ASSIGNMENT-1

NAME-MADHAV JHA

REGISTRATION NO.-12107418

SECTION-D2112

ROLL. NO.-RD2112A41

COURSE CODE-CAP447

DATE OF SUBMISSION-29-09-2021

Q1.Explain briefly the main components of RAPID MINER Studio.

Answer:-

RapidMiner Studio is a where you build and edit analytic processes. RapidMiner Studio and RapidMiner Server connect and interact with each other.RapidMiner Studio is a visual workflow designer for predictive analytics that brings data science and machine learning to everyone on the analytics team.When you're working on a new project of any kind, often the first step will be to go to a whiteboard, where you will plan the workflow and identify the key steps on the way to your goal. If you're a data scientist, the workflow will usually include one or more of the following steps:

- Import data
- Prepare data
- Build a model
- Validate the model
- Apply the model

RapidMiner Studio implements your whiteboard workflow in software, in the *Design View*. The Design View includes numerous *panels*.

- Repository

  Repository: The place where your data, processes, and results are stored, either locally or remotely.

Also known as: folder, workspace, project

When working with RapidMiner Studio, you need a place to save your work. The Repository can be used to store:

- Data
- Processes
- Results
- Operators

Operators: The elements of a Process, each Operator takes input and creates output, depending on the choice of parameters.Also known as: function, formula, node.To use RapidMiner Studio effectively, you have to learn about its Operators. RapidMiner Studio includes hundreds of Operators but all we not need to know,we have to know only a few operators.

- Parameters

Parameters: Options for configuring the behavior of an Operator.The content of the Parameters Panel is context-dependent. Select any Operator that is displayed in the Process Panel, and the Parameters Panel displays the options for configuring that Operator. Because RapidMiner Studio includes many Operators, each with its own unique functionality, the range of parameters is also quite diverse. By default, RapidMiner Studio will show you only the more commonly used parameters. To see all of the available parameters, click Show advanced parameters.

- Help:-

Help: Displays a help text for the current Operator. The content of the *Help Panel* is also context-dependent. Select any Operator that is displayed in the Process Panel, and the Help Panel displays a help text for that Operator. The Help Panel provides useful background informations.

- Process

Process: A connected set of Operators that help you to transform and analyze your data.Also known as: flow, program, pipeline, diagram .Your goal is to create a finished process, a connected set of Operators that produce a result. For example, your process might read a data set and build a predictive model. When you have connected all your Operators and set their parameters, press the Run Run arrow button at the top of the user interface, and the results will be displayed in the Results View.

Q2(A). Explain any 5 operators along with its usage and snapshots in rapidminer.

Answer:-The five operators along with uses and snapshot are below:-

- Sort:-

This operator sorts the input data set in ascending or descending Order according to several attributes.The attributes to sort by are specified using the sort by Parameter. For each attribute, sorting is done in ascending or descending order, depending on the setting of the sorting order parameter. The resulting data set is sorted by the first attribute,Then subsets of the same value in the first attribute are sorted by the second attribute.

Figure :-1 ,shows the short operator preview.

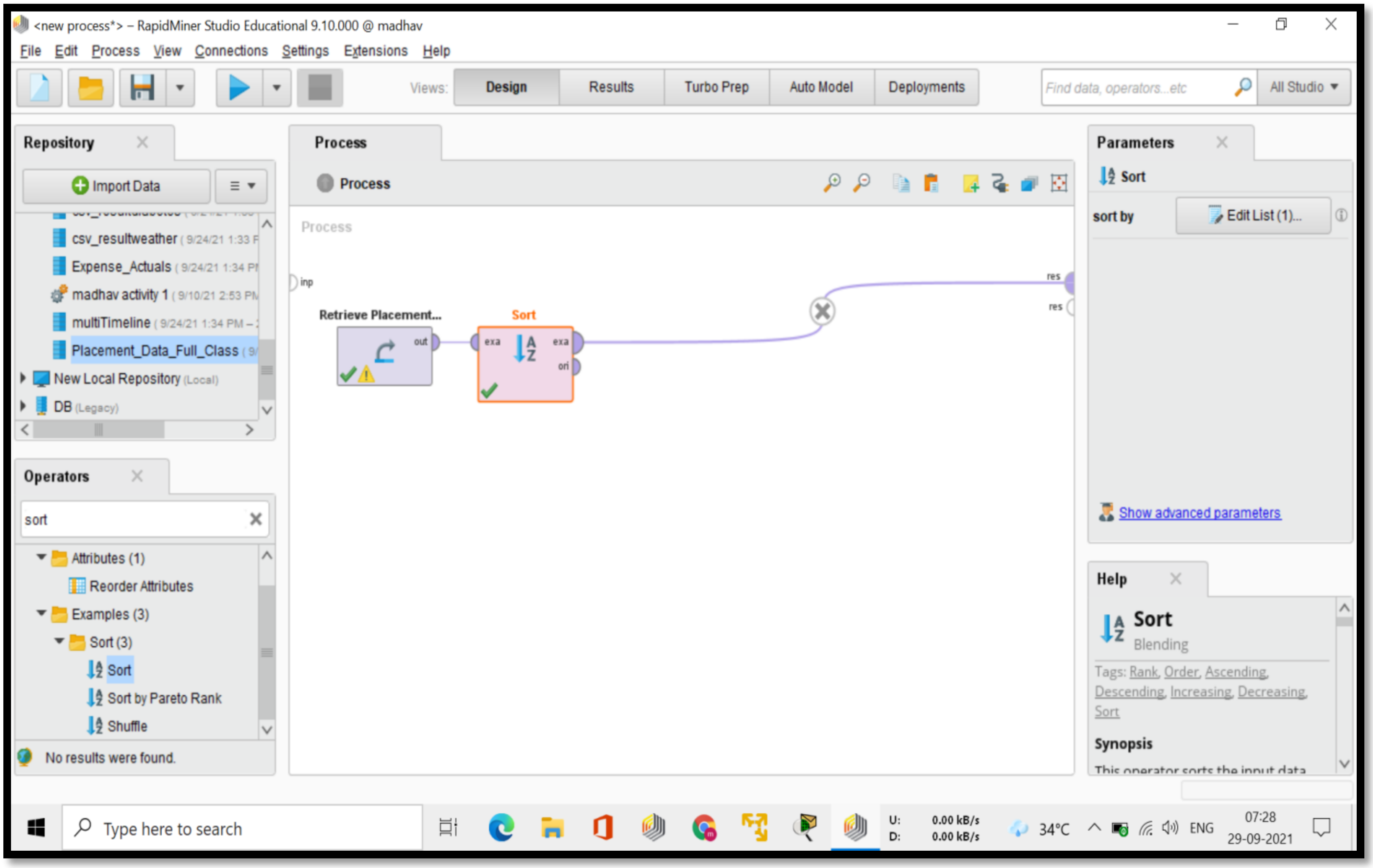
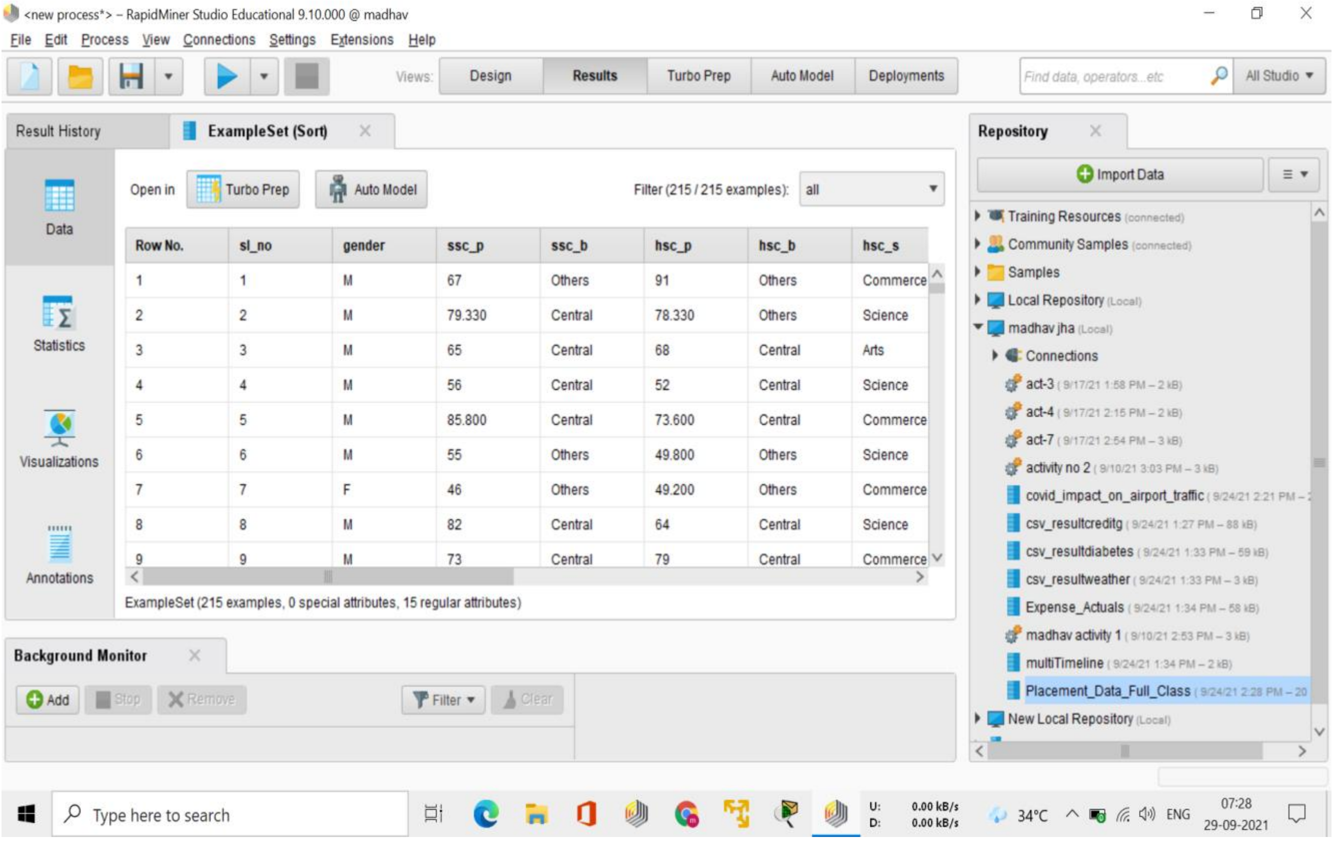

<new process*> – RapidMiner Studio Educational 9.10.000 @ madhav
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import Data
csv_resultweather (9/24/21 1:33 P
Expense_Actuals (9/24/21 1:34 PM
madhav activity 1 (9/10/21 2:53 PM
multiTimeline (9/24/21 1:34 PM –
Placement_Data_Full_Class (9/
New Local Repository (Local)
DB (Legacy)
Operators
sort
Attributes (1)
Reorder Attributes
Examples (3)
Sort (3)
Sort
Sort by Pareto Rank
Shuffle
No results were found.
Process
Process
Process
inp
res
res
Retrieve Placement...
out
Sort
exa
exa
ori
Parameters
Sort
sort by
Edit List (1)...
Show advanced parameters
Help
Sort
Blending
Tags: Rank, Order, Ascending, Descending, Increasing, Decreasing, Sort
Synopsis
This operator sorts the input data
Type here to search
U: 0.00 kB/s
D: 0.00 kB/s
34°C
ENG
07:28
29-09-2021

Figure :-2 ,shows the short operatoration on sl_no attribute in assending order.



<new process*> – RapidMiner Studio Educational 9.10.000 @ madhav
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
ExampleSet (Sort)
Data
Statistics
Visualizations
Annotations
Open in Turbo Prep Auto Model
Repository
Import Data
Training Resources (connected)
Community Samples (connected)
Samples
Local Repository (Local)
madhav jha (Local)
Connections
act-3 (9/17/21 1:58 PM – 2 kB)
act-4 (9/17/21 2:15 PM – 2 kB)
act-7 (9/17/21 2:54 PM – 3 kB)
activity no 2 (9/10/21 3:03 PM – 3 kB)
covid_impact_on_airport_traffic (9/24/21 2:21 PM –
csv_resultcreditg (9/24/21 1:27 PM – 88 kB)
csv_resultdiabetes (9/24/21 1:33 PM – 59 kB)
csv_resultweather (9/24/21 1:33 PM – 3 kB)
Expense_Actuals (9/24/21 1:34 PM – 58 kB)
madhav activity 1 (9/10/21 2:53 PM – 3 kB)
multiTimeline (9/24/21 1:34 PM – 2 kB)
Placement_Data_Full_Class (9/24/21 2:28 PM – 20
New Local Repository (Local)
Background Monitor
Add Stop Remove Filter Clear
Type here to search
U: 0.00 kB/s
D: 0.00 kB/s
34°C ENG
07:28
29-09-2021

Filter (215 / 215 examples): all

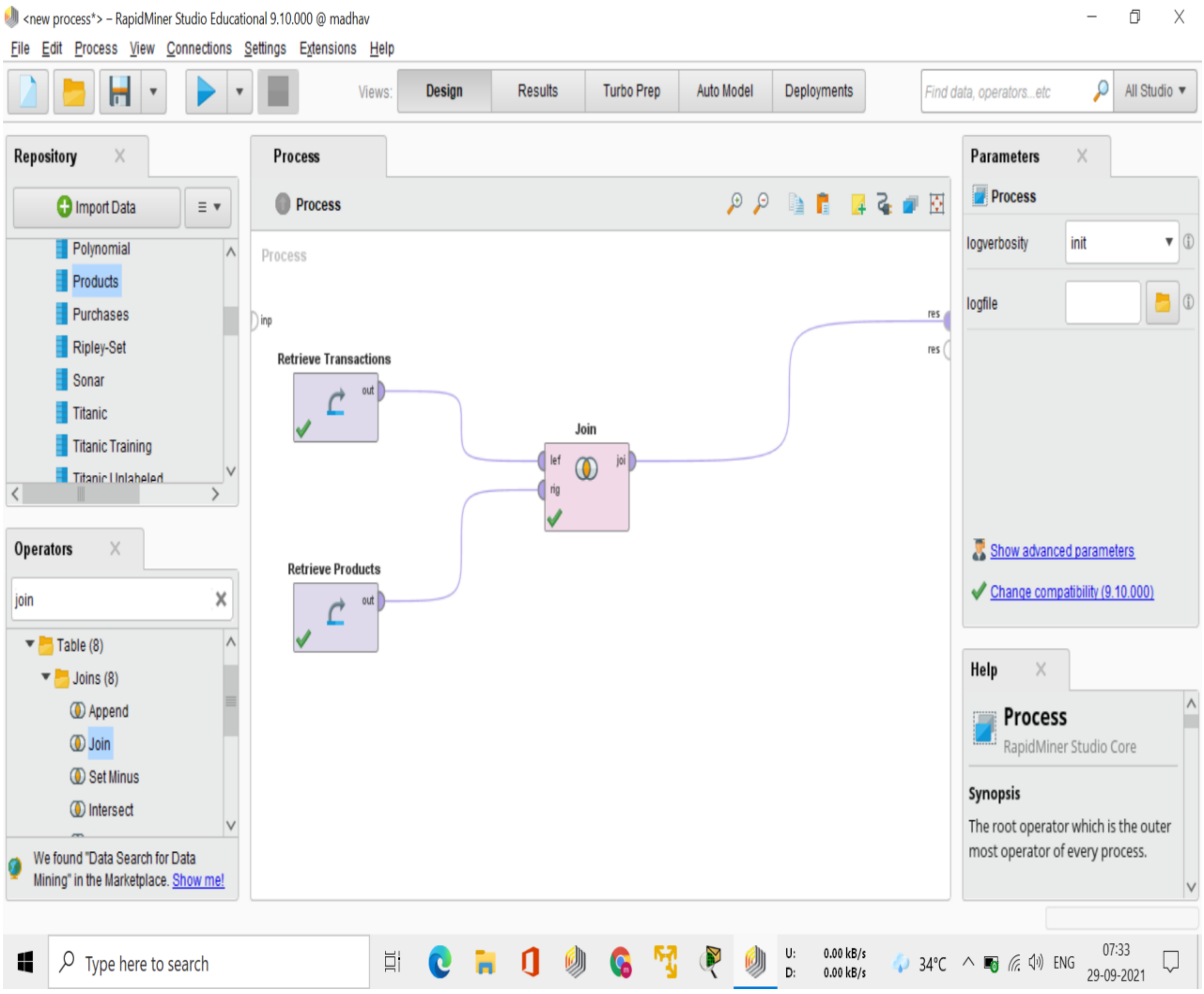
Row No.	sl_no	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s
1	1	M	67	Others	91	Others	Commerce
2	2	M	79.330	Central	78.330	Others	Science
3	3	M	65	Central	68	Central	Arts
4	4	M	56	Central	52	Central	Science
5	5	M	85.800	Central	73.600	Central	Commerce
6	6	M	55	Others	49.800	Others	Science
7	7	F	46	Others	49.200	Others	Commerce
8	8	M	82	Central	64	Central	Science
9	9	M	73	Central	79	Central	Commerce

ExampleSet (215 examples, 0 special attributes, 15 regular attributes)

- Join:-

  This Operator joins two ExampleSets using one or more Attributes of the input ExampleSets as key attributes.Identical values of the key attributes indicate matching Examples. An Attribute with id role is selected as key by default but an arbitrary set of one or more Attributes can be chosen as key. Four types of joins are possible: inner, left, right and outer join. All these types of joins are explained In the parameters section.

Figure :-3 ,shows the join operator preview. which combining two example sets i.e transaction and product.



join

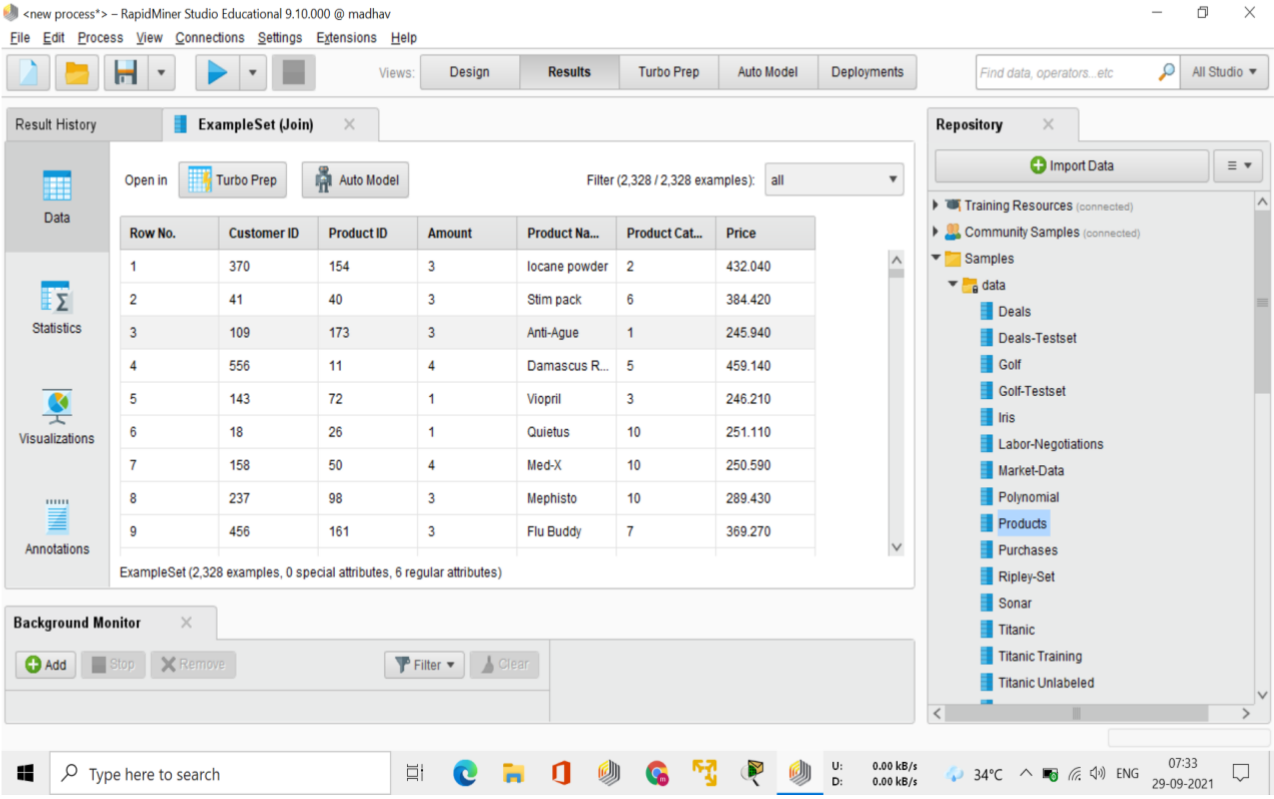
logverbosity: init

We found "Data Search for Data Mining" in the Marketplace. Show me!

Synopsis

The root operator which is the outer most operator of every process.

Figure :-4 ,shows the short operatoration on product id,which combining two example sets i.e transaction and product.



<new process*> – RapidMiner Studio Educational 9.10.000 @ madhav
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
ExampleSet (Join)
Data
Statistics
Visualizations
Annotations
Open in Turbo Prep Auto Model
Repository
Import Data
Training Resources (connected)
Community Samples (connected)
Samples
data
Deals
Deals-Testset
Golf
Golf-Testset
Iris
Labor-Negotiations
Market-Data
Polynomial
Products
Purchases
Ripley-Set
Sonar
Titanic
Titanic Training
Titanic Unlabeled
Background Monitor
Add Stop Remove Filter Clear
Type here to search
34°C ENG 07:33 29-09-2021

Filter (2,328 / 2,328 examples): all

Row No.	Customer ID	Product ID	Amount	Product Na...	Product Cat...	Price
1	370	154	3	Iocane powder	2	432.040
2	41	40	3	Stim pack	6	384.420
3	109	173	3	Anti-Ague	1	245.940
4	556	11	4	Damascus R...	5	459.140
5	143	72	1	Viopril	3	246.210
6	18	26	1	Quietus	10	251.110
7	158	50	4	Med-X	10	250.590
8	237	98	3	Mephisto	10	289.430
9	456	161	3	Flu Buddy	7	369.270

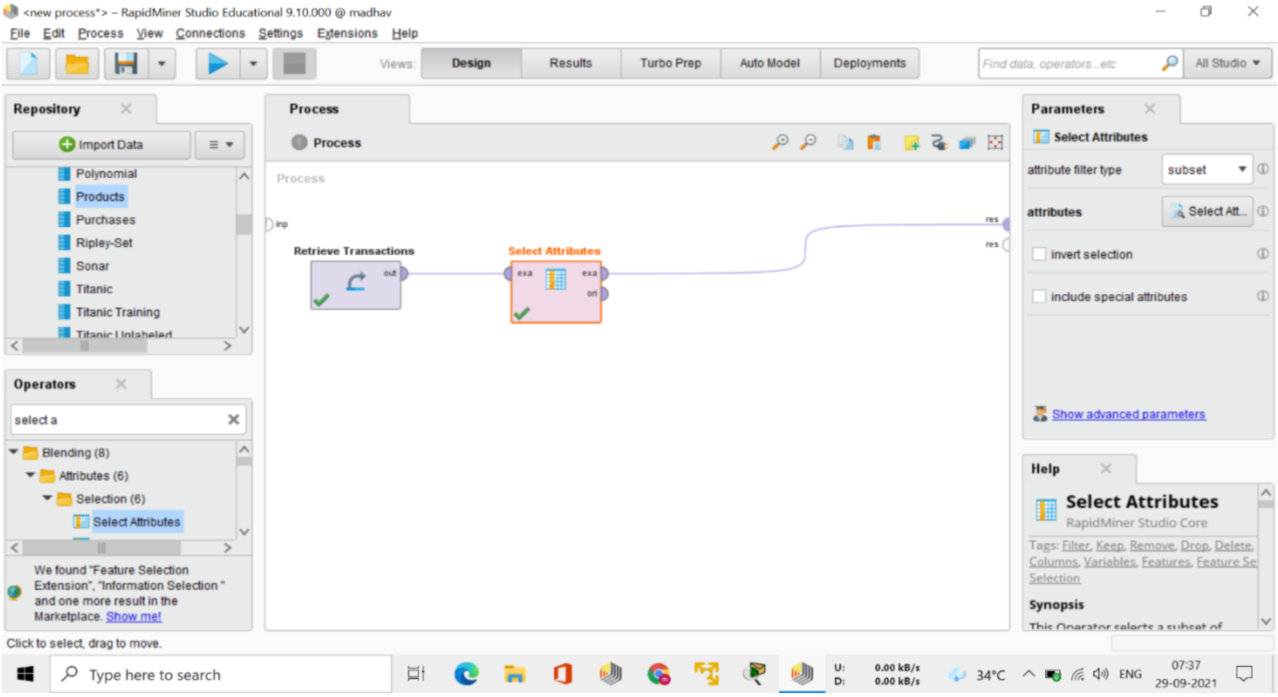
ExampleSet (2,328 examples, 0 special attributes, 6 regular attributes)

- Replace:-
- Retrieve:-This Operator can access stored information in the Repository and Load them into the Process.The Retrieve Operator loads a RapidMiner Object into the Process. This Object is often an ExampleSet but it can also be a

Collection or a Model. Retrieving data this way also provides the Meta data of the RapidMiner Object.

- Select Attributes:-This Operator selects a subset of Attributes of an ExampleSet and Removes the other Attributes.The Operator provides different filter types to make Attribute selection easy. Possibilities are For example: Direct selection of attributes. Selection by a regular expression or selecting only attributes without missing values. See parameter attribute filter type for a detailed description of the different filter types.

Figure :-5 ,shows the select attribute operator preview.



We found "Feature Selection Extension", "Information Selection " and one more result in the Marketplace. Show me!

Click to select, drag to move.

Select Attributes

attribute filter type	subset
attributes	Select Att...

☐ invert selection

☐ include special attributes

Show advanced parameters

Select Attributes

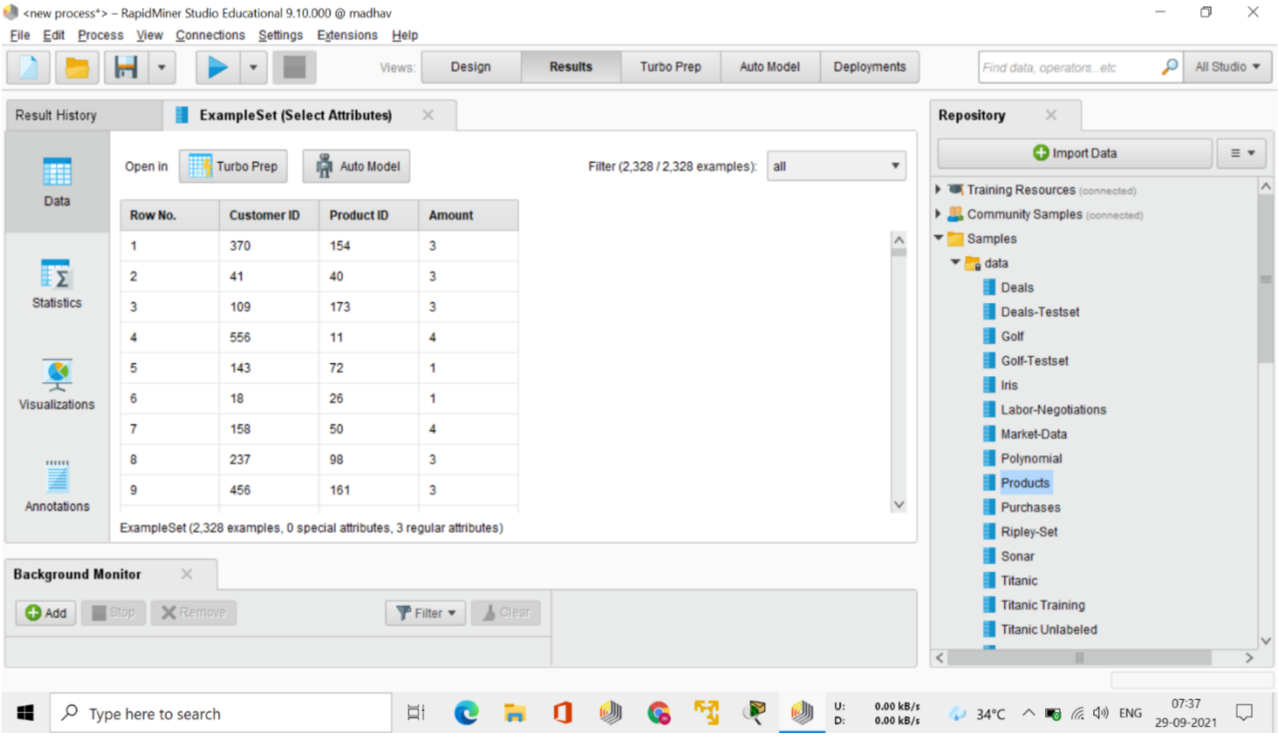
RapidMiner Studio Core

Tags: Filter, Keep, Remove, Drop, Delete, Columns, Variables, Features, Feature Selection

Synopsis

This Operator selects a subset of

Figure :-6 ,shows the selection of attributes operatoration on transaction data set.



ExampleSet (Select Attributes)

Filter (2,328 / 2,328 examples): all

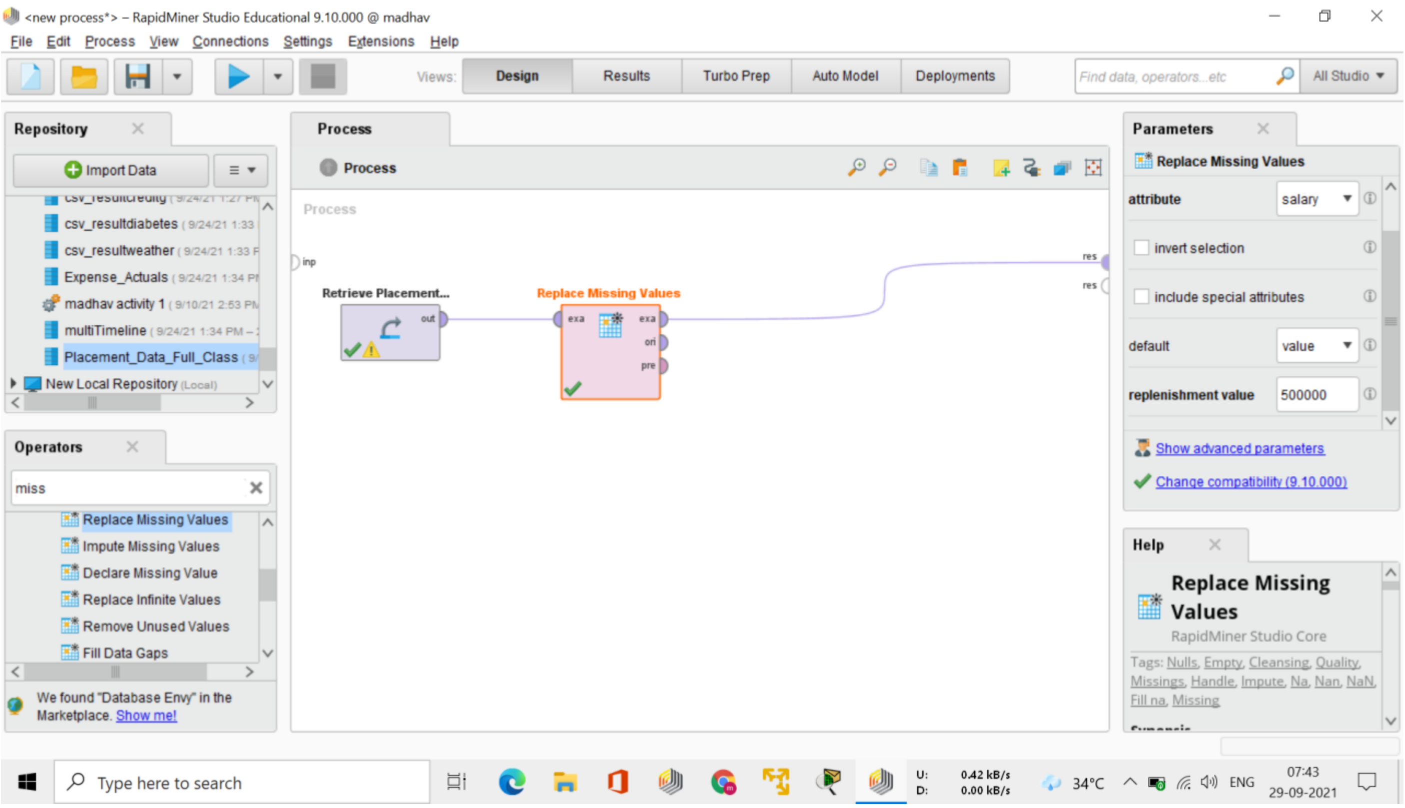
Row No.	Customer ID	Product ID	Amount
1	370	154	3
2	41	40	3
3	109	173	3
4	556	11	4
5	143	72	1
6	18	26	1
7	158	50	4
8	237	98	3
9	456	161	3

ExampleSet (2,328 examples, 0 special attributes, 3 regular attributes)

- Replace Missing Values:-

This Operator replaces missing values in Examples of selected Attributes by a specified replacement.Missing values can be replaced by the minimum, maximum or average value of that Attribute.zero can also be used to replace missing values. Any replenishment value can also be specified as a replacement of missing values.

Figure :-7 ,shows the replace missing values operator preview.



<new process*> – RapidMiner Studio Educational 9.10.000 @ madhav
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import Data
csv_resultcredit
csv_resultdiabetes (9/24/21 1:33
csv_resultweather (9/24/21 1:33
Expense_Actuals (9/24/21 1:34 PM
madhav activity 1 (9/10/21 2:53 PM
multiTimeline (9/24/21 1:34 PM
Placement_Data_Full_Class
New Local Repository (Local)
Operators
miss
Replace Missing Values
Impute Missing Values
Declare Missing Value
Replace Infinite Values
Remove Unused Values
Fill Data Gaps
Process
Process
Process
inp
res
res
Retrieve Placement...
out
Replace Missing Values
exa
exa
ori
pre
Parameters
Replace Missing Values
Show advanced parameters
Change compatibility (9.10.000)
Help
Replace Missing Values
RapidMiner Studio Core
Synopsis
Type here to search
U: 0.42 kB/s
D: 0.00 kB/s
34°C
ENG
07:43
29-09-2021

We found "Database Envy" in the Marketplace. Show me!

attribute: salary

□ invert selection

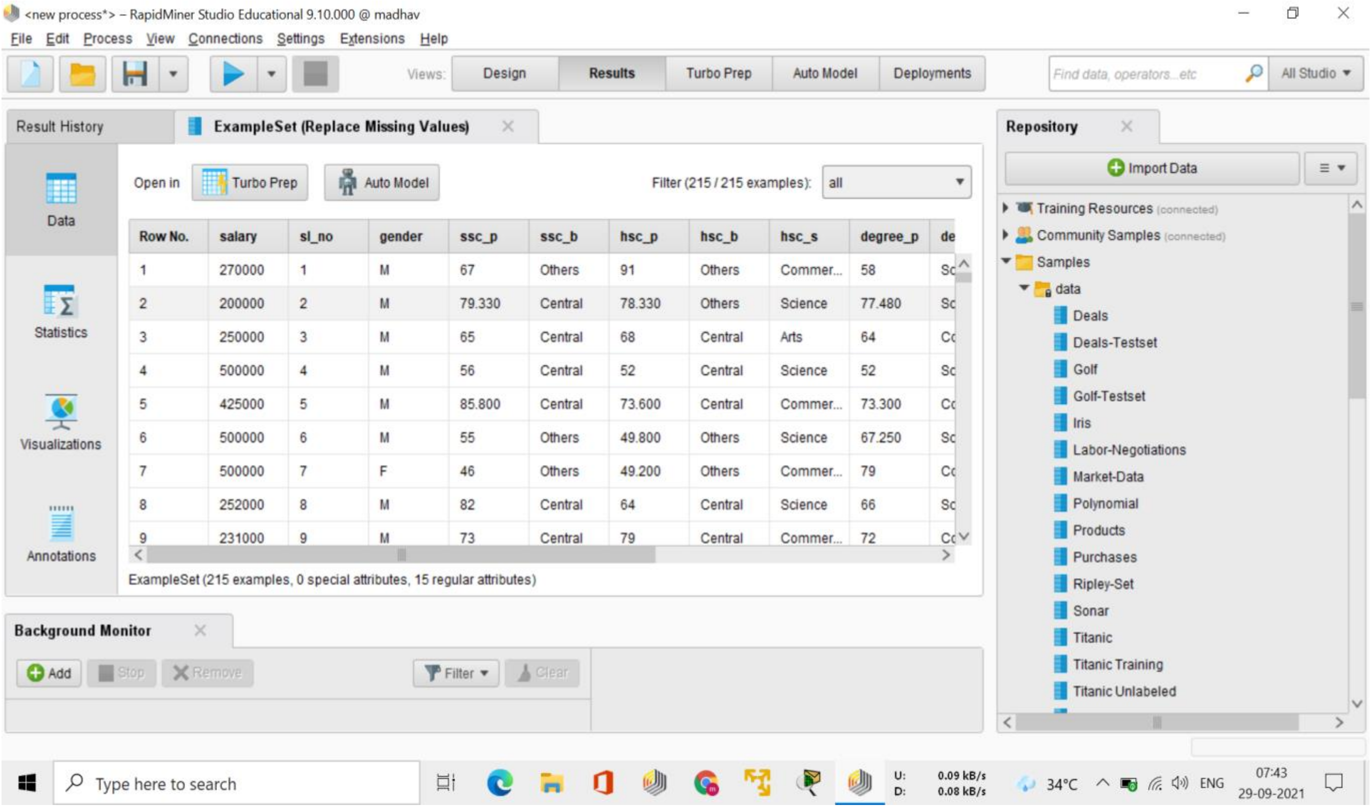
□ include special attributes

default: value

replenishment value: 500000

Tags: Nulls, Empty, Cleansing, Quality, Missings, Handle, Impute, Na, Nan, NaN, Fill na, Missing

Figure :-8 ,shows the select missing values operatoration on placement data set,in which i had some missing values in salary column and replacing missing values with 50000.

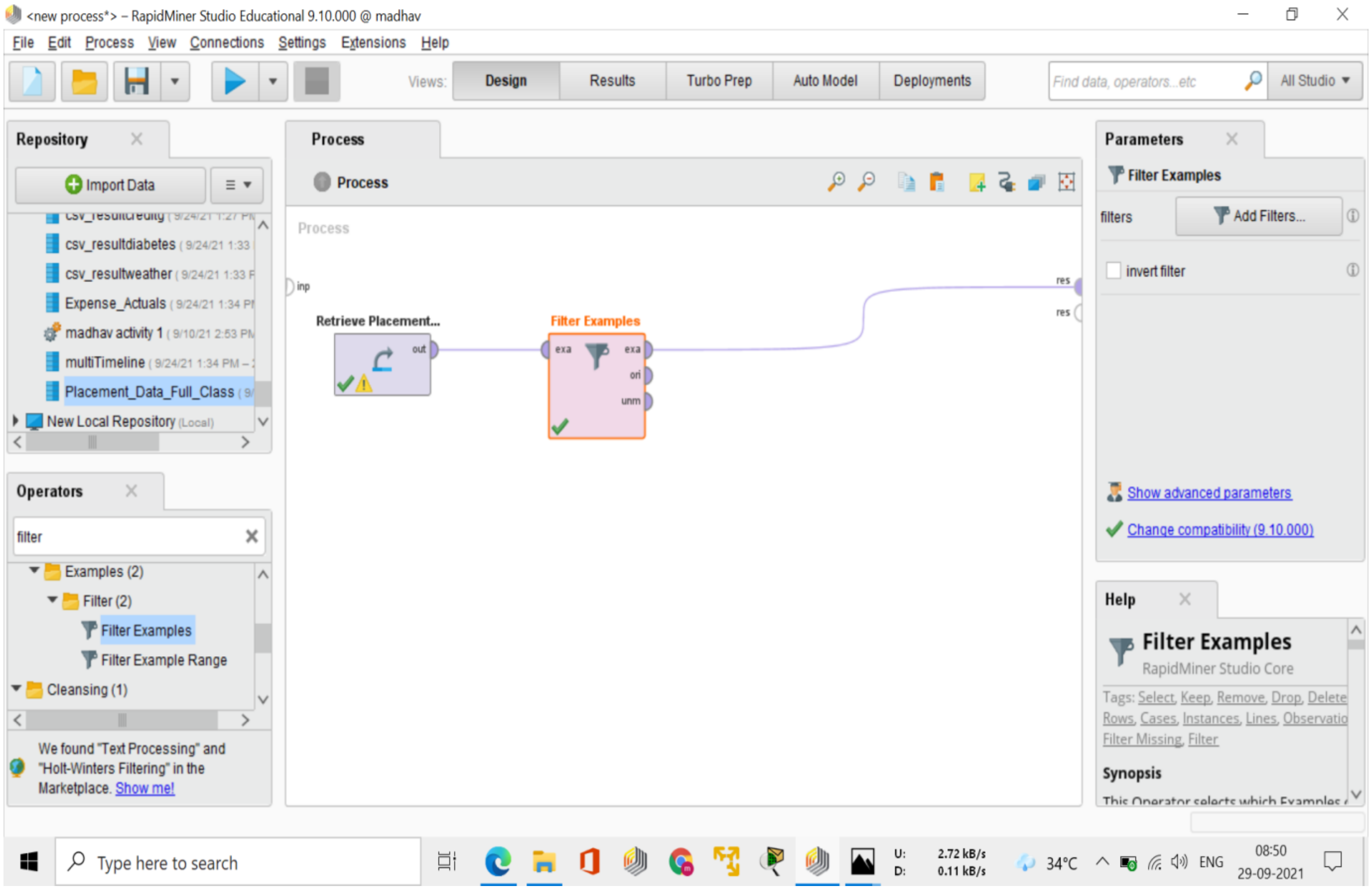

<new process*> – RapidMiner Studio Educational 9.10.000 @ madhav
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
ExampleSet (Replace Missing Values)
Data
Statistics
Visualizations
Annotations
Open in Turbo Prep Auto Model
Filter (215 / 215 examples): all
Repository
Import Data
Training Resources (connected)
Community Samples (connected)
Samples
data
Deals
Deals-Testset
Golf
Golf-Testset
Iris
Labor-Negotiations
Market-Data
Polynomial
Products
Purchases
Ripley-Set
Sonar
Titanic
Titanic Training
Titanic Unlabeled
Background Monitor
Add Stop Remove Filter Clear
Type here to search
U: 0.09 kB/s
D: 0.08 kB/s
34°C ENG
07:43
29-09-2021

Row No.	salary	sl_no	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	de
1	270000	1	M	67	Others	91	Others	Commer...	58	Sc
2	200000	2	M	79.330	Central	78.330	Others	Science	77.480	Sc
3	250000	3	M	65	Central	68	Central	Arts	64	Cc
4	500000	4	M	56	Central	52	Central	Science	52	Sc
5	425000	5	M	85.800	Central	73.600	Central	Commer...	73.300	Cc
6	500000	6	M	55	Others	49.800	Others	Science	67.250	Sc
7	500000	7	F	46	Others	49.200	Others	Commer...	79	Cc
8	252000	8	M	82	Central	64	Central	Science	66	Sc
9	231000	9	M	73	Central	79	Central	Commer...	72	Cc

ExampleSet (215 examples, 0 special attributes, 15 regular attributes)

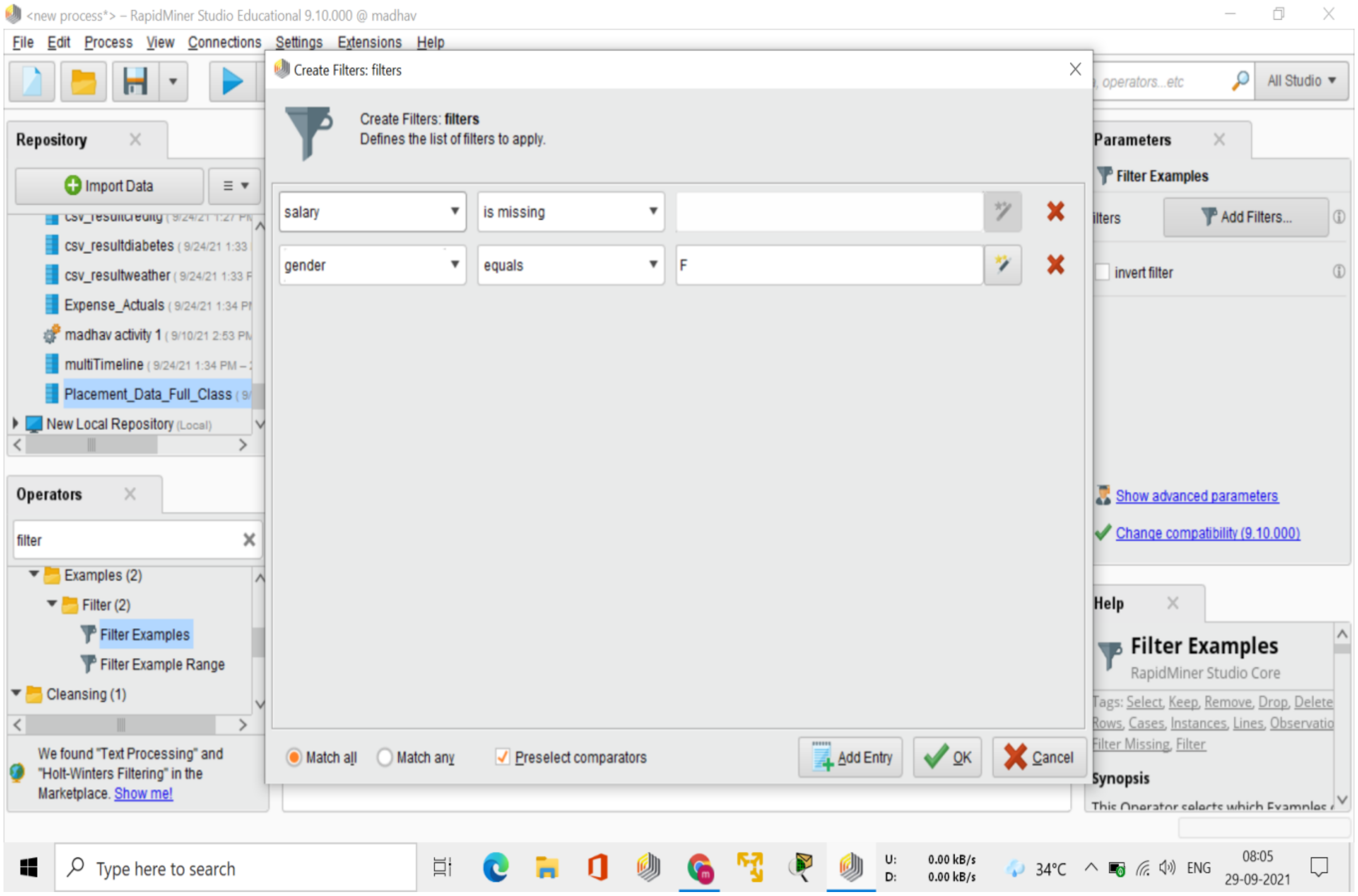
- Filter examples:-

Figure :-9 ,shows the filter examples operator preview.



We found "Text Processing" and "Holt-Winters Filtering" in the Marketplace. Show me!

Figure :-10 ,shows the filter examples operatoration on salary attribute whose values are missing and gender is equal to female.



<new process*> – RapidMiner Studio Educational 9.10.000 @ madhav
File Edit Process View Connections Settings Extensions Help
Repository
Import Data
csv_resultcredit
csv_resultdiabetes (9/24/21 1:33
csv_resultweather (9/24/21 1:33
Expense_Actuals (9/24/21 1:34 PM
madhav activity 1 (9/10/21 2:53 PM
multiTimeline (9/24/21 1:34 PM
Placement_Data_Full_Class
New Local Repository (Local)
Operators
filter
Examples (2)
Filter (2)
Filter Examples
Filter Example Range
Cleansing (1)
We found "Text Processing" and "Holt-Winters Filtering" in the Marketplace. Show me!
Parameters
Filter Examples
filters
Add Filters...
invert filter
Show advanced parameters
Change compatibility (9.10.000)
Help
Filter Examples
RapidMiner Studio Core
Tags: Select, Keep, Remove, Drop, Delete Rows, Cases, Instances, Lines, Observatio Filter Missing, Filter
Synopsis
This Operator selects which Examples
Add Entry
OK
Cancel
Type here to search
U: 0.00 kB/s
D: 0.00 kB/s
34°C
ENG
08:05
29-09-2021

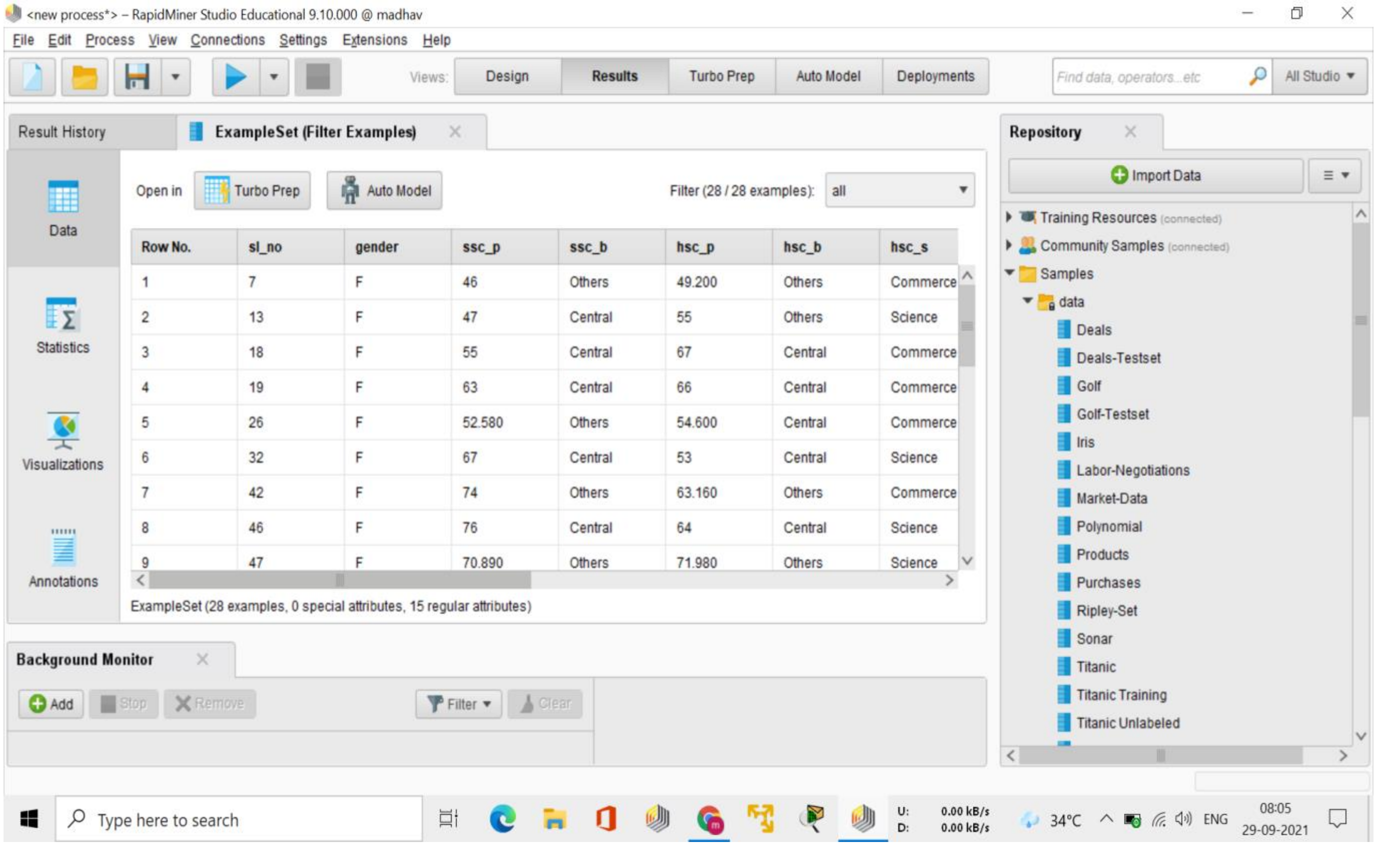
Create Filters: filters

Create Filters: **filters**
Defines the list of filters to apply.

salary	is missing	
gender	equals	F

● Match all ○ Match any ☑ Preselect comparators

Figure :-11 ,shows the filter examples operatoration on gender values which is equal to female.

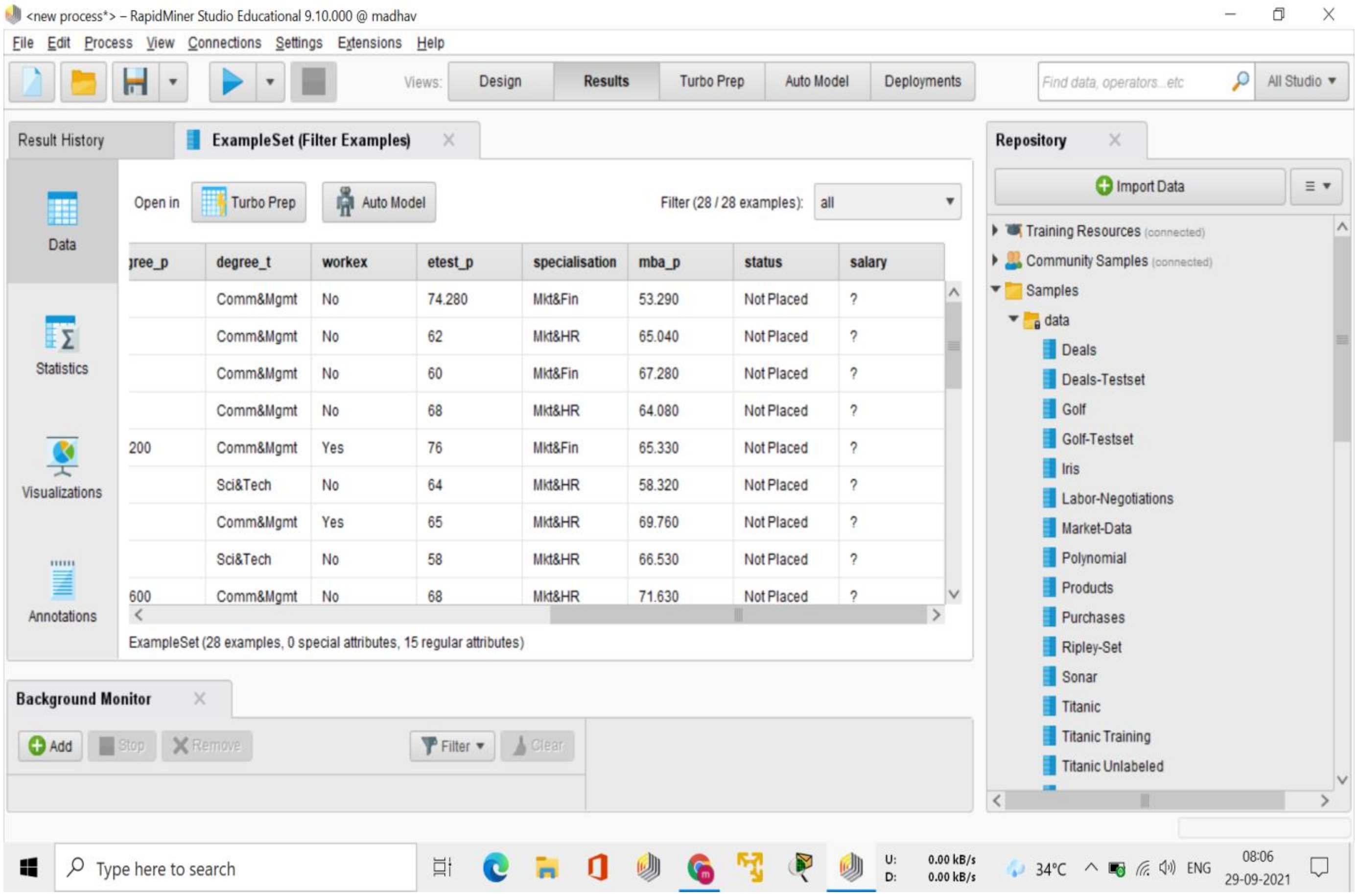

<new process*> – RapidMiner Studio Educational 9.10.000 @ madhav
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
ExampleSet (Filter Examples)
Data
Statistics
Visualizations
Annotations
Open in Turbo Prep Auto Model
Repository
Import Data
Training Resources (connected)
Community Samples (connected)
Samples
data
Deals
Deals-Testset
Golf
Golf-Testset
Iris
Labor-Negotiations
Market-Data
Polynomial
Products
Purchases
Ripley-Set
Sonar
Titanic
Titanic Training
Titanic Unlabeled
Background Monitor
Add Stop Remove Filter Clear
Type here to search
U: 0.00 kB/s
D: 0.00 kB/s
34°C ENG
08:05
29-09-2021

Filter (28 / 28 examples): all

Row No.	sl_no	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s
1	7	F	46	Others	49.200	Others	Commerce
2	13	F	47	Central	55	Others	Science
3	18	F	55	Central	67	Central	Commerce
4	19	F	63	Central	66	Central	Commerce
5	26	F	52.580	Others	54.600	Central	Commerce
6	32	F	67	Central	53	Central	Science
7	42	F	74	Others	63.160	Others	Commerce
8	46	F	76	Central	64	Central	Science
9	47	F	70.890	Others	71.980	Others	Science

ExampleSet (28 examples, 0 special attributes, 15 regular attributes)

Figure :-12 ,shows the filter examples operatoration on missing values of salary.


<new process*> – RapidMiner Studio Educational 9.10.000 @ madhav
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
ExampleSet (Filter Examples)
Data
Statistics
Visualizations
Annotations
Open in Turbo Prep Auto Model
Filter (28 / 28 examples): all
Repository
Import Data
Training Resources (connected)
Community Samples (connected)
Samples
data
Deals
Deals-Testset
Golf
Golf-Testset
Iris
Labor-Negotiations
Market-Data
Polynomial
Products
Purchases
Ripley-Set
Sonar
Titanic
Titanic Training
Titanic Unlabeled
Background Monitor
Add Stop Remove Filter Clear
Type here to search
U: 0.00 kB/s
D: 0.00 kB/s
34°C ENG
08:06
29-09-2021

ree_p	degree_t	workex	etest_p	specialisation	mba_p	status	salary
	Comm&Mgmt	No	74.280	Mkt&Fin	53.290	Not Placed	?
	Comm&Mgmt	No	62	Mkt&HR	65.040	Not Placed	?
	Comm&Mgmt	No	60	Mkt&Fin	67.280	Not Placed	?
	Comm&Mgmt	No	68	Mkt&HR	64.080	Not Placed	?
200	Comm&Mgmt	Yes	76	Mkt&Fin	65.330	Not Placed	?
	Sci&Tech	No	64	Mkt&HR	58.320	Not Placed	?
	Comm&Mgmt	Yes	65	Mkt&HR	69.760	Not Placed	?
	Sci&Tech	No	58	Mkt&HR	66.530	Not Placed	?
600	Comm&Mgmt	No	68	Mkt&HR	71.630	Not Placed	?

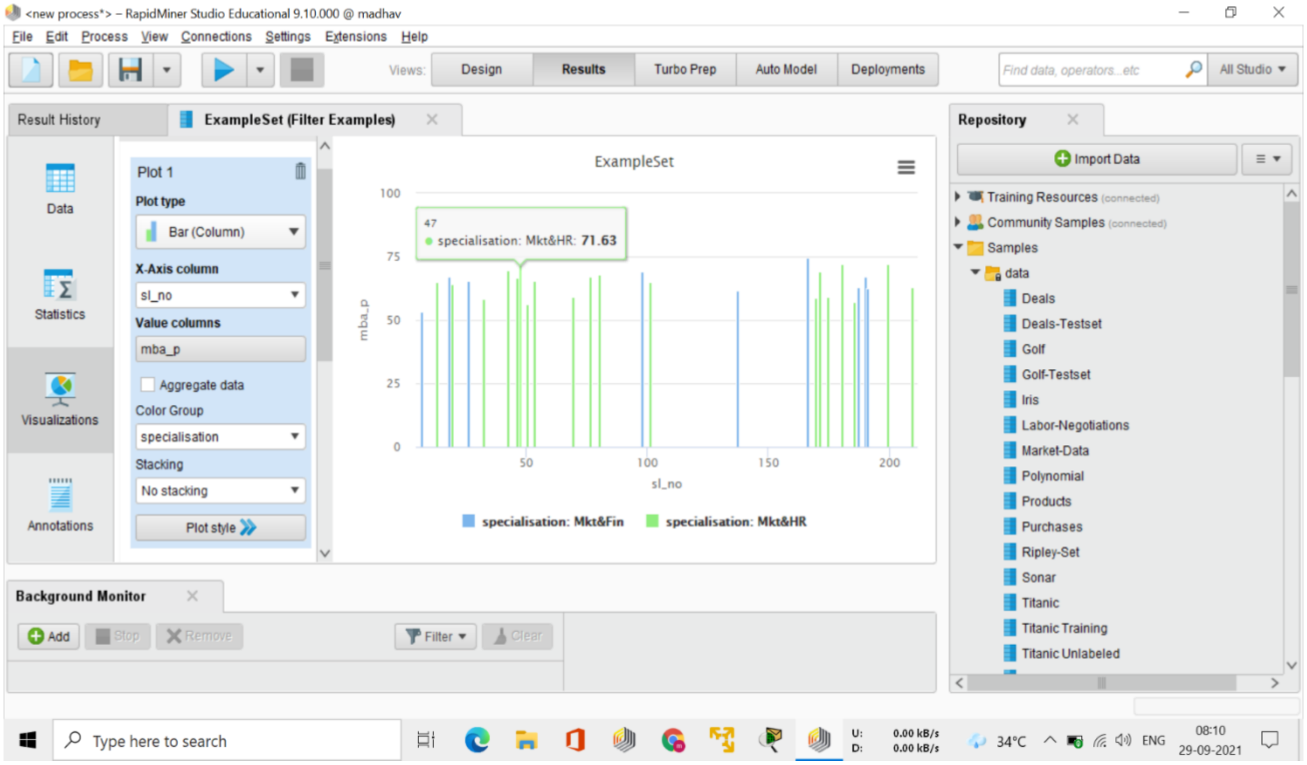
ExampleSet (28 examples, 0 special attributes, 15 regular attributes)

Q2(B).What do you understand by graphical representation and statistics in rapidminer. Attach any 5 different types of graph and their interpretation.

Answer:-

A statistical graph or chart is defined as the pictorial representation of statistical data in graphical form. The statistical graphs are used to represent a set of data to make it easier to understand and interpret statistical information.

Graphical representation and statical based analysis helps us to view data in different view.


Plot 1

Plot type: Bar (Column)

X-Axis column: sl_no

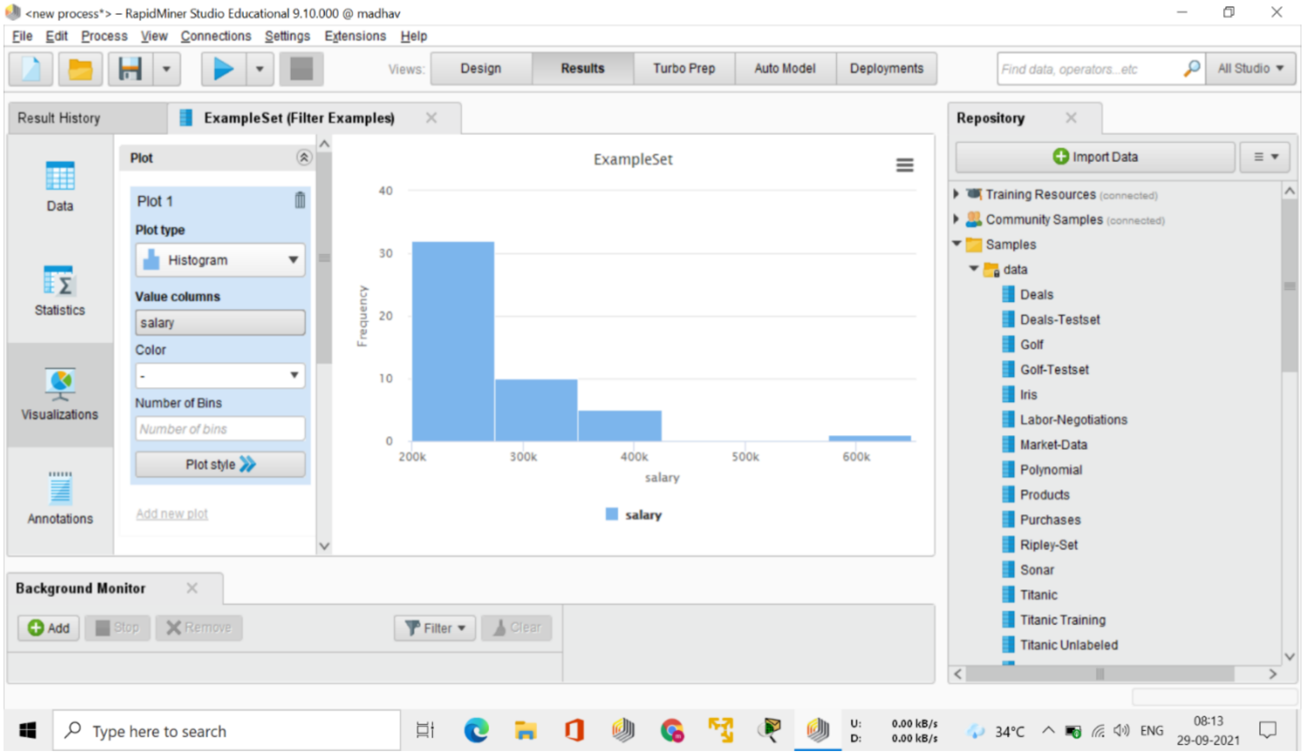
Value columns: mba_p

□ Aggregate data

Color Group: specialisation

Stacking: No stacking

47
specialisation: Mkt&HR: 71.63



<new process*> – RapidMiner Studio Educational 9.10.000 @ madhav

ExampleSet (Filter Examples)

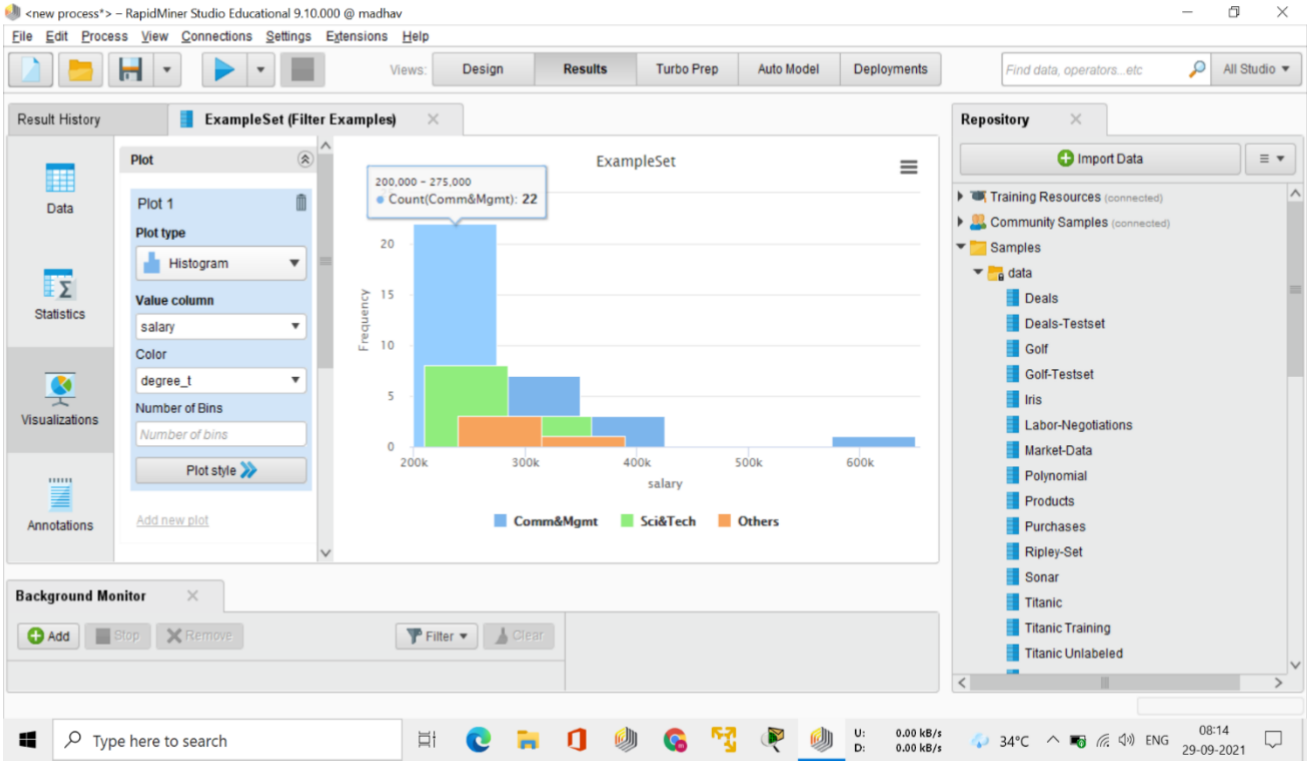
Plot 1

Plot type: Histogram

Value columns: salary

Color: -

Number of Bins: Number of bins



<new process*> – RapidMiner Studio Educational 9.10.000 @ madhav

ExampleSet (Filter Examples)

Plot 1

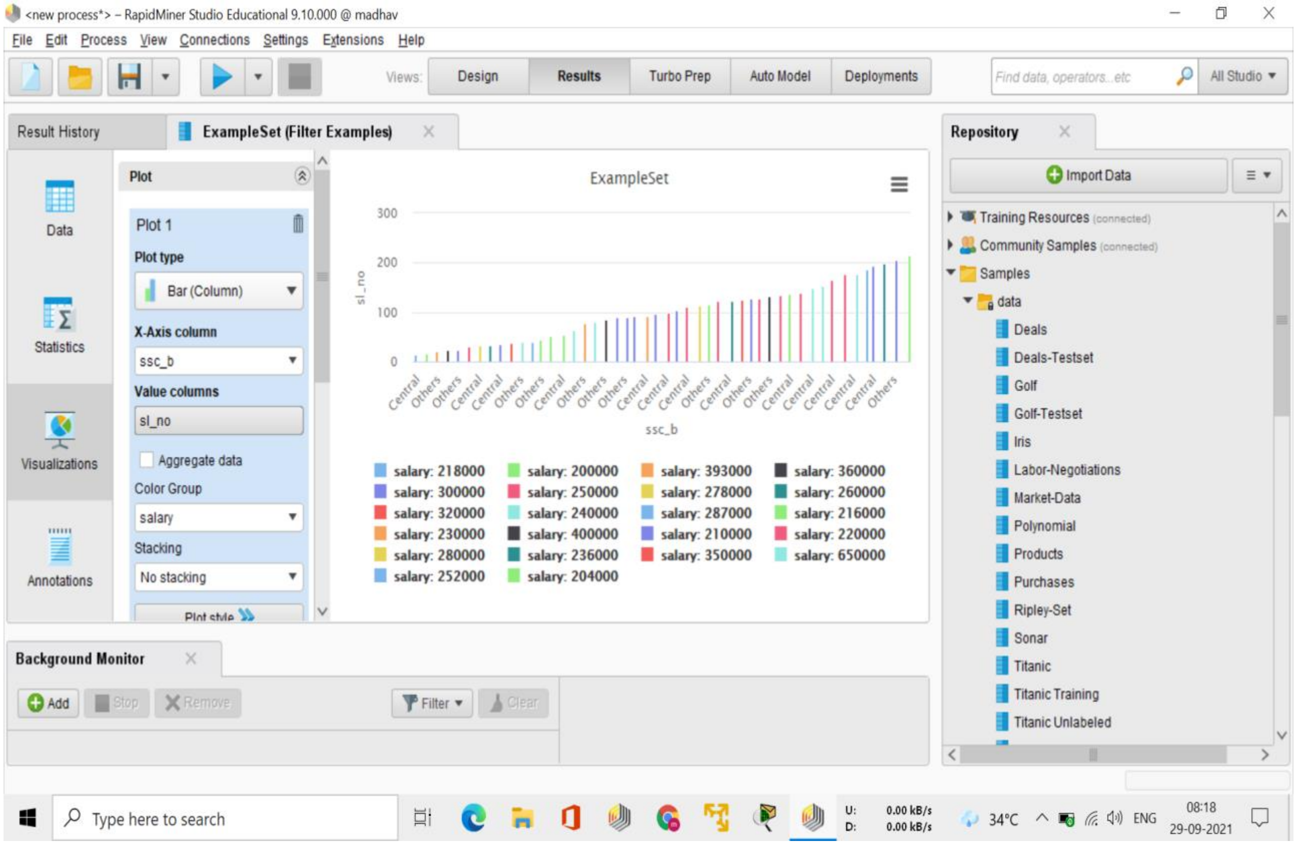
Plot type: Histogram

Value column: salary

Color: degree_t

Number of Bins: Number of bins

200,000 – 275,000
● Count(Comm&Mgmt): 22



Plot type: Bar (Column)

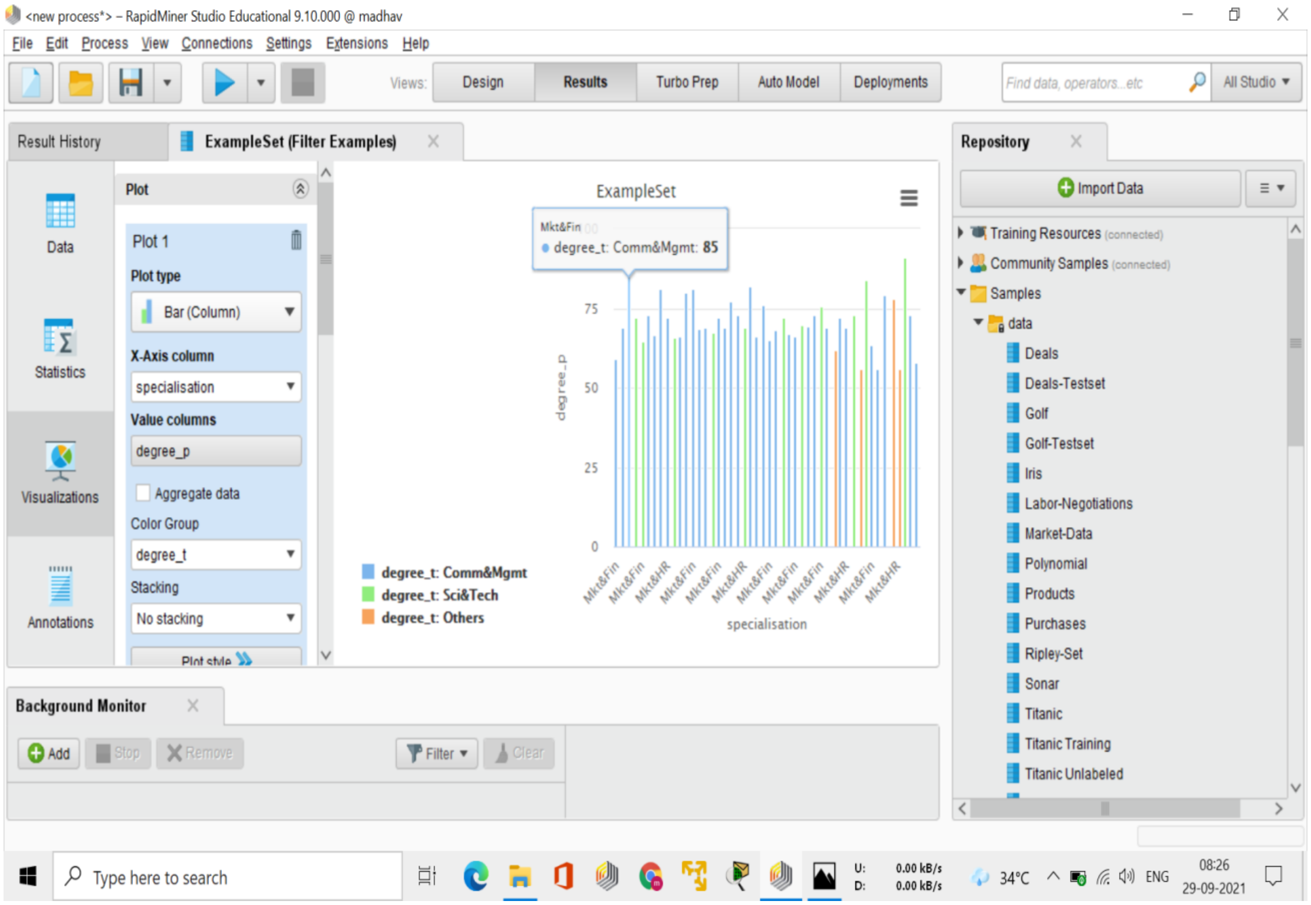
X-Axis column: ssc_b

Value columns: sl_no

□ Aggregate data

Color Group: salary

Stacking: No stacking



Plot 1

Plot type: Bar (Column)

X-Axis column: specialisation

Value columns: degree_p

□ Aggregate data

Color Group: degree_t

Stacking: No stacking

Plot style

Q3.What do you mean by preprocessing in Data Mining. Explain any 5 operations to handle missing values and attach screenshots of same.

Answer:-

Data preprocessing involves transforming raw data to well-formed data sets so that data mining analytics can be applied. Raw data is often incomplete and has inconsistent formatting .

Preprocessing involves both data validation and data imputation. The goal of data validation is to assess whether the data in question is both complete and accurate. The goal of data imputation is to correct errors and input missing values , either manually or automatically.

- There are few steps to preprocess:-
- Data cleaning
- Data integration
- Data transformation
- Data reduction
- Data DDiscretization

>Data cleaning:-

In data cleaning process we fill the missing,identify outliers and remove them,smooth out noisy data,and correct inconsistent data sets.

>Data integration: combines data from multiple sources into a coherent store.Careful integration can help reduce and avoid redundancies and inconsistencies in resulting data set.This can help improve the accuracy and speed of the subsequent data mining process.This is referred to as the entity identification problem.

>Data Transformation:-

Data is normalized and generalized. Normalization is a process that ensures that no data is redundant, it is all stored in a single place, and all the dependencies are logical.

>Data Reduction:-

When the volume of data is huge, databases can become slower, costly to access, and challenging to properly store. Data reduction aims to present a reduced representation of the data in a data warehouse.

Data reduction strategies:-

1)Dimensionality reduction

2)Numerosity reduction

3)Data compression

>Discretization :-
Reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values.

>Five operations to handle missing values:-