BUDS Training: R Visualization Lab

Andrea Dominique O. Cristobal

Using the `diamonds` data set contained in the `ggplot2` package, we will examine factors that determine the price of a diamond. Use `?diamonds` to get the definitions of the variables. Upload your knitted pdf file to EdX when you are finished.

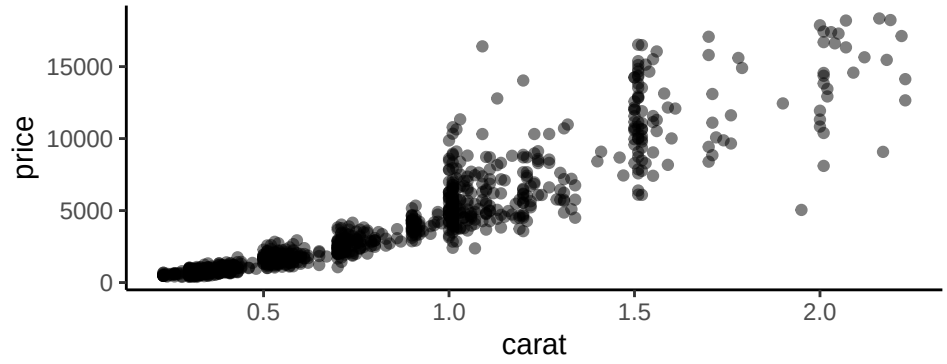
```
# Load the packages
suppressMessages(library(tidyverse))
```

```
# Take a sample of the diamond  with carats less than 2.5 to speed up plotting.
set.seed(1313)
diamond_samp <- diamonds %>%
  filter(carat <= 2.5) %>%
  sample_n(1000)
```

Problem 1

Using the `diamond_samp` data frame generate a scatterplot of diamond price versus size as measured in carats, where price is our response variable and carat is the explanatory variable. Use `theme_classic()`. Describe the nature of the relationship.

```
ggplot(data= diamond_samp, mapping = aes(x = carat, y= price))+
  geom_point(alpha=0.5)+
  theme_classic()
```



- The plot above shows that increasing carat is related to increasing price, with observations clustered in carat < 1.0.

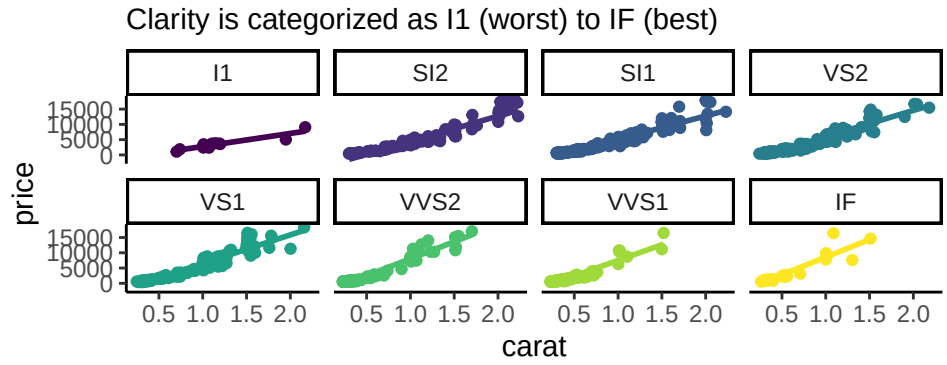
Problem 2

Examine the relationship between price and carats by clarity by creating one plot per color. Fit a straight line for each clarity level. Describe what you can learn from this plot.

```
ggplot(data= diamond_samp, mapping = aes(x = carat, y= price, color= clarity))+
  geom_point(position="jitter",show.legend=FALSE)+
  geom_smooth(method="lm",fill=NA,show.legend=FALSE)+
```

```
  facet_wrap(~clarity, nrow=2)+
  labs(
  subtitle= "Clarity is categorized as I1 (worst) to IF (best)")+
  theme_classic()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



- Even though we have established a positive relationship between price and carat, we can also use clarity as a determinant of diamond price. Steepness of the trend line varies per level of clarity, such that diamonds with better clarity tend to approach higher prices (with respect to carat) at a faster rate compared to lower clarity diamonds.

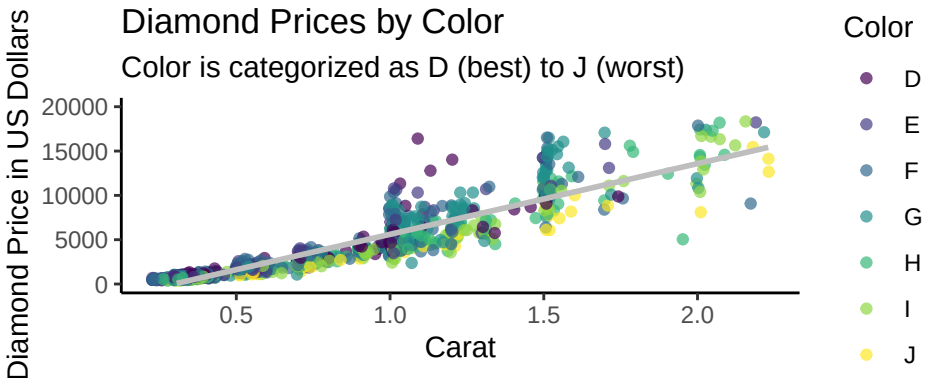
Problem 3

Create what you consider to be a really cool plot that tells an interesting story about how the different variables are related. Play with various arguments and aesthetics. Describe what insights your plot reveals. Use good axis labels, titles, legend labels, etc.

```
ggplot(data= diamond_samp, mapping = aes(x= carat, y= price))+
  geom_point(position="jitter",alpha=0.7, aes(color=color))+
  geom_smooth(method="lm",fill=NA,color="grey")+
  scale_y_continuous(limits = c(0, 20000), breaks = seq(0, 20000, 5000)) +
  labs(
    title= "Diamond Prices by Color",
    subtitle= "Color is categorized as D (best) to J (worst)",
    x= "Carat",
    y= "Diamond Price in US Dollars",
    color= "Color")+
  theme_classic()
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 3 rows containing missing values (geom_smooth).
```



- Adding another dimension to the carat and price relationship is diamond color. Diamonds with better color are priced higher, in relation to carat. Moreover, for this sample, better-colored diamonds are concentrated in lower carats. Those with relatively higher carats (>1.5) tend to have lower quality in terms of color, but with higher price still.