Assignment 1

Kayla Choi - MA615 2021

9/20/2021

```
#include the necessary library
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

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

```
#source the R script that holds our functions
source(file="hand_functions.R", echo = TRUE)
```

```
##
## > sum_special <- function(df_x) {
## +     try(if (!is.data.frame(df_x))
## +         stop("Input data must be a data frame."))
## +     sp_means <- apply(df_ .... [TRUNCATED]
```

```
# call built-in data mtcars.
data(mtcars)
```

```
# Select only car models where mpg<20
mtcars_mpg2 <- mtcars[mtcars$mpg < 20,]

# Reduce the variables to mpg, cyl, disp, hp, gears
mtcars_mpg2 <- mtcars_mpg2[, c(1,2,3,4,10)]

#look at the data frame and its summary
head(mtcars_mpg2)
```

```
##                   mpg cyl  disp  hp gear
## Hornet Sportabout 18.7   8 360.0 175    3
## Valiant           18.1   6 225.0 105    3
## Duster 360        14.3   8 360.0 245    3
## Merc 280          19.2   6 167.6 123    4
## Merc 280C         17.8   6 167.6 123    4
## Merc 450SE        16.4   8 275.8 180    3
```

```
summary(mtcars_mpg2)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :6.000   Min.   :145.0   Min.   :105.0
##  1st Qu.:14.78   1st Qu.:8.000   1st Qu.:275.8   1st Qu.:156.2
##  Median :15.65   Median :8.000   Median :311.0   Median :180.0
##  Mean   :15.90   Mean   :7.556   Mean   :313.8   Mean   :191.9
##  3rd Qu.:18.02   3rd Qu.:8.000   3rd Qu.:360.0   3rd Qu.:226.2
##  Max.   :19.70   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       gear
##  Min.   :3.000
##  1st Qu.:3.000
##  Median :3.000
##  Mean   :3.444
##  3rd Qu.:3.750
##  Max.   :5.000
```

Looking at the summary of mtcars_mpg2 data frame, it is evident that the descriptive statistics for cyl (number of cylinders) and gear are not meaningful. This is because they are factor variables and not continuous.

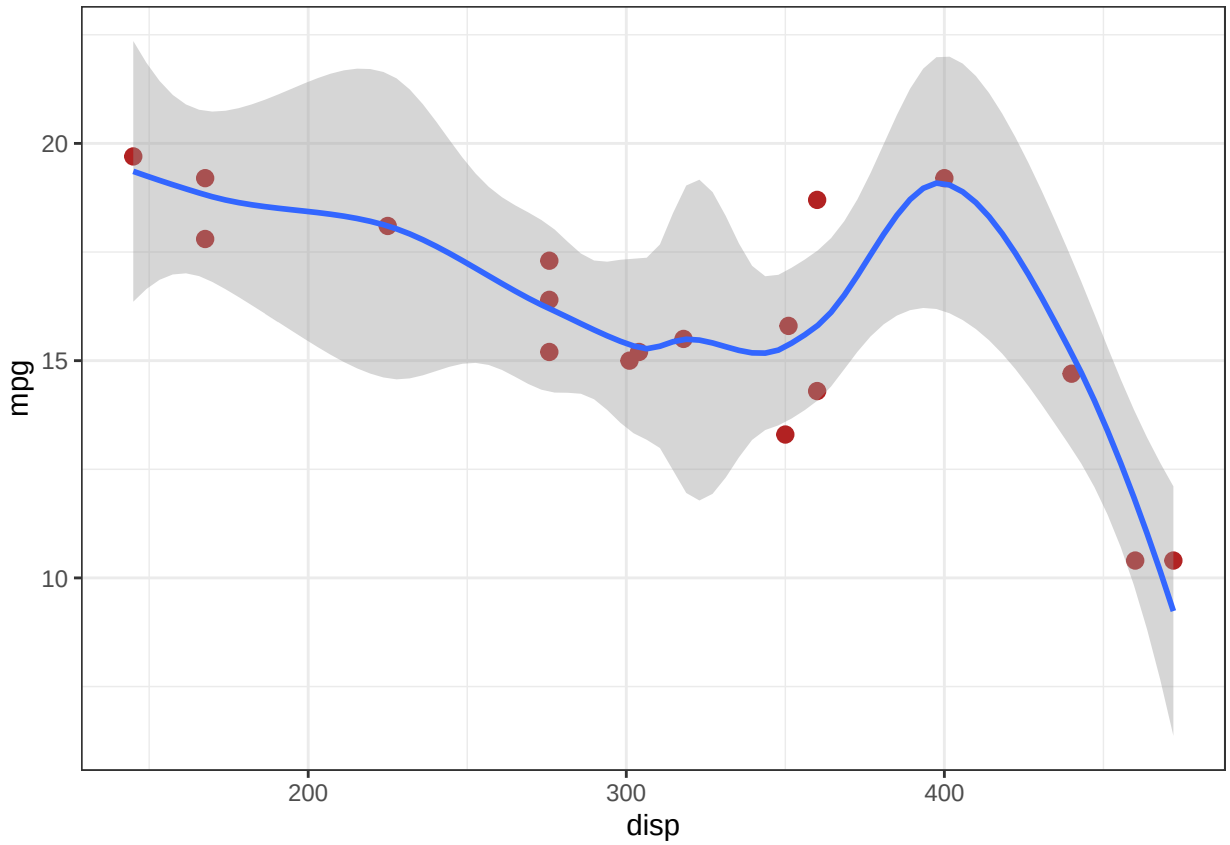
```
# Now use the function from hand_functions.R and store in a variable
sp_out <- sum_special(mtcars_mpg2)
sp_out
```

```
## $sp_means
##    mpg    cyl   disp     hp   gear
##  15.90   7.56 313.81 191.94   3.44
##
## $sp_var
##     mpg     cyl    disp      hp    gear
##    7.53    0.73 9438.76 3253.58    0.61
##
## $sp_cov
##          mpg   cyl    disp      hp   gear
## mpg     7.53 -1.32 -188.80  -75.81   0.64
## cyl    -1.32  0.73   64.71   28.44  -0.26
## disp -188.80 64.71 9438.76 2679.60 -34.19
## hp    -75.81 28.44 2679.60 3253.58  15.20
## gear    0.64 -0.26  -34.19   15.20   0.61
##
## $sp_cor
##        mpg   cyl  disp    hp  gear
## mpg   1.00 -0.56 -0.71 -0.48  0.30
## cyl  -0.56  1.00  0.78  0.58 -0.39
## disp -0.71  0.78  1.00  0.48 -0.45
## hp   -0.48  0.58  0.48  1.00  0.34
## gear  0.30 -0.39 -0.45  0.34  1.00
```

This function outputs a list of summary statistics, including mean, variance, covariance, and correlation. It is interesting to note that there is a high negative correlation between disp and mpg. Let's explore that in the next ggplot.

```
#explore visualization of disp vs mpg
ggplot(mtcars_mpg2) +
  aes(x = disp, y = mpg) +
  geom_point(shape = "bullet", size = 4L, colour = "#B22222") +
  geom_smooth(span = 0.5) +
  theme_bw()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



From this plot, we can see that as displacement (cu.in.) increases, mpg generally decreases. However, I notice two points that do not follow this general trend. I want to look at which ones these are.

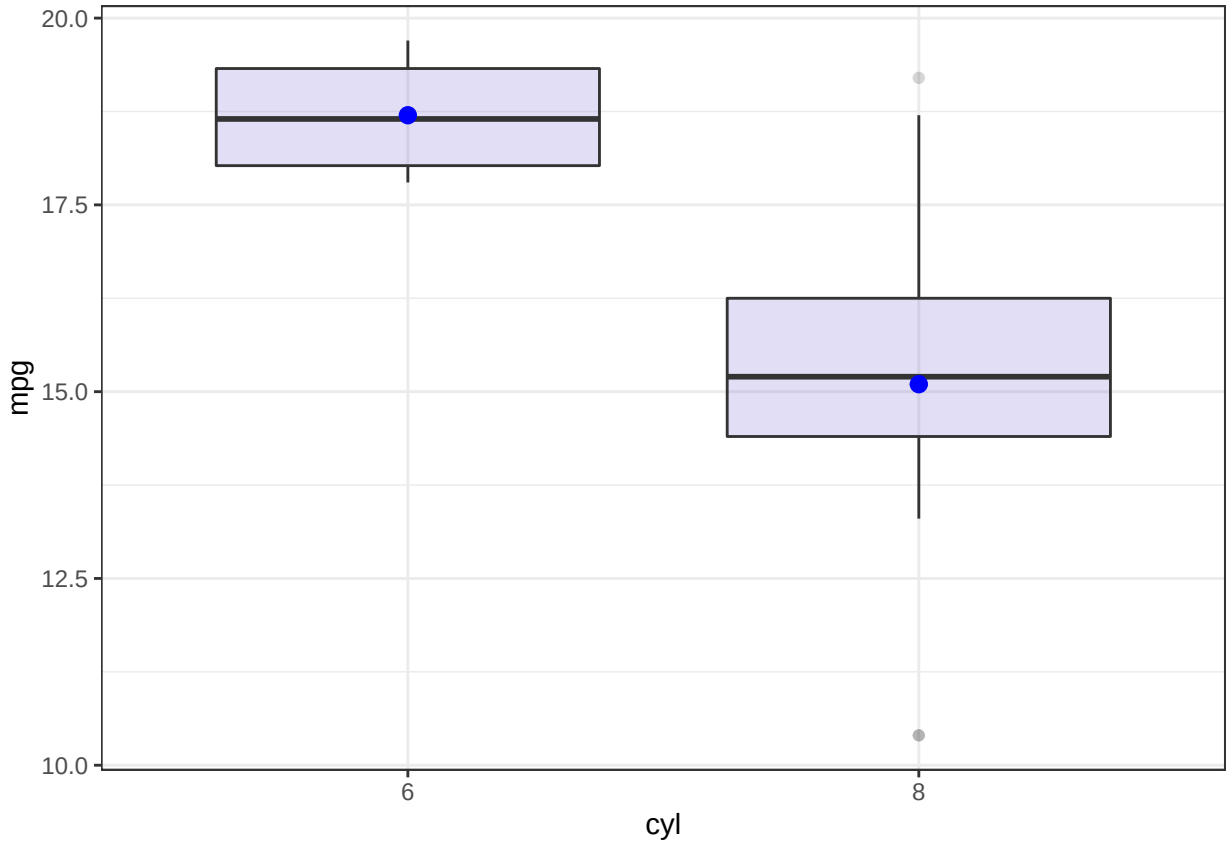
```
#look at the two unusual points
disp_over_350 <- mtcars_mpg2[(mtcars_mpg2$disp) >= 350, ]
unusual_mpg <- disp_over_350[(disp_over_350$mpg) >= 17, ]
unusual_mpg
```

```
##                    mpg cyl disp  hp gear
## Hornet Sportabout 18.7   8  360 175    3
## Pontiac Firebird  19.2   8  400 175    3
```

They are the Hornet Sportabout and the Pontiac Firebird.

Next, let's look at cylinders vs mpg. Because cylinder is a factor, as I mentioned above when we looked at the summary statistics, we need to make sure we convert it via as.factor().

```
#explore visualization of cyl vs mpg
ggplot(mtcars_mpg2, aes(x=as.factor(cyl), y=mpg)) +
    geom_boxplot(fill="slateblue", alpha=0.2) +
    stat_summary(fun=mean, geom="point", shape=20, size=4, color = "blue") +
    xlab("cyl") +
    theme_bw()
```



For cars with 8 cylinders, there is a larger spread in the mpg and the median is lower. I have added the mean as well in the visualization. There are also two outliers with 8 cylinder cars. I want to look at which ones these are.

```
cyl_8 <- mtcars_mpg2[(mtcars_mpg2$cyl) ==8 , ]
outlier_cyl_large <- cyl_8[(cyl_8$mpg) >= 18.75, ]
outlier_cyl_small <- cyl_8[(cyl_8$mpg) <= 12.5, ]
outlier_cyl_large
```

```
##                   mpg cyl disp  hp gear
## Pontiac Firebird 19.2   8  400 175    3
```

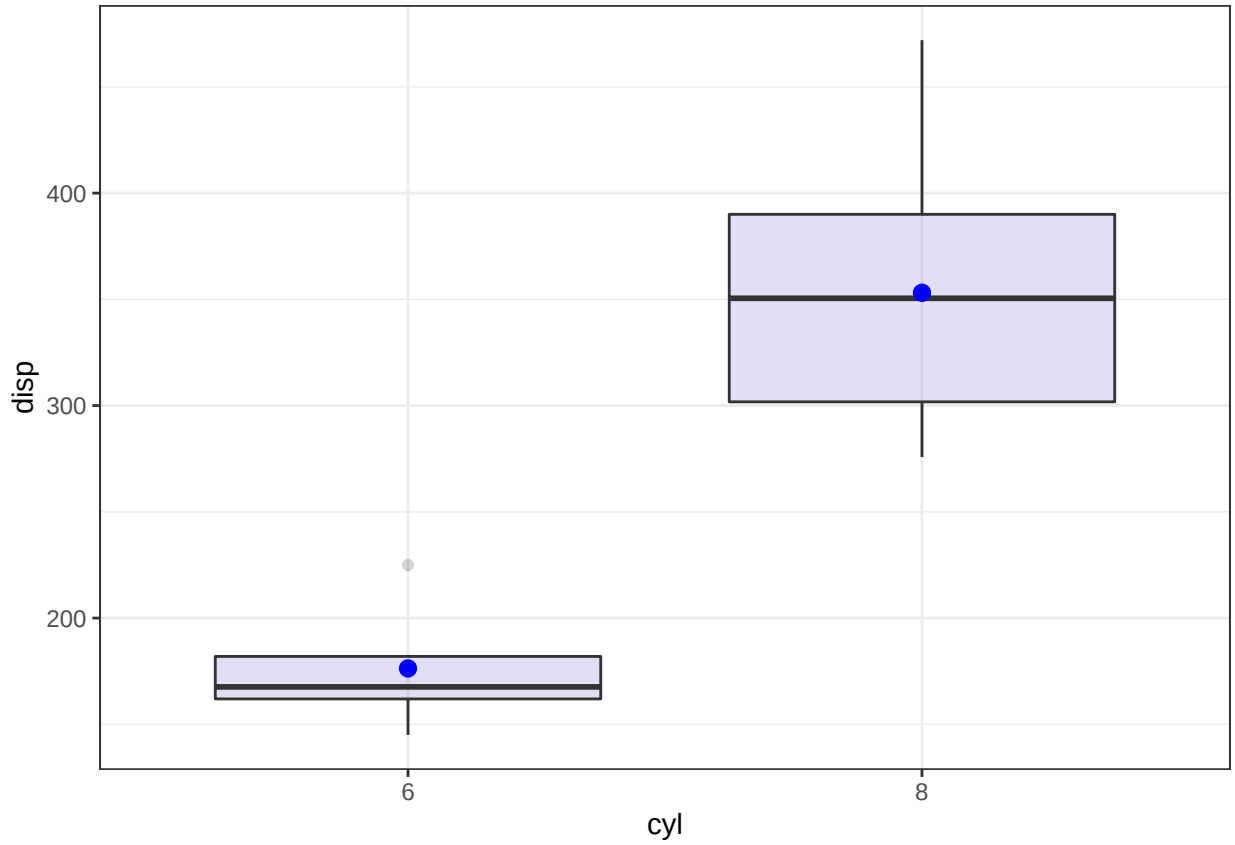
```
outlier_cyl_small
```

```
##                      mpg cyl disp  hp gear
## Cadillac Fleetwood  10.4   8  472 205    3
## Lincoln Continental 10.4   8  460 215    3
```

The Pontiac Firebird has a high mpg compared to other 8-cylinder models.
I only saw one distinct point that was a low outlier, but it turns out that there are actually two models that have the same mpg: Cadillac Fleetwood and Lincoln Continental.

We also saw from the sum_special function that there is a high positive correlation between cylinder and displacement. Let's explore that.

```
#explore visualization of cyl vs mpg
ggplot(mtcars_mpg2, aes(x=as.factor(cyl), y=disp)) +
    geom_boxplot(fill="slateblue", alpha=0.2) +
    stat_summary(fun=mean, geom="point", shape=20, size=4, color = "blue") +
    xlab("cyl") +
    theme_bw()
```



For cars with 6 cylinders, the displacement is much lower, has a smaller spread, and is left-skewed. The 8 cylinder displacement has a significantly higher median displacement and a larger spread.