Modeling Basics / Intro to R

READING:	0.1, 0.2
EXERCISES:	ALL CHAPTER 0
ASSIGNED:	HW 1
PRODUCER:	DR. MARIO

Course Website / Syllabus

- Access Course Website Through Canvas
- Cover Syllabus
  - Office Hours
  - Grading and Curving
  - Attendance: UNC Check-In App
  - Homework
  - Quizzes
  - Exams
  - PDFs and Gradescope
  - Grade Disputes
  - Honor Code
- Usage of Course Website and Canvas

Course Website / Syllabus

- Access Course Website Through Canvas
- Cover Syllabus
  - Office Hours
  - Grading and Curving
  - Attendance: UNC Check-In App
  - Homework
  - Quizzes
  - Exams
  - PDFs and Gradescope
  - Grade Disputes
  - Honor Code
- Usage of Course Website and Canvas

umbers with a context.

n + Departures from a Pattern

How do we identify the actual pattern?

How do we characterize the departures (errors)?

Statistical Modeling

Find a model for a relationship between a response variable (Y) and one (or more) predictor/explanatory variables (X_1, X_2, ..., X_k).

Purposes for Statistical Modeling

1. Making predictions
2. Understanding relationships
3. Assessing differences

Types of Variables

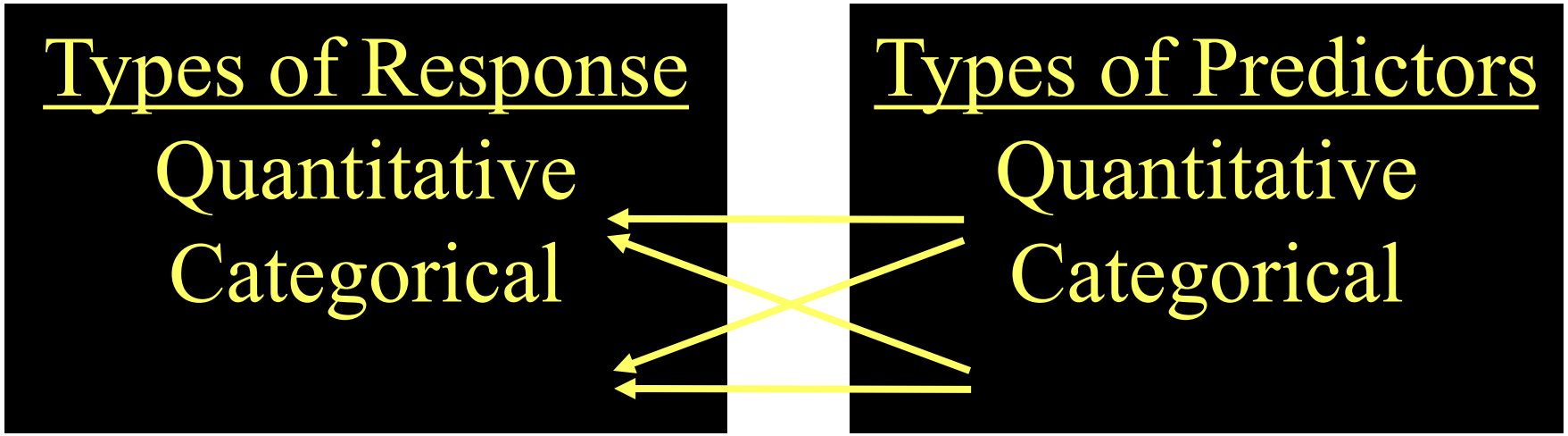
Quantitative: expressible as numbers for which arithmetic makes sense

Categorical: divides sample points into groups

Binary = categorical with just two groups

Two Main Themes of STOR 455

(1)



(2) Allow for models with *multiple* predictors.

Building a Statistical Model: Four Step Process

1. CHOOSE – Pick a form for the model
2. FIT – Estimate any parameters
3. ASSESS – Is the model adequate? Could it be simpler? Are conditions met?
4. USE – Answer the question of interest

General form of a model:

$$Y = f(X) + \varepsilon$$

Random Error

"Expected" Y for some combination of predictors

Data

Model

Error

Example: Lego Prices

Question:
How can we predict the price of a Lego set?

Predictor variables: Start with *none*.

Example: Constant Model

$$Y = c + \varepsilon$$

where c is an (unknown) constant

Terminology:
The constant c is a parameter of this model.
We use data to provide a sample estimate of c.

How should we estimate c from data?

Predicted Value for Response

Get an for Y using the predictors and the model with estimated parameters.

Notation: The predicted y is denoted $\hat{y}$

For the constant model: $\hat{y} = \hat{c}$

Examples: (sample mean)

(sample median)

Questions

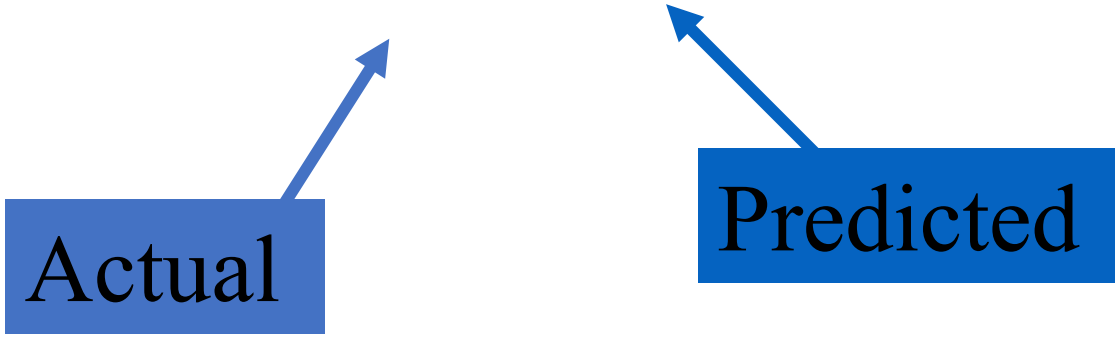
(1) Which estimator (mean or median) is better?

(2) Is either model any good?

Residuals

Using the predicted value for each sample case the residual is

Residual = $y - \hat{y}$



Criteria to Minimize Residuals?

Sum of residuals:

Sum of absolute deviations:

Sum of squared errors:

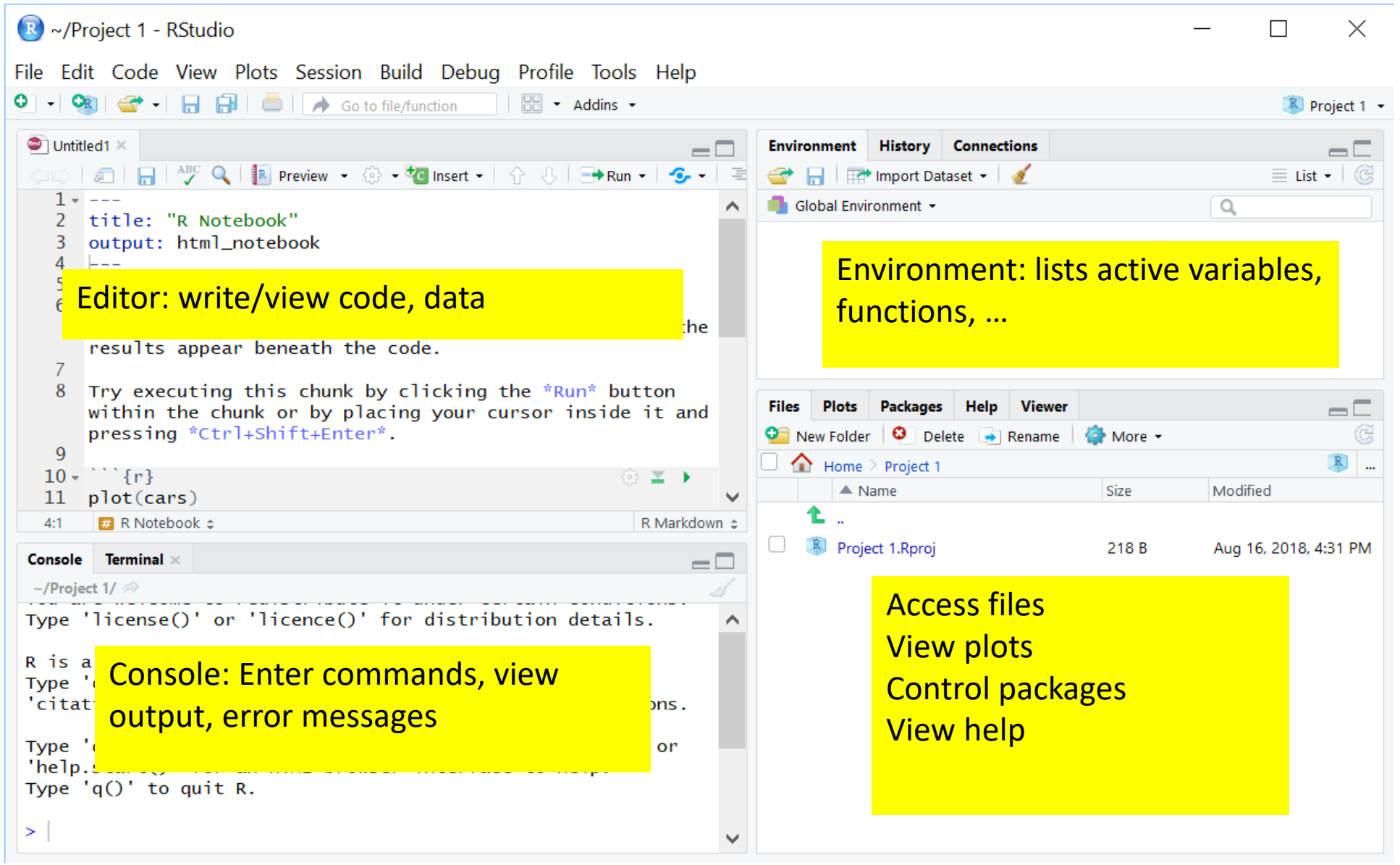
Technology

We need software to automate computations…

R – a free, widely used, open source statistics package

Rstudio – an interface for R

RStudio


~/Project 1 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project 1
Untitled1
Preview
Insert
Run
---
title: "R Notebook"
output: html_notebook

Editor: write/view code, data
results appear beneath the code.
Try executing this chunk by clicking the *Run* button
within the chunk or by placing your cursor inside it and
pressing *Ctrl+Shift+Enter*.
```{r}
plot(cars)
4:1
R Notebook
R Markdown
Console
Terminal
~/Project 1/
Type 'license()' or 'licence()' for distribution details.
Console: Enter commands, view output, error messages
Type 'q()' to quit R.
Environment
History
Connections
Import Dataset
List
Global Environment
Environment: lists active variables, functions, …
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home
Project 1
Name
Size
Modified
Project 1.Rproj
218 B
Aug 16, 2018, 4:31 PM
Access files
View plots
Control packages
View help
```

A First RStudio Session

- Load the *Lego* data into R
- Summarize the *Amazon_Price* variable
  - Numerical: mean and median
  - Graphical: histogram, boxplot
- Compute and evaluate residuals

A First RStudio Session

Load the *lego* data into R

```r
# loads a package needed to use the read_csv() function
# install package before first using it for the first time

library(readr)

# loads the lego dataframe into the environment from GitHub

lego <- read_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/lego.csv")

# Alternative way to load dataframe (remove # to use)
# lego.csv must be saved in the same folder as this notebook!

#lego <- read_csv("lego.csv")


# Shows the variables and first 6 cases (by default)

head(lego)
```

A First RStudio Session

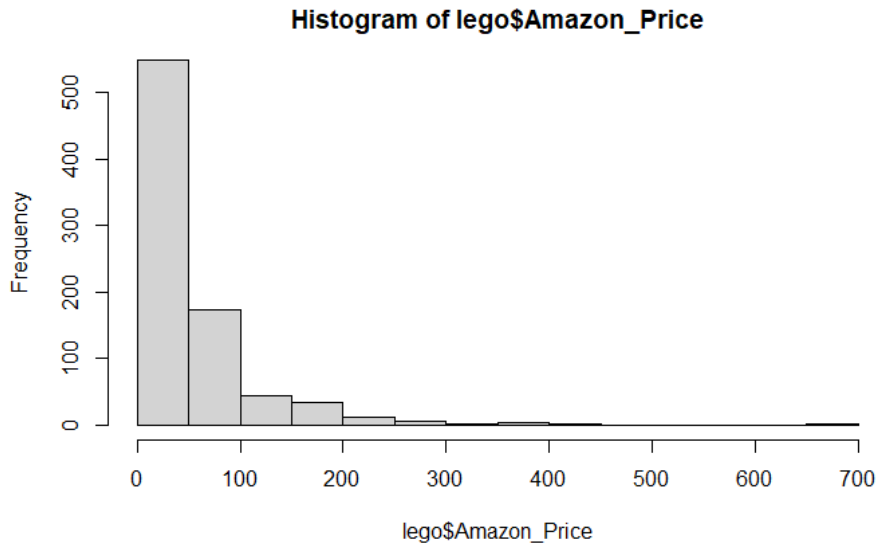
Summarize the *Amazon_Price* variable - Numerical: mean and median

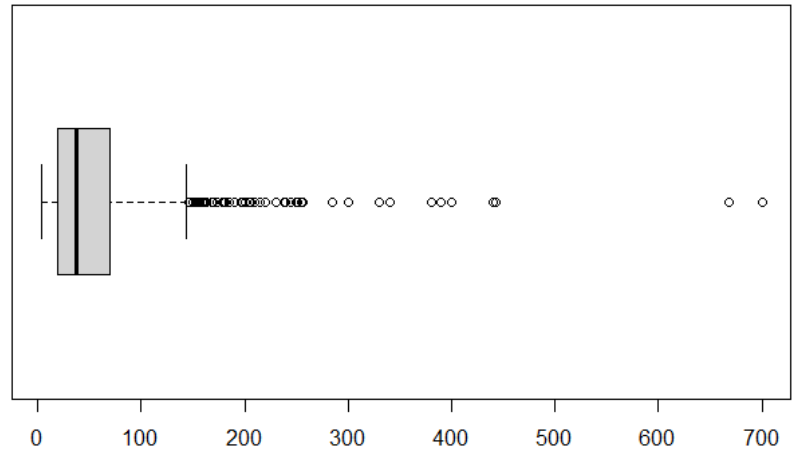
```
# dataframe$variable_name

mean(lego$Amazon_Price, na.rm = TRUE)
median(lego$Amazon_Price, na.rm = TRUE)

[1] 57.8232
[1] 37.325
```

A First RStudio Session

Summarize the *Amazon_Price* variable - Graphical: histogram, boxplot

```
hist(lego$Amazon_Price)

boxplot(lego$Amazon_Price, horizontal = TRUE)
```





A First RStudio Session

Compute and evaluate residuals

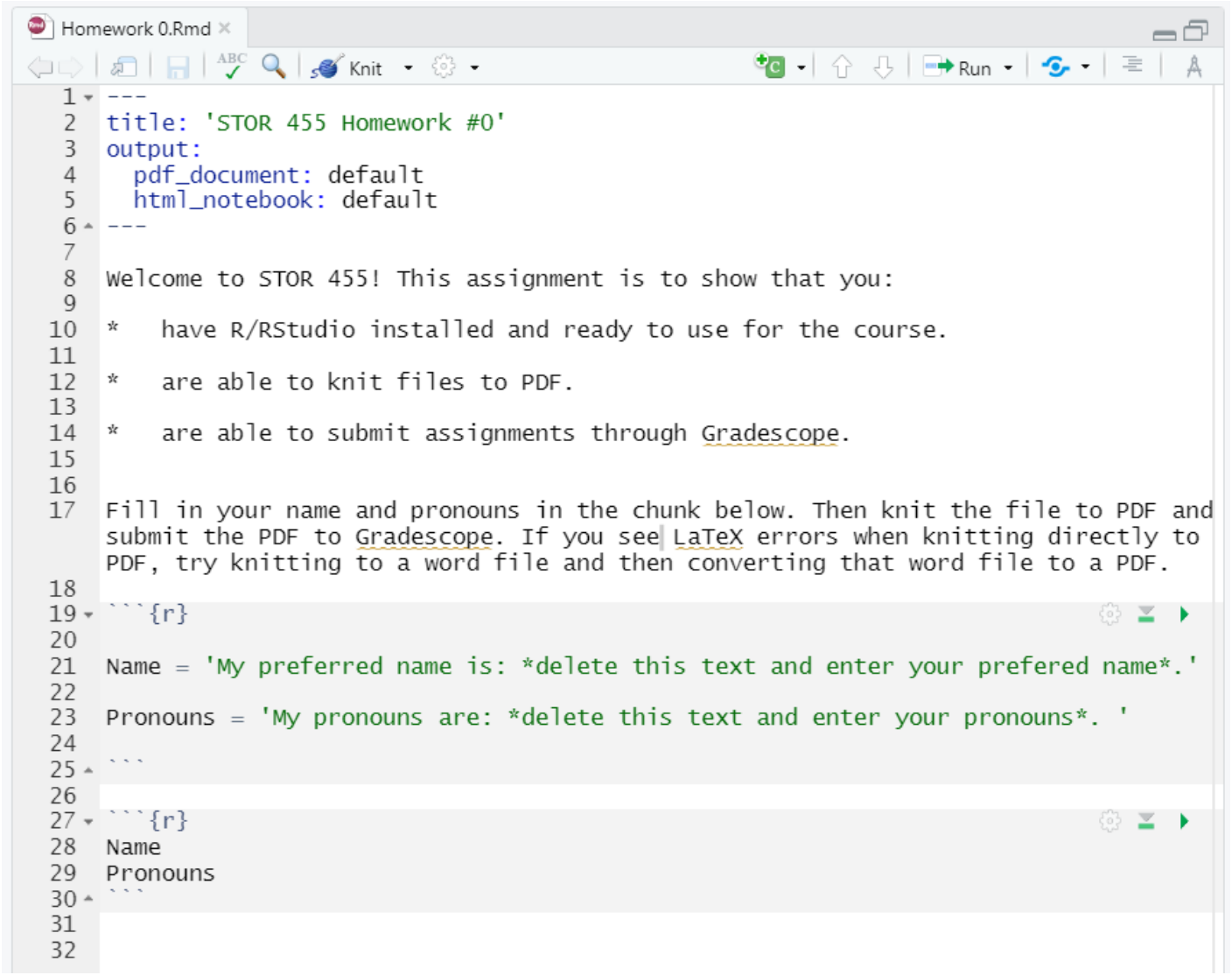
```
# removes NA Amazon_Prices

lego_rm_AP_na = subset(lego, is.na(Amazon_Price) == FALSE)

# Assignment operators in R: = vs. <-

xbar = mean(lego_rm_AP_na$Amazon_Price)
m = median(lego_rm_AP_na$Amazon_Price)

residxbar = lego_rm_AP_na$Amazon_Price - xbar
residm = lego_rm_AP_na$Amazon_Price - m

sum(residxbar^2)
sum(residm^2)
```

Can we use a predictor to improve the model?

See Chapter 1

Two-sample t-test for a difference in means
(comparing only two themes)

Homework 0


Homework 0.Rmd
Knit
Run
---
title: 'STOR 455 Homework #0'
output:
pdf_document: default
html_notebook: default

Welcome to STOR 455! This assignment is to show that you:
* have R/RStudio installed and ready to use for the course.
* are able to knit files to PDF.
* are able to submit assignments through Gradescope.
Fill in your name and pronouns in the chunk below. Then knit the file to PDF and submit the PDF to Gradescope. If you see LaTeX errors when knitting directly to PDF, try knitting to a word file and then converting that word file to a PDF.
```{r}
Name = 'My preferred name is: *delete this text and enter your prefered name*.'
Pronouns = 'My pronouns are: *delete this text and enter your pronouns*. '
```
```{r}
Name
Pronouns
```