MAJOR PROJECT

TO PERFORM EXPLORATORY DATA ANALYSIS (EDA) AND APPLY A REGRESSOR AND CALCULATE THE ACCURACY OF THE MODEL OF EMPLOYEE DETAILS DATA SET

-M.MOUNIKA

ABSTRACT

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations. Exploratory Data Analysis, or EDA, is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset. EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better. We will use **Python** language (**Pandas** library) for this purpose.

Regression Analysis is a Statistical method to explore relationships between one dependent (criterion) variable and two or more independent (predictor) variables.It explores the strength of the relationship and models future relationships using mean, median, and normal distributions.In other words, Regression is a method used to establish which factors are most important for the problem, which variables to ignore (the outliers), and how they impact each other.It discovers patterns in the data by analyzing the relationship between variables and makes a “best guess” to make a prediction.

Accuracy of a model is defined as a number of points classified correctly to a total number of points.

INTRODUCTION

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations. Exploratory Data Analysis, or EDA, is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset. EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better. We will use **Python** language (**Pandas** library) for this purpose.

IMPORTING LIBRARIES :

We will start by importing the libraries we will require for performing EDA. These include NumPy, Pandas, Matplotlib, and Seaborn.

READING DATA :

We will now read the data from a CSV file into a Pandas DataFrame. You can download the dataset for your reference.

EDA build a robust understanding of the data, issues associated with either the info or process. it's a scientific approach to get the story of the data.

TYPES OF EXPLORATORY DATA ANALYSIS:

1. Univariate Non-graphical
2. Multivariate Non-graphical
3. Univariate graphical
4. Multivariate graphical

- **Univariate Non-graphical:** this is the simplest form of data analysis as during this we use just one variable to research the info. The standard goal of univariate non-graphical EDA is to know the underlying sample

distribution/ data and make observations about the population. Outlier detection is additionally part of the analysis. The characteristics of population distribution include:

- **Central tendency:** The central tendency or location of distribution has got to do with typical or middle values. The commonly useful measures of central tendency are statistics called mean, median, and sometimes mode during which the foremost common is mean. For skewed distribution or when there's concern about outliers, the median may be preferred.
- **Spread:** Spread is an indicator of what proportion distant from the middle we are to seek out the find the info values. the quality deviation and variance are two useful measures of spread. The variance is that the mean of the square of the individual deviations and therefore the variance is the root of the variance
- **Skewness and kurtosis:** Two more useful univariates descriptors are the skewness and kurtosis of the distribution. Skewness is that the measure of asymmetry and kurtosis may be a more subtle measure of peakedness compared to a normal distribution

2. Multivariate Non-graphical: Multivariate non-graphical EDA technique is usually wont to show the connection between two or more variables within the sort of either cross-tabulation or statistics.

- For categorical data, an extension of tabulation called cross-tabulation is extremely useful. For 2 variables, cross-tabulation is preferred by making a two-way table with column headings that match the amount of one-variable and row headings that match the amount of the opposite two variables, then filling the counts with all subjects that share an equivalent pair of levels.

- For each categorical variable and one quantitative variable, we create statistics for quantitative variables separately for every level of the specific variable then compare the statistics across the amount of categorical variable.
- Comparing the means is an off-the-cuff version of ANOVA and comparing medians may be a robust version of one-way ANOVA.

3.Univariate graphical: Non-graphical methods are quantitative and objective, they are doing not give the complete picture of the data; therefore, graphical methods are more involve a degree of subjective analysis, also are required. Common sorts of univariate graphics are:

- **Histogram:** The foremost basic graph is a histogram, which may be a barplot during which each bar represents the frequency (count) or proportion (count/total count) of cases for a variety of values. Histograms are one of the simplest ways to quickly learn a lot about your data, including central tendency, spread, modality, shape and outliers.
- **Stem-and-leaf plots:** An easy substitute for a histogram may be stem-and-leaf plots. It shows all data values and therefore the shape of the distribution.
- **Boxplots:** Another very useful univariate graphical technique is that the boxplot. Boxplots are excellent at presenting information about central tendency and show robust measures of location and spread also as providing information about symmetry and outliers, although they will be misleading about aspects like multimodality. One among the simplest uses of boxplots is within the sort of side-by-side boxplots.
- **Quantile-normal plots:** The ultimate univariate graphical EDA technique is that the most intricate. it's called the quantile-normal or QN plot or more generally the quantile-quantile or QQ plot. it's wont to see

how well a specific sample follows a specific theoretical distribution. It allows detection of non-normality and diagnosis of skewness and kurtosis

4.Multivariate graphical: Multivariate graphical data uses graphics to display relationships between two or more sets of knowledge. The sole one used commonly may be a grouped barplot with each group representing one level of 1 of the variables and every bar within a gaggle representing the amount of the opposite variable.

Other common sorts of multivariate graphics are:

- **Scatterplot:** For 2 quantitative variables, the essential graphical EDA technique is that the scatterplot , sohas one variable on the x-axis and one on the y-axis and therefore the point for every case in your dataset.
- **Run chart:** It's a line graph of data plotted over time.
- **Heat map:** It's a graphical representation of data where values are depicted by color.
- **Multivariate chart:** It's a graphical representation of the relationships between factors and response.
- **Bubble chart:** It's a data visualization that displays multiple circles (bubbles) in two-dimensional plot.

In a nutshell: You ought to always perform appropriate EDA before further analysis of your data. Perform whatever steps are necessary to become more conversant in your data, check for obvious mistakes, learn about variable distributions, and study about relationships between variables. EDA is not an exact science- It is very important are!

TOOLS REQUIRED FOR EXPLORATORY DATA ANALYSIS:

Some of the most common tools used to create an EDA are:

1. R: An open-source programming language and free software environment for statistical computing and graphics supported by the R foundation for statistical computing. The R language is widely used among statisticians in developing statistical observations and data analysis.

2. Python: An interpreted, object-oriented programming language with dynamic semantics. Its high level, built-in data structures, combined with dynamic binding, make it very attractive for rapid application development, also as to be used as a scripting or glue language to attach existing components together. Python and EDA are often used together to spot missing values in the data set, which is vital so you'll decide the way to handle missing values for machine learning.

Apart from these functions described above, EDA can also:

- **Perform k-means clustering:** Perform k-means clustering: it's an unsupervised learning algorithm where the info points are assigned to clusters, also referred to as k-groups, k-means clustering is usually utilized in market segmentation, image compression, and pattern recognition
- EDA is often utilized in predictive models like linear regression, where it's wont to predict outcomes.
- It is also utilized in univariate, bivariate, and multivariate visualization for summary statistics, establishing relationships between each variable, and understanding how different fields within the data interact with one another.

Regression Analysis is a Statistical method to explore relationships between one dependent (criterion) variable and two or more independent (predictor) variables.It explores the strength of the relationship and models future relationships using mean, median, and normal distributions.In other words, Regression is a method used to establish which factors are most important for the problem, which variables to ignore (the outliers), and how they impact each other.It discovers patterns in the data by analyzing the relationship between variables and makes a "best guess" to make a prediction.

Regression in data Science analysis is used for prediction, forecasting, and inferring causal relationships between independent and dependent variables.

There are various types of regressions used in data science and machine learning. Each type has its importance in different scenarios and is selected for the best way it can solve the problem.

The important types of Regression used in Data Science are:

- Linear Regression
- Polynomial Regression
- Logistic Regression
- Decision Tree Regression
- Random Forest Regression
- Support Vector Regression
- Ridge Regression
- Lasso Regression

The type of Regression technique used depends upon the existing variables and the outcomes required.

Regression is also critical for any Machine Learning problem that crunches continuous numbers.

The advantages of using Regression are predictive analytics and optimization operations to understand what variables are significant and what to disregard.

It cuts down guesswork and hypotheses in decision-making by executing a scientific way of analyzing data and predicting outcomes.

Regression Analysis is widely used in Finance for use cases like forecasting revenues and expenses, in Marketing campaigns to determine target customer groups, in Logistics and Supply Chains for predicting inventory levels, in Market Research and Sales as a forecasting tool, and so on.

Accuracy of a model is defined as a number of points classified correctly to a total number of points. By Performance measure of a model what I mean is to know how well our model(classification model or Regression model) is performing with the test data or live data.

Performance of a model is always measured on test data, not training data or validation data.Performance measure is also called a Performance metric.

Amongst all the available performance measures, **accuracy** is the most easy-to-understand metric.

EXISTING SYSTEM

During this analysis of EXPLORATORY DATA ANALYSIS (EDA) , We will start by importing the libraries we will require for performing EDA. These include NumPy, Pandas, Matplotlib, and Seaborn.

In Python, there are several libraries and corresponding modules that can be used to perform regression depending on a specific problem that one encounters and its complexity. Here I assume that the reader knows Python and some of its most important libraries.

- polyfit of NumPy
- linregress of SciPy
- OLS and ols of statsmodels
- LinearRegression of scikit-learn

SYSTEM REQUIREMENTS

SOFTWARE REQUIREMENTS

- Operating system : windows XP/7
- Platform or software : Juypiter pythone NoteBook
- Coding language : Python
- Database : mySql

HARDWARE REQUIREMENTS

- Minimum 4GB RAM is required
- 64-bit versions of Microsoft Windows 10,8,7
- 2.5GB hard disk space and another 1GB for caches

ADVANTAGES AND DISADVANTAGES OF EXPLORATORY DATA ANALYSIS

Advantages of EDA:

- It gives us valuable insights into the data.
- It helps us with feature selection (i.e using PCA)
- Visualization is an effective way of detecting outliers.

Disadvantages of EDA:

- If not perform properly EDA can misguide a problem.
- EDA does not effective when we deal with high-dimensional data.

EXPLORATORY DATA ANALYSIS

Dataset Used: employees.csv

Let's read the dataset using the Pandas module and print the 1st five rows. To print the first five rows we will use the **head()** function.

CODE :

```
import pandas as pd
import numpy as np
df = pd.read_csv("C:\Users\Mounika M\OneDrive\Documents\Desktop\employees.csv")
df.head()
```

OUTPUT:

	First Name	Gender	Start Date	Last Login Time	Salary	Bonus %	Senior Management	Team
0	Douglas	Male	8/6/1993	12:42 PM	97308	6.945	True	Marketing
1	Thomas	Male	3/31/1996	6:53 AM	61933	4.170	True	NaN
2	Maria	Female	4/23/1993	11:17 AM	130590	11.858	False	Finance
3	Jerry	Male	3/4/2005	1:00 PM	138705	9.340	True	Finance
4	Larry	Male	1/24/1998	4:47 PM	101004	1.389	True	Client Services

Let’s see the shape of the data using the shape.

CODE:

```
df.shape
```

OUTPUT:

(1000, 8)

This means that this dataset has 1000 rows and 8 columns.

Let’s get a quick summary of the dataset using the describe method. The describe() function applies basic statistical computations on the dataset like extreme values, count of data points standard deviation, etc. Any missing value or NaN value is automatically skipped. describe() function gives a good picture of the distribution of data.

CODE:

```
df.describe()
```

OUTPUT:

	Salary	Bonus %
count	1000.000000	1000.000000
mean	90662.181000	10.207555
std	32923.693342	5.528481
min	35013.000000	1.015000
25%	62613.000000	5.401750
50%	90428.000000	9.838500
75%	118740.250000	14.838000
max	149908.000000	19.944000

Now, let’s also the columns and their data types. For this, we will use the info() method.

CODE:

```
df.info()
```

OUTPUT:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   First Name         933 non-null    object
 1   Gender             855 non-null    object
 2   Start Date         1000 non-null   object
 3   Last Login Time    1000 non-null   object
 4   Salary             1000 non-null   int64
 5   Bonus %            1000 non-null   float64
 6   Senior Management  933 non-null    object
 7   Team               957 non-null    object
dtypes: float64(1), int64(1), object(6)
memory usage: 62.6+ KB
```

Now let's check if there are any missing values in our dataset or not.

CODE:

```
df.isnull().sum()
```

OUTPUT:

```
First Name           67
Gender              145
Start Date            0
Last Login Time       0
Salary                0
Bonus %               0
Senior Management    67
Team                 43
dtype: int64
```

We can see that every column has a different amount of missing values. Like Gender as 145 missing values and salary has 0. Now for handling these missing values there can be several cases like dropping the rows containing NaN or replacing NaN with either mean, median, mode, or some other value.

let’s try to fill the missing values of gender with the string “No Gender”.

CODE:

```
df["Gender"].fillna("No Gender", inplace = True)

df.isnull().sum()
```

OUTPUT:

```
First Name           67
Gender                0
Start Date            0
Last Login Time       0
Salary                0
Bonus %               0
Senior Management    67
Team                 43
dtype: int64
```

We can see that now there is no null value for the gender column. Now, Let’s fill the senior management with the mode value.

CODE:

```
mode = df['Senior Management'].mode().values[0]
df['Senior Management']= df['Senior Management'].replace(np.nan, mode)

df.isnull().sum()
```

OUTPUT:

```
First Name           67
Gender                0
Start Date            0
Last Login Time       0
Salary                0
Bonus %               0
Senior Management     0
Team                 43
dtype: int64
```

Now for the first name and team, we cannot fill the missing values with arbitrary data, so, let's drop all the rows containing these missing values.

CODE:

```
df = df.dropna(axis = 0, how ='any')

print(df.isnull().sum())
df.shape
```

OUTPUT:

```
First Name           0
Gender               0
Start Date           0
Last Login Time      0
Salary               0
Bonus %              0
Senior Management    0
Team                 0
dtype: int64

(899, 8)
```

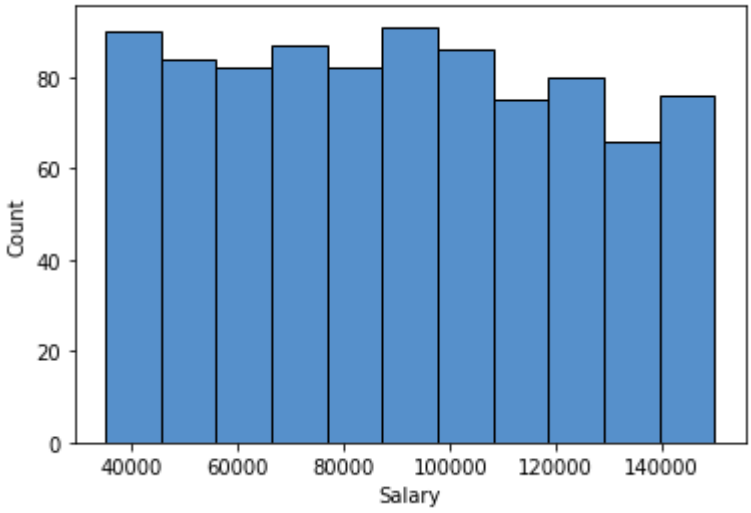
We can see that our dataset is now free of all the missing values and after dropping the data the number of also reduced from 1000 to 899.

DATA VISUALIZATION

CODE:

```
# importing packages

import seaborn as sns

import matplotlib.pyplot as plt

sns.histplot(x='Salary', data=df, )

plt.show()
```

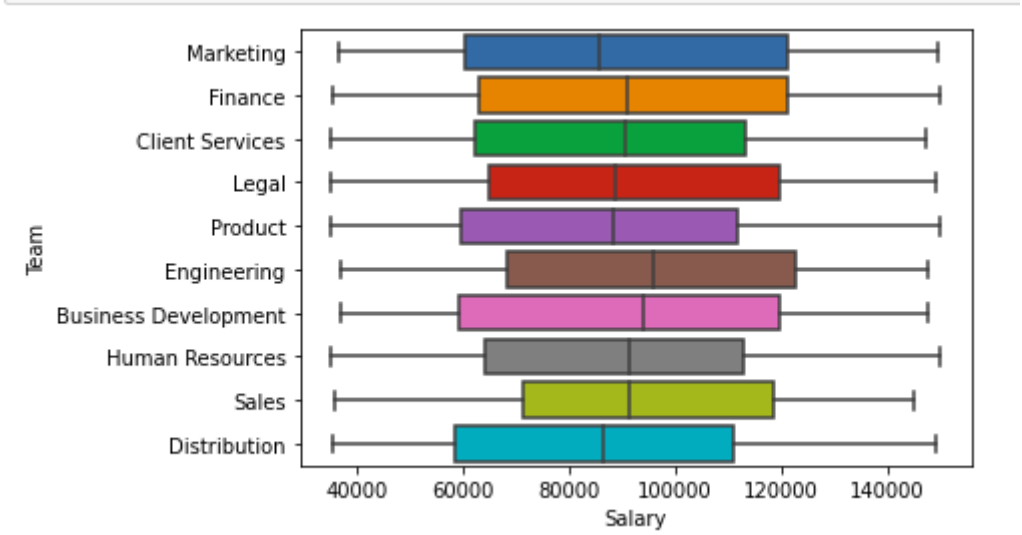
OUTPUT:



CODE :

```
# importing packages
import seaborn as sns
import matplotlib.pyplot as plt
sns.boxplot( x="Salary", y='Team', data=df, )
plt.show()
```

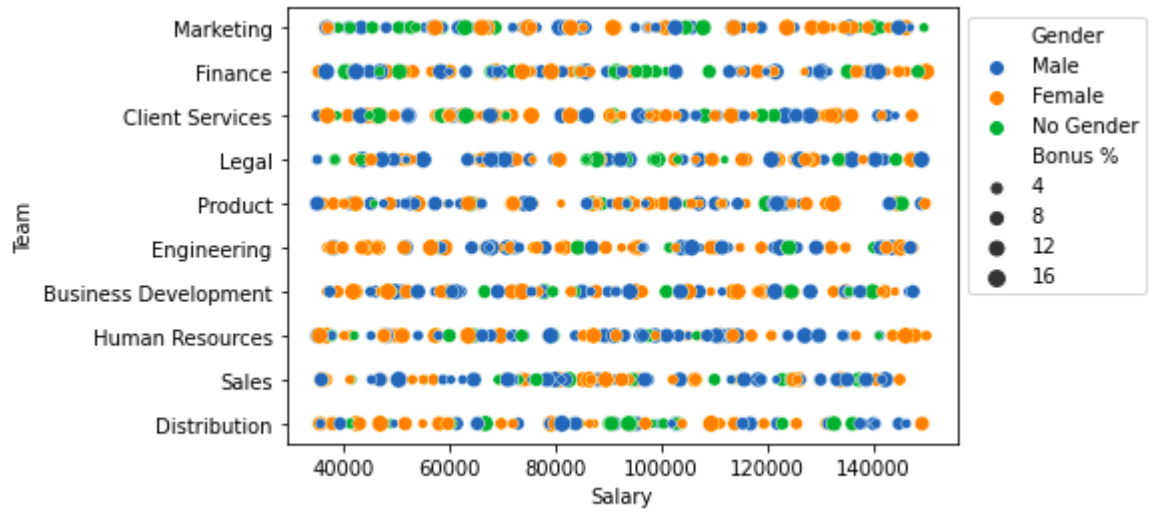
OUTPUT:



CODE :

```
# importing packages
import seaborn as sns
import matplotlib.pyplot as plt
sns.scatterplot( x="Salary", y='Team', data=df,
                                    hue='Gender', size='Bonus %')
# Placing Legend outside the Figure
plt.legend(bbox_to_anchor=(1, 1), loc=2)
plt.show()
```

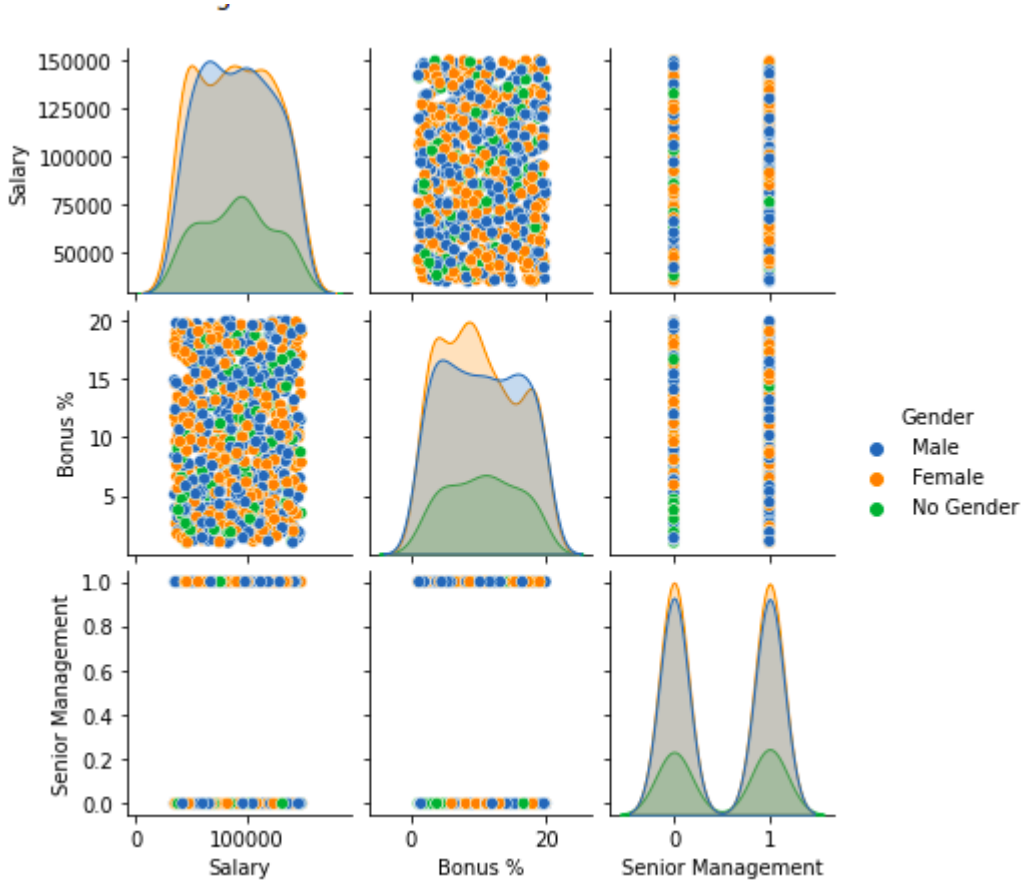
OUTPUT:



CODE :

```
# importing packages
import seaborn as sns
import matplotlib.pyplot as plt
sns.pairplot(df, hue='Gender', height=2)
```

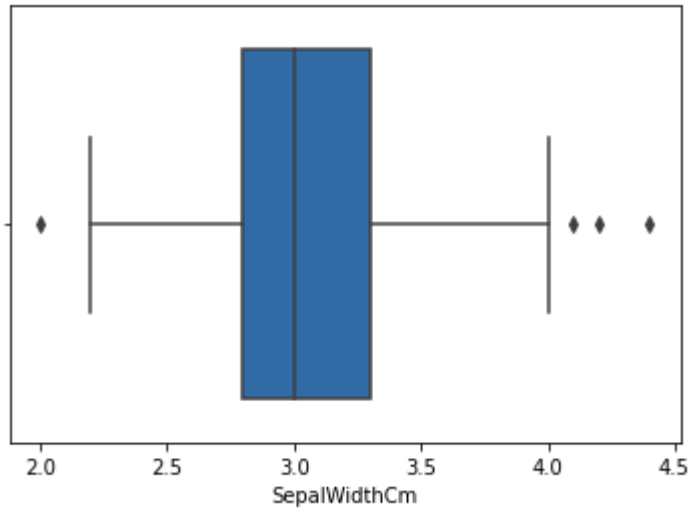
OUTPUT:



CODE :

```
# importing packages
import seaborn as sns
import matplotlib.pyplot as plt
# Load the dataset
df = pd.read_csv('Iris.csv')
sns.boxplot(x='SepalWidthCm', data=df)
```

OUTPUT:



In the above graph, the values above 4 and below 2 are acting as outliers.

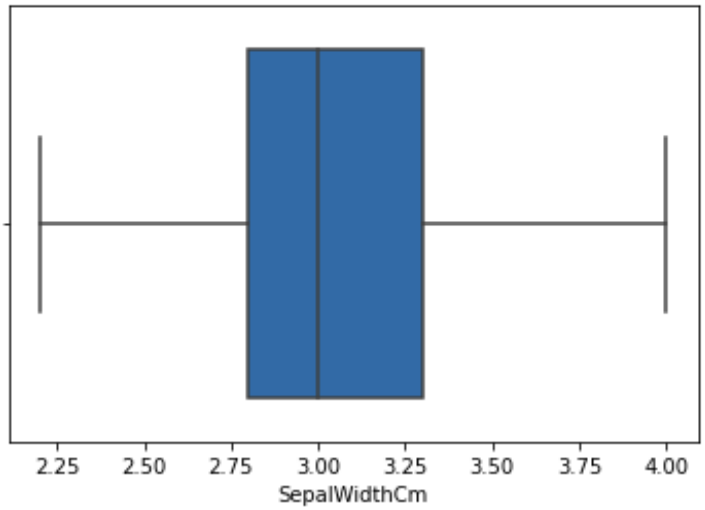
REMOVING OUTLIERS :

CODE:

```
# Importing
import sklearn
from sklearn.datasets import load_boston
import pandas as pd
import seaborn as sns
# Load the dataset
df = pd.read_csv('Iris.csv')
# IQR
Q1 = np.percentile(df['SepalWidthCm'], 25,
                                    interpolation = 'midpoint')
Q3 = np.percentile(df['SepalWidthCm'], 75,
                                    interpolation = 'midpoint')
IQR = Q3 - Q1
print("Old Shape: ", df.shape)
# Upper bound
upper = np.where(df['SepalWidthCm'] >= (Q3+1.5*IQR))
# Lower bound
lower = np.where(df['SepalWidthCm'] <= (Q1-1.5*IQR))
# Removing the Outliers
df.drop(upper[0], inplace = True)
df.drop(lower[0], inplace = True)
print("New Shape: ", df.shape)
sns.boxplot(x='SepalWidthCm', data=df)
```

OUTPUT:

```
Old Shape:  (150, 6)
New Shape:  (146, 6)

<AxesSubplot:xlabel='SepalWidthCm'>
```



CONCLUSION

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.We will use **Python** language (**Pandas** library) for this purpose.Regression is a method used to establish which factors are most important for the problem, which variables to ignore (the outliers), and how they impact each other.It discovers patterns in the data by analyzing the relationship between variables and makes a “best guess” to make a prediction.


Thank
You