DSC 424
Assignment #4
Dr. John McDonald

Note: For each of the analysis problems, include a copy of the full analysis in your report along with your conclusions. You should include all relevant output and code snippets from R, but you should make your answers and analysis based on each part of the R output clear and easy to read. Analyze each part carefully and explain your analysis in full sentences. Do not simply include the output from R with no explanation.

1) **(20 points) Paper Review:** An academic paper from a conference or Journal will be posted to the Homework 2 content section of D2L. It contains a usage of Correspondence Analysis. Review the paper and evaluate their usage of the technique. In particular, address in detail the following points. You should be able to fill at least a page with your review and analysis, and each point should be answered in a complete paragraph with several sentences. If you claim something about the paper, you should be able to back it up with a quote or evidence from the paper.

    a) How suitable is their data for CA? Is their overall use of the technique appropriate? If not, explain why, and if so, explain what they hope to discover and how CA can help.
    b) How are they applying CA? What variables are being analyzed and what types of categorical levels do they contain?
    c) How did they use graphs from the CA in their analysis?
    d) Did they use any techniques to evaluate goodness of fit? If not, was it appropriate that they did not? How would it have helped their exposition if they had? If they did, wheat were the results?
    e) What conclusions does CA allow them to draw? How impactful are those conclusions? Are there any practical, actionable implications from their conclusions?
    f) Are there any issues that you can identify in their analysis? Do they make any assumptions that seem unwarranted? Have they mis-applied any techniques or statistical tests that we've gone over in class (note, you do not have to dive deep here, just see if on the surface that assumptions are being met and statistical tests properly applied).

2) **(20 points)** The file "Survey.csv" contains survey responses to a questionnaire about Wikepedia pages. Each question's responses are on a 5 point likert scale. Perform an ordinal Principal Factor Analysis (exploratory) on this data addressing the following points. The questions are

QU1: Articles in Wikipedia are reliable

QU2: Articles in Wikipedia are updated

QU3: Articles in Wikipedia are comprehensive

QU4: In my area of expertise, Wikipedia has a lower quality than other educational resources

QU5: I trust in the editing system of Wikipedia

VIS1: Wikipedia improves visibility of students' work

VIS2: It is easy to have a record of the contributions made in Wikipedia

VIS3: I cite Wikipedia in my academic papers

IM1: The use of Wikipedia is well considered among colleagues

IM2: In academia, sharing open educational resources is appreciated

IM3: My colleagues use Wikipedia

a) Compute the correlation matrix in three ways, first with the pearson correlation, second with the spearman rank correlation, third with the Kendall tau correlation, and visualize all three with the corrplot function. How do the correlation matrices differ.
b) Compute the KMO test for sample adequacy and interpret it.
c) Use the Spearman correlation matrix to conduct a first PCA for the data to choose a number of factors to extract in your factor analysis. Explain your reasoning for your choice of number.
d) Use the Spearman correlation with the "principal" function from the psych package to compute a principal factor analysis on the data using the number of components that you found.
e) Print the loadings with a proper cutoff and then attempt to interpret the results. Give a name to each of the factors that describes what it encapsulates. Note any surprising connections between the variables in their contributions. In particular, are there any variable groupings that are different than the label groupings for the question above
f) Evaluate the goodness of fit with the Chi-square and the RMSEA.
g) (5 points e.c.) Repeat the analysis with the polychoric correlation. The pysch package has a nice polychoric function you can use.

3) **(20 points)** Perform a correspondence analysis on the stores and ages data in StoresAndAges.csv. In this file you are provided with the table for the two sets of categories. In particular perform the following

   a) Create a mosaic plot of the two categorical variables.
   b) Plot the results of the correspondence analysis
   c) With each store, create an age profile for the store. Which customer ages are most highly and least highly represented. For each store, draw the scale for that store and demonstrate that age profile on the graph. What you are doing here is evaluating, for each store, the correspondence of each of the ages just like we did in class.
   d) What patterns can you discern form the plot? In particular look at the age category for each of the staff groups. Are there major differences in shopping patterns between the age groups? Review the class notes on how to read these plots.
   e) From the summary of the result, what percentage of the "inertia" do the first two eigenvectors account for? How many eigenvectors would we need to get to 80 of the inertia? How easy would it be to plot the data with this many dimensions?

4) (20 points): A common application of Discriminant Analysis is the classification of bonds into various bond rating classes. These ratings are intended to reflect the risk of the bond and influence the cost of borrowing for companies that issue bonds. Various financial ratios culled from annual reports are often used to help determine a company's bond rating.

The Excel spreadsheet BondRating.xls (XLS) contains two sheets named Training data and Validation data. These are data from a sample of 95 companies selected from COMPUSTAT financial data tapes. The company bonds have been classified by Moody's Bond Ratings (1980) into seven classes of risk ranging from AAA, the safest, to C, the most risky. The data include ten financial variables for each company. These are:

LOPMAR: Logarithm of the operating margin,
LFIXMAR: Logarithm of the pretax fixed charge coverage,
LTDCAP: Long-term debt to capitalization,
LGERRAT: Logarithm of total long-term debt to total equity,
LLEVER: Logarithm of the leverage,
LCASHLTD: Logarithm of the cash flow to long-term debt,
LACIDRAT: Logarithm of the acid test ratio,
LCURRAT: Logarithm of the current assets to current liabilities,
LRECTURN: Logarithm of the receivable turnover,
LASSLTD: Logarithm of the net tangible assets to long-term debt.

The data are divided into 81 observations in the Training data sheet and 14 observations in the Validation data sheet. The bond ratings have been coded into numbers in the column with the title CODERTG, with AAA coded as 1, AA as 2, etc. Develop a Linear Discriminant Analysis model to classify the bonds in the Validation data sheet.

   a) What is the performance of the classifier on the training data? Notice that there is order in the class variables (i.e., AAA is better than AA, which is better than A,…).
   b) What is the performance of the classifier on the validation data?
   c) Would certain misclassification errors be worse than others? If so, how would you suggest measuring this?