Indicators of unemployment

36103 - STATISITCAL THINKING FOR DATA SCIENCE Assessment Task 2 – Part A, Project Proposal

28 April 2019

Group Name – Dole Bludgers? lol PLEASE THINK OF SOMETHING BETTER!!!!

Group Members - Htet Naing Aung 13385531, Susannah Gynther, Robert Kell, William Kent 13285337, James Tesoriero, Reasmey Tith 10845345, Xiaojun Zeng

# Project Proposal

# Indicators of unemployment

Rationale

The Australian Bureau of Statistics (ABS) uses the internationally agreed standards in defining unemployment. To be classified as unemployed a person needs to meet the following three criteria (Abs.gov.au, 2019):

* Not working more than one hour in the reference week;
* Actively looking for work in previous four weeks; and
* Be available to start work in the reference week.

Gleeson (2019) states that unemployment impacts on the economic, social and mental health of not only the person who is unemployed but their family and community in the short-term and can have impact for decades to come. Additionally, the longer a person remains unemployed the more difficult it can become to find employment as skills and abilities deteriorate over the time of unemployment. Hudson (2019) found that unemployment can cause a ripple effect across the economy. As the proportion of unemployed persons increases, less tax is collected, and government spending will rise accordingly to pay more in unemployment benefits, affecting the ongoing financial stability of the economy.

Research has revealed several misnomers about unemployment and resulting factors such as crime and domestic violence. Janko and Popli (2015) in their analysis of Canadian data showed that there was no relationship between unemployment and crime. Another study showed that gender-based unemployment played a part in the increase of domestic violence although did not increase domestic violence over all (Anderberg et al, 2013).

Our research aims to broaden the scope for factors that could affect unemployment in NSW. The unemployment rate in Australia for March 2019 is 5.0% (Tradingeconomics.com, 2019), while in NSW it was 4.3% for the same period (Taffa, 2019). To maintain NSW’s low unemployment and resulting high prosperity, detailed and region specific research can potentially unlock hidden characteristics about each area and if they impact unemployment rates. The outcomes of our research are targeted towards governmental policy makers and social welfare groups in NSW. Any new information can be used to assess existing services and their effectiveness, as well as highlight new areas where more services are needed.

**Our questions**

What factors predict unemployment rates in New South Wales? Of these factors, are there any that are unique or unexpected? What social demographics are related to unemployment? Does level of education affect unemployment and if so what level is the highest contributor?

Data Sources

This analysis will bring together a range of data sources and information covering geographical, educational and biographical data. Data will predominantly be obtained from Australian State and Federal agencies and departments.

Unemployment figures will be obtained from the Australian Government Department of Jobs and Small Business, SA2 Data tables — Small Area Labour Markets, December quarter 2018 (Should we be using the Census 2016 figures?). This research will further explore data from the ABS Census of Population and Housing 2016 (Census 2016) and a range of datasets from other areas of the ABS and non-ABS sources summarised in Table 1.

Table 1 – Sources of Data

|  |  |  |  |
| --- | --- | --- | --- |
| Indicator/variable | Description of dataset | Source | Geographic Level |
| Unemployment  Will to fill in |  | ABS | SA2 |
| Socio economic status | Socio- economic indexes for areas (SEIFA) 2016 | ABS | SA2 |
| Crime Statistics | Annual incident counts, rates per 100,000 population and ranks for selected offences (2011-2018) | Bureau of Crime and Statistics research (BOSCAR), NSW Department of Justice | LGA |
| Drug and alcohol? Probably leave this out for now |  |  |  |
| Education Level | Census of Population and Housing: Reflecting Australia - Stories from the Census, 2016 | NSW Department of Education?  ABS | SA2 |
| Demographics  Age  Gender  Dwelling type  Household composition | Estimated Resident Population (ERP) by SA2 (ASGS 2016) Age and Sex, 2001 Onwards | ABS | SA2 |
| Justin to split above section |  |  |  |
| Race and Ethnic background | Ancestry, Migrant arrivals, Place of birth, Indigenous population | ABS | SA2 |
| Air quality readings | Site Air Quality Index | NSW Office of Environment and Heritage | Reading station |
| Access to employment? Can delete this |  |  |  |
| Family Status | Marriages and divorces, Australia, 2016 | ABS | SA2 |
| Commute to work | Census of Population and Housing: Commuting to Work - More Stories from the Census, 2016 | ABS | SA2 |
| Access to Green Space | NSW Mesh blocks ASGS Edition 2016 | ABS | Mesh Block |

This project will collate data at the Australian Statistical Geographic Standard (ASGS) Statistical Area 2 (SA2) level. An SA2 has an average population of 10,000 people and can include one or more related suburbs that interact socially and economically (Abs.gov.au, 2018).

Using data at the SA2 level will allow analysis of over 570 distinct geographical areas in NSW. This will allow interpretation of any trends found to answer the research question. Data from the year 2016 was chosen as it was the year with the most data available. Coding examples of how we have acquired and merged our data are included in the appendices.

Modelling

An individual that wants to work is either employed or not employed; this is a binary outcome. As such, as part of this project, a multivariate logistic regression on grouped data will be performed in order to help answer the research question.

The unemployment data is a proportion of the population that is unemployed for a particular SA2, which is the total number of unemployed over the population as a percentage.

Issues

Issues currently experienced and expected include:

* **Diverse datasets** – Datasets have been gathered in a variety of formats making merging more difficult.
* **Level of Granularity** – Data is not always held at the SA2 level. It may be captured at a more granular level and will need to be rolled up, or data may be captured at larger geographical areas and need to be broken down to the SA2 level.
* **Data inconsistencies** – Data from government agencies is often input by humans with different levels of understanding and standards with can lead to data inconsistencies.
* **Data Reliability** – For some predictors the data will rely heavily on the ABS Census 2016. The census was completed by individuals who may provide false or non-sensical answers due to not understanding the question, systems and staff failed to interpret an answer, or an individual did not want to provide an answer due to data privacy concerns.

Summary

This research aims to provide new information about the factors affecting unemployment in NSW. It is hoped that this additional information will aid those making decisions that impact the provision of support and services for the unemployed, as well as better help those involved in improving employment prospects.

References

Abs.gov.au. (2019). *6105.0 - Australian Labour Market Statistics, July 2014*. [online] Available at: http://www.abs.gov.au/ausstats/abs@.nsf/products/FBE517ECA9B07F63CA257D0E001AC7D4 [Accessed 26 Apr. 2019].

Crowe, L. and Butterworth, P. (2016). The role of financial hardship, mastery and social support in the association between employment status and depression: results from an Australian longitudinal cohort study. *BMJ Open*, [online] 6(5), p.e009834. Available at: https://bmjopen.bmj.com/content/6/5/e009834?utm\_source=trendmd&utm\_medium=cpc&utm\_campaign=jnis&trendmd-shared=1&utm\_term=TrendMDPhase4&utm\_content=Journalcontent.

Anderberg, D., Rainer, H., Wadsworth, J. and Wilson, T. (2013). *Unemployment and Domestic Violence: Theory and Evidence (Discussion Paper 7515)*. [online] Ftp.iza.org. Available at: http://ftp.iza.org/dp7515.pdf [Accessed 26 Apr. 2019].

Gleeson PhD, P. (2019). *The Overall Effects of Unemployment*. [online] Smallbusiness.chron.com. Available at: https://smallbusiness.chron.com/overall-effects-unemployment-37104.html [Accessed 26 Apr. 2019].

Hudson, P. (2019). *How Unemployment Rates Affect The Economy*. [online] Elite Daily. Available at: https://www.elitedaily.com/news/business/how-unemployment-rates-affect-the-economy [Accessed 26 Apr. 2019].

Taffa, V. (2019). *New South Wales Unemployment Rate At 4.3 % March 2019 | The Southern Thunderer*. [online] Southernthunderer.com.au. Available at: https://www.southernthunderer.com.au/new-south-wales-unemployment-rate-at-4-3-march-2019/ [Accessed 26 Apr. 2019].

Tradingeconomics.com. (2019). *Australia Unemployment Rate | 2019 | Data | Chart | Calendar | Forecast*. [online] Available at: https://tradingeconomics.com/australia/unemployment-rate [Accessed 26 Apr. 2019].

Abs.gov.au (2018). *Australian Statistical Geography Standard (ASGS),* July 2018.[online] Available at: https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Australian+Statistical+Geography+ Standard+(ASGS) [Accessed 24 Apr. 2019]

## Hanrahan, C. (2017), *Census results are out, but can we trust the data?,* [online] ABC News, Available at: https://www.abc.net.au/news/2017-06-26/census-results-are-coming-out-can-we-trust-them/8594132 [Accessed 27 April 2019]

Appendices

**Appendix 1 - Example of Australian Bureau of Statistics Census 2016 data extract R code**

|  |
| --- |
| **library**(rsdmx)  **library**(tidyverse)  *# getwd()*  **setwd**(**dirname**(rstudioapi**::getActiveDocumentContext**()**$**path))  *# Check to make sure the ABS folder is available*  *# and, if not, create it. Saving file to right*  *# location will fail without the required folder*  **if** (**!dir.exists**("../Data Files/ABS")) {  **create.dir**("../Data Files/ABS")  }  *# Get the ABS Census 2016 Data on Dwelling Type*  dwelling\_data <- **as.data.frame**(**readSDMX**(providerId = "ABS",  resource = "data", flowRef = "ABS\_C16\_T24\_SA",  key = "TOT.TOT+11+21+22+31+32+33+34+91+92+93+94+Z+NA.0+1+2+3+4+5+6+7+8+9.SA2",  key.mode = "SDMX", start = 2016, end = 2016))  **summary**(dwelling\_data)  **head**(dwelling\_data)  **str**(dwelling\_data)  *# MISSING 9 SA2 Codes*  dwelling\_data **%>% distinct**(ASGS\_2016)  *# Distinct dimension values*  dwelling\_data **%>% distinct**(DWTD\_2016)  ## Retrieve Metadata to help with decoding values.  ds\_url = "http://stat.data.abs.gov.au/restsdmx/sdmx.ashx/GetDataStructure/ABS\_C16\_T24\_SA"  dataStructure <- **readSDMX**(ds\_url)  codeList <- **slot**(dataStructure, "codelists")  *# Dwelling Type*  dwelling\_type <- **as.data.frame**(codeList, codelistId = "CL\_ABS\_C16\_T24\_SA\_STRD\_2016")  *# Get Required Data and put in meaningful*  *# descriptions*  dwelling\_data\_final <- dwelling\_data **%>% inner\_join**(dwelling\_type,  by = **c**(STRD\_2016 = "id")) **%>% select**(SA2\_CODE = ASGS\_2016,  DWELLING\_TYPE = label.en, obsValue)  *# getwd()*  **write\_csv**(dwelling\_data\_final, "../Data Files/ABS/Dwelling\_Type\_SA2\_2016.csv") |

**Appendix 2 - Example of NSW Government Air Quality data download using R**

|  |
| --- |
| *# getwd()*  **setwd**(**dirname**(rstudioapi**::getActiveDocumentContext**()**$**path))  **if** (**!dir.exists**("../Data Files/NSWGovt/")) {  **dir.create**("../Data Files/NSWGovt/")  }  *## Download NSW Air Quality File if it doesn't*  *## already exist*  **if** (**!file.exists**("../Data Files/NSWGovt/AirQuality\_Data.xls")) {  aq = "https://airquality.environment.nsw.gov.au/aquisnetnswphp/tmp/tmp\_table\_21553\_1555911469.xls"  **download.file**(aq, destfile = "../Data Files/NSWGovt/AirQuality\_Data.xls",  mode = "wb")  }  *## Download NSW Air Quality Stations if it doesn't*  *## already exist*  **if** (**!file.exists**("../Data Files/NSWGovt/AirQuality\_Station\_Data.xlsx")) {  stations = **paste0**("https://datasets.seed.nsw.gov.au/dataset/",  "ee5fd225-ab54-49c4-8c91-930219018cd0/resource/",  "e09a1918-af2b-4375-ad04-00fabce72a10/download/",  "air-quality-monitoring-sites-summary.xlsx")  **download.file**(stations, destfile = "../Data Files/NSWGovt/AirQuality\_Stations\_Data.xlsx",  mode = "wb")  } |

**Appendix 3 - Example of Australian Bureau of Statistics Socio-Economic Indexes for Areas**

**data extract using R**

|  |
| --- |
| **library**(rsdmx)  *# getwd()*  **setwd**(**dirname**(rstudioapi**::getActiveDocumentContext**()**$**path))  **if** (**!dir.exists**("../Data Files/ABS/")) {  **dir.create**("../Data Files/ABS/")  }  data <- **as.data.frame**(**readSDMX**(providerId = "ABS",  resource = "data", flowRef = "ABS\_SEIFA2016\_SA2",  key.mode = "SDMX", start = 2016, end = 2016))  **write.csv**(data, "../Data Files/ABS/SEIFA\_2016\_Data.csv") |

**Appendix 4 - Example of Australian Bureau of Statistics Census 2016 data cleaning using R**

**(Dwelling Type)**

|  |
| --- |
| **library**(tidyverse)  **library**(data.table)  *# getwd()*  **setwd**(**dirname**(rstudioapi**::getActiveDocumentContext**()**$**path))  ## Read the raw data csv  raw\_data <- **read.csv**("../../Raw Data/Data Files/ABS/Dwelling\_Type\_SA2\_2016.csv",  quote = "\"")  **head**(raw\_data)  **str**(raw\_data)  *# Clean the data - Band Dwelling Type and*  *# create percentages - Note Total is not*  *# always the sum of the breakdown*  clean\_data <- raw\_data **%>% filter**(DWELLING\_TYPE **!=**  "Total") **%>% mutate**(DWELLING\_BAND = **case\_when**(DWELLING\_TYPE **==**  "Separate house " **~** "DWELLING\_HOUSE",  DWELLING\_TYPE **%like%** "Semi-detached, row or terrace house" **~**  "DWELLING\_SEMI", DWELLING\_TYPE **%like%**  "Flat or apartment" **~** "DWELLING\_FLAT",  DWELLING\_TYPE **%like%** "House or flat attached to a shop" **~**  "DWELLING\_FLAT", TRUE **~** "DWELLING\_OTHER")) **%>%**  **select**(SA2\_CODE, DWELLING\_BAND, obsValue) **%>%**  **group\_by**(SA2\_CODE, DWELLING\_BAND) **%>%**  **summarise**(Total\_Value = **sum**(obsValue)) **%>%**  **spread**(DWELLING\_BAND, Total\_Value) **%>%**  **mutate**(PERC\_DWELLING\_HOUSE = DWELLING\_HOUSE**/**(DWELLING\_HOUSE **+**  DWELLING\_FLAT **+** DWELLING\_SEMI **+** DWELLING\_OTHER),  PERC\_DWELLING\_FLAT = DWELLING\_FLAT**/**(DWELLING\_HOUSE **+**  DWELLING\_FLAT **+** DWELLING\_SEMI **+**  DWELLING\_OTHER), PERC\_DWELLING\_SEMI = DWELLING\_SEMI**/**(DWELLING\_HOUSE **+**  DWELLING\_FLAT **+** DWELLING\_SEMI **+**  DWELLING\_OTHER), PERC\_DWELLING\_OTHER = DWELLING\_OTHER**/**(DWELLING\_HOUSE **+**  DWELLING\_FLAT **+** DWELLING\_SEMI **+**  DWELLING\_OTHER))  *# Write cleaned data set to csv getwd()*  **write\_csv**(clean\_data, "../Data Files/ABS/Dwelling\_Type\_SA2.csv") |

**Appendix 5 - Example of Australian Bureau of Statistics Census 2016 data cleaning using R**

**(Demographics)**

|  |
| --- |
| **library**(tidyverse)  **library**(janitor)  **setwd**(**dirname**(rstudioapi**::getActiveDocumentContext**()**$**path))  erp <- **read\_csv**("../../Raw Data/Data Files/ABS/ERP/ABS\_ERP\_ASGS2016\_25042019132433480.csv")  **names**(erp)  erp <- erp **%>% select**(**-c**("MEASURE", "Measure",  "SEX\_ABS", "AGE", "FREQUENCY", "Frequency",  "TIME", "Flag Codes", "Flags", "REGIONTYPE",  "Geography Level"))  erp <- erp **%>% rename**(sa2\_code = ASGS\_2016) **%>%**  **clean\_names**()  erp\_by\_sex <- erp **%>% group\_by**(sa2\_code,  sex) **%>% summarise**(total\_value = **sum**(value)) **%>%**  **spread**(sex, total\_value) **%>% clean\_names**()  erp\_by\_age <- erp **%>% group\_by**(sa2\_code,  age) **%>% summarise**(total\_value = **sum**(value)) **%>%**  **spread**(age, total\_value) **%>% clean\_names**()  erp <- erp\_by\_sex **%>% left\_join**(erp\_by\_age)  **write\_csv**(erp, "../Data Files/ABS/ERP\_SA2\_2016.csv") |

**Appendix 6 - Example of finding percent of SA2 area covered by parkland using R**

|  |
| --- |
| **library**(tidyverse)  *# getwd()*  **setwd**(**dirname**(rstudioapi**::getActiveDocumentContext**()**$**path))  *# Create destination folder if it doesn't*  *# already exist*  **if** (**!dir.exists**("../Data Files/ABS/")) {  **dir.create**("../Data Files/ABS/")  }  *# Read the raw data csv*  mesh\_blocks <- **read.csv**("../../Raw Data/Data Files/ABS/Mesh\_Blocks/MB\_2016\_NSW.csv")  **str**(mesh\_blocks)  mesh\_blocks **%>% distinct**(MB\_CATEGORY\_NAME\_2016)  *# Find % of space allocated to Parkland*  *# for each mesh block*  open\_space <- mesh\_blocks **%>% filter**(STATE\_NAME\_2016 **==**  "New South Wales") **%>% select**(MB\_CODE\_2016,  MB\_CATEGORY\_NAME\_2016, SA2\_CODE = SA2\_MAINCODE\_2016,  AREA\_SQKM = AREA\_ALBERS\_SQKM) **%>% group\_by**(SA2\_CODE,  MB\_CATEGORY\_NAME\_2016) **%>% summarise**(SUM\_AREA\_SQKM = **sum**(AREA\_SQKM)) **%>%**  **spread**(MB\_CATEGORY\_NAME\_2016, SUM\_AREA\_SQKM,  fill = 0) **%>% mutate**(PERC\_OPEN\_SPACE = Parkland**/**(Commercial **+**  Education **+** `Hospital/Medical` **+** Industrial **+**  MIGRATORY **+** NOUSUALRESIDENCE **+** OFFSHORE **+**  Other **+** Parkland **+** `Primary Production` **+**  Residential **+** SHIPPING **+** Transport **+**  Water))  *# Write data to csv*  **write\_csv**(open\_space, "../Data Files/ABS/Open\_Space\_SA2.csv") |

**Appendix 7 - Example of Australian Bureau of Statistics Socio-Economic Indexes for Areas**

**data cleaning using R**

|  |
| --- |
| **library**(tidyverse)  *# Set directory to my the location where*  *# this file is*  **setwd**(**dirname**(rstudioapi**::getActiveDocumentContext**()**$**path))  *# getwd()*  seifa\_data <- **read\_csv**("../../Raw Data/Data Files/ABS/SEIFA\_2016\_Data.csv")  *# Review data - AsGS\_2016 field is an INT*  **head**(seifa\_data)  *# Summarise the data - obsValue has 48*  *# NA's but non of these have a*  *# SEIFA\_MEASURE == SCORE*  **summary**(seifa\_data)  seifa\_nas <- seifa\_data **%>% filter**(**is.na**(obsValue) **==** TRUE)  *# Create a clean data set for use - only*  *# want high level scores and remove*  *# records with an obsValue of NA*  clean\_seifa\_data <- seifa\_data **%>% filter**(SEIFA\_MEASURE **==**  "SCORE") **%>% select**(SA2\_CODE = ASGS\_2016,  SEIFAINDEXTYPE, obsValue) **%>% spread**(SEIFAINDEXTYPE,  obsValue) **%>% select**(SA2\_CODE, SEIFA\_Edu\_Occ\_Index = IEO,  SEIFA\_Economic\_Res\_Index = IER, SEIFA\_Rel\_SocioEco\_Adv\_Disadv\_Index = IRSAD,  SEIFA\_Rel\_SocioEco\_Disadv\_Index = IRSD)  *# Write Clean Data to disk*  **write\_csv**(clean\_seifa\_data, "../Data Files/ABS/SEIFA\_2016\_Data.csv")  *# Check for duplicates - nope none, only*  *# 1 record per SA2*  clean\_seifa\_data **%>% group\_by**(SA2\_CODE) **%>%**  **mutate**(total = **n**()) **%>% filter**(total **>**  1) |

**Appendix 8 - Example of Exploratory Data Analysis of Australian Bureau of Statistics Census**

**2016 data**

|  |
| --- |
| **library**(tidyverse)  **library**(Hmisc)  **library**(corrplot)  *# getwd()*  **setwd**(**dirname**(rstudioapi**::getActiveDocumentContext**()**$**path))  *# read cleaned data set*  dwelling\_type <- **read\_csv**("../Data Files/ABS/Dwelling\_Type\_SA2.csv")  mesh\_blocks <- **read\_csv**("../../Raw Data/Data Files/ABS/Mesh\_Blocks/MB\_2016\_NSW.csv")  **str**(mesh\_blocks)  *# Get mesh block data at SA2 level*  sa2\_data <- mesh\_blocks **%>% distinct**(SA2\_MAINCODE\_2016,  SA2\_NAME\_2016, STATE\_CODE\_2016, STATE\_NAME\_2016)  *# No duplicate SA2 Codes*  sa2\_data **%>% group\_by**(SA2\_MAINCODE\_2016) **%>% summarise**(cnt = **n**()) **%>%**  **filter**(cnt **>** 1)  *# Some SA2's don't have any dwellings - positive*  *# skew*  dwelling\_type **%>% mutate**(TOTAL = DWELLING\_HOUSE **+** DWELLING\_FLAT **+**  DWELLING\_SEMI **+** DWELLING\_OTHER) **%>% ggplot**() **+**  **geom\_histogram**(**aes**(x = TOTAL), bins = 50)  *# 62 SA2's have no dwellings - 7 in NSW, a military*  *# base, centennial park, a NP, a cemetry, and*  *# Industrial area, Banksmeadow is whaves and*  *# industry*  dwelling\_type **%>% mutate**(TOTAL = DWELLING\_HOUSE **+** DWELLING\_FLAT **+**  DWELLING\_SEMI **+** DWELLING\_OTHER) **%>% filter**(TOTAL **==**  0) **%>% left\_join**(sa2\_data, by = **c**(SA2\_CODE = "SA2\_MAINCODE\_2016")) **%>%**  **select**(SA2\_CODE, TOTAL, SA2\_NAME\_2016, STATE\_NAME\_2016) **%>%**  **filter**(**between**(SA2\_CODE, 1e+08, 2e+08))  *# There are a couple of areas with high numbers of*  *# dwellings - Waterloo/Beaconsfield in NSW is high*  *# density*  dwelling\_type **%>% mutate**(TOTAL = DWELLING\_HOUSE **+** DWELLING\_FLAT **+**  DWELLING\_SEMI **+** DWELLING\_OTHER) **%>% filter**(TOTAL **>**  15000) **%>% left\_join**(sa2\_data, by = **c**(SA2\_CODE = "SA2\_MAINCODE\_2016")) **%>%**  **select**(SA2\_CODE, TOTAL, SA2\_NAME\_2016, STATE\_NAME\_2016) **%>%**  **filter**(**between**(SA2\_CODE, 1e+08, 2e+08))  *# Remove SA2's with no dwellings and only show NSW*  *# SA2's*  dwelling\_type\_filtered <- dwelling\_type **%>% mutate**(TOTAL = DWELLING\_HOUSE **+**  DWELLING\_FLAT **+** DWELLING\_SEMI **+** DWELLING\_OTHER) **%>%**  **filter**(TOTAL **!=** 0) **%>% inner\_join**(sa2\_data, by = **c**(SA2\_CODE = "SA2\_MAINCODE\_2016"))  *## DWELLING HOUSE Some areas in NSW have no houses -*  *## data may be slightly skewed*  dwelling\_type\_filtered **%>% ggplot**() **+ geom\_histogram**(**aes**(x = DWELLING\_HOUSE),  bins = 50)  *# Standardise data and confirm data has a long tail*  house\_std <- **scale**(dwelling\_type\_filtered**$**DWELLING\_HOUSE)  **qqnorm**(house\_std)  **abline**(a = 0, b = 1, col = "grey")  ## DWELLING FLAT There are 21 no flat SA2's in NSW -  ## industrial areas, offshore shipping, Rural areas  dwelling\_type\_filtered **%>% filter**(DWELLING\_FLAT **==**  0) **%>% select**(SA2\_CODE, SA2\_NAME\_2016, PERC\_DWELLING\_HOUSE,  PERC\_DWELLING\_FLAT, PERC\_DWELLING\_SEMI, PERC\_DWELLING\_OTHER)  *# Some areas in NSW have no houses - data skewed*  dwelling\_type\_filtered **%>% ggplot**() **+ geom\_histogram**(**aes**(x = DWELLING\_FLAT),  bins = 100)  *# Standardise data and confirm data is not normally*  *# distributed*  flat\_std <- **scale**(dwelling\_type\_filtered**$**DWELLING\_FLAT)  **qqnorm**(flat\_std)  **abline**(a = 0, b = 1, col = "grey")  ## DWELLING SEMI There are 19 no semi SA2's in NSW -  ## industrial areas, airport, offshore shipping,  ## rural areas  dwelling\_type\_filtered **%>% filter**(DWELLING\_SEMI **==**  0) **%>% select**(SA2\_CODE, SA2\_NAME\_2016, PERC\_DWELLING\_HOUSE,  PERC\_DWELLING\_FLAT, PERC\_DWELLING\_SEMI, PERC\_DWELLING\_OTHER)  *# Some areas in NSW have no houses - data skewed*  dwelling\_type\_filtered **%>% ggplot**() **+ geom\_histogram**(**aes**(x = DWELLING\_SEMI),  bins = 100)  *# Standardise data and confirm data is not normally*  *# distributed*  semi\_std <- **scale**(dwelling\_type\_filtered**$**DWELLING\_SEMI)  **qqnorm**(semi\_std)  **abline**(a = 0, b = 1, col = "grey")  *# Check correlation between variables*  dwelling\_matrix <- dwelling\_type\_filtered **%>% select**(DWELLING\_FLAT,  DWELLING\_HOUSE, DWELLING\_OTHER, DWELLING\_SEMI) **%>%**  **as.matrix**()  *# Show values - nothing really high*  **rcorr**(dwelling\_matrix, type = "pearson")  *# And a plot for good measure*  **corrplot**(**cor**(dwelling\_matrix), method = "ellipse") |

**Appendix 9 - Example of merging cleaned datasets for modelling using R**

|  |
| --- |
| **library**(tidyverse)  *# getwd()*  **setwd**(**dirname**(rstudioapi**::getActiveDocumentContext**()**$**path))  *# Read csv's*  dwelling\_type <- **read\_csv**("../../Clean Data/Data Files/ABS/Dwelling\_Type\_SA2.csv")  hh\_composition <- **read\_csv**("../../Clean Data/Data Files/ABS/HouseHold\_Composition\_SA2.csv")  place\_of\_birth <- **read\_csv**("../../Clean Data/Data Files/ABS/Place\_Of\_Birth\_SA2.csv")  seifa <- **read\_csv**("../../Clean Data/Data Files/ABS/SEIFA\_2016\_Data.csv")  mesh\_blocks <- **read\_csv**("../../Raw Data/Data Files/ABS/Mesh\_Blocks/MB\_2016\_NSW.csv")  *# Get mesh block data at SA2 level*  sa2\_data <- mesh\_blocks **%>% distinct**(SA2\_MAINCODE\_2016,  SA2\_NAME\_2016, STATE\_CODE\_2016, STATE\_NAME\_2016)  *# Join Datasets together*  model\_data <- dwelling\_type **%>% inner\_join**(hh\_composition,  by = **c**("SA2\_CODE")) **%>% inner\_join**(place\_of\_birth,  by = **c**("SA2\_CODE")) **%>% inner\_join**(seifa, by = **c**("SA2\_CODE")) **%>%**  **semi\_join**(sa2\_data, by = **c**(SA2\_CODE = "SA2\_MAINCODE\_2016")) **%>%**  **select**(**-starts\_with**("PERC\_")) |