Lecture Assignment 8

Viraj Vijaywargiya

2022-04-27

```
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()
```

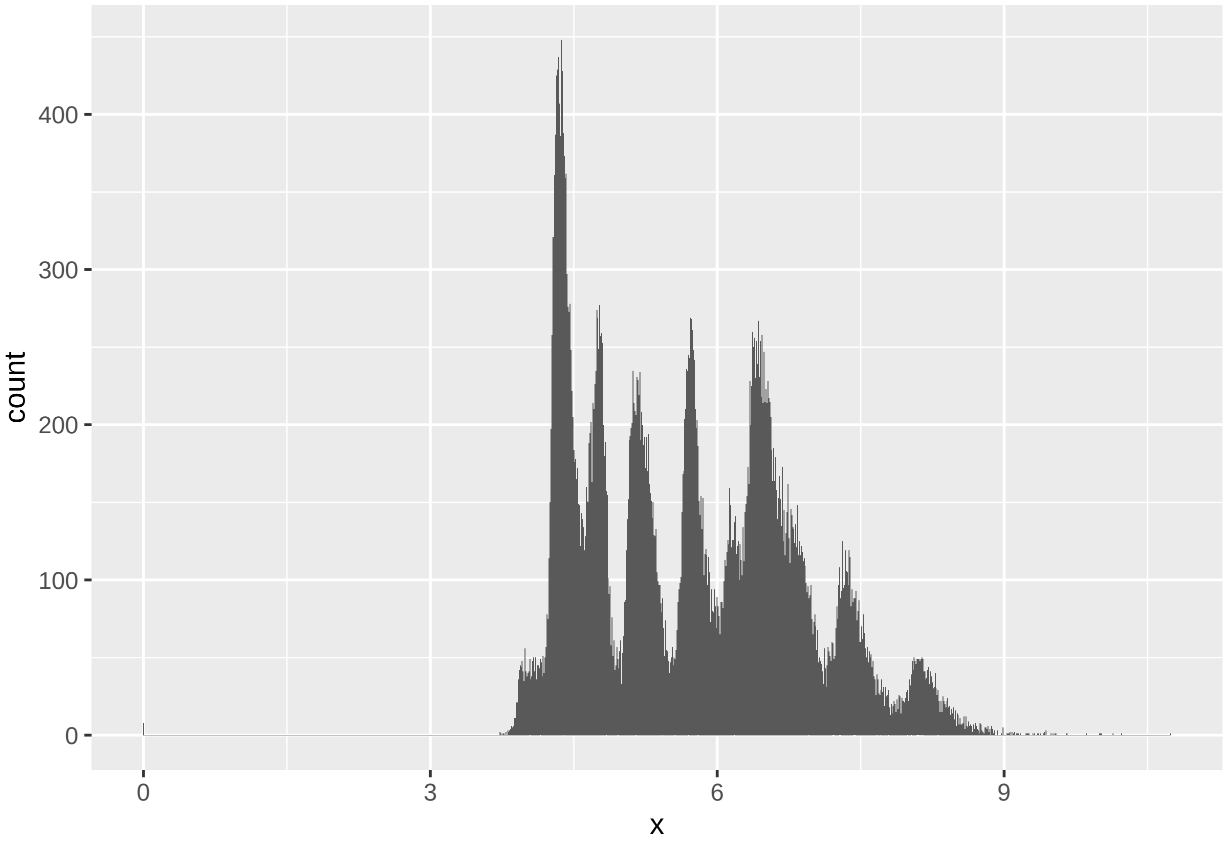
Part 7.3.4

Question 1

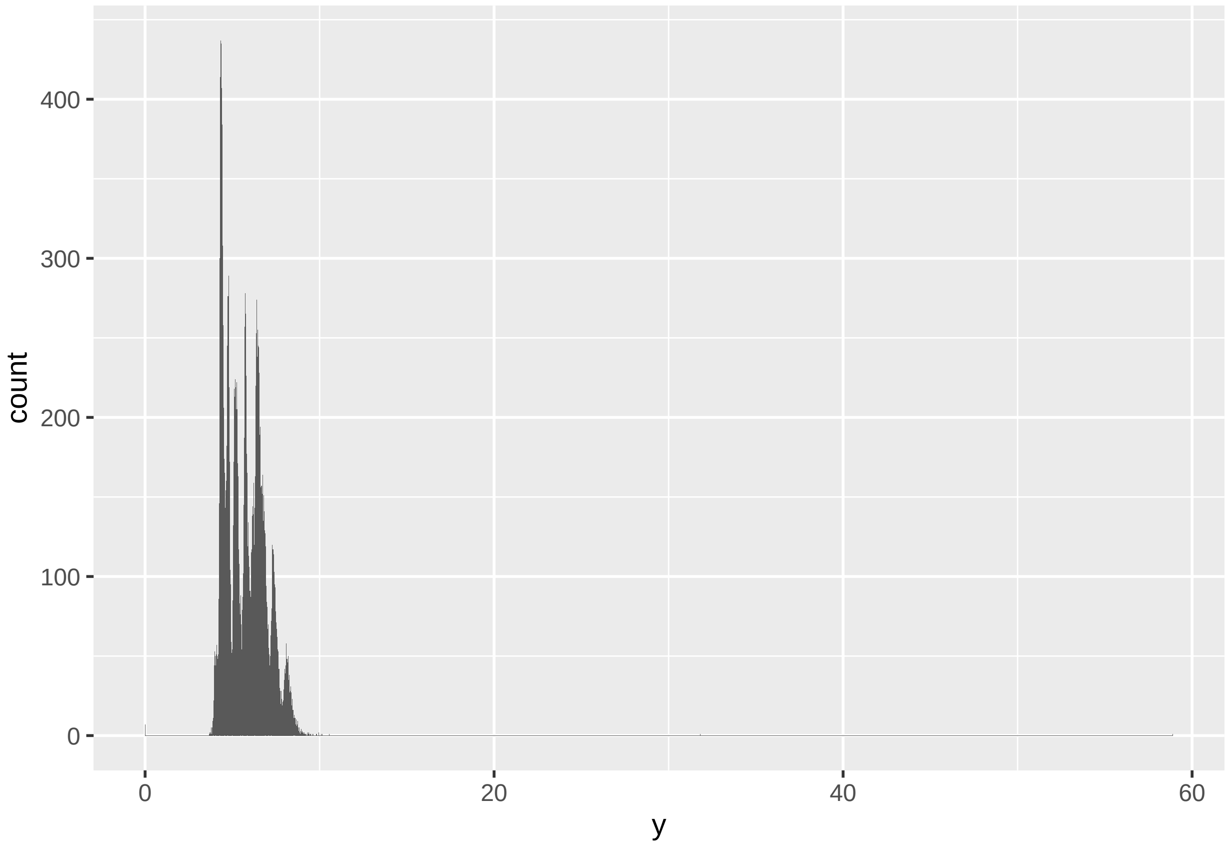
```
summary(select(diamonds, x, y, z))
```

```
##        x                y                z
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910
##  Median : 5.700   Median : 5.710   Median : 3.530
##  Mean   : 5.731   Mean   : 5.735   Mean   : 3.539
##  3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040
##  Max.   :10.740   Max.   :58.900   Max.   :31.800
```

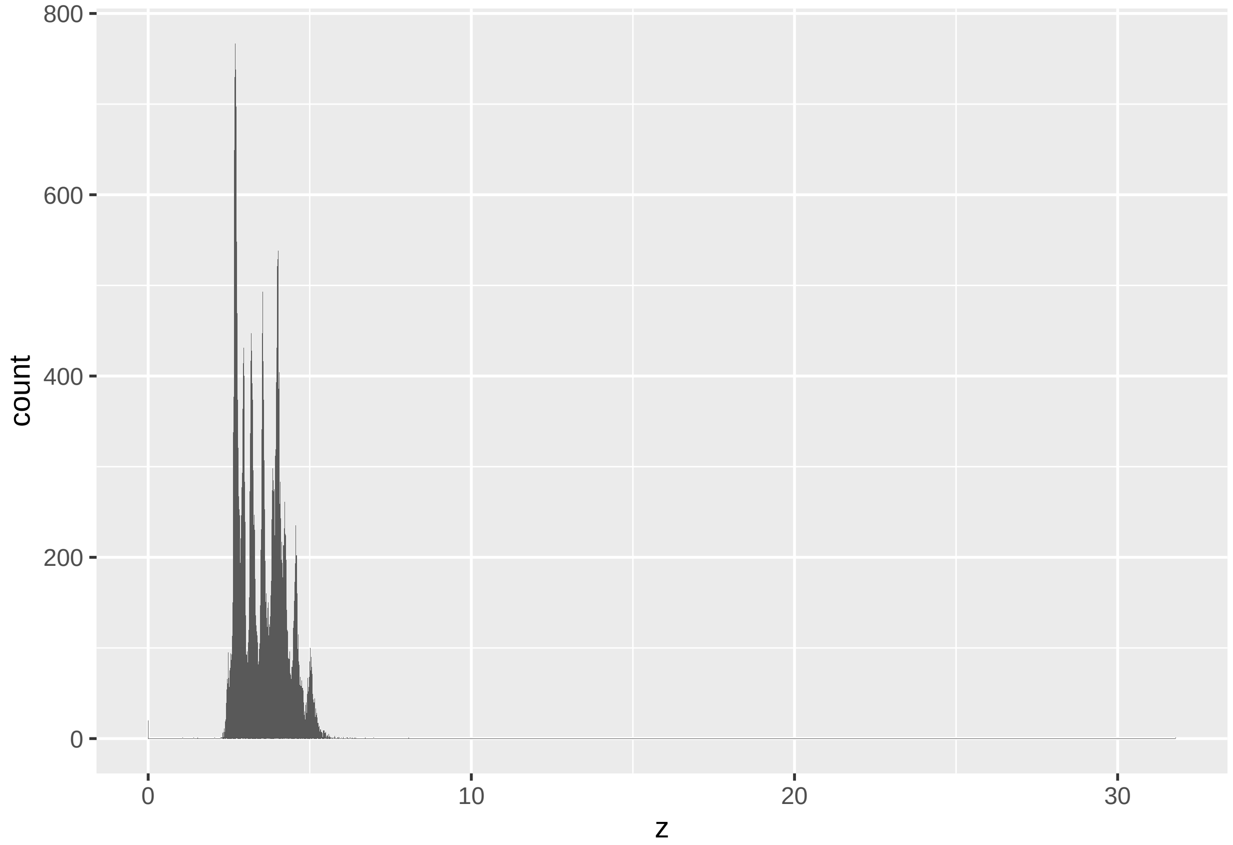
```
ggplot(data = diamonds) +
  geom_histogram(mapping = aes(x = x), binwidth = 0.01)
```



```
ggplot(data = diamonds) +
  geom_histogram(mapping = aes(x = y), binwidth = 0.01)
```



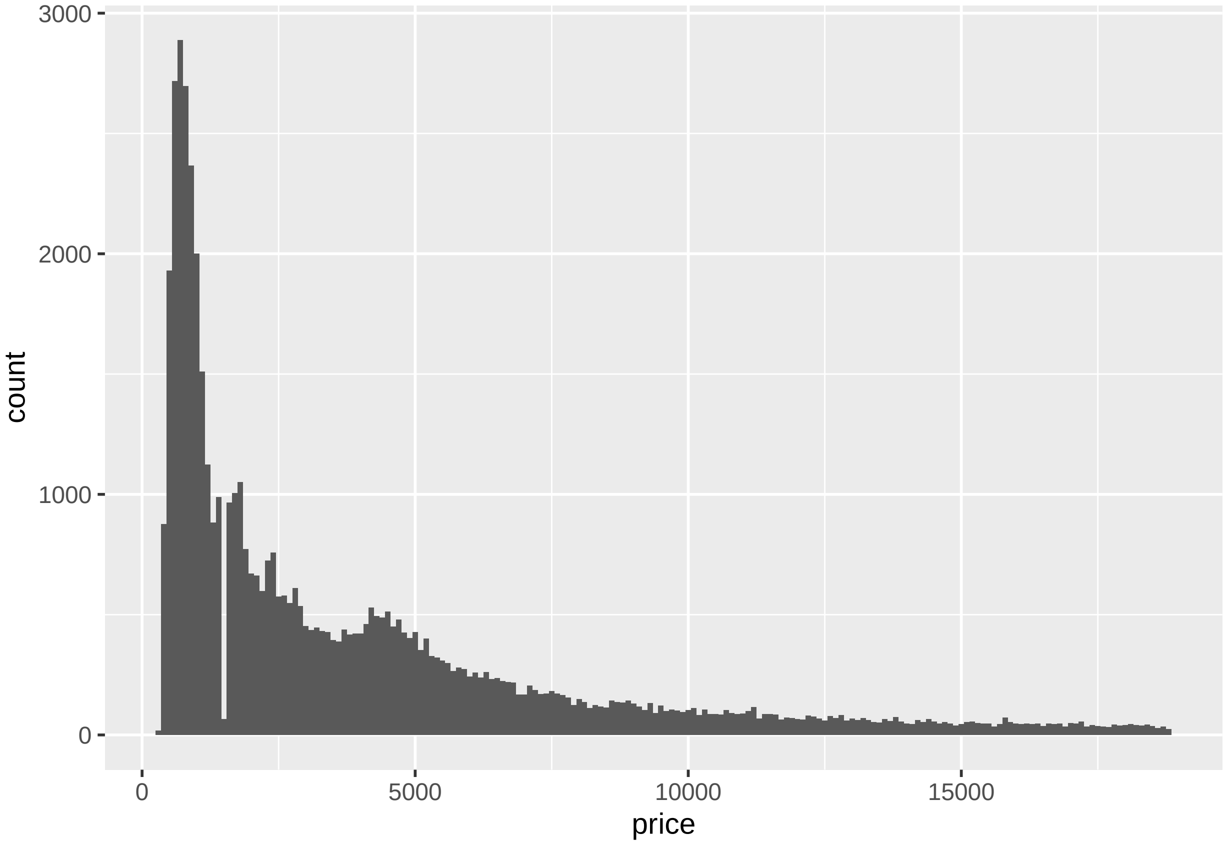
```r
ggplot(data = diamonds) +
  geom_histogram(mapping = aes(x = z), binwidth = 0.01)
```



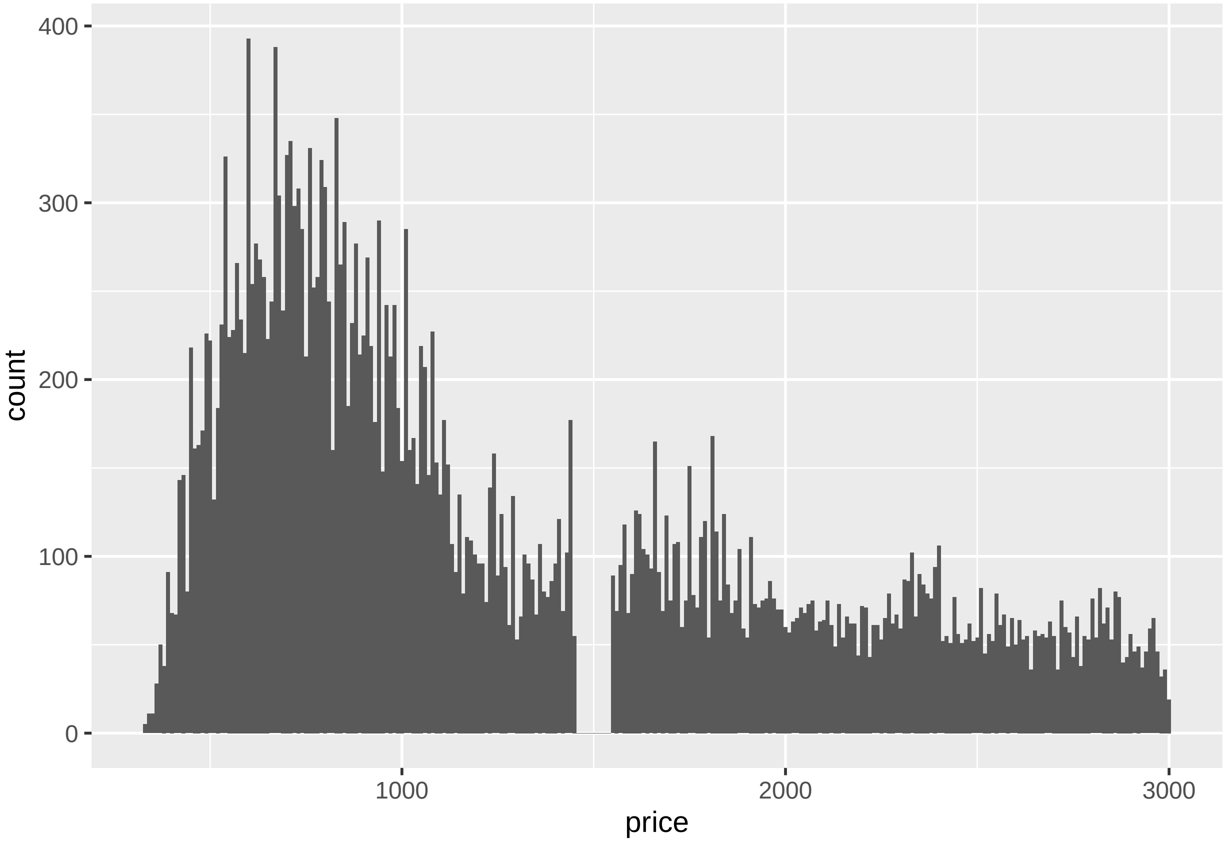
From the distributions, we can learn that x and y are larger than z (x and y having inter-quartile range of 4.71-6.54, whereas, z having inter-quartile range of 2.91-4.04), all right skewed, multimodal, and there are outliers (there are some diamonds with values of zero and some with unusual large values of x, y, or z). For length, width, and depth, I would compare the variables values. The length would be less than the width, and depth should be less than both the length and width since it is expressed as a percentage of length/width of the diamond.

Question 2

```
ggplot(data = diamonds) +
  geom_histogram(mapping = aes(x = price), binwidth = 100, center = 0)
```



```
ggplot(filter(diamonds, price < 3000)) +
  geom_histogram(mapping = aes(x = price), binwidth = 10, center = 0)
```



The data, price, has many spikes. There isn't much difference in distributions in the last one or two digits from the plots above. The distribution has a bulge around $750, and there are no diamonds around the price $1500.

Question 3

```
diamonds %>%
  filter(carat == 0.99) %>%
  count(carat)
```

```
## # A tibble: 1 x 2
##   carat     n
##   <dbl> <int>
## 1  0.99    23
```

```
diamonds %>%
  filter(carat == 1) %>%
  count(carat)
```

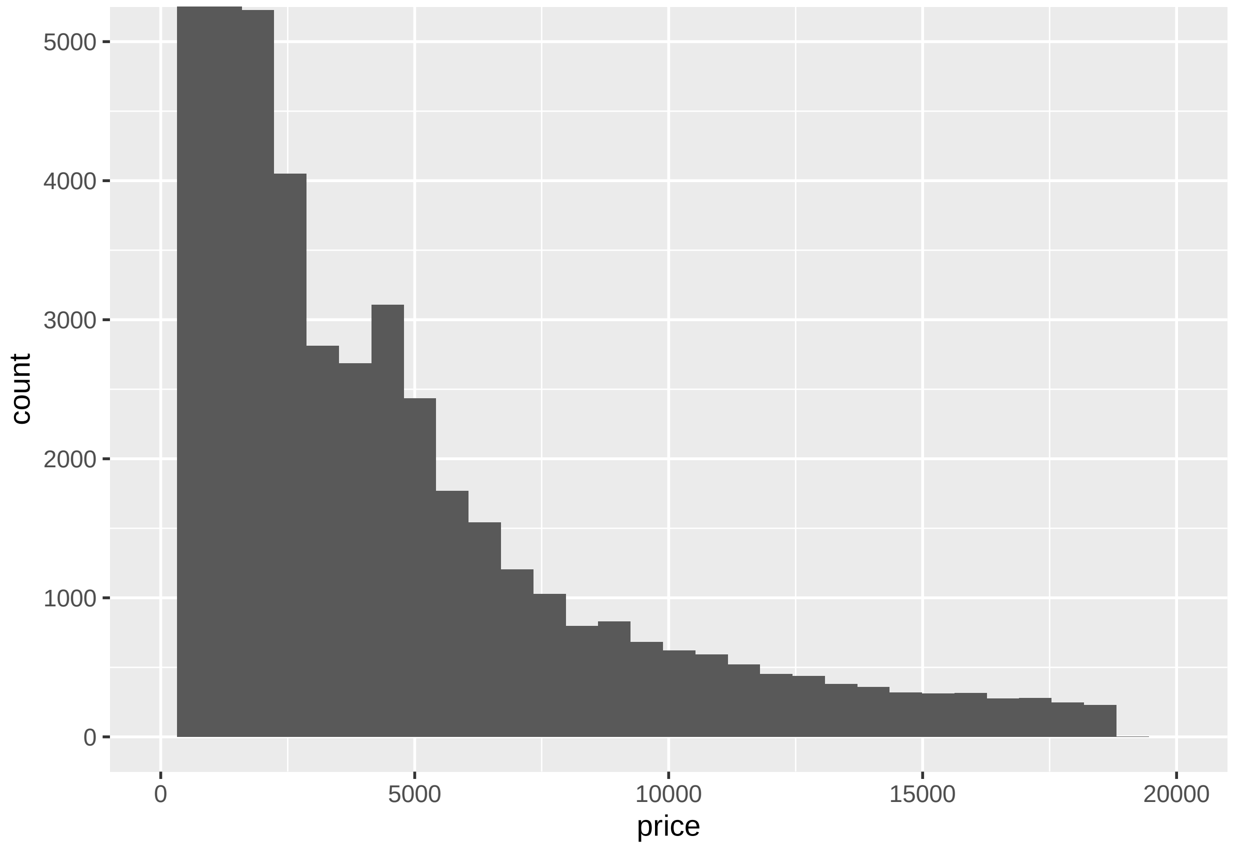
```
## # A tibble: 1 x 2
##   carat     n
##   <dbl> <int>
## 1     1  1558
```

There are 23 diamonds that are 0.99 carat, and there are 1558 diamonds that are 1 carat. This is possibly because sellers usually round up the carat value that is close to a better number, which is 1 in this case. So, there is a higher number of 1s than 0.99s, and this causes the difference.

Question 4

```
ggplot(diamonds) +
  geom_histogram(mapping = aes(x = price)) +
  coord_cartesian(xlim = c(0, 20000), ylim = c(0, 5000))
```

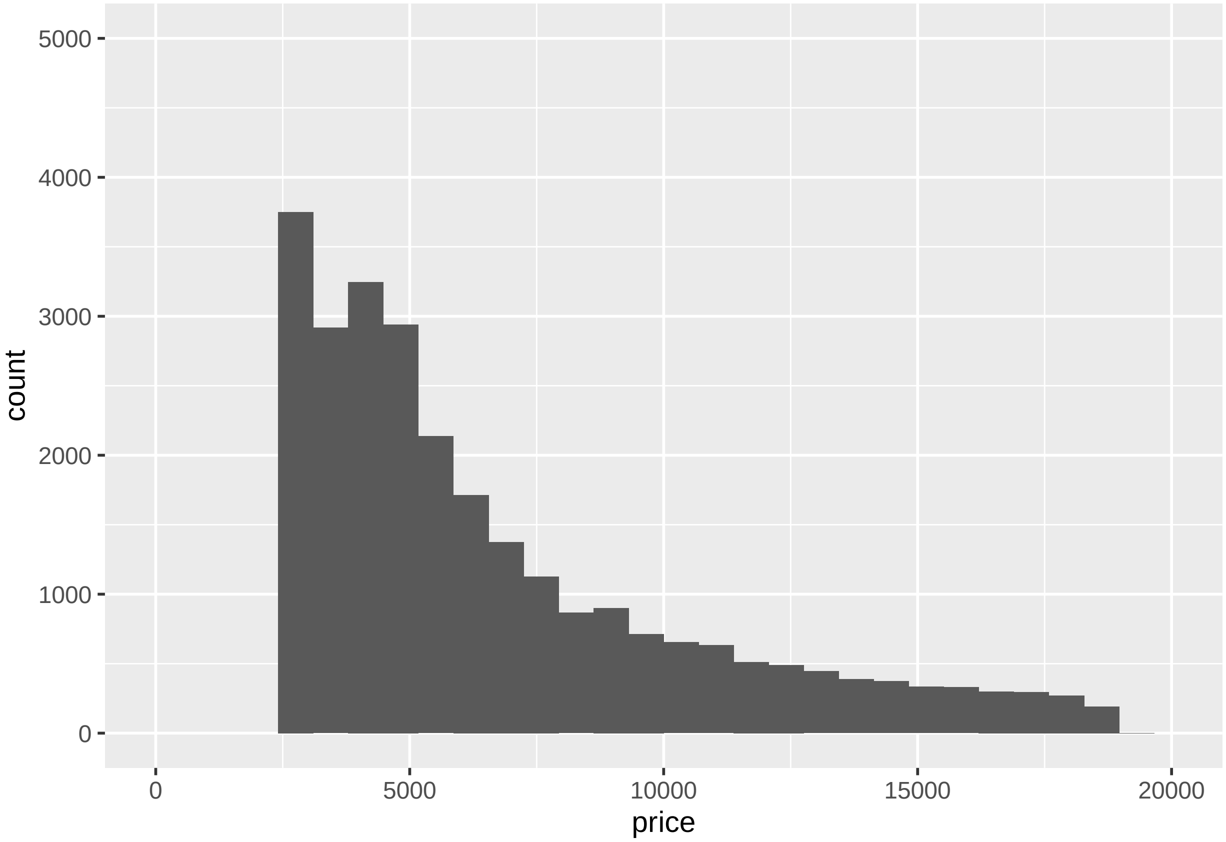
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(diamonds) +
  geom_histogram(mapping = aes(x = price)) +
  xlim(0, 20000) +
  ylim(0, 5000)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 5 rows containing missing values (geom_bar).
```



The coord_cartesian() function is more zoomed in than just xlim() and ylim() functions since it zooms in on the area specified by the limits. Leaving binwidth unset gives a message saying that the default binwidth R has taken for this plot is “bins = 30”, and it states that a better binwidth value can be picked.

Part 7.4.1

Question 1

geom_histogram() removes rows with NA values. Missing values in histogram are removed when the no. of observations in each bin are calculated. Whereas, a geom_bar doesn't remove NA values but is treated as another factor or category instead.

Question 2

na.rm = TRUE removes NA values from the vector when calculating mean and sum.