Class Assignment 1

Course Code: CAP447 **Set A**

Course Title: Data Warehousing And Mining Laboratory **Section: D2112**

Name: Jayshri lal Pandit **Roll No. : RD2112A103**

Reg. No. : 12111670

---

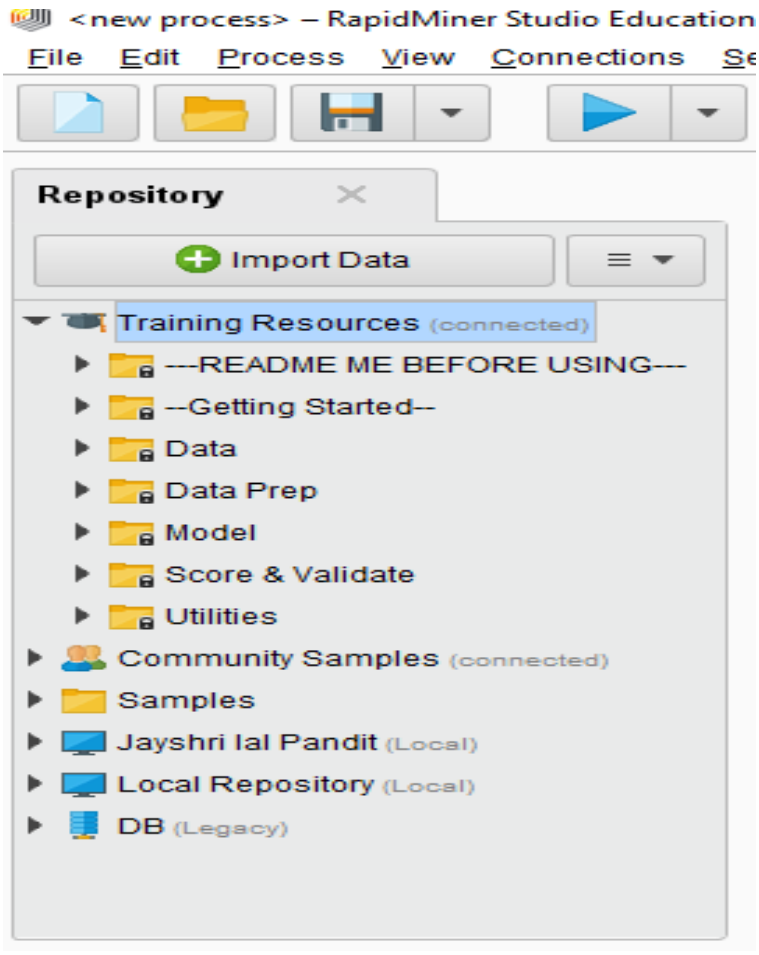
1. **Explain briefly the main components of RAPID MINER studio**

Ans: - The following are the main components of RAPID MINER Studio is explained briefly:-

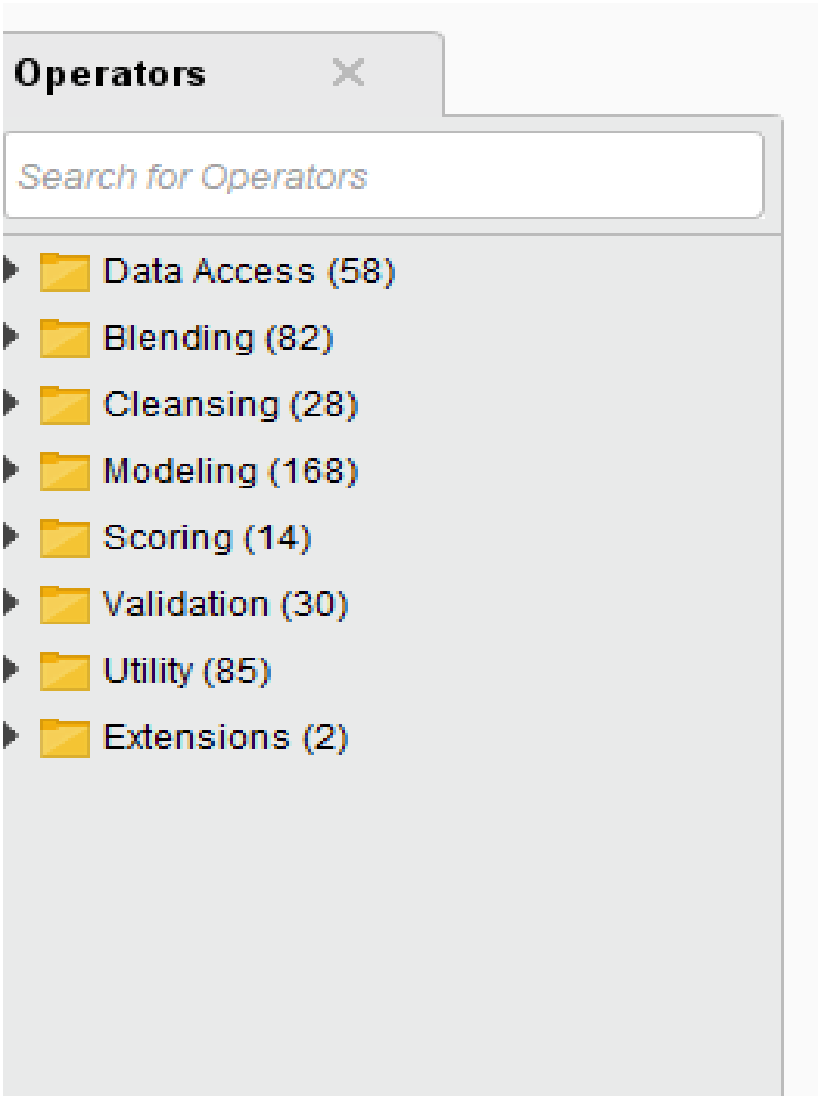
(i) **Repository:-**

A repository is simply a folder that holds all of our Rapid Miner data sets (we call them "Example Sets), processes, and other file objects that you will create using Rapid Miner Studio. This folder can be stored locally on your computer, or on a Rapid Miner Server.

When we start Rapid Miner Studio for the first time, it will automatically create a "Local Repository" for us. If we want to see where it is, just go to our Rapid Miner folder on our computer, go into "repositories", and then into "Local Repository". We should see two folders here: 'data' and 'processes'. These correspond EXACTLY to what we see in the Repository panel in Rapid Miner Studio.



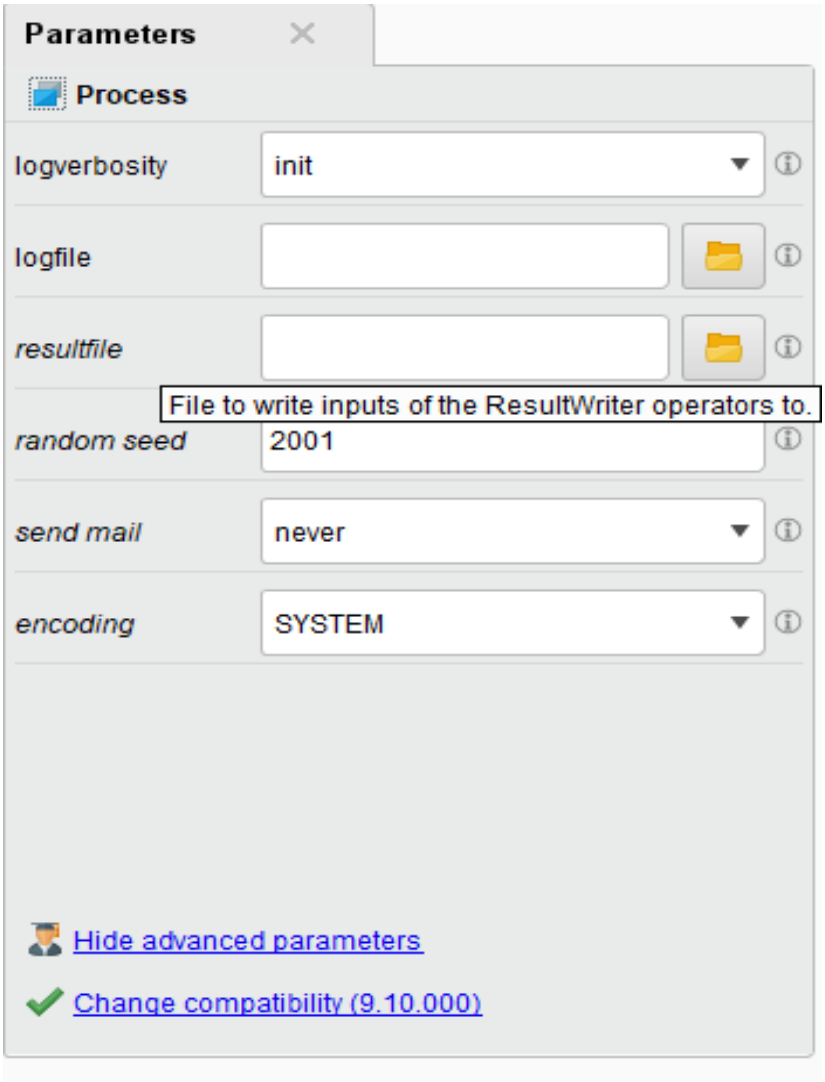
(ii) **Operators:-**The building blocks, grouped by function, used to create rapid Miner processes. An *operator* has input and output ports. The action performed on the input ultimately leads to what is supplied to the output. Operator parameter control those actions. There are more than 1500 operators available in rapid Miner. Operators, in the **Operators** panel of the **Design** view, are both browsable and searchable.



(iii) **Parameters:-**

The settings whose values determine the characteristics or behaviour of an operator. Rapid Miner presents parameters in the **Parameters** panel of the **Design** view. There are regular parameters and expert parameters. The expert parameters are indicated by italic names and are displayed or hidden by clicking the **Show/Hide advanced parameters** link at the bottom of the panel.

As part of the Wisdom of Crowds capabilities, Rapid Miner Studio provides parameter recommendations based on the knowledge and best practices of other Rapid Miner users. The recommender helps configure operators by providing recommendations on which parameters to change and by suggesting appropriate parameter values



etc.

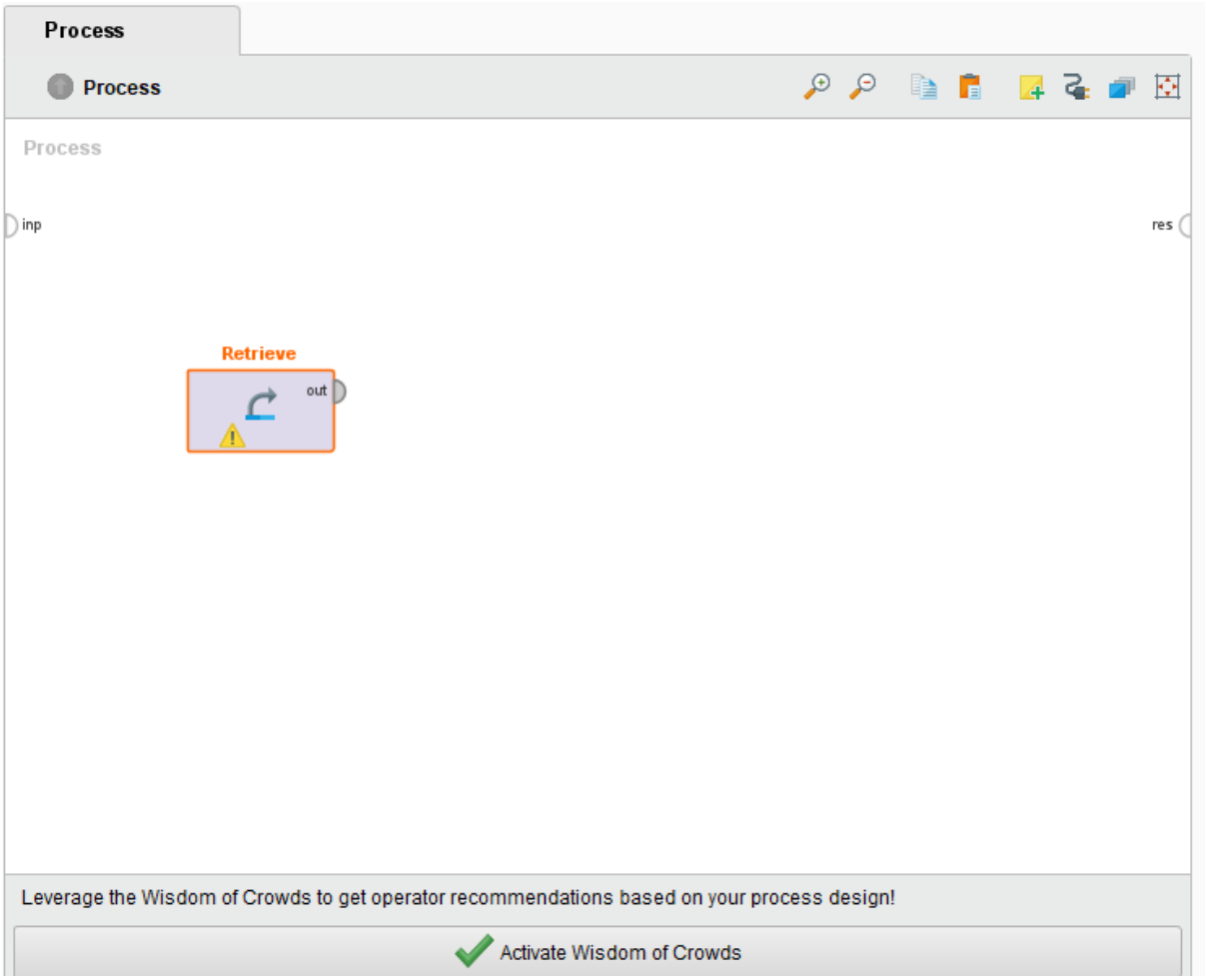
2.

a. Explain any 5 operators along with its usage and snapshots in rapid miner.

Ans: - Following are any 5 operators along with its usage explained and snapshots in rapid miner:-

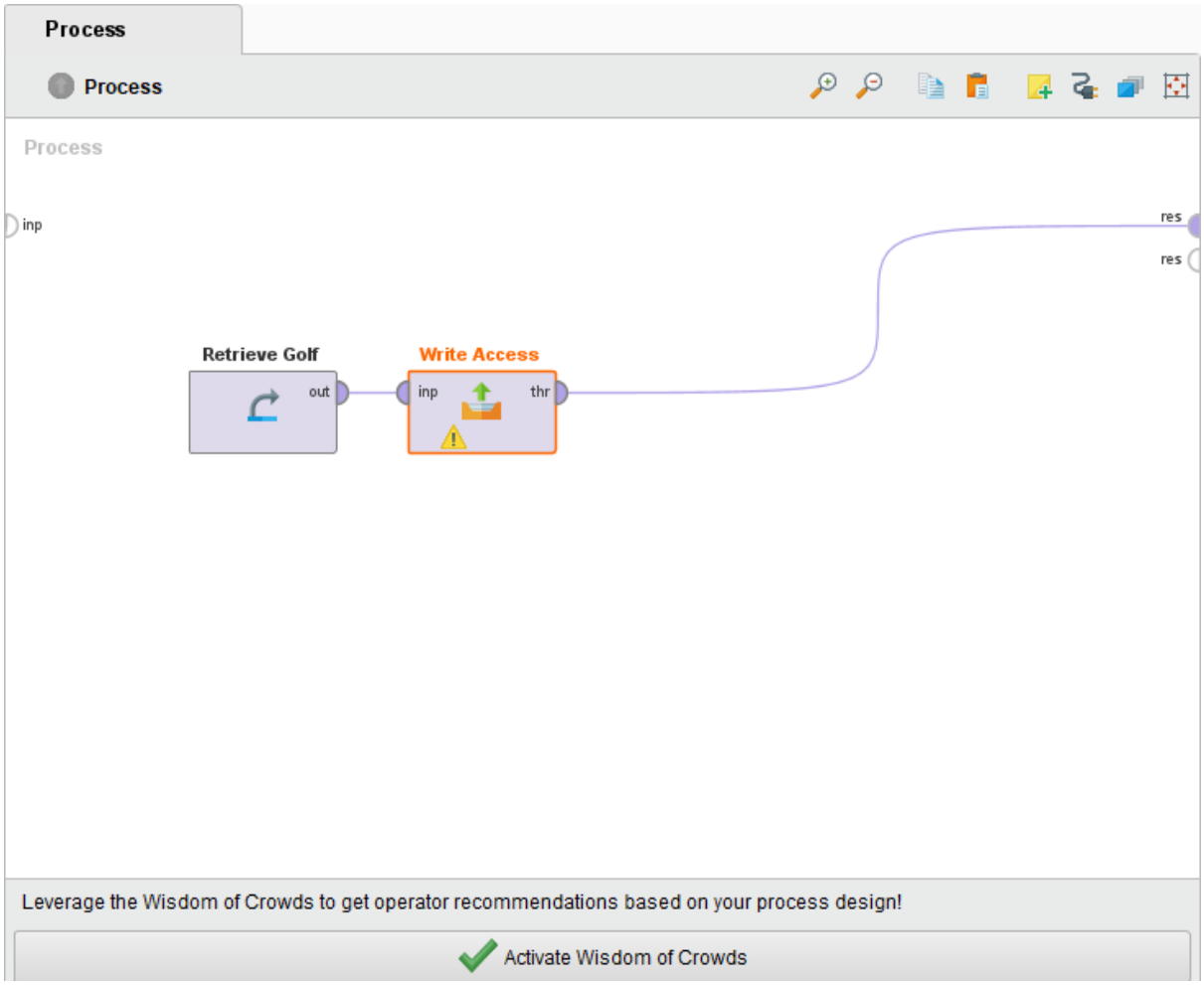
(i)**Retrieve: -** The Retrieve Operator loads a Rapid Miner Object into the Process. This Object is often an Example Set but it can also be a Collection or a Model. Retrieving data this way also provides the meta data of the Rapid Miner Object.

This Operator is like the different Read Operators in the Data Access group. Storing the data inside a repository gives one the advantage that meta data properties are stored as well. Meta data gives you additional information about the Rapid Miner Object you retrieve. For an ExampleSet this is e.g. the names and types of Attributes, their range and how many missing values there are. Meta data allows you to easily configure parameters of other Operators, for example you can select Attributes from a list of available Attributes. The data stored in the Repository can only be changed within a Rapid Miner Process. Data stored on disk or within database can be changed by other means.



(ii)Write Access:- The Write Access operator is used for writing an ExampleSet into the specified Microsoft Access database. We0 only need to have a basic understanding of databases in order to use this operator properly. Please go through the parameters and the attached Example Process to understand the working of this operator.

Input (inp) this input port expects an ExampleSet. It is output of the Retrieve operator in the attached Example Process.

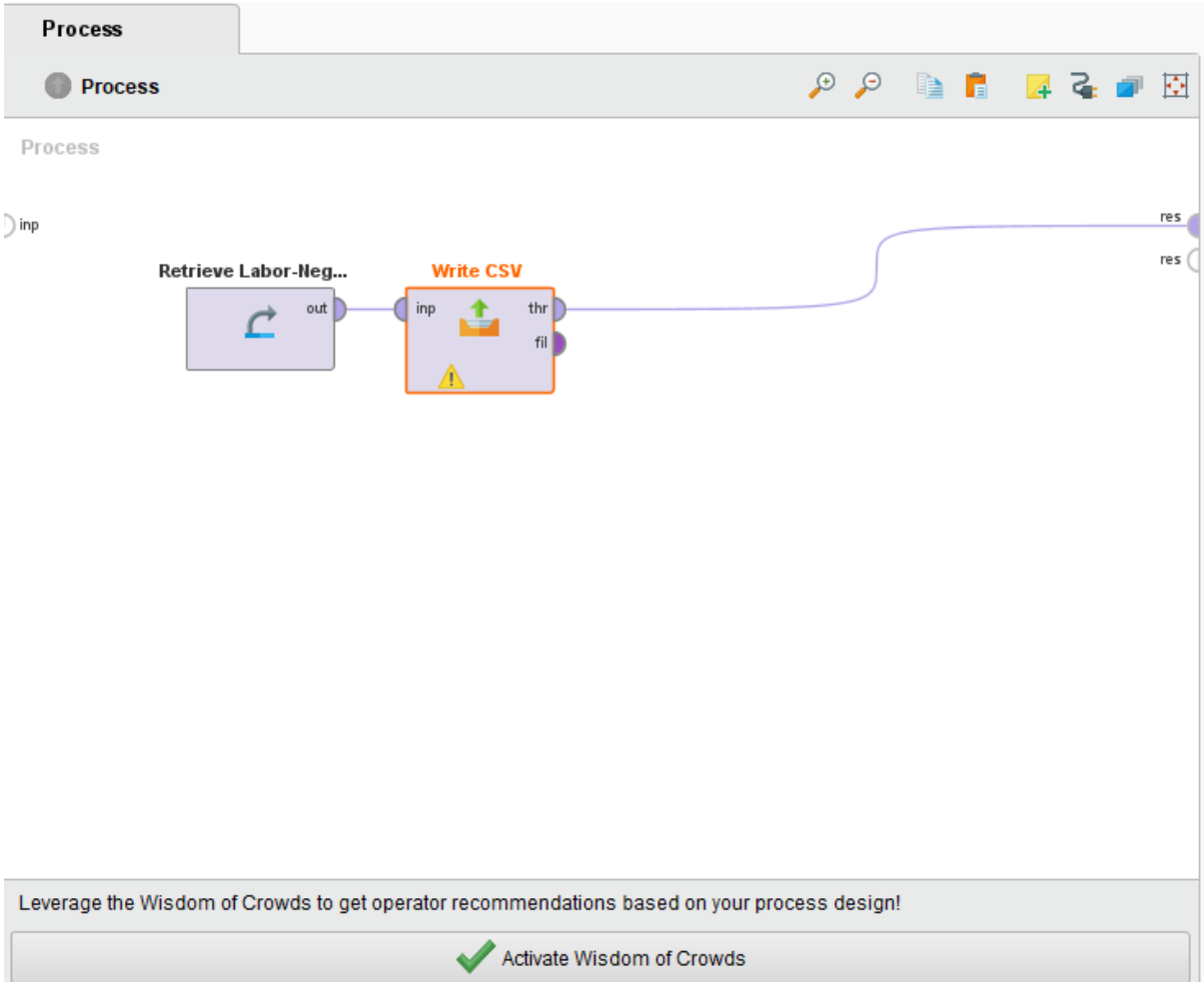


(iii)Write CSV:-

A comma-separated values (CSV) file stores tabular data (numbers and text) in plain-text form. CSV files have all values of an example in one line. Values for different attributes are separated by a constant separator. It may have many rows. Each row uses a constant separator for separating attribute values. The name suggests that the attributes values would be separated by commas, but other separators can also be used. This separator can be specified using the column separator parameter. Missing data values are indicated by empty cells.

Input (inp) This input port expects an ExampleSet. It is output of the Retrieve operator in the attached Example Process.

Through (thr) The ExampleSet that was provided at the input port is delivered through this output port without any modifications. This is usually used to reuse the same ExampleSet in further operators of the process.



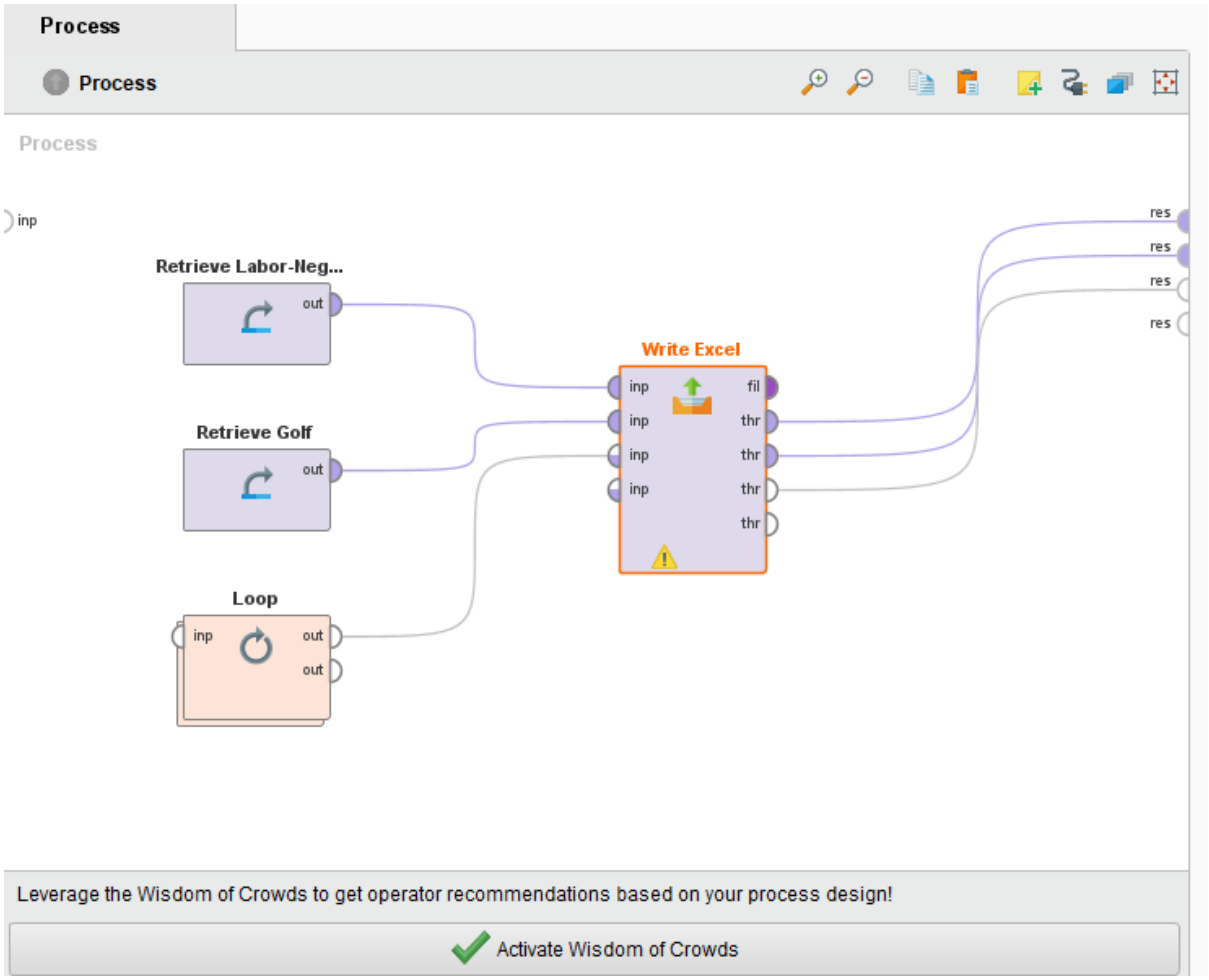
(iv)Write Excel:-

The Write Excel operator can be used for writing ExampleSets into a Microsoft Excel spreadsheet file. This operator creates Excel files that are readable by Excel 95, 97, 2000, XP, 2003 and newer versions. Missing data values in the ExampleSet are indicated by empty cells in the Excel file. The first row of the resultant Excel file has the names of attributes of the input ExampleSet. Files written by the Write Excel operator can be loaded in RapidMiner using the Read Excel operator. Multiple input ExampleSets can be specified and this will lead to one excel file with multiple sheets. By default the sheets will be named after their corresponding ExampleSet's source. Sheet names can optionally be specified using the sheet names parameter.

Input (inp) This input port expects an ExampleSet or a Collection of ExampleSets. The specified ExampleSets will become sheets in the resulting Excel file. The Write Excel operator can have multiple inputs. When one input is connected, another input port becomes available which is ready to accept another input (if any). The order of inputs remains the same in the resulting Excel file.

file (fil) The created Excel file is provided as a file object that can be used with other operators with file input ports like 'Write File'.

through (thr) The ExampleSet or collection of ExampleSets that was provided at the corresponding input port is delivered through this output port without any modifications. This is usually used to reuse the same ExampleSets in further operators of the process.

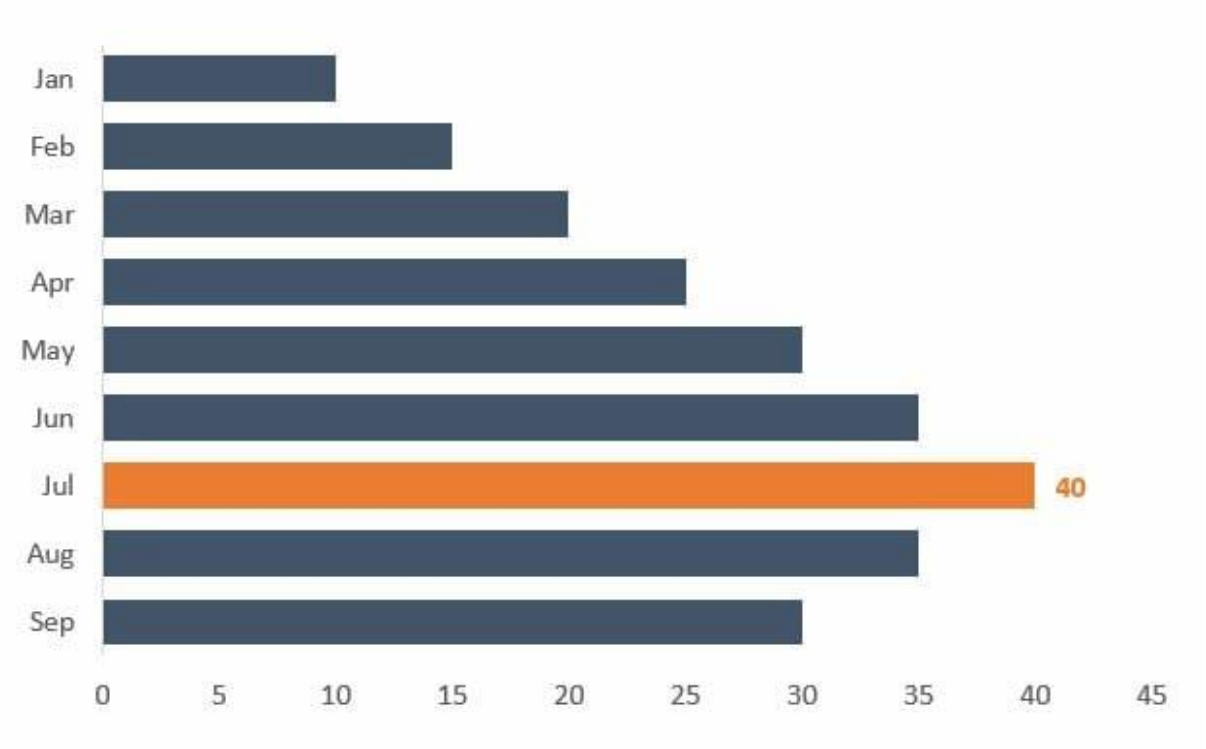


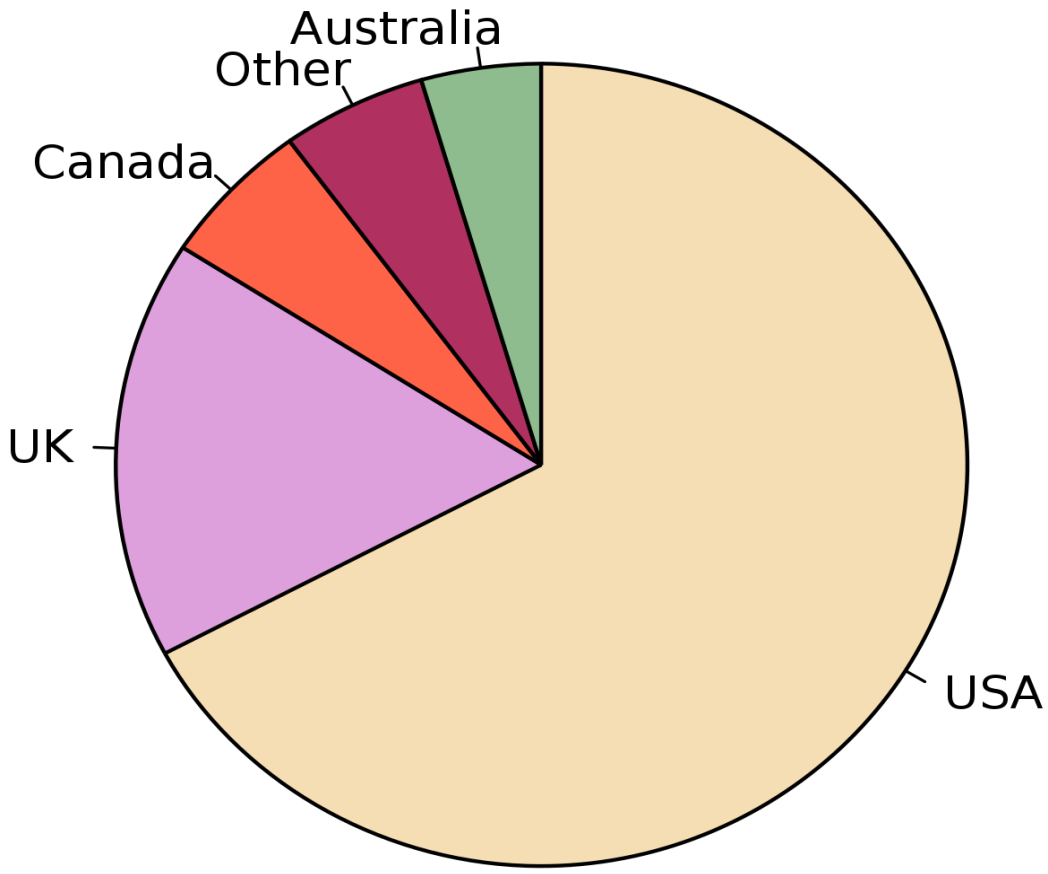
b. What do you understand by graphical representation and statistics in rapidminer. Attach any 5 different types of graph and their interpretation

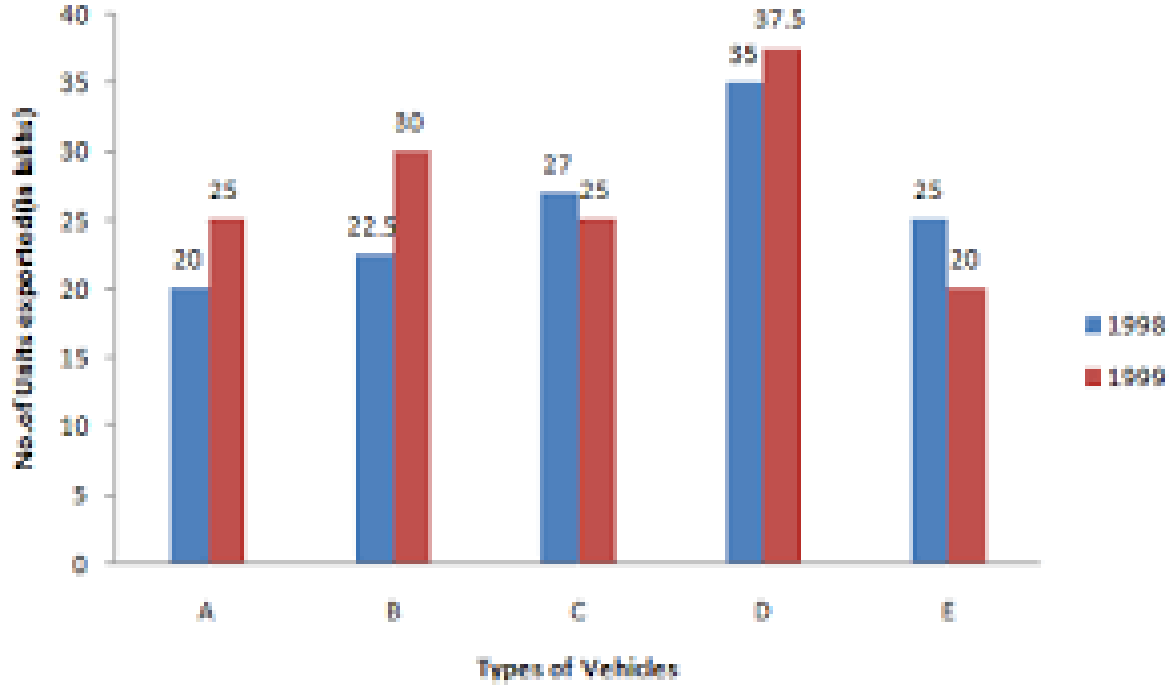
Ans:- Graphical Representation is a way of analysing numerical data. It exhibits the relation between data, ideas, information and concepts in a diagram. It is easy to understand and it is one of the most important learning strategies. It always depends on the type of information in a particular domain. There are different types of graphical representation. Some of them are as follows:

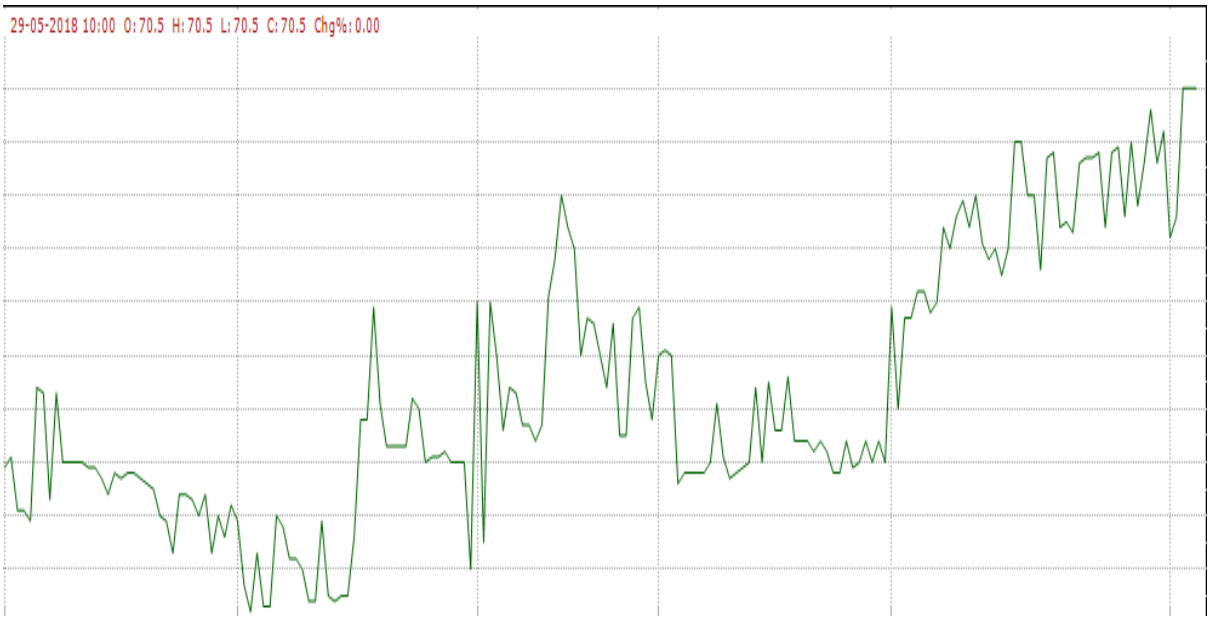
- **Line Graphs** – Line graph or the linear graph is used to display the continuous data and it is useful for predicting future events over time.
- **Bar Graphs** – Bar Graph is used to display the category of data and it compares the data using solid bars to represent the quantities.
- **Histograms** – The graph that uses bars to represent the frequency of numerical data that are organised into intervals. Since all the intervals are equal and continuous, all the bars have the same width.
- **Line Plot –** It shows the frequency of data on a given number line. ‘ x ‘ is placed above a number line each time when that data occurs again.
- **Frequency Table** – The table shows the number of pieces of data that falls within the given interval.

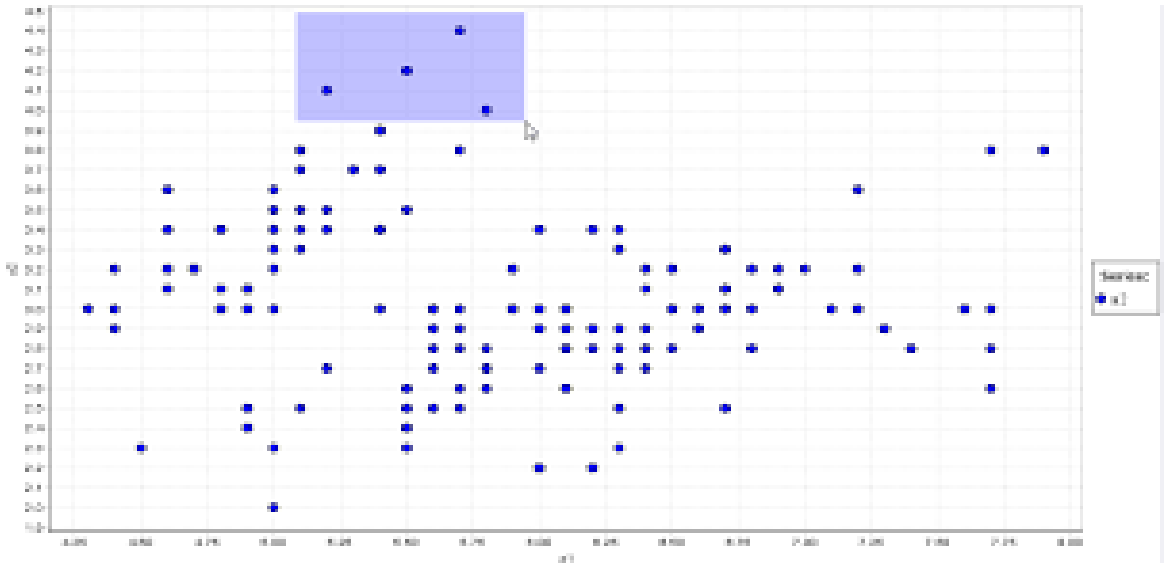
- **Circle Graph** – Also known as the pie chart that shows the relationships of the parts of the whole. The circle is considered with 100% and the categories occupied is represented with that specific percentage like 15%, 56%, etc.
- **Stem and Leaf Plot –** In the stem and leaf plot, the data are organised from least value to the greatest value. The digits of the least place values from the leaves and the next place value digit forms the stems.
- **Box and Whisker Plot** – The plot diagram summarises the data by dividing into four parts. Box and whisker show the range (spread) and the middle (median) of the data.






40
35
30
25
20
15
10
5
0
20
25
22.5
30
27
25
35
37.5
25
20
A
B
C
D
E
1998
1999
Types of Vehicles


29-05-2018 10:00 O:70.5 H:70.5 L:70.5 C:70.5 Chg%:0.00



3. What do you mean by preprocessing in Data Mining? Explain any 5 operations to handle missing values and attach screenshots of same.

Ans: - Data preprocessing is a **data** mining technique that involves transforming raw **data** into an understandable format. Real-world **data** is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and is likely to contain many errors. **Data preprocessing** is a proven method of resolving such issues.

In the real world **data** are generally incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate **data**. Noisy: containing errors or outliers. Inconsistent: containing discrepancies in codes or names.