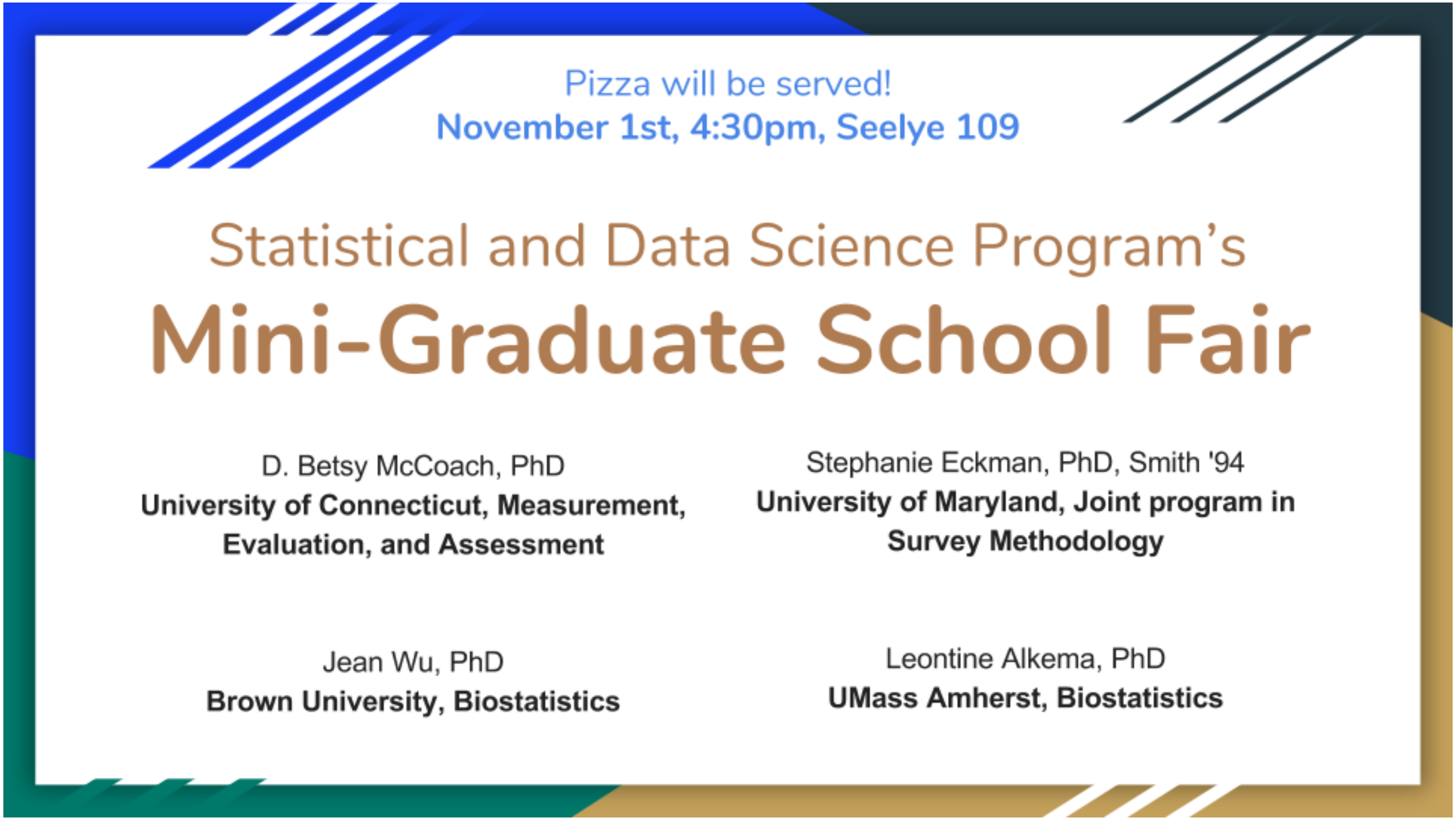
LECTURE 15:
FINAL PROJECT WORKSHOP PT. 1

November 1, 2017
SDS 293: Machine Learning

Announcements 1/3



Pizza will be served!

November 1st, 4:30pm, Seelye 109

Statistical and Data Science Program's

Mini-Graduate School Fair

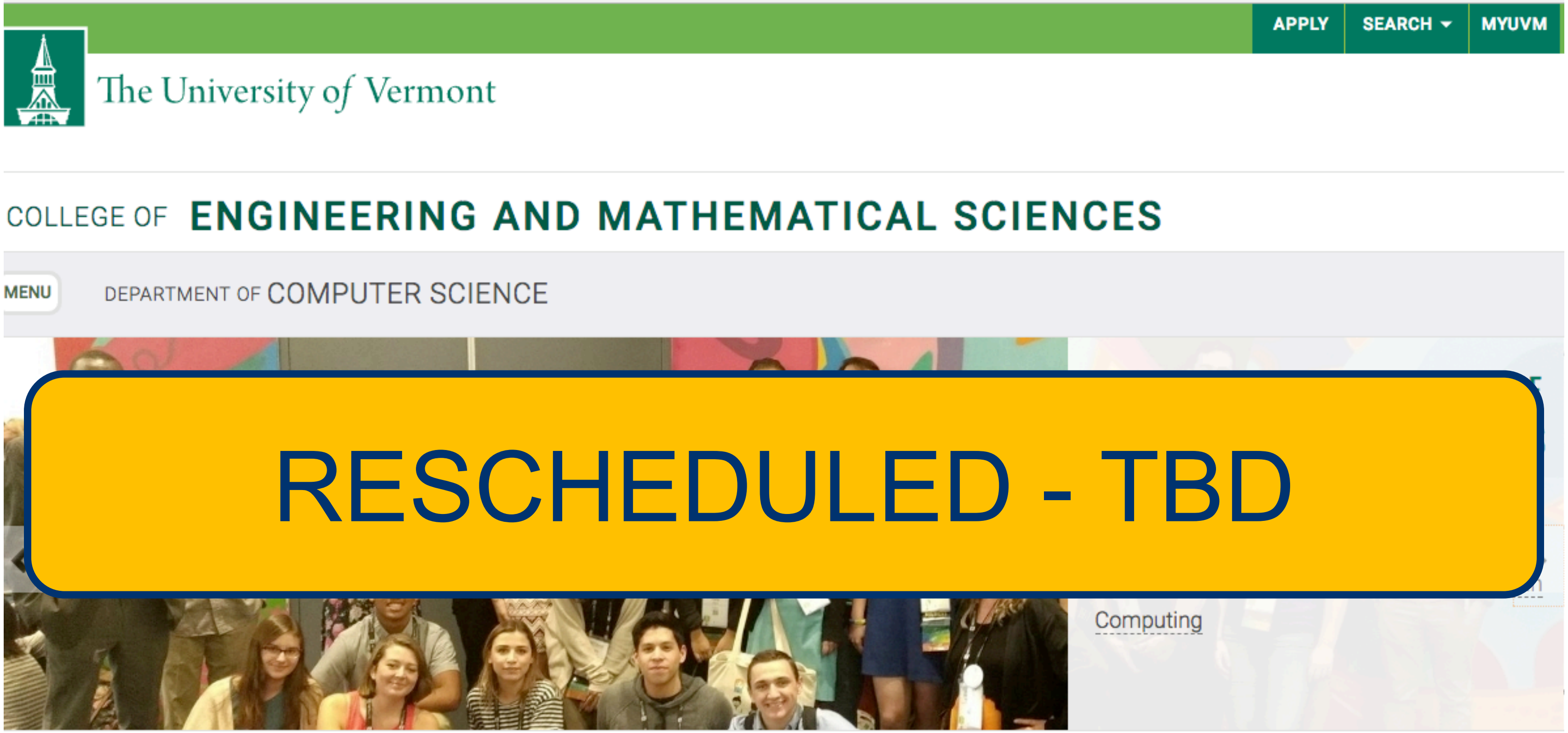
D. Betsy McCoach, PhD
University of Connecticut, Measurement, Evaluation, and Assessment

Stephanie Eckman, PhD, Smith '94
University of Maryland, Joint program in Survey Methodology

Jean Wu, PhD
Brown University, Biostatistics

Leontine Alkema, PhD
UMass Amherst, Biostatistics

Announcements 2/3


APPLY
SEARCH
MYUVM
The University of Vermont
COLLEGE OF ENGINEERING AND MATHEMATICAL SCIENCES
MENU
DEPARTMENT OF COMPUTER SCIENCE
Computing

RESCHEDULED - TBD

UVM CS Graduate Program Recruiting Event
Friday November 3rd at noon (location TBD)
RSVP to Jordan on Slack

Announcements 3/3

- Known grading bug in A3 – will be fixed by tomorrow

Outline

- Final Project Overview
  - Big picture (recap)
  - Some possible datasets
  - Deliverables timeline
- Topic Brainstorm Activity
- EDA

Final project (recap)

- **Goal:** apply the ML techniques we’ve learned to solve a real-world problem you care about

Final deliverable: a poster (or interactive visualization) that will be demonstrated during our end-of-semester reception and a 2-page write up of the methods you used

- Example problems

yelp dataset challenge

The Yelp dataset is a subset of our businesses, reviews, and user data for use in personal, educational, and academic purposes. Available in both JSON and SQL files, use it to teach students about databases, to learn NLP, or for sample production data while you learn how to make mobile apps.

4,700,000 reviews

156,000 businesses

200,000 pictures

12 metropolitan areas

1,000,000 tips by 1,100,000 users
Over 1.2 million business attributes like hours, parking, availability, and ambience
Aggregated check-ins over time for each of the 156,000 businesses

yelp dataset challenge: example ?s

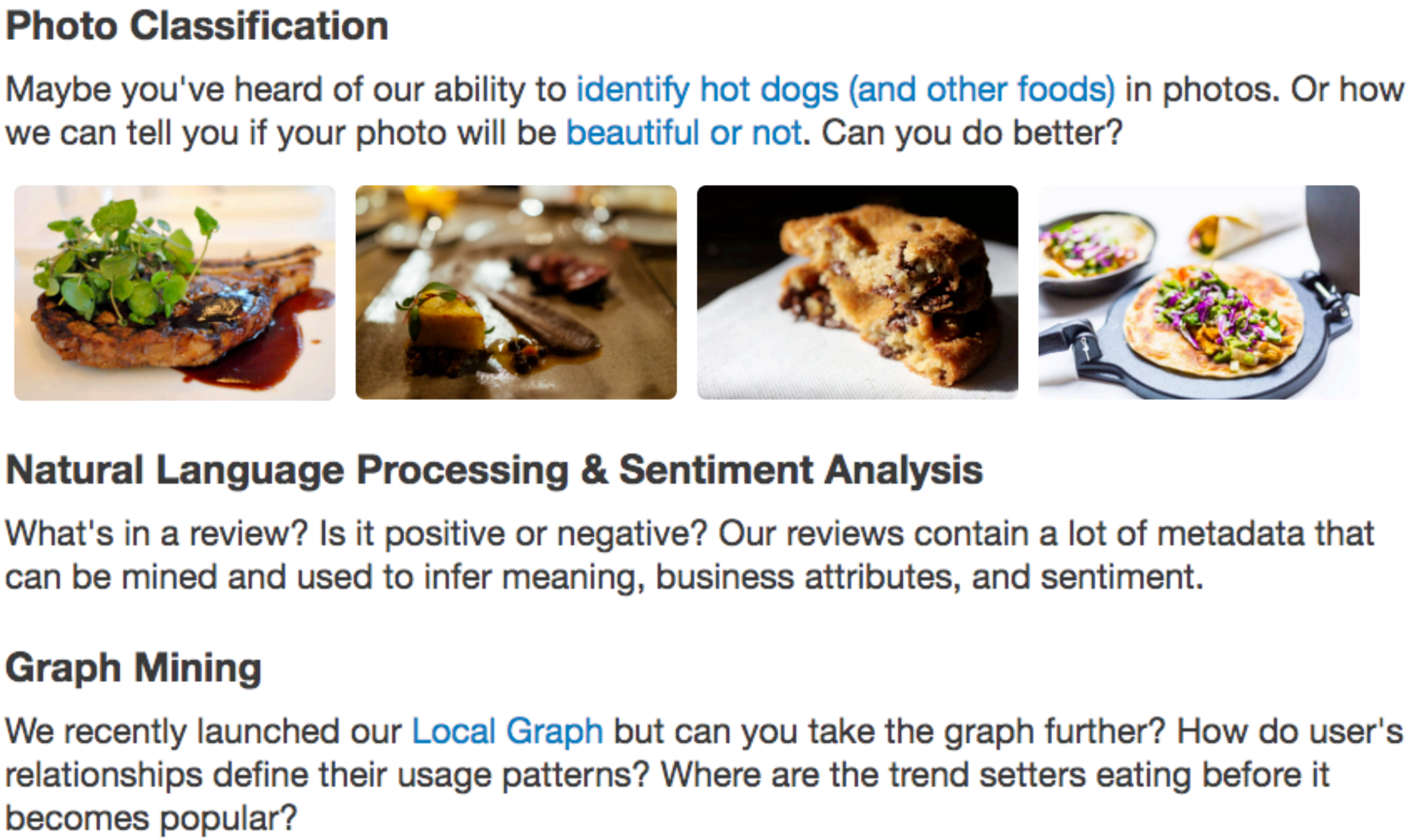


Photo Classification

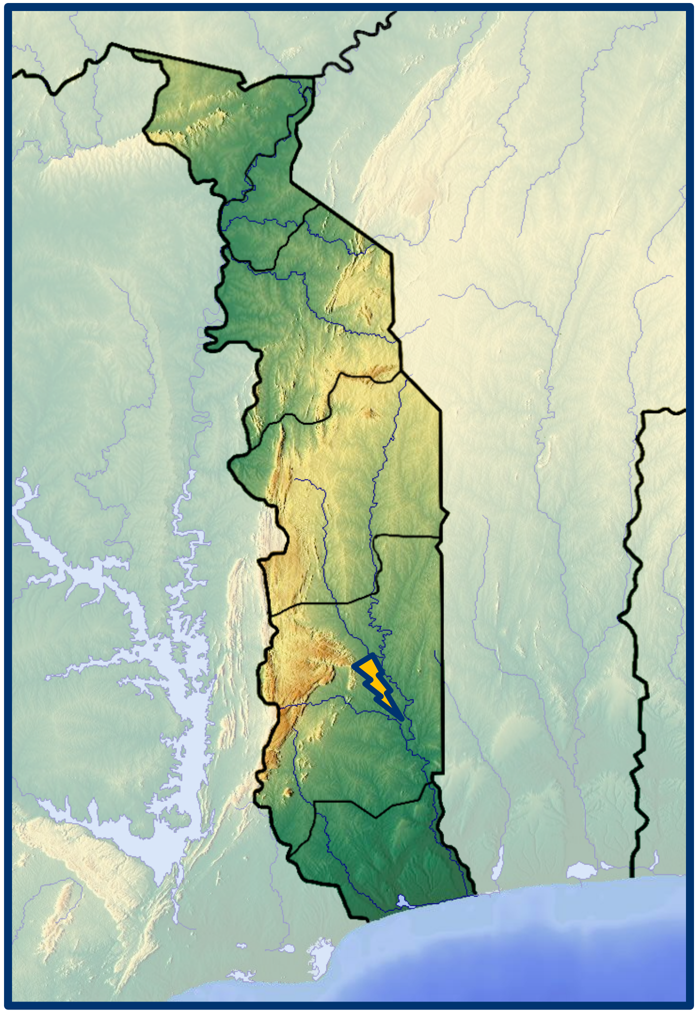
Maybe you've heard of our ability to identify hot dogs (and other foods) in photos. Or how we can tell you if your photo will be beautiful or not. Can you do better?

Natural Language Processing & Sentiment Analysis

What's in a review? Is it positive or negative? Our reviews contain a lot of metadata that can be mined and used to infer meaning, business attributes, and sentiment.

Graph Mining

We recently launched our Local Graph but can you take the graph further? How do user's relationships define their usage patterns? Where are the trend setters eating before it becomes popular?



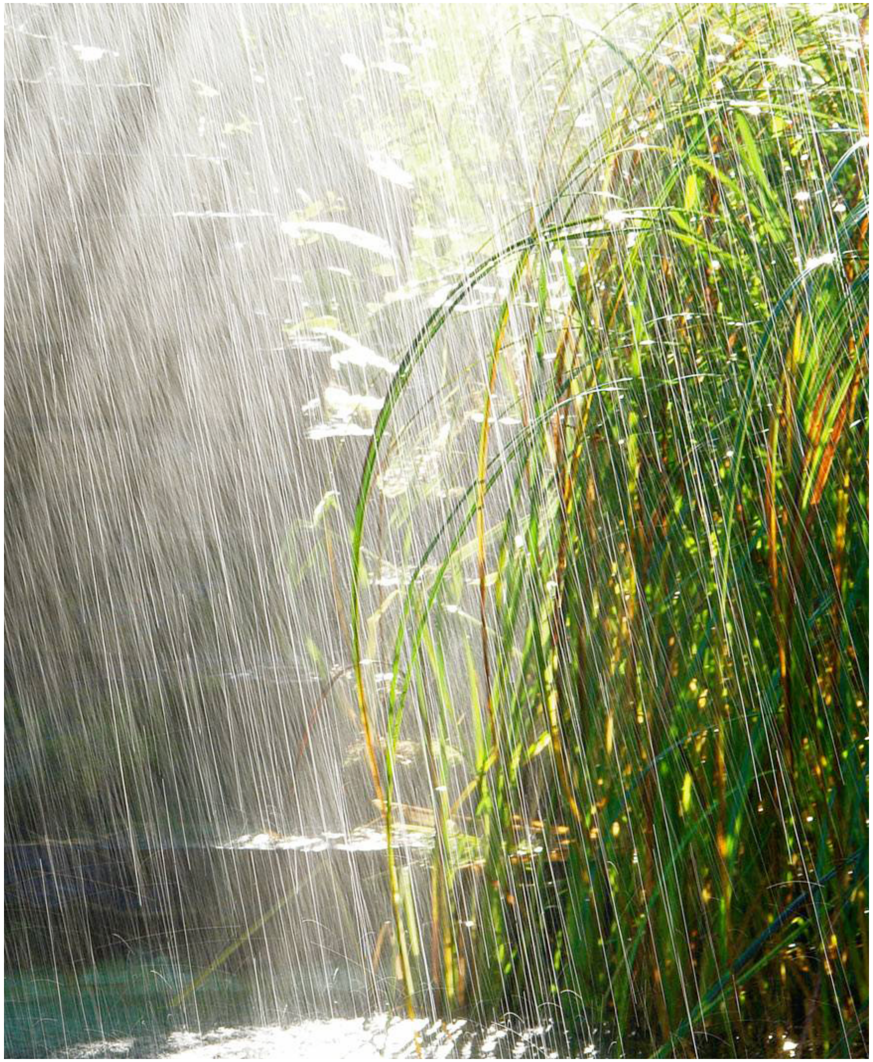
Nangbéto Hydropower Dam in Togo

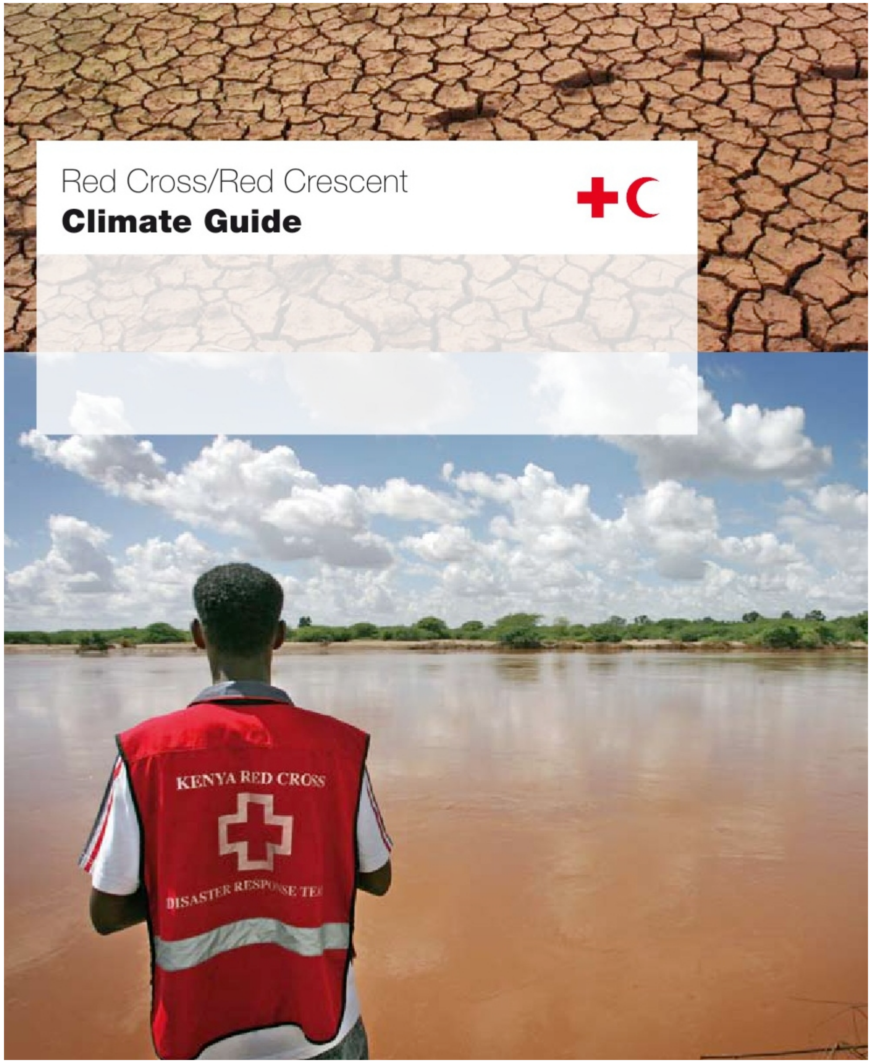
- Embankment dam on the Mono River
- Located in the Plateaux Region (2nd largest population in Togo)
- Supplies power, and as a secondary goal fish and water reserves

RED CROSS/RED CRESCENT
CLIMATE CENTRE

International Federation
of Red Cross and Red Crescent Societies

The Netherlands Red Cross





Red Cross/Red Crescent
Climate Guide

CREDIT SUISSE

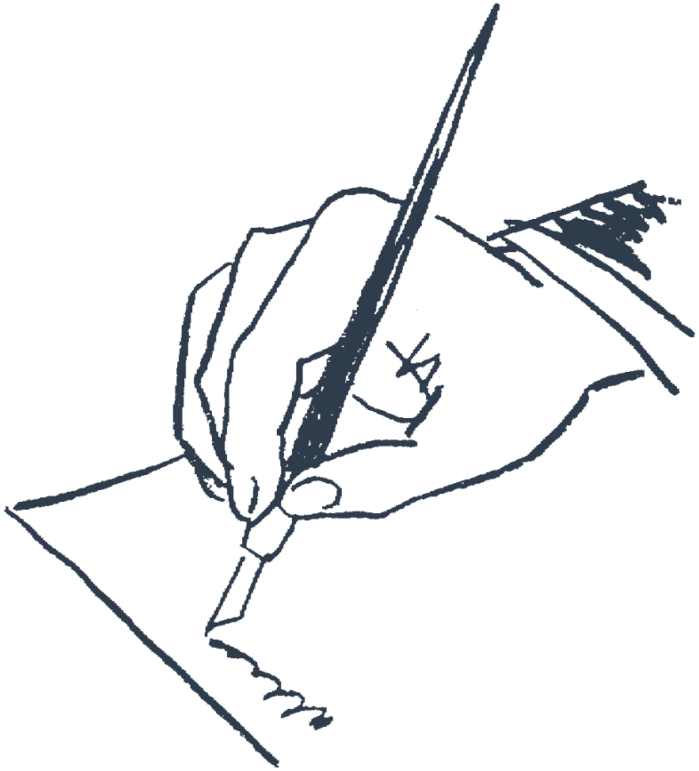
Past Projects

- Predicting Student Debt Upon College Graduation
- NYC 311 - What are some of the factors that determine the number of days spent to solve non-emergency issues in New York City?
- “What Are Your Odds?”: An Interactive Web Application to Visualize Health Outcomes
- Predicting Airbnb Listing Prices in the NYC Area
- …and many more!

Final Project Deliverables

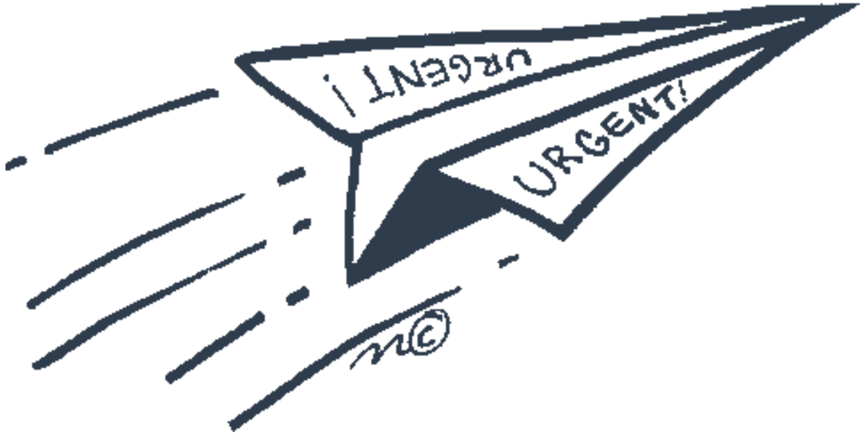
- Nov. 8th - FP1: Data Appendix
- Nov. 27th – FP2: Initial Model
- Dec. 4th – FP3: Revised Model
- Dec 13th – FP4: Poster (Final Project Reception)
- Dec. 22nd - FP5: Final Write-Up

Activity: topic brainstorming



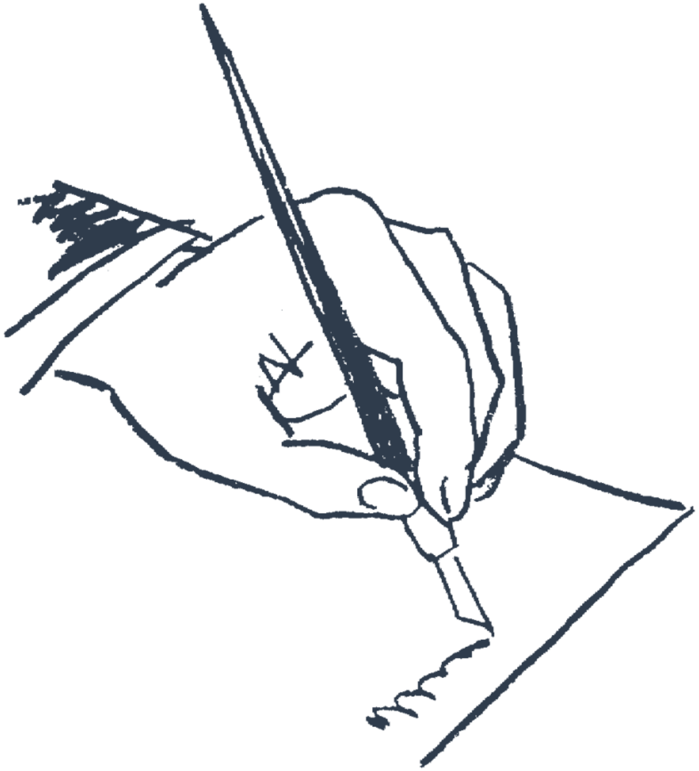
Step 1: Write a quick description of a data set you think would be interesting to explore at the top of the page, and write your 99 number at the bottom

Activity: real world problems



Step 2: Pass your description clockwise to the next person

Activity: real world problems



Step 3: Read the description of the dataset, and underneath the description, write a question you think someone might want to answer using it

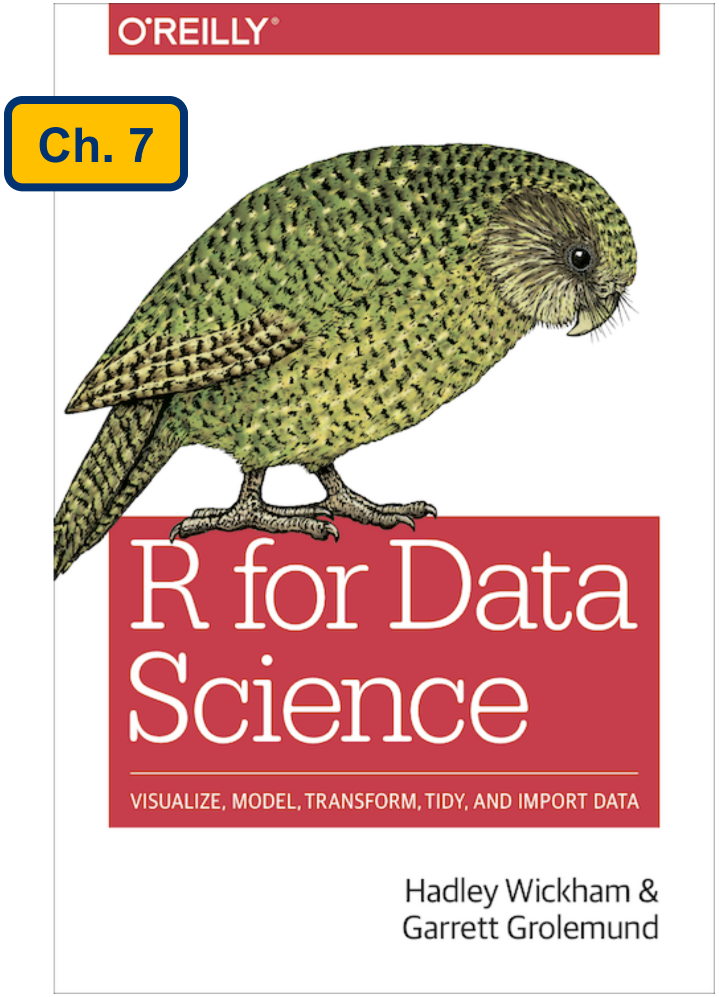
Activity: real world problems

Step 4: Fold over the top of the paper (leaving just your question visible), and pass it on. Now repeat!

To get credit for this activity

- **By the end of class**: write a quick Slack post about a potential final project topic
- Please include:
  - A description of the domain / dataset
  - The problem(s) you're trying to solve / question(s) you're trying to answer
  - The audience (who would care about your results?)
  - Where to find the data (if you know)
- Not 100% sure? Try a couple and get some feedback
- See a topic you like? Reply to the post and form a team!

Exploratory data analysis (EDA)



Ch. 7

http://r4ds.had.co.nz/

Our usual goal:

model some phenomenon using a dataset

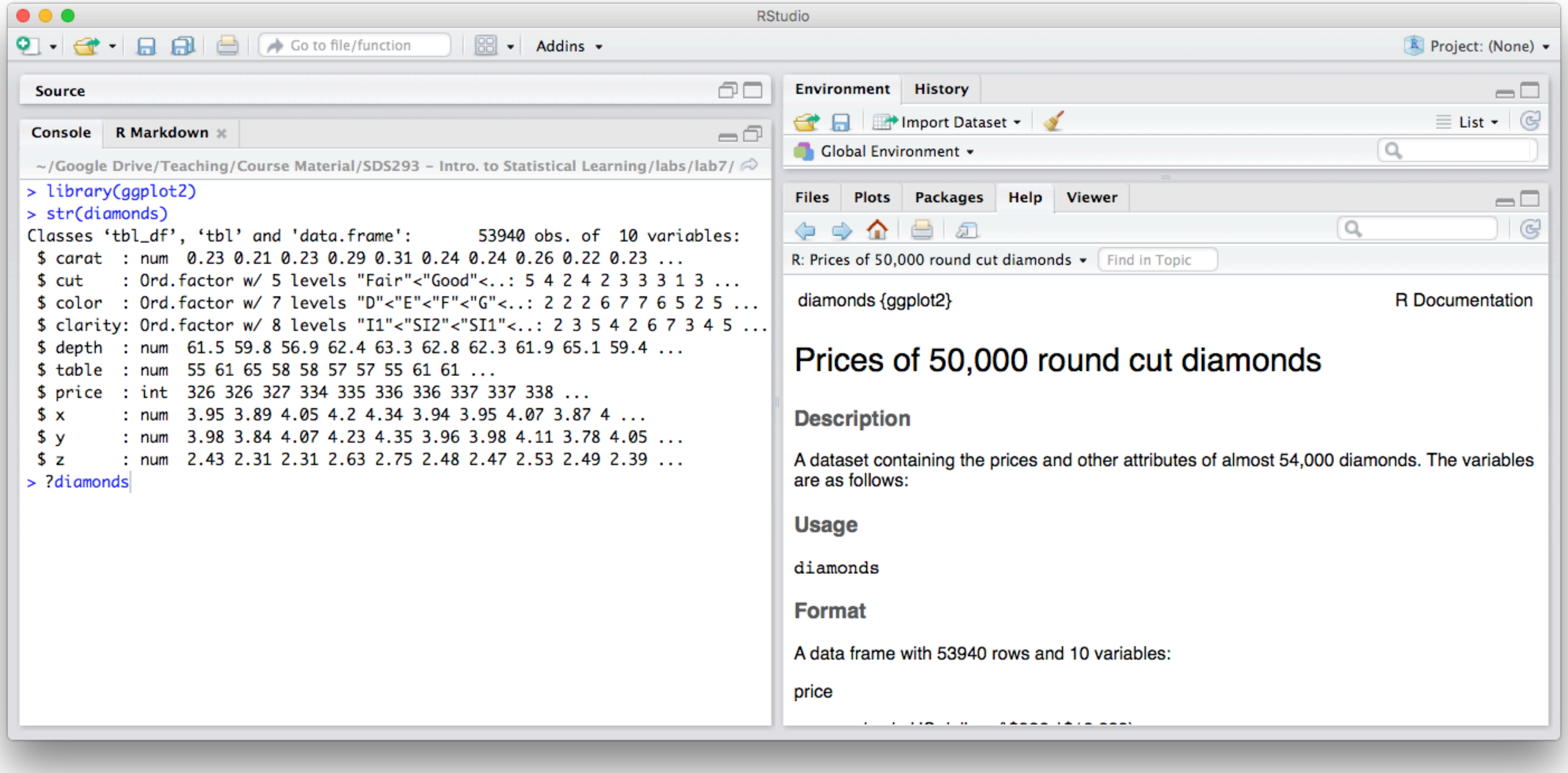
Goal of EDA:

develop an understanding of a dataset

Quick EDA walkthrough: `diamonds`

Useful starting questions

- **Q1:** what's in my data?
- **How to find out**: `?, str()`



~/Google Drive/Teaching/Course Material/SDS293 – Intro. to Statistical Learning/labs/lab7/

```
> library(ggplot2)
> str(diamonds)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':	53940 obs. of  10 variables:
 $ carat  : num  0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
 $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
 $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
 $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
 $ depth  : num  61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
 $ table  : num  55 61 65 58 58 57 57 55 61 61 ...
 $ price  : int  326 326 327 334 335 336 336 337 337 338 ...
 $ x      : num  3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
 $ y      : num  3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
 $ z      : num  2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
> ?diamonds
```

diamonds {ggplot2} R Documentation

Prices of 50,000 round cut diamonds

Description

A dataset containing the prices and other attributes of almost 54,000 diamonds. The variables are as follows:

Usage

```
diamonds
```

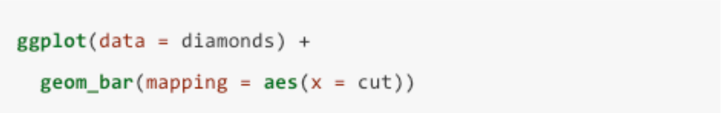
Format

A data frame with 53940 rows and 10 variables:

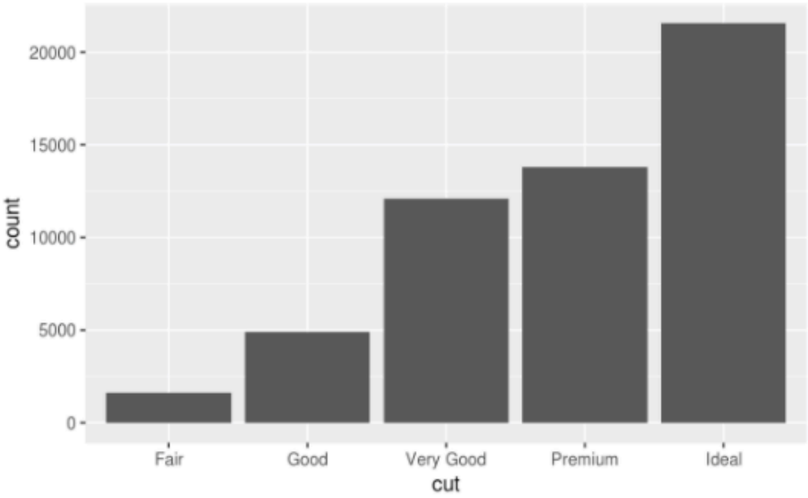
price

Useful starting questions

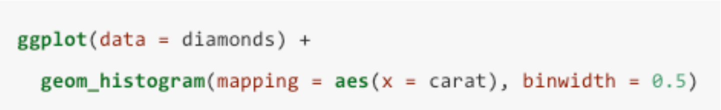
- **Q2**: what type of **variation** occurs within my variables?
- **How to find out**: visualize the distribution



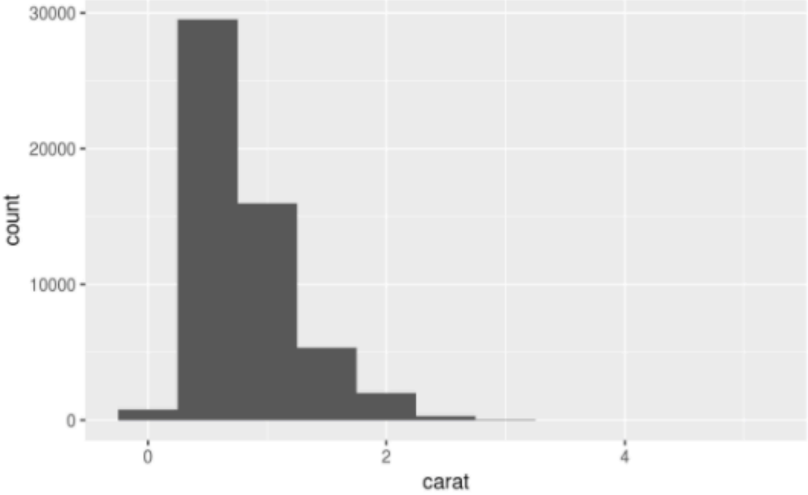
```
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut))
```



Categorical

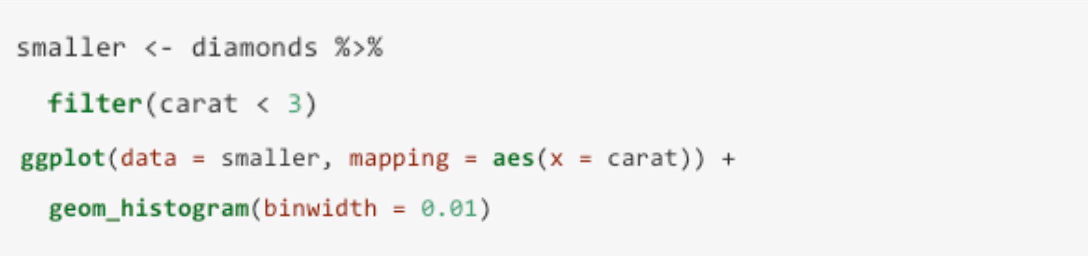


```
ggplot(data = diamonds) +
  geom_histogram(mapping = aes(x = carat), binwidth = 0.5)
```

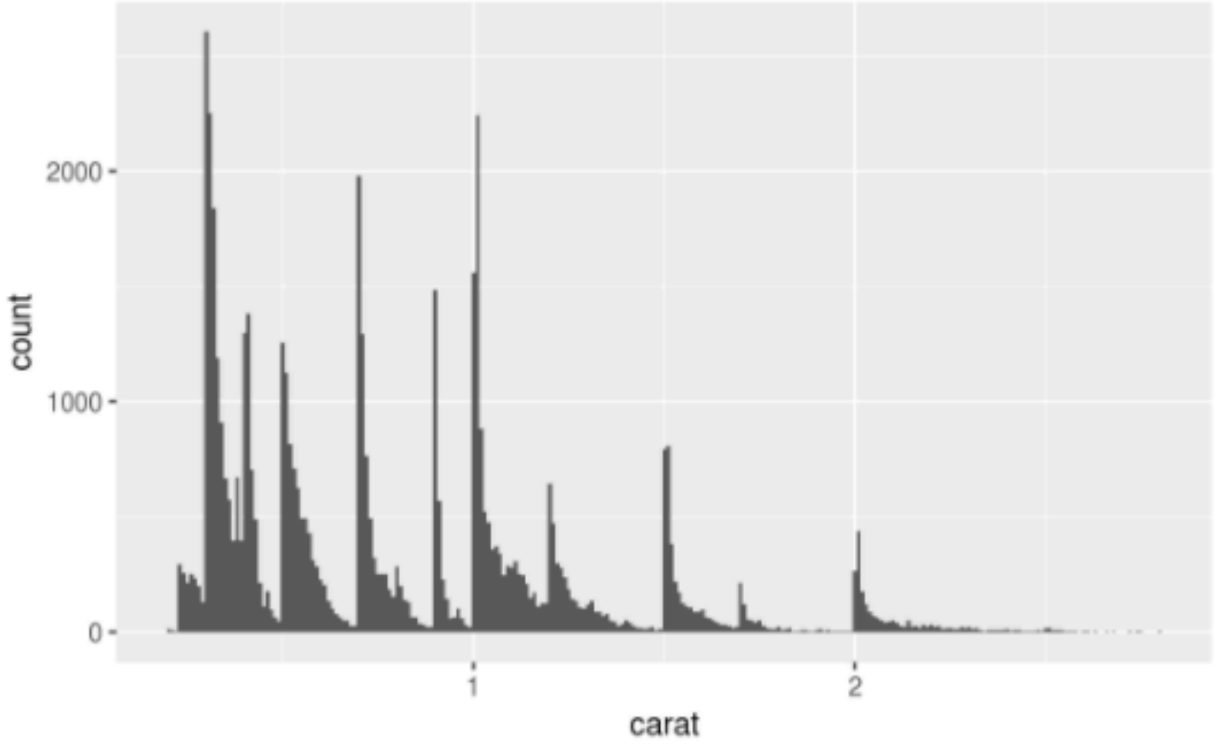


Continuous

#protip 1: adjust bin size for more detail

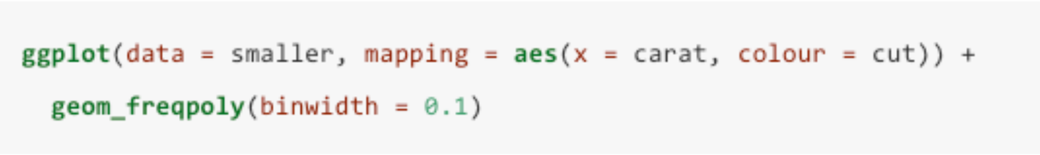


```
smaller <- diamonds %>%
  filter(carat < 3)
ggplot(data = smaller, mapping = aes(x = carat)) +
  geom_histogram(binwidth = 0.01)
```

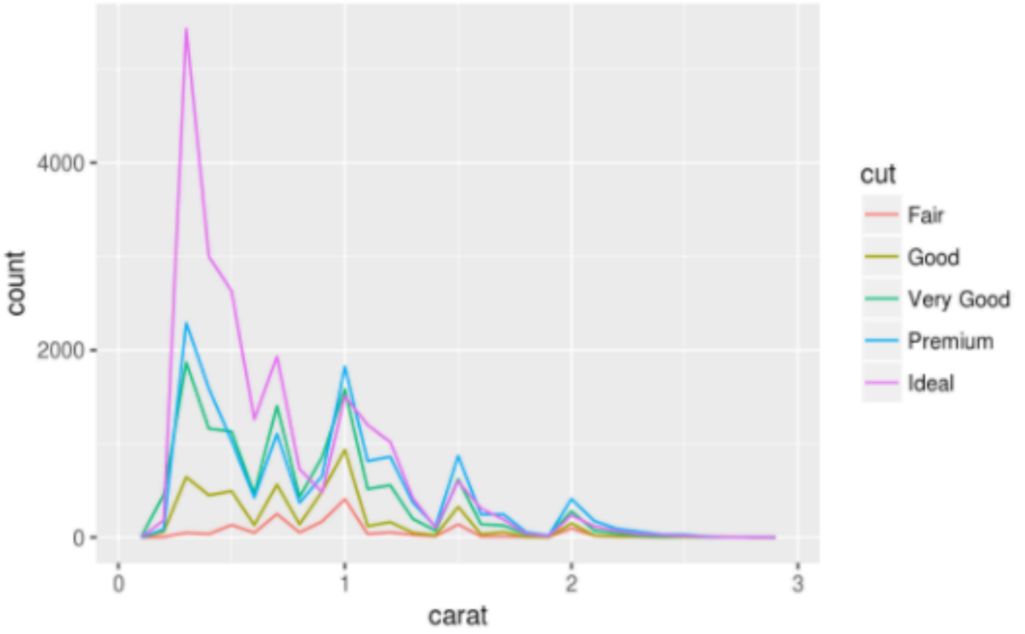


#protip 2: visualizing multiple distributions

- Use `geom_freqpoly()` instead of `geom_histogram()`

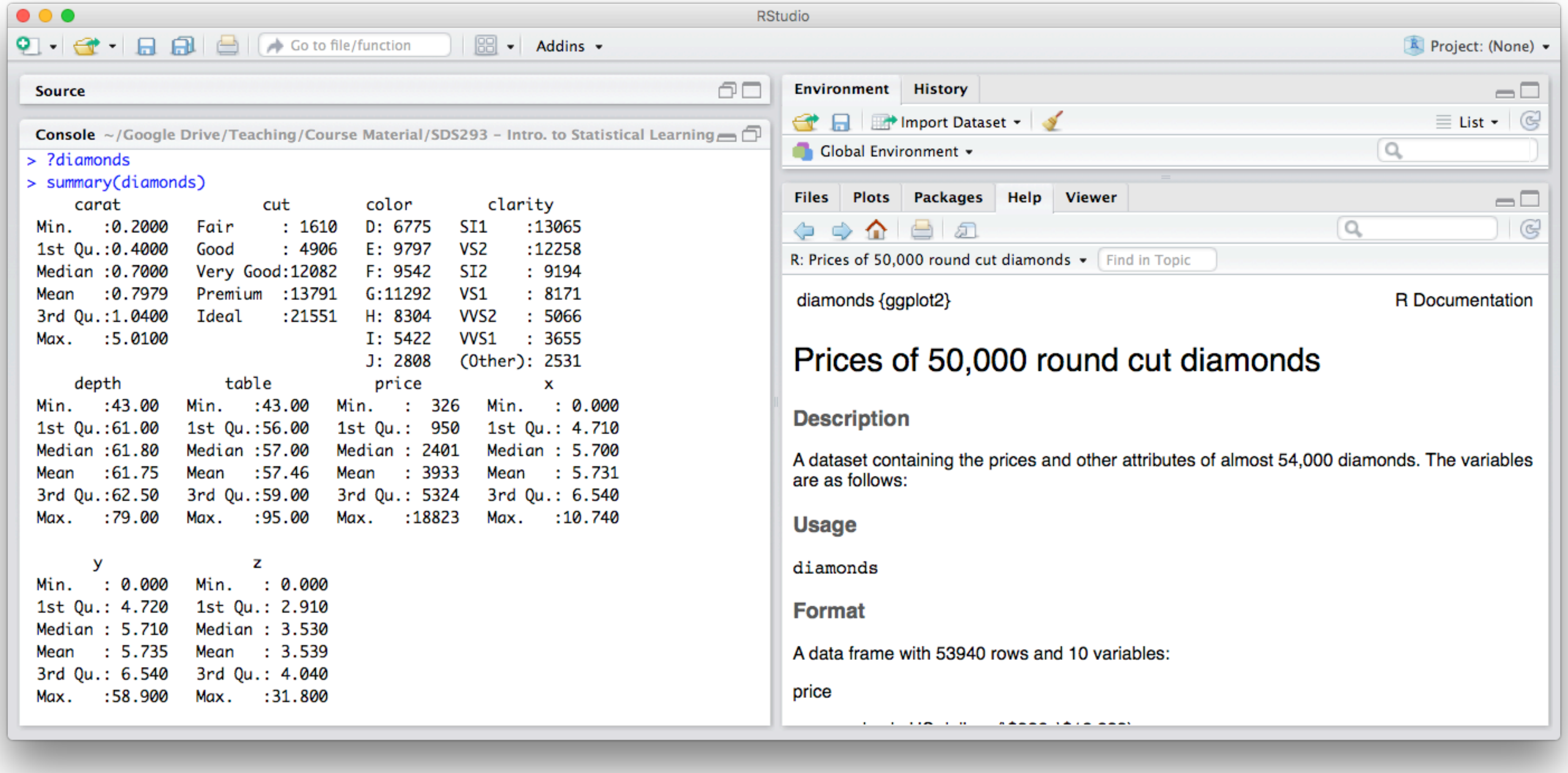


```
ggplot(data = smaller, mapping = aes(x = carat, colour = cut)) +
  geom_freqpoly(binwidth = 0.1)
```



Useful starting questions

- **Q3**: am I missing any data?
- **How to find out**: `summary()`



```
> ?diamonds
> summary(diamonds)
     carat               cut        color       clarity     
 Min.   :0.2000   Fair     : 1610   D: 6775   SI1    :13065  
 1st Qu.:0.4000   Good     : 4906   E: 9797   VS2    :12258  
 Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194  
 Mean   :0.7979   Premium  :13791   G:11292   VS1    : 8171  
 3rd Qu.:1.0400   Ideal    :21551   H: 8304   VVS2   : 5066  
 Max.   :5.0100                     I: 5422   VVS1   : 3655  
                                    J: 2808   (Other): 2531  
     depth           table           price             x         
 Min.   :43.00   Min.   :43.00   Min.   :  326   Min.   : 0.000  
 1st Qu.:61.00   1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710  
 Median :61.80   Median :57.00   Median : 2401   Median : 5.700  
 Mean   :61.75   Mean   :57.46   Mean   : 3933   Mean   : 5.731  
 3rd Qu.:62.50   3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540  
 Max.   :79.00   Max.   :95.00   Max.   :18823   Max.   :10.740  

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

diamonds {ggplot2} R Documentation

Prices of 50,000 round cut diamonds

Description

A dataset containing the prices and other attributes of almost 54,000 diamonds. The variables are as follows:

Usage

```
diamonds
```

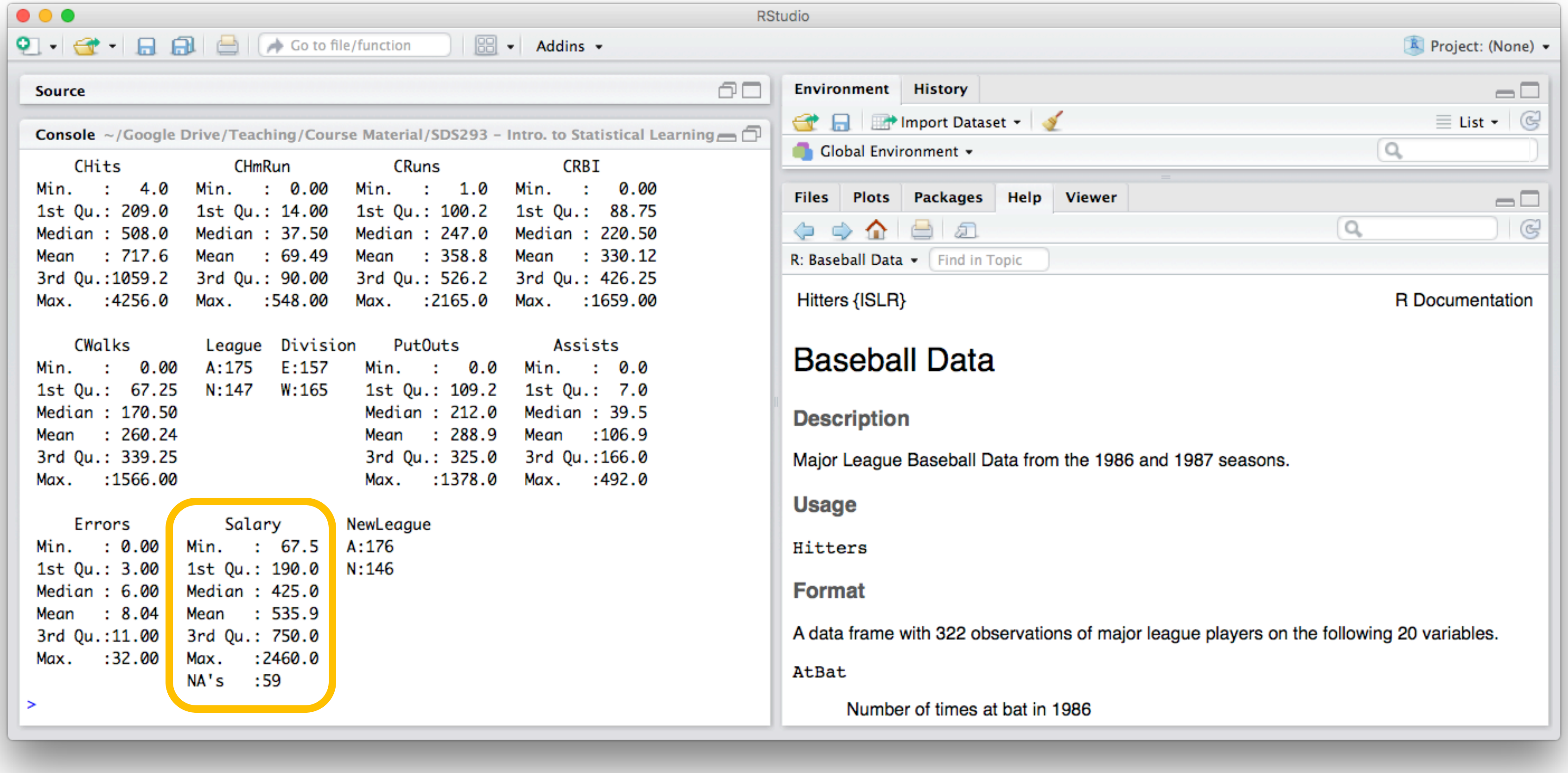
Format

A data frame with 53940 rows and 10 variables:

price

Useful starting questions

- **Q3**: am I missing any data?
- **How to find out**: `summary()`



```
    CHits            CHmRun          CRuns             CRBI
 Min.   :   4.0   Min.   :  0.00   Min.   :   1.0   Min.   :   0.00
 1st Qu.: 209.0   1st Qu.: 14.00   1st Qu.: 100.2   1st Qu.:  88.75
 Median : 508.0   Median : 37.50   Median : 247.0   Median : 220.50
 Mean   : 717.6   Mean   : 69.49   Mean   : 358.8   Mean   : 330.12
 3rd Qu.:1059.2   3rd Qu.: 90.00   3rd Qu.: 526.2   3rd Qu.: 426.25
 Max.   :4256.0   Max.   :548.00   Max.   :2165.0   Max.   :1659.00

    CWalks         League  Division    PutOuts          Assists
 Min.   :   0.00   A:175   E:157    Min.   :   0.0   Min.   :  0.0
 1st Qu.:  67.25   N:147   W:165    1st Qu.: 109.2   1st Qu.:  7.0
 Median : 170.50                    Median : 212.0   Median : 39.5
 Mean   : 260.24                    Mean   : 288.9   Mean   :106.9
 3rd Qu.: 339.25                    3rd Qu.: 325.0   3rd Qu.:166.0
 Max.   :1566.00                    Max.   :1378.0   Max.   :492.0

    Errors          Salary        NewLeague
 Min.   : 0.00   Min.   :  67.5   A:176
 1st Qu.: 3.00   1st Qu.: 190.0   N:146
 Median : 6.00   Median : 425.0
 Mean   : 8.04   Mean   : 535.9
 3rd Qu.:11.00   3rd Qu.: 750.0
 Max.   :32.00   Max.   :2460.0
                 NA's   :59
>
```

Hitters {ISLR} R Documentation

Baseball Data

Description

Major League Baseball Data from the 1986 and 1987 seasons.

Usage

`Hitters`

Format

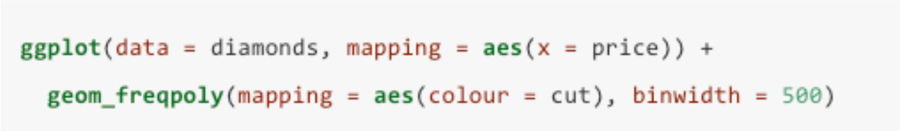
A data frame with 322 observations of major league players on the following 20 variables.

`AtBat`

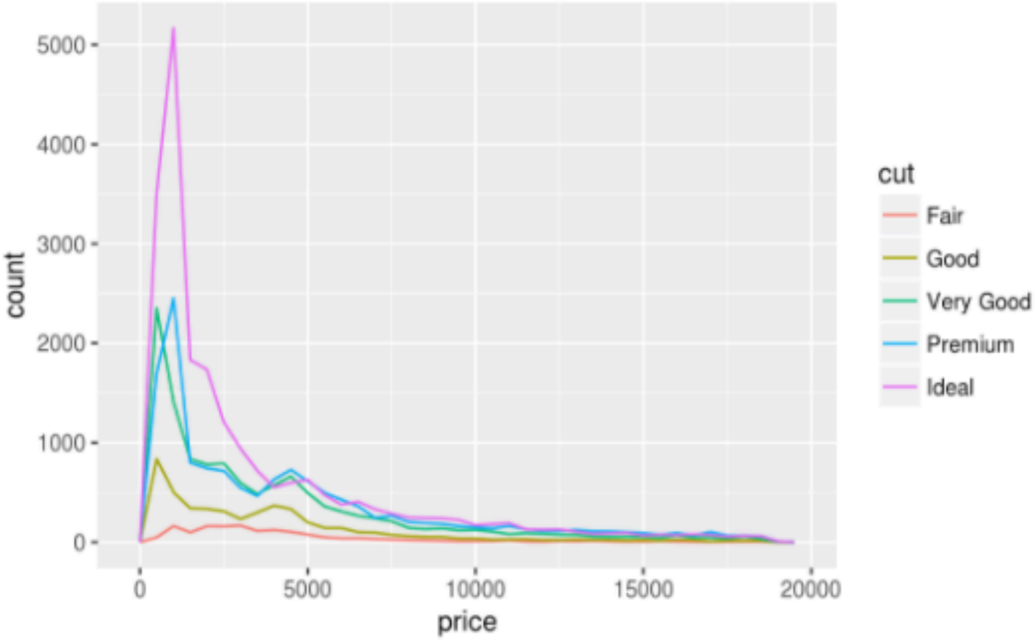
Number of times at bat in 1986

Useful starting questions

- **Q4**: what type of **covariation** occurs within my variables?
- **How to find out**: visualize the (relative) distribution

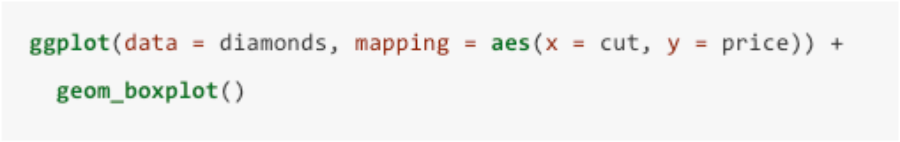


```
ggplot(data = diamonds, mapping = aes(x = price)) +
  geom_freqpoly(mapping = aes(colour = cut), binwidth = 500)
```

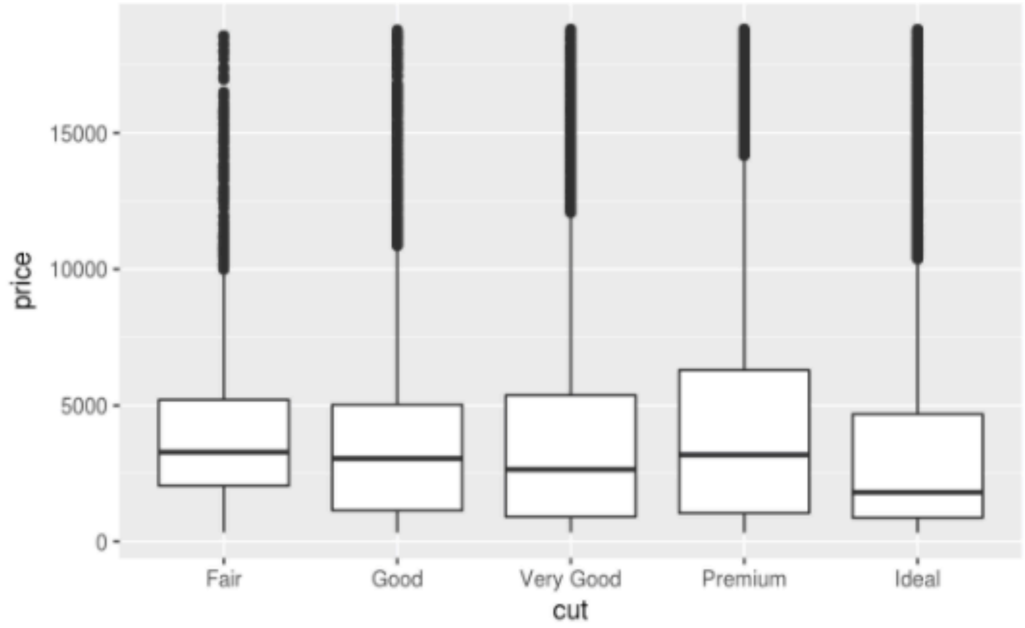


Useful starting questions

- **Q4**: what type of **covariation** occurs within my variables?
- **How to find out**: visualize the (relative) distribution



```
ggplot(data = diamonds, mapping = aes(x = cut, y = price)) +
  geom_boxplot()
```



Your turn!

- Find some data and perform EDA
- **By the end of class**: write a quick Slack post about a potential final project topic, including:
  - A description of the domain / dataset
  - The problem(s) you're trying to solve / question(s) you're trying to answer
  - The audience (who would care about your results?)
  - Where to find the data (if you know)
- See a topic you like? Reply to the post and form a team!

Coming up

- A6 due tonight
- FP1 released this afternoon