Data Mining
IT 270
Homework 2

Dr. Alexander Pelaez
alexander.pelaez@hofstra.edu

$1^{st} March, 2023$

Instructions

Please make sure if you use R you copy and paste it into Word using Courier Font (makes it easier to Read). For each of the problems that are looking for a response (not just a calculation), be sure to explain and interpret the results. If you are not sure... ASK PLEASE. Please start each question on a new page and clearly label that start of each problem (Maybe slightly larger font, bold face , underline...) anything that will help me find the problem you are working on.

Please remember if you submit a word document, you must copy your R Code and results (generally just copy what the output in the console is and it will include your code). Don't provide me R code with original data output please - it makes it too long. When you copy and paste it to word please make sure the R Code / Output is in courier font (10 pt).

1 Question 1

Using the **Domotic1.txt** dataset (See Data Files Link in Blackboard)

(a) Describe the data in the data set including necessary summary statistics. What are some issues with the data set and how do you handle them? (Then create a new dataset where you have handled these problems).

(b) The chief analyst wants to decompose the data into something more manageable, i.e. reduce the number of dimensions. Using principal component analysis, reduce the dimensions, using the appropriate technique .

(c) How many dimensions did you decide on, please indicate and show all code and results, and explain how you choose the dimensions. In addition describe the dimensions based on the variables (like our example in class where we called one dimension "size")

(d) PCA produces a chart with a unit circle. Produce this chart for this example and explain the pieces of the chart.

(e) Write the linear function for each of the components and calculate the answer for each observation, i.e. if you have 5 dimensions you will create 5 variables in the dataset, and using the linear function you create (hint the loadings of the PCA dimensions), you will generate values for each of the observations.

(f) Interpret all your findings and provide a small summary of all the findings.

2 Question 2

Using the **drugconsupmtion** dataset (See Data Files Link in Blackboard)

(a) UGH THIS DATASET STINKS!!! Do you agree / disagree, why?

(b) The chief researcher believes that there are a number of factors that can be used in explaining the data. Tell the researcher how many factors are appropriate and how you would label the factors, and use the appropriate rotation?

(c) Which variables don't seem to be explained well by the factors (what is this called and what are the values)?

(d) Explain why we might do rotation of the factors. What is the difference between orthogonal and oblique rotations?

(e) Run your factor analysis with three rotations (none, oblique and orthogonal). Is there any difference in your final results, ultimately, what are your final results.

(f) Provide a chart of your final answer including all the factors, loadings and unique variances. *While there is a function in R that will do this for you ,* ***do not use it****. Either draw it by hand and take a picture, or use my PowerPoint (hint much better idea), and just change some of the numbers and lines.*

3 Question 3

Using the **Mpg.csv** dataset (See Data Files Link in Blackboard), answer the questions below.

(a) Create an appropriate training set (80%) and test set (20%) for the data.

(b) Run a basic linear regression and identify the important variables against the training set.

(c) Once you have a model, provide measures of accuracy (e.g. ME, MAE, MPE, MAPE, RMSE, etc) for the training set

(d) Run the model against the test set and provide measures of accuracy.

(e) **[Optional]** run the model with the training set then iteratively add one row to the training set, and test measures of accuracy against the test test, i.e. allow the model to change coefficients as new data comes in. Store the measures of accuracy in a data frame for each run, and interpret the results. *This requires some deep coding, but for those who want to give it a try, I will help you with it.*