Homework #2 Documentation

DSC 200 - Data Science I

2022-03-14

Deadline: 23:59 on Monday, 21 March 2022

Total Points: 10

Getting started

Prerequisites

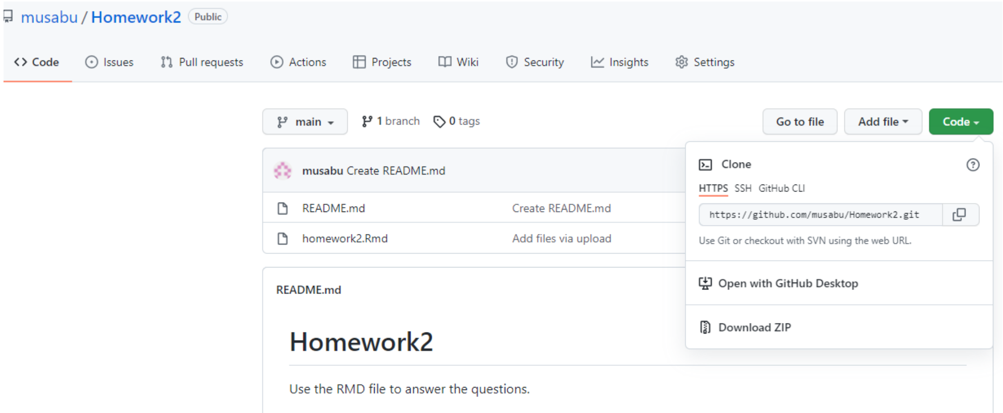
This Homework assumes that you have reviewed the lectures titled "Meet the toolkit: Programming" and "Meet the toolkit: version control and collaboration".

If you haven't yet done so, please pause and read through the lectures before continuing. I am always available to help via email if you get stuck.

Starting slow

As the course progresses, you are encouraged to explore beyond what the Homework dictates; a willingness to experiment will make you a much better programmer. There are excercises included in the Lecture slides on Blackboard for you to practise.

Step 1. Get URL of the repo to be cloned and clone it.



musabu Create README.md

README.md Create README.md

homework2.Rmd Add files via upload

https://github.com/musabu/Homework2.git

Use Git or checkout with SVN using the web URL.

README.md

Homework2

Use the RMD file to answer the questions.

Go to the GitHub page https://github.com/musabu/Homework2, click on the green **Code** button, select **HTTPS** (this might already be selected by default, and if it is, you'll see the text *Use Git or checkout with SVN using the web URL* in the image on the right). Click on the clipboard icon to copy the repo URL.

#To Clone it

Go to you github account, click the + sign on top-right corner of the page and select *import repository*. Paste the URL copied and write ‘Homework2’ in the *Repository Name* field. Make it *Private* and click *Begin import*.

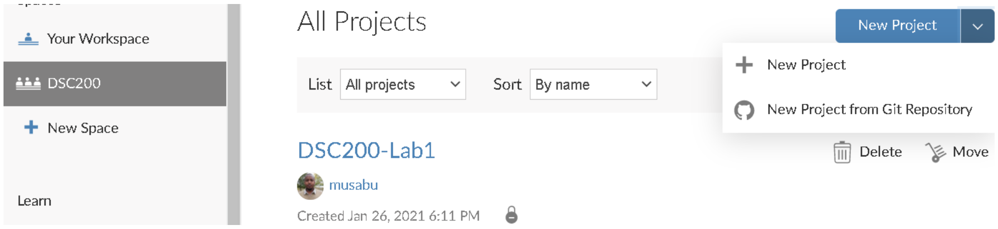
Step 2. Go to RStudio Cloud

Go to rstudio.cloud and then **navigate to DSC200 workspace** via the left sidebar. If you don’t have DSC200 workspace, click on ‘New Space,’ type the name and click on ‘Create.’

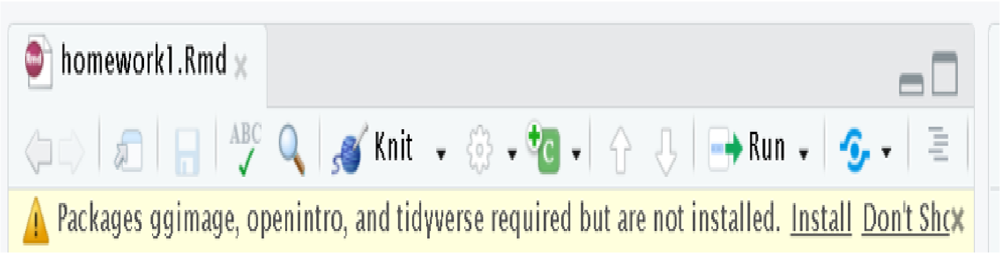
Step 3. Download the cloned repo to your RStudio

In RStudio, click on the **down arrow** next to **‘New Project’** and then choose **New Project from Git Repository**.

In the textfield “URL of your Git Repository,” enter the following link https://github.com/YourUsername/Homework2.git replacing the text ‘YourUsername’ with the username you used in creating your GitHub account. Make sure the box for **Add packages from the base project** is checked (it should be, by default) and then click **OK**.



After the project is cloned, click on the Rmd file **homework2.Rmd** from the file viewer on buttom-right of RStudio. You may see the text *packages ggimage, openintro, and tidyverse required but are not installed* as shown in the image below. Please click *Install* to download the packages.



Packages ggimage, openintro, and tidyverse required but are not installed. Install Don't Sho

Warm up

Before we introduce the data, let’s warm up with some simple exercises.

Step 1. Update the YAML

Open homework2.Rmd, change the author name to your name, write your Student ID (remove the following text: ‘Type your Student ID here’), and knit the document.

Step 2: Create a Personal Access Token on Github and store it in your RStudio

See Lab 1 document on how to complete this process.

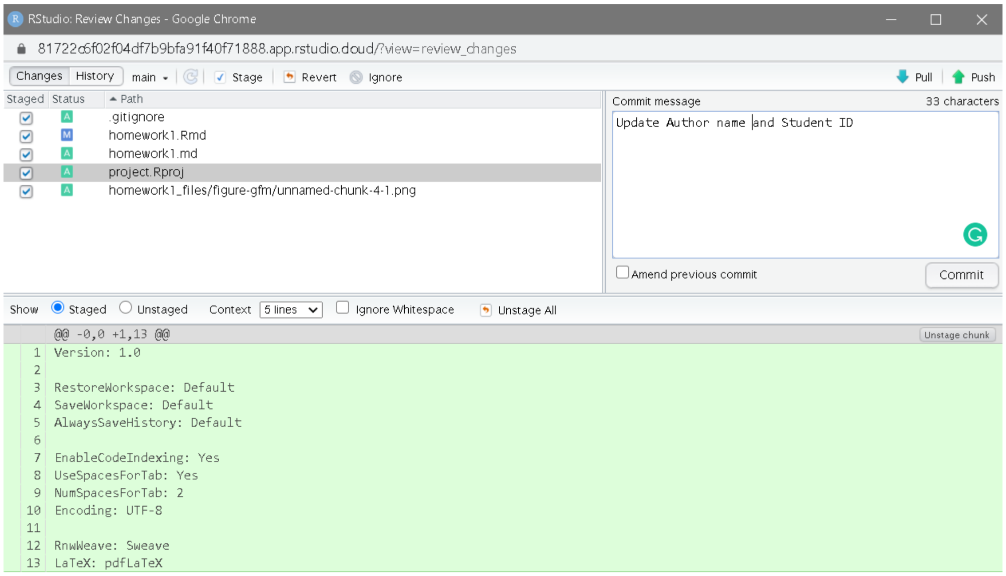
Step 3: Commit

Then Go to the **Git pane** in your RStudio.

You should see that your Rmd (R Markdown) file and its output, your md file (Markdown), file are listed there as recently changed files.

Next, click on **Diff**. This will pop open a new window that shows you the **diff**erence between the last committed state of the document and its current state that includes your changes. If you're happy with these changes, click on the checkboxes of all files in the list, and type *"Update author name and student ID"* in the **Commit message** box and hit **Commit**.

Note: the names and number of files shown on your RStudio may be slightly different from what is shown in the image below. That's fine.


RStudio: Review Changes - Google Chrome
81722c6f02f04df7b9bfa91f40f71888.app.rstudio.cloud/?view=review_changes
Changes
History
main
Stage
Revert
Ignore
Pull
Push
Commit
Unstage All
Unstage chunk

Staged	Status	Path
☑	A	.gitignore
☑	M	homework1.Rmd
☑	A	homework1.md
☑	A	project.Rproj
☑	A	homework1_files/figure-gfm/unnamed-chunk-4-1.png

Commit message: Update Author name and Student ID (33 characters)

□ Amend previous commit

Show ● Staged ○ Unstaged Context: 5 lines □ Ignore Whitespace

```
@@ -0,0 +1,13 @@
Version: 1.0

RestoreWorkspace: Default
SaveWorkspace: Default
AlwaysSaveHistory: Default

EnableCodeIndexing: Yes
UseSpacesForTab: Yes
NumSpacesForTab: 2
Encoding: UTF-8

RnwWeave: Sweave
LaTeX: pdfLaTeX
```

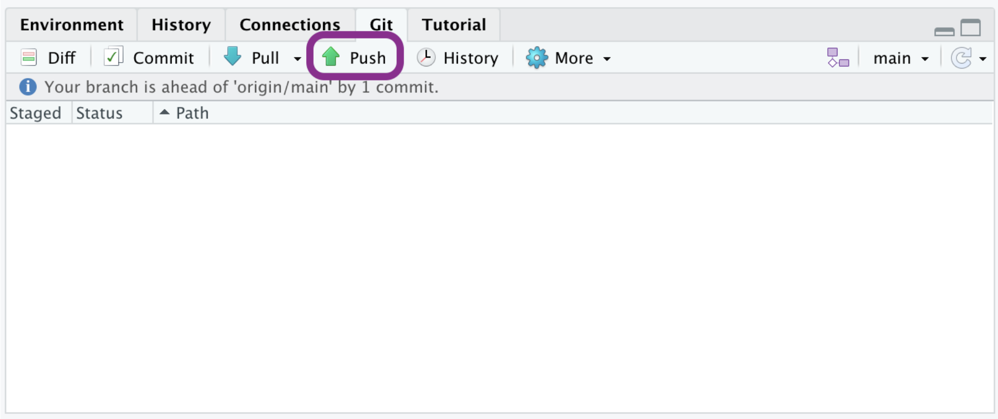
You don't have to commit after every change, this would get quite cumbersome. You should consider committing states that are *meaningful to you* for inspection, comparison, or restoration.

Step 4: Push

Now that you have made an update and committed this change, it's time to push these changes to the web! Or more specifically, to your repo on GitHub. Why? So that your instructor can see your changes.

You can invite me to see your work by going into your repository and clicking on ‘Settings’ –> ‘Collaborators’ –> ‘Add People’ –> Type my username ‘musabu’. Select my username from the list and add.

In order to push your changes to GitHub, click on **Push** after doing the commit in the previous step or in the Environment Pane under the Git tab as shown in the image below.



Your branch is ahead of 'origin/main' by 1 commit.

Packages

R is an open-source language, and developers contribute functionality to R via packages. In this Homework we will use the following packages:

- **tidyverse**: a collection of packages for doing data analysis in a “tidy” way
- **openintro**: a package that contains the datasets from OpenIntro resources
- **ggrpel**: a package that contains extra geoms for ggplot2

Data

The city of Seattle, WA has an open data portal that includes pets registered in the city.

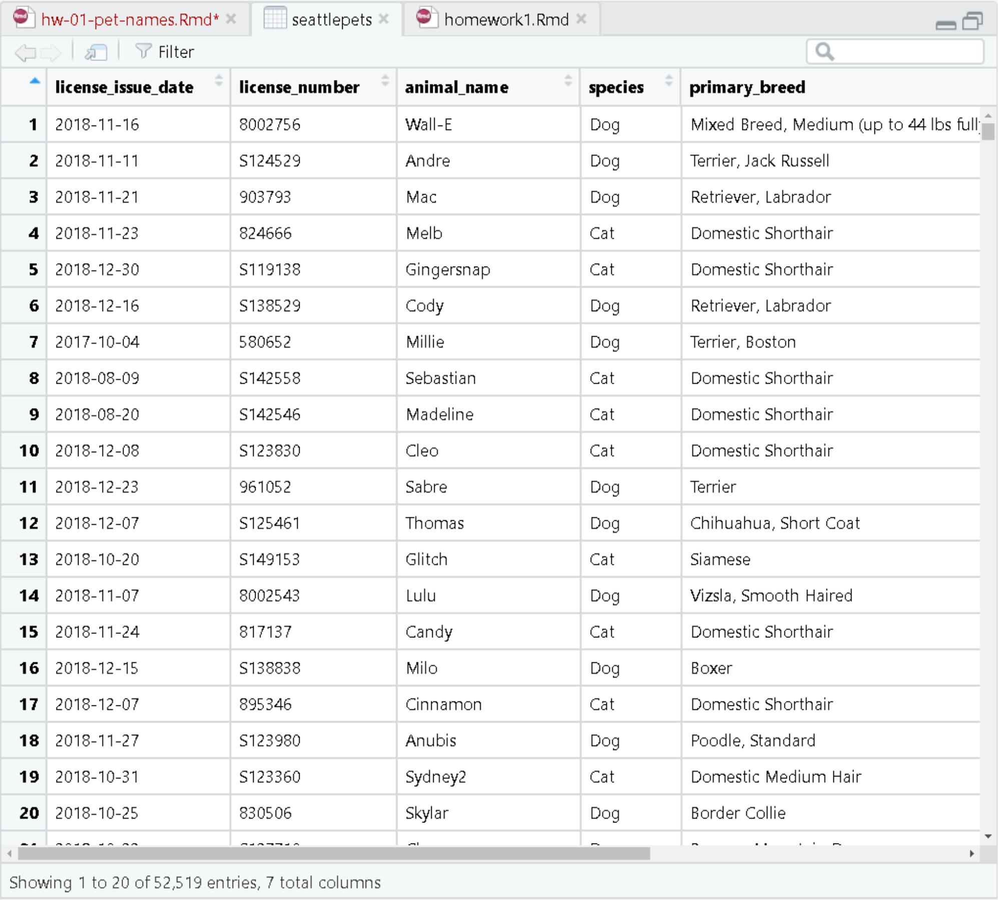
For each registered pet, we have information on the pet’s name and species. The data used in this exercise can be found in the **openintro** package, and it’s called `seattlepets`.

Since the dataset is distributed with the package, we don’t need to load it separately; it becomes available to us when we load the package.

You can view the dataset as a spreadsheet using the `View()` function. Note that you should not put this function (`View()`) in your R Markdown document, but instead type it directly in the Console, as it pops open a new window (and the concept of popping open a window in a static document doesn’t really make sense…).

When you run this in the console, you’ll see the following **data viewer** window pop up.

```
View(seattlepets)
```



	license_issue_date	license_number	animal_name	species	primary_breed
1	2018-11-16	8002756	Wall-E	Dog	Mixed Breed, Medium (up to 44 lbs full
2	2018-11-11	S124529	Andre	Dog	Terrier, Jack Russell
3	2018-11-21	903793	Mac	Dog	Retriever, Labrador
4	2018-11-23	824666	Melb	Cat	Domestic Shorthair
5	2018-12-30	S119138	Gingersnap	Cat	Domestic Shorthair
6	2018-12-16	S138529	Cody	Dog	Retriever, Labrador
7	2017-10-04	580652	Millie	Dog	Terrier, Boston
8	2018-08-09	S142558	Sebastian	Cat	Domestic Shorthair
9	2018-08-20	S142546	Madeline	Cat	Domestic Shorthair
10	2018-12-08	S123830	Cleo	Cat	Domestic Shorthair
11	2018-12-23	961052	Sabre	Dog	Terrier
12	2018-12-07	S125461	Thomas	Dog	Chihuahua, Short Coat
13	2018-10-20	S149153	Glitch	Cat	Siamese
14	2018-11-07	8002543	Lulu	Dog	Vizsla, Smooth Haired
15	2018-11-24	817137	Candy	Cat	Domestic Shorthair
16	2018-12-15	S138838	Milo	Dog	Boxer
17	2018-12-07	895346	Cinnamon	Cat	Domestic Shorthair
18	2018-11-27	S123980	Anubis	Dog	Poodle, Standard
19	2018-10-31	S123360	Sydney2	Cat	Domestic Medium Hair
20	2018-10-25	830506	Skylar	Dog	Border Collie

Showing 1 to 20 of 52,519 entries, 7 total columns

You can find out more about the dataset by inspecting its documentation (which contains a **data dictionary**, name of each variable and its description), which you can access by running `?seattlepets` in the Console or using the Help menu in RStudio to search for `seattlepets`.

Exercises

1. How many pets are included in this dataset? (Simply count the number of rows) 2 Points

Write your answer in your R Markdown document under Exercise 1, knit the document, commit your changes with a commit message that says "Completed Exercise 1", and push. Make sure to commit and push all changed files so that your Git pane is cleared up afterwards.

2. How many variables do we have for each pet? (Simply count the number of columns) 2 Points

Write your answer in your R Markdown document under Exercise 1, knit the document, commit your changes with a commit message that says "Completed Exercise 2", and push. Make sure to commit and push all changed files so that your Git pane is cleared up afterwards.

3. What are the three most common pet names in Seattle? To do this you will need to count the frequencies of each pet name and display the results in descending order of frequency so that you can easily see the top three most popular names. 6 Points

Write your answer in your R Markdown document under Exercise 3. In this exercise you will not only provide a written answer but also include some code and output. You should insert the code in the code chunk provided for you, knit the document to see the output, and then write your narrative for the answer based on the output of this function, and knit again to see your narrative, code, and output in the resulting document. Then, commit your changes with a commit message that says "Completed Exercise 3", and push. Make sure to commit and push all changed files so that your Git pane is cleared up afterwards.