­­­

|  |
| --- |
| ARDEN UNIVERSITY |
| BSC (HONS) COMPUTING |
| DATA ANALYSIS AND VISUALISATION |
| DATA ANALYSIS AND VISUALISATION |
| STU |
|  |
| 1384 |

Contents

[1.TASK 1 3](#_Toc158715966)

[1.1 DATA CLEANING 3](#_Toc158715967)

[1.2 Remove duplicate values 4](#_Toc158715968)

[1.3 Handling missing values 5](#_Toc158715969)

[2.TASK 2 7](#_Toc158715970)

[2.1 DESCRIPTIVE ANALYSIS 7](#_Toc158715971)

[2.2 SKEWNESS 8](#_Toc158715972)

[2.3 CORRELATION 8](#_Toc158715973)

[2.4 HISTOGRAM 9](#_Toc158715974)

[2.5 HYPOTHESIS TESTING 10](#_Toc158715975)

[2.6 REPORT 12](#_Toc158715976)

[TASK 3 12](#_Toc158715977)

[Largest and the Smallest grape 13](#_Toc158715978)

[Largest and Smallest perimeter 14](#_Toc158715979)

# 1.TASK 1

# 1.1 DATA CLEANING

The practice of repairing or eliminating inaccurate, corrupted, improperly formatted, duplicate, or incomplete data from a dataset is known as data cleaning. It is common for data to be mislabeled or duplicated when integrating different data sources. Even though results and algorithms appear correct, they are unreliable if the data is inaccurate. Since the procedures will differ depending on the dataset, there is no one set method that can be used to prescribe the precise steps in the data cleaning process. However, in order to ensure that you are cleaning data correctly each and every time, you must create a template for your procedure.

How to clean data

Step 1: Remove duplicate or irrelevant observations

Step 2: Fix structural errors

Step 3: Handle missing data

1.2 Finding duplicates

1. Select the range A1:C10.

2. On the Home tab, in the Styles group, click Conditional Formatting.

3. Click Highlight Cells Rules, Duplicate Values.

4. Select a formatting style and click OK.

Result. Excel highlights the duplicate names.

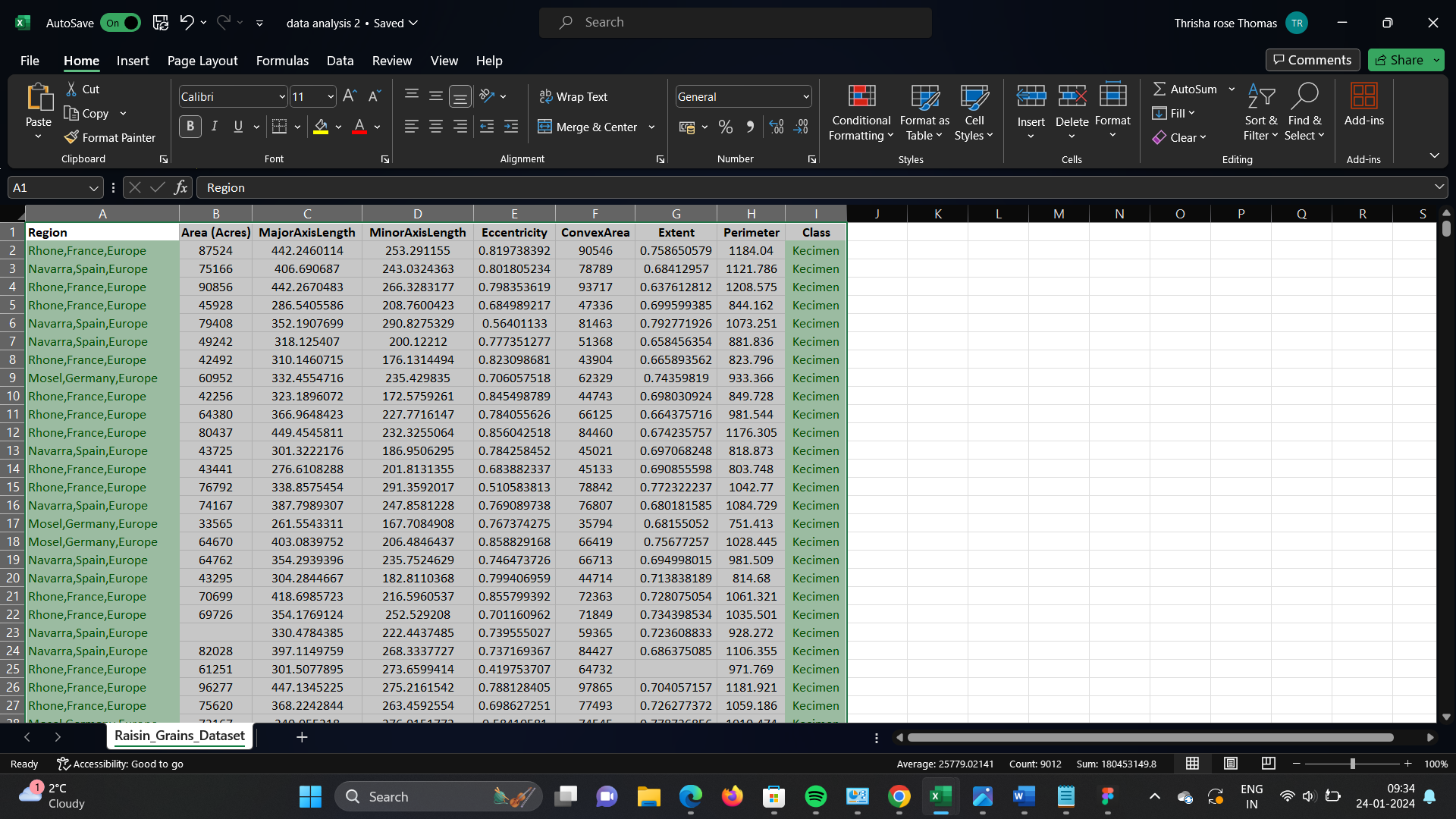


Fig 1.1

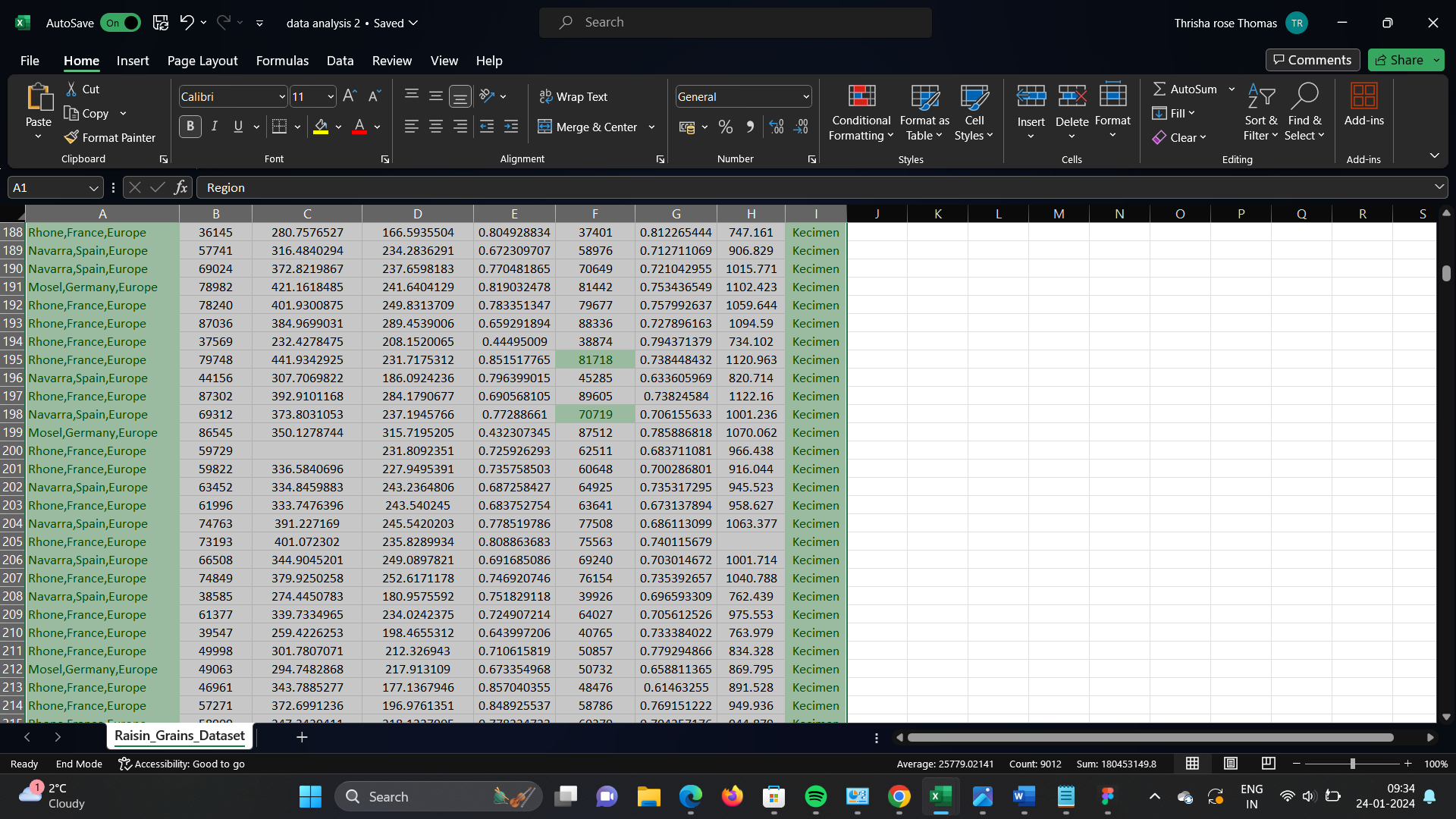


Fig 1.2

# 1.2 Remove duplicate values

1.Select the range of cells that has duplicate values you want to remove.

2.Select **Data** > **Remove Duplicates**, and then under **Columns**, check or uncheck the columns where you want to remove the duplicates.

There were 1003 data before duplication, however there are now just 903 data left.

# 1.3 Handling missing values

Highlighting the missing value

1. Highlight the cells you want, and click the Conditional Formatting button and click New Rule.
2. Choose the Format only cells that contain option and change the dropdown to Blanks.
3. Set your Format.

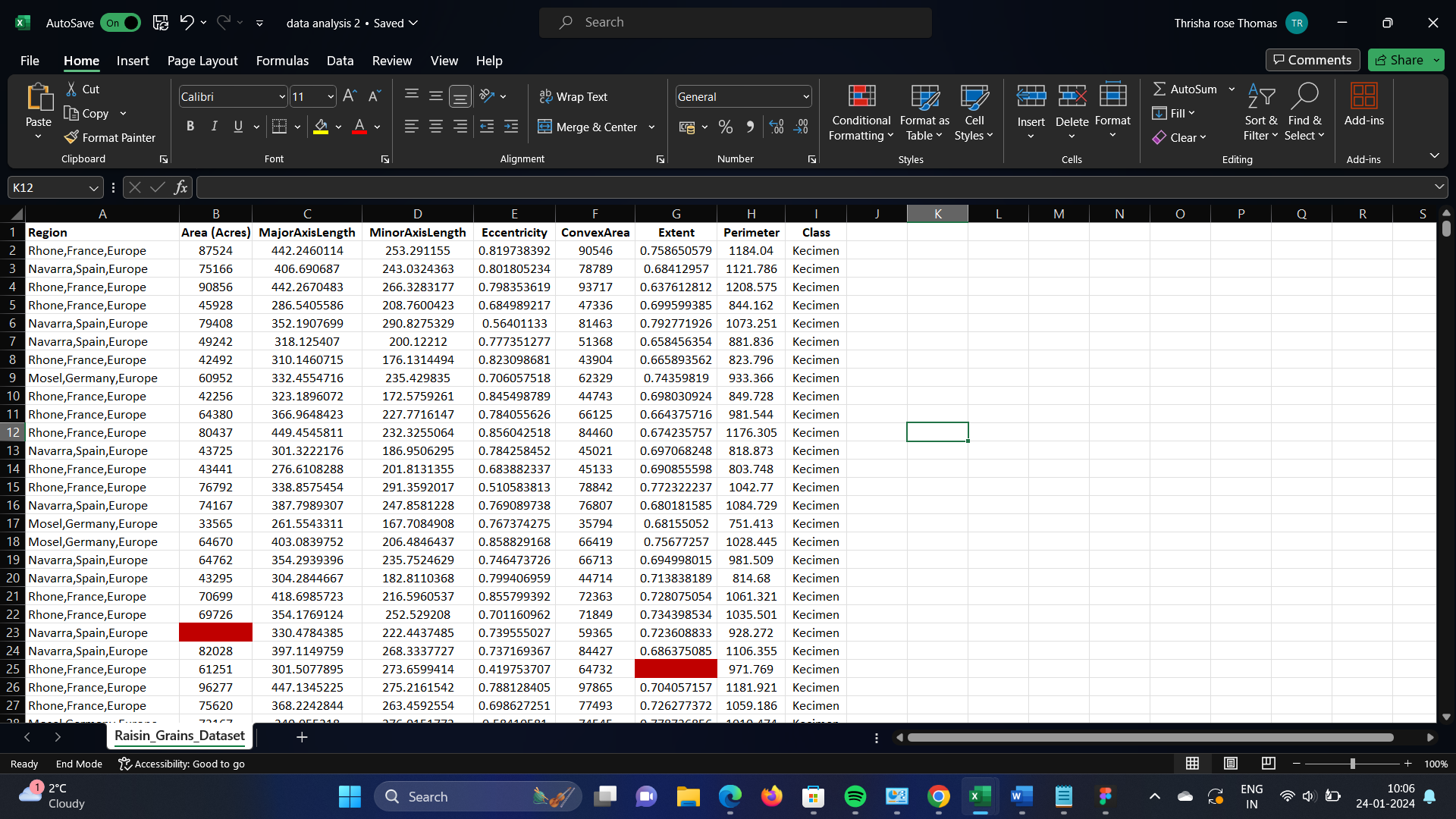


Fig 1.3

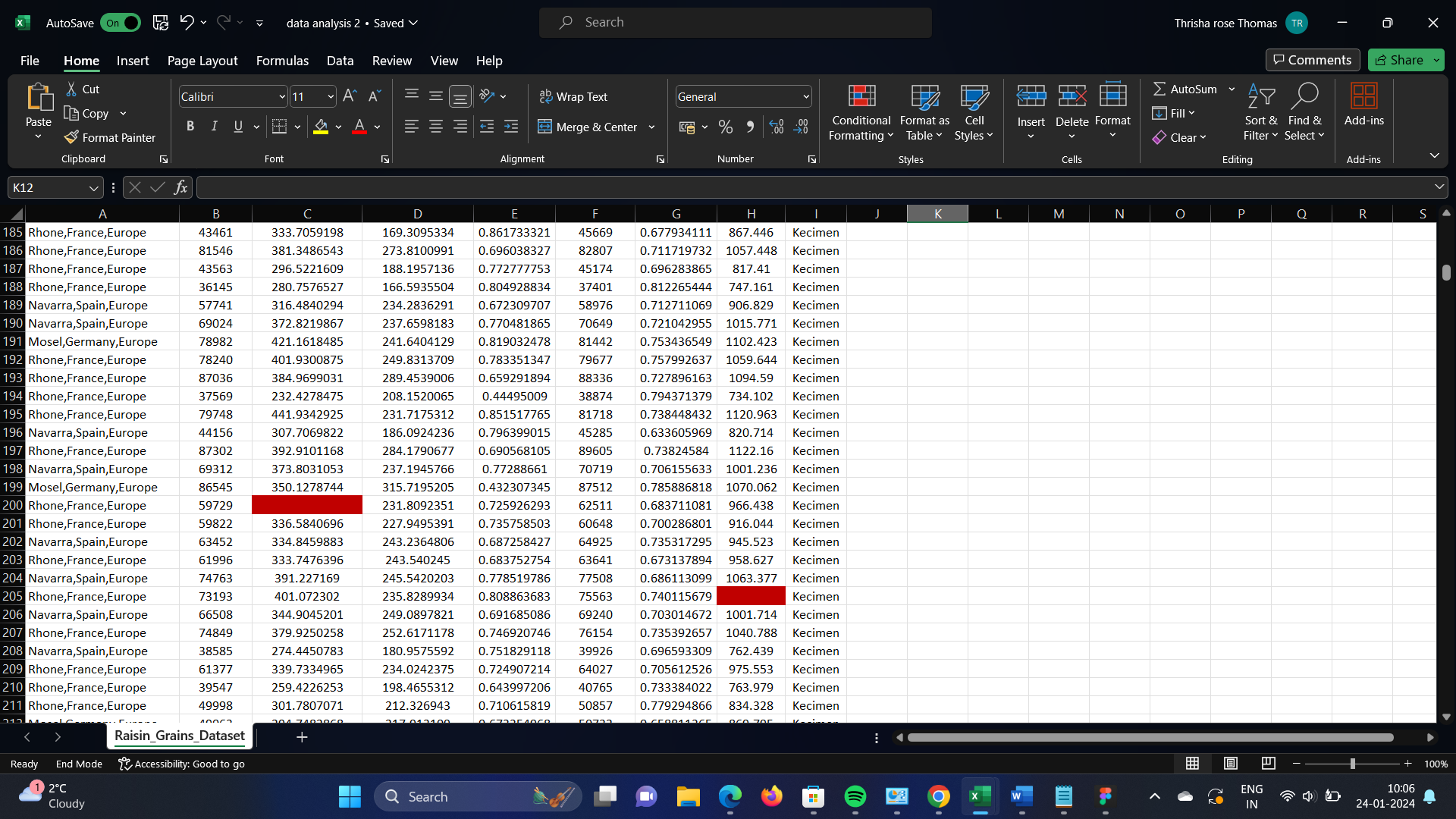


Fig 1.4

To finding the missing cells I used the formula

=COUNTBLANK(A2:I2)

"A2:I2" is the range of cells we want to count.

When we Press "Enter" to execute the formula, and we can see the number of blank cells in the range displayed in the cell.

Once the blank cell has been located, the mean value must be entered.

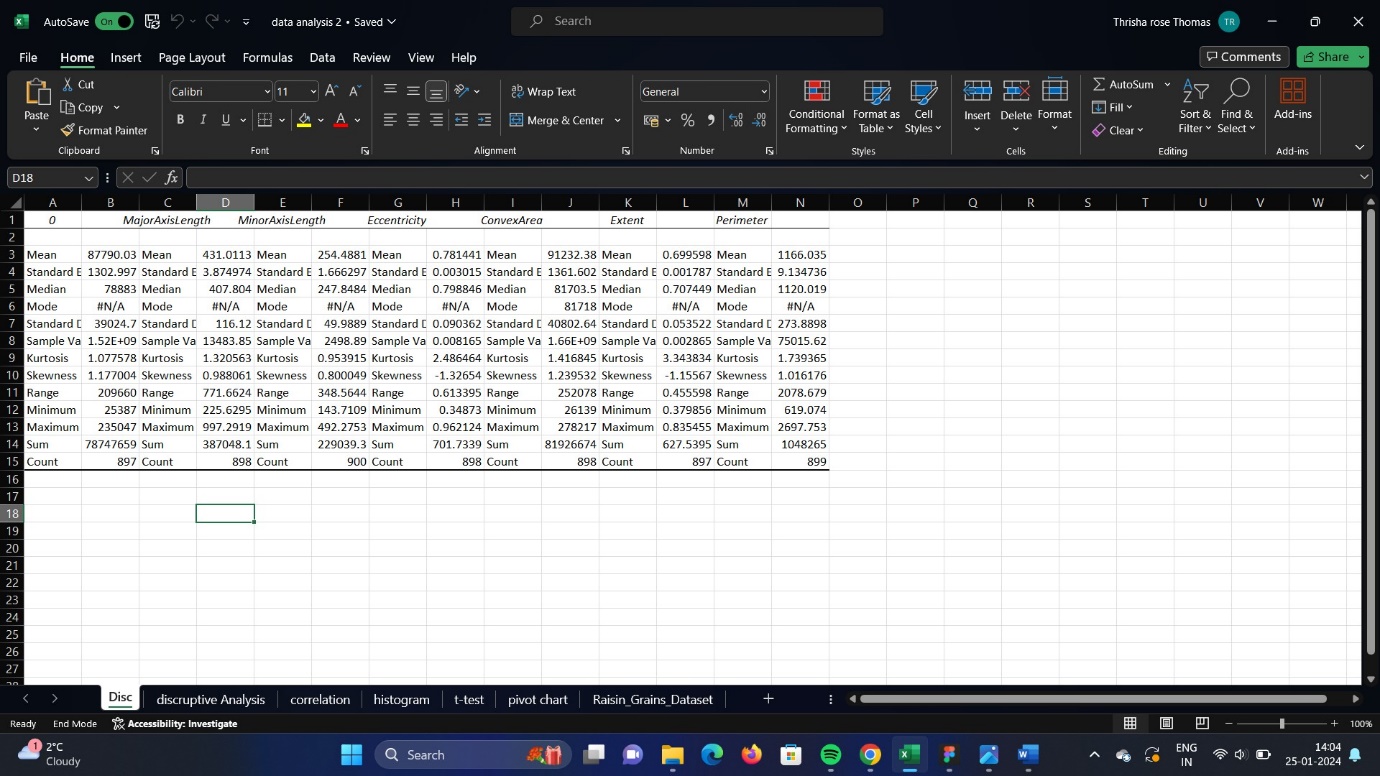


Fig 1.5

We can now replace the blank cell with the descriptive analysis of the entire set of data. To do this, select the range, press Ctrl+F, then enter the mean value and replace it.

# 2.TASK 2

# 2.1 DESCRIPTIVE ANALYSIS

Using the information present in the dataset, Excel's descriptive statistics provide a summary of the statistical data. We can also choose from a variety of output possibilities with this strategy.

We must perform a fresh descriptive analysis after substituting the mean value.

Steps to calculate descriptive analysis

1. Click on Data Analysis in the Analysis section
2. Select Descriptive Statistics, then click OK
3. Click on the Input Range selection button, then select the range of cells for the column
4. Check the Summary Statistics checkbox

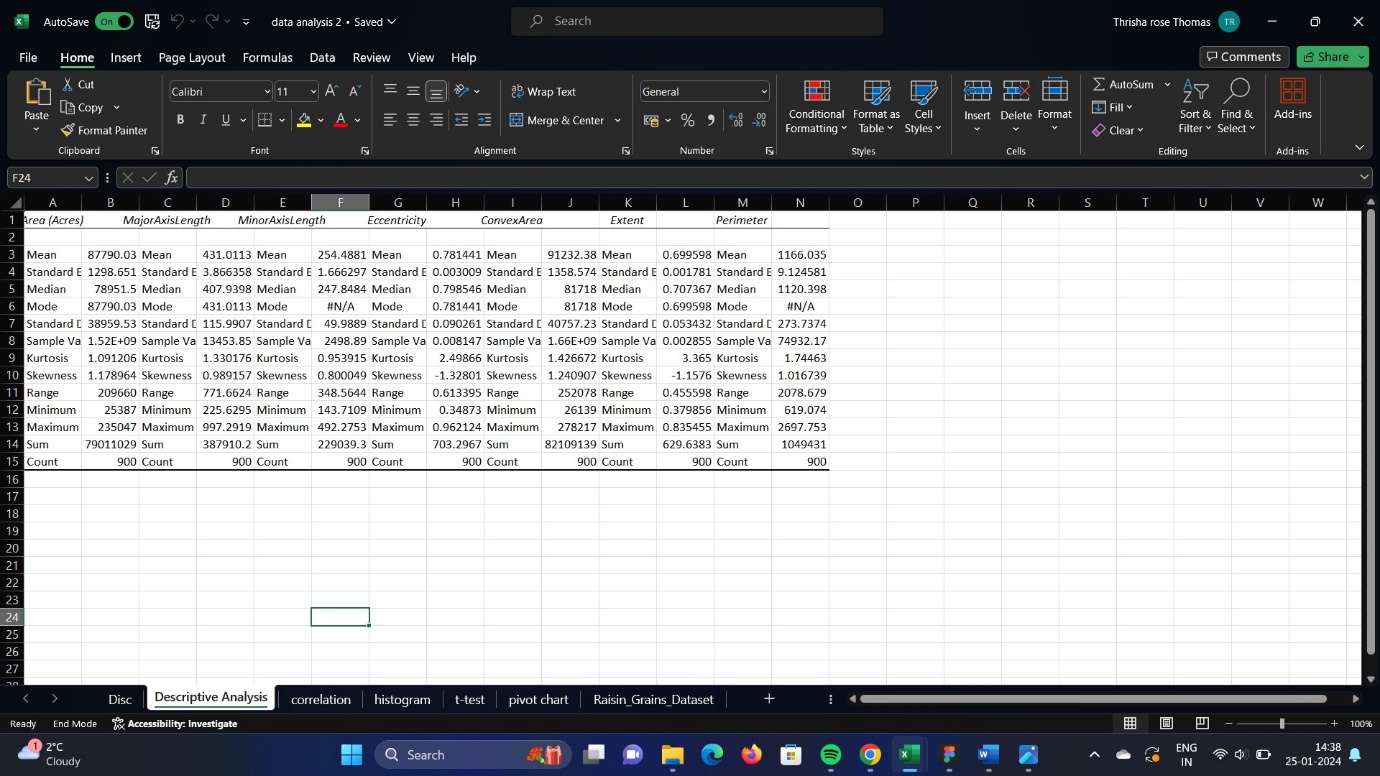


Fig 2.1

# 2.2 SKEWNESS

A severely skewed distribution is one where the skewness value is either more than or less than -1. Moderately skewed distributions occur when skewness falls between 0.5 and 1 or between -1 and -0.5. If the distribution's skewness falls between -0.5 and 0.5, it is roughly symmetric.

We can infer from this data that the major and minor axes lengths are not skewed because their values are less than 1, but the eccentricity may be skewed negatively. Thus, we can use the Excel formula "=k-(E2)" to convert to positive if we so choose, where k is the maximum value. We convert because we are unable to compute negatively skewed data directly; instead, we must make it positive. Once it is positive, we can use the Excel formula "=SQRT(F2)" to find the square root of that specific range and make it normal.

# 2.3 CORRELATION

1. Select “Data” from the top bar menu.
2. Select “Data Analysis” in the top right-hand corner.
3. Select Correlation.
4. Define your data range and output.
5. Evaluate your correlation coefficient.

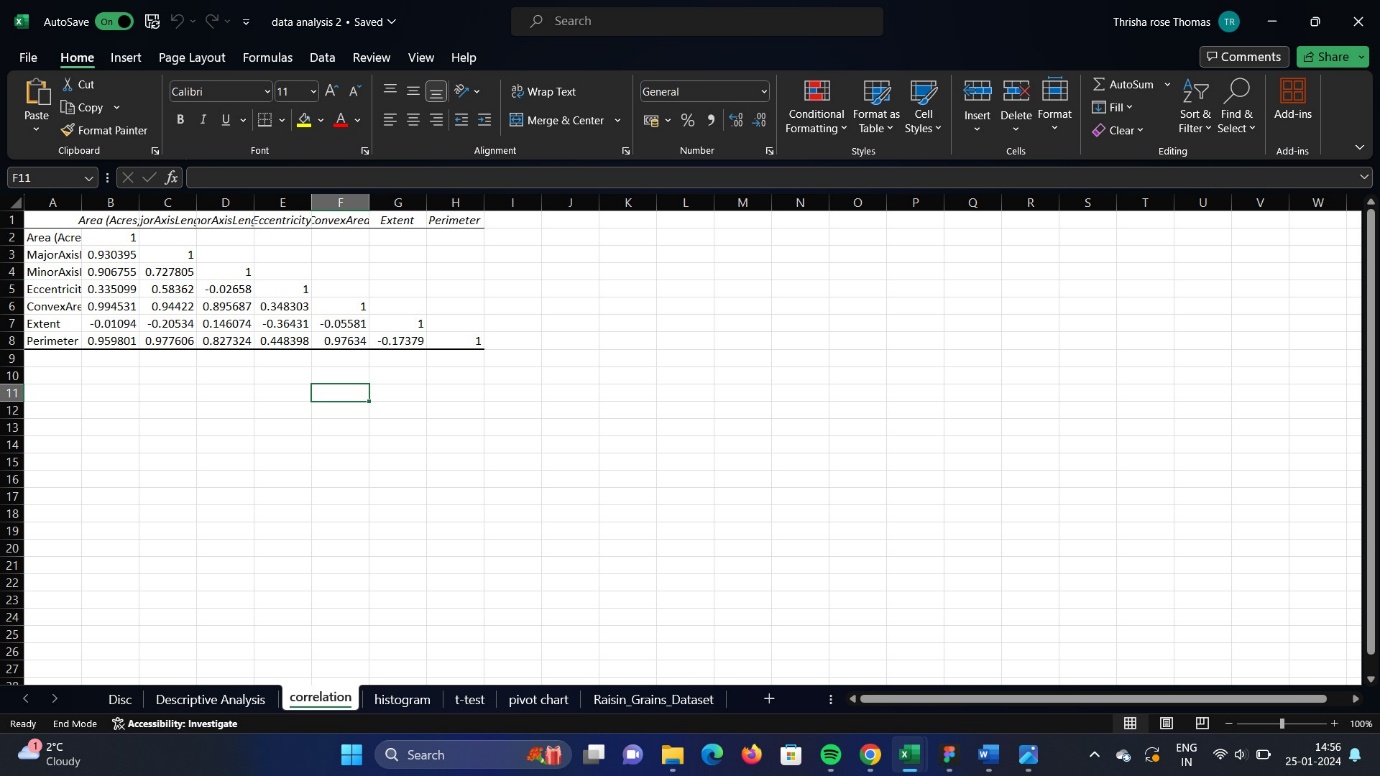


Fig 2.2

The linear Relationship between two variables is measured via correlation. Correlation provides an indicator of the strength of the link by measuring and relating the variance of each variable.

The correlation coefficient, which ranges from -1 to +1, indicates the degree of relationship between two variables.

We draw the conclusion from the data that eccentricity does not correlate with minor axis length but does correlate with major axis length because of the value -0.02 and for Major Axis Length its 0.58. So, we can say that they are not depended to each other because their relationship is not strong

# 2.4 HISTOGRAM

1. Click the Data tab and choose the Data Analysis command.
2. Select the Histogramtool and click OK.
3. Type A2:A11 for the input range and B2:B6 for the bin range.
4. Under Output Options, pick the Output Range radio box and enter the name of the cell where you’d like the histogram output table to be presented. We have now entered cell D2 in our example.
5. Check the box labeled "Chart Output."
6. click ok

histogram is a visual aid that arranges a collection of data points into user-specified ranges. Its layout is akin to that of a bar chart. The histogram groups several data points into logical ranges or bins to simplify a data series into a visually understandable representation.

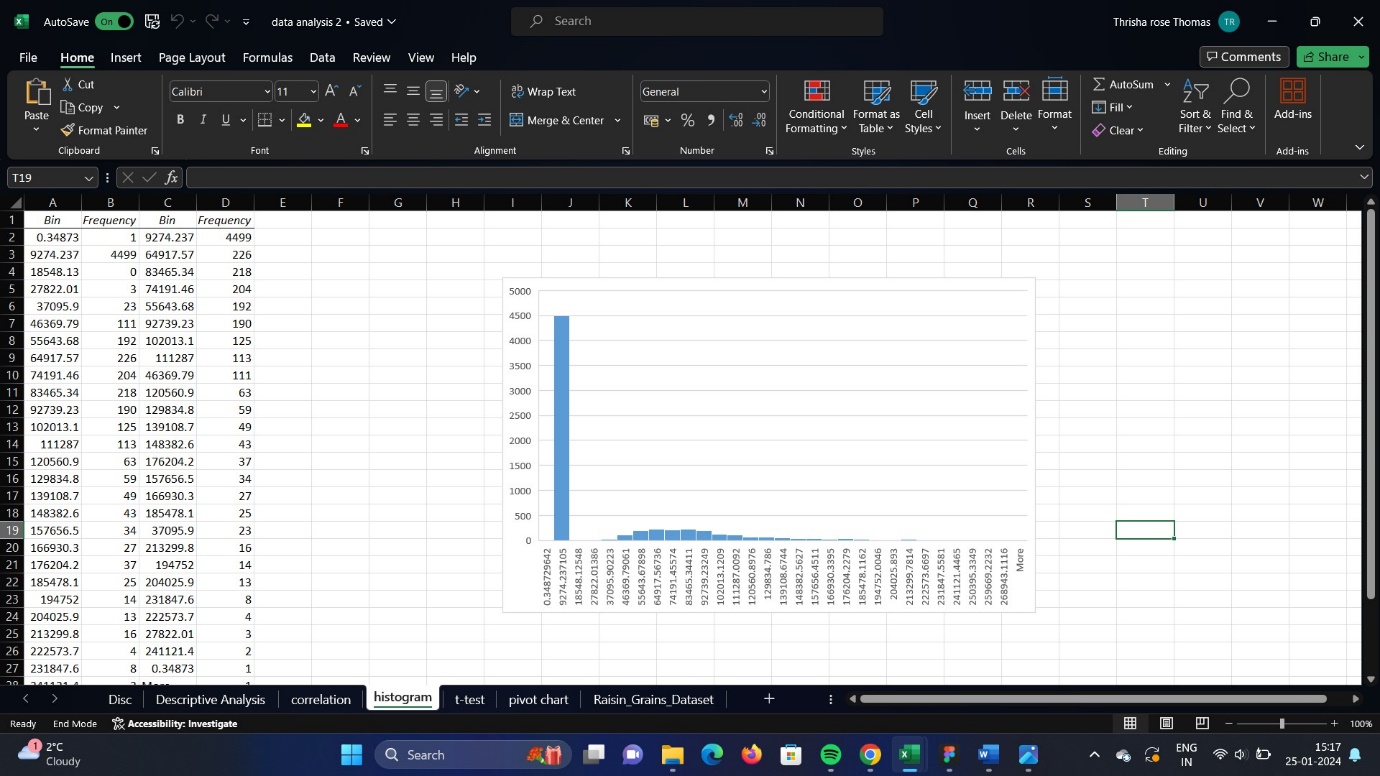


FIG 2.3

# 2.5 HYPOTHESIS TESTING

T test: Two-sample Assuming Unequal Variances

When the following conditions are met:

(1) the samples are normally distributed

(2) the standard deviations of the two populations are unknown and presumed to be unequal

(3) the sample size is large enough (more than thirty), the two-sample T-test with unequal variance can be used.

We can determine that we must perform T test since we have met the testing requirements: Two-sample Assuming Unequal Variances for the Reasons Below:

* We have only 2 groups
* We have range more than 30
* We have different sample variances
* The data is normally distributed

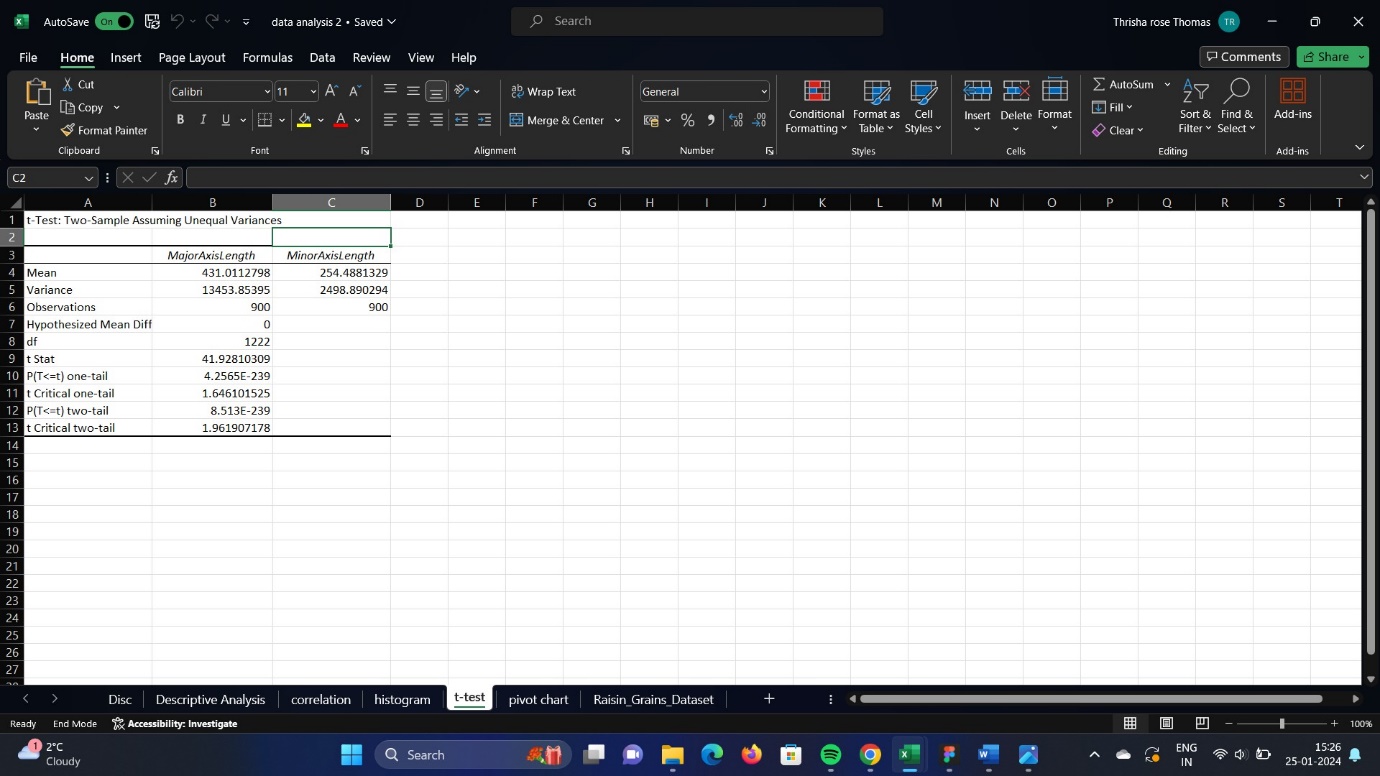


Fig 2.4

From the test we can conclude that p value is 8.513E-239.So, if the p<=5 then we need to reject null hypothesis and accept alternative hypothesis so this case we need to do the same

Since the major and minor axes Length are independent variables and the mean value of the two groups differs, we may deduce that eccentricity depends on the length of the major and minor axes Length and affects the grape's shape.

# 2.6 REPORT

My research indicates that the Major Axis Length and Minor Axis Length have a significant impact on Eccentricity because they have distinct mean values. For instance, if the Minor Axis Length has larger values, the grape would be larger and become more juicy, making it more expensive.

Additionally, the grape will be longer and appear thinner if the Major Axis Length has greater value, which will result in a lower price.

# TASK 3

PIVOT CHART

Insert pivot chart

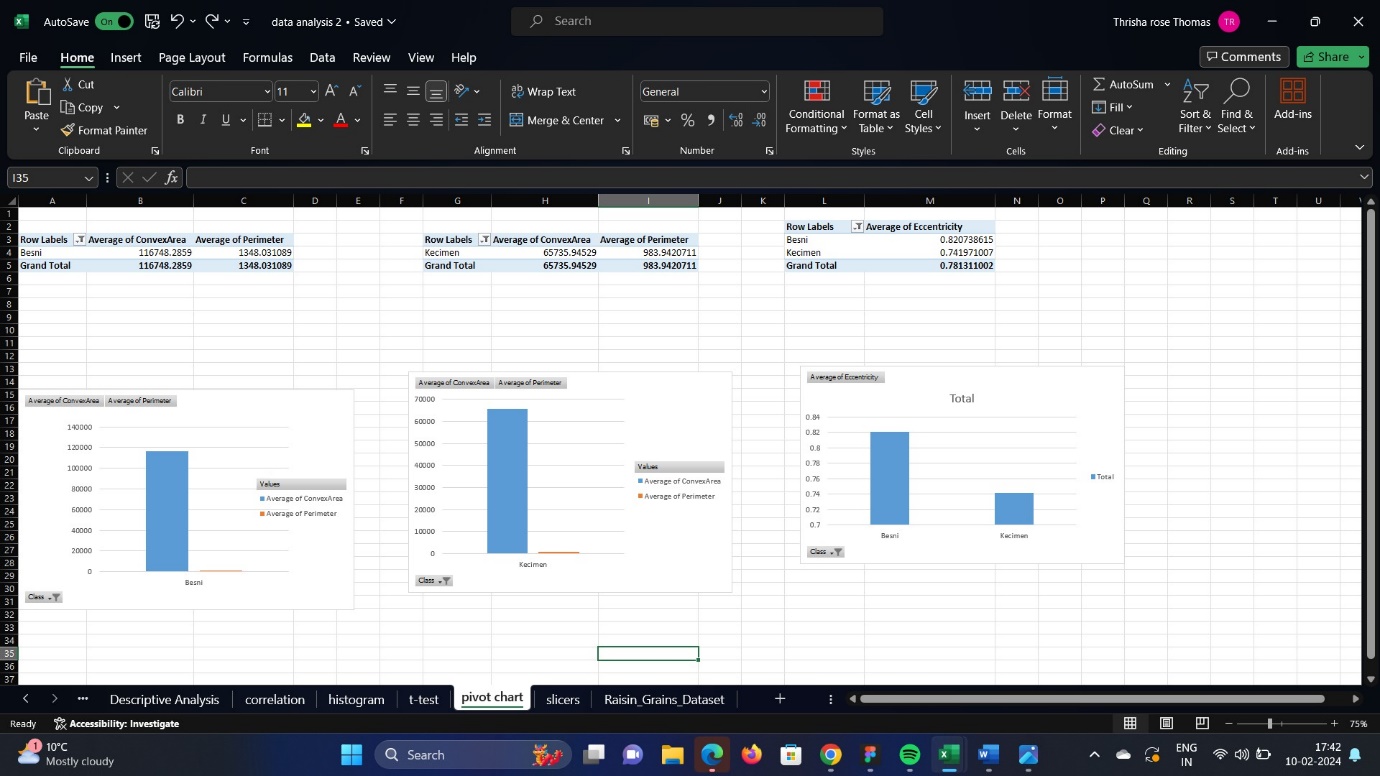
Step 1: click on pivot table from insert

Step 2: insert the range from the option table/range

Step 3: select the option new worksheet

Step 4: choose the fields to be selected in Pivot Table Fields

Step 5: From pivot Table Analyse choose any pivot chart

Fig 3.1

These are the three pivot charts I created for task 3, The class Besni is on the first chart, followed by Kecimen on the second, and each class's eccentricity is displayed on the third chart.

# Largest and the Smallest grape

The table below lists the regions with the smallest and largest grapes for each class that we analysed using convex area. Convex area is defined as the entire area bounded by a shape's outermost boundary. If the specific class's convex area is large, the grape is large; if the value is small, the grape is small.

|  |  |  |  |
| --- | --- | --- | --- |
| Region | Class | largest | Smallest |
| Mosel,Germany,Europe | Besni | 221527 | 50748 |
|  | Kecimen | 178818 | 28607 |
| Navarra,spain,Europe | Besni | 239093 | 50103 |
|  | Kecimen | 221396 | 33699 |
| Rhone,France,Europe | Besni | 278217 | 41934 |
|  | Kecimen | 107423 | 30316 |

# Largest and Smallest perimeter

|  |  |  |
| --- | --- | --- |
| Region | Smallest | Largest |
| Morsel,Germany,Europe | Kecimen | Besni |
| Navarra,spain,Europe | Besni | Kecimen |
| Rhone,France,Europe | Kecimen | Besni |

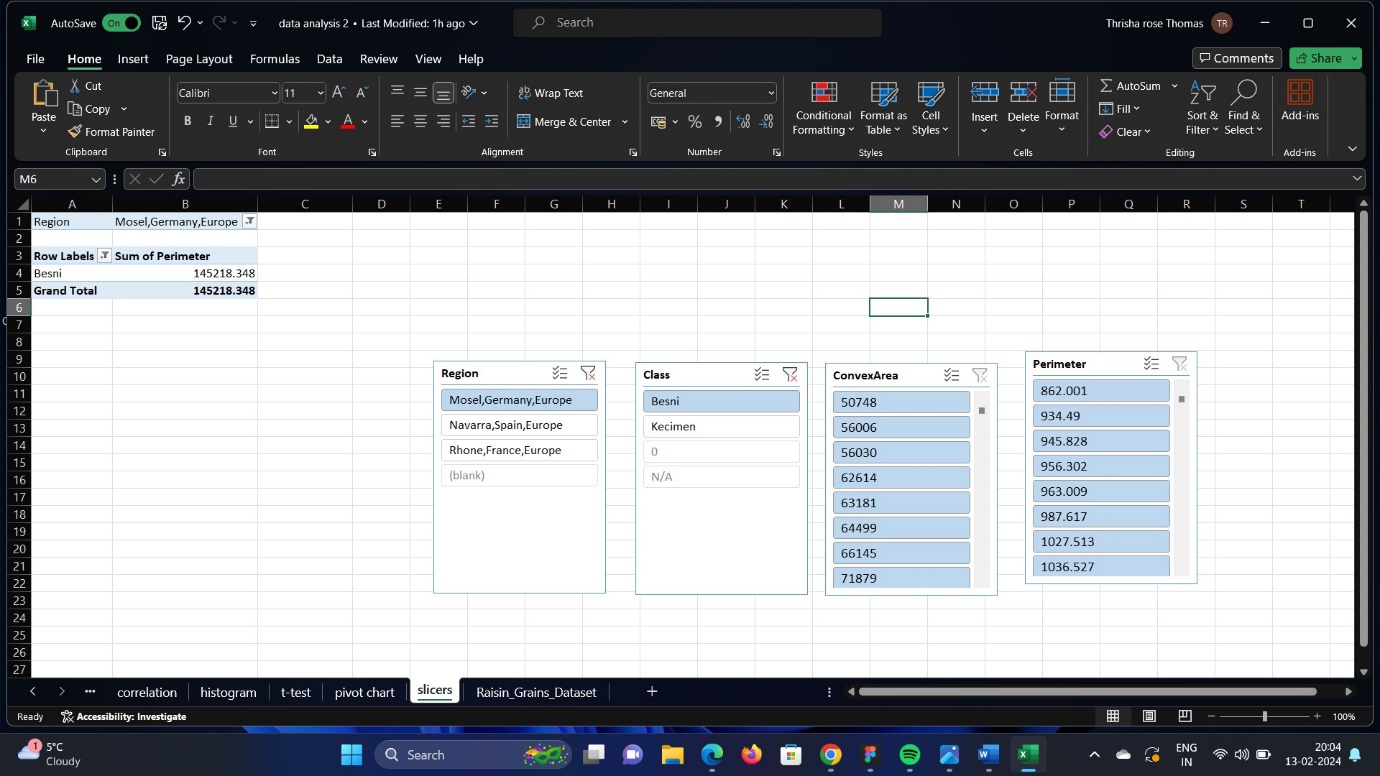


Fig 3.2

References

<https://blog.hubspot.com/marketing/how-to-create-pivot-table-tutorial-ht>

<https://www.excel-easy.com/examples/descriptive-statistics.html>

<https://www.excel-easy.com/examples/correlation.html>