CSC-593 Data Mining & Data Warehousing (Lab)

A LAB REPORT
SUBMITTED FOR THE PARTIAL FULFILLMENT OF AWARD OF THE DEGREE OF
<Name of Programme>

By
<Name of Student in Capital><Roll>

Faculty In-Charge
Atul Nag, Associate Professor
KISS-DU

KIIT SCHOOL OF SOCIAL SCIENCES(KSSS)
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY (KIIT)
Bhubaneswar, Odisha
December 2022

TABLE OF CONTENTS

Exercise – 1 (Microsoft Excel Basics)

Aim: To Learn the basics of Microsoft Excel

Workout 1: change the name of the current worksheet

Procedure: To change the name of the current worksheet, we take the following steps:

Step 1. Step 1. Right-click on the worksheet tab named Sheet1

Step 2. Step 2. Select the Rename option

Step 3. Step 3. Enter Nowlin to rename the worksheet and press Enter

Workout 2: create a copy of the newly renamed Nowlin worksheet

Procedure: create a copy of the newly renamed Nowlin worksheet by following these steps:

Step 1. Step 1. Right-click the worksheet tab named Nowlin

Step 2. Step 2. Select the Move or Copy… option

Step 3. Step 3. When the Move or Copy dialog box appears, select the checkbox for Create a Copy, and click OK

Workout 3: create additional worksheets

Procedure: To create additional worksheets follow these steps:

Step 1. Step 1. Right-click on the tab of any existing worksheet

Step 2. Step 2. Select Insert…

Step 3. Step 3. When the Insert dialog box appears, select Worksheet from the General area, and click OK

Workout 4: save the workbook

Procedure: To save the workbook using the filename Nowlin, we perform the following steps:

Step 1. Click the File tab on the Ribbon

Step 2. Click Save in the list of options

Step 3. Select This PC under Save As, and click Browse

Step 4. When the Save As dialog box appears:

a. Select the location where you want to save the file

b. Enter the filename Nowlin in the File name: box Click Save

Workout 5: calculate the average monthly revenue and cost at Nowlin

Data File: ***NowlinFinancial***

Procedure: to calculate the average monthly revenue and cost at Nowlin. To do so, we execute the following steps:

Step 1. Select cell B17 in the DATAfile NowlinFinancial

Step 2. Click the Insert Function button fx . Select Statistical in the Or select a category: box Select AVERAGE from the Select a function: options

Step 3. When the Function Arguments dialog box appears: Enter B2:B13 in the Number1 box Click OK

Step 4. Repeat Steps 1 through 3 for the cost data in column C

Workout 5: calculating the monthly profit at Nowlin Plastics

Data File: ***NowlinFinancial***

Procedure: calculating the monthly profit at Nowlin Plastics using the following steps:

Step 1. Enter the label Profit in cell D1 and press the Bold button in the Font group of the Home tab

Step 2. Enter the formula =B2-C2 in cell D2

Step 3. Copy the formula from cell D2 by selecting cell D2 and clicking Copy from the Clipboard group of the Home tab

Step 4. Select cells D3:D13 Step 5. Paste the formula from cell D2 by clicking Paste from the Clipboard group of the Home tab

Workout 6: calculate the after-tax profit for Nowlin in each month

Data File: ***NowlinFinancial***

Procedure: calculate the after-tax profit for Nowlin in each month by using the following steps:

Step 1. Enter the label After-Tax Profit in cell E1 and press the Bold button in the Font group of the Home tab.

Step 2. Enter the formula =D2*(1-$B$19) in cell E2

Step 3. Copy the formula from cell E2 by selecting cell E2 and clicking Copy from the Clipboard group of the Home tab

Step 4. Select cells E3:E13 Step 5. Paste the formula from cell E2 by clicking Paste from the Clipboard group of the Home tab

Exercise – 2 (Sort, Filter & Conditional Formatting)

Aim: To Learn the sort, filter and conditional formatting methods of Microsoft Excel

Workout 1: calculate the percentage in sales

Data File: ***1_Top20Cars2019***

Procedure: The per cent change in sales for each model from February 2018 to February 2019 has been calculated.:

Step 1. This is done by entering the formula =(D2-E2)/E2 in cell F2 and then copying the contents of this cell to cells F3 to F20.

Workout 2: sort the given dataset on 2018 sales

Data File: ***1_Top20Cars2019***

Procedure: To do this, we use Excel's Sort function, as shown in the following steps.

Step 1. Select cells A1:F21

Step 2. Click the Data tab in the Ribbon

Step 3. Click Sort in the Sort & Filter group

Step 4. Select the check box for My data has headers

Step 5. In the first Sort by dropdown menu, select Sales (February 2018)

Step 6. In the Order dropdown menu, select Largest to Smallest

Step 7. Click OK

Workout 3: Filter the given dataset on sales to see only the models made by Nissan

Data File: ***1_Top20Cars2019***

Procedure: To only in seeing the sales of models made by Nissan. We can do this using Excel's Filter function:

Step 1. Select cells A1:F21

Step 2. Click the Data tab in the Ribbon

Step 3. Click Filter in the Sort & Filter group

Step 4. Click on the Filter Arrow in column B, next to Manufacturer

Step 5. If all choices are checked, you can easily deselect all choices by unchecking (Select All). Then select only the check box for Nissan.

Step 6. Click OK

Workout 4: Conditional formatting in Excel can make it easy to identify data that satisfy certain conditions in a data set.

Data File: ***1_Top20Cars2019***

Procedure: To quickly identify the automobile models for which sales had decreased from February 2018 to February 2019. We can quickly highlight these models::

Step 1. Starting with the original data, select cells F1:F21

Step 2. Click the Home tab in the Ribbon

Step 3. Click Conditional Formatting in the Styles group

Step 4. Select Highlight Cells Rules, and click Less Than . . . from the dropdown menu

Step 5. Enter 0% in the Format cells that are LESS THAN: box

Step 6. Click OK

Instead of highlighting only models with decreasing sales, we could instead choose Data Bars from the Conditional Formatting dropdown menu in the Styles group of the Home tab in the Ribbon.

Exercise – 3 (Frequency Distributions of Categorical Data)

Aim: To Learn the create frequency distributions from categorical data

Workout 1: To calculate the frequency, relative frequency and percent frequency of categorical observations

Data File: ***2_SoftDrinks***

Procedure: to calculate the frequency of categorical observations occurring in a data set using the COUNTIF function.

Step 1. In cell E2, we enter the formula =COUNTIF($A$2:$B$26, D2), where A2:B26 is the range for the sample data, and D2 is the bin (Coca-Cola) that we are trying to match.

Step 2. In cell E8, we enter the formula =SUM(E2:E6) and add the data label in D8 as Total

Step 3. To calculate the relative frequency, we enter the Data column label – Relative Frequency on F1

a. In F2, we enter the formula =E2/E8 and the change the formatting of the cell to "General"

Step 4. To calculate the Percent Frequency, we enter the data column label – Percent Frequency in G1

a. In G2, we enter the formula =F2*100

Exercise – 4 (Frequency Distributions of Quantitative Data)

Aim: To Learn the creation of frequency distributions from quantitative data and generate a histogram

Workout 1: To calculate the frequency distribution of quantitative observations

Data File: ***3_AuditTime***

Procedure: to calculate the frequency of quantitative observations occurring in the given data set.

Step 1. The upper limits of the defined bins in the cells C2:C6

We can use the FREQUENCY function in Excel to count the number of observations in each bin.

Step 2. Select cells D2:D6

Step 3. Type the formula =FREQUENCY(A2:A21, C2:C6). The range A2:A21 defines the data set, and the range C2:C6 defines the bins.

Step 4. Press CTRL+SHIFT+ENTER after typing the formula in Step 3

A common graphical presentation of quantitative data is a **histogram**.

This graphical summary can be prepared for data previously summarized in either a frequency, a relative frequency, or a percent frequency distribution.

Histograms can be created in Excel using the Data Analysis ToolPak.

Workout 2: To create histogram using the Data Analysis ToolPak

Data File: ***3_AuditTime***

Procedure: Histograms can be created in Excel using the Data Analysis ToolPak.

Step 1. Step 1. Click the Data tab in the Ribbon

Step 2. Step 2. Click Data Analysis in the Analyze group

Step 3. Step 3. When the Data Analysis dialog box opens, choose Histogram from the list of Analysis Tools, and click OK

a. In the Input Range: box, enter A2:A21

b. In the Bin Range: box, enter C2:C6

c. Under Output Options:, select New Worksheet Ply: Select the check box for Chart Output
Click OK

We have also removed the gaps between the columns in the histogram in Excel to match the traditional format of histograms.

To remove the gaps between the columns in the histogram created by Excel, follow these steps:

Step 4. Step . Right-click on one of the columns in the histogram. Select Format Data Series

Step 5. When the Format Data Series pane opens, click the Series Options button, Set the Gap Width to 0%

Exercise – 4 (Measures of Location & Variability)

Aim: To Learn to calculate the measures of location (mean, median and mode) and variability (range, variance, standard deviation and coefficient of variation)

Workout 1: To calculate the mean of a given dataset

Data File: ***4_HomeSales***

Procedure: For calculating the mean of the Home Sales data

Step 1. The value for the mean in cell E2 is calculated using the formula $=AVERAGE(B2{:}B13)$.

Workout 2: To calculate the median of a given dataset

Data File: ***4_HomeSales***

Procedure: For calculating the median of the Home Sales data, the median of a data set can be found in Excel using the function MEDIAN

Step 1. The value for the median in cell E3 is found using the formula =MEDIAN(B2:B13).

Workout 3: To calculate the mode of a given dataset

Data File: ***4_HomeSales***

Procedure: For calculating the mode of the Home Sales data. The Excel MODE.SNGL function will return only a single most-often-occurring value. For multimodal distributions, we must use the MODE.MULT command in Excel to return more than one mode.

The given dataset is multimodal. To find both modes in Excel, we take these steps:

Step 1. Select cells E4 and E5

Step 2. Type the formula =MODE.MULT(B2:B13)

Step 3. Press CTRL+SHIFT+ENTER after typing the formula in Step 2.

Workout 4: To calculate the range of a given dataset

Data File: *4_HomeSales*

Procedure: The range can be found by subtracting the smallest value from the largest value in a data set.

Step 1. The range can be calculated in Excel using the MAX and MIN functions. The range value in cell E7 calculates the range using the formula $=MAX(B2{:}B13) - MIN(B2{:}B13)$.

Workout 5: To calculate the variance of a given dataset

Data File: *4_HomeSales*

Procedure: The variance is based on the deviation from the mean, which is the difference between the value of each observation and the mean..

Step 1. In Excel, you can find the variance for sample data using the VAR.S function The variance in cell E8 is calculated using the formula $=VAR.S(B2{:}B13)$.

Workout 6: To calculate the standard deviation & coefficient of variation of a given dataset

Data File: *4_HomeSales*

Procedure: The standard deviation is defined to be the positive square root of the variance.

Step 1. Sample standard deviation of the home sales data, which can be calculated using Excel's STDEV.S function. The sample standard deviation in cell E9 is calculated using the formula $=STDEV.S(B2{:}B13)$.

Step 2. The coefficient of variation for the home sales data is calculated in cell E11 using the formula $=E9/E2$, which divides the standard deviation by the mean.

Exercise – 5 (Analysis of Distributions)

Aim: To Learn to calculate the percentile, quartiles, z-scores and construct boxplot.

Workout 1: To calculate the 85th percentile of a given dataset

Data File: ***4_HomeSales***

Procedure: The pth percentile can also be calculated in Excel using the function PERCENTILE.EXC.

Step 1. The value in cell E13 is calculated using the formula $=PERCENTILE.EXC(B2{:}B13,0.85)$

a. B2:B13 defines the data set for which we are calculating a percentile, and 0.85 defines the percentile of interest.

Workout 2: To calculate the different quartiles of a given dataset

Data File: ***4_HomeSales***

Procedure: These division points are referred to as the quartiles and are defined as follows:

- Q1 = first quartile, or 25th percentile
- Q2 = second quartile, or 50th percentile (also the median)
- Q3 = third quartile, or 75th percentile

A quartile can be computed in Excel using the function QUARTILE.EXC. For the calculations for first, second, and third quartiles for the home sales data

Step 1. The formula used in cell E15 is $=QUARTILE.EXC(B2{:}B13,1)$. The range B2:B13 defines the data set, and 1 indicates that we want to compute the first quartile.

Step 2. Cells E16 and E17 use similar formulas to compute the second and third quartiles.

Workout 3: To calculate the z-scores of each data item in a given dataset

Data File: *4_HomeSales*

Procedure: The z-score is often called the standardized value. The z-score can be calculated in Excel using the function STANDARDIZE. To calculate the z-scores, we must provide the mean and standard deviation for the data set in the arguments of the STANDARDIZE function.

Step 1. The z-score in cell C2 is calculated with the formula $=STANDARDIZE(B2, \$B\$15, \$B\$16)$, where cell B15 contains the mean of the home sales data and cell B16 contains the standard deviation of the home sales data.

Step 2. We can then copy and paste this formula into cells C3:C13.

Workout 3: To construct a boxplot for a given dataset

Data File: *4_HomeSales*

Procedure: A boxplot is a graphical summary of the distribution of data. A boxplot is developed from the quartiles for a data set. we will create a boxplot for a single variable using the HomeSales file.

Step 1. Select cells B1:B13

Step 2. Click the Insert tab on the Ribbon

a. Click the Insert Statistic Chart button in the Charts group

b. Choose the Box and Whisker chart from the drop-down menu

Excel orients the boxplot vertically, and by default, it also includes a marker for the mean.

Data File: *5_HomeSalesComparison*

Procedure: we will use the HomeSalesComparison file to create boxplots in Excel for multiple variables.

Step 1. Step 1. Select cells B1:F11

Step 2. Step 2. Click the Insert tab on the Ribbon

a. Click the Insert Statistic Chart button in the Charts group

b. Choose the Box and Whisker chart from the drop-down menu

Excel again orients the boxplot vertically.

The different selling locations are shown in the Legend at the top of the figure, and different colours are used for each boxplot.

Exercise – 6 (Measures of Association Between Two Variables)

Aim: To Learn to calculate covariance & correlation coefficient.

Workout 1: To calculate covariance for two variables whose measure of association is to be measured

Data File: ***6_BottledWater***

Procedure: The sample covariance can also be calculated in Excel using the COVARIANCE.S

Step 1. The covariance is calculated in cell B17 using the formula $=COVARIANCE.S(A2{:}A15, B2{:}B15)$.

a. A2:A15 defines the range for the x variable (high temperature)

b. B2:B15 defines the range for the y variable (sales of bottled water).

Workout 2: To calculate the correlation coefficient for two variables whose measure of association is to be measured

Data File: ***6_BottledWater***

Procedure: We can compute correlation coefficients using the Excel function CORREL.

Step 1. The correlation coefficient is computed in cell B18 for the sales of bottled water using the formula $=CORREL(A2{:}A15, B2{:}B15)$,

a. A2:A15 defines the range for the x variable

b. B2:B15 defines the range for the y variable.

Exercise – 7 (Crosstabulation)

Aim: To Learn to create crosstabulation using Pivot Tables.

Workout 1: To create crosstabulation to summarize given dataset

Data File: ***8_Restaurant***

Procedure: A crosstabulation in Microsoft Excel is known as a PivotTable.

Step 1. Click the Insert tab on the Ribbon

Step 2. Click PivotTable in the Tables group

Step 3. When the Create PivotTable dialog box appears:

a. Choose Select a table or range

b. Enter A1:D301 in the Table/Range: box

c. Select New Worksheet as the location for the PivotTable Report

d. Click OK

The following steps show how to use Excel's PivotTable Field List to assign the Quality Rating field to the rows, the Meal Price ($) field to the columns, and the Restaurant field to the body of the PivotTable report.

Step 4. In the PivotTable Fields task pane, go to Drag fields between areas below:

a. Drag the Quality Rating field to the ROWS area

b. Drag the Meal Price ($) field to the COLUMNS area

c. Drag the Restaurant field to the VALUES area

Step 5. Click on Sum of Restaurant in the VALUES area

Step 6. Select Value Field Settings from the list of options

Step 7. When the Value Field Settings dialog box appears:

a. Under Summarize value field by, select Count

b. Click OK

To complete the PivotTable, we need to group the columns representing meal prices and place the row labels for a quality rating in the proper order:

Step 8. Right-click in cell B4 or any other cell containing a meal price column label

Step 9. Select Group from the list of options

Step 10. When the Grouping dialog box appears:

a. Enter 10 in the Starting at: box

b. Enter 49 in the Ending at: box

c. Enter 10 in the By: box

d. Click OK

Step 11. Right-click on "Excellent" in cell A5

Step 12. Select Move and click Move "Excellent" to End

Workout 2: To create frequency distribution from pivot tables for given dataset

Data File: ***8_Restaurant***

Procedure: We can also use a PivotTable to create per cent frequency distributions, as shown in the following steps:

Step 1. To invoke the PivotTable Fields task pane, select any cell in the pivot table

Step 2. In the PivotTable Fields task pane, click the Count of Restaurant in the VALUES area

Step 3. Select Value Field Settings . . . from the list of options

Step 4. When the Value Field Settings dialog box appears, click the tab for ShowValues As

Step 5. In the Show values as area, select % of Grand Total from the drop-down menu

Step 6. Click OK

Workout 3: To modify created pivot table to display new information

Data File: ***8_Restaurant***

Procedure: PivotTables in Excel are interactive, and they may be used to display statistics other than a simple count of items. As an illustration, we can easily modify the PivotTable to display summary information on wait times instead of meal prices.

Step 1. To invoke the PivotTable Fields task pane, select any cell in the pivot table

Step 2. In the PivotTable Fields task pane, click the Count of Restaurant field in the VALUES area

a. Select Remove Field

Step 3. Drag the Wait Time (min) to the VALUES area

Step 4. Click on Sum of Wait Time (min) in the VALUES area

Step 5. Select Value Field Settings… from the list of options

Step 6. When the Value Field Settings dialog box appears:

a. Under Summarize value field by, select Average

b. Click Number Format

c. In the Category: area, select Number

d. Enter 1 for Decimal places:

e. Click OK

f. When the Value Field Settings dialog box reappears, click OK

We can also examine only a portion of the data in a PivotTable using the Filter option in Excel.

Exercise – 8 (Scatter Charts)

Aim: To Learn to create Scatter Charts to analyse relationship between variables.

Workout 1: To create Scatter Charts for given dataset

Data File: ***9_Electronics***

Procedure: A scatter chart is a graphical presentation of the relationship between two quantitative variables. Consider the advertising/sales relationship for an electronics store in San Francisco in the given dataset.

Step 1. Select cells B2:C11

Step 2. Click the Insert tab in the Ribbon

Step 3. Click the Insert Scatter (X,Y) or Bubble Chart button in the Charts group

Step 4. When the list of scatter chart subtypes appears, click the Scatter button

Step 5. Click the Design tab under the Chart Tools Ribbon

Step 6. Click Add Chart Element in the Chart Layouts group

a. Select Chart Title, and click Above Chart

b. Click on the text box above the chart, and replace the text with Scatter Chart for the San Francisco Electronics Store

Step 7. Click Add Chart Element in the Chart Layouts group

a. Select Axis Title, and click Primary Horizontal

b. Click on the text box under the horizontal axis, and replace "Axis Title" with Number of Commercials

Step 8. Click Add Chart Element in the Chart Layouts group

a. Select Axis Title, and click Primary Vertical

b. Click on the text box next to the vertical axis, and replace "Axis Title" with Sales ($100s)

Step 9. Right-click on one of the horizontal grid lines in the body of the chart, and

a. click Delete

Step 10. Right-click on one of the vertical grid lines in the body of the chart, and

a. click Delete

A trendline is a line that provides an approximation of the relationship between the variables. To add a linear trendline using Excel, we use the following steps:

Step 1. Right-click on one of the data points in the scatter chart and select

a. Add Trendline…

Step 2. When the Format Trendline task pane appears, select Linear under Trendline Options

Exercise – 9 (Line, Bar and Column Charts)

Aim: To Learn to create Line, Bar and Column Charts to analyse relationship between variables.

Workout 1: To create Line Charts for given dataset

Data File: ***10_Kirkland***

Procedure: Line charts are very useful for time series data collected over a period (minutes, hours, days, years, etc.) To create the line chart in Excel, we follow these steps:

Step 1. Select cells A2:B13

Step 2. Click the Insert tab on the Ribbon

Step 3. Click the Insert Line Chart button in the Charts group

Step 4. When the list of line chart subtypes appears, click the Line with Markers button under 2-D Line (This creates a line chart for sales with a basic layout and minimum formatting)

Step 5. Select the line chart that was just created to reveal the Chart Buttons

Step 6. Click the Chart Elements button

a. Select the check boxes for Axes, Axis Titles, and Chart Title

b. Deselect the check box for Gridlines

c. Click on the text box next to the vertical axis, and replace “Axis Title” with Sales ($100s)

d. Click on the text box next to the horizontal axis and replace “Axis Title” with Month

e. Click on the text box above the chart, and replace “Sales ($100s)” with Line Chart for Monthly Sales Data

Workout 2: To create Bar and Column Charts for given dataset

Data File: ***12_AccountsManaged***

Procedure: Bar charts use horizontal bars to display the magnitude of the quantitative variable:

Step 1. Select cells A2:B9

Step 2. Click the Insert tab on the Ribbon

Step 3. Click the Insert Column or Bar Chart button in the Charts group

Step 4. When the list of bar chart subtypes appears:

a. Click the Clustered Bar button in the 2-D Bar section

Step 5. Select the bar chart that was just created to reveal the Chart Buttons.

Step 6. Step 6. Click the Chart Elements button

a. Select the check boxes for Axes, Axis Titles, and Chart Title

b. Deselect the check box for Gridlines

c. Click on the text box next to the horixontal axis, and replace “Axis Title” with Accounts Managed

d. Click on the text box next to the vertical axis, and replace “Axis Title” with Manager

e. Click on the text box above the chart, and replace “Chart Title” with Bar Chart of Accounts Managed

We can make this bar chart even easier to read by ordering the results by the number of accounts managed. We can do this with the following steps:

Step 1. Select cells A1:B9

Step 2. Right-click any of the cells A1:B9

a. Select Sort

b. Click Custom Sort

Step 3. When the Sort dialog box appears:

a. Make sure that the check box for My data has headers is checked

b. Select Accounts Managed in the Sort by box under Column

c. Select Smallest to Largest under Order

d. Click OK

In the completed bar chart in Excel, we can easily compare the relative number of accounts managed for all managers. However, note that it is difficult to interpret from the bar chart exactly how many accounts are assigned to each manager. If this information is necessary, these data are better presented as a table or by adding data labels to the bar chart, which is created in Excel using the following steps:

Step 1. Select the chart to reveal the Chart Buttons

Step 2. Click the Chart Elements button

a. Select the check box for Data Labels

Exercise – 10 (Heat Maps and Stacked Column Charts)

Aim: To Learn to create Heat Maps and Stacked Column Charts to analyse relationship between variables.

Workout 1: To create HeatMaps for given dataset

Data File: ***14_SameStoreSales***

Procedure: A heat map is a two-dimensional graphical representation of data that uses different shades of colour to indicate magnitude. Can be created in Excel

Step 1. Select cells B2:M17

Step 2. Click the Home tab on the Ribbon

Step 3. Click Conditional Formatting in the Styles group

a. Select Color Scales and click on Blue–White–Red Color Scale.

To add the sparklines in column N, we use the following steps:

Step 4. Select cell N2

Step 5. Click the Insert tab on the Ribbon

Step 6. Click Line in the Sparklines group

Step 7. When the Create Sparklines dialog box appears:

a. Enter B2:M2 in the Data Range: box

b. Enter N2 in the Location Range: box

c. Click OK

Step 8. Copy cell N2 to N3:N17

Workout 1: To create Stacked-Column Charts for given dataset

Data File: ***11_KirklandRegional***

Procedure: To create the stacked-column chart for the given dataset, we use the following steps:

Step 1. Select cells A2:C14

Step 2. Click the Insert tab on the Ribbon

Step 3. In the Charts group, click the Insert Column or Bar Chart button Select Stacked Column under 2-D Column.

An alternative chart for these same data is called a clustered-column (or clustered-bar) chart. It is created in Excel following the same steps but selecting Clustered Column under the 2-D Column in Step 3.