Lab 2

Enter the names of the group members here: Austine Do, Graceanne Becker, Catherine Zhong

This assignment is due by the end of the lab. Only one student in the group submits a pdf file on Gradescope.

For all questions, include the R commands/functions that you used to find your answer (show R chunk). Answers without supporting code will not receive credit. Write full sentences to describe your findings.

In this lab, you will explore the dataset `diamonds` contained in the package `tidyverse`. Let's first upload the funnctions and objects available through that package:

```r
# Upload the package
library(tidyverse)
```

The dataset consists of prices and quality information from about 54,000 diamonds. The first few observations are listed below.

```r
head(diamonds)
```

```
## # A tibble: 6 x 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
## 2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
## 3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
## 4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
## 5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
## 6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
```

Question 1: (3 pts)

Save the dataset `diamonds` in the environment and name it using the initials of all team members. *Remember that you can get more details about the dataset by running `?diamonds` in the console.*

```r
# this code saves the diamonds dataset into the environment as our team intials
czadgb <- diamonds
```

How many rows are there in the dataset? How many columns?

```r
# this code gets the dimension of the dataset
dim(diamonds)
```

```
## [1] 53940    10
```

```r
str(diamonds)
```

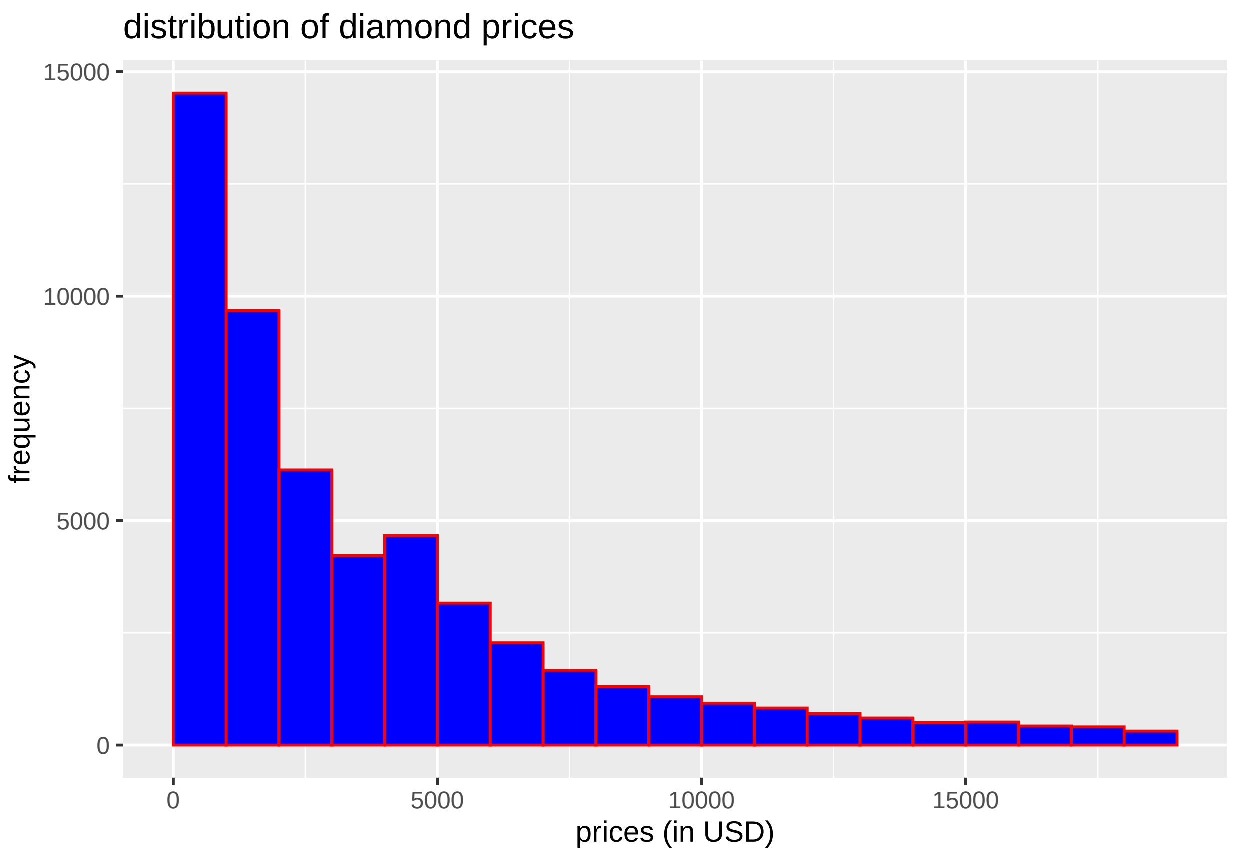
```
## tibble [53,940 x 10] (S3: tbl_df/tbl/data.frame)
##  $ carat  : num [1:53940] 0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
##  $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
##  $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
##  $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
##  $ depth  : num [1:53940] 61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
##  $ table  : num [1:53940] 55 61 65 58 58 57 57 55 61 61 ...
##  $ price  : int [1:53940] 326 326 327 334 335 336 336 337 337 338 ...
##  $ x      : num [1:53940] 3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
##  $ y      : num [1:53940] 3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
##  $ z      : num [1:53940] 2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

The dataset has 53940 rows and 10 columns.

Question 2: (5 pts)

Consider the variable `price` in US dollars. Represent the distribution of this variable with an appropriate graph using `ggplot()` (include a title and label). Comment on the shape of the distribution and report the appropriate statistics. Write a sentence to interpret these statistics.

```r
# This the distribution of the diamond prices as a histogram
diamonds |>
    ggplot(aes(x = price)) +
    geom_histogram(binwidth = 1000, bins = 20, center = 500, color = 'red', fill = 'blue') +
    labs(title = 'distribution of diamond prices',
         x = 'prices (in USD)',
         y = 'frequency')
```



```
summary(diamonds$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     326     950    2401    3933    5324   18823
```

```
IQR(diamonds$price)
```

```
## [1] 4374.25
```

The distribution of the diamond prices are positively skewed. The median of diamond price is $2401 and the IQR is $4374.25

Question 3: (6 pts)

The “4 Cs” of diamonds are traditionally `carat`, `cut`, `color`, and `clarity`. Create a new variable in your dataset, called `topfourC`, that has a `TRUE` value when satisfying ALL of these conditions (and is FALSE otherwise):

- the diamond’s cut is Ideal or Premium
- the color is D, E, or F (colorless)

- the clarity is IF, VVS1 or VVS2 (internally flawless or with very very slight inclusions)
- the diamond is in the top 25 percent for carat (i.e., carat is above the 3rd quartile).

```
# This code adds the variable 'topfourC' variable to the data frame
summary(diamonds$carat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.2000  0.4000  0.7000  0.7979  1.0400  5.0100
```

```
czadgb <- mutate(czadgb, topfourC = ifelse((carat >= 1.0400) & (cut == 'Ideal' | cut == 'Premium') & (c
```

Find the number of diamonds that meet these criteria. Is it rare for a diamond to meet this criteria?

```
# This counts the number of diamonds that satisfies the top four C's conditions
sum(czadgb$topfourC == TRUE)
```

```
## [1] 319
```

```
319/53490
```

```
## [1] 0.005963732
```

Yes it is rarely because based on this dataset only 0.5% of diamonds meet this expectation

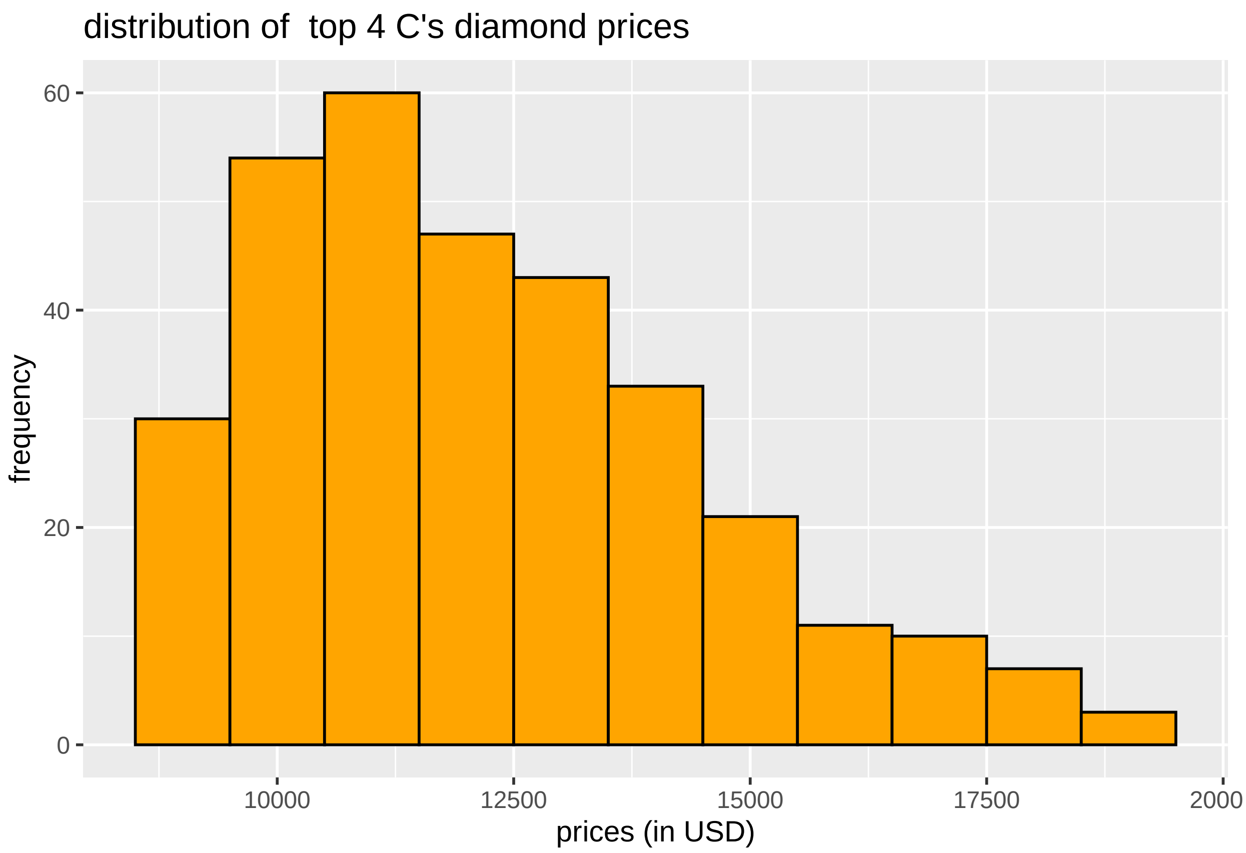
---

Question 4: (4 pts)

Focusing on the diamonds meeting the conditions for `topfourC`, represent the distribution of `price` with the same type of graph you used in question 3 (include a title and label). How do the two distributions (distribution of `price` for all diamonds vs distribution of `price` for top diamonds) compare? *Hint: refer to shape, center, and spread.*

```
# This is the price distribution of the top four C's diamonds as a histogram

czadgb |>
    filter(topfourC == TRUE) |>
    ggplot(aes(x = price)) +
    geom_histogram(binwidth = 1000, bins = 20, color = 'black', fill = 'orange') +
    labs(title = 'distribution of  top 4 C\'s diamond prices',
         x = 'prices (in USD)',
         y = 'frequency')
```



```
top4c <- filter(czadgb, czadgb$topfourC == TRUE)
summary(top4c$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8610   10442   11846   12217   13592   18700
```

```
IQR(top4c$price)
```

```
## [1] 3149
```

Both are positively skewed but the median for the top four C's diamond price was much higher at $11846 and the IQR was $3149.

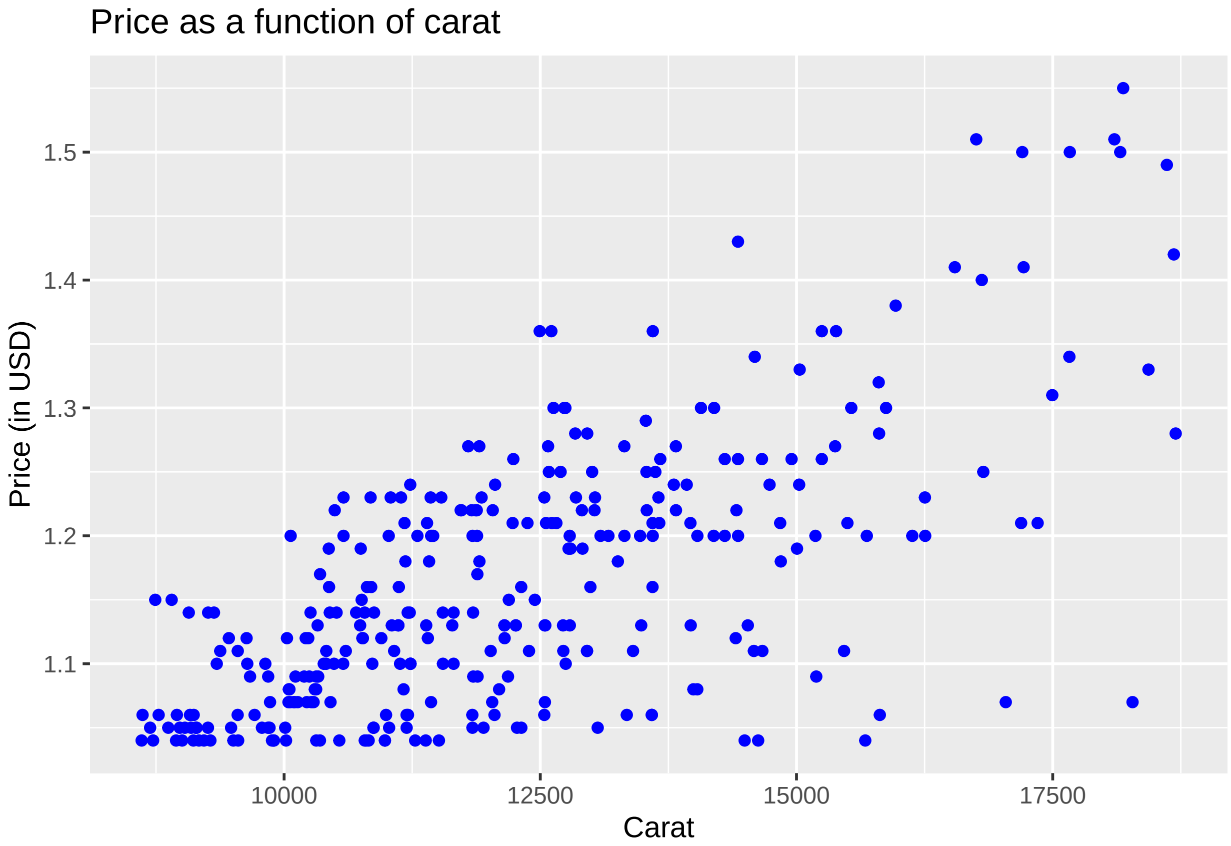
Question 5: (5 pts)

Still focusing on the diamonds meeting the conditions for `topfourC`, choose a numeric variable that you think might affect the `price` of a top diamond. Write a question you would like to investigate using these variables (could be a question one of you suggested in the intro lab):

Does the weight of the diamond usually lead to a higher price point?

Using a `ggplot` with `geom_point()`, make a visualization to answer your question (include a title and labels).

```
# A scatter plot of the price vs. carat
top4c |>
    ggplot() +
    geom_point(aes(x = price, y = carat), color = 'blue') +
    labs(title = 'Price as a function of carat',
         x = 'Carat',
         y = 'Price (in USD)')
```



How would you interpret what you see in this visualization?

There is a weak and positive correlation between the carat (weight) of top four C's diamonds and the price of top four C's diamonds.

Formatting: (2 pts)

Make sure the names of all group members are included at the beginning of the document.

Knit your file! You can knit into pdf directly or into html. Once it knits in html, click on `Open in Browser` at the top left of the window pops out. Print your html file into pdf from your browser.

Any issue? Ask other classmates or TA!

Finally, remember to select pages for each question when submitting your pdf to Gradescope and to identify your group members.