Monash University: Assessment Cover Sheet

Student name	Hussain	Syed Jazil	
School/Campus		**Student's I.D. number**	28900766
Unit name	ETF3500 - ETF5500 - High Dimensional Data Analysis S2 2020		
Lecturer's name		**Tutor's name**	
Assignment name	ETF3500 - Exam - 23 November 2020, 10:00am (Melbourne, Australian time)	**Group Assignment: No** **Note, each student must attach a coversheet**	

Lab/Tute Class:	**Lab/Tute Time:**	**Word Count:**
Due date: 23-11-2020	**Submit Date:**	**Extension granted** □

If an extension of work is granted, specify date and provide the signature of the lecturer/tutor. Alternatively, attach an email printout or handwritten and signed notice from your lecturer/tutor verifying an extension has been granted.

Extension granted until (date): ..../......./........... Signature of lecturer/tutor: ....

Late submissions policy	**Days late**	**Penalty applied**
Penalties apply to late submissions and may vary between faculties. Please refer to your faculty's late assessment policy for details.		

Patient/client confidentiality: Where a patient/client case study is undertaken a signed Consent Form must be obtained.

Intentional plagiarism or collusion amounts to cheating under Part 7 of the Monash University (Council) Regulations

Plagiarism: Plagiarism means to take and use another person's ideas and or manner of expressing them and to pass these off as one's own by failing to give appropriate acknowledgement. This includes material from any source, staff, students or the Internet - published and unpublished works.

Collusion: Collusion means unauthorised collaboration on assessable written, oral or practical work with another person. Where there are reasonable grounds for believing that intentional plagiarism or collusion has occurred, this will be reported to the Associate Dean (Education) or nominee, who may disallow the work concerned by prohibiting assessment or refer the matter to the Faculty Discipline Panel for a hearing.

Student Statement:

- I have read the university's Student Academic Integrity Policy and Procedures
- I understand the consequences of engaging in plagiarism and collusion as described in Part 7 of the Monash University (Council) Regulations (academic misconduct).
- I have taken proper care to safeguard this work and made all reasonable efforts to ensure it could not be copied.
- No part of this assignment has been previously submitted as part of another unit/course.
- I acknowledge and agree that the assessor of this assignment may, for the purposes of assessment, reproduce the assignment and:

  i. provide it to another member of faculty and any external marker; and/or

  ii. submit to a text matching/originality checking software; and/or

  iii. submit it to a text matching/originality checking software which may then retain a copy of the assignment on its database for the purpose of future plagiarism checking.

- I certify that I have not plagiarised the work of others or participated in unauthorised collaboration or otherwise breached the academic integrity requirements in the Student Academic Integrity Policy.

Date: ..../......./............ Signature:... *

Privacy Statement:

For information about how the University deals with your personal information go to http://privacy.monash.edu.au/guidelines/collection-personal-information.html#enrol

High Dimensiona Data Analysis

Syed Jazil Hussain 28900766

A Standardisation and Distance (10 Marks)

The following question only requires you to use the variables `income`, `experience` and `age`.

1. Standardise `income`, `experience` and `age` by centering (subtracting the mean) and scaling (dividing by the standard deviation) using the `scale` function. Print out the first 5 observations. **(1 Marks)** Beer%>% dplyr::select(cost,calories,alcohol)%>% summarise_all(mean)->means print(means)

Beer%>% dplyr::select(cost,calories,alcohol)%>% summarise_all(sd)->sds print(sds)

PabstEL_std<-(PabstEL-means)/sds Augs_std<-(Augs-means)/sds

dif<-PabstEL_std-Augs_std print(dif)

```
data<-read.csv("data28900766.csv")

data%>%
  select(income,experience,age)%>%
  scale%>%
head(Data, n=5)
```

```
##          income  experience         age
## [1,]  1.2516141 -0.09505763  0.05490783
## [2,]  1.5201830  1.49986899  1.72337785
## [3,] -0.5635220  0.70240568  0.50994511
## [4,] -1.3489094 -1.21150626 -0.70348763
## [5,] -0.2820516  2.13783964  2.17841513
```

2. From your answer to Q1, what is the standardised value of `income` for the first observation (Nichols) in your data **(1 Mark)**

The standardised value of income for the first observtion (Nichols) is 1.2516

*3. The government proposes a universal basic income meaning that $10000 is added to every income. Create a variable `NewIncome` which is equal to `income` plus 10000 (**`NewIncome` is only to be used for question A**).* **(1 Mark)**

```
data%>%
  mutate(NewIncome=10000+income)->datanew
head(datanew)
```

```
##   surname    income experience age gender        sector second_language
## 1 Nichols 166637.64          7  33   Male Manufacturing         Spanish
## 2  Fisher 177771.84         17  44 Female  Construction            None
## 3 Mcbride  91386.65         12  36   Male        Retail            None
## 4   Noble  58826.46          0  28   Male        Health            None
## 5    Park 103055.71         21  47   Male  Construction            None
## 6   Ponce 142631.38         11  35 Female  Construction          German
##   education_years siblings NewIncome
## 1               9        1 176637.64
## 2              11        1 187771.84
## 3               1        4 101386.65
## 4               0        4  68826.46
## 5               0        4 113055.71
## 6               1        3 152631.38
```

4. Find the Euclidean Distance between the first and second observation (Nichols and Fisher) using `income`*,* `experience` *and* `age` *as the variables. Do NOT standardise the data* **(1 Marks)**

Nichols: Income=166637.64, Experience=7, Age=33

Fisher: Income=177771.84, Experience=17, Age 44

Euclidean distance = sqrt[(income1-income2)$^{2+(\text{experience1-experience2})}$2+(age1-age2)ˆ2] =11134.21

5. Find the Euclidean Distance between the first and second observation (Nichols and Fisher) using `NewIncome`*,* `experience` *and* `age` *as the variables. Do NOT standardise the data* **(1 Mark)**

Nichols: NewIncome=176637.64, Experience=7, Age=33

Fisher: NewIncome=187771.84, Experience=17, Age 44

Euclidean distance = sqrt[(income1-income2)$^{2+(\text{experience1-experience2})}$2+(age1-age2)ˆ2] =11134.18

6. Are the answers to Question 4 and Question 5 the same? Why or why not? **(1 Marks)**

The answers between question 4 and 5 are the same because the new income increases by the same ammount resulting in the same euclidean distance being calculated

7. Consider that you are working for a business that streams movies. You have access to data on a list of movies that each customer has seen. How could you use this data to define a distance between two different customers? **(2 Marks)**

Jaccard similarity is used to determine how close two sets of data are, in this case, the two different customers and the list of movies they have seen. The jaccard distance will be calculated by subtracting the jaccard similarity from 1.

8. For the example in the previous question, describe how collaborative filtering can be used to make recommendations of movies to customers. **(2 Marks)**

Collaborative filtering is a recommendation system, it is an algorithm where similar users or items are based off of similar users. In the case of the customers, it will see which movies they have in common, and based off of that conslusion it will recommend them similar movies. If they have movies in common and customer 1 has seen a specific movie, it will recommend that movie to customer 2.

B Principal Components Analysis (10 Marks)

1. *Carry out Principal Components on the data using all numeric variables.* **(2 Marks)**

```
data%>%
  select_if(is.numeric)%>%
  prcomp(scale.=TRUE)->pca

pca
```

```
## Standard deviations (1, .., p=5):
## [1] 1.5593534 1.3570942 0.6960287 0.4404186 0.2197446
##
## Rotation (n x k) = (5 x 5):
##                        PC1        PC2        PC3         PC4         PC5
## income          -0.5131940 -0.1041249 -0.8275091 -0.20139326 -0.02143521
## experience      -0.4165148  0.5409187  0.1757315 -0.01583418  0.70907696
## age             -0.4135222  0.5423141  0.1932394  0.05420454 -0.70328884
## education_years -0.4649236 -0.4373680  0.1593295  0.75215440  0.03785689
## siblings         0.4195150  0.4595320 -0.4707651  0.62491252  0.02649646
```

2. *Did you standardise the variables? Why or why not?* **(2 Marks)**

Yes, we standardised the variables because they are calcaulted in different units (i.e income is in thousands). This ensures that the data is sensitive to the units of measurement.

3. *What is the weight on number of siblings for the 4th principal component?* **(1 Mark)**

The weight of on the number of siblings for the 4th principal component is 0.6249

4. *What is the standard deviation of the 3rd principal component?* **(1 Mark)**

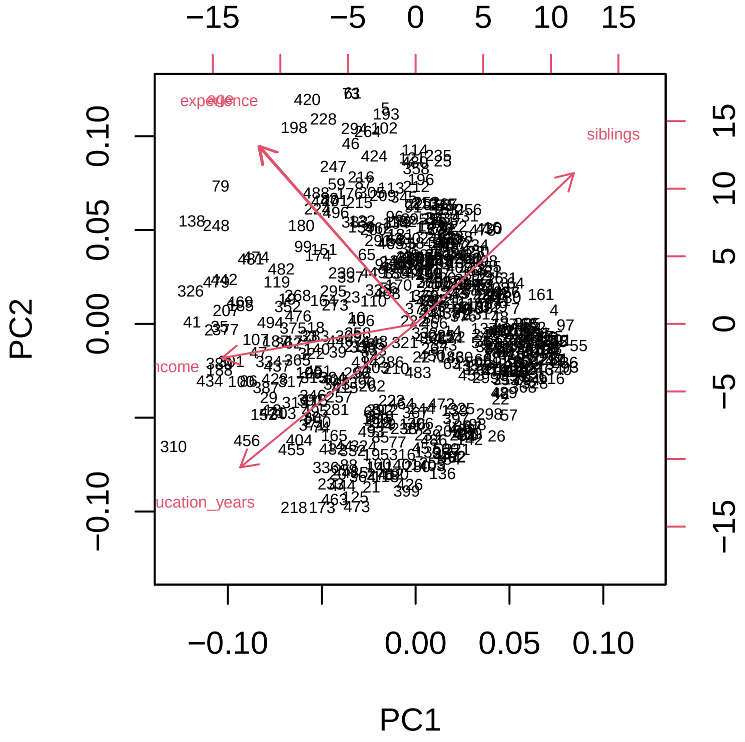
```
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5
## Standard deviation     1.5594 1.3571 0.69603 0.44042 0.21974
## Proportion of Variance 0.4863 0.3683 0.09689 0.03879 0.00966
## Cumulative Proportion  0.4863 0.8547 0.95155 0.99034 1.00000
```

The standard deviation of the 3rd principal component is 0.69603

5. *Make a distance biplot.* **(1 Marks)**

```
biplot(pca,cex=0.5)
```



6. Pick two variables that according to the biplot are highly postively correlated with one another. If there are no such variables for your dataset, then describe what you would be looking for in the biplot to indicate that two variables are postively correlated. **(1 Mark)**

Age and Experience, they are close together in the same direction.

7. Pick two variables that according to the biplot are uncorrelated. If there are no such variables for your dataset, then describe what you would be looking for in the biplot to indicate that two variables are uncorrelated. **(1 Mark)**

Education years and siblings, they are far apart in the opposite direction.

8. What proportion of overall variation in the data is explained by the biplot? **(1 Mark)**

Proportion of variance of PC1 + PC2 = 0.4863 + 0.3683 = 0.8547, 85.47% of the variation of the data is explained by the biplot.

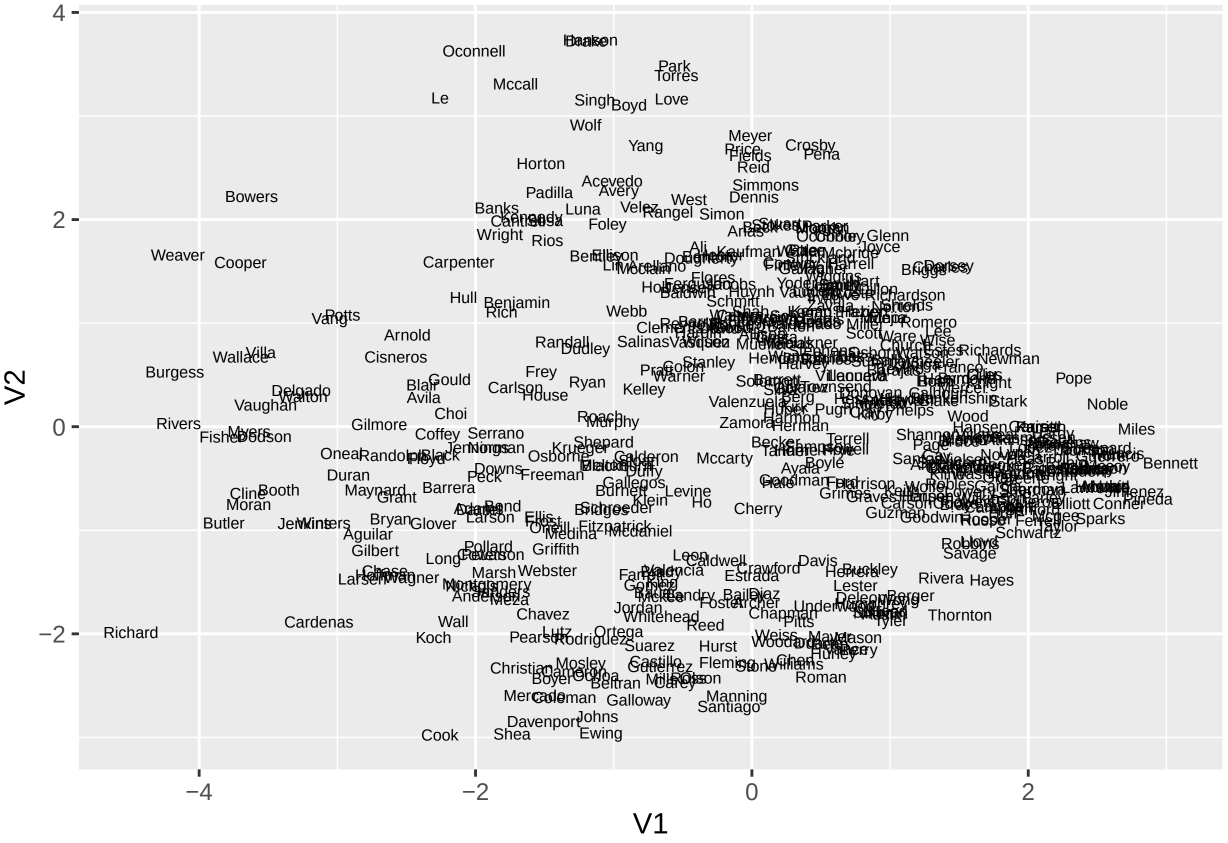
C Multidimensional Scaling (15 Marks)

1. *Using only those observations for which* `second_language` *is French, carry out classical multidimensional scaling. Find a two dimensional representation and use standardised value of* `income`*,* `experience`*,* `age`*,* `education_years` *and* `siblings` *as the variables.* **(4 Marks)**

```
data%>%
  select(income,experience,age,education_years,siblings)%>%
  scale%>%
  dist->dd

rownames(data)->attributes(dd)$Labels

cmds<-cmdscale(dd,eig=T)

cmds$points%>%
  as.data.frame()->df

df<-add_column(df,Surnames=data$surname)
```

2. *Plot a 2-dimensional representation of this data. Rather than plot the observations as points use the individuals' surnames.* **(3 Marks)**

```
ggplot(df, aes(x=V1,y=V2,label=‘Surnames‘))+geom_text(size=2)
```



3. Name two individuals (by surname) who are similar according to your plot in Question 2, and two individuals (by surname) who are different. If you were unable to generate the plot in Question 2, then describe how you would answer this question. **(1 Mark)**

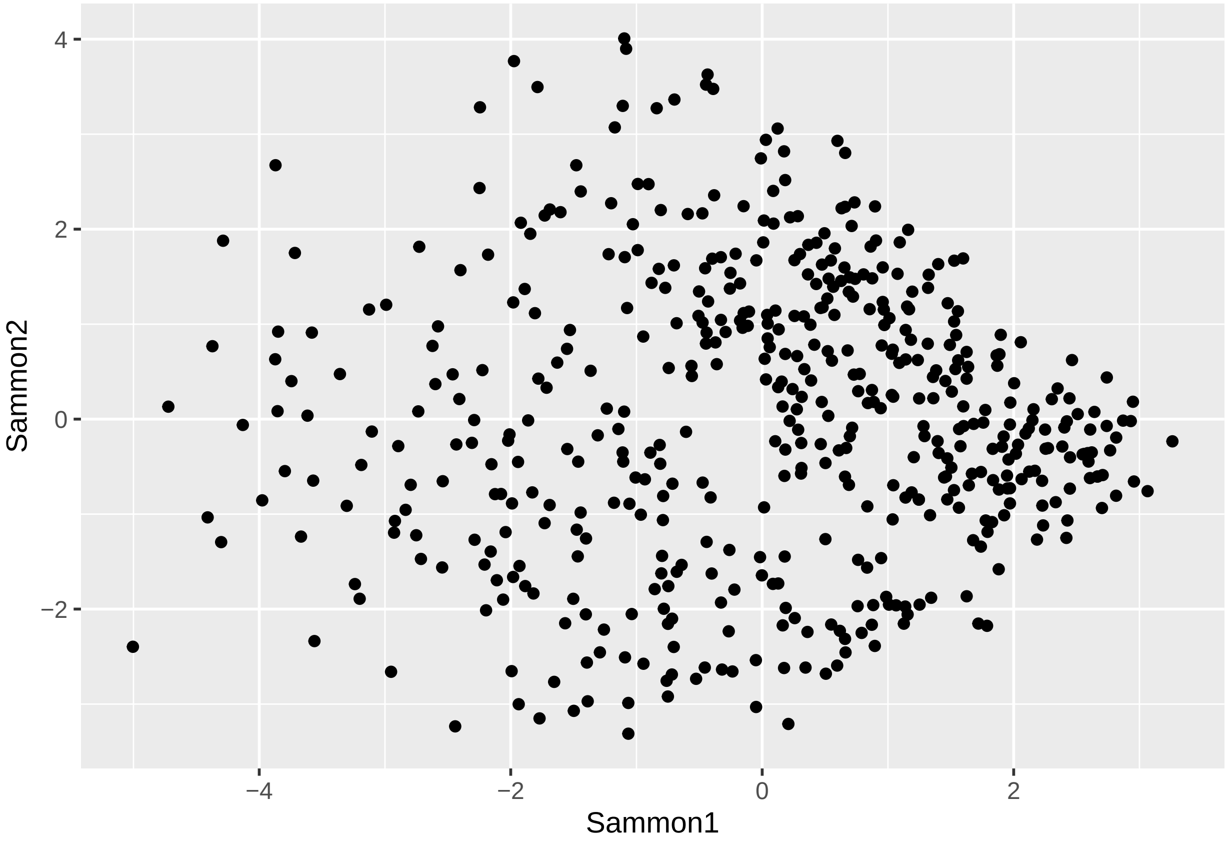
INCLUDE YOUR ANSWER HERE

4. Plot the same plot as in Question 2 using the Sammon mapping. **(3 Marks)**

```
smds<-sammon(dd)
```

```
## Initial stress        : 0.02825
## stress after  10 iters: 0.01883, magic = 0.461
## stress after  20 iters: 0.01780, magic = 0.500
## stress after  30 iters: 0.01779, magic = 0.500
```

```
df<-add_column(df,Sammon1=smds$points[,1],
               Sammon2=smds$points[,2])

ggplot(df,aes(x=Sammon1,y=Sammon2,label='surname'))+geom_point()
```



5. Are you conclusions in Question 3 robust to using a different multidimensional scaling method? If you were unable to generate the plot in Question 2 and/or Question 4, then describe how you would answer this question. **(1 Mark)**

The conclusion in question 3 is robust because sammon mapping and classical MDS give us the same solutions.

6. Describe the differences between classical multidimensional scaling and the Sammon mapping. **(3 Marks)**

Sammon mapping is not based on eigenvalue decomposition, it is not based on rotation and it is non linear mapping unlike classical MDS which is based on eigenvalue decomposition, is based on rotation and is linear mapping.

D Correspondence analysis (ETF3500 students only) (10 Marks)

1. *Construct a contingency table between the* `sector` *and* `second_language` *variables.* **(1 Mark)**

```
data%>%
  select(sector,second_language)%>%
  table%>%
  addmargins()->crosstab
print(crosstab)
```

```
##                               second_language
## sector                         Chinese French German Greek Hindi None Other
##   Accommodation                      0      4      1     7     7    0     4
##   Administrative_Support             0      0      2     2    15    0     1
##   Agriculture                        2      0      1     2     0    0     2
##   Construction                       2      0      8     0     6   10    12
##   Education                         14     20      4     3     1    4     1
##   Finance_Insurance                  6      4      0     0     1    2     0
##   Health                             0      0      3     1     4   49     0
##   Manufacturing                      6      3     12     0     0    1     1
##   Not_stated                         6      0      0     5     1   15     1
##   Other                             10      0      0     0     0    1    13
##   PublicAdministration_Safety        3      0      0     0     4   19    12
##   Retail                             4     15      0     1     6   12     1
##   Scientific_Technical               3     12      1     2     1    8    11
##   Transport_Postal_Warehousing       0      5      1     2     9   11     2
##   Wholesale                          0      1      0     1     3    0     1
##   Sum                               56     64     33    26    58  132    62
##                               second_language
## sector                         Spanish Sum
##   Accommodation                      9  32
##   Administrative_Support             0  20
##   Agriculture                        1   8
##   Construction                       0  38
##   Education                          2  49
##   Finance_Insurance                  3  16
##   Health                             1  58
##   Manufacturing                     18  41
##   Not_stated                         0  28
##   Other                             25  49
##   PublicAdministration_Safety        1  39
##   Retail                             6  45
##   Scientific_Technical               1  39
##   Transport_Postal_Warehousing       0  30
##   Wholesale                          2   8
##   Sum                               69 500
```

2. *Using the contingency table in point 1, perform correspondance analysis on the* `sector` *and* `second_language` *variables and visualise the results.* **(2 Marks)**

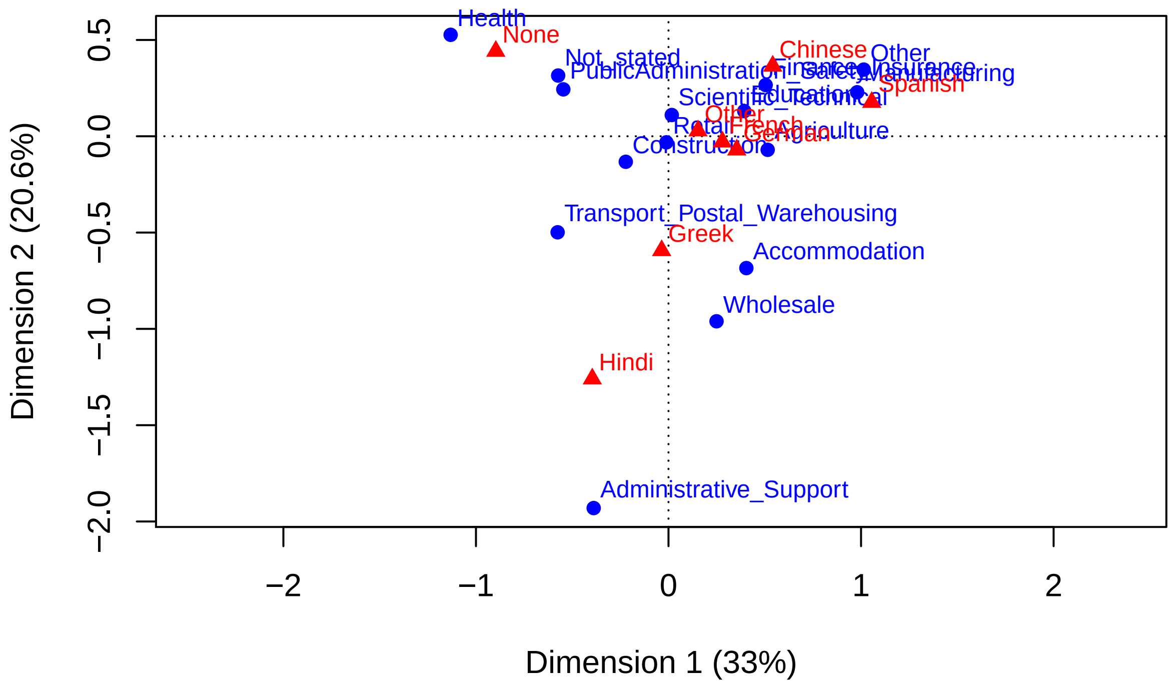
```
summary(data$sector)
```

```
##    Length     Class      Mode
##       500 character character
```

```r
summary(data$second_language)
```

```
##    Length     Class      Mode 
##       500 character character
```

```r
table(data$sector,data$second_language)%>%
  ca()%>%
  plot(cex=0.2)
```



3. Based on the results in point 2, which sector is most associated to people that speak Spanish as a second language? **(1 Mark)**

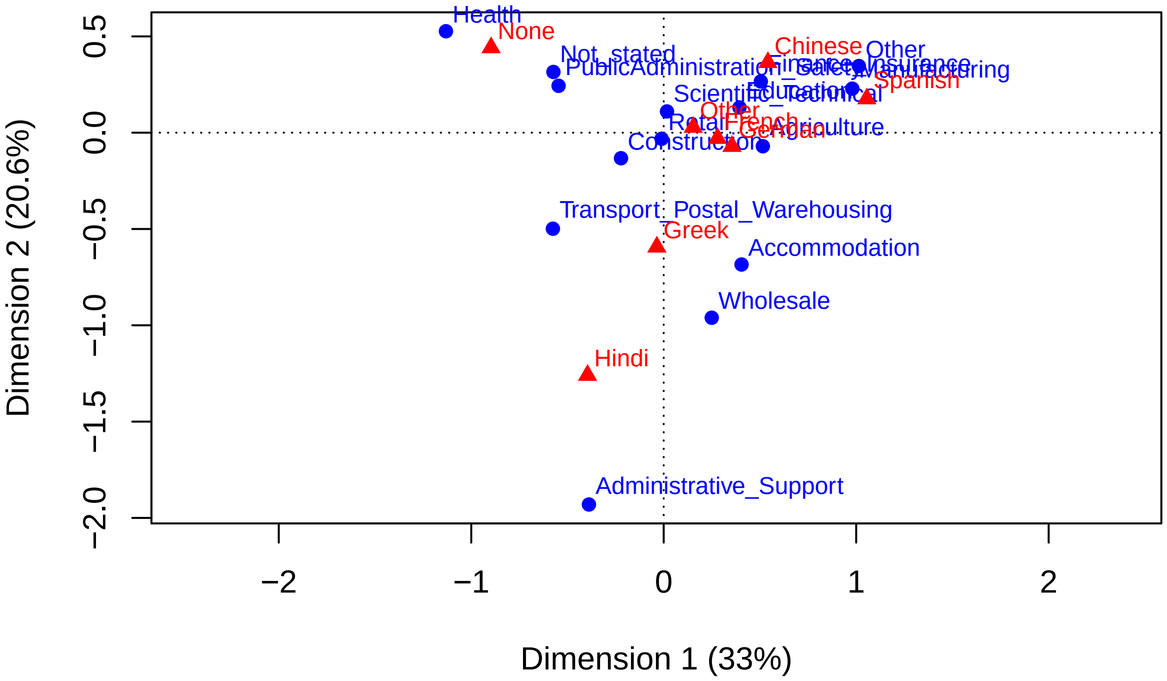
```r
summary(data$sector)
```

```
##    Length     Class      Mode 
##       500 character character
```

```r
summary(data$second_language)
```

```
##    Length     Class      Mode 
##       500 character character
```

```
table(data$sector,data$second_language)%>%
  ca()%>%
  plot(cex=0.2)
```



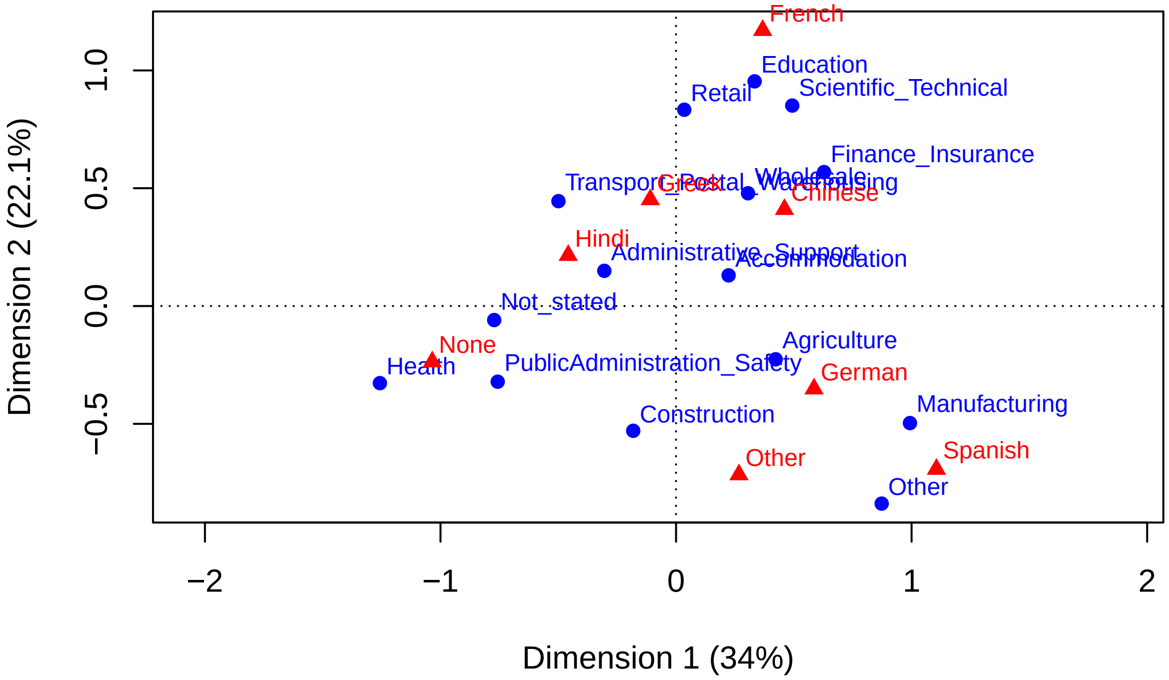
The sector Manufacturing is highly associated with people that speak Spanish.

4. Based on the results in point 2, how much inertia is explained by the first dimension? **(1 Mark)**

33% of the inertia is explained by the first dimension.

5. Repeat point 2, but this time, only consider those individuals whose `income` *is greater than 100000 and* `age` *is greater than 25.* **(2 Marks)**

```
data_filtered<-filter(data,income>100000 & age>25)

table(data_filtered$sector,data_filtered$second_language)%>%
  ca%>%
  plot()
```



6. Based on the results in point 5, how much inertia is explained by the second dimension? **(1 Mark)**

22.1% of the inertia is explained by the second dimension.

7. Compute how much inertia is explained overall by the figures in points 2 and 5. Discuss in which of these two exercises CA helps explain a larger amount of inertia. **(2 Marks)**

points 2 = 20.6 + 33 =53.6% Points 2 helps explain a total of 53.6% of the total inertia. points 5 = 22.1 + 34 =56.1% points 5 helps explain a total of 56.1% of the total inertia.

The correspondence analysis for points 5 helps explaining 2.5% more inertia than points 2.

E Correspondence analysis (ETF5500 students only) (10 Marks)

1. Using only individuals whose `gender` *is Female and whose* `income` *is less than $200000, construct a contingency table between the* `sector` *and* `second_language` *variables.* **(1 Mark)**

```
#INCLUDE YOUR R CODE HERE
```

2. Using the contingency table in point 1, perform correspondance analysis on the `sector` *and* `second_language` *variables and visualise the results.* **(1 Marks)**

```
#INCLUDE YOUR R CODE HERE
```

3. Based on the results in point 2, which sector is most associated to people that speak Spanish as a second language? **(1 Mark)**

```
#INCLUDE YOUR R CODE HERE
```

INCLUDE YOUR ANSWER HERE

4. Based on the results in point 2, how much inertia is explained by the first dimension? **(1 Mark)**

INCLUDE YOUR ANSWER HERE

5. Repeat point 2, but this time, only consider those individuals whose `gender` *is Male and whose* `income` *is less than $200000.* **(1 Marks)**

```
#INCLUDE YOUR R CODE HERE
```

6. Based on the results in point 5, how much inertia is explained by the second dimension? **(1 Mark)**

INCLUDE YOUR ANSWER HERE

7. Compute how much inertia is explained overall by the figures in points 2 and 5. Discuss in which of these two figures CA helps explain a larger amount of inertia. **(1 Marks)**

INCLUDE YOUR ANSWER HERE

8. Disscuss the differences or similarities between the results obtained in points 2 and 5, for example, are the associations between `sector` *and* `second_language` *consistent?* **(1 Mark)**

INCLUDE YOUR ANSWER HERE

9. In your own words, describe the role that the sigular value decompostion (SVD) of a matrix plays in correspondace analysis. **(2 Marks)**

INCLUDE YOUR ANSWER HERE

F Factor Modelling (5 Marks)

1. Fit a 2-factor model to the numerical variables in the dataset (set `rotation`*=‘none’).* **(1 Mark)**

```
data%>%
  select_if(is.numeric)%>%
  factanal(factors = 2,rotation = ’none’,scores = ’none’)->fa_n
fa_n
```

```
##
## Call:
## factanal(x = ., factors = 2, scores = "none", rotation = "none")
##
## Uniquenesses:
##          income      experience             age education_years        siblings
##           0.568           0.009           0.087           0.013           0.368
##
## Loadings:
##                  Factor1 Factor2
## income            0.394   0.526
## experience        0.992
## age               0.952
## education_years   0.133   0.985
## siblings                 -0.792
##
##                 Factor1 Factor2
## SS loadings       2.070   1.886
## Proportion Var    0.414   0.377
## Cumulative Var    0.414   0.791
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 1.94 on 1 degree of freedom.
## The p-value is 0.164
```

2. For each of the two factors, list the variables whose factor loadings are greater than 0.1 in absolute value. **(1 Mark)**

For factor 1, the loadings that are greater than 0.1 in absolute value are income (0.394), experience (0.992), age(0.952), education years (0.133)

For factor 1, the loadings that are greater than 0.1 in absolute value are income (0.526), education years (0.985), siblings (-0.791).

3. Provide a plot that visualises the association between factors and variables. **(1 Mark)**

```
plot(fa_n)
```

```
## Error in xy.coords(x, y, xlabel, ylabel, log): ’x’ is a list, but does not have components ’x’ and ’
```

4. Fit a 2-factor model to the numerical variables in the dataset, but now setting `rotation = "promax"`. **(1 Mark)**

```
data%>%
  select_if(is.numeric)%>%
  factanal(factors = 2,rotation = ’promax’,scores = ’none’)->fa_p
fa_p
```

```
##
## Call:
## factanal(x = ., factors = 2, scores = "none", rotation = "promax")
##
## Uniquenesses:
##          income      experience             age education_years        siblings
##           0.568           0.009           0.087           0.013           0.368
##
## Loadings:
##                 Factor1 Factor2
## income           0.280   0.545
## experience       1.006
## age              0.964
## education_years          1.005
## siblings                -0.807
##
##                Factor1 Factor2
## SS loadings      2.033   1.966
## Proportion Var   0.407   0.393
## Cumulative Var   0.407   0.800
##
## Factor Correlations:
##         Factor1 Factor2
## Factor1   1.000  -0.186
## Factor2  -0.186   1.000
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 1.94 on 1 degree of freedom.
## The p-value is 0.164
```

5. *Disscuss the differences between the two factor modelling approaches used in questions 1 and 4.* **(1 Mark)**

The approach used in question 4 was an oblique rotation compared to no rotation in question 1. The ‘promax’ rotation give us a more accurate observation. The result of the oblique rotation is a set of loadings that reflect the simple structure better. As seen from the loadings, there are more missing values for the promax rotation and the variables that do have loadings are highly correlated with the the factors. (i.e, Factor 1 could not be properly determined by the loadings with no rotation, however, with promax we can clearly see the only two variables that are strongly correlated with factor 1 are experience and age.)