THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Introduction to predictive analytics

Lecture 1

STA 371G

Course goals

- Use regression and time series analysis to build predictive models
- Create decision tree models to systematize business decision-making
- Utilize simulations to forecast outputs based on uncertain inputs
- Given a new business situation, select an appropriate analysis, carry it out, and effectively communicate the results
- This is a practical course!

About the course staff

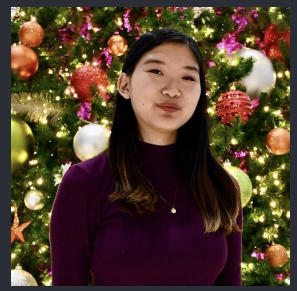
- Instructor: **Brian Lukoff, Ph.D.** (he/him/his)
  - Office hours: MW 9:30-10:30 AM
  - Course questions: `sta371g-lukoff@austin.utexas.edu`
  - Personal contact: `brian.lukoff@utexas.edu` or 415-652-8853
- TAs and office hours:

Holden Archer

M 3:30-6:30 PM

Michelle Chahda

Th 3:30-6:30 PM



Meryl Xiong

MW 11-12:30 PM

Leyang Xu

F 2:00-5:00 PM

Who am I?

- **Educator:** Teaching statistics at McCombs since 2014; previously taught at Harvard University and Boston University
- **Entrepreneur:** Currently co-founder and CEO of Perusall; formerly co-founder and CEO of Learning Catalytics (acquired by Pearson)
- **Engineer/statistician:** Software engineering/analytics background

1. Find someone who...

2. Course logistics

3. Let's do some statistics, yo

Your group gets points if someone in your breakout room matches the characteristic below. The goal is for your group to get the most points. Each member of your group can only be used once!

The winning group will be crowned the STA 371G Find Someone Who Champion™ (SFSWC™).

1 pt Has been to a UT football game

1 pt Has been to Torchy's Tacos

1 pt Plays a musical instrument

1 pt Doesn't like asparagus

1 pt Has ridden a bike

1 pt Has worn earrings

1 pt Is an only child

2 pts Has been to a honky tonk

2 pts Has met a celebrity

2 pts Was born outside of Texas

2 pts Was born outside of the US

2 pts Has a tattoo

2 pts Can do a Rubik's cube

3 pts Watches *Jersey Shore: Family Vacation*

3 pts Can wiggle their ears

3 pts Thinks statistics is awesome (really!)

3 pts Voted in a national election (US or non-US) in 2020

1. Find someone who…

2. Course logistics

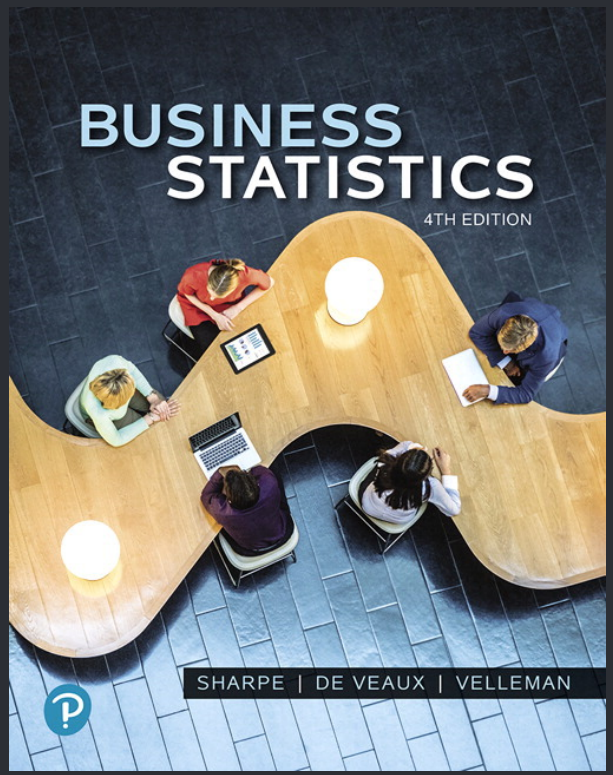
3. Let’s do some statistics, yo

Canvas

- Access at `canvas.utexas.edu`
- This is your home base for the course
- Make sure you can log in and are enrolled in STA 371G in Canvas

Textbook

- Access the textbook as an eBook through MyStatLab OR buy a hardcover or looseleaf textbook (Sharpe, *Business Statistics*, 4th ed) that includes a MyStatLab access code



Using Zoom

- Make sure to join using your UT Zoom account (you should *not* find yourself in the waiting room)
- Stay muted during class, unless you are called on
- Raise your hand (in the Participant window) to ask questions
- Feel free to use the chat (Leyang will monitor during class)

Class structure

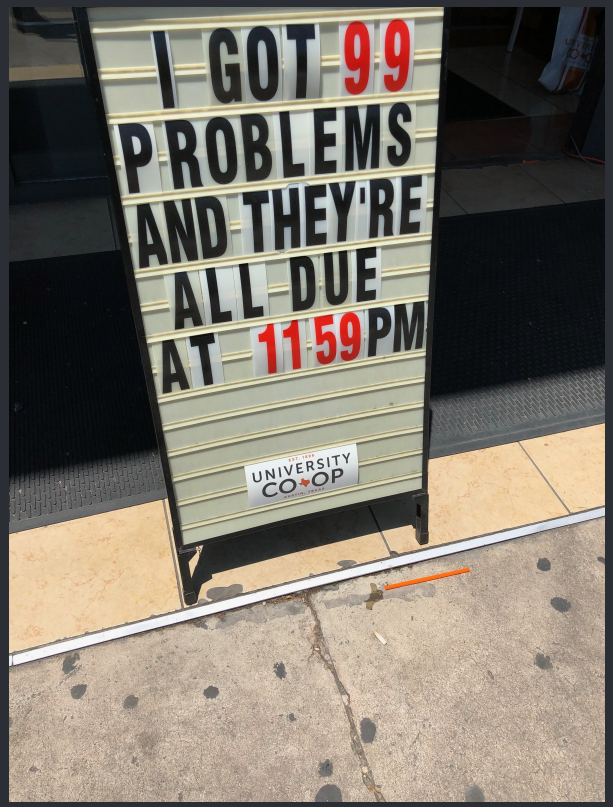
- Understanding the concepts really only comes from practice
- Class will be synchronous; class time will be divided between lecture and practice
- We will use Learning Catalytics so you can practice the concepts during class
- No cost to use this
- Graded on participation, not correctness; answer 75% of the questions to get 100% of the credit

Let's try out Learning Catalytics

1. Go to Canvas, and click MyLab and Mastering in the left sidebar. Log in with your Pearson account, or create one.
2. In the Learning Catalytics box, click Join Session in Progress.
3. Answer the question there!

Homework

- Why homework?
- 12 homework assignments during the semester, due Mondays
- We will use MyStatLab for online homework; purchase and register through Canvas (you'll get a 2-week free trial)



Statistical computing

- We will use R for statistical analysis throughout the course
- This is industrial-strength, state-of-the-art, and free software for statistical computing
- We will access R through RStudio, a graphical interface for R
- Download R and RStudio at `rstudio.com`

Quizzes

- Short weekly quizzes on Tuesdays at 6:30 PM
- No midterm or final exams!
- Quizzes are cumulative and may include questions from earlier in the semester
- Take quizzes on your own (quizzes are automatically analyzed for patterns of suspicious behavior)
- You'll have access to R during every quiz
- I will drop the lowest of your quiz scores

Team project

- You will apply regression techniques (we'll learn about this!) to build a predictive model for an interesting data set
- Six deliverables throughout the semester:
    1. Initial proposal (March 4)
    2. Literature review (March 11)
    3. Final proposal (April 1)
    4. Exploratory data analysis (April 15)
    5. PowerPoint deck (April 29)
    6. Presentation to class (May 3-5 & 18)

Academic integrity

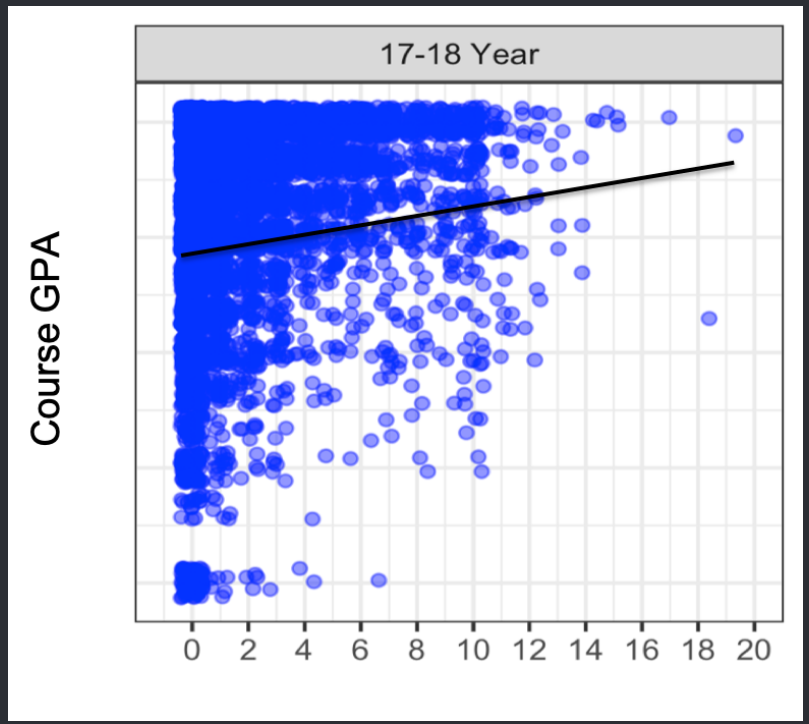
- Academic integrity (scholastic honesty) is an important prerequisite for academic achievement (a degree obtained by cheating doesn't mean anything, and cheating devalues the whole process for everyone!)
- Do not collaborate on projects (outside of your team project group) or on quizzes
- Don't be afraid to ask if you have a question about what is allowed

Grading

Quizzes	**50%**
Project	**25%**
Homework	**15%**
Learning Catalytics	**10%**

PLUS (Peer-Led Undergraduate Studying)

- Weekly, student-run study groups
- You can apply to be a facilitator or just participate in any of the sessions
- Students who attend more PLUS sessions tend to get higher grades!
- PLUS will begin in the third week of the semester



How to get an A in STA 371G

- Work on more problems than are assigned in the homework; can find problems in the book and on MyStatLab
- Consider attending the PLUS sessions
- Get help when you need it
  - My office hours (after every class, until 10:30 AM)
  - TA office hours (see syllabus for schedule)
  - Post questions on Piazza (accessible through Canvas)
  - E-mail the staff address (`sta371g-lukoff@austin.utexas.edu`) to ask questions
  - E-mail me directly (`brian.lukoff@utexas.edu`) for private questions or to set up an appointment

1. Find someone who…

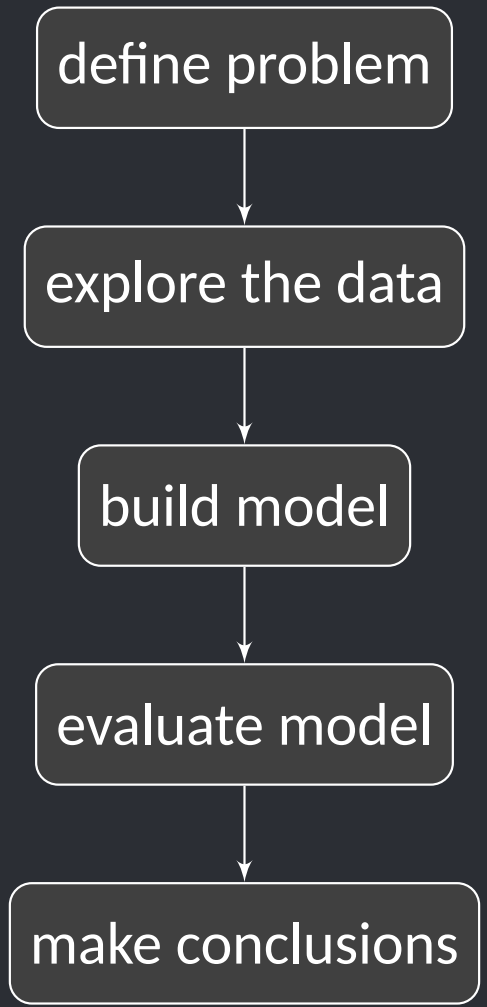
2. Course logistics

3. Let’s do some statistics, yo

Purpose of a model

- **Make a prediction** about one variable based on the others
- **Understand the relationships** between the variables

Data analysis process



Define the problem

What personal characteristics about an instructor do you think are predictive of the scores they receive on student evaluations?

Economics of Education Review

Volume 24, Issue 4, August 2005, Pages 369–376

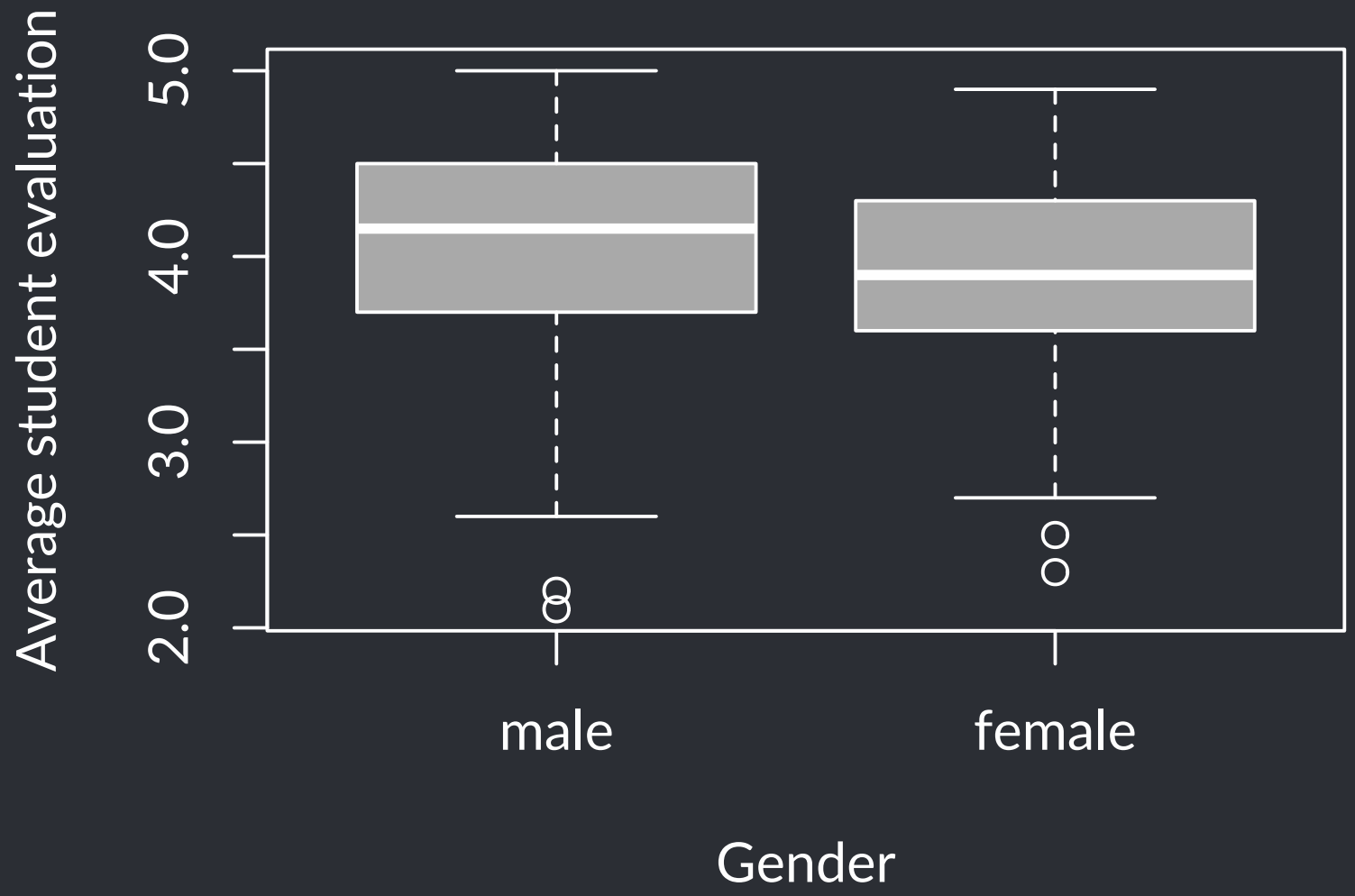
Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity

Daniel S. Hamermesh, Amy Parker

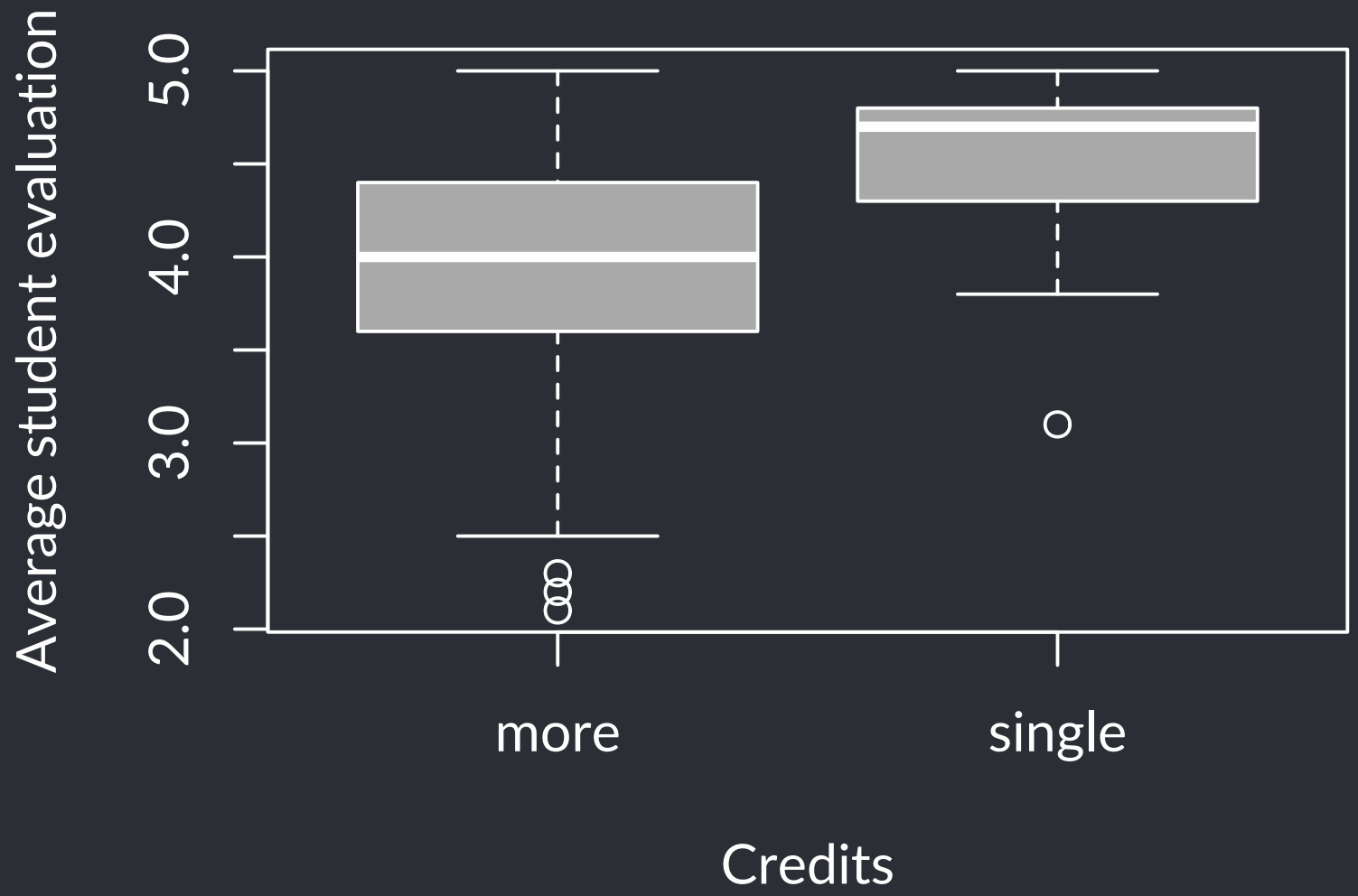
Hamermesh & Parker (2005) data set

- Student evaluations of $N = 463$ instructors at UT Austin, 2000-2002
- For each instructor:
  - **beauty**: average score from a six-student panel)
  - **gender**: male or female
  - **credits**: single- or multi-credit course
  - **age**: age of instructor
  - (and more...)

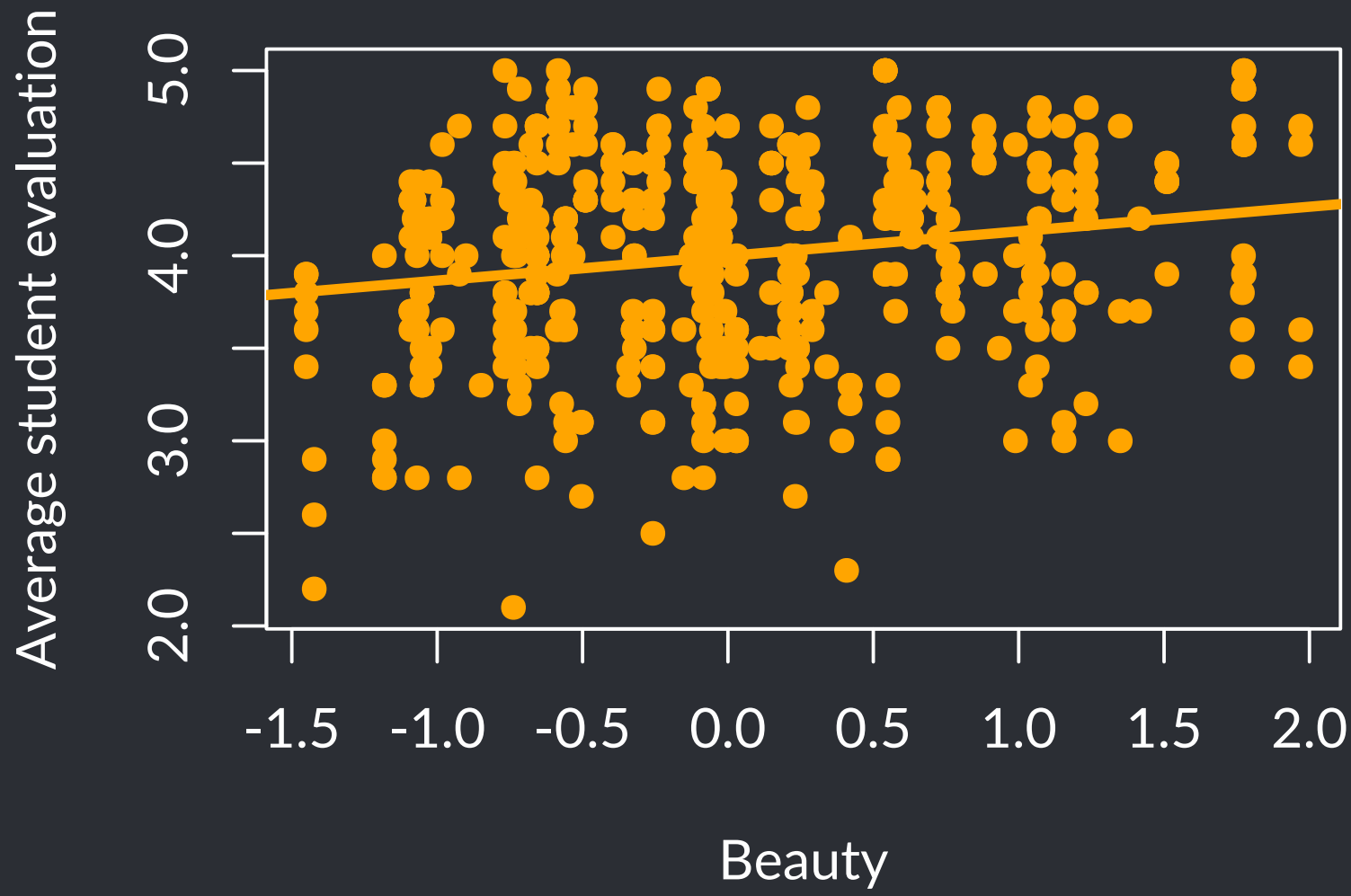
Explore the data



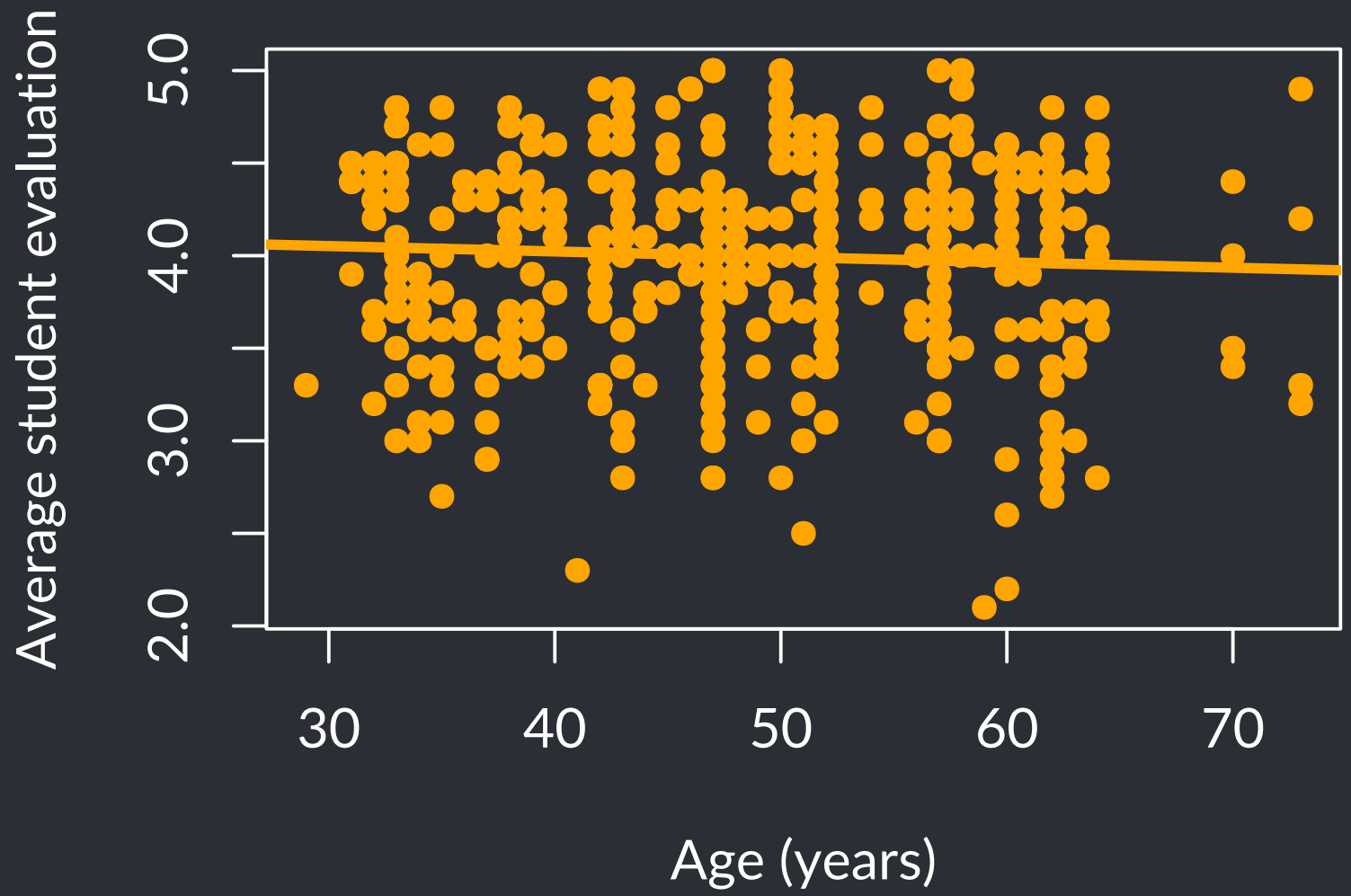
Explore the data



Explore the data



Explore the data



Build the model

A regression model lets us create a model that incorporates all of these relationships to best predict evaluation scores:

$$\widehat{\text{eval}} = 4.13 + 0.16 \cdot \text{beauty} - 0.2 \cdot \text{female} + 0.58 \cdot \text{credits} + 0 \cdot \text{age}$$

Build the model

A regression model lets us create a model that incorporates all of these relationships to best predict evaluation scores:

$$\widehat{\text{eval}} = 4.13 + 0.16 \cdot \text{beauty} - 0.2 \cdot \text{female} + 0.58 \cdot \text{credits} + 0 \cdot \text{age}$$

We predict a 40-year-old female, with a beauty score of 2, teaching a multi-credit course would get an evaluation score of

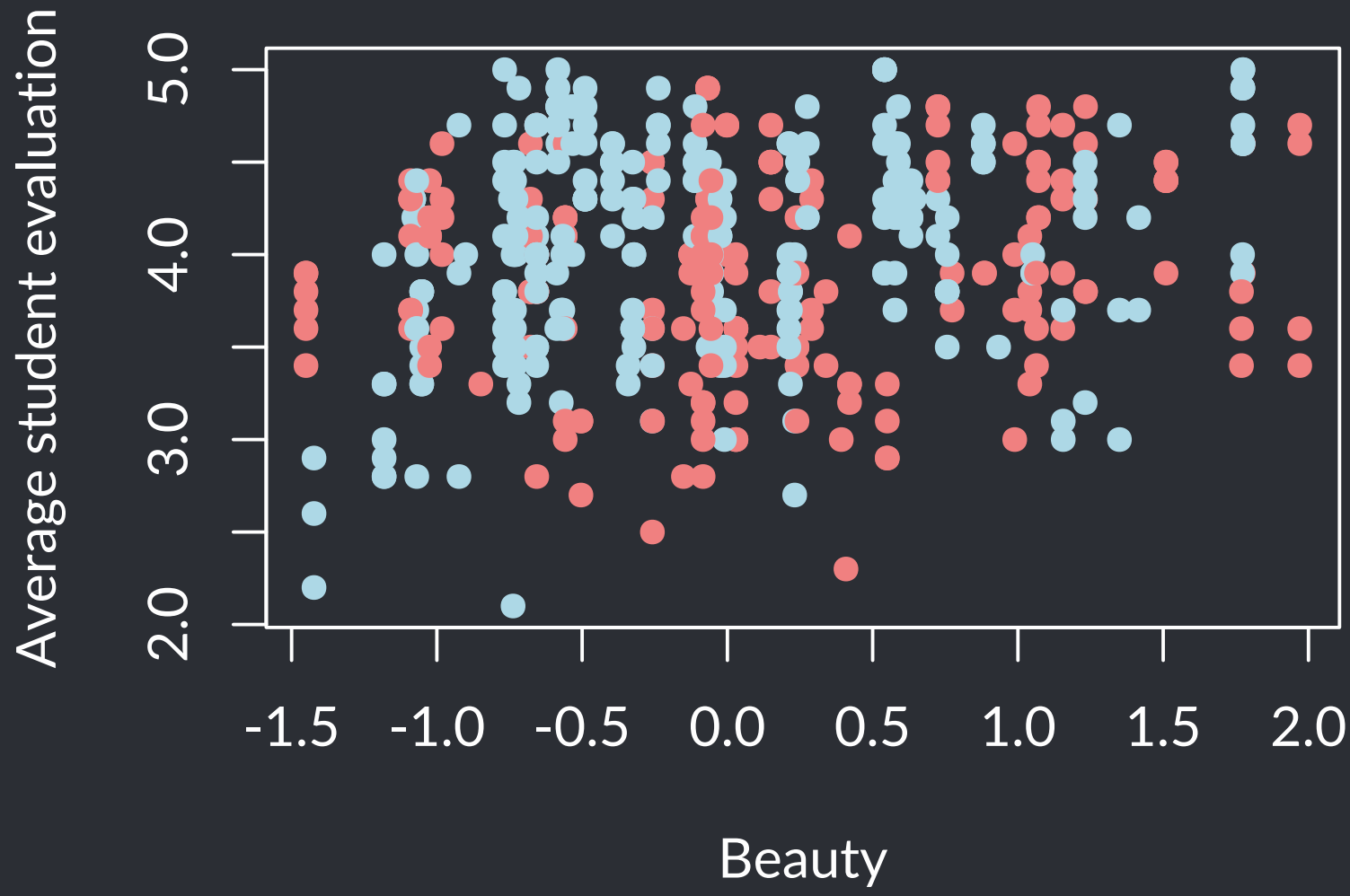
$$\widehat{\text{eval}} = 4.13 + 0.16 \cdot 2 - 0.2 \cdot 1 + 0.58 \cdot 0 = 4.18.$$

Evaluate the model

How could you evaluate the quality of this model?

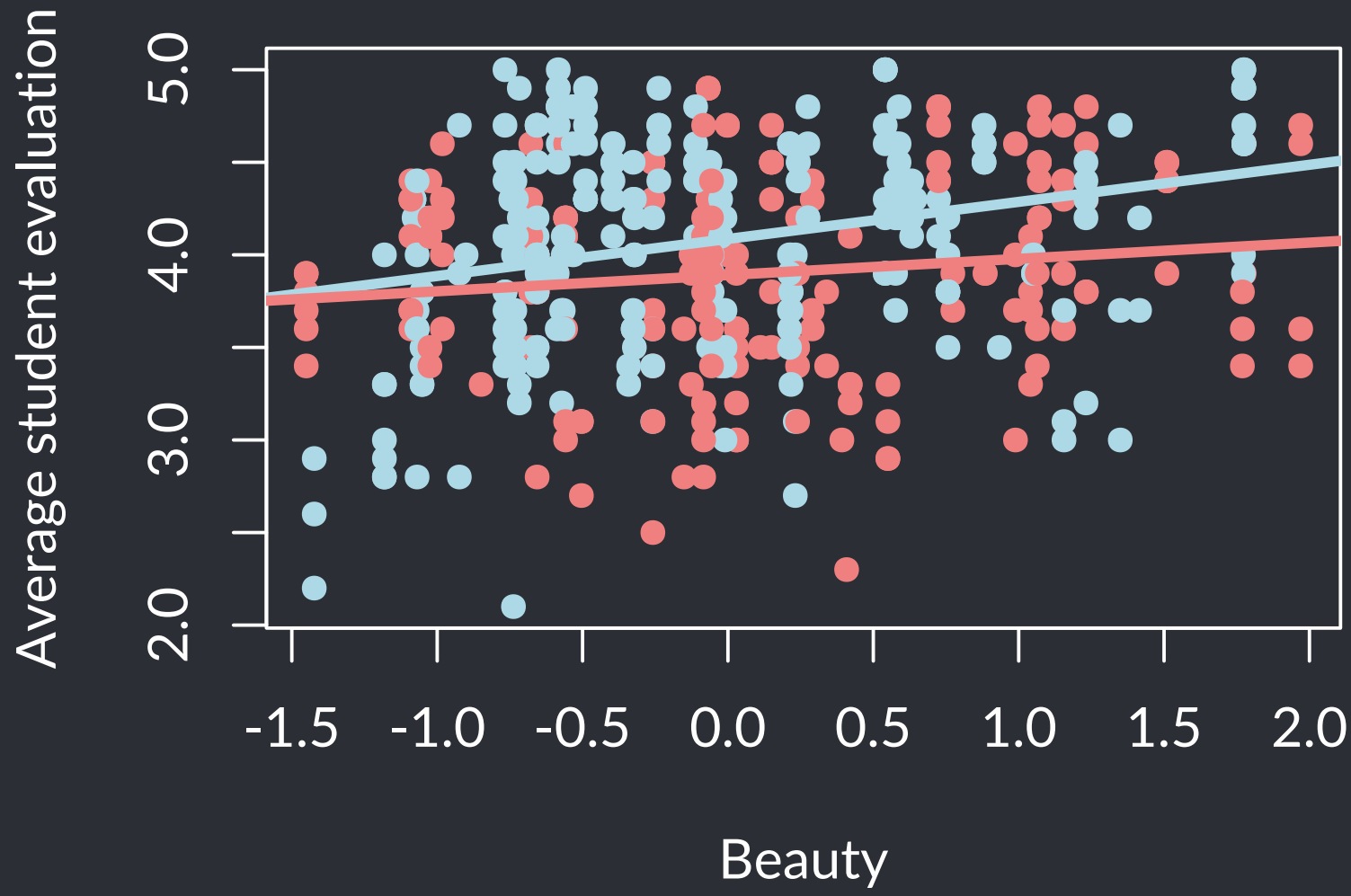
Can we do better?

Do you see a difference between men (blue) and women (red)?



Can we do better?

Do you see a difference between men (blue) and women (red)?



Five for the weekend

1. Read the syllabus
2. Do the first reading assignment (§5.1-5.8, on probability)
3. Install R and RStudio on your computer (instructions on Canvas)
4. Get registered with MyStatLab (access through Canvas)
5. Watch the video message from Michelle about PLUS, and apply to be a facilitator if you are interested