1a. (10 points) What is the difference between discrimination and classification? Between characterization and clustering? Between classification and prediction? For each of these pairs of tasks, how are they similar?

1.Discrimination: is about understanding patterns that can lead to treating people or groups unfairly because of certain things like their race or age. Its goal is to divide individuals into different groups based on their measurements or characteristics, and it can be called separation.

Classification: observing new cases along with numeric values and assigning them to groups base on their number values, and it involves building models that put data into different categories or groups based on certain features.

Both discrimination and classification involve categorizing or sorting items based on certain criteria.

2. Characterization summarize a dataset, or a group of data based on various statistical measures or features. Its goal is to provide insights into the underlying patterns or properties of the data.

Clustering: is a data analysis technique where data points are grouped together based on their similarity. It is goal is to find natural groupings or clusters within the data without any prior information about the class labels.

Both characterization and clustering involve organizing data based on similarities or patterns

3.Classification: is a supervised learning, usually you have a set of predefined classes, classifies new data based on observation from training set, K-Nearest neighbor and decision tree algorithms are most popular, and it is more complex comp

Predication estimates future outcome based on historical data or patterns. It can be either supervised or unsupervised, depending on the availability of labeled data.

Both classification and prediction involve making decisions based on available information or patterns in the data.

1b. (10 points) Please described the three major stages of Knowledge Discovery Process.

Data preparation: Includes data preprocessing to ensure that the data is suitable, integration, transformation, reduction, and removal of noise or outliers, collecting necessary information to model or account for noise.

Data mining: Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth. Its goal is to extract valuable information that can be used for decision-making and problem-solving.

Interpretation of the extracted units: analyzing the results obtained from data mining algorithms, interpreting the discovered patterns in the context of the problem domain, and evaluating their relevance and reliability.

2a. (10 points) Explain the definition of Minkowski Distance and list 3 special cases of it.

Minkowski helps us figure out how far apart or similar two points are from each other in a space with multiple dimensions. is distance is a mix of Euclidean distance and Manhattan distance
Special cases: Manhattan, Euclidean, and Chebyshev distance.

2b. (10 points) Compute the cosine similarity between the two vectors below:
x1 = (4, 0, 1, 3, 5)
x2 = (−1, 2, 3, 0, 1)

$$.\frac{((4*(-1))+(0*2)+(1*3)+(3*0)+(5*1))}{\sqrt{4^2+0^2+1^2+3^2+5\text{^}2}*\sqrt{(-1)^2+2^2+3^2+0^2+1^2}} = 0.145$$

$$4^2 + 0^2 + 1^2 + 3^2 + 5^2$$

3. (20 points) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) What is the mean of the data? What is the median?
The mean = 809/24 = 29.963
The median is the 14th = 25

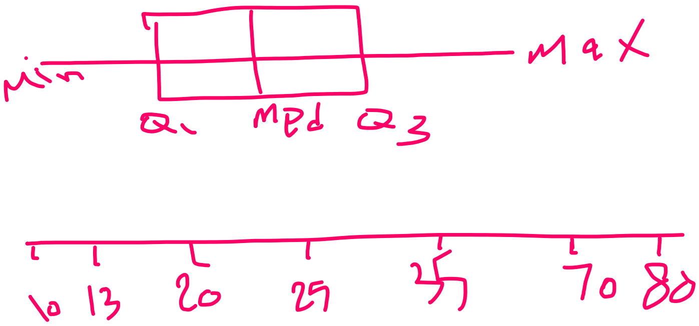
(b) What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).
Mode: 25, 35 and it’s bimodal

(c) What is the midrange of the data?
Midrange = (13 + 70)/2= 41.5

(d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?
13/2 = 7.5, will be number 7, Q1 = 20, Q3 = 35

(e) Give the five-number summary of the data.
Min = 13, Q1 =20, Median =25, Q3 =25, Max =70

(f) Show a boxplot of the data.



(g) How is a quantile-quantile plot different from a quantile plot?

A quantile plot shows the quantiles of one dataset against the quantiles of another dataset without comparing them.

A quantile-quantile plot compares the quantiles of the dataset to the quantiles of a theoretical distribution.

4. (20 points) Using the data for age given in question 3, answer the following.

(a) Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

Step 1: sort the data in increasing order (already sorted).

Step2: group the date into bins of size 3.

Step3: calculate the mean for each bin and replace the values in the bin with the means.

Effect: Smoothing by bin means reduces the variability in the data by replacing individual values with the mean of nearby values, it helps to identify trending but may cause the loss of information if the data has no outliers.

Step2:
Bin 1: 13, 15, 16
Bin 2: 16, 19, 20
Bin 3: 20, 21, 22
Bin 4: 22, 25, 25
Bin 5: 25, 25, 30
Bin 6: 33, 33, 35
Bin 7: 35, 35, 35
Bin 8: 36, 40, 45
Bin 9: 46, 52, 70

Step3:
Bin 1: (13 + 15 + 16) / 3 = 14.67
Bin 2: (16 + 19 + 20) / 3 = 18.33
Bin 3: (20 + 21 + 22) / 3 = 21.00
Bin 4: (22 + 25 + 25) / 3 = 24.00

Bin 5: (25 + 25 + 30) / 3 = 26.67
Bin 6: (33 + 33 + 35) / 3 = 33.67
Bin 7: (35 + 35 + 35) / 3 = 35.00
Bin 8: (36 + 40 + 45) / 3 = 40.33
Bin 9: (46 + 52 + 70) / 3 = 56.00

Smoothed data: 14.67, 14.67, 14.67, 18.33, 18.33, 18.33, 21.00, 21.00, 21.00, 24.00, 24.00, 24.00, 26.67, 26.67, 26.67, 33.67, 33.67, 33.67, 35.00, 35.00, 35.00, 40.33, 40.33, 40.33, 56.00

(b) How might you determine outliers in the data?
Z-score: Calculate the z-score of each data point and identify points with z-scores beyond a certain threshold as outliers.
Boxplot: Plot the data using a boxplot and identify points beyond the whiskers as outliers.
Modified Z-score: A variation of the z-score method that is more robust to outliers.

(c) What other methods are there for data smoothing?
Moving Averages: Calculate the average of data points within a sliding window. This smooths out short-term fluctuations.

Kernel Smoothing: Use a kernel function to assign weights to neighboring data points based on their distance from the point being smoothed.

Moving average: Replace each data point with the average of itself and its neighboring points within a specified window size

Exponential Smoothing: Assign different weights to past data points, giving more importance to recent data. This method is useful for time-series data.

Polynomial Regression: Fit a polynomial function to the data, which provides a smooth curve through the points.

Savitzky-Golay Filter: A type of linear smoothing filter that fits polynomial functions to data segments.

5. (20 points) Consider a dataset with six variables: X1, X2, X3, X4, X5, and X6. The covariance matrix of this dataset is given by:

x1 = [2.5, 2.3, 2.1, 0.7]
x2 = [0.5, 0.8, 1.5, 3.6]
x3 = [2.2, 3.0, 0.3, 2.0]
x4 = [1.9, 2.2, 1.8, 0.3]
x5 = [3.1, 2.5, 3.2, 1.0]
x6 = [2.3, 2.8, 0.5, 4.0]
Calculate the covariance matrix for X1, X2, X3, X4, X5 and X6. Show your steps in the computation and provide the final covariance matrix.

1. we Compute the mean for each variable:

X1 mean = (2.5 + 2.3 + 2.1 + 0.7) / 4 = 1.9
Repeat the same steps for x2, x3, x4, x5, and x 6, we end up with:
The means of each variable are as follows:
Mean of X1: 1.9
Mean of X2: 1.6
Mean of X3: 1.875
Mean of X4: 1.55
Mean of X5: 2.45
Mean of X6: 2.4

2. We calculate the covariance between each pair of variables using:
cov(Xi, Xj) = $\frac{1}{n-1}\sum_{k=1}^{n}(\mathrm{Xik} - \mathrm{Xi})(\mathrm{Xjk} - \mathrm{Xj})$.

$$\left[\begin{bmatrix} 0.667 & -1.133 & 0.06 & 0.673 & 0.767 & -0.733 \\ -1.133 & 1.953 & -0.283 & -1.143 & -1.243 & 1.06 \\ 0.06 & -0.283 & 1.289 & 0.098 & -0.365 & 1.203 \\ 0.673 & -1.143 & 0.098 & 0.723 & 0.753 & -0.75 \\ 0.767 & -1.243 & -0.365 & 0.753 & 1.03 & -1.263 \\ -0.733 & 1.06 & 1.203 & -0.75 & -1.263 & 2.113 \end{bmatrix}\right]$$