YES') AS dep_delay
191    FROM flights_data
192    WHERE NOT MISSING(year_manufactured)
193    ORDER BY before_2003;
194  QUIT;
195  PROC SGPLOT DATA=flights_year_mean;
196      VBAR before_2003 / RESPONSE=dep_delay STAT=MEAN DATALABEL GROUP=before_2003;
197      XAXIS LABEL="Year of Manufacture (Before 2003)";
198      YAXIS LABEL="Mean Departure Delay (minutes)";
199      TITLE "(Fig. 6A) Mean Departure Delay by Plane Manufacture Year";
200  RUN;
```



7. Model

We will now look at different models of aircrafts and their ability to recover departure delay time in transit.

Fig. 7A highlights which plane models are able to recover from a delayed departure and still arrive on time or early. Of all 120 types of planes that flew from NYC, only 6 types were able to make up departure delay time in transit. These were the *737-524*, *737-8FH*, *757-212*, *757-2B7*, *767-201*, and *777-222* (Fig. 7A). Of those, the *757-212* had the earliest average arrival time of 23 minutes ahead of schedule (Fig. 7B). These models likely benefit from better operational efficiency, higher cruising speeds, or optimized flight paths that allow for time recovery. The ability to arrive on time or early despite a late departure is crucial for maintaining overall schedule integrity, especially for airlines with tight schedules or during peak travel seasons. These plane models are particularly valuable in high-stakes situations where connecting flights or turnaround times are critical. Their ability to exceed expectations, even after delays, offers a competitive advantage in maintaining passenger satisfaction and operational efficiency.

Fig. 7D provides an overarching comparison between planes that experience early arrivals and those with mean delays. The results show a clear divide between models that tend to be on time or early after a delayed departure and those that face persistent delays, and highlights models *310Q*, *737-990*, *767-432ER*, *767-324*, *MYSTERE FALCON 900*, and *DC-7BF* as the most notorious for delays even after an on time or early departure. This comparison underlines the need for strategic deployment based on performance in delay-prone scenarios. The contrasting performance highlights that

airlines should invest in using more resilient plane models for routes where on-time arrivals are a priority. This ensures minimal schedule disruption, even under challenging conditions.

Fig. 7E introduces a carrier-specific dimension, showing which plane models are most frequently used by major carriers. Interestingly, the analysis reveals that none of the most popular plane models used by these carriers are among the top performers in early arrival after delayed departures. This could suggest that while these planes may be favored for other operational reasons (cost-efficiency, capacity, etc.), they are not the best choice for mitigating the impact of delays. No carrier's most popular plane model is one of the early arrival leaders in delayed departure scenarios.

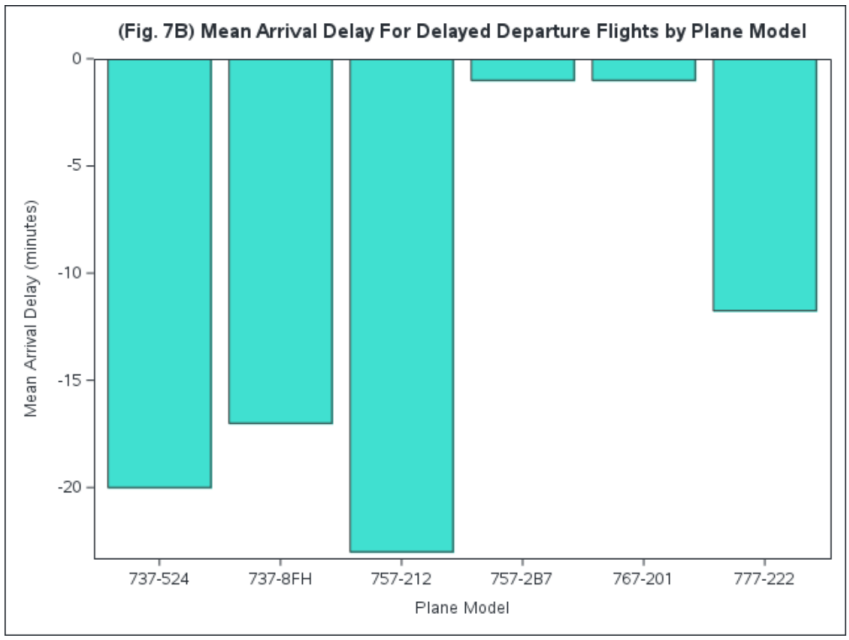
In summation, plane models that manage on-time or early arrivals despite delays (as seen in 7A and 7C) are operationally efficient and should be prioritized for delay-prone routes. Airlines should consider diversifying their fleets to include plane models with better time recovery performance for critical routes. Models that frequently suffer from longer delays (7B) may need to be deployed on less time-sensitive routes. The lack of alignment between the most popular plane models (7E) and the best performers in early arrivals (7C) suggests a missed opportunity for airlines to optimize their fleet based on performance under delay conditions.

```
/* 7A. Models of Planes with On Time or Early Arrival After Late Departure */
PROC SQL;
    CREATE TABLE flights_model AS
    SELECT model,
           MEAN(arr_delay) AS arr_delay
    FROM flights_data
    WHERE model IS NOT NULL AND dep_delay > 0
    GROUP BY model;
QUIT;
PROC CONTENTS DATA=flights_model;
RUN;
DATA flights_model_bar;
    SET flights_model;
    IF arr_delay <= 0;
RUN;
ODS HTML FILE='flights_model_bar.html';
PROC PRINT DATA=flights_model_bar NOOBS;
    TITLE "(Fig. 7A) Models of Planes with On Time or Early Arrival After Late Departure";
RUN;
ODS HTML CLOSE;
```

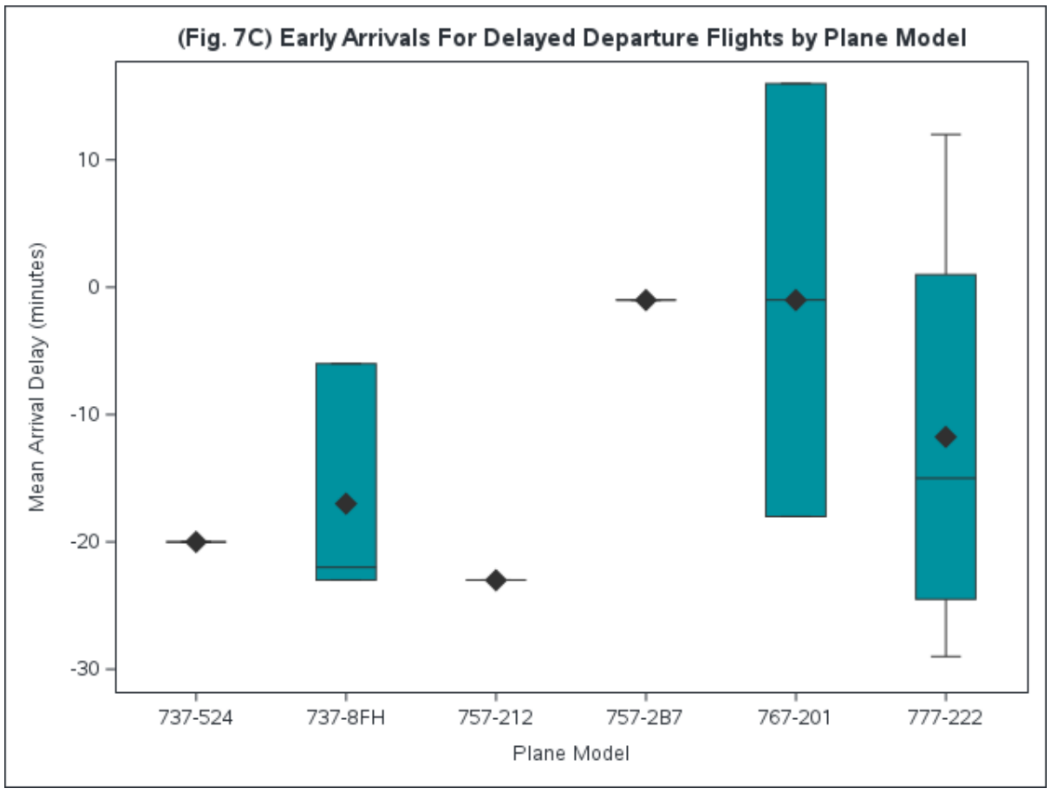
(Fig. 7A) Models of Planes with On Time or Early Arrival After Late Departure

model	arr_delay
737-524	-20.00
737-8FH	-17.00
757-212	-23.00
757-2B7	-1.00
767-201	-1.00
777-222	-11.75

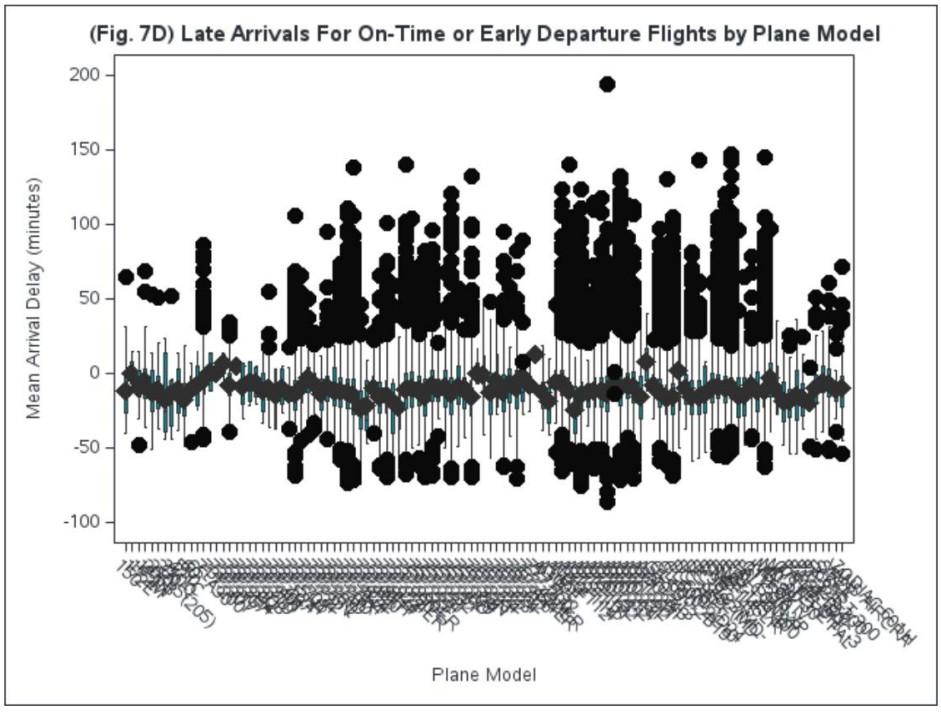
```
/* 7B. Mean Arrival Delay For Delayed Departure Flights by Plane Model */
PROC SGPLOT DATA=flights_model_bar;
    VBAR model / RESPONSE=arr_delay FILLATTRS=(COLOR=turquoise) STAT=MEAN;
    TITLE "(Fig. 7B) Mean Arrival Delay For Delayed Departure Flights by Plane Model";
    XAXIS LABEL="Plane Model";
    YAXIS LABEL="Mean Arrival Delay (minutes)";
RUN;
```




```sql
/* 7C. Early Arrivals For Delayed Departure Flights by Plane Model */
PROC SQL;
  CREATE TABLE flights_model_box AS
  SELECT model,
         arr_delay
    FROM flights_data
   WHERE model IS NOT NULL
     AND dep_delay > 0
     AND model IN (SELECT model FROM flights_model_bar)
  ;
QUIT;
PROC SGPLOT DATA=flights_model_box;
  VBOX arr_delay / CATEGORY=model;
  XAXIS LABEL="Plane Model";
  YAXIS LABEL="Mean Arrival Delay (minutes)";
  TITLE "(Fig. 7C) Early Arrivals For Delayed Departure Flights by Plane Model";
RUN;
```



```sas
/* 7D. Late Arrivals For On-Time or Early Departure Flights by Plane Model */
PROC SQL;
    CREATE TABLE flights_model_box_worst6 AS
    SELECT model, MEAN(arr_delay) AS arr_delay
    FROM flights_data
    WHERE model IS NOT NULL AND arr_delay > 0 AND dep_delay <= 0
    GROUP BY model
    ORDER BY arr_delay DESC;
QUIT;
PROC SORT DATA=flights_data;
    BY model;
RUN;
PROC SORT DATA=flights_model_box_worst6;
    BY model;
RUN;
DATA flights_model_box_worst6_data;
    MERGE flights_data(in=a) flights_model_box_worst6(in=b);
    BY model;
    IF a AND b AND dep_delay <= 0;
RUN;
PROC SGPLOT DATA=flights_model_box_worst6_data;
    VBOX arr_delay / CATEGORY=model;
    TITLE "(Fig. 7D) Late Arrivals For On-Time or Early Departure Flights by Plane Model";
    XAXIS LABEL="Plane Model";
    YAXIS LABEL="Mean Arrival Delay (minutes)";
RUN;
```




```sas
/* 7E. Most Popular Model For Each Carrier  */
PROC SQL;
  CREATE TABLE flights_model_carrier AS
  SELECT carrier, model, COUNT(*) AS count
  FROM flights_data
  WHERE model IS NOT NULL AND carrier IS NOT NULL
  GROUP BY carrier, model
  ORDER BY carrier, count DESC;
QUIT;
PROC SORT DATA=flights_model_carrier;
  BY carrier descending count;
RUN;
DATA flights_model_carrier_top;
  SET flights_model_carrier;
  BY carrier;
  RETAIN top_model top_count;
  IF FIRST.carrier THEN top_model = model;
  IF FIRST.carrier THEN top_count = count;
  IF LAST.carrier;
RUN;
ODS HTML FILE="flights_model_carrier_top.html";
PROC PRINT DATA=flights_model_carrier_top NOOBS LABEL;
  TITLE "(Fig. 7E) Most Popular Model For Each Carrier";
RUN;
ODS HTML CLOSE;
```

(Fig. 7E) Most Popular Model For Each Carrier

carrier	model	count	top_model	top_count
9E	CL-600-2B19	6836	CL-600-2D24	10580
AA	172E	13	767-223	4257
AS	737-8FH	16	737-890	346
B6	A321-231	53	A320-232	34063
DL	A330-323	1	MD-88	10191
EV	EMB-145	274	EMB-145LR	28027
F9	A319-112	18	A320-214	617
FL	AT-5	3	717-200	2774
HA	A330-243	342	A330-243	342
MQ	CF-5D	103	G1159B	486
OO	CL-600-2B19	3	CL-600-2C10	25
UA	737-524	4	737-824	13809
US	757-2B7	2	A319-112	5844
VX	A319-115	94	A320-214	4859
WN	737-3A4	1	737-7H4	10389
YV	MYSTERE FAL	4	CL-600-2C10	311

8. Number of Engines

We will now explore the relationship between the number of engines on an aircraft and its ability to manage delayed departures, focusing on the recovery time and arrival performance. Specifically, we will examine whether planes with a different number of engines have any significant impact on how well flights recover from late departures and arrive either on time or early.

While three-engine planes showed the best performance in Fig. 8A, with an average arrival time of 11 minutes early, it is important to note that this conclusion is drawn from a small dataset of only 7 data points. One of these planes, the *MYSTERE FALCON 900* (Fig. 8C), was found to be one of the latest arrivals for on-time or early departures in the earlier analysis (**7D**), which complicates the interpretation of its performance. The limited data for three-engine planes suggests that these results might not be fully representative or as reliable as those for planes with more data.

The overwhelming majority of planes in the dataset, with 276,939 values (Fig. 8B), have two engines. These planes showed a solid performance in terms of time recovery but didn't outperform the three-engine planes in Fig. 8A. However, given the size of the data sample, the results for two-engine planes are more statistically reliable. These planes form the backbone of commercial aviation, with a balance of speed, fuel efficiency, and reliability, making them well-suited for various routes.

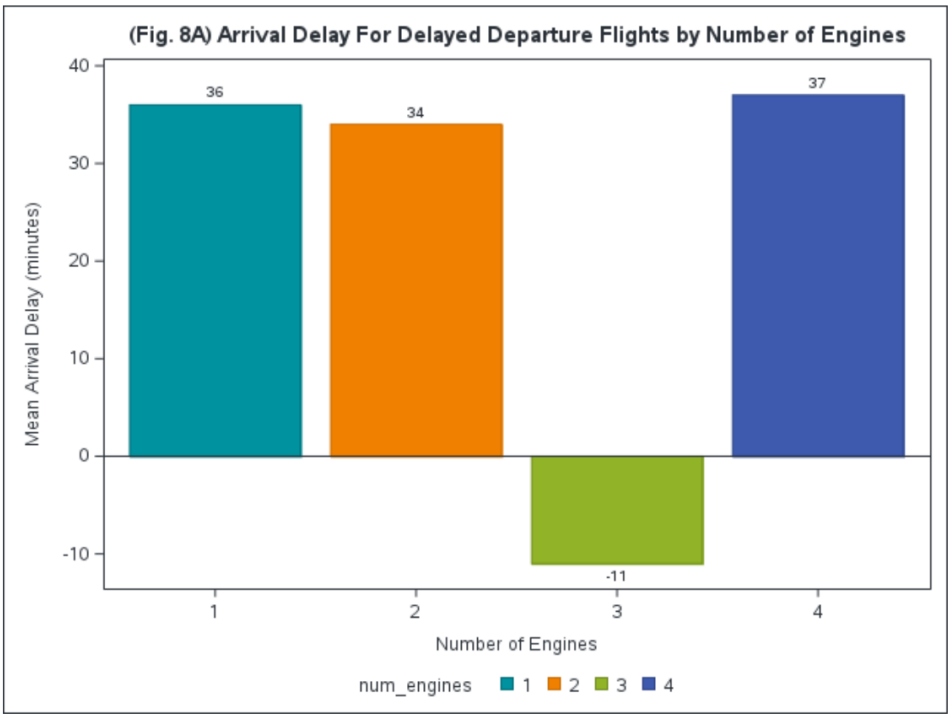
There are 1,938 values for one-engine planes and 133 values for four-engine planes (Fig. 8B). The one-engine planes are less common in commercial passenger flights and typically serve specific purposes like cargo or private flights, so their performance is less central to the overall analysis. Four-engine planes, primarily used for long-haul and high-capacity flights, performed less favorably in terms of time recovery. Given the small number of data points for both one-engine and four-engine planes, their statistical significance is limited, and results should be interpreted with caution.

The standout performance of three-engine planes in arriving early may not reflect a general trend due to the very small dataset. Additionally, the fact that the *MYSTERE FALCON 900* was identified as one of the latest arrivals for on-time or

early departure flights (**7D**) suggests that even within this limited dataset, the performance of three-engine planes may be inconsistent.

While Fig. 8A shows that three-engine planes had the best performance in terms of recovery from delayed departures, this insight must be taken with caution due to the very small dataset of only 7 values. The results are more reliable for two-engine planes, which dominate the dataset and provide consistent, predictable performance in time recovery. For airlines, the data suggests that two-engine planes remain the most dependable for maintaining schedules, given their large sample size and solid performance. While three-engine planes show promise, further data is needed to draw conclusions about their efficiency.

```
/* 8A. Arrival Delay For Delayed Departure Flights by Number of Engines */
PROC SQL;
  CREATE TABLE flights_num_engine_bar AS
  SELECT num_engines, ROUND(MEAN(arr_delay), 1) AS arr_delay
  FROM flights_data
  WHERE num_engines IS NOT NULL AND dep_delay > 0
  GROUP BY num_engines;
QUIT;
PROC SGPLOT DATA=flights_num_engine_bar;
  VBAR num_engines / RESPONSE=arr_delay STAT=SUM GROUP=num_engines DATALABEL;
  XAXIS LABEL="Number of Engines";
  YAXIS LABEL="Mean Arrival Delay (minutes)";
  TITLE "(Fig. 8A) Arrival Delay For Delayed Departure Flights by Number of Engines";
RUN;
```



```sql
/* 8B. Count of Each Number of Engines */
PROC SQL;
    CREATE TABLE num_engines_table AS
    SELECT num_engines, COUNT(*) AS count
    FROM flights_data
    WHERE num_engines > 0
    GROUP BY num_engines;
QUIT;
ODS HTML FILE="num_engines_table.html";
PROC PRINT DATA=num_engines_table;
    TITLE "(Fig. 8B) Count of Each Number of Engines";
RUN;
ODS HTML CLOSE;
```

(Fig. 8B) Count of Each Number of Engines

Obs	num_engines	count
1	1	2014
2	2	282005
3	3	7
4	4	144

```sql
/* 8C. Models of Three Engine Planes */
PROC SQL;
  CREATE TABLE models_table AS
  SELECT model, COUNT(*) AS count
  FROM flights_data
  WHERE num_engines = 3
  GROUP BY model;
QUIT;
ODS HTML FILE="models_table.html";
PROC PRINT DATA=models_table NOOBS;
    TITLE "(Fig. 8C) Models of Three Engine Planes";
RUN;
ODS HTML CLOSE;
```

(Fig. 8C) Models of Three Engine Planes

model	count
A330-223	3
MYSTERE FAL	4

9. Engine Type

Fig. 9A shows that Turbo-fan engines dominate the fleet, reflecting their widespread use in modern commercial aviation due to their fuel efficiency and operational flexibility. These engines power a broad range of aircraft, from short-haul to long-haul flights, and are preferred for their ability to balance fuel economy and high thrust. In contrast, Turbo-jet engines are far less common, indicating their use in niche markets such as older aircraft or specific military applications. Turbo-jets are less fuel-efficient and noisier compared to turbo-fans, explaining their reduced presence in commercial fleets.

Fig. 9B highlights that Turbo-jet engines have the lowest mean arrival delay at 26.6 minutes, which is notable given their limited use. Several factors likely contribute:

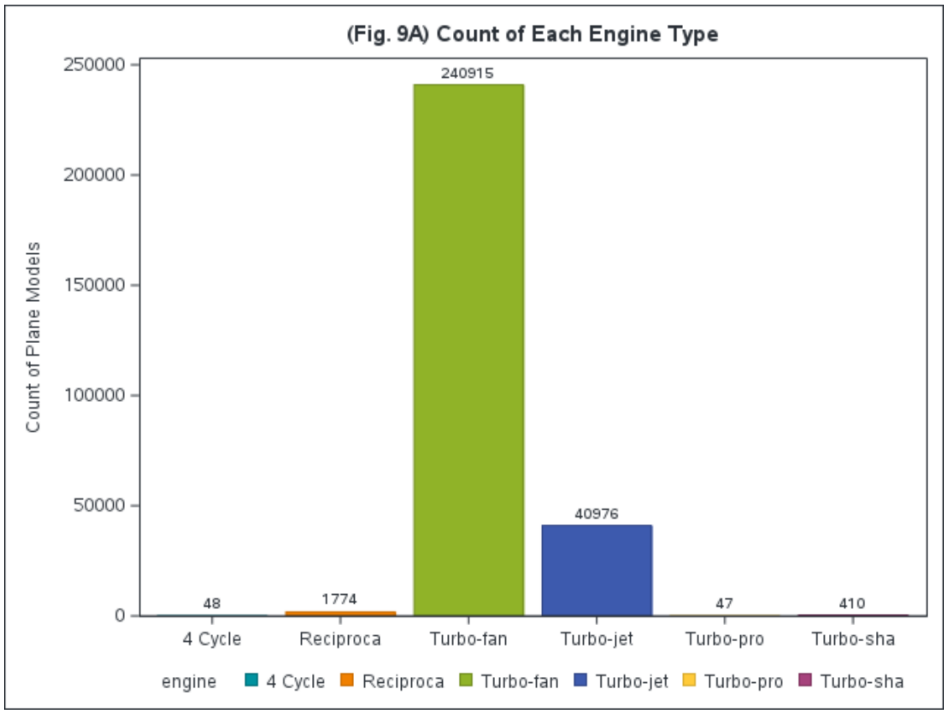
- Operational Context: Turbo-jet-powered flights may operate in less congested airports or on shorter routes, reducing their exposure to delays.
- Flight Characteristics: These engines might be used in specialized aircraft with faster recovery from delays due to fewer passengers or less complex logistics.

Conversely, Turbo-fan engines exhibit slightly higher mean delays, potentially due to:

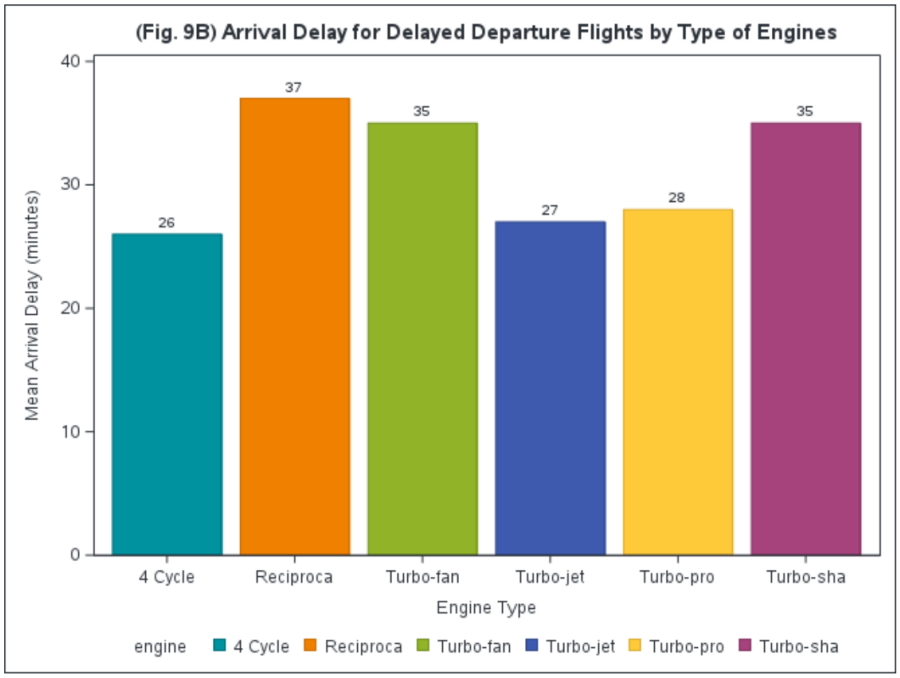
- Longer Routes and Utilization: These engines are widely used in larger aircraft on high-traffic, long-haul routes, which are more prone to delays due to air traffic congestion and logistical complexity.
- Operational Complexity: Larger planes face greater scheduling challenges and longer turnaround times, contributing to their higher delay averages.

The lower delays for turbo-jet engines likely reflect their use in less complex, shorter routes rather than superior engine performance. Turbo-fan engines, despite having higher delays, remain critical to commercial aviation due to their versatility in both short and long-haul operations. The data suggests that engine performance alone doesn't dictate delay recovery; route complexity, airspace congestion, and operational demands also play a significant role. To improve delay management, it is essential to consider how these broader factors intersect with engine type and aircraft operation.

```sql
/* 9A. Count of Each Engine Type */
PROC SQL;
    CREATE TABLE engines_table AS
    SELECT engine, COUNT(*) AS count
    FROM flights_data
    WHERE engine IS NOT NULL
    GROUP BY engine;
QUIT;
PROC SGPLOT DATA=engines_table;
    VBAR engine / RESPONSE=count GROUP=engine DATALABEL;
    XAXIS LABEL="Engine Type" DISPLAY=(NOLABEL);
    YAXIS LABEL="Count of Plane Models";
    TITLE "(Fig. 9A) Frequency of Each Engine Type";
RUN;
```



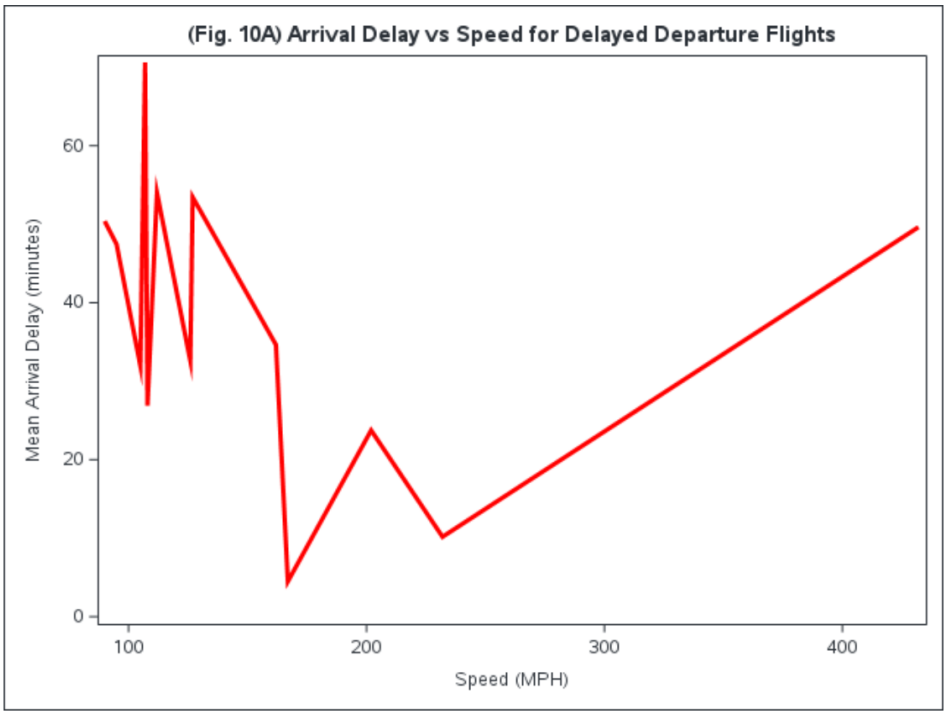
```sql
/* 9B. Arrival Delay for Delayed Departure Flights by Type of Engines */
PROC SQL;
   CREATE TABLE flights_engine_arr_delay AS
   SELECT engine, ROUND(MEAN(arr_delay), 1) AS arr_delay
   FROM flights_data
   WHERE engine IS NOT NULL AND dep_delay > 0
   GROUP BY engine;
QUIT;
PROC SGPLOT DATA=flights_engine_arr_delay;
   VBAR engine / RESPONSE=arr_delay STAT=SUM GROUP=engine DATALABEL;
   XAXIS LABEL="Engine Type";
   YAXIS LABEL="Mean Arrival Delay (minutes)";
   TITLE "(Fig. 9B) Arrival Delay for Delayed Departure Flights by Type of Engines";
RUN;
```



10. Speed

In Fig. 10A, the relationship between arrival delay and flight speed is explored for flights that departed late. The data shows a noticeable variation in mean arrival delay as speed increases. However, it's important to highlight that there are only 13 unique speed values represented in the dataset, which limits the robustness of the findings. This small sample size makes it difficult to generalize conclusions about the impact of speed on arrival delays across different flights. The figure shows that while some speeds are associated with shorter delays, the overall pattern is not strong enough to establish a clear, reliable relationship due to the limited and potentially non-representative dataset. As such, further investigation with more comprehensive data would be needed to verify any potential trends.

```sql
/* 10A. Arrival Delay vs Speed for Delayed Departure Flights */
PROC SQL;
   CREATE TABLE flights_speed_arr_delay AS
   SELECT speed, MEAN(arr_delay) AS arr_delay
   FROM flights_data
   WHERE speed IS NOT NULL AND dep_delay > 0
   GROUP BY speed;
QUIT;
PROC SQL;
   SELECT COUNT(*) AS row_count, COUNT(DISTINCT speed) AS unique_speeds
   FROM flights_speed_arr_delay;
QUIT;
PROC SGPLOT DATA=flights_speed_arr_delay;
    SERIES X=speed Y=arr_delay / LINEATTRS=(COLOR=RED);
    TITLE "(Fig. 10A) Arrival Delay vs Speed for Delayed Departure Flights";
    XAXIS LABEL="Speed (MPH)";
    YAXIS LABEL="Mean Arrival Delay (minutes)";
RUN;
```



Conclusion

This study has provided a comprehensive analysis of the factors contributing to flight delays across New York City's three major airports: John F. Kennedy International (JFK), LaGuardia (LGA), and Newark Liberty International (EWR). By leveraging the nycflights13 dataset, we identified key patterns and trends in flight performance, helping us understand how environmental conditions, aircraft characteristics, and operational challenges interplay to influence punctuality. The results not only confirm existing knowledge about the complexity of air travel delays but also present unique findings with implications for airport management, airlines, and policy planning.

Key Findings

1. *Airport Comparison*: Newark (*EWR*) emerged as the airport with the highest frequency of delays, reflecting its status as a busy hub managing both international and domestic operations. The airport's dual-function increases operational stress, leading to frequent bottlenecks. In contrast, LaGuardia (*LGA*) and JFK exhibited fewer delays, which could be attributed to better scheduling practices, fewer long-haul disruptions at LGA, and improved international traffic management at *JFK*. However, *JFK*'s high variability

in flight volume suggests that even major airports with robust infrastructures remain vulnerable to seasonal surges in demand.

2. *Weather Impact*: Weather was identified as a significant external factor influencing flight delays. Extreme temperatures—particularly below 30°F or above 85°F—were associated with higher delay probabilities. In winter, operations such as de-icing cause longer ground times, and in summer, heat-induced turbulence can delay departures and arrivals. Moreover, seasonal variability adds complexity to flight schedules, as winter storms or summer thunderstorms can ripple across multiple airports, amplifying delays throughout the network.
3. *Aircraft Age and Performance*: One of the more surprising insights was that newer aircraft (post-2003) exhibited slightly longer average delays (13.43 minutes) than older planes (12.88 minutes). This finding challenges the assumption that modern aircraft always deliver superior punctuality. The reason might lie in factors like airline scheduling policies or over-reliance on newer aircraft for high-traffic routes, which are more prone to congestion-related delays. A detailed engine analysis showed that while two-engine aircraft remain the most common and consistent performers, three-engine planes recorded the best recovery times, arriving 11 minutes earlier on average than scheduled. However, the limited sample size for three-engine planes makes it difficult to draw robust conclusions about their performance.

Implications and Recommendations

The data reveals that delays are shaped by a combination of environmental, infrastructural, and operational factors. To improve punctuality, airlines and airport authorities can adopt several measures:

- *Dynamic Scheduling*: Adjusting flight schedules based on historical weather patterns and airport congestion trends can mitigate delays.
- *Data-Driven Maintenance*: Insights from aircraft performance should guide predictive maintenance to ensure optimal fleet readiness.
- *Resource Allocation*: Airports like Newark, with heavier traffic, could benefit from enhanced air traffic management systems to reduce bottlenecks.

Additionally, this analysis suggests that collaboration among airports, airlines, and meteorological services is essential. Integrating machine learning models to predict delays and disruptions based on multiple variables could further improve decision-making processes.

Future Research

The limitations of this study, such as small sample sizes for certain aircraft types and limited time scope (2013), point to areas for future research. Expanding the analysis to multiple years and including more airports would enhance the validity of the findings. Moreover, exploring passenger satisfaction in relation to these delays could provide a more holistic understanding of the impact on airline operations.

Biography

I am **Saransh Rakshak**, a data scientist with a strong foundation in data analysis, machine learning, and programming. I have a Bachelor of Arts in Data Science from the *University of California, Berkeley*, and am currently pursuing a Master of Science in Data Science and Analytics at *Clemson University*.

I have always been fascinated by airplane engineering and amazed at airports efficiency despite serving high volumes of people. As a result, when I came upon this dataset I was instantly interested!

SAS Code

```
/* SAS Data Analysis: Factors Affecting Flight Delays and Flight Times */

libname flights '/home/sraksha/final_project/data';

/* Importing Data */

proc import datafile="/home/sraksha/final_project/data/flights_processed.csv"
   OUT=flights_data
   DBMS=CSV
   REPLACE;
   GETNAMES=YES;
RUN;

/* Adding was_delayed boolean column */

DATA flights_data;
   SET flights_data;
   was_delayed = (arr_delay > 0 or dep_delay > 0);
RUN;
```

```
PROC PRINT DATA=flights_data (OBS=5); RUN;



/* 1A. Carrier Delay Count */
PROC SQL;
   CREATE TABLE flights_carrier_count AS
   SELECT carrier, COUNT(*) AS count
   FROM flights_data
   WHERE was_delayed = 1
   GROUP BY carrier
   ORDER BY count DESC;
QUIT;
PROC PRINT DATA=flights_carrier_count; RUN;
PROC SGPLOT DATA=flights_carrier_count;
   VBAR carrier / RESPONSE=count DATALABEL GROUP=carrier STAT=SUM;
   TITLE "(Fig. 1A) Count of Delay by Carrier";
   XAXIS LABEL="Carrier";
   YAXIS LABEL="Count";
RUN;

/* 1B. Carrier Delay Frequency */
PROC SQL;
 CREATE TABLE flights_delay_freq AS
 SELECT carrier,
      SUM(was_delayed) AS delayed,
      SUM(CASE WHEN was_delayed = 0 THEN 1 ELSE 0 END) AS notdelayed,
      SUM(was_delayed) / (SUM(was_delayed) + SUM(CASE WHEN was_delayed = 0 THEN 1 ELSE 0 END)) AS delay_freq
 FROM flights_data
 GROUP BY carrier;
QUIT;
PROC PRINT DATA=flights_delay_freq; RUN;
PROC SGPLOT DATA=flights_delay_freq;
   VBAR carrier / RESPONSE=delay_freq DATALABEL GROUP=carrier STAT=SUM;
   TITLE "(Fig. 1B) Frequency of Delay by Carrier";
   XAXIS LABEL="Carrier";
   YAXIS LABEL="Frequency";
RUN;
```

```
/* 2A. Count of delays grouped by origin */
PROC SQL;
 CREATE TABLE flights_origin_count AS
 SELECT origin,
      COUNT(*) AS count
 FROM flights_data
 WHERE was_delayed = 1
 GROUP BY origin
 ORDER BY count DESC;
QUIT;
PROC SGPLOT DATA=flights_origin_count;
 VBAR origin / RESPONSE=count GROUP=origin DATALABEL;
 TITLE "(Fig. 2A) Count of Delays by Origin";
 XAXIS LABEL="Origin";
 YAXIS LABEL="Count";
 STYLEATTRS DATACOLORS=(blue);
RUN;

/* 2B. Origin Delay Frequency */
PROC SQL;
 CREATE TABLE flights_origin_freq AS
 SELECT origin,
      SUM(was_delayed) AS delayed,
      SUM(CASE WHEN was_delayed = 0 THEN 1 ELSE 0 END) AS notDelayed,
      100 * SUM(was_delayed) /
         (SUM(was_delayed) + SUM(CASE WHEN was_delayed = 0 THEN 1 ELSE 0 END)) AS delay_freq
 FROM flights_data
 GROUP BY origin;
QUIT;
PROC SGPLOT DATA=flights_origin_freq;
 VBAR origin / RESPONSE=delay_freq GROUP=origin DATALABEL;
 TITLE "(Fig. 2B) Frequency of Delay by Origin";
 XAXIS LABEL="Origin";
 YAXIS LABEL="% of Delayed Flights";
RUN;


/* 3A. Mean Departure Delay by Temperature */
PROC SQL;
 CREATE TABLE flights_temp_mean AS
```

```
  SELECT temp,
        MEAN(dep_delay) AS mean_dep_delay
  FROM flights_data
  WHERE temp IS NOT NULL
  GROUP BY temp;
QUIT;
PROC SGPLOT DATA=flights_temp_mean;
  SERIES X=temp Y=mean_dep_delay / LINEATTRS=(THICKNESS=2 COLOR=blue);
  TITLE "(Fig. 3A) Mean Departure Delay by Temperature";
  XAXIS LABEL="Temperature (°F)" GRID;
  YAXIS LABEL="Mean Departure Delay (minutes)" GRID;
RUN;

/* 3B. Mean Departure Delay by Temperature and Carrier */
PROC SQL;
   CREATE TABLE flights_temp_carrier_mean AS
   SELECT carrier,
        temp,
        MEAN(dep_delay) AS mean_dep_delay
   FROM flights_data
   WHERE temp IS NOT NULL
   GROUP BY carrier, temp;
QUIT;
PROC SORT DATA=flights_temp_carrier_mean;
   BY carrier;
RUN;
PROC SGPLOT DATA=flights_temp_carrier_mean;
   SERIES X=temp Y=mean_dep_delay;
   TITLE "(Fig. 3B) Mean Departure Delay by Temperature and Carrier";
   XAXIS LABEL="Temperature (°F)";
   YAXIS LABEL="Mean Departure Delay (minutes)";
   BY carrier;
RUN;

/* 3C. Mean Departure Delay by Temperature and Origin */
PROC SQL;
   CREATE TABLE flights_temp_origin_mean AS
   SELECT origin, temp,
        MEAN(dep_delay) AS mean_dep_delay
   FROM flights_data
```

```
    WHERE temp IS NOT NULL
    GROUP BY origin, temp;
QUIT;
PROC SGPLOT DATA=flights_temp_origin_mean;
    SERIES X=temp Y=mean_dep_delay / GROUP=origin LINEATTRS=(THICKNESS=2);
    XAXIS LABEL="Temperature (°F)";
    YAXIS LABEL="Mean Departure Delay (minutes)";
    TITLE "(Fig. 3C) Mean Departure Delay by Temperature and Origin";
RUN;


/* 4A. Mean Departure Delay by Precipitation */
PROC SQL;
    CREATE TABLE flights_precip_mean AS
    SELECT precip,
        MEAN(dep_delay) AS mean_dep_delay
    FROM flights_data
    WHERE precip IS NOT NULL
    GROUP BY precip;
QUIT;
PROC SQL;
    SELECT COUNT(*) INTO :row_count
    FROM flights_precip_mean;
QUIT;
%PUT Number of rows in flights_precip_mean: &row_count;
PROC SGPLOT DATA=flights_precip_mean;
    SERIES X=precip Y=mean_dep_delay / LINEATTRS=(COLOR=BLACK);
    TITLE "(Fig. 4A) Mean Departure Delay by Precipitation";
    XAXIS LABEL="Precipitation (in)";
    YAXIS LABEL="Mean Departure Delay (minutes)";
RUN;


/* 5A. Mean Departure Delay by Visibility */
PROC SQL;
    CREATE TABLE flights_visib_mean AS
    SELECT visib, MEAN(dep_delay) AS mean_dep_delay
    FROM flights_data
    WHERE visib IS NOT NULL
    GROUP BY visib;
```

```
QUIT;
PROC SGPLOT DATA=flights_visib_mean;
   SERIES X=visib Y=mean_dep_delay / LINEATTRS=(COLOR=BLACK);
   XAXIS LABEL="Visibility (mi)";
   YAXIS LABEL="Mean Departure Delay (minutes)";
   TITLE "(Fig. 5A) Mean Departure Delay by Visibility";
RUN;



/* 6A. Mean Departure Delay by Plane Manufacture Year */
PROC SQL;
 CREATE TABLE flights_year_mean AS
 SELECT
  (CASE WHEN year_manufactured <= 2003 THEN 'Before 2003' ELSE 'After 2003' END) AS before_2003,
  MEAN(dep_delay, 'na.rm'='YES') AS dep_delay
 FROM flights_data
 WHERE NOT MISSING(year_manufactured)
 ORDER BY before_2003;
QUIT;
PROC SGPLOT DATA=flights_year_mean;
   VBAR before_2003 / RESPONSE=dep_delay STAT=MEAN DATALABEL GROUP=before_2003;
   XAXIS LABEL="Year of Manufacture (Before 2003)";
   YAXIS LABEL="Mean Departure Delay (minutes)";
   TITLE "(Fig. 6A) Mean Departure Delay by Plane Manufacture Year";
RUN;



/* 7A. Models of Planes with On Time or Early Arrival After Late Departure */
PROC SQL;
   CREATE TABLE flights_model AS
   SELECT model,
        MEAN(arr_delay) AS arr_delay
   FROM flights_data
   WHERE model IS NOT NULL AND dep_delay > 0
   GROUP BY model;
QUIT;
PROC CONTENTS DATA=flights_model;
RUN;
DATA flights_model_bar;
   SET flights_model;
```

```
   IF arr_delay <= 0;
RUN;
ODS HTML FILE='flights_model_bar.html';
PROC PRINT DATA=flights_model_bar NOOBS;
   TITLE "(Fig. 7A) Models of Planes with On Time or Early Arrival After Late Departure";
RUN;
ODS HTML CLOSE;

/* 7B. Mean Arrival Delay For Delayed Departure Flights by Plane Model */
PROC SGPLOT DATA=flights_model_bar;
   VBAR model / RESPONSE=arr_delay FILLATTRS=(COLOR=turquoise) STAT=MEAN;
   TITLE "(Fig. 7B) Mean Arrival Delay For Delayed Departure Flights by Plane Model";
   XAXIS LABEL="Plane Model";
   YAXIS LABEL="Mean Arrival Delay (minutes)";
RUN;

/* 7C. Early Arrivals For Delayed Departure Flights by Plane Model */
PROC SQL;
  CREATE TABLE flights_model_box AS
  SELECT model,
        arr_delay
    FROM flights_data
   WHERE model IS NOT NULL
     AND dep_delay > 0
     AND model IN (SELECT model FROM flights_model_bar)
  ;
QUIT;
PROC SGPLOT DATA=flights_model_box;
  VBOX arr_delay / CATEGORY=model;
  XAXIS LABEL="Plane Model";
  YAXIS LABEL="Mean Arrival Delay (minutes)";
  TITLE "(Fig. 7C) Early Arrivals For Delayed Departure Flights by Plane Model";
RUN;

/* 7D. Late Arrivals For On-Time or Early Departure Flights by Plane Model */
PROC SQL;
   CREATE TABLE flights_model_box_worst6 AS
   SELECT model, MEAN(arr_delay) AS arr_delay
   FROM flights_data
   WHERE model IS NOT NULL AND arr_delay > 0 AND dep_delay <= 0
```

```
   GROUP BY model
   ORDER BY arr_delay DESC;
QUIT;
PROC SORT DATA=flights_data;
   BY model;
RUN;
PROC SORT DATA=flights_model_box_worst6;
   BY model;
RUN;
DATA flights_model_box_worst6_data;
   MERGE flights_data(in=a) flights_model_box_worst6(in=b);
   BY model;
   IF a AND b AND dep_delay <= 0;
RUN;
PROC SGPLOT DATA=flights_model_box_worst6_data;
   VBOX arr_delay / CATEGORY=model;
   TITLE "(Fig. 7D) Late Arrivals For On-Time or Early Departure Flights by Plane Model";
   XAXIS LABEL="Plane Model";
   YAXIS LABEL="Mean Arrival Delay (minutes)";
RUN;

/* 7E. Most Popular Model For Each Carrier  */
PROC SQL;
 CREATE TABLE flights_model_carrier AS
 SELECT carrier, model, COUNT(*) AS count
 FROM flights_data
 WHERE model IS NOT NULL AND carrier IS NOT NULL
 GROUP BY carrier, model
 ORDER BY carrier, count DESC;
QUIT;
PROC SORT DATA=flights_model_carrier;
 BY carrier descending count;
RUN;
DATA flights_model_carrier_top;
 SET flights_model_carrier;
 BY carrier;
 RETAIN top_model top_count;
 IF FIRST.carrier THEN top_model = model;
 IF FIRST.carrier THEN top_count = count;
 IF LAST.carrier;
```

```
RUN;
ODS HTML FILE="flights_model_carrier_top.html";
PROC PRINT DATA=flights_model_carrier_top NOOBS LABEL;
 TITLE "(Fig. 7E) Most Popular Model For Each Carrier";
RUN;
ODS HTML CLOSE;



/* 8A. Arrival Delay For Delayed Departure Flights by Number of Engines */
PROC SQL;
 CREATE TABLE flights_num_engine_bar AS
 SELECT num_engines, ROUND(MEAN(arr_delay), 1) AS arr_delay
 FROM flights_data
 WHERE num_engines IS NOT NULL AND dep_delay > 0
 GROUP BY num_engines;
QUIT;
PROC SGPLOT DATA=flights_num_engine_bar;
 VBAR num_engines / RESPONSE=arr_delay STAT=SUM GROUP=num_engines DATALABEL;
 XAXIS LABEL="Number of Engines";
 YAXIS LABEL="Mean Arrival Delay (minutes)";
 TITLE "(Fig. 8A) Arrival Delay For Delayed Departure Flights by Number of Engines";
RUN;

/* 8B. Count of Each Number of Engines */
PROC SQL;
   CREATE TABLE num_engines_table AS
   SELECT num_engines, COUNT(*) AS count
   FROM flights_data
   WHERE num_engines > 0
   GROUP BY num_engines;
QUIT;
ODS HTML FILE="num_engines_table.html";
PROC PRINT DATA=num_engines_table;
   TITLE "(Fig. 8B) Count of Each Number of Engines";
RUN;
ODS HTML CLOSE;

/* 8C. Models of Three Engine Planes */
PROC SQL;
 CREATE TABLE models_table AS
```

```
 SELECT model, COUNT(*) AS count
 FROM flights_data
 WHERE num_engines = 3
 GROUP BY model;
QUIT;
ODS HTML FILE="models_table.html";
PROC PRINT DATA=models_table NOOBS;
            TITLE "(Fig. 8C) Models of Three Engine Planes";
RUN;
ODS HTML CLOSE;



/* 9A. Count of Each Engine Type */
PROC SQL;
  CREATE TABLE engines_table AS
  SELECT engine, COUNT(*) AS count
  FROM flights_data
  WHERE engine IS NOT NULL
  GROUP BY engine;
QUIT;
PROC SGPLOT DATA=engines_table;
  VBAR engine / RESPONSE=count GROUP=engine DATALABEL;
  XAXIS LABEL="Engine Type" DISPLAY=(NOLABEL);
  YAXIS LABEL="Count of Plane Models";
  TITLE "(Fig. 9A) Count of Each Engine Type";
RUN;

/* 9B. Arrival Delay for Delayed Departure Flights by Type of Engines */
PROC SQL;
  CREATE TABLE flights_engine_arr_delay AS
  SELECT engine, ROUND(MEAN(arr_delay), 1) AS arr_delay
  FROM flights_data
  WHERE engine IS NOT NULL AND dep_delay > 0
  GROUP BY engine;
QUIT;
PROC SGPLOT DATA=flights_engine_arr_delay;
  VBAR engine / RESPONSE=arr_delay STAT=SUM GROUP=engine DATALABEL;
  XAXIS LABEL="Engine Type";
  YAXIS LABEL="Mean Arrival Delay (minutes)";
  TITLE "(Fig. 9B) Arrival Delay for Delayed Departure Flights by Type of Engines";
```

```
RUN;



/* 10A. Arrival Delay vs Speed for Delayed Departure Flights */
PROC SQL;
  CREATE TABLE flights_speed_arr_delay AS
  SELECT speed, MEAN(arr_delay) AS arr_delay
  FROM flights_data
  WHERE speed IS NOT NULL AND dep_delay > 0
  GROUP BY speed;
QUIT;
PROC SQL;
  SELECT COUNT(*) AS row_count, COUNT(DISTINCT speed) AS unique_speeds
  FROM flights_speed_arr_delay;
QUIT;
PROC SGPLOT DATA=flights_speed_arr_delay;
   SERIES X=speed Y=arr_delay / LINEATTRS=(COLOR=RED);
   TITLE "(Fig. 10A) Arrival Delay vs Speed for Delayed Departure Flights";
   XAXIS LABEL="Speed (MPH)";
   YAXIS LABEL="Mean Arrival Delay (minutes)";
RUN;
```