Keiland Pullen
Homework 4



**1.)** The article is "Hospital Image: A Correspondence Analysis Approach" borrowed from the "Journal of Health Care Marketing.

**a.)** In the article, the author states that the data was composed of "yes/no" answers from the questionnaire. It states "This information, whether a hospital is associated with a feature or not (binary data), became the input to correspondence analysis." For this case, yes, correspondence analysis (CA) is appropriate as they were able to look at the relationship between two groups of variables – hospitals and features. The purpose of the research was for the hospitals to get a better understanding of how their various hospital services measured up against their patients view of the hospitals. This is demonstrated in the contingency table displayed in the article.

**b.)** The research identifies 13 hospital feature variables such as "cancer treatment", "laser treatment", and "women's health services". Other variables are "expert emergency treatment", "heart disease prevention and treatment", "rehabilitation services", "call-in health information services", and several others. The objects were identified as the 16 different hospitals. Each of the variables is a "Yes" or "No" categorical variable.

**c.)** The article features 2 tables and 1 image map. The image map is used as a two-dimensional display to how both dimensions Factor 1 and Factor 2 correspond to the principal components of the data. From the article, "From this image map, which shows the correspondence between the 13 features and the 16 hospitals, one, can glean interesting information." For example, the image shows that the "Cleveland Clinic" hospital is more associated with "heart disease", "cancer treatment" and "technological equipment". The "Cleveland Clinic" is not associated with or maybe loosely associated with "community programs" and "programs for seniors". The map allowed the hospital to compare their current skills with competing hospitals and allows them to look at what skills may need development.

**d.)** In regard to evaluating the goodness of fit for the model, there didn't seem to be any noticeable discussion. The article discussed the results of the questionnaire and how those were interpreted using the tables and contingency table. The authors did mention the frequency of 2 variables, but should have included more information along with a Chi-Square test. More discussion on the goodness of fit would have added more reliability and credibility to the research and could close any gaps regarding misinterpretation.

**e.)** Using corresponding analysis (CA), the researchers state that they were able to "visualize their hospitals' comparative advantages and disadvantages in relation to their competitors' positions of strength and weakness." The article also states that CA allowed the hospital system to develop strategies to identify the strengths of their clinical programs. This allowed them to market those programs to a larger audience. Furthermore, CA allowed the hospital to create defensive strategies to improved their list of services, that were strongly associated with other competing hospitals.

**f.)** For this article, it seems that the researchers were able to draw a number of positive conclusions that should ultimately help the hospital develop and improve its services. The main issue that I have are the values of the variables. The values are only "Yes" and "Not". If this is a hospital questionnaire, my preference for data would from the use of a Likert Scale. I believe that this would have allowed the hospital to gain a greater insight into the data. At most, maybe one-half of the variables could have been "Yes/No". One item that I found missing was the Chi-Square test. I'm not sure if this could have been used, but it should have at least been mentioned. Also, I believe that a correlation matrix could have been useful. Overall, there are a few techniques that could have been included that would have added more validity to the research, unless they were purposely left out.

Keiland Pullen DSC 424 | Fall 2021
Homework 4

**2.)**

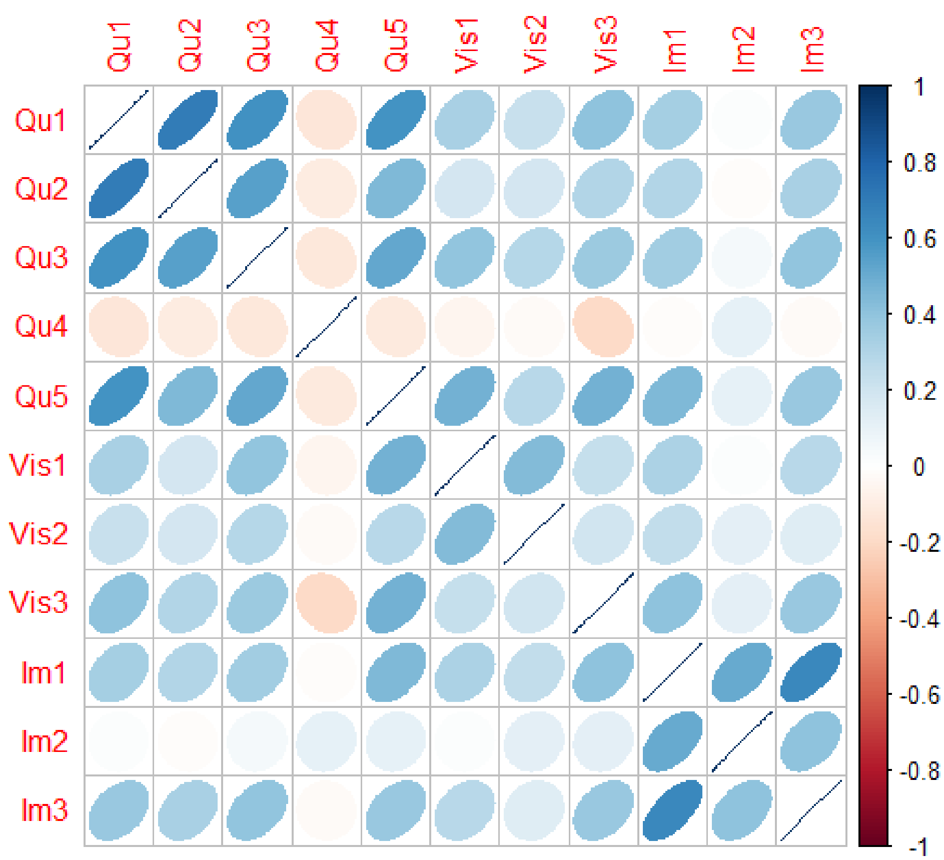
```
setwd("C:/Users/Home/Desktop/DePaul/DSC-424-AdvancedDataAnalysis/week-5/Homework")
library(corrplot)
ds = read.table("Survey.csv", sep=",", header=T)
```

**a.)**

Pearson:

```
> c = cor(ds)
> print(c)
>corrplot(c, method="ellipse")
```

| | Qu1 | Qu2 | Qu3 | Qu4 | Qu5 | Vis1 | Vis2 | Vis3 | Im1 | Im2 | Im3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Qu1 | 1.00000000 | 0.69056585 | 0.6045452 | -0.13423485 | 0.5938894 | 0.32105528 | 0.22360135 | 0.4045556 | 0.33072146 | 0.01198666 | 0.37003681 |
| Qu2 | 0.69056585 | 1.00000000 | 0.5438652 | -0.10242737 | 0.4426255 | 0.18909481 | 0.18383070 | 0.2940380 | 0.29357109 | -0.01458967 | 0.32722094 |
| Qu3 | 0.60454520 | 0.54386525 | 1.0000000 | -0.12643342 | 0.5191393 | 0.39564033 | 0.28017242 | 0.3673545 | 0.34073109 | 0.04966370 | 0.39478640 |
| Qu4 | -0.13423485 | -0.10242737 | -0.1264334 | 1.00000000 | -0.1142571 | -0.05060630 | -0.02370919 | -0.1908196 | -0.01731659 | 0.10751462 | -0.02454507 |
| Qu5 | 0.59388936 | 0.44262554 | 0.5191393 | -0.11425710 | 1.0000000 | 0.47341640 | 0.27708105 | 0.4747601 | 0.44354852 | 0.10102656 | 0.37941833 |
| Vis1 | 0.32105528 | 0.18909481 | 0.3956403 | -0.05060630 | 0.4734164 | 1.00000000 | 0.43011806 | 0.2334773 | 0.31276682 | 0.01343525 | 0.27387663 |
| Vis2 | 0.22360135 | 0.18383070 | 0.2801724 | -0.02370919 | 0.2770811 | 0.43011806 | 1.00000000 | 0.1938247 | 0.24335827 | 0.11873267 | 0.13160090 |
| Vis3 | 0.40455560 | 0.29403799 | 0.3673545 | -0.19081962 | 0.4747601 | 0.23347727 | 0.19382473 | 1.0000000 | 0.40507794 | 0.11837845 | 0.37003731 |
| Im1 | 0.33072146 | 0.29357109 | 0.3407311 | -0.01731659 | 0.4435485 | 0.31276682 | 0.24335827 | 0.4050779 | 1.00000000 | 0.50239918 | 0.64631560 |
| Im2 | 0.01198666 | -0.01458967 | 0.0496637 | 0.10751462 | 0.1010266 | 0.01343525 | 0.11873267 | 0.1183785 | 0.50239918 | 1.00000000 | 0.40696969 |
| Im3 | 0.37003681 | 0.32722094 | 0.3947864 | -0.02454507 | 0.3794183 | 0.27387663 | 0.13160090 | 0.3700373 | 0.64631560 | 0.40696969 | 1.00000000 |



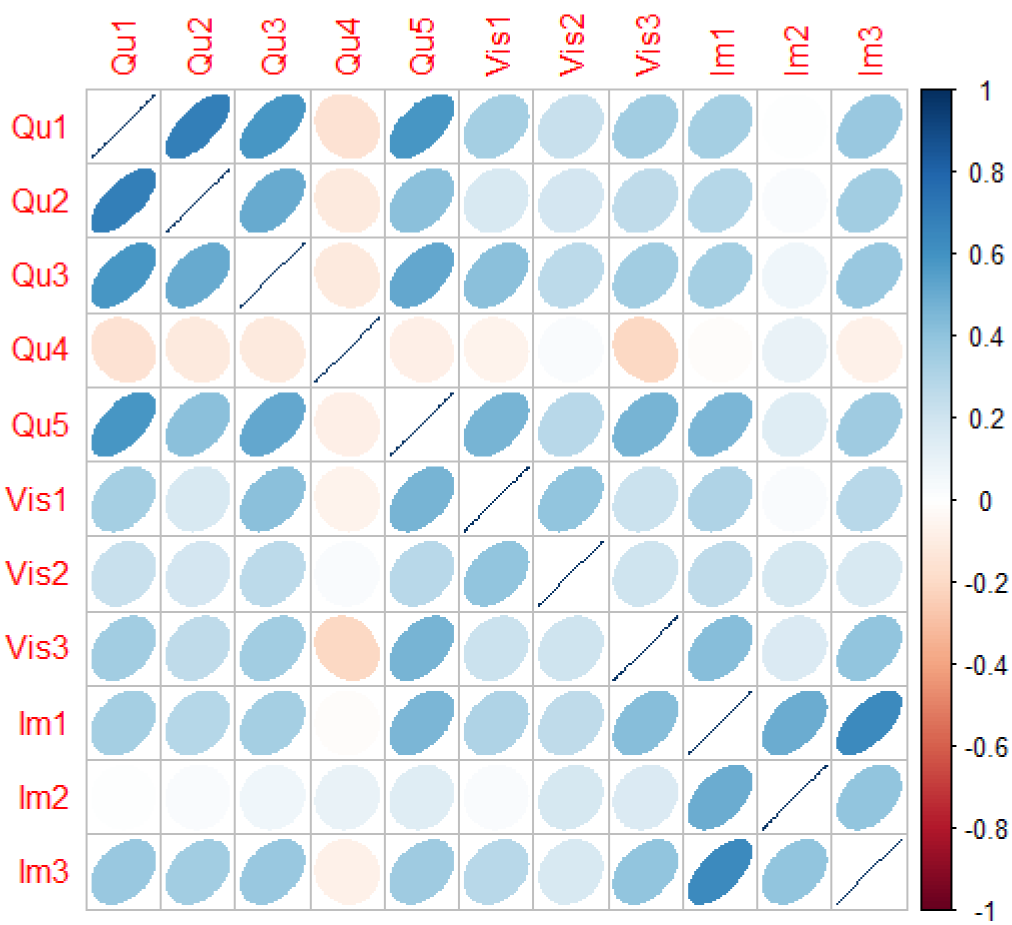
Keiland Pullen
Homework 4

DSC 424 | Fall 2021

Spearman

```
cS = cor(ds, method = "spearman")
print(cS)
corrplot(cS, method="ellipse")
```

| | Qu1 | Qu2 | Qu3 | Qu4 | Qu5 | Vis1 | Vis2 | Vis3 | Im1 | Im2 | Im3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Qu1 | 1.00000000 | 0.68509748 | 0.5878939 | -0.15157709 | 0.58292798 | 0.33347765 | 0.22178799 | 0.3459261 | 0.33000524 | 0.00234924 | 0.37027240 |
| Qu2 | 0.68509748 | 1.00000000 | 0.5000471 | -0.11572989 | 0.41042527 | 0.16656887 | 0.18625036 | 0.2540890 | 0.28848734 | 0.02371458 | 0.34092829 |
| Qu3 | 0.58789389 | 0.50004710 | 1.0000000 | -0.11666848 | 0.51933422 | 0.41825553 | 0.26847594 | 0.3469657 | 0.33864249 | 0.06842650 | 0.37859426 |
| Qu4 | -0.15157709 | -0.11572989 | -0.1166685 | 1.00000000 | -0.08708786 | -0.06675524 | 0.02550185 | -0.2087290 | -0.01613649 | 0.09051871 | -0.07473327 |
| Qu5 | 0.58292798 | 0.41042527 | 0.5193342 | -0.08708786 | 1.00000000 | 0.46154225 | 0.27545371 | 0.4617664 | 0.45909492 | 0.13662731 | 0.35818408 |
| Vis1 | 0.33347765 | 0.16656887 | 0.4182555 | -0.06675524 | 0.46154225 | 1.00000000 | 0.39566510 | 0.2125064 | 0.30678507 | 0.02205894 | 0.27216535 |
| Vis2 | 0.22178799 | 0.18625036 | 0.2684759 | 0.02550185 | 0.27545371 | 0.39566510 | 1.00000000 | 0.2094084 | 0.25985435 | 0.17221415 | 0.16738022 |
| Vis3 | 0.34592609 | 0.25408895 | 0.3469657 | -0.20872898 | 0.46176637 | 0.21250638 | 0.20940837 | 1.0000000 | 0.42102453 | 0.15940691 | 0.39638793 |
| Im1 | 0.33000524 | 0.28848734 | 0.3386425 | -0.01613649 | 0.45909492 | 0.30678507 | 0.25985435 | 0.4210245 | 1.00000000 | 0.49823021 | 0.63119066 |
| Im2 | 0.00234924 | 0.02371458 | 0.0684265 | 0.09051871 | 0.13662731 | 0.02205894 | 0.17221415 | 0.1594069 | 0.49823021 | 1.00000000 | 0.39779965 |
| Im3 | 0.37027240 | 0.34092829 | 0.3785943 | -0.07473327 | 0.35818408 | 0.27216535 | 0.16738022 | 0.3963879 | 0.63119066 | 0.39779965 | 1.00000000 |



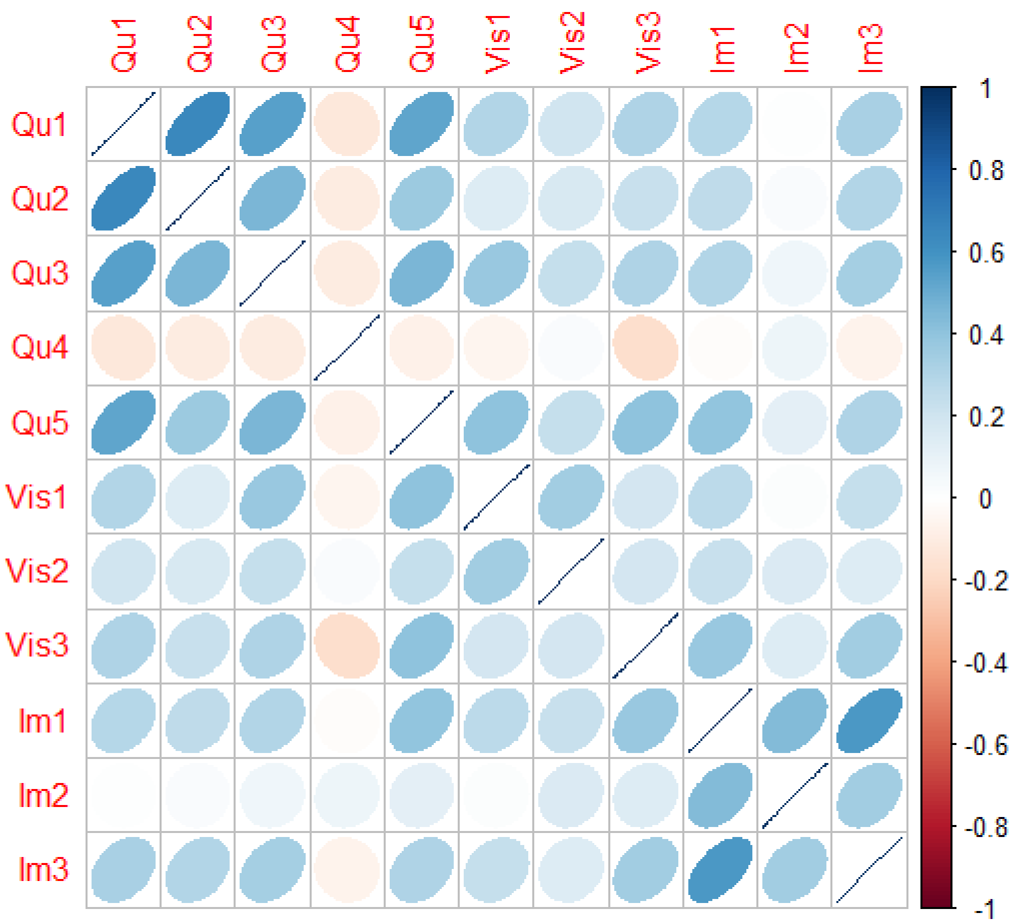
Keiland Pullen
Homework 4

DSC 424 | Fall 2021

Kendall Tau

```
k = cor(ds, method = "kendall")
print(k)
corrplot(k, method="ellipse")
```

| | Qu1 | Qu2 | Qu3 | Qu4 | Qu5 | Vis1 | Vis2 | Vis3 | Im1 | Im2 | Im3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Qu1 | 1.000000000 | 0.64461404 | 0.54725264 | -0.12921819 | 0.52782167 | 0.29897964 | 0.1970511 | 0.3074402 | 0.2885985 | 0.001646515 | 0.32951909 |
| Qu2 | 0.644614036 | 1.00000000 | 0.45811342 | -0.10273724 | 0.36475198 | 0.14551909 | 0.1627692 | 0.2236363 | 0.2563052 | 0.022736849 | 0.29679407 |
| Qu3 | 0.547252636 | 0.45811342 | 1.00000000 | -0.10527826 | 0.45912244 | 0.37056427 | 0.2383310 | 0.3080751 | 0.2974953 | 0.062326727 | 0.33230667 |
| Qu4 | -0.129218195 | -0.10273724 | -0.10527826 | 1.00000000 | -0.07370463 | -0.05865705 | 0.0226063 | -0.1767891 | -0.0142836 | 0.076512682 | -0.06944346 |
| Qu5 | 0.527821675 | 0.36475198 | 0.45912244 | -0.07370463 | 1.00000000 | 0.40132047 | 0.2384228 | 0.4038848 | 0.3982152 | 0.116576277 | 0.30689697 |
| Vis1 | 0.298979641 | 0.14551909 | 0.37056427 | -0.05865705 | 0.40132047 | 1.00000000 | 0.3499661 | 0.1842321 | 0.2663199 | 0.018058772 | 0.23331690 |
| Vis2 | 0.197051102 | 0.16276922 | 0.23833103 | 0.02260630 | 0.23842279 | 0.34996606 | 1.0000000 | 0.1813221 | 0.2237267 | 0.152578723 | 0.14157197 |
| Vis3 | 0.307440178 | 0.22363632 | 0.30807507 | -0.17678909 | 0.40388479 | 0.18423207 | 0.1813221 | 1.0000000 | 0.3741080 | 0.140684633 | 0.34014762 |
| Im1 | 0.288598533 | 0.25630519 | 0.29749533 | -0.01428360 | 0.39821524 | 0.26631994 | 0.2237267 | 0.3741080 | 1.0000000 | 0.431886688 | 0.57035106 |
| Im2 | 0.001646515 | 0.02273685 | 0.06232673 | 0.07651268 | 0.11657628 | 0.01805877 | 0.1525787 | 0.1406846 | 0.4318867 | 1.000000000 | 0.34587874 |
| Im3 | 0.329519086 | 0.29679407 | 0.33230667 | -0.06944346 | 0.30689697 | 0.23331690 | 0.1415720 | 0.3401476 | 0.5703511 | 0.345878736 | 1.00000000 |



The following is the range for each correlation:

```
> range(k) # Kendall Tau
[1] -0.1767891  1.0000000
> range(c) # Pearson
[1] -0.1908196  1.0000000
> range(cS) # Spearman
[1] -0.208729  1.000000
```

The following displays the max and min of subtracting the matrices from each other.

```
> max(cS - c)  # Spearman - Pearson
[1] 0.05348148
> min(cS - c)  # Spearman - Pearson
[1] -0.05862952
> max(k - c)  # Kendall - Pearson
[1] 0.04631548
> min(k - c)  # Kendall - Pearson
[1] -0.09711543
```

The above demonstrate that the differences between the Matrices are rather small.

**b.)**

KMO test on Pearson

```
> KMO(c)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = c)
Overall MSA =  0.82
MSA for each item =
 Qu1  Qu2  Qu3  Qu4  Qu5 Vis1 Vis2 Vis3  Im1  Im2  Im3
0.82  0.79   0.91  0.75  0.88 0.76 0.77 0.89 0.79 0.65 0.83
```

The Overall KMO value of 0.82 along with variable values of 0.6 suggests that we can proceed with factor analysis.

**c.)**

```
p = prcomp(cor(ds, method = "spearman"))
summary(p)
```

Importance of components:

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | PC10 | PC11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Standard deviation | 0.6571 | 0.4339 | 0.3288 | 0.23042 | 0.22013 | 0.16332 | 0.15158 | 0.13413 | 0.10079 | 0.08371 | 2.874e-17 |
| Proportion of Variance | 0.4721 | 0.2059 | 0.1182 | 0.05805 | 0.05298 | 0.02917 | 0.02512 | 0.01967 | 0.01111 | 0.00766 | 0.000e+00 |
| Cumulative Proportion | 0.4721 | 0.6780 | 0.7962 | 0.85429 | 0.90727 | 0.93644 | 0.96156 | 0.98123 | 0.99234 | 1.00000 | 1.000e+00 |

Using the above PCA, I would use 4 factors, because it is at PC4 where we see that 85% of the variance is accumulated. It possible to use 3 factors because PC3 is just a hair under 80%.

**d.)**

```
p2 = principal(cor(ds, method = "spearman"), nfactors=4)
summary(p2)
```

Factor analysis with Call: principal(r = cor(ds, method = "spearman"), nfactors = 4)

Test of the hypothesis that 4 factors are sufficient.
The degrees of freedom for the model is 17 and the objective function was 1.09

The root mean square of the residuals (RMSA) is 0.08

**e.)**

Loadings:

| | RC1 | RC2 | RC3 | RC4 |
|---|---|---|---|---|
| Qu1 | 0.870 | | | |
| Qu2 | 0.868 | | | |
| Qu3 | 0.696 | | | |
| Qu4 | | | | -0.912 |
| Qu5 | 0.574 | | 0.432 | |
| Vis1 | | | 0.807 | |
| Vis2 | | | 0.779 | |
| Vis3 | | 0.420 | | 0.509 |
| Im1 | | 0.796 | | |
| Im2 | | 0.832 | | |
| Im3 | | 0.725 | | |

| | RC1 | RC2 | RC3 | RC4 |
|---|---|---|---|---|
| SS loadings | 2.663 | 2.162 | 1.691 | 1.207 |
| Proportion Var | 0.242 | 0.197 | 0.154 | 0.110 |
| Cumulative Var | 0.242 | 0.439 | 0.592 | 0.702 |

There are several distinct groupings, however, the “Qu5” and “Vis3” variables are both contributors to 2 factor groups.

**f.)**

```
fit = factanal(ds, 4, scores="regression")
print(fit$loadings, cutoff=.4, sort=T)
print(fit)
```

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 7.62 on 17 degrees of freedom.
The p-value is 0.974

The above tells us that we can fail to reject this hypothesis

The following RMSEA was reported in the following:

```
> hw.model = 'QU =~ Qu1 + Qu2 + Qu3 + Qu4 + Qu5
+           Vis =~ Vis1 + Vis2 + Vis3
+           IM =~ Im1 + Im2 + Im3'
> fit = cfa(hw.model, data=ds)
> summary(fit, fit.measures=TRUE)
lavaan 0.6-9 ended normally after 34 iterations

  Estimator                         ML
  Optimization method           NLMINB
  Number of model parameters        25

  Number of observations           119
```

```
Model Test User Model:

  Test statistic                                71.985
  Degrees of freedom                                41
  P-value (Chi-square)                           0.002

Model Test Baseline Model:

  Test statistic                               465.601
  Degrees of freedom                                55
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.925
  Tucker-Lewis Index (TLI)                       0.899

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -1540.300
  Loglikelihood unrestricted model (H1)      -1504.308

  Akaike (AIC)                                3130.601
  Bayesian (BIC)                              3200.079
  Sample-size adjusted Bayesian (BIC)         3121.044

Root Mean Square Error of Approximation:

  RMSEA                                          0.080
  90 Percent confidence interval - lower         0.048
  90 Percent confidence interval - upper         0.110
  P-value RMSEA <= 0.05                          0.060

Standardized Root Mean Square Residual:

  SRMR                                           0.080

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Expected
  Information saturated (h1) model          Structured

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  QU =~
    Qu1               1.000
    Qu2               0.904    0.107    8.449    0.000
    Qu3               0.884    0.104    8.497    0.000
    Qu4              -0.245    0.150   -1.635    0.102
    Qu5               1.054    0.128    8.238    0.000
  Vis =~
    Vis1              1.000
    Vis2              0.708    0.203    3.484    0.000
    Vis3              1.405    0.321    4.374    0.000
  IM =~
    Im1               1.000
    Im2               0.697    0.129    5.387    0.000
    Im3               0.861    0.122    7.040    0.000

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)
  QU ~~
    Vis               0.264    0.064    4.157    0.000
    IM                0.252    0.061    4.124    0.000
  Vis ~~
    IM                0.270    0.071    3.818    0.000

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .Qu1               0.159    0.035    4.569    0.000
```

```
 .Qu2          0.297    0.047    6.327    0.000
 .Qu3          0.278    0.044    6.298    0.000
 .Qu4          0.953    0.124    7.683    0.000
 .Qu5          0.440    0.068    6.448    0.000
 .Vis1         0.635    0.097    6.550    0.000
 .Vis2         0.591    0.083    7.152    0.000
 .Vis3         0.933    0.155    6.019    0.000
 .Im1          0.131    0.072    1.828    0.068
 .Im2          0.794    0.111    7.159    0.000
 .Im3          0.425    0.076    5.558    0.000
  QU           0.415    0.077    5.402    0.000
  Vis          0.251    0.095    2.640    0.008
  IM           0.634    0.120    5.287    0.000
```

**g.)**

```
library(polycor)

het = hetcor(ds)

ds$Qu1 = factor(ds$Qu1, levels = c(1,2,3,4,5), ordered = T)
ds$Qu2 = factor(ds$Qu2, levels = c(1,2,3,4,5), ordered = T)
ds$Qu3 = factor(ds$Qu3, levels = c(1,2,3,4,5), ordered = T)
ds$Qu4 = factor(ds$Qu4, levels = c(1,2,3,4,5), ordered = T)
ds$Qu5 = factor(ds$Qu5, levels = c(1,2,3,4,5), ordered = T)
ds$Vis1 = factor(ds$Vis1, levels = c(1,2,3,4,5), ordered = T)
ds$Vis2 = factor(ds$Vis2, levels = c(1,2,3,4,5), ordered = T)
ds$Vis3 = factor(ds$Vis3, levels = c(1,2,3,4,5), ordered = T)
ds$Im1 = factor(ds$Im1, levels = c(1,2,3,4,5) , ordered = T)
ds$Im2 = factor(ds$Im2, levels = c(1,2,3,4,5), ordered = T)
ds$Im3 = factor(ds$Im3, levels = c(1,2,3,4,5), ordered = T)

het = hetcor(ds)
summary(het)


hetCor = het$correlations
hetCor

phet = princomp(covmat = hetCor, cor=T)
summary(phet)

phet2 = principal(hetCor, nfactors=4)
summary(phet2)
print(phet2$loadings, cutoff=.4)
```

```
> hetCor = het$correlations
> hetCor
            Qu1         Qu2         Qu3           Qu4        Qu5        Vis1        Vis2       Vis3         Im1         Im2          Im3
Qu1   1.00000000  0.77558379  0.70743795 -0.1349949439  0.6794376  0.34816205  0.25887420  0.4713950 0.383893709  0.01330986 0.4190272802
Qu2   0.77558379  1.00000000  0.63731787 -0.1021300447  0.4871091  0.20350306  0.19206985  0.3151390 0.335052127 -0.02996818 0.3634029867
Qu3   0.70743795  0.63731787  1.00000000 -0.1291010967  0.5926328  0.44094215  0.31922546  0.4178973 0.391150258  0.03935248 0.4466261280
Qu4  -0.13499494 -0.10213004 -0.12910110  1.0000000000 -0.1187339 -0.03618364 -0.02073868 -0.2231590 0.001161944  0.12907978 0.0006384136
Qu5   0.67943765  0.48710914  0.59263281 -0.1187339327  1.0000000  0.53098019  0.31432708  0.5543858 0.494675605  0.11650656 0.4195876923
Vis1  0.34816205  0.20350306  0.44094215 -0.0361836418  0.5309802  1.00000000  0.47308340  0.2383451 0.360756287  0.01023501 0.3073081859
Vis2  0.25887420  0.19206985  0.31922546 -0.0207386814  0.3143271  0.47308340  1.00000000  0.2235509 0.280137397  0.12291945 0.1445080985
Vis3  0.47139500  0.31513897  0.41789728 -0.2231590115  0.5543858  0.23834512  0.22355087  1.0000000 0.473244283  0.13772079 0.4261632798
Im1   0.38389371  0.33505213  0.39115026  0.0011619439  0.4946756  0.36075629  0.28013740  0.4732443 1.000000000  0.56032915 0.7243453271
Im2   0.01330986 -0.02996818  0.03935248  0.1290797838  0.1165066  0.01023501  0.12291945  0.1377208 0.560329155  1.00000000 0.4594743343
Im3   0.41902728  0.36340299  0.44662613  0.0006384136  0.4195877  0.30730819  0.14450810  0.4261633 0.724345327  0.45947433 1.0000000000

> phet = princomp(covmat = hetCor, cor=T)
```

Keiland Pullen DSC 424 | Fall 2021
Homework 4

```
> summary(phet)
Importance of components:
                       Comp.1    Comp.2    Comp.3     Comp.4    Comp.5     Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     2.1302856 1.2856415 1.0724218 0.99940419 0.8083078 0.77029388 0.65258257 0.58738035 0.49561255 0.48235202
Proportion of Variance 0.4125561 0.1502613 0.1045535 0.09080079 0.0593965 0.05394115 0.03871491 0.03136506 0.02233016 0.02115122
Cumulative Proportion  0.4125561 0.5628174 0.6673709 0.75817167 0.8175682 0.87150932 0.91022423 0.94158929 0.96391946 0.98507068
                       Comp.11
Standard deviation     0.40524377
Proportion of Variance 0.01492932
Cumulative Proportion  1.00000000
> phet2 = principal(hetCor, nfactors=4)
> summary(phet2)

Factor analysis with Call: principal(r = hetCor, nfactors = 4)

Test of the hypothesis that 4 factors are sufficient.
The degrees of freedom for the model is 17  and the objective function was  1.25

The root mean square of the residuals (RMSA) is  0.07
> print(phet2$loadings, cutoff=.4)

Loadings:
     RC1    RC2    RC3    RC4
Qu1   0.897
Qu2   0.896
Qu3   0.782
Qu4                 -0.924
Qu5   0.630         0.422
Vis1         0.820
Vis2         0.831
Vis3  0.403                0.506
Im1          0.823
Im2          0.856
Im3          0.764

               RC1   RC2   RC3   RC4
SS loadings    3.105 2.288 1.752 1.195
Proportion Var 0.282 0.208 0.159 0.109
Cumulative Var 0.282 0.490 0.650 0.758
```

**3.)**

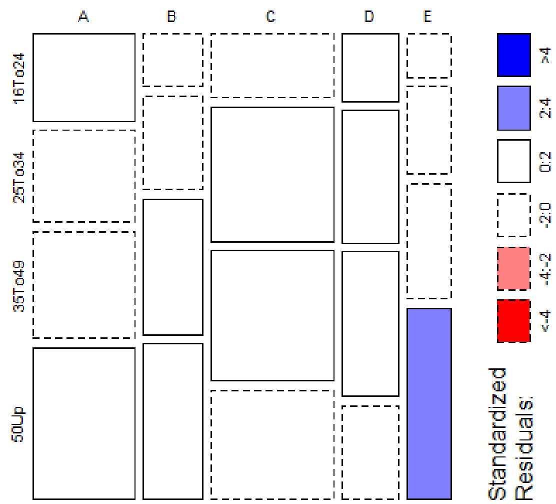
```
library(ca)
data = read.table("StoresAndAges.csv", sep=",", header=T)
data
head(data)
storesName = substr(data$X, 1,1)
stores = paste(storesName)
stores
data = data[,c(2:5)]
rownames(data) = stores
names(data) <- c( "16To24", "25To34", "35To49", "50Up")
head(data)

mosaicplot(data, shade=T, main="")
```

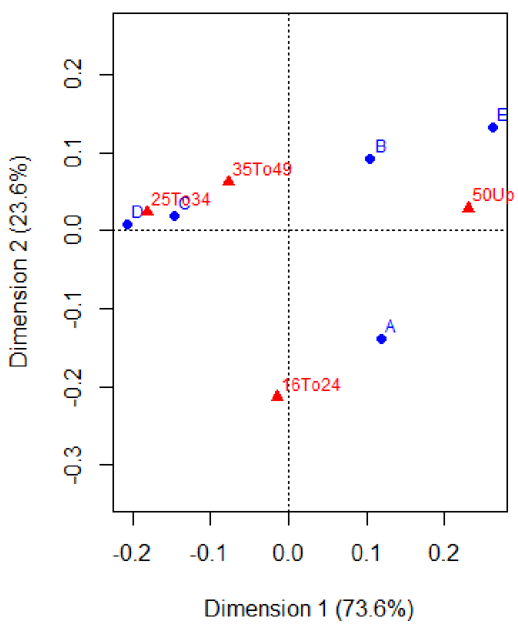
Keiland Pullen
Homework 4

DSC 424 | Fall 2021

**a.)**

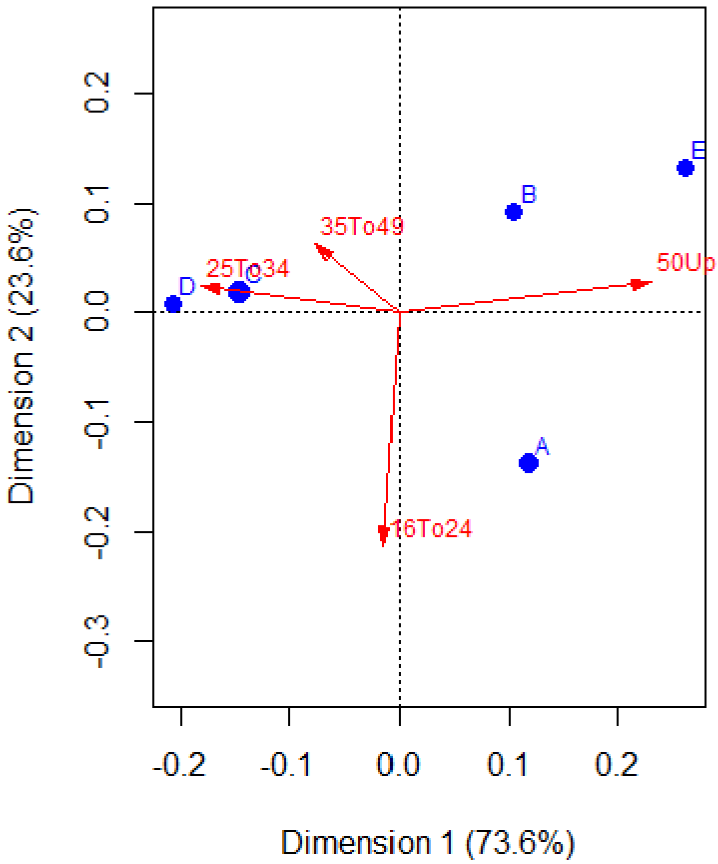



**b.)**
c = ca(data)
summary(c)
plot(c)



**c.)**

For this particular question, we're asked to create an "age profile". I wasn't sure if this plot is exactly what the question is asking or if it is asking for an actual table.




**d.)**

Reading the plot above tells us that Stores C and D are more likely to have ages "25 – 34". Additionally, store C is associated with the age "35 to 49" range as well. Store E has a relationship with the "50 and Up age group. For store A, its associated with the "16 to 24" age group. Store B is associated with both the "35 to 49" and "50 and Up" age group.

**e.)**

```
Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.026345  73.6  73.6  ******************
 2      0.008443  23.6  97.2  ******
 3      0.001008   2.8 100.0  *
        -------- -----
 Total: 0.035797 100.0


Rows:
    name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1 |    A |  264 1000  245 |  119 430 143 | -138 570 592 |
2 |    B |  153  889   93 |  104 496  63 |   93 393 155 |
3 |    C |  321  961  203 | -146 946 261 |   18  15  13 |
4 |    D |  147  966  181 | -206 965 237 |    8   1   1 |
5 |    E |  114  986  278 |  261 784 296 |  133 202 239 |

Columns:
    name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1 | 16T2 |  153  997  196 |  -15   5   1 | -213 992 822 |
2 | 25T3 |  254  954  250 | -182 937 318 |   24  16  17 |
3 | 35T4 |  286  843   93 |  -77 512  65 |   62 332 131 |
4 | 50Up |  307  997  461 |  230 982 615 |   28  15  29 |
```

The percentage of the “inertia” of the first two eignvectors accounts for 97%. It this is correct, then both should be used to get to the 80% mark, as the first is at approximately 74%. With 2 dimensions, the plots should be fairly simple using R.

**4.)**

**a.)**

```
DA <- import_list("C:/Users/Home/Desktop/DePaul/DSC-424-AdvancedDataAnalysis/week-5/Homework/BondRating.xls")
DATrain = DA$training
DATrain

LDAModel <- lda(CODERTG ~ LOPMAR + LFIXCHAR + LGEARRAT + LTDCAP + LLEVER + LCASHLTD + LACIDRAT +
              LCURRAT + LRECTURN + LASSLTD , data=DATrain)
LDAModel

Call:
lda(CODERTG ~ LOPMAR + LFIXCHAR + LGEARRAT + LTDCAP + LLEVER +
   LCASHLTD + LACIDRAT + LCURRAT + LRECTURN + LASSLTD, data = DATrain)

Prior probabilities of groups:
        1         2         3         4         5         6         7
0.1111111 0.1604938 0.1481481 0.1604938 0.1604938 0.1358025 0.1234568

Group means:
     LOPMAR  LFIXCHAR    LGEARRAT    LTDCAP      LLEVER   LCASHLTD     LACIDRAT   LCURRAT LRECTURN  LASSLTD
1 -1.738889 1.6637778 -0.99555556 0.2881111  0.12388889 -0.3940000  0.059888889 0.6932222 1.943889 1.804000
2 -2.094385 1.8042308 -1.05315385 0.2641538 -0.08338462 -0.3925385 -0.003692308 0.6640769 2.266308 1.733462
3 -2.017917 1.7306667 -0.94075000 0.3034167  0.04291667 -0.4003333  0.017500000 0.6387500 2.074250 1.693417
4 -2.213923 1.3204615 -1.01200000 0.2704615 -0.02153846 -0.5720769 -0.063230769 0.7600769 2.032077 1.721769
5 -1.981846 1.7073077 -0.75800000 0.3272308  0.07430769 -0.7765385  0.137076923 0.7471538 1.950000 1.510077
6 -2.078545 0.9529091 -0.07790909 0.4812727  0.44972727 -1.4103636 -0.033181818 0.7031818 1.818182 1.103182
7 -1.783600 0.5873000  0.10860000 0.5248000  0.64370000 -1.4720000 -0.031600000 0.4642000 1.650000 0.993700

Coefficients of linear discriminants:
                 LD1        LD2        LD3         LD4          LD5        LD6
LOPMAR   -0.7720156  -2.993776 -1.0902999   1.19056396   0.003079991 -1.0907388
LFIXCHAR  0.3309649  -1.032219  2.0342609  -0.17225468  -0.566130362  0.4446614
```

```
LGEARRAT  2.0228900 -13.206606  4.3603205  30.56370258  19.296973115 -8.6572293
LTDCAP   27.6725970  15.434851  1.0663233 -30.15183168   0.636947862 22.5703473
LLEVER   -5.2113899   4.540020 -5.2197916 -13.97013291 -12.485287860  4.5123115
LCASHLTD -0.8040312   3.684976 -0.6103313  -1.47884309   2.343115368  2.1285439
LACIDRAT -0.2978150  -3.360777 -0.7014467  -0.09884748   0.507853522 -0.9383520
LCURRAT  -2.0007312   2.040593 -1.1419790   1.51718949  -2.677213623  3.2930473
LRECTURN -1.1369903  -2.245231 -0.6432160   0.81809242   0.686713979 -0.9182123
LASSLTD   5.2328461 -14.461158  1.3481935  26.33072526  16.502239043 -5.7011832

Proportion of trace:
   LD1    LD2    LD3    LD4    LD5    LD6
0.6309 0.1209 0.1005 0.0705 0.0587 0.0186

> pred <- predict(LDAModel, newdata=DATrain[,4:13])$class
> table(pred, DATrain$CODERTG)

pred  1  2  3  4  5  6  7
   1  4  1  0  0  1  1  0
   2  3  7  3  1  1  0  0
   3  0  1  6  0  1  0  2
   4  1  2  2 11  2  0  1
   5  0  2  1  1  8  1  0
   6  1  0  0  0  0  8  1
   7  0  0  0  0  0  1  6
```

On the training data, it seems that the companies are where they should be.

**b.)**

```
DAValidate = DA$validation
DAValidate
head(DAValidate)

LDAModeVall <- lda(CODERTG ~ LOPMAR + LFIXCHAR + LGEARRAT + LTDCAP + LLEVER + LCASHLTD + LACIDRAT +
            LCURRAT + LRECTURN + LASSLTD , data=DAValidate)
LDAModeVall
pred2 <- predict(LDAModeVall, newdata=DAValidate[,4:13])$class
pred2

table(pred2,DAValidate$CODERTG)

   > LDAModeVall <- lda(CODERTG ~ LOPMAR + LFIXCHAR + LGEARRAT + LTDCAP + LLEVER + LCASHLTD + LACIDRAT +
+              LCURRAT + LRECTURN + LASSLTD , data=DAValidate)
Warning message:
In lda.default(x, grouping, ...) : variables are collinear

> head(DAValidate)
  OBS  RATING  CODERTG  LOPMAR  LFIXCHAR  LGEARRAT  LTDCAP  LLEVER  LCASHLTD  LACIDRAT  LCURRAT  LRECTURN
LASSLTD
1   8    AAA      1 -1.323    0.998   -0.936  0.281 -0.042   -0.187    0.001   0.863    1.349   1.704
2   9    AAA      1 -2.100    1.516   -1.654  0.159  0.251    0.342   -0.077   0.347    1.762   2.515
3  23     AA      2 -1.743    1.626   -1.207  0.230 -0.066   -0.266   -0.229   0.543    1.718   1.917
4  24     AA      2 -1.776    1.153   -0.450  0.389  0.171   -0.898   -0.073   0.440    2.227   1.251
5  37      A      3 -1.704    3.691   -3.155  0.040 -0.936    1.573    0.122   0.998    2.033   3.493
6  38      A      3 -1.774    0.887   -0.532  0.369  0.013   -0.929    0.070   0.781    1.891   1.232
>
> LDAModeVall <- lda(CODERTG ~ LOPMAR + LFIXCHAR + LGEARRAT + LTDCAP + LLEVER + LCASHLTD + LACIDRAT +
+              LCURRAT + LRECTURN + LASSLTD , data=DAValidate)
Warning message:
In lda.default(x, grouping, ...) : variables are collinear
> LDAModeVall
Call:
```

```
lda(CODERTG ~ LOPMAR + LFIXCHAR + LGEARRAT + LTDCAP + LLEVER +
    LCASHLTD + LACIDRAT + LCURRAT + LRECTURN + LASSLTD, data = DAValidate)

Prior probabilities of groups:
        1         2         3         4         5         6         7 
0.1428571 0.1428571 0.1428571 0.1428571 0.1428571 0.1428571 0.1428571 

Group means:
   LOPMAR LFIXCHAR LGEARRAT LTDCAP  LLEVER LCASHLTD LACIDRAT LCURRAT LRECTURN LASSLTD
1 -1.7115   1.2570  -1.2950 0.2200  0.1045   0.0775   -0.038  0.6050   1.5555  2.1095
2 -1.7595   1.3895  -0.8285 0.3095  0.0525  -0.5820   -0.151  0.4915   1.9725  1.5840
3 -1.7390   2.2890  -1.8435 0.2045 -0.4615   0.3220    0.096  0.8895   1.9620  2.3625
4 -2.0750   0.8125  -1.0790 0.2530 -0.0750  -0.4920   -0.467  0.6315   2.2220  1.7525
5 -2.1440   1.5530  -1.0440 0.2605 -0.1340  -0.5590    0.008  0.6475   2.1930  1.6740
6 -2.3700   0.9170  -0.0330 0.4900  0.3950  -1.5985   -0.244  0.7755   1.9855  1.0140
7 -1.8125   0.2815  -0.0375 0.4890  0.3355  -1.3485   -0.151  0.1900   2.1310  0.9390

Coefficients of linear discriminants:
                 LD1        LD2        LD3        LD4        LD5        LD6
LOPMAR   -2.69120927  1.5293389 -1.2026581 -0.1541835 -0.8582843 -1.6587618
LFIXCHAR  0.01650485 -1.4655423  0.4252668  1.0702619  1.5550367 -0.5075890
LGEARRAT  0.42984985 -1.1265425  0.4333757  0.5991812  0.7036927 -0.1831204
LTDCAP   -9.30916052  3.6096960 -1.2511995  1.8919072 -5.1330331 -2.0303814
LLEVER    1.89143444  2.6101123 -4.5476300  2.0071943 -0.6385808 -0.2136826
LCASHLTD -0.22607207  0.4307177 -0.3604615  0.2167872  0.1257777 -0.1415353
LACIDRAT -6.82442048  4.1009148  0.3525305  2.2692405 -2.0724229  1.6638000
LCURRAT   5.46079189  4.8637491  0.1367320  1.5815373 -2.6709414 -1.0739454
LRECTURN -2.78633528  0.3970923  0.3304194 -0.3107691 -1.6483496 -0.4048196
LASSLTD  -0.37738469  2.1209424 -1.2296289 -0.5837193 -1.2200023  0.2157965

Proportion of trace:
   LD1    LD2    LD3    LD4    LD5    LD6 
0.4849 0.2924 0.1060 0.0961 0.0128 0.0077 
> pred2 <- predict(LDAModeVall, newdata=DAValidate[,4:13])$class
> pred2
 [1] 1 1 2 7 3 3 4 4 5 5 6 6 7 7
Levels: 1 2 3 4 5 6 7
> 
> table(pred2,DAValidate$CODERTG)
     
pred2 1 2 3 4 5 6 7
    1 2 0 0 0 0 0 0
    2 0 1 0 0 0 0 0
    3 0 0 2 0 0 0 0
    4 0 0 0 2 0 0 0
    5 0 0 0 0 2 0 0
    6 0 0 0 0 0 2 0
    7 0 1 0 0 0 0 2
```

On the validation worksheet, there is one company in level 7 that appears to be in at a risk level. The majority of the companies listed, appear to all be at the AA level.

**c.)**

In this case, this is an example of where domain knowledge will prove helpful. Would certain misclassification errors be worse? This really depends on the companies that borrow the bond. On the other hand, misclassification could prove dramatic for companies that lend the bonds.