Lecture Assignment 8

Taiki Yamashita

2024-04-29

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
##
## Attaching package: 'ggstance'
##
##
## The following objects are masked from 'package:ggplot2':
##
##     geom_errorbarh, GeomErrorbarh
```

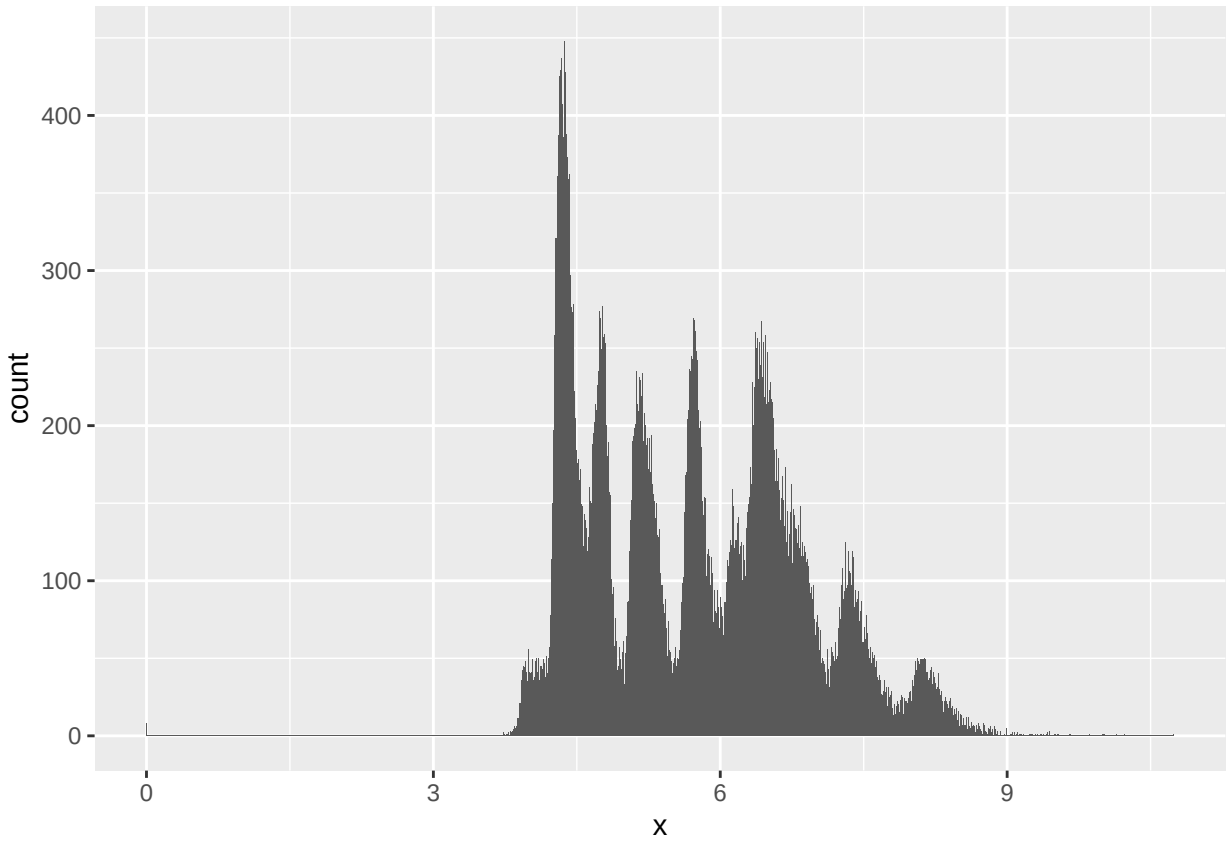
7.3.4 Question 1 ————

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.

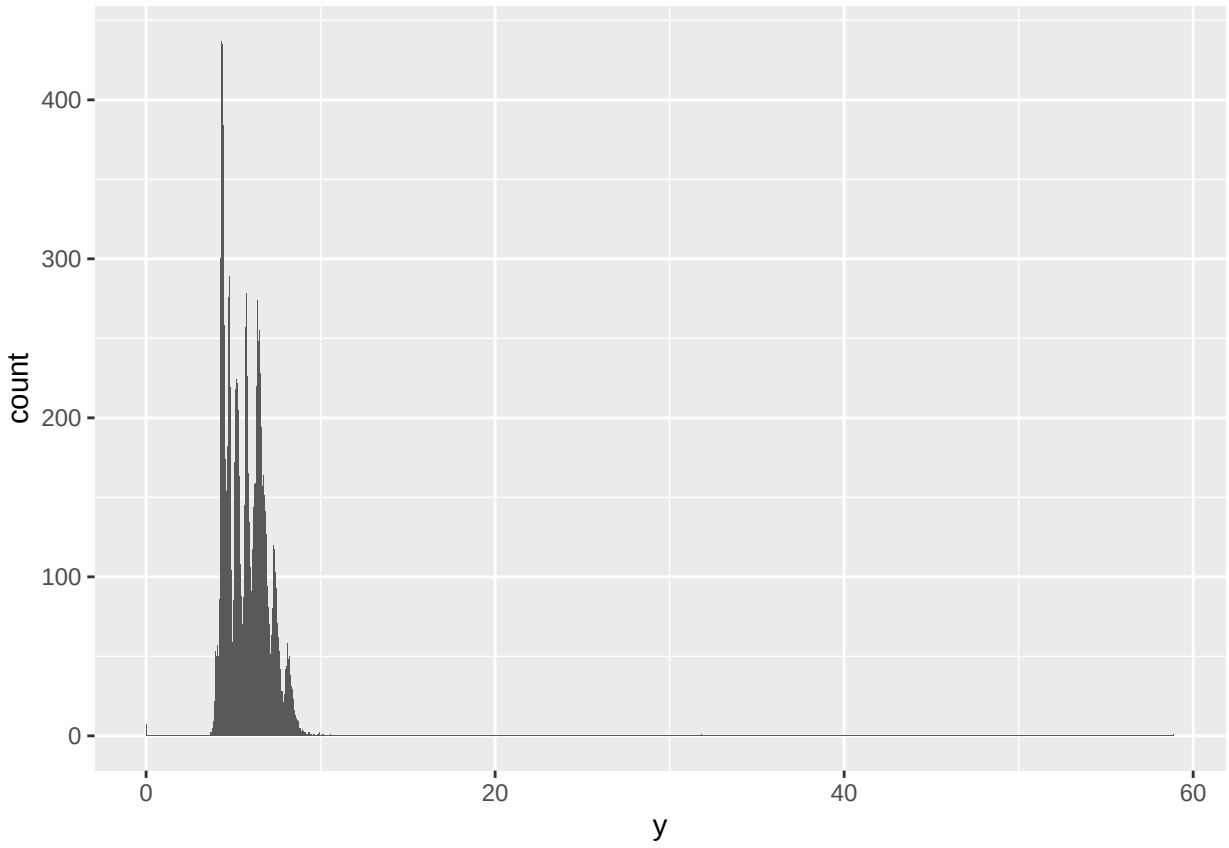
```
# first, we must calculate the summary statistics for these variable and plot their distributions.
summary(select(diamonds, x, y, z))
```

```
##        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
```

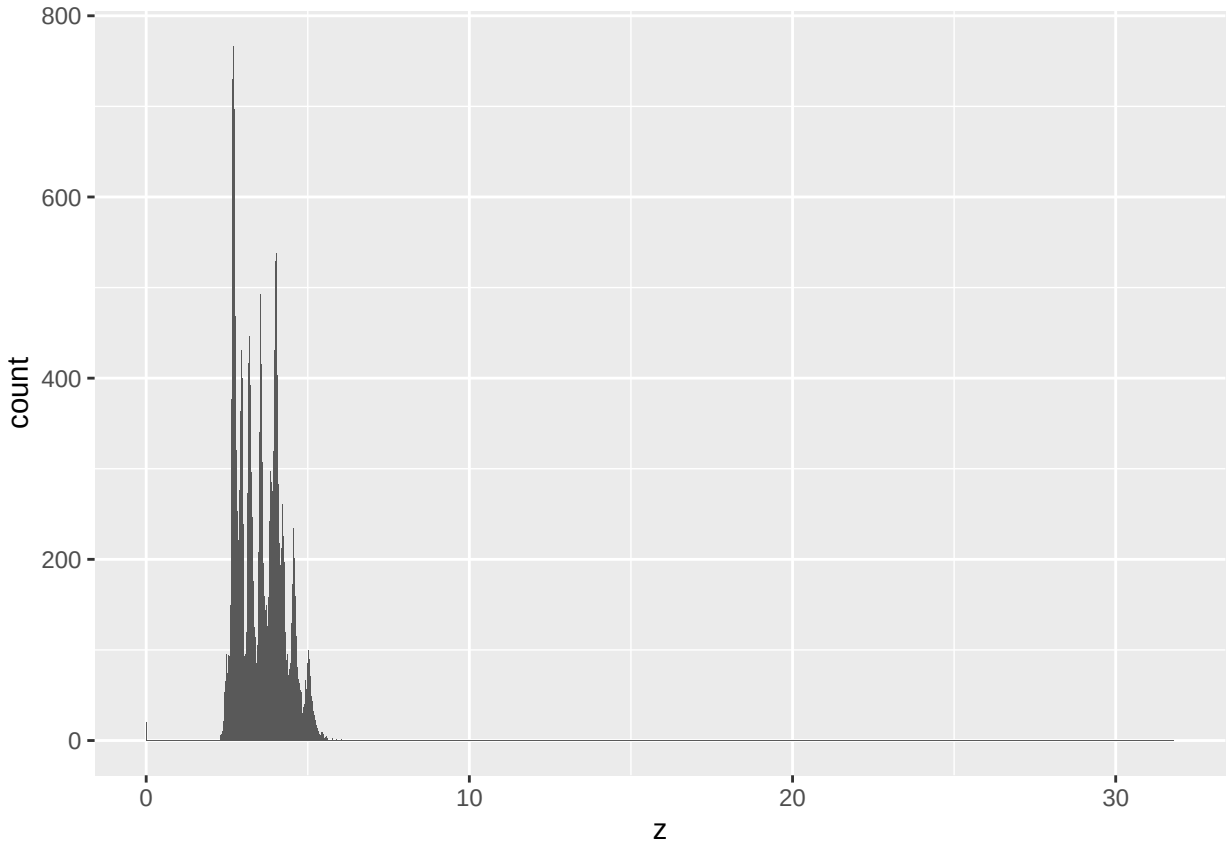
```
# for x
ggplot(diamonds) +
  geom_histogram(mapping = aes(x=x), binwidth = 0.01)
```




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




```
# for z
ggplot(diamonds) +
  geom_histogram(mapping = aes(x=z), binwidth = 0.01)
```

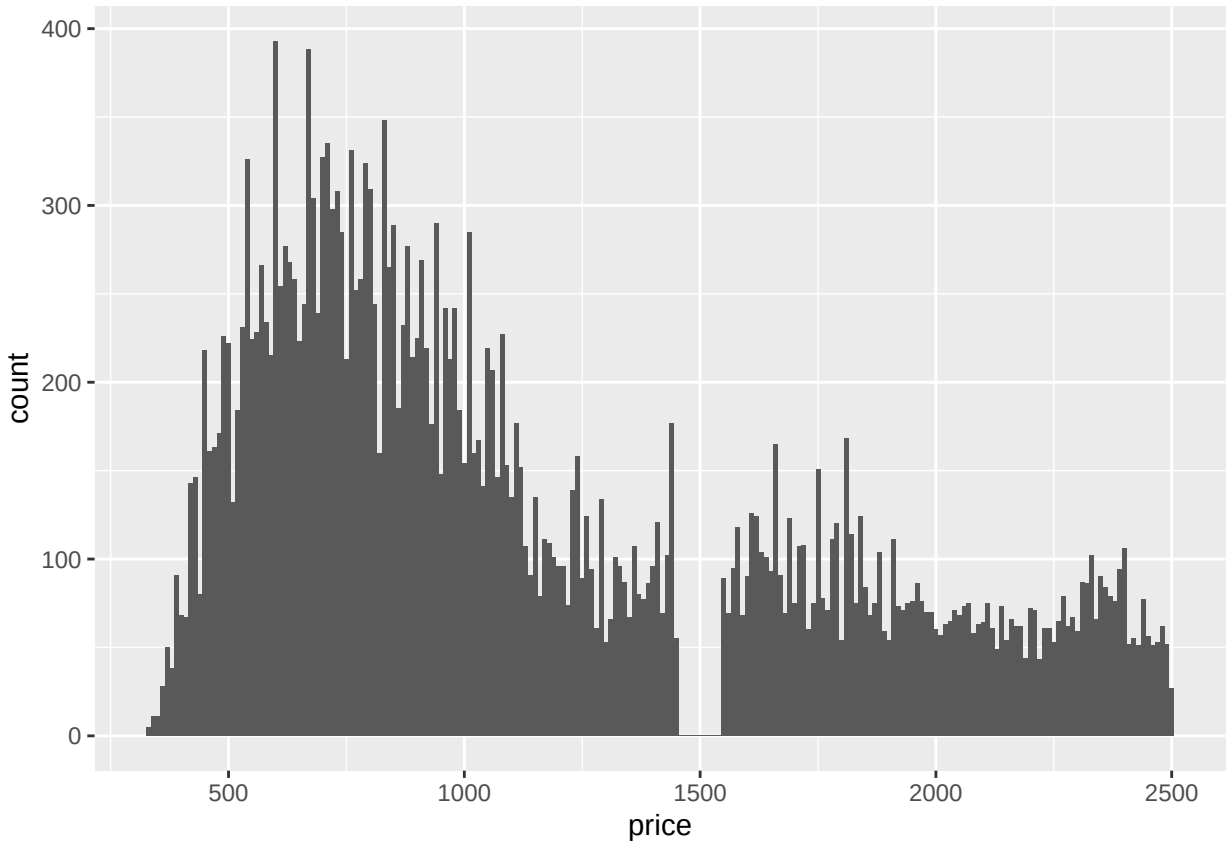


What we can see after exploring each of the distributions with different variables is that x and y are larger than z, all of the distributions are right skewed, there are outliers for each of the graphs, and they are multimodal.

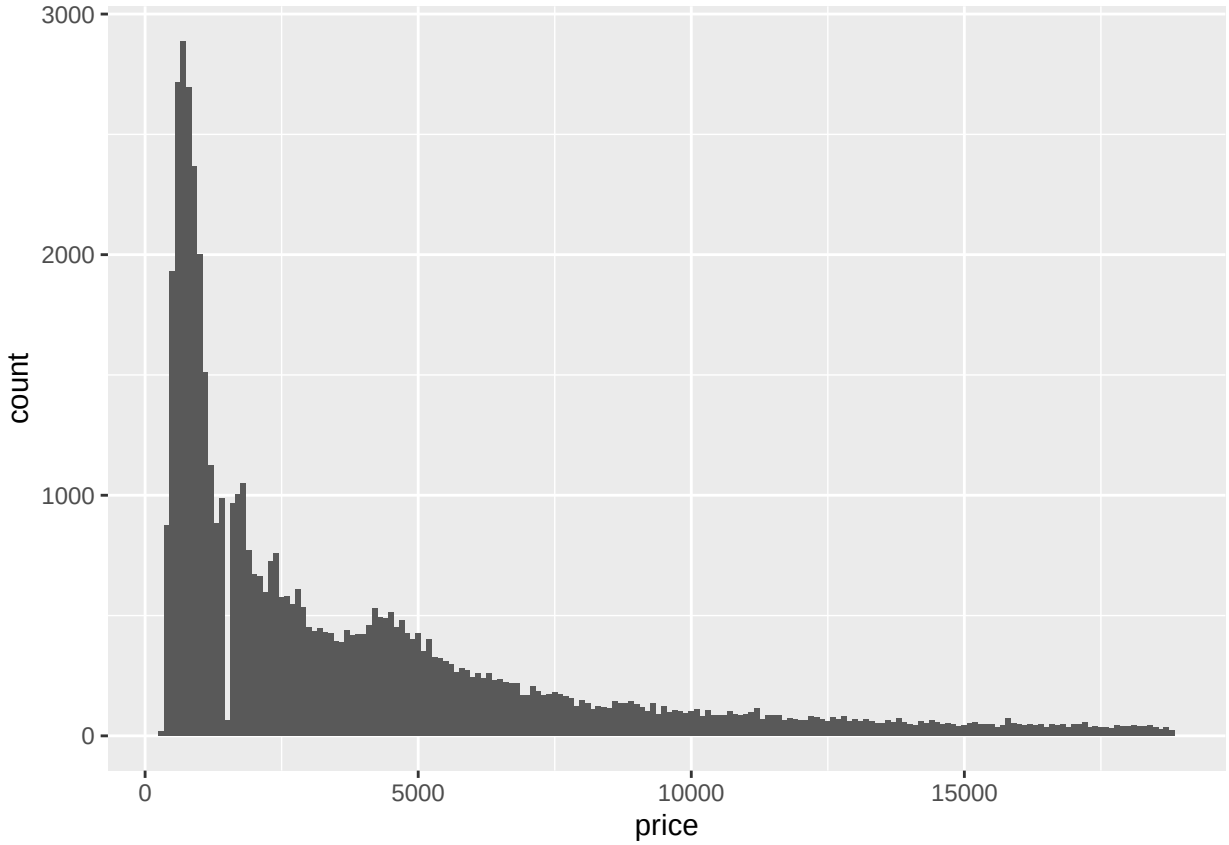
7.3.4 Question 2 ————

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.)

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

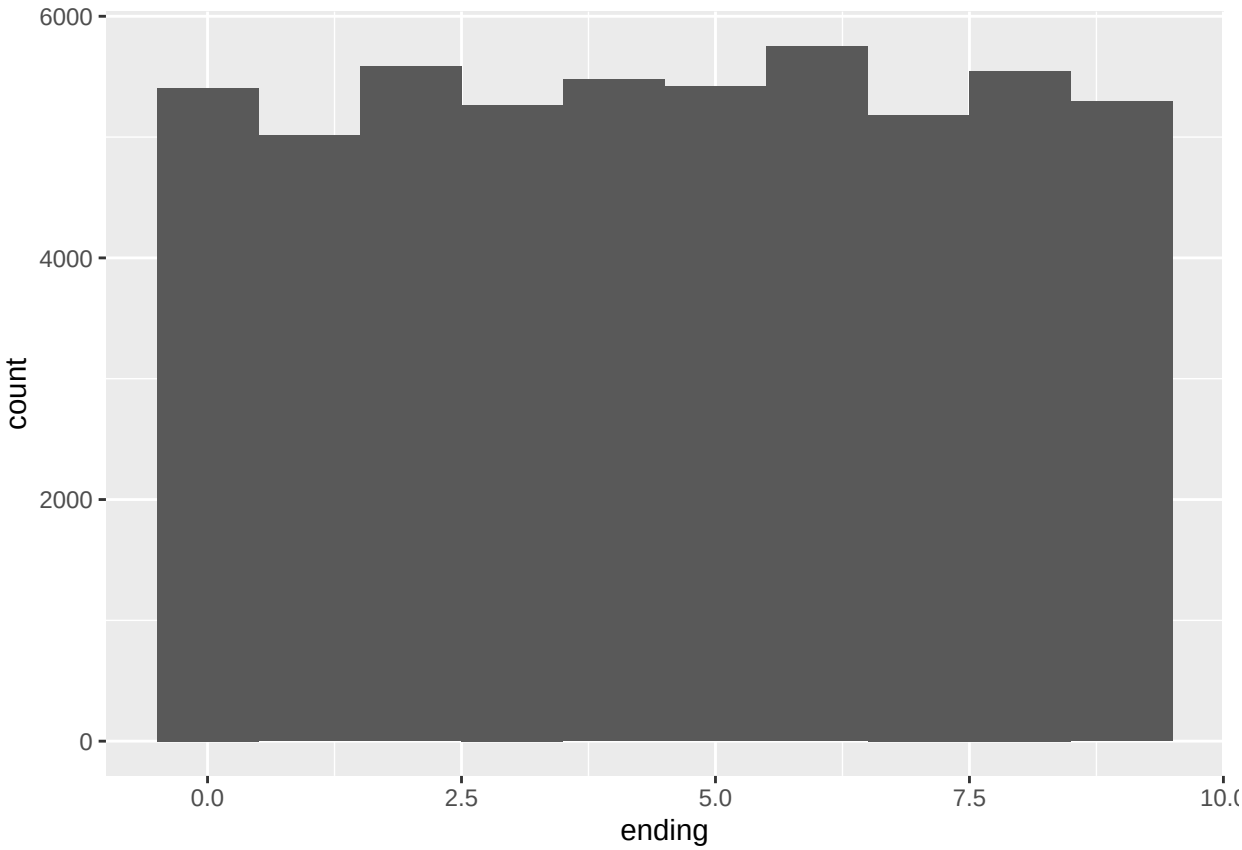



```
ggplot(filter(diamonds), aes(x=price)) +
  geom_histogram(binwidth=100, center=0)
```



The price data has many spikes, but it is hard to tell what each spike corresponds to. The plots do not show much difference in the distribution in the last one or two digits. There are no diamonds with a price of $1500!

```
diamonds %>%
  mutate(ending = price %% 10) %>%
  ggplot(aes(x=ending)) +
  geom_histogram(binwidth=1, center=0)
```



This way it is easier to visualize the difference in the distribution by looking at the last one or two digits specifically.

7.3.4 Question 3 ———

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

7.3.4 Question 4 ———

Compare and contrast coord_cartesian() vs xlim() or ylim() when zooming in on a histogram. What happens if you leave binwidth unset? What happens if you try and zoom so only half a bar shows?

7.4.1 Question 1 ———

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

7.4.1 Question 2 ———

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