DSE – 2159 DATA ANALYTICS LABORATORY

Lab 1 – SECTION A , BATCH 1 Date:8th 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 "Boston Crab Meat" , customer wise.
5. Tabulate the customer's region wise sales of products in each category.
6. Visualize the customer's region wise sales of products in each category using an appropriate chart.
7. Visualize the total sales of each product, employee 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 A , BATCH 2 Date:10th Nov 2021

EXERCISE 1

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

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

EXERCISE 2:

Data frame creation and manipulation

8. Create a data frame with details of 10 students and columns as Roll Number, Name, Gender, Marks1, Marks2, Marks3.
9. Create a new column with total marks
10. Find the lowest marks in Marks1
11. Find the Highest marks in Marks2
12. Find the average marks in Marks3
13. Find student name with highest average
14. Find how many students failed in Marks2 (<40)

EXERCISE 3:

- **Exer 2 – Data Analysis using mtcars**
  12. Find the car with the best mpg
  13. Find the car with the worst mpg
  14. Find the car with the best horsepower
  15. Find 5 number summary of displacement
  16. Find median horse power
  17. What is average mpg for manual vs. automatic cars
  18. Draw a histogram of miles per gallon
  19. Boxplot of mpg for each cylinder type
  20. Create a crosstab displaying count of automatic vs. manual cars
  21. Create a crosstab displaying count of "am vs cyl"
  22. What is the correlation between the weight of the car and mpg

Lab 2 – SECTION A , BATCH 1 Date:15 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 2 – SECTION A , BATCH 2 Date:17 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 Release Time? Sort number of releases in each Release Time 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 were least 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 low budgeted movies, that is, movies with budget less than 25 crore ?
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 holiday season? 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 views? 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 Box office Collection for every Genre.
15. During 2013−2015, Visualize the number of Youtube views, Youtube likes released in every year. Also, visualize with best fit graph.

Lab 2 – SECTION A , BATCH 1 Date:22 nd 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 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 a 80 ,20 split.