Supplemental Exercise SOLUTIONS - Data Manipuation & EDA

Jillian Morrison

October 1, 2019

Below are exercises you can use to practice data manipulation and EDA. See solutions on moodle.

Data Manipulation

`Fertility` dataset

1. Load the dplyr package. load the `Fertility` dataset from the `{AER}` package. Use `glimpse()` to see what is in the dataset.

```
> #install.packages("AER")
> library(AER)
> data("Fertility")
```

2. Save rows 35 to 50 of the age and work variables to a new dataset calles `Fert`. Hint: Use `slice()` and `%>%`

```
> library(dplyr)
> Fertility %>%  select(age, work) %>%slice(35:50)
   age work
1   28   20
2   33   12
3   32    0
4   26   52
5   32   52
6   28    0
7   32   40
8   35    0
9   33    0
10  32   42
11  29    0
12  29   52
13  31    0
14  30   51
15  28    0
16  29    0
```

3. Count how many women proceeded to have a third child.

```
> Fertility %>%  filter(morekids == "yes") %>%  count()
# A tibble: 1 x 1
      n
  <int>
1 96912
```

4. There are four possible gender combinations for the first two children. Which is the most common?

```
> Fertility %>%  group_by(gender1, gender2) %>%  count()
# A tibble: 4 x 3
# Groups:   gender1, gender2 [4]
  gender1 gender2     n
  <fct>   <fct>   <int>
1 female  female  60946
2 female  male    62724
3 male    female  63185
4 male    male    67799
```

5. By racial composition what is the proportion of woman working four weeks or less in 1979?

```
> Fertility %>%  group_by(afam, hispanic, other) %>%  summarise(mean(work <= 4))
# A tibble: 6 x 4
# Groups:   afam, hispanic [4]
  afam  hispanic other `mean(work <= 4)`
  <fct> <fct>    <fct>             <dbl>
1 no    no       no                0.509
2 no    no       yes               0.470
3 no    yes      no                0.524
4 no    yes      yes               0.506
5 yes   no       no                0.303
6 yes   yes      no                0.454
```

6. Filter out a subset of woman between the age 22 and 24 and calculate the proportion who had a boy as their firstborn

```
> Fertility %>%
+   filter(between(age, 22, 24)) %>%
+   summarise(mean(gender1 == "male"))
  mean(gender1 == "male")
1               0.5036608
```

7. Add a new column, age squared, to the dataset.

```
> Fertility <- Fertility %>%mutate(age_sq = age^2)
```

8. Calculate the proportion of women who have a third child by gender combination of the first two children?

```
> Fertility %>%
+   group_by(gender1, gender2) %>%
+   summarise(mean(morekids == "yes"))
# A tibble: 4 x 3
# Groups:   gender1 [2]
  gender1 gender2 `mean(morekids == "yes")`
  <fct>   <fct>                       <dbl>
1 female  female                      0.425
2 female  male                        0.347
3 male    female                      0.346
4 male    male                        0.404
```

9. Out of all the racial composition in the dataset which had the lowest proportion of boys for their firstborn.

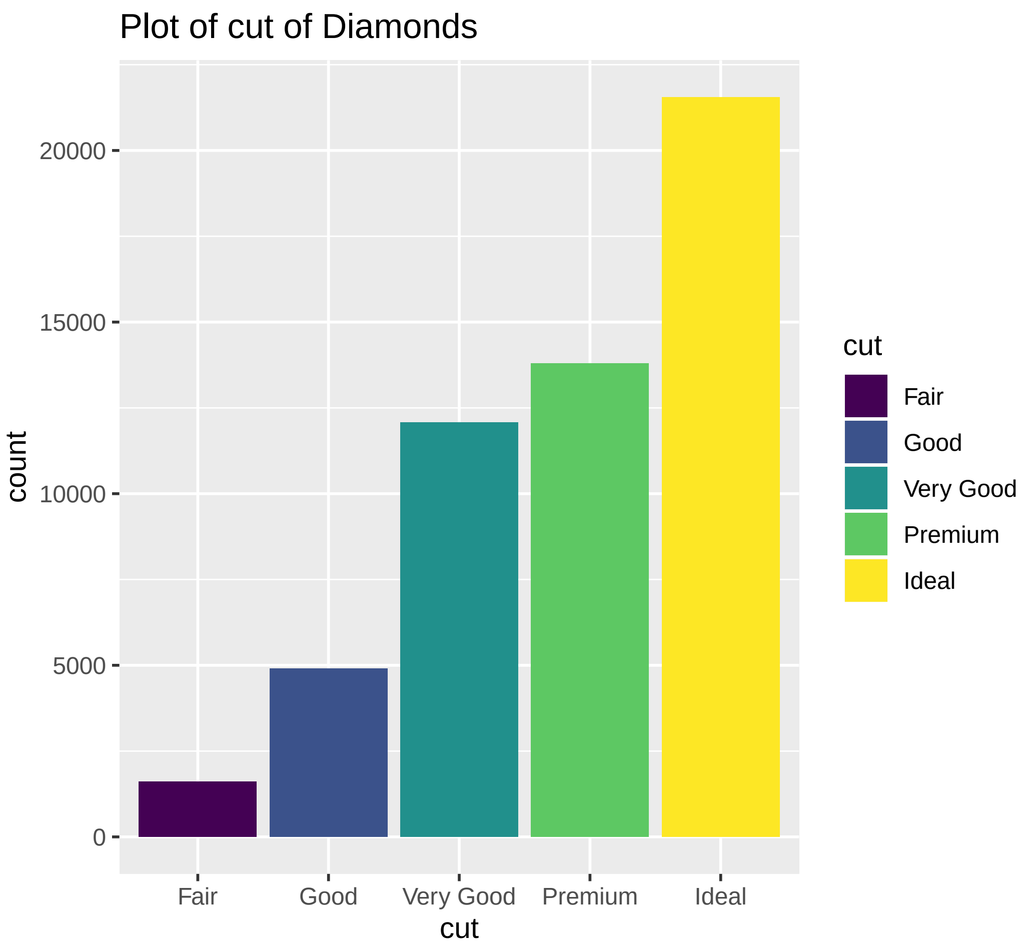
```
> Fertility %>%
+   group_by(afam, hispanic, other) %>%
+   summarise(prop_boys_fb = mean(gender1 == "male"), n = n()) %>%
+   arrange(prop_boys_fb)
# A tibble: 6 x 5
# Groups:   afam, hispanic [4]
  afam  hispanic other prop_boys_fb      n
  <fct> <fct>    <fct>        <dbl>  <int>
1 yes   no       no           0.509  12960
2 no    yes      no           0.512  11117
3 no    yes      yes          0.513   7584
4 no    no       no           0.515 216033
5 no    no       yes          0.520   6764
6 yes   yes      no           0.561    196
```

Exploratory Data Analysis

`diamonds` dataset

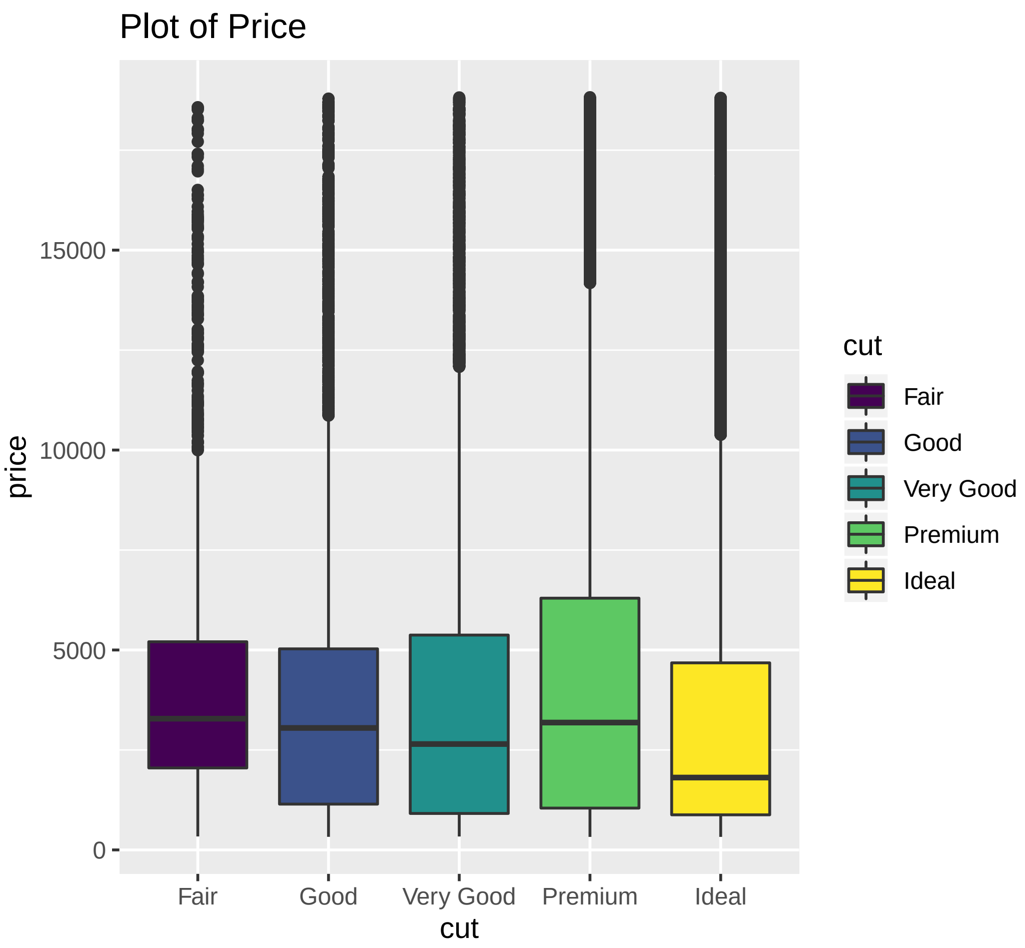
1.Using `diamonds` dataset in `{datasets}` package, COnstruct a barplot of cut. Add colors, a legend, and titles to the plot.

```
> library(ggplot2)
> ggplot(data = diamonds, mapping = aes(x = cut, fill = cut)) + geom_bar() +
+   ggtitle("Plot of cut of Diamonds")
```



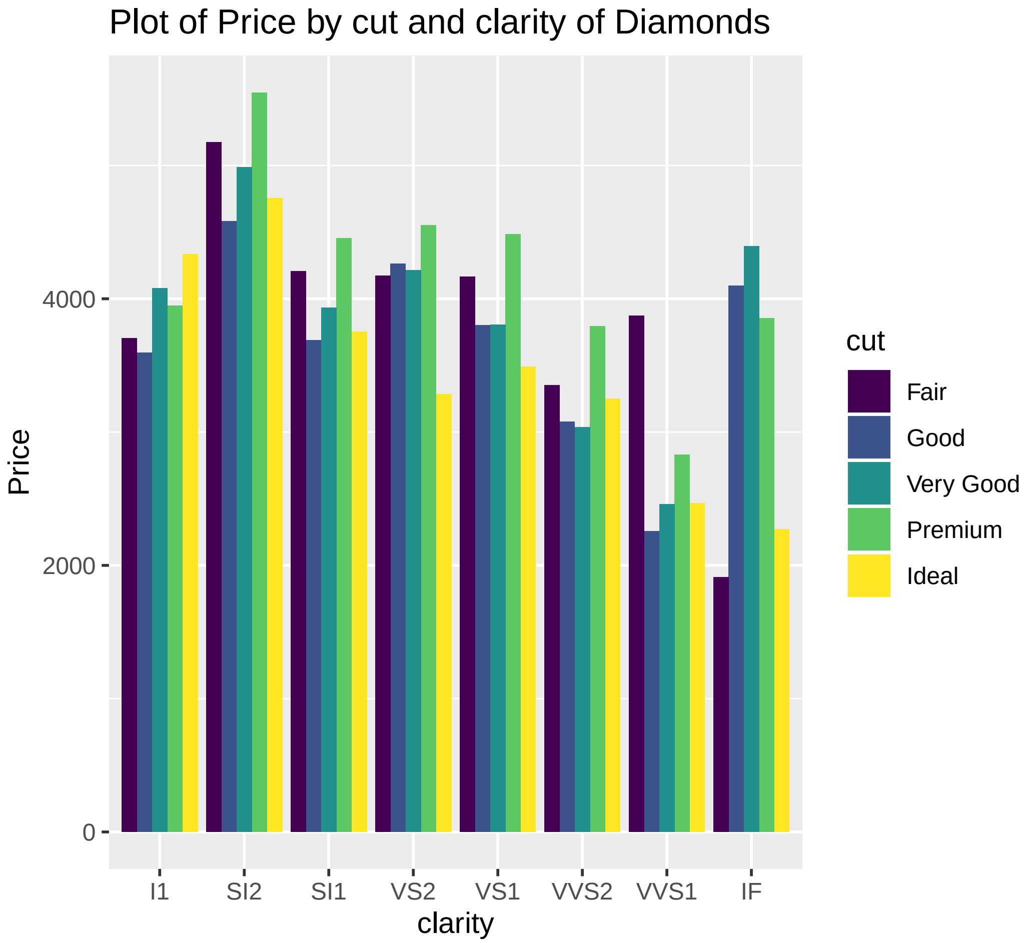
2. Create boxplots of Price by cut of diamonds. Add titles and labels

```
> ggplot(data=diamonds, aes(x=cut, y=price, fill=cut))+ geom_boxplot()+ggtitle("Plot of Price")
```



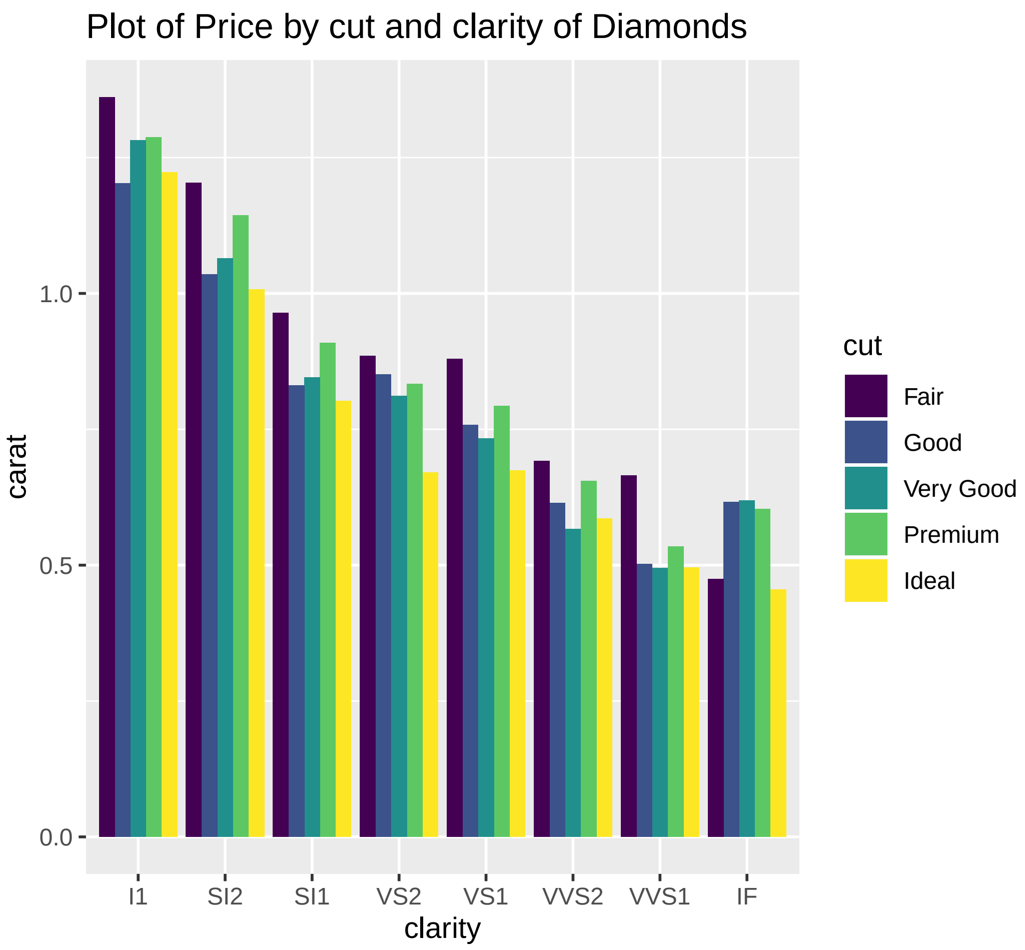
3. Construct barplot of mean Price by cut and clarity

```
> dia=diamonds%>%group_by(clarity,cut)%>% summarise(Price=mean(price), carat=mean(carat))
> ggplot(data = dia, mapping = aes(x = clarity, y=Price, fill = cut)) +
+   geom_bar(stat="identity", position="dodge") +
+   ggtitle("Plot of Price by cut and clarity of Diamonds")
```



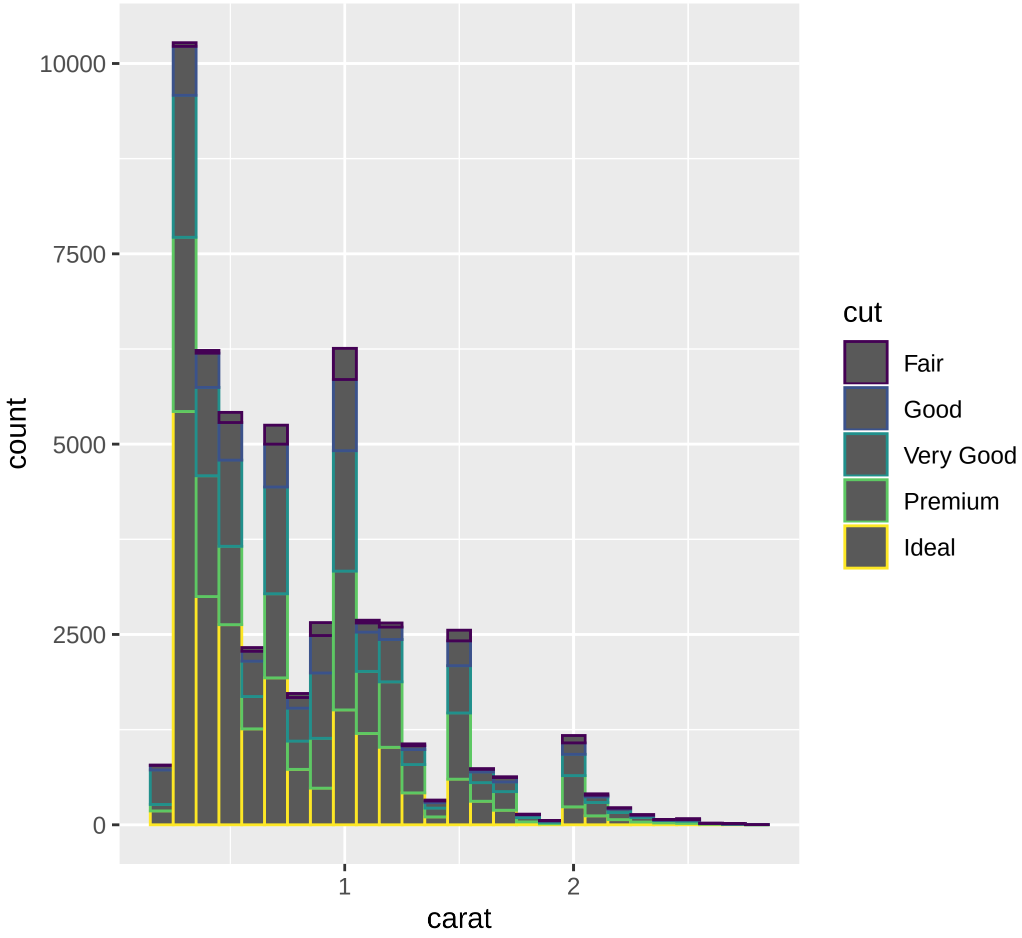
4. Construct barplot of mean carat by cut and clarity. Rearange the order of the grouping variables and choose the order that makes the most sense

```
> ggplot(data = dia, mapping = aes(x = clarity, y=carat, fill = cut)) +
+   geom_bar(stat="identity", position="dodge") +
+   ggtitle("Plot of Price by cut and clarity of Diamonds")
```



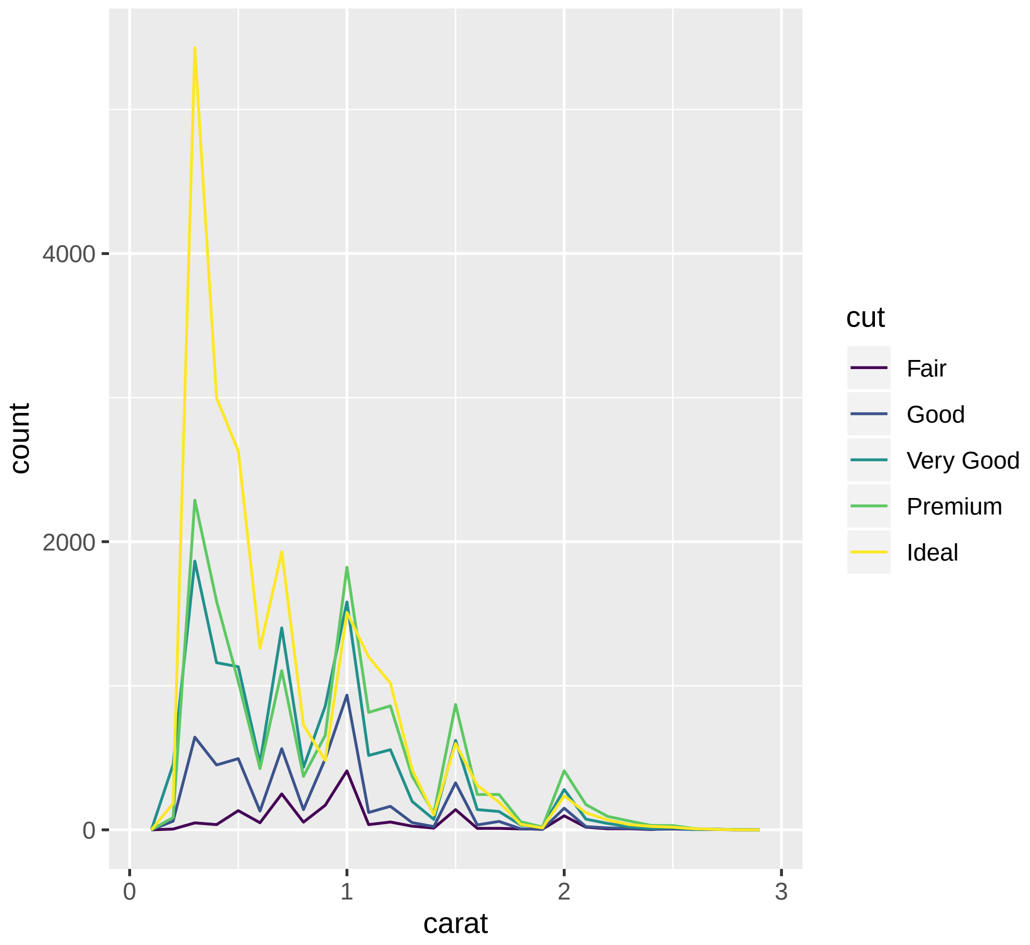
5. Select the observations/diamonds that have carat less than 3. construct histograms of carat and group by cut.

```
> smaller <- diamonds %>%   filter(carat < 3)
>
> ggplot(data = smaller, mapping = aes(x = carat, colour = cut)) +
+   geom_histogram(binwidth = 0.1)
```



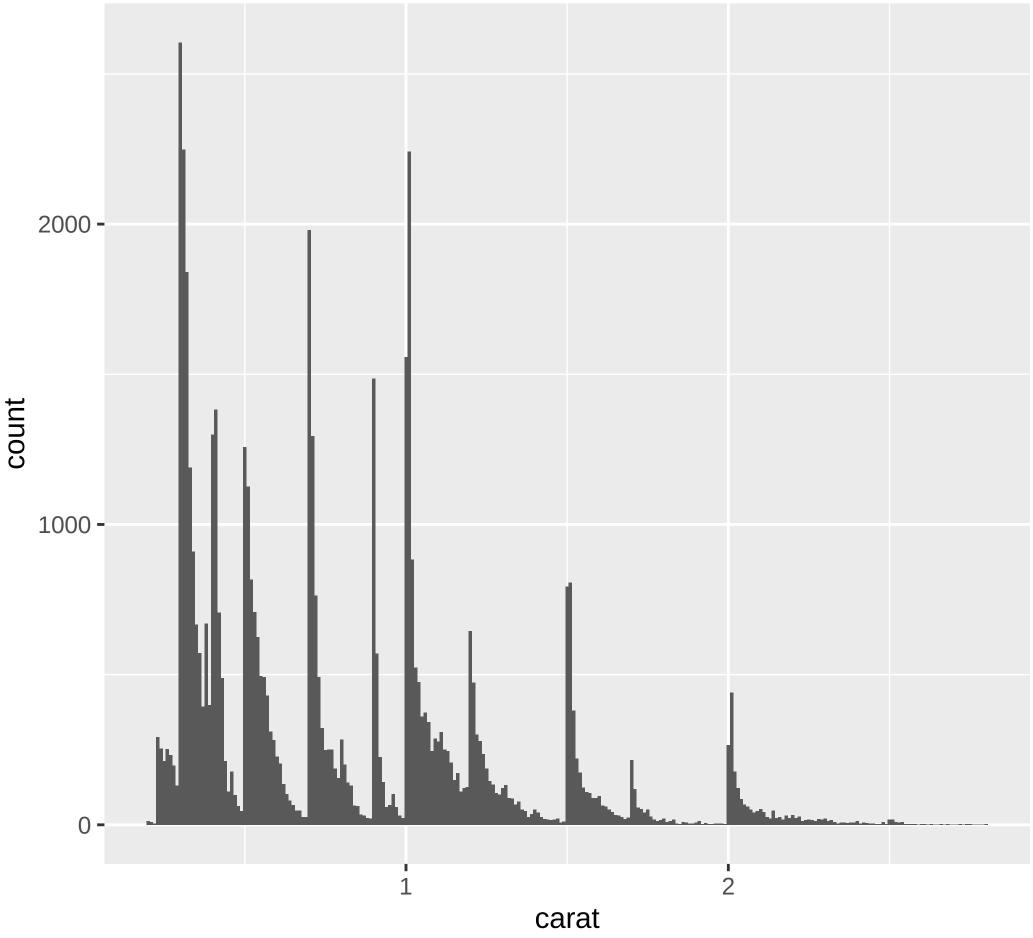
6. Notice that you cannot say much about the histogram in 5 above. try using `geom_freqpoly()` instead of `geom_histogram()`. Compare the result to the result in 5.

```
> ggplot(data = smaller, mapping = aes(x = carat, colour = cut)) +
+   geom_freqpoly(binwidth = 0.1)
```



7. Try the following code for a histogram:

```
> ggplot(data = smaller, mapping = aes(x = carat)) +
+   geom_histogram(binwidth = 0.01)
```



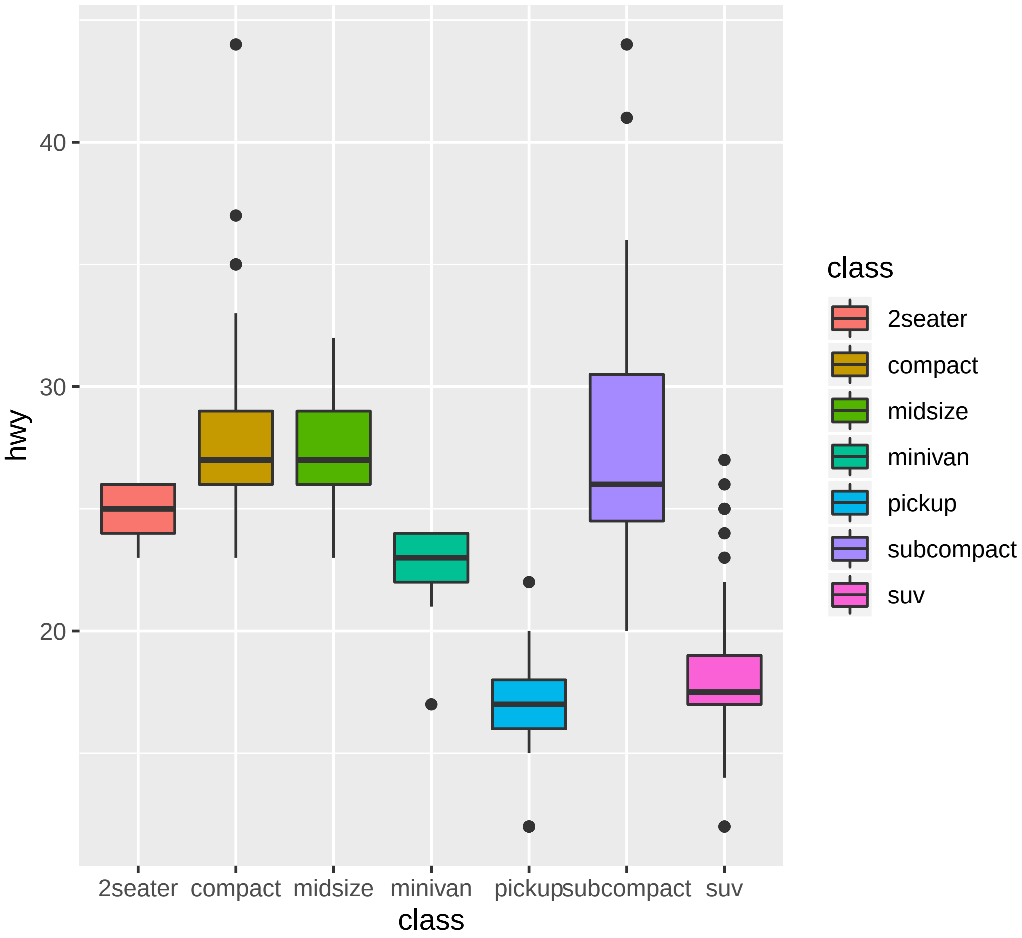
What kinds of questions does this generate about diamonds?

```
> #Why are there more diamonds at whole carats and common fractions of carats?
>
> #Why are there more diamonds slightly to the right of each peak than there are slightly to the left o
>
> #Why are there no diamonds bigger than 3 carats?
```

mpg dataset

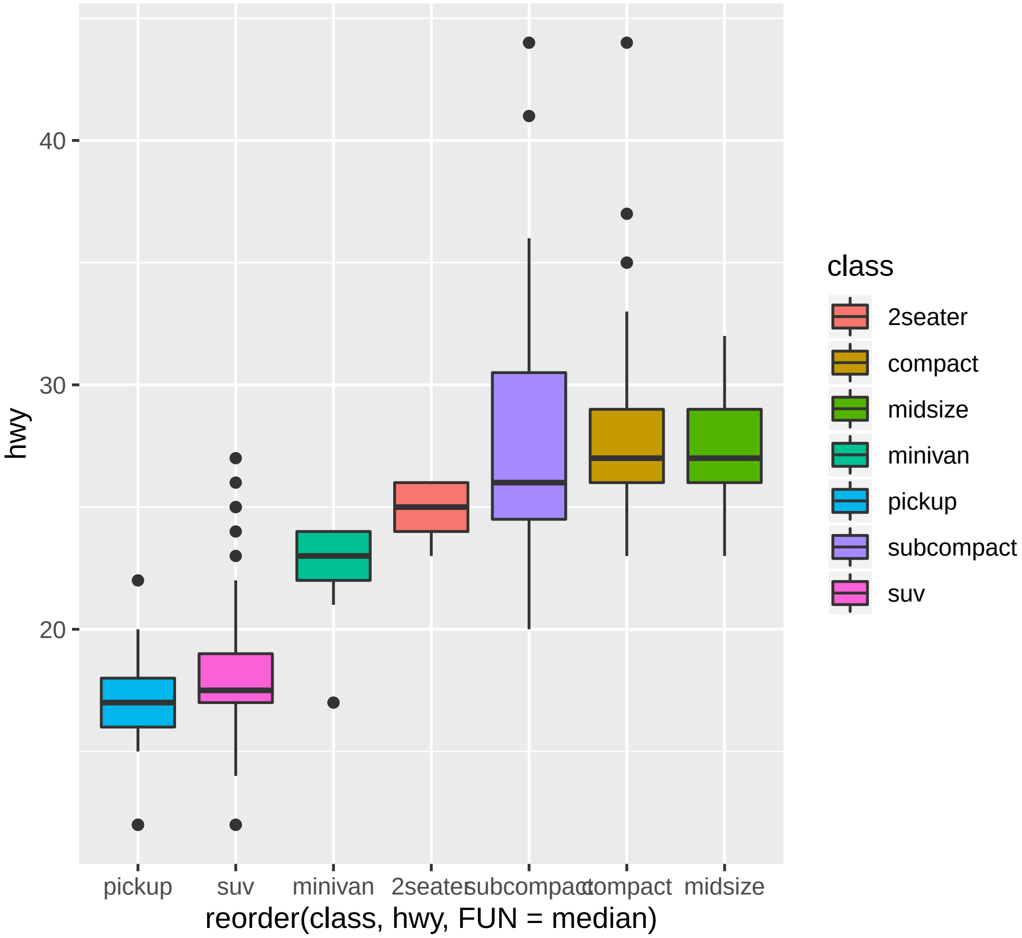
8. Create boxplots of highway mileage by class

```
> ggplot(data = mpg, mapping = aes(x = class, y = hwy, fill=class)) +  geom_boxplot()
```



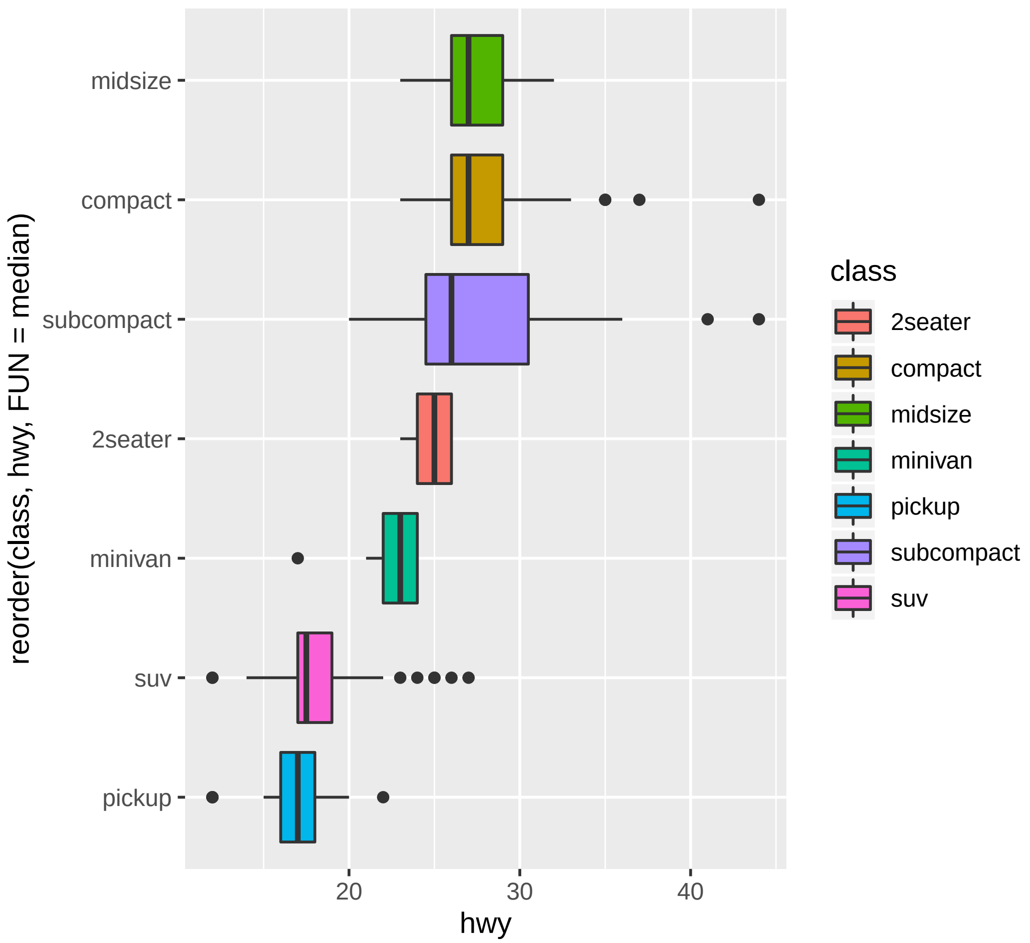
9. Re-order the plot in 8 by the median for each class. Hint: for the x variable, use x = `reorder(class, hwy, FUN = median)`

```
> ggplot(data = mpg) +
+   geom_boxplot(mapping = aes(x = reorder(class, hwy, FUN = median), y = hwy, fill=class))
```



10. Use layering (i.e. `+ coord_flip()`) to flip the plots 90 degrees to horizontal boxplots instead.

```
> ggplot(data = mpg) +
+   geom_boxplot(mapping = aes(x = reorder(class, hwy, FUN = median), y = hwy, fill=class)) +
+   coord_flip()
```


midsize
compact
subcompact
2seater
minivan
suv
pickup
reorder(class, hwy, FUN = median)
20
30
40
hwy
class
2seater
compact
midsize
minivan
pickup
subcompact
suv