Student Name: Yanli Dong

FINA 5376 Financial Data Analytics

Professor Dr.Diltz

Spring 2021

Project 1 – The Layered Grammar of Graphics and Package “ggplot2”

Step 1:Execute the following commands in the RStudio console:> library(ggplot2)> data(package = “ggplot2”)This returns a list of datasets that reside in the ggplot2 package

```
Data sets in package 'ggplot2':

diamonds                        Prices of over 50,000 round cut diamonds
economics                       US economic time series
economics_long                  US economic time series
faithfuld                       2d density estimate of Old Faithful data
luv_colours                     'colors()' in Luv space
midwest                         Midwest demographics
mpg                             Fuel economy data from 1999 to 2008 for 38 popular models of cars
msleep                          An updated and expanded version of the mammals sleep dataset
presidential                    Terms of 11 presidents from Eisenhower to Obama
seals                           Vector field of seal movements
txhousing                       Housing sales in TX
```

Step 2:Select a dataset from the list, and study the data:Number of variables and observationsVariable names and definitionsYou can google on “package ggplot2” or “datasets contained in ggplot2” for more information

*Dataset:* `diamonds`

*A dataset containing the prices and other attributes of almost 54,000 diamonds.*

*It is a data frame with 53940 rows and 10 variables.*

***price***

*price in US dollars (\$326--\$18,823)*

***carat***

*weight of the diamond (0.2--5.01)*

***cut***

*quality of the cut (Fair, Good, Very Good, Premium, Ideal)*

***color***

*diamond colour, from D (best) to J (worst)*

***clarity***

a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))

**x**

length in mm (0--10.74)

**y**

width in mm (0--58.9)

**z**

depth in mm (0--31.8)

**depth**

total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)

**table**

width of top of diamond relative to widest point (43--95)

Step 3:Explore your data by plotting various combinations of data in various ways. Create new variables with mutate() if necessary. Run any kind of statistics you see fit. Run regressions using function lm().

```
> ggplot(data=diamonds)+geom_bar(mapping=aes(x=cut))+ggtitle("Count verse Cut")

> ggplot(diamonds) +geom_histogram(mapping = aes(x = x), binwidth = 0.01)

 > ggplot(diamonds) +geom_histogram(mapping = aes(x = y), binwidth = 0.01)

> ggplot(diamonds) +geom_histogram(mapping = aes(x = z), binwidth = 0.01)


> ggplot(data =diamonds)+geom_histogram(mapping = aes(x=carat),binwidth = 0.5)+ggtitle("Count Versus Carat Size")
```

**summary statistics for diamond**

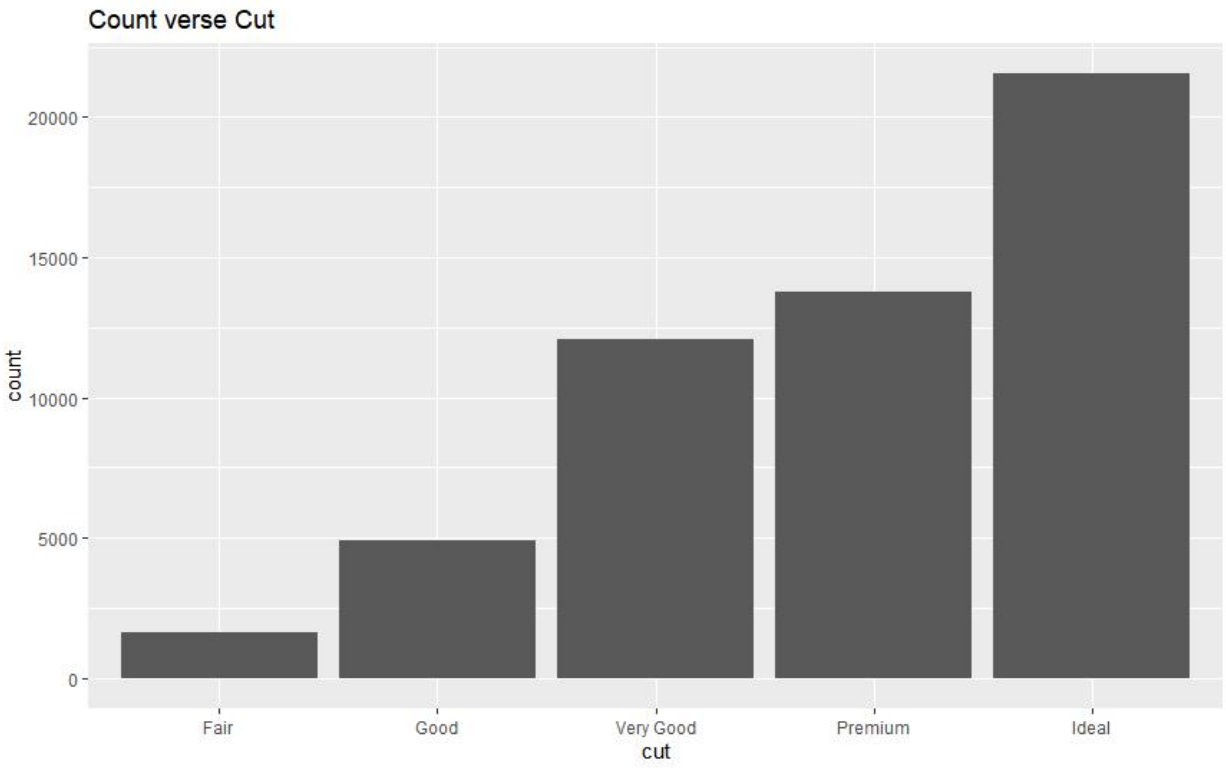
```
> summary(diamonds)
```

```
> summary(diamonds)
     carat               cut        color        clarity          depth           table           price
 Min.   :0.2000   Fair      : 1610   D: 6775   SI1     :13065   Min.   :43.00   Min.   :43.00   Min.   :  326
 1st Qu.:0.4000   Good      : 4906   E: 9797   VS2     :12258   1st Qu.:61.00   1st Qu.:56.00   1st Qu.:  950
 Median :0.7000   Very Good:12082   F: 9542   SI2     : 9194   Median :61.80   Median :57.00   Median : 2401
 Mean   :0.7979   Premium   :13791   G:11292   VS1     : 8171   Mean   :61.75   Mean   :57.46   Mean   : 3933
 3rd Qu.:1.0400   Ideal     :21551   H: 8304   VVS2    : 5066   3rd Qu.:62.50   3rd Qu.:59.00   3rd Qu.: 5324
 Max.   :5.0100                     I: 5422   VVS1    : 3655   Max.   :79.00   Max.   :95.00   Max.   :18823
                                    J: 2808   (Other): 2531
       x                y                z
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910
 Median : 5.700   Median : 5.710   Median : 3.530
 Mean   : 5.731   Mean   : 5.735   Mean   : 3.539
 3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040
 Max.   :10.740   Max.   :58.900   Max.   :31.800
```

Step 4 :

Plot1:

> ggplot(data=diamonds)+geom_bar(mapping=aes(x=cut))+ggtitle("Count verse Cut")

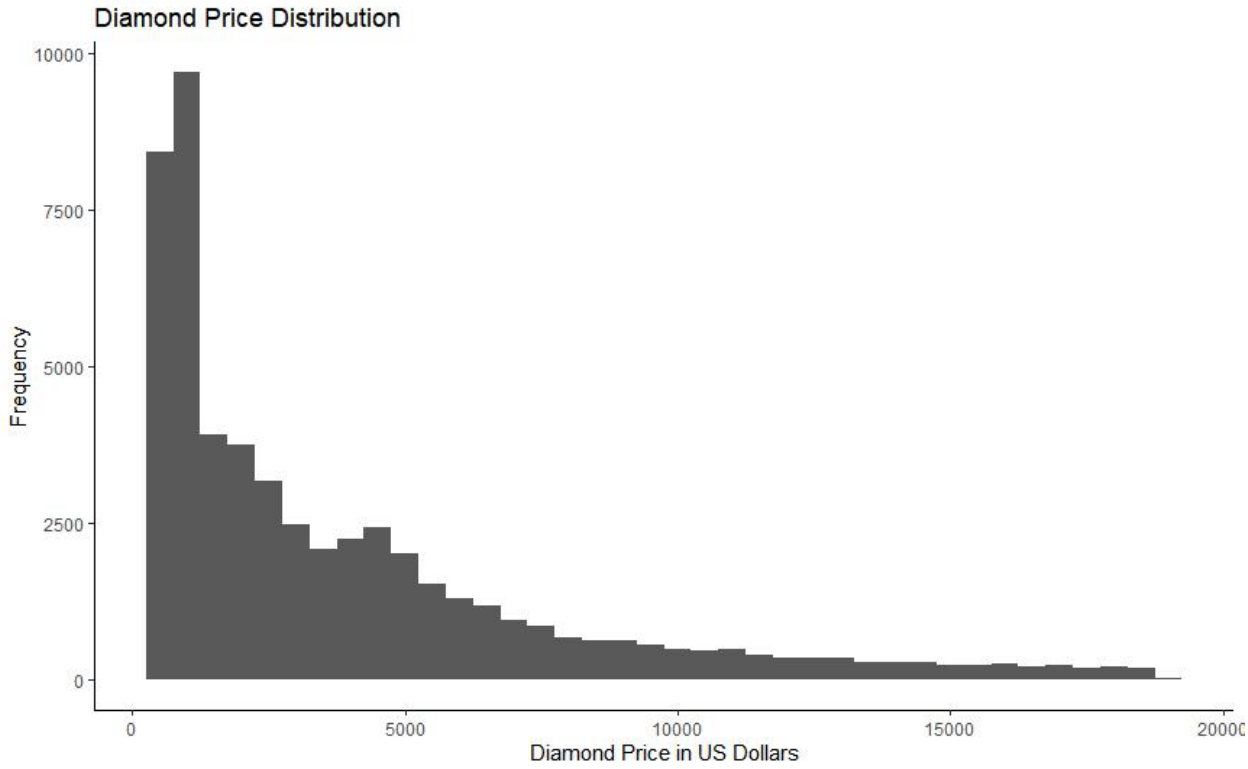



1. This plot is about frequency about different cut, on the x-axis, the chart displays cut, a variable from diamonds, on the y-axis, it displays count, the plot title is "Count verse Cut".The plot is based on counts of diamonds in each bin, the idea cut has the most counts, followed by Premium, very Good, and Good, the smallest count is Fair cut.

   geom_bar() begins with the diamonds dataset, and transforms the data with the count stat, which return a data set of cut values and counts, then use the transformed data to build the plot, cut is mapped to the x axis, count is mapped to the y axis. the aes argument specifies the visual properties(x).

2. > ggplot(data = diamonds, aes(x = depth)) + geom_histogram(binwidth = 0.2) + facet_wrap(~ cut)

> ggplot(data=diamonds) + geom_histogram(binwidth = 500, aes(x=price)) + ggtitle("Diamond Price Distribution") + xlab ("Diamond Price in US Dollars") + ylab("Frequency") + theme_classic()

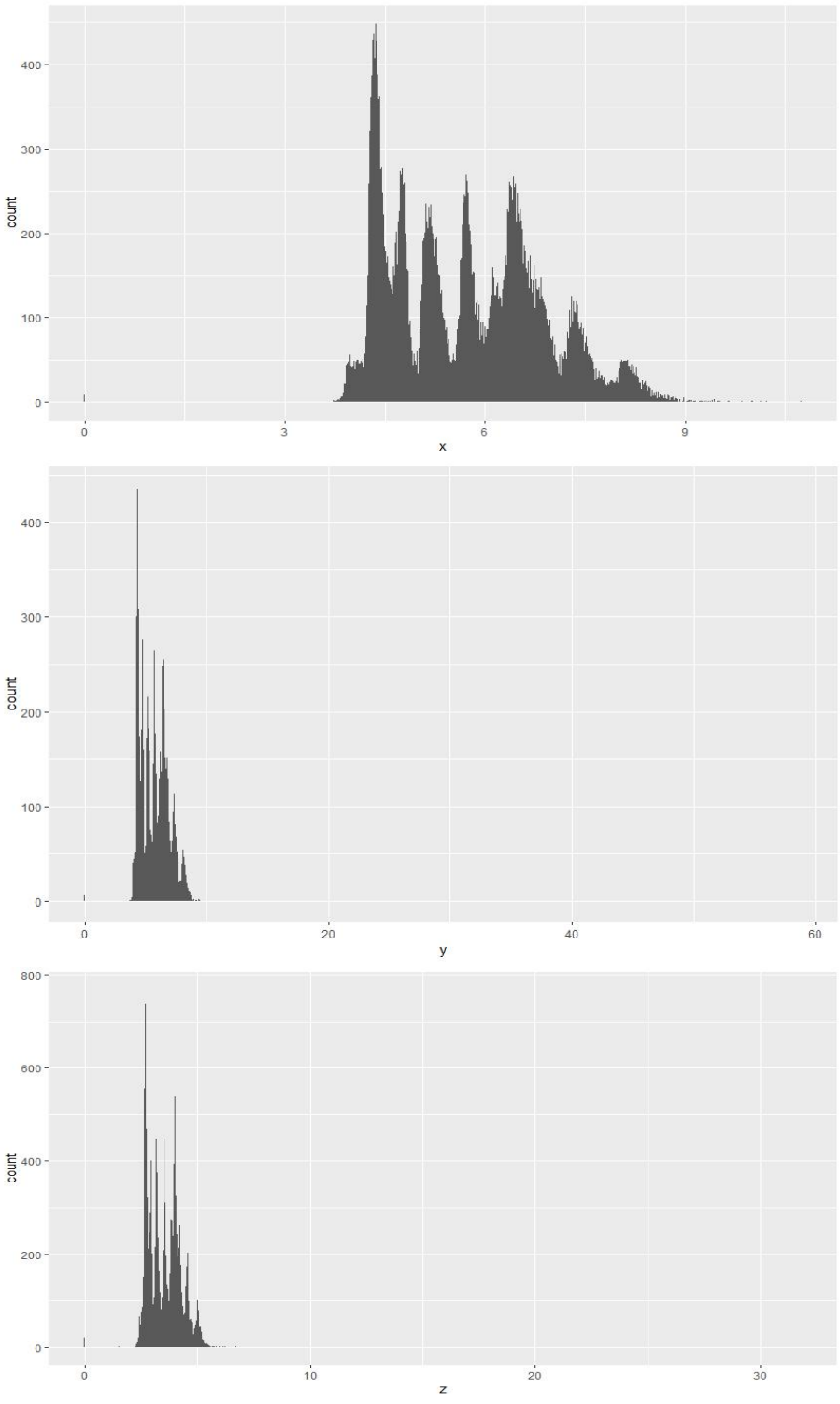



The plot is about the diamond price distribution which is a long tail distribution. It has a very high concentration of observations below US $5,000 mark. It seems to show that there is a demand for the higher quality diamonds. The frequency decease with diamond prices increases, which means less people who are willing to pay more for such higher quality diamonds.

3. > ggplot(diamonds) +geom_histogram(mapping = aes(x = x), binwidth = 0.01)

> ggplot(diamonds) +geom_histogram(mapping = aes(x = y), binwidth = 0.01)

> ggplot(diamonds) +geom_histogram(mapping = aes(x = z), binwidth = 0.01)

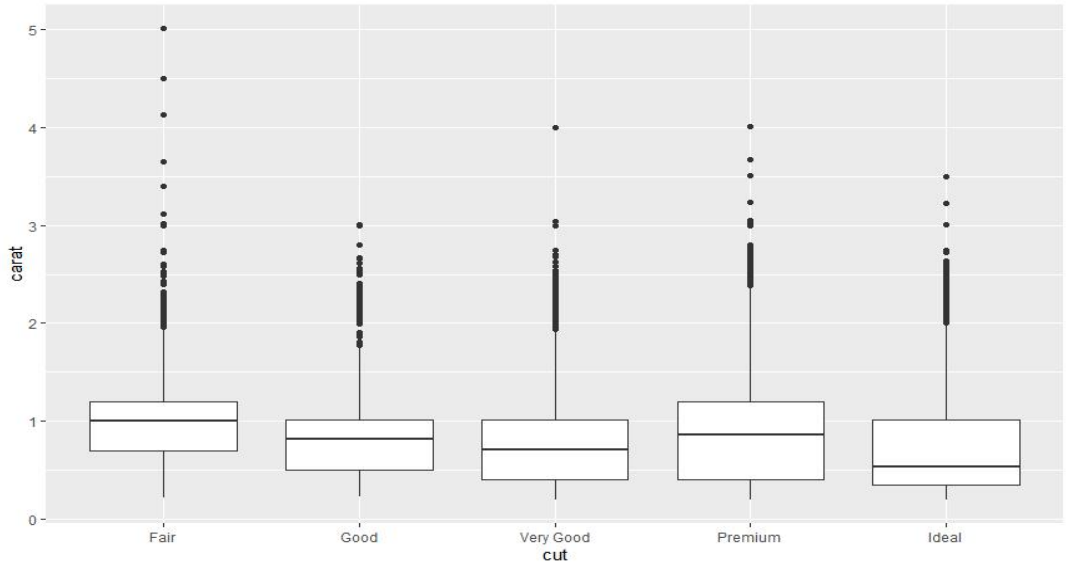



The three plots explore the distribution of each of the x, y, and z variables in diamonds. There several noticeable features of the distributions: x and y are larger than z,there are outliers,they are all right skewed, and they are multimodal or “spiky”. The typical values of x and y are larger than z, with x and y having inter-quartile ranges of 4.7–6.5,

while z has an inter-quartile range of 2.9–4.0. There are two types of outliers in this data. Some diamonds have values of zero and some have abnormally large values of x, y, or z.

4.

```
> ggplot(diamonds, aes(x = cut, y = carat)) + geom_boxplot()
```




The plot is about the relationship between cut and carat, since carat is a continuous variable and cut is a categorical variable, it can be visualized with a box plot.

There is a lot of variability in the distribution of carat sizes within each cut category. There is a slight negative relationship between carat and cut. Noticeably, the largest carat diamonds have a cut of “Fair” (the lowest).

This negative relationship can be due to the way in which diamonds are selected for sale. A larger diamond can be profitably sold with a lower quality cut, while a smaller diamond requires a better cut.

> ggplot(data =diamonds)+geom_histogram(mapping = aes(x=carat),binwidth = 0.5)+ggtitle("Count Versus Carat Size")