

Date – 11/02/2022

ASSIGNMENT NO 1
PRN- 2020BTECS00211
Name – Aashita Gupta
Course – SET LAB

Q1. 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/1A3Fxsfm6BSfhFZGDrl47RTe45bSgYP/view>

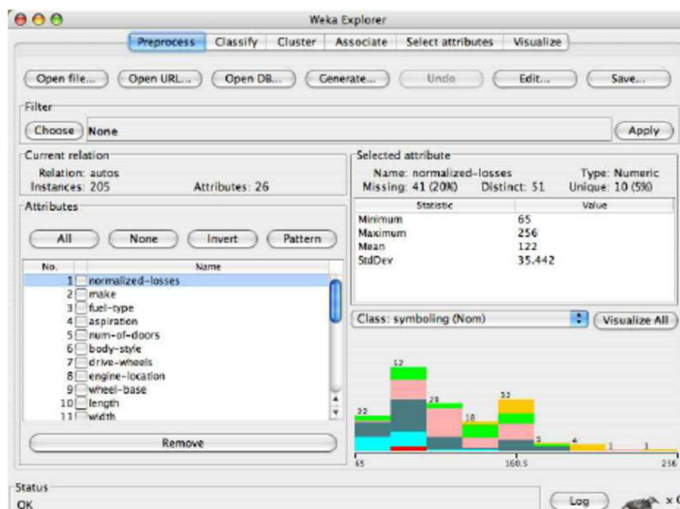
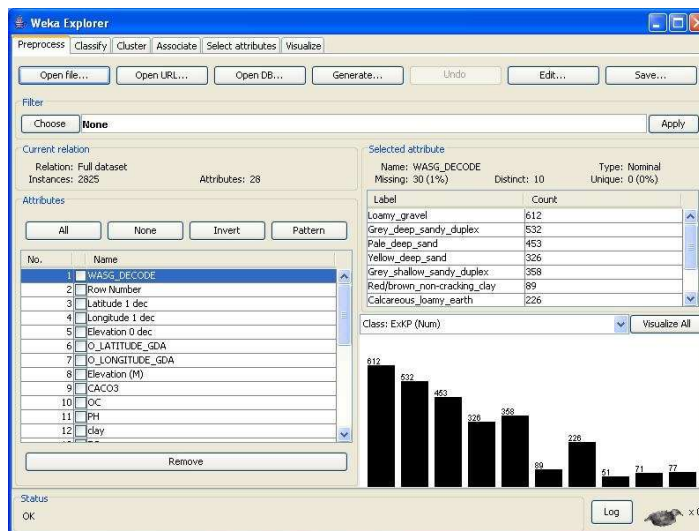
Note-provide screen shots for every task

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)

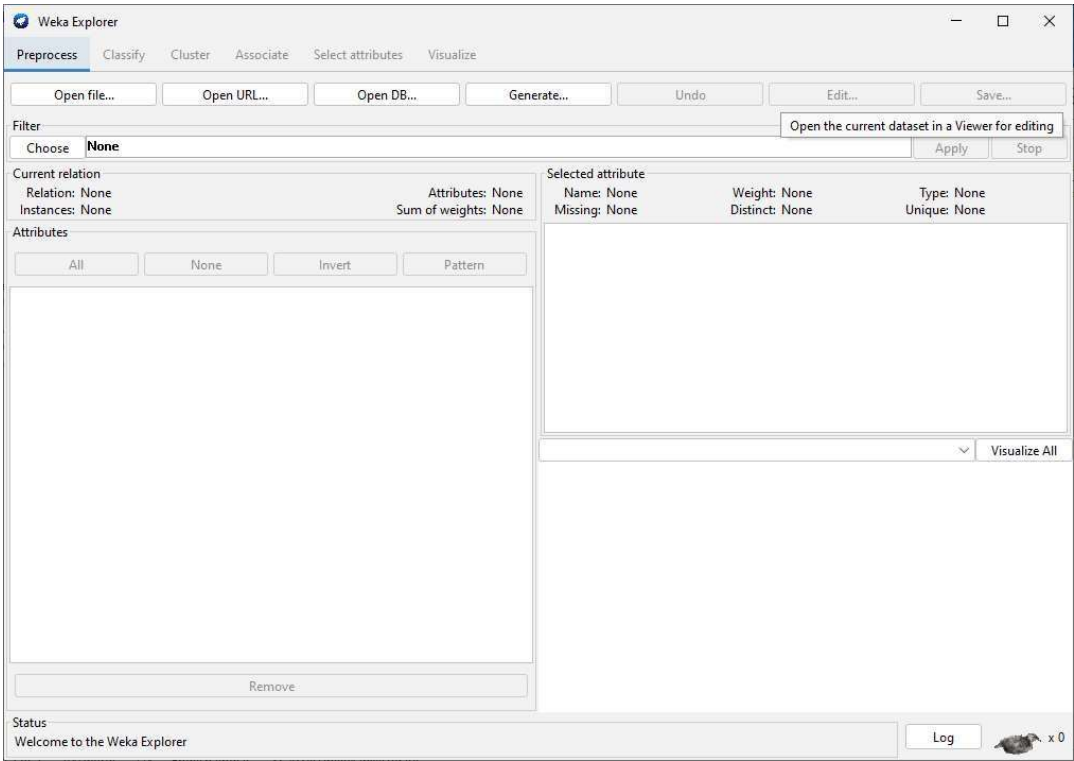
→

Open Weka

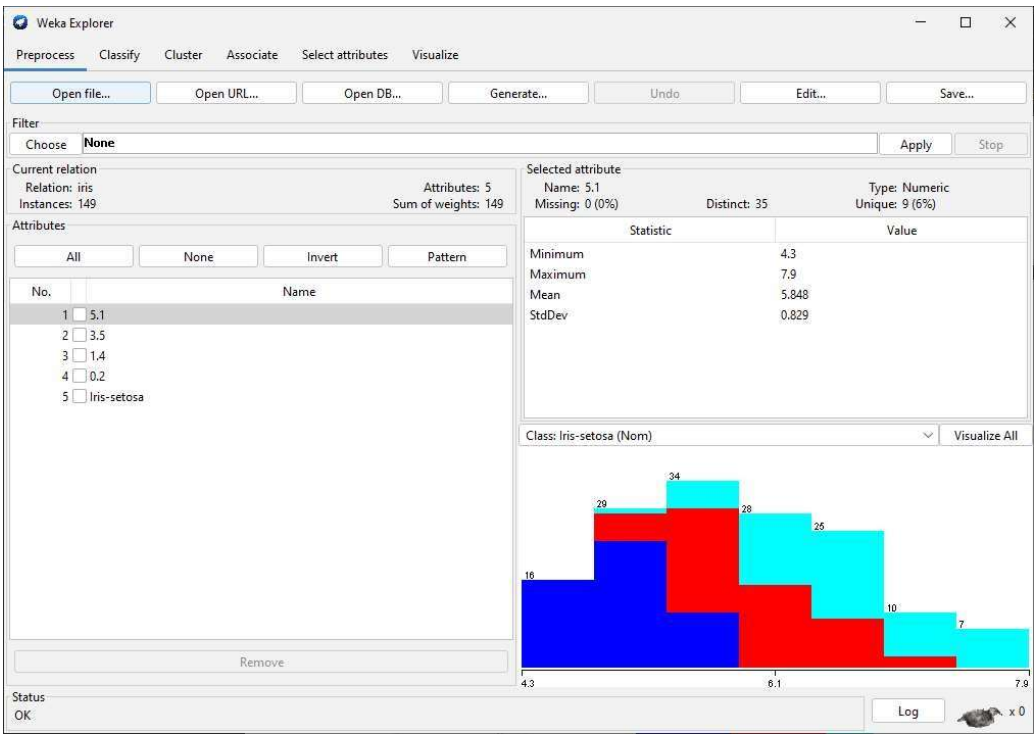




Select Explorer

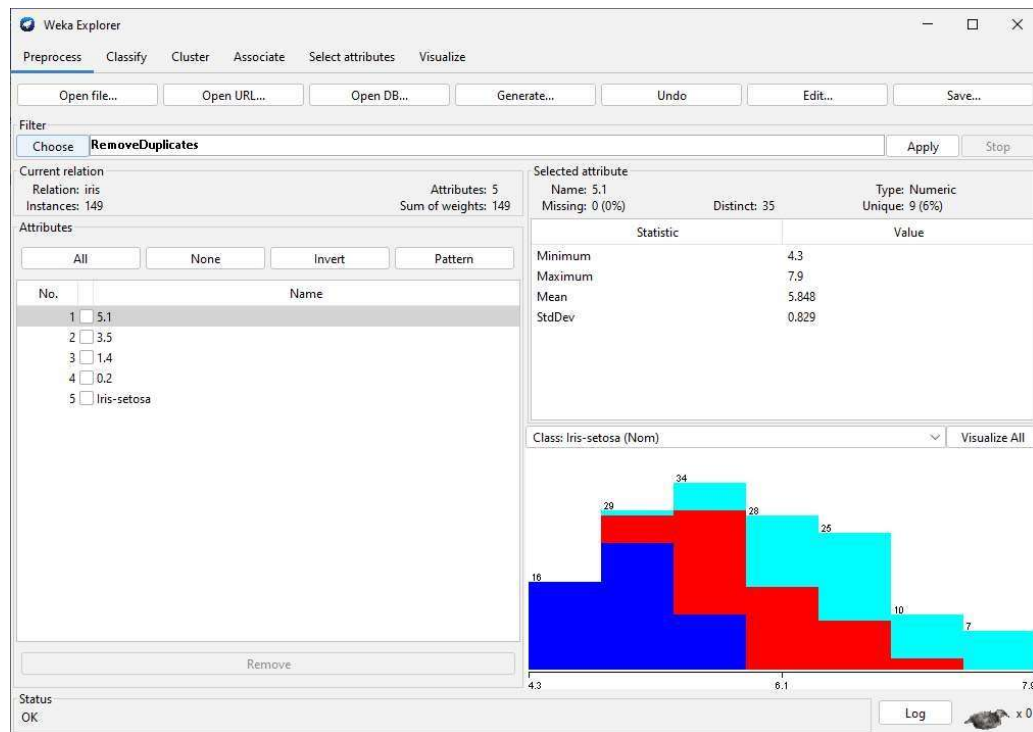


Open iris.csv file

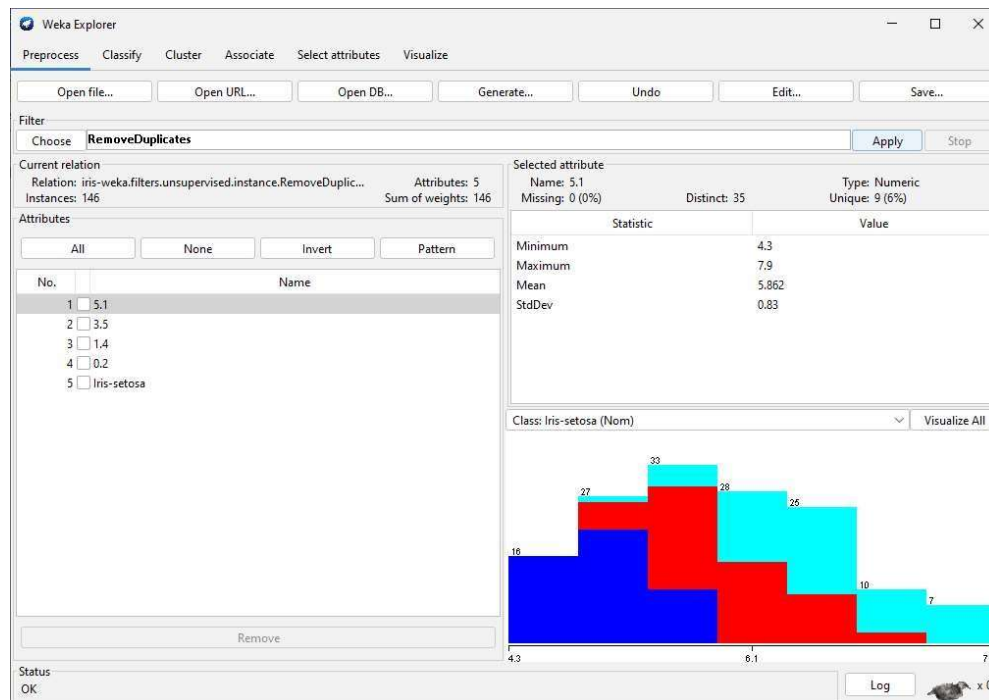


Preprocessing

For adding preprocess filter, click on filter and select filter



After applying



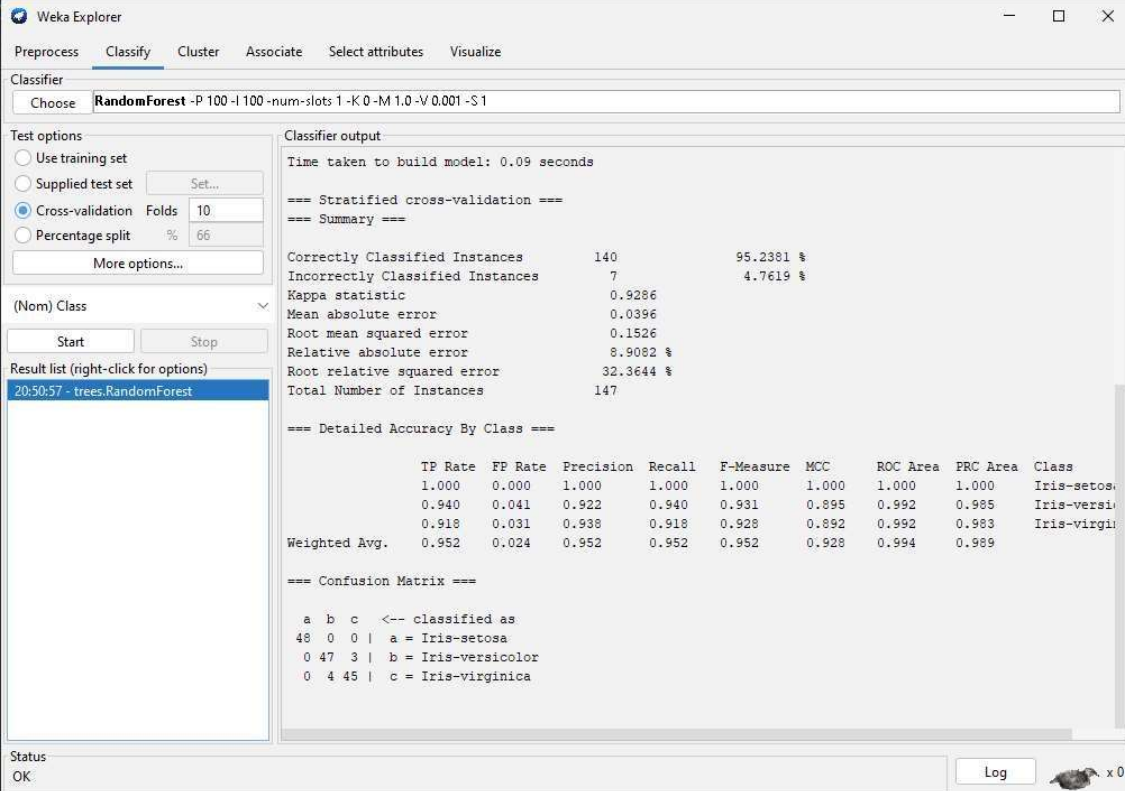
Classification

Select the classify option and select appropriate method from filter

I have chosen Random Forest under Trees

After that we can adjust the options and hit start to see result

I have selected cross validation with 10 folds



Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **RandomForest** -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Nom) Class

Start Stop

Result list (right-click for options)

20:50:57 - trees.RandomForest

Classifier output

Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	140	95.2381 %
Incorrectly Classified Instances	7	4.7619 %
Kappa statistic	0.9286	
Mean absolute error	0.0396	
Root mean squared error	0.1526	
Relative absolute error	8.9082 %	
Root relative squared error	32.3644 %	
Total Number of Instances	147	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Iris-setosa
	0.940	0.041	0.922	0.940	0.931	0.895	0.992	0.985	Iris-versicol
	0.918	0.031	0.938	0.918	0.928	0.892	0.992	0.983	Iris-virginica
Weighted Avg.	0.952	0.024	0.952	0.952	0.952	0.928	0.994	0.989	

=== Confusion Matrix ===

```
a b c <-- classified as
48 0 0 | a = Iris-setosa
0 47 3 | b = Iris-versicolor
0 4 45 | c = Iris-virginica
```

Status OK

Log x 0

You can see the classification result on the right window. we can see confusion matrix, algorithm used, error, accuracy etc.

Q2. 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.

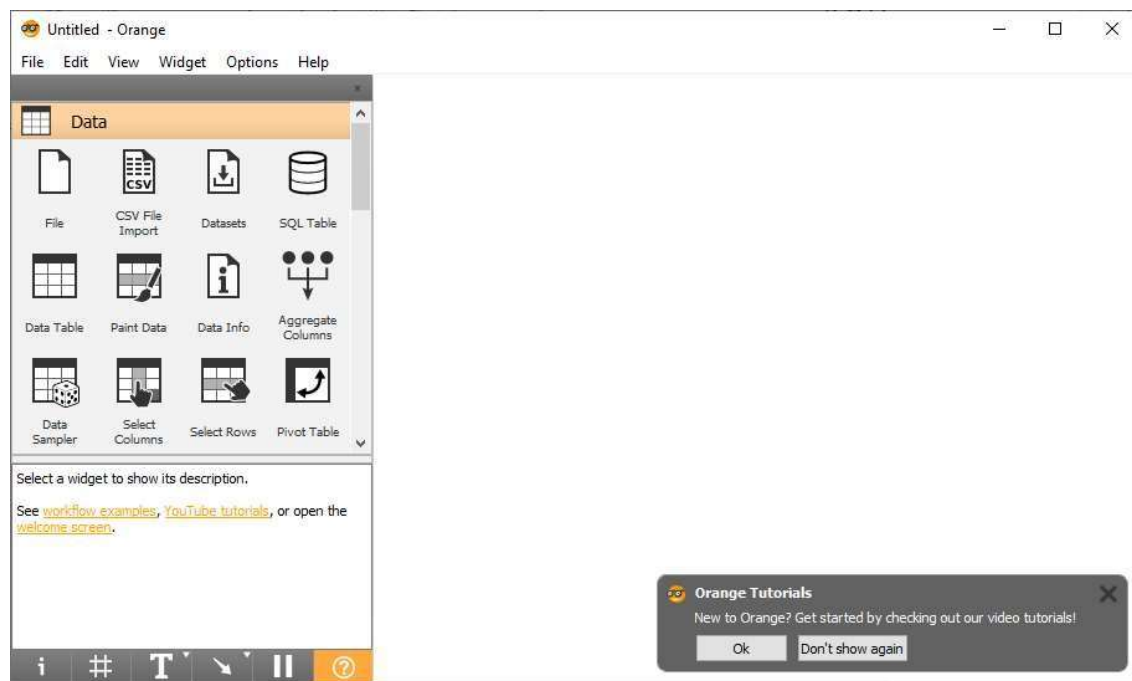
- a. Install orange
- b. Show data distribution
- c. Show linear projection
- d. Show FreeViz

Use dataset

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

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

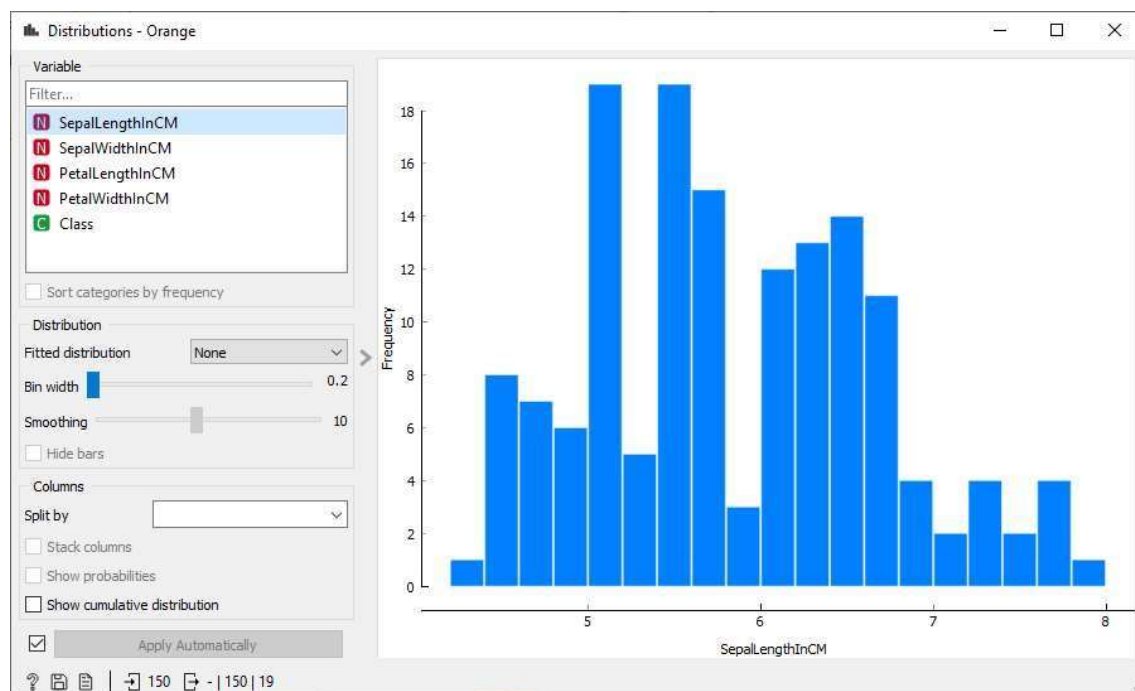
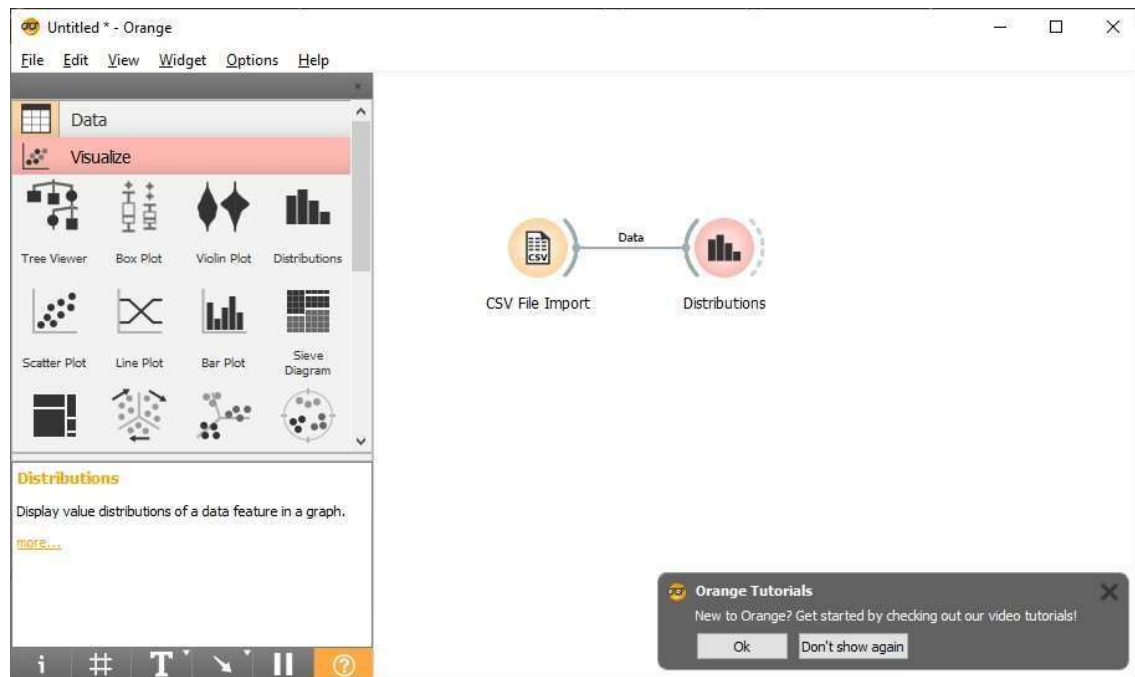
Install and Open Orange



Data Distribution

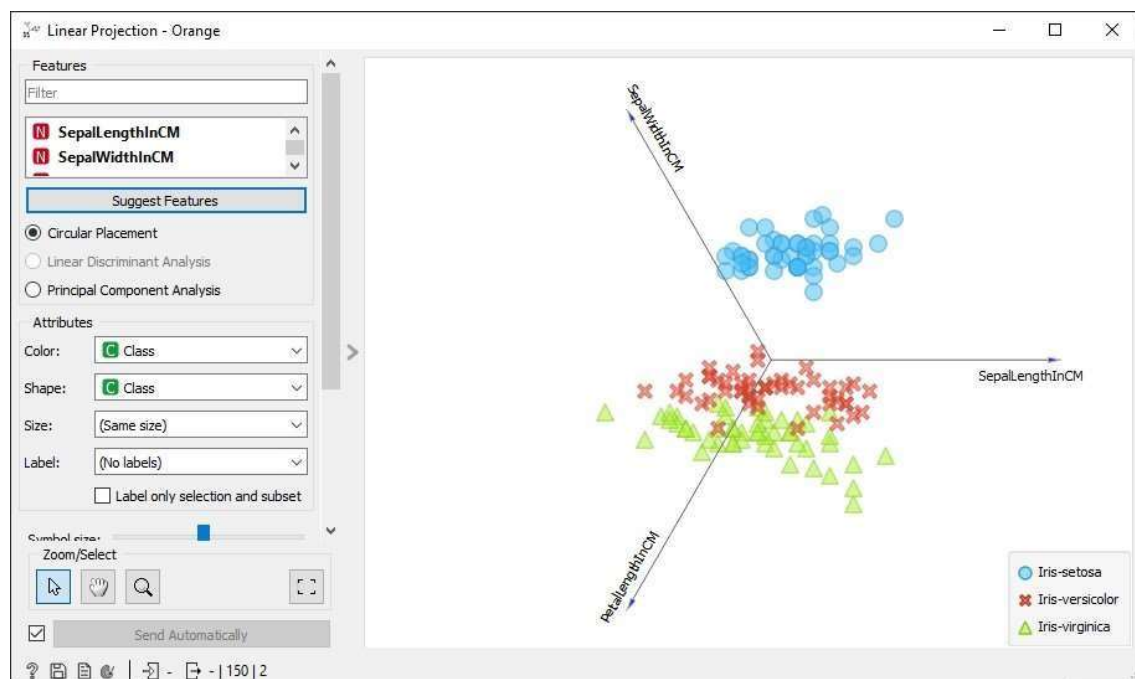
Select CSV import and choose dataset file

Select CSV import icon and then drag and search distribution



Linear Projection

Select Import CSV file and drag and search projection

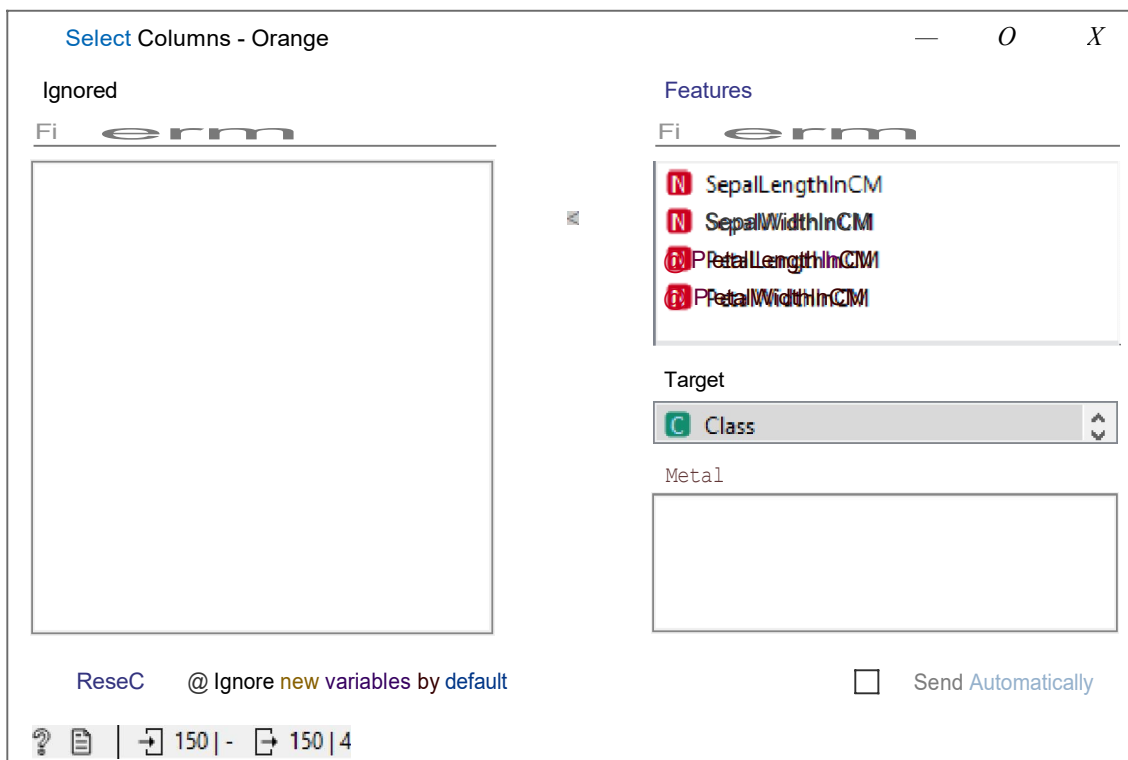
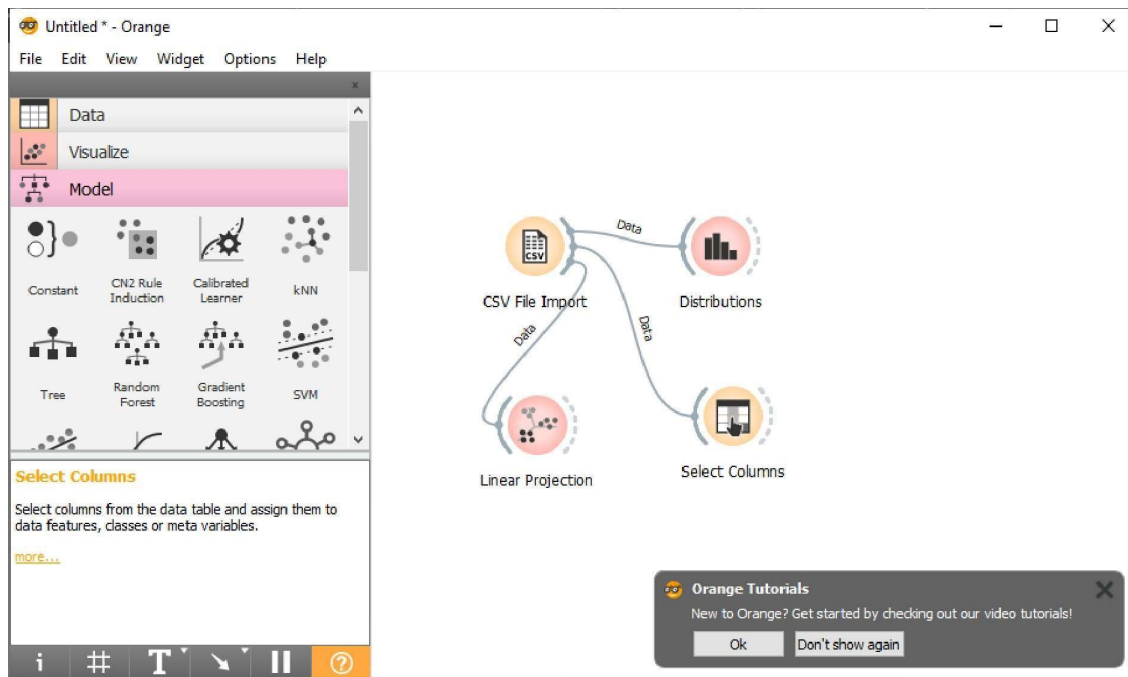


FreeViz

Drag from Import CSV and select Select Column

Select target variable

Drag from Select Column and search FreeViz



Optimize

Initialization: Circular

Rene

Color: Class

Shape: Class

Size: (Size)

Label: (No labels)

@ Label only selection and subset

Symbol size:

Opacity:

Hide radius:

@ 'show' color regions

Zoom/Select



SetJcl Automaticall

PetalLengthInCP1

SepalLengthInCP1

nCM

Iris-setosa
\$g Iris-versicolor
/M Iris-virginica

Q3. Differentiate in between free software, Open source software and proprietary software with respect to its properties.

1. Free Software:

Free software (or libre software is computer software distributed under terms that allow users to run the software for any purpose as well as to study, change, and distribute it and any adapted versions. Free software is a matter of liberty, not price; all users are legally free to do what they want with their copies of a free software (including profiting from them) regardless of how much is paid to obtain the program. Computer programs are deemed "free" if they give end-users (not just the developer) ultimate control over the software and, subsequently, over their devices.

The right to study and modify a computer program entails that source code—the preferred format for making changes—be made available to users of that program. While this is often called "access to source code" or "public availability", the Free Software Foundation (FSF) recommends against thinking in those terms, because it might give the impression that users have an obligation (as opposed to a right) to give non-users a copy of the program.

2. Open-source Software:

Open-source software is computer software whose source code is available openly on the internet and programmers can modify it to add new features and capabilities without any cost. Here the software is developed and tested through open collaboration. This software is managed by an open-source community of developers. It provides community support as well as commercial support if available for maintenance. We can get it for free of cost. This software also sometimes comes with a license and sometimes does not. This license provides some rights to users.

- Software can be used for any purpose
- Allows studying how the software works
- Freedom to modify and improve the program
- No restrictions on redistributions

Some examples of Open source software includes Android, Ubuntu, Firefox, Open Office etc.

3. Proprietary Software:

Proprietary software is computer software where the source codes are publicly not available only the company that has created can modify it. Here the software is developed and tested by the individual or organization by which it is owned not by the public. This software is managed by a closed team of individuals or groups that developed it. We have to pay to get this software and its commercial support is available for maintenance. The company gives a valid and authenticated license to the users to use this software. But this license put some restrictions on users also like.

- the number of installations of this software into computers
- Restrictions on sharing of software illegally
- Time period up to which software will operate
- Number of features allowed to use

S.No.	OPEN SOURCE SOFTWARE	PROPRIETARY SOFTWARE
01.	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.
02.	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.
03.	In open source software the source code is public.	In proprietary software the source code is protected.
04.	Open source software can be installed into any computer.	Proprietary software can be installed into any computer without valid license.
05.	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.
06.	Open source software is managed by an open source community of developers.	Proprietary software is managed by an closed team of individuals or groups that developed it.
07.	It is more flexible and provides more freedom which encourages innovation.	It is not much flexible so here is very limited innovation scope with the restrictions.
08.	Users can get open software for free of charge.	Users must have to pay to get the proprietary software.

09.	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.
10.	Examples are Android, Linux, Firefox, Open Office, GIMP, VLC Media player etc.	Examples are Windows, MacOS, Internet Explorer, Google earth

Q4. Using Anaconda Python create Histogram, Scatter plot and Bar plot for the dataset given below.

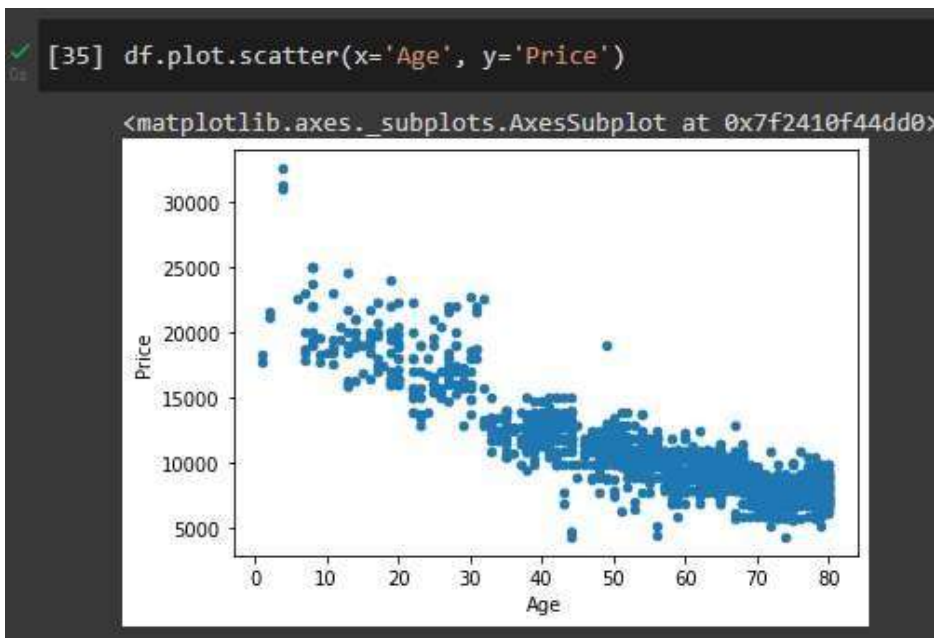
Dataset-

https://drive.google.com/file/d/1i11BZFe8Xj9kNq7eeE9KOa_lz1KhEdXJ/view

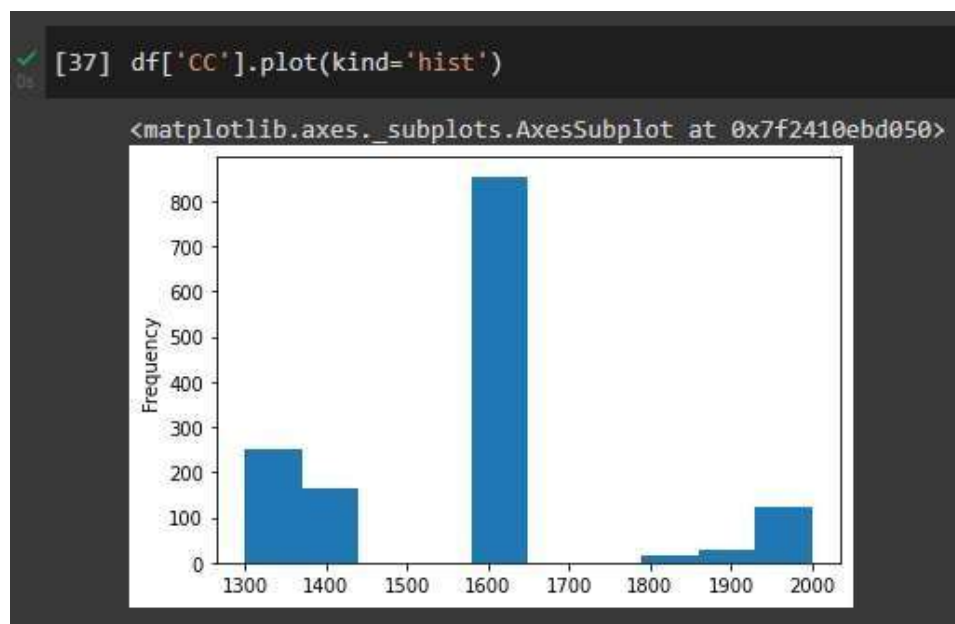
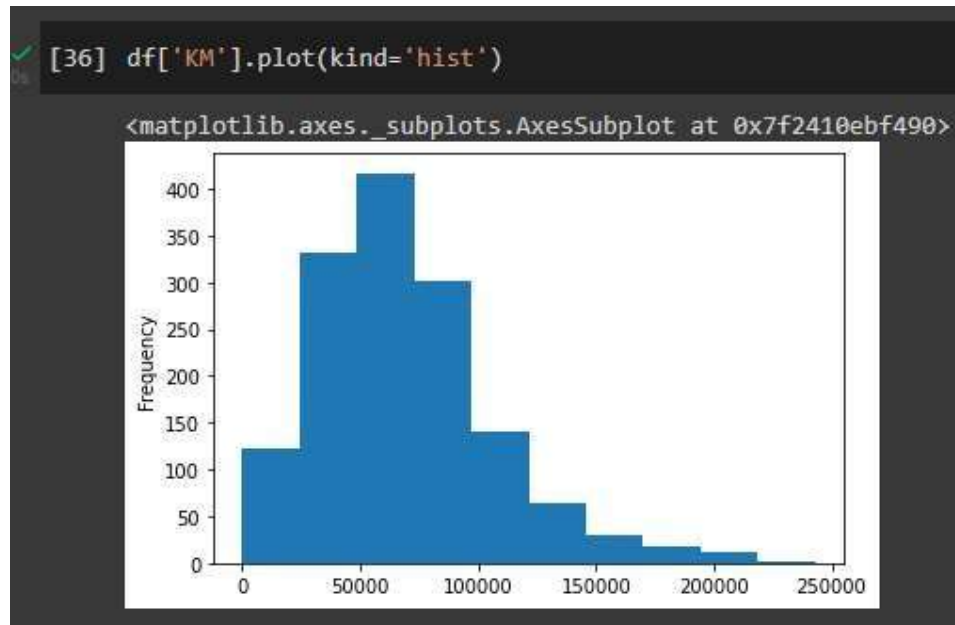
- Scatter plot- Scatter plot of Price Vs Age
- Histogram- for Kilometer and CC
- Bar plot- Bar plot for different fuel types

```
✓ [34] import pandas as pd  
      df = pd.read_csv('/content/toyota.csv')
```

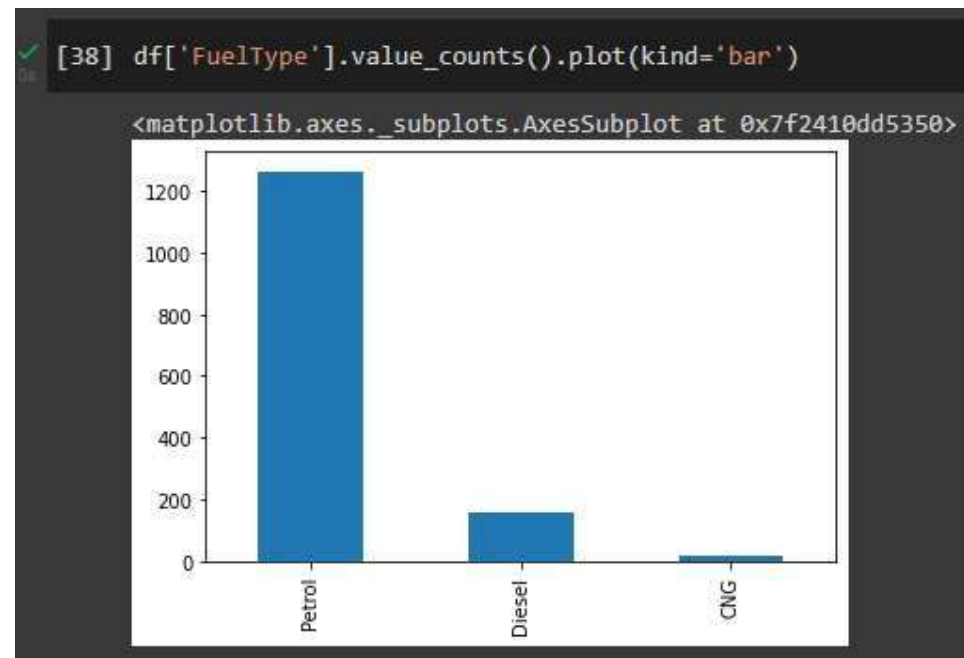
Scatter plot – Price vs Age



Histogram – Kilometer and CC



Bar Plot – Different Fuel Types



Q5. Enlist some examples along with its purpose and properties (at least 10) of FOSS and proprietary software with respect to database.

1. PostgreSQL

This relational database software has been around since 1997 and is the top choice in communities like Ruby, Python, Go, etc.

2. MariaDB

MariaDB was created as a replacement for MySQL by the same person who developed MySQL.

3. CockroachDB

The idea behind “cockroach” is that it’s an insect built for survival. No matter what happens — predators, floods, eternal darkness, rotting food, bombing, the cockroach finds a way to survive and multiply.

4. ClickHouse

It uses every hardware to its maximum potential to approach each query faster. The peak performance of processing a query usually remains more than two terabytes each second.

5. Neo4j

Support for transactional applications and graph analytics. Data transformation abilities for digesting large-scale tabular data into graphs. Specialized query language (Cypher) for querying the graph database Visualization and discovery features

6. Redis

When it comes to databases, it’s almost too easy to overlook the existence of Redis. That’s because Redis is an in-memory database and is mostly used in support functions like caching.

7. SQLite

SQLite is a lightweight C library that provided a relational database storage engine. Everything in this database lives in a single file (with a .sqlite extension) that you can put anywhere in your filesystem. And that’s all you need to use it! Yes, no “server” software to install and no service to connect to.

8. Cassandra

Cassandra belongs to what's known as the "columnar" family of databases. The storage abstraction in Cassandra is a column rather than a row. The idea here is to store all the data in a column physically together on the disk, minimizing seek time.

9. Timescale

The timescale is a type of what's called a "time series" database. It's different from a traditional database in that time is the primary axis of concern, and the analytics and visualization of massive data sets is a top priority

10. CouchDB

is a neat little database solution that sits quietly in a corner and has a small but dedicated following. It was created to deal with the problems of a net loss and eventual resolution of data, which happens to be a problem so messy that developers would instead switch jobs than deal with it

TASK:

Data Visualization softwares used in Seniors Company:

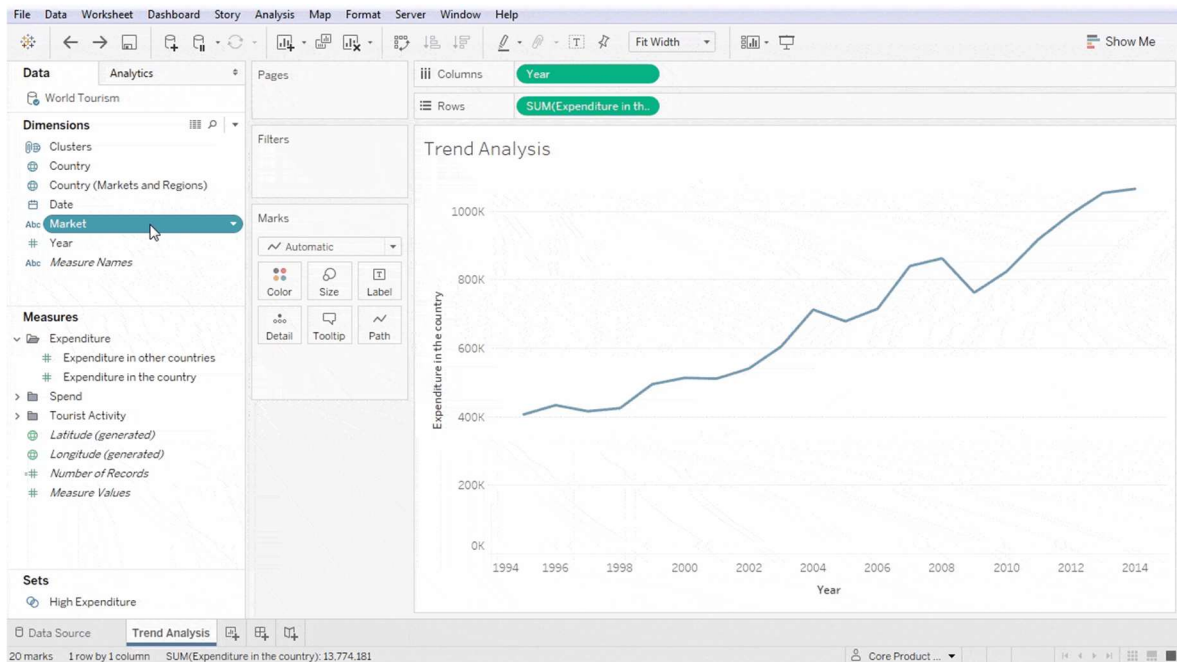
1. Name – Aditya Gadadhani
Company – Mindbody Inc.
Software – Tableau (Open Source Tool)

Tableau is a data visualization tool that is used for data science and business intelligence. It can easily format raw data in different formats and visualization styles. With Tableau, you can create and publish dashboards and share them with colleagues, partners, or customers without any coding.

Tableau is an innovator in the field of data visualization with its ease of use, stunning visualizations, vibrant community, and more contributing to success.

Tableau recognizes that the way we use data today and the field of analytics is much different now than even a decade ago. There are currently strong offerings in both closed and open source software, and with Tableau, you get the best of both options as developers. Tableau recognizes the importance of these open source tools in the data space. The tools themselves are impressive, and more importantly, they come with a large network of experts and enthusiasts that know how to leverage them. Rather than trying to develop similar tools, Tableau has opened up critical parts of its ecosystem to them. And this enables the large community of open source contributors to extend Tableau while allowing Tableau customers to best integrate the tools and technologies they choose to use.





2. Name – Megha Kesare

Company – Deloitte Touche Tohmatsu Limited

Software – QlikView

This visualization tool brought to you by Qlik, is a simple and easy tool that lets the user put business in total control. It lets you consolidate, search, visualize and analyze all the data sources for fetching useful business insights. QlikView is an easy way to get answers to the most critical business questions in quick time.

One of the highest used tools, it also ranks high in terms of customer loyalty, performance, features and quality. It offers integrated BI platform along with highly insightful demos, training manuals and tutorials, helping the users to get easily acquainted with the tool in no time. It comes with several components such as QlikView SERVER (QVS), QlikView Publisher (QVP).

