# Experiment no. 01

**Aim:** Data preparation using NumPy and Pandas

1. Derive an index field and add it to the data set.
2. Obtain a listing of all records that are outliers according to the any field.

## Derive an index field and add it to the data set.

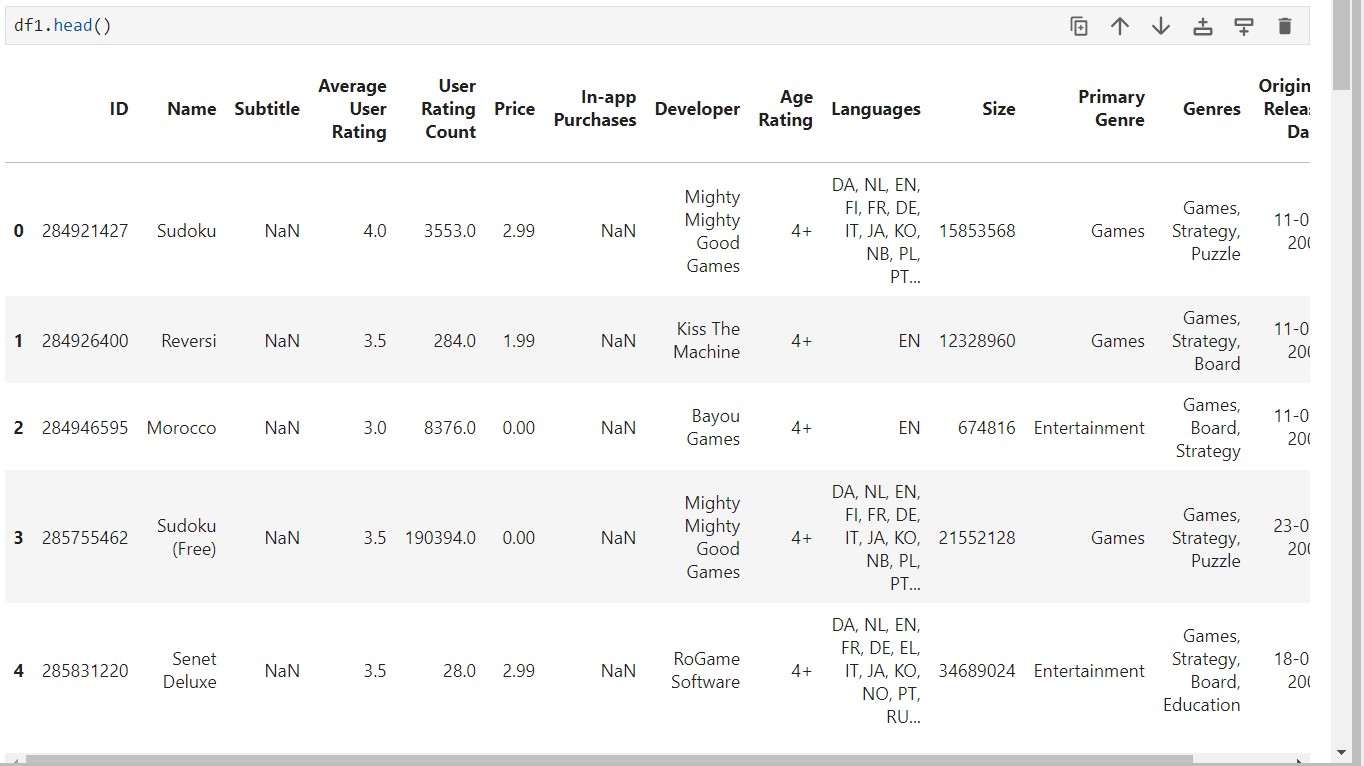
* + 1. Import the Pandas and NumPy:



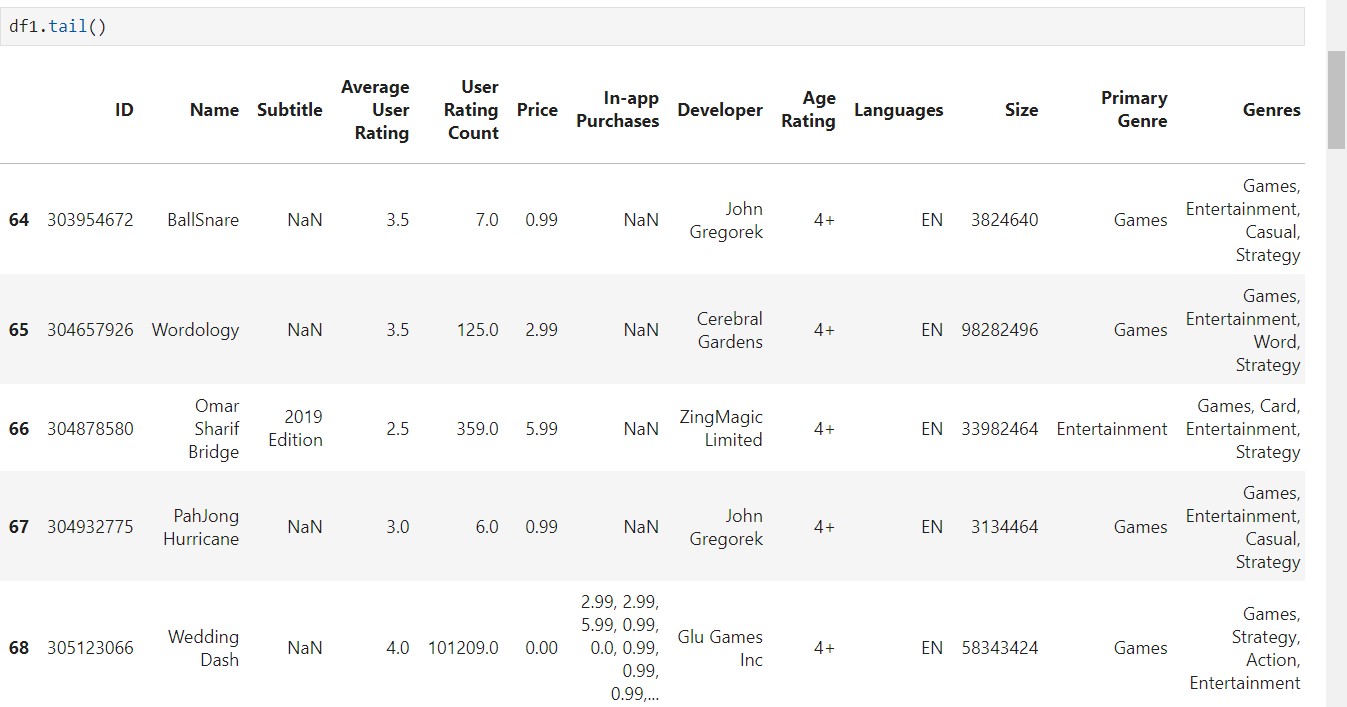
* + 1. **Read:**to read data stored as a csv file into a Pandas DataFrame.



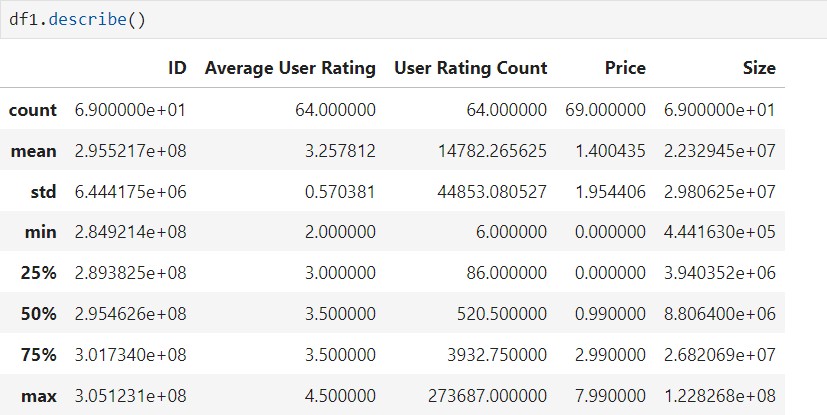
* + 1. **Head:** returns a specified number of rows, string from the top. This method returns the first 5 rows if a number is not specified.



* + 1. **Tail**: returns a specified number of last rows. This method returns the last 5 rows if a number is not specified



* + 1. **Describe:** It returns description of the data in the DataFrame



* + 1. **Isnull:**Detect missing values for an array-like object.



* + 1. **Isnull().sum():**Calling the sum() method on the isnull() series returns the count of True values which actually corresponds to the number of NaN values.

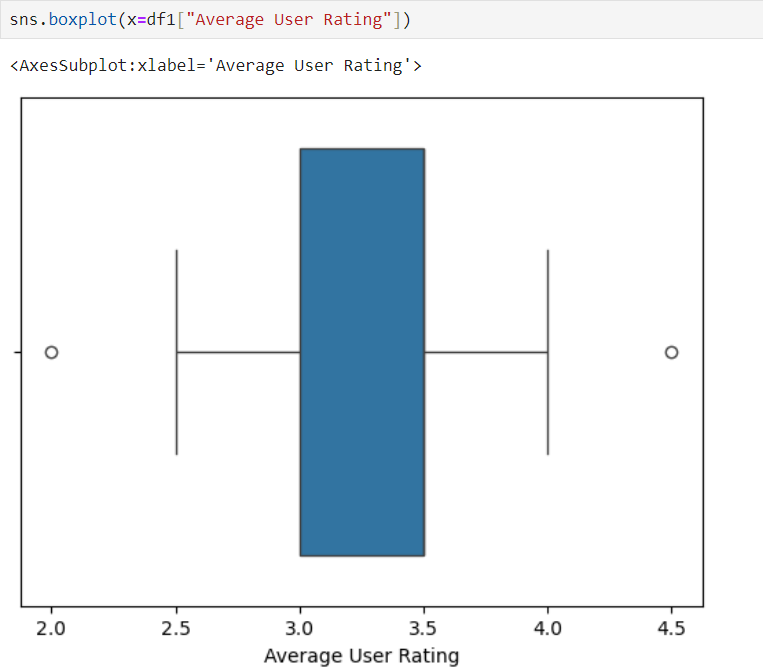


## Obtain a listing of all records that are outliers according to the any field.

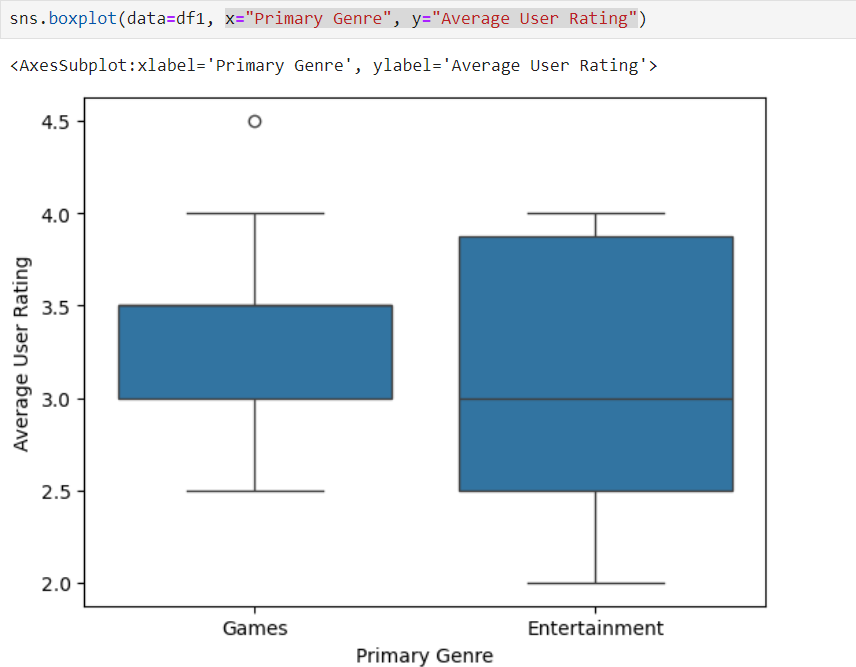
* + 1. Import seaborn:

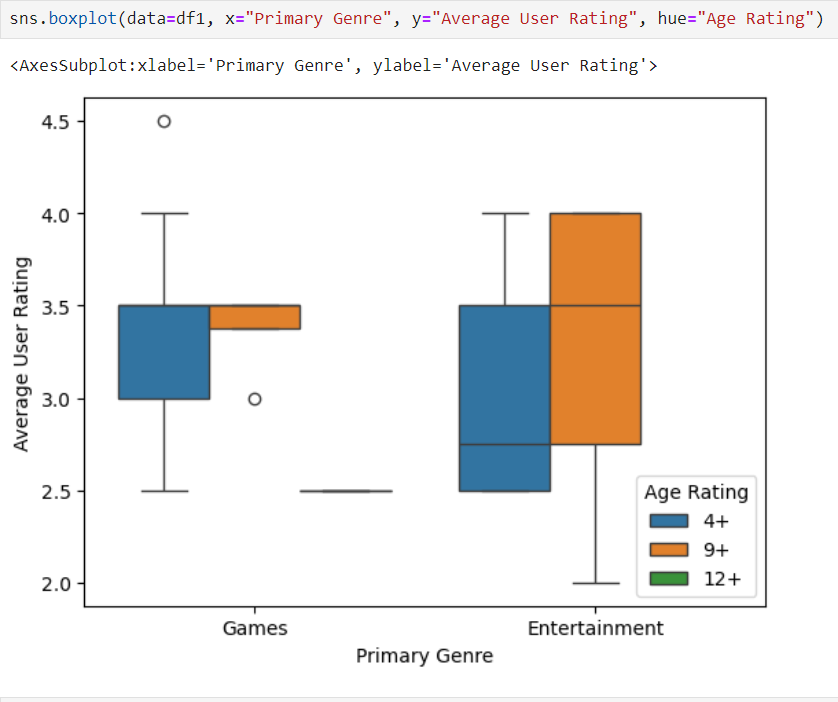


* + 1. Draw a single Horizontal boxplot:



* + 1. Group by Categorical variables:



* + 1. Draw a vertical boxplot with nested grouping by 2 variables:

## Conclusion:

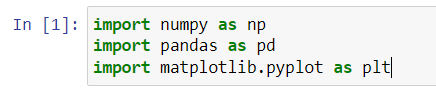
Thus, in this experiment, we have successfully implemented data preparation using NumPy and Pandas.

# Experiment no. 02

**Aim:** Data Visualization/Exploratory data Analysis for the selected data set using Matplotlib and Seaborn

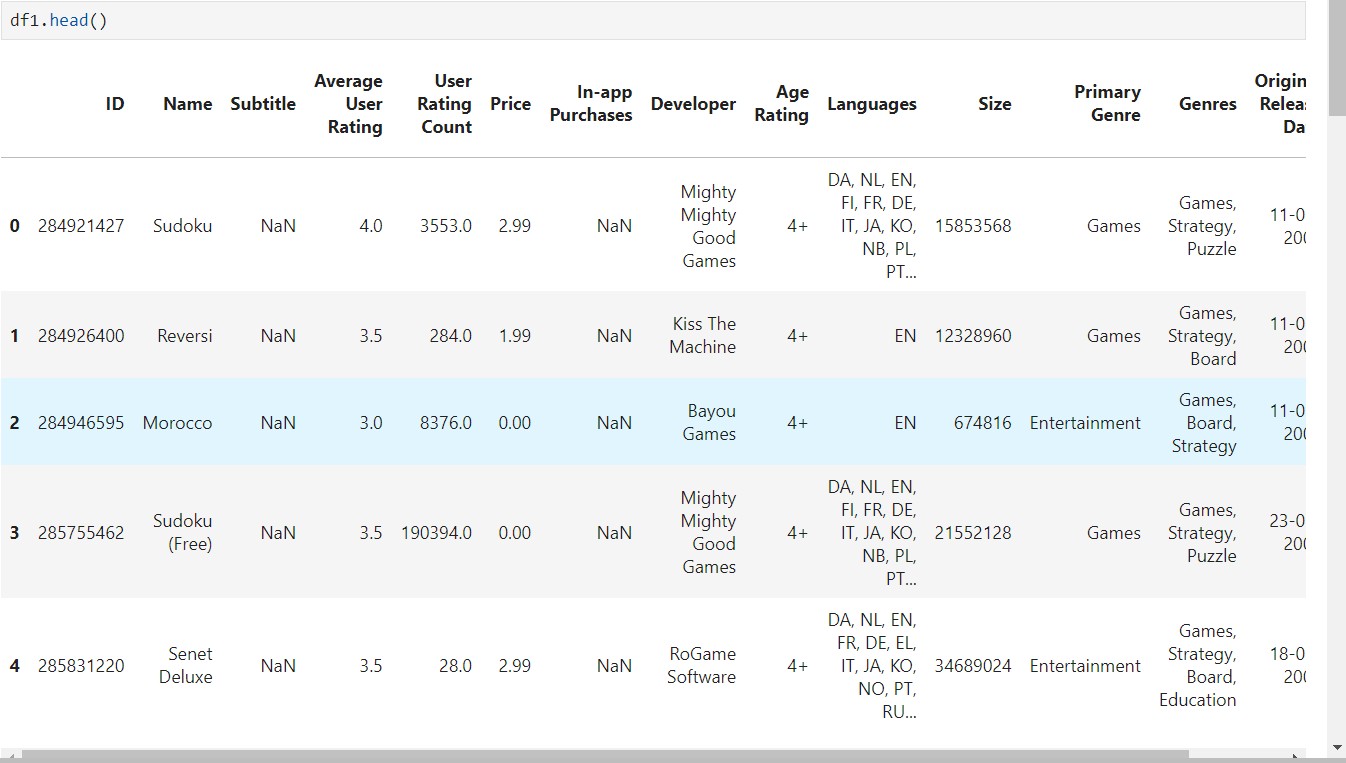
1. Create a bar graph, contingency table using any 2 variables.
2. B. create a normalized histogram

## Import the NumPy, Pandas and Matplotlib libraries:

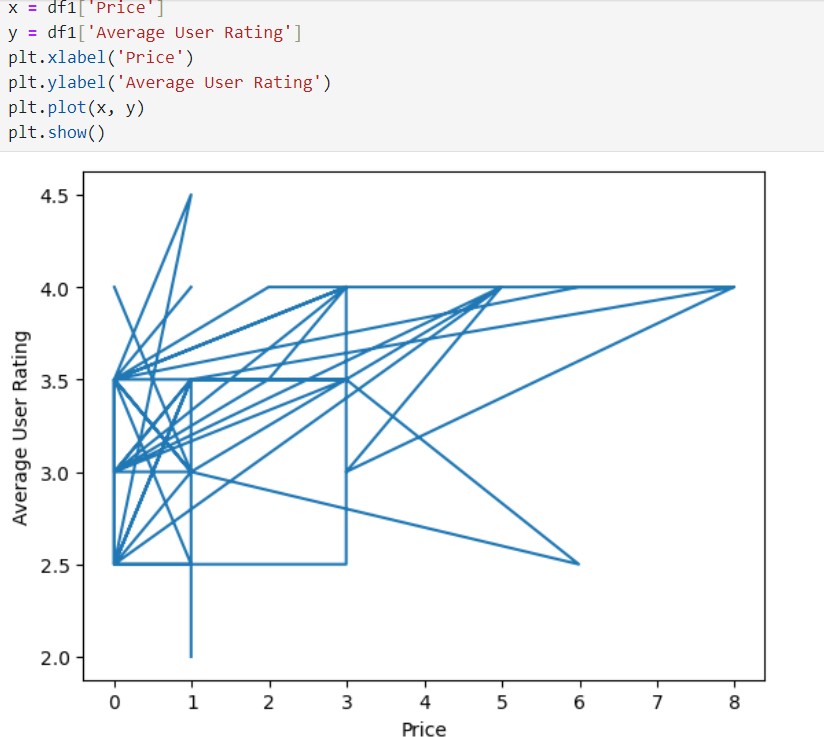


**Open the dataset and read the data:**

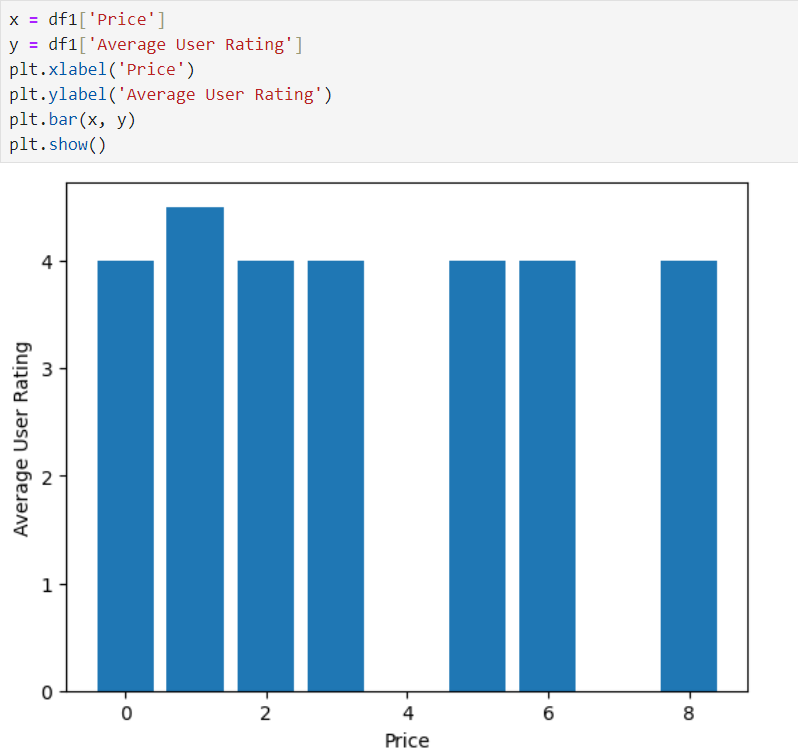




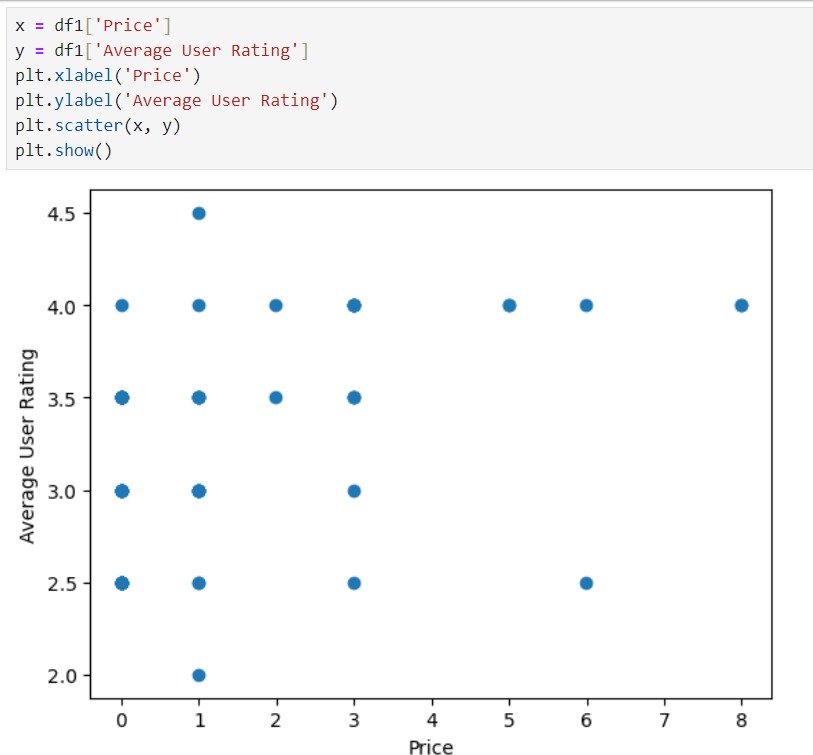
## Line plots:



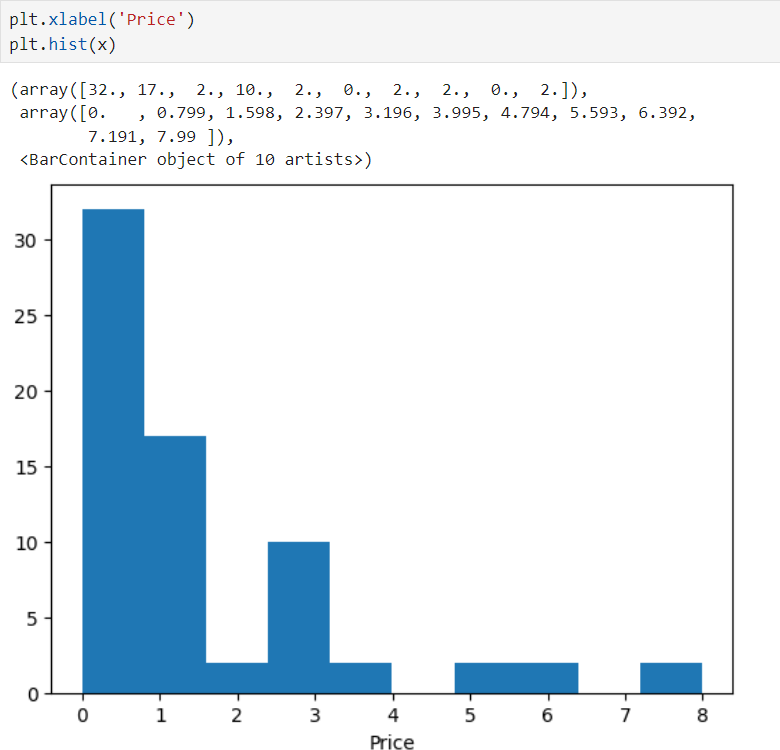
**Bar Plot:**



## Scatter plot:



**Histogram plot:**



## Conclusion:

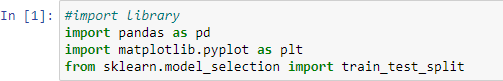
Thus, in this experiment we have successfully implemented Data Visualization/Exploratory data Analysis for the selected data set using Matplotlib and Seaborn.

# Experiment no.03

**Aim: Data Modeling:**

* 1. Identify the total number of records in the training data set.
  2. Validate your partition by performing a two-sample z-test.

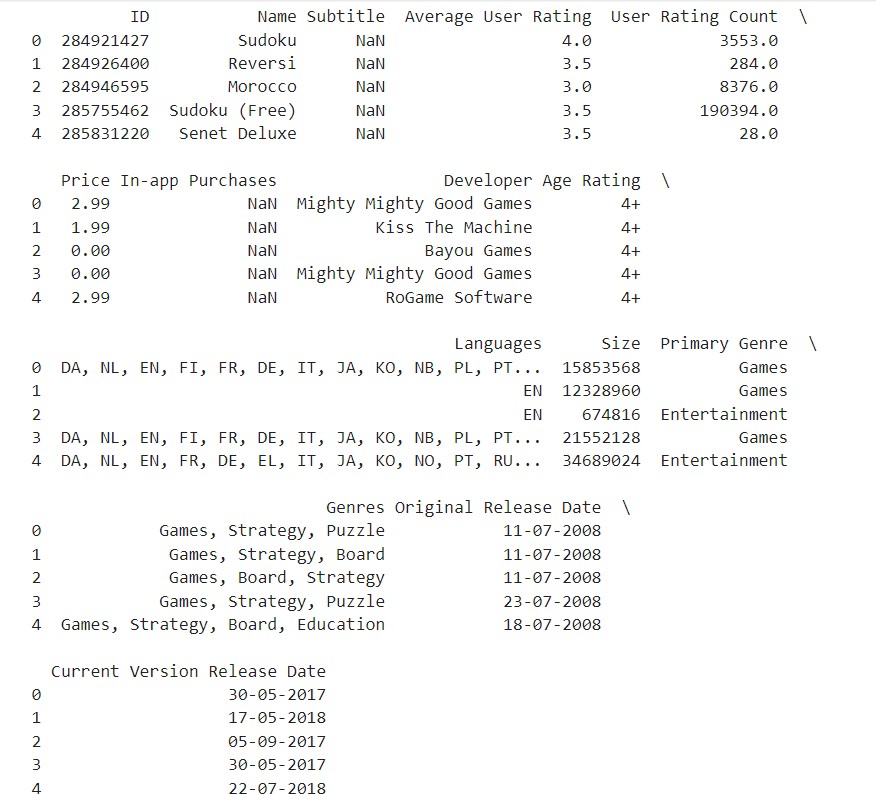
## Import Libraries:



**Load the titanic dataset:**



## Display the fist few rows of the dataset:



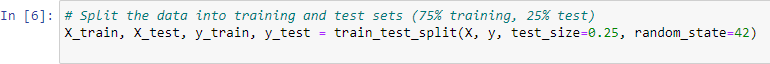
**Define features x and target variables y:**

* X: contains the features by dropping the survived column
* Y: contains the target variable which is survived



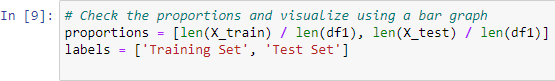
## Split data into Training and Test sets:

* test\_size: It specifies proportion of the dataset to include in the test split.
* random\_state: It sets a seed for reproducibility



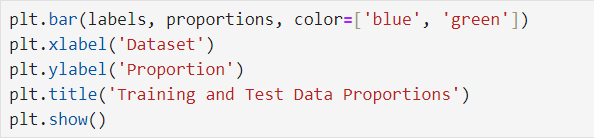
## Check proportions and visualize using bar graph:

* Proportions: It id a list containing the calculated proportions
* Labels: It is containing labels for the bar graph.



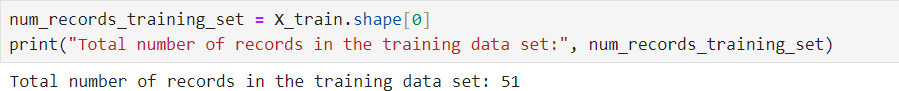
## Plot the bar Graph:

* + Plt.bar: Creates the bar graph.
  + Plt.xlable, plt.ylabel. plt.title: Adds labels and title to the graph.
  + Plt.show(): Display the graph



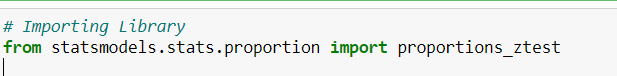


## Display the total no. of records in the training set:



**B. Validate your partition by performing a two-sample z-test.**

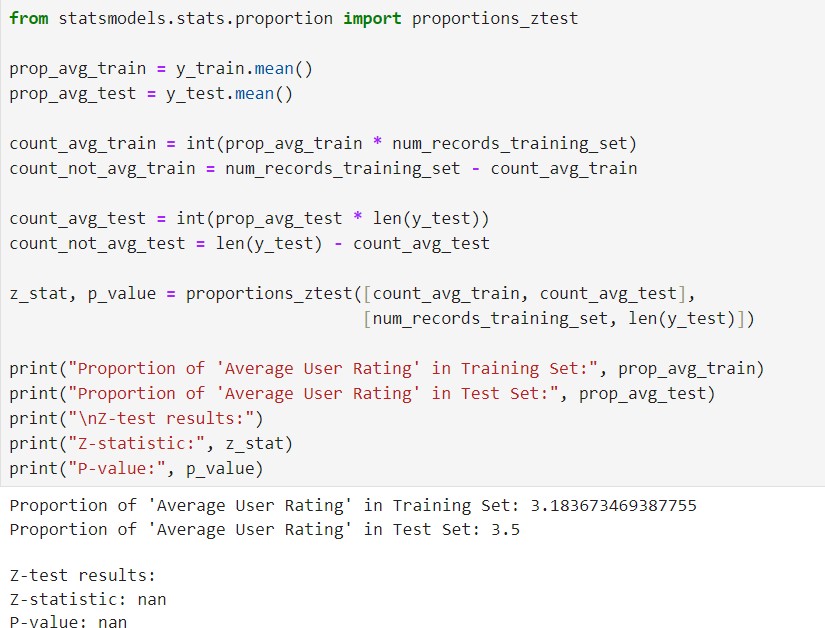
## Importing Library:



**Calculate the proportion of 'Survived' in training and test sets:**



**Perform and display two sample z-test:**



**Conclusion:**

Hence, we have successfully studied to implement Data Modelling.

# Experiment no.: 04

## Aim: Implementation of Statistical Hypothesis Test using Scipy and Sci-kit learn:

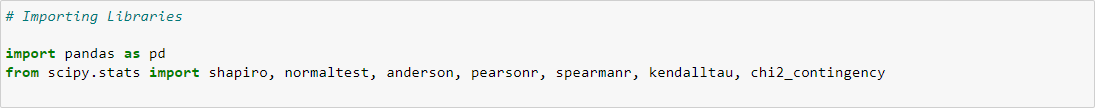
1. **Pearson’s Correlation Coefficient**

## Spearman’s Rank Correlation

1. **Kendall’s Rank Correlation**

## Chi-Squared Test

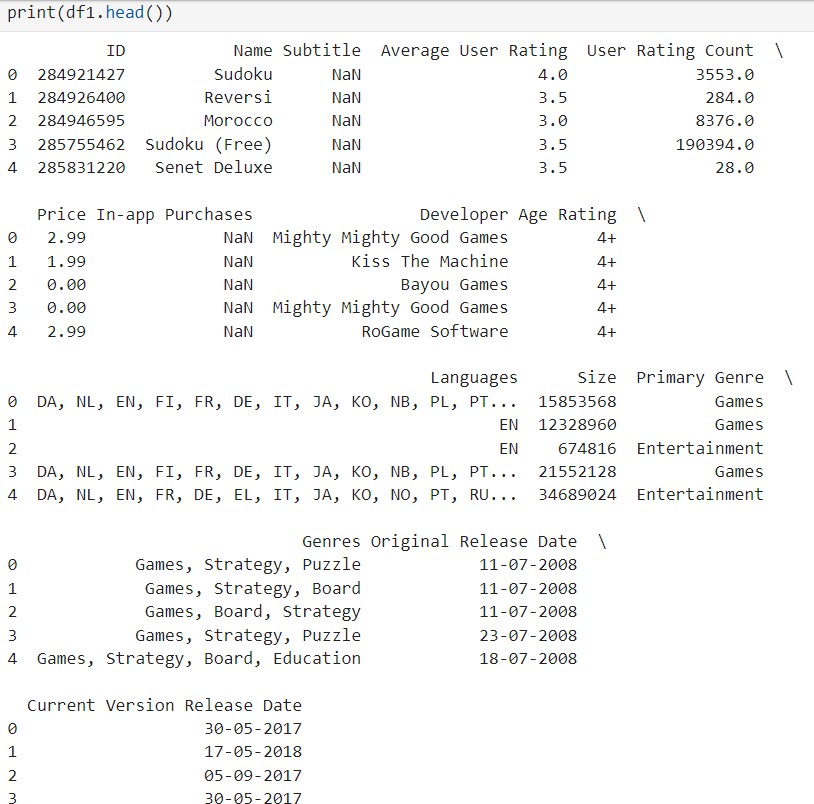
**Import Libraries:**



## Load the titanic dataset:

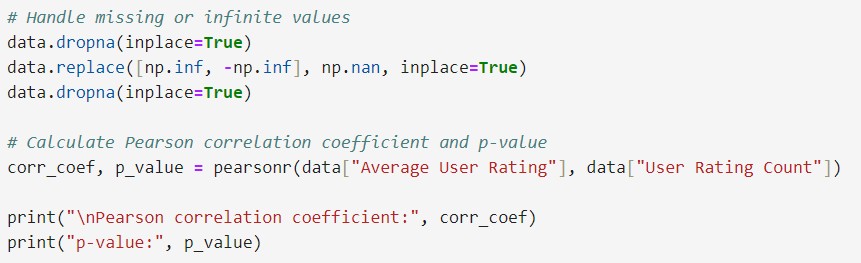


**Display the first few rows of the dataset:**

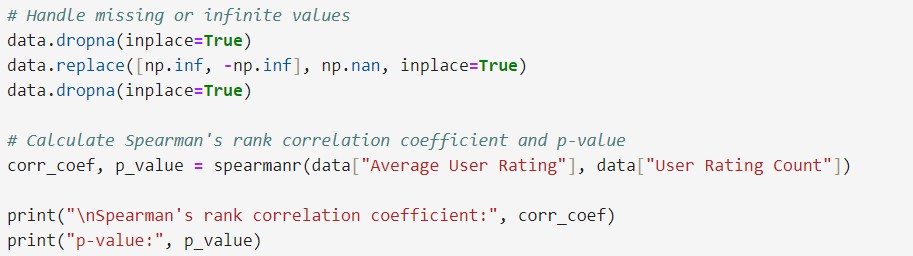


## Display Correlation Tests:

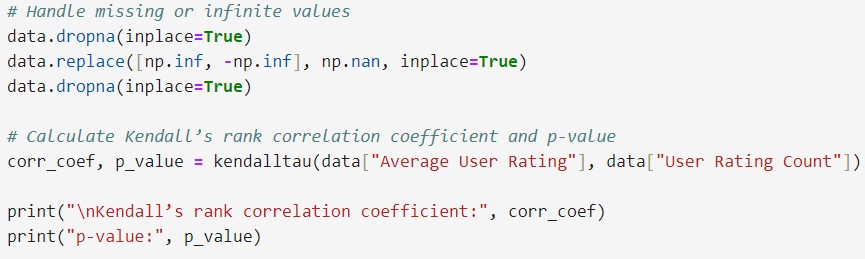
1. **Pearson’s Correlation Coefficient:**



## Spearman’s Rank Correlation:



1. **Kendall’s Rank Correlation:**



## Chi-Squared Test:



**Conclusion:**

Hence, we have successfully studied to implement Statistical Hypothesis Test using Scipy and Sci-kit learn.

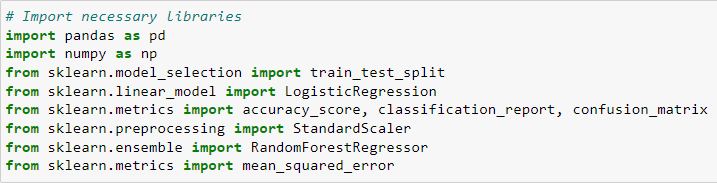
# Experiment no.: 05

## Aim: Regression Analysis:

1. **Perform Logistic Regression to find out relation between variables.**

## Apply regression Model techniques to predict the data on above dataset

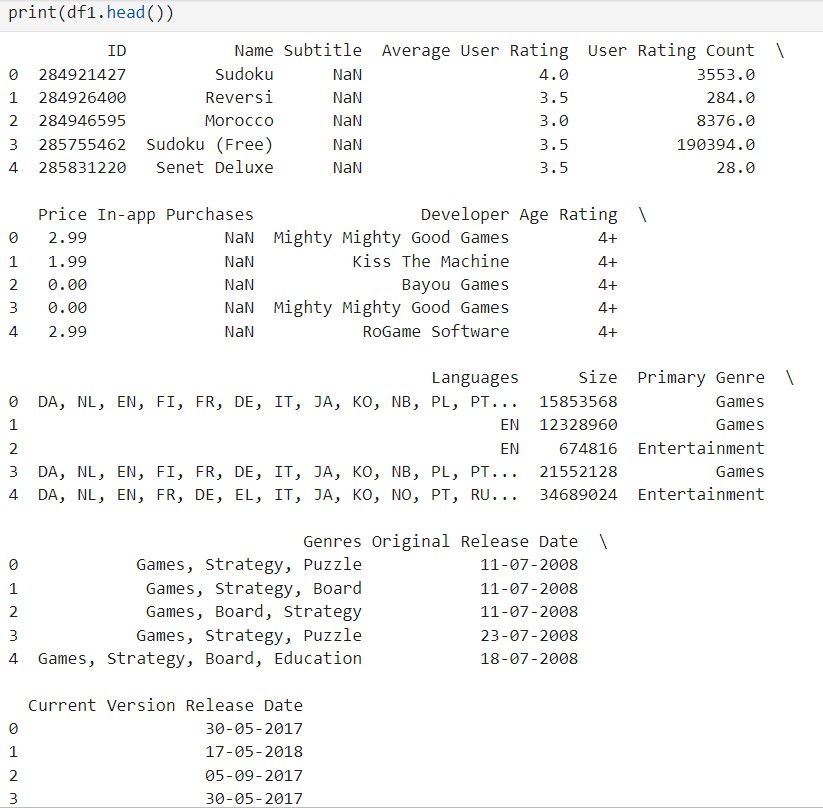
**Import Libraries:**



## Load the titanic dataset:



**Display the first few rows of the dataset:**



## Drop irrelevant columns:



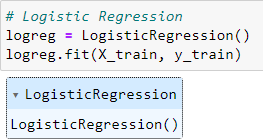
**Handle missing values:**



## Split the data into training and testing sets:



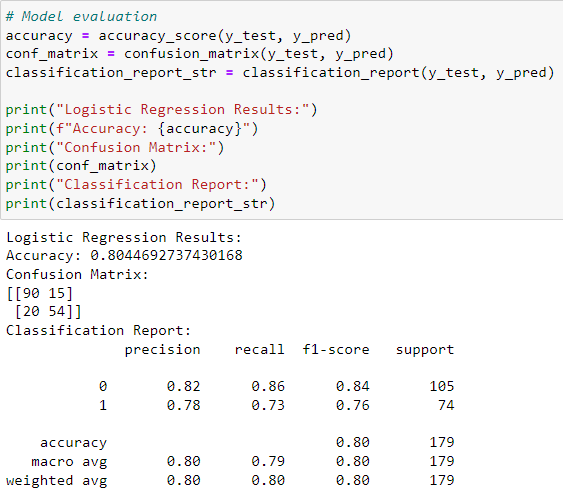
**Logistic Regression:**



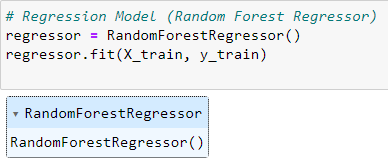
## Predictions on the test set:



**Model evaluation:**



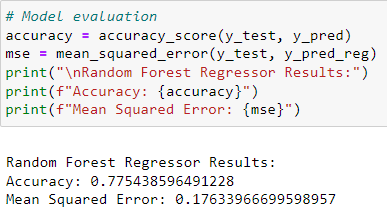
## Regression Model (Random Forest):



**Predictions on the test set:**



## Model evaluation:



**Conclusion:**

Hence, we have successfully studied to implement Statistical Hypothesis Test

using Scipy and Sci-kit learn.