Stat 437 HW1

Nam Jun Lee (11606459)

General rule

Please show your work and submit your computer codes in order to get points. Providing correct answers without supporting details does not receive full credits. This HW covers:

- The basics of `dplyr`
- Creating scatter plot using `ggplot2`
- Elementary Visualizations (via ggplot2): density plot, histogram, boxplot, barplot, pie chart
- Advanced Visualizations via ggplot2: faceting, annotation

You DO NOT have to submit your HW answers using typesetting software. However, your answers must be legible for grading. Please upload your answers to the course space.

Problem 1

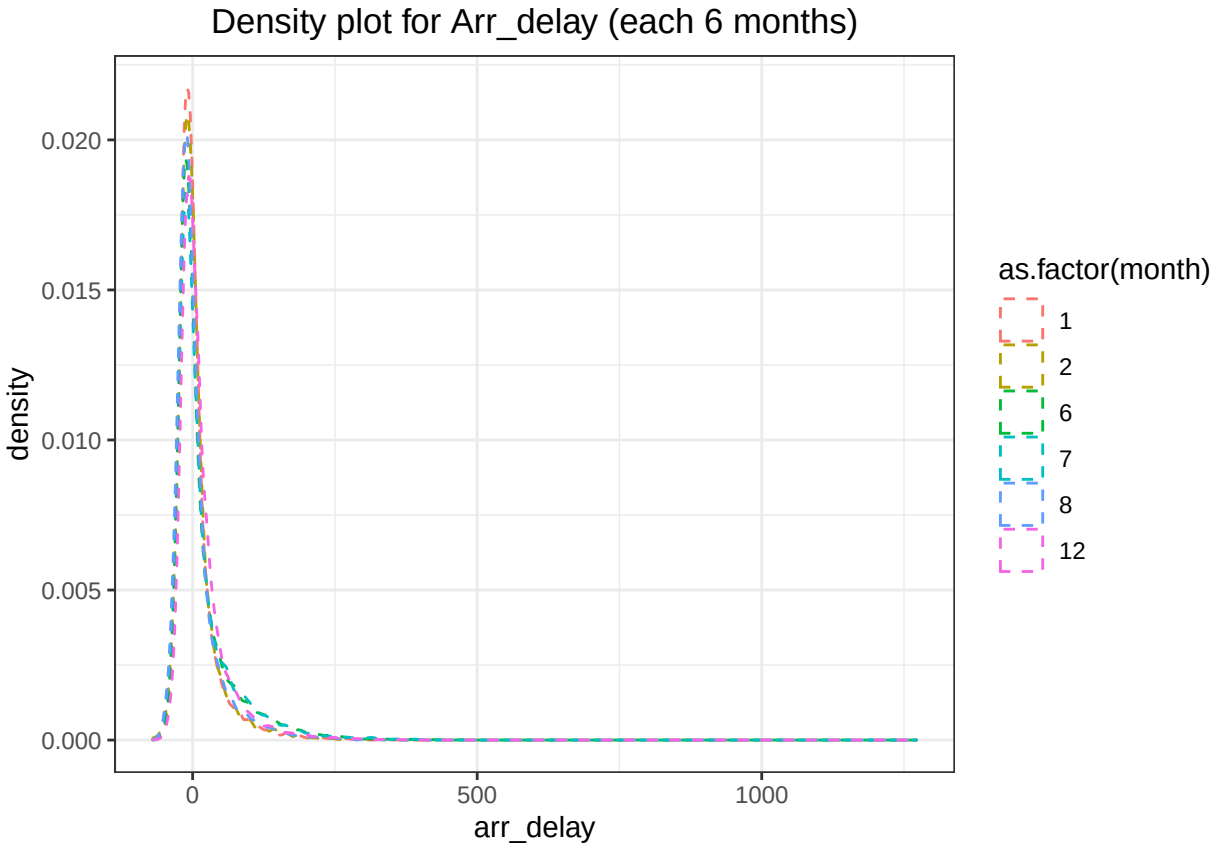
Please refer to the NYC flight data `nycflights13` that has been discussed in the lecture notes and whose manual can be found at https://cran.r-project.org/web/packages/nycflights13/index.html. We will use `flights`, a tibble from `nycflights13`.

You are interested in looking into the average `arr_delay` for 6 different `month` 12, 1, 2, 6, 7 and 8, for 3 different `carrier` "UA", "AA" and "DL", and for `distance` that are greater than 700 miles, since you suspect that colder months and longer distances may result in longer average arrival delays. Note that you need to extract observations from `flights` and obtain the needed sample means for `arr_delay`, and that you are required to use `dplyr` for this purpose.

The following tasks and questions are based on the extracted observations.

(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` into a factor in order to create the plot. What can you say about the average `arr_delay` across the 6 months?

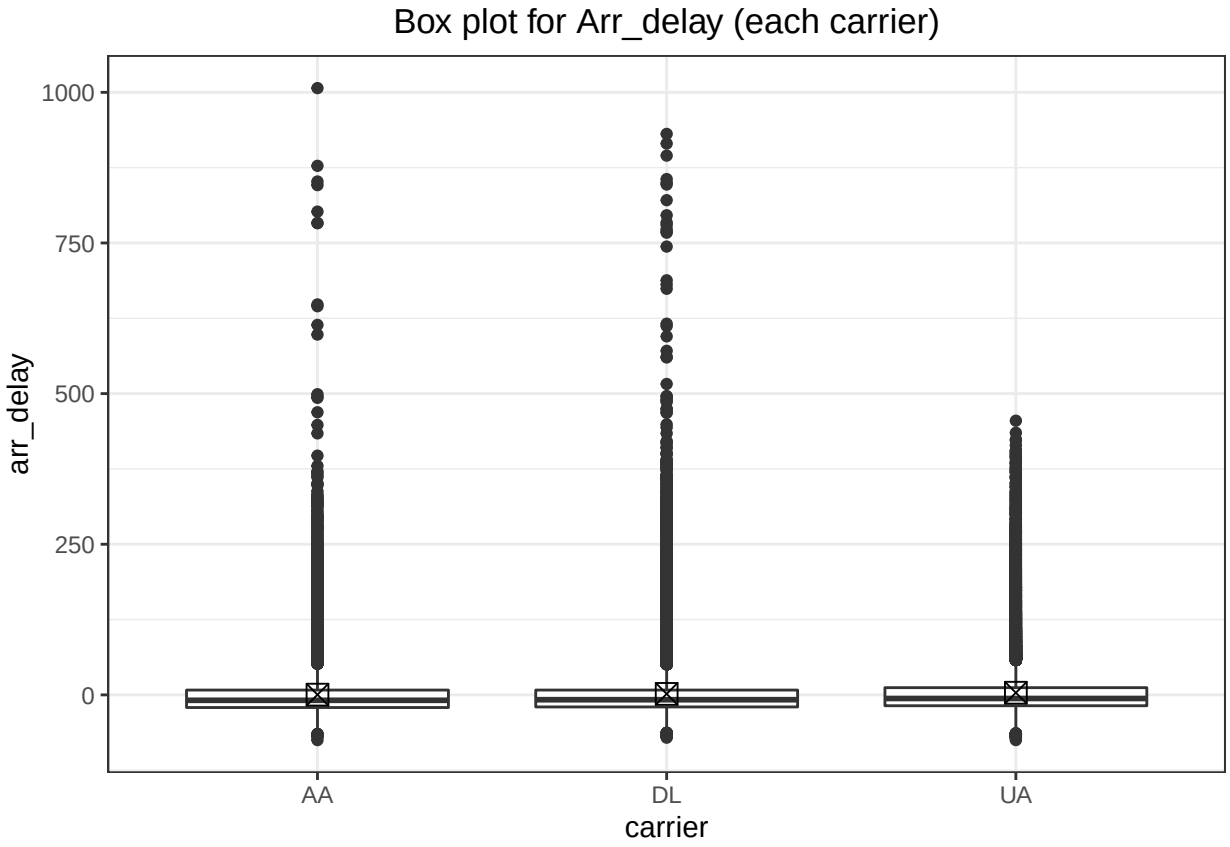
```
# select rows from flights, for which month is
# 12,1,2,6,7,8.
a = flights %>%
    filter(month %in% c(12, 1, 2, 6, 7, 8))
# remove rows that have any NA
a = na.omit(a)
# density plot
a1 = ggplot(a, aes(x = arr_delay, color = as.factor(month))) +
    geom_density(linetype = "dashed") + theme_bw() + theme(plot.title = element_text(hjust = 0
    ggtitle("Density plot for Arr_delay (each 6 months)")
a1
```



When checking the average arr_delay for 6 months, it can be seen that all have the highest density at **zero**.

(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 the 3 carriers?

```
# select rows from flights, for which carrier is UA, AA,
# DL.
b = flights %>%
    filter(carrier %in% c("UA", "AA", "DL"))
# remove rows that have any NA
b = na.omit(b)
# box plot (find mean using stat_summary)
b1 = ggplot(b, aes(x = carrier, y = arr_delay)) + geom_boxplot() +
    theme_bw() + stat_summary(fun = mean, geom = "point", shape = 7,
    size = 3.5) + ggtitle("Box plot for Arr_delay (each carrier)") +
    theme(plot.title = element_text(hjust = 0.5))
b1
```



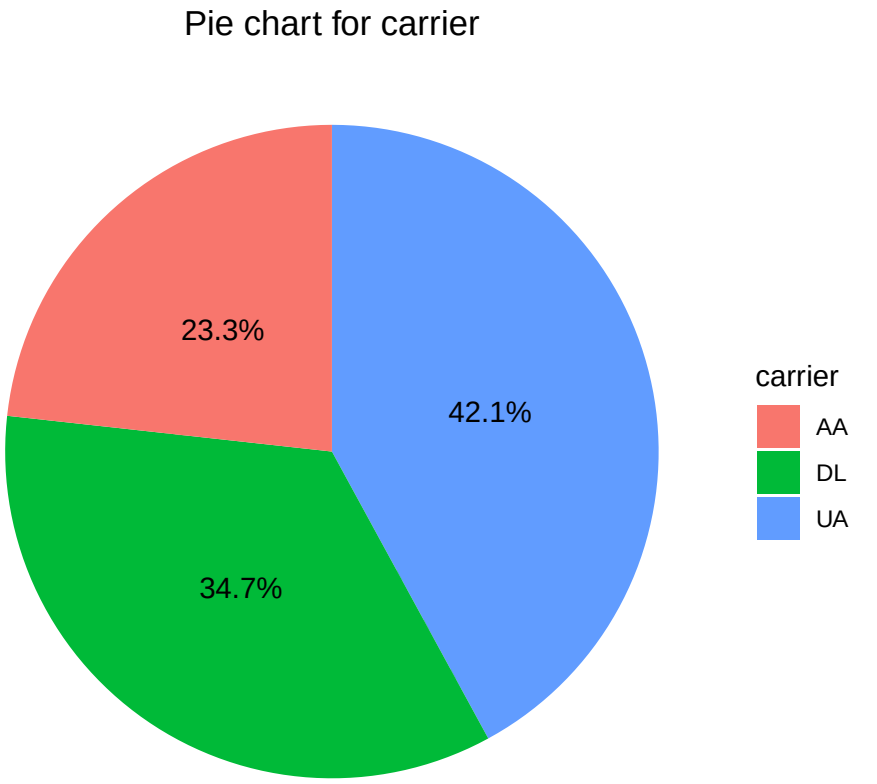
As a result of checking the box plot for 'arr_delay' for each of the three carriers, it can be seen that the average of the three carriers is **zero** and that there are many outliers for each of the three carriers.

(1.c) Create a pie chart for the 3 carriers where the percentages are the proportions of observations for each carrier and where percentages are superimposed on the sectors of the pie chart disc.

```
# group data via carrier, obtain counts for each carrier
# and compute percentages from counts
c = b %>%
    group_by(carrier) %>%
    count() %>%
    ungroup() %>%
    mutate(percentage = n/sum(n)) %>%
    arrange(desc(carrier))
# creates labels using the percentage
c$labels <- scales::percent(c$percentage)
c
```

```
## # A tibble: 3 x 4
##   carrier     n percentage labels
##   <chr>   <int>      <dbl> <chr>
## 1 UA      57782      0.421 42.1%
## 2 DL      47658      0.347 34.7%
## 3 AA      31947      0.233 23.3%
```

```
# pie chart
c1 = ggplot(c) + geom_bar(aes(x = "", y = percentage, fill = carrier),
    stat = "identity", width = 1) + coord_polar("y", start = 0) +
    theme_void() + geom_text(aes(x = 1, y = cumsum(percentage) -
    percentage/2, label = labels)) + ggtitle("Pie chart for carrier") +
    theme(plot.title = element_text(hjust = 0.5))
c1
```

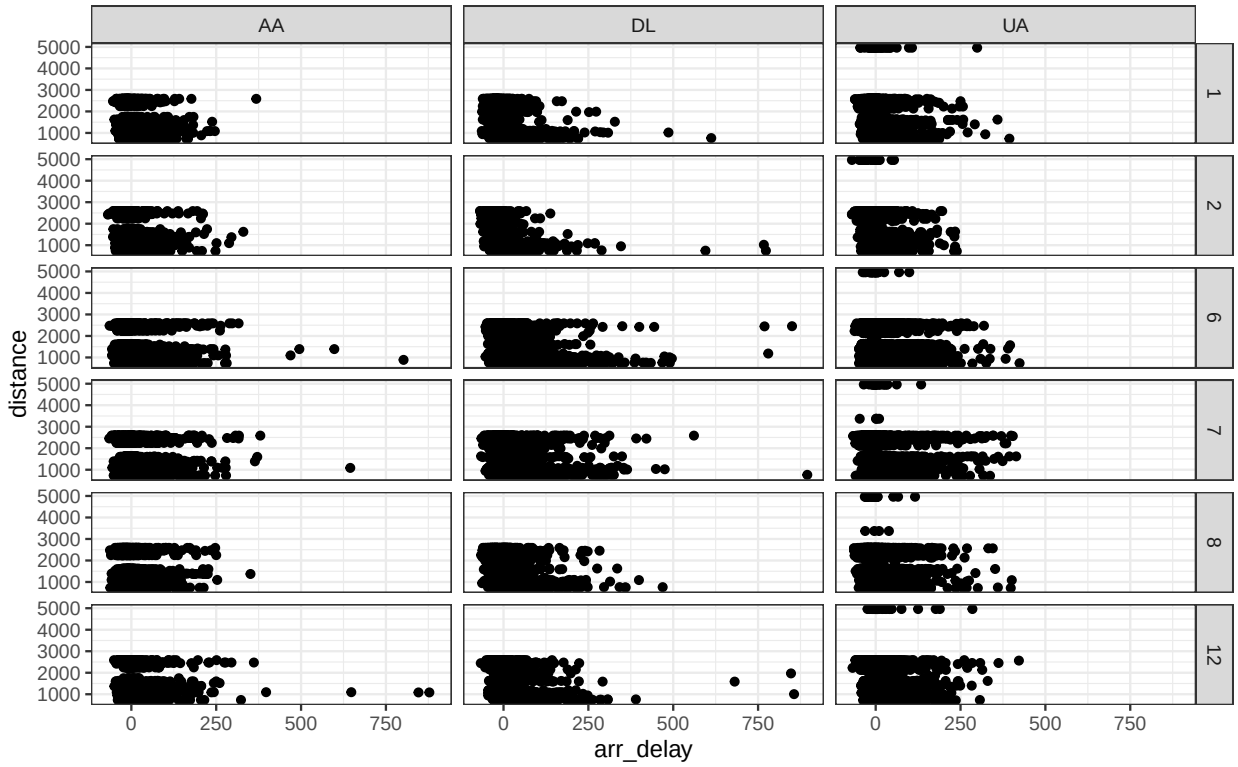


As a result of checking the proportion:

$$UA > DL > AA$$

(1.d) Plot `arr_delay` against `distance` with `facet_grid` designated by `month` and `carrier`.

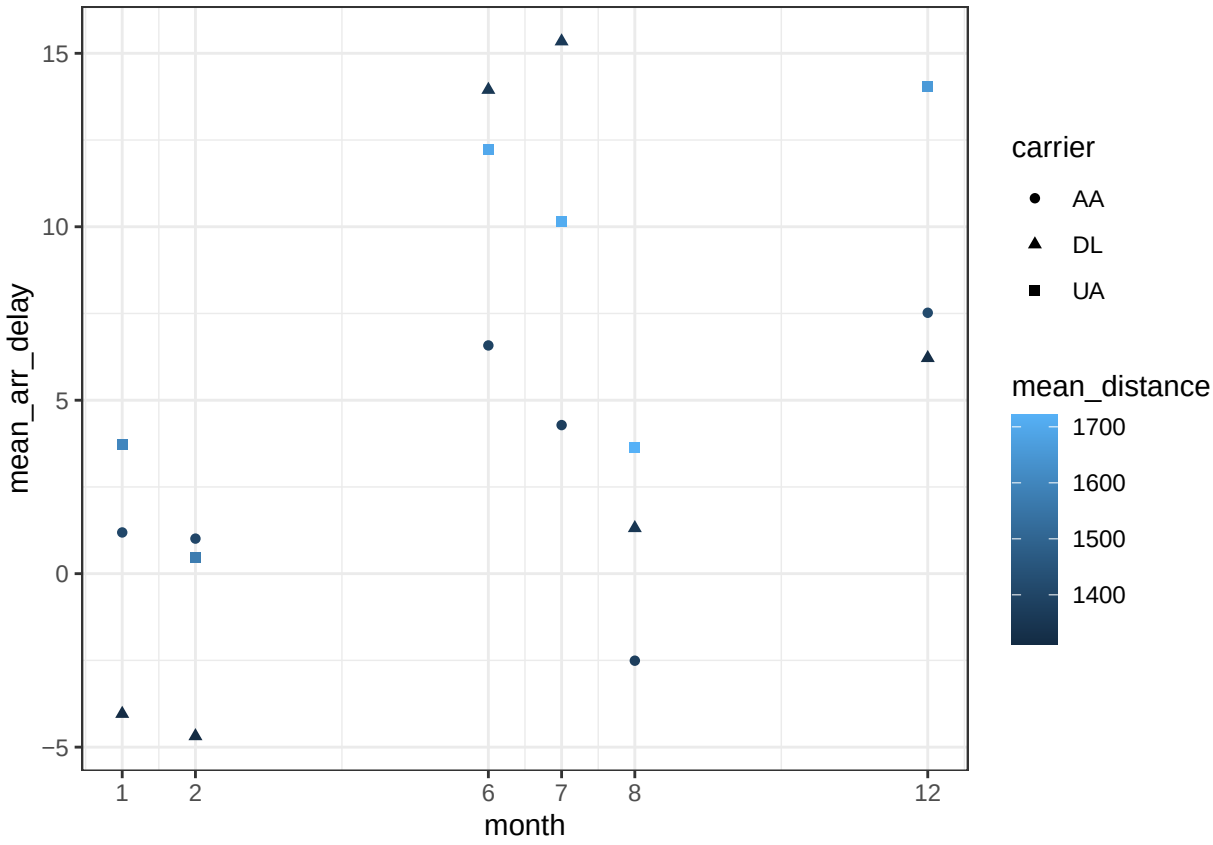
```
# select rows from flights, for which month is
# 12,1,2,6,7,8, and carrier is UA, AA, DL, and distance
# higher than 700.
d = flights %>%
    filter(month %in% c(12, 1, 2, 6, 7, 8), carrier %in% c("UA",
        "AA", "DL"), distance > 700)
# remove rows that have any NA
d = na.omit(d)
# show plot
d1 = ggplot(data = d) + theme_bw() + geom_point(mapping = aes(x = arr_delay,
    y = distance)) + facet_grid(month ~ carrier)
d1
```



As a result of checking the graph, it was found that the UA traveled the longest distance and the departure delay time did not exceed 500. Conversely, it can be seen that AA and DL have several departure delays of more than 500.

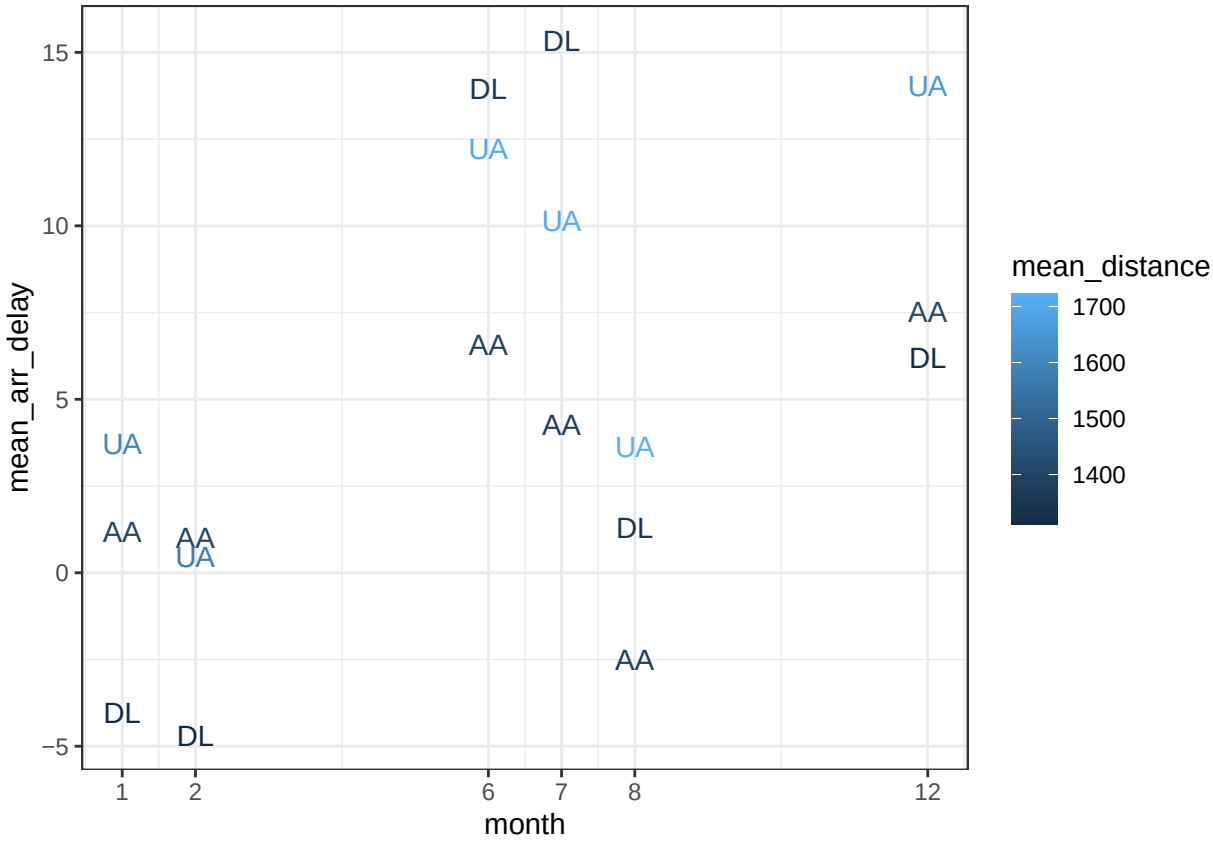
(1.e) For each feasible combination of values of `month` and `carrier`, compute the sample average of `arr_delay` and save them into the variable `mean_arr_delay`, and compute the sample average of `distance` and save these averages into the variable `mean_distance`. Plot `month` against `mean_arr_delay` with `shape` designated by `carrier` and `color` by `mean_distance`, and 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.

```
# compute the sample average of arr_delay and sample
# average of distance
e = d %>%
    group_by(month, carrier) %>%
    summarise(mean_arr_delay = mean(arr_delay), mean_distance = mean(distance),
        .groups = "keep") %>%
    as.data.frame()
# plot 1
e1 = ggplot(e, aes(x = month, y = mean_arr_delay)) + theme_bw() +
    geom_point(aes(shape = carrier, colour = mean_distance)) +
    scale_x_continuous(breaks = c(1, 2, 6, 7, 8, 12))
e1
```




```
# plot 2 (annotate)
e2 = ggplot(e, aes(x = month, y = mean_arr_delay)) + theme_bw() +
    geom_text(aes(label = carrier, colour = mean_distance)) +
```

```
    scale_x_continuous(breaks = c(1, 2, 6, 7, 8, 12))
e2
```

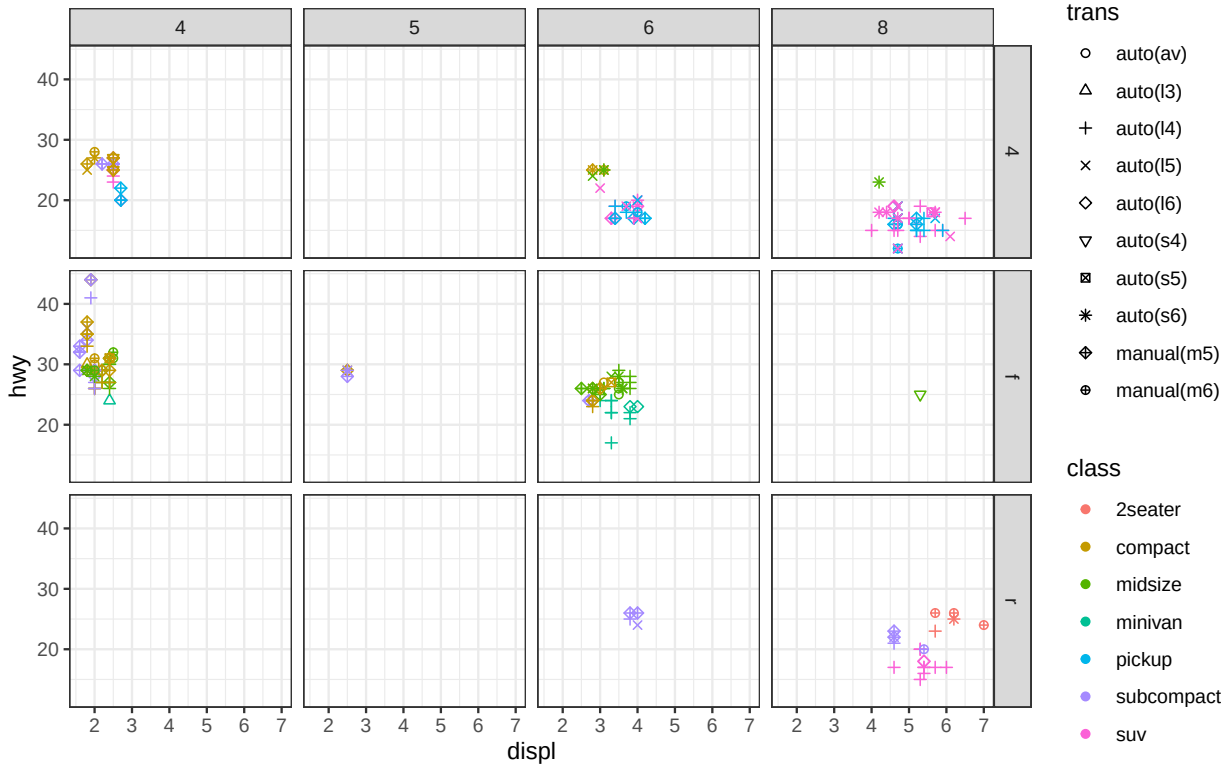


As a result of checking the two graphs, it can be seen that the average distance of UA is the longest, and DL and AA are almost similar. In addition, it can be seen that the average arrival delay of the three carriers was prolonged in June, July, and December. Conversely, considering that the average arrival delay in January, February, and August is less than in other months, **it is not reasonable that the colder the cold, the longer the average arrival delay may be**.

Problem 2

Please refer to the data set `mpg` that is available from the `ggplot2` package. Plot `displ` against `hwy` with faceting by `drv` and `cyl`, `color` disgnated by `class`, and `shape` by `trans`. This illustrates visualization with 4 factors.

```
# levels in mpg aftering converting character into factors
mpg1 = mpg %>%
    dplyr::mutate_if(is.character, as.factor)
# show plot (more than 6 class level so using
# scale_shape_manual)
m = ggplot(mpg1, aes(x = displ, y = hwy)) + theme_bw() + geom_point(aes(colour = class,
    shape = trans)) + scale_shape_manual(values = 1:length(unique(mpg$trans))) +
    facet_grid(drv ~ cyl)
m
```



As a result of checking the graph, there are no elements where drv is 4, cyl is 5, drv is r, and cyl is 4, and 5. In addition, **it can be seen that the higher the hwy, the lower the displ**.