Descriptive Statistics Using R
Assignment
Batch 23
Answers
By
Vridhi Wadhawan (12420090)

Prior to analysis, the data was cleaned and explored. The dataset, stored as flights_data, initially contained 336,776 observations across 19 variables. Given the size of the dataset, outliers were removed and further examination revealed several columns (specifically, 'year', 'day', 'hour', 'min', and others) that were deemed irrelevant to the analysis. These were subsequently removed, leaving a final dataset of 327,346 observations across 12 variables.

Q1.

To calculate flight delays, we considered both departure and arrival delays. While a two-minute departure delay might ideally result in a two-minute arrival delay, this is rarely the case. In-flight delays often contribute to the overall arrival delay. Therefore, we calculated in-flight delay by subtracting departure delay from arrival delay. Making a dataset of 327,346 observations across 13 variables.

```
Console  Terminal  Jobs
~/
> # Question 1
> # (a)
> # Calculate in-flight delay
> flight_data <- flight_data %>%
+   mutate(in_flight_delay = arr_delay - dep_delay)
>
> # Function to calculate mode
> getmode <- function(v) {
+   uniqv <- unique(v)
+   uniqv[which.max(tabulate(match(v, uniqv)))]
+ }
```

a) First we calculate the overall benchmark for all 3 categories of dep_delay, arr_dely and in_flight_delay. The overall benchmark includes the mean, median, mode, standard deviation, variance, minimum, maximum, quartile 1 , quartile3 and count of airlines under each of the three categories.

```
Console  Terminal  Jobs
~/
> # Benchmark Summary
> benchmark_summary <- data.frame(
+   Statistic = c("Mean", "Median", "Mode", "Standard Deviation", "Variance", "Min", "Q1", "Q3", "Max", "Count"),
+
+   Departure_Delay = c(mean(flight_data$dep_delay, na.rm = TRUE),
+                       median(flight_data$dep_delay, na.rm = TRUE),
+                       getmode(flight_data$dep_delay),
+                       sd(flight_data$dep_delay, na.rm = TRUE),
+                       var(flight_data$dep_delay, na.rm = TRUE),
+                       min(flight_data$dep_delay, na.rm = TRUE),
+                       quantile(flight_data$dep_delay, 0.25, na.rm = TRUE),
+                       quantile(flight_data$dep_delay, 0.75, na.rm = TRUE),
+                       max(flight_data$dep_delay, na.rm = TRUE),
+                       nrow(flight_data)),  # Fixed Count
+
+   In_Flight_Delay = c(mean(flight_data$in_flight_delay, na.rm = TRUE),
+                       median(flight_data$in_flight_delay, na.rm = TRUE),
+                       getmode(flight_data$in_flight_delay),
+                       sd(flight_data$in_flight_delay, na.rm = TRUE),
+                       var(flight_data$in_flight_delay, na.rm = TRUE),
+                       min(flight_data$in_flight_delay, na.rm = TRUE),
+                       quantile(flight_data$in_flight_delay, 0.25, na.rm = TRUE),
+                       quantile(flight_data$in_flight_delay, 0.75, na.rm = TRUE),
+                       max(flight_data$in_flight_delay, na.rm = TRUE),
+                       nrow(flight_data)),  # Fixed Count
+
+   Arrival_Delay = c(mean(flight_data$arr_delay, na.rm = TRUE),
+                     median(flight_data$arr_delay, na.rm = TRUE),
+                     getmode(flight_data$arr_delay),
+                     sd(flight_data$arr_delay, na.rm = TRUE),
+                     var(flight_data$arr_delay, na.rm = TRUE),
+                     min(flight_data$arr_delay, na.rm = TRUE),
+                     quantile(flight_data$arr_delay, 0.25, na.rm = TRUE),
+                     quantile(flight_data$arr_delay, 0.75, na.rm = TRUE),
+                     max(flight_data$arr_delay, na.rm = TRUE),
+                     nrow(flight_data))  # Fixed Count
+ )
```

```
> cat("\nBenchmark Summary:\n")

Benchmark Summary:
> print(kable(benchmark_summary, format = "markdown"))
```

Statistic	Departure_Delay	In_Flight_Delay	Arrival_Delay
Mean	12.55516	-5.659779	6.895377
Median	-2.00000	-7.000000	-5.000000
Mode	-5.00000	-9.000000	-13.000000
Standard Deviation	40.06569	18.043648	44.633292
Variance	1605.25932	325.573244	1992.130727
Min	-43.00000	-109.000000	-86.000000
Q1	-5.00000	-17.000000	-17.000000
Q3	11.00000	3.000000	14.000000
Max	1301.00000	196.000000	1272.000000
Count	327346.00000	327346.000000	327346.000000

Here, we see the following:
Average Trends:

- Departure Delay: On average, flights depart 12.56 minutes late. But the median (middle value) is -2 minutes, meaning more than half of the flights actually left earlier than scheduled. Thus, the mean is sensitive to outliers as many flights might leave on time or early, but a few significantly delayed flights increase the mean departure delay significantly. The most common delay (mode) was -5 minutes, meaning many flights left 5 minutes before schedule.
- In-Flight Delay: On average, flights make up for lost departure time (-5.66 minutes) while in the air. The median is -7 minutes, meaning most flights speed up while being in the air to reach their destination earlier than expected. The most common in-flight adjustment was -9 minutes, so it's normal for flights to speed up and arrive ahead of time.
- Arrival Delay: On average, flights arrive 6.89 minutes late. But the median (-5 minutes) shows that more than half of the flights actually arrive early. The most common arrival delay is -13 minutes, meaning many flights land 13 minutes ahead of schedule.

Thus, most flights actually depart close to or earlier than scheduled but still show an average delay because a few extremely late departures skew the data. Also in all the cases since mean is greater than the median, it is a left skewed data.

Variation & Consistency:
Standard deviation tells us how much the delay varies from the mean value. Departure delays have a high standard deviation (40 minutes), meaning delays fluctuate a lot. Arrival delays also vary widely (44 minutes), but in-flight delays are more consistent (18 minutes). Thus, flight departure and arrival delays are very unpredictable, but delays during the flight itself are more consistent.

Extreme Cases

Some flights depart 43 minutes early, and in-flight adjustments can reduce time by up to 109 minutes whereas some flights depart 1,301 minutes having the worst arrival delay of 1,272 minutes. While most flights run close to schedule, extreme delays (probably due to weather, technical issues, or air traffic) can be very severe.

Hence as a measure of the middle value for median is preferred over mean as median is in general robust to outliers.

```
Console  Terminal  Jobs
~/
> # Add airline names to flight data
> flight_data <- left_join(flight_data, airlines, by = "carrier")
> names(flight_data)[names(flight_data) == "name"] <- "carrier_name"
> # Airline-Specific Summary
> flight_data_summary <- flight_data %>%
+   group_by(carrier) %>%
+   summarize(
+     mean_dep_delay = mean(dep_delay, na.rm = TRUE),
+     median_dep_delay = median(dep_delay, na.rm = TRUE),
+     mode_dep_delay = getmode(dep_delay),
+     sd_dep_delay = sd(dep_delay, na.rm = TRUE),
+     var_dep_delay = var(dep_delay, na.rm = TRUE),
+     min_dep_delay = min(dep_delay, na.rm = TRUE),
+     q1_dep_delay = quantile(dep_delay, 0.25, na.rm = TRUE),
+     q3_dep_delay = quantile(dep_delay, 0.75, na.rm = TRUE),
+     max_dep_delay = max(dep_delay, na.rm = TRUE),
+     flights_count = n()
+   ) %>%
+   arrange(desc(median_dep_delay))
```

Now we calculate for dep_delay for each airline after we append the data to include the respective airline names through the airline dataset as mentioned in the meta data

```
Console  Terminal  Jobs
~/
> cat("\nAirline-Specific Summary:\n")

Airline-Specific Summary:
> print(kable(flight_data_summary, format = "markdown"))
```

carrier	mean_dep_delay	median_dep_delay	mode_dep_delay	sd_dep_delay	var_dep_delay	min_dep_delay	q1_dep_delay	q3_dep_delay	max_dep_delay	flights_count
FL	18.605984	1	-4	52.49106	2755.3113	-22	-4	17.00	602	3175
WN	17.661657	1	-1	43.23745	1869.4774	-13	-2	17.00	471	12044
F9	20.201175	0	0	58.40434	3411.0668	-27	-4	18.00	853	681
UA	12.016908	0	-3	35.54792	1263.6547	-20	-4	11.00	483	57782
VX	12.756646	0	-1	44.01625	1937.4307	-20	-4	8.00	653	5116
B6	12.967548	-1	-4	38.38022	1473.0409	-43	-5	12.00	502	54049
EV	19.838929	-1	-5	46.44617	2157.2471	-32	-5	25.00	548	51108
9E	16.439574	-2	-5	45.48751	2069.1139	-24	-6	16.00	747	17294
DL	9.223950	-2	-5	39.65630	1572.6218	-33	-5	5.00	960	47658
YV	18.898897	-2	-7	49.16484	2417.1813	-16	-7	22.25	387	544
AA	8.569130	-3	-4	37.36527	1396.1632	-24	-6	4.00	1014	31947
AS	5.830748	-3	-6	31.42680	987.6436	-21	-7	3.00	225	709
MQ	10.445381	-3	-7	39.02520	1522.9661	-26	-7	9.00	1137	25037
HA	4.900585	-4	-5	74.10990	5492.2775	-16	-7	-1.00	1301	342
US	3.744693	-4	-6	27.93911	780.5937	-19	-7	0.00	500	19831
OO	12.586207	-6	-6	43.06599	1854.6798	-14	-9	4.00	154	29

From above we see that:

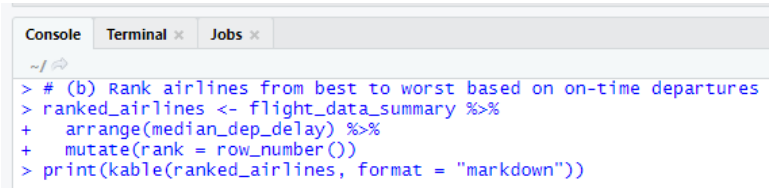
- Some airlines have more delays on average as Frontier Airlines (F9) and Envoy Air (EV) have the highest average departure delays of around 20 minutes. In contrast, Hawaiian

Airlines (HA) and US Airways (US) typically tend to depart ahead of schedule as their median departure delay is -4 minutes.

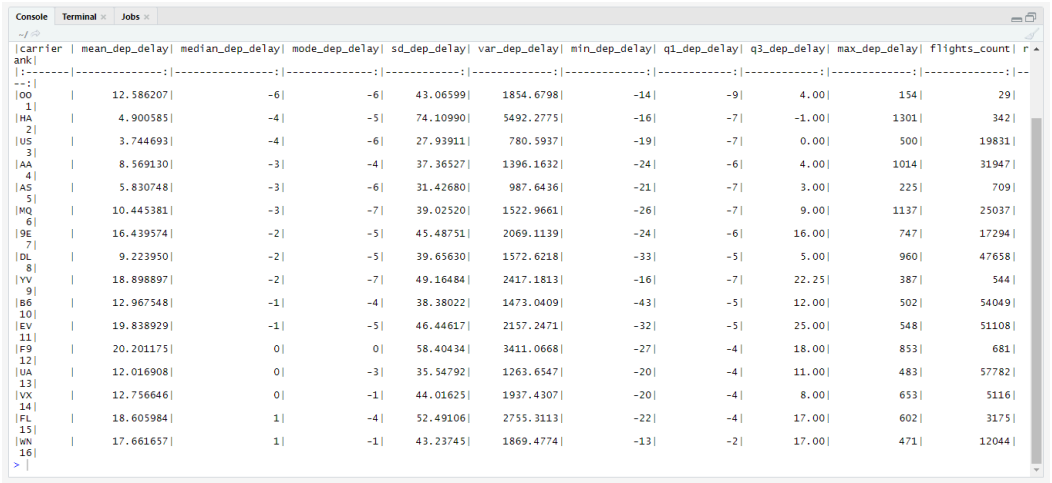
- ❖ Most flights leave on time or early as the median departure delay for most airlines is close to 0 or negative, meaning half of their flights leave early or on time.
- ❖ Extreme delays exist, but they are extremely rare as the maximum departure delays reach over 21 hours or 1301 minutes for some airlines, but since the median delay is low, these are outliers.
- ❖ Smaller airlines tend to have more variation in delays as Hawaiian Airlines (HA) has the highest variability i.e., standard deviation of 74 minutes, suggesting unpredictable delays, while US Airways (US) has the most stable schedule as it is having a standard deviation of 28 minutes.
- ❖ Regional carriers tend to be slightly worse as airlines like Envoy Air (EV), ExpressJet (OO), and Mesa Airlines (YV) show higher average departure delays compared to larger carriers like Delta (DL) or American Airlines (AA).
- ❖ All the carriers' mean is greater than their median thus they are left skewed.

Most flights depart on time or early, but a few long delays skew the averages. Larger airlines generally have more stable schedules, while smaller/regional carriers tend to face more unpredictable delays.

b) Ranking airlines based on their median departure delay time where airlines with best performance is at the top and the worst performance at the bottom.



Ranking specifically based on the median because median shows a true picture and is not skewed by extremities unlike mean.



carrier	mean_dep_delay	median_dep_delay	mode_dep_delay	sd_dep_delay	var_dep_delay	min_dep_delay	q1_dep_delay	q3_dep_delay	max_dep_delay	flights_count	rank
OO	12.586207	-6	-6	43.06599	1854.6798	-14	-9	4.00	154	29	1
HA	4.900585	-4	-5	74.10990	5492.2775	-16	-7	-1.00	1301	342	2
US	3.744693	-4	-6	27.93911	780.5937	-19	-7	0.00	500	19831	3
AA	8.569130	-3	-4	37.36527	1396.1632	-24	-6	4.00	1014	31947	4
AS	5.830748	-3	-6	31.42680	987.6436	-21	-7	3.00	225	709	5
MQ	10.445381	-3	-7	39.02520	1522.9661	-26	-7	9.00	1137	25037	6
9E	16.439574	-2	-5	45.48751	2069.1139	-24	-6	16.00	747	17294	7
DL	9.223950	-2	-5	39.65630	1572.6218	-33	-5	5.00	960	47658	8
YV	18.898897	-2	-7	49.16484	2417.1813	-16	-7	22.25	387	544	9
B6	12.967548	-1	-4	38.38022	1473.0409	-43	-5	12.00	502	54049	10
EV	19.838929	-1	-5	46.44617	2157.2471	-32	-5	25.00	548	51108	11
F9	20.201175	0	0	58.40434	3411.0668	-27	-4	18.00	853	681	12
UA	12.016908	0	-3	35.54792	1263.6547	-20	-4	11.00	483	57782	13
VX	12.756646	0	-1	44.01625	1937.4307	-20	-4	8.00	653	5116	14
FL	18.605984	1	-4	52.49106	2755.3113	-22	-4	17.00	602	3175	15
WN	17.661657	1	-1	43.23745	1869.4774	-13	-2	17.00	471	12044	16

SkyWest Airlines Inc. (OO) exhibits the best median departure time, while Southwest Airlines Co. (WN) has the poorest. However, despite SkyWest's strong median, its mean departure delay of 12.58 minutes and a high standard deviation of 43 minutes indicate significant variability, with potential delays of approximately 40 minutes. Similarly,

Southwest's mean delay of 17.66 minutes and a comparable standard deviation suggest consistent lateness, also with the possibility of delays around 40 minutes.

Additional analysis:

```
> ## For Extra Information
> flight_data_summary_inflight <- flight_data %>%
+   group_by(carrier) %>%
+   summarize(
+     mean_inflight_delay = mean(in_flight_delay, na.rm = TRUE),
+     median_inflight_delay = median(in_flight_delay, na.rm = TRUE),
+     mode_inflight_delay = getmode(in_flight_delay),
+     sd_inflight_delay= sd(in_flight_delay, na.rm = TRUE),
+     var_inflight_delay = var(in_flight_delay, na.rm = TRUE),
+     min_inflight_delay = min(in_flight_delay, na.rm = TRUE),
+     q1_inflight_delay = quantile(in_flight_delay, 0.25, na.rm = TRUE),
+     q3_inflight_delay = quantile(in_flight_delay, 0.75, na.rm = TRUE),
+     max_inflight_delay = max(in_flight_delay, na.rm = TRUE),
+     flights_count = n()
+   ) %>%
+   arrange(desc(median_inflight_delay))
>
> cat("\nAirline-Specific Summary (inflight delay):\n")

Airline-Specific Summary (inflight delay):
> print(kable(flight_data_summary_inflight, format = "markdown"))
```

carrier	mean_inflight_delay	median_inflight_delay	mode_inflight_delay	sd_inflight_delay	var_inflight_delay	min_inflight_delay	q1_inflight_delay	q3_inflight_delay	max_inflight_delay	flights_count
FL	1.5099213	-1	-2	15.84287	250.9967	-36	-9.00	9	110	3175
F9	1.7195301	-2	-7	22.48619	505.6286	-44	-14.00	13	135	681
MQ	0.3293526	-2	-6	16.83482	283.4112	-49	-11.00	8	157	25037
OO	-0.6551724	-2	-2	13.77958	189.8768	-17	-11.00	6	40	29
US	-1.6150976	-4	-8	16.18303	261.8905	-64	-12.00	6	148	19831
B6	-3.5095746	-6	-7	17.31688	299.8742	-69	-14.00	4	146	54049
EV	-4.0424982	-6	-10	15.16047	229.8397	-109	-13.00	3	150	51108
YV	-3.3419118	-6	-13	17.03417	290.1628	-38	-14.25	4	108	544
DL	-7.5796089	-9	-11	18.88386	356.6001	-79	-19.00	2	161	47658
AA	-8.2048393	-10	-10	19.24815	370.4913	-71	-21.00	2	181	31947
UA	-8.4588972	-10	-12	19.06615	363.5182	-74	-20.00	1	165	57782
WN	-8.0125374	-10	-9	16.85397	284.0564	-58	-19.00	1	131	12044
9E	-9.0599052	-11	-14	18.61474	346.5086	-64	-21.00	0	112	17294
HA	-11.8157895	-11	-11	23.19845	538.1683	-87	-25.00	4	52	342
VX	-10.9921814	-12	-19	20.60936	424.7456	-72	-24.00	1	196	5116
AS	-15.7616361	-17	-17	19.96150	398.4615	-70	-31.00	-4	65	709

```
> flight_data_summary_arr <- flight_data %>%
+   group_by(carrier) %>%
+   summarize(
+     mean_arr_delay = mean(arr_delay, na.rm = TRUE),
+     median_arr_delay = median(arr_delay, na.rm = TRUE),
+     mode_arr_delay = getmode(arr_delay),
+     sd_arr_delay= sd(arr_delay, na.rm = TRUE),
+     var_arr_delay = var(arr_delay, na.rm = TRUE),
+     min_arr_delay = min(arr_delay, na.rm = TRUE),
+     q1_arr_delay = quantile(arr_delay, 0.25, na.rm = TRUE),
+     q3_arr_delay = quantile(arr_delay, 0.75, na.rm = TRUE),
+     max_arr_delay = max(arr_delay, na.rm = TRUE),
+     flights_count = n()
+   ) %>%
+   arrange(desc(median_arr_delay))
>
> cat("\nAirline-Specific Summary (arr):\n")

Airline-Specific Summary (arr):
> print(kable(flight_data_summary_inflight, format = "markdown"))
```

carrier	mean_inflight_delay	median_inflight_delay	mode_inflight_delay	sd_inflight_delay	var_inflight_delay	min_inflight_delay	q1_inflight_delay	q3_inflight_delay	max_inflight_delay	flights_count
FL	1.5099213	-1	-2	15.84287	250.9967	-36	-9.00	9	110	3175
F9	1.7195301	-2	-7	22.48619	505.6286	-44	-14.00	13	135	681
MQ	0.3293526	-2	-6	16.83482	283.4112	-49	-11.00	8	157	25037
OO	-0.6551724	-2	-2	13.77958	189.8768	-17	-11.00	6	40	29
US	-1.6150976	-4	-8	16.18303	261.8905	-64	-12.00	6	148	19831
B6	-3.5095746	-6	-7	17.31688	299.8742	-69	-14.00	4	146	54049
EV	-4.0424982	-6	-10	15.16047	229.8397	-109	-13.00	3	150	51108
YV	-3.3419118	-6	-13	17.03417	290.1628	-38	-14.25	4	108	544
DL	-7.5796089	-9	-11	18.88386	356.6001	-79	-19.00	2	161	47658
AA	-8.2048393	-10	-10	19.24815	370.4913	-71	-21.00	2	181	31947
UA	-8.4588972	-10	-12	19.06615	363.5182	-74	-20.00	1	165	57782
WN	-8.0125374	-10	-9	16.85397	284.0564	-58	-19.00	1	131	12044
9E	-9.0599052	-11	-14	18.61474	346.5086	-64	-21.00	0	112	17294
HA	-11.8157895	-11	-11	23.19845	538.1683	-87	-25.00	4	52	342
VX	-10.9921814	-12	-19	20.60936	424.7456	-72	-24.00	1	196	5116
AS	-15.7616361	-17	-17	19.96150	398.4615	-70	-31.00	-4	65	709

Departure Delays: Most delayed airlines are Frontier Airlines (F9) and AirTran Airways (FL) have the highest average departure delays of 20 minutes.Least delayed airlines are Hawaiian Airlines (HA) and US Airways (US) experience the least departure delays of about 4–5 minutes on average. In terms of variability, some airlines have extreme outliers such as AirTran (FL) had a max departure delay of 602 minutes. Majority of flights departing on time. Many airlines have a median departure delay of 0 or negative values, meaning most flights leave on time or even earlier.

In-Flight Delays: Smallest delays mid-air airlines are Alaska Airlines (AS) and Virgin America (VX) show negative in-flight delays, meaning they often make up time en route. Biggest mid-air delays are Frontier (F9) and AirTran (FL) tend to have positive in-flight delays, suggesting they lose time while flying. Most airlines recover time as they have a negative median in-flight delay, meaning they speed up in the air to compensate for departure delays.

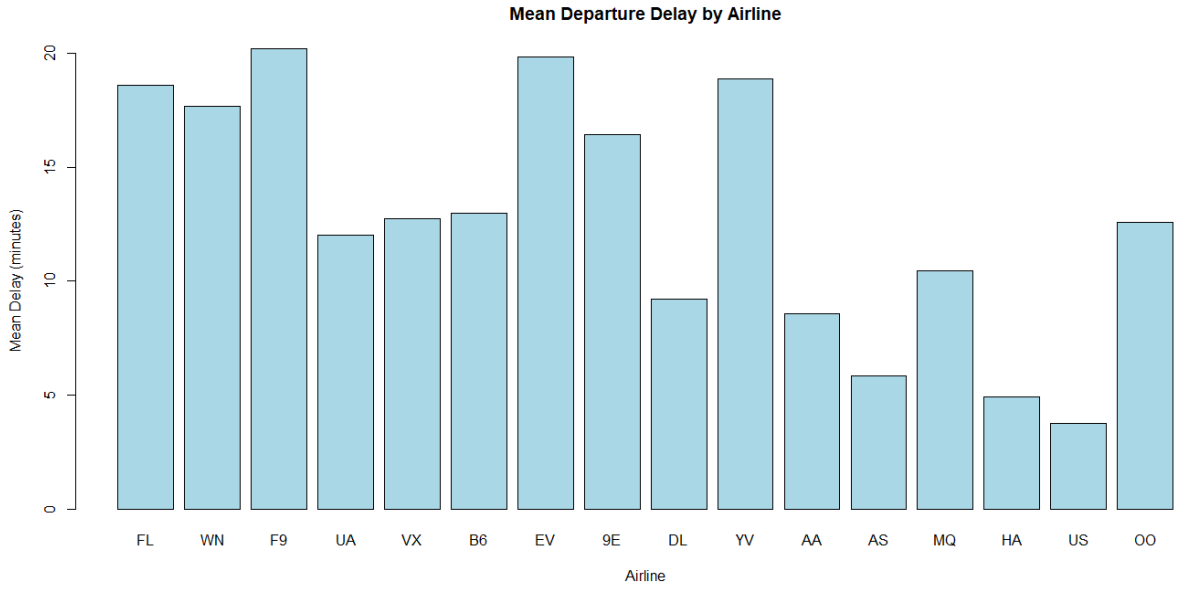
Arrival Delays: Final arrival performance of Hawaiian Airlines (HA) and US Airways (US) is the best as they tend to arrive earliest, while Frontier (F9) and AirTran (FL) often have longer arrival delays. Despite departure delays, many flights manage to arrive closer to schedule due to mid-air adjustments.

Thus, Hawaiian Airlines (HA) has the best performance across all metrics, with low departure, in-flight, and arrival delays. Frontier Airlines (F9) and AirTran (FL) struggle with delays across all stages—departures, in-flight, and arrivals.Most airlines reduce delays while in the air, helping mitigate late departures.

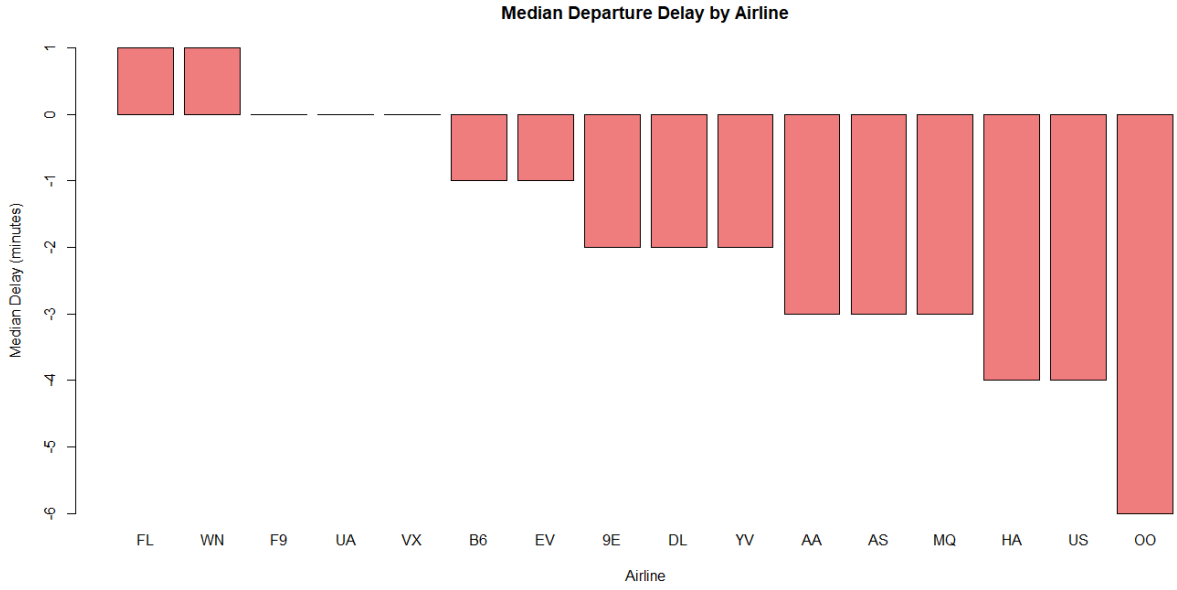
c) Mean and median delays offer different stories about the distribution of departure delays. While both mean and median delays describe central tendency, they offer different insights. The mean, or average delay, can be skewed by a few extremely long delays, making it less representative of the typical passenger experience. The median, representing the middle value, is less sensitive to these outliers and provides a more realistic view of what most passengers encounter.

SkyWest is a good example of how a few extreme delays can skew the average. As the average delay is quite high being 12.58 minutes, but their typical delay is -6 minutes implying that half their flights are early. This large difference shows how a few very long delays can distort the average, making SkyWest look worse than they are for most passengers whereas the median tells a different story as most of SkyWest flights are on time or early, but there's a risk of encountering one of those significant delays.

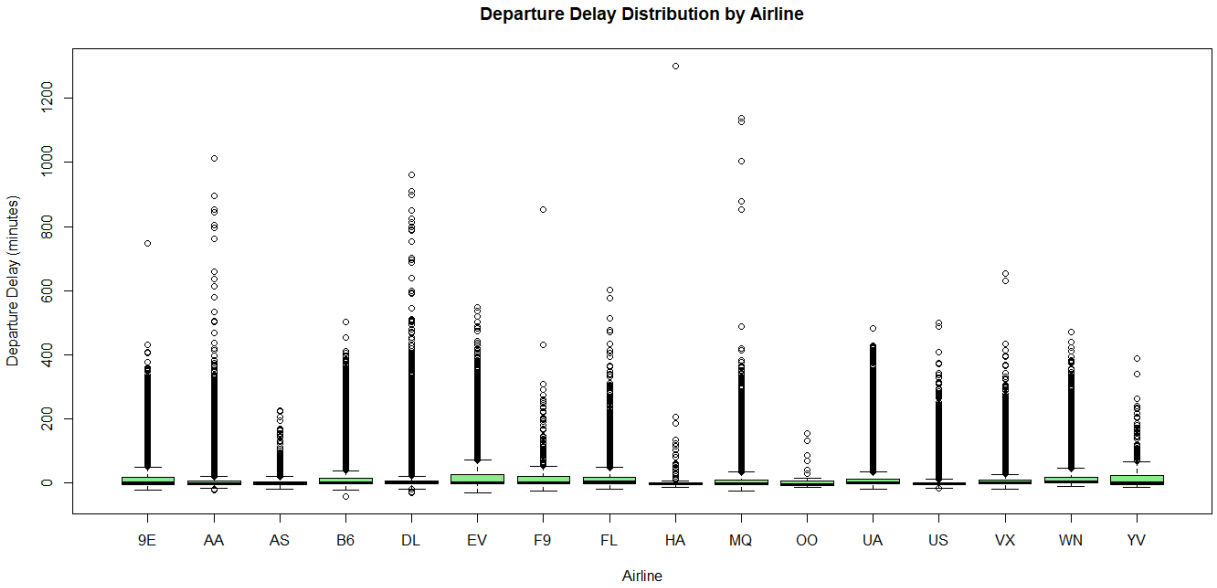
Therefore, the median delay is generally more informative for understanding typical passenger experiences with airline delays.



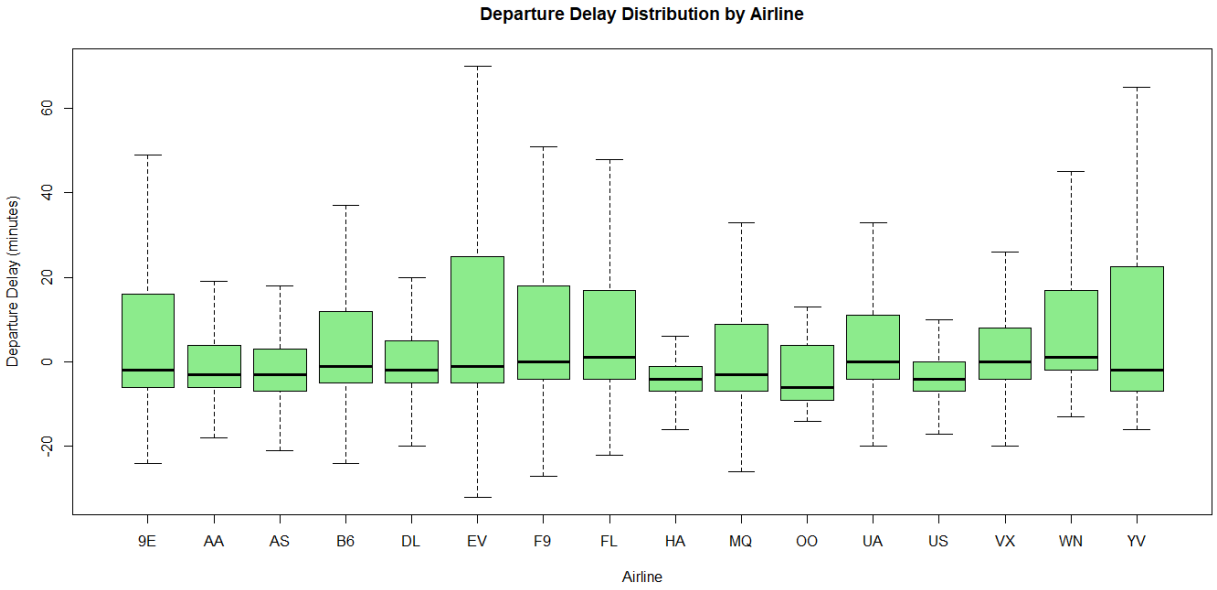
Mean Departure By Airline



Median Departure By Airline



Departure Delay Distribution by Airline (Including Outliers)



Departure Delay Distribution by Airline (Excluding Outliers)

Therefore, taking median the airline with the most delay is Southwest Airlines Co. having a median delay of 1 minute ie. departing 1 minute late for majority of its flights and the airline with the least delays is SkyWest Airlines Inc. having a median delay of -6 minutes ie. departing 6 minutes early for majority of the flights

P.S Note: Request you to kindly see the R code pdf attached for other visualizations

Q2.

a) Using the following R code, the departure delays and arrival delays for JFK, LGA, and EWR airports were calculated.

```
> # Question 2
> #(a) Compare average departure delay for JFK, LGA, and EWR
> airport_dep_delays <- flight_data %>%
+   filter(origin %in% c("JFK", "LGA", "EWR")) %>%
+   group_by(origin) %>%
+   summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE),
+             median_dep_delay = median(dep_delay, na.rm = TRUE),
+             mode_dep_delay = getmode(dep_delay),
+             sd_dep_delay = sd(dep_delay, na.rm = TRUE),
+             var_dep_delay = var(dep_delay, na.rm = TRUE),
+             min_dep_delay = min(dep_delay, na.rm = TRUE),
+             q1_dep_delay = quantile(dep_delay, 0.25, na.rm = TRUE),
+             q3_dep_delay = quantile(dep_delay, 0.75, na.rm = TRUE),
+             max_dep_delay = max(dep_delay, na.rm = TRUE),
+             flights_count = n()
+   ) %>%
+   arrange(desc(median_dep_delay))
> cat("\nAverage Departure Delay by Airport:\n")

Average Departure Delay by Airport:
> print(kable(airport_dep_delays, format = "markdown"))


|origin | mean_dep_delay| median_dep_delay| mode_dep_delay| sd_dep_delay| var_dep_delay| min_dep_delay| q1_dep_delay| q3_dep_delay| max_dep_delay| flights_count|
|:------|--------------:|----------------:|--------------:|------------:|-------------:|-------------:|------------:|------------:|------------:|-------------:|
|EWR    |       15.00911|               -1|             -4|     41.18521|      1696.221|           -25|           -4|           15|         1126|        117127|
|JFK    |       12.02361|               -1|             -3|     38.82710|      1507.544|           -43|           -5|           10|         1301|        109079|
|LGA    |       10.28658|               -3|             -5|     39.91130|      1592.912|           -33|           -6|            7|          911|        101140|
```

```
> #Arrivial Delays
> airport_arr_delays <- flight_data %>%
+   filter(origin %in% c("JFK", "LGA", "EWR")) %>%
+   group_by(origin) %>%
+   summarize(mean_arr_delay = mean(arr_delay, na.rm = TRUE),
+             median_arr_delay = median(arr_delay, na.rm = TRUE),
+             mode_arr_delay = getmode(arr_delay),
+             sd_arr_delay= sd(arr_delay, na.rm = TRUE),
+             var_arr_delay = var(arr_delay, na.rm = TRUE),
+             min_arr_delay = min(arr_delay, na.rm = TRUE),
+             q1_arr_delay = quantile(arr_delay, 0.25, na.rm = TRUE),
+             q3_arr_delay = quantile(arr_delay, 0.75, na.rm = TRUE),
+             max_arr_delay = max(arr_delay, na.rm = TRUE),
+             flights_count = n()) %>%
+   arrange(desc(median_arr_delay))
> cat("\nAverage Arrivial Delay by Airport:\n")

Average Arrivial Delay by Airport:
> print(kable(airport_arr_delays, format = "markdown"))


|origin | mean_arr_delay| median_arr_delay| mode_arr_delay| sd_arr_delay| var_arr_delay| min_arr_delay| q1_arr_delay| q3_arr_delay| max_arr_delay| flights_count|
|:------|--------------:|----------------:|--------------:|------------:|-------------:|-------------:|------------:|------------:|------------:|-------------:|
|EWR    |       9.107055|               -4|            -13|     45.52918|      2072.907|           -86|          -16|           16|         1109|        117127|
|LGA    |       5.783488|               -5|            -13|     43.86227|      1923.899|           -68|          -17|           12|          915|        101140|
|JFK    |       5.551481|               -6|            -13|     44.27745|      1960.492|           -79|          -18|           13|         1272|        109079|
>
> # Function to count outliers
> count_outliers <- function(data) {
+   Q1 <- quantile(data, 0.25, na.rm = TRUE)
+   Q3 <- quantile(data, 0.75, na.rm = TRUE)
+   IQR_value <- Q3 - Q1
+   lower_bound <- Q1 - 1.5 * IQR_value
+   upper_bound <- Q3 + 1.5 * IQR_value
+   sum(data < lower_bound | data > upper_bound, na.rm = TRUE)
+ }
```

```
> # Count outliers for each airport for dep_delay
> outlier_counts <- flight_data %>%
+   filter(origin %in% c("JFK", "LGA", "EWR")) %>%
+   group_by(origin) %>%
+   summarize(outlier_count = count_outliers(dep_delay), .groups = "drop")
> cat("\nOutlier Counts by Airport for departure delays:\n")

Outlier Counts by Airport for departure delays:
> print(kable(outlier_counts, format = "markdown"))


|origin | outlier_count|
|:------|-------------:|
|EWR    |         14998|
|JFK    |         14440|
|LGA    |         14361|
>
> # Count outliers for each airport for arr_delay
> outlier_counts <- flight_data %>%
+   filter(origin %in% c("JFK", "LGA", "EWR")) %>%
+   group_by(origin) %>%
+   summarize(outlier_count = count_outliers(arr_delay), .groups = "drop")
> cat("\nOutlier Counts by Airport for arrivial delays:\n")

Outlier Counts by Airport for arrivial delays:
> print(kable(outlier_counts, format = "markdown"))


|origin | outlier_count|
|:------|-------------:|
|EWR    |         10382|
|JFK    |          9148|
|LGA    |          8557|
```

From the above we observe that:

Newark (EWR) Has the Most Unpredictable Delays

- ❖ EWR has the highest average departure delay of 15 minutes and arrival delay of 9 minutes.
- ❖ However, the median departure delay is -1 minute and the median arrival delay is -4 minutes, implying that more than half the flights actually leave and arrive early.
- ❖ Since the maximum departure delay is 1,126 minutes,which is among some of the extreme delays which heavily skew the average.
- ❖ Departure delays have high variability since Standard Deviation is 41.18 (mins sq) and variance is 1,696 mins, indicating that delays at EWR fluctuate significantly.

JFK and LGA Have More Stable Delays

- ❖ JFK's average departure delay of 12 minutes which is slightly higher than LGA's 10 minutes, but their median departure delays are -1 and -3 minutes which show that most flights are still on time or early.
- ❖ JFK has the highest maximum departure delay 1,301 minutes, making it one of the worst case scenarios for extreme delays.
- ❖ LGA has the most reliable performance overall, with a lower max delay of 911 minutes and a lower variability since standard deviation is 39.91 (mins sq) and variance is 1,592 mins.

Arrival Delays: EWR Again Performs the Worst

- ❖ EWR's arrival delays are the most unpredictable with the highest average of 9 minutes and largest spread of extreme delay of 1,109 minutes.
- ❖ LGA and JFK have similar median arrival delays of -5 and -6 minutes, implying that flights usually arrive early or on time.
- ❖ The standard deviation and variance for EWR's arrival delays is 45.52 (mins sq) and 2,072 mins respectively are the highest, showing large inconsistencies in when flights actually arrive.

- Extreme delays are common at all airports.The presence of many outliers suggests that rare but severe delays are a major problem at all three airports, especially at EWR.
- Since, the first quartile for EWR is -4 minutes, JFK is -5 minutes and LGA is -6 minutes which implies that at least 25% of flights depart early.
- Since, the third quartile for EWR is 15 minutes, JFK is 10 minutes and LGA is 5 minutes which implies that EWR has the longest delays for 75% of flights.
- Also note that all airports have a high volume, but EWR handles the most flights.

b) Using the R code, boxplots were generated to visualize the distribution of departure and arrival delays across JFK, LGA, and EWR airports. The key observations are:

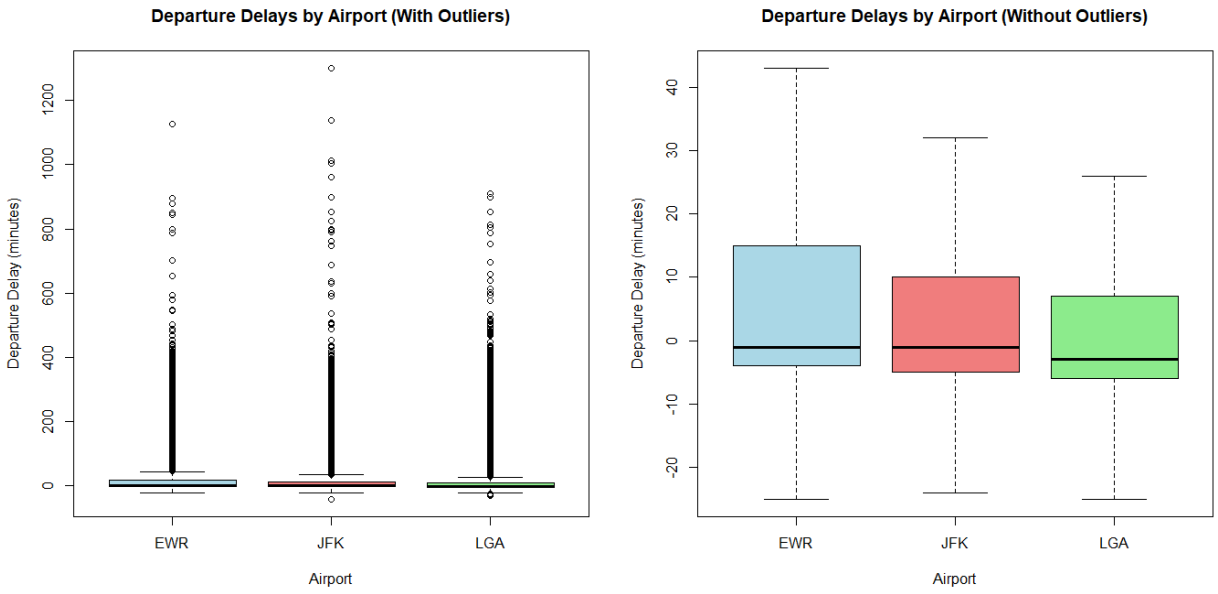
i) Outliers lie predominantly on the upper side for all three airports, indicating that delayed departures are more common than early ones.

ii) EWR has the highest median delay, more extreme outliers, and a wider interquartile range (Q1 to Q3), suggesting frequent long delays and greater variability in departure times.

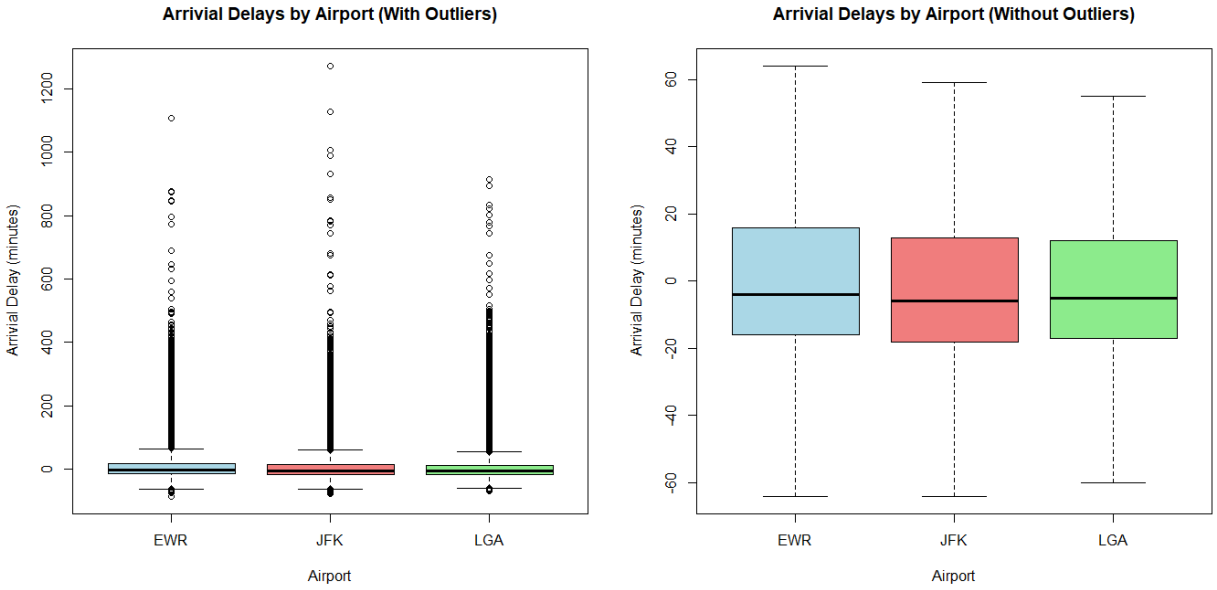
iii) JFK and LGA have similar medians, but LGA shows a slightly smaller spread in delays between Q1 and Q3, implying more consistent departures compared to JFK.

iv) Outliers are present at all three airports, indicating occasional extreme delays. However, EWR experiences the highest number of delays, with 15,145 delayed flights out of approx. 337000 total flights.

```
Console  Terminal  Jobs
~/
> # (b) Boxplot for delays by airport
> par(mfrow = c(1, 2)) # Arrange plots in 1 row, 2 columns
> boxplot(dep_delay ~ origin,
+         data = flight_data,
+         col = c("lightblue", "lightcoral", "lightgreen"),
+         main = "Departure Delays by Airport (With Outliers)",
+         xlab = "Airport",
+         ylab = "Departure Delay (minutes)",
+         outline = TRUE)
> boxplot(dep_delay ~ origin,
+         data = flight_data,
+         col = c("lightblue", "lightcoral", "lightgreen"),
+         main = "Departure Delays by Airport (Without Outliers)",
+         xlab = "Airport",
+         ylab = "Departure Delay (minutes)",
+         outline = FALSE)
> par(mfrow = c(1, 1))
> #For more information
> par(mfrow = c(1, 2)) # Arrange plots in 1 row, 2 columns
> boxplot(arr_delay ~ origin,
+         data = flight_data,
+         col = c("lightblue", "lightcoral", "lightgreen"),
+         main = "Arrivial Delays by Airport (With Outliers)",
+         xlab = "Airport",
+         ylab = "Arrivial Delay (minutes)",
+         outline = TRUE)
> boxplot(arr_delay ~ origin,
+         data = flight_data,
+         col = c("lightblue", "lightcoral", "lightgreen"),
+         main = "Arrivial Delays by Airport (Without Outliers)",
+         xlab = "Airport",
+         ylab = "Arrivial Delay (minutes)",
+         outline = FALSE)
> par(mfrow = c(1, 1))
```



Departure Delays by the Airport (Including & Excluding Outliers)



Arrival Delays by the Airport (Excluding Outliers)

c) From the above airline operators can gain the following insights:

i) EWR faces the most severe departure delays, with the highest average delay of 15.01 minutes and having the highest outliers with 14,998 delayed flights and the highest variance in delays, indicating inconsistency Thus, it requires better scheduling, operational improvements, and congestion management.

ii) LGA is the most reliable, with the shortest average delay of 12.02 minutes and fewer outliers of 14,361 delayed flights when compared to the other two, making it ideal for passengers.

iii) JFK has moderate delays of 12.02 minutes with a high number of outliers of 14,440 delayed flights, requiring peak-hour congestion management and operational optimizations.

iv) Arrival delays show a similar pattern with EWR has the highest arrival delays, with an average of 9.11 minutes and the most extreme cases. LGA and JFK have lower arrival delays, with LGA slightly outperforming JFK.

Thus, delays at the origin significantly impact arrival punctuality, emphasizing the need for improving departure processes.

To improve overall efficiency the airline operators can work on the following suggestion to reduce the delays at the airports.

i. Airlines should also work with airport authorities to improve runway efficiency and reduce bottlenecks across all three airports.

ii. Airlines should allocate buffer time for flights departing from EWR, as delays are more common.

iii. Airlines should optimize flight schedules and ground handling at EWR to reduce extreme delays

iv. Inform passengers flying from EWR about the higher probability of delays

v. Airlines should consider optimizing gate assignments, staffing, or ground operations at EWR to mitigate delays

vi. Since LGA performs best by having fewer delays which could possibly be due to better scheduling, fewer congestion issues, or improved turnaround times. Adopting the similar policies at EWR and JFK could help improve their efficiency.

Thus, we conclude that departure delays do vary by airport. From the above subset we see EWR experiencing the most delays and LGA performing the best. Airlines should focus on reducing delays at EWR while maintaining efficiency at JFK and LGA.

Q3.

a) Using the R code, the average departure delay for each month was calculated below. Further, apart from the median, mode, standard deviation and variance for each month was calculated.

Console Terminal Jobs

~/

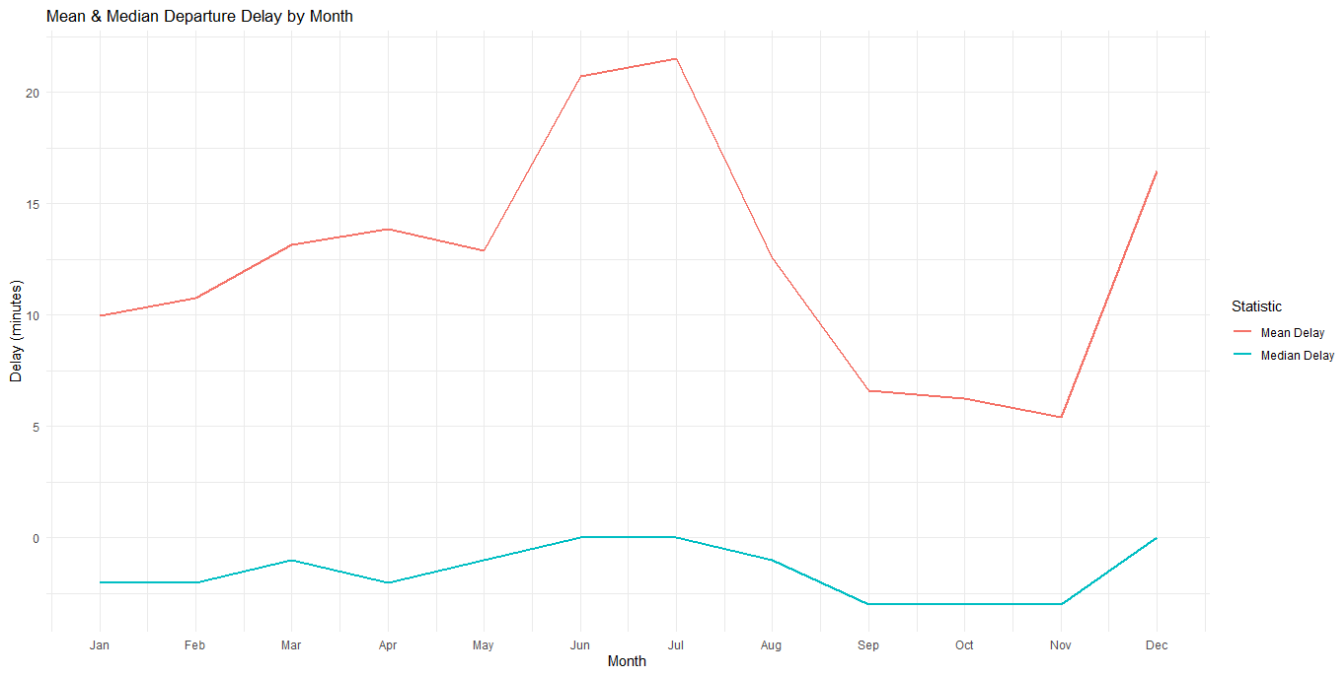
```
> # Question 3
> # (a) Calculate average departure delay for each month
> monthly_delays <- flight_data %>%
+   group_by(month) %>%
+   summarize(
+     mean_dep_delay = mean(dep_delay, na.rm = TRUE),
+     median_dep_delay = median(dep_delay, na.rm = TRUE),
+     mode_dep_delay = getmode(dep_delay),
+     sd_dep_delay = sd(dep_delay, na.rm = TRUE),
+     var_dep_delay = var(dep_delay, na.rm = TRUE),
+   ) %>%
+   arrange(month)
> cat("\nMonthly Departure Delay Summary:\n")

Monthly Departure Delay Summary:
> print(kable(monthly_delays, format = "markdown"))
```

month	mean_dep_delay	median_dep_delay	mode_dep_delay	sd_dep_delay	var_dep_delay
1	9.985491	-2	-5	36.30906	1318.3479
2	10.760239	-2	-3	36.16785	1308.1132
3	13.164289	-1	-4	40.04933	1603.9492
4	13.849187	-2	-5	42.89274	1839.7870
5	12.891709	-1	-4	39.18334	1535.3342
6	20.725614	0	-3	51.29158	2630.8259
7	21.522179	0	-3	51.24325	2625.8706
8	12.570524	-1	-3	37.59760	1413.5796
9	6.630285	-3	-5	35.47187	1258.2537
10	6.233175	-3	-5	29.66357	879.9275
11	5.420340	-3	-5	27.55734	759.4068
12	16.482161	0	-2	41.73258	1741.6079

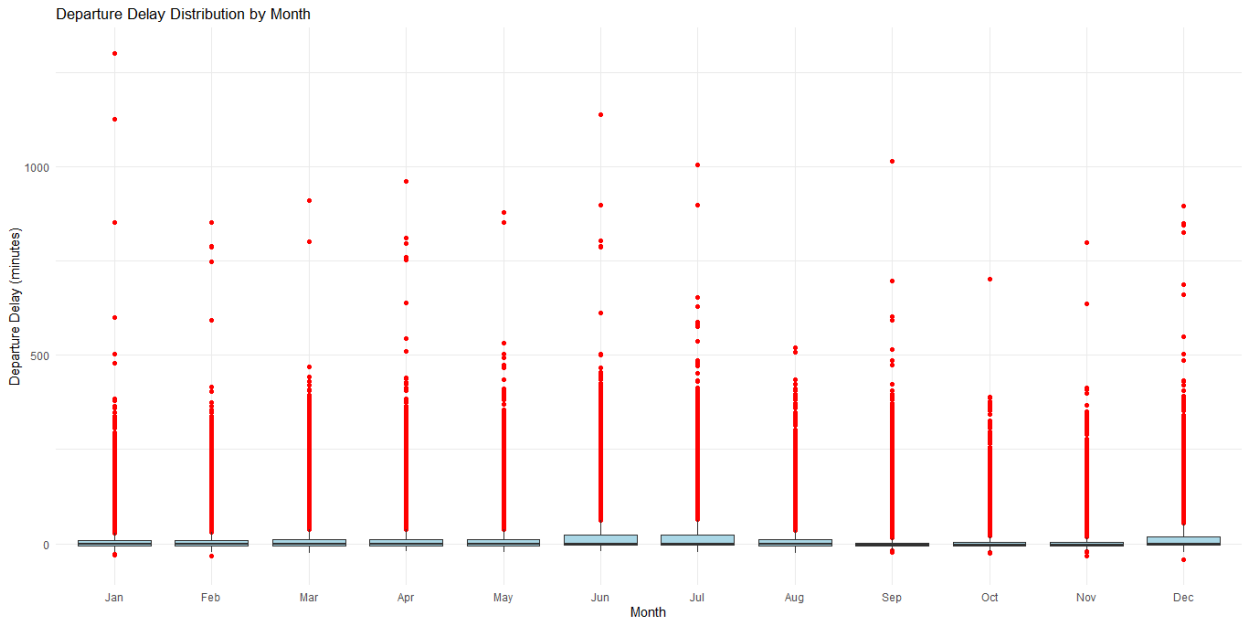
b) Using the R code, a combined mean and median line plot and boxplot is generated to visualize the delays over the months to observe the trends.

Console | Terminal | Jobs

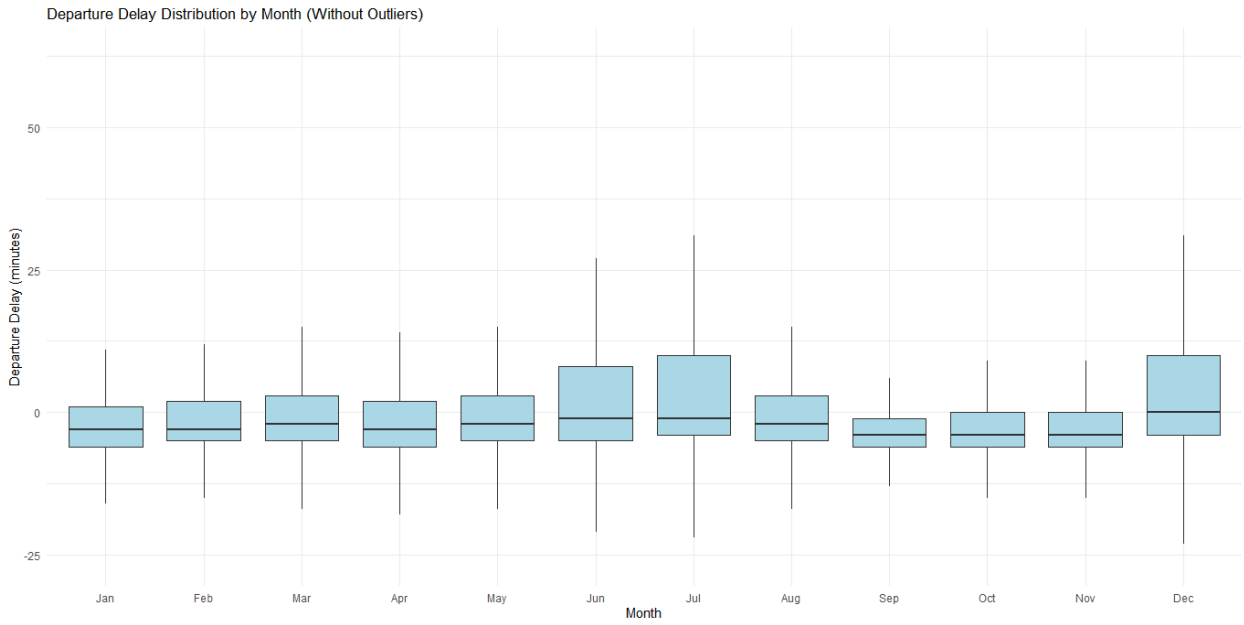
```
> # (b) Plots for departure delays by month
> # Line plot for mean and median delays
> ggplot(monthly_delays, aes(x = month)) +
+   geom_line(aes(y = mean_dep_delay, color = "Mean Delay"), size = 1) +
+   geom_line(aes(y = median_dep_delay, color = "Median Delay"), size = 1) +
+   scale_x_continuous(breaks = 1:12, labels = month.abb) +
+   labs(title = "Mean & Median Departure Delay by Month",
+        x = "Month",
+        y = "Delay (minutes)",
+        color = "Statistic") +
+   theme_minimal()
> # Boxplot for departure delays by month
> ggplot(flight_data, aes(x = factor(month), y = dep_delay)) +
+   geom_boxplot(fill = "lightblue", outlier.color = "red") +
+   scale_x_discrete(labels = month.abb) +
+   labs(title = "Departure Delay Distribution by Month",
+        x = "Month",
+        y = "Departure Delay (minutes)") +
+   theme_minimal()
> # without outliers
> ggplot(flight_data %>%
+          group_by(month) %>%
+          filter(
+            dep_delay >= quantile(dep_delay, 0.25, na.rm = TRUE) - 1.5 * IQR(dep_delay, na.rm = TRUE) &
+              dep_delay <= quantile(dep_delay, 0.75, na.rm = TRUE) + 1.5 * IQR(dep_delay, na.rm = TRUE)
+          ),
+        aes(x = factor(month), y = dep_delay)) +
+   geom_boxplot(fill = "lightblue", outlier.shape = NA) +  # Hide outliers
+   scale_x_discrete(labels = month.abb) +
+   labs(title = "Departure Delay Distribution by Month (Without Outliers)",
+        x = "Month",
+        y = "Departure Delay (minutes)") +
+   theme_minimal()
```



Line Plot Mean & Median Departure Delay by Month



Boxplot Mean & Median Departure Delay by Month (with outliers)



Boxplot Mean & Median Departure Delay by Month (without outliers)

Flight delays are influenced by seasonal demand, weather conditions, and airport operational efficiency. The summer months of June and July, along with December, experience the worst delays, averaging over 16 to 21 minutes. These months mark peak travel periods, with increased airport congestion due to vacationers, families traveling together, and holiday travelers flying home for Christmas. This surge in passengers

leads to longer security lines, baggage handling delays, and boarding inefficiencies, all of which contribute to extended departure delays.

Severe weather patterns also play a significant role in flight delays. Hurricanes and thunderstorms, which are more frequent in the summer months, can disrupt flight schedules, especially for coast-to-coast travel. In December, winter storms, icy runways, and reduced visibility due to snow or fog further delay departures. These weather-related disruptions often lead to flight rerouting, cancellations, and extended waiting times for travelers.

The off-peak months from September through November show the lowest average delays, with departures typically under six minutes late. Fewer travelers during this period lead to smoother airport operations, reduced congestion, and more efficient airline scheduling. Many flights during these months even depart on time or early, as fewer aircraft are waiting for takeoff slots.

A key observation is the difference between mean and median delays, particularly in peak months. The mean delay is significantly higher than the median, indicating the presence of extreme outlier delays. These severe delays—caused by storms, air traffic congestion, or airline operational issues—skew the average, making delays seem worse than they are for most travelers.

Thus, flight delays are highly dependent on the time of year. Peak summer months (June-July) and the December holiday season experience increased delays due to high passenger volumes, weather disruptions, and operational challenges. Conversely, the off-season months (January-May and September-November) see better on-time performance due to improved efficiency, fewer travelers, and more stable weather conditions.

Q4.

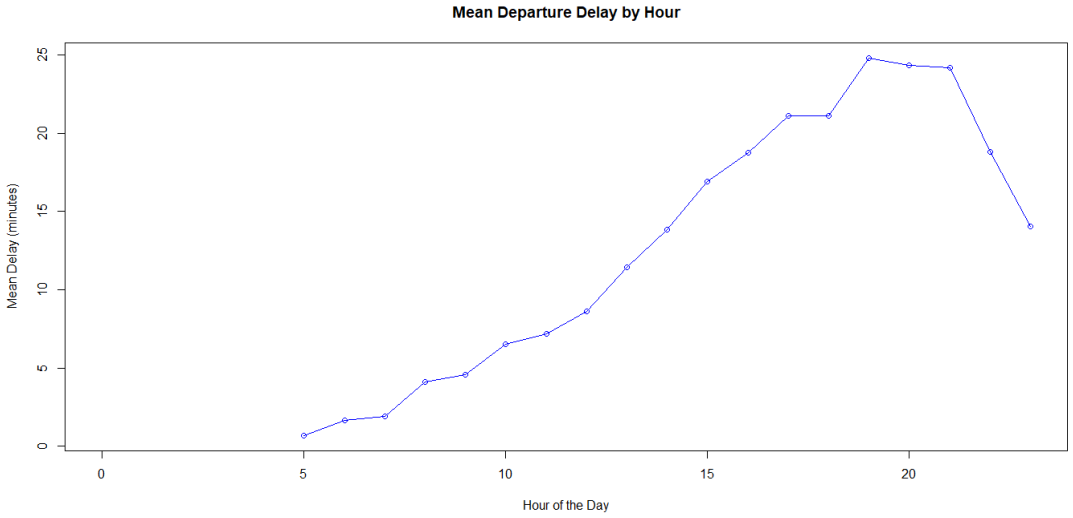
a) Using the R code, the flights are grouped by their departure hour and the average delay (mean) and median is calculated for each hour.

Console | Terminal | Jobs

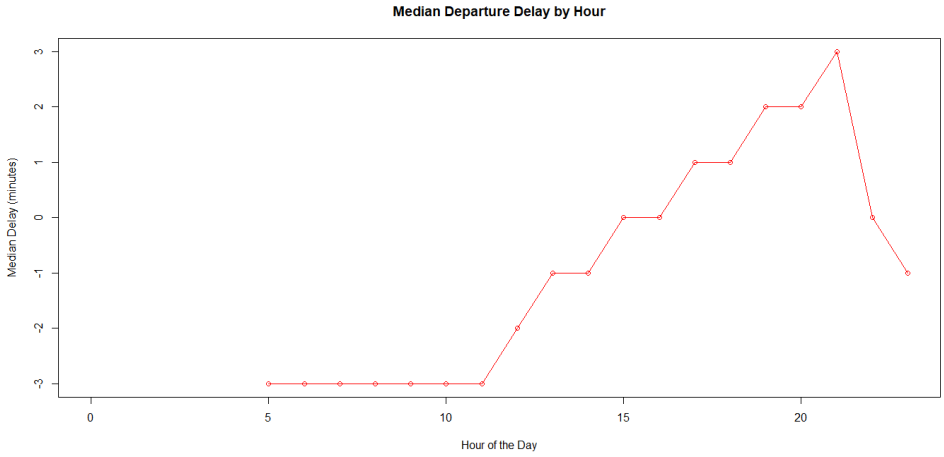
```
> # Question 4
> # (a) Group flights by departure hour and compute the average delay
> # Creating a summary dataframe for hourly delays
> hourly_delays <- data.frame(Hour = 0:23,
+                             Mean_Delay = numeric(24),
+                             Median_Delay = numeric(24))
> # Calculating mean and median departure delay for each hour
> for (i in 0:23) {
+   hour_flights <- flights[flights$hour == i, ]
+   hourly_delays$Mean_Delay[i + 1] <- mean(hour_flights$dep_delay, na.rm = TRUE)
+   hourly_delays$Median_Delay[i + 1] <- median(hour_flights$dep_delay, na.rm = TRUE)
+ }
> hourly_delays
   Hour Mean_Delay Median_Delay
1     0        NaN           NA
2     1        NaN           NA
3     2        NaN           NA
4     3        NaN           NA
5     4        NaN           NA
6     5  0.6877572           -3
7     6  1.6427956           -3
8     7  1.9140778           -3
9     8  4.1279478           -3
10    9  4.5837378           -3
11   10  6.4982946           -3
12   11  7.1916503           -3
13   12  8.6148485           -2
14   13 11.4376504           -1
15   14 13.8188742           -1
16   15 16.8945646            0
17   16 18.7570165            0
18   17 21.1006059            1
19   18 21.1100818            1
20   19 24.7847911            2
21   20 24.3041048            2
22   21 24.1957431            3
23   22 18.7910972            0
24   23 14.0171756           -1
```

b) Using the R code, line plots were generated to visualize the delays over the hours during days to observe the trends.

```
> # b) Line plots for delays by hour
> plot(hourly_delays$Hour,
+      hourly_delays$Mean_Delay,
+      type = "o",
+      col = "blue",
+      xlab = "Hour of the Day",
+      ylab = "Mean Delay (minutes)",
+      main = "Mean Departure Delay by Hour")
> plot(hourly_delays$Hour,
+      hourly_delays$Median_Delay,
+      type = "o",
+      col = "red",
+      xlab = "Hour of the Day",
+      ylab = "Median Delay (minutes)",
+      main = "Median Departure Delay by Hour")
```



Mean Departure Delay by the Hour



Median Departure Delay by the Hour

c) From the above graph and table we can divide the time into four segments and interpret the following about each:

   i) Between the hours of 12 am to 4 am we see that the data for this period is unavailable which is why we see the “Na” in the table and see no graph plotted against it. This could likely be due to lack of scheduled flights due to low demand, or potentially inconsistent data collection.
   ii) Early Bird Gets the Worm On-Time: The given data clearly shows that early morning flights between 5 am to 9 am experience significantly less delay likely due to fewer planes in the air, less traffic at the airport from both passengers and ground crew’s side, and thus, less chances of cascading delays from earlier in the day.
   iii) The Afternoon Rush: Starting from around 10 am, we start seeing delays beginning to increase noticeably as more and more flights might be scheduled during that time. This trend tends to accelerate throughout the day and has a peak in the late afternoon and evening time between 4pm to 8pm. Being a classic rush hour for air travel.
   iv) Evening Doesn't Equal Relief: While delays do decrease somewhat after the evening peak is hit between 7pm to 9 pm, the delay doesn't return to the low levels as seen during the early morning hours. Indicating a likely backlog of delayed flights and potentially crew scheduling issues that contribute to continued delays.

   We also see a large difference between the mean and median delay, especially during peak hours. The median delay is consistently much lower than the mean which suggests that while most flights might experience a relatively small delay closer to the median, a few flights that experience very long delays skew the average mean upwards.

   Looking at the given data, the best times to schedule flights in order to minimize delays would be between 5 am and 9 am assuming that there is sufficient demand for the same as well. While the absolute lowest mean delay would be at 5 am, the delays remain very low till 8 am. We see that there is a relation between the departure delays and the time of day. The trend is observed is that as the day advances, delays accumulate and often cascade into subsequent hours. After 9 am, we see that there is a steady climb in mean delays, indicating increasing congestion and likely other contributing factors as the day progresses and peaks during 7pm to 9 pm and then gradually decreases as the day comes to an end.

Q5.

a) Using the R code, we quantified the relationship, using correlation, between flight distance ie, how far a plane is going to traveling and its departure delay ie, how late the plane leaves, as follows:

```
Console  Terminal  Jobs
~/
> # Question 5
> # (a) Correlation between flight distance and departure delay
> distance_dep_correlation <- cor(flight_data$distance,
+                                 flight_data$dep_delay, use = "complete.obs")
> distance_dep_correlation
[1] -0.0216809
> print(paste("Correlation between distance and departure delay: ",
+             distance_dep_correlation))
[1] "Correlation between distance and departure delay:  -0.0216809043516393"
> distance_arr_correlation <- cor(flight_data$distance,
+                                 flight_data$arr_delay, use = "complete.obs")
> distance_arr_correlation
[1] -0.06186776
> print(paste("Correlation between distance and departure delay: ",
+             distance_arr_correlation))
[1] "Correlation between distance and departure delay:  -0.0618677560887851"
> distance_inflight_correlation <- cor(flight_data$distance,
+                                 flight_data$in_flight_delay, use = "complete.obs")
> distance_inflight_correlation
[1] -0.1048957
> print(paste("Correlation between distance and departure delay: ",
+             distance_inflight_correlation))
[1] "Correlation between distance and departure delay:  -0.10489570798776"
```

Correlation tells us how strongly two variables are related to one and other.The values range from -1 to 1 .A positive correlation i.e, a correlation closer to +1, would mean that as flight distance increases, departure delays also tend to increase. A negative correlation i.e, a correlation closer to -1, would means that as flight distance increases, departure delays tend to decrease

From the data, we checked how flight distance (short vs. long flights) is related to delays at different stages:

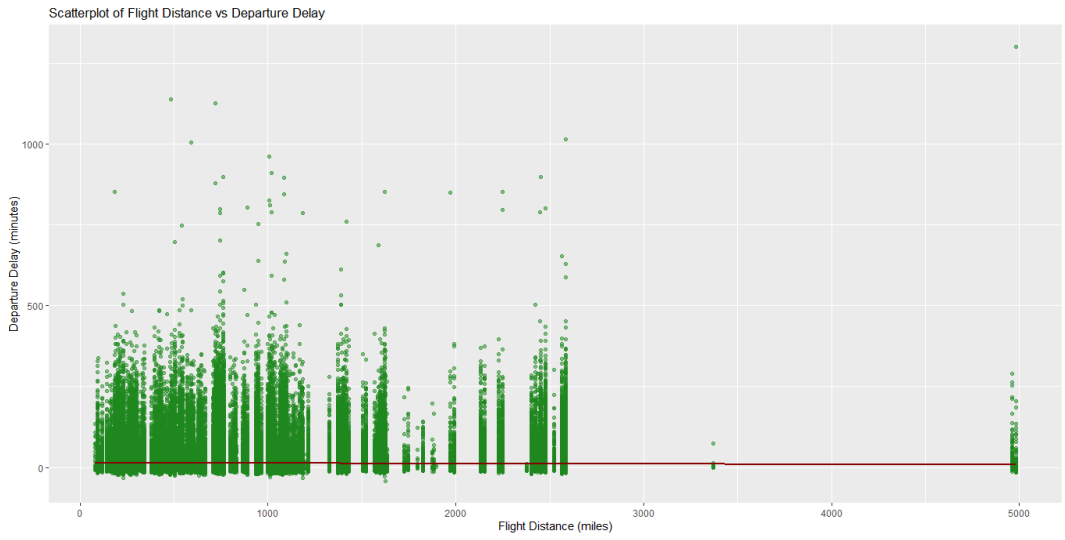
- Before Takeoff (Departure Delay) → Correlation: -0.02
- At Arrival (Arrival Delay) → Correlation: -0.06
- During Flight (In-Flight Delay) → Correlation: -0.10

Since our correlation is very close to zero and negative (-0.02 to -0.10), it suggests that longer flights are not significantly more delayed than shorter flights. If anything, the weak negative correlation hints that longer flights might have slightly fewer delays as it has to cover a longer distance, but the negative effect is so minor that it's practically negligible.

This indicates that delays are mostly caused by other factors like weather, airport congestion, time of day etc. Also there might be the case that longer flights get priority, have better planning, or less congestion at their departure times. Also one more important thing to note here is that in-flight delays have the highest negative correlation (-0.10). This means longer flights might actually make up time in the air, reducing overall delays.

b) Using the R code, scatter plots were generated to visualize the correlation between the flight distance and the departure delays.

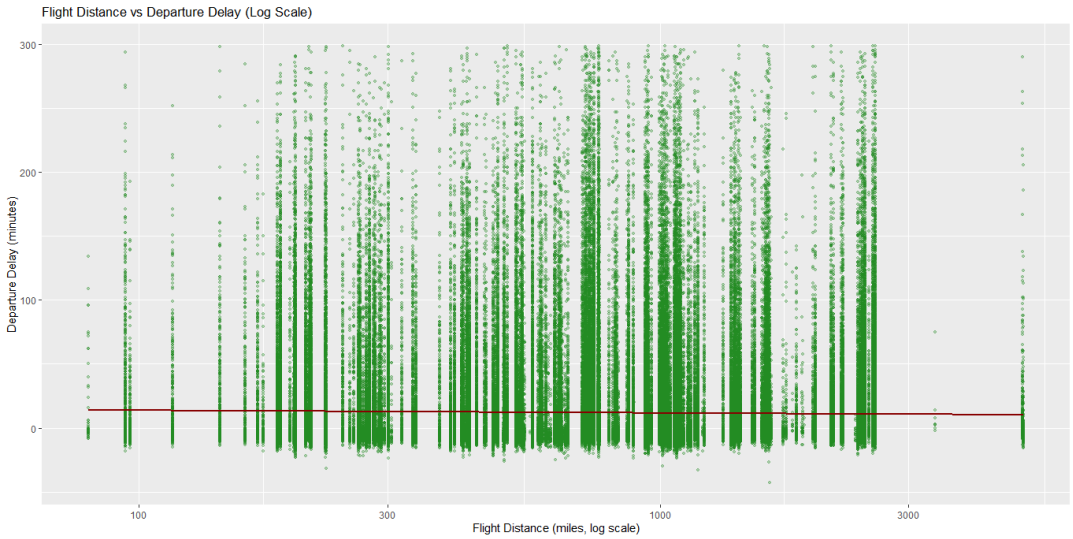
```
Console  Terminal  Jobs
~/
> # (b) Scatterplot for distance vs departure delay
> ggplot(flight_data, aes(x = distance,
+                         y = dep_delay)) +
+   geom_point(alpha = 0.5,
+              color = "forestgreen") +
+   geom_smooth(method = "lm",
+               color = "darkred",
+               se = FALSE) +
+   labs(title = "Scatterplot of Flight Distance vs Departure Delay",
+        x = "Flight Distance (miles)",
+        y = "Departure Delay (minutes)")
`geom_smooth()` using formula = 'y ~ x'
> #Scatterplot using log scale for better distance visualization with trend line
> flights_filtered <- subset(flight_data, dep_delay < 300)
> ggplot(flights_filtered,
+        aes(x = distance,
+            y = dep_delay)) +
+   geom_point(alpha = 0.3,
+              size = 1,
+              color = "forestgreen") +
+   geom_smooth(method = "lm",
+               color = "darkred",
+               se = FALSE) +
+   scale_x_log10() +
+   labs(title = "Flight Distance vs Departure Delay (Log Scale)",
+        x = "Flight Distance (miles, log scale)",
+        y = "Departure Delay (minutes)")
`geom_smooth()` using formula = 'y ~ x'
```



Flight Distance Vs Departure Delay

The above scatterplot visually represents the relationship between flight distance and its delay, where each dot represents a flight, with:

- X-axis : Flight distance (in miles)
- Y-axis : Departure delay (in minutes)



Flight Distance Vs Departure (Log scale)

The above scatterplot visually represents the relationship between flight distance and departure delay where each dot represents a flight, with:

- ❖ X-axis: Flight distance (in miles, log scale)
- ❖ Y-axis: Departure delay (in minutes)

Observation:

i. There is no clear trend as departure delays appear scattered randomly across all distances without any noticeable pattern
ii. The trend lines are flat showing no strong pattern between distance and delays
iii. If longer flights had significantly more delays, we would expect to see an upward trend i.e., a pattern where dots rise as distance increases, instead, we see no such pattern
iv. Delays are highly variable at all distances, ie, both short and long flights can be delayed, but not in any predictable way
v. There are clusters of many flights near 0 delay, showing that a significant portion departs on time

Since delays occur across all distances, factors like weather, airport congestion, airline scheduling, or air traffic control likely play a bigger role than flight length. This further supports the idea that flight length does not strongly influence departure delays.

c) When we think about flight delays, it's natural to assume that longer flights might experience longer delays—maybe due to more complex logistics, increased chances of weather disruptions as the carrier is in the air longer, or stricter safety checks. However, the data does not support this idea.

The correlation between flight distance and delay at all three stages is weakly negative (-0.02167 for departure delay, -0.06186 for arrival delay, and -0.10489 for in-flight delay).

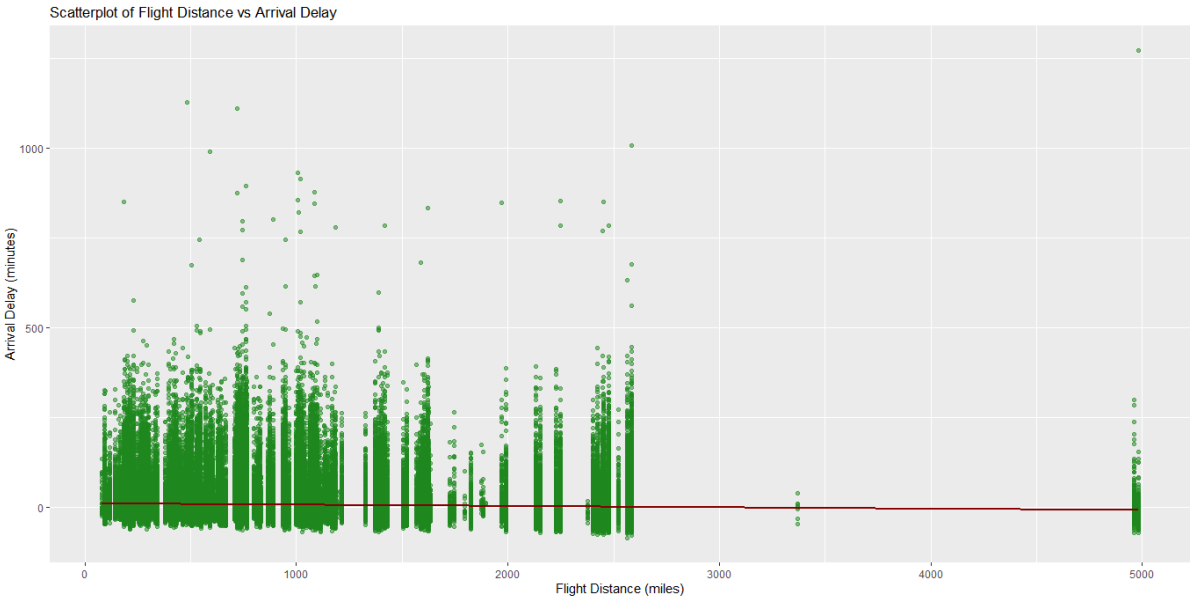
These values are close to zero, meaning that flight distance has almost no impact on delays. If anything, the slight negative correlation suggests that longer flights may actually experience slightly fewer delays, but the effect is tiny.

Delays happen randomly—some long flights are on time, some short flights are delayed, and vice versa. The scatter plot confirms that delays occur across all distances without a clear relationship, reinforcing the weak correlation. If longer flights were consistently experiencing more delays, we would expect to see a positive correlation and a noticeable upward pattern in the scatterplot, but that is not the case. Other factors like weather, airport congestion, airline scheduling, and mechanical issues are likely to play a bigger role in determining delays than flight distance. Thus, longer flights are not necessarily more prone to delays based on this dataset.

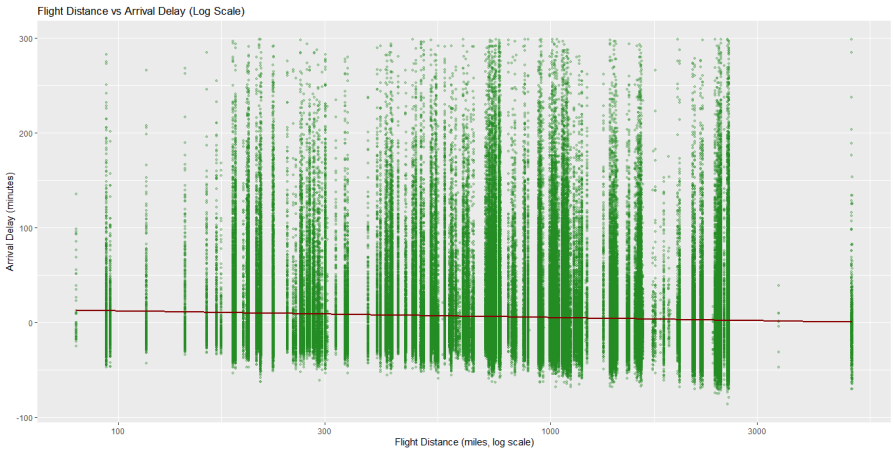
Interestingly, in-flight delays have the highest negative correlation (-0.10), suggesting that longer flights might make up for lost time while in the air. This could be due to more opportunities for speed adjustments, priority handling, or less congestion at cruising altitude.

Thus to answer the question if longer flights have longer delays we can confidently say that we do not see a correlation between them given the current data set implying that longer flights are not necessarily more prone to delays based on the dataset. Other factors, such as airport congestion, weather, airline scheduling, and mechanical issues, play a much bigger role in determining delays than the distance of the flight itself.

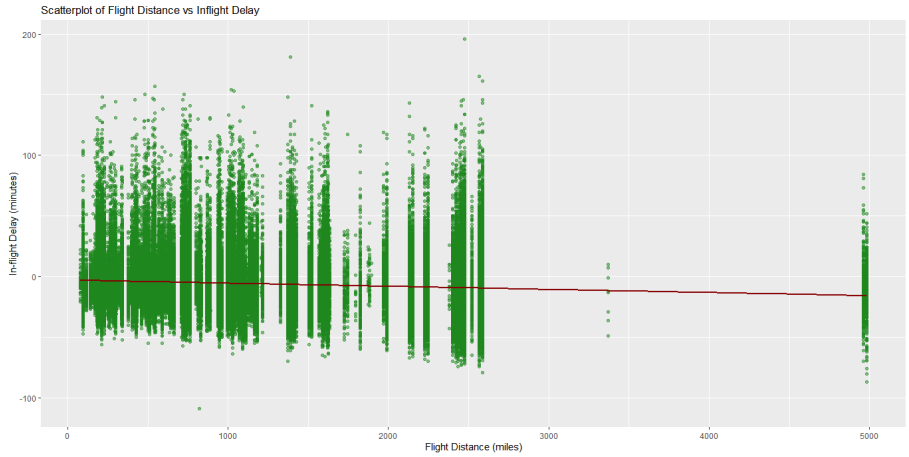
For additional information



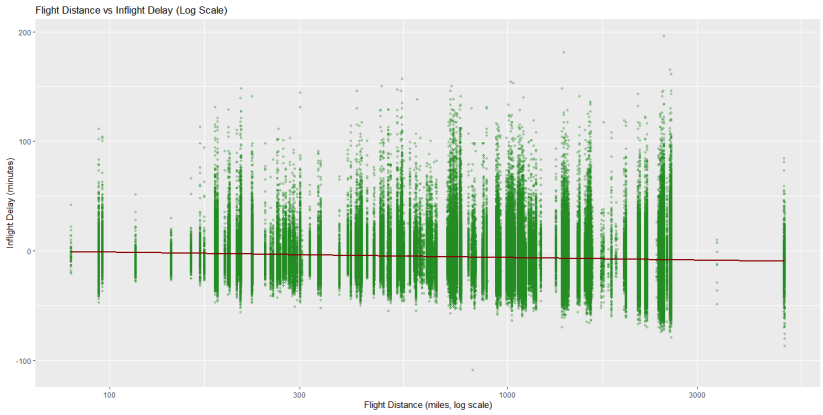
Flight Distance VS Arrival Delay



Flight Distance VS Arrival Delay (Log Scale)



Flight Distance VS Inflight Delay



Flight Distance VS Inflight Delay (Log Scale)

P.S Note: Request you to kindly see the R code pdf attached for code of the visualizations

Q6.

a) Airline efficiency here is defined by how much air time an airline takes to fly each mile of its route. This helps us understand which airlines operate their flights more effectively.

- ❖ Lower value would mean higher efficiency as the airline would covers more ground in less time
- ❖ Higher value means lower efficiency as the airline would take longer to cover same distance

Using the R code, we computed the ratio of air time to distance for each airline, as follows:

```
Console  Terminal  Jobs
~/
> # Question 6
> # (a) Compute ratio of air time to distance for each airline
> airline_efficiency <- flight_data %>%
+   mutate(air_time_per_mile = air_time / distance) %>%
+   group_by(carrier) %>%
+   summarize(avg_air_time_per_mile = mean(air_time_per_mile, na.rm = TRUE)) %>%
+   left_join(airlines, by = "carrier") %>%
+   arrange(avg_air_time_per_mile)
> cat("\nAirline Efficiency (Avg Air Time per Mile):\n")

Airline Efficiency (Avg Air Time per Mile):
> airline_efficiency
# A tibble: 16 x 3
   carrier avg_air_time_per_mile name
   <chr>                   <dbl> <chr>
 1 HA                      0.125 Hawaiian Airlines Inc.
 2 VX                      0.135 Virgin America
 3 AS                      0.136 Alaska Airlines Inc.
 4 F9                      0.142 Frontier Airlines Inc.
 5 UA                      0.145 United Air Lines Inc.
 6 DL                      0.145 Delta Air Lines Inc.
 7 AA                      0.146 American Airlines Inc.
 8 WN                      0.151 Southwest Airlines Co.
 9 FL                      0.153 AirTran Airways Corporation
10 B6                      0.155 JetBlue Airways
11 OO                      0.165 SkyWest Airlines Inc.
12 MQ                      0.167 Envoy Air
13 EV                      0.169 ExpressJet Airlines Inc.
14 9E                      0.182 Endeavor Air Inc.
15 US                      0.185 US Airways Inc.
16 YV                      0.189 Mesa Airlines Inc.
```

The code above calculates the efficiency of each airline by dividing air time by distance for every flight and then averaging this ratio per airline. The result is stored in airline_efficiency, which stores airlines by their “average air time per mile”. This is then sorted and ranked from best to worst based on their average air time per mile. The full name of the airlines is appended to the set as well.

From the sorted table, the most efficient airline i.e., the one with the lowest average air time per mile are:

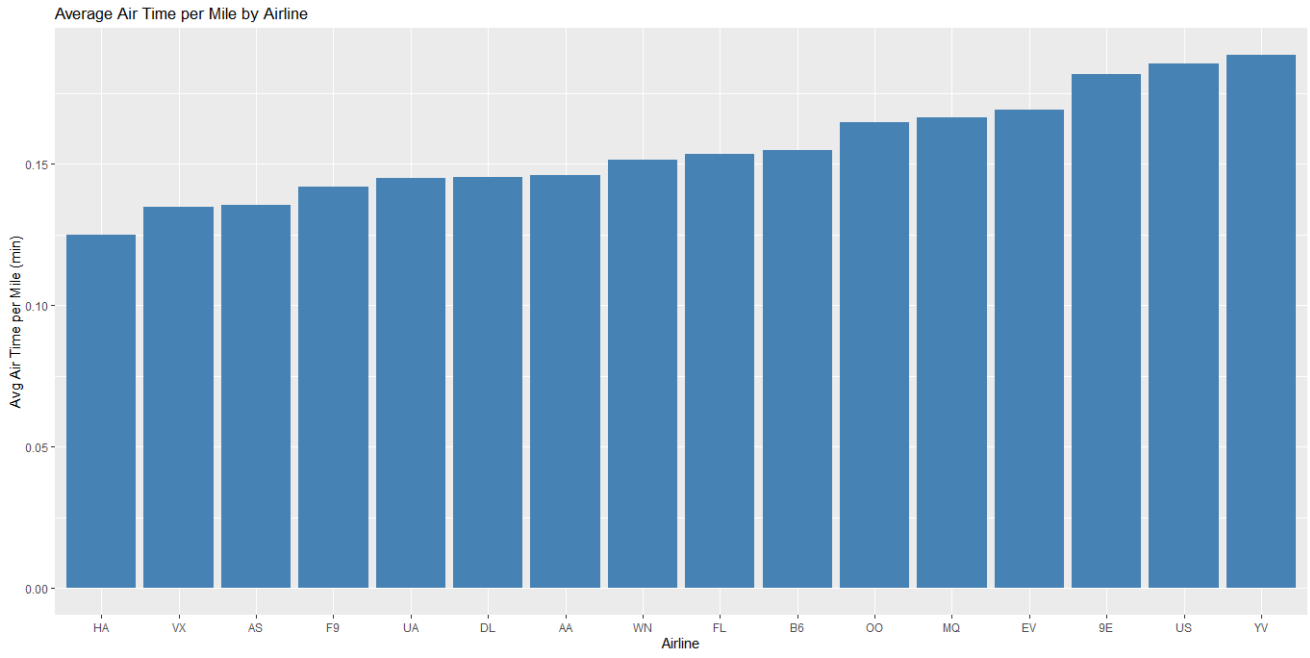
i) Hawaiian Airlines (HA) → 0.125 min/mile
ii) Virgin America (VX) → 0.132 min/mile
iii) Alaska Airlines (AS) → 0.135 min/mile

On the other end, the least efficient airlines with the highest air time per mile include:

i) Mesa Airlines (YV) → 0.188 min/mile
ii) US Airways (US) → 0.186 min/mile
iii) Endeavor Air (9E) → 0.186 min/mile

b) Using the R code, bar plots were generated to visualize which airlines are the most efficient, i.e., taking the shortest time per mile.

```
Console   Terminal   Jobs
~/
> # (b) Plot air time per mile by airline
> ggplot(airline_efficiency, aes(x = reorder(carrier, avg_air_time_per_mile),
+                                y = avg_air_time_per_mile)) +
+   geom_bar(stat = "identity", fill = "steelblue", show.legend = FALSE) +
+   labs(title = "Average Air Time per Mile by Airline",
+        x = "Airline",
+        y = "Avg Air Time per Mile (min)")
```



Average Air Time per Mile by Airline

To visualize the table, the bar plot displays the average air time per mile for each airline.

From the data, we see that Hawaiian Airlines (HA), Virgin America (VX), and Alaska Airlines (AS) are the most efficient, while Mesa Airlines (YV), US Airways (US), and Endeavor Air (9E) are the least efficient.

Thus, Hawaiian Airlines is the most efficient in terms of travel time having a travel time of 0.125 min/mile.

Q7. Using the R code, we computed the best airline that values punctuality.

```
> # Question 7: Best airline for business travelers
> # Rank airlines based on average air time per mile
> ranked_airlines <- airline_efficiency %>%
+   arrange(avg_air_time_per_mile)
> ranked_airlines
# A tibble: 16 x 3
   carrier avg_air_time_per_mile name
   <chr>                   <dbl> <chr>
 1 HA                      0.125 Hawaiian Airlines Inc.
 2 VX                      0.135 Virgin America
 3 AS                      0.136 Alaska Airlines Inc.
 4 F9                      0.142 Frontier Airlines Inc.
 5 UA                      0.145 United Air Lines Inc.
 6 DL                      0.145 Delta Air Lines Inc.
 7 AA                      0.146 American Airlines Inc.
 8 WN                      0.151 Southwest Airlines Co.
 9 FL                      0.153 AirTran Airways Corporation
10 B6                      0.155 JetBlue Airways
11 OO                      0.165 SkyWest Airlines Inc.
12 MQ                      0.167 Envoy Air
13 EV                      0.169 ExpressJet Airlines Inc.
14 9E                      0.182 Endeavor Air Inc.
15 US                      0.185 US Airways Inc.
16 YV                      0.189 Mesa Airlines Inc.
```

Airlines in terms of efficiency

```
> # Airline-Specific Delay Summary (dep_delay)
> flight_data_summary <- flight_data %>%
+   group_by(carrier) %>%
+   summarize(
+     mean_dep_delay = mean(dep_delay, na.rm = TRUE),
+     median_dep_delay = median(dep_delay, na.rm = TRUE),
+     mode_dep_delay = getmode(dep_delay),
+     sd_dep_delay = sd(dep_delay, na.rm = TRUE),
+     var_dep_delay = var(dep_delay, na.rm = TRUE),
+     min_dep_delay = min(dep_delay, na.rm = TRUE),
+     q1_dep_delay = quantile(dep_delay, 0.25, na.rm = TRUE),
+     q3_dep_delay = quantile(dep_delay, 0.75, na.rm = TRUE),
+     max_dep_delay = max(dep_delay, na.rm = TRUE),
+     iqr_dep_delay = IQR(dep_delay, na.rm = TRUE),
+     flights_count = n()
+   ) %>%
+   arrange(desc(median_dep_delay))
> cat("\nAirline-Specific (dep_delay) Summary:\n")

Airline-Specific (dep_delay) Summary:
> print(kable(flight_data_summary, format = "markdown"))
```

carrier	mean_dep_delay	median_dep_delay	mode_dep_delay	sd_dep_delay	var_dep_delay	min_dep_delay	q1_dep_delay	q3_dep_delay	max_dep_delay	iqr_dep_delay	flights_count
FL	18.605984	1	-4	52.49106	2755.3113	-22	-4	17.00	602	21.00	3175
WN	17.661657	1	-1	43.23745	1869.4774	-13	-2	17.00	471	19.00	12044
F9	20.201175	0	0	58.40434	3411.0668	-27	-4	18.00	853	22.00	681
UA	12.016908	0	-3	35.54792	1263.6547	-20	-4	11.00	483	15.00	57782
VX	12.756646	0	-1	44.01625	1937.4307	-20	-4	8.00	653	12.00	5116
B6	12.967548	-1	-4	38.38022	1473.0409	-43	-5	12.00	502	17.00	54049
EV	19.838929	-1	-5	46.44617	2157.2471	-32	-5	25.00	548	30.00	51108
9E	16.439574	-2	-5	45.48751	2069.1139	-24	-6	16.00	747	22.00	17294
DL	9.223950	-2	-5	39.65630	1572.6218	-33	-5	5.00	960	10.00	47658
YV	18.898897	-2	-7	49.16484	2417.1813	-16	-7	22.25	387	29.25	544
AA	8.569130	-3	-4	37.36527	1396.1632	-24	-6	4.00	1014	10.00	31947
AS	5.830748	-3	-6	31.42680	987.6436	-21	-7	3.00	225	10.00	709
MQ	10.445381	-3	-7	39.02520	1522.9661	-26	-7	9.00	1137	16.00	25037
HA	4.900585	-4	-5	74.10990	5492.2775	-16	-7	-1.00	1301	6.00	342
US	3.744693	-4	-6	27.93911	780.5937	-19	-7	0.00	500	7.00	19831
OO	12.586207	-6	-6	43.06599	1854.6798	-14	-9	4.00	154	13.00	29

Airline Departure Delay Stats

```
> #inflight delay
> flight_data_summary_inflight <- flight_data %>%
+   group_by(carrier) %>%
+   summarize(
+     mean_inflight_delay = mean(in_flight_delay, na.rm = TRUE),
+     median_inflight_delay = median(in_flight_delay, na.rm = TRUE),
+     mode_inflight_delay = getmode(in_flight_delay),
+     sd_inflight_delay= sd(in_flight_delay, na.rm = TRUE),
+     var_inflight_delay = var(in_flight_delay, na.rm = TRUE),
+     min_inflight_delay = min(in_flight_delay, na.rm = TRUE),
+     q1_inflight_delay = quantile(in_flight_delay, 0.25, na.rm = TRUE),
+     q3_inflight_delay = quantile(in_flight_delay, 0.75, na.rm = TRUE),
+     max_inflight_delay = max(in_flight_delay, na.rm = TRUE),
+     iqr_inflight_delay = IQR(in_flight_delay, na.rm = TRUE),
+     flights_count = n()
+   ) %>%
+   arrange(desc(median_inflight_delay))
> cat("\nAirline-Specific Summary (inflight delay):\n")

Airline-Specific Summary (inflight delay):
> print(kable(flight_data_summary_inflight, format = "markdown"))
```

```
> #arrival delay
> flight_data_summary_arr <- flight_data %>%
+   group_by(carrier) %>%
+   summarize(
+     mean_arr_delay = mean(arr_delay, na.rm = TRUE),
+     median_arr_delay = median(arr_delay, na.rm = TRUE),
+     mode_arr_delay = getmode(arr_delay),
+     sd_arr_delay= sd(arr_delay, na.rm = TRUE),
+     var_arr_delay = var(arr_delay, na.rm = TRUE),
+     min_arr_delay = min(arr_delay, na.rm = TRUE),
+     q1_arr_delay = quantile(arr_delay, 0.25, na.rm = TRUE),
+     q3_arr_delay = quantile(arr_delay, 0.75, na.rm = TRUE),
+     max_arr_delay = max(arr_delay, na.rm = TRUE),
+     iqr_arr_delay = IQR(arr_delay, na.rm = TRUE),
+     flights_count = n()
+   ) %>%
+   arrange(desc(median_arr_delay))
> cat("\nAirline-Specific Summary (arr):\n")

Airline-Specific Summary (arr):
> print(kable(flight_data_summary_inflight, format = "markdown"))
```

Console Terminal Jobs

```
Airline-Specific Summary (inflight delay):
> print(kable(flight_data_summary_inflight, format = "markdown"))
```

carrier	mean_inflight_delay	median_inflight_delay	mode_inflight_delay	sd_inflight_delay	var_inflight_delay	min_inflight_delay	q1_inflight_delay	q3_inflight_delay	max_inflight_delay	iqr_inflight_delay	flights_count
FL	1.5099213	-1	-2	15.84287	250.9967	-36	-9.00	9	110	18.00	3175
F9	1.7195301	-2	-7	22.48619	505.6286	-44	-14.00	13	135	27.00	681
MQ	0.3293526	-2	-6	16.83482	283.4112	-49	-11.00	8	157	19.00	25037
OO	-0.6551724	-2	-2	13.77958	189.8768	-17	-11.00	6	40	17.00	29
US	-1.6150976	-4	-8	16.18303	261.8905	-64	-12.00	6	148	18.00	19831
B6	-3.5095746	-6	-7	17.31688	299.8742	-69	-14.00	4	146	18.00	54049
EV	-4.0424982	-6	-10	15.16047	229.8397	-109	-13.00	3	150	16.00	51108
YV	-3.3419118	-6	-13	17.03417	290.1628	-38	-14.25	4	108	18.25	544
DL	-7.5796089	-9	-11	18.88386	356.6001	-79	-19.00	2	161	21.00	47658
AA	-8.2048393	-10	-10	19.24815	370.4913	-71	-21.00	2	181	23.00	31947
UA	-8.4588972	-10	-12	19.06615	363.5182	-74	-20.00	1	165	21.00	57782
WN	-8.0125374	-10	-9	16.85397	284.0564	-58	-19.00	1	131	20.00	12044
9E	-9.0599052	-11	-14	18.61474	346.5086	-64	-21.00	0	112	21.00	17294
HA	-11.8157895	-11	-11	23.19845	538.1683	-87	-25.00	4	52	29.00	342
VX	-10.9921814	-12	-19	20.60936	424.7456	-72	-24.00	1	196	25.00	5116
AS	-15.7616361	-17	-17	19.96150	398.4615	-70	-31.00	-4	65	27.00	709

Airline InFlight Delay Stats

carrier	mean_arr_delay	median_arr_delay	mode_arr_delay	sd_arr_delay	var_arr_delay	min_arr_delay	q1_arr_delay	q3_arr_delay	max_arr_delay	iqr_arr_delay	flights_count
F9	21.9207048	6	0	61.64600	3800.229	-47	-9.00	31.00	834	40.00	681
FL	20.1159055	5	2	54.08767	2925.476	-44	-7.00	24.00	572	31.00	3175
EV	15.7964311	-1	-13	49.86147	2486.166	-62	-14.00	26.00	577	40.00	51108
MQ	10.7747334	-1	-6	43.17431	1864.021	-53	-13.00	18.00	1127	31.00	25037
YV	15.5569853	-2	-13	52.92223	2800.763	-46	-16.00	24.25	381	40.25	544
B6	9.4579733	-3	-10	42.84230	1835.462	-71	-14.00	17.00	497	31.00	54049
WN	9.6491199	-3	-12	46.87770	2197.519	-58	-15.00	15.00	453	30.00	12044
UA	3.5580111	-6	-14	40.98434	1679.716	-75	-18.00	12.00	455	30.00	57782
US	2.1295951	-6	-11	33.06695	1093.423	-70	-15.00	8.00	492	23.00	19831
9E	7.3796692	-7	-15	50.08678	2508.685	-68	-21.00	15.00	744	36.00	17294
OO	11.9310345	-7	-24	48.58493	2360.495	-26	-16.00	6.00	157	22.00	29
DL	1.6443409	-8	-14	44.40229	1971.563	-71	-20.00	8.00	931	28.00	47658
AA	0.3642909	-9	-15	42.51618	1807.626	-75	-21.00	8.00	1007	29.00	31947
VX	1.7644644	-9	-18	49.96645	2496.646	-86	-23.00	8.00	676	31.00	5116
HA	-6.9152047	-13	-21	75.12942	5644.430	-70	-27.75	2.75	1272	30.50	342
AS	-9.9308886	-17	-18	36.48263	1330.983	-74	-32.00	2.00	198	34.00	709

Airline Arrival Delay Stats

For a business traveler, punctuality is paramount. Missing a meeting or connection due to flight delays can have significant professional consequences. Therefore, selecting an airline with a proven track record of on-time performance is essential.

While efficiency, measured by average air time per mile, is a factor, it doesn't tell the whole story. We must also consider departure, arrival, and in-flight delays to get a comprehensive view of an airline's punctuality.

Hawaiian Airlines (HA) does stand out as a strong contender. Here's why:

- Departure Punctuality: HA demonstrates the lowest median departure delay of -4 minutes. This means that, more often than not, their flights leave on time or even ahead of schedule.
- Minimal In-Flight Delays: HA also excels in minimizing in-flight delays, with a median of -11 minutes. This suggests that once airborne, HA flights tend to make up for any potential initial delays.
- Reliable Arrival Times: HA's median arrival delay of -13 minutes indicates that most of their flights arrive on time or early, which is crucial for business travelers with tight schedules.
- Flight Efficiency: HA has the best average air time per mile of 0.125 minutes/mile, indicating that they operate efficiently.

However, it is important to note that HA has a very high standard deviation in departure and arrival delays. This means that when they are delayed, they tend to be very delayed. Also, HA has a very small number of flights in the data set, so the data might not be as robust as other airlines.

Alternative considerations to keep in mind:

- Delta Air Lines (DL) and American Airlines (AA): These airlines show strong performance with low median arrival delays of -8 and -9 minutes, respectively and low median in flight delays of -7.5 and -8.2 respectively. They also have a large amount of flights in the data set, which creates a more robust set of data. They are also large airlines that cover many major airports.
- Alaska Airlines (AS): AS has the best median arrival delay of all of the airlines of -17 minutes. It also has a very low median departure delay of -3 minutes. However, it also has a high standard deviation, like HA, and a low number of flights in the data set.

Considering all factors, Delta Air Lines (DL) or American Airlines (AA) are the most reliable choices for business travelers who prioritize punctuality. While Hawaiian Airlines shows impressive numbers, the high standard deviation, and low amount of flights, means that they are less consistent. DL and AA offer a balance of efficient operations, minimal delays, and extensive coverage, making them dependable for time-sensitive travel.

Q8. As an airport manager, my primary goal is to minimize delays and enhance operational efficiency. Delays negatively impact passenger satisfaction, increase airline costs, and strain

airport infrastructure. Analyzing the nycflights dataset reveals key trends that inform two actionable strategies:

- Strategic Slot Management and Operational Flow Optimization:
  - Problem: The data clearly demonstrates a cascading delay effect. Early morning flights generally experience fewer disruptions, while afternoon and evening flights suffer from accumulated delays. This indicates a need for proactive slot management to break the chain reaction.
- Recommendations:
  - Allocate premium slots (before 8 AM) to high-priority routes having business-heavy, long-haul, and those prone to downstream impacts. This minimizes the risk of significant delays propagating throughout the day.
  - Create a system that allows for flexible reallocation of slots based on real-time data. For example, if a morning flight is significantly delayed, its slot could be temporarily reallocated to a later, potentially less critical flight, to minimize the impact.
  - Introduce a tiered incentive program for airlines. Offer preferential gate assignments, reduced landing fees, or expedited turnaround services for carriers that consistently achieve on-time departures, particularly in the morning.
  - Work with airlines to optimize turnaround times. Invest in technology and infrastructure that speeds up baggage handling, refueling, and aircraft cleaning. Enforce stricter turnaround time metrics, and work with airlines to find ways to reduce them.
  - Encourage airlines to schedule routine maintenance during off peak hours, and reward airlines who do so.

- Proactive Winter Weather Mitigation and Air Traffic Flow Management:
  - Problem: The dataset highlights a significant increase in delays during winter months (December, January and February) due to adverse weather conditions. This necessitates a robust winter operations strategy.
- Recommendations:
  - Strengthen collaboration with the FAA and air traffic control to implement dynamic rerouting, optimized departure sequences, and collaborative decision-making during harsh weather.
  - Install systems that monitor and manage ground traffic flow, especially during de-icing operations. This will help to reduce bottlenecks and improve the efficiency of ground operations.
  - When delays are unavoidable, improve communication with passengers. Provide real time updates, and multiple ways to access that information.

By strategically managing flight slots and proactively mitigating winter weather impacts, an airport can significantly improve on-time performance. These strategies require a collaborative approach with airlines, air traffic control, and other stakeholders. While external factors will always play a role, data-driven decision-making and operational excellence are crucial for minimizing delays and enhancing the passenger experience.