Stat 437 HW1

Yu-Tung(Jenny), Cheng (11678647)

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

For an assignment or project, you DO NOT have to submit your answers or reports using typesetting software. However, your answers must be well organized and well legible for grading. Please upload your answers in a document to the course space. Specifically, if you are not able to knit a .Rmd/.rmd file into an output file such as a .pdf, .doc, .docx or .html file that contains your codes, outputs from your codes, your interpretations on the outputs, and your answers in text (possibly with math expressions), please organize your codes, their outputs and your answers in a document in the format given below:

```
Problem or task or question ...
Codes ...
Outputs ...
Your interpretations ...
```

It is absolutely not OK to just submit your codes only. This will result in a considerable loss of points on your assignments or projects.

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 (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 requird to use `dplyr` for this purpose.

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

##install and set up R markdown

```
library(knitr)
```

(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?

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(nycflights13)
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.8     v purrr   0.3.4
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
head(flights)
```

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	▸
<int>	<int>	<int>	<int>	<int>	<dbl>	<int>	
2013	1	1	517	515	2	830	
2013	1	1	533	529	4	850	
2013	1	1	542	540	2	923	
2013	1	1	544	545	-1	1004	
2013	1	1	554	600	-6	812	
2013	1	1	554	558	-4	740	

6 rows | 1-7 of 19 columns

```
#create a small data frame 'temp' from flights and select 6 months and 3 carriers from 'flights'
temp = flights %>% select(month, arr_delay, carrier, distance) %>% filter(month %in% c(12,1,2,6,
7,8), carrier %in% c("UA","AA","DL"), distance > 700)

#remove rows having any NA
temp = na.omit(temp)

#obtain mean by each combination of month value and carrier
sm = temp %>% group_by(month, carrier) %>% summarise( mean_arr_delay = mean(arr_delay)) %>% as.d
ata.frame()
```

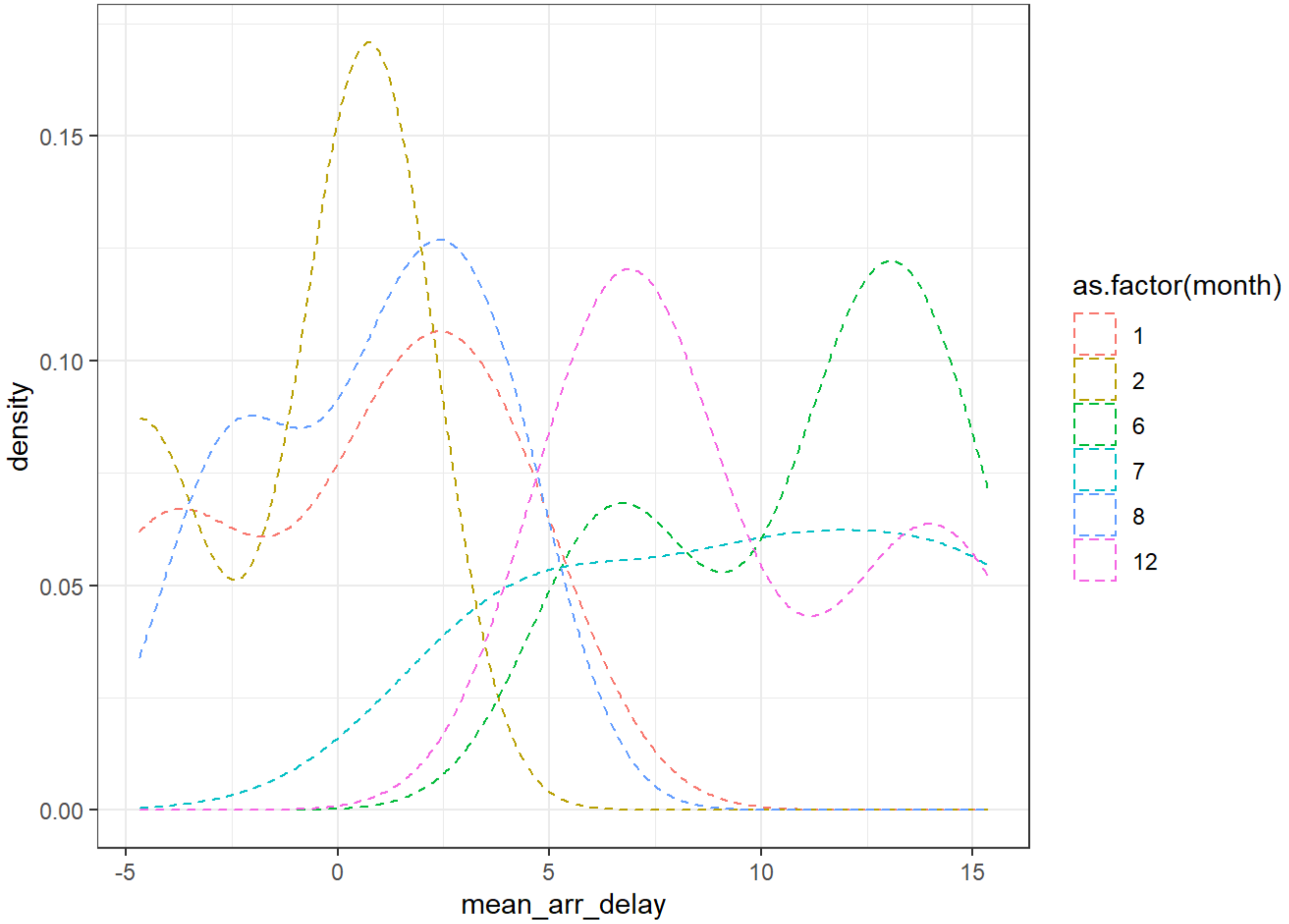
```
## `summarise()` has grouped output by 'month'. You can override using the
## `.groups` argument.
```

```
head(sm)
```

	month <int>	carrier <chr>	mean_arr_delay <dbl>
1	1	AA	1.1883167
2	1	DL	-4.0404784
3	1	UA	3.7168572
4	2	AA	1.0104530
5	2	DL	-4.6836803
6	2	UA	0.4696726

6 rows

```
#apply for the density plot with the select rows of arr_delay for each of the 6 months with colo
r by month as factor
p2 = ggplot(sm, aes(x=mean_arr_delay, color = as.factor(month))) + geom_density(linetype = "dash
ed") + theme_bw()

print(p2)
```



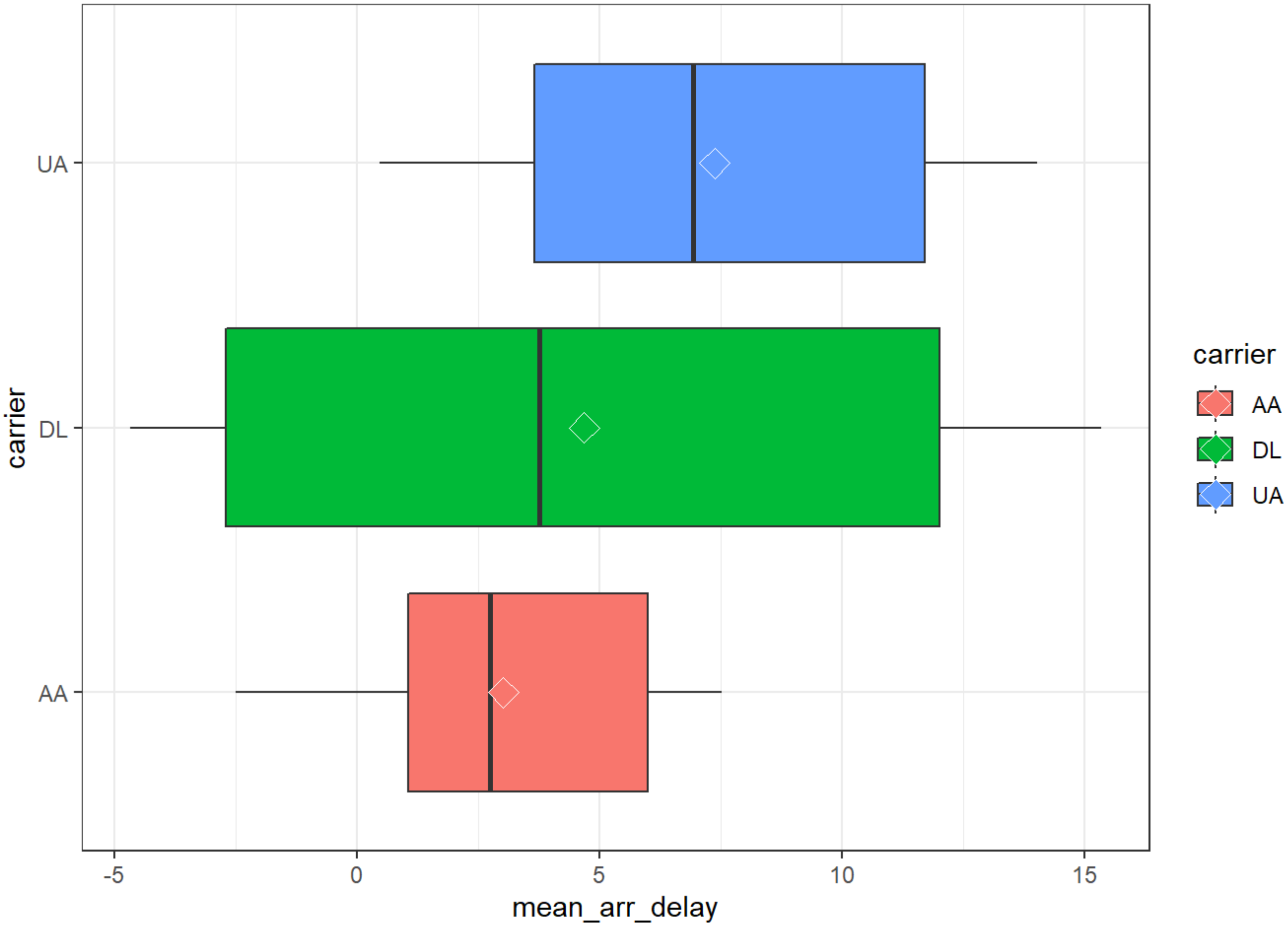
```
  ##Interpretation: As the density plot with these 6 months as factors for computing the average
arrival of delay, we can observe that February relatively stands the least frequency of arrival
 delay in 2013. The brown dashed line which is February with the positively skewed right distrib
ution mostly gathered on the left which its mode is smaller than mean.
```

(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?

```
#create a boxplot

p3= ggplot(sm, aes(x=mean_arr_delay, y= carrier, fill = carrier)) + geom_boxplot() + theme_bw()+
stat_summary(fun.y=mean,geom="point",shape=23, size=4, col = "white")
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```

```
print(p3)
```



```
 ##Interpretation: The median of AA carrier is relatively small on the left-most point which is
 the less frequency of arrival delay as the box plot with x-axis of  arrival delay. However, as
 the scatter degree with its median of AA is more concentrated from the left side which means it
s data distribution is left-skewed. For DL carrier, there is a huge range  between the upper qua
rtile and the lower quartile compared with AA and UA, and it shows the arrival delay of DL is un
stable. As for UA, the upper quartile, lower quartile, and the median are bigger than others sin
ce the arrival delay of UA is more than others.
```

(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.

```
library(scales)
```

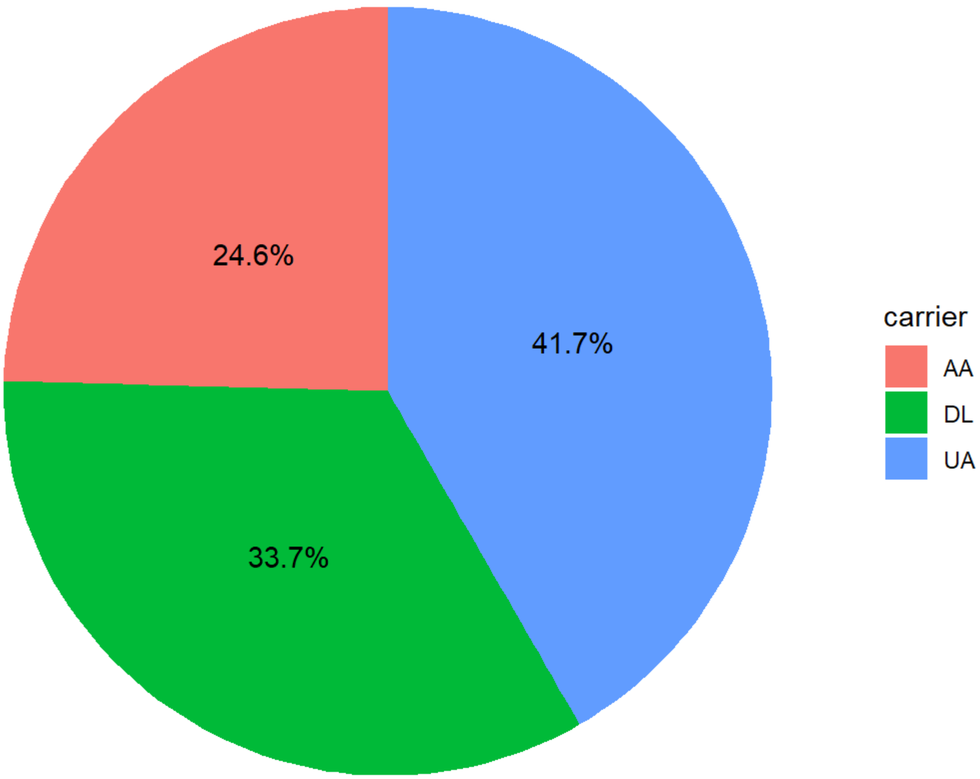
```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
sm2 = temp %>% group_by(carrier) %>% count() %>% ungroup() %>% mutate(percentage=`n`/sum(`n`))
%>% arrange(desc(carrier))

sm2$labels <- scales::percent(sm2$percentage)

print(sm2)
```

```
## # A tibble: 3 x 4
##   carrier     n percentage labels
##   <chr>   <int>      <dbl> <chr>
## 1 UA      25921      0.417 41.7%
## 2 DL      20982      0.337 33.7%
## 3 AA      15285      0.246 24.6%
```

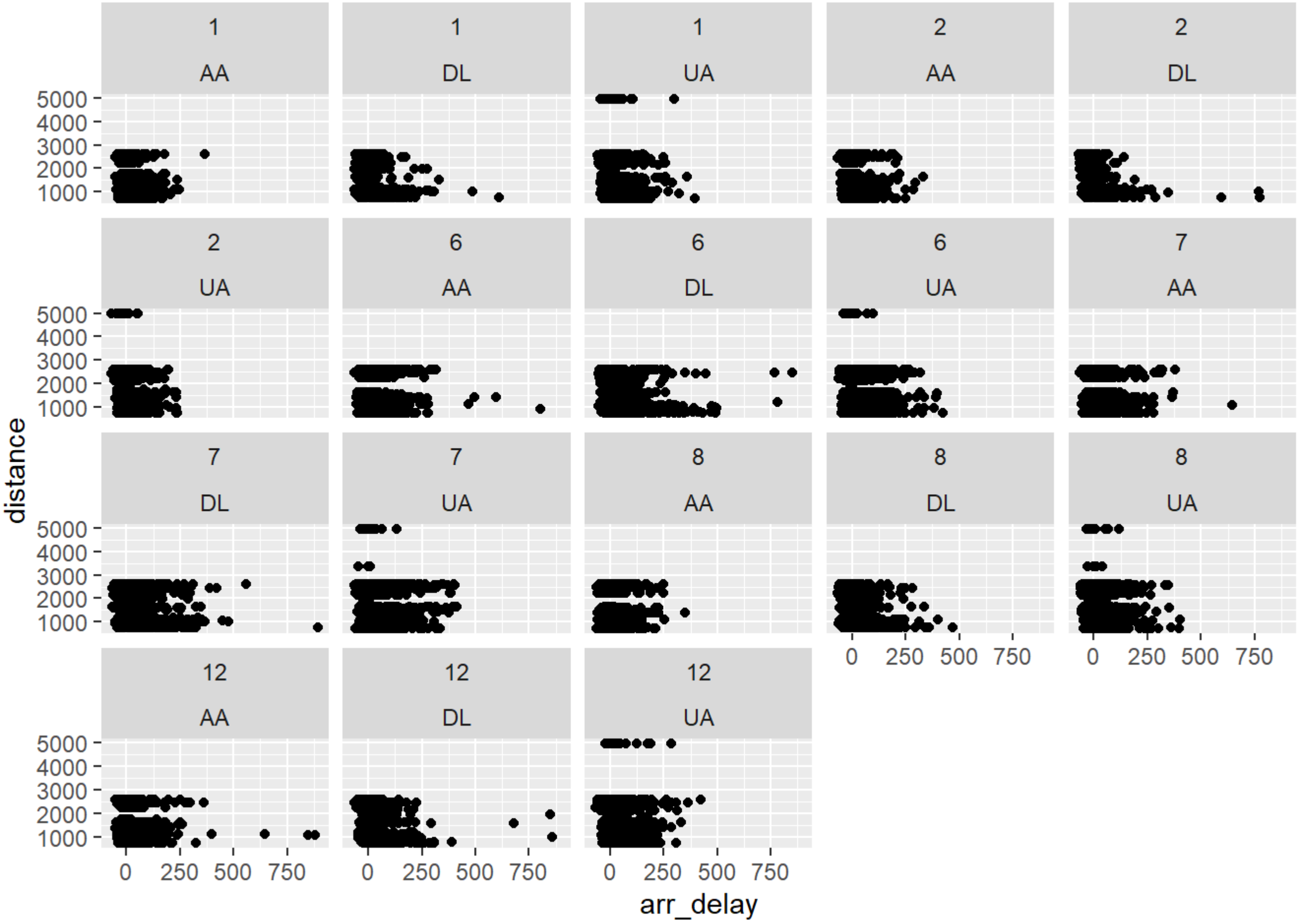
```
pie = ggplot(sm2) + geom_bar(aes(x="", y =percentage, fill=carrier), stat="identity", width = 1)
+coord_polar("y", start = 0) + theme_void() + geom_text(aes(x=1, y= cumsum(percentage)-percentag
e/2, label=labels))

print(pie)
```



```
##Interpretation: The pie chart expresses that UA has 41.7%, DL has 33.7%, and AA has 24.6% of t
heir observations.
```

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

```
plot1 = ggplot(data = temp) + geom_point(mapping = aes(x = arr_delay, y = distance))

plot2 = plot1 + facet_wrap(month~carrier, nrow = 4)

print(plot2)
```



```
##Interpretation: Only UA carrier has the records of arriving delay with the long-distance fligh
ts.
```

(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.

```
#obtain mean by each combination of month value and carrier
mean_arr_delay = temp %>% group_by(month, carrier) %>% summarise( mean_arr_delay = mean(arr_dela
y)) %>% as.data.frame()
```

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

```
print(mean_arr_delay)
```

```
##    month carrier mean_arr_delay
## 1      1      AA      1.1883167
## 2      1      DL     -4.0404784
## 3      1      UA      3.7168572
## 4      2      AA      1.0104530
## 5      2      DL     -4.6836803
## 6      2      UA      0.4696726
## 7      6      AA      6.5790090
## 8      6      DL     13.9477945
## 9      6      UA     12.2224950
## 10     7      AA      4.2829618
## 11     7      DL     15.3451183
## 12     7      UA     10.1627702
## 13     8      AA     -2.5087199
## 14     8      DL      1.3142706
## 15     8      UA      3.6292620
## 16    12      AA      7.5205479
## 17    12      DL      6.2163475
## 18    12      UA     14.0423901
```

```
mean_distance = temp %>% group_by(month, carrier) %>% summarise( mean_distance = mean(distance))
%>% as.data.frame()
```

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

```
print(mean_distance)
```

```
##    month carrier mean_distance
## 1      1      AA      1407.774
## 2      1      DL      1317.234
## 3      1      UA      1601.036
## 4      2      AA      1404.979
## 5      2      DL      1312.187
## 6      2      UA      1570.462
## 7      6      AA      1392.524
## 8      6      DL      1356.091
## 9      6      UA      1698.521
## 10     7      AA      1381.290
## 11     7      DL      1362.119
## 12     7      UA      1708.035
## 13     8      AA      1378.601
## 14     8      DL      1352.149
## 15     8      UA      1722.198
## 16    12      AA      1415.697
## 17    12      DL      1323.626
## 18    12      UA      1657.657
```

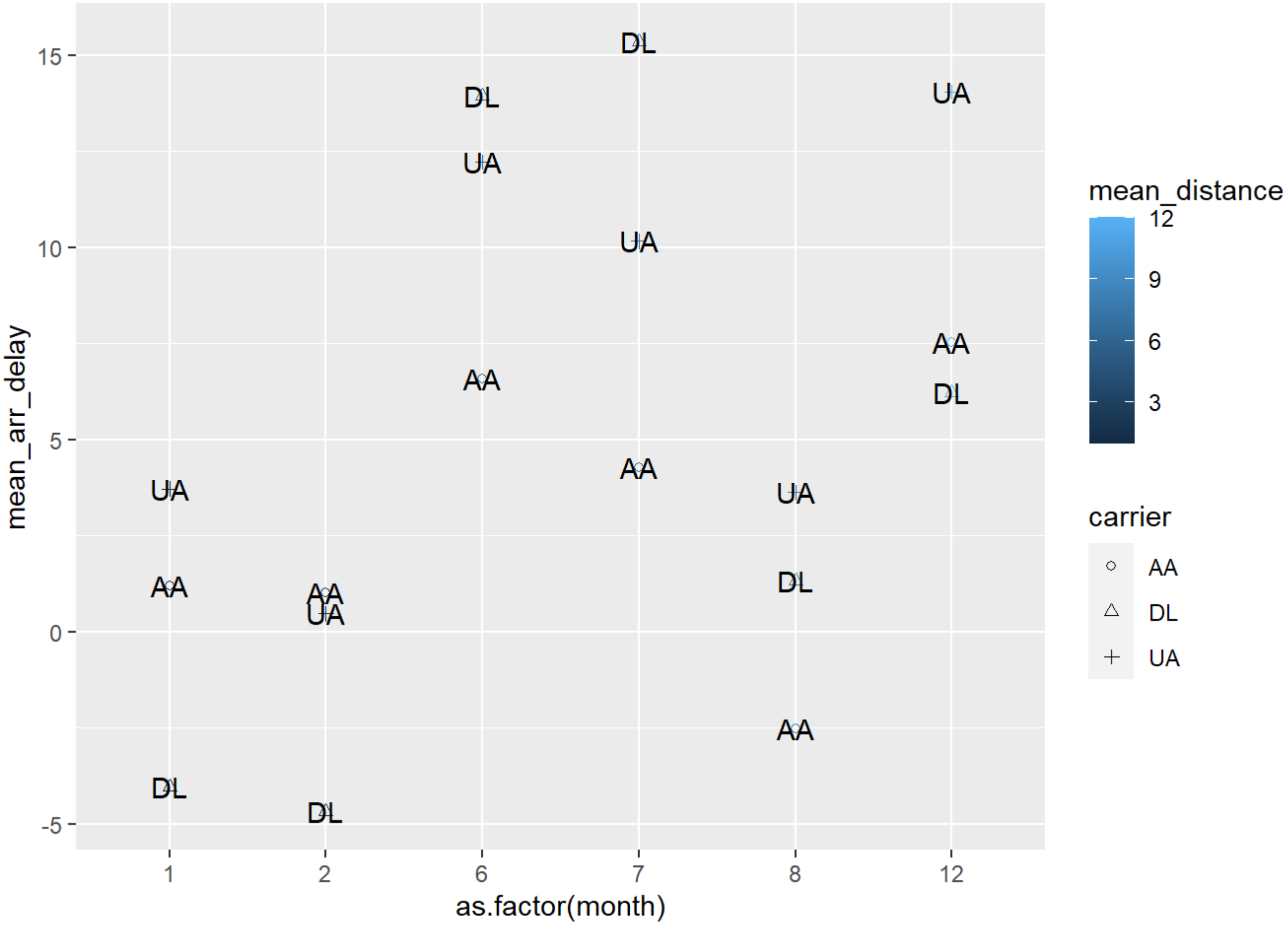
```
sm['mean_distance'] <- c(mean_distance)
```

```
## Warning in `[<-.data.frame`(`*tmp*`, "mean_distance", value = list(month =
## c(1L, : provided 3 variables to replace 1 variables
```

```
print(sm)
```

```
##    month carrier mean_arr_delay mean_distance
## 1      1      AA      1.1883167             1
## 2      1      DL     -4.0404784             1
## 3      1      UA      3.7168572             1
## 4      2      AA      1.0104530             2
## 5      2      DL     -4.6836803             2
## 6      2      UA      0.4696726             2
## 7      6      AA      6.5790090             6
## 8      6      DL     13.9477945             6
## 9      6      UA     12.2224950             6
## 10     7      AA      4.2829618             7
## 11     7      DL     15.3451183             7
## 12     7      UA     10.1627702             7
## 13     8      AA     -2.5087199             8
## 14     8      DL      1.3142706             8
## 15     8      UA      3.6292620             8
## 16    12      AA      7.5205479            12
## 17    12      DL      6.2163475            12
## 18    12      UA     14.0423901            12
```

```
#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
pp3 = ggplot(sm, aes(x=as.factor(month), y = mean_arr_delay)) + geom_point(aes(as.factor(month),
mean_arr_delay, colour = mean_distance, shape = carrier)) + geom_text(data=sm, aes(x=as.factor(m
onth) , y=mean_arr_delay, label=carrier)) + scale_fill_distiller(palette = "YlOrBr") + scale_sha
pe_manual(values = 1:length(unique(sm$carrier)) )

print(pp3)
```

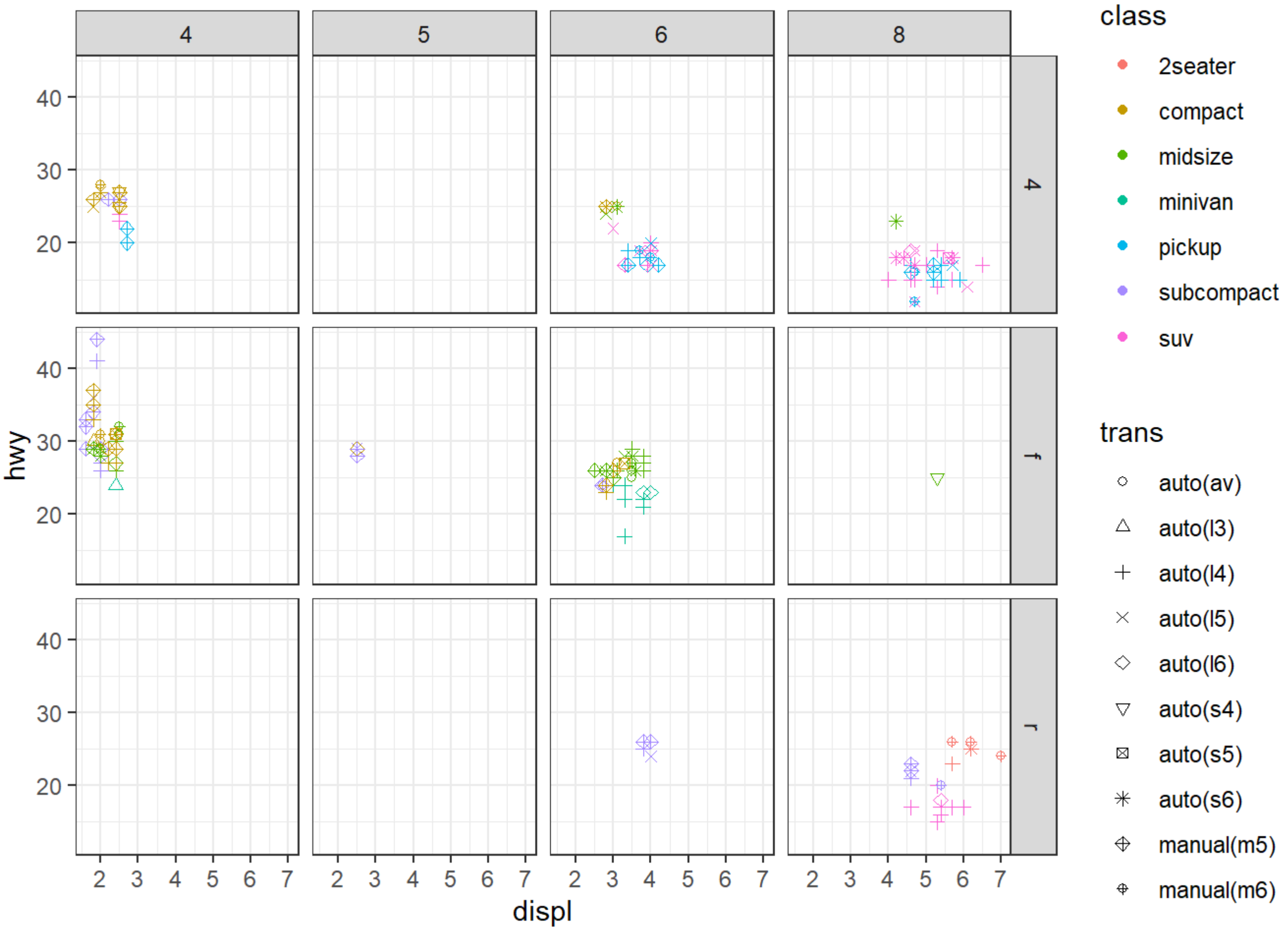


```
##Interpretation: DL is unstable and it delays pretty often in July.
```

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.

```
#visualization with >= 3 factors

p1c = ggplot(mpg, aes(x = displ, y = hwy)) + theme_bw() + geom_point(aes(color=class, shape=trans)) + facet_grid(drv~cyl) + scale_shape_manual(values=c(1:10))

print(p1c)
```



```
##Interpretation: Larger cars has more cylinders and higher engine displacement in litres, and the smaller cars requires less.
```