BizViz How-to-Guide

Predictive Analysis

Connect Data Preparation Components to a Query Service

Release :	2.0
Date :	March 3, 2016

Table of Contents

1. Document Purpose

The purpose of this document is to guide users on how to connect Data Preparation components to a query service. It is recommended that users follow the step-by-step process given below.

2. Prerequisites

2.1. Software

- Browser that supports HTML5
- Operating System: Windows 7

2.2. Knowledge of BizViz Server

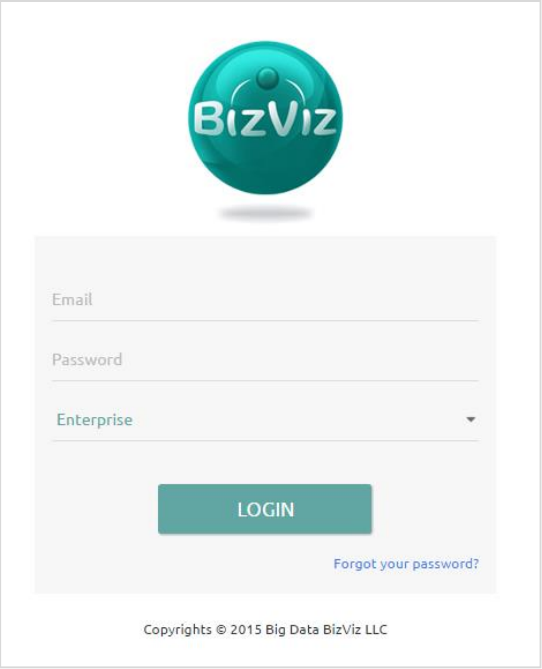
The user should have a basic understanding of the BizViz Server.

3. Step-by-Step Process

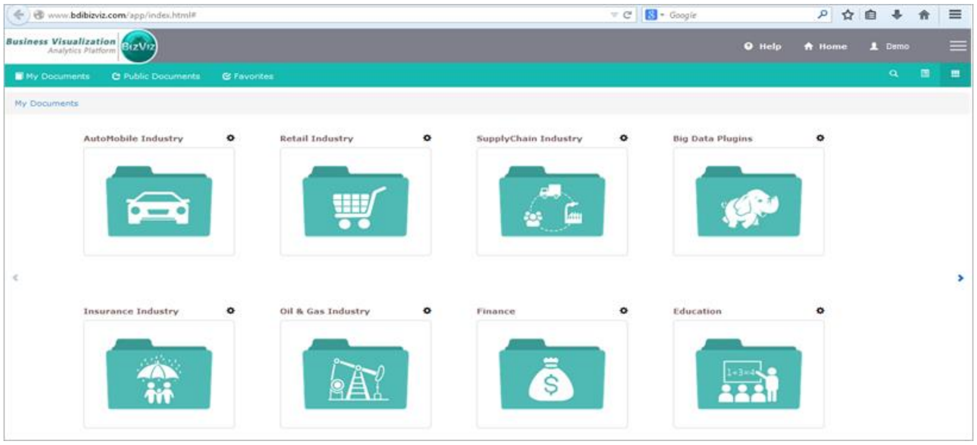
3.1. Login to the BizViz Portal

i) In the URL bar, enter → http://apps.bdbizviz.com/app/index.html

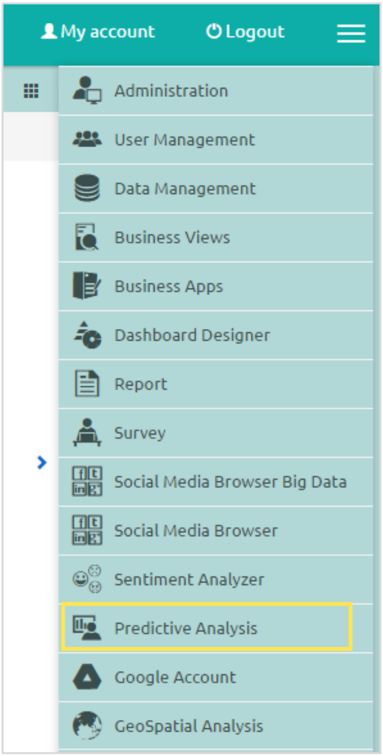
ii) Enter your credentials to Login



iii) Click on '**Login**' to view the BizViz Portal Home Screen

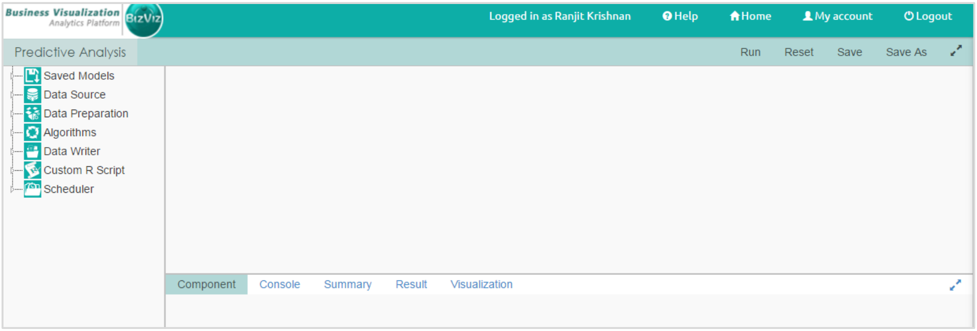


iv) Click on the ’**Menu**’ ≡ button to display a list of installed applications.



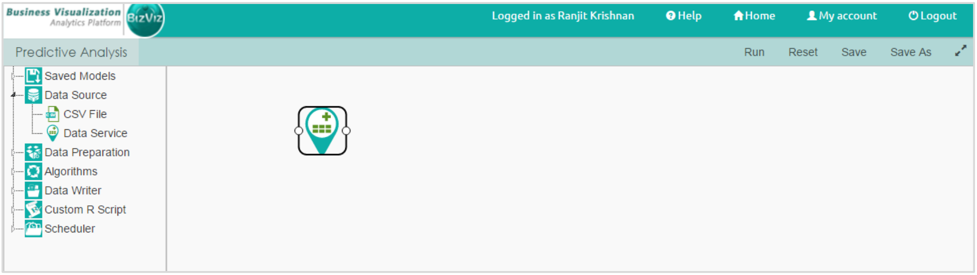
v) Click on ‘**Predictive Analysis**’, as shown above.

vi) Users will be redirected to the predictive analysis home screen.

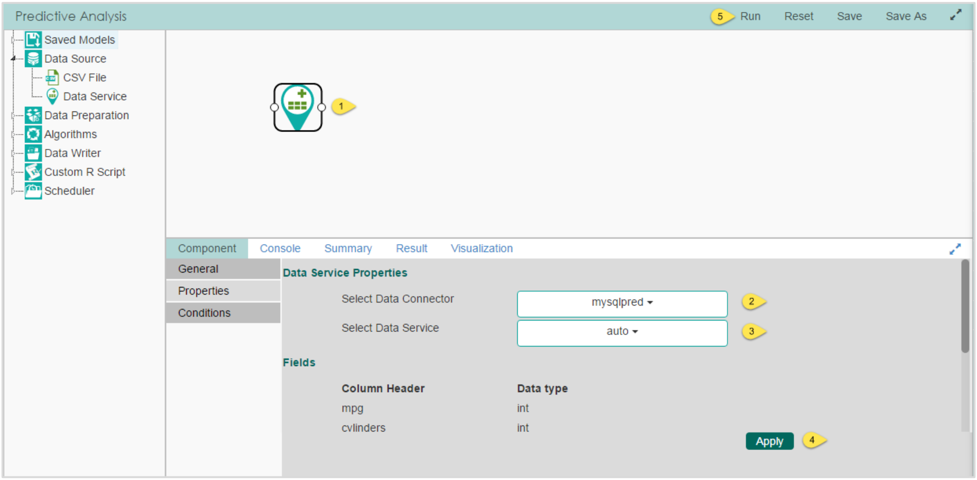


3.2. Steps to Create and Configure a Query Service Connection

i) Navigate to the Predictive Analysis Home Screen and drag and drop the Query Service component onto the workspace.



ii) Click on the '**Query Service**' component and configure the fields as shown below.



Data Service Properties

Select Data Connector: mysqlpred

Select Data Service: auto

Fields

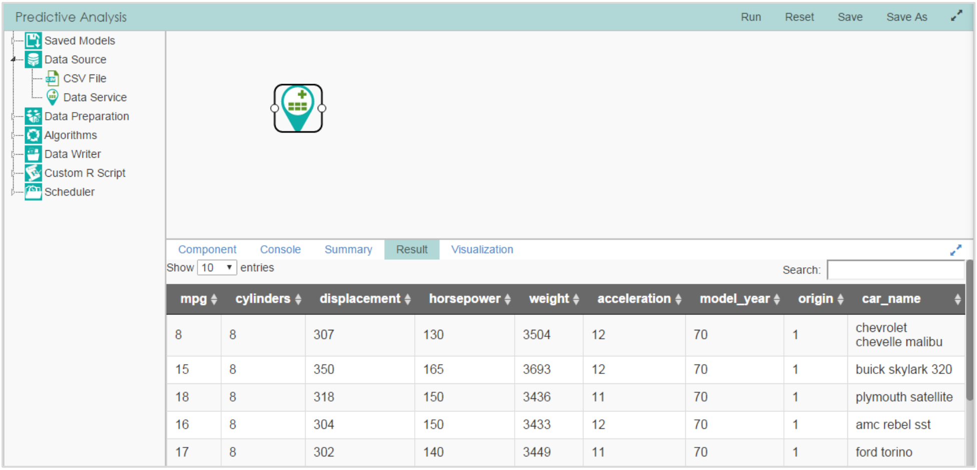
Column Header	Data type
mpg	int
cylinders	int

- **General**
  a. **Component Name**: Default name of the query service
  b. **Alias**: Provide an Alias name if required
  c. **Description**:

- **Properties**
  a. **Select Data Source**: Select a data source from the drop-down list
  b. **Select Query Service**: Select a Query Service from the drop-down list

- **Conditions**
  a. **Filter Name:** The name of the filter (‘**Where Clause**’) that will be applied to the query.
  b. **Control Type:** The type of filter that will be applied (**ex**. Text value, List of Values (LOV))

iii) Click on ‘**Apply**’ and select ‘**Run**’ to view the data retrieved using the selected query service.



mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin	car_name
8	8	307	130	3504	12	70	1	chevrolet chevelle malibu
15	8	350	165	3693	12	70	1	buick skylark 320
18	8	318	150	3436	11	70	1	plymouth satellite
16	8	304	150	3433	12	70	1	amc rebel sst
17	8	302	140	3449	11	70	1	ford torino

3.3. Connecting a Data Preparation to a Query Service

Data preparation components are used to make changes to or limit retrieved data, such as changing the data-type, applying filters, implementing normalization, and implementing sampling techniques.

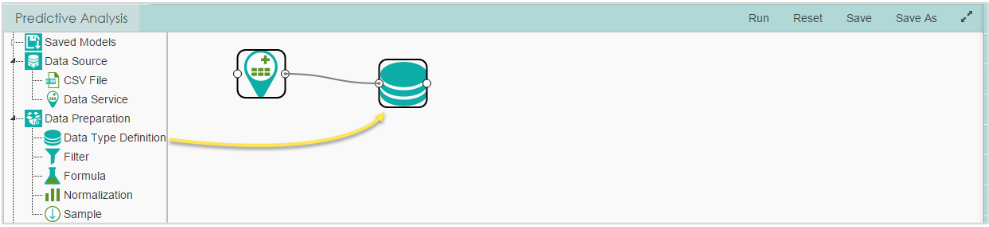
Data Preparation Type	Description
Data Type Definition	User can change the source columns data-type
Filter	• Row Filer: Conditions can be given to limit the retrieved data-set by applying operators, functions, etc.… • Column Filter: Users can select the columns they want to retrieve from the database table
Formula	Users can create a formula using available columns, functions, and operators
Normalization	Normalizing data enables variables with different scales of measurement or different value ranges to be

	converted to a common scale for comparison. This is a necessary data preparation step before applying some predictive algorithms.
Sampling	Data sampling is an analysis technique used to select, manipulate, and analyze a representative subset of data points in order to identify patterns and trends in the larger data-set being examined.

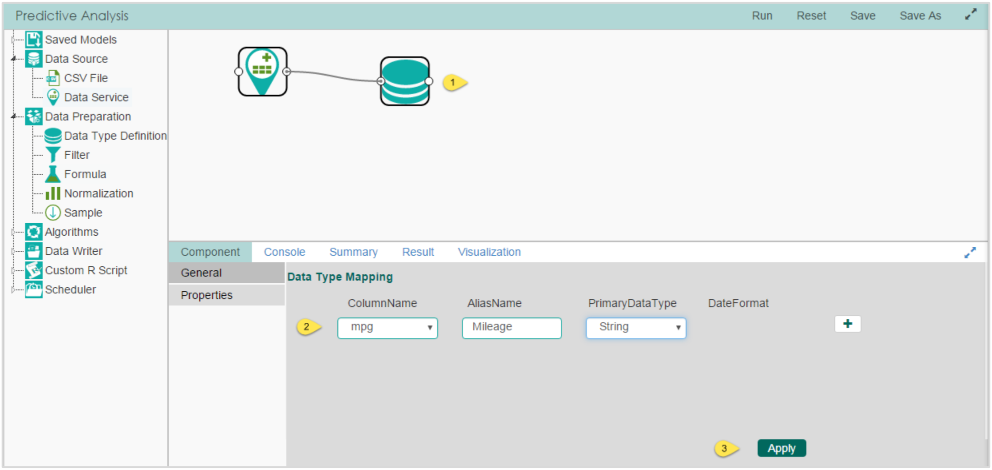
The data preparation components listed above are explained here with examples:

a. Data Type Definition

Drag and drop the ‘**Data Type Definition**’ component onto the workspace and connect it with the Query Service component.



i) Click on the ‘**Data Type Definition**’ component. From the drop-down list, select the column which you wish to change the data type for, make the necessary change, and click on ‘**Apply**’ (follow the steps as shown below):

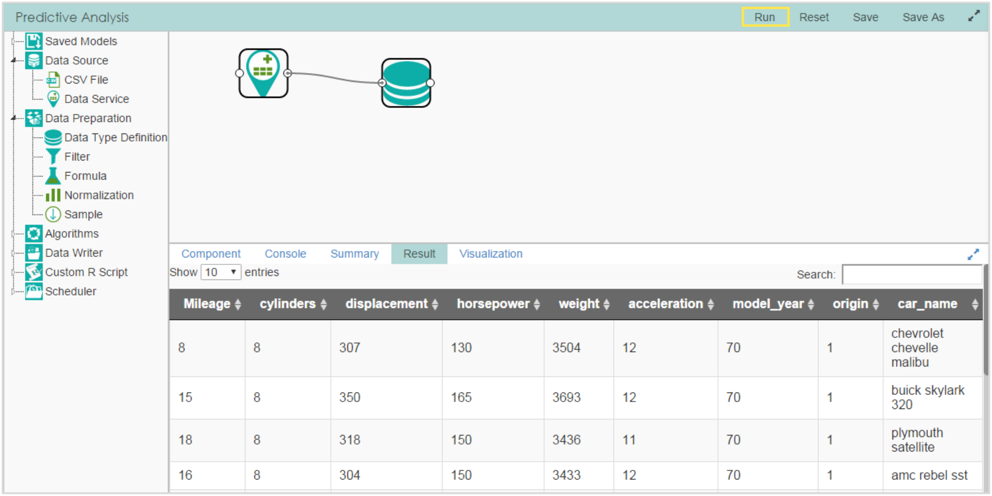


Data Type Mapping

	ColumnName	AliasName	PrimaryDataType	DateFormat
2	mpg	Mileage	String	

Ex: In the above figure, we have selected the column ‘**mpg**’ and given ‘**Mileage**’ as an Alias Name and changed the data type to ‘**String**’.

ii) Click on ‘**Run**’ to view the results.



Mileage	cylinders	displacement	horsepower	weight	acceleration	model_year	origin	car_name
8	8	307	130	3504	12	70	1	chevrolet chevelle malibu
15	8	350	165	3693	12	70	1	buick skylark 320
18	8	318	150	3436	11	70	1	plymouth satellite
16	8	304	150	3433	12	70	1	amc rebel sst

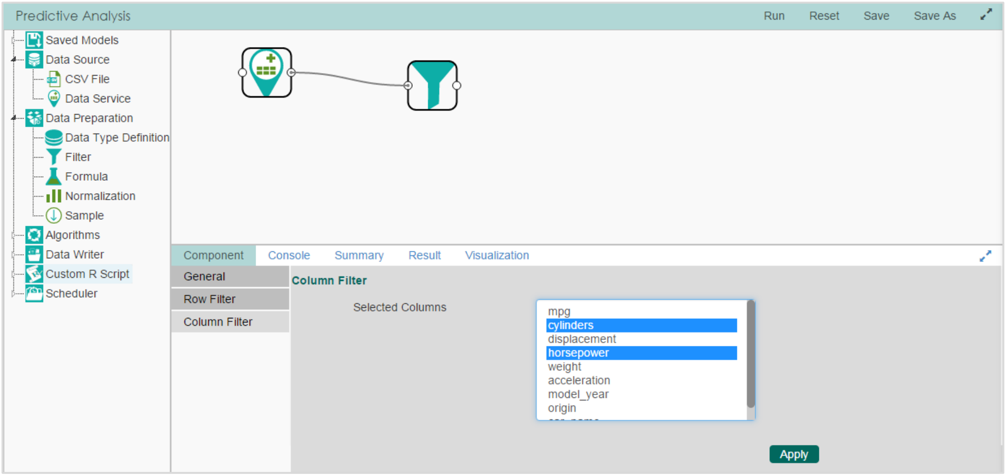
Note:

- Alias Names can be displayed in the result data-set
- Column types can only be viewed at the database and algorithm level.

b. Filter

Row Filter: Functions and operators can be applied to any columns to limit the returned data-set.

Column Filter: You can select the columns you wish to view in the report.

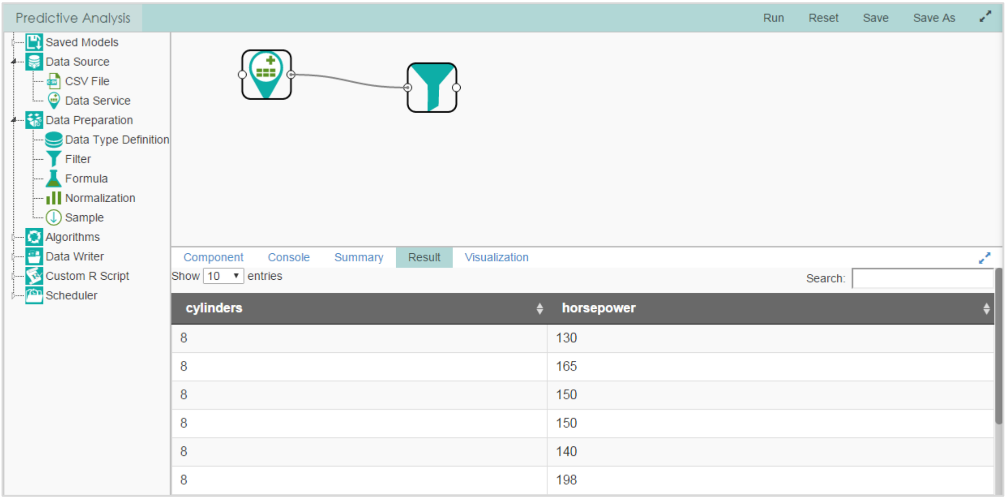


Column Filter

Selected Columns: mpg, **cylinders** (selected), displacement, **horsepower** (selected), weight, acceleration, model_year, origin

Note: In the figure above, we can see that only the ‘**cylinders**’ and ‘**horsepower**’ columns have been selected.

On clicking ‘**Apply**’, these selected columns will be displayed in the target, as shown below.



Show 10 entries

cylinders	horsepower
8	130
8	165
8	150
8	150
8	140
8	198

Row Filter:

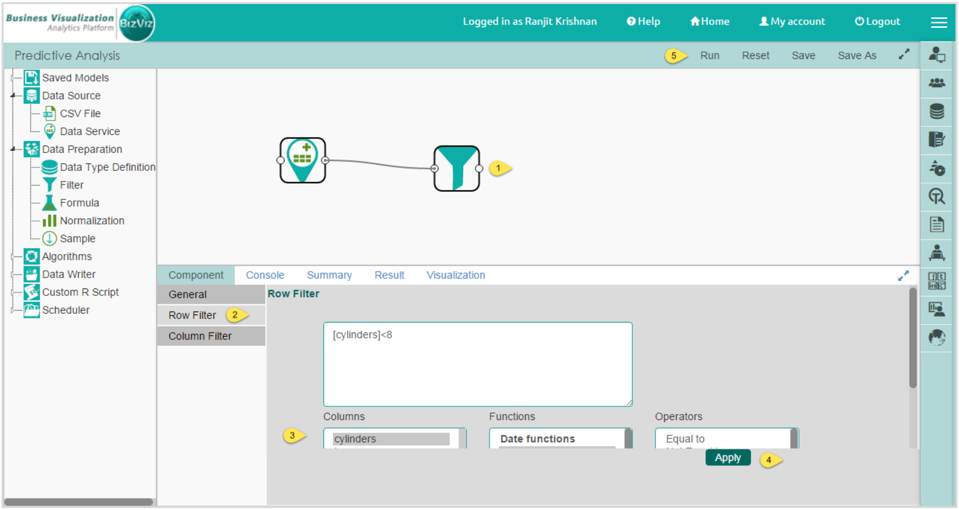
Using ‘**Row Filter**’, data can be filtered by using functions and operators.

i) Click on the ‘**Filter**’ component.

ii) Click on the ‘**Row Filter**’ tab which is available under ‘**Component**’ section.

iii) Build a formula by double clicking on the fields available under ‘**Columns**’, ‘**Functions**’, and ‘**Operators**’.

iv) Click on ‘**Apply**’ and then click on ‘**Run**’ to view the result set.



Row Filter

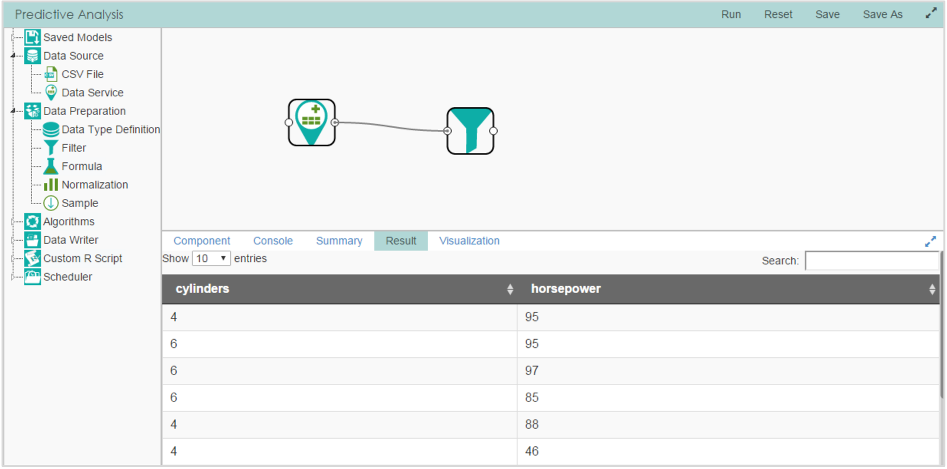
[cylinders]<8

Columns: cylinders

Functions: Date functions

Operators: Equal to

Result:



cylinders	horsepower
4	95
6	95
6	97
6	85
4	88
4	46

Note: Based on the conditions we set in the Row Filter section, only data where the **'cylinder'** value is less than 8 are displayed.

c. Formula

Users can define a formula which will be applied to the source data.

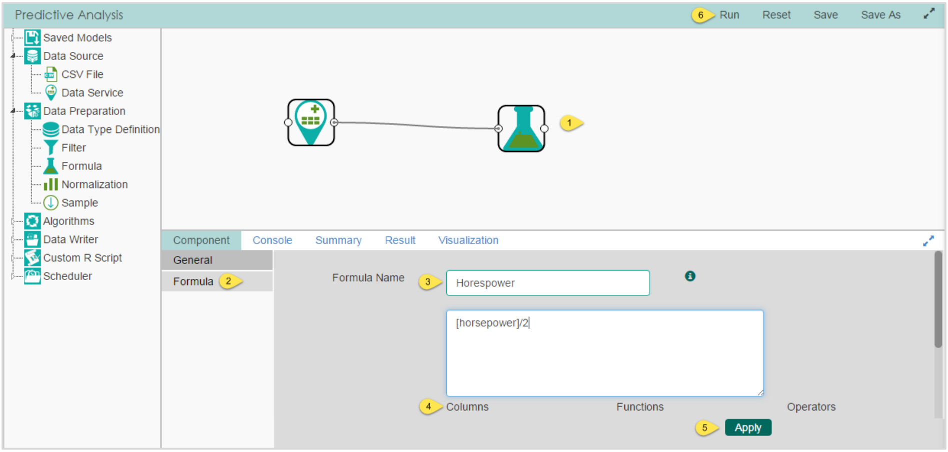
i) Click on the '**Formula**' component.

ii) Click on the '**Formula**' tab which is under '**Component**' section.

iii) Build a formula by double clicking on the fields available under Columns, Functions, and Operators.

iv) Enter a name for the formula in the '**Formula Name**' text field.

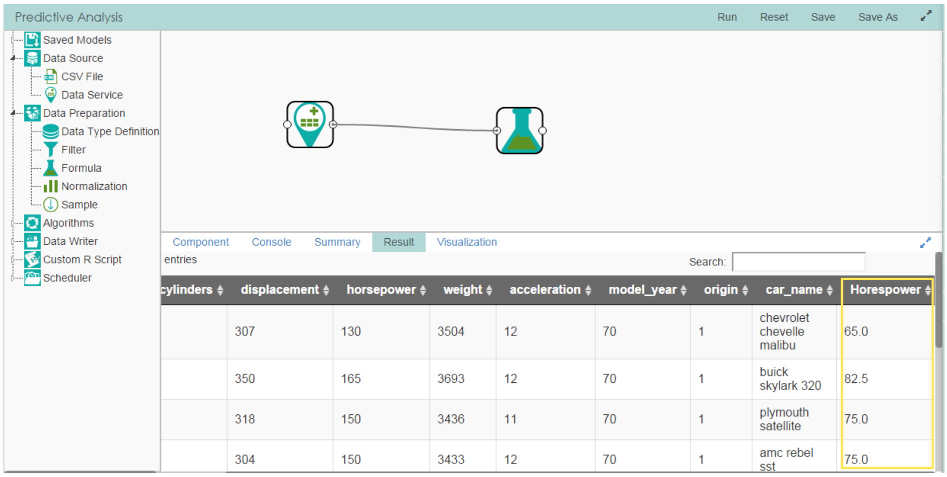
v) Click on '**Apply**' and then click on '**Run**' to view the result set.



Formula Name: Horespower

[horsepower]/2

Result:



cylinders	displacement	horsepower	weight	acceleration	model_year	origin	car_name	Horespower
	307	130	3504	12	70	1	chevrolet chevelle malibu	65.0
	350	165	3693	12	70	1	buick skylark 320	82.5
	318	150	3436	11	70	1	plymouth satellite	75.0
	304	150	3433	12	70	1	amc rebel sst	75.0

Note: The newly created column, named '**Horespower**', will be displayed with data filtered using the given condition, as shown above.

d. Normalization

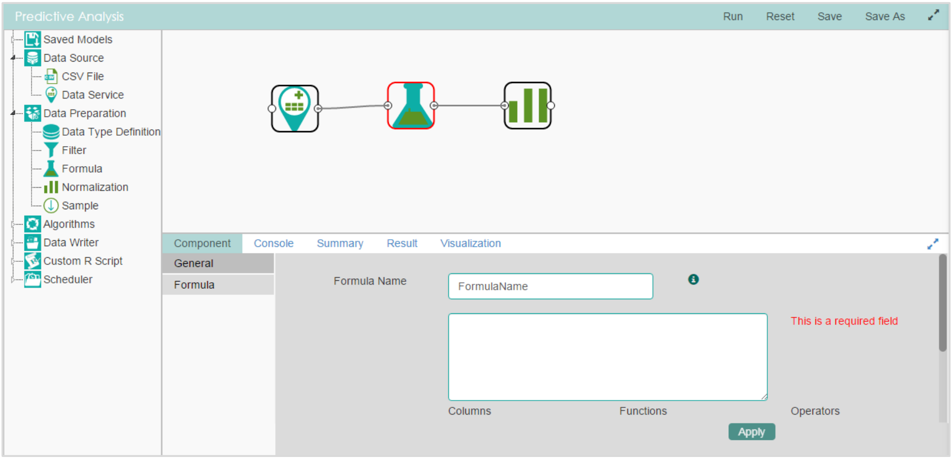
Min-max Normalization: Performs a linear transformation on the original data values. Suppose that minX and maxX are the minimum and maximum of feature X. We would like to map interval [minX,maxX] into a new interval [new_minX, new_maxX]. Consequently, every value v from the original interval will be mapped into value new_v using the following formula.

$$new\_v = \frac{v - min_X}{max_X - min_X} \cdot (new\_max_X - new\_min_X) + new\_min_X$$

Min-max normalization preserves the relationships among the original data values. A problem may occur if a value of an unseen data point (that we would like to predict on) is out of [*minX, maxX*] interval.

Follow the step by step process given below to implement Min-max Normalization:

i) Drag and drop the Query, Formula, and Normalization components onto the workspace.
ii) Click on the Query component and upload the file.
iii) Connect the Query component to the Filter component and configure it as shown below.

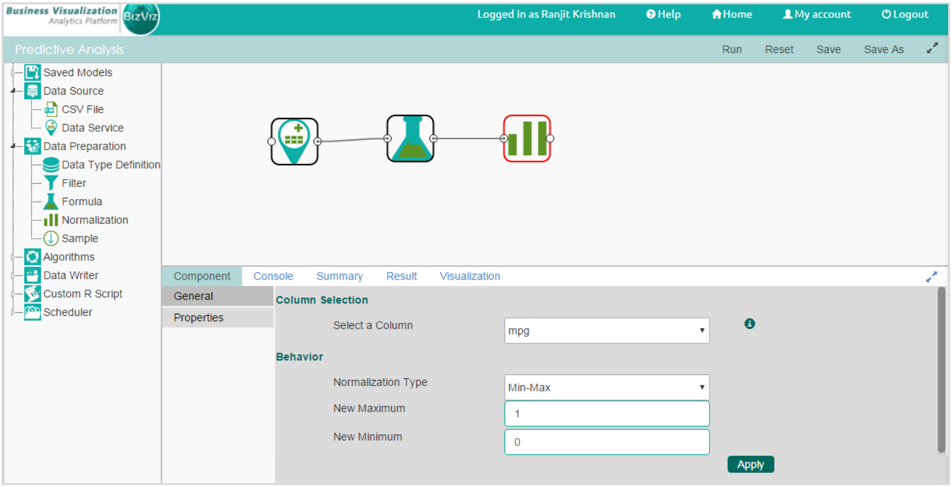


Formula Name: FormulaName

This is a required field

Columns Functions Operators

iv) Build a formula by double clicking on the fields available under ‘**Columns**’, ‘**Functions**’, and ‘**Operators**’.



Column Selection

Select a Column: mpg

Behavior

Normalization Type: Min-Max

New Maximum: 1

New Minimum: 0

v) Enter the following fields:

- **Select a Column:** Select a column on which normalization should be performed
- **Normalization Type:** Select the Normalization type from the dropdown list
- **New Maximum:** Enter a Maximum value
- **New Minimum:** Enter a Minimum value

vi) Click on ‘**Apply**’, and then click on ‘**Run**’ to view the result set.

Zero Score Normalization: also called ‘Zero-mean Normalization’. The values of attribute X are normalized using the mean and standard deviation of X. A new value new_v is obtained using the following expression:

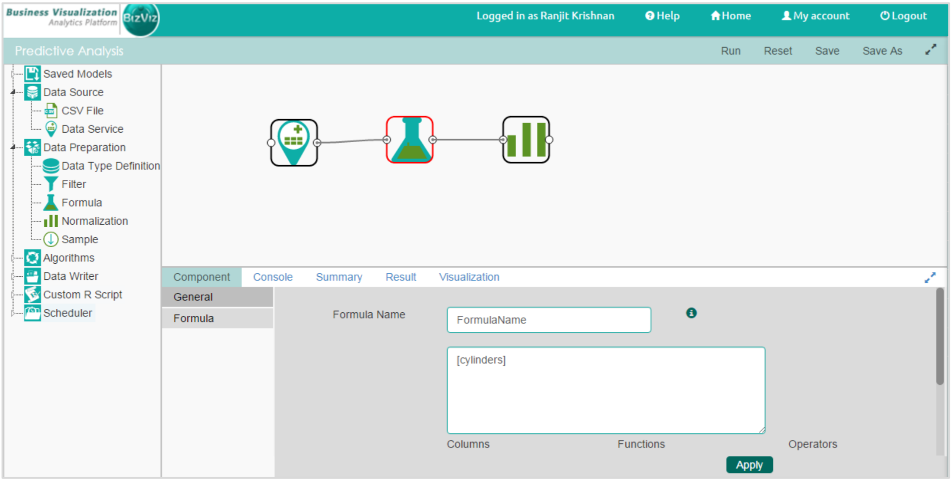
$$new\_v = \frac{v - \mu_X}{\sigma_X}$$

where ⍰X and ⍰X are the mean and standard deviation of attribute X, respectively. If ⍰X and ⍰X are not known, they can be estimated from the sample (column in D that corresponds to feature X). In such a case, we can substitute X ⍰ˆ for ⍰X and X ⍰ˆ for ⍰X in the expression above.

After zero-mean normalizing each feature will have a mean value of 0. Also, the unit of each value will be the number of (estimated) standard deviations away from the (estimated) mean. Note that z-score normalization may be sensitive to small values of ⍰X.

Follow the step by step process given below to implement Zero Score Normalization:

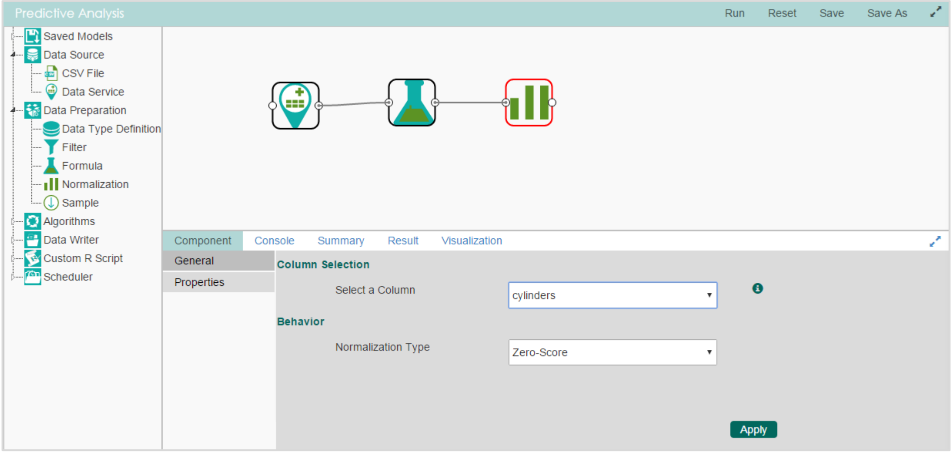
i) Drag and drop the Query, Formula, and Normalization components onto the workspace.
ii) Click on the Query component and configure it
iii) Connect the Query component to the Filter component and configure it as shown below.



Formula Name: FormulaName

[cylinders]

iv) Connect the Formula component with the Normalization component and configure as shown below:



Column Selection

Select a Column: cylinders

Behavior

Normalization Type: Zero-Score

v) Enter the following fields:
- **Select a Column:** Select a field on which normalization should be performed
- **Normalization Type:** Select a Normalization type

vi) Click on '**Apply**' and then click on '**Run**' to view the result set.

Decimal Scaling: Normalizes by moving the decimal point of values of feature X. The number of decimal points moved depends on the maximum absolute value of X. A modified value new v corresponding to v is obtained using

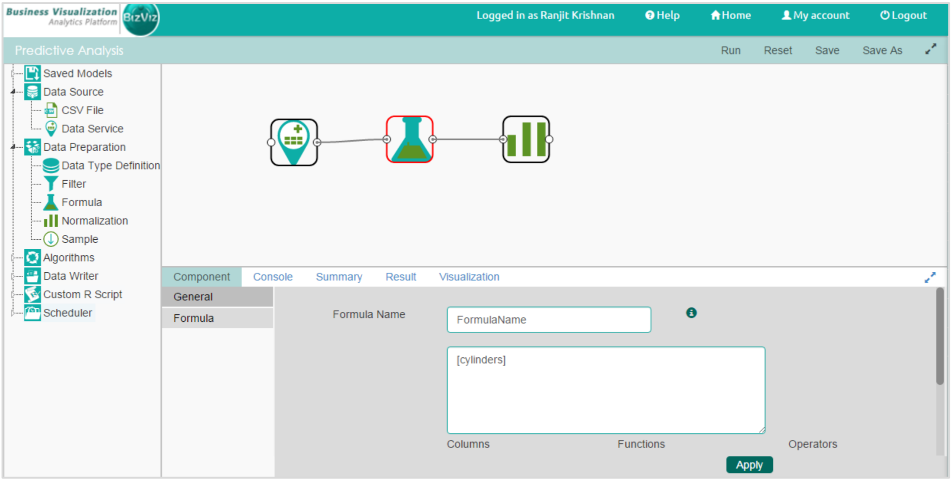
$$new\_v = \frac{v}{10^c}$$

Where, *c* is the smallest integer such that max(|*new v*|) < 1.

Example: suppose the range of attribute X is ⍰500 to 45. The maximum absolute value of X is 500. To normalize by decimal scaling we will divide each value by 1,000 (c = 3). In this case, ⍰500 becomes ⍰0.5 while 45 will become 0.045.

Follow the step by step process given below to implement Zero Score Normalization:

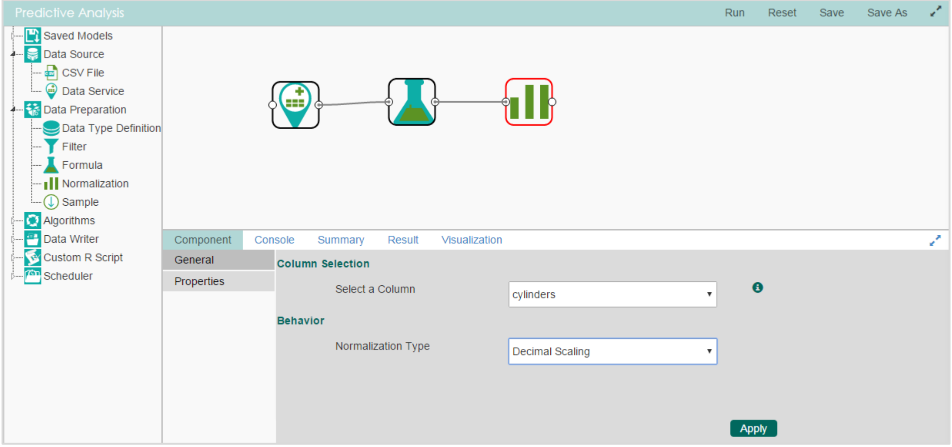
i) Drag and drop the CSV, Formula, and Normalization components onto the workspace.
ii) Click on the CSV component and upload the file.
iii) Connect the CSV component to the Formula component and configure it as shown below:



Formula Name: FormulaName

[cylinders]

iv) Connect the Formula component to the Normalization component and configure as shown below:



Column Selection

Select a Column: cylinders

Behavior

Normalization Type: Decimal Scaling

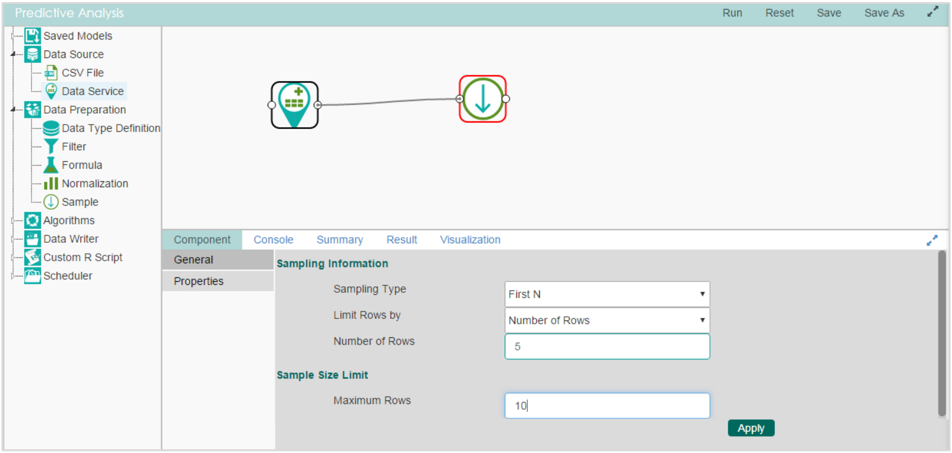
v) Click on '**Apply**' and then click '**Run**' to view the result set.

e. Sampling

Sampling is used to filter rows based on conditions.

Follow the step by step process given below to implement Sampling:

i) Drag and drop the Query and sampling component on to the workspace.
ii) Click on the Query component and configure it.
iii) Connect the Query component to the '**Sampling component**' and configure it as shown below:



Sampling Information

Sampling Type: First N

Limit Rows by: Number of Rows

Number of Rows: 5

Sample Size Limit

Maximum Rows: 10

iv) Enter the following fields:

- **Sampling Type:** Select a sampling type
- **Limit Rows by:** Option to restrict the row count
- **Number of Rows:** The number of rows the user wishes to view
- **Maximum Rows:** The maximum number of rows to be viewed from your data-set