DSC 200 - Data Science I
Lab Project Documentation
Term 2232

Deadline: 23:59 on Monday, 20 May 2024

Contents

1. Project Objectives

The goal of this project is to explore the dataset ***seattlepets***, a dataset of registered pets in Seattle, WA, between 2003 and 2018, provided by the city's Open Data Portal. It is also to evaluate the skills you acquired in the use of R, Posit Could, Git, and GitHub.

2. Workflow

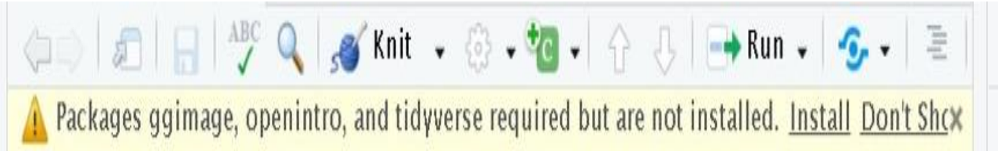
The next few steps will walk you through the process of getting information of the GitHub repo to be cloned, cloning your repo in a new Posit Cloud project, and getting started with the analysis.

Step 1. Clone the Project files to your github account.

1. Import the repository **https://github.com/massayony/Project_Repo_Term2232.git** to our github account.
2. Name the repository **Lab_Project_T2232**.
3. Make it private.

Step 2. Import the project to Posit Cloud

1. In your Posit Cloud account, crate a **New Project from Git Repository**.
2. Make sure to provide the URL of your own github repository (i.e., https://github.com/<your_github_username>/Lab_Project_T2232), DO NOT use https://github.com/massayony/Project_Repo_Term2232.git
3. Open the Rmd **Lab_project.Rmd**. If you see the text *packages ggimage, openintro, and tidyverse required but are not installed* as shown in the image below, click *Install* to download the packages.



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

Step 3. Update the YAML

1. Update **Lab_project.Rmd**, by your name and ID, then
2. knit the document.

Step 3: Commit

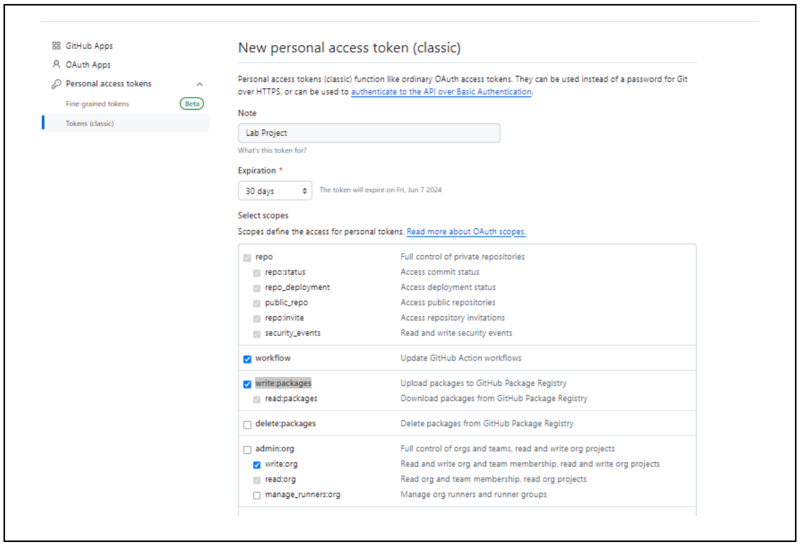
1. Go to the **Git pane**, click on **Diff**. This will pop open a new window that shows you the **diff**erence between the last committed state of the document and its current state that includes your changes.
2. Click on the checkboxes of all files in the list.
3. Type *"Update name and ID"* in the **Commit message** box, and
4. hit **Commit**.

Step 4: Push changes

To push your changes to GitHub, you need to create and add personal access token (PAT) to your project.

To create a PAT on your Github account:

1. Go to Settings ➔ Developer settings ➔ Personal Access Token, then choose Tokens (classic) and click on Generate new token (classic).
2. In the Note text filed, type "Lab Project Token"
3. Make sure to check ALL boxes for write privileges, specifically, the first one **write:packages**.



New personal access token (classic)

Personal access tokens (classic) function like ordinary OAuth access tokens. They can be used instead of a password for Git over HTTPS, or can be used to authenticate to the API over Basic Authentication.

Note: Lab Project

What's this token for?

Expiration *: 30 days — The token will expire on Fri, Jun 7 2024

Select scopes

Scopes define the access for personal tokens. Read more about OAuth scopes.

☑ repo	Full control of private repositories
☑ repo:status	Access commit status
☑ repo_deployment	Access deployment status
☑ public_repo	Access public repositories
☑ repo:invite	Access repository invitations
☑ security_events	Read and write security events
☑ workflow	Update GitHub Action workflows
☑ write:packages	Upload packages to GitHub Package Registry
☑ read:packages	Download packages from GitHub Package Registry
☐ delete:packages	Delete packages from GitHub Package Registry
☐ admin:org	Full control of orgs and teams, read and write org projects
☑ write:org	Read and write org and team membership, read and write org projects
☑ read:org	Read org and team membership, read org projects
☐ manage_runners:org	Manage org runners and runner groups

4. Copy the your personal access token to clipboard (for safety, you may also copy it to a text file and save it in case you need to reuse it)

To add the PAT to your Posit project:

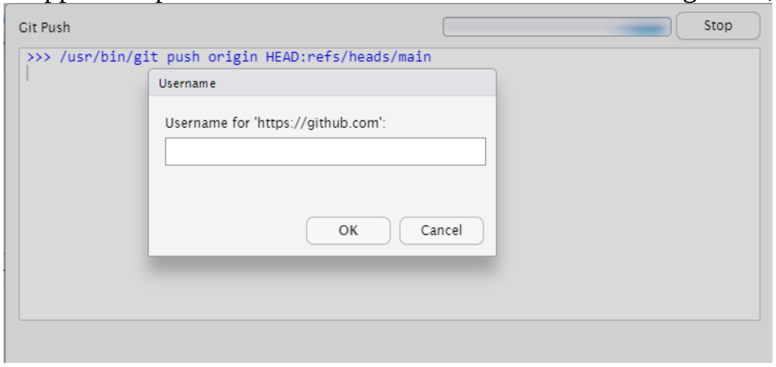
1. Go to the Console tab and type the following commands:

```
> install.packages("gitcreds")
> library(gitcreds)
> gitcreds_set()
```

2. It should prompt you for your access token. Paste the PAT created on Github earlier and hit ENTER.

Note: if the system asked you what to do or to make a select from some options, type 2, and hit ENTER, and then paste the PAT, and ENTER.

Note that you need to run the command `gitcreds_set()` every time you are asked to provide username for github.com as the support for password authentication was removed on August 13, 2021!



```
>>> /usr/bin/git push origin HEAD:refs/heads/main
```

Username for 'https://github.com':

If you encounter this message while you are trying to push, close the window and go to the Console to add the PAT to your project

Step 5: Confirm

1. Go to your repo on GitHub and confirm that your changes are visible in your rmd **and** md files.
2. If anything is missing, commit and push again.

3. Packages

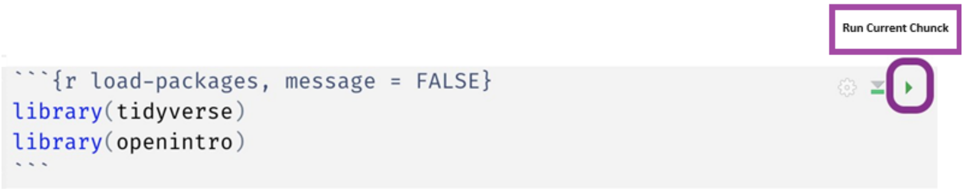
In this Project, you will use the following packages:

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

Note that these packages would be already downloaded if you execute step2.3 above!

You must also load the packages to R Markdown environment as well as to the Console.

1. Note that **Lab_project.Rmd** already contains an R chunk labelled ***load-packages*** which has the necessary code for loading both packages.
2. These packages can be loaded to R Markdown environment when you **Knit** the ***rmd*** document.
3. To load the packages in the Console, jsut send the code to the Console by clicking on the **Run Current Chunk** icon (green arrow pointing right icon).



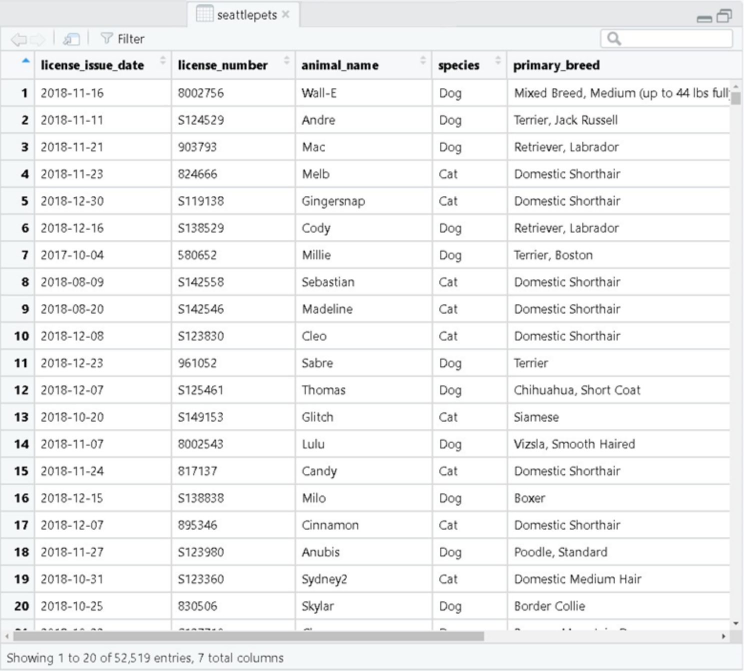
~~~
```{r load-packages, message = FALSE}
library(tidyverse)
library(openintro)
```
~~~

4. 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.

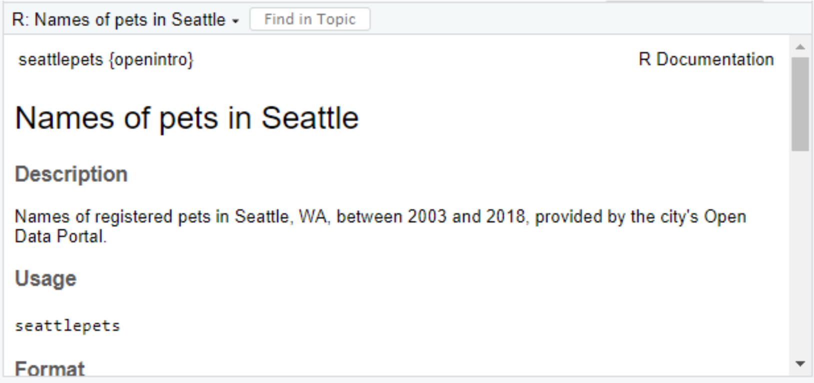
Type the command `view(seattlepets)` directly in the Console to view the dataset as a spreadsheet in the **data viewer** window pop up.



	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), by running `?seattlepets` in the Console.



seattlepets {openintro} R Documentation

Names of pets in Seattle

Description

Names of registered pets in Seattle, WA, between 2003 and 2018, provided by the city's Open Data Portal.

Usage

```
seattlepets
```

Format

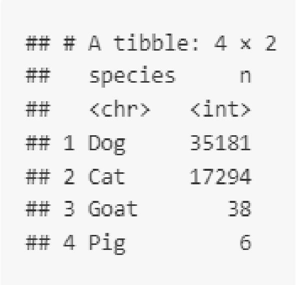
5. Tasks

For each of the following tasks, do the following:

1. write your answer (R chunk and narrative(explanation)) in the R Markdown document under the section ##Tasks.
2. knit the document,
3. commit your changes with an appropriate commit message like "*Completed Task 1*", and
4. push changes to your GitHub repo.

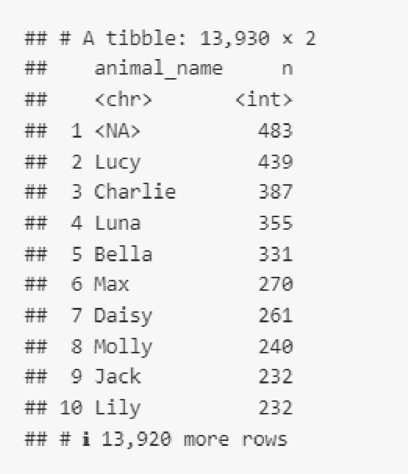
Make sure to commit and push all changed files so that your Git pane is cleared up afterwards.

1. How many pets are included in this dataset?
2. How many variables do we have for each pet?
3. What are the pet species in Seattle? How many pets are there for each species?
   Hint: To do this, you need to *count the frequencies* of each species and display the result.



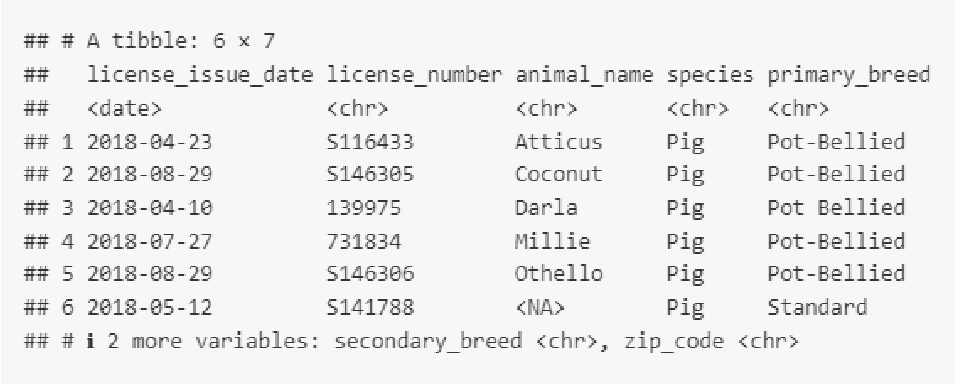
```
## # A tibble: 4 × 2
##   species     n
##   <chr>   <int>
## 1 Dog     35181
## 2 Cat     17294
## 3 Goat       38
## 4 Pig         6
```

4. What are the ten most common pet names (animal_name) in Seattle?
   Hint: To do this, you 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.



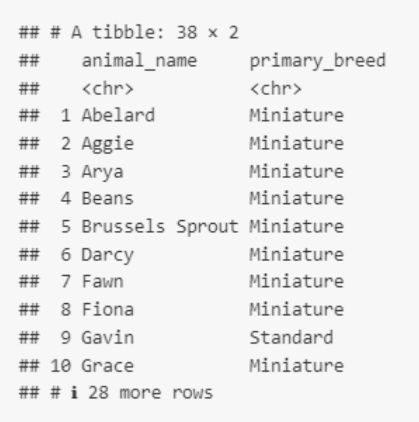
```
## # A tibble: 13,930 × 2
##    animal_name     n
##    <chr>       <int>
##  1 <NA>          483
##  2 Lucy          439
##  3 Charlie       387
##  4 Luna          355
##  5 Bella         331
##  6 Max           270
##  7 Daisy         261
##  8 Molly         240
##  9 Jack          232
## 10 Lily          232
## # ℹ 13,920 more rows
```

5. Retrieve and display all the 6 records for the species *Pig* sorted by pet names (animal_name) (use *filter* and *arrange* functions)



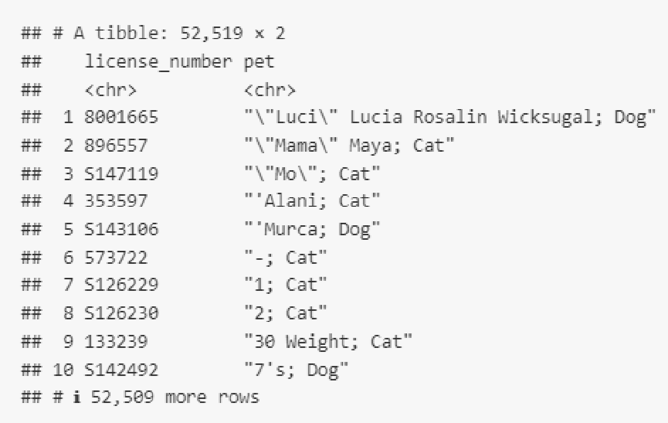
```
## # A tibble: 6 × 7
##   license_issue_date license_number animal_name species primary_breed
##   <date>             <chr>          <chr>       <chr>   <chr>
## 1 2018-04-23         S116433        Atticus     Pig     Pot-Bellied
## 2 2018-08-29         S146305        Coconut     Pig     Pot-Bellied
## 3 2018-04-10         139975         Darla       Pig     Pot Bellied
## 4 2018-07-27         731834         Millie      Pig     Pot-Bellied
## 5 2018-08-29         S146306        Othello     Pig     Pot-Bellied
## 6 2018-05-12         S141788        <NA>        Pig     Standard
## # ℹ 2 more variables: secondary_breed <chr>, zip_code <chr>
```

6. Retrieve and display ONLY the pet name (animal_name) and primary_breed of the species *Goat* sorted by pet names (animal_name) (use *select*, *filter* and *arrange* functions)



```
## # A tibble: 38 × 2
##    animal_name     primary_breed
##    <chr>           <chr>
##  1 Abelard         Miniature
##  2 Aggie           Miniature
##  3 Arya            Miniature
##  4 Beans           Miniature
##  5 Brussels Sprout Miniature
##  6 Darcy           Miniature
##  7 Fawn            Miniature
##  8 Fiona           Miniature
##  9 Gavin           Standard
## 10 Grace           Miniature
## # ℹ 28 more rows
```

7. Concatenate the two columns animal_name and species into a single column named *pet*, then display *license_number* an *pet* sorted by *pet* as in the below snapshot.

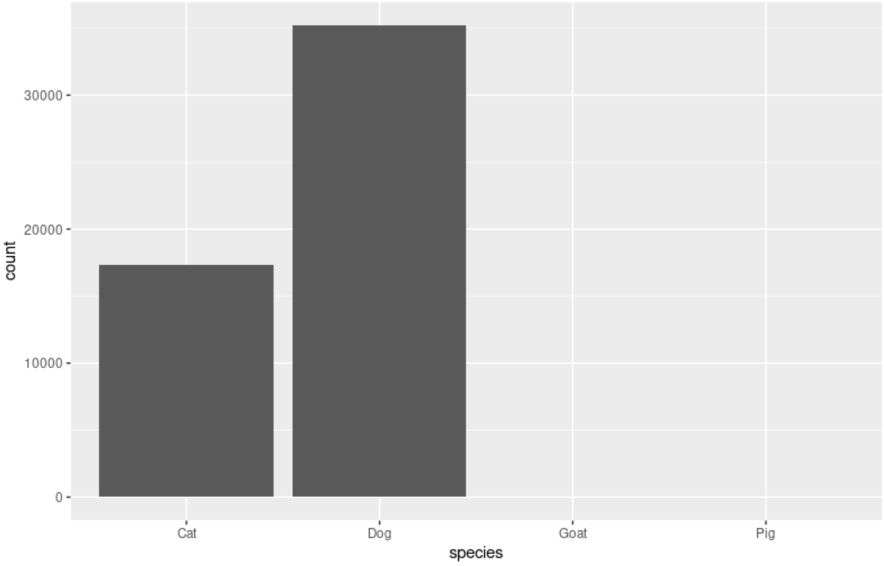


```
## # A tibble: 52,519 × 2
##    license_number pet
##    <chr>          <chr>
##  1 8001665        "\"Luci\" Lucia Rosalin Wicksugal; Dog"
##  2 896557         "\"Mama\" Maya; Cat"
##  3 S147119        "\"Mo\"; Cat"
##  4 353597         "'Alani; Cat"
##  5 S143106        "'Murca; Dog"
##  6 573722         "-; Cat"
##  7 S126229        "1; Cat"
##  8 S126230        "2; Cat"
##  9 133239         "30 Weight; Cat"
## 10 S142492        "7's; Dog"
## # ℹ 52,509 more rows
```

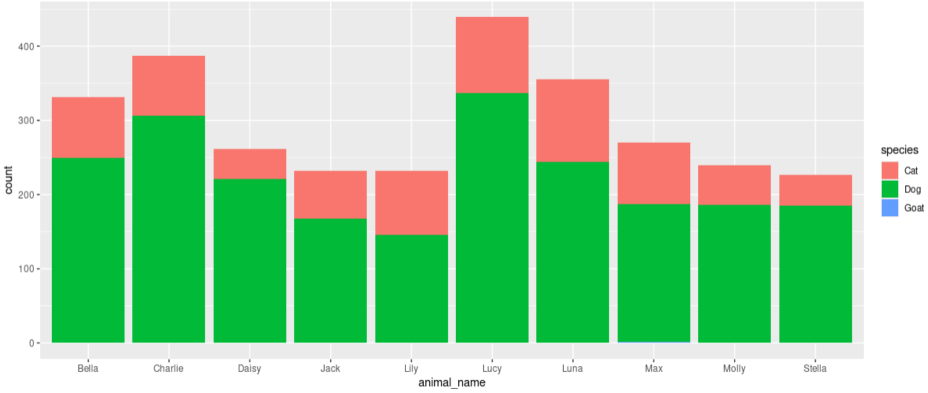
Hint: To do this, use `mutate` to add the new column *pet*. To concatenate the two columns, use the function `paste` with `mutate` as follows:

```
mutate(pet = paste( animal_name , species, sep = "; ")
```

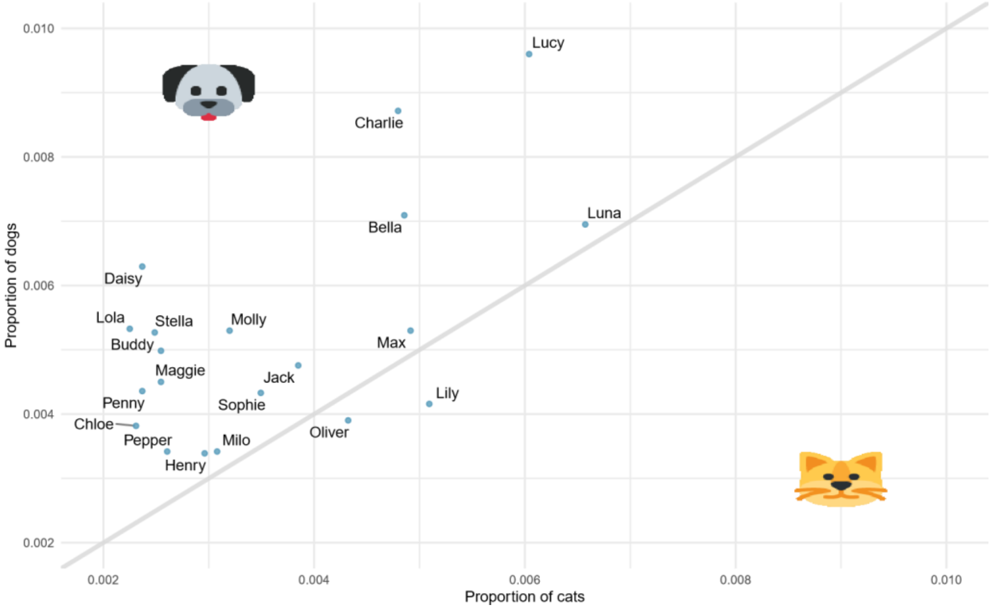
8. Plot the counts of the species as bars (use *geom_bar()*)



9. Study the code chunck under Task9.
    a. What does the code chunk do?
       Hint: Note that the category
       `c("Lucy","Charlie","Luna, ....................,"Stella")`
       represents the list of the 10 most pet names in the data set. (Refer to Task4 above.)
    b. Plot the counts of the pet names (`animal_name`) in `top_10_names` as bar plot segmented by `species` as below.



10. Study the code chunk under Task #10.
    The code plots the proportion of dogs with a given name versus the proportion of cats with the same name. The 20 most common cat and dog names are displayed. The diagonal line on the plot is the $x = y$ line; if a name appeared on this line, the name's popularity would be exactly the same for dogs and cats.



    a. What names are more common for cats than dogs? The ones above the line or the ones below the line?
    b. Is the relationship between the two variables (proportion of cats with a given name and proportion of dogs with a given name) positive or negative? What does this mean in context of the data? (Click on positive correlation and negative correlation to read and understand what the relation between variables mean before you answer question 6.

6. Solution Submission

The submission is two steps:

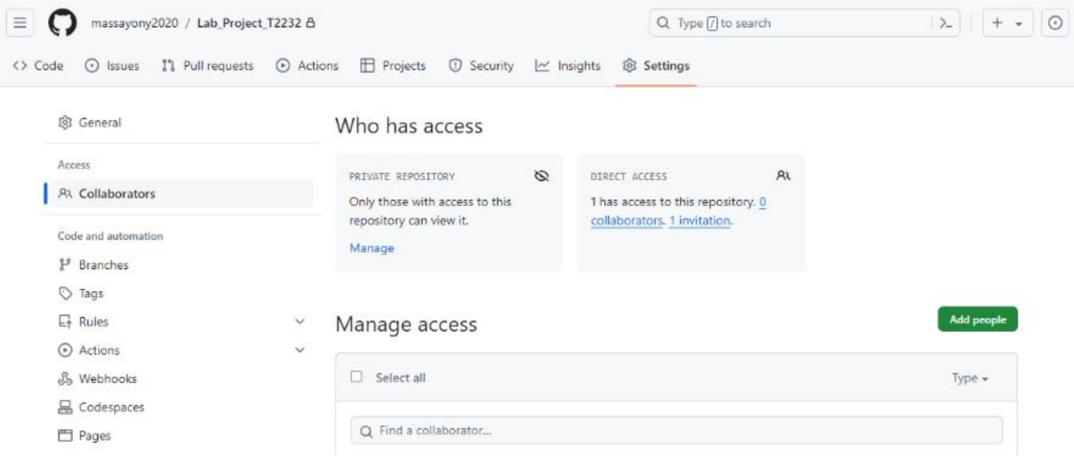
Step 1: Upload file Lab_project.Rmd to Blackboard:

1. Download the Final version of the file **Lab_project.Rmd** from your GitHub repo
2. Upload **Lab_project.Rmd** to Blackboard as a solution to this Assessment in Blackboard.

> Note that contains your answer (R chunk and narrative(explanation)) to ALL the tasks of this project.
>
> Make sure to download the **latest version** of the file after you commit and push all the updates from Posit Cloud to GitHub.

Step 2: Add your instructor as Collaborator to your GitHub Repo:

1. Go to your private repo on GitHub and click on ***Settings***.
2. To the left of the screen, under ***Access***, click on **Collaborator**
3. Then Click on **Add people** and type the instructor's username **m.assayony@gmail.com**



Who has access

PRIVATE REPOSITORY

Only those with access to this repository can view it.

Manage

DIRECT ACCESS

1 has access to this repository. 0 collaborators. 1 invitation.

Manage access