**Software Engineering Tools Lab**

Assignment No-1

(Module 1- Introduction to OSS)

**Name: Muskan Raju Attar**

**PRN:2019BTECS00017**

**Batch:T1**

**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.**

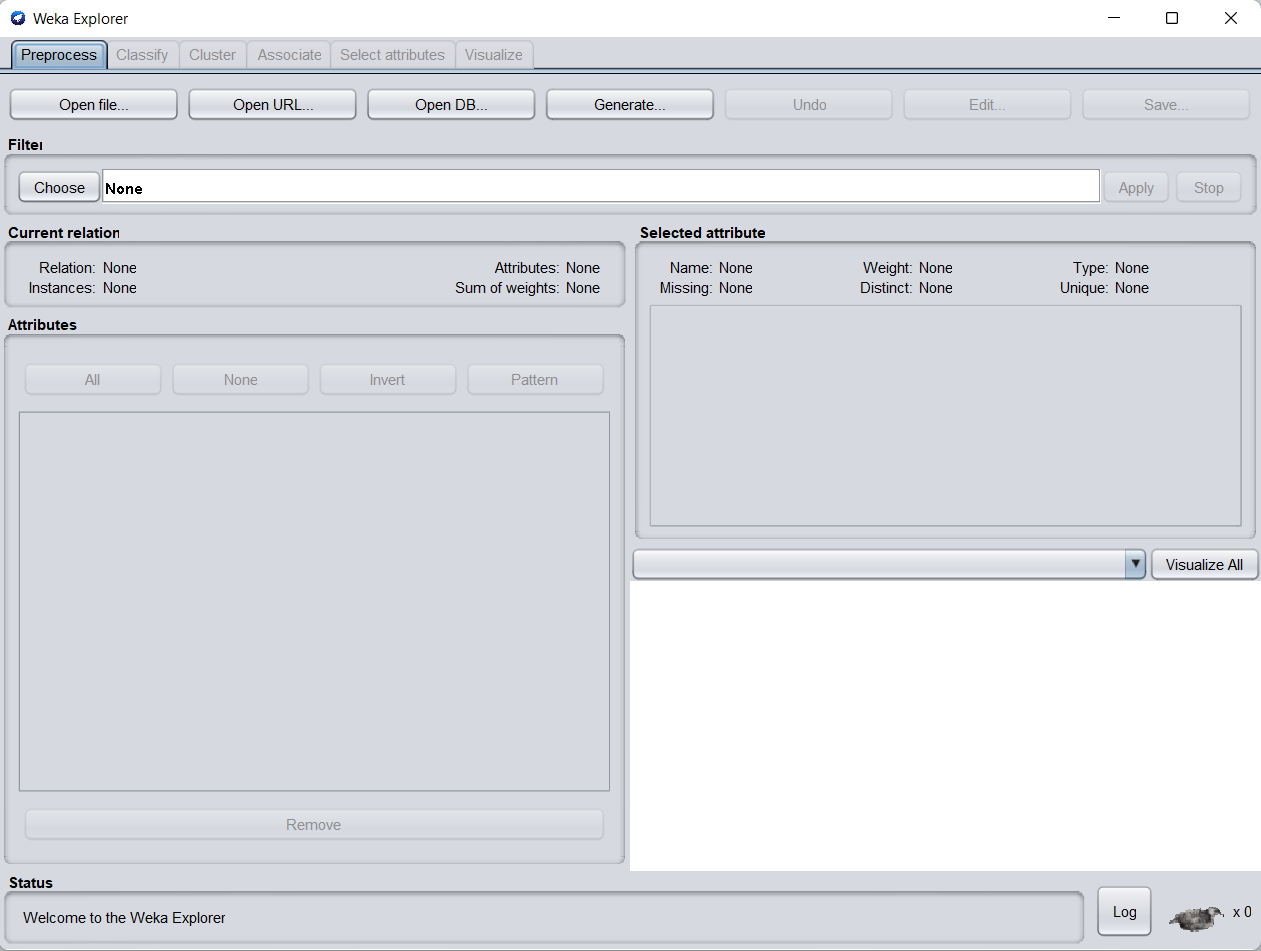
**Use the Iris dataset**

**Given https://drive.google.com/file/d/1A3Fxsfzm6BSfhFZGDrjI47RTe45bSgYP/view**

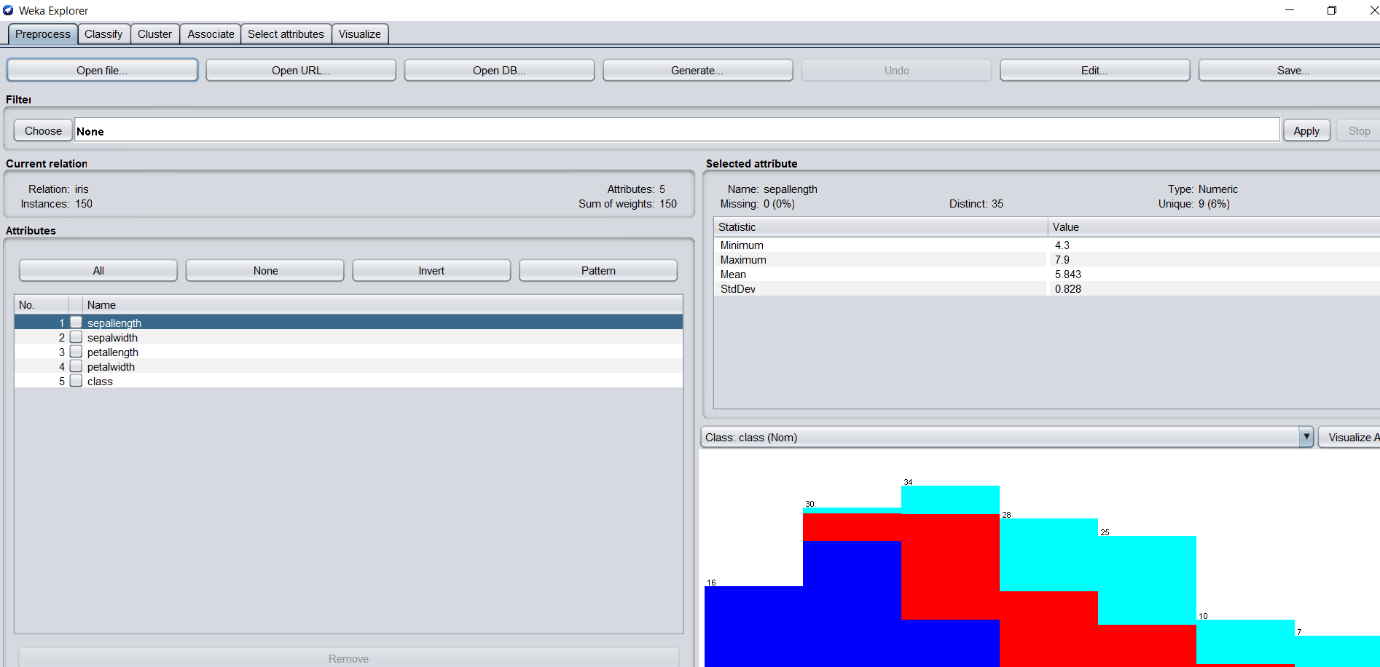
**Note-provide screen shots for every task**

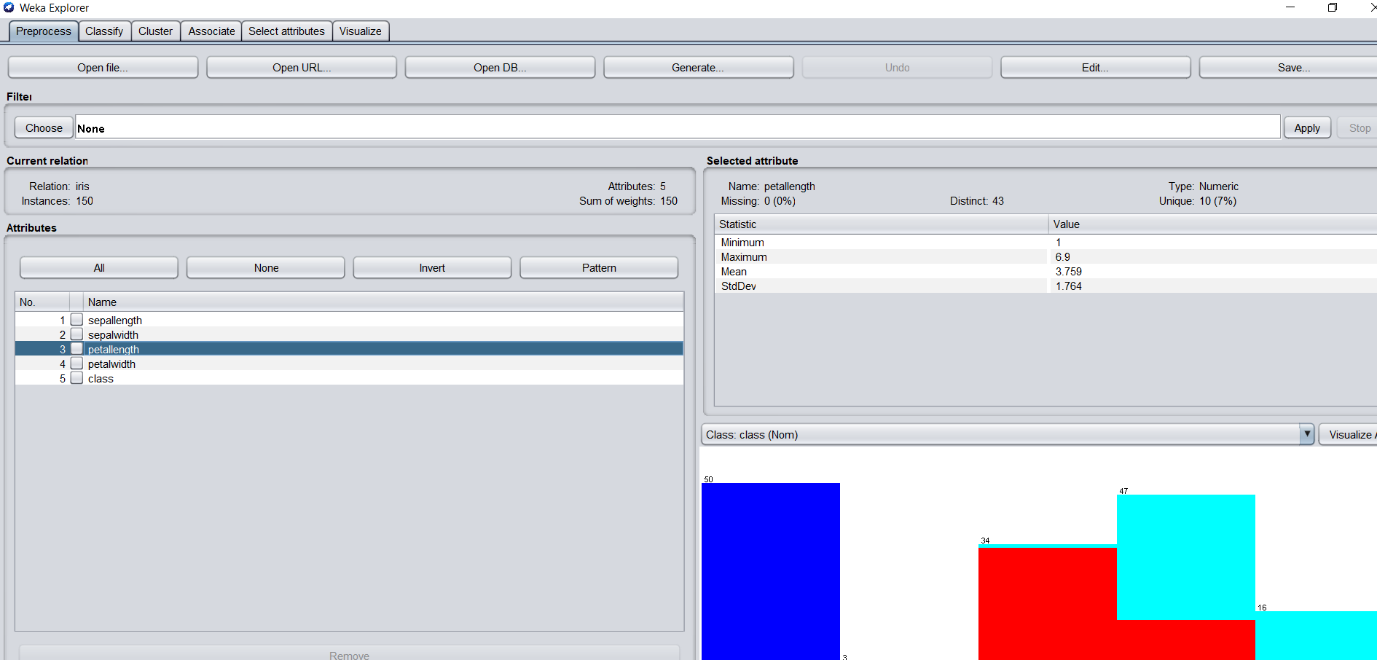
**Create a report which will illustrate the details of tasks performed (for e.g to perform Preprocessing of data provide details of navigation and selection of appropriate parameters)**

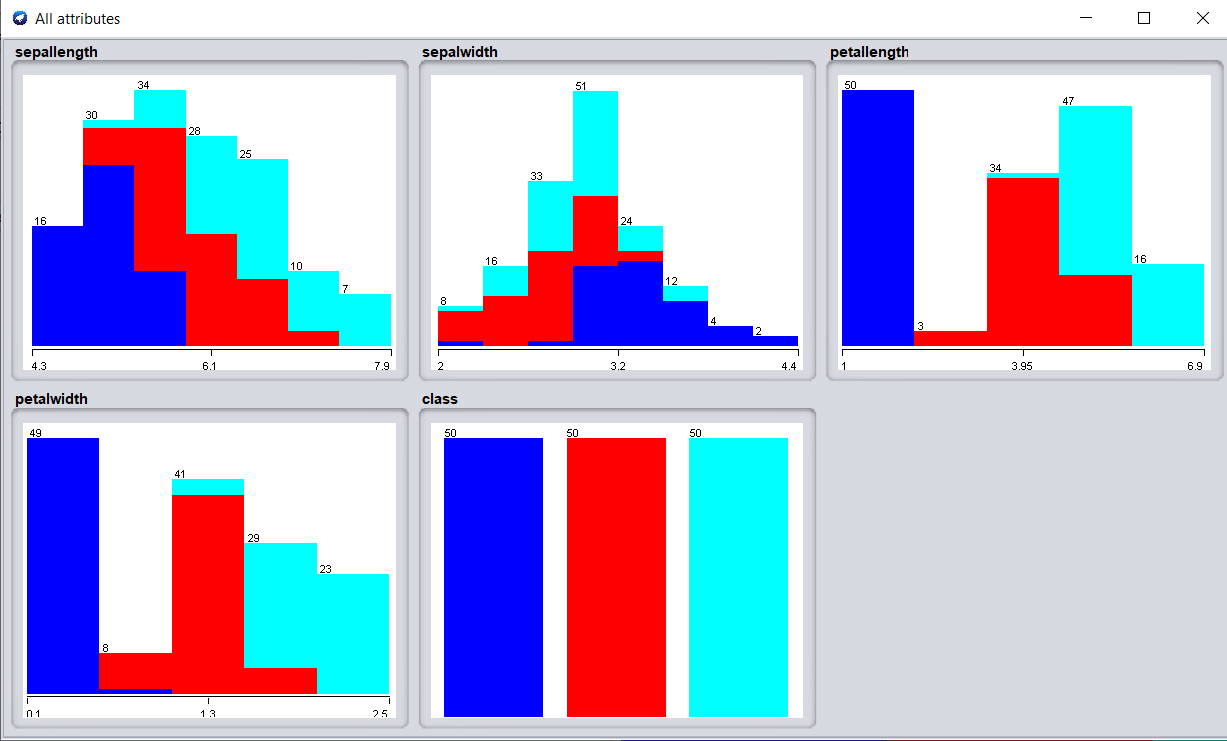
**Installation:**



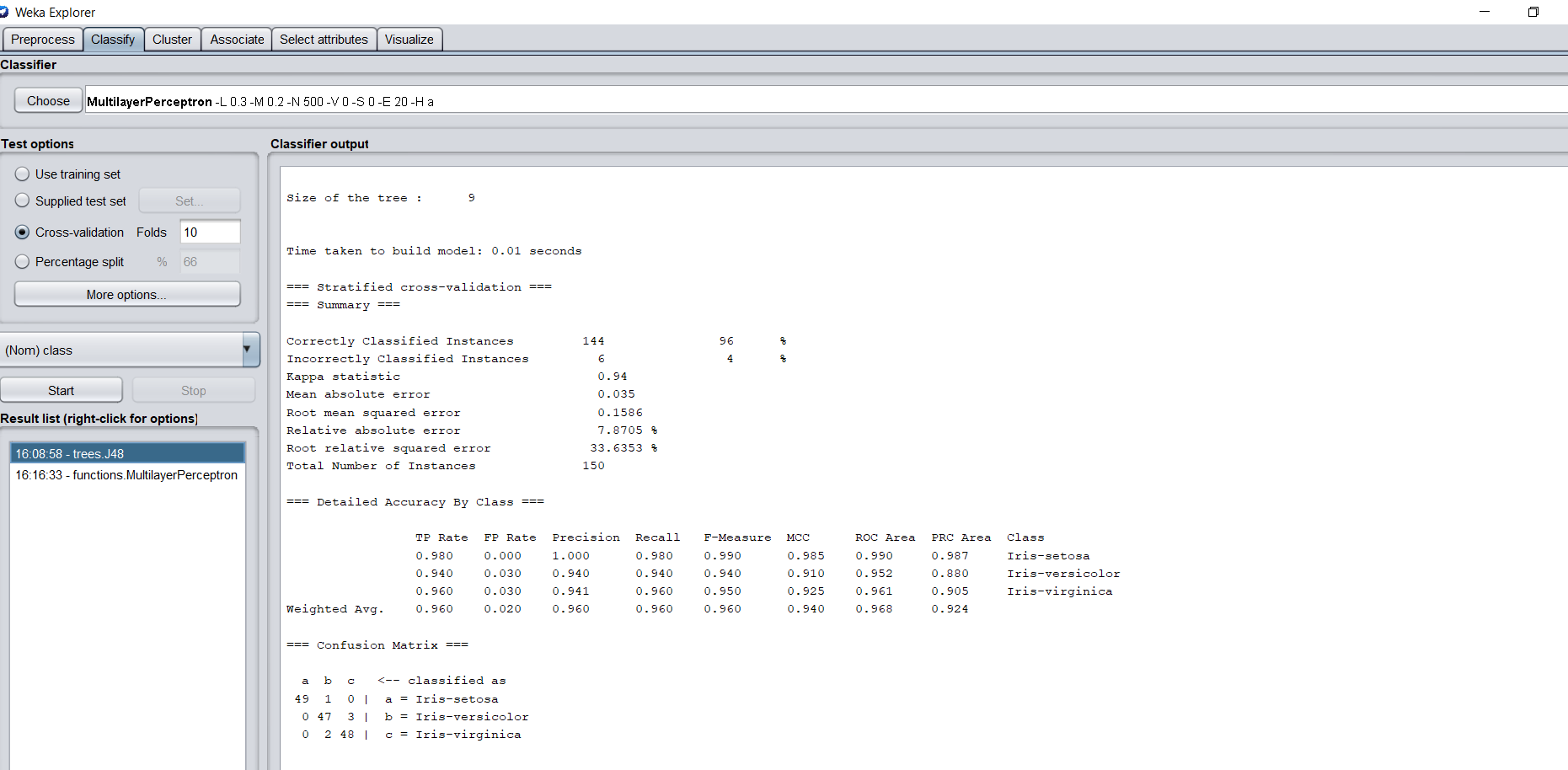
**Preprocessing**

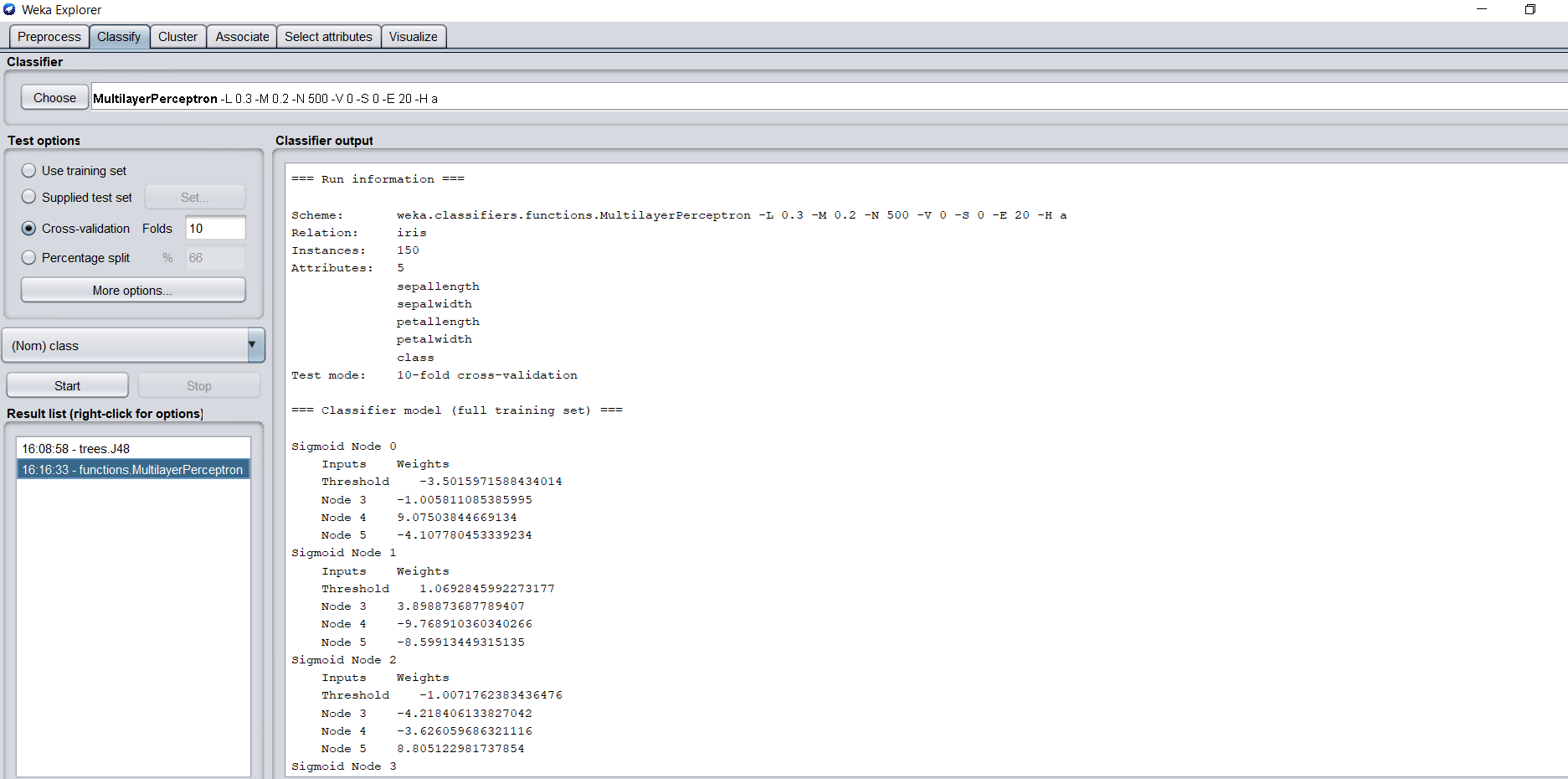


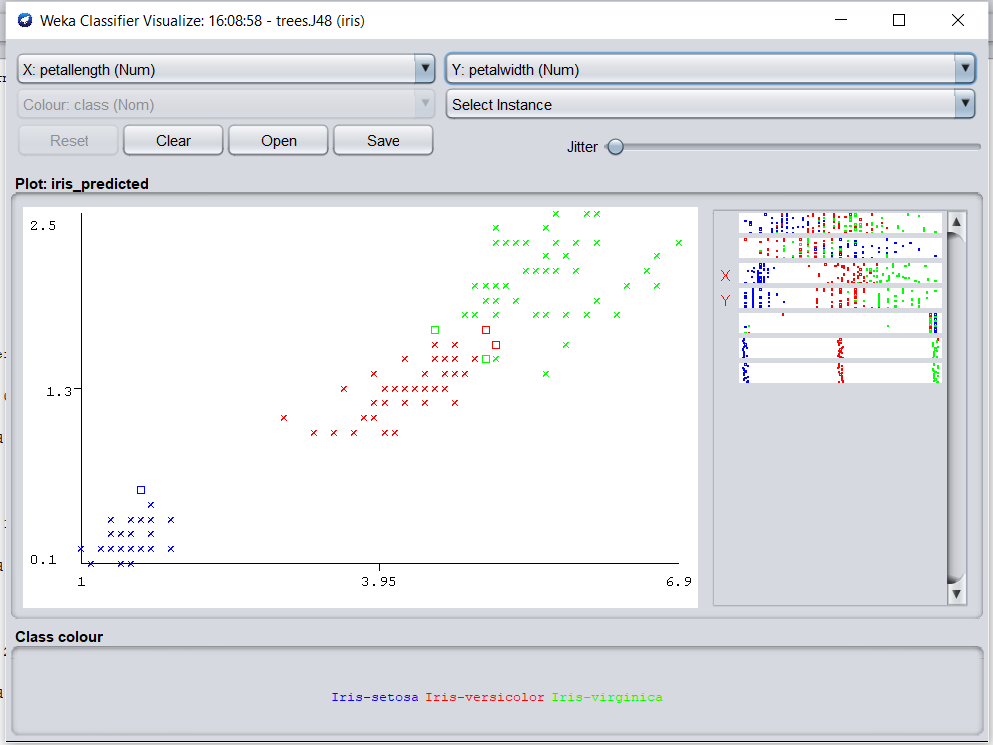


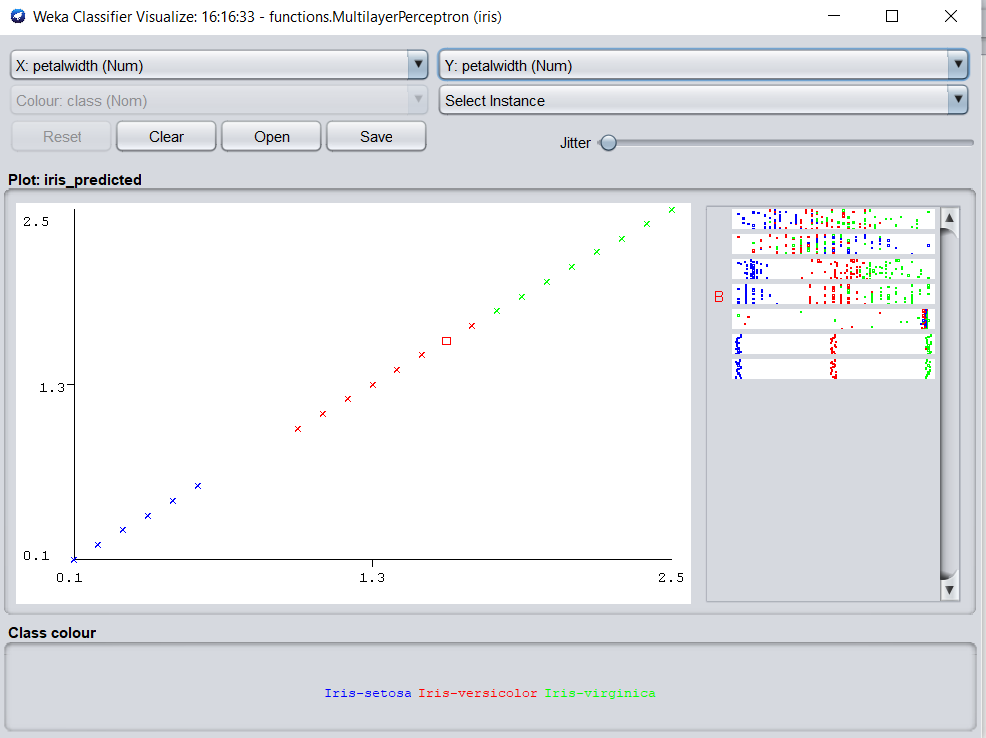


**Data Classification:**

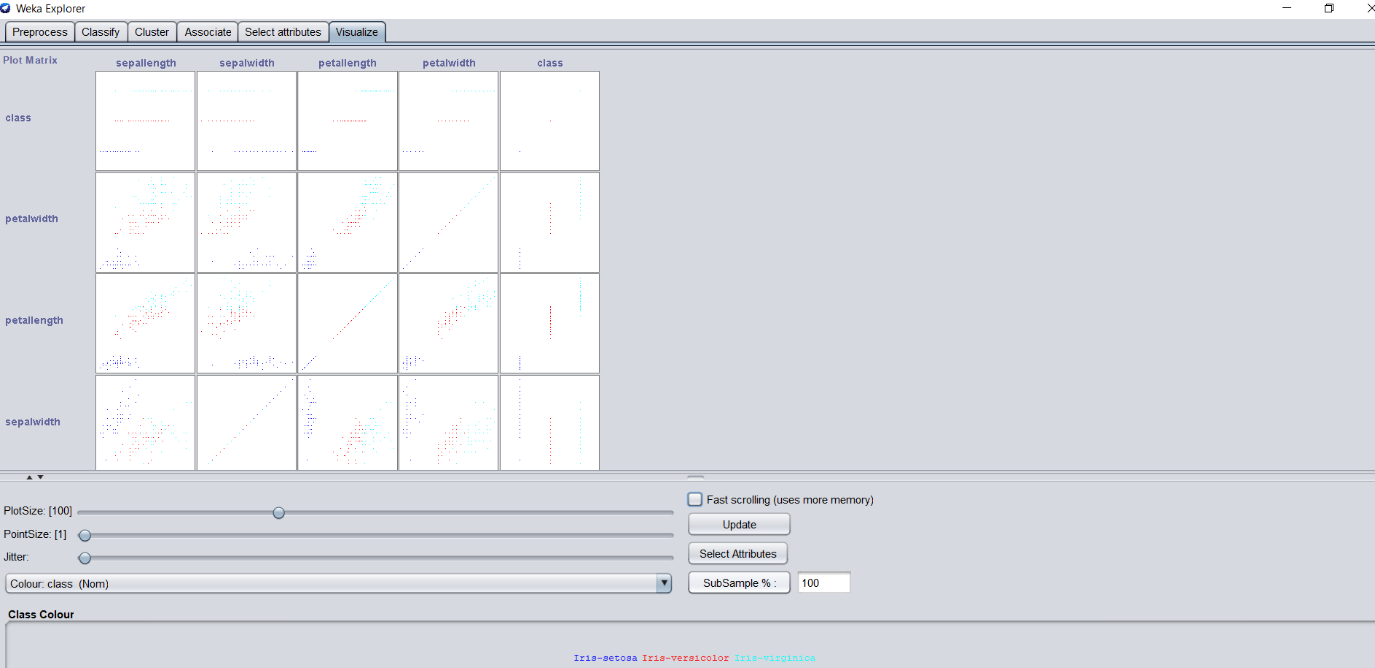


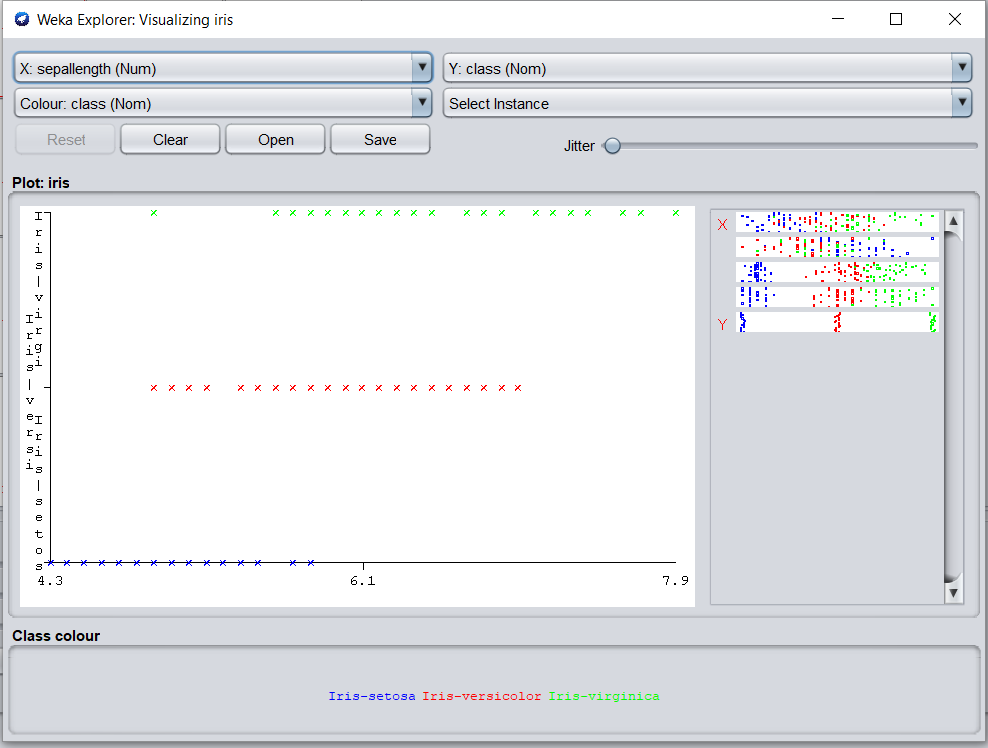






**Data Visualization:**





**Report :**

**Abstract:**

IRIS is an open access flower based dataset and is normally available on UCI dataset. The major objective of this research work is to examine the IRIS data using data mining techniques available supported in WEKA. In this work, four different classifier viz. Bayes Network Classifier, J48, Random Forest and OneR has been succefully used to classify the IRIS dataset. The dataset consist of five different attributes viz. sepallength, sepalwidth, petallength, petalwidth and class. The number of instaces in 150. It has been observed that the rate of correctly classified instances using J48 is better than bayes network, random forest and oneR classifier. The use of J48 assist us in getting 96% of accuracy. Whereas, the mininmum rate of classification achieved is with bayes network classifier.

**Keywords**: Data Mining, Classfiers, IRIS data set and Kappa statistics.

1. **Introduction**

Data mining is an important area for computer sciencists and researchers. Nowadays, there is no problem of data. however, the main problem lies in extracting meaningful information from the large volume of data. data mining techniques assists in mining large volume information and converting data into meaningful information so that the data can be classified, grouped or past and future prediction can be made[1][2]. In last few years, lot of research work has been done using differet data mining techniques in the area of agriculture[3][4][5], business & marketing [6][7], medical science [8-15], stock market[16][17] and pharmaceutical products [18][19]. The root of data minig techniques lie in three different subjects viz. Statistics, Artificial Intelligence and Machine Learning. Several heuristics have been projected to perk up the competence of the data mining process. As stated earier, clustering, association mining and prediction are four major tasks of data mining technique.

1. **Literature Review**

Data mining is playing significant role in the current days. In general, data mining techniques can be described into different categories known as classification, clustering, association, regression and prediction. These different techniques have been successfully used in different area viz. agriculture, health science, business, fincance, engineering, weather forecasting etc. It has been found that different researchers have used different classification clustering, association and predictive techniques for mining their massive data of different domain. In agriculture, people work on finding the relationship between spray and food/vegetables, Prediction of problematic wine fermentation, plants disease diagnosis, optimizing pesticides etc [18][19][20]. In medical science, different classification and clustering techniques have been used to diagnose different human diseases like diabetes, cardio, stroke, stress, cancer etc. [21][22][23]. Moreover, some of the people have also find the association between medicine and health of the person. In-text processing, opinion mining, web mining and sentiment analysis are on the top list[24][25].

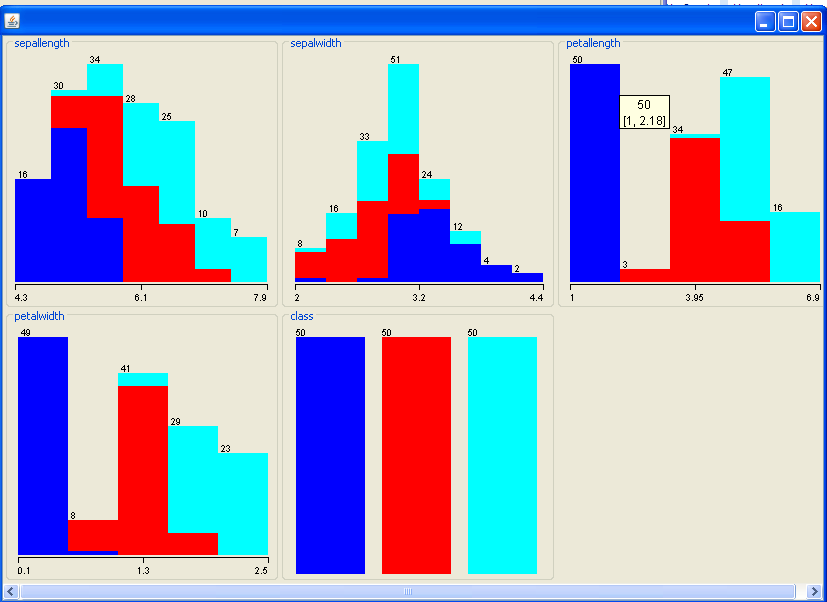
1. **Methods and Results**

**IRIS is flower based multivariate dataset.** This is perhaps the best known database to be found in the pattern recognition literature. It has 150 instances and 4 attributes. In this dataset, there are three different classes of 50 instances each, where each class refers to a type of iris plant.

Table 1: Characterstics of Dataset

|  |  |
| --- | --- |
| **Attribute** | **Value** |
| Data Set Characteristics | **multivariate** |
| Attribute Characteristics: | **Real** |
| Number of instances | **150** |
| Number of attributes | **4** |
| Missing value | **No** |
| Domain | **Life science** |

The visualization of all five different attributes viz. sepallength, sepalwidth, petallength, petalwidth and class are shown in Figure 1.



**Figure 1: Visualization of all IRIS attributes**

The four different classifier viz. Bayes Network Classifier, J48, Random Forest and OneR have been succesufully employed using WEKA for IRIS dataset. Table 2, 3, 4 and 5 presents the performance of these classifier in classifying the IRIS dataset.

**Table 2: Performance metric of Bayes Network Classifier**

|  |  |
| --- | --- |
| **Attribute** | Value |
| **Total Number of Instances** | 150 |
| **Correctly Classified Instances** | 139 (92.6667 %) |
| **Incorrectly Classified Instances** | 11 (7.3333 %) |
| **Kappa statistic** | 0.89 |
| **Mean absolute error** | 0.0454 |
| **Root mean squared error** | 0.1828% |
| **Root relative squared error** | 38.7793 % |
| **Time Taken** | 0.02 seconds |

**Table 3: Performance metric of J48**

|  |  |
| --- | --- |
| **Attribute** | Value |
| **Total Number of Instances** | 150 |
| **Correctly Classified Instances** | 144 (96 %) |
| **Incorrectly Classified Instances** | 06 (4%) |
| **Kappa statistic** | 0.94 |
| **Mean absolute error** | 0.035 |
| **Root mean squared error** | 0.1586 |
| **Root relative squared error** | 33.6353 % |
| **Time Taken** | 0.02 seconds |

**Table 4: Performance metric of Random Forest**

|  |  |
| --- | --- |
| **Attribute** | Value |
| **Total Number of Instances** | 150 |
| **Correctly Classified Instances** | 143 (95.3333 %) |
| **Incorrectly Classified Instances** | 07 (4.66%) |
| **Kappa statistic** | 0.93 |
| **Mean absolute error** | 0.04 |
| **Root mean squared error** | 0.1655 |
| **Root relative squared error** | 35.0999 % |
| **Time Taken** | 0.02 seconds |

**Table 5: Performance metric of OneR**

|  |  |
| --- | --- |
| **Attribute** | Value |
| **Total Number of Instances** | 150 |
| **Correctly Classified Instances** | 141 (94 %) |
| **Incorrectly Classified Instances** | 09 (6%) |
| **Kappa statistic** | 0.91 |
| **Mean absolute error** | 0.04 |
| **Root mean squared error** | 0.2 |
| **Root relative squared error** | 42.4264 % |
| **Time Taken** | 0.02 seconds |

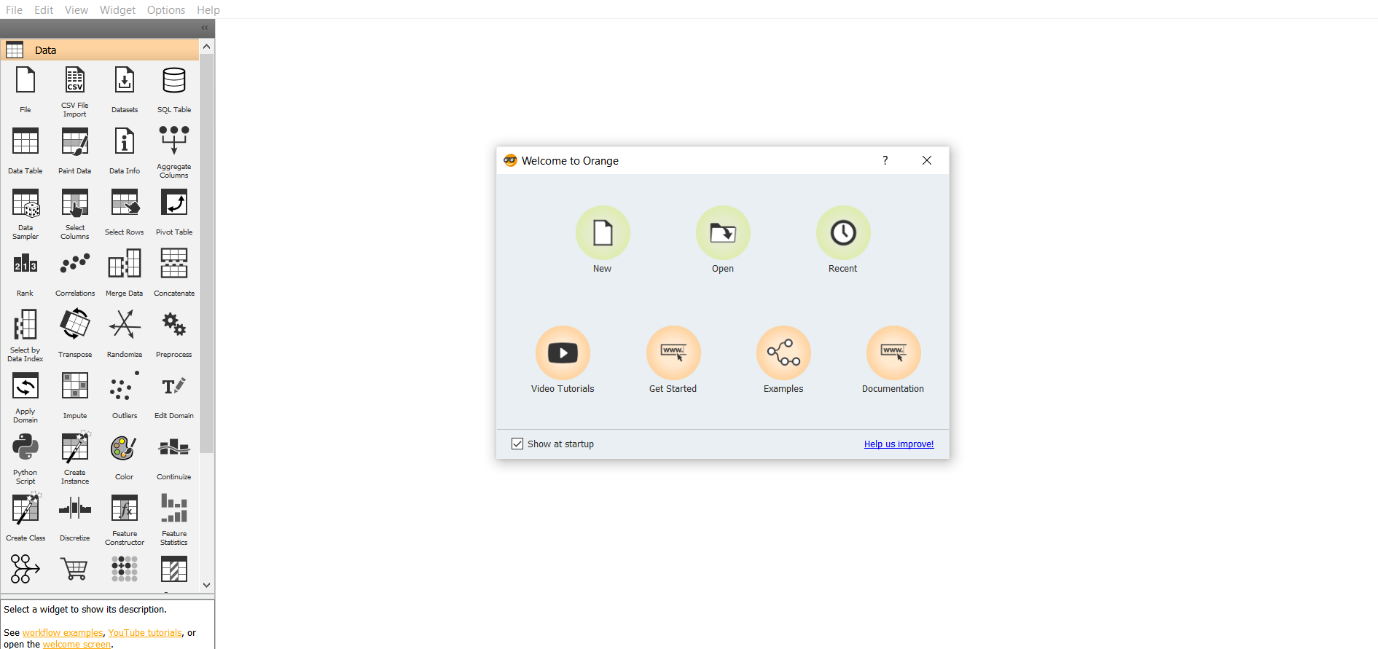
1. **Conclusion**

The objective of this research work is to present the use of WEKA classifiers in categorizing the IRIS dataset. In this work, four different classifier viz. Bayes Network Classifier, J48,

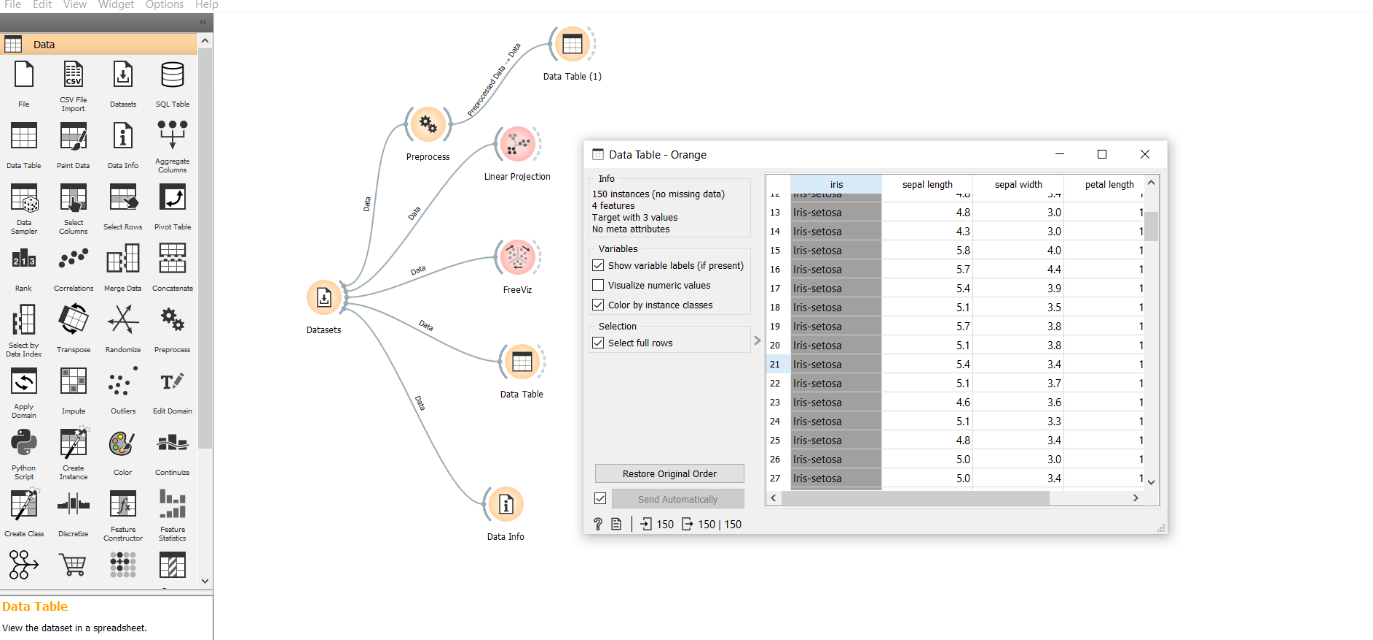
Random Forest and OneR has been succefully used to classify the IRIS dataset. different performance metric such as correctly classified instances, incorrectly classified instances, kappa statistics, mean absolute error, root mean squared error, Root relative squared error along with execution time has been computed and examined. The rate of correctly classified instances using J48 is better than bayes network, random forest and oneR classifier. The use of J48 assist us in getting 96% of accuracy. Whereas, the mininmum rate of classification achieved is with bayes network classifier.

**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.**

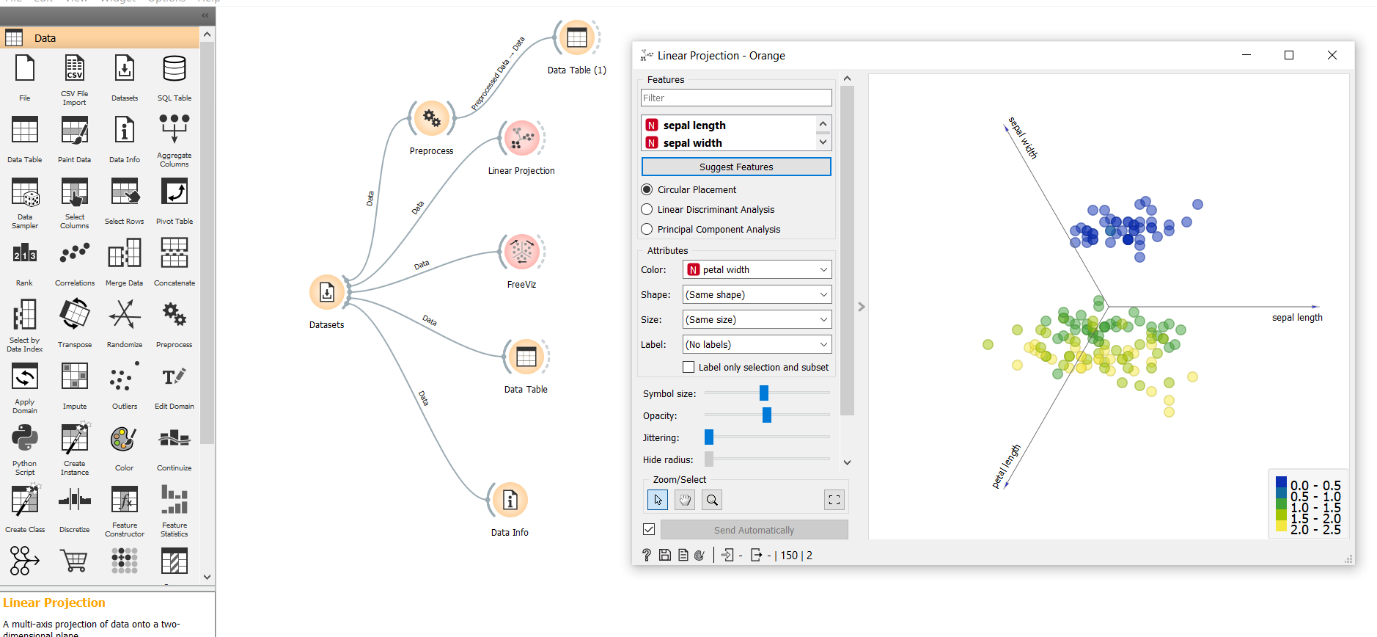
1. Install orange



1. Show data distribution

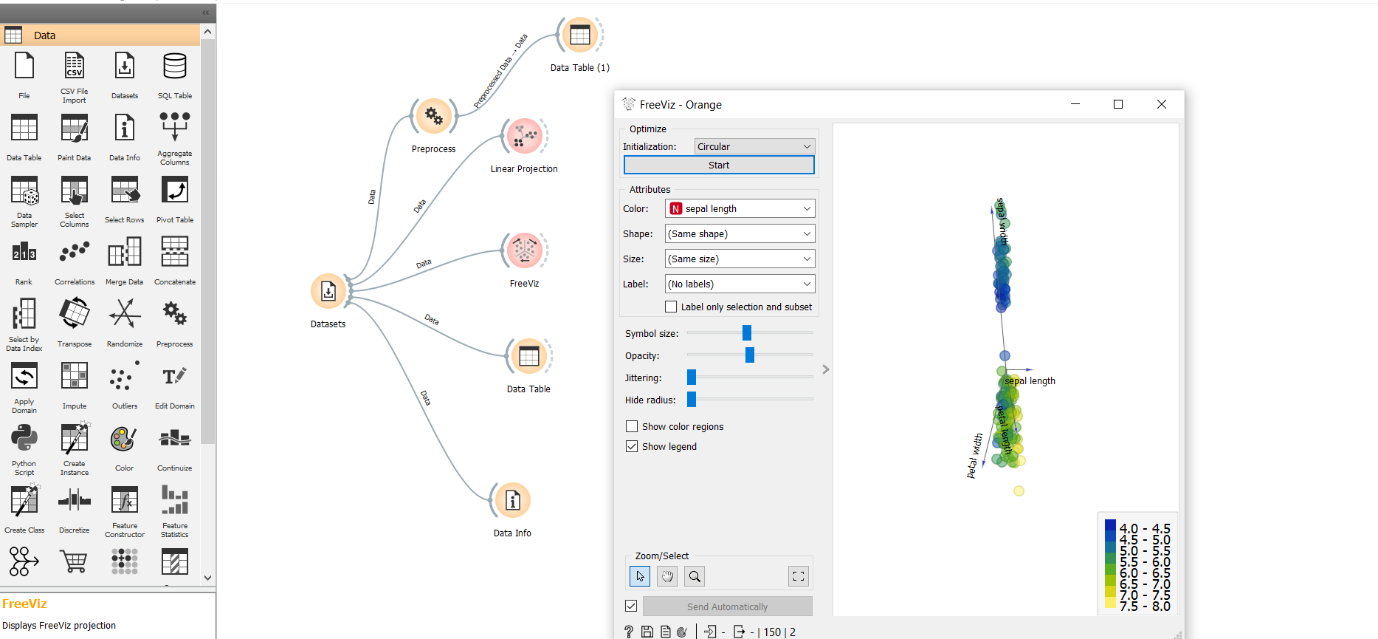


1. Show linear projection

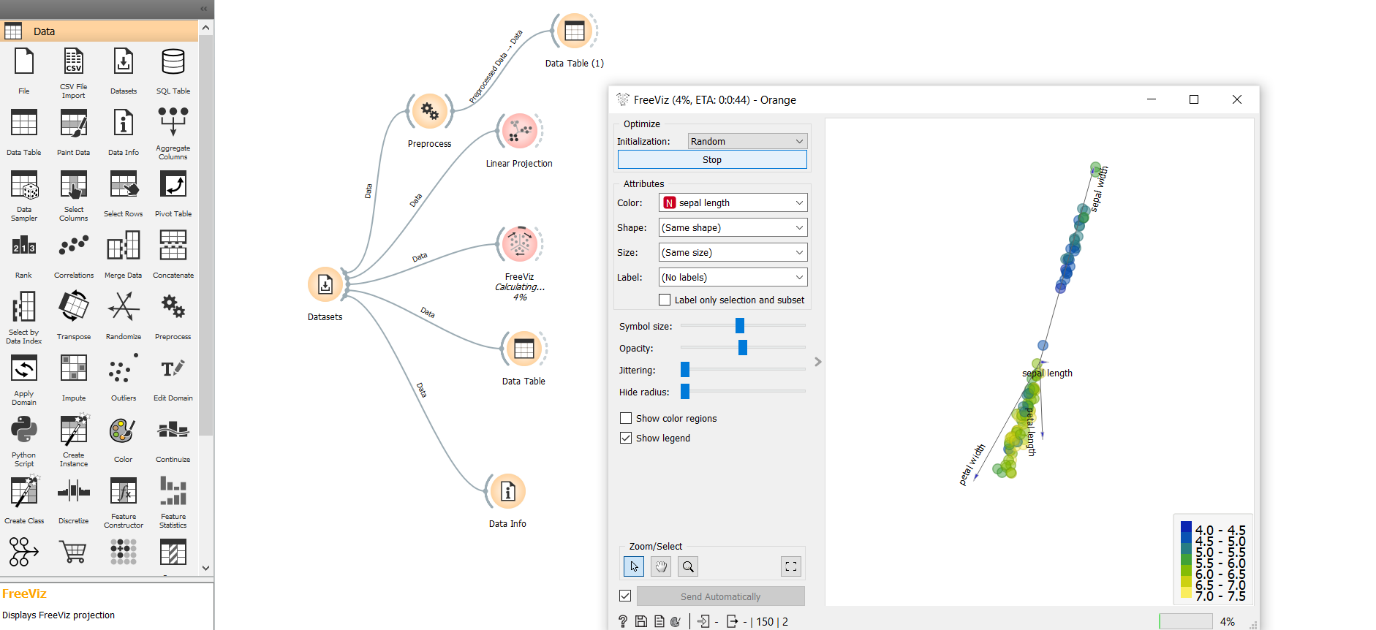


1. Show FreeViz

1)circular:



2)Random:



**Use dataset**

**https://drive.google.com/file/d/1m6sKI1Dap0XK6Bw1edUd5PohwpPwXnd9/view**

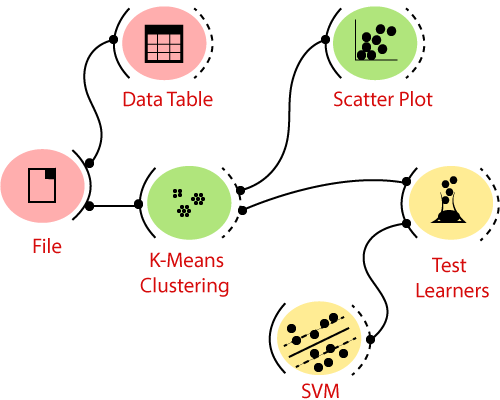
**Create a report for this task and upload screenshots for the same.**

**Report:**

Orange supports a flexible domain for developers, analysts, and data mining specialists. Python, a new generation scripting language and programming environment, where our data mining scripts may be easy but powerful. Orange employs a component-based approach for fast prototyping.

Orange Widgets:

Orange widgets give us a graphical user interface to orange's data mining and machine learning techniques. They incorporate widgets for data entry and preprocessing, classification, regression, association rules and clustering a set of widgets for model assessment and visualization of assessment results, and widgets for exporting the models into PMML.



Widgets convey the data by tokens that are passed from the sender to the receiver widget. For example, a file widget outputs the data objects, that can be received by a widget classification tree learner widget. The classification tree builds a classification model that sends the data to the widget that graphically shows the tree. An evaluation widget may get a data set from the file widget and objects.

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

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter of Comparison | Free software | Open source Software | Proprietary Software |
| Definition | “Free software” means software that respects users’ freedom and community. Roughly, it means that the users have the freedom to run, copy, distribute, study, change and improve the software. | Open Source software refers to software that contains a source code with license which is open to free use, modification and distribution. | Proprietary software refers to the type of software that contains a licensed source code and is copyrighted for use. |
| Collaboration | The term “free software” is sometimes misunderstood—it has nothing to do with price. It is about freedom | Open Source software is developed for open collaboration. | Proprietary software is not meant for open collaboration, but only for the creator and users who have paid for it. |
| Access | Software freedom translates to social freedom. | Open Source software has open access, that is, can be accessed by anyone. | Proprietary software can be accessed only by those who developed it and those who have paid for it. |
| Flexibility | Free Software  Is most Flexible ,  it can used by anyone. | Open Source software is flexible, that is, it can be used, modified and distributed by anyone | Proprietary software has restricted flexibility, that is, there are restrictions on its usage. |
| Examples | The Free Software Directory maintains a large database of free-software packages. Some of the best-known examples include the Linux kernel, the BSD and Linux operating systems, the GNU Compiler Collection and C library; the MySQL relational database; the Apache web server; and the Sendmail mail transport agent. | FreeBSD (Berklee Software Distribution), Android, LibreOffice, [**Ubuntu**](https://askanydifference.com/difference-between-fedora-and-ubuntu/) are a few examples of Open Source software. | Windows, Microsoft, macOS, [**Adobe Photoshop**](https://askanydifference.com/difference-between-apple-aperture-and-adobe-photoshop/), Adobe Flash Player are a few examples of Proprietary software. |

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

**given below.**

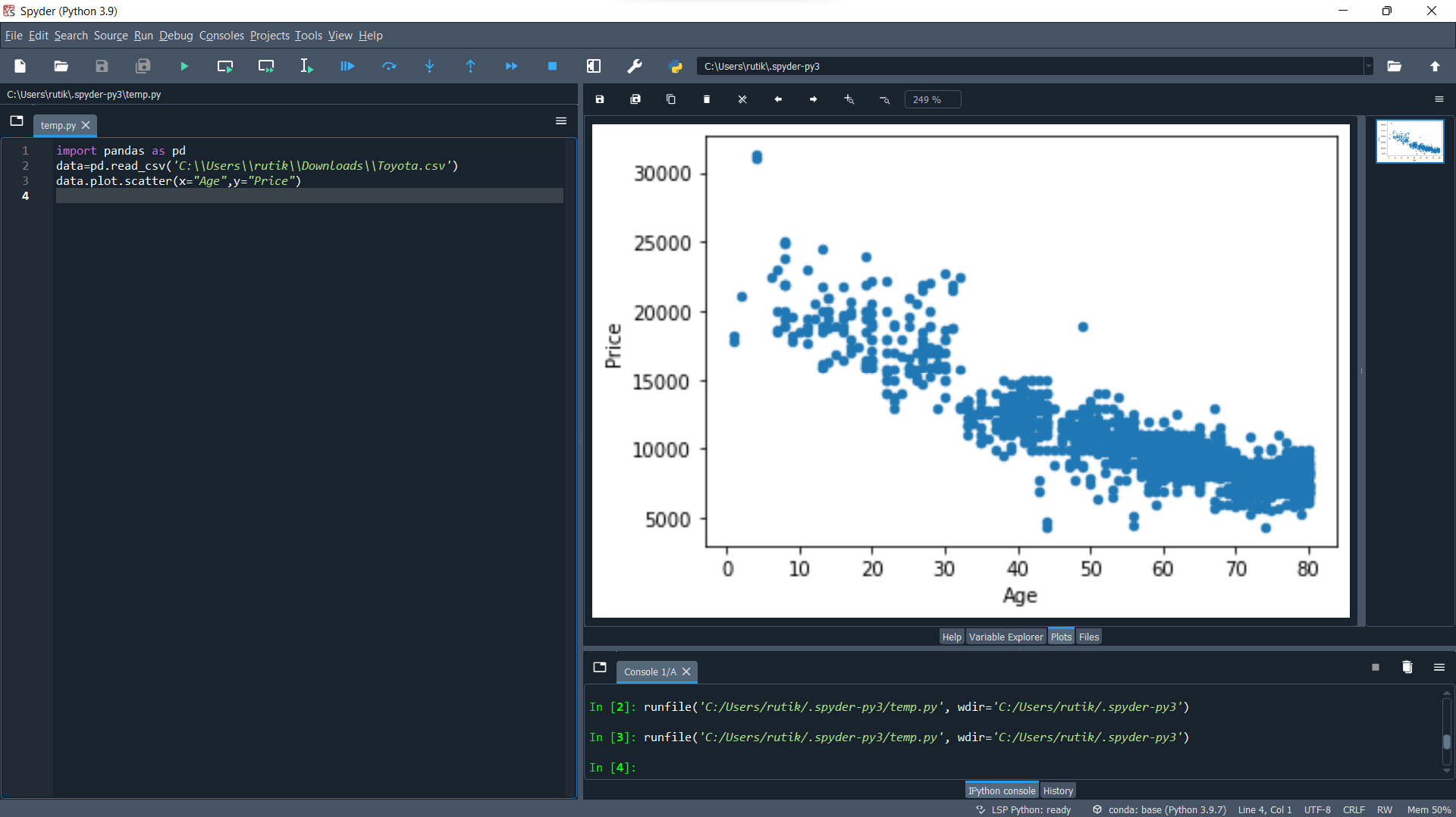
**Dataset- https://drive.google.com/file/d/1i11BZFe8Xj9kNq7eeE9KOa\_Iz1KhEdXJ/view**

**a. Scatter plot- Scatter plot of Price Vs Age**

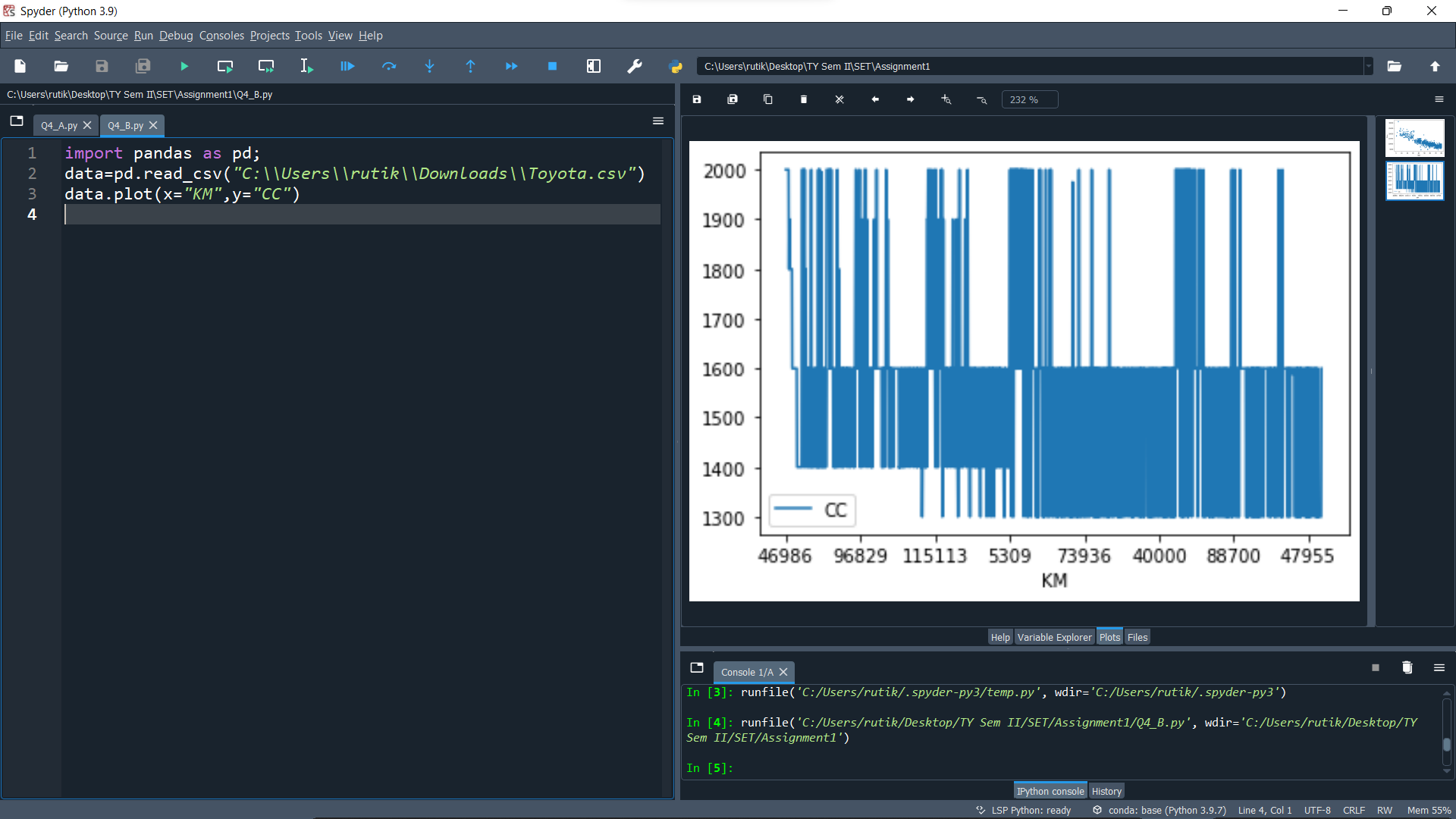
**b. Histogram- for Kilo meter and CC**

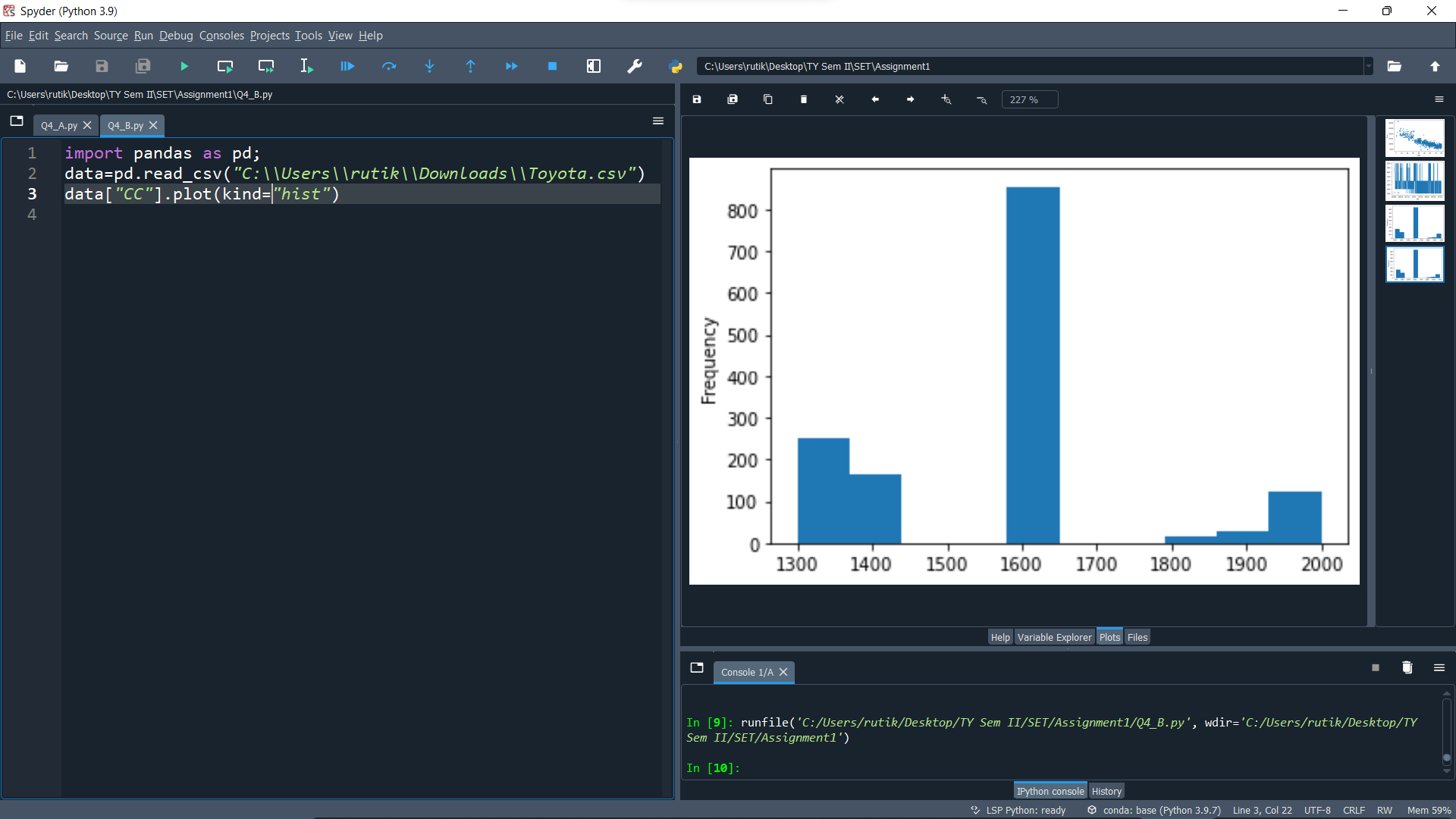
**c. Bar plot- Bar plot for different fuel types**

1. Scatter plot- Scatter plot of Price Vs Age

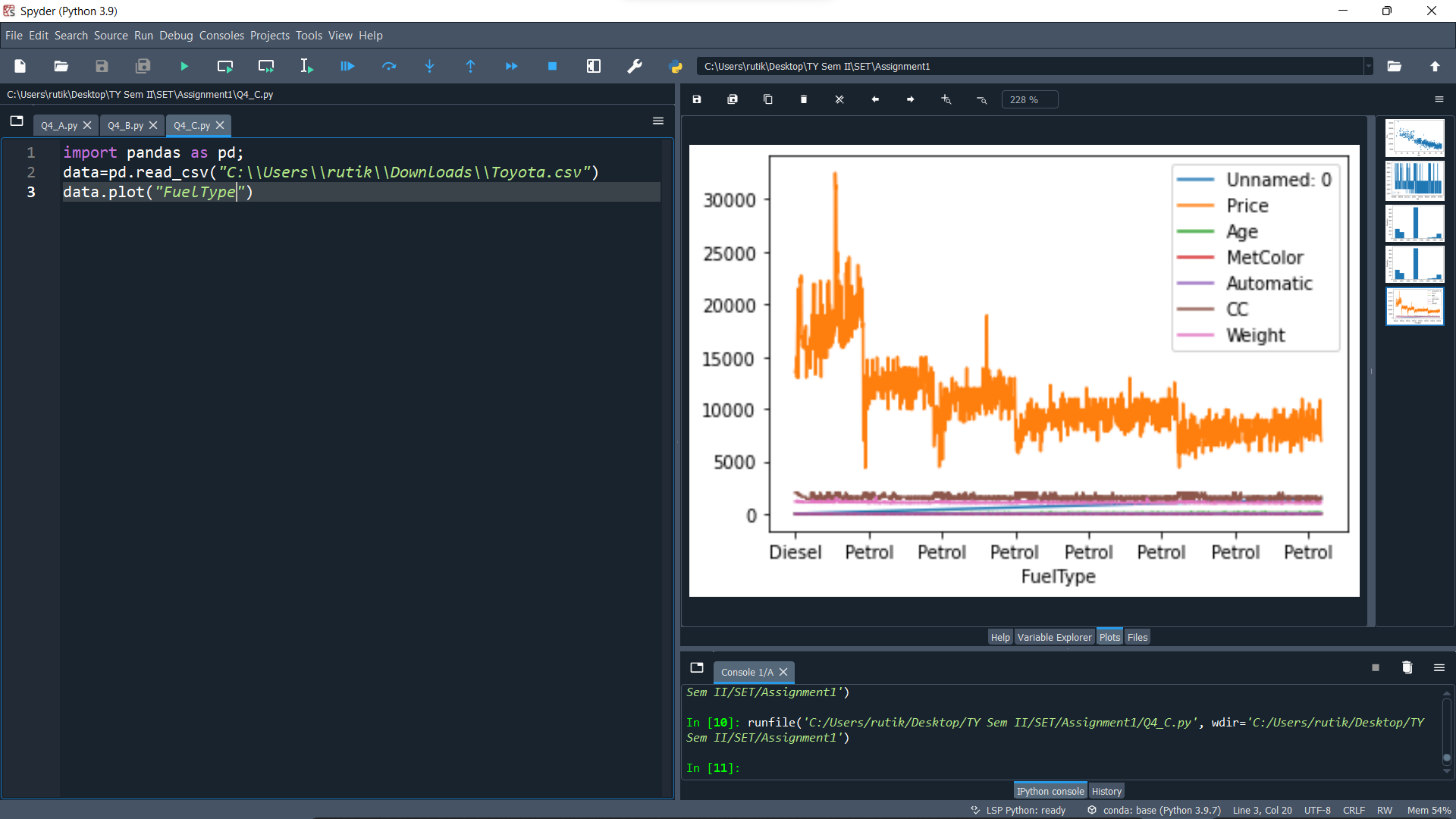


1. Histogram- for Kilometer and CC





1. 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.**

**FOSS:**

Free and open-source software (FOSS) is a software that can be classified as both free software and open-source software.

Examples: Ubuntu, VLC Player, Android system, etc.

**Advantages of FOSS:**

1. Expansive licensing: Proprietary software licenses are usually quite restrictive in terms of use, number of users, type of machine and other. There is fee to own license of a proprietary software. Open-source software are free to own there is no restriction on how we use the software, we can install it on unlimited machines.

2. Transparency: Open-source development is carried out openly. As software code is openly available anyone who finds some bug can fix it for others. As development process is carried out publicly its development process is transparent. Users can easily communicate with product developers to understand their product decisions and offer opinions for betterment of software.

3. Source Code inspection: As source code is openly available anyone can view code of the software for better understanding of how the application works.

4. Source Code modification: We can also modify code after inspecting the code. One can modify code and for himself and also, he can add those changes to main version of software so that changes will be available for everyone for use.

5. Community: Foundation of open source projects is community, it includes developers of software and also users. Users in community can easily share there feedback with developers so that developers can improve the software.

6. Redistribution rights: Open-source licenses allow users to perform some changes in software and again redistribute the software without taking permission of the original product owner.

**Proprietary Software:**

This type of software requires licenses for their use. Company or organization that owns the software provides rights to use the software to customer. Users can only install software only on limited number of machines and cannot redistribute it.

1. Increased Functionality and Convenience

Proprietary systems are easier to use and learn, leading to faster work processes. Skype, for example, is used by organizations worldwide. It takes minutes to sign up for an account and make international phone calls or conduct video interviews online. On top of that, your customers, suppliers and employees may already have a Skype account, so they know how to use it.

Open-source programs are trickier to use and may lack user-friendly features, affecting productivity in the workplace. Unskilled end users may find it difficult to navigate them and take full advantage of what they have to offer. After all, there is a reason so few people use Linux.

2. Superior Customer Support

Open-source software can be difficult to install and set up. Customizing it isn't easier either. Plus, your staff may not be familiar with the program and may need additional training.

The average employee lacks the expertise to use open-source programs. Therefore, your team members may need help with most tasks. They will spend hours trying to figure things out instead of focusing on the tasks at hand.

Proprietary software is more accessible and includes technical support. Most companies offering these programs provide dedicated sources, 24/7 assistance, live chat and user manuals. The antivirus program Bitdefender, for example, offers online resources, technical support around the clock and security-configuration services for enterprises. If your employees experience any issues, they can simply call or email the service provider.

3. Lower Maintenance Costs

As a small-business owner, you may prefer open-source software due to its low cost. Most programs are free or cost next to nothing. The downside is that you may end up paying a lot more for setup, maintenance and customization than you'd pay with proprietary software.

Some open-source programs are difficult to install and set up, so you may need to call an expert to do the job. In some cases, new hardware may be necessary to use the software. If your employees are not familiar with the program, they will need support and training, which may further increase the costs. Updating the software, testing new versions and applying patches isn't cheap either.

4. Stronger Competitive Advantage

Proprietary technology enables organizations to be more profitable, productive and innovative. This is particularly true for software-development companies, which often use custom programs at the core of their business model.

Even if you're not a software developer, you can still benefit from using proprietary systems. For example, you may hire a team to create software programs that integrate with your existing technology. This may improve work performance and productivity in your organization, streamline business processes and increase production. Furthermore, you may customize the program and add new features as your business changes.

5. Secure Financing for Your Business

Nearly one-third of startups close their doors because they run out of capital. Developing proprietary technology doesn't guarantee success, but it could make it asier to secure financing for your small business. Plus, you will be able to charge higher prices because no other company offers the same product as you do.

As it turns out, big data investors prefer to put their money in companies selling proprietary software — or at least something other than open-source software, such as proprietary add-ons. This kind of technology isn't restricted by what already exists in the market.

**Data visualization tools used in IT company:**

1. Excel
2. MS power Bi
3. Tableau
4. Zoho Analytics
5. Sisense
6. Qlik Sense
7. Domo
8. Klipfolio
9. SAP Analytics Cloud