Textbook Homework Problems

Amanda Liu PID:730042603

7.3.4 #1,2,3

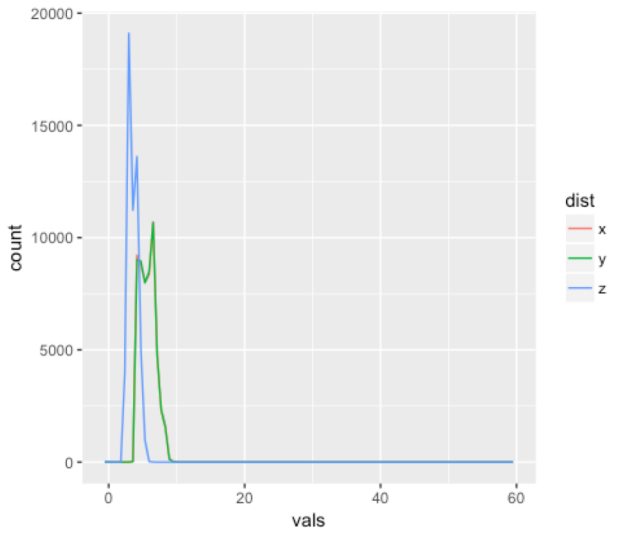
1. Explore the distribution of each of the x, y, and z variables in diamonds. What do you learn? Think about a diamond and how you might decide which dimension is the length, width, and depth.

From the distribution plot, I find that the x-distribution and y-distribution are nearly the same. And all of them are right-skewed.

From the data, the length is directly proportional to the x value.

Thus, I think the length is x, width is y, depth is z.

```
> diamonds %>% gather(key = dist, vals, x, y, z) %>% ggplot(aes(
vals, colour = dist)) + geom_freqpoly(bins = 100)
```




```
> diamonds %>% filter(y < 30) %>% select(x, y, z)
# A tibble: 53,938 x 3
       x     y     z
   <dbl> <dbl> <dbl>
 1  3.95  3.98  2.43
 2  3.89  3.84  2.31
 3  4.05  4.07  2.31
 4  4.20  4.23  2.63
 5  4.34  4.35  2.75
 6  3.94  3.96  2.48
 7  3.95  3.98  2.47
 8  4.07  4.11  2.53
 9  3.87  3.78  2.49
10  4.00  4.05  2.39
# ... with 53,928 more rows
```

x

length in mm (0–10.74)

y

width in mm (0–58.9)

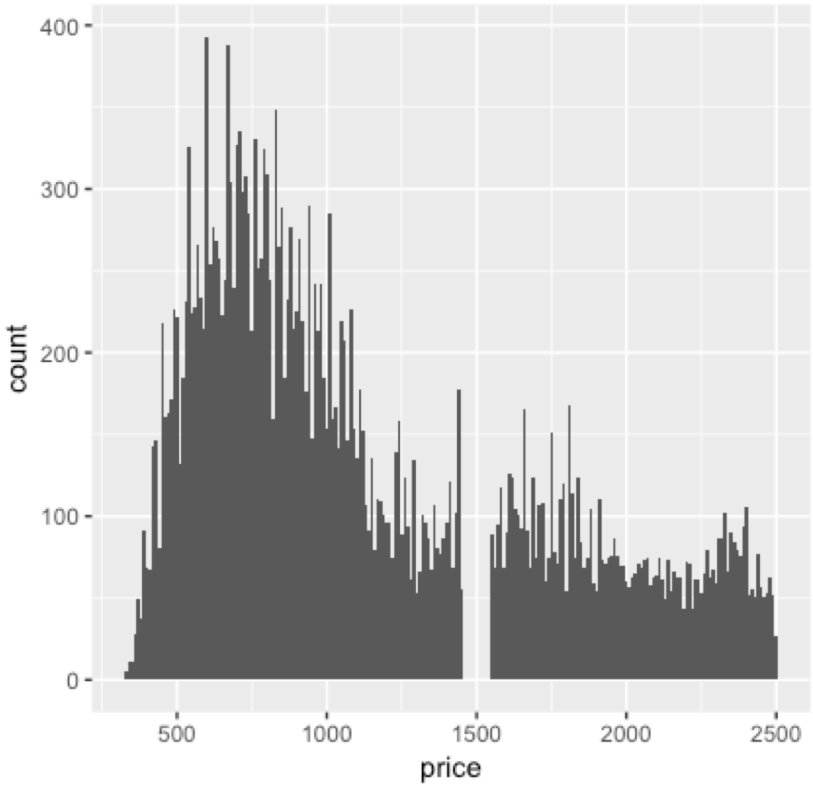
z

depth in mm (0–31.8)

2. Explore the distribution of price. Do you discover anything unusual or surprising? (Hint: Carefully think about the binwidth and make sure you try a wide range of values.)

From the graph, I find that there is no diamond with a price of $1500, which is suprising.

```
> ggplot(filter(diamonds, price < 2500), aes(x = price)) + geom_histogr
m(binwidth = 10, center = 0)
```



3. How many diamonds are 0.99 carat? How many are 1 carat? What do you think is the cause of the difference?

```
> diamonds %>% filter(carat %in% c(0.99, 1)) %>% count(carat)
# A tibble: 2 x 2
  carat     n
  <dbl> <int>
1  0.99    23
2  1.00  1558
```

23 diamonds are 0.99 carat. 1558 diamonds are 1 carat.

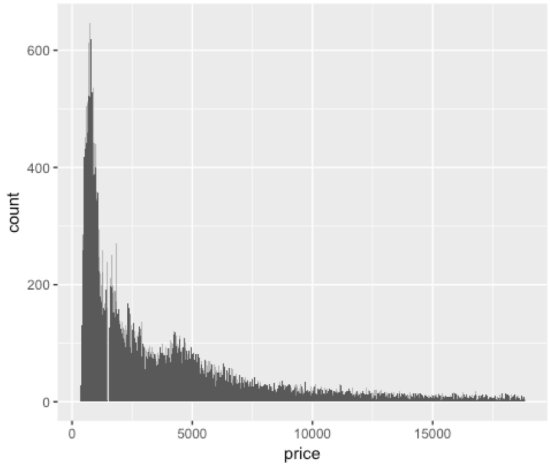
There might exist some diamonds between 0.99 carat and 1 carat and then being rounded up to 1 carat, which makes more 1 carat diamonds. And costumers always prefer to buy 1-carat diamonds.

7.4.1 #1,2

1. What happens to missing values in a histogram? What happens to missing values in a bar chart? Why is there a difference?

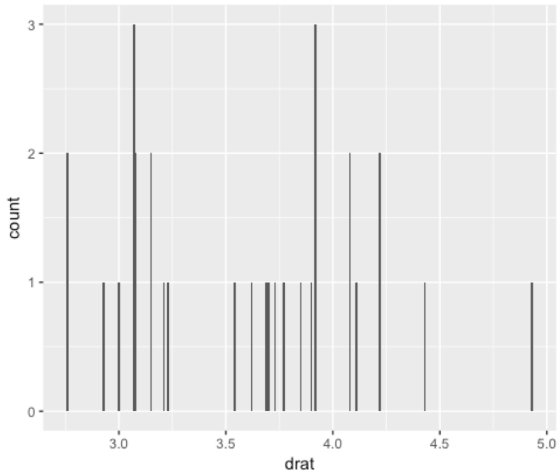
In a histogram, a missing value will cause a gap in the graph.

```
> diamonds %>% ggplot(aes(price)) + geom_histogram(bins = 1000)
```



In the bar chart, a missing value will directly be removed.

```
> mtcars[1,5] <- NA
> mtcars %>% ggplot(aes(drat)) + geom_bar()
Warning message:
Removed 1 rows containing non-finite values
(stat_count).
```



In histogram, the x variable can only be numeric. R would directly drop the value if it is not numeric. While in a bar chart, R can treat NA in a separate category and displays them in the graph.

2. What does na.rm = TRUE do in mean() and sum()?

"na.rm = True" means to remove the missing values (NA) when calculating mean() and sum().

7.5.1.1 #2

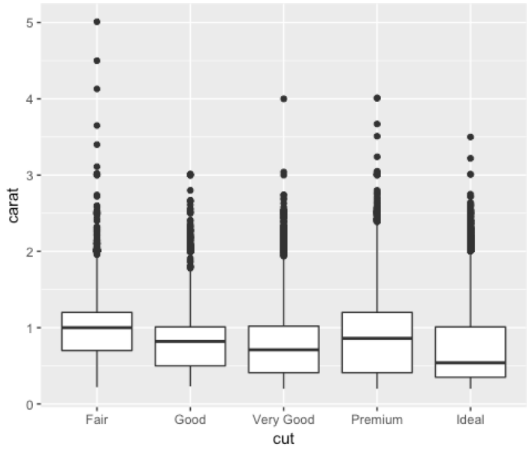
2. What variable in the diamonds dataset is most important for predicting the price of a diamond? How is that variable correlated with cut? Why does the combination of those two relationships lead to lower quality

diamonds being more expensive?

I believe the most important variable in the diamonds dataset for predicting the price of a diamond is **carat**.

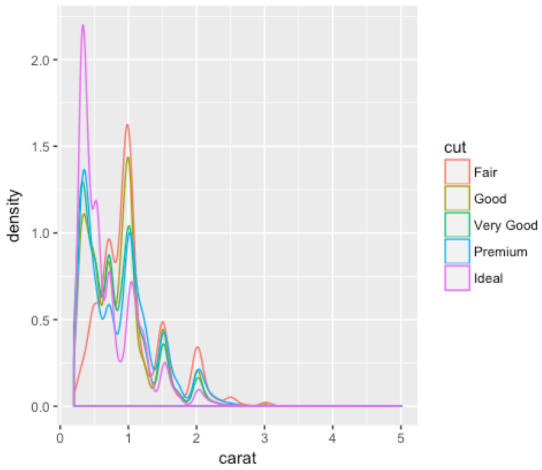
Cut: It seems like carat is negatively correlated with cut, which means lower quality diamonds have greater carat.

```
> diamonds %>% ggplot(aes(cut, carat)) + geom_boxplot()
>
```



Carat: It seems that the fair diamonds have the highest quantity of carat.

```
> diamonds %>% ggplot(aes(carat, colour = cut)) + geom_density
(position = "dodge")
```



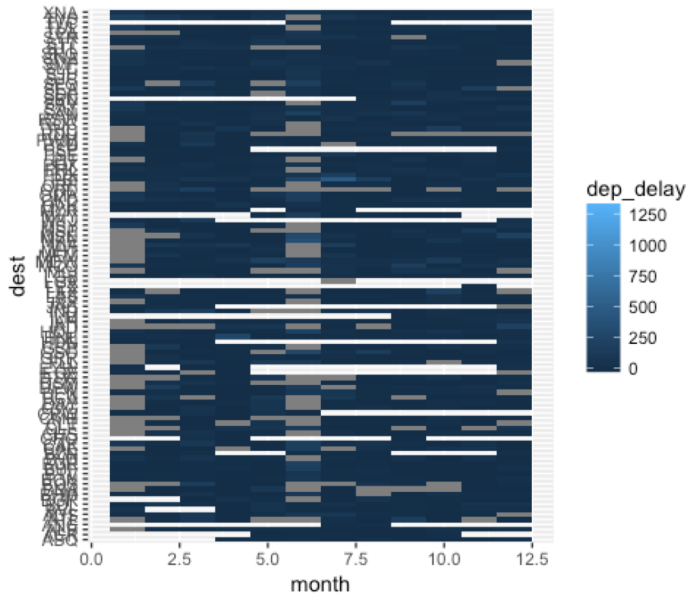

```
> diamonds %>% group_by(cut) %>% summarise(cor(carat, price))
# A tibble: 5 x 2
        cut `cor(carat, price)`
      <ord>               <dbl>
1      Fair           0.8592985
2      Good           0.9224716
3 Very Good           0.9263704
4   Premium           0.9250047
5     Ideal           0.9311760
```

From these two relationships, I find that the lower quality diamonds usually have much more carat, which is the most important variable for predicting the prices. Higher carats will have a higher price. Thus, lower quality diamonds may be more expensive.

7.5.2.1 #2,3

2. Use geom_tile() together with dplyr to explore how average flight delays vary by destination and month of year. What makes the plot difficult to read? How could you improve it?

```
> flights %>% ggplot(aes(x = month, y = dest, fill = dep_delay
)) + geom_tile()
```



The graph is almost filled by dark colors and we cannot see the data which has a dep_delay higher than 1250 very clearly.

Improvement:

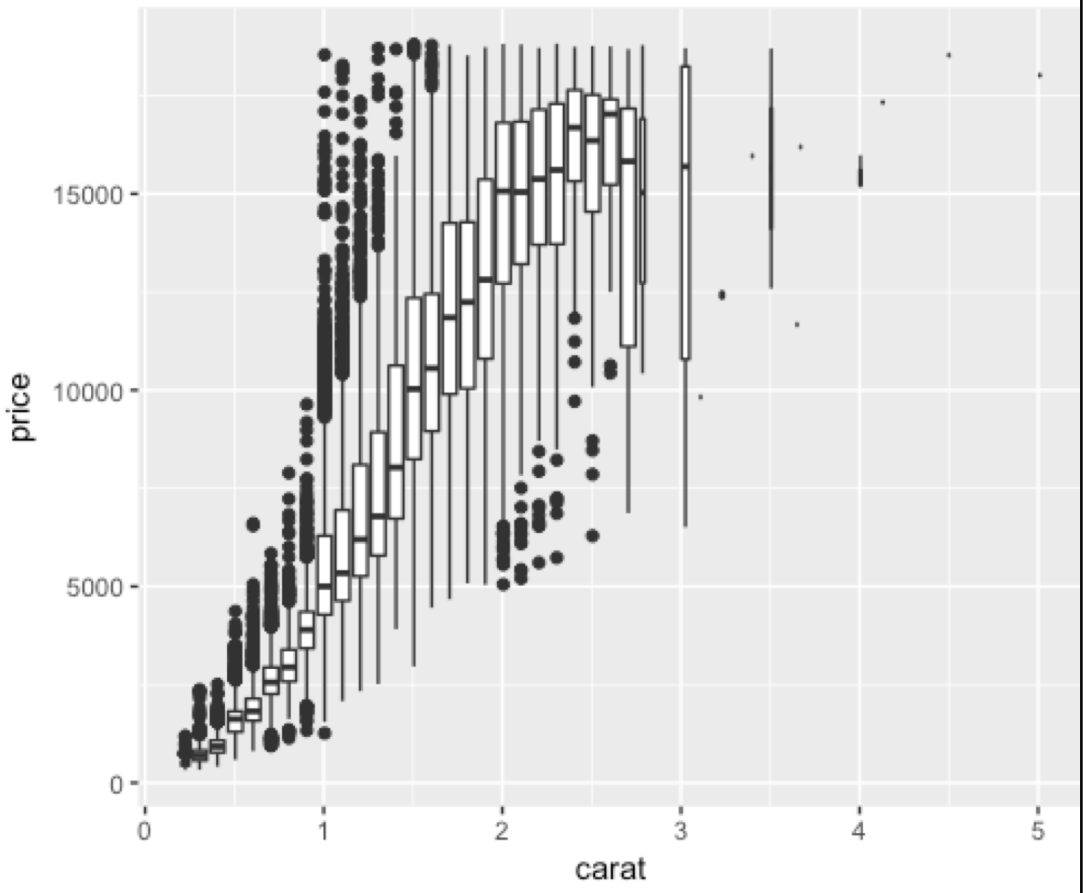
Make colors for dep_delay more different.

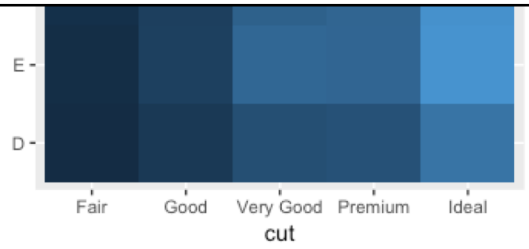
Remove those missing values

3. Why is it slightly better to use aes(x = color, y = cut) rather than aes(x = cut, y = color) in the example above?

The price range for very large diamonds is wider than that of those smaller diamonds. Although the larger the diamonds are, the more expensive they are, there still exist some large diamonds that have relatively lower price. I am not very surprised by the result. Because there are many other factors which would also affect the prices.

```
> smaller<- diamonds %>% filter(carat < 3)
> ggplot(data = diamonds, mapping = aes(x = carat, y = price)
) + geom_boxplot(mapping = aes(group = cut_width(carat, 0.1))
)
```





7.5.3.1 #3

3. How does the price distribution of very large diamonds compare to small diamonds. Is it as you expect, or does it surprise you?