# Using GitHub and RStudio to Facilitate Authentic Learning Experiences in a Regression Analysis Course

Maria Tackett
Duke University

WSDS 2019

bit.ly/wsds2019-regression

Introduction

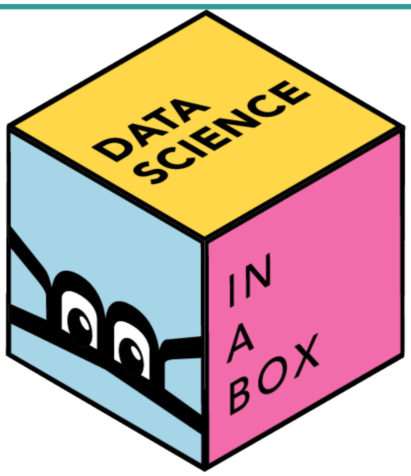
Assignments

Challenges
+
Learning
Outcomes

# Motivation

> *"To be prepared for statistics and data science careers,* ***students need facility with professional statistical analysis software****, the ability to access and wrangle data in various ways, and the ability to perform algorithmic problem solving."*
>
> *2014 ASA Curriculum Guidelines for Undergraduate Programs in Statistical Science*

Innovations in intro statistics and data science courses



**Goal:** Create learning experiences to continue cultivating these skills in intermediate courses

# The Course

75 students who have taken introductory statistics, data science, or probability course

Applied course focusing on linear regression, analysis of variance and logistic regression

In-class activities  + computing labs + homework assignments

# Computing Tools

R + R Markdown for analysis and write up

Run Git commands through Git pane

Assign and submit assignments

Platform for collaboration on group assignments

RStudio Cloud

Packages installed + Git configured on Day 1

Çetinkaya-Rudel, M., & Rundel, C. (2017). Infrastructure and Tools for Teaching Computing Throughout the Statistical Curriculum. *The American Statistician,* 72, 58-65.

Introduction

Assignments

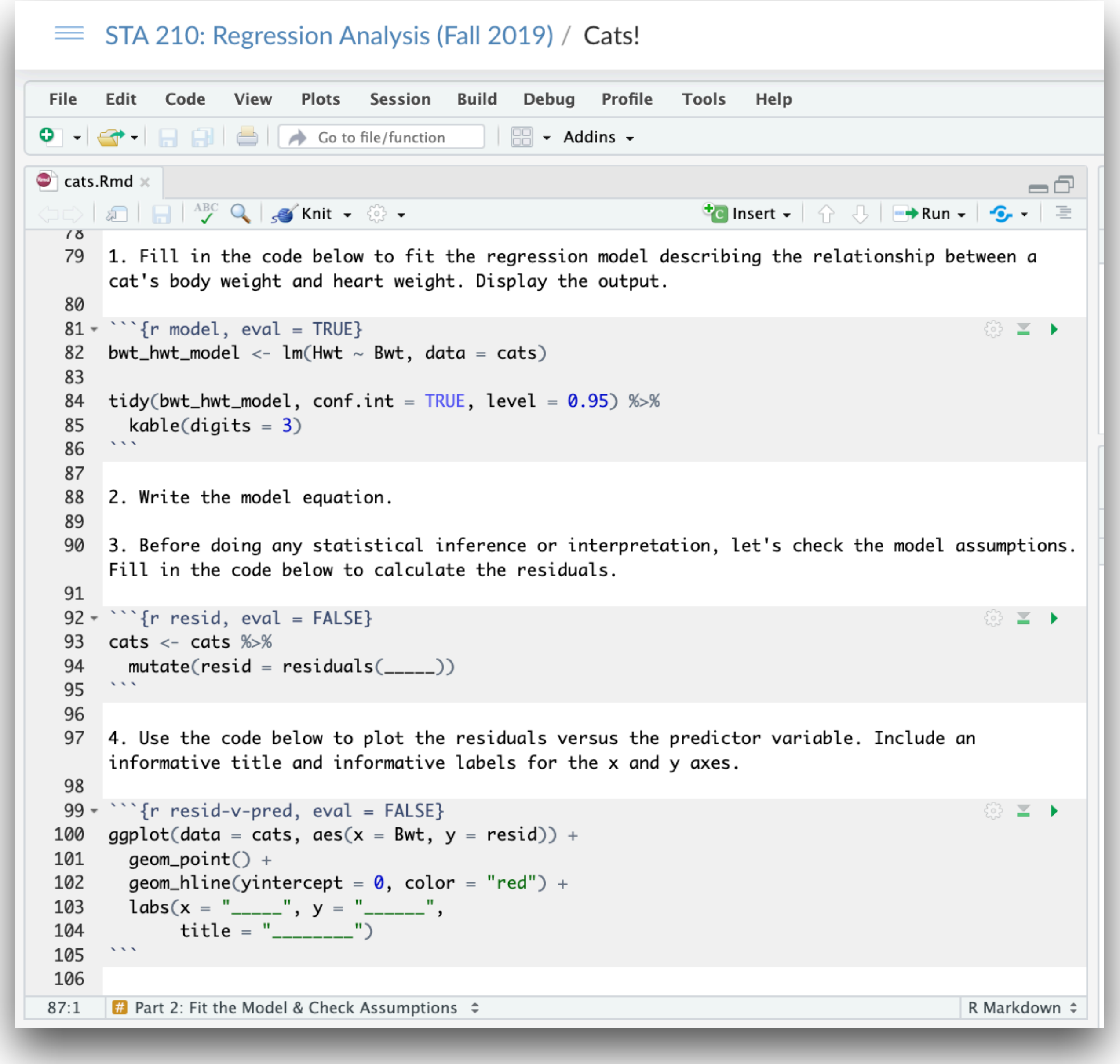
Challenges
+
Learning
Outcomes

# In-Class Activities

1. Create RStudio Cloud project with R Markdown template and data.

2. Students make copy of RStudio Cloud project and answer questions by writing short lines of code and interpreting the results.

3. Display a student's RStudio Cloud project on the classroom screen as the student (or group) shares their work with the class.



~~~
1. Fill in the code below to fit the regression model describing the relationship between a
cat's body weight and heart weight. Display the output.

```{r model, eval = TRUE}
bwt_hwt_model <- lm(Hwt ~ Bwt, data = cats)

tidy(bwt_hwt_model, conf.int = TRUE, level = 0.95) %>%
  kable(digits = 3)
~~~

2. Write the model equation.

3. Before doing any statistical inference or interpretation, let's check the model assumptions.
Fill in the code below to calculate the residuals.

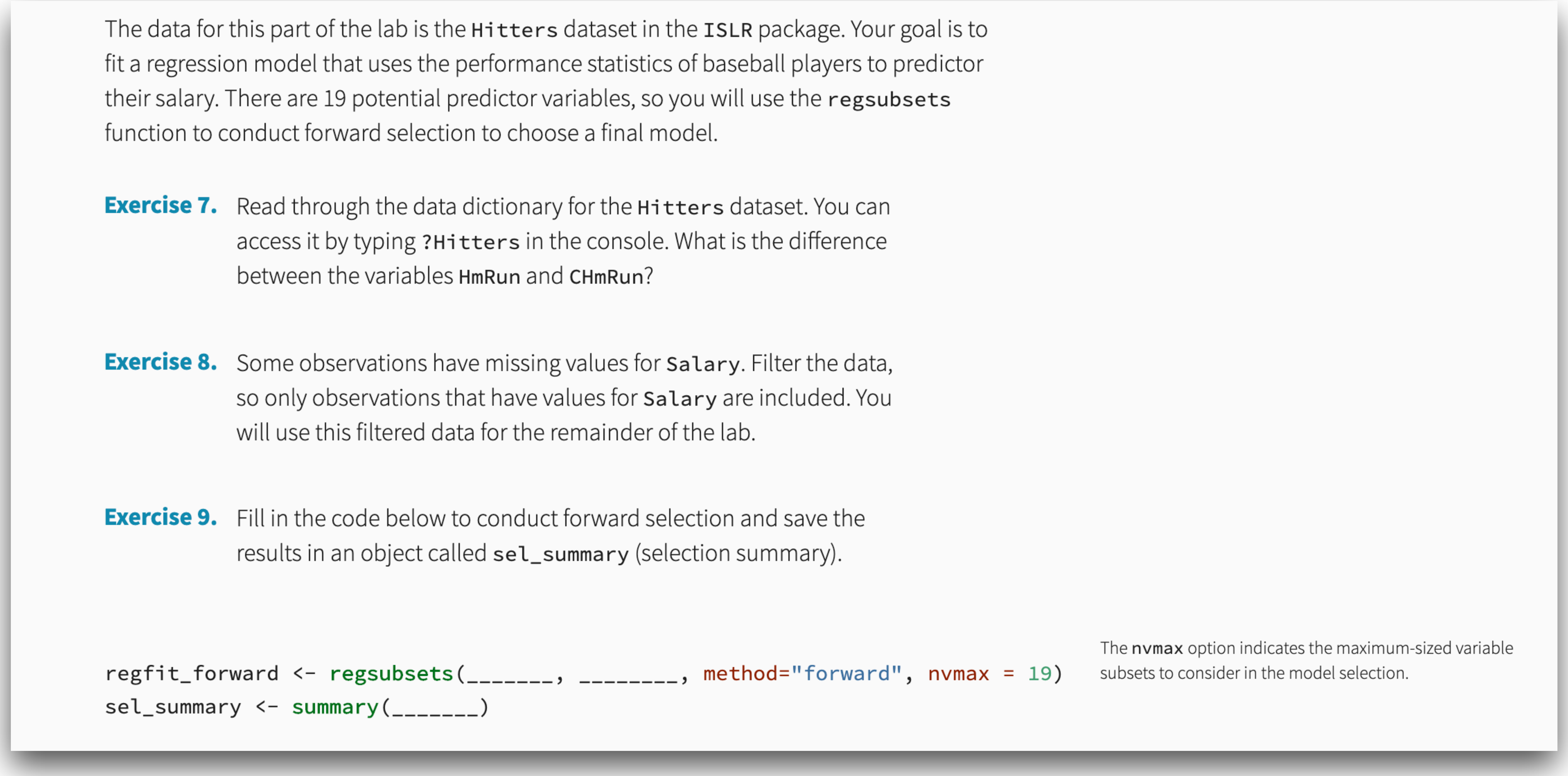
```{r resid, eval = FALSE}
cats <- cats %>%
  mutate(resid = residuals(_____))
```

4. Use the code below to plot the residuals versus the predictor variable. Include an
informative title and informative labels for the x and y axes.

```{r resid-v-pred, eval = FALSE}
ggplot(data = cats, aes(x = Bwt, y = resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red") +
  labs(x = "_____", y = "______",
       title = "________")
```

```
```

# Computing Assignments



The data for this part of the lab is the `Hitters` dataset in the `ISLR` package. Your goal is to fit a regression model that uses the performance statistics of baseball players to predictor their salary. There are 19 potential predictor variables, so you will use the `regsubsets` function to conduct forward selection to choose a final model.

**Exercise 7.** Read through the data dictionary for the `Hitters` dataset. You can access it by typing `?Hitters` in the console. What is the difference between the variables `HmRun` and `CHmRun`?

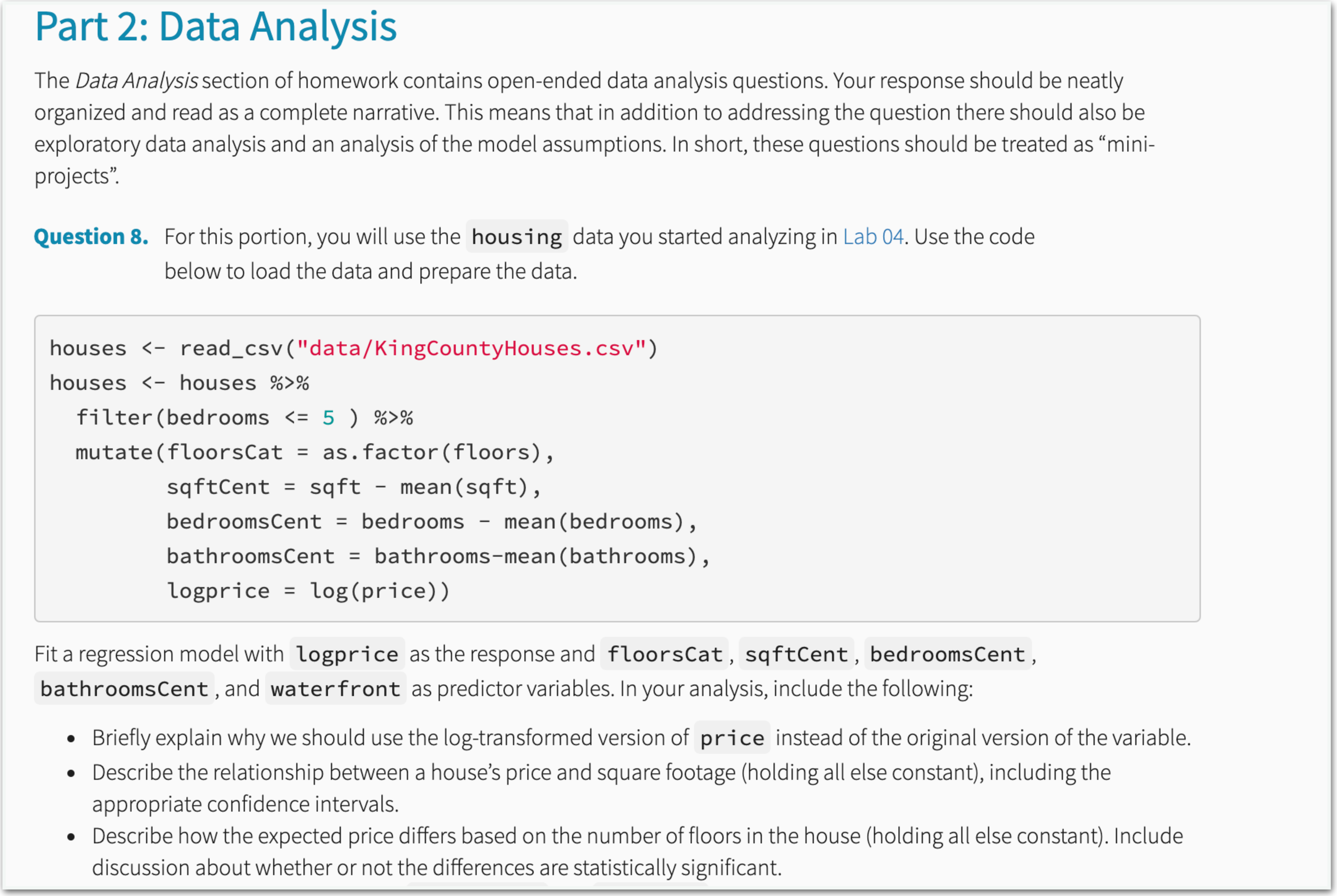
**Exercise 8.** Some observations have missing values for `Salary`. Filter the data, so only observations that have values for `Salary` are included. You will use this filtered data for the remainder of the lab.

**Exercise 9.** Fill in the code below to conduct forward selection and save the results in an object called `sel_summary` (selection summary).

```
regfit_forward <- regsubsets(_______, ________, method="forward", nvmax = 19)
sel_summary <- summary(_______)
```

The `nvmax` option indicates the maximum-sized variable subsets to consider in the model selection.

# Homework Assignments



## Part 2: Data Analysis

The *Data Analysis* section of homework contains open-ended data analysis questions. Your response should be neatly organized and read as a complete narrative. This means that in addition to addressing the question there should also be exploratory data analysis and an analysis of the model assumptions. In short, these questions should be treated as "mini-projects".

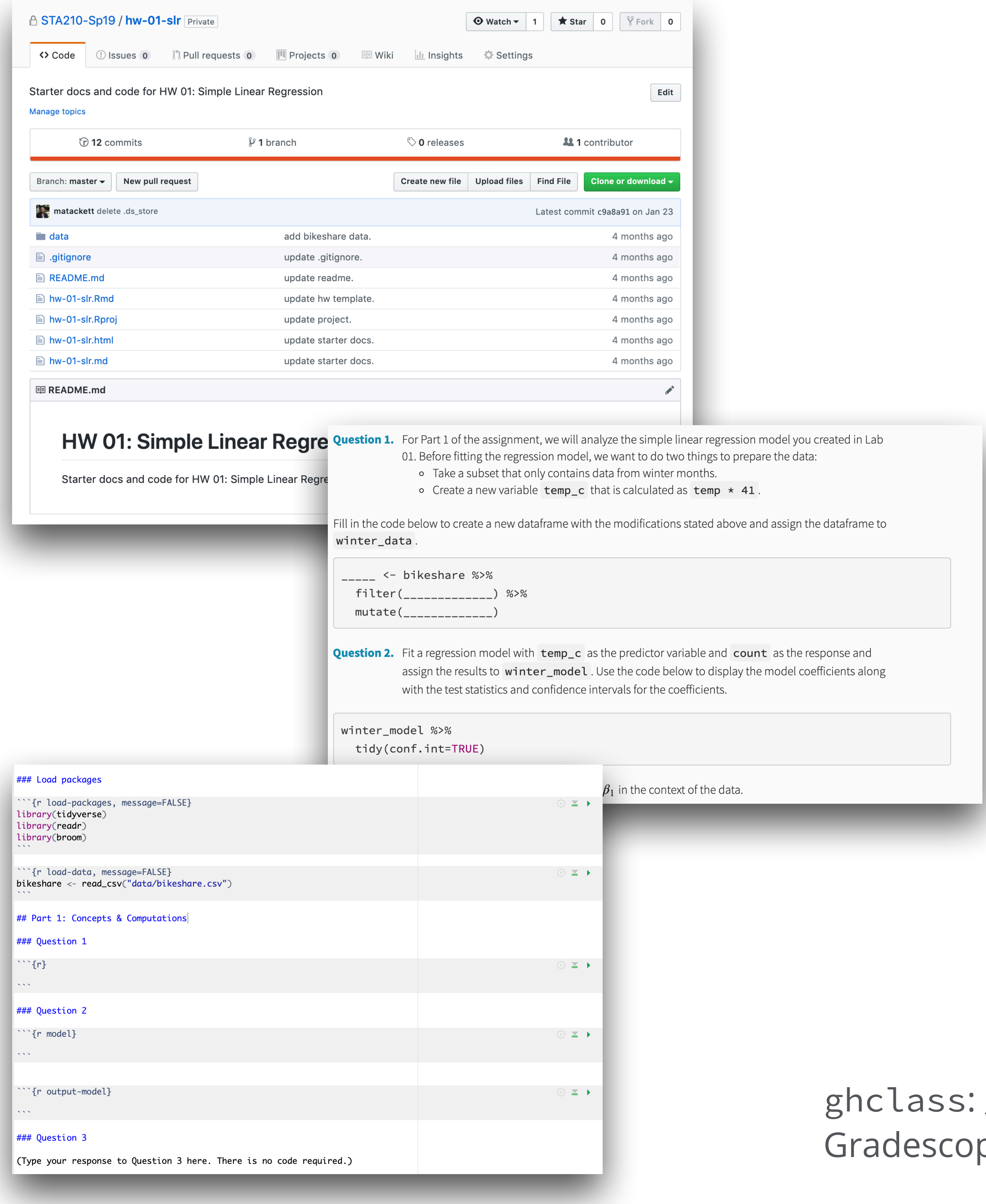
**Question 8.** For this portion, you will use the `housing` data you started analyzing in Lab 04. Use the code below to load the data and prepare the data.

```
houses <- read_csv("data/KingCountyHouses.csv")
houses <- houses %>%
  filter(bedrooms <= 5 ) %>%
  mutate(floorsCat = as.factor(floors),
         sqftCent = sqft - mean(sqft),
         bedroomsCent = bedrooms - mean(bedrooms),
         bathroomsCent = bathrooms-mean(bathrooms),
         logprice = log(price))
```

Fit a regression model with `logprice` as the response and `floorsCat`, `sqftCent`, `bedroomsCent`, `bathroomsCent`, and `waterfront` as predictor variables. In your analysis, include the following:

- Briefly explain why we should use the log-transformed version of `price` instead of the original version of the variable.
- Describe the relationship between a house's price and square footage (holding all else constant), including the appropriate confidence intervals.
- Describe how the expected price differs based on the number of floors in the house (holding all else constant). Include discussion about whether or not the differences are statistically significant.

# Assignment Workflow



Starter docs and code for HW 01: Simple Linear Regression

matackett delete .ds_store — Latest commit c9a8a91 on Jan 23

| | | |
|---|---|---|
| data | add bikeshare data. | 4 months ago |
| .gitignore | update .gitignore. | 4 months ago |
| README.md | update readme. | 4 months ago |
| hw-01-slr.Rmd | update hw template. | 4 months ago |
| hw-01-slr.Rproj | update project. | 4 months ago |
| hw-01-slr.html | update starter docs. | 4 months ago |
| hw-01-slr.md | update starter docs. | 4 months ago |

README.md

HW 01: Simple Linear Regre

Starter docs and code for HW 01: Simple Linear Regre

**Question 1.** For Part 1 of the assignment, we will analyze the simple linear regression model you created in Lab 01. Before fitting the regression model, we want to do two things to prepare the data:

- Take a subset that only contains data from winter months.
- Create a new variable `temp_c` that is calculated as `temp * 41`.

Fill in the code below to create a new dataframe with the modifications stated above and assign the dataframe to `winter_data`.

```
_____ <- bikeshare %>%
  filter(_____________) %>%
  mutate(_____________)
```

**Question 2.** Fit a regression model with `temp_c` as the predictor variable and `count` as the response and assign the results to `winter_model`. Use the code below to display the model coefficients along with the test statistics and confidence intervals for the coefficients.

```
winter_model %>%
  tidy(conf.int=TRUE)
```

$\beta_1$ in the context of the data.

~~~
### Load packages

```{r load-packages, message=FALSE}
library(tidyverse)
library(readr)
library(broom)
```

```{r load-data, message=FALSE}
bikeshare <- read_csv("data/bikeshare.csv")
```

## Part 1: Concepts & Computations

### Question 1

```{r}

```

### Question 2

```{r model}

```


```{r output-model}

```

### Question 3

(Type your response to Question 3 here. There is no code required.)
~~~

1. Create starter repo in GitHub.

2. Make a copy of the starter repo for each student (or team) using **`ghclass`**, an R package for managing the GitHub organization for a course.

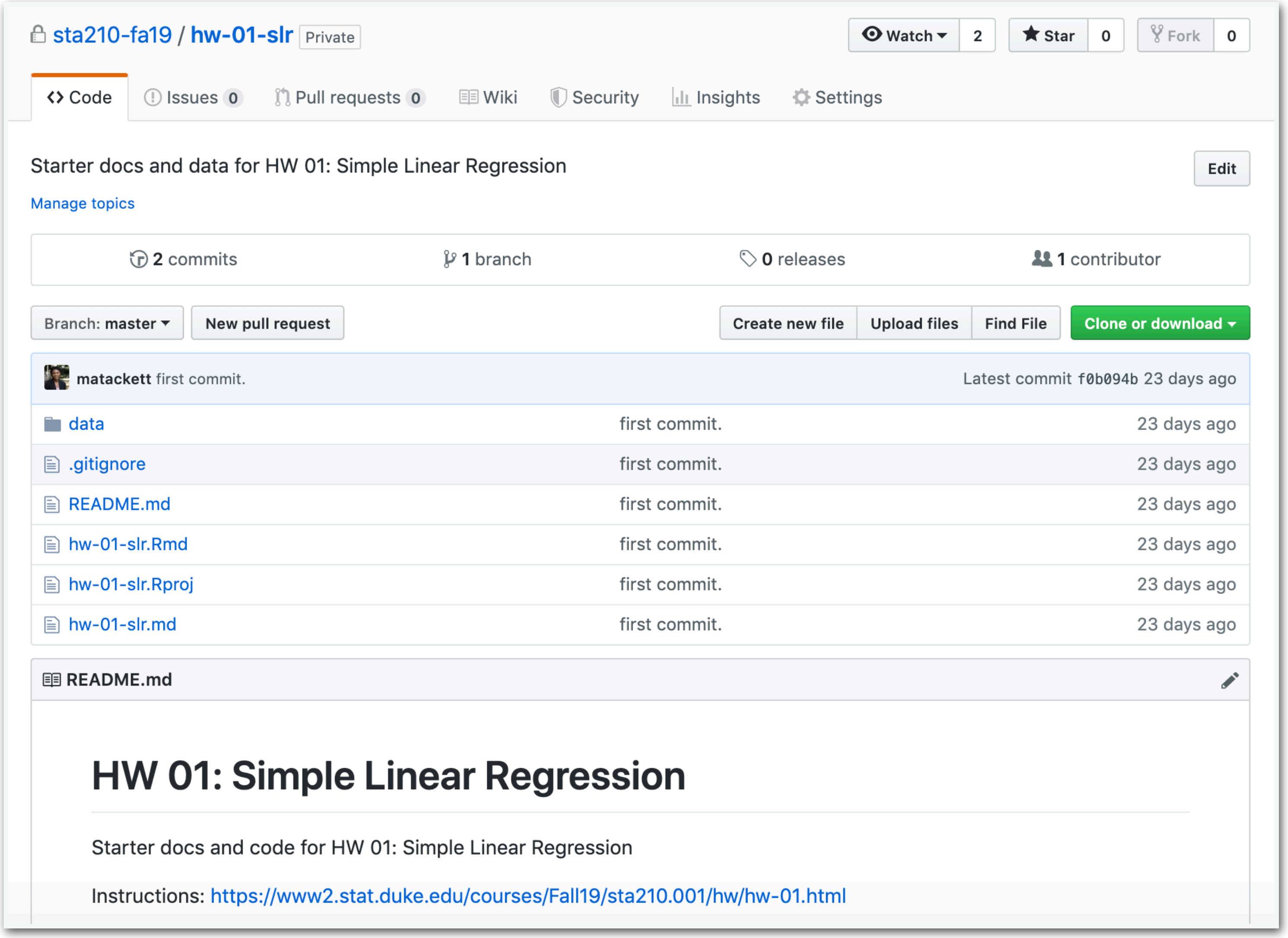
3. Students clone repo into RStudio, write code and narrative in R Markdown, and push their work to GitHub.

4. Clone repos and grade the completed Markdown files in Gradescope.

`ghclass`: https://rundel.github.io/ghclass/articles/ghclass.html
Gradescope: https://www.gradescope.com/

# Starter Repo

Create starter repo as a R project and push to GitHub.
Make copy of repo for each student (or team).



sta210-fa19 / hw-01-slr Private

Starter docs and data for HW 01: Simple Linear Regression

2 commits 1 branch 0 releases 1 contributor

matackett first commit. Latest commit `f0b094b` 23 days ago

| | | |
|---|---|---|
| data | first commit. | 23 days ago |
| .gitignore | first commit. | 23 days ago |
| README.md | first commit. | 23 days ago |
| hw-01-slr.Rmd | first commit. | 23 days ago |
| hw-01-slr.Rproj | first commit. | 23 days ago |
| hw-01-slr.md | first commit. | 23 days ago |

README.md

## HW 01: Simple Linear Regression

Starter docs and code for HW 01: Simple Linear Regression

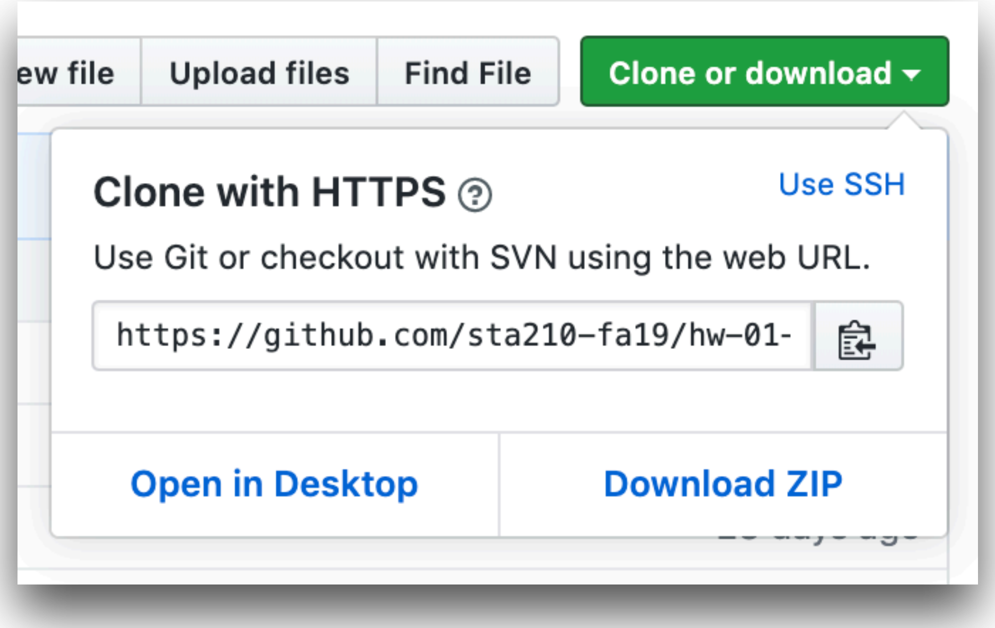
Instructions: https://www2.stat.duke.edu/courses/Fall19/sta210.001/hw/hw-01.html

### Starter repo includes...

- R Markdown template for student responses
- README with link to assignment instructions
- Folder containing the datasets

# Start New Project

Students start a new project in RStudio Cloud from their GitHub repo



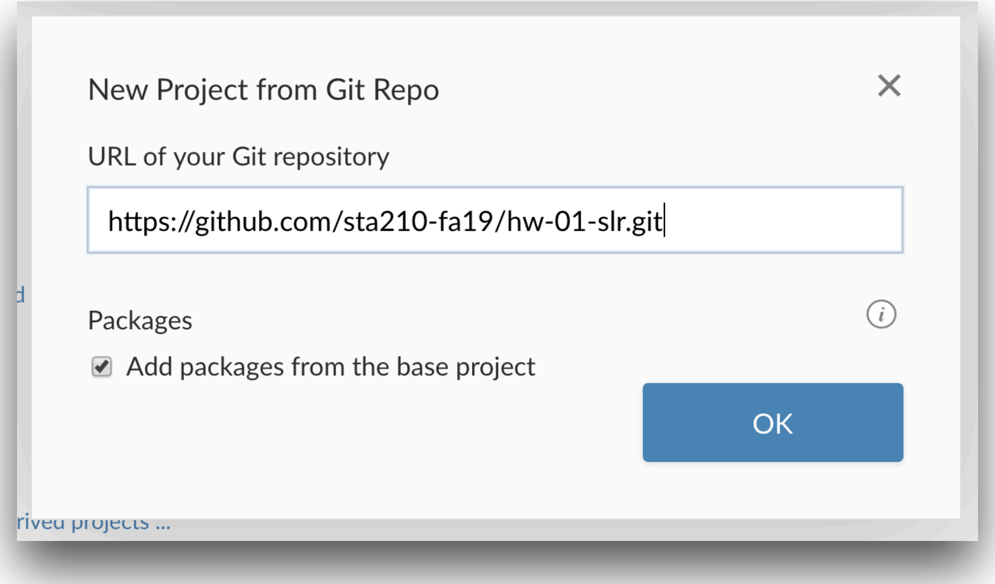
ew file | Upload files | Find File | Clone or download ▾

**Clone with HTTPS** ⓘ Use SSH

Use Git or checkout with SVN using the web URL.

```
https://github.com/sta210-fa19/hw-01-
```

Open in Desktop | Download ZIP



New Project from Git Repo ✕

URL of your Git repository

https://github.com/sta210-fa19/hw-01-slr.git

Packages ⓘ

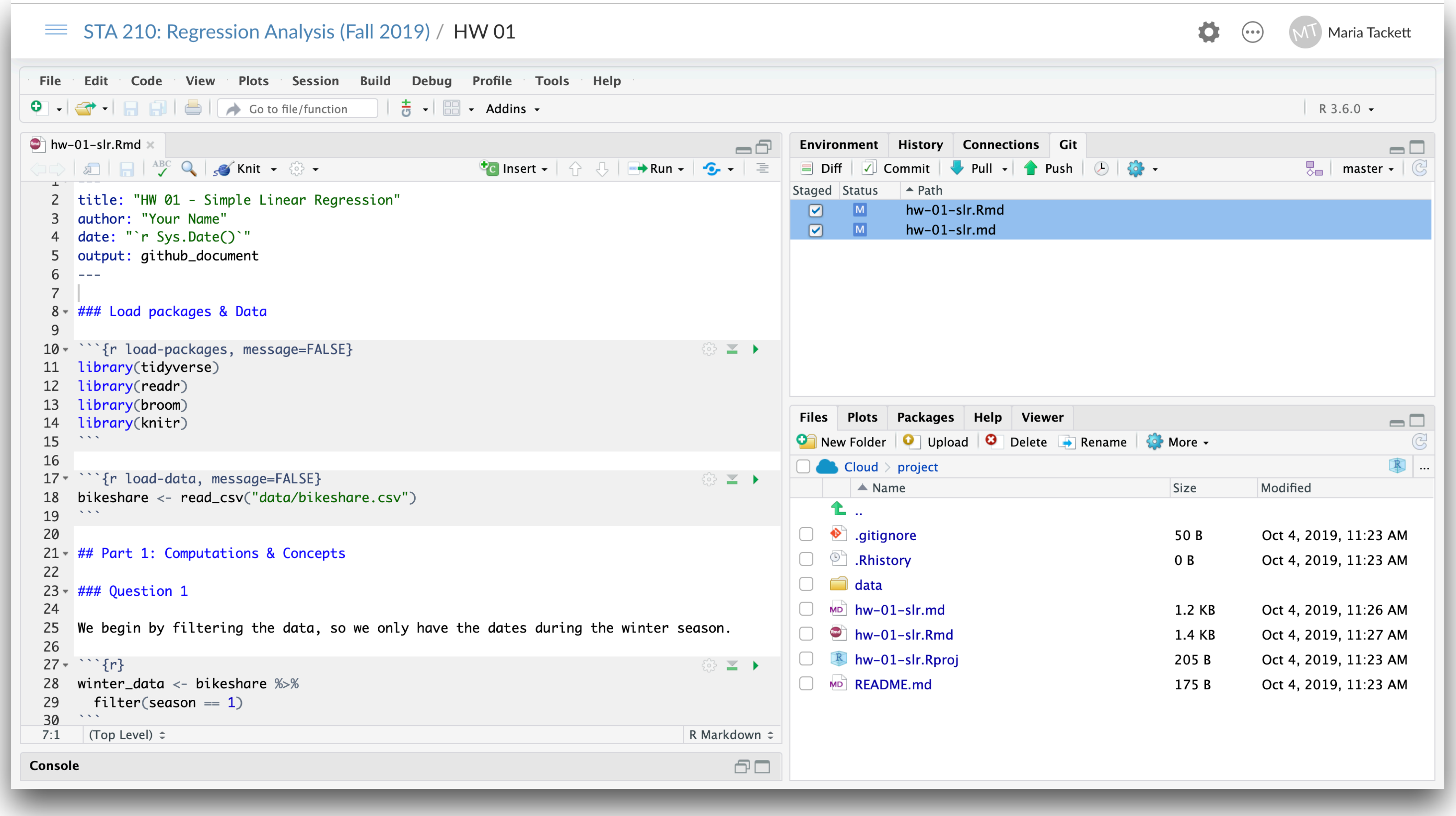
☑ Add packages from the base project

OK

rived projects ...

# Work on Assignment

Students write code and narrative in R Markdown and push work to GitHub periodically as they work on the assignment.



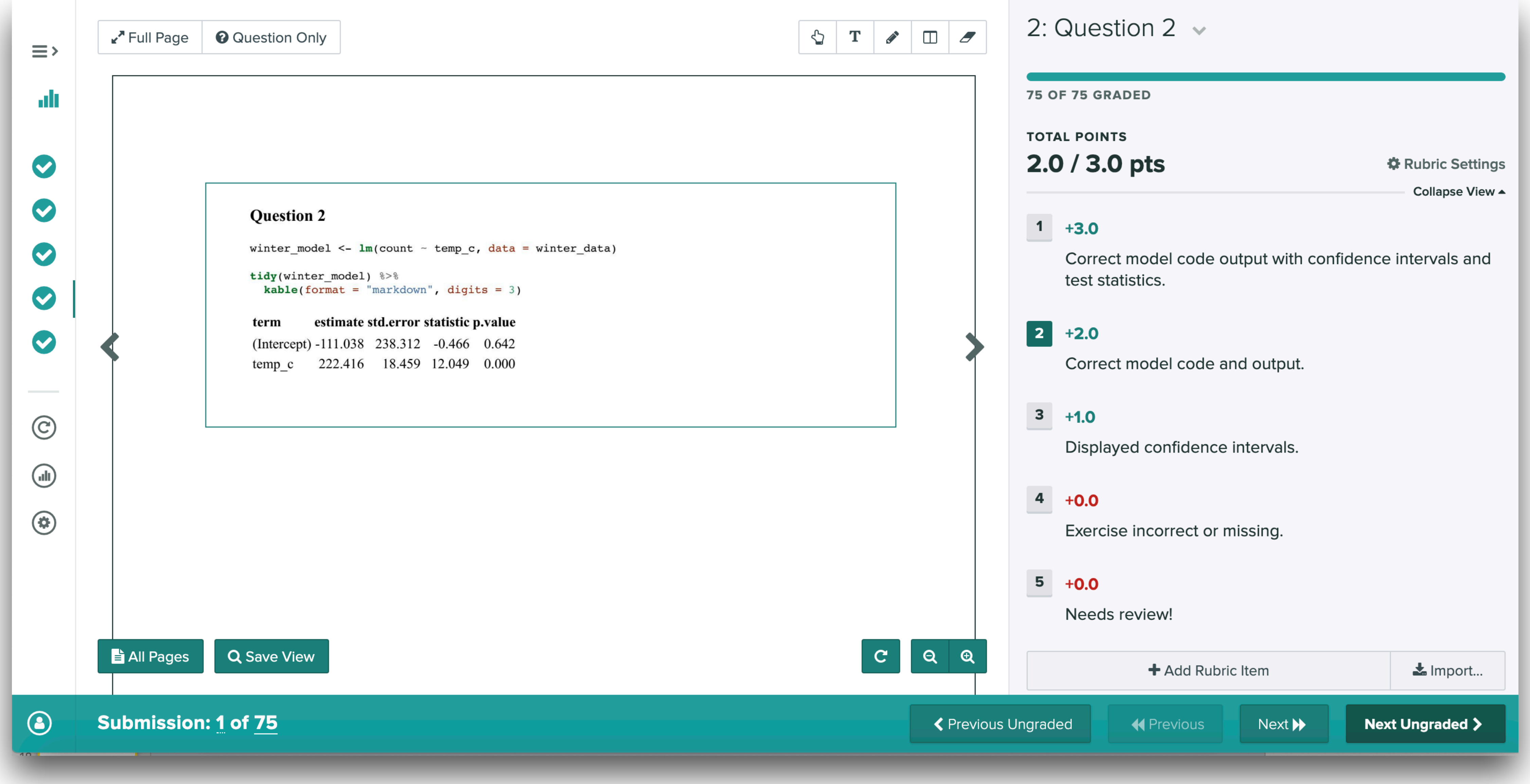
~~~
---
title: "HW 01 - Simple Linear Regression"
author: "Your Name"
date: "`r Sys.Date()`"
output: github_document
---

### Load packages & Data

```{r load-packages, message=FALSE}
library(tidyverse)
library(readr)
library(broom)
library(knitr)
```

```{r load-data, message=FALSE}
bikeshare <- read_csv("data/bikeshare.csv")
```

## Part 1: Computations & Concepts

### Question 1

We begin by filtering the data, so we only have the dates during the winter season.

```{r}
winter_data <- bikeshare %>%
  filter(season == 1)
```
~~~

| Staged | Status | Path |
|---|---|---|
| ☑ | M | hw-01-slr.Rmd |
| ☑ | M | hw-01-slr.md |

| | Name | Size | Modified |
|---|---|---|---|
| | .. | | |
| □ | .gitignore | 50 B | Oct 4, 2019, 11:23 AM |
| □ | .Rhistory | 0 B | Oct 4, 2019, 11:23 AM |
| □ | data | | |
| □ | hw-01-slr.md | 1.2 KB | Oct 4, 2019, 11:26 AM |
| □ | hw-01-slr.Rmd | 1.4 KB | Oct 4, 2019, 11:27 AM |
| □ | hw-01-slr.Rproj | 205 B | Oct 4, 2019, 11:23 AM |
| □ | README.md | 175 B | Oct 4, 2019, 11:23 AM |

# Grading

Clone the student repos and upload to Gradescope for grading.



### Question 2

```r
winter_model <- lm(count ~ temp_c, data = winter_data)

tidy(winter_model) %>%
  kable(format = "markdown", digits = 3)
```

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | -111.038 | 238.312 | -0.466 | 0.642 |
| temp_c | 222.416 | 18.459 | 12.049 | 0.000 |

2: Question 2

75 OF 75 GRADED

TOTAL POINTS

2.0 / 3.0 pts

1 +3.0
Correct model code output with confidence intervals and test statistics.

2 +2.0
Correct model code and output.

3 +1.0
Displayed confidence intervals.

4 +0.0
Exercise incorrect or missing.

5 +0.0
Needs review!

Introduction

Assignments

Challenges
+
Learning
Outcomes

# Challenges

### Different computing backgrounds

- Begin semester with short in-class exercises to expose students to RStudio and GitHub early-on
- Have students work in groups that include diverse computing experiences

### Learning curve for Git/GitHub

- Use Git through Git pane in RStudio
- Only focus on basic Git functions (push, pull, commit)

### Handling large classes

- RStudio Cloud (or other server) to reduce computing differences
- **`ghclass`** package for easy course management

# Learning Outcomes

**Iterative Workflow**

- Students "submit" work frequently by pushing it to GitHub
- Easier for students to incorporate feedback, make changes and include new ideas along the way

**Professional Computing Skills**

- R and GitHub are popular tools in industry
- Students gain computing skills that are applicable for internships, jobs, and higher-level classes

**Professional Collaboration**

- GitHub makes it easier to truly collaborate on group assignments
- Students gain skills for collaboration and sharing work in the workplace

# Thank You!

maria.tackett@duke.edu

@MT_statistics

bit.ly/wsds2019-regression

www.introregression.org

Maria Tackett
Duke University

# References

- Çetinkaya-Rudel, M., & Rundel, C. (2017). Infrastructure and Tools for Teaching Computing Throughout the Statistical Curriculum. *The American Statistician*, 72, 58-65.
- Cloud and GitHub icons from www.flaticon.com
- Course Management with ghclass. Retrieved from https://rundel.github.io/ghclass/
- Curriculum Guidelines for Undergraduate Programs in Statistical Science. Retrieved from https://www.amstat.org/asa/files/pdfs/EDU-guidelines2014-11-15.pdf
- Data Science in a Box. Retrieved from https://datasciencebox.org/