IS507 - Data, Statistical Models, and Information - Assignment 6

1) Choose a technique that we have covered so far in this course, and try applying that technique to your data. You may choose any of:
a) Model building and Multiple Regression
b) PCA
c) CFA
d) CCA
e) CA (correspondence analysis)
If you technique, are working as a group, each member of your group should try a different or the same technique with different aspects of the data.

Solution:

- I have chosen to perform the Principal Component Analysis on the Online News Popularity dataset.
- It has 61 columns of which 1st is the columns of URLs and the last column is the target variable for prediction. So, I will drop these 2 columns. Remaining all columns are numeric in nature.
- We now have 59 columns which can be used for performing PCA and then obtain the different components.
- The major aim of the PCA techniques is to reduce the dimensionality of the dataset and explaining the same amount of information with fewer variables.
- The dataset has not been cleaned and I have run just a basic PCA model.

```
> # Load library readr to read the dataset
> library(readr)
> library(plyr)
> library(dplyr)
> library(tidyr)
> library(stringr)
>
> # Set working directory
> setwd("~/Desktop/IS507 - Data, Statistical Models and Information/Assignments/Assignment_6")
>
> # 1) PCA on OnlineNewsPopularity Dataset
>
> # Import the dataset using read_csv()
> online_news_raw_dataset <- read.csv("OnlineNewsPopularity.csv", header=TRUE)
> online_news_dataset <- online_news_raw_dataset
>
> # Shape of dataset
> dim(online_news_dataset)
[1] 39644    61
>
> # Count of missing values
> sum(is.na(online_news_dataset))
[1] 0
>
> # Delete rows with missing values
> dataset <- na.omit(online_news_dataset)
>
> # Shape of new dataset after listwise deletion
> dim(online_news_dataset)
[1] 39644    61
>
```

```
29  # Dropping the 1st URL column and last target column
30  new_online_news_dataset <- online_news_dataset[, c(2:60)]
```

Before performing PCA, let us run a few tests to check the data fit.

- We run the KMO sampling adequacy test. For our case, the overall KMO value is 0.5 which indicates that the data is poor in terns of correlation within the data.
- We then check the Bartlett's test of sphericity. For our case, the p-value < 0.001 < 0.05. This tells us that the data has enough correlations to run PCA and the correlation matrix is not an identity matrix.
- The Cronbach's Alpha coefficient is 0.416. This is very less than the default value of 0.7. This tells us that the variables within the components are not consistent as well as the reliability is also very low.
- Nevertheless, we run the PCA test.

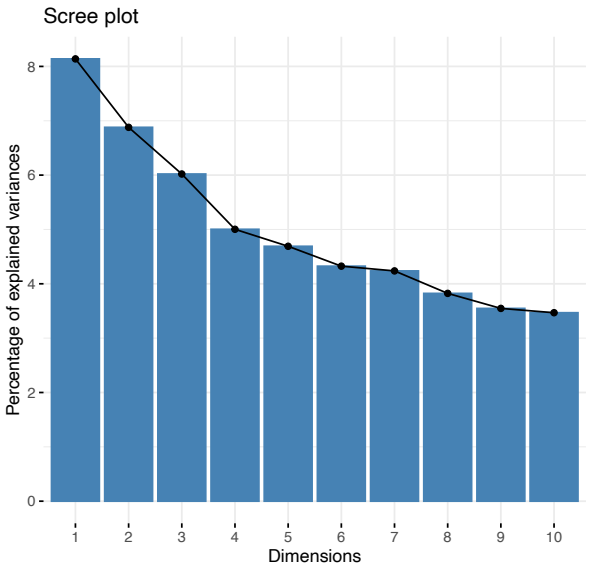
```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = new_online_news_dataset)
Overall MSA =  0.5
```

```
> bartlett.test(new_online_news_dataset)

        Bartlett test of homogeneity of variances

data:  new_online_news_dataset
Bartlett's K-squared = 42896066, df = 58, p-value < 2.2e-16
```

```
> CronbachAlpha(new_online_news_dataset)
[1] 0.4161898
```

Initial PCA:

- We first use the entire dataset to perform an initial PCA, just to understand and get an idea as to how many components can be used to explain majority of the information.
- We use the scree plot with the knee method to identify the components. We understand and can estimate that the knee bends at the 5th component and flattens.
- So, we end up choosing 5 components for the PCA model.



First Actual PCA:

- We use the principal() function with 'promax' rotation as the size of the dataset is high ~ 40000 rows. I also pass 5 factors to build the PCA model.
- When we check the cumulative variance, we see that the 5 component model is able to explain around approximately 31.3% of the variance in the data.
- We check the loadings and notice that there are certain variables which do not have any loadings. So, we remove them from the dataset and then check again.
- We can also observe the cross-loadings and so we increase the cut-off value to remove them.

```
> comps = print(pca_with_5_components_promax_1$loadings, sort=T)

Loadings:
                           RC1    RC2    RC3    RC5    RC4
average_token_length        0.709  0.136 -0.235 -0.115
global_subjectivity         0.753  0.257
avg_positive_polarity       0.708  0.296
max_positive_polarity       0.731  0.306
avg_negative_polarity      -0.666  0.272
min_negative_polarity      -0.697  0.395
global_sentiment_polarity   0.131  0.852
global_rate_positive_words  0.428  0.575
global_rate_negative_words  0.498 -0.558
rate_positive_words         0.343  0.778 -0.163
rate_negative_words         0.407 -0.750
kw_max_min                 -0.116         0.700 -0.305
kw_avg_min                 -0.107         0.688 -0.398
kw_max_avg                                0.797
kw_avg_avg                                0.814  0.259
timedelta                         0.167         -0.648
kw_min_min                                      -0.688
kw_max_max                                       0.716
kw_avg_max                                0.138  0.839
n_unique_tokens                                         1.000
n_non_stop_words                                        0.999
n_non_stop_unique_tokens                                1.000
n_tokens_title                                   0.200
n_tokens_content            0.449         -0.144
num_hrefs                   0.411
num_self_hrefs              0.271  0.107
num_imgs                    0.227         0.148
num_videos                  0.194         0.144  0.109
num_keywords                              0.127 -0.348
data_channel_is_lifestyle                        -0.164
data_channel_is_entertainment  0.166 -0.126  0.133
data_channel_is_bus                0.325 -0.157  0.234
data_channel_is_socmed             0.122
data_channel_is_tech               0.236         -0.274
data_channel_is_world             -0.462 -0.319
kw_min_max                                       0.389
kw_min_avg                                0.228  0.443
self_reference_min_shares                 0.253
self_reference_max_shares                 0.324
self_reference_avg_sharess                0.336
```

```
weekday_is_monday
weekday_is_tuesday
weekday_is_wednesday
weekday_is_thursday
weekday_is_friday
weekday_is_saturday
weekday_is_sunday
is_weekend
LDA_00                                  0.367 -0.140  0.206
LDA_01                          0.128
LDA_02                                -0.444 -0.347
LDA_03                                        0.471  0.212
LDA_04                                 0.242        -0.337
min_positive_polarity           0.164
max_negative_polarity          -0.220
title_subjectivity              0.116         0.203
title_sentiment_polarity               0.294
abs_title_subjectivity                       -0.106
abs_title_sentiment_polarity    0.119         0.202

                 RC1   RC2   RC3   RC5   RC4
SS loadings    4.545 4.063 3.462 3.415 3.002
Proportion Var 0.077 0.069 0.059 0.058 0.051
Cumulative Var 0.077 0.146 0.205 0.262 0.313
>
```

Process:

- We keep on performing this process 6 times till we get no cross-loadings and also there are no variables with no loadings.
- In the final dataset we have 21 variables and we get 5 components. The cumulative variance explained by these components is close to 76%. The RC1 component explains the maximum variance of 19% followed by RC2 (14.5%), RC5 (14.4%), RC3 (14.3%), and RC4 (13.7%).

```
> comps = print(pca_with_5_components_promax_6$loadings, cutoff=0.59, sort=T)

Loadings:
                          RC1    RC2    RC5    RC3    RC4
average_token_length       0.737
global_subjectivity        0.817
global_rate_positive_words 0.679
rate_positive_words        0.719
avg_positive_polarity      0.815
max_positive_polarity      0.812
global_sentiment_polarity         -0.711
global_rate_negative_words         0.860
rate_negative_words                0.947
min_negative_polarity             -0.639
timedelta                                0.780
kw_min_min                               0.884
kw_max_max                              -0.916
kw_avg_max                              -0.763
n_unique_tokens                                1.000
n_non_stop_words                               1.000
n_non_stop_unique_tokens                       1.000
kw_max_min                                            0.894
kw_avg_min                                            0.884
kw_max_avg                                            0.855
kw_avg_avg                                            0.731

                 RC1   RC2   RC5   RC3   RC4
SS loadings    4.026 3.036 3.023 3.000 2.867
Proportion Var 0.192 0.145 0.144 0.143 0.137
Cumulative Var 0.192 0.336 0.480 0.623 0.760
```

Naming the components:

Since, we have the 5 components, it will be very difficult to provide a meaningful name for the components as the variables are not related to each other.

The principal component RC1 —> Positive Influence of News : In this component, all the variables are describing the effect of the content of the news articles in terms of the positive effect on the masses due to the shares.

The principal component RC2 —> Negative Influence of News : In this component, all the variables are describing the effect of the content of the news articles in terms of the negative effect on the masses due to the shares. The global sentiment polarity and min_negative_polarity are negative indicating the inverse effect.

The principal component RC5 —> Shareability on basis of worst keywords : In this component, all the variables are describing the worst keywords that are present in news articles and their shares. The kw_max_max and kw_avg_max are describing the best keywords and hence are negative.

The principal component RC3 —> Frequency of unique and non-stop words : In this component, all the variables are describing the proportion of the distinct words and the non-stop/significant words inside a document.

The principal component RC4 —> Shareability on basis of worst and average keywords : In this component, all the variables are describing the average as well as the worst keywords that are present in news articles and their shares.

2) Paper Review: An academic paper from a conference or Journal will be posted to the Moodle. It contains a usage of Canonical Correlation Analysis. Review the paper and evaluate their usage of Canonical Correlation Analysis. In particular, address (Vacation Benefits and Activities Understanding Chinese Family Travellers)

- **How suitable is their data for CC?**

Solution:

The study has 2 major objectives:

a) First, was to understand the type of benefits pursued by Chinese family travellers because exploring benefits can lead to an understanding of the vacation experience.
b) Second, was to examine the relationship between the benefits obtained from taking part into the different activities for the Chinese family travellers.

The final dataset consists of 19 benefit items and 32 activities of participation. Once the Factor Analysis is performed, the data is checked for linearity, multicollinearity and the the CCA technique is performed.

- **How are they applying CC? What two groups of variables are being correlated? Are they metric, ordinal, nominal?**

Solution:

- Firstly, Exploratory Factor Analysis is performed on the dataset and it helps us to get 4 major Factors which can be useful in explaining the benefits sought by the Chinese family travellers.
- After this, the Canonical Correlation Analysis is performed on the benefit items of the 4 factors and the 32 activities of participation.
- They come up with 4 Canonical Variates (CVs) out of which only the 1st pair of the CV is of significance as it gives/explain the major relationship between the benefits and the activities.

The 2 groups of variables that are being correlated are:

1st - The benefits obtained by the Chinese family travellers after going on trips - There were 19 benefits and each was rated on a 5-point scale. So, this is a numeric variable.
2nd - The activities that the Chinese family travellers take part into - There were 32 activity items and each was rated on frequency of participation on a 5-point scale. So, this is a numeric variable.

- **What methods do they use to judge the quality of the correlation? Do they evaluate, and how do they evaluate the stability of the components?**

Solution:

Before performing Canonical Correlation Analysis, the study undertakes Exploratory Factor Analysis.
To check the fit of the data for Factor Analysis, they perform the KMO test for sampling adequacy and Bartlett's test of sphericity.

- The KMO value was 0.912 which indicates that the data has a good amount of correlation among the variables.
- The Bartlett's test metric was high which indicates that there are atleast some variables which contribute and the matrix of correlations is not an identity matrix. It means that there are enough correlations which can be broken up into different factors.

The Factor Analysis resulted in 4 factors for the benefits. Before performing CCA, the data is checked for linearity, sample size, and multicollinearity. Stability of the components is assessed using Cronbach's alpha test.

- Cronbach's alpha coefficient for all the 4 factors have values > 0.7 which indicate that the factors chosen are credible and demonstrate reliability and internal consistency.
- The 4 factors combined explains close to 62% of the total variance in the data.

Table 2
Dimensions of Family Vacation Benefit Sought: Factor Analysis

Items	Mean	Standard Deviation	Factor Loading	Cronbach's Alpha	Eigenvalue	Variance Explained (%)
Factor 1: Communication and Togetherness	3.78			0.883	8.167	21.919
Increasing communication with family members	3.83	1.104	.752			
Building a stronger family bond	3.83	1.111	.698			
Making memories together	3.67	1.148	.655			
Doing things together	3.65	1.148	.650			
Having fun with family members	4.07	1.013	.637			
Finding more things in common	3.60	1.184	.613			
Sharing quality time together	3.95	1.074	.609			
Respecting family members' decision	3.64	1.128	.578			
Factor 2: Shared Exploration	3.69			0.831	1.414	14.043
Experiencing a different culture	3.75	1.040	.780			
Tasting authentic local food	3.47	1.033	.701			
Sharing the same experiences with family members	3.80	1.068	.649			
Experiencing new things together	3.75	1.059	.606			
Factor 3: Escape and Relaxation	3.59			0.780	1.160	13.929
Getting away from the demands of home	3.32	1.242	.786			
Escaping from the daily routine	3.36	1.121	.784			
Getting a change from a busy job or life	3.75	1.041	.629			
Relaxing outside the home	3.93	1.053	.623			
Factor 4: Experiential Learning for Children	3.84			0.760	1.047	12.153
Broadening children's horizon	4.00	1.040	.824			
Extending children's knowledge	3.84	1.141	.708			
Children can learn about culture, history, and people	3.69	1.132	.675			
Total variance explained (%)						62.044

Note: The benefit items were measured based on a 5-point scale (1 = *never important*; 2 = *low importance*; 3 = *occasionally important*; 4 = *very important*; 5 = *always important*).

Four different analyses are performed between the 4-factor benefit items and the 32 activities.

- **How many correlates do they concentrate on in their analysis, and do they attempt to interpret the correlates in terms of the original variables?**

Solution:

- After performing the CCA, the researchers come up with 4 canonical variates.
- But after conducting the multivariate tests, they found out that only the *1st pair of CV (canonical variates for both X and Y) is of importance* as it explains the statistically significant proportion of the relationship between the benefits obtained and activity participation.

- *Yes, they attempt to interpret the variates in terms of original variables* and try to establish the relationship between the benefits perceived and the activities.
- The CV1 explains us the significant relationship between the activity of 'Taking pictures and videos' and the 4 items under the benefits under the factor of 'Communication and Togetherness': 'having fun with family members', 'respecting family members' decisions', 'finding more things in common' and 'sharing quality time together'. So, this means that these 4 benefits under the single factor of 'Communication and Togetherness'. This tells us that the benefit of communicating and being with one another is related to the activity when the family takes videos and photos.
- The CV2 explains us the relationship between the activities and the benefit of 'Shared Exploration'. There are many activities which contribute to shared experiences like 'tasting local authentic food', 'visiting historical site' and so on. This means that the benefit of 'Shared Exploration' is related to various activities which the family members do together and serves in happiness.
- The CV3 explains us the 'Escape and Relaxation' which is related to activities like 'canoeing and kayaking' and 'farm visits and agritourism sites'. This explains us that families which consider 'Escape and Relaxation' as benefit generally take part in outdoor activities.
- The CV4 explains us the relation between the benefit factor of 'Experiential Learning for Children' appear to show a good correlation with the seven activities like 'visiting a historical site', 'visiting a natural or ecological site', 'shopping for art and crafts', 'enjoying local tastes and delicacies', 'buying local specialities and souvenirs' and so on. This means that the benefit of 'Experiential Learning for Children' is related to activities which are related to nature and culture or in materialistic pleasures.

- **What conclusions does CC allow them to draw?**

Solution:

There are multiple conclusions that have been derived from the study:

- The 19 benefits obtained from participating in different activities were condensed into 4 major factors which can explain the majority of the data.
- The CV1 explains us the significant relationship between the activity of 'Taking pictures and videos' and the benefits under the factor of 'Communication and Togetherness'.
- CV4 which denotes the 'Experiential Learning for Children' is the most important according to the Chinese family followed by 'Communication and Togetherness', 'Shared Exploration' and 'Escape and Relaxation'.
- Benefit obtained is influenced by making digital memories via photos and videos. This process of capturing the experiences in forms of photos and videos instills a sense of family and unity.

- The benefit seeking and going on vacations is influenced by experiences which can benefit the children which can broaden their horizons and help them learn. It is critical because the Chinese system of education does not ensure holistic learning.
- Also, the choice of destination is important because family members want to seek peace and relaxation away from the daily routine.

3) CCA in R: Perform the following Canonical Correlation Analysis on the Young People Survey from R PCA and FA Lab - Young People Survey. Perform a canonical correlation analysis describing the relationships between the spending and phobias variables using the data under the R PCA and FA Lab - Young People Survey in the content folder).

1. Answer the following questions regarding the canonical correlations.

- **Test the null hypothesis that the canonical correlations are all equal to zero. Give your test statistic, d.f., and p-value.**

Solution:

Null Hypothesis: The canonical correlations are all equal to zero.
Alternate Hypothesis: All the canonical correlations are not equal to zero.

Let's assume the level of significance is $\alpha = 0.05$.

We will use the Bartlett's Chi-Squared test to find the optimum number of the variates to be used.

```
Bartlett's Chi-Squared Test:

          rho^2      Chisq df    Pr(>X)
CV 1 1.6131e-01 2.3364e+02 70 < 2.2e-16 ***
CV 2 9.3337e-02 1.1473e+02 54 2.873e-06 ***
CV 3 3.5065e-02 4.8488e+01 40    0.1679
CV 4 1.7883e-02 2.4358e+01 28    0.6625
CV 5 1.0335e-02 1.2160e+01 18    0.8389
CV 6 5.8135e-03 5.1373e+00 10    0.8818
CV 7 1.7676e-03 1.1959e+00  4    0.8788
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The degrees of freedom for the CV1 is 7*10 = 70.
The degrees of freedom for the CV2 is (7-1)*(10-1) = 54.

The p-value for the CV1 is very less ~ 0.001 < 0.05.
The p-value for the CV2 is very less ~ 0.001 < 0.05.

Since the p-value for the CV1 and CV2 is less than 0.05, we reject the null hypothesis. It means that there is atleast 1 pair of canonical correlations which is not equal to zero.

- **How many significant canonical variates are there?**

Solution:

When we perform the Bartlett's Chi-squared test, we see that for the first 2 CV's we get the p-value < 0.05 meaning that these 2 are significant and can be used to explain the data.

There are 2 significant canonical variates CV1 and CV2.

- **Present the first two canonical correlations (Cancor)?**

Solution:

The first canonical correlation CV1 is 0.402.
The second canonical correlation CV2 is 0.306.

```
Canonical Correlations

      CV 1       CV 2
0.40163175 0.30551101
```

- **What can you conclude from the above analyses?**

Solution:

- Approximately 40.2% of the overlapping variance is the relationship explained by the CV1. This means that close to 40.2% of the data is explaining the spending type with the phobias.
- Approximately 30.6% of the overlapping variance is the relationship explained by the CV2. This means that close to 30.6% of the data is explaining the spending type with the phobias.
-

2. Answer the following questions regarding the canonical variates.

- **Give the formulae for the first canonical variate for the spending and phobias variables.**

Solution:

Formula for the CV1 for the spending variable:

CV1 = (0.124) * Finances +
(-0.739) * Shopping.centres +
(0.153) * Branded.clothing +
(0.224) * Entertainment.spending +
(-0.495) * Spending.on.looks +
(0.396) * Spending.on.gadgets +
(-0.009) * Spending.on.healthy.eating

Formula for the CV1 for the phobias variable:

CV1 = (-0.016) * Flying +
(-0.270) * Storm +
(-0.123) * Darkness +
(0.116) * Heights +
(-0.201) * Spiders +
(-0.414) * Snakes +
(-0.182) * Rats +
(-0.147) * Ageing +
(-0.266) * Dangerous.dogs +
(0.279) * Fear.of.public.speaking

- **Give the correlations between the first canonical variate for spending and the phobias variables.**

Solution:

Correlations for Canonical variate for the spending variable:

```
Structural Correlations (Loadings):

        X Vars:
                                 CV 1       CV 2         CV 3        CV 4        CV 5       CV 6        CV 7
Finances                    0.12403017 -0.3563334  0.263921240 -0.33497651 -0.17723554 -0.2171724 -0.77278914
Shopping.centres           -0.73950257  0.3180932  0.486374598  0.04450706  0.28134478  0.0391196 -0.18090414
Branded.clothing           -0.15336570  0.6214466 -0.005850975  0.51090064 -0.03559943 -0.5370173 -0.19893333
Entertainment.spending      0.10328499  0.7798820 -0.376137189  0.03711544  0.20053265  0.3773020 -0.23598615
Spending.on.looks          -0.61272133  0.4810484 -0.104906259  0.19502256 -0.50197052  0.2550620 -0.16460515
Spending.on.gadgets         0.34878536  0.5788693  0.528537546  0.31037858 -0.33408358  0.2353319 -0.02406983
Spending.on.healthy.eating -0.02051828  0.5498327  0.129296258 -0.66655974 -0.32725050 -0.3075805  0.18586271
```

Correlations for Canonical variate for the phobias variable:

```
        Y Vars:
                                CV 1        CV 2         CV 3        CV 4       CV 5        CV 6        CV 7
Flying                   -0.2346319 -0.70983536 -0.308965001 -0.08653951 -0.2446722 -0.23004415 -0.32158972
Storm                    -0.6282086 -0.28356869 -0.227607576 -0.13766660 -0.1041981  0.30821310  0.46129352
Darkness                 -0.5105372 -0.09916868  0.064254593 -0.41057663  0.1981478  0.57327144 -0.14596262
Heights                  -0.1523808 -0.12108441 -0.568244681  0.03308342  0.5032095  0.09342092 -0.10957325
Spiders                  -0.5956342  0.03222097  0.471342158  0.19882091  0.1851932 -0.12823661 -0.16412378
Snakes                   -0.5506376 -0.05461118  0.105444396 -0.15488168 -0.1835584 -0.32909698  0.19704936
Rats                     -0.6460022 -0.11212721  0.005965679  0.50987045 -0.1199007  0.12015863 -0.04020317
Ageing                   -0.4268573  0.32236168 -0.255496786 -0.16277142 -0.4308472  0.14858695 -0.52089032
Dangerous.dogs           -0.6481888 -0.30698919 -0.024720641 -0.13329280  0.2646902 -0.09895365 -0.05630875
Fear.of.public.speaking   0.1399566 -0.49761653  0.338772216 -0.03865834 -0.1699801  0.38332053 -0.13247187
```

```
Canonical Communalities (Fraction of Total Variance
Explained for Each Variable, Within Sets):

       X Vars:
          Finances      Shopping.centres      Branded.clothing Entertainment.spending
                 1                     1                     1                     1
   Spending.on.looks   Spending.on.gadgets Spending.on.healthy.eating
                 1                     1                     1

       Y Vars:
          Flying         Storm      Darkness       Heights       Spiders
       0.8780716     0.8644588     0.8323923     0.6358315     0.6951893
         Snakes          Rats        Ageing Dangerous.dogs Fear.of.public.speaking
       0.5221180     0.7203253     0.8569314     0.6157925             0.5768478


Canonical Variate Adequacies (Fraction of Total Variance
Explained by Each CV, Within Sets):


       X Vars:
     CV 1      CV 2      CV 3      CV 4      CV 5      CV 6      CV 7
0.15627657 0.29876884 0.10782898 0.13646539 0.08896102 0.09921171 0.11248749

       Y Vars:
     CV 1      CV 2      CV 3      CV 4      CV 5      CV 6      CV 7
0.24308108 0.10711480 0.08882691 0.05653508 0.07301539 0.08032753 0.07089506
```

```
Redundancy Coefficients (Fraction of Total Variance
Explained by Each CV, Across Sets):


       X | Y:
       CV 1         CV 2         CV 3         CV 4         CV 5         CV 6         CV 7
0.0252086706 0.0278861795 0.0037810333 0.0024403678 0.0009194381 0.0005767644 0.0001988280

       Y | X:
       CV 1         CV 2         CV 3         CV 4         CV 5         CV 6         CV 7
0.0392109365 0.0099977715 0.0031147239 0.0010109992 0.0007546354 0.0004669818 0.0001253110


Aggregate Redundancy Coefficients (Total Variance
Explained by All CVs, Across Sets):

       X | Y: 0.06101128
       Y | X: 0.05468136
```

- **What can you conclude from the above analyses?**

Solution:

In the above screenshots, we can see that there are different types of correlation coefficients.

Firstly, we have the Structural Coefficients or Loadings and they represent the paired correlations between the variates.
Square of the structural coefficients represents the proportion of the variance for the variables to the canonical variates.

- The communality coefficients are obtained and they explain the proportion of total variance for each Canonical variate. If we square the value of communality coefficient and add them across a

variable across all the CVs we get the sum as 1. For example close to 12% of the variance is explained by the Finances in the CV1 and close to 35% in the CV2 but in a negatively correlated manner. For spending, the communality coefficient for CV1 is 1 and for phobias the CV1 communality coefficient is 0.87.

- The adequacy coefficients are obtained and they explain the proportion of total variance for each variable within each Canonical variate. If we take the average of them after squaring the loading values then we get the value of these adequacy coefficients. For example, the adequacy coefficient for CV1 of spending variable is 0.156. For the phobias variable CV1 adequacy coefficient is 0.243.

EXTRA CREDIT (10 points) Perform a correspondence analysis on the Reading Level and Education Level Completed liking data in readers.csv. In this file you are provided with the table for the two sets of categories. In particular perform the following

E1-Some Primary
E2-Primary Completed
E3-Some Secondary
E4-Secondary Completed
E5-Some tertiary

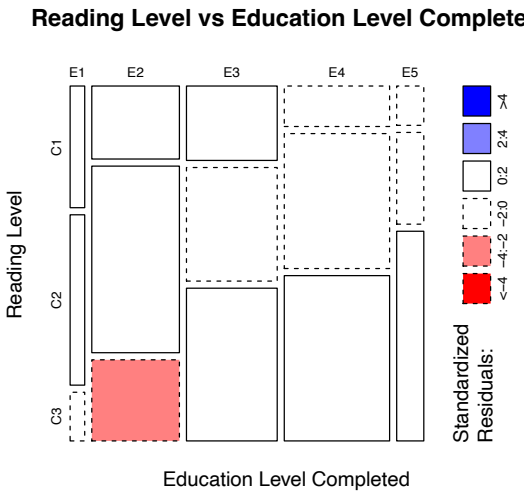
C1-Reading at a Glance
C2-Read Fairly Thoroughly
C3-Read Very Through

- **Create a mosaic plot of the two categorical variables.**

Solution:

- We have converted the data into a matrix with Education Level as rows and Reading level as the columns.
- We use the mosaicplot() function to visualise the 2 categorical variables of Education Level Completed and Reading Level.
- We observe that we have a red box for E2 education and C3 reading which clearly explains that readers who have just completed primary may have very less chance of reading very thoroughly.

```
# 4.1) Mosaic Plot of the 2 categorical variables

install.packages("vcd")
library(vcd)

mosaicplot(final_dataset, shade=TRUE, ylab="Reading Level", xlab="Education Level Completed",
           main="Reading Level vs Education Level Completed")
```



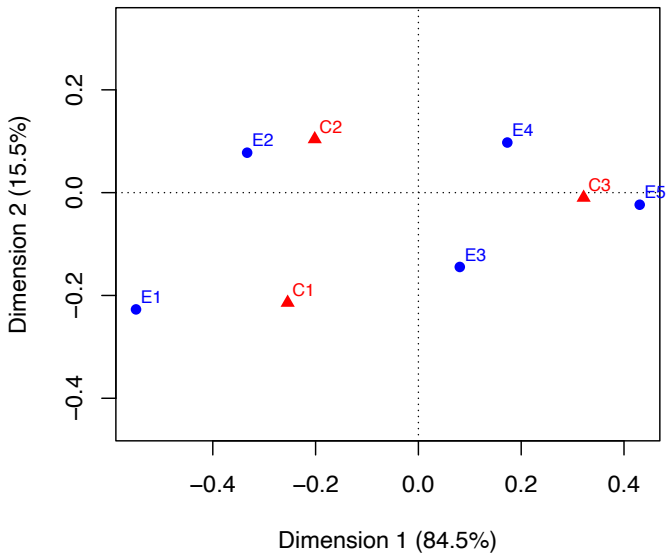
- **Plot the results of the correspondence analysis.**

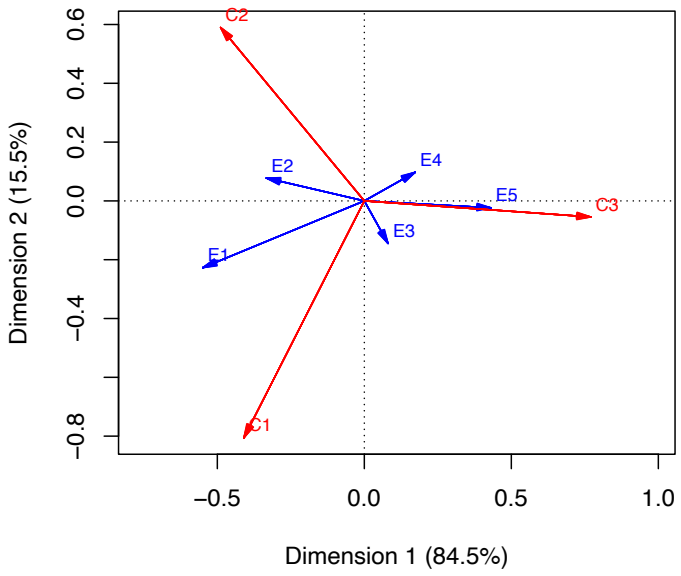
Solution:

- We use the 'ca' library for performing the Correspondence Analysis on the dataset.
- When we read the summary of the Correspondence Analysis, we can understand that there are 2 dimensions that we can use and the 1st dimension explains the majority of the data as it has the highest variance - close to 84.5%. The 2nd dimension explain around 15.5% of the data.
- When we just use the plot() on the correspondence analysis fit, we see that we have lines through the origin/average (0,0) and the items on the left have lesser relationship with a variable as compared to the items on right.

```
178  # 4.2) Correspondence Analysis and summary
179
180  fit=ca(final_dataset)
181  fit
182  summary(fit)
183
184  # Plotting correspondence analysis
185  plot(fit)
186  plot(fit, map='rowgreen', arrows=c(T, T))
187
```

```
> summary(fit)

Principal inertias (eigenvalues):

 dim    value      %    cum%   scree plot
 1      0.070369  84.5  84.5   ********************
 2      0.012892  15.5 100.0   ****
        -------- -----
 Total: 0.083260 100.0


Rows:
     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1 |   E1 |   45 1000  190 | -549 854 192 | -227 146 180 |
2 |   E2 |  269 1000  378 | -333 948 425 |   78  52 126 |
3 |   E3 |  279 1000   92 |   81 237  26 | -145 763 452 |
4 |   E4 |  324 1000  153 |  173 759 138 |   97 241 239 |
5 |   E5 |   83 1000  186 |  431 997 220 |  -24   3   4 |

Columns:
     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1 |   C1 |  183 1000  242 | -254 585 168 | -214 415 649 |
2 |   C2 |  413 1000  256 | -202 790 239 |  104 210 348 |
3 |   C3 |  404 1000  502 |  321 999 593 |  -10   1   3 |
>
```



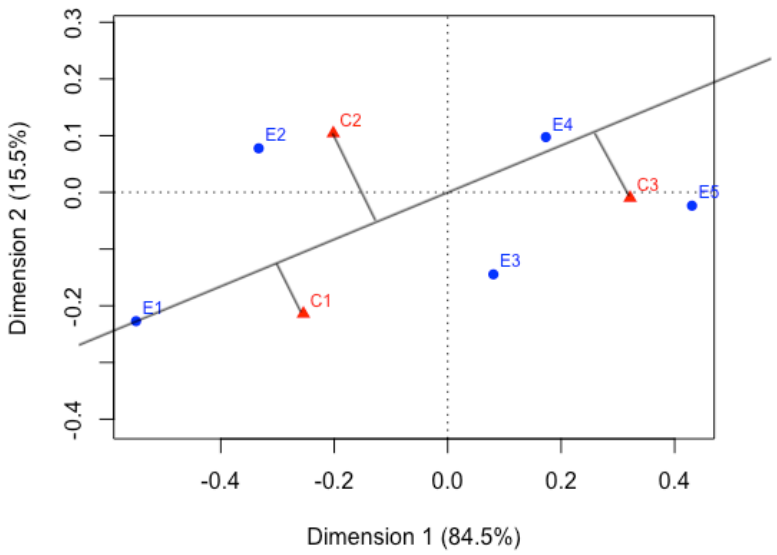


- When we just use the plot() with arrows drawn on the correspondence analysis fit, we see that if the angle between the items is acute (<90) then those 2 variables have a higher relationship. For example the variables E5 and C3 have very less angle between them and so they have high relation. This is true since C3 represents very thorough reading and E5 represents Tertiary education which is the case.
- But, if the angle is obtuse (>90), then the relationship between those variables is less. For example the variables E1 and C3 have very high angle between them and so they have less relation. This is true since C3 represents very thorough reading and E1 represents some primary education which is possible.

- **With each Education Level Completed, create a profile for the Reading Level. Which Reading Level are most highly and least highly represented. For each Education Level Completed, draw the scale for that Education Level completed and demonstrate that Reading Level profile on the graph.**

Solution:

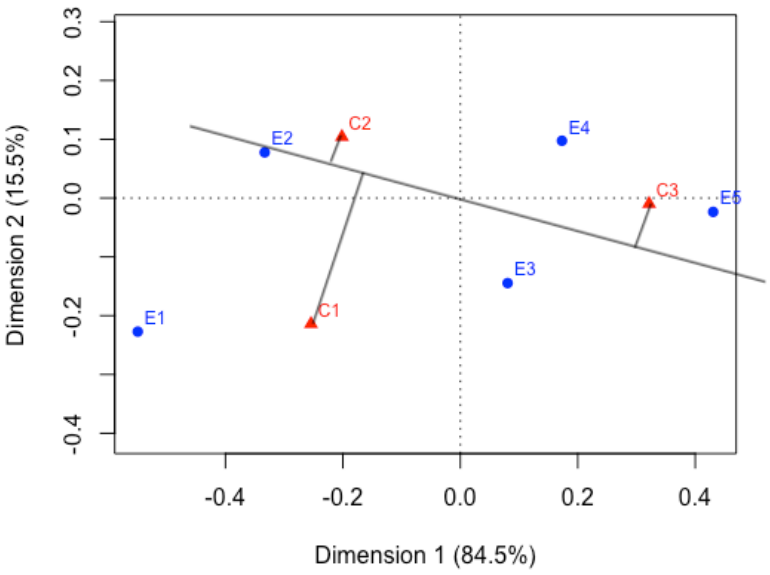
Profile with E1:



In this case when we draw a line from E1 through the origin, we see that the C1 reading is highly correlated which is obvious because anyone with some form of education can read texts at a glance. The C2 variable is having the least correlation as compared to C3 which is little absurd because people having primary education level should be able to read fairly thoroughly as compared to very thoroughly.

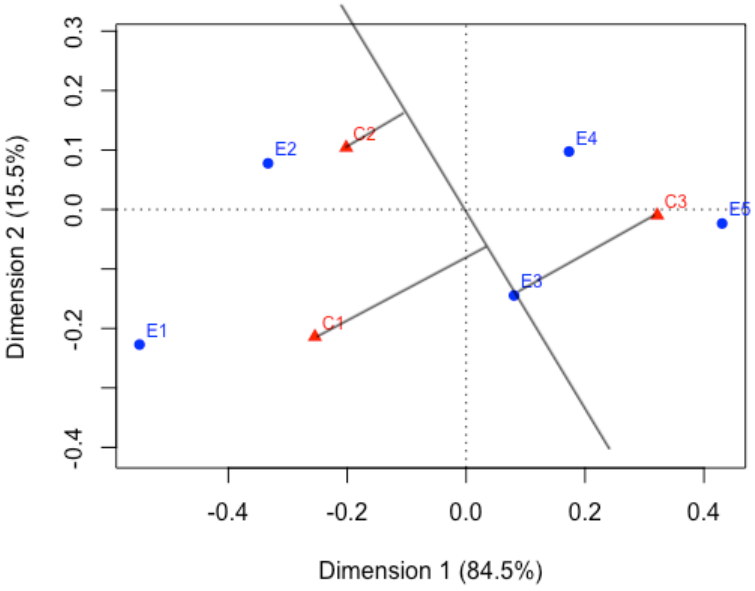
Profile with E2:

In this case when we draw a line from E2 through the origin, we see that the C1 reading is least correlated which is not what we expect because anyone with some form of completed primary education should be able to read texts at a glance. The C2 variable is having the highest correlation as compared to C3 which makes sense because people having primary education level should be able to read fairly thoroughly as compared to very thoroughly.



Profile with E3:

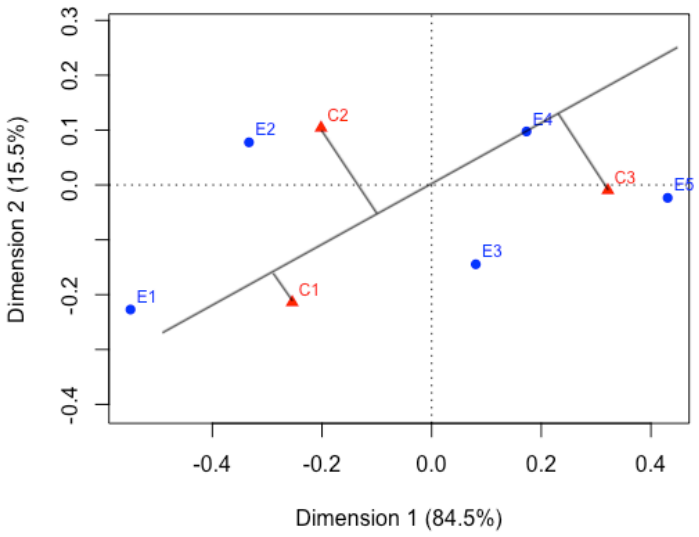
In this case when we draw a line from E3 through the origin, we see that the C1 reading is least correlated which is not what we expect because anyone with some form of secondary education should be able to read texts at a glance. The C2 variable is having the highest correlation as compared to C3 which makes sense because people having secondary education level should be able to read fairly thoroughly. Also, the C3 is correlated less as compared to C2 which indicates that people with some secondary education can still read texts very thoroughly.



Profile with E4:

In this case when we draw a line from E4 through the origin, we see that the C1 reading is highly correlated which is obvious because we expect anyone with completed secondary education level to

be able to read texts at a glance. The C2 variable is having similar correlation as C3 which makes sense because people having completed secondary education level should have the habit to read fairly thoroughly and also have the practice of reading very thoroughly.



Profile with E5:

In this case when we draw a line from E5 through the origin, we see that the C1 reading is the least correlated which is not correct because we expect anyone with some kind of tertiary education level to be able to read texts at a glance. The C3 variable is having the highest correlation which makes sense because people having some tertiary education level should have the habit to read very thoroughly. C2 is having moderate correlation with E5 which is also partly fair because people with some tertiary education should be able to read texts fairly.