Australian National University

Assignment 2

Ethan Yifan ZHU
u7560434

COMP 8410

School of Computing,
College of Engineering, Computing and Cybernetics,
The Australian National University

1 Introduction

Nowadays data is being gathered and used widely. Using data mining methods to extract interesting patterns is becoming more and more important. This course offers a solid foundation for comprehending the complicated processes in data mining, from the initial stages of data preparation to the complicated analysis and result interpretation. It is expected to critically apply these concepts to the real world. This assignment focuses on using data mining techniques and technologies in real-world settings.

This assignment has two goals. The first one is to use a variety of mining techniques to find hidden patterns and relationships in the dataset, which will provide further insight into the driving attributes of the referendum. The second goal is to evaluate performance, checking whether they can forecast outcomes well and make insightful conclusions.

2 Data Description

2.1 Data Source and Population

The data originates from The Australian Data Archive's ANU Poll Dataverse (Biddle and McAllister, 2023). It was a survey conducted in late 2023 related to the 14th October 2023 Australian Constitutional Referendum on the Aboriginal and Torres Strait Islander Voice to Parliament.

The survey gathered feedback from a wide range of respondents representing different ages, genders, occupations, economic levels, and political views and addressed several topics, including general political and social views and public opinion on the referendum.

2.2 Data Attributes

The dataset consists of several sections, each containing a unique set of questions. Below is a summary of all attributes.

General Questions: These questions focus on general satisfaction with political institutions and government, as well as plans to vote for parties and satisfaction with Australian life. General questions are noted as “A” plus the question number.

Mental Health: This section evaluates emotional and mental health. The responses indicate the frequency of the following questions: how frequently did respondents feel anxious, despairing, restless, or lonely over four weeks? Mental health questions are noted as “D” plus the question number.

Employment and Income: The employment status, type, hours worked, and income levels of the respondents are covered in this section. Besides it also includes questions about financial hardship. Employment and income questions are noted as “E” plus the question number.

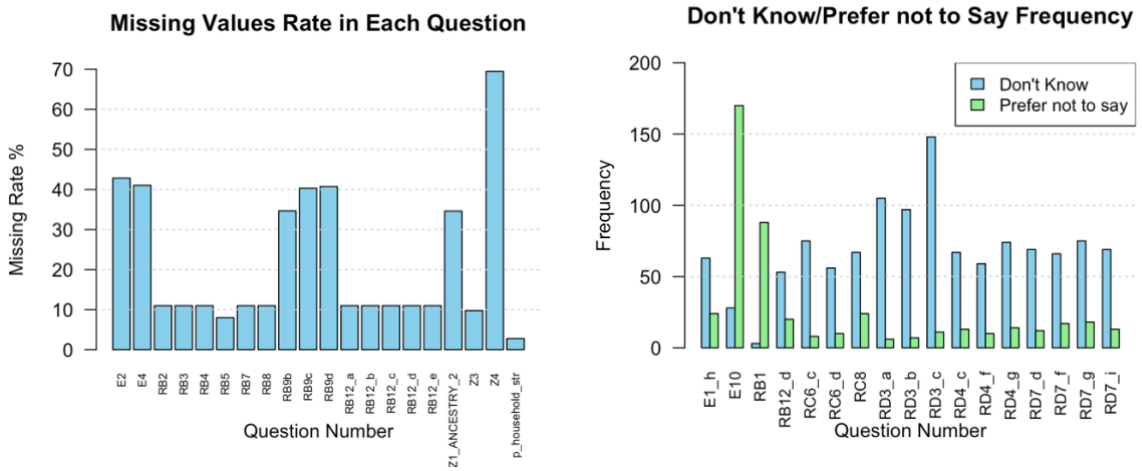
Referendum Related: These are questions regarding the referendum, including whether or not respondents cast votes, how they voted, and what factors affected their choice. It also asks whether voting should be required and how much interest they had in the referendum campaign. Referendum-related questions are noted as “RA” (The Referendum Campaign) or “RB” (Voting) plus the question number.

Political and General Views: This section reflects the respondents' political inclinations, levels of contentment with democracy, and opinions on different political parties. Additionally, respondents' levels of trust in several institutions, including the federal government, political parties, and the legal system are questioned. Political and general views questions are noted as “RC” or “RD” plus the question number.

Demographic Information: Data on the ancestry, religious affiliation, and other demographic characteristics of the respondents such as regions and states, backgrounds, citizenships, languages, households, ages, educations, and genders, are gathered in this section. It inquires about the significance of the respondents' ancestry and whether or not they follow any particular religion. Demographic information questions are noted as “Z” plus the question number or “p” plus descriptions.

2.3 Data Quality

The dataset has consistent coding and descriptions, and it is generally well-structured. However, there are still some missing values. In order to check for missing values, certain attributes could require closer examination (Brick and Kalton, 1996).



(a) (b)

Figure 2.1 Data quality for the provided data (a) Missing value rate in each attribute (b) Don’t know/Prefer not to say counts in each attribute

Figure 2.1 (a) summarised missing value rates for all questions. The question “Z4 Which religion or denomination do you belong to?” has the highest missing rate, approaching 70%. This might be attributed to “Z3: Do you consider yourself as

belonging to any particular religion or denomination?", in which if one answered "yes", Z4 would be skipped and left as missing value. Questions E2 and E4 (Section E: Employment and Income) also have relatively high missing rates, each approaching or exceeding 40%. Followed by RB9b, RB9c and RB9d, around 40%. These questions are the same as Z4, for they highly depended on the answers from E1 and RB9a respectively.

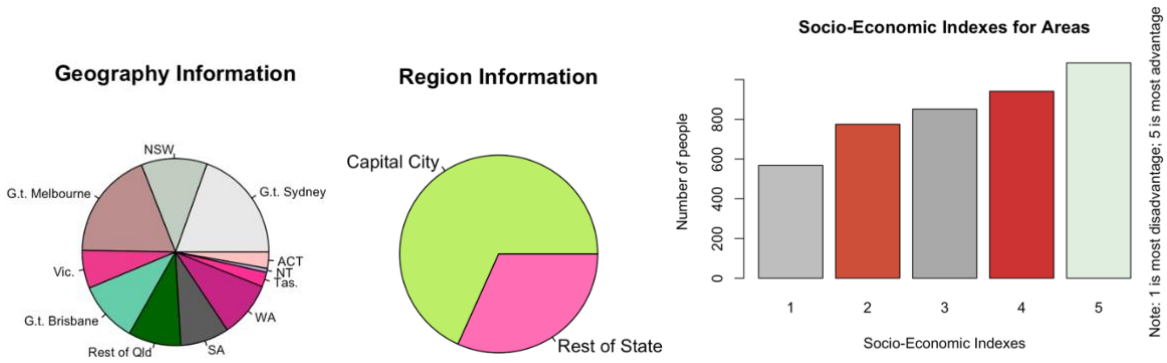
From the respondents' points of view, the minimum number of missing values is 1, while the maximum number is 18, with median value of 3 and mean value of 4.406 for a single questionnaire.

For all questions, Attributes like "Don't know", or "Prefer not to Say" might suggest areas where participants are uncertain or unwilling to respond. There are some reasons why they choose these options. Initially, in cases when they are unsure about the meaning of a question (Feick, 1989). Second, not to reflect or make a commitment (Oppenheim, 1993). Third, when the survey exceeds their capacity or level of motivation (Krosnick, 1991).

Figure 2.1(b) shows the frequency of "Don't know" or "Prefer not to say" in each question. Notably, the "E10: What is your household's total income, after tax and compulsory deductions, from all sources?" has the highest count for "Prefer not to say," approaching 170. This can lead to further analysis of why certain attributes have higher rates and how they might affect the dataset's overall quality.

2.4 Basic Statistical Summary

In this part, statistical summary will be made according to the demographic information of respondents.

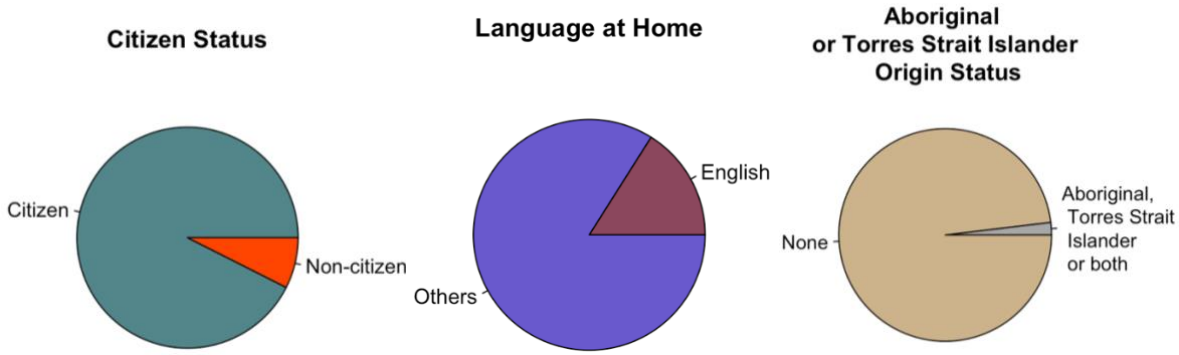


(a) (b) (c)

Figure 2.2 Demographic information I (a) Geography information (b) Region information (c) Socio-economic indexes for areas

Figure 2.2 (a) suggests the geographic distribution of respondents. Majority of them are from New South Wales (19.53% from Great Sydney and 11.42% from remaining area of NSW), followed by Victoria (18.74% from Great Melbourne and 6.58% from remaining area of Vic.). Figure 2.2 (b) shows region distribution. 68.26% of them are

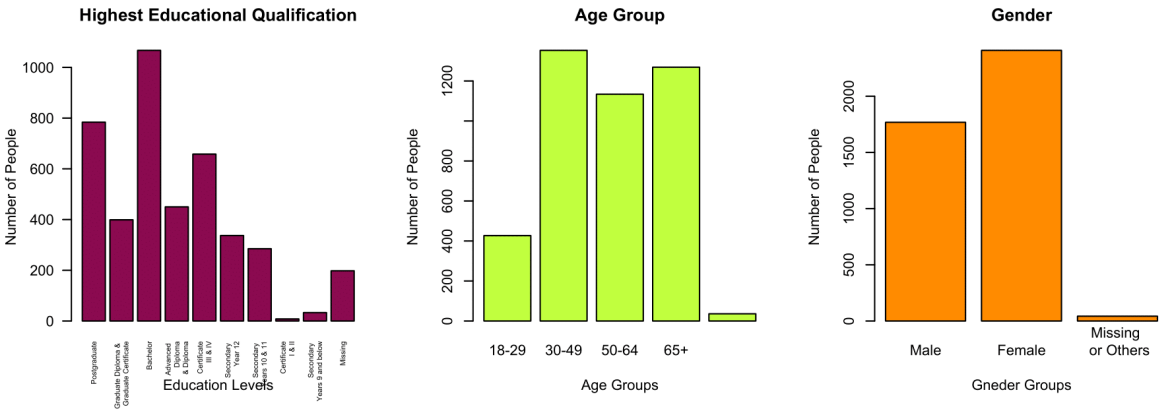
living in capital city, while 31.74% are staying in non-capital city. Figure 2.2 (c) demonstrates the number of people across five socio-economic indexes. From the most advantage index to the most disadvantage, the number of people is decreasing.



(a) (b) (c)

Figure 2.3 Demographic information II (a) Citizen status (b) Language spoken at home (c) Aboriginal or Torres Strait Islander origin status

Figure 2.3 (a) compares citizens to non-citizens. There is a much larger proportion of being citizens (92.55%). Figure 2.3 (b) illustrates the languages spoken at home, with a larger portion speaking other languages (83.95%) rather than English. Figure 2.3 (c) shows the minority portion of Aboriginal or Torres Strait Islander Origin of 80 people (1.90%).



(a) (b) (c)

Figure 2.4 Demographic information III (a) Highest educational qualification level distribution (b) Age distribution (c) Gender distribution

Figure 2.4 (a) shows the number of people with various levels of education. Most of them have highest education level of Bachelor's degree (25.29%), followed by Postgraduate degree (18.58%) and Certificate III & IV (15.60%). Figure 2.4 (b) illustrates the number of people in different age brackets. The majority is in their 30-

49 (32.07%), followed by 65+ (30.08%) and 50-64 (26.88%). Figure 2.4 (c) shows the number of people by gender, with more females (57.07%) than males (41.91%).

3 Methods Description Results Presentation

3.1 Software and Packages

R (version 4.3.2 based on Mac with M2 chips) with RStudio was selected as the primary data analysis tool. Additionally, following packages were used during analysis: arules, arulesViz, rpart, rpart.plot, and MASS. Before using each package, there was code for installation in the provided code in the appendix.

3.2 Data Mining Method

3.2.1 Association Rules

Association rule mining is ideal for discovering patterns and relationships within categorical data (Kumbhare and Santosh, 2014).

Step 1: Data Pre-processing

The first step of pre-processing was eliminating certain columns from the dataset, including SRCID, IntDate, s_order, Mode, and A6_VERB. Because they either contain uniform data (A6_VERB with a constant value of 99) that does not contribute to pattern discovery, or administrative data (such IDs and timestamps) that are unnecessary to the research. To avoid biases on the questions' order rather than their content, columns listing the questions' order were also eliminated, i.e. questions A4_order, RA2_order etc.

Apart from the column 'E4' which indicates working hours and was kept as a numeric type, all other columns were transformed to factor type in order to prepare them for categorical analysis in association rules.

Step 2: Method

The 'RB1' column, which indicates voter decision, was chosen at the Righthand side (RHS) in the rules, and only focused on binary results ('Yes' or 'No'). The Apriori method was used to extract rules because of its effectiveness in handling big datasets and its capacity to identify common itemsets. The parameters were set up with 0.3 support and 0.4 confidence to guarantee the quality of the obtained rules. To maintain the rules' readability, the number of items in each rule could only be in range 1 to 10. The strongest correlations were then found by sorting the rules based on their lift values.

Step 3: Interpreting Result

According to the top rules in Table 3.1, there are two strong indicators for voting "yes": strong agreement that First Nations people should have a voice in decisions that affect them (RD7_c=1) and beliefs about the referendum's positive effect on First Nations communities (RB12_d=1). Strong, dependable regulations are indicated by

the high lift and confidence values. Together with personal citizenship status (p_citizen=1) and non-Indigenous origin (p_atsi=4), these opinions further reinforce a pattern where a "Yes" vote is significantly predicted by the general public view.

Table 3.1 Top 5 association rules for {RB1=1}

LHS	RHS	Support	Confidence	Coverage	Lift
{RB12_d=1,RD7_c=1}	{RB1=1}	0.304	0.950	0.320	1.933
{RB5=1,RB12_d=1}	{RB1=1}	0.317	0.931	0.340	1.894
{RB5=1,RB12_d=1,p_citizen=1}	{RB1=1}	0.312	0.931	0.335	1.893
{RB5=1,RB12_d=1,p_atsi=4}	{RB1=1}	0.311	0.930	0.334	1.893
{RB5=1,RB12_d=1,p_citizen=1,p_atsi=4}	{RB1=1}	0.306	0.930	0.329	1.892

Top Rules with {RB1=2} are shown in Table 3.2. Economic concerns may have influenced the 'No' vote, as evidenced by denying both unemployment and actively seeking employment (E1_c=2), which were paired with citizenship (p_citizen) and language spoken at home (p_lote=2). The 'No' vote appears to be composed of a less coherent group of people, as evidenced by the lower levels of support and confidence in these rules compared to the 'Yes' vote.

Table 3.2 Top 5 association rules for {RB1=2}

LHS	RHS	Support	Confidence	Coverage	Lift
{E1_b=2,E1_c=2,p_citizen=1,p_lote=2,p_atsi=4}	{RB1=2}	0.308	0.436	0.705	1.096
{E1_b=2,p_citizen=1,p_lote=2}	{RB1=2}	0.321	0.437	0.736	1.096
{E1_b=2,E1_c=2,p_citizen=1,p_lote=2}	{RB1=2}	0.315	0.437	0.721	1.096
{E1_b=2,E1_c=2,p_citizen=1,p_lote=2,p_atsi=4}	{RB1=2}	0.308	0.436	0.705	1.096
{E1_b=2,E1_d=2,p_citizen=1,p_lote=2,p_atsi=4}	{RB1=2}	0.306	0.436	0.701	1.093

These results give professionals in socio-political studies and political campaigns a practical understanding of voter psychology and demographic factors. Comprehending these patterns may encourage the creation of more sophisticated, data-driven, and focused campaign strategies. This investigation also shows how data mining techniques can be essential in strategic decision-making processes in political fields, as well as how powerful they are in revealing hidden patterns.

3.2.2 Decision Tree

Decision Trees offers a simple graphical depiction of the decision pathways impacted by opinion and demographic factors (Song and Lu, 2015).

Step 1: Data Pre-processing

Columns irrelevant to the analysis were eliminated, except those that contained information on demographics, RC Political Views, and RD General Views. To prevent sequence bias, columns with question orders were removed, as indicated before. To improve data quality, rows with missing values, rows indicating as ‘don’t know’ (-98) or ‘refused’ (-99), and specifically in ‘p_geography’ tagged as ‘Unable to establish’ (-97) were eliminated. Besides, categories such as ‘Informal/didn't vote’ and ‘Not eligible’ were removed, leaving only replies where ‘RB1’ was 1 or 2 (voted “Yes” or “No”).

Step 2: Modelling

A 70% training set and a 30% test set were randomly divided from the cleaned data. Using the training data, a decision tree was built to simulate the probability of voting depending on the factors. Then test set was used to evaluate the model. This approach offers an intuitive understanding of the various aspects that influence voting behaviour and maintains a simple visual representation.

Step 3: Interpreting Result

Figure 3.1 shows the decision tree model built based on training data. The decision tree starts with the respondents' opinions regarding the significance of First Nations peoples having a voice (RD7_c).

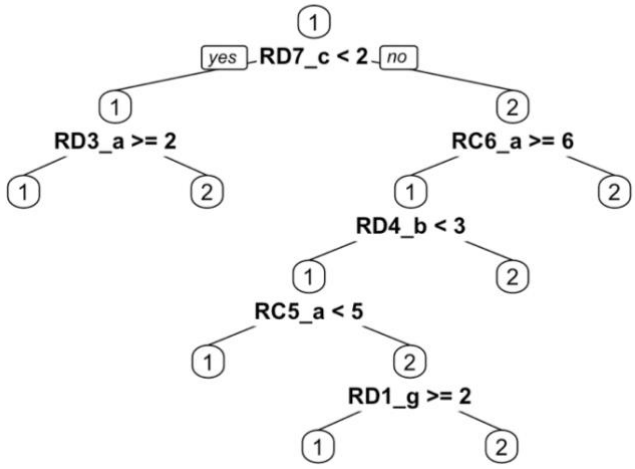


Figure 3.1 Decision Tree

For those who are likely to vote “Yes”, they might hold the view that (1) they agree First Nations people have their voice and government help for their land rights has not gone far enough or about right, or (2) they are in favour of Anthony Albanese and the Liberal Party, government should provide more assistance for First Nations people, and they feel they are not very close or not at all close to Australian people.

For those who are likely to vote “No”, they might hold the view that (1) they agree First Nations people have their voice but government help for their land rights has gone too far, and/or (2) they are not in favour of Anthony Albanese.

The contingency table in Table 3.3, yet having some misclassification, validates the model's great predictive ability, especially for 'Yes' votes. Voter decisions are mostly influenced by major opinion patterns that the tree identifies. True positive rate (recall) is 89.19%. Accuracy is 85.08%.

Table 3.3 Test result for the training model

		RB1 origin		
prediction		1	2	Total
	1	454 TP	55 FP	509
	2	81 FN	322 TN	403
	Total	535	377	912

The decision tree shows the complex relationships between specific political beliefs, views on First Nations peoples, and larger demographic issues. It provides insightful information for politicians and political analysts. Understanding these may help in creating more effective and targeted communication in political campaigns and policy discussions.

3.2.3 Stepwise Regression

Stepwise regression's fundamental concept is to add variables to the model one at a time. An F-test and a t-test are performed on the chosen predictor variables one at a time after each predictor variable is introduced (Johnsson, 1992).

Step 1: Data Pre-processing

Pre-processing procedures were similar to those in Decision Tree, with an emphasis on general questions (Section A), mental health (Section D) and employment, income and financial hardship (Section E). To prevent sequence bias, columns with question orders were removed, as indicated before. Rows with missing values, rows indicating ‘don’t know’ (-98 or 98) or ‘refused’ (-99 or 99) were eliminated. The purpose was to investigate the relationship between voting behaviour and socio-economic characteristics.

Step 2: Modelling

Firstly a full model was fitted using linear regression including all of the predictors. Then it was refined by applying “stepAIC” function, which used both forward and backward stepwise regression to remove statistically insignificant predictors. By focusing only on essential factors, this procedure improves the model's interpretability and prediction accuracy.

Step 3: Interpreting Result

The final regression formula is

RB1 = 1.2958 + (-0.0584) * A1 + 0.2161 * A4_a + 0.0484 * A4_c + (-0.0230) * D1_d + (-0.1939) * E1_f + (-0.1211) * E1_h + 0.0013 * E4 + (-0.0152) * E10 + (-0.0225) * RA2_a + (-0.0385) * RA2_b + 0.1396 * RA4 + 0.1160 * RA8

Positive coefficients for satisfaction of institutions (A4_a and A4_c) suggest that higher levels of dissatisfaction and distrust are associated with a higher probability of voting “No”. Besides, negative coefficients for the employment status (E1_f) suggest that retired people are more likely to vote “No”, which possibly indicates different priorities or more conservative opinions. Furthermore, the fewer people show their interest in Referendum campaign (RA4) and outcome (RA8), the more likelihood they would vote “No”.

Table 3.4 Final coefficients

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.2958235	0.1279957	10.124	< 2e-16	***
A1	-0.0584175	0.0122355	-4.774	2.04e-06	***
A4_a	0.2161411	0.0236859	9.125	< 2e-16	***
A4_c	0.0484359	0.0235940	2.053	0.040312	*
D1_d	-0.0229782	0.0116217	-1.977	0.048264	*
E1_f	-0.1938700	0.0547754	-3.539	0.000417	***
E1_h	-0.1210949	0.0327070	-3.702	0.000224	***
E4	0.0013198	0.0008069	1.636	0.102184	
E10	-0.0152436	0.0055752	-2.734	0.006351	**
RA2_a	-0.0225216	0.0129837	-1.735	0.083082	.
RA2_b	-0.0384978	0.0133764	-2.878	0.004077	**
RA4	0.1396382	0.0189109	7.384	2.97e-13	***
RA8	0.1159516	0.0276833	4.189	3.03e-05	***

Notes: Signif. codes: 0, ‘***’ 0.001, ‘**’ 0.01, ‘*’ 0.05, ‘.’ 0.1, ‘ ’ 1. Residual standard error: 0.4157 on 1133 degrees of freedom. Multiple R-squared: 0.2634, Adjusted R-squared: 0.2556. F-statistic: 33.76 on 12 and 1133 DF, p-value: < 2.2e-16

The model's overall F-statistic is significant, indicating that the predictors together account for roughly 25.56% of the variance. It highlights the connections between psychological, societal, and personal aspects impacting voting behaviour.

4 Conclusion and Future Work

Key elements affecting voter decisions for the Australian Constitutional Referendum in 2023 have been identified by using data mining techniques such as association rules, decision trees, and stepwise regression. By highlighting demographic and opinion-based elements, the findings may influence campaign advertising. By addressing voter concerns, an understanding of these attributes may persuade hesitant voters. To guarantee inclusion with a variety of voter categories, it is necessary to convert these findings into successful strategies for campaigns.

Further studies might explore social media data to obtain public mood and improve adaptability. Besides, prediction accuracy may also be increased by using more advanced machine learning techniques or, looking more deeply at the current outcome. Evaluating the effectiveness of the model would be improved by the long-term study that tracks changes in opinion over time, for the data provided only cover half a month. Furthermore, it is essential to consider the moral effects of data mining in political campaigns, including privacy, consent, and the prospect of manipulation.

Reference

Biddle, N. and McAllister, I. (2023). ANU Poll 57/Australian Constitutional Referendum Survey (ACRS) (October 2023): Aboriginal and Torres Strait Islander Voice to Parliament. [online] Available at: doi:10.26193/13NPGQ.

Brick, J. and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5(3), pp.215–238. doi:https://doi.org/10.1177/096228029600500302.

Feick, L.F. (1989). Latent Class Analysis of Survey Questions that Include Don't Know Responses. *The Public Opinion Quarterly*, [online] 53(4), pp.525–547. Available at: https://www.jstor.org/stable/2749357 [Accessed 27 Apr. 2024].

Johnsson, T. (1992). A procedure for stepwise regression analysis. *Statistical Papers*, 33(1), pp.21–29. doi:https://doi.org/10.1007/bf02925308.

Krosnick, J.A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3), pp.213–236. doi:https://doi.org/10.1002/acp.2350050305.

Kumbhare, T. and Santosh, S. (2014). *An Overview of Association Rule Mining Algorithms*. [online] International Journal of Computer Science and Information Technologies. Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d4058d9f3f66c53ddea776c974fbd740afd994b4.

Oppenheim, A.N. (1993). Questionnaire Design, Interviewing and Attitude Measurement. *Journal of Marketing Research*, 30(3), p.393. doi:https://doi.org/10.2307/3172892.

Song, Y.-Y. and Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, [online] 27(2), pp.130–5. doi:https://doi.org/10.11919/j.issn.1002-0829.215044.

Appendix

```
##############################################
# ANU COMP 8410 Assignment 2 code
# Ethan Y. ZHU
# u7560434
##############################################
# Load a CSV file
data <- read.csv("/Users/ethanyifanzhu/Desktop/2024 S1/COMP8410 - Data Mining/Assignment/Assignment 2/02_ANUPoll_57_CSV_100150_general.csv")
data <- as.data.frame(data)

# Number of rows and columns
n_row <- nrow(data)
n_col <- ncol(data)

##############################################
############### Data summary ###############
options(max.print = .Machine$integer.max)
str(data, list.len=ncol(data))
summary(data)

##############################################
############### Data Quality ###############
# Display column with missing values
missing_counts <- colSums(is.na(data))
missing_counts <- missing_counts[missing_counts > 0] / n_row * 100
missing_counts
quartz()
# Create a bar plot
barplot(
  missing_counts,
  ylim = c(0, 70),
  main = "Missing Values Rate in Each Question",
  xlab = "Question Number",
  ylab = "Missing Rate %",
  col = "skyblue",
  border = "black",
  las = 2,  # Rotate x-axis labels for better readability
  cex.names = 0.6  # Adjust x-axis label font size
)
grid(0, 7, lwd = 1, col = "azure3")

# Display row with missing values
missing_counts_row <- rowSums(is.na(data))
missing_counts_row <- missing_counts_row[missing_counts_row > 0]
```

```r
summary(missing_counts_row)

# Display column with code 98 or 99
code98_counts <- colSums(data == -98, na.rm = TRUE)
code99_counts <- colSums(data == -99, na.rm = TRUE)
reject_count <- rbind(code98_counts, code99_counts)
reject_count <- reject_count[, colSums(reject_count > 50) > 0]
reject_count
quartz()
# Create a bar plot
barplot(
  reject_count,
  beside = TRUE,
  ylim = c(0, 200),
  main = "Don't Know/Prefer not to Say Frequency",
  xlab = "Question Number",
  ylab = "Frequency",
  col = c("skyblue", "lightgreen"),
  border = "black",
  las = 2,  # Rotate x-axis labels for better readability
  cex.names = 0.8  # Adjust x-axis label font size
)
grid(0, 4, lwd = 1, col = "azure3")
# Add a legend
legend(
  "topright",
  legend = c("Don't Know", "Prefer not to say"),
  fill = c("skyblue", "lightgreen"),
  cex = 0.9
)

##############################################
############ Statistical Summary ############
# p_geography. Capital city / rest of state by state
geography <- table(data$p_geography)
geography <- cbind(geography[1], geography[2], geography[3], geography[4],
                   geography[5], geography[6], geography[7] + geography[8],
                   geography[9] + geography[10], geography[11] + geography[12],
                   geography[13] + geography[14], geography[15])
geography / sum(geography) * 100
geo_label <- c("G.t. Sydney", "NSW", "G.t. Melbourne", "Vic.",
               "G.t. Brisbane", "Rest of Qld", "SA",
               "WA", "Tas.", "NT", "ACT")
pie(geography, labels = geo_label, main = "Geography Information", cex = 0.7, col
= sample(colors(TRUE), 11))
```

```r
# p_region. Capital city / rest of state
region <- table(data$p_region)
region / sum(region) * 100
reg_label <- c("Capital City", "Rest of State")
pie(region, labels = reg_label, main = "Region Information", col = sample(colors(TRUE), 2))


# p_seifa. Socio-Economic Indexes for Areas (IRSD 2021)
barplot(table(data$p_seifa), main = "Socio-Economic Indexes for Areas", col = sample(colors(TRUE), 5), xlab = "Socio-Economic Indexes", ylab = "Number of people")
mtext("Note: 1 is most disadvantage; 5 is most advantage", side = 4)


# p_citizen. Are you an Australian citizen?
citizen <- table(data$p_citizen)
citizen / sum(citizen) * 100
pie(citizen, labels = c("Citizen", "Non-citizen"), main = "Citizen Status", col = sample(colors(TRUE), 2))


# p_lote. Do you use a language other than English at home?
language_using <- table(data$p_lote)
language_using <- cbind(language_using[3], language_using[1]+language_using[2]+language_using[4])
language_using / sum(language_using) * 100
pie(language_using, labels = c("English", "Others"), main = "Language at Home", col = sample(colors(TRUE), 2))


# p_atsi. Are you of Aboriginal or Torres Strait Islander origin?
asti <- table(data$p_atsi)
asti / sum(asti) * 100
asti <- cbind("Aboriginal, Torres Strait Islander or both" = asti[3] + asti[4] + asti[5], "None of Both" = asti[6])
pie(asti, main = "Aboriginal \n or Torres Strait Islander \n Origin Status", labels = c("Aboriginal, \n Torres Strait \n Islander \n or both", "None"), col = sample(colors(TRUE), 2))


quartz()
par(mfrow = c(1, 3))
# d_education. Highest educational qualification (Confidentialised Variable)
edu <- table(data$p_education_sdc)
edu <- cbind("Postgraduate" = edu[2], "Graduate Diploma & \n Graduate Certificate" = edu[3],
             "Bachelor" = edu[4],"Advanced \n Diploma\n & Diploma" = edu[5],
             "Certificate\n III & IV" = edu[6], "Secondary \n Year 12" = edu[7],
             "Secondary \n Years 10 & 11" = edu[8],"Certificate\n I & II" = edu[9],
             "Secondary \n Years 9 and below" = edu[10], "Missing" = edu[1])
edu / sum(edu) * 100
barplot(edu, main = "Highest Educational Qualification", col = sample(colors(TRUE), 10), las = 2, cex.names = 0.5, xlab = "Education Levels", ylab = "Number of People")
```

```r
# p_age_group_ADA. Age group as of 1st March 2023 - ADA groupings (Confidentialised Variable)
age <- table(data$p_age_group_sdc)
age <- cbind("18-29" = age[2], "30-49" = age[3], "50-64" = age[4], "65+" = age[5], "Missing \n or Others" = age[1])
age / sum(age) * 100
barplot(age, main = "Age Group", col = sample(colors(TRUE), 5), xlab = "Age Groups", ylab = "Number of People")


# p_gender. How do you describe your gender? (Confidentialised Variable)
gender <- table(data$p_gender_sdc)
gender <- cbind("Male" = gender[2], "Female" = gender[3], "Missing \n or Others" = gender[1])
gender / sum(gender) * 100
barplot(gender, main = "Gender", col = sample(colors(TRUE), 3), xlab = "Gneder Groups", ylab = "Number of People")


# Other summary information
summary(data$A1)
tab1 <- table(data$A1)
per1 <- (tab1 / sum(tab1)) * 100
tab1
per1


tab2 <- table(data$A6)
per2 <- (tab2 / sum(tab2)) * 100
tab2
per2


tab3 <- table(data$RB1)
per3 <- (tab3 / sum(tab3)) * 100
tab3
per3


tab4 <- table(data$RC2)
per4 <- (tab4 / sum(tab4)) * 100
tab4
per4


##############################################
############ Association Rules ##############
# Installing Packages
install.packages("arules")
install.packages("arulesViz")
# Loading package
library(arules)
library(arulesViz)
```

```r
# All attributes
data_asso <- data[c(5, 6, 8, 10:32, 34:54, 56, 57, 60:64, 66:69, 71:74, 76:79,
81:88, 90:93, 95:105, 107:111, 113:115, 117:125, 127:143, 145:152, 155:158)]

# Convert all columns (except E4) to factors
data_asso[] <- lapply(data_asso, as.factor)
data_asso$E4 <- as.integer(data_asso$E4)
str(data_asso)

# Convert dataframe into a 'transactions' object.
data_asso_transaction <- as(data_asso, "transactions")
data_asso_transaction

# Perform association rule mining using the Apriori algorithm.
rules1 <- apriori(data = data_asso_transaction,
                  parameter = list(support = 0.3, confidence = 0.4, minlen = 1,
maxlen = 10),
                  appearance = list(rhs = c("RB1=1", "RB1=2"), default = "lhs"))
summary(rules1)

# Display rules
inspectDT(sort(rules1,by = "lift"))
quartz()
plot(rules1, method = "graph")

##########################################
############### Decision Tree ###############
# Installing Packages
install.packages("rpart")
install.packages("rpart.plot")

# Loading package
library(rpart)
library(rpart.plot)

# Demographic + RC Political Views + RD General Views => RB1 yes/no
data_tree <- data[c(45, 145:152, 155:158, 66:69, 71:74, 76:79, 81:88, 90:93, 95:105,
107:111, 113:115, 117:125, 127:136)]
# Remove rows containing missing values
data_tree <- na.omit(data_tree)
# Remove rows containing value -97, -98, -99
data_tree <- data_tree[!apply(data_tree == -97, 1, any), ]
data_tree <- data_tree[!apply(data_tree == -98, 1, any), ]
data_tree <- data_tree[!apply(data_tree == -99, 1, any), ]

# Include rows which RB1 is either 1 or 2
```

```r
data_tree <- data_tree[data_tree$RB1 == 1 | data_tree$RB1 == 2, ]
data_tree[1:13] <- lapply(data_tree[1:13], as.factor)

set.seed(84103425)
# Randomly sample 70% of the data for training set
trainID <- sample(nrow(data_tree), 0.7*nrow(data_tree))
train <- data_tree[trainID, ]
test <- data_tree[-trainID, ]

# Decision tree model
fit <- rpart(RB1~., data = train)
fit

# Plot decision tree
prp(fit, type = 2)

# Predict and create a contingency table
pred <- predict(fit, test, type = "class")
RB1_origin <- test[, "RB1"]
table(pred, RB1_origin)

##########################################
########## Stepwise Regression #############
install.packages("MASS")
library(MASS)

# A General Questions + D MENTAL HEALTH + E EMPLOYMENT, INCOME AND FINANCIAL HARDSHIP
=> RB1 yes/no
data_step <- data[c(45, 5, 6, 8, 10:31, 32, 34:44)]

# Data cleaning as before
data_step <- na.omit(data_step)
data_step <- data_step[!apply(data_step == -98, 1, any), ]
data_step <- data_step[!apply(data_step == -99, 1, any), ]
data_step <- data_step[!apply(data_step == 98, 1, any), ]
data_step <- data_step[!apply(data_step == 99, 1, any), ]
data_step <- data_step[data_step$RB1 == 1 | data_step$RB1 == 2, ]

# Fit a linear model to predict RB1. This full model includes all attributes
model_full <- lm(RB1~., data = data_step)
summary(model_full)

# Perform stepwise regression to optimize the model
model_step <- stepAIC(model_full, direction = "both", trace = FALSE)
summary(model_step)
```