Maral Nourimand
Data Mining – Week 02 Exercise
09.11.2023

Question #1

```
% Load the data from the text file
data = readtable('inco13par.txt');

% Identify the types of variables:
% they are nominal variables(like DIAGNOSI, UVA, US, SS, AGE, ...) and
% numeric variables under each category of these nominal variables.

% Display the summary of the variables
summary(data);
```

```
Variables:

  NO: 529×1 double

    Values:

      Min         1
      Median    344
      Max       727

  DIAGNOSI: 529×1 double

    Values:

      Min        0
      Median     0
      Max        4

  UVA: 529×1 double

    Values:

      Min            0
      Median         0
      Max            1
      NumMissing     1

  US: 529×1 double

    Values:
      Min          0
      Median        7
      Max          18
      NumMissing     144
...
```

For the other categories Min/Median/Max are calculated also. I did not copy all of them to keep the solution simple. Some categories have no Missing values.

```
% we can calculate the mean of the entire dataset using nanmean
% AGE is the last column
mean_age = nanmean(data(:, 16))

% Find unique diagnoses and count the number of unique diagnoses
% DIAGNOSI are in the second column.
% or by name of the column: diagnoses = unique(data.DIAGNOSI)
diagnoses = unique(data(:, 2))
num_diagnoses = height(diagnoses)
```

```
  AGE: 529×1 double

      Values:

        Min          26
        Median        51.5
        Max          89
        NumMissing     7


mean_age =

  table

    AGE
   _______

   52.328


diagnoses =

  5×1 table

   DIAGNOSI
   ________

      0
      1
```

```
     2
     3
     4

num_diagnoses =

     5
```

Question #2

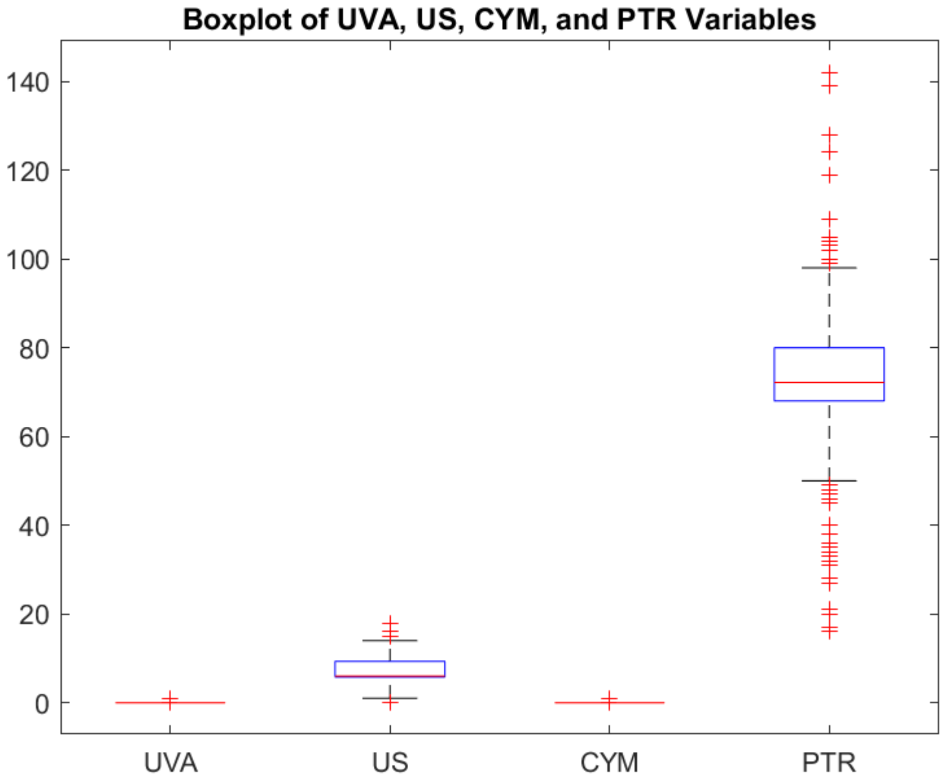
```matlab
% Loop through each DIAGNOSI[0,1,2,3,4] and replace NaNs with the mean of
% the respective DIAGNOSI
data_new = data;  % initialize new table
for i = 0:num_diagnoses-1
    idx = find(data.DIAGNOSI == i); % Find the indices of the current diagnosis
    for j = 1:width(data) % Loop through each variable
        current_column = data{idx, j};
      % Calculate the mean without NaNs
        current_mean = mean(current_column, 'omitnan');
      % Replace NaNs with the mean
        current_column(isnan(current_column)) = current_mean;
      % Assign the modified column to the new table
        data_new{idx, j} = current_column;
    end
end
```

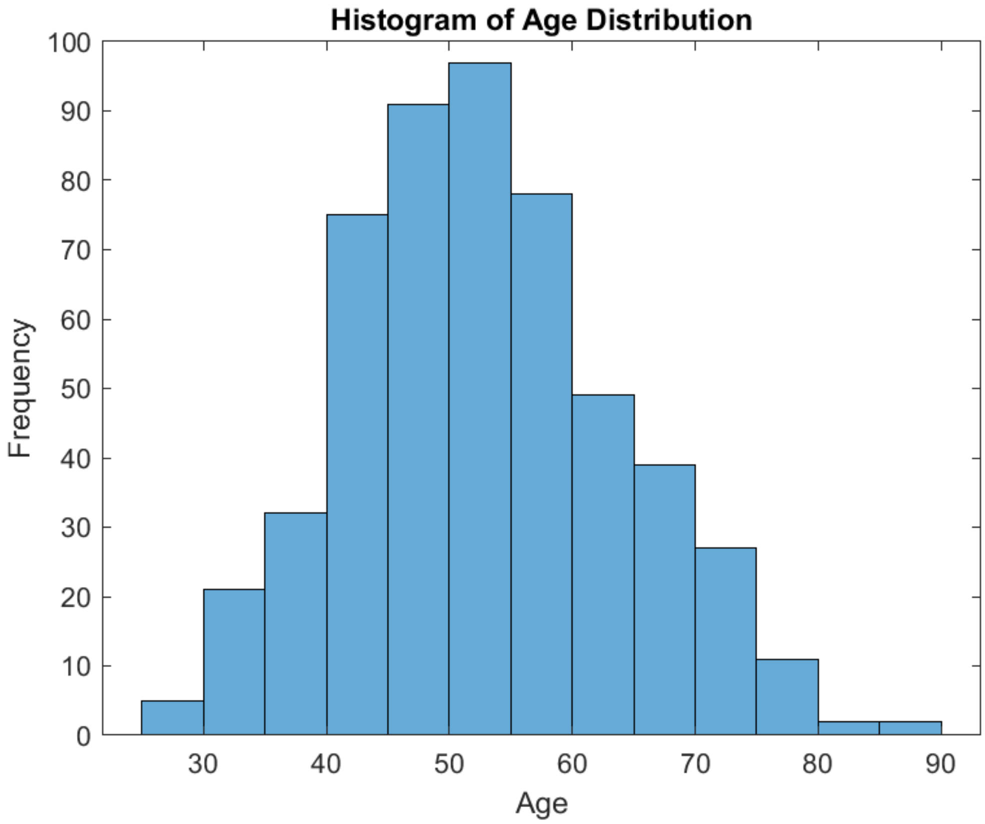
Question #3

```matlab
UVA = data_new.UVA; % Extracting the UVA variable
US = data_new.US;   % Extracting the US variable
CYM = data_new.CYM; % Extracting the CYM variable
PTR = data_new.PTR; % Extracting the PTR variable
Age = data_new.AGE; % Extracting the Age variable

% Create a box plot for the selected variables
boxplot([UVA, US, CYM, PTR], 'Labels', {'UVA', 'US', 'CYM', 'PTR'});
title('Boxplot of UVA, US, CYM, and PTR Variables');

% Create a histogram for the Age variable
histogram(Age);
title('Histogram of Age Distribution');
xlabel('Age');
ylabel('Frequency');
```





```
% some skewness in the tail of the AGE histogram can be seen
```

Question #4

```
% Extract rows 2, 269, and 393 and assign them into a table
rows_of_interest = [2, 269, 393];
selected_rows = data_new(rows_of_interest, :);

% Extract the variables for calculation from table into double
data_selected = selected_rows{:, 1:end};

% Calculate Euclidean distances
% pdist(X) returns the Euclidean distance between pairs of observations in X
distances = pdist(data_selected, 'euclidean');

% Reshape the distances into a square matrix
num_rows = size(data_selected, 1);
distances_matrix = squareform(distances)
```

```
distances_matrix =

         0  456.1220  466.1793
  456.1220         0   25.5926
  466.1793   25.5926         0
```

```
% Find the minimum distances
min_distances = min(distances_matrix(distances_matrix > 0));
disp("Minimum distances:");
disp(min_distances);
```

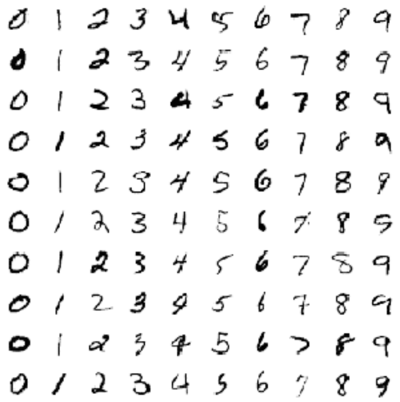
```
Minimum distances:
   25.5926
```

It means that the case of the row 269 and row 393 are the closest case to each other.
When using the Euclidean distance measure, one potential problem is that it assumes that all features are equally important and that the relationships between features are linear. So it may not be accurate representation to compare observations.

Some other distance measures include:

- ✓ Manhattan distance (L1 distance)
- ✓ Cosine distance (it measures the angle between two vectors)
- ✓ Hamming distance (for categorical data)
- ✓ Mahalanobis distance (accounts for correlations between variables)

Question #5



```
% Calculate the sum of absolute differences between pixel values
difference_num1_num2 = sum(abs(num1 - num2), 'all');
difference_num1_num3 = sum(abs(num1 - num3), 'all');
difference_num2_num3 = sum(abs(num2 - num3), 'all');

% Display the results
disp("Sum of absolute differences between (5,5) and (1,5):");
disp(difference_num1_num2);

disp("Sum of absolute differences between (5,5) and (1,6):");
disp(difference_num1_num3);

disp("Sum of absolute differences between (1,5) and (1,6):");
disp(difference_num2_num3);
```

```
Sum of absolute differences between (5,5) and (1,5):
    66

Sum of absolute differences between (5,5) and (1,6):
    81

Sum of absolute differences between (1,5) and (1,6):
    67
```

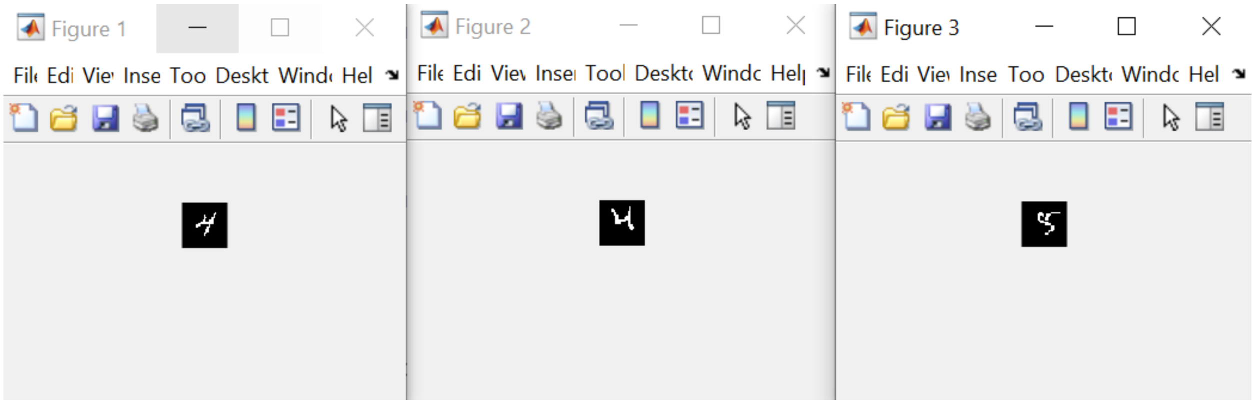


Figure 1
Fil Edi Vie Inse Too Deskt Wind Hel
Figure 2
File Edi View Inse Tool Deskt Windc Help
Figure 3
File Edi View Inse Too Deskt Windc Hel

The main factor that affects the result obtained is the visual dissimilarity between the handwritten numbers at the specified positions. The calculated differences between the pixel values represent the dissimilarity between the corresponding images. The lower the sum of absolute differences, the more similar the images are. Here, we can see that Figure1 and Figure2 are the closest as their distance is the minimum also.