THE CURIOSITY CUP 2022
A Global SAS® Student Competition

Two Possible Approaches for the DASS Dataset

– AristoSAS Team –

ABSTRACT

The paper aims to implement two possible approaches to analyse the DASS Dataset, containing answers to a questionnaire designed to study depression, anxiety, and stress along with demographical information. The first approach consists of a multiple correspondence analysis that allows reducing the dimensionality by producing scores on two new dimensions that are interpreted according to the available information and are used to cluster observations; the second approach relies on multiple logistic regression that is applied for each of the three above-mentioned emotional states, whose scores are derived according to the already validated and available rules of the questionnaire. The results of the two approaches are then compared.

INTRODUCTION

This analysis, performed with SAS® Studio on SAS® OnDemand for Academics, is based on the DASS Dataset retrieved from (kaggle.com, 2019); it collects the answers of an online survey whose goal is to measure depression, anxiety, and stress (DAS). The authors aim to reduce the dimensionality of such a dataset and to find some groups of observations with similar patterns and possibly with specific demographical features. In chapter 1, a preliminary analysis is implemented and the dataset is cleaned; in chapter 2, a multiple correspondence analysis (mca) is performed producing scores used to cluster observations; in chapter 3, three multiple logistic regression models allow to understand the effect of demographical and personality domain variables on a response, represented by each of the DAS emotional states. Finally, in the conclusion, the authors realise some comparisons among the obtained results.

1. PRELIMINARY ANALYSIS

1.1. DATA STRUCTURE

The downloaded dataset consists of 39775 observations and the following 172 variables[1]: 42 items related to as many statements concerning DAS (to which the respondent has to provide a self-degree of applicability[2] referring to the previous week) and for each of them the relative order in the survey and answering time (for a total of 126 variables); technical information regarding the provenience of the questionnaire (ISO country code), where the user found the test (*source*) and 3 variables for the compilation time; 10 variables (*TIPI_*) used to measure[3] the user's level of agreement relying on the "Ten Item Personality Inventory" test (Gosling, Rentfrow, & Swann, 2003); a checklist of words in the survey, i.e. 16 dummy variables (*VCL_*) codified in such a way that 1 is reported if the subject knows the word definition and 0 otherwise; 13 demographical variables (education, gender, living area, native language, age, religion, sexual orientation, writing hand, race, vote in a national election, marital status, number of siblings, major degree) and 2 technical details (the device type and a dummy –*uniquenetworklocation*– indicating if the survey comes from an already processed network).

[1] The authors will write variable names in *italic*.

[2] On a 1 to 4 scale, where 1 represents "Did not apply to me at all" and 4 "Apply to me very much, or most of the time".

[3] On a 1 to 7 scale, where 1 represents "Disagree Strongly" and 7 "Agree Strongly"; this scale also allows a neutral position "Neither agree nor disagree" represented by value 4.

1.2. DATA CLEANING

1.2.1. Searching for Missing Values and Validating Observations

For some variables, the value 0 is found even if not included in the codebook as a possible answer. Therefore, the authors, in a context like this, assume it as missing information.

Handling of Missing Values

None of the DAS 126 variables shows missing information. On the contrary, variables *TIPI_* manifest some missing values: the writers decide that the observations (447) who presented less than 50% of them should be removed. Instead, for the remaining units, the value 0 is replaced with the neutral answer (4). Concerning the demographical variables, since most of them are categorical, the writers do not impute or remove missing values but include them into an additional and reserved level not to lose any further information. Finally, specifically for the question "Including you, how many children did your mother have?" that could be misunderstood, the authors recodify this numerical variable replacing 0 with value 1, assuming that at least the respondent should contribute to the family size.

Validating Observations according to the Word Check List

As introduced before, each variable *VCL_* refers to a word; considering that three of them do not exist, the writers decide that units who do not check any word (554) and those who check all the words or at least 2 out of the 3 non-existing words (1142) in the list should be deleted from the dataset because it is expected that either the respondent's level of attention is not particularly high or that they do not know sufficiently well the English language to succussing in understanding the other questions in the survey.

Validating Observations according to the Answering Time

Since the dataset contains 3 variables that measure the time spent in each part of the questionnaire (the introduction, the DAS questions, and the demographical ones), the authors decide to use the total time to complete the survey– information obtained by summing for each subject the 3 variables – to select only the observations who spend a reasonable amount of time. In particular, using the quantiles answers with a total compilation time smaller than 344 seconds (25% quantile, around 4-5 seconds for question) and greater than 2062 seconds (95% quantile) are removed (11305).

Validating Observations according to Demographical Specifics

Information concerning *education* and *age* is crossed to check if any inconsistent pattern is present. Therefore, units younger than 15 with *education* equal or superior to "High school" (179) and units younger than 18 with *education* equal or superior to "University degree" (12) are deleted.

Possible Presence of Duplicates

uniquenetworklocation, described above, is equal to 2 when several questionnaires are submitted from the same network location. However, no duplicated records are present in the dataset for those with value 2 of this variable.

1.2.2. Managing variables

The authors implement some manipulation to the dataset in order to reach a structure that is as adequate as possible to perform the analysis.

Depression, Anxiety, Stress Scores and Personality Domain Variables

Concerning the first 126 variables, only the degree of applicability to the 42 statements is considered (*Q1A-Q42A)*. Relative to them and referring to (neurocogsystem.com, p. 1), three groups of questions can be defined regarding depression, anxiety, and stress.

Since in this document, the answers are reported in a 0 to 3 instead of a 1 to 4 scale, the writers decide to recodify them uniformly to the above-mentioned source. Scores for each group of questions are thus computed and categorized as shown in (neurocogsystem.com, p. 3), producing the *dep_cat*, *anx_cat* and *str_cat* variables.
Also, starting from the 10 variables *TIPI_*, 5 new numerical variables are derived representing the Big Five Personality Domains which are reported in (Gosling, Rentfrow, & Swann, 2003, pp. 504-528): Extraversion (*extr*), Agreeableness (*agree*), Conscientiousness (*consc*), Emotional Stability (*emot*) and Openness to Experience (*open*).

Categorizing Variables

- Technical information in *country* is categorized referring to the continents; correspondence among countries and ISO code is taken from (Wikipedia, 2021).
- *fam_cat* is created starting from *familysize* using the following categorization: 0="Only child", 1="One-Two Siblings", 2="Three-Five Siblings", 3="More than 6 Siblings"; a special category 4="Unlikely" is also considered (54 and 133 are included here).

Finally, for most of the remaining demographical variables, a FORMAT procedure is implemented to facilitate comprehension and visualization without losing any information.

Detecting and eventually Correcting Atypical Values

- For the variable *age* (in years) wrong values such as 1991,1993,1996 that in practice express the year of birth are transformed into the corresponding age at the time of the survey, which is assumed to be 2018[4] (e.g. 2018-1991=27). Also, values of *age* greater than 100 (115 and 223) are recodified using the median (21).
- Very large values for the variable *familysize* lie in the dataset; as said before, such values are not imputed or removed but put into the dedicated class "Unlikely".

Creation of ID Variable

Once the authors end up cleaning the dataset (final number of observations: 26136), the variable *ID* is finally added to identify each record of the validated dataset.

2. MULTIPLE CORRESPONDENCE ANALYSIS AND CLUSTERING

2.1. MULTIPLE CORRESPONDENCE ANALYSIS

The 42 items *Q1A-Q42A* are involved in a dimension-reduction technique to implement the following analysis. To achieve this aim, mca is performed, by exploiting demographical-categorical variables as supplementary information.

Relying on the Greenacre Adjusted Inertia Decomposition, the authors decide to consider 2 dimensions since they are able to explain 94.56% of the inertia. The CORRESP procedure outputs the left plot in Figure 1 from which, through the position of the categories of the items, an interpretation of the dimensions is possible. The first one represents inclination to depression/anxiety/stress such that from left to right it increases; the second can be considered as the strength of the respondent's opinion: in the top, it is possible to detect the 0 and 3 profiles – that represent the more extreme answers – while in the bottom there are the 1 and 2 profiles (intermediate answers).

The following code[5] allows to project the observations into the 2-dimensional space obtained according to the previously exposed axes; moreover, also the levels of the supplementary items can be represented in such a space to support and complement the interpretation.

[4] The data are collected between 2017 and 2019: since no further information is available 2018 is used.

[5] For lack of space format statement, exploited for better visualization, is reported in the appendix section "Supporting Code".

```
PROC CORRESP DATA=DASS.VALIDATED SHORT NOPRINT BINARY OUT=DASS.SCORES_Q;
    TABLE Q1A--Q42A EDUCATION GENDER FAM_CAT URBAN ENGNAT HAND RELIGION
          VOTED ORIENTATION RACE MARRIED CONT;
    SUPPLEMENTARY EDUCATION GENDER FAM_CAT URBAN ENGNAT HAND RELIGION VOTED
                  ORIENTATION RACE MARRIED CONT;
RUN;
```

Using *ID*, the writers merge the unit scores and the other information collected and derived from the survey.

2.2. CLUSTERING PROCEDURES

The authors perform several clustering procedures relying only on the projections on the 2 dimensions found with mca. Supplementary variables along with *age*, *extr*, *agree*, *consc*, *emot* and *open* are then used to profile the clusters.

The considered clustering analyses are hierarchical (according to average and complete linkage and Ward's method) and k-means (which instead produces a partition).

CLUSTER procedure allows the implementation of the above-mentioned hierarchical methods and provides the information for the choice of the best number of clusters[6]. According to such a choice, the TREE procedure cuts the dendrogram and offers the corresponding clustering. The average linkage method returns 6 clusters and performs the best result in terms of interpretability, separation, and homogeneity of the groups.

Besides, the writers rely on the available information present in (SAS® Help Center, s.d.) to perform k-means analysis. HPCLUS procedure automatically provides the partition related to the best number of clusters (4) according to the aligned box criterion (ABC).

Figure 2 shows the representation of the two mentioned clustering results; as you can see, average linkage allows an interpretation on both the dimensions while k-means defines groups basically relying just on the first. For this reason and for lack of space, the authors choose to interpret only the hierarchical result.

The mca dimensions allow a first explanation of the clusters. The dark-green cluster[7] represents subjects that seem uncertain about their condition (both dimensions are around 0). The light-green cluster contains units with high strength of opinion (high values for the second dimension) but not in a precise direction (values around 0 for the first dimension); mca axes indeed rely on 42 items but a person may rate diversely questions about DAS (e.g. value 3 for questions about depression, but value 0 for questions about anxiety)[8]. The blue and brown clusters share a low value for the inclination to the three conditions and differ for the second dimension (higher for the blue and lower for the brown). You can see a similar structure for the purple and red clusters; they are both characterised by higher values for the first axis but units in the red cluster show a stronger opinion than observations in the purple one. Afterwards, clusters may also be interpreted by exploiting demographical information. FREQ procedure allows displaying the joint distribution of the cluster membership and the levels of each categorical item, while PROC MEANS shows summary statistics of numerical variables by cluster. Table 1 collects the main information derived from this analysis.

3. MULTIPLE LOGISTIC REGRESSION

3.1. IMPLEMENTING THREE MODELS

The authors also implement three distinct multiple logistic regression models with dependent variable *dep_cat*, *anx_cat* and *str_cat* respectively. For each of them, all the demographical variables (apart from major degree) and personality domain variables (*extr*, *agree*, *consc*, *emot* and *open*) are used as regressors: for each of the categorical items, a reference level is defined.

[6] The authors rely on the cubic clustering criterion, the pseudo-F, and the pseudo-t^2 statistics.

[7] Henceforth, clusters are named using the colors in Figure 2.

[8] This is only a possible explanation for the behavior of the observations who belong to this cluster.

To simplify the interpretation of the results, the proportional odds assumption is considered (common slopes for the different levels of the dependent variable), which is the default option of the MODEL statement in the LOGISTIC procedure; an additional option of this statement allows to specify a selection criterion for the explanatory variables (BACKWARD is chosen for these models). Since the writers aim for an interpretation in terms of risk of a higher level of the response w.r.t. a lower one, the option DESCENDING is applied.

The code to fit the model with *dep_cat* as response follows:

```
PROC LOGISTIC   DATA=DASS.DASS42 ORDER=INTERNAL;
    CLASS EDUCATION(PARAM=REF REF='LESS THAN HIGH SCHOOL') GENDER
           (PARAM=REF REF='MALE') URBAN(PARAM=REF REF='URBAN') ENGNAT
           (PARAM=REF REF='NO') HAND (PARAM=REF REF='RIGHT') RELIGION
           (PARAM=REF REF='ATHEIST') ORIENTATION (PARAM=REF
           REF='HETEROSEXUAL') RACE (PARAM=REF REF='WHITE') VOTED (PARAM=REF
           REF='NO') MARRIED (PARAM=REF REF='NEVER MARRIED') FAM_CAT
           (PARAM=REF REF='ONLY CHILD');
    MODEL DEP_CAT (DESCENDING) = EDUCATION GENDER AGE FAM_CAT URBAN ENGNAT
           HAND RELIGION ORIENTATION RACE VOTED MARRIED EXTR AGREE CONSC
           EMOT OPEN/ SELECTION=BACKWARD;
RUN;
```

For lack of space, only one of the fitted models is displayed; in particular, the one obtained using the above code. The choice of such a model is related to interpretability reasons.

3.2. INTERPRETATION OF THE RESULTS

Table 2 shows only the numeric variables or the levels of the items that have a significant effect[9]. Looking at odds ratios, it is possible to state that: a graduate education leads to a lower risk of being depressed w.r.t. a less-than-high-school education; older people tend to have a slightly higher probability of showing depression; living in a suburban area seems a protective factor w.r.t. an urban area; Muslim and Buddhist religions have a decreasing effect on depression w.r.t. the Atheism; all the considered sexual orientations lead to a higher risk of the condition w.r.t. heterosexuality; also Arab shows this behavior w.r.t. the white race; currently married people tend to have a lower probability of being depressed than never-married subjects; finally, all the five personality domain variables are protective factors.

CONCLUSION

Looking at the results of both the analyses and at the position of the supplementary profiles[10], it is possible to draw some conclusions: a lower education seems related to a higher level of DAS (confirmed by the position of the less-than-high-school profile); even if the model in section 3.2. shows *age* as a risk factor, in the other two models (with *anx_cat* and *str_cat* as response) *age* is a protective factor, consistently with the description of the blue and red clusters in Table 1; living in suburban area may lead to a lower level of DAS; sexual orientations other than heterosexuality may be risk factors as confirmed by both the analyses and the position of the bisexual profile; Arab profile shows a higher value for both mca axes and this is confirmed by logistic regression; results concerning the effect of being currently married agree: logistic regression suggests that it is a protective factor; purple and red clusters contain a smaller proportion of such units, while blue cluster a higher percentage and finally the first-dimension score in the plot is smaller; the protective effect of four out of the five personalities is confirmed by the clustering; nevertheless, the *agree* score average is similar among all the groups.

To conclude, the authors suggest for future studies to work also on the variable *major*, that may be an interesting source for the explanation of the DAS scores but due to its open-ended nature requires an important effort in the codification and validation.

[9] Henceforth, such effect has to be considered *ceteris paribus*.

[10] In the second plot of Figure 1.

APPENDIX

SUPPORTING TABLES

Cluster	**Major characteristics** (cluster vs population percentage - if present)
Dark-green Cluster (*Neutral Uncertain*)	• Lower % of males (19.14 vs 21.27) • **Cluster with largest % of Asian subjects** (66.36)
Light-green Cluster (*Neutral Decided*)	• Larger % of response "Other" in the sexual orientation (11.67 vs 8.86), of subjects with education lower than high school (13.33 vs 8.23), and subjects that write with the left hand (14.67 vs 10.14) • Lower % of observations who do not have English as natural language (56.00 vs 70.44) • Half of the observations have 1/2 siblings (50.33 vs 43.19) • **Cluster with lowest % of females** (65.33)
Blue Cluster *(NO-DAS Decided)*	• Larger % of currently married (20.04 vs 11.46), of heterosexual (70.39 vs 62.63), of male (28.16 vs 21.27), and of people who voted in the last national election (36.79 vs 28.02) • Lower % of never married (76.09 vs 85.42), of bisexual (6.24 vs 10.67), of female (71.38 vs 77.59), and of white subjects (18.64 vs 21.27) • 64.55% have at least a university degree • **Oldest cluster** (mean=26.35 and median=23); however 75% of subjects are less than 29 years old • **Largest average values** for *extr* (4.22 vs 3.53), *consc* (5 vs 4.26), *open* (5.13 vs 4.60), *emot* (5.05 vs 3.30)
Brown Cluster *(NO-DAS Uncertain)*	• Almost **1/3 of black people are in this cluster** • **Largest number of observations** (7344 out of 26136)
Purple Cluster *(DAS Uncertain)*	• Larger % of observations with an education lower than high school (51.7 vs 44.61), of bisexual (13.55 vs 10.67) and of units who did not vote in the last election (76.62 vs 71.22) • Lower % of currently married (7.77 vs 11.46), of heterosexual (57.84 vs 62.63)
Red Cluster *(DAS Decided)*	• Larger % of never married (91.42 vs 85.42), of bisexual and other (14.81 vs 10.67 and 12.16 vs 8.86 respectively), of female (83.40 vs 77.59), of subjects with lower education (53.94 with at most a high school education vs 44.61) • Lower % of currently married (5.97 vs 11.46), of heterosexual (52.29 vs 62.63), of subjects living in a suburban area (26.93 vs 32.33), of people who did not vote in the last election (79.17 vs 71.22) • **Youngest cluster** (mean=21.83, median=20) • **Lowest average values** for *extr* (2.98 vs 3.53), *consc* (3.67 vs 4.26), *open* (4.03 vs 4.60), *emot* (1.95 vs 3.30)

Table 1. Description of the Average Linkage Clusters w.r.t. Supplementary Variables

Odds Ratio (OR) Estimates	
Effect	**Point Estimates (Wald C.I.)**
Graduate degree w.r.t. Less than high school (*education*)	**0.870** (0.773 – 0.978)
Age	**1.006** (1.002 – 1.011)
Suburban w.r.t. Urban (*urban*)	**0.895** (0.849 – 0.944)

<table>
<tr><td colspan="2"><b>Muslim w.r.t. Atheist</b> (<i>religion</i>)</td><td colspan="2"><b>0.829</b> (0.738 – 0.932)</td></tr>
<tr><td colspan="2"><b>Asexual w.r.t. Heterosexual</b> (<i>orientation</i>)</td><td colspan="2"><b>1.178</b> (1.060 – 1.309)</td></tr>
<tr><td colspan="2"><b>Bisexual w.r.t. Heterosexual</b> (<i>orientation</i>)</td><td colspan="2"><b>1.293</b> (1.195 – 1.398)</td></tr>
<tr><td colspan="2"><b>Homosexual w.r.t. Heterosexual</b> (<i>orientation</i>)</td><td colspan="2"><b>1.233</b> (1.101 – 1.380)</td></tr>
<tr><td colspan="2"><b>NA orient w.r.t. Heterosexual</b> (<i>orientation</i>)</td><td colspan="2"><b>1.104</b> (1.014 – 1.202)</td></tr>
<tr><td colspan="2"><b>Other w.r.t. Heterosexual</b> (<i>orientation</i>)</td><td colspan="2"><b>1.172</b> (1.079 1.273)</td></tr>
<tr><td colspan="2"><b>Arab w.r.t. White</b> (<i>race</i>)</td><td colspan="2"><b>1.685</b> (1.288 – 2.204)</td></tr>
<tr><td colspan="2"><b>Currently married w.r.t. Never married</b> (<i>married</i>)</td><td colspan="2"><b>0.637</b> (0.583 – 0.696)</td></tr>
<tr><td colspan="2"><b>Extr</b></td><td colspan="2"><b>0.767</b> (0.755 – 0.780)</td></tr>
<tr><td><b>Agree</b></td><td><b>0.922</b> (0.904 – 0.941)</td><td><b>Emot</b></td><td><b>0.535</b> (0.525 – 0.545)</td></tr>
<tr><td><b>Consc</b></td><td><b>0.823</b> (0.809 – 0.837)</td><td><b>Open</b></td><td><b>0.898</b> (0.881 – 0.915)</td></tr>
</table>

Table 2. LOGISTIC Procedure: Significative OR Estimates with *dep_cat* as response

SUPPORTING FIGURES

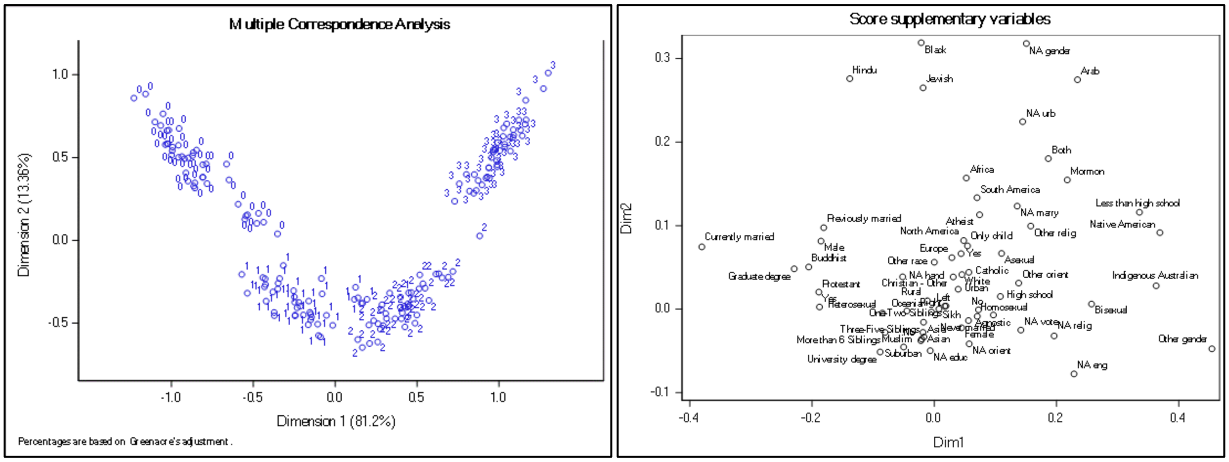

Figure 1. CORRESP Procedure: Profiles projected in the first 2 dimensions

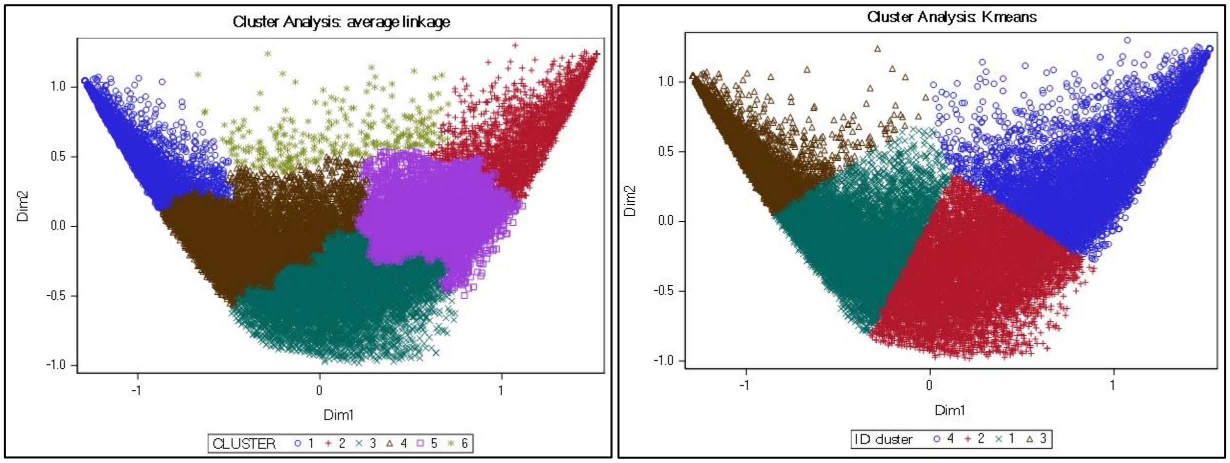

Figure 2. SGPLOT Procedure: Average linkage (K=6) **and k-means** (K=4) **clusterings**

SUPPORTING CODE

The format statement of the CORRESP procedure follows:

```
FORMAT      EDUCATION F_EDUCATION. GENDER F_GENDER. FAM_CAT F_FAMCAT. URBAN F_URBAN.
            ENGNAT F_ENGNAT. HAND F_HAND. RELIGION F_RELIGION. VOTED F_VOTED.
            ORIENTATION F_ORIENTATION. RACE F_RACE. MARRIED F_MARRIED. CONT F_COUNTRY.;
```

REFERENCES

Article in Website neurocogsystem.com (n.d.). Accessed January 29, 2022. Available at https://neurocogsystem.com/wp-content/uploads/2021/02/DASS-42-Scoring.pdf

Website kaggle.com (2019). Accessed January 29, 2022. Available at https://www.kaggle.com/yamqwe/depression-anxiety-stress-scales/download

Article in Website Gosling, S. D., Rentfrow, P. J., & Swann, W. J. (2003). "A Very Brief Measure of the Big Five Personality Domains". *Journal of Research in Personality,* Volume no. 37, Page 504-528. Accessed January 29, 2022. Available at http://gosling.psy.utexas.edu/scales-weve-developed/ten-item-personality-measure-tipi/

Website SAS® Help Center. (n.d.). Accessed January 29, 2022. Available at https://documentation.sas.com/doc/en/emhpprcref/14.2/emhpprcref_hpclus_syntax01.htm

Website Wikipedia. (2021, December 29). Accessed January 29, 2022. Available at https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes