# Dog Gone: Project Diary

## Lab Meeting 1 16/9/20

This lab session was the official commencement of the project and subsequent group work.

Initially we all worked together to identify an interesting data source. The ideas consisted of:

- Use Spotify to find the most popular artist in terms of number of streams and compare to other music ratings.
- Ascertain import and export data for New Zealand wineries / vineyards and view on a global perspective.
- Obtain Accident Compensation Corporation (ACC) falls data to determine which region in New Zealand was the clumsiest.
- Gather space flight data (https://nextspaceflight.com/).

A decision was made to have each individual of the group focus on a particular idea, and report back in a week with the results of their individual research and preliminary analysis.

The ideas were distributed as such:

- Andrew - Spotify/ music
    - 21/9/20 Looked further into the idea, Musixmatch.com is a good website for scraping song lyrics from. Found a Wikipedia article with artists with the most songs in the top 100, but thought that this could be punishing to older artists (anyone popular before the top 100 came into existence).
- Oscar - Vineyards/ wineries
- Torie - ACC falls data
    - Also explored the idea of looking at death rates in NZ in 2020 vs other years, but realized this would focus more on analysis than wrangling.
- Helen - Space flight data

# Group Meeting 1 23/9/20

The group met over lunch to discuss each individual's findings from their respective ideas.

In this session we came to a consensus that the ACC data was the most promising avenue for this project.

After making this call the group decided to get input from Giulio and the lab tutors in the lab session following lunch.

# Lab Meeting 2 23/9/20

In this lab we discussed our idea with the lab tutors. After receiving the nod of approval, we divided up the work and began wrangling the data.

The wrangling was divided up by the following roles:

- Andrew - E Scooter Injuries
    - 27/9/20 extracted the info from the excel sheets into a dataframe that would be useful for the project, ready to bring to the next group meeting.
- Oscar - Mobility Scooter Injuries
    - 27/9: put CSV data into Jupyter Notebook data frame using R and wrangled into usable format
- Torie - Fall Injuries
    - 27/9: put CSV data into Jupyter Notebook data frame and wrangled into usable format
- Helen - Hot Water Bottle Injuries
    - 27/9/20 Data was on a PDF so used an online PDF to excel convertor to retrieve data. Using R and wrangled data into 5 separate data frames relating to hot water bottle injuries.

The following resources were used to obtain the initial datasets and help wrangle them:

- https://catalogue.data.govt.nz/dataset/fall-data
- https://catalogue.data.govt.nz/dataset/hot-water-bottle-injuries
- https://catalogue.data.govt.nz/dataset/mobility-scooter-injuries
- https://catalogue.data.govt.nz/dataset/e-scooter-injuries/resource/aa5837a6-b743-499f-8e2e-ca1e0f3bbc1c
- https://catalogue.data.govt.nz/dataset?q=ACC&sort=score+desc%2C+metadata_modified+desc
- Sports injuries - data.govt.nz - discover and use data

- Motor Vehicle Accident claims - data.govt.nz - discover and use data

All data was kept in excel spreadsheets, and each group member used R to extract data from the displayed excel spreadsheets.

## Lab Meeting 3 30/9/20

First, the group shared the wrangled data frames. After a brief analysis one main obstacle to merging and associating the data sets became apparent:

- No consistent annual measurement.
    - Namely: financial year, calendar year and monthly reports.

At this point, Torie suggested exploring a topic related to pets, and came across the figures.nz website - https://places.figure.nz/. This website summarised a variety of data, including pet related information, per region and district in New Zealand. A second website was obtained that contained information on various dog breeds - https://www.akc.org/expert-advice/dog-breeds/2020-popular-breeds-2019/ (AKC). The group decided this would be a good opportunity to handle multiple types of data: i.e. data we scraped as well as data we found available in CSVs on the website.

In investigating the Figures NZ data, we decided to focus on the pets data section which lists a variety of statistics for each region, including:

- De-sexed registered dog by breed
- ACC claims
- Registered dogs by sex
- Total registered dogs

The data was available in CSV format; however, each CSV included a multitude of data for that region including:

- Demographics
- Income
- Households
- Wellbeing
- Health
- Education
- Work
- Work safety
- Transport
- Gambling

- Agriculture
- Election
- Council Finances

Additionally, the regions were split into 67 separate districts. We decided to consolidate the regions and districts into 14 distinct regions:

- Northland
- Auckland
- Waikato
- Bay of Plenty
- Gisborne
- Hawke's Bay
- Taranaki
- Manawatu/Wanganui
- Wellington
- Nelson/Marlborough
- West Coast
- Canterbury
- Otago
- Southland

Once the data we had was fully discussed as above, we distributed tasks for the next week:

- Andrew - Design a script in R that wrangles data frames from each district into a regional data frame.
    - 1/10/20: Wrote the first R code to get all of the pet information that we needed out of 1 specific csv. This needed running each time with a new CSV
    - 6/10/20: Combined the code I had written previously with code that could read the name of all files in a specific directory, and made it into a function that could read all of the files in a directory and take the pet information from each file and add it to a blank dataframe.
- Oscar - Determine the districts associated with each region. Assist Andrew with R code.
- Helen - Organise a record of events up to this point and a medium with which we can keep a journal for the duration of this project. Assist Andrew with R code.
- Torie - Begin scraping the AKC website into a Julia data frame.
    - 1/10/20: Went to another lab meeting to ask for help with Julia scraping. Had trouble with CSS selectors. From Emil's suggestion I scraped the AKC site in R to make sure it would work in Julia. Continued working on the site for the next few days.

# Lab Meeting 4 7/10/20

The group discussed the results of the previous weeks work.

Andrew had developed a code that could take the input of multiple district CSV data files and concatenate them into a singular data frame.

Oscar had created a table identifying the districts in each region using the following maps:

- www.planetware.com/map/new-zealand-north-island-regions-and-districts-map-nz-nz025.htm.
- www.planetware.com/map/new-zealand-south-island-regions-and-districts-map-nz-nz024.htm.

The Auckland region was not divided into its substituent districts by the figures.nz website.

Torie had successfully scraped the AKC website into a data frame in Julia. She saved the information to a CSV format which was then exported to R for further analysis.

Helen had constructed an accurate journal in google docs - a central hub for the project - and we kept the majority of resources and information present in this document.

In this lab the group presented the wrangling we had completed, were due to complete, and the direction of the project. Feedback was given suggesting we make sure to include a focus on another variable in relation to the dog data. This was discussed within the group prior to presentation but reinforced by the feedback. This meant adjusting the code Andrew had created to accommodate a extra columns of data from the figures.nz CSV files.

The group organised a meeting on Saturday, October 10th and distributed tasks to each achieve by that point:

- Andrew - Adjust the R code to include columns with 'Median Personal Income' and 'Household Ownership'.
- Oscar - Pull together a plan/ set of code to visualise the data wrangled.
- Torie - Merge the figures.nz dataframe with the scraped AKC data frame
- Helen - Creat a github repository and uploaded all data assembled so far, as well as other necessary documents so that all group members would have access.

# Group Meeting 3 10/10/20

The group met at Torie's flat to spend the best part of a Saturday afternoon on the project. The group had achieved the following in the few days since last meeting:

- Andrew had designed an R code that would take an input of a region folder consisting of figures.nz CSV data for each district and would combine the districts to get totals for each variable desired in the figures.nz CSV.

- Oscar had identified an R package with which an interactive map visualisation could be constructed for a New Zealand broken down into regions. However, after hours of research he determined that it was not possible to run one of the critical libraries - rgdal - on his computer system.

- Torie had merged the data but encountered a discrepancy between figures.nz data and AKC scraped data when it came to the dog breeds present. Specifically, there were some New Zealand breeds not mentioned in the top dog breeds data as it was scraped from an American website.

- Helen had created a Github repository and made sure everyone had access and everything was present in the repository. She also had begun attempting to download the necessary packages for data visualisation that Oscar had identified.

The group spent several hours together going over the assembled data, helping each other out with various problems, and organising future tasks. At the end of the meeting it was determined that the data visualisation would not be realistically achievable after each group member attempted downloading the package to no avail. Oscar had created and organised the figures.nz CSV files for each district into the appropriate regional folders ready for wrangling using Andrew's code. Helen and Andrew had further refined the R code for use on these folders, and also assisted Torie with a solution for merging the dog data between websites.

The AKC data for weight and height was presented in a non-uniform format with weight sometimes being represented between two values, this was also the case for height with difference in genders being stated eg. ‘Height 22.5-24.5 inches (male), 21.5-23.5 inches (female)’. Torie found it challenging to filter and alter the heights due to the formatting difference. We decided instead to focus on weight as the size measurement. Torie handled this wrangling.

Helen compiled a file of how dog breed names were represented within each data frame and which were missing. We found that New Zealand specific breeds were not represented in the AKC website such as the huntaway and heading dog. The heading dog was originally breed from Collies so we used collie temperament and weight to estimate its attributes (Differences can be found in the Breed_name_difference.csv. We also filtered out the AKC dog list down to the

dogs contained in the figures data and arranged them in the same order. Another challenge we found was the fact that some breeds were represented in more subspecies, for example, Collies in the figures data included border, smooth, rough and bearded Collies, whilst the AKC data only contained, a Collie, border and bearded breed. In this case we combined the column data for smooth and rough to Collie and matched it to the 'Collie' data from AKC.

The group distributed tasks and organised a future meeting date before the presentation. The tasks were distributed as follows:

- Andrew - Wrangle regions Taranaki, Tasman, Waikato, Wellington, West Coast
- Oscar - Wrangle regions Manawatu-Whanganui, Marlborough, Nelson, Northland, Otago, Southland
- Helen - Wrangle regions Auckland, BOP, Canterbury, Gisbourne, Hawke's Bay
- Torie - Get the Julia CSV exactly like the R CSV

## Group Meeting 4 12/10/20

The group met briefly following another lecture to go over what was achieved in each individual task, and what was left to do before the presentation.

Andrew, Oscar and Helen had all managed to upload a CSV data frame containing information on population, dog weight, total registered dogs, median personal income, house owned and house not owned for each region and for each breed of dog. Each individual had written code in R that further wrangled the regional data into a data frame containing information on all regions using the sum of each region's constituent districts.

Andrew then proceeded to stitch all these data frames together so that we had a final table with regional data for the entirety of New Zealand collated from figures.nz district data. This data was normalised using the population column. This population data was obtained from the 2013 census by figures.nz, meanwhile the dog data was obtained in 2019. In researching for an useable set of recent population data from the 2018 census, no population breakdown matched that of the figures.nz website, and so the 2013 census population data had to be used.

Torie and Helen had together managed to find a solution for the problems with the AKC data. This was to be done by manually matched up the appropriate dog breeds. It was determined, after a conversation with Thomas Li, that a function to do this job would be far more labour intensive than it is actually worth, and even he would say that it would be tricky to achieve. This knowledge confirmed to the group that manually adjusting the CSV was an appropriate measure.

In this group meeting we discussed what needed to be done for the presentation of data, and the roles we were to take in the presentation. Helen had also discovered an effective way to

graphically display an analysis of the wrangled data using an interactive bubble plot. The following tasks were distributed to each group member:

- Andrew - Presenting the data wrangling portion of the presentation and was to create powerpoint slides for this section.
- Oscar - Create the skeleton of the powerpoint to be used in the presentation, and would be presenting the websites used and ethical considerations for scraping these websites and was to create powerpoint slides for this section. Also, was to use the information ascertained by Helen for the bubble plot and write code/ create graphs for the presentation.
- Helen - Presenting the analysis of the wrangled data and conclusion of the project and was to create powerpoint slides for this section.
- Torie - Presenting the introductory portion of the presentation and was to create powerpoint slides for this section.

## Group Meeting 5 14/10/20

This was the final group meeting before the presentation to go over the powerpoint slides as a group and practice the presentation for time.

Going into the meeting Oscar had created a working bubble plot code that could be readily modified to compare different variables in the data frame. Helen and Torie had modified the dog breeds so that chihuahua and collie breeds were condensed. Torie had also made the decision to use the second most popular dog breed in any graphical analysis because the Labrador was the most popular in every region. Each group member had created slides ready for presentation.

At the meeting Andrew and Oscar worked together with the plot code Oscar had created to finalise the most interesting variable comparison in the graphs used before coming to a consensus on the graphs that would best display our work in the presentation.

The group practiced with a timer and helped each other out with ideas for the presentation. Once ready, the group left for the presentation itself.

# Lab Presentations 14/10/20

The group presented in lab in the following order:

1. Introduction - Torie
2. Data Sources and ethics - Oscar
3. Wrangling - Andrew
4. Analysis and conclusion - Helen

Following the presentation the group held a discussion with Giulio and it was decided to include a further analysis into the temperament of each dog breed for the final report.

The group set up a new time to meet and organise the report following a hectic week of assignment due dates for multiple courses.

# Group Meeting 6 19/10/20

The group met to assign roles for completing the full project report.

In the meeting everyone went through their respective code and began cleaning it up and annotating each step within the code. A group discussion was had as to how best to include an analysis of the dog breed temperament. The temperament was recorded by the top dogs website as three identifying characteristics of each breed (i.e. loyal, energetic, obedient). It was decided that a sentiment analysis of each breeds characteristics would be the best way to examine this data.

Tasks for the final report were distributed as follows:

- Andrew - Organise the R code used in wrangling the figures.nz data into a neat Jupyter notebook.
- Oscar - Organise the project diary into a presentable format and fill in any missed details. Also, perform a sentiment analysis using R on each dog breed's characeristics.
- Helen - Write the first draft report for the final project.
- Torie - Organise the Julia and R code used in wrangling the top dog breeds data into a neat Jupyter notebook.

# Group Meeting 7 22/10/20

The group met for a final time to finalise the report for submission.

During the days prior the group communicated online between each other so as to ensure each individual was performing their assigned task in the right manner, and to receive input or information from other group members for their respective tasks.

Andrew had successfully created a clean and annotated Jupyter notebook with all the R code used to wrangle data from the figures.nz CSV's. This concatenated code produced by Oscar and Helen as well as his own code, thus providing markers with all the information in one place.

Oscar had successfully ascertained an R package that could be used for sentiment analysis. He altered an existing data frame in R to append an extra column with sentiment scores for each dog breed. He also used the same graph function created earlier to perform a graphical analysis of the sentiment scores. Although there were no meaningful trends observed, this data was still included to show just that.

Torie had successfully created a clean and annotated Jupyter notebook with all the Julia and R code used to wrangle data from the top dog breeds website.

Helen had written a complete first draft of the final report document. She had also sent it out to everyone in the group prior to meeting so they would have time to edit and provide feedback in the group meeting.

In the meeting the final notebooks were run through and refined to improve clarity for the markers and ensure they performed as expected. Helen created a ReadMe file to make running the code even clearer. Edits were made to the final report as well as all supporting documentation to be included in the meeting and online over the next 24 hours. This diary was exported from endnote into word to submit as a PDF document.

The final project was submitted on Friday, October 3.