Lab 4 – SECTION A, BATCH 2 Date: 17th Sept. 2022

Exercise 1 – Data Preprocessing, Regression

Using the given **CEREALS** dataset, perform data preprocessing and answer the following questions.

1) Create a table with the 5-number summary of all the numeric attributes.
2) For each of the numeric attributes (proteins up to vitamins), identify and replace all missing data (indicated with -1) with the arithmetic mean of the attribute.
3) Create a table with the 5-number summary of all the numeric attributes after treating missing values. Do you think the strategy used in dealing with missing values was effective?
4) For each of the numeric attributes (proteins up to vitamins), identify and replace all noisy data with the median of attribute.
5) Create a table with the 5-number summary of all the numeric attributes after treating noisy values. Do you think the strategy used in dealing with noisy values was effective?

Use the prepared or preprocessed data to answer the following:

6) Cross tabulate the type of cereal (hot vs cold) against the manufacturer
7) Which is the cereal with the best rating, worst rating?
8) Plot a side-by-side boxplot comparing the consumer rating of hot vs. cold cereals.
9) Is there a relation between sugars, calories, carbs, and fat?
10) Which manufacturers produce cereal with highest calories?
11) Use correlation tests and visualization to identify if the two variables calories and consumer rating associated?
12) Use correlation tests and visualization to identify if the two variables shelf and consumer rating associated?
13) Is there a relation between manufacturer and rating?
14) Which nutrients are essential for a good rating for a cereal?
15) Design a Linear regression model to predict the rating of a cereal based on top 3 related nutrients. Tabulate the accuracy of the model using an 80, 20 split.

Lab 5 – SECTION A, BATCH 1 Date: 21st Sept. 2022

Exercise 1 – Descriptive Analytics and Visualization using Matplotlib, Seaborn: (Cross tabulation, distributions, Multi-variate analysis, Various Plots)

Use the IPL datasets and answer the following:

1) Count the total number of matches conducted in the year 2008
2) Find the city name where maximum and minimum number of matches conducted.
3) Find total count of matches city wise.
4) Find the Team which is maximum and minimum toss winner.
5) Check the toss decision that the team has taken.
6) Count the total number of normal and tie matches.
7) Find the team names where the match result is tie.
8) Find the team name who won the match by highest runs.
9) Find the team name who won the match by lowest runs.
10) Find the players who was awarded "Player of the match" more than 3 times.
11) Find the player who was awarded as player of the match maximum times.
12) Find the Venue where the team won the match by highest runs.
13) Find the Venue where the team won the match by lowest runs.
14) Find the Umpires who did umpiring maximum times.
15) Find the Total matches played in each season
16) Find the Total runs in each season
17) No. of tosses won by each team
18) Visualize the Toss decision across seasons
19) Find the Dismissal Kind and Visualize using best fit graph
20) Find the Top 10 run scorers in IPL and Visualize using best fit graph
21) Visualize the Highest MOM award winners
22) Find Total Number of Played Matches by each team
23) Compare Total Played Matches vs Winning Matches vs Win Rate
24) Find the Distribution of Won the Matches
25) Ratio between Total Matches and Win Matches
26) What is the choice of each team after winning the toss?

Lab 6 – SECTION B, BATCH 3 Date: 22nd Sept. 2022

Exercise 1 – Time Series Analysis

Use the "employment.csv" data set and perform time series analysis and visualization through the following questions.

1. Convert datestamp column to a datetime object and Set the datestamp columns as the index of your DataFrame. Check if there are missing values in each column.
2. Generate a boxplot to find the distribution of unemployment rate for every industry.
3. Using line chart Visualize the unemployment rate of workers by industry.
4. Plot the monthly and yearly trends.
5. Apply time series decomposition to your dataset to visualize the trend and seasonality.
6. Visualize the seasonality of Agriculture, Health and Finance sector.
7. Visualize the seasonality of multiple time series and the correlation between each time series in the dataset.

Exercise 2: Text Analysis

Download the amazon_baby.zip file and answer the following:

1. Check the number of the reviews received for each product.
2. Check the products that have more than 15 reviews.
3. Find any missing review are present or not, if present remove those data.
4. Clean the data and remove the special characters and replace the contractions with its expansion by converting the uppercase character to lower case. Also, remove the punctuations.
5. Add the Polarity, length of the review, the word count and average word length of each review.
6. Visualize the distribution of the word count, review length, and polarity.
7. Visualize polarity considering the rating.
8. Visualize the count of the reviews of each rating available in the dataset.
9. List the Top 20 products based on the polarity.
10. Visualize to check whether the review length changes with rating.
11. Visualize the distribution of Top 25 Unigram, Bigram and Trigram.

Lab 7 – SECTION A, BATCH 1 Date: 19th OCT. 2022

Exer 1: Association Rule Mining

1. Use the "groceries.csv" dataset and answer the following:
2. How many transactions and items are there in the data set?
3. Prepare the data for finding association rules. Each transaction will contain a list of item in the transaction.
   [['citrus fruit', 'semi-finished bread', 'margarine', 'ready soups'],
   ['tropical fruit', 'yogurt', 'coffee'],…….
   ['whole milk']]

4. Use Python library *mlxtend* and convert the transactions into a format that can be used in the Apriori method for finding frequent itemsets.
   pip install mlxtend
   from mlxtend.preprocessing import TransactionEncoder
   from mlxtend.frequent_patterns import apriori, association_rules
5. Find top selling items with minimum support of 2%.
6. Find all frequent itemsets with minimum support of 5%.
7. Find all frequent itemsets of length 2 with minimum support of 2%.
8. Find the top 10 association rules with minimum support of 2%, sorted by confidence in descending order.
9. Find association rules with minimum support of 2% and lift of more than 1.0.

Lab 8 – SECTION A, BATCH 1 Date: 26th Oct. 2022

Exer 1: Collaborative Filtering

1. Read about the movielens dataset and write down a summary of metadata.

User-Based Similarity

2. Read the "ratings.csv" file and create a pivot table with index='userId', columns='movieId', values = "rating.
3. sklearn.metrics.pairwise_distances can be used to compute distance between all pairs of users. pairwise_distances() takes a metric parameter for what distance measure to use. Use cosine similarity for finding similarity among users. Use the following packages.

```
4. from sklearn.metrics import pairwise_distances
5. from scipy.spatial.distance import cosine, correlation
```

6. Find the 5 most similar user for user with user Id 25.
7. Use the "movies" dataset to find out the names of movies, user 1 and user 338 have watched in common and how they have rated each one of them.
8. Use the movies dataset to find out the common movie names between user 2 and user 338 with least rating of 4.0

Item-Based Similarity

9. Create a pivot table for representing the similarity among movies using correlation.
10. Find the top 5 movies which are similar to the movie "Godfather".

Lab 9 – SECTION A, BATCH 1 Date:2nd Nov. 2022

Exer 1: Clustering

Download the data set "*Online Retail.xlsx"* from
https://archive.ics.uci.edu/ml/datasets/online+retail

1. Read and write a summary of the metadata .
2. Select only the transactions that have occurred from 01/04/ 2011 and 09/12/2011 and create a dataset.
3. Calculate the RFM values for each customer (by customer id). RFM represents:
    - R (Recency) − Recency should be calculated as the number of months before he or she has made a purchase from the online store. If he/she made a purchase in the month of December 2011, then the Recency should be 0. If purchase is made in November 2011 then Recency should be 1 and so on and so forth.
    - F (Frequency) − Number of invoices by the customer from 01/04/ 2011 and 09/12/2011.
    - M (Monetary Value) − Total spend by the customer from 01/04/ 2011 and 09/12/2011.
4. Use the elbow method to identify how many customer segments exist, using the RFM values for each customer.
5. Create the customer segments with K-means algorithm by using number of clusters is suggested by elbow method.

    ```
    from sklearn.cluster import KMeans
    ```
6. Plot the clusters in a scatter plot and mark each segment differently using lmplot.
7. Print the cluster centers of each customer segment and explain them intuitively.
8. Create the customer segments with Agglomerative algorithm by using number of clusters is suggested by elbow method.

    ```
    from sklearn.cluster import AgglomerativeClustering
    ```
9. Visualize the clusters using the dendrogram.
10. Compare the clusters obtained using KMeans vs. Agglomeration.