**INTEL UNNATI INDUSTRIAL TRAINING**

**Topic 12**

## **Knowledge Representation and Insights Generation from Structured Datasets**

**Team Name**

TECH TITANS

**Authors**

GADADASU GANESH

RAGURU NARENDRAKUMAR REDDY

GAJJALA ROHIT SAI

HARSHITA KUMARI

DODDA APARNA

**INTERNAL GUIDE**

THOTA SHIVA RATNA SAI

**Date**

**01-06-2023 to 15-07-2023**

**MALLAREDDY COLLEGE OF ENGINEERING AND TECHNOLOGY**

**ABSTRACT**

This report is written with purpose of narrating the project

completed during the one-month period of the Intel Unnati

Industrial Training program with Malla Reddy College of

Engineering and Technology, Hyderabad, Telangana.

The program commenced on 1st June 2023 and ended on 15th

July 2023.

The **Intel® Unnati Program** is focused on technology

inclusion, and advancing students’ skills in emerging

technology.

. The “**Knowledge representation and insights generation from a structured dataset”** project gain the insights from a structured dataset and represent the data in graphical form.

The project ended in success period of one month and results are shared through GItHub community.

**Introduction**

## Problem Statement:

## Knowledge Representation and Insights Generation from Structured Datasets

## Objectives:

* The primary objective of this project is to develop an Al-based solution that can effectively represent knowledge and generate insights from any structured dataset.

## The solution should be capable of processing and analyzing structured data, identifying patterns, and generating meaningful insights that can aid in decision-making processes

## Dataset Description

## Dataset source:

The dataset is taken from “UCI Machine Learning Repository” and the name of the dataset is “iris”.

## Key features:

* The iris dataset contains total of 150 samples with 5 features namely Sepal\_Length, Sepal\_Width, Petal\_Length, Petal\_Width, Species
* The Iris dataset is relatively clean and straightforward, making it an excellent choice for beginners in data science and machine learning

**Methodology**

## Methods used:

* Gradient Boosting Machine(GBM) Algorithm is used for training and testing the dataset
* Matplotlib and Seaborn are the libraries used to represent the data.
* Line plots, heat maps, pair plot, box plots and count plot are the specific visualizations used.
* Pattern identification techniques used in the code are discriptive statistics, data visualization and correlation anlysis

## Tools:

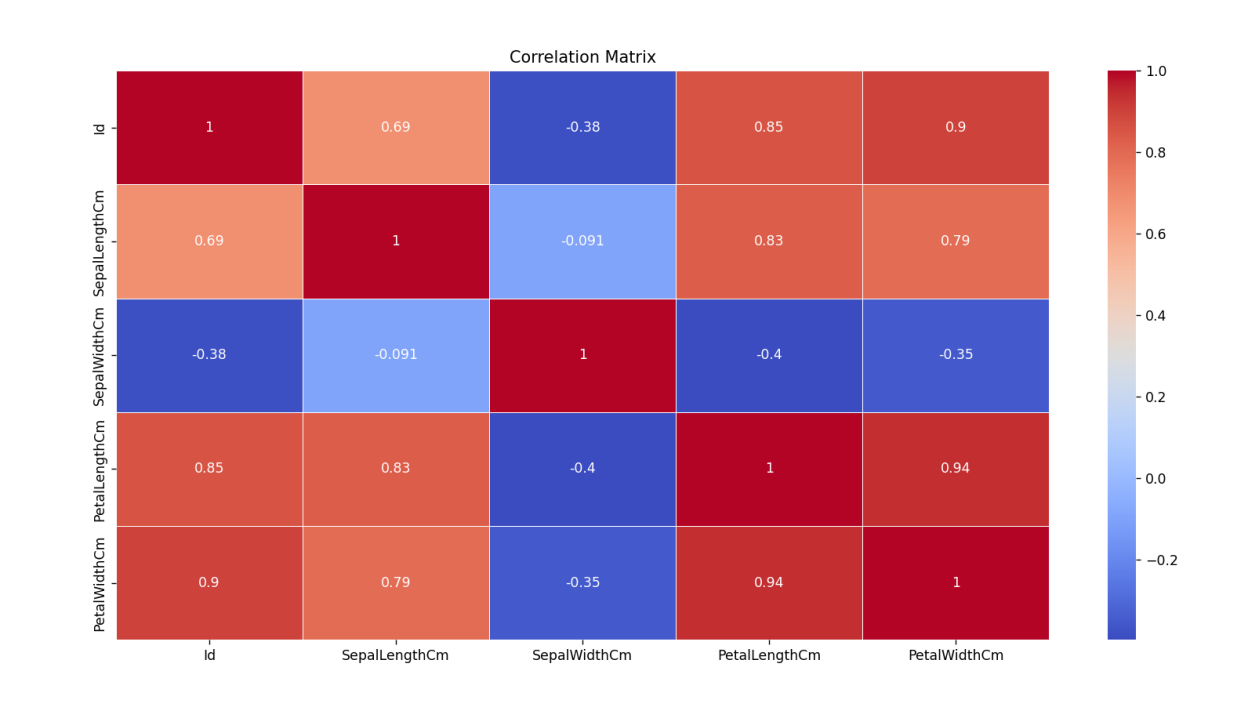
Python and python libraries

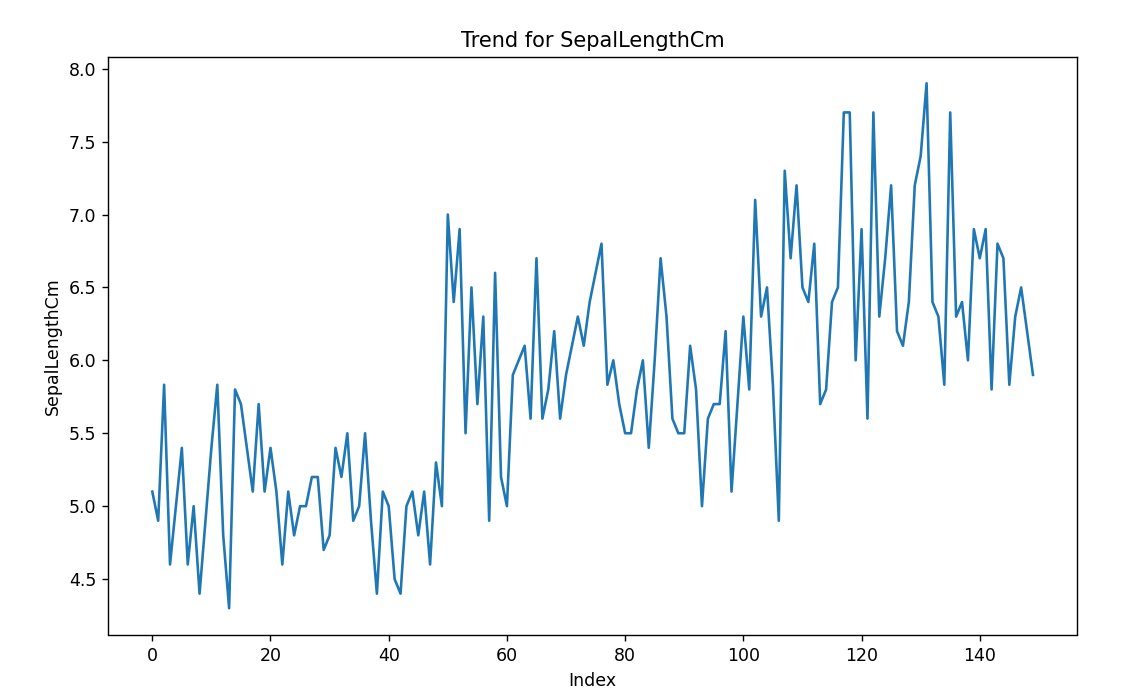
**Results and Discussion**

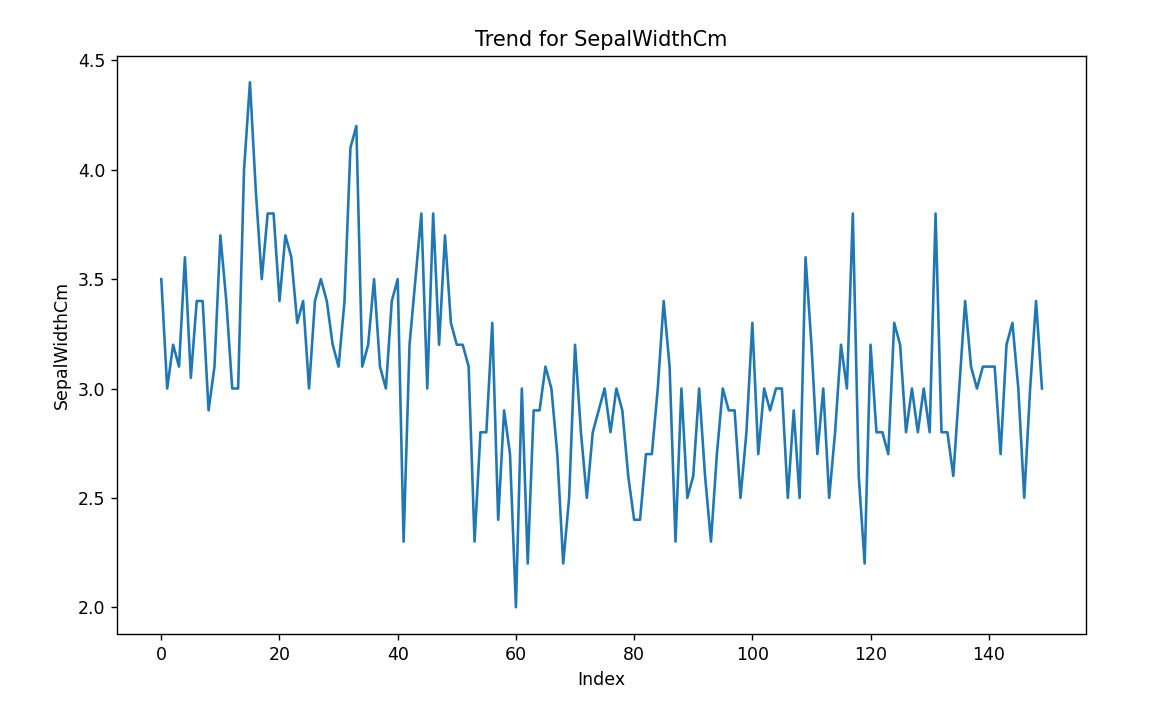
**Results:**

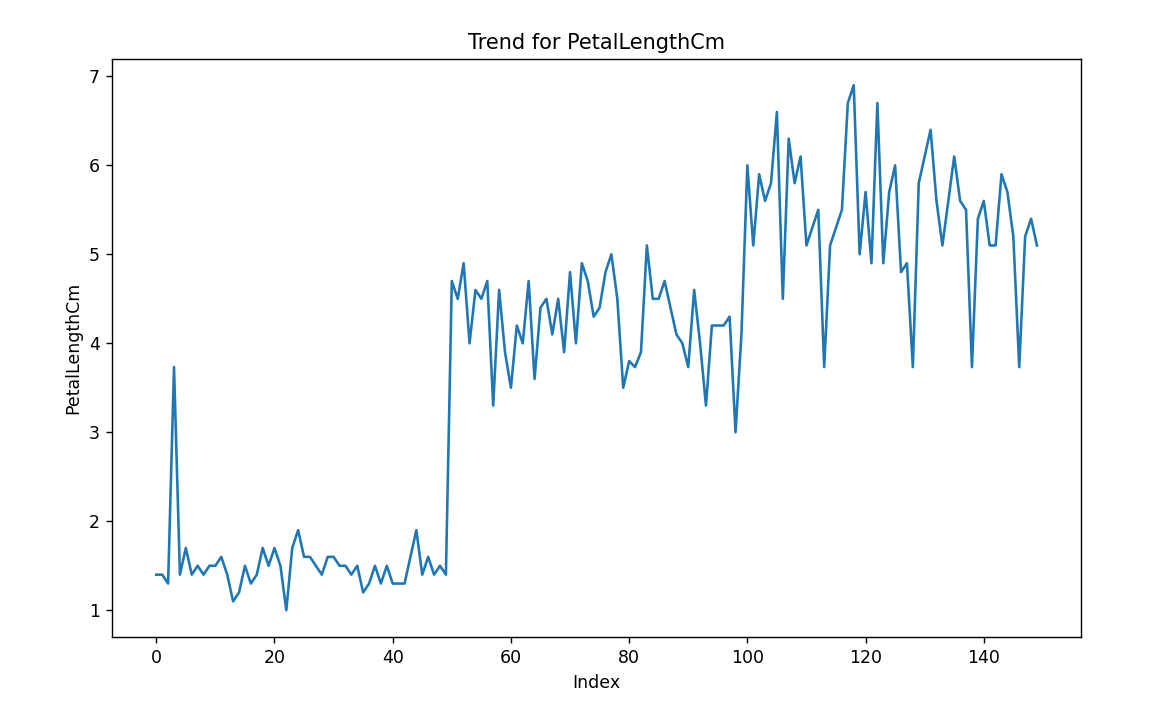
* *All columns are not having any Null Entries*
* *Four columns are numerical type only Single column categorical type*
* There are 3 duplicates and each species ( Iris virginica, setosa, versicolor) has 50 as it’s count
* Sepal Length and Sepal Width features are slightly correlated with each other
* Setosa is having smaller feature and less distributed.
* Versicolor is distributed in a average manner and average features
* Virginica is highly distributed with large no .of values and features

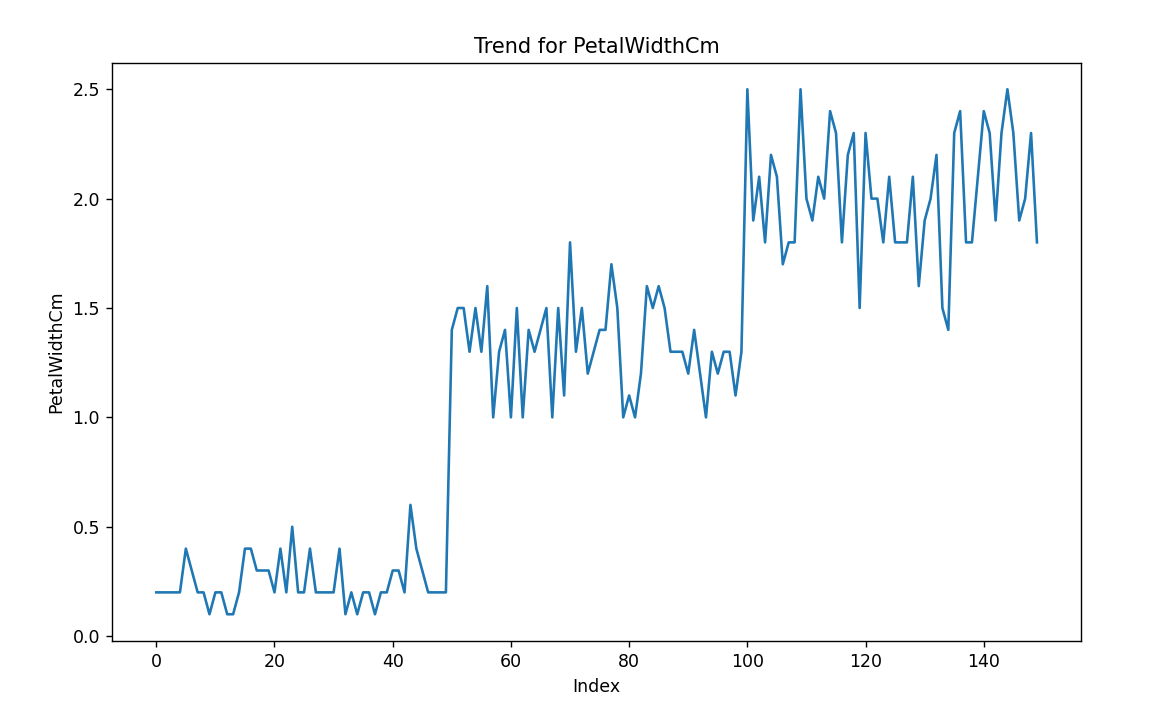
## Visualizations:

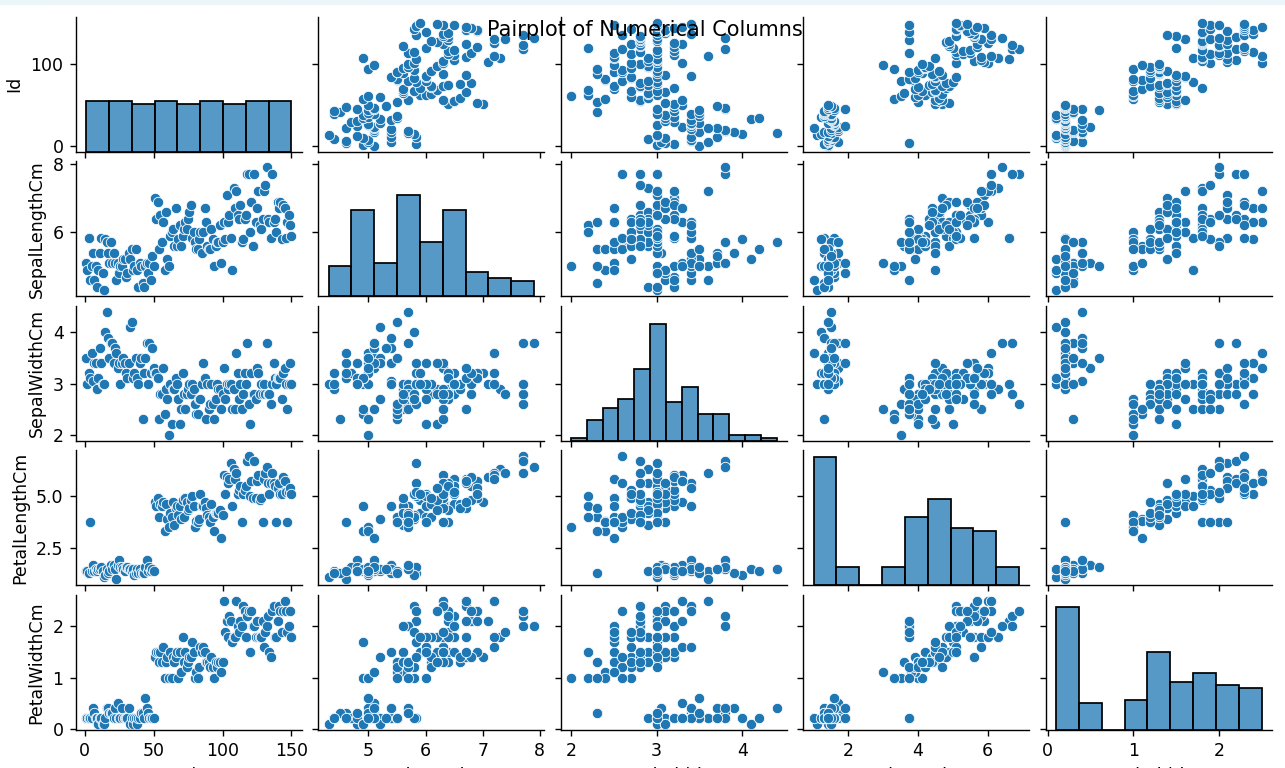


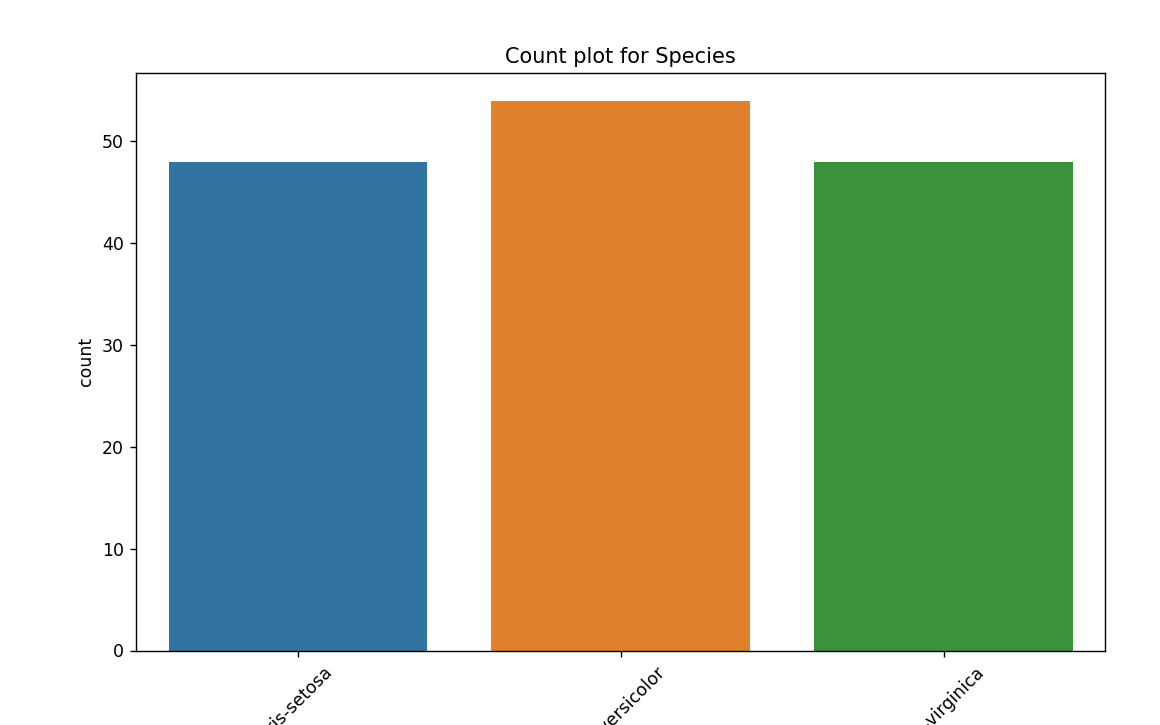


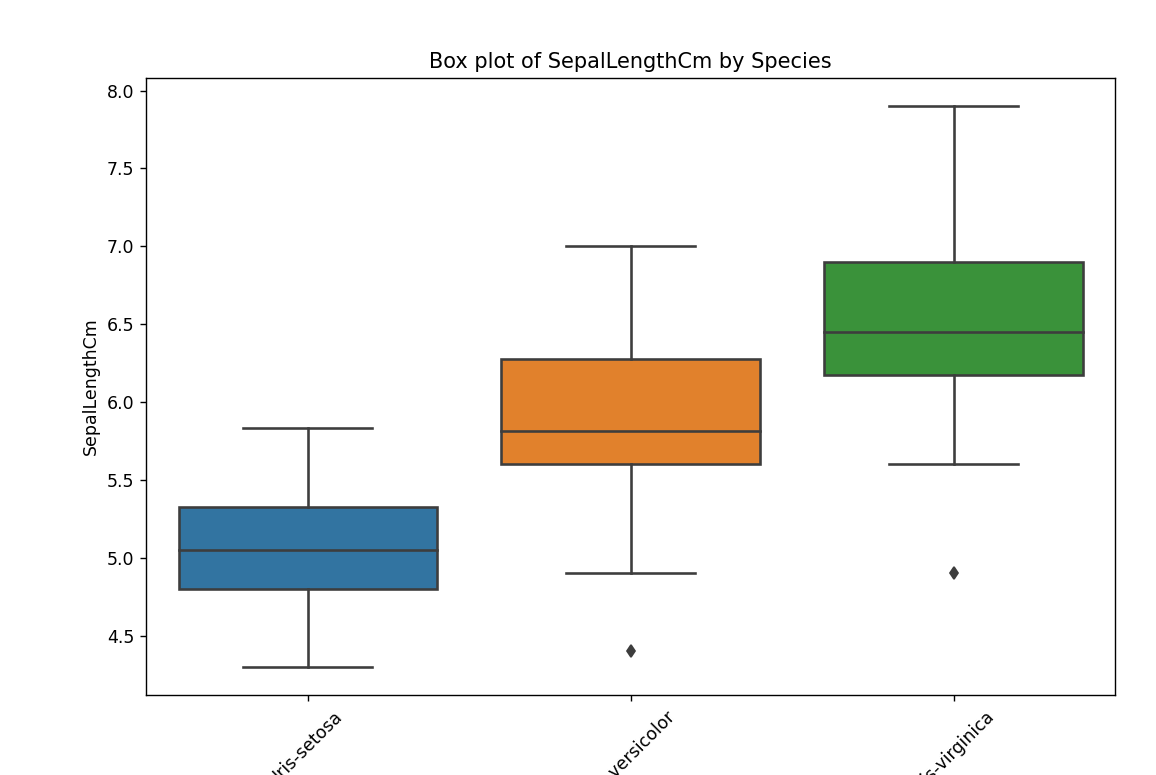
****

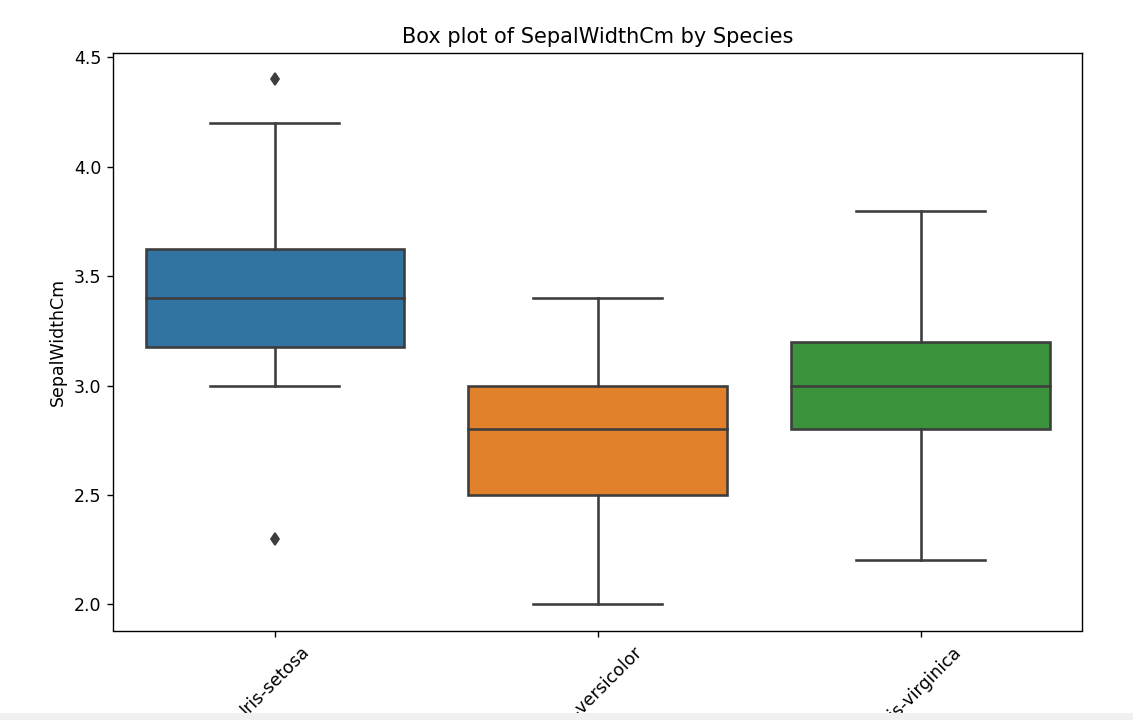
****

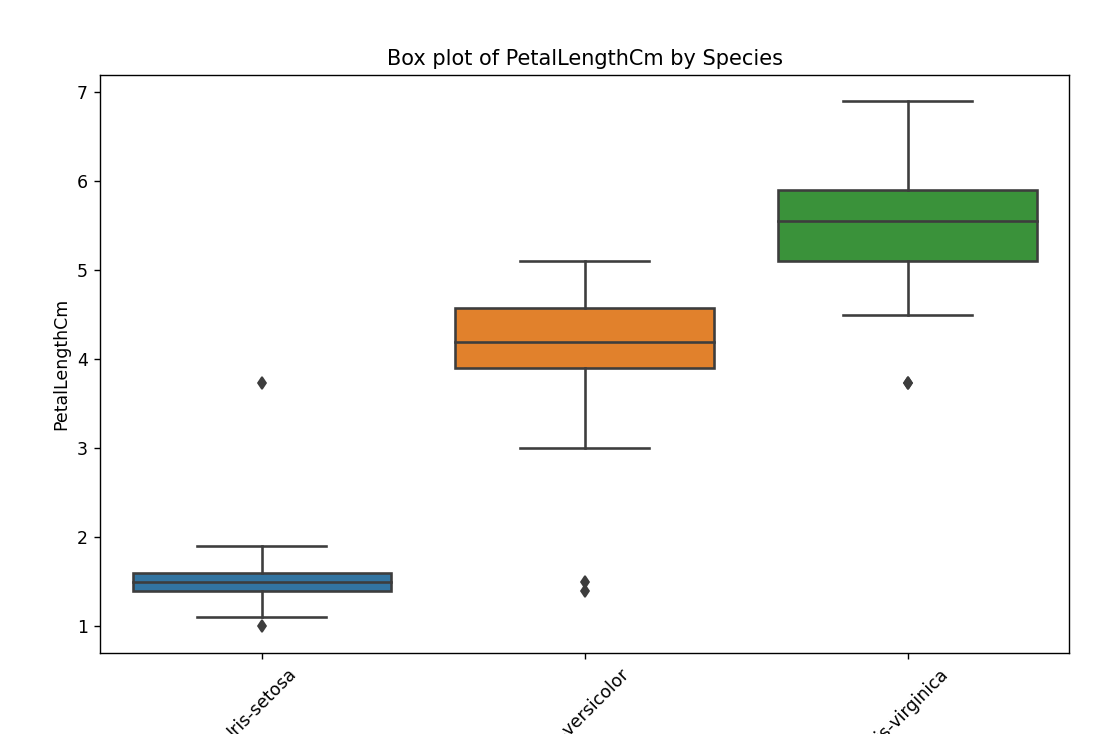
****

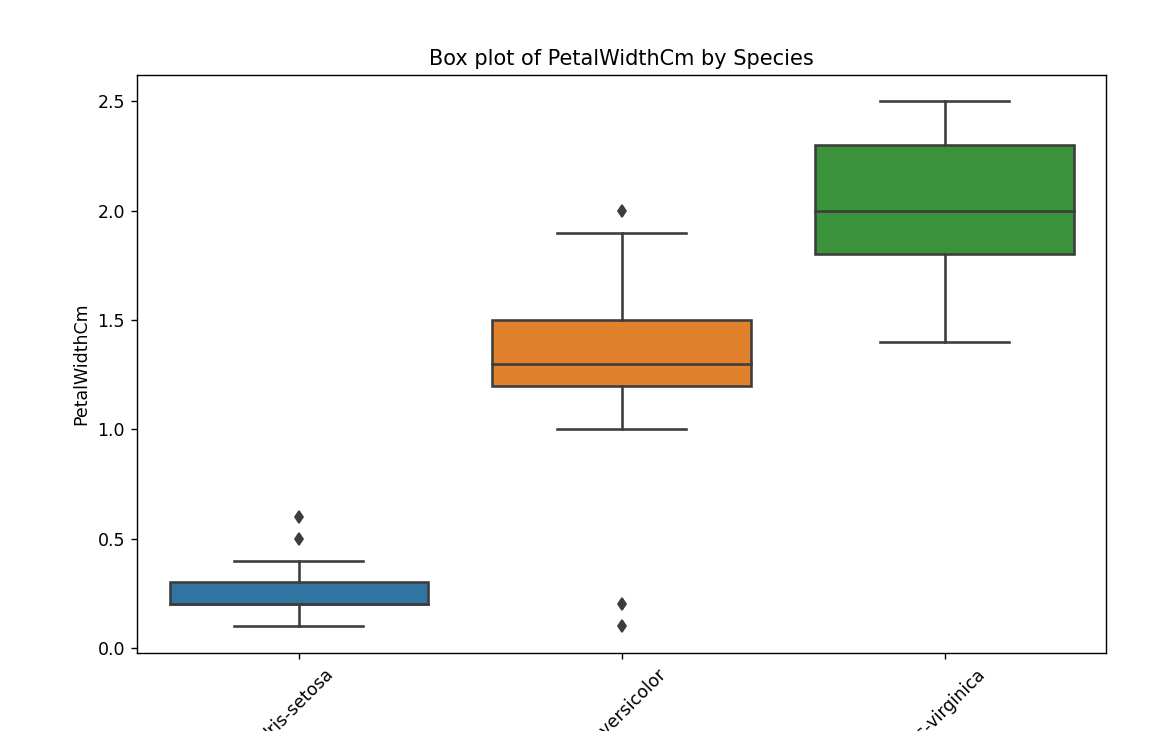
****

****

****

****

****

****

**Conclusion**

**Findings:**

* The analysis of the Iris dataset reveals that petal length and petal width are highly effective for differentiating species.
* Petal measurements show strong positive correlation and distinct clusters in pair plots, making them excellent for species classification
* Setosa is easily distinguishable with smaller petal dimensions, while Versicolor and Virginica, although overlapping, generally have larger measurements
* Overall, petal measurements provide significant discriminatory power, making them crucial for accurate species identification.

## 