Third Year B. Tech., Sem VI 2021-22

Software Engineering & Tools Lab

Assignment / Journal submission

PNR/ Roll No:2020BTECS00203

Full name: Tejal Satish Pansare

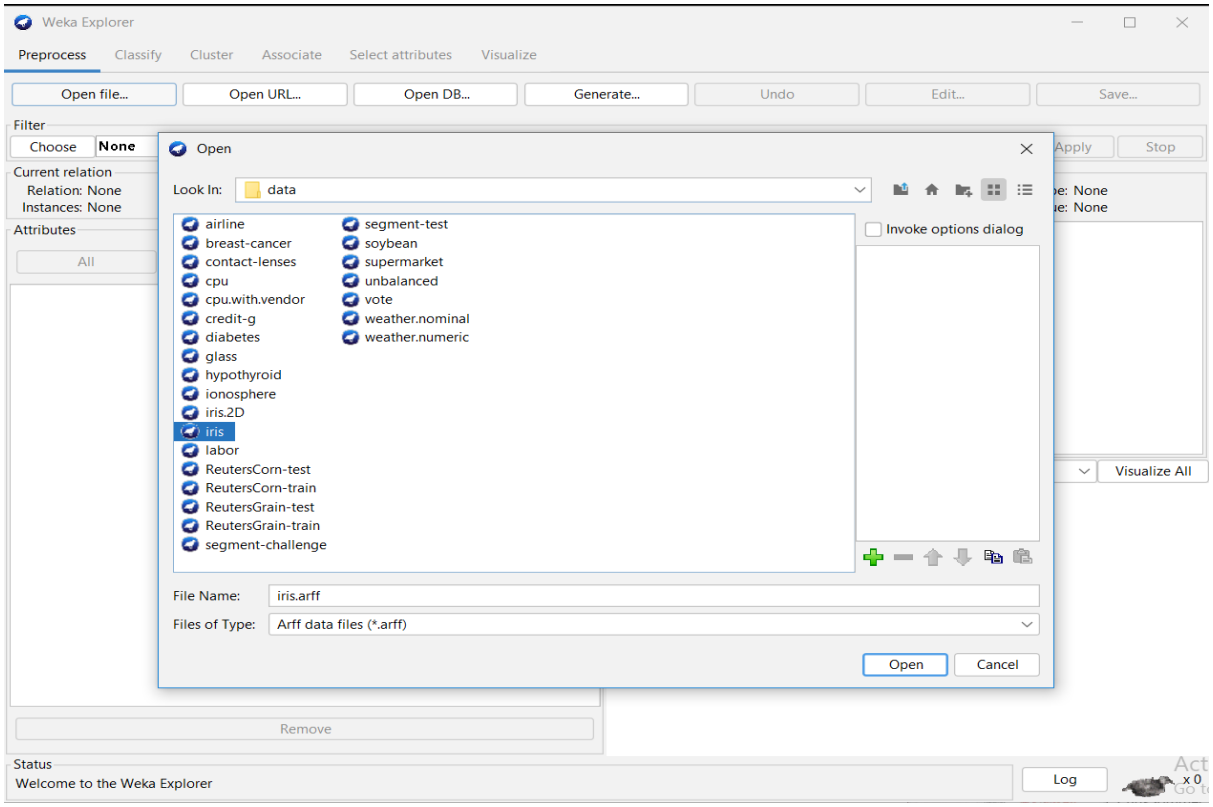
Batch: T8

Assignment: 1

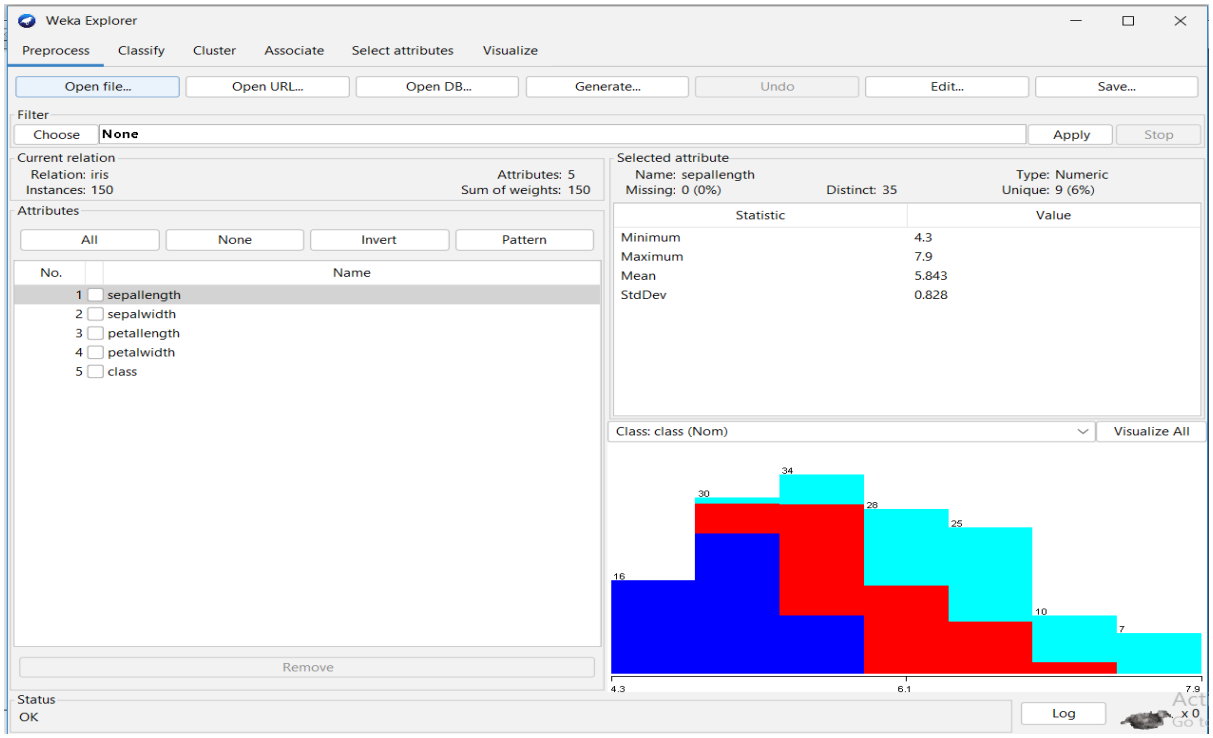
Title of assignment: Introduction to OSS

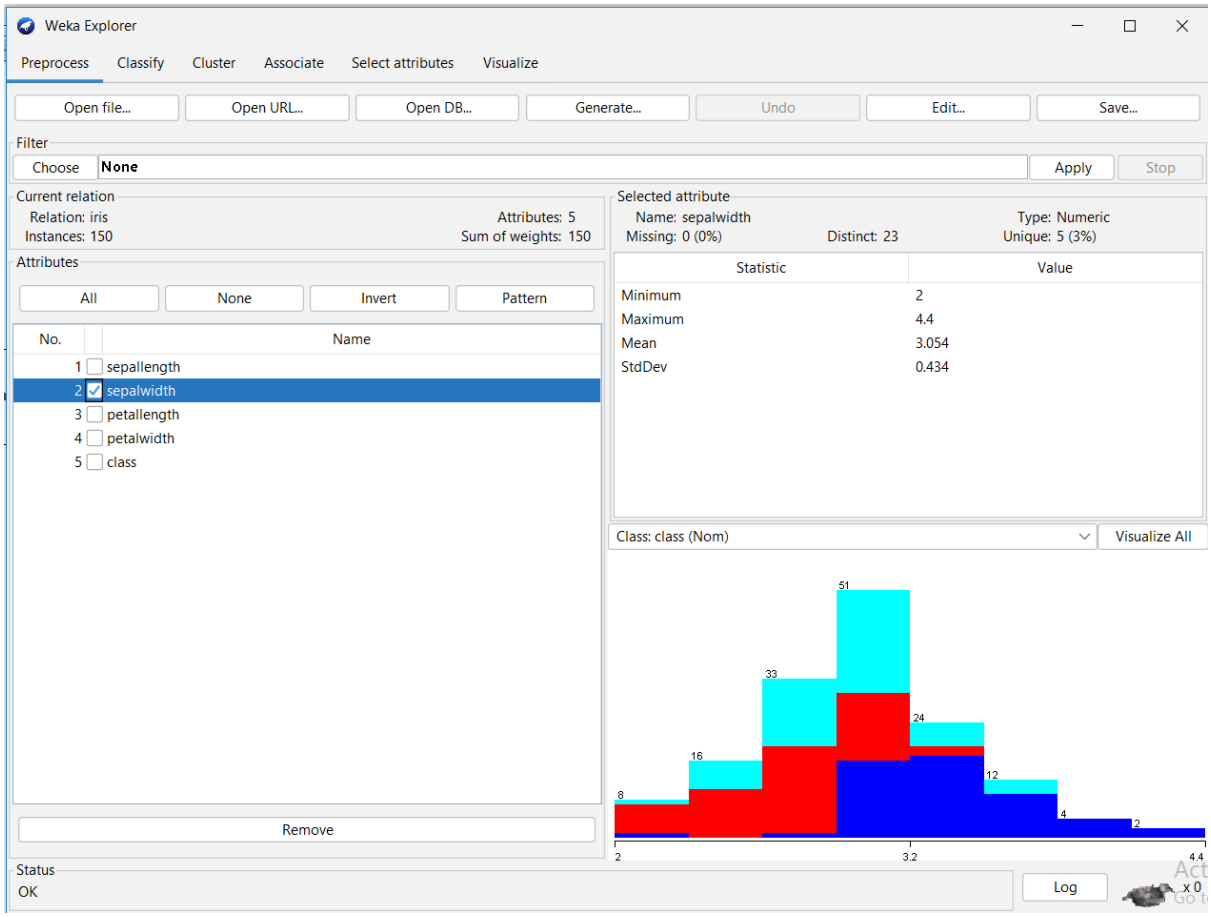
---

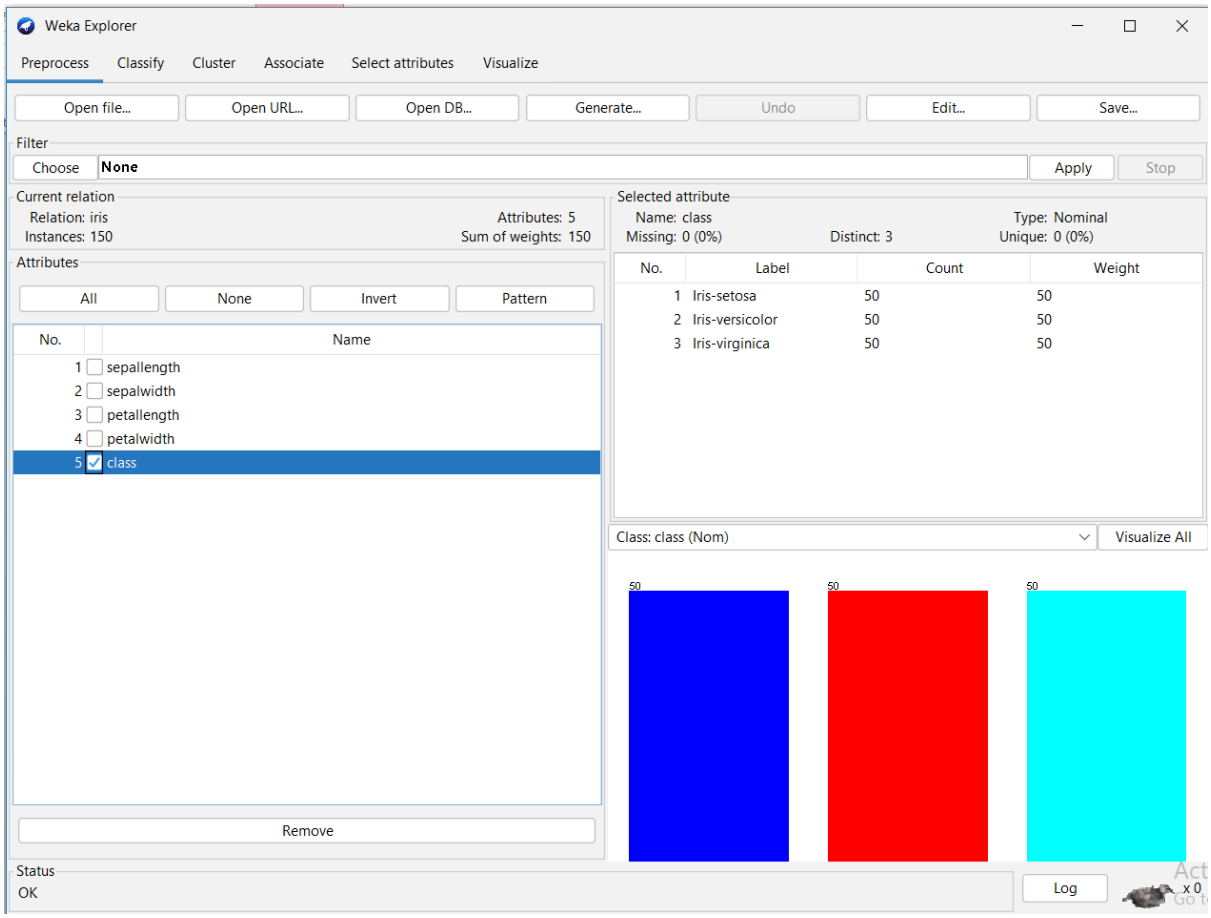
1. Weka is a GUI workbench that empowers data wranglers to assemble machine learning pipelines, train models, and run predictions without having to write code.
Using Weka tool perform below tasks such as data preprocessing, data classification (use any appropriate ML algorithm) and data visualization efficiently on given dataset.

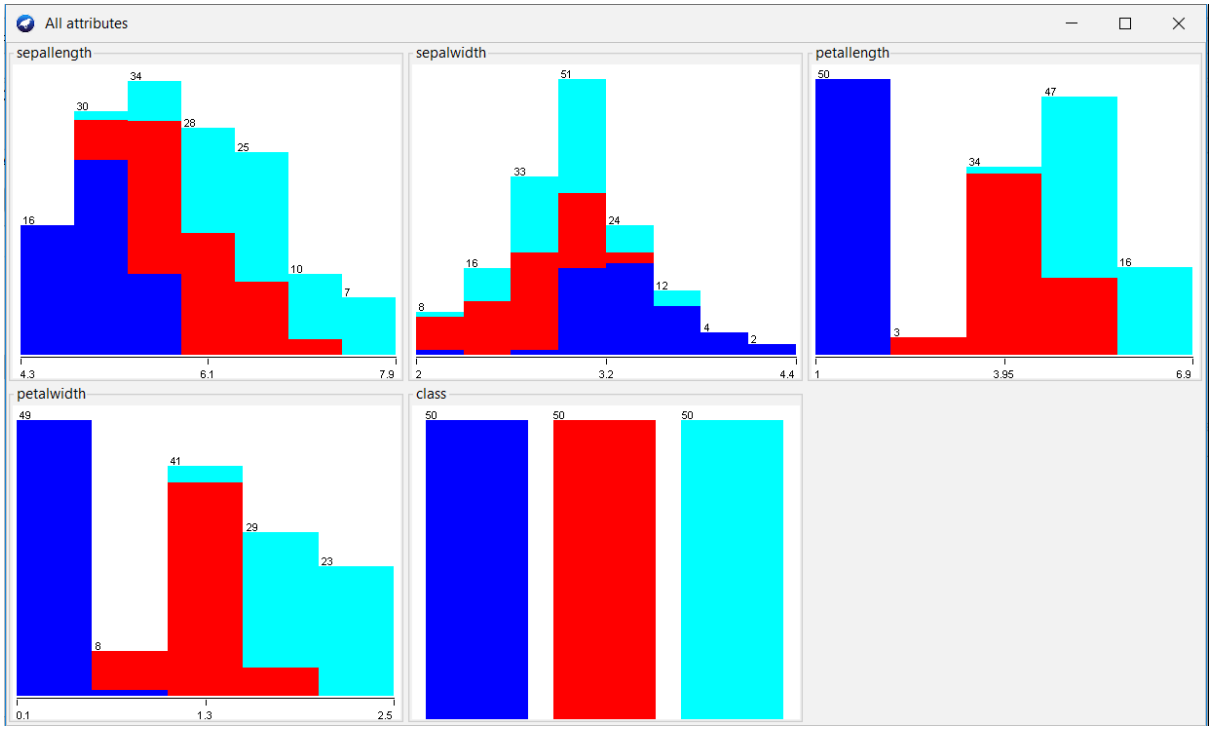


1) Processing of Data




Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: iris
Instances: 150
Attributes: 5
Sum of weights: 150
Attributes
All
None
Invert
Pattern
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Remove
Selected attribute
Name: sepalwidth
Missing: 0 (0%)
Distinct: 23
Type: Numeric
Unique: 5 (3%)
Statistic
Value
Minimum 2
Maximum 4.4
Mean 3.054
StdDev 0.434
Class: class (Nom)
Visualize All
51
33
24
16
12
8
4
2
2
3.2
4.4
Status
OK
Log
x 0

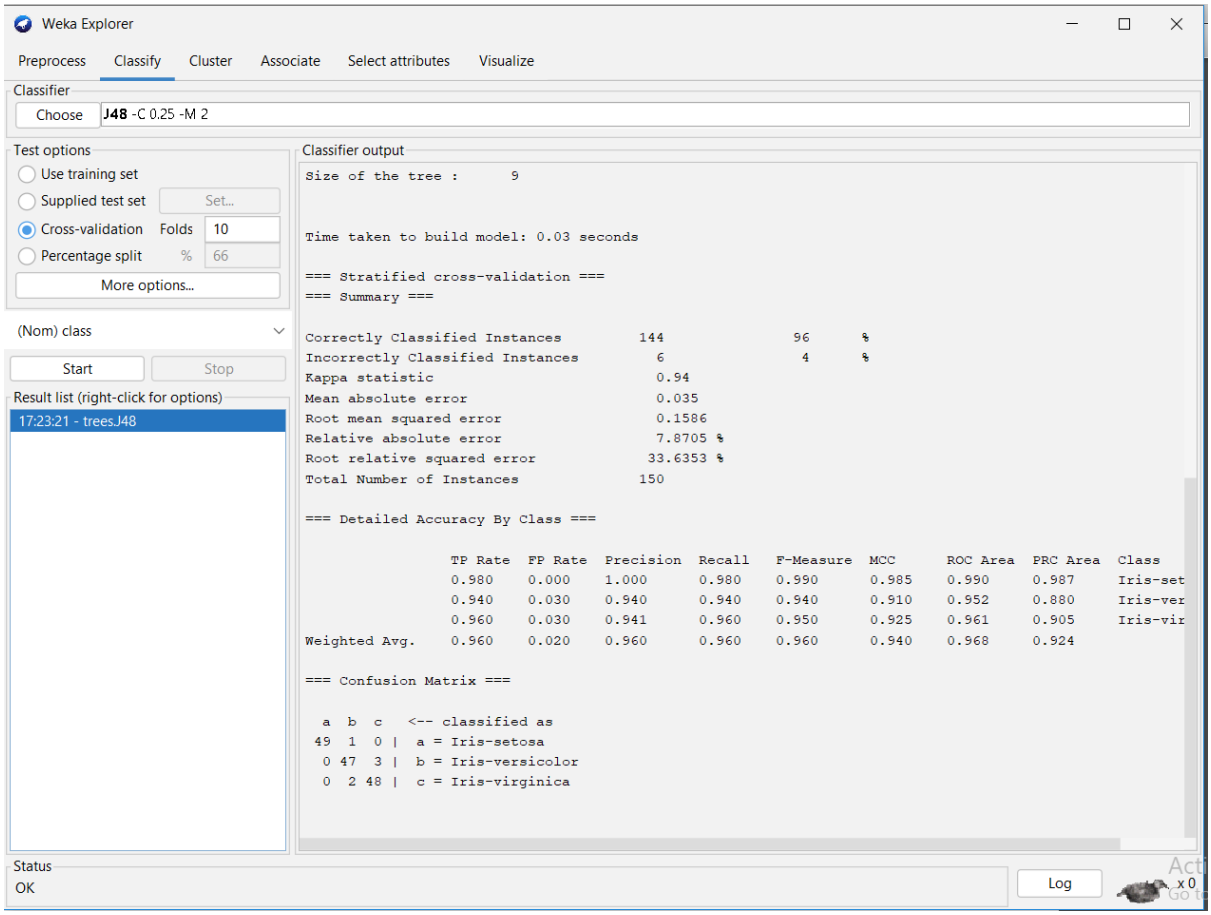

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: iris
Instances: 150
Attributes: 5
Sum of weights: 150
Attributes
All
None
Invert
Pattern
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Remove
Selected attribute
Name: class
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)
No.
Label
Count
Weight
1 Iris-setosa 50 50
2 Iris-versicolor 50 50
3 Iris-virginica 50 50
Class: class (Nom)
Visualize All
50
50
50
Status
OK
Log
x 0



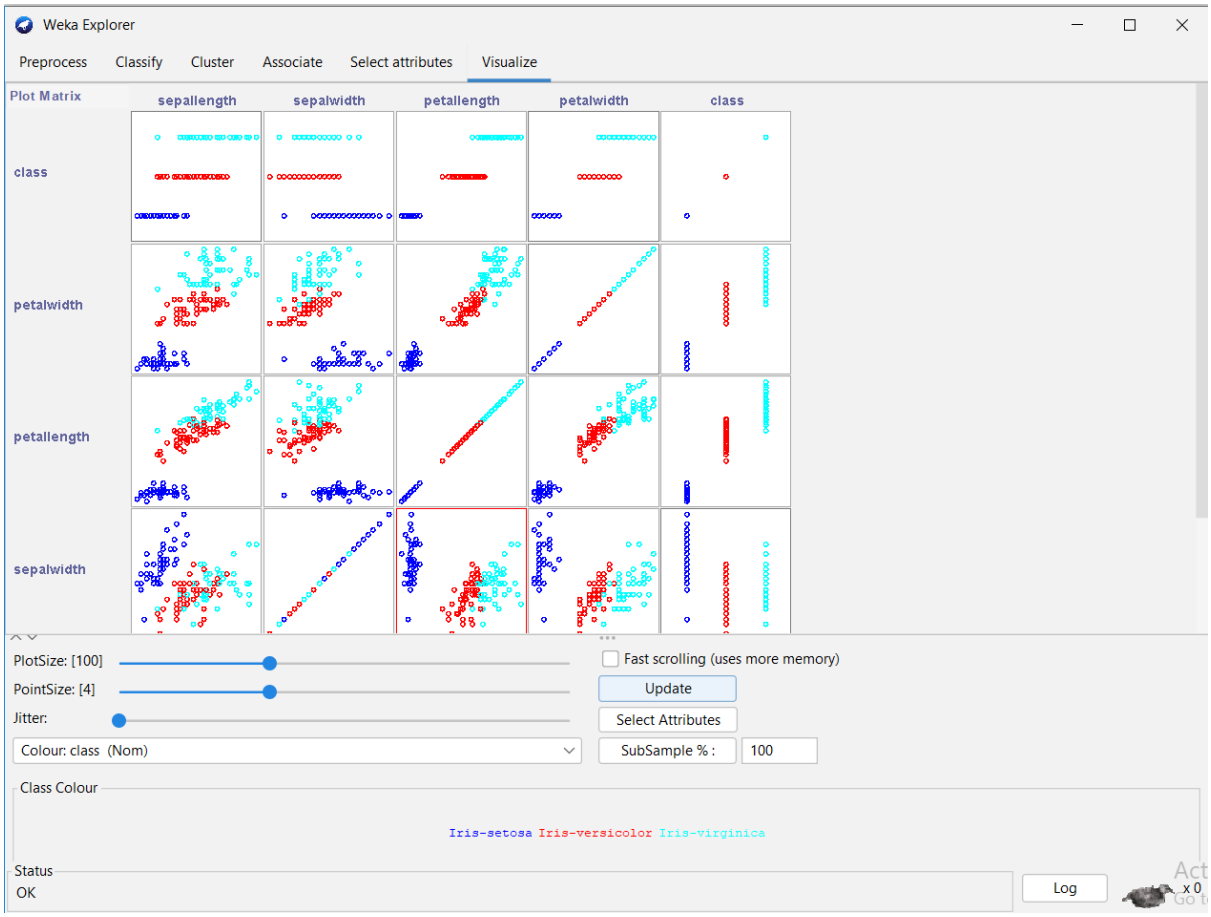
From basic visualisation of all 5 parameters, we observe that **petallength** gives a relatively clear separation of classes. Hence we take petallength as our parameter. We will be using **Linear Regression,** so we need only one parameter.

2) Data Classification.

We make our model based on **Linear Regression** and get following analysis of the model:

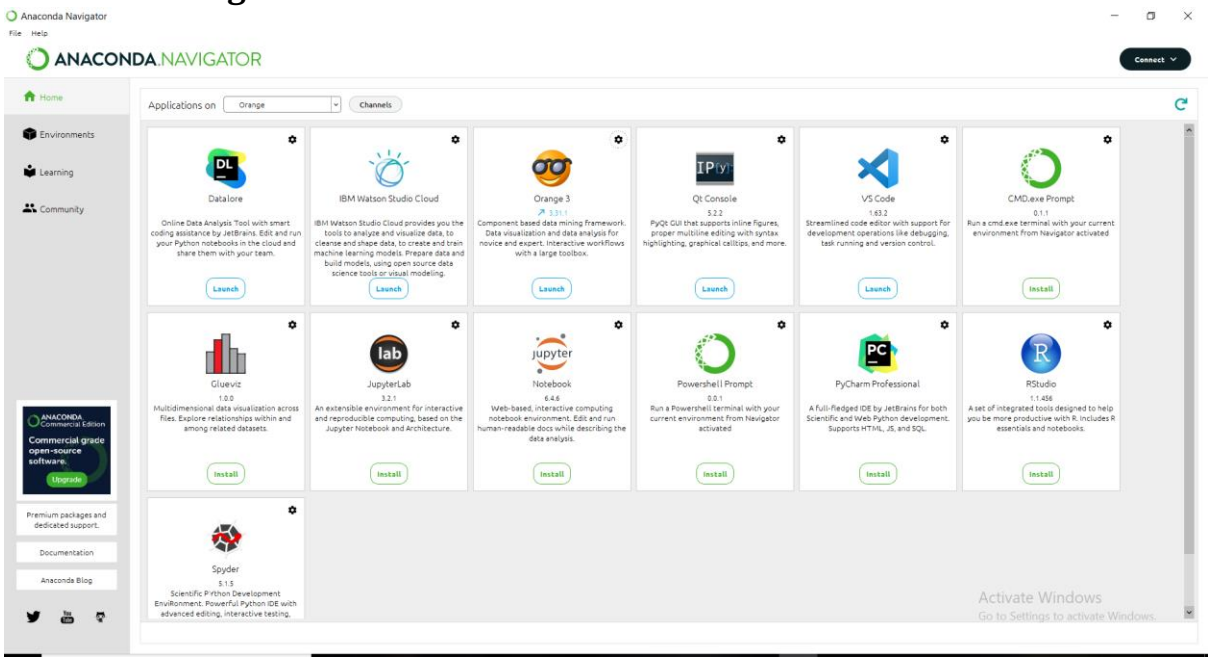


3) Data Visualization

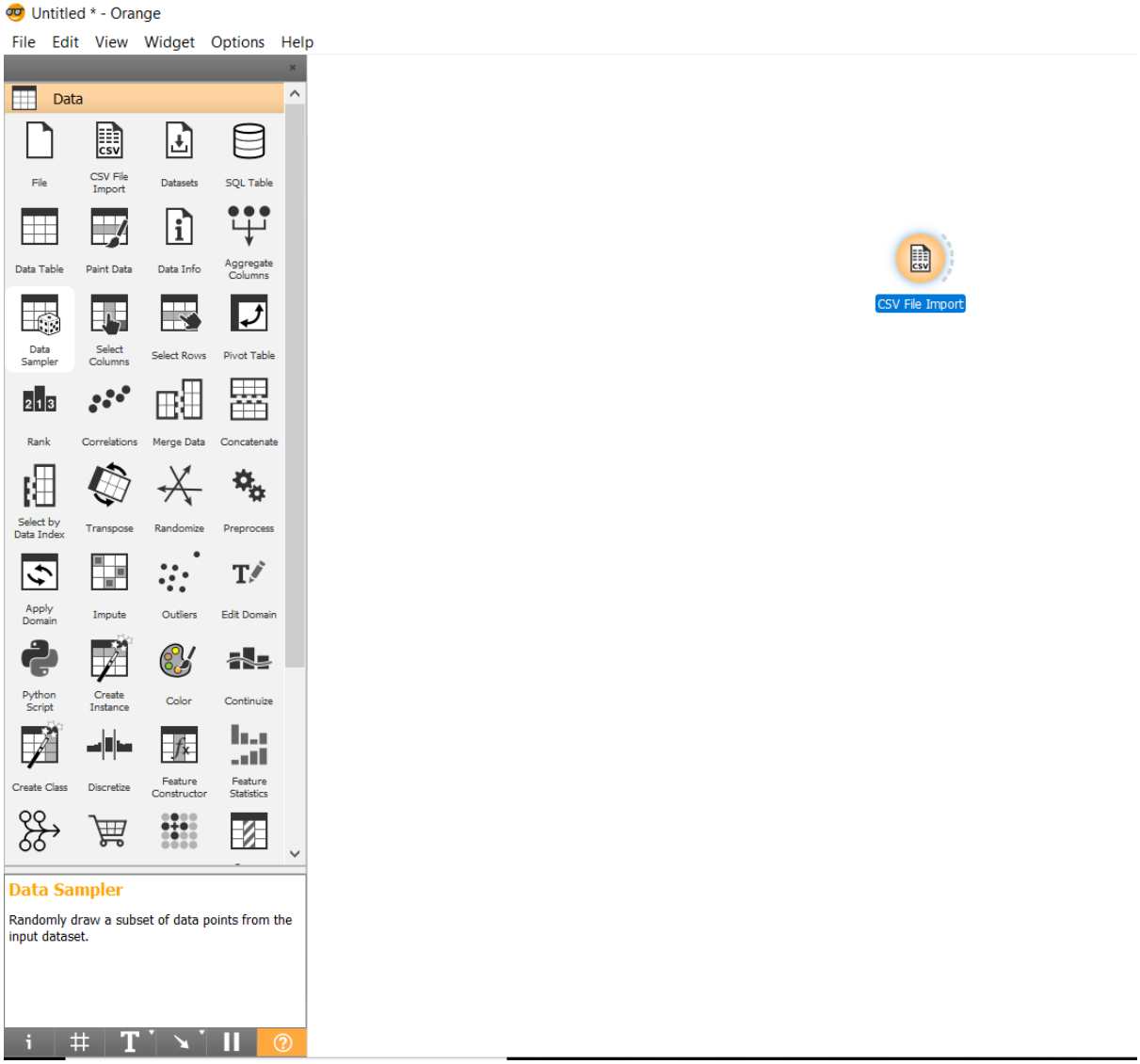


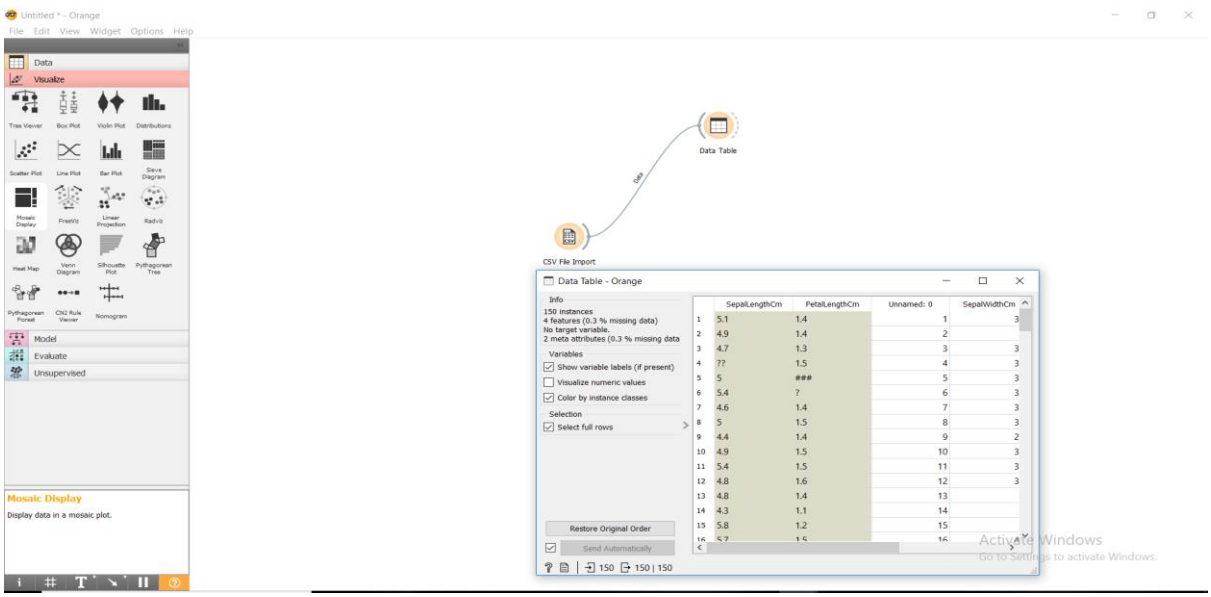
2. Orange is an easy to use data visualization tool with a large toolkit. In spite of being a GUI-based beginner-friendly tool, you mustn't mistake it for a light-weight one. It can do statistical distributions and box plots as well as decision trees, hierarchical clustering and linear projections.

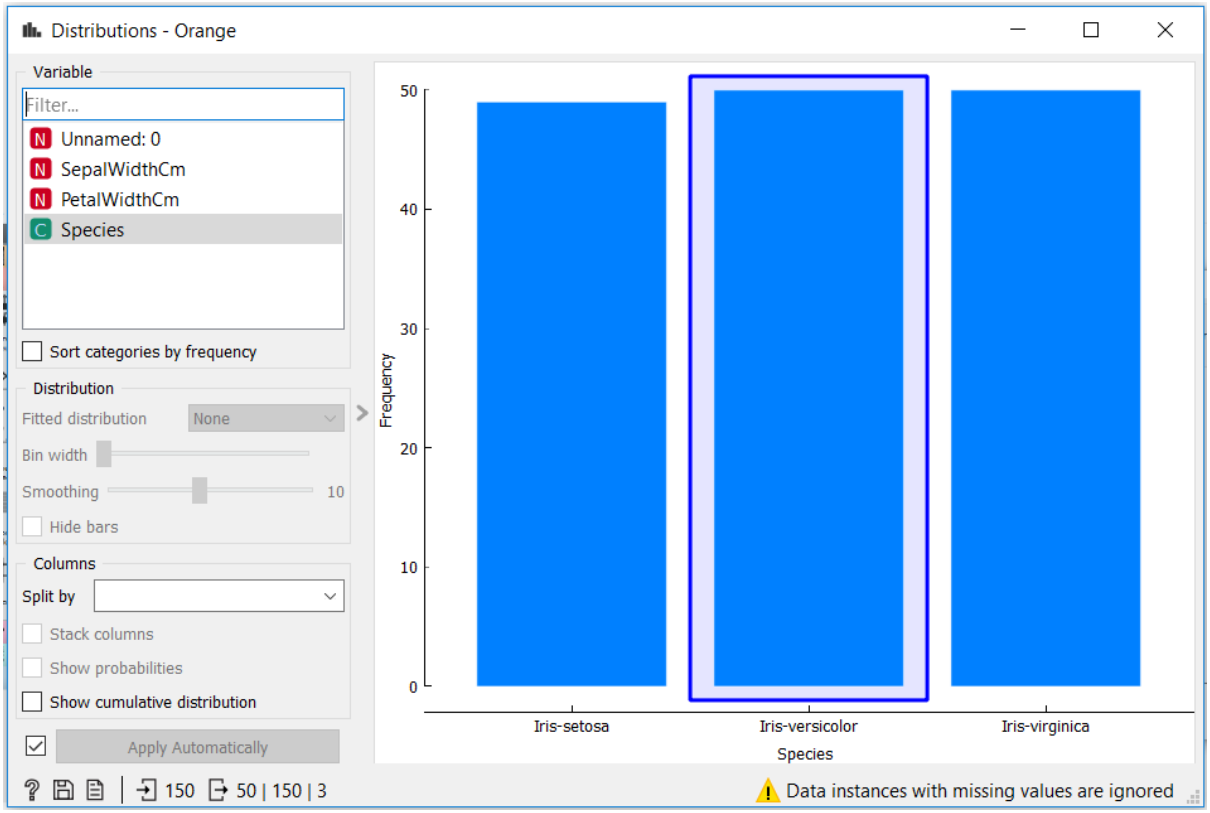
a. Install orange



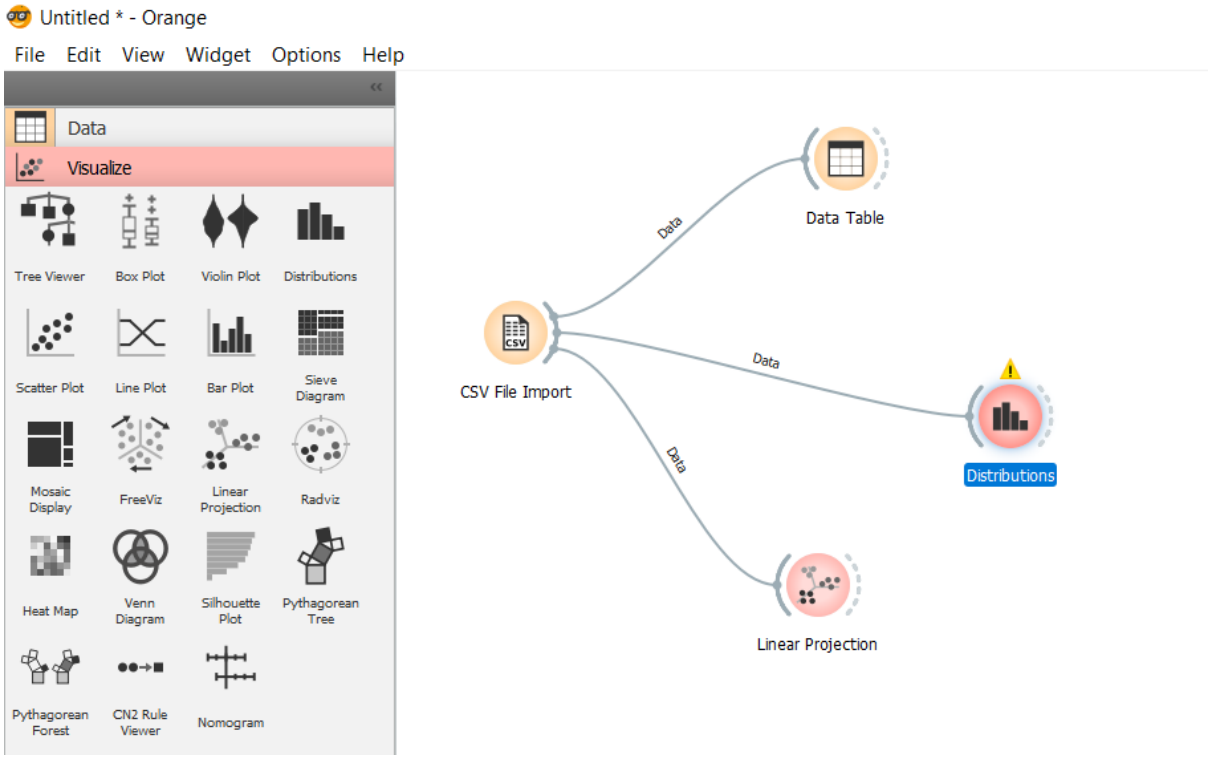
b. Show data distribution

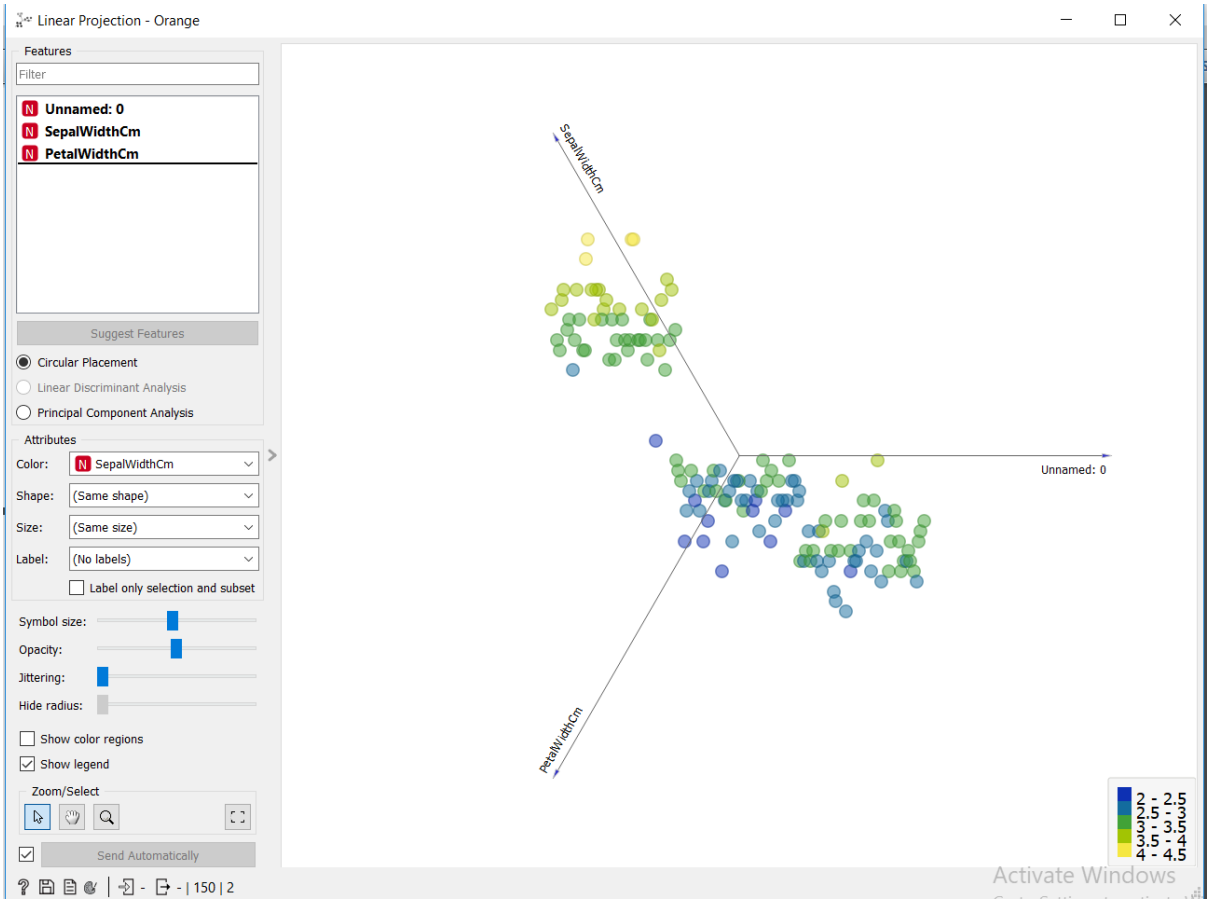






c. Show linear projection





3. Differentiate in between free software, Open source software and proprietary software with respect to its properties.

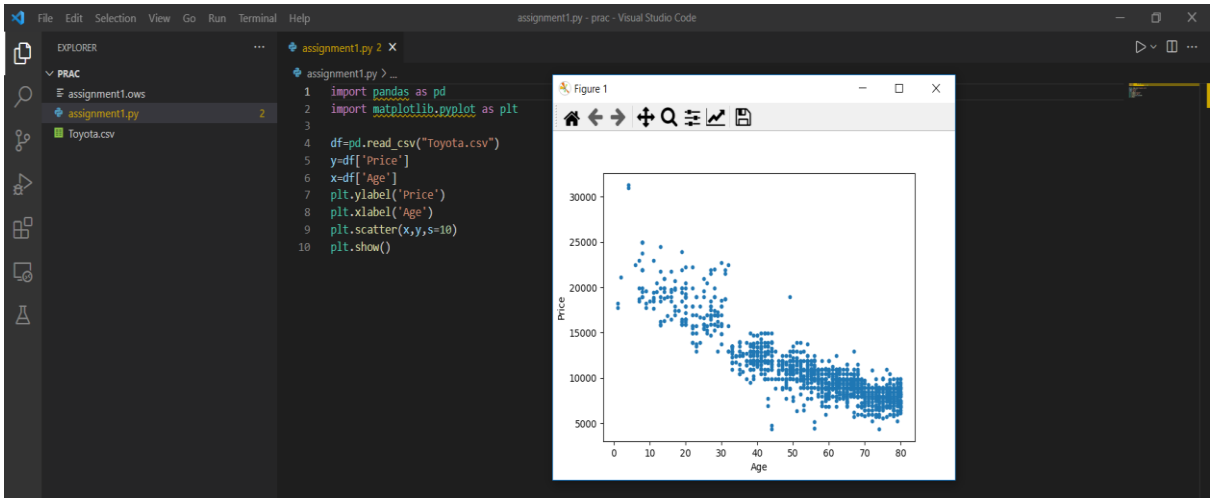
Open Source software	**Proprietary Software**
Open source software is a computer software whose source code is available openly in internet and programmers can modify it to add new features and capabilities without any cost.	Proprietary software is a computer software where the source codes are not publicly not available only the company which has created can modify it.
Here the software is developed and tested through open collaboration.	Here the software is developed and tested by the individual or organization by which it is owned not by public.
In open source software the source code is public.	In proprietary software the source code is protected
Open source software can be installed into any computer.	Proprietary software can be installed into any computer without valid license.

Users do not need to have any authenticated license to use this software.	Users need to have a valid and authenticated license to use this software.
Users can get open software for free of charge.	Users must have to pay to get the proprietary software.
In open source software faster fixes of bugs and better security is availed due to the community.	In proprietary software the vendor is completely responsible for fixing of malfunctions.

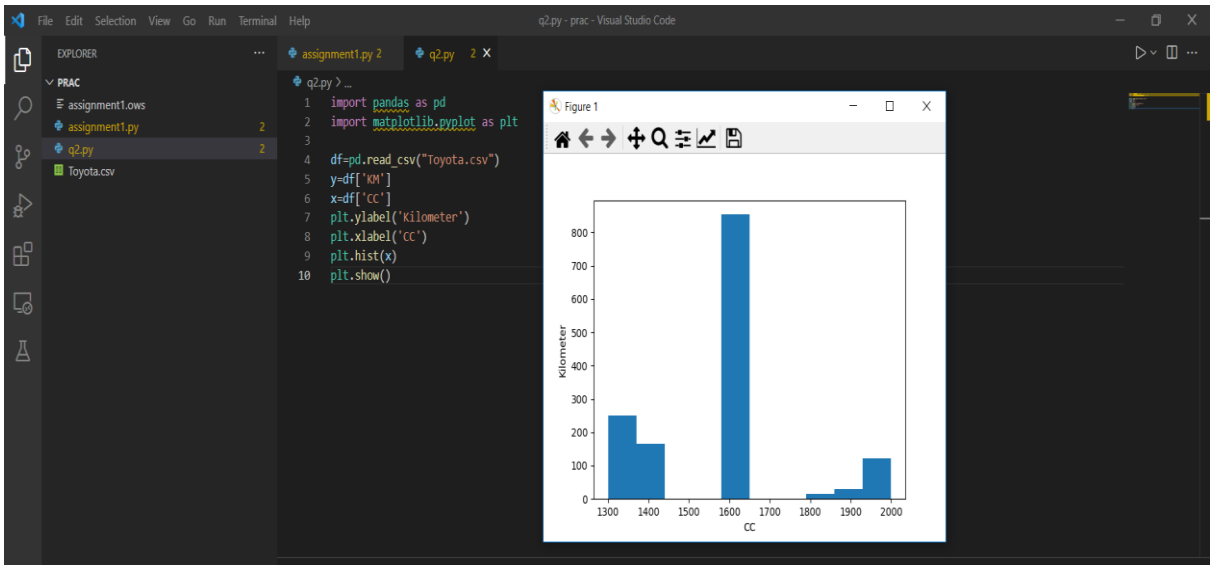
4. Using Anaconda Python create Histogram, Scatter plot and Bar plot for the dataset given below.

Dataset-
https://drive.google.com/file/d/1i11BZFe8Xj9kNq7eeE9KOa_Iz1KhEdXJ/view

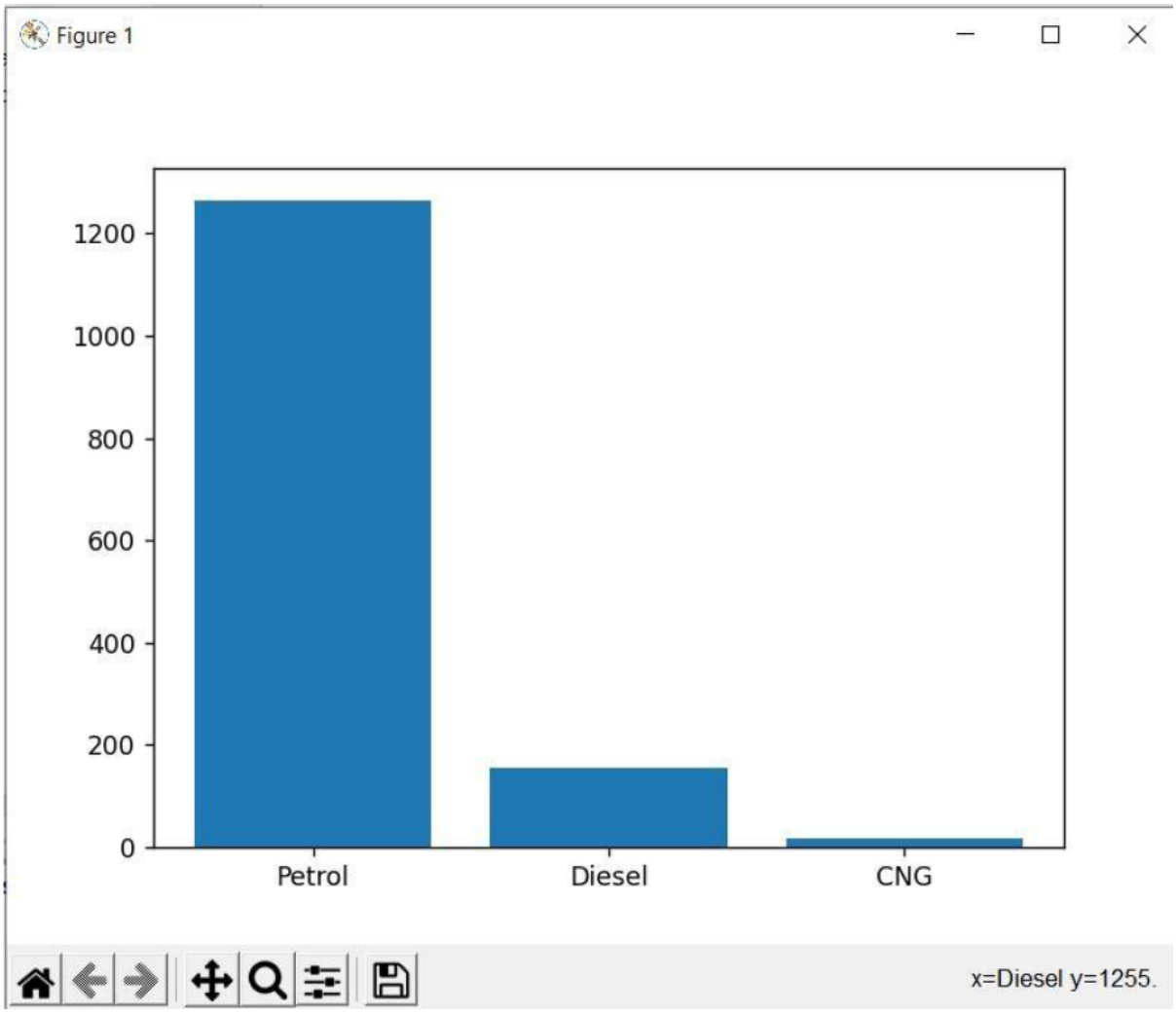
a. Scatter plot- Scatter plot of Price Vs Age



b. Histogram- for Kilometer and CC



c. Bar plot- Bar plot for different fuel types



5. Enlist some examples along with its purpose and properties (at least 10) of FOSS and proprietary software with respect to database.

1. FOSS

Examples : MySQL, SQLite, MongoDB

a) MySQL

1. **Relational Database Management System (RDBMS):** MySQL is a relational database management system. This database language is based on the SQL queries to access and manage the records of the table.
2. **Easy to use:** MySQL is easy to use. We have to get only the basic knowledge of SQL. We can build and interact with MySQL by using only a few simple SQL statements.
3. **It is secure:** MySQL consists of a solid data security layer that protects sensitive data from intruders. Also, passwords are encrypted in MySQL.

4. **Client/ Server Architecture:** MySQL follows the working of a client/server architecture. There is a database server (MySQL) and arbitrarily many clients (application programs), which communicate with the server; that is, they can query data, save changes, etc.
5. **Free to download:** MySQL is free to use so that we can download it from MySQL official website without any cost.
6. **It is scalable :**MySQL supports multi-threading that makes it easily scalable. It can handle almost any amount of data, up to as much as 50 million rows or more. The default file size limit is about 4 GB. However, we can increase this number to a theoretical limit of 8 TB of data.
7. **Speed:** MySQL is considered one of the very fast database languages, backed by a large number of the benchmark test.
8. **High Flexibility:** MySQL supports a large number of embedded applications, which makes MySQL very flexible.
9. **Compatible on many operating systems:** MySQL is compatible to run on many operating systems, like Novell NetWare, Windows* Linux*, many varieties of UNIX* (such as Sun* Solaris*, AIX, and DEC* UNIX), OS/2, FreeBSD*, and others.
10. **Allows roll-back:** MySQL allows transactions to be rolled back, commit, and crash recovery.
11. **Memory efficiency:** Its efficiency is high because it has a very low memory leakage problem.

b) MongoDB

1. **Support ad hoc queries**: In MongoDB, you can search by field, range query and it also supports regular expression searches.
2. **Indexing:** You can index any field in a document.
3. **Replication:** MongoDB supports Master Slave replication. A master can perform Reads and Writes and a Slave copies data from the master and can only be used for reads or back up (not writes)
4. **Duplication of data:** MongoDB can run over multiple servers. The data is duplicated to keep the system up and also keep its running condition in case of hardware failure.
5. **Load balancing:**It has an automatic load balancing configuration because of data placed in shards.
6. Supports map reduce and aggregation tools
7. Uses JavaScript instead of Procedures.
8. It is a schema-less database written in C++.
9. Provides high performance.
10. Stores files of any size easily without complicating your stack.