Assignment 1 Set A
Couse Code : CAP447
Subject Name : Data Warehouse and Data Mining LAB
Date of Submission : 09/28/2021
Name : Sirjanpreet Kaur
Roll no. : B56
Section : D2112
Group : 2
Registration no. : 12107974

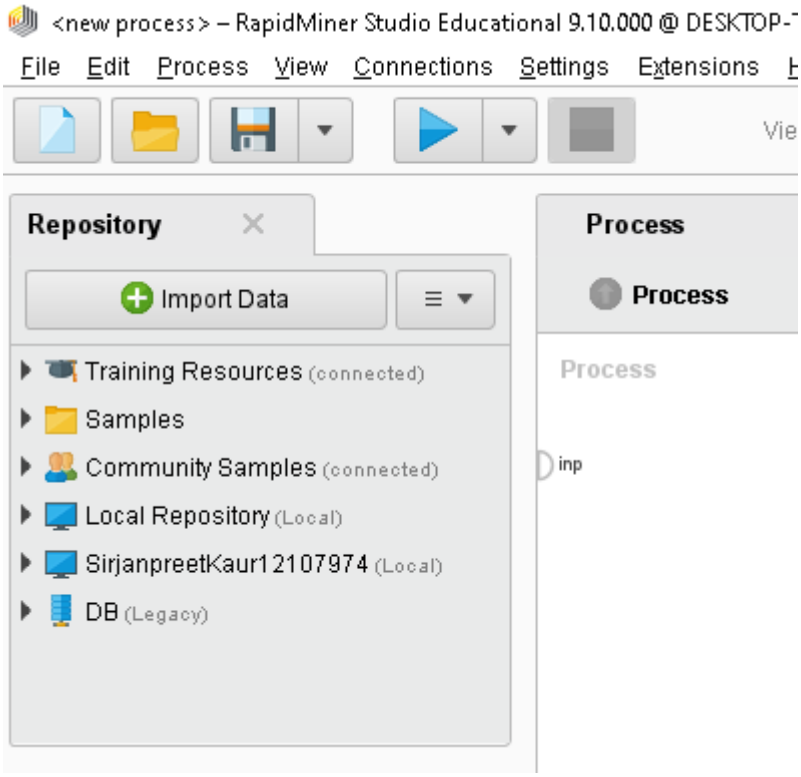
Data Set : Education in India

Link : https://www.kaggle.com/rajanand/education-in-india

Ques 1: Explain briefly the main components of RAPID MINER Studio.

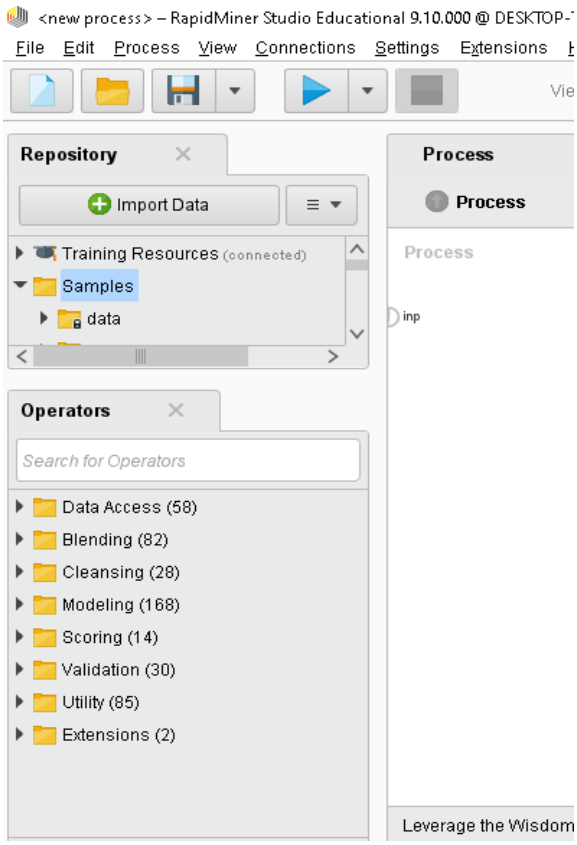
Ans : Rapid Miner Studio is a visual data science software platform that provides an integrated environment for data preparation ,machine learning ,text mining and predictive analysis. It is workflow designer that runs the prototyping & validation of models. There are various components of Rapid Miner Studio that is given below : -

(i)Repository :- A repository is a folder that holds all RapidMiner data sets, processes and other files that we create using Rapid Miner Studio. We can import or upload anything from our local repository or by default inside the Samples folder, it provides us with a lot of data. When you first start the Rapid Miner, it will automatically create a local repository. You can create your own repository in which you want to store your data.



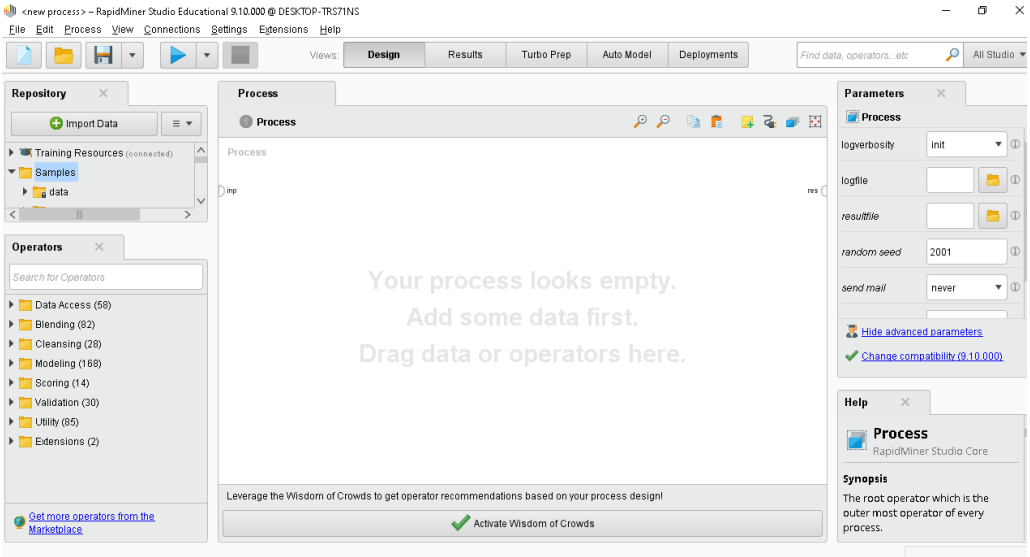
This is the repository column which is on the left of the interface.

(ii)Operators:-Operators are used to perform operations. Operators are the building blocks that are used to create Rapid Miner Processes. It has input and outport ports. The action is performed in input and after running the process, it will show the output. There are more than 1000 operators in Rapid Miner Studio.

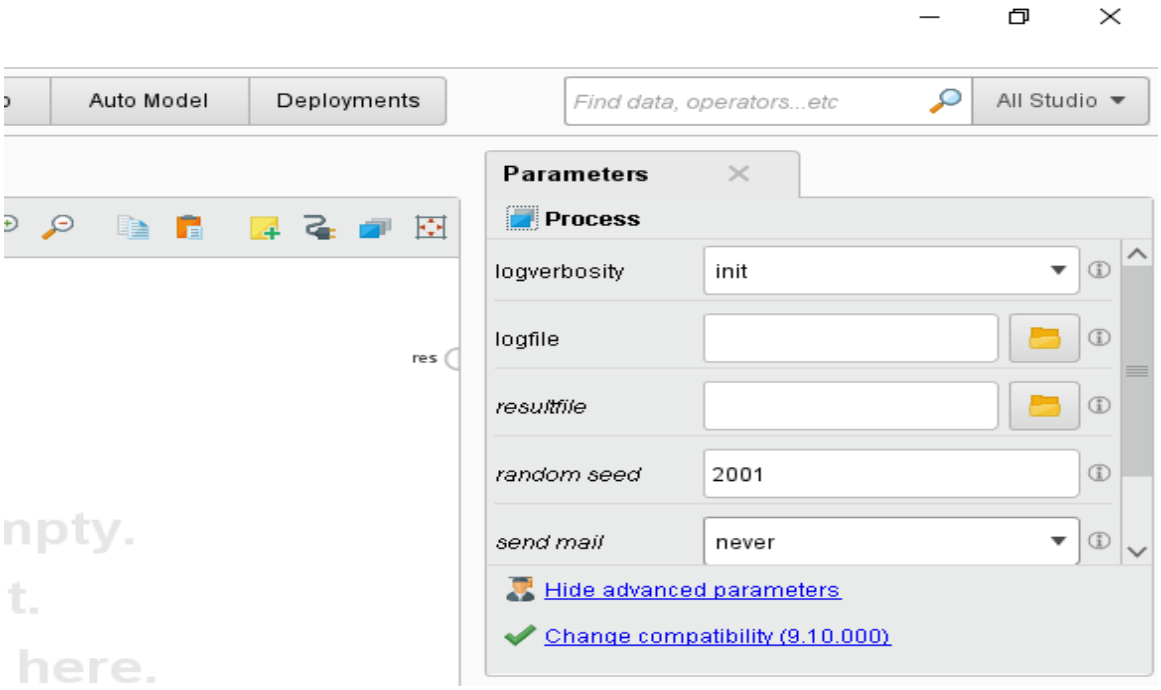


These are operators on the left side of the panel. There are many operators which are used to perform different functions on the processes.

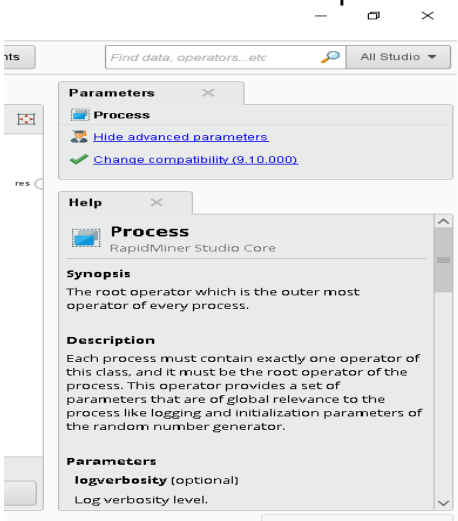
(iii)Process View:-Process view is the working area for building processes. This is the area in the design view where we can drag processes and operators. By double clicking on the process, you can drag and drop the process on the panel view. In the same way, operators are also dragged and dropped and can perform operations by using various operators. Here below is the screenshot of the process view.



(iv)Parameters:- Parameters define the characters and behaviours of an operator. By The setting(s) whose value(s) determine the characteristics or behaviour of an operator. Parameters are present in the parameter panel .In parameters, we can alter the values of parameters that what function we want to perform or what data we want to view in the results. There are many types of parameters available such as for defining average, minimum, maximum, mean value. Parameters are available for defining real or integer numbers, strings, and collections of string. Below is the screenshot of Parameters panel:-



(vi)Help:-Help is used to check the behavior of the operator. It tells the synopsis, description of the particular operator. It is at the bottom right side of the panel.Below is the screenshot of Help :-

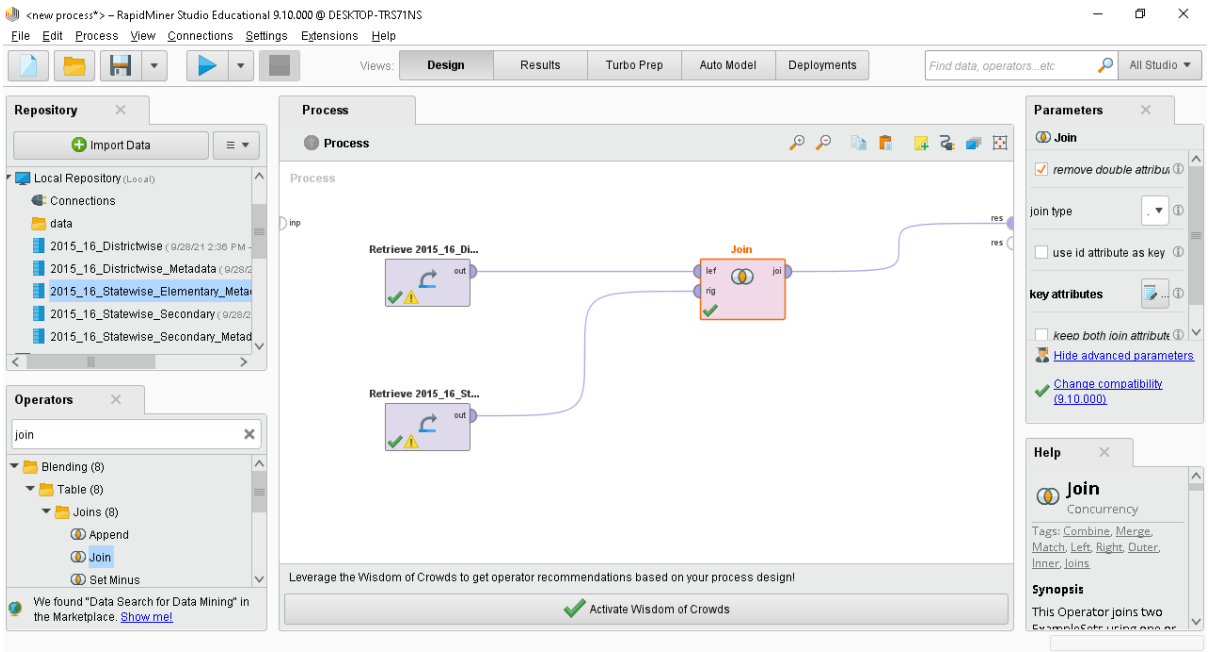


Ques 2:-a.)Explain any 5 operators along with its usage and snapshots in RapidMiner.

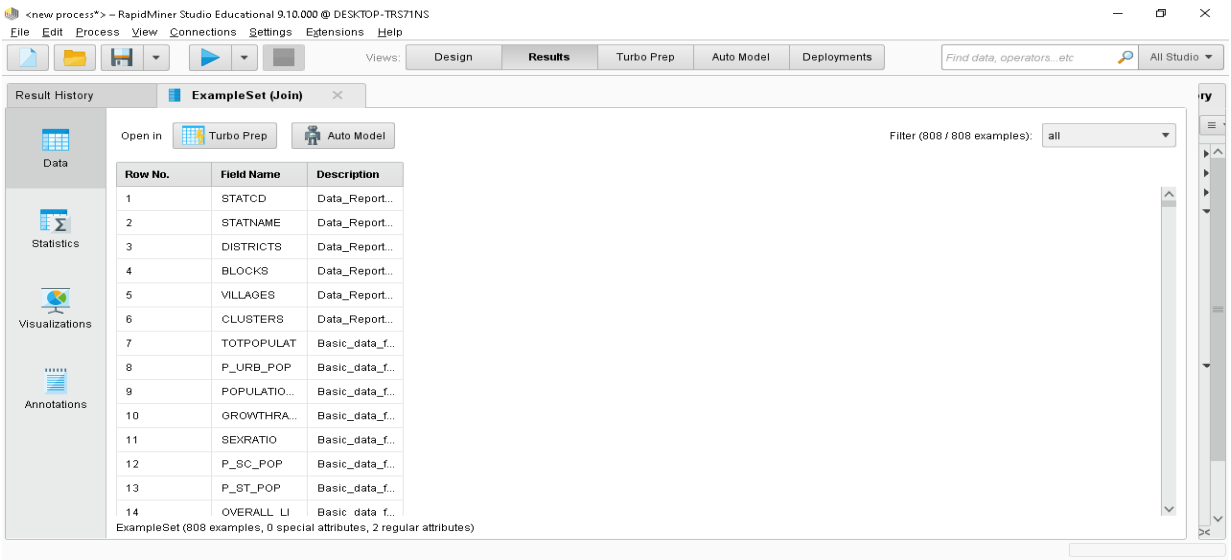
Ans:-In this, We are going to use **Education in India Dataset**.
Link : https://www.kaggle.com/rajanand/education-in-india

a)There are many types of operators used in RapidMiner, some of them are given below:-

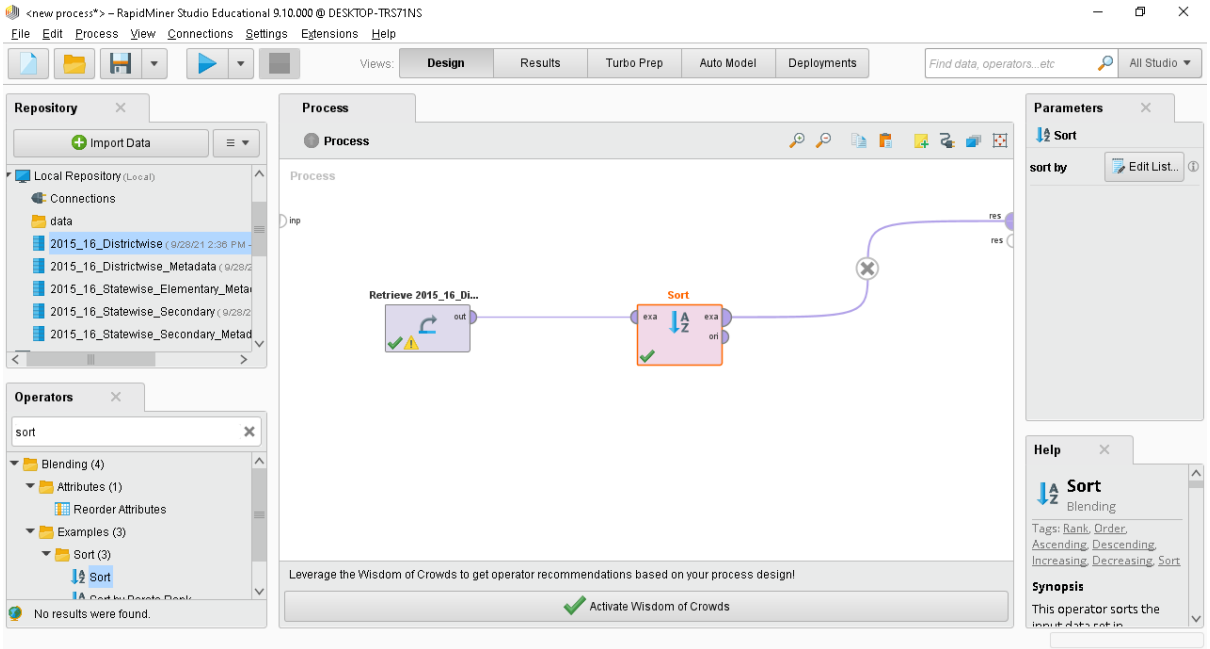
(i)Join:-Join Operator is used to join two Datasets using one or more Attributes of the input Example Sets as key Attributes. In the given snapshot , we first dragged the Products data and the Transactions data from the Samples. Then drag Join operator into the Process panel. Then we connect the output port of Retrieve Products to an input port of Join. Then connect to the output:-



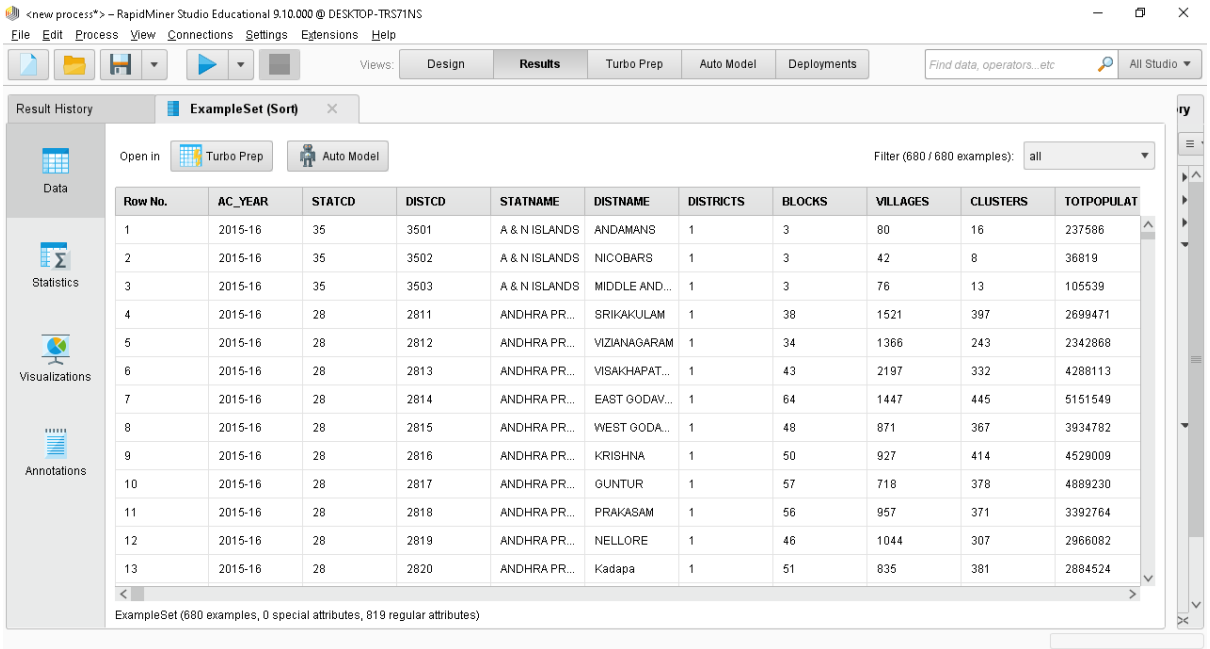
After running, following output is shown :-



(ii)Sort:-Sort Operator is used to sort any number of columns in a dataset. The 'Education in India' data set is loaded. Then Sort operator is applied on it. The sort by parameter is used to set the first attribute name parameter to 'STATNAME' and the associated sorting order to 'ascending'. It will show the state names in ascending order. Then connect to the output:-

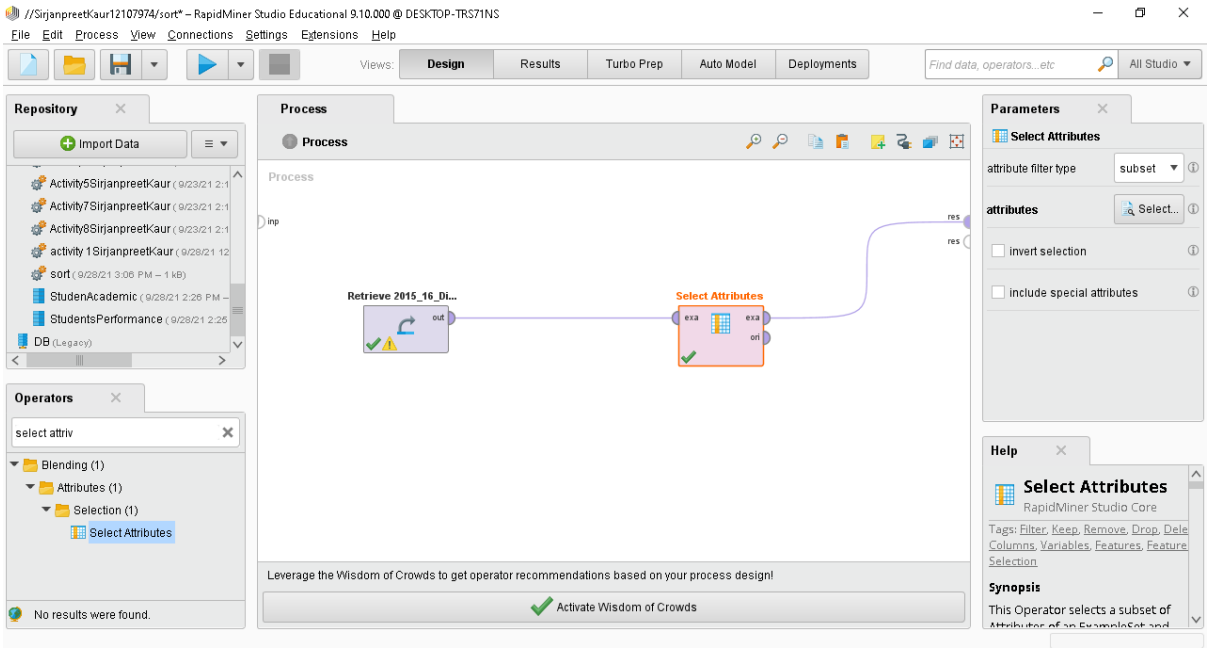


After running, following output is shown :-

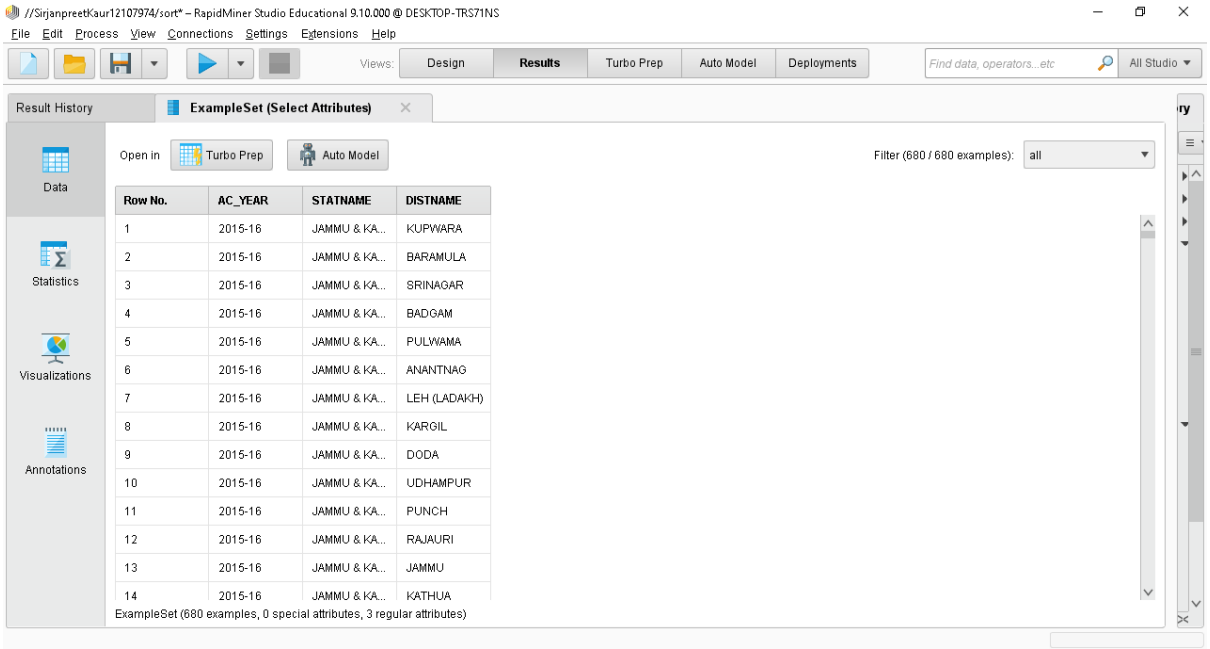


Row No.	AC_YEAR	STATCD	DISTCD	STATNAME	DISTNAME	DISTRICTS	BLOCKS	VILLAGES	CLUSTERS	TOTPOPULAT
1	2015-16	35	3501	A & N ISLANDS	ANDAMANS	1	3	80	16	237586
2	2015-16	35	3502	A & N ISLANDS	NICOBARS	1	3	42	8	36819
3	2015-16	35	3503	A & N ISLANDS	MIDDLE AND...	1	3	76	13	105539
4	2015-16	28	2811	ANDHRA PR...	SRIKAKULAM	1	38	1521	397	2699471
5	2015-16	28	2812	ANDHRA PR...	VIZIANAGARAM	1	34	1366	243	2342868
6	2015-16	28	2813	ANDHRA PR...	VISAKHAPAT...	1	43	2197	332	4288113
7	2015-16	28	2814	ANDHRA PR...	EAST GODAV...	1	64	1447	445	5151549
8	2015-16	28	2815	ANDHRA PR...	WEST GODA...	1	48	871	367	3934782
9	2015-16	28	2816	ANDHRA PR...	KRISHNA	1	50	927	414	4529009
10	2015-16	28	2817	ANDHRA PR...	GUNTUR	1	57	718	378	4889230
11	2015-16	28	2818	ANDHRA PR...	PRAKASAM	1	56	957	371	3392764
12	2015-16	28	2819	ANDHRA PR...	NELLORE	1	46	1044	307	2966082
13	2015-16	28	2820	ANDHRA PR...	Kadapa	1	51	835	381	2884524

(iii)Select Attributes:-Select Attributes operators is to select particular attributes. The original output port is connected to the input port of the Select Attributes Operator .In select Attribute operator ,we select attribute filter type to subset and select attributes AC_YEAR,DISTNAME and STATNAME. When we run, it shows the three columns AC_YEAR,DISTNAME and STATNAME.Then connect to the output:-

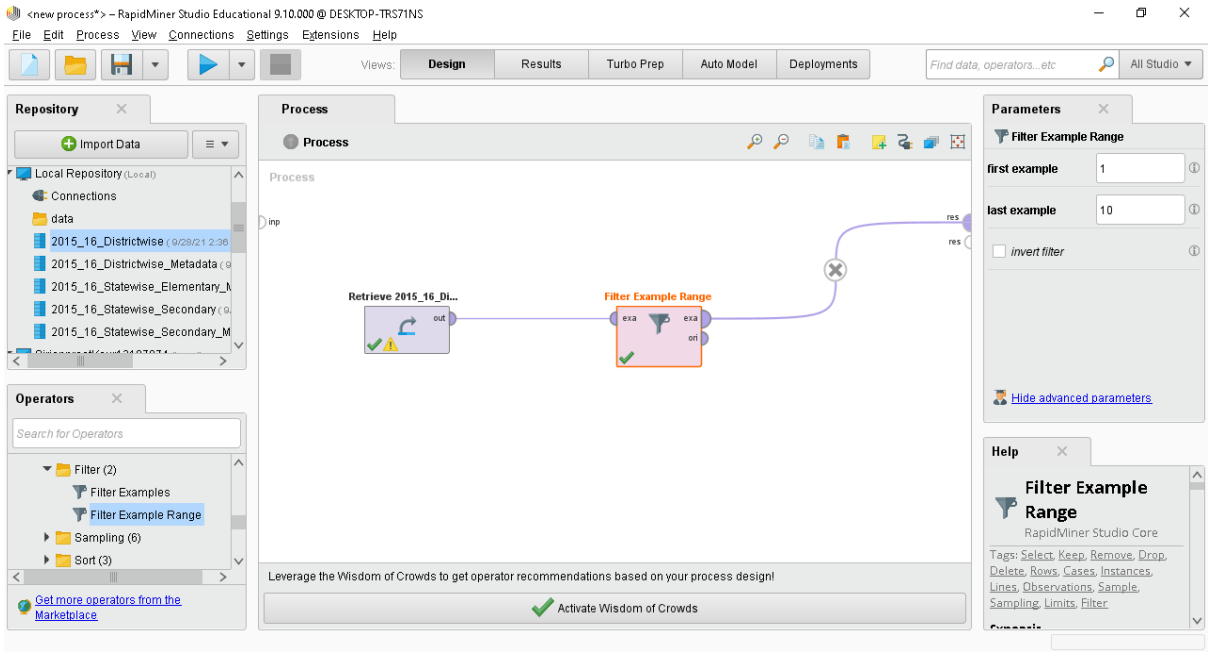


After running, following output is shown :-

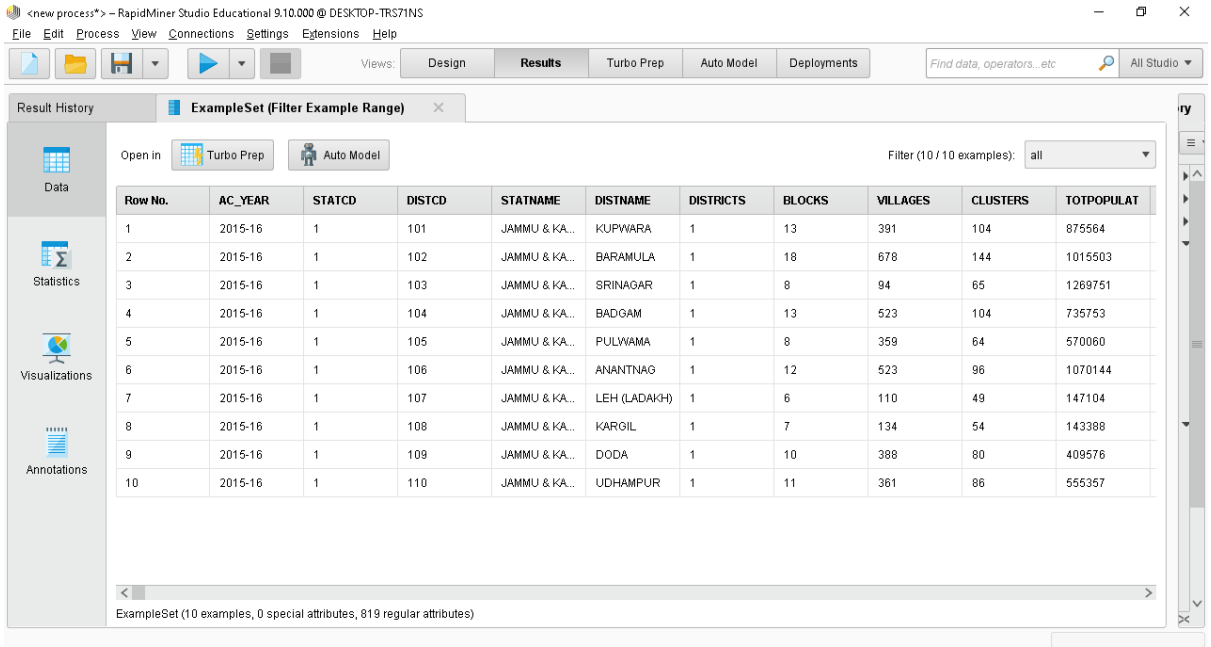


Row No.	AC_YEAR	STATNAME	DISTNAME
1	2015-16	JAMMU & KA...	KUPWARA
2	2015-16	JAMMU & KA...	BARAMULA
3	2015-16	JAMMU & KA...	SRINAGAR
4	2015-16	JAMMU & KA...	BADGAM
5	2015-16	JAMMU & KA...	PULWAMA
6	2015-16	JAMMU & KA...	ANANTNAG
7	2015-16	JAMMU & KA...	LEH (LADAKH)
8	2015-16	JAMMU & KA...	KARGIL
9	2015-16	JAMMU & KA...	DODA
10	2015-16	JAMMU & KA...	UDHAMPUR
11	2015-16	JAMMU & KA...	PUNCH
12	2015-16	JAMMU & KA...	RAJAURI
13	2015-16	JAMMU & KA...	JAMMU
14	2015-16	JAMMU & KA...	KATHUA

(iv)Filter Examples Range:-Filter Examples Range operator is to select the number of columns that lie in the specified index range attributes. The original output port is connected to the input port of the Filter Example Range Operator .In this operator , in parameters columns. ,we select first example 1 and last example 10.This means that the column from 1 to 10 will show in the result after running. Then connect to the output:-

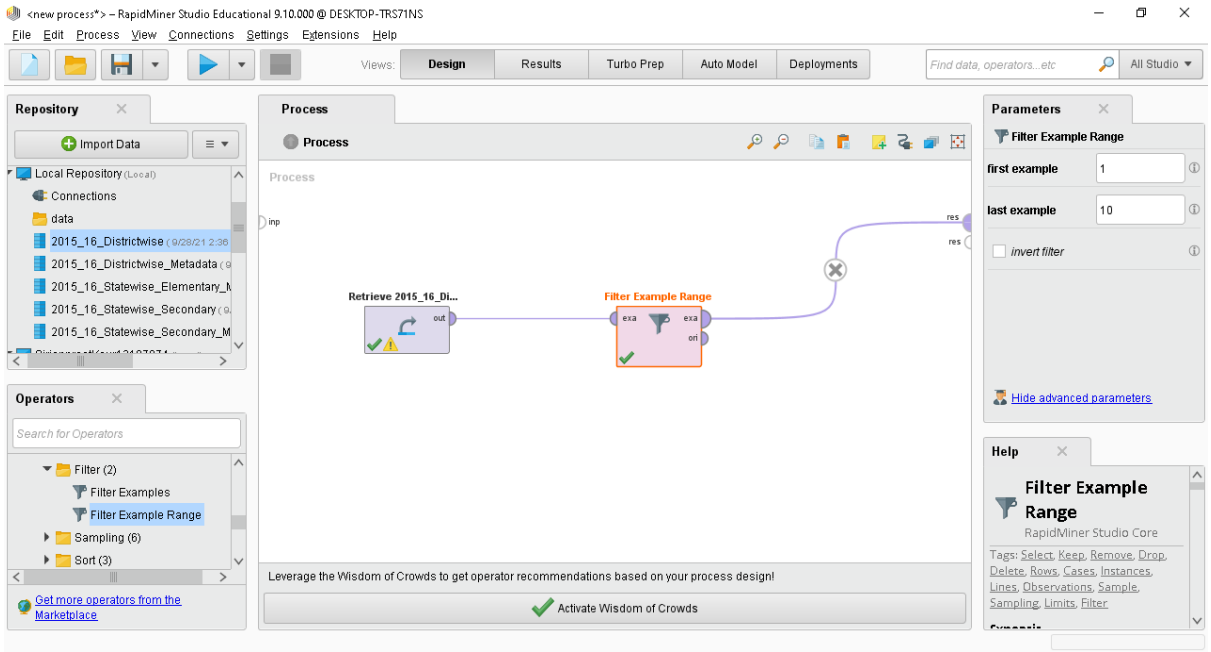


After running, following output is shown :-

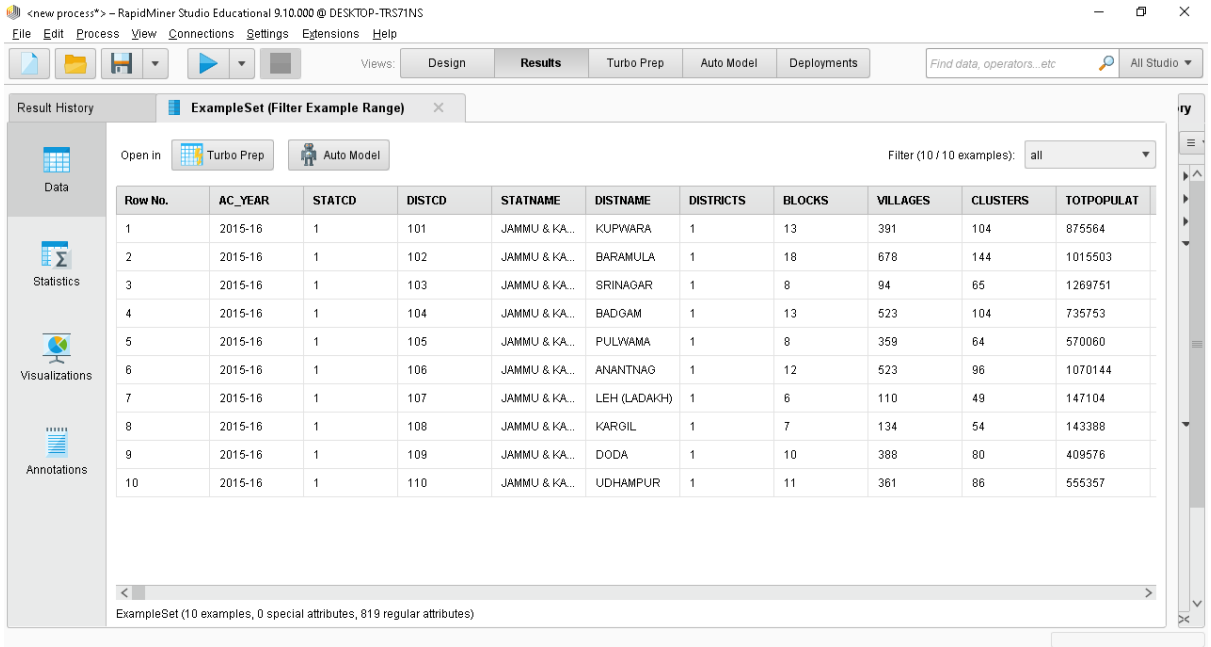


Row No.	AC_YEAR	STATCD	DISTCD	STATNAME	DISTNAME	DISTRICTS	BLOCKS	VILLAGES	CLUSTERS	TOTPOPULAT
1	2015-16	1	101	JAMMU & KA...	KUPWARA	1	13	391	104	875564
2	2015-16	1	102	JAMMU & KA...	BARAMULA	1	18	678	144	1015503
3	2015-16	1	103	JAMMU & KA...	SRINAGAR	1	8	94	65	1269751
4	2015-16	1	104	JAMMU & KA...	BADGAM	1	13	523	104	735753
5	2015-16	1	105	JAMMU & KA...	PULWAMA	1	8	359	64	570060
6	2015-16	1	106	JAMMU & KA...	ANANTNAG	1	12	523	96	1070144
7	2015-16	1	107	JAMMU & KA...	LEH (LADAKH)	1	6	110	49	147104
8	2015-16	1	108	JAMMU & KA...	KARGIL	1	7	134	54	143388
9	2015-16	1	109	JAMMU & KA...	DODA	1	10	388	80	409576
10	2015-16	1	110	JAMMU & KA...	UDHAMPUR	1	11	361	86	555357

(v)Filter Examples Range:-Filter Examples Range operator is to select the number of columns that lie in the specified index range attributes. The original output port is connected to the input port of the Filter Example Range Operator .In this operator , in parameters columns. ,we select first example 1 and last example 10.This means that the column from 1 to 10 will show in the result after running. Then connect to the output:-



After running, following output is shown :-



Row No.	AC_YEAR	STATCD	DISTCD	STATNAME	DISTNAME	DISTRICTS	BLOCKS	VILLAGES	CLUSTERS	TOTPOPULAT
1	2015-16	1	101	JAMMU & KA...	KUPWARA	1	13	391	104	875564
2	2015-16	1	102	JAMMU & KA...	BARAMULA	1	18	678	144	1015503
3	2015-16	1	103	JAMMU & KA...	SRINAGAR	1	8	94	65	1269751
4	2015-16	1	104	JAMMU & KA...	BADGAM	1	13	523	104	735753
5	2015-16	1	105	JAMMU & KA...	PULWAMA	1	8	359	64	570060
6	2015-16	1	106	JAMMU & KA...	ANANTNAG	1	12	523	96	1070144
7	2015-16	1	107	JAMMU & KA...	LEH (LADAKH)	1	6	110	49	147104
8	2015-16	1	108	JAMMU & KA...	KARGIL	1	7	134	54	143388
9	2015-16	1	109	JAMMU & KA...	DODA	1	10	388	80	409576
10	2015-16	1	110	JAMMU & KA...	UDHAMPUR	1	11	361	86	555357

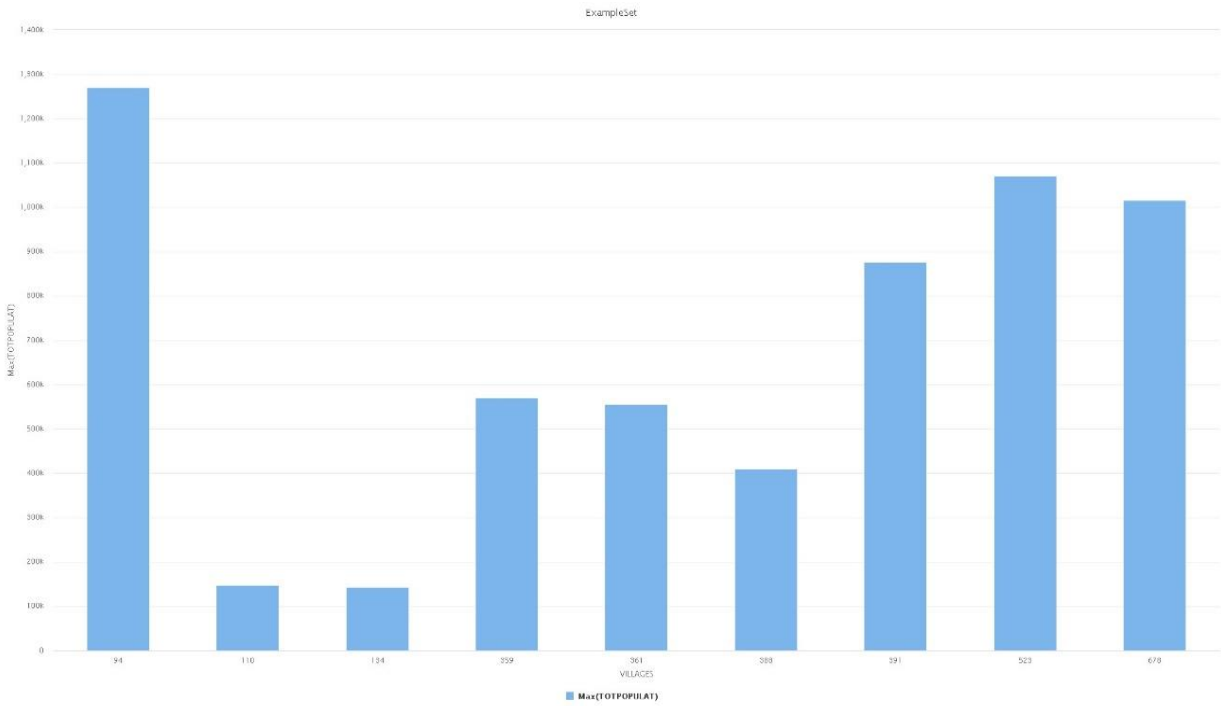
b.) What do you understand by graphical representation and statistics in RapidMiner. Attach any 5 different types of graph and their interpretation.

Ans:- A Graph is a representation of pictorial form of data in RapidMiner. Graphical representation is the visualization of data sets in the form of various graphs.The graphs are used to represent a set of data to make it easier to understand and interpret statistical information.By visualizing, it is easy to understand the data from large collection of Data. Graphical Representation is the depiction of data in the form of graphs, charts ,pie-charts etc. It means that we can visualize data ,get summary of data, aggregate data in various forms i.e. in pictorial form. It provides summary representation .With this, there are many tools in RapidMiner that are used to visualize data for more understanding. Bar Graphs, histogram, pyramid, pie-charts, scattergram and many more tools. For depicting relationship between two variables, we usually use Scattergram. In visualization, we can choose any tool for depiction of attributes. We can choose the attribute to which we want to view data in form of visualization. We can perform various operations such as aggregate data, change the colors of legend, hide titles, change color, change font, change background color and many more. So Graphical representation helps us to view data in simpler form means in aggregated or summary form through which one can easily understand the whole data.

The Statistics is the collection, analyzation and produce conclusion of data .Statistics means the representation of data in statistical form. It means Statics tab gives you sort of basic summary statistics of the different attributes in the data set. In Rapid Miner, after running the process, there is statistics tab. On clicking statistics tab, it provides the basic summary of the attributes. In the first column, there is name of the attribute. In the second column, it provides you the data type of the column. Data types depicts that whether the attribute is polynomial, alphanumeric, numeric , binomial etc. The third column counts the missing value in each attribute that how many values are missing in particular attribute. The fourth and fifth column tells the minimum and maximum value of the particular attribute. For example:-The data entry of particular country is minimum 0 and maximum 765.Sixth column shows the values of the attribute according to the operation performed. Basically, statistics involves the collection, description, analysis, and inference of conclusions from quantitative data. So Statistics represents the summary of the huge data.

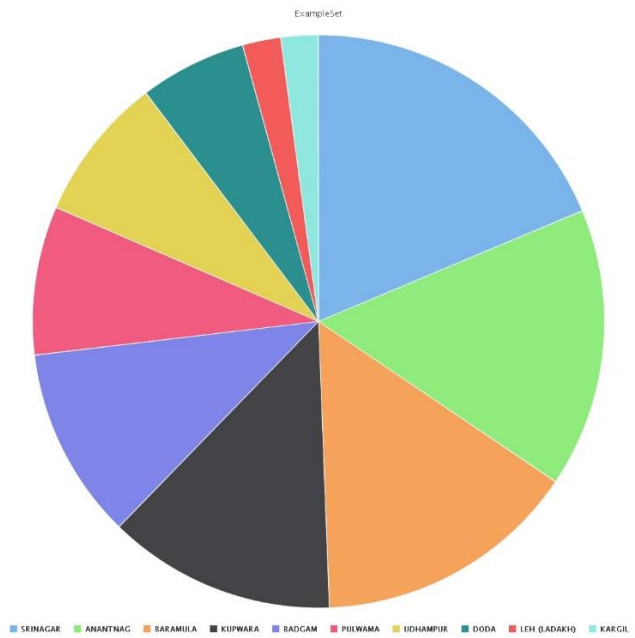
The various types of graphs are given below:-
(i)Bar-Column
(ii)Pie-Chart
(iii)Histogram
(iv)Sunburst
(v)Wordcloud

(i)Bar (Column)



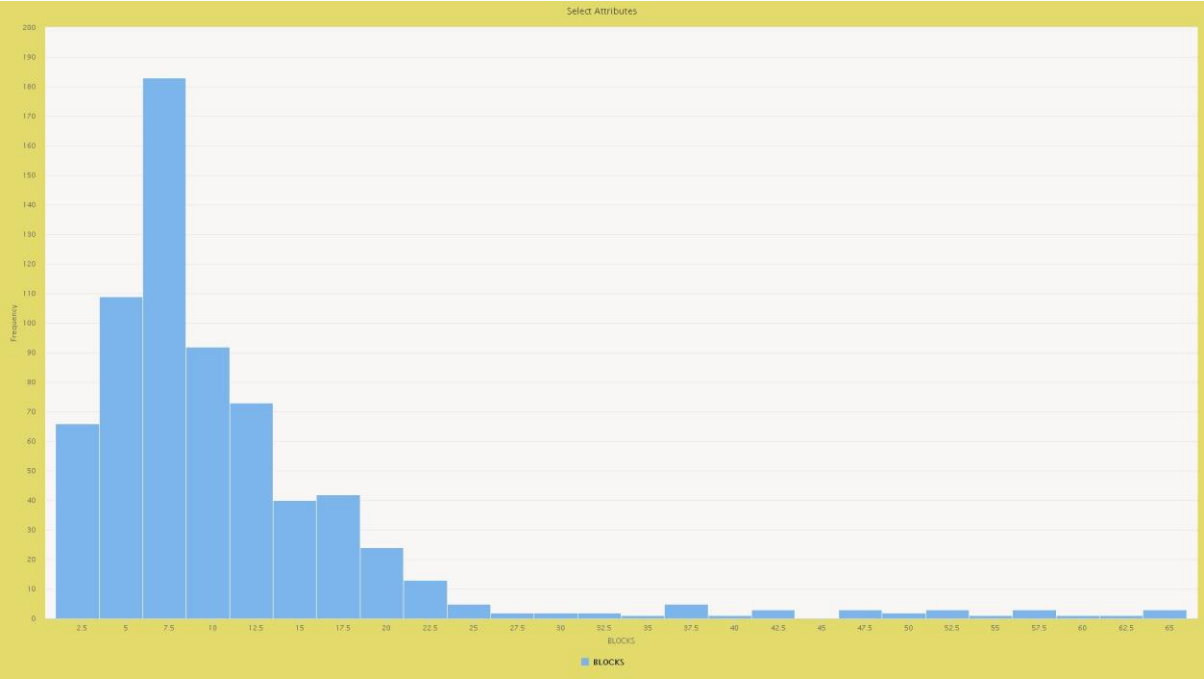
Interpretation:-This is the Bar-graph. In this graph, x-axis shows the number of villages and Y-axis shows the total population of the villages. The graph shows that there are 94 villages that have highest number of population that is MAX(TOTALPOPULAT):1,269,751.

(ii)Pie-chart:-



Interpretation:-This is the Pie-chart. It shows the number of Districts. The districts are given in different colors. It shows that Srinagar has the highest population i.e. MAX(TOTALPOPULAT):1,269,751 and the total number of population of Kargil is MAX(TOTALPOPULAT):143,3888.

(iii)Histogram



Interpretation:-This is the Histogram. In this graph, x-axis shows the number of Blocks and Y-axis shows the frequency. The graph shows that there are total 183 blocks that lies in the range of 6 – 8.5. villages that have highest number of population that is COUNT(BLOCKS):183.

(iv)Sunburst:-

Interpretation:-This is the Sunburst. It shows the number of states in different colors. It represents the number of entries of particular state. It represents the total number of entries in the data set .The total number of entries of states are 32,525.When you place the cursor on any state, It shows the number of entries of particular state.

(v)Wordcloud:-

Interpretation:-This is the Wordcloud. This graph represents the states. The states are represented according to the area. This graph represents the height of word according to the area of state. The Rajasthan has the largest area that is 342,239 sq.km i.e.AREA_SQKM:342,239.

Ques 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 in the real world is dirty means data is incomplete. It lacks attribute values, contain aggregate data, lacks certain attributes,noisy, containing errors or outliers. Data in data warehouse is noisy that is inconsistent, contains redundant data. Data has no quality in Data Warehouse and there is no mining results. So to overcome these issues, there is need of Data Preprocessing.

Data preprocessing is the transformation of raw data into meaningful data. We first need to check the data after that we should apply various data mining algorithms. Data preprocessing is used to check the quality of data. The quality of data can be checked by: accuracy, completeness,consistency,timeliness,believability,valueadded,interoperability,accessibility.

The tasks performed in Data preprocessing are:-

1.Data Cleaning:-It refers to the process of removing the inconsistent data from the data set. It removes the noisy data, remove outliers and fill missing values. It can be handled by filling the average or estimated value/

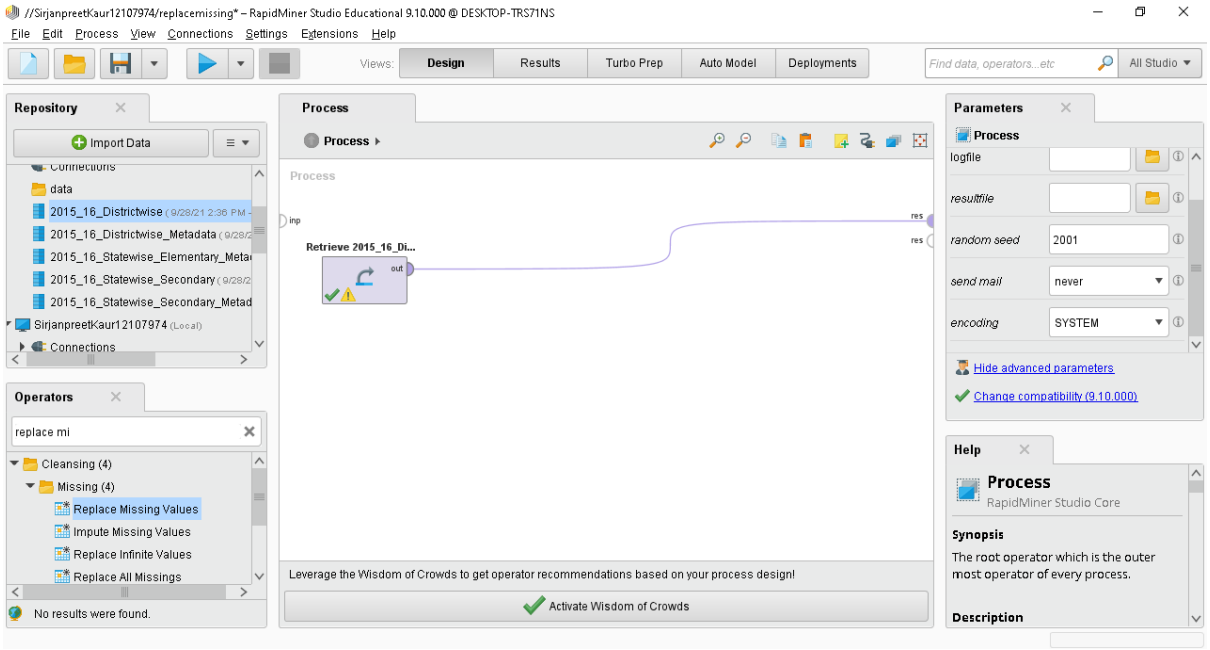
2.Data Integration:-It refers to the integration of multiple databases, data cubes and data files into single datasheet. Normalization is used to handle data discretization.

3.Data Transformation:-It refers to the change made in structure of data. It normalize and aggregate the data from the dataset. There are some of the techniques in data transformation are Smoothing, aggregation, discretization and normalization.

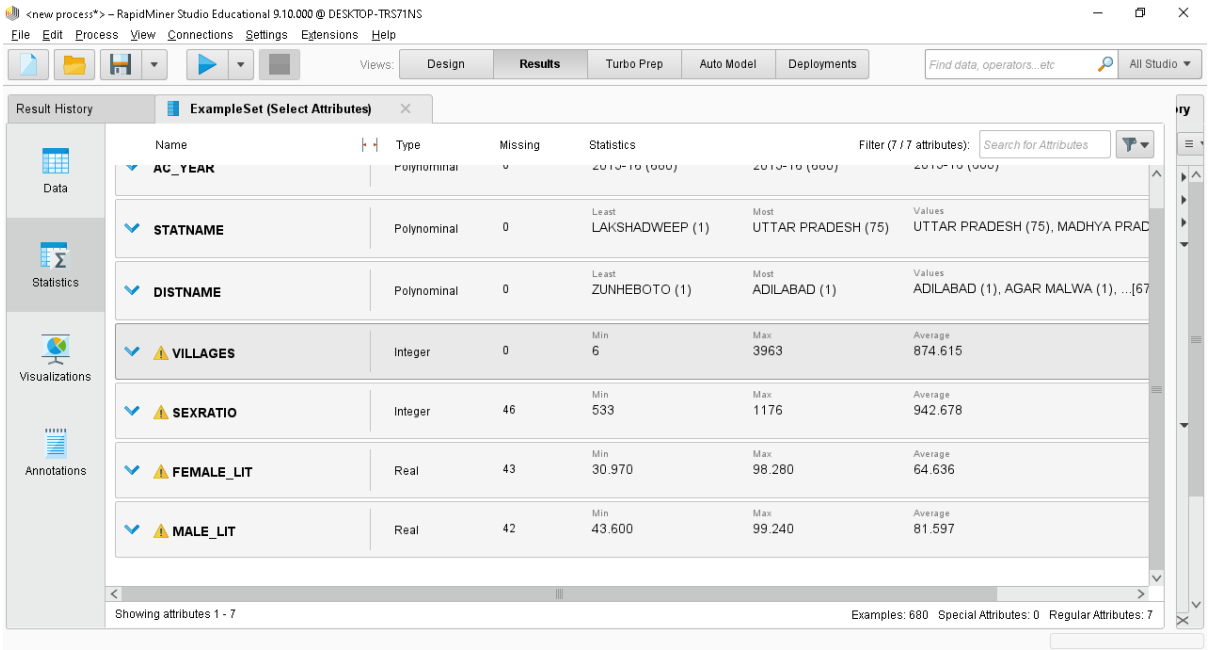
4.Data Reduction:-It refers to the reduction of volume of data that makes analysis easier but produces the same or similar result. There are some of the techniques in data reduction are Dimensionality reduction, Numerosity reduction, Data compression.

5.Data Discretization:-It refers to the reduction of part of data especially for numerical data. The Binning method is used to handle data discretization.

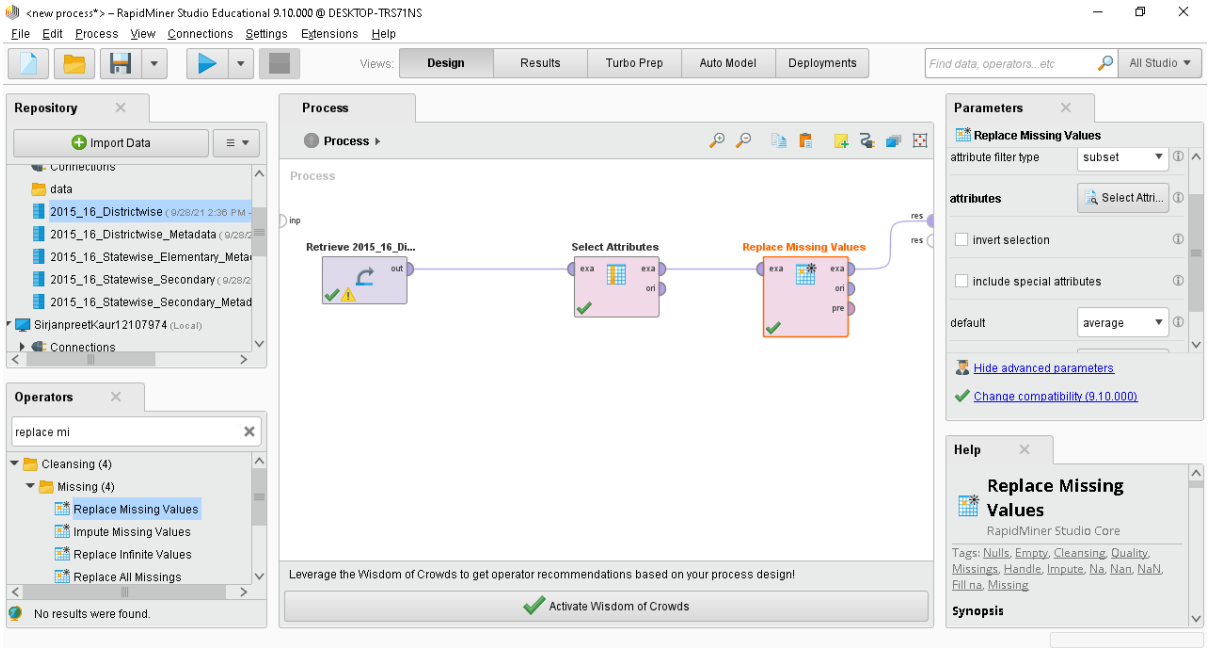
We use the dataset ;
Education in India
Link : https://www.kaggle.com/rajanand/education-in-india



There are the columns in which values are missing. FEMALE_LIT, MALE_LIT and SEXRATIO attributes have missing values.

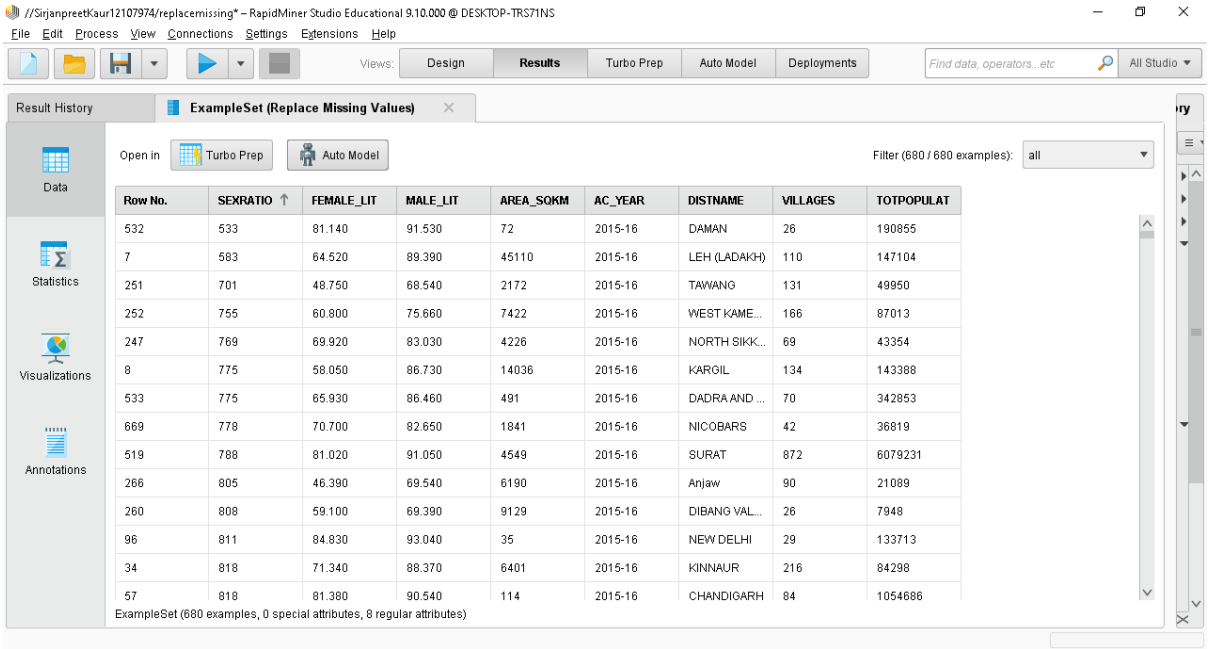


We use Replace Missing Values operator to replace the missing value as shown in snapshot :-



The operations to handle missing values are given below:-

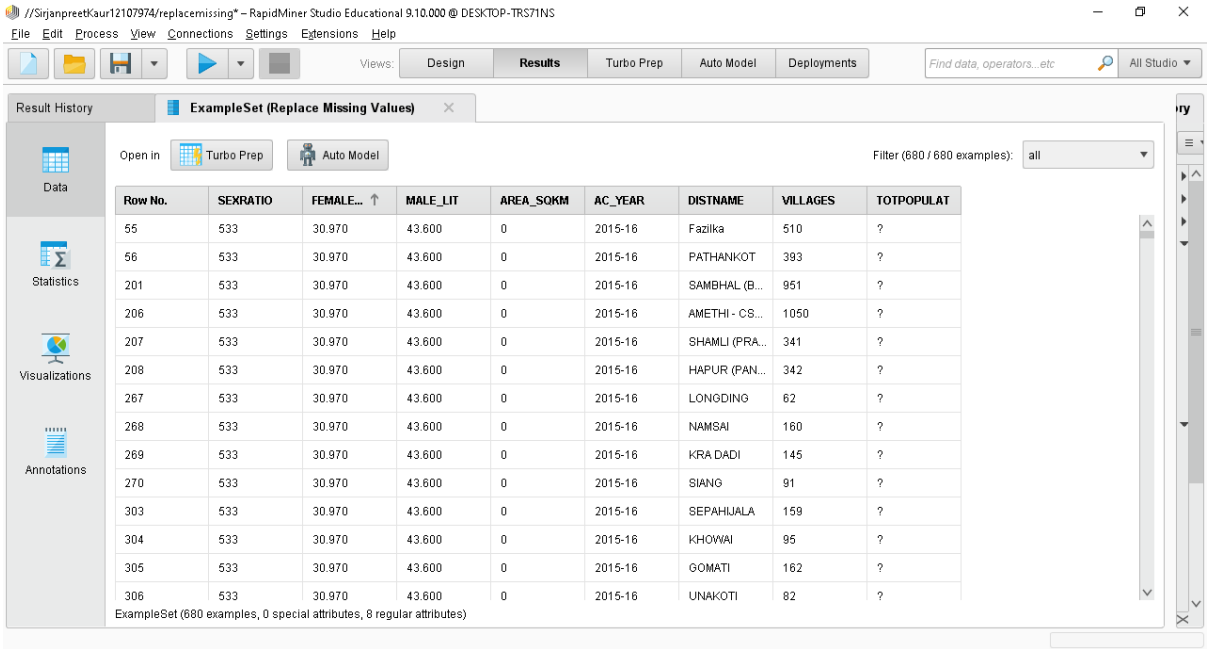
(i)average:-The average is used to replace the missing value with the average of the values of particular column. So the missing values in these columns are replaced by average values as shown in the given screenshot::-



Row No.	SEXRATIO	FEMALE_LIT	MALE_LIT	AREA_SQKM	AC_YEAR	DISTNAME	VILLAGES	TOTPOPULAT
532	533	81.140	91.530	72	2015-16	DAMAN	26	190855
7	583	64.520	89.390	45110	2015-16	LEH (LADAKH)	110	147104
251	701	48.750	68.540	2172	2015-16	TAWANG	131	49950
252	755	60.800	75.660	7422	2015-16	WEST KAME...	166	87013
247	769	69.920	83.030	4226	2015-16	NORTH SIKK...	69	43354
8	775	58.050	86.730	14036	2015-16	KARGIL	134	143388
533	775	65.930	86.460	491	2015-16	DADRA AND ...	70	342853
669	778	70.700	82.650	1841	2015-16	NICOBARS	42	36819
519	788	81.020	91.050	4549	2015-16	SURAT	872	6079231
266	805	46.390	69.540	6190	2015-16	Anjaw	90	21089
260	808	59.100	69.390	9129	2015-16	DIBANG VAL...	26	7948
96	811	84.830	93.040	35	2015-16	NEW DELHI	29	133713
34	818	71.340	88.370	6401	2015-16	KINNAUR	216	84298
57	818	81.380	90.540	114	2015-16	CHANDIGARH	84	1054686

In this, the missing value in SEX_RATIO, FEMALE_LIT, MALE_LIT and AREA_SQKM are replaced by the average value of the particular attribute.

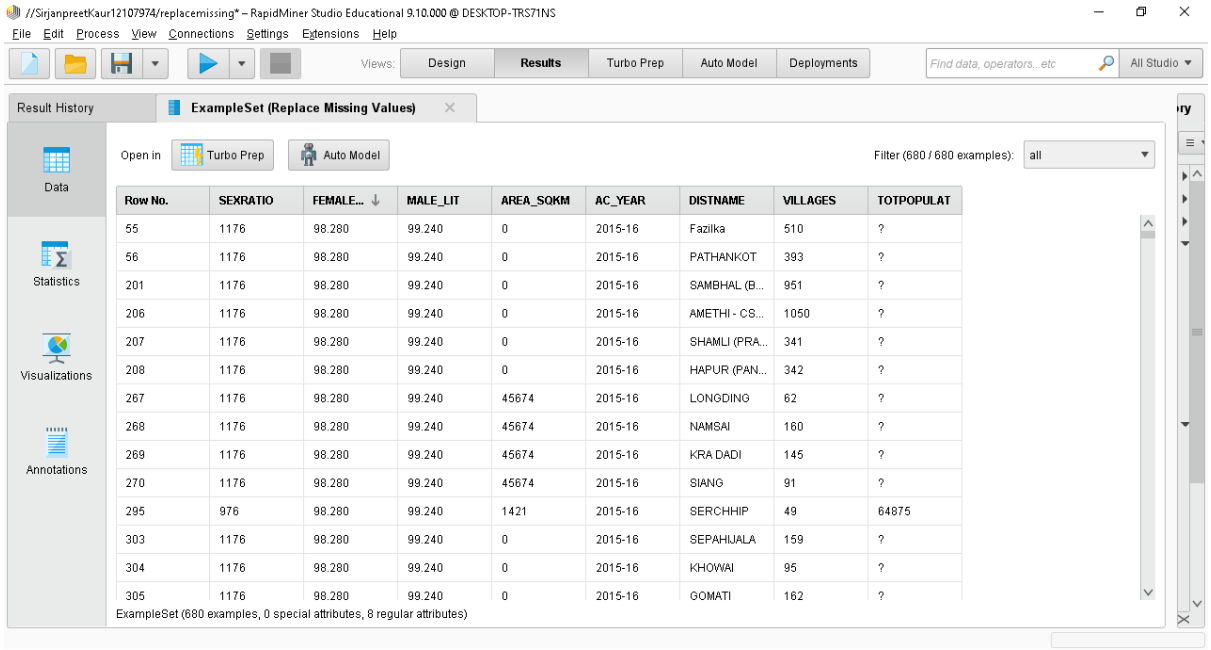
(ii)minimum :- The minimum parameter is used to replace the missing value with the minimum values of particular column. So the missing values in these columns are replaced by minimum values as shown in the given screenshot:-



Row No.	SEXRATIO	FEMALE...	MALE_LIT	AREA_SQKM	AC_YEAR	DISTNAME	VILLAGES	TOTPOPULAT
55	533	30.970	43.600	0	2015-16	Fazilka	510	?
56	533	30.970	43.600	0	2015-16	PATHANKOT	393	?
201	533	30.970	43.600	0	2015-16	SAMBHAL (B...	951	?
206	533	30.970	43.600	0	2015-16	AMETHI - CS...	1050	?
207	533	30.970	43.600	0	2015-16	SHAMLI (PRA...	341	?
208	533	30.970	43.600	0	2015-16	HAPUR (PAN...	342	?
267	533	30.970	43.600	0	2015-16	LONGDING	62	?
268	533	30.970	43.600	0	2015-16	NAMSAI	160	?
269	533	30.970	43.600	0	2015-16	KRA DADI	145	?
270	533	30.970	43.600	0	2015-16	SIANG	91	?
303	533	30.970	43.600	0	2015-16	SEPAHIJALA	159	?
304	533	30.970	43.600	0	2015-16	KHOWAI	95	?
305	533	30.970	43.600	0	2015-16	GOMATI	162	?
306	533	30.970	43.600	0	2015-16	UNAKOTI	82	?

In this, the missing value in SEX_RATIO, FEMALE_LIT, MALE_LIT and AREA_SQKM are replaced by the minimum value of the particular attribute as 533,30.970,43.600 and 0 respectively.

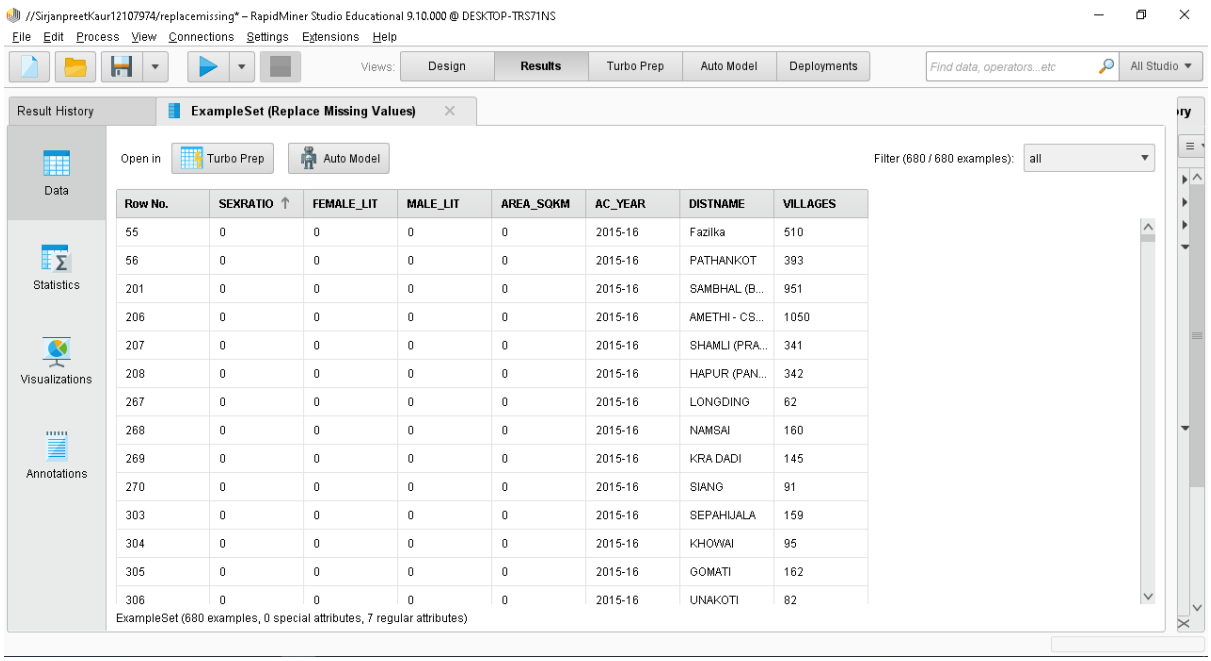
(iii)maximum :- The maximum parameter is used to replace the missing value with the maximum values of particular column. So the missing values in these columns are replaced by maximum values as shown in the given screenshot:-



Row No.	SEXRATIO	FEMALE...	MALE_LIT	AREA_SQKM	AC_YEAR	DISTNAME	VILLAGES	TOTPOPULAT
55	1176	98.280	99.240	0	2015-16	Fazilka	510	?
56	1176	98.280	99.240	0	2015-16	PATHANKOT	393	?
201	1176	98.280	99.240	0	2015-16	SAMBHAL (B...	951	?
206	1176	98.280	99.240	0	2015-16	AMETHI - CS...	1050	?
207	1176	98.280	99.240	0	2015-16	SHAMLI (PRA...	341	?
208	1176	98.280	99.240	0	2015-16	HAPUR (PAN...	342	?
267	1176	98.280	99.240	45674	2015-16	LONGDING	62	?
268	1176	98.280	99.240	45674	2015-16	NAMSAI	160	?
269	1176	98.280	99.240	45674	2015-16	KRA DADI	145	?
270	1176	98.280	99.240	45674	2015-16	SIANG	91	?
295	976	98.280	99.240	1421	2015-16	SERCHHIP	49	64875
303	1176	98.280	99.240	0	2015-16	SEPAHIJALA	159	?
304	1176	98.280	99.240	0	2015-16	KHOWAI	95	?
305	1176	98.280	99.240	0	2015-16	GOMATI	162	?

In this, the missing value in SEX_RATIO, FEMALE_LIT, MALE_LIT and AREA_SQKM are replaced by the maximum value of the particular attribute as 1176,398.280,99.240 and 45674 respectively.

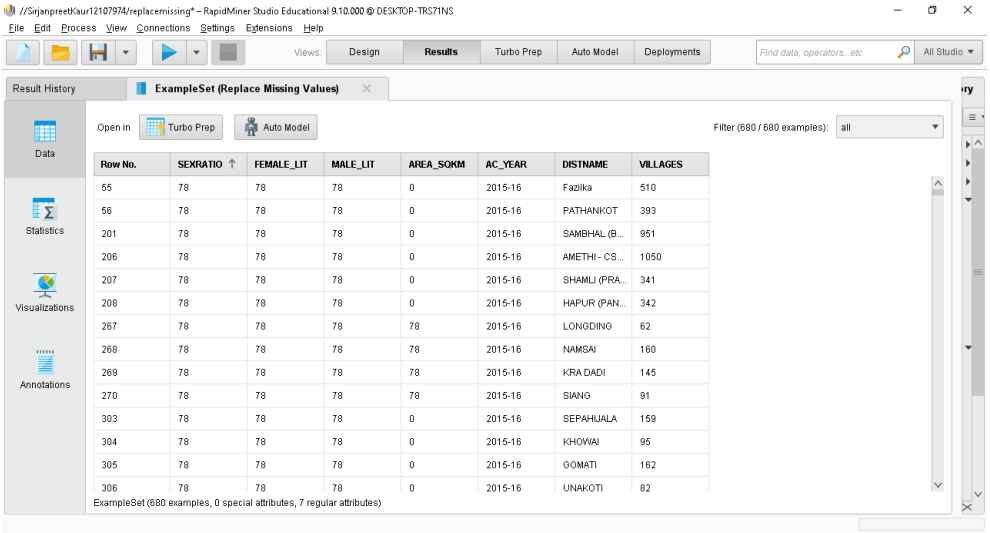
(iv)zero :- Zero is used to replace the missing value with zero.So the missing values in these columns are replaced by zero as shown in the given screenshot:-



Row No.	SEXRATIO	FEMALE_LIT	MALE_LIT	AREA_SQKM	AC_YEAR	DISTNAME	VILLAGES
55	0	0	0	0	2015-16	Fazilka	510
56	0	0	0	0	2015-16	PATHANKOT	393
201	0	0	0	0	2015-16	SAMBHAL (B...	951
206	0	0	0	0	2015-16	AMETHI - CS...	1050
207	0	0	0	0	2015-16	SHAMLI (PRA...	341
208	0	0	0	0	2015-16	HAPUR (PAN...	342
267	0	0	0	0	2015-16	LONGDING	62
268	0	0	0	0	2015-16	NAMSAI	160
269	0	0	0	0	2015-16	KRA DADI	145
270	0	0	0	0	2015-16	SIANG	91
303	0	0	0	0	2015-16	SEPAHIJALA	159
304	0	0	0	0	2015-16	KHOWAI	95
305	0	0	0	0	2015-16	GOMATI	162
306	0	0	0	0	2015-16	UNAKOTI	82

In this, the missing value in SEX_RATIO, FEMALE_LIT, MALE_LIT and AREA_SQKM are replaced by 0 as 0,0,0 and 0 respectively.

(v)value :- value is used to replace the missing value with the value defined. We can provide any value to the missing values. So the missing values in these columns are replaced by 78 as i have given value 78as shown in the given screenshot:-



Row No.	SEXRATIO	FEMALE_LIT	MALE_LIT	AREA_SQKM	AC_YEAR	DISTNAME	VILLAGES
55	78	78	78	0	2015-16	Fazilka	510
56	78	78	78	0	2015-16	PATHANKOT	393
201	78	78	78	0	2015-16	SAMBHAL (B...	951
206	78	78	78	0	2015-16	AMETHI - CS...	1050
207	78	78	78	0	2015-16	SHAMLI (PRA...	341
208	78	78	78	0	2015-16	HAPUR (PAN...	342
267	78	78	78	78	2015-16	LONGDING	62
268	78	78	78	78	2015-16	NAMSAI	160
269	78	78	78	78	2015-16	KRA DADI	145
270	78	78	78	78	2015-16	SIANG	91
303	78	78	78	0	2015-16	SEPAHIJALA	159
304	78	78	78	0	2015-16	KHOWAI	95
305	78	78	78	0	2015-16	GOMATI	162
306	78	78	78	0	2015-16	UNAKOTI	82

In this, the missing value in SEX_RATIO, FEMALE_LIT, MALE_LIT and AREA_SQKM are replaced by 78 as 78,78,78 and 78 respectively.