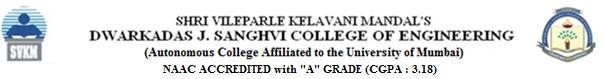
Academic Year 2024-25



**DEPARTMENT OF INFORMATION TECHNOLOGY**

**COURSE CODE: DJ19ITL503 SAP ID: 60003220045**

**COURSE NAME: Data Warehousing and Mining CLASS: T Y B. TECH**

**NAME : Anish Sharma**

# LAB EXPERIMENT NO. 5

**CO/LO: Interpret data for a given dataset.**

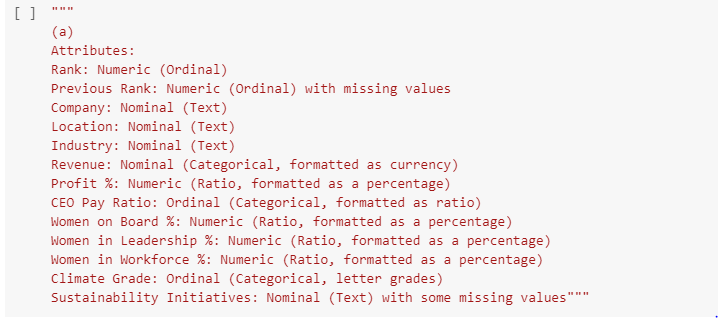
**Aim: To perform data exploration and preprocessing on selected data set.**

**Instructions:**

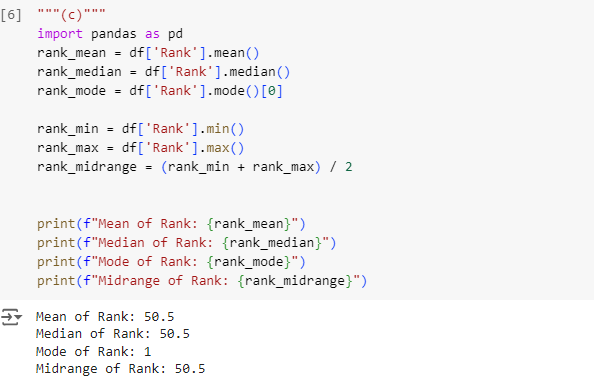
Select any data set. (Kaggle, UCI Machine Learning Repository, Google data sets)

**Exercise 1:**

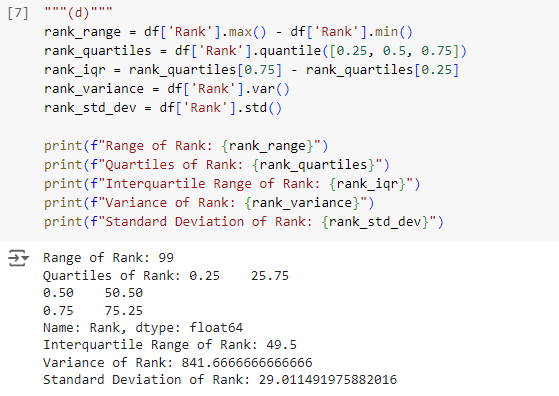
1. List all the attributes.
2. Identify the type of each attribute i.e. nominal, binary (symmetric or asymmetric), ordinal, numeric (interval scaled or ratio scaled).



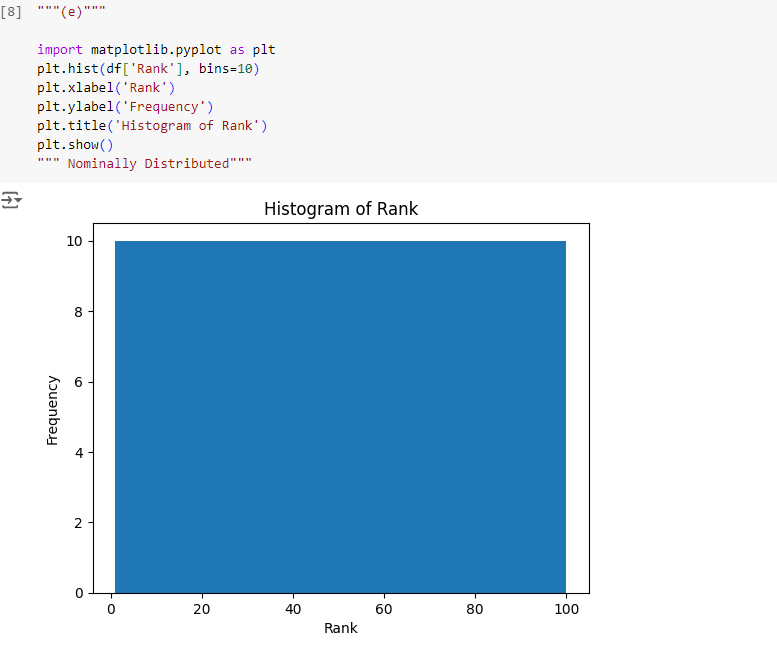
1. Calculate the mean, median, mode and midrange for required attribute to measure the center of data.



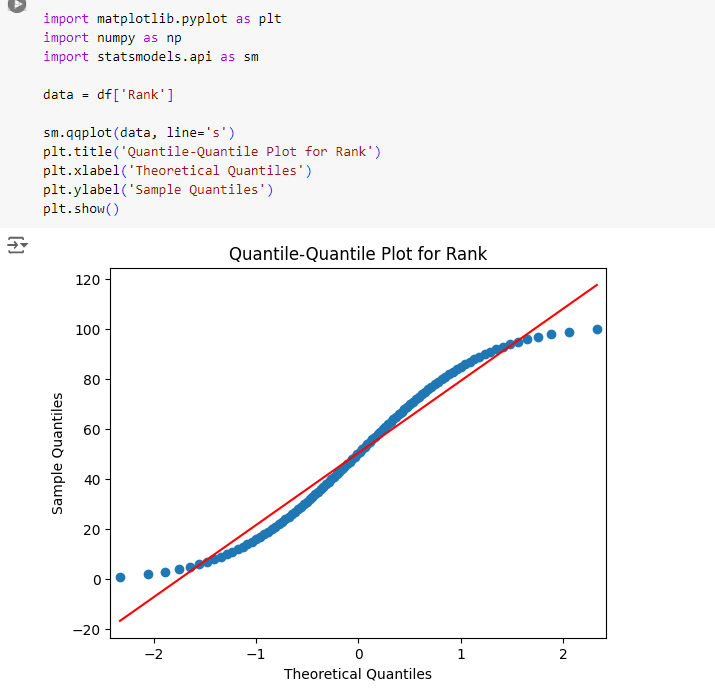
1. Calculate range, quartiles, interquartile range, variance and standard deviation for required attribute to measure the dispersion of data.



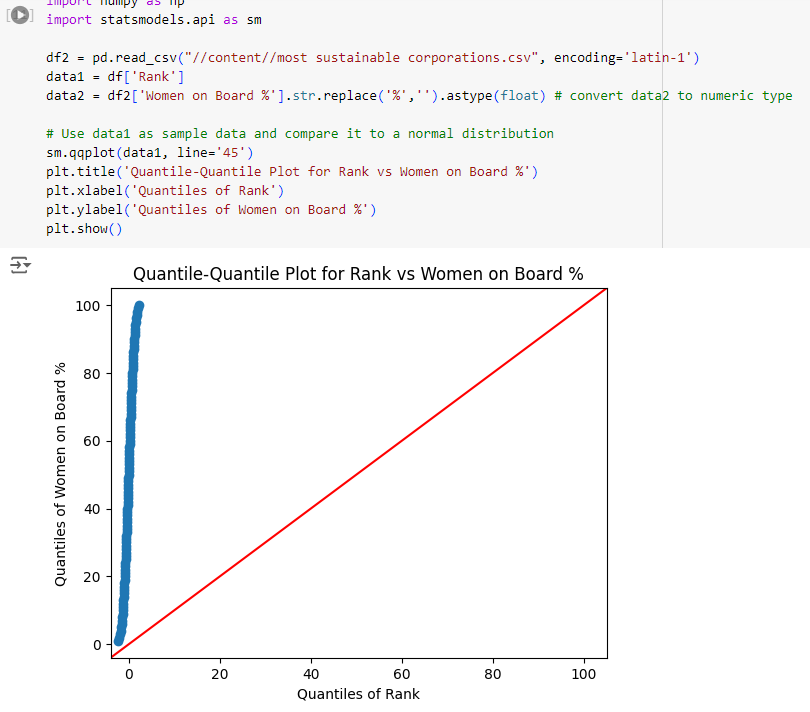
1. Identify whether the data is normally distributed or left and right skewed by plotting the required attribute using histogram.



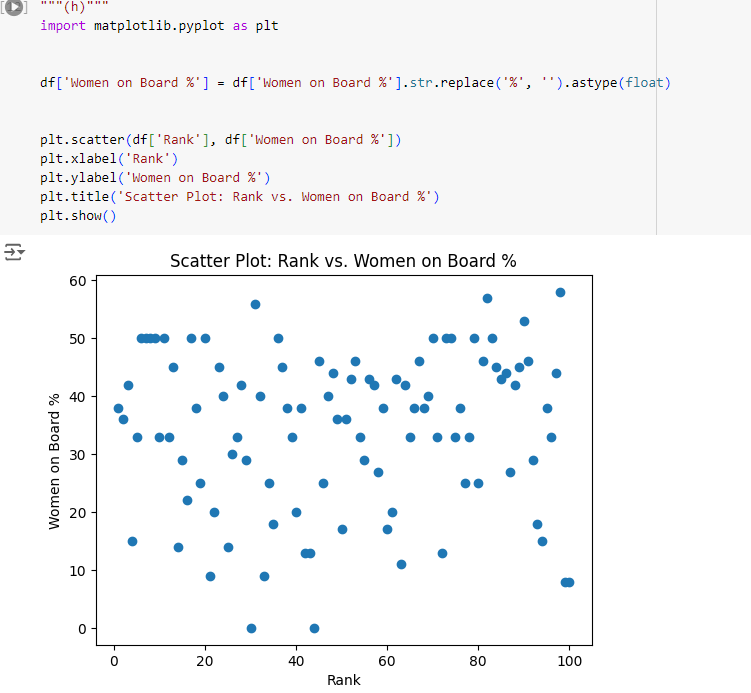
1. Draw the Quantile plot to check the distribution of data based on quantiles (use univariate data distribution).



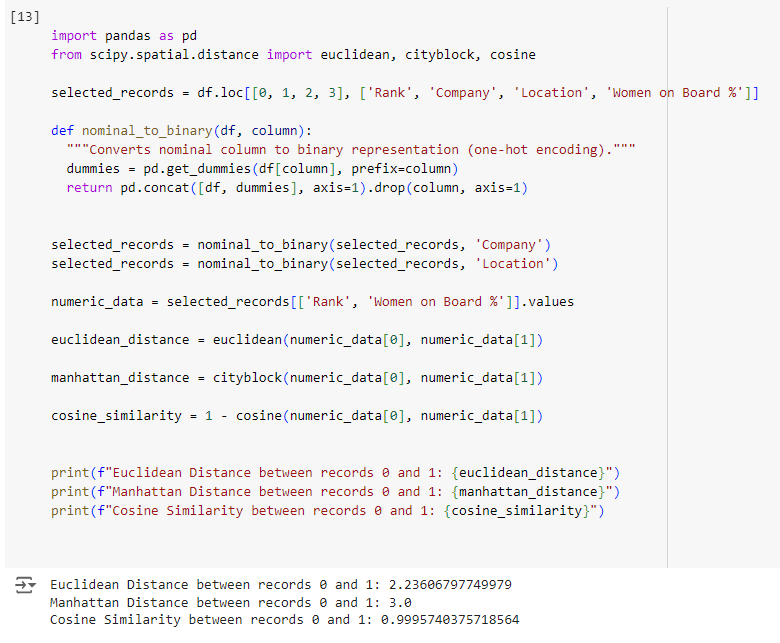
Draw Quantile-Quantile plot and determine whether the data follows the same distribution (use attributes from two different data sets)



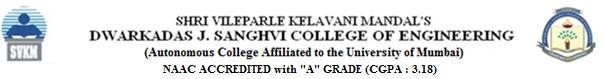
1. Draw a Scatter plot for any two attributes to identify whether the attributes have positive, negative or no correlation.



1. Calculate the dissimilarity measure for the selected dataset by selecting few records (should include nominal, binary, ordinal and numeric attributes).



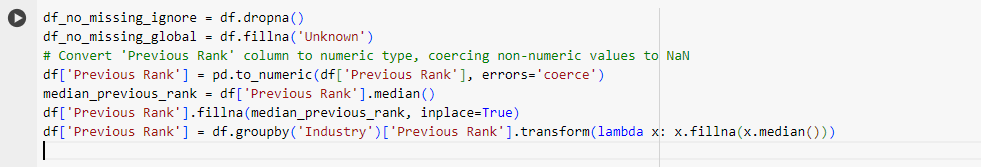
Academic Year 2024-25



**Exercise 2:**

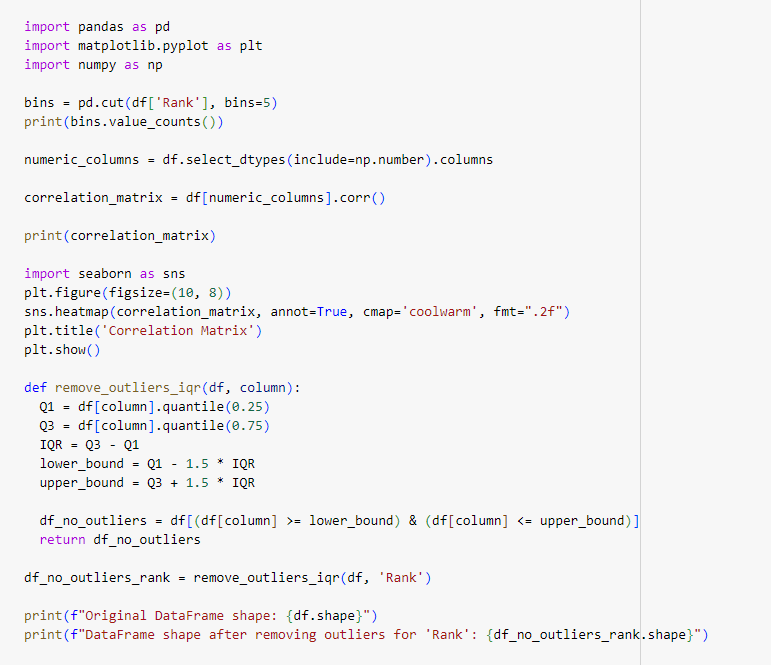
I. Preprocess the selected data set

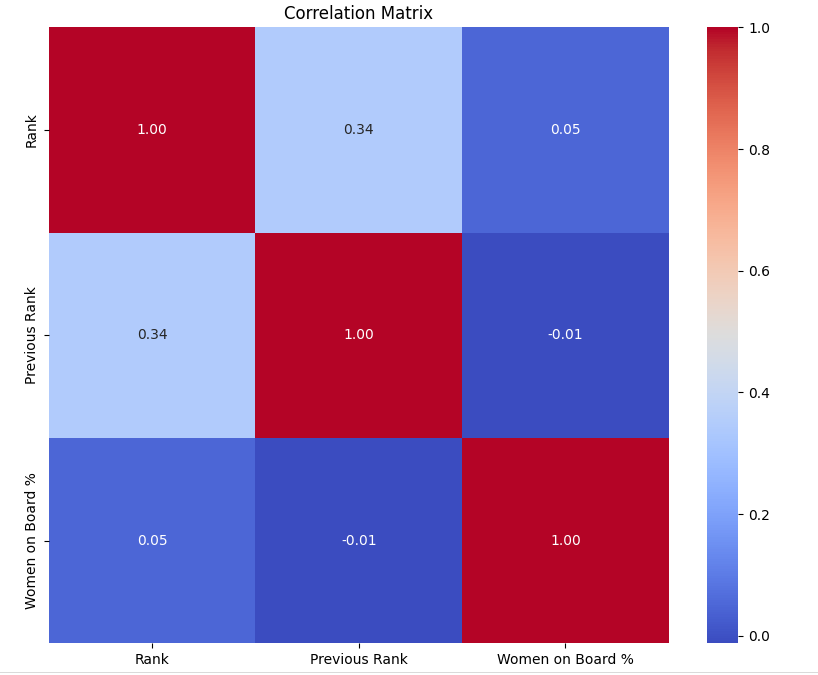
1. Remove missing values using any one suitable technique of the following:
   1. Ignore the tuple
   2. Fill in the missing value manually
   3. Use a global constant to fill in the missing value
   4. Use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value
   5. Use the attribute mean or median for all samples belonging to the same class as the given tuple
   6. Use the most probable value to fill in the missing value



1. Remove noisy data if any.

1. Perform data binning on the selected data.
2. Calculate Correlation between variables and create a correlation matrix.





**Note: You can use different data sets for each question as per the requirement and specify the links of the dataset used for each question or select an appropriate dataset that can work for all the above.**