Exploratory data analysis

- overview

- EDA goals and benefits

Univariate data analysis

- Characterizing data with descriptive statistics

- Univariate distribution plots

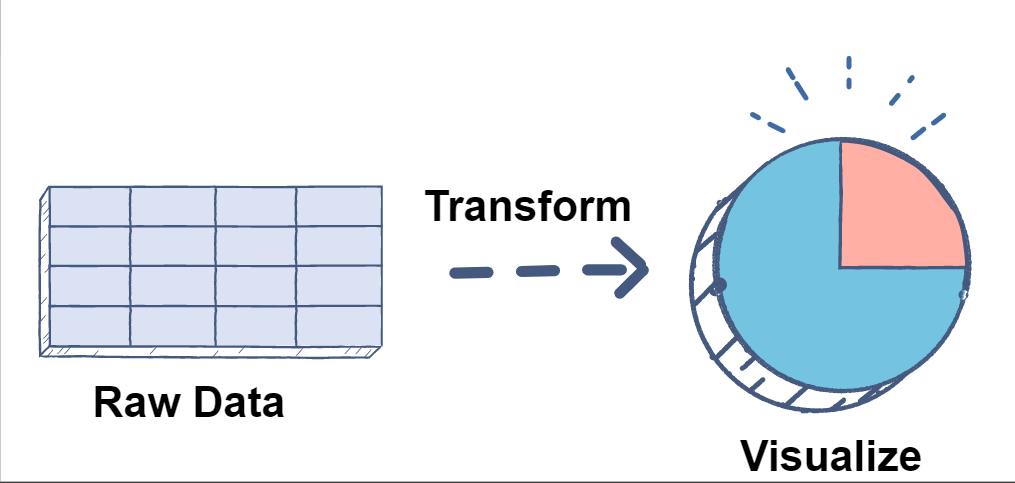
- Univariate comparison plots

- Univariate composition plots

Exploratory data analysis

- overview

Exploratory Data Analysis (EDA) is a way to investigate datasets and find preliminary information, insights, or uncover underlying patterns in the data. Instead of making assumptions, data can be processed in a systematic method to gain insights and make informed decisions.



Why Exploratory Data Analysis?

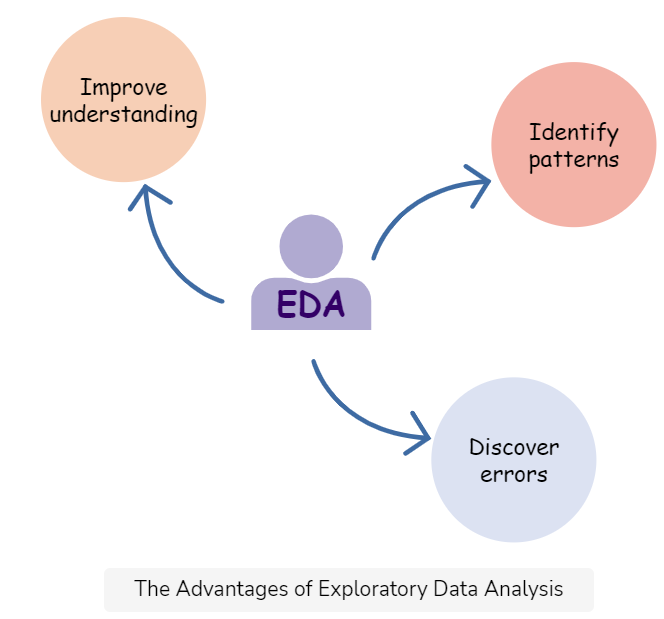
Some advantages of Exploratory Data Analysis include:

Improve understanding of variables by extracting averages, mean, minimum, and maximum values, etc.

Discover errors, outliers, and missing values in the data.

Identify patterns by visualizing data in graphs such as box plots, scatter plots, and histograms.

Hence, the main goal is to understand the data better and use tools effectively to gain valuable insights or draw conclusions.



Statistics

The first step in data analysis is to observe the statistical values of the data to decide if it needs to be preprocessed in order to make it more consistent

Describe

The describe() method of a pandas data frame gives us important statistics of the data like min, max, mean, standard deviation, and quartiles.

For example, we want to verify the minimum and maximum values in our data. This can be done by invoking the describe() method:

Data cleaning

Removing nulls

In order to identify the number of nulls within each column, we can invoke the isnull() method on each column of the pandas data frame.

If null values are found within a column, they can be replaced with the column mean using the fillna() method:

Data visualizations

As human beings, it is difficult to visualize statistical values. As an alternative, visualizations can be utilized in order to better understand the data and detect patterns.

- EDA goals and benefits

Goals

The goal of EDA is to allow data scientists to get deep insight into a data set and at the same time provide specific outcomes that a data scientist would want to extract from the data set. It includes:

* List of outliers
* Estimates for parameters
* Uncertainties about those estimates
* List of all important factors
* Conclusions or assumptions as to whether certain individual factors are statistically essential
* Optimal settings
* A good predictive model

The following are some benefits of an EDA:

Detecting missing or inaccurate data

A data clean-up in the early stages of Exploratory Data Analysis may help you discover any faults in the dataset during the analysis. It will alert you if you need to modify the data or collect new data entirely before continuing with the deep analysis.

Testing your hypothesis

It is critical to ensure that any assumptions or hypotheses you are working on can withstand inspection. It will assist you in determining if you are inferring the correct results based on your knowledge of the facts. If not, you know your assumptions are incorrect or you’re asking the wrong questions about the dataset.

Developing the most effective model

During the analysis, any unnecessary information must be removed. This is due to the fact that extraneous data might either distort your results or just hide crucial insights with unneeded noise.

Error detection

Exploratory Data Analysis assists in determining whether data may result in inevitable mistakes in your subsequent analysis. Knowing which facts will have an influence on your results can assist you to avoid accepting erroneous conclusions or mistakenly identifying an outcome.

Assisting in choosing the right tool

Exploratory Data Analysis will assist you in determining which approaches and statistical models will assist you in extracting the information you want from your dataset. Intuition and reflection are essential abilities for doing exploratory data analysis. While EDA may entail the execution of predefined tasks, it is the interpretation of the outcomes of these activities that is the true talent.

Exploratory Data Analysis greatly helps data scientists guarantee that the results they create are legitimate and appropriate to any targeted business outcomes and goals. When EDA is finished and insights are obtained, its characteristics can be used for more complex data analysis or modeling, including machine learning. Aspiring data analysts might consider taking a complete curriculum in data analytics to gain critical skills relating to tools, methodologies, strategies, and frequently used computer languages for exploratory data analysis.

Univariate data analysis

Data visualization is the process of representing data using visual elements like charts, graphs, etc. that helps in deriving meaningful insights from the data. It is aimed at revealing the information behind the data and further aids the viewer in seeing the structure in the data.

Data visualization will make the scientific findings accessible to anyone with minimal exposure in data science and helps one to communicate the information easily. It is to be understood that the visualization technique one employs for a particular data set depends on the individual’s taste and preference. Let’s first understand univariate analysis in python and how to perform univariate analysis in python.

Need for visualizing data :

Understand the trends and patterns of data

Analyze the frequency and other such characteristics of data

Know the distribution of the variables in the data.

Visualize the relationship that may exist between different variables

The number of variables of interest featured by the data classifies it as univariate, bivariate, or multivariate. For eg., If the data features only one variable of interest then it is a uni-variate data. Further, based on the characteristics of data, it can be classified as categorical/discrete and continuous data.

In this article, the main focus is on univariate data visualization(data is visualized in one-dimension). For the purpose of illustration, the ‘iris’ data set is considered. The iris data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. The different variables involved in the data set are Sepal Length, Sepal Width, Petal Length, Petal width which is continuous and Variety which is a categorical variable. Though the data set is multivariate in nature, for univariate analysis, we consider one variable of interest at a time.

Univariate Analysis

Analysis of a single variable in your data

A variable can be a column in your dataset

We can use various graphs like

Histograms, Barcharts, Density plots, Distribution plots, and so on

We can use pandas plotting functionality, seaborn library for plotting data.

These plots help in understanding the location/position of observations in the data variable, its distribution, and dispersion. Uni-variate plots are of two types: 1)Enumerative plots and 2)Summary plots These plots enumerate/show every observation in data and provide information about the distribution of the observations on a single data variable.

- Characterizing data with descriptive statistics

Types of variables

* Categorical variable:
* These have values in the form of string or text
* These even include values like dates, gender and so on
* Numerical variable
* These have numerical values like age, stock value, and so on
* Basically any value that can be measured

**Categorical values** can be further divided into two types:

Nominal variables:

Simply put, these are values which have no order amongst them, that is no value is greater than the other or can be arranged in any order with respect to one another.

For example weather, country, and so on.

Ordinal variable:

These categorical variables can be arranged in some order with respect to one another

For example, rank in a competition, educational level and so on.

**Numerical variable** can be further divided into two types:

Continuous variable:

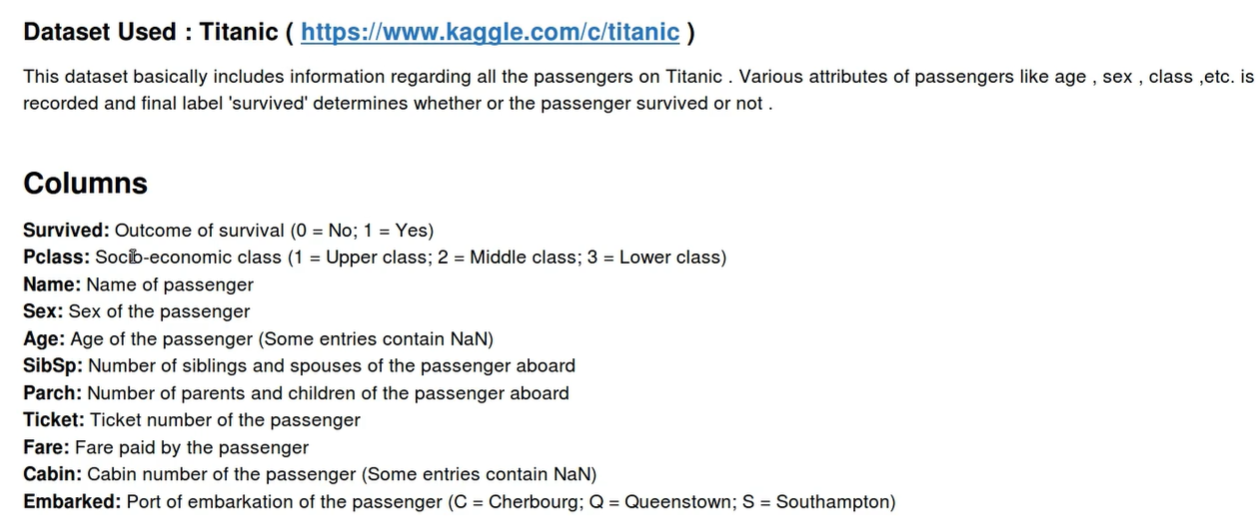
These have infinite values

For example, height, weight, age, and so on

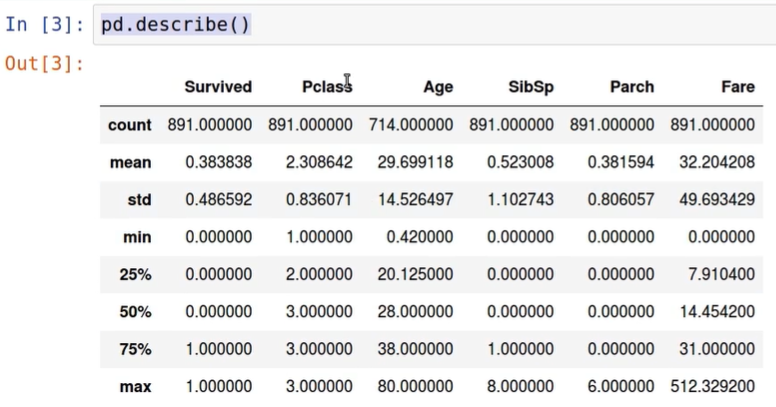
Discrete variable:

These have finite values

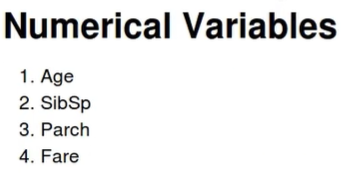
For example the number of gold medalswon by an athlete etc

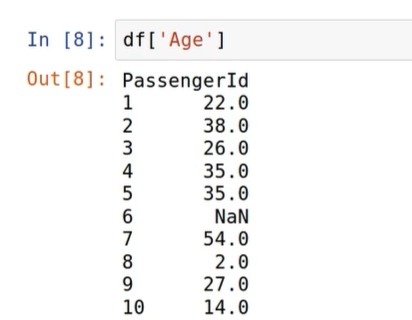


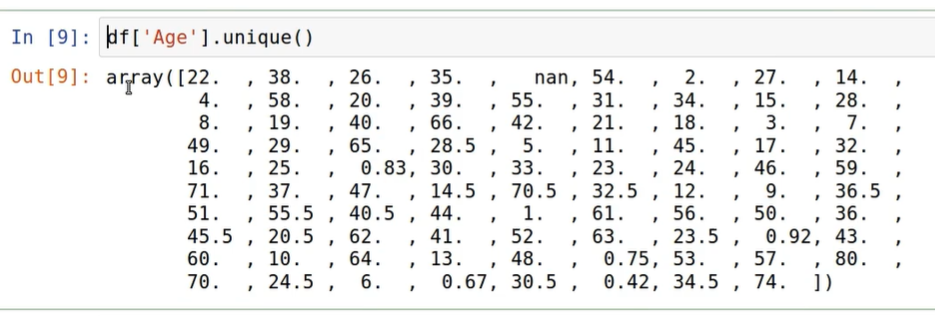




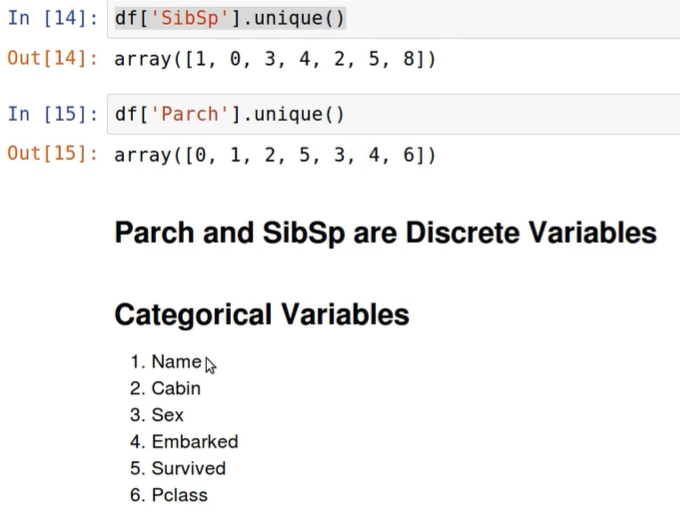




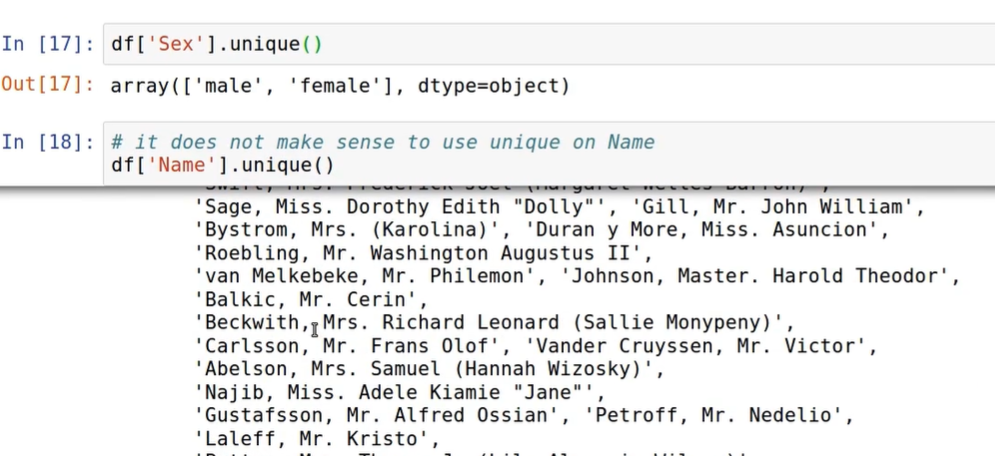




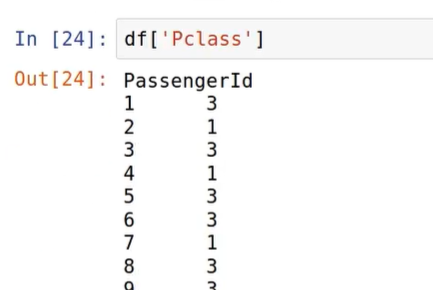
Age and Fare are Continuous variables

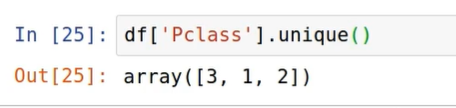


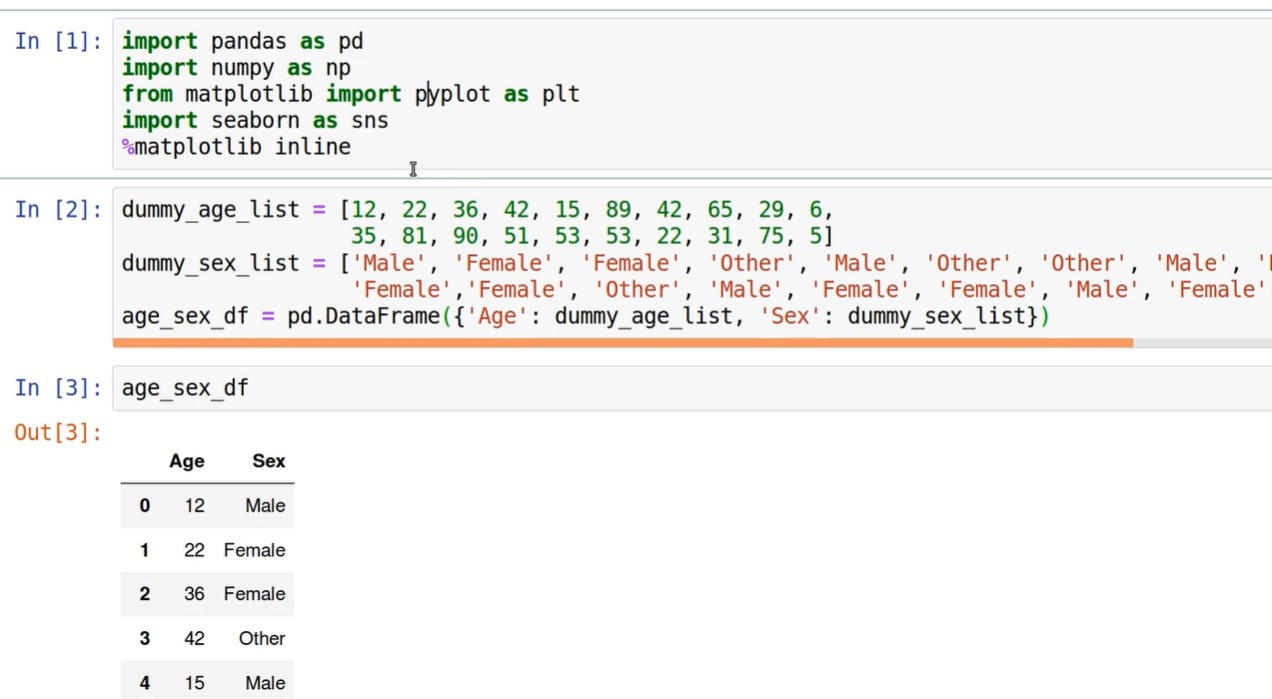
Nominal values



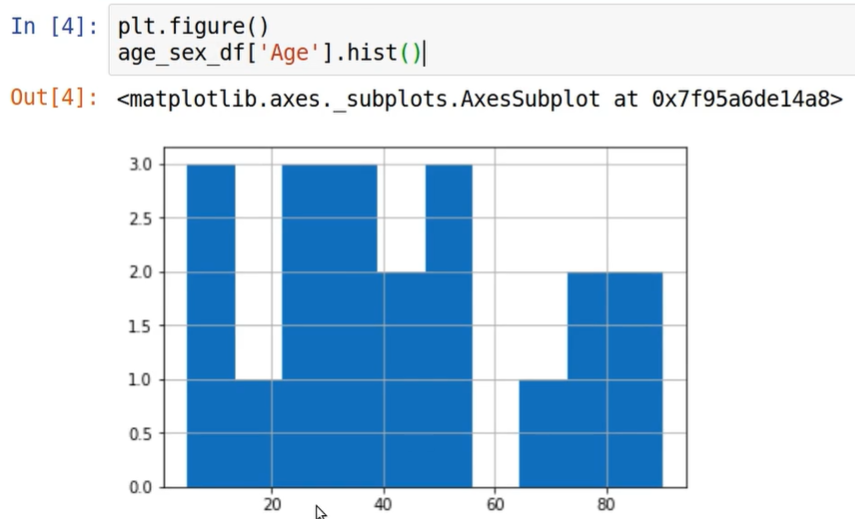
Ordinal Variable



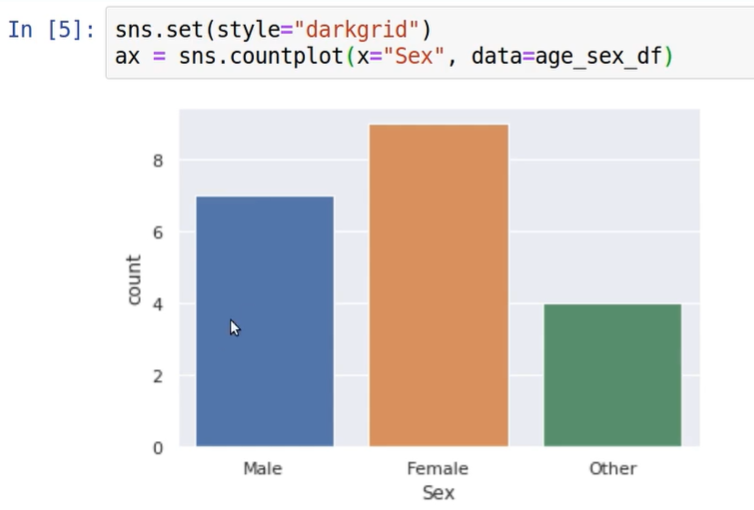


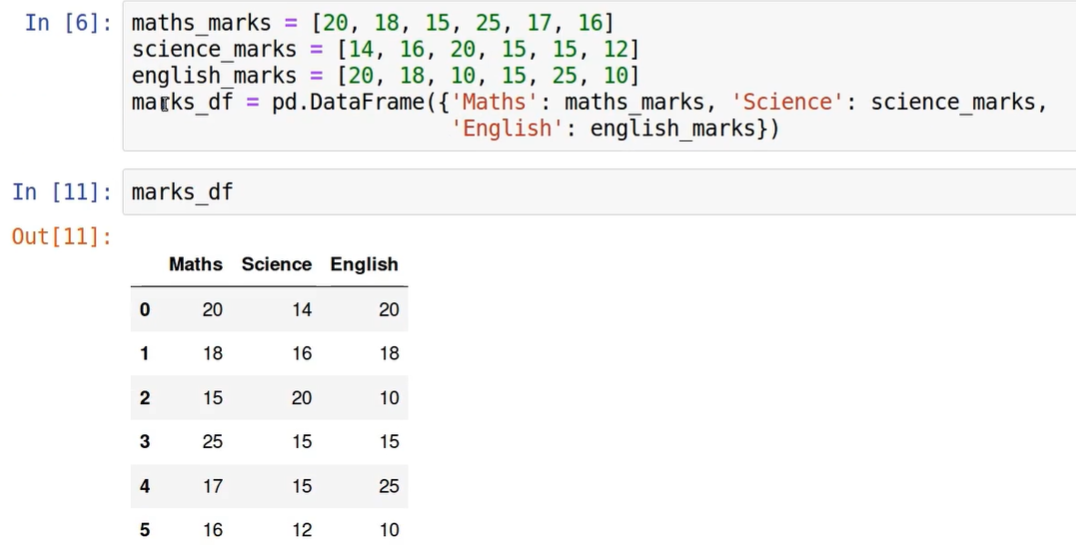


Histogram

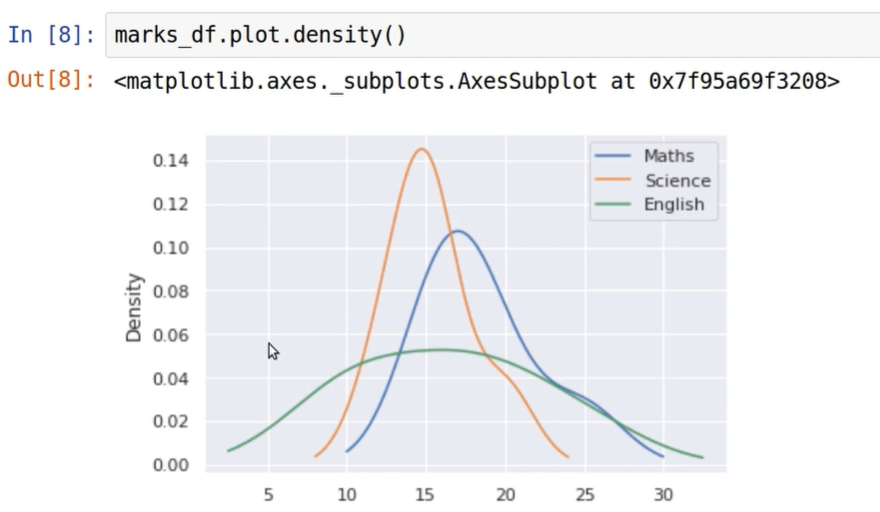


Bar chart

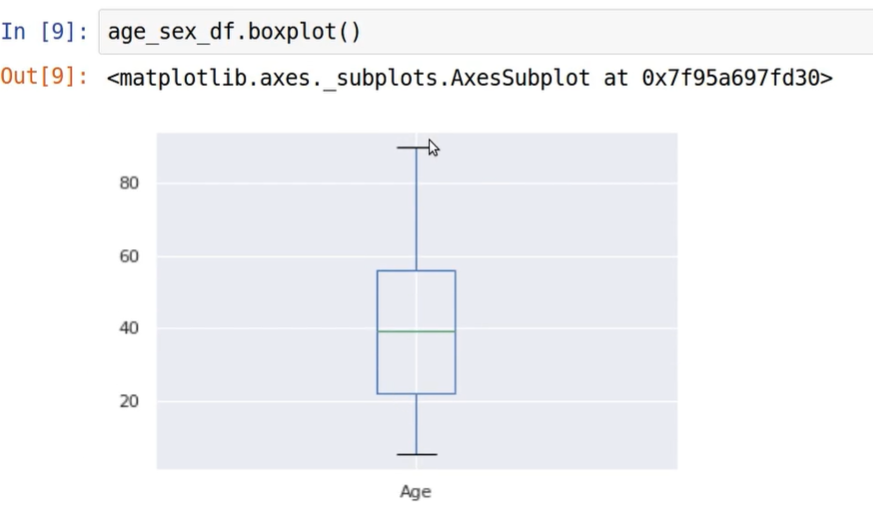


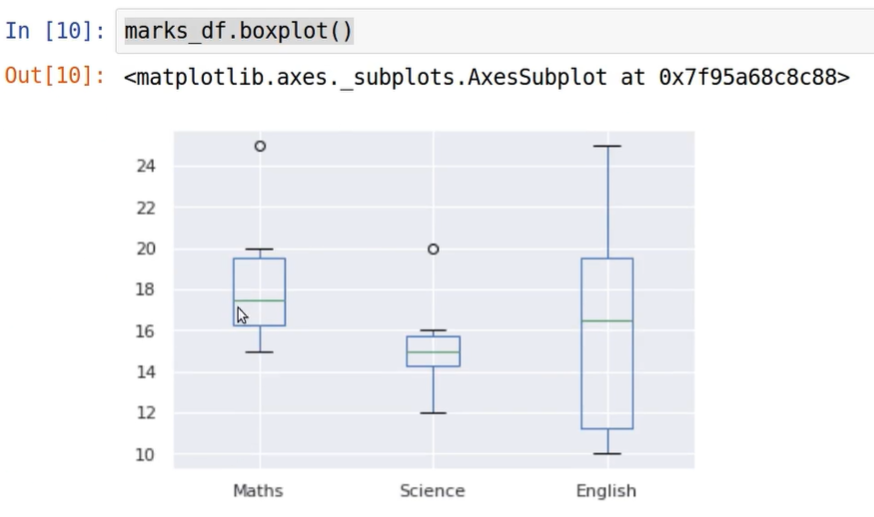


Density plots

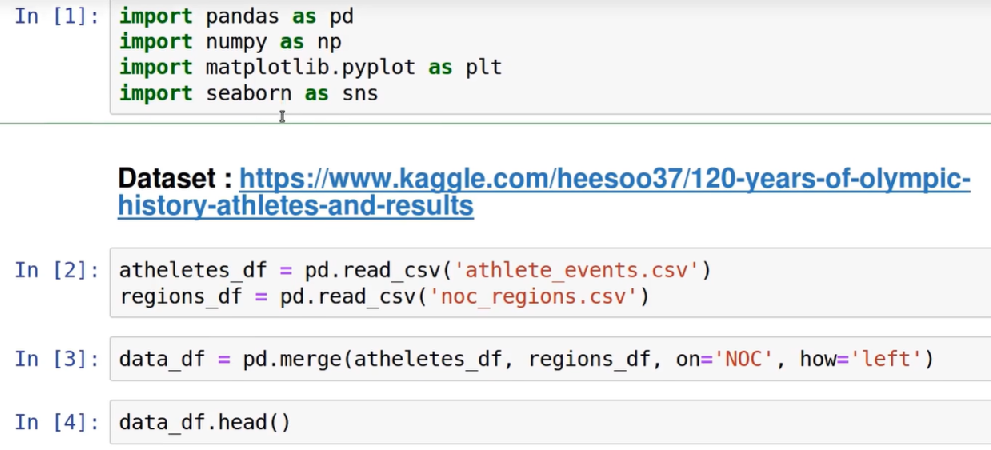


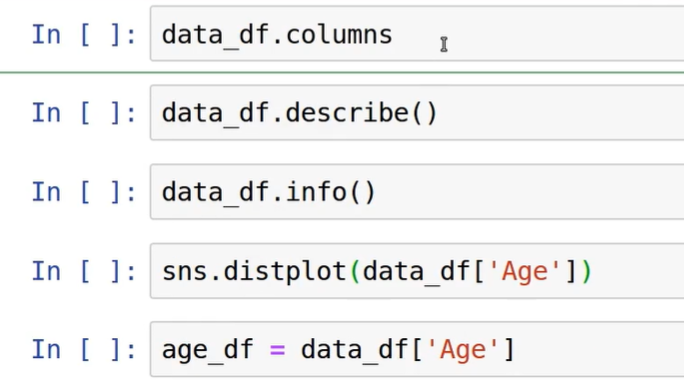
Box plot





Univariate Analysis over Olympic dataset





Springboard: Exploratory Data Analysis with Pandas and Python 3.x