Welcome to DATA 151

I'm so glad you're here!

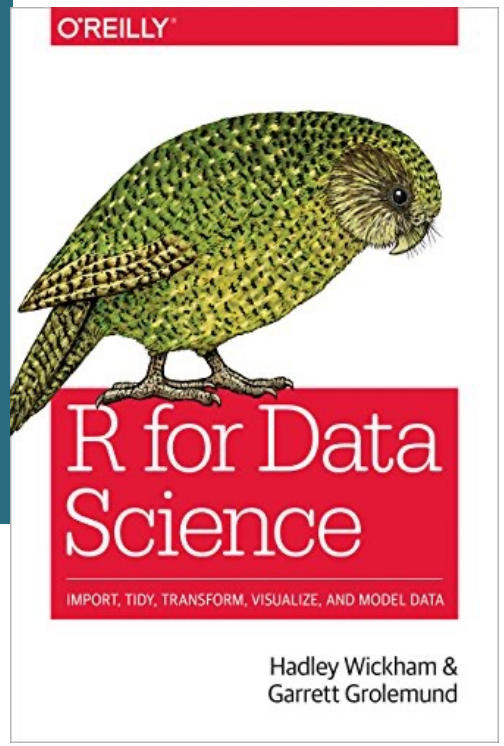
DATA 151: CLASS 6A
INTRODUCTION TO DATA SCIENCE (WITH R)

EXPLORATORY DATA ANALYSIS

NOTES PREPARED BY PROF. KITADA SMALLEY (FALL 2022)

ANNOUNCEMENTS

RELEVANT READING



Introduction to Data Science:

- Tuesday:
  - Introduction to Data Science
  - Ch 7: Introduction to data viz
  - Ch 8: ggplot2
- Thursday:
  - R for Data Science
  - Ch 7: Exploratory Data Analysis

HOMEWORK REMINDER

Due this week: (DUE 10/6)

- *HW #5: DC Importing Data in R*
  - *Only one chapter (not the whole course)*
  - ***No submission on WISE necessary, do on DataCamp***
- *Project Milestone #2: Project Meetings*
  - Each group must have their data set approved by class on Thursday

HOMEWORK REMINDER

Due next week: (DUE 10/13)

- *HW #6: DC Introduction to Data Visualization in ggplot2*
  - ***No submission on WISE necessary, do on DataCamp***
- *Project Milestone #3: EDA Step 1*
  - Ask questions and form hypotheses

EXPLORATORY DATA ANALYSIS

AN ITERATIVE CYCLE

EDA is an iterative cycle. You:

1. Generate questions about your data.
2. Search for answers by visualising, transforming, and modelling your data.
3. Use what you learn to refine your questions and/or generate new questions.

AN ITERATIVE CYCLE

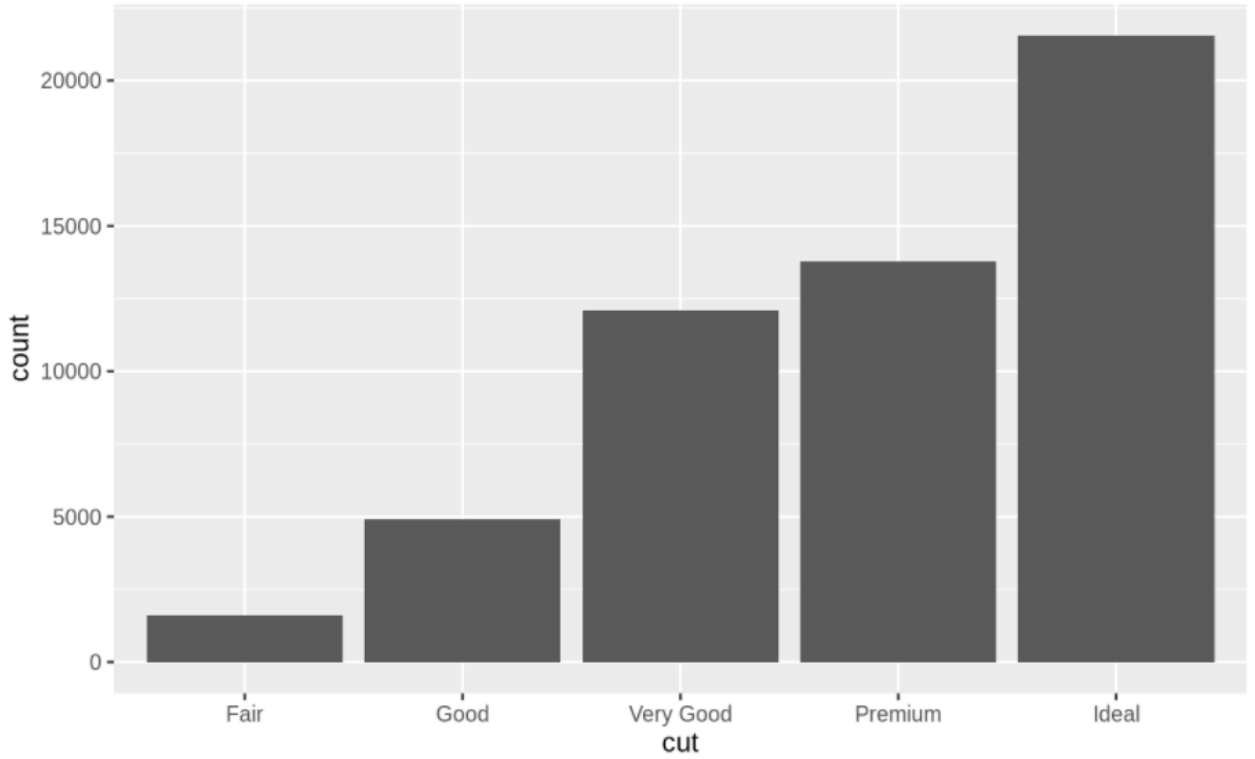
"EDA is not a formal process with a strict set of rules. More than anything, EDA is a state of mind."

QUESTIONS TO ASK YOURSELF

1. What type of variation occurs within my variables?
2. Which values are the most common? Why?
3. Which values are rare? Why? Does that match your expectations?
4. Can you see any unusual patterns? What might explain them?
5. What type of covariation occurs between my variables?

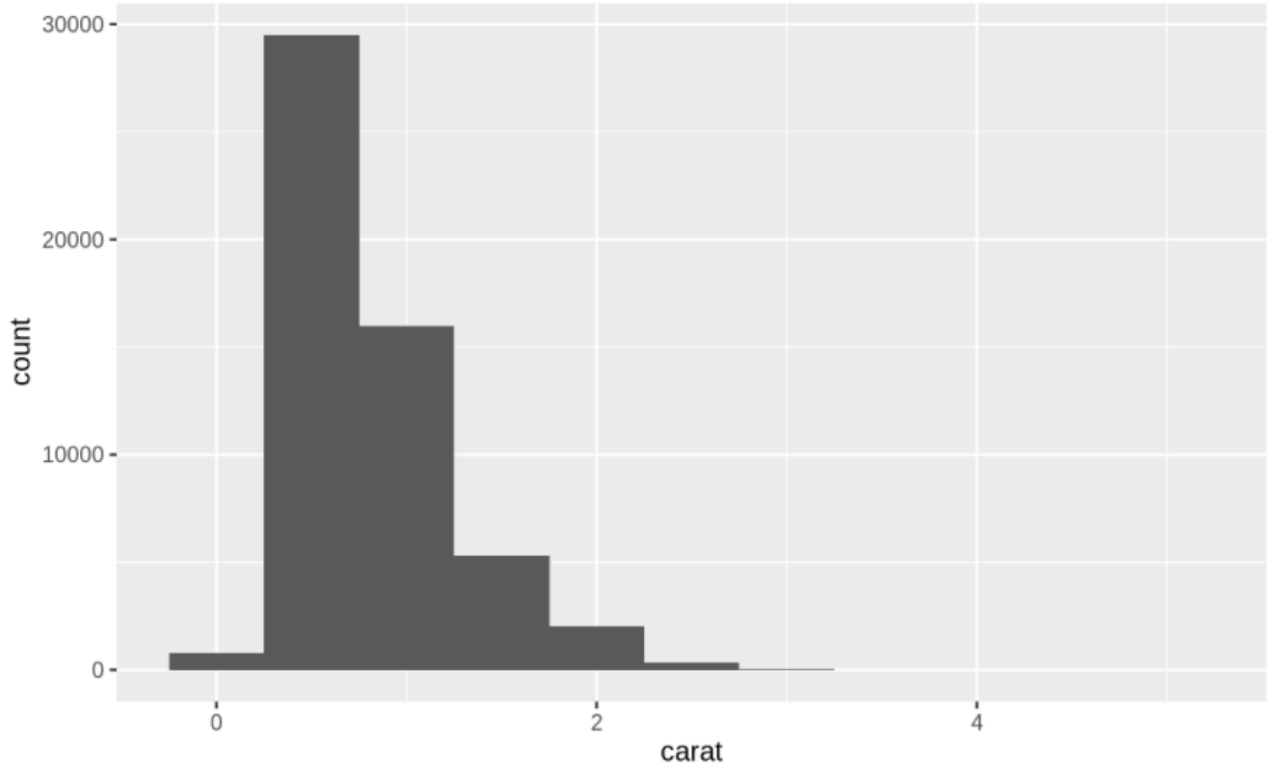
VISUALIZING DISTRIBUTIONS

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



VISUALIZING DISTRIBUTIONS

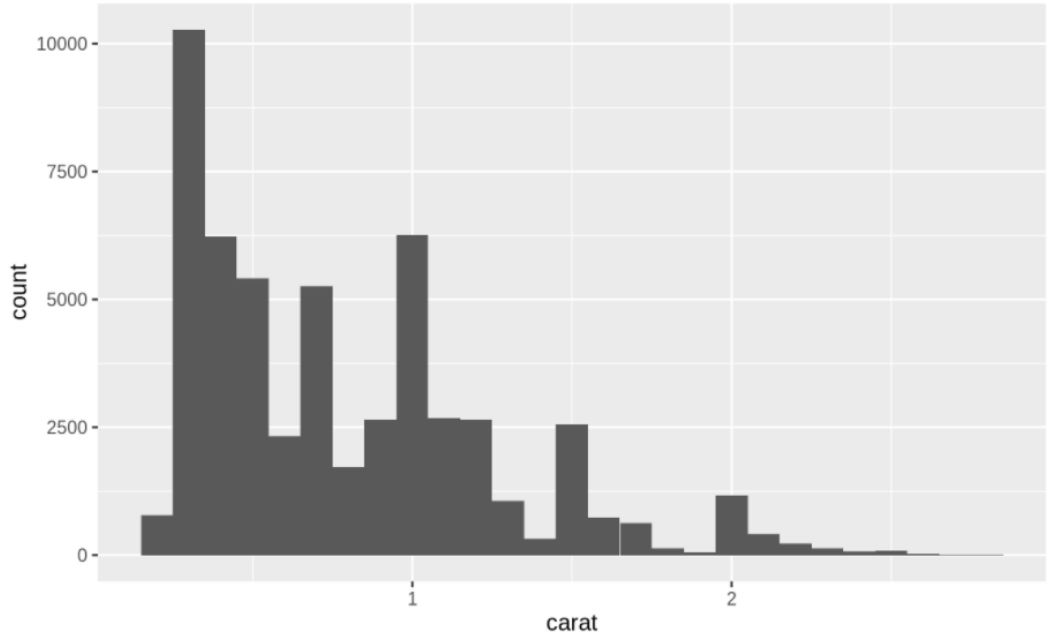
```
ggplot(data = diamonds) +
  geom_histogram(mapping = aes(x = carat), binwidth = 0.5)
```



VISUALIZING DISTRIBUTIONS

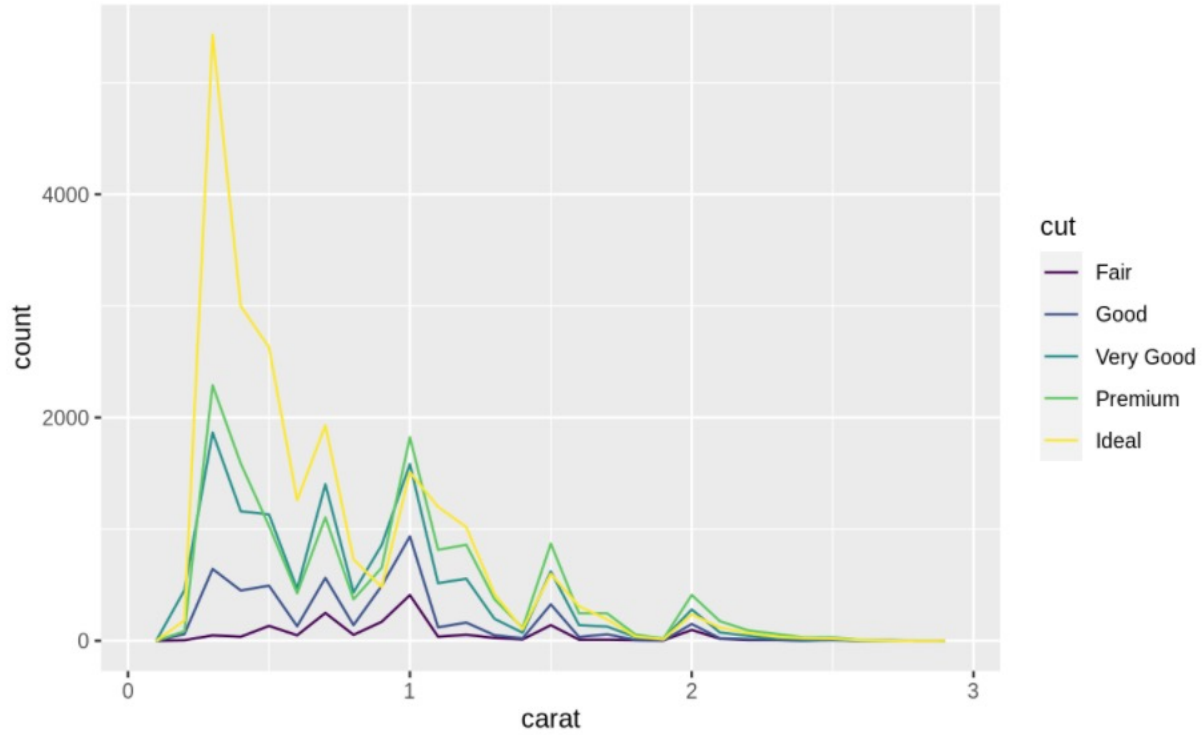
```r
smaller <- diamonds %>%
  filter(carat < 3)

ggplot(data = smaller, mapping = aes(x = carat)) +
  geom_histogram(binwidth = 0.1)
```



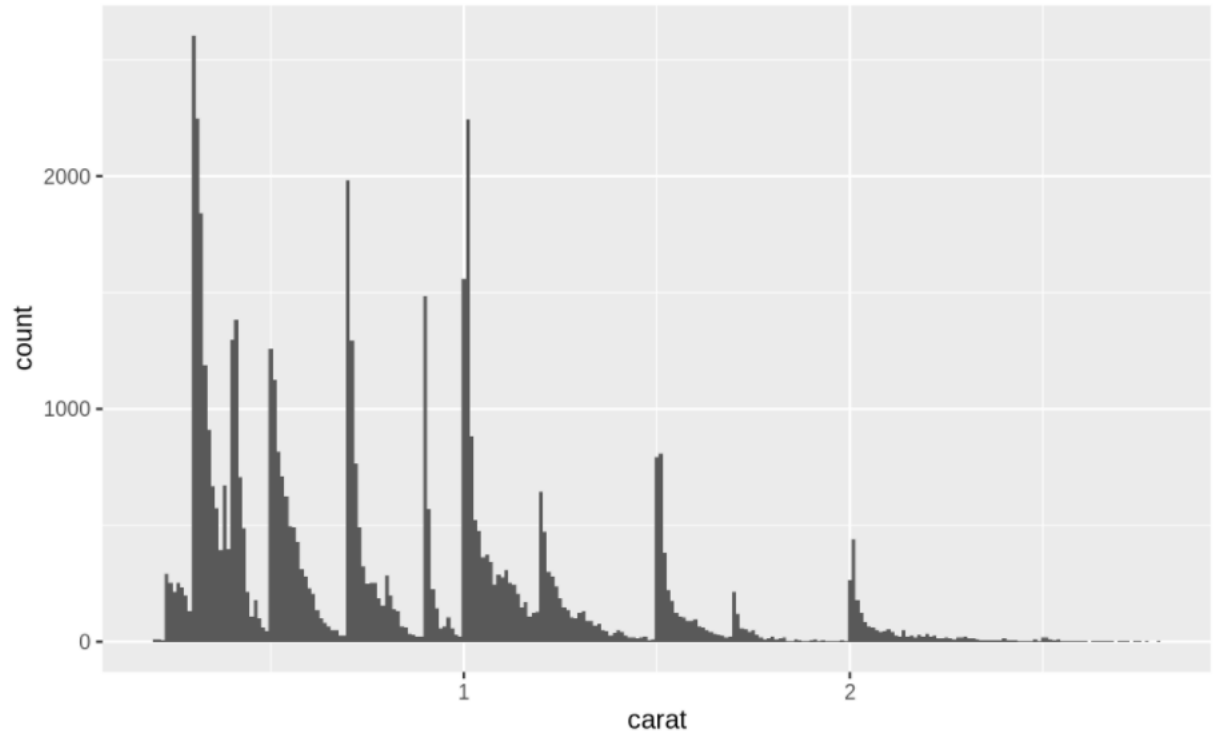
VISUALIZING DISTRIBUTIONS

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



VISUALIZING DISTRIBUTIONS

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



PHILOSOPHY AND STRATEGY OF EDA

Watch (after class) the following excerpt (~ 12mins) from a workshop on EDA given at the UN. Di Cook talks about EDA with respect to an OECD data set on education.

LINK

What strategies does she suggest for Exploratory Data Analysis?

PHILOSOPHY AND STRATEGY OF EDA

Di suggests two key strategies:

1. **Write down your expectations ahead of time** This gives you a starting point for things to look at. Try to verify your expectations of the data, but be prepared to be surprised.

PHILOSOPHY AND STRATEGY OF EDA

2. **Show the data** Don't over-process the data. Start with the rawest data possible, then refine it according to what you see (either to refine a question, or make a clearer display).

PHILOSOPHY AND STRATEGY OF EDA

3. **Note what surprises you** You can sometimes get pretty involved in an analysis and forgot how you got where you did. It's important to make notes along the way.

MEET WITH YOUR GROUP

MILESTONE #3

Due 10/13 - Milestone #3: Exploratory Data Analysis Step #1

Write at least **5 well defined questions** that you want to explore from your approved dataset.

- Note what variables from the dataset you plan to use.
- There must be at least one question for a categorical variable, at least one question for a numeric variable, at least one question compares a numeric variable across groups (from a categorical variable) and at least one question for the relationship between two numeric variables.
- Write hypotheses for what you expect to find from your questions, respectively. Note that these hypotheses need not be scientific.