STOR 320 Introduction to Data Science

Lecture 1

Yao Li

Department of Statistics and Operations Research

UNC Chapel Hill

Instructor

- Name: Yao Li
- Email: yaoli@email.unc.edu
- Office hours: Wednesday, Friday 9:00AM to 10:00AM
- Personal website: https://liyao880.github.io/yaoli/
- Course website: https://liyao880.github.io/stor320/
- Research interest: adversarial deep learning, large-scale recommender systems, model compression
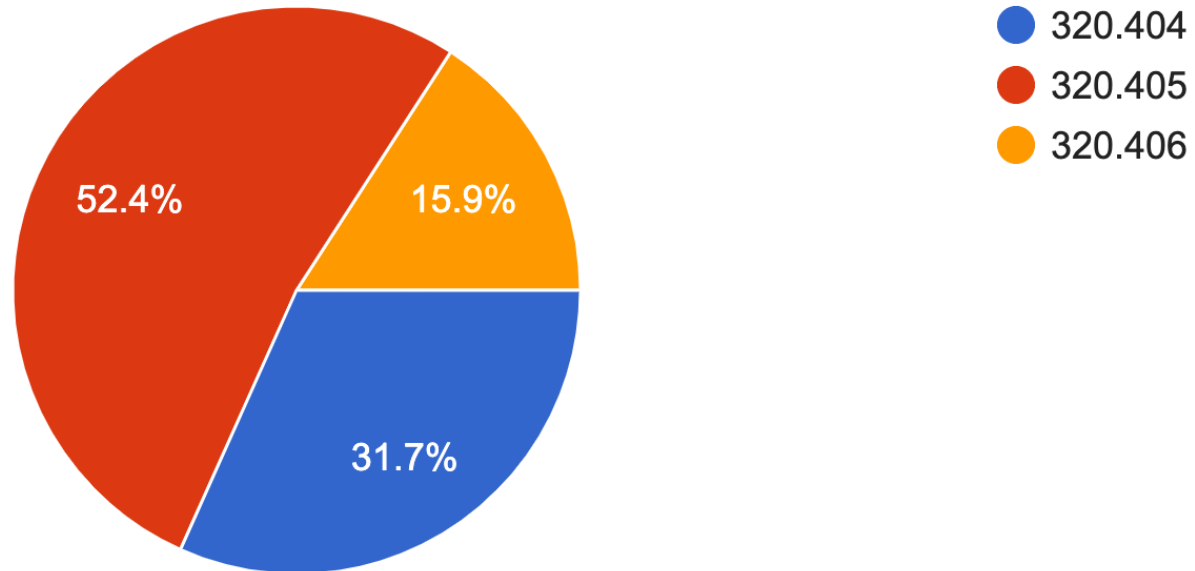
Get to know your instructor

- Join at **www.kahoot.it**

Survey Results

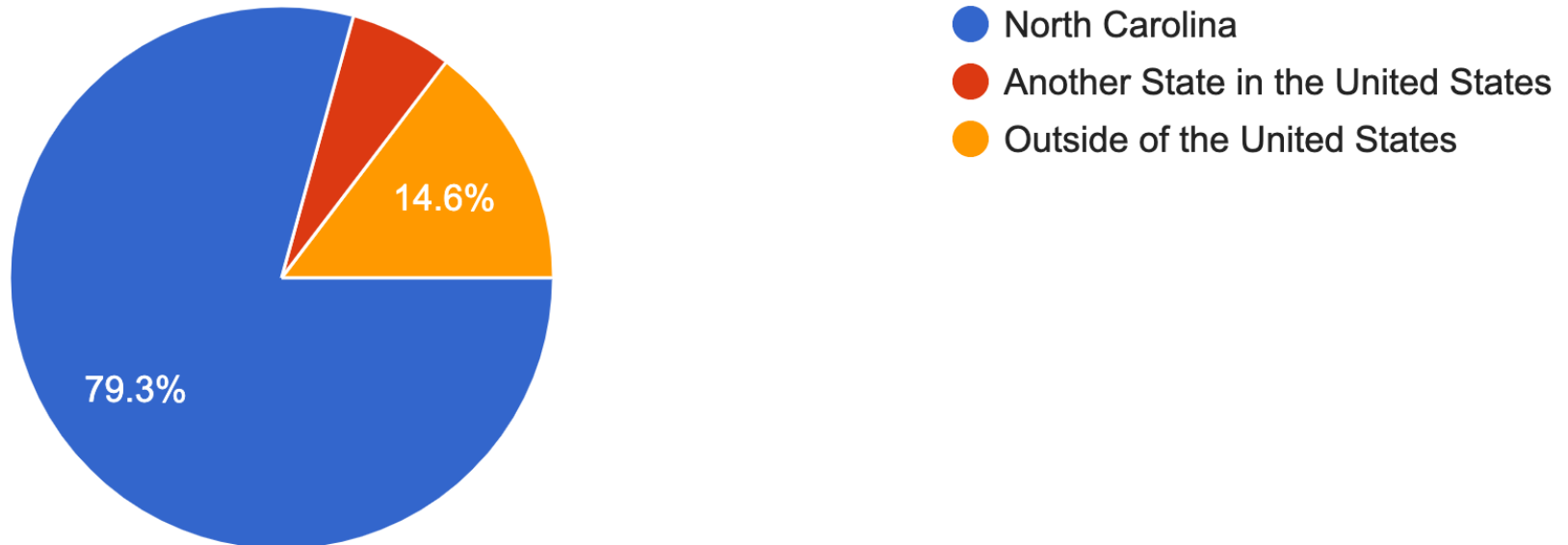In what lab section are you registered?

82 responses

- 320.404
- 320.405
- 320.406

52.4%

31.7%

15.9%

Survey Results

Where will you be for the majority of your time this spring semester?

82 responses

- North Carolina
- Another State in the United States
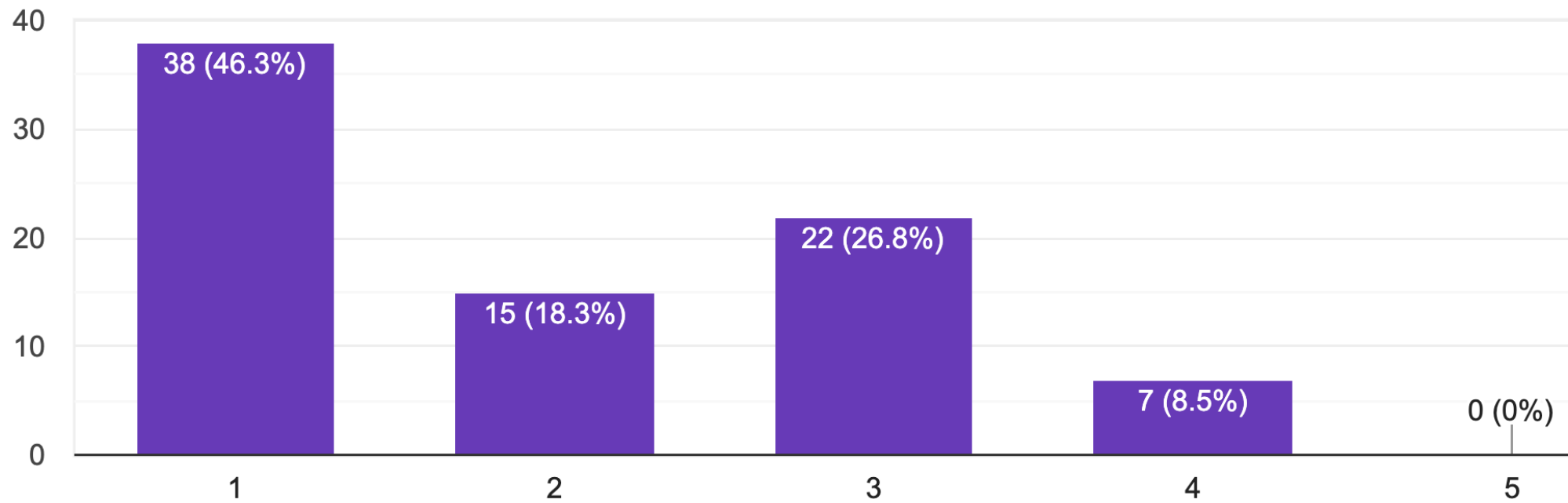- Outside of the United States

14.6%

79.3%

Survey Results

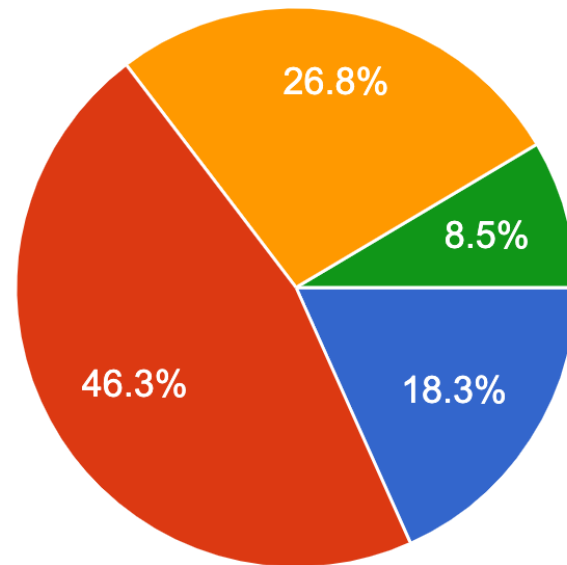What is your experience with R?

82 responses

Survey Results

What do you plan to pursue after graduation?

82 responses

- 🔵 Job Not in Field Related to Data Science
- 🔴 Job in Field Related to Data Science
- 🟠 Master's Degree
- 🟢 Doctoral Degree
- 🟣 Medical, Pharmaceutical, Dental, Nursing, or Veterinary Degree

Pie chart values: 26.8%, 8.5%, 46.3%, 18.3%

Instructional Assistant

- Pavlos Zoubouloglou (404)
  - Email: pavlos@live.unc.edu
  - Office Hours: Monday 10:00AM-11:00AM, Friday 1:00PM-2:00PM

- Taylor Petty (405)
  - Email: tmpetty@live.unc.edu
  - Office Hours: Monday 2:30PM-3:30PM, Wednesday 1:30PM-2:30PM

- Sam Booth (406)
  - Email: slbooth@live.unc.edu
  - Office Hours: Tuesday, Thursday 4:00PM-5:00PM

Lectures and Labs

- Lectures   MWF   8:00 AM – 8:50 AM

- Labs
  - 404 Tuesday 9:30AM – 10:20AM
  - 405 Tuesday 5:00PM – 5:50PM
  - 406 Thursday 9:30AM – 10:20AM

- Email Christine (crikeat@email.unc.edu)

Outline

- Administrative details

- What's the course about?

- Introduction to R

Ask Questions in Class

- By default, your microphone will be muted.

- If you have a question, feel free to unmute yourself and ask questions.

- Also, you can type your question in the in-meeting chat window.

Remote Instruction

- This will be a hybrid course:
  a) lectures will be held live online during the scheduled time and recorded so that you can watch them later;
  b) All lab sessions will be online but not recorded;
  c) some of the lectures might be prerecorded if there are connection issues and livestreaming is not possible;
  d) office hours will be held online but not recorded;
  e) all assignments will be done remotely.

Questions

- Three ways to ask questions:
  - ➢ post questions on Sakai forum;

  - ➢ come to the virtual office hours on Zoom;

  - ➢ send an email to the instructor or the IAs.

Grading

Lab Attendance	10%
Labs	15%
Homework	45%
Final Project	30%

A	94 to 100	B	83 to 86.99	C	73 to 76.99	D	60 to 66.99
A-	90 to 93.99	B-	80 to 82.99	C-	70 to 72.99	F	0 to 59.99
B+	87 to 89.99	C+	77 to 79.99	D+	67 to 69.99		

Homework and Labs

- Around 7 homework assignments and 4 data analysis assignments. They will be posted on Gradescope and there will be about one week to complete the homework and about two weeks to complete data analysis assignments.

- Lab assignment:
  - Due 30 minutes after the lab ends.
  - No late submission will be accepted.
  - will be based on the topics discussed in lecture or related to your final project.

Project

- For the final project, each section of STOR 320 will be divided into research groups of size 4 or 5. To ensure fairness, students will be assigned randomly based on lab session.

- The groups will be assigned by Feb 5, 2021 (Friday) and you can find your group on shared via [google sheet](#).
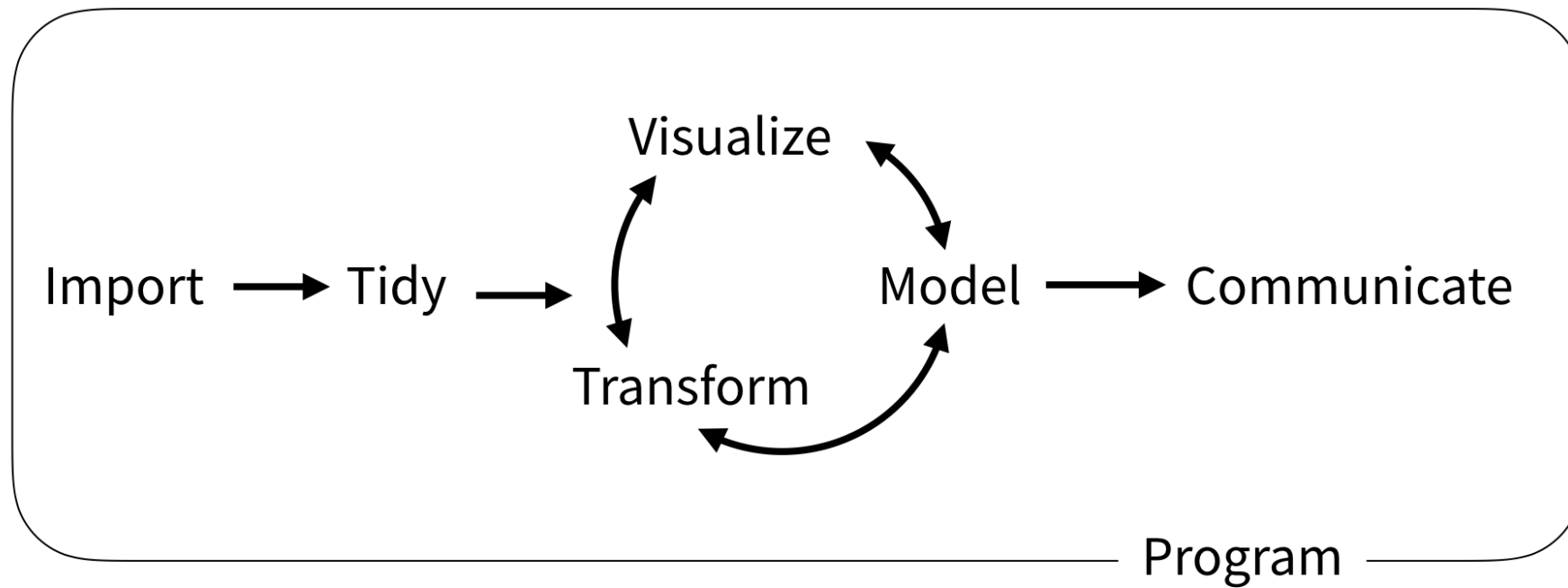
Project

Project proposal	10%
Exploratory data analysis	20%
Final report	40%
Final presentation	30%

Important dates

|	
Project proposal	February 17
Exploratory data analysis	March 26
Final report	May 5
Final Presentation	April 30, May 3, May 5

What is data science?

Wickham and Grolemund (2017)

The model of data science
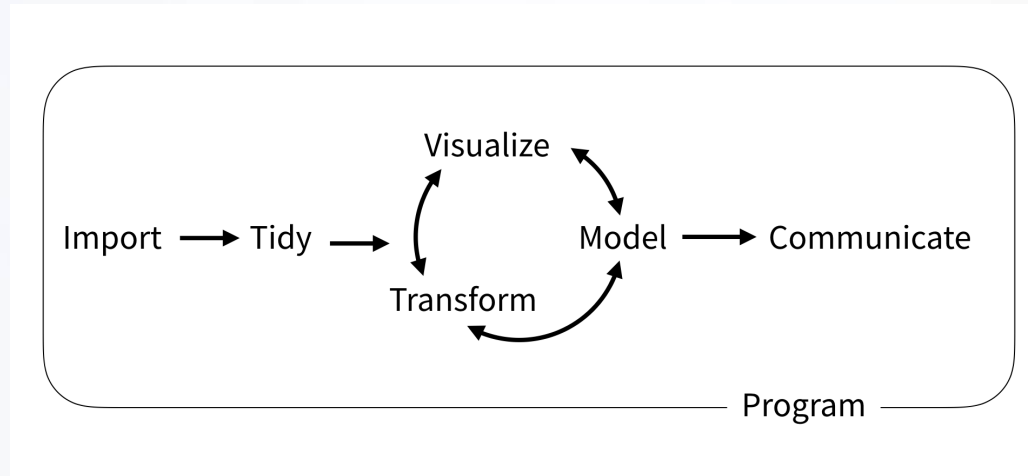
- First we must *import* our data.
- *Tidy* data → consistent structure
- Transformation:
  - narrowing in on observations of interest
  - creating new variables
  - calculating a set of summary statistics

The model of data science
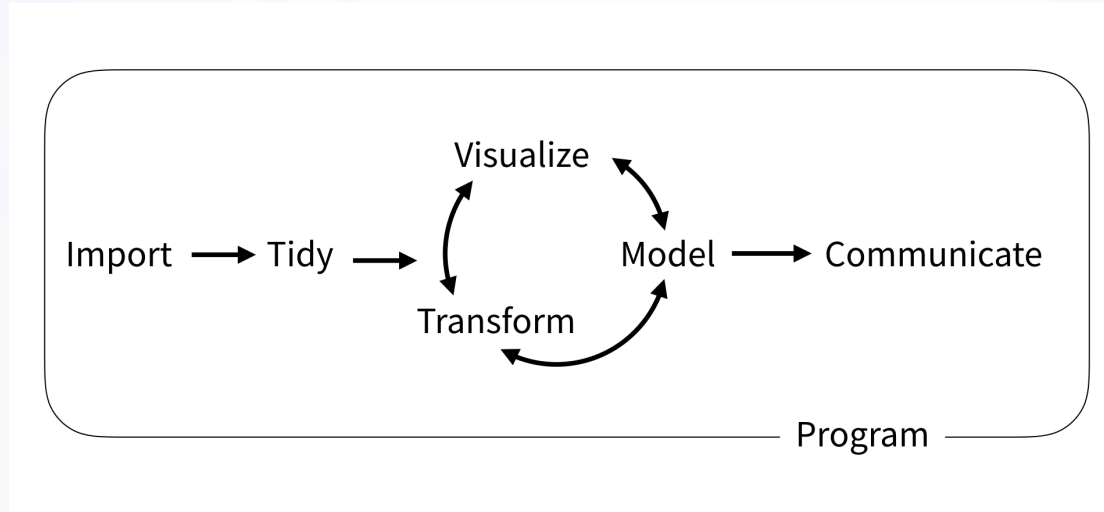
- *Visualization:* show you things that you did not expect or raise new questions about the data.

- Use a *model* to answer your questions

- *Communication:* an absolutely critical part of any data analysis project.

- Surrounding all these tools is programming.

R and RStudio

Why R?

- Easy to learn and easy to use.
- Very popular and one of the standard languages for statistics, data science, computational biology, finance, industry, etc.
- Free and open-source.
- A lot of high-quality packages.
- New technology and ideas often appear first in R.
- Supported by a vast community that maintains and updates R.
- Runs on basically any platform.

Learning Programming

- Transfer the concepts to other languages

- How you approach a computational task and reason about the computations is similar

- Learning another programming language will be much easier in the future

Statistical Learning

- Linear regression.

- Classification (logistic regression, LDA, K-nearest neighbors).

- Cross-validation and bootstrap.

- Principal component analysis.

- Clustering methods (K-means clustering and hierarchical clustering).

- Recommender systems.

- Neural networks.

Textbooks

- *R for Data Science.* Hadley Wickham. Legally free online, but can be purchased for less than $40 on Amazon. Additional suggested texts are provided on the website. All texts used in this course are free and downloadable from course website.

- *The elements of statistical learning: data mining, inference, and prediction*. Hastie, Trevor, Robert Tibshirani, and Jerome Friedman.