we’ll there, slowly but surely

mine-cetinkaya-rundel
minebocek
✉ mine@stat.duke.edu

mine çetinkaya-rundel
duke university
statistical science
😱

tl;d~~r~~l

3
main
points

1

You only get one first day of class.

Start with something that excites students, teach the necessary evils later.

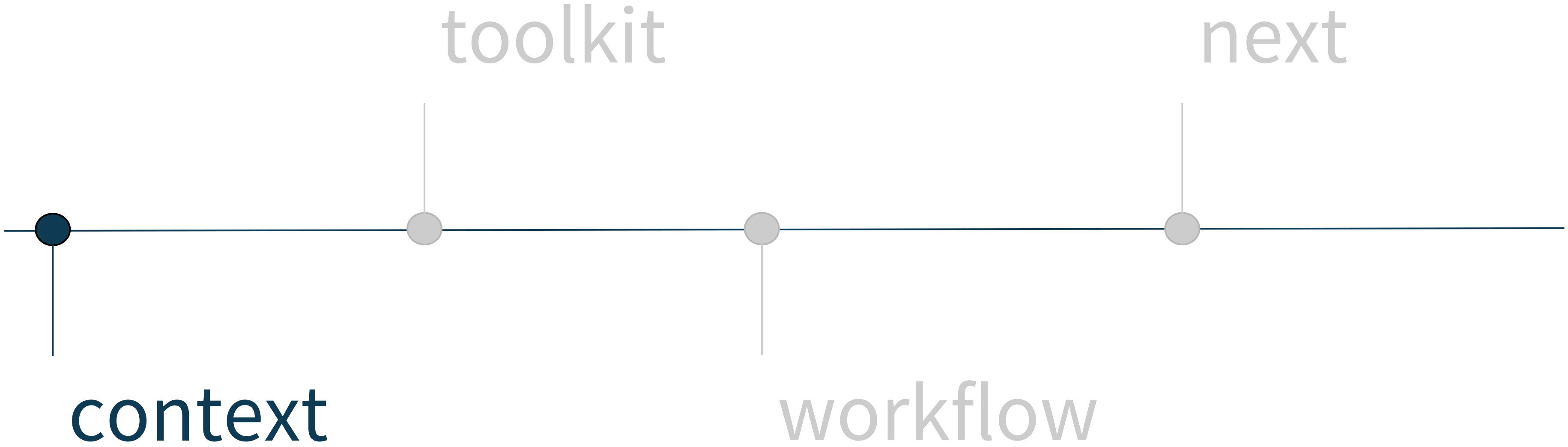
2

First micro-manage, then set free.

If you want students to have well organized repos with well thought out commits, teach best practices early.

3 Git doesn't just happen.

Carve out instructional time, especially for failure prone situations.



context

toolkit

workflow

next

students ready to tackle data head on in a statistical and/or computational context

intro to data science

students with little to no background in computing, data science, or statistics (but enthusiasm to learn!)

first-year undergrad
seminar
18 students

open to all
(targeted at first two years)
80 students

emphasize modern and multivariate EDA + data viz

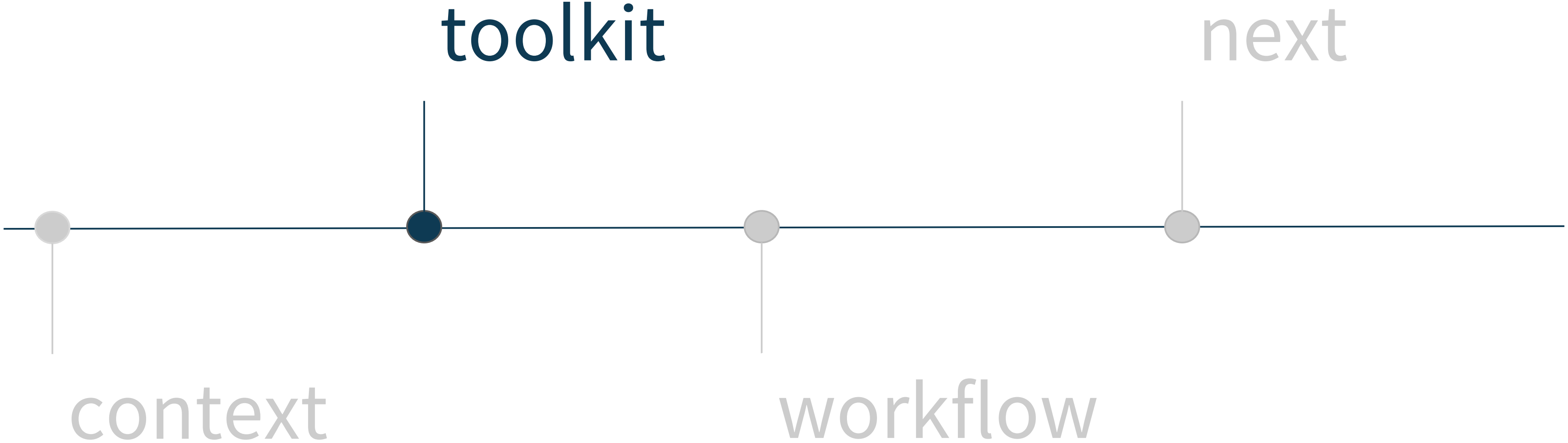
start at the beginning of data analysis cycle with data collection and cleaning

approach statistics from a model based perspective

underscore effective communication of findings

teach (not just expect) reproducible computation

encourage + enforce working collaboratively (think, code, write, present)



context

toolkit

workflow

next

language R

integrated development environment RStudio®

literate programming rmarkdown

version control & collaboration

version control: lots of mistakes along the way, need ability keep track of history (and revert)

collaboration: platform and interface designed to enable collaboration

accountability: transparent commit history

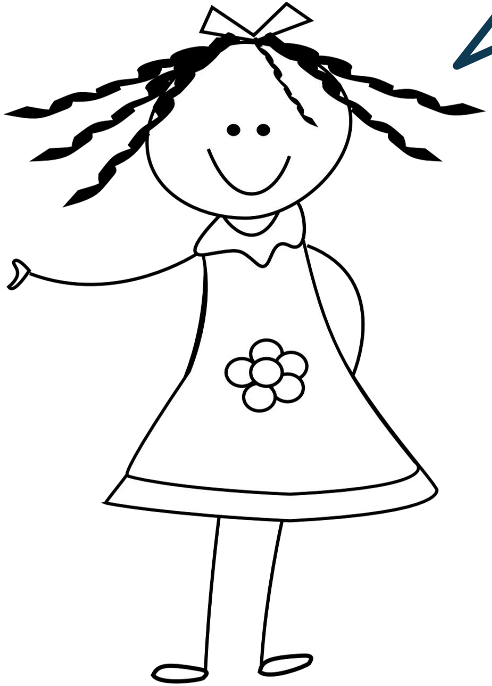
early intro: mastery takes time, start early (day 1), good for marketability + discoverability one organization per course

one repo per student (or team)
per assignment

weekly team assignments,
biweekly individual assignments

1

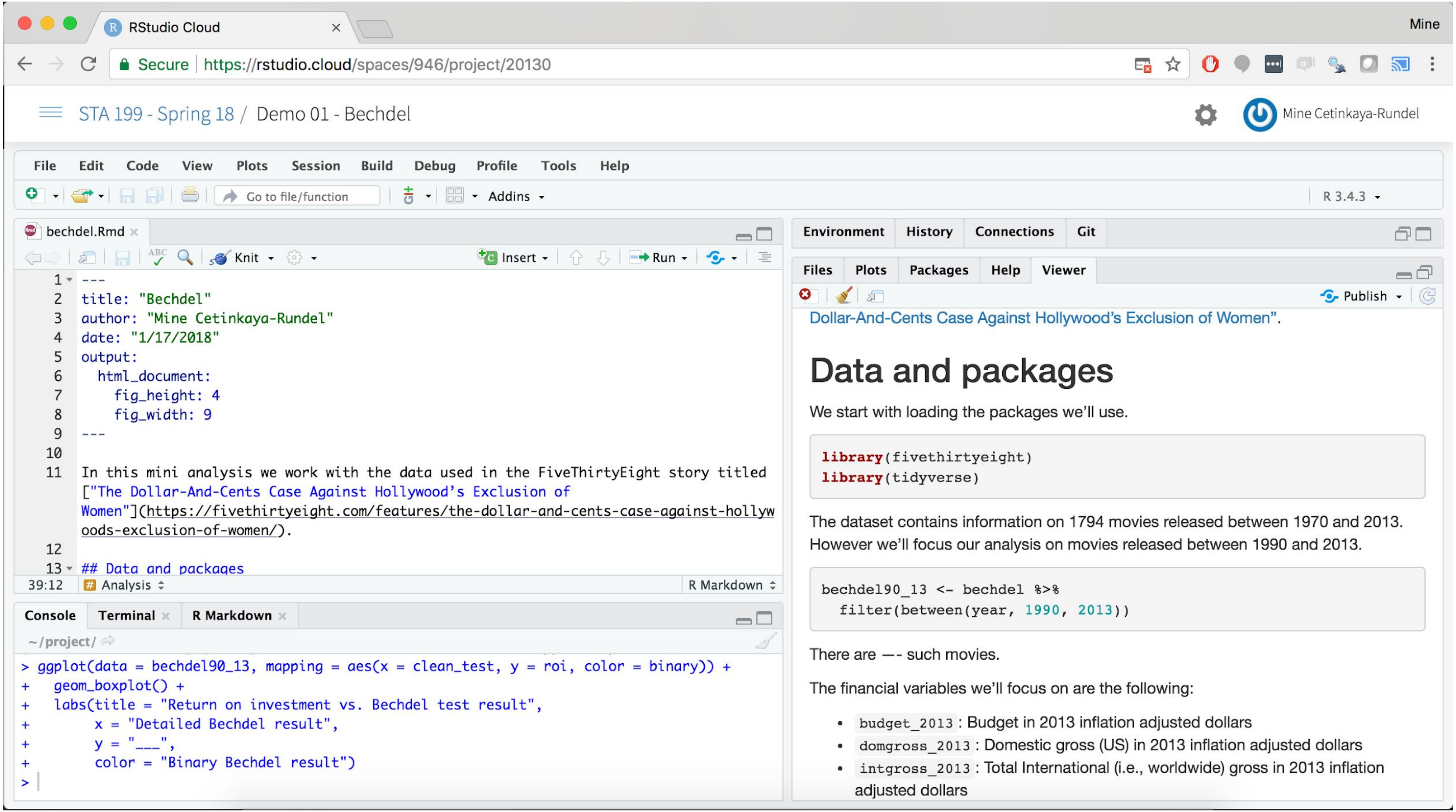
You only get one first day of class.



How do you prefer to spend first day of class?

01 - Install R
02 - Install RStudio
03 - Install git
04 - Install packages

01 - Go to rstudio.cloud
02 - Log in
03 - Create a compelling data visualization



```
---
title: "Bechdel"
author: "Mine Cetinkaya-Rundel"
date: "1/17/2018"
output:
  html_document:
    fig_height: 4
    fig_width: 9
---

In this mini analysis we work with the data used in the FiveThirtyEight story titled
["The Dollar-And-Cents Case Against Hollywood’s Exclusion of
Women"](https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollyw
oods-exclusion-of-women/).

## Data and packages
```

```
> ggplot(data = bechdel90_13, mapping = aes(x = clean_test, y = roi, color = binary)) +
+   geom_boxplot() +
+   labs(title = "Return on investment vs. Bechdel test result",
+        x = "Detailed Bechdel result",
+        y = "___",
+        color = "Binary Bechdel result")
>
```

Dollar-And-Cents Case Against Hollywood’s Exclusion of Women”.

Data and packages

We start with loading the packages we’ll use.

```
library(fivethirtyeight)
library(tidyverse)
```

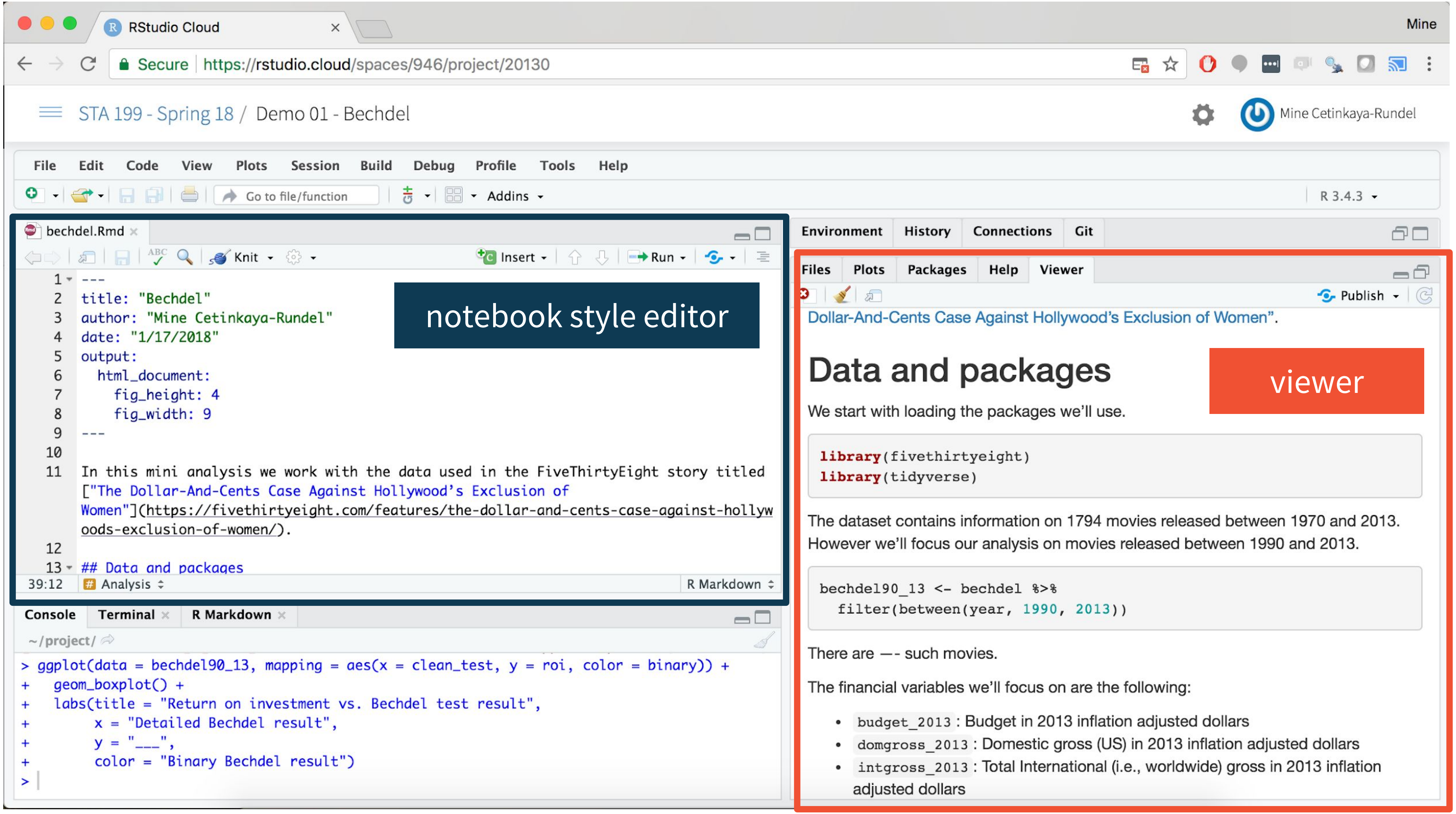
The dataset contains information on 1794 movies released between 1970 and 2013. However we’ll focus our analysis on movies released between 1990 and 2013.

```
bechdel90_13 <- bechdel %>%
  filter(between(year, 1990, 2013))
```

There are —- such movies.

The financial variables we’ll focus on are the following:

- `budget_2013`: Budget in 2013 inflation adjusted dollars
- `domgross_2013`: Domestic gross (US) in 2013 inflation adjusted dollars
- `intgross_2013`: Total International (i.e., worldwide) gross in 2013 inflation adjusted dollars



RStudio Cloud
Secure https://rstudio.cloud/spaces/946/project/20130
Mine
STA 199 - Spring 18 / Demo 01 - Bechdel
Mine Cetinkaya-Rundel
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 3.4.3
bechdel.Rmd
Knit
Insert
Run
Environment History Connections Git
Files Plots Packages Help Viewer
Publish
39:12 Analysis
R Markdown
Console Terminal R Markdown
~/project/

notebook style editor

```
---
title: "Bechdel"
author: "Mine Cetinkaya-Rundel"
date: "1/17/2018"
output:
  html_document:
    fig_height: 4
    fig_width: 9
---

In this mini analysis we work with the data used in the FiveThirtyEight story titled ["The Dollar-And-Cents Case Against Hollywood's Exclusion of Women"](https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/).

## Data and packages
```

```
> ggplot(data = bechdel90_13, mapping = aes(x = clean_test, y = roi, color = binary)) +
+   geom_boxplot() +
+   labs(title = "Return on investment vs. Bechdel test result",
+        x = "Detailed Bechdel result",
+        y = "___",
+        color = "Binary Bechdel result")
>
```

viewer

Dollar-And-Cents Case Against Hollywood's Exclusion of Women".

Data and packages

We start with loading the packages we'll use.

```
library(fivethirtyeight)
library(tidyverse)
```

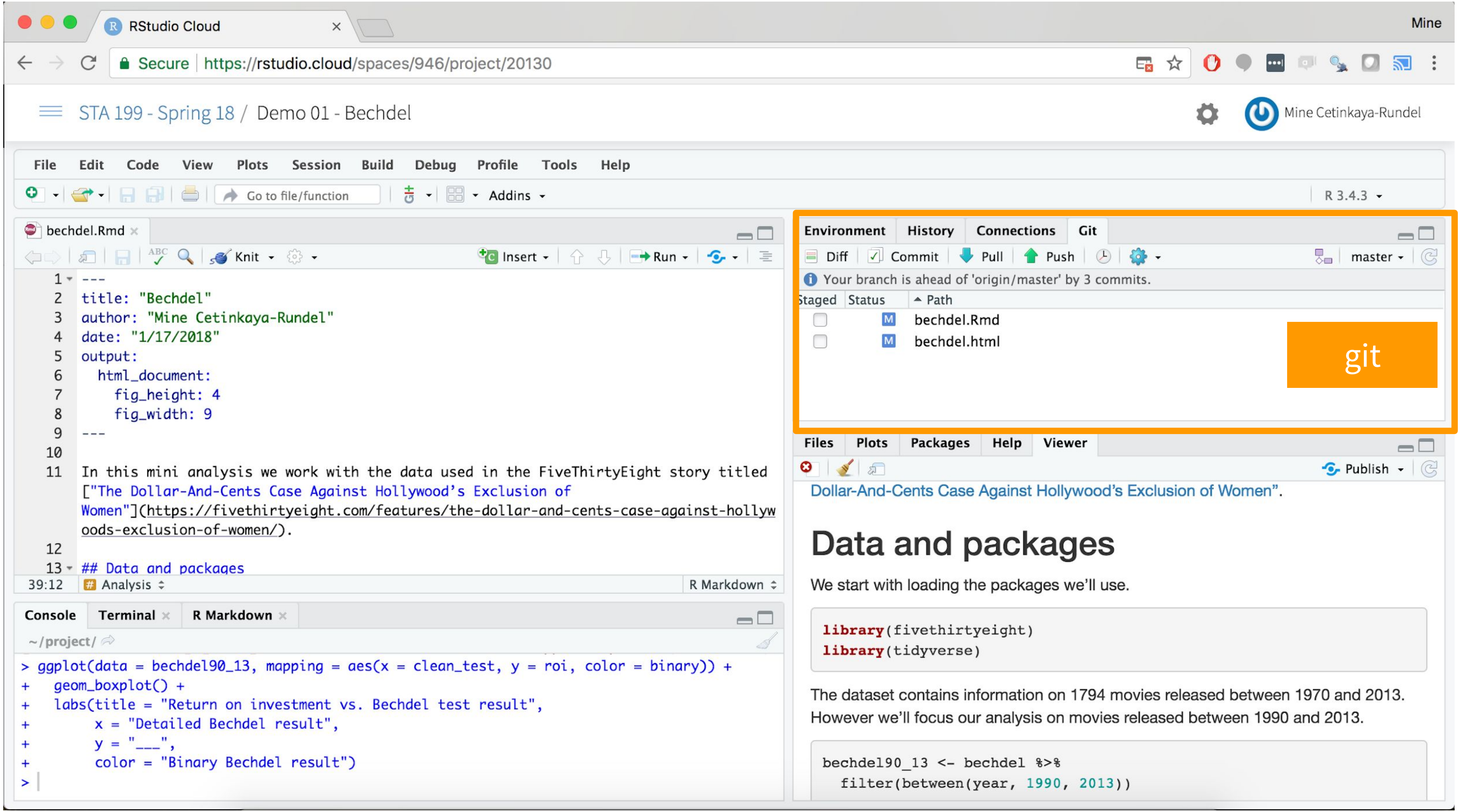
The dataset contains information on 1794 movies released between 1970 and 2013. However we'll focus our analysis on movies released between 1990 and 2013.

```
bechdel90_13 <- bechdel %>%
  filter(between(year, 1990, 2013))
```

There are –- such movies.

The financial variables we'll focus on are the following:

- `budget_2013`: Budget in 2013 inflation adjusted dollars
- `domgross_2013`: Domestic gross (US) in 2013 inflation adjusted dollars
- `intgross_2013`: Total International (i.e., worldwide) gross in 2013 inflation adjusted dollars



```
---
title: "Bechdel"
author: "Mine Cetinkaya-Rundel"
date: "1/17/2018"
output:
  html_document:
    fig_height: 4
    fig_width: 9
---

In this mini analysis we work with the data used in the FiveThirtyEight story titled ["The Dollar-And-Cents Case Against Hollywood’s Exclusion of Women"](https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/).

## Data and packages
```

Your branch is ahead of 'origin/master' by 3 commits.

Staged	Status	Path
□	M	bechdel.Rmd
□	M	bechdel.html

git

```
> ggplot(data = bechdel90_13, mapping = aes(x = clean_test, y = roi, color = binary)) +
+   geom_boxplot() +
+   labs(title = "Return on investment vs. Bechdel test result",
+        x = "Detailed Bechdel result",
+        y = "___",
+        color = "Binary Bechdel result")
>
```

Dollar-And-Cents Case Against Hollywood’s Exclusion of Women”.

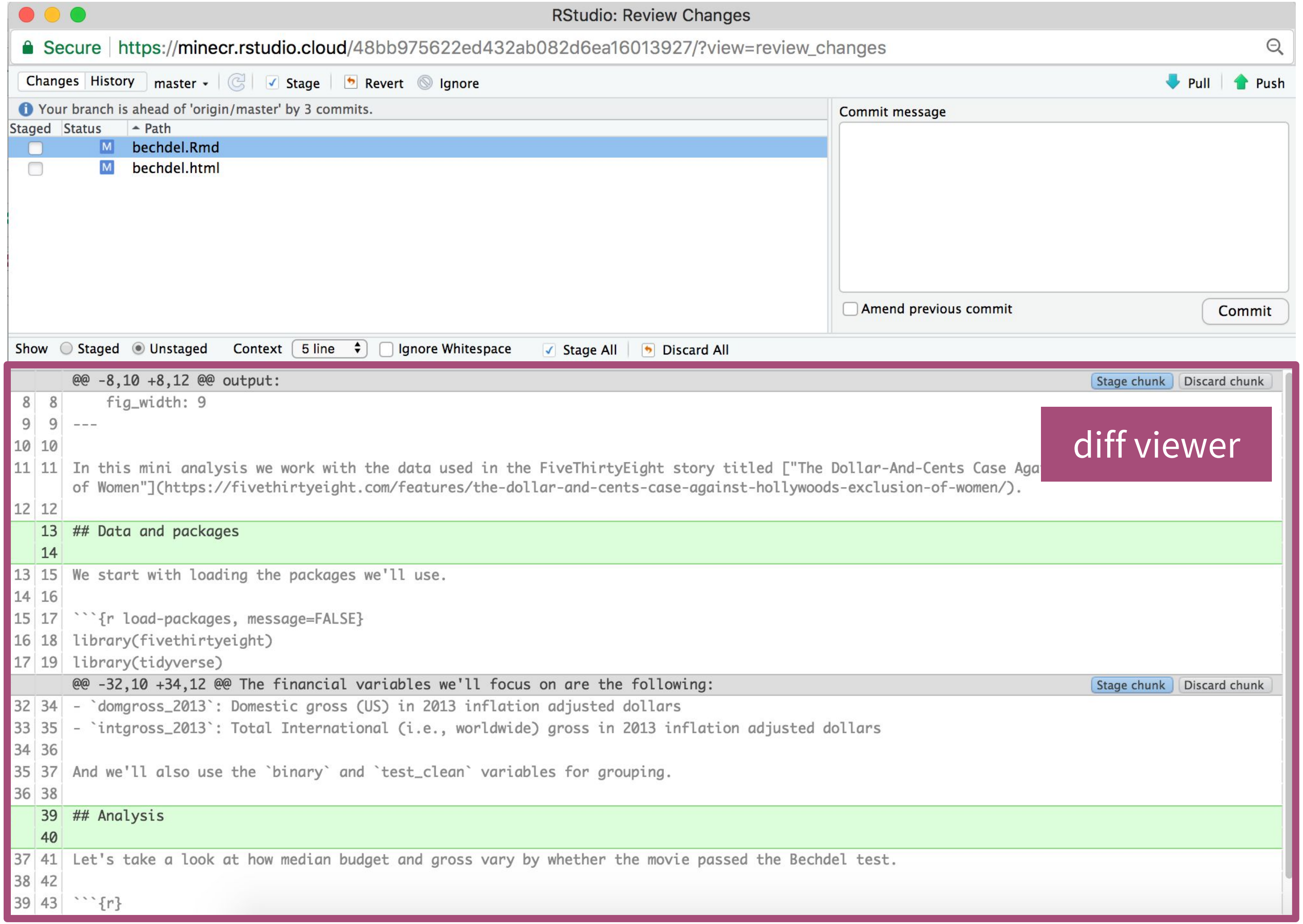
Data and packages

We start with loading the packages we’ll use.

```
library(fivethirtyeight)
library(tidyverse)
```

The dataset contains information on 1794 movies released between 1970 and 2013. However we’ll focus our analysis on movies released between 1990 and 2013.

```
bechdel90_13 <- bechdel %>%
  filter(between(year, 1990, 2013))
```

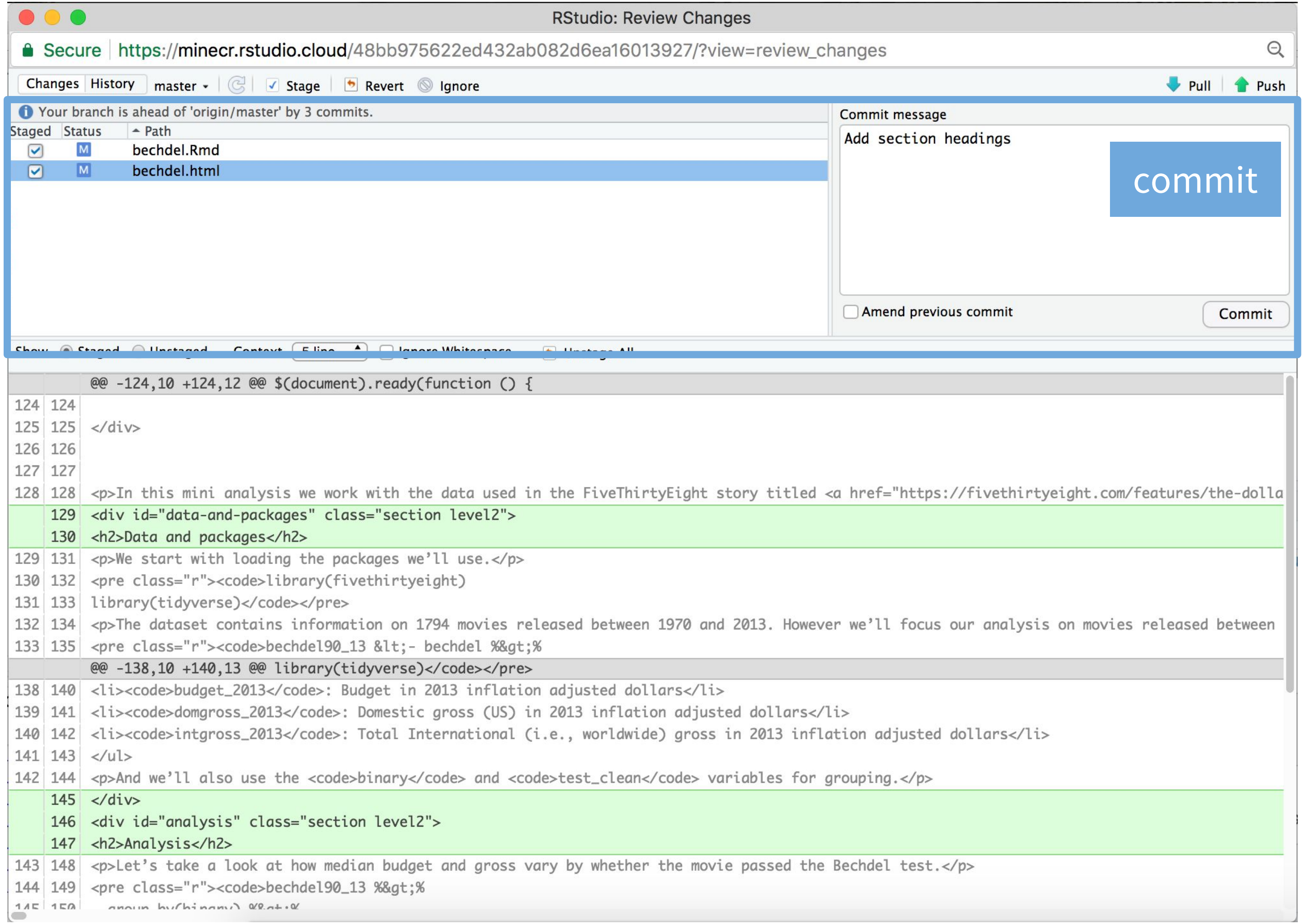


Your branch is ahead of 'origin/master' by 3 commits.

Staged	Status	Path
☐	M	bechdel.Rmd
☐	M	bechdel.html

diff viewer

~~~
@@ -8,10 +8,12 @@ output:
    fig_width: 9
---

In this mini analysis we work with the data used in the FiveThirtyEight story titled ["The Dollar-And-Cents Case Aga
of Women"](https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/).

## Data and packages

We start with loading the packages we'll use.

```{r load-packages, message=FALSE}
library(fivethirtyeight)
library(tidyverse)
@@ -32,10 +34,12 @@ The financial variables we'll focus on are the following:
- `domgross_2013`: Domestic gross (US) in 2013 inflation adjusted dollars
- `intgross_2013`: Total International (i.e., worldwide) gross in 2013 inflation adjusted dollars

And we'll also use the `binary` and `test_clean` variables for grouping.

## Analysis

Let's take a look at how median budget and gross vary by whether the movie passed the Bechdel test.

```{r}
~~~



commit

Your branch is ahead of 'origin/master' by 3 commits.

Staged	Status	Path
☑	M	bechdel.Rmd
☑	M	bechdel.html

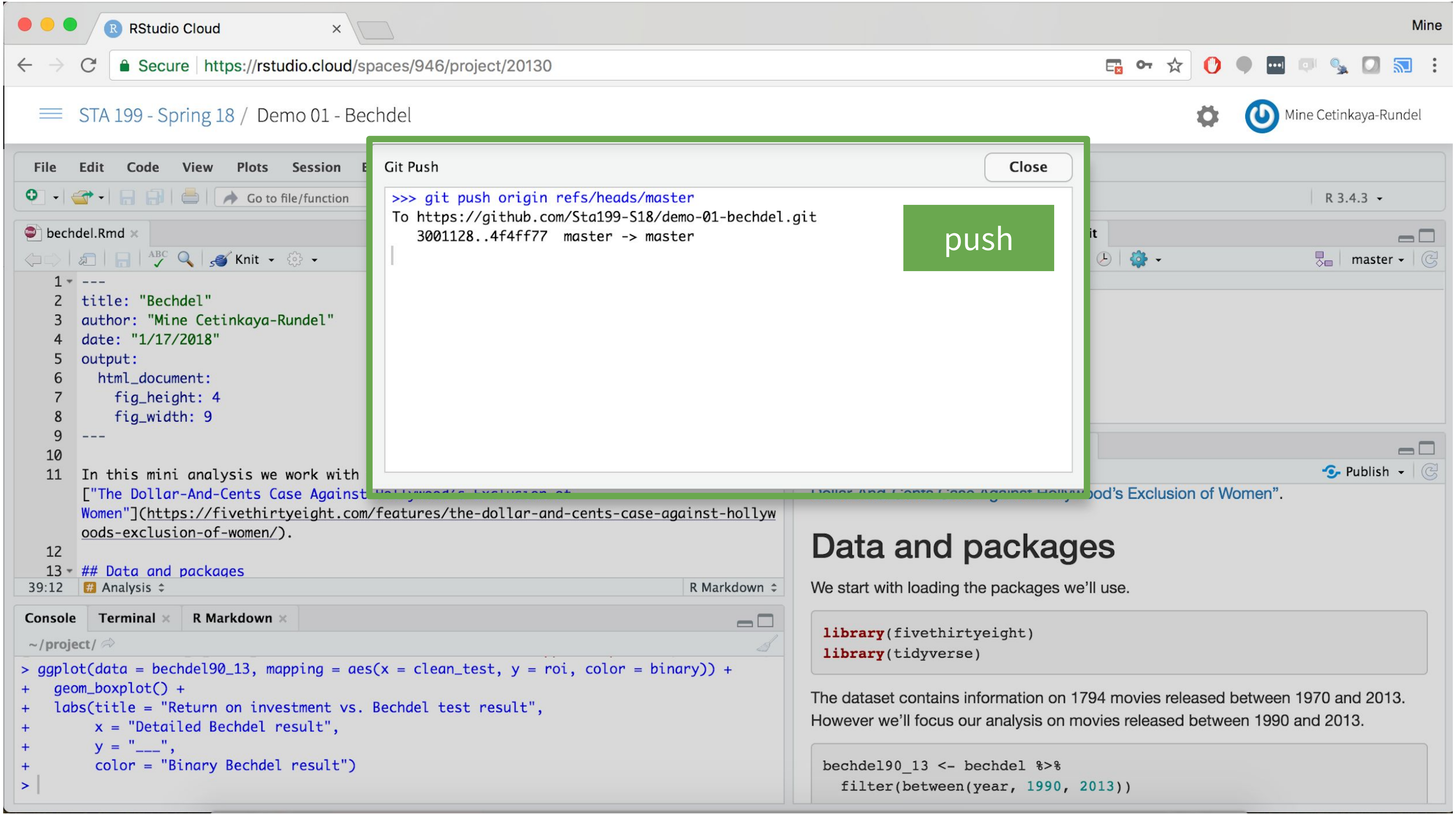
Commit message

Add section headings

□ Amend previous commit

Commit

```
@@ -124,10 +124,12 @@ $(document).ready(function () {

</div>


<p>In this mini analysis we work with the data used in the FiveThirtyEight story titled <a href="https://fivethirtyeight.com/features/the-dolla
<div id="data-and-packages" class="section level2">
<h2>Data and packages</h2>
<p>We start with loading the packages we’ll use.</p>
<pre class="r"><code>library(fivethirtyeight)
library(tidyverse)</code></pre>
<p>The dataset contains information on 1794 movies released between 1970 and 2013. However we’ll focus our analysis on movies released between
<pre class="r"><code>bechdel90_13 &lt;- bechdel %&gt;%
@@ -138,10 +140,13 @@ library(tidyverse)</code></pre>
<li><code>budget_2013</code>: Budget in 2013 inflation adjusted dollars</li>
<li><code>domgross_2013</code>: Domestic gross (US) in 2013 inflation adjusted dollars</li>
<li><code>intgross_2013</code>: Total International (i.e., worldwide) gross in 2013 inflation adjusted dollars</li>
</ul>
<p>And we’ll also use the <code>binary</code> and <code>test_clean</code> variables for grouping.</p>
</div>
<div id="analysis" class="section level2">
<h2>Analysis</h2>
<p>Let’s take a look at how median budget and gross vary by whether the movie passed the Bechdel test.</p>
<pre class="r"><code>bechdel90_13 %&gt;%
  group_by(binary) %&gt;%
```


RStudio Cloud
Mine
Secure https://rstudio.cloud/spaces/946/project/20130
STA 199 - Spring 18 / Demo 01 - Bechdel
Mine Cetinkaya-Rundel
File Edit Code View Plots Session
Go to file/function
R 3.4.3
bechdel.Rmd
Knit
master
Publish
39:12 Analysis
R Markdown
Console Terminal R Markdown
~/project/
Git Push
Close

```
>>> git push origin refs/heads/master
To https://github.com/Sta199-S18/demo-01-bechdel.git
   3001128..4f4ff77  master -> master
```

push

```
---
title: "Bechdel"
author: "Mine Cetinkaya-Rundel"
date: "1/17/2018"
output:
  html_document:
    fig_height: 4
    fig_width: 9
---

In this mini analysis we work with
["The Dollar-And-Cents Case Against
Women"](https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollyw
oods-exclusion-of-women/).

## Data and packages
```

```
> ggplot(data = bechdel90_13, mapping = aes(x = clean_test, y = roi, color = binary)) +
+   geom_boxplot() +
+   labs(title = "Return on investment vs. Bechdel test result",
+        x = "Detailed Bechdel result",
+        y = "___",
+        color = "Binary Bechdel result")
>
```

ood's Exclusion of Women".

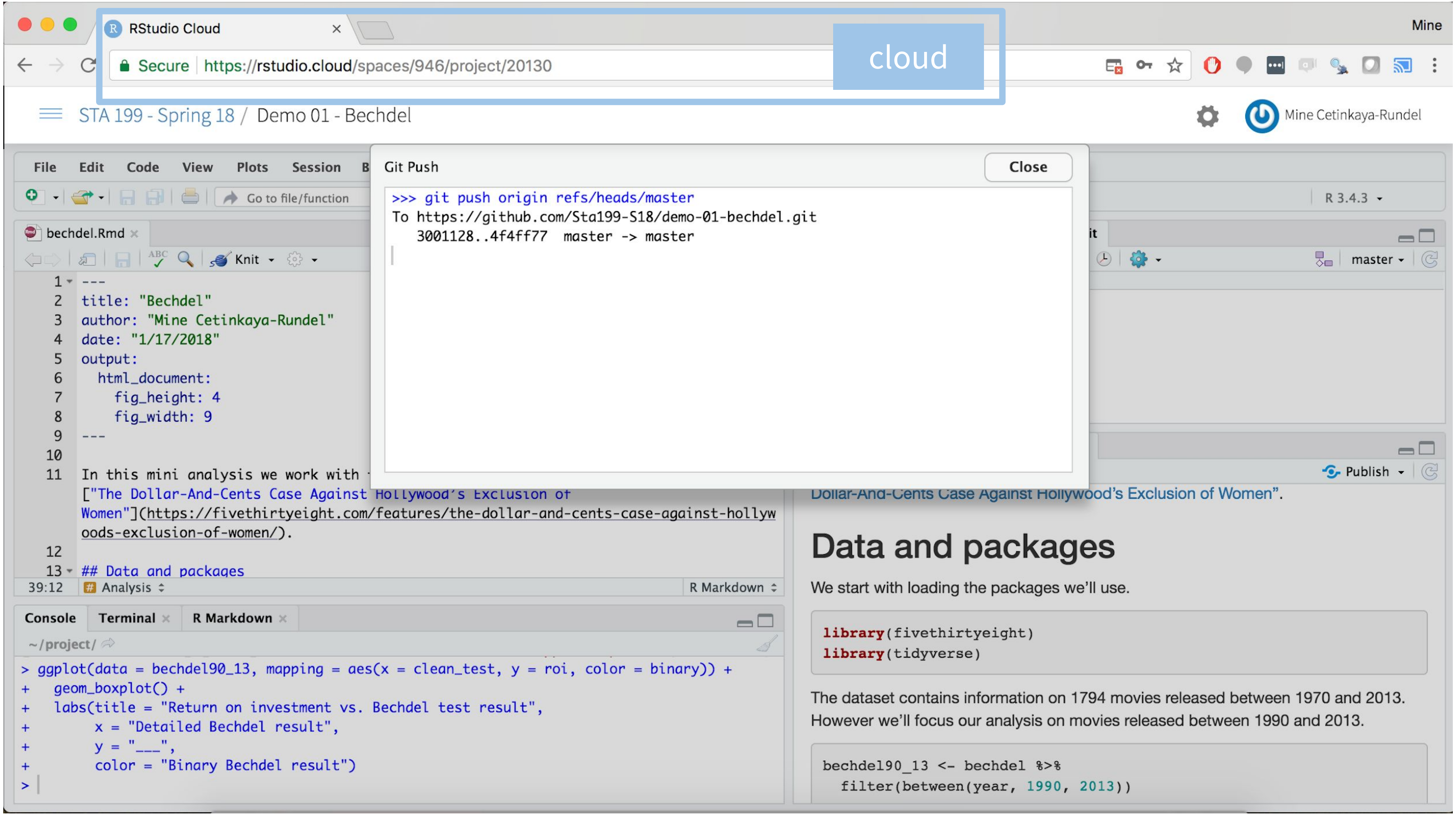
Data and packages

We start with loading the packages we'll use.

```
library(fivethirtyeight)
library(tidyverse)
```

The dataset contains information on 1794 movies released between 1970 and 2013. However we'll focus our analysis on movies released between 1990 and 2013.

```
bechdel90_13 <- bechdel %>%
  filter(between(year, 1990, 2013))
```



```
>>> git push origin refs/heads/master
To https://github.com/Sta199-S18/demo-01-bechdel.git
   3001128..4f4ff77  master -> master
```

```
---
title: "Bechdel"
author: "Mine Cetinkaya-Rundel"
date: "1/17/2018"
output:
  html_document:
    fig_height: 4
    fig_width: 9
---

In this mini analysis we work with
["The Dollar-And-Cents Case Against Hollywood's Exclusion of
Women"](https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollyw
oods-exclusion-of-women/).

## Data and packages
```

```
> ggplot(data = bechdel90_13, mapping = aes(x = clean_test, y = roi, color = binary)) +
+   geom_boxplot() +
+   labs(title = "Return on investment vs. Bechdel test result",
+        x = "Detailed Bechdel result",
+        y = "___",
+        color = "Binary Bechdel result")
>
```

Dollar-And-Cents Case Against Hollywood's Exclusion of Women".

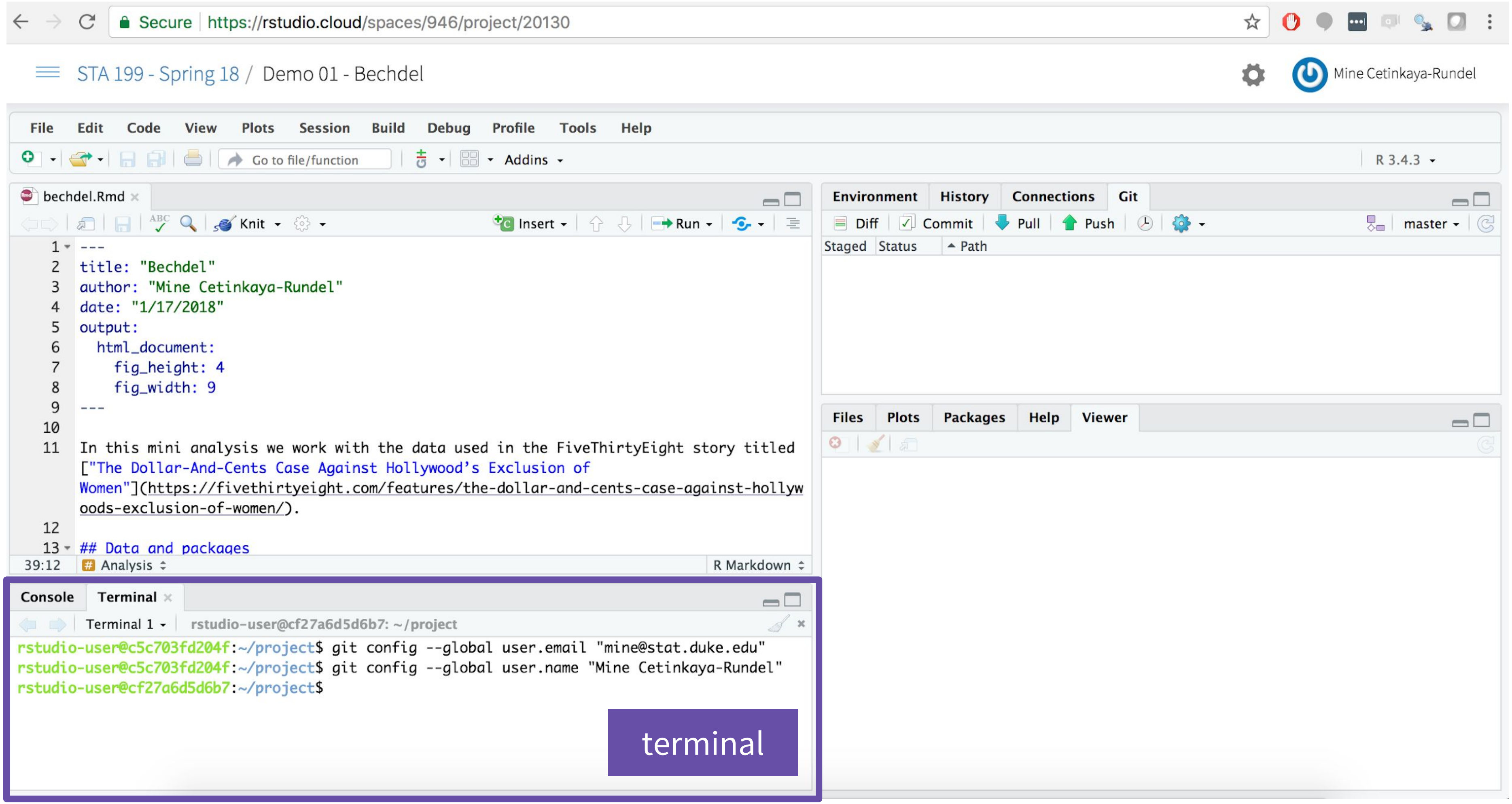
Data and packages

We start with loading the packages we'll use.

```
library(fivethirtyeight)
library(tidyverse)
```

The dataset contains information on 1794 movies released between 1970 and 2013. However we'll focus our analysis on movies released between 1990 and 2013.

```
bechdel90_13 <- bechdel %>%
  filter(between(year, 1990, 2013))
```



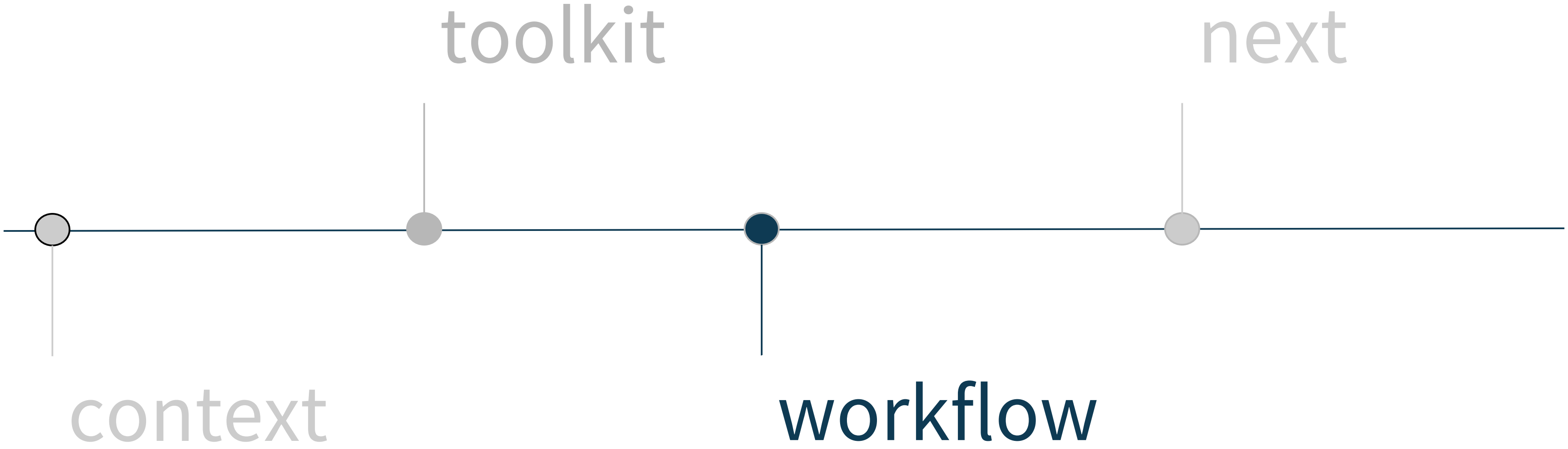
```
---
title: "Bechdel"
author: "Mine Cetinkaya-Rundel"
date: "1/17/2018"
output:
  html_document:
    fig_height: 4
    fig_width: 9
---

In this mini analysis we work with the data used in the FiveThirtyEight story titled ["The Dollar-And-Cents Case Against Hollywood’s Exclusion of Women"](https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/).

## Data and packages
```

```
rstudio-user@c5c703fd204f:~/project$ git config --global user.email "mine@stat.duke.edu"
rstudio-user@c5c703fd204f:~/project$ git config --global user.name "Mine Cetinkaya-Rundel"
rstudio-user@cf27a6d5d6b7:~/project$
```

terminal

- Having the same setup facilitates peers helping each other, especially early on.

- Hearing others articulate questions around infrastructure helps students better articulate their own questions.



context

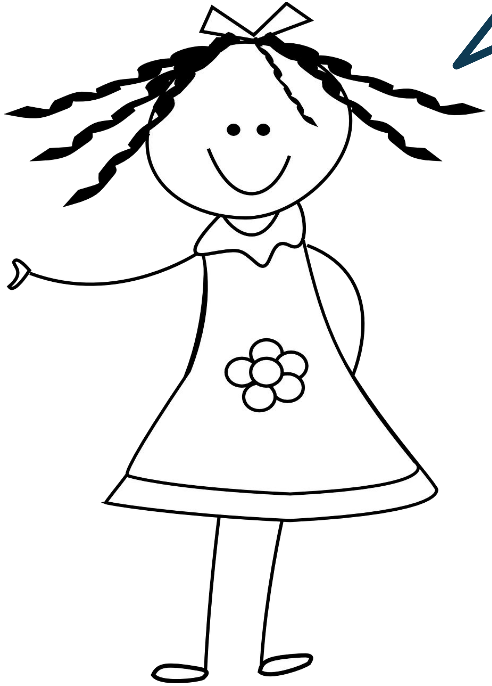
toolkit

workflow

next

2

First micromanage, then set free.



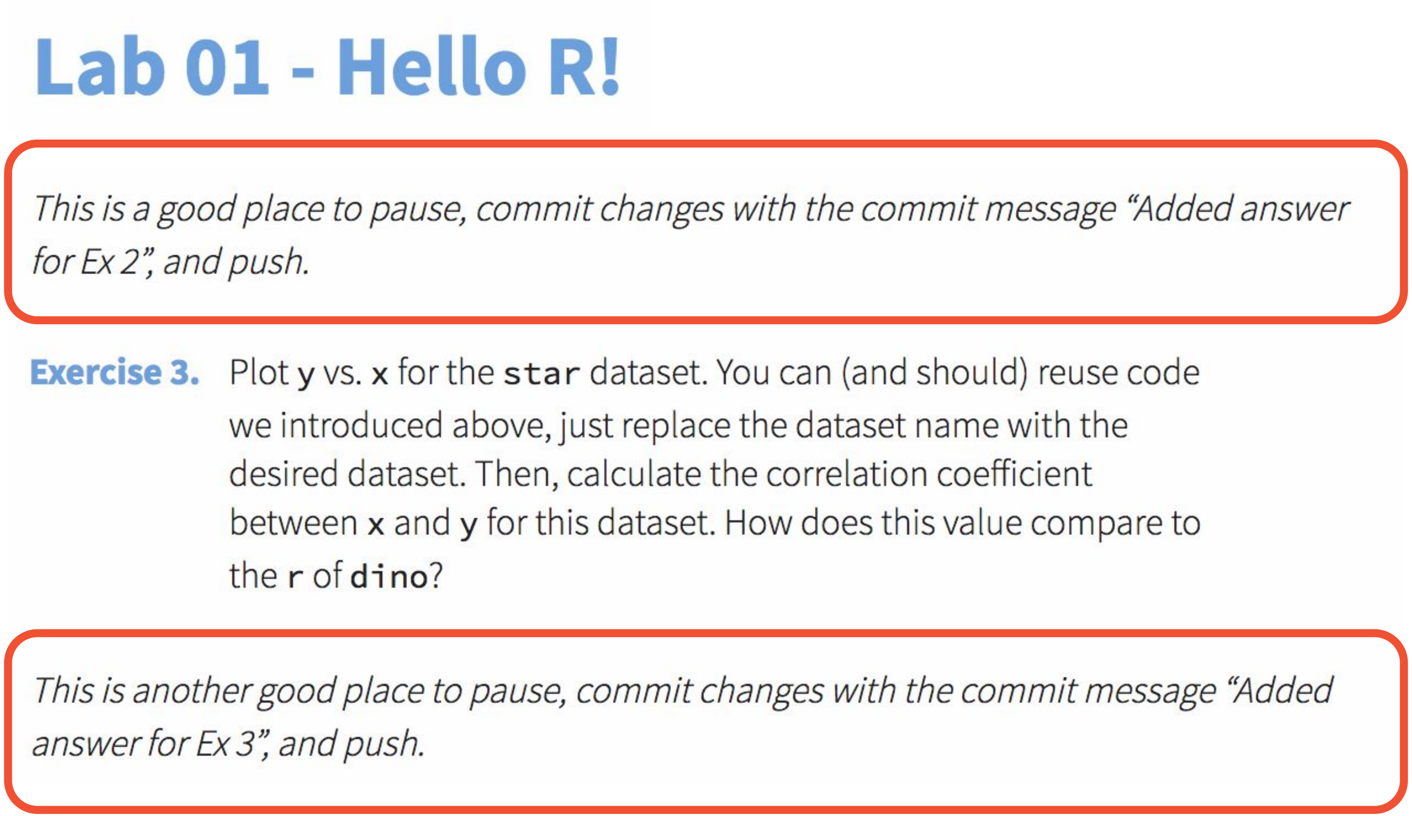
How do you prefer to spend your time?

teaching

grading

managing team conflicts

First micro-manage...



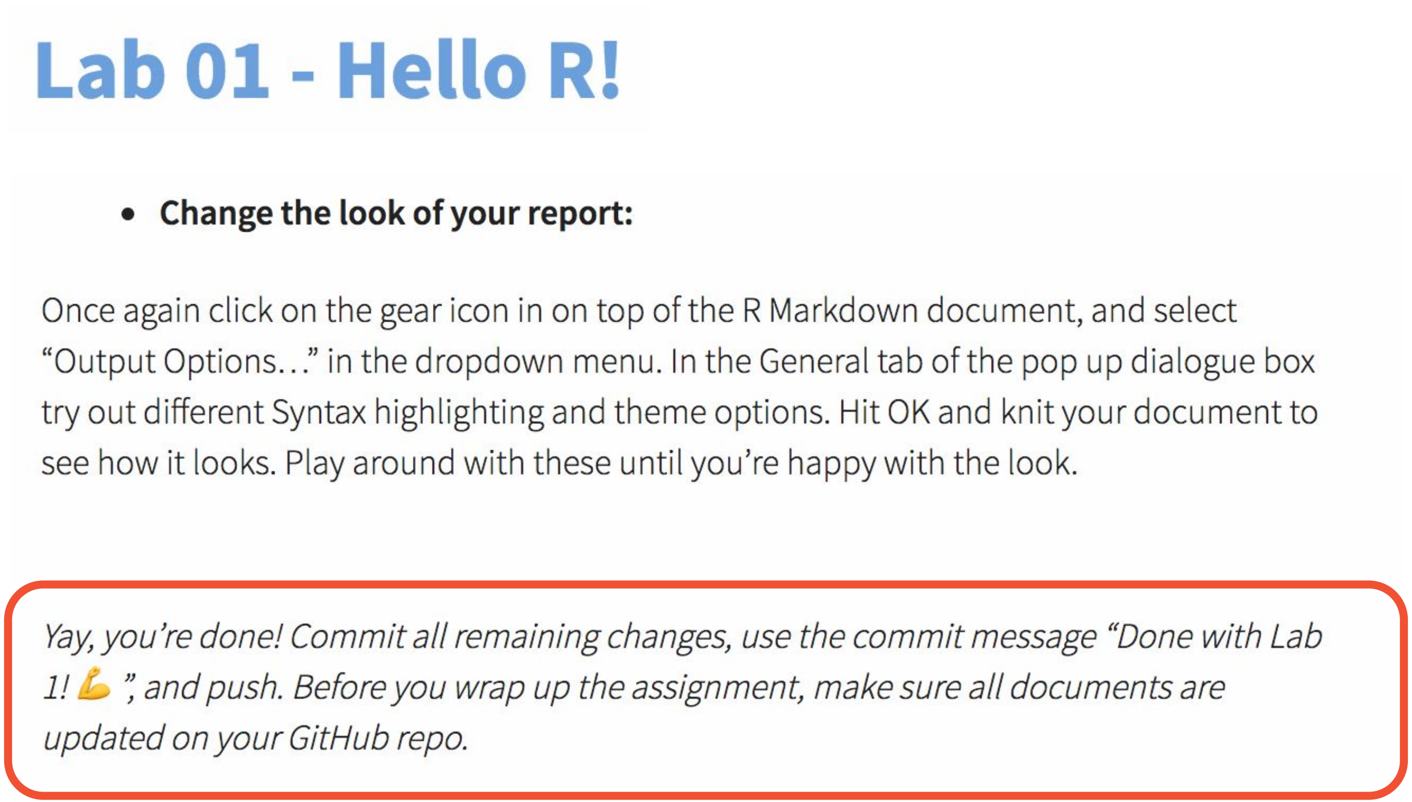
Lab 01 - Hello R!

This is a good place to pause, commit changes with the commit message "Added answer for Ex 2", and push.

Exercise 3. Plot `y` vs. `x` for the `star` dataset. You can (and should) reuse code we introduced above, just replace the dataset name with the desired dataset. Then, calculate the correlation coefficient between `x` and `y` for this dataset. How does this value compare to the `r` of `dino`?

This is another good place to pause, commit changes with the commit message "Added answer for Ex 3", and push.

First micro-manage...



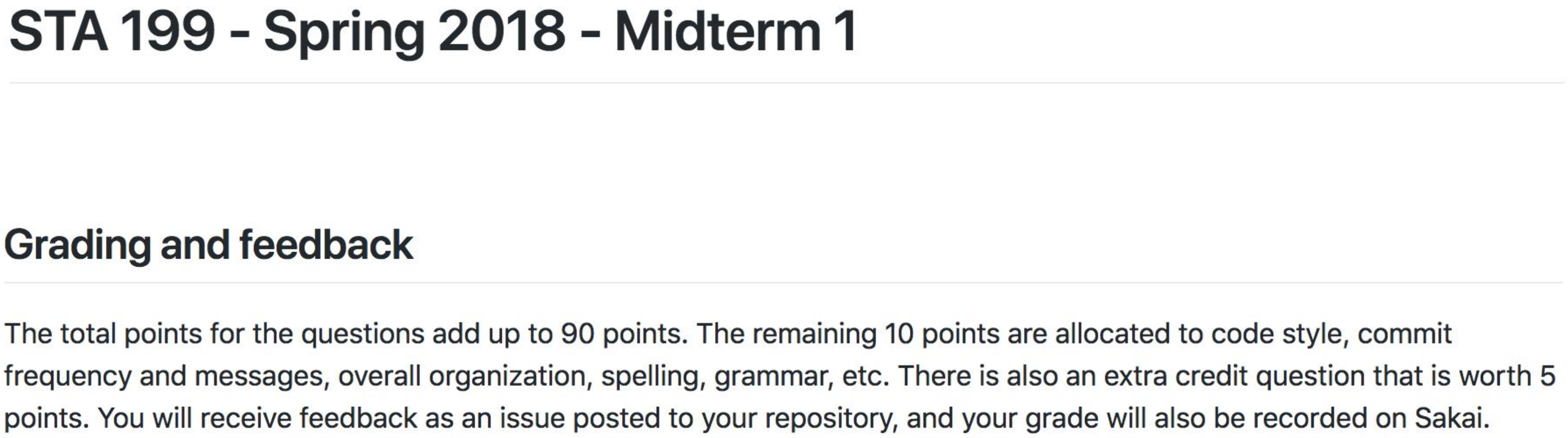
Lab 01 - Hello R!

- **Change the look of your report:**

Once again click on the gear icon in on top of the R Markdown document, and select “Output Options…” in the dropdown menu. In the General tab of the pop up dialogue box try out different Syntax highlighting and theme options. Hit OK and knit your document to see how it looks. Play around with these until you’re happy with the look.

Yay, you’re done! Commit all remaining changes, use the commit message “Done with Lab 1! 💪 ”, and push. Before you wrap up the assignment, make sure all documents are updated on your GitHub repo.

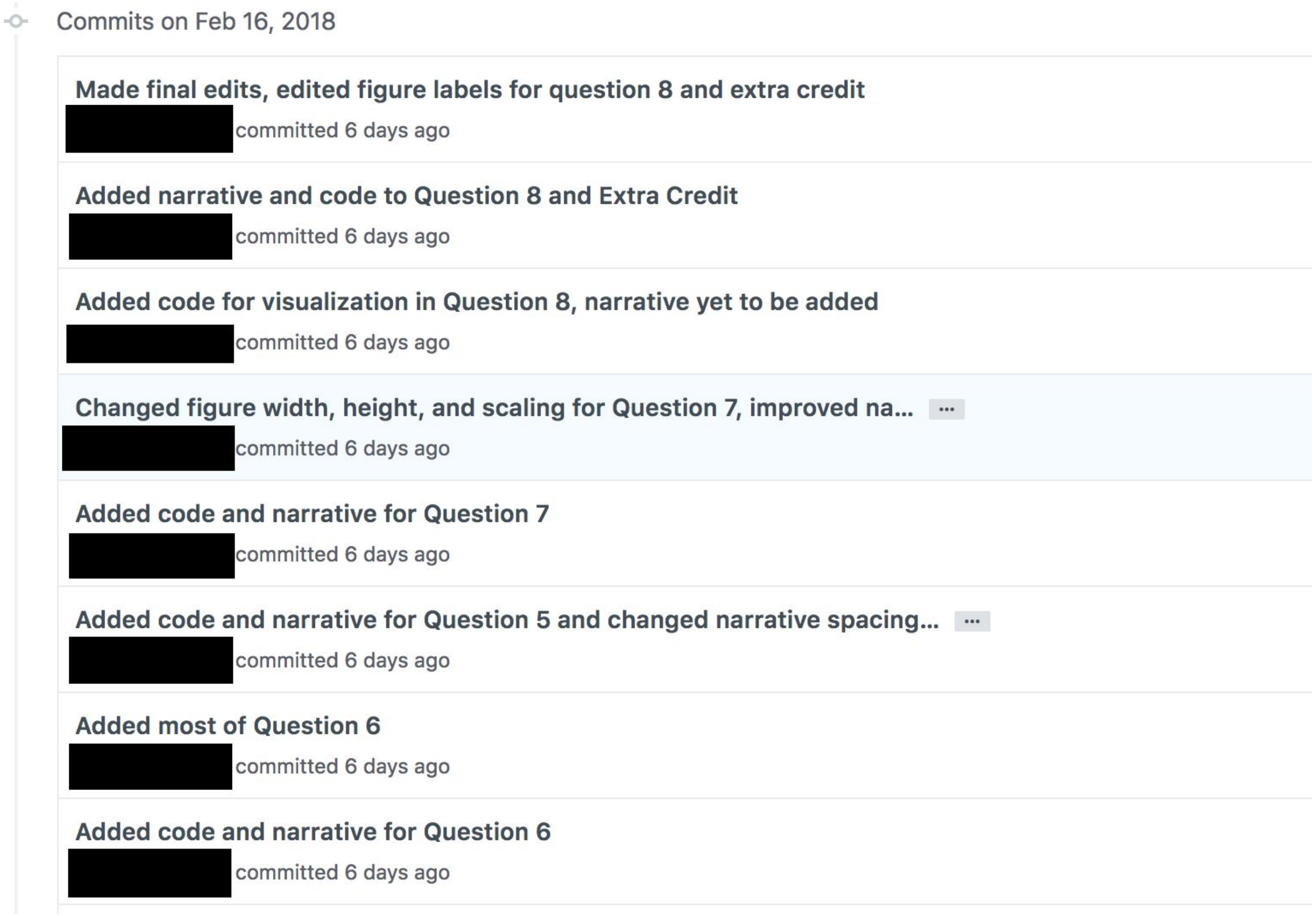
... then set free



STA 199 - Spring 2018 - Midterm 1

Grading and feedback

The total points for the questions add up to 90 points. The remaining 10 points are allocated to code style, commit frequency and messages, overall organization, spelling, grammar, etc. There is also an extra credit question that is worth 5 points. You will receive feedback as an issue posted to your repository, and your grade will also be recorded on Sakai.



Commits on Feb 16, 2018

Made final edits, edited figure labels for question 8 and extra credit
[redacted] committed 6 days ago

Added narrative and code to Question 8 and Extra Credit
[redacted] committed 6 days ago

Added code for visualization in Question 8, narrative yet to be added
[redacted] committed 6 days ago

Changed figure width, height, and scaling for Question 7, improved na...
[redacted] committed 6 days ago

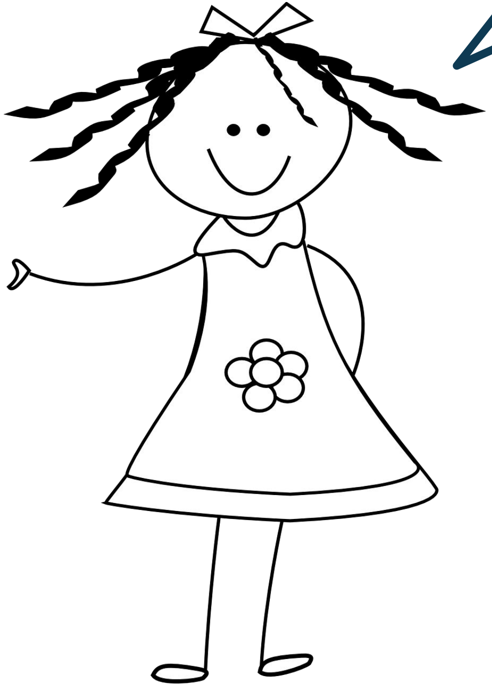
Added code and narrative for Question 7
[redacted] committed 6 days ago

Added code and narrative for Question 5 and changed narrative spacing...
[redacted] committed 6 days ago

Added most of Question 6
[redacted] committed 6 days ago

Added code and narrative for Question 6
[redacted] committed 6 days ago

What does this have to do with collaboration?



- If students have graded team assignments / assessments, early pointers for best practices help establish common expectations.

- Being meticulous about regularly and informatively committing work makes them better collaborators.

3 Git doesn't just happen.

Resolving merge conflicts

Option 1:

Team activity where we cause and resolve merge conflicts **during class**.

Works well in small classes, with established teams.

Setup:

- Start with identical repos, one for each team.
- Assign numbers (1), (2), (3), and (4) to team members. Going forward only one member at a time touches their computer.

- **Member 1 -** Change the team name placeholder to your actual team name in the YAML of your R Markdown file, save, commit, and push (with an informative commit message!)
- **Member 2 -** Change the team name to some other word, save, commit, push.
    - You should get an error. Read the error!
    - Pull.
    - Locate the merge conflict in the R Markdown file (it should be on top, but you can also search for the word HEAD)
    - Resolve the merge conflict by choosing the correct/preferred change.
    - Commit with a message “Resolving merge conflict”, and push.
- **Member 3 -** Add a label to the first code chunk, save, commit, push. You should get an error. Pull. No merge conflicts should occur. Now push.
- **Member 4 -** Add a different name to the first code chunk, save, commit, push. You should get an error. Read the error! Pull. Locate the merge conflict in the R Markdown file. Resolve the merge conflict by choosing the correct/preferred label. Commit with a message “Resolving merge conflict”, and push.

Resolving merge conflicts

Option 2:

Individual activity where we cause and resolve merge conflicts **during class**.

Works well in larger classes, where Option 1 can be difficult to manage.

Setup:

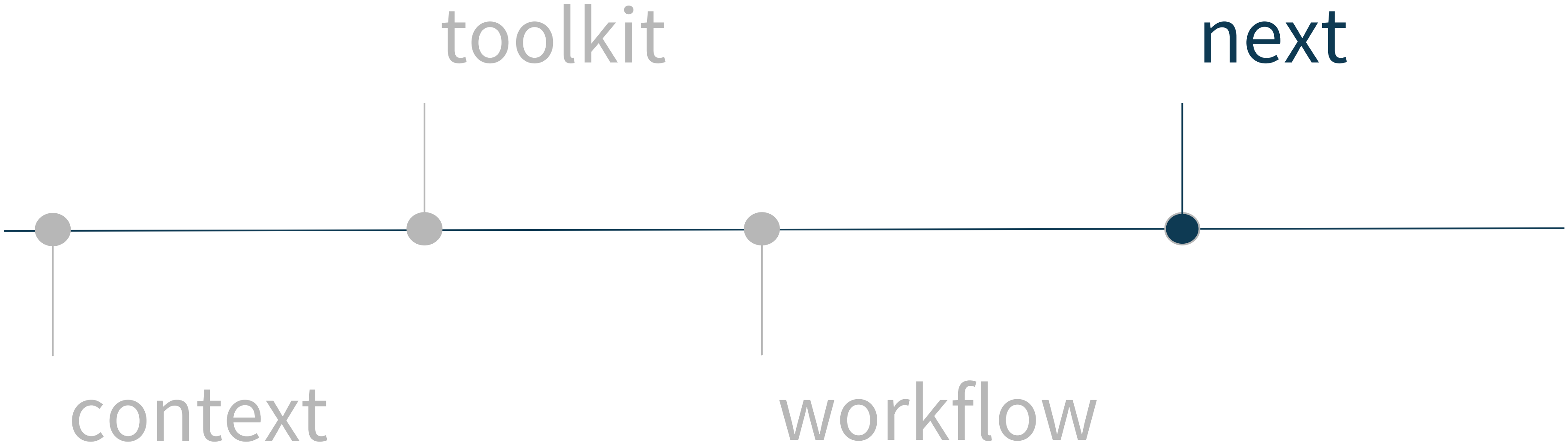
- Start with identical repos, one for each student.

- **Students -** Each student should update their R Markdown file in their repo to change the placeholder author name to their name. Then, commit (with an informative message) and push the change.
- **Me -** Push a file with the same name and (almost) identical content as the students’ R Markdown file to their repositories, with the only difference being my name as the author name.

```r
install_github("rundel/ghclass")
library(ghclass)

amc_repos <- get_repos(org = "Sta199-S18",
                       filter =  "demo-02-merge-conflict-")

add_files(amc_repos,
          message = "I’m going to cause a merge conflict, watch out!",
          files = "demo-02-merge-conflict.Rmd")
```

- Merge conflicts can be hard, let the first time students struggle with them be in an organized fashion and in class where you / TAs can help in person.

- Observation: If students *reeeeaaally* want to avoid merge conflicts, they physically get together and work!



context

toolkit

workflow

next

1

Branch / PR / inline code review model

2

Peer review

we'll there,
slowly but surely

course web: bit.ly/sta199-s18

course org: https://github.com/Sta199-S18

mine-cetinkaya-rundel
minebocek
mine@stat.duke.edu

mine çetinkaya-rundel
duke university
statistical science