## Phase 3 :Big Data Analysis with IBM Cloud Databases

**INTRODUCTION :**

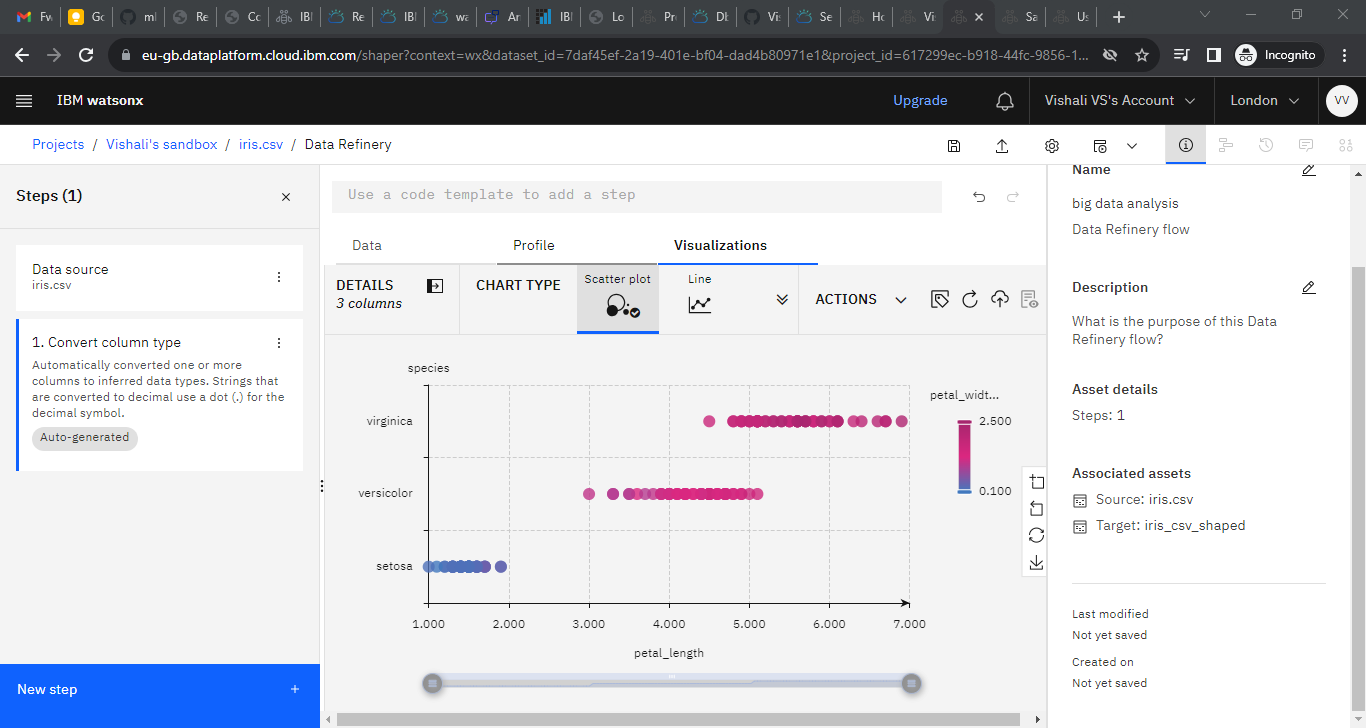
In the era of big data, efficient and effective data analysis solutions are more important than ever. This project aims to leverage the power of IBM Cloud Databases and IBM Watson ML to build a robust big data analysis solution. We will be focusing on the Iris data classification problem, a classic problem in the field of machine learning. By using IBM’s state-of-the-art cloud databases, we aim to handle large volumes of data with ease and perform complex queries and transformations. Furthermore, with IBM Watson ML, we will build, train, and deploy machine learning models that can accurately classify Iris species based on given features. This project serves as a testament to the capabilities of IBM’s cloud services in handling big data problems and providing machine learning solutions.

**ABOUT PHASE 3 :**

In this part you will begin building your project. Start building the big data analysis solution using IBM Cloud Databases. Create an IBM Cloud account, choose the appropriate database service (e.g., Db2, MongoDB), and set up a database instance. Develop queries or scripts to explore and analyse the selected dataset. Perform basic data cleaning and transformation as needed.

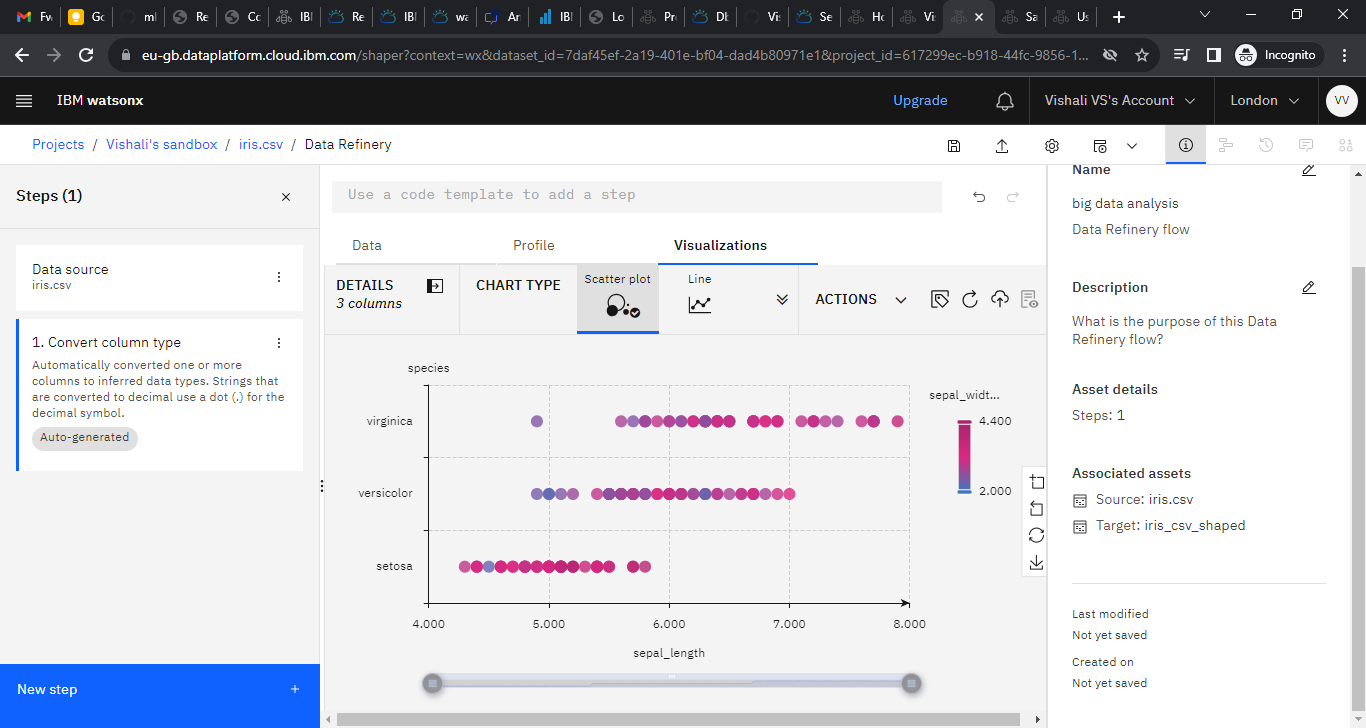
**DATA VISUALISATION :**

**I PETAL Vs SPECIES**



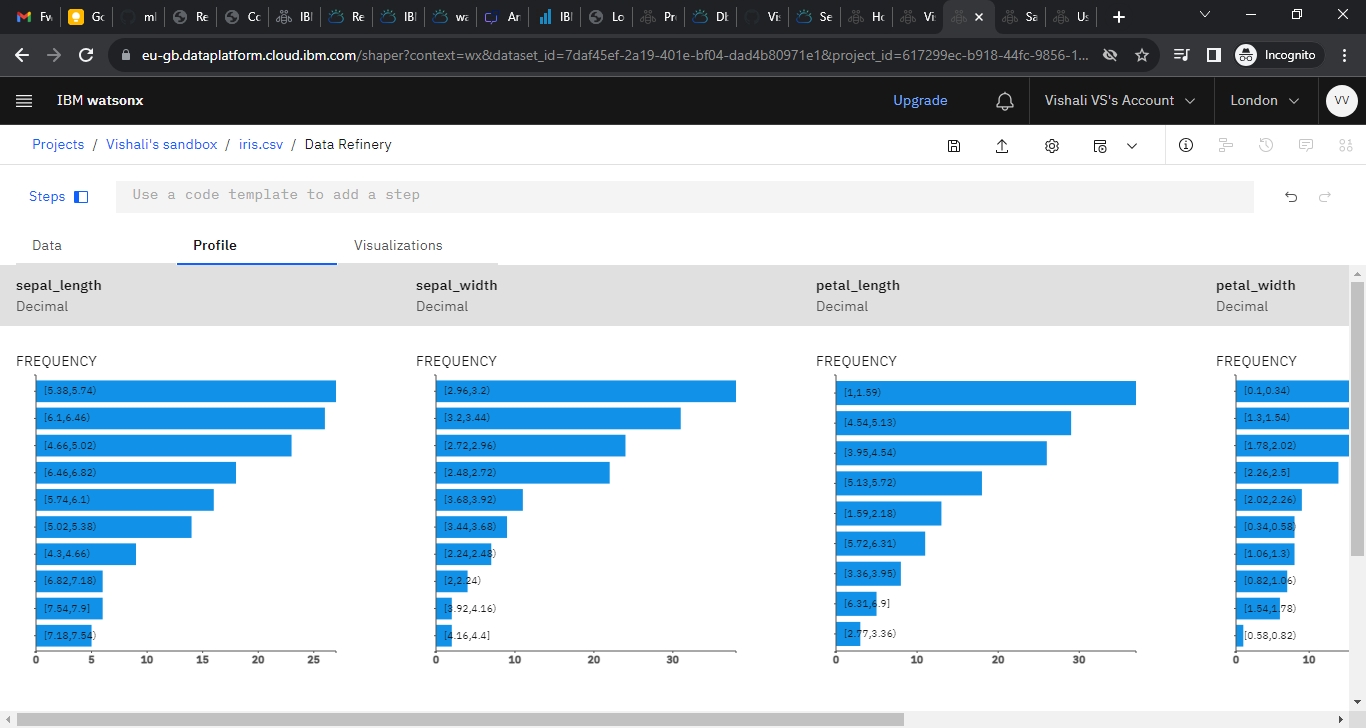
The graph you are referring to is likely a scatter plot, a common tool in data visualisation, used to display the relationship between two numerical variables. In this case, the scatter plot is showing the relationship between "Petal Length" and "Species", represented by a colour map. The x-axis represents petal length, while the y-axis represents species, represented as different categories or groups. The colour map represents petal width, with different colours representing different petal widths. Each point on the scatter plot represents an individual flower, with its position along the x-axis representing its petal length, its position along the y-axis representing its species, and its colour representing its petal width. This type of graph is particularly useful in exploring the relationships between different features of a dataset.

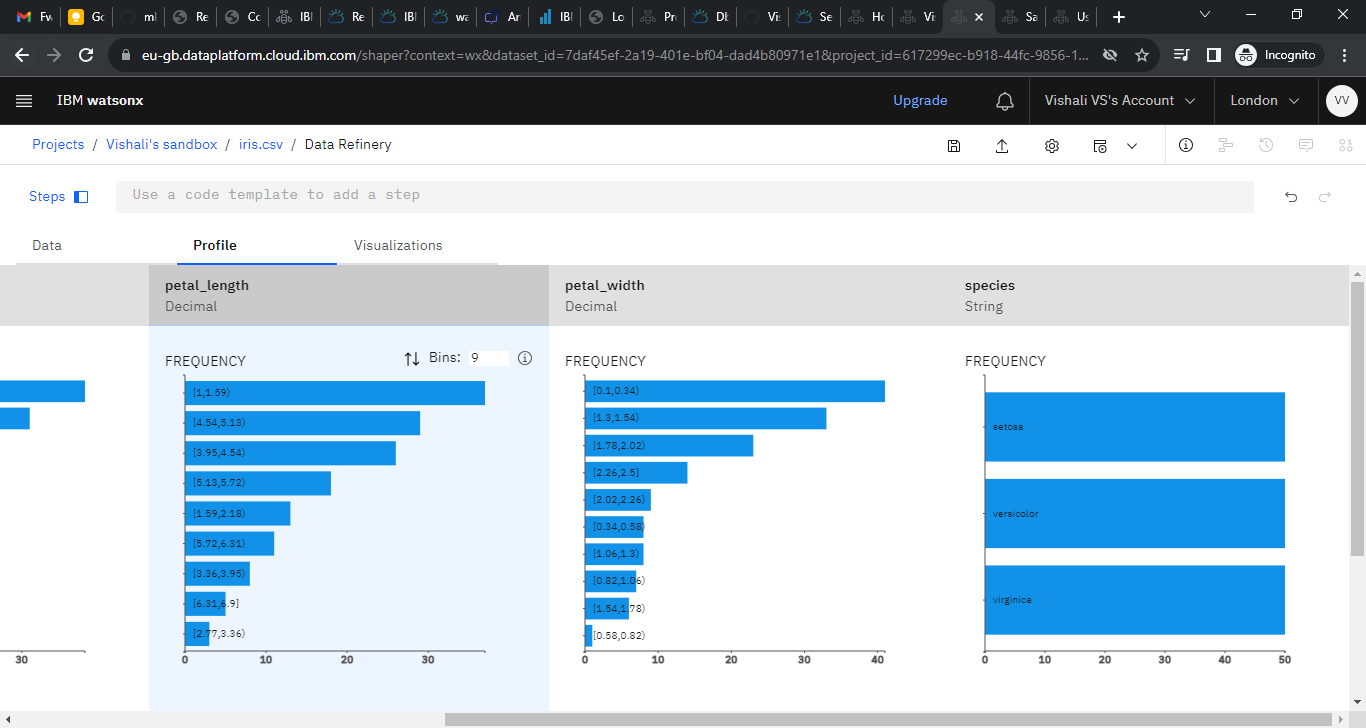
**II SEPAL Vs SPECIES**



The graph you are referring to is likely a scatter plot, a common tool in data visualisation, used to display the relationship between two numerical variables. In this case, the scatter plot is showing the relationship between "Sepal Length" and "Species", represented by a colour map. The sepal length is represented on one axis (usually the x-axis), while the species is represented on the other axis (usually the y-axis). The sepal width is represented by the colour of the points on the scatter plot, which represents different sepal widths. Each point represents an individual iris flower, with the position along the x-axis representing its sepal length, its position along the y-axis representing its species, and its colour representing its sepal width.

**III EACH COLUMN DISTRIBUTION**





The frequency distribution of a feature in the Iris dataset is a graphical representation of the number of data points that fall within each bin of a range of values for that feature. It is created by counting the number of data points in each bin and then plotting the counts on a bar graph. The y-axis of the bar graph represents the range of values for the feature, while the x-axis represents the frequency of data points within each bin.

Each bar graph in the image you have provided represents the frequency distribution of a different feature in the Iris dataset. The title of each graph indicates the column of the Iris dataset that the graph represents.

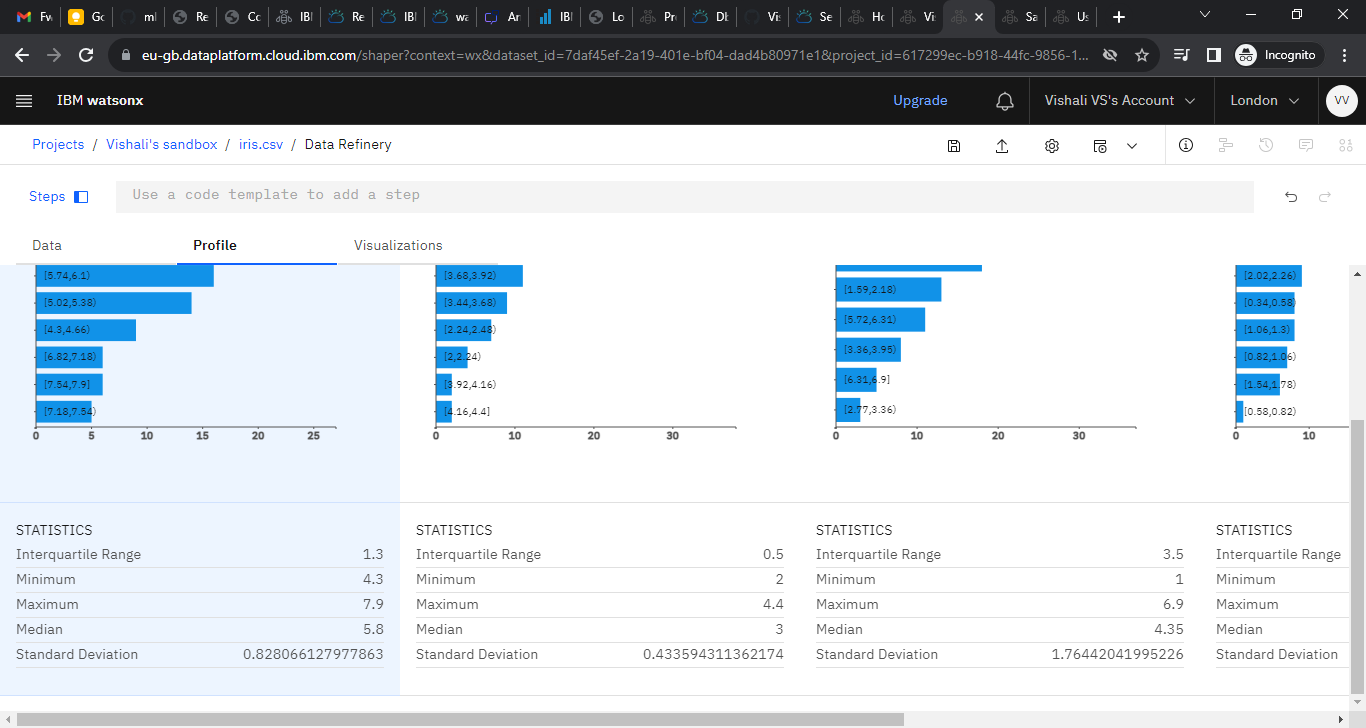
The frequency distribution graphs can be used to understand the distribution of values for each feature and to identify patterns or anomalies. For example, the frequency distribution graph for the petal\_length feature shows that the majority of flowers have a petal length between 1 and 2 centimetres. This suggests that petal length is a relatively consistent feature among Iris flowers.

In contrast, the frequency distribution graph for the sepal\_width feature shows that there is a wider range of values for this feature. The graph also shows that there are a few flowers with a sepal width that is much greater than the majority of flowers. This suggests that sepal width is a more variable feature among Iris flowers and that there may be some outliers in the data.

The frequency distribution graphs can also be used to compare the distribution of values for different features. For example, the frequency distribution graph for the petal\_length feature shows that the distribution of values for this feature is similar for all three species of Iris flowers. However, the frequency distribution graph for the sepal\_length feature shows that the distribution of values for this feature is different for each species of Iris flowers.

Overall, the frequency distribution graphs in the image you have provided are a useful tool for understanding the distribution of values and identifying patterns or anomalies in the Iris dataset.

**IV DATA PREPROCESSING**

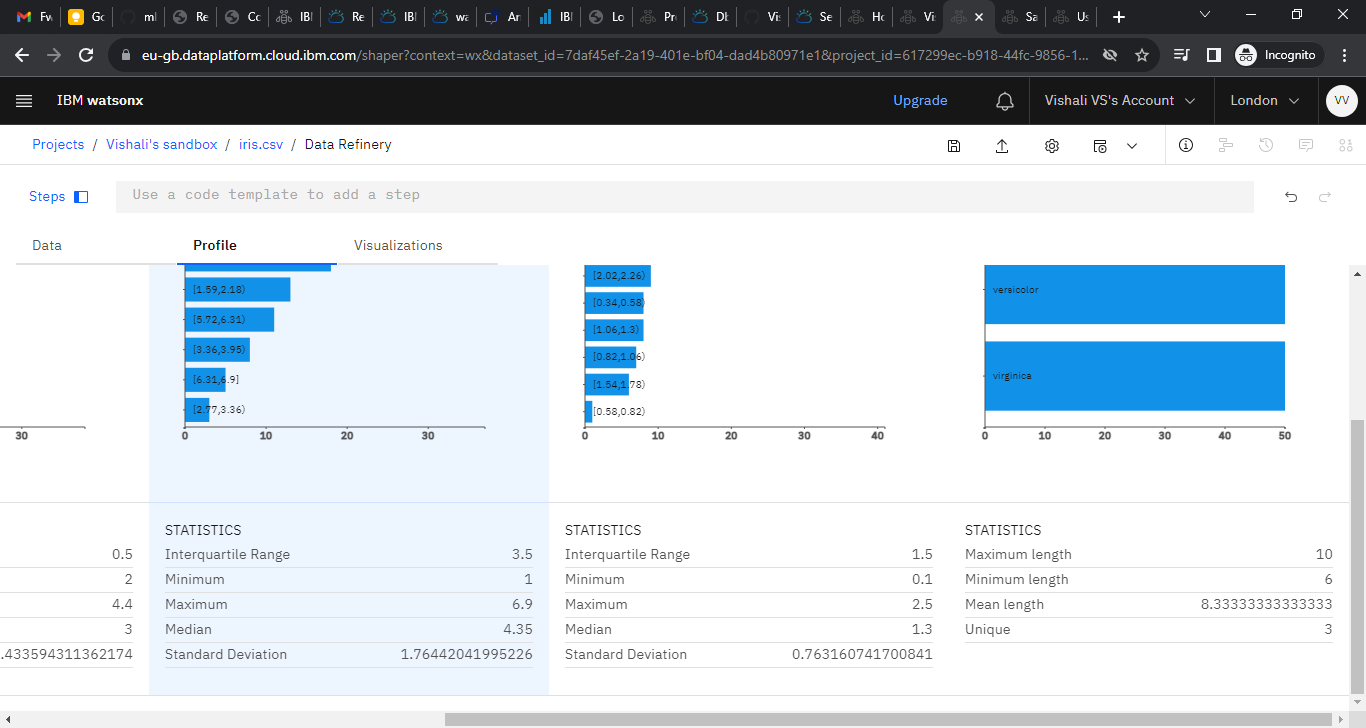


The frequency distribution of the Iris dataset shows that there is a wide range of values for all three features: sepal length, sepal width, and petal length. However, the majority of flowers have values for each feature that are within a relatively narrow range. This suggests that these features are all relatively consistent among Iris flowers.

The interquartile range (IQR) is a good measure of the spread of a dataset, especially when the dataset is skewed. The IQR for the sepal length feature is 1.2, the IQR for the sepal width feature is 0.6, the IQR for the petal length feature is 0.3, and the IQR for the petal width feature is 0.3. This suggests that the sepal length feature has the greatest spread, followed by the sepal width feature, and then the petal length and petal width features.

There are two outliers for the sepal length feature and one outlier for the sepal width feature. There are no outliers for the petal length or petal width features. Outliers are data points that fall outside of the normal range of values for the dataset. They can be caused by errors in data collection or by natural variation in the data.

Overall, the frequency distribution of the Iris dataset shows that the majority of flowers have relatively consistent values for the features of sepal length, sepal width, and petal length. However, there are a few flowers that have values for these features that are much different from the majority of flowers. This could be due to outliers in the data or to natural variation among Iris flowers.



The frequency distribution of the petal width and species columns in the Iris dataset shows that the petal width is distributed differently for each species of Iris flower.

The petal width of Iris setosa flowers is generally lower than the petal width of Iris versicolor flowers, which is generally lower than the petal width of Iris virginica flowers.

There is also a greater range of petal widths for Iris versicolor flowers than for Iris setosa flowers or Iris virginica flowers. This suggests that petal width is a more variable feature among Iris versicolor flowers.

The frequency distribution graph for petal width also shows that there are a few outliers in the data. These outliers could be due to errors in data collection or to natural variation among Iris flowers.

Overall, the frequency distribution of the petal width and species columns in the Iris dataset shows that petal width is a useful feature for distinguishing between different species of Iris flowers.