**Questions**

**(Real Data Set)**

**Q. Write a Python code for loading a data set through Pandas library. Your code should include functions for both CSV and EXCEL files.**

**Sample Answer:**

First of all, we have to import the Pandas library and set an alliance to access it. Through this alliance (here ‘pd’ is the alliance) we can read the file by calling ‘read\_csv’ or ‘read\_excel’ and pass the name of the file as an argument along with the file format.

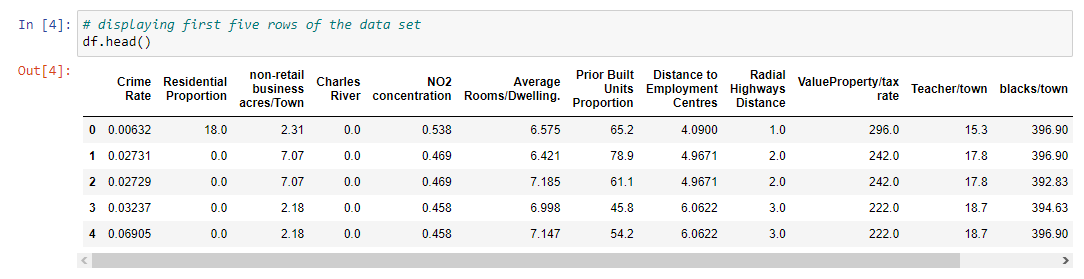
****

****

**Q. Write a Python code for displaying first five rows of a data set.**

**Sample Answer:**

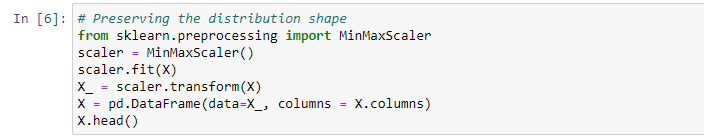
Using the Pandas library, we can display/print first five rows of the data set through the built-in function head().

****

**Q. How can we normalize dispersed data in Python? Use any Python library function in your code.**

**Sample Answer:**

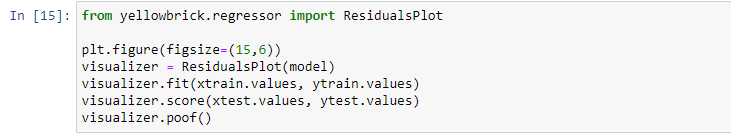
Normalization of data is a preprocessing method. A function, MinMaxScaler, from the sci-kit learn library is imported and used that normalizes the data set features by subtracting the minimum value and dividing it by the range of the data. The range is the difference between the original maximum and minimum values of the data set.

****

**Q. Why is a residual plot used in regression? How can a residual plot be plotted using Python?**

**Sample Answer:**

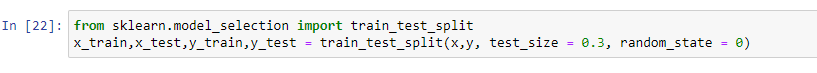
A residual plot is a plot that is used to display fitted values against the residual values in a regression model. It is also used to check or assess whether or not a regression model is appropriate for a data set.

****

**Q. What is model training in Python? Write a code for training a model using train/test methods.**

**Sample Answer:**

Model training is a method to measure the accuracy of your model. It is also called Train/Test because you split the data set into two sets: a training set and a testing set. We train the model using the training set and test the model using the testing set. We use the ‘train\_test\_split’ function from the *sci-kit learn* library for this purpose in Python.



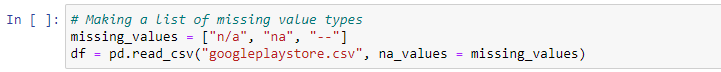
**Q. Suppose your boss has given you a task of identifying missing values in a data set. Using a Python code search and count the missing or null values in the data set.**

**Sample Answer:**

We can use .isnull() function from the Pandas library to recognize missing values.



If there are multiple null or missing values that differ in format from each other, we can first save all the occurring formats in a list, then past the list in to .isnull() function as an argument.

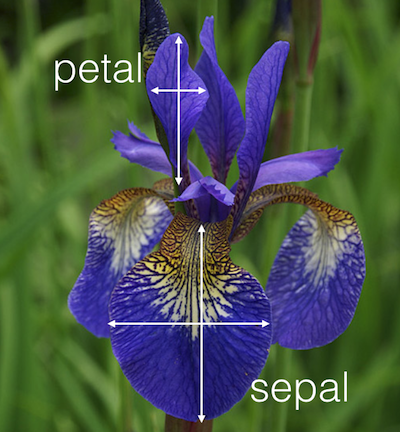


**Q. You are provided with a data set and asked to carry out the following procedure:**

* **Import libraries that are needed for linear regression**
* **Load your data set into the Python compiler**
* **Preprocess your data by removing/handling missing values**
* **Create a regression model and fit the data into it**
* **Train and Test your model**
* **Get results from your model**

**Use proper Python code using most suited libraries for the procedure.**

**Q. According to the given picture:**

****

* **Use iris flower dataset from sklearn library and try to form clusters of flowers using petal width and length features. Drop other two features for simplicity.**
* **Figure out if any preprocessing such as scaling would help here**
* **Draw elbow plot and from that figure out optimal value of k**

# **Pandas Exercises for Data Analysis**

## **1. How to create a series from a list, numpy array and dictionary?**

## **2. How to convert the index of a series into a column of a data frame?**

## **3. How to combine many series to form a data frame?**

## **4. How to get the items not common to both series A and series B?**

## **5. How to compute the mean squared error on a truth and predicted series?**

## **6. How to get the day of month, week number, day of year and day of week from a series of date strings?**

## **7. How to import only every nth row from a csv file to create a data frame?**

## **8. How to change column values when importing csv to a data frame?**

## **9. How to check if a data frame has any missing values?**

## **10. How to count the number of missing values in each column?**

## **11. How to replace missing values of multiple numeric columns with the mean?**

## **12. How to change the order of columns of a data frame?**

## **13. How to format all the values in a data frame as percentages?**

## **14. How to swap two rows of a data frame?**

## **15. How to create a column containing the minimum by maximum of each row?**

## **16. How to normalize all columns in a data frame?**

## **17. How to join two data frames by 2 columns so they have only the common rows?**

## **18. How to get the positions where values of two columns match?**