Data Mining
IT 270
Homework 1

Prof. Alexander Pelaez
apelaez@hofstra.edu

$8^{th} Feburuary, 2023$

1 Question 1

For parts (a-f) please show all work. You may use R for parts (g and h).

Given:

$$\mathbf{A} = \begin{bmatrix} -3 & 4 \\ -1 & 3 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 0 & -5 \\ 2 & 6 \\ 4 & 2 \end{bmatrix}, \mathbf{C} = \begin{bmatrix} 2 \\ 9 \\ 3 \end{bmatrix}, \mathbf{D} = \begin{bmatrix} 1 & 2 \\ 2 & -2 \end{bmatrix}, \mathbf{E} = \begin{bmatrix} 9 & 13 & 12 \\ 4 & 3 & 10 \\ 4 & 6 & 12 \end{bmatrix}, \mathbf{\Sigma} = \begin{bmatrix} 4 & 2 & 0 \\ 2 & 8 & 3 \\ 0 & 3 & 12 \end{bmatrix}$$

(a) $\mathbf{BA}$

(b) $\mathbf{A'DB'}$

(c) Find the determinant of $\mathbf{D}$. Show all work.

(d) $(\mathbf{AA})\mathbf{A'}$

(e) Find the trace of matrix $\mathbf{E}$.

(f) Assume Sigma $\mathbf{\Sigma}$ is the covariance matrix of the matrix $\mathbf{X}$; calculate $\mathbf{V}^{\frac{1}{2}}$.

(g) Calculate the correlation matrix of Sigma $\mathbf{\Sigma}$ (do not use the `cov2cor()` function).

(h) Using `R` calculate the determinant of matrix Sigma $\mathbf{\Sigma}$.

2 Question 2

Using the ForestFires dataset (See Data Files Link in Blackboard)

(a) Describe the data in the data set including column atttributes, necessary summary statistics, missing values and outliers.

(b) Creating groups from the data is a key task for an analyst. What groups can be created from the dataset, i.e. look at a column and identify new columns with categorical grouping variables that can be derived from another. You dont need to create them in R just describe what the column would be and what it would contain. If you want to do it in R you are welcome to.

(c) Analyze the correlations of the columns, is there anything interesting? (Please round to 3 decimals). Be sure to only include numeric variables. Also include only variables where correlation would have meaning.

(d) Lookup the function corrplot in the corrplot package (provide a nice correlation chart).

3 Question 3

Using the dataset `red-wine.csv`

(a) Given any dataset, such as this one, what are some of the methods to handle outliers. What would you use for this dataset, if anything?

(b) Given any dataset, such as this one, what are some of the methods to handle missing observations. In the red-wine.csv dataset, how many records have missing observations, what would you do with these? Hint: Look up the `complete.cases()` function in `R`.

(c) For each column in the red-wine.csv, examine the column and determine, not necessarily by statistical method, the number of outliers, if any, and how you might consider handling them?

(d) Normalize the data set and produce a covariance and correlation matrix. Only use the columns where a covaraince or correlation would have meaning

Note to answer the question above you could consider creating a table in TeXor LaTeXe.g., or any other typesetting system or a Word Processing software:

Example: Column	Outliers	Missing Observations
X	15 Outliers, max outlier is observation 22 with a value of 300. Explanation of why this is considered an outlier	22 missing observations, Consider removing all because...
Y		
Y		