Homework 2 - Data Visualization

Rodrigo De Luna Lara

September 15, 2016

1 mpg Dataset Analysis

It was requested to describe the relationship between highway mpg and car manufacturer using the mpg dataset.For this, a box plot was used grouping the data by manufacturer. As can be seen in Figure 1, the manufacturer with the highest highway mpg is Honda, showing almost complete separation from the rest of the manufacturers with an efficiency between 32 and 34 miles per gallon.

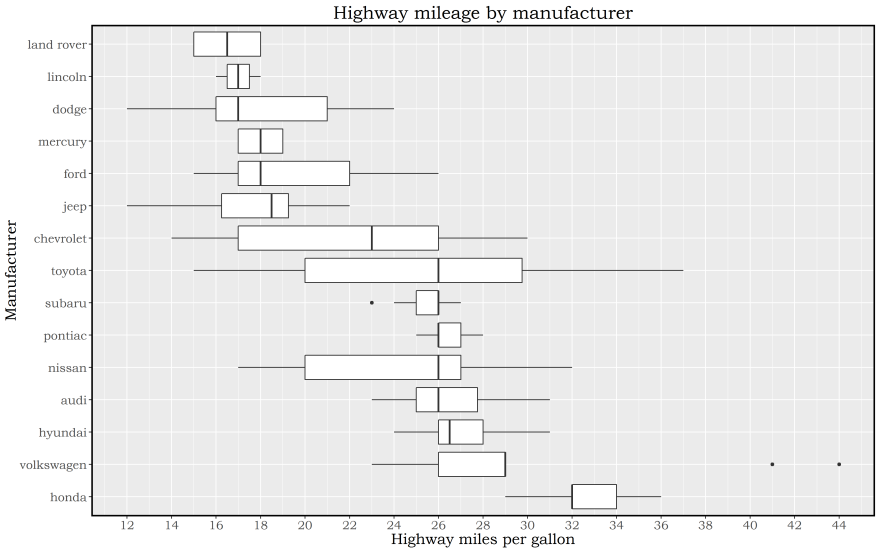

Figure 1: Box Plots for Highway mpg by Manufacturer

The least efficient vehicles are the ones manufactured by Land Rover, having between 15 and 18 miles per gallon. It is also interesting to note that the manufacturer with the highest spread is Toyota, closely followed by Chevrolet. On the other hand, Lincoln, Subaru and Pontiac have the least spread, having an IQR of 1 mile per gallon.

It can also be observed that Volkswagen has a pair of extremely efficient cars, at 41 and 44 miles per gallon; however, they are outliers from the Volkswagen population and the third quartile for the Volkswagen box is 3 units away from the first quartile of the Honda box, so Honda prevails as the manufacturer with the most efficient cars in highway.

Next it was requested to describe the three-way relationship between city efficiency, highway efficiency and the vehicle class. To do this the individual relationships between each efficiency and the vehicle class were plotted in box plots (Figures 2 and 3).

Both plots are ordered by decreasing median, so it can be seen that in both cases the compact class has the best efficiency, while SUVs and pickups have the worst overall performance. This is expected due to the fact that SUVs and pickups tend to have more cylinders and are heavier than compact cars.

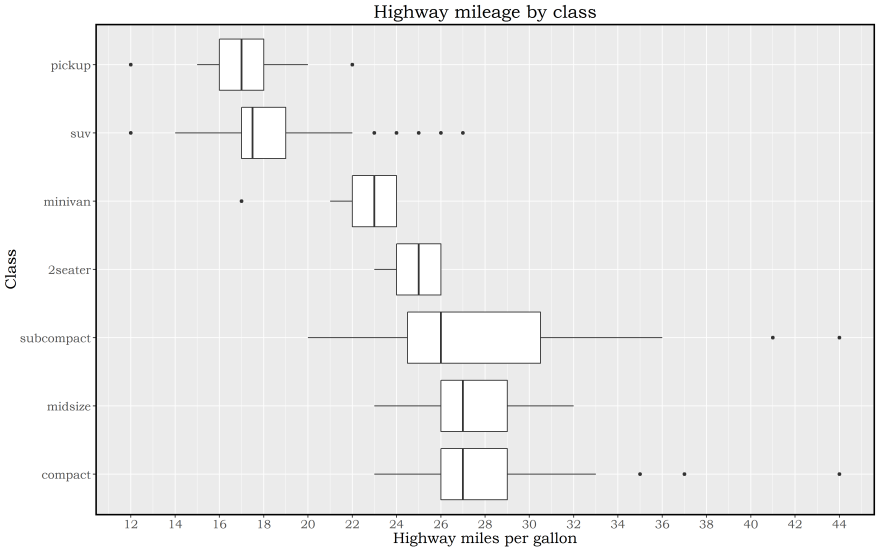

Figure 2: Box Plots for Highway mpg by Vehicle Class

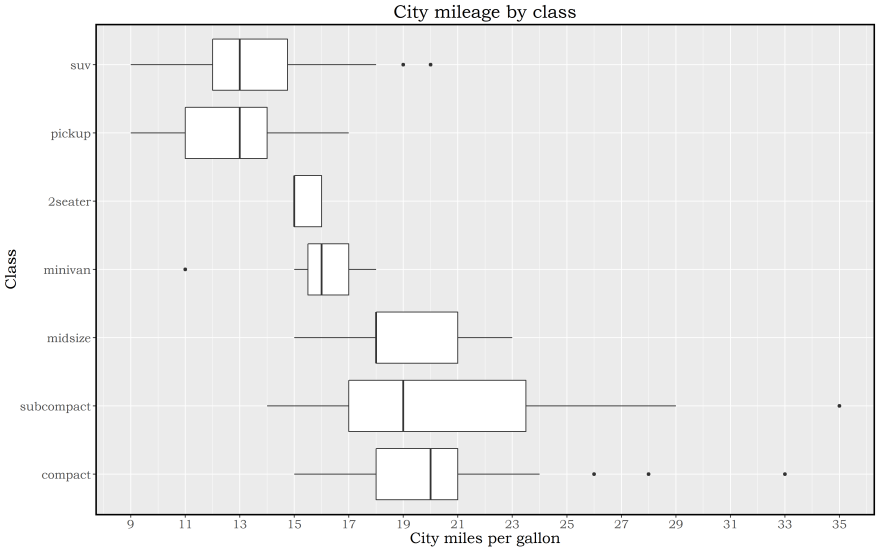

Figure 3: Box Plots for City mpg by Vehicle Class

To analyze the three way relationship, a plot of city mpg vs highway mpg was created, with symbols and colors indicating the class (Figure 4). This plot shows an expected trend; highway efficiency is better than city efficiency by several units and the separation between them is dependent on the class of the vehicle. Heavier vehicles, or vehicles with more cylinders, like SUVs and pickups have significantly less improvement than lighter, more efficient vehicles like compacts and subcompacts.

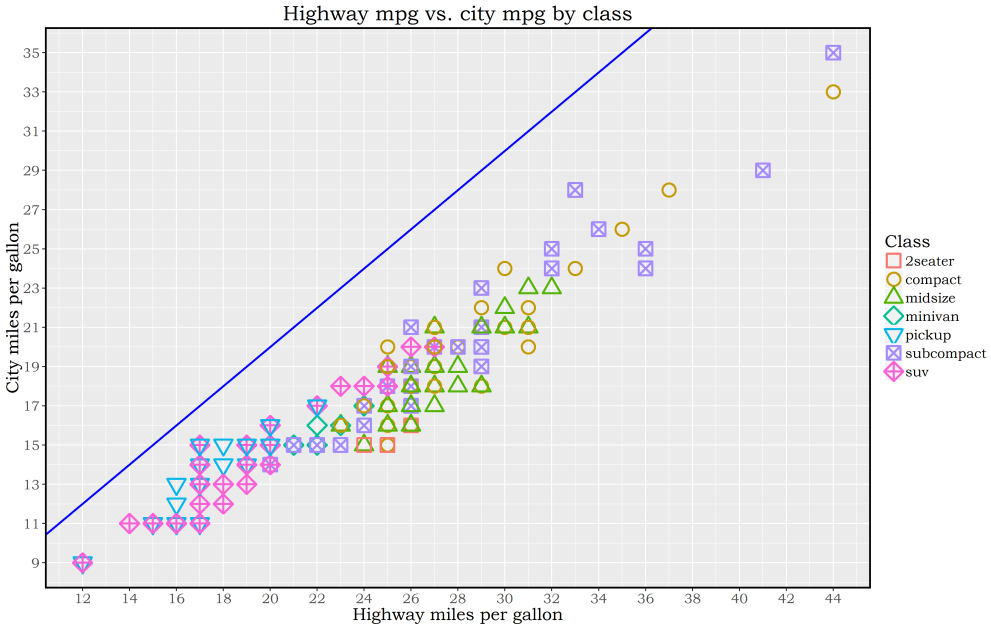

Figure 4: Three Way Relationship Between City/Highway Efficiency and Class

The difference in efficiency between subcompacts and compacts on one end and SUVs and pickups on the other end is quite clear from this plot, which is in agreement with the box plots in Figures 2 and 3. Minivans, midsize cars and 2 seaters are in between the extremes of the aforementioned groups, having decent improvement of highway mileage versus city mileage with regards to the most and least efficient vehicles.

The relationship is clear, highway efficiency is always higher than city efficiency irrespective of the class of the vehicle. However, the class does determine the expected improvement between both. The plot clearly shows higher margins for a significant part of the compacts and subcompacts. Finally, the median highway and city miles per gallon were plotted by class, as shown in Figure 5, to confirm what was seen in the previous plot. Compact cars are on top with the best highway and city fuel efficiency, while pickups are at the bottom with the worst overall efficiency.

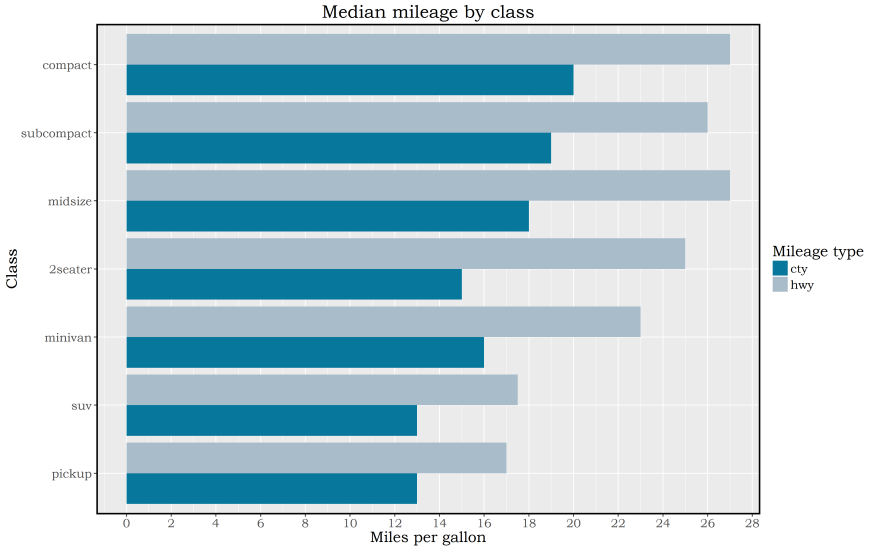

Figure 5: Three Way Relationship Between City/Highway Efficiency and Class

Figure 5 allows for easier visualization of the difference between highway mpg and city mpg for each class. Figure 4 allowed to see the overall trend for the three-way relationship of these 2 parameters and the class, but Figure 5 gives additional insights on the median efficiency. For example, it can be seen that the class with the highest difference is the 2-seater class with a difference of 10 mpg, which was not easily detectable in Figure 4.

2 Discussion on Histogram vs Boxplots

Whether to choose a histogram or a box plot depends on what is required to analyze. Both of them are helpful but only when used in the appropriate way.

For a single variable or class the histogram provides more detail on the distribution of the data, its number of modes, its skewness and its kurtosis, giving precise statistical measures that describe the data. A boxplot of a single variable allows for easier visualization of outliers and quartiles, giving easy to interpret measures on the distribution of the data.

Therefore, for a single variable, the choice depends on the extent of the analysis that wants to be done, histograms are much more complex, give significantly more information but are sometimes harder to understand visually, they are also extremely dependent on the bin width, an incorrect bin width can alter completely the interpretation of the data; boxplots on the other hand are quick and easy to understand, have no tunable parameters and are easy to make but have the drawback of giving only an overview of the distribution of the data.

For several classes or variables, histograms are definitely not a good option when the number is relatively large. Stacking histogram of up to 4 variables is possible, but the visualization becomes hard after that point. Boxplots are better for large number of classes or variables because they are easier to stack without losing information or making it harder to analyze. For example, if Figure 1 were to be of histograms instead of boxplots it would look like this:

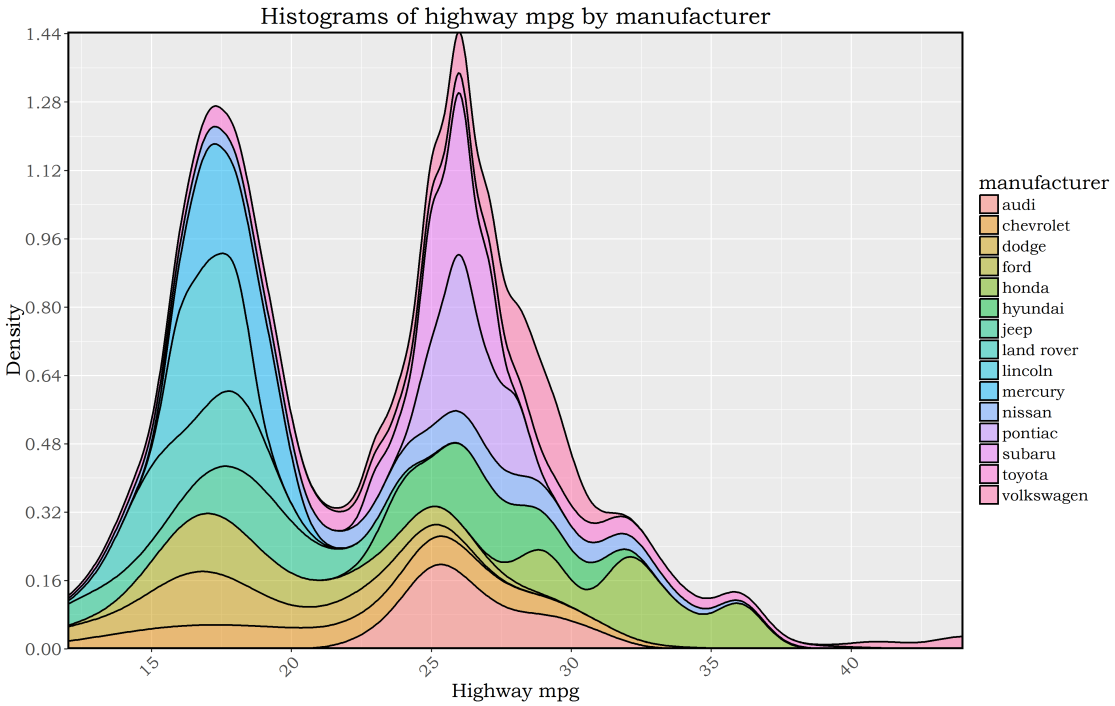

Figure 6: Example of Stacked Histogram Density Plots

Which is certainly not easy to interpret as the boxplots, given the complexity involved in presenting the histograms in an understandable manner. On the other hand, for a single variable a histogram is much more informative:
So in summary, for single variables a histogram may be more valuable than a boxplot, and for multiple variables boxplots can be easier to interpret, although histograms could be used up to certain number of variables before it's very difficult to get information from them, as seen in Figure 7, in which case it's possible to see that the data is bimodal from the histogram, information that the boxplot lacks to provide.

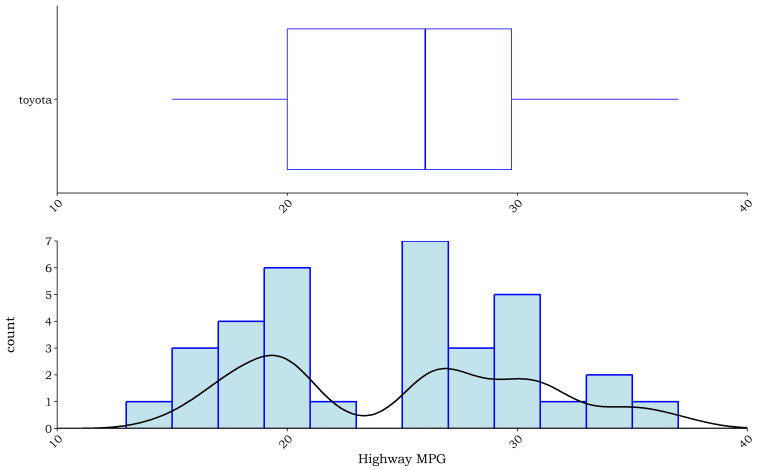

Figure 7: Example of Histogram vs Boxplot for a Single Variable

3 Plot Image File Size Analysis

It was requested to generate 2 sets of random numbers with the *runif* function, plot them in a scatter plot and save the plot with 4 different file extensions: PostScript (.ps), Portable Document Format (.pdf), Portable Network Graphics (.png) and Joint Photographic Experts Group Format (.jpeg). This was done iteratively for 20 points from $N = 0$ to $N = 100000$ (almost to the point of saturation of the plot), saving each plot to file and getting the file size automatically.

The resulting file sizes were plotted, as shown in Figure 8.

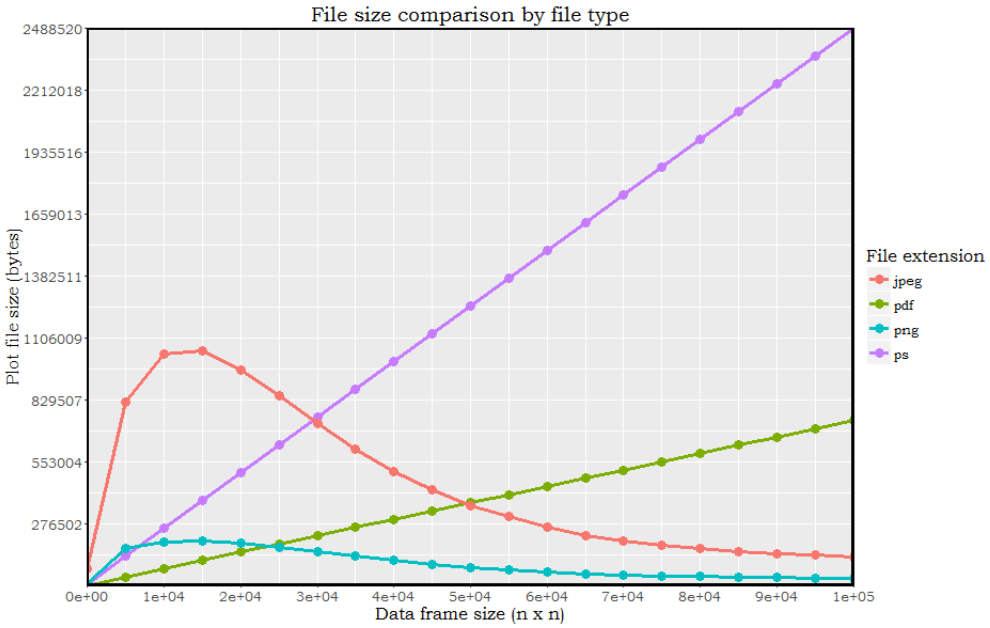

Figure 8: File Size Trend by File Format

It can be seen that for both the *.pdf* and *.ps* file formats the file size is a linear function of the size of the data frame, in which the slope for the *.pdf* format is significantly lower than the slope of the *.ps* format, which is expected as they both involve some vectorization of the image for visualization purposes, i.e. they are not meant for image compression.

On the other hand, both *.png* and *.jpeg* extensions show an interesting positively-skewed bell curve. The file size has a maximum point and then it starts to decay to an asymptote. Even when the size of the data frame used in the plot increases by several orders of magnitude, the file size remains almost constant. This can be explained with the image compression both of these formats use, as more points are added to the scatter plot, it starts to resemble the image of a black square, which is easier to compress than the image of several scattered points.

4 Diamonds Dataset Analysis

With the diamonds dataset included in the ggplot package in R, it was requested to plot histograms for color, carat and price, commenting on their shapes. It was also requested to analyze the three-way relationship between carat, cut and price and provide insights on their relationship. Figure 9 shows the requested histograms for the first part, with the addition of the histogram for the *cut* variable.

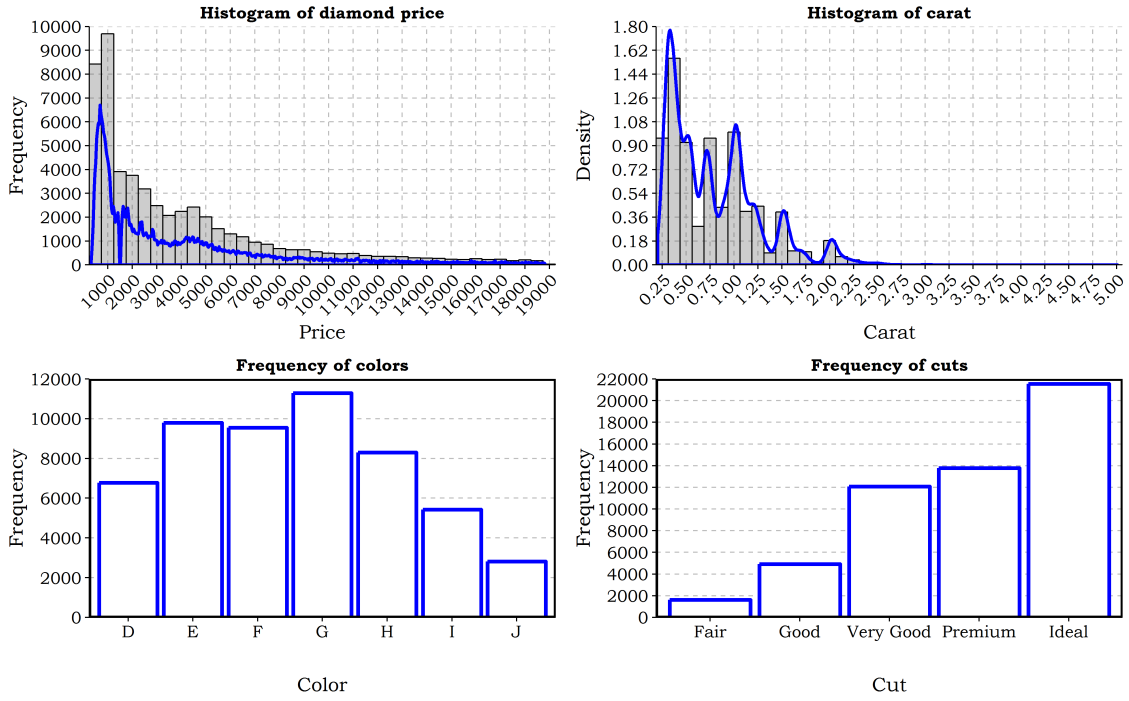

Figure 9: Selected Histograms

It can be seen that for the price the distribution could be seen as bimodal, having a first mode up to 1500 and a second mode at around 2000; however, it looks more like an inverted Gauss distribution and the appearance of 2 modes could be due to the choice of bin width.

The distribution for the carat shows that most of the values lie below 2.5 carats, diamonds larger than are extremely uncommon as evidenced by the distribution. Carats larger than 2.5 do exist, the maximum carat in the dataset is of 5.01, but there are only a handful of diamonds between that maximum and the value of 2.5 shown in the histogram.

The frequency histogram of the colors show an even distribution of colors, except for the J color, which is significantly less frequent than the rest of the colors. Color D corresponds to a colorless diamond, while the range from $G - J$ indicates near-colorless diamonds. The rarest color is therefore color J, which has the least clear color.

The frequency histogram for cuts show that surprisingly the majority of diamonds have more than *Good* cut, the mode being the $Ideal$ cut and $Fair$ diamonds being the least common. The histogram could easily fit a half-normal distribution, with the values tending towards an $Ideal$ cut.

To analyze the requested three-way relationship, the plots in Figure 10 were produced to analyze the estimated effect of cut, color and clarity on the price of the diamond as the carat increases.
The plots show that the carat is one of the main drivers of the diamond price, as most trend lines show increasing price with increasing carat. Of the three factors used (cut, color and clarity), the one that drives the price up more steeply with respect to carat size is clarity.

The clarity of the diamond seems to be limited by the carat, the largest diamonds (carat > 3.5) have only a clarity rating of $I1$, which indicates that there are clearly visible inclusions in the diamond. For diamonds between 2.5 and 3.5 carats two additional clarity ratings are available, $SI1$ and $SI2$, which indicate that the diamond is slightly included. For diamonds smaller than 2.5 carats the rest of the clarity ratings become available, with each improvement in clarity resulting in a different limit to the carat of the diamond.

Better clarity ratings also carry an increase in price, which is very evident between the curves for the worst clarity ($I1$, diamonds with inclusions) and the best clarity (IF, internally flawless diamonds)

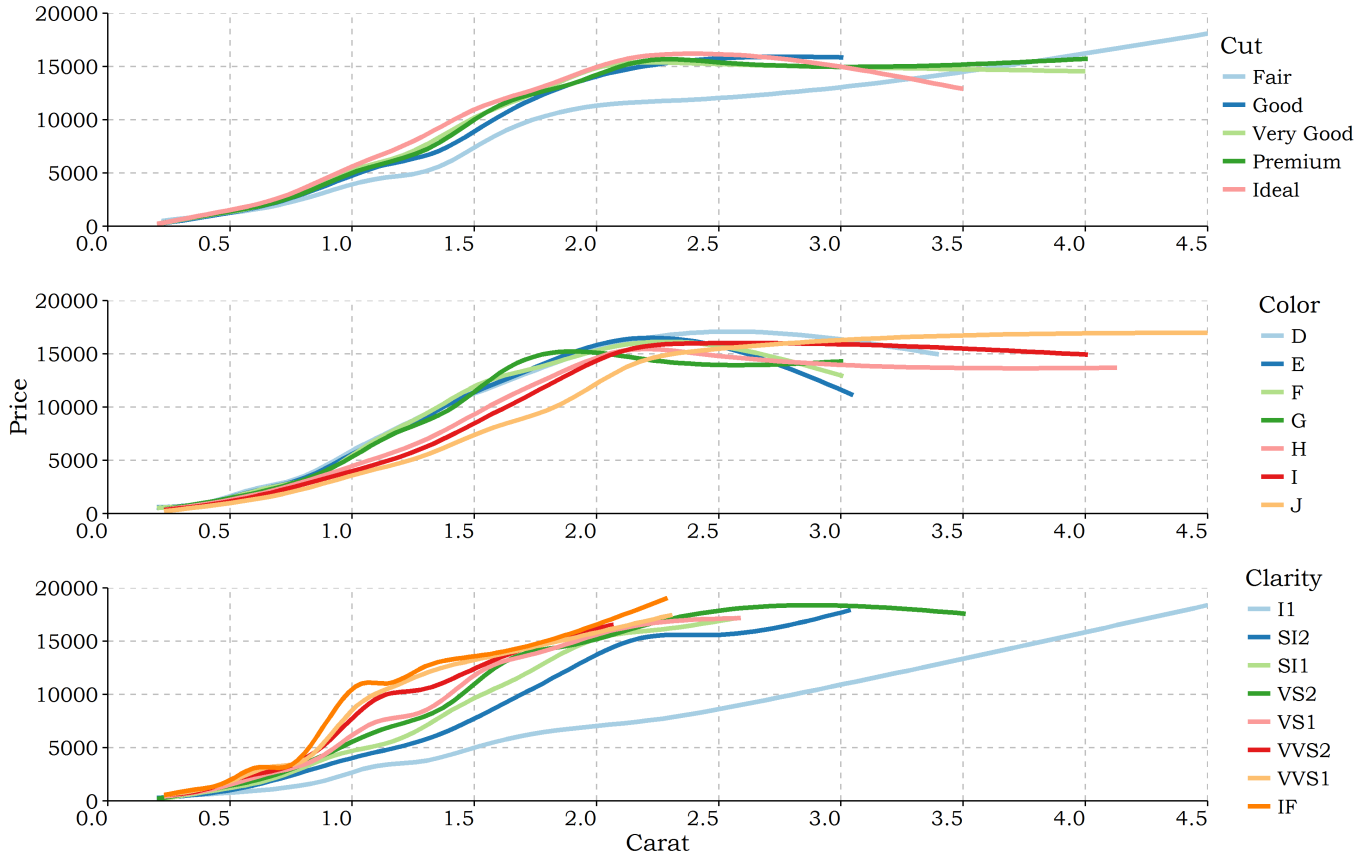

Figure 10: Price vs Carat by Cut, Color and Clarity

As for the relationship with the color, the opposite trend from the clarity is observed. The color of the diamond seems to be also limited by the carat, colorless diamonds (Dcolor) tend to be smaller than near colorless diamonds (J color). There is also separation between curves for the color, although it seems to have less weight on the price.

The trends for diamonds with D through G colors and with H and I colors almost overlap for carats < 2.0, showing significant separation only for diamonds between 2.0 and 3.0 carats. The difference in price between the best and worst colors ranges between 2500 and 5000; in contrast, the price difference between the best and worst clarity ranges between 5000 and 7500, showing that the clarity has a higher effect on the price than the color.

Finally, it is clear that the cut has no significant effect on the price, except for Fair cut diamonds. For Good to Ideal cut diamonds the difference between trends is almost negligible. The trends by cut also have a correlation with the carat, being cut off a maximum carat.

Overall, looking at the plots it can be concluded that the clarity has the most effect on the price of the diamond, followed by the color. The cut doesn't seem to have a significant effect. Considering all of these variables it can be said that the main driver for the cost of the diamonds is the carat, followed by the clarity.

The carat is the main driving factor because larger carat diamonds are extremely rare, as was evidenced by the histogram of the carat in Figure 9. The rarer the diamond, the more expensive it is expected to be, regardless of color, cut and clarity. The largest diamonds, between 4.0 and 5.0 carats are restricted to fair cut, near-colorless color (J) and a clarity with clear inclusions ($I1$). Despite having the lowest factors of the cut, color and clarity categories they are still the most expensive diamonds, providing evidence towards the assertion that the main driving factor on the cost is the carat.

Knowing that the cut is not very significant, a plot of price vs carat was created for each combination of color and clarity, to provide further support to the previous claims.

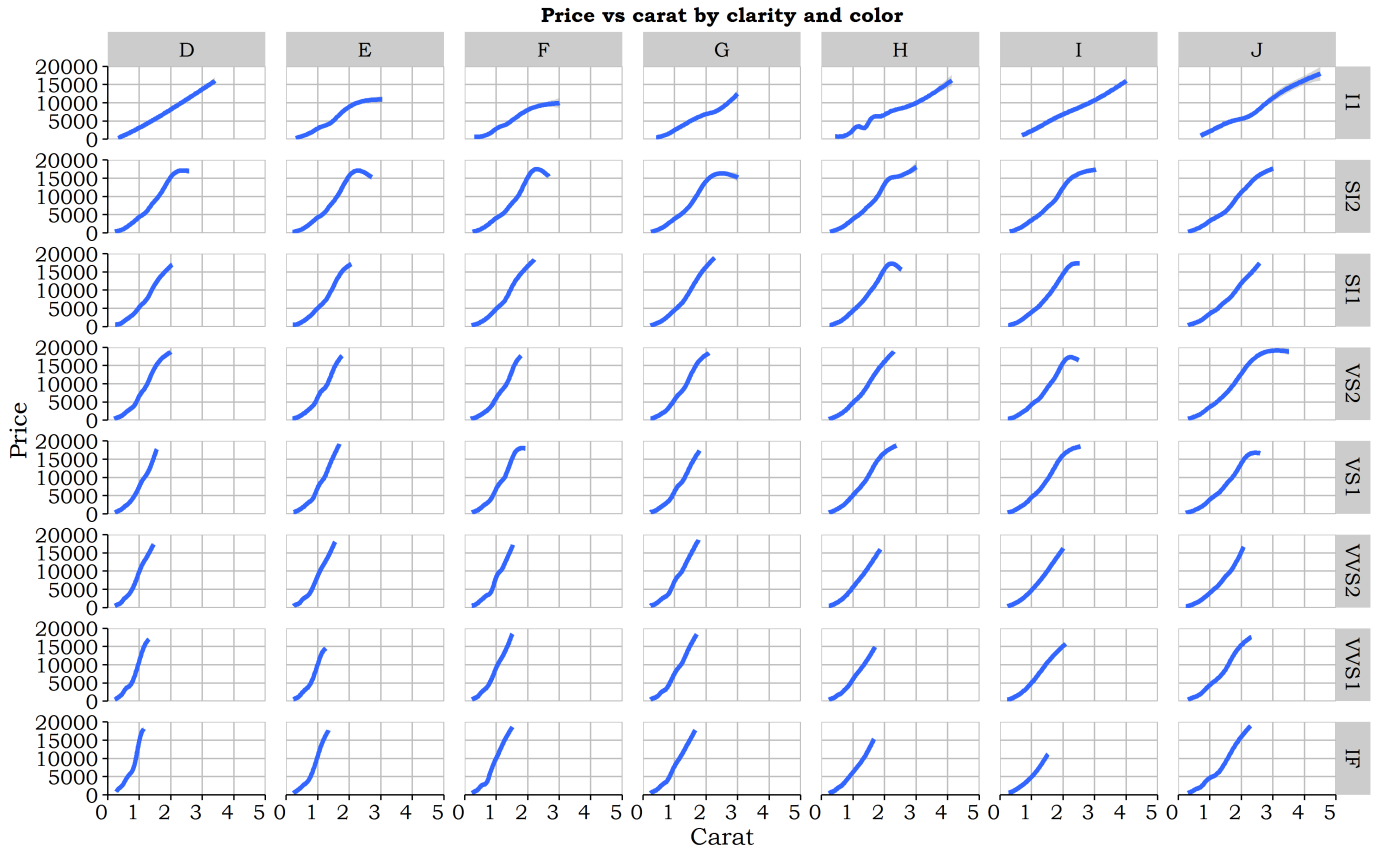

Figure 11: Plots of Price vs Carat by Clarity and Color

It can be appreciated that the slopes for the regressions are very uniform, showing decay in the slope with decreasing quality in a diagonal way (from left to right and bottom to top). The carat limit of each combination is also evident, the maximum carat is increased with the decay in the slope of the regressions in the same direction. The maximum price also increases in the same direction, providing support to the claim that the main driver of the price is the carat size, even with decreasing quality in color and clarity.

The plot can be summarized briefly as follows: with decreasing clarity and color the maximum allowable carat is increased, driving the prices up in the same direction, indicating that the main factor of the price of diamonds is the carat.

5 R Code

5.1 Code for Figure 1

```
plot1 = ggplot(mpg, aes(reorder(manufacturer, -hwy, median),hwy)) +
  geom_boxplot() +
  coord_flip() +
  scale_x_discrete() +
  scale_y_continuous(breaks = round(seq(min(mpg$hwy), max(mpg$hwy), by = 2))) +
  labs(title="Highway mileage by manufacturer", x = "Manufacturer", y = "Highway miles per gallon") +
  theme(text = element_text(size=16, family="Bookman"),
        panel.border = element_rect(colour = "black", size=1.5, fill=NA))
#print(plot1)
ggsave("Plot1.png",plot=plot1)
```

5.2 Code for Figure 2

```
plot2 = ggplot(mpg, aes(reorder(class, -hwy, median), hwy)) +
  geom_boxplot() +
  coord_flip() +
  scale_x_discrete() +
  scale_y_continuous(breaks = round(seq(min(mpg$hwy), max(mpg$hwy), by = 2))) +
  labs(title="Highway mileage by class", x = "Class", y = "Highway miles per gallon") +
  theme(text = element_text(size=16, family="Bookman"),
        panel.border = element_rect(colour = "black", size=1.5, fill=NA))
#print(plot2)
ggsave("Plot2.png",plot=plot2)
```

5.3 Code for Figure 3

```
plot3 = ggplot(mpg, aes(reorder(class, -cty, median), cty)) +
  geom_boxplot() +
  coord_flip() +
  scale_x_discrete() +
  scale_y_continuous(breaks = round(seq(min(mpg$cty), max(mpg$cty), by = 2))) +
  labs(title="City mileage by class", x = "Class", y = "City miles per gallon") +
  theme(text = element_text(size=16, family="Bookman"),
        panel.border = element_rect(colour = "black", size=1.5, fill=NA))
#print(plot3)
ggsave("Plot3.png",plot=plot3)
```

5.4 Code for Figure 4

```
plot4 = qplot(x = hwy, y = cty , data = mpg ,
              main = "Highway mpg vs. city mpg by class",
              pch = mpg$class,
              colour = mpg$class,
              size=I(5),
              stroke=1.5) +
  scale_shape_manual(values = c(0,1,2,5,6,7,9)) +
  scale_x_continuous(breaks = round(seq(min(mpg$hwy), max(mpg$hwy), by = 2))) +
  scale_y_continuous(breaks = round(seq(min(mpg$cty), max(mpg$cty), by = 2))) +
  labs(x = "Highway miles per gallon", y = "City miles per gallon", pch="Class", color="Class") +
  geom_abline(slope=1,intercept=0,color="blue",size=1) +
  theme(text = element_text(size=16, family="Bookman"),
        panel.border = element_rect(colour = "black", size=1.5, fill=NA))
#print(plot4)
ggsave("Plot4.png",plot=plot4)
```

5.5 Code for Figure 5

```
library(reshape2)
DF = melt(aggregate(list(cty = mpg$cty,hwy = mpg$hwy),by=list(class = mpg$class),FUN=median))
plot5 = ggplot(DF,aes(reorder(DF$class,DF$value,sum),value,fill=variable)) +
  geom_bar(stat="identity",position = "dodge") +
  coord_flip() +
  scale_fill_manual(values = c("#07779C","#A9BCCA")) +
  scale_y_continuous(breaks = round(seq(0, max(mpg$hwy), by = 2))) +
  labs(title="Median mileage by class", x = "Class", y = "Miles per gallon", fill="Mileage type") +
  theme(text = element_text(size=16, family="Bookman"),
        panel.border = element_rect(colour = "black", size=1.5, fill=NA))
#print(plot5)
ggsave("Plot5.png",plot=plot5)
```

5.6 Code for Figure 6

```
plot6 = ggplot(mpg,aes(x=mpg$hwy,y=..density..,fill=manufacturer)) +
  geom_density(alpha=0.5,position="stack",size=0.75) +
  scale_x_continuous(expand = c(0,0),breaks = (seq(0,50,by=5))) +
  scale_y_continuous(expand = c(0,0),breaks = (seq(0,1.6,length.out = 11))) +
  labs(title = "Histograms of highway mpg by manufacturer",
       x = "Highway mpg",
       y = "Density") +
  theme(text = element_text(size=16, family="Bookman"),
        panel.border = element_rect(colour = "black", size=1.5, fill=NA),
        axis.text.x = element_text(angle = 45,size=14,hjust=1),
        axis.text.y = element_text(size=14))
#print(plot6)
ggsave("Plot6.png",plot)
```

5.7 Code for Figure 7

```
DF = mpg[which(mpg$manufacturer=="toyota"),]
plot7_1 = ggplot(DF,aes(manufacturer,hwy)) + geom_boxplot(color="blue") + coord_flip() +
  scale_x_discrete(expand = c(0,0)) +
  scale_y_continuous(expand = c(0,0),limit=c(10,40)) +
  labs(y = "",x ="") +
  theme(text = element_text(size=16, family="Bookman"),
        panel.border = element_rect(colour = "black", size=1.5, fill=NA),
        axis.text.x = element_text(angle = 45,size=14,hjust=1),
        axis.text.y = element_text(size=14))

plot7_2 = ggplot(DF,aes(x = hwy, y = ..count..)) +
  geom_histogram(color="blue",fill="light blue",size=1,alpha=0.75,binwidth = 2) +
  geom_density(size=1,color="black",adjust=1/2) +
  scale_y_continuous(expand = c(0,0),breaks=seq(0,7,by=1)) +
  scale_x_continuous(expand = c(0,0),limit=c(10,40)) +
  labs(x = "Highway MPG") +
  theme(text = element_text(size=16, family="Bookman"),
        panel.border = element_rect(colour = "black", size=1.5, fill=NA),
        axis.text.x = element_text(angle = 45,size=14,hjust=1),
        axis.text.y = element_text(size=14))

require(cowplot)
plot7 = plot_grid(plot7_1,plot7_2,align='vh',nrow=2)
#print(plot7)
ggsave("Plot7.png",plot7)
```

5.8 Code for Figure 8

```
n.sample = seq(0,100000,length.out = 21)

ps.size = pdf.size = jpeg.size = png.size = rep(0,length(n.sample))

for (i in 1:length(n.sample)){
  x = runif(n.sample[i])
  y = runif(n.sample[i])
  plot = qplot(x=x,y=y)

  ggsave("plot.ps",plot=plot)
  ps.size[i] = file.size("plot.ps")

  ggsave("plot.pdf",plot=plot)
  pdf.size[i] = file.size("plot.pdf")

  ggsave("plot.jpeg",plot=plot)
  jpeg.size[i] = file.size("plot.jpeg")

  ggsave("plot.png",plot=plot)
  png.size[i] = file.size("plot.png")

}

df = data.frame(N = n.sample,
                ps = ps.size,
                pdf = pdf.size,
                jpeg = jpeg.size,
                png = png.size)

plot8 = ggplot(df, aes(N, y = value, color = variable)) +
  geom_point(aes(y = ps, col = "ps"),size=4) +
  geom_line(aes(y = ps, col = "ps"),size=1.25) +
  geom_point(aes(y = pdf, col = "pdf"),size=4) +
  geom_line(aes(y = pdf, col = "pdf"),size=1.25) +
  geom_point(aes(y = jpeg, col = "jpeg"),size=4) +
  geom_line(aes(y = jpeg, col = "jpeg"),size=1.25) +
  geom_point(aes(y = png, col = "png"),size=4) +
  geom_line(aes(y = png, col = "png"),size=1.25) +
  scale_x_continuous(expand = c(0, 0),breaks = seq(0,100000,length.out = 11)) +
  scale_y_continuous(expand = c(0, 0),breaks = round(seq(0,max(df),length.out = 10))) +
  labs(title = "File size comparison by file type",
       x = "Data frame size (n x n)",
       y = "Plot file size (bytes)",
       color = "File extension") +
  theme(text = element_text(size=16, family="Bookman"),
        panel.border = element_rect(colour = "black", size=1.5, fill=NA))
ggsave("format_size.png",plot8)
```

5.9 Code for Figure 9

```
hist1 = ggplot(diamonds,aes(x = price, y = ..count..)) +
  geom_histogram(alpha=0.3, binwidth = 500, colour="black") +
  geom_density(adjust=1/5000,size=1.25,color="blue") +
  scale_x_continuous(expand = c(0,0),breaks = (seq(0,20000,by=1000))) +
  scale_y_continuous(expand = c(0,0),breaks = (seq(0,10000,by=1000)),limits=c(0,10000)) +
  labs(title = "Histogram of diamond price",
       x = "Price",
       y = "Frequency") +
  theme(text = element_text(size=16, family="Bookman"),
        panel.border = element_rect(colour = "black", size=1.5, fill=NA),
        axis.text.x = element_text(angle = 45,size=14,hjust=1),
        axis.text.y = element_text(size=14),
        panel.grid.major = element_line(linetype="dashed",color="grey"))
#print(hist1)

hist2 = ggplot(diamonds,aes(x = carat, y = ..density..)) +
  geom_histogram(alpha=0.3, binwidth = 0.125, colour="black") +
  geom_density(size = 1.25, color = "blue") +
  scale_x_continuous(expand = c(0,0),breaks = (seq(0,5,by=0.25))) +
  scale_y_continuous(expand = c(0,0),breaks = (seq(0,1.8,length.out = 11)),limits=c(0,1.8)) +
  labs(title = "Histogram of carat",
       x = "Carat",
       y = "Density") +
  theme(text = element_text(size=16, family="Bookman"),
        panel.border = element_rect(colour = "black", size=1.5, fill=NA),
        axis.text.x = element_text(angle = 45,size=14,hjust=1),
        axis.text.y = element_text(size=14),
        panel.grid.major = element_line(linetype="dashed",color="grey"))
#print(hist2)

hist3 = ggplot(diamonds,aes(color)) +
  geom_bar(fill="white",size=1.5,color="blue") +
  scale_y_continuous(expand = c(0,0),breaks = round(seq(0,13000,by=2000)), limits=c(0,12000)) +
  labs(title = "Frequency of colors",
       x = "Color",
       y = "Frequency") +
  theme(text = element_text(size=16, family="Bookman"),
        panel.border = element_rect(colour = "black", size=1.5, fill=NA,linetype = 1),
        axis.text.x = element_text(size=14),
        axis.text.y = element_text(size=14),
        panel.grid.major = element_line(linetype="dashed",color="grey"),
        panel.grid.major.x = element_blank())
#print(hist3)

hist4 = ggplot(diamonds,aes(cut)) +
  geom_bar(fill="white",size=1.5,color="blue") +
  scale_y_continuous(expand = c(0,0),breaks = round(seq(0,22000,by=2000)), limits=c(0,22000)) +
  labs(title = "Frequency of cuts",
       x = "Cut",
       y = "Frequency") +
  theme(text = element_text(size=16, family="Bookman"),
        panel.border = element_rect(colour = "black", size=1.5, fill=NA,linetype = 1),
        axis.text.x = element_text(size=14),
        axis.text.y = element_text(size=14),
        panel.grid.major = element_line(linetype="dashed",color="grey"),
        panel.grid.major.x = element_blank())
#print(hist4)

plot9 = plot_grid(hist1,hist2,hist3,hist4,nrow=2,ncol=2,align='vh');
library("cowplot");
ggsave("Plot9.png",plot9,scale=1);
```

5.10 Code for Figure 10

```
plot10_1 = ggplot(diamonds, aes(x=carat, y=price, color=cut)) + geom_smooth(se=FALSE, size=1.5) +
  scale_y_continuous(expand=c(0,0),limits=c(0,20000),breaks=seq(0,20000,by=5000)) +
  scale_x_continuous(expand=c(0,0),limits=c(0,4.5),breaks=seq(0,4.5,by=0.5)) +
  labs(x = "", y = "",color="Cut") +
  theme(text = element_text(size=16, family="Bookman"),
        panel.border = element_rect(colour = "black", size=1.5, fill=NA),
        axis.text.x = element_text(size=14,hjust=1),
        axis.text.y = element_text(size=14),
        panel.grid.major = element_line(linetype="dashed",color="grey"),
        panel.border = element_rect(colour="black",fill=NA,linetype=1)) +
  scale_color_brewer(palette = "Paired")

plot10_2 = ggplot(diamonds, aes(x=carat, y=price, color=color)) + geom_smooth(se=FALSE, size=1.5) +
  scale_y_continuous(expand=c(0,0),limits=c(0,20000),breaks=seq(0,20000,by=5000)) +
  scale_x_continuous(expand=c(0,0),limits=c(0,4.5),breaks=seq(0,4.5,by=0.5)) +
  labs(x = "", y = "Price",color="Color") +
  theme(text = element_text(size=16, family="Bookman"),
        panel.border = element_rect(colour = "black", size=1.5, fill=NA),
        axis.text.x = element_text(size=14,hjust=1),
        axis.text.y = element_text(size=14),
        panel.grid.major = element_line(linetype="dashed",color="grey"),
        panel.border = element_rect(colour="black",fill=NA,linetype=1)) +
  scale_color_brewer(palette = "Paired")

plot10_3 = ggplot(diamonds, aes(x=carat, y=price, color=clarity)) + geom_smooth(se=FALSE, size=1.5) +
  scale_y_continuous(expand=c(0,0),limits=c(0,20000),breaks=seq(0,20000,by=5000)) +
  scale_x_continuous(expand=c(0,0),limits=c(0,4.5),breaks=seq(0,4.5,by=0.5)) +
  labs(x = "Carat", y = "",color="Clarity") +
  theme(text = element_text(size=16, family="Bookman"),
        panel.border = element_rect(colour = "black", size=1.5, fill=NA),
        axis.text.x = element_text(size=14,hjust=1),
        axis.text.y = element_text(size=14),
        panel.grid.major = element_line(linetype="dashed",color="grey"),
        panel.border = element_rect(colour="black",fill=NA,linetype=1)) +
  scale_color_brewer(palette = "Paired")


library(cowplot)
plot10 = plot_grid(plot10_1,plot10_2,plot10_3,nrow=3,ncol=1,align='vh')
#print(plot10)
ggsave("Plot10.png",plot10)
```

5.11 Code for Figure 11

```
plot11 = ggplot(diamonds, aes(x=carat, y=price))+ facet_grid(clarity~color) + geom_smooth(size=1.5)+
  scale_y_continuous(expand=c(0,0),limits=c(0,20000)) +
  scale_x_continuous(expand=c(0,0),limits=c(0,5)) +
  labs(title = "Price vs carat by clarity and color", x = "Carat", y = "Price") +
  theme(text = element_text(size=16, family="Bookman"),
        panel.border = element_rect(colour = "black", size=1.5, fill=NA),
        axis.text.x = element_text(size=14,hjust=1),
        axis.text.y = element_text(size=14),
        panel.grid.major = element_line(linetype=1,color="grey"),
        panel.border = element_rect(colour="black",fill=NA,linetype=1),
        panel.margin = unit(c(0.5),"cm"))
#print(plot11)
ggsave("Plot11.png",plot11,scale=1)
```