Group Members:
Andrew Doodson: 45329683
Helen Morris: 18427458
Oscar Evans: 21466194
Torie Owers: 34511575

Dog Gone: Wrangling American Kennel Club data with Figure NZ Registered Dog Data

I. Introduction & Data Sources

The purpose of the data collected was to provide an overview of dog ownership in New Zealand including breed specific statistics from multiple sources. Another focal point was understanding socioeconomic factors within a region that may affect dog ownership including, median personal income and home ownership. The data collected was from the Figures NZ website (*Figures NZ*, 2019) which has data for each district and region in New Zealand, this was collected to inform voters during the 2019 council elections. The second data source we decided to use was the American Kennel Club (AKC) frome their article "The Most Popular Dog Breeds of 2019" . The data we collected from this site was the ranking data of dog breeds, dog temperament, height and weight (*Staff, A*, 2020). The AKC data was not in a CSV or excel format to download, as such the data was scraped using Julia (*Bezanson, J.* 2017). Once scraping was complete we processed and wrangled the pets data from Figures and AKC dog breed information using R (*R Core Team*, 2013). When choosing these data sources we considered multiple factors including: the html format of AKC, ethical considerations of using the data from each site, comparability of each data source, and usefulness of the data.

In terms of ethical considerations for the data collection we used the Maori view of the world framework provided by Caleb Moses (*Moses, C.,* 2020), including whanaungatanga, rangatiratanga, manaakitanga, kotahitanga and kaitiakitanga. Whanaungatanga represents centrality of kinship; where did this data come from and from who. The figures NZ data was collated from a variety of data sources available to the public, their purpose "is to get the people of New Zealand using data to thrive", they wanted to make data more accessible to more New Zealanders. The American Kennel Clubs' main goal is to inform responsible dog ownership. As such the information is widely for anyone wanting to learn more about different breeds. Rangatiratanga addresses whether the power to use this data been granted by the right people. Figures NZ only publishes data that was already open to the public, or data an organisation wants published or found. Each sites' terms and conditions allow scraping and promote data sharing. With respect to kanaakitanga (extending love, compassion and hospitality), the data was used to give back interesting information back to the public on dog ownership within New Zealand. Regarding kotahitanga (power held by a collective), the findings we aimed to collate have collective benefits to the public as a whole and could be interesting for New Zealanders. Finally, kaitiakitanga (acting as a guardian over precious things), all data from Figures NZ held no personal information and during the wrangling process we aimed to present the data as accurately as it was collected.

When selecting AKC to web scrape we first inspected the elements of the ranked dogs to determine how easily the data could be scraped. The page had a uniform layout with specific CSS selectors for the dog name and rank. Furthermore, on each breed page the temperament, height and weight were similarly formatted with specific CSS selectors. This matched our skills in scraping and was chosen as our secondary source. The Figures NZ data was all collected in CSV format and all districts and regions were collected with temporal considerations in place. All data was from 2019 apart from census data, with 2013 being the most recent census population data included in the Figures NZ data. Although there were breed differences in terms of naming and missing information on New Zealand specific breed, we determined the two data sources would be comparable for the purpose of the final database. The data is useful in providing an overview of New Zealand's most

Group Members:
Andrew Doodson: 45329683
Helen Morris: 18427458
Oscar Evans: 21466194
Torie Owers: 34511575

owned dog per capita by region. An interesting use of the data was finding limiting factors involved in owning a dog in New Zealand such as income and home ownership.

II. Wrangling & Challenges

i. Figures CSV Wrangling

Figures NZ data was presented in 67 CSV files, one for each district and region in New Zealand. It was decided to combine the CSV data by 16 regions as defined by the 2013 census data (Table 1.). Each CSV contained a section on pets. The CSVs contained data including demographics, income, households, wellbeing, health, education, work, work safety, transport, gambling, agriculture, election and council finances. These CSV files for the districts were then sorted into folders based on what region of New Zealand they would fall under. Our initial notebook (Notebook 1 > 'Get Info From Figures NZ.ipynb') was then made to read through these folders and extract the information required from each CSV. Our first function in the aforementioned notebook iterated through each CSV in a folder and extracts the "Pets – Total Registered Dogs" for each CSV, then binds each iteration to a blank data frame that was initially created and then supplied to the function. Three similar functions were used to extract "Demographics – Population at 2013 Census", "Households – Home Ownership by households" and "Income – Median personal income". For pet's data we now had counts for each dog for each district in that region, this required use of the aggregate function to combine all district total registered dog counts into a single sum of each region (using the compress_pets function). The aggregate functions were used additionally for home ownership, and population. The aggregate function was also used for median personal income, this time however instead of taking a sum of all districts in a region we used it to take the mean value for each region. The columns were appropriately renamed to better reflect the data. Category was renamed to Breed and Value to Region. Each region was combined into a single data frame (Notebook 1 > 'All_Pets_FiguresNZ.csv') showing the total counts of each breed in all regions. Home ownership was summarised in a separate data frame (Notebook 1 > "All_Homes_FiguresNZ.csv"), income (Notebook 1 > "All_Incomes_FiguresNZ.csv") and population (Notebook 1 > "All_Populations_FiguresNZ.csv"). Below in Table 2 is a description of each R package used throughout the wrangling process using R, these packages were used across multiple notebooks and were appropriately loaded in when required.

Region	Districts Included
Northland	Far North, Whangarei, Kaipara
Auckland	Auckland
Waikato	Thames Coromandel, Hauraki, Waikato, Matamata-Piako, Hamilton City, Waipa, Otorohanga, South Waikato, Waitomo, Taupo
Bay of Plenty	Western Bay of Plenty, Tauranga, Kawerau, Whakatane, Opotoki, Rotorua
Gisborne	Gisborne
Hawkes Bay	Wairoa, Hastings, Napier City, Central Hawkes Bay
Taranaki	New Plymouth, Stratford, South Taranaki

Group Members:
Andrew Doodson: 45329683
Helen Morris: 18427458
Oscar Evans: 21466194
Torie Owers: 34511575

Manawatu- Whanganui	Ruapehu, Rangitikei, Manawatu, Tararua, Palmerston North City, Horowhenua
Wellington	Masterton, Carterton, South Wairarapa, Upper Hutt City, Porirua City, Lower Hutt City, Wellington, Kapiti Coast
Nelson	Nelson City
Marlborough	Marlborough
Tasman	Tasman
West Coast	Buller, Grey, Westland
Canterbury	Hurunui, Selwyn, Waimakariri, Christchurch City, Banks Peninsula, Ashburton, Timaru, Mackenzie, Waitaki, Waimate, Chatham Islands, Kaikoura
Otago	Queenstown Lakes, Central Otago, Dunedin City, Clutha
Southland	Southland, Gore, Invercargill
Table 1. Distribution of districts within the final regions used for data wrangling	

R Packages	Brief Description
tidyverse	Group of packages specifically used in data wrangling, included dplyr, purrr, stringr and more. *Wickham, H. et al*., (2019).
janitor	Was used to transpose columns to rows *Firke, S* (2020)
hablar	Retype functions allows user to easily change from type not useful for analysis e.g. char to int. *Sjoberg, D.* (2020)
rvest	
magrittr	Offers operators that make the code more readable by structuring operations left to right. *Milton Bache, S. & Wickham, H.* (2014).
tidyr	Helps create tidy data, has gather and spread functions to go from long to wide data frames *Wickham, H.* (2020)
Table 2. R packages used for wrangling figures and AKC data frames	

In order to have data that is comparable and is useful for analysis we used the population data retrieved to determine pets per capita and homes per capita. In Notebook 2 > "Getting Per Capita Information.ipynb", using the data frames produced in the first Notebook we used the mutate function and divided each dog count per region by the population in that region. The same process was followed for the Homes per capita data frame.

ii. American Kennel Club Wrangling

One of the first steps was to scrape the required data from the AKC website. This was done using Julia, the code for this is found in the Notebook 4 > 'Julia_Scraping.ipnyb' notebook. The packages used to do this scraping are detailed in Table 3. The first step in scraping involved requesting the page using the HTTP package (Table 3.). The body of the kennel page is then parsed using the Gumbo package. For each breed the CSS selector was identified as ".in-cell-link", the function

'eachmatch' was then used to retrieve each node pertaining to the specified selector. Next using 'getattr()' allowed us to iterate through the parsed HTML code to extract all URLs with the 'href' attribute. This created our list of breeds, using the created list of URLs and HTML format on each breed page we selected for breed height, weight and temperament using our own functions.

Julia Package	Brief Description
HTTP	Package used for scraping websites in Julia (*J.*, 2016)
Gumbo	A Julia wrapper around Googles gumbo library for parsing HTML (*J.,* 2015)
Cascadia	CSS selector library, use the parsed HTML string to create a selector from a string, then use each match to get nodes within the document(*A.,* 2015)

Table 3. Packages used in Julia web scraping of the AKC website

The data frame produced from the AKC website included the page URL, Breed, height, weight and temperament for 193 dog breeds. The breeds present in AKC were not all present in the Figures NZ data, as such we filtered the data frame to represent all the breeds that were present in the Figures NZ data frame additionally ordering them to reflect order in the Figures NZ data frames. The differences in breed names was manually compared between each site and collected in Notebook 3 > "Breed_Differences.csv". There were a lot of differences in the way a breed was named across both sites e.g. Collie, Border vs Border Collie in the Figures NZ and AKC data frames respectively. This was challenging in terms of joining the AKC and Figures NZ website. We opted to manually go through and rename the AKC breed names to match those in figures NZ. We had a couple of ideas on how to do this using code, including the package "R Selenium" or creating a list that could be iterated through the breed names to match names that were the same. Unfortunately we were unable to execute this successfully and manually updated the names in the AKC data frame to match the breeds in figures NZ (match in the sense of formatted the same way, as document in the 'Breed_Diffrences.csv', the same breeds were written differently in each data frame).

iii. Final Data Frames & Limitations

Our final step was to combine dog breeds to be comparable to the AKC data frame. A variety of breeds were not represented over each data frame. The heading and huntaway dogs are NZ specific breeds that were not represented in the AKC data frame were approximated with similar breeds, e.g. Heading dogs were bred from collies. Another factor of the figures data was the over-representation of mastiff breeds, such as *Dogo argentino*, *Japanese tosa* and *Perro de presa canario*, these dogs are classed as menacing breeds in New Zealand. The CSV data also included information on menacing dogs which could be why it was overrepresented although most had zero counts in most regions. As such the breeds *Japanese tosa* and *Perro de presa canario* were filtered out due to neither breed having registered dogs in any of the regions.

Notebook 5 > 'Breed_Combining_DF.ipynb' was preparation for merging the AKC data with figures NZ, this involved removing any leading and trailing whitespace from the breed name, filtering out breeds not present in Figures NZ and ordering them in a similar position. As mentioned above it was a struggle to combine these two data frames, we manually changed the names in AKC to match

Group Members:
Andrew Doodson: 45329683
Helen Morris: 18427458
Oscar Evans: 21466194
Torie Owers: 34511575

those in the figures NZ. Finally, 'left_join' was used to insert AKC breed weight into the figures data frame. Furthermore, median personal income and homes owned per capita were joined to the data frame.

The below diagram (Fig. 1.) represents the final data frames as relational data. From here we are able to investigate different relationships between the data frames. With all the data needed from each data source our final step involved joining the data frames for analysis. The main interest with the data we collected is comparing the types of breeds in each regions and how this might be related to home ownership or median personal income.

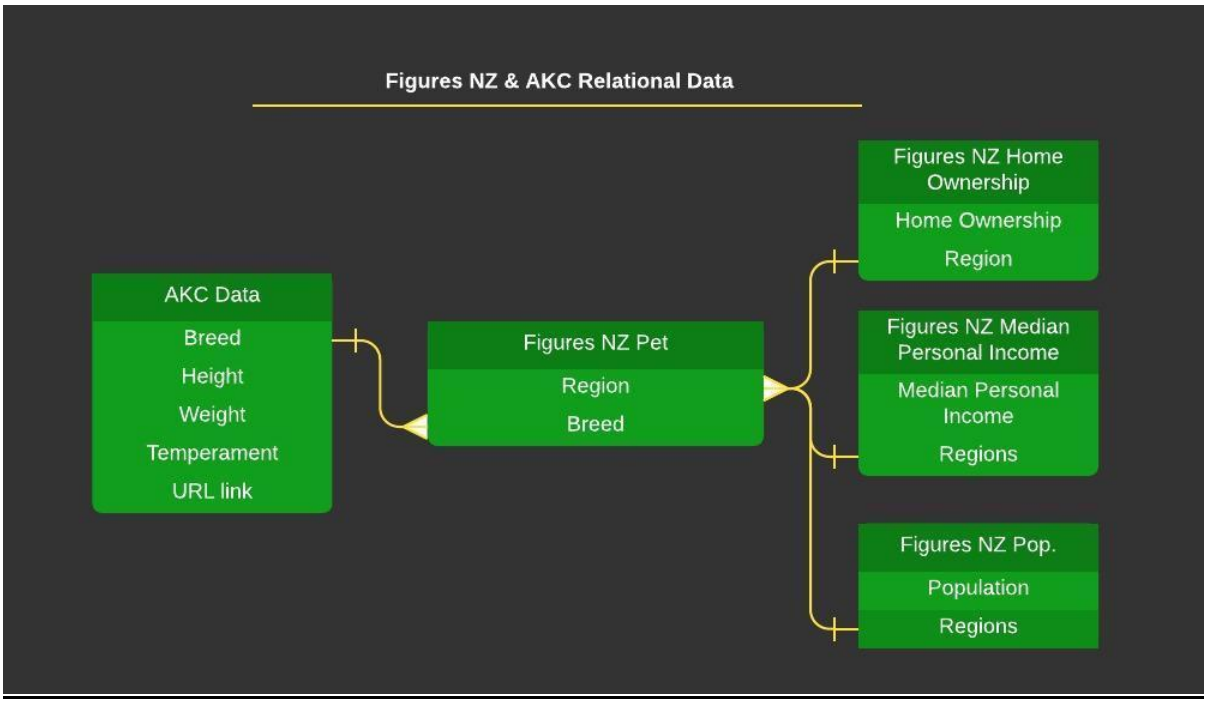

Fig. 1. Relational data representation of data collected from the figures NZ website and American kennel club. Figure generated using

For analysis we performed joins between the Figures pet per capita data, home ownership per capita and median personal income, along with the weight of the dog from the AKC data frame. With this comparisons between regional differences in dog ownership will be easier to perform.

Some of the limitations in the data included missing information on a lot of popular breeds in New Zealand (such as Dachshunds and Great Danes, neither represented in the total registered dogs). Hence, the data gives a general view of dog ownership in New Zealand due to gaps in the data. There is a National Dog Database which may have a more complete image of dogs registered in New Zealand, however this requires special access to the database, which we could not obtain. Another issue that became apparent was the difference in the New Zealand specific breeds – i.e. Huntaway and Heading, that were not represented in the AKC data frame. As a way to mitigate this we investigated each breed to find a similar approximation, the Heading dog was bred from collies as such we used the collie characteristics to represent the Heading dog.

Group Members:
Andrew Doodson: 45329683
Helen Morris: 18427458
Oscar Evans: 21466194
Torie Owers: 34511575

One of our biggest considerations was choosing to use the population data included in the Figures NZ CSVs which was collected during the 2013 census. The most recent census, 2018, had considerable issues with collection of data with a much lower percentage of participants. According to *Stats NZ*. (2018) there was a mix in terms of data quality, with very high, high, poor and very poor quality. As figures NZ only provided 2013 census data but 2019 data for all non-census data this could be a result of the poor quality of the 2018 census data. Hence we opted to use the 2013 census data provided through figures NZ.

III. Analysis

Using our final data frames a brief analysis was performed. This was carried out using the R plotly (*Plotly*, 2020) package which is used to produce interactive plots. The interactive bubble plots allows the user to plot using five separate variables (Notebook 6 > "Project_Bubble_Graph.ipnyb") in Fig 2. A comparison between dogs per capita and home ownership per capita were compared in each region (represented as separate colours) with weight of the second most popular dog (size of bubble). The second most popular dog was used as the most popular dog in all regions is the Labrador Retriever. This produced an upward trend in the number of dogs per capita with an increase in number of houses owned in a region. This suggests a possible correlation between the two factors, although more robust analysis would need to be done to obtain a clear picture of the relationship.

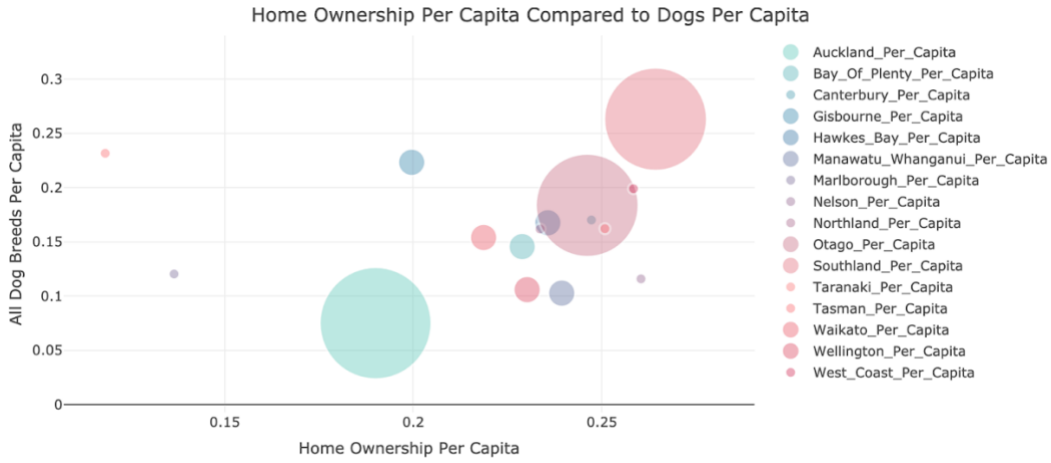

Fig 2. *Home ownership per capita compared to dogs per capita, size of bubble indicates the weight of the dog, along with regions plotted by colour.*

Group Members:
Andrew Doodson: 45329683
Helen Morris: 18427458
Oscar Evans: 21466194
Torie Owers: 34511575

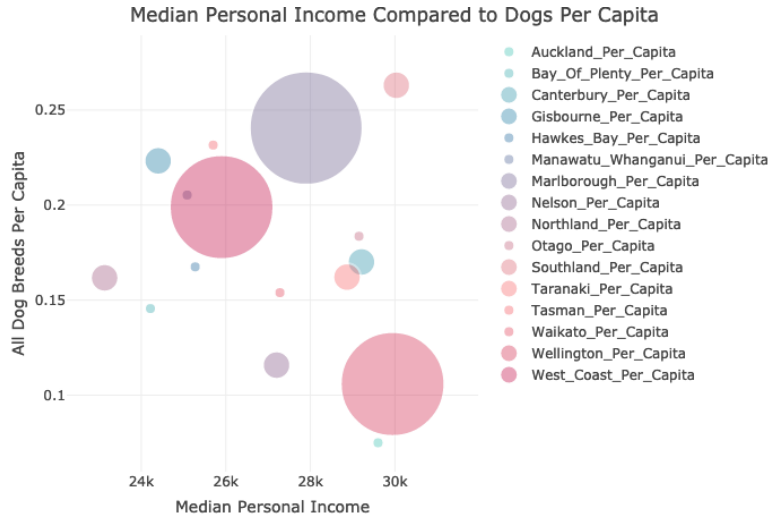

Fig 3. *Median personal income compared to dogs per capita, size of bubble indicates the weight of the dog, along with regions plotted by colour.*

The final plot used the same variables as *Fig.2* but replaced home ownership per capita with the median personal income. This resulted in *Fig 3*. where not much of a trend was found between any of the variables plotted.

Finally, we performed sentiment analysis on the temperament words assigned to each breed. This was carried out using the R package syuzhet (*Jockers, ML.* 2015). A higher positive scores indicates a more positive sentiment score. As none of the dogs were described with negative attributes all scores were positive, although there was still significant differences how positively some breeds were viewed. A summed score of each word was the final output, when plotting there were no significant trends in sentiment scores.

Rank	Weight	Temperament	Breed	Sentiment Score
7	20	Friendly, Curious, Merry	Beagle	2.15
43	12	Playful, Curious, Peppy	Bichon Frise	2.4
14	65	Bright, Fun-Loving, Active	Boxer	2.25
55	35	Alert, Curious, Pleasant	Australian Cattle Dog	1.15
35	6	Charming, Graceful, Sassy	Chihuahua	1.75
136	45	Smart, Bouncy, Charismatic	Bearded Collie	1.5
33	30	Affectionate, Smart, Energetic	Border Collie	1.75
38	60	Devoted, Graceful, Proud	Collie	2.3
162	65	Gentle, Independent, Noble	Greyhound	1.75
36	7	Gentle, Playful, Charming	Maltese	3
6	60	Active, Proud, Very Smart	Poodle (Standard)	1.5
3	65	Friendly, Intelligent, Devoted	Golden Retriever	2.55
1	65	Friendly, Active, Outgoing	Labrador Retriever	1

Group Members:
Andrew Doodson: 45329683
Helen Morris: 18427458
Oscar Evans: 21466194
Torie Owers: 34511575

8	95	Loyal, Loving, Confident Guardian	Rottweiler	2.65
18	11	Friendly, Smart, Obedient	Miniature Schnauzer	1.25
2	65	Confident, Courageous, Smart	German Shepherd Dog	2.3
20	9	Affectionate, Playful, Outgoing	Shih Tzu	1.5
16	13	Affectionate, Gentle, Graceful	Cavalier King Charles Spaniel	2.25
28	25	Gentle, Smart, Happy	Cocker Spaniel	2.25
27	50	Friendly, Playful, Obedient	English Springer Spaniel	1.75
61	50	Playful, Charming, Mischievous	Bull Terrier	1.5
122	18	Friendly, Independent, Amusing	Smooth Fox Terrier	1.25
75	9	Alert, Inquisitive, Lively	Russell Terrier	1
82	28	Clever, Brave, Tenacious	Staffordshire Bull Terrier	1.75
44	15	Loyal, Happy, Entertaining	West Highland White Terrier	2.5

Table 4. Sentiment analysis performed using the R package 'syuzhet' on the temperaments scraped from American Kennel Club (AKC).

IV. Conclusions & Future Work

Although the Figures NZ data came in a tidy CSV format it still provided wrangling challenges that were unforeseen, including the challenge of removing redundant empty information. Additionally, mutating columns to match up with breeds in the AKC data frame. Such as the Rough and Smooth Collies, neither of which were present in the AKC data frame. The final challenge we did not manage to complete was joining the two frames with a created function or other packages instead opting to mechanically rename the breeds. It was found that Labradors were the most common breed in New Zealand and that home ownership (Fig. 2) may be a factor in dog ownership. Southland had the highest number of dogs per capita with 0.26 followed closely by Tasman with 0.23.

Future work may include finding some New Zealand specific breed information to include more precise data from the Heading and Huntaway breeds. Gaining access to the National Dog Database would provide a more complete view of dog ownership in New Zealand. There are many interesting factors to consider; the Figures NZ data also included wellbeing data such as loneliness statistics which could be an interesting comparison, as well as examining whether a higher number of dogs per capita is linked to a lower percentage of reported loneliness. A further interesting factor to consider could be the use of working dogs within New Zealand. Labradors for instance are often used as seeing eye dogs, beagles as police sniffer dogs and huntaway and heading dogs as sheep herding dogs. This project was an illuminating exercise in collaboratively collecting and managing data, and there is certainly more work we can achieve regarding this in the future.

Group Members:
Andrew Doodson: 45329683
Helen Morris: 18427458
Oscar Evans: 21466194
Torie Owers: 34511575

V. Reference

Note On References: R packages were cited using the citation function in R, a similar Julia function was not found, hence the github URL for the corresponding packages was used for citation. APA referencing.

*A., (*2015). *Algocircle/Cascadia.jl*. GitHub. https://github.com/Algocircle/Cascadia.jl

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, *59*(1), 65–98. https://doi.org/10.1137/141000671

Customer update on data quality of 2018 Census | Stats NZ. (2018). Stats NZ. https://www.stats.govt.nz/news/customer-update-on-data-quality-of-2018-census

Figures NZ. (2019). Figures NZ. https://places.figure.nz/

Firke, S. (2020). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.0.1. https://CRAN.R-project.org/package=janitor

J. (2016). *JuliaWeb/HTTP.jl*. GitHub. https://github.com/JuliaWeb/HTTP.jl

J. (2015). *JuliaWeb/Gumbo.jl*. GitHub. https://github.com/JuliaWeb/Gumbo.jl

Jockers, ML. (2015). Syuzhet: Extract Sentiment and Plot Arcs from Text, https://github.com/mjockers/syuzhet

Milton Bache, S. & Wickham, H. (2014). magrittr: A Forward-Pipe Operator for R. R package version 1.5. https://CRAN.R-project.org/package=magrittr

Moses, C (2020). Māori Data Sovereignty [power point].

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Sievert, C. (2020). Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC Florida

Sjoberg, D. (2020). hablar: Non-Astonishing Results in R. R package version 0.3.0. https://CRAN.R-project.org/package=hablar

Staff, A. (2020, May 1). *The Most Popular Dog Breeds of 2019*. American Kennel Club. https://www.akc.org/expert-advice/dog-breeds/2020-popular-breeds-2019/

Wickham, H. et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Wickham, H. (2020). tidyr: Tidy Messy Data. R package version 1.1.2. https://CRAN.R-project.org/package=tidyr

Group Members:
Andrew Doodson: 45329683
Helen Morris: 18427458
Oscar Evans: 21466194
Torie Owers: 34511575