More on Exploratory Data Analytics (EDA)

ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960 MGMT-4962/6962 BCBP 4960

Group 1 Module 3(b) Feb 1st, 2021

Graphs (Visualization)..

- *The simple graph has brought more information to the data analyst's mind than any other device.* ~John Tukey

Plot tools/ tips

Combining Plots:

http://statmethods.net/advgraphs/layout.html

How to Read and Use Histograms in R:

http://flowingdata.com/2014/02/27/how-to-read-histograms-and-use-them-in-r/

pairs, gpairs, scatterplot.matrix, clustergram, etc.

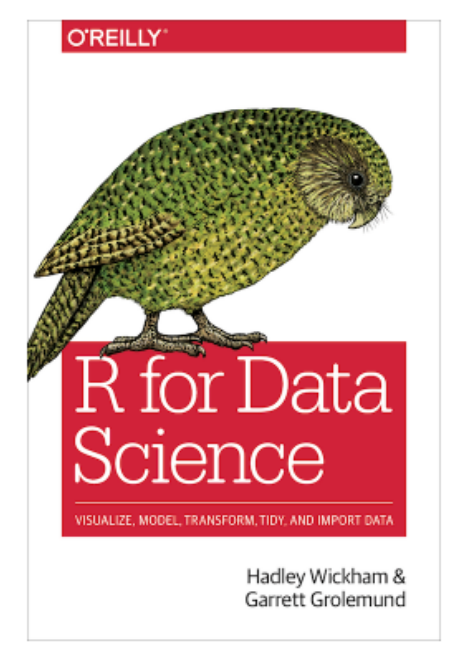
data()

precip, presidents, iris, swiss, sunspot.month (!), environmental, ethanol, ionosphere

More script fragments in R will be available on the web site (http://aquarius.tw.rpi.edu/html/DA)

More on EDA (Resource: **R for Data Science)**

EDA lecture notes are based on the Chapter 5



https://r4ds.had.co.nz/

https://proquestcombo-safaribooksonline-com.libproxy.rpi.edu/9781491910382

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Exploratory **D**ata **A**nalysis

- How to use visualization and transformation to explore your data in a systematic way, a task that statisticians call **E**xploratory **D**ata **A**nalysis, or **EDA** for short. EDA is an iterative cycle.

- You will hear this term “EDA” a lot, throughout the semester…

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

You will:

- **Generate questions about your data**.
- **Search for answers** by **visualizing, transforming, and modeling your data**.
- Use what you learn to refine your questions and/or generate new questions.

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- **EDA is not a formal process with a strict set of rules**. More than anything**, EDA is a state of mind**. During the initial phases of EDA you should feel free to investigate every idea that occurs to you.
- Some of these ideas will pan out, and some will be dead ends.
- As your exploration continues, you will hone in on a few particularly productive areas that you'll eventually write up and communicate to others.

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- **EDA is an important part of any data analysis**, even if the questions are handed to you on a platter, because you always need to investigate the quality of your data.

- **Data cleaning is just one application of EDA**

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- *There are no routine statistical questions, only questionable statistical routines.*

  ~ Sir David Cox

- *Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.*

  ~John Tukey

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Your Goal:

- **Your goal during EDA is to develop an understanding of your data**.
- The easiest way to do this is to use questions as tools to guide your investigation. When you ask a question, **the question focuses your attention on a specific part of your dataset and helps you decide which graphs, models, or transformations to make.**

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

creative process..

- **EDA is fundamentally a creative process**. And like most creative processes, the key to asking *quality* questions is to generate a large *quantity* of questions.
- At the beginning, It is difficult to ask revealing questions at the start of your analysis because you do not know what insights are contained in your dataset.

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

No rule on which question..

- **There is no rule about which questions you should ask to guide your research**.
- However, two types of questions will always be useful for making discoveries within your data. You can loosely word these questions as:

1) What type of variation occurs within my variables?

2) What type of covariation occurs between my variables?

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Variable, Value, Observation

- A ***variable*** is a quantity, quality, or property that you can measure.
- A ***value*** is the state of a variable when you measure it. The value of a variable may change from measurement to measurement.
- An ***observation***, or a *case*, is a set of measurements made under similar conditions (you usually make all of the measurements in an observation at the same time and on the same object). An observation will contain several values, each associated with a different variable. **Sometimes refer to an observation as a data point**.

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Variation…

- ***Variation*** is the tendency of the values of a variable to change from measurement to measurement. You can see variation easily in real life; if you measure any continuous variable twice, you will get two different results. This is true even if you measure quantities that are constant, like the speed of light.
- Each of your measurements will include a small amount of error that varies from measurement to measurement.

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Variation…

- **Categorical variables can also vary if you measure across different subjects (e.g., the eye colors of different people**), or different times (e.g., the energy levels of an electron at different moments).
- Every variable has its own pattern of variation, which can reveal interesting information.
- **The best way to understand the pattern is to visualize the distribution of variables' values**.

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Visualizing Distributions

- How you visualize the distribution of a variable will depend on whether the variable is categorical or continuous. A **variable is *categorical* if it can only take one of a small set of values**.
- **In R, categorical variables are usually saved as factors or character vectors**. To examine the distribution of a categorical variable, use a bar chart:

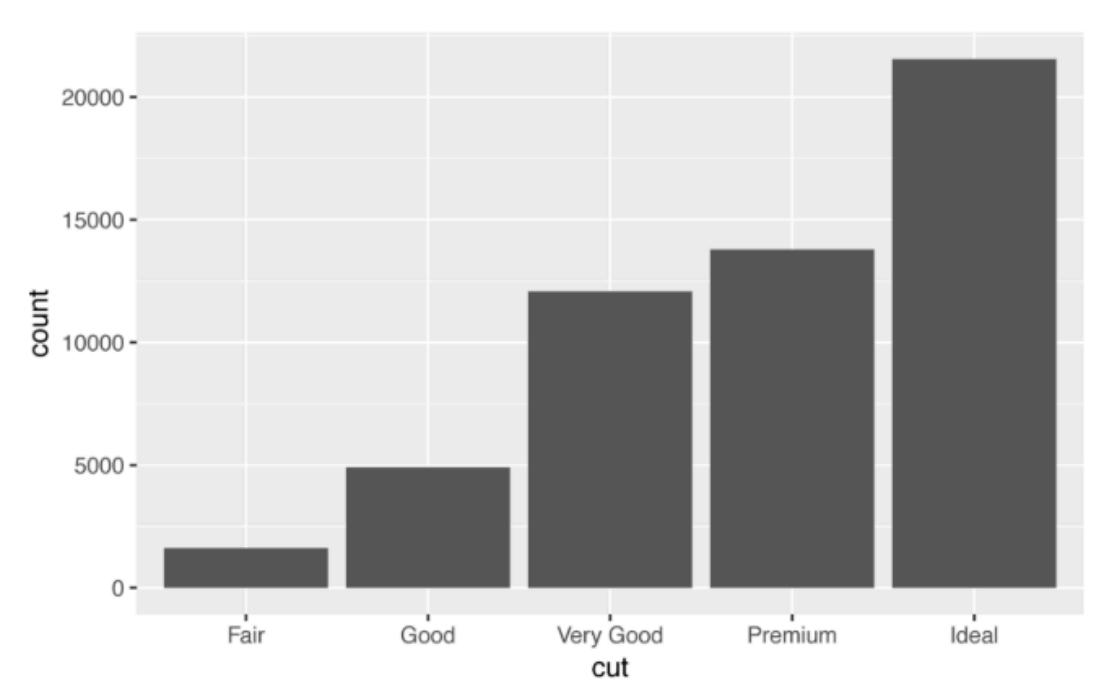
Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Categorical Variables…

- A variable is *categorical* if it can only take one of a small set of values.
- In R, categorical variables are usually saved as factors or character vectors.
- **To examine the distribution of a categorical variable, use a Bar Chart**

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut))



The height of the bars displays how many observations occurred with each x value.

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- The height of the bars displays how many observations occurred with each x value. You can compute these values manually with

dplyr::count():

dplyr::count():

```
diamonds %>%
  count(cut)
#> # A tibble: 5 × 2
#>         cut     n
#>       <ord> <int>
#> 1      Fair  1610
#> 2      Good  4906
#> 3 Very Good 12082
#> 4   Premium 13791
#> 5     Ideal 21551
```

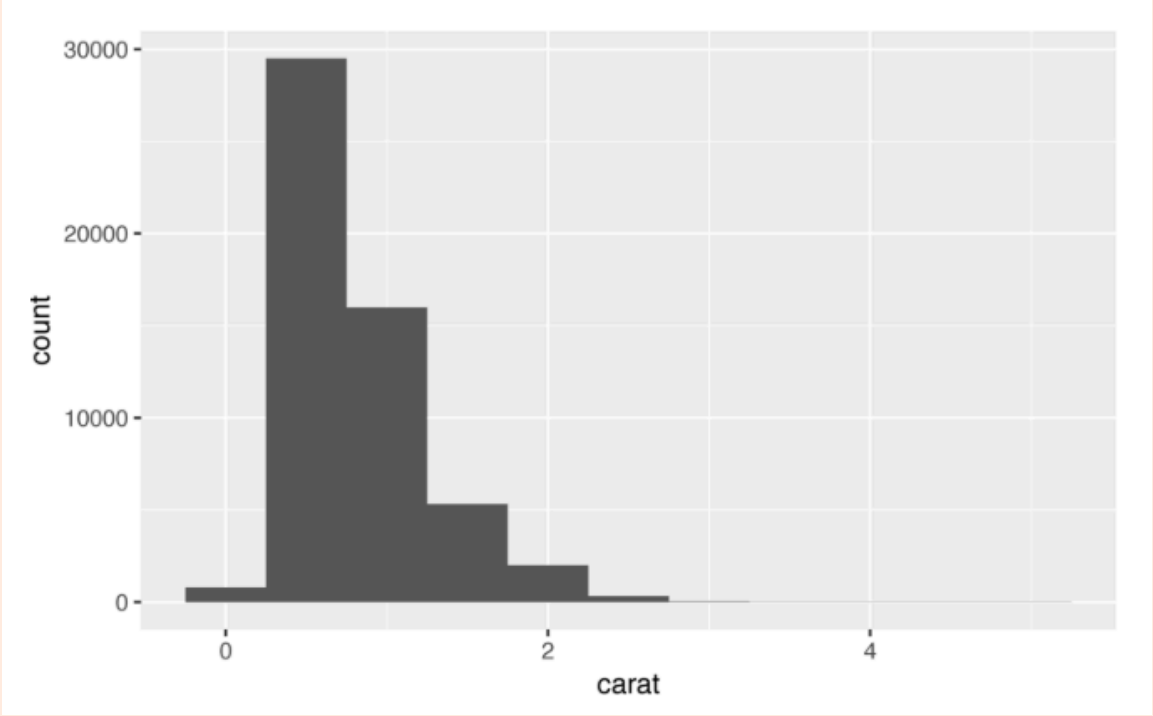

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Continuous Variable…

- A **variable is *continuous* if it can take any of an infinite set of ordered values**.
- Numbers and date-times are two examples of continuous variables.
- **To examine the distribution of a continuous variable, use a histogram**

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- ggplot(data = diamonds) + geom_histogram(mapping = aes(x = carat), binwidth = 0.5)



Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- You can compute this by hand by combining dplyr::count() and ggplot2::cut_width():

diamonds %>% count(cut_width(carat, 0.5))

```
diamonds %>%
  count(cut_width(carat, 0.5))
#> # A tibble: 11 × 2
#>    `cut_width(carat, 0.5)`     n
#>                     <fctr> <int>
#> 1            [-0.25,0.25]   785
#> 2             (0.25,0.75] 29498
#> 3             (0.75,1.25] 15977
#> 4             (1.25,1.75]  5313
#> 5             (1.75,2.25]  2002
#> 6             (2.25,2.75]   322
#> # ... with 5 more rows
```


Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- **A histogram divides the x-axis into equally spaced bins** and then uses the height of each bar to display the number of observations that fall in each bin.
- In the preceding graph, the tallest bar shows that almost 30,000 observations have a carat value between 0.25 and 0.75, which are the left and right edges of the bar.

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- You can set the width of the intervals in a histogram with the ***binwidth*** argument, which is measured in the units of the x variable.
- You should always **explore a variety of *binwidths*** when working with histograms, as **different *binwidths* can reveal different patterns**.

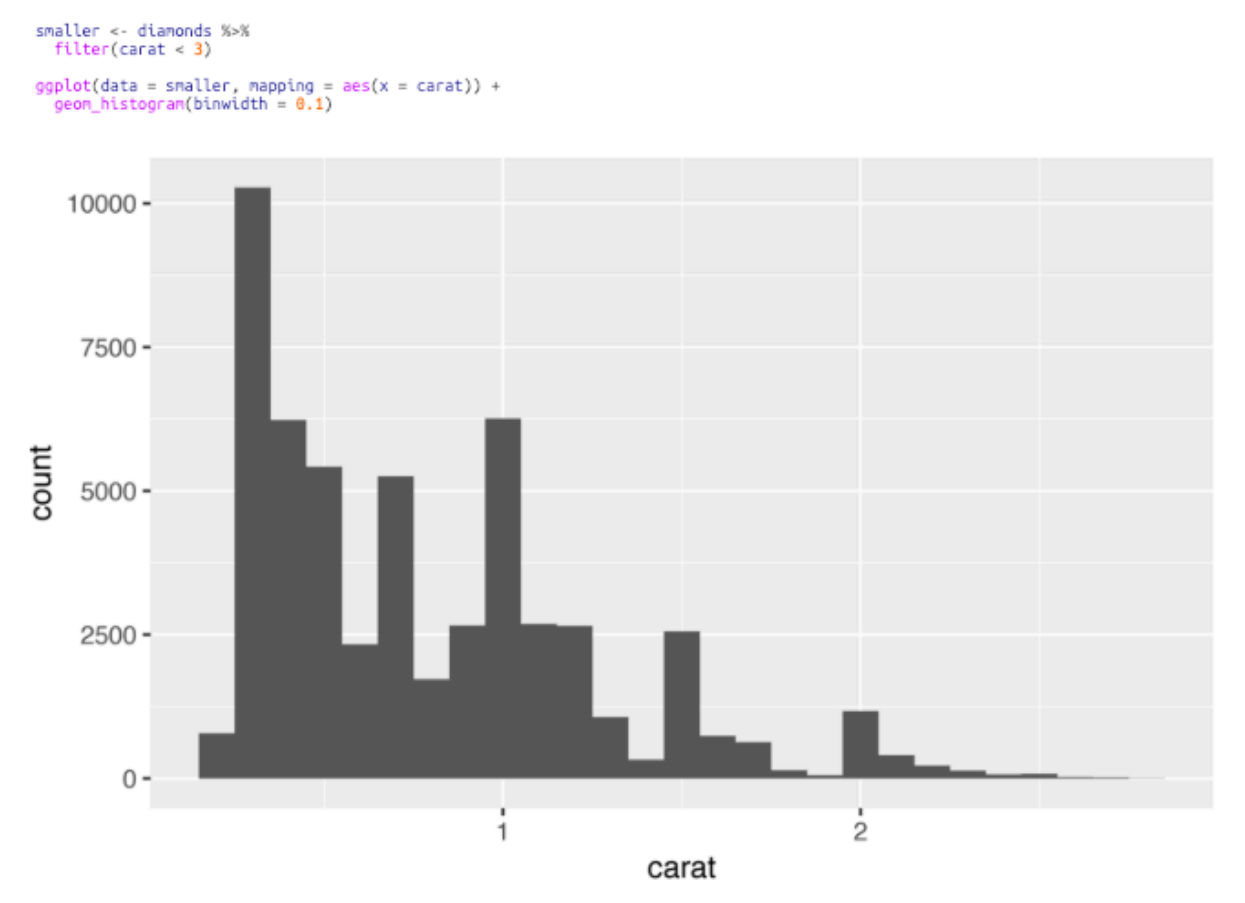
Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- For example, here is how the preceding graph looks when we zoom into just the diamonds with a size of less than three carats and choose a smaller binwidth:

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


Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

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

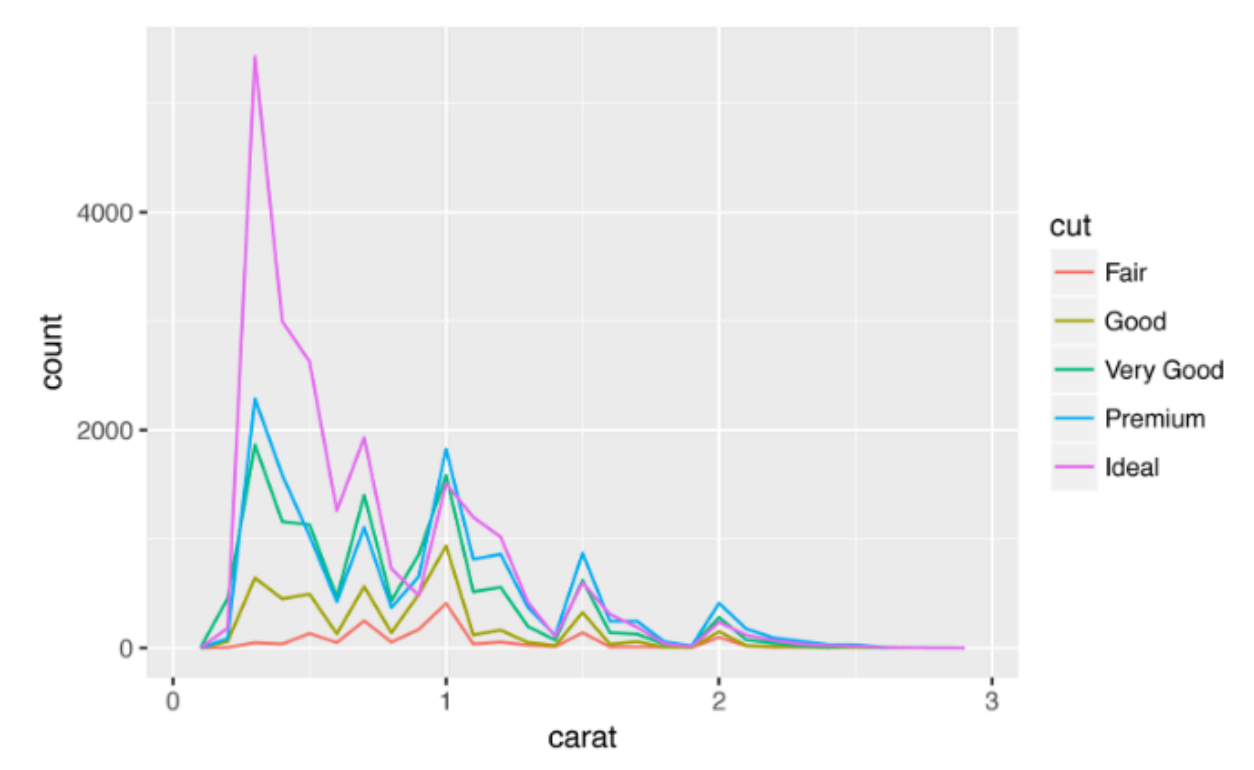


Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- If **you wish to overlay multiple histograms in the same plot**,
  use geom_freqpoly() instead
  of geom_histogram().
- geom_freqpoly() performs the same calculation as geom_histogram(), but instead of displaying the counts with bars, uses lines instead. **It's much easier to understand overlapping lines than bars**:

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- ggplot(data = smaller, mapping = aes(x = carat, color = cut)) + geom_freqpoly(binwidth = 0.1)



Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- There are a few challenges with this type of plot, which we will come back to in "A Categorical and Continuous Variable" later..

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Typical Values

- In both bar charts and histograms, **tall bars show the common values of a variable, and shorter bars show less-common values**.
- Places that do not have bars reveal values that were not seen in your data.
- To turn this information into useful questions, look for anything unexpected:

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Questions…

- Which values are the most common? Why?
- Which values are rare? Why? Does that match your expectations?
- Can you see any unusual patterns? What might explain them?

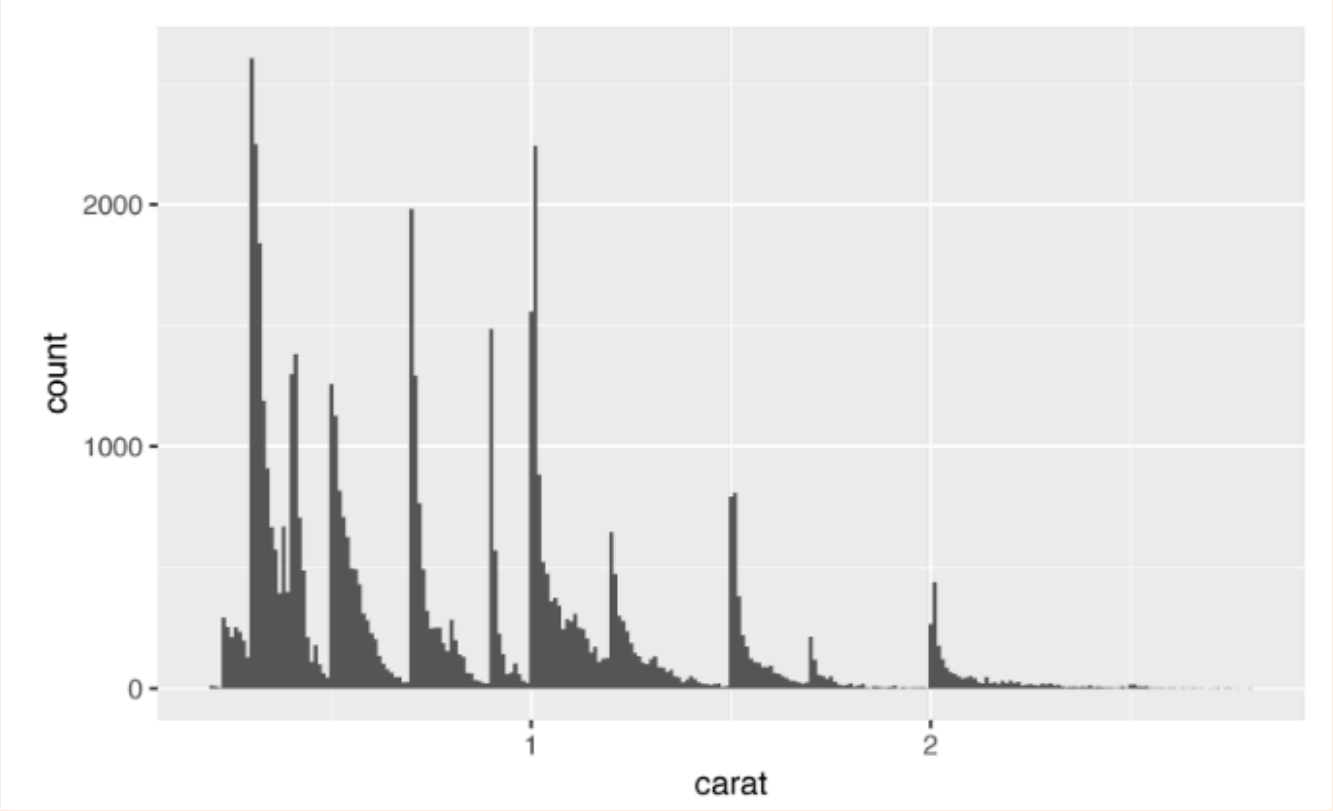
Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Example…

- As an example, the following histogram suggests several interesting questions:
- 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 of each peak?
- Why are there no diamonds bigger than 3 carats?

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

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



- In general, clusters of similar values suggest that subgroups exist in your data

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Clusters…

- In general, clusters of similar values suggest that subgroups exist in your data. **To understand the subgroups**,

Ask:

- **How are the observations within each cluster similar to each other**?
- **How are the observations in separate clusters different from each other**?
- **How can you explain or describe the clusters**?
- **Why might the appearance of clusters be misleading**?

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Relationship *between* variables

- Many of the preceding questions will prompt you to explore:

  a **relationship *between* variables**, for example, to see if the **values of one variable can explain the behavior of another variable**

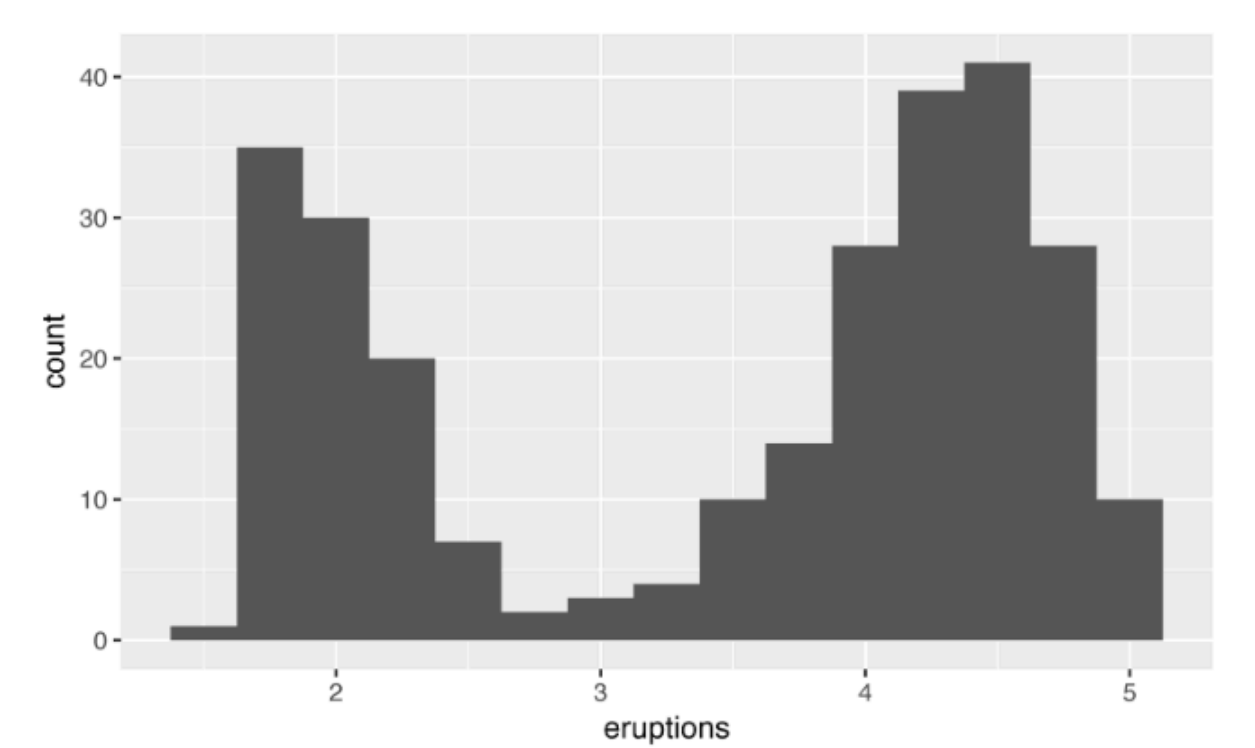
Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- The following histogram shows the length (in minutes) of 272 eruptions of the Old Faithful Geyser in Yellowstone National Park. Eruption times appear to be clustered into two groups: there are short eruptions (of around 2 minutes) and long eruptions (4–5 minutes), but little in between

Image/Photo Credit: yellowestonepark.com

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- ggplot(data = faithful, mapping = aes(x = eruptions)) + geom_histogram(binwidth = 0.25)



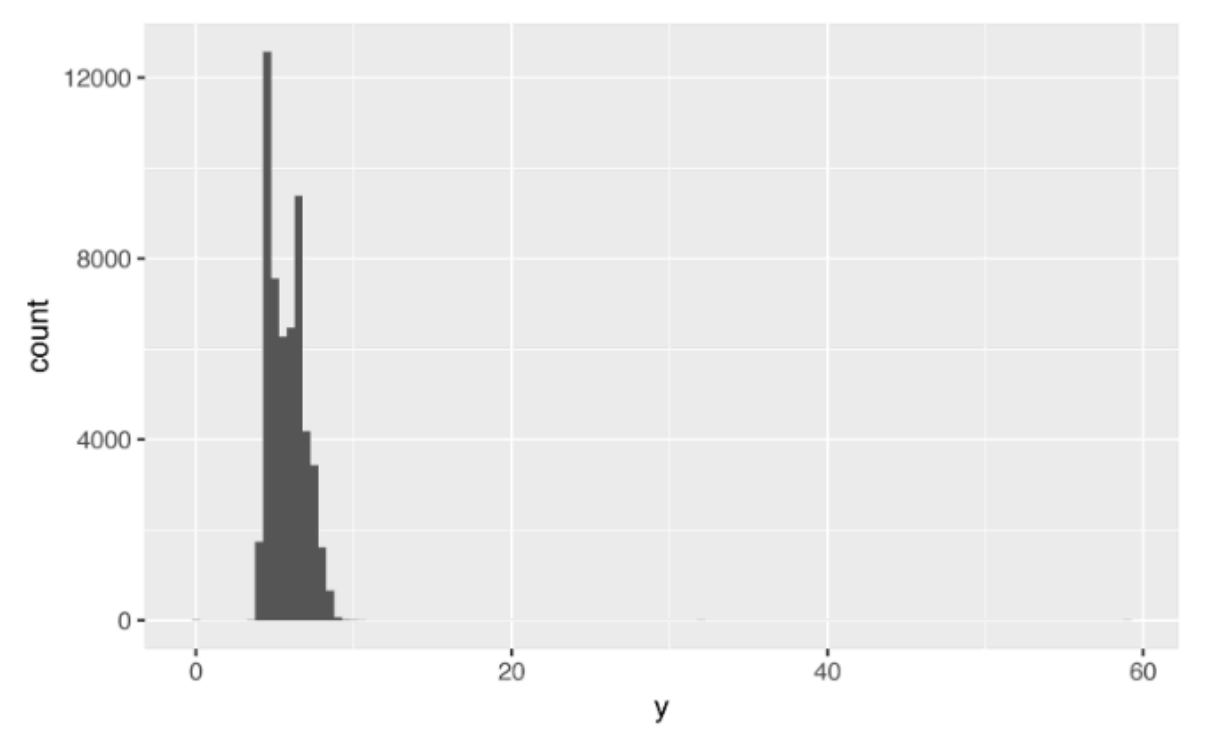
Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Unusual Values…

- **Outliers are observations that are unusual; data points that don't seem to fit the pattern**. Sometimes outliers are data entry errors; other times outliers suggest important new science.
- When you have a lot of data, outliers are sometimes difficult to see in a histogram. For example, take the distribution of the y variable from the diamonds dataset. The only evidence of outliers is the unusually wide limits on the y-axis:

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- ggplot(diamonds) + geom_histogram(mapping = aes(x = y), binwidth = 0.5)



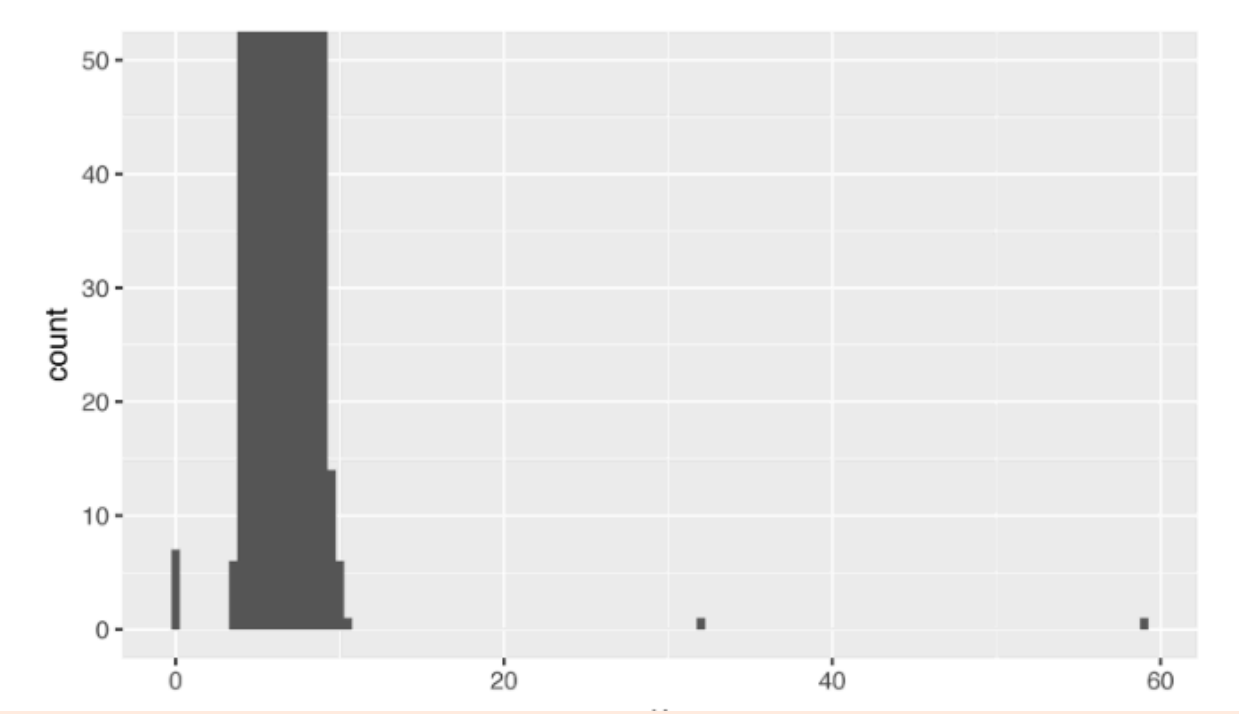
Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- There are so many observations in the common bins that the rare bins are so short that you can't see them (although maybe if you stare intently at 0 you'll spot something). **To make it easy to see the unusual values, we need to zoom in to small values of the y-axis with**

coord_cartesian():

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- ggplot(diamonds) + geom_histogram(mapping = aes(x = y), binwidth = 0.5) + coord_cartesian(ylim = c(0, 50))



Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- (coord_cartesian() also has an xlim() argument for when you need to zoom into the x-axis. **ggplot2** also has xlim() and ylim()functions that work slightly differently: they throw away the data outside the limits.)
- This allows us to see that there are three unusual values: 0, ~30, and ~60. We pluck them out with **dplyr**:

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

```
unusual <- diamonds %>%
  filter(y < 3 | y > 20) %>%
  arrange(y)
unusual
```

```
#> # A tibble: 9 × 10
#>   carat       cut color clarity depth table price     x
#>   <dbl>     <ord> <ord>   <ord> <dbl> <dbl> <int> <dbl>
#> 1  1.00 Very Good     H     VS2  63.3    53  5139  0.00
#> 2  1.14      Fair     G     VS1  57.5    67  6381  0.00
#> 3  1.56     Ideal     G     VS2  62.2    54 12800  0.00
#> 4  1.20   Premium     D    VVS1  62.1    59 15686  0.00
#> 5  2.25   Premium     H     SI2  62.8    59 18034  0.00
#> 6  0.71      Good     F     SI2  64.1    60  2130  0.00
#> 7  0.71      Good     F     SI2  64.1    60  2130  0.00
#> 8  0.51     Ideal     E     VS1  61.8    55  2075  5.15
#> 9  2.00   Premium     H     SI2  58.9    57 12210  8.09
#> # ... with 2 more variables:
#> #   y <dbl>, z <dbl>
```


Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- The y variable measures one of the three dimensions of these diamonds, in mm. **We know that diamonds can't have a width of 0mm**, **so these values must be incorrect**.
- We might also suspect that measurements of 32mm and 59mm are implausible: those diamonds are over an inch long, but don't cost hundreds of thousands of dollars!

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Outliers…

- **It's good practice to repeat your analysis with and without the outliers.** If they have minimal effect on the results, and you can't figure out why they're there, it's reasonable to replace them with missing values and move on.
- However, if they have a substantial effect on your results, you shouldn't drop them without justification. You'll need to figure out what caused them (e.g., a data entry error) and disclose that you removed them in your write-up.

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Exercises (In-Class Work)

- 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.

- 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.)

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

- Compare and contrast coord_cartesian() versus 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?

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Missing Values

- If you've encountered unusual values in your dataset, and simply want to move on to the rest of your analysis, you have two options:

1) Drop the entire row with the strange values

2) Replacing the unusual values with missing values

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- Drop the entire row with the strange Values

```
diamonds2 <- diamonds %>%
filter(between(y, 3, 20))
```

(Not recommended this option as just because one measurement is invalid, doesn't mean all the measurements are. Additionally, if you have low-quality data, by time that you've applied this approach to every variable you might find that you don't have any data left!)

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- Replacing the unusual values with missing values

```
diamonds2 <- diamonds %>% mutate(y = ifelse(y < 3 | y > 20, NA, y))
```

- (recommended replacing the unusual values with missing values. The easiest way to do this is to use mutate() to replace the variable with a modified copy. You can use the ifelse() function to replace unusual values with NA)
- ifelse() has three arguments. The first argument test should be a logical vector. The result will contain the value of the second argument, yes, when test is TRUE, and the value of the third argument, no, when it is false.

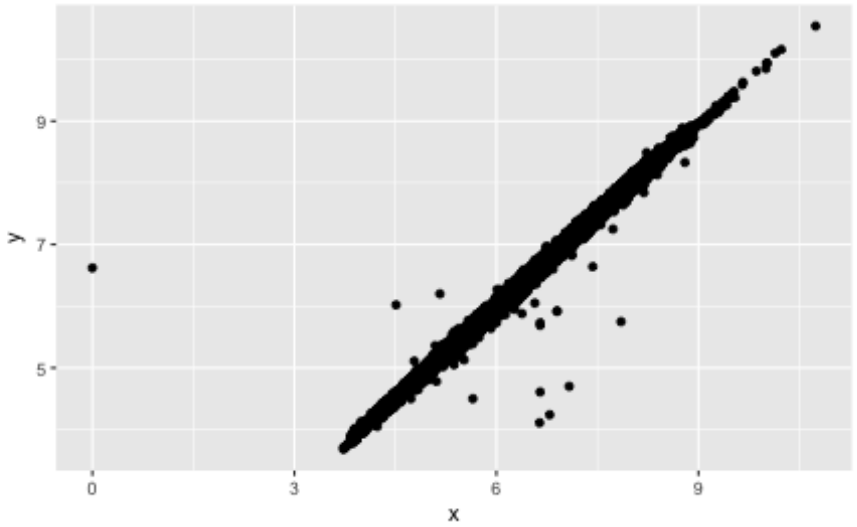
More on this later….

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- Like R, **ggplot2** subscribes to the philosophy that missing values should never silently go missing. It's not obvious where you should plot missing values, so **ggplot2** doesn't include them in the plot, but it does warn that they've been removed:

- ggplot(data = diamonds2, mapping = aes(x = x, y = y)) + geom_point()

```
ggplot(data = diamonds2, mapping = aes(x = x, y = y)) +
  geom_point()
#> Warning: Removed 9 rows containing missing values
#> (geom_point).
```



Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- To suppress that warning, set **na.rm = TRUE**:

```
ggplot(data = diamonds2, mapping = aes(x = x,
y = y)) + geom_point(na.rm = TRUE)
```


Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- Other times you want to understand what makes observations with missing values different from observations with recorded values.
- For example, in nycflights13::flights, missing values in the dep_time variable indicate that the flight was cancelled. So you might want to compare the scheduled departure times for cancelled and noncancelled times. You can do this by making a new variable with is.na():

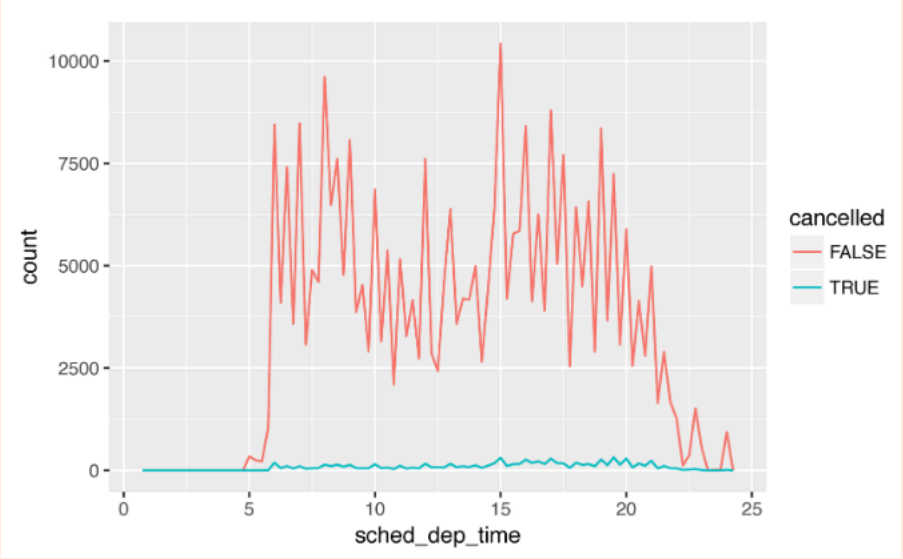
Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

NYC Flights

- 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(mapping = aes(sched_dep_time)) + geom_freqpoly(mapping = aes(color = cancelled), binwidth = 1/4)

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

```r
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(mapping = aes(sched_dep_time)) +
    geom_freqpoly(
      mapping = aes(color = cancelled),
      binwidth = 1/4
    )
```



However, this plot isn't great because there are many more non-cancelled flights than cancelled flights. Later, we'll explore some techniques for improving this comparison.

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Exercises

1. What happens to missing values in a histogram? What happens to missing values in a bar chart? Why is there a difference?
2. What does na.rm = TRUE do in mean() and sum()?

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/