Reference notes

# Basic Git workflow

https://swcarpentry.github.io/git-novice/reference

--- open git bash on desired folder (i.e. Version 2 - Manuscript)

> git status

> git add .

> git commit –m “<commit message goes here>”

> git push origin master

> nano file.txt } open in a text editor

> cd . } change directory to parent folder?

> ls -a } list all contents of the folder, -a show all including hidden folders

Lab notebook

# Nov 14

## Notes about project structuring

* Pick a naming convention and stick with it. Do not mix them. Perhaps consider writing your own style guide.
* Use functions whenever you can, and keep them short.
* Use **packrat** to manage dependencies to keep things consistent and reproducible.
* Versioning systems like git keep you from holding onto old versions out of fear of needing them in the future. Use versioning software.

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000424

In practice, therefore, the scripts that I write tend to fall into these four categories:

1. **Driver script.** This is a top-level script; hence, each directory contains only one or two scripts of this type.
2. **Single-use script.** This is a simple script designed for a single use. For example, the script might convert an arbitrarily formatted file associated with this project into a format used by some of your existing scripts. This type of script resides in the same directory as the driver script that calls it.
3. **Project-specific script.** This type of script provides a generic functionality used by multiple experiments within the given project. I typically store such scripts in a directory immediately below the project root directory (e.g., the msms/bin/parse-sqt.py file in [Figure 1](https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000424#pcbi-1000424-g001)).
4. **Multi-project script.** Some functionality is generic enough to be useful across many projects. I maintain a set of these generic scripts, which perform functions such as extracting specified sequences from a FASTA file, generating an ROC curve, splitting a file for *n*-fold cross-validation, etc.

* I’m interested to try writing my manuscript as an r markdown file. One of the biggest barriers seems to be sharing with collaborators, but apparently rmd files can be knit as a word document: https://rmarkdown.rstudio.com/articles\_docx.html

## Setting up Git

Links from the Software carpentry collaborate document

**Resources**

<https://swcarpentry.github.io/git-novice/reference>

<https://www.atlassian.com/git>

**Git cheat sheets**

<https://www.atlassian.com/git/tutorials/atlassian-git-cheatsheet>

<https://github.github.com/training-kit/downloads/github-git-cheat-sheet.pdf>

Following the first link

* Set user name, email, line endings
* Come back to setting the default text editor if it’s not what I like when I try it

-I can click on a folder and say Git Bash here to start in that directory

-Could use GUI or Bash. I think for now I will use Bash since that’s what I learned before and what I know I have good instructions for. Maybe when I’m ready I can write down a list of the commands that I would use regularly when committing etc

-Created a .gitignore file. Set it up to ignore everything that is in the “data” folder

-what adding all of the files that I had in the repository (at this point, just folders, r project files, and the git ignore file), I got the following warning message from git bash. I am going to ignore it based on what I read on the internet.

warning: LF will be replaced by CRLF in .Rhistory.

The file will have its original line endings in your working directory

-set up and connected folder to a github repository

# Nov 15, 2019

## Setting up Git

-added a README file

-did another commit/push of edits

-created the reference notes at the beginning of this document

# Nov 19, 2019

## Preparing PRISM data – getting started

-added the new version of the raw prism data to data folder: unedited excel file and CSV file with the top two rows that were full of Jennie and Isabelle’s comments removed.

-Laurent gave me his script for cleaning the PRISM data

-created a library script for attaching packages, user defined functions, defining objects that will be used regularly such as coordinate reference systems

-started working on a prism data cleaning script, based on scripts Laurent and I have written already

# Nov 20, 2019

## Preparing PRISM data – just looking around

-started a list of things that I don’t understand about the PRISM data, or errors, @ top of this document

-looking through columns I might want to find inconsistencies/surprises (recorded above)

# Nov 21, 2019

## Preparing PRISM data – getting on the same page as Laurent and Natalie

-Added temporary files starting with ~ and .Rproj.user to my git ignore file

-tried to add git configuration setting to make comparing version of .docx files within git more useful. Not 100% sure if it worked or if it matters. Instructions found here:

<https://stackoverflow.com/questions/22439517/view-docx-file-on-github-and-use-git-diff-on-docx-file-format>

<https://github.com/vigente/gerardus/wiki/Integrate-git-diffs-with-word-docx-files>

-spent a while talking with Laurent to figure out why prism GIS files of zones and plots don’t align well, and possible solutions

-spent an hour on the phone with Natalie and Laurent to talk about what we are finding with the PRISM data. Tried running one of Natalie’s scripts for her afterwards

-moved all of my PRISM notes from top of this document into a shared google doc with Natalie and Laurent.

-read the document describing the GIS analyses that Tyler did

# Nov 22, 2019

-worked on filtering prism dataset, particularly time and date columns. Some of them are annoying to deal with and I’m going to put it off until I’m sure that I actually need those columns.

# Nov 25, 2019

-making a plan of what I need to do next. Finished making this document so that I could talk to Paul about it

-this seems like a good reference for SDM process: <https://damariszurell.github.io/SDM-Intro/>

-worked a bit on filtering prism dataset

-received another new version of the PRISM dataset. I had to re-check some of the columns that I had checked last week. Some of the have changed slightly

# Nov 26, 2019

## Bi-Weekly Meeting with Paul

-came up with the idea of using Path or SEM modelling to look at the interrelationships between the predictor variables

-made me realize (again) that my grid doesn’t fully need to line up with the plots, I can just make my own

-better idea to start modelling, then fix my layers once I know what I really need

-talked about time lags ideas but we both were confused about what you actually measure and how it would all work

-talked about computational issues I might run into – options if I need more power are 5th floor, Carleton physics, ECCC supercomputers, ACEnet

-which one has the shortest queue?

## Other tasks

-Made notes about SEM and Path Analysis in literature review document

# Nov 27, 2019

## Trying real hard to finish cleaning up the PRISM data

-changing variables classes

-fixing mistakes (species in wrong group etc)

-trying to edit mutate\_cond function so that it would work with NA values. The test script I made suggested to me that it didn’t work the way I thought it did. Did not figure this out.

# Nov 28, 2019

## Finally finished cleaning up PRISM data

-finally finished going through all of the columns: looking for errors, filtering out plots that don’t meet my criteria for inclusion, cleaning formatting

-recorded all of these errors on shared doc / to send to Isabel

# Nov 29, 2019

## Comparing and extracting data from PRISM GIS

-checked the specs of the INGO environment so that I can compare to other options:

-24GB RAM, 6 processors

-GIS files and excel file have different plot names. GIS file names are likely in some of the historic plot name columns

-maybe the proportion of plot surveyed is more reliable than the areas?

-SIR 38D isn’t in any of the GIS files. Would be useful to write a script to match gis files with excel file plots to see where they match up

-first create my list of all plots in the excel file so that I don’t have to look at multiple rows?

# Dec 2, 2019

## Itemizing all the plots that we need to make decisions about how/whether to include them

-Made lists of plots that were surveyed multiple times in the same year, plots that were surveyed in multiple years, and some of the plots that were field selected

# Dec 3, 2019

## Continue list of plots that need decisions made about them

-Finished lists of plots that were field selected and field modified

-Read a bunch of the different comment columns to try to understand what’s happening

# Dec 4, 2019

## Continue list of plots that need decisions made about them

-started trying to write down my opinions and questions about decisions

-looking at specific plot examples to understand them

-wrote a script to count the number of surveyors in each plot

\*\*\*This video about tidyverse tricks looks worthwhile: <https://www.youtube.com/watch?v=NDHSBUN_rVU>

# Dec 5, 2019

## Finished list of plots that need decisions made about them

-tried to group plots together to see if there were logical years to keep, but part way through realized this didn’t make that much sense. There will always be a reasonable sized group created during that field season

-created summary table and models looking at the differences in how the plots were selected (GIS vs field)

-watched the video I found yesterday

# Dec 12, 2019

## Moving on to modelling

-since last week: added a couple of other decisions to my list of decisions, which I thought I had finished. Went through it with Paul and Laurent, looked at some of the GIS files in a bit more depth, long conversation with the PRISM team on the phone to discuss all of our issues and divide up jobs

# Dec 16, 2019

## Goals this week

* Monday: Finish prepping a temporary version of prism data to play around with in modelling – 1 day
* Tuesday: Extract environmental values onto current version of dataset – 1 day
* Wednesday: Review ecography paper – 1 day
* Thursday: Sketch of my ideas for moving forward with hypotheses and analyses – 1 day
* Friday: Tidy up my office, prepare for next semester, attend to little things that I might be ignoring, test out version control to see how it works

## Finish temporary version of dataset

* Merging plots that were surveyed multiple times in the same and different years into a single observation value per plot
* Merging all plot dataset with shorebird dataset

# Dec 17, 2019

* Made a map of PCI plots for Jean-Louis
* Added goose counts to my shorebird observation dataset

# Jan 7, 2020

* Was reviewing my notes from the WHSG conference. These r function and packages sound useful; takes a raster and calculates a number of landscape metrics (patch size, patch density, shapes, areas, edges). Similar to FRAGSTATS

<https://www.rdocumentation.org/packages/spatialEco/versions/1.2-1/topics/land.metrics>

<https://cran.r-project.org/web/packages/landscapemetrics/landscapemetrics.pdf>

# Jan 13, 2020

## Goal for the week

Monday: ~~Meet with Paul~~, ~~Finish EE review (this took all week, I wrote way too much),~~ ~~read through revisions, ~~TAing~~

-prepare work from Brandon

~~-Register for nature museum symposium~~

-finish my paper outline

-finish my prism dataset that I was working on before Christmas

-Start attempting to do some test stats/ analysis

# Jan 21, 2020

## Goals for the week

-prepare work for Brandon

-finish my paper outline

## Current project priorities

-test if my modelling ideas work (Use Yukon PRISM data?)

-assess my computational needs

-start running habitat layers in case they take a long time

-did some stuff for TAing on Mon, Wed. On Tuesday I did a bunch of little tasks that built up while I was working on the review during the previous week. On Thursday I stayed home sick and read the path analysis book. On Friday I worked on sketch of hypotheses document, did my prep for the discussion group.

# Jan 28, 2020

## Goals for the week

-finish my paper outline

-prepare work for Brandon

## Current project priorities

-test if my modelling ideas work (Use Yukon PRISM data?)

-assess my computational needs

-start running habitat layers in case they take a long time

Tuesday: worked on my sketch document for about an hour and a half. I’m getting a bit lost. What concrete steps do I need to do?

* 1. I’m trying to flesh out my hypotheses a bit more. I think I want about a paragraph explaining the reasoning behind each one
  2. I need to read more about shorebirds and figure out what I would guess would be the most important factors are that shape their distribution. Traits, predictions relating to those traits.
  3. I want to have a clearer idea of what my predictions are and what my analyses would be that would produce those predictions. See the diagrams I drew on paper – path, interpretations

# Week of March 2

-getting back into my project after spending February working on my masters paper

-looked at the ENM course that Debbie sent me. I’m at 40 mins in the 1st week’s video. I wonder if I can download the audio onto my phone to listen to in the car

## Current project priorities

-finish my plan of what I’m going to do

-brainstorm a plan for my future projections chapter

## Conclusions of meeting with Paul on March 11

Chapter 1 – prepare land cover data

Chapter 2 – Shelve CTI idea

-figure out how I could fit ground based work into PCI and Rasmussen comparison chapter

Chapter 3 – Email Richard Pearson and ask him if there are any plans to revisit his projection project

Field work – figure out if I have science needs for chapter 2, figure out how involved I would get in shaping data collection

# Week of March 16

## Current priorities

-get some version of landcover data ready for testing

To do:

~~-go to office to get files and equipment that I’m missing~~

~~-see if I can get the tower computer running faster~~

-finish reading the things I’m thinking about for spatial scale (see my motes in the landcover document)

-re-assess what environmental layers I have and which ones I’m missing

-download data layers that I’m missing

-NDVI

-NDWI

What do I want to send to Paul tomorrow:

~~-A list of what I have for the land cover variables, some characteristics of them, what I’m planning to do with them~~

~~-write an email to richard pearson~~ <- I’m going to do this in a month when things have simmered down a bit

~~-write an email to Debbie about land cover maps newer?~~

Found in:

**SUPPLEMENTAL MATERIALS (for raster cavm map)**

Table S1. Detailed legend for Raster Circumpolar Arctic Vegetation Map. Species authority is the Circumpolar Artic Flora and Fauna (CAFF) Group’s Pan-Arctic Species List (PASL) (Walker et al 2016)

Idea: In this notebook, I should make notes about what I made notes about that day and where I stored them, so that at least I can find everything that I’ve been thinking about

Is this too much work?

**Conclusions of meeting with Paul on March 20**

-Gave me instructions on how to connect to the network from home:

Eps file

Password

Ciscob anyconnect

Gov password

Enter email with ec.gc.ca

Say I agree

An option the prevents when physically connect (later), after done setting stuff up and downloading

Paul – get advice from other people on remote sensing and sdm stuff would be a good idea. Suggested Karine mallard (?), Cheryl Johnson etc

Talk to Laurent about Landcover data, and status of the GIS coordinate stuff. Says one of us should tell him to switch to that if that’s what I need to progress

-and ask him what is john doing to see if it’s similar to what I’m doing

## Getting back into the PRISM R prep

-What are the NA’s in the PRISM quality column? Why don’t they have a quality?

-I removed 2 years (1996 PCI, 2019). I might want to add these back in if I’m trying to verify the coordinates of all plots. There might be other paring down that I might want to remove.

---look at what Scott published with the PRISM data

# Week of March 23

## Monday

-talk on the phone with Laurent. He wrote a script that matches up the names of Isabelle’s plots with Tyler’s plots. There are ~260ish without any match, so we probably have to assume that Isabelle’s are correct

-he loaded only the attribute tables into R, not full shapefile. Smart. Easier and lighter.

He is sending me:

-script that matches plots names

-list of plot names

-list of plots that he removed for whatever reason

-google doc and google drive of what he was working on with John

TO DO:

~~-read the google doc and google drive and see if there is any useful information them about what he is doing with the Landcover data~~

-when reading johns notes on PRISM 2 (on dropbox), it seems to be that the good/medium/poor classification system isn’t as clear as they make it out to be. I wonder if my model based approach could be less subjective than this?

-when I get Laurent’s script, I want to use it to see if the coordinates are the same between Isabelle’s file and Tyler’s files

-I will need to figure out how to have them in the same projection. Figure out how to do this. What projection are Tyler’s files in? What projection are Isabelle’s files in?

~~-make script 3: adding GIS plot coordinates all in a uniform projections system, from UTM~~

~~-what coordinate system are all the tyler gis files in? check this~~

## Tuesday

-made the rscript that adds coordinates to plots in my datasets I’ve made

-downloaded qgis, but it pisses me off. Not showing gdb files on my, slow as fuck on pc

-started extracting shapefiles from the geodatabases, but its so slow. Try to do it in R before proceeding

## Wednesday and Thursday

-figured out how to extract shapefiles form geodatabases in R

-Got Laurent’s script working to see what he did and what I could use

# Week of March 29

-Goals:

1. verify all coordinates for plots, and make a dataset that is all of the prism plots with their proper coordinates

2. extract habitat proportions in these plots

Steps to complete 1:

~~-add Laurent’s plot name corrections to my script shapefile dataset~~

~~-find all the plots that match between Isabel’s files and Tylers files~~

-I could try to match more by looking in other GIS files. There are 700 missing

-compare their GIS coordinates to see if they’re the same

-I will either have to add all of Isabel’s coordinates to that dataset, or set it up to find within all of Tyler’s, or set it up to pick one corner of each of them and compare those

-check how many of them don’t match, how much they don’t match, and if there’s any obvious reason why they don’t match, if discrepancies can be grouped by year or region.

## Sunday

-added Laurent’s plot name corrections to my script shapefile dataset

-Found which plots matched between Isabel and Tyler’s. There are many that don’t match. Could end up being worthwhile to check the “Selected” gis files as well

-Part way through figuring out how to get the coordinates from shapefile attribute to data frame column so that I can try to match them between the two datasets

## Monday

-Backed my work up on git

-prepared code that allows me to compare the coordinates from GIS and DB. At a first glace, they look pretty different.

-I re-did it using WGS84 so that I could actually measure the distance between the two set of coordinates. Distances ranged from 0-3000 km.

-Now investigating if it was my transformations that caused the issue

## Tuesday and Wednesday morning

-Re-reviewed Ecography paper

## Wednesday afternoon

-trying a different approach. If I keep the files I created as polygons in the same crs, I could probably find a way to measure the distance between those

-seems like the best way to do this would be to use the sf package

https://r-spatial.github.io/sf/articles/sf1.html

-still seems like it would be worthwhile to look at the two different version of the plots in GIS. However, the fact the I had made the Observation one just a point seemed to be problematic. In retrospect though, it shouldn’t be? I should be able to look at where those points are

-if no one answers my GIS question, I will have to export some of the un transformed versions of plots by hand

- I think should just go ahead and do the habitat cover calculations with the GIS data, and then check the coordinates afterwards.

## Thursday and Friday

-I used the GIS plots to extract percent cover

-Convo with Paul

## Friday April 10

-to do before next Friday: finish assessing if the coordinates from PRISM match

-finish getting all my other environmental variables ready

# Week of April 13

## Monday April 13

-getting Isabel’s polygons ready for matching with Tyler’s

## Tuesday April 14

-look at both sets of shapefiles in GIS

-realized that my code was associating the wrong name with plots

-posted my question on stack overflow

-started researching potential temperature data

## Wednesday April 15

-finished comparing some of the GIS shapefiles. There are still some left to do though

-prepared to send this to others

## Thursday April 16

-sent work above to Paul

-going through other predictor variables. What data do I have, what do I still need to get

## Friday April 17

-finished summarizing my current status of predictor variables

-talked on the phone with Paul

-worked on clarifying GIS terminology

-checked the CRS’s I used. I did change datums. Try doing it without changing datums as much as possible

-talked to Laurent on the phone about questions he has about plot locations

-checked the plot coordinates in ArcGIS, they are the same as what I did in R

# Week of April 20

## Monday April 20

-Re-did some of the parts that needed quality control based on conversation with R

-changed the map\_datum of some of the UTMs and using a different final projection for all of the final products so that there was less switching between datums. This didn’t change my results

-sent Laurent the list of plot’s ive checked so far

-trying to look at the other plots that I didn’t check yet. I realized my initial categorizations don’ t split them up perfectly. Re-check the coords sets/ re-think how to plot them up.

To do:

Check the rest of the plots

Send questions to Tyler and Isabel

Continue preparing data layers

-have all of this done my when I talk to Paul 2 weeks from now

## Tuesday April 21

*Notes from “What They Forgot To Teach You About R”:*

<https://rstats.wtf/get-to-know-your-r-installation.html>

-wasn’t actually complete when I was looking at it. Check again in the future.

-I could have all my project files in one file, and then the actual project folders would be elsewhere

-file names should be both machine and human readable, takes advantage of default ordering

-gobbling: including a keyword in file name so that they are easy to find

-can put metadata in filenames. Separate within a unit with – and betweenunits with\_

-2013-04-28\_SAMPLES\_location-name\_file-name.csv

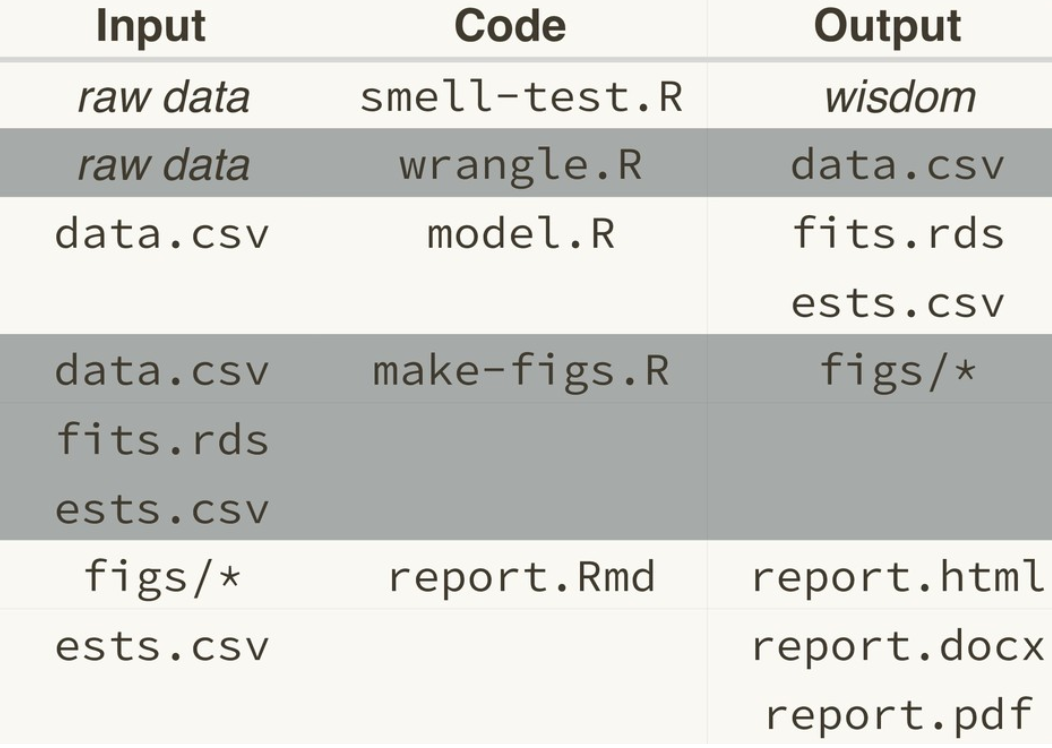
-descriptive names with – separating them is called a slug (from URLs)

-make all number series have the same number of digits (i.e. 03, not 3)

-put something numeric first to make sorting easier

-dates: YYYY-MM-DD

-have different scripts and outputs for different steps:



-you can edit .Rprofile to set a default CRAN mirror, customize options

-they suggested keeping base/recommended packages separately from other package which would allow you to update R without affecting your packages. I don’t fully understand and probably won’t do this right now.

-Mac and PC have different tools you need to install separately if you want to use source packages (includes non-R code) instead of binary packages (from CRAN, nothing extra) – Need to set up a a development environment, use devtools

-you can install packages to a temporary library if you don’t want them to affect your normal library

Do this course after: <https://happygitwithr.com/>

* Worked all week on trying to match up Isabel and Tyler’s GIS coordinates
* I kept making small alterations and alternative files until I got confused about what I had done. Next time I should work slower, and make more documentation. I got a very different answer than last time, and when I looked in GIS many plots were in the wrong locations. Need to start from the top and work through it more slowly and carefully with more checks.
* This would have been a good situation to use version control. I wish I could go back a fix what I had done.

-set up version control for this computer

-rename all my files according to the conventions above

# Week of May 11

## Tuesday May 12

* I notice that I stop being organized and taking notes when I get too deep into something. It feels like it takes too much time, but probably saves me time in the long run
* The past couple of week, I finished assessing the state of all the coordinates, and trying to figure out the likely best location of the coordinates. I made a document, discussed with Paul, Jon, and Laurent, and am now putting it into action. Last week I found lost plot coordinates with Isabel.
* Yesterday I confirmed if the locations I was creating in R matched up with Tyler’s shapefiles for the plots the plots that we will use his coordinates. There was an error when the names were getting mixed up in region 10. I also noticed some Somerset Island plots were duplicated. Had a phone meet with Paul , Jon, Laurent
* I have a set of coordinates for about 75% of the plots, based on the shapefiles

## Wednesday May 13

* Started worked on adding the remaining 25% of coordinates that need to come from the excel file. I started changed my script so that I save all the UTM coordinates, not just the first 5
* Goal: convert all the files that require coordinates from the excel file
* I came across this and it seemed useful: <https://tidyeval.tidyverse.org/sec-why-how.html#unquoting-code>
  + I learned that if a function wants to interpret an unquoted expression as a character, but you want to force it to interpret the expression as an object, you can use !! before it aka the tidy evaluation operator

# Week of May 11

## Tuesday May 19

* Sent a list of plots that were missing some corners to Isabel
* Weekly phone call meeting

## Wednesday May 20

* Try more of these Firefox extensions: <https://blog.mozilla.org/firefox/firefox-extensions-work-from-home/?utm_source=desktop-snippet&utm_medium=snippet&utm_campaign=moz-blogs-h1-2020&utm_term=23102&utm_content=REL>

## Friday May 22

* Converted g2 to lat long (plots that don’t have a GIS shapefile, and are straightforward to deal with). Visually checked them, outlined issues, sent to Isabel for verification.

# Week of May 25

* Finally finished proofing coordinates!!!!! Now to go back through a tidy up lose ends that I’ve ignored for the last month or so
  + ~~Back up all of my work and files somewhere~~
  + Get github working on my current work set up
  + Rename and re-organize files
  + Save useful code, delete useless code
  + Re-think coastline buffer / study area boundary, try to include the plots on islands. Will the plots outside the coastline buffer have Landcover data? Landcover boundary could be an alternative
  + Create a version to send to Isabel with two different kinds of coordinates and other relevant geographical info
    - Re-project all the shapefiles into UTMS, include Lat/Long and UTM
    - Include a comment what the same coord is used for multiple plots
    - I suspect CLGT regions/subregions are incorrect

-to some cleaning and organizing before bit. I have big files in areas that I will upload. Also, make sure to not upload some potential new folders like exported

-move some of the notes I’ve made from SDM file to here?