DSE – 2159 DATA ANALYTICS LABORATORY

Lab 1 – SECTION B , BATCH 3 Date:10th Nov 2021

EXERCISE 1

Perform analysis on the NORTHWIND (COMBINED) data set using the pivot tables and charts in MS Excel.

1. Identify the top 5 and bottom 5 selling products in the company.
2. Identify the top 5 selling products and the salesmen who sell them.
3. Tabulate the total sales of each product, ship country wise.
4. Tabulate the total sales of "Sea weed and fish" , customer wise.
5. Tabulate the employee's region wise sales of products in each category.
6. Visualize the employees's region wise sales of products in each category using an appropriate chart.
7. Visualize the total sales of each product, customer wise with an appropriate chart.
8. Tabulate the total sales of each product, category-wise as a percentage of the entire sales.
9. Visualize the total sales of each product, category-wise as a percentage of the entire sales.
10. Summarize the sales for each product, year wise and visualize the same in an appropriate chart.

EXERCISE 2:

Data frame creation and manipulation

1. Create a data frame with details of 10 students and columns as Roll Number, Name, Gender, Marks1, Marks2, Marks3.
2. Create a new column with total marks
3. Find the lowest marks in Marks1
4. Find the Highest marks in Marks2
5. Find the average marks in Marks3
6. Find student name with highest average
7. Find how many students failed in Marks2 (<40)

EXERCISE 3:

- **Exer 2 – Data Analysis using mtcars**
  1. Find the car with the best mpg
  2. Find the car with the worst mpg
  3. Find the car with the best horsepower
  4. Find 5 number summary of displacement
  5. Find median horse power
  6. What is average mpg for manual vs. automatic cars
  7. Draw a histogram of miles per gallon
  8. Boxplot of mpg for each cylinder type
  9. Create a crosstab displaying count of automatic vs. manual cars
  10. Create a crosstab displaying count of "am vs cyl"
  11. What is the correlation between the weight of the car and mpg

DSE – 2159 DATA ANALYTICS LABORATORY

Lab 1 – SECTION B , BATCH 4 Date:11th Nov 2021

EXERCISE 1

Perform analysis on the NORTHWIND (COMBINED) data set using the pivot tables and charts in MS Excel.

1. Identify the top 5 and bottom 5 selling products in the company.
2. Identify the top 5 selling products and the salesmen who sell them.
3. Tabulate the total sales of each product, ship country wise.
4. Tabulate the total sales of "Chocolade" , customer wise.
5. Tabulate the customer's region wise sales of products in each category.
6. Visualize the ship's region wise sales of products in each category using an appropriate chart.
7. Visualize the total sales of each product, customer wise with an appropriate chart.
8. Tabulate the total sales of each product, category-wise as a percentage of the entire sales.
9. Visualize the total sales of each product, category-wise as a percentage of the entire sales.
10. Summarize the sales for each product, year wise and visualize the same in an appropriate chart.

EXERCISE 2:

Data frame creation and manipulation

1. Create a data frame with details of 10 students and columns as Roll Number, Name, Gender, Marks1, Marks2, Marks3.
2. Create a new column with total marks
3. Find the lowest marks in Marks1
4. Find the Highest marks in Marks2
5. Find the average marks in Marks3
6. Find student name with highest average
7. Find how many students failed in Marks2 (<40)

EXERCISE 3:

- **Exer 2 – Data Analysis using mtcars**

1. Find the car with the best mpg
2. Find the car with the worst mpg
3. Find the car with the best qsec
4. Find 5 number summary of mpg
5. Find median horse power
6. What is average mpg for manual vs. automatic cars
7. Draw a histogram of miles per gallon
8. Boxplot of mpg for each automatic vs manual cars
9. Create a crosstab displaying count of automatic vs. manual cars and carburators
10. Create a crosstab displaying count of "am vs cyl"
11. What is the correlation between the weight of the car and mpg

Lab 2 – SECTION B , BATCH 3 Date:17th Nov 2021

The data file bollywood.csv contains box office collection and social media promotion information about movies released in 2013−2015 period. Following are the columns and their descriptions. :

- SlNo
- Release Date
- MovieName – Name of the movie
- ReleaseTime – Mentions special time of release. LW (Long weekend), FS (Festive Season), HS (Holiday Season), N (Normal)
- Genre – Genre of the film such as Romance, Thriller, Action, Comedy, etc
- Budget – Movie creation budget
- BoxOfficeCollection – Box office collection
- YoutubeViews – Number of views of the YouTube trailers
- YoutubeLikes – Number of likes of the YouTube trailers
- YoutubeDislikes – Number of dislikes of the YouTube trailers

Use Python code to answer the following questions:

1. How many records are present in the dataset? 1
2. How many movies were released in each Release Time? Sort number of releases in Release Time in descending order. 1
3. Which genre had highest number of releases during the Festive Season? 1
4. How many movies in each genre got released in different release times like long weekend, festive season, etc. (Note: Do a cross tabulation between Genre and ReleaseTime.) 1
5. In which year were maximum number movie released? (Note: Extract a new column called year from ReleaseDate column.) 1
6. Which month of the year typically sees most releases of high budgeted movies, that is, movies with budget of 30 crore or more? 1
7. Which are the top 10 flop movies with minimum return on investment (ROI)? Calculate return on investment (ROI) as (BoxOfficeCollection – Budget) / Budget. 1
8. Do the movies have higher ROI if they get released on festive seasons or long weekend? Calculate the average ROI for different release times. 1
9. Is there a correlation between box office collection and YouTube likes? Is the correlation positive or negative? 1
10. Which genre of movies typically sees more YouTube views? Draw boxplots for each genre of movies to compare. 2
11. Which of the variables among Budget, BoxOfficeCollection, YoutubeView, YoutubeLikes, YoutubeDislikes are highly correlated? Note: Draw pair plot or heatmap. 2
12. During 2013−2015 period, highlight the genre of movies and their box office collection? Visualize with best fit graph. 2
13. During 2013−2015, find the number of movies released in every year. Also, visualize with best fit graph. 2
14. Find the distribution of movie budget for every Genre. 1
15. During 2013−2015, Visualize the number of YouTube likes and YouTube dislikes every year. Also, visualize with best fit graph. 2

Lab 2 – SECTION B , BATCH 4 Date:18 th Nov 2021

The data file bollywood.csv contains box office collection and social media promotion information about movies released in 2013−2015 period. Following are the columns and their descriptions. :

- SlNo
- Release Date
- MovieName – Name of the movie
- ReleaseTime – Mentions special time of release. LW (Long weekend), FS (Festive Season), HS (Holiday Season), N (Normal)
- Genre – Genre of the film such as Romance, Thriller, Action, Comedy, etc
- Budget – Movie creation budget
- BoxOfficeCollection – Box office collection
- YoutubeViews – Number of views of the YouTube trailers
- YoutubeLikes – Number of likes of the YouTube trailers
- YoutubeDislikes – Number of dislikes of the YouTube trailers

Use Python code to answer the following questions:

1. How many records are present in the dataset?
2. How many movies got released in each genre? Sort number of releases in each genre in descending order.
3. Which genre had highest number of releases?
4. How many movies in each genre got released in different release times like long weekend, festive season, etc. (Note: Do a cross tabulation between Genre and ReleaseTime.)
5. Which month of the year, maximum number movie releases are seen? (Note: Extract a new column called month from ReleaseDate column.)
6. Which month of the year typically sees most releases of high budgeted movies, that is, movies with budget of 25 crore or more?
7. Which are the top 10 movies with maximum return on investment (ROI)? Calculate return on investment (ROI) as (BoxOfficeCollection – Budget) / Budget.
8. Do the movies have higher ROI if they get released on festive seasons or long weekend? Calculate the average ROI for different release times.
9. Is there a correlation between box office collection and YouTube likes? Is the correlation positive or negative?
10. Which genre of movies typically sees more YouTube likes? Draw boxplots for each genre of movies to compare.
11. Which of the variables among Budget, BoxOfficeCollection, YoutubeView, YoutubeLikes, YoutubeDislikes are highly correlated? Note: Draw pair plot or heatmap.
12. During 2013−2015 period, highlight the genre of movies and their box office collection? Visualize with best fit graph.
13. Visualize the Budget and Box office collection based on Genre.
14. Find the distribution of movie budget for every Genre.
15. During 2013−2015, find the number of movies released in every year. Also, visualize with best fit graph.

Lab 3 – SECTION B , BATCH 3 Date:24th Nov 2021

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 upto 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 upto 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 calories 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 a 70 ,30 split.

Lab 3 – SECTION B , BATCH 4 Date:25th Nov 2021

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 upto 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 upto 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 calories 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 a 70 ,30 split.

Lab 4 – SECTION B , BATCH 3 Date: 1ST DEC 2021

Exer 1: Use the Cereals.csv data set and WEKA to answer the following questions:

Perform the following preprocessing steps

1. Replace missing value with mode or means.
2. Convert vitamins, shelf to nominal attribute.
3. Design a Linear regression model to predict the rating of a cereal based on all nutrients. Tabulate the accuracy of the model using
    a. 80 % , 20 % split
    b. Using cross validation.
4. Use select attributes tab and CfsSubsetEval to identify the top 5 related nutrients.
5. Design a Linear regression model to predict the rating of a cereal based on top 5 related nutrients. Tabulate the accuracy of the model using
    a. 80 % , 20 % split
    b. Using cross validation.
6. Discretise rating into 5 bins using the filters.
7. Design a Logistic regression model to predict the rating of a cereal based on top 5 related nutrients. Tabulate the accuracy of the model using
    a. 70 % , 30 % split
    b. Using cross validation.

Exer 2: Using the SUPERMARKET.ARFF data set and WEKA to answer the following questions:

1.Split the dataset into 2 datasets , 1 containing items and the other containing departments.

2.For the item data set, find the 5 most frequent itemsets ranked as per support.

3. Which are the top 5 selling items in the dataset ?

4. For the top selling item, find association rules with the item on the RHS of the rule. Tabulate the support, confidence and lift of the rule.

5. Find top 5 association rules for the department store. Tabulate the support, confidence and lift of the rule.

Lab 4 – SECTION B , BATCH 4 Date: 2nd DEC 2021

Exer 1: Use the Cereals.csv data set and WEKA to answer the following questions:

Perform the following preprocessing steps

8. Replace missing value with mode or means.
9. Convert vitamins, shelf to nominal attribute.
10. Design a Linear regression model to predict the rating of a cereal based on all nutrients. Tabulate the accuracy of the model using
    c. 80 % , 20 % split
    d. Using cross validation.
11. Use select attributes tab and CfsSubsetEval to identify the top 5 related nutrients.
12. Design a Linear regression model to predict the rating of a cereal based on top 5 related nutrients. Tabulate the accuracy of the model using
    c. 80 % , 20 % split
    d. Using cross validation.
13. Discretise rating into 5 bins using the filters.
14. Design a Logistic regression model to predict the rating of a cereal based on top 5 related nutrients. Tabulate the accuracy of the model using
    c. 70 % , 30 % split
    d. Using cross validation.

Exer 2: Using the SUPERMARKET.ARFF data set and WEKA to answer the following questions:

1.Split the dataset into 2 datasets , 1 containing items and the other containing departments.

2.For the item data set, find the 5 most frequent itemsets ranked as per support.

3. Which are the top 5 selling items in the dataset ?

4. For the top selling item, find association rules with the item on the RHS of the rule. Tabulate the support, confidence and lift of the rule.

5. Find top 5 association rules for the department store. Tabulate the support, confidence and lift of the rule.

Lab 5 – SECTION B , BATCH 3 Date: 8th Dec 2021

Exer 1:

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.

Exer 2:

Use the "groceries.csv" dataset and answer the following:

1. How many transactions and items are there in the data set?
2. Prepare the data for finding association rules. Each transaction will contain a list of item in each transaction.
   [['citrus fruit', 'semi-finished bread', 'margarine', 'ready soups'],
   ['tropical fruit', 'yogurt', 'coffee'],…….
   ['whole milk']]
3. 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
4. Find top selling items with minimum support of 2%.
5. Find all frequent itemsets with minimum support of 5%.
6. Find all frequent itemsets of length 2 with minimum support of 2%.
7. Find the top 10 association rules with minimum support of 2%, sorted by confidence in descending order.
8. Find association rules with minimum support of 2% and lift of more than 1.0.

Lab 5 – SECTION B , BATCH 4 Date: 9th Dec 2021

Exer 1:

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.

Exer 2:

Use the "groceries.csv" dataset and answer the following:

1. How many transactions and items are there in the data set?
2. Prepare the data for finding association rules. Each transaction will contain a list of item in each transaction.

[['citrus fruit', 'semi-finished bread', 'margarine', 'ready soups'],

['tropical fruit', 'yogurt', 'coffee'],…….

['whole milk']]

3. 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

4. Find top selling items with minimum support of 2%.
5. Find all frequent itemsets with minimum support of 5%.
6. Find all frequent itemsets of length 2 with minimum support of 2%.
7. Find the top 10 association rules with minimum support of 2%, sorted by confidence in descending order.
8. Find association rules with minimum support of 2% and lift of more than 1.0.

Lab 6 – SECTION B , BATCH 3 Date: 21ST Dec 2021

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 10.
7. Use the "movies" dataset to find out the names of movies, user 2 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 6 – SECTION B , BATCH 4 Date: 23rd Dec 2021

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 7 – SECTION B , BATCH 3 Date: 28th Dec 2021

Exer 1: Clustering

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

a. Read and write a summary of the metadata .
b. Select only the transactions that have occurred from 01/04/ 2011 and 09/12/2011 and create a dataset.
c. Calculate the RFM values for each customer (by customer id). RFM represents:

2. 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.
3. F (Frequency) − Number of invoices by the customer from 01/04/ 2011 and 09/12/2011.
4. M (Monetary Value) − Total spend by the customer from 01/04/ 2011 and 09/12/2011.
    a. Use the elbow method to identify how many customer segments exist, using the RFM
5. values for each customer.
    a. Create the customer segments with K-means algorithm by using number of clusters is suggested by elbow method.
6. `from sklearn.cluster import KMeans`
    a. Plot the clusters in a scatter plot and mark each segment differently using lmplot.
    b. Print the cluster centers of each customer segment and explain them intuitively.
    c. Create the customer segments with Agglomerative algorithm by using number of clusters is suggested by elbow method.
7. `from sklearn.cluster import AgglomerativeClustering`
    a. Visualize the clusters using the dendrogram.
    b. Compare the clusters obtained using KMeans vs. Agglomeration.

Exer 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 B , BATCH 4 Date: 30th Dec 2021

Exer 1: Clustering

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

a. Read and write a summary of the metadata .
b. Select only the transactions that have occurred from 01/04/ 2011 and 09/12/2011 and create a dataset.
c. Calculate the RFM values for each customer (by customer id). RFM represents:

2. 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.
3. F (Frequency) − Number of invoices by the customer from 01/04/ 2011 and 09/12/2011.
4. M (Monetary Value) − Total spend by the customer from 01/04/ 2011 and 09/12/2011.
    a. Use the elbow method to identify how many customer segments exist, using the RFM
5. values for each customer.
    a. Create the customer segments with K-means algorithm by using number of clusters is suggested by elbow method.
6. `from sklearn.cluster import KMeans`
    a. Plot the clusters in a scatter plot and mark each segment differently using lmplot.
    b. Print the cluster centers of each customer segment and explain them intuitively.
    c. Create the customer segments with Agglomerative algorithm by using number of clusters is suggested by elbow method.
7. `from sklearn.cluster import AgglomerativeClustering`
    a. Visualize the clusters using the dendrogram.
    b. Compare the clusters obtained using KMeans vs. Agglomeration.

Exer 2: Text Analysis

Download the amazon_baby.zip file and answer the following:

12. Check the number of the reviews received for each product.
13. Check the products that have more than 15 reviews.
14. Find any missing review are present or not, If present remove those data.
15. 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.
16. Add the Polarity, length of the review, the word count and average word length of each review.
17. Visualize the distribution of the word count, review length, and polarity.
18. Visualize polarity considering the rating.
19. Visualize the count of the reviews of each rating available in the dataset.
20. List the Top 20 products based on the polarity.
21. Visualize to check whether the review length changes with rating.
22. Visualize the distribution of Top 25 Unigram, Bigram and Trigram.