學號:108753208 姓名:葉冠宏

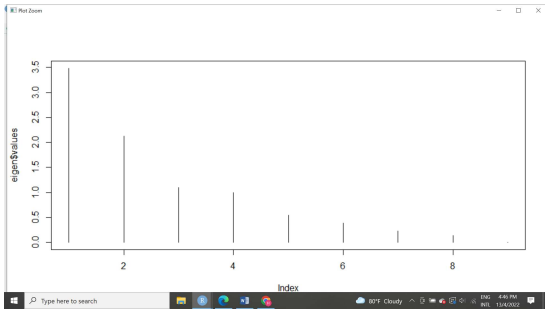
多變量分析期中考

1.

Q1:

我們執行 PCA，觀察每一個 component 的貢獻是如何。可以發現到如果只取前兩個 component，則其 cumulative variance 的貢獻為 62%。但如果取至 5 個 component，對 cumulative variance 的貢獻則可以達到 91%。

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5     Comp.6     Comp.7     Comp.8
Standard deviation     1.8673916 1.4595113 1.0483118 0.9972377 0.73703306 0.61921536 0.47513583 0.36985122
Proportion of Variance 0.3874613 0.2366859 0.1221064 0.1104981 0.06035753 0.04260307 0.02508378 0.01519888
Cumulative Proportion  0.3874613 0.6241472 0.7462536 0.8567517 0.91710919 0.95971227 0.98479605 0.99999493
                           Comp.9
Standard deviation     6.754636e-03
Proportion of Variance 5.069456e-06
Cumulative Proportion  1.000000e+00
```



接著，我們去把 data 執行 permutation test ，set p-value=0.05，發現只有前兩個 component 是顯著(看下圖 pval)。我們可以看到

Component1=0.524 Arg -0.347 Man -0.256 PS-0.325 Con-0.379 SI -0.387 SPS-0.367 TC

Component2=0.618 Min+0.355 Man+0.261 PS-0.350 SI-0.454 Fin-0.222 SPS+0.203 TC

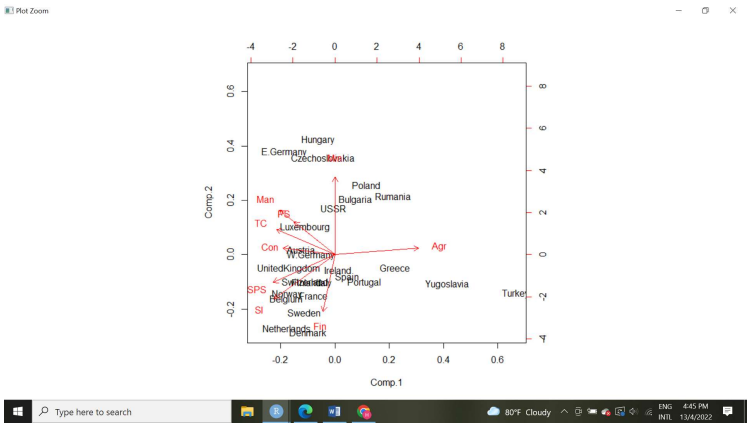
可以觀察到 component1 詮釋的比較是各個國家綜觀的產業結構，含農業、工業、服務業等。而 component2 的組成則似乎更聚焦於營造業、製造業有關的加成。

但如果考量到 cumulative proportion，如果取到第四個 component，則可以解釋約 85%的 variance。因此我們還是傾向於取四個 component。

```
$pve
     Comp.1      Comp.2      Comp.3      Comp.4      Comp.5      Comp.6      Comp.7      Comp.8
3.874613e-01 2.366859e-01 1.221064e-01 1.104981e-01 6.035753e-02 4.260307e-02 2.508378e-02 1.519888e-02
     Comp.9
5.069456e-06

$pval
[1] 0.000 0.000 0.996 0.897 1.000 1.000 1.000 1.000 1.000


Loadings:
     Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
Agr   0.524                       0.213  0.153                0.806
Min          0.618 -0.201        -0.164 -0.101 -0.726
Man  -0.347  0.355 -0.150 -0.346 -0.385 -0.288  0.479  0.126  0.366
PS   -0.256  0.261 -0.561  0.393  0.295  0.357  0.256 -0.341
Con  -0.325         0.153 -0.668  0.472  0.130 -0.221 -0.356
SI   -0.379 -0.350 -0.115        -0.284  0.615 -0.229  0.388  0.238
Fin         -0.454 -0.587         0.280 -0.526 -0.187  0.174  0.145
SPS  -0.387 -0.222  0.312  0.412 -0.220 -0.263 -0.191 -0.506  0.351
TC   -0.367  0.203  0.375  0.314  0.513 -0.124         0.545

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var  0.111  0.111  0.111  0.111  0.111  0.111  0.111  0.111  0.111
Cumulative Var  0.111  0.222  0.333  0.444  0.556  0.667  0.778  0.889  1.000
```

我們取 component 1 和 component 2 去做 biplot。我們可以發現 Greece, Turkey…等的國家產業比較傾向於農業，而 USSR, Poland, Hungary, East Germany, Rumania…等前蘇聯國家其產業比較

傾向於 Mining, Manufacturing 等重工業。至於西歐國家，英國、瑞典、荷蘭、丹麥…等國家其產業比較偏向於服務業，例如:金融業、個人服務業等。



2.

Q2:

首先，我們用 Mardia's Test 去 check the assumption of “multivariate

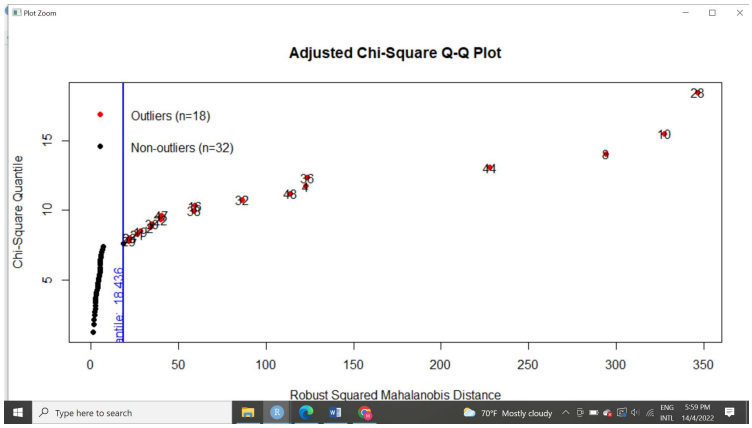
normality” for all variables，發現 data 並沒有符合 multivariate

normality，可以發現 test 的 result 都呈現 No。

```
$multivariateNormality
           Test         Statistic              p value Result
1 Mardia Skewness 138.478863289408 0.000169934283751673     NO
2 Mardia Kurtosis 2.53225049191323   0.0113332995941031     NO
3             MVN              <NA>                 <NA>     NO

$univariateNormality
          Test  Variable Statistic   p value Normality
1 Shapiro-Wilk    V1        0.9401    0.0135    NO
2 Shapiro-Wilk    V2        0.9557    0.0589    YES
3 Shapiro-Wilk    V3        0.9781    0.4732    YES
4 Shapiro-Wilk    V4        0.9604    0.0921    YES
5 Shapiro-Wilk    V5        0.9723    0.2870    YES
6 Shapiro-Wilk    V6        0.9482    0.0287    NO
7 Shapiro-Wilk    V7        0.9768    0.4255    YES

$Descriptives
    n     Mean   Std.Dev Median  Min   Max   25th   75th        Skew   Kurtosis
V1 50  98.836  7.337345 100.65 81.5 110.8 93.550 105.05 -0.55165506 -0.7234453
V2 50 106.622 10.124315 106.25 87.3 122.3 99.500 114.75 -0.06282027 -1.2015880
V3 50 102.810  4.712218 103.15 94.3 115.3 99.075 106.45  0.21227040 -0.4979168
V4 50  11.220  3.950149  10.00  1.0  18.0  8.250  14.00  0.05576043 -0.4944719
V5 50  14.180  3.384780  15.00  5.0  20.0 12.000  17.00 -0.31586390 -0.4007025
V6 50  10.560  2.139617  11.00  5.0  15.0  9.000  12.00 -0.49747110 -0.1726375
V7 50  29.760 10.537707  31.50  9.0  51.0 21.500  37.00 -0.10100835 -0.8971707
```

接著，我們看是否有 outlier，結果如下圖，共有 18 個 outlier。於是我們移除掉那些 outliers，再做一次 Mardia's Test，結果 new data 有符合 assumption of “multivariate

normality”，可以發現 test 的 result 都呈現 YES。




```
$multivariateNormality
             Test         Statistic           p value Result
1 Mardia Skewness  81.0033745310097 0.572359604838906    YES
2 Mardia Kurtosis -1.25941203682144 0.207881544021272    YES
3             MVN              <NA>              <NA>    YES

$univariateNormality
          Test  Variable Statistic   p value Normality
1 Shapiro-Wilk    V1        0.9229    0.0249    NO
2 Shapiro-Wilk    V2        0.9674    0.4311    YES
3 Shapiro-Wilk    V3        0.9496    0.1405    YES
4 Shapiro-Wilk    V4        0.9293    0.0375    NO
5 Shapiro-Wilk    V5        0.9562    0.2150    YES
6 Shapiro-Wilk    V6        0.8906    0.0036    NO
7 Shapiro-Wilk    V7        0.9571    0.2282    YES

$Descriptives
    n      Mean  Std.Dev Median  Min   Max    25th   75th        Skew    Kurtosis
V1 32 100.70937 5.023516 102.25 92.0 108.3  95.725 104.55 -0.30277958 -1.33996268
V2 32 108.73125 7.883238 108.80 94.5 122.3 102.375 113.85  0.07625644 -1.09073378
V3 32 103.80000 3.899959 103.65 97.8 110.8 100.175 106.80  0.13920941 -1.14886241
V4 32  12.37500 3.498848  12.00  7.0  18.0  10.000  15.25  0.26073484 -1.26445834
V5 32  15.15625 2.713281  15.00 10.0  20.0  12.750  17.00 -0.11279724 -0.86163299
V6 32  10.56250 1.740180  11.00  6.0  13.0   9.750  12.00 -0.86645442 -0.09885294
V7 32  31.31250 7.045646  32.00 18.0  43.0  25.750  37.00 -0.26962463 -1.06588447
```

我們開始去做 Canonical Correlation Analysis。我們把 Sales growth, Sales profitability, New account sales 放在 first group，令其為 U 變數。Creativity test, Mechanical reasoning test ,Abstract reasoning test ,Mathematics test 放在 second group ，令其為 V 變數。發現 U1, V1 ，即所有係數組合中有最大的 correlation 的組合，其 correlation 為 0.9996328。而 U2,V2，即有第二大的 correlation 的組合，其 correlation 為 0.9628392。

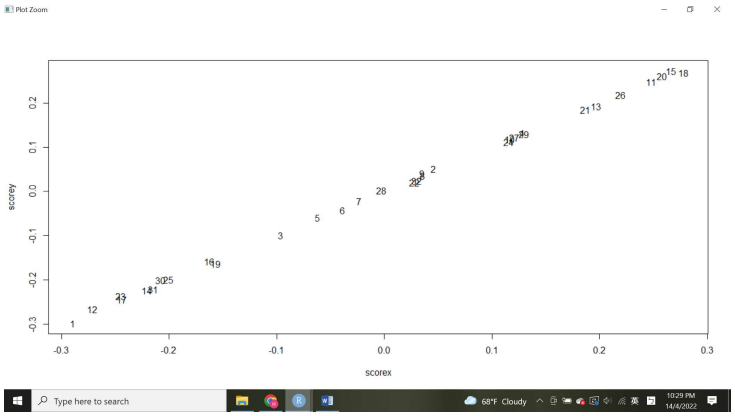
接著我們看到 xcoef, ycoef，得知 U1= 0.07425*Sales growth+ 0.01787*Sales profitability+ 0.09517*New account sales。而 V1=0.06573* Creativity test+ 0.01592*Mechanical reasoning test +0.04469*Abstract reasoning test+ 0.11015*Mathematics test。

```
$cor
[1] 0.9996328 0.9628392 0.8123569

$xcoef
           [,1]       [,2]       [,3]
[1,] 0.07425533  0.1916537 -0.4629000
[2,] 0.01787914 -0.4504927  0.2229265
[3,] 0.09517193  0.2450483  0.2431576

$ycoef
           [,1]       [,2]        [,3]         [,4]
[1,] 0.06573757  0.1530667  0.14836496  0.001607011
[2,] 0.01592337 -0.1335848  0.02416174 -0.149347801
[3,] 0.04469725  0.1656971 -0.09010886 -0.094542051
[4,] 0.11015724 -0.1465739 -0.06723228  0.133340375

$xcenter
[1] -8.673617e-19 -1.994932e-17  3.470802e-17

$ycenter
[1]  1.084202e-17 -8.239937e-18 -1.214306e-17  1.127570e-17
```

我們去執行 Rao's F，可以看到如果設 significance level 為 0.05，則三個 canonical correlations(U1,V1),(U2,V2),(U3,V3)都是 significant 的。

```
Wilks' Lambda, using F-approximation (Rao's F):
                stat     approx df1     df2      p.value
1 to 3:  1.821193e-05 1043.85644   9 107.235 0.000000e+00
2 to 3:  2.480537e-02  120.35969   4  90.000 0.000000e+00
3 to 3:  3.400762e-01   89.26379   1  46.000 2.399636e-12
```

我們去把 U1 和 V1 變數畫成 regression，呈現結果如下，呈現高度相關。



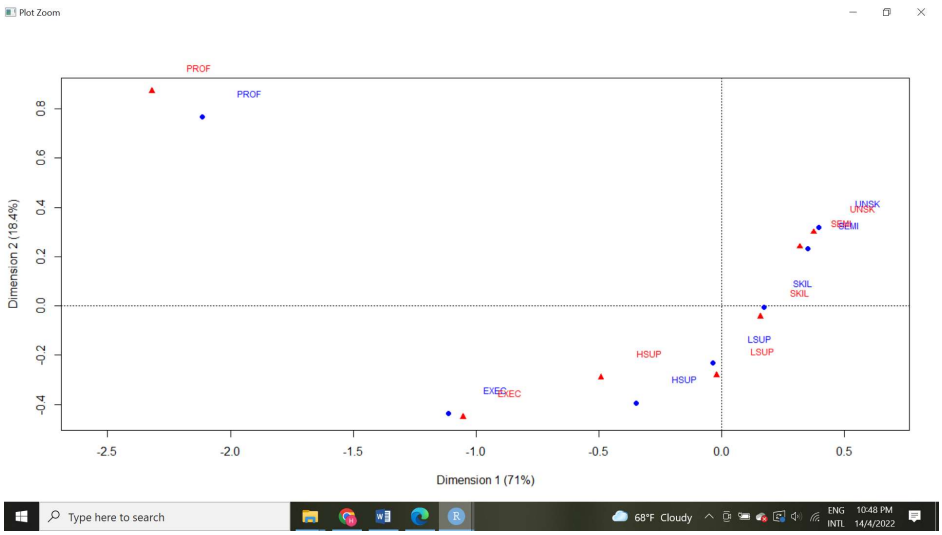
3.

Q3:

我們去做 Simple Correspondence Analysis，得知 the first dimension explains 70.95% of the total inertia, 而 the 2nd dimension explains 18.37% of the total inertia.

```
 Principal inertias (eigenvalues):
           1        2        3        4        5        6
Value      0.276286 0.071521 0.027298 0.008999 0.004459 0.000841
Percentage 70.95%   18.37%   7.01%    2.31%    1.15%    0.22%


 Rows:
             PROF      EXEC      HSUP      LSUP      SKIL      SEMI      UNSK
Mass     0.036889  0.042894  0.098656  0.148127  0.431799 0.130969 0.110666
ChiDist  2.258286  1.314930  0.538738  0.315517  0.196717 0.476462 0.580472
Inertia  0.188127  0.074165  0.028634  0.014746  0.016709 0.029732 0.037289
Dim. 1  -4.020749 -2.112958 -0.660239 -0.063348  0.329794 0.667769 0.755530
Dim. 2   2.870021 -1.638986 -1.474111 -0.865061 -0.017029 0.865303 1.192999


 Columns:
             PROF      EXEC      HSUP      LSUP      SKIL      SEMI      UNSK
Mass     0.029454  0.045468  0.094367  0.131255  0.408636 0.169574 0.121247
ChiDist  2.491591  1.262278  0.579226  0.345704  0.201781 0.442813 0.551462
Inertia  0.182850  0.072446  0.031660  0.015686  0.016638 0.033251 0.036872
Dim. 1  -4.411679 -1.999905 -0.934570 -0.036656  0.303974 0.606559 0.715927
Dim. 2   3.261269 -1.687748 -1.086582 -1.053766 -0.163358 0.900114 1.118780
```

接著，我們把 row variable 和 column variable 投影到 dimension1 和 dimension2。我們可以觀察到 managerial and executive 、high supervisory 主要集中在 dimension1 的中間和 dimension2 的低值。professional and high administrative 位在 dimension2 的高值和 dimension1 的低值。其餘的職業技能類別則大約在 dimension1 的高值和 dimension2 的中間值。而 father 從事的 occupation 大約和兒子從事的 occupation 是差不多的，因為 row variable 的投影和相同職業 column variable 的投影位置大概在差不多的位置。



4.

Q4:

MCA using indicator matrix:=> 2-dimension explains 24.2% of inertia

```
> mammals.mca<-mjca(dt, nd=2, lambda="indicator")
> summary(mammals.mca)

Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.541698  14.0  14.0  ****
 2      0.397120  10.2  24.2  ***
```

MCA using Burt Table:=> 2-dimension explains 43.2% of inertia

```
> mammals.burt<-mjca(dt, nd=2, lambda="Burt")
> #print(mammals.mca)
>
> summary(mammals.burt)

Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.293437  28.1  28.1  *******
 2      0.157704  15.1  43.2  ****
```

Adjusted MCA:=> 2-dimension explains 50.6% of inertia

```
> mammals.adj<-mjca(dt, nd=2, lambda="adjusted")
> #print(mammals.mca)
> summary(mammals.adj)

Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.226792  35.5  35.5  ***********
 2      0.096717  15.1  50.6  *****
```

Joint Correspondence Analysis:=> 2-dimension explains 52.6% of inertia

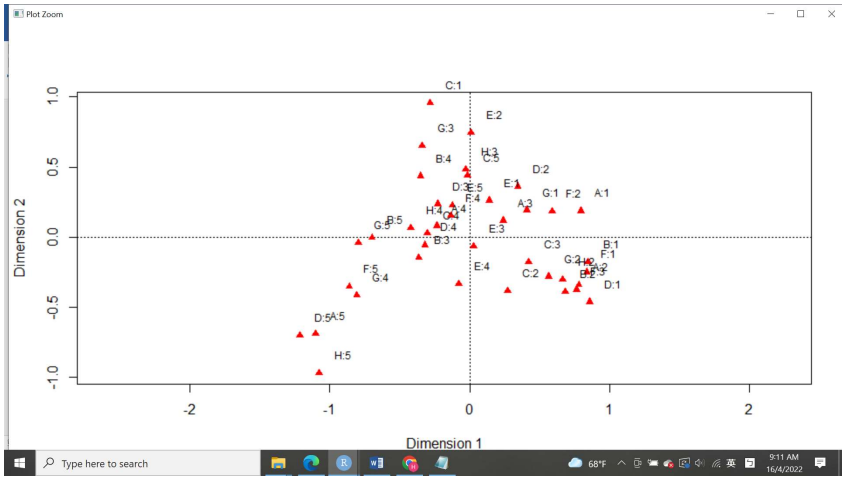
```
Diagonal inertia discounted from eigenvalues: 0.0580428
Percentage explained by JCA in 2 dimensions: 52.6%
(Eigenvalues are not nested)
[Iterations in JCA: 12 , epsilon = 7.98e-05]
```

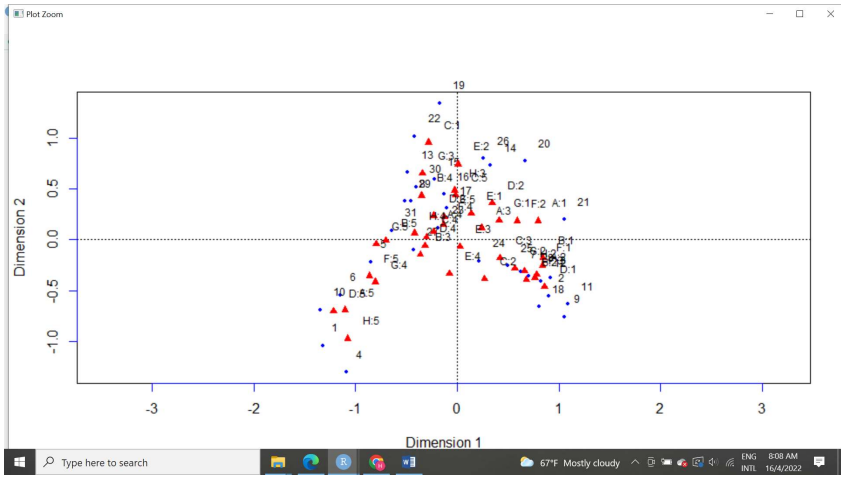
我們可以發現% of inertia explained by 2-dimension 隨著 methods 的變換越來越高。JCA 有最高的 %of inertia explained.

解釋:

2D-plot of JCA:

我們可以從圖中看出回答 disagree 的偏向 dimension 1 的右半邊(正越多的)，而回答 agree 的較傾向於 dimension1 的左半邊(負越多的)。而對於 dimension2 來說，越上面(正越多)的越傾向於 disagree，越下面(負越多)的越 agree。而我們綜觀所有 student 回答的分布，第 9, 11, 18,19 大家似乎更傾向於回答 disagree，而第 1,4, 5,6 ,10 題大家似乎更傾向於回答 agree。可以觀察得到對於課程評價的問題比較傾向於正面評價，或是評估自我感覺比較有學到東西的問題，學生的回答比較傾向於 agree，而對課程評價較負面或是自我學習評估比較沒學到東西的問題比較傾向於回答 disagree。可見這堂課應該教的還不錯。





Q5:

根據上圖，以及 Q4 的分析，可見大家回答 agree 或是 disagree 的回答都還蠻 consistent 的，因為可以看到針對分布在右半邊的某個問題，有許多 student 都回答 strongly disagree 所代表的 1，而針對左半邊的問題，大家都比較傾向於回答 5 所代表的 strongly agree。如 Q4 提到的，針對課程或是自我學習評估較負面表述的問題，大家會比較 consistent 的回答 disagree，而對於持較正面表述的問題，大家會更傾向於回答 agree。

5.

Q6:

我們可以看到，經過 rotation 後，取兩個 factor，其可以解釋約 52.8% 的 data(從 cumulative var 那邊看)。而對於個別的變數，對 Percentage employed in power supply industries 可解釋約 81.5%的 data，對 Percentage employed in finance 可解釋約 79%的 data，都蠻高的。但對於

Agr(Percentage employed in agriculture), Man(Percentage employed in manufacturing), SPS(Percentage employed in social and personal services)的解釋能力則較弱，僅分別有 0.5%, 0.5%, 22.7%而已。而 factor1 和 factor2 都有 Arg 的變數。

如果降低 factor 的數量為 1，雖然對於 Man, SPS 的解釋能力會有顯著的上升，但對於整體的解釋能力降低至 37.3%。

而如果我們取 3 個 factor，雖然對於整體的解釋能力可以上升至 64.4%，但對於 PS 的解釋能力降至 0.5%。而且即便設 cutoff=0.4，factor1 和 factor2 還是都有 Agr 變數。

在沒有 rotate 之前，其可以解釋約 54.5%的 data，比 rotate 後高了一點，但即使設 cutoff=0.4，Factor1 和 Factor2 都有 Man, SI,SPS 的變數，會使解釋上比較困難，因此雖然這樣會失去 orthogonal 的性質，我們還是選擇 rotate 後的結果。

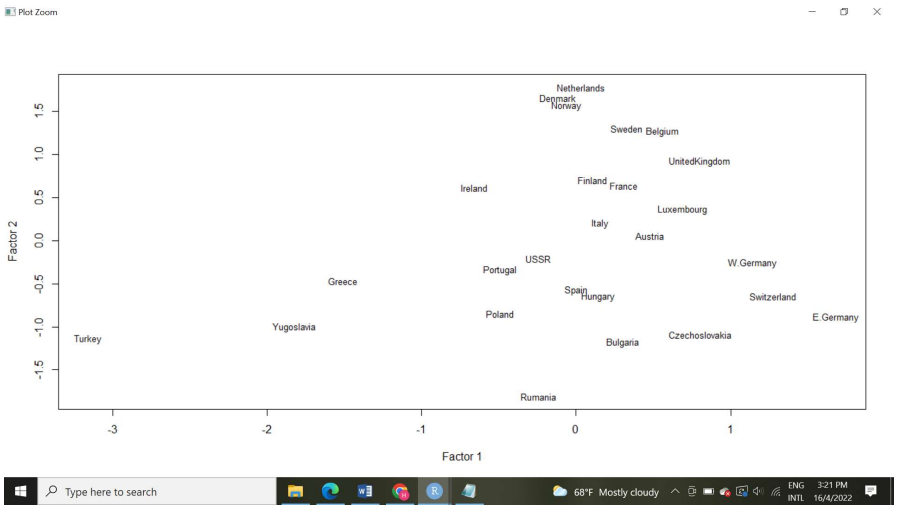
考量到上述的因素，如果我們 rotate，取兩個 factor，並設置 cutoff =0.4，則

factor1=-0.807*Agr+1.003*Man+0.424*PS+0.548*Con+0.435*TC

factor2=-0.444*Agr-0.604*Min+0.688*SI+0.467*Fin+0.762*SPS。

可以觀察到 factor1 似乎比較和營造業、工業有關，而 factor2 比較和服務業有關。因此從 Biplot 中可以觀察到荷蘭、丹麥、挪威、英國、法國等發達國家除了工業很發達之外，其產業也有高度的服務業，因此整體的分布偏向於右上。而波蘭、保加利亞、東德等前蘇聯國家，其工業、營造業的比重比服務業等高出了許多。至於希臘、土耳其等國，其工業化程度和服務業發達程度遠不如西歐等發達國家，位於 biplot 的左下角。

```
Call:
factanal(x = sub, factors = 2, scores = "Bartlett", rotation = "promax")

Uniquenesses:
  Agr   Min   Man    PS   Con    SI   Fin   SPS    TC
0.005 0.590 0.005 0.815 0.679 0.307 0.790 0.227 0.681

Loadings:
    Factor1 Factor2
Agr -0.807  -0.444
Min         -0.604
Man  1.003
PS   0.424
Con  0.548
SI           0.688
Fin          0.467
SPS          0.762
TC   0.435

               Factor1 Factor2
SS loadings      2.689   2.061
Proportion Var   0.299   0.229
Cumulative Var   0.299   0.528

Factor Correlations:
        Factor1 Factor2
Factor1   1.000   0.204
Factor2   0.204   1.000

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 158.11 on 19 degrees of freedom.
The p-value is 5.91e-24
```



Q7:

根據 Q1，

Component1=0.524 Arg -0.347 Man -0.256 PS-0.325 Con-0.379 SI -0.387 SPS-0.367 TC

Component2=0.618 Min+0.355 Man+0.261 PS-0.350 SI-0.454 Fin-0.222 SPS+0.203 TC

component1 詮釋的比較是各個國家綜觀的產業結構，含農業、工業、服務業等。而 component2 的組成則似乎更聚焦於營造業、製造業有關的加成。

根據 Q6，

factor1=-0.807*Agr+1.003*Man+0.424*PS+0.548*Con+0.435*TC

factor2=-0.444*Agr-0.604*Min+0.688*SI+0.467*Fin+0.762*SPS

factor1 似乎比較和營造業、工業有關，而 factor2 比較和服務業有關。

=>可以觀察到在 PCA 的分析中，雖然我們可以稍微看出 component1 和 component2 各自詮釋的意義是什麼，例如: component1 較著重綜觀每個 variable 都考量的解釋。但當中還是有許多 variable 是兩個 component 都有的，在解釋能力上比 factor analysis 還弱。

而 factor analysis 可以明確分出 factor1 和 factor2 各自傾向的產業解釋是什麼。但同樣是取兩個維度，PCA 可以解釋約 62%的 variance，但 factor analysis 僅能解釋約 52.8% 的 variance，比 PCA 稍微弱一點。可見兩個分析工具所關心的統計分析重點並不一樣。