Group Coursework Submission Form

Specialist Masters Programme

Please list all names of group members: 1. Attagrish, Phu 2. Curro, Simona 3. Gooda, Anusha 4. Perera, Joshua	**GROUP NUMBER:** **7**
MSc in: Business Analytics	
Module Code: SMM634	
Module Title: Analytics Methods for Business	
Lecturer: Dr Rosalba Radice	**Submission Date: 13/02/2021**

Declaration:
By submitting this work, we declare that this work is entirely our own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work, we acknowledge that we have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. We also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked on-line):

Deduction for Late Submission: | **Final Mark:** **%**

Group Coursework 1

ANALYTICS METHODS FOR BUSINESS SMM634

FEBRUARY 13, 2021

CASS BUSINESS SCHOOL
GROUP 7

Table of Contents

Question 1

a) Dataset Analysis

Since variances differ vastly among variables, with some clearly dominating others due to the design, the data should be scaled before implementing the PCA. As, if failing to do so, could lead to a domination of *Var1, Var2,* and *Var3* as they are characterized by the highest variances and far from those of other variables. Consequently, the interpretability and validity of the result could be hindered. However, it can be argued that one can choose to not standardise the variables, if they have the same unit of measurement (e.g., kilogram).

As a consequence of the lack of this contextual information, the variables have been standardised, so that the PCA Biplot, which shows the first and second principal components of the variables and the location of each observation, can be easily interpretable (Appendix A). However, if variables would have kept unstandardised, the Biplot would be more difficult to interpret (Appendix B), as it is dominated by variables with relatively high variances**.** Predictably, the Biplot (Appendix A) shows a strong correlation among *Var4, Var5* and *Var6,* and among *Var7, Var8* and *Var9.* Particularly, the former group of variables is mostly and dominantly explained by the first principal component (PC1) and the latter by the second principal component (PC2). The former variables' arrows of PC1 also homogeneously point to the left, as with the latter variables' arrows point to the top. The direction is not important, but their homogeneity is. As the correlations were originally set to be positive among variables within each group, a directional similarity of variables in each group is expected.

Additionally, the PCA loadings for the variables within each group are approximately equalled and hence comparable. However, the two groups of variables are found to be marginally correlated as each group is mostly explained by different principal components. *Var1, Var2* and *Var3* are internally less correlated compared to the other two groups and the directions of the arrows are not as homogenous as the previous two groups, spanning to both PC1 and PC2, with *Var1* and *Var2* mostly explained by PC2 and *Var3* by PC1. Further, it seems that for both PC1 and PC2, *Var1, Var2,* and *Var3* are not as important as their counterparts and, by design, it should not be as outer correlations are approximately zeros. However, since PC3 explains most of the *Var1, Var2,* and *Var3* (Appendix C), the Biplot should not be used to reliably evaluate these three variables (*Var1, Var2,* and *Var3*) as it probably presents an incomplete picture.

Lastly, it can be shown that the calculated cumulative variance explained (Appendix D) shows that it is probably optimal to include just PC1, PC2, and PC3 and ignore the rest, as we can clearly identify an elbow of the plot (at PC = 3). Including all the first three principal components indicates that we capture approximately 84.7% of the original variation. Hence, low-dimensional representation of the data has been achieved, while maintaining as much information as possible. In sum, meaningful insights can be derived from this analysis, by looking at where each observation is located, if the variables are contextually identified.

c) Dataset Analysis

Similar to reasons mentioned in the previous analysis, we therefore perform the PCA on the standardised variables. Inherently designed to have all zero intercorrelations, a Biplot (Appendix E) predictably shows variables spreading across all directions, with few having an approximately same direction (e.g., *Var3* and *Var4*). Indeed, albeit designed to have zero correlations, in practice, the correlations are not exactly equalled to zeros, as illustrate by a modest correlation (0.25; see R-script lines 28-30) between *Var3* and *Var4* which could help explain why they are seemingly correlated. In contrast to the previous analysis, the cumulative variance explained (Appendix F) seems to have no clear elbow, and it takes from PC1 to PC7 to explain the 85.85% of the original variance, which is approximately equivalent to the level achieved by PC1 to PC3 in the previous analysis. In addition, the contribution of variance explained by each consecutive PC decreases only slightly (Appendix G), which led to the same conclusions resulted in the cumulative one as well. This stems from the fact that the variables are approximately uncorrelated, and hence there is limited justification to use PCA in the first place.

Question 2

Method

In the analysis, classical MDS has been employed as the given data are numerically hard facts, not subjective datasets, which may require ordinal MDS instead. Additionally, standardisation has been applied to the dataset, minimising the effect of significant differences of the GDP variable, which could adversely affect the result and its interpretability. Then, given the confidence that the problem incurred from large differences has been reduced through standardisation, the Euclidian distance has been used to calculate the required distance matrix in order to confidently use the PCA to guide the results obtained from the classical MDS application. Afterwards, the MDS has been performed. Next, PCA has been performed on the scaled dataset to aid the interpretation of visualisations generated by the conducted classical MDS. Lastly, in the analysis, we abbreviate the variables in accordance with the ones given. Mainly, Increase as Annual Percentage Population Growth rate and GDP as GDP per capita in US dollars.

Analysis

The one-dimensional classical MDS graph (Appendix H) differentiates developed countries' clusters on the left-hand side and less developed countries placed on the right-hand side. The cluster on the extreme left is made up of countries with a high level of GDP per capita and life expectancy and low rates of infant mortality and fertility, which are characteristics of highly developed countries. Albeit, a PC1 loading for GDP is slightly less in absolute value than the other variables, overall PC1 loadings are approximately equaled in absolute terms (Appendix I). Hence, it can be argued that the first dimension represents an overall development of a country. Hence, the right-hand cluster is categorized as less developed countries. Indeed, Malawi, located on the far right, shows one of the highest IMR, TFR, and Increase, and one of the lowest Life and GDP values. Overall, the less developed countries show a higher population growth rate, IMR, and TFR, but lower Life and GDP values.

The two-dimensional configuration (Appendix L) illustrates an additional insight. Dimension 1 can be interpreted as overall countries' development as previously stated. However, Dimension 2 mainly represents GDP as shown by its dominant value of PC2 loading over the rest (Appendix I). Mainly, Romania and Croatia, which are located on the top left, show a lower level of GDP, compared to the cluster on the bottom left consisting of European countries, Australia and USA, but share those

other characteristics, such as low values for IMR, TFR, and Increase, and high values Life, that highly developed countries have. Another example is China and Albania, which have roughly the same values of all variables except GDP, with the former locating vertically and directly above the latter, and thus predictably has lower GDP.

Consequently, it can be said that dimension 2 is mainly a function of GDP, but cautions have to be applied as if a country locating above another does not necessarily mean lower GDP. In fact, Romania has a higher value of the second dimension but has a higher GDP compared to Malawi. Dimension 1 also has to be taken into account, as exemplified by the aforementioned examples.

In terms of dimensional adequacy, it can be shown that the goodness of fit value (Appendix M) for the one dimension is approximately 0.8, which equals an acceptance level of 0.8 (ETH Zurich, 2013). An increase to two-dimension has yielded a GOF of 0.92, which is fairly better. However, it is essential to notice a decreasing marginal return of GOF for the additional dimension (i.e., decreased slope). Hence, the optimal number of dimensions is arguably one or two, depending on whether the objective is to achieve interpretability or possible additional insights. As two-dimensional results do not considerably inhibit the result's interpretability from the one-dimensional option, we advocate the two-dimensional approach as the optimal number of dimensions for this particular dataset.

Question 3

Method

In the new technological era, the evolution of the social communication era caused researchers to investigate the increasing prevalence of online dating phenomena and the influence of technology among emerging adults who belong to the 18-29 age group. Social communication technology allows individuals to lower anxiety related to dating practices and strengthen relationships, promoting communication-based on nonverbal nature. Recent studies explored online dating services' characteristics and motivation, but few studies empirically validated instruments to measure online dating services use. The following research aims to create an empirically supported instrument to assess online dating services' intensity based on an existing tool designed to measure the emerging adults' intensity Facebook. The Facebook Intensity Scale (FBI) has been modified to create the Online Dating Intensity Scale (ODI). The FBI is a one-factor consisting of nine items designed that showed strong psychometric properties to measure frequency, engagement level, and duration indices of an individual's Facebook usage. Similarly, the ODI would include items related to attitudes solutions about online dating and specific activities measurement in terms of quantity, frequency, and duration related to intensity solutions. The ODI resulted in a 10-item instrument based on a 5-points Likert-type scale, and the total scores have been calculated as participants' mean scores. The ODI represented a measurement tool to assess individuals' engagement level, emotional connection, and intensity in online dating activities.

Data Source and Analysis

Emerging adults represent the population under analysis to assess the relationship between users' intensity of online dating activity and the related scores of social desirability and the relationship between online dating services use and the related demographic variables. Data collection occurred through a web-based survey and face-to-face administration forms. The survey has been designed to be accessible by avoiding technical language and decreasing the perceived cost of participation by minimizing private information. A general demographic questionnaire has been included to collect data related to demographic variables such as race, ethnicity, gender, and age. Participants were also required to specify from the 16 listed services whether they used online dating services or telephone applications. Notably, 241 individuals have been recruited through face-to-face selection, and 253 individuals have been selected through online recruitment for data collection. The sample contained students' data enrolled in multiple universities (n = 8).

The final sample contained 494 observations: 27.3% of individuals who currently use online dating services, 48.4% used online dating services more than one year ago, and 23.1% used the platforms more than a year ago. Thus, rigorous data collection procedures have been applied to guarantee heterogeneity and geographic representation in the sample. A listwise deletion has been used to remove missing cases, but outliers have been included in the analysis to ensure consistency within the dataset. Furthermore, the Q-Q plots, histograms, and boxplots showed that data were not normally distributed. Principal Axis Factoring (PAF) has been applied to the dataset, subject to Oblimin rotation with Kaiser normalization, to evaluate the shared variance among the set of X variables (items) to get information about the smaller set of latent variables called factors. In fact, the first step was the evaluation of the commonalities, defined as sums of squared factor loadings. Oblique rotation served to make the output more interpretable. The Kaiser normalization allows to apply the rotation on normalized factor loadings to maximize the variance of the squared loadings of the factor in the factor matrix. The data screening process demonstrated that the data collected were not normally distributed but adequate for the analysis. The Kaiser-Mayer-Olkin, which measures the sampling adequacy, indicated that the underlying factor caused 81.9% of the variables' variance. Furthermore, Bartlett's test of sphericity tests the hypothesis that the correlation matrix is an identity matrix. However, Bartlett's test-related p-value was less than 0.001 allowing to conclude the variables included in the dataset are unrelated. The first analysis led to factor loadings greater than 0.32, problematic cross-loading on four variables across two factors, and low communality values. As a result, items 7, 4, 10, 9 with low communalities and similar online dating characteristics have been removed.

Finally, Explanatory Factor Analysis has been applied to the final data set. The factor structure analysed resulted in a five-item instrument with a one-factor solution, which demonstrated adequate internal consistency reliability (0.83). Notably, a one-factor structure has been chosen as the final model because the related eigenvalue was the highest (3.04), demonstrating the model's robustness. Also, all the related five items showed high factor loadings ranging from 0.65 to 0.85. The one-factor structure analysis, accounting for 60.79% of the total variance, was a parsimonious model consistent with the analysis's theoretical goal. The model succeeded in capturing the individual's intensity of online dating services use. The research showed no correlation among the nonparametric variables as the related p-value was greater than the threshold set (0.001),

demonstrating no statistically significant relationship between participants' ODI scores and the demographic variables. Furthermore, the analysis aimed to measure the social desirability in the participants' responses. The bivariate correlations between the modified ODI and the Marlowe–Crowne Social Desirability Scale–Short Form A, showed that participants' responses were not influenced by social desirability.

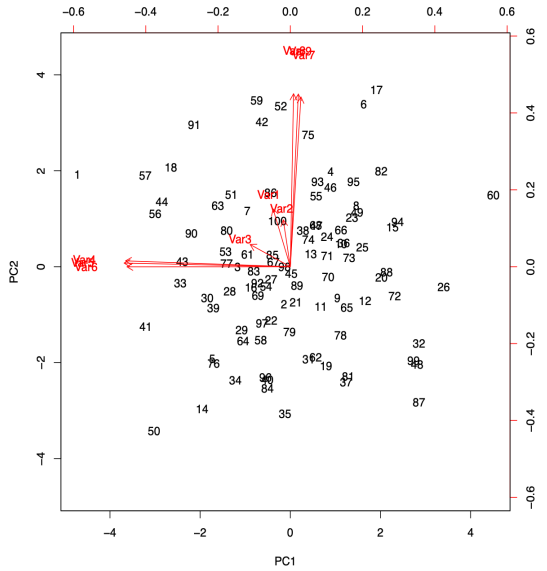
In conclusion, the ODI showed strong psychometric properties. The preliminary analysis using PAF and Oblimin rotation with Kaiser normalization supported ODI's potential to produce reliable scores about the intensity of an individual's online dating services use. The ODI can be used in clinical practices as an instrument to measure the clients' intensity of technology use. Mainly, online dating services present adverse effects concerning the potentially addictive nature involving technologies. Clinicians could use the ODI to measure clients' potential reductions in their online dating activities or patterns of obsessive behaviours during different therapy periods.

Limitations and Recommendations

The analysis presented some limitations. The majority of individuals were heterosexual and belonged to White and Non-Hispanic background, which negatively impacted the analysed sample's homogeneity. Also, 68% of the sample's emerging adults belonged to the 18-21 age group, which caused the age group's unequal distribution. The results lacked generality as most participants were recruited from six schools located in the southern-eastern United States. Further researches should explore the factor structure and the psychometric properties of the ODI on more diverse populations to evaluate the ODI's potentialities as a valuable instrument to measure online dating services use. Furthermore, the self-report nature of the data instrument, which could be subject to bias or influenced by a basement effect due to wide intervals' responses to the ODI items, could cause results to be inconsistent or show a lack of level discrimination for low-intensity use of online dating services. Further researches should use alternative intervals for potential responses to the ODI items and collect data through qualitative investigation. In conclusion, the current generation of emerging adults can be identified as the first technological cohort to grow up using online technology regularly. Further analysis should involve an equal adult age representation and a more diverse population in terms of gender, race, ethnicity, sexual orientation and practiced behaviours in the sample to make reliable inferences and valuable predictions about the future generations.

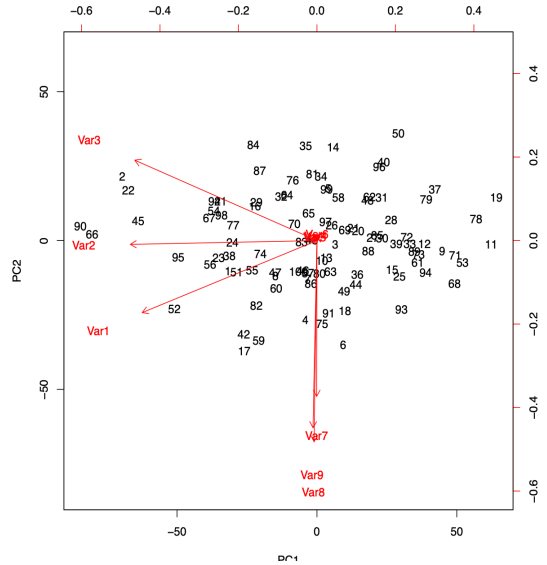
Appendix A

Graph 1: Biplot (Scaled Dataset a)



Appendix B

Graph 2: Biplot (Unscaled Dataset a)



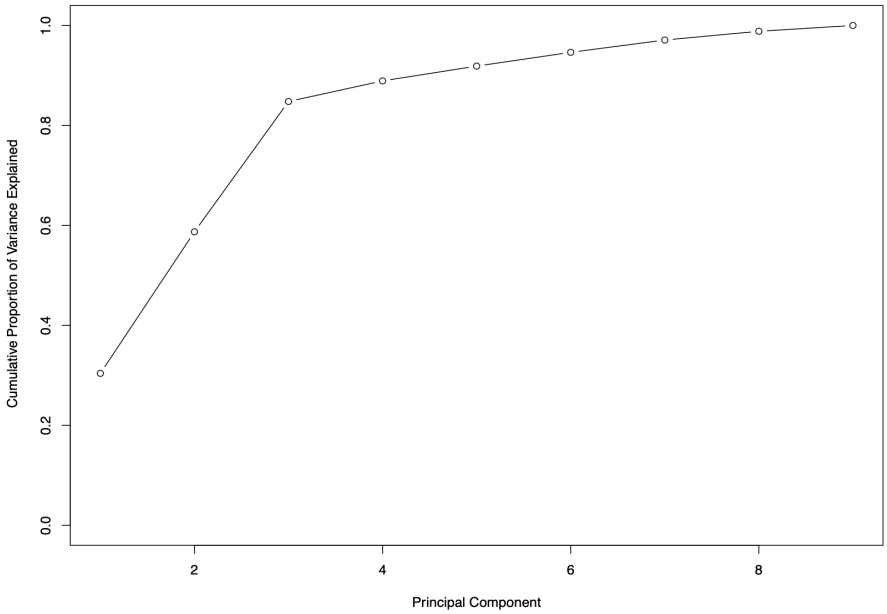
Appendix C

Table 1: PCA Loadings Analysis

	PC1	PC2	PC3
Var1	-0.05969388	0.1878542105	-0.52832125
Var2	-0.02430125	0.1523281042	-0.56592493
Var3	-0.13878033	0.0721083208	-0.57221958
Var4	-0.56983362	0.0189910590	0.08003140
Var5	-0.57480804	0.0118144989	0.06665655
Var6	-0.56498270	-0.0002723545	0.09041398
Var7	0.03686505	0.5519471109	0.14528552
Var8	0.01133637	0.5622574749	0.13410870
Var9	0.02790109	0.5612734950	0.12265081

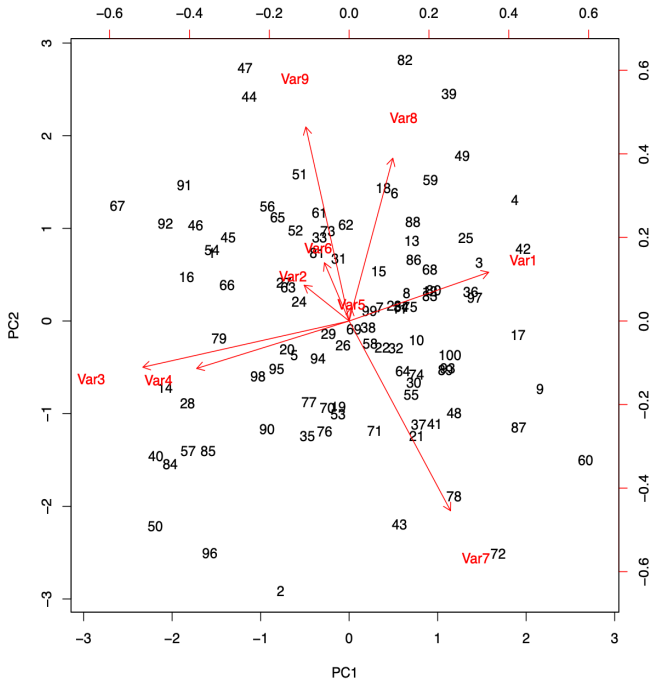
Appendix D

Graph 3: Cumulative Variance Explained (Scaled Dataset a)



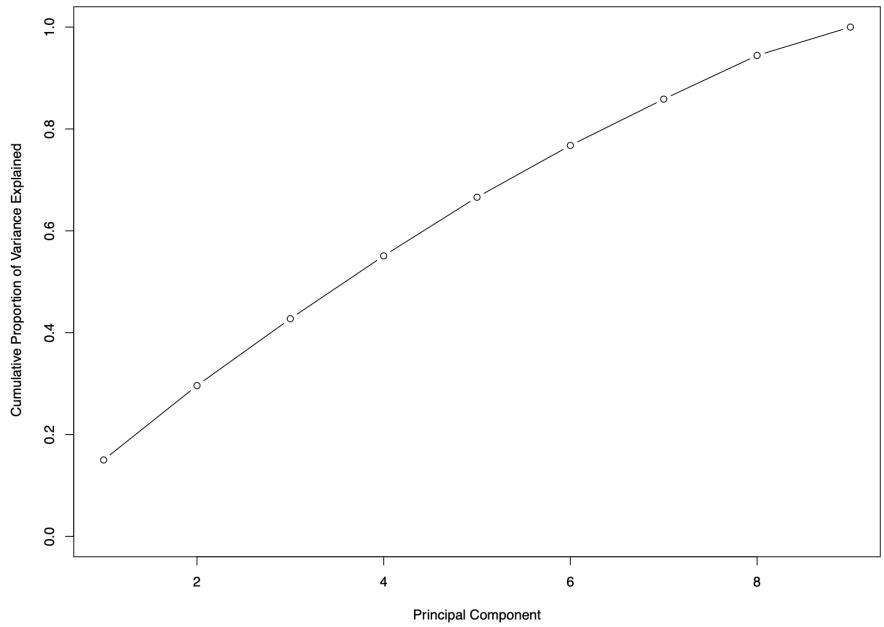
Appendix E

Graph 4: Biplot (Scaled Dataset c)



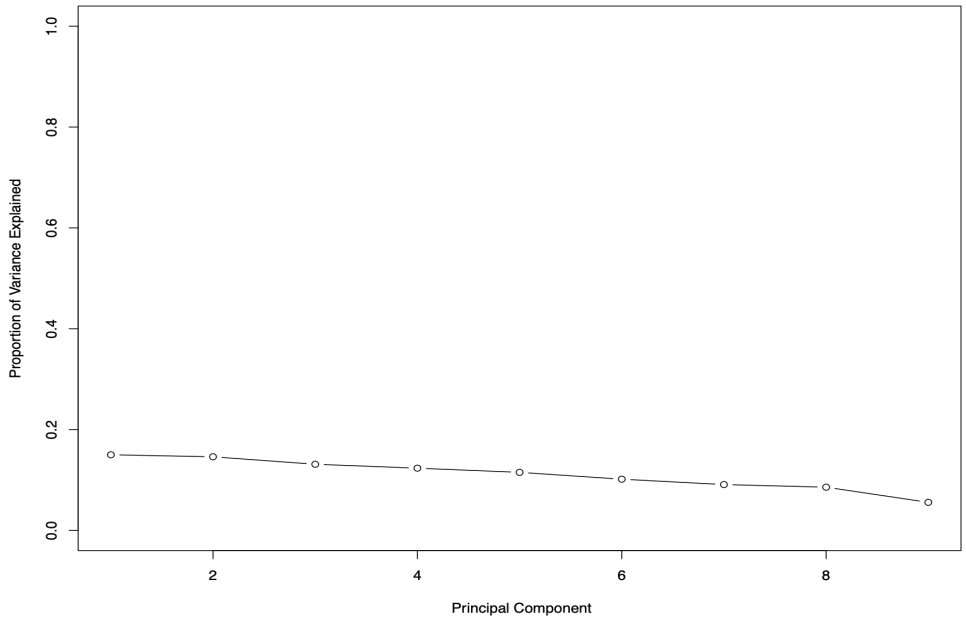
Appendix F

Graph 5: Cumulative Variance Explained (Scaled Dataset c)



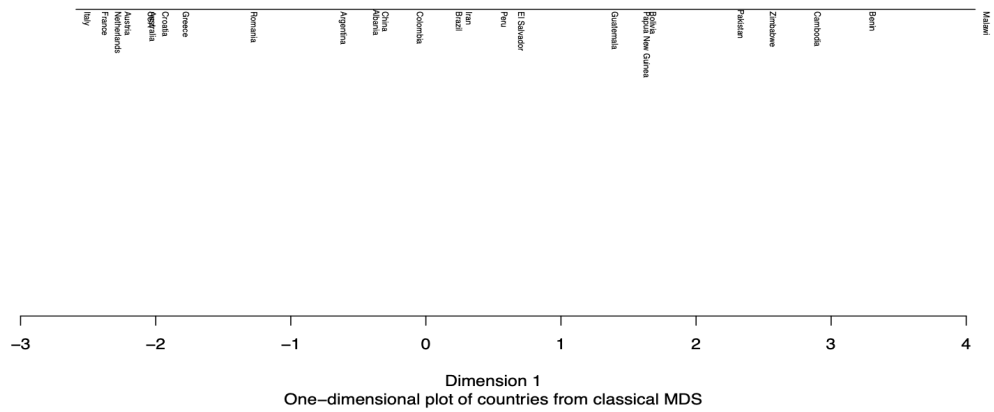
Appendix G

Graph 6: Individual Variance Explained (Scaled Dataset c)



Appendix H

Graph 7: One-Dimensional Plot of Countries from Classical MDS



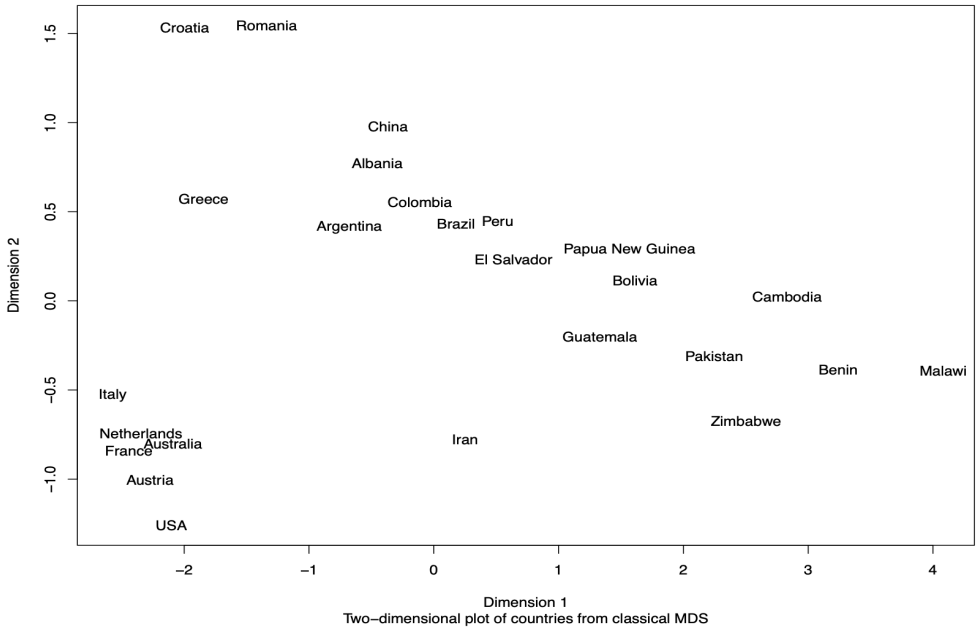
Appendix I

Table 2: PCA Loadings Analysis

	PC1	PC2
Increase	0.4264716	0.52382004
Life	-0.4748670	0.05079789
IMR	0.4751735	-0.01529870
TFR	0.4747086	0.24145120
GDP	-0.3761451	0.81516833

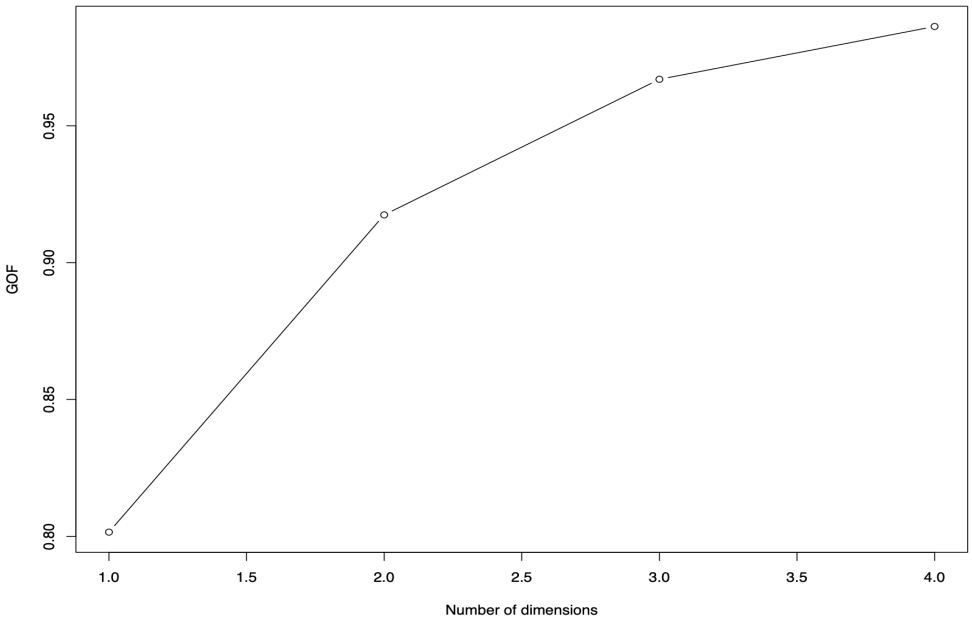
Appendix L

Graph 8: Two-Dimensional Plot of Countries from Classical MDS



Appendix M

Graph 9: GOF vs Number of Dimensions Plot



Appendix N

R-Script Question 1

```
# install important library
install.packages('rospca')

# load library
library(rospca)

############# QUESTION 1 #############
# a.)
## generate the instructed data
df_all = dataGen(m = 1, n = 100, p = 9 , a = c(0.7, 0.9, 0.8), bLength = 3, SD = c(20,5,10))
df_all
df_1 = df_all$data[[1]] # get the generated data

# b.)
corr_df = cor(df_1) # check correlation among the variables of the generated data
sd_df = apply(df_1, 2 , sd)  # check standard deviation values of the generated data
corr_df # close but not exactly equal to the stated
sd_df # close but not exactly equal to the stated

# c.)
## generate the same data, except the intercorrelations are zeros
df2_all = dataGen(m = 1, n = 100, p = 9 , a = c(0, 0, 0), bLength = 3, SD = c(20,5,10))
df2_all
df_2 = df2_all$data[[1]]



## check correlation and sd
corr_df2 = cor(df_2) # check correlation among the variables of the generated data
sd_df2 = apply(df_2, 2 , sd)  # check standard deviation among the variables of the generated data
```

```
corr_df2 # see some variables correlate, albeit not significantly
sd_df2 # close but not exactly equal to the stated

# d.)
## perform PCA on the first generated data (a.) ##
### convert data into data frame format, to make the analysis easier
df_1 = data.frame(df_1)
### put the column names to ease the readers in the document
colnames(df_1) = c('Var1','Var2','Var3','Var4','Var5','Var6','Var7','Var8','Var9')
df_1
### because variances differ vastly among variables, we should scale the data before doing PCA
pr_1.out = prcomp(df_1, scale = TRUE) # we can use this directly as our variables are quantitative
pr_1.out
### plot the first two principal components
biplot(pr_1.out, scale = 0) # we see a high clustering of variables 7,8 and 9, and of variables 4,5,
and 6
                # which is expected as each variable within these two groups are highly correlated
                # (0.8, 0.9, respectively). Also, since the outer correlations are relatively close
                # to zero, we clearly see that the group with variables 7,8 and 9 explains a lot for
PC2, while little for PC1
                # whereas the group 4,5 and 6 exhibits a vice versa phenomenon.

### test when we did not standardize the variables
pr_1_uc.out = prcomp(df_1)
biplot(pr_1_uc.out, scale = 0) # Var1,2,3,7,8, and 9 clearly dominate Var4,5, and 6 due to variance
domination

### calculate variance explained by each principal component for the standardized case
pr_1.out$sdev
pr_1.var = pr_1.out$sdev ^2
pr_1.var
pve_1 = pr_1.var/sum(pr_1.var)
```

```
pve_1 # individual variance explained of each principal component as of the total (9), as each
variable is standardized and
     # hence the sum of total variance = 9 (1+1+1+ ...+ 1)
### plot variance explained by each principal component
plot(pve_1, xlab = " Principal Component", ylab = "Proportion of Variance Explained",
    ylim = c(0,1), type = "b")
### plot cumulative variance explained
plot(cumsum(pve_1), xlab = "Principal Component",
    ylab =" Cumulative Proportion of Variance Explained", ylim = c(0,1), type = "b")
### it seems that in this case, we should focus on only three principal components
### as we can clearly identify an elbow of the plot

## perform PCA on the second generated data (c.) ##
### convert data into data frame format, to make the analysis easier
df_2 = data.frame(df_2)
### put the column names to ease the readers in the document
colnames(df_2) = c('Var1','Var2','Var3','Var4','Var5','Var6','Var7','Var8','Var9')
df_2
### examine the variance of the data
apply(df_2, 2, var)
### hence we should scale the data due to huge differences of variances before doing PCA
pr_2.out = prcomp(df_2, scale = TRUE) # we can use this directly as our variables are quantitative
pr_2.out
### plot the first two principal components
biplot(pr_2.out, scale = 0) # this is so contrast with pr_1, as the red arrows, which are the first
                    # two component loading vectors, spread out in all directions
                    # This is expected, as all the variables are marginally correlated
### test when we did not standardize the variables
pr_2_uc.out = prcomp(df_2)
biplot(pr_2_uc.out, scale = 0) # clearly difficult to interpret
### calculate variance explained by each principal component for standardized case
pr_2.out$sdev
```

```
pr_2.var = pr_2.out$sdev ^2
pr_2.var
pve_2 = pr_2.var/sum(pr_2.var)
pve_2
### plot variance explained by each principal component
plot(pve_2, xlab = " Principal Component", ylab = "Proportion of Variance Explained",
    ylim = c(0,1), type = "b") # relatively, all principal components have similar variance explained
### plot cumulative variance explained
plot(cumsum(pve_2), xlab = "Principal Component",
    ylab =" Cumulative Proportion of Variance Explained", ylim = c(0,1), type = "b") # no clear elbow
# it seems that in this case, we are faced with difficulties, as the more principal component
# we included, the more complexities in terms of interpretation
# also, it is more difficult to visualize when we have more dimensions
# To illustrate, we have to include up to 4 principals, to have the cumulative variance explained
# well-above 50%, and this hinders interpretability and makes visualization almost impossible to
# interpret by eyes

### In sum, we know that df_2 is just marginally linearly correlated by construction, and hence the
justification
### of doing PCA analysis is limited, if not non-existence.
```

\

Appendix O

R-Script Question 2

```
############ QUESTION 2 ############

# load data
df3 = read.table('UNSY97.txt', header = TRUE)
df3
str(df3)
summary(df3)

# get the name of each rows
nation_names = ("Albania, Argentina, Australia, Austria, Benin, Bolivia, Brazil, Cambodia, China,
Colombia, Croatia, El Salvador, France, Greece, Guatemala, Iran, Italy, Malawi,
Netherlands,Pakistan, Papua New Guinea, Peru, Romania, USA, Zimbabwe")
nation_names = strsplit(nation_names, ',') # split the names into single entity
nation_names = nation_names[[1]] # extract the required element in the list
rownames(df3) = nation_names # name the dataframe columns
df3
# check variance
apply(df3, 2 , var) # we see variables such as GDP dominating others in terms of variance
# check mean
apply(df3, 2 , mean) # we see variables such as GDP dominating others in terms of mean
# hence, we standardize the variables first to minimise the distance effect of GDP when using
Euclidean distances
df3.sc = scale(df3)
# calculate Euclidean distances
df3.sc.dist = dist(df3.sc)
#analysis
## 1 dimension
df3_1.sc.mds = cmdscale(df3.sc.dist, k = 1)

### 1-dimensional visualisation
```

```
plot(c(df3_1.sc.mds),rep(1,25), type = 'l', ylab = '',xlab = 'Dimension 1', axes = F, xlim = c(-3,4),
    sub = "One-dimensional plot of countries from classical MDS")
axis(side =1)
text(c(df3_1.sc.mds),rep(1,25), c(nation_names), cex=0.5, srt = 270, pos = 4, xpd = T)
# it appears that European countries and USA are located in the same cluster
# whereas the South American countries are also located in the same cluster

## 2 dimension
df3_2.sc.mds = cmdscale(df3.sc.dist)

### 2-dimensional visualisation
plot(df3_2.sc.mds, type = "n", xlab = "Dimension 1", ylab = "Dimension 2",
    sub = "Two-dimensional plot of countries from classical MDS")
text(df3_2.sc.mds, rownames(df3_2.sc.mds), cex=1)

# conduct PCA to aid the analysis
pr_3.out = prcomp(df3.sc) # we can use this directly as our variables are quantitative
pr_3.out # these loadings can be used to evaluate previous two graphs

# evaluation of dimensional competency
## conduct a GOF test
### create an empty vector to store values
gof = c()
k = c(1,2,3,4)
### create a for loop to extract GOF values
for (i in k){
  to_store = cmdscale(df3.sc.dist, k = i, eig = TRUE)$GOF[1]
  gof = append(gof, to_store)

}
### plot GOF vs. dimensions
plot(k, gof, type = "b", xlab = "Number of dimensions", ylab = "GOF")
```

References

Stat.ethz.ch. 2013. *Multidimensional Scaling*. [online] Available at: <https://stat.ethz.ch/education/semesters/ss2013/ams/slides/v4.1.pdf> [Accessed 8 February 2021].