Assignment 1: Due Friday, September 24 at 11:59pm

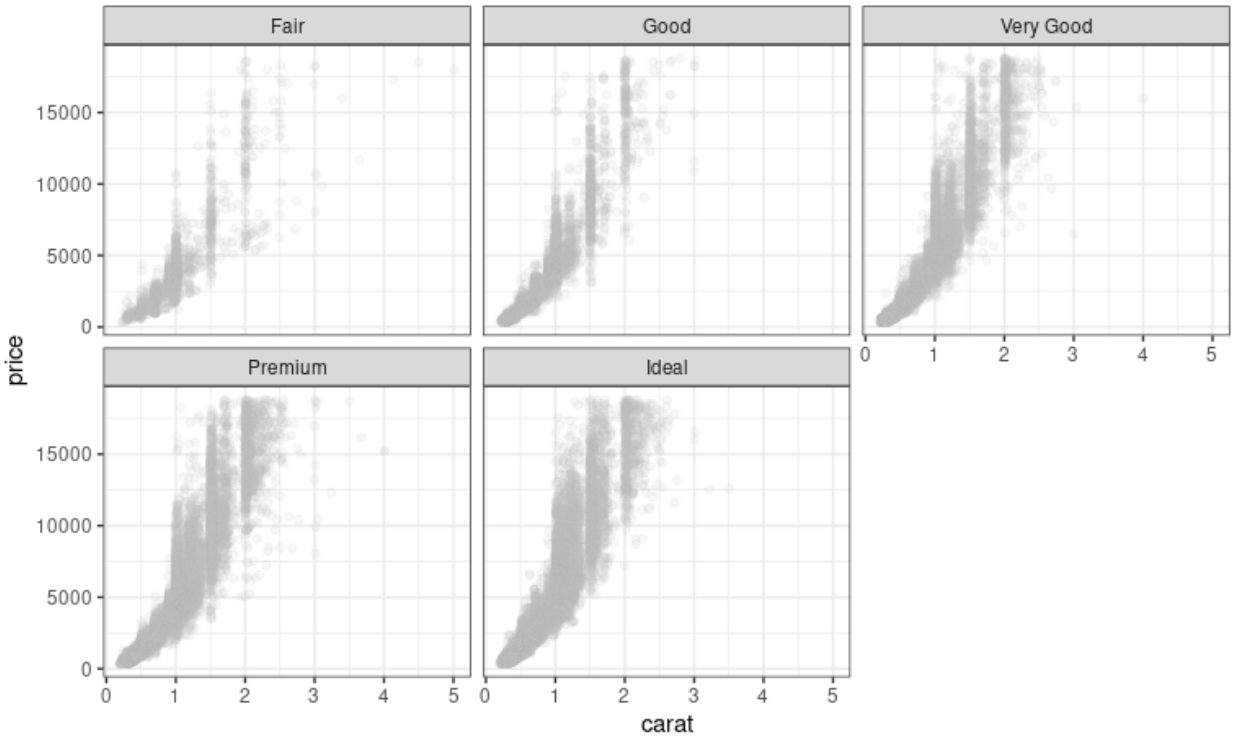
For this assignment, you will be analyzing the **`diamonds`** data frame included in the **`tidyverse`** package, which contains 53,930 observations on 10 features related to diamonds. The goal of this assignment is to become more familiar with R, and so you will need to use resources such as the **`ggplot2`** cheat sheet on Canvas and help menus in R to learn how to make different types of plots. Provide all responses in the designated spaces in this Word document, then save it as a pdf and upload it to Canvas.

1. **[10%]** Generate a jittered scatterplot of diamond price as a function of diamond weight that is broken into five sub-panels across two rows, in which each sub-panel represents a different cut quality. Set the color of the points to gray with a transparency alpha value of 0.1, and the background to white rather than the default color of gray.

   Provide code below:

```
ggplot(diamonds,aes(x=carat, y=price) ) +
  geom_jitter(color= "grey", alpha=0.1) +
  facet_wrap(~ cut, nrow= 2) +
  theme_bw()
```

   Provide figure below:



2. **[5%]** What is a rug plot, and what geometric layer (function) in **`ggplot2`** can be used to generate one?

   Provide answer below:

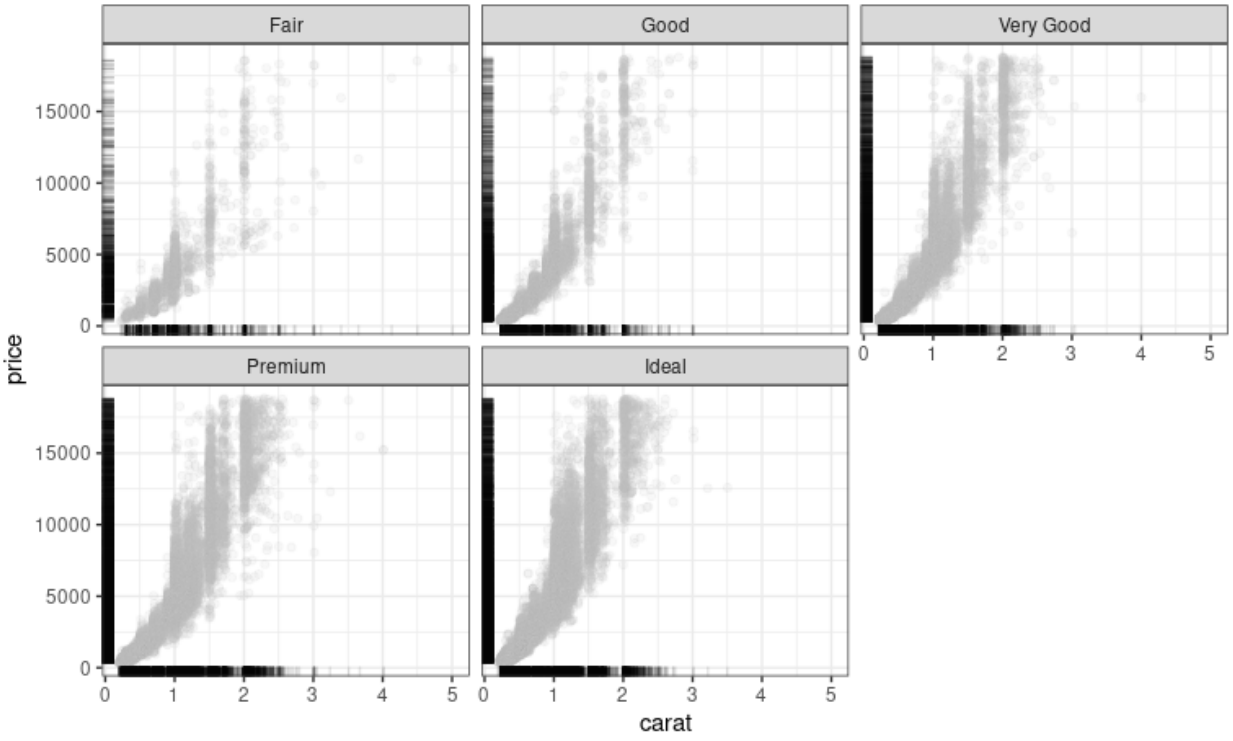
   A rug plot creates a 1D density plot and simply adds distribution marks along both the x and y axis of a graph. The function in ggplot2 to generate a rug plot is: geom_rug()

3. **[10%]** Add a rug plot to the figure from question 1.

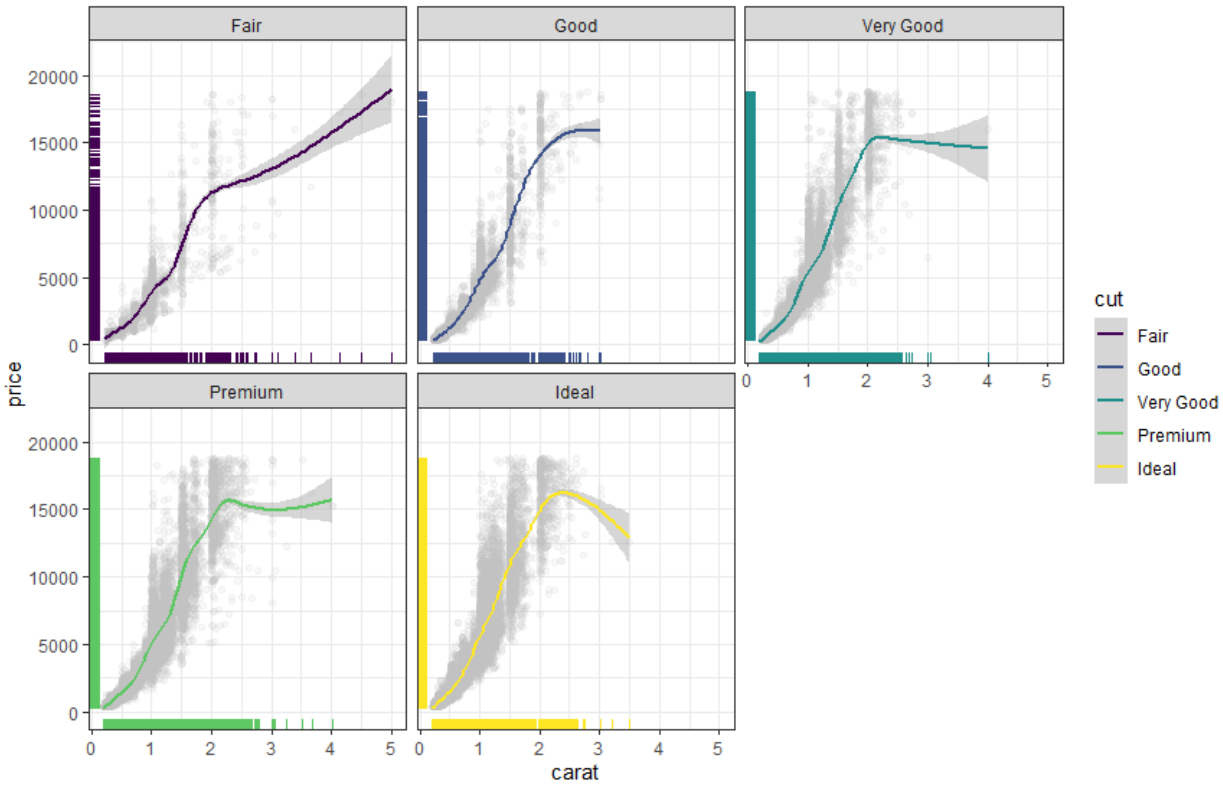
   Provide code below:

```
ggplot(diamonds,aes(x=carat, y=price) ) +
  geom_jitter(color= "grey", alpha=0.1) +
  geom_rug(color= "black", alpha=0.1) +
  facet_wrap(~ cut, nrow= 2) +
  theme_bw()
```

   Provide figure below:



4. **[20%]** Provide the code to generate the following figure, in which the bands around the fitted colored lines are 95% confidence intervals.



Provide code below:

```
ggplot(diamonds,aes(x=carat, y=price, color = cut) ) +
  geom_jitter(color= "grey", alpha=0.1) +
  geom_rug(alpha=0.1) +
  geom_smooth(level=0.95) +
  facet_wrap(~ cut, nrow= 2) +
  theme_bw()
```

5. **[5%]** In the figure from question 4, does there appear to be a relationship between diamond price and diamond weight? If there is a relationship, then what is it?

Provide answer below:

Yes, the heavier the weight (more carats) the higher the price and therefore more expensive the diamond.

6. **[5%]** In the figure from question 4, why are the confidence intervals much narrower for diamonds weighing less than three carats than for those weighing more than three carats?

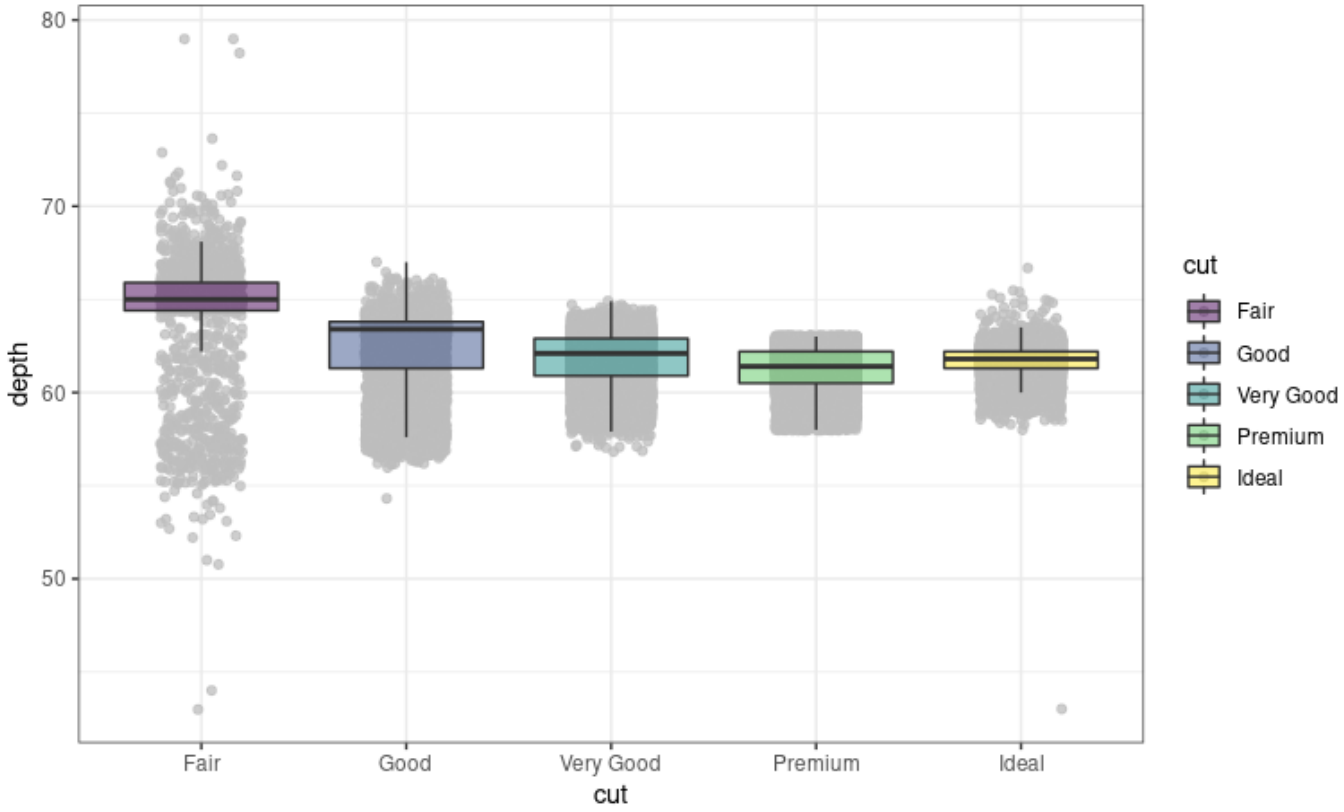
   Provide answer below:

   The confidence interval is narrower because larger sample size (more data) in that region (3 carats or less). There is more data on diamonds of 3 carats or less and therefore the confidence interval is narrower due to it being precise.

7. **[5%]** Use the code provided below to create a box plot without outliers. Note that the argument **`outlier.shape = NA`** to remove outliers from the plot.

```
ggplot(data = diamonds,
   mapping = aes(x = cut, y = depth, fill = cut)) +
 geom_jitter(width = 0.2, color = "gray", alpha = 0.75) +
 geom_boxplot(alpha = 0.5, outlier.shape = NA) +
 theme_bw()
```

   Provide figure below:



8. [5%] In your plot from question 7, the *x* axis is categorical. Therefore, what is the purpose of the following piece of code?

```
geom_jitter(width = 0.2, color = "gray", alpha = 0.75)
```

Provide answer below:

The purpose of geom_jitter, is it shows the scattered data points. If it is removed, only the box plots are shown.

9. [5%] What is a violin plot, and what geometric layer (function) in **`ggplot2`** can be used to generate one?

Provide answer below:

A violin plot is similar to a box plot however they also show the probability density of the data at different values. A violin plot creates peaks, valleys and tails to show the density/distribution of the data. The function is geom_violin.

10. [5%] How are violin plots different from box plots?

Provide answer below:

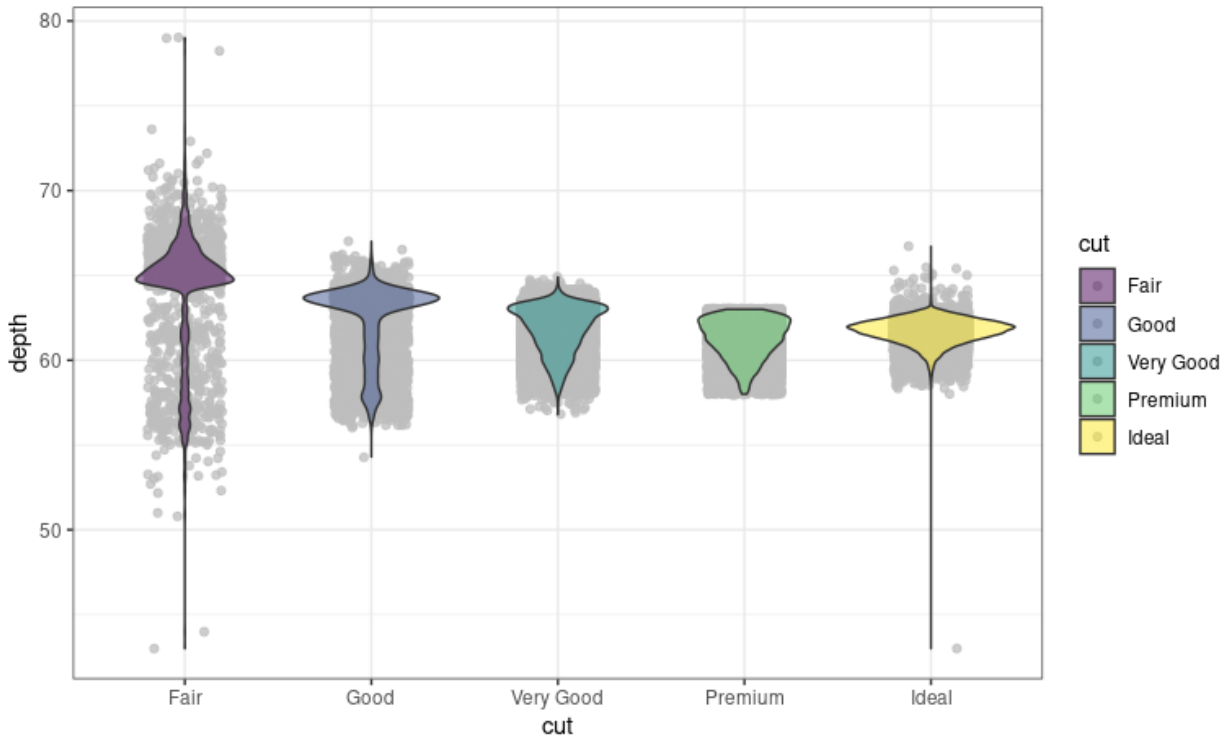
A violin plot reveals the structure in the data. It shows the density distribution. Box plot does not. Overall, a violin plot is more detailed.

11. [10%] Replace the box plots with violin plots in the figure from question 7, giving them the same level of transparency as the box plots.

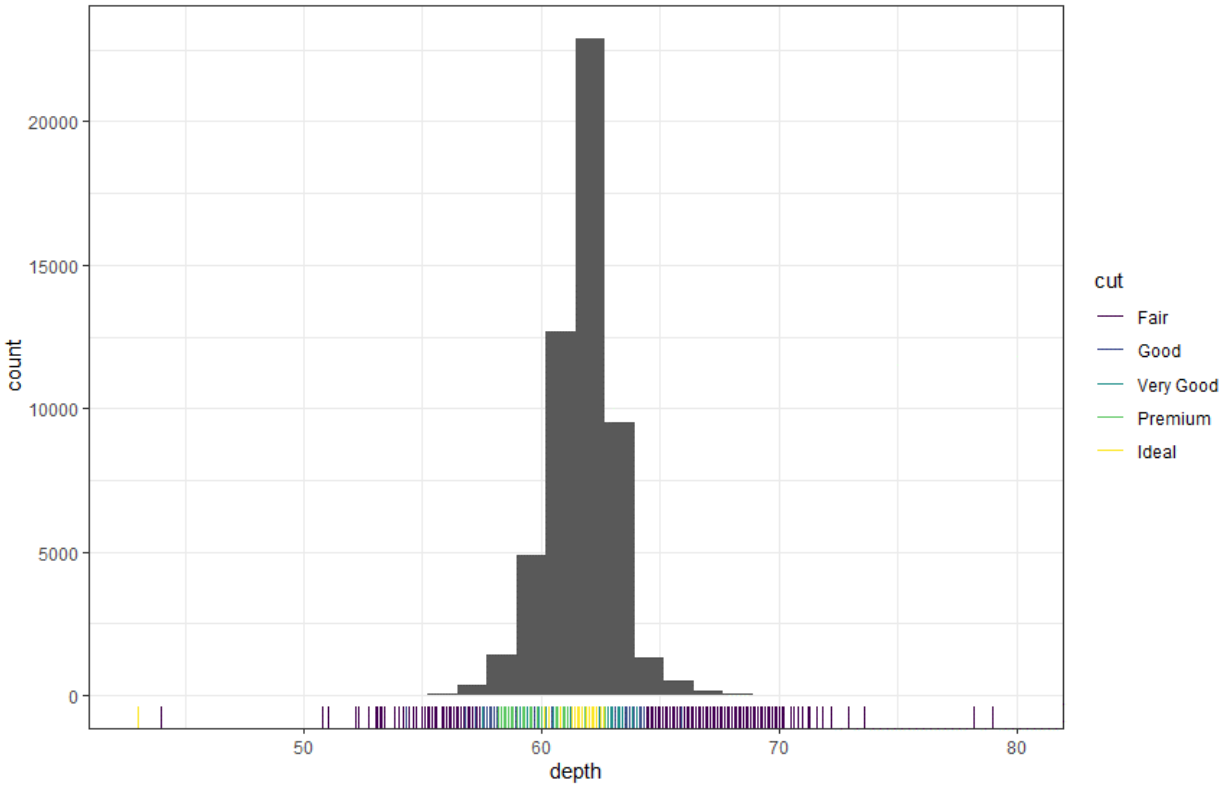
Provide code below:

```
ggplot(data = diamonds,
       mapping = aes(x = cut, y = depth, fill = cut)) +
  geom_jitter(width = 0.2, color = "gray", alpha = 0.75) +
  geom_violin(alpha = 0.5, outlier.shape = NA) +
  theme_bw()
```

Provide figure below:



12. [15%] Provide the code to generate the following plot.



Provide code below:

```
ggplot(data = diamonds,
       mapping = aes(x = cut, y = depth, fill = cut)) +
  geom_jitter(width = 0.2, color = "gray", alpha = 0.75) +
  geom_violin(alpha = 0.5, outlier.shape = NA) +
  theme_bw()
```