Assignment 1

Course Code - CAP447

Course Teacher – Dr. Punam Rattan

Name – Aniket Kumar

**Registration Number -** 12108348

Section - D2112

Roll Number - RD2112B81

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

Rapid Miner studio is a platform to analyze and perform various operations on a data set to obtain the useful informations, graphical representations of data like line chart, bar chart, pie chart, etc.

Following are some main components of RapidMiner studio:

- **Repository**
    - It is the place where the data, processes and results are stored, either locally or remotely.
    - It is the collection of datasets, also known as folder, workspace or project.
    - We can import external datasets too and add to the repository.
    - The Repository can be used to store:
        - Data
        - Processes
   - Results
        - Operators
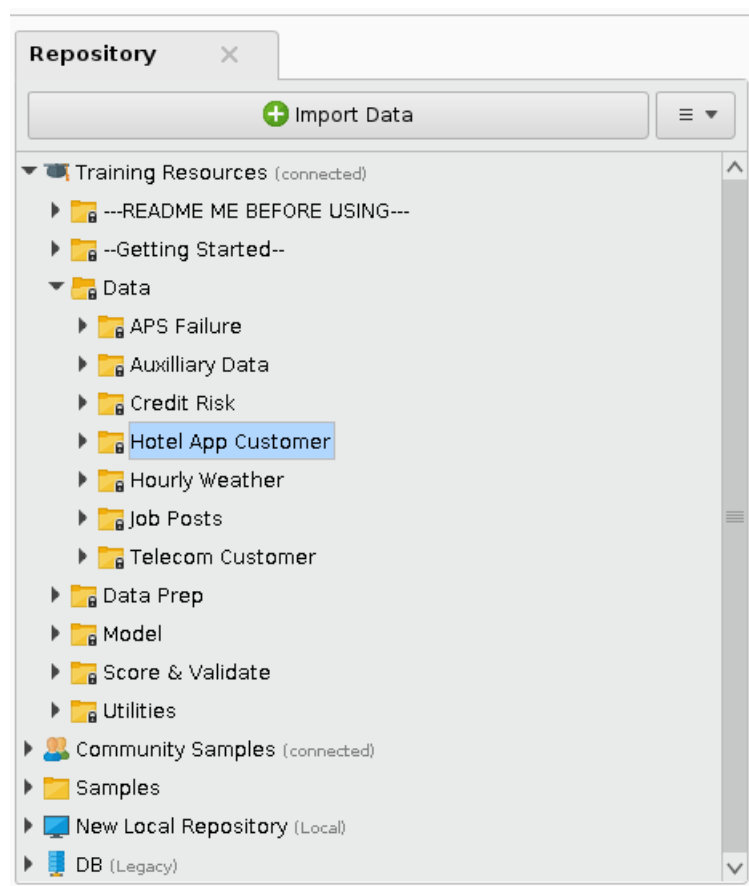
- **Operators**
    - This section contains various kinds of operators that are used to perform different operations.
    - Some of the operators are sort, join, aggregate, select attribute, etc.
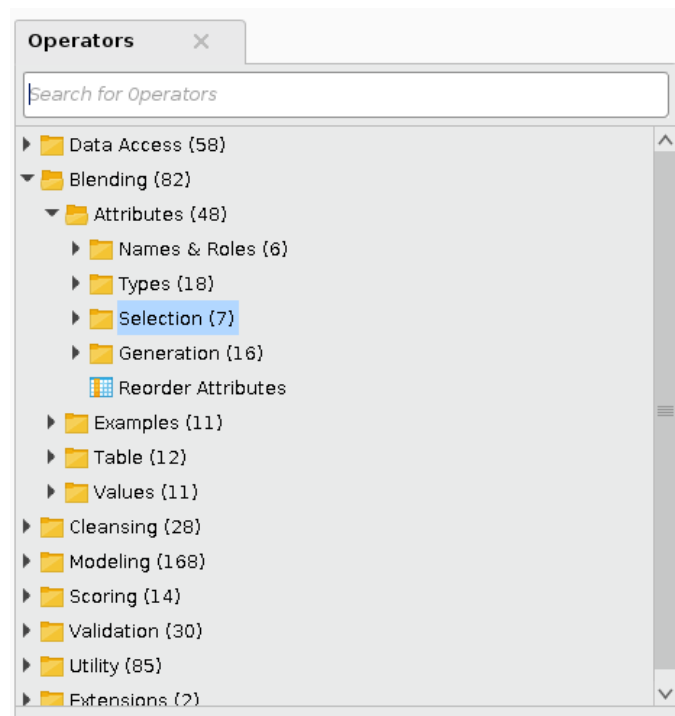
- **Views**
  There are five kinds of view in Rapidminer:
  - **Design**
    - It is the default view, which contains the **Process.**
    - In Process, we drag different datasets and operators and connect them.
  - **Results**
    - After Running the process, we get the results in Results section, in which there are Data, Statistics, Visualization and annotation.
  - **Turbo Prep**
    - It is used for blending and cleaning the data as well as to build data preparation process automatically.
  - **Auto Model**
    - It is used to create models.
  - **Deployments**
   - It is used to manage projects and deployed models.
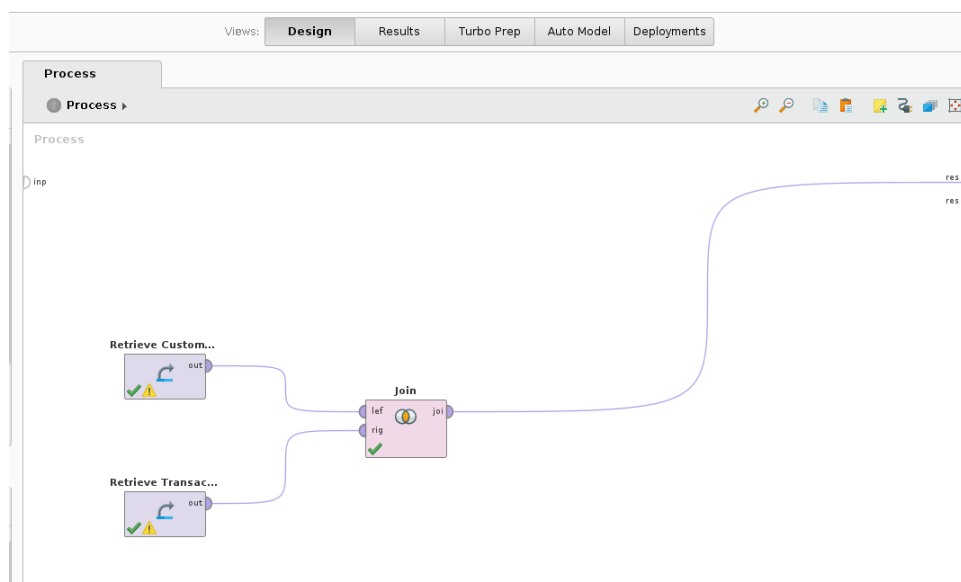
- **Parameters**
  - This section is used to configure the behavior of different operators.
  - It's content is displayed when we select a particular operator.
  - By default, RapidMiner Studio shows only the more commonly used parameters.
  - To see all of the available parameters, we have to click Show advanced parameters.
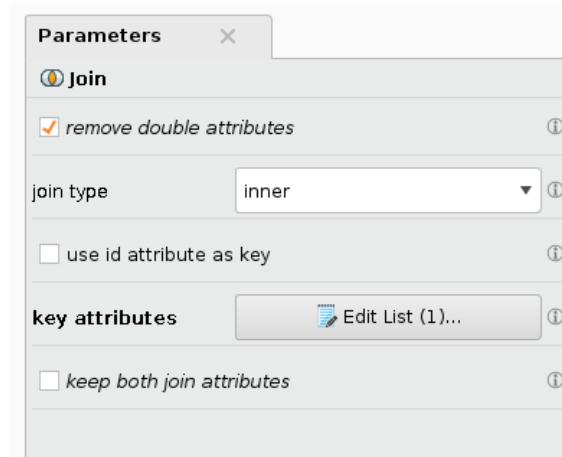
- **Help**
  - It displays a help text for the current Operator.
  - The content of the Help Panel is also context-dependent.
  - If we select any Operator that is displayed in the Process Panel, the Help Panel displays a help text for that Operator.
  - The Help Panel provides useful background informations.
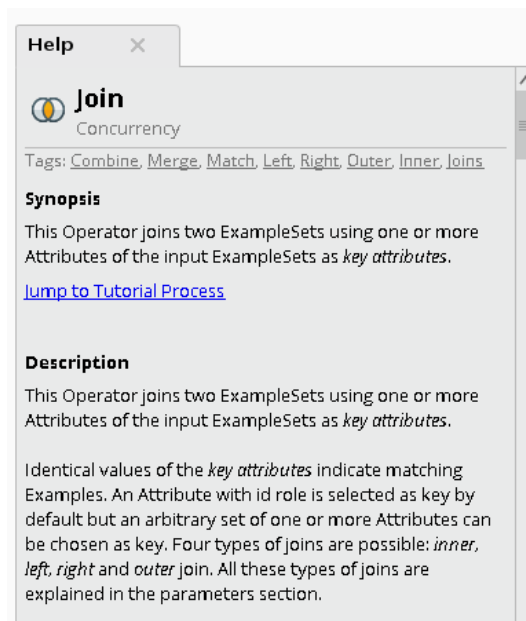
These were the components of the RapidMiner that are used together to perform various operations on the datasets and get the meaningful result.

2.

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

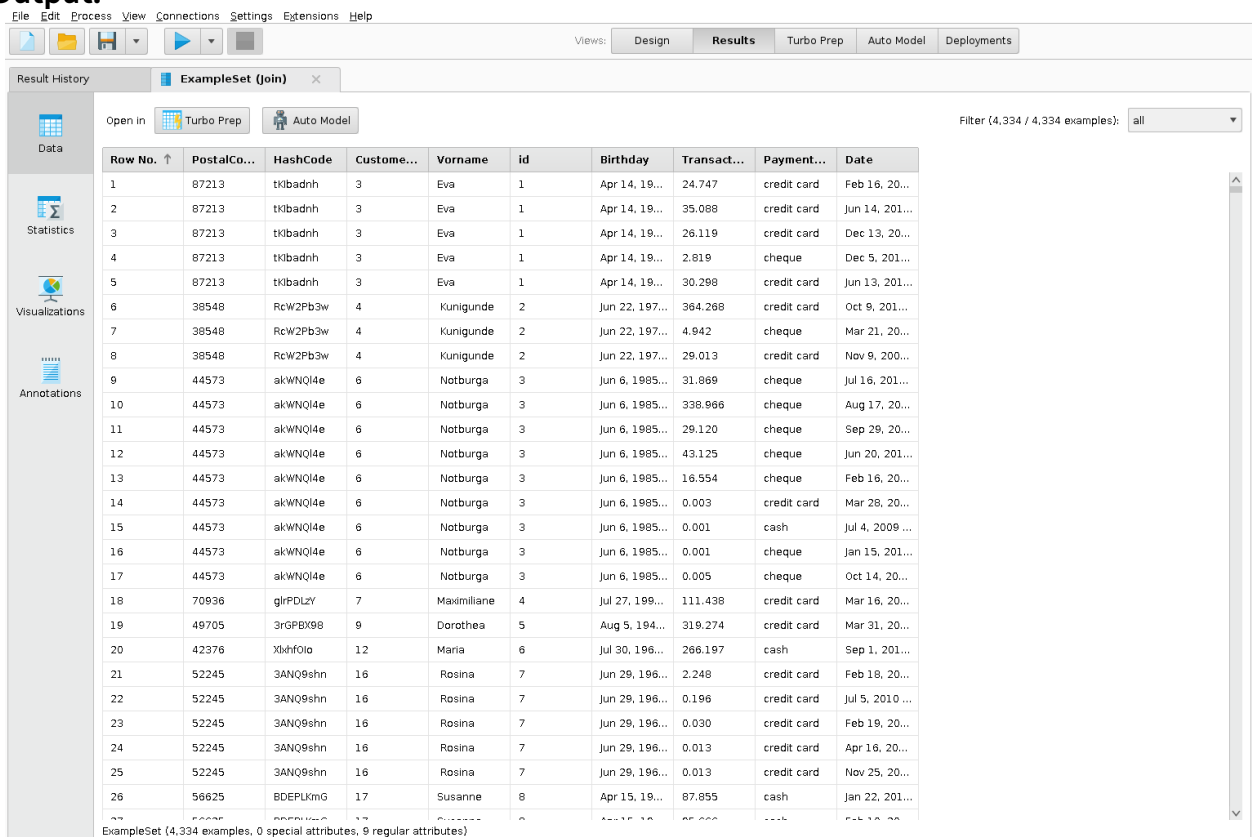**Data sets used :** Customer Data and Transaction Data
Path: Repository → Training Resources → Data → Hotel App Customer

(I). Join – This operator joins two Data Sets using one or more attributes that are key attributes in both the Data sets.
Steps:
- Drag the Customer Data and Transaction Data into the process.
- Search for sort operator and drag it into the process.
- Connect the out ports of Customer Data and Transaction Data into the input port of Join operator.
- Click on join to select it.
- In the Parameters of join, go to the edit list of key attributes.
- Select the key attributes in both side, here CustomerId.
- Now, connect the output port of join operator to the result port.
- Press Run to execute the process.

Output:

(ii). Aggregate - The Aggregate operator creates a new ExampleSet from the input ExampleSet showing the results of the selected aggregation functions. Many aggregation functions are supported including SUM, COUNT, MIN, MAX, AVERAGE and many other similar functions known from SQL.

Steps:

- Drag the Customer Data and Transaction Data into the process.
- Search for sort operator and drag it into the process.
- Connect the out ports of Customer Data and Transaction Data into the input port of Join operator.
- Click on join to select it.
- In the Parameters of join, go to the edit list of key attributes.
- Select the key attributes in both side, here CustomerId.
- Search for Aggregate operator and drag it to the process.
- Now, connect the output port of join operator to the input port of Aggregate operator.
- In the Parameters of Aggregate, go to edit list of group by attributes and select the postal code.
- Now, connect the output port of Aggregate to the result port.
- Press Run to execute the process.

Output:
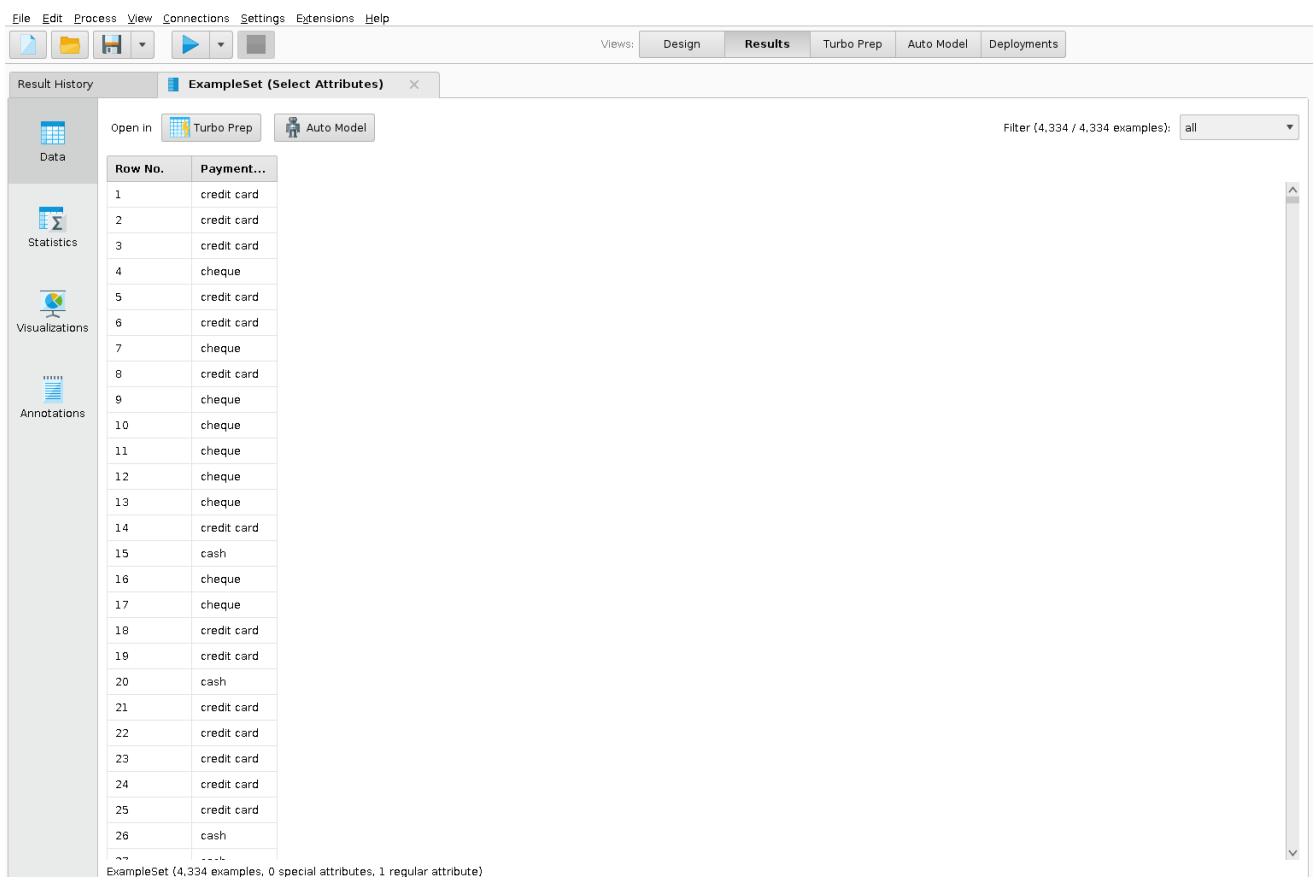
(iii). Select Attributes - This Operator selects a subset of Attributes of an ExampleSet and removes the other Attributes.

Steps:

- Drag the Customer Data and Transaction Data into the process.
- Search for sort operator and drag it into the process.
- Connect the out ports of Customer Data and Transaction Data into the input port of Join operator.
- Click on join to select it.
- In the Parameters of join, go to the edit list of key attributes.
- Select the key attributes in both side, here CustomerId.
- Search for Select Attribute operator and drag it to the process.
- Now, connect the output port of join operator to the input port of Select Attribute operator.
- In the Parameters of Select Attribute, set attribute filter type to single and attribute to Payment Method.
- Now, connect the output port of Select Attribute to the result port.
- Press Run to execute the process.

Output:

(iv). Sort - This operator sorts the input data set in ascending or descending order according to several attributes.

Steps:

- Drag the Customer Data and Transaction Data into the process.
- Search for sort operator and drag it into the process.
- Connect the out ports of Customer Data and Transaction Data into the input port of Join operator.
- Click on join to select it.
- In the Parameters of join, go to the edit list of key attributes.
- Select the key attributes in both side, here CustomerId.
- Search for Sort operator and drag it to the process.
- Now, connect the output port of join operator to the input port of sort operator.
- In the Parameters of sort, go to the edit list of sort by and select one attribute and ascending or descending.
- Now, connect the output port of sort to the result port.
- Press Run to execute the process.

Output:

(v). Descretize by Binning - This operator discretizes the selected numerical attributes into user-specified number of bins. Bins of equal range are automatically generated, the number of the values in different bins may vary.

Steps:

- Drag the Customer Data and Transaction Data into the process.
- Search for sort operator and drag it into the process.
- Connect the out ports of Customer Data and Transaction Data into the input port of Join operator.
- Click on join to select it.
- In the Parameters of join, go to the edit list of key attributes.
- Select the key attributes in both side, here CustomerId.
- Search for Descreticize by Binning operator and drag it to the process.
- Now, connect the output port of join operator to the input port of sort operator.
- In the Parameters of Descreticize by Binning, select attribute type to single, attribute to TransactionValue and number of bins to 3.
- Now, connect the output port of Descreticize by Binning to the result port.
- Press Run to execute the process.

Output:

2b. What do you understand by graphical representation and statistics in rapidminer. Attach any 5 different types of graph and their interpretation.

Graphical Represenation:
- It 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:

- **Data sets used :** Customer Data and Transaction Data
- **Path:** Repository → Training Resources → Data → Hotel App Customer

  - **Line Graphs** – It s a type of chart which displays information as a series of data points called 'markers' connected by a straight line segments. Line graph or the linear graph is used to display the continuous data and it is useful for predicting future events over time.
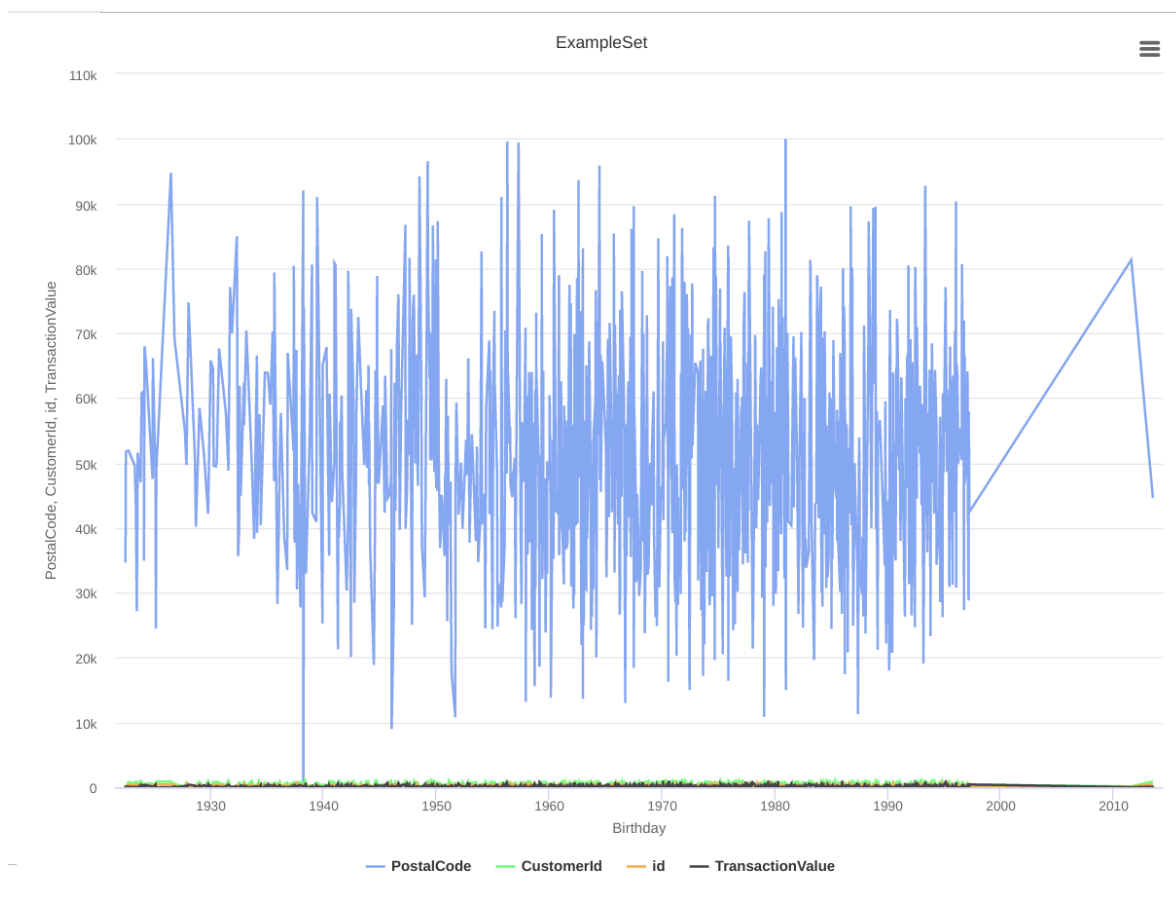
  - **Bar Graphs** – Bar Graph is used to display the category of data and it compares the data using solid bars to represent the quantities.
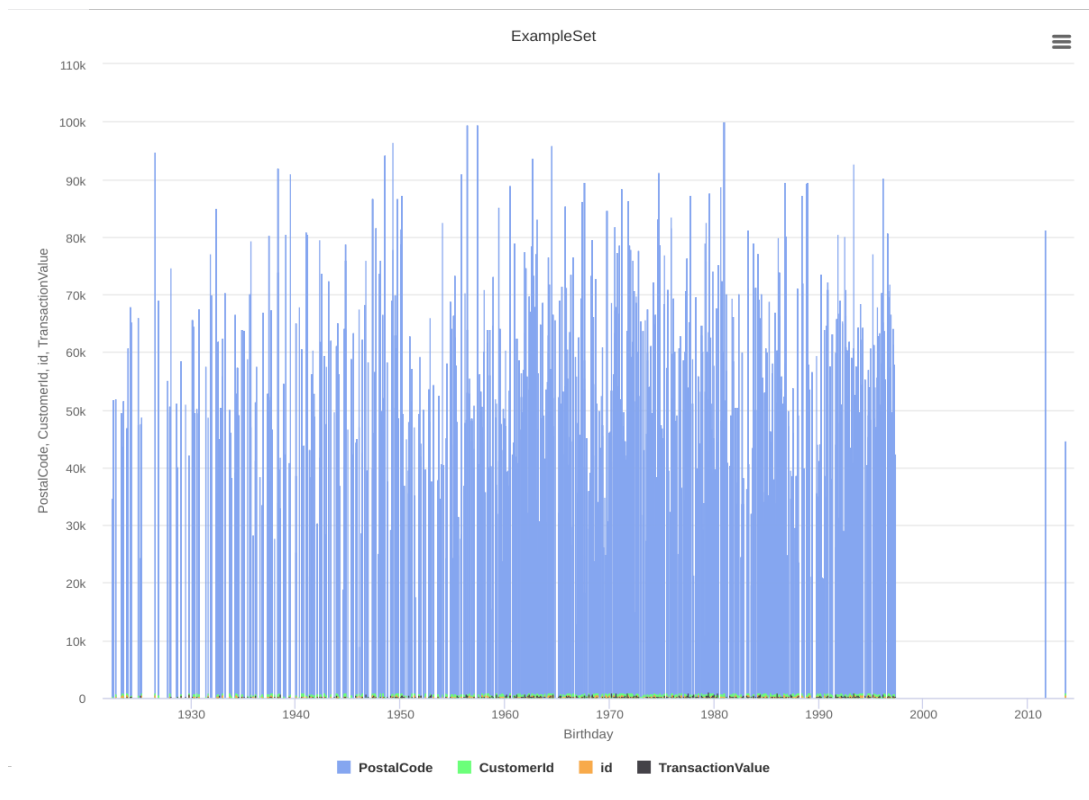
- **Histograms** – The graph that uses bars to represent the frequency of numerical data that are organised into intervals. Since all the intervals are equal and continuous, all the bars have the same width.
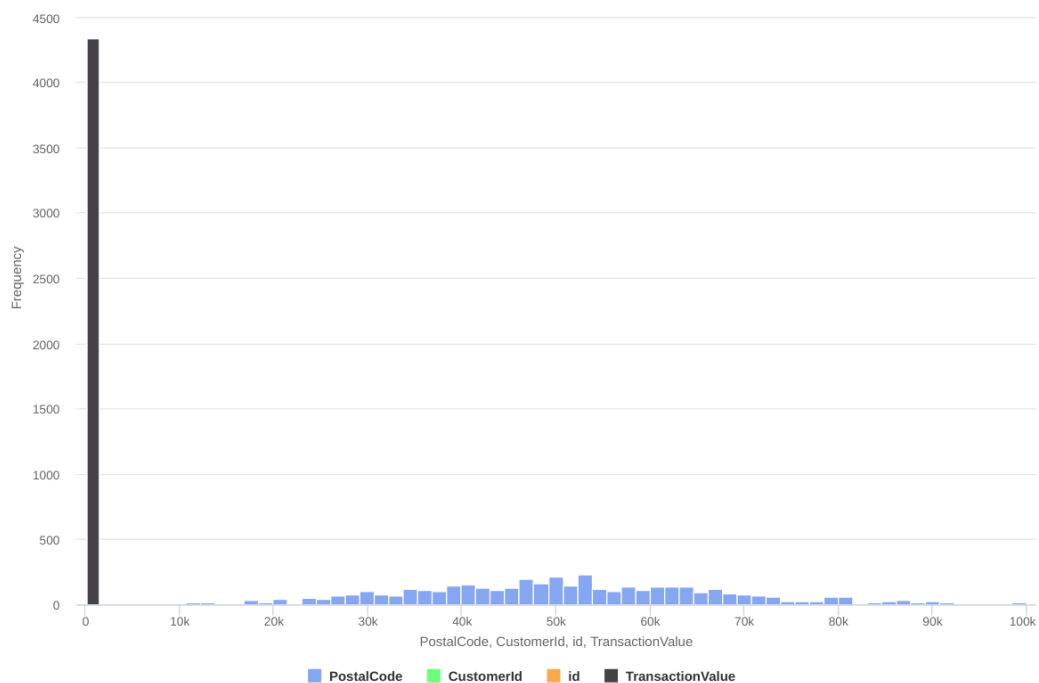
- **Scatter plot** - A scatter plot uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe

relationships between variables.
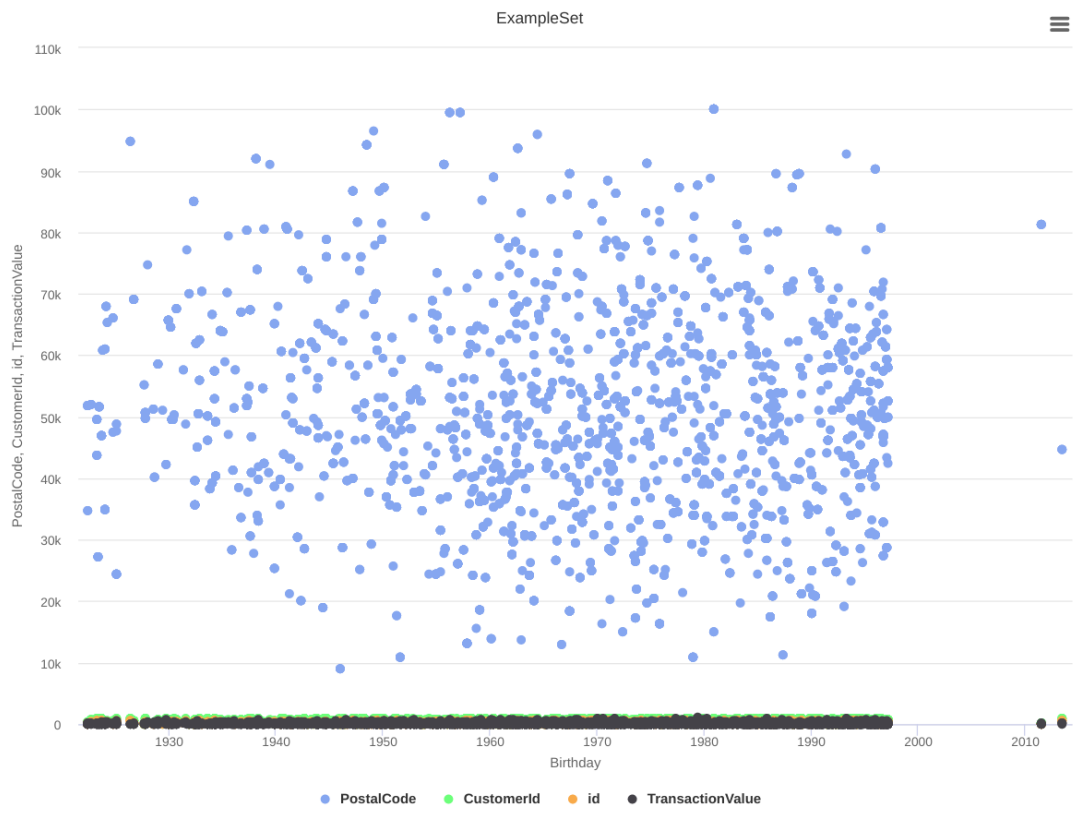
- **Pie Chart –** It shows the relationships of the parts of the whole. The circle is considered with 100% and the categories occupied is represented with that specific percentage like 15%, 56%, etc.
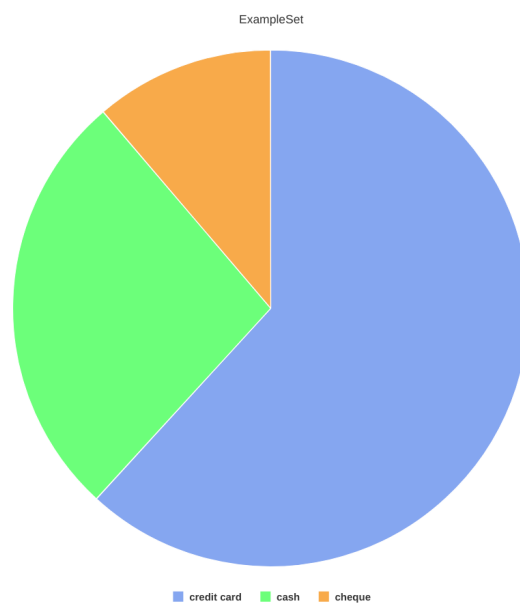
Statistics in RapidMiner:

Statistics in Rapidminer gives the output after analyzing the data set, which contains class labels, no of missing values, minimum, maximum and average of a particular attribute.

Name	Type	Missing	Statistics		
PostalCode	Integer	0	Min 1	Max 99999	Average 51297.960
HashCode	Nominal	0	Least zPlL3DcR (1)	Most cujNb6GQ (20)	Values cujNb6GQ (20), wHh275mF (20), ...[869 more]
CustomerId	Integer	0	Min 1	Max 1000	Average 489.938
Vorname	Nominal	0	Least Benedikt (3)	Most Georg (80)	Values Georg (80), Leopold (64), ...[166 more]
id	Integer	0	Min 1	Max 551	Average 249.098
⚠ **Birthday**	Date-time	1	Earliest date Jun 8, 1922 5:55 AM	Latest date Jun 24, 2013 4:55 AM	Duration 33253d 23h 0m 0s
TransactionValue	Real	0	Min 0.000	Max 1051.191	Average 77.944
PaymentMethod	Nominal	0	Least cheque (487)	Most credit card (2679)	Values credit card (2679), cash (1168), ...[1 more]
Date	Date-time	0	Earliest date Mar 12, 2008 5:46 PM	Latest date Feb 27, 2014 7:52 AM	Duration 2177d 14h 5m 34s

Filter (9 / 9 attributes): Search for Attribut

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

Preprocessing in Data Mining:
- Data preprocessing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data.
- The quality of the data should be checked before applying machine learning or data mining algorithms.
- Preprocessing of data is mainly to check the data quality. The quality can be checked by the following

    - **Accuracy**: To check whether the data entered is correct or not.
    - **Completeness**: To check whether the data is available or not recorded.
    - **Consistency:** To check whether the same data is kept in all the places that do or do not match.
    - **Timeliness**: The data should be updated correctly.
    - **Believability**: The data should be trustable.
    - **Interpretability**: The understandability of the data.

- Following are some major tasks in pre-processing:

    - **Data Cleaning –** It is the process to remove incorrect data, incomplete data and inaccurate data from the datasets, and it also replaces the missing values.

    - **Data Integration –** It is the process of combining multiple sources into a single dataset.

    - **Data reduction -** This process helps in the reduction of the volume of the data which makes the analysis easier yet produces the same or almost the same result.

    - **Data transformation -** The change made in the format or the structure of the data is called data transformation.

Handling missing values

There are several ways to handle missing values, some of them are:
- **Ignore the data row.**

    - If there are lots of missing values in a particular row, we just ignore that row.

- **Use a global constant to fill in for missing values.**

    - We just a put a common value to all those missing values.

- **Use attribute mean.**

    - Replacing missing values of an attribute with the mean (or median if its discrete) value for that attribute in the database.

- **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.

- **Use a data mining algorithm to predict the most probable value.**

  ○ The value can be determined using regression, inference based tools using Bayesian formalism, decision trees, clustering algorithms (K-Mean\Median etc.).

We can use **Replace Missing values** operator to perform these operations to handle missing values in the datasets.

- **Data sets used :** Customer Data and Transaction Data
- **Path:** Repository → Training Resources → Data → Hotel App Customer

Steps:

- Drag the Customer Data and Transaction Data into the process.
- Search for sort operator and drag it into the process.
- Connect the out ports of Customer Data and Transaction Data into the input port of Join operator.
- Click on join to select it.
- In the Parameters of join, go to the edit list of key attributes.
- Select the key attributes in both side, here CustomerId.
- Hover your mouse pointer over the output port and wait for the tooltip to show the meta data.
- Press F3 while the tooltip is shown. It is now transformed into a window and you can scroll down to see the information about all the columns.
- Check for columns with missing values.
- Add a new operator **Select Attributes**.
- Connect the new operator to the **Retrieve** operator and the output to the result port "res" on the right.
- In Parameters, change the **attribute filter type** to *Subset* and select all attributes but *Birthday*. This means that this will be removed by the operator.
- Run the process.
- Click on the **Statistics** tab and check which attributes with missing values are still left.
- Here, in this dataset, there's no missing value left, but to understand this, unselect the Birthday and perform the below operations.
- Search for the operator **Replace Missing Values** and add it to the process. You can drop it on the connection line between **Select Attributes** and the result port (move the mouse pointer until the connection is highlighted before you drop the new operator). This way you do not need to manually reconnect the operators.
- In the Parameters of this operator, use *single* for the **attribute filter type** and select *Birthday* for **attribute**.
- Run the process again and interpret results

Name		Type	Missing	Statistics				Filter (9 / 9 attributes):	Search for Attribut
⚠ Birthday		Date time	1	Earliest date Jun 8, 1922 5:55 AM	Latest date Jun 24, 2013 4:55 AM	Duration 33253d 23h 0m 0s			
PostalCode		Integer	0	Min 1	Max 99999	Average 51297.960			
HashCode		Polynominal	0	Least zPlL3DcR (1)	Most cujNb6GQ (20)	Values cujNb6GQ (20), wHh275mF (20), ...[869 more]			
CustomerId		Integer	0	Min 1	Max 1000	Average 489.938			
Vorname		Polynominal	0	Least Benedikt (3)	Most Georg (80)	Values Georg (80), Leopold (64), ...[166 more]			
id		Integer	0	Min 1	Max 551	Average 249.098			
TransactionValue		Real	0	Min 0.000	Max 1051.191	Average 77.944			
PaymentMethod		Polynominal	0	Least cheque (487)	Most credit card (2679)	Values credit card (2679), cash (1168), ...[1 more]			
Date		Date time	0	Earliest date Mar 12, 2008 5:46 PM	Latest date Feb 27, 2014 7:52 AM	Duration 2177d 14h 5m 34s			

- Search for **Filter Examples** and drop it again on the connection line to the result port. In case you miss the connection, you can of course always manually connect the operator instead.
- Note the **link** at the bottom of the **Parameters** panel which shows/hides advanced parameters Click on **Show advanced parameters** to display all operator parameters.
- New parameters should appear. Set the **condition class** to *no_missing_attributes*.
- Run the process again.
- Now there's no missing value left in the dataset.

Name		Type	Missing	Statistics				Filter (9 / 9 attributes):	Search for Attribut
⚠ Birthday		Date time	0	Earliest date Jun 8, 1922 5:55 AM	Latest date Jun 24, 2013 4:55 AM	Duration 33253d 23h 0m 0s			
PostalCode		Integer	0	Min 1	Max 99999	Average 51291.322			
HashCode		Polynominal	0	Least EOGj6tDH (0)	Most cujNb6GQ (20)	Values cujNb6GQ (20), wHh275mF (20), ...[869 more]			
CustomerId		Integer	0	Min 1	Max 1000	Average 490.048			
Vorname		Polynominal	0	Least Benedikt (3)	Most Georg (80)	Values Georg (80), Leopold (64), ...[166 more]			
id		Integer	0	Min 1	Max 551	Average 249.153			
TransactionValue		Real	0	Min 0.000	Max 1051.191	Average 77.879			
PaymentMethod		Polynominal	0	Least cheque (487)	Most credit card (2678)	Values credit card (2678), cash (1168), ...[1 more]			
Date		Date time	0	Earliest date Mar 12, 2008 5:46 PM	Latest date Feb 25, 2014 8:47 AM	Duration 2175d 15h 0m 36s			