HW 1

Zach Fechko (011711215)

1/22/23

```
library(tidyverse)
library(ggplot2)
library(dplyr)
library(nycflights13)
library(plotly)
```

Problem 1

```
# create data frame df from flights with the following attributes:
# - months 12, 1, 2, 6, 7, and 8
# - carriers UA, AA, and DL
# - distance greater than 700

flights_sml <- flights %>% select(month, carrier, distance, arr_delay)

#dataframe grouped by month
month.df <- flights_sml %>%
    filter(month %in% c(12, 1, 2, 6, 7, 8) & carrier %in% c("UA", "AA", "DL") & distance >
    group_by(month)


#dataframe grouped by carrier
carrier.df <- flights_sml %>%
    filter(month %in% c(12, 1, 2, 6, 7, 8) & carrier %in% c("UA", "AA", "DL") & distance >
    group_by(carrier)
```

Problem 1.a

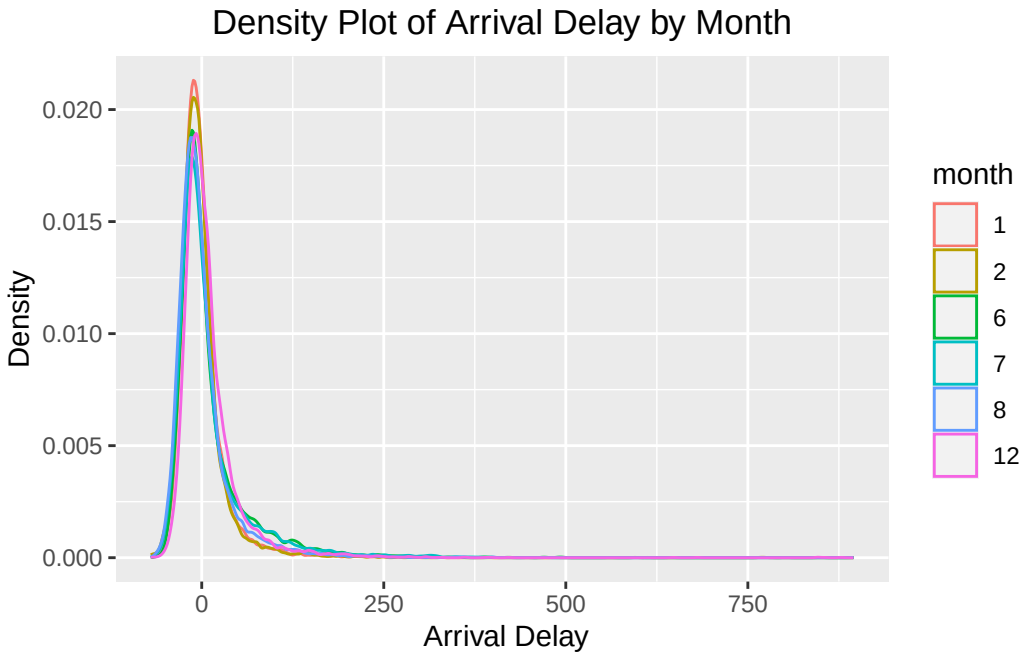
In a single plot, create a density plot for `arr_delay` for each of the 6 months with `color` aesthetic designated by month. Note that you need to convert `month` to a factor in order to create the plot. What can you say about the average `arr_delay` for each month?

```
#convert month to factor
month.df$month <- as.factor(month.df$month)

# create density plot
p1a <- ggplot(month.df) +
    geom_density(mapping = aes(x = arr_delay, color = month)) +
    labs(title = "Density Plot of Arrival Delay by Month", x = " Arrival Delay", y = "Dens
    theme(plot.title = element_text(hjust = 0.5))

p1a
```

```
Warning: Removed 1307 rows containing non-finite values (`stat_density()`).
```



In the density plot, we can see that the average arrival delay for each month centers around 0.

Problem 1.b

In a single plot, create a boxplot for `arr_delay` for each of the 3 carriers. What can you say about the average `arr_delay` for each carrier?

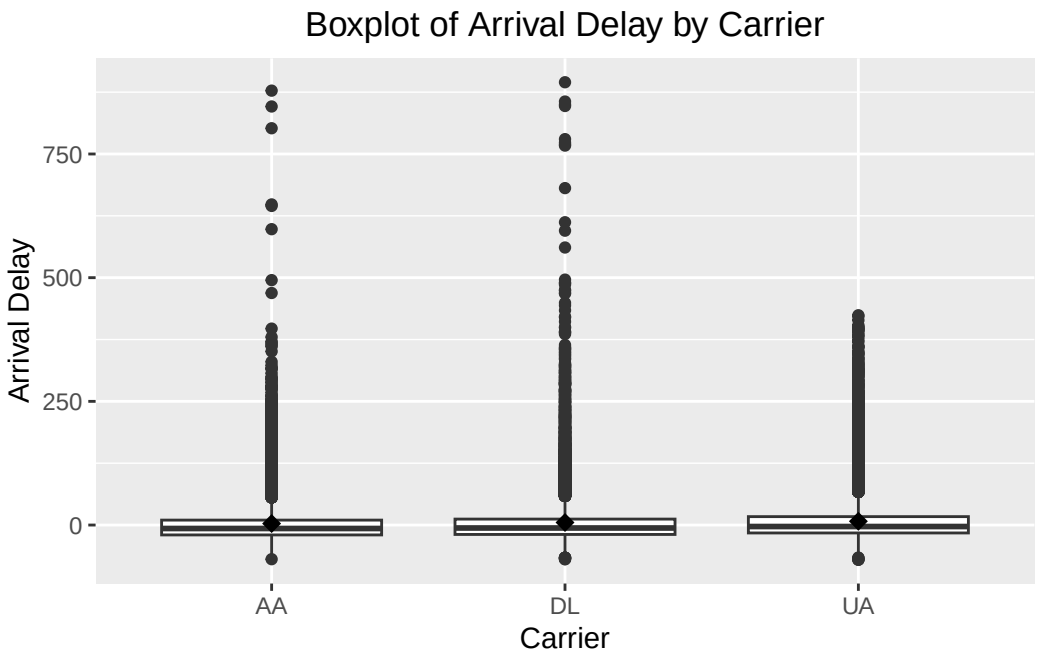
```
p1b <- ggplot(carrier.df, aes(x = carrier, y = arr_delay)) +
    geom_boxplot() +
    stat_summary(fun.y = mean, geom = "point", shape = 18, size = 3) +
    labs(title = "Boxplot of Arrival Delay by Carrier", x = "Carrier", y = "Arrival Delay"
    theme(plot.title = element_text(hjust = 0.5))
```

```
Warning: The `fun.y` argument of `stat_summary()` is deprecated as of ggplot2 3.3.0.
i Please use the `fun` argument instead.
```

```
p1b
```

```
Warning: Removed 1307 rows containing non-finite values (`stat_boxplot()`).
```

```
Warning: Removed 1307 rows containing non-finite values (`stat_summary()`).
```



Carrier	Mean Arrival Delay
AA	2.978
DL	5.112
UA	7.599

United airlines had the highest average arrival delay, followed by Delta, and then American.

Problem 1.c

Create a pie chart for the 3 carriers where the percentages are the proportions of observations and where percentages are superimposed on the sectors of the pie chart.

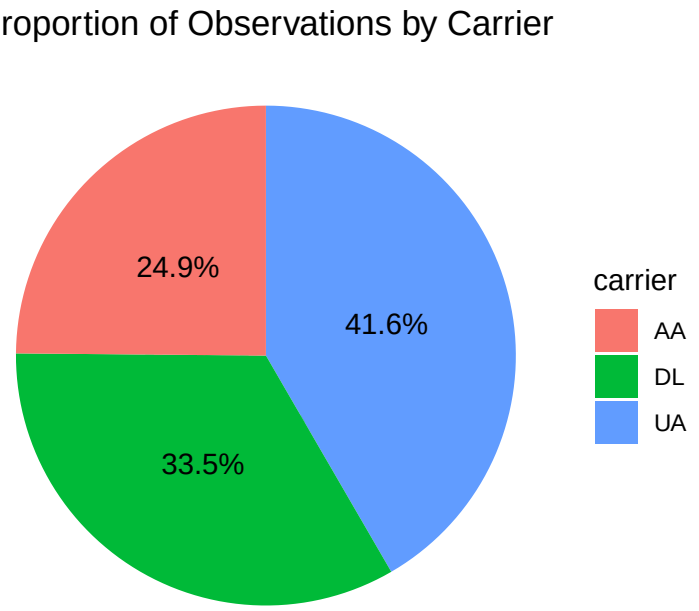
```
library(scales)

df1c <- carrier.df %>%
    group_by(carrier) %>%
    dplyr::count() %>% ungroup() %>%
    mutate(percentage = n/sum(n)) %>%
    dplyr::arrange(desc(carrier))

df1c$labels <- scales::percent(df1c$percentage)

# create pie chart
p1c <- ggplot(df1c, aes(x = "", y = percentage, fill = carrier)) +
    geom_bar(width = 1, stat = "identity") +
    geom_text(aes(label = labels), position = position_stack(vjust = 0.5)) +
    coord_polar("y", start = 0) +
    labs(title = "Proportion of Observations by Carrier", x = "", y = "Proportion") +
    theme_void() +
    theme(plot.title = element_text(hjust = 0.5))

p1c
```



Problem 1.d

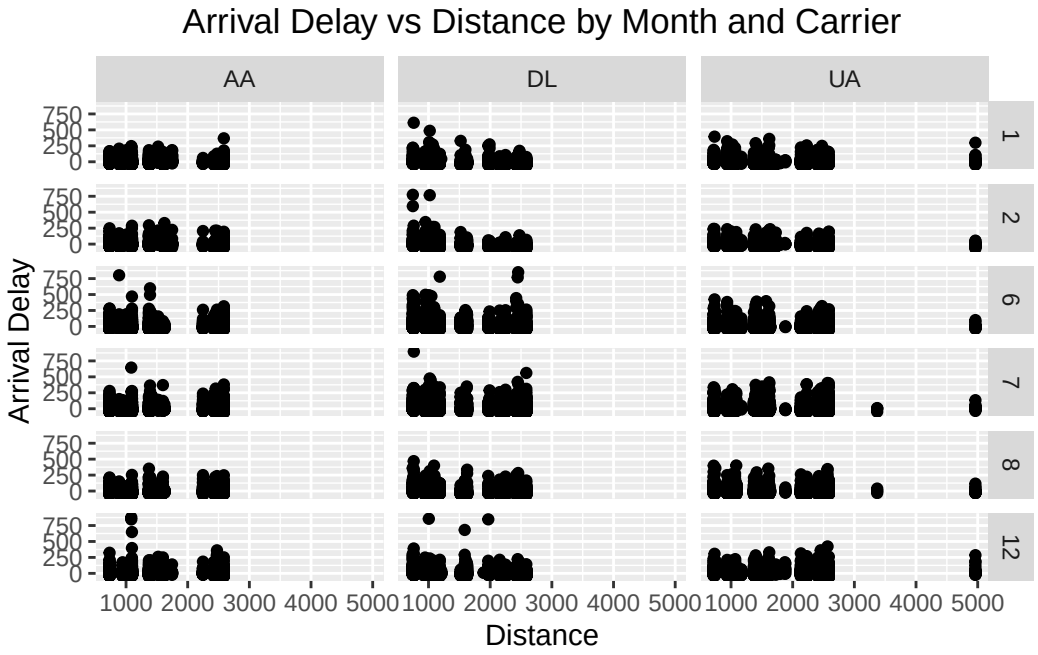
Plot `arr_delay` against `distance` with `facet_grid` designated by `month` and `carrier`

```
df1d <- flights_sml %>%
    filter(month %in% c(12, 1, 2, 6, 7, 8) & carrier %in% c("UA", "AA", "DL") & distance >

p1d <- ggplot(df1d, aes(x = distance, y = arr_delay)) +
    geom_point() +
    facet_grid(month ~ carrier) +
    labs(title = "Arrival Delay vs Distance by Month and Carrier", x = "Distance", y = "Ar
    theme(plot.title = element_text(hjust = 0.5))

p1d
```

```
Warning: Removed 1307 rows containing missing values (`geom_point()`).
```



Problem 1.e

For each feasible combination of values of `month` and `carrier`, compute the sample average of `arr_delay` and save them into the variable `mean_arr_delay`, and compute the sample average of `distance` and save these averages into the variable `mean_distance`.

Plot `month` against `mean_arr_delay` with shape designated by `carrier` and color by `mean_distance` and annotate each point by its associated carrier name.

```
df1e <- flights_sml %>%
    filter(month %in% c(12, 1, 2, 6, 7, 8) & carrier %in% c("UA", "AA", "DL") & distance >
    group_by(month, carrier) %>%
    summarise(mean_arr_delay = mean(arr_delay, na.rm = TRUE), mean_distance = mean(distanc
```

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

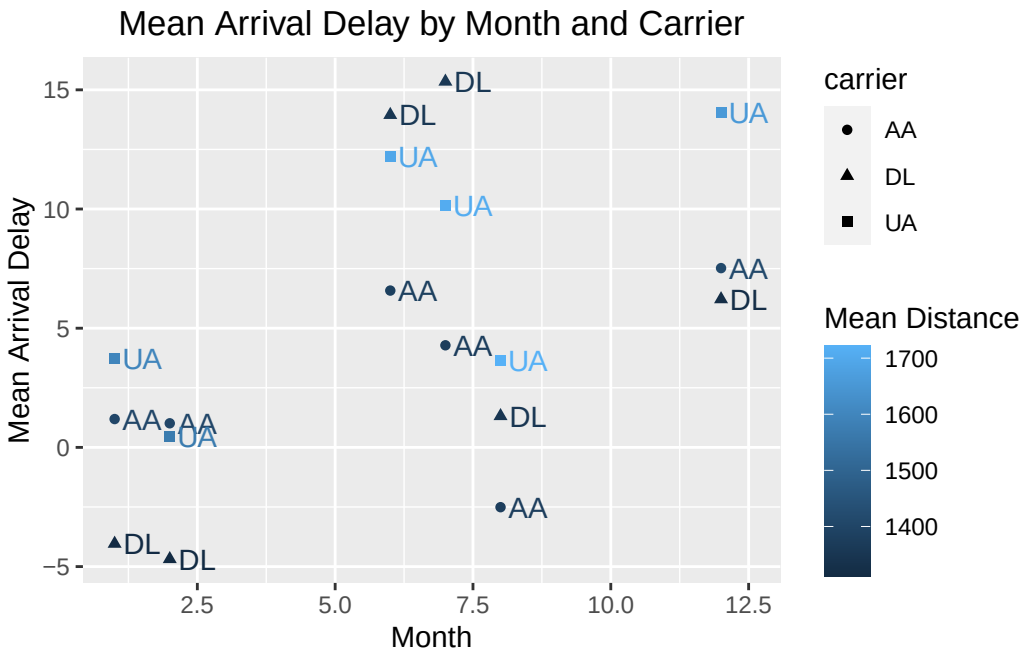
```
df1e
```

```
# A tibble: 18 x 4
# Groups:   month [6]
```

```
   month carrier mean_arr_delay mean_distance
   <int> <chr>            <dbl>         <dbl>
 1     1 AA               1.19          1404.
 2     1 DL              -4.04          1314.
 3     1 UA               3.72          1598.
 4     2 AA               1.01          1404.
 5     2 DL              -4.68          1312.
 6     2 UA               0.470         1569.
 7     6 AA               6.58          1382.
 8     6 DL              13.9           1353.
 9     6 UA              12.2           1693.
10     7 AA               4.28          1376.
11     7 DL              15.3           1357.
12     7 UA              10.2           1708.
13     8 AA              -2.51          1378.
14     8 DL               1.31          1352.
15     8 UA               3.63          1722.
16    12 AA               7.52          1412.
17    12 DL               6.22          1324.
18    12 UA              14.0           1655.
```

```r
p1e <- ggplot(df1e, aes(x = month, y = mean_arr_delay, shape = carrier, color = mean_dista
    geom_point() +
    labs(title = "Mean Arrival Delay by Month and Carrier", x = "Month", y = "Mean Arrival
    scale_color_continuous(name = "Mean Distance") +
    geom_text(aes(label = carrier), nudge_x = 0.5) +
    theme(plot.title = element_text(hjust = 0.5))

p1e
```



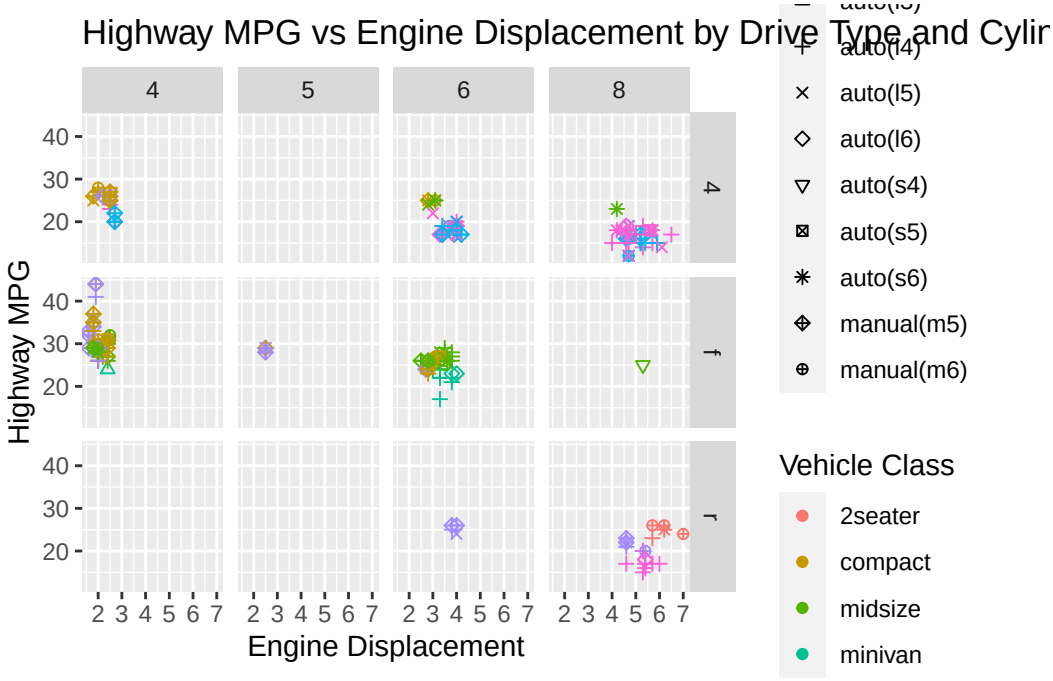
Problem 2

Refer to the `mpg` dataset. Plot `displ` against `hwy` with faceting by `drv` and `cyl`, color designated by `class`, and `shape` by `trans`

```
mpg_sml <- mpg %>% select(displ, hwy, drv, cyl, class, trans)


p2 <- ggplot(mpg_sml, aes(x = displ, y = hwy, color = class, shape = trans)) +
    geom_point() +
    facet_grid(drv ~ cyl) +
    labs(title = "Highway MPG vs Engine Displacement by Drive Type and Cylinders", x = "En
    scale_color_discrete(name = "Vehicle Class") +
    scale_shape_manual(values = 1:length(unique(mpg_sml$trans)), name = "Transmission Type

p2
```


Highway MPG vs Engine Displacement by Drive Type and Cylir
4
5
6
8
4
f
r
Highway MPG
40
30
20
2 3 4 5 6 7
Engine Displacement
auto(l4)
auto(l5)
auto(l6)
auto(s4)
auto(s5)
auto(s6)
manual(m5)
manual(m6)
Vehicle Class
2seater
compact
midsize
minivan