Phase 3: Big Data Analysis with IBM Db2 Cloud Database

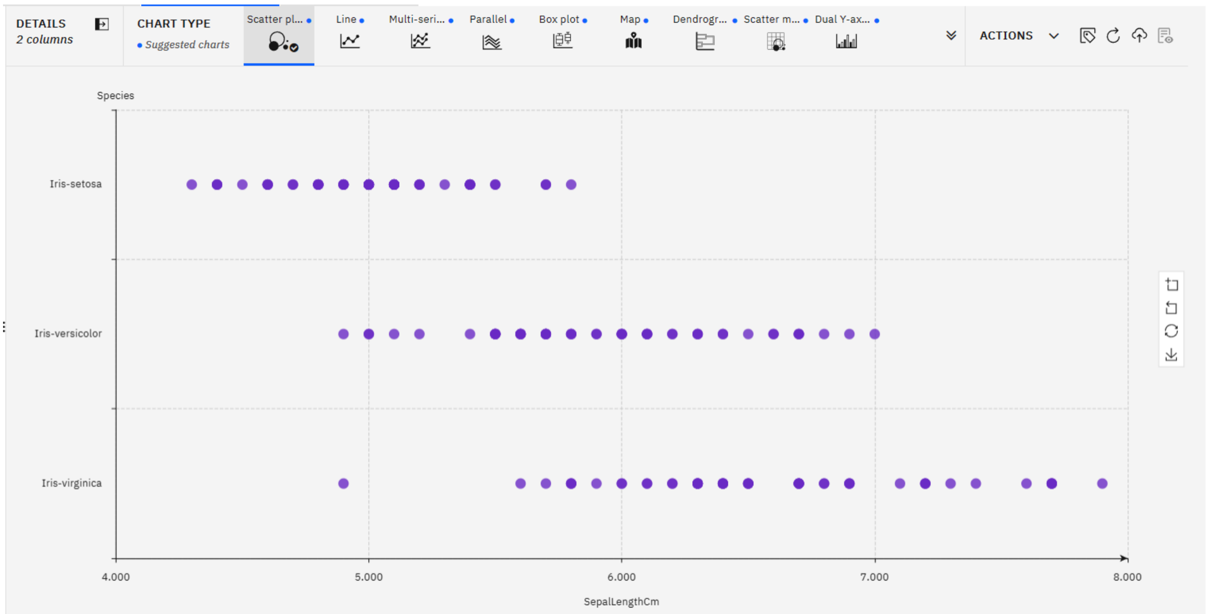
Introduction:

In today's data-driven world, having efficient data tools is crucial. Our project uses IBM Cloud Databases and IBM Watson ML to create a strong solution for analysing big data. We're focusing on the Iris data classification problem, a classic in machine learning. Using IBM's top-notch cloud databases, we'll handle lots of data easily and run complex queries. With IBM Watson ML and Db2, we'll load dataset, build, train, learning models to classify Iris species accurately. This project shows how well IBM's cloud services handle big data and provide practical machine learning solutions.

Now, let's get started on your project. Use IBM Cloud Databases to build the big data analysis solution. Make an IBM Cloud account, pick the right database service (like Db2 or MongoDB), and set up a database. Write queries or scripts to explore and analyse the data. Clean and transform the data as needed.

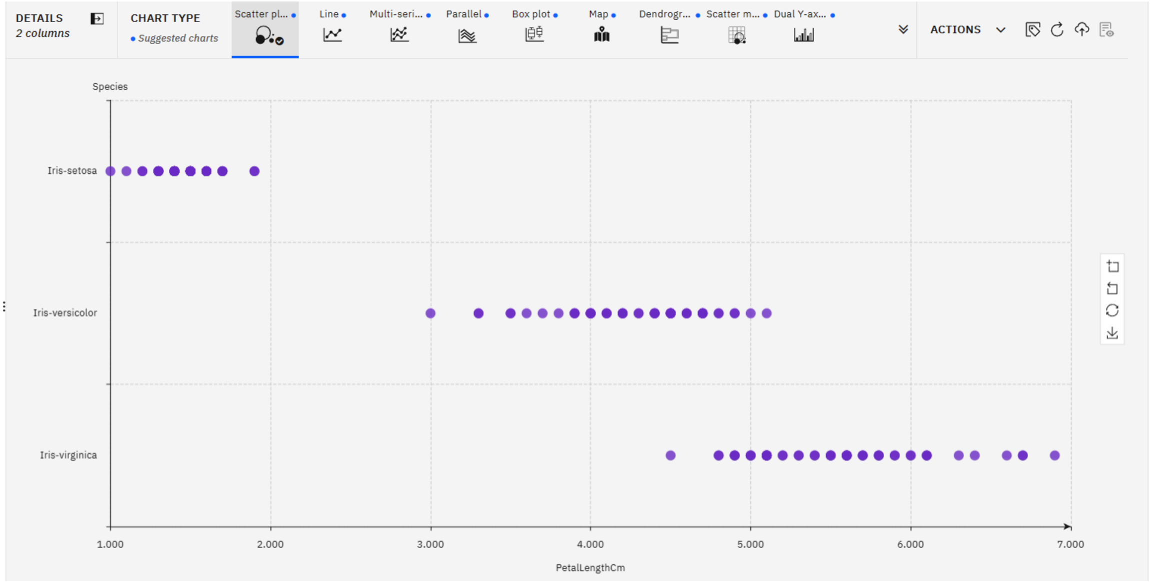
Data Visualisation:

1. Sepal Length and Species:



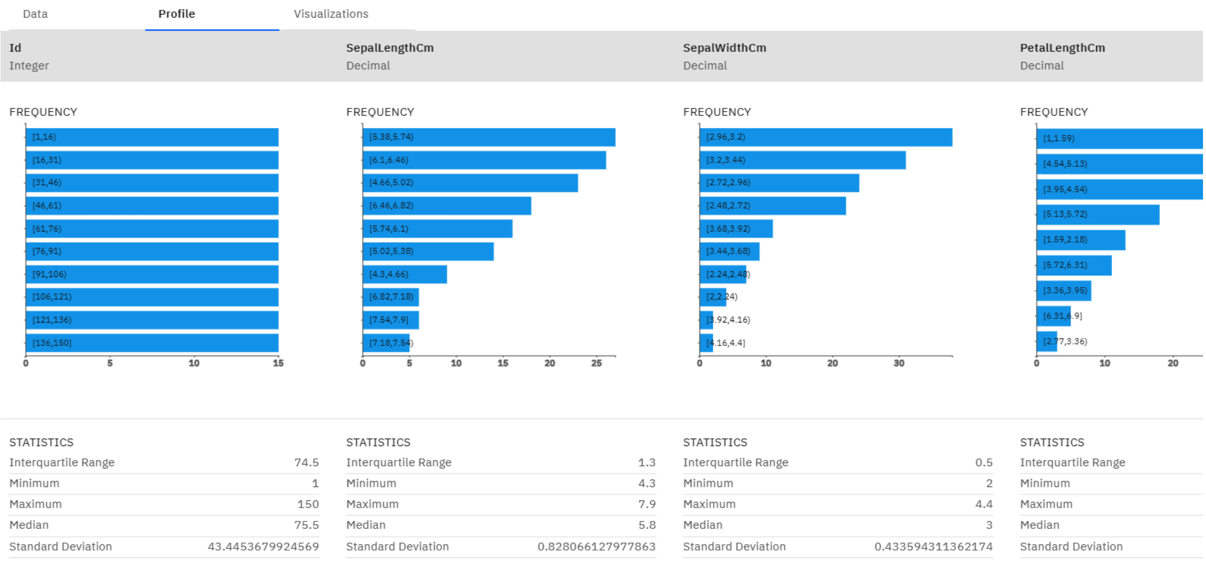
The graph you're seeing right now is a scatter plot, a common tool for showing the link between two numbers. This one display how "Sepal Length" and "Species" are connected, using colours to show different sepal widths. Sepal length is on one side (usually the x-axis), species on the other (usually the y-axis), and the dot colours show various sepal widths. Each dot is a single iris flower. Where it is on the x-axis is its sepal length, on the y-axis is its species, and its colour is its sepal width.

2. Petal Length and Species:

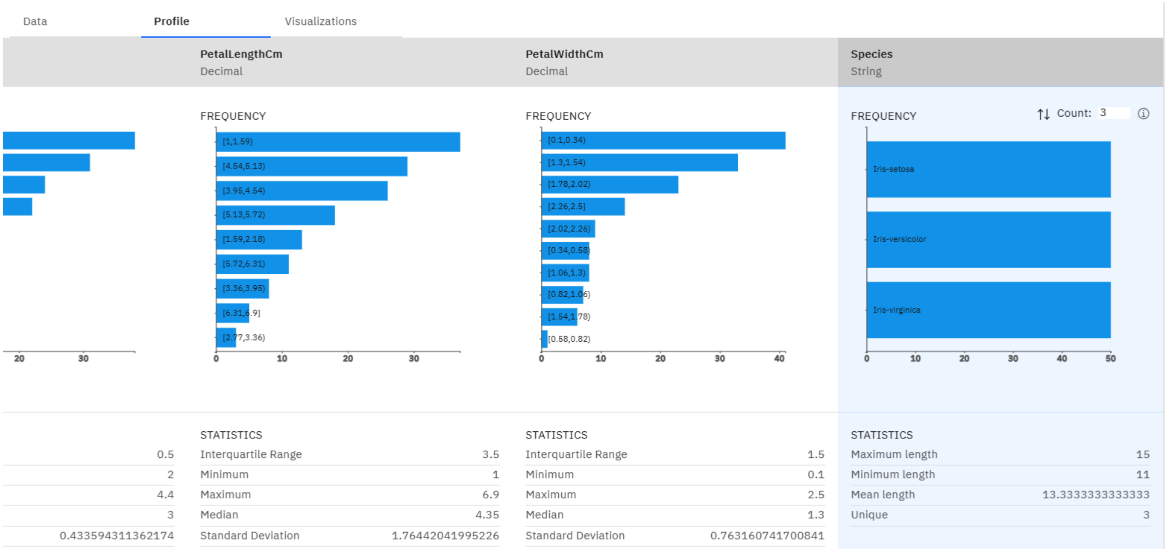


This one illustrates how "Petal Length" and "Species" are related, using colours to indicate different petal widths. The x-axis shows petal length, and the y-axis shows species as various groups. The colours show petal width. Each dot on the scatter plot is a single flower. Where it is on the x-axis is its petal length, on the y-axis is its species, and its colour is its petal width. This kind of graph is handy for exploring relationships in a dataset.

3. Each Attribute Distribution:



STATISTICS	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm
Interquartile Range	74.5	1.3	0.5	
Minimum	1	4.3	2	
Maximum	150	7.9	4.4	
Median	75.5	5.8	3	
Standard Deviation	43.4453679924569	0.828066127977863	0.433594311362174	



	PetalLengthCm Decimal		PetalWidthCm Decimal		Species String	
	STATISTICS		STATISTICS		STATISTICS	
0.5	Interquartile Range	3.5	Interquartile Range	1.5	Maximum length	15
2	Minimum	1	Minimum	0.1	Minimum length	11
4.4	Maximum	6.9	Maximum	2.5	Mean length	13.3333333333333
3	Median	4.35	Median	1.3	Unique	3
0.433594311362174	Standard Deviation	1.76442041995226	Standard Deviation	0.763160741700841		

The frequency distribution in the Iris dataset shows how many data points fall into different value ranges for a feature. It's made by counting the data points in each range and putting the counts on a bar graph. The y-axis on the graph shows the feature's value range, and the x-axis shows how often data points fall into each range.

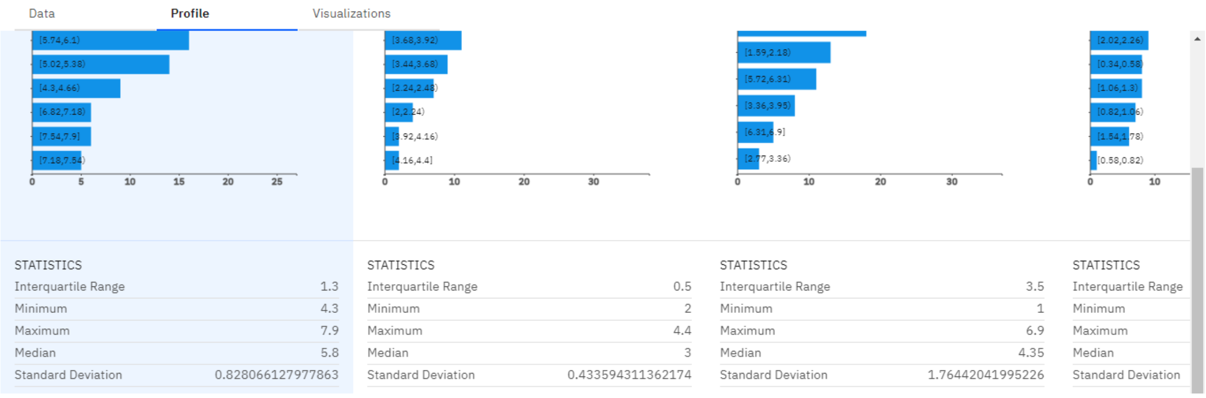
Each bar graph in the image represents the frequency distribution of a different feature in the Iris dataset. The graph's title tells you which column of the Iris dataset it represents.

These graphs help us understand how values are spread for each feature and spot patterns or anomalies. For example, looking at the petal length graph, most flowers have a petal length between 1 and 2 cms. This suggests that petal length is quite consistent among Iris flowers.

On the other hand, the sepal width graph shows a wider range of values. It also reveals a few flowers with much greater sepal width than the majority, hinting at variability and potential outliers.

These graphs also let us compare value distributions for different features. The petal length graph indicates similar distributions for all three Iris species, while the sepal length graph shows variations between species.

4. Preprocessing Data:



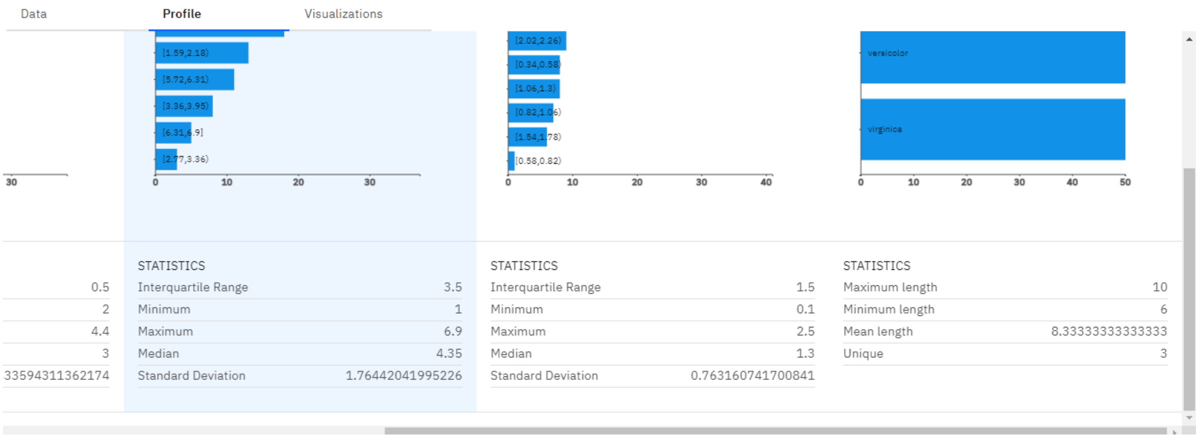
STATISTICS		STATISTICS		STATISTICS		STATISTICS
Interquartile Range	1.3	Interquartile Range	0.5	Interquartile Range	3.5	Interquartile Range
Minimum	4.3	Minimum	2	Minimum	1	Minimum
Maximum	7.9	Maximum	4.4	Maximum	6.9	Maximum
Median	5.8	Median	3	Median	4.35	Median
Standard Deviation	0.828066127977863	Standard Deviation	0.433594311362174	Standard Deviation	1.76442041995226	Standard Deviation

The Iris dataset's frequency distribution reveals a broad range of values for sepal length, sepal width, and petal length. However, most flowers have values within a relatively narrow range for each feature, suggesting consistency among Iris flowers.

To measure data spread, especially in skewed datasets, we use the interquartile range (IQR). For sepal length, the IQR is 1.2, for sepal width, it's 0.6, for petal length, it's 0.3, and for petal width, it's 0.3. This indicates that sepal length has the greatest spread, followed by sepal width, and then petal length and petal width.

There are two outliers for sepal length and one for sepal width, but none for petal length or petal width. Outliers are data points outside the normal value range, possibly caused by errors or natural variation.

In summary, the Iris dataset's frequency distribution shows that most flowers have consistent values for sepal length, sepal width, and petal length. Yet, a few flowers have significantly different values, potentially due to outliers or natural variation among Iris flowers.



	0.5
	2
	4.4
	3
	33594311362174

STATISTICS

Interquartile Range	3.5
Minimum	1
Maximum	6.9
Median	4.35
Standard Deviation	1.76442041995226

STATISTICS

Interquartile Range	1.5
Minimum	0.1
Maximum	2.5
Median	1.3
Standard Deviation	0.763160741700841

STATISTICS

Maximum length	10
Minimum length	6
Mean length	8.33333333333333
Unique	3

The frequency distribution of petal width and species in the Iris dataset reveals distinct patterns for each Iris flower species.

Generally, Iris setosa flowers have lower petal width compared to Iris versicolor, which, in turn, tends to have lower petal width than Iris virginica.

Moreover, Iris versicolor flowers exhibit a broader range of petal widths compared to Iris setosa and Iris virginica, suggesting more variability in this feature for Iris versicolor.

Examining the frequency distribution graph for petal width, there are a few outliers. These outliers might be caused by data collection errors or natural variations among Iris flowers.

In summary, the frequency distribution of petal width and species in the Iris dataset highlights petal width as a valuable feature for distinguishing between different Iris flower species.