Homework 1 (STAT 5860)

Eric Pettengill

Insturctions:

1. Download the `Homework1.Rmd` file from the course Elearning.
2. Open `Homework1.Rmd` in RStudio.
3. Replace the *"Your Name Here"* text in the `author` with your own name.
4. Write your answer to each problem by editing `Homework1.Rmd`.
5. After you finish all the problems, click `Knit to PDF` to create a pdf file. Upload your pdf file to Homework 1 Dropbox in the course Elearning.

Loading packages

```
library(MASS)
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
## select(): dplyr, MASS
```

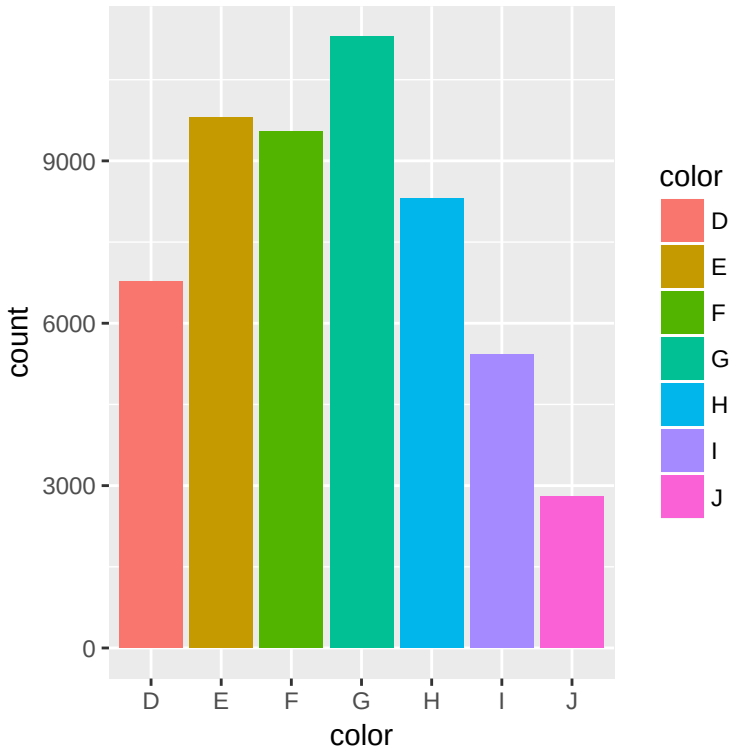
```
library(forcats)
```

Problem 1. For this problem we'll use the `diamonds` dataset from the `ggplot2` package. Use the functions inside `dplyr` and `ggplot2` packages to answer the following questions.

```
# Change the object of diamonds dataset from tibble to data frame.
diamonds <- as.data.frame(diamonds)
```

(a) Draw the bar chart to display the total number of diamonds in the `diamonds` dataset, grouped by `color` grading.

```
ggplot(diamonds) +
  geom_bar(aes(x = color, fill = color))
```



(b) We call diamond is colorless if color grading is D, E, or F. Select colorless diamonds from the `diamonds` dataset and assign it to `colorless_diamonds` for later use.

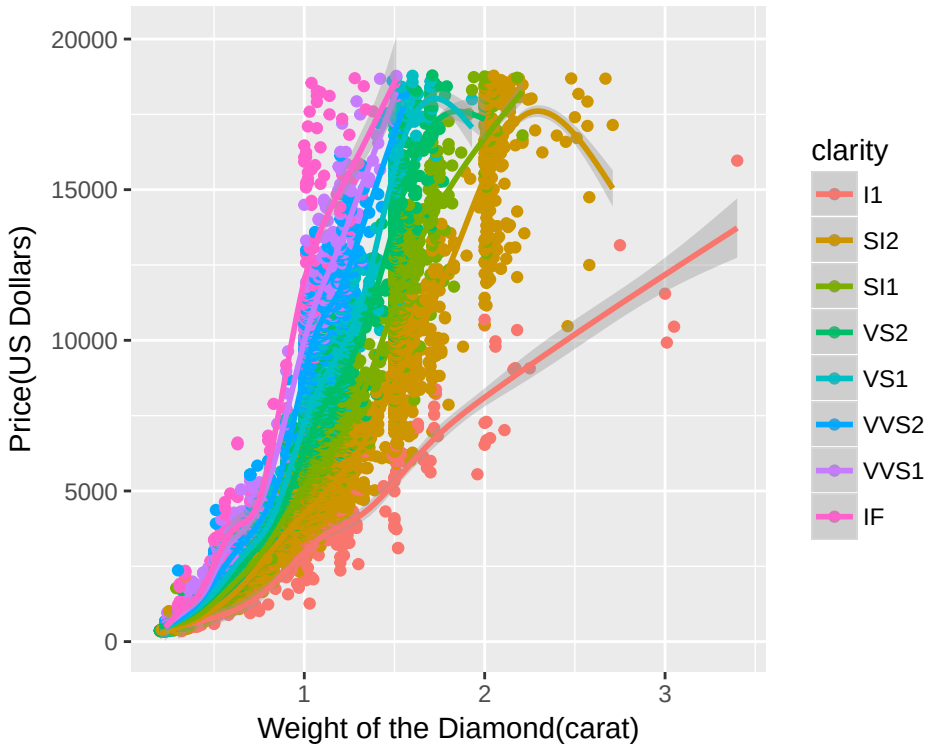
```
colorless_diamonds <- diamonds %>%
  filter(color %in% c("D", "E", "F"))

levels(colorless_diamonds)
```

```
## NULL
```

(c) Use `colorless_diamonds` data frame to recreate the uploaded *"The Value of Colorless Diamond"* plot.

```
ggplot(colorless_diamonds, aes(x=carat, y=price, color=clarity)) +
  geom_point() +
  geom_smooth() +
  ylab("Price(US Dollars)") +
  xlab("Weight of the Diamond(carat)")
```

```
## `geom_smooth()` using method = 'gam'
```



(d) The ideal depth for round diamond is between two percentages 59.5% and 62.9%, inclusive. Select the ideal depth diamonds from the `diamonds` dataset and assign it to `ideal_depth_diamonds` for later use.

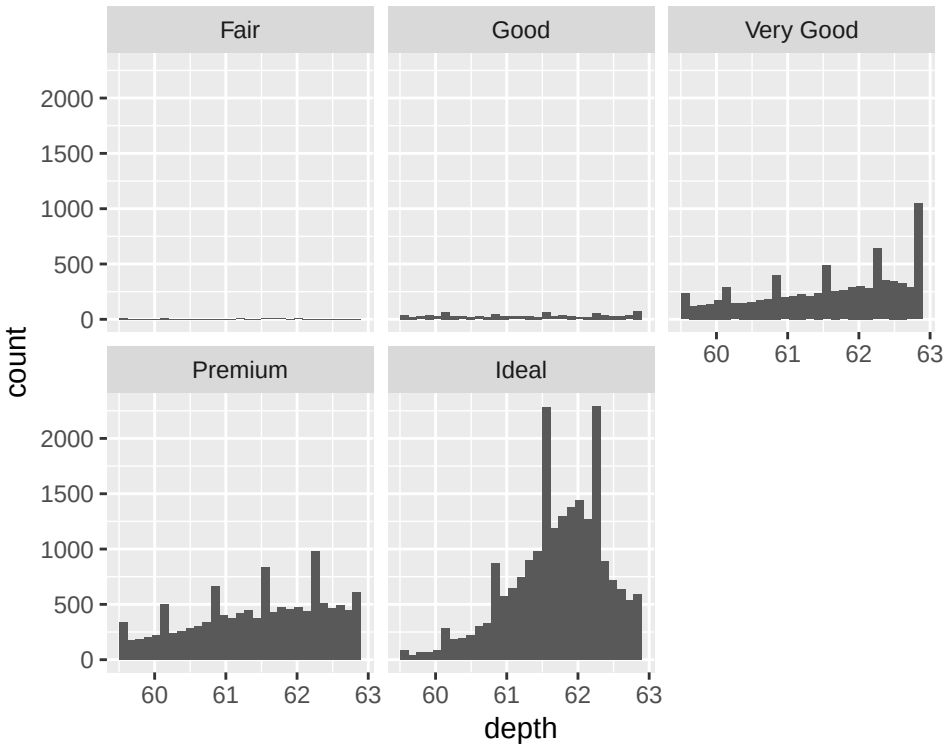
```
ideal_depth_diamonds <- diamonds %>%
  filter(depth >= 59.5 & depth <= 62.9)

range(ideal_depth_diamonds$depth)
```

```
## [1] 59.5 62.9
```

(e) Use `ideal_depth_diamonds` data frame to draw the histogram of the `depth` for each quality of the `cut` separately in one grpahic window.

```
ggplot(ideal_depth_diamonds, aes(x=depth)) +
  geom_histogram() +
  facet_wrap(~cut)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



(f) The length and width ratio gives an idea how big or small the diamond will appear, also defining its shape. To calculate the length and width ratio just divide the length of the diamond by its width. Add a new variable named `LWratio` to the `diamonds` dataset and assign it to `LW_diamonds`.

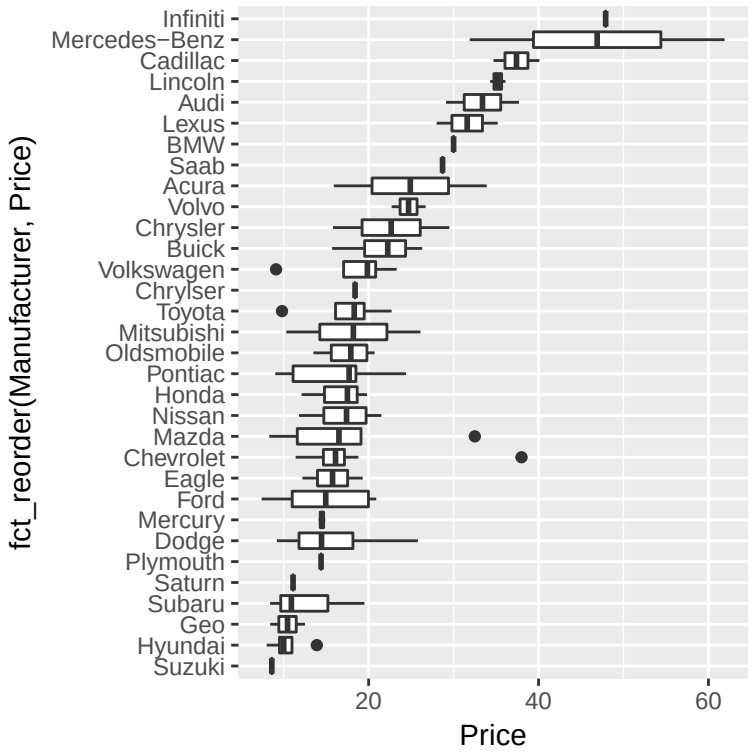
```
LW_diamonds <- diamonds %>%
  mutate(LWratio = x/y)

head(LW_diamonds)
```

```
##   carat       cut color clarity depth table price    x    y    z   LWratio
## 1  0.23     Ideal     E     SI2  61.5    55   326 3.95 3.98 2.43 0.9924623
## 2  0.21   Premium     E     SI1  59.8    61   326 3.89 3.84 2.31 1.0130208
## 3  0.23      Good     E     VS1  56.9    65   327 4.05 4.07 2.31 0.9950860
## 4  0.29   Premium     I     VS2  62.4    58   334 4.20 4.23 2.63 0.9929078
## 5  0.31      Good     J     SI2  63.3    58   335 4.34 4.35 2.75 0.9977011
## 6  0.24 Very Good     J    VVS2  62.8    57   336 3.94 3.96 2.48 0.9949495
```

Problem 2. For this problem we'll use the `Cars93` dataset from the `MASS` package. Write down two interesting questions that you could answer with `Cars93` dataset, and use appropriate visualizations using `ggplot2` to answer them. Make sure at least one question should involve multiple variables.

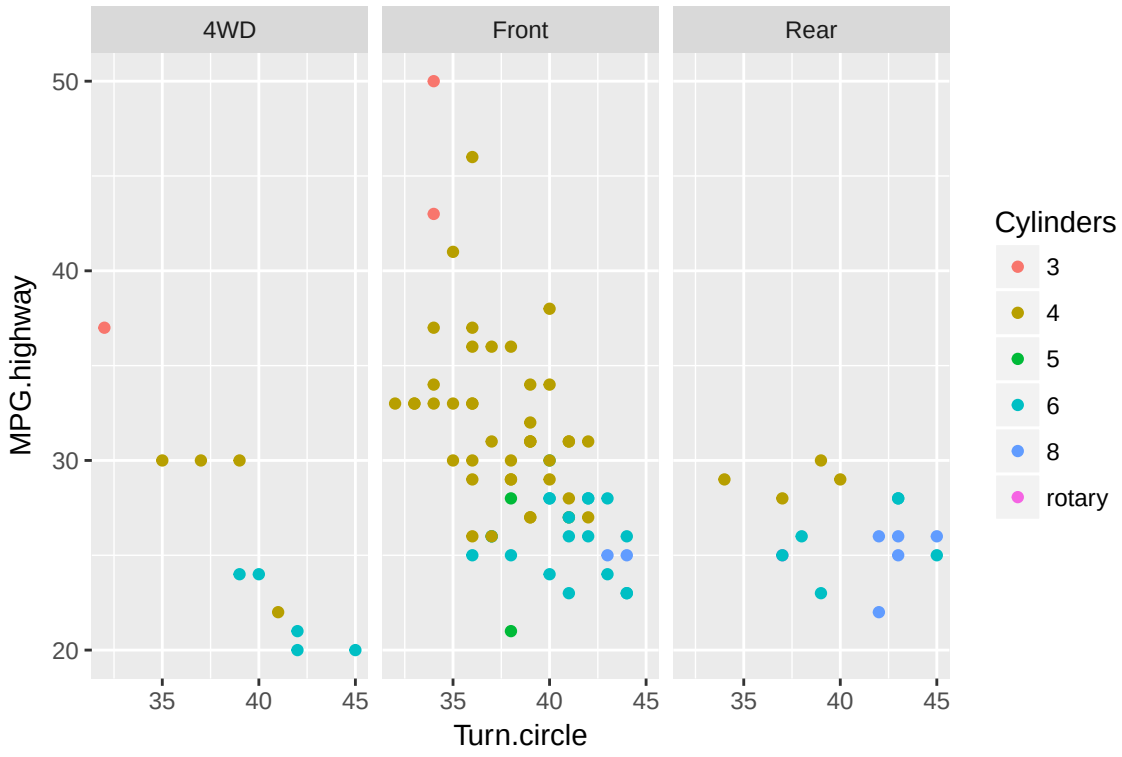
(a) Question 1: Which brands sell for more, i.e., what is the price distribution by brand?

```
ggplot(data = Cars93, aes(x = fct_reorder(Manufacturer, Price), y = Price)) +
  geom_boxplot() +
  coord_flip()
```



(b) Question 2: Are Turn.circle and MPG.highway correlated and by what DriveTrain and Cylinders

```
ggplot(Cars93, aes(x = Turn.circle, y = MPG.highway, color = Cylinders)) +
  geom_point() +
  facet_wrap(~DriveTrain)
```




```
ggplot(Cars93, aes(x = Turn.circle, y = MPG.city, color = Cylinders)) +
  geom_point() +
  facet_wrap(~DriveTrain)
```