

DATA MINING PROJECT

Submission 1

Topic : Data Preprocessing

Data set : Ecommerce dataset(Opticals)

Group 26:

Boby Aloysius Johnson	(B130698CS)
Akshay Babu	(B130165CS)
M Nimesh Reddy	(B130536CS)
K Manpreeth Sai	(B130400CS)
Balmukund Sinha	(B130168CS)

INTRODUCTION

→ Project Overview:

This data mining project is about what kind of a spectacle frame does a customer actually want to buy. It is the manufacturer's analysis point of view in which the data mining is going to happen.

This project demonstrates the Data Preprocessing carried out on a uncleaned data using Open Refine and the data mining techniques are applied on it using the RapidMiner.

→ Project Deliverables:

Data Cleaning

- .Missing Values
- . Noisy Data

Data Integration

- . Redundancy and Correlation Analysis
- . Data Value Conflict Resolution

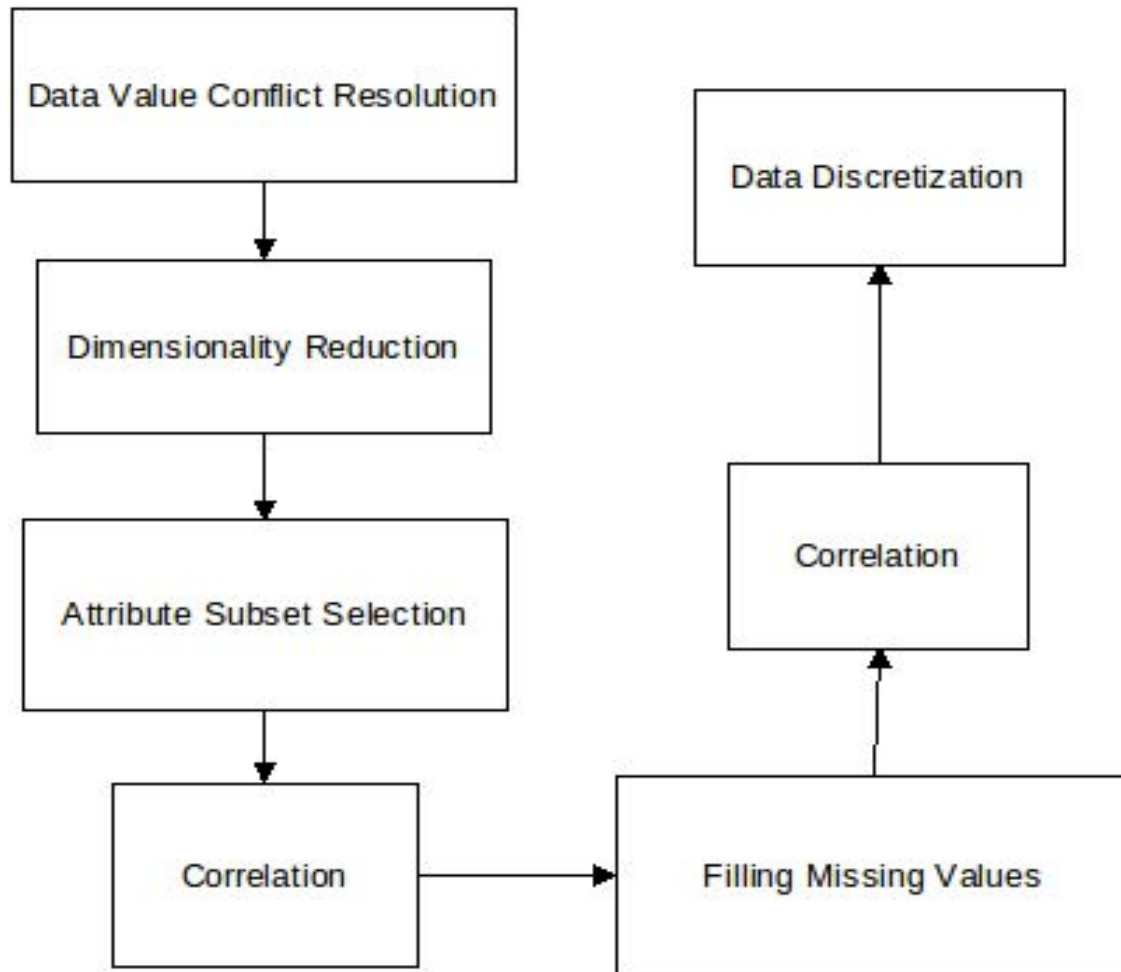
Attribute Subset Selection

- .Stepwise Backward Elimination

Discretization

- . Binning

→ **PROCESS MODEL**



→ ROLE AND RESPONSIBILITY :

.Everyone in our team got involved in selecting the data set.

.We divided the work amongst us equally and at the last we discussed how each of us progressed with our work .

.We have identified the appropriate techniques to be applied and chose the corresponding operators using RapidMiner and OpenRefine.

→ DATA CLEANING TOOLS

- 1) OpenRefine
- 2) Rapidminer

→ PROJECT MANAGEMENT PLAN

PREPROCESSING

a)Tasks planned:

Attributes reduction and cleaning the dirty data set given

b)Description of the plan

1.Attribute number being huge, we have planned to reduce the redundant and correlated attributes

2.Data given being dirty was planned to be cleansed by filling blanks and replacing missing values by appropriate measures and operations.

→ **DATA SET DESCRIPTION**

The data set is about spectacle dataset of an e-commerce website. The dimensionality of the data set is 19 and the number of entires is 13299.

ATTRIBUTE	TYPE	DESCRIPTION
ProductID	NOMINAL	ID for each product
Title	NOMINAL	Name of the product
Description	NOMINAL	Description about the product
ImageURL	NOMINAL	URL for the image of the product
MRP	NUMERIC	MRP of the product
Price	NUMERIC	Price of the product
ProductURL	NOMINAL	URL of the product
Categories	NOMINAL	Category of the product
ProductBrand	NOMINAL	Brand name of the product
DeliveryTime	NUMERIC	Time taken to deliver the product
InStock	BINARY	Information regarding whether the product is in stock
CODAvailable	BINARY	Information regarding whether the product has cash-on-delivery
Discount	NUMERIC	Discount available for the product
CashBack	NUMERIC	Cashback for the product
Size	NOMINAL	Size of the product
Color	NOMINAL	Color of the product
SizeVariant	NOMINAL	
StyleCode	NOMINAL	Style code for the product

→ TASKS:

1. Data Value Conflict Resolution
2. Dimensionality Reduction
3. Attribute Subset Selection
4. Correlation
5. Filling Missing Values
6. Discretization
7. Dependency Reduction

1.Data Value Conflict Resolution:

The data set contains lot of data value conflicts for many attributes. This conflict resolution was done using OpenRefine operator “Text Facet” for the attributes Color, Category, CodAvailable, CashBack. RapidMiner was not feasible to make this conflict resolutions , so OpenRefine was preferred to carry this step. Now the data set’s inconsistency is reduced.

Google refine 1 csv Permalink

Open... Export Help

Facet / Filter Undo / Redo

13298 rows

Show as: rows records Show: 5 10 25 50 rows

Extensions: Freebase

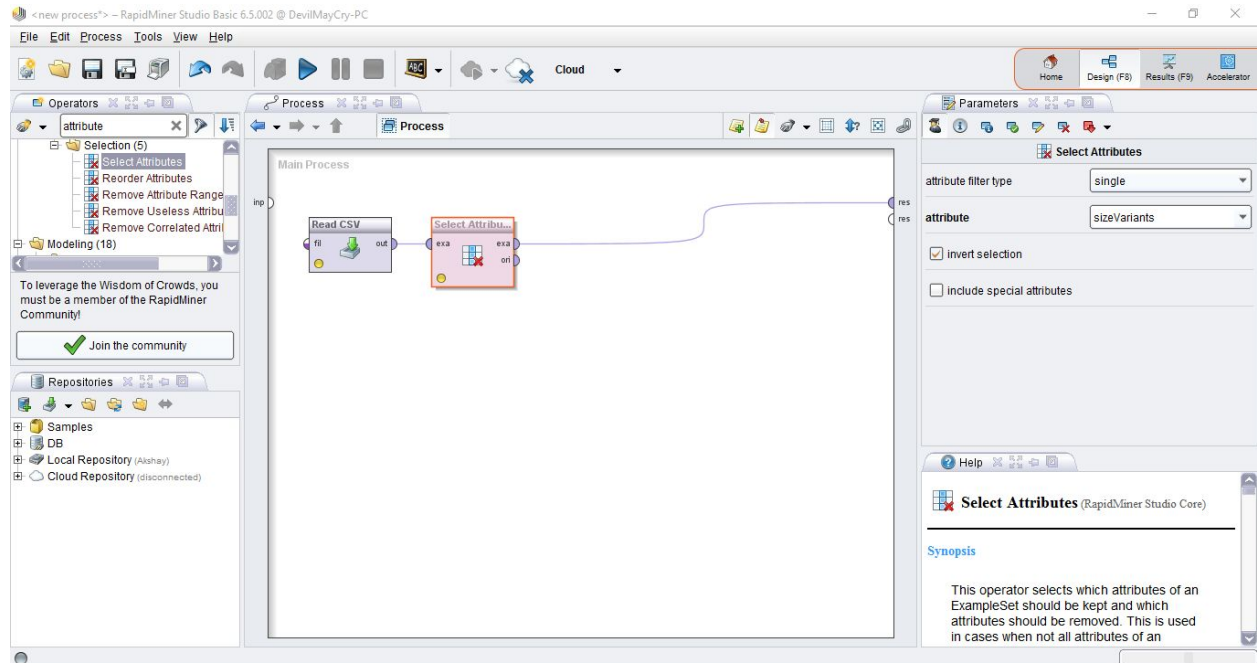
Refresh Reset All Remove All

categories 16 choices Sort by: name count Cluster

productid	title	description	imageUrl	mrp	price	productUrl	categories	productBrand
1	Ryan Half Rim Rectangle Frame		http://img5a.flixcart.com/image/frame/sh/q/r/45-ryan-47-40b400- imgae26ajwzvegrvf.jpeg,http://img5a.flixcart.com/image/frame/s/h/q/r/45-ryan-47-original- imgae26ajwzvegrvf.jpeg,http://img6a.flixcart.com/image/frame/s/h/q/r/45-ryan-47-75x75- imgae26ajwzvegrvf.jpeg,http://img5a.flixcart.com/image/frame/s/h/q/r/45-ryan-47-275x275- imgae26ajwzvegrvf.jpeg,http://img5a.flixcart.com/image/frame/s/h/q/r/45-ryan-47-125x125- imgae26ajwzvegrvf.jpeg,http://img6a.flixcart.com/image/frame/s/h/q/r/45-ryan-47-40x40- imgae26ajwzvegrvf.jpeg,http://img6a.flixcart.com/image/frame/s/h/q/r/45-ryan-47-1100x1100- imgae26ajwzvegrvf.jpeg,http://img6a.flixcart.com/image/frame/s/h/q/r/45-ryan-47-100x100- imgae26ajwzvegrvf.jpeg,http://img5a.flixcart.com/image/frame/s/h/q/r/45-ryan-47-200x200- imgae26ajwzvegrvf.jpeg	1500	599	http://dl.flipkart.com/dl/ryan-half-rim-rectangle-frame/p /tme292icemjdn3x?pid=FRAE28YHUK4GPSHQ	mens	Ryan
2	Ryan Half Rim		http://img5a.flixcart.com/image/frame/d/h/a/r/26-ryan-	1400	599	http://dl.flipkart.com/dl/ryan-half-rim-rectangle-frame/p	gents	Ryan

2.Dimensionality Reduction:

The data set also had another attribute “Size Variant” which had many inconsistent garbage values. So it was also removed using “Invert Select Attribute” operator in RapidMiner.



3.Attribute Subset Selection:

Selecting the required attributes and remove any redundant attributes liked derived attributes.Stepwise Backward Elimination is used for removing “Description” since it is not relevant for our data mining process since it had many null values which are counted to be more than 5500 out of 13297 data entries . This is removed use “Inverting the Select Attribute” operator in Rapidminer tool.

Name	Type	Miss.	Statistics
label rating	Polynomial	13195	Least: poor (4), Most: good (49), Values: good (49), excellent (35), ...[2 more]
productId	Integer	0	Min: 1, Max: 13298, Average: 6648.726, Deviation: 3838.716
title	Polynomial	0	Least: nu look [...] Frame (1), Most: Red Knot [...] ame (774), Values: Red Knot [...] gle Frame (774), Vincent [...] gle Frame (62)
description	Polynomial	5855	Least: ÅÿæšÄÄ [...] \$ÄÄ (1), Most: Whether [...] st. (401), Values: Whether [...] the rest. (401), Frames f [...] ily wear. (209), ...
imageUrl	Polynomial	6	Least: http://i [...] .jpeg; (1), Most: http://i [...] .jpeg; (5), Values: http://i [...] q76.jpeg; (5), http://i [...] n5y.jpeg; (4), ...[13010]
mrp	Integer	0	Min: 0, Max: 32400, Average: 2159.173, Deviation: 2330.941
price	Integer	0	Min: 0, Max: 13900, Average: 1107.601, Deviation: 1318.890
productUrl	Polynomial	0	Least: http://d [...] FR9ZG (1), Most: http://d [...] YT8MG (1), Values: http://d [...] F5RGYT8MG (1), http://d [...] FZ3MHWDQD (1), ...
categories	Polynomial	400	Least: KIDS (82), Most: GENTS (7572), Values: GENTS (7572), LADIES (5242), ...[1 more]
productBrand	Polynomial	0	Least: nu look (1), Most: Red Knot (1548), Values: Red Knot (1548), Vincent Chase (1448), ...[225 more]

Showing attributes: 1 - 19 Examples: 13,296 Special Attributes: 1 Regular Attributes: 18

4.Correlation Analysis:

From the Correlation matrix which was obtained by “Correlation Matrix” operator in RapidMiner Tool , the attributes which had correlation coefficient greater than 0.95 or less than -0.95 have been reduced to a single attribute. In our analysis , “productUrl” and “productId” had a correlation of 100% , so productUrl attribute has been removed. The attributes “productId” and “ImageUrl” are also correlated , for which “ImageUrl” has been removed. The attributes “productId” and “StyleCode” are also correlated , for which “StyleCode” has been removed.

5.Cleaning and Filling:

Data tend to be incomplete, noisy and inconsistent. We attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

Missing values for the following attributes are found:

1. CodAvailable – 810 missing values

The missing values here are replaced by **Mode** since it is a Binary Attribute. The operator “Replace Missing Values” is used in Rapidminer tool.

2. Category -- 400 missing values

The categories available are Gents , Ladies and Kids. It is Nominal Attribute hence we are using **Mode** to replace using operator “Replace Missing Values” is used in Rapidminer tool.

3. Size --201 missing values

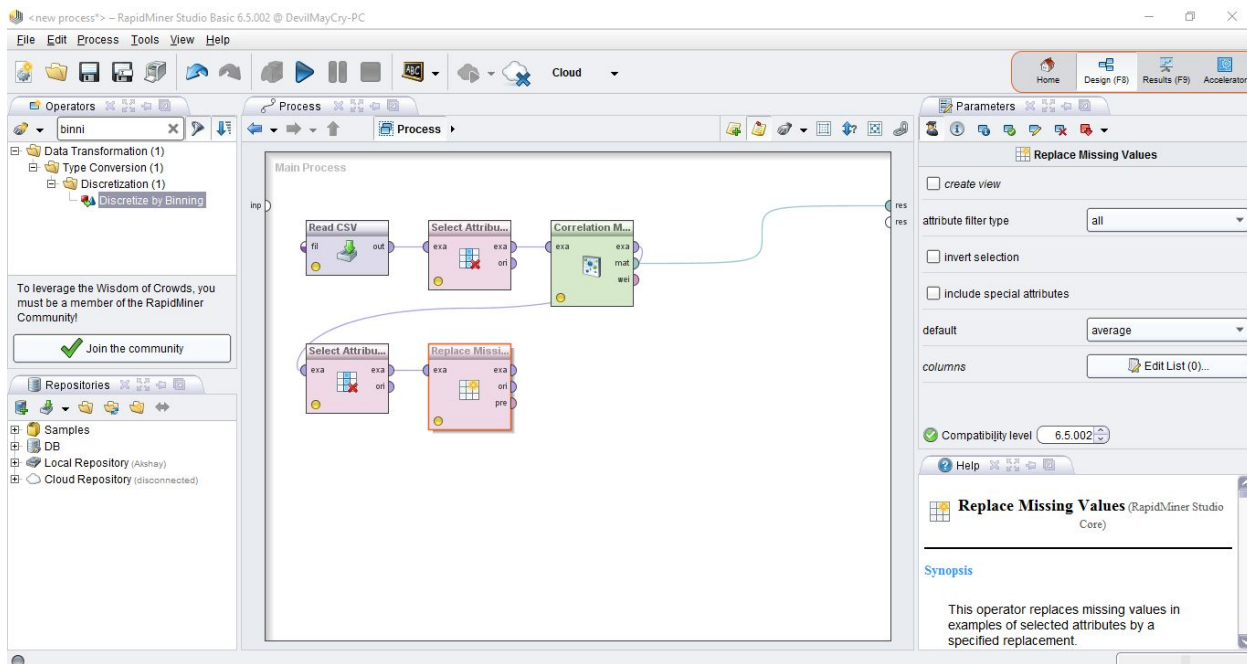
The categories available are S(small), M (medium) , L (large). It is Nominal Attribute hence we are using **Mode** to replace using operator “Replace Missing Values” is used in Rapidminer tool.

4. InStock ---635 missing values

The categories available are True or False. It is Binary Attribute hence we are using **Median** to replace using operator “Replace Missing Values” is used in Rapidminer tool.

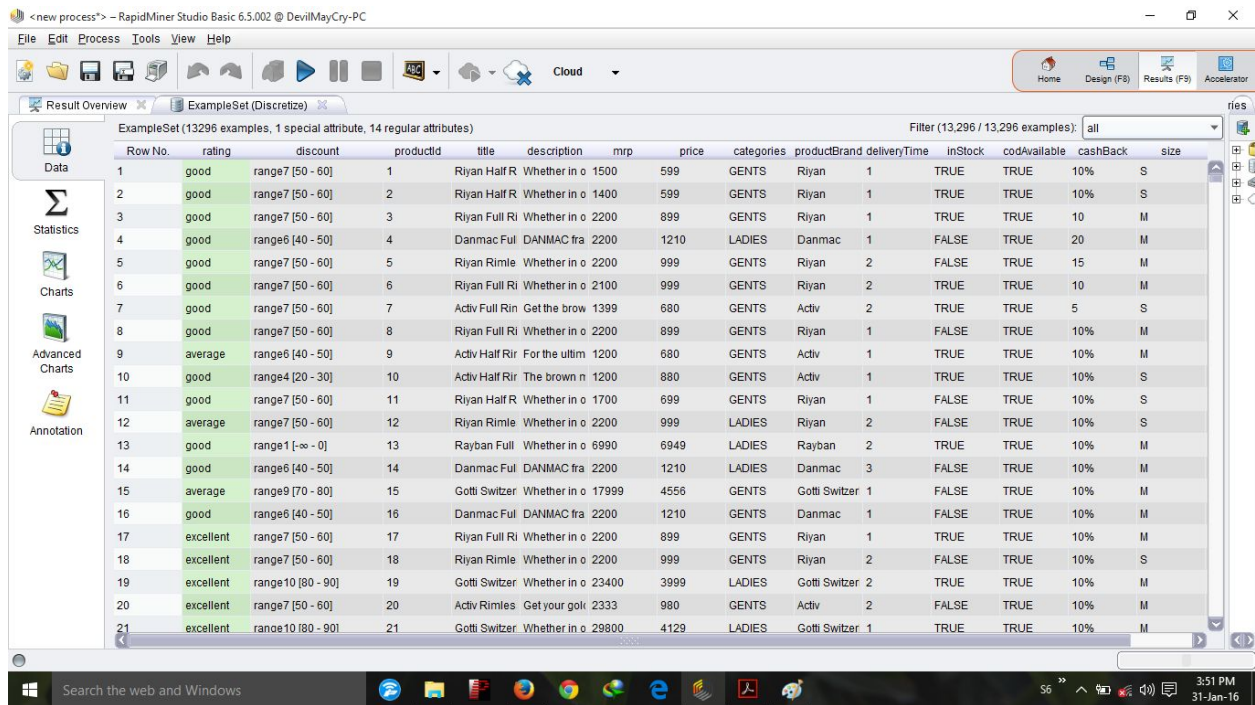
5. Color --12 missing values

The categories available are Black , Grey and many. It is Nominal Attribute hence we are using **Mode** to replace using operator “Replace Missing Values” is used in Rapidminer tool.



6. Discretization:

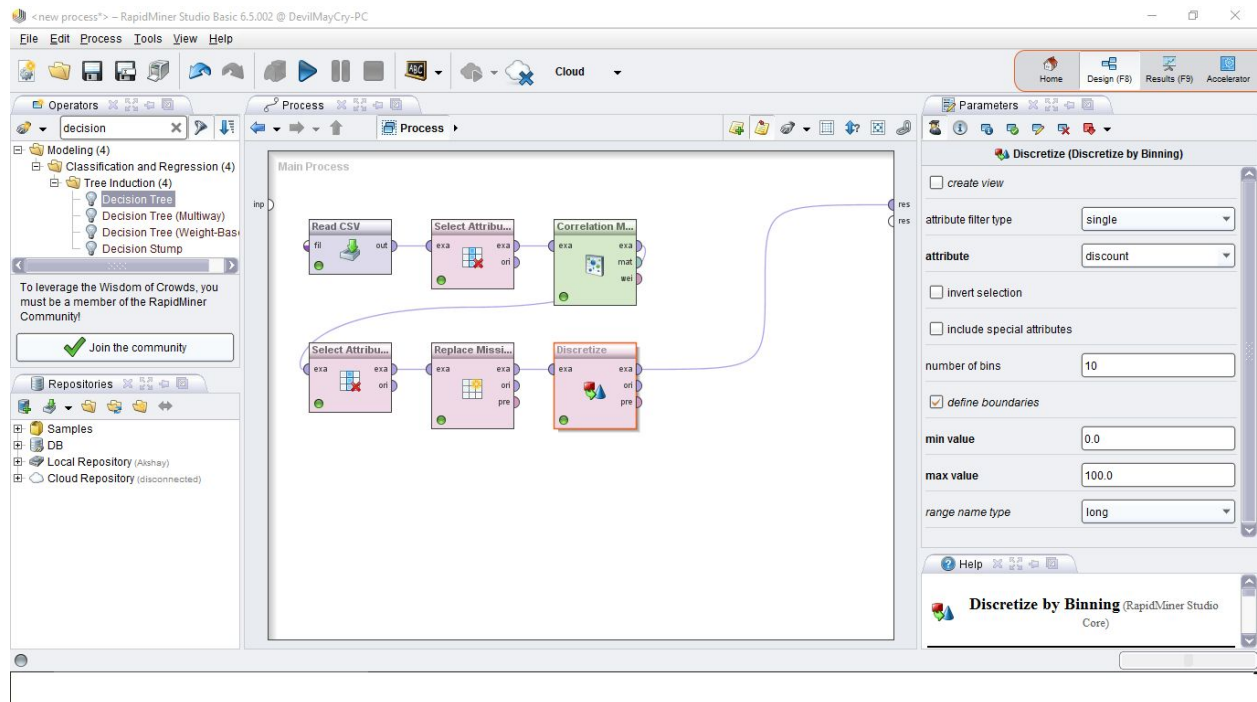
As missing values were filled, the correlation analysis was performed so to check if any other correlated attributes were present. Binning was performed as a part of discretization for the attribute discount as its values were scattered and to make it uniform, the attribute values were divided into 10 bins. Each bin is replaced by that bin's median.



ExampleSet (13296 examples, 1 special attribute, 14 regular attributes)

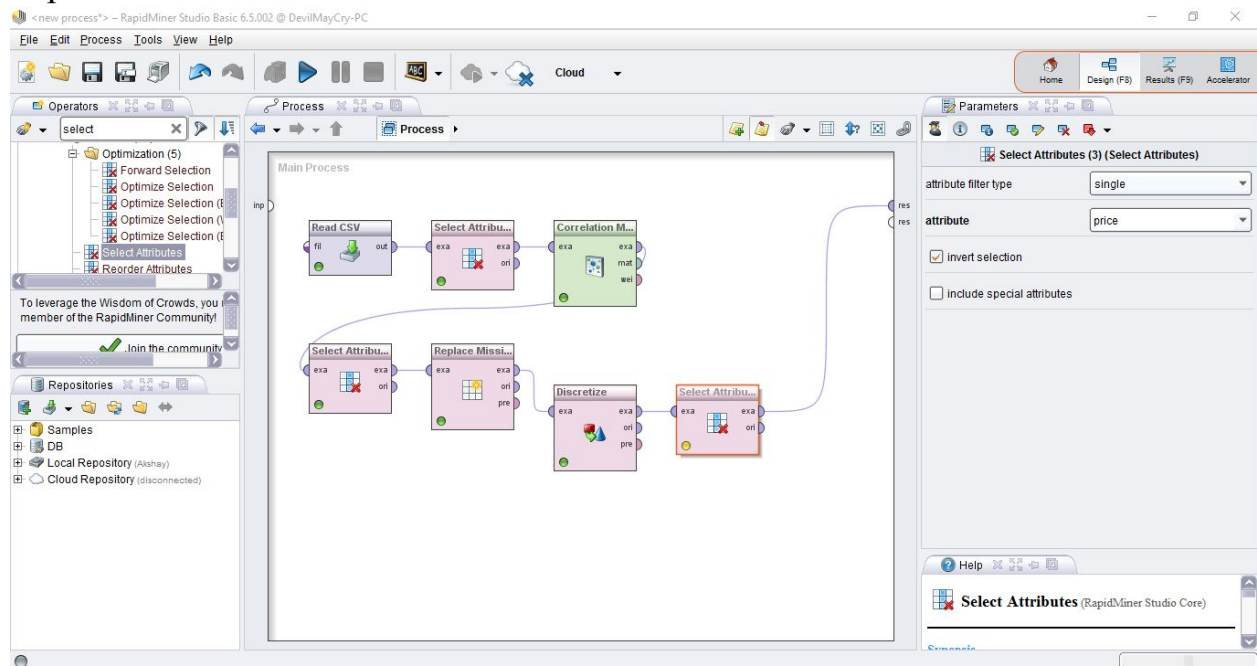
Filter (13,296 / 13,296 examples): all

Row No.	rating	discount	productId	title	description	mrp	price	categories	productBrand	deliveryTime	InStock	codAvailable	cashBack	size
1	good	range7 [50 - 60]	1	Riyan Half R	Whether in o	1500	599	GENTS	Riyan	1	TRUE	TRUE	10%	S
2	good	range7 [50 - 60]	2	Riyan Half R	Whether in o	1400	599	GENTS	Riyan	1	TRUE	TRUE	10%	S
3	good	range7 [50 - 60]	3	Riyan Full Ri	Whether in o	2200	899	GENTS	Riyan	1	TRUE	TRUE	10%	M
4	good	range6 [40 - 50]	4	Danmac Ful	DANMAC fra	2200	1210	LADIES	Danmac	1	FALSE	TRUE	20%	M
5	good	range7 [50 - 60]	5	Riyan Rimle	Whether in o	2200	999	GENTS	Riyan	2	FALSE	TRUE	15%	M
6	good	range7 [50 - 60]	6	Riyan Full Ri	Whether in o	2100	999	GENTS	Riyan	2	TRUE	TRUE	10%	M
7	good	range7 [50 - 60]	7	Activ Full Rin	Get the brow	1399	680	GENTS	Activ	2	TRUE	TRUE	5%	S
8	good	range7 [50 - 60]	8	Riyan Full Ri	Whether in o	2200	899	GENTS	Riyan	1	FALSE	TRUE	10%	M
9	average	range6 [40 - 50]	9	Activ Half Rir	For the ultim	1200	680	GENTS	Activ	1	TRUE	TRUE	10%	M
10	good	range4 [20 - 30]	10	Activ Half Rir	The brown n	1200	880	GENTS	Activ	1	TRUE	TRUE	10%	S
11	good	range7 [50 - 60]	11	Riyan Half R	Whether in o	1700	699	GENTS	Riyan	1	FALSE	TRUE	10%	S
12	average	range7 [50 - 60]	12	Riyan Rimle	Whether in o	2200	999	LADIES	Riyan	2	FALSE	TRUE	10%	S
13	good	range1 [-∞ - 0]	13	Rayban Full	Whether in o	6990	6949	LADIES	Rayban	2	TRUE	TRUE	10%	M
14	good	range6 [40 - 50]	14	Danmac Ful	DANMAC fra	2200	1210	LADIES	Danmac	3	FALSE	TRUE	10%	M
15	average	range9 [70 - 80]	15	Gotti Switzer	Whether in o	17999	4556	GENTS	Gotti Switzer	1	FALSE	TRUE	10%	M
16	good	range6 [40 - 50]	16	Danmac Ful	DANMAC fra	2200	1210	GENTS	Danmac	1	FALSE	TRUE	10%	M
17	excellent	range7 [50 - 60]	17	Riyan Full Ri	Whether in o	2200	899	GENTS	Riyan	1	TRUE	TRUE	10%	M
18	excellent	range7 [50 - 60]	18	Riyan Rimle	Whether in o	2200	999	GENTS	Riyan	2	FALSE	TRUE	10%	S
19	excellent	range10 [80 - 90]	19	Gotti Switzer	Whether in o	23400	3999	LADIES	Gotti Switzer	2	TRUE	TRUE	10%	M
20	excellent	range7 [50 - 60]	20	Activ Rimles	Get your golr	2333	980	GENTS	Activ	2	FALSE	TRUE	10%	M
21	excellent	range10 [80 - 90]	21	Gotti Switzer	Whether in o	29800	4129	LADIES	Gotti Switzer	1	TRUE	TRUE	10%	M



7. Dependency Reduction:

In the given data set, the attributes “MRP”, “Price” and “Discount” are interrelated. Since MRP can be derived from discount attribute, the “price” attribute is redundant so it has been removed using “invert select attribute” operator in RapidMiner.

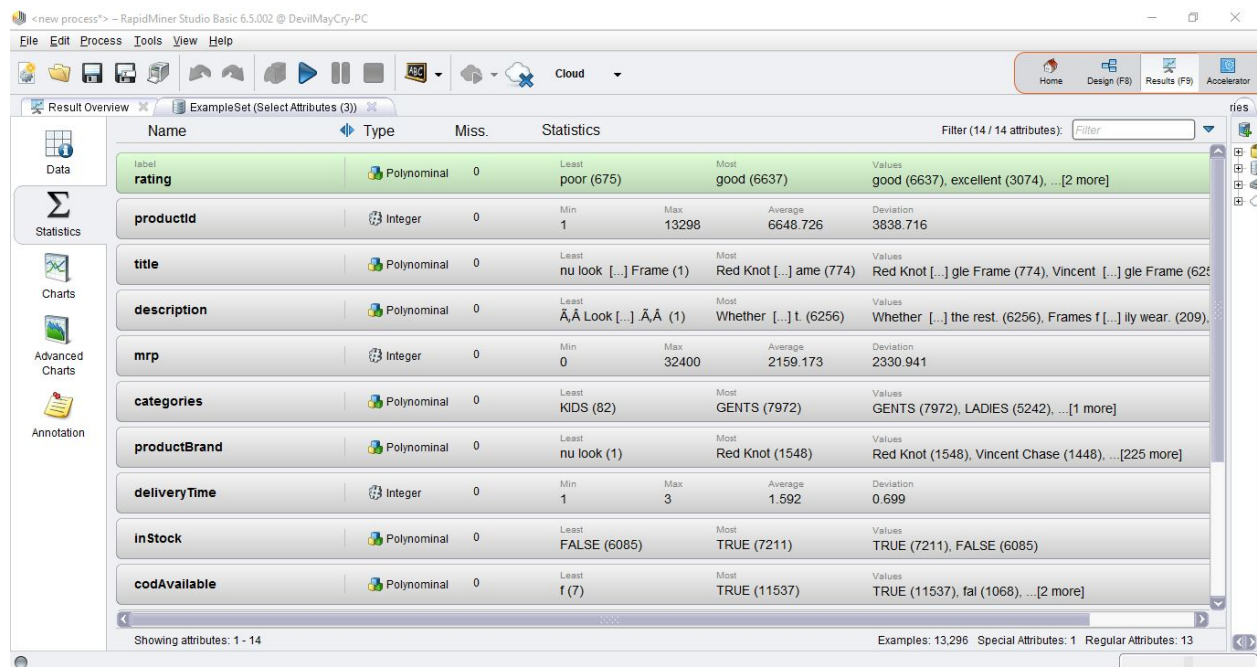


→ Limitations of the tool:

RapidMiner tool was not feasible to clear data value conflicts, so OpenRefine was used as it is very easy to clear these conflicts.

→ Problem Statement and Result:

Conduct data preprocessing for given uncleaned e-commerce data set.
>>The resultant data set was found clean and consistent after performing the above tasks.



Name	Type	Miss.	Statistics
label	Polynomial	0	Least poor (675) Most good (6637) Values good (6637), excellent (3074), ...[2 more]
rating	Polynomial	0	Least poor (675) Most good (6637) Values good (6637), excellent (3074), ...[2 more]
productid	Integer	0	Min 1 Max 13298 Average 6648.726 Deviation 3838.716
title	Polynomial	0	Least nu look [...] Frame (1) Most Red Knot [...] ame (774) Values Red Knot [...] gle Frame (774), Vincent [...] gle Frame (625)
description	Polynomial	0	Least ÃÃ Look [...] ÃÃ (1) Most Whether [...] t. (6256) Values Whether [...] the rest. (6256), Frames f [...] ily wear. (209),
mrp	Integer	0	Min 0 Max 32400 Average 2159.173 Deviation 2330.941
categories	Polynomial	0	Least KIDS (82) Most GENTS (7972) Values GENTS (7972), LADIES (5242), ...[1 more]
productBrand	Polynomial	0	Least nu look (1) Most Red Knot (1548) Values Red Knot (1548), Vincent Chase (1448), ...[225 more]
deliveryTime	Integer	0	Min 1 Max 3 Average 1.592 Deviation 0.699
inStock	Polynomial	0	Least FALSE (6085) Most TRUE (7211) Values TRUE (7211), FALSE (6085)
codAvailable	Polynomial	0	Least f (7) Most TRUE (11537) Values TRUE (11537), fal (1068), ...[2 more]

Showing attributes: 1 - 14 Examples: 13,296 Special Attributes: 1 Regular Attributes: 13

REQUIREMENT SPECIFICATION

Hardware Requirements

RAPID MINER formerly known as YALE (Yet Another Learning Environment), is a software platform that provides an integrated environment for Machine learning, Data Mining, Predictive analytics and Business analytics. It runs on PC if it satisfies the minimum system requirements like Dual core with 2GHz processor, 4GB Ram and free disk space of >1GB provided operating systems of windows 7, 8, 8.1, Linux or Mac OS X 10.8-10.10 with 64-bit of java7 or java8 platform. Rapid Miner is written in the Java programming language. Rapid Miner provides a GUI to design and execute analytical work flows.

Open Refine is a standalone open source desktop application for data cleanup and transformation to other formats, the activity known as Data Wrangling. It is similar to spreadsheet applications (and can work with spreadsheet file formats); however, it behaves more like a database. It requires a reasonably modern PC running Linux, Windows (XP and later), Mac OSX. Sophisticated graphics hardware is not needed.

Specific Requirements

User Interface

- 1) Google Chrome browser.
- 2) Rapid Miner

Software Interface

- 1) Google Chrome browser.
- 2) Rapid Miner

Other Details

Java platform is required for running RapidMiner.

Software

Introduction :

OPEN REFINE, formerly Google Refine is a powerful tool for working with messy data, cleaning it, transforming it from one format to another, extending it with web services and linking it to databases like Freebase.

RAPID MINER (formerly known as YALE) written in the Java Programming language, this tool offers advanced analytics through template-based frameworks. Rapid Miner also provides functionality like data preprocessing and visualization, predictive analytics and statistical modeling, evaluation and deployment.

Reliability

OPEN REFINE, the only problem is the usage of Heap memory that limit the real use cases with huge files.

RAPID MINER is much stronger when it comes to analytical ETL, data and text mining and especially with enterprise edition of the server RapidAnalytics-predict to reporting dashboard.

Availability

OPEN REFINE, increases its availability by extending it with web services and linking it to databases like Freebase.

RAPID MINER is No.1 open source platform for predictive analysis. Rapid miner will change its model to business source, which means that older versions of software are available under OSI-certified open source license while the latest version, although open source for most parts of the product, will be available only as trial version or under a commercial purpose.

Security

OPEN REFINE, only listens to TCP requests coming from localhost (127.0.0.1) for security issues.

RAPIDMINER stores all the information used during processing of data set on the local system and thus there are no security issues it has to take care of.

Maintainability

OPEN REFINE, the most surprising architectural feature for many is the fact that OpenRefine has no database, it runs off of its own inmemory datastore that is built upfront to be optimized for the operations required by faceted browsing and infinite undo.

RAPID MINER, It has no maintainability issues as the software is installed locally on the system as it is easy to handle the data provided.

Portability

OPEN REFINE, The clientside part of OpenRefine is implemented in Javascript and uses jQuery and jQueryUI to simplify portability across modern browsers.

RAPID MINER, the present popularity of the software owes much to the existence of Java Virtual Machines for all important platforms, along with the fact that all code necessary to compile and run rapid miner is included in the distribution.

Performance

RAPID MINER, There is no performance issue regarding to this software as this most used software tool for data mining, predictive analytics and other fields. It provides easy integration of process as web services and offers additional extensional with scripts if something is missing. Rapid

miner is much stronger in performance when it comes to analytical ETL, data and text mining. Depending on used algorithms, Rapid Miner can use much more data. The graphical user interface is really powerful but sometimes this also adds to the complexity.

OPEN REFINE, The Open Refine server is just a web server implemented in Java. The use of Java strikes a balance between performance and portability across operating system.