STAT 362 R for Data Science

Assignment 3

Please follow the general instructions as in Assignment 1.

Due: Feb 24, 2023 (11:59pm)

Using the following code to load the required packages:

```
library(tidyverse)
library(nycflights13)
library(gcookbook)
library(MASS)
```

Q1(a): Use `filter()` and `ggplot()` to create a histogram of the departure delay of the flights that departed in Feb. Use the choice `binwidth = 10` in `geom_histogram()`.

Note: include the flights with negative values in departure delay.

Q1(b): Use `filter()` and `ggplot()` to create a histogram of the departure delay of the flights that departed in Feb with departure delay less than 100 minutes. Use the choice `binwidth = 5` in `geom_histogram()`.

Note: include the flights with negative values in departure delay.

Q2(a): Use `filter()` and `ggplot()` to create a scatterplot of `arr_delay` (y-axis) versus `dep_delay` (x-axis) using the flights that departed on Jan 1 with departure delay strictly less than 3 hours.

Q2(b) Use `filter()` and `ggplot()` to create a scatterplot of `arr_delay` versus `dep_delay` using the flights that departed on Jan 1 with departure delay strictly less than 10 minutes.

Q2(c): Which graph in (a) and (b) shows a more apparent trend between the two delay times? Answer this as a comment in R.

Q2(d): Compute the correlation (use `cor`) for the points that you plot in (a) and (b). Do the results make sense in view of your observation in (c)?

Hint: when there are missing values (`NA`) in the data, you can use `cor(x, y, use = "complete.obs")` to compute the correlation of `x` and `y`. This will use only the complete observations for the calculation. Otherwise, you will get `NA` as the output.

Example:

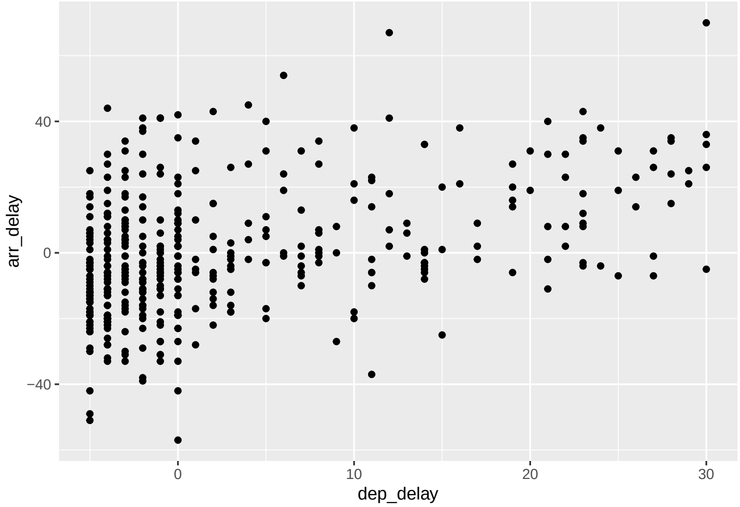
```
x <- c(1:5, NA)
y <- c(NA, 1, 2, 3, 6, 6)
cor(x, y) # NA
## [1] NA

cor(x, y, use = "complete.obs")
## [1] 0.9561829

# the complete cases are (2,1), (3,2), (4,3) and (5,6)
cbind(x, y)
##       x  y
## [1,]  1 NA
## [2,]  2  1
## [3,]  3  2
## [4,]  4  3
## [5,]  5  6
## [6,] NA  6
# compute the correlation for these points
cor(c(2, 3, 4, 5), c(1, 2, 3, 6)) # same as cor(x,y, use="complete.obs")
## [1] 0.9561829
```

Q3: Write a function called `plot_delay` with arguments `which_month`, `which_day`, `lower_range` and `upper_range` to create the scatterplot (using `ggplot()`) of `arr_delay` versus `dep_delay` using the flights that departed in month being equal to `which_month` and day being equal to `which_day` with departure delay between `lower_range` and `upper_range` (inclusively).

For example,

```
plot_delay(which_month = 1, which_day = 30, lower_range = -5, upper_range = 30)
```

should give you the following plot:



Q4(a): Use `ggplot()` and `geom_bar()` to create a bar chart of counts of the number of flights in each month.

Q4(b): Use `ggplot()` and `geom_bar()` to create a bar chart of counts of the number of flights in each day in January.

Q4(c): Use `ggplot()` and `geom_col()` to create a bar chart of values of the average arrival delay in each day in January.

Q4(d): Use `ggplot()` to create a bar chart of values of the average arrival delay in each Saturday in Jan and Feb.

Q5: Create a line graph of the average departure delay in the dataset `flights`. The x-axis is the time, represented by `1:365`. The y-axis is the average departure delay on each day.

Q6: The package `MASS` contains a dataset called `birthwt`. From the `birthwt` dataset, create a plot of a kernel density estimate of the density of the birth weight when the mother's age is greater than or equal to 25 using `ggplot()` with `geom_density()`.

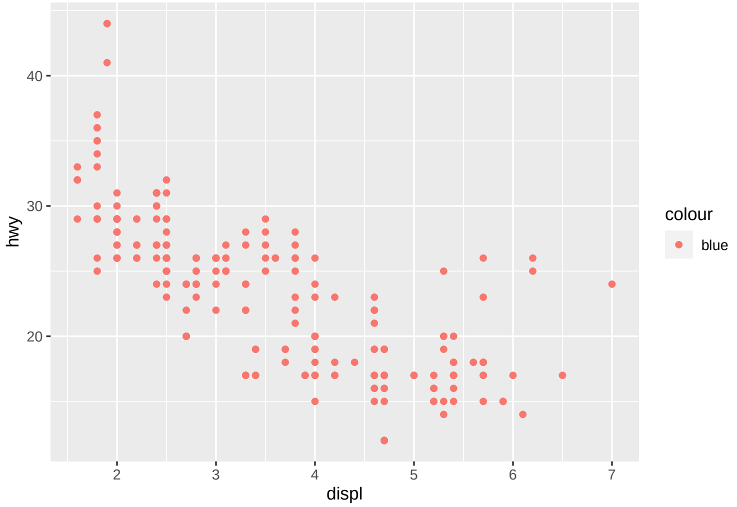
For Q7, consider the dataset `mpg`.

Q7(a): Run `ggplot(data = mpg)`. What do you see?

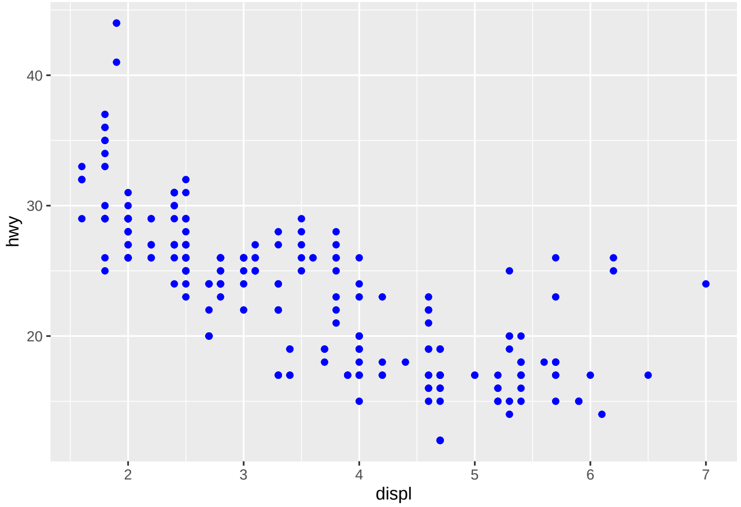
Q7(b): Use `ggplot` to create a scatterplot of `class` vs `drv` (class on the y-axis and `drv` on the `x-axis`). Why is the plot not useful? Any reasonable answer is ok.

Q7(c) What's gone wrong with the following code? Why are the points not blue?

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, color = "blue"))
```

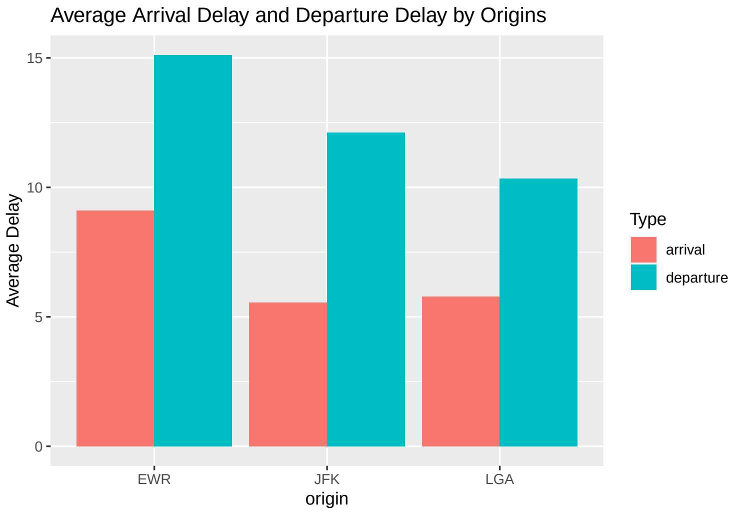


Q7(d): Fix the code to obtain the following plot:



Q8: Consider the `flights` dataset in the package `nycflights13`. Recreate the R code necessary to generate the following graph, which shows the average arrival delay and departure delay by origins.

Hint: use `group_by()` and `summarize()`.

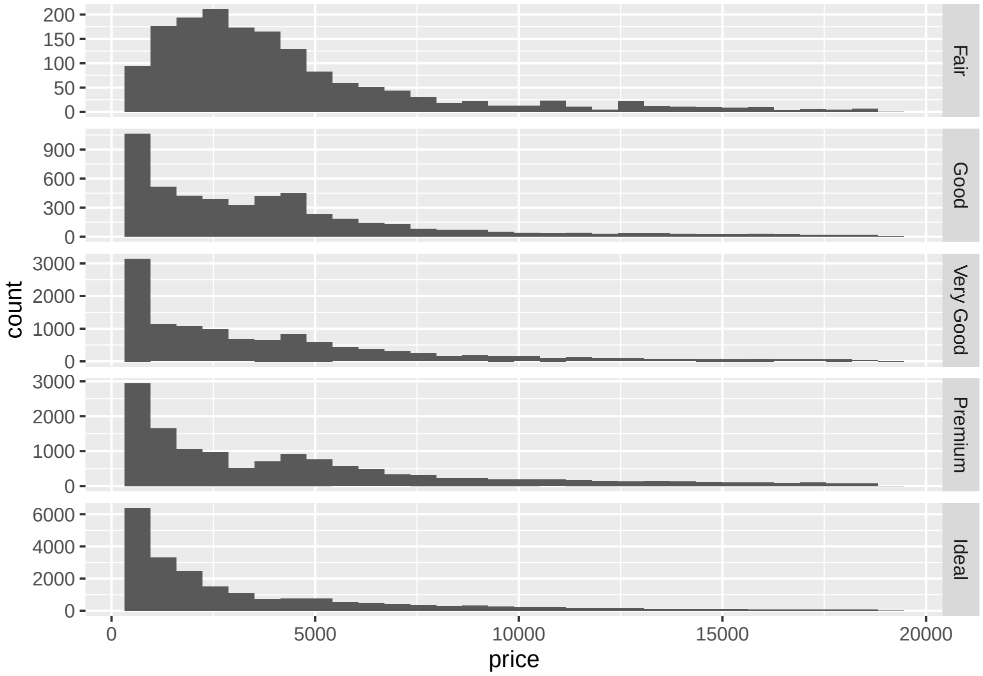


For Q9, consider the `diamonds` dataset in the package `ggplot2`. If you have loaded `tidyverse`, you will have this dataset.

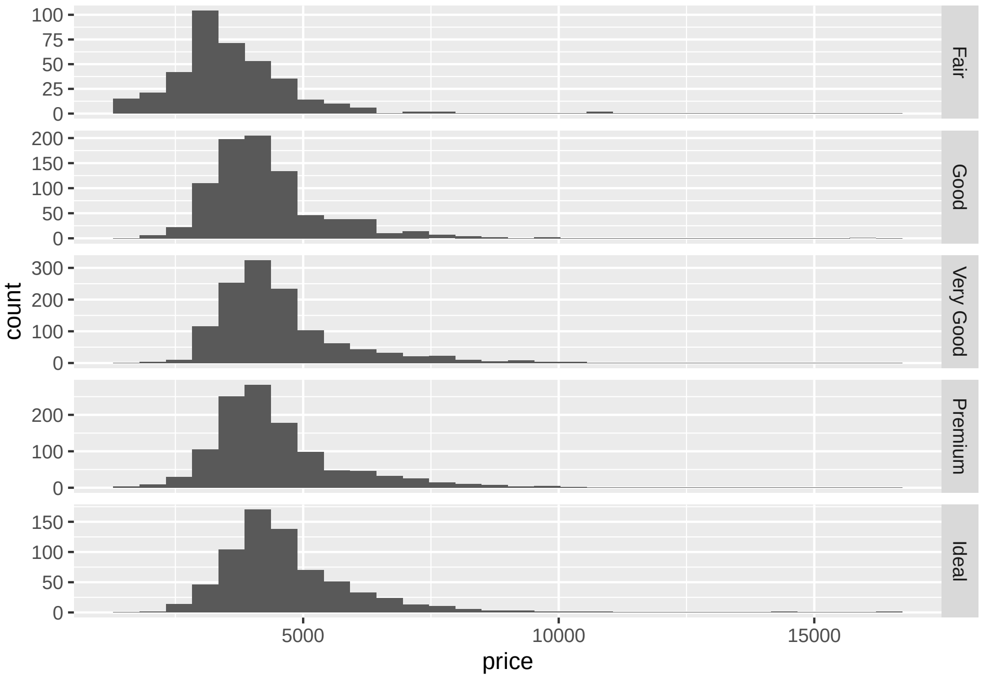
Q9(a): Compute the average price of the diamonds grouped by the quality of the cut. Do you think the results are reasonable? Explain.

Hint: use `?diamonds` to find out which variable is for the quality of the cut.

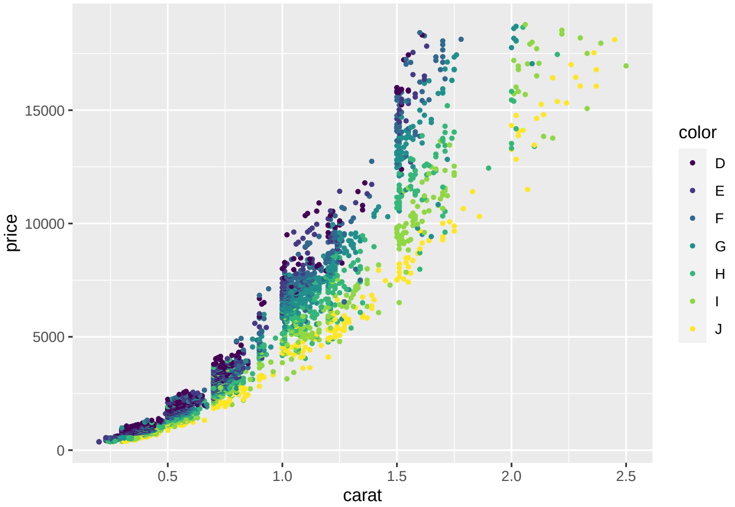
Q9(b): Create histograms of the price of the diamonds by the variable `cut`. Use `facet_grid()` and set `scales = "free_y"` to obtain the following graph (this option sets the scales on the y-axis vary across rows). Do you expect the observation that a fair diamond tends to be more expensive than an ideal diamond? Why or why not?



Q9(c): Recreate the plot in (b) using only diamonds with `carat <= 1` and `carat >= 0.9`. You should obtain the following graph.



Q9(d): Create a scatterplot of `price` vs `carat` using diamonds with `cut` equals `"Ideal"` and `clarity` equals `"VS2"`. Map the colors of the points to the `color` variable in `diamonds` to reveal the diamond color. Set the point size to `0.9`. You should obtain the following plot. Write down two features that you observe (any reasonable answers are ok).



Q10: Before you work on this question, study Section 3.6 in one of our reference books **R for data science** first.
https://r4ds.had.co.nz/data-visualisation.html

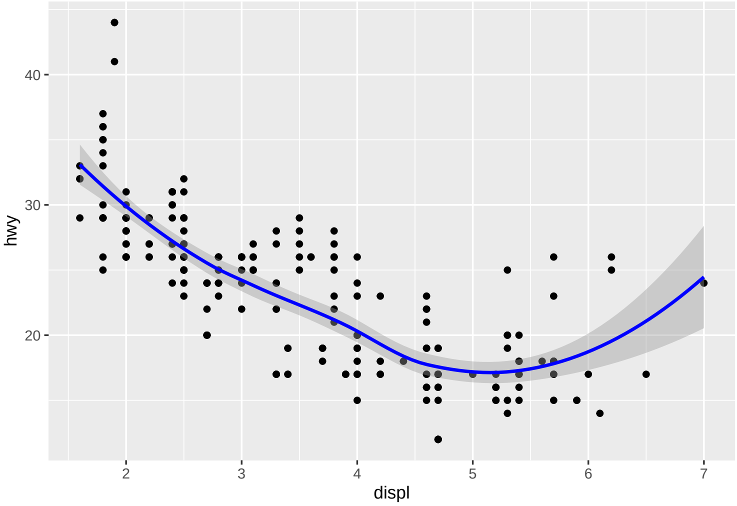
Note that the following two ways give the same result:

```
# in the notes, we only discussed using data in ggplot()
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point() +
  geom_smooth()
```

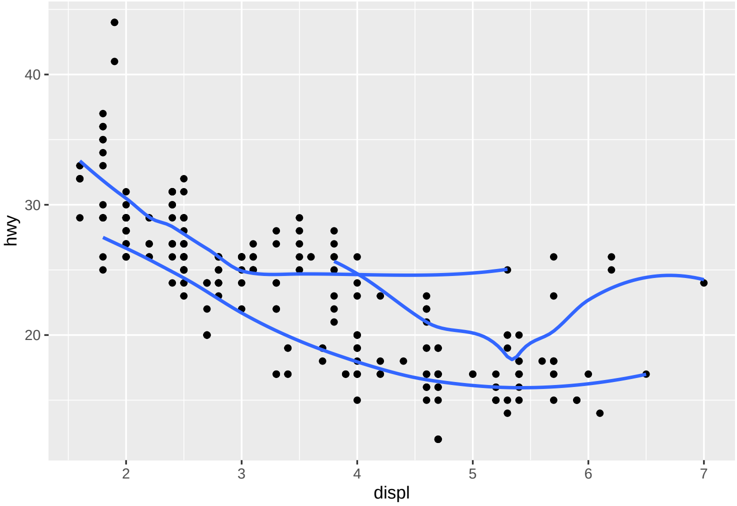
```r
# we can also use data in the geom objects
ggplot() +
  geom_point(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_smooth(data = mpg, mapping = aes(x = displ, y = hwy))
```

Q10(a): What does the `se` argument to `geom_smooth()` do? Hint: try `?geom_smooth`.

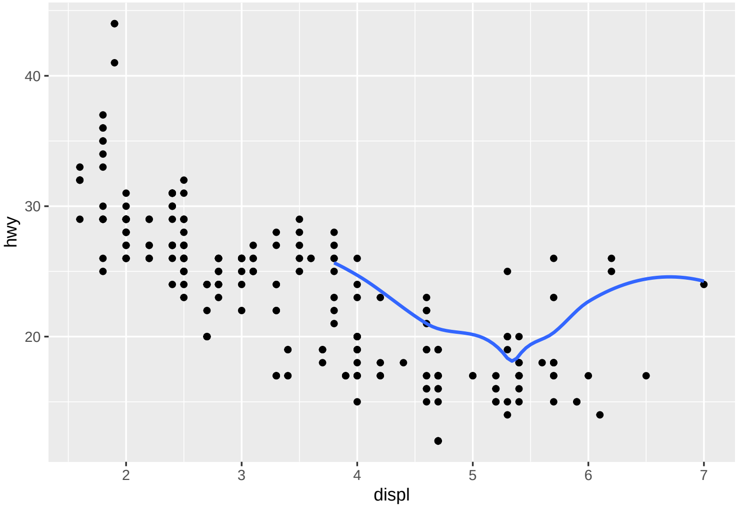
Q10(b): Recreate the R code necessary to generate the following graph.



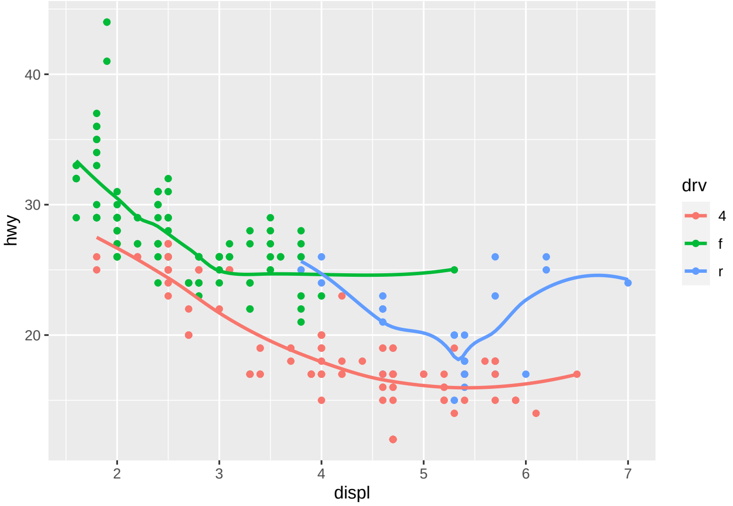
Q10(c): Recreate the R code necessary to generate the following graph. The smooth lines correspond to the data points with different values of `drv`.



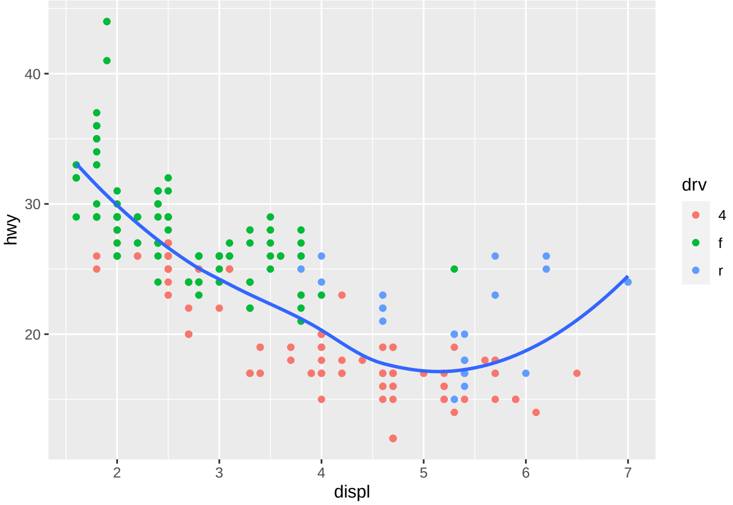
Q10(d): Recreate the R code necessary to generate the following graph. The smooth line corresponds to the data points with `drv` equals `"r"`.



Q10(e): Recreate the R code necessary to generate the following graph.



Q10(f): Recreate the R code necessary to generate the following graph.


hwy
displ
drv
f
r