**DS USING PYTHON LAB**

**EXPERIMENT: 04**

**AIM:** Implementation of Statistical Hypothesis Test using Scipy and Sci-kit learn.

**PROBLEM STATEMENT:**

Perform the following Tests:

1. Pearson’s Correlation Coefficient
2. Spearman’s Rank Correlation
3. Kendall’s Rank Correlation
4. Chi-Squared Test

**THEORY:**

**Scipy**

SciPy, a scientific library for Python, is an open source, BSD-licensed library for mathematics, science and engineering. The SciPy library depends on NumPy, which provides convenient and fast N-dimensional array manipulation. The main reason for building the SciPy library is that it should work with NumPy arrays. It provides many user-friendly and efficient numerical practices such as routines for numerical integration and optimization.

**Sci-kit learn**

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistent interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

**Statistical Hypothesis Test**

Hypothesis testing is a statistical analysis that uses sample data to assess two mutually exclusive theories about the properties of a population. Statisticians call these theories the null hypothesis and the alternative hypothesis. A hypothesis test assesses your sample statistic and factors in an estimate of the sample error to determine which hypothesis the data support.

When you can reject the null hypothesis, the results are statistically significant, and your data support the theory that an effect exists at the population level.

**Pearson’s Correlation Coefficient**

Pearson’s correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

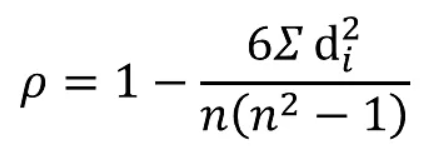
Degree of correlation:

* Perfect: If the value is near ± 1, then it is said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).
* High degree: If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.
* Moderate degree: If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.
* Low degree: When the value lies below + .29, then it is said to be a small correlation.
* No correlation: When the value is zero.

**Spearman’s Rank Correlation**

Spearman’s rank correlation measures the strength and direction of association between two ranked variables. It basically gives the measure of monotonicity of the relation between two variables i.e. how well the relationship between two variables could be represented using a monotonic function.

The formula for Spearman’s rank coefficient is:



𝝆 = Spearman’s rank correlation coefficient

di = Difference between the two ranks of each observation

n = Number of observations

The Spearman Rank Correlation can take a value from +1 to -1 where,

A value of +1 means a perfect association of rank

A value of 0 means that there is no association between ranks

A value of -1 means a perfect negative association of rank

**Kendall’s Rank Correlation**

Kendall rank correlation is used to test the similarities in the ordering of data when it is ranked by quantities. Kendall’s correlation coefficient uses pairs of observations and determines the strength of association based on the patter on concordance and discordance between the pairs.

* Concordant: Ordered in the same way (consistency). A pair of observations is considered concordant if (x2 — x1) and (y2 — y1) have the same sign.
* Discordant: Ordered differently (inconsistency). A pair of observations is considered concordant if (x2 — x1) and (y2 — y1) have opposite signs.

Kendall rank correlation (non-parametric) is an alternative to Pearson’s correlation (parametric) when the data you’re working with has failed one or more assumptions of the test. This is also the best alternative to Spearman correlation (non-parametric) when your sample size is small and has many tied ranks.

**Chi-Squared Test**

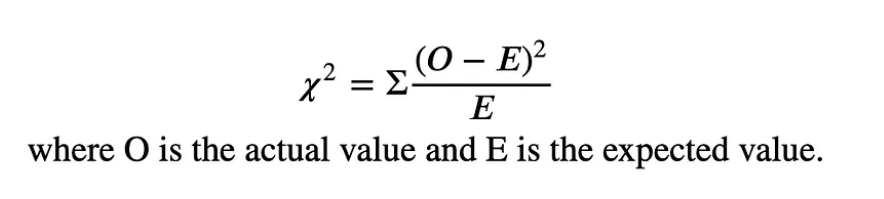
A chi-squared test (symbolically represented as χ2) is basically a data analysis on the basis of observations of a random set of variables. Usually, it is a comparison of two statistical data sets. A hypothesis is a consideration that a given condition or statement might be true, which we can test afterwards. The chi-square test is used to estimate how likely the observations that are made would be, by considering the assumption of the null hypothesis as true.

Chi-squared tests are usually created from a sum of squared falsities or errors over the sample variance.

P stands for probability here. To calculate the p-value, the chi-square test is used in statistics. The different values of p indicates the different hypothesis interpretation, are given below:

* P ≤ 0.05; Hypothesis rejected
* P > 0.05; Hypothesis accepted

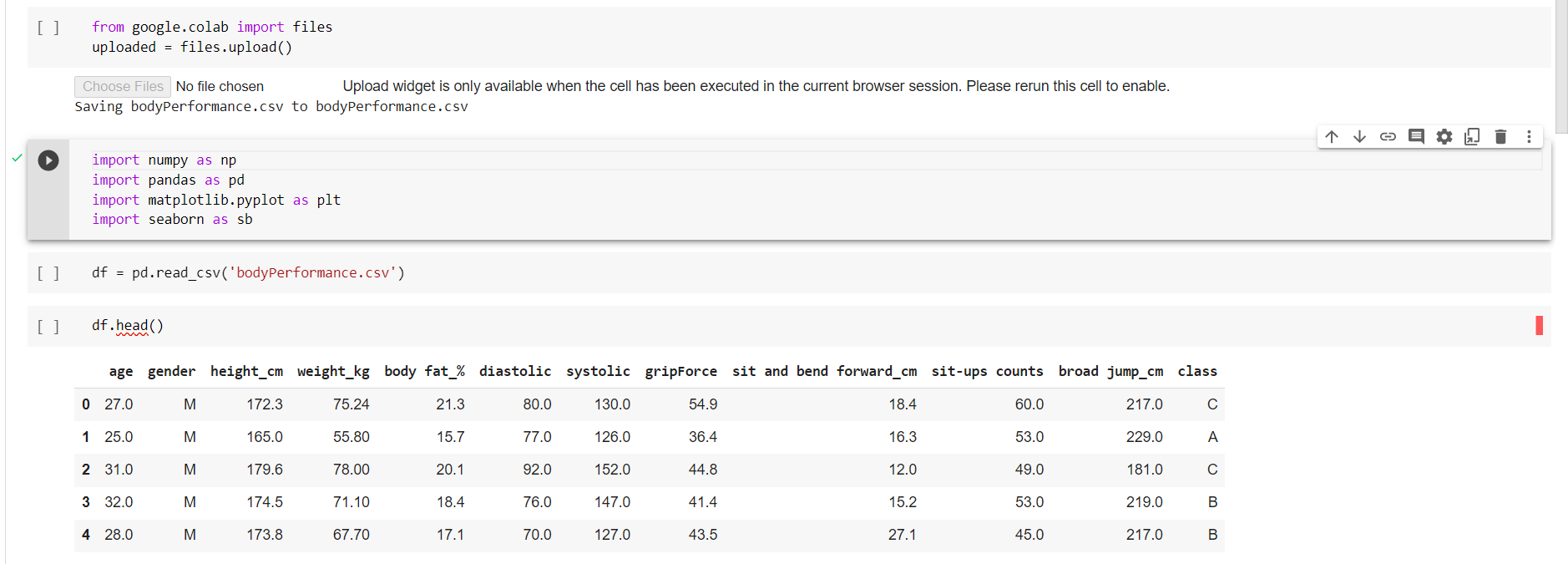
The formula for the Chi-square is:



The 𝜒2 equation tells us to find the square of the difference between the actual value and expected value and divide it by the expected value. Then add all together to find the 𝜒2 value.

**IMPLEMENTATION:**

1. **Loading dataset into Google collab.**



1. **Pearson’s Correlation Coefficient**

Pandas dataframe.corr() is used to find the pairwise correlation of all columns in the Pandas Dataframe in Python. Any NaN values are automatically excluded. Any non-numeric data type or columns in the Dataframe, it is ignored.

Syntax: DataFrame.corr(self, method=’pearson’, min\_periods=1)

Parameters:

a]method :

pearson: standard correlation coefficient

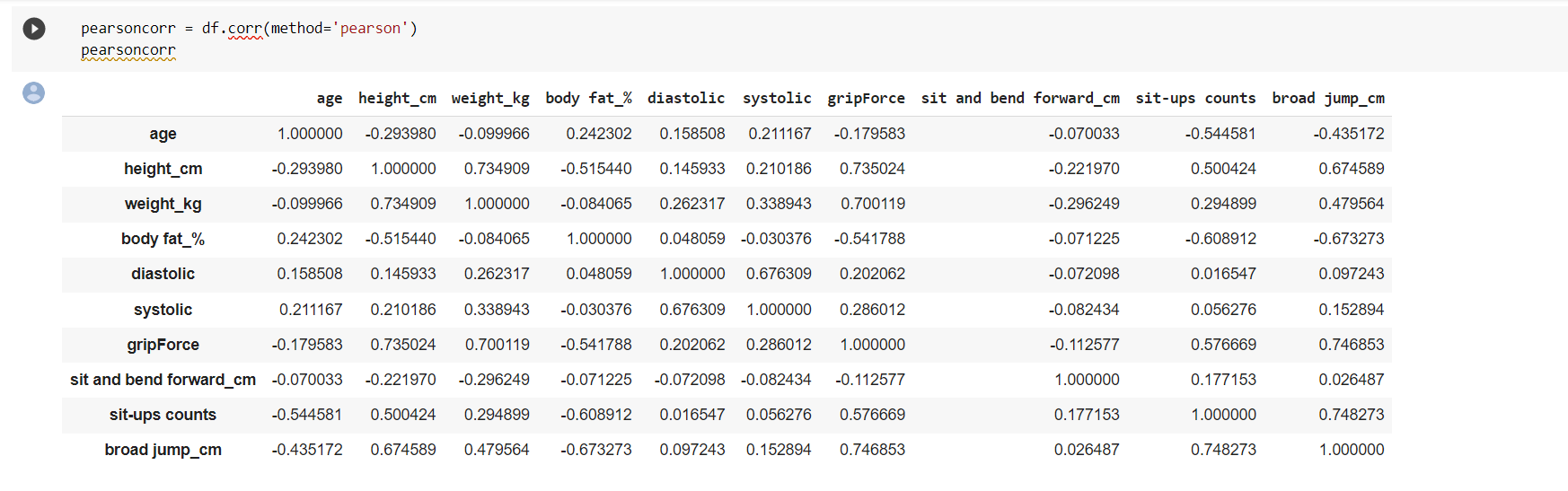
kendall: Kendall Tau correlation coefficient

spearman: Spearman rank correlation

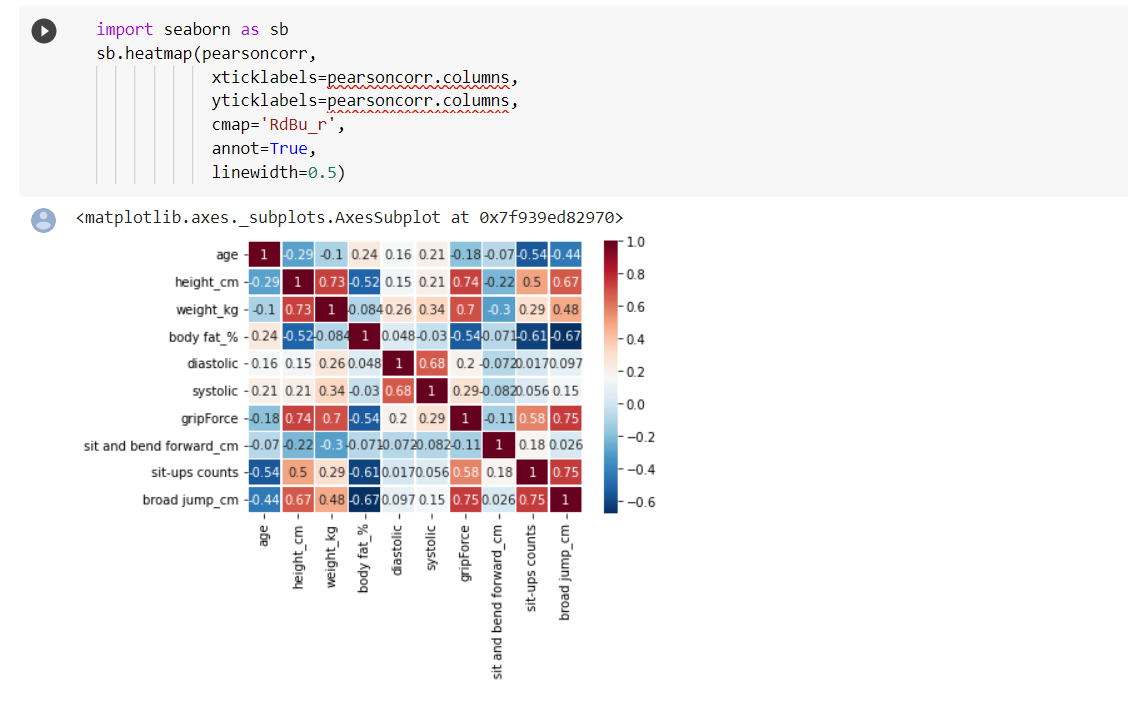
b]min\_periods :

Minimum number of observations required per pair of columns to have a valid result. Currently only available for pearson and spearman correlation

Returns: count :y : DataFrame



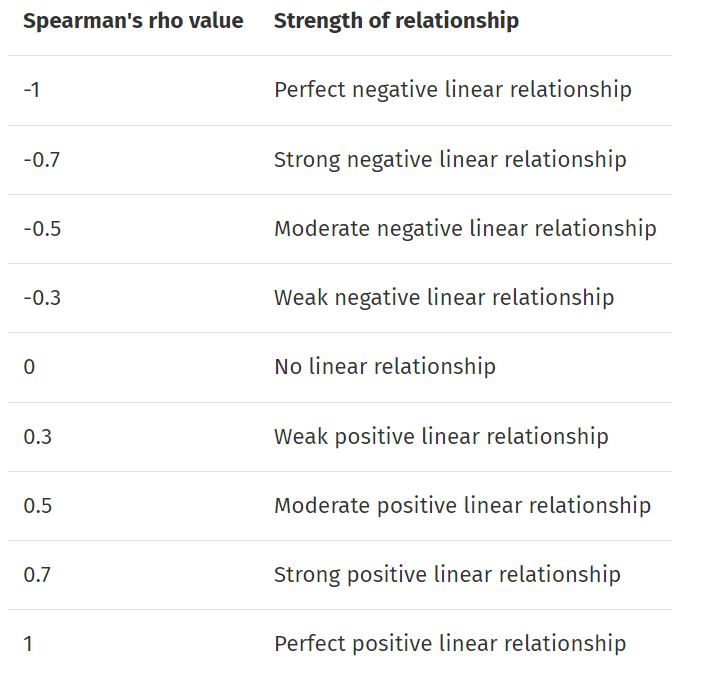
To make this look beautiful and easier to interpret, we have made a heat map after calculating the Pearson coefficient of correlation.



1. **Spearman’s Rank Correlation**

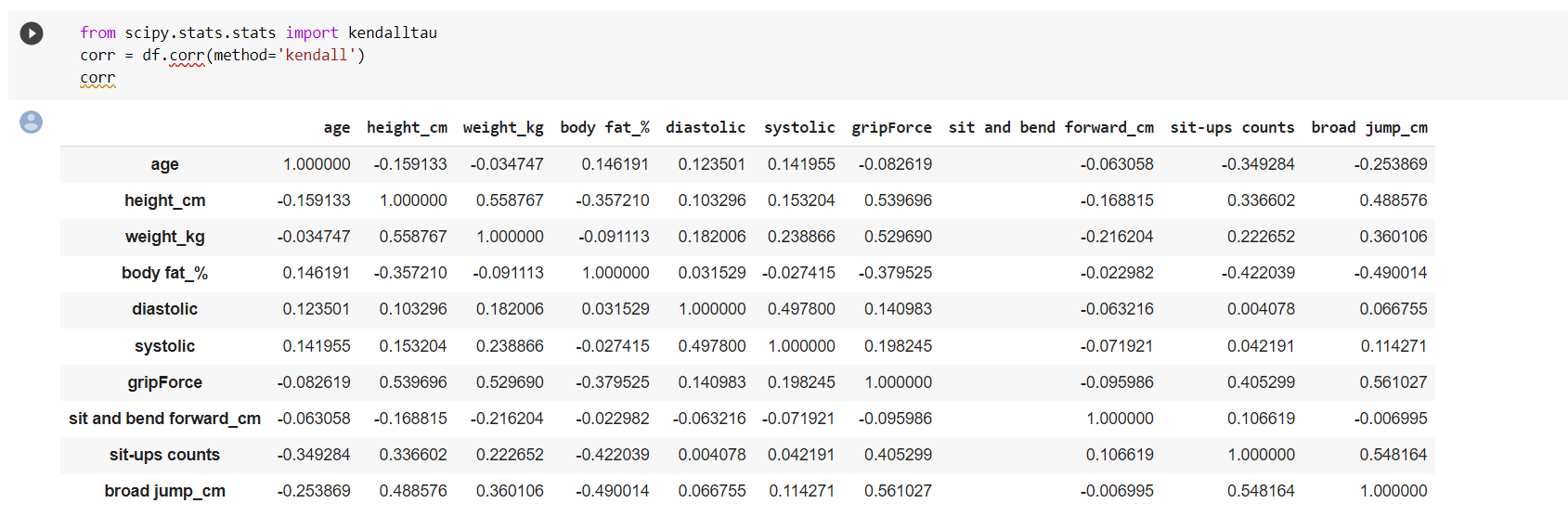


The correlation coefficient returned, will be a value between -1 and +1. Here’s how you can interpret what these coefficients mean:

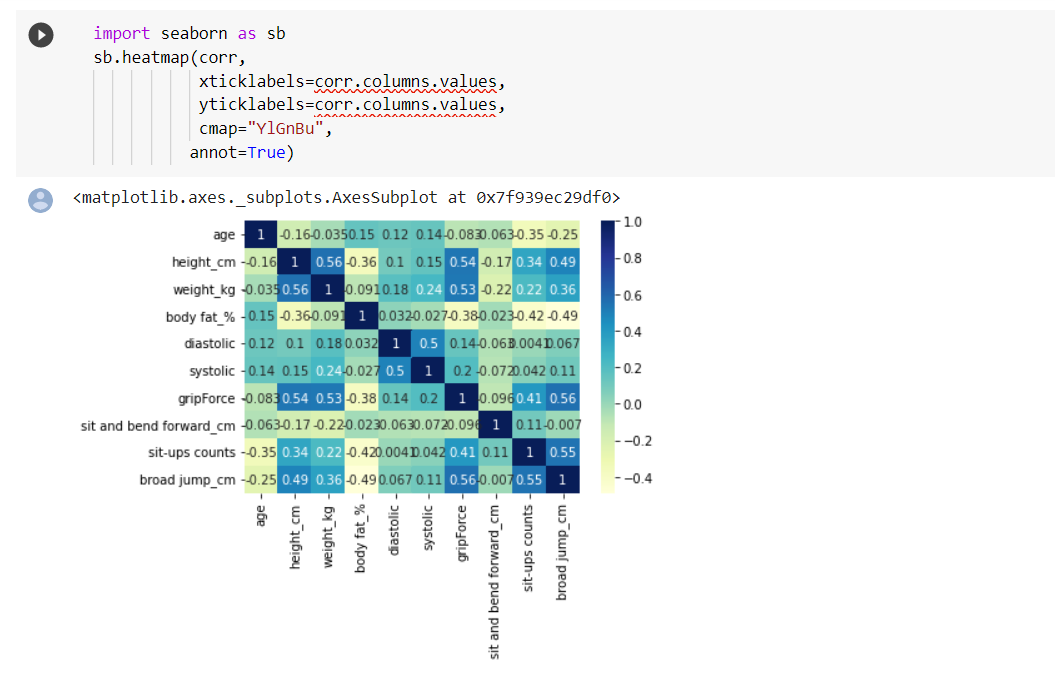


1. **Kendall’s Rank Correlation**

Pandas dataframe.corr() is used to find the pairwise correlation of all columns in the dataframe.

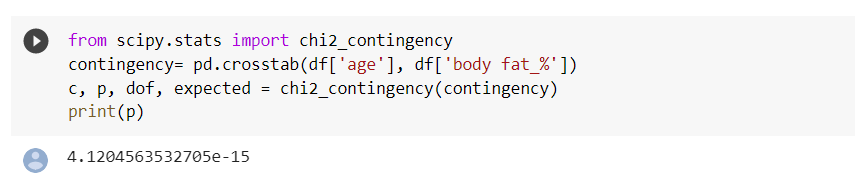


Visualize using a Heat-map



1. **Chi-Squared Test**

SciPy’s chi2\_contingency() returns four values, 𝜒2 value, p-value, degree of freedom and expected values.



**CONCLUSION:**

We have studied and implemented different correlation tests. These correlation tests were used to identify the relationships between different variables in the dataset. We assumed the hypothesis and the correlation tests helped us verify if they were accepted or rejected.