Kubra Iqbal
Assignment 3

1B) (20 points, for literature review projects) Choose a technique that you will be covering in your literature review, research it, and write two paragraphs discussing when it is used, how does the technique work, and how do you interpret its results.

Clustering is a broad set of techniques that is used for finding subgroups of observations within a data set. When we are conducting cluster observations, we want the observations in this particular group to be similar and observations in different groups to be dissimilar. There is no response variable that is used, which implies that it seeks to find relationships between the N observations without being trained by a response variable. The main thing Clustering Analysis does is it allows us to identify which observations are alike and potentially categorize them.

To perform a cluster analysis in R, generally the data has to be prepared in steps. Rows and observations(individuals) and columns are variables. Any missing values in the data must be removed or estimated before the analysis is conducted. The data must be standardized to make variables comparable. Recall that, standardization usually consists of transforming the variables such that they have mean zero and standard deviation one.

2) Paper Review (10 points): An academic paper from a conference or Journal will be posted to the Homework 3 content section of D2L. It contains a usage of Canonical Correlation. Review the paper and evaluate their usage of Canonical Correlation. In particular, address (Vacation Benefits and Activities Understanding Chinese Family Travelers)

a) How suitable is their data for CC?

When the data was analyzed, first descriptive statistics was presented and then exploratory factor analysis was performed. After doing that process, canonical correlation analysis was used to assess the nature and magnitude of the relationship between benefits sought and vacation activities. Canonical correlation is used to evaluate the correlation between two sets of variables. In this study, Chinese family members were treated as one set whereas activities that participated represent the other set. Canonical correlation is appropriate to be used when the researched has limited knowledge about whether the two sets of variables are related and how strong is the relationship between them two.

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

Using CC the study examined the relationship between benefits sought and the activity participation of Chinese family travelers. Before conducting the analysis, baseline statistical assumptions including, sample size, linearity and multicollinearity were checked to ensure that the CC will be a good fit for this particular study. To continue with the CC, four separate analyses were performed between the benefit items of each of the four benefit factors and 32 activities. Each benefit dimension was treated as one set and activity items were constituted the other set.

c) 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?

There were a few different methods used to judge the quality of the correlation. The first canonical variate pair shows a significant relationship between taking pictures and videos and four items under the factor of Communication and Togetherness: having gun with family members, respecting family members decision, finding more things in common and sharing quality time together. In simpler words they are associated with communication and togetherness and seem to be related to the activity of digitally capturing family trip experiences.
The second one was, canonical variate reveals a relationship between Shared exploration and activity participation The benefits that it includes are mostly, sharing experiences, spending more time with family.
The third significant pair consists of two benefit items, escaping from the routine life and relaxing and both of them are about families that consider escaping and relaxation as important.
Lasly, the canonical variate, it's a significant relationship between the three items in the benefit dimension of Experiential Learning for Children and seven activity items.

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

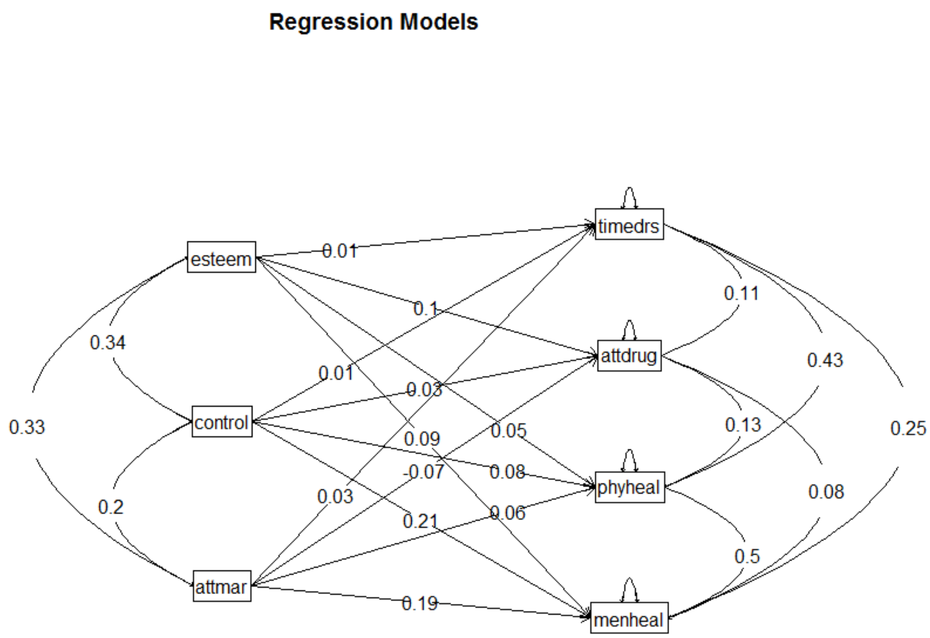
As a multivariate technique, canonical correlation analysis simultaneously evaluates the correlation between two sets of variables. In this study, benefits sought by Chinese family travelers were treated as one set, whereas activities participated in represented the other set.

e) What conclusions does CC allow them to draw?

The outcome of the study provide practical insights for destination marketers looking to tap into this particular segment. Marketing strategies should be aligned with three separate yet intertwining aspects of vacation benefits sought, including child-centric learning and experience, family-level interactions and personal - level relaxation. It is important to change the nature of the Chinese family and its implications for tourism. The tension between an individualistic view of tourism and the social reality of the family holiday deserves explicit managerial attention. Although a lot of people do travel with their families, both tourism and hospitality

studies have ignored the family, preferring instead to focus on individual travelers and group tours. For Chinese people it's not more about the destination but more about spending time with the family. There are also some implications that were drawn from the study - results from this study demonstrated that benefits pursued during family vacations may coincide with many aspects of family life in general, including relationship, education, family legacy and search for continuity.
Overall the study explains that Chinese family segment is fastly growing one to both domestic and international destinations. This population has its unique needs, desires and wants to spend quality time with their family during vacations.

3. Answer the following questions regarding the canonical correlations.



```
Call: setCor(y = y, x = x, data = data, z = z, n.obs = n.obs, use = use,
    std = std, square = square, main = main, plot = plot, show = show)

Multiple Regression from raw data

 DV =  timedrs
        slope   se    t    p  VIF
esteem   0.01 0.05 0.25 0.80 1.25
control  0.01 0.05 0.14 0.89 1.15
attmar   0.03 0.05 0.53 0.60 1.14

 Multiple Regression
           R R2  Ruw R2uw Shrunken R2 SE of R2 overall F df1 df2     p
timedrs 0.04  0 0.03    0       -0.01        0       0.2   3 461 0.899

 DV =  attdrug
        slope   se     t     p  VIF
esteem   0.10 0.05  1.96 0.051 1.25
control  0.03 0.05  0.68 0.500 1.15
attmar  -0.07 0.05 -1.39 0.170 1.14

 Multiple Regression
           R   R2  Ruw R2uw Shrunken R2 SE of R2 overall F df1 df2     p
attdrug 0.11 0.01 0.11 0.01        0.01     0.01      1.96   3 461 0.119

 DV =  phyheal
        slope   se    t     p  VIF
esteem   0.05 0.05 1.05 0.300 1.25
control  0.08 0.05 1.67 0.095 1.15
attmar   0.06 0.05 1.14 0.250 1.14

 Multiple Regression
           R   R2  Ruw R2uw Shrunken R2 SE of R2 overall F df1 df2     p
phyheal 0.14 0.02 0.14 0.02        0.01     0.01      3.09   3 461 0.027

 DV =  menheal
        slope   se    t       p  VIF
esteem   0.09 0.05 1.94 5.3e-02 1.25
control  0.21 0.05 4.48 9.5e-06 1.15
attmar   0.19 0.05 4.16 3.9e-05 1.14

 Multiple Regression
           R   R2  Ruw R2uw Shrunken R2 SE of R2 overall F df1 df2        p
menheal 0.36 0.13 0.35 0.13        0.12     0.03     23.06   3 461 6.35e-14

Various estimates of between set correlations
Squared Canonical Correlations
[1] 0.13522 0.01178 0.00019
Chisq of canonical correlations
[1] 66.827  5.449  0.088

 Average squared canonical correlation =  0.05
 Cohen's Set Correlation R2 =  0.15
 Shrunken Set Correlation R2 =  0.12
 F and df of Cohen's Set Correlation  6.13 12 1201.46
Unweighted correlation between the two sets =  0.22
```

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

```
[1] 0.8544348
> ## [1] 0.6963021
> # df (n - 1) where n = 465
> a$df[1] + a$df[2]
[1] 464
> ## [1] 465
> print("p-value from txt, Sig. F = 0.000")
[1] "p-value from txt, Sig. F = 0.000"
> ## [1] "p-value from txt, Sig. F = 0.000"
> # ct = corr.test(attitudnal, health)
> # ct$p
```

b. Test the null hypothesis that the second canonical correlations equal zero.

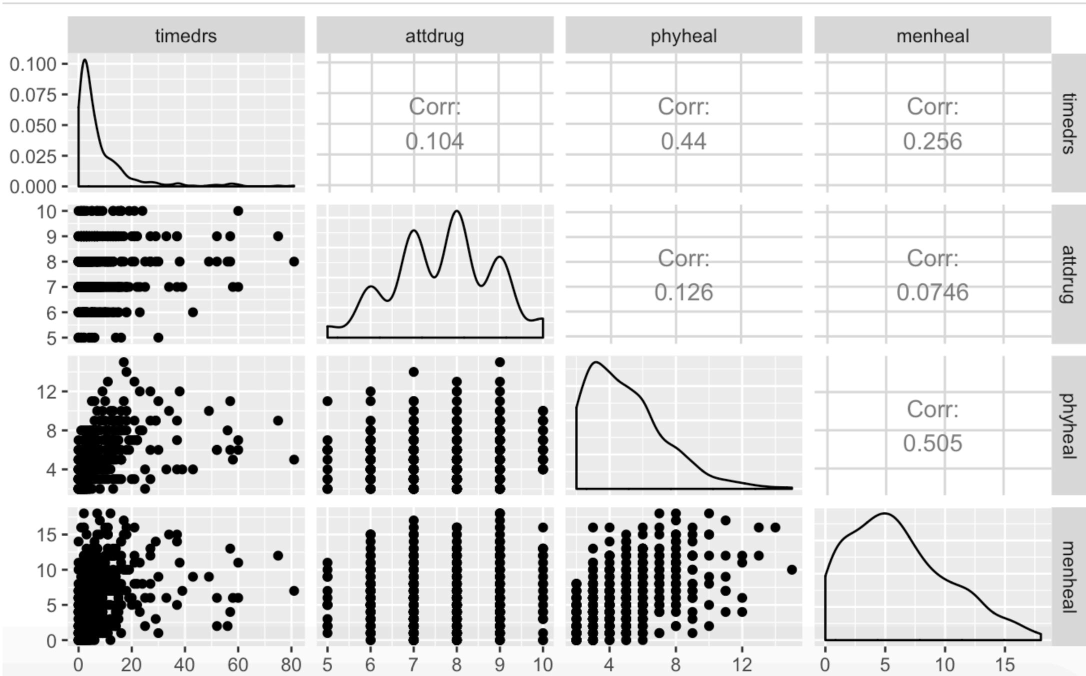
Give your test statistic, d.f., and p-value.

```
[1] "p-value from txt, Sig. F = 0.000"
> ## [1] "p-value from txt, Sig. F = 0.000"
> # the test statistic - part b
> (1 - a$cancor2[2]) * (1 - a$cancor2[3])
[1] 0.9880335
> a$df[1] + a$df[2]
[1] 464
> ## [1] 465
> print("p-value from txt, Sig. F = 0.000")
[1] "p-value from txt, Sig. F = 0.000"
> ## [1] "p-value from txt, Sig. F = 0.000"
```

c. Present the two canonical correlations

```
> # correlations between the two groups of variables
> matcor(attitudnal,health)
$Xcor
           esteem   control    attmar
esteem  1.0000000 0.3430881 0.3111108
control 0.3430881 1.0000000 0.1956356
attmar  0.3111108 0.1956356 1.0000000

$Ycor
          timedrs    attdrug   phyheal    menheal
timedrs 1.0000000 0.10429935 0.4395293 0.25557025
attdrug 0.1042993 1.00000000 0.1256049 0.07463548
phyheal 0.4395293 0.12560492 1.0000000 0.50494642
menheal 0.2555703 0.07463548 0.5049464 1.00000000

$XYcor
             esteem    control      attmar     timedrs     attdrug    phyheal    menheal
esteem  1.000000000 0.34308805  0.31111077 0.005161781  0.10630534 0.08808581 0.21187992
control 0.343088054 1.00000000  0.19563563 0.016727610  0.05484031 0.11207370 0.27851693
attmar  0.311110770 0.19563563  1.00000000 0.040998792 -0.03818892 0.09489314 0.27275821
timedrs 0.005161781 0.01672761  0.04099879 1.000000000  0.10429935 0.43952926 0.25557025
attdrug 0.106305335 0.05484031 -0.03818892 0.104299345  1.00000000 0.12560492 0.07463548
phyheal 0.088085806 0.11207370  0.09489314 0.439529262  0.12560492 1.00000000 0.50494642
menheal 0.211879923 0.27851693  0.27275821 0.255570252  0.07463548 0.50494642 1.00000000

> |
```

```
> # display the canonical correlations
> cc1 <- cc(attitudnal, health)
> cc1$cor
[1] 0.367463665 0.135672061 0.009701075
>
> cc1[3:4]
$xcoef
               [,1]        [,2]        [,3]
esteem  -0.05773755 -0.22178979  0.16095259
control -0.46723726 -0.06432804 -0.70191929
attmar  -0.06412407  0.09363553  0.04908886

$ycoef
               [,1]        [,2]        [,3]
timedrs  0.01270721  0.03117169  0.08349010
attdrug -0.03939882 -0.83656961  0.24615563
phyheal  0.04472619 -0.06985953 -0.40187317
menheal -0.25466365  0.03054455  0.06081644

>
```



d. What can you conclude from the above analyses?

The analysis above shows that the Canonical correlation coefficients test for the existence of the relationships between two sets of the given variables. The coefficients are low and this means that the health variables and the attitudinal variables are not positively correlated with each other. The R squared values are also low.

2. Answer the following questions regarding the canonical variates.

a. Give the formulae for the significant canonical variates for the attitudinal

and health variables.

The linear combination of the sets of variables (predictor and DV). Significant canonical variates will have a low p-value (< 0.05).

$$\mathrm{H0}: \rho_1^* = \rho_2^* = \cdots = \rho_p^* = 0$$

b. Give the correlations between the significant canonical variates for

attitudinal and the attitudinal variables, and the correlations between the

significant canonical variates for health and the health variables.

```
> # question 2
>
> # compute canonical loadings
> cc2 <- comput(attitudnal, health, cc1)
>
> # display canonical loadings/latent variables
> cc2[3:6]
$corr.X.xscores
              [,1]       [,2]       [,3]
esteem  -0.6011354 -0.6563911  0.4600423
control -0.7779911 -0.2260033 -0.5888231
attmar  -0.7341083  0.5124392  0.4428459

$corr.Y.xscores
               [,1]         [,2]          [,3]
timedrs -0.03351197  0.026783335  0.0055485474
attdrug -0.03604043 -0.127721918  0.0029821685
phyheal -0.13829624 -0.010802913 -0.0040382375
menheal -0.36005461  0.005716298 -0.0004422981

$corr.X.yscores
              [,1]        [,2]         [,3]
esteem  -0.2209931 -0.08860846  0.004464254
control -0.2856098 -0.03039026 -0.005696040
attmar  -0.2740008  0.06969372  0.004278934

$corr.Y.yscores
               [,1]        [,2]       [,3]
timedrs -0.09161342  0.19981276  0.5870960
attdrug -0.09732843 -0.94295266  0.2783981
phyheal -0.37701291 -0.07364105 -0.3934967
menheal -0.98186138  0.05888152  0.0252500
```

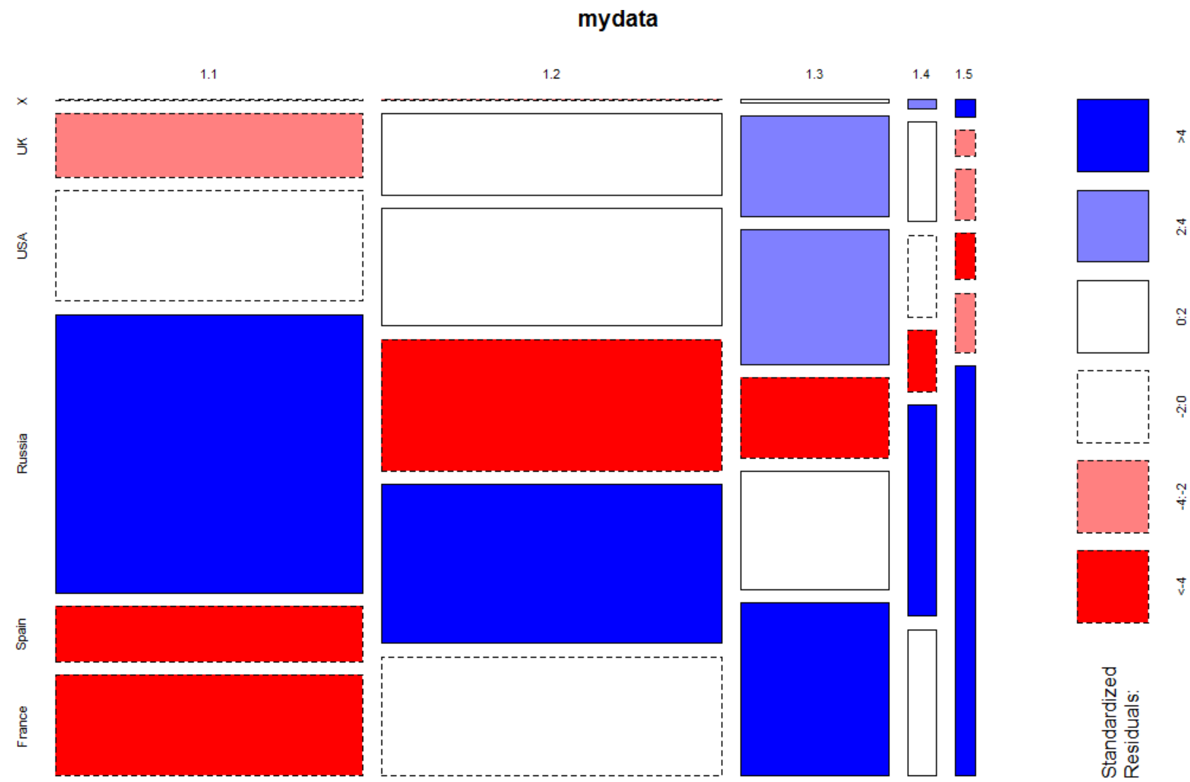
```
>
> pv <- pf(f, d1, d2, lower.tail = FALSE)
> (dmat <- cbind(WilksL = w, F = f, df1 = d1, df2 = d2, p = pv))
        WilksL        F df1      df2            p
[1,] 0.8489691 6.448088  12 1212.046 3.203242e-11
[2,] 0.9815007 1.435138   6  918.000 1.980243e-01
[3,] 0.9999059      NaN   2      NaN          NaN
>
```

c. What can you conclude from the above analyses?

The analysis above that the health variable is more related to each other when compared to attitudinal variable group. The canonical variates are interesting enough to use to represent the relationship that is being seen above.

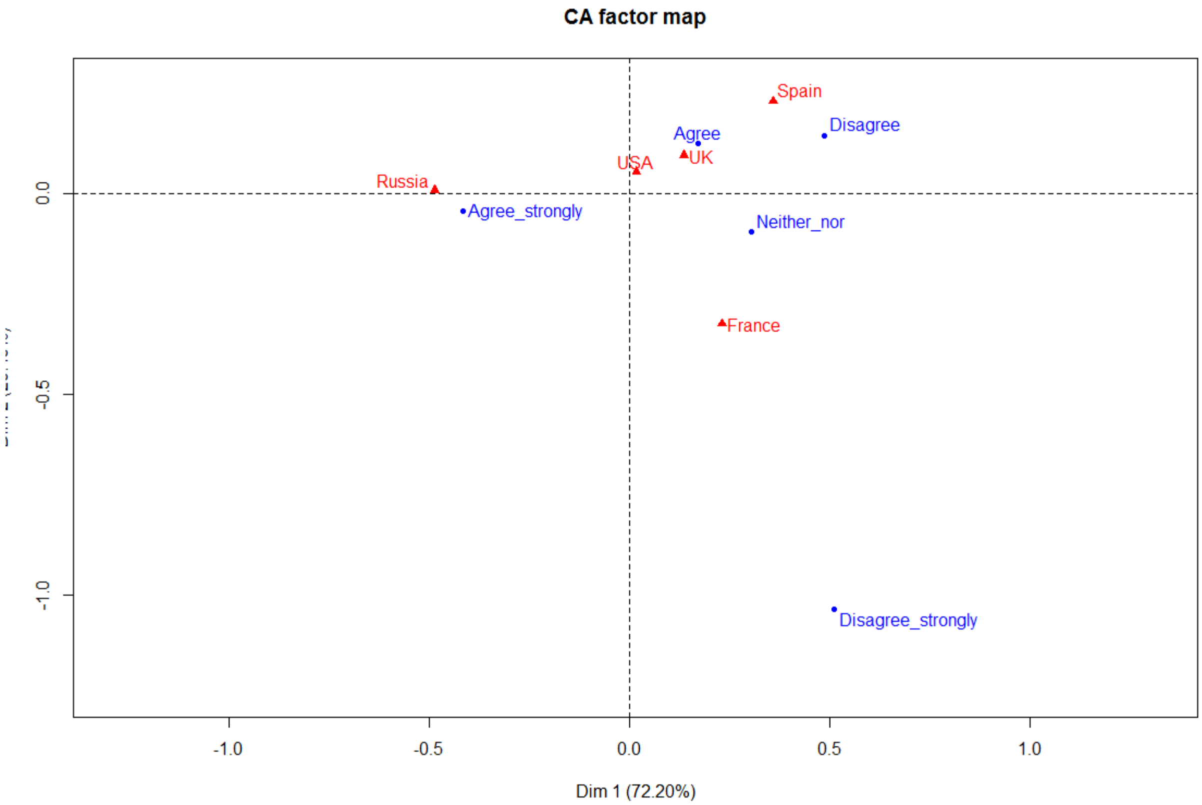
EXTRA CREDIT (10 points)

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

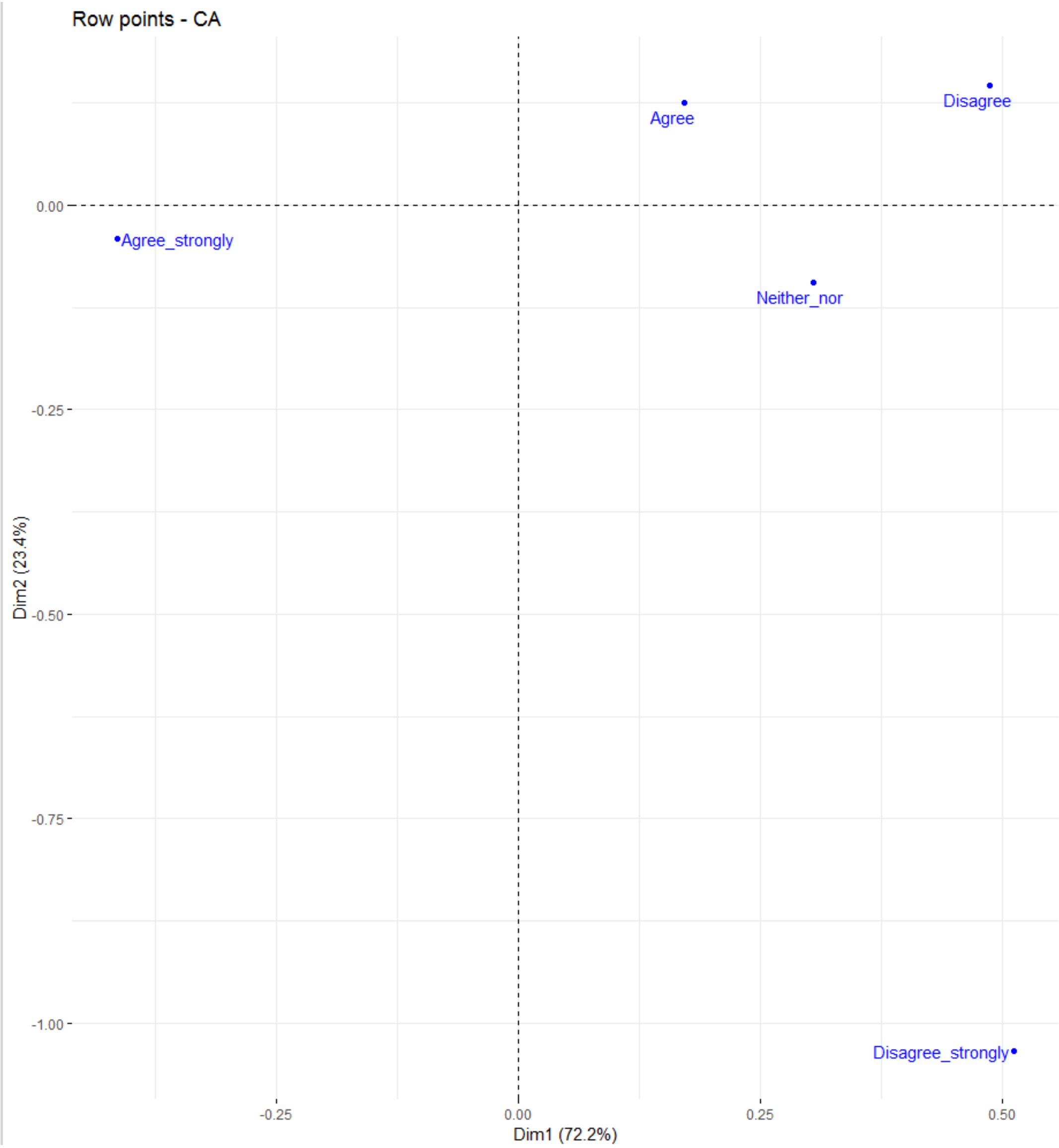


b) Plot the results of the correspondence analysis

```
Call:
CA(X = mydata[, 1:5])

The chi square of independence between the two variables is equal to 879.2675 (p-value =  7.510022e-177 ).

Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4
Variance               0.105   0.034   0.006   0.000
% of var.             72.204  23.445   4.021   0.330
Cumulative % of var.  72.204  95.649  99.670 100.000

Rows
                   Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2  
Agree_strongly    |    63.147 | -0.415 59.733  0.990 | -0.041  1.826  0.010 |  0.001  0.004  0.000 |
Agree             |    18.842 |  0.171 11.264  0.626 |  0.126 18.699  0.337 | -0.039 10.386  0.032 |
Neither_nor       |    21.401 |  0.304 15.553  0.761 | -0.095  4.655  0.074 |  0.142 60.642  0.165 |
Disagree          |     9.776 |  0.487  7.667  0.821 |  0.147  2.140  0.074 | -0.140 11.379  0.068 |
Disagree_strongly |    31.784 |  0.512  5.782  0.190 | -1.035 72.680  0.777 | -0.211 17.589  0.032 |

Columns
                   Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2  
UK                |     5.379 |  0.137  2.287  0.445 |  0.096  3.500  0.221 |  0.109 26.285  0.285 |
USA               |     2.938 |  0.018  0.056  0.020 |  0.056  1.737  0.201 |  0.105 35.704  0.708 |
Russia            |    67.195 | -0.486 63.773  0.993 |  0.010  0.082  0.000 | -0.039  7.190  0.006 |
Spain             |    35.611 |  0.358 22.915  0.673 |  0.231 29.318  0.280 | -0.094 28.535  0.047 |
France            |    33.827 |  0.232 10.970  0.339 | -0.323 65.363  0.657 | -0.025  2.286  0.004 |
> dimdesc(res)
$`Dim 1`
$`Dim 1`$row
                       coord
Agree_strongly    -0.4146165
Agree              0.1709538
Neither_nor        0.3040600
Disagree           0.4872861
Disagree_strongly  0.5120719

$`Dim 1`$col
             coord
Russia -0.48616128
USA     0.01756578
UK      0.13679137
France  0.23190124
Spain   0.35817156


$`Dim 2`
$`Dim 2`$row
                        coord
Disagree_strongly -1.03450563
Neither_nor       -0.09478680
Agree_strongly    -0.04131211
Agree              0.12551026
Disagree           0.14670250

$`Dim 2`$col
              coord
France -0.322568380
Russia  0.009916321
USA     0.055850022
UK      0.096429881
Spain   0.230858890


$`Dim 3`
$`Dim 3`$row
                          coord
Disagree_strongly -0.2107439093
Disagree          -0.1400812750
Agree             -0.0387363104
Agree_strongly     0.0007682517
Neither_nor        0.1416795851

$`Dim 3`$col
             coord
Spain  -0.09431693
Russia -0.03852068
France -0.02497957
USA     0.10485520
UK      0.10942796
```



c) **With each country, create a profile for the sports likings. Which sports liking are most highly and least highly represented. For each country, draw the scale for that country and demonstrate that sports liking profile on the graph.**



Row points - CA
Disagree
Agree
Agree_strongly
Neither_nor
Disagree_strongly
0.00
-0.25
-0.50
-0.75
-1.00
Dim2 (23.4%)
-0.25
0.00
0.25
0.50
Dim1 (72.2%)