Group Diary

Allan

22/09/21

Allan chose to wrangle rainfall data (https://catalogue.data.govt.nz/dataset/river-flow-and-stage-monitoring-sites/resource/eb2e5666-9601-476c-bac4-82973574b7f6)

25/09/21

Allan changed dataset to wells and bores dataset because it has more rows than rain fall data (https://catalogue.data.govt.nz/dataset/river-flow-and-stage-monitoring-sites/resource/eb2e5666-9601-476c-bac4-82973574b7f6)

01/10/21

Tried convert NZTM coordinates in the wells and bores dataset to latitude/longitude (using Jupyter Notebook R). Initially tried this with sf package but had installation issues. Used R map library to do this.

02/10/21

Then tried to convert latitude/longitude into suburbs. But was unable to do this due to lack of access to Google Maps API.

05/10/21

Change dataset to Victimisation Time and Place dataset (https://www.police.govt.nz/about-us/publications-statistics/data-and-statistics/policedatanz/victimisation-time-and-place?nondesktop) to wrangle in R (because was unable to convert latitude/longitude into suburbs with wells and bores dataset). Downloaded Victimisation Time and Place dataset as tableu file, then converted to CSV.

11/10/21

Victimisation Time and Place dataset wrangling in R. Filter dataset by Canterbury suburbs only.

Remove rows with no suburb. Remove full stops at end of suburb column names. Group by suburb, and create column get number of crime records, number of victims, and count of crime types per suburb. Attempt to groupby suburb and types of crime to get a separate column for number of each crime type in each suburb (was unsuccessful).

15/10/21

Data wrangling in Julia with Subnational population estimates dataset (consisting of suburb population data) (http://nzdotstat.stats.govt.nz/wbos/Index.aspx?DataSetCode=TABLECODE7993&_ga=2.104052355.2017186628.1634336143-719107025.1632194594&_gac=1.187780826.1634341688.CjwKCAjwzaSLBhBJEiwAJSRokp1ezN-rHwcbENfgmCOaqB5ZV1PjF8PHuwo4KUjET-ASdu_WBRRo5xoCnIgQAvD_BwE)

Convert csv file into DataFrame. Indexing was used to select only the suburb, and 2016-2020 suburb population columns. Rename was used to rename a column from "June" to "Suburb". Save dataframe as a new CSV.

16/10/21

Left join Subnational population estimates dataset onto Victimisation Time and Place dataset. This resulted in a dataframe with unique suburb per row, with additional columns for crime statistics and suburb population statistics. Normalised crime data per 1000 people using transform function. This was done by dividing the number of crimes (or victims) divided to the year 2020 population times 1000. Rename columns so they are more user friendly to read. Create histogram of Canterbury Number of crime records per 1000 people between Sep2016-Aug2021.

18/10/21

Start writing report

30/10/21 until 01/10/21

Finish writing report. Changed my Subnational population estimates dataset that originally included all New Zealand suburbs to only Canterbury suburbs. Left join Victimisation Time and Place on Subnational population estimates dataset to include all Canterbury suburb population data (previously Subnational population estimates was left joined onto Victimisation Time and Place dataset).

Cathy

15/09/21

Join the group, discuss some topics of our group project. Such as Christchurch residents' health and environment analysist.

21/09/21

Started searching the datasets for data wrangle. Download the data from Statistic NZ: 2013_mb_dataset_Canterbury_Region, statsnzmeshblock-2021-generalised-CSV and statsnzmeshblock-2021-generalised-FGDB (tried to make a ArcGIS map of our study area)

28/09/21

Found the Unit-Area datasets to instead of meshblock code for next data wrangle work

04/10/21

Tried covert Unit-Area data in Jupyter Notebook R. output the suburb names Suburbs length/ Km and suburbs square /Km^2 in Christchurch

05/10/21

Discussed with Jonathan about which information or column we should keep. In wrangle Unit-Area dataset. Because we were doing a similar part but used different datasets. Therefore, we decided to add the local council details and region distinct in dataframe.

12/10/21

Discussed with Daniel of my census datasets, these this too old to our research, because the data from 2001 to 2013, some of the values are changed a lot until now, so we decided to use the most nearly census dataset and I would focus on wrangling regional households.

15/10/21

Had problems in R part: cannot read the xlsx format file in R, read Daniel's code. Selected the 2006 to 2018 Total households in occupied private dwellings. Dropped off total owned, total stated and total. The total stated is the same value as total. The total owned values cannot clearly identify, because it has another column named not elsewhere included.

17/10/21

Had some issues of plotting the local household diagram, tried to review the dataset. Head the top 10 suburbs, statistic summary the local household in r and Jupyter Notebook. To prepare the presentation and report structure.

22/10/21

Renew the households output file, I prefer to add dwelling owned or partly owned and Dwelling held in a family. But the final output file looks like it makes sense. In the Statistical area 1 dataset for 2018 Census about households in the Canterbury region. It also has dwellings not owned and not held in a family trust column. I tried to use reorder function in Jupyter Notebook R, the output results have several values trend wrong.

23/10/21

Prepared for the presentation script and fix a little bit of presentation slides. Practiced presentation with group members

26/10/21

Failed to figure the line of for locations and households by APIs,

29/10/21

Started writing the report. Discussed the report structures with Allan and Daniel.

30/10/21

Finished writing the introduction limitation and conclusion in my parts of report. Discussed with other members'of their dataframe.

Daniel

14/09/21

Created our group. It is composed of Daniel Peach, Allan Ou, Cathy Yin, Logan Lawson, and Jonathan Coachman.

21/09/21

We decided that our project is going to look at environmental data in suburbs around the Canterbury region, such as weather, water quality, air quality, etc. We've found a number of decent-looking data sources, particularly from ECAN and statsNZ. I am going to wrangle a dataset related to water quality, obtained from (Groundwater Quality Monitoring Sites - Current - Datasets - data.govt.nz - discover and use data)

26/09/21

I managed to convert some of the dataset to a tidy format, but there's an issue in that each row contains a url link that directs to a website which contains a number of tabs of water quality data over multiple recordings. This may be difficult to automate scraping.

28/09/21

We had the idea to try use regional/suburb house price data against the environmental data, so people using the dataset could look at how environmental features affect house price. There's a database run by REINZ which has substantial data, but its access is restricted. We are applying to see if we can use it.

30/09/21

I was unable to successfully wrangle the water quality dataset. Attempted to set up a scraping pipeline to scrape the contents of the csv links, however each link contains multiple part-csv files and so this would have required an extreme amount of work.

01/20/21

Found a new dataset, which can be obtained from the link (https://www.stats.govt.nz/information-releases/statistical-area-1-dataset-for-2018-census-updated-march-2020). This datasets consist of a number of sheets, each corresponding to a different collection of data.

03/10/21

Started to wrangle data using the readxl and tidyxl packages to extract information from individual sheets.

05/10/21 until 14/10/21

Extracted a number of datasets from the census data: income, employment status, ethnicity, age, gender. Converted column names to dataframe-appropriate values. Age has yet to be wrangled properly as there are a substantial number of columns involved, breaking up age into 5-year groups by gender.

12/10/21

We drew the schema for our ideal final dataset. It looks like it will be a star-type schema, centred on the suburb column, with branches for house price, crime rates, nearby schools, and census data (education level, income, rent, gender, etc.). We also signed a form to provide us access REINZ data. Decided to drop ethnicity data due to data ethics issues.

15/10/21 until 17/10/21

Dropped age data as the number of columns made the dataset difficult to work with. Implemented data wrangling steps using Julia – was successful in extracting income data from sheets in the .xlsx file by utilising the "XLSX" package. "!rename" was used to change column names to dataframe-appropriate values.

19/10/21

Established which group members were working on the report, presentation and visualisations. Attempted to visualise data by mapping through post code and suburb.

20/10/21

Cleaned datasets into final forms with dataframe-appropriate column headings and correct datatypes. Was unsuccessful in creating a map visualisation using postcode or suburb – Tableau only maps region, not suburb, and this process would be very difficult using a different visualisation software.

22/10/21

Visualised several dataset features in Tableau. Created line charts exploring median personal income and median personal rent by suburb.

30/10/21 until 01/11/21

Started working on the final copy of the report. Visualised several new features: median rent against median income by suburb, average school decile against median rent by suburb

Jonathan

21/09/21

We teamed up as group and came up with a temporary group name: TBA.

We then brainstormed about an interesting topic we can tackle as a group with Canterbury in mind. We have chosen our topic and decided everyone should look into interesting datasets until the next lab session.

24/09/21 until 28/09/21

We looked into various datasets. I looked into meshblocks and thought we could use that as a common join key for all future datasets. I tried to find real estate data, I looked into Propy (https://www.proppy.co.nz/) and Trade Me and whether they have a public API to access property prices (https://developer.trademe.co.nz/) with no luck.

I also tried to find weather data and stumbled upon metservice and their API (https://www.metocean.co.nz/news/2021/8/31/metocean-solutions-forecast-api-allows-easy-customer-access-to-accurate-marine-and-atmospheric-weather-data). Unfortunately, their API is meant for commercial applications and acquiring a license is costly.

01/10/21 until 05/10/21

I found a Stats meshblock dataset that seemed promising since they included not only the meshblock ids, but also the corresponding suburb in Statistical Area 2 standard and various other geographic data: (https://www.stats.govt.nz/information-releases/2013-census-meshblock-dataset). I also looked into other interesting and relevant datasets and found the crime dataset, which Alan decided to wrangle.

12/10/21 until 15/10/21

Wrangled the Stats dataset into two datasets, meshblocks.csv and areas.csv using R by first importing the Excel file using the “XLSX” package. Both have a common key that allows any other dataset with geographic data in either meshblocks or suburbs to be joined. I also found another dataset, the school directory from EducationCounts (https://www.educationcounts.govt.nz/directories/list-of-nz-schools) and decided to wrangled it. I chose relevant columns only, dropped the ones that are irrelevant, renamed columns, changed

values of 99 to none and added suburb data using Julia. Discussed with Daniel possible way of visualizing the data in Tableau.

19/10/21

We worked on the presentation this week. I prepared the data model structure using Oracle's MySQL Workbench and prepared the talking points for that part of the presentation. Practiced the presentation with group members.

26/10/21 until 01/10/21

We finalized the report writing and delegated various parts for the final submission. I wrote the achievements section and added to various parts that required input from the whole group. I also created a Github repository where we will upload the final data for the public to access the datasets. Formatted documents such as this group diary.

Logan

21/09/21

Create a relational database of geographic variables. The gist of our group's many ideas is relating geographical data with the goal of identifying novel relationships. I think of it like layering multiple factors of possibly connected data (on a map), with the purpose of later analyzing the dataset to determine where these variables interact. I have personal interest in the issues of systematic inequality and sustainability. My ideas are generally focused by these issues. I am also very interested in making something that is worthwhile doing. This may only be because it is worthwhile me learning how to do it; but! I would really like to make something that could be useful to someone in the future. ie. not just smash some tables together for the fun of it. If we have to do something like this, why not make it useful.

Prototyping: Here the questions I am posing are for the purpose of directing the way we collect and pre-process our data. I am aware we will do no analysis.

CLIFLO: Here I am calling upon my knowledge of NIWA's cliflo database of weather and other variables. An idea I had is summarized below. ![Wrangling Project Ideas-CLIFLO](D:\uni\mads\DATA422\groupProject\Wrangling Project Ideas-CLIFLO.png)

> Question: Is there a way classify geographical areas for use in an optimal way?
>
> Challenges: cliflo database only allows 40000 rows at a time, and there is a lot to get through, enormous database, data will require some engineering processing to convert weather data to a renewable quotient
>
> Thinking ahead: Could the classifications consider the environmental changes associated with development? Would the classifications still be accurate and useful post-development? Would the act of development change the classifications? Is there a type of algorithm for this type of simulation problem?

REINZ (property value data): Emailed REINZ - David Shaw, The REINZ database would unlock a very important variable in property prices. My idea is summarized below. ![Wrangling Project Ideas-REINZ](D:\uni\mads\DATA422\groupProject\Wrangling Project Ideas-REINZ.png)

> Question: What is the relationship between house prices and other socia-economic factors? Is it the case that property values and socio-economic factors have enforcing relationships? Are there thresholds or limits to these relationships? *Could these relationships be leveraged to improve an area's socio-economic health?*

Challenges: confidentiality issues, extra work explaining to REINZ

Thoughts: Everyone in the group is at this stage working on their own ideas. It will be important to get everyone together and nail down one thing next week. Prototyping did not get far today, I will have to work on this over the week. The little progress I made was: the above brainstorming, emailing reinz, downloading some data from cliflo

22/09/21

David Shaw replies

> However, there is something about your project which is a little bit different and very topical to REINZ. Data wrangling or the manipulation and linking of our sales database to other databases is an ongoing topic of interest to us. We don't have an in-house Data Scientist currently and so there would be a benefit to REINZ in having some fresh pairs of trained eyes looking at the data in the way you describe. I am thinking there is an opportunity here to provide you with the data you require in exchangefor a copy of your findings and a walk-through of those findings by your project team. Thinking bigger, there may be an opportunity for REINZ to develop a relationship/partnership with the University of Canterbury to work similarly.

The way that David has described this makes it sound perfect for this assignment. There may be some issues with confidentiality. I will discuss this with Thomas and Giulio.

28/09/2021

CLIFLO: Still haven't got far with a prototype - I've been a bit under the pump lately. I am working on it before our lab, hopefully I get somewhere.

REINZ: Meeting has been set up for tomorrow with REINZ. Will discuss questions to ask with the team today.

Giulio mentioned that it is unlikely this will work. He mentioned that if REINZ is overly secretive with their data - ie I cannot share my code with anyone, then this opportunity is not worth it for me. I agree.

Quick Meeting: Everyone to finish collecting data and sort it into regions/mesh blocks by next week, I am to get weather data sorted, downloading may require scraping, maybe take steps towards renewable quotient, maybe just estimate the time required, there will be a load of documentation required for this

TODO: Quick write up about cliflo, write a scraper to source the data, convert x,y to meshblock/region

29/09/21

Present: David Shaw, Ian ... (David's manager), Nick Ward, Logan Lawson, Giulio Dalla Riva

Logan's Agenda: What will we get from you? What are we allowed to do with it? How restrictive is REINZ with regards to our code? Can we publish our code? Github? What is REINZ looking to gain? What would you expect delivered? Giulio - Does this seem acceptable to you? What are your thoughts?

Meeting Notes: What will we get from you? Email to come regarding available data fields

What are we allowed to do with it? Join it with other interesting geographical features

How restrictive is REINZ with regards to our code? Can we publish our code? Github? REINZ is very cautious with handing over any large amount of data, however, David and Ian mentioned any IP we create in this data wrangling project belongs to us. We are not going to do any analysis but if we did, I imagine REINZ would be much more restrictive about what we could do with our findings. The fact that this is simply a wrangling project makes it much less risky for them. What is REINZ looking to gain? What would you expect delivered? Just a fresh set of eyes to look at the data and join it to other interesting datasets. We need to nail this down in more context with them, but this will likely mean a report from us and a walkthrough of this report Giulio - Does this seem acceptable to you? What are your thoughts? Giulio seemed receptive to the project. He offered some time to anonymize data if this was.

General: Wrangling only, Can publish scripts, Helpful for REINZ that we don't need much actual data - just something to work with, significantly reduces risk for them, Both UC and REINZ seem keen to develop an ongoing relationship through the DATA601 project program

Actions: David to provide list of variables, Logan to determine required variables

05/10/21

via email, David provided a list of variables and Logan determined required variables. We are just waiting on his reply with housing price data. At this stage, the team has decided to run with environmental data, weather data and housing data. I still haven't wrangled any data for this course and this is quite slack. I might focus on getting at least one dataset done today.

cliflo database (lpl251234 12F8X76R), weather data by date and area

Rethinking this Idea, Canterbury may not be big enough to create meaningful comparisons using weather data. David mentioned he may include Dunedin and Auckland data in the REINZ database. Giulio gave me some comments during our lab today about data security. He said that as long as the data is used on a secure machine and not published anywhere, there should be no issues. I think what I will do is purchase a thumb drive or something for the data and store it on that, maybe with help from REINZ. Another thing that could be worth doing is keeping a document detailing what happens with the data, when it is copied moved deleted changed etc. A hashtag in this journal may be suffficient. ->#REINZ<-

Scraper: Quick script to rip data off cliflo. found a library clifro that will do the job. And its for R!!!, #citation @unknown{unknown, author = {Seers, Blake and Shears, Nick}, year = {2015}, month = {04}, pages = {}, title = {New Zealand's climate data in R --- An introduction to clifro}, doi = {10.13140/RG.2.1.3689.7121} }

Meeting with other team: doing suburb related data, important locations, crime, crashes, Could link up afterwards with ours for REINZ

From the team: meetup next week

12/10/21

422 Presentation- important points: challenges, decision, background, what we tried, what worked what didn't, not bs syntax talk, Technical talk not a pitch, knowledgeaable audience, relational database diagram, really easy and explicit, beware tmi, Havent done anything in the lab today because ive been doing my other assignment, getting very frustrated with julia haha. I'll get onto this in the morning after I am caught up for digi405. Will catch up with team on friday or Saturday.

15/10/21

Finally working on wrangling for this project. Today I aim to complete wrangling of cliflo data into suburb level accuracy and build a scraping script capable of collecting any data required by 1 pm. I want to verify that this can be joined with the fake data supplied by REINZ. create a google project for mapping api Wrangling documentation, Variables, Important for REINZ, SaleID - Main Join to REINZ

Important for TBA: Suburb, SuburbID, SaleDate, SalePrice

Interesting: cliflo data - the data itself is not in a very accessible format. We should find/come up with a more easily understood aggregation of weather. Maybe something like a "good weather quotient" or something similar. image-20211015132512360

This is what it looks like now and it is pretty much unusable for any layperson (non meteorologically trained person).

Scraping: Location data from Cliflo data - lat long, Here I am saving to a .feather filetype to improve speed when working with large files down the track. When we look to join all our data hopefully this will improve performance.

Reverse geocoding cliflo data: The team has expressed interest in joining location by suburb. This seems like an appropriate promary level of granularity. cliflo gives location data in the form of latitude and longitude. The process of getting suburb information from coordinates is called reverse geocoding. This artice shows the process that I took to reverse geocode cliflo location data to suburb level. The script I have created should be able to be used for any coordinates. At this stage it may not work well on large datasets. I need to incorporate sleep times and memory management into the script to allow for bulk revgcing work.

Revgeo: Issue, revgeo is not working because it is getting information from http://photon.komoot.de/reverse?lon=171.7472&lat=-43.89658\. Using the address http://photon.komoot.io/reverse?lon=171.7472&lat=-43.89658 does work and returns:

{"features":[{"geometry":{"coordinates":[171.7473393,-43.8964029],"type":"Point"},"type":"Feature","properties":{"osm_id":5362353014,"country":"New Zealand","city":"Ashburton","countrycode":"NZ","postcode":"7700","county":"Ashburton District","type":"house","osm_type":"N","osm_key":"place","housenumber":"68","street":"Grig
g
Street","district":"Ashburton","osm_value":"house","state":"Canterbury"}}],"type":"FeatureCollection"}

If there is no alternative, I could write a scraping function for this website. Photon API is free but has throttling features. A solution could be a script that gets required information on demand. Otherwise, we'll just have to leave a scraper running for a long time to perform the reverse geo coding.

Tidygeocoder: I'll try this one since revgeo didnt work. The documentation will form the basis of my design. This seems to work fast and easy. Suburb data can be extracted from addresses.

17/10/21

Finished off a mock up ascript with a few stations from the chchc region. Works pretty well. Will run a cliflo scrape for all chch now and then run it through the script. There should be plenty of other environmental data that would be fine to run through this script too.

Decisions

Using feather file type for storage: Here we will have large datasets and be using dataframes in all parts of our wrangling. Feather is designed for this application. Final output could be a csv. Seems like feather is a better option for working on stuff pre join and export.

Choosing weather variables: Date, Wspeed_kmhr, Rain_mm, MeanTemp_C, Sun_hrs, These seem like the variables that the average person would be most interested in. If we were to be looking at agricultural property this may need to change.

Still jsut waiting on Giulio for the real data.

Team Meeting: We decided important actions, getting started on report, I'm going to get structure sorted - I've written plenty of technical reports, Need to have a look at the requirements and get the structure of the report, Need to write up our individual wrnagling parts

19/10/21

missed class today on account of upcoming test, currently writing my methodology for scraping the cliflo data and joining - will upload tonight, Datasheet info, Team doc

My approach: Scraping CliFlo weather data, CliFlo is a web portal the New Zealand National Climate Database. CliFlo requires a free subscription. To obtain weather data, the package clifro was used. clifro makes accessing CliFlo via R trivial. The steps involved in aquiring weather data are set out in table XX.

Step Package/functions What was involved

1 CliFlo web portal The CliFlo web tool was used to extract the station numbers for all weather stations in the Christchurch region.

2 cf_station Collected station numbers were supplied to cf_station to create a list of station objects for clifro to query CliFlo with.

3 cf_datatype A datatype . clifro uses a sequence of numbers to select required data fields to query from cliflo, a little bit like a phone dial menu (press 1 to go here). Using cf_datatype without supplying an input walks the user through the selection process.

4 cf_query The station and datatype objects were supplied to clifro's query function and the response assigned to to a variable.

5 tidyverse The response clifro object was converted to a tibble object and columns were renamed where required.

6 feather The resulting dataframe was written to file for further processing

Reverse geocoding CliFlo location data: Location data associated with CliFlo weather stations is provided in the form of latitude and longitude coordinates. In this project it was decided that suburb was the best primary key for joining datasets. To obtain suburb data, reverse geocoding is required. For this application, the tidygeocoder package was used. This package has capability to be supplied with coordinates and return a street adress. Suburb data was obtained from the returned structured address. The steps involved are detailed in table XX.

Step Package/functions What was involved

1 reverese_geocode The lat and lon columns from the CliFlo weather database were supplied to the reverse geocoding function and adresses returned.

2 separate The list of addresses returned by reverse geocoding was separated into columns by adress field and the resulting columns named appropriately.

3 mutate A new column was created for mean temperature. The column was created by mutating the provided maximum and minimum daily temperatures.

4 select Interesting meteorological variables were chosen and selected. The remaining variables were dropped.

5 renameThe columns in the final dataframe were renamed to more user friendly names by removing brackets.

Joining: The main dataset for this project is the REINZ property sales database. This database includes suburb and region columns. The weather data can be joint simply by these two columns.

22/10/21

We decided to present a slide each. Everyone will have to contribute information for everyone to put together.

Presentation slides: Objectives Logan, data sources Allan, techniques Jonathon, from Schema Jonathon, Visualisations Daniel, Challenges Cathy

Turns out my data didnt work properly, I'll fix this this weekend. Data dictionary: col name, dtype, description. For presentation, I'm going to write objectives into the report and come up witht the overall idea for the project. Ill confirm this with the team and go from there.

We discussed the issue of which suburb encoding to use. We made the decision to use general suburb names over the SA2 names present in some datasets. Daniel is going to make a script to clean up these names.

25/10/21

haven't been able to fix data - will try to do after I've written out the project objectives, Completed the slides in our meeting today. Going to catch up at 10am tomorrow for practice. Sent an email to David Shaw about a list of Suburb names. I may have to make up a solution in the meantime. We also haven't recieved the full dataset yet. Chasing up on that, hopefully giulio has the data. Possible approach to reverse geocoding.

30/10/21

Finished all other assignments now. I will be stuck into this all weekend.

To Do: fix reverse geocoding, Create suburb extractor script, compare adress entries with NZES database and extract suburb string -> didnt work, investigate GIS solution, Write up, The reverse geocoding process, this is just a band aid solution, Its pretty likely that there are much faster ways of doing this - but this works for now, geocoding services have been known to change policies - google now requires you pay - could pop up later, Reinz includes lat long, so applying the same process to their data could validate this method - have not received, REINZ may already have a method of RGC that we are not aware of, REINZ contains address data, Using the NZES suburb definition, Reverse geocoding is hard - with some literature, Design database location spine, Lat long, address, suburb, SA2, Integrate LINZ data into the joining process so SA3 etc. can be integrated down the track, Determine appropriate weather statistic aggregation, Mention in write up that reinz includes rural property data, so including weather could be very valuable for these properties, Recreate code in Julia, create an integration framework, Change paths and make all the scripts work together later on to make this reproducible, Talk to giulio and thomas about summer time

Nice to Have: implement a natural resources heatmap, Areas with good growing characteristics,areas with good renewable characteristics,areas that could do both,aggregate data or collect monthly data

Suburb extraction script: Dumb string method, splitting addresses, checking against another database, feather justification, This kind of worked but is enherently flawed. These data bases are too complex and address formatting is not standard. This can, however be adapted to fit the SA2 names to . Here it seems like the GIS approach may still be superior.

cool GIS version: Using the band aid approach mentioned earlier does nto seem appropriate. Here a GIS approach was taken. GIS software solution - not really what i want, How to do it in R, this is what i want also I figured it out from the above references. I'll write this up in the morning.

Database structure: I've decided to keep everything tidy, I would create an R environment and keep everything in there.

31/20/21

Spatial spine:

Lat lon meshblock sa2 fenz region

Coords id name name name

Here we will work completely in NZTM 2000 crs 2193

Main script notes: Cliflo station list, Manual collection from Cliflo web interface, Manual deletion of header rows in cliflo station list, Cliflo weather data, need cliflo account credentials - free account, Assign suburb function, need coord columns to be named lat and lon, Assign region function, need coord columns to be named lat and lon, Station list manual wrangle, some of the names came with extra comma's so fixed that, could be avoided by exporting to tab delimited on cliflo web app,Split up queries, Cliflo stops after 10 consecutive queries - maybe this is because of a free account, cliflo has a max of 40000 rows per query for a free account, Appropriate spatial granularity here is suburb, weather wont change much across more granular spatial encodings, weather stations are to sparse to warrant more detail, include localities, Next steps, aggregate data or collect monthly data

Station list procurement: image-20211031105522261, Started here and selected the daily datatype. Then went to the choose stations page. image-20211031105634025, Chose the center of NZ as a coordinate, then smpecified a radius of 1000 km. This should get all of NZ. Tried this with 1500 the first time and it picked up some stations outside of NZ. image-20211031105749907, Copied and pasted the resulting list. Ran the following lines to get rid of extra commas. readLines("collectedData/clifloStationList-Active.csv") %>% str_replace(pattern = ", ", replacement = "") %>% writeLines("collectedData/clifloStationList-Active.csv")#fix station list

01/11/21

Game time, write wrangling steps check! write REINZ procurement steps, write introduction, rewrite structure section, Format, Recreate something in Julia, write data dictionaries for all my things