Assignment 1

Name – Akash Raj **Subject code – CAP 447**

Roll no – RD2112B78 **Section – D2112**

Reg no – 12108382 **Submitted to – Poonam Rattan Mam**

---

1. Explain briefly the main components of Rapid miner.

Ans- The main components of rapid minor are –

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

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

iii) Parameters - The setting(s) whose value(s) determine the characteristics or behavior of an operator. RapidMiner presents parameters in the Parameter 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.

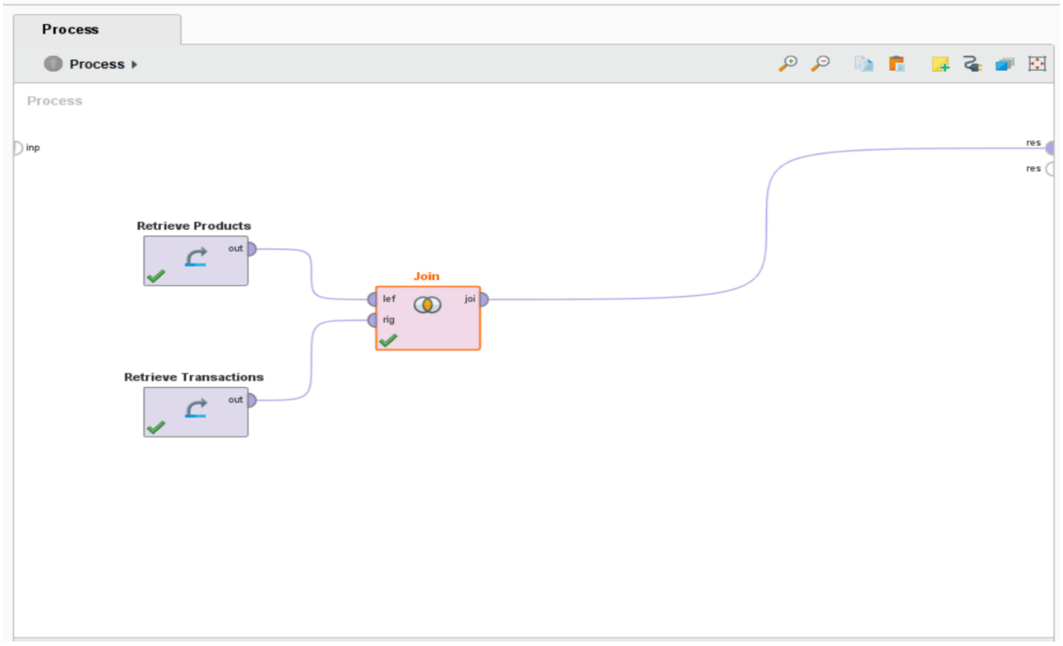
iv) Help – It gives the definition and description of the operators or the attributes which are being used.

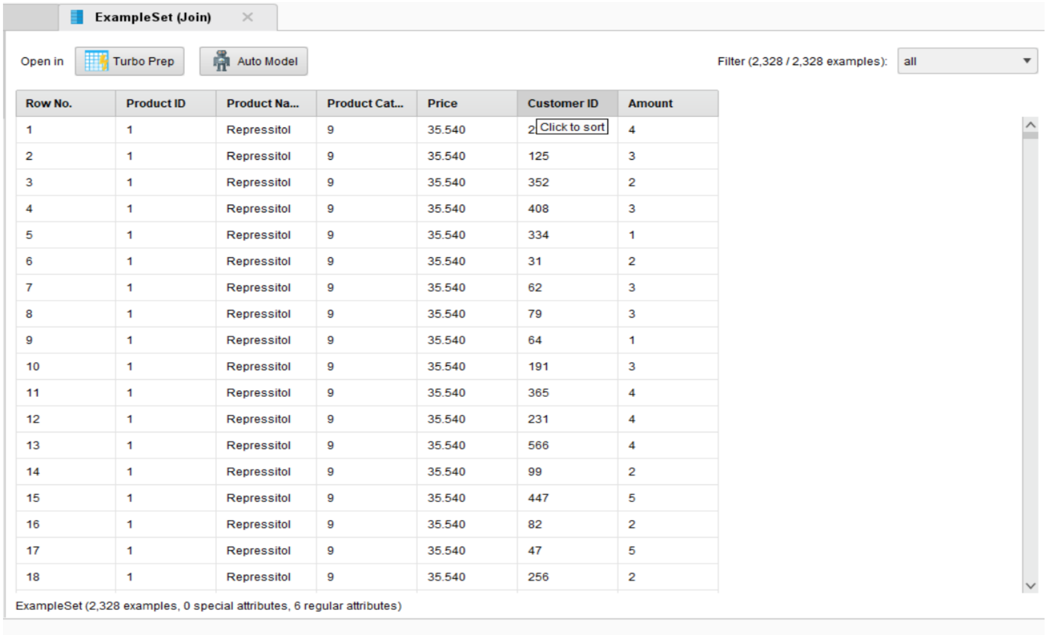
v) Views – This section consists of few important elements which are used in rapid miner. It consist of design, result, turboprop, auto model and deployment. The design and result options are used to switch between the visualization part and the result part.

2. a) Explain any 5 operators with its usage with its snapshot in rapid miner.

Ans- Five operators with its usage are –

1) **Join Operator –** The join operator is used to combine two or more data sets into a single using one or more attributes of the given data set known as key attributes.



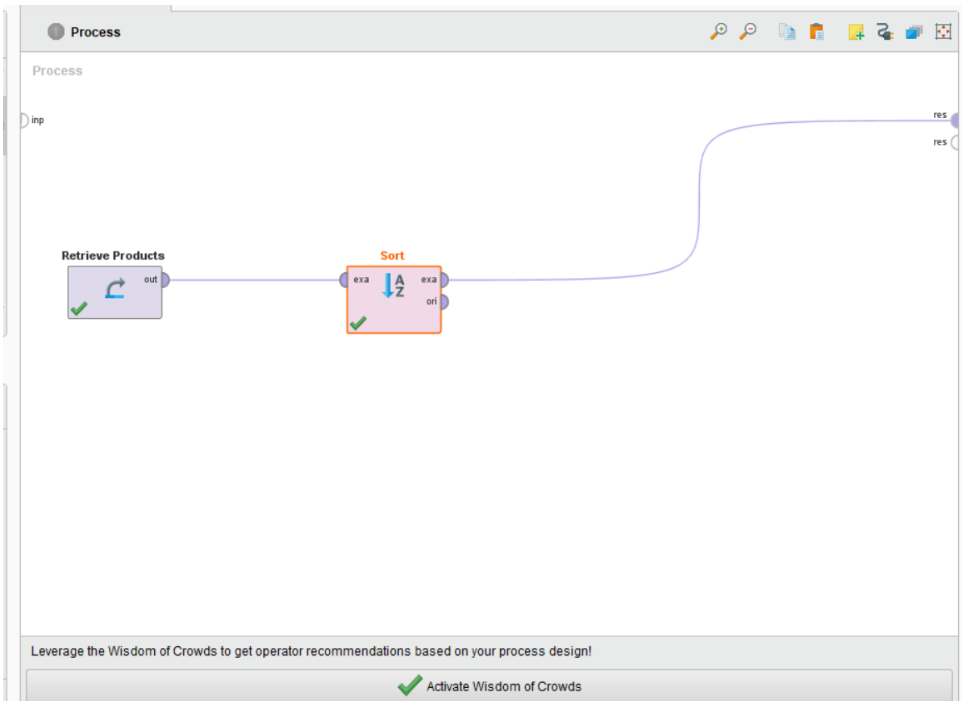


Row No.	Product ID	Product Na...	Product Cat...	Price	Customer ID	Amount
1	1	Repressitol	9	35.540	2	4
2	1	Repressitol	9	35.540	125	3
3	1	Repressitol	9	35.540	352	2
4	1	Repressitol	9	35.540	408	3
5	1	Repressitol	9	35.540	334	1
6	1	Repressitol	9	35.540	31	2
7	1	Repressitol	9	35.540	62	3
8	1	Repressitol	9	35.540	79	3
9	1	Repressitol	9	35.540	64	1
10	1	Repressitol	9	35.540	191	3
11	1	Repressitol	9	35.540	365	4
12	1	Repressitol	9	35.540	231	4
13	1	Repressitol	9	35.540	566	4
14	1	Repressitol	9	35.540	99	2
15	1	Repressitol	9	35.540	447	5
16	1	Repressitol	9	35.540	82	2
17	1	Repressitol	9	35.540	47	5
18	1	Repressitol	9	35.540	256	2

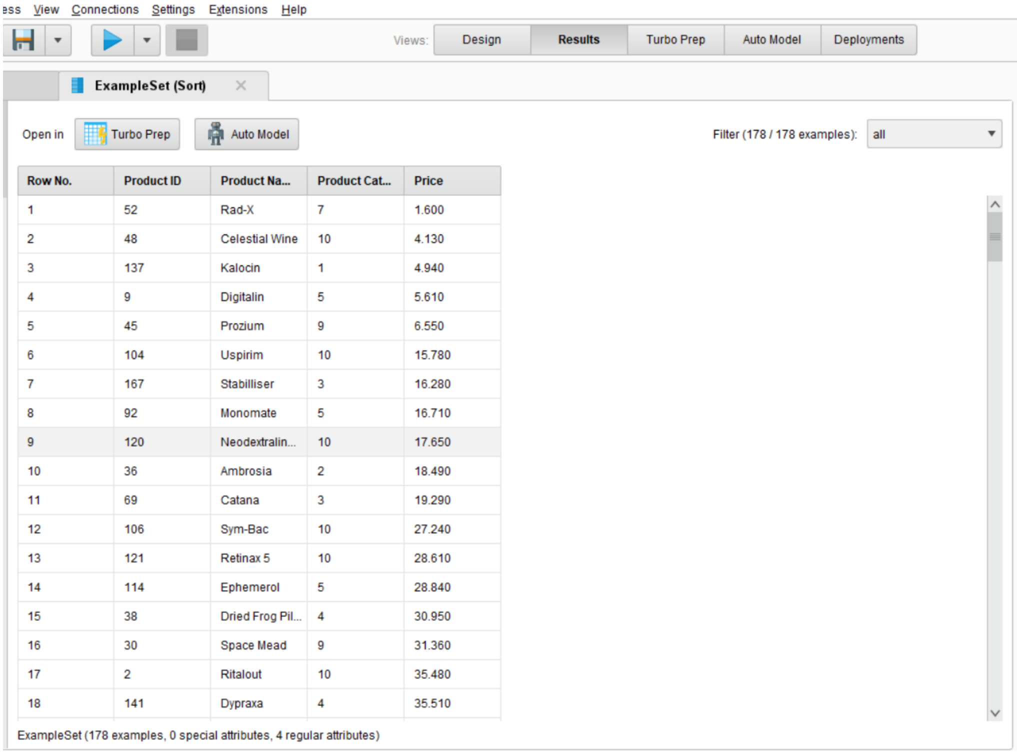
ExampleSet (2,328 examples, 0 special attributes, 6 regular attributes)

Product and transaction are combined using join operator

2) **Sort Operator –** The sort operator is used to sort the data set provided at the input port. The complete data set is sorted according to a single or more attributes. The sorting is done in either ascending or descending order.



Process
Process
inp
res
res
Retrieve Products
out
Sort
exa
exa
ori
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds



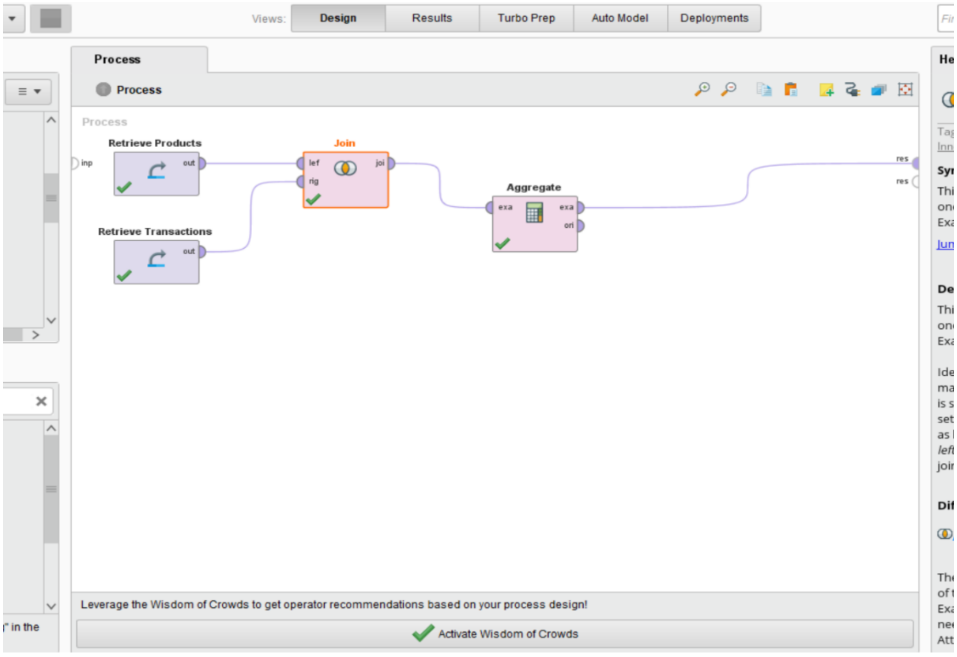
ess View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
ExampleSet (Sort)
Open in Turbo Prep Auto Model
Filter (178 / 178 examples): all

Row No.	Product ID	Product Na...	Product Cat...	Price
1	52	Rad-X	7	1.600
2	48	Celestial Wine	10	4.130
3	137	Kalocin	1	4.940
4	9	Digitalin	5	5.610
5	45	Prozium	9	6.550
6	104	Uspirim	10	15.780
7	167	Stabilliser	3	16.280
8	92	Monomate	5	16.710
9	120	Neodextralin...	10	17.650
10	36	Ambrosia	2	18.490
11	69	Catana	3	19.290
12	106	Sym-Bac	10	27.240
13	121	Retinax 5	10	28.610
14	114	Ephemerol	5	28.840
15	38	Dried Frog Pil...	4	30.950
16	30	Space Mead	9	31.360
17	2	Ritalout	10	35.480
18	141	Dypraxa	4	35.510

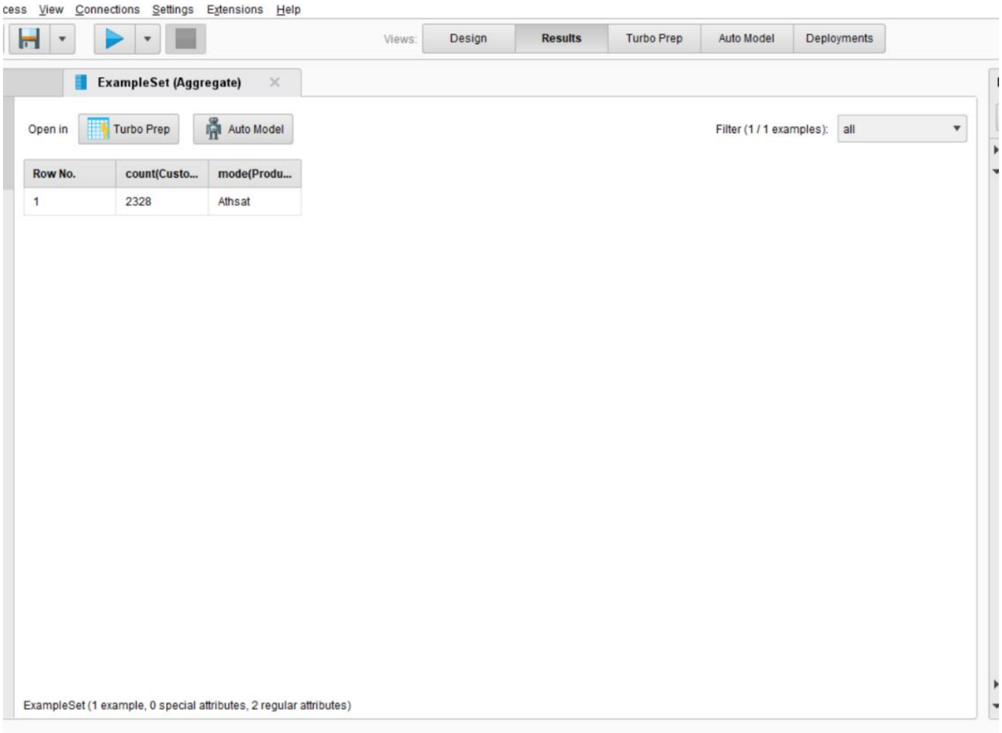
ExampleSet (178 examples, 0 special attributes, 4 regular attributes)

The price of the product dataset has been sorted using sort operator in ascending order.

3) **Aggregate operator –** The aggregate operator performs the aggregation function known from SQL. This operator provides a lot of functionalities in the same format as provided by the SQL aggregation function. It performs mathematical operations like average, count, aggregate, Max, Min and Sum.



Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

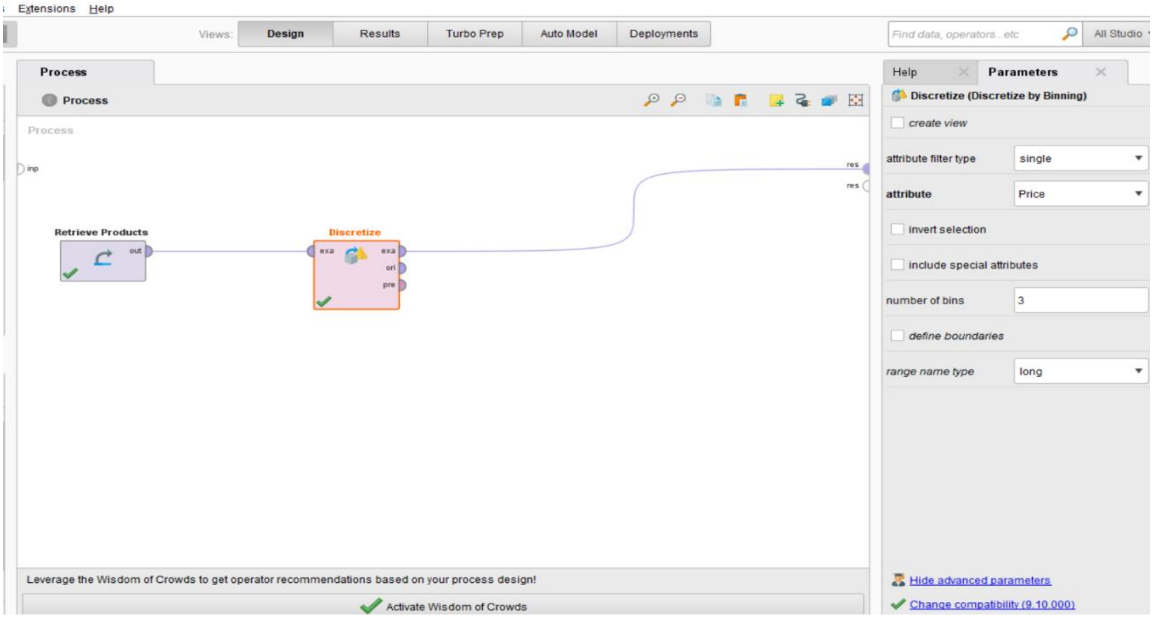


Row No.	count(Custo...	mode(Produ...
1	2328	Athsat

ExampleSet (1 example, 0 special attributes, 2 regular attributes)

The Count and mode functions are used from the Aggregate function

4) **Discretize By Binning -** This operator discretizes the selected numerical attributes to nominal attributes. The number of bins parameter is used to specify the required number of bins. This discretization is performed by simple binning. The range of numerical values is partitioned into segments of equal size.


Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Process
Process
inp
res
res
Retrieve Products
out
Discretize
exa
exa
ori
pre
Help
Parameters
Hide advanced parameters
Change compatibility (9.10.000)
Activate Wisdom of Crowds

Discretize (Discretize by Binning)

☐ create view

attribute filter type	single
attribute	Price

☐ invert selection

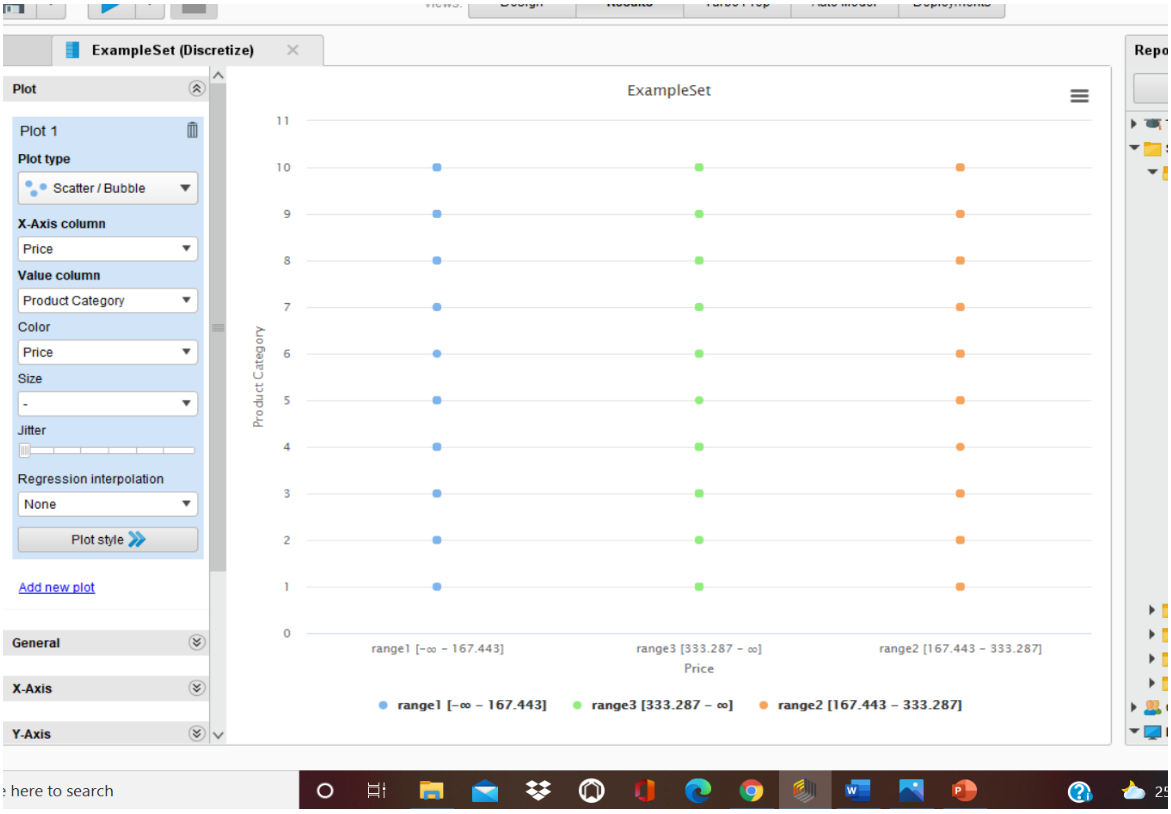
☐ include special attributes

number of bins	3

☐ define boundaries

range name type	long

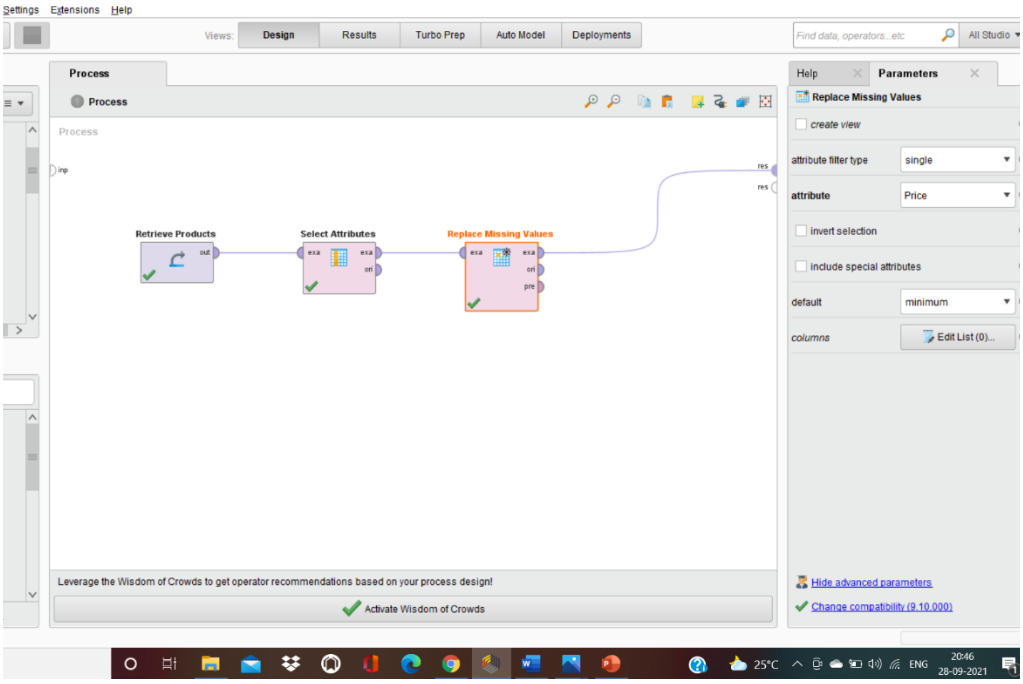
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!



Plot type	Scatter / Bubble
X-Axis column	Price
Value column	Product Category
Color	Price
Size	-
Jitter	
Regression interpolation	None

<u>The price has been discretized using the Discretize by binning method</u>

5) **Replace Missing Values –** This operator is use to replace missing value in example of selected attributes by a specified replacement. The missing values can be replaced by minimum, maximum or average value of that attribute. Zero can also be used to replace the missing values. Any replenishment value can also be specified as the replacement of the value.



Replace Missing Values

□ create view

attribute filter type: single

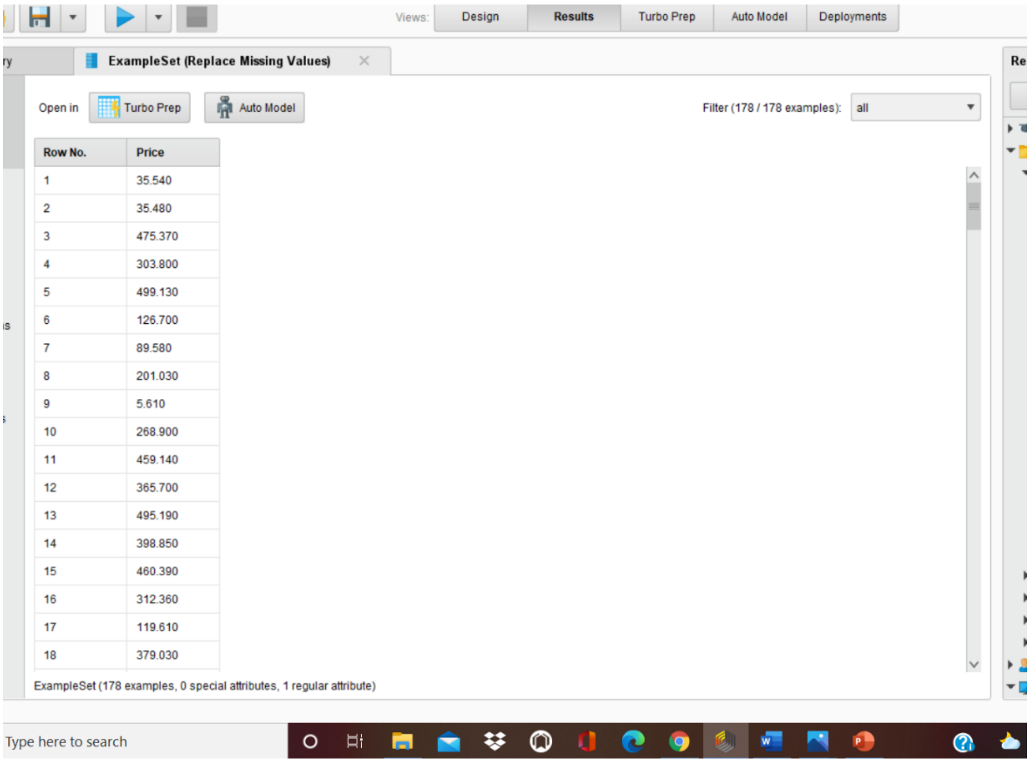
attribute: Price

□ invert selection

□ include special attributes

default: minimum

columns: Edit List (0)...



Filter (178 / 178 examples): all

Row No.	Price
1	35.540
2	35.480
3	475.370
4	303.800
5	499.130
6	126.700
7	89.580
8	201.030
9	5.610
10	268.900
11	459.140
12	365.700
13	495.190
14	398.850
15	460.390
16	312.360
17	119.610
18	379.030

ExampleSet (178 examples, 0 special attributes, 1 regular attribute)

Replacing missing values of price

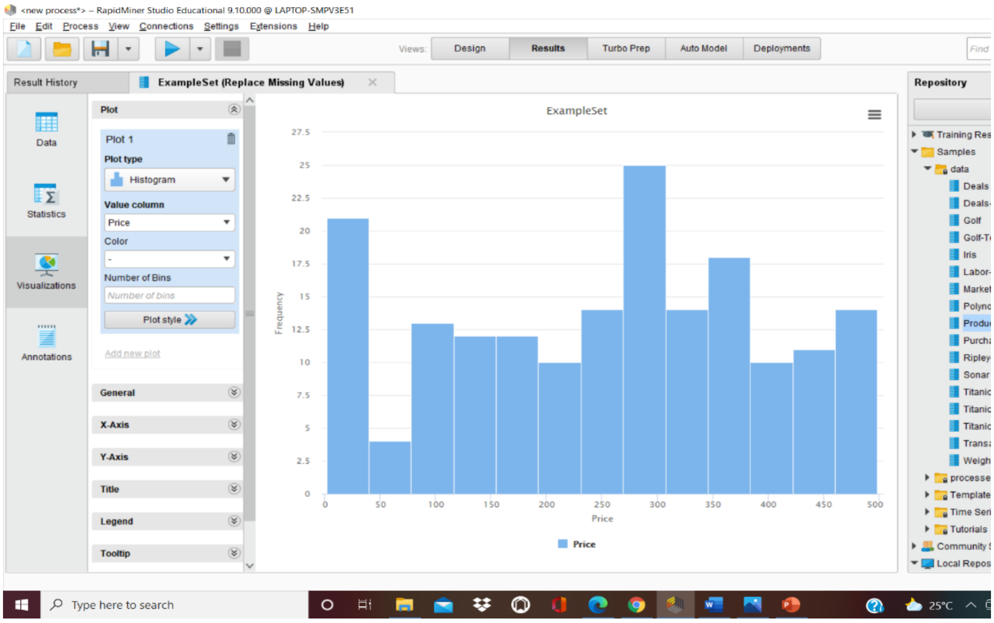
b. What do you understand by graphical representation and statistics in rapid miner. Attach any 5 different types of graphs and their interpretation.

Ans- Graphical representation is a method to show and represent values, increases, decreases, comparisons. It is the visual representation of data through charts and graphs. It helps to make prediction or show report of how certain situation was yesterday and now.

Statistics provides meaning to the data represented by the graphs.

Five different types of graphs are-

i) **Histogram –** Histogram is a non-cumulative frequency graph. It is drawn on a natural scale in which the representative frequencies of the different class of values are represented through vertical rectangles drawn closed to each other. Measure of central tendency, mode can be easily determined with the help of this graph.



Plot 1

Plot type: Histogram

Value column: Price

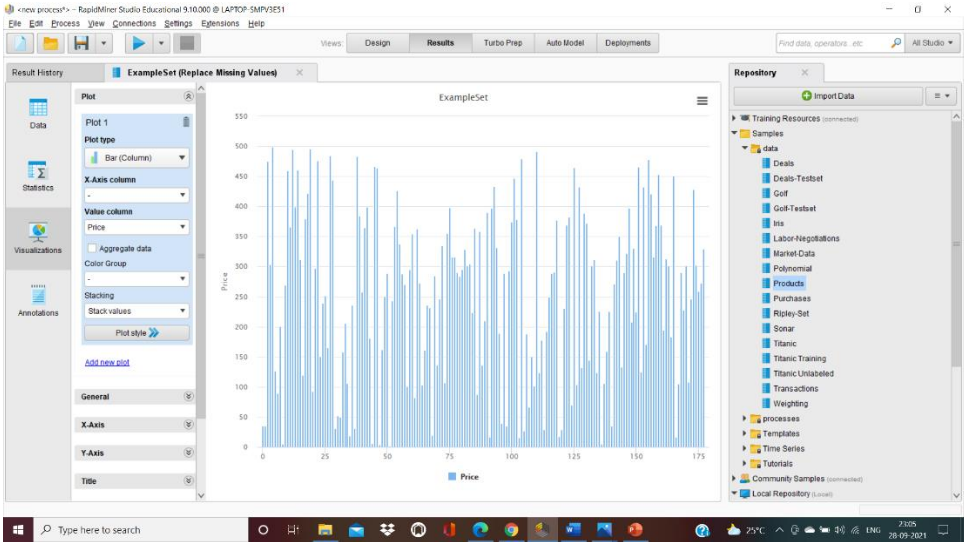
Color: -

Number of Bins: Number of bins

Plot style

A Histogram

ii) **Bar-Chart –** A bar chart or bar graph is a chart or a graph that presents categorical data with rectangular bars with height or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar graph is sometimes called a line graph.



Plot type: Bar (Column)

X-Axis column: -

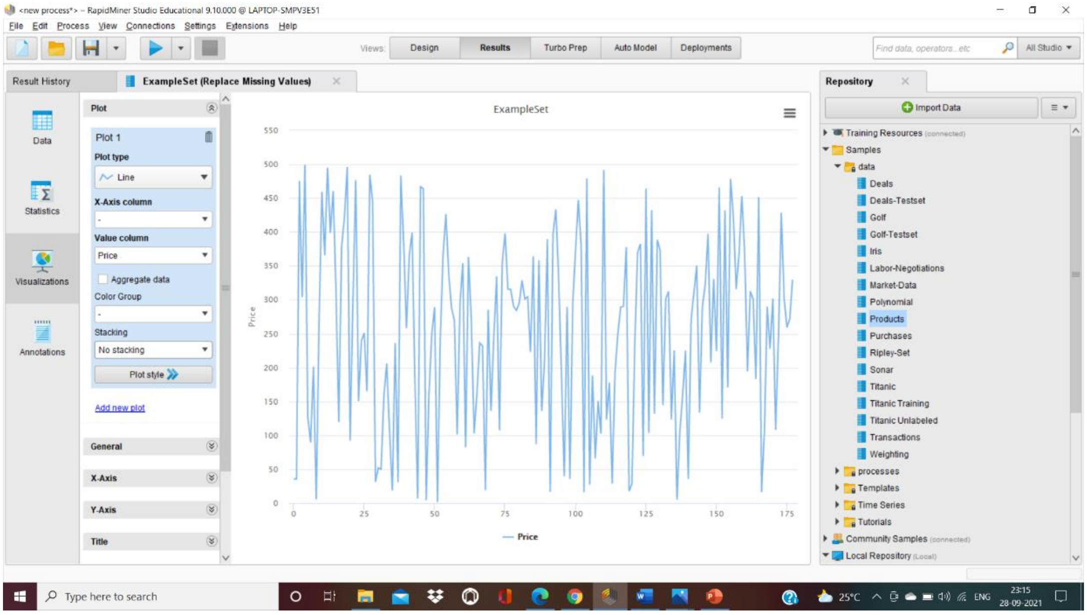
Value column: Price

Color Group: -

Stacking: Stack values

A Vertical Bar Graph

iii) **Line Graph –** A line graph is a type of chart which displays information as a series of data points called 'markers' connected by a straight line segments.



Plot type: Line

X-Axis column: -

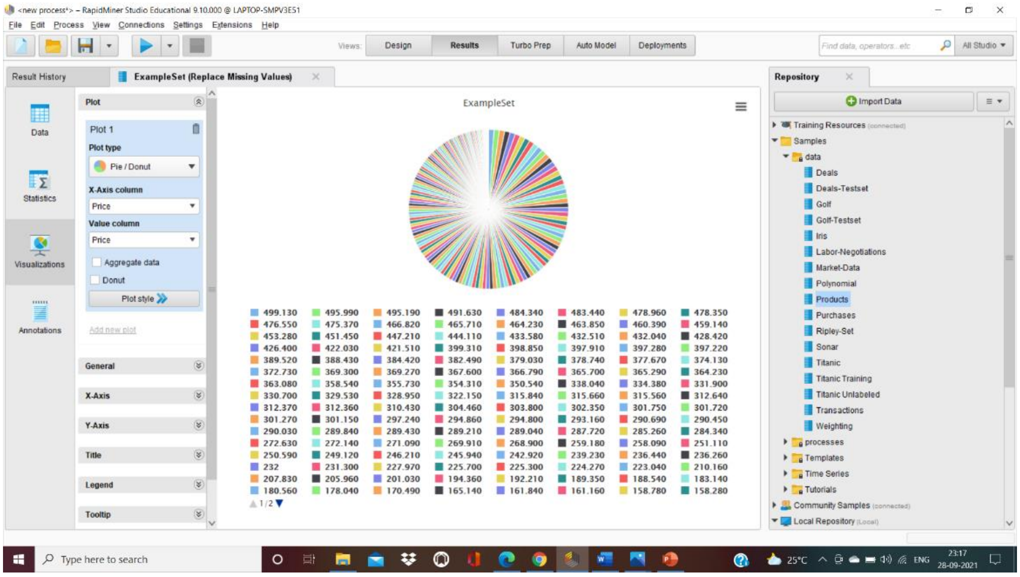
Value column: Price

Color Group: -

Stacking: No stacking

A Line Graph

iv) **Pie-Chart -** A pie chart is a circular statistical graph, which is divided into slices to illustrate numerical proportion. In this, the arc length of each slice, is proportional to the quantity it represents.



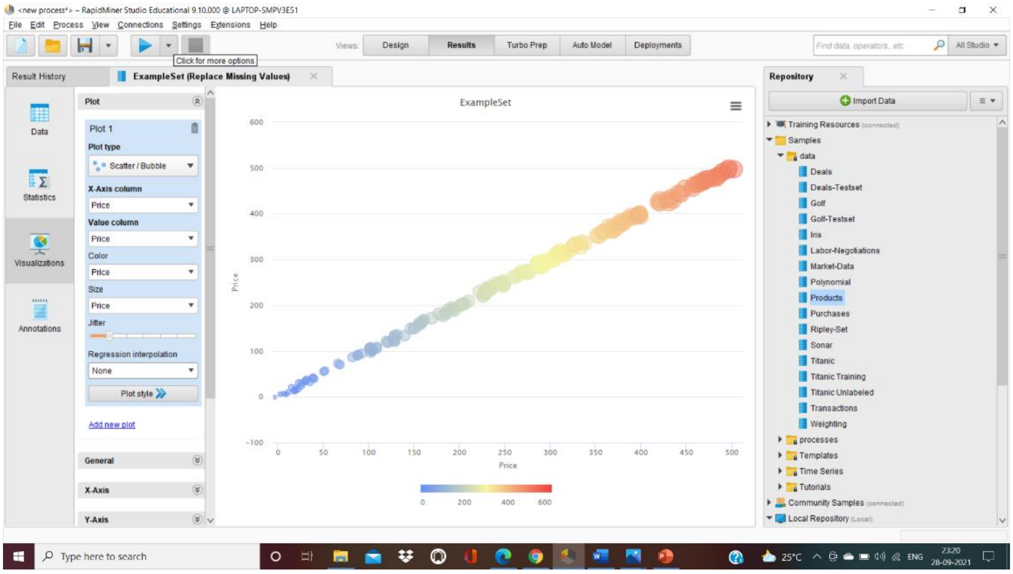
Plot type: Pie / Donut

X-Axis column: Price

Value column: Price

A Pie Chart

v) **Scatter plot -** A scatter plot uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.



Plot type: Scatter / Bubble

X-Axis column: Price

Value column: Price

Color: Price

Size: Price

Regression interpolation: None

A Scatter Plot

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

Ans – Data preprocessing is use to improve the quality of data in the data warehouse. Using this technique we remove the noisy data. It also helps in correcting the inconsistency in data as well as the incomplete data, due to which the efficiency of data is increased as well as it is easy to mine the data.

Steps Involved in Data Preprocessing:

1. **Data Cleaning:** The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.
2. **Data Transformation:**
   This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:
   1. **Normalization:**
      It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

   2. **Attribute Selection:**
      In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

   3. **Discretization:**
      This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

   4. **Concept Hierarchy Generation:**
      Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

3. **Data Reduction:** Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we use data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

   Five operations to handle missing values are-

1) **Ignore the Data Row-** This is usually done when the class label is missing (assuming your data mining goal is classification), or many attributes are missing from the row (not just one). However, you’ll obviously get poor performance if the percentage of such rows is high**.**
2) **Use a global constant to fill in for missing values -** Decide on a new global constant value, like “unknown“, “*N/A*” or minus infinity, that will be used to fill all the missing values. This technique is used because sometimes it just doesn’t make sense to try and predict the missing value.
3) **Use attribute mean –** Replacing missing values of an attribute with the mean (or median if its discrete) value for that attribute in the database.
4) **Use attribute mean for all samples belonging to the same class** - Instead of using the mean (or median) of a certain attribute calculated by looking at all the rows in a database, we can limit the calculations to the relevant class to make the value more relevant to the row we’re looking at.

5) **Using a data mining algorithm to predict most probable value -** The value can be determined using regression, inference based tools using Bayesian formalism, decision trees, clustering algorithms (K-Mean\Median etc.).