Stats15 Final Project: Causes of Flight Delays

Eliyah Dawson, Ethan Dao, Frank Hobson, Stephanie Lei, Will Bodeau

2022-12-02

Contents

Section 1: Introduction

1.1 Why Airline Data and Our Research Question

Traveling around the world can be a very stressful event. From rushing through airport security to ensuring your arrival accommodation, many things can go wrong. Over the last few years during the arrival and slow departure of COVID-19, airline flights have been constantly delayed and/or canceled. Although COVID may have impacted flight proceedings, delays have been happening for years before that. In this project, we aim to answer exactly what factors affect flight delays.

1.2 Background Information

Basic Structure: This data was collected from the Bureau of Transportation Statistics' Office of Airline Information (found in the TranStats Library). The data is self-reported by US certified air carriers' flight logs (specifically airlines that account for at least one percent of domestic scheduled passenger revenues), and is compiled into a data table. Data is collected from the US only, and initially had 318,017 observations with 21 variables. There are 33 airlines in this dataset. Not all airlines existed for the entirety of the time period that this data covers, as carriers frequently merge or dissolve completely, represented by NA values when appropriate.

What is Considered a Delay? A flight is labeled as a delay when it arrives at (or departs) the gate 15 minutes or more after its scheduled arrival/departure time as shown in a Computerized Reservation System (CRS). There is no standardized CRS, and each airline is able to manage their own, or use an already created system. Carries report delays as one of five broad categories. There is an overlap between NAS Delays and weather delays, as weather delays accounts for extreme weather conditions, while NAS also includes some weather delays that are less extreme but still hinder aircraft or airport operations. These weather delays within the NAS delay category account for 45% of delays in that category.

Airport Listings This data set uses data from 420 airports between June 2003 and June 2022. To further categorize the data, we joined this dataset with a dataset that provided the two letter code for a state, the region it is located in (North, South, Southwest, etc.), and its division (Pacific, New England, etc) found on GitHub and originally determined by the US Census Bureau. There are 4 distinct regions and 9 divisions.

1.3 Explanation of Variables

- **year**: The year in which these flights occurred
- **month**: The month in which these flights occurred
- **carrier**: A unique two character alphanumeric code used to identify airline companies. Each carrier has an airline designator that is used by air traffic controllers, assigned by the International Air Transport Association (IATA).
- **carrier_name**: The full name of each carrier. An air carrier is a company certified by the US Department of Transportation that provides air transport services for both freight and passengers. There are 97 operating air carriers as of March 2022, and this dataset focuses on commercial airliners with 1% or more of domestic passenger revenues, for a total of 33 companies.
- **airport**: Designated three letter code used to identify airports by the IATA. In addition to using the codes to identify airports, the IATA also uses these codes to attribute the number of flight delays and cancellations to each airport.
- **airport_name**: The full name of the airport where the flights arrived at
- **arr_flights**: The number of flights arriving at their respective arrival airports
- **arr_del15**: The number of flights that arrived to their landing airports by 15 minutes or more, or delayed by 15 minutes or more

- **carrier_ct**: The total number of flights that are delayed due to the fault of the air carrier. Air carrier delays are determined by whether the flight was delayed by circumstances within the airline's control (ex. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.)
- **weather_ct**: The total number of flights that are delayed due to extreme weather conditions. Extreme weather conditions include actual or forecasted weather conditions that, in the judgment of the carrier, delay the operation of a flight (ex. tornadoes, blizzards, hurricanes)
- **nas_ct**: The total number of flights that are delayed due to the National Airspace System (NAS). The NAS is the network of the airspace, navigation facilities, and airports of the United States, which regulates the safe operation of aircrafts and records information about flight departures, arrivals, and delays. NAS delays cover a broad range of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.
- **security_ct**: The total number of flights that are delayed due to security reasons. Different security reasons may include evacuation of a terminal or concourse, a security breach in the aircraft, inoperative screening equipment, and/or lines longer than 29 minutes at screening areas.
- **late_aircraft_ct**: The total number of flights that are delayed due to late aircraft arrival, or because a previous aircraft arrived late, which in turn caused the present aircraft to depart late.
- **arr_cancelled**: The number of canceled flights. A cancellation occurs when an airline does not operate the flight at all for a certain reason.
- **arr_diverted**: The number of diverted flights. A flight is classified as diverted when it has been routed from its original arrival destination to a new, typically temporary, arrival destination.
- **arr_delay**: The total time (in number of minutes) that flights have been delayed.
- **carrier_delay^**: The total time (in number of minutes) that flights have been delayed due to the fault of the air carrier.
- **weather_delay^**: The total time (in number of minutes) that flights have been delayed due to extreme weather conditions.
- **nas_delay^**: The total time (in number of minutes) that a flight is delayed due to the National Airspace System (NAS).
- **security_delay^**: The total time (in number of minutes) that a flight is delayed due to security reasons.
- **late_aircraft_delay^**: The total time (in number of minutes) that a flight is delayed due to late aircraft arrival, or because a previous aircraft arrived late, which in turn caused the present aircraft to depart late.

Section 2: Data Loading and Cleanup

2.1 Loading libraries and the dataset

```
library(dplyr)
library(tidyverse)
library(ggstatsplot)
options(warn=-1)
```

```
Airline_Delay_Cause <- read.csv(file = "Airline_Delay_Cause.csv")
```

```
Airline_Delay_Cause %>%
  glimpse()
```

```
## Rows: 318,017
## Columns: 21
## $ year                <int> 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 20~
## $ month               <int> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5,~
## $ carrier             <chr> "9E", "9E", "9E", "9E", "9E", "9E", "9E", "9E", "9~
## $ carrier_name        <chr> "Endeavor Air Inc.", "Endeavor Air Inc.", "Endeavo~
## $ airport             <chr> "ABE", "ABY", "ACK", "AEX", "AGS", "ALB", "ATL", "~
## $ airport_name        <chr> "Allentown/Bethlehem/Easton, PA: Lehigh Valley Int~
## $ arr_flights         <dbl> 136, 91, 19, 88, 181, 134, 3042, 118, 53, 58, 91, ~
## $ arr_del15           <dbl> 7, 16, 2, 14, 19, 18, 453, 13, 8, 14, 15, 2, 40, 3~
## $ carrier_ct          <dbl> 5.95, 7.38, 0.13, 7.26, 13.84, 4.42, 142.96, 4.83,~
## $ weather_ct          <dbl> 0.00, 0.00, 0.00, 0.76, 0.00, 1.00, 13.53, 0.00, 1~
## $ nas_ct              <dbl> 0.05, 2.54, 1.00, 4.35, 3.07, 6.48, 106.98, 6.85, ~
## $ security_ct         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ late_aircraft_ct    <dbl> 1.00, 6.09, 0.88, 1.64, 2.09, 6.09, 189.53, 1.32, ~
## $ arr_cancelled       <dbl> 0, 0, 1, 0, 0, 5, 2, 0, 1, 1, 0, 0, 15, 2, 0, 12, ~
## $ arr_diverted        <dbl> 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,~
## $ arr_delay           <dbl> 255, 884, 138, 947, 808, 917, 38397, 914, 582, 996~
## $ carrier_delay       <dbl> 222, 351, 4, 585, 662, 224, 20775, 578, 445, 290, ~
## $ weather_delay       <dbl> 0, 0, 0, 35, 0, 78, 1205, 0, 20, 0, 35, 0, 0, 0, 0~
## $ nas_delay           <dbl> 4, 81, 106, 125, 87, 398, 3610, 171, 19, 618, 87, ~
## $ security_delay      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ late_aircraft_delay <dbl> 29, 452, 28, 202, 59, 217, 12807, 165, 98, 88, 18,~
```

```
nrow(Airline_Delay_Cause) %>%
  print()
```

```
## [1] 318017
```

For reference, this shows the general structure of the original dataframe. It has 21 variables and 318,017 rows. However, we also need to check for NA values.

```
map_dfr(Airline_Delay_Cause, ~sum(is.na(.))) %>%
  glimpse()
```

```
## Rows: 1
## Columns: 21
## $ year                <int> 0
## $ month               <int> 0
## $ carrier             <int> 0
## $ carrier_name        <int> 0
## $ airport             <int> 0
## $ airport_name        <int> 0
## $ arr_flights         <int> 488
## $ arr_del15           <int> 728
## $ carrier_ct          <int> 488
## $ weather_ct          <int> 488
## $ nas_ct              <int> 488
## $ security_ct         <int> 488
## $ late_aircraft_ct    <int> 488
## $ arr_cancelled       <int> 488
## $ arr_diverted        <int> 488
## $ arr_delay           <int> 488
## $ carrier_delay       <int> 488
## $ weather_delay       <int> 488
## $ nas_delay           <int> 488
## $ security_delay      <int> 488
## $ late_aircraft_delay <int> 488
```

There are 15 columns that have NA values. 14 out of the 15 columns with NA values have 488 NA observations.

```
NA_Rows <- Airline_Delay_Cause[rowSums(is.na(Airline_Delay_Cause)) > 0,]
glimpse(NA_Rows)
```

```
## Rows: 728
## Columns: 21
## $ year                <int> 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 20~
## $ month               <int> 4, 4, 4, 4, 1, 1, 1, 1, 1, 1, 1, 12, 12, 10, 10, 1~
## $ carrier             <chr> "OO", "YV", "YV", "YX", "MQ", "OO", "OO", "YV", "Y~
## $ carrier_name        <chr> "SkyWest Airlines Inc.", "Mesa Airlines Inc.", "Me~
## $ airport             <chr> "LNK", "MLB", "SHV", "GRB", "COU", "BOS", "LFT", "~
## $ airport_name        <chr> "Lincoln, NE: Lincoln Airport", "Melbourne, FL: Me~
## $ arr_flights         <dbl> NA, 1, 1, NA, NA, 7, NA, 2, 1, 1, 2, NA, NA, NA, N~
## $ arr_del15           <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ carrier_ct          <dbl> NA, 0, 0, NA, NA, 0, NA, 0, 0, 0, 0, NA, NA, NA, N~
## $ weather_ct          <dbl> NA, 0, 0, NA, NA, 0, NA, 0, 0, 0, 0, NA, NA, NA, N~
## $ nas_ct              <dbl> NA, 0, 0, NA, NA, 0, NA, 0, 0, 0, 0, NA, NA, NA, N~
## $ security_ct         <dbl> NA, 0, 0, NA, NA, 0, NA, 0, 0, 0, 0, NA, NA, NA, N~
## $ late_aircraft_ct    <dbl> NA, 0, 0, NA, NA, 0, NA, 0, 0, 0, 0, NA, NA, NA, N~
## $ arr_cancelled       <dbl> NA, 1, 1, NA, NA, 7, NA, 2, 1, 1, 2, NA, NA, NA, N~
## $ arr_diverted        <dbl> NA, 0, 0, NA, NA, 0, NA, 0, 0, 0, 0, NA, NA, NA, N~
## $ arr_delay           <dbl> NA, 0, 0, NA, NA, 0, NA, 0, 0, 0, 0, NA, NA, NA, N~
## $ carrier_delay       <dbl> NA, 0, 0, NA, NA, 0, NA, 0, 0, 0, 0, NA, NA, NA, N~
## $ weather_delay       <dbl> NA, 0, 0, NA, NA, 0, NA, 0, 0, 0, 0, NA, NA, NA, N~
## $ nas_delay           <dbl> NA, 0, 0, NA, NA, 0, NA, 0, 0, 0, 0, NA, NA, NA, N~
## $ security_delay      <dbl> NA, 0, 0, NA, NA, 0, NA, 0, 0, 0, 0, NA, NA, NA, N~
## $ late_aircraft_delay <dbl> NA, 0, 0, NA, NA, 0, NA, 0, 0, 0, 0, NA, NA, NA, N~
```

After doing some more research, we found out that the NA values were caused by airlines discontinuing their service in certain airports at the time for a variety of reasons.

```
air_data <- Airline_Delay_Cause
slice(air_data, 2873)
```

```
##   year month carrier          carrier_name airport                 airport_name
## 1 2022     4      OO SkyWest Airlines Inc.     LNK Lincoln, NE: Lincoln Airport
##   arr_flights arr_del15 carrier_ct weather_ct nas_ct security_ct
## 1          NA        NA         NA         NA     NA          NA
##   late_aircraft_ct arr_cancelled arr_diverted arr_delay carrier_delay
## 1               NA            NA           NA        NA            NA
##   weather_delay nas_delay security_delay late_aircraft_delay
## 1            NA        NA             NA                  NA
```

For example, the NA values in this case are caused by pilot shortages in the area, causing SkyWest to halt services at the airport for the month of April 2022 in Lincoln, Nebraska. Although this is interesting, these NA values involve one-time events that cannot be attributed to a specific cause (airline, carrier, delay, etc.)

Therefore, we decided to drop these NA values, since they only represent 0.23% of our data and would not be significant in our data analysis.

```
air_data <- Airline_Delay_Cause %>%
  mutate(month = month.name[c(month)]) %>%
  drop_na()
head(air_data)
```

```
##   year month carrier      carrier_name airport
## 1 2022   May      9E Endeavor Air Inc.     ABE
## 2 2022   May      9E Endeavor Air Inc.     ABY
## 3 2022   May      9E Endeavor Air Inc.     ACK
## 4 2022   May      9E Endeavor Air Inc.     AEX
## 5 2022   May      9E Endeavor Air Inc.     AGS
## 6 2022   May      9E Endeavor Air Inc.     ALB
##                                                 airport_name arr_flights
## 1 Allentown/Bethlehem/Easton, PA: Lehigh Valley International         136
## 2                      Albany, GA: Southwest Georgia Regional          91
## 3                           Nantucket, MA: Nantucket Memorial          19
## 4                      Alexandria, LA: Alexandria International          88
## 5                  Augusta, GA: Augusta Regional at Bush Field         181
## 6                             Albany, NY: Albany International         134
##   arr_del15 carrier_ct weather_ct nas_ct security_ct late_aircraft_ct
## 1         7       5.95       0.00   0.05           0             1.00
## 2        16       7.38       0.00   2.54           0             6.09
## 3         2       0.13       0.00   1.00           0             0.88
## 4        14       7.26       0.76   4.35           0             1.64
## 5        19      13.84       0.00   3.07           0             2.09
## 6        18       4.42       1.00   6.48           0             6.09
##   arr_cancelled arr_diverted arr_delay carrier_delay weather_delay nas_delay
## 1             0            0       255           222             0         4
## 2             0            0       884           351             0        81
## 3             1            0       138             4             0       106
## 4             0            0       947           585            35       125
```

```
## 5             0            0       808           662             0        87
## 6             5            0       917           224            78       398
##   security_delay late_aircraft_delay
## 1              0                  29
## 2              0                 452
## 3              0                  28
## 4              0                 202
## 5              0                  59
## 6              0                 217
```

We dropped NA values and also changed the month column from integer values to String names of the individual months, which will make visualizing the data more clear.

Now we can also join the dataset to a new dataframe to determine what state and region each airport is located in.

```
# load region csv and add it to the airline data
library(readr)
urlfile = "https://raw.githubusercontent.com/cphalpert/census-regions/master/us%20census%20bureau%20reg
region_data <- read_csv(url(urlfile))

# Extract state and join columns
data_state <- air_data %>%
  mutate(State = str_extract(air_data$airport_name, paste0(state.abb, collapse = "|")))

air_data <- data_state %>%
  full_join(region_data, by = c("State" = "State Code"), suffix = c("_abb", "_name"))

air_data %>%
  select(year, airport_name, 22:25) %>%
  glimpse()
```

```
## Rows: 317,290
## Columns: 6
## $ year         <int> 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 202~
## $ airport_name <chr> "Allentown/Bethlehem/Easton, PA: Lehigh Valley Internatio~
## $ State        <chr> "PA", "GA", "MA", "LA", "GA", "NY", "GA", "WI", "TX", "NC~
## $ State_name   <chr> "Pennsylvania", "Georgia", "Massachusetts", "Louisiana", ~
## $ Region       <chr> "Northeast", "South", "Northeast", "South", "South", "Nor~
## $ Division     <chr> "Middle Atlantic", "South Atlantic", "New England", "West~
```

This can be used to explain flight delay causes in relation to weather patterns and climate in each individual region, determine air traffic per region, etc.

```
air_data %>%
  select(7:21) %>%
  summary()
```

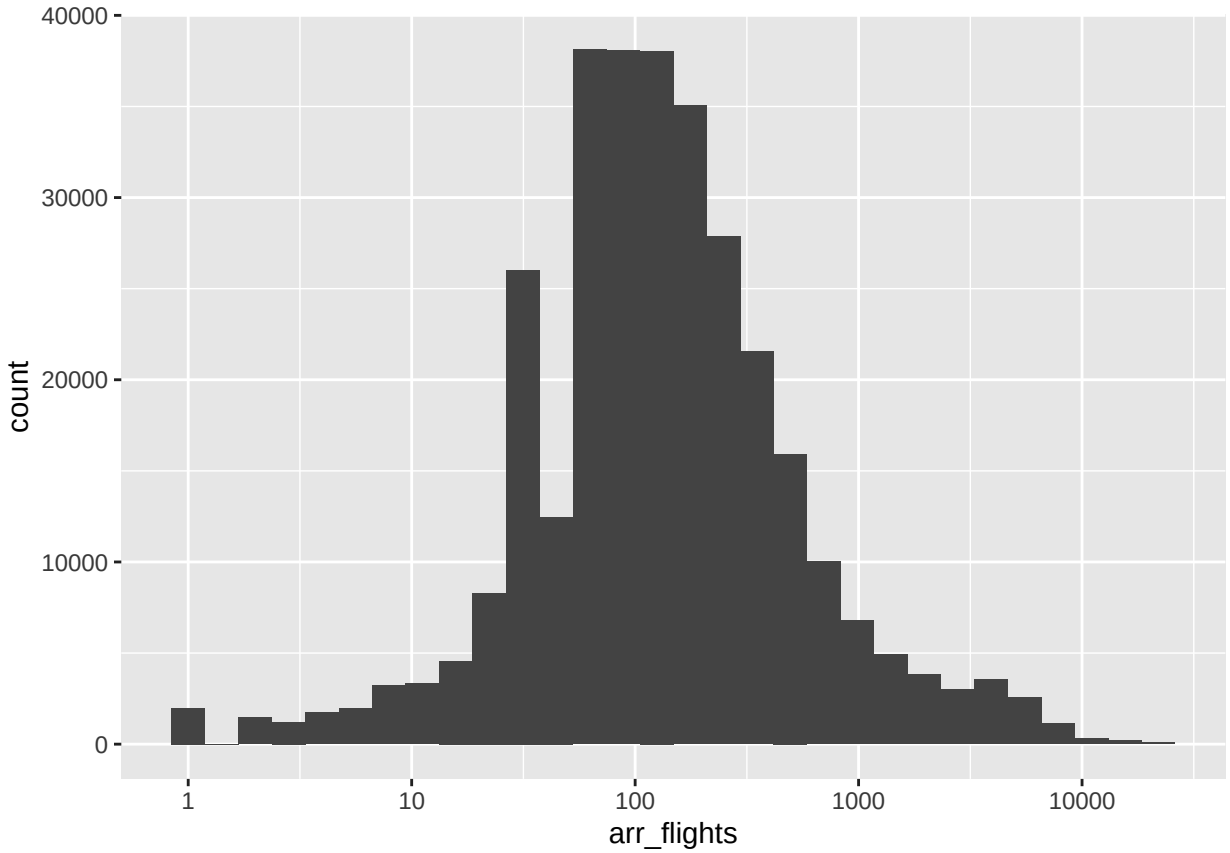
```
##   arr_flights      arr_del15        carrier_ct        weather_ct
##  Min.   :    1   Min.   :   0.0   Min.   :   0.00   Min.   : 0.000
##  1st Qu.:   59   1st Qu.:   9.0   1st Qu.:   3.01   1st Qu.: 0.000
##  Median :  120   Median :  23.0   Median :   8.21   Median : 0.580
##  Mean   :  382   Mean   :  72.9   Mean   :  21.09   Mean   : 2.618
```

```
##  3rd Qu.:  274   3rd Qu.:  56.0   3rd Qu.:  19.71   3rd Qu.:  2.000
##  Max.   :21977   Max.   :6377.0   Max.   :1792.07   Max.   :717.940
##  NA's   :1       NA's   :1        NA's   :1         NA's   :1
##      nas_ct         security_ct      late_aircraft_ct  arr_cancelled
##  Min.   :  -0.01   Min.   : 0.0000   Min.   :   0.00   Min.   :   0.000
##  1st Qu.:   1.69   1st Qu.: 0.0000   1st Qu.:   1.65   1st Qu.:   0.000
##  Median :   5.50   Median : 0.0000   Median :   5.87   Median :   1.000
##  Mean   :  24.02   Mean   : 0.1792   Mean   :  24.99   Mean   :   7.206
##  3rd Qu.:  15.37   3rd Qu.: 0.0000   3rd Qu.:  17.05   3rd Qu.:   4.000
##  Max.   :4091.27   Max.   :80.5600   Max.   :1885.47   Max.   :4951.000
##  NA's   :1         NA's   :1         NA's   :1         NA's   :1
##   arr_diverted        arr_delay      carrier_delay    weather_delay
##  Min.   :  0.0000   Min.   :     0   Min.   :     0   Min.   :    0.0
##  1st Qu.:  0.0000   1st Qu.:   437   1st Qu.:   149   1st Qu.:    0.0
##  Median :  0.0000   Median :  1203   Median :   437   Median :   25.0
##  Mean   :  0.8682   Mean   :  4213   Mean   :  1288   Mean   :  220.7
##  3rd Qu.:  1.0000   3rd Qu.:  3083   3rd Qu.:  1101   3rd Qu.:  159.0
##  Max.   :256.0000   Max.   :433687   Max.   :196944   Max.   :57707.0
##  NA's   :1          NA's   :1        NA's   :1        NA's   :1
##    nas_delay      security_delay    late_aircraft_delay
##  Min.   :   -19   Min.   :   0.00   Min.   :     0
##  1st Qu.:    56   1st Qu.:   0.00   1st Qu.:    79
##  Median :   203   Median :   0.00   Median :   352
##  Mean   :  1100   Mean   :   7.22   Mean   :  1597
##  3rd Qu.:   602   3rd Qu.:   0.00   3rd Qu.:  1111
##  Max.   :238440   Max.   :3760.00   Max.   :148181
##  NA's   :1        NA's   :1         NA's   :1
```

This shows a basic summary of all numeric rows, and shows the extreme values present in each row, and shows that our data will consistently will be full of outliers. In general, it is not reliable to outright remove these disproportionately large data points as we do not suspect that they were input incorrectly or are false data, but they should be taken into account for, as they highly skew calculations such as the average.

We also want to assign airports a "size" based on the number of incoming flights they receive because there is such a large discrepancy between airports. To create a variable denoting the size of an airport, first we have to find a metric to determine the size of an airport (small, medium, or large).

```
ggplot(air_data, aes(x=arr_flights)) + geom_histogram() + scale_x_log10()
```



This data is skewed quite heavily considering that it has a logarithmic scale on the x-axis. Unsurprisingly, there is not an even distribution of data, which makes sense since airports vary greatly in capacity, and major airports will have lots of arriving flights. The extreme distribution explains why we need to categorize airports by their size for future analysis.

```
summary_data <- air_data %>%
  group_by(airport) %>%
  summarise(total = sum(arr_flights)) %>%
  summary(total)
print(summary_data)
```

```
##    airport              total
##  Length:419         Min.   :      1
##  Class :character   1st Qu.:   6131
##  Mode  :character   Median :  31398
##                     Mean   : 289998
##                     3rd Qu.: 145904
##                     Max.   :7259395
##                     NA's   :1
```

This summary shows the summary data for flights per airport. Using the 1st quartile, 3rd quartile, and IQR, we can create size brackets to categorize airports to.

```
size_data <- air_data %>%
  group_by(airport) %>%
  summarise(total = sum(arr_flights)) %>%
```

```
  ungroup() %>%
  mutate(airport_size = case_when(total < 6131 ~ "Small",
                                  total > 6131 & total < 145904 ~ "Medium",
                                  total > 145904 ~ "Large")) %>%
  select(airport, airport_size)
```

Small airports are when the total sum of arriving flights falls within the bottom 25%, medium is the interquartile range between 1st and 3rd, and large airports are anything beyond the third quartile, within the top 25%.

```
#Joins size_data to air_data
air_data <- air_data %>%
  full_join(size_data, by = "airport")
```

```
# Segment data for pre-COVID vs. post-COVID
air_data <- air_data %>%
  mutate(covid_yesorno = case_when(year <= 2019 ~ "no",
                                   year > 2019 ~ "yes")) %>%
  drop_na()
```

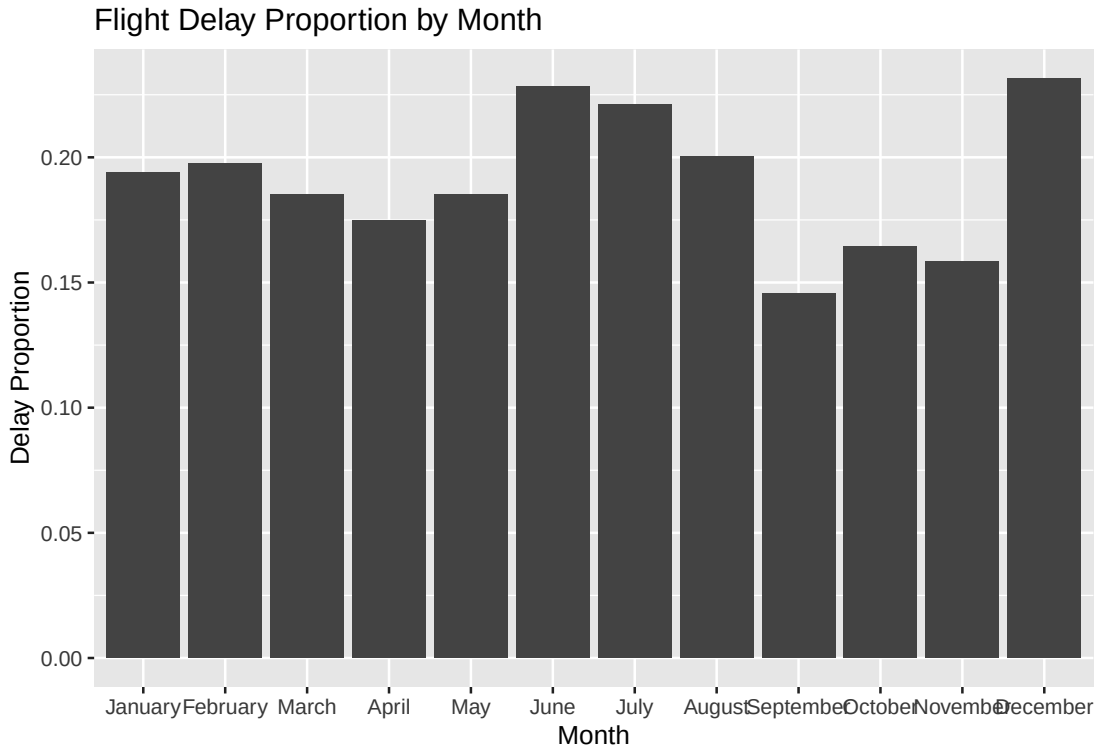
Section 3 - Exploratory Data Analysis

3.1: Delays by Time of Year

Let's look at the flight delays over time to look for any trends where the number of flight delays might have been higher or lower.

3.1.1: Number of flight delays vs. time of year:

```
air_data$month <- factor(air_data$month, levels = month.name)
air_data %>%
  group_by(month) %>%
  summarise(delayproportion = sum(arr_del15) / sum(arr_flights)) %>%
  ggplot(aes(x = month, y = delayproportion)) +
  geom_col() +
  labs(x = "Month", y = "Delay Proportion", title = "Flight Delay Proportion by Month")
```

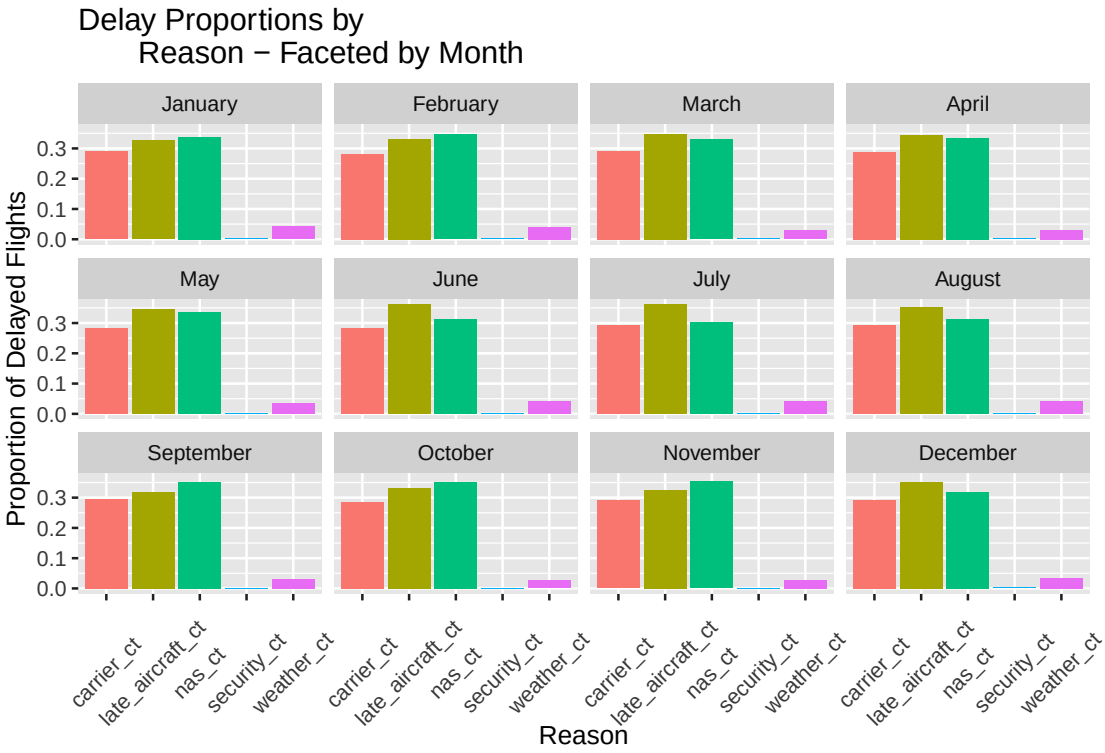


Here we can see that the months that generally have the most delays are June, July, and December, likely due to summer vacation and traveling for Christmas. However, to investigate further into why this is happening, we can look deeper into the reasons for flight delays during these months to draw conclusions.

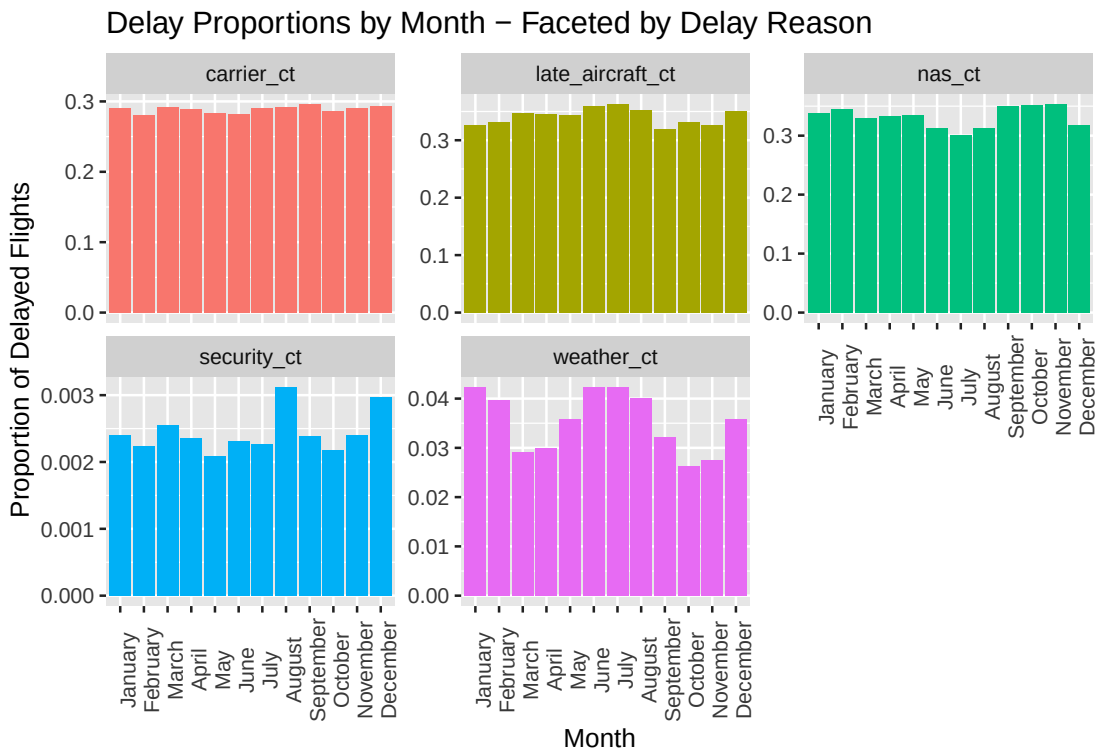
By plotting the relationship between reasons for flight delays and the time of year, we can try to see if there is a certain time of year where a certain flight delay reason occurs more and see if the greater proportion of flight delays in June, July, and December is due to a specific reason.

3.1.2: Reasons for Flight Delays vs. Time of year

```
delay_reasons_1 <- air_data %>% ## Calculates proportions of flight delays by reason
  pivot_longer(cols = "carrier_ct":"late_aircraft_ct", names_to = "reason", values_to = "amount") %>%
  group_by(reason, month) %>%
  summarise(total_delay_prop = sum(amount) / sum(arr_del15))

delay_reasons_1 %>% ## Shows the proportion of flight delay reasons by month
  ggplot(aes(x = reason, y = total_delay_prop, fill = reason)) +
  geom_col() +
  facet_wrap(~month) +
  labs(x = "Reason", y = "Proportion of Delayed Flights", title = "Delay Proportions by
       Reason - Faceted by Month") +
  guides(fill = "none") +
  theme(axis.text.x = element_text(angle = 45)) +
  theme(axis.text.x = element_text(vjust= 0.4))
```




```
delay_reasons_1 %>% ## Compares the monthly proportions of flight delay reasons
  group_by(reason) %>%
  ggplot(aes(x = month, y = total_delay_prop, fill = reason)) +
  geom_col() +
  facet_wrap(~reason, scales = "free_y") +
  labs(x = "Month", y = "Proportion of Delayed Flights", title = "Delay Proportions by Month - Faceted
  guides(fill = "none") +
  theme(axis.text.x = element_text(angle = 90))
```



Looking at the uniform distribution in the proportions of flight delays due to the different flight delay reasons, there is no significant flight delay reason to explain the increase in flight delays in June, July, and December, aside from security and weather delays.

We can look at the less uniform distributions of security and weather delays to see if this can explain any events or relationships over the months. There is an increase in security delays in the months of August and December, as well as an increase in weather delays in the summer (June, July, August) and winter (December, January, February) months.

The increase in security delays in August can be explained by the volume of returning passengers due to end of summer travel. Airlines and airports have to deal with a greater amount of luggage and bagging, which ultimately leads to more delays due to security issues when checking the bags and with TSA.

```
library(knitr)
#include_graphics("augustsecurity.png")
```

Taken from: https://www.tsa.gov/news/press/releases/2017/07/27/seattle-tacoma-international-airport-prepared-busiest-august-record

The security delays in December can also be explained with similar reasoning. December is a popular time for people to travel for the holiday season, and the increased number of bags and carry-ons is the likely reason for the increase in security and TSA lines. As a result, we can expect a greater amount of security delays in December.

```
library(knitr)
#include_graphics("christmastravel.png")
#include_graphics("christmastsa.png")
```

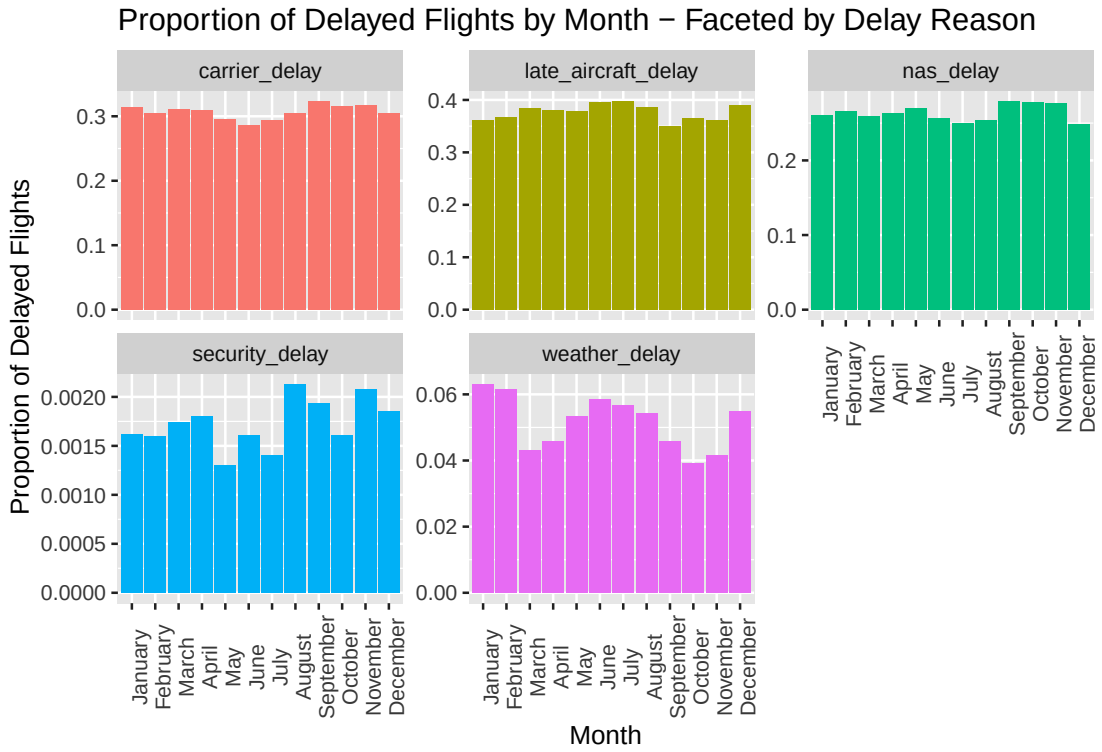
Taken from: https://www.nerdwallet.com/article/travel/flying-for-christmas-best-days#:~:text=Is%20Christmas%20Day%20a%20good,from%20the%20holiday%20you%20get. https://www.foxbusiness.com/lifestyle/tsa-busiest-travel-days-christmas-holiday

On the other hand, the weather delays can be attributed to the yearly patterns of weather. Summer weather has a greater amount of thunderstorms and hot weather, both of which can delay or cancel flights, while winter weather generally means colder temperatures and inclement weather.

This also varies by region, which will be covered further in Section 3.2 (Weather Delays).

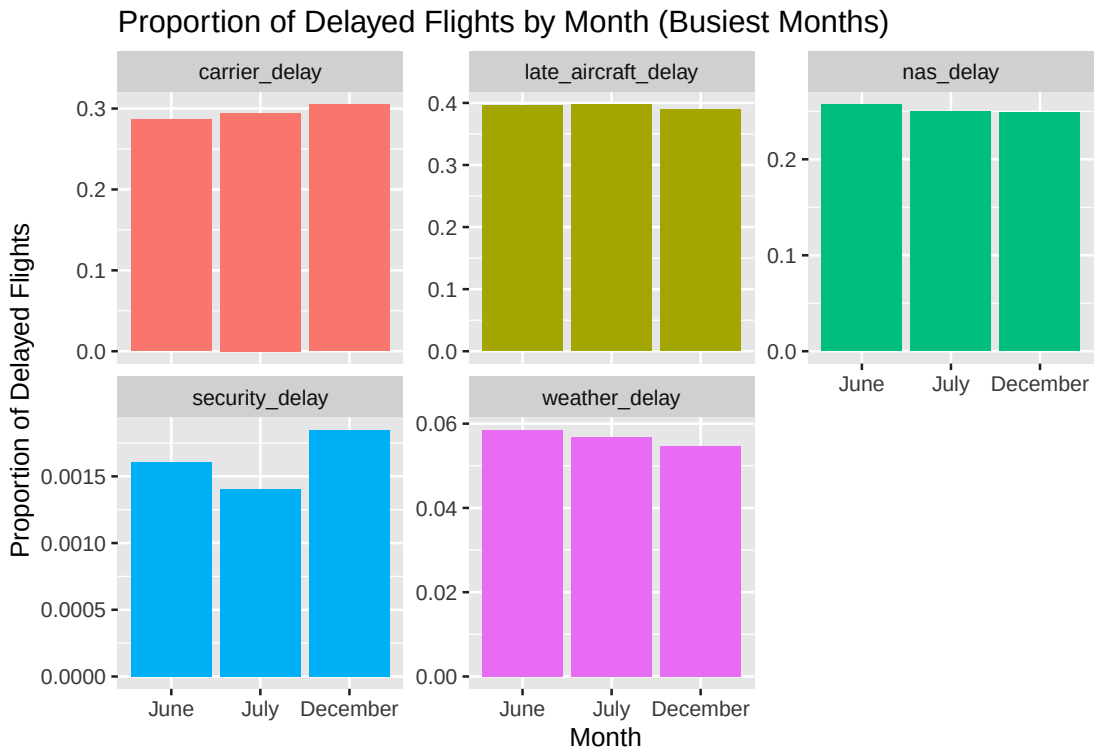
3.1.3: Length of delay vs. time of year:

```
delay_time <- air_data %>% ## Gets delay times
  pivot_longer(cols = "carrier_delay":"late_aircraft_delay", names_to = "reason", values_to = "amount")
  select(month, reason, amount, arr_delay) %>%
  group_by(reason, month) %>%
  summarise(total_delay_time_prop = sum(amount) / sum(arr_delay))

delay_time %>%
  group_by(reason) %>%
  ggplot(aes(x = month, y = total_delay_time_prop, fill = reason)) +
  geom_col(position = "dodge") +
  facet_wrap(~reason, scale = "free_y") +
  labs(x = "Month", y = "Proportion of Delayed Flights", title = "Proportion of Delayed Flights by Month
  guides(fill = "none") +
  theme(axis.text.x = element_text(angle = 90))
```




```
delay_time %>%
  filter(month %in% c("June", "July", "December")) %>%
  group_by(reason) %>%
  ggplot(aes(x = month, y = total_delay_time_prop, fill = reason)) +
```

```
  geom_col(position = "dodge") +
  facet_wrap(~reason, scale = "free_y") +
  labs(x = "Month", y = "Proportion of Delayed Flights", title = "Proportion of Delayed Flights by Month
  guides(fill = "none")
```



After analyzing the distribution of flight delay proportions by time delayed instead of the number of flights delayed, the findings are very similar to the previous graphs. The only notable difference is an increase in security delays in November, which indicates that the proportion of minutes spent in security delays in November was greater relative to the proportion of flights that were delayed due to security reasons.

This increase could be explained by Thanksgiving travel, where TSA officers and other security workers at airports must deal with a greater volume of travelers.

```
library(knitr)
#include_graphics("novembertsa.png")
```

Taken from: https://www.tsa.gov/news/press/releases/2018/11/13/tsa-administrator-2018-holiday-travel-period-expected-be-busiest

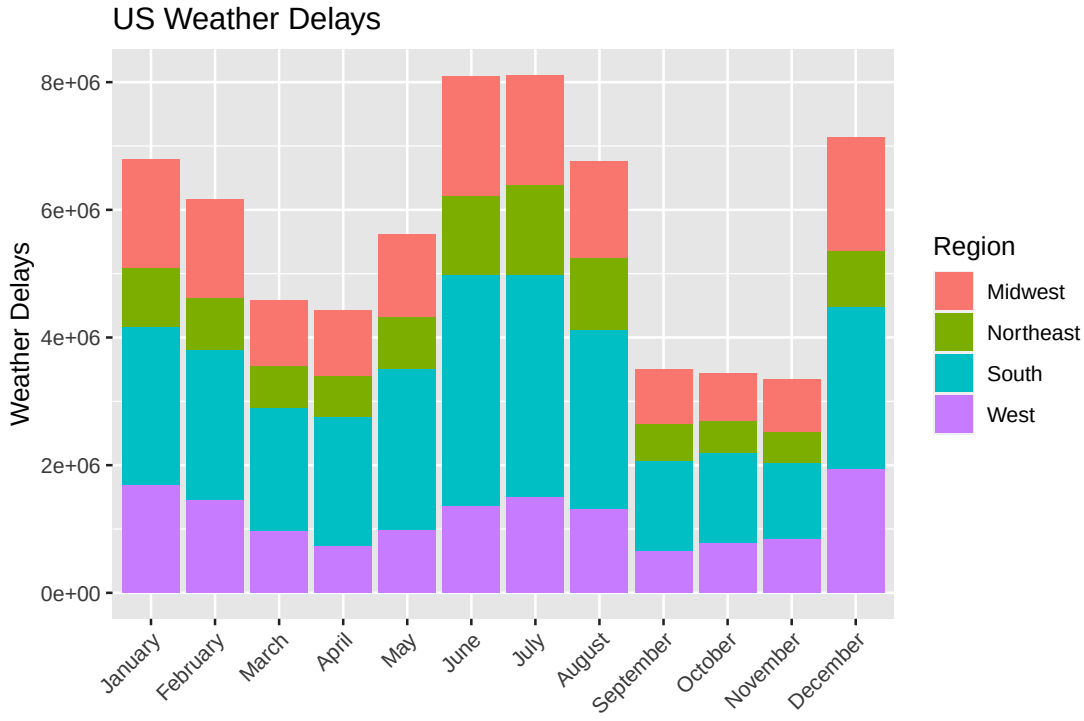
Overall, we found a few interesting relationships between variables by looking at flight delays by time of year. - The summer and holiday months have a greater amount of flight delays than other months, presumably due to the greater number of travelers. - The flight delay reasons that varied the most through time of year were delays due to security and weather. - While other flight delay reasons were relatively unaffected, security delays increased during busy travel times (August) and during the holiday season (November, December). - Weather delays were most prominent during summer and winter months due to a greater likelihood of inclement weather affecting flights.

3.2: Weather Delays and Regional Differences

3.2.1: Weather Delays Overall

We want to check how weather affects delays. Because weather can fluctuate drastically across the US, we also need to look at delays by region, by month, and by year for a more reliable analysis.

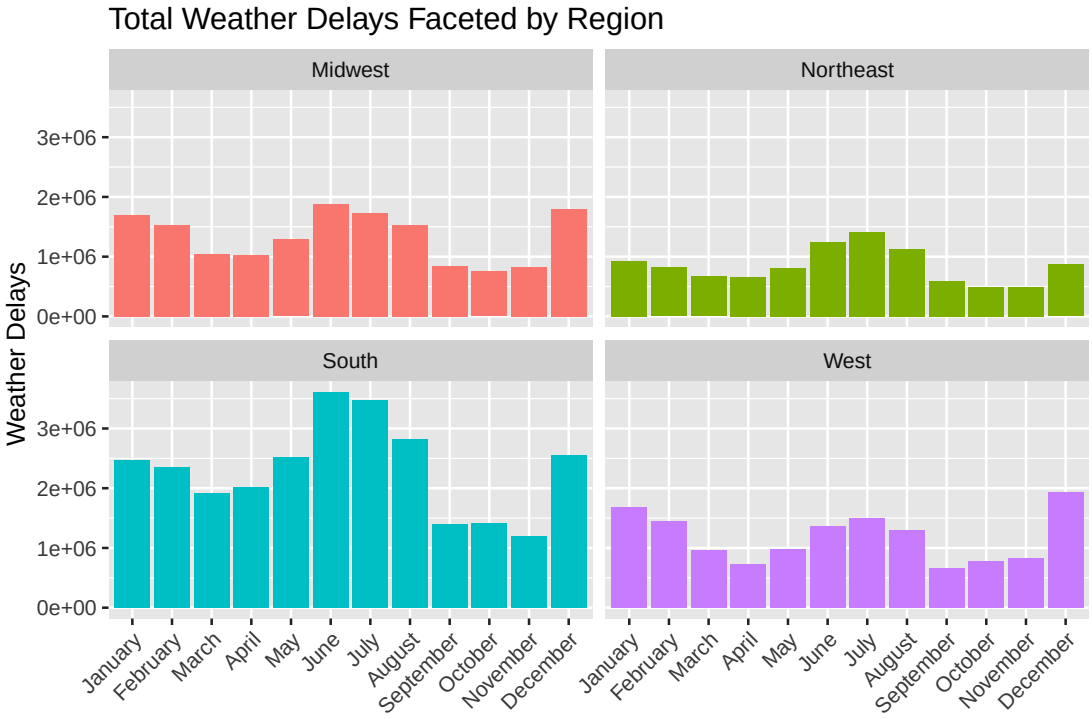
```
air_data %>%
  group_by(Region, month) %>%
  summarise(total_weather_delay = sum(weather_delay)) %>%
  ggplot(aes(x=reorder(month, +total_weather_delay), y=total_weather_delay, fill = Region)) + geom_col()
  xlab("") + ylab("Weather Delays") + labs(title = "US Weather Delays")
```



This analyzes average weather delays from 2003-2022 by region. Because it covers nearly two decades of data, we can see a general pattern in the delays induced by a region's typical climate, as well as the trends for the entire country. There are large peaks in June-July and December-January for all regions.

```
air_data %>%
  group_by(Region, month) %>%
  summarise(total_weather_delay = sum(weather_delay)) %>%
  ggplot(aes(x=reorder(month, +total_weather_delay), y=total_weather_delay, fill = Region)) + geom_col()
  xlab("") + ylab("Weather Delays") + facet_wrap(~ Region) +
  guides(fill = "none") + labs(title = "Total Weather Delays Faceted by Region")
```

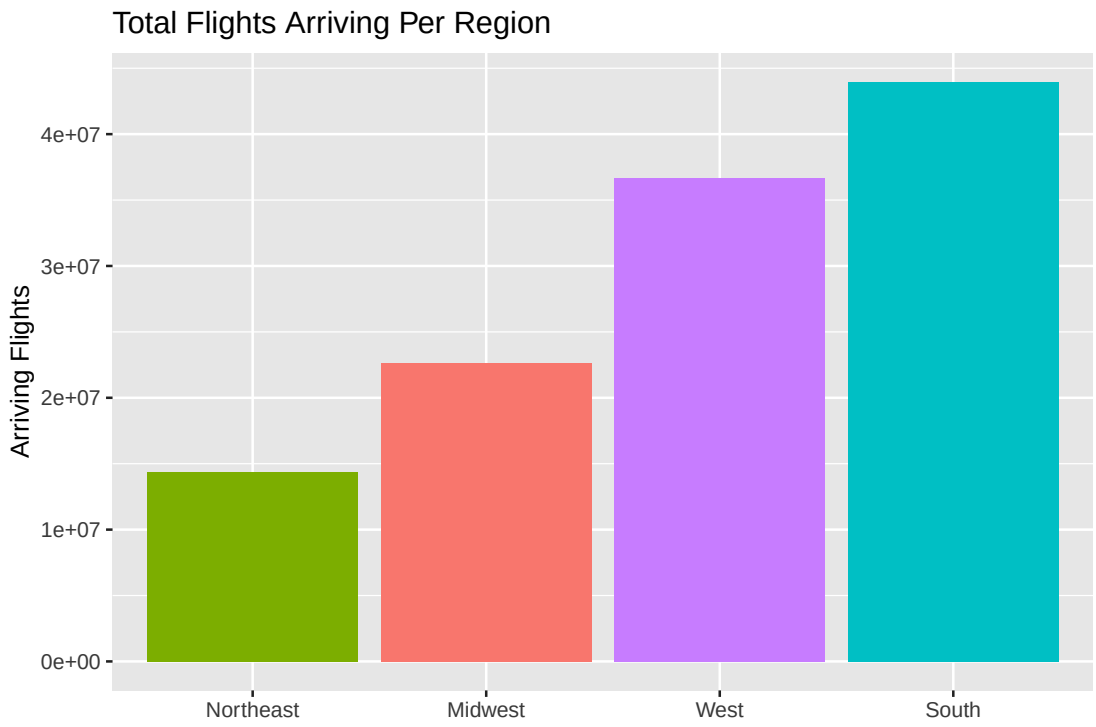
```
## `summarise()` has grouped output by 'Region'. You can override using the
## `.groups` argument.
```



The second graph shows weather delays, but now faceted by region. We see that the South generally has more weather related delays, especially in June and July. All regions have the most weather delays in these months, except for the West, which has more delays in January and December. The Midwest has weather delay peaks in both June-August and December-February.

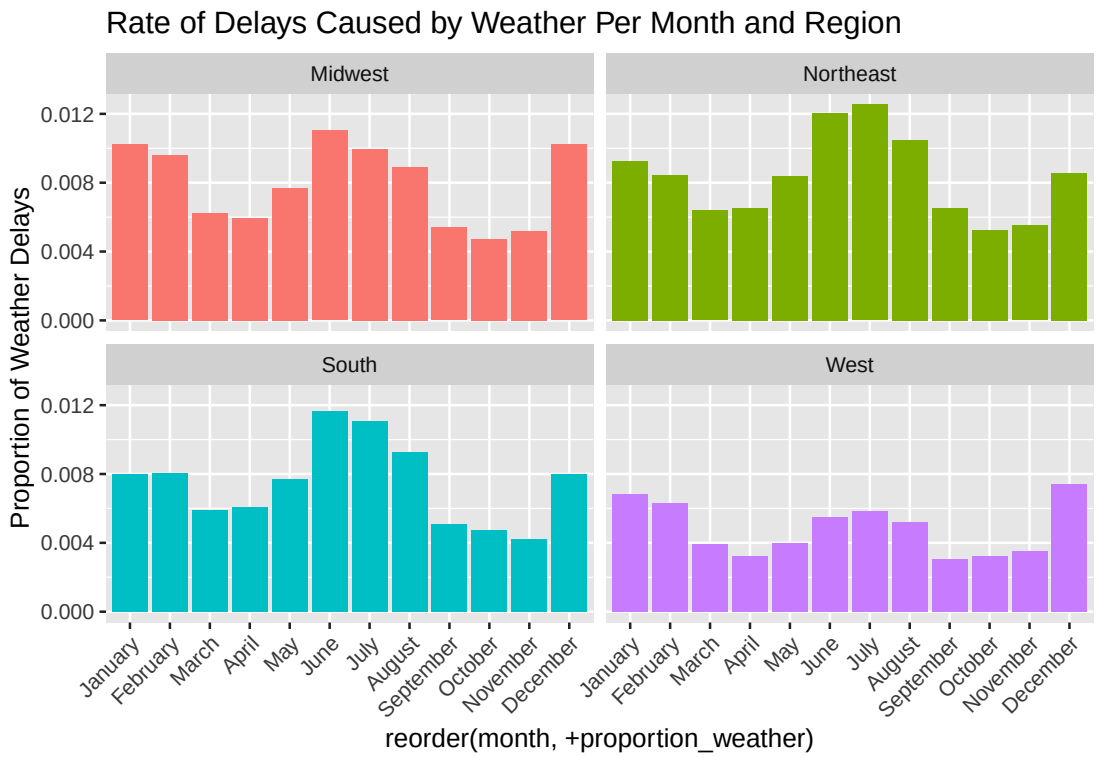
One question that arises from this is whether or not there is a significant variance in flights per year and per region. Clearly more flights over a certain period of time and/or in a certain region would result in a greater amount of delays.

```
air_data %>%
  group_by(Region) %>%
  summarise(total_arr_flights = sum(arr_flights)) %>%
  ggplot(aes(x=reorder(Region, +total_arr_flights), y=total_arr_flights, fill = Region)) + geom_col() +
  guides(fill = "none") + labs(x="", y="Arriving Flights", title="Total Flights Arriving Per Region")
```

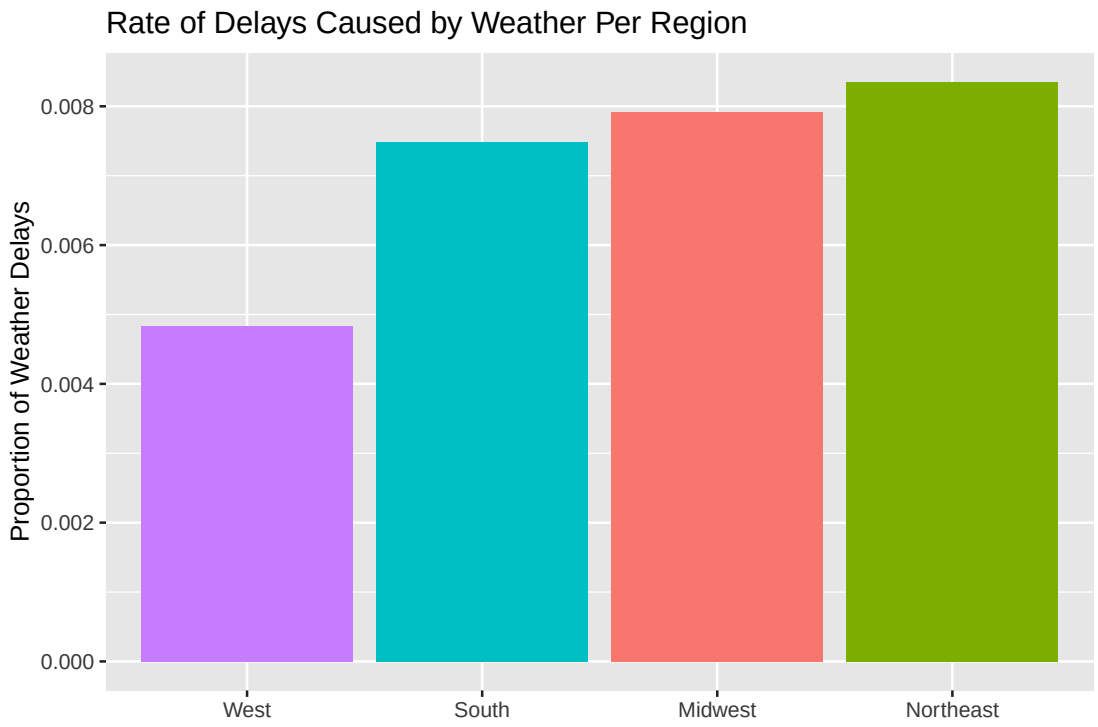


As expected, the South has the greatest amount of arriving flights, so it makes sense for the region to also have the greatest amount of delays. We need to calculate the proportion of weather delays instead.

```
proportion_weather_delay <- air_data %>%
  group_by(Region, month) %>%
  summarise(across(c(arr_flights:late_aircraft_delay), sum, na.rm = TRUE), proportion_weather = weather_

ggplot(proportion_weather_delay, aes(x=reorder(month, +proportion_weather), y=proportion_weather, fill=R
  guides(fill = "none") + labs(y="Proportion of Weather Delays", title = "Rate of Delays Caused by Weath
```



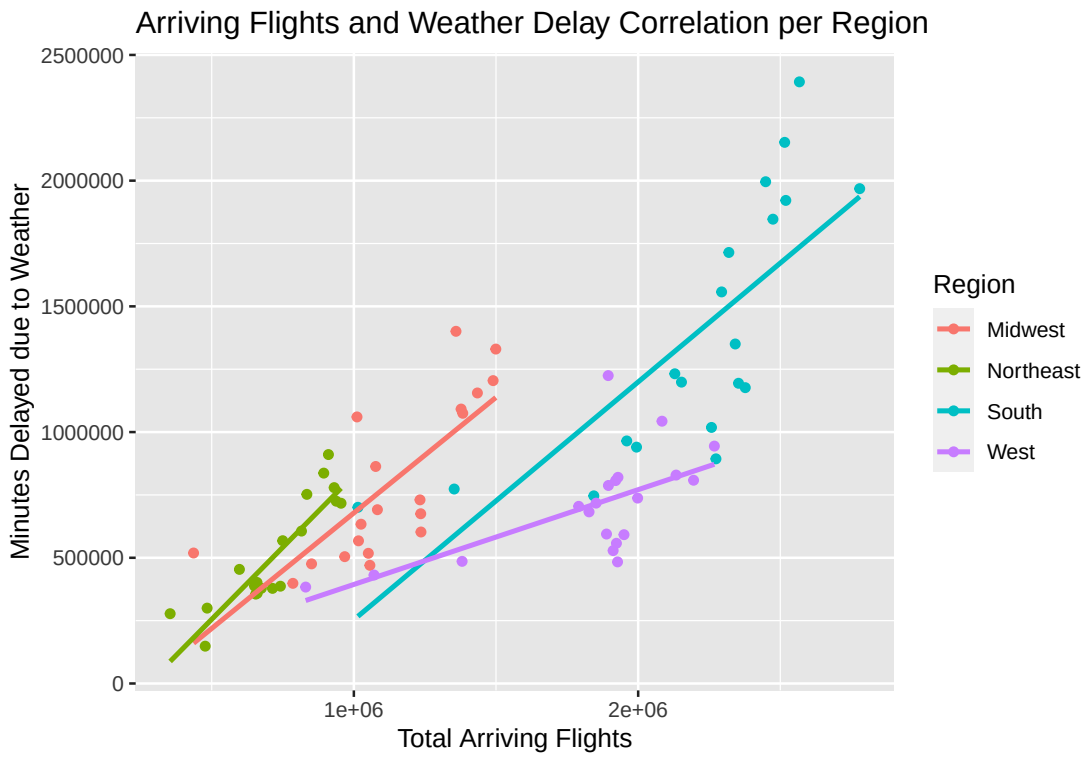

```
air_data %>%
  group_by(Region) %>%
  summarise(total_proportion_weather = sum(weather_ct)/sum(arr_flights)) %>%
  ggplot(aes(x=reorder(Region, +total_proportion_weather), y=total_proportion_weather, fill = Region)) +
  geom_col() +
  guides(fill = "none") +
  labs(x="", y="Proportion of Weather Delays", title= "Rate of Delays Caused by Weather Per Region")
```



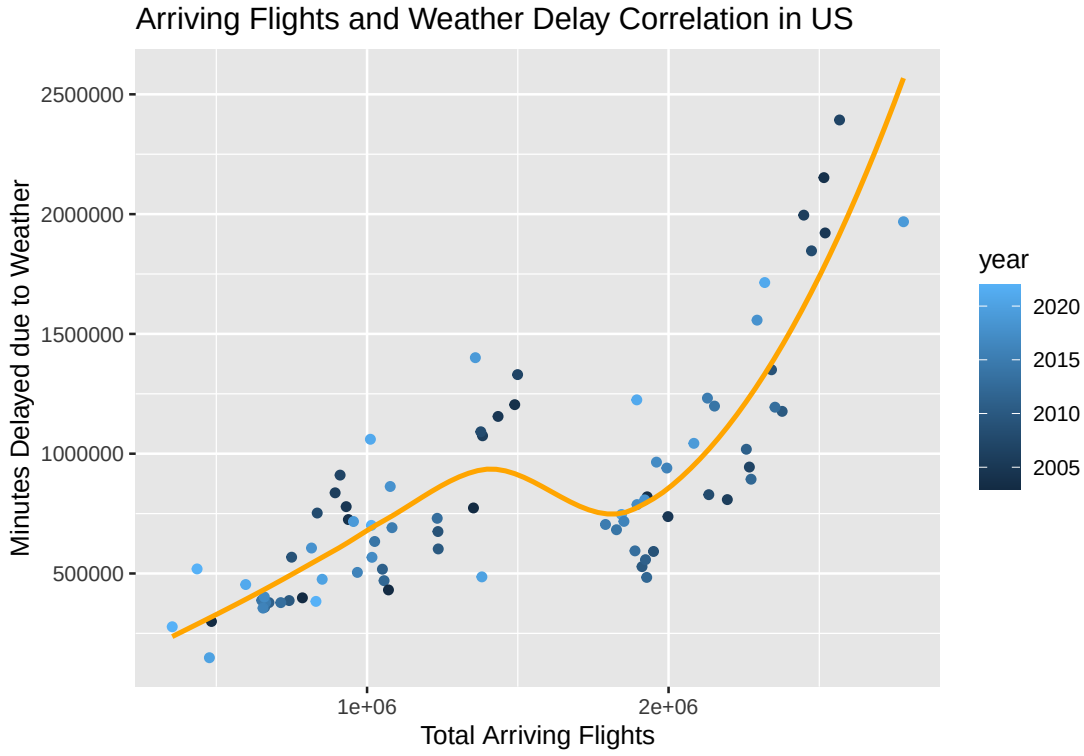
After calculating the proportions, we see that although the South has the most flights delayed due to weather, that is because the South has the most incoming flights overall, and not because the South has a higher rate of flights being delayed. Instead, the Northeast and Midwest have the highest total rate of weather delays.

Because we cannot directly calculate a proportion for the number of arriving flights to weather delay (which is measured in minutes) we want to verify the correlation between these two variables, and expect a positive correlation.

```
arr_flights_weather_delay_comp <- air_data %>%
  group_by(Region, year) %>%
  summarise(total_weather_delay = sum(weather_delay), total_arr_flights = sum(arr_flights))

  ggplot(arr_flights_weather_delay_comp, aes(x=total_arr_flights, y=total_weather_delay, color = Region)
  labs(x = "Total Arriving Flights", y = "Minutes Delayed due to Weather", title = "Arriving Flights and
```




```
ggplot(arr_flights_weather_delay_comp, aes(x=total_arr_flights, y=total_weather_delay, color=year)) +
labs(x = "Total Arriving Flights", y = "Minutes Delayed due to Weather", title = "Arriving Flights an
```

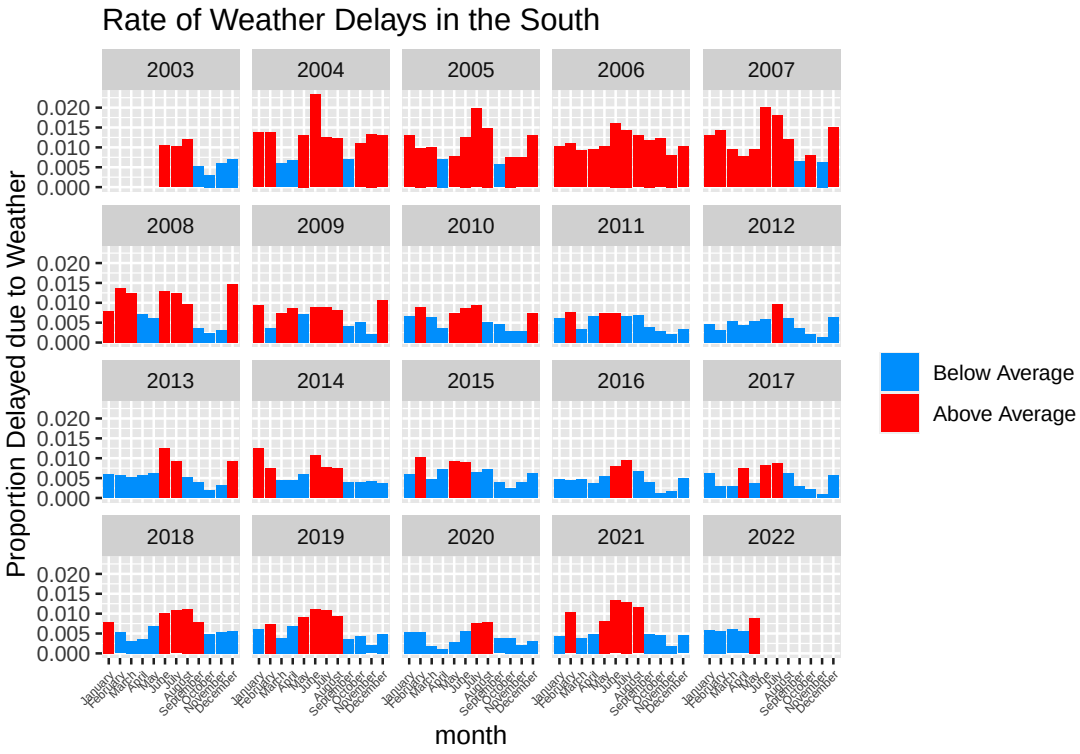


As each color represents a region, the fact that the points are clustered roughly by color shows that the number of

arriving flights is generally consistent throughout the 18 years of data provided.As expected, more flights equates to more delays.

3.2.2: Weather Delays by Region

```
air_data %>%
  filter(Region == "South") %>%
  group_by(year, month) %>%
  summarise(proportion_delayed = sum(weather_ct)/sum(arr_flights)) %>%
  ggplot(aes(x=month, y = proportion_delayed, fill = proportion_delayed > mean(proportion_delayed))) +
  geom_col() +
  facet_wrap(~year) +
  scale_x_discrete(limits = month.name) +
  theme(axis.text.x = element_text(angle=45, hjust=1, size = 5)) +
  ggtitle("South") +
  scale_fill_discrete(limits = month.name) +
  scale_fill_manual(values = c("#008EFC","#FF0000"), labels=c('TRUE'='Above Average','FALSE'='Below Ave
  labs(fill='', title = "Rate of Weather Delays in the South", y="Proportion Delayed due to Weather" )
```



South: Delays generally occur in June or July (summer)

Midwest: Delays generally occur in December and January (Autumn) and sometimes June-August

Northeast: Delays generally occur in June and July, and also occasionally December-February.

West: Delays generally occur in December and January.

Now with data of the rates of weather delays visualized for each month, year, and region, we can very easily see outliers. The weather_ct category in this dataset represents extreme weather related events like

hurricanes, ice storms, and tornados. A spike in the proportion of flights delayed for weather related reasons in a specific month indicates that there was critically disruptive weather.

South: June 2004 We see that June 2004 has abnormally high amounts of weather delays in the South.

```
air_data %>%
  filter(Region == "South", year == 2004) %>%
  group_by(month) %>%
  summarise(
    total_weather_delay = sum(weather_delay),
    total_arr_flights = sum(arr_flights),
    ) %>%
  print()
```

```
## # A tibble: 12 x 3
##    month     total_weather_delay total_arr_flights
##    <fct>                   <dbl>             <dbl>
##  1 January                197690            207714
##  2 February               177232            197246
##  3 March                   79470            214323
##  4 April                   83648            207875
##  5 May                    222585            210658
##  6 June                   367094            205358
##  7 July                   177324            213167
##  8 August                 184473            215049
##  9 September              100211            203770
## 10 October                145511            214697
## 11 November               207755            208379
## 12 December               209369            217164
```

This shows us that June had the most weather delays (367094 minutes of delays, which is equivalent to 6118 hours of delays, or about 255 days of collective delays). Why?

https://en.wikipedia.org/wiki/2004_Atlantic_hurricane_season

The 2004 Atlantic hurricane was exceptionally destructive in the US, with over half of the 16 tropical cyclones in the Atlantic hitting the US. This was a weather anomaly due to the formation of a rare type of El Niño in the Pacific.

Midwest: December 2007

```
air_data %>%
  filter(Region == "Midwest", year == 2007 | year == 2008, month == "December") %>%
  group_by(year, month) %>%
  summarise(proportion_delayed = sum(weather_ct)/sum(arr_flights)) %>%
  print()
```

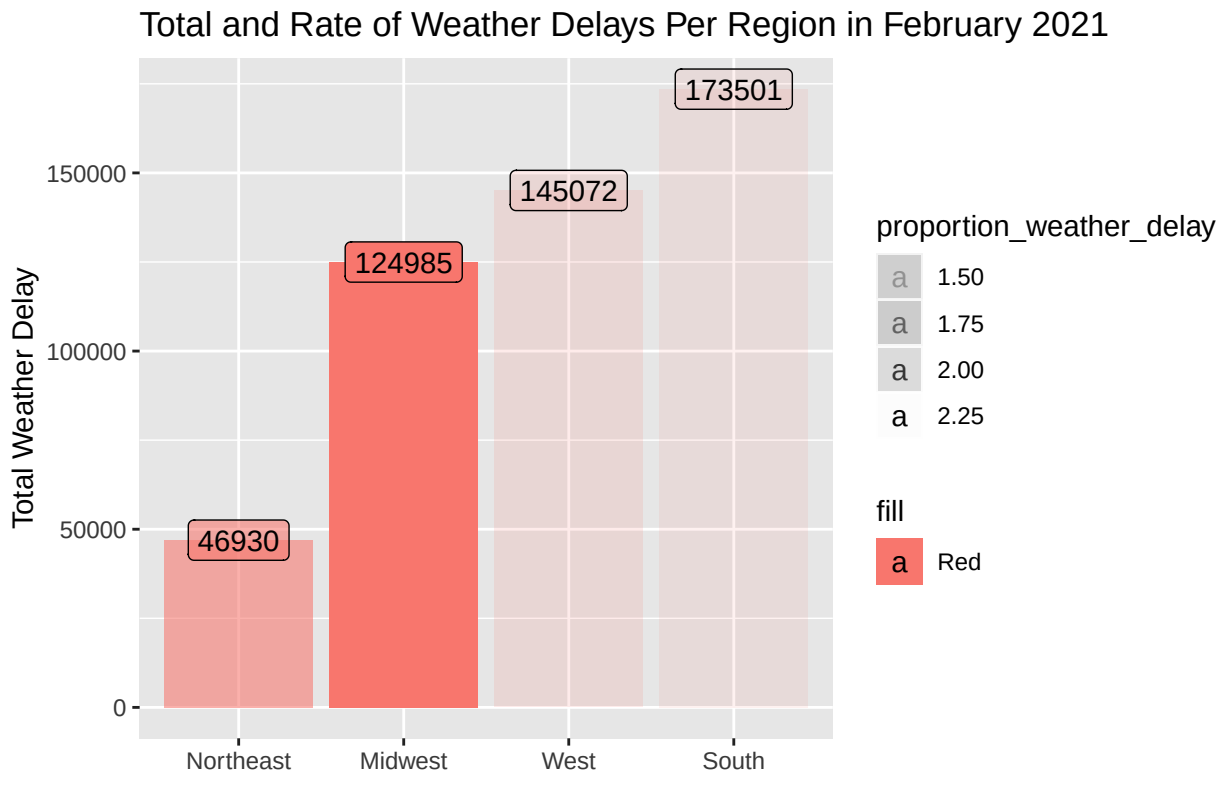
```
## # A tibble: 2 x 3
## # Groups:   year [2]
##    year month    proportion_delayed
##   <int> <fct>                 <dbl>
## 1  2007 December             0.0190
## 2  2008 December             0.0190
```

We see that January December of 2007 and 2008 in the Midwest had extreme delay rates that were nearly identical. By searching online for reports on (what we assume) are winter storms in the Midwest during this time, we see that December 2007 had two severe winter storms that killed over 50 people, and primarily affected the Midwest. https://en.wikipedia.org/wiki/Mid-December_2007_North_American_winter_storms

2008 had a similar season of winter storms. For example, in Rochester, Wisconsin, the winter storms made December 2008 the third snowiest month on record. https://www.weather.gov/arx/dec2008

Another unique major spike occurs simultaneously across the country from a much more recent weather anomaly in 2021; Winter Storm Uri (https://en.wikipedia.org/wiki/February_13âĂŞ17,_2021_North_American_winter_storm) which killed at least 290 people and cost nearly $200 billion in damage making it the costliest natural disaster in US history, and the deadliest storm since 1993. This storm affected all parts of the US, as seen with how each Region has weather delays spike during the month of February 2021.

```
air_data %>%
  filter(year == 2021, month == "February") %>%
  group_by(month, Region) %>%
  summarise(
    total_weather_delay = sum(weather_delay),
    total_arr_flights = sum(arr_flights),
    proportion_weather_delay = total_weather_delay/total_arr_flights
    ) %>%
  ggplot(aes(x=reorder(Region, +total_weather_delay), y=total_weather_delay, alpha = proportion_weather
```



A darker shade of red indicates a greater proportion of delayed flights

3.3: Carrier Delays

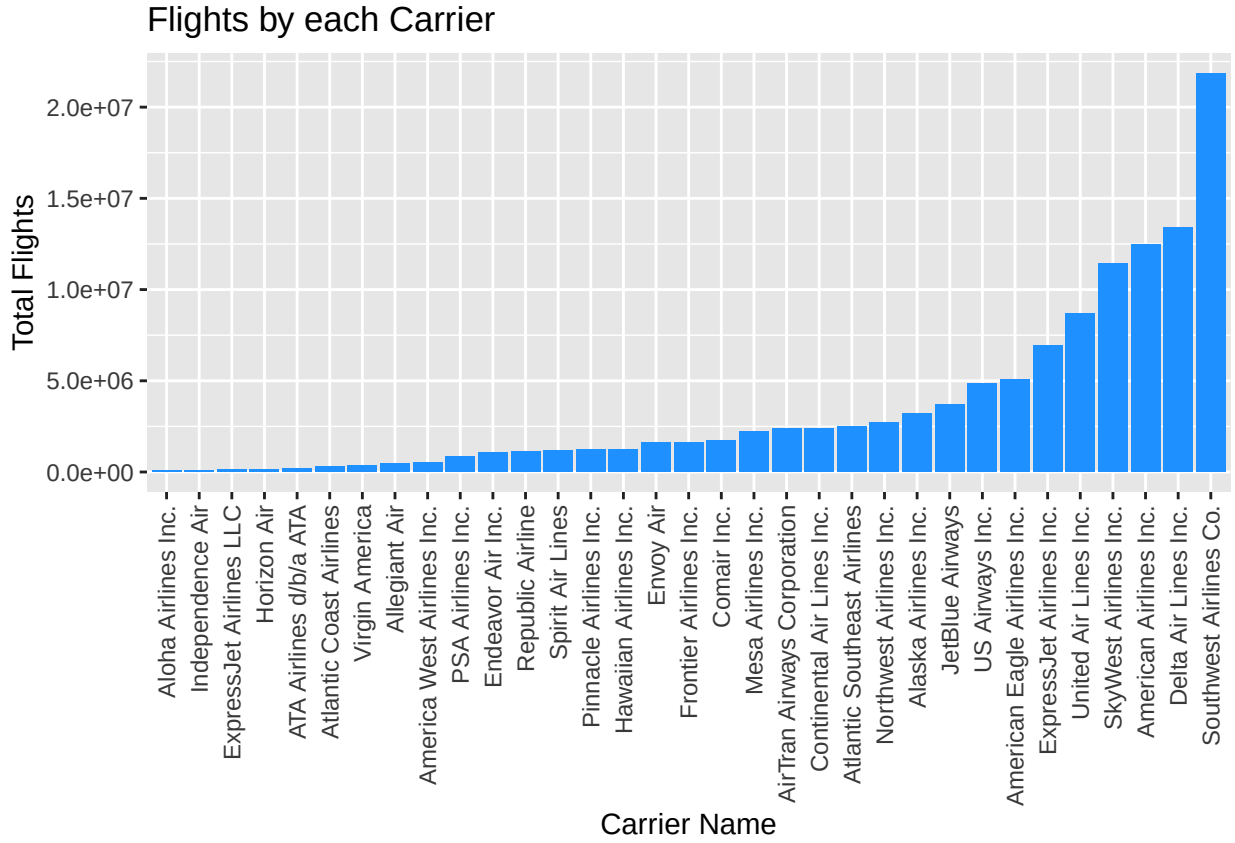
This section explores the variation in carrier delays, uncovering which carriers have the most delays.

```
# Carrier delays account for 30.5% of total delayed minutes
air_data %>%
  summarize(percent_carrier_delay_minutes = sum(carrier_delay)/sum(arr_delay)*100)
```

```
##   percent_carrier_delay_minutes
## 1                      30.52184
```

3.3.1: Flights per Carrier

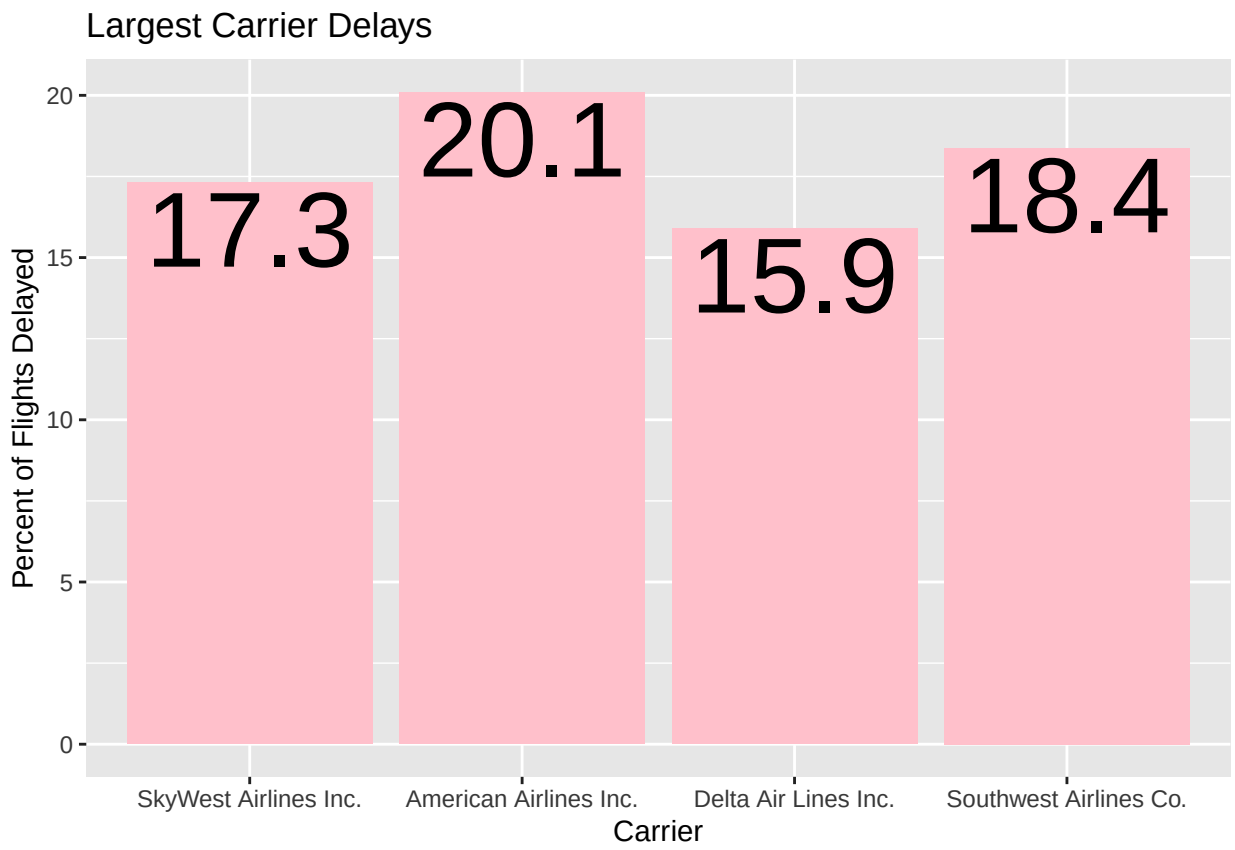
```
air_data %>%
  group_by(carrier_name) %>%
  summarise(total = sum(arr_flights)) %>%
  arrange(desc(total)) %>%
  ggplot(aes(x=reorder(carrier_name, +total), y=total)) +
  geom_col(fill = "dodgerblue") +
  labs(x = "Carrier Name", y = "Total Flights", title = "Flights by each Carrier") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



The graph above shows that Southwest Airlines is the largest carrier by far (with over 18% of flights), followed by Delta Air Lines, American Airlines, and SkyWest Airlines. Together, these four carriers accounted for 50.27% of all US flights from June 2003 to May 2022.

3.3.2: Largest Carriers

```
large_carriers_name <- c("Southwest Airlines Co.", "Delta Air Lines Inc.",
                         "American Airlines Inc.", "SkyWest Airlines Inc.")
air_data %>%
  filter(carrier_name %in% large_carriers_name) %>%
  group_by(carrier_name) %>%
  summarize(proportion_delayed = sum(arr_del15)/sum(arr_flights), flights = sum(arr_flights)) %>%
  ggplot(aes(x = reorder(carrier_name, +flights), y = proportion_delayed * 100)) +
  geom_col(fill = "pink") +
  geom_text(aes(label = round(proportion_delayed, digits = 3) * 100), vjust = 1.1, size = 14) +
  labs(x = "Carrier", y = "Percent of Flights Delayed", title = "Largest Carrier Delays")
```



While Southwest has the most flights, Delta Air Lines comes first for the most reliability.

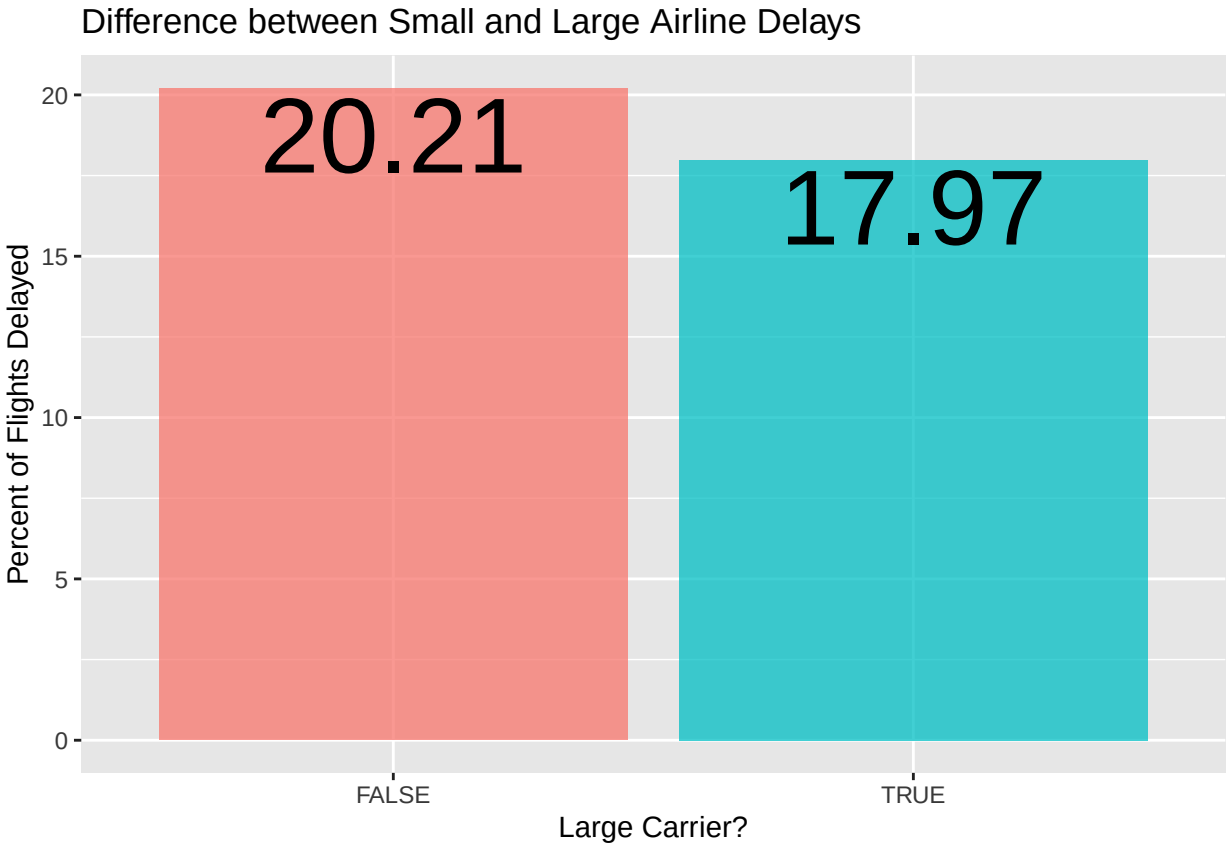
```
air_data <- air_data %>%
  mutate(large_carrier = carrier_name %in% large_carriers_name)
```

```
air_data %>%
  filter(large_carrier == TRUE) %>%
  summarize(large_carrier_average_delay = sum(arr_del15)/sum(arr_flights))
```

```
##   large_carrier_average_delay
## 1                   0.1797482
```

Large airlines have delayed flights roughly 17.97% of the time.

```
air_data %>%
  group_by(large_carrier) %>%
  summarize(proportion_delayed = sum(arr_del15)/sum(arr_flights)) %>%
  ggplot(aes(x = large_carrier, y = proportion_delayed*100)) +
  geom_col(aes(fill = large_carrier), alpha = 0.75) +
  geom_text(aes(label = round(proportion_delayed, digits = 4)*100), vjust = 1.1, size = 14) +
  labs(x = "Large Carrier?", y = "Percent of Flights Delayed",
       title = "Difference between Small and Large Airline Delays") +
  guides(fill = "none")
```

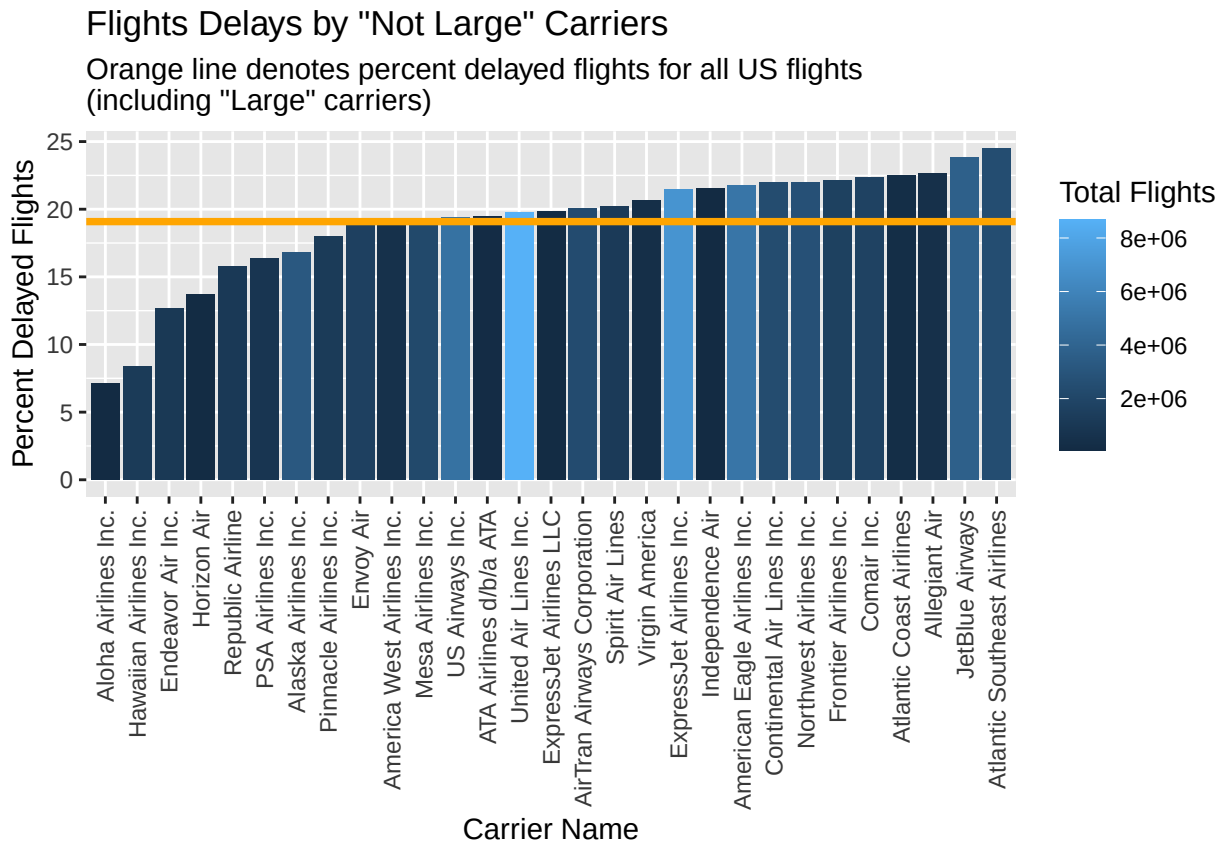


As a whole, the four largest air carriers (roughly half of US flights) are more reliable than the rest.

For curious minds wondering about all the other airline carriers:

```
air_data %>%
  filter(!carrier_name %in% large_carriers_name) %>%
  group_by(carrier_name) %>%
  summarise(proportion_delayed2 = sum(arr_del15)/sum(arr_flights), flights = sum(arr_flights)) %>%
  ggplot(aes(x=reorder(carrier_name, +proportion_delayed2), y=proportion_delayed2*100)) +
  geom_col(aes(fill = flights)) +
  geom_hline(yintercept = 19.084, color = "orange", lwd = 1.3) +
  labs(x = "Carrier Name", y = "Percent Delayed Flights", fill = "Total Flights",
       title = "Flights Delays by \"Not Large\" Carriers",
       subtitle = "Orange line denotes percent delayed flights for all US flights
```

```
(including \"Large\" carriers)") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



3.3.3: Unreliable Carriers

The four least reliable carriers are Atlantic Southeast Airlines, JetBlue Airways, Allegiant Air, and Atlantic Coast Airlines.

```
unreliable_carrier <- c("Atlantic Southeast Airlines", "JetBlue Airways",
                        "Allegiant Air", "Atlantic Coast Airlines")
air_data %>%
  filter(carrier_name %in% unreliable_carrier) %>%
  select(year, month, carrier_name, airport, State, arr_flights, arr_del15) %>%
  arrange(desc(arr_del15/arr_flights)) %>%
  filter(arr_flights >= 50) %>% #This just to skip past the rows with very few flights
  head()
```

```
##   year  month                carrier_name airport State arr_flights arr_del15
## 1 2006 August Atlantic Southeast Airlines     ISO    NC          62        53
## 2 2007   July Atlantic Southeast Airlines     ISP    NY          60        46
## 3 2007   July Atlantic Southeast Airlines     ILG    DE          56        42
## 4 2006   July Atlantic Southeast Airlines     ISO    NC          62        46
## 5 2007 August Atlantic Southeast Airlines     ISP    NY          57        42
## 6 2005   July              JetBlue Airways     MSY    LA          93        68
```

```
air_data %>%
  group_by(carrier_name) %>%
  summarize(proportion_delayed = sum(arr_del15)/sum(arr_flights)) %>%
  arrange(desc(proportion_delayed)) %>%
  head(4)
```

```
## # A tibble: 4 x 2
##   carrier_name                proportion_delayed
##   <chr>                                    <dbl>
## 1 Atlantic Southeast Airlines              0.245
## 2 JetBlue Airways                          0.238
## 3 Allegiant Air                            0.226
## 4 Atlantic Coast Airlines                  0.225
```

3.3.4: Duplicated Airline Codes

There is an interesting conundrum with airlines carriers and carrier codes. Some of the carriers share the same carrier code. This happens generally when one carrier is bought out by another, or goes bankrupt.

```
#Notice "9E" and "DH" are doubled. There are also "EV", "MQ", and "OH".
air_data %>%
  group_by(carrier_name, carrier) %>%
  summarize(flights = sum(arr_flights)) %>%
  arrange(carrier)
```

```
## # A tibble: 35 x 3
## # Groups:   carrier_name [33]
##    carrier_name               carrier  flights
##    <chr>                      <chr>      <dbl>
##  1 Endeavor Air Inc.          9E       1066795
##  2 Pinnacle Airlines Inc.     9E       1209189
##  3 American Airlines Inc.     AA      12443139
##  4 Aloha Airlines Inc.        AQ         89547
##  5 Alaska Airlines Inc.       AS       3206011
##  6 JetBlue Airways            B6       3702115
##  7 Continental Air Lines Inc. CO       2389612
##  8 Atlantic Coast Airlines    DH        291890
##  9 Independence Air           DH        103998
## 10 Delta Air Lines Inc.       DL      13388446
## # ... with 25 more rows
```

The complete list of duplicate codes is as follows:

- "9E" : Endeavor Air Inc., Pinnacle Airlines Inc.
- "DH" : Atlantic Coast Airlines, Independence Air
- "EV" : Atlantic Southeast Airlines, ExpressJet Airlines Inc., ExpressJet Airlines LLC
- "MQ" : American Eagle Airlines Inc., Envoy Air
- "OH" : Comair Inc., PSA Airlines Inc.

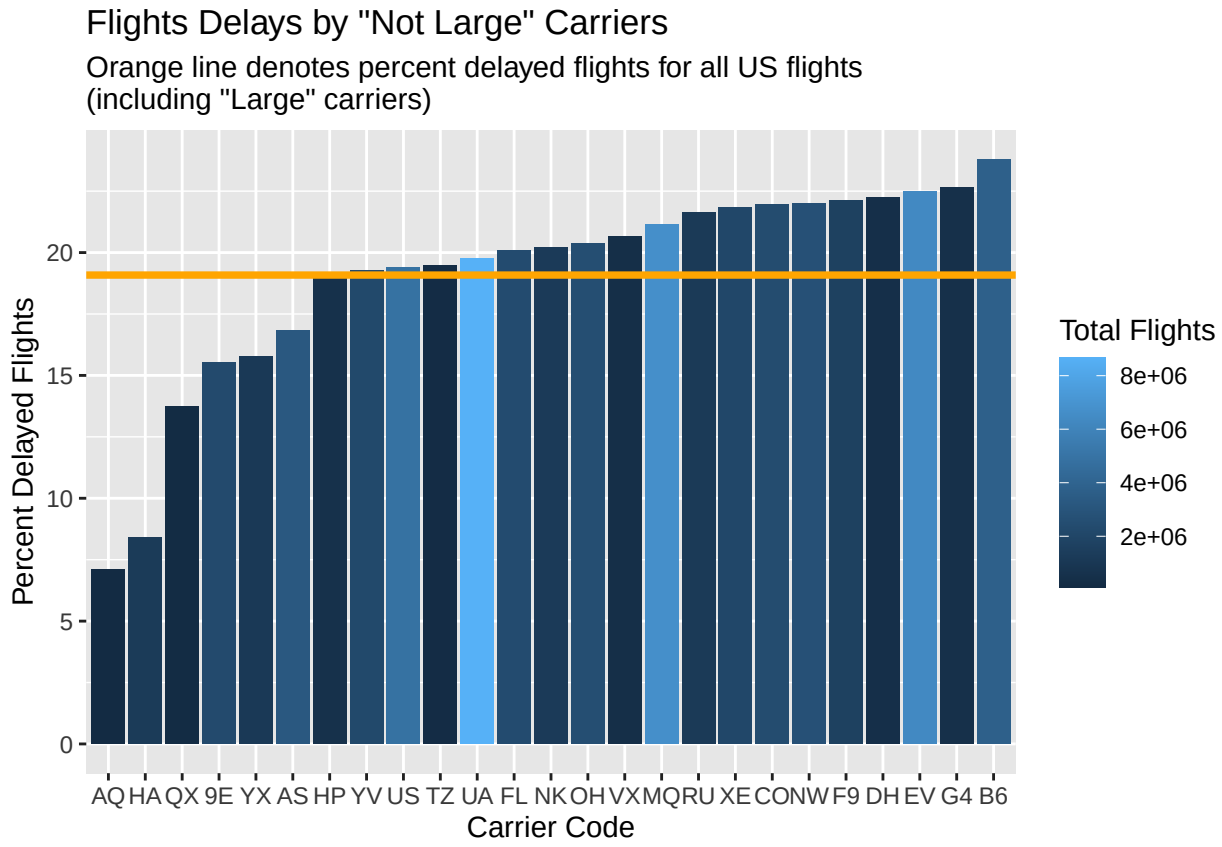
```
duplicate_codes <- c("9E", "DH", "EV", "MQ", "OH")
air_data %>%
  filter(carrier %in% duplicate_codes) %>%
  group_by(year, carrier, carrier_name) %>%
  summarize(proportion_delayed = sum(arr_del15)/sum(arr_flights)) %>%
  ggplot(aes(x = year, y = proportion_delayed, fill = carrier_name)) +
  geom_col(position = "dodge", alpha = 0.9) +
  facet_wrap(~carrier) +
  labs(y = "Proportion Flights Delayed", x = "Year",
       title = "Carrier Codes Reused for Different Airlines",
       subtitle = "Change in color represents new airline") +
  guides(fill = "none")
```

Acknowledging that COVID had a significant impact on flights for all airlines (as will be shown in 3.5.2), none of the airlines had a significant change in reliability when their names changed.

That said, two of the least reliable carriers were Atlantic Southeast Airlines and Atlantic Coast Airlines, which were regional airlines and Delta Air Lines subsidiaries. If you seek out the least reliable airlines by code instead of by name, i.e. grouping together airlines with the same code, there is a new list of least reliable airlines:

```
air_data %>%
  filter(!carrier_name %in% large_carriers_name) %>%
  group_by(carrier) %>%
  summarise(proportion_delayed2 = sum(arr_del15)/sum(arr_flights), flights = sum(arr_flights)) %>%
  ggplot(aes(x=reorder(carrier, +proportion_delayed2), y=proportion_delayed2*100)) +
  geom_col(aes(fill = flights)) +
```

```
  geom_hline(yintercept = 19.084, color = "orange", lwd = 1.3) +
  labs(x = "Carrier Code", y = "Percent Delayed Flights", fill = "Total Flights",
       title = "Flights Delays by \"Not Large\" Carriers",
       subtitle = "Orange line denotes percent delayed flights for all US flights
(including \"Large\" carriers)")
```

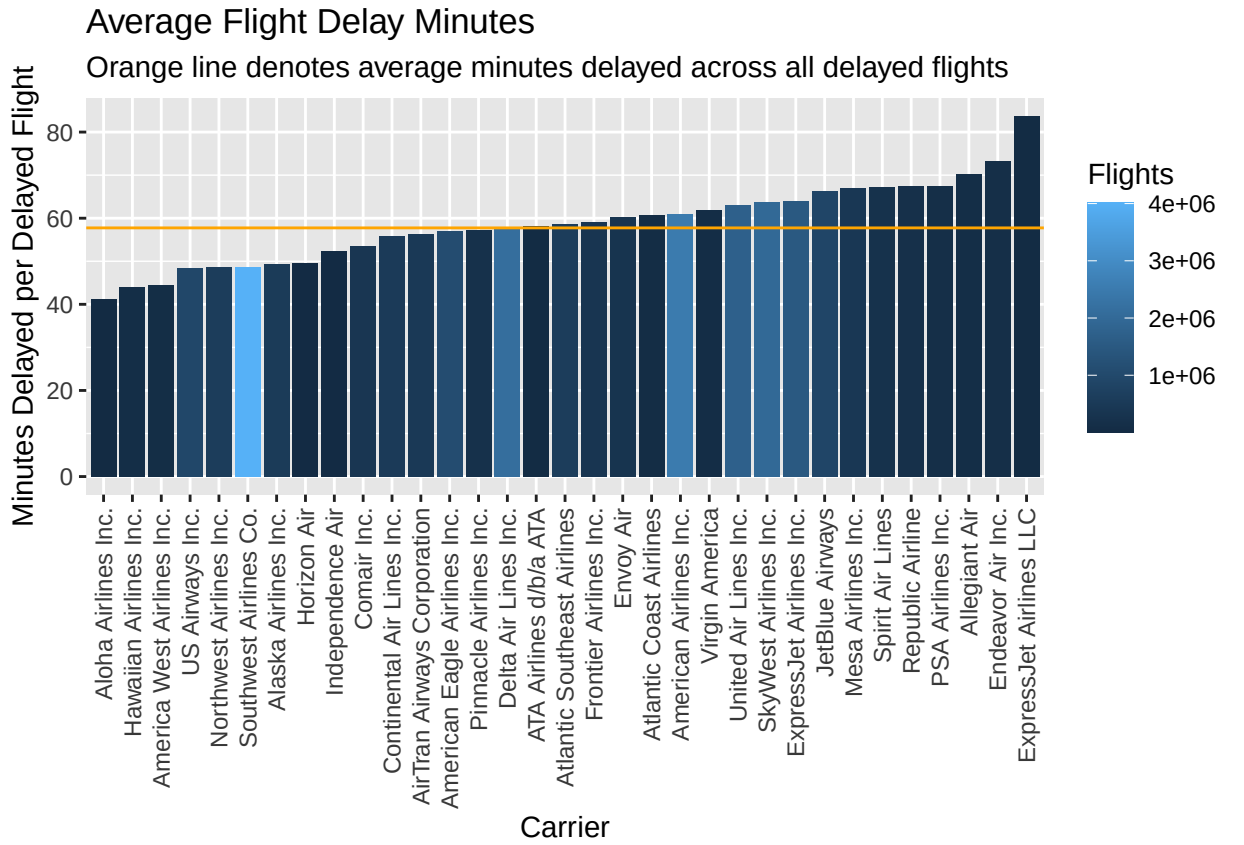


The least reliable airline codes are B6 (JetBlue Airways), G4 (Allegiant Air), EV (Atlantic Southeast Airlines, ExpressJet Airlines Inc., ExpressJet Airlines LLC), and DH (Atlantic Coast Airlines, Independence Air).

3.3.5: Minutes Delayed

Beyond the number of flights delayed, we can examine the amount of time each flight was delayed:

```
air_data %>%
  group_by(carrier_name) %>%
  summarize(minutes_delay = sum(arr_delay)/sum(arr_del15), delayed_flights = sum(arr_del15)) %>%
  ggplot(aes(y = minutes_delay, x = reorder(carrier_name, +minutes_delay), fill = delayed_flights)) +
  geom_col() +
  geom_hline(yintercept = 57.75, color = "orange") +
  labs(y = "Minutes Delayed per Delayed Flight", x = "Carrier", fill = "Flights",
       title = "Average Flight Delay Minutes",
       subtitle = "Orange line denotes average minutes delayed across all delayed flights") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



This shows that ExpressJet Airlines LLC *was* clearly the worst airline when its flights were delayed, with its average delayed flight being over 80 minutes delayed. It also shows that Aloha Airlines Inc. and Hawaiian Airlines Inc. were the best, suggesting that flights to/from Hawaii tend to be less delayed (note: both ExpressJet Airlines and Aloha Airlines are no longer operating due to bankruptcy).

```
#As seen, Hawaii has the fewest delays
air_data %>%
  group_by(State_name) %>%
  summarize(percent_of_flights_delayed = sum(arr_del15)/sum(arr_flights)*100) %>%
  arrange(percent_of_flights_delayed) %>%
  head(3)
```

```
## # A tibble: 3 x 2
##   State_name percent_of_flights_delayed
##   <chr>                           <dbl>
## 1 Hawaii                           11.3
## 2 Utah                             14.1
## 3 Montana                          15.4
```

3.3.6: Summary

Southwest Airlines is the largest airline in the United States, and the top four airlines make up 50% of all US flights. We find that the top four largest airlines have fewer delays than the smaller airlines. Atlantic Southeast Airlines and JetBlue Airways have the highest frequency of delayed flights. During our investigation we found that carrier codes did not necessarily align with carrier names. However, we found that the carrier codes maintained a consistent proportion of delayed flights despite the name changes. Across all the

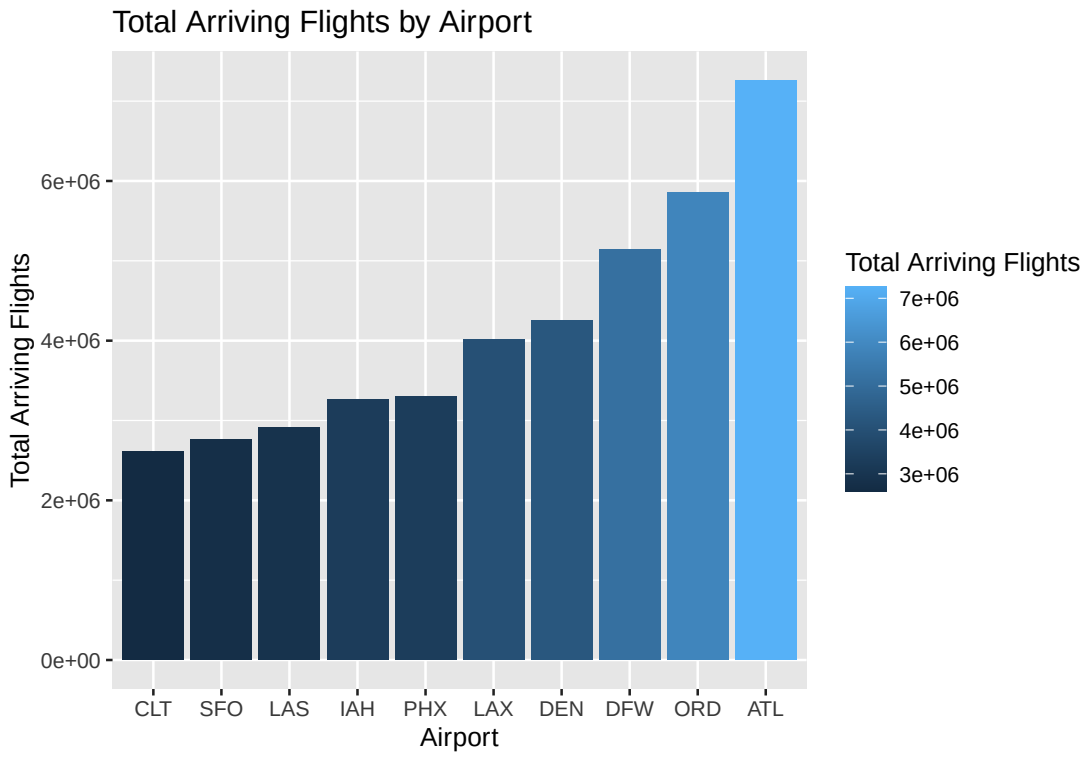
airlines, the average delayed flight is just under one hour delayed. Southwest Airlines has a notably lower average delay time of 48.7 minutes. Passengers should strongly consider which carrier they fly with if they wish to limit their delay time.

3.4: Delays by Airport and Region

Although the numbers and causes of delays differ in each airport, we can look at a few major airports to analyze the frequency and types of delays that they have and compare them with delay averages across all of the airports. This will allow us to get a feel of major airports with a large amount of flight data.

Filter to find airport with most arriving flights:

```
most_arriving_flights <- air_data %>%
  group_by(airport, airport_name) %>%
  summarise(total_arr_flights = sum(arr_flights, na.rm = TRUE)) %>%
  arrange(desc(total_arr_flights))
most_arriving_flights %>%
  head(10) %>%
  ggplot(aes(x=reorder(airport, +total_arr_flights), y=total_arr_flights, fill = total_arr_flights)) +
  geom_col() +
  labs(x = "Airport", y = "Total Arriving Flights", title = "Total Arriving Flights by Airport", fill =
```



It looks like the airport with the most arriving flights is Hartsfield-Jackson Atlanta International Airport in Atlanta, GA. Let's look at the information for this airport to find any interesting relationships and compare it to other airports in the US.

```
air_data %>%
  filter(airport == "ATL") %>%
```

```
  group_by(airport) %>%
  summarise(total_delayproportion = sum(arr_del15) / sum(arr_flights))
```

```
## # A tibble: 1 x 2
##   airport total_delayproportion
##   <chr>                   <dbl>
## 1 ATL                     0.184
```

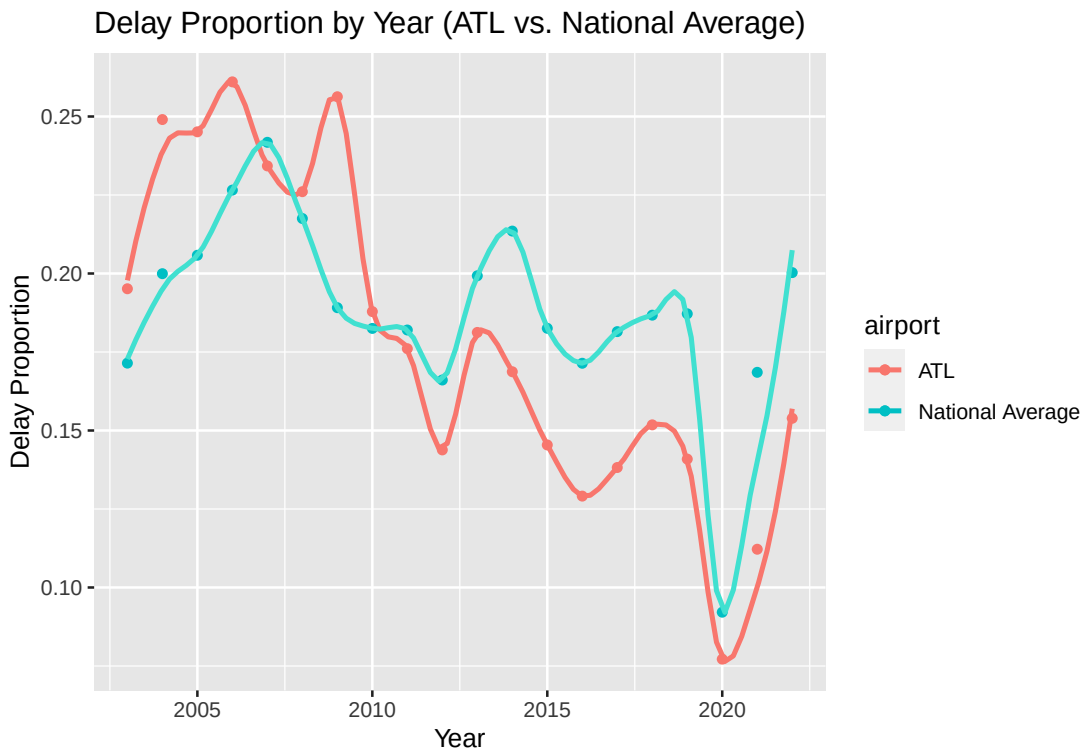
```
air_data %>%
  summarise(total_delayproportion = sum(arr_del15) / sum(arr_flights))
```

```
##   total_delayproportion
## 1             0.1908397
```

Compared to the national average, Atlanta had a lower proportion of flight delays overall. To see if there are any further trends, we can plot the amount of delays over time for Atlanta vs. the national average.

3.4.1: Flight delays vs. time:

```
natl_air_data <- air_data %>% ## Proportion of delays in all airports over time
  group_by(year) %>%
  summarise(delayproportion_year = sum(arr_del15) / sum(arr_flights))

ATL_air_data <- air_data %>% ## Plots the proportion of delays in ATL over time
  filter(airport == "ATL") %>%
  group_by(year, airport) %>%
  summarise(delayproportion_year = sum(arr_del15) / sum(arr_flights))

ggplot(ATL_air_data, aes(x = year, y = delayproportion_year)) +
  geom_point(aes(color = airport)) +
  geom_smooth(span = 0.25, se = FALSE, aes(color = airport)) +
  scale_y_continuous("Delay Proportion") +
  geom_point(data = natl_air_data, aes(x = year, y = delayproportion_year, color = "National Average")) +
  geom_smooth(data = natl_air_data, aes(x = year, y = delayproportion_year), span = 0.25, se = FALSE, co
  labs(x = "Year", y = "Delay Proportion", title = "Delay Proportion by Year (ATL vs. National Average)
```



The data shows that Atlanta Hartsfield-Jackson Airport had a greater than average proportion of flight delays up to 2006-2007, but steadily decreased their flight delay proportion to go below the national average up until today. What happened in 2006?

```
library(knitr)
#include_graphics("delta2007.png")
```

Taken from: https://news.delta.com/26-million-renovation-atlanta-hub-support-future-growth-continues

A possible explanation could be the $26 million renovation project by Delta in 2006 and 2007 to support growth of the airport, which resulted in more resources and funding to decrease the amount of flight delays at the airport.

The number of delays in Atlanta increased against the decreasing trend in 2009, but this can be attributed to a mass technology glitch that affected the airport's aviation system in November.

```
library(knitr)
#include_graphics("atlanta2009.png")
```

Taken from: https://www.latimes.com/archives/la-xpm-2009-nov-20-la-na-flight-delay20-2009nov20-story.html

```
#Arriving flights to ATL by airline
air_data %>%
  filter(airport=="ATL") %>%
  group_by(carrier_name) %>%
  summarise(total = sum(arr_flights)) %>%
  arrange(desc(total)) %>%
  head(1)
```

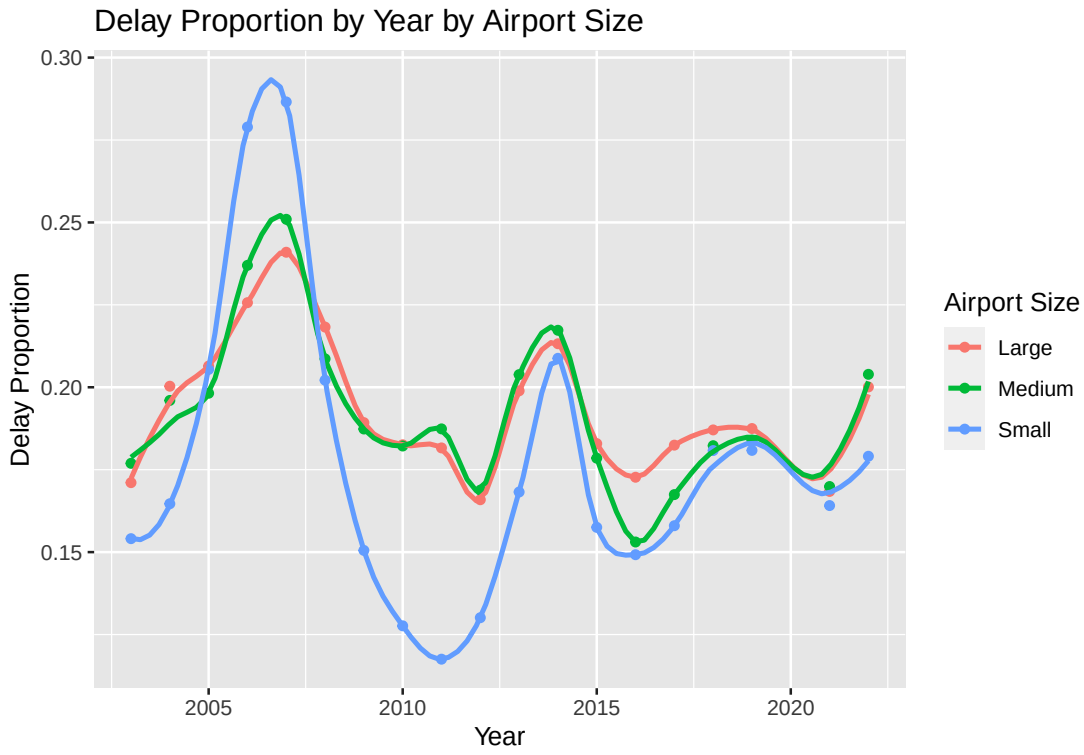
```
## # A tibble: 1 x 2
##   carrier_name           total
##   <chr>                  <dbl>
## 1 Delta Air Lines Inc. 3797408
```

Another thing to note is the renovation by Delta, which makes up the biggest number of arriving flights at the Hartsfield-Jackson Atlanta International Airport. The Atlanta airport is a hub for Delta Airlines, meaning that Delta uses Atlanta as a central transfer point for their flights. This would encourage Delta to invest their funds into the airport, leading to things such as the renovation in 2007 to improve customer service and flight experience.

Now that we've looked at the most popular airport in the US, let's see if airport size is a significant factor in the amount of flight delays using the size data from earlier.

3.4.2: Airport size vs. flight delays over time:

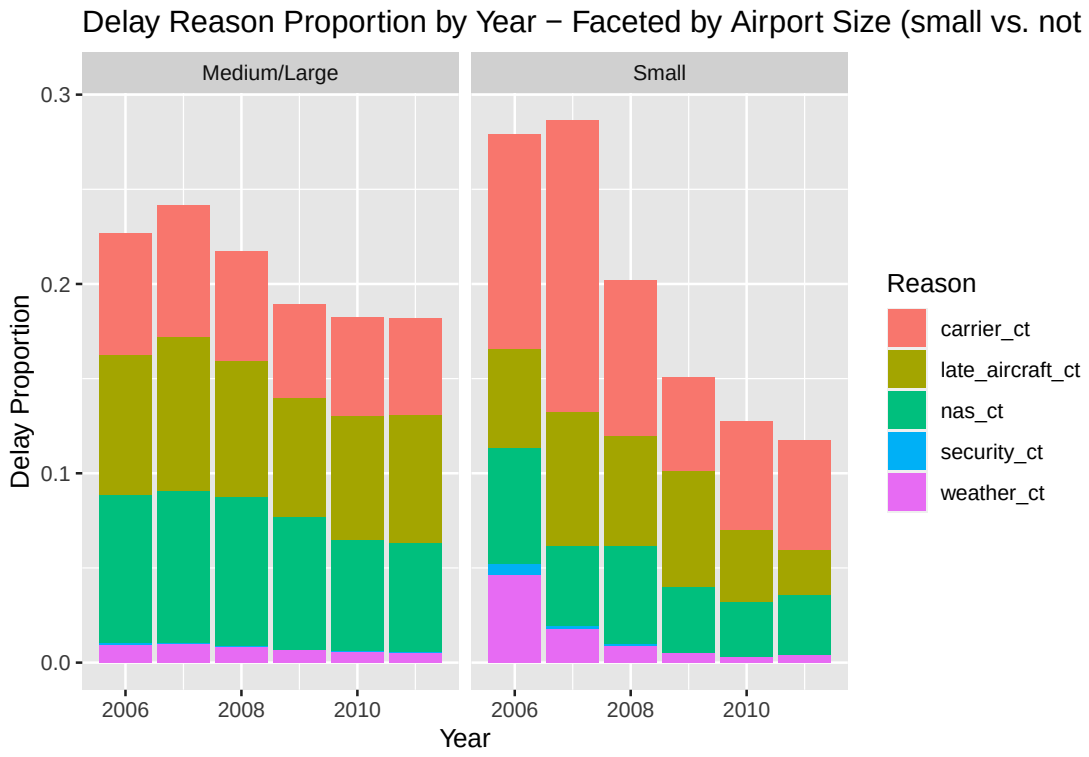
```
air_data %>%
  filter(year!=2020) %>%
  group_by(airport_size, year) %>%
  summarise(delayproportion = sum(arr_del15) / sum(arr_flights)) %>%
  ggplot(aes(x = year, y = delayproportion, color = airport_size)) +
    geom_point() +
    scale_y_continuous('Delay Proportion') +
    geom_smooth(span = 0.3, se = FALSE, aes(color = airport_size)) +
  labs(x = "Year", y = "Delay Proportion", title = "Delay Proportion by Year by Airport Size", color =
```



From this graph, we can see that small airports have consistently had less flight delays than medium and large airports over time, with the exception of years 2006-2010. To investigate these years, we can look at the different reasons for flight delays to see what caused the delays.

3.4.3: Delay Cause vs. Airport Size

```
airport_size_delays <- air_data %>% ## Separates airport size by small vs. medium/large
  mutate(small_yesorno = ifelse(airport_size == "Small", "Small", "Medium/Large")) %>%
  filter(year %in% c(2006:2011)) %>%
  pivot_longer(cols = "carrier_ct":"late_aircraft_ct", names_to = "reason", values_to = "amount") %>%
  group_by(reason, year, small_yesorno) %>%
  summarise(total_delay_proportion = sum(amount) / sum(arr_flights))

airport_size_delays %>% ## Plots delay reason proportions by airport size
  ggplot(aes(x = year, y = total_delay_proportion, fill = reason)) +
  geom_col() +
  facet_wrap(~small_yesorno) +
  labs(x = "Year", y = "Delay Proportion", fill = "Reason", title = "Delay Reason Proportion by Year - I
```



Compared to the medium and large airports, the small airports had a noticeably higher proportion of carrier delays than medium and large airports did from 2006-2011.

One reason for this could be a national pilot shortage that occurred in 2006 and 2007.

```
library(knitr)
#include_graphics("pilotshortage.png")
#include_graphics("pilotgraph.png")
```

Taken from: https://www.csmonitor.com/2007/0802/p01s08-ussc.html (Article) https://www.statista.com/statistics/537863/number-of-pilots-in-the-united-states/ (Graph)

Although this pilot shortage affected all airports regardless of size, smaller airports are more likely to be impacted by events such as these, both in positive and negative ways. Their lack of resources makes them

more prone to pilot shortages, but their small size allows them to avoid issues such as overcrowding and mass technical issues. This also explains the greater fluctuation in flight delays in smaller airports compared to medium or large airports.

3.5 Case study: 2008 Recession and Covid-19

3.5.1: 2008 Recession

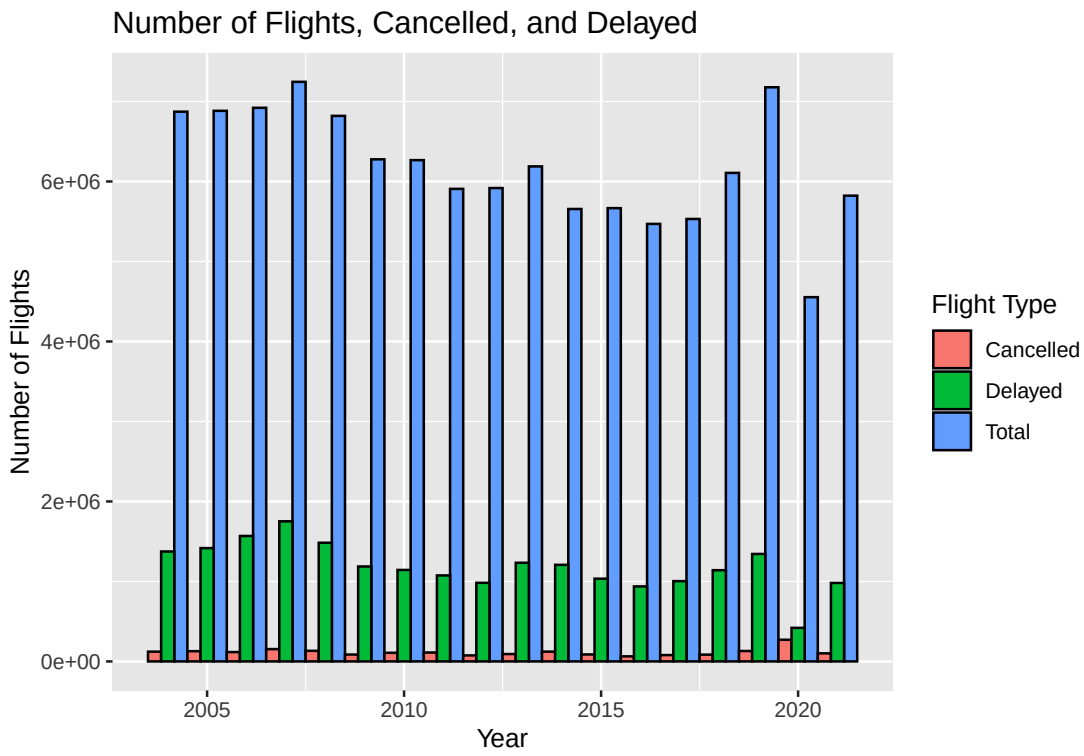
Our first case study involves the 2008 Recession. The 2008 Recession began in December of 2007 and lasted about halfway through 2009. We expected airline delays and/or cancellations to increase during this time due to cost issues within airlines and people's reluctance to fly on tight budgets.

Part 1: Grand Overlook First, we looked at a comparison of total arrival flights, arrival delays/proportion of flights delayed, and arrival cancellations/proportion of flights cancelled across the entire data set (2003-2021). We excluded 2003 and 2022 because the data only goes through half of those years.

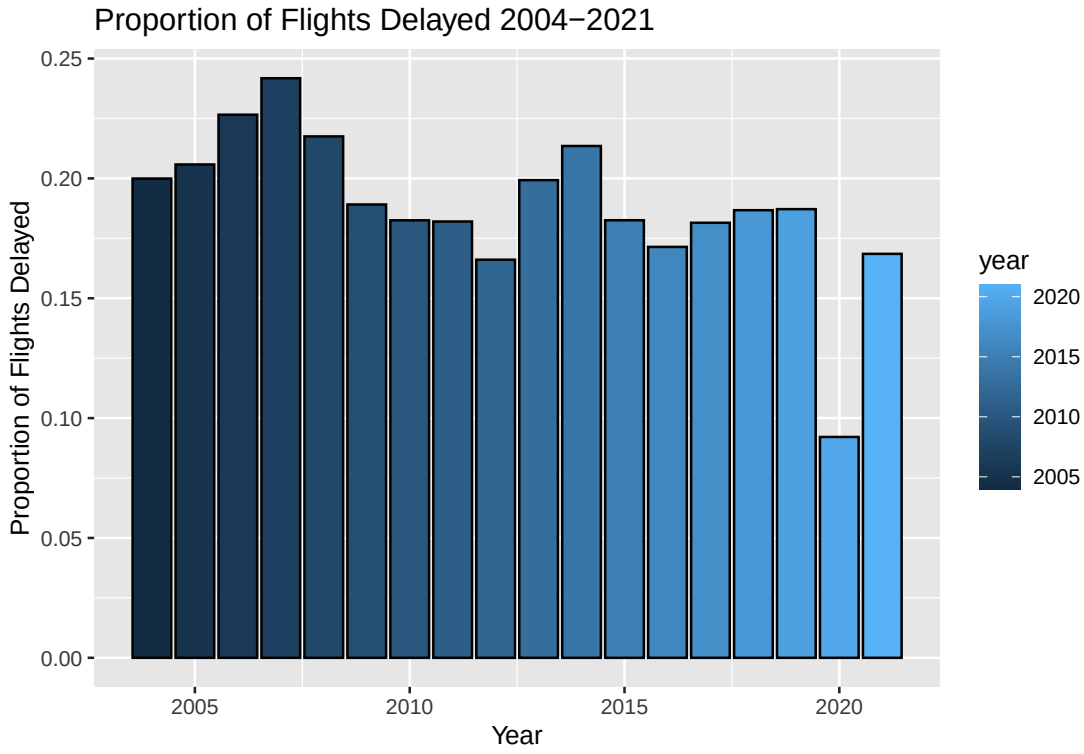
```
air_dataSum <- air_data %>%
  filter(year %in% 2004:2021) %>%
  group_by(year) %>%
  summarise(
    across(c(arr_flights:late_aircraft_delay), sum, na.rm = TRUE),
    proportion_delayed = arr_del15/arr_flights,
    proportion_cancelled = arr_cancelled/arr_flights
  )
glimpse(air_dataSum)
```

```
## Rows: 18
## Columns: 18
## $ year                <int> 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2~
## $ arr_flights         <dbl> 6872218, 6882922, 6920671, 7245207, 6819642, 6276~
## $ arr_del15           <dbl> 1374054, 1416577, 1567955, 1751743, 1483345, 1186~
## $ carrier_ct          <dbl> 352965.4, 400814.3, 443835.8, 505208.8, 398141.0,~
## $ weather_ct          <dbl> 69650.42, 65044.53, 64781.25, 70272.04, 55160.85,~
## $ nas_ct              <dbl> 542258.0, 519666.1, 540309.8, 580498.3, 536219.6,~
## $ security_ct         <dbl> 4526.93, 3660.63, 5748.68, 4807.29, 3166.62, 2269~
## $ late_aircraft_ct    <dbl> 404655.2, 427391.7, 513279.8, 590957.4, 490658.2,~
## $ arr_cancelled       <dbl> 122063, 128047, 116824, 154232, 132356, 86047, 10~
## $ arr_diverted        <dbl> 13299, 13582, 15610, 16752, 16917, 15023, 15053, ~
## $ arr_delay           <dbl> 70499296, 73868337, 84506041, 97969447, 84228590,~
## $ carrier_delay       <dbl> 18266925, 20773722, 23451737, 27921762, 23226638,~
## $ weather_delay       <dbl> 4902172, 4593449, 4715611, 5577929, 4519613, 3185~
## $ nas_delay           <dbl> 23677625, 23242238, 24969373, 27443190, 25571855,~
## $ security_delay      <dbl> 173826, 135801, 215443, 172291, 111023, 76989, 10~
## $ late_aircraft_delay <dbl> 23478748, 25123127, 31153877, 36854275, 30799461,~
## $ proportion_delayed  <dbl> 0.19994331, 0.20581041, 0.22656112, 0.24177957, 0~
## $ proportion_cancelled <dbl> 0.01776181, 0.01860358, 0.01688044, 0.02128745, 0~
```

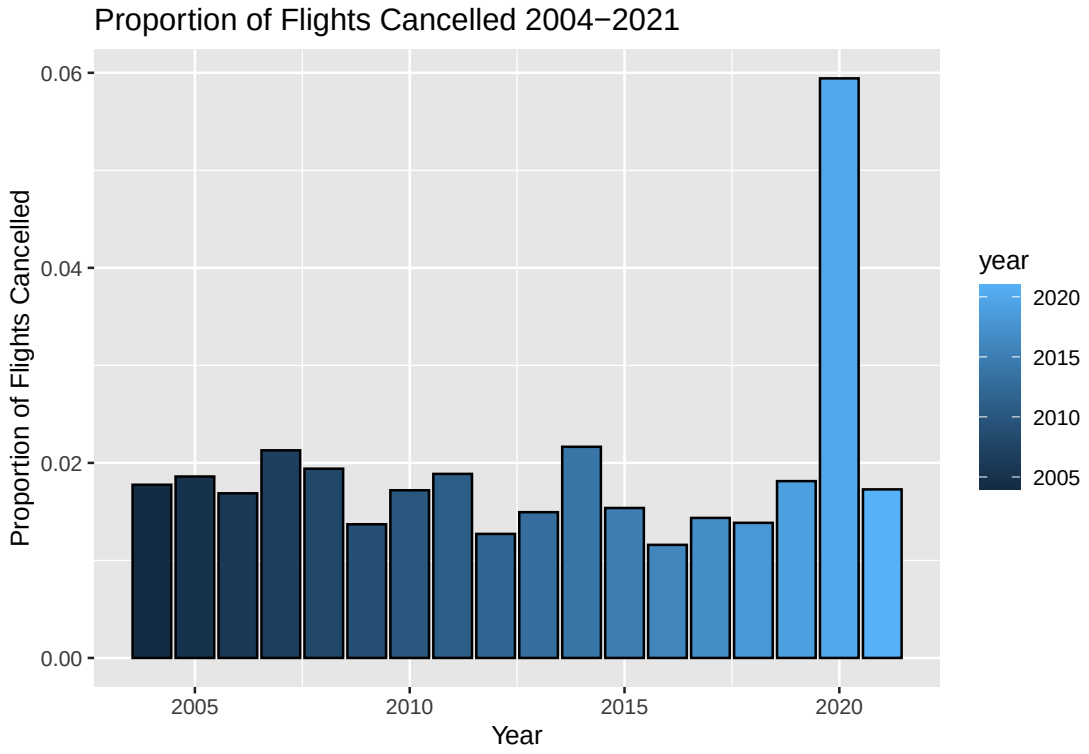
```
AirDataPivoted <- pivot_longer(air_dataSum, cols = c("arr_flights","arr_cancelled", "arr_del15"), names_

ggplot(AirDataPivoted, aes(x = year, y = Number)) + geom_col(aes(fill = comparison), position = "dodge"
```




```
ggplot(air_dataSum, aes(x = year, y = proportion_delayed, fill = year)) + geom_col(color= "black") + xla
```



```
ggplot(air_dataSum, aes(x = year, y = proportion_cancelled, fill = year)) + geom_col(color= "black") + 
```



In making these bar graphs, we can see a few things:

- First, overall arrival flights per year tend to trend higher before 2010 than after.
- Second, both total flights delayed and proportion of flights delayed trend higher during the period of 2005-2010, peaking during the recession.
- Third, by excluding 2020 (which we will investigate later in the Covid-19 case study), we can see that total flights cancelled also trends higher during the period of 2005-2010, although less prominently.
- Finally, however, the proportion of cancelled flights seems to have no large difference around the time of the recession to that of other years.

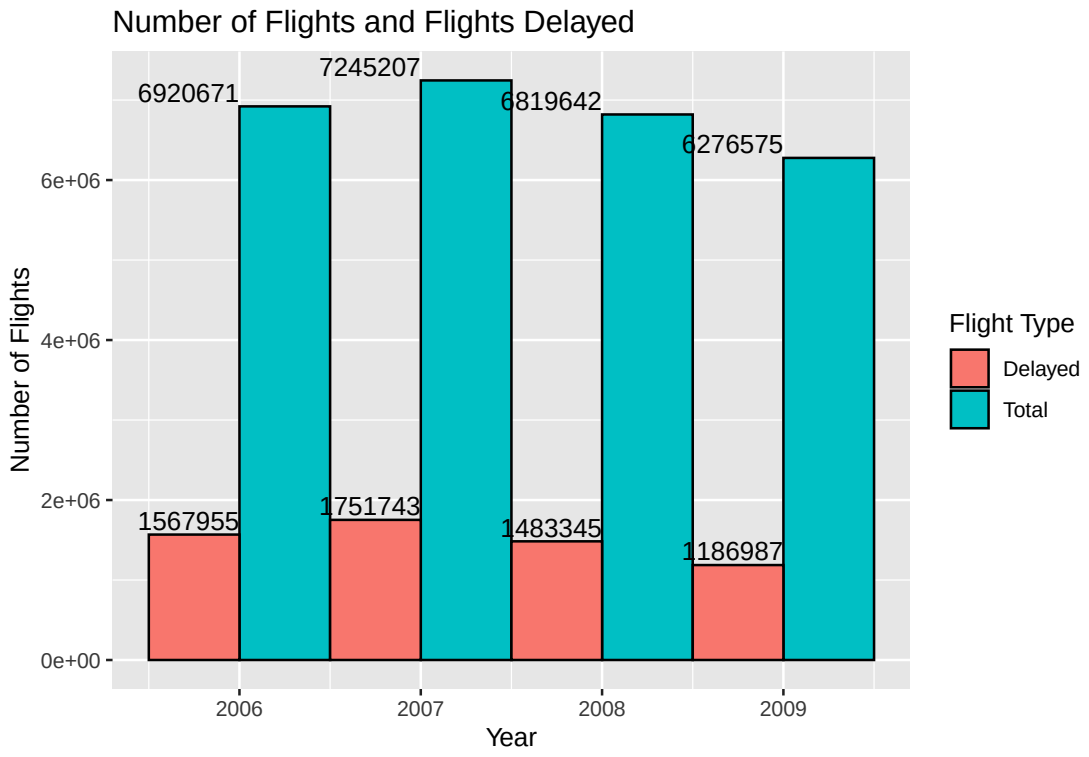
Part 2: Zooming In To investigate whether the recession had an impact, we took a closer look at flights and flight delays in the period of 2007-2009.

```
# loads proportion of delayed flights into data set
air_data <- air_data %>%
  mutate(
    proportion_delayed = arr_del15/arr_flights
  )

#zooms in to period of recession when talking about summarized totals by year
data_recession <- air_data %>%
  filter(year %in% 2006:2009) %>%
  group_by(year) %>%
  summarize(
    across(c(arr_flights:late_aircraft_delay), sum, na.rm = TRUE),
    proportion_delayed = arr_del15/arr_flights,
```

```
    proportion_cancelled = arr_cancelled/arr_flights
  )
  head(data_recession)
```

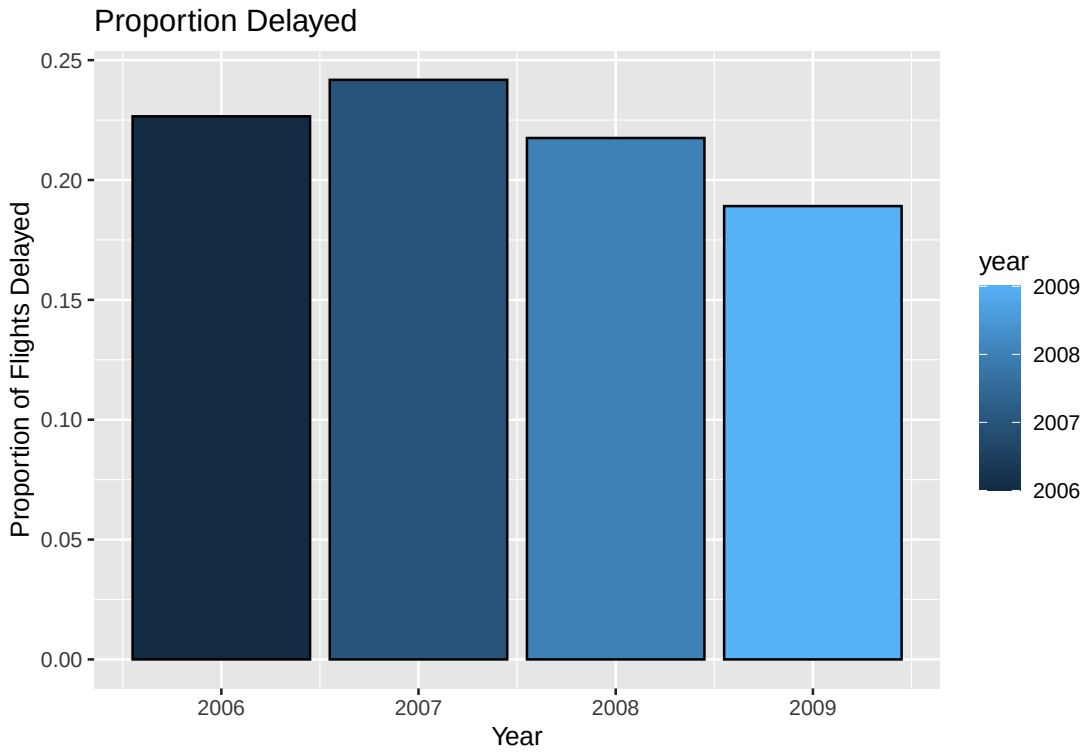
```
## # A tibble: 4 x 18
##    year arr_fli~1 arr_d~2 carri~3 weath~4 nas_ct secur~5 late_~6 arr_c~7 arr_d~8
##   <int>     <dbl>   <dbl>   <dbl>   <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1  2006   6920671 1567955 443836.  64781. 5.40e5   5749. 513280.  116824   15610
## 2  2007   7245207 1751743 505209.  70272. 5.80e5   4807. 590957.  154232   16752
## 3  2008   6819642 1483345 398141   55161. 5.36e5   3167. 490658.  132356   16917
## 4  2009   6276575 1186987 313005.  40526. 4.40e5   2269. 391307.   86047   15023
## # ... with 8 more variables: arr_delay <dbl>, carrier_delay <dbl>,
## #   weather_delay <dbl>, nas_delay <dbl>, security_delay <dbl>,
## #   late_aircraft_delay <dbl>, proportion_delayed <dbl>,
## #   proportion_cancelled <dbl>, and abbreviated variable names 1: arr_flights,
## #   2: arr_del15, 3: carrier_ct, 4: weather_ct, 5: security_ct,
## #   6: late_aircraft_ct, 7: arr_cancelled, 8: arr_diverted
```

```
  recessionDataPivoted <- pivot_longer(data_recession, cols = c("arr_flights", "arr_del15"), names_to =

ggplot(recessionDataPivoted, aes(x = year, y = Number)) + geom_col(aes(fill = comparison), position = "
```



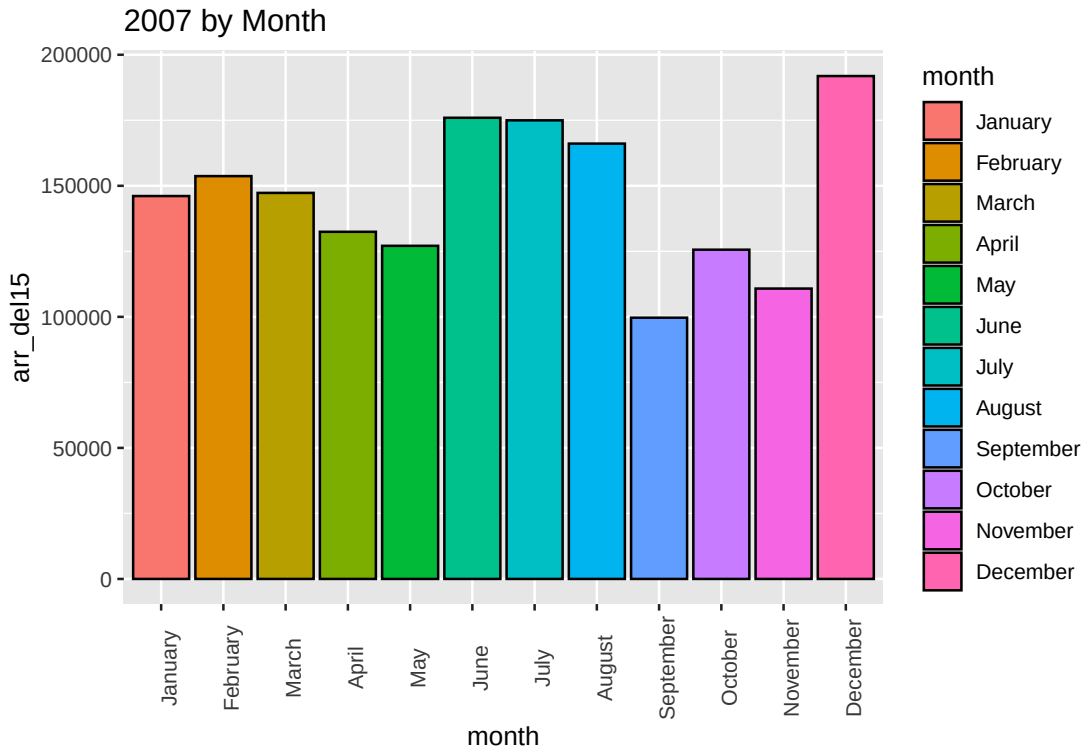
The total flights seem pretty consistent in the context of the recession times, despite later declines in the 2010s. Flight delays also looked like it followed a similar pattern, prompting us to look at delay proportions.

```
  ggplot(data_recession, aes(x = year, y = proportion_delayed, fill = year)) + geom_col(color= "black")
```



Looking at these graphs we can see that the delay proportion peaks in 2007. We then looked at delays in 2007 monthly to see if delays happened at the end of the year (when the recession hit = possible correlation) or in the beginning of the year (uncorrelated to recession).

```
data_2007_monthly <- air_data %>%
  filter(year == 2007) %>%
  group_by(month, year) %>%
  summarise(
    across(c(arr_flights:late_aircraft_delay), sum, na.rm = TRUE),
    proportion_delayed = arr_del15/arr_flights
   ) %>%
  mutate(month = factor(month, levels = month.name)) %>%
  arrange(month)

ggplot(data_2007_monthly, aes(x = month, y = arr_del15, fill = month)) + geom_col(color= "black") + ther
```




```
head(data_2007_monthly)
```

```
## # A tibble: 6 x 18
## # Groups:   month [6]
##   month     year arr_fl~1 arr_d~2 carri~3 weath~4 nas_ct secur~5 late_~6 arr_c~7
##   <fct>    <int>    <dbl>   <dbl>   <dbl>   <dbl>  <dbl>   <dbl>   <dbl>   <dbl>
## 1 January   2007   602479  146103  40776.   6893. 50524.    342.  47569.   15361
## 2 February  2007   548344  153708  44700.   7555. 48635.    404.  52414.   24017
## 3 March     2007   619857  147306  45066.   4950. 46992.    388.  49911.   16175
## 4 April     2007   596157  132461  37830.   4179. 45997.    375.  44080.   10693
## 5 May       2007   613203  127106  35360.   4692. 45512.    339.  41202.    6534
## 6 June      2007   611769  175962  49819.   8686. 55307.    570.  61580.   16432
## # ... with 8 more variables: arr_diverted <dbl>, arr_delay <dbl>,
## #   carrier_delay <dbl>, weather_delay <dbl>, nas_delay <dbl>,
## #   security_delay <dbl>, late_aircraft_delay <dbl>, proportion_delayed <dbl>,
## #   and abbreviated variable names 1: arr_flights, 2: arr_del15, 3: carrier_ct,
## #   4: weather_ct, 5: security_ct, 6: late_aircraft_ct, 7: arr_cancelled
```

Based on this graph, we can see that the most delays happened in December. This is when the recession hit. There is an increase in delays from June-August, suggesting a spike in travel in summer, but the increase is even greater in December. However, we had to look closer at December to see if this was actually correlated to the recession or another factor (e.g. winter weather).

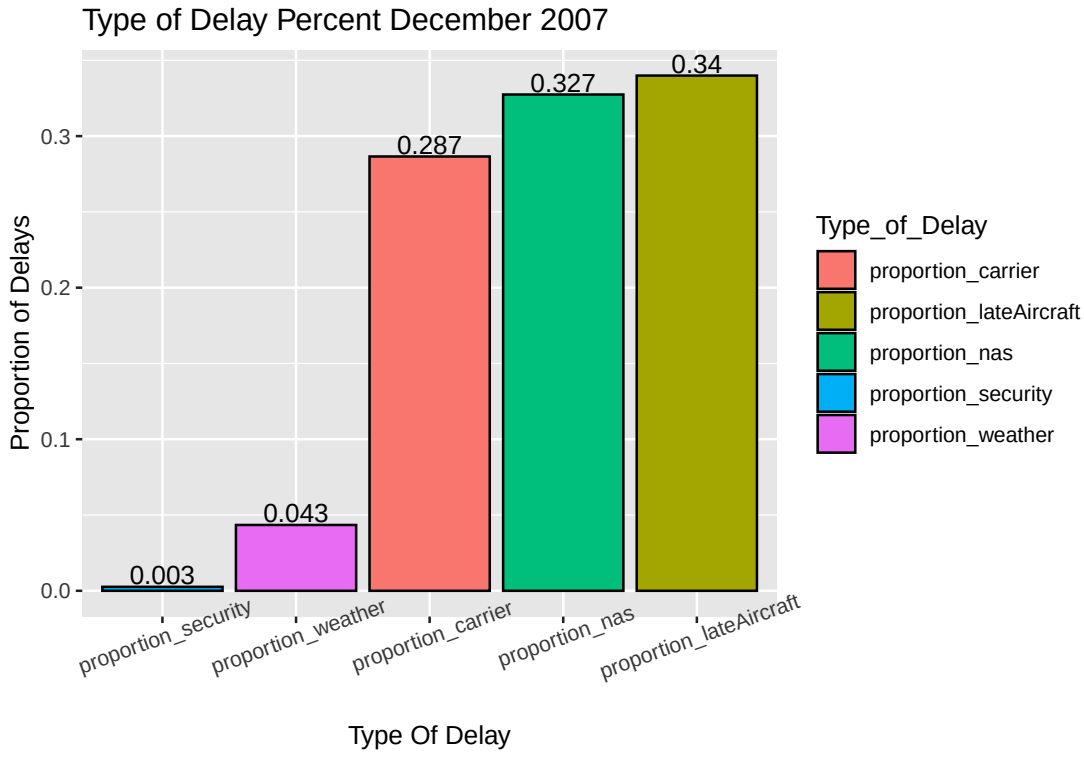
```
December2007 <- air_data %>%
  filter(year == 2007, month == "December") %>%
```

```
  summarise(
    month = "December",
    year = 2007,
    across(c(arr_flights:late_aircraft_delay), sum, na.rm = TRUE),
    proportion_delayed = arr_del15/arr_flights,
    proportion_carrier = carrier_ct/arr_del15,
    proportion_weather = weather_ct/arr_del15,
    proportion_nas = nas_ct/arr_del15,
    proportion_security = security_ct/arr_del15,
    proportion_lateAircraft = late_aircraft_ct/arr_del15,
    total = sum(proportion_carrier, proportion_weather, proportion_nas, proportion_security, proportion
  )
december2007pivoted <- pivot_longer(December2007, cols = "proportion_carrier":"proportion_lateAircraft"
  select(month, year, Type_of_Delay, Percent) %>%
  arrange(Percent)
  print(december2007pivoted)
```

Part 3: Types of Delay

```
## # A tibble: 5 x 4
##   month     year Type_of_Delay           Percent
##   <chr>    <dbl> <chr>                     <dbl>
## 1 December  2007 proportion_security     0.00263
## 2 December  2007 proportion_weather      0.0434
## 3 December  2007 proportion_carrier      0.287
## 4 December  2007 proportion_nas          0.327
## 5 December  2007 proportion_lateAircraft 0.340
```

```
  ggplot(december2007pivoted, aes(x = reorder(Type_of_Delay, +Percent), y = Percent, fill = Type_of_Dela
```

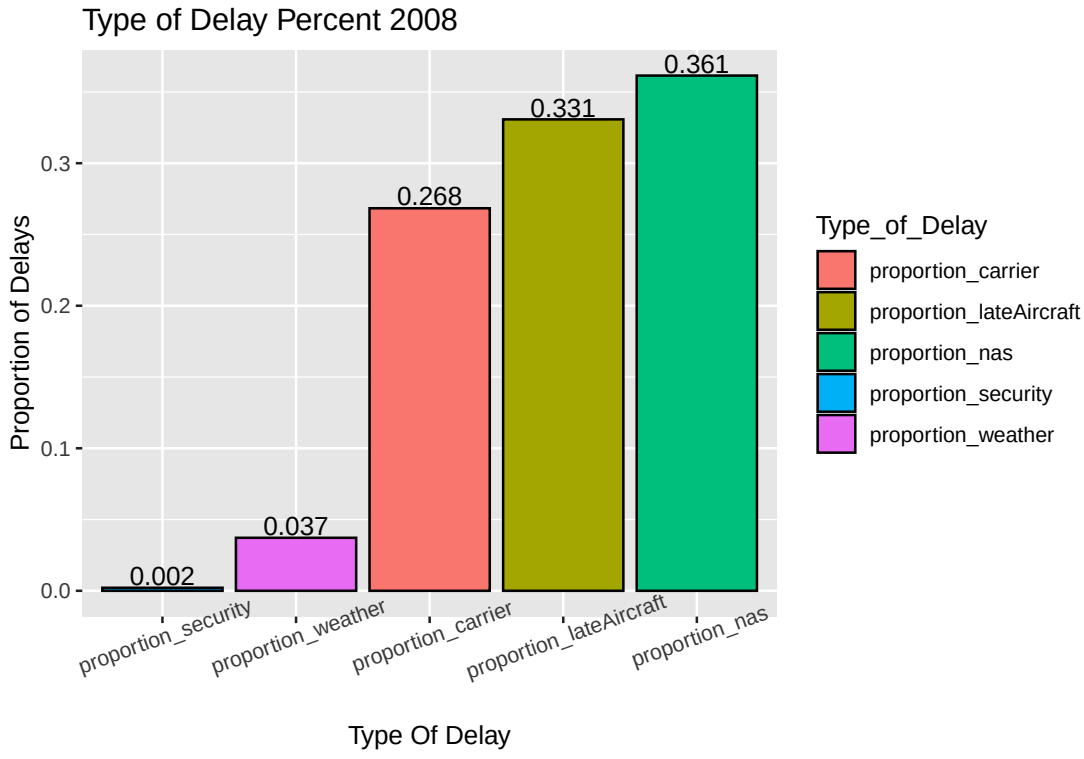


```
# air_data %>%
#   filter(year == 2007, month == "December") %>%
#   mutate(lateAircraftProp = late_aircraft_ct/arr_del15) %>%
#   arrange(desc(lateAircraftProp)) %>%
#   head()
```

Based on this graph, we can see that the largest contributor to delays in December 2007 was late aircraft delays, followed by NAS delays and carrier delays. We wanted to compare this to the rest of the recession.

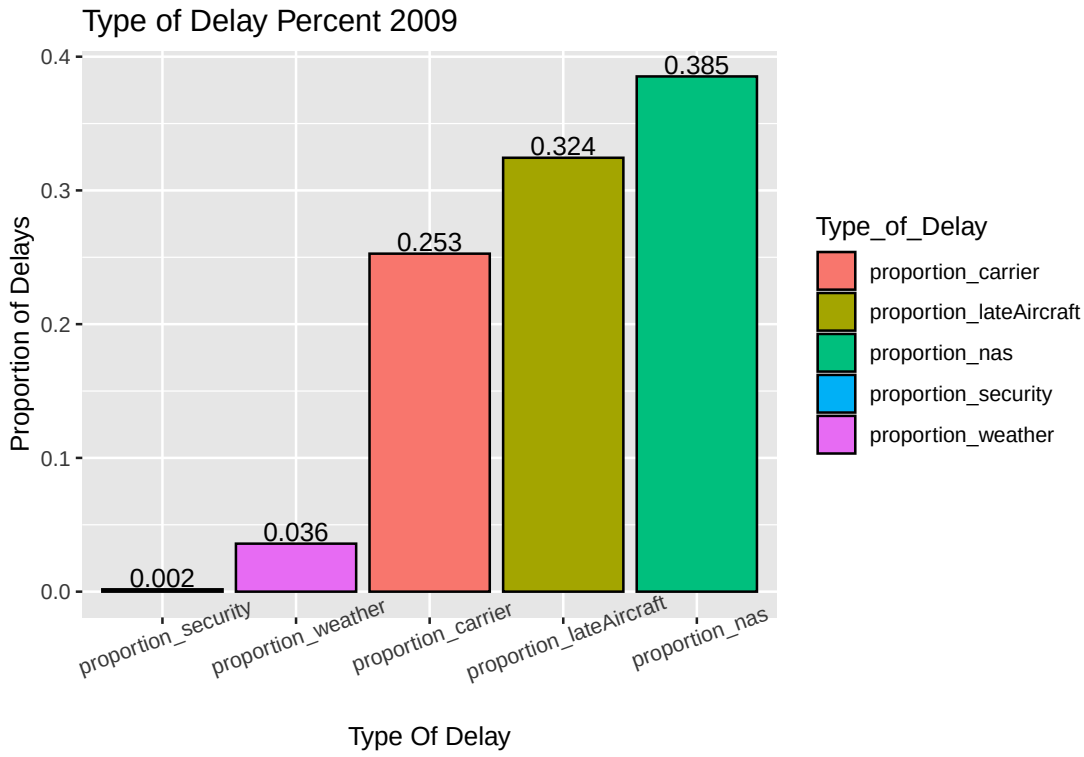
First we looked at the types of delay percentages for all of 2008:

```
Delays2008 <- air_data %>%
  filter(year == 2008) %>%
  summarise(
    year = 2008,
    across(c(arr_flights:late_aircraft_delay), sum, na.rm = TRUE),
    proportion_delayed = arr_del15/arr_flights,
    proportion_carrier = carrier_ct/arr_del15,
    proportion_weather = weather_ct/arr_del15,
    proportion_nas = nas_ct/arr_del15,
    proportion_security = security_ct/arr_del15,
    proportion_lateAircraft = late_aircraft_ct/arr_del15,
    total = sum(proportion_carrier, proportion_weather, proportion_nas, proportion_security, proportion
  )
delays2008pivoted <- pivot_longer(Delays2008, cols = "proportion_carrier":"proportion_lateAircraft", na
  select( year, Type_of_Delay, Percent) %>%
  arrange(Percent)

  ggplot(delays2008pivoted, aes(x = reorder(Type_of_Delay, +Percent), y = Percent, fill = Type_of_Delay
```



Then we looked at types of delay percentages for the first half of 2009, the months when the recession was still happening:

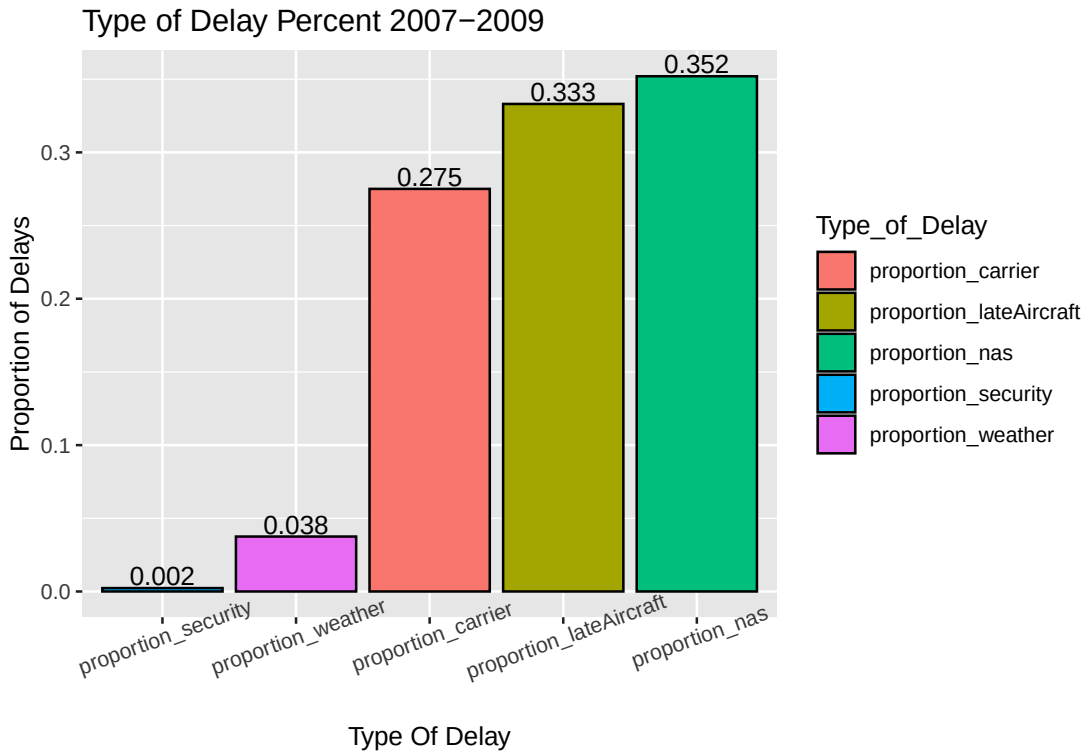
```
Delays2009 <- air_data %>%
  filter(year == 2009, month %in% c("January", "February", "March", "April", "May", "June")) %>%
  summarise(
    year = 2009,
    across(c(arr_flights:late_aircraft_delay), sum, na.rm = TRUE),
    proportion_delayed = arr_del15/arr_flights,
    proportion_carrier = carrier_ct/arr_del15,
    proportion_weather = weather_ct/arr_del15,
    proportion_nas = nas_ct/arr_del15,
    proportion_security = security_ct/arr_del15,
    proportion_lateAircraft = late_aircraft_ct/arr_del15,
    total = sum(proportion_carrier, proportion_weather, proportion_nas, proportion_security, proportion
  )
delays2009pivoted <- pivot_longer(Delays2009, cols = "proportion_carrier":"proportion_lateAircraft", na
  select(year, Type_of_Delay, Percent) %>%
  arrange(Percent)
 # print(delays2009pivoted)

  ggplot(delays2009pivoted, aes(x = reorder(Type_of_Delay, +Percent), y = Percent, fill = Type_of_Delay
```



Finally, we looked at the overall delay percentages for 2007-2009, the entire recession:

```
RecessionDelays <- air_data %>%
  filter(year %in% 2007:2009) %>%
  summarise(
    across(c(arr_flights:late_aircraft_delay), sum, na.rm = TRUE),
```

```
    proportion_delayed = arr_del15/arr_flights,
    proportion_carrier = carrier_ct/arr_del15,
    proportion_weather = weather_ct/arr_del15,
    proportion_nas = nas_ct/arr_del15,
    proportion_security = security_ct/arr_del15,
    proportion_lateAircraft = late_aircraft_ct/arr_del15,
    total = sum(proportion_carrier, proportion_weather, proportion_nas, proportion_security, proportion
  )
recessionDelayspivoted <- pivot_longer(RecessionDelays, cols = "proportion_carrier":"proportion_lateAir
  select(Type_of_Delay, Percent) %>%
  arrange(Percent)

  ggplot(recessionDelayspivoted, aes(x = reorder(Type_of_Delay, +Percent), y = Percent, fill = Type_of_
```



Based on these graphs, we can see that the three top reasons for delays remain constant: NAS delays, late aircraft delays and carrier delays; however, they differ in order from December 2007. Because the overall trend suggested NAS Delays made up the most of the delays, we researched into what NAS Delays actually entail. NAS Delays are a combination of non-extreme weather conditions, airport operations, heavy traffic volume, air traffic control, etc. Here is what we found on NAS Delays:

```
library(knitr)
#include_graphics("4Pies.png")
```

Part 4: Division of NAS Delay Causes Taken from:

- 2007: https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?20=E

- 2008: https://www.transtats.bts.gov/OT_Delay/ot_delaycause1.asp?qv52ynB=pun46&20=E
- 2009: https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?20=E
- 2007-2009: https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?20=E

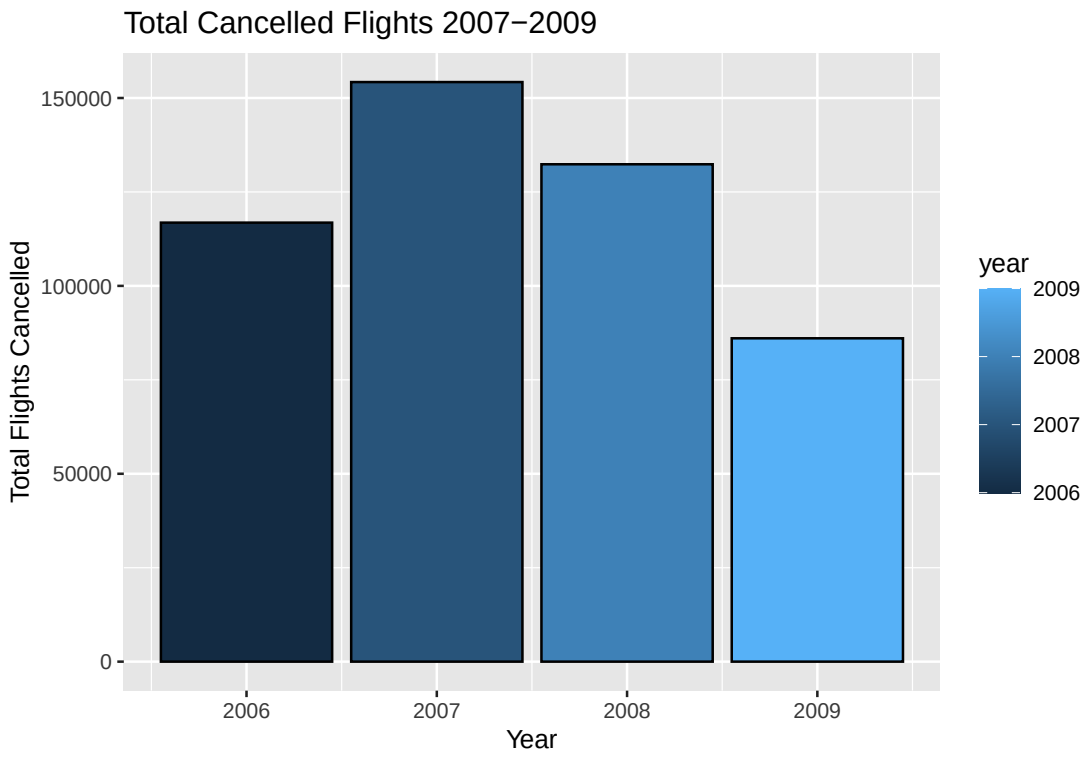
Analysis This data shows that a majority of NAS delays are caused by non-extreme weather and volume delays. Because a all the charts suggest a large majority is due to non-extreme weather, we can assume that a lot of delays are due to weather. Heavy traffic, the other chunk of delay causes, is due to build up of flights leaving at once. Departing flights follow cues on taxiways before they are allowed to take off. During the winter, flights have to undergo de-icing, which can lead to delays in take off as well.*

Based on the numbers and information provided, we concluded that the Recession itself was not a major delay component during these years. Despite our predictions, we found that a majority of delays were actually caused by weather, both extreme and non extreme, as well as heavy traffic volume delays (unexpected as we assumed air traffic would be down if people's finances weren't flourishing). Overall, the financial crisis in 2008 did not have as large of an effect on delays as we assumed it would.

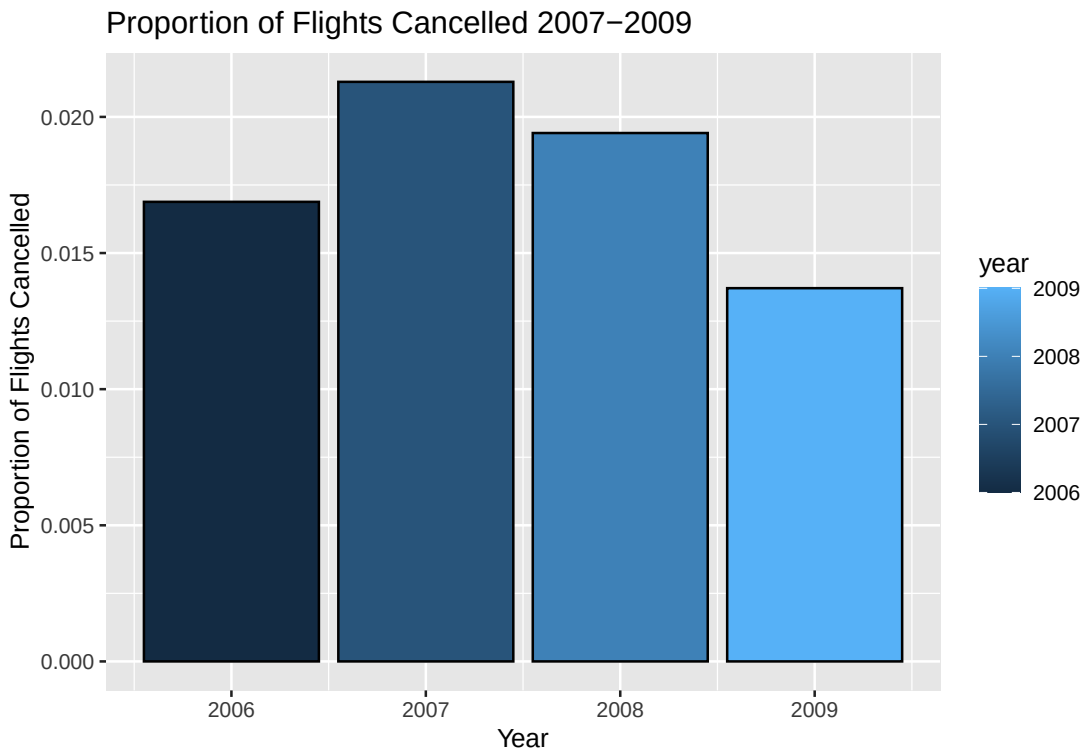
*Resource: https://simpleflying.com/en-route-delays-us-flights-causes/

Part 5: Flight Cancellations Despite these conclusions, we also wanted look at flight cancellations around the time of the financial crisis.

```
ggplot(data_recession, aes(x = year, y = arr_cancelled, fill = year)) + geom_col(color = "black") + g
```

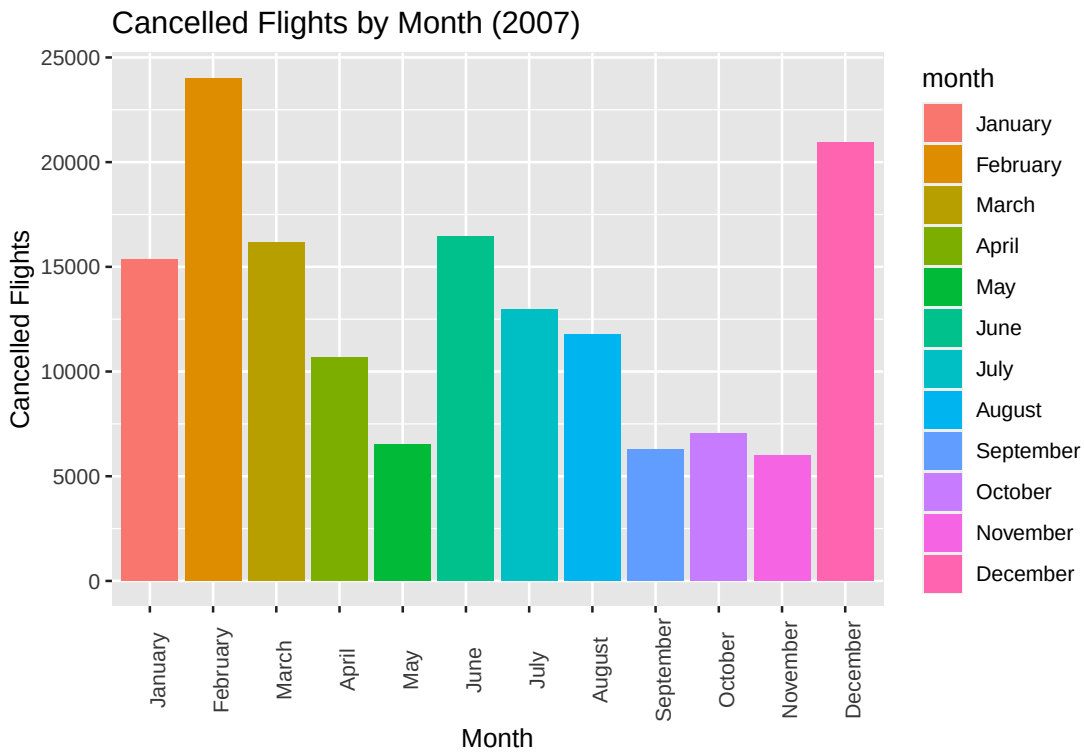



```
ggplot(data_recession, aes(x = year, y = proportion_cancelled, fill = year)) + geom_col(color = "black
```



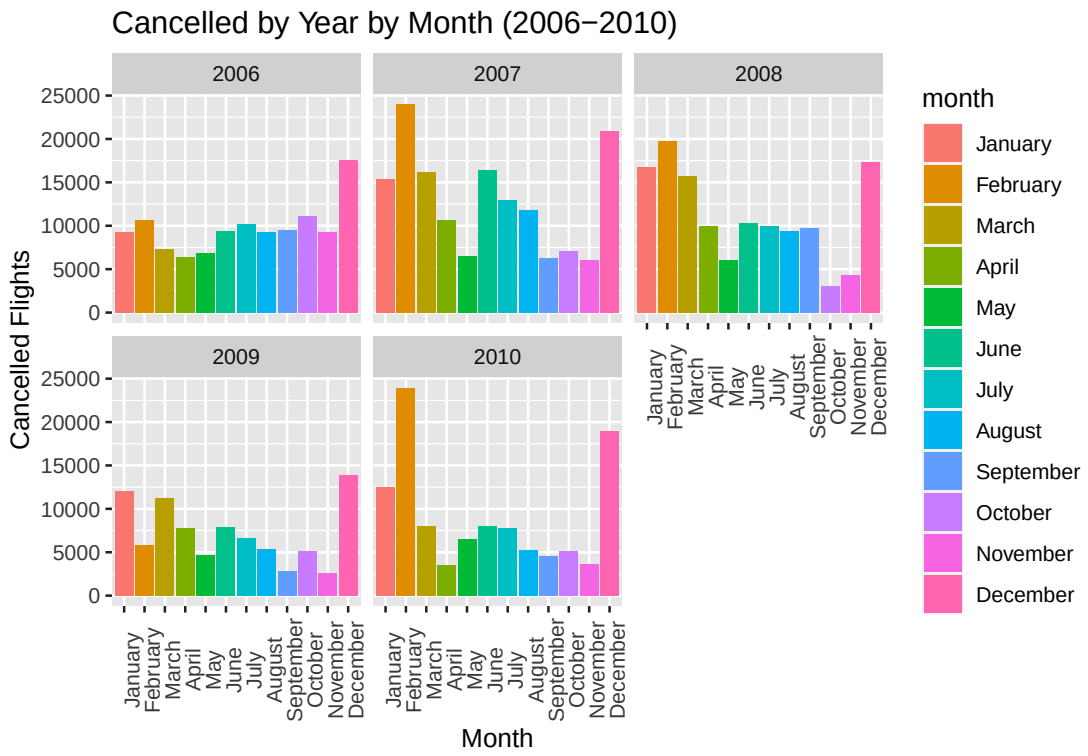
We took a look at 2007, which had the most cancellations, to see if it was possible that there was an increase in December when the recession first hit.

```
data_2007_monthly <- air_data %>%
   filter(year %in% 2007) %>%
  group_by(month) %>%
  summarise(
    across(c(arr_flights:late_aircraft_delay), sum, na.rm = TRUE),
    proportion_delayed = arr_del15/arr_flights
   ) %>%
  mutate(month = factor(month, levels = month.name)) %>%
  arrange(month)

ggplot(data_2007_monthly, aes(x = month, y = arr_cancelled, fill = month)) + geom_col() + theme(axis.te
```



Based on this graph, we can see that the largest number of cancelled flights occurred in February 2007, before the recession even began. We do see an influx again in December, but we wanted to look at this in larger context to see if this is a pattern or specific to this year. We can see that this tends to be the trend across the entire recession (2007-2009), as well as in the years surrounding this period.

```
BeforeAndAfter <- air_data %>%
  filter(year %in% 2006:2010) %>%
  group_by(month, year) %>%
  summarise(
    across(c(arr_flights:late_aircraft_delay), sum, na.rm = TRUE),
    proportion_delayed = arr_del15/arr_flights
   ) %>%
  mutate(month = factor(month, levels = month.name)) %>%
  arrange(month)

ggplot(BeforeAndAfter, aes(x = month, y = arr_cancelled, fill = month)) + geom_col() + theme(axis.text.
```



Overall, these trends suggest that the recession did not have as much of an effect on flight cancellations either.

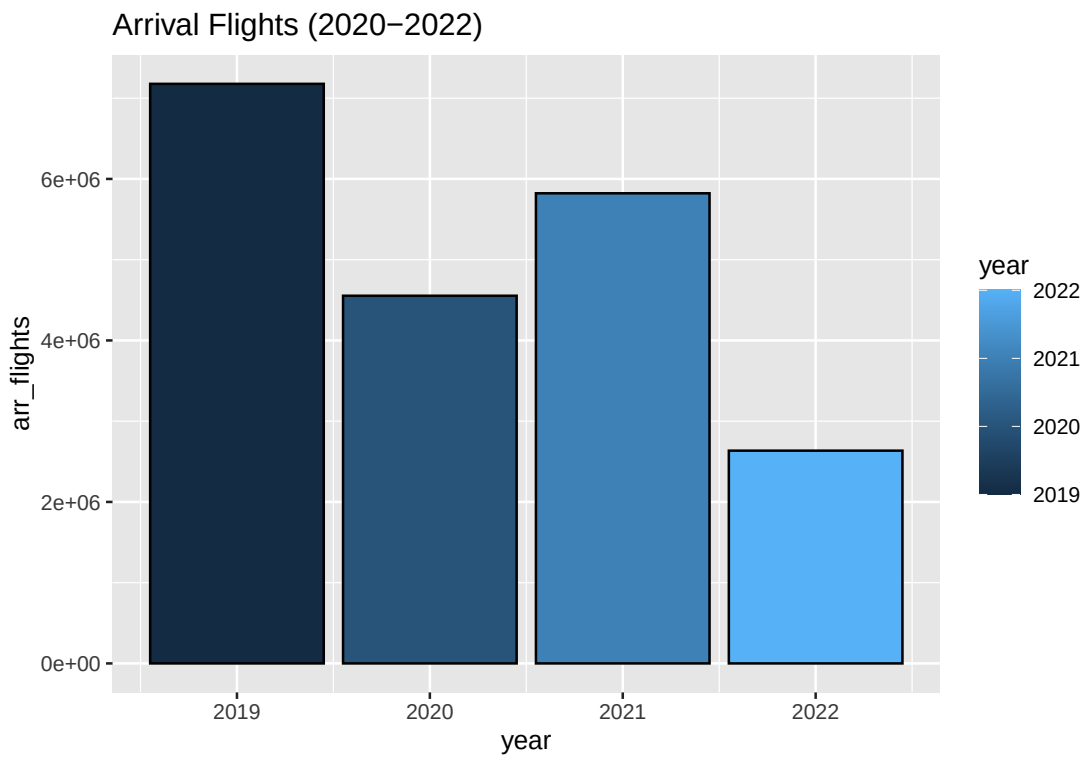
3.5.2: Covid-19 (2020-Present)

Over the last few years, Covid-19 has greatly effected everyone's lives. Based on what we know from current event news, the airline industry in 2020 was nearly decimated as people refused to travel. We wanted to look at the data and see if it matched our previous knowledge.

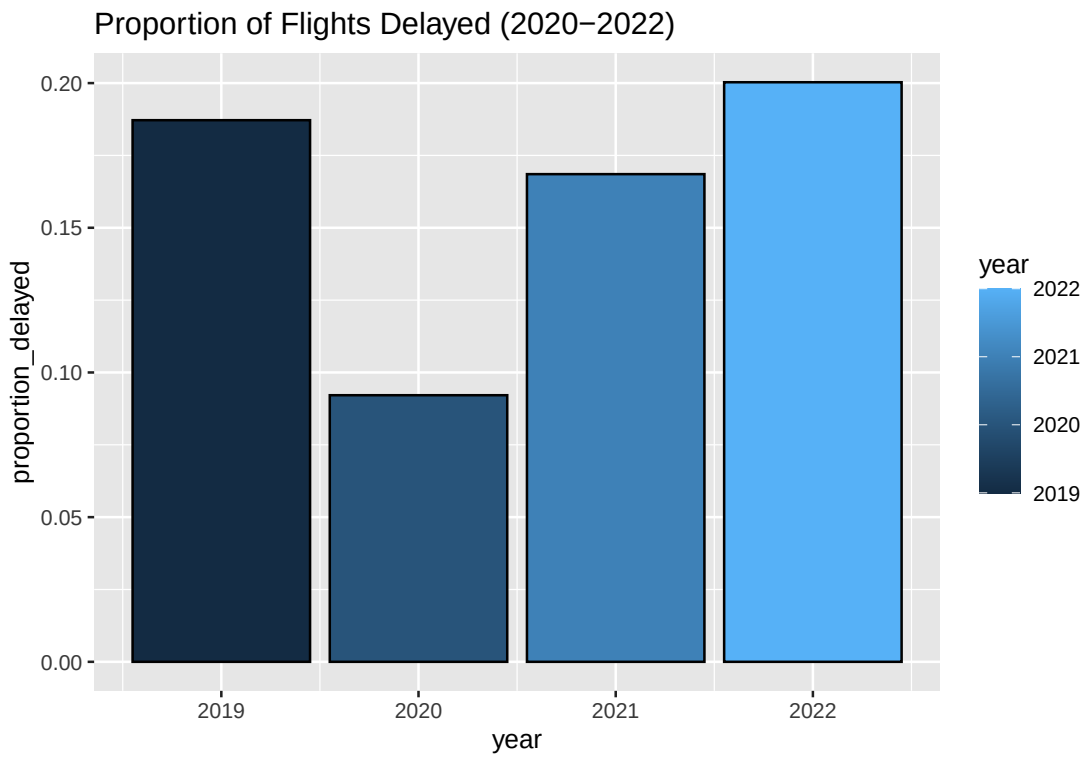
Based on the graphs in part 1 of the 2008 recession section, we can see that flight cancellations spiked while flight delays went down in 2020.

First we wanted to look at the overall number of arrival flights, as well as the proportion of flights delayed and cancelled (2019 is for "Before-Covid" reference).

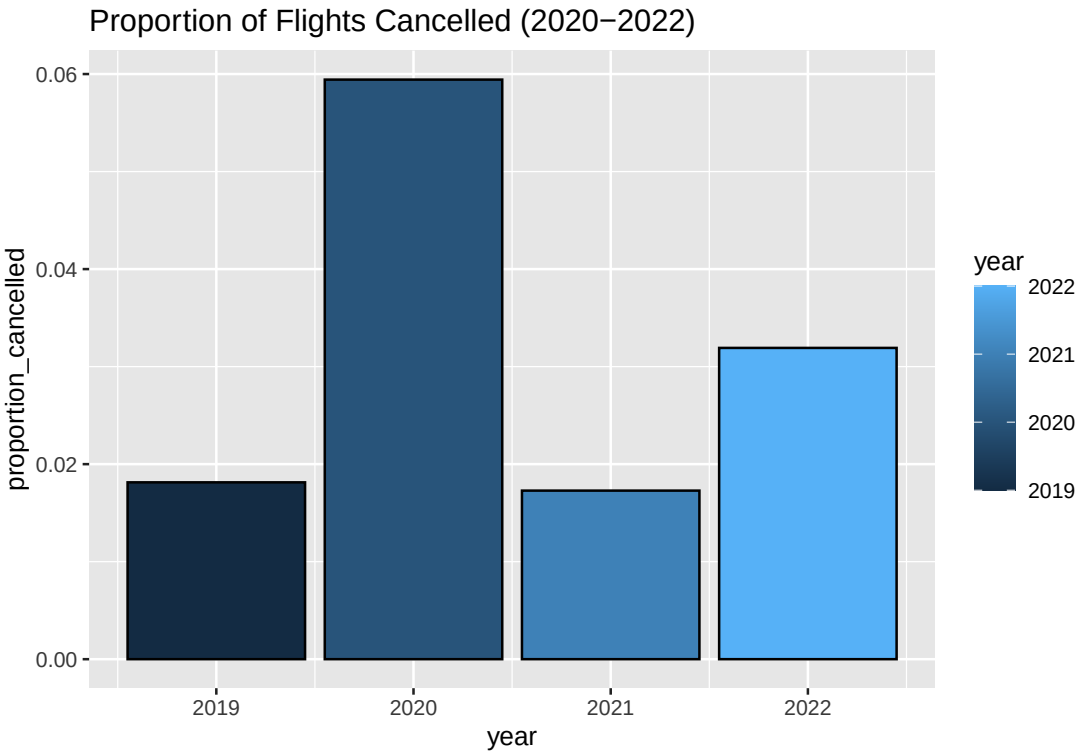
```
CovidYears <- air_data %>%
  filter(year %in% 2019:2022) %>%
  group_by(year) %>%
  summarise(
    across(c(arr_flights:late_aircraft_delay), sum, na.rm = TRUE),
    proportion_delayed = arr_del15/arr_flights,
    proportion_cancelled = arr_cancelled/arr_flights
  )

ggplot(CovidYears, aes(x = year, y = arr_flights, fill = year)) + geom_col(color = "black") + ggtitle("
```




```
ggplot(CovidYears, aes(x = year, y = proportion_delayed, fill = year)) + geom_col(color = "black") + gg
```



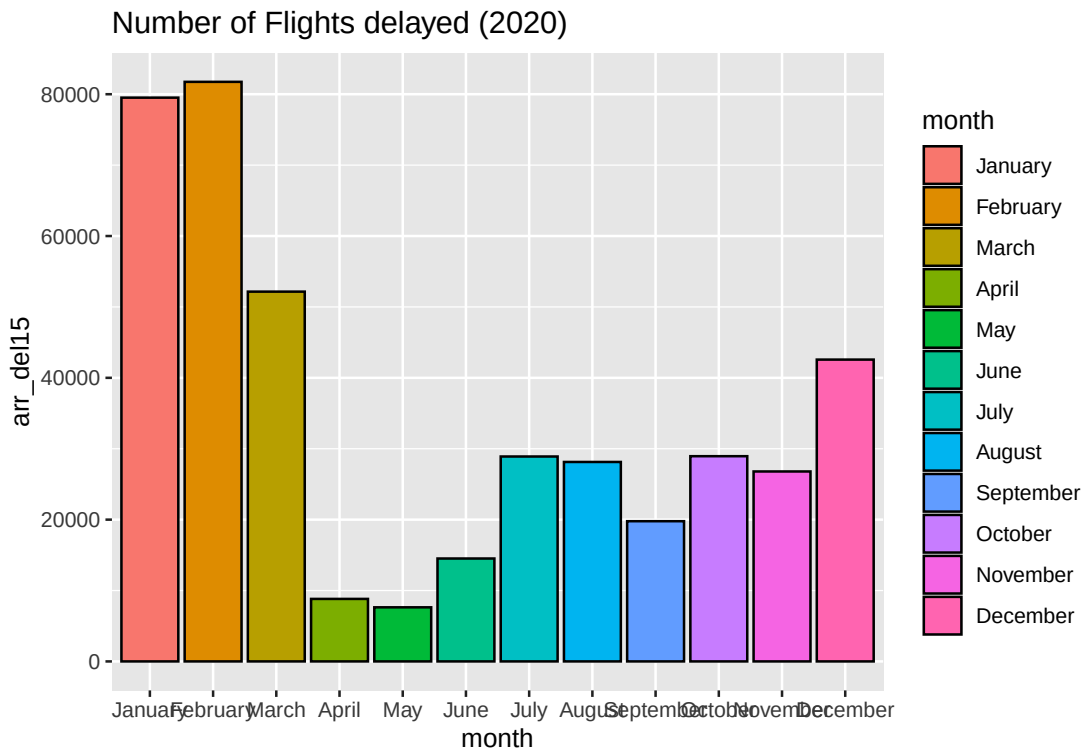
```
ggplot(CovidYears, aes(x = year, y = proportion_cancelled, fill = year)) + geom_col(color = "black") + g
```



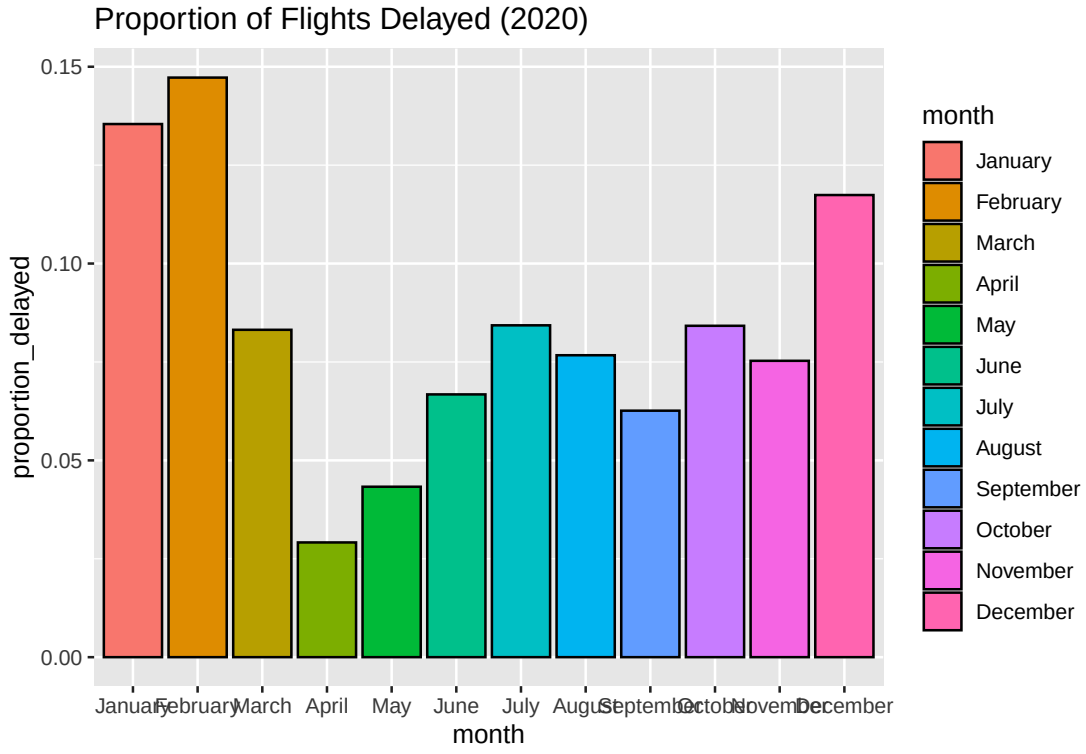
From these graphs, we can see that the Covid years have very different flight stats from 2019 (before Covid). The overall number of flights was much higher in 2019 than 2020 and 2021, suggesting something caused a large drop. We see a really large spike in the proportion of flight cancellations in 2020, which we wanted to look into to see if it was caused by Covid. Lastly, delays dropped in 2020 (perhaps because of the lack of flights) but began to rise again in 2021. *Important to note: 2022 flight stats only go through the first half of the year, accounting for the smaller number of flights.

First, we looked at delays in 2020 by month.

```
Covid2020 <- air_data %>%
  filter(year == 2020) %>%
  group_by(month) %>%
  summarise(
    across(c(arr_flights:late_aircraft_delay), sum, na.rm = TRUE),
    proportion_delayed = arr_del15/arr_flights,
    proportion_cancelled = arr_cancelled/arr_flights
   ) %>%
  mutate(month = factor(month, levels = month.name)) %>%
  arrange(month)

ggplot(Covid2020, aes(x = month, y = arr_del15, fill = month)) + geom_col(color = "black") + ggtitle("Nu
```



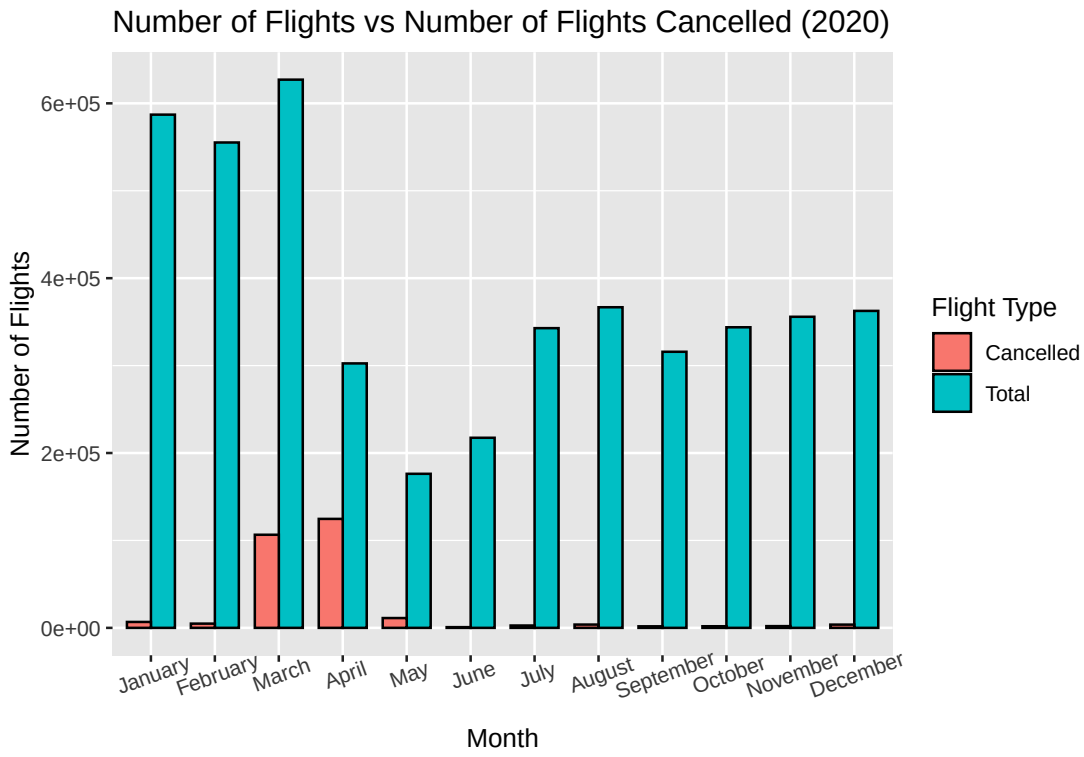

```
ggplot(Covid2020, aes(x = month, y = proportion_delayed, fill = month)) + geom_col(color = "black") + g
```



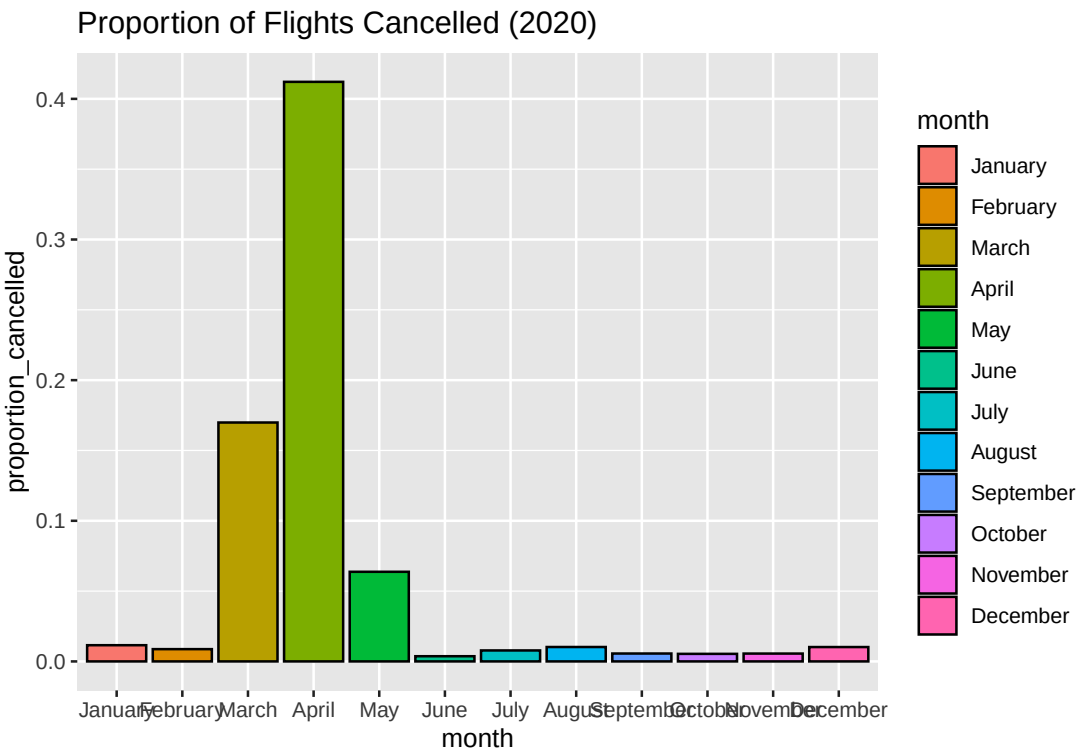
We can see a sharp decrease in the number of flight delays, as well as in the proportion of flight delays. Knowing how Covid affected flights, we thought that maybe this was because flight numbers were down, so

we decided to look at cancellations in 2020.

```
Covid2020 <- air_data %>%
  filter(year == 2020) %>%
  group_by(month) %>%
  summarise(
    across(c(arr_flights:late_aircraft_delay), sum, na.rm = TRUE),
    proportion_delayed = arr_del15/arr_flights,
    proportion_cancelled = arr_cancelled/arr_flights
   ) %>%
  mutate(month = factor(month, levels = month.name)) %>%
  arrange(month)

Covid2020Pivoted <- pivot_longer(Covid2020, cols = c("arr_flights","arr_cancelled"), names_to = "compari

ggplot(Covid2020Pivoted, aes(x = month, y = Number)) + geom_col(aes(fill = comparison), position = "dodg
```




```
ggplot(Covid2020, aes(x = month, y = proportion_cancelled, fill = month)) + geom_col(color = "black") +
```



As we can see, flight cancellations spiked in March and April of 2020. This is when the pandemic first hit and everyone was in lock down. We can see these results corroborated by news articles.

```
library(knitr)
#include_graphics("Flight_Cancelled_March.png")
```

Taken From: https://www.vox.com/the-goods/2020/4/20/21224080/coronavirus-air-travel-decline-charts

```
library(knitr)
#include_graphics("TSA_Travelers_March.png")
```

Taken From: https://www.vox.com/the-goods/2020/4/20/21224080/coronavirus-air-travel-decline-charts

We also see that according to the BTS, airlines started scheduling fewer flights for the rest of the year, starting in April.

```
library(knitr)
#include_graphics("Scheduled_Flights.png")
```

Taken from: https://www.bts.gov/data-spotlight/april-operated-flights-hit-record-lows

Clearly, Covid-19 affected flights a lot. It caused a massive number of cancellations to occur, resulting in fewer flights over the rest of the year in 2020. Although we see flight numbers start to go up in 2021, we can attribute this to people's gradual transition back into normal life. More people started to travel again in 2021, leading to an increase in flights and air traffic and an increase in delays. As we continue through 2022, numbers adjust even more to account for the number of people getting back into traveling. Overally, Covid increased flight cancellations which caused a decrease in flights and a decrease in delays in 2020. As we move on in time, the numbers return closer to how they were pre-Covid.

Section 4: Data Modeling

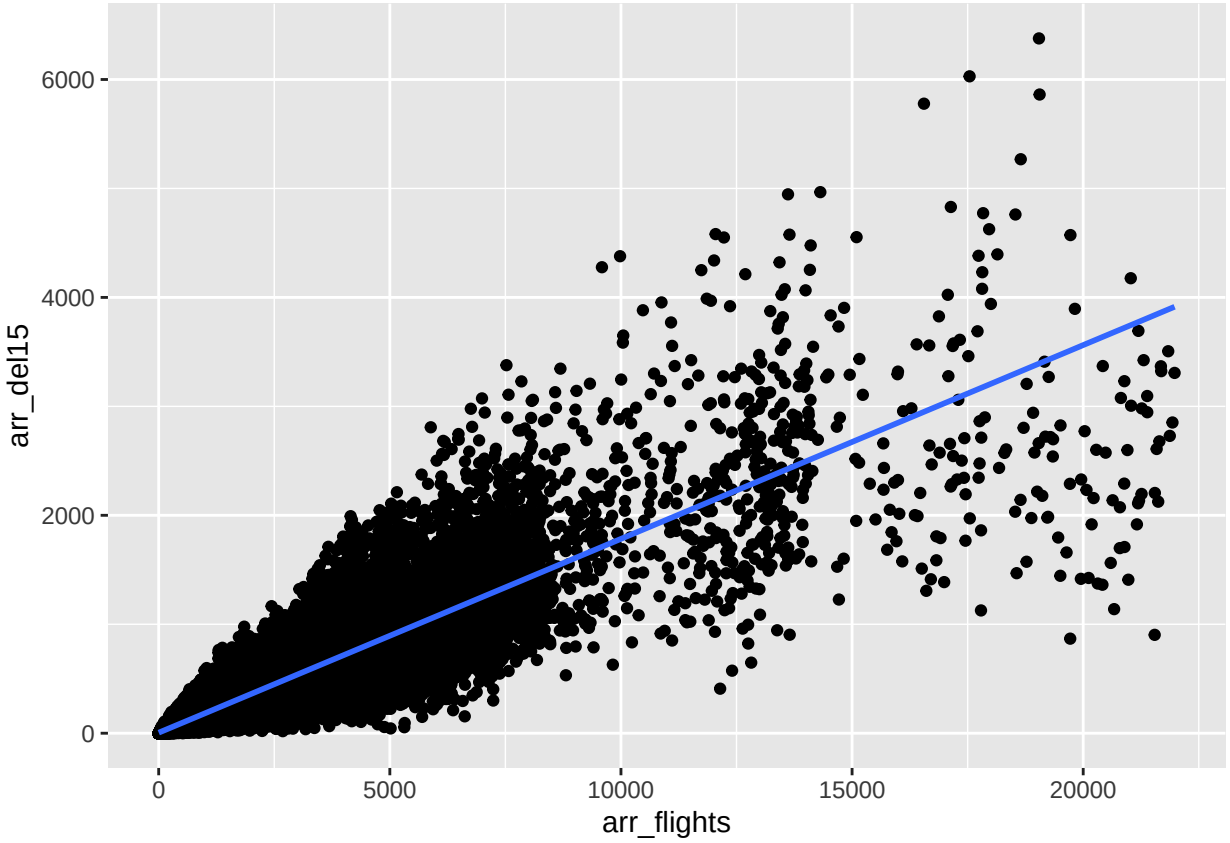
4.1: Linear Regression Analysis

For our analysis, we decided to use a linear regression model on variables explored previously.By analyzing the relationship between the explanatory and response variables, we can see how each of the variables we analyzed affected the proportion or number of flight delays. We need to check that these variables have a linear relationship.

```
air_data_lm <- air_data %>%
  mutate(propdelays = arr_del15 / arr_flights)
```

```
qplot(arr_flights, arr_del15, data = air_data, geom="point") + geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
cor(air_data_lm$arr_flights, air_data_lm$arr_del15)
```

```
## [1] 0.9186992
```

Because there seems to be a positive correlation between the number of arriving flights and the number of delayed flights, we can perform a linear regression. We will be modeling data that has been shown to have a similar correlation in our exploration of the data.

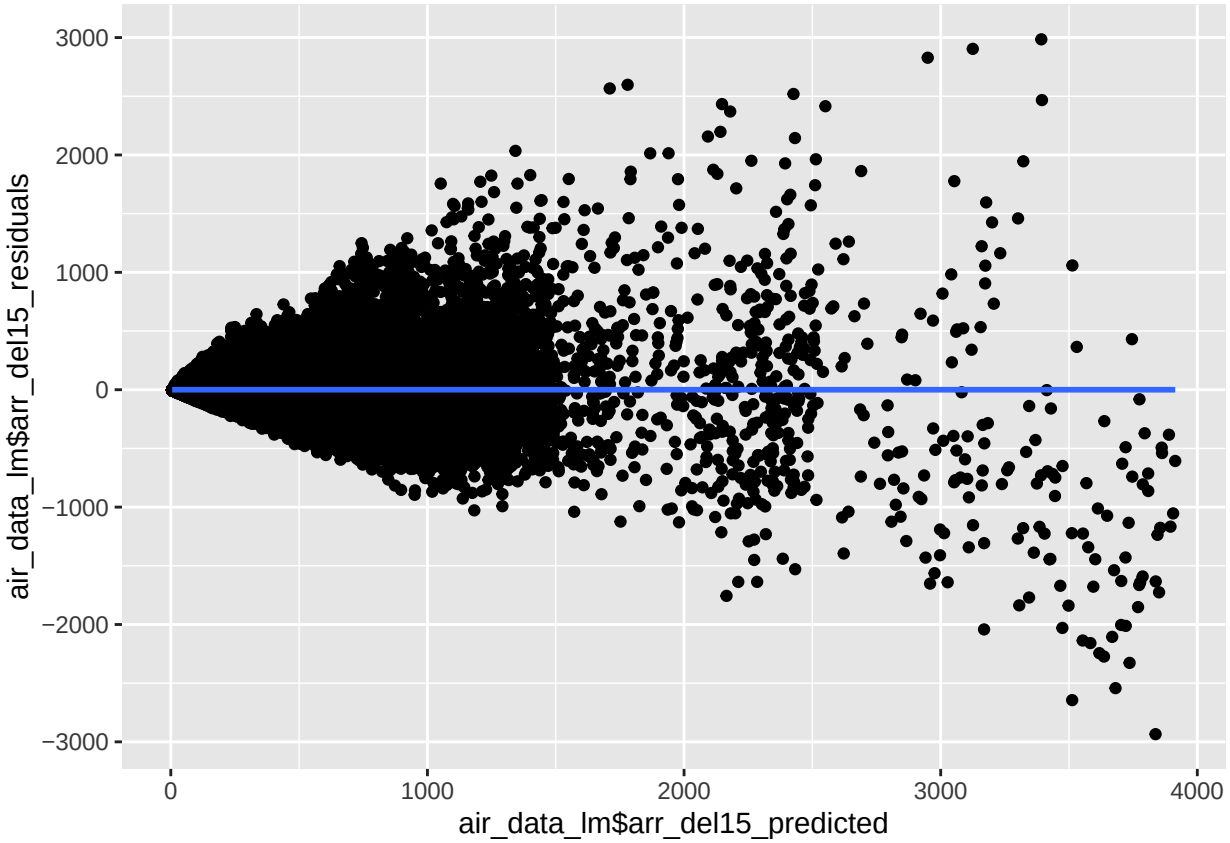
LM of Total Delays by Arriving Flights:

```
lm_arr_del15_vs_arr_flights <- lm(arr_del15 ~ arr_flights, data = air_data_lm)
summary(lm_arr_del15_vs_arr_flights)
```

```
##
## Call:
## lm(formula = arr_del15 ~ arr_flights, data = air_data_lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2934.30    -9.98    -4.44     5.16  2984.90
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.9636448  0.1525968   32.53   <2e-16 ***
## arr_flights 0.1778677  0.0001379 1289.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 79.38 on 307352 degrees of freedom
## Multiple R-squared:  0.844,  Adjusted R-squared:  0.844
## F-statistic: 1.663e+06 on 1 and 307352 DF,  p-value: < 2.2e-16
```

```
air_data_lm$arr_del15_predicted <- lm_arr_del15_vs_arr_flights$fitted.values
air_data_lm$arr_del15_residuals <- lm_arr_del15_vs_arr_flights$residuals

qplot(air_data_lm$arr_del15_predicted, air_data_lm$arr_del15_residuals) + geom_smooth(method = "lm", se
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Hear we see that the p-value is less than 0.00000000000000022, which indicates a significant result that is unlikely to be only due to chance. It confirms that the positive correlation between arriving flights and delayed flights is present. In the residuals plot, we see that there is a linear relationship due to the fact that the line of best fit is horizontal, showing a generally constant variance. Having a pattern would indicate that the relationship is not completely linear, and may be biased as a result.

LM of Weather Delay by Arriving Flights

```
lm_weather_delay_vs_arr_flights <- lm(weather_delay ~ arr_flights, data = air_data_lm)
summary(lm_weather_delay_vs_arr_flights)
```

```
##
## Call:
## lm(formula = weather_delay ~ arr_flights, data = air_data_lm)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10161    -86    -41     11  51395
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.365879   1.311442   16.29   <2e-16 ***
## arr_flights  0.522102   0.001185  440.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 682.2 on 307352 degrees of freedom
## Multiple R-squared:  0.3869, Adjusted R-squared:  0.3869
## F-statistic: 1.94e+05 on 1 and 307352 DF,  p-value: < 2.2e-16
```

```
air_data_lm$weather_delay_predicted <- lm_weather_delay_vs_arr_flights$fitted.values
air_data_lm$weather_delay_residuals <- lm_weather_delay_vs_arr_flights$residuals

cor(air_data_lm$arr_flights, air_data_lm$weather_delay)
```

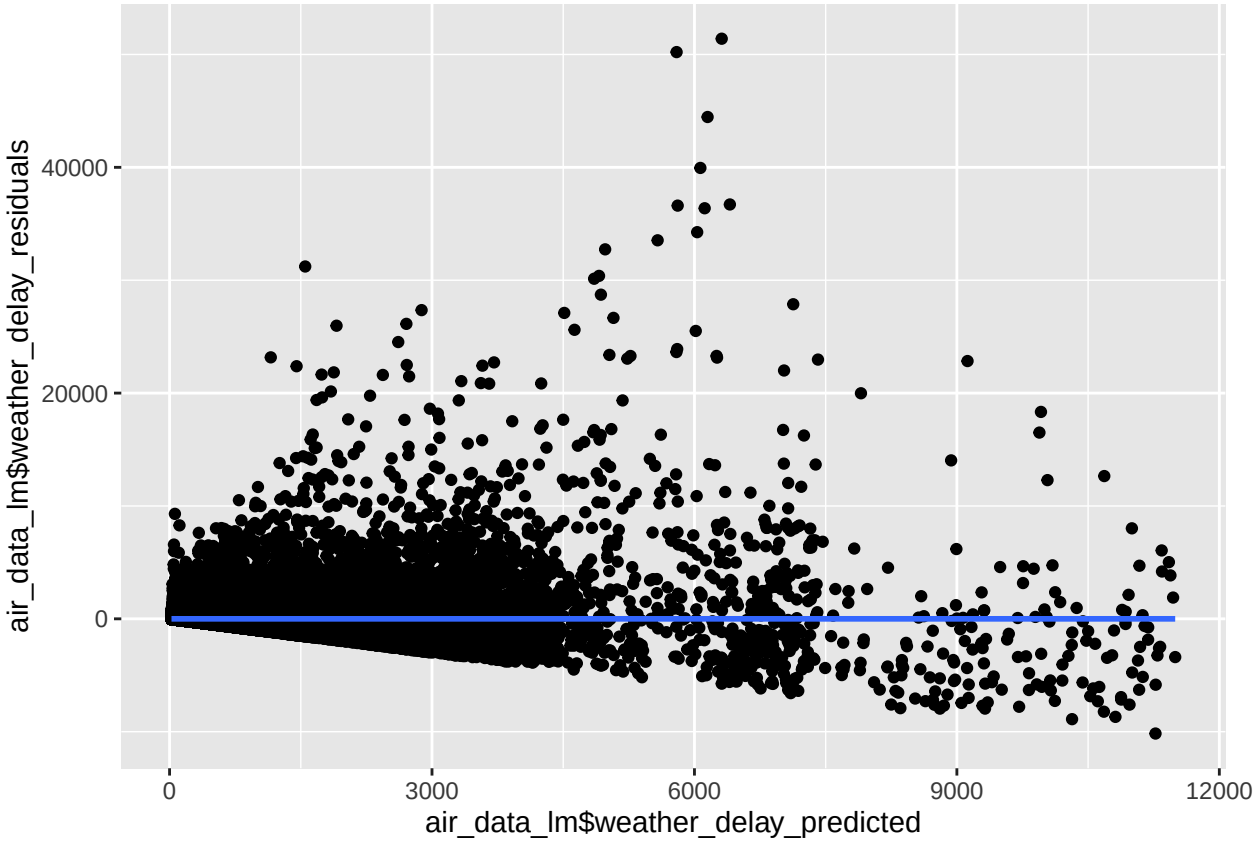
```
## [1] 0.6220512
```

```
coef(lm_weather_delay_vs_arr_flights)
```

```
## (Intercept) arr_flights
##  21.3658792   0.5221022
```

```
qplot(air_data_lm$weather_delay_predicted, air_data_lm$weather_delay_residuals) + geom_smooth(method =
```

```
## `geom_smooth()` using formula 'y ~ x'
```



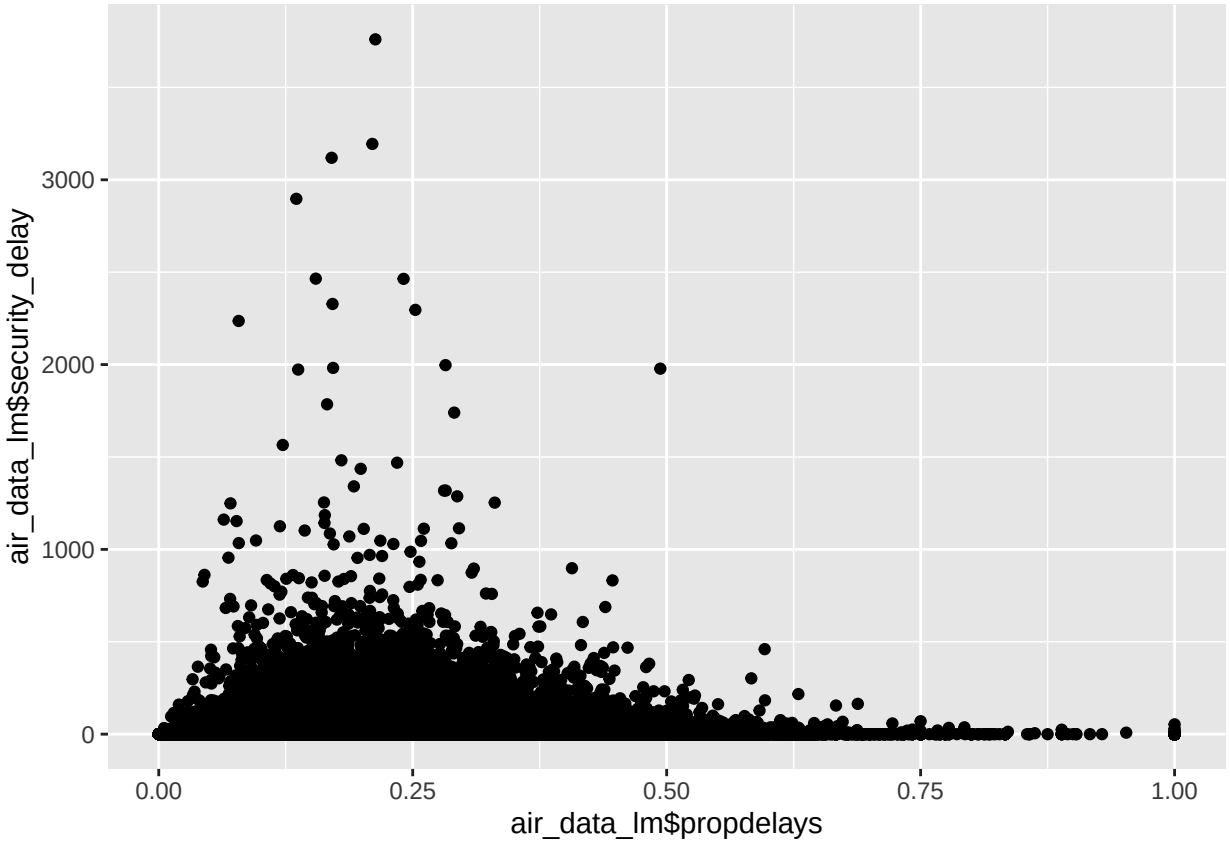
This residual plot reveals a lower linear relationship. Because the R-squared value is 0.3869, the data has a subpar linear relationship.

LM of Security Delay by Arriving Flights

```
lm_security_delay_vs_arr_flights <- lm(security_delay ~ propdelays, data = air_data_lm)
summary(lm_security_delay_vs_arr_flights)
```

```
##
## Call:
## lm(formula = security_delay ~ propdelays, data = air_data_lm)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -14.3   -7.6   -6.9   -6.2 3752.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.4551     0.1429   38.19   <2e-16 ***
## propdelays    8.8717     0.6288   14.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.08 on 307352 degrees of freedom
## Multiple R-squared:  0.0006473,  Adjusted R-squared:  0.000644
## F-statistic: 199.1 on 1 and 307352 DF,  p-value: < 2.2e-16
```

```
air_data_lm$security_delay_predicted <- lm_security_delay_vs_arr_flights$fitted.values
air_data_lm$security_delay_residuals <- lm_security_delay_vs_arr_flights$residuals

qplot(air_data_lm$propdelays, air_data_lm$security_delay)
```




```
cor(air_data_lm$propdelays, air_data_lm$security_delay)
```

```
## [1] 0.02544191
```

```
coef(lm_security_delay_vs_arr_flights)
```

```
## (Intercept)  propdelays
##    5.455143    8.871726
```

Here we see a very low correlation between the proportion of delays and security delays. The r-squared value is incredibly low at 0.000644, meaning the data has a negligible linear relationship.

LM of Proportion of Delays by Month:

```
lm_propdelays_vs_month <- lm(propdelays ~ month, data = air_data_lm)
summary (lm_propdelays_vs_month)
```

```
##
## Call:
## lm(formula = propdelays ~ month, data = air_data_lm)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.23989 -0.07031 -0.01098  0.05718  0.84812
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.2075567  0.0006725 308.638   <2e-16 ***
## monthFebruary   0.0019796  0.0009559   2.071   0.0384 *
## monthMarch     -0.0161623  0.0009543 -16.936   <2e-16 ***
## monthApril     -0.0319267  0.0009539 -33.468   <2e-16 ***
## monthMay       -0.0243437  0.0009569 -25.440   <2e-16 ***
## monthJune       0.0291067  0.0009563  30.438   <2e-16 ***
## monthJuly       0.0258021  0.0009582  26.927   <2e-16 ***
## monthAugust     0.0017681  0.0009533   1.855   0.0636 .
## monthSeptember -0.0556769  0.0009527 -58.442   <2e-16 ***
## monthOctober   -0.0398090  0.0009563 -41.629   <2e-16 ***
## monthNovember  -0.0422952  0.0009553 -44.274   <2e-16 ***
## monthDecember   0.0323339  0.0009520  33.963   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1084 on 307342 degrees of freedom
## Multiple R-squared:  0.06449,    Adjusted R-squared:  0.06446
## F-statistic:  1926 on 11 and 307342 DF,  p-value: < 2.2e-16
```

```
air_data_lm$propdelays_month_predicted <- lm_propdelays_vs_month$fitted.values
air_data_lm$propdelays_month_residuals <- lm_propdelays_vs_month$residuals

coef(lm_propdelays_vs_month)
```

```
##    (Intercept)  monthFebruary     monthMarch     monthApril       monthMay
##    0.207556739    0.001979567   -0.016162257   -0.031926668   -0.024343705
##      monthJune      monthJuly    monthAugust monthSeptember   monthOctober
##    0.029106688    0.025802051    0.001768113   -0.055676870   -0.039809018
##  monthNovember  monthDecember
##   -0.042295172    0.032333932
```

```
explanatory_data <- tibble(month = unique(air_data$month))
prediction_data <- explanatory_data %>%
  mutate(delay_proportion = predict(lm_propdelays_vs_month, explanatory_data)) %>%
  arrange(desc(delay_proportion))
prediction_data
```

```
## # A tibble: 12 x 2
##    month     delay_proportion
##    <fct>                <dbl>
##  1 December             0.240
##  2 June                 0.237
##  3 July                 0.233
##  4 February             0.210
##  5 August               0.209
##  6 January              0.208
##  7 March                0.191
##  8 May                  0.183
```

```
##  9 April                0.176
## 10 October              0.168
## 11 November             0.165
## 12 September            0.152
```

```
#Additionally, we used 3 lines of AirBNB example's code that creates
#a table showing the intercept plus the estimate.
#Using this table in addition to our linear regression summary made the visualization
#of our linear regression model a lot clearer.
```

Our model predicted that the time of year has a significant effect on the proportion of flight delays. The p-value for each month of the year is extremely small, showing that the proportion of flight delays varies significantly each month. The month with the greatest predicted delay proportion was December, while the month with the least predicted delay proportion was September. The predicted delay proportion in December was almost 0.09 greater than the predicted proportion in September!

LM of Proportion of Delays by Airport Size:

```
lm_delays_vs_airport_size <- lm(propdelays ~ airport_size, data = air_data_lm)
summary (lm_delays_vs_airport_size)
```

```
##
## Call:
## lm(formula = propdelays ~ airport_size, data = air_data_lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20251 -0.07538 -0.01319  0.05989  0.83066
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.2025056  0.0002419  837.25   <2e-16 ***
## airport_sizeMedium -0.0151892  0.0004467  -34.01   <2e-16 ***
## airport_sizeSmall  -0.0331668  0.0016363  -20.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1119 on 307351 degrees of freedom
## Multiple R-squared:  0.004748,   Adjusted R-squared:  0.004742
## F-statistic: 733.2 on 2 and 307351 DF,  p-value: < 2.2e-16
```

```
air_data_lm$propdelays_airport_size_predicted <- lm_propdelays_vs_month$fitted.values
air_data_lm$propdelays_airport_size_residuals <- lm_propdelays_vs_month$residuals

coef(lm_delays_vs_airport_size)
```

```
##        (Intercept) airport_sizeMedium  airport_sizeSmall
##         0.20250565        -0.01518925        -0.03316680
```

```
explanatory_data <- tibble(airport_size = unique(air_data$airport_size))
prediction_data <- explanatory_data %>%
  mutate(delay_proportion = predict(lm_delays_vs_airport_size, explanatory_data)) %>%
  arrange(desc(delay_proportion))
prediction_data
```

```
## # A tibble: 3 x 2
##   airport_size delay_proportion
##   <chr>                   <dbl>
## 1 Large                   0.203
## 2 Medium                  0.187
## 3 Small                   0.169
```

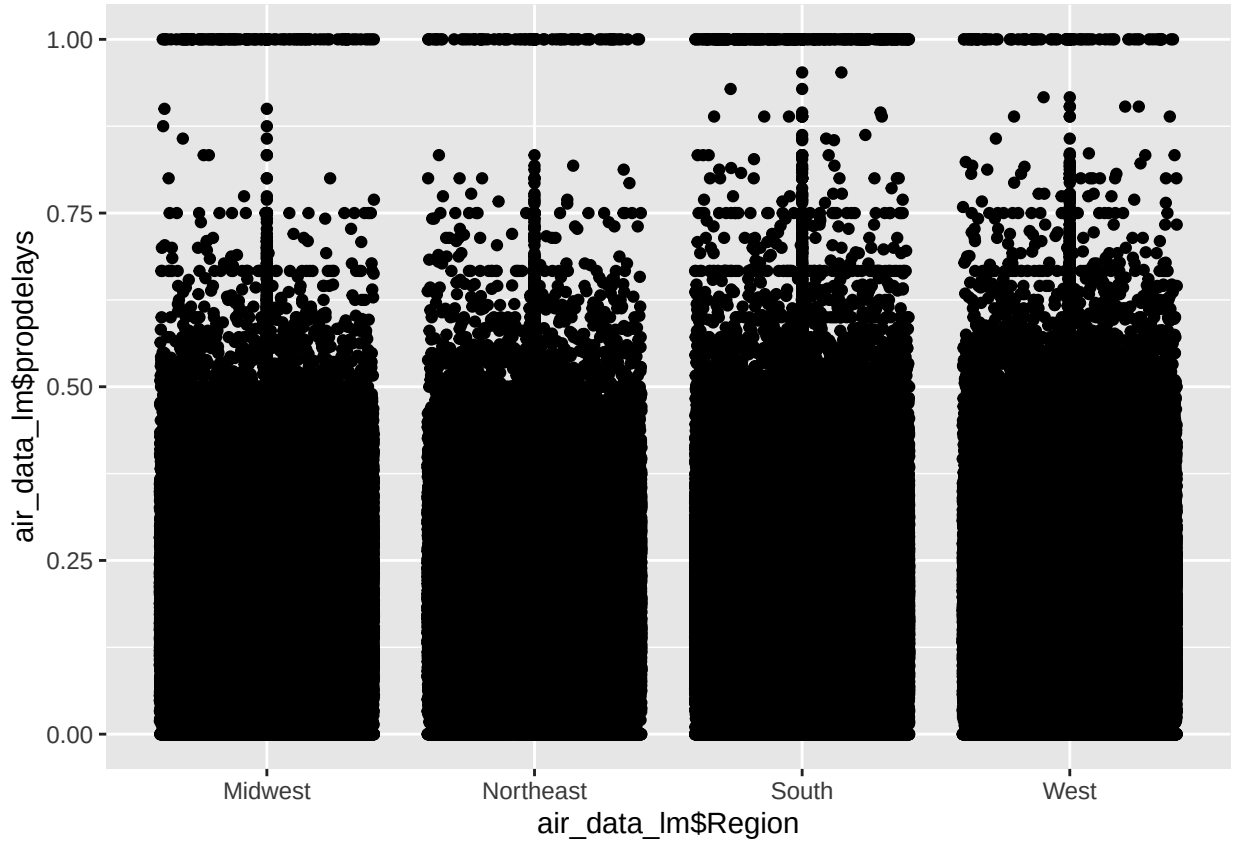
The p-values for delay proportions by airport size are also extremely small, signifying that there is variance in the proportion of flight delays depending on size of the airport. Large airports had the greatest predicted delay proportion, while small airports had the least predicted delay proportion.

LM of Proportion of Delays by Region:

```
lm_delays_vs_region <- lm(propdelays ~ Region, data = air_data_lm)
summary (lm_delays_vs_region)
```

```
##
## Call:
## lm(formula = propdelays ~ Region, data = air_data_lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21807 -0.07470 -0.01295  0.05960  0.81483
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.1984638  0.0004337 457.629   <2e-16 ***
## RegionNortheast  0.0196021  0.0006964  28.147   <2e-16 ***
## RegionSouth      0.0001895  0.0005440   0.348    0.728
## RegionWest      -0.0132928  0.0005810 -22.880   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1117 on 307350 degrees of freedom
## Multiple R-squared:  0.007935,   Adjusted R-squared:  0.007925
## F-statistic: 819.5 on 3 and 307350 DF,  p-value: < 2.2e-16
```

```
air_data_lm$propdelays_region_predicted <- lm_propdelays_vs_month$fitted.values
air_data_lm$propdelays_region_residuals <- lm_propdelays_vs_month$residuals

qplot(air_data_lm$Region, air_data_lm$propdelays) + geom_jitter()
```




```
coef(lm_delays_vs_region)
```

```
##     (Intercept) RegionNortheast     RegionSouth      RegionWest
##    0.1984638089    0.0196021253    0.0001895369   -0.0132928326
```

```
explanatory_data <- tibble(Region = unique(air_data$Region))
prediction_data <- explanatory_data %>%
  mutate(delay_proportion = predict(lm_delays_vs_region, explanatory_data)) %>%
  arrange(desc(delay_proportion))
prediction_data
```

```
## # A tibble: 4 x 2
##   Region    delay_proportion
##   <chr>                <dbl>
## 1 Northeast            0.218
## 2 South                0.199
## 3 Midwest              0.198
## 4 West                 0.185
```

The p-values for each region were all significantly small except for the Southern region. This is due to the fact that the delay proportion for the South does not significantly vary from that of the Midwest. Therefore, the delay proportion varies significantly by each region, with the exception of the South and Midwest regions.

LM of Proportion of Delays by COVID:

```
lm_delays_vs_covid <- lm(propdelays ~ covid_yesorno, data = air_data_lm)
summary (lm_delays_vs_covid)
```

```
##
## Call:
## lm(formula = propdelays ~ covid_yesorno, data = air_data_lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20633 -0.07447 -0.01384  0.05782  0.85230
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.2063292  0.0002154     958   <2e-16 ***
## covid_yesornoyes -0.0586269  0.0005583    -105   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1102 on 307352 degrees of freedom
## Multiple R-squared:  0.03463,    Adjusted R-squared:  0.03463
## F-statistic: 1.103e+04 on 1 and 307352 DF,  p-value: < 2.2e-16
```

```
air_data_lm$propdelays_covid_predicted <- lm_propdelays_vs_month$fitted.values
air_data_lm$propdelays_covid_residuals <- lm_propdelays_vs_month$residuals

coef(lm_delays_vs_covid)
```

```
##      (Intercept) covid_yesornoyes
##       0.20632923      -0.05862689
```

```
explanatory_data <- tibble(covid_yesorno = unique(air_data$covid_yesorno))
prediction_data <- explanatory_data %>%
  mutate(delay_proportion = predict(lm_delays_vs_covid, explanatory_data)) %>%
  arrange(desc(delay_proportion))
prediction_data
```

```
## # A tibble: 2 x 2
##   covid_yesorno delay_proportion
##   <chr>                   <dbl>
## 1 no                      0.206
## 2 yes                     0.148
```

The p-value of the flight delay proportion before and after COVID-19 is very small, showing a significant difference in the number of flight delays before and after COVID-19. This can be seen in the table that shows our model's predictions, with the delay proportion before COVID-19 being almost 0.06 higher than after the start of the pandemic!

Our model predicted that flight delay proportions varied significantly by time of year, airport size, region, and COVID-19.

LM of Proportion of Delays by Carrier:

```
lm_delays_vs_carrier <- lm(propdelays ~ carrier_name, data = air_data_lm)
summary (lm_delays_vs_carrier)
```

```
##
## Call:
## lm(formula = propdelays ~ carrier_name, data = air_data_lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25670 -0.07171 -0.01260  0.05750  0.87424
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          0.200170   0.001332 150.235  < 2e-16
## carrier_nameAlaska Airlines Inc.    -0.038320   0.001647 -23.269  < 2e-16
## carrier_nameAllegiant Air           -0.009274   0.001907  -4.863 1.16e-06
## carrier_nameAloha Airlines Inc.     -0.085831   0.006956 -12.339  < 2e-16
## carrier_nameAmerica West Airlines Inc.  0.023738   0.003090   7.682 1.57e-14
## carrier_nameAmerican Airlines Inc.   0.015885   0.001552  10.236  < 2e-16
## carrier_nameAmerican Eagle Airlines Inc.  0.028043   0.001597  17.561  < 2e-16
## carrier_nameATA Airlines d/b/a ATA   0.013007   0.003970   3.276 0.001054
## carrier_nameAtlantic Coast Airlines  0.024485   0.003379   7.246 4.30e-13
## carrier_nameAtlantic Southeast Airlines  0.056529   0.001662  34.020  < 2e-16
## carrier_nameComair Inc.              0.030687   0.001829  16.782  < 2e-16
## carrier_nameContinental Air Lines Inc.  0.016638   0.001902   8.748  < 2e-16
## carrier_nameDelta Air Lines Inc.    -0.036215   0.001490 -24.302  < 2e-16
## carrier_nameEndeavor Air Inc.       -0.074405   0.001988 -37.428  < 2e-16
## carrier_nameEnvoy Air               -0.025271   0.001760 -14.357  < 2e-16
## carrier_nameExpressJet Airlines Inc.  0.009277   0.001505   6.166 7.03e-10
## carrier_nameExpressJet Airlines LLC -0.018090   0.003188  -5.674 1.39e-08
## carrier_nameFrontier Airlines Inc.   0.022974   0.001663  13.815  < 2e-16
## carrier_nameHawaiian Airlines Inc.  -0.035495   0.002272 -15.624  < 2e-16
## carrier_nameHorizon Air             -0.058385   0.003981 -14.666  < 2e-16
## carrier_nameIndependence Air         0.019940   0.004451   4.480 7.45e-06
## carrier_nameJetBlue Airways          0.030369   0.001705  17.809  < 2e-16
## carrier_nameMesa Airlines Inc.      -0.002553   0.001613  -1.583 0.113477
## carrier_nameNorthwest Airlines Inc.  0.046777   0.001811  25.829  < 2e-16
## carrier_namePinnacle Airlines Inc.  -0.013176   0.001908  -6.905 5.02e-12
## carrier_namePSA Airlines Inc.       -0.030787   0.002153 -14.301  < 2e-16
## carrier_nameRepublic Airline        -0.051306   0.002120 -24.198  < 2e-16
## carrier_nameSkyWest Airlines Inc.   -0.024154   0.001445 -16.718  < 2e-16
## carrier_nameSouthwest Airlines Co.  -0.005945   0.001569  -3.789 0.000151
## carrier_nameSpirit Air Lines        -0.013349   0.002269  -5.883 4.04e-09
## carrier_nameUnited Air Lines Inc.    0.008146   0.001560   5.222 1.77e-07
## carrier_nameUS Airways Inc.         -0.004482   0.001716  -2.611 0.009027
## carrier_nameVirgin America          -0.013521   0.003313  -4.082 4.47e-05
##
## (Intercept)                         ***
## carrier_nameAlaska Airlines Inc.    ***
## carrier_nameAllegiant Air           ***
## carrier_nameAloha Airlines Inc.     ***
```

```
## carrier_nameAmerica West Airlines Inc.   ***
## carrier_nameAmerican Airlines Inc.       ***
## carrier_nameAmerican Eagle Airlines Inc. ***
## carrier_nameATA Airlines d/b/a ATA       **
## carrier_nameAtlantic Coast Airlines      ***
## carrier_nameAtlantic Southeast Airlines  ***
## carrier_nameComair Inc.                  ***
## carrier_nameContinental Air Lines Inc.   ***
## carrier_nameDelta Air Lines Inc.         ***
## carrier_nameEndeavor Air Inc.            ***
## carrier_nameEnvoy Air                    ***
## carrier_nameExpressJet Airlines Inc.     ***
## carrier_nameExpressJet Airlines LLC      ***
## carrier_nameFrontier Airlines Inc.       ***
## carrier_nameHawaiian Airlines Inc.       ***
## carrier_nameHorizon Air                  ***
## carrier_nameIndependence Air             ***
## carrier_nameJetBlue Airways              ***
## carrier_nameMesa Airlines Inc.
## carrier_nameNorthwest Airlines Inc.      ***
## carrier_namePinnacle Airlines Inc.       ***
## carrier_namePSA Airlines Inc.            ***
## carrier_nameRepublic Airline             ***
## carrier_nameSkyWest Airlines Inc.        ***
## carrier_nameSouthwest Airlines Co.       ***
## carrier_nameSpirit Air Lines             ***
## carrier_nameUnited Air Lines Inc.        ***
## carrier_nameUS Airways Inc.              **
## carrier_nameVirgin America               ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1086 on 307321 degrees of freedom
## Multiple R-squared:  0.06192,    Adjusted R-squared:  0.06183
## F-statistic:   634 on 32 and 307321 DF,  p-value: < 2.2e-16
```

```
air_data_lm$propdelays_carrier_predicted <- lm_propdelays_vs_month$fitted.values
air_data_lm$propdelays_carrier_residuals <- lm_propdelays_vs_month$residuals

coef(lm_delays_vs_carrier)
```

```
##                               (Intercept)
##                               0.200169552
##         carrier_nameAlaska Airlines Inc.
##                              -0.038319914
##                carrier_nameAllegiant Air
##                              -0.009273794
##          carrier_nameAloha Airlines Inc.
##                              -0.085830637
##   carrier_nameAmerica West Airlines Inc.
##                               0.023737746
##       carrier_nameAmerican Airlines Inc.
##                               0.015885079
## carrier_nameAmerican Eagle Airlines Inc.
```

```
##                              0.028042637
##       carrier_nameATA Airlines d/b/a ATA
##                              0.013006654
##      carrier_nameAtlantic Coast Airlines
##                              0.024485004
##  carrier_nameAtlantic Southeast Airlines
##                              0.056529366
##                 carrier_nameComair Inc.
##                              0.030686512
##   carrier_nameContinental Air Lines Inc.
##                              0.016638158
##         carrier_nameDelta Air Lines Inc.
##                             -0.036214703
##            carrier_nameEndeavor Air Inc.
##                             -0.074405089
##                   carrier_nameEnvoy Air
##                             -0.025270971
##     carrier_nameExpressJet Airlines Inc.
##                              0.009277468
##      carrier_nameExpressJet Airlines LLC
##                             -0.018089548
##       carrier_nameFrontier Airlines Inc.
##                              0.022973619
##       carrier_nameHawaiian Airlines Inc.
##                             -0.035495207
##                 carrier_nameHorizon Air
##                             -0.058385157
##             carrier_nameIndependence Air
##                              0.019939955
##              carrier_nameJetBlue Airways
##                              0.030369006
##           carrier_nameMesa Airlines Inc.
##                             -0.002552765
##      carrier_nameNorthwest Airlines Inc.
##                              0.046776608
##       carrier_namePinnacle Airlines Inc.
##                             -0.013175777
##            carrier_namePSA Airlines Inc.
##                             -0.030786835
##             carrier_nameRepublic Airline
##                             -0.051305930
##        carrier_nameSkyWest Airlines Inc.
##                             -0.024154169
##       carrier_nameSouthwest Airlines Co.
##                             -0.005945369
##             carrier_nameSpirit Air Lines
##                             -0.013349111
##        carrier_nameUnited Air Lines Inc.
##                              0.008145891
##              carrier_nameUS Airways Inc.
##                             -0.004481815
##               carrier_nameVirgin America
##                             -0.013521067
```

```
explanatory_data <- tibble(carrier_name = unique(air_data$carrier_name))
prediction_data <- explanatory_data %>%
  mutate(delay_proportion = predict(lm_delays_vs_carrier, explanatory_data)) %>%
  arrange(desc(delay_proportion))
prediction_data
```

```
## # A tibble: 33 x 2
##    carrier_name                 delay_proportion
##    <chr>                                   <dbl>
##  1 Atlantic Southeast Airlines             0.257
##  2 Northwest Airlines Inc.                 0.247
##  3 Comair Inc.                             0.231
##  4 JetBlue Airways                         0.231
##  5 American Eagle Airlines Inc.            0.228
##  6 Atlantic Coast Airlines                 0.225
##  7 America West Airlines Inc.              0.224
##  8 Frontier Airlines Inc.                  0.223
##  9 Independence Air                        0.220
## 10 Continental Air Lines Inc.              0.217
## # ... with 23 more rows
```

The predicted data is the expected value based on the linear modelling, and the residual shows how far away the predicted value is from the actual value. Now we can create residuals plots:

4.2: Coefficient Analysis:

Although all of our variables have small p-values, we should also look at the coefficient values of the variable estimates to see how much these variables affect the proportion of flight delays.

```
lm_delays_vs_variables <- lm(propdelays ~ month + airport_size + Region + covid_yesorno, data = air_data
summary(lm_delays_vs_variables)
```

```
##
## Call:
## lm(formula = propdelays ~ month + airport_size + Region + covid_yesorno,
##     data = air_data_lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27040 -0.06850 -0.01311  0.05354  0.90635
##
## Coefficients:
##                      Estimate Std. Error  t value Pr(>|t|)
## (Intercept)         0.2254899  0.0007753  290.856  < 2e-16 ***
## monthFebruary       0.0019240  0.0009325    2.063 0.039078 *
## monthMarch         -0.0159846  0.0009310  -17.170  < 2e-16 ***
## monthApril         -0.0324904  0.0009306  -34.913  < 2e-16 ***
## monthMay           -0.0248703  0.0009335  -26.642  < 2e-16 ***
## monthJune           0.0248292  0.0009337   26.592  < 2e-16 ***
## monthJuly           0.0218771  0.0009355   23.386  < 2e-16 ***
## monthAugust        -0.0023041  0.0009308   -2.475 0.013307 *
```

```
## monthSeptember     -0.0597011  0.0009301  -64.189  < 2e-16 ***
## monthOctober       -0.0440024  0.0009336  -47.132  < 2e-16 ***
## monthNovember      -0.0465520  0.0009327  -49.913  < 2e-16 ***
## monthDecember       0.0283592  0.0009295   30.511  < 2e-16 ***
## airport_sizeMedium -0.0132850  0.0004255  -31.225  < 2e-16 ***
## airport_sizeSmall  -0.0204154  0.0015547  -13.131  < 2e-16 ***
## RegionNortheast     0.0165494  0.0006630   24.962  < 2e-16 ***
## RegionSouth        -0.0017225  0.0005179   -3.326 0.000881 ***
## RegionWest         -0.0162825  0.0005525  -29.468  < 2e-16 ***
## covid_yesornoyes   -0.0571359  0.0005406 -105.683  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1058 on 307336 degrees of freedom
## Multiple R-squared:  0.1097, Adjusted R-squared:  0.1097
## F-statistic:  2228 on 17 and 307336 DF,  p-value: < 2.2e-16
```

The months of December, June, and July, and the years during and after COVID-19 had the greatest absolute value coefficients, meaning they had the greatest effect on the proportion of flight delays. December, June, and July had respective coefficients of 0.061, 0.057, and 0.054, meaning the model predicted that the proportion of flight delays would be 0.061, 0.057, and 0.054 greater than average in these months. Additionally, in the years of COVID-19, the coefficient was -0.057, meaning the model predicted the proportion of flight delays to be 0.057 lower than the mean.

On the other hand, the month of May and the South region had the lowest absolute value coefficients, meaning they had the least effect on the proportion of flight delays. They had respective coefficients of 0.008 and -0.002, meaning our model predicted the flight delay proportion to be 0.008 greater than the mean during May and 0.002 lower than the mean in the South.