Batch 6 - Exploratory Data Analysis

Sawanya Watanakijcharoenman

```
library(tidyverse)
```

Load library

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)
library(ggthemes)
library(RColorBrewer)
```

```
glimpse(diamonds)
```

Diamonds Data Overview

```
## Rows: 53,940
## Columns: 10
## $ carat   <dbl> 0.23, 0.21, 0.23, 0.29, 0.31, 0.24, 0.24, 0.26, 0.22, 0.23, 0.~
## $ cut     <ord> Ideal, Premium, Good, Premium, Good, Very Good, Very Good, Ver~
## $ color   <ord> E, E, E, I, J, J, I, H, E, H, J, J, F, J, E, E, I, J, J, J, I,~
## $ clarity <ord> SI2, SI1, VS1, VS2, SI2, VVS2, VVS1, SI1, VS2, VS1, SI1, VS1, ~
## $ depth   <dbl> 61.5, 59.8, 56.9, 62.4, 63.3, 62.8, 62.3, 61.9, 65.1, 59.4, 64~
## $ table   <dbl> 55, 61, 65, 58, 58, 57, 57, 55, 61, 61, 55, 56, 61, 54, 62, 58~
## $ price   <int> 326, 326, 327, 334, 335, 336, 336, 337, 337, 338, 339, 340, 34~
## $ x       <dbl> 3.95, 3.89, 4.05, 4.20, 4.34, 3.94, 3.95, 4.07, 3.87, 4.00, 4.~
## $ y       <dbl> 3.98, 3.84, 4.07, 4.23, 4.35, 3.96, 3.98, 4.11, 3.78, 4.05, 4.~
## $ z       <dbl> 2.43, 2.31, 2.31, 2.63, 2.75, 2.48, 2.47, 2.53, 2.49, 2.39, 2.~
```

```
summary(diamonds)
```

Data Summary of Diamonds Data

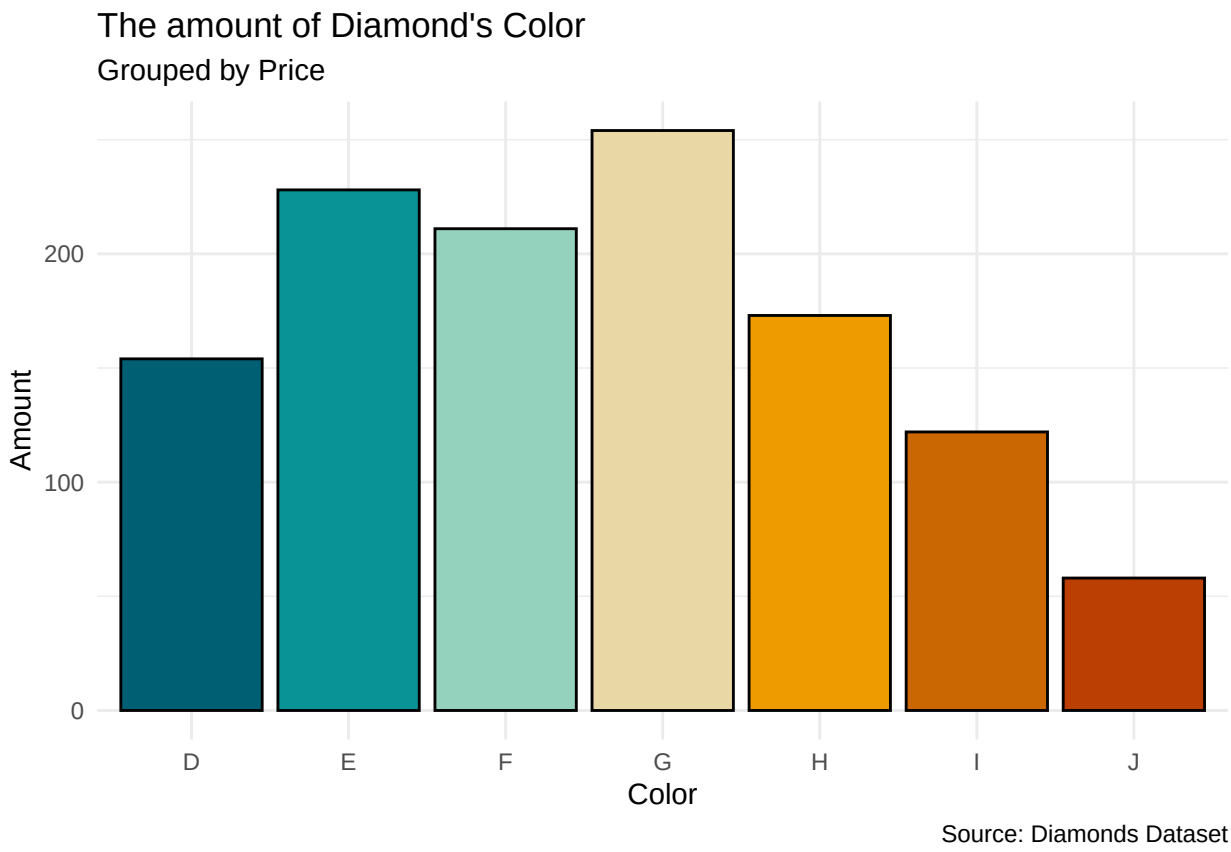
```
##      carat               cut        color        clarity          depth
##  Min.   :0.2000   Fair      : 1610   D: 6775   SI1    :13065   Min.   :43.00
##  1st Qu.:0.4000   Good      : 4906   E: 9797   VS2    :12258   1st Qu.:61.00
##  Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194   Median :61.80
```

```
##  Mean   :0.7979   Premium  :13791   G:11292   VS1    : 8171   Mean   :61.75
##  3rd Qu.:1.0400   Ideal    :21551   H: 8304   VVS2   : 5066   3rd Qu.:62.50
##  Max.   :5.0100                     I: 5422   VVS1   : 3655   Max.   :79.00
##                                     J: 2808   (Other): 2531
##      table           price             x                y
##  Min.   :43.00   Min.   :  326   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720
##  Median :57.00   Median : 2401   Median : 5.700   Median : 5.710
##  Mean   :57.46   Mean   : 3933   Mean   : 5.731   Mean   : 5.735
##  3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540
##  Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :58.900
##
##        z
##  Min.   : 0.000
##  1st Qu.: 2.910
##  Median : 3.530
##  Mean   : 3.539
##  3rd Qu.: 4.040
##  Max.   :31.800
##
```

The Amount of Diamond's Color

- According to the bar chart, the color scale of diamonds sorted by price has G as the highest number, followed by E, and J as the lowest.

```
set.seed(64)
ggplot(sample_n(diamonds, 1200), aes(color, fill = price)) +
  geom_bar(color = "black", fill = c("#005F73", "#0A9396", "#94D2BD",
                   "#E9D8A6", "#EE9B00", "#CA6702",
                   "#BB3E03")) +
  labs(
    title = "The amount of Diamond's Color",
    x = "Color",
    y = "Amount",
    subtitle = "Grouped by Price",
    caption = "Source: Diamonds Dataset"
  ) +
  theme_minimal()
```

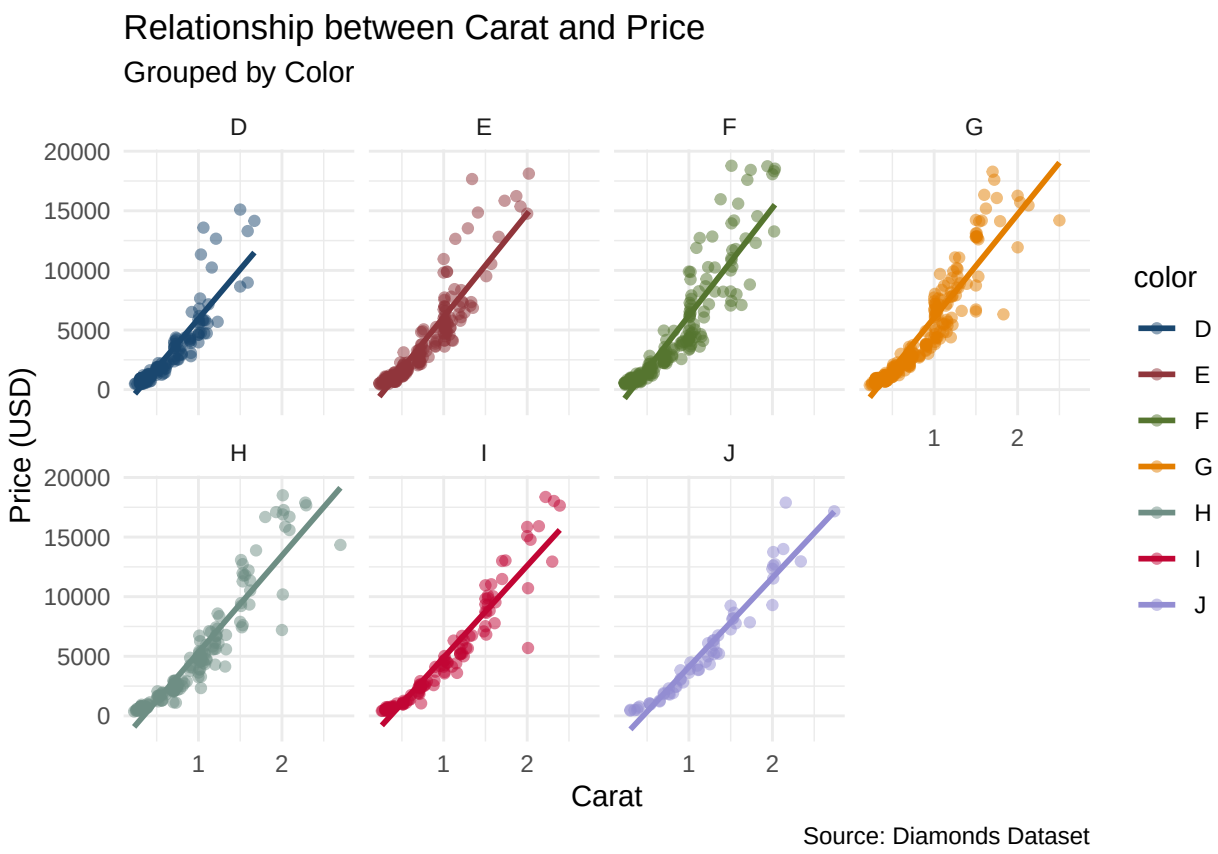


Relationship between Carat and Price

- As the carat size grows, prices rise, as shown by the graph.

```
## using sample 500
set.seed(64)
ggplot(sample_n(diamonds, 1200), aes(carat,price, color = color)) +
  geom_point(alpha=0.5) +
  geom_smooth(method = "lm", se=F) +
  labs(
    title = "Relationship between Carat and Price",
    x = "Carat",
    y = "Price (USD)",
    subtitle = "Grouped by Color",
    caption = "Source: Diamonds Dataset"
  ) +
  theme_minimal()+
  scale_color_stata()+
  facet_wrap(~ color, ncol = 4)
```

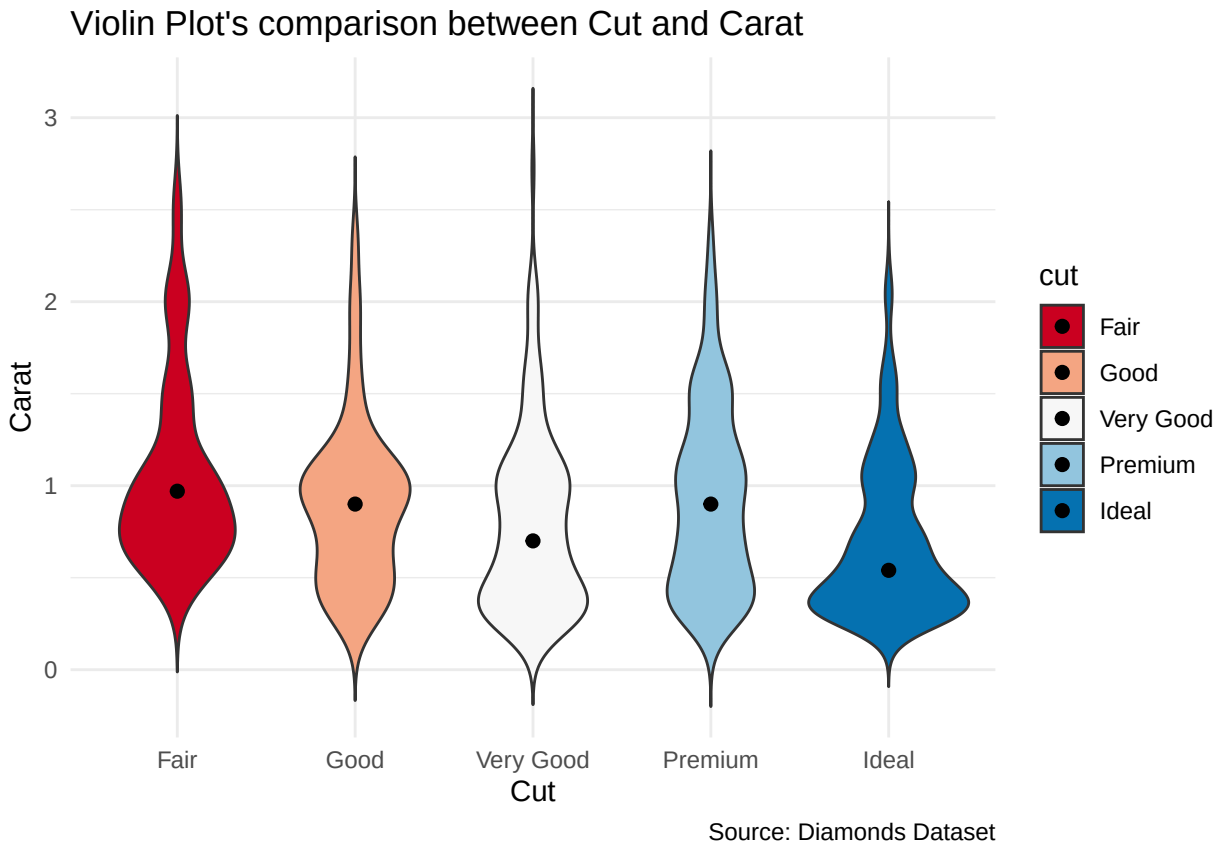
```
## `geom_smooth()` using formula 'y ~ x'
```



Distribution of each carat

- The violin plot, which shows a density of observations for the cut by carat, The dot represents the median for each cut. The median carat weight of good, very good, and ideal diamonds is less than one carat. Few fair and premium cut diamonds have median weights that are less than or equal to 1 carat but significantly higher than that.

```
set.seed(64)
ggplot(sample_n(diamonds, 1200), aes(cut, carat, fill = cut)) +
  geom_violin(trim = F) +
  stat_summary(fun = median, geom = "point", size = 2, color = "black") +
  theme_minimal() +
  labs(
    title = "Violin Plot's comparison between Cut and Carat",
    x = "Cut",
    y = "Carat",
    caption = "Source: Diamonds Dataset"
  ) +
  scale_fill_brewer(palette = "RdBu")
```




```
install.packages("nycflights13")
```

Flights Data Overview

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(nycflights13)
glimpse(flights)
```

```
## Rows: 336,776
## Columns: 19
## $ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2~
## $ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ dep_time       <int> 517, 533, 542, 544, 554, 554, 555, 557, 557, 558, 558, ~
## $ sched_dep_time <int> 515, 529, 540, 545, 600, 558, 600, 600, 600, 600, 600, ~
## $ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2, -2, -1~
## $ arr_time       <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 753, 849,~
## $ sched_arr_time <int> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 745, 851,~
## $ arr_delay      <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -3, 7, -1~
## $ carrier        <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV", "B6", "~
## $ flight         <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79, 301, 4~
## $ tailnum        <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN", "N394~
## $ origin         <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR", "LGA",~
## $ dest           <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL", "IAD",~
## $ air_time       <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138, 149, 1~
```

```
## $ distance       <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 944, 733, ~
## $ hour           <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 5, 6, 6, 6~
## $ minute         <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, 59, 0~
## $ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013-01-01 0~
```

```
summary(flights)
```

Data Summary of Flights Data

```
##       year          month             day           dep_time    sched_dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1   Min.   : 106
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907   1st Qu.: 906
##  Median :2013   Median : 7.000   Median :16.00   Median :1401   Median :1359
##  Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349   Mean   :1344
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744   3rd Qu.:1729
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400   Max.   :2359
##                                                  NA's   :8255
##    dep_delay          arr_time    sched_arr_time   arr_delay
##  Min.   : -43.00   Min.   :   1   Min.   :   1   Min.   : -86.000
##  1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1124   1st Qu.: -17.000
##  Median :  -2.00   Median :1535   Median :1556   Median :  -5.000
##  Mean   :  12.64   Mean   :1502   Mean   :1536   Mean   :   6.895
##  3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1945   3rd Qu.:  14.000
##  Max.   :1301.00   Max.   :2400   Max.   :2359   Max.   :1272.000
##  NA's   :8255      NA's   :8713                  NA's   :9430
##    carrier              flight       tailnum             origin
##  Length:336776      Min.   :   1   Length:336776      Length:336776
##  Class :character   1st Qu.: 553   Class :character   Class :character
##  Mode  :character   Median :1496   Mode  :character   Mode  :character
##                     Mean   :1972
##                     3rd Qu.:3465
##                     Max.   :8500
##
##      dest              air_time        distance         hour
##  Length:336776      Min.   : 20.0   Min.   :  17   Min.   : 1.00
##  Class :character   1st Qu.: 82.0   1st Qu.: 502   1st Qu.: 9.00
##  Mode  :character   Median :129.0   Median : 872   Median :13.00
##                     Mean   :150.7   Mean   :1040   Mean   :13.18
##                     3rd Qu.:192.0   3rd Qu.:1389   3rd Qu.:17.00
##                     Max.   :695.0   Max.   :4983   Max.   :23.00
##                     NA's   :9430
##      minute        time_hour
##  Min.   : 0.00   Min.   :2013-01-01 05:00:00.00
##  1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00.00
##  Median :29.00   Median :2013-07-03 10:00:00.00
##  Mean   :26.23   Mean   :2013-07-03 05:22:54.64
##  3rd Qu.:44.00   3rd Qu.:2013-10-01 07:00:00.00
##  Max.   :59.00   Max.   :2013-12-31 23:00:00.00
##
```

Top 5 carriers dominate from NYC Airports

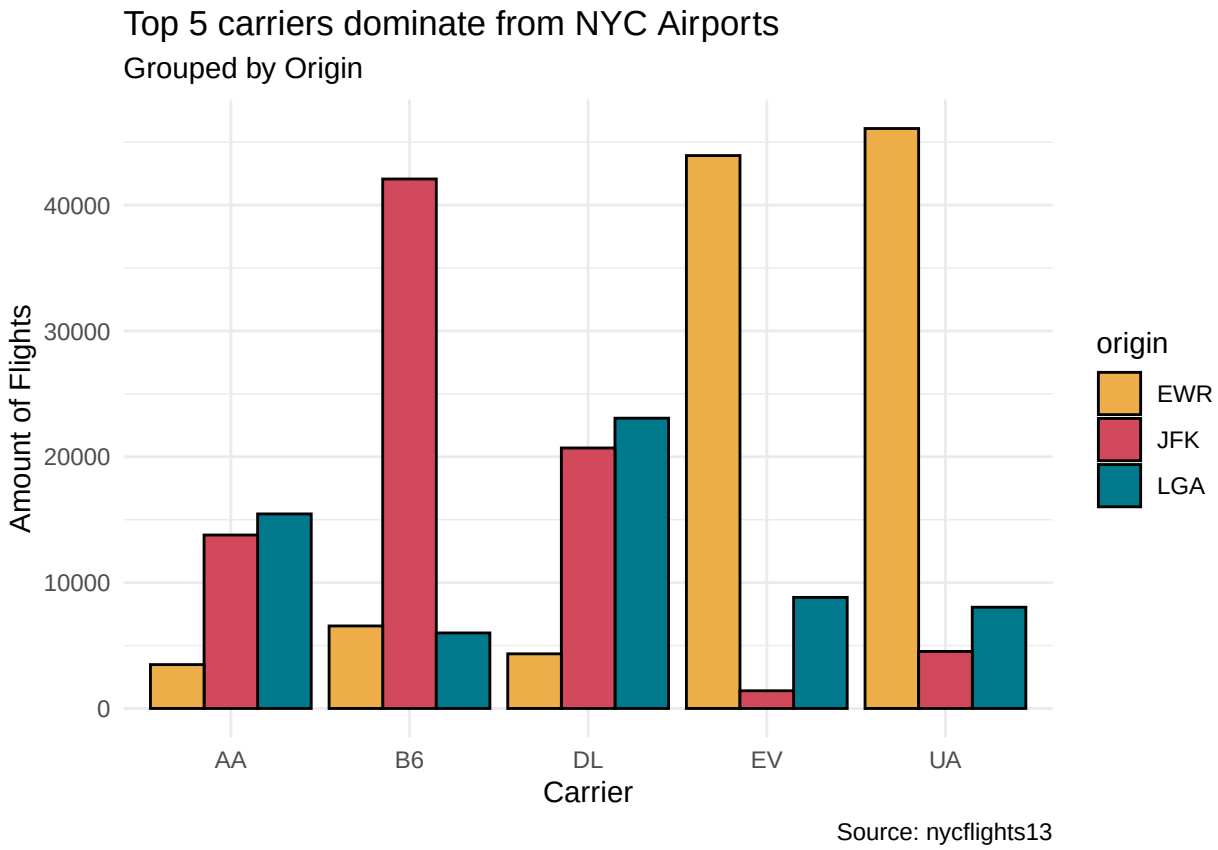
- The mass of flights out of NYC airports were operated by five large airlines:

1. United Air Lines (UA)
2. JetBlue Airways (B6)
3. ExpressJet Airlines (EV)
4. Delta Air Lines (DL)
5. American Airlines (AA)

origin	carrier	n	name
1 EWR	UA	46087	United Air Lines Inc.
2 EWR	EV	43939	ExpressJet Airlines Inc.
3 JFK	B6	42076	JetBlue Airways
4 LGA	DL	23067	Delta Air Lines Inc.
5 JFK	DL	20701	Delta Air Lines Inc.
6 LGA	AA	15459	American Airlines Inc.
7 JFK	AA	13783	American Airlines Inc.
8 LGA	EV	8826	ExpressJet Airlines Inc.
9 LGA	UA	8044	United Air Lines Inc.
10 EWR	B6	6557	JetBlue Airways
11 LGA	B6	6002	JetBlue Airways
12 JFK	UA	4534	United Air Lines Inc.
13 EWR	DL	4342	Delta Air Lines Inc.
14 EWR	AA	3487	American Airlines Inc.
15 JFK	EV	1408	ExpressJet Airlines Inc.

```
top <- flights %>%
  select(origin, carrier) %>%
  group_by(origin, carrier) %>%
  count(carrier) %>%
  arrange(desc(n)) %>%
  left_join(airlines,"carrier") %>%
  filter(carrier == "UA" |
         carrier == "B6"|
         carrier == "EV"|
         carrier == "DL"|
         carrier == "AA"
         )

ggplot(top, aes(carrier, n, fill = origin)) +
  geom_col(position = "dodge", color = "black") +
  labs(
    title = "Top 5 carriers dominate from NYC Airports",
    x = "Carrier",
    y = "Amount of Flights",
    subtitle = "Grouped by Origin",
    caption = "Source: nycflights13"
  ) +
  theme_minimal() +
  scale_fill_manual(values = c("#edae49", "#d1495b", "#00798c"))
```

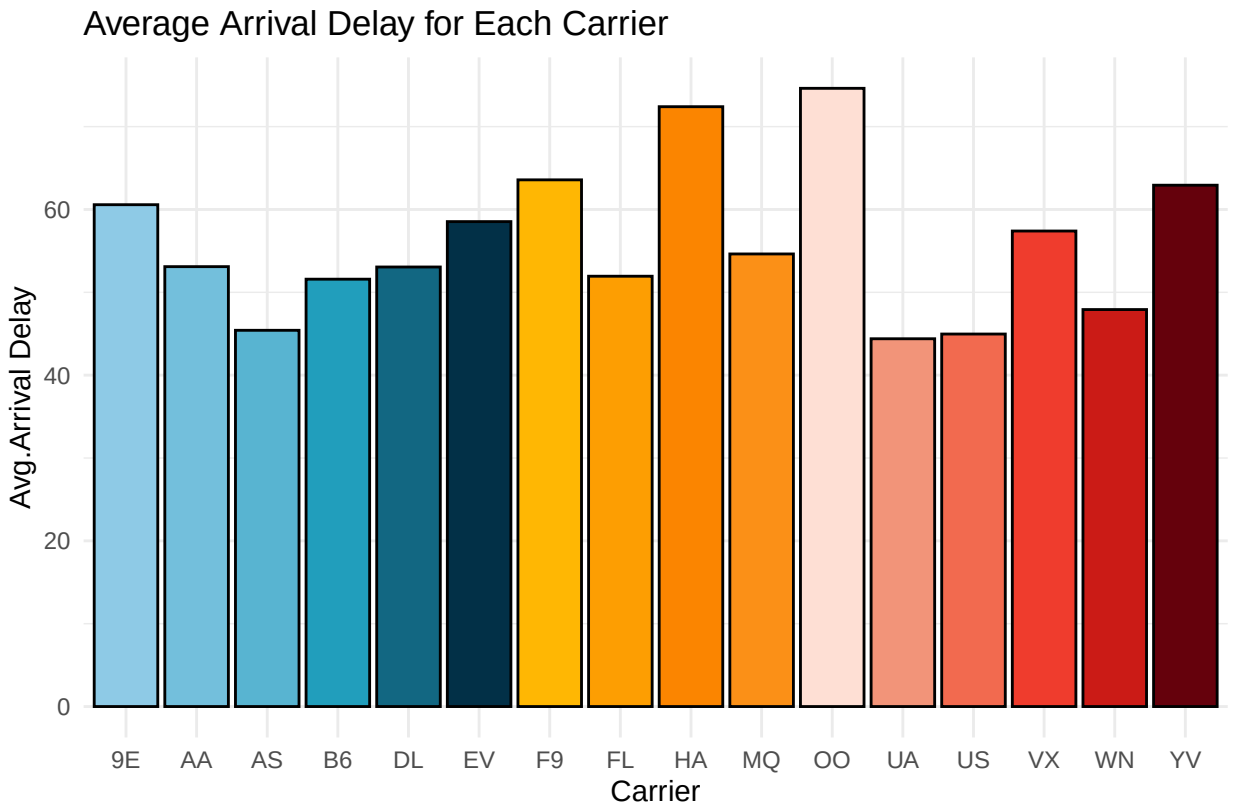


Average Arrival Delay for Each Carrier

- According to the chart, OO and HA have higher arrival delays than other carriers, whereas UA and US carriers perform best.

```
delayed <- flights %>%
  filter(dep_delay > 0 & arr_delay > 0) %>%
  group_by(carrier) %>%
  summarize(m = mean(arr_delay))

ggplot(delayed, aes(carrier, m)) +
  geom_col(color = "black",
           fill= c("#8ecae6","#73bfdc","#58b4d1","#219ebc","#126782",
                   "#023047","#ffb703","#fd9e02","#fb8500","#fb9017",
                   "#fedfd4","#f29479","#f26a4f","#ef3c2d","#cb1b16",
                   "#65010c")) +
  labs(
    title = "Average Arrival Delay for Each Carrier",
    x = "Carrier",
    y = "Avg.Arrival Delay",
    caption = "Source: nycflights13"
  ) +
  theme_minimal()
```


Average Arrival Delay for Each Carrier
Avg.Arrival Delay
0
20
40
60
9E
AA
AS
B6
DL
EV
F9
FL
HA
MQ
OO
UA
US
VX
WN
YV
Carrier
Source: nycflights13