Australian National University

COMP8410-Data Mining

Semester 1, 2024

Factors Influencing Public Satisfaction with National Development: A Comparative Analysis Across Political Parties Based on the Life in Australia™ Wave 83 Questionnaire

Course Code: COMP8410

Course Name: Data Mining

Student number	Student name	Contribution	Signature
u7564091	Yue Zheng	100%	Yue Zheng

Factors Influencing Public Satisfaction with National Development: A Comparative Analysis Across Political Parties Based on the Life in Australia™ Wave 83 Questionnaire

1. Why Public Satisfaction?

Public satisfaction is a cornerstone of effective governance and reflects the general well-being of a society. It measures how content the populace is with their life circumstances and the performance of their government. This concept is vital because it bridges the gap between governmental policies and the actual needs and experiences of the citizens. Analyzing public satisfaction allows governments to gauge the impact of their policies and adjust their strategies to better align with the public's expectations.

1.1 The importance of Public Satisfaction.

Public satisfaction is crucial for several reasons. Firstly, it serves as a direct indicator of the effectiveness of government actions and policies. When citizens are satisfied, it typically indicates that the government is successfully meeting their needs and expectations. Secondly, high levels of public satisfaction are often associated with greater social harmony and reduced social unrest. This environment fosters a more stable society where individuals feel valued and involved in the developmental processes of their nation.

1.2 The influence of public Satisfaction

Since public satisfaction's importance in the social life, the impact of it might extends across various domains in the society, includes:

Political Legitimacy: Satisfied citizens are more likely to view their government as legitimate, which strengthens democratic processes and stabilizes political systems.

Economic Development: There is a strong correlation between public satisfaction and economic performance. Satisfied citizens contribute more actively to the economy, which can lead to increased productivity and economic growth.

Social Cohesion: Public satisfaction can lead to greater social cohesion, as citizens who are content with their lives and governance are less likely to engage in disruptive behaviors and more likely to participate in community-building activities.

1.3 How we can benefit from finding out the factors that influence public Satisfaction

Based on the importance and far-reaching influence of public satisfaction, understanding the factors that influence public satisfaction provides multiple benefits:

Enhanced Policy-Making: By identifying the drivers of satisfaction, policymakers can design targeted interventions that directly address the areas of concern, leading to more effective and efficient governance.

Predictive Insights: Analyzing these factors can also offer predictive insights into future trends in public behavior and expectations, allowing governments to proactively adjust to changing dynamics.

Resource Allocation: Knowing what influences public satisfaction helps in the optimal allocation of resources. Governments can prioritize spending and initiatives in areas that will most improve public satisfaction, thereby maximizing the impact of public expenditures.

By delving into the causes and effects of public satisfaction, this study aims to provide actionable insights that can lead to more responsive governance and an improved quality of life for citizens. This comprehensive understanding is essential for building a society where the government and its citizens are in a continuous and constructive dialogue.

2. Data Resource and Methodology for Empirical Analysis

2.1 Data Resource

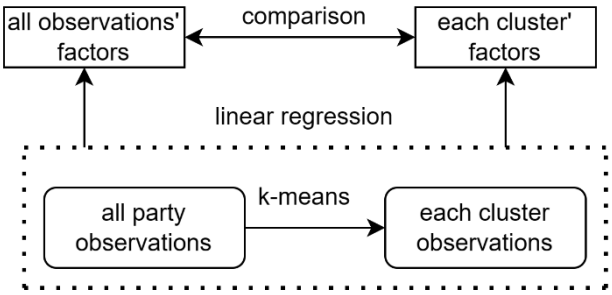
This paper aims to explore the factors influencing public satisfaction, making it essential to utilize data closely related to this topic. For our empirical analysis, we have chosen the "Life in Australia™ Wave 83 Questionnaire," a dataset developed and administered by the Australian National University. Known for its strict academic and ethical standards, the survey's design features robust methodological practices, including randomized sampling, which ensures that the data is representative of the entire Australian population. Moreover, it undergoes a thorough ethical review process, especially important given its focus on sensitive issues, including those concerning Aboriginal and Torres Strait Islander peoples. These rigorous standards not only bolster the credibility and reliability of the findings but also establish the questionnaire as a trusted resource in academic and policy-making circles.

2.2 Methodology: An overall view

The primary methodology employed in this paper will be linear regression, which is well-suited for elucidating the relationships between various factors. To examine the differences in influence factors across different political parties, we will initially use the k-means clustering algorithm to categorize the parties. This involves clustering the parties into three distinct groups based on the averages of certain attributes reflecting party support. After establishing these groups, we will apply linear regression to the dataset encompassing all observations. Subsequently, separate linear regression analyses will be conducted for each cluster to do the parallel experiment to further understand the specific dynamics within each group. This approach allows for a detailed comparative analysis across different political affiliations.

Additionally, in the linear regression analysis, we will conduct a robustness test by reducing the number of control variables to assess the stability of the regression results which is crucial for ensuring the reliability of our findings.

Hence, the overview of the methodology of this paper will be shown below:



3. Data Preprocessing and Variable Selecting

Before going into the empirical analysis, it is essential to get and brief understanding of our dataset, try to preprocess the data into the way we want and select those variables that content information we would like to analyze, this part will give a view of how the paper is doing data preprocessing and variable selecting for cluster part and linear regression part.

3.1 Overview of Dataset

The dataset 'Life in Australia™ Wave 83 Questionnaire' contains 4,219 observations, each with 152 attributes. Most of the attributes measure the extent of agreement or disagreement with various statements. Each attribute can have up to 2,932 missing values. Besides missing data, some attributes include values such as [-99, -98, -97, 97, 98, 99], which represent responses like 'don’t know' or 'refuse to answer' and do not provide much

useful information. Additionally, certain attributes, such as Z1, which have a large number of unique values, pose challenges for empirical analysis.

3.2 Data Preprocessing

Due to the presence of low-information values in our dataset and the large number of observations, we have decided to convert values such as [-99, -98, -97, 97, 98, 99] to NA. We will focus our analysis solely on observations that are complete and contain useful information.

I used Python to transform values such as [-99, -98, -97, 97, 98, 99] into NAs. By making this change, we improved the quality of our dataset, although it required sacrificing some observations.

3.3 Variable Selecting for Party Clustering

For the party clustering task, we want to select variables that could indicate why people might support a particular party. Consequently, police trends are an excellent factor that reflects the reasons for their support. Therefore, we have chosen variables related to respondents' trust in different groups of people (RC7X) and the degree of political inclination they perceive in themselves (RC8) for this task.

After selecting the variables, we need to examine the range of each to determine if further data transformation is necessary to equalize their importance. For the variables under RC7X, the range is from 1 to 4, while for RC8, it extends from 0 to 10. These ranges are relatively close, making rescaling unnecessary.

In conclusion, there is 8 variables selected in party clustering task, RC7_a, RC7_b, RC7_c, RC7_d, RC7_e, RC7_f, RC7_g and RC8.

3.4 Variable Selecting for Factor Analyze

For the linear regression component of our factor analysis, we aim to select variables that provide a wealth of information while avoiding issues of multicollinearity. Therefore, we are choosing variables that encompass a broad range of information and will use a correlation matrix to assess the extent of multicollinearity.

For the dependent variable, we have selected A1, which reflects public satisfaction with national development and meets the requirements of our analysis tasks.

For the independent variables, we have chosen:

A3 to reflect life satisfaction,

E10 to reflect income,

E11a to indicate satisfaction with income,

RA1 to reflect political engagement,

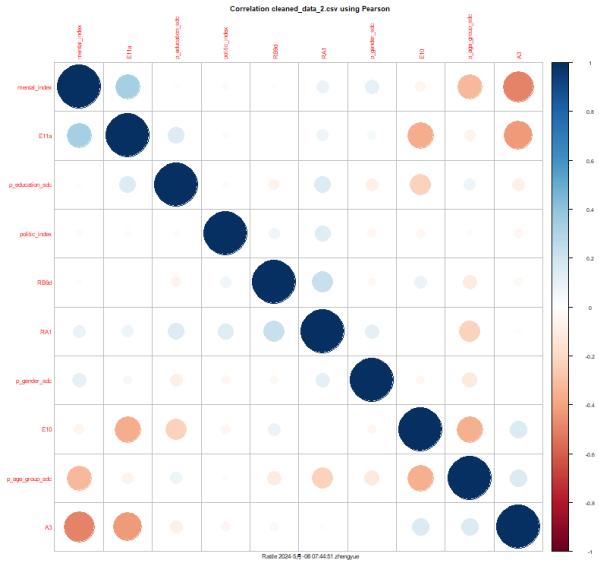
RB9d to reflect closeness to a political party,

p_gender_sdc, p_education_sdc, p_age_group_sdc to capture general personal information.

Additionally, we have created two new variables, mental_index and political_index, as control factors. We will test the robustness of our model by omitting these variables in subsequent experiments to see if the results of the linear regression remain stable. These indices are derived by selecting the maximum values from the sets [D1_a, D1_b, D1_c, D1_d, D1_e, D1_f] and [RA2_a, RA2_b, RA2_c, RA2_d, RA2_e], respectively.

Similarly, the value range of the selected variables is also close that we do not need to implement, and kind of data rescale tasks.

After selecting the variables for linear regression, I used the Pearson correlation matrix to check for multicollinearity. The correlation graphs shown below indicate that multicollinearity is not significant, confirming that the choice of variables is reasonable.



4 Party Clustering

4.1 Party Attributes Compute

Based on the experiment we have designed, we need to compute the mean attributes of the selected variables for each party. Using Python, we can calculate the mean attributes for each party as follows:

PartyID	RC7_a	RC7_b	RC7_c	RC7_d	RC7_e	RC7_f	RC7_g	RC8
1	2.37	2.57	2.13	2.63	2.57	2.61	3.29	6.66
2	2.53	2.83	2.26	2.72	2.73	2.64	3.27	6.95
3	2.14	2.1	1.92	2.5	2.55	2.99	3.29	3.31
4	2.37	2.36	2.1	2.8	2.68	3.38	3.3	2.1
5	2.36	2.53	2.1	2.66	2.66	2.54	3.29	6.81
6	3.24	3.3	2.92	3.27	3.41	3	3.16	6.69
7	2.17	2.33	1.67	3	2.5	3.08	3.42	4.17
8	3	3	3	4	2	3	2	4
9	2.5	3	2.5	3.5	3.5	3	2.5	7
10	3	4	3	3	4	2	4	5
11	2.2	2.2	2	2.8	2.2	3.4	2.8	2.2
15	2.8	3.4	2.2	3.4	3.8	2.6	2.8	8
16	2	2	2	3	3	4	4	1
18	3	2.67	2	2.67	3	1.67	3	6.33
22	3	3	2	4	4	4	3	5
23	1.5	1.5	1.5	2.5	2	3	4	4.5
32	3.5	3.5	3	3.5	3.5	4	3.5	0.5
33	2.33	3	2.33	2.67	3	2.67	3.67	3.33
34	2.5	3.5	3.5	3	3	2.5	3	7.5
36	4	4	2	4	3	4	3	10
37	3.33	3.33	2.67	3.67	3	3.67	3.67	2

4.2 Party Clustering by Mean Attributes

After computing the mean attributes, we can perform k-means clustering on the data we obtained in Section 4.1. We will cluster the 18 different parties into 3 distinct groups as follows:

Group Number	Party ID
0	6, 9, 10, 15, 18, 34
1	1, 2, 3, 4, 5, 7, 11, 16, 23, 33
2	8, 22, 32, 36, 37

Which is:

Cluster 0: Pauline Hanson's One Nation, Katter's Australian Party (KAP), Liberal Democratic Party, United Australia Party, Australian Christians, The Great Australian Party.

Cluster 1: Liberal, Nationals, Australian Labor Party, The Greens, Liberal National Party of Queensland, Independent, Animal Justice Party, FUSION: Science, Pirate, Secular, Climate Emergency, Australian Democrats, Sustainable Australia Party - Stop Overdevelopment / Corruption.

Cluster 2: Shooters, Fishers and Farmers Party, Australian Citizens Party, Socialist Alliance, TNL, Victorian Socialists, Western Australia Party.

5 Linear Regression Factor Analyze

5.1 Linear Regression for All Observations

Basing for the variables we just selected, we implement linear regression using numerical regression model via Rattle, getting the result:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.063980	0.337868	14.988	< 2e-16 ***
A3	-0.287182	0.019166	-14.984	< 2e-16 ***
E10	-0.015828	0.012893	-1.228	0.21979
E11a	0.045722	0.037756	1.211	0.22610
RA1	-0.200438	0.041187	-4.867	0.00000126 ***
RB9d	0.024022	0.050110	0.479	0.63174
p_gender_sdc	0.174957	0.056649	3.088	0.00205 **
p_age_group_sdc	-0.068032	0.033711	-2.018	0.04377 *
p_education_sdc	-0.004587	0.015488	-0.296	0.76717
mental_index	0.030395	0.030887	0.984	0.32525
politic_index	-0.032652	0.034464	-0.947	0.34359

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.033 on 1418 degrees of freedom

Multiple R-squared: 0.2379, Adjusted R-squared: 0.2325

F-statistic: 44.26 on 10 and 1418 DF, p-value: < 2.2e-16

We can see that A3, RA1, p_gender_sdc, and p_age_group_sdc are statistically significant in relation to A1, with an adjusted R-squared of 0.2325.

After getting the regression result, we do robustness test by not performing linear regression without control variables mental_index and politic_index.

Getting the result:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.078532	0.276326	18.379	< 2e-16 ***
A3	-0.293574	0.017578	-16.701	< 2e-16 ***

E10	-0.015439	0.012852	-1.201	0.22983
E11a	0.053624	0.037272	1.439	0.15046
RA1	-0.204734	0.040879	-5.008	0.000000618 ***
RB9d	0.021046	0.050058	0.420	0.67424
p_gender_sdc	0.182292	0.056384	3.233	0.00125 **
p_age_group_sdc	-0.076122	0.032545	-2.339	0.01947 *
p_education_sdc	-0.004948	0.015481	-0.320	0.74931

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.033 on 1418 degrees of freedom
Multiple R-squared: 0.2379, Adjusted R-squared: 0.2325
F-statistic: 44.26 on 10 and 1418 DF, p-value: < 2.2e-16

In the robustness test, the R-squared value remains nearly constant, and the significance of the variables changes only slightly. Based on these findings, we can conclude that the model remains robust when altering the control variables. The model has passed the robustness test and is therefore valid.
In conclusion, we identified three variables that are negatively correlated with public satisfaction: A3, which reflects life satisfaction; RA1, which relates to political engagement; and p_age_group_sdc, which pertains to age. Additionally, we found that one variable, p_gender_sdc, which reflects gender, is also related to public satisfaction.

5.2 Linear Regression for Each Cluster

Implementing linear regression for each cluster respectively, we get the result above.
For cluster0:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.94618	2.81552	1.402	0.1764
A3	-0.23764	0.13225	-1.797	0.0875
E10	-0.08417	0.13500	-0.623	0.5400
E11a	-0.46006	0.44494	-1.034	0.3135
RA1	0.16551	0.23901	0.692	0.4966
RB9d	0.42947	0.38380	1.119	0.2764
p_gender_sdc	0.78377	0.48375	1.620	0.1208
p_age_group_sdc	0.11946	0.32206	0.371	0.7146
p_education_sdc	0.06784	0.15209	0.446	0.6604

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.153 on 20 degrees of freedom
Multiple R-squared: 0.2966, Adjusted R-squared: 0.01523
F-statistic: 1.054 on 8 and 20 DF, p-value: 0.4313

For cluster1:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.021720	0.271658	18.485	< 2e-16 ***
A3	-0.284249	0.017562	-16.186	< 2e-16 ***
E10	-0.009727	0.012542	-0.776	0.43813
E11a	0.044306	0.036504	-1.034	0.3135
RA1	-0.180627	0.038095	-4.741	0.00000232

				***
RB9d	-0.018271	0.048681	-0.375	0.70748
p_gender_sdc	0.170723	0.054984	3.105	0.00194 **
p_age_group_sdc	-0.058942	0.031941	-1.845	0.06518
p_education_sdc	-0.011204	0.014748	-0.760	0.44755

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.039 on 1517 degrees of freedom
Multiple R-squared: 0.2111, Adjusted R-squared: 0.207
F-statistic: 50.75 on 8 and 1517 DF, p-value: < 2.2e-16

For cluster2:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.0000	NA	NA	NA
A3	-0.3333	NA	NA	NA
E10	0.3333	NA	NA	NA
E11a	0.6667	NA	NA	NA
RA1	NA	NA	NA	NA
RB9d	NA	NA	NA	NA
p_gender_sdc	NA	NA	NA	NA
p_age_group_sdc	NA	NA	NA	NA
p_education_sdc	NA	NA	NA	NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared: 1, Adjusted R-squared: NaN
F-statistic: NaN on 3 and 0 DF, p-value: NA

We can see that only the results from Cluster 1 align with the results from all observations, while the other linear regression analyses do not show any reasonable results. Returning to the party clustering section, the parties have been divided into three groups. However, for the parties in Clusters 0 and 2, the number of respondents is very limited, providing insufficient data for linear regression. Consequently, the results from Cluster 1 dominate the linear regression analysis as it represents the 'majority'.

5.3 Experiment Conclusion

For the experiment given above, the idea of separating party and seeing the difference between influence factor in different kinds of party falls. However, that is still some conclusion we can draw from the general linear regression for all observations. That are: 1. Life satisfaction is negatively correlated with public satisfaction, indicating that as life satisfaction increases, public satisfaction decreases. 2. Political engagement trendency is negatively correlated with public satisfaction, suggesting that higher levels of political involvement may lead to lower public satisfaction. This finding indicates that more politically engaged individuals could be more critical of national developments. 3. Age also shows a negative correlation with public satisfaction, implying that older age groups may experience lower levels of satisfaction. This trend highlights possible generational differences in expectations and perceptions of national progress. 4. Gender, reflected by the variable p_gender_sdc, shows a relationship with public satisfaction, reflect that male is more likely to be satisfied with national development.

6. Influence Factor Analysis and Insight

Based on the influencing factors identified from the experiments above, aside from immutable characteristics

such as age and gender, we find that there are two additional factors that can influence public satisfaction with national development: life satisfaction and political engagement. Although it is challenging to determine exactly how these factors relate, we can hypothesize. It might be that individuals who are more satisfied with their personal lives could be more critical of public development, perhaps due to higher expectations or a greater awareness of societal issues. Additionally, greater willingness to engage in politics might correlate with increased dissatisfaction, as those more involved may be more aware of and sensitive to policy failures.

As the hypotheses suggest, we can gain valuable insights for public management: it is crucial to manage people's expectations and increase their political engagement. Implementing transparent communication strategies and educational initiatives can set realistic expectations about the timelines and complexities of national development. Furthermore, creating accessible platforms for political engagement and establishing robust feedback mechanisms can empower citizens to actively participate in the political process. These efforts can foster a more informed and involved citizenry, ultimately leading to higher satisfaction with public development initiatives.

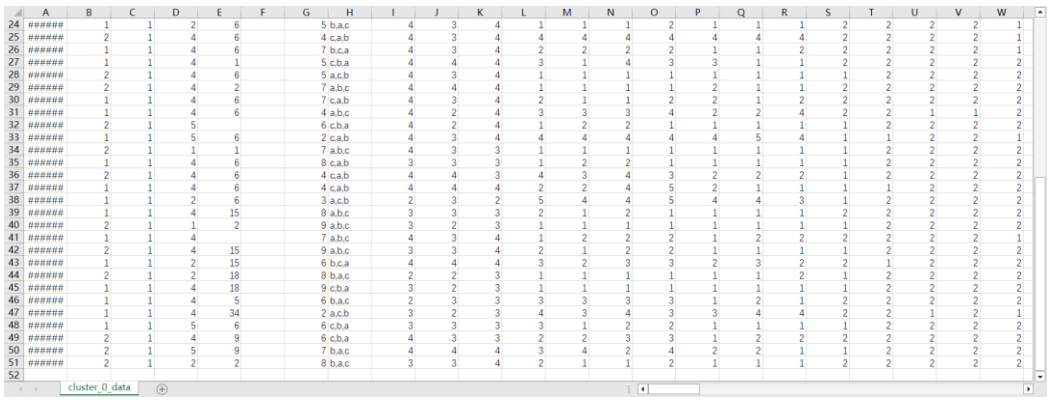
Appendix:

1 Cluster result for party cluster:

```
RB9c
1.0     1
2.0     1
3.0     1
4.0     1
5.0     1
6.0     0
7.0     1
8.0     2
9.0     0
10.0    0
11.0    1
15.0    0
16.0    1
18.0    0
22.0    2
23.0    1
32.0    2
33.0    1
34.0    0
36.0    2
37.0    2
Name: cluste
```

2. Observations in the minority clusters:

Cluster0:



Cluster2:

	IntDate	s_order	Mode	A1	A6	A6_VERB	A3	A4_order	A4_a	A4_b	A4_c	D1_a	D1_b	D1_c	D1_d	D1_e	D1_f	D3	E1_a	E1_b	E1_c	E1_d	E1_e	E
2	######	1	1	4	22		6	a,c,b	3	3	3	2	1	1	2	1	1	1	2	2	2	2	2	
3	######	1	1	5	3		1	c,b,a	3	3	3	2	4	4	4	3	4	2	2	2	2	2	1	
4	######	1	1	4	32		7	b,c,a	4	2	3	2	1	2	3	1	1	4	2	1	1	2	1	
5	######	1	1	4	36		7	c,b,a	4	3	3	2	1	2	2	1	1	1	1	2	2	2	2	
6	######	1	1	4	37		5	c,b,a	3	3	3	3	3	2	2	2	3	3	1	2	2	2	2	
7	######	2	1	5	4		2	b,c,a	4	3	3	4	5	4	5	5	5	4	2	2	2	2	2	
8	######	1	1	2	37		6	c,b,a		2		1	2	2	3	3	3	4	2	2	2	2		
9	######	2	1	4	8		2	b,c,a	3	3	3	3	4	3	5	4	5	4	1	2	1	2	2	

3. Linear regression for the main model by Rattle

```
Summary of the Linear Regression model (built using lm):

Call:
lm(formula = A1 ~ ., data = crs$dataset[crs$train, c(crs$input,
    crs$target)])

Residuals:
    Min      1Q  Median      3Q     Max
-4.0129 -0.7539 -0.3248  0.9377  2.8282

Coefficients:
                 Estimate Std. Error t value   Pr(>|t|)
(Intercept)      5.063980   0.337868  14.988    < 2e-16 ***
A3              -0.287182   0.019166 -14.984    < 2e-16 ***
E10             -0.015828   0.012893  -1.228    0.21979
E11a             0.045722   0.037756   1.211    0.22610
RA1             -0.200438   0.041187  -4.867 0.00000126 ***
RB9d             0.024022   0.050110   0.479    0.63174
p_gender_sdc     0.174957   0.056649   3.088    0.00205 **
p_age_group_sdc -0.068032   0.033711  -2.018    0.04377 *
p_education_sdc -0.004587   0.015488  -0.296    0.76717
mental_index     0.030395   0.030887   0.984    0.32525
politic_index   -0.032652   0.034464  -0.947    0.34359
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.033 on 1418 degrees of freedom
Multiple R-squared:  0.2379,    Adjusted R-squared:  0.2325
F-statistic: 44.26 on 10 and 1418 DF,  p-value: < 2.2e-16
```

```
==== ANOVA ====

Analysis of Variance Table

Response: A1
                Df  Sum Sq Mean Sq  F value    Pr(>F)
A3               1  426.64  426.64 399.5074 < 2.2e-16 ***
E10              1    0.98    0.98   0.9151 0.3389269
E11a             1    2.37    2.37   2.2232 0.1361724
RA1              1   20.75   20.75  19.4330 0.0000112 ***
RB9d             1    0.08    0.08   0.0769 0.7815397
p_gender_sdc     1   13.84   13.84  12.9633 0.0003286 ***
p_age_group_sdc  1    5.91    5.91   5.5298 0.0188312 *
p_education_sdc  1    0.11    0.11   0.1021 0.7493260
mental_index     1    1.02    1.02   0.9513 0.3295580
politic_index    1    0.96    0.96   0.8976 0.3435914
Residuals     1418 1514.32    1.07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
[1] "\n"
Time taken: 0.02 secs

Rattle timestamp: 2024-05-06 05:35:43 zhengyue
=====================================================================
```

3. Linear regression for the robustness test model by Rattle

```
Summary of the Linear Regression model (built using lm):

Call:
lm(formula = A1 ~ ., data = crs$dataset[crs$train, c(crs$input,
    crs$target)])

Residuals:
    Min      1Q  Median      3Q     Max
-4.0189 -0.7650 -0.3141  0.9443  2.8425

Coefficients:
                 Estimate Std. Error t value    Pr(>|t|)
(Intercept)      5.078532   0.276326  18.379     < 2e-16 ***
A3              -0.293574   0.017578 -16.701     < 2e-16 ***
E10             -0.015439   0.012852  -1.201     0.22983
E11a             0.053624   0.037272   1.439     0.15046
RA1             -0.204734   0.040879  -5.008 0.000000618 ***
RB9d             0.021046   0.050058   0.420     0.67424
p_gender_sdc     0.182292   0.056384   3.233     0.00125 **
p_age_group_sdc -0.076122   0.032545  -2.339     0.01947 *
p_education_sdc -0.004948   0.015481  -0.320     0.74931
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.033 on 1420 degrees of freedom
Multiple R-squared:  0.2369,    Adjusted R-squared:  0.2326
F-statistic:  55.1 on 8 and 1420 DF,  p-value: < 2.2e-16
```

```
==== ANOVA ====

Analysis of Variance Table

Response: A1
                Df  Sum Sq Mean Sq  F value     Pr(>F)
A3               1  426.64  426.64 399.5499 < 2.2e-16 ***
E10              1    0.98    0.98   0.9152  0.3389010
E11a             1    2.37    2.37   2.2234  0.1361512
RA1              1   20.75   20.75  19.4350 0.00001119 ***
RB9d             1    0.08    0.08   0.0769  0.7815283
p_gender_sdc     1   13.84   13.84  12.9647  0.0003283 ***
p_age_group_sdc  1    5.91    5.91   5.5304  0.0188247 *
p_education_sdc  1    0.11    0.11   0.1021  0.7493130
Residuals     1420 1516.29    1.07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
[1] "\n"
Time taken: 0.02 secs

Rattle timestamp: 2024-05-06 08:05:12 zhengyue
=====================================================================
```

4. Linear regression for the cluster test model by Rattle

Cluster0:

```
Summary of the Linear Regression model (built using lm):

Call:
lm(formula = Al ~ ., data = crs$dataset[crs$train, c(crs$input,
    crs$target)])

Residuals:
    Min      1Q  Median      3Q     Max
-2.3849 -0.4623  0.2290  0.6913  1.4367

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.60857    3.06627   0.525   0.6063
A3              -0.06704    0.15842  -0.423   0.6772
E10             -0.01879    0.13468  -0.140   0.8906
E11a            -0.60218    0.43521  -1.384   0.1834
RA1              0.21589    0.23399   0.923   0.3684
RB9d             0.52611    0.38142   1.379   0.1847
p_gender_sdc     0.83471    0.46702   1.787   0.0907 .
p_age_group_sdc  0.08814    0.31618   0.279   0.7836
p_education_sdc  0.11734    0.15457   0.759   0.4576
mental_index     0.50945    0.27118   1.879   0.0766 .
politic_index   -0.10459    0.33267  -0.314   0.7568
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.11 on 18 degrees of freedom
Multiple R-squared:  0.4128,    Adjusted R-squared:  0.0866
F-statistic: 1.265 on 10 and 18 DF,  p-value: 0.3184

==== ANOVA ====

Analysis of Variance Table

Response: Al
                Df  Sum Sq Mean Sq F value  Pr(>F)
A3               1  2.7741  2.7741  2.2502 0.15094
E10              1  0.4018  0.4018  0.3259 0.57515
E11a             1  2.7073  2.7073  2.1959 0.15567
RA1              1  0.4772  0.4772  0.3870 0.54166
RB9d             1  1.1982  1.1982  0.9719 0.33728
p_gender_sdc     1  3.1460  3.1460  2.5518 0.12757
p_age_group_sdc  1  0.2400  0.2400  0.1947 0.66428
```

Cluster1:

```
Summary of the Linear Regression model (built using lm):

Call:
lm(formula = Al ~ ., data = crs$dataset[crs$train, c(crs$input,
    crs$target)])

Residuals:
    Min      1Q  Median      3Q     Max
-3.8777 -0.7521 -0.3684  0.9514  3.2754

Coefficients:
                 Estimate Std. Error t value   Pr(>|t|)
(Intercept)      5.021720   0.271658  18.485    < 2e-16 ***
A3              -0.284249   0.017562 -16.186    < 2e-16 ***
E10             -0.009727   0.012542  -0.776    0.43813
E11a             0.044306   0.036504   1.214    0.22503
RA1             -0.180627   0.038095  -4.741 0.00000232 ***
RB9d            -0.018271   0.048681  -0.375    0.70748
p_gender_sdc     0.170723   0.054984   3.105    0.00194 **
p_age_group_sdc -0.058942   0.031941  -1.845    0.06518 .
p_education_sdc -0.011204   0.014748  -0.760    0.44755
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.039 on 1517 degrees of freedom
Multiple R-squared:  0.2111,    Adjusted R-squared:  0.207
F-statistic: 50.75 on 8 and 1517 DF,  p-value: < 2.2e-16

==== ANOVA ====

Analysis of Variance Table

Response: Al
                Df  Sum Sq Mean Sq  F value     Pr(>F)
A3               1  398.34  398.34 369.3450  < 2.2e-16 ***
```

Cluster2:

```
Summary of the Linear Regression model (built using lm):

***Note*** Singularities were found in the modeling
and are indicated by an NA in the following table.
This is often the case when variables are linear
combinations of other variables, or the variable
has a constant value.  These variables will be ignored
when using the model to score new data and will not be
included as parameters in the exported scoring routine.

Call:
lm(formula = Al ~ ., data = crs$dataset[crs$train, c(crs$input,
    crs$target)])

Residuals:
ALL 4 residuals are 0: no residual degrees of freedom!

Coefficients: (5 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.0000        NaN     NaN      NaN
A3               -0.3333        NaN     NaN      NaN
E10               0.3333        NaN     NaN      NaN
E11a              0.6667        NaN     NaN      NaN
RA1                   NA         NA      NA       NA
RB9d                  NA         NA      NA       NA
p_gender_sdc          NA         NA      NA       NA
p_age_group_sdc       NA         NA      NA       NA
p_education_sdc       NA         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:    NaN
F-statistic:   NaN on 3 and 0 DF,  p-value: NA

==== ANOVA ====

Analysis of Variance Table

Response: Al
          Df Sum Sq Mean Sq F value Pr(>F)
A3         1 5.4915  5.4915     NaN    NaN
E10        1 0.4132  0.4132     NaN    NaN
E11a       1 0.0952  0.0952     NaN    NaN
Residuals  0 0.0000     NaN
[1] "\n"
Time taken: 0.01 secs

Rattle timestamp: 2024-05-06 06:59:41 zhengyue
```