**Experiment 1: Exploratory data analysis using Python**

**Aim:** To understand the data through Exploratory data analysis

* Data cleaning- Missing Values, remove outliers
* Data transformation- Min-max normalization, Z-score normalization, Decimal Scaling
* Data Discretization- Binning
* Data analysis and Visualization

**Theory:**

Data Cleaning:

Data cleaning, also known as data cleansing or data scrubbing, is a crucial step in the data preparation process of any data analysis or machine learning project. It involves identifying and correcting errors, inconsistencies, and inaccuracies in datasets to ensure that the data is accurate, reliable, and ready for analysis.

Some common problems found in datasets are:

* Missing Values
* Outliers
* Inconsistent Formats, etc.

Some functions used for data cleaning:

* dropna()
* fillna()
* drop\_duplicates()

Data transformation:

Data transformation refers to the process of converting or altering the raw data in a way that makes it more suitable for analysis, modeling, or visualization. It involves applying various mathematical, statistical, or logical operations to the data to achieve specific objectives, such as improving data quality, normalizing scales, handling outliers, or making the data conform to assumptions required by certain analytical methods.

Some of the techniques involved in data transformation are:

* Normalization
* Standardization
* Min-Max normalization
* Decimal Scaling

Data discretization:

Data discretization, also known as binning or discretization, is the process of converting continuous or numeric data into discrete bins or intervals. In other words, it involves dividing a continuous variable's range into smaller, non-overlapping intervals and assigning data points to the corresponding interval. This transformation is particularly useful when working with data analysis, visualization, or certain types of machine learning algorithms that benefit from reduced data granularity or when data is naturally presented in grouped or categorized form.

Some functions that help us perform data discretization:

* pandas.qcut
* pandas.cut

Data visualisation:

Data visualization is the graphical representation of information and data. It involves using visual elements like charts, graphs, and maps to present complex data in a more accessible and understandable format. Data visualization is a powerful tool for exploring, analyzing, and communicating insights from data, making it an essential part of data analysis

The data can be presented in various forms via data visualisation for a better understanding of the data and some of those forms are:

* Pie-chart
* Histogram
* Scatterplot, etc.

Data analysis:

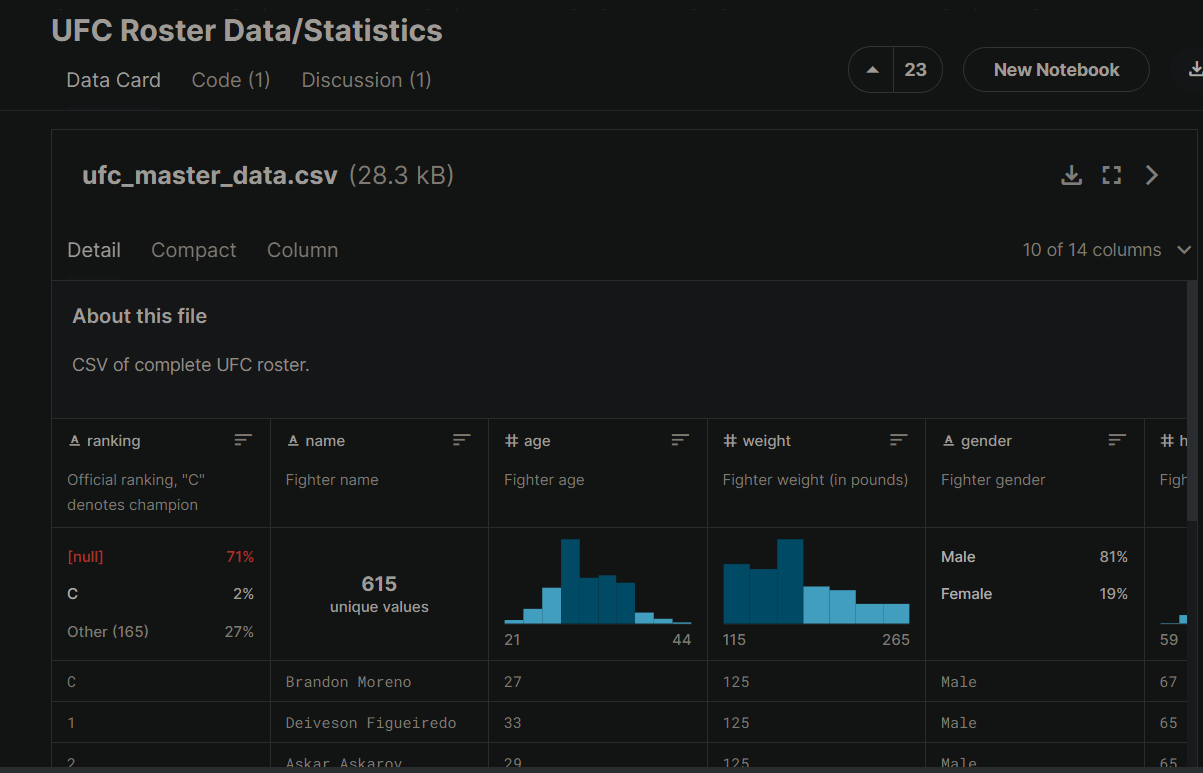
Data analysis is the process of inspecting, cleaning, transforming, and interpreting data to extract meaningful insights, discover patterns, and make informed decisions. It involves using various techniques, tools, and methodologies to understand the underlying structure of data, identify trends, relationships, and anomalies, and derive actionable information from the data.

Some functions used for data analysis are:

* .head()
* .info()
* .describe()

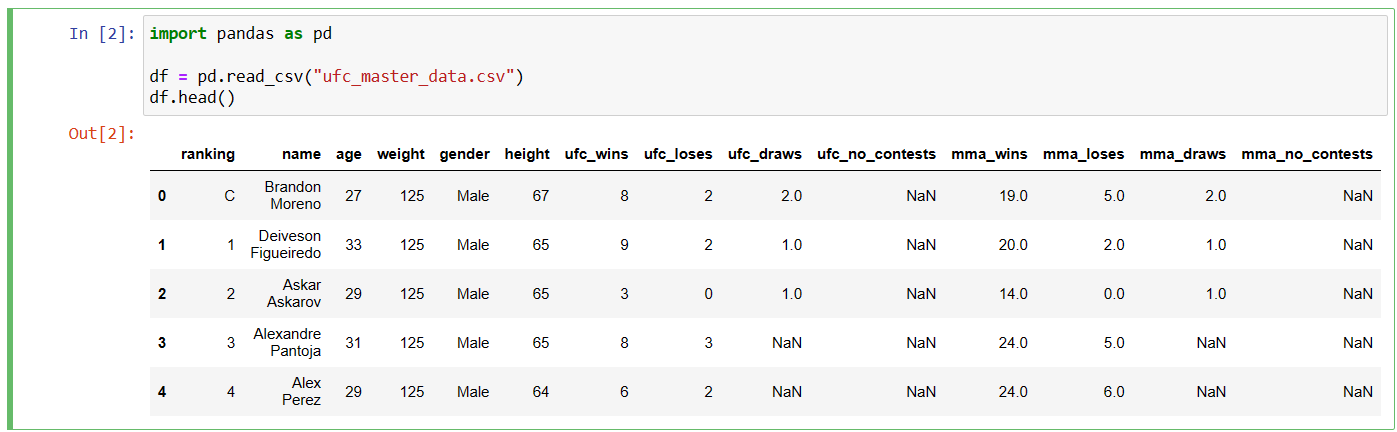
**Steps:**

1. **Load the libraries Download the data set from kaggle/ other sources**

****

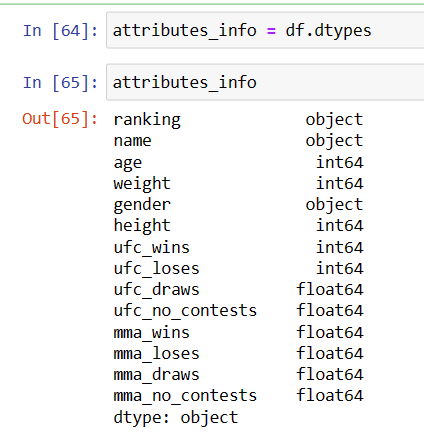
****

1. **Read the file –select appropriate file read function according to data type of file**

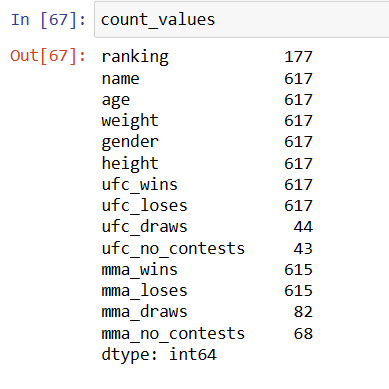
****

1. **Describe the attributes name, count no of values, and find min, max, data type, range, quartile, percentile, box plot and outliers.**

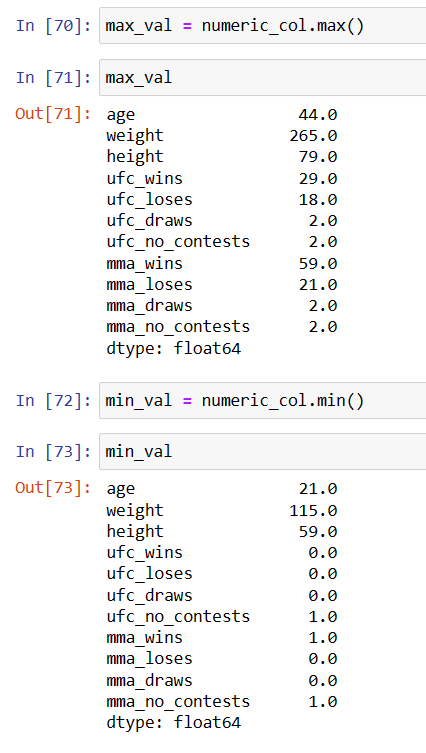
Attribute names and data-type:



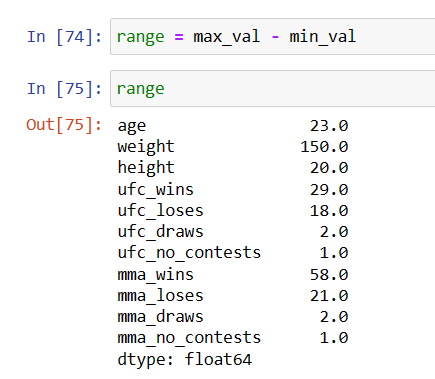
No. of values:



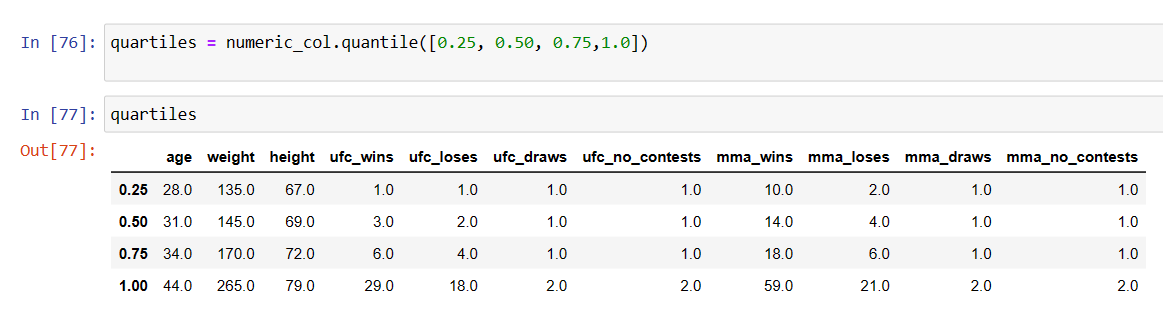
Min and Max:



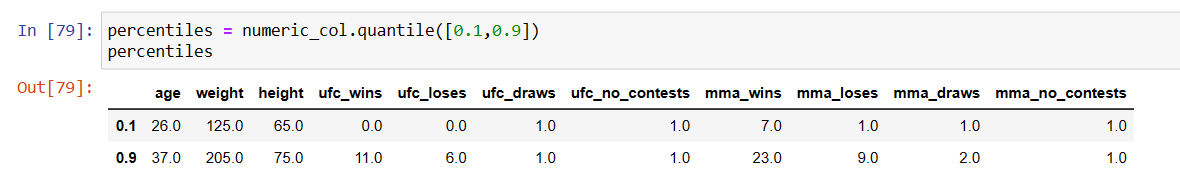
Range:



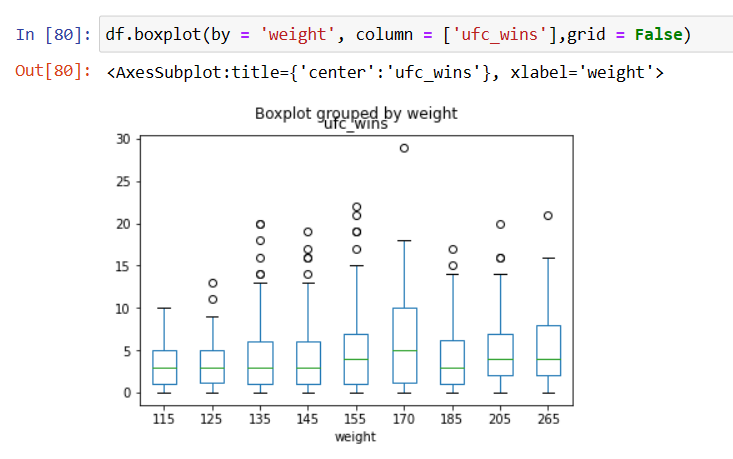
Quartile:



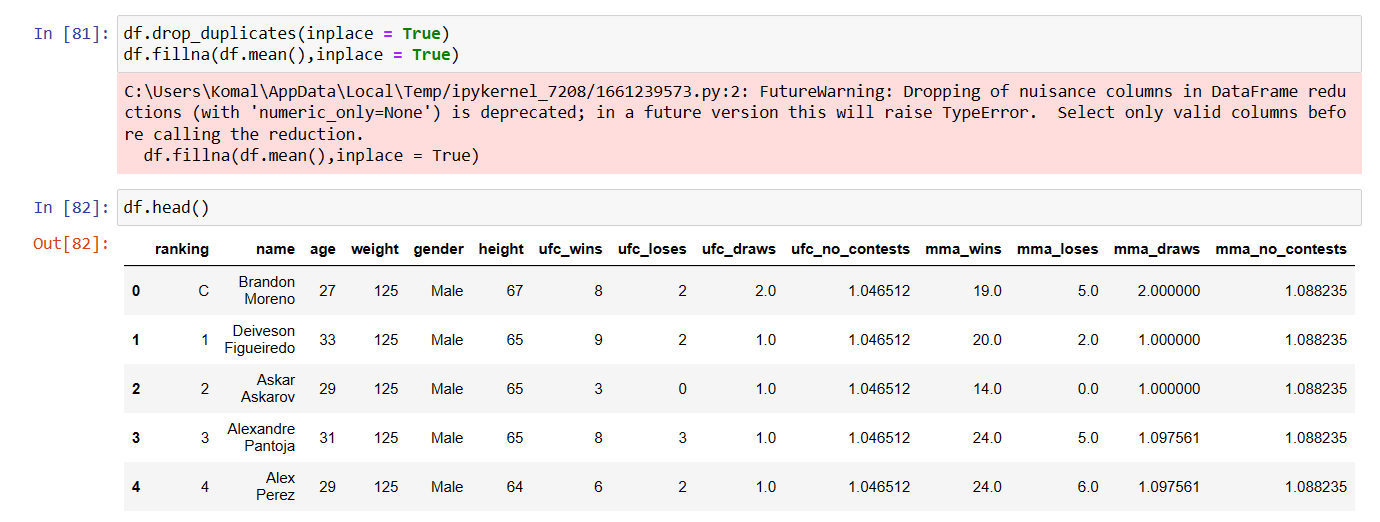
Percentile:

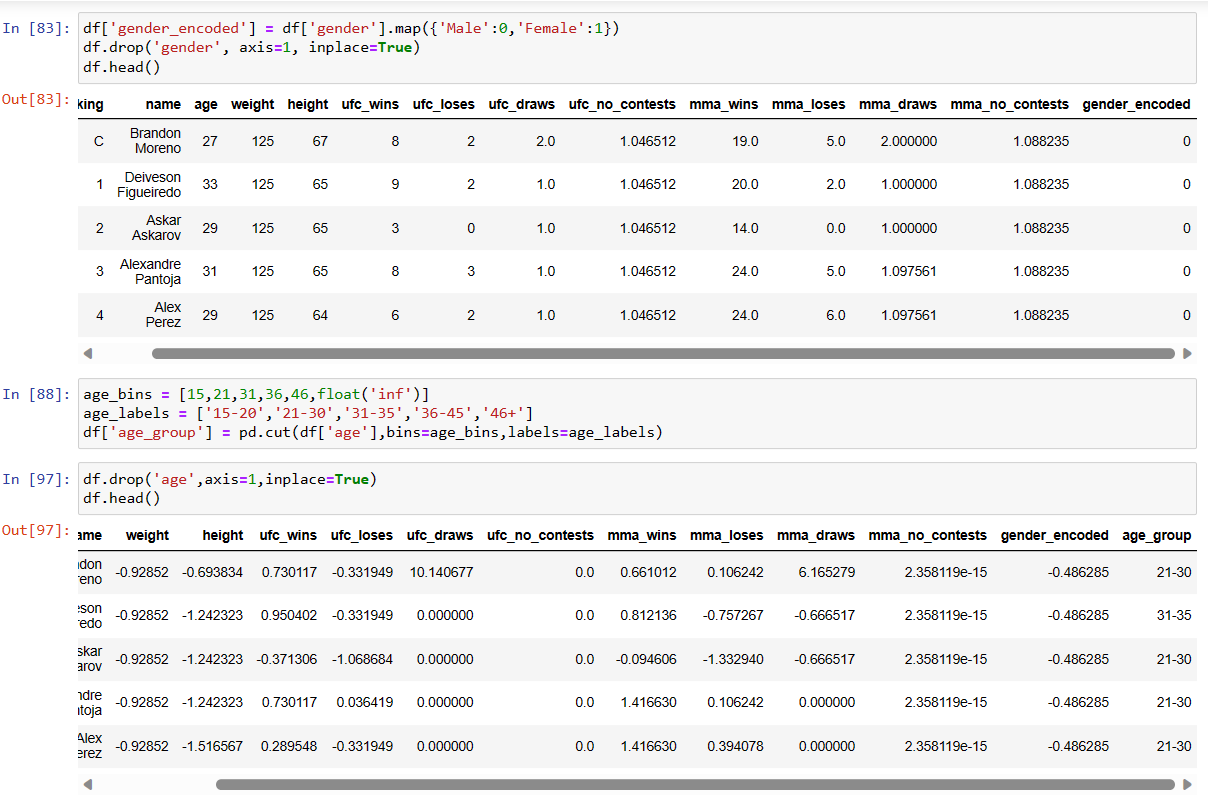


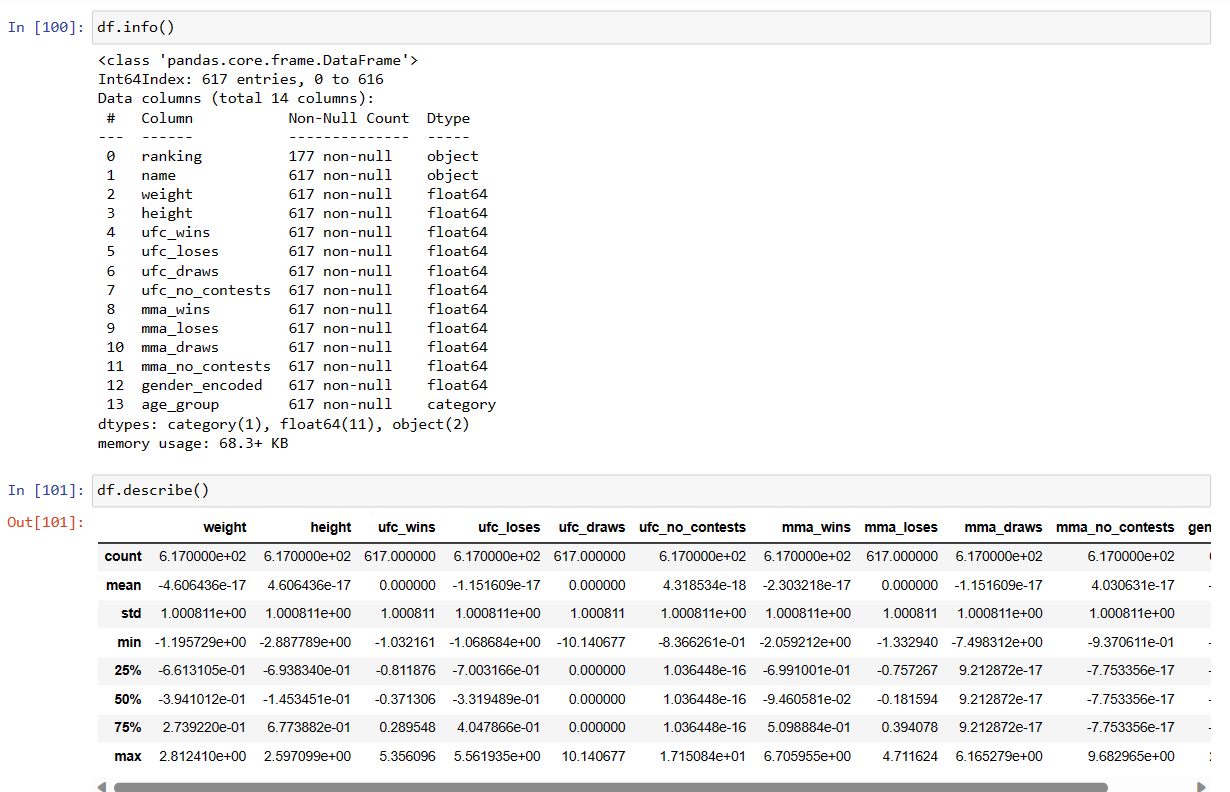
Boxplot and outliers:



1. **Perform cleaning,transformation , discretization and analysis**

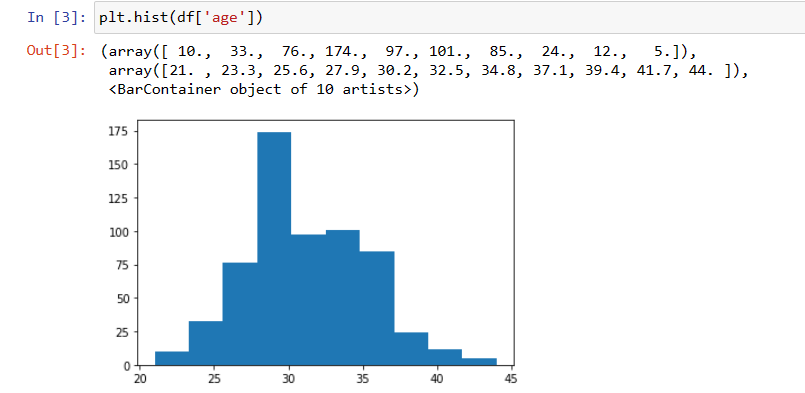




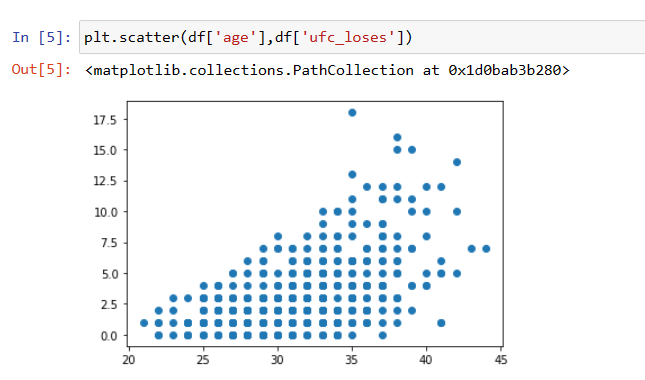


1. **Give visualization of statistical description of data – in form of histogram, scatter plot, pie chart,Give correlation matrix**

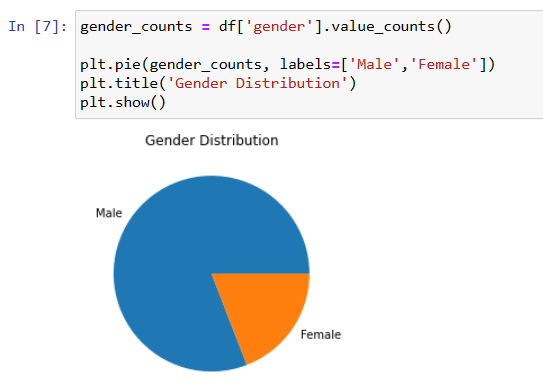
Histogram:



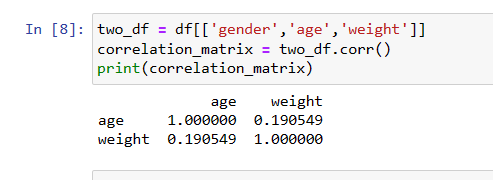
Scatter plot:



Piechart:



Correlation matrix:



Frequency table:

