Big Data Analysis with IBM cloud Databases

1. Project Objective

In this project, our primary objective is to harness the power of IBM Cloud services to analyze and visualize business data using Python. Through a combination of data exploration, machine learning modelling, and data visualization, we aim to extract actionable insights from the Iris dataset. This project serves as a comprehensive example of how IBM Cloud services can be employed to make data-driven decisions and solve real-world business challenges.

2. Design Thinking Process

a. Visualizing Data with IBM Cloud Services

To embark on our data analysis journey, we kickstart the project by leveraging IBM Cloud services, specifically Watson Studio, to visualize our business data. Watson Studio is a robust platform that empowers us to efficiently explore, clean, and prepare the data for analysis. Using Python as our programming language, we conduct a systematic and iterative process of data visualization.

b. Analysing Data with Machine Learning Models

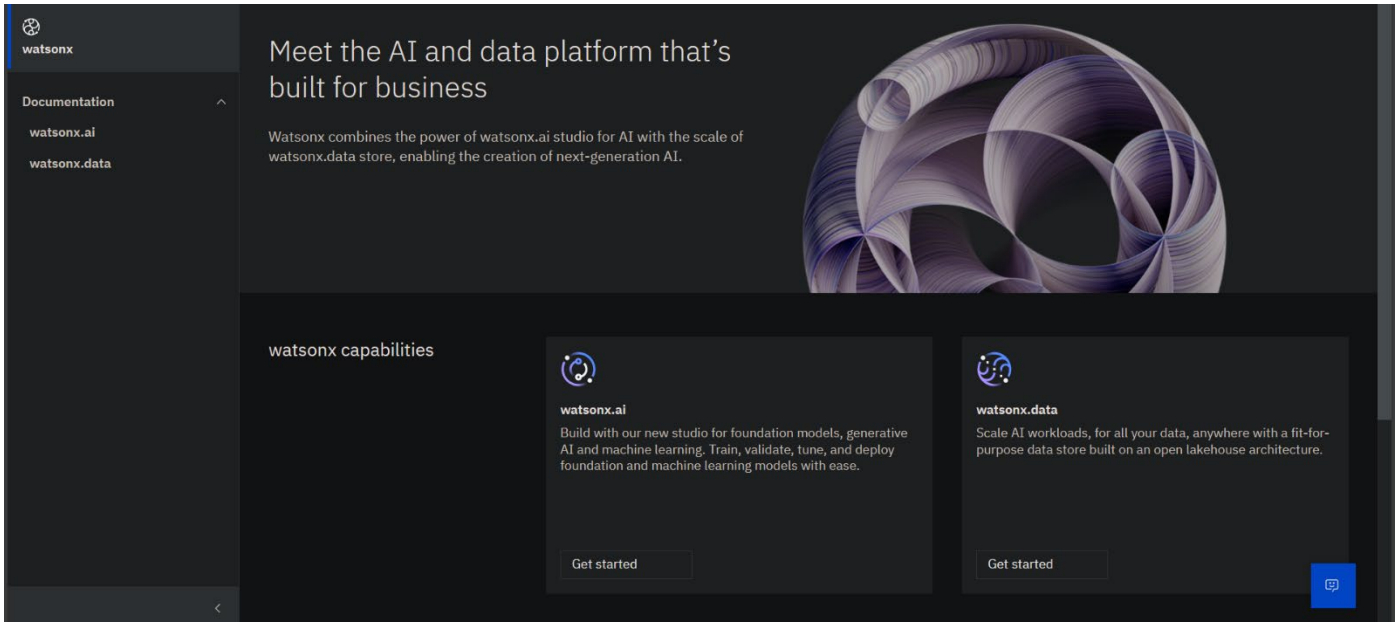
Our project employs two prominent machine learning models: the Decision Tree Classifier and the K-Nearest Neighbours (KNN) algorithm. These models, implemented using Python, act as the engines that drive our data analysis. The goal is to unearth patterns, relationships, and predictive insights within the Iris dataset. The following sections elaborate on the various phases of development and how we use these models to extract knowledge from the data.

3. Development Phases

The development of this project is a structured process, comprising the following phases:

1. **Data Collection:** We commence by acquiring the Iris dataset, renowned for its simplicity and relevance to classification tasks.

2. **Data Preprocessing:** Employing Python within Watson Studio (WatsonX), we undertake the essential data cleaning and preprocessing. This step is crucial for ensuring data quality, consistency, and reliability.



3. **Model Building:** Our analysis takes shape through the creation of the Decision Tree Classifier and K-Nearest Neighbours models. These models lay the foundation for the subsequent steps.

4. **Model Training:** Using the Iris dataset, we train our machine learning models. The models learn from the provided data, enabling them to make informed predictions.

5. **Evaluation and Validation:** We rigorously assess the performance of our models to verify their accuracy and reliability.

6. **Visualization:** The data analysis process reaches its pinnacle as we employ Python libraries, such as Matplotlib and Seaborn, to craft insightful visual representations of the data. Visualizations are invaluable for conveying complex information effectively.

7. **Insights Generation:** At this stage, we extract valuable insights from the analysis results, revealing patterns, trends, and outliers within the Iris dataset.

4. Selected Dataset

For the purpose of this project, the Iris dataset was thoughtfully chosen. This dataset comprises measurements of three distinct species of iris flowers, providing a diverse and manageable dataset for demonstration. The Iris dataset's simplicity allows us to focus on the core principles of data analysis and the capabilities of IBM Cloud services.

5. Creating an Account in IBM

To kickstart your journey with IBM Cloud services, the initial step involves:

1. **Sign up for IBM Cloud:** If you don't already have an IBM Cloud account, you'll need to sign up for one. You can sign up at IBM Cloud.
2. **Log into IBM Cloud:** Once you have an IBM Cloud account, log in to the IBM Cloud console.
3. **Choose your payment plan:** Choose your payment plan to proceed further to use IBM cloud services.

6. Analysis & Database Setup

In this section, we will detail the steps to create a WatsonX environment within the IBM Cloud ecosystem. WatsonX is a powerful platform that streamlines data analysis and visualization tasks.

Steps to Create a Watson X Environment:

1. **Access the IBM Cloud Catalog:** To get started, navigate to the IBM Cloud platform and access the catalog of available services.

2. **Search for WatsonX:** Within the catalog, search for "WatsonX" and select it. This will initiate the process of creating your WatsonX environment.

3. **Create a WatsonX Environment:** Follow the prompts to create a WatsonX environment. You can choose the specifications that match your project's requirements.

4. **Create a Sandbox in WatsonX:** Once your WatsonX environment is set up, you can create a sandbox within it. A sandbox is a controlled environment for data analysis and experimentation. It ensures that your work is organized and isolated.

5. **Upload/Import the Iris Dataset:** In your WatsonX sandbox, you can upload or import the Iris dataset. This dataset will serve as the foundation for your data analysis and visualization tasks.

6. **Create a New Task:** Next, create a new task within WatsonX. Choose a task that is related to data analysis and visualization. WatsonX offers a variety of task options to suit your needs.

7. **Analyze the Data:** WatsonX simplifies data analysis by providing several options for graph plotting based on the data. You can leverage these built-in tools to explore and visualize the data automatically.

Analysis and Visualization Using WatsonX:

WatsonX offers a user-friendly interface that empowers you to interact with your data and perform analyses without the need for extensive coding. It provides a seamless environment to visualize your data and gain insights. However, it's important to note that you can also analyze the data by

writing Python code in an IBM Cloud notebook or proceed to analyze it locally in a Jupyter notebook, as described in the next section.

7. Analysis Techniques

While WatsonX provides a convenient platform for data analysis and visualization, you can also choose to harness the power of Python for a more tailored approach to analysis. Here are the steps to perform data analysis using Python in an IBM Cloud notebook:

1. **IBM Cloud Notebook:** Within your IBM Cloud environment, you have the option to create a notebook. These notebooks are Jupyter-based and allow you to write and execute Python code seamlessly.

2. **Data Analysis with Python:** Write Python code within the notebook to perform customized data analysis. This approach gives you full control over the analysis process and allows you to implement advanced techniques and models, such as the Decision Tree Classifier and K-Nearest Neighbors (KNN), which were previously mentioned.

3. **Data Visualization:** In your Python-based analysis, you can use libraries like Matplotlib and Seaborn to create informative visualizations. This level of customization enables you to tailor your graphs and charts to specific business requirements.

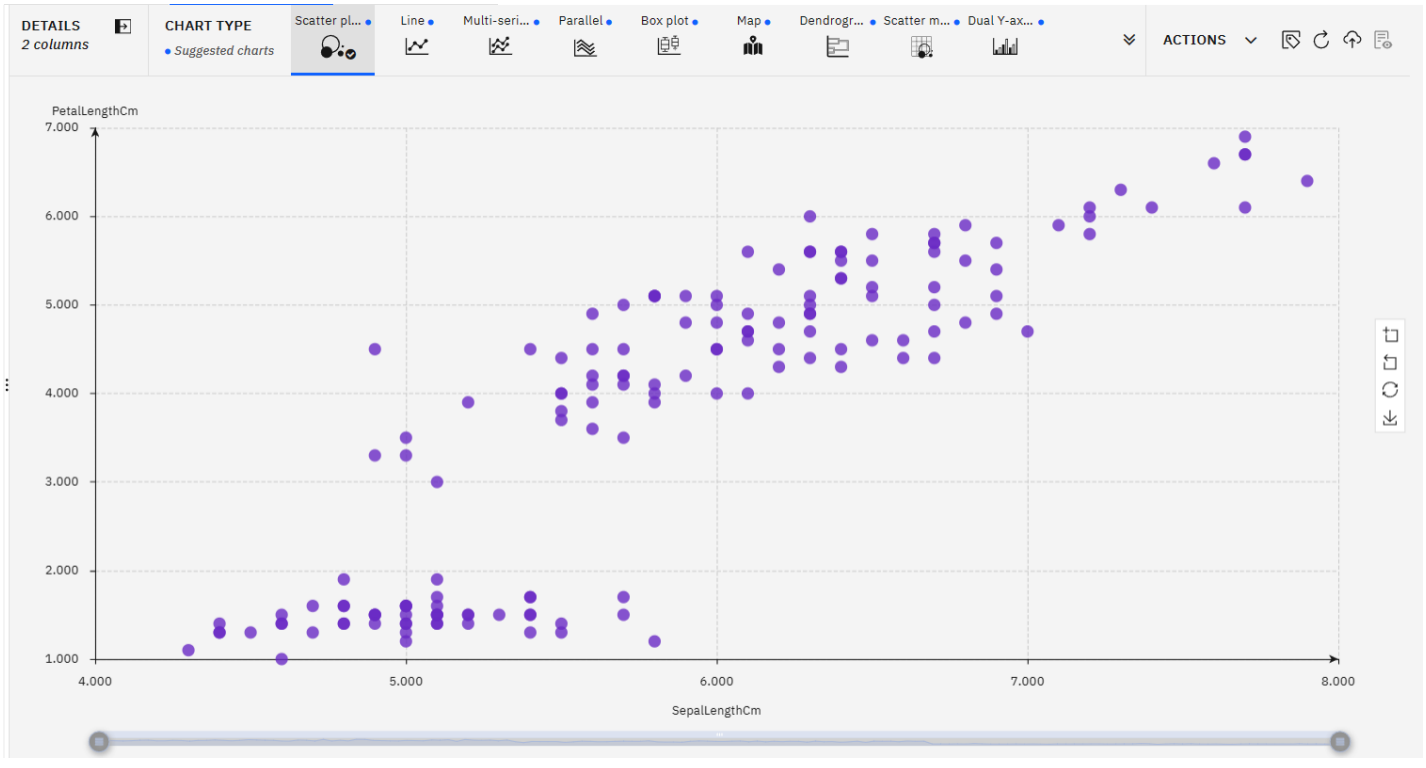
By providing these steps, we showcase the flexibility of IBM Cloud services. You have the choice to use WatsonX for a user-friendly, automated approach to data analysis and visualization, or you can opt for the versatility of Python for more customized and complex analyses. IBM Cloud empowers you to select the best approach to suit your project's needs.

8. Visualization Methods

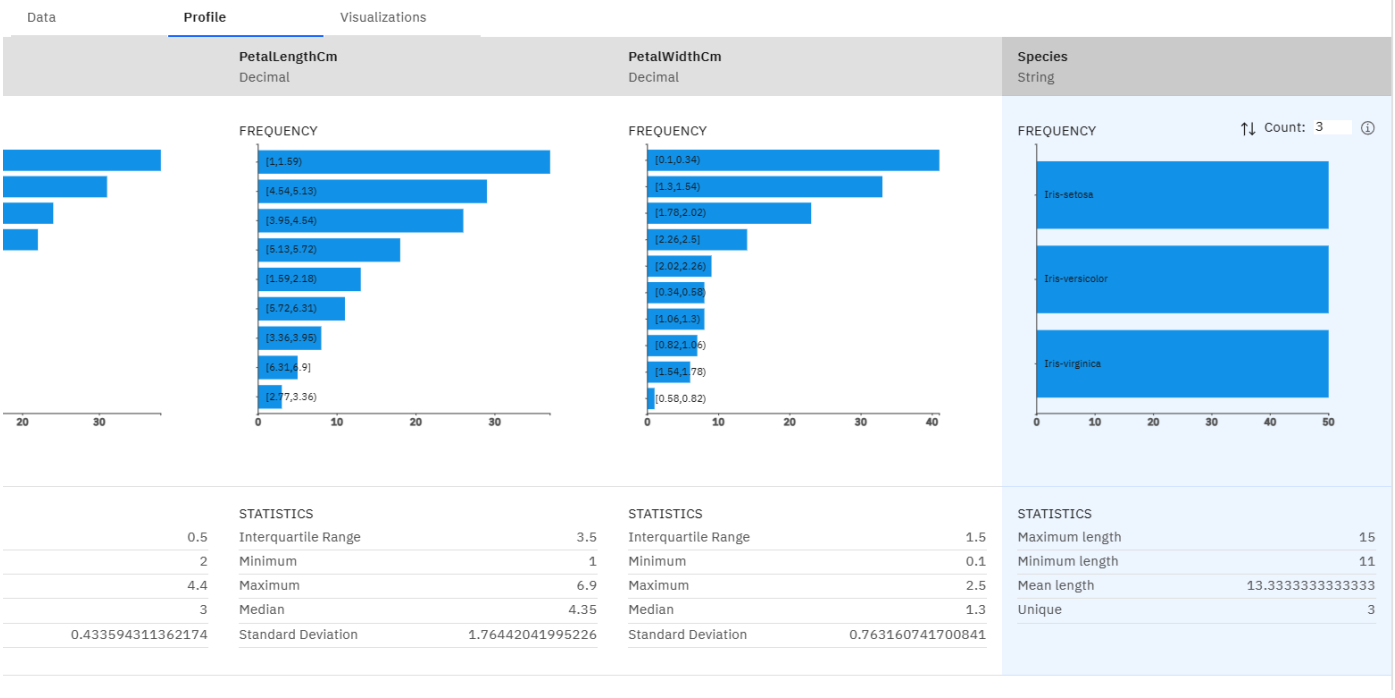
To make the data more comprehensible and to communicate our findings clearly, we employ Python libraries like **Matplotlib** and **Seaborn** for data visualization. These libraries enable us to create a wide range of informative charts, graphs, and plots that enhance our understanding of the data.

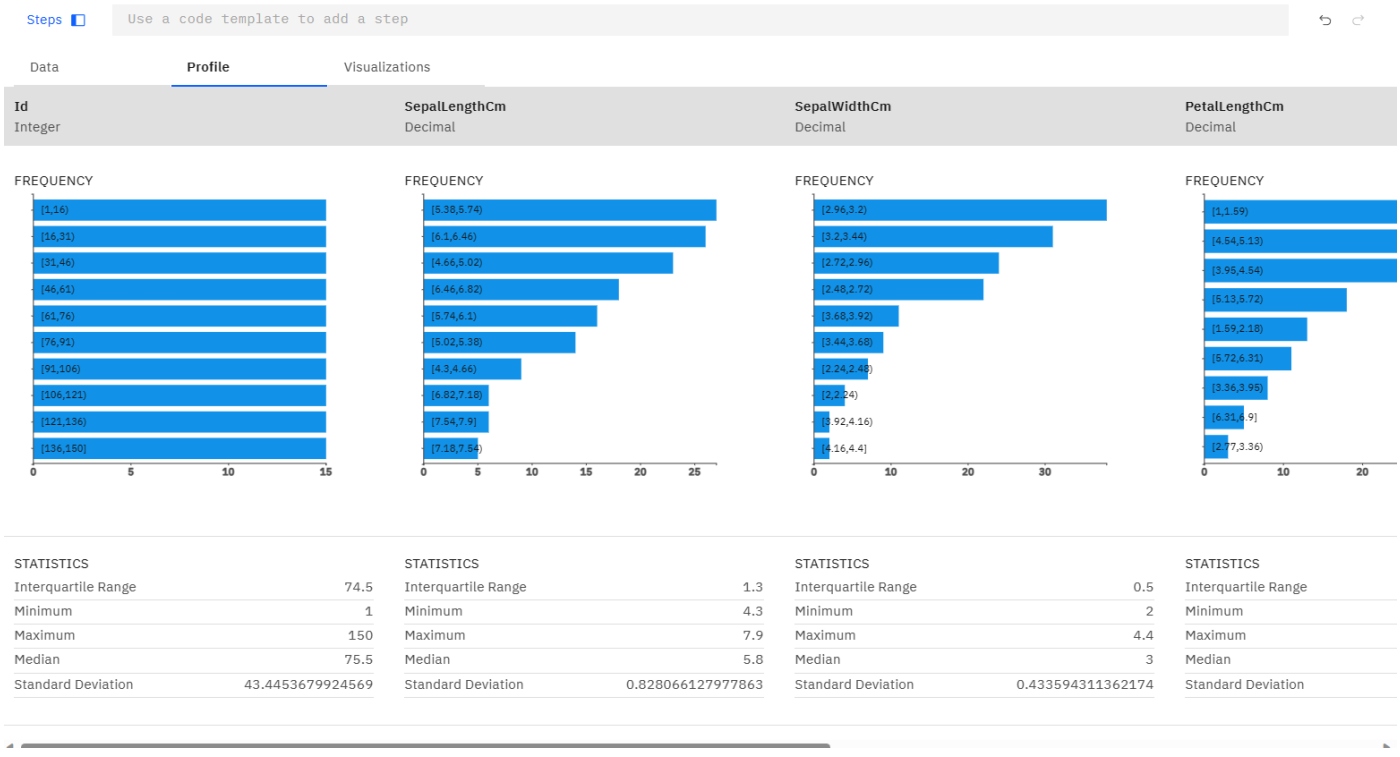
Below are some of the charts generated while finding relationship between 2 attributes:

Sepal Length vs Petal Length (Scatter Plot):

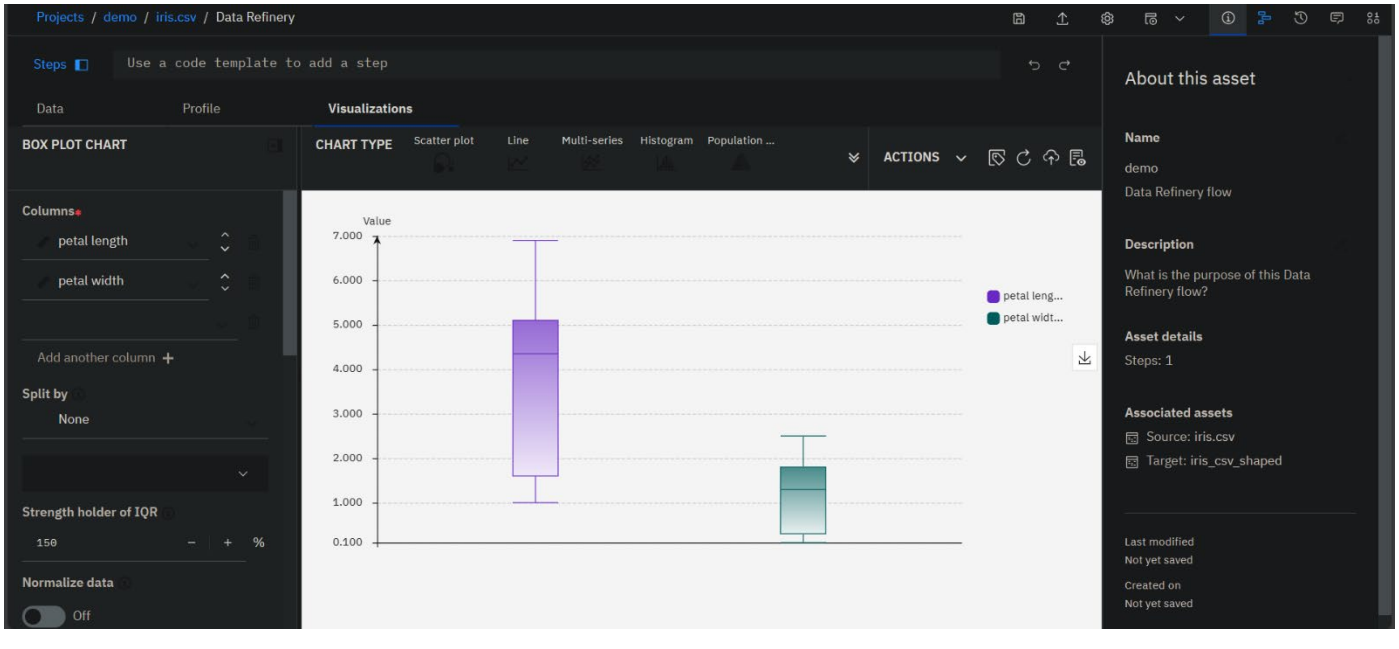


Summary of the Dataset (Bar Graph):

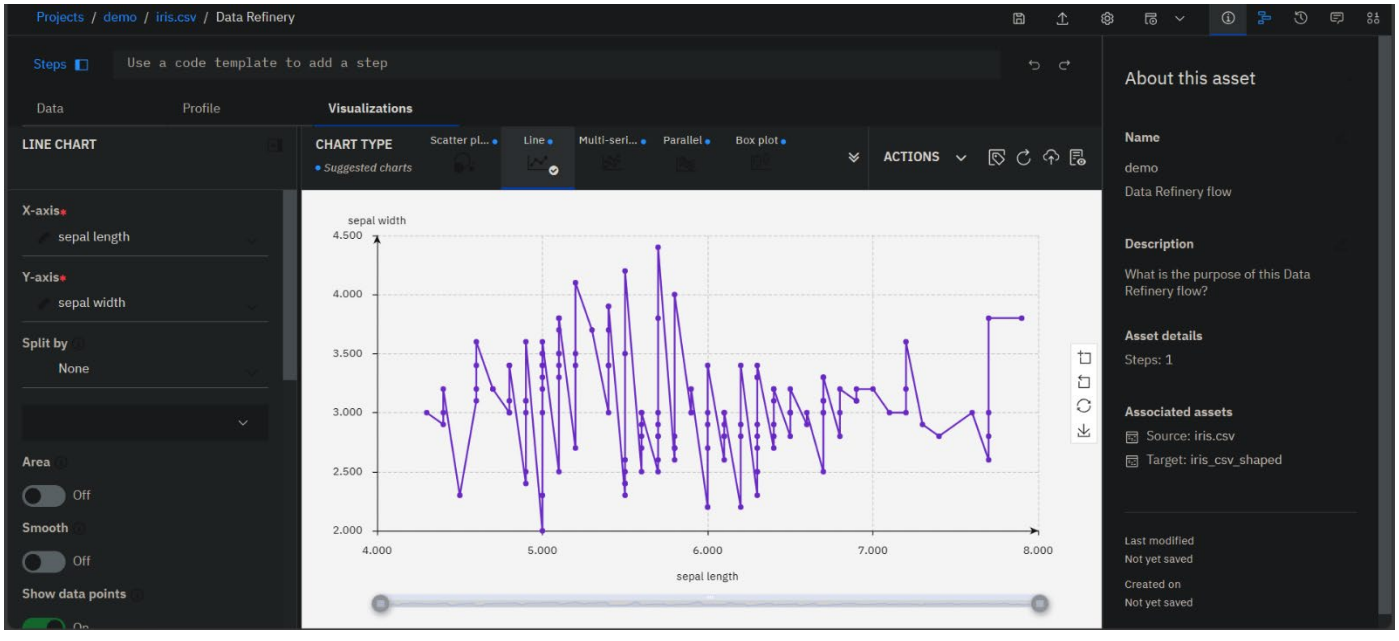




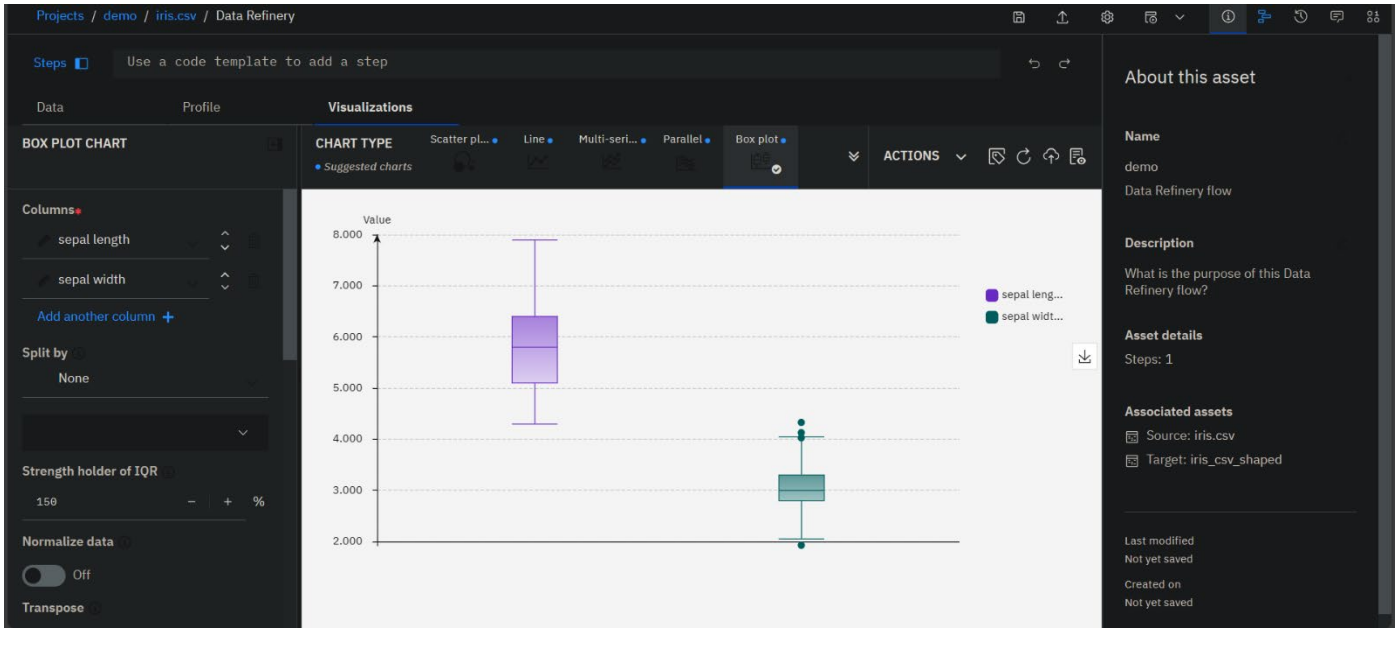
Petal Length vs Petal Width (Box plot):



Sepal Length vs Sepal Width (Line plot):



Sepal Length vs Sepal Width (Box plot):



9. Translating Findings into Business Insights

The ultimate goal of data analysis is to translate findings into actionable business insights. Through the use of IBM Cloud services and the Iris dataset, we demonstrate how raw data can be transformed into valuable knowledge. These insights have the potential to drive strategic decisions, optimize processes, and gain a competitive edge in the business world.

10. Conclusion

In conclusion, this project showcases the effective utilization of IBM Cloud services, Python, and machine learning models to analyse and visualize business data. The Iris dataset serves as a tangible example, illustrating the step-by-step process of turning raw data into actionable insights. By following the design thinking process and the structured development phases, we demonstrate how data-driven decision-making can be harnessed to solve real-world business challenges.