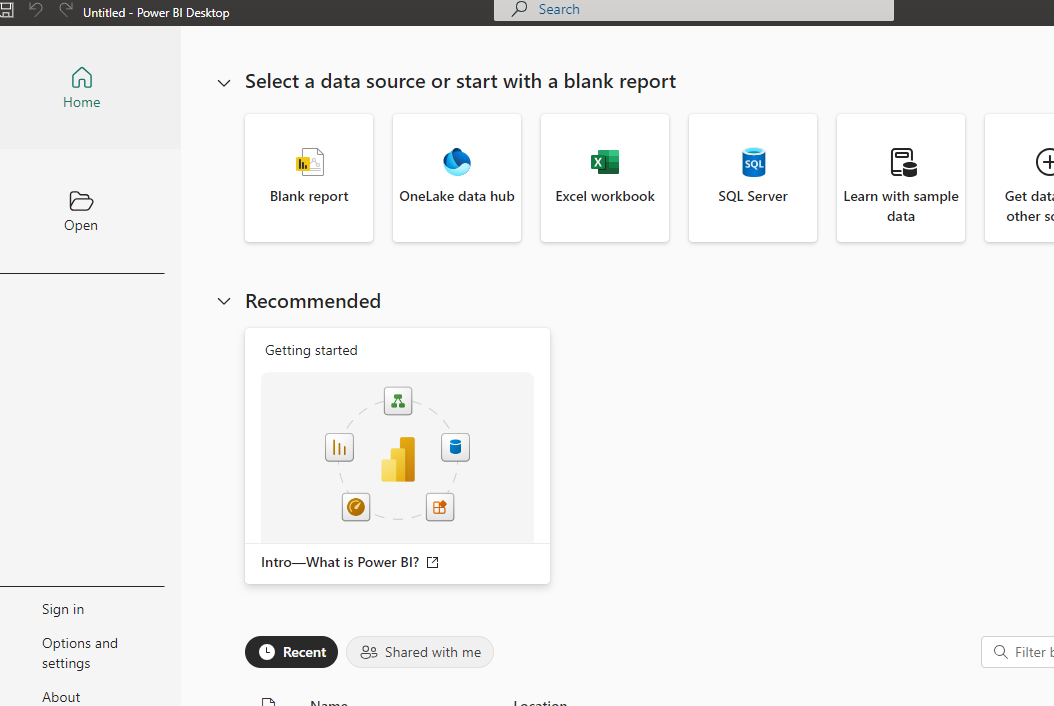
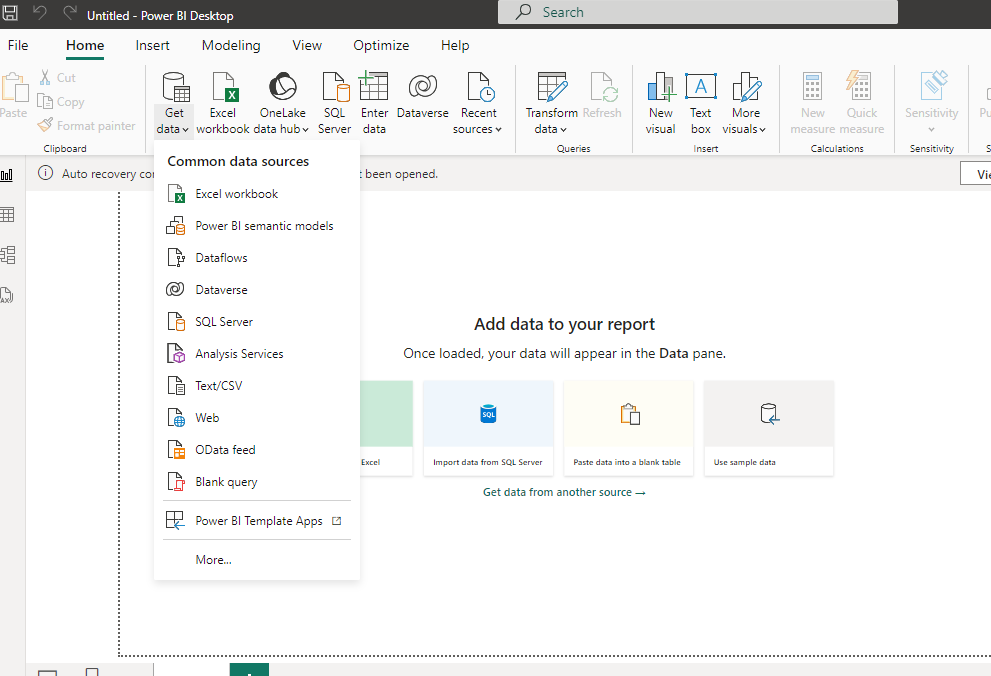
### Capstone Project: Exploring Coffee Quality Data with Power BI

**Name: Sanjiban Hati**

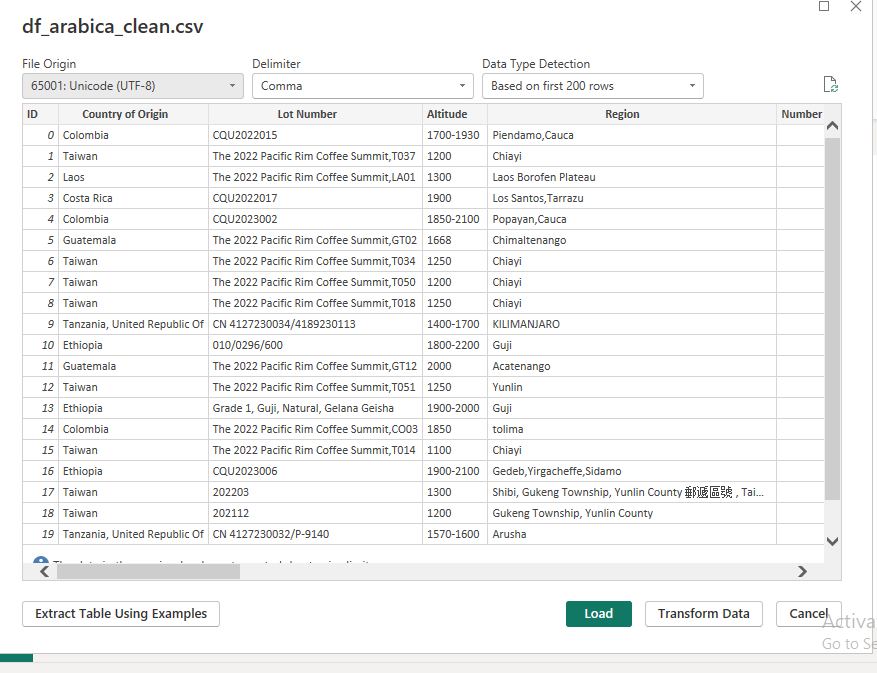
**Student ID: S8895**

#### **Introduction:** The Coffee Quality Institute (CQI) is a non-profit organization that works to improve the quality and value of coffee worldwide. ‘df\_arabica\_clean.csv’, the dataset we are using for the project consists of several columns including ID, Country of Origin, Regions, different sensory attributes like Aroma, Flavor etc, Defects, Overall, Total Cup Points and others. As part of this exercise, we analyzed the data to derive conclusions which would guide the vendor decide better for their business and thus help to grow the same over period.

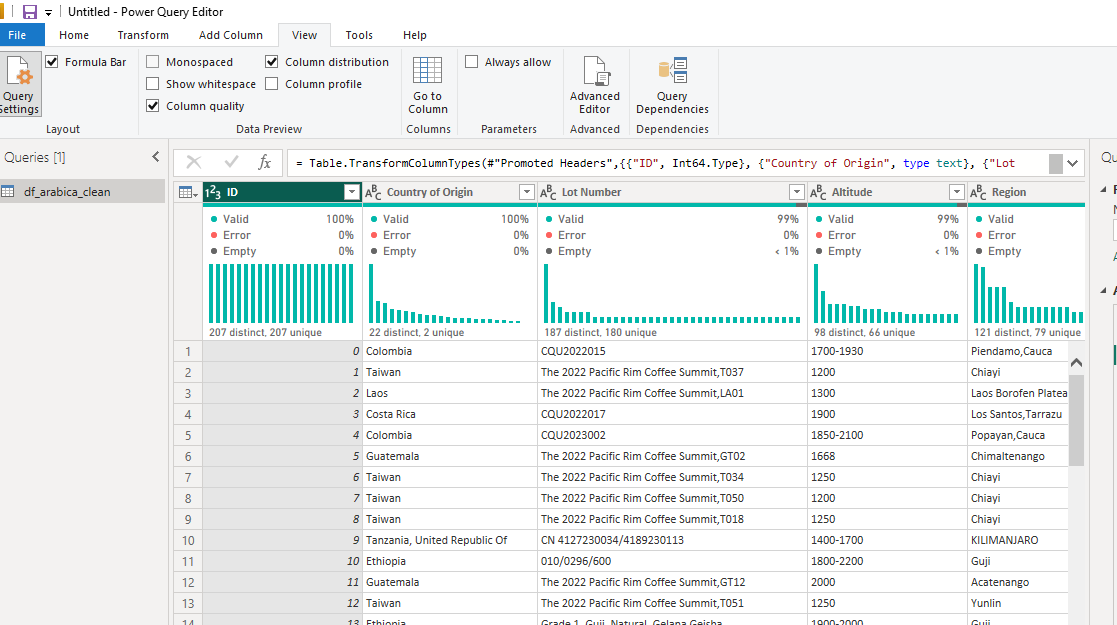
**Step 1: Database Importing**   
The database, df\_arabica\_clean.csv was first downloaded from the Odinschool platform to the system. On opening Power BI, the following flow was followed to import the database into the application.  
**Blank report -> Get data -> Text/CSV** (Choose the file from the location and click on Open)   
  




After the file is chosen, there come three options i.e. **Load**, **Transform** **data** and **Cancel** among which the second one is to be selected as the data cleaning and preprocessing is needed for the database.



**Step 2: Data Cleaning & Pre-processing**  
Now the data was opened in the **Power Query Editor** mode. On selecting the boxes against the options **Column quality** and **Column distribution** under **View** ribbon, it was seen whether the rows for a particular column of that table were having errors or empty values in themselves.

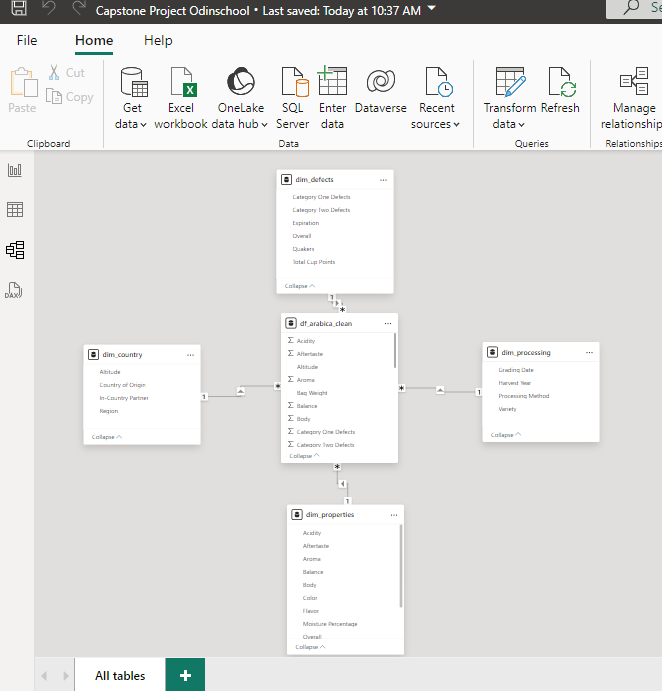
  
  
Based on that, the preprocessing steps were executed.   
Below are shared the steps which were needed to be performed as part of the data cleaning part of this particular database.   
A. Lot Number, Altitude, Region, Variety and Processing Method - all these 5 columns had blank values, so they were first replaced with Null.  
B. Then each of them were filled with the values from the cells right above them by using the Fill down option inside Fill button under Transform ribbon.   
C. Following columns were removed from the table as they were having a single value for all the records and hence didn’t contribute to the analysis.   
  
Status -> Completed   
Defects -> 0   
Clean Cup -> 10  
Sweetness -> 10   
  
with the above steps, the cleaning and pre-processing part was completed for the database.

**Step 3: Data Modelling**  
In this segment, there were created four **dimension tables** which are **dim\_country**, **dim\_processing**, **dim\_defects** and **dim\_properties** respectively from **df\_arabica\_clean** as the **fact table**. Please find the column details for them below:

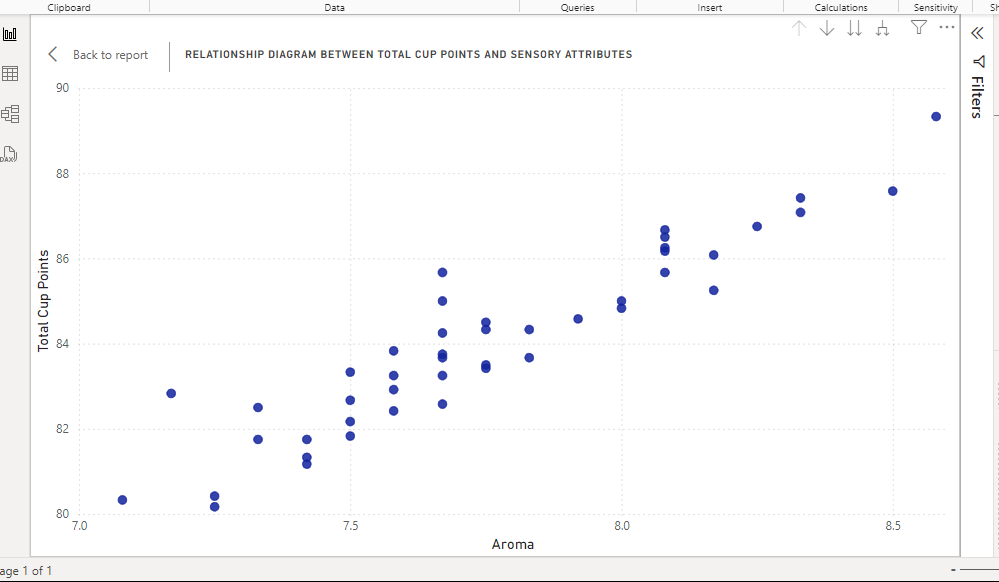
|  |  |
| --- | --- |
| Column Name | Type |
| Country of Origin | Dimension(dim\_country) |
| Altitude | Dimension(dim\_country) |
| Region | Dimension(dim\_country) |
| In-Country Partner | Dimension(dim\_country) |
| Processing Method | Dimension(dim\_processing) |
| Harvest Year | Dimension(dim\_processing) |
| Grading Date | Dimension(dim\_processing) |
| Variety | Dimension(dim\_processing) |
| Total Cup Points | Dimension(dim\_defects) |
| Overall | Dimension(dim\_defects) |
| Category One Defects | Dimension(dim\_defects) |
| Quakers | Dimension(dim\_defects) |
| Category Two Defects | Dimension(dim\_defects) |
| Expiration | Dimension(dim\_defects) |
| Variety | Dimension(dim\_properties) |
| Aroma | Dimension(dim\_properties) |
| Flavor | Dimension(dim\_properties) |
| Aftertaste | Dimension(dim\_properties) |
| Acidity | Dimension(dim\_properties) |
| Body | Dimension(dim\_properties) |
| Balance | Dimension(dim\_properties) |
| Uniformity | Dimension(dim\_properties) |
| Overall | Dimension(dim\_properties) |
| Total Cup Points | Dimension(dim\_properties) |
| Moisture Percentage | Dimension(dim\_properties) |
| Color | Dimension(dim\_properties) |

**One to Many Relationships** were retained between the following pairs as part of the **Star Schema**.

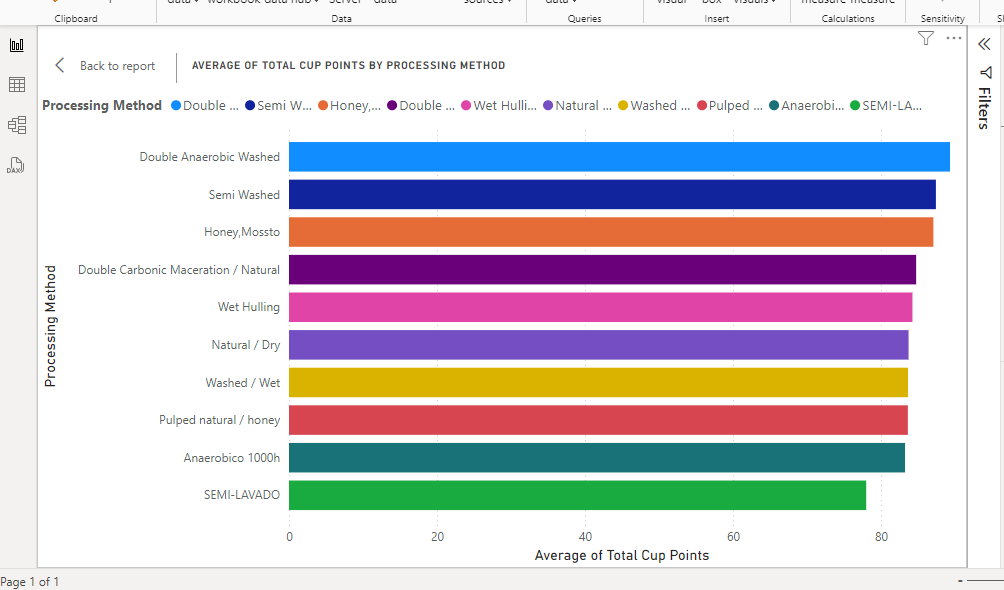
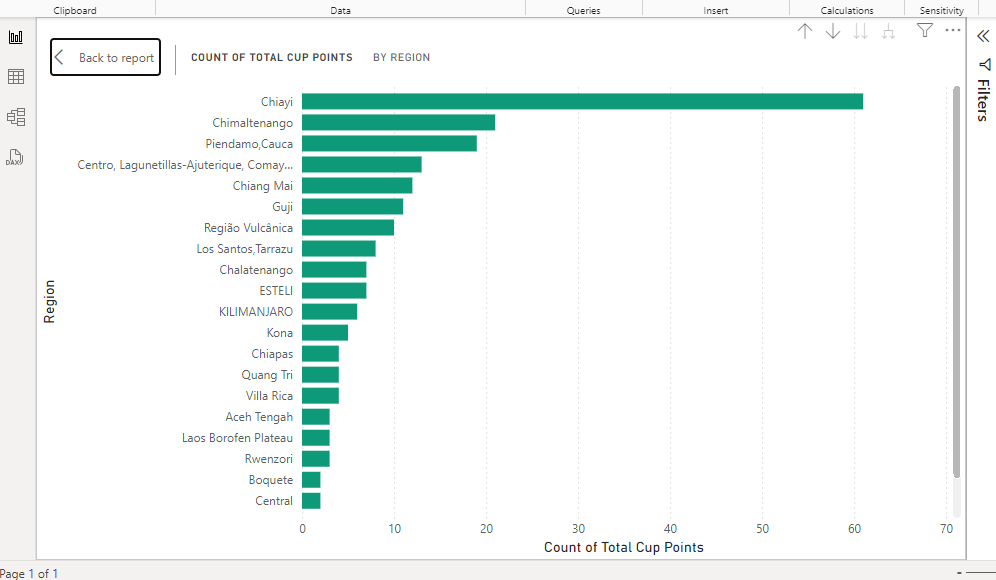
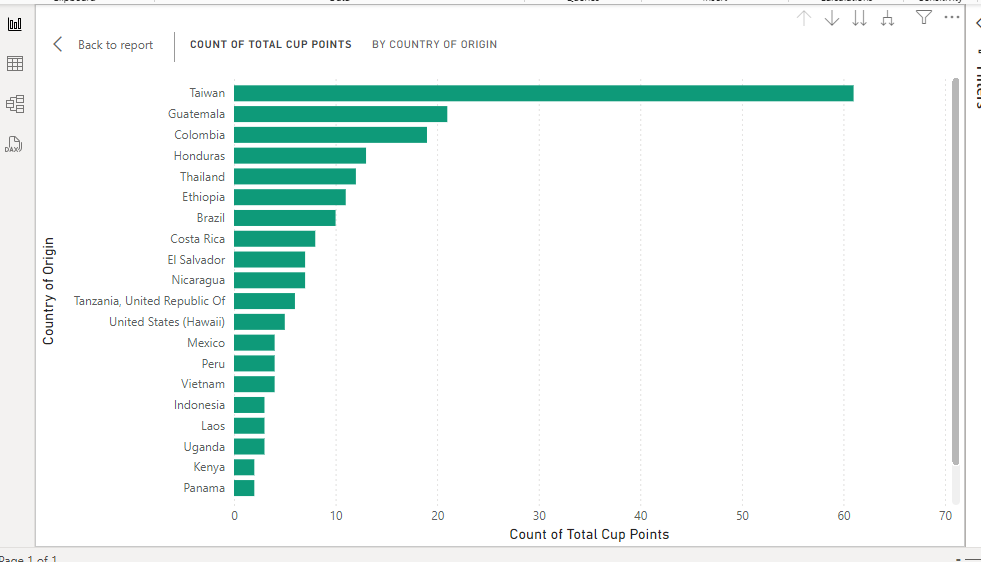
Country of Origin (dim\_country) -> Country of Origin (df\_arabica\_clean)  
Processing Method (dim\_processing) -> Processing Method (df\_arabica\_clean)  
Total Cup Points (dim\_defects) -> Total Cup Points (df\_arabica\_clean)  
Variety (dim\_properties) -> Variety (df\_arabica\_clean)

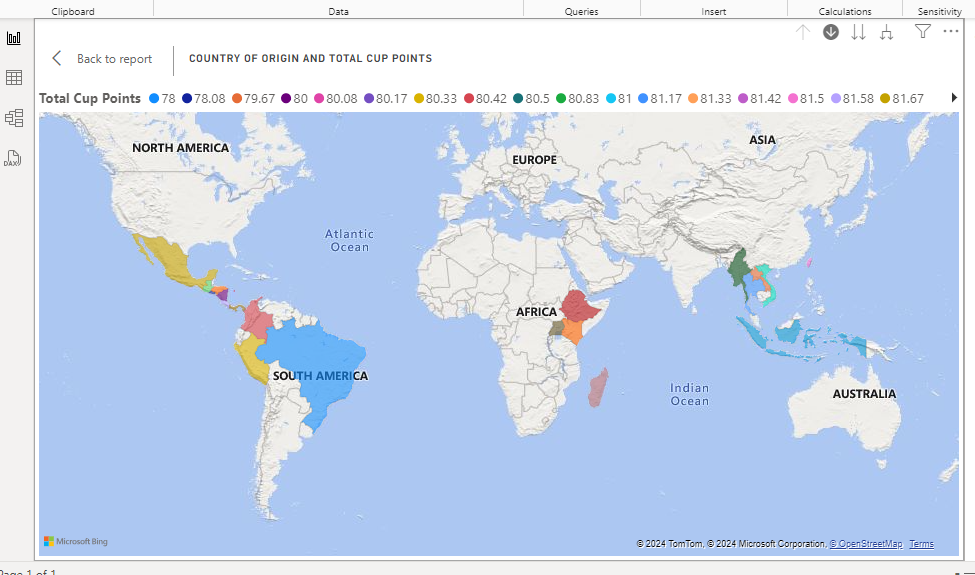
Please find the Star Schema from the Model View below:  


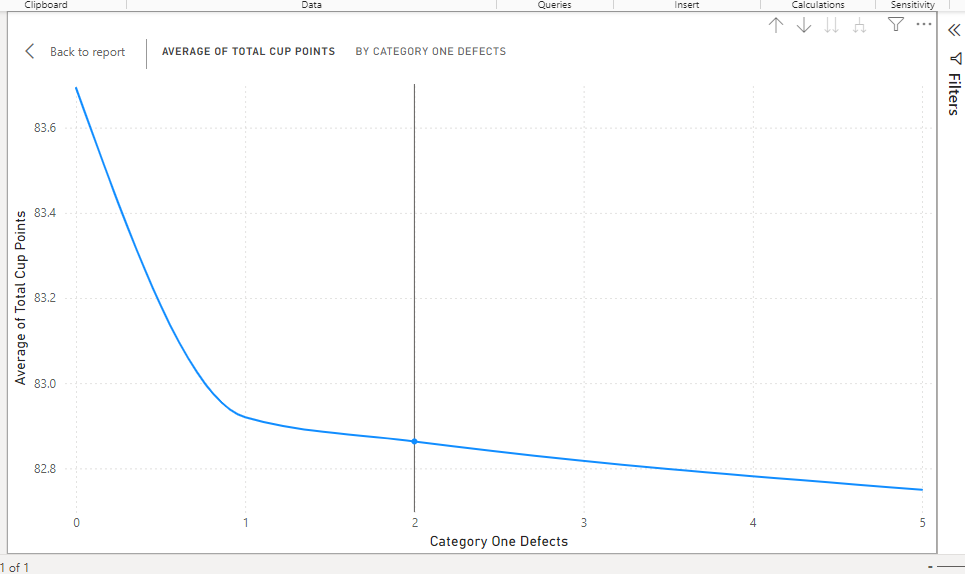
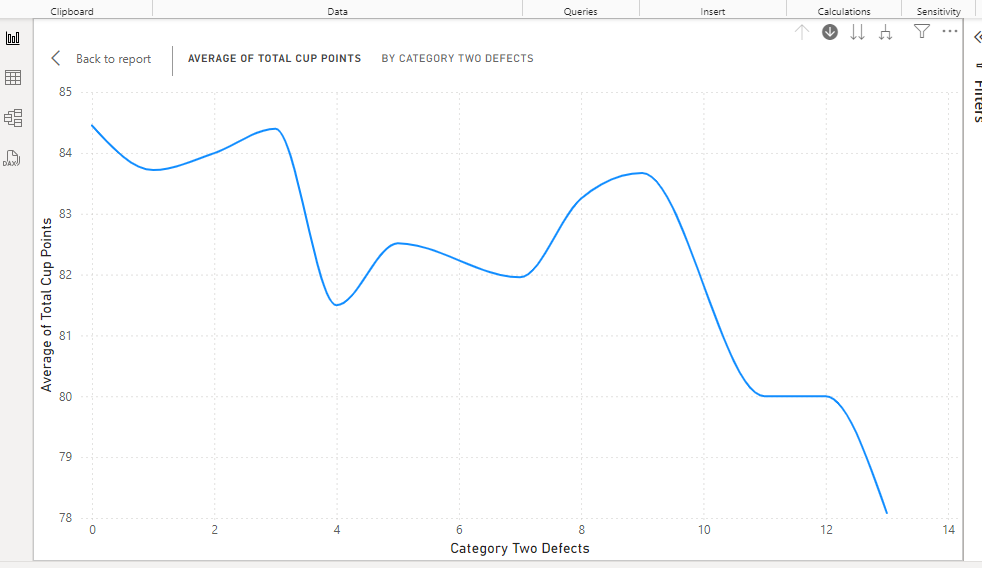
**Step 4: Data Analysis**  
As part of the Data Analysis for the given dataset, a number of approaches were taken. They are being shared below.   
  
**A. Influence of Sensory Attributes on the Coffee Quality:**  
To showcase how the coffee quality is determined on the basis of several sensory attributes like Aroma, Aftertaste, Acidity, Balance, Body and Flavor, **Scatter Plot** was introduced to the report where the sensory attributes were placed on the X axis and Total Cup Points on the Y axis.



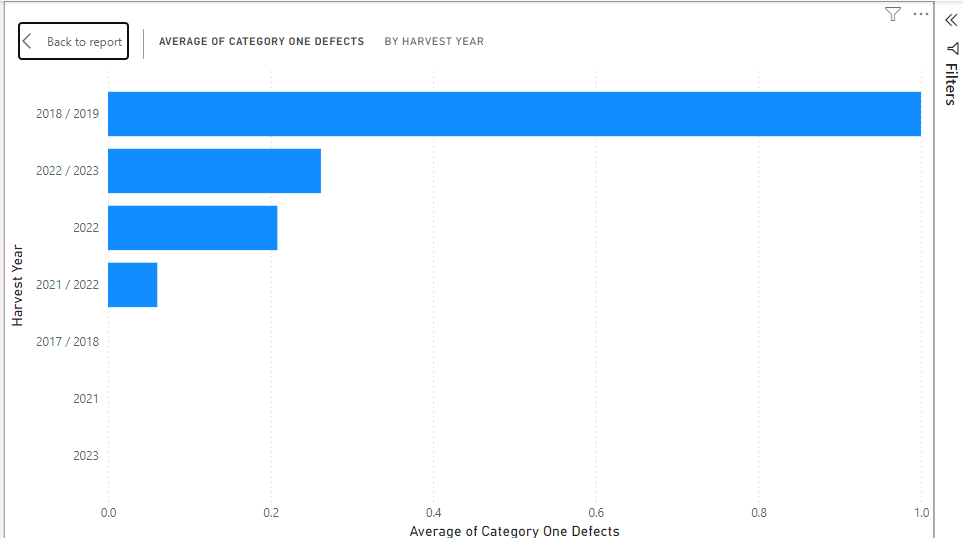
As it can be seen in the graph, the Coffee Quality (Total Cup Points) shows a directly proportional relationship with each of the sensory attributes.

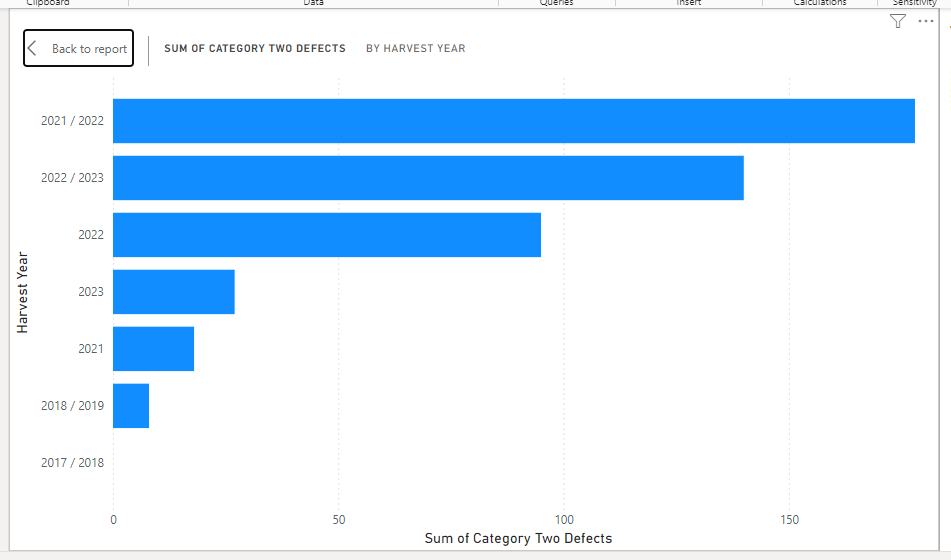
**B. Correlation between processing methods, origin regions, and coffee quality scores:**   
In order to find the correlation between processing methods, origin regions, and coffee quality, there were used **Bar Chart** and **Filled Map** respectively.   
The first Bar chart had Processing Method on Y axis, Legend and Average of Total Cup Points on X axis.   
  
  
From the graph, it’s evident that for the method, Double Anaerobic Washed, the average value of Total Cup Points is the highest one.   
  
In the second Bar Chart, Country of Origin and Region were placed on the Y axis whereas Total Cup Points on the X axis.   
  
  
  
  
It was discovered from this graph that the Total Cup Points has the highest value for the region Chiayi and the Country of Origin Taiwan respectively.

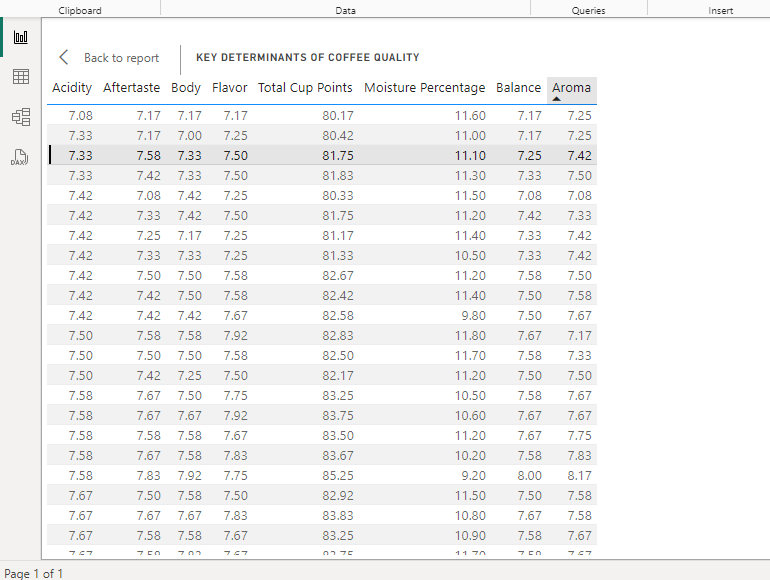
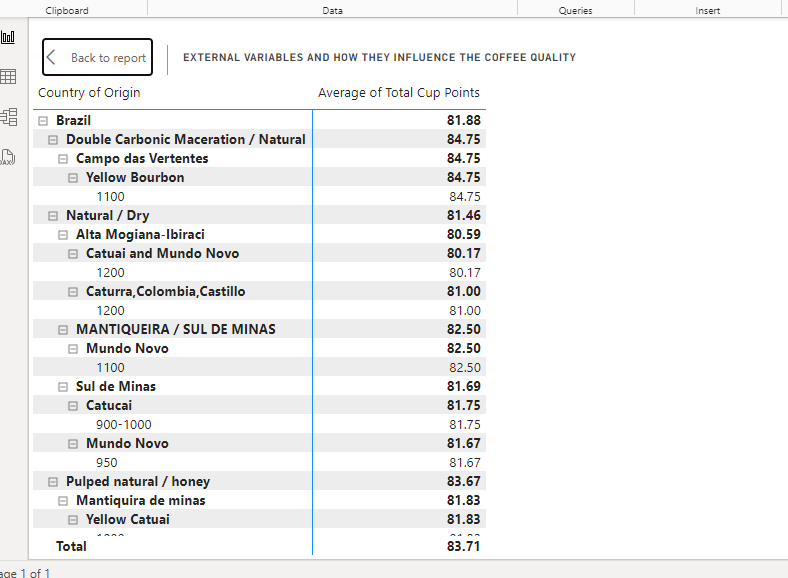
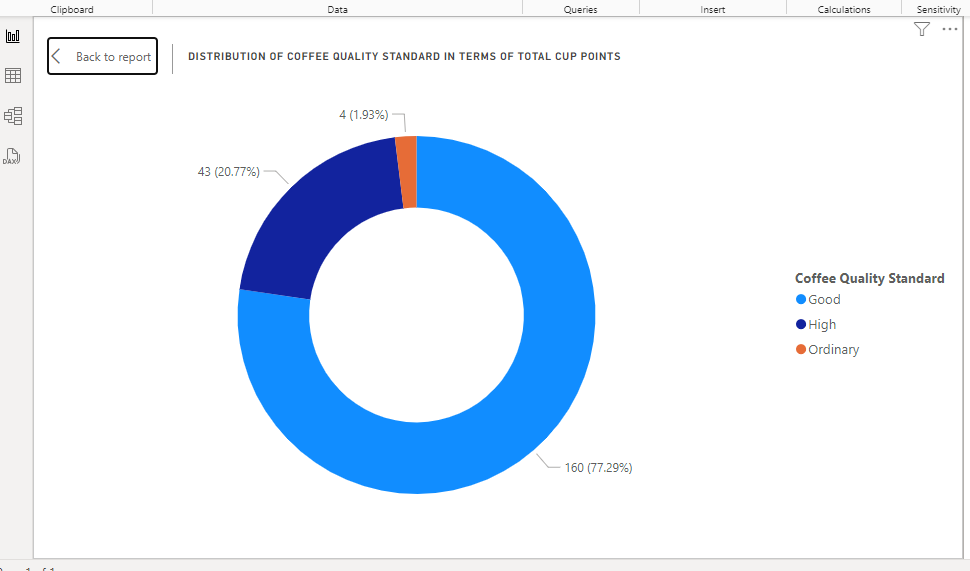
To further emphasize the variation of coffee quality along with the Geographical changes, **Filled Map** was incorporated in the report.   
  
  
  
**C. Relationship between the occurrence of defects and coffee quality along with change in the same over the years:**   
For portraying the distribution of coffee quality along with the occurrence of defects, **Line Chart** was used in the report where Category One Defects and Category Two Defects were kept on the X axis and Average of Total Cup Points on the Y axis.

As it can be seen in the graphs above, Defects (both the categories) retain an inversely proportional relationship with the Coffee Quality (Total Cup Points), i.e. higher quality stands for lower values for defects.



  
In the graphs shared above, the change in both types of Defects over the period can be observed. As it can be seen, Category One and Category Two Defects are the maximum for the years 2018/2019 and 2021/2022 respectively.

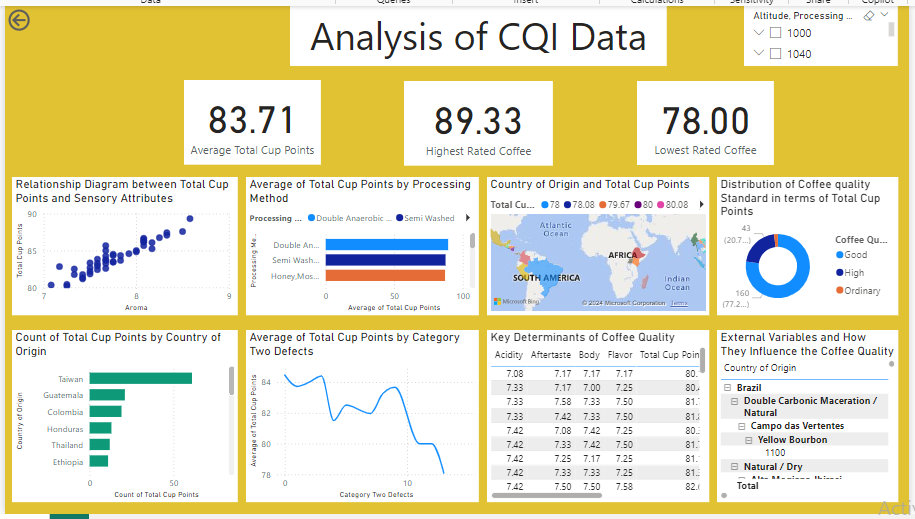
**D. Influence of Different Variables on the Coffee Quality:**   
To share how different variables impact the coffee quality, a Table along with a Matrix were included in the report.   
The following **Table** represents the various factors deciding the coffee quality written against the Total Cup Points.   
  
  
  
Along with this find the **Matrix** shared below. Here all the deciding components have been gathered together against the Total Cup Points to give a comprehensive idea about the impact of the same on the Total Cup Points.   
  
  
  
**E. Miscellaneous:**   
After the earlier analysis, in this step there was created a **conditional column** in the table df\_arabica\_clean. **Coffee Quality Standard**, the column was created based on the values from the column Total Cup Points.   
If Total Cup Points <= 80, Coffee Quality Standard was considered as Ordinary  
If Total Cup Points <= 85, Coffee Quality Standard was considered as Good  
If Total Cup Points > 85, Coffee Quality Standard was considered as High   
  
Afterwards, in the report was incorporated a Donut Chart that shows the percentage for each type of Coffee quality standard recorded in the given dataset.   
  
  
  
As shown in the visualization, major part of the coffee quality belongs to the standard Good, i.e. Total Cup Points resides between 80 and 85.   
  
Apart from these, as the sole purpose of this analysis was to have a thorough understanding of the Coffee quality and the impact of several attributes on the same, the following **DAX** expressions were considered using **New Measure** for portraying a high level view of the quality as well.   
  
**Average Total Cup Points = AVERAGE(df\_arabica\_clean[Total Cup Points])**

**Highest Rated Coffee = CALCULATE(MAX(df\_arabica\_clean[Total Cup Points]), df\_arabica\_clean[Total Cup Points] = MAX(df\_arabica\_clean[Total Cup Points]))**

**Lowest Rated Coffee = CALCULATE(MIN(df\_arabica\_clean[Total Cup Points]), df\_arabica\_clean[Total Cup Points] = MIN(df\_arabica\_clean[Total Cup Points]))**

The values were documented in the report with the help of **Card** visuals.   
  
  
  
Thereafter, **Slicer** was added to the report so that the variation of the coffee quality in different visualizations can be observed along with the changes in the determining factors like Altitude, Processing Method, Country of Origin, Region and Variety.

That’s how the complete analysis for the CQI data was executed and the below is being shared the end report for the same.



Tools which were used as part of this project are:

1. Power BI
2. Microsoft Excel