Lecture Assignment 9

Viraj Vijaywargiya

2022-04-29

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

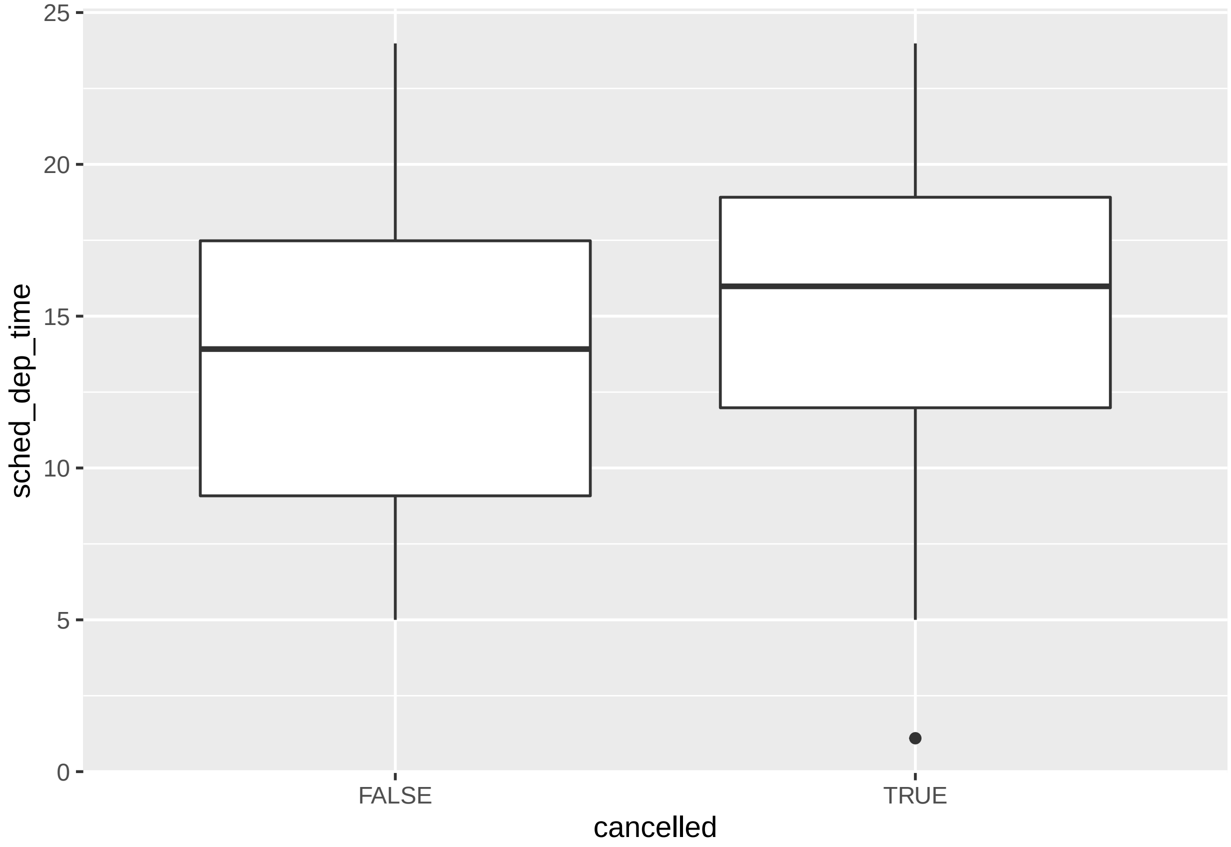
```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Part 7.5.1.1

Question 1

To improve the visualization, we can use boxplot instead of freqpoly.

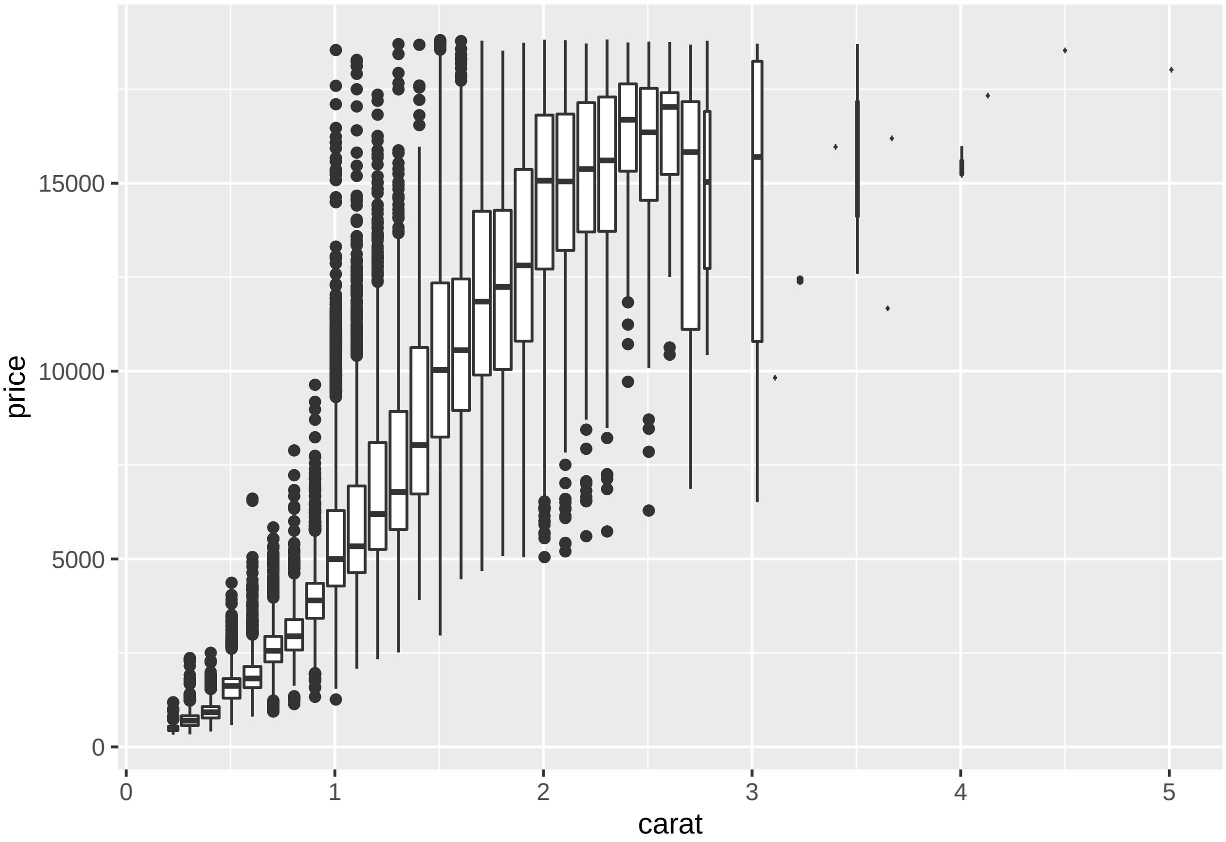
```
nycflights13::flights %>%
  mutate(cancelled = is.na(dep_time), sched_hour = sched_dep_time %/% 100,
    sched_min = sched_dep_time %% 100, sched_dep_time = sched_hour + sched_min / 60) %>%
  ggplot() +
  geom_boxplot(mapping = aes(y = sched_dep_time, x = cancelled))
```



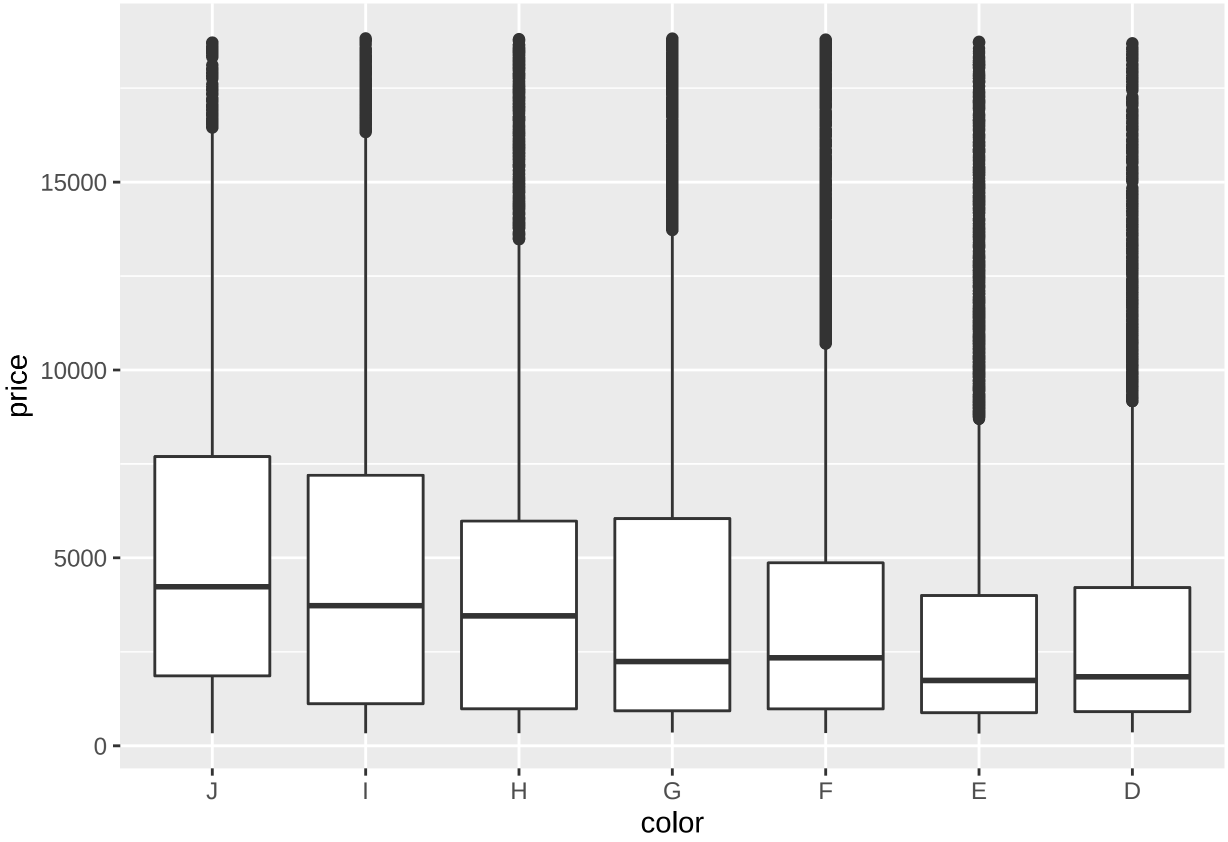
Question 2

Comparing the relationships of variables, carat, clarity, color, and cut, with price using boxplot.

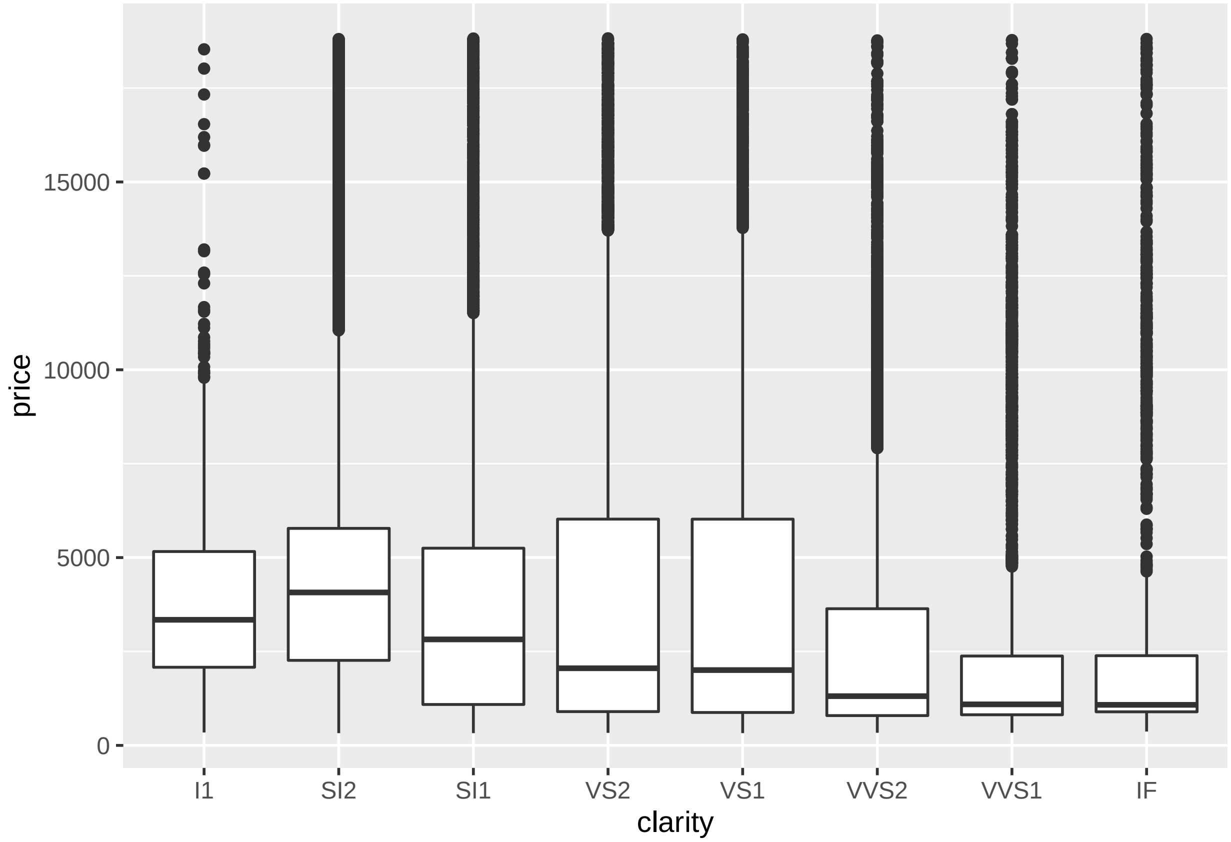
```
ggplot(data = diamonds, mapping = aes(x = carat, y = price)) + #both continuous
  geom_boxplot(mapping = aes(group = cut_width(carat, 0.1)), orientation = "x")
```



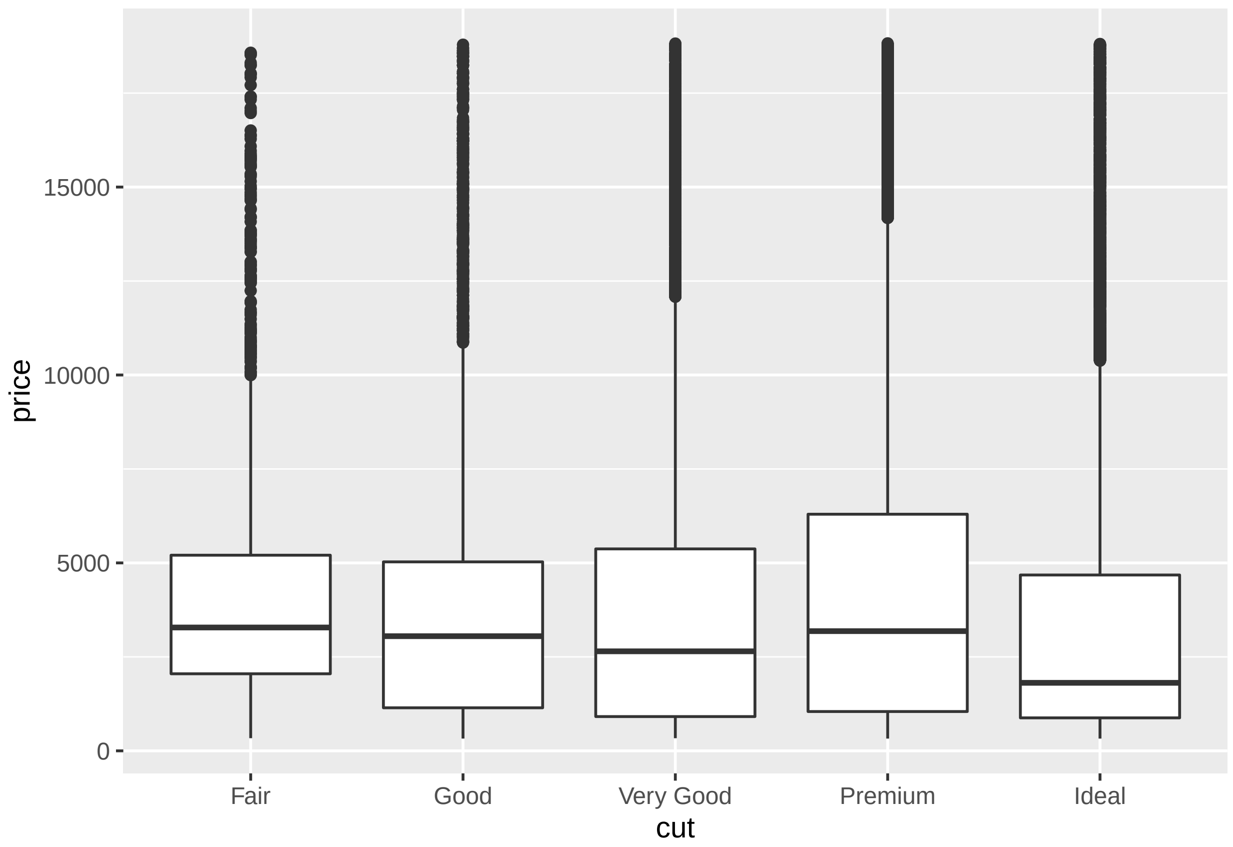
```
diamonds %>%
  mutate(color = fct_rev(color)) %>%
  ggplot(aes(x = color, y = price)) +
    geom_boxplot()
```



```
ggplot(data = diamonds) +
  geom_boxplot(mapping = aes(x = clarity, y = price))
```

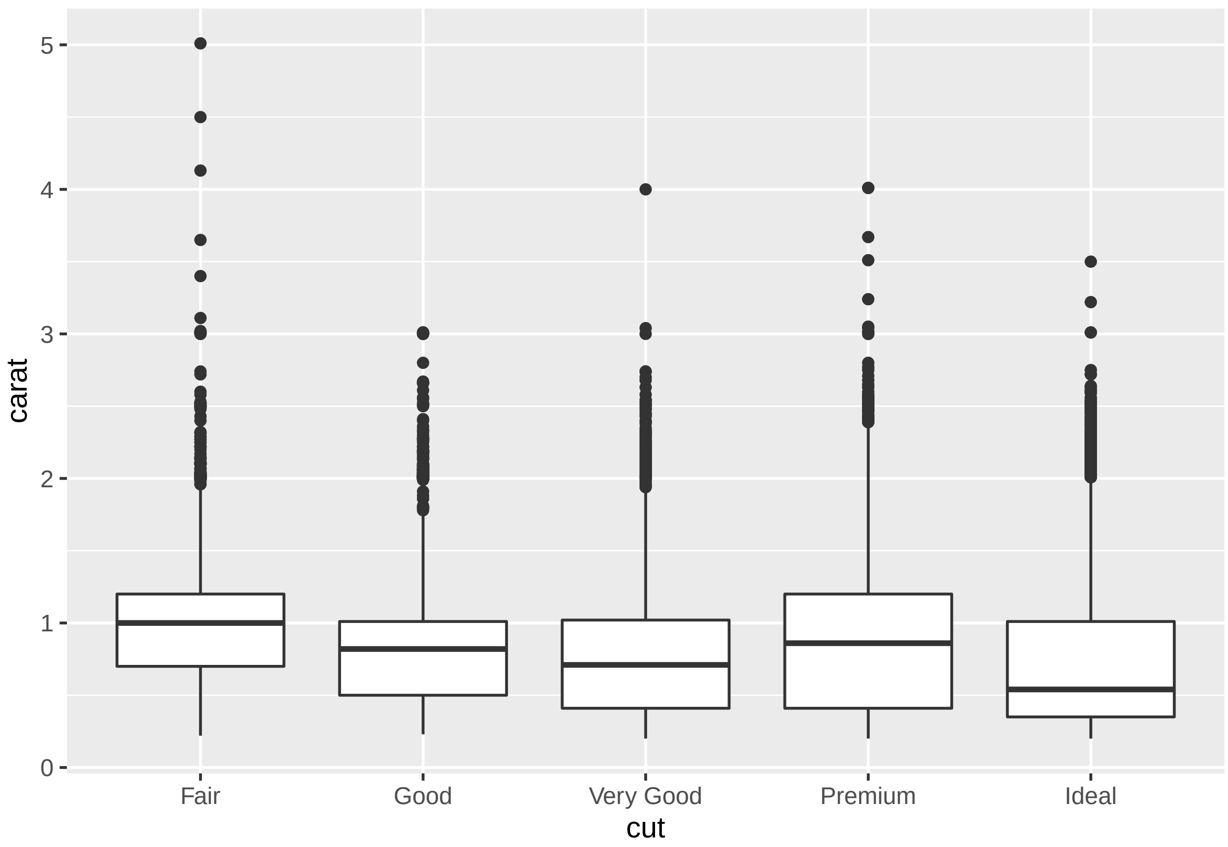


```
ggplot(data = diamonds) +
  geom_boxplot(mapping = aes(x = cut, y = price))
```



From the plots above, clearly the variable, carat, is most important for predicting the price of a diamond. Relationship between carat and cut using a boxplot.

```
ggplot(data = diamonds) +
  geom_boxplot(mapping = aes(x = cut, y = carat))
```



A slight negative relation between carat and cut can be seen from the plot above, with largest carat diamond having a cut of Fair. This could be because a large diamond can be sold with lower quality, whereas, a small diamond needs to have a better cut.

Part 7.5.2.1

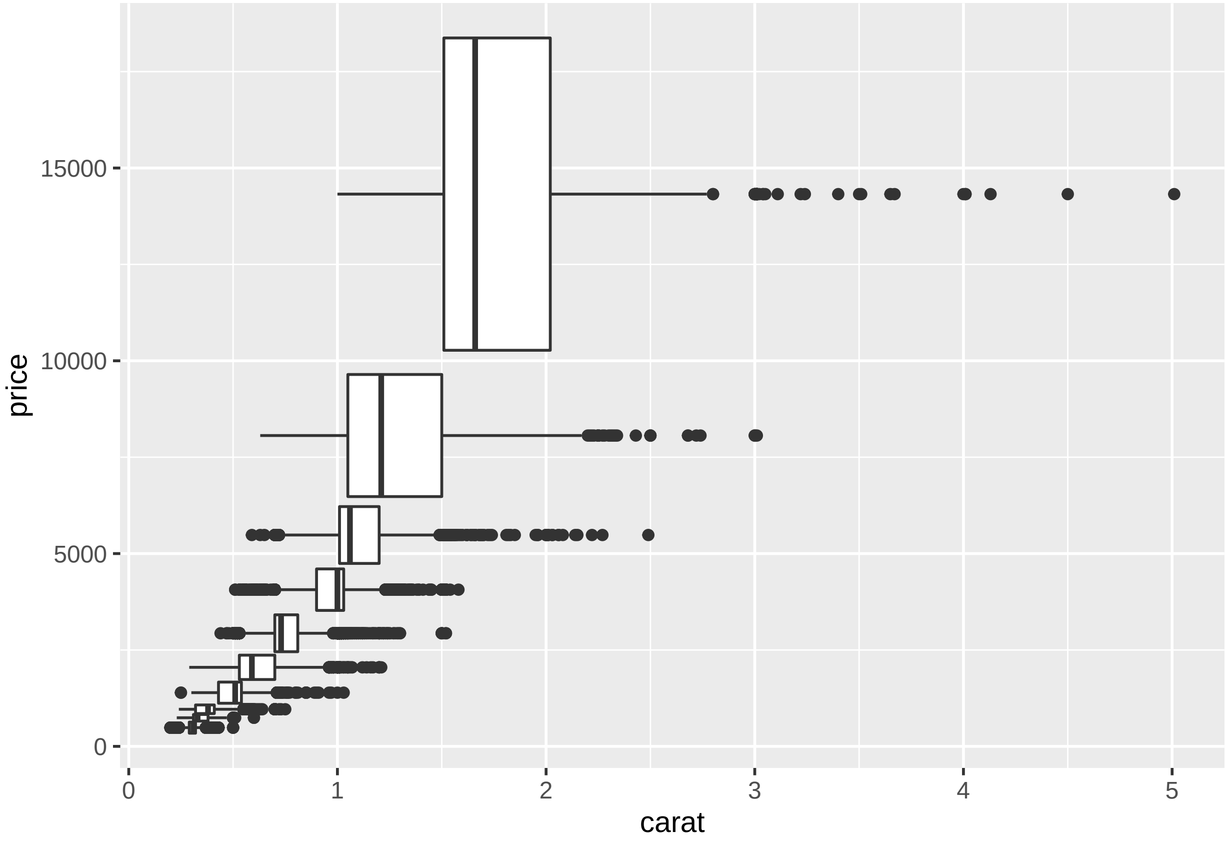
Question 3

Categorical variables with larger number of categories or longer labels should kept on the y axis, which makes it easier to read. However, this is just slightly better in this case because the labels don't overlap when the order is switched. Also, larger numbers are at the top when using x = color and y = cut, and this also makes the plot easier to read.

Part 7.5.3.1

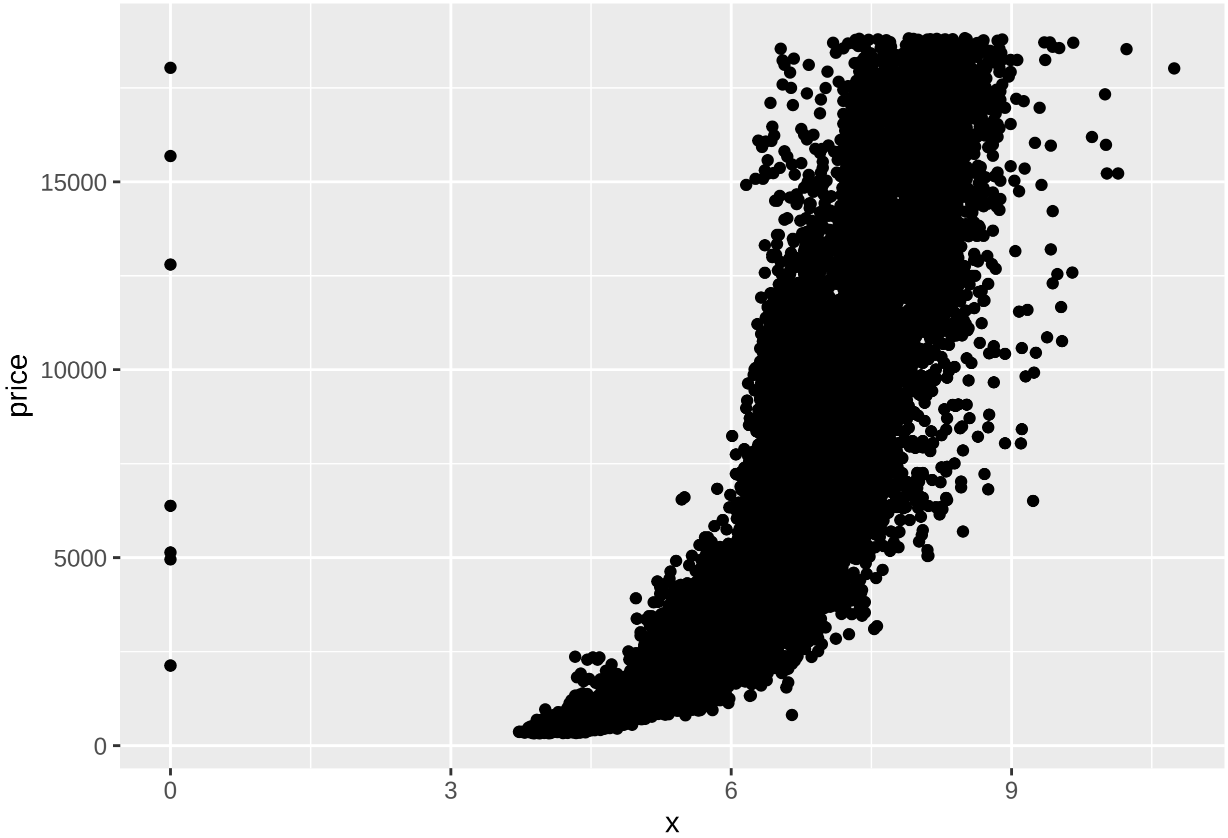
Question 2

```
ggplot(data = diamonds, mapping = aes(x = price, y = carat)) +
  geom_boxplot(mapping = aes(group = cut_number(price, 10))) +
  coord_flip()
```



Question 3

```
ggplot(data = diamonds, mapping = aes(x = x, y = price)) +
  geom_point()
```



There is more variation in prices for larger diamonds than for smaller ones. I didn't really know what to expect as I have very little knowledge about diamonds, however, it does make sense as larger diamonds would have more variety (in terms of cut, color, clarity) and therefore, vary more in price than those of smaller ones.

Question 5

There is a strong relation between x and y in this case, and the outliers aren't extreme in either x or y. Therefore, a scatterplot is a better display because a binned plot would not reveal these outliers, which may leave us to believe that the largest value of x was an outlier even if it fits the bivariate pattern well.