

# FIT3152 Data analytics – 2023: Assignment 1

<b>Your task</b>	<ul style="list-style-type: none"> <li>Analyse the country level predictors of pro-social behaviours to reduce the spread of COVID-19 during the early stages of the pandemic.</li> <li>This is an individual assignment.</li> </ul>
<b>Value</b>	<ul style="list-style-type: none"> <li>This assignment is worth <b>25%</b> of your total marks for the unit.</li> <li>It has 40 marks in total.</li> </ul>
<b>Suggested Length</b>	<ul style="list-style-type: none"> <li>8 – 10 A4 pages (for your report) + extra pages as appendix (for your R script and clustering table).</li> <li>Font size 11 or 12pt, single spacing</li> </ul>
<b>Due Date</b>	<b>11.55pm Monday 17<sup>th</sup> April 2023</b>
<b>Submission</b>	<ul style="list-style-type: none"> <li>Submit a single PDF file and single video file on Moodle.</li> <li>Use the naming convention: <i>FirstnameSecondnameID.{pdf, mp4, mov etc.}</i></li> <li>Turnitin will be used for similarity checking of all written submissions.</li> </ul>
<b>Generative AI Use</b>	<ul style="list-style-type: none"> <li>In this assessment, you must not use generative artificial intelligence (AI) to generate any materials or content in relation to the assessment task.</li> </ul>
<b>Late Penalties</b>	<ul style="list-style-type: none"> <li>10% (4 mark) deduction per calendar day for up to one week.</li> <li>Submissions more than 7 calendar days after the due date will receive a mark of zero (0) and no assessment feedback will be provided.</li> </ul>

## Instructions

Address each of the research questions below and report the results of your analysis and your interpretation of those results.

You are expected to include at least one high quality multivariate graphic summarising key results. You may also include other simpler graphs and tables. Report any assumptions you've made in modelling and include your R code as an appendix. Your R code must be machine readable text as the university requires all student submissions to be processed by plagiarism detection software.

There are two options for compiling your written report:

- (1) You can create your report using any word processor with your R code pasted in as machine-readable text as an appendix, and save as a pdf, or
- (2) As an R Markup document that contains the R code with the discussion/text interleaved. Render this as an HTML file and save as a pdf.

Your video report should be less than 100MB in size. You may need to reduce the resolution of your original recording to achieve this. Use a standard file format such as .mp4, or mov for submission.

## Software

It is expected that you will use R for your data analysis and graphics and tables. You are free to use any R packages you need but must document these in your report and include in your R code. You may use other software, such as Excel, to create the table of clustering data for Question 3(a).

## Questions

During the early stages of the COVID-19 pandemic, researchers surveyed participants around the globe. A baseline study was conducted with the aim of identifying the most important predictors of pro-social COVID-19 behaviours, that is, actions that would reduce the spread of the virus. You can read a more detailed description of the research and results in Van Lissa (2022), see references.

The aim of this assignment is to understand country-level differences in predictors of pro-social behaviours, reported by participants as: "I am willing to:

- help others who suffer from coronavirus." (c19ProSo01)
- make donations to help others that suffer from coronavirus." (c19ProSo02)
- protect vulnerable groups from coronavirus even at my own expense." (c19ProSo03)
- make personal sacrifices to prevent the spread of coronavirus." (c19ProSo04)

Your task is to analyse the baseline survey data overall, with a focus on the country you have been assigned. You may make use of any additional data you require to answer the following questions.

### 1. Descriptive analysis and pre-processing. (6 Marks)

(a) Describe the data overall, including things such as dimension, data types, distribution of numerical attributes, variety of non-numerical (text) attributes, missing values, and anything else of interest or relevance.

(b) Comment on any pre-processing or data manipulation required for the following analysis.

### 2. Focus country vs all other countries as a group. (12 Marks)

(a) Identify your focus country from the accompanying list (FocusCountryByID.pdf). How do participant responses for your focus country differ from the other countries in the survey as a group?

(b) How well do participant responses (attributes) predict pro-social attitudes (c19ProSo01, 2, 3 and 4) for your focus country? Which attributes seem to be the best predictors? Explain your reasoning.

(c) Repeat Question 2(b) for the other countries as a group. Which attributes are the strongest predictors? How do these attributes compare to those of your focus country?

### 3. Focus country vs cluster of similar countries. (10 Marks)

(a) Using several social, economic, health, political or other indicators, identify between 3 and 7 countries (in the baseline data) that are similar to your focus country using clustering. Van Lissa (2022) refers to several indicators you might consider, among others. Some of these are listed in the references, but these are not exhaustive. State the indicators used and describe how you calculated/identified similar countries. Copy and paste the table of values you used for your clustering into your report as an Appendix.

(b) How well do participant responses predict pro-social attitudes (c19ProSo01, 2, 3 and 4) for this cluster of similar countries? Which attributes are the strongest predictors? How do these attributes compare to those of your focus country? Comment on the similarity and/or difference between your results for this question and Question 2(c). That is, does the

group of all other countries 2(c), or the cluster of similar countries 3(b) give a better match to the important attributes for predicting pro-social attitudes in your focus country? Discuss.

#### 4. Video Presentation: (Submission Hurdle and 4 Marks)

Record a short presentation using your smart phone, Zoom, or similar method. Your presentation should be approximately 5 minutes in length and summarise your main findings for Sections 1 – 3, as well as describing how you conducted your research and any assumptions made. Pay particular emphasis to your results in Questions 2(c) and 3(b)

#### 5 Overall considerations (8 Marks)

This includes: the quality and clarity of your reasoning and assumptions; the strength of support for your findings; the quality of your writing in general and communication of results; the quality of your graphics throughout, including at least one high-quality multivariate graphic; the quality of your R coding.

## Data

The data for this assignment is a reduced version of that collected for the PsyCorona baseline study, Van Lissa et al. (2022). The filename is “PsyCoronaBaselineExtract.csv”. The data includes ordinal data coded on a numerical scale. For this assignment assume it is reasonable to treat these responses as numerical.

Create your individual data as follows:

```
rm(list = ls())
set.seed(12345678) # XXXXXXXX = your student ID
cvbase = read.csv("PsyCoronaBaselineExtract.csv")
cvbase <- cvbase[sample(nrow(cvbase), 40000), ] # 40000 rows
```

Locate your focus country using the accompanying document FocusCountryByID.pdf.

## References and web links

C. J. Van Lissa, et al., (2022) Using machine learning to identify important predictors of COVID-19 infection prevention behaviors during the early phase of the pandemic. Patterns 3, 100482.  
<https://doi.org/10.1016/j.patter.2022.100482>

The World Bank Data Collections (and Governance Indicators)  
<https://datacatalog.worldbank.org/collections>  
<http://info.worldbank.org/governance/wgi/>

Organisation for Economic Co-operation and Development (OECD)Data  
<https://data.oecd.org/>

Global Health Security Index: Reports and Data  
<https://www.ghsindex.org/report-model/>

World Health Organization  
<https://www.who.int/>

Data fields and brief descriptor (note AD = Agree/Disagree). See BaselineCodebookExtract for full description.)

Concept	Variable Name	Label
Affect	affAnx	How did you feel over the last week? - Anxious
	affCalm	...Calm
	affContent	...Content
	affBor	...Bored
	affEnerg	...Energetic
	affDepr	...Depressed
	affExc	...Excited
	affNerv	...Nervous
	affExh	...Exhausted
	affInsp	...Inspired
	affRel	...Relaxed
Likelihood	PLRAC19	How likely is it that... in the next few months? - You will get infected with coronavirus.
	PLRAEco	... Your personal situation will get worse due to economic consequences of coronavirus.
Societal Discontent	disc01	AD - I fear that things will go wrong in society.
	disc02	AD - I feel concerned when I think about the future of society.
	disc03	AD - I am satisfied with society.
Job Insecurity	jblInsec01	AD - Chances are, I will soon lose my job.
	jblInsec02	AD - I am sure I can keep my job.
	jblInsec03	AD - I feel insecure about the future of my job.
	jblInsec04	AD - I already lost my job.
Employment Status	employstatus_1	Which best describes your employment status during the last week (multiple may apply)?-Employed, working 1-24 hours per week
	employstatus_2	... Employed, working 24-39 hours per week
	employstatus_3	... Employed, working 40 or more hours per week
	employstatus_4	... Not employed, looking for work
	employstatus_5	... Not employed, not looking for work
	employstatus_6	... Homemaker
	employstatus_7	... Retired
	employstatus_8	... Disabled, not able to work
	employstatus_9	... Student
	employstatus_10	... Volunteering
Perceived Financial Strain	PFS01	AD - I am financially strained.
	PFS02	AD - I often think about my current financial situation.
	PFS03	AD - Due to my financial situation, I have difficulties paying for my expenses.
Disempowerment	fail01	AD - Not a lot is done for people like me in this country.
	fail02	AD - If I compare people like me against other people in this country, my group is worse off.

	fail03	AD - Recent events in society have increased my struggles in daily life.
Life Satisfaction	happy	In general, how happy would you say you are?
	lifeSat	In general, how satisfied are you with your life?
	MLQ	AD - "My life has a clear sense of purpose."
Corona Community Injunctive norms	c19NormShould	AD - "Right now, people in my area..."-...should self-isolate and engage in social distancing.
	c19NormDo	AD - "Right now, people in my area..."-...do self-isolate and engage in social distancing.
	c19IsStrict	To what extent is your community....-...developing strict rules in response to the Coronavirus?
	c19IsPunish	...punishing people who deviate from the rules that have been put in place in response to the Coronavirus?
	c19IsOrg	...well organized in responding to the Coronavirus?
Trust in Government	trustGovCtry	In general, how much do you trust the government of your country to take the right measures to deal with the coronavirus pandemic?
	trustGovState	In general, how much do you trust your community to take the right measures to deal with the coronavirus pandemic?
Gender	gender	What is your gender?
Age	age	What is your age?
Education	edu	What is your highest level of education?
Country Self Report	coded_country	In which country do you currently live in?
Corona ProSocial Behavior	c19ProSo01	AD - "I am willing to..."-...help others who suffer from coronavirus.
	c19ProSo02	AD - "I am willing to..."-...make donations to help others that suffer from coronavirus.
	c19ProSo03	AD - "I am willing to..."-...protect vulnerable groups from coronavirus even at my own expense.
	c19ProSo04	AD - "I am willing to..."-...make personal sacrifices to prevent the spread of coronavirus.