

Project Proposal

Due November 16 at 11:59pm

ALL group member names here

Load Packages

```
library(tidyverse)
```

Dataset 1 (top choice)

Data source:

The dataset was manually downloaded from the U.S. Department of Transportation's Bureau of Transportation Statistics (BTS) On-Time Performance delay-cause database (https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?20=E). During the download process, I selected All Carriers and specified the time window from July 2020 through July 2025.

Brief description:

This dataset is maintained by the U.S. Department of Transportation's Bureau of Transportation Statistics (BTS). Airlines are required by federal regulation (14 CFR Part 234) to report monthly on-time performance for all domestic scheduled flights operated by major U.S. carriers. Data are collected directly from each airline's operational system, verified by BTS, and aggregated at the carrier-airport-month level. Each observation represents a monthly record for a specific airline and airport in the U.S. It includes both categorical variables (e.g., carrier, airport, month, year) and numerical variables (e.g., number of flights, delay counts, and average delay minutes by cause such as weather or carrier).

Research question 1:

- What are the primary causes of flight delays among major U.S. airlines, and how do these causes vary by season or airport?

Load the data and provide a glimpse():

```

library(tidyverse)
library(lubridate)

df <- read_csv("Airline_Delay_Cause2020_2025.csv")

Rows: 126165 Columns: 21
-- Column specification -----
Delimiter: ","
chr (4): carrier, carrier_name, airport, airport_name
dbl (17): year, month, arr_flights, arr_del15, carrier_ct, weather_ct, nas_c...
i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.

```

```
glimpse(df)
```

```

Rows: 126,165
Columns: 21
$ year                  <dbl> 2025, 2025, 2025, 2025, 2025, 2025, 2025, 2025, 20~
$ month                 <dbl> 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, ~
$ carrier                <chr> "YV", "YV", "YV", "YV", "YV", "YV", "YV", "Y~
$ carrier_name           <chr> "Mesa Airlines Inc.", "Mesa Airlines Inc.", "Mesa ~
$ airport                <chr> "BWI", "CHS", "CLE", "CLT", "CMH", "COS", "CRP", "~~
$ airport_name            <chr> "Baltimore, MD: Baltimore/Washington International~
$ arr_flights             <dbl> 18, 48, 65, 134, 61, 31, 19, 75, 97, 31, 31, 148, ~
$ arr_del15               <dbl> 2, 16, 10, 31, 12, 5, 4, 14, 23, 1, 4, 38, 32, 47, ~
$ carrier_ct              <dbl> 1.43, 4.06, 3.53, 13.19, 6.44, 1.84, 0.71, 7.64, 8~
$ weather_ct              <dbl> 0.00, 3.15, 1.00, 2.43, 0.28, 0.00, 0.30, 1.00, 0.~
$ nas_ct                  <dbl> 0.57, 5.49, 3.04, 10.63, 3.59, 1.15, 0.70, 3.57, 4~
$ security_ct              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ late_aircraft_ct         <dbl> 0.00, 3.30, 2.43, 4.74, 1.69, 2.01, 2.29, 1.79, 9.~
$ arr_cancelled            <dbl> 0, 2, 2, 8, 0, 0, 1, 0, 1, 0, 0, 6, 1, 4, 1, 1, 0, ~
$ arr_diverted              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, ~
$ arr_delay                <dbl> 44, 1282, 735, 2264, 630, 668, 823, 722, 2125, 50, ~
$ carrier_delay             <dbl> 32, 331, 151, 798, 211, 154, 45, 485, 655, 41, 32, ~
$ weather_delay             <dbl> 0, 438, 172, 303, 17, 0, 6, 24, 65, 0, 51, 569, 27~
$ nas_delay                 <dbl> 12, 259, 139, 702, 123, 67, 14, 81, 437, 9, 10, 47~
$ security_delay             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ late_aircraft_delay        <dbl> 0, 254, 273, 461, 279, 447, 758, 132, 968, 0, 94, ~

```

```

# Missing values per column
df %>%
  summarise(across(everything(), ~ sum(is.na(.)))) %>%
  pivot_longer(everything(), names_to = "column", values_to = "na_count")

# A tibble: 21 x 2
  column      na_count
  <chr>        <int>
1 year          0
2 month         0
3 carrier       0
4 carrier_name  0
5 airport        0
6 airport_name   0
7 arr_flights    261
8 arr_del15     497
9 carrier_ct     261
10 weather_ct    261
# i 11 more rows

# Number of unique categories for categorical variables
df %>% summarise(
  n_carriers = n_distinct(carrier),
  n_airports = n_distinct(airport),
  n_years = n_distinct(year),
  n_months = n_distinct(month)
)

# A tibble: 1 x 4
  n_carriers n_airports n_years n_months
  <int>       <int>     <int>     <int>
1         25         389       6        12

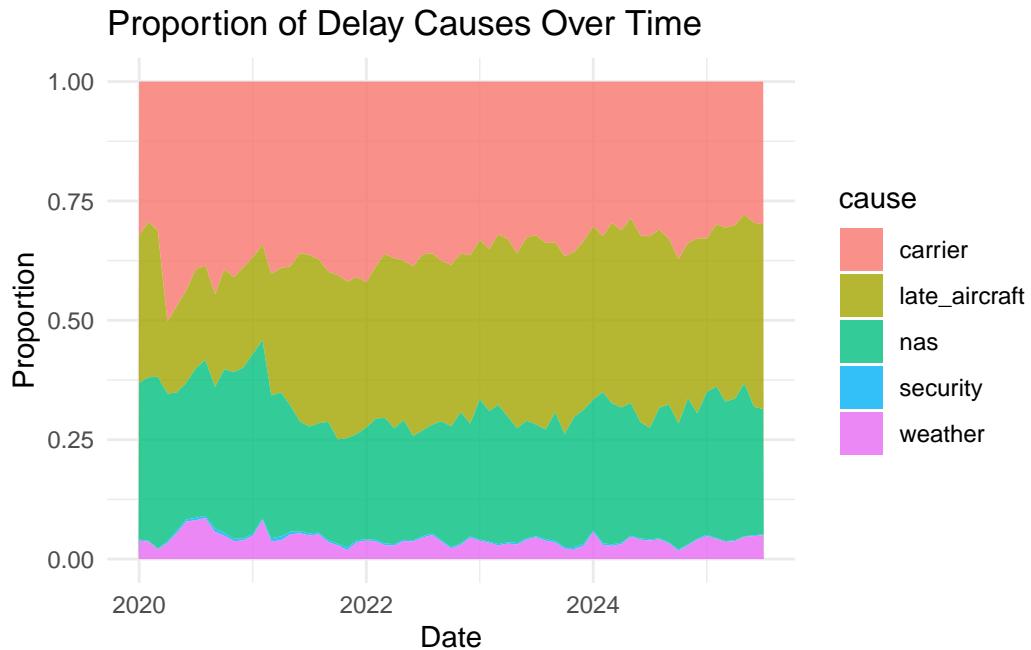
# Quick summary for key numerical columns
df %>%
  select(arr_flights, arr_del15, carrier_ct, weather_ct, nas_ct, security_ct, late_aircraf
summary()

arr_flights      arr_del15      carrier_ct      weather_ct
```

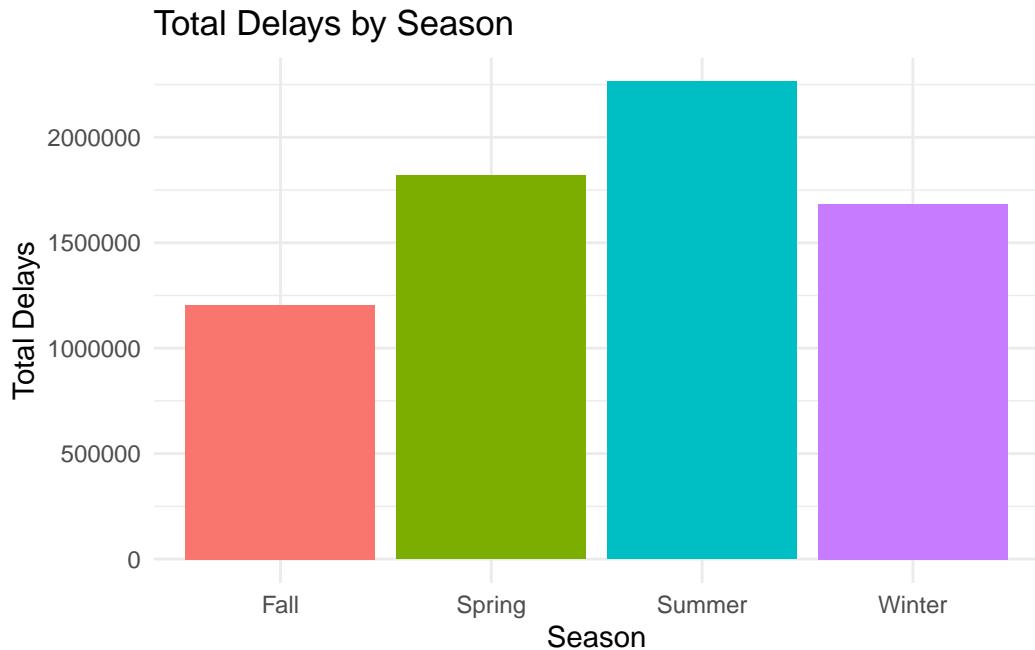
Min. : 1.0	Min. : 0.00	Min. : 0.00	Min. : 0.000
1st Qu.: 35.0	1st Qu.: 5.00	1st Qu.: 1.76	1st Qu.: 0.000
Median : 82.0	Median : 13.00	Median : 5.00	Median : 0.200
Mean : 299.3	Mean : 55.47	Mean : 19.01	Mean : 2.168
3rd Qu.: 197.0	3rd Qu.: 37.00	3rd Qu.: 14.30	3rd Qu.: 1.680
Max. :21854.0	Max. :5544.00	Max. :1886.58	Max. :343.380
NA's :261	NA's :497	NA's :261	NA's :261
nas_ct security_ct late_aircraft_ct			
Min. : 0.00	Min. : 0.0000	Min. : 0.00	
1st Qu.: 0.67	1st Qu.: 0.0000	1st Qu.: 0.91	
Median : 2.88	Median : 0.0000	Median : 3.52	
Mean : 14.77	Mean : 0.1658	Mean : 19.26	
3rd Qu.: 8.62	3rd Qu.: 0.0000	3rd Qu.: 11.54	
Max. :1685.74	Max. :58.6900	Max. :2588.13	
NA's :261	NA's :261	NA's :261	

Exploratory Plots:

```
# Delay Causes proportion
df %>%
  mutate(date = ymd(paste(year, month, "01"))) %>%
  group_by(date) %>%
  summarize(
    carrier = sum(carrier_ct, na.rm=TRUE),
    weather = sum(weather_ct, na.rm=TRUE),
    nas = sum(nas_ct, na.rm=TRUE),
    security = sum(security_ct, na.rm=TRUE),
    late_aircraft = sum(late_aircraft_ct, na.rm=TRUE)
  ) %>%
  pivot_longer(-date, names_to="cause", values_to="count") %>%
  group_by(date) %>%
  mutate(prop = count / sum(count)) %>%
  ggplot(aes(date, prop, fill = cause)) +
  geom_area(alpha=0.8) +
  labs(title = "Proportion of Delay Causes Over Time",
       x = "Date", y = "Proportion") +
  theme_minimal()
```



```
# Seasonal Delay Totals
df %>%
  mutate(season =
    case_when(
      month %in% c(12,1,2) ~ "Winter",
      month %in% c(3,4,5) ~ "Spring",
      month %in% c(6,7,8) ~ "Summer",
      TRUE ~ "Fall"
    )) %>%
  group_by(season) %>%
  summarize(total_delays = sum(arr_del15, na.rm = TRUE)) %>%
  ggplot(aes(season, total_delays, fill = season)) +
  geom_col(show.legend = FALSE) +
  labs(title = "Total Delays by Season",
       x = "Season", y = "Total Delays") +
  theme_minimal()
```



```
# Delays by airports
df %>%
  group_by(airport) %>%
  summarize(total_delays = sum(arr_del15, na.rm = TRUE)) %>%
  slice_max(total_delays, n = 10) %>%
  ggplot(aes(x = reorder(airport, total_delays),
             y = total_delays)) +
  geom_col(fill = "darkblue") +
  coord_flip() +
  scale_y_continuous(labels = scales::comma) +
  labs(title = "Top 10 Airports by Total Delays",
       x = "Airport",
       y = "Total Delays") +
  theme_minimal()
```

Top 10 Airports by Total Delays

