ent No.8

ion of ETL
ess

Batch:A2 Roll No.:16010421063 Experiment No.:8

Aim: Execution of ETL process

Resources needed: Different RDBMS such as MySQL, Postgres and Excel, CSV, Rapidminer 5.3/ Latest vision

Theory

Data Warehouse:

An analytics-focused type of data management system called a data warehouse is intended to assist and allow business intelligence (BI) activities. Large amounts of historical data are frequently included in data warehouses, which are only designed to be used for queries and analysis. Application log files and transaction apps are only two examples of the many different sources from which the data in a data warehouse often comes.

Big data from various sources is centralised and combined in a data warehouse. Because of its analytical skills, businesses can get more out of their data and make better decisions. It gradually compiles a historical record that data scientists and business analysts can find quite useful. Because to these features, a data warehouse can be regarded as an organization's "single source of truth."

ETL:

Extract, Transform, Load (ETL) refers to a process in database usage and especially in data warehousing. Data extraction is where data is extracted from homogeneous or heterogeneous data sources; data transformation where the data is transformed for storing in the proper format or structure for the purposes of querying and analysis; data loading where the data is loaded into the final target database, more specifically, an operational data store, data mart, or data warehouse.

One may improve their chances of achieving better connection and scalability by employing a well-established ETL framework. A decent ETL tool must be able to interface with the several different relational databases and read the various file formats employed by a business. ETL solutions have started to move into Enterprise Application Integration, or even Enterprise Service Bus, systems that now encompass a lot more than simply the extraction, transformation, and loading of data. Converting CSV files into formats usable by relational databases is one frequent use case for ETL technologies. ETL solutions make it feasible for users to input csv-like data feeds/files and import it into a database with as little code as possible, facilitating a typical translation of millions of records. ESTL instruments

RapidMiner:

RapidMiner provides data mining and machine learning procedures including: data loading and transformation (Extract, transform, load (ETL)), data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment. RapidMiner is written in the Java programming language. RapidMiner provides a GUI to design and execute analytical workflows. Those workflows are called “Processes” in RapidMiner and they consist of multiple “Operators”. Each operator performs a single task within the process, and the output of each operator forms the input of the next one.

Alternatively, the engine can be called from other programs or used as an API. Individual functions can be called from the command line. RapidMiner provides learning schemes, models and algorithms and can be extended using R and Python scripts.

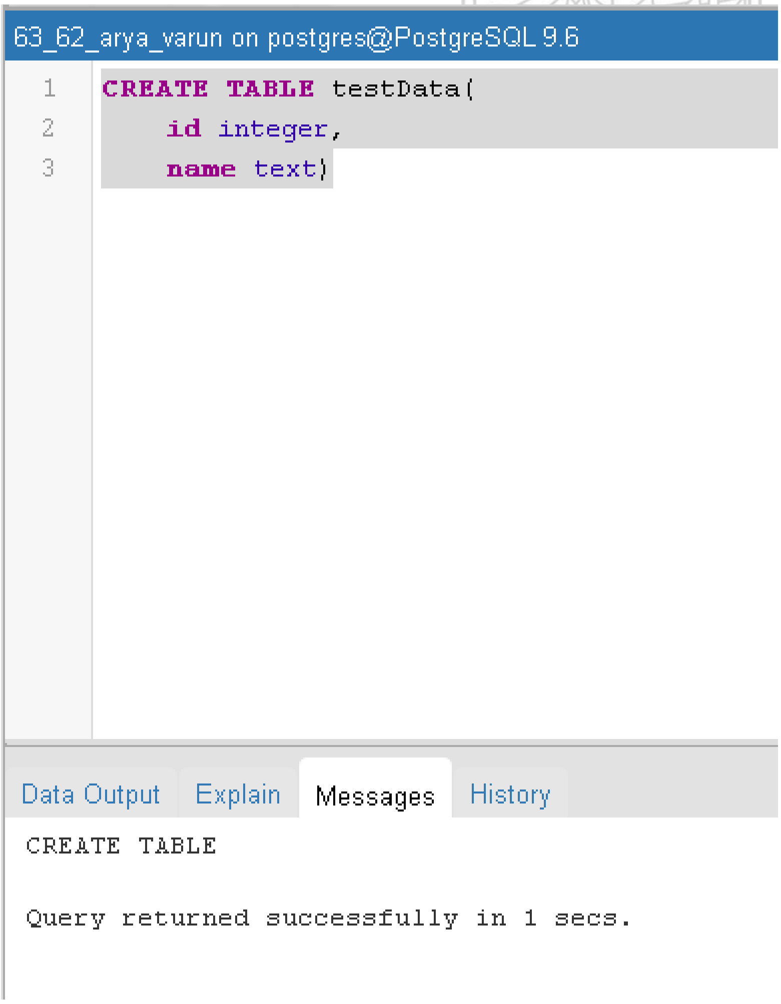
Activities:

For ETL:

1. Go through the tutorial provided by RapidMiner
2. Install https://rapidminer.software.informer.com/download/#downloading
3. Extract data from 2 to 3 heterogeneous sources such as excel, MYSQL, Postgres etc.
4. Download any data set from *https://www.kaggle.com/datasets* or similar website
5. Apply five different transformations and filters to the data with specific requirement
6. Prepare a report for the activities 2 and 4 (ETL part) with steps and visualisations applied.

Results:

Report for ETL

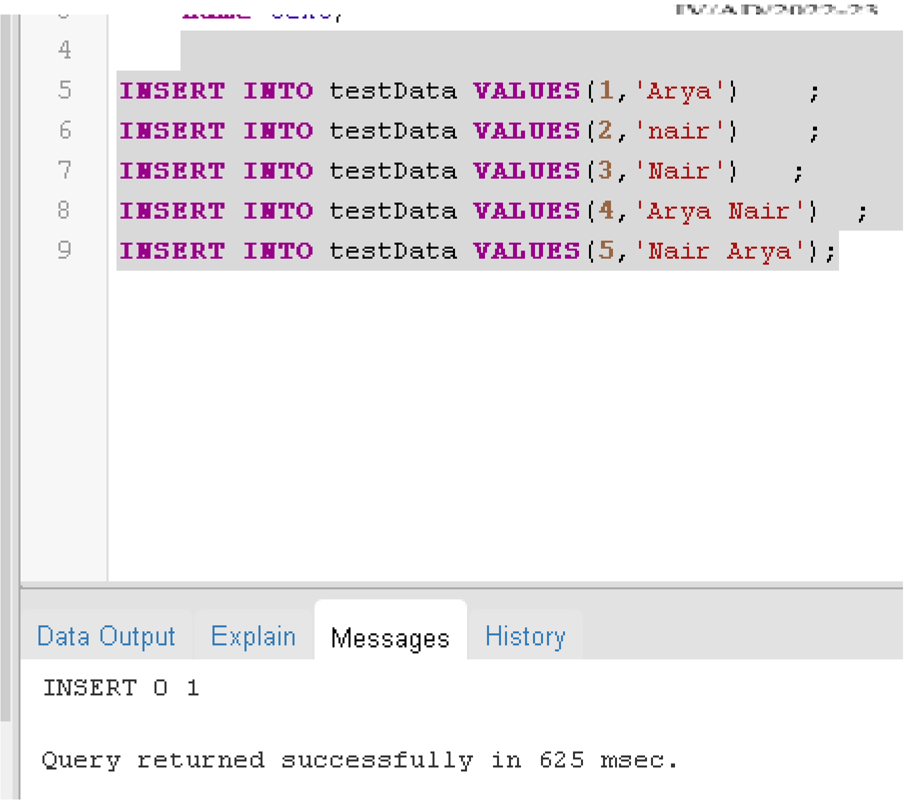


63_62_arya_varun on postgres@PostgreSQL 9.6

```
CREATE TABLE testData(
    id integer,
    name text)
```

```
CREATE TABLE

Query returned successfully in 1 secs.
```

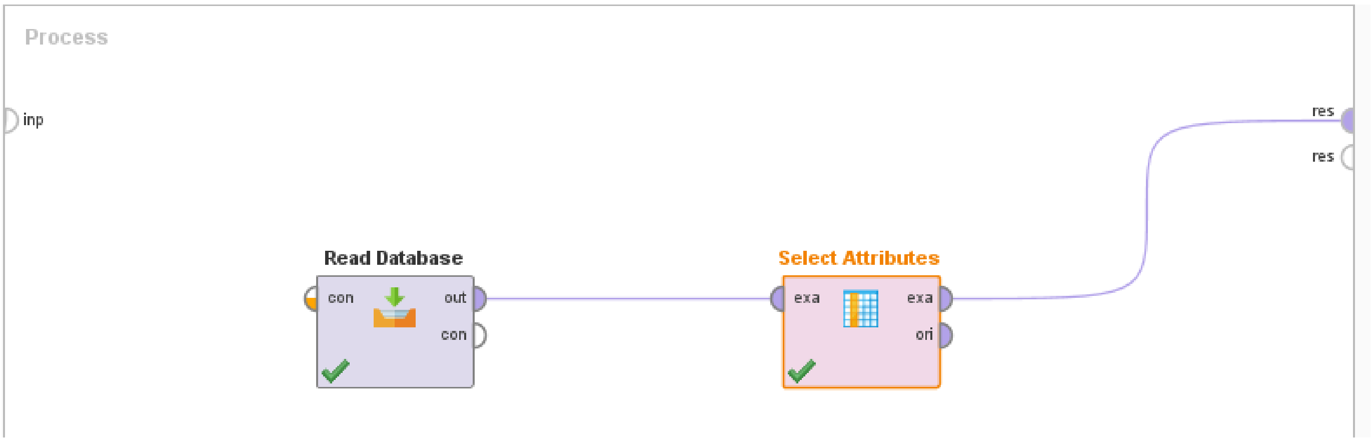


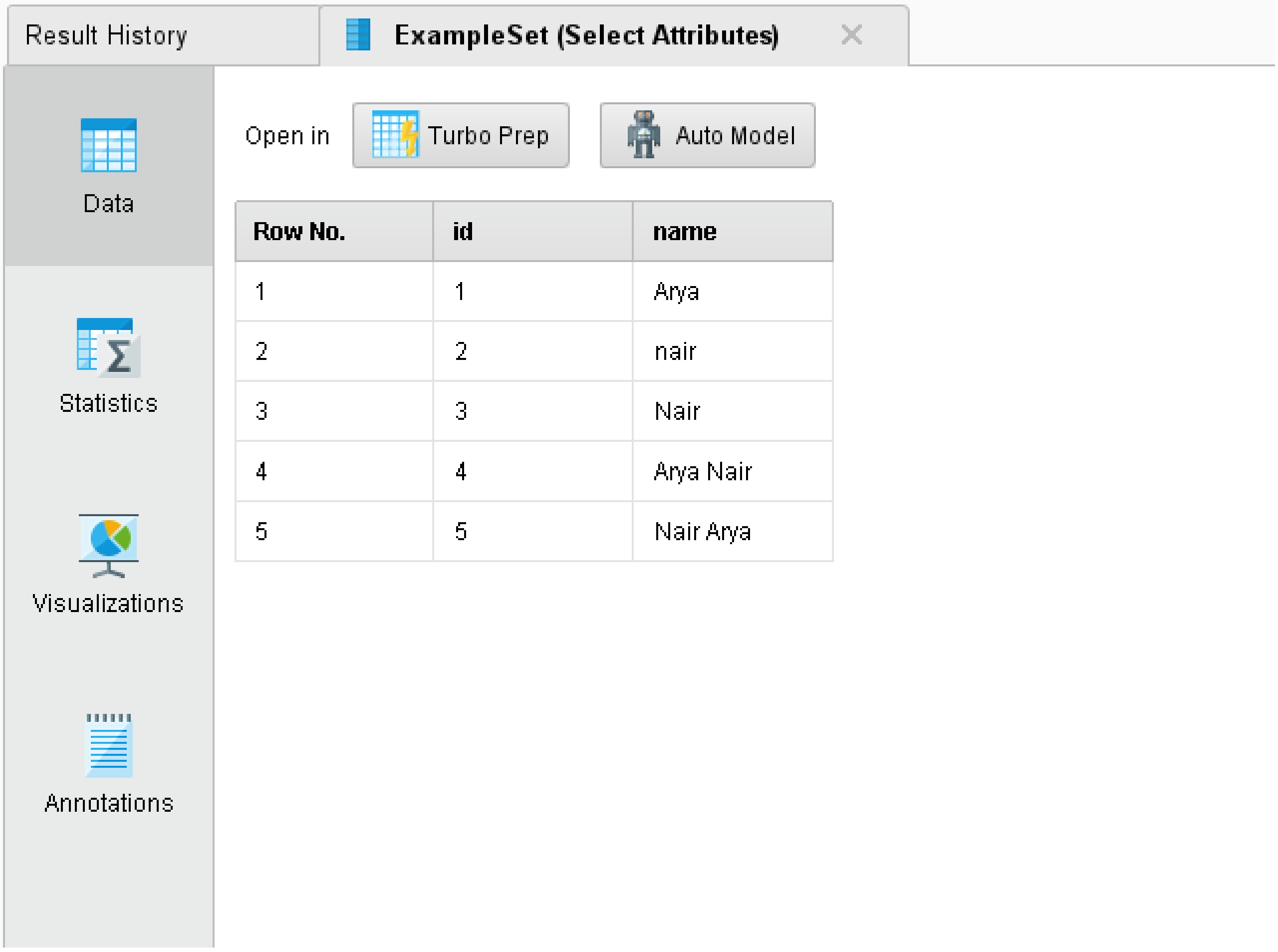
```
    name text);

INSERT INTO testData VALUES(1,'Arya')    ;
INSERT INTO testData VALUES(2,'nair')    ;
INSERT INTO testData VALUES(3,'Nair')   ;
INSERT INTO testData VALUES(4,'Arya Nair')  ;
INSERT INTO testData VALUES(5,'Nair Arya');
```

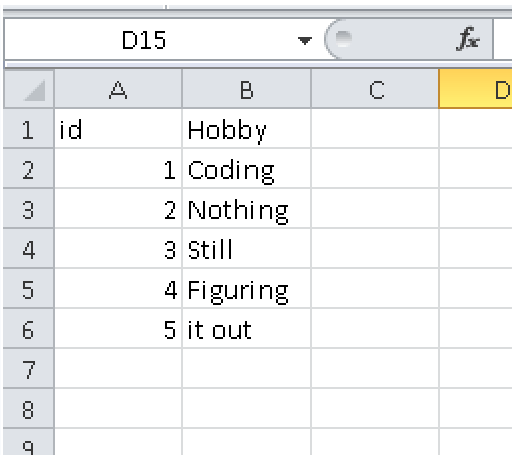
```
INSERT 0 1

Query returned successfully in 625 msec.
```

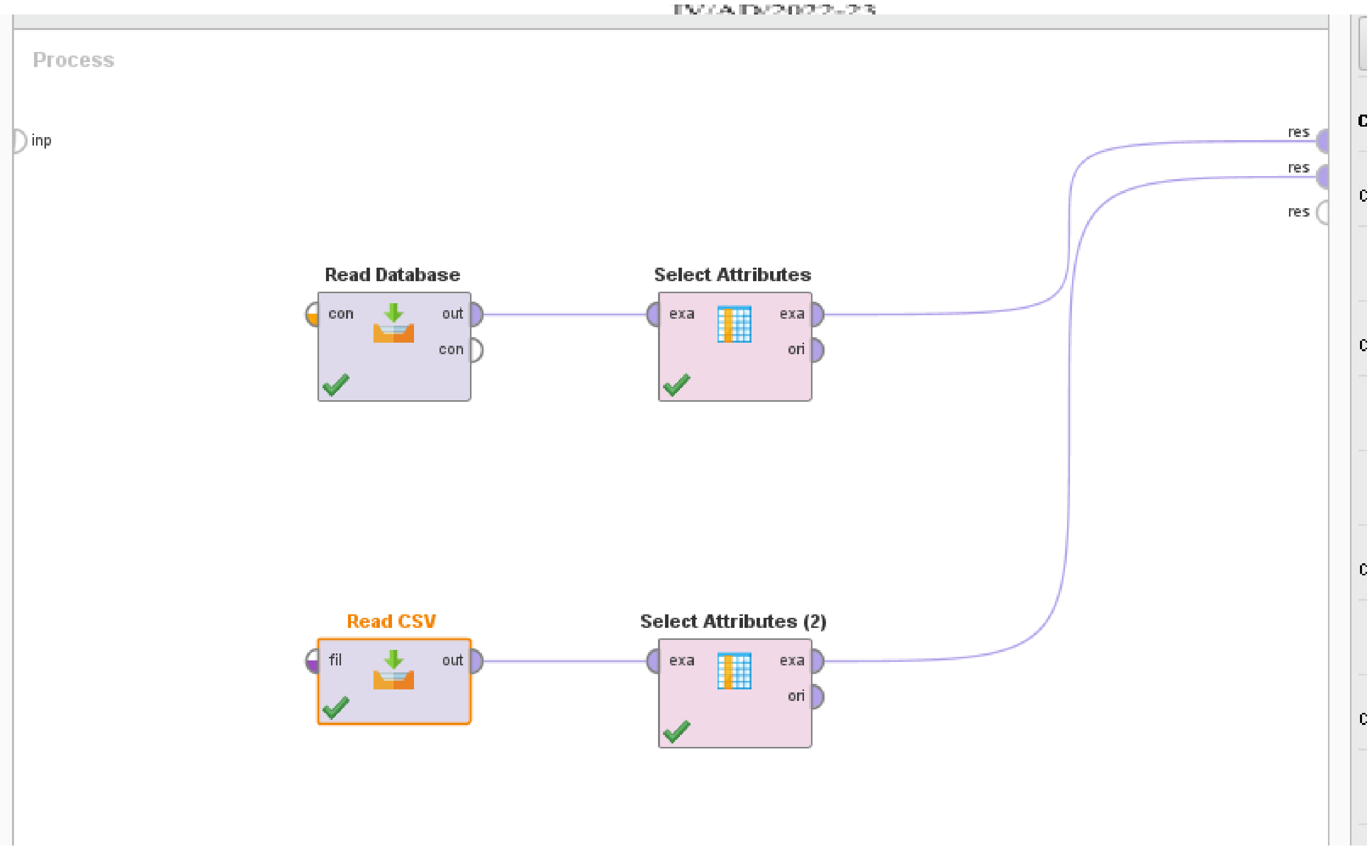


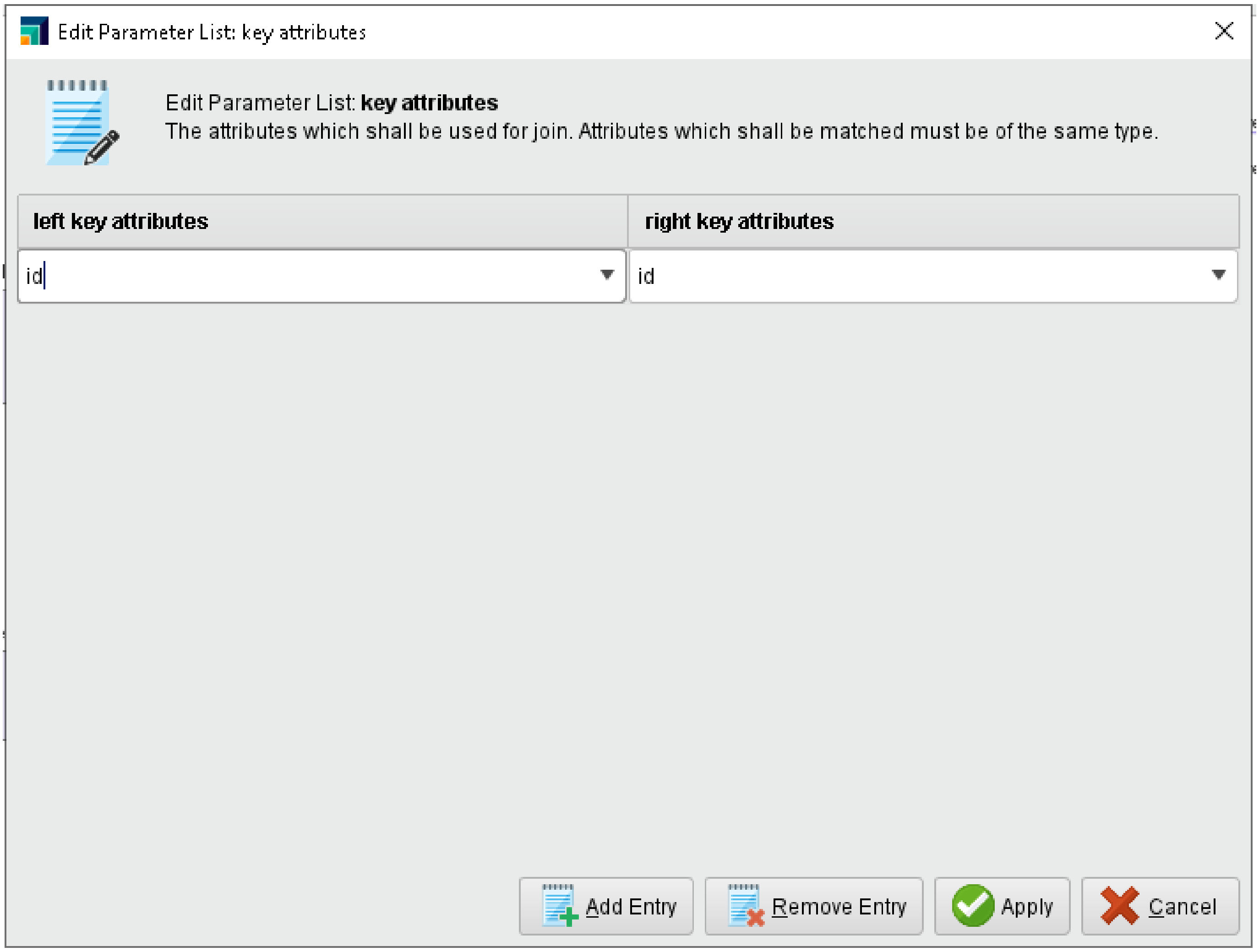


Row No.	id	name
1	1	Arya
2	2	nair
3	3	Nair
4	4	Arya Nair
5	5	Nair Arya



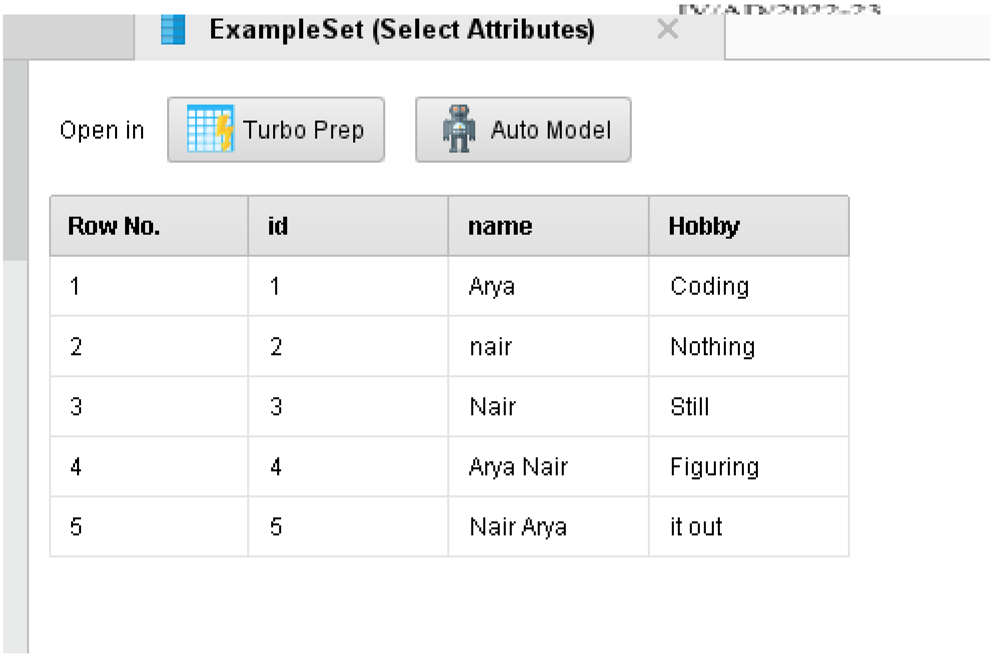
	A	B	C	D
1	id	Hobby		
2	1	Coding		
3	2	Nothing		
4	3	Still		
5	4	Figuring		
6	5	it out		
7				
8				
9				



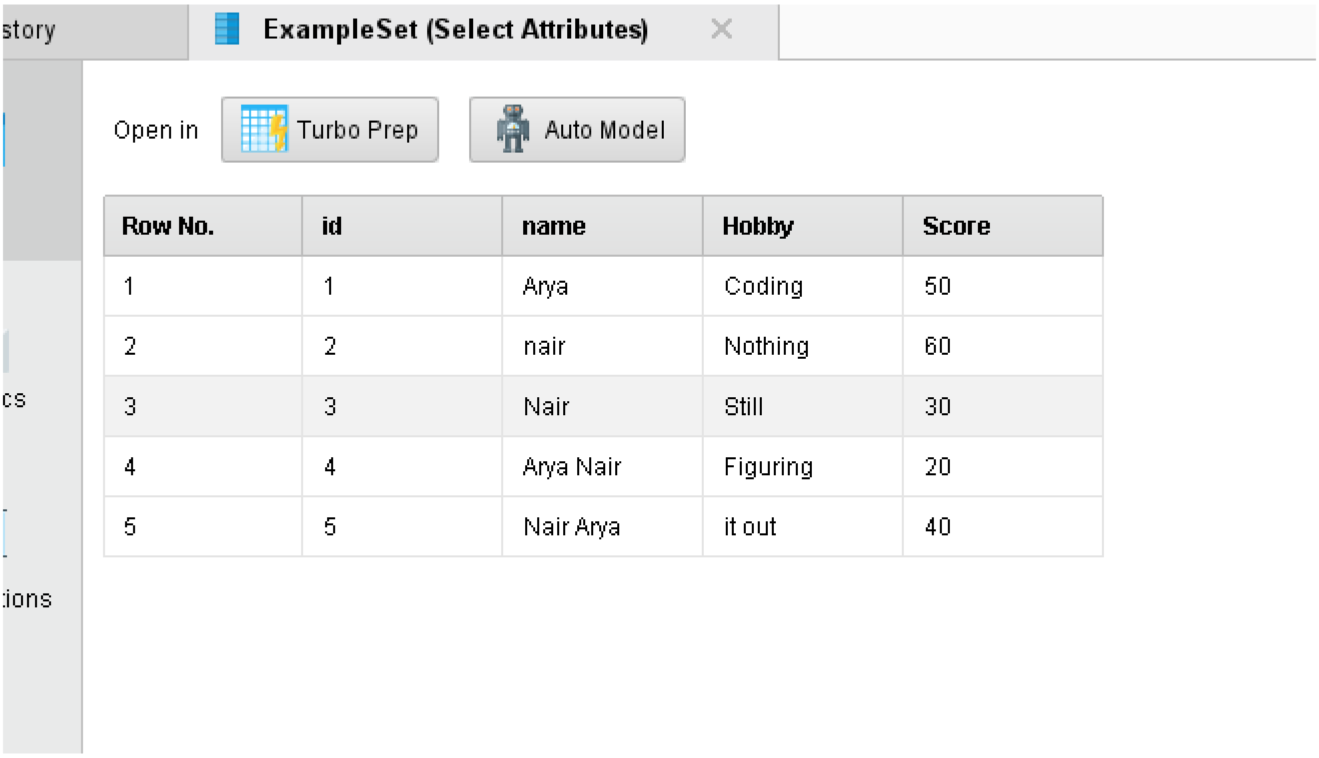

Edit Parameter List: key attributes

Edit Parameter List: **key attributes**
The attributes which shall be used for join. Attributes which shall be matched must be of the same type.

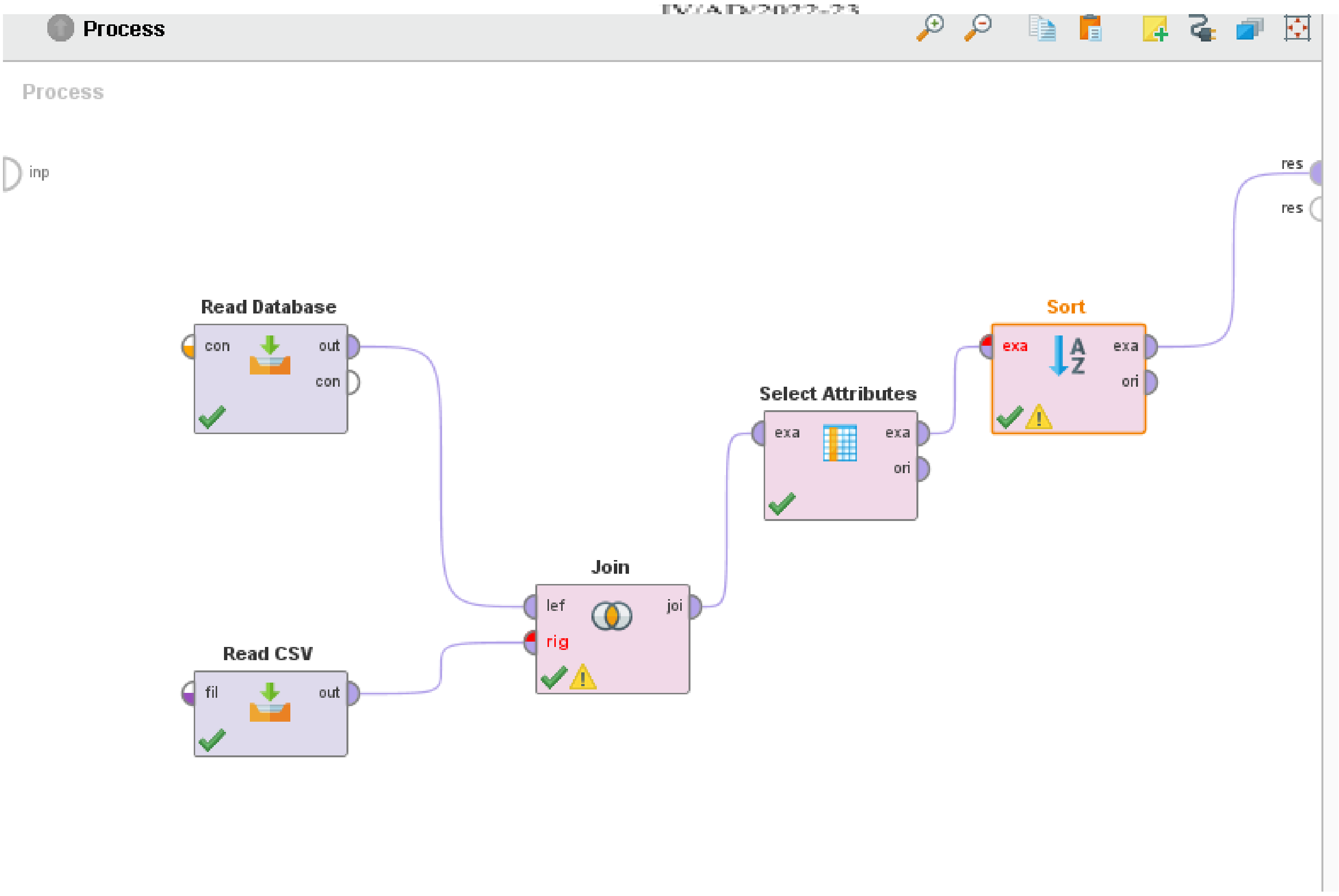
left key attributes	right key attributes
id	id

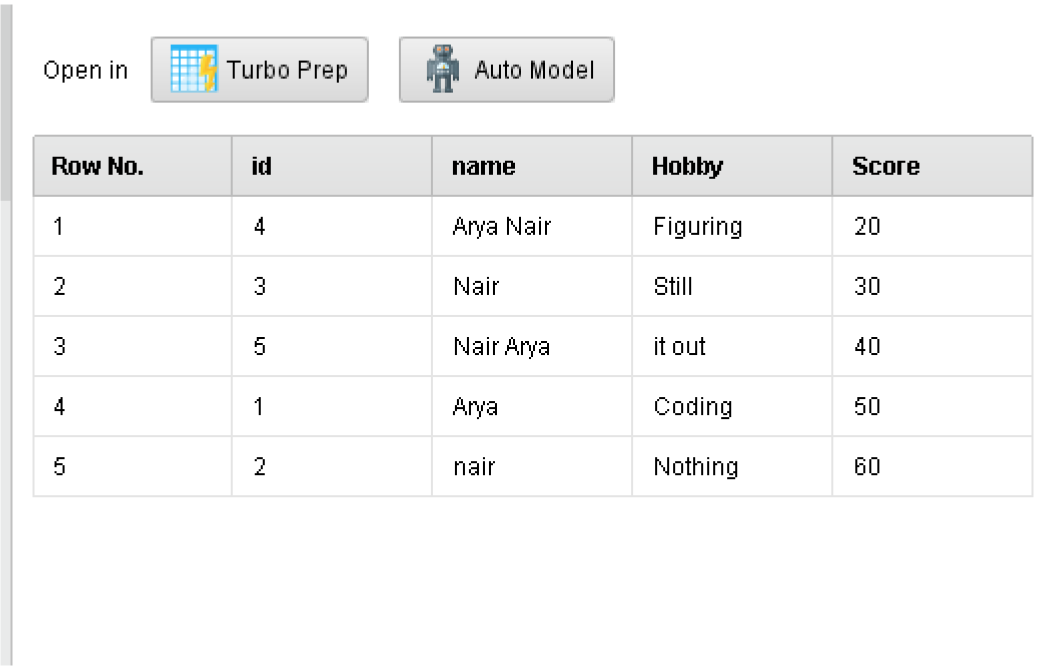


Row No.	id	name	Hobby
1	1	Arya	Coding
2	2	nair	Nothing
3	3	Nair	Still
4	4	Arya Nair	Figuring
5	5	Nair Arya	it out

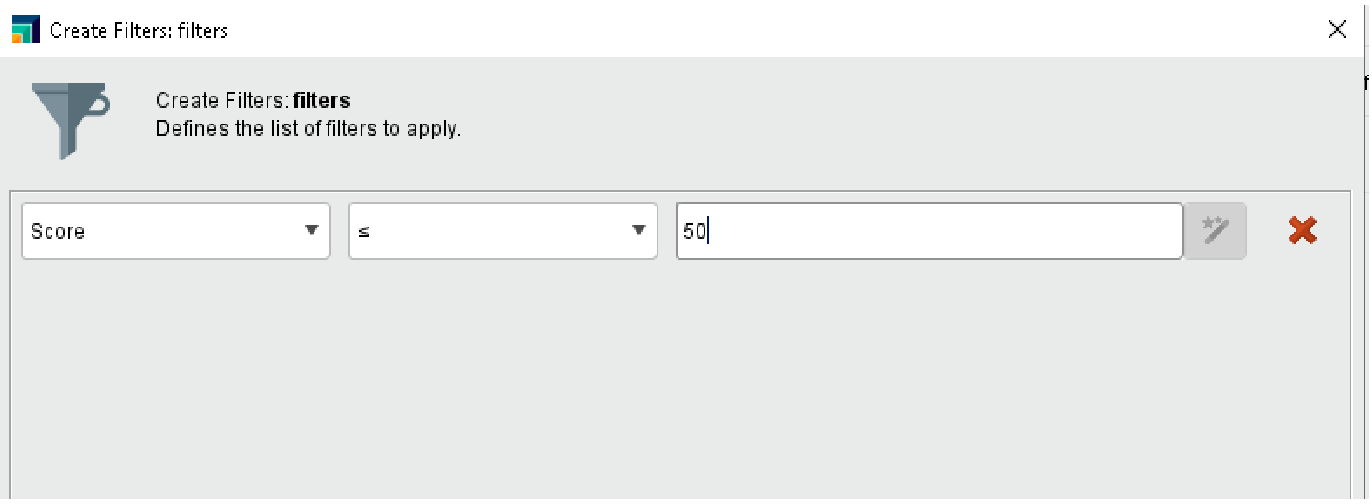


Row No.	id	name	Hobby	Score
1	1	Arya	Coding	50
2	2	nair	Nothing	60
3	3	Nair	Still	30
4	4	Arya Nair	Figuring	20
5	5	Nair Arya	it out	40



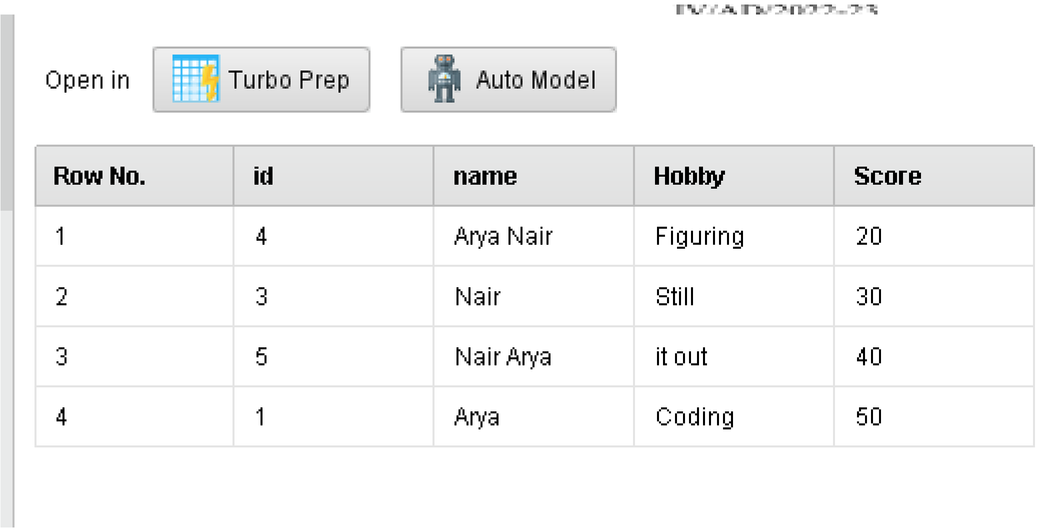


Row No.	id	name	Hobby	Score
1	4	Arya Nair	Figuring	20
2	3	Nair	Still	30
3	5	Nair Arya	it out	40
4	1	Arya	Coding	50
5	2	nair	Nothing	60

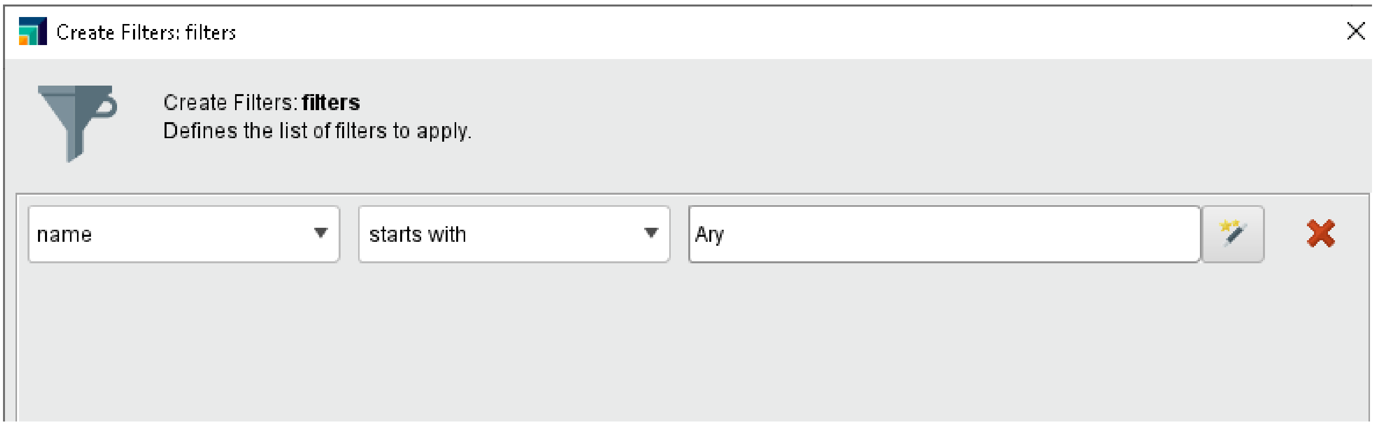

Create Filters: filters

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

Score	≤	50



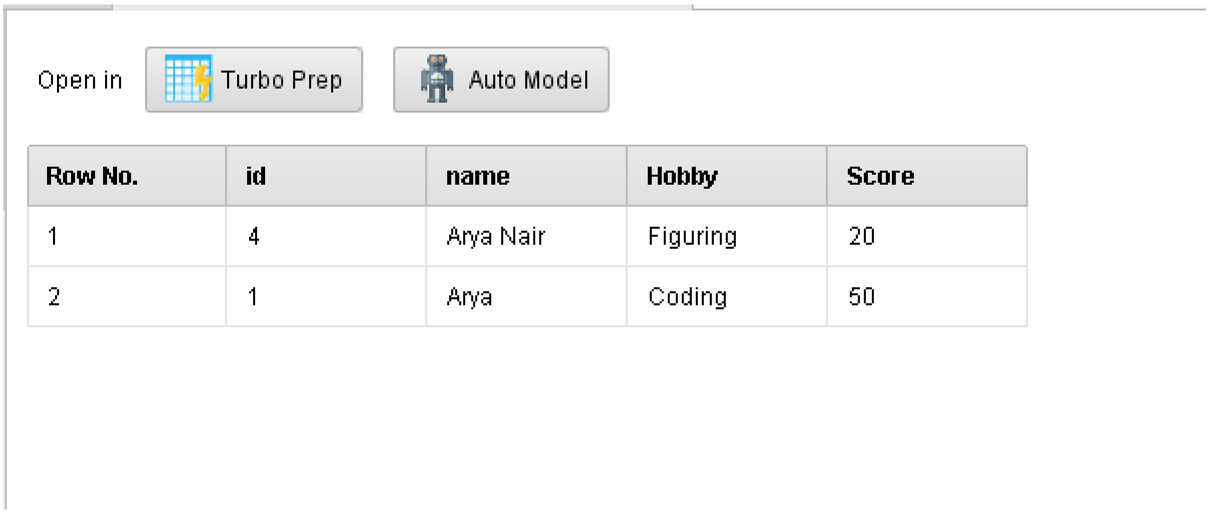
Row No.	id	name	Hobby	Score
1	4	Arya Nair	Figuring	20
2	3	Nair	Still	30
3	5	Nair Arya	it out	40
4	1	Arya	Coding	50



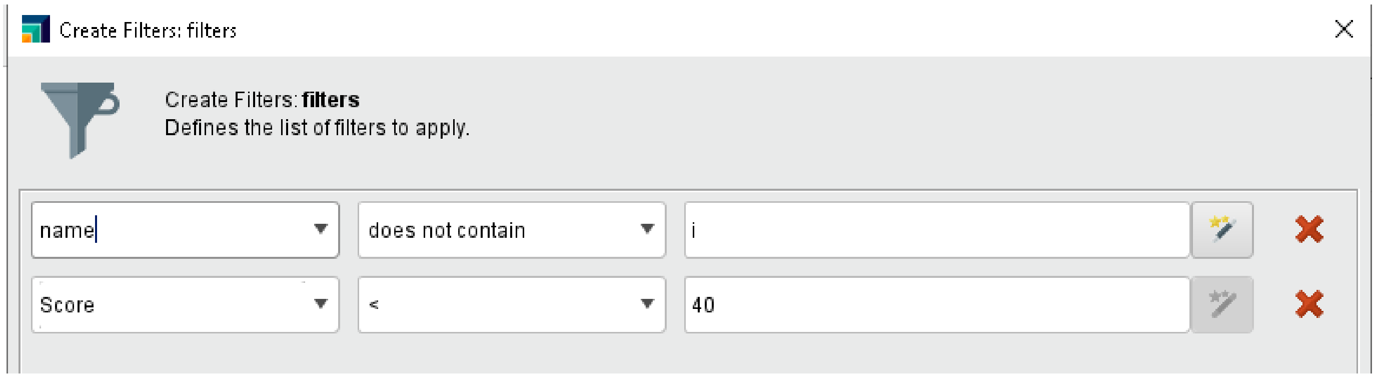
Create Filters: filters

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

name | starts with | Ary



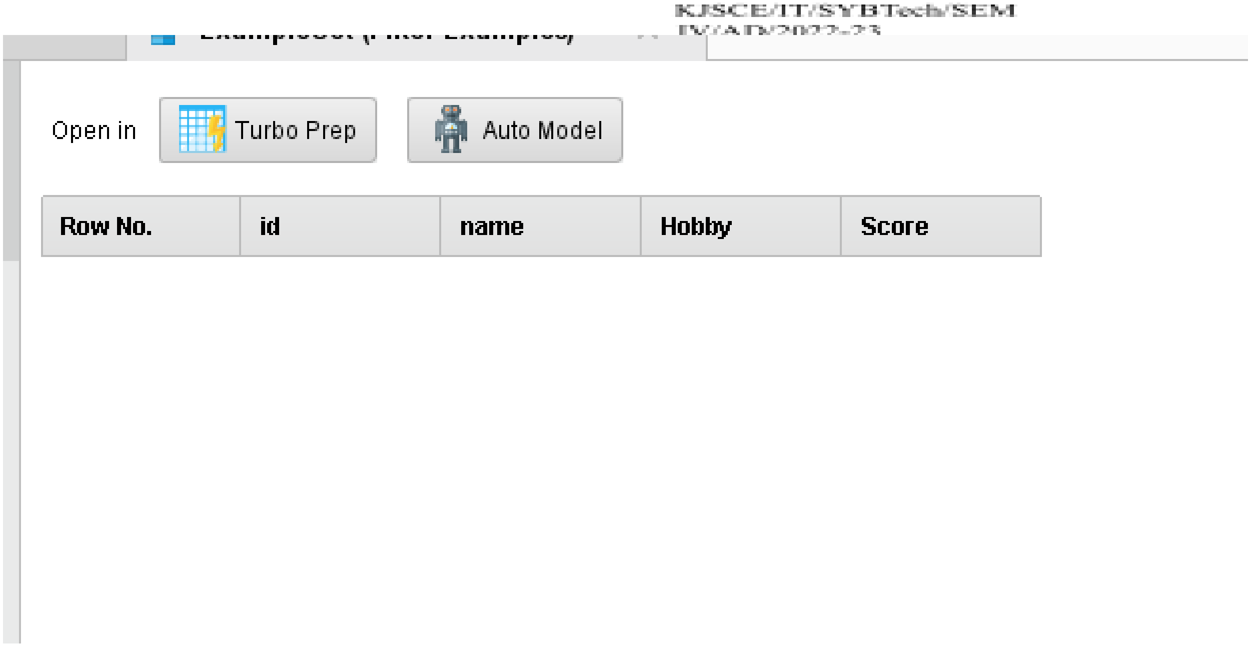
Row No.	id	name	Hobby	Score
1	4	Arya Nair	Figuring	20
2	1	Arya	Coding	50



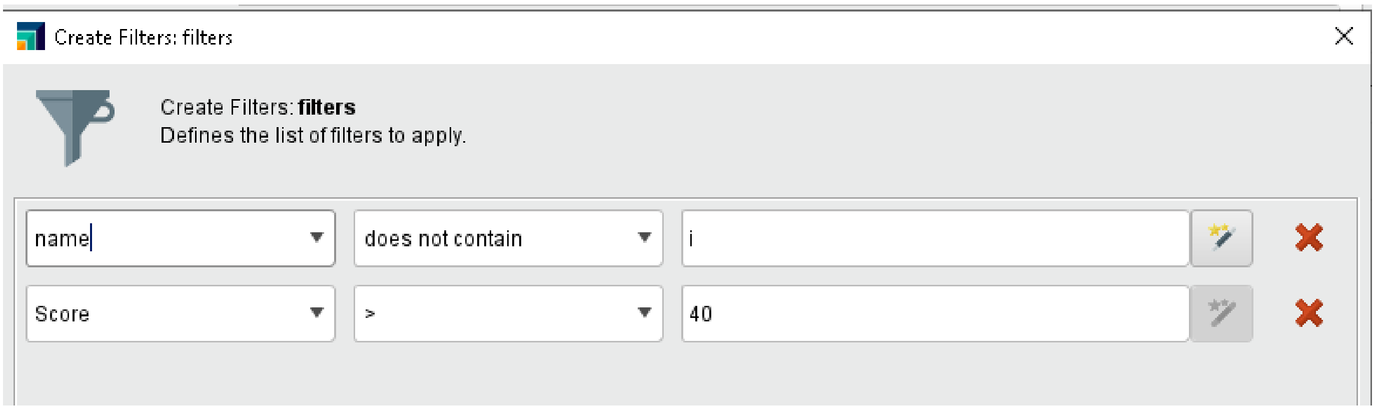
Create Filters: filters

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

name | does not contain | i
Score | < | 40

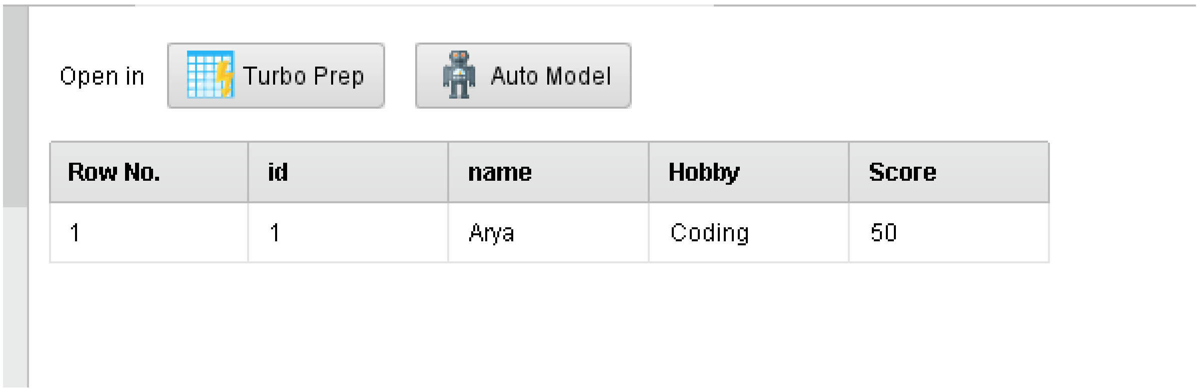


Row No.	id	name	Hobby	Score

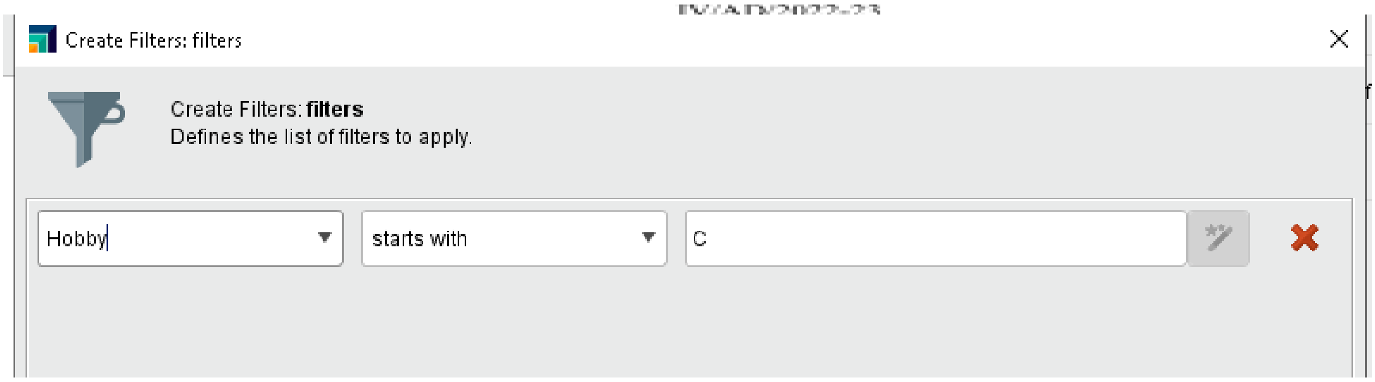


Create Filters: **filters**
Defines the list of filters to apply.

name	does not contain	i
Score	>	40



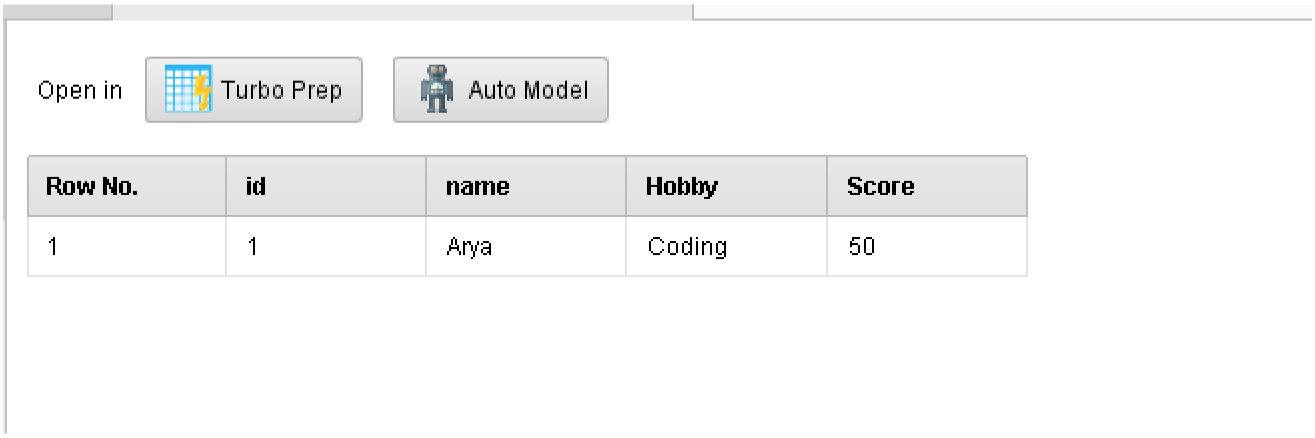
Row No.	id	name	Hobby	Score
1	1	Arya	Coding	50



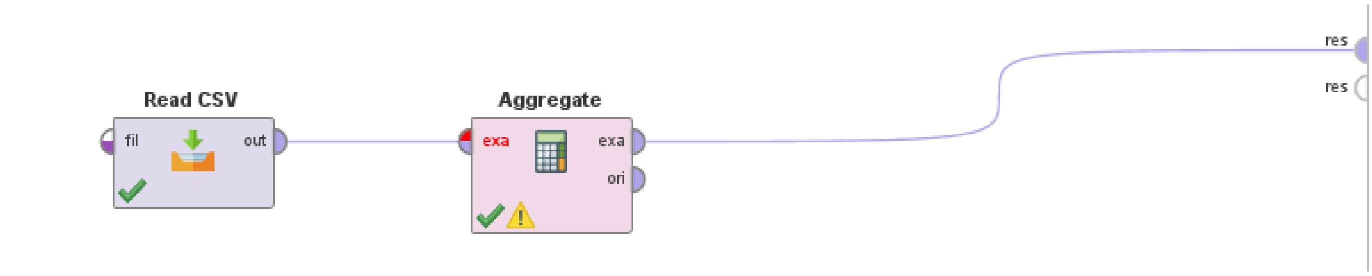
Create Filters: filters

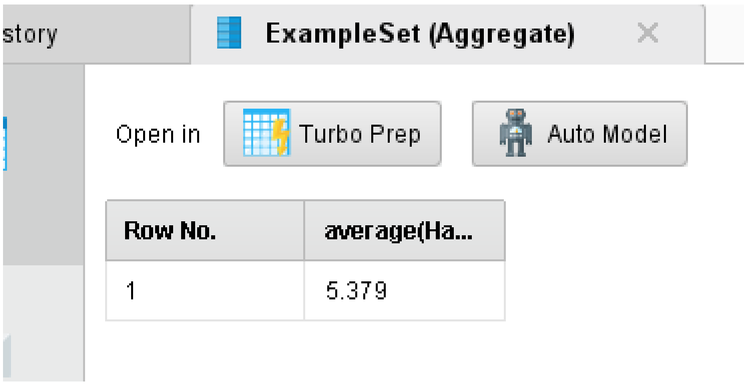
Create Filters: **filters**
Defines the list of filters to apply.

Hobby | starts with | C

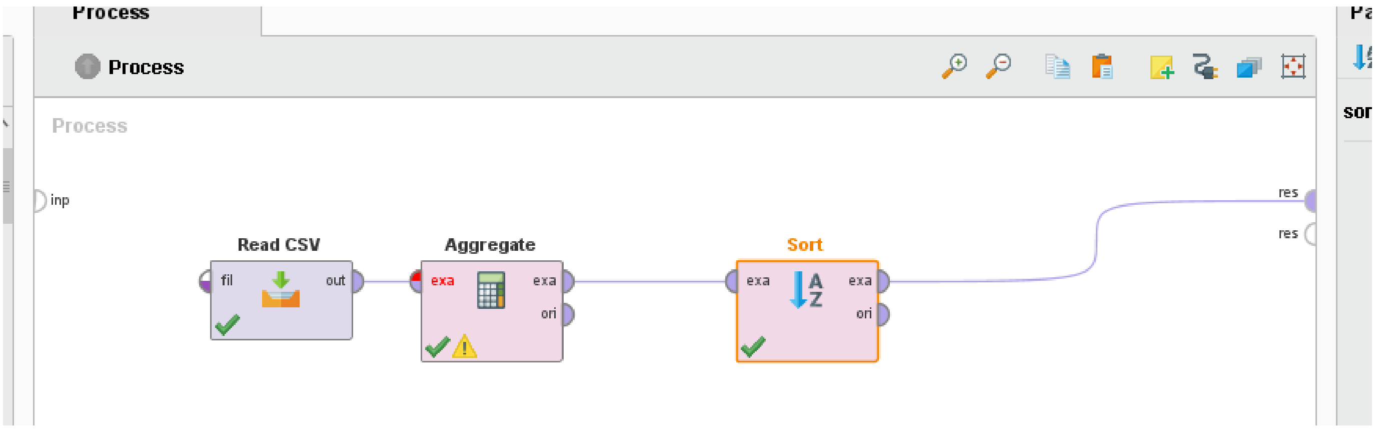


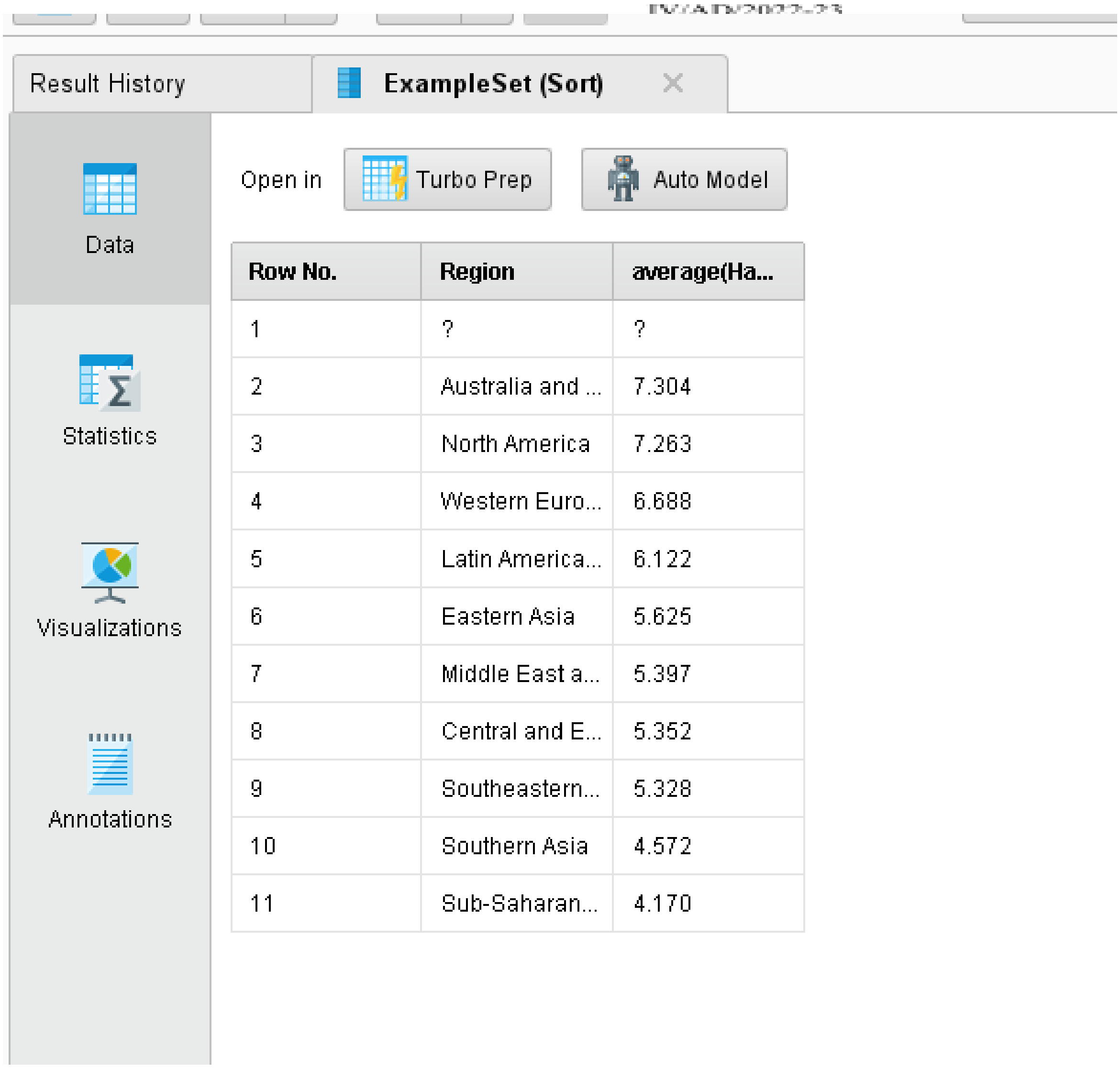
Row No.	id	name	Hobby	Score
1	1	Arya	Coding	50



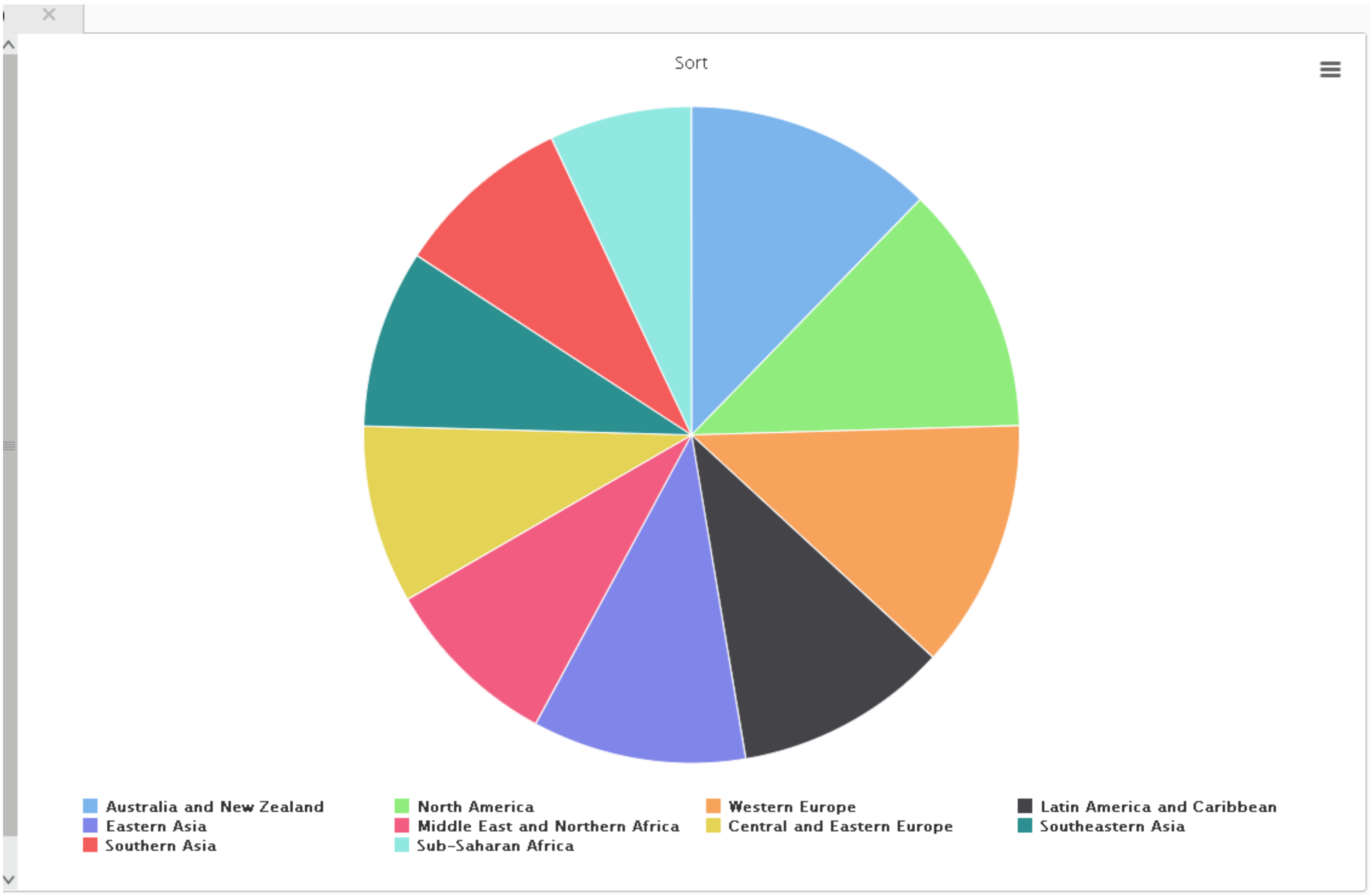


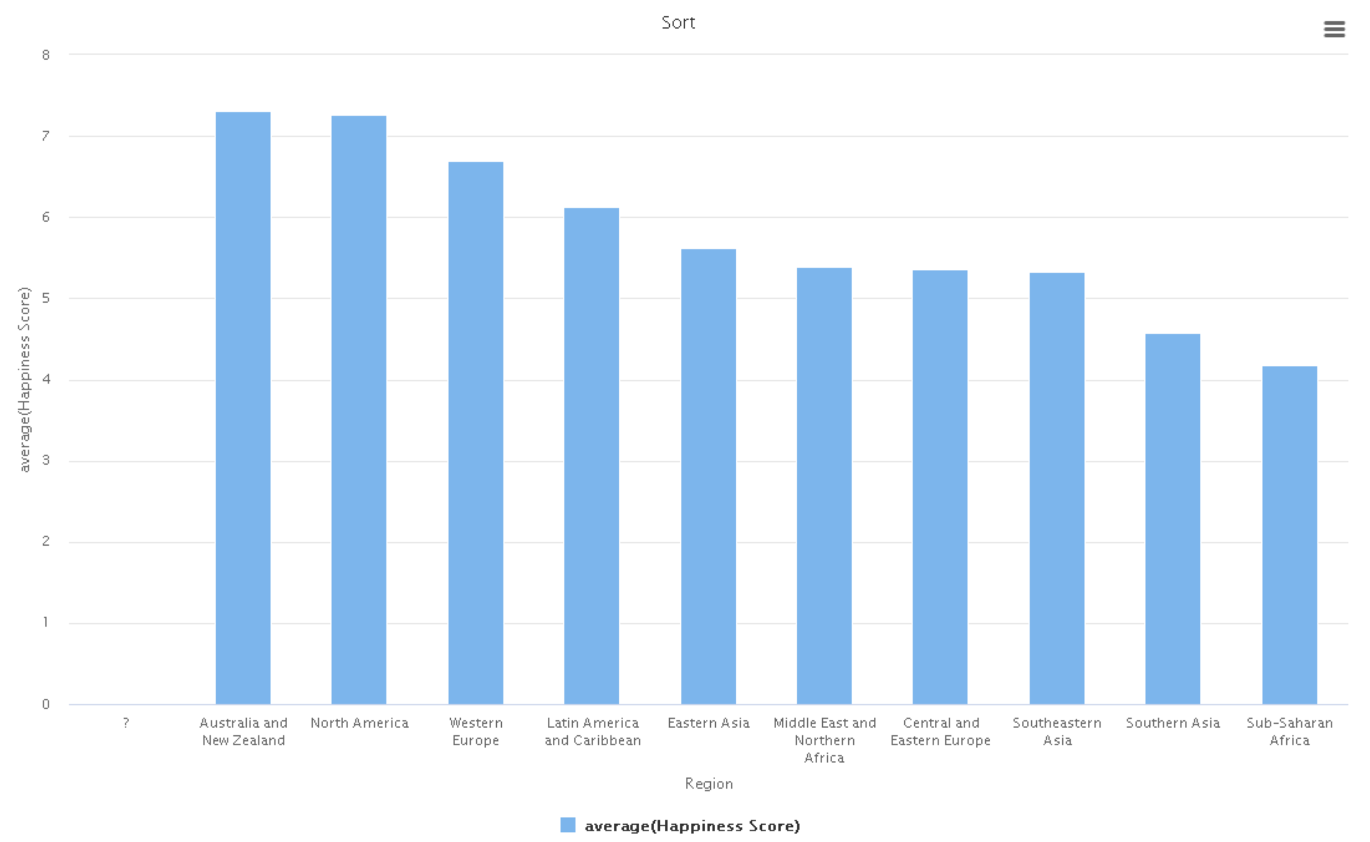
Row No.	average(Ha...
1	5.379

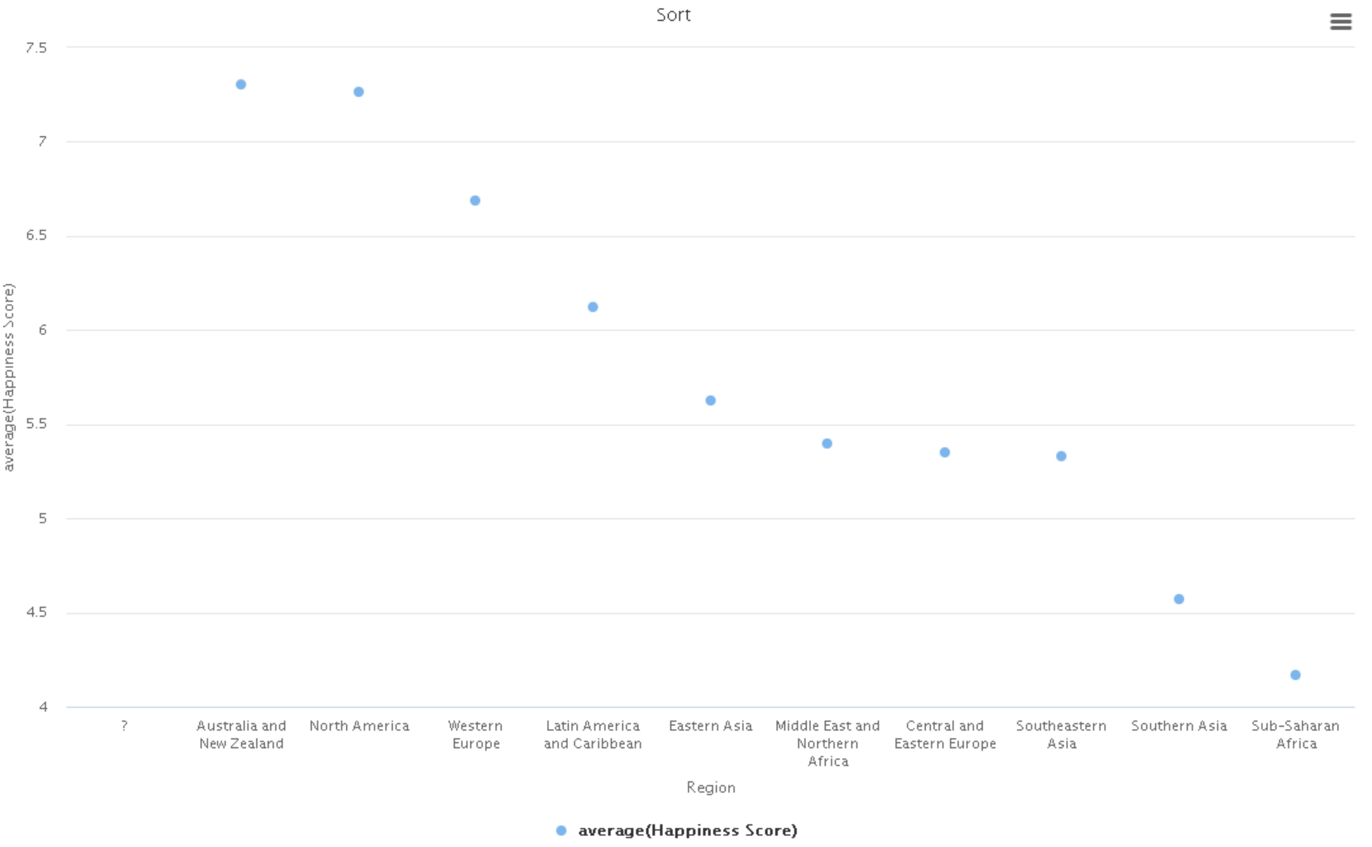




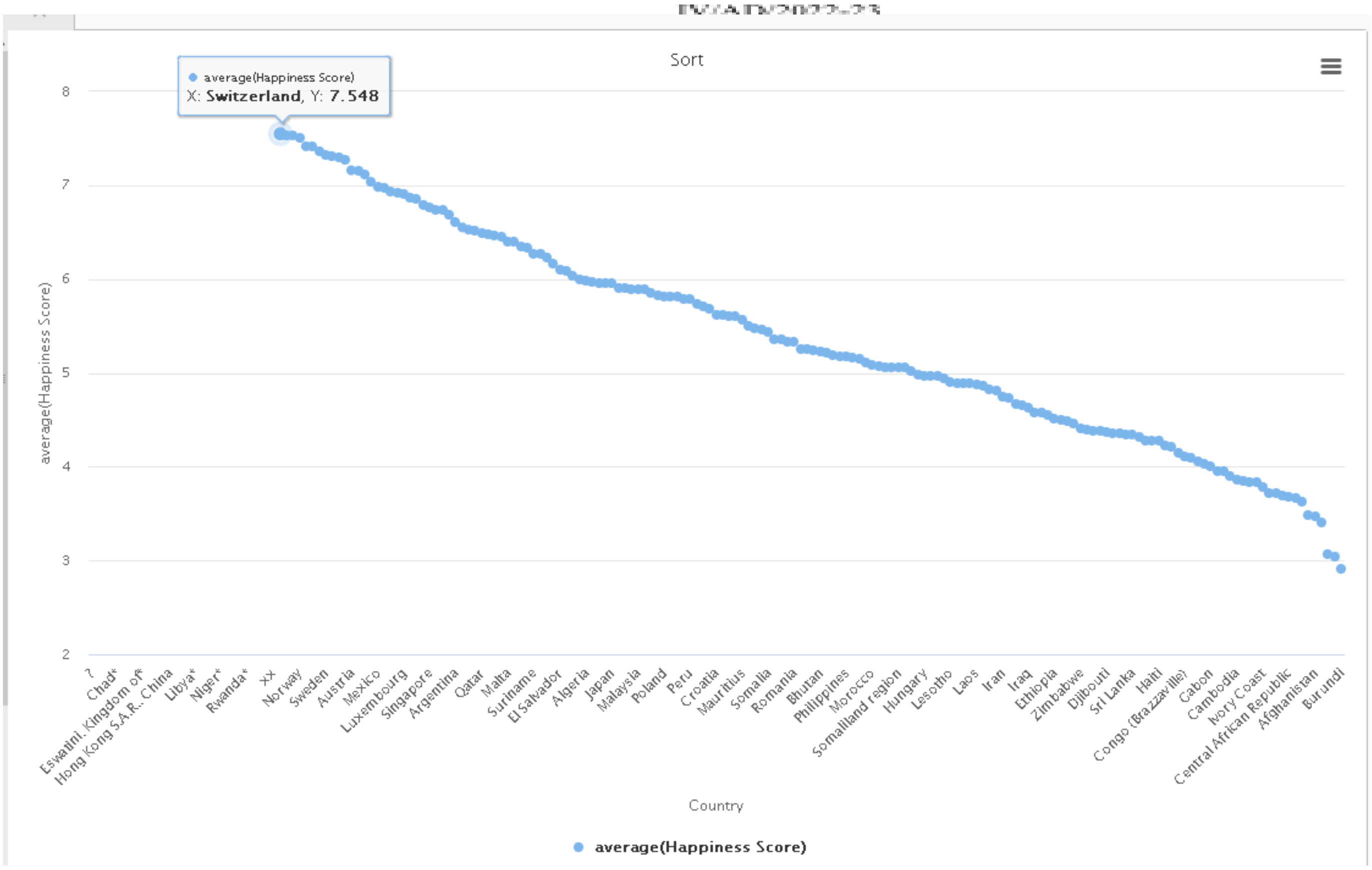
Row No.	Region	average(Ha...
1	?	?
2	Australia and ...	7.304
3	North America	7.263
4	Western Euro...	6.688
5	Latin America...	6.122
6	Eastern Asia	5.625
7	Middle East a...	5.397
8	Central and E...	5.352
9	Southeastern...	5.328
10	Southern Asia	4.572
11	Sub-Saharan...	4.170

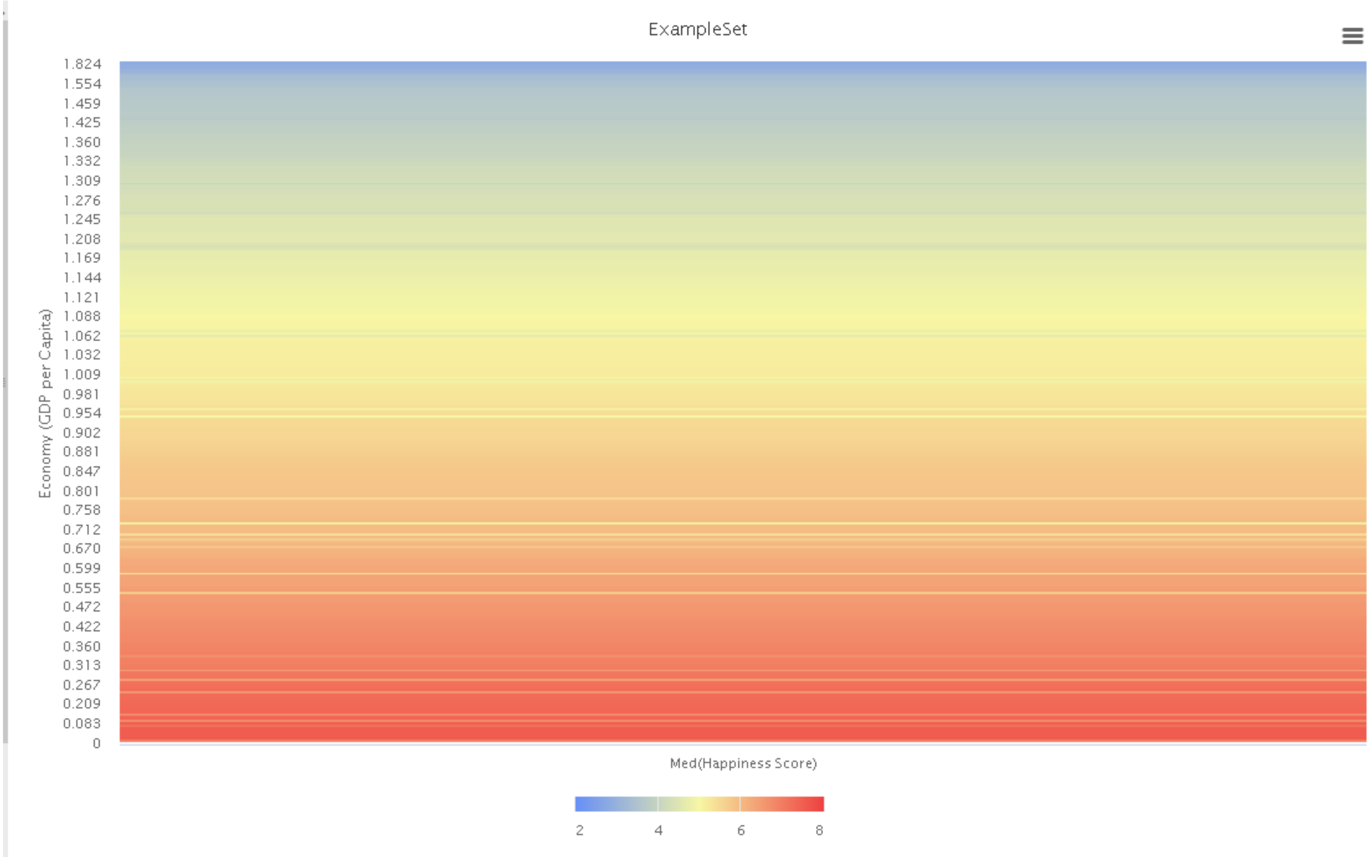


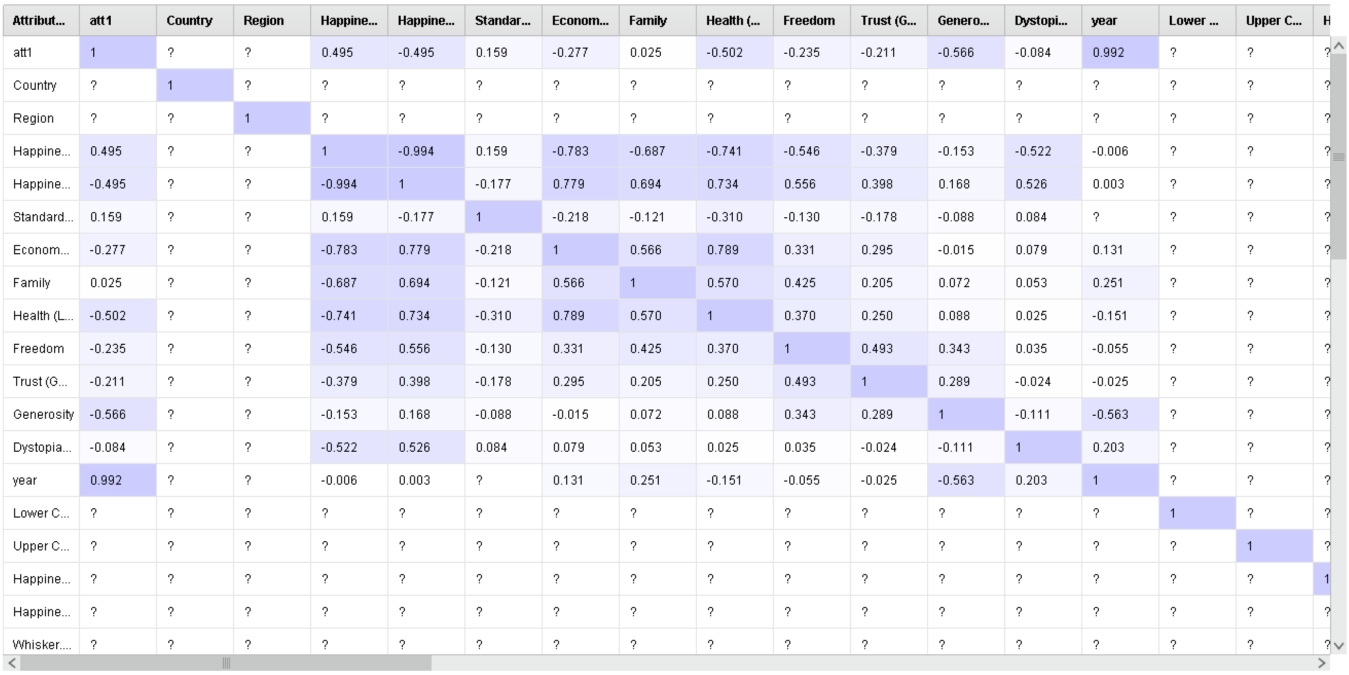




Australia is most happy


Sort
average(Happiness Score)
X: Switzerland, Y: 7.548
8
7
6
5
4
3
2
average(Happiness Score)
?
Chad*
Eswatini, Kingdom of*
Hong Kong S.A.R., China
Libya*
Niger*
Rwanda*
xx
Norway
Sweden
Austria
Mexico
Luxembourg
Singapore
Argentina
Qatar
Malta
Suriname
El Salvador
Algeria
Japan
Malaysia
Poland
Peru
Croatia
Mauritius
Somalia
Romania
Bhutan
Philippines
Morocco
Somaliland region
Hungary
Lesotho
Laos
Iran
Iraq
Ethiopia
Zimbabwe
Djibouti
Sri Lanka
Haiti
Congo (Brazzaville)
Gabon
Cambodia
Ivory Coast
Central African Republic
Afghanistan
Burundi
Country
average(Happiness Score)


ExampleSet
1.824
1.554
1.459
1.425
1.360
1.332
1.309
1.276
1.245
1.208
1.169
1.144
1.121
1.088
1.062
1.032
1.009
0.981
0.954
0.902
0.881
0.847
0.801
0.758
0.712
0.670
0.599
0.555
0.472
0.422
0.360
0.313
0.267
0.209
0.083
0
Economy (GDP per Capita)
Med(Happiness Score)
2
4
6
8



Attribut...	att1	Country	Region	Happine...	Happine...	Standar...	Econom...	Family	Health (...	Freedom	Trust (G...	Genero...	Dystopi...	year	Lower ...	Upper C...	H
att1	1	?	?	0.495	-0.495	0.159	-0.277	0.025	-0.502	-0.235	-0.211	-0.566	-0.084	0.992	?	?	?
Country	?	1	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Region	?	?	1	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Happine...	0.495	?	?	1	-0.994	0.159	-0.783	-0.687	-0.741	-0.546	-0.379	-0.153	-0.522	-0.006	?	?	?
Happine...	-0.495	?	?	-0.994	1	-0.177	0.779	0.694	0.734	0.556	0.398	0.168	0.526	0.003	?	?	?
Standard...	0.159	?	?	0.159	-0.177	1	-0.218	-0.121	-0.310	-0.130	-0.178	-0.088	0.084	?	?	?	?
Econom...	-0.277	?	?	-0.783	0.779	-0.218	1	0.566	0.789	0.331	0.295	-0.015	0.079	0.131	?	?	?
Family	0.025	?	?	-0.687	0.694	-0.121	0.566	1	0.570	0.425	0.205	0.072	0.053	0.251	?	?	?
Health (L...	-0.502	?	?	-0.741	0.734	-0.310	0.789	0.570	1	0.370	0.250	0.088	0.025	-0.151	?	?	?
Freedom	-0.235	?	?	-0.546	0.556	-0.130	0.331	0.425	0.370	1	0.493	0.343	0.035	-0.055	?	?	?
Trust (G...	-0.211	?	?	-0.379	0.398	-0.178	0.295	0.205	0.250	0.493	1	0.289	-0.024	-0.025	?	?	?
Generosity	-0.566	?	?	-0.153	0.168	-0.088	-0.015	0.072	0.088	0.343	0.289	1	-0.111	-0.563	?	?	?
Dystopia...	-0.084	?	?	-0.522	0.526	0.084	0.079	0.053	0.025	0.035	-0.024	-0.111	1	0.203	?	?	?
year	0.992	?	?	-0.006	0.003	?	0.131	0.251	-0.151	-0.055	-0.025	-0.563	0.203	1	?	?	?
Lower C...	?	?	?	?	?	?	?	?	?	?	?	?	?	?	1	?	?
Upper C...	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	1	?
Happine...	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	1
Happine...	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Whisker....	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?

Outcomes:CO3 Understanding of data warehouse and its multi-dimensional modeling

Conclusion: (Conclusion to be based on the outcomes achieved)

Successfully understood and implemented rapidminer and made analysis using the same

Grade: AA / AB / BB / BC / CC / CD /DD

Signature of faculty in-charge with date

References:

- https://www.oracle.com/in/database/what-is-a-data-warehouse
- Paulraj Ponniah, “Data Warehousing: Fundamentals for IT Professionals”, Wiley India