**EXPERIMENT : 0**

**Aim :** To study basics of dataset, method of statistics and analysis of iris data set using pandas library in ipython.

**Reqiurements :** Jupiter notebook.

**Theory :**

**1. Basic Defintions :**

* **Dataset:** Dataset is a collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer. In other words, data set is a combination of data having common features
* **Features of Dataset :** A [**feature**](https://www.datarobot.com/wiki/feature/) is a measurable property of the object you’re trying to analyze. In datasets, features appear as columns.
* **Inferential statistics :** inferential Statistics are produced by more complex mathematical calculations, and allow us to infer trends and make assumptions and predictions about a population based on a study of a sample taken from it. Inferential statistics is one of the two main branches of statistics.Inferential statistics use a random sample of data taken from a population to describe and make inferences about the population. Inferential statistics are valuable when examination of each member of an entire population is not convenient or possible. For example, to measure the diameter of each nail that is manufactured in a mill is impractical. You can measure the diameters of a representative random sample of nails. You can use the information from the sample to make generalizations about the diameters of all of the nails.
* **Descriptive analysis :**In Descriptive Statistics your are describing, presenting, summarizing and organizing your data (population), either through numerical calculations or graphs or tables. Descriptive analysis contrasts with *inferential statistics* which draws conclusions about a population based on a sample of data. Descriptive analysis also contrasts with *predictive analytics*where the focus is on extrapolating from the past to predict future trends and outcomes. And descriptive analysis contrasts with prescriptive analytics, which takes prediction a step further to not only predict what will happen, but also recommend a course of action based on the most rational response to the distribution of future events.
* **ARITHMETIC MEAN** : Sum of all the observations and divided by the number of observations results the arithmetic mean. Symbolically, the arithmetic mean, also called simply mean is

x = ∑x/n , where x is simple mean.

* **MEDIAN** Median is defined as the value of the middle item (or the mean of the values of the two middle items) when the data are arranged in an ascending or descending order of magnitude. Thus, in an ungrouped frequency distribution if the n values are arranged in ascending or descending order of magnitude, the median is the middle value if n is odd. When n is even, the median is the mean of the two middle values.
* **MODE :** The mode is another measure of central tendency. It is the value at the point around which the items are most heavily concentrated.
* **STANDARD DEVIATION** : The standard deviation is similar to the mean deviation in that here too the deviations are measured from the mean. The *standard deviation* of a population of values is computed as:

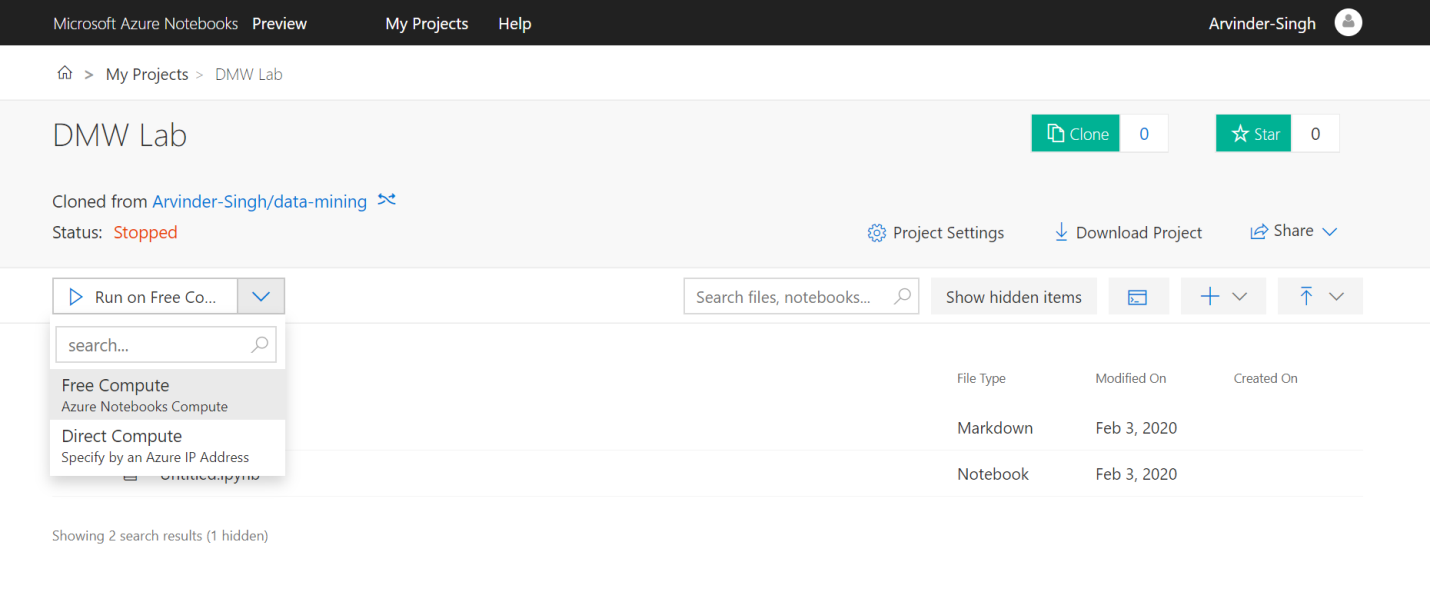
 = [(xi-)2/N]1/2

* **Correlation :** The correlation between variables is a measure of the nature and degree of association between the variables.
* **Variance :** This mean of the squared deviations is known as the variance. Symbolically,

Var (X) = 2 = (xi-)2/N

**Procedure :**

1. Open Jupiter notebook. Here we have used Microsoft Azure Notebook to access Jupiter notebook.
2. Create a new project and open it .
3. To start working, click on “Free Compute” .



1. In Now select Python 3.6 in Notebook section and write following code on interface provided.

**CODE:**

import pandas as pd

df=pd.read\_csv("https://gist.githubusercontent.com/curran/a08a1080b88344b0c8a7/raw/639388c2cbc2120a14dcf466e85730eb8be498bb/iris.csv")

print(df.shape)

print(df.info)

df.mean()

df.median()

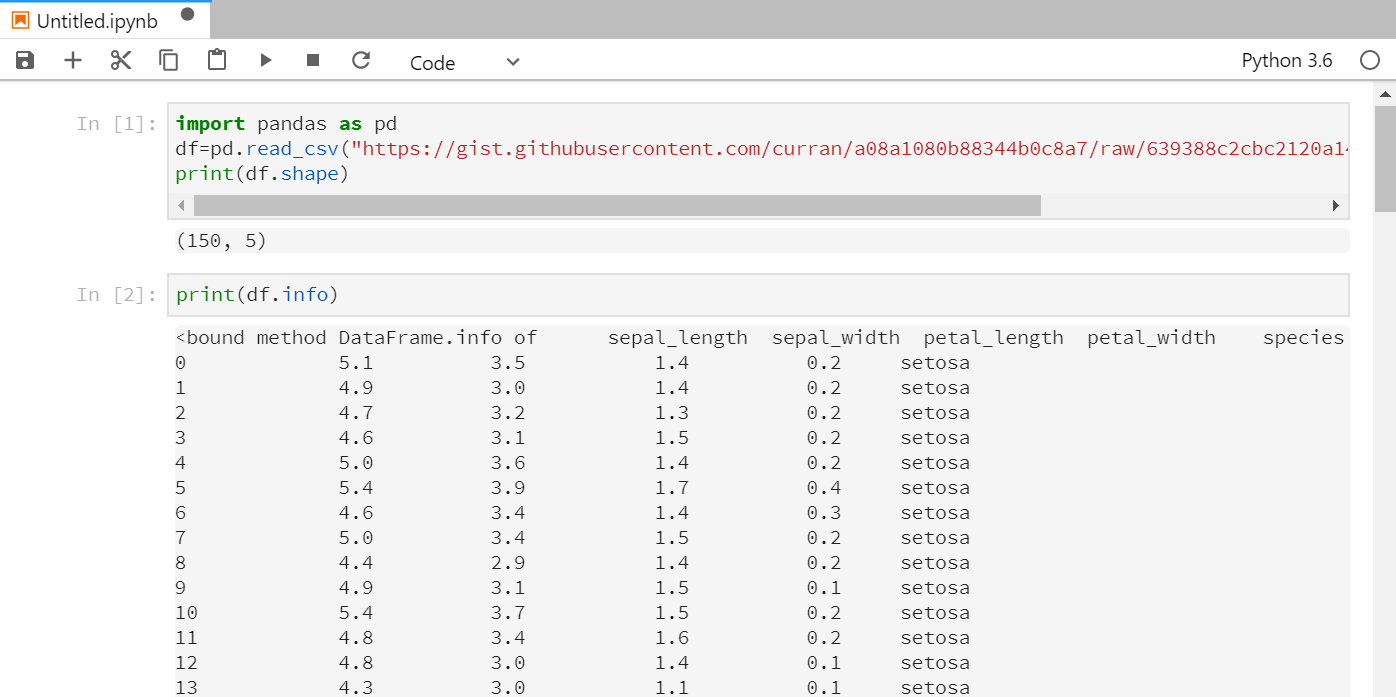
df.mode()

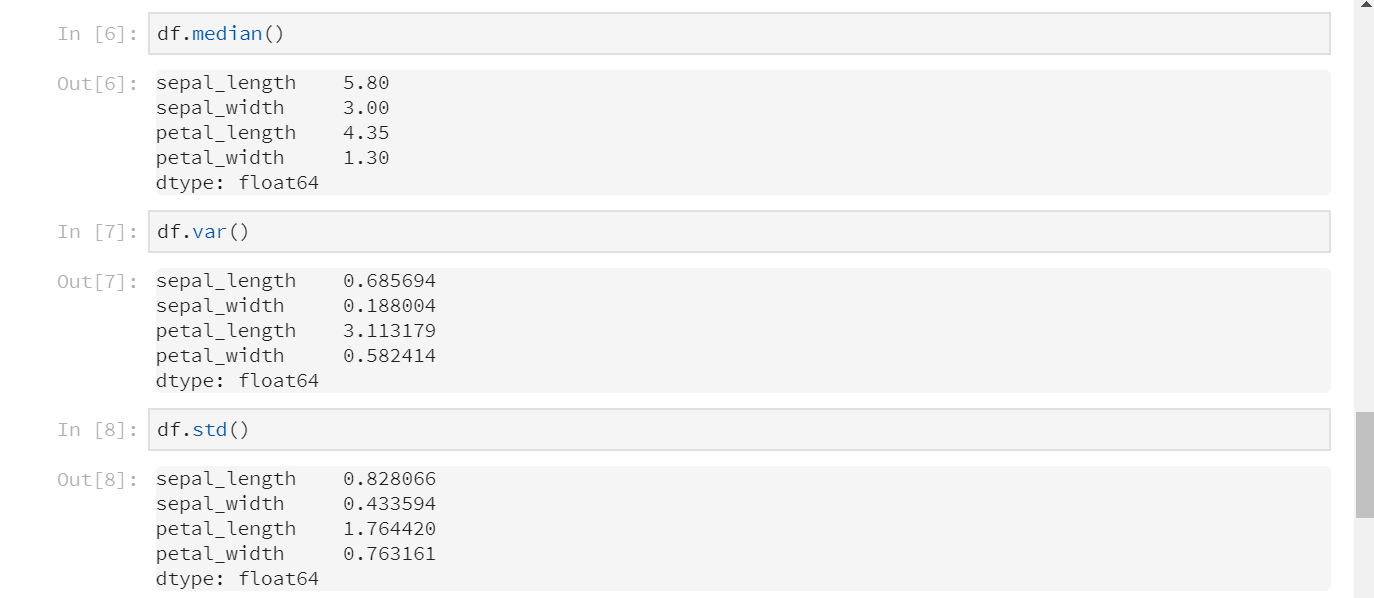
df.std()

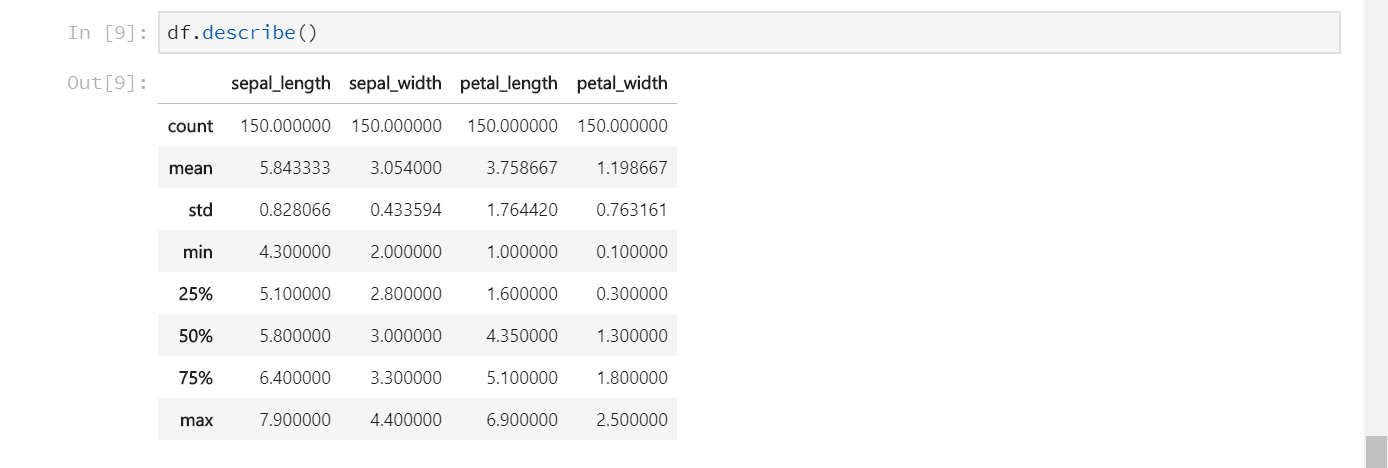
df.var()

df.describe()

1. The output will appear as :



** **

****

**EXPERIMENT : 1**

**Aim :** To study and implement Descriptive Analysis using statistica.

**Reqiurements :** Statistica, iris data set (irirs.csv).

**Theory :**

* **Inferential statistics :** inferential Statistics are produced by more complex mathematical calculations, and allow us to infer trends and make assumptions and predictions about a population based on a study of a sample taken from it. Inferential statistics is one of the two main branches of statistics.Inferential statistics use a random sample of data taken from a population to describe and make inferences about the population. Inferential statistics are valuable when examination of each member of an entire population is not convenient or possible. For example, to measure the diameter of each nail that is manufactured in a mill is impractical. You can measure the diameters of a representative random sample of nails. You can use the information from the sample to make generalizations about the diameters of all of the nails.
* **Descriptive analysis :**In Descriptive Statistics your are describing, presenting, summarizing and organizing your data (population), either through numerical calculations or graphs or tables. Descriptive analysis contrasts with *inferential statistics* which draws conclusions about a population based on a sample of data. Descriptive analysis also contrasts with *predictive analytics*where the focus is on extrapolating from the past to predict future trends and outcomes. And descriptive analysis contrasts with prescriptive analytics, which takes prediction a step further to not only predict what will happen, but also recommend a course of action based on the most rational response to the distribution of future events.
* **ARITHMETIC MEAN** : Sum of all the observations and divided by the number of observations results the arithmetic mean. Symbolically, the arithmetic mean, also called simply mean is

x = ∑x/n ,

where x is simple mean.

* **MEDIAN** Median is defined as the value of the middle item (or the mean of the values of the two middle items) when the data are arranged in an ascending or descending order of magnitude. Thus, in an ungrouped frequency distribution if the n values are arranged in ascending or descending order of magnitude, the median is the middle value if n is odd. When n is even, the median is the mean of the two middle values.
* **MODE :** The mode is another measure of central tendency. It is the value at the point around which the items are most heavily concentrated.
* **STANDARD DEVIATION** : The standard deviation is similar to the mean deviation in that here too the deviations are measured from the mean. The *standard deviation* of a population of values is computed as:

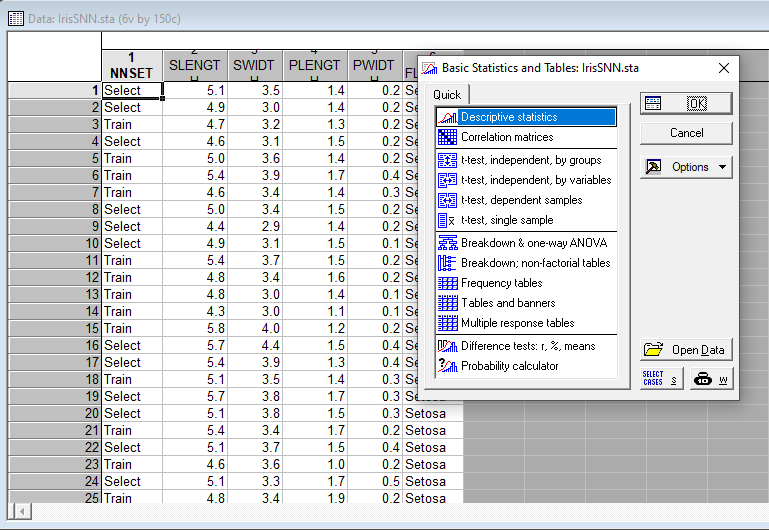
 = [(xi-)2/N]1/2

* **Correlation :** The correlation between variables is a measure of the nature and degree of association between the variables.
* **Variance :** This mean of the squared deviations is known as the variance. Symbolically,

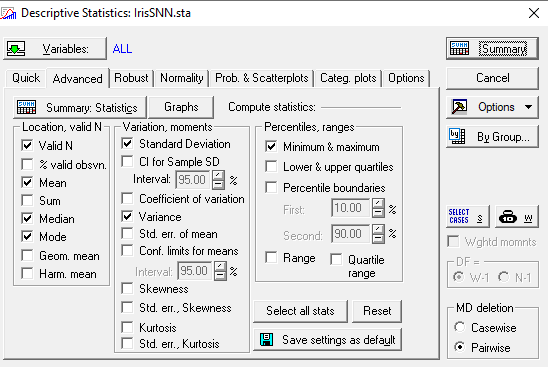
Var (X) = 2 = (xi-)2/N

**Procedure :**

1. Load the dataset ( here IrisSNN.sta is used as dataset) on which the analysis is to be performed.
2. Choose “Basic statistics/Tables” option from Statistics menu in menu bar. Following window will apper after this.

****

1. Choose Descriptive statistcs and click OK.
2. Select from the various options avai**EXPERIMENT**le under advanced tab and make chioces as per requirement.



1. Click on Summary tab to view the result result in tabular form

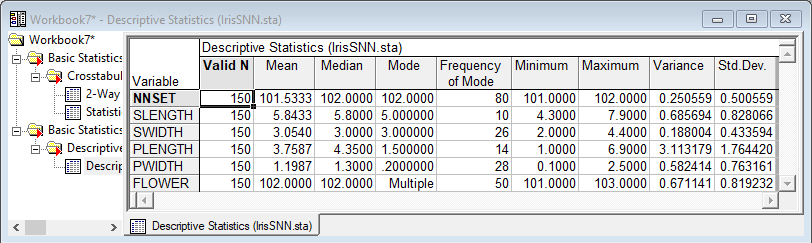
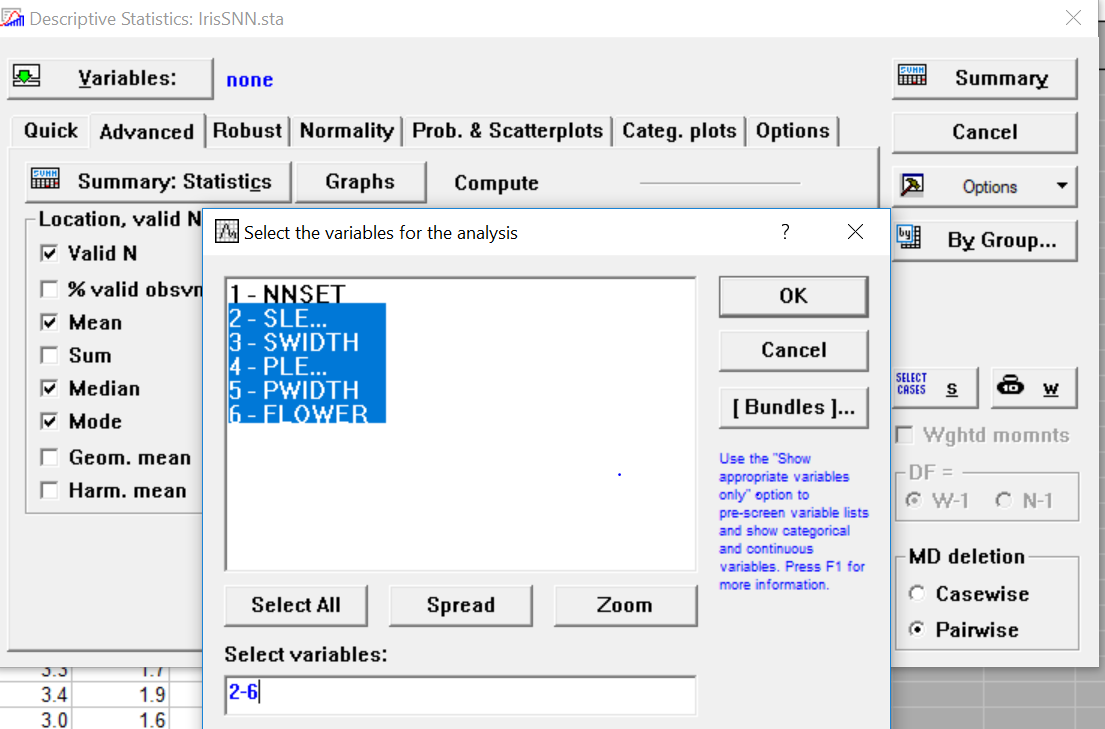
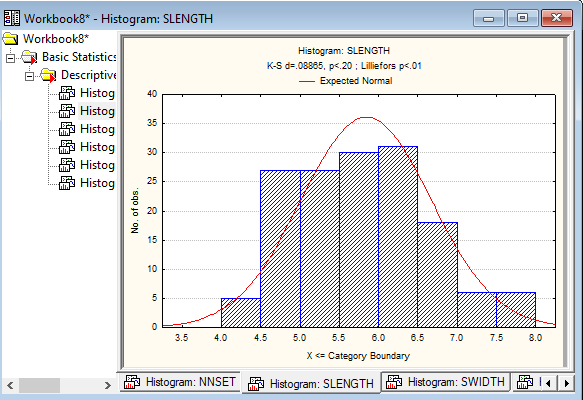
****

Table showing mean, median, mode, variance and standard deviation of the dataset across various attributes like sepal length, sepal width, etc.

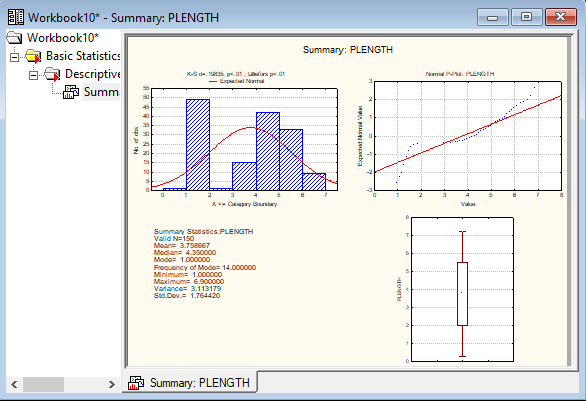
1. We can also choose Graphs tab and select the variables as per requirement to get results in graphical form.

****

various attributes like sepal length, sepal width, etc.



Representation of sepal length in the form of a histogram



Detailed summary of all the flowers based upon petal length

**EXPERIMENT : 2**

**Aim :** To study the basics of

* Dataset, features of dataset and types of attributes.
* Database, Data-Warehouse and Data-Marts.

**Theory :**

**1. Data-set and it’s features :**

* **Dataset:** Dataset is a collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer. In other words, data set is a combination of data having common features
* **Features of Dataset :** A [**feature**](https://www.datarobot.com/wiki/feature/) is a measurable property of the object you’re trying to analyze. In datasets, features appear as columns.
* **Attributes :** An attribute is a data field, representing a characteristic or feature of a data object. The nouns attribute, dimension, feature, and variable are often used interchangeably in the literature. The term dimension is commonly used in data warehousing. Machine learning literature tends to use the term feature, while statisticians prefer the term variable. Data mining and database professionals commonly use the term attribute. The distribution of data involving one attribute (or variable) is called univariate. A bivariate distribution involves two attributes, and so on.
* **Types of Attributes :**
* **Nominal Attributes :** Nominal means “relating to names.” The values of a nominal attribute are symbols or names of things. Each value represents some kind of category, code, or state, and so nominal attributes are also referred to as categorical. The values do not have any meaningful order
* **Binary Attributes :** A binary attribute is a nominal attribute with only two categories or states: 0 or 1, where 0 typically means that the attribute is absent, and 1 means that it is present. Binary attributes are referred to as Boolean if the two states correspond to true and false. A binary attribute is symmetric if both of its states are equally valuable and carry the same weight; that is, there is no preference on which outcome should be coded as 0 or 1. Simiarly a binary attribute is asymmetric if the outcomes of the states are not equally important.
* . **Ordinal Attributes :** An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive values is not known.
* **Numeric Attributes :** A numeric attribute is quantitative; that is, it is a measurable quantity, represented in integer or real values. Numeric attributes can be interval-scaled or ratio-scaled.
* **Attribute vector :** A set of attributes used to describe a given object is called an attribute vector (or feature vector).

**2. Database, Data-Warehouse and Data-Marts :**

* **Database :** Database is a systematic collection of interrelated data that represents the current status of the stored data. There could be multiple heterogeneous databases where the schema of one database may not agree with the schema of another. A database system supports ad-hoc query and on-line transaction processing.
* **Data-Warehouse :** A data warehouse is a repository of information collected from multiple sources, over a history of time, stored under a uniﬁed schema, and used for data analysis and decision support.
* **Data-Marts :** A data mart is focused on a single functional area of an organization and contains a subset of data stored in a Data Warehouse.A data mart is a condensed version of Data Warehouse and is designed for use by a specific department, unit or set of users in an organization. E.g., Marketing, Sales, HR or finance. It is often controlled by a single department in an organization.Data Mart usually draws data from only a few sources compared to a Data warehouse. Data marts are small in size and are more flexible compared to a Datawarehouse.

**EXPERIMENT : 3**

**Aim :** To study various distance measures (Euclidean distance, Weighted Euclidean distance, Manhattan distance) and implimenting them in a programming language.

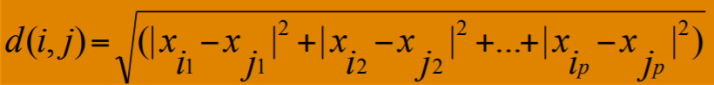
**Reqiurements :**Sublime Text editor, GCC

**Theory :**

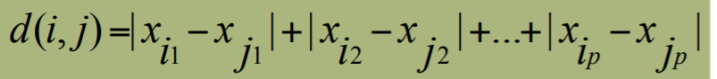
we describe distance measures that are commonly used for computing the dissimilarity of objects described by numeric attributes. In some cases, the data are normalized before applying distance calculations. This involves transforming the data to fall within a smaller or common range, such as [−1, 1] or [0.0, 1.0].

These measures include the Euclidean, Manhattan and weighted Euclidean.

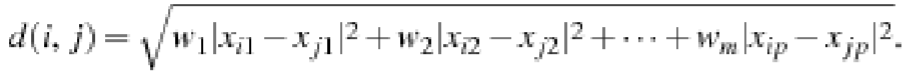
* **Euclidean distance :**The Euclidean distance between objects i and j is defined as :



* **Manhattan distance :**The Manhattan distance between objects i and j is defined as :



* **Weighted Euclidean distance :**The Weighted Euclidean distance between objects i and j is defined as :



Where w₁, w₂, w₃, ….. , wₘ are corresponding weights.

**Procedure :**

**Code to implement :** Euclidean distance, Weighted Euclidean distance, Manhattan distance

#include <bits/stdc++.h>

#include<cmath>

using namespace std;

double Euclidean(double \*x,double \*y,int n){

int i;

double dis;

for(i=0;i<n;i++){

dis+=(x[i]-y[i])\*(x[i]-y[i]);

}

return sqrt(dis);

}

double WeightedEuclidean(double \*x,double \*y,int n){

int i;

double dis;

for(i=0;i<n;i++){

cout<<"Enter the weight for "<<i+1<<"th attribute- ";

int w;

cin>>w;

dis+=w\*(x[i]-y[i])\*(x[i]-y[i]);

}

return sqrt(dis);

}

double Manhattan(double \*x,double \*y,int n){

int i;

double dis;

for(i=0;i<n;i++){

dis+=abs(x[i]-y[i]);

}

return dis;

}

int main()

{

cout << "Enter no of itemset :";

int n;cin >> n;

double x[n],y[n];

cout << "Enter data for X :";

for (int i = 0; i < n; i++)

{ cin >> x[i]; }

cout << "Enter data for Y :";

for (int i = 0; i < n; i++){

cin >> y[i];

} ;

int ch;

cout<<"\n1. Display Data\n2. Eucledean distance.\n3. Manhatten Distance\n4. Weighted Eucledean distance\n5. Press 5 to exit\n\nEnter a valid option ";

cin>>ch;

while(ch!=5){

switch(ch){

case 0:

break;

case 1:

cout<<"\n\n"<<setw(10)<<"X"<<setw(10)<<"Y";

for(int p=0;p<n;p++)

cout<<"\n"<<setw(10)<<x[p]<<setw(10)<<y[p];

cout<<"\n";

break;

case 2:

cout<<"Required distance "<<Euclidean(x,y,n)<<"\n";

break;

case 3:

cout<<"Required distance "<<Manhattan(x,y,n)<<"\n";

break;

case 4:

cout<<"Required distance "<<WeightedEuclidean(x,y,n)<<"\n";

break;

case 5:

break;

default:

cout<<"Enter a valid choice";

break;

}

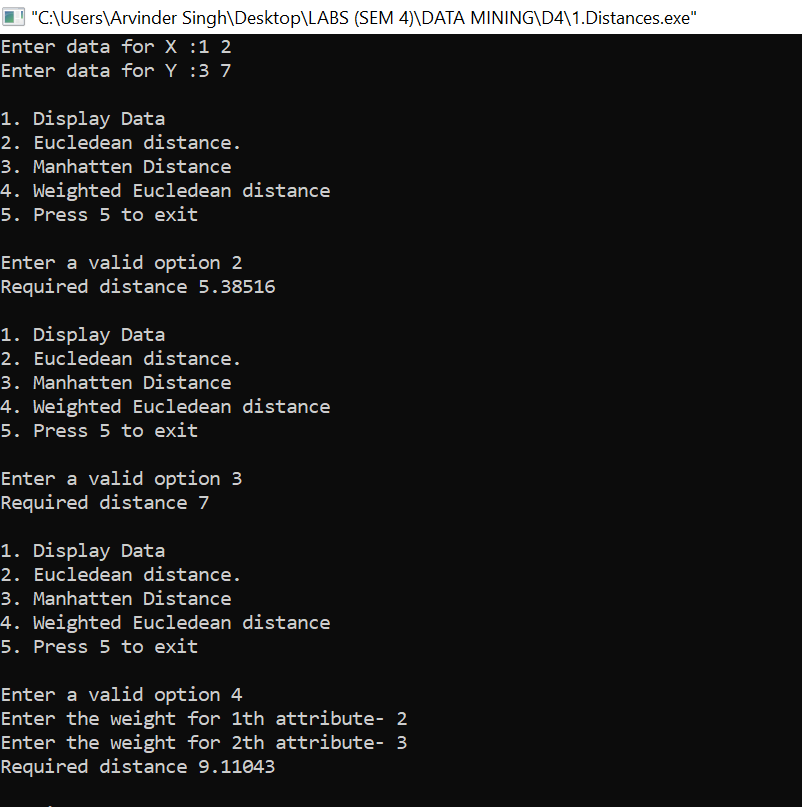
cout<<"\n1. Display Data\n2. Eucledean distance.\n3. Manhatten Distance\n4. Weighted Eucledean distance\n5. Press 5 to exit\n\nEnter a valid option ";

cin>>ch;

}

return 0;

}



**EXPERIMENT : 4**

**Aim :** To study and implement CHI-Square method in a programming language

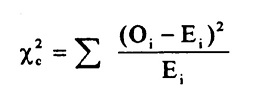
**Reqiurements :**Sublime Text editor, GCC compiler

**Theory :**

The Chi-Squared test ( also known as Pearson’s Chi-Squared Test ) is a [statistical hypothesis test](https://machinelearningmastery.com/statistical-hypothesis-tests/) that assumes (the null hypothesis) that the observed frequencies for a categorical variable match the expected frequencies for the categorical variable. The test calculates a statistic that has a chi-squared distribution, named for the Greek capital letter Chi (X) pronounced “ki” as in kite.

The Chi-Squared test does this for a contingency table, first calculating the expected frequencies for the groups, then determining whether the division of the groups, called the observed frequencies, matches the expected frequencies.

The formula for the chi-square statistic used in the chi square test is:

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/chi-square-formula.jpg)

Where

subscript “c” are the degrees of freedom

O is your observed value

E is your expected value

The result of the test is a test statistic that has a chi-squared distribution and can be interpreted to reject or fail to reject the assumption or null hypothesis that the observed and expected frequencies are the same.

The variables are considered independent if the observed and expected frequencies are similar, that the levels of the variables do not interact, are not dependent.

We can interpret the test statistic in the context of the chi-squared distribution with the requisite number of degress of freedom as follows:

* **If Statistic >= Critical Value**: significant result, reject null hypothesis (H0), dependent.
* **If Statistic < Critical Value**: not significant result, fail to reject null hypothesis (H0), independent.

The degrees of freedom for the chi-squared distribution is calculated based on the size of the contingency table as:

degrees of freedom: (rows - 1) \* (cols - 1)

In terms of a p-value and a chosen significance level (alpha), the test can be interpreted as follows:

* **If p-value <= alpha**: significant result, reject null hypothesis (H0), dependent.
* **If p-value > alpha**: not significant result, fail to reject null hypothesis (H0), independent.

For the test to be effective, at least five observations are required in each cell of the contingency table.

**Procedure :**

**Code to implement** CHI-Square method

#include <iostream>

#include <cmath>

using namespace std;

int main(){

int n;

cout<<"CHI-Square analysis for two Variables\n\n";

cout<<"Enter the no. of attributes\n";

cin>>n;

double arr[n+1][3];

cout<<"Enter the observed values\n";

int i;

for(i=0;i<n;i++){

cin>>arr[i][0]>>arr[i][1];

arr[n][0]+=arr[i][0];

arr[n][1]+=arr[i][1];

arr[i][2]+=arr[i][0]+arr[i][1];

}

arr[n][2]=arr[n][0]+arr[n][1];

double exp[n][2];

for(i=0;i<n;i++){

exp[i][0]=arr[i][2]\*arr[n][0]/arr[n][2];

exp[i][1]=arr[i][2]\*arr[n][1]/arr[n][2];

}

double chi\_square=0;

for(i=0;i<n;i++){

chi\_square+=((arr[i][0]-exp[i][0]))\*((arr[i][0]-exp[i][0]))/exp[i][0];

chi\_square+=((arr[i][1]-exp[i][1]))\*((arr[i][1]-exp[i][1]))/exp[i][1];

}

cout<<"Enter the tabular value\n";

double t;

cin>>t;

cout<<"Value of CHI Square is: "<<chi\_square<<"\n";

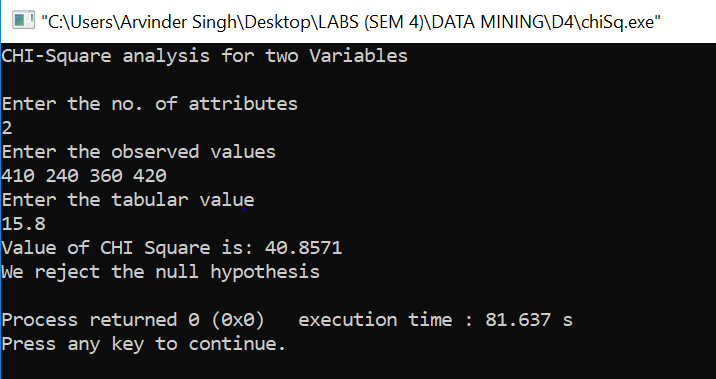
if(t<chi\_square)

cout<<"We reject the null hypothesis\n";

else cout<<"The null hypothesis is correct\n";

return 0;

}



**EXPERIMENT : 5**

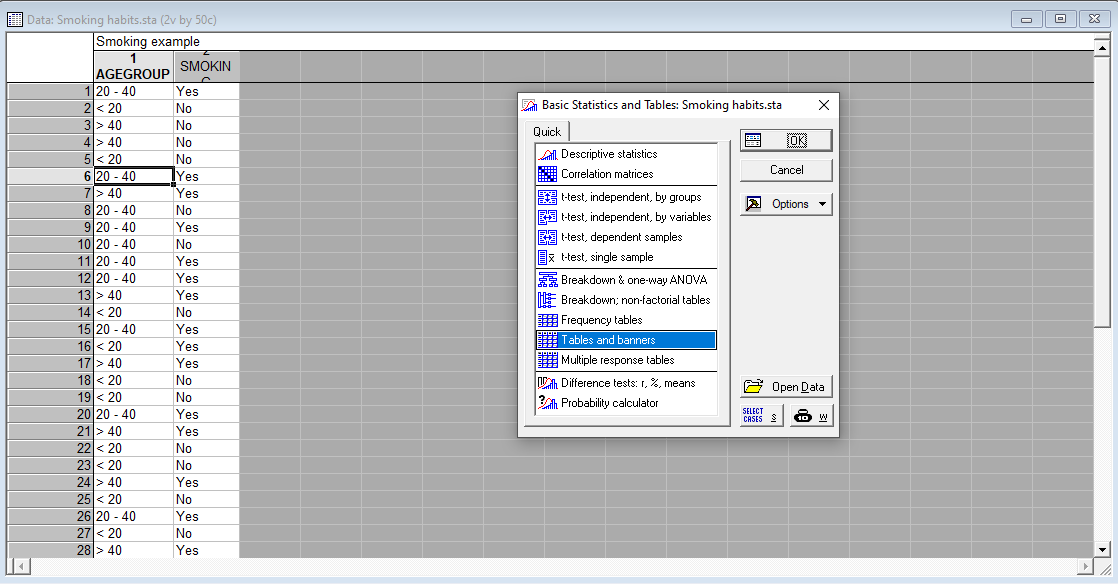
**Aim :**To perform CHI-Square method on Smoking Habits dataset (SmokingHabts.sta) in Statistica.

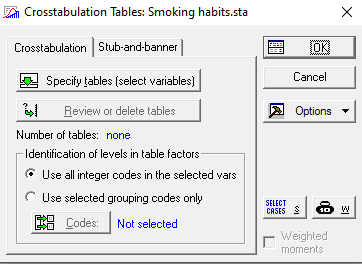
**Reqiurements :** Statistica, SmokingHabts.sta data set (SmokingHabts.sta).

**Procedure :**

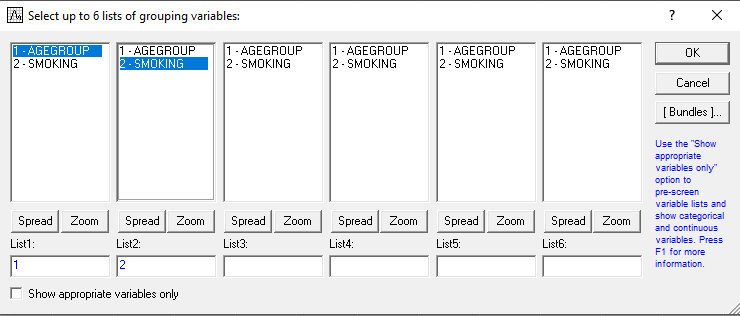
**Implimentation of** CHI-Square method on Statistca :

1. Select the dataset and then, tables/banners option as shown:-

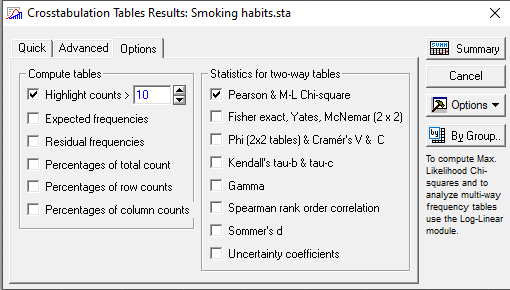


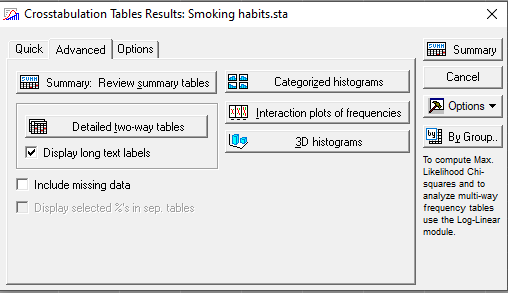


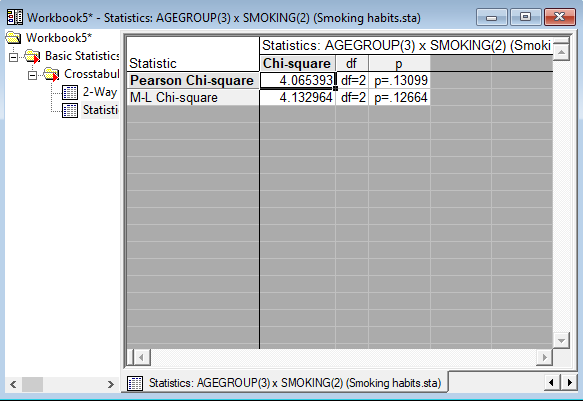
1. Select the list one and list two variables:-



1. Select Pearson & M-L Chi-square:-







Result

**EXPERIMENT : 6**

**Aim :** To study and implement Feature selection in Statistica.

**Reqiurements :** Statistica, 30 best predictors of Yield (selected from 1950 predictors) data set (30PredictorsOfYield.sta) .

**Theory :**

Feature selection refers to the process of reducing the inputs for processing and analysis, or of finding the most meaningful inputs. A related term, feature engineering (or feature extraction), refers to the process of extracting useful information or features from existing data.

Feature selection is critical to building a good model for several reasons. One is that feature selection implies some degree of cardinality reduction, to impose a cutoff on the number of attributes that can be considered when building a model.

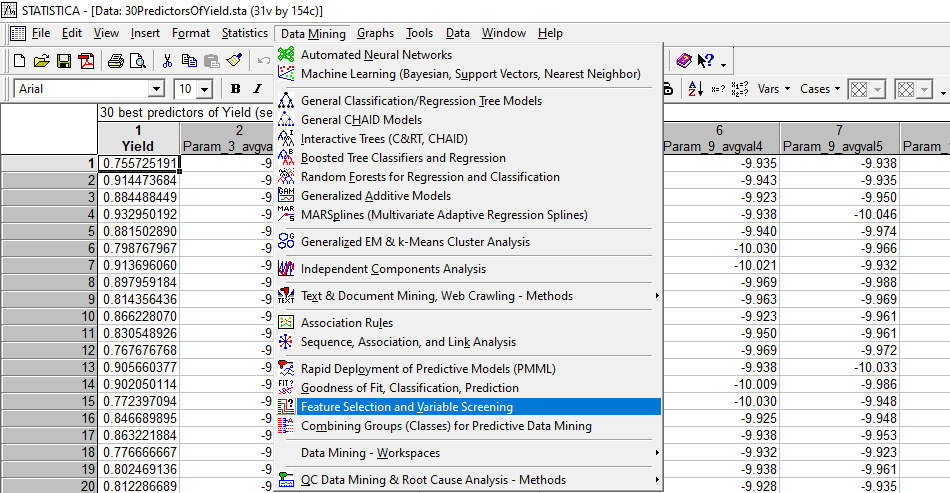
Even if resources were not an issue, you would still want to perform feature selection and identify the best columns, because unneeded columns can degrade the quality of the model in several ways:

* Noisy or redundant data makes it more difficult to discover meaningful patterns.
* If the data set is high-dimensional, most data mining algorithms require a much larger training data set.

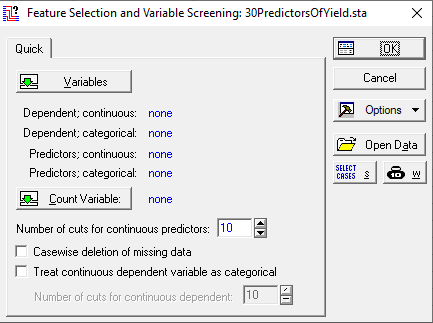
During the process of feature selection, either the analyst or the modeling tool or algorithm actively selects or discards attributes based on their usefulness for analysis.

**Procedure :**

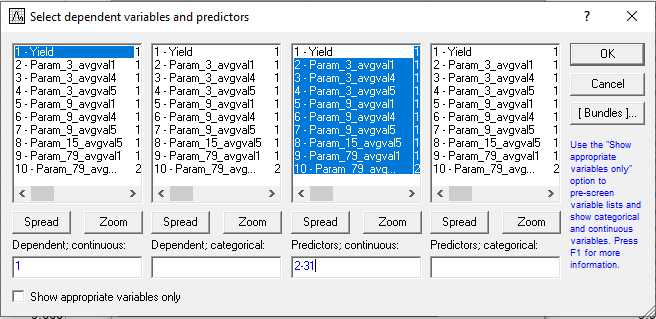
Step 1. Choose Feature Selection and Variable Screening option from the Data Mining drop down button.



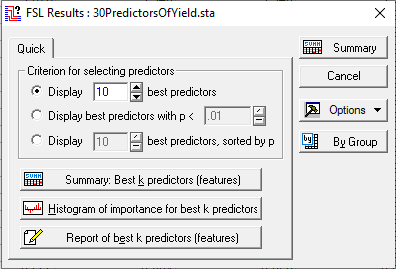
Step 2. Click on the variables button.



Step 3. Select dependent and predictor variables as shown.



Step 4. Select the type of output needed (summary, histogram, etc).



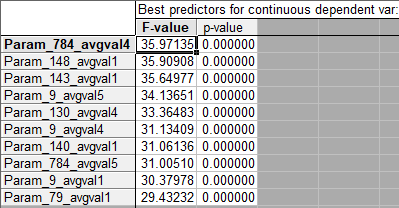
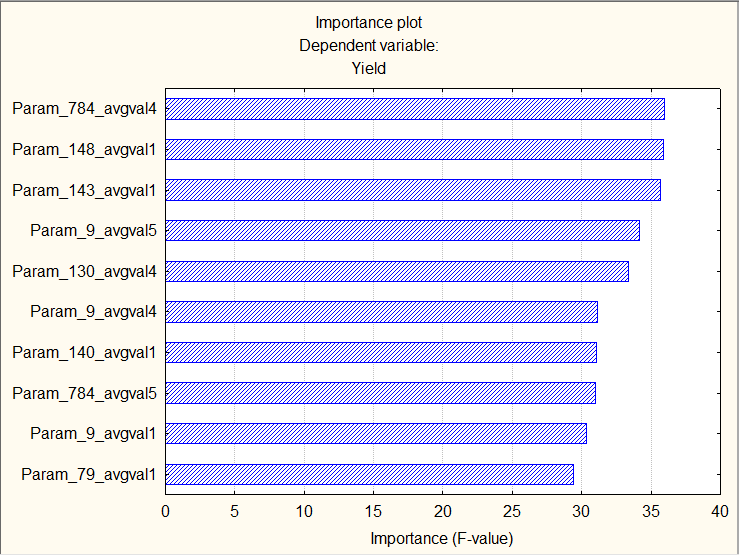


Table containing the 10 best predictors



Histogram of 10 best predictors along with their F-value

**EXPERIMENT : 6**

**Aim :** To study and implement Aprirori in Statistica.

**Reqiurements :** Statistica, 30 best predictors of Yield (selected from 1950 predictors) data set (30PredictorsOfYield.sta) .

**Theory :**

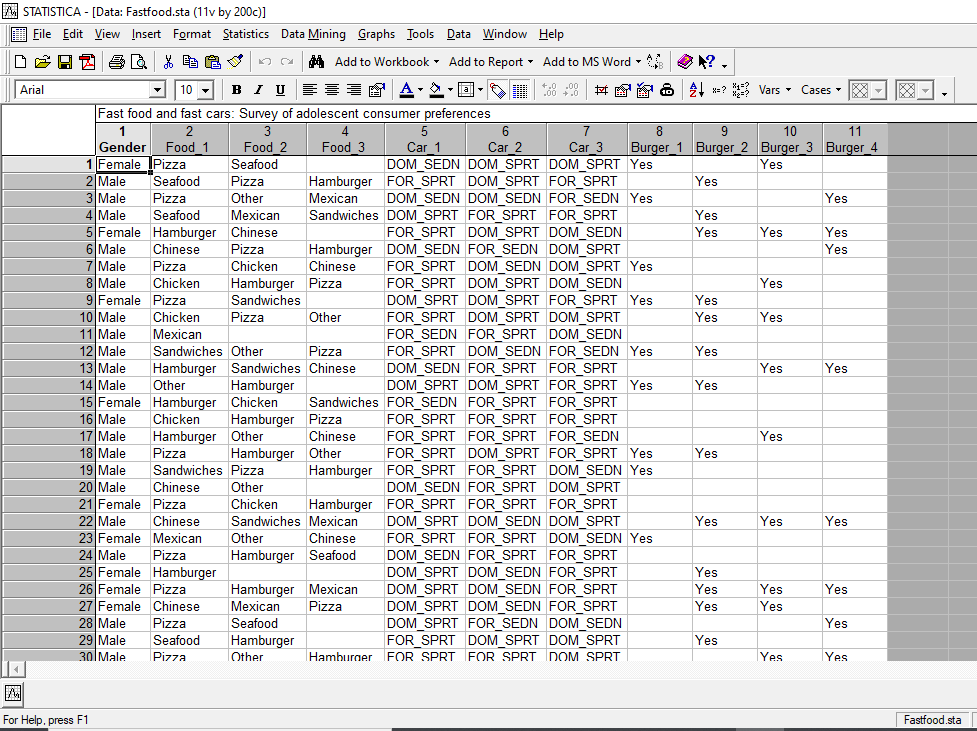
**Apriori algorithm** : Name of the algorithm is Apriori because it uses prior knowledge of frequent itemset properties. We apply an iterative approach or level-wise search where k-frequent itemsets are used to find k+1 itemsets.

To improve the efficiency of level-wise generation of frequent itemsets, an important property is used called *Apriori property* which helps by reducing the search space.

**Apriori Property :***All subsets of a frequent itemset must be frequent(Apriori propertry).  
If an itemset is infrequent, all its supersets will be infrequent.*

**Procedure :**

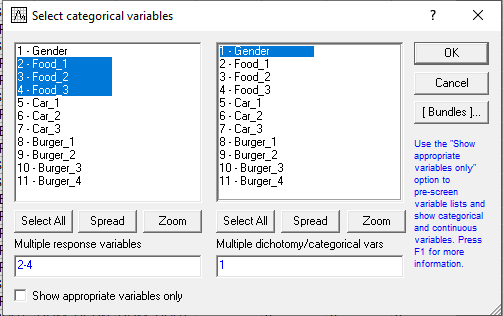
1. Open Statistica and go on File-Open Examples-Datasets and open the Fastfood.sta dataset.



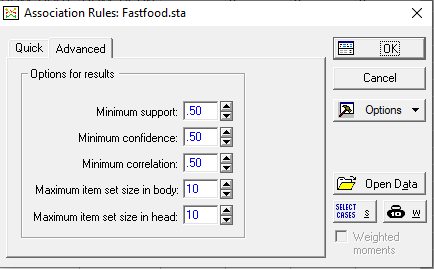
1. Go to Data mining -Association rules. You get a dialog box as shown below



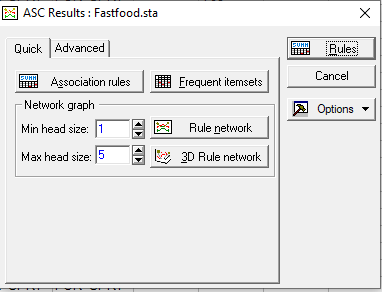
1. Click on Variables and select the required variables from the dialog box.



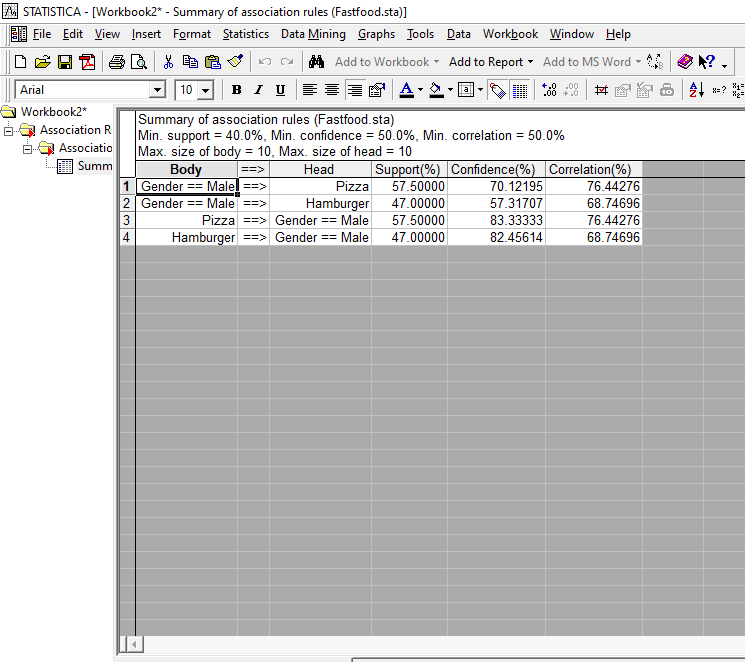
1. Click on OK and go to Advanced option of the previous dialog box.



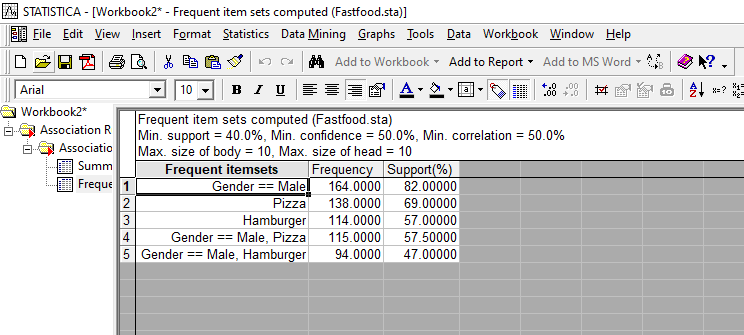
1. Set the Minimum support and Minimum confidence as you require and click on OK. You get a dialog box as shown.



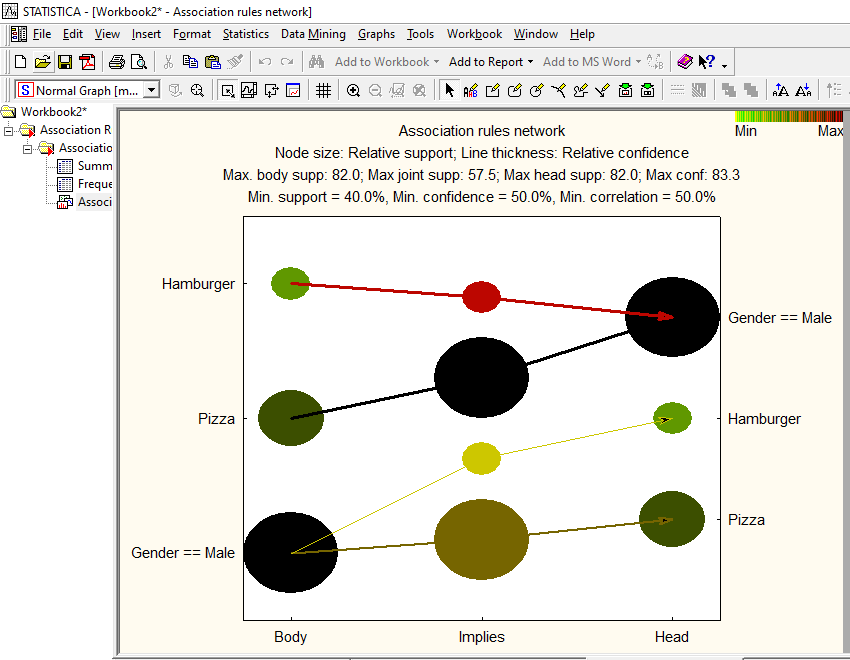
1. Click on Association Rules to get the shown details.



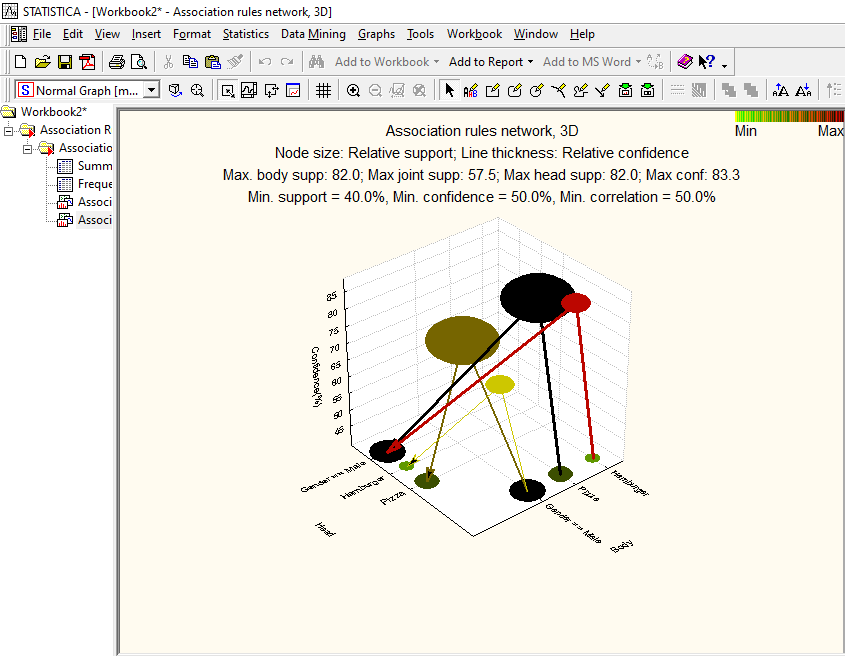
1. Go on Data mining-Association rules and click continue current. Now click on The Frequent itemset option.



1. Go on Data mining-Association rules and click continue current. Now click on Rule Network.



1. Go on Data mining -Association rules and click continue current. Now click on 3D Rule Network.



Font : has a chi-squared distribution, named for the Greek capital letter Chi (X) pronounced “ki” as in kite.