**DATA422- Group Project**

**Project Diary**

Group Members:

Andrew Doodson:

Helen Morris: 18427458

Oscar Evans:

Torie Owers:

Initially we all worked together to identify an interesting data source. Our ideas varied from: Use Spotify to find the most popular artist in terms of number of streams and compare to other music ratings, looking at Accident Compensation Corporation (ACC) data to determine which region in New Zealand was the clumsiest and using space flight data (<https://nextspaceflight.com/>).

We settled on looking into ACC data, comparing different types of accidents, such as E-scooters, hot water bottle, falls and mobility scooter accidents. We thought we could use this data to compare to a variety of statistics such as sunshine hours and GDP information:

* + <https://catalogue.data.govt.nz/dataset?q=ACC&sort=score+desc%2C+metadata_modified+desc>
  + <https://catalogue.data.govt.nz/dataset/e-scooter-injuries/resource/aa5837a6-b743-499f-8e2e-ca1e0f3bbc1c>
  + [Sports injuries - data.govt.nz - discover and use data](https://catalogue.data.govt.nz/dataset/sports-injuries)
  + [Motor Vehicle Accident claims - data.govt.nz - discover and use data](https://catalogue.data.govt.nz/dataset/motor-vehicle-accident-claims)
  + <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>

Each of us focused on a specific dataset, Torie: Fall data, Helen: Hot Water Bottle, Andrew: E-scooters and Oscar: Mobility Scooters. We each extracted the data for each of these datasets into data frames using R. However, we realised there was no uniform measurement of years which was going to be a primary key, there was a variety with financial year, calendar year and monthly reports.

We found another site which summarises a variety of data per region and district in New Zealand (<https://places.figure.nz/>). In investigating the data, we decided to focus on the pet’s data section which lists a variety of statistics for each region, including de-sexed registered dog by breed, ACC claims, registered dogs by sex and total registered dogs. The data was available in a CSV format; however, each CSV included all data for that region including Demographics, Income, Households, Wellbeing, Health, Education, Work, Work safety, Transport, Gambling, Agriculture, Election and Council Finances. Andrew worked on writing an R script to filter out the unnecessary information for each region.

Additionally, the regions were split into 67 separate districts and regions. 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 and Southland. Oscar worked on assigning the regions and districts to the 14 regions aforementioned.

As this data source is quite clean it will not require a huge amount of data wrangling, therefore we decided to compare the regional NZ data to this breeding site <https://www.akc.org/expert-advice/dog-breeds/2020-popular-breeds-2019/>

Which lists the most popular dogs and attributes of each breed such as temperament and size. We decided these would be interesting to compare to different regions, which region prefers smaller dogs and who likes playful vs stoic dogs. Torie, is working on how to scrape this site using Julia, although ran into some issues.

We had been keeping a rough journal prior to this, a more concise summary of work to date was written up in this document. Additionally, we set up a GitHub repository (<https://github.com/HelionPeros/Dog_Gone>) for future work and code to be tracked. Using Andrews R script with the downloaded CSV files from <https://places.figure.nz/>, we began to transport the Pet CSV data for each region to a Jupyter Notebook. Helen worked on these items.