Australian National University

COMP8410-Data Mining

Semester 1, 2022

The Analysis and Discovery of Data Mining in COVID-19 vaccinations attitudes and behaviors

Course Code: COMP8410

Course Name: Data Mining

Lecturer's Name: Kerry Taylor & Pouya Omran

Tutor's Name: Cheng Xue and Zhangcheng Qiang

Date/time due: Monday Week 10 (9th of May 2022) / 9am

Student number	Student name	Contribution	Signature
u7167784	Man Jin	100%	Man Jin

Table of Contents

1. Problem Description

1.1 Goals

In recent years, COVID-19 has been a popular topic, and the virus has been evolving. Every day, the number of cases around the world grows. The most efficient method of preventing infectious diseases is vaccination. Efforts to produce vaccinations against COVID-19 types are being stepped up by governments. People's attitudes towards vaccines, on the other hand, have long been a global issue. Vaccine apprehension and rejection, particularly for new pandemic vaccines, can be significant hurdles to vaccination (Danchin, Biezen, Manski-Nankervis, Kaufman & Leask, 2020). Using association mining, decision trees, neural networks, and logistic regression technologies, this report uses ANU poll 35 (August 2020): COVID-19 attitudes and behaviors (wave 3) as the data source to forecast people's attitudes regarding COVID-19 vaccination.

1.2 Expected Impact

To improve people's vaccine aspirations, the health-care system should increase the risks and benefits of COVID-19 vaccines, as well as provide free vaccine to everyone. The first stage is to learn about society's attitudes towards the vaccine. A questionnaire can be used to gain direct feedback from the participants. However, the classification technique can also be used to predict data and predict future outcomes. In other words, vaccine attitudes can be forecasted to better understand the features of people with varied vaccine attitudes. The findings will assist governments and health authorities in encouraging hesitant people to get vaccinated and ensuring that these new vaccines are widely distributed (Seale et al., 2021).

1.3 Data Mining

By learning the training set to map each input feature to the model, some classification algorithms are used as models. The model establishment stage, or training stage, is the first step in the classification process, and the evaluation stage is the second. The training phase's goal is to define the classification model for pre-defined data classes or concept sets. We should now select a sample of the data from the known data set as the model's training set and the rest as the testing set. In the evaluation stage, we need to use the model established in the first stage to classify the testing set data tuples, to evaluate the prediction accuracy of the classification model.

First, association mining is used in this report to discover associations between items in the data set, which can lead to the discovery of multiple frequently associated data items. Then, as the training set for the classification model, we can use these features that are highly related to vaccination attitude, and then apply them to the testing set for model evaluation.

2. Data Description

2.1 Basic Description and Deep Understanding

This report uses the original data version of the ANU (Australian National University) Poll 35 (August 2020): COVID-19 attitudes and behaviors (Wave 3). The data was collected in August 2020 to assess Australians' views on some important or topical issues and can be used for providing a timely update during the COVID-19 pandemic. The questionnaire was divided into seven sections: ANU questions, Experiences with COVID-19, Opinions and behavioral responses to COVID-19 vaccine, Mental health, Employment, income and financial hardship, Opinions and behavioural responses to COVID-19 policy, and Opinions and behavioural responses to bushfire policy. There are 335 attributes and 3061 data records in the original data file. The table below summarises the data attributes and population. Please see the following table for more information. For the detailed statistics, please see figures on the appendix, and these figures are described by Tableau.

	Basic Information	Section A	Section B	Section V	Section D
Description	*The unique serial number, Date of survey completion, Mode of survey completion, Order of code frames for specified questions*	*ANU questions*	*Experiences with COVID-19*	*Opinions and behavioral responses to COVID-19 vaccine*	*Mental health*
Number of topics	0 topic	4 topics	5 topics	2 (1 for web users) topics	2 topics
Type	Numeric and Date	Numeric	Numeric	Numeric	Numeric
Attributes	Column 1-4	Column 5-15	Column 16-30	Column 31-116	Column 117-123
Quality assessment	High level	Medium-high level	Medium level	Medium level	High level
Statistics (completion ratio)	100%	99%	100%	100%	100%

Table 1. General Summary of the original data set

	Section E	Section G	Section F	Confidential Data (Confidential)	Confidential Data
Description	*Employment, income and financial hardship*	*Opinions and behavioural responses to COVID-19 policy*	*Opinions and behavioural responses to bushfire policy*	*Paticipants personal information (All are restricted value)*	*Paticipants personal information (Gender, Age group, State, StateMap)*
Number of topics	18 topics	1 (All for web users) topics	7 topics	0 topic	0 topic
Type	Numeric	Numeric	Numeric	Numeric	Numeric
Attributes	Column 124-155	Column 156-288	Column 289-316	Column 317-329	Column 331-335
Quality assessment	Medium-low level	Low level	High level	Low level	High level
Statistics (completion ratio)	100%	100%	100%	Confidential	100%

Table 1. General Summary of the original data set

For the quality assessment, the table only shows various levels for the basic view of the data quality. The following is a more detailed explanation. The **basic information** is marked as high level because every data can be used and there are not contains useless information such as “-99” and “-98”. **Section A** gets a medium-high level due to the column named “A2_oth” only has one record and others are empty. **Section B** has 5 questions but the values of B6a-B6e are “-99” and they cannot be used in this report. In the **section V**, the V1 is treated as target value, but the V2 (column 32-116) does not have enough contextual information. **Section D** and **Section F** have a high level of data quality and they can be used in data analysis and data mining. There are many empty values appear in **section E**, E10 and E14 are questions about total income and household population respectively, but the table shows "- 99". Obviously, due to many null and bad values, its quality level is medium-low level. For **section G**, this part data can be ignored that does not have enough understandable metadata. For confidential data is about participants personal information, so the column 317-329 shows “-99”. However, the personal information about participants’ gender, age group, state, and state map open for the users. After data cleaning, only 95 attributes left.

The following is a more specific explanation of the basic statistical description, and it is based on ignorance of understandable and meaningless metadata. This report uses some data of the Section A, Section B, Section V, and Section D, which to do a detailed explanation. The basic information shows that there are 3061 participants attended the questionnaire during the 10/08/2020 to 24/08/2020 period, 94% chose the online mode and 6% chose the phone mode to finish the questionnaire.

About section A, 1,886 people chose to be satisfied with the development direction of Australia, accounting for the largest proportion. The largest number of people chose the Liberal Party, and the Labor Party ranked second. Life satisfaction presents a skewed distribution, with satisfaction increasing from 0 to 8, reaching the peak at 8, and then decreasing (Figure 1). Australians are most satisfied with the federal government in Canberra, with an average confidence rating of 2.27. Less confidence in organizations responsible for firefighting in regional or rural areas (Figure 2). Section B shows most people have had close contact with someone who has had a confirmed infection of COVID-19 (Figure 3). However, 57% of participants believed that they would not be very likely to be infected with COVID-19 in the next six months. People think they need to wear a mask when there is a high level of COVID-19, and they agree that can prevent an increase in COVID-19 (Figure 4). For Section V: response to COVID-19 vaccine, most people are willing to be vaccinated. Only 158 people are completely reluctant to be vaccinated (Figure 5). For Section D, 1905 respondents thought that relatively or none of the time (less than 1 day) had felt lonely in the past week. The average value of negative attitude towards life is 1.85 (the most negative is 5) (Figure 6).

For the deep understanding, the attitude of COVID-19 vaccines as the target value of this report, it has important research significance. According to the basic statistics, we can infer that woman over the age of 65, living in Queensland and with the highest education level of secondary school or less are more willing to be vaccinated. However, this is obviously inaccurate. We cannot predict what affects people's acceptance of the vaccine only by simple basic statistics. It is also

important to look at Australians' satisfaction with national development direction, life, and confidence in institutions as input features to predict people's views and attitudes towards vaccines. In addition, this report will also study COVID-19 experience and mental health attitudes towards vaccines to classify and predict by data mining technology.

3. Data Preprocessing

The data is properly preprocessed and then applied to the data mining method in the fourth part. In this report, V1 column (options and behavioral responses to COVID-19 vaccine) are selected as the target value for prediction. There are 1, 2, 3 and 4 in the source data, and the degree of vaccination consent increases in turn. Therefore, participants who choose 4 are recoded as 1, while those who choose 1, 2 and 3 are recoded as 0. Then do the data normalization.

Because A1, A3, A4, B1, B4, B6, B7, D1 and D3 are used as input features in this report, their options should be preprocessed next. To begin, it is discovered that A4, B1, B6, and D1 have multiple options, and each option is chosen separately. When the characteristics of each option are examined, they are not black and white, but have degree or frequency characteristics. The average value can be calculated by combining the options from multiple topics into one value and then normalizing them. You can directly normalize the remaining questions based on the number of options. For the association mining, the numeric value also needs to be transformed to categorical value, which can be done by Rattle under Transform-Recode tab.

4. Methods Description and Results

4.1 Association Mining

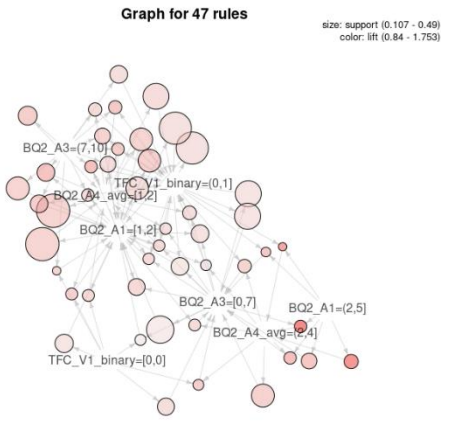
4.1.1 Parameter Setting

The default support is 0.1, but the confidence should be higher than the default, so I set it as 0.6. The amount of occurrences/frequency of occurrences of an item set in the entire transaction set is referred to as support. The frequency of simultaneous occurrence of the left and right components of association rules is referred to as confidence. The higher the value, the more likely it is to happen at the same time.

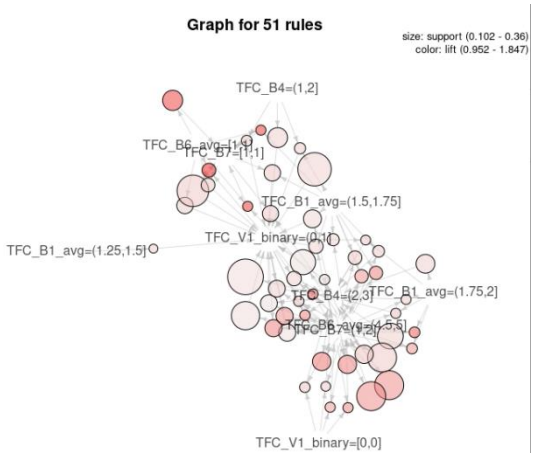
4.1.2 Applicability and Limitations

Although the association rule algorithm's principle is simple and straightforward to implement, the operation time increases significantly as the size of the frequent item set grows. Furthermore, the unique support is used without considering the relative importance of each attribute.

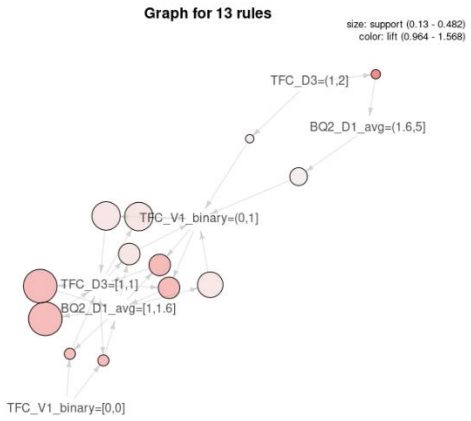
4.1.3 Results



Graph A. Association Mining for Section A



Graph B. Association Mining for Section B



Graph C. Association Mining for Section D

TFC_V1_binary = [0,0] means people definitely not get the vaccine, and TFC_V1_binary = (0,1] means people want to get the vaccine. Graph A presents the higher the satisfaction with the national development direction (A1) / life (A3) / institutions (A4), the more likely they are to be vaccinated. As shown in graph B, people who have had COVID-19 (B1), believe they will be infected with COVID-19 in the next six months (B4), follow epidemic prevention measures (B6), and are willing to wear masks (B7) are more likely to be vaccinated. However, graph C shows a relation between mental health and vaccine attitudes, but based on the size of the pink circle, people who have no negative emotions (D1) and do not feel lonely (D3) are more willing to be vaccinated.

4.2 Decision Tree

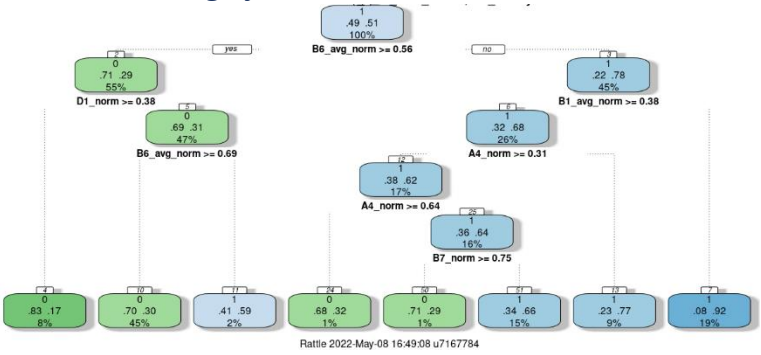
4.2.1 Parameter Setting

For the decision tree model, A1, A3, A4, B1, B4, B6, B7, D1 and D3 are used as input. I chose the default parameters (Min Split: 20, Max Depth: 30, Min Bucket: 7) except complexity, because if the default complexity is selected, the generated decision tree is small, so I chose to increase the complexity to 0.003 to get more branches.

4.2.2 Applicability and Limitations

Decision tree is simple to understand and use, and it can produce feasible results for large data sources in a short amount of time. It is also extremely efficient, which only needs to be built once, and each prediction's maximum calculation time does not exceed the decision tree's depth. However, the decision tree is easy to ignore the correlation between data. For example, B4 is related to vaccine attitude, but it is not displayed in the decision tree.

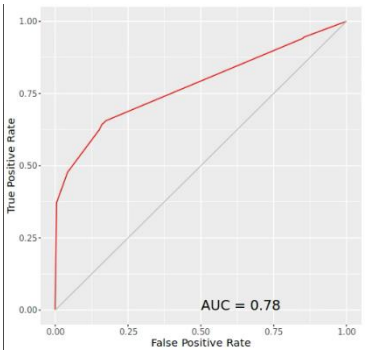
4.2.3 Model Training by Rattle and Results



Graph D. Decision Tree

B6 is the root node. The lower the B6 value, the more people insist on wearing masks both indoors and outdoors. Participants who do not wear masks and have negative emotions (D1_norm > 0.38) will reach the first leaf node, indicating that 83% of them do not get vaccinated. Conversely, if participants tend to wear masks, that is, from the root node to B1 to the right, if the value of B1 is smaller, it indicates that participants have recently experienced COVID-19 and will reach the last leaf node, indicating that 92% of people think they should be vaccinated.

4.2.4 Model Evaluation



Graph E. ROC

Graph E is the ROC curve for the decision tree, it shows the AUC is 0.78. The area under the ROC curve, between 0.1 and 1, can be used as a value to directly evaluate the quality of the classifier. The larger the value, the better.

4.3 Neural Networks

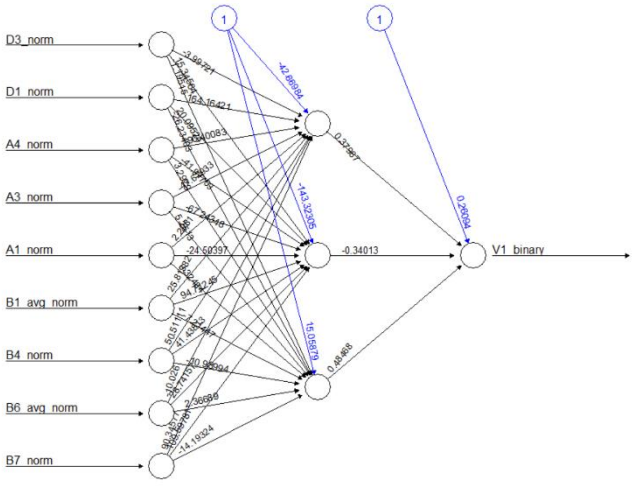
4.3.1 Parameter Setting

A hidden layer and three hidden units (hidden = 3) are set in the neural network model after a series of experiments. The activation function is “logistic” and linear.output=f. The training set contains 75% of the data, while the testing set comprises 25% of the data.

4.3.2 Applicability and Limitations

In a neural network, we usually do not have to worry about the data structure. It can learn almost any type of characteristic variable relationship with ease. These models, however, are difficult to explain and comprehend due to their complexity. In addition, the training period is longer.

4.3.3 Model Training by R and Results



Graph F. Neural Network

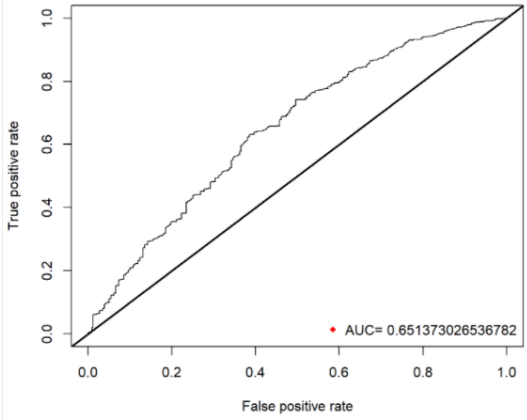
The above figure is the neural network prediction graph. The leftmost node (i.e., input node) is the original data variable. Black arrows (and related numbers) are weights, which are the

contribution of the variable to the next node. The blue line is the offset weight. Intermediate nodes are hidden nodes. Each of these nodes constitutes the component that the network is learning to recognize. The rightmost (output node) node is the final output of the neural network.

4.3.4 Model Evaluation

```
> table(predict,real)
       real
predict   0   1
      0  48  26
      1 212 432
```

```
> print(accuracy)
[1] 0.6685237
> print(precision)
[1] 0.6708075
> print(recall)
[1] 0.9432314
> print(F_measure)
[1] 0.784029
```



Graph G. Evaluation for NN

As can be seen from the above figures, we can see that the confusion matrix is used to determine the number of truths and errors generated by our prediction. The model generated 48 true negatives (0) and 432 true positives (1). Therefore, the accuracy is (TP+TN)/(P+N), which is 0.6685237. The precision is TP/(TP+FP) = 0.6708075. The recall is TP/P = 0.9432314. In addition, the AUC is 0.65, which is not a bad outcome.

4.4 Logistic Regression

4.4.1 Parameter Setting

We split the data into two sets: a training set (75% used to develop the prediction model) and a test set (25% for evaluating model). The fitting model uses GLM () function and the family = binomial option should be selected.

4.4.2 Applicability and Limitations

Logistic regression is straightforward and simple. The model has an elevated level of interpretability. The impact of different features on the results can be seen by looking at their coefficients. However, it can only handle binary classification issues.

4.4.3 Model Training by R and Results

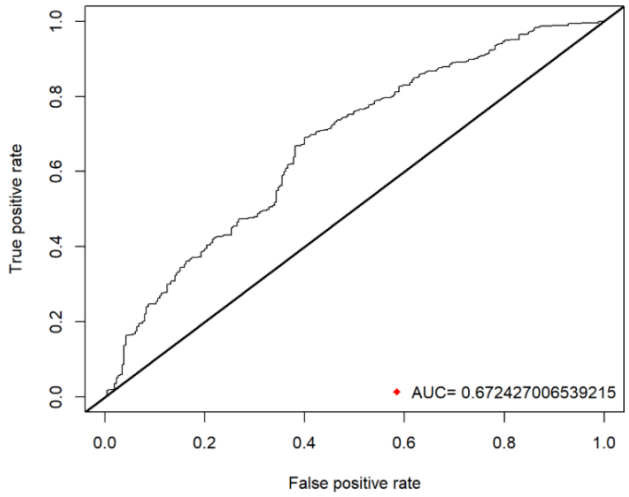
```
> coefficients(pre)
(Intercept)     D3_norm     D1_norm     A4_norm     A3_norm
 2.30522450 -0.03087562 -0.39012357 -1.16788692  0.26241849
    A1_norm B1_avg_norm     B4_norm B6_avg_norm     B7_norm
 0.19340580 -1.09186529 -0.18589606 -1.54161724 -0.46758883
```

We can see from the coefficients that D3's contribution to the equation is negligible, as evidenced by the association mining rules, and there is no D3 node in the decision tree. There was no significant association between whether participants felt lonely and their attitude towards the vaccine because D3 was about loneliness. Participants who wore masks indoors and outdoors, as well as their trust in Australian institutions, were the most linked to vaccine attitudes. People's attitudes towards vaccines will be acceptable if they chose to wear masks and have great trust in Australian government agencies.

4.4.4 Model Evaluation

```
> predict=ifelse(predict_>0.5,1,0)
>
> table(predict,real)
       real
predict   0   1
      0  68  45
      1 197 408
```

```
> print(accuracy)
[1] 0.6629526
> print(precision)
[1] 0.6743802
> print(recall)
[1] 0.9006623
> print(F_measure)
[1] 0.7712665
```



The accuracy is 0.6629526 and it is close to the accuracy value of neural network.

5. Conclusion and Further Work

5.1 Conclusion and Challenges

In Conclusion, although billions of dollars will be invested in the development of the COVID-19 vaccine, the arrival of these expected vaccines does not guarantee the acceptance of the vaccine. To increase vaccine trust in general practice, the government must invest in learning about the factors that influence COVID-19 vaccine acceptance and developing measures with the community to maximize absorption once these vaccines are accessible. This report found that those with high confidence in Australian institutions and high satisfaction with the direction of life and national development tend to receive vaccines, and those who have COVID-19 experience, such as have a COVID-19 testing, also have higher willingness. However, mental health has negligible effect on vaccine attitude.

The challenge of this report is to avoid politicizing Australia's COVID-19 vaccination program, which may lead to partisan differences in people's views on the vaccine. Moreover, people who cannot communicate in English are excluded from the sample, which may affect the representation of multiculturalism. Finally, this data source requires respondents to have access to the Internet, which may limit the participation of some community members. However, given Australia's network infrastructure, this should not be a significant issue.

5.2 Further Work

At present, vaccination is still one of the most effective tools to combat COVID-19 pandemic. However, there are still some people who are reluctant to participate in vaccination. This vaccine hesitation is a complex phenomenon driven by individuals, such as whether there are negative emotions and lack of confidence in national institutions. Therefore, it is important to ensure that it is not related to other factors such as loneliness, fear, and social stress. Therefore, the government should continue to improve people's confidence in the government. A series of solutions can achieve these improvements, including improving government transparency and community cooperation. Additionally, we can continue to investigate people's current attitude towards booster vaccination.

6. References

Biddle, N., Edwards, B., Gray, M., & Sollis, K. ANU Poll 35 (August 2020): COVID-19 attitudes and behaviours (Wave 3). *ADA Dataverse, V2* (2020). https://doi:10.26193/ZFGFNE

Danchin, M., Biezen, R., Manski-Nankervis, J., Kaufman, J., & Leask, J. (2020). Preparing the public for COVID 19 vaccines. Retrieved 8 May 2022, from https://www1.racgp.org.au/ajgp/2020/october/preparing-the-public-for-covid-19-vaccines

Seale, H., Heywood, A.E., Leask, J. et al. Examining Australian public perceptions and behaviors towards a future COVID-19 vaccine. *BMC Infect Dis* 21, 120 (2021). https://doi.org/10.1186/s12879-021-05833-1

7. Appendix

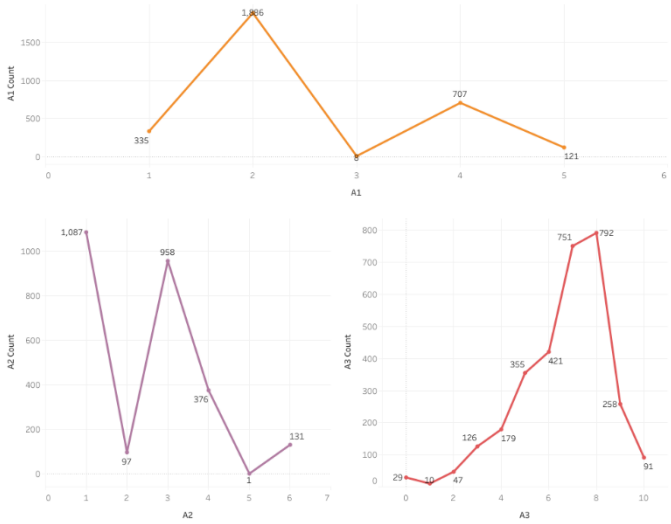


Figure 1. Section A statistics

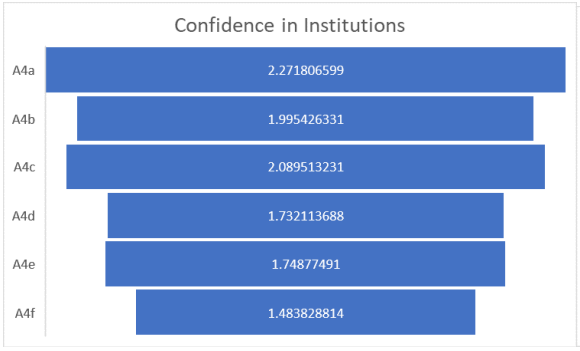


Figure 2. Section A statistics

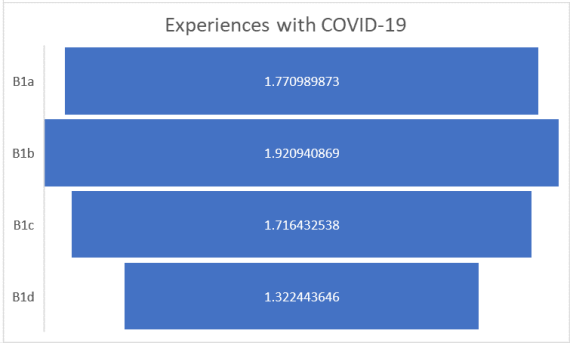


Figure 3. Section B statistics

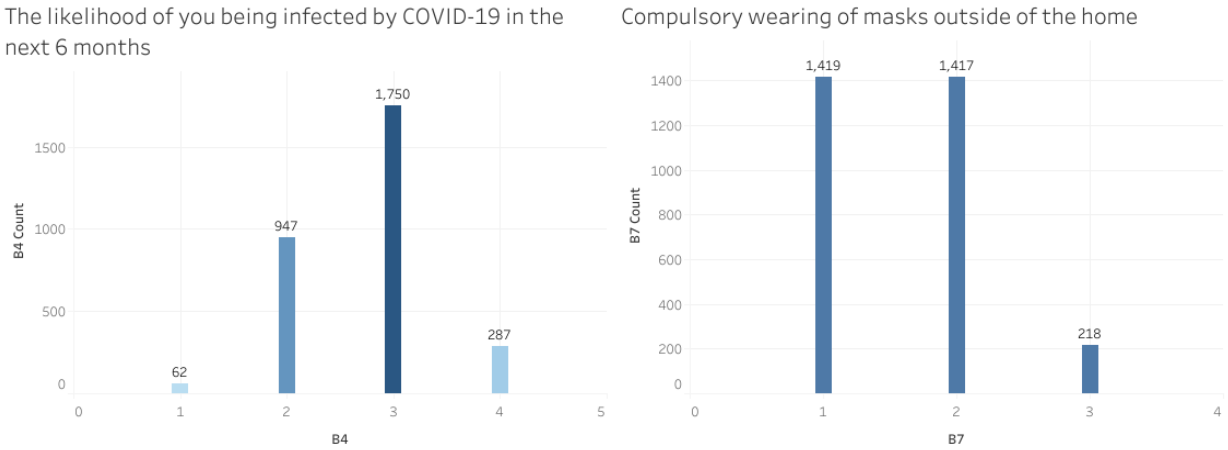


Figure 4. Section B statistics

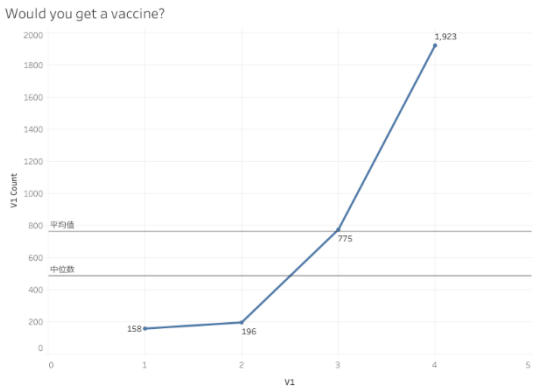


Figure 5. Section V statistics

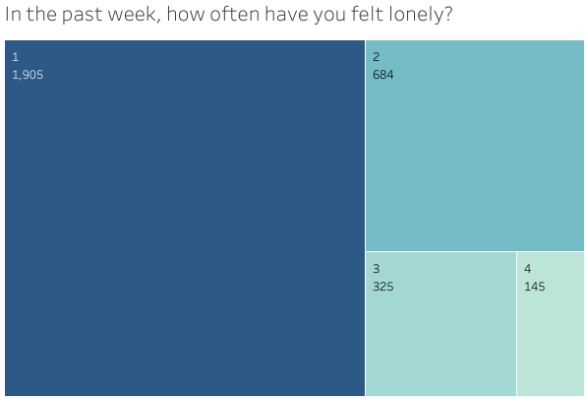


Figure 6. Section D statistics