**Data Exploration and Visualization**

***Assignment#1***

*Exploration and Visualization of dataset “diamonds” using ggplot2.*

***Tamoor Akbar F2018313032***

***Hamza Kashif F2018313002***

***Summary:***

ggplot2 was installed and dataset of “diamonds” was uploaded to it. Using ggplot2, data was explored and visualized through different chart techniques like ***Histogram, Bar Chart, line chart, scattered charts*** ectc***.*** Operations like ***Fecet*** and ***fill*** was used to further explore the data on a deeper lever.

The dataset consists of 53940 entries with 10 columns. Our main focal during this study were price, carat(weight) and the characteristics like cut, clarity and color. In visualization we ignored dimensions (table, x,y,z, depth) because these are ultimately linked to weight (carat). Our findings are as below:

* Most of the diamonds’ price is less than $5000.
* Impact of color and clarity of diamond do not have strong impact on price; both of these categories were almost equally distributed in dataset. However, when we explored “cut”, we found there are more diamonds in data set with “Ideal” cut.
* Expensive diamonds (price < $5000) are either “ideal” or “premium” cut. Therefore we concluded *better the cut type, higher the price of diamond will be.*
* *Few trends that we notices by exploring:*
* Most of the diamonds are cut with ideal category followed by premium and very good.
* SI1,SI2 and VS2 are the most populated clarity categories.
* Color E,F,G are most common.
* There is a linear relationship between weight of diamond and its price. However, once diamond weight exceeds 2.25 Carat weight, the line gets bended with a flatter progression. Interestingly in most of the case, prices increase starts to slow down when the weight of diamond increase from 2.5 carat.
* Most of the diamonds weight less than 3 carats. The ones with more than 4 carat weight are only fair cut diamonds although fewer in quantity.

**Procedure and Findings:**

**# ggplot2 library**

To upload the ggplot2 library:

library(ggplot2)

install.packages("ggplot2")

**#Uploading dataset and review:**

diamonds

str(diamonds)

View(diamonds)

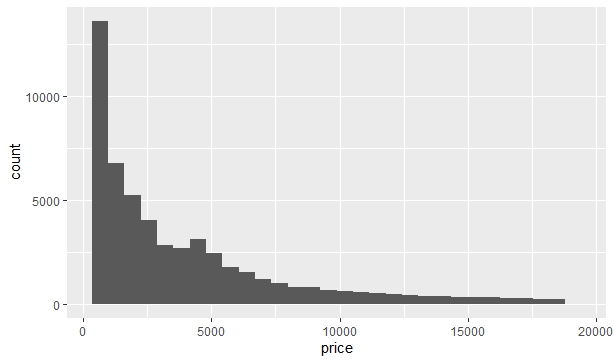
summary(diamonds)

This exploration gives some basic insight about the data. There are 53940 entries in data. Headers covers the attributes like weight(caret), weight(table, depth, x,y,z), price, cut, clarity and color. Among those cut, clarity and color are ordinary factors, dimension and weights are numbers and price is integer.

**# Histogram:**

Let’s start with basic Histogram to see the overall distribution of data w.r.t price:

ggplot(diamonds,aes(x=price))+geom\_histogram()



*The graphs shows most of the diamonds’ cost is less than $5000.*

**#Histogram – Diamonds’ color Classification:**

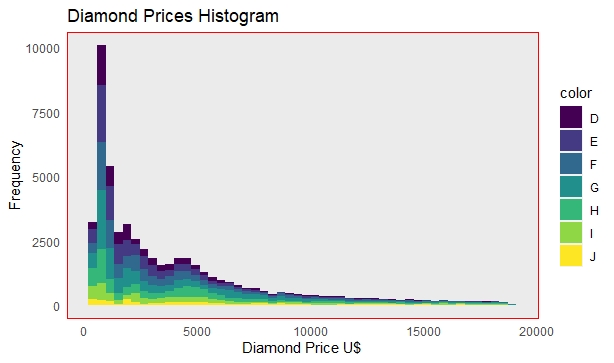
To explore the data, histogram was drawn by filling it by the color of diamond. Also, the graph was themed with background, border & titles to make it more visually aesthetic. Bin size was adjusted at 50 (after some hit & trail) for the appropriate distribution of data:

ggplot(diamonds,aes(x=price,fill=color))+geom\_histogram(bins = 50)+

ggtitle("Diamond Prices Histogram") +

xlab("Diamond Price U$") + ylab("Frequency")+

theme(panel.background = (element\_rect(color="red")),(element\_blank()))



*Through this we can see that price of diamond is not strongly dependent of its color because we can see in graph that all the colors are almost equally distributed in whole price range.*

Since the color classes are seven and the observations are more than 50000, therefore the data looks merged. Therefore to ensure the right trend reading it’s better to use facet:

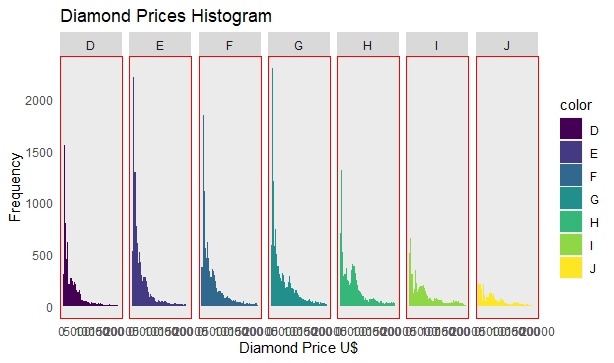
ggplot(diamonds,aes(x=price,fill=color))+geom\_histogram(bins = 50)+

ggtitle("Diamond Prices Histogram") +

xlab("Diamond Price U$") + ylab("Frequency")+

theme(panel.background = (element\_rect(color="red")),(element\_blank()))+

facet\_grid(~color)



*Here we can easily identify that the trend in all colors is same across price index.*

**#Histogram – diamonds’ clarity Classification:**

Now the same steps repeated for the other ordinal factor of clarity to identify the relation between diamond’s clarity and the price. Histogram of price was plotted by filling it with clarity. In the second graph faceting was done for better visualization:

ggplot(diamonds,aes(x=price,fill=clarity))+geom\_histogram(bins = 50)+

ggtitle("Diamond Prices Histogram") +

xlab("Diamond Price U$") + ylab("Frequency")+

theme(panel.background = (element\_rect(color="red")),(element\_blank()))

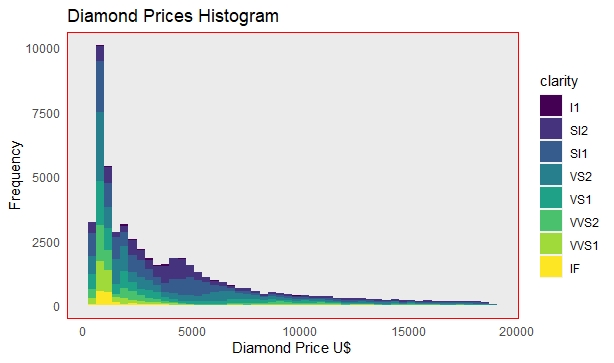
ggplot(diamonds,aes(x=price,fill=clarity))+geom\_histogram(bins = 50)+

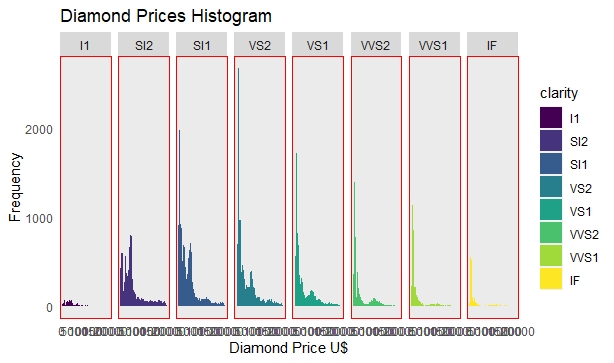
ggtitle("Diamond Prices Histogram") +

xlab("Diamond Price U$") + ylab("Frequency")+

theme(panel.background = (element\_rect(color="red")),(element\_blank()))+

facet\_grid(~clarity)





*Again, through this we can see that price of diamond is not strongly dependent of its clarity because we can see in graph that all the colors are almost equally distributed in whole price range.*

**#Histogram – diamonds’ cut Classification:**

Now we will see if there is any relationship between diamond’s cut type and its price. Let’s draw again a histogram but this time it will be filled with ordinal factor of “cut”:

ggplot(diamonds,aes(x=price,fill=cut))+geom\_histogram(bins = 50)+

ggtitle("Diamond Prices Histogram") +

xlab("Diamond Price U$") + ylab("Frequency")+

theme(panel.background = (element\_rect(color="red")),(element\_blank()))

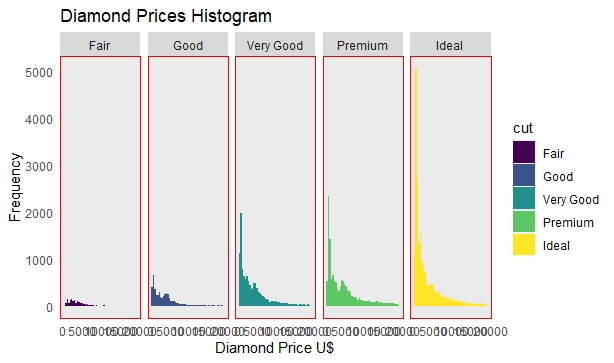
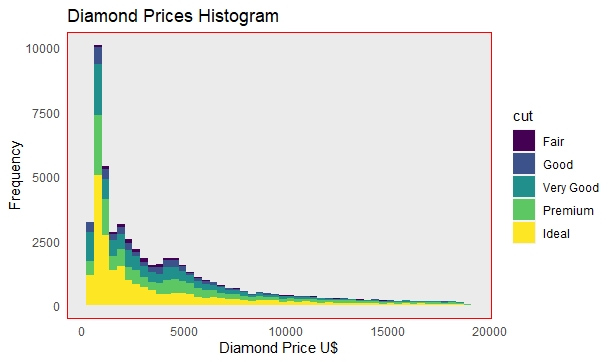
ggplot(diamonds,aes(x=price,fill=cut))+geom\_histogram(bins = 50)+

ggtitle("Diamond Prices Histogram") +

xlab("Diamond Price U$") + ylab("Frequency")+

theme(panel.background = (element\_rect(color="red")),(element\_blank()))+

facet\_grid(~cut)



*Here we can easily identify that there are more diamonds in dataset with ideal cut (yellow). Also, in high price range (>$1000) most of the diamonds are either Ideal or premium cut. Just a few are with “very good” but no case with “fair” category. Therefore we can infer, better the cut type, higher the price of diamond will be.*

**Bar Charts – Cut, Clarity, Color:**

We will use the bar charts now to see the frequency of different categorical factors like cut, clarity and color:

ggplot(data=diamonds,aes(x=cut))+geom\_bar(fill="aquamarine3")+

ggtitle("Bar Chart - Cut") +

xlab("Cut") + ylab("Frequency")+

theme(panel.background = (element\_rect(color="red")),(element\_blank()))

ggplot(data=diamonds,aes(x=clarity))+geom\_bar(fill="aquamarine2")+

ggtitle("Bar Chart - Clarity") +

xlab("clarity") + ylab("Frequency")+

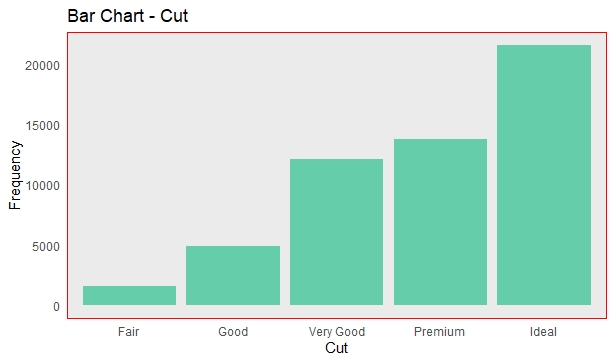
theme(panel.background = (element\_rect(color="red")),(element\_blank()))

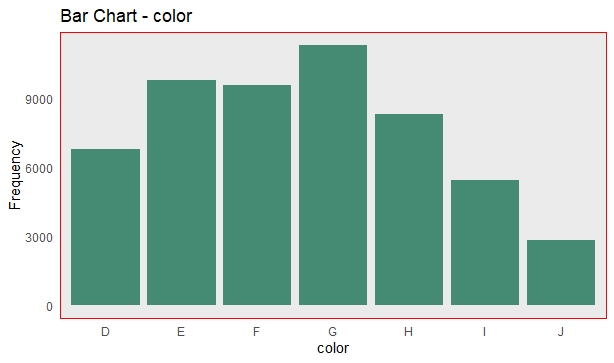
ggplot(data=diamonds,aes(x=color))+geom\_bar(fill="aquamarine4")+

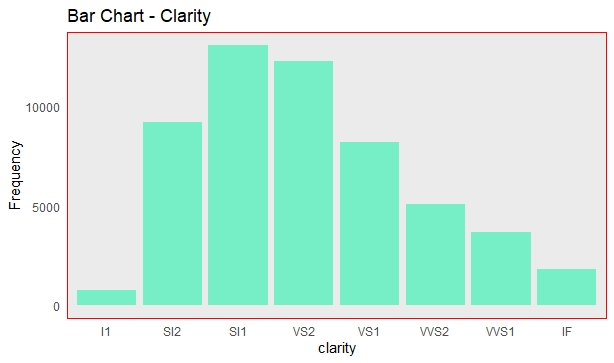
ggtitle("Bar Chart - color") +

xlab("color") + ylab("Frequency")+

theme(panel.background = (element\_rect(color="red")),(element\_blank()))

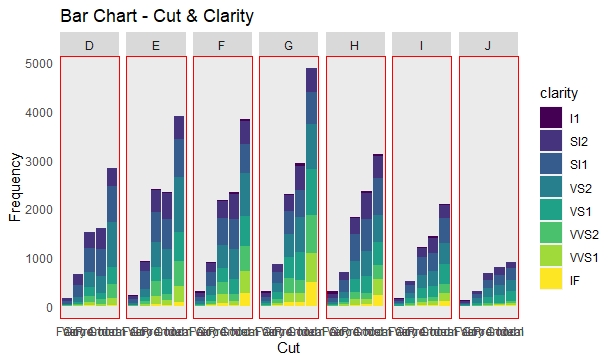






These bar charts indicate that:

* *Most of the diamonds are cut with ideal category followed by premium and very good.*
* *SI1,SI2 and VS2 are the most populated clarity categories.*
* *Color E,F,G are most common.*

We can also infer the same results using facet and filling:

**Line Chart – Price Vs Carat:**

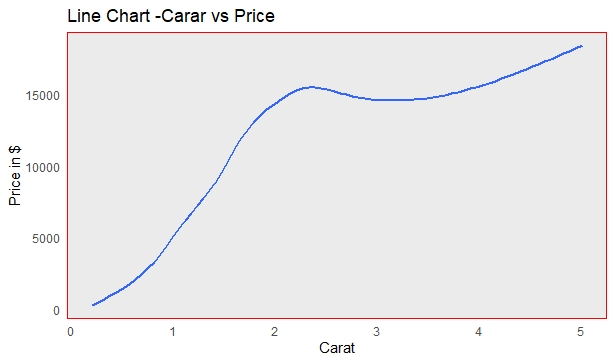
Let’s use line charts to identify the relation between carat and the price:

ggplot(data=diamonds, aes(x=carat,y=price))+geom\_smooth(se=F)+

ggtitle("Line Chart -Carar vs Price") +

xlab("Carat") + ylab("Price in $")+

theme(panel.background = (element\_rect(color="red")),(element\_blank()))



*We can see a linear relationship between weight of diamond and its price. However, once diamond weight exceeds 2-2.25 Carta weight, the line gets bended with a flatter progression.*

To explore its reason, let’s split the line chart on the three categories of cut, clarity and color:

ggplot(data=diamonds, aes(x=carat,y=price,fill=color,color=color))+geom\_smooth(se=F)+

ggtitle("Line Chart -Carar vs Price - class Color") +

xlab("Carat") + ylab("Price in $")+

theme(panel.background = (element\_rect(color="red")),(element\_blank()))

ggplot(data=diamonds, aes(x=carat,y=price,fill=clarity,color=clarity))+geom\_smooth(se=F)+

ggtitle("Line Chart -Carar vs Price - class Clarity") +

xlab("Carat") + ylab("Price in $")+

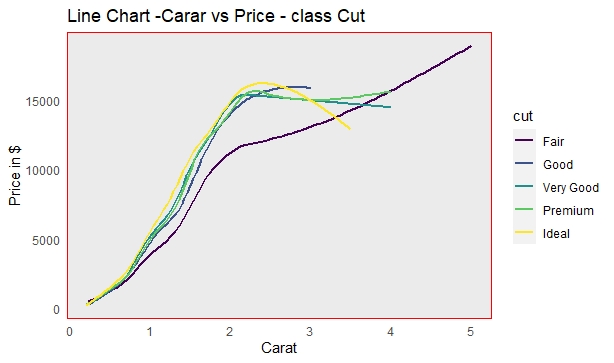
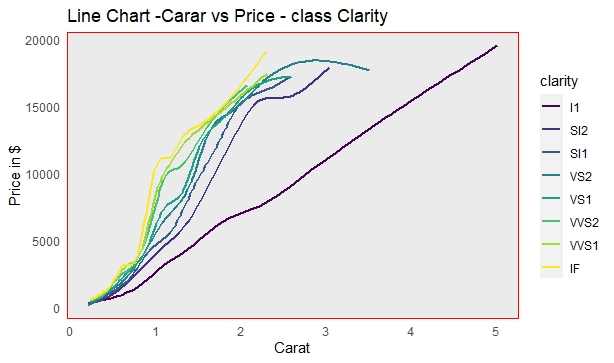
theme(panel.background = (element\_rect(color="red")),(element\_blank()))

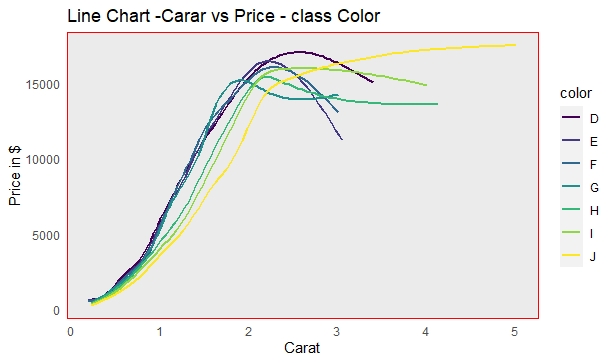
ggplot(data=diamonds, aes(x=carat,y=price,fill=cut,color=cut))+geom\_smooth(se=F)+

ggtitle("Line Chart -Carar vs Price - class Cut") +

xlab("Carat") + ylab("Price in $")+

theme(panel.background = (element\_rect(color="red")),(element\_blank()))





*We see in these graphs that color and clarity are playing vital role in bending the carat vs price curve. Interestingly in most of the case, prices increase starts to slow down when the weight of diamond increase from 2.5 carat.*

**#Scattered & Line Chart – Carat vs Price – Class Cut**

*One more interesting fact which was revealed in last section was,, heavy weight diamonds are mostly fair cut instead of superior quality cut of ideal or premium.*

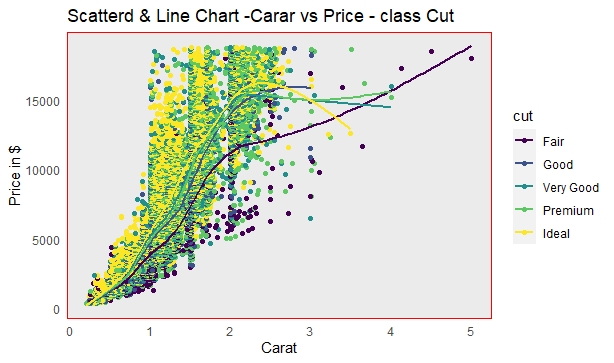
To reinforce this point, we can draw scattered & line chart:

ggplot(data=diamonds, aes(x=carat,y=price,col=cut))+geom\_point()+geom\_smooth(se=F)+

ggtitle("Scatterd & Line Chart -Carar vs Price - class Cut") +

xlab("Carat") + ylab("Price in $")+

theme(panel.background = (element\_rect(color="red")),(element\_blank()))



*Here it’s clear that, most of the diamonds weight less than 3 carats. The ones with more than 4 carat weight are only fair cut diamonds although fewer in quantity.*