

Data cleaning and screening Protocol

Due to this survey being an online survey, which will be downloaded electronically into SPSS, it is unlikely there will be errors within the dataset as there is less possibility of human error. However, we still need to screen and clean the data.

1. Checking categorical variables

- In the SPSS menu bar, select the following in the following order
Analyse – Descriptive Statistics – Frequencies
- Select the categorical variables
Age-gender-education-ethnicity-past vac-had covid-source of info
- Select the following
Statistic - Minimum and Maximum – Continue- OK
- Check the categorical variables are correct.
Age between XX – Gender between 1-5 – Education between 1-7 – Ethnicity between 1-21 – Past Vac 1-3 – Had Covid 1-2 – Source of info – 1-7.
Missing data will show up in the tables created by the above action.

- **Additional Check for missing data**

Analysis- Descriptive Statistics- Frequency- Select all variables – OK

This will provide you with a table indicating all variables and each variable will have a total below indicating how many participants took part, along with how many did not answer that question.

There may be a few missing data fields within the data set as there was no forced answer questions.

2. Checking continues variables

- In the SPSS menu bar, select the following in the order shown.
Analysis - Descriptive Statistics – Descriptive
- Select the continues variables
- Select the following
Options – Minimum – Maximum – Standard Deviation – Continue – OK
- Check the minimum and maximum values are correct and fall within the correct scores.
ST variables between 1-5 – VHS between 1-7 – TBP between 1-7.

3. Outliers and Errors

- In the SPSS menu bar, select the following in the order shown.
Analyse – Descriptive Statistics – Explore – Both – Select Variable (move to dependent list) – select ID (move to Label Cases) – Statistics – Outliers – Continue – Plots – Histogram – Stem and Leaf – Options – Exclude cases pairwise – Continue – OK

Check the output box plot for outliers, check if the outlier is an error, compare trimmed mean and original mean for differences, if there is a notable difference, we will have to revisit the data scores using the extreme values table.