Student Name & ID: Akhil Montrose || 81788

Data Science Assignment 1: Basics & Titanic Dataset [50 marks]

$$\boldsymbol{R55} = \begin{bmatrix} \mathbf{3} & \mathbf{5} & \mathbf{7} & \mathbf{9} & \mathbf{11} \\ \mathbf{7} & \mathbf{8} & \mathbf{9} & \mathbf{10} & \mathbf{10} \\ \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} \\ \mathbf{6} & \mathbf{9} & \mathbf{8} & \mathbf{5} & \mathbf{12} \\ \mathbf{1} & \mathbf{4} & \mathbf{5} & \mathbf{7} & \mathbf{5} \end{bmatrix}$$

Figure 1: Elements of Matrix R88

Basics: Write code in Matlab to perform the following

1. **Data Creation and Storage:** Create the following matrices:
   a. A 7 x 7 matrix containing ones, and name it O7 [1 mark]
   b. Create the R55 matrix as shown in Figure 1 above [1 mark]
   c. Save matrix R55 at location C:\Matlab\Work [1 mark]

a) O7 = ones (7,7);
b) a = [3 5 7 9 11];
   b = [7 8 9 10 10];
   c = [1 2 3 4 5];
   d = [6 9 8 5 12];
   e = [1 4 5 7 5];
   a1 = [a;b];
   a2 = [a1;c];
   a3 = [a2;d];
   R55 = [a3;e];

2. **Data Extraction:** Extract the following items from R55 created in part(a):
    a. Element located on the 2nd row and 4th column, and name it E24 [1 mark]
    b. The entire second row, and name it Row2 [1 mark]
    c. The entire 4th column, and name it Col4 [1 mark]
    d. The Subsection of elements starting from the element located on the 1st row and 2nd column, and ending at the element located on the 4th row and 5th column, and name it Sub1245 [1 mark]

a) E24 = R55(2,4);
b) Row2 = R55(2);
c) Col4 = R55(:,4);
d) Sub1245 = R55([1 2 3 4],[2 3 4 5:end-1]);

3. Data Modification: Replace the subsection, Sub1245, in matrix R55, with the corresponding subsection from matrix O7, and name it R55Mod [1 mark]

```
O7Mod = O7([1 2 3 4],[1 2 3]);
R55Mod = R55;
R55Mod(Sub1245) = O7Mod;
```

4. **Searching Through Data:** Search through matrix R55 to find the specific element of: SE = 5, and indicate where these elements are located, using:
    a. For loops and if statements [3 marks]
    b. Logical matrices [3 marks]
    c. The find function in Matlab [3 marks]

a)
```
Size = 0;
for I = 1; 1:25
   if(R55(i) ==5)
   size = size +I;
   location(size,:) = I;
end
end
```

b)
```
R55(R55==5) = 1;
```

c)
```
[h,j] = find (R55.' == 5);
output = [h,j];
```

5. **Sorting Data:** sort matrix R55 by:
    a. Rows (you can use the sortrows function in Matlab) [1 mark]
    b. Columns (you can use the sortrows function in Matlab) [1 mark]

    a) sortArray = sortrows(R55);
    b) sortMatrixC = sort(R55);

6. **Obtain Basic Statistics:** obtain the following:
    a. Mean value from the entire matrix R55 [1 mark]
    b. Minimum value from Row 4 of matrix R55 [1 mark]
    c. Maximum value from Col 4 of matrix R55 [1 mark]
    d. Standard deviation of matrix Sub1245 (in part 2-d) [1 mark]

    a) M = mean(R55);
    3.60000000000000 5.60000000000000 6.40000000000000 7
    8.60000000000000

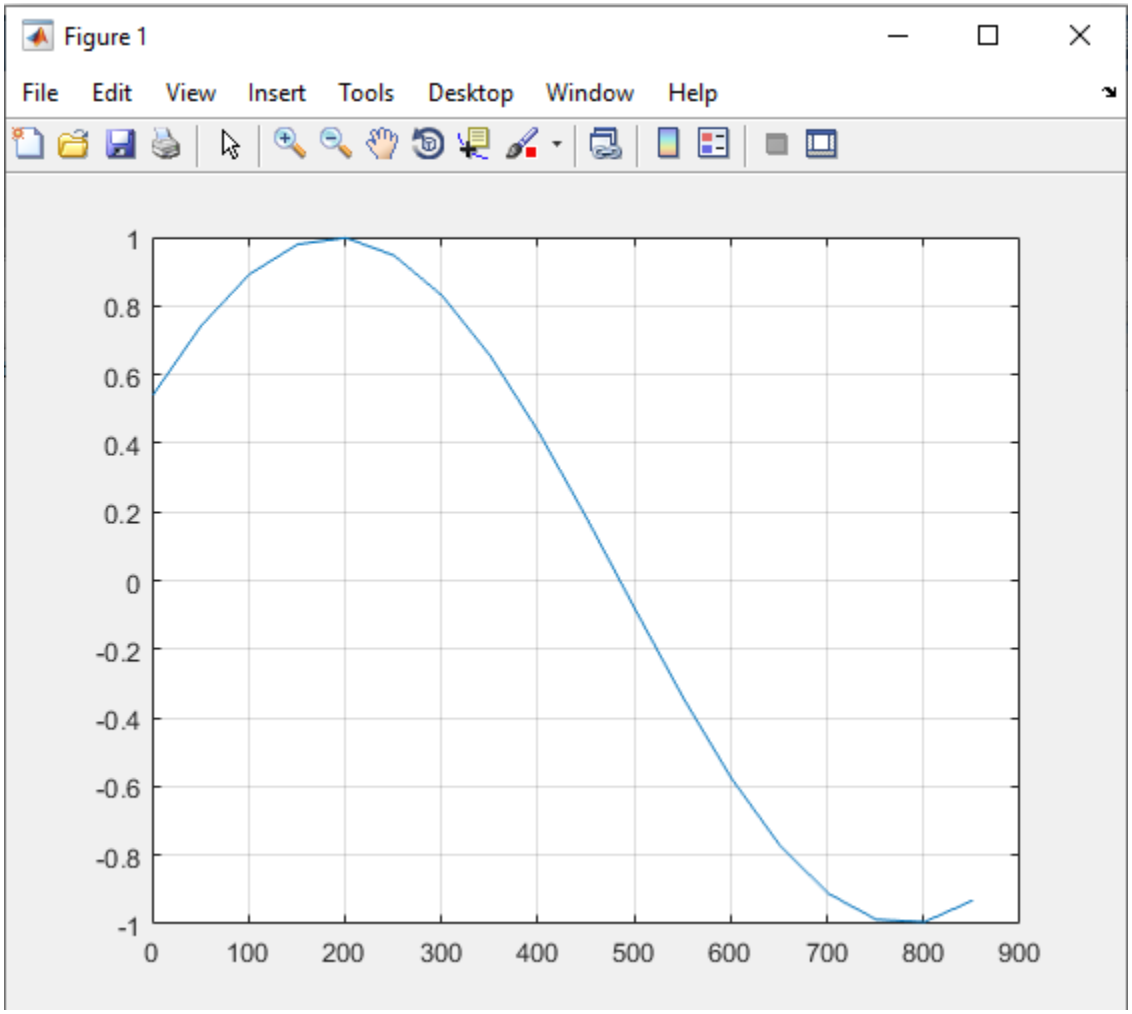
    b) minimum = min(R55);
    1 2 3 4 5

    c) maximum = max(R55);
    7 9 9 10 12

    d) StandardD = std(Sub1245);
    3.16227766016838 2.62995563967658 2.94392028877595

7. **Plotting Data:** plot a graph of the cos function using 900 points [1 mark]

```
x = 1:50:900;
y = cos(x);
plot(x,y);
```



8. **Create a Function:** to find the sum of four numbers [3 marks]

```
function total = tot(a,b,c,d)
    total = a+b+c+d
end
a =1; b=2; c =4; d=9;
total = tot(a, b, c, d);
```

Analysis of the Titanic Dataset: Write code in Matlab to perform the following:

9. Fill-in Missing Data: fill the missing ages with the average age. [5 marks]

```
% Loading age from column 6
ageData = TitanicDataSet(:,6);

%%%%% convert to type double
ageDataNum = cell2mat(ageData);

%%%%% Search dataset to find NaN values
ageMissL = isnan(ageDataNum);

%%%% find the mean age in the dataset
ageExist = ageDataNum(~ageMissL);
ageMean = mean(ageExist);

%%% replace missing age data with mean age
% make a copy
ageFull = ageDataNum;

% replace NaN (missing data) with zeros
ageFull(isnan(ageFull)) = 0;

% replace zeros (missing data) with median age
MeanVector = ageMean*ageMissL;
ageFull = ageFull + MeanVector;
```

10. Locating Data: find the names of persons who payed fares between 75 and 200 dollars [8 marks]

```
fareData = TitanicDataSet(:,10);
fareDataNum = cell2mat(fareData);
fareDataLow = fareDataNum>=75;
fareDataHigh = fareDataNum<=200;
fareDataRange = fareDataLow.*fareDataHigh;
fareDataRange = fareDataRange>0;

nameData = TitanicDataSet(:,4);
correctNames = nameData(fareDataRange);
```

11. Finding Simple Probability: find the probability of survival given the guest is a mid-aged man (i.e. with age between 30-45) staying in a second class cabin [10 marks]

```
%%11
survData = TitanicDataSet(:,2);
survDataNum = cell2mat(survData);
noGuests = length(survDataNum);

sexData = TitanicDataSet(:,5);
sexMl = strcmp(sexData,'male');

%2 vector for class=2
classData = TitanicDataSet(:,3);
classDataNum = cell2mat(classData);
class2L = classDataNum==2;


%2-a-iii) vector for 30<Age<45

ageData = TitanicDataSet(:,6);
```

```
ageDataNum = cell2mat(ageData);
ageDataLow = ageDataNum>=30;
ageDataHigh = ageDataNum<=45;
ageDataRange = ageDataLow.*ageDataHigh;


%2-a-iv) vector for sex=male AND class=2 AND 30<Age<45
YFC2 = sexMl.*class2L.*ageDataRange;


%2-a-v) P(18<Age<35,Sex=female,Class=1)
pYFC2 = sum(YFC2)/noGuests;
```

Please type out your solutions in this word document, then convert it to a pdf (file ➔ save as ➔ type pdf, make sure you have Adobe Acrobat Reader installed) submit your solutions as a pdf file on canvas