Section 1

a Preprocessing (Orange Tool) Class Works

1) Perform imputation on Heart Disease dataset.

Input

Data Info - Orange

Data table properties

Name: heart_disease

Size: 303 rows, 14 columns

Features: 7 categorical, 6 numeric

Targets: categorical outcome with 2 classes

Additional attributes

Name: Heart Disease dataset

Description: Data on the presence of heart disease in patients.

Author: A. Janosi, W. Steinbrunn, M. Pfisterer, R. Detrano

Year: 1988

Reference: Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64, 304-310.

303

Info

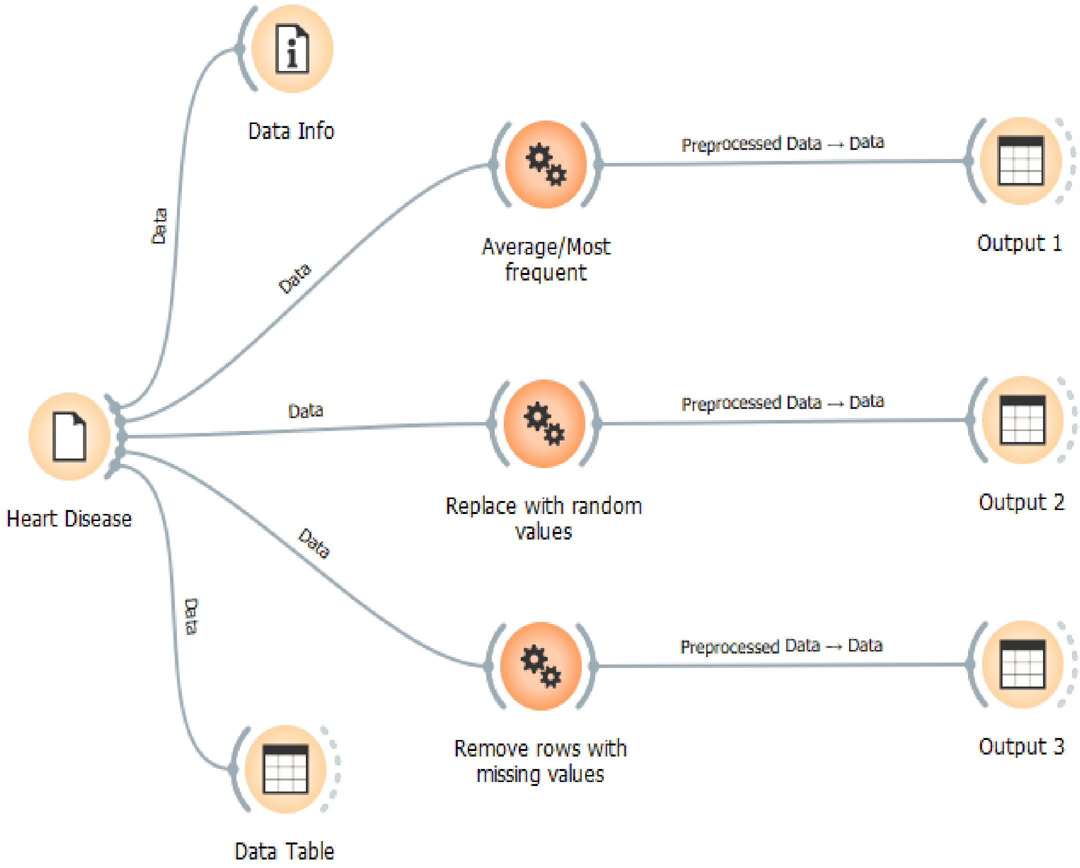
303 instances
13 features (0.2 % missing data)
Target with 2 values
No meta attributes.

	diameter narrowing	major vessels colored
303	0	?
288	0	?
193	1	?
167	0	?

Process

Imputation

- Replace Missing values with average or most frequent value.
- Replace with a random value.
- Remove Rows with missing values.



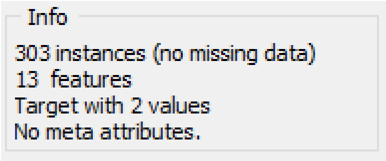
Output

Output 1 (Average/Most frequent value)

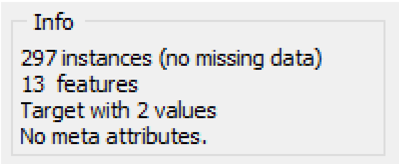
Info

303 instances (no missing data)
13 features
Target with 2 values
No meta attributes.

Output 2 (Replace with random value)

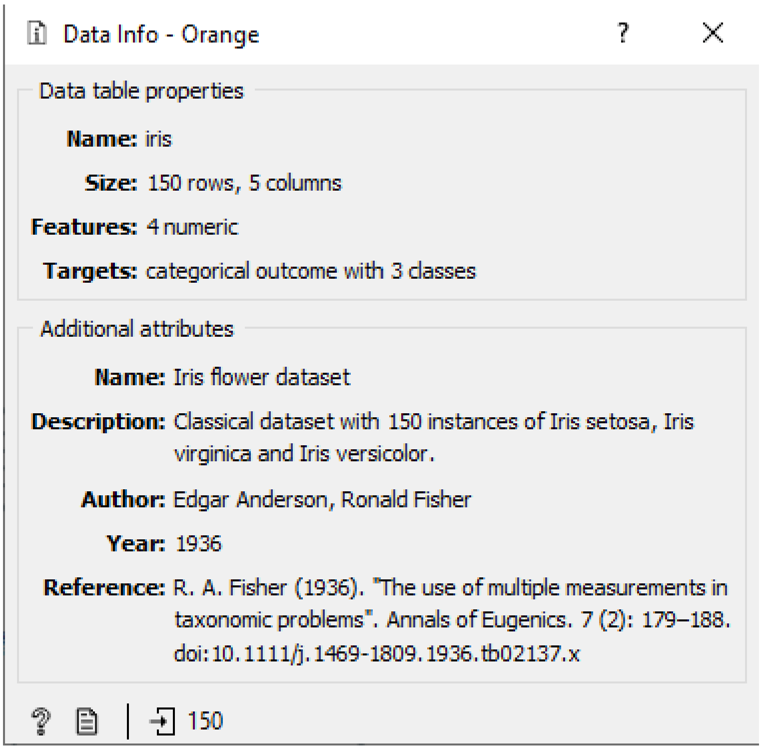


Output 3- Remove rows with missing values.



2) Perform Discretization on Iris dataset

Input



Info

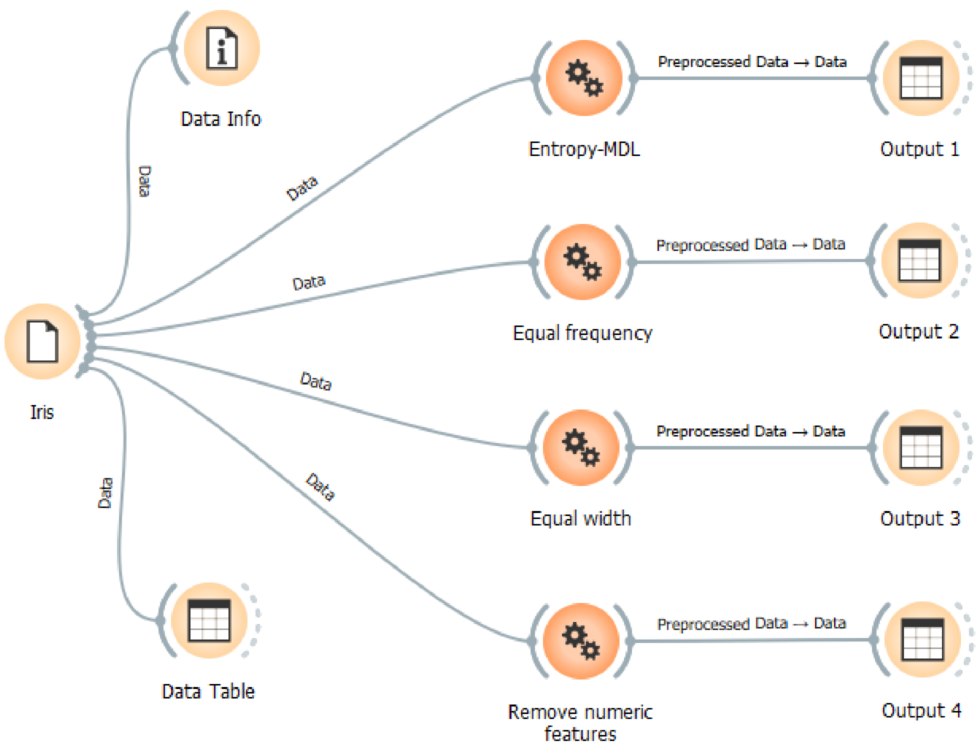
150 instances (no missing data)
4 features
Target with 3 values
No meta attributes.

	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	5.1	3.5	1.4	0.2
2	Iris-setosa	4.9	3.0	1.4	0.2
3	Iris-setosa	4.7	3.2	1.3	0.2
4	Iris-setosa	4.6	3.1	1.5	0.2
5	Iris-setosa	5.0	3.6	1.4	0.2

Process

Discretization

- Equal frequency discretization
- Equal width discretization
- Remove numeric values



Output

Output 1 (Equal frequency discretization)

	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	< 5.45	≥ 3.15	< 2.45	< 0.8
2	Iris-setosa	< 5.45	2.85 - 3.15	< 2.45	< 0.8
3	Iris-setosa	< 5.45	≥ 3.15	< 2.45	< 0.8
4	Iris-setosa	< 5.45	2.85 - 3.15	< 2.45	< 0.8
5	Iris-setosa	< 5.45	≥ 3.15	< 2.45	< 0.8

Output 2 (Equal width discretization)

	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	< 5.5	2.8 - 3.6	< 2.967	< 0.9
2	Iris-setosa	< 5.5	2.8 - 3.6	< 2.967	< 0.9
3	Iris-setosa	< 5.5	2.8 - 3.6	< 2.967	< 0.9
4	Iris-setosa	< 5.5	2.8 - 3.6	< 2.967	< 0.9
5	Iris-setosa	< 5.5	2.8 - 3.6	< 2.967	< 0.9

Output 3 (Remove numeric values)

	iris
1	Iris-setosa
2	Iris-setosa
3	Iris-setosa
4	Iris-setosa
5	Iris-setosa

3. Perform continuization on Titanic dataset.

Input

Data Info - Orange

Data table properties

Name: titanic

Size: 2201 rows, 4 columns

Features: 3 categorical

Targets: categorical outcome with 2 classes

Additional attributes

Name: Titanic dataset

Description: Data from the real Titanic passengers and their survival.

Author: Kaggle

Year: 1912

2201

Info

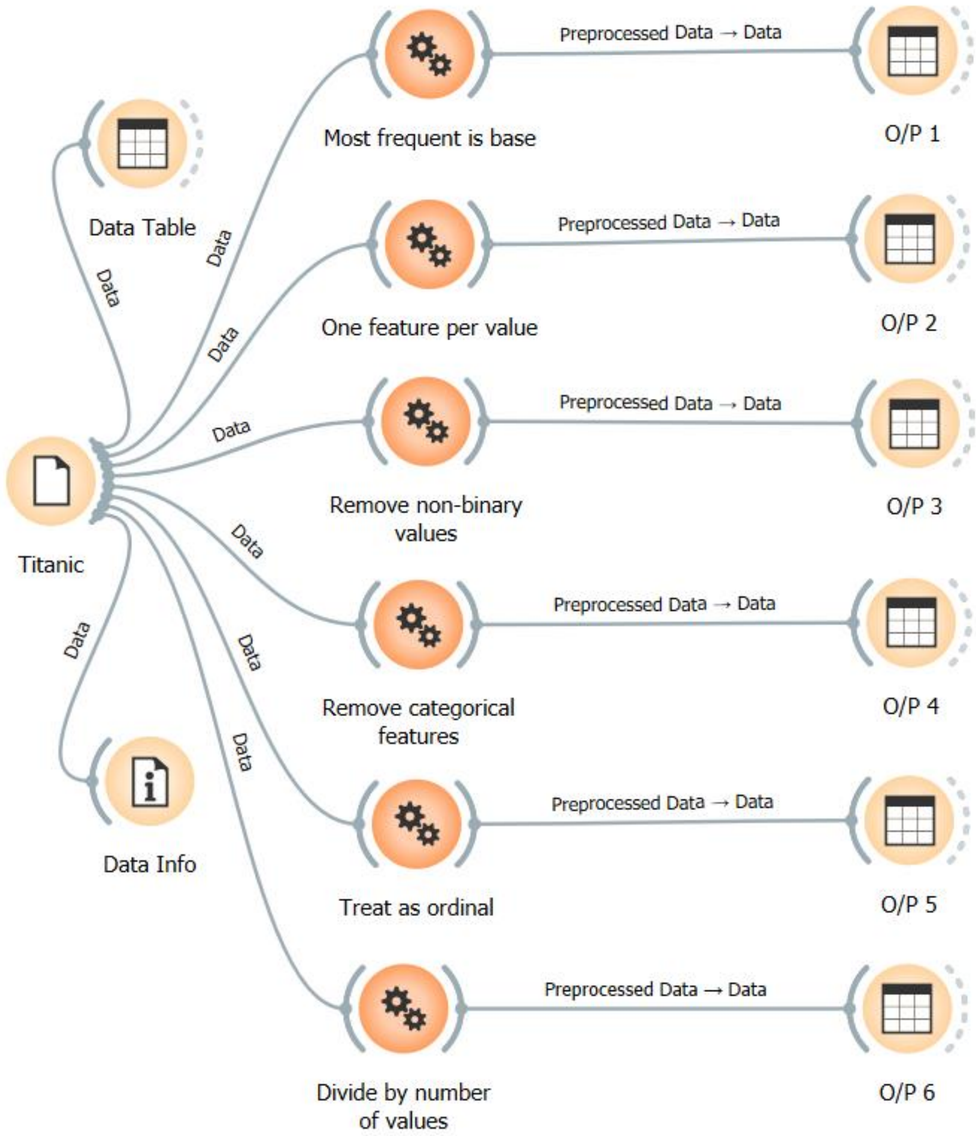
2201 instances (no missing data)
3 features
Target with 2 values
No meta attributes.

	survived	status	age	sex
1	yes	first	adult	male
2	yes	first	adult	male
3	yes	first	adult	male
4	yes	first	adult	male
5	yes	first	adult	male

Process

Continuization

- Most frequent is base.
- One feature per value.
- Remove non-binary values.
- Remove categorical features.
- Treat as ordinal.
- Divide by number of values.



Output

Output 1 (Most frequent is base)

	survived	status=first	status=second	status=third	age=child	sex=female
1	yes	1	0	0	0	0
2	yes	1	0	0	0	0
3	yes	1	0	0	0	0
4	yes	1	0	0	0	0
5	yes	1	0	0	0	0

Output 2 (One feature per value)

	survived	status=crew	status=first	status=second	status=third	age=adult	age=child	sex=female	sex=male
1	yes	0	1	0	0	1	0	0	1
2	yes	0	1	0	0	1	0	0	1
3	yes	0	1	0	0	1	0	0	1
4	yes	0	1	0	0	1	0	0	1
5	yes	0	1	0	0	1	0	0	1

Output 3 (Remove non-binary values)

	survived	age=child	sex=male
1	yes	0	1
2	yes	0	1
3	yes	0	1
4	yes	0	1
5	yes	0	1

Output 4 (Remove categorical features)

	survived
1	yes
2	yes
3	yes
4	yes
5	yes

Output 5 (Treat as ordinal)

	survived	status	age	sex
1	yes	1	0	1
2	yes	1	0	1
3	yes	1	0	1
4	yes	1	0	1
5	yes	1	0	1

Output 6 (Divide by number of values)

	survived	status	age	sex
1	yes	0.333333	0	1
2	yes	0.333333	0	1
3	yes	0.333333	0	1
4	yes	0.333333	0	1
5	yes	0.333333	0	1

4. Perform normalization on Iris dataset

Input

Data Info (1) - Orange ?

Data table properties

Name: iris

Size: 150 rows, 5 columns

Features: 4 numeric

Targets: categorical outcome with 3 classes

Additional attributes

Name: Iris flower dataset

Description: Classical dataset with 150 instances of Iris setosa, Iris virginica and Iris versicolor.

Author: Edgar Anderson, Ronald Fisher

Year: 1936

Reference: R. A. Fisher (1936). "The use of multiple measurements in taxonomic problems". Annals of Eugenics. 7 (2): 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x

150

Info

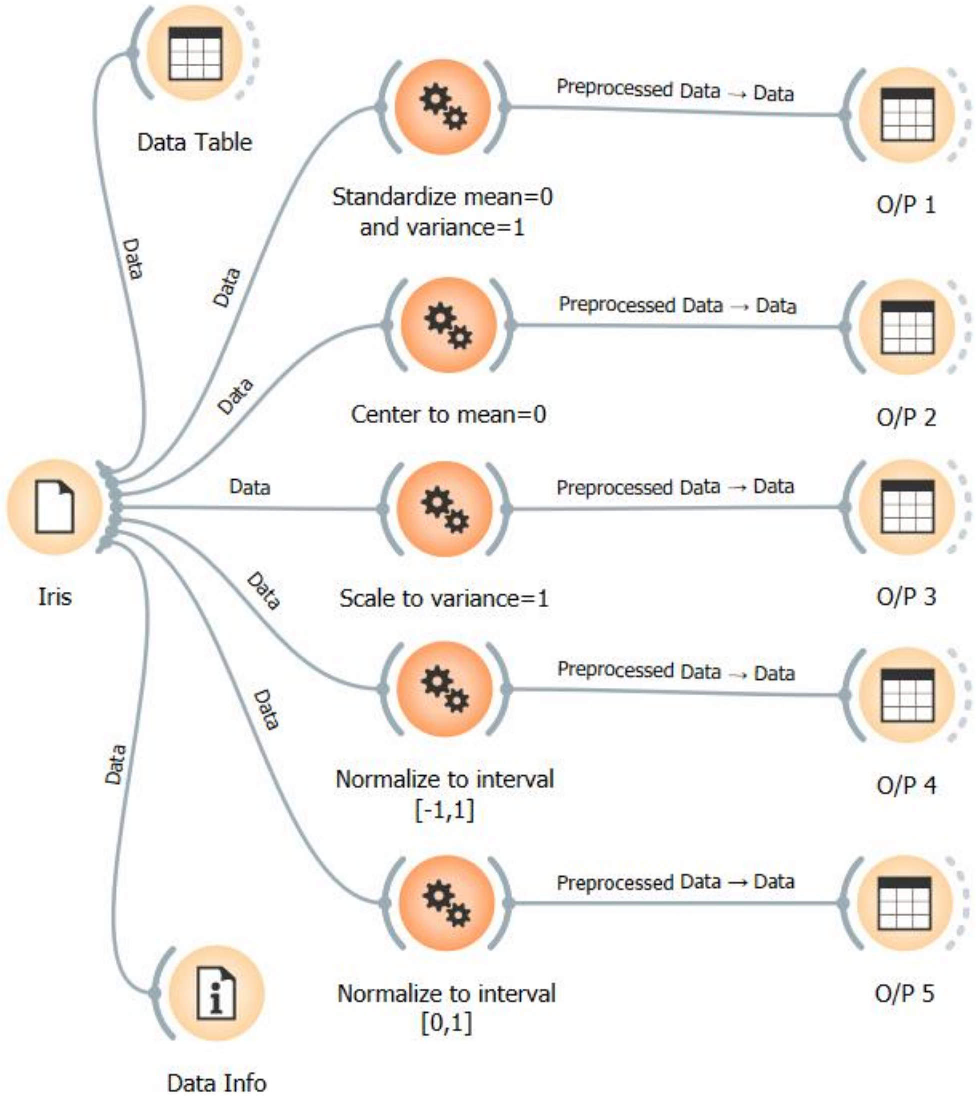
150 instances (no missing data)
4 features
Target with 3 values
No meta attributes.

	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	5.1	3.5	1.4	0.2
2	Iris-setosa	4.9	3.0	1.4	0.2
3	Iris-setosa	4.7	3.2	1.3	0.2
4	Iris-setosa	4.6	3.1	1.5	0.2
5	Iris-setosa	5.0	3.6	1.4	0.2

Process

Normalization

- Standardize μ=0 and variance = 1.
- Center to $\mu = 0$.
- Scale to variance = 1.
- Normalize to interval [-1,1].
- Normalize to interval [0,1].



Output

Output 1 (Standardize μ=0 and variance = 1)

	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	-0.901	1.032	-1.341	-1.313
2	Iris-setosa	-1.143	-0.125	-1.341	-1.313
3	Iris-setosa	-1.385	0.338	-1.398	-1.313
4	Iris-setosa	-1.507	0.106	-1.284	-1.313
5	Iris-setosa	-1.022	1.263	-1.341	-1.313

Output 2 (Center to μ=0.)

	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	-0.743	0.446	-2.359	-0.999
2	Iris-setosa	-0.943	-0.054	-2.359	-0.999
3	Iris-setosa	-1.143	0.146	-2.459	-0.999
4	Iris-setosa	-1.243	0.046	-2.259	-0.999
5	Iris-setosa	-0.843	0.546	-2.359	-0.999

Output 3 (Scale to variance = 1)

	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	6.180	8.099	0.796	0.263
2	Iris-setosa	5.937	6.942	0.796	0.263
3	Iris-setosa	5.695	7.405	0.739	0.263
4	Iris-setosa	5.574	7.173	0.853	0.263
5	Iris-setosa	6.058	8.331	0.796	0.263

Output 4 (Normalize to interval [-1,1])

	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	-0.5556	0.25	-0.8644	-0.9167
2	Iris-setosa	-0.6667	-0.1667	-0.8644	-0.9167
3	Iris-setosa	-0.7778	0.00	-0.8983	-0.9167
4	Iris-setosa	-0.8333	-0.0833	-0.8305	-0.9167
5	Iris-setosa	-0.6111	0.3333	-0.8644	-0.9167

Output 5 (Normalize to interval [0,1])

	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	0.2222	0.6250	0.0678	0.0417
2	Iris-setosa	0.1667	0.4167	0.0678	0.0417
3	Iris-setosa	0.1111	0.50	0.0508	0.0417
4	Iris-setosa	0.0833	0.4583	0.0847	0.0417
5	Iris-setosa	0.1944	0.6667	0.0678	0.0417

5) Perform Randomization on Iris dataset.

Input

Data Info (1) - Orange

Data table properties

Name: iris

Size: 150 rows, 5 columns

Features: 4 numeric

Targets: categorical outcome with 3 classes

Additional attributes

Name: Iris flower dataset

Description: Classical dataset with 150 instances of Iris setosa, Iris virginica and Iris versicolor.

Author: Edgar Anderson, Ronald Fisher

Year: 1936

Reference: R. A. Fisher (1936). "The use of multiple measurements in taxonomic problems". Annals of Eugenics. 7 (2): 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x

150

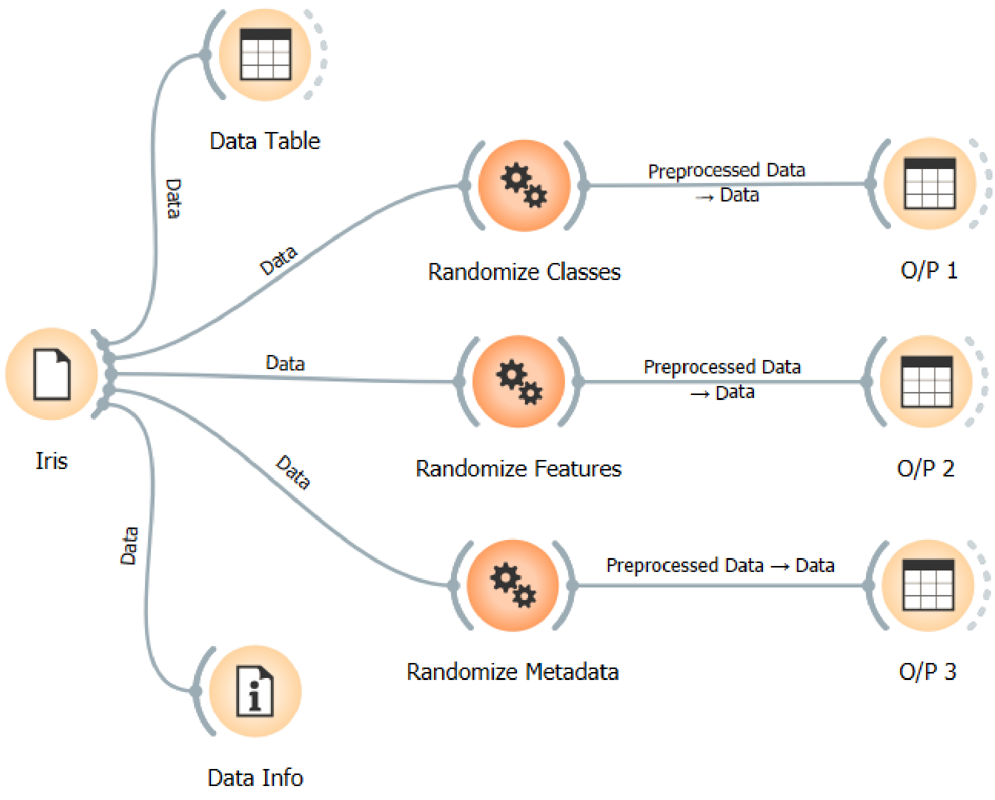
Info

150 instances (no missing data)
4 features
Target with 3 values
No meta attributes.

	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	5.1	3.5	1.4	0.2
2	Iris-setosa	4.9	3.0	1.4	0.2
3	Iris-setosa	4.7	3.2	1.3	0.2
4	Iris-setosa	4.6	3.1	1.5	0.2
5	Iris-setosa	5.0	3.6	1.4	0.2

Process

Randomization



Output

Output 1 (Randomize Classes)

	iris	sepal length	sepal width	petal length	petal width
1	Iris-versicolor	5.1	3.5	1.4	0.2
2	Iris-versicolor	4.9	3.0	1.4	0.2
3	Iris-setosa	4.7	3.2	1.3	0.2
4	Iris-setosa	4.6	3.1	1.5	0.2
5	Iris-setosa	5.0	3.6	1.4	0.2

Output 2 (Randomize Features)

	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	6.7	2.7	4.8	0.2
2	Iris-setosa	6.3	2.6	6.7	1.2
3	Iris-setosa	4.4	3.1	1.5	2.3
4	Iris-setosa	6.7	3.4	5.3	0.2
5	Iris-setosa	4.8	3.4	5.0	1.8

Output 3 (Randomize Metadata)

	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	5.1	3.5	1.4	0.2
2	Iris-setosa	4.9	3.0	1.4	0.2
3	Iris-setosa	4.7	3.2	1.3	0.2
4	Iris-setosa	4.6	3.1	1.5	0.2
5	Iris-setosa	5.0	3.6	1.4	0.2

6) Perform Remove Sparse on zoo data set

Input

Data Info - Orange ? ×

Data table properties

Name: zoo

Size: 101 rows, 18 columns

Features: 16 categorical

Targets: categorical outcome with 7 classes

Metas: 1 text

Additional attributes

Name: Zoo dataset

Description: This dataset consists of 101 animals with various traits to describe them.

Author: Richard Forsyth

Year: 1990

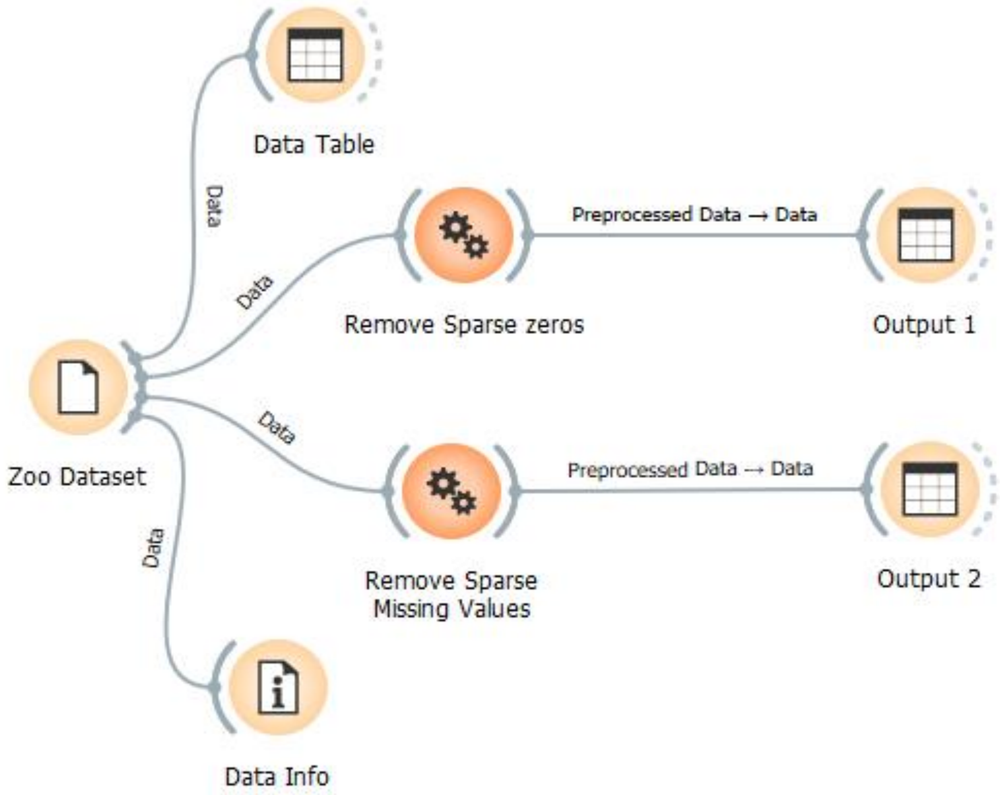
101

Info

101 instances (no missing data)
16 features
Target with 7 values
1 meta attribute

	type	name	hair	feathers	eggs	milk
1	mammal	aardvark	1	0	0	1
2	mammal	antelope	1	0	0	1
3	fish	bass	0	0	1	0
4	mammal	bear	1	0	0	1
5	mammal	boar	1	0	0	1

Process



Output

Output 1 (Remove sparse zeros)

	type	name	eggs	predator	toothed	backbone	breathes	legs	tail
1	mammal	aardvark	0	1	1	1	1	4	0
2	mammal	antelope	0	0	1	1	1	4	1
3	fish	bass	1	1	1	1	0	0	1
4	mammal	bear	0	1	1	1	1	4	0
5	mammal	boar	0	1	1	1	1	4	1

Output 2 (Remove sparse missing values)

	type	name	eggs	predator	toothed	backbone	breathes	legs	tail
1	mammal	aardvark	0	1	1	1	1	4	0
2	mammal	antelope	0	0	1	1	1	4	1
3	fish	bass	1	1	1	1	0	0	1
4	mammal	bear	0	1	1	1	1	4	0
5	mammal	boar	0	1	1	1	1	4	1

7) Perform Feature Selection on Wine dataset.

Input

Data Info - Orange ? ×

Data table properties

Name: wine

Size: 178 rows, 14 columns

Features: 13 numeric

Targets: categorical outcome with 3 classes

Additional attributes

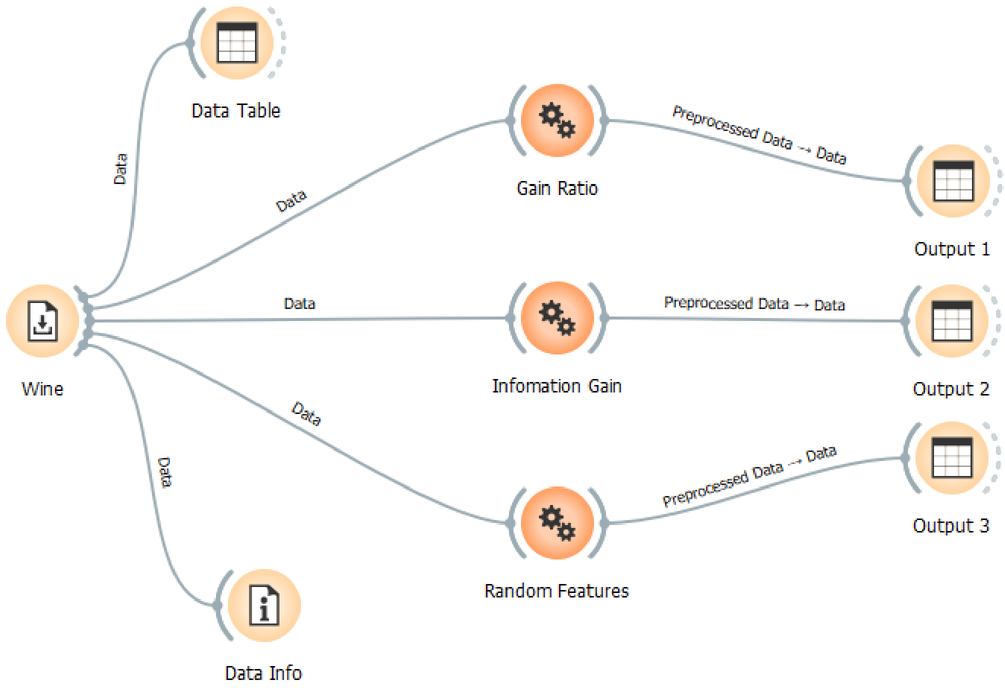
178

Info

178 instances (no missing data)
13 features
Target with 3 values
No meta attributes.

	Wine	Alcohol	Malic Acid	Ash	Alcalinity of ash	Magnesium
1	1	14.23	1.71	2.43	15.6	127
2	1	13.20	1.78	2.14	11.2	100
3	1	13.16	2.36	2.67	18.6	101
4	1	14.37	1.95	2.50	16.8	113
5	1	13.24	2.59	2.87	21.0	118

Process



Output

Output 1 (Gain Ratio)

	Wine	Flavanoids	Proline	Color intensity	D/OD315 of diluted	Alcohol
1	1	3.06	1065	5.64	3.92	14.23
2	1	2.76	1050	4.38	3.40	13.20
3	1	3.24	1185	5.68	3.17	13.16
4	1	3.49	1480	7.8	3.45	14.37
5	1	2.69	735	4.32	2.93	13.24

Output 2 (Information Gain)

	Wine	Flavanoids	Proline	Color intensity	D/OD315 of diluted	Alcohol
1	1	3.06	1065	5.64	3.92	14.23
2	1	2.76	1050	4.38	3.40	13.20
3	1	3.24	1185	5.68	3.17	13.16
4	1	3.49	1480	7.8	3.45	14.37
5	1	2.69	735	4.32	2.93	13.24

Output 3 (Random Features)

	Wine	Total phenols	Color intensity	Proline	Proanthocyanins	D/OD315 of diluted
1	1	2.80	5.64	1065	2.29	3.92
2	1	2.65	4.38	1050	1.28	3.40
3	1	2.80	5.68	1185	2.81	3.17
4	1	3.85	7.8	1480	2.18	3.45
5	1	2.80	4.32	735	1.82	2.93

8) Perform Feature Selection on Lenses dataset

Input

Data Info - Orange ? ×

Data table properties

Name: lenses

Size: 24 rows, 5 columns

Features: 4 categorical

Targets: categorical outcome with 3 classes

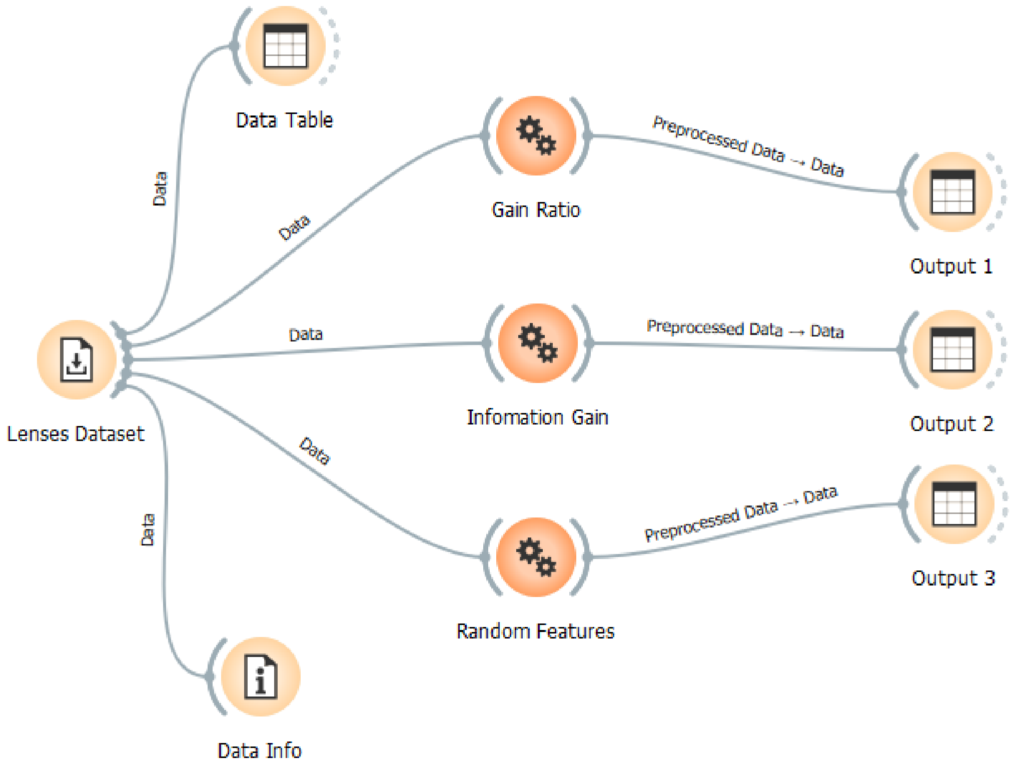
Additional attributes

24

Info

24 instances (no missing data)
4 features
Target with 3 values
No meta attributes.

	lenses	age	prescription	astigmatic	tear_rate
1	none	young	myope	no	reduced
2	soft	young	myope	no	normal
3	none	young	myope	yes	reduced
4	hard	young	myope	yes	normal
5	none	young	hypermetrope	no	reduced



Output

Output 1 (Gain Ratio)

	lenses	tear_rate	astigmatic	prescription	age
1	none	reduced	no	myope	young
2	soft	normal	no	myope	young
3	none	reduced	yes	myope	young
4	hard	normal	yes	myope	young
5	none	reduced	no	hypermetrope	young

Output 2 (Information Gain)

	lenses	tear_rate	astigmatic	prescription	age
1	none	reduced	no	myope	young
2	soft	normal	no	myope	young
3	none	reduced	yes	myope	young
4	hard	normal	yes	myope	young
5	none	reduced	no	hypermetrope	young

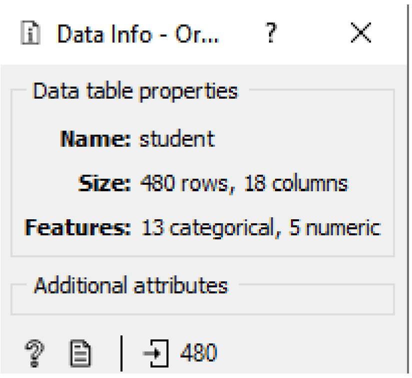
Output 3 (Random Features)

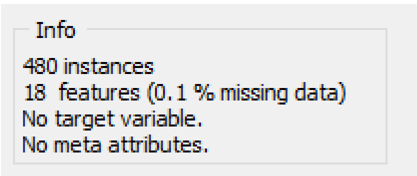
	lenses	astigmatic	tear_rate	prescription	age
1	none	no	reduced	myope	young
2	soft	no	normal	myope	young
3	none	yes	reduced	myope	young
4	hard	yes	normal	myope	young
5	none	no	reduced	hypermetrope	young

b. Preprocessing (Orange Tool) Exercises

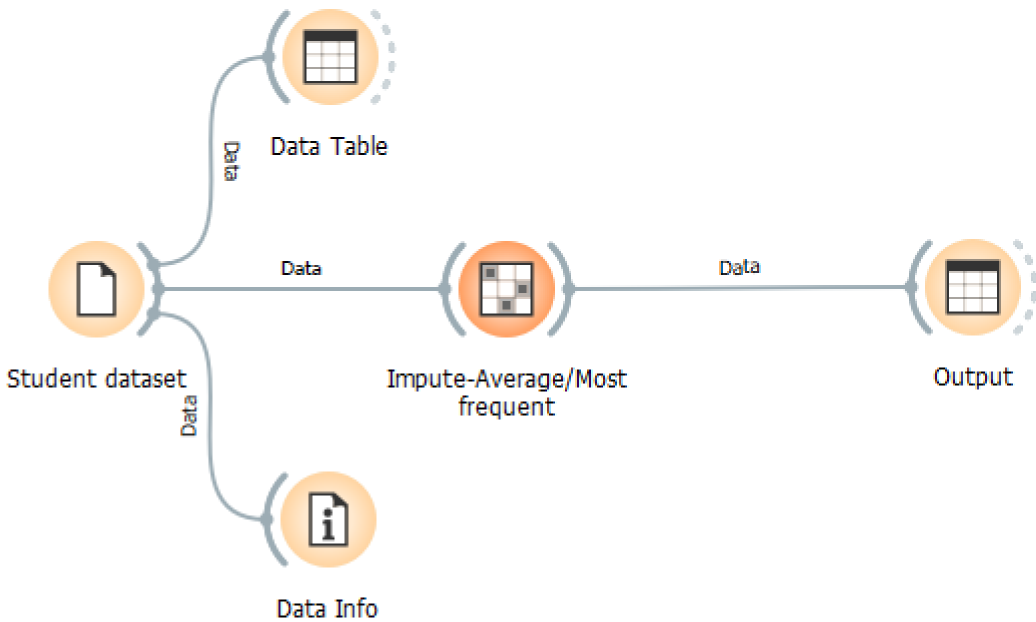
1) Replace missing values by the mean of the values of records having same class value. Display the entire data after replacement.

Input





Process



Output

Info
480 instances (no missing data)
18 features
No target variable.
No meta attributes.

2) Perform binning(3 bins) for the attribute AnnouncementsView.

Input

Data Info - Or... ? ×

Data table properties

Name: student

Size: 480 rows, 18 columns

Features: 13 categorical, 5 numeric

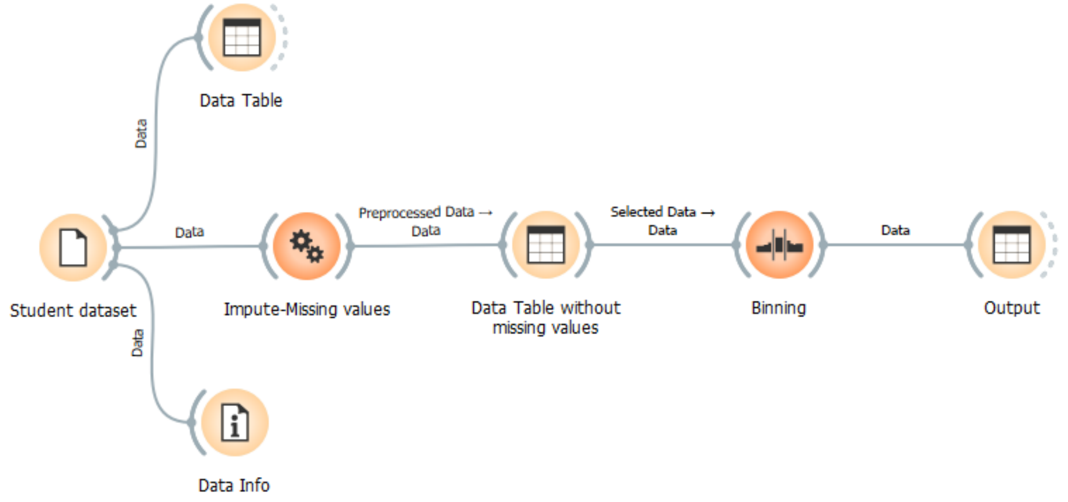
Additional attributes

480

Info
480 instances
18 features (0.1 % missing data)
No target variable.
No meta attributes.

	Gender	nnouncementsView	NationalITy	PlaceofBirth	StageID	GradeID	SectionID
1	M	2	KW	KuwaIT	lowerlevel	G-04	A
2	M	3	KW	KuwaIT	lowerlevel	G-04	A
3	M	0	KW	KuwaIT	lowerlevel	G-04	A
4	M	5	KW	KuwaIT	lowerlevel	G-04	A
5	M	12	KW	KuwaIT	lowerlevel	G-04	A
6	F	13	KW	KuwaIT	lowerlevel	G-04	A
7	M	?	KW	KuwaIT	MiddleSchool	G-07	A

Process



Output

Info
480 instances (no missing data)
18 features
No target variable.
No meta attributes.

	Gender	nnouncementsView	NationalITy	PlaceofBirth	StageID	GradeID	SectionID
1	M	2	KW	KuwaIT	lowerlevel	G-04	A
2	M	3	KW	KuwaIT	lowerlevel	G-04	A
3	M	0	KW	KuwaIT	lowerlevel	G-04	A
4	M	5	KW	KuwaIT	lowerlevel	G-04	A
5	M	12	KW	KuwaIT	lowerlevel	G-04	A
6	F	13	KW	KuwaIT	lowerlevel	G-04	A
7	M	38.03	KW	KuwaIT	MiddleSchool	G-07	A

	Gender	nnouncementsView	NationalITy	PlaceofBirth	StageID	GradeID	SectionID
1	M	< 20	KW	KuwaIT	lowerlevel	G-04	A
2	M	< 20	KW	KuwaIT	lowerlevel	G-04	A
3	M	< 20	KW	KuwaIT	lowerlevel	G-04	A
4	M	< 20	KW	KuwaIT	lowerlevel	G-04	A
5	M	< 20	KW	KuwaIT	lowerlevel	G-04	A
6	F	< 20	KW	KuwaIT	lowerlevel	G-04	A
7	M	20 - 40	KW	KuwaIT	MiddleSchool	G-07	A

3) Remove redundant variables/features having high correlation.

Input

Data Info - Or... ? ×

Data table properties

Name: student

Size: 480 rows, 18 columns

Features: 13 categorical, 5 numeric

Additional attributes

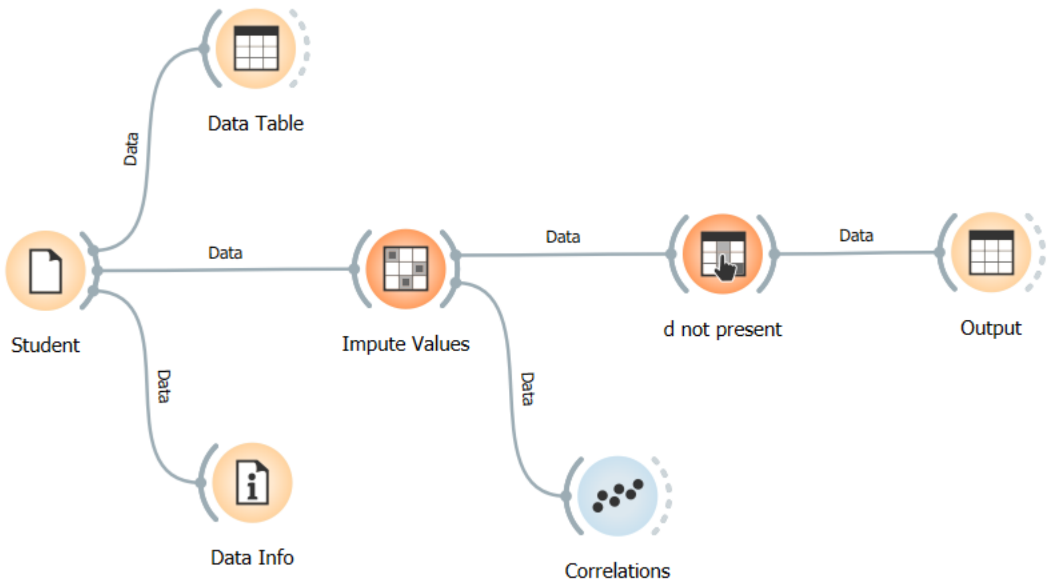
480

Info

480 instances
18 features (0.1 % missing data)
No target variable.
No meta attributes.

	Gender	NationalITy	d
1	M	KW	20
2	M	KW	25
3	M	KW	30
4	M	KW	35
5	M	KW	50

Process



Output

Correlations - Orange

Pearson correlation

(All combinations)

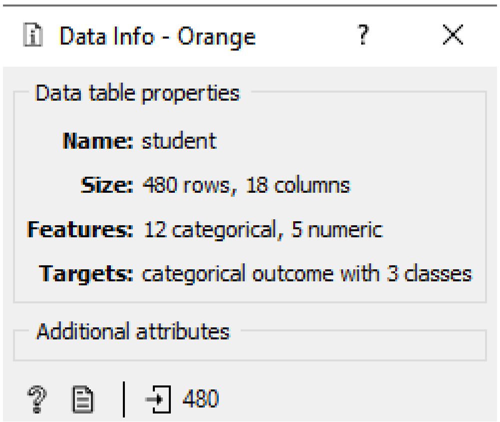
Filter ...

1	+1.000	Discussion	d
2	+0.692	VisITedResources	raisedhands
3	+0.643	AnnouncementsView	raisedhands
4	+0.590	AnnouncementsView	VisITedResources
5	+0.414	AnnouncementsView	d
6	+0.414	AnnouncementsView	Discussion
7	+0.339	d	raisedhands
8	+0.339	Discussion	raisedhands
9	+0.243	VisITedResources	d
10	+0.243	Discussion	VisITedResources

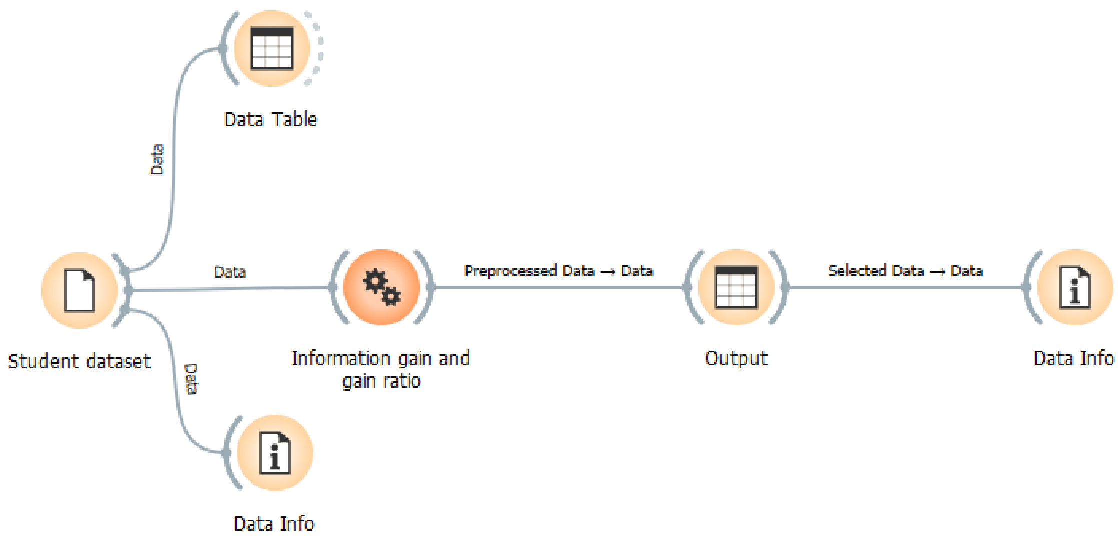
	Gender	NationalITy	PlaceofBirth
1	M	KW	KuwaIT
2	M	KW	KuwaIT
3	M	KW	KuwaIT
4	M	KW	KuwaIT
5	M	KW	KuwaIT

4) Select important variables/features using Information gain and gain ratio.

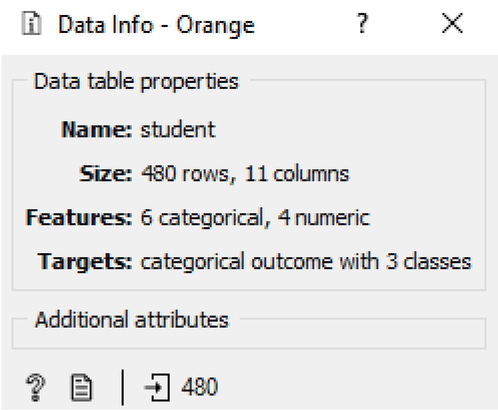
Input



Process

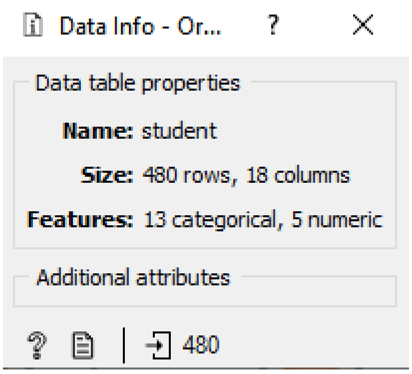


Output



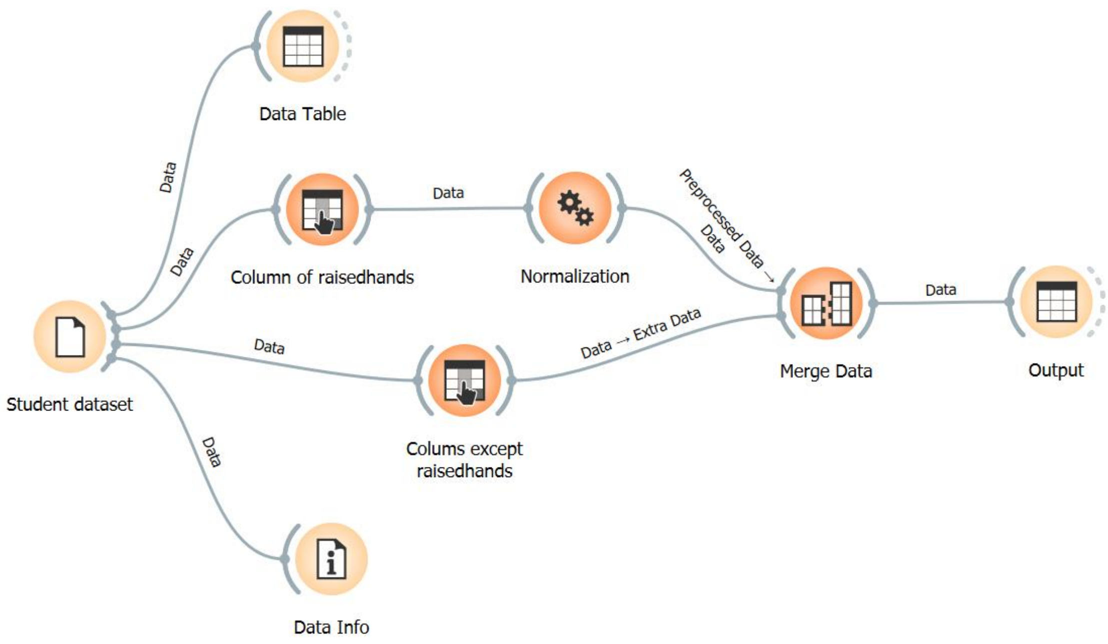
5) Perform normalization [-1,1] on the attribute raisedhands.

Input



	raisedhands	Gender	NationalITy	PlaceofBirth	StageID	GradeID
1	15	M	KW	KuwaIT	lowerlevel	G-04
2	20	M	KW	KuwaIT	lowerlevel	G-04
3	10	M	KW	KuwaIT	lowerlevel	G-04
4	30	M	KW	KuwaIT	lowerlevel	G-04
5	40	M	KW	KuwaIT	lowerlevel	G-04

Process

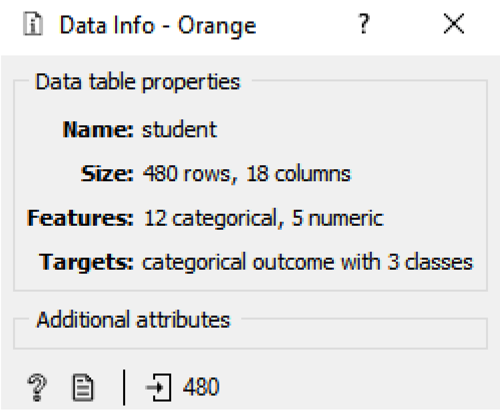


Output

	raisedhands	Gender	NationalITy	PlaceofBirth	StageID	GradeID
1	-0.70	M	KW	KuwaIT	lowerlevel	G-04
2	-0.60	M	KW	KuwaIT	lowerlevel	G-04
3	-0.80	M	KW	KuwaIT	lowerlevel	G-04
4	-0.40	M	KW	KuwaIT	lowerlevel	G-04
5	-0.20	M	KW	KuwaIT	lowerlevel	G-04

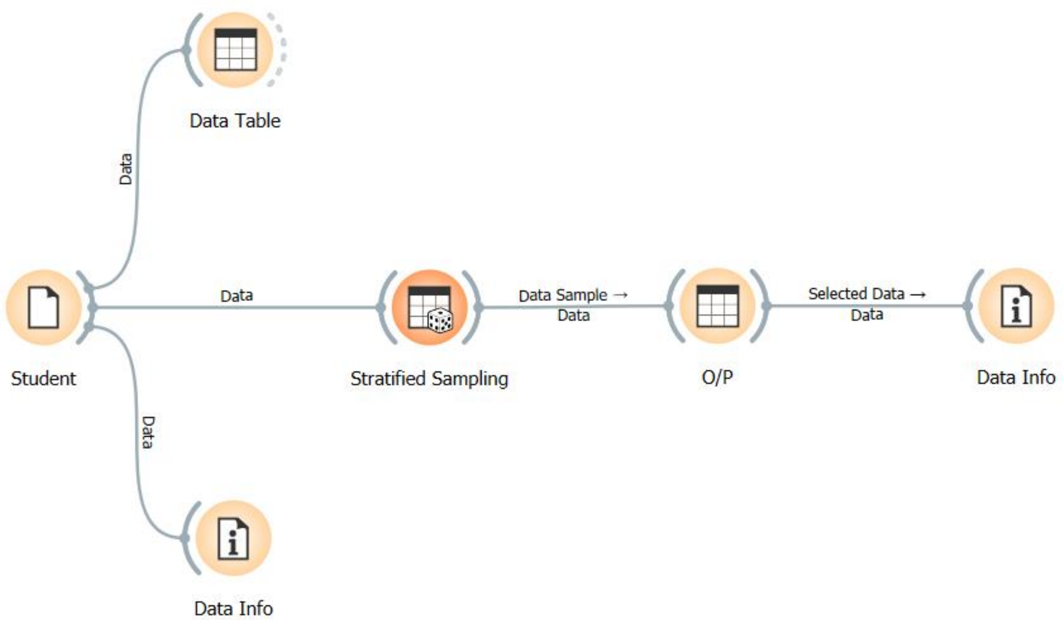
6) Do a stratified random sampling to draw a sample size of approximately 100 out of the total records.

Input

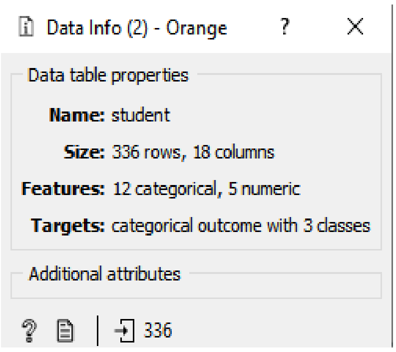


	Class	Gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID
1	M	M	KW	KuwaIT	lowerlevel	G-04	A
2	M	M	KW	KuwaIT	lowerlevel	G-04	A
3	L	M	KW	KuwaIT	lowerlevel	G-04	A
4	L	M	KW	KuwaIT	lowerlevel	G-04	A
5	M	M	KW	KuwaIT	lowerlevel	G-04	A

Process



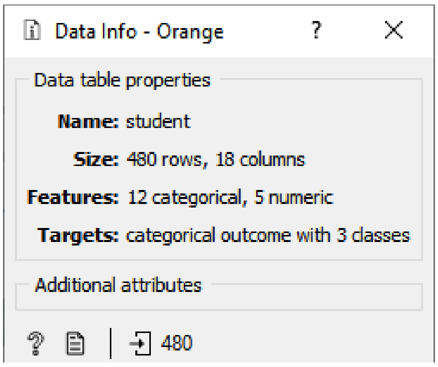
Output



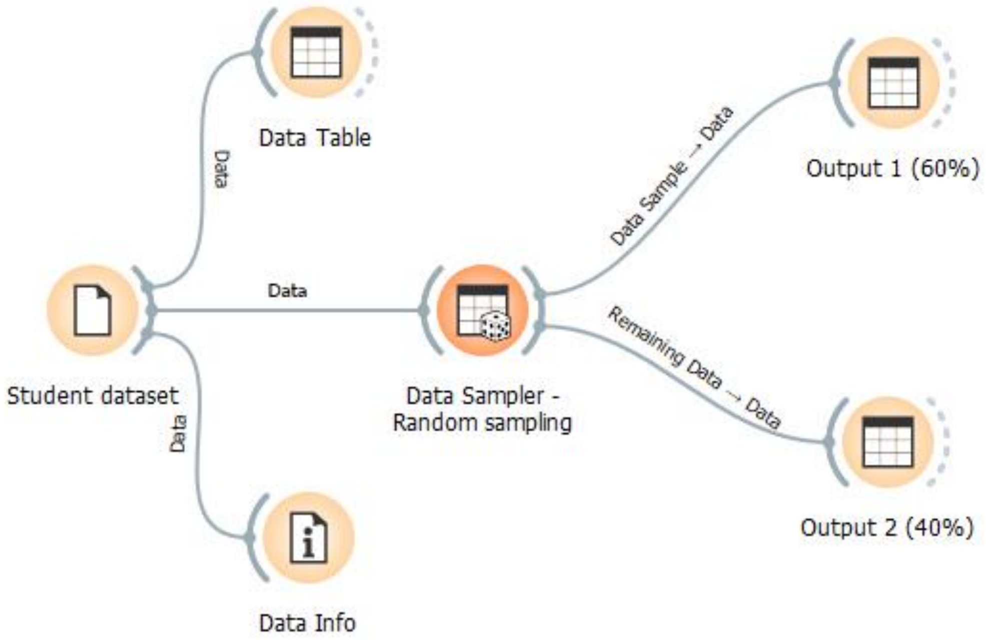
	Class	Gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID
1	H	F	Jordan	Jordan	MiddleSchool	G-07	B
2	M	M	Jordan	Jordan	MiddleSchool	G-07	B
3	L	M	KW	KuwaIT	lowerlevel	G-02	B
4	L	M	Jordan	Jordan	MiddleSchool	G-07	A
5	M	F	Jordan	Jordan	MiddleSchool	G-06	A

7) Partition the data into 2 data sets (60:40) using random partitioning.

Input



Process



Output

Info

288 instances
17 features (0.1 % missing data)
Target with 3 values
No meta attributes.

	Class	Gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID
1	M	M	Jordan	Jordan	lowerlevel	G-02	A
2	M	M	Iraq	Iraq	MiddleSchool	G-08	A
3	M	M	Jordan	Jordan	MiddleSchool	G-08	A
4	M	M	Jordan	Jordan	MiddleSchool	G-08	A
5	H	F	Jordan	Jordan	MiddleSchool	G-08	A

Info

192 instances
17 features (0.1 % missing data)
Target with 3 values
No meta attributes.

	Class	Gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID
1	M	M	KW	KuwaIT	MiddleSchool	G-07	B
2	H	F	lebanon	lebanon	lowerlevel	G-02	B
3	H	M	Palestine	Jordan	lowerlevel	G-02	A
4	L	M	Jordan	Jordan	lowerlevel	G-02	A
5	L	M	KW	KuwaIT	lowerlevel	G-02	B

Use mtcars data set to

1) Replace the missing data with the average/median of the feature wt.

Input

Data Info - ... ?

Data table properties

Name: mtcars

Size: 32 rows, 11 columns

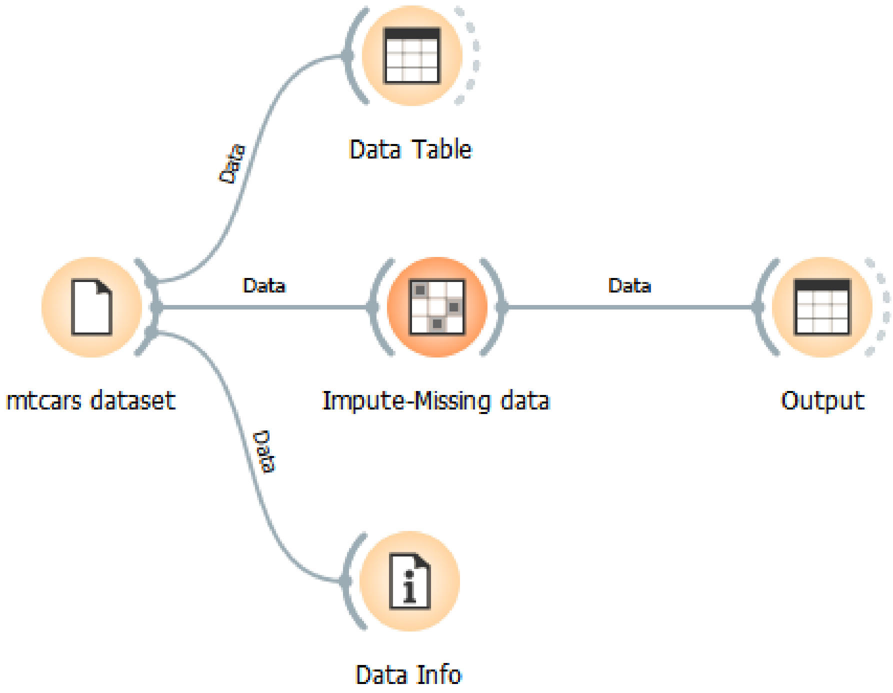
Features: 2 categorical, 9 numeric

Additional attributes

32

	mpg	cyl	disp	hp	drat	wt
1	21.0	6	160.0	110	3.90	2.620
2	21.0	6	160.0	110	3.90	2.875
3	22.8	4	108.0	93	3.85	2.320
4	21.4	6	258.0	110	3.08	3.215
5	18.7	8	360.0	175	3.15	3.440

Process



Output

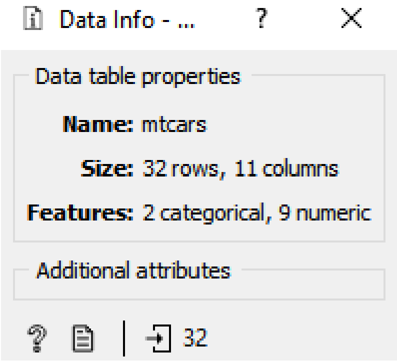
Info

32 instances (no missing data)
11 features
No target variable.
No meta attributes.

	mpg	cyl	disp	hp	drat	wt
1	21.0	6	160.0	110	3.90	2.620
2	21.0	6	160.0	110	3.90	2.875
3	22.8	4	108.0	93	3.85	2.320
4	21.4	6	258.0	110	3.08	3.215
5	18.7	8	360.0	175	3.15	3.440

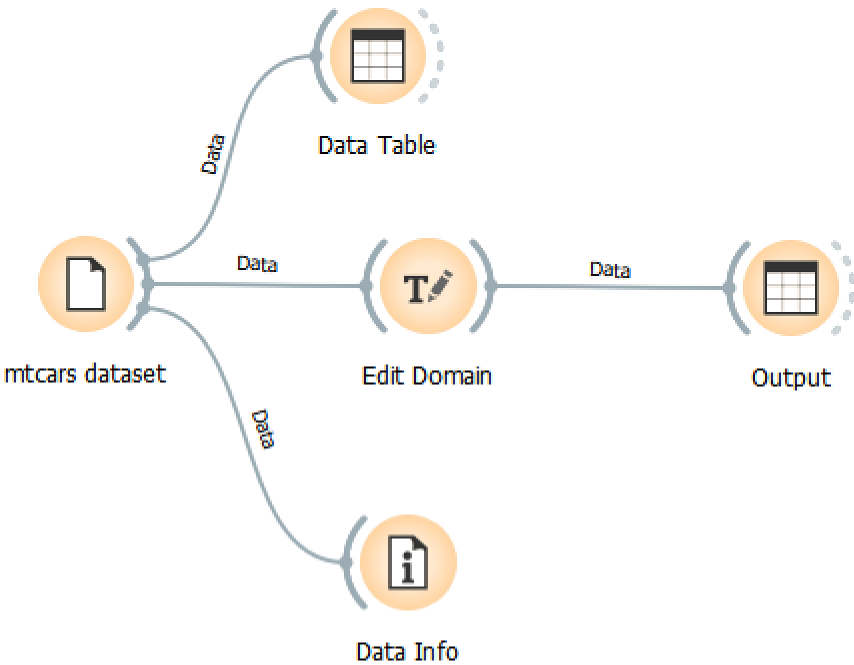
2) Transform the numerical variable am to manual-0 and automatic-1.

Input



	am	mpg	cyl	disp	hp
1	1	21.0	6	160.0	110
2	1	21.0	6	160.0	110
3	1	22.8	4	108.0	93
4	0	21.4	6	258.0	110
5	0	18.7	8	360.0	175

Process

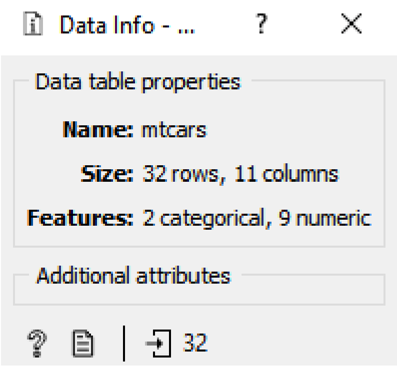


Output

	am	mpg	cyl	disp	hp
1	Automatic	21.0	6	160.0	110
2	Automatic	21.0	6	160.0	110
3	Automatic	22.8	4	108.0	93
4	Manual	21.4	6	258.0	110
5	Manual	18.7	8	360.0	175

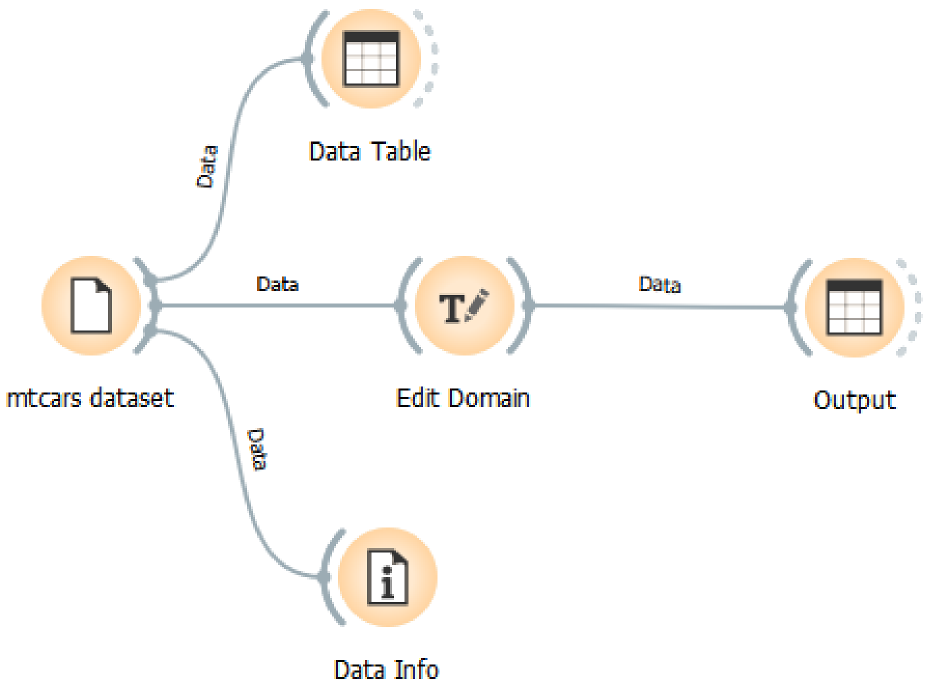
3) Transform the numerical variable gear by appending "gear" to the number of gears given in the feature.

Input



	gear	mpg	cyl	disp	hp
1	4	21.0	6	160.0	110
2	4	21.0	6	160.0	110
3	4	22.8	4	108.0	93
4	3	21.4	6	258.0	110
5	3	18.7	8	360.0	175

Process

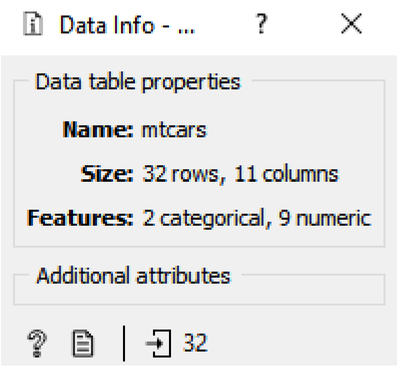


Output

	gear	mpg	cyl	disp	hp
1	Gear 4	21.0	6	160.0	110
2	Gear 4	21.0	6	160.0	110
3	Gear 4	22.8	4	108.0	93
4	Gear 3	21.4	6	258.0	110
5	Gear 3	18.7	8	360.0	175

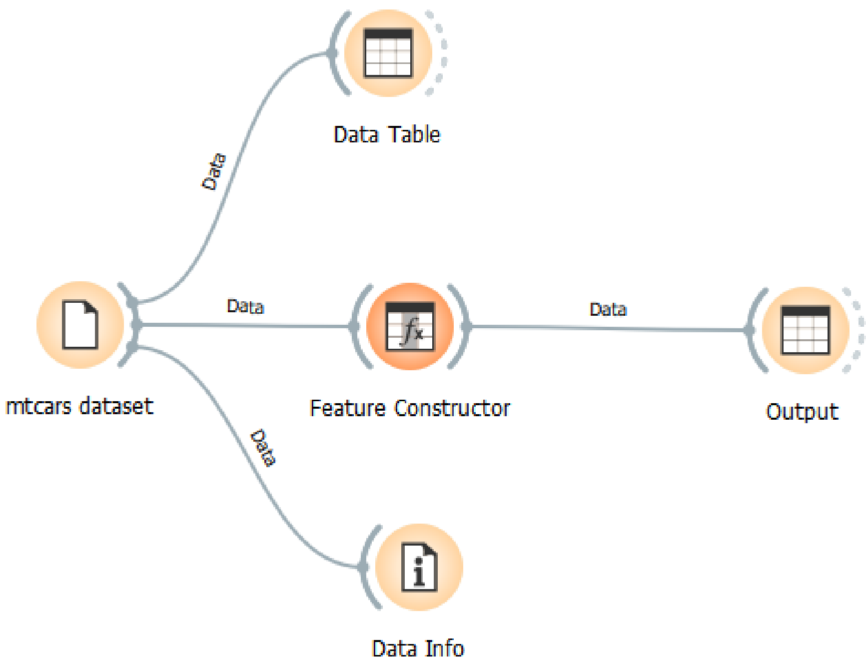
4) Add a new attribute Engine type based on the condition for the attribute vs (0 = V-shaped, 1 = straight).

Input



	mpg	cyl	disp	hp	drat
1	21.0	6	160.0	110	3.90
2	21.0	6	160.0	110	3.90
3	22.8	4	108.0	93	3.85
4	21.4	6	258.0	110	3.08
5	18.7	8	360.0	175	3.15

Process



Output

Info
32 instances (no missing data)
12 features
No target variable.
No meta attributes.

	Engine Type	mpg	cyl	disp	hp
1	V-Shaped	21.0	6	160.0	110
2	V-Shaped	21.0	6	160.0	110
3	Straight	22.8	4	108.0	93
4	Straight	21.4	6	258.0	110
5	V-Shaped	18.7	8	360.0	175

5) Scale the feature disp.

Input

Data Info - ... ?

Data table properties

Name: mtcars

Size: 32 rows, 11 columns

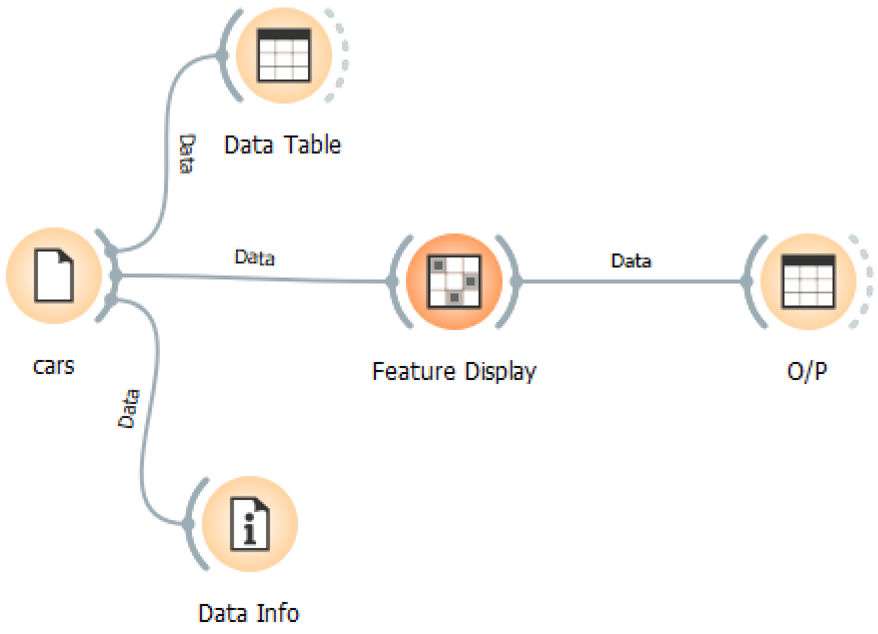
Features: 2 categorical, 9 numeric

Additional attributes

32

	mpg	cyl	disp	hp	drat
1	21.0	6	160.0	110	3.90
2	21.0	6	160.0	110	3.90
3	22.8	4	108.0	93	3.85
4	21.4	6	258.0	110	3.08
5	18.7	8	360.0	175	3.15

Process



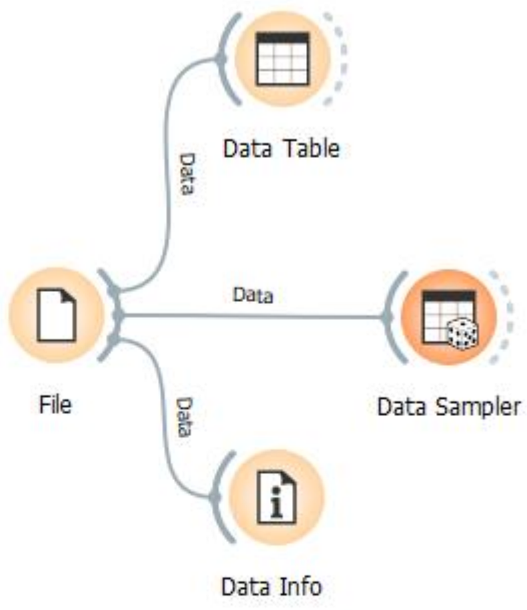
6) Split the dataset into 70%training data set and 30% test dataset

Info

32 instances (no missing data)
11 features
No target variable.
No meta attributes.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
1	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
2	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
3	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
4	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
5	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
6	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1

Process



Output

Info

23 instances (no missing data)
11 features
No target variable.
No meta attributes.

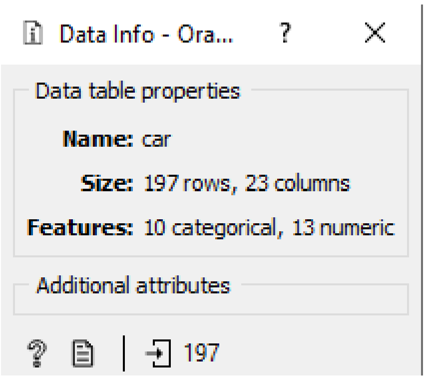
Section II

a) Data Visualization (Orange Tool) Class Work

Use car.csv data set to

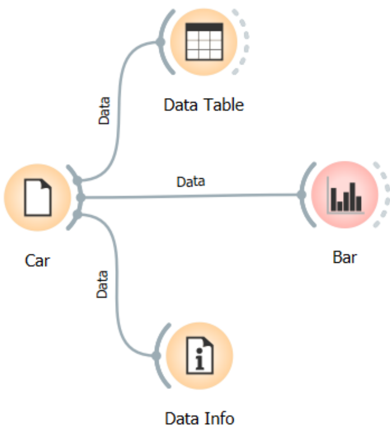
1) Plot a bar chart to compare the price of different makes of car.

Input

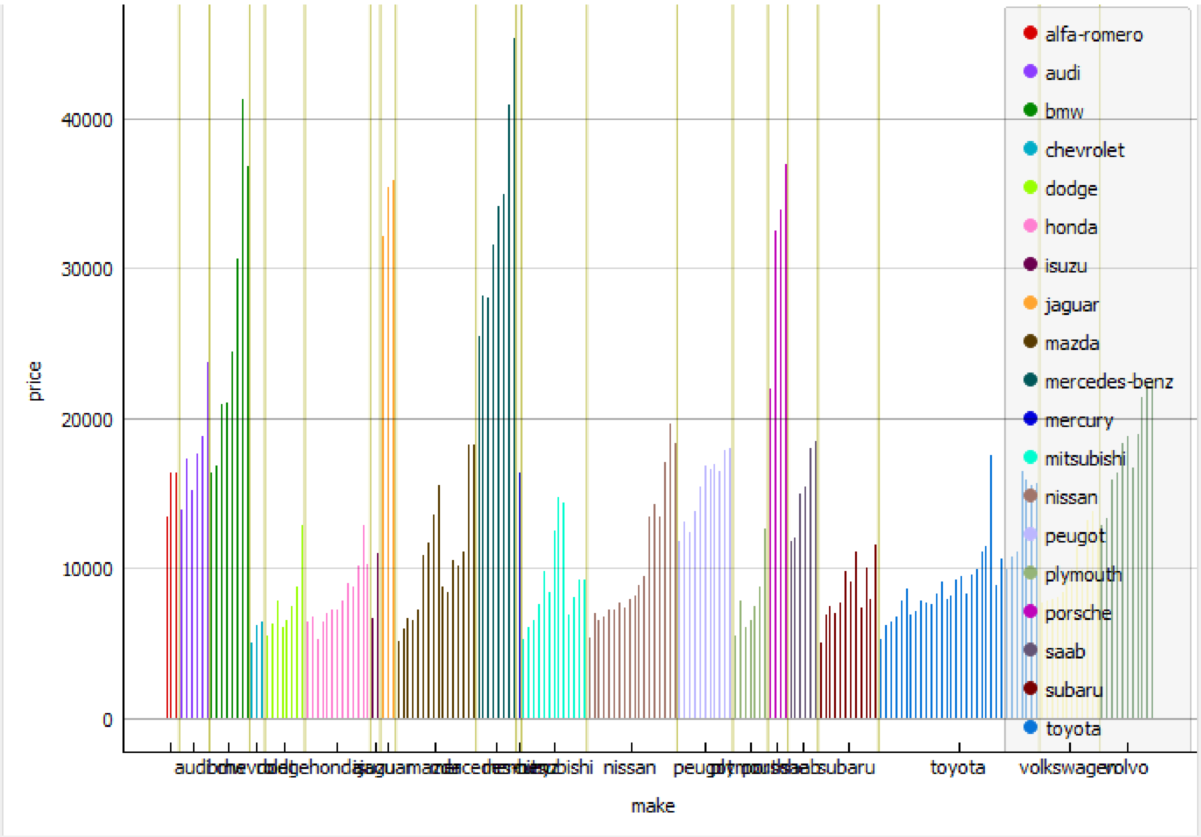


	Feature 1	make	fuel_type	aspiration	num_of_doors	body_style	drive_wheels
1	1	alfa-romero	gas	std	two	convertible	rwd
2	2	alfa-romero	gas	std	two	convertible	rwd
3	3	alfa-romero	gas	std	two	hatchback	rwd
4	4	audi	gas	std	four	sedan	fwd
5	5	audi	gas	std	four	sedan	4wd

Process



Output

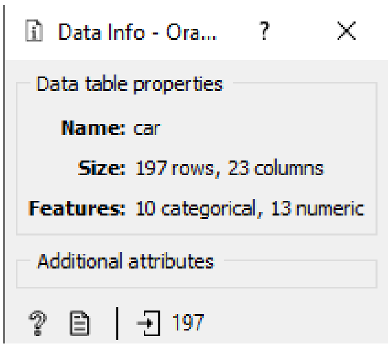


Interpretation

- Chevrolet is the cheapest car out of the lot
- Average price of the cars is around 18000

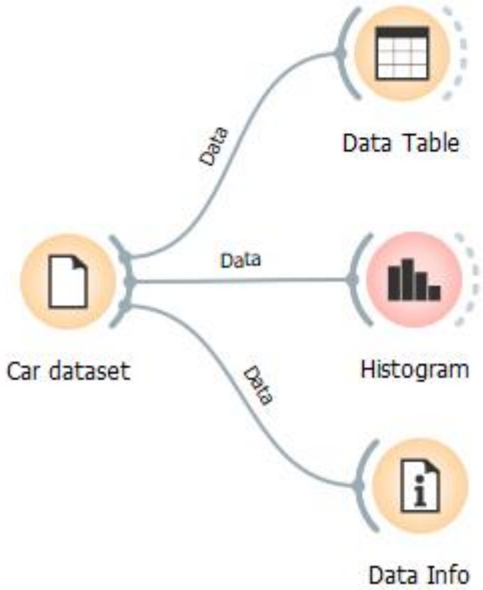
2) Create a histogram for analyzing city mileage.

Input

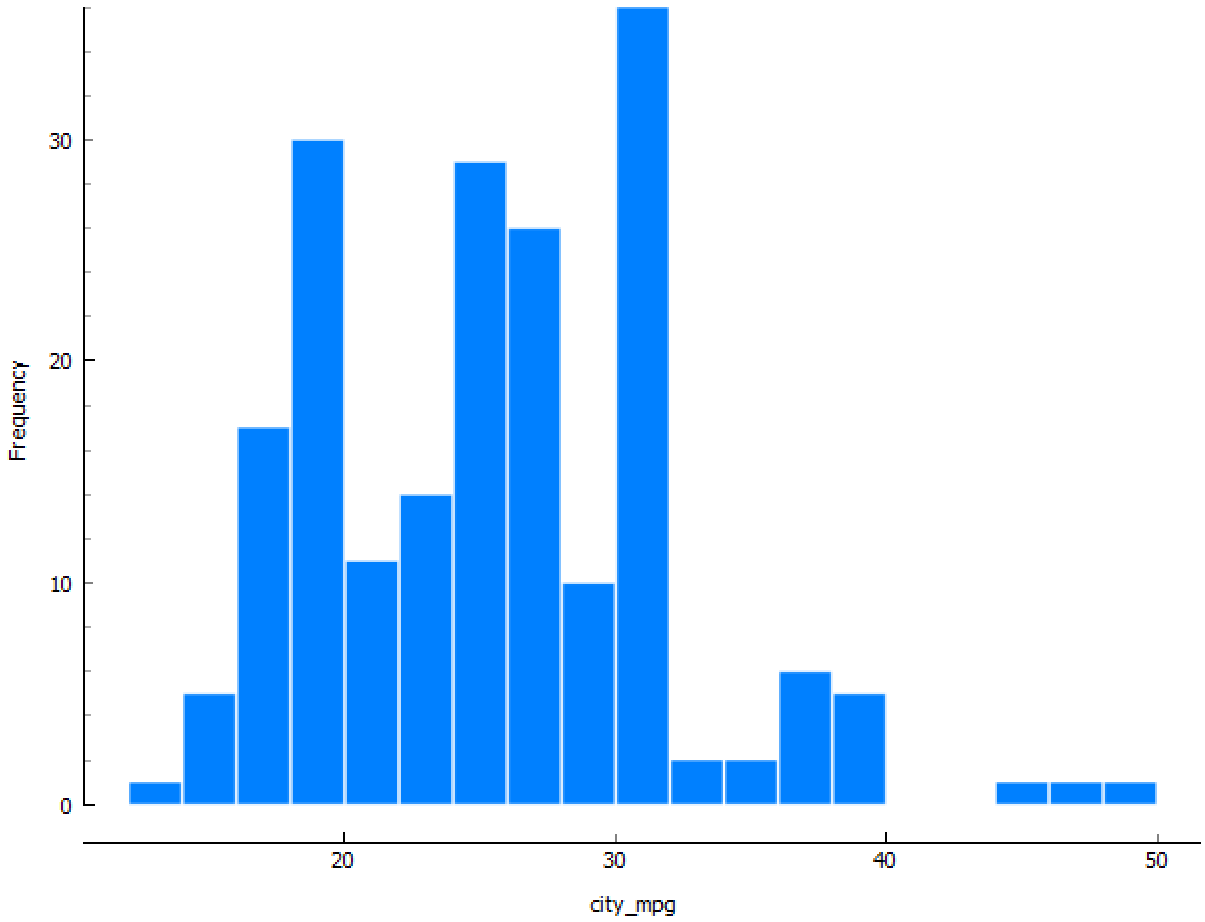


	Feature 1	make	fuel_type	aspiration	num_of_doors	body_style	drive_wheels
1	1	alfa-romero	gas	std	two	convertible	rwd
2	2	alfa-romero	gas	std	two	convertible	rwd
3	3	alfa-romero	gas	std	two	hatchback	rwd
4	4	audi	gas	std	four	sedan	fwd
5	5	audi	gas	std	four	sedan	4wd

Process



Output

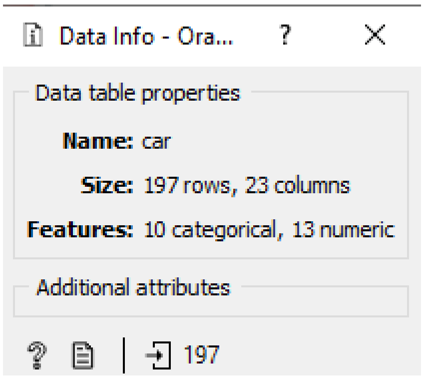


Interpretation

- Less frequency of cars having mileage > 40
- Average Mileage ranges from 28 to 32

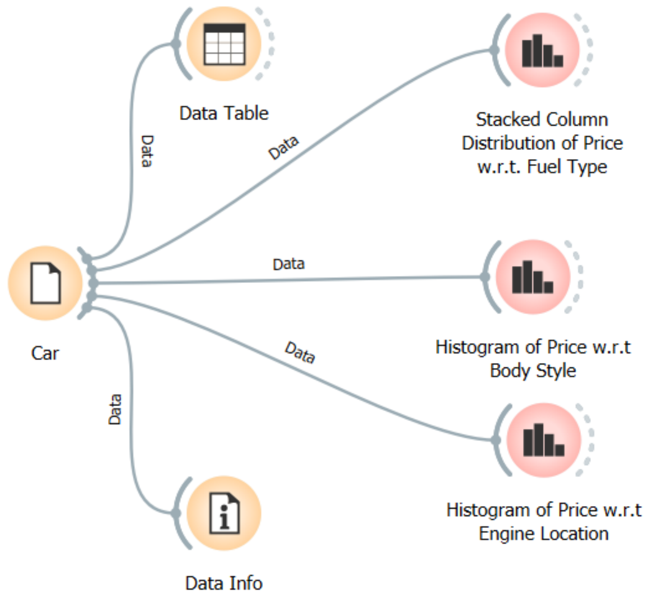
3) Create a histogram for analyzing price. Show a stacked column distribution with respect to fuel_type. Similarly create a histogram for price w.r.t body_style and price w.r.t engine_location. Write your inferences for price of cars w.r.t the above variables.

Input



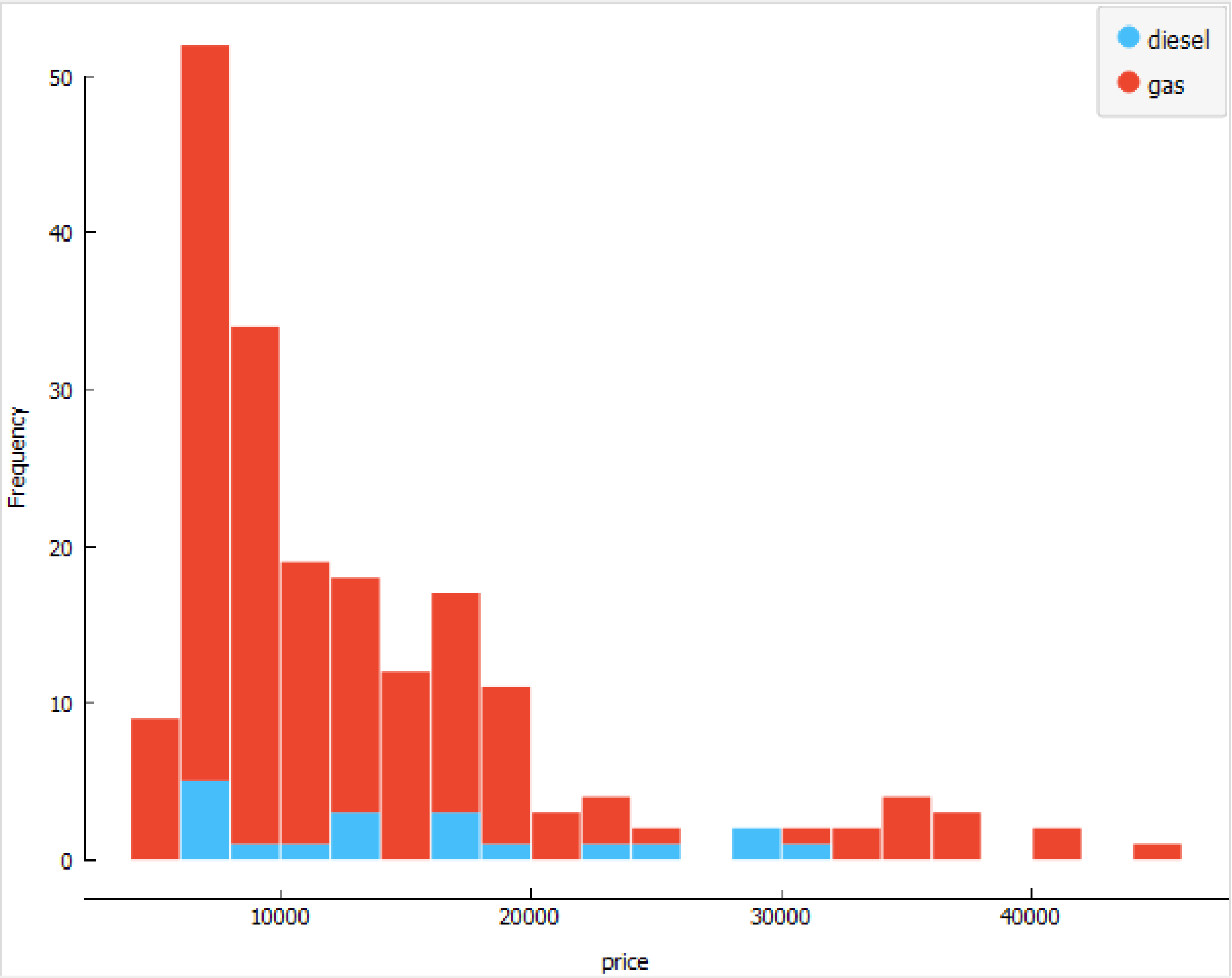
	Feature 1	make	fuel_type	aspiration	num_of_doors	body_style	drive_wheels
1	1	alfa-romero	gas	std	two	convertible	rwd
2	2	alfa-romero	gas	std	two	convertible	rwd
3	3	alfa-romero	gas	std	two	hatchback	rwd
4	4	audi	gas	std	four	sedan	fwd
5	5	audi	gas	std	four	sedan	4wd

Process



Output

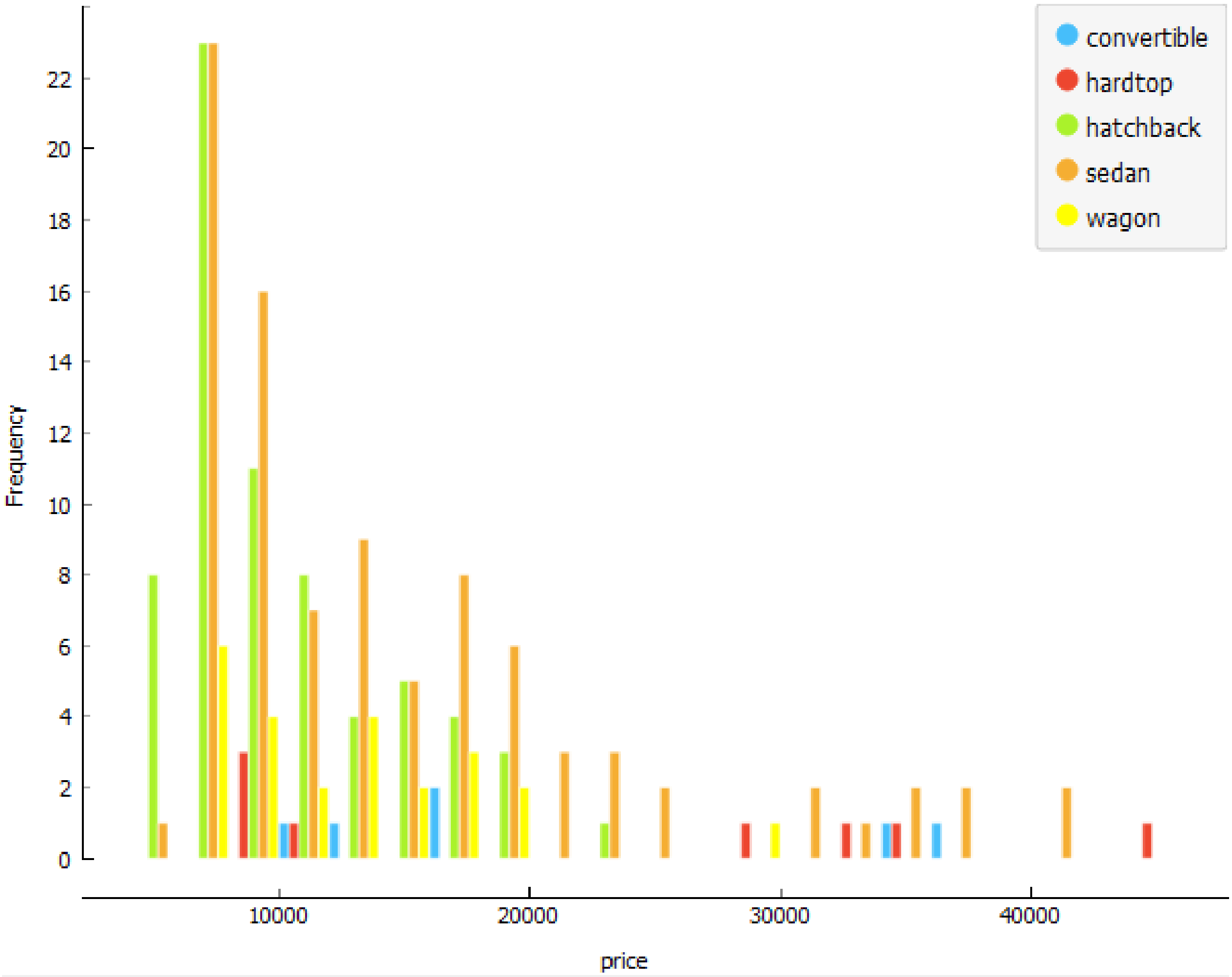
a) Stacked column distribution of price w.r.t fuel type.



Interpretation

- Low demand for diesel types
- Cost of gas type are higher

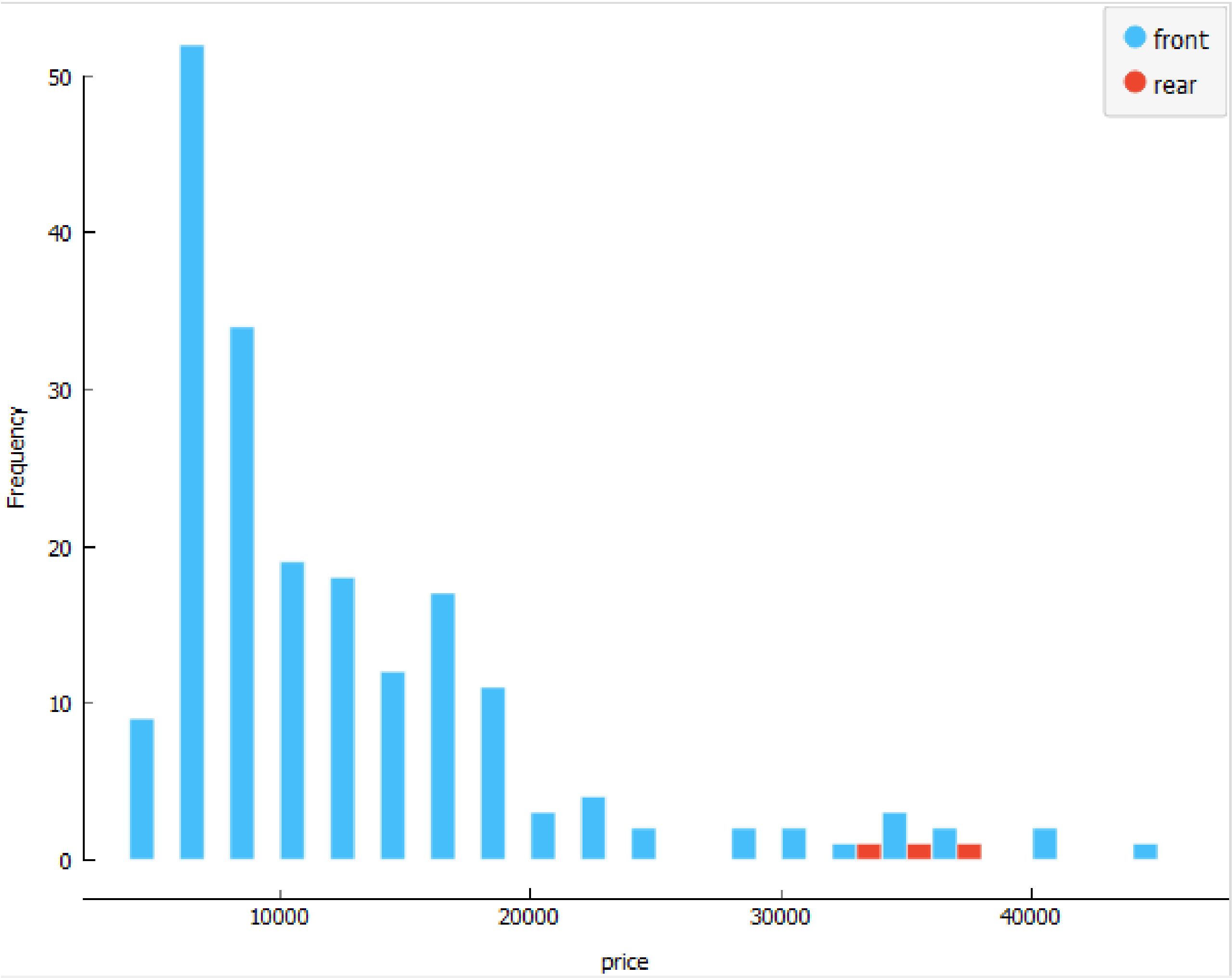
Histogram for price w.r.t body style.



Interpretation

- Hatchback and Sedan body styles has higher demand than others
- As price increases the demand for the hatchback decreases

Histogram of price with respect to engine location.

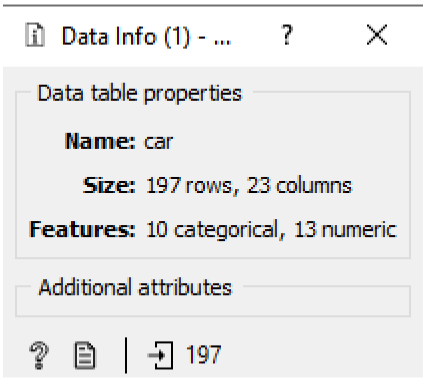


Interpretation

- Most cars has engine located at front
- Price of cars with engine at front are higher

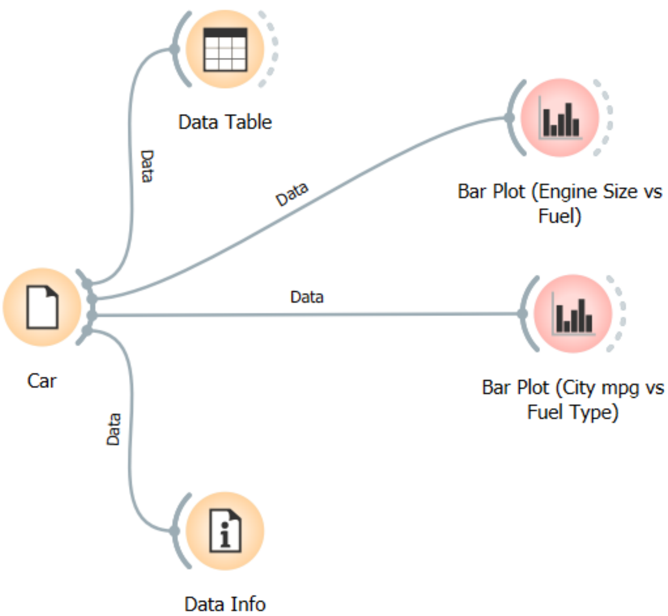
4) Visualize a bar plot for engine_size Vs make. Similarly visualize a bar plot for city_mpg vs fuel_type and write your inferences.

Input



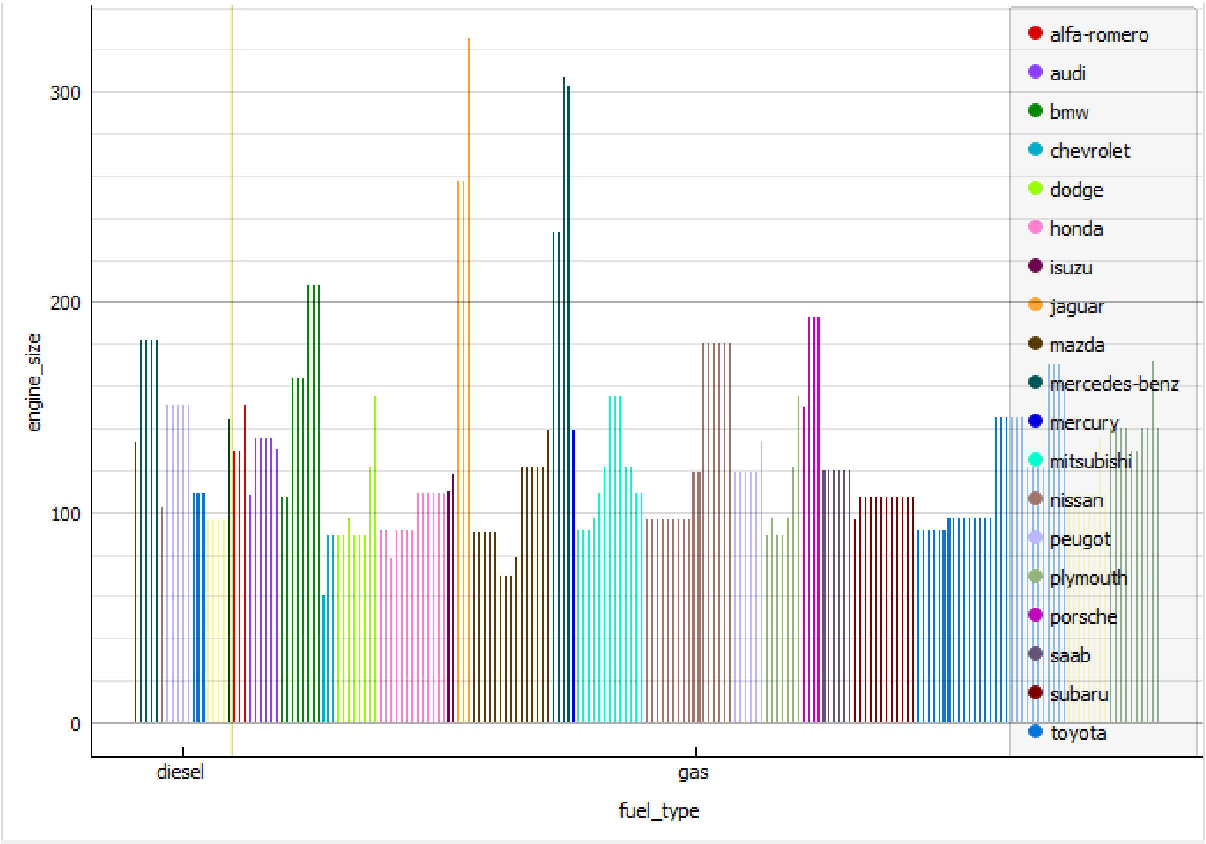
	Feature 1	make	fuel_type	aspiration	num_of_doors	body_style	drive_wheels
1	1	alfa-romero	gas	std	two	convertible	rwd
2	2	alfa-romero	gas	std	two	convertible	rwd
3	3	alfa-romero	gas	std	two	hatchback	rwd
4	4	audi	gas	std	four	sedan	fwd
5	5	audi	gas	std	four	sedan	4wd

Process

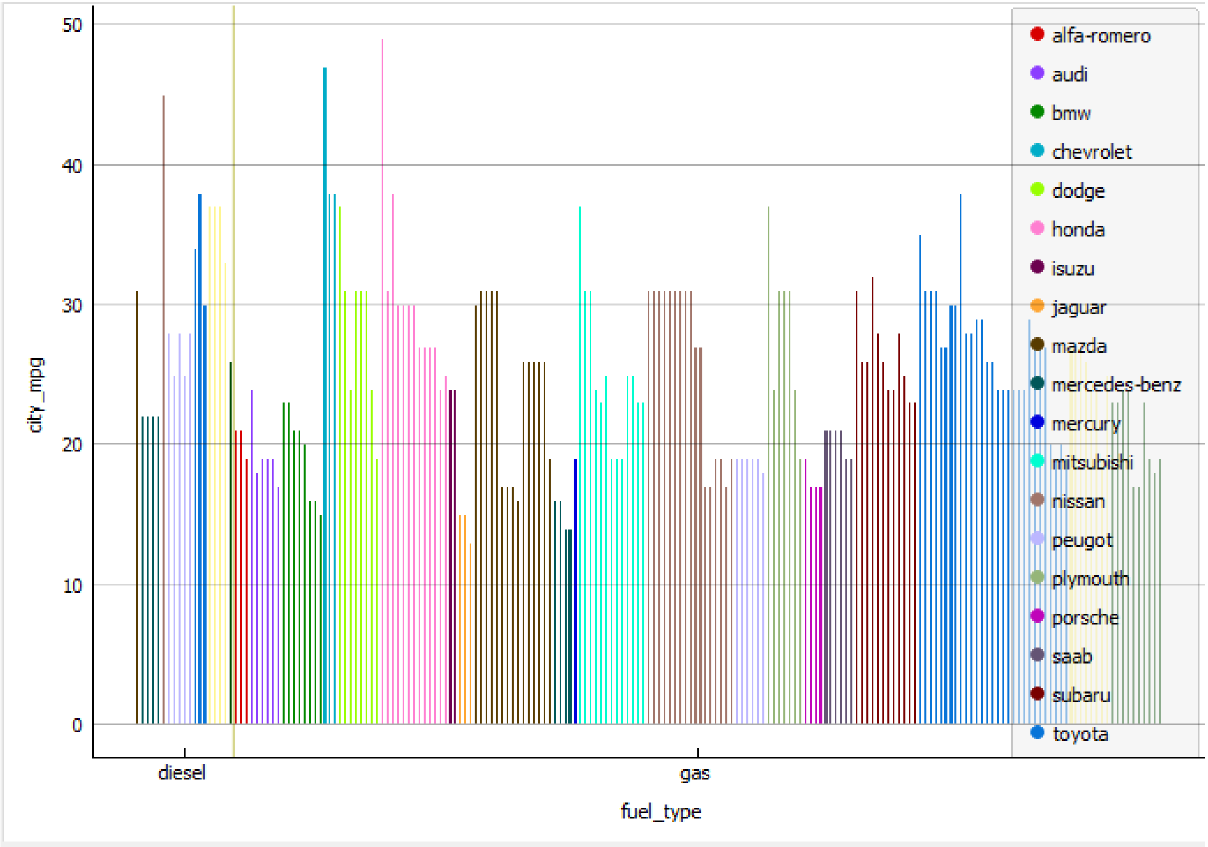


Output

Output 1 (Bar plot for engine_size vs fuel_type.)



Output 2 (Bar plot for city_mpg vs fuel_type.)

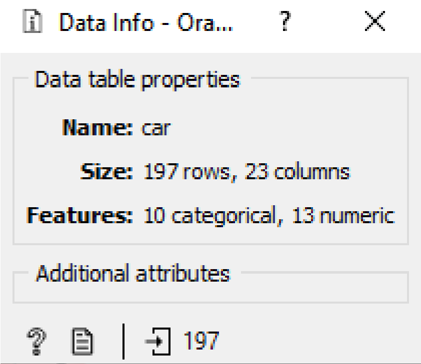


Interpretation

- Diesel cars have average mileage
- Cars on gas has mileage variations

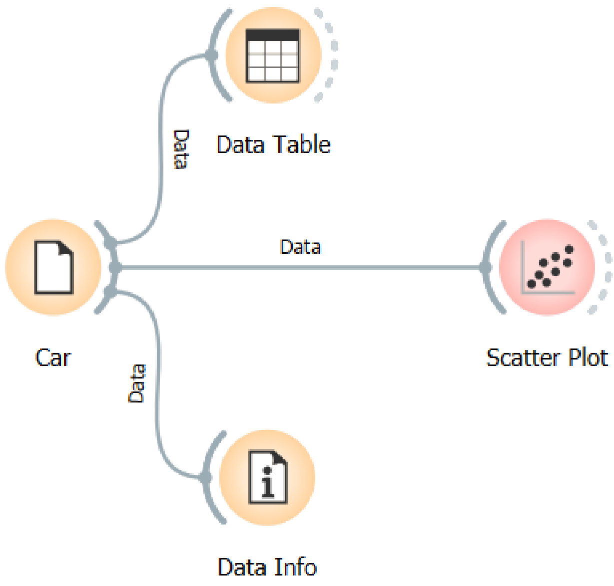
5) Create a scatter plot for price, vs engine_size, w.r.t num_of_cylinders(color), aspiration(shape), wheel_base(size).

Input



	Feature 1	make	fuel_type	aspiration	num_of_doors	body_style	drive_wheels
1	1	alfa-romero	gas	std	two	convertible	rwd
2	2	alfa-romero	gas	std	two	convertible	rwd
3	3	alfa-romero	gas	std	two	hatchback	rwd
4	4	audi	gas	std	four	sedan	fwd
5	5	audi	gas	std	four	sedan	4wd

Process



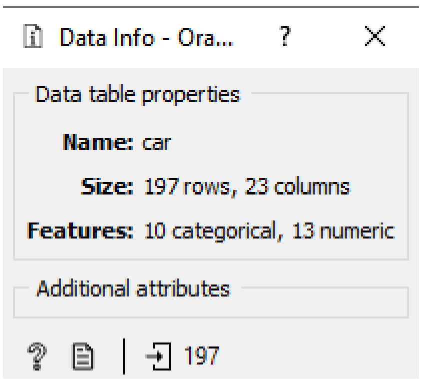
Output

Interpretation

- Here, the size of the engine increases as the price increases
- Positive correlation

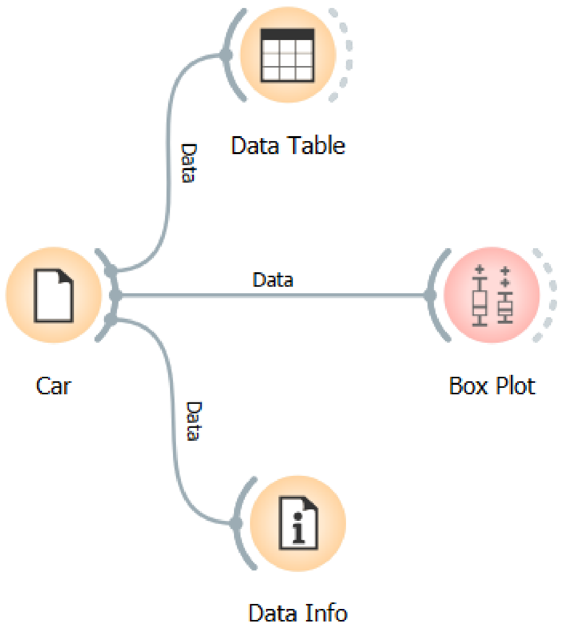
6) Create a boxplot for price w.r.t body_styles.

Input

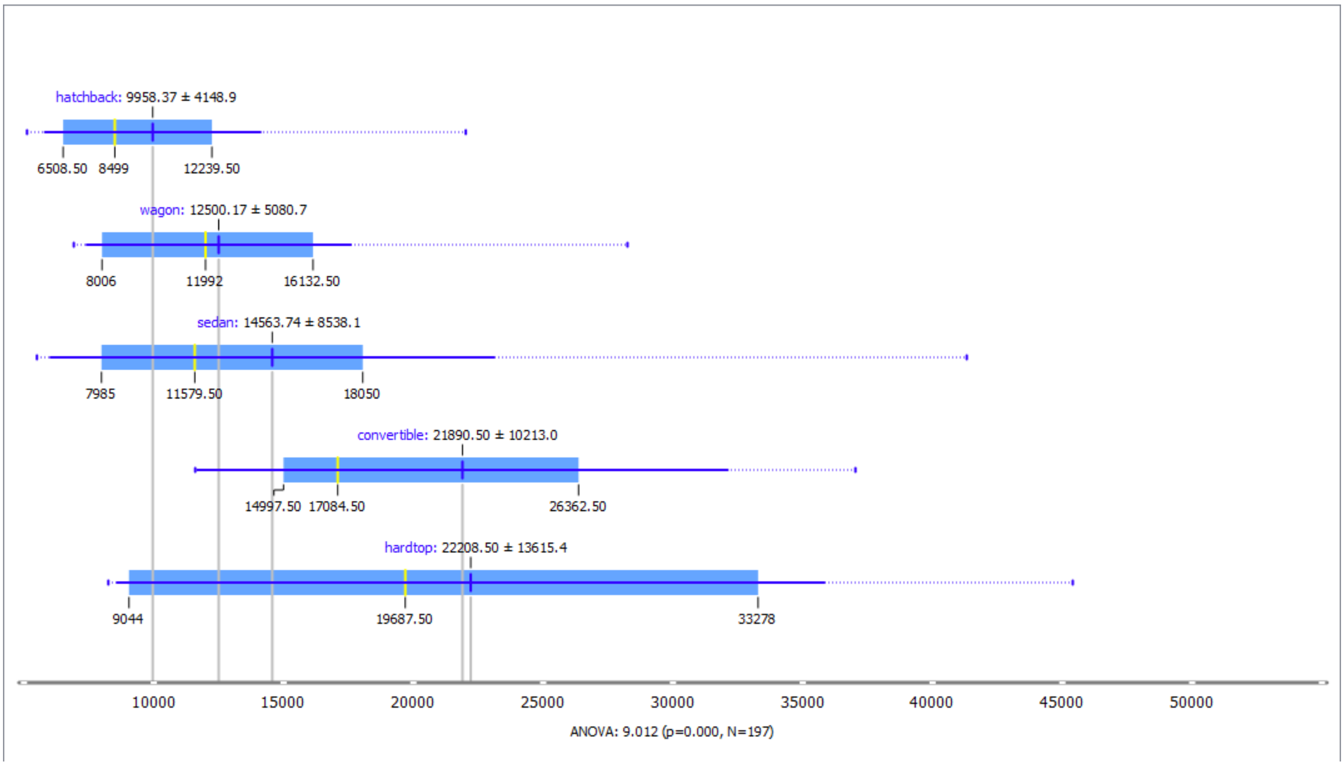


	Feature 1	make	fuel_type	aspiration	num_of_doors	body_style	drive_wheels
1	1	alfa-romero	gas	std	two	convertible	rwd
2	2	alfa-romero	gas	std	two	convertible	rwd
3	3	alfa-romero	gas	std	two	hatchback	rwd
4	4	audi	gas	std	four	sedan	fwd
5	5	audi	gas	std	four	sedan	4wd

Process

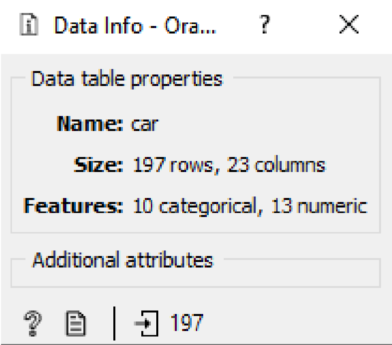


Output



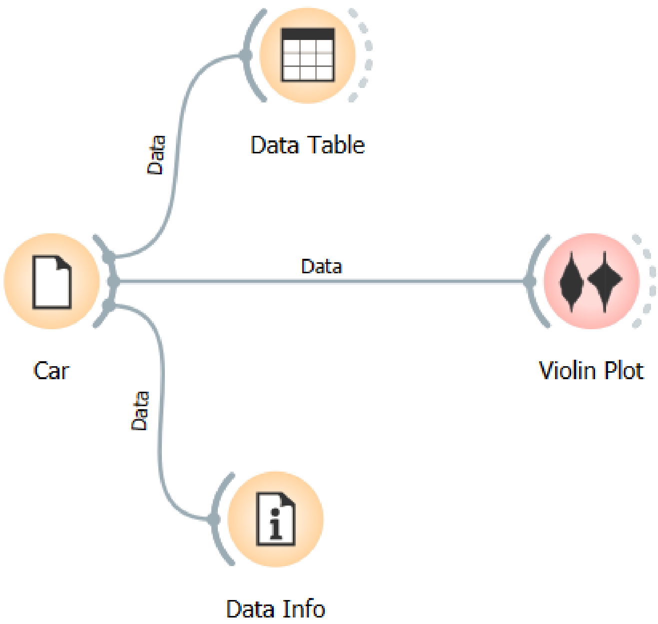
7) Create a violin plot for price w.r.t aspiration.

Input

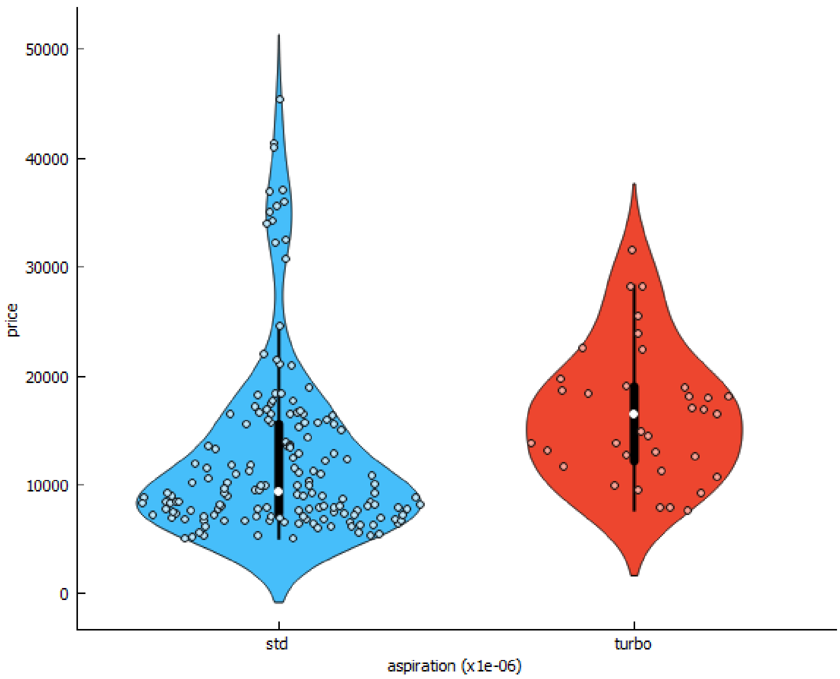


	Feature 1	make	fuel_type	aspiration	num_of_doors	body_style	drive_wheels
1	1	alfa-romero	gas	std	two	convertible	rwd
2	2	alfa-romero	gas	std	two	convertible	rwd
3	3	alfa-romero	gas	std	two	hatchback	rwd
4	4	audi	gas	std	four	sedan	fwd
5	5	audi	gas	std	four	sedan	4wd

Process

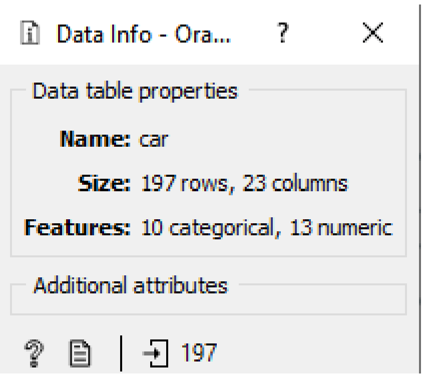


Output



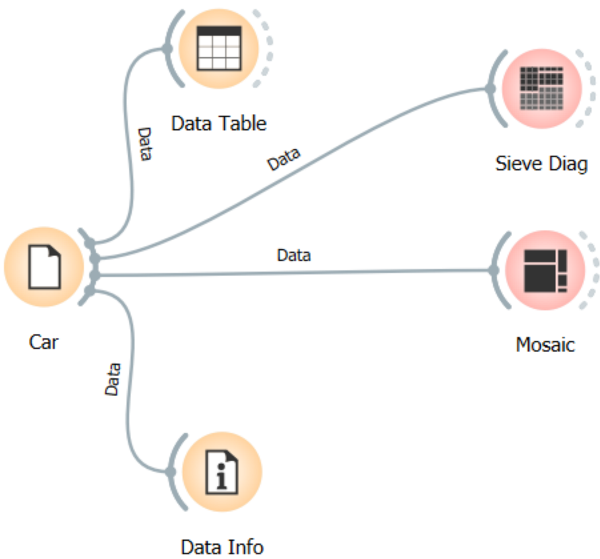
8) Illustrate sieve diagram and mosaic display for city_mpg vs highway_mpg.

Input



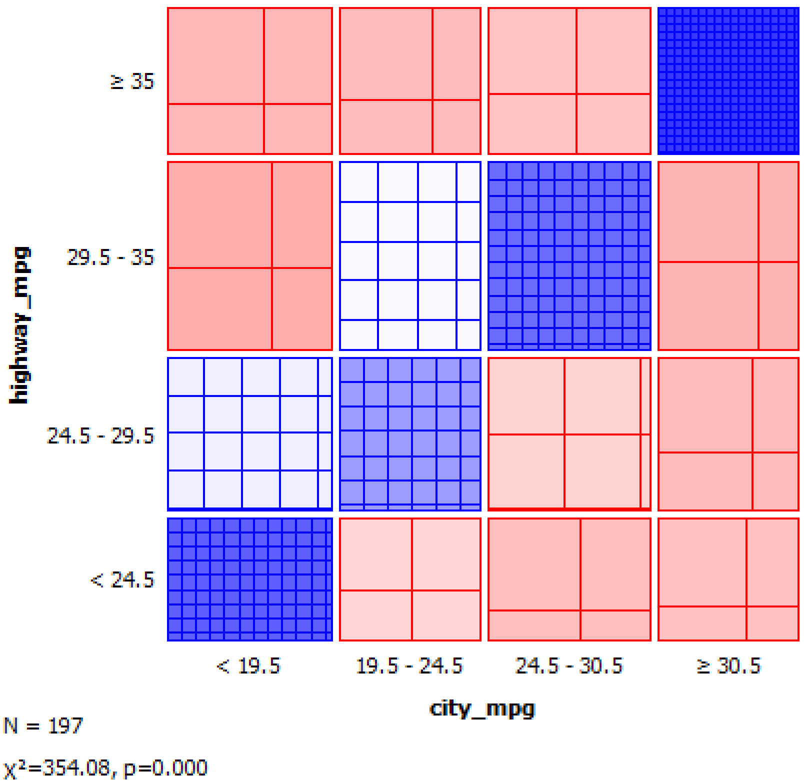
	Feature 1	make	fuel_type	aspiration	num_of_doors	body_style	drive_wheels
1	1	alfa-romero	gas	std	two	convertible	rwd
2	2	alfa-romero	gas	std	two	convertible	rwd
3	3	alfa-romero	gas	std	two	hatchback	rwd
4	4	audi	gas	std	four	sedan	fwd
5	5	audi	gas	std	four	sedan	4wd

Process

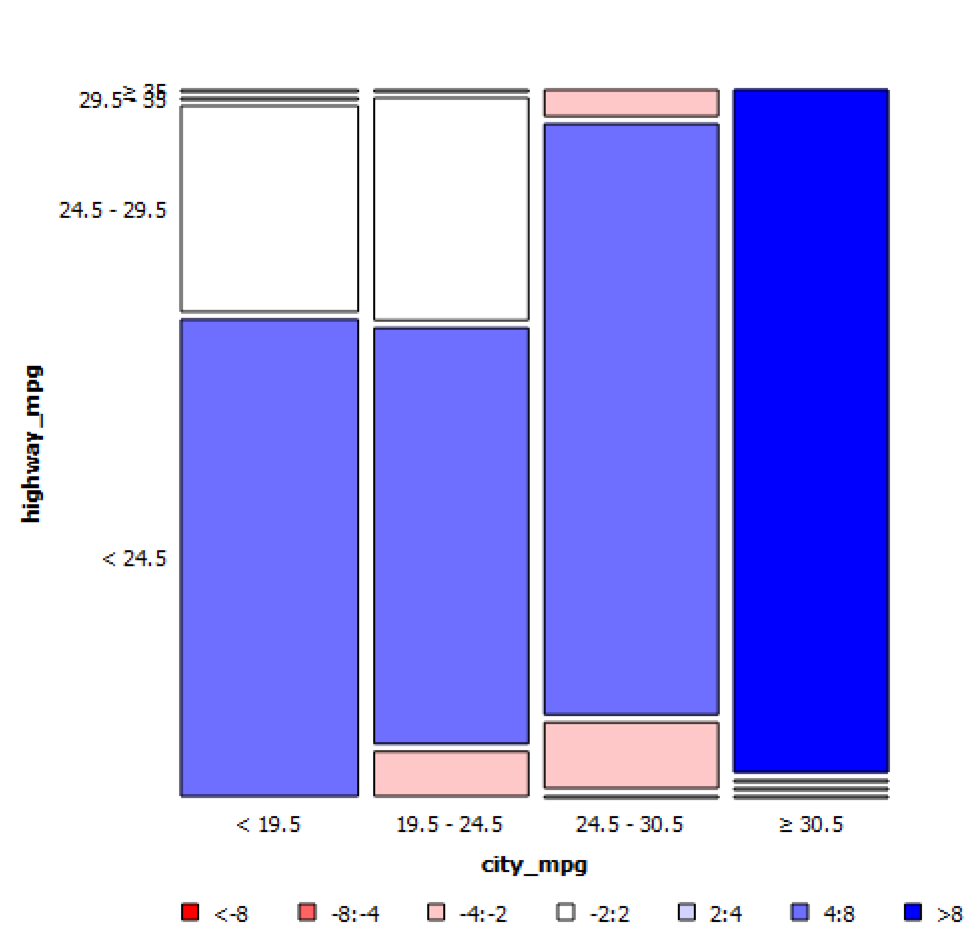


Output

Output 1 (Sieve Diagram for city_mpg vs highway_mpg)



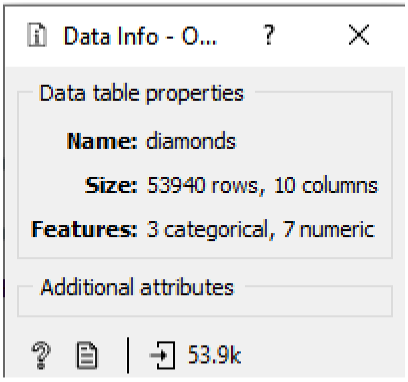
Output 2 (Mosaic Display for city_mpg vs highway_mpg)



Illustrate the following using *diamonds* data set

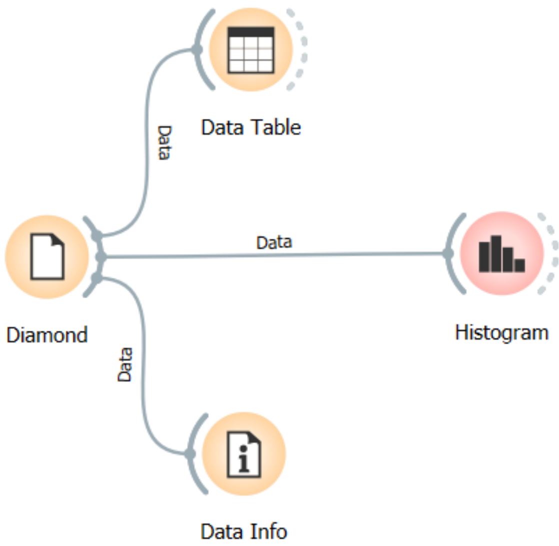
1) Create a histogram of "carat" w.r.t cut.

Input

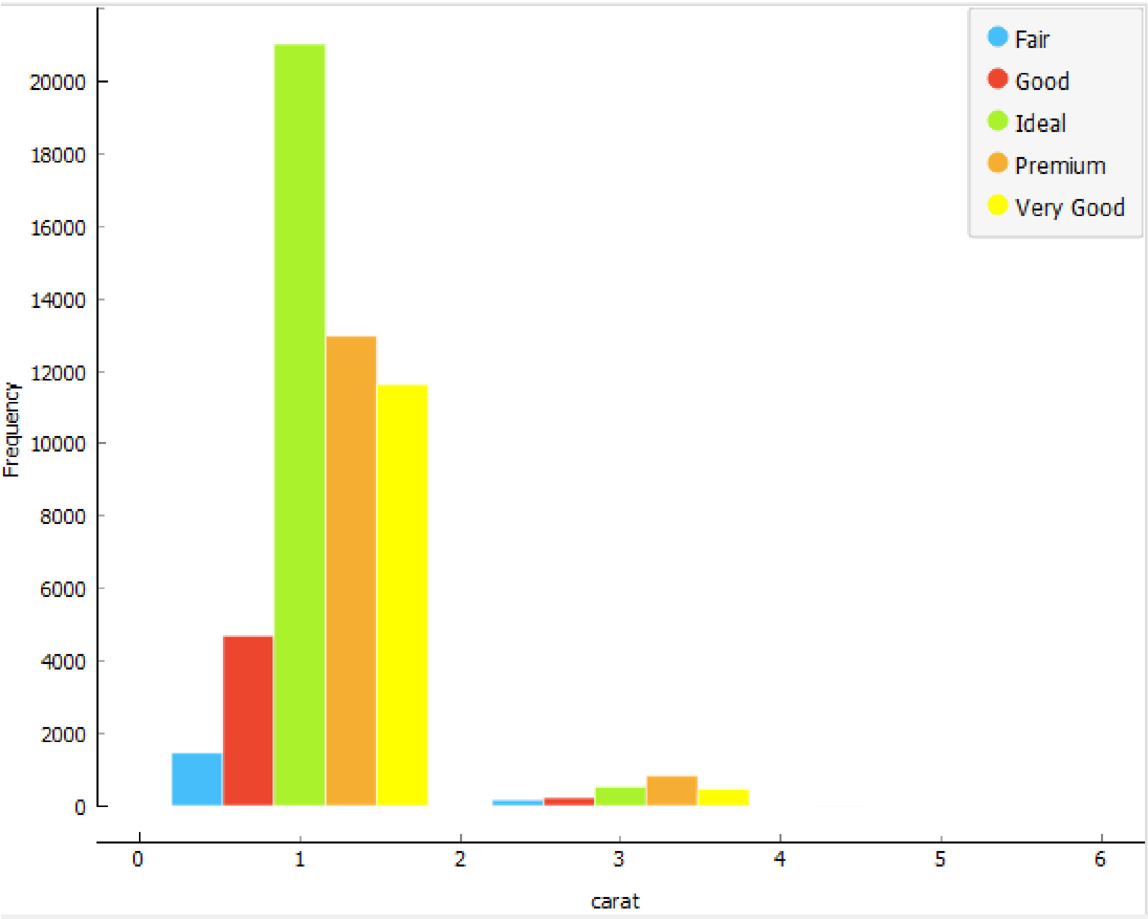


	carat	cut	color	clarity	depth	table	price
1	0.23	Ideal	E	SI2	61.5	55.0	326
2	0.21	Premium	E	SI1	59.8	61.0	326
3	0.23	Good	E	VS1	56.9	65.0	327
4	0.29	Premium	I	VS2	62.4	58.0	334
5	0.31	Good	J	SI2	63.3	58.0	335

Process



Output

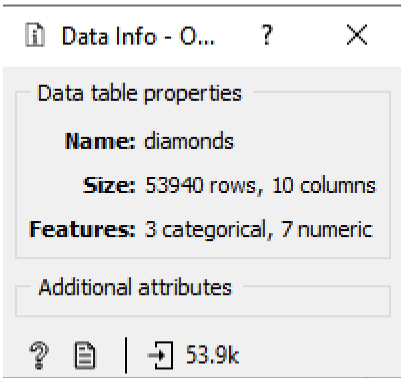


Interpretation

- Frequency of Ideal cuts are higher

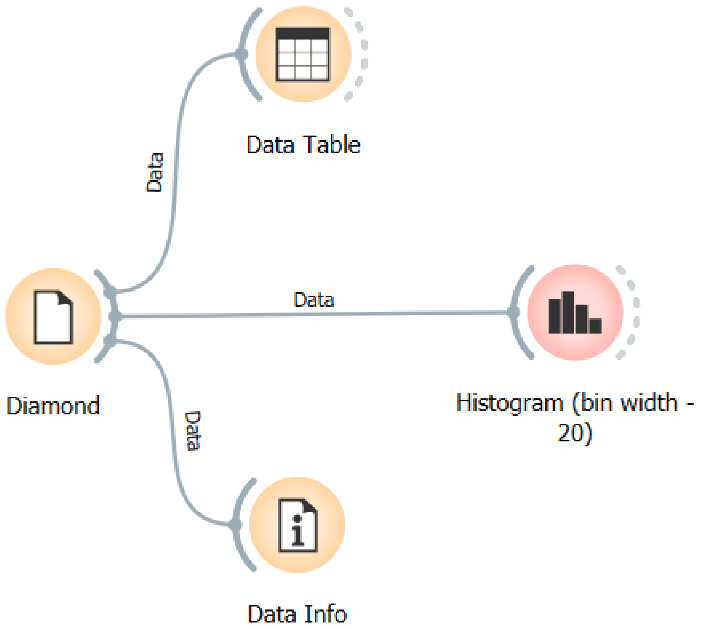
2) Set the bin width of the histogram to 20.

Input

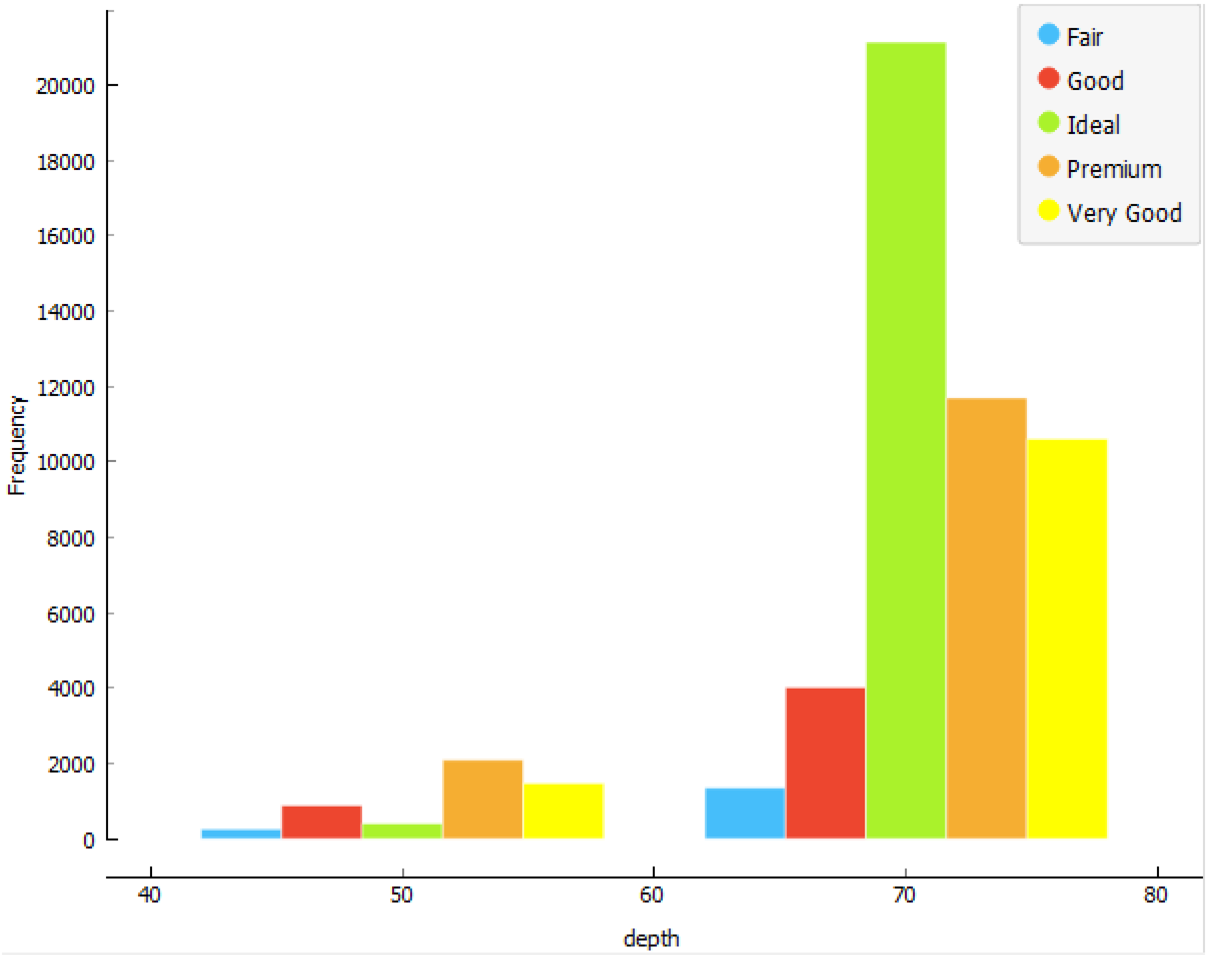


	carat	cut	color	clarity	depth	table	price
1	0.23	Ideal	E	SI2	61.5	55.0	326
2	0.21	Premium	E	SI1	59.8	61.0	326
3	0.23	Good	E	VS1	56.9	65.0	327
4	0.29	Premium	I	VS2	62.4	58.0	334
5	0.31	Good	J	SI2	63.3	58.0	335

Process

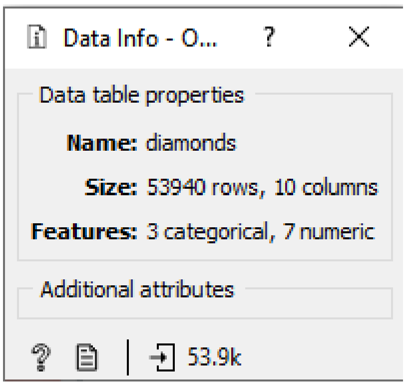


Output



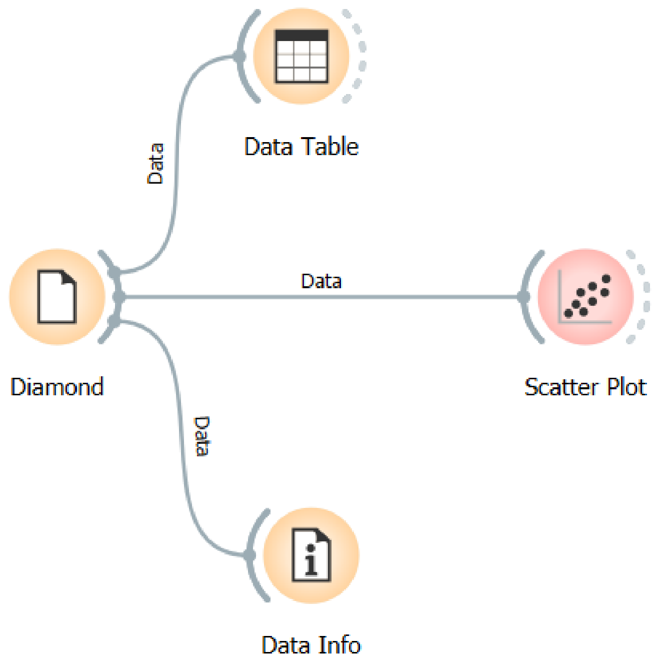
3) Make a scatterplot: carat vs price, set the color to clarity.

Input

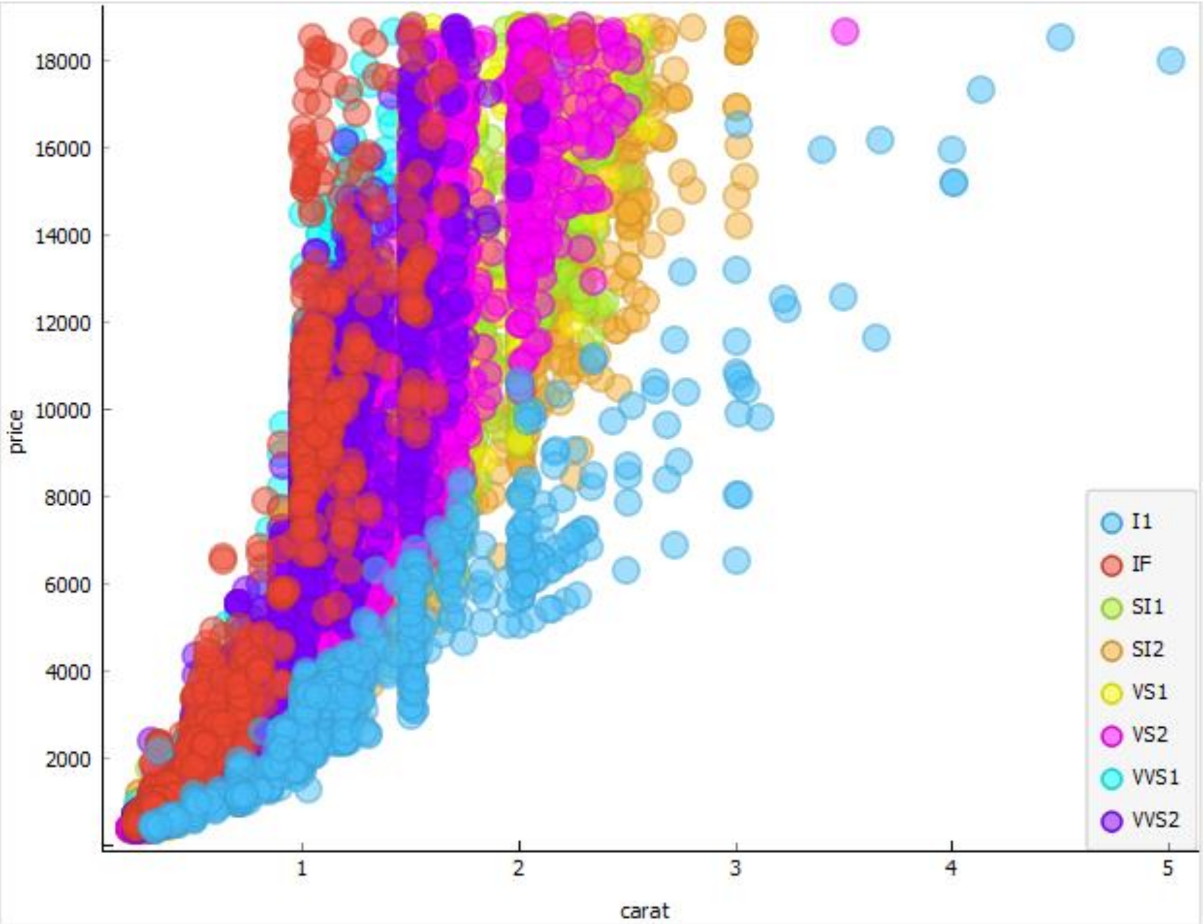


	carat	cut	color	clarity	depth	table	price
1	0.23	Ideal	E	SI2	61.5	55.0	326
2	0.21	Premium	E	SI1	59.8	61.0	326
3	0.23	Good	E	VS1	56.9	65.0	327
4	0.29	Premium	I	VS2	62.4	58.0	334
5	0.31	Good	J	SI2	63.3	58.0	335

Process

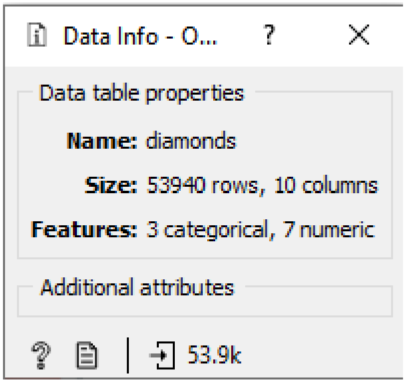


Output



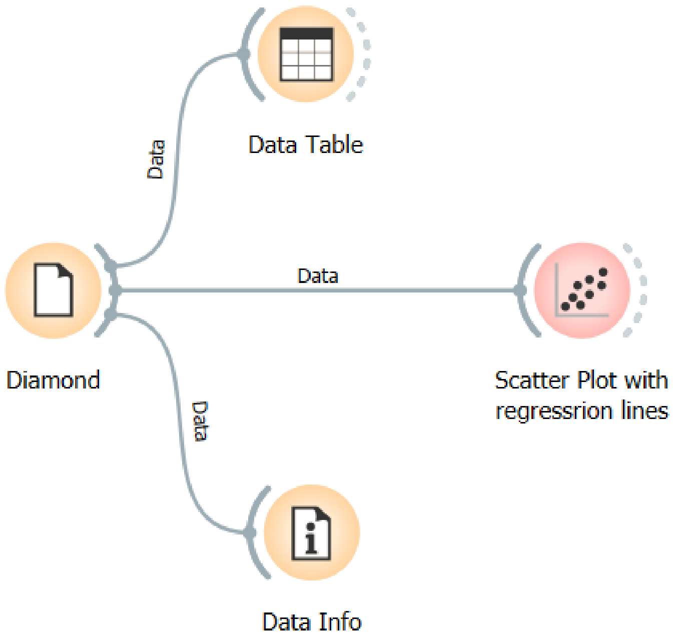
4) Make a scatterplot: carat vs price, set the color to clarity. Also add regression line to the plot.

Input

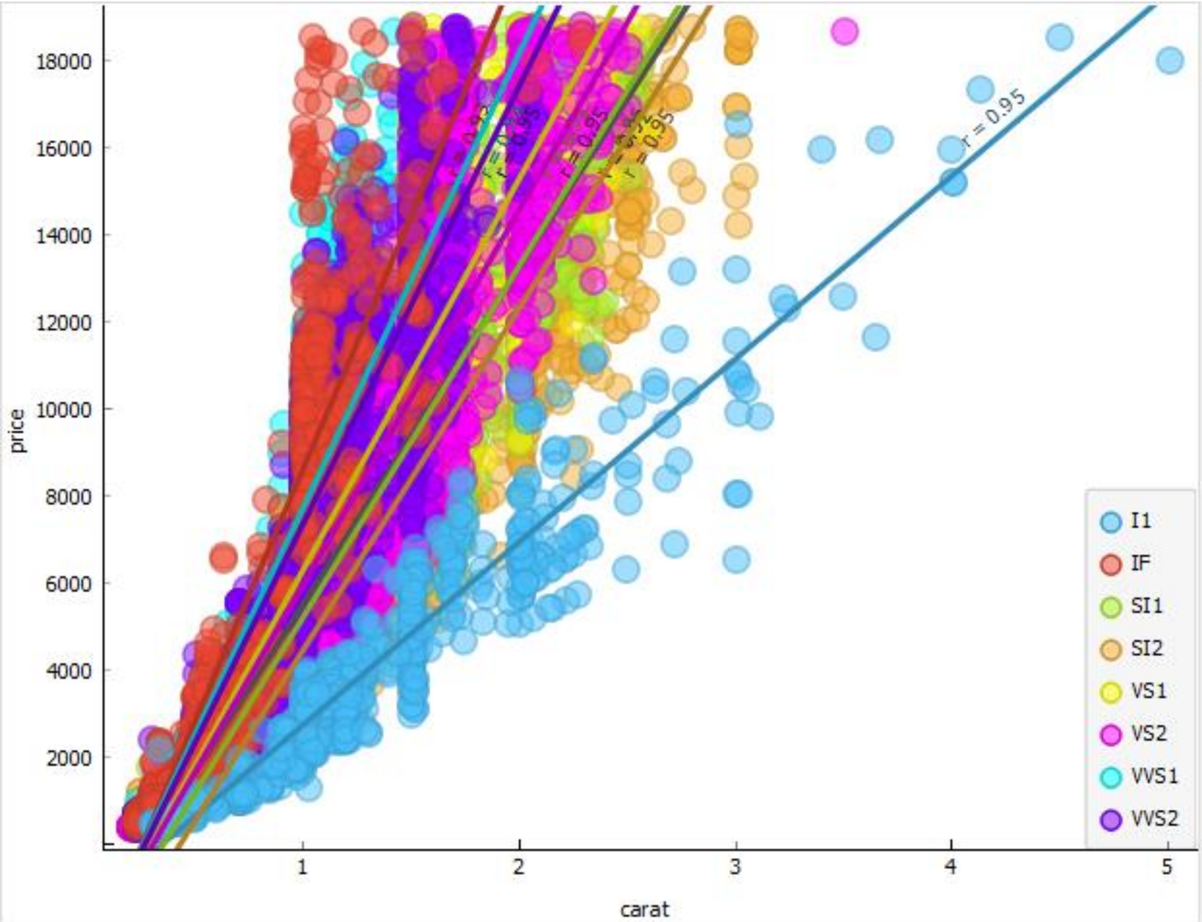


	carat	cut	color	clarity	depth	table	price
1	0.23	Ideal	E	SI2	61.5	55.0	326
2	0.21	Premium	E	SI1	59.8	61.0	326
3	0.23	Good	E	VS1	56.9	65.0	327
4	0.29	Premium	I	VS2	62.4	58.0	334
5	0.31	Good	J	SI2	63.3	58.0	335

Process

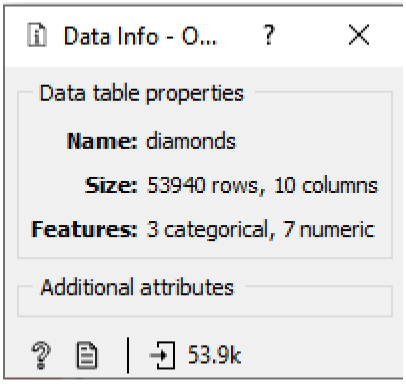


Output



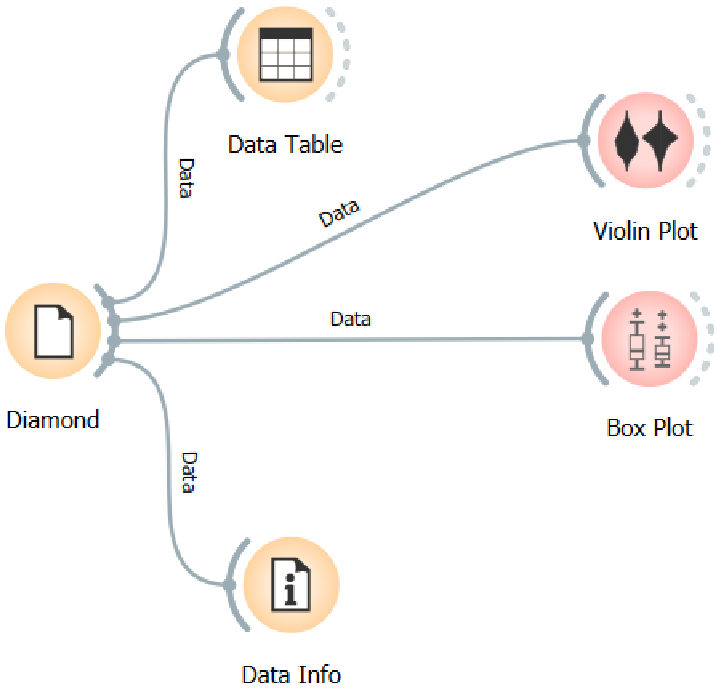
5) For carat vs cut, make a violin and a boxplot.

Input



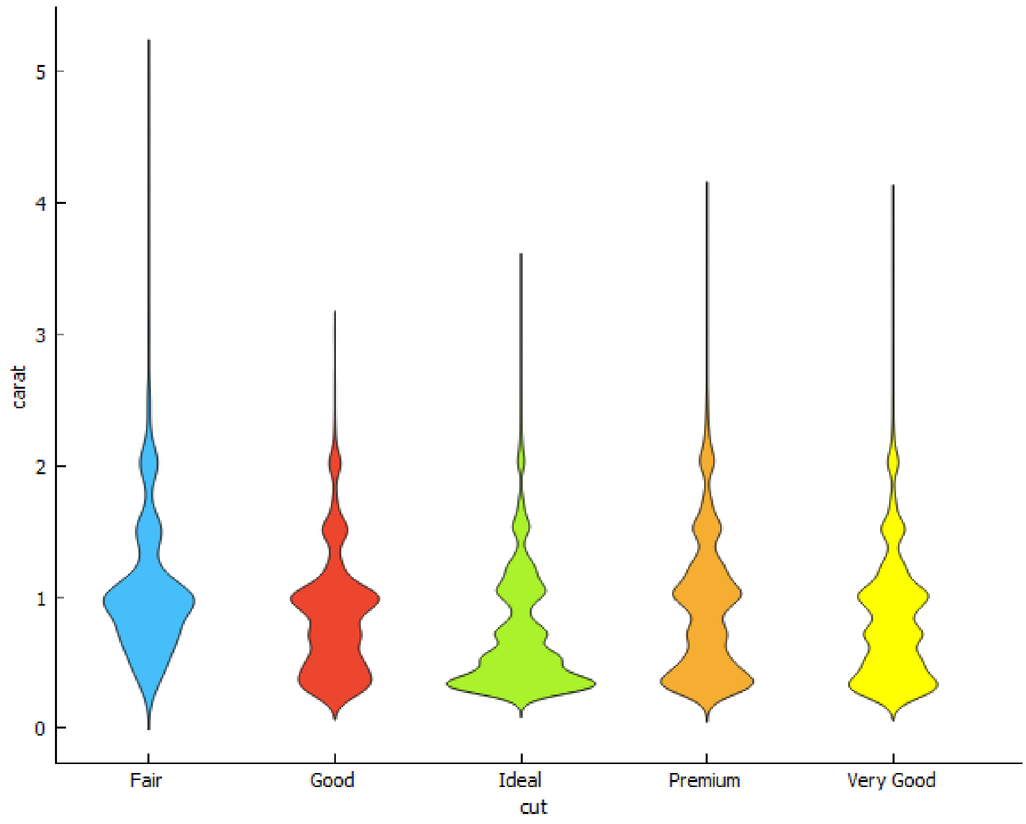
	carat	cut	color	clarity	depth	table	price
1	0.23	Ideal	E	SI2	61.5	55.0	326
2	0.21	Premium	E	SI1	59.8	61.0	326
3	0.23	Good	E	VS1	56.9	65.0	327
4	0.29	Premium	I	VS2	62.4	58.0	334
5	0.31	Good	J	SI2	63.3	58.0	335

Process



Output

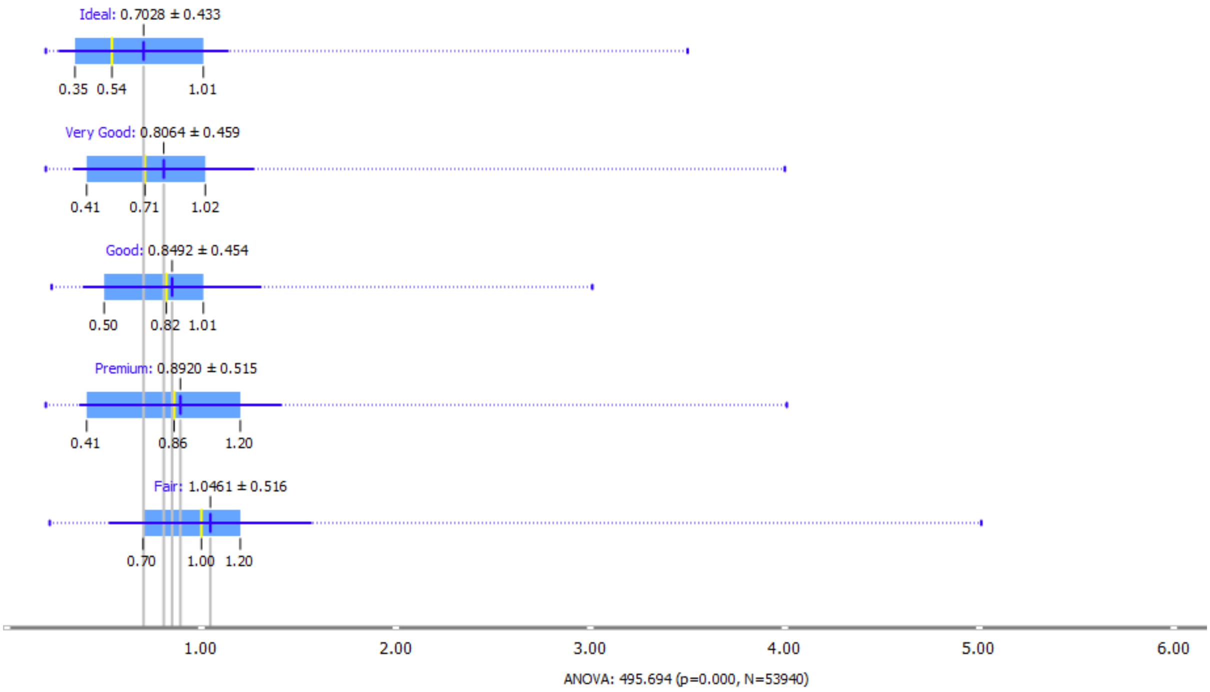
Output 1 (Violin plot)



Interpretation

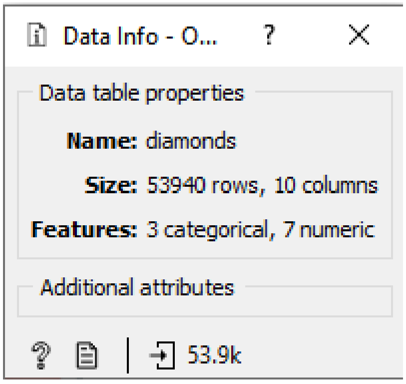
- Most diamonds have average average quality ideal cut

Output 2 (Box plot)



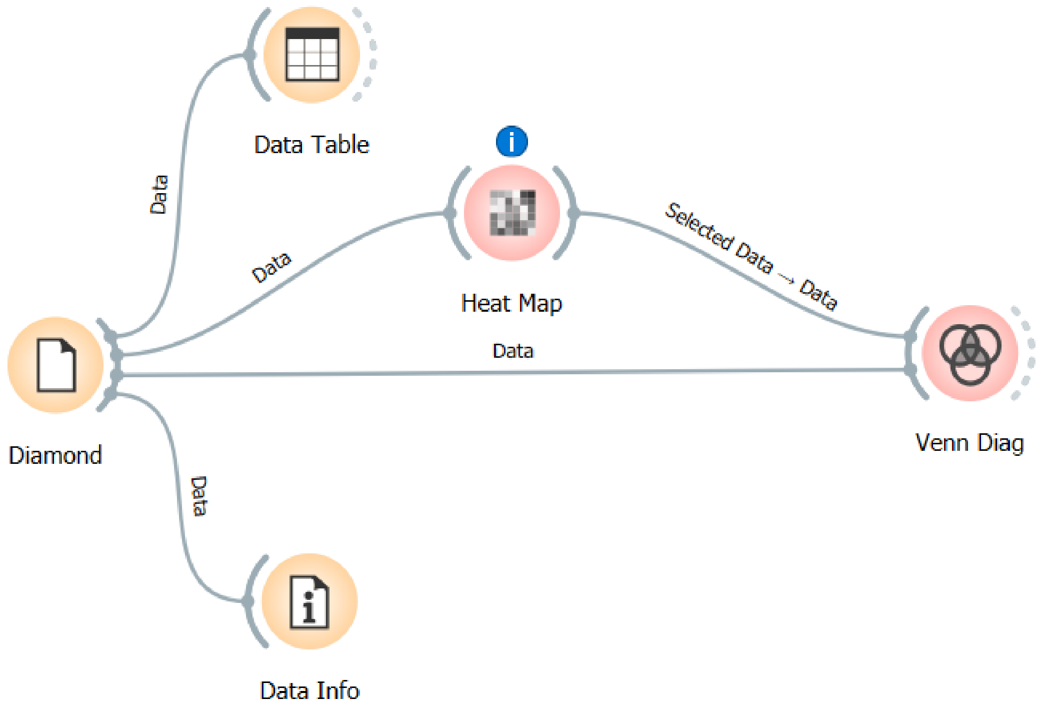
6) Illustrate Heat map and Venn Diagram using the data set.

Input



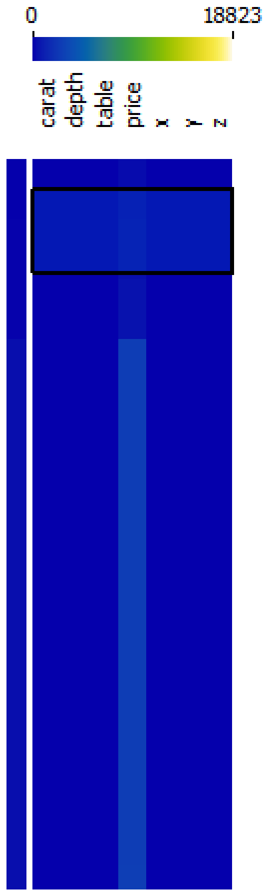
	carat	cut	color	clarity	depth	table	price
1	0.23	Ideal	E	SI2	61.5	55.0	326
2	0.21	Premium	E	SI1	59.8	61.0	326
3	0.23	Good	E	VS1	56.9	65.0	327
4	0.29	Premium	I	VS2	62.4	58.0	334
5	0.31	Good	J	SI2	63.3	58.0	335

Process



Output

Output 1 (Heat map)



Output 2 (Venn Diagram)

7) Illustrate freeviz, linear projection and radviz using the data set.

Input

Data Info - O... ?

Data table properties

Name: diamonds

Size: 53940 rows, 10 columns

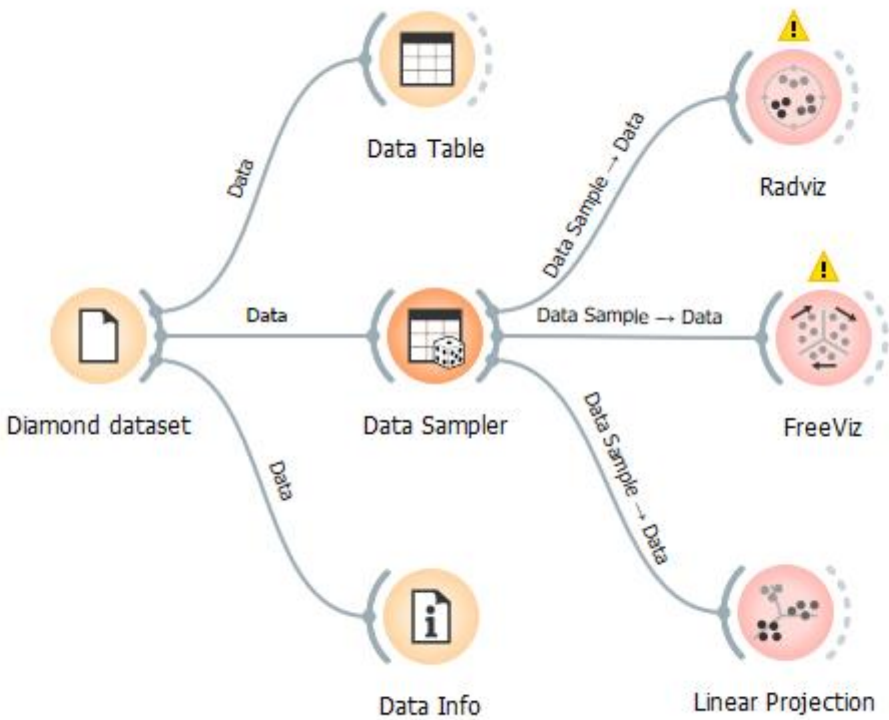
Features: 3 categorical, 7 numeric

Additional attributes

53.9k

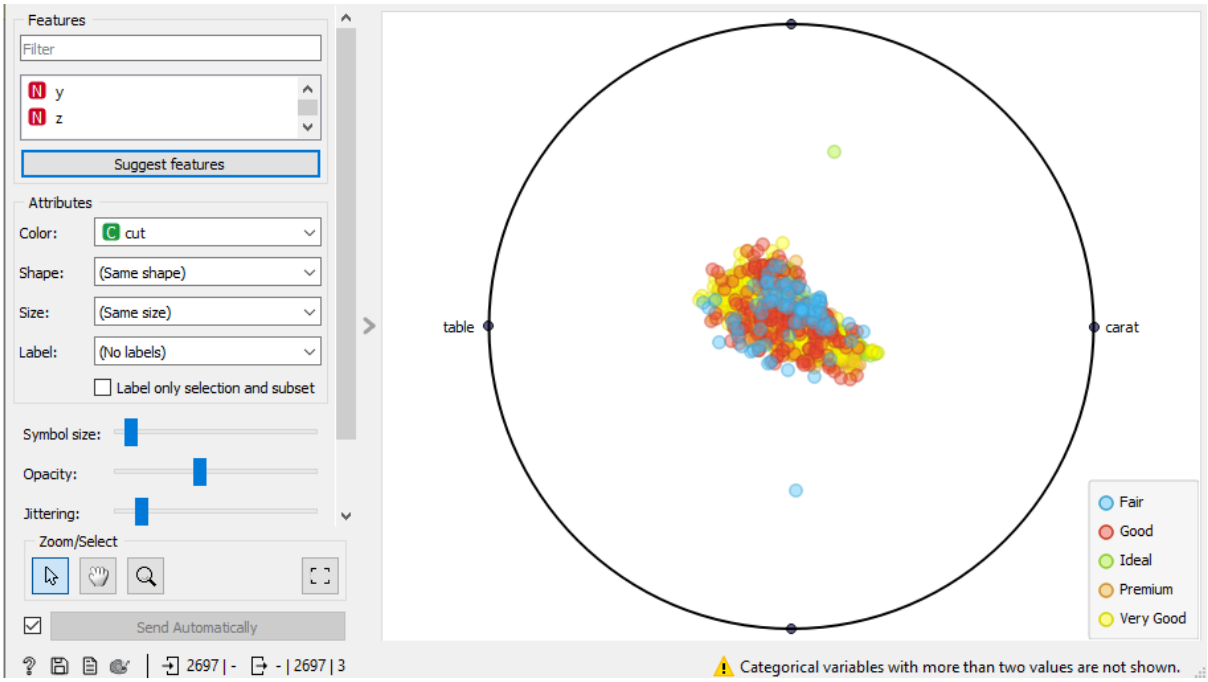
	carat	cut	color	clarity	depth	table	price
1	0.23	Ideal	E	SI2	61.5	55.0	326
2	0.21	Premium	E	SI1	59.8	61.0	326
3	0.23	Good	E	VS1	56.9	65.0	327
4	0.29	Premium	I	VS2	62.4	58.0	334
5	0.31	Good	J	SI2	63.3	58.0	335

Process

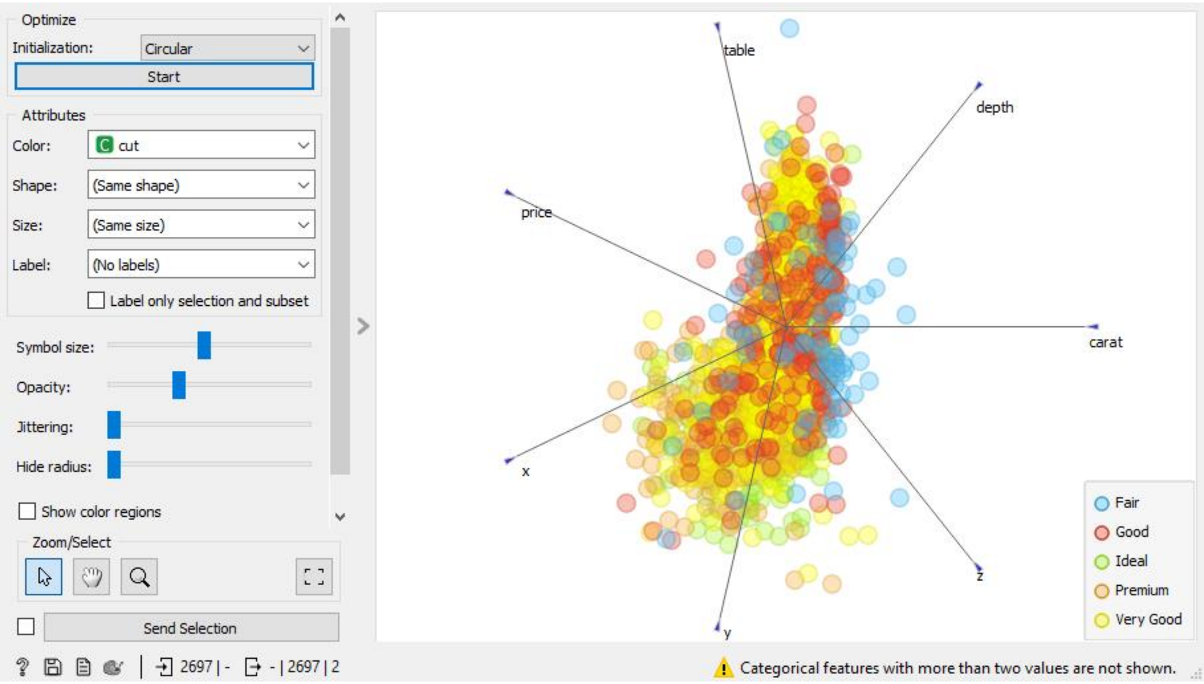


Output

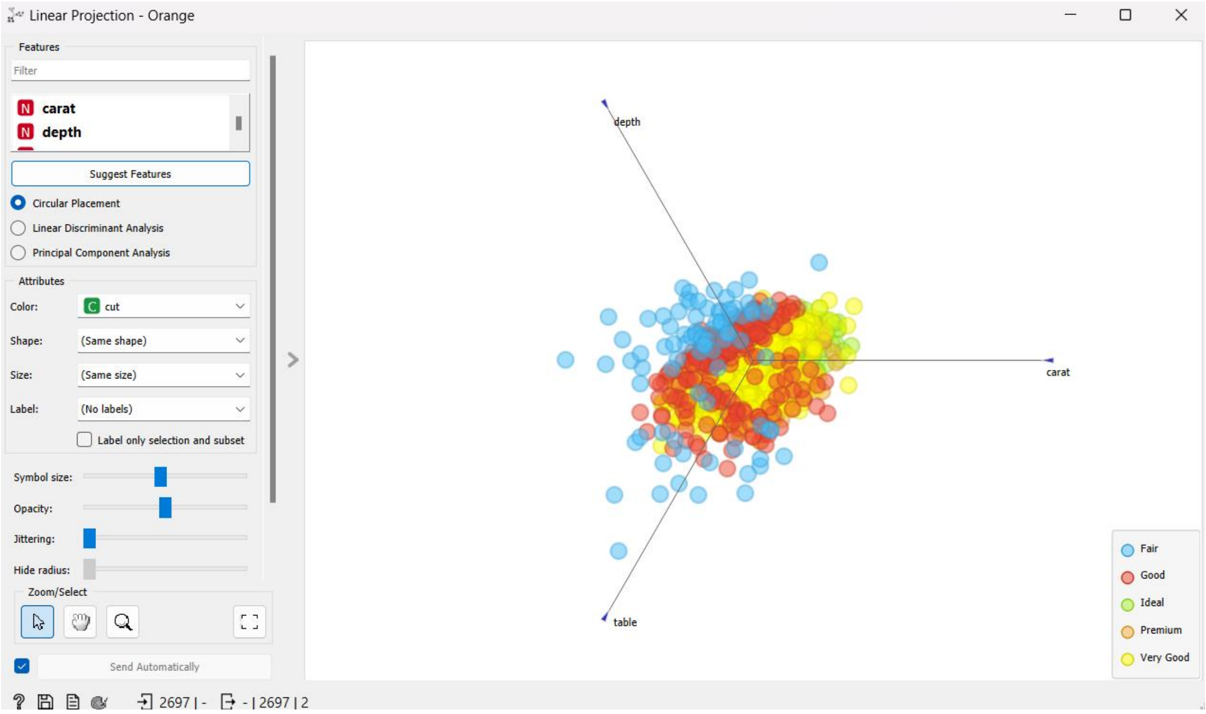
Radviz



Freeviz



Linear projection



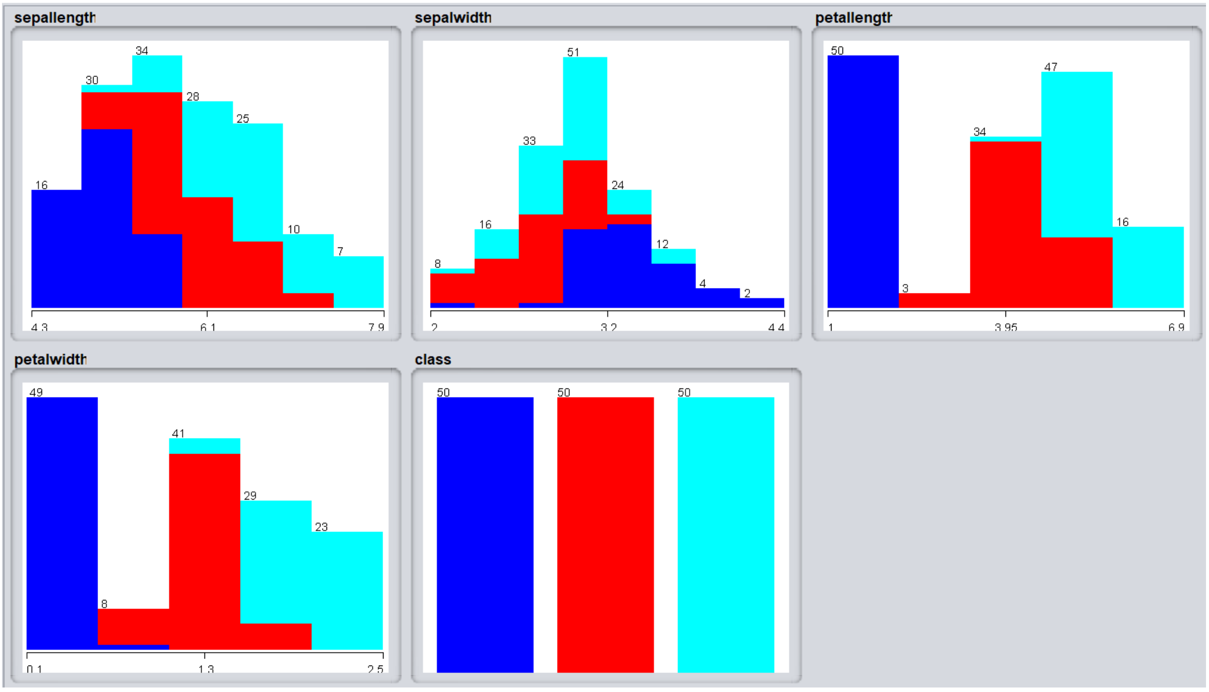
a Data Visualization (Weka)

1) Give a visualization of the distribution of Iris dataset w.r.t all the features

Relation: iris

No.	1: sepallength Numeric	2: sepalwidth Numeric	3: petallength Numeric	4: petalwidth Numeric	5: class Nominal
1	5.1	3.5	1.4	0.2	Iris-s...
2	4.9	3.0	1.4	0.2	Iris-s...
3	4.7	3.2	1.3	0.2	Iris-s...
4	4.6	3.1	1.5	0.2	Iris-s...
5	5.0	3.6	1.4	0.2	Iris-s...
6	5.4	3.9	1.7	0.4	Iris-s...
7	4.6	3.4	1.4	0.3	Iris-s...

Output



2) Display the plot matrix for the Iris data set

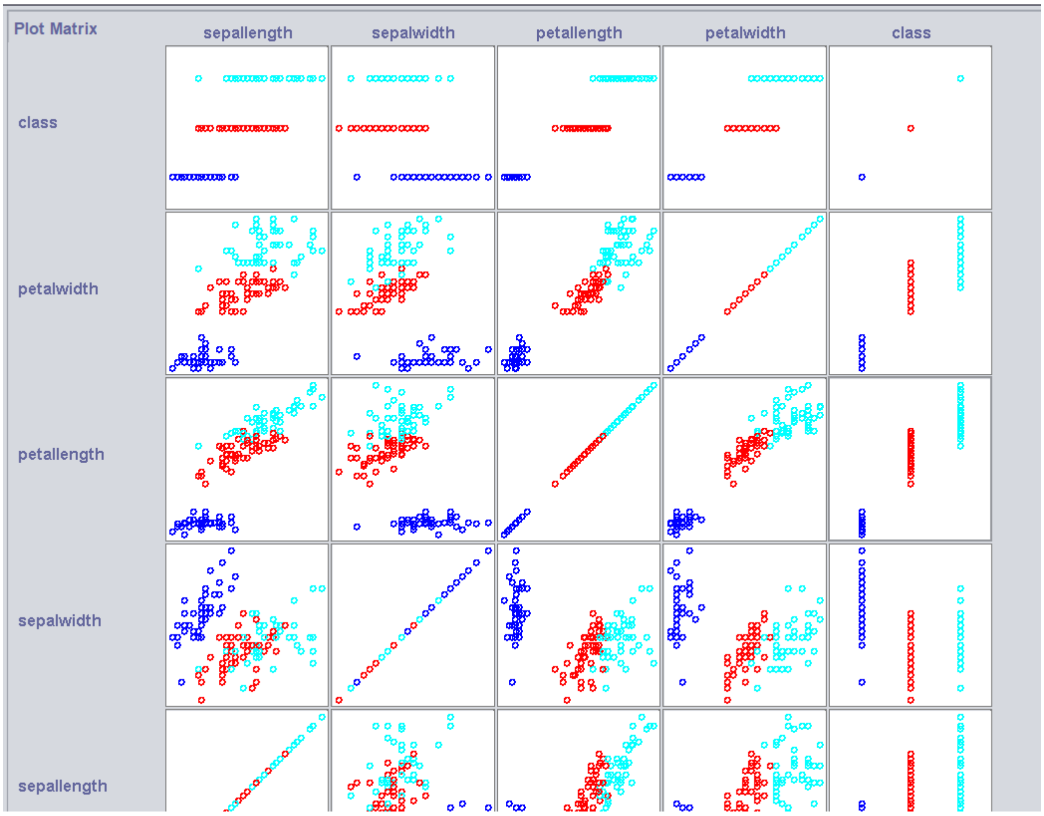
Input

Relation: iris

No.	1: sepallength Numeric	2: sepalwidth Numeric	3: petallength Numeric	4: petalwidth Numeric	5: class Nominal
1	5.1	3.5	1.4	0.2	Iris-s...
2	4.9	3.0	1.4	0.2	Iris-s...
3	4.7	3.2	1.3	0.2	Iris-s...
4	4.6	3.1	1.5	0.2	Iris-s...
5	5.0	3.6	1.4	0.2	Iris-s...
6	5.4	3.9	1.7	0.4	Iris-s...
7	4.6	3.4	1.4	0.3	Iris-s...

Process

Output

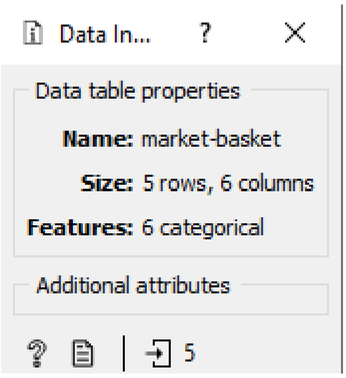


Section III

a (Association Rule Mining -Class Work)

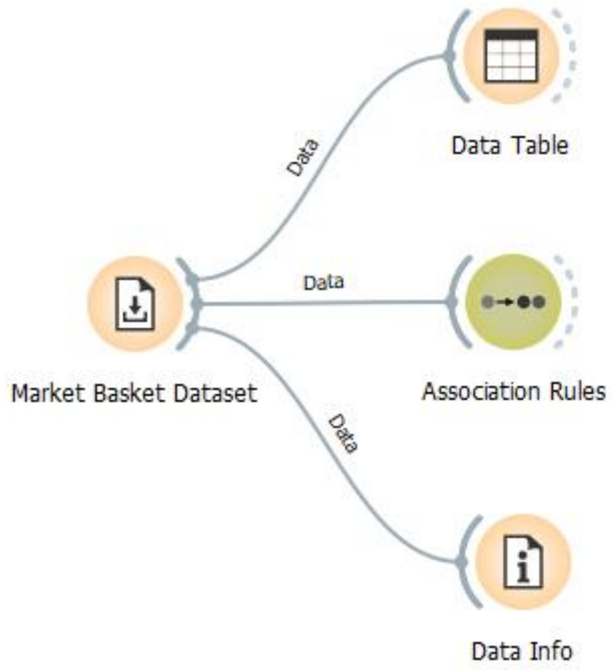
1) Generate association rules using Market Basket Data set in Orange Tool. Compare the different measures to assess the quality of rules.

Input

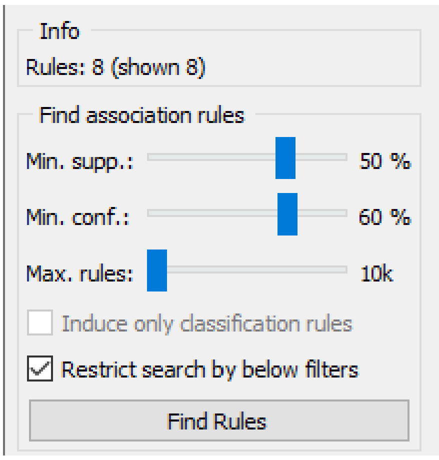


	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	?	?	?	?
2	1	?	1	1	1	?
3	?	1	1	1	?	1
4	1	1	1	1	?	?
5	1	1	1	?	?	1

Process



Output



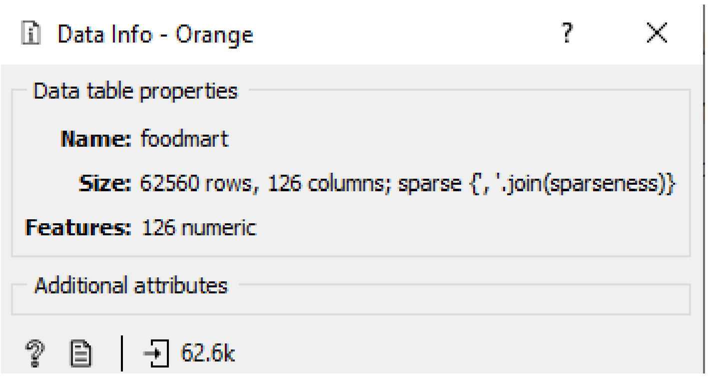
Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.600	0.750	0.800	1.000	0.938	-0.040	Milk=1	→	Bread=1
0.600	0.750	0.800	1.000	0.938	-0.040	Bread=1	→	Milk=1
0.600	0.750	0.800	1.000	0.938	-0.040	Diapers=1	→	Bread=1
0.600	0.750	0.800	1.000	0.938	-0.040	Bread=1	→	Diapers=1
0.600	0.750	0.800	1.000	0.938	-0.040	Diapers=1	→	Milk=1
0.600	0.750	0.800	1.000	0.938	-0.040	Milk=1	→	Diapers=1
0.600	1.000	0.600	1.333	1.250	0.120	Beer=1	→	Diapers=1
0.600	0.750	0.800	0.750	1.250	0.120	Diapers=1	→	Beer=1

Interpretation

- There is a high chance that a person buying bread,diapers and egg will also buy beer.

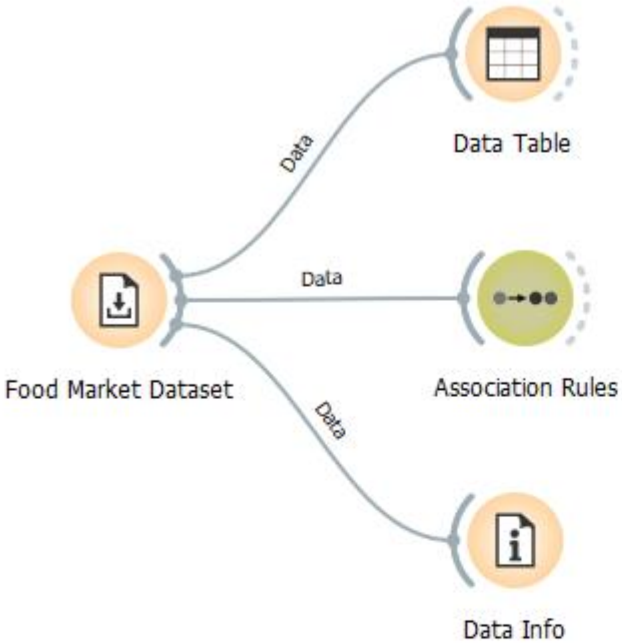
2) Generate association rules using Food mart Data set in Orange Tool. Compare the different measures to assess the quality of rules.

Input

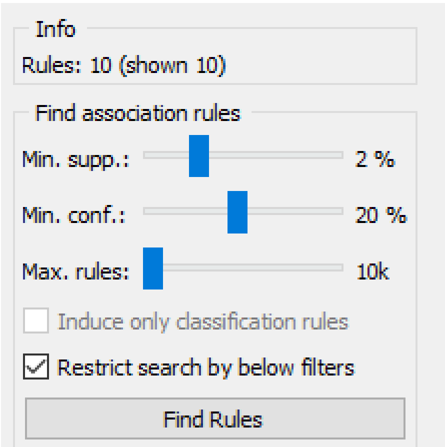


	{...}
62560	Frozen Vegetables=3, Clams=3, STORE_ID_24=1
62559	Flavored Drinks=4, Waffles=4, Canned Vegetables=3, Frozen Chicken=3, STORE_ID_24=1
62558	Soup=3, Fresh Vegetables=3, Donuts=3, STORE_ID_24=1
62557	Cleaners=4, Eggs=4, Fresh Fruit=3, Muffins=4, Tools=3, Sour Cream=5, Wine=5, STORE_ID_24=1
62556	Fresh Vegetables=3, Deli Meats=2, Flavored Drinks=3, Beer=3, Personal Hygiene=3, Lightbulbs=3, Computer M...
62555	Milk=3, Eggs=5, Paper Wipes=2, Cottage Cheese=3, Pot Scrubbers=4, STORE_ID_24=1
62554	Cereal=3, Fresh Fruit=3, Popcorn=4, Muffins=3, Candles=4, Chocolate=3, STORE_ID_24=1

Process



Output



Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.050	0.287	0.175	1.619	1.017	0.001	Fresh Fruit	→	Fresh Vegetables
0.035	0.299	0.117	2.421	1.059	0.002	Dried Fruit	→	Fresh Vegetables
0.035	0.293	0.119	2.375	1.035	0.001	Soup	→	Fresh Vegetables
0.031	0.262	0.118	2.405	0.926	-0.002	Cheese	→	Fresh Vegetables
0.028	0.279	0.099	2.854	0.987	-0.000	STORE_ID_13	→	Fresh Vegetables
0.027	0.260	0.105	2.691	0.921	-0.002	Cookies	→	Fresh Vegetables
0.025	0.278	0.089	3.160	0.982	-0.000	STORE_ID_17	→	Fresh Vegetables
0.022	0.284	0.079	3.577	1.004	0.000	Paper Wipes	→	Fresh Vegetables
0.022	0.278	0.078	3.625	0.985	-0.000	Canned Vegetables	→	Fresh Vegetables
0.020	0.253	0.080	3.523	0.894	-0.002	Wine	→	Fresh Vegetables

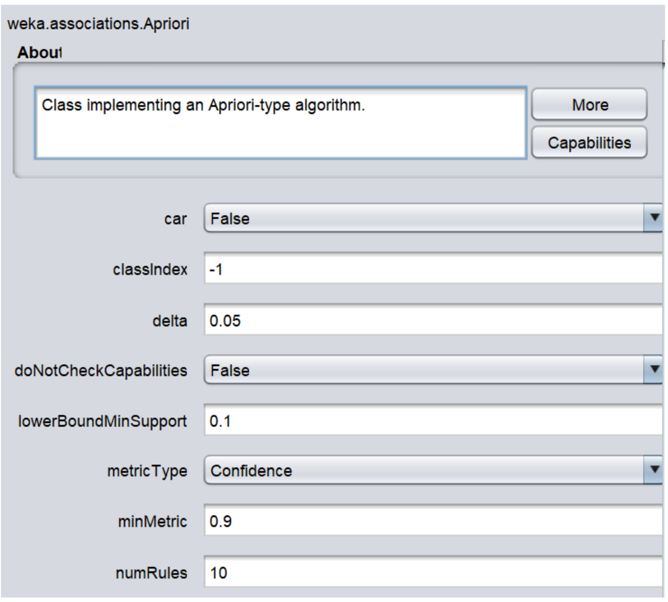
3) Generate association rules using supermarket Data set in WEKA using Apriori algorithm.

Input

Relation: supermarket-weka.filters.unsupervised.attribute.Remove-R1-9,11,57,70,79-81,88-89,98,100-102,107-114,116-120,

No.	1: grocery misc Nominal	2: baby needs Nominal	3: bread and cake Nominal	4: baking needs Nominal	5: coupons Nominal	6: juice-sat-cord-ms Nominal	7: tea Nominal	8: biscu Nomin
1		t	t	t		t		t
2								
3			t	t		t		t
4			t	t		t		t
5			t	t		t	t	
6			t	t		t	t	t
7			t	t		t	t	t

Process



Output

```
Size of set of large itemsets L(2): 498

Size of set of large itemsets L(3): 1959

Size of set of large itemsets L(4): 2888

Size of set of large itemsets L(5): 1679

Size of set of large itemsets L(6): 317

Size of set of large itemsets L(7): 11

Best rules found:

 1. biscuits=t frozen foods=t pet foods=t milk-cream=t vegetables=t 516 ==> bread and cake=t 475    <conf:(0.92)
 2. baking needs=t biscuits=t milk-cream=t margarine=t fruit=t vegetables=t 505 ==> bread and cake=t 464    <co
 3. biscuits=t frozen foods=t milk-cream=t margarine=t vegetables=t 585 ==> bread and cake=t 537    <conf:(0.92)
 4. biscuits=t canned vegetables=t frozen foods=t fruit=t vegetables=t 536 ==> bread and cake=t 492    <conf:(0
 5. baking needs=t frozen foods=t milk-cream=t margarine=t fruit=t vegetables=t 517 ==> bread and cake=t 474
 6. biscuits=t frozen foods=t pet foods=t milk-cream=t fruit=t 511 ==> bread and cake=t 468    <conf:(0.92)> li
 7. biscuits=t frozen foods=t tissues-paper prd=t milk-cream=t vegetables=t 575 ==> bread and cake=t 526    <co
 8. biscuits=t frozen foods=t beef=t fruit=t vegetables=t 536 ==> bread and cake=t 490    <conf:(0.91)> lift:(1
 9. baking needs=t biscuits=t frozen foods=t cheese=t fruit=t 538 ==> bread and cake=t 491    <conf:(0.91)> lif
10. biscuits=t frozen foods=t milk-cream=t margarine=t fruit=t 592 ==> bread and cake=t 540    <conf:(0.91)> li
```

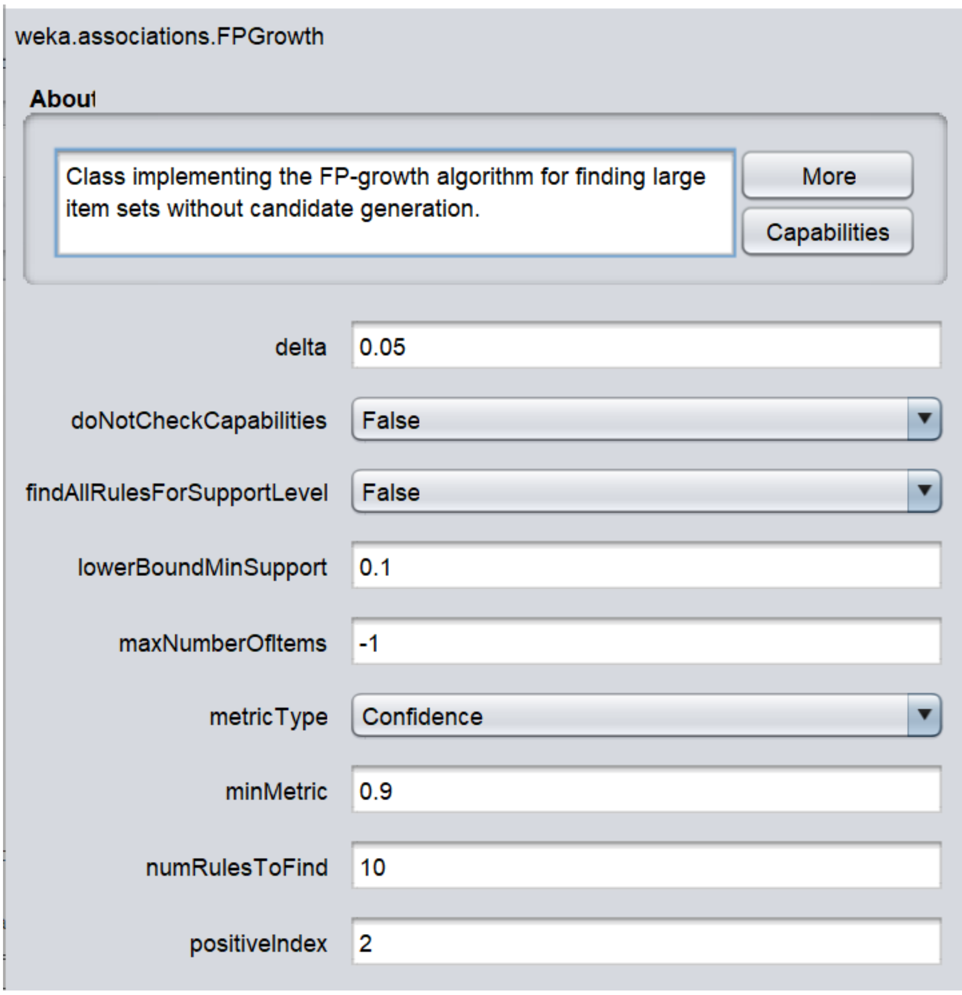
4) Generate association rules using supermarket Data set in WEKA using FP – growth Algorithm

Input

Relation: supermarket-weka.filters.unsupervised.attribute.Remove-R1-9,11,57,70,79-81,88-89,98,100-102,107-114,116-120,

No.	1: grocery misc Nominal	2: baby needs Nominal	3: bread and cake Nominal	4: baking needs Nominal	5: coupons Nominal	6: juice-sat-cord-ms Nominal	7: tea Nominal	8: biscu Nomin
1		t	t	t		t		t
2								
3			t	t		t		t
4			t	t		t		t
5			t	t		t	t	
6			t	t		t	t	t
7			t	t		t	t	t

Process



Output

```
=== Run information ===

Scheme:       weka.associations.FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1
Relation:     supermarket-weka.filters.unsupervised.attribute.Remove-R1-9,11,57,70,79-81,88-89,98,100-102,107-1
Instances:    4627
Attributes:   105
              [list of attributes omitted]
=== Associator model (full training set) ===

FPGrowth found 101 rules (displaying top 10)

 1. [vegetables=t, milk-cream=t, frozen foods=t, biscuits=t, pet foods=t]: 516 ==> [bread and cake=t]: 475   <c
 2. [fruit=t, vegetables=t, milk-cream=t, baking needs=t, biscuits=t, margarine=t]: 505 ==> [bread and cake=t]:
 3. [vegetables=t, milk-cream=t, frozen foods=t, biscuits=t, margarine=t]: 585 ==> [bread and cake=t]: 537   <c
 4. [fruit=t, vegetables=t, frozen foods=t, biscuits=t, canned vegetables=t]: 536 ==> [bread and cake=t]: 492
 5. [fruit=t, vegetables=t, milk-cream=t, baking needs=t, frozen foods=t, margarine=t]: 517 ==> [bread and cake
 6. [fruit=t, milk-cream=t, frozen foods=t, biscuits=t, pet foods=t]: 511 ==> [bread and cake=t]: 468   <conf:(
 7. [vegetables=t, milk-cream=t, frozen foods=t, biscuits=t, tissues-paper prd=t]: 575 ==> [bread and cake=t]:
 8. [fruit=t, vegetables=t, frozen foods=t, biscuits=t, beef=t]: 536 ==> [bread and cake=t]: 490   <conf:(0.91)
 9. [fruit=t, baking needs=t, frozen foods=t, biscuits=t, cheese=t]: 538 ==> [bread and cake=t]: 491   <conf:(0
10. [fruit=t, milk-cream=t, frozen foods=t, biscuits=t, margarine=t]: 592 ==> [bread and cake=t]: 540   <conf:(
```

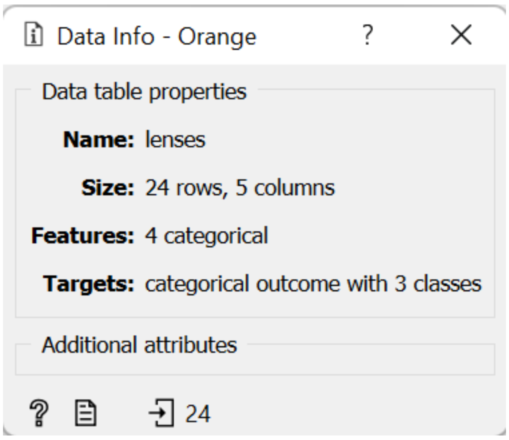
Interpretation

- There is a 92% probability that a person purchasing fruits,frozen food & biscuits may also purchase bread and cake.
- There is a 92% probability that a person purchasing fruits,baking needs and biscuits may also buy bread and cake.
- There is a 91% probability that a person buying fruits,snacks may also buy bread and cake.

b (Association Rule Mining -Class Work)

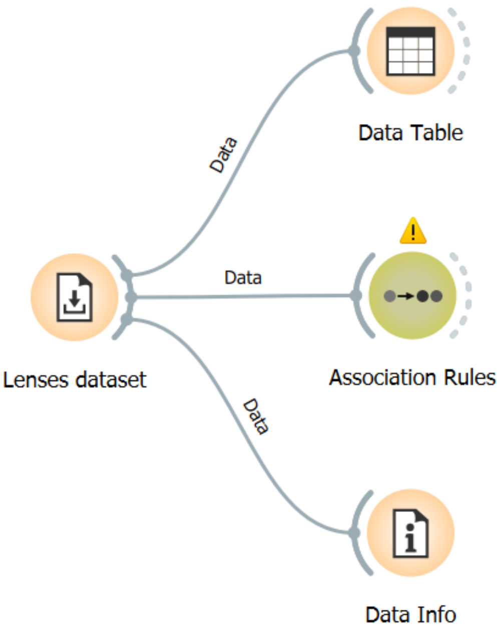
1) Generate association rules using Lenses Data set in Orange Tool

Input

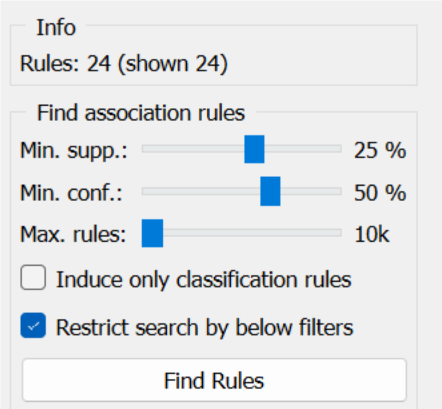


	lenses	age	prescription	astigmatic	tear_rate
1	none	young	myope	no	reduced
2	soft	young	myope	no	normal
3	none	young	myope	yes	reduced
4	hard	young	myope	yes	normal
5	none	young	hypermetrope	no	reduced

Process



Output



Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		
0.250	0.500	0.500	1.000	1.000	0.000	astigmatic=no	→	prescription=hypermetrope
0.250	0.500	0.500	1.000	1.000	0.000	prescription=hypermetrope	→	astigmatic=no
0.250	0.500	0.500	1.000	1.000	0.000	astigmatic=no	→	prescription=myope
0.250	0.500	0.500	1.000	1.000	0.000	prescription=myope	→	astigmatic=no
0.250	0.500	0.500	1.000	1.000	0.000	astigmatic=yes	→	prescription=hypermetrope
0.250	0.500	0.500	1.000	1.000	0.000	prescription=hypermetrope	→	astigmatic=yes
0.250	0.500	0.500	1.000	1.000	0.000	astigmatic=yes	→	prescription=myope

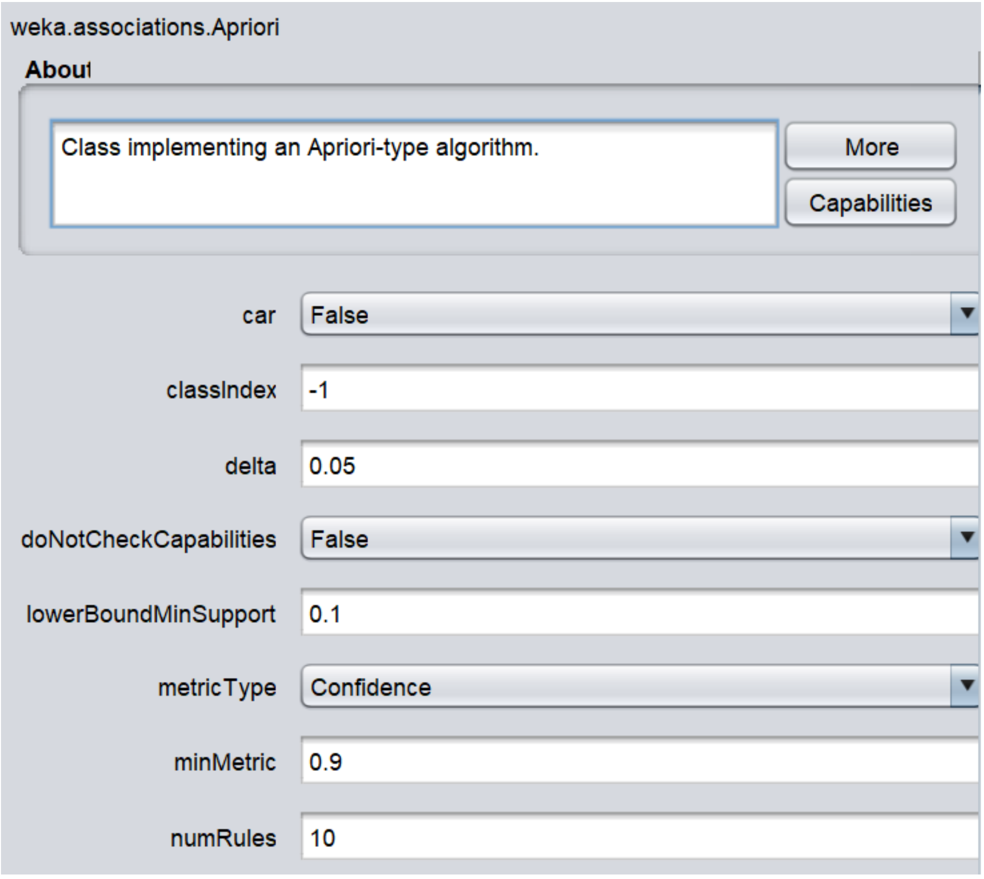
2) Generate association rules using Lenses Data set in WEKA using Apriori algorithm.

Input

Relation: contact-lenses

No.	1: age Nominal	2: spectacle-prescrip Nominal	3: astigmatism Nominal	4: tear-prod-rate Nominal	5: **contact-lenses** Nominal
1	young	myope	no	reduced	none
2	young	myope	no	normal	soft
3	young	myope	yes	reduced	none
4	young	myope	yes	normal	hard
5	young	hypermetrope	no	reduced	none
6	young	hypermetrope	no	normal	soft
7	young	hypermetrope	yes	reduced	none

Process



Output

```
Minimum support: 0.2 (5 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsets L(1): 11

Size of set of large itemsets L(2): 21

Size of set of large itemsets L(3): 6

Best rules found:

 1. tear-prod-rate=reduced 12 ==> contact-lenses=none 12    <conf:(1)> lift:(1.6) lev:(0.19) [4] conv:(4.5)
 2. spectacle-prescrip=myope tear-prod-rate=reduced 6 ==> contact-lenses=none 6    <conf:(1)> lift:(1.6) lev:(0.
 3. spectacle-prescrip=hypermetrope tear-prod-rate=reduced 6 ==> contact-lenses=none 6    <conf:(1)> lift:(1.6)
 4. astigmatism=no tear-prod-rate=reduced 6 ==> contact-lenses=none 6    <conf:(1)> lift:(1.6) lev:(0.09) [2] cc
 5. astigmatism=yes tear-prod-rate=reduced 6 ==> contact-lenses=none 6    <conf:(1)> lift:(1.6) lev:(0.09) [2] c
 6. contact-lenses=soft 5 ==> astigmatism=no 5    <conf:(1)> lift:(2) lev:(0.1) [2] conv:(2.5)
 7. contact-lenses=soft 5 ==> tear-prod-rate=normal 5    <conf:(1)> lift:(2) lev:(0.1) [2] conv:(2.5)
 8. tear-prod-rate=normal contact-lenses=soft 5 ==> astigmatism=no 5    <conf:(1)> lift:(2) lev:(0.1) [2] conv:(
 9. astigmatism=no contact-lenses=soft 5 ==> tear-prod-rate=normal 5    <conf:(1)> lift:(2) lev:(0.1) [2] conv:(
10. contact-lenses=soft 5 ==> astigmatism=no tear-prod-rate=normal 5    <conf:(1)> lift:(4) lev:(0.16) [3] conv:
```

Interpretation

- Tear production rate has been reduced significantly with a support of 100%

3) Generate association rules using Lenses Data set in WEKA using FP – growth Algorithm

Relation: contact-lenses

No.	1: age Nominal	2: spectacle-prescrip Nominal	3: astigmatism Nominal	4: tear-prod-rate Nominal	5: contact-lenses Nominal
1	young	myope	no	reduced	none
2	young	myope	no	normal	soft
3	young	myope	yes	reduced	none
4	young	myope	yes	normal	hard
5	young	hypermetrope	no	reduced	none
6	young	hypermetrope	no	normal	soft
7	young	hypermetrope	yes	reduced	none