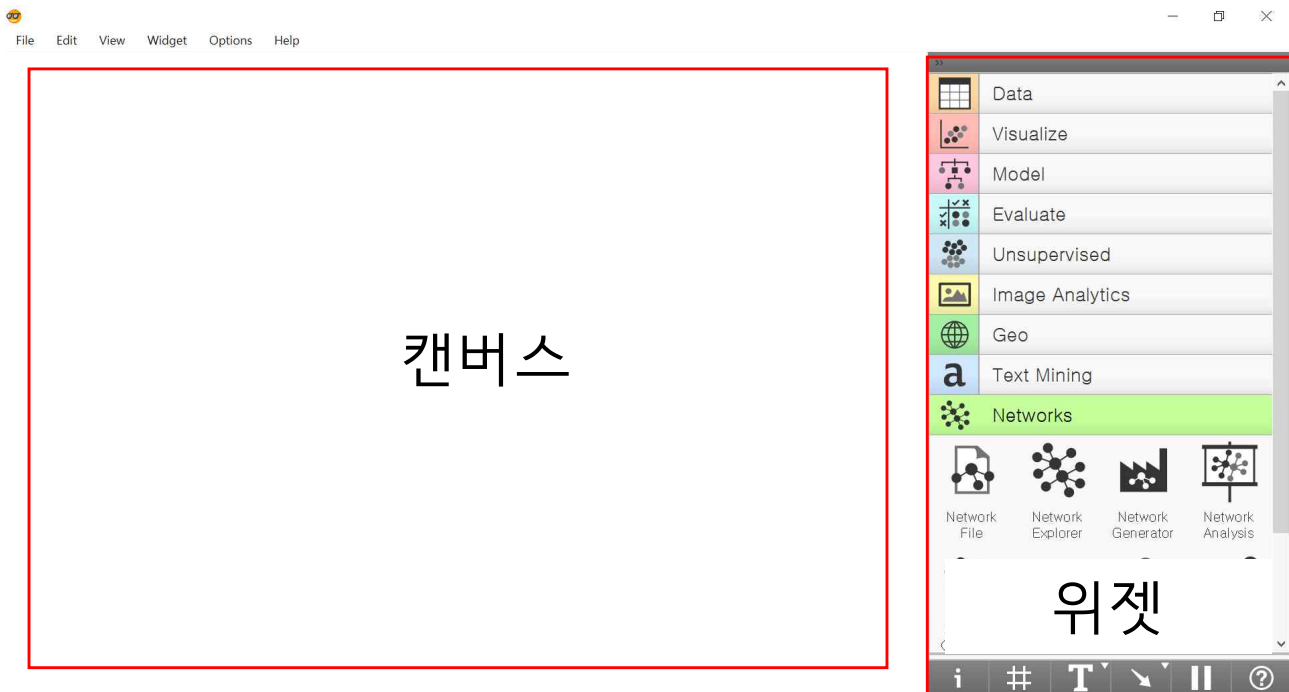


# Orange3로 머신러닝 모델 만들기

## 1 Orange3란?

- 오렌지3은 오픈소스 프로그램으로 원하는 위젯을 선택하고, 서로 연결하여 간단히 머신러닝 모델을 만들 수 있는 프로그램입니다. azure와는 달리 설치해서 사용하는 프로그램이고, 필요에 따라 업데이트를 하거나 add-on을 추가하여 이미지 분석, 텍스트 마이닝, 네트워크 분석, 지도를 활용한 데이터 분석 등 다양한 분야에서 활용할 수 있습니다. 카테고리는 기본적으로 데이터, 시각화, 지도/비지도 알고리즘, 평가로 구성되어 있고 위젯의 기능을 설명해주는 아이콘으로 구성되어 있어 좀 더 이해하기 쉽고 친숙하게 사용할 수 있습니다.
- 위젯을 클릭하거나 캔버스에 끌어올 수 있고, 캔버스에서 마우스 우클릭하여 원하는 위젯을 선택할 수도 있습니다.
- 더 다양한 위젯을 사용하기 위해서는 [Options] - [add-ons..]을 클릭하여 부가 기능을 설치할 수 있습니다.

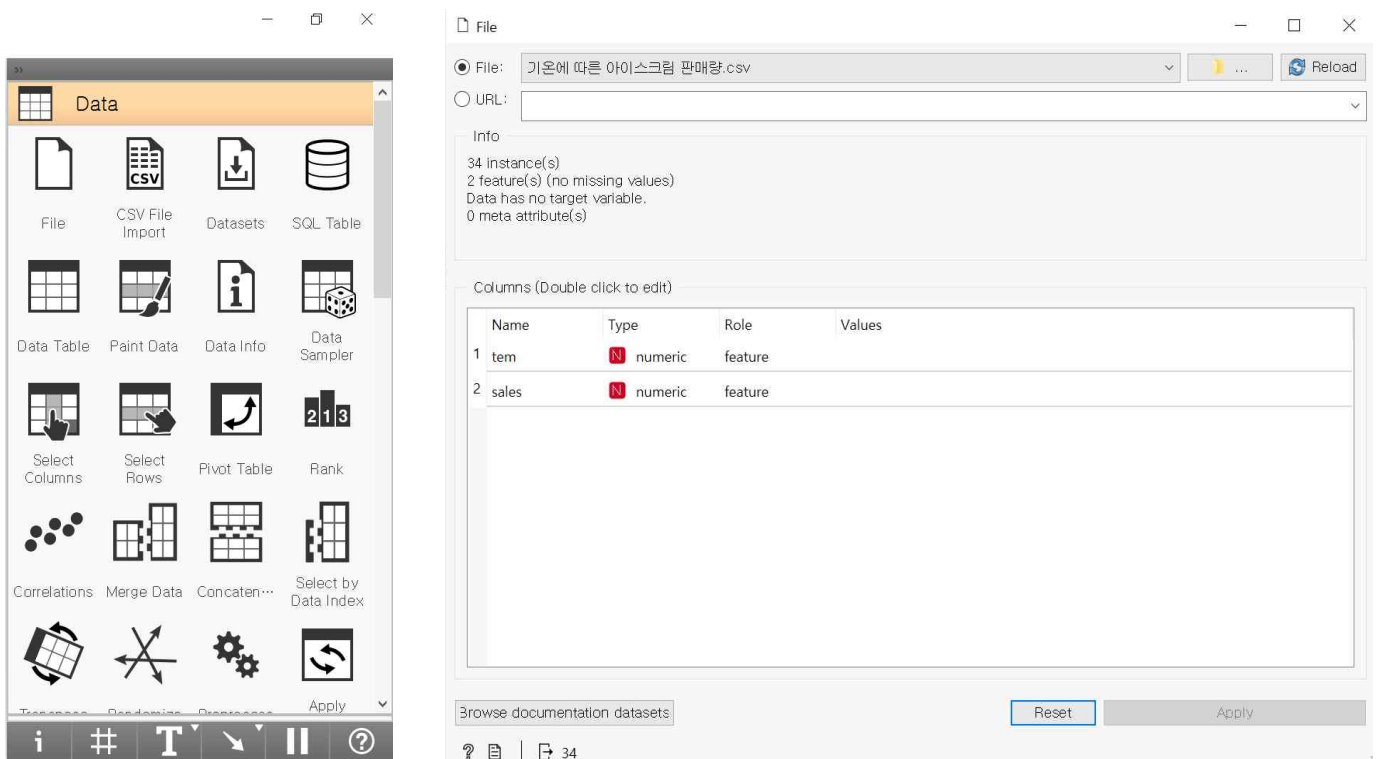


## 2 Orange3 무작정 따라하기

### 기온에 따른 아이스크림 판매량 예측 모델 만들기 [Linear Regression]

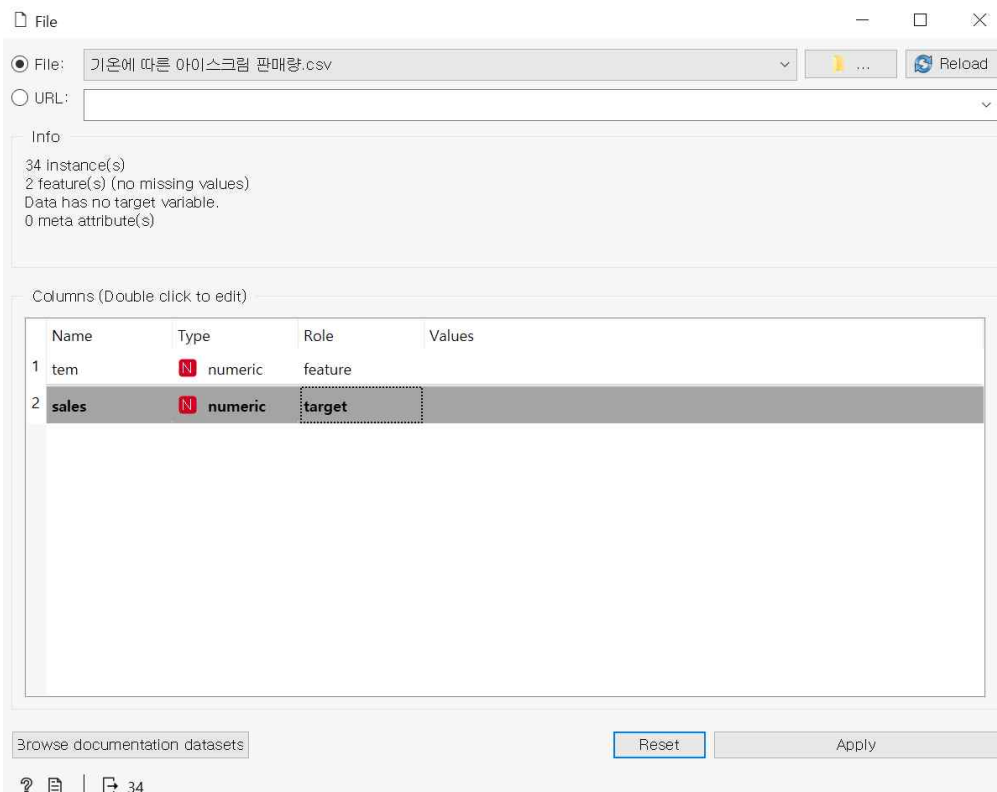
#### 1. 데이터 가져오기

- [Data] - [File] 위젯을 클릭합니다.
- 파일 위젯을 더블 클릭하고, “기온에 따른 아이스크림 판매량” 파일을 불러옵니다.



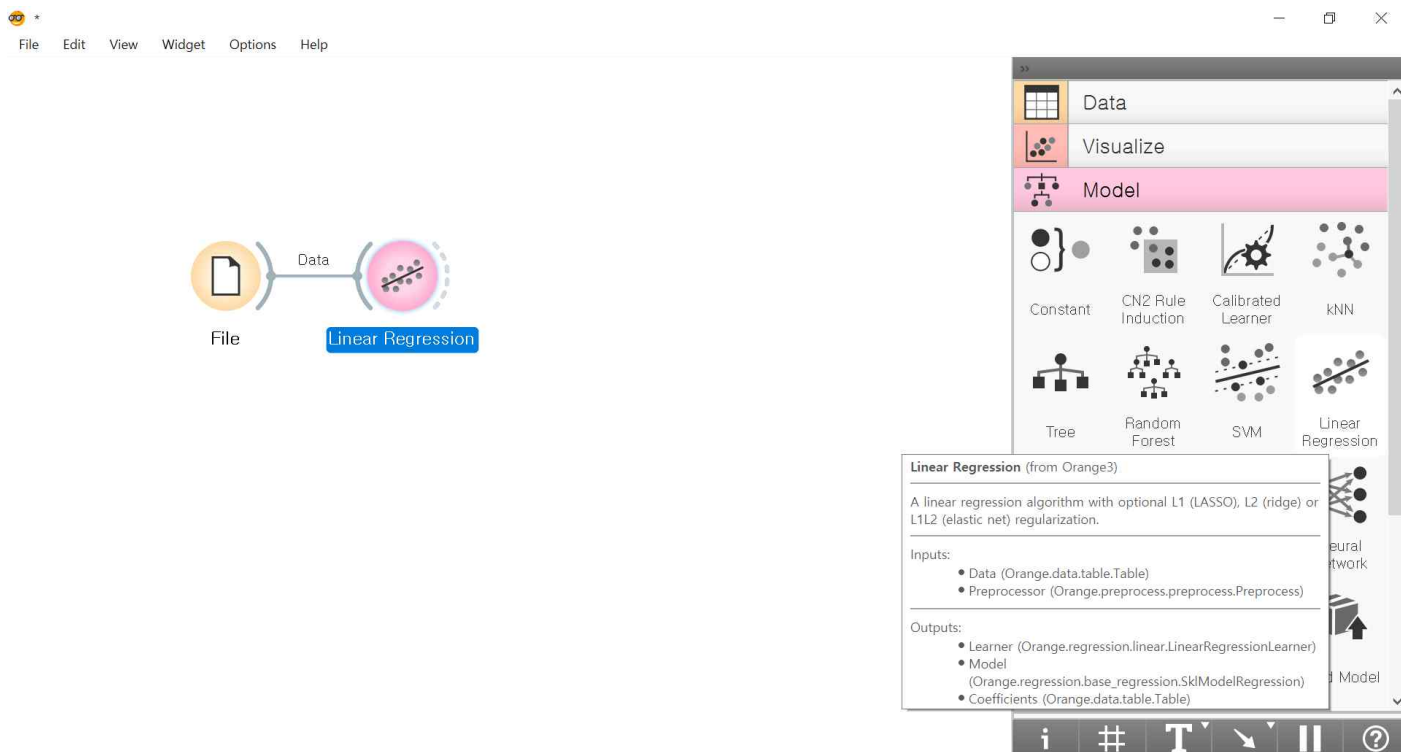
## 2. 데이터 전처리하기

- 예측하고자 하는 값이 판매량(sales)이므로 feature로 되어 있는 판매량의 Role을 클릭하여 [target]으로 바꿔줍니다. 그리고 Apply를 클릭합니다. 설정이 완료됐으면 닫기를 누릅니다.



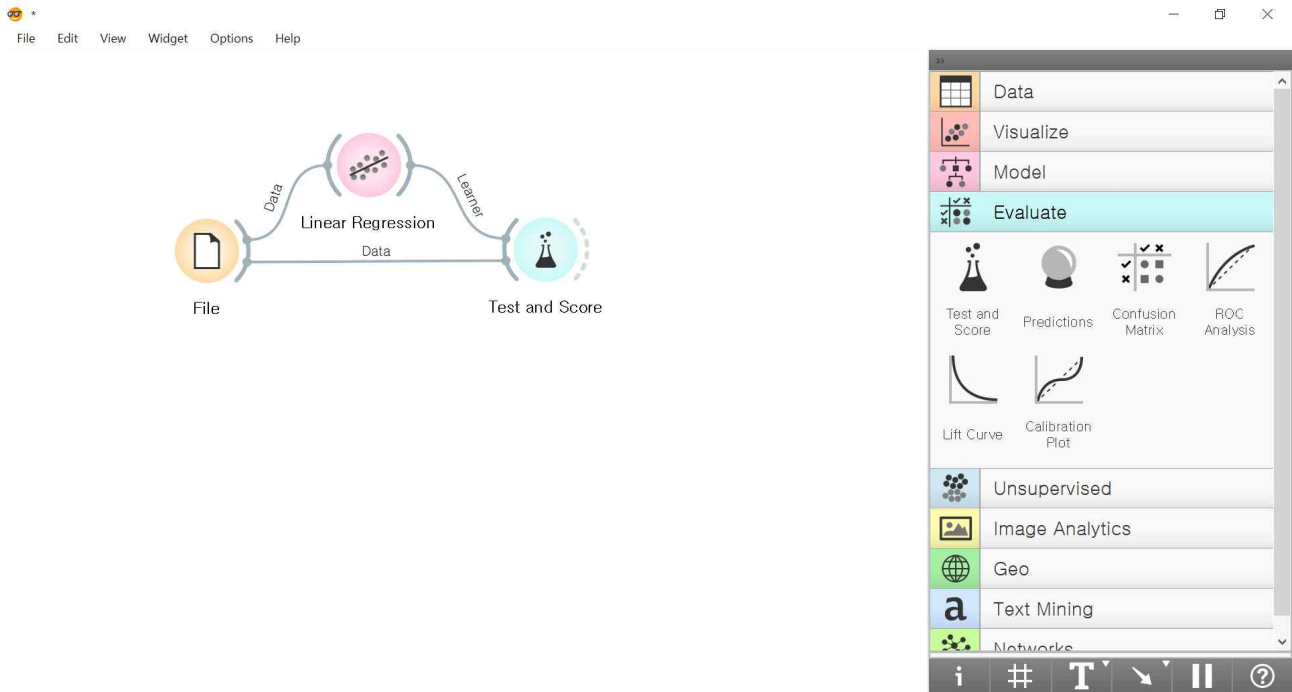
## 3. 알고리즘 선택

- [Model] 카테고리를 클릭합니다.
- [Linear regression] 알고리즘을 선택합니다.
- 파일 위젯과 Linear regression 알고리즘 위젯을 연결합니다.



#### 4. 모델 평가하기

- [Evaluate] 카테고리를 선택합니다.
- [Test and Score] 위젯을 클릭합니다.
- Linear regression 위젯과 Test and Score 위젯을 연결합니다. 학습된 모델이 Test and Score 위젯과 연결되었습니다.
- 그리고 파일 위젯과 Test and Score 위젯을 연결합니다. 모델을 테스트하기 위한 데이터가 연결되었습니다.



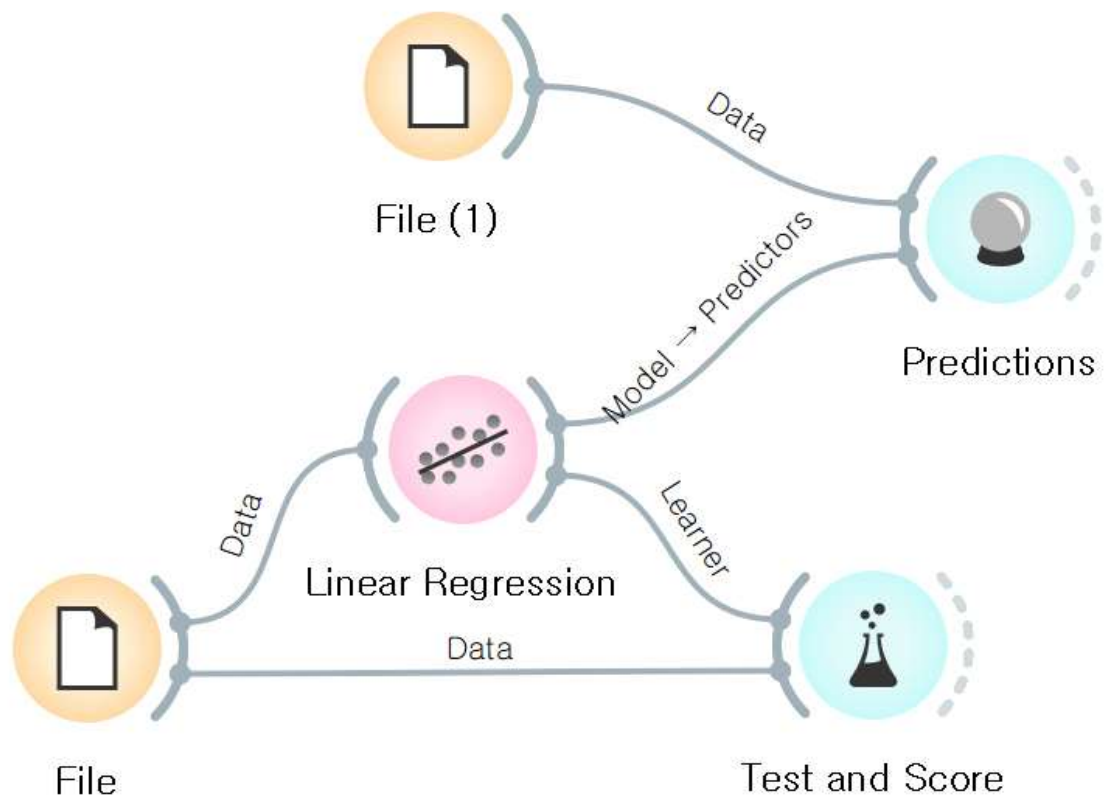
- [Test and Score] 위젯을 더블클릭하여 샘플링 방법을 “Random sampling”으로 선택합니다.
- 그리고 평가 지표를 확인합니다.

The screenshot shows the 'Test and Score' widget configuration window. The 'Sampling' section is active, showing options for 'Cross validation' and 'Random sampling'. The 'Random sampling' option is selected. The 'Evaluation Results' section displays a table with the following data:

Model	MSE	RMSE	MAE	R2
Linear Regression	636.376	25.226	18.420	0.826

## 5. 예측하고자 하는 데이터 불러오기

- 기존만 입력한 “아이스크림 테스트.csv” 파일을 파일 위젯을 사용하여 불러옵니다.
- [Evaluate] 카테고리에서 [Predictions] 위젯을 선택합니다.
- 그리고 새로운 테스트 파일과 [Predictions] 위젯을 연결하면 데이터가 입력됩니다. Linear Regression 위젯과 Predictions 위젯을 연결하여 학습된 모델을 [Predictions] 위젯에 연결합니다.
- Predictions에 훈련된 모델과 새로운 데이터 연결이 다 되었습니다.



- [Predictions] 위젯을 더블클릭하여 모델이 예측한 기온에 따른 판매량을 확인합니다.

	Linear Regression	tem
1	24	5
2	41	8
3	52	10
4	68	13
5	80	15
- 6	96	18
7	107	20

# 타이타닉 생존자 예측(분류)하기

## [Logistic Regression]

### 1. 데이터 불러오기

- [Data] 카테고리 - [File]위젯을 클릭합니다.
- “타이타닉.csv” 데이터셋을 가져옵니다.

### 2. 데이터 전처리하기

- **데이터 유형(Type) 수정하기**: 대부분 데이터의 유형이 자동으로 선택되지만 바꿔줘야 할 데이터가 있다면 적절한 유형으로 바꿔줘야 합니다.















1) “categorical”(범주형) 로 설정: pclass / survived / gender / embarked

2) “numeric”(숫자): age / sibsp / parch / fare

3) “meta”(참고만 할 예정): name / ticket / cabin / home.dest

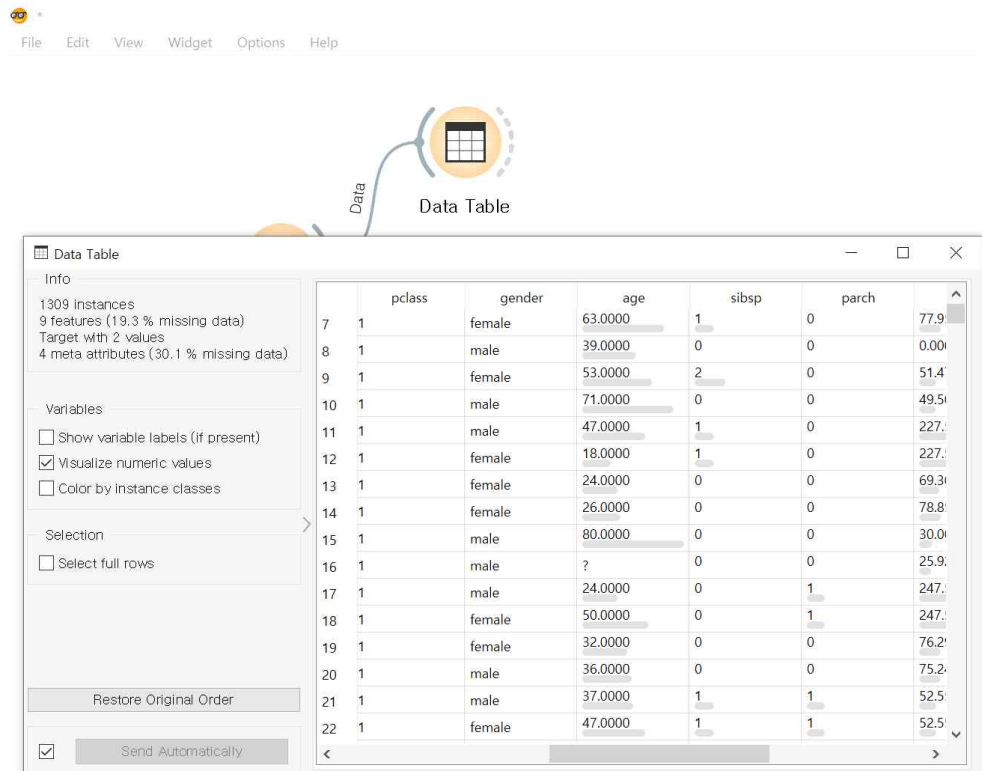
\*boat/body 비어있는 값이 너무 많아 삭제

- 예측하고자 하는 값이 생존 여부이기 때문에 “survived” 데이터의 역할(Role)을 [target]으로 바꿔줍니다.

	Name	Type	Role	Values
1	<b>pclass</b>	 <b>categorical</b>	<b>feature</b>	
2	<b>survived</b>	 <b>categorical</b>	<b>target</b>	<b>0, 1</b>
3	gender	 categorical	feature	female, male
4	age	 numeric	feature	
5	sibsp	 numeric	feature	
6	parch	 numeric	feature	
7	fare	 numeric	feature	
8	embarked	 categorical	feature	C, Q, S
9	boat	 categorical	feature	1, 2, 3, 4, 5, 5 7, 5 9, 6, 7, 8, 8 10, 9, 10, 11, 12, 13, 13 15, 13 15 B, 14, 15, ...
10	body	 numeric	feature	
11	name	 text	meta	
12	ticket	 text	meta	
13	cabin	 text	meta	
14	home.dest	 text	meta	

## - 비어있는 값 처리하기

1) [Data] - [Data table] 위젯을 클릭하고, [Data Table] 위젯을 더블클릭합니다. 그러면 업로드된 데이터를 표 형식으로 확인할 수 있습니다. 데이터를 확인해보니 비어있는 값이 대략 19.3%입니다. 비어있는 값이 있을 경우 모델의 훈련 결과에 영향을 미칠 수 있기 때문에 비어있는 값을 평균/최빈값으로 채워보겠습니다.



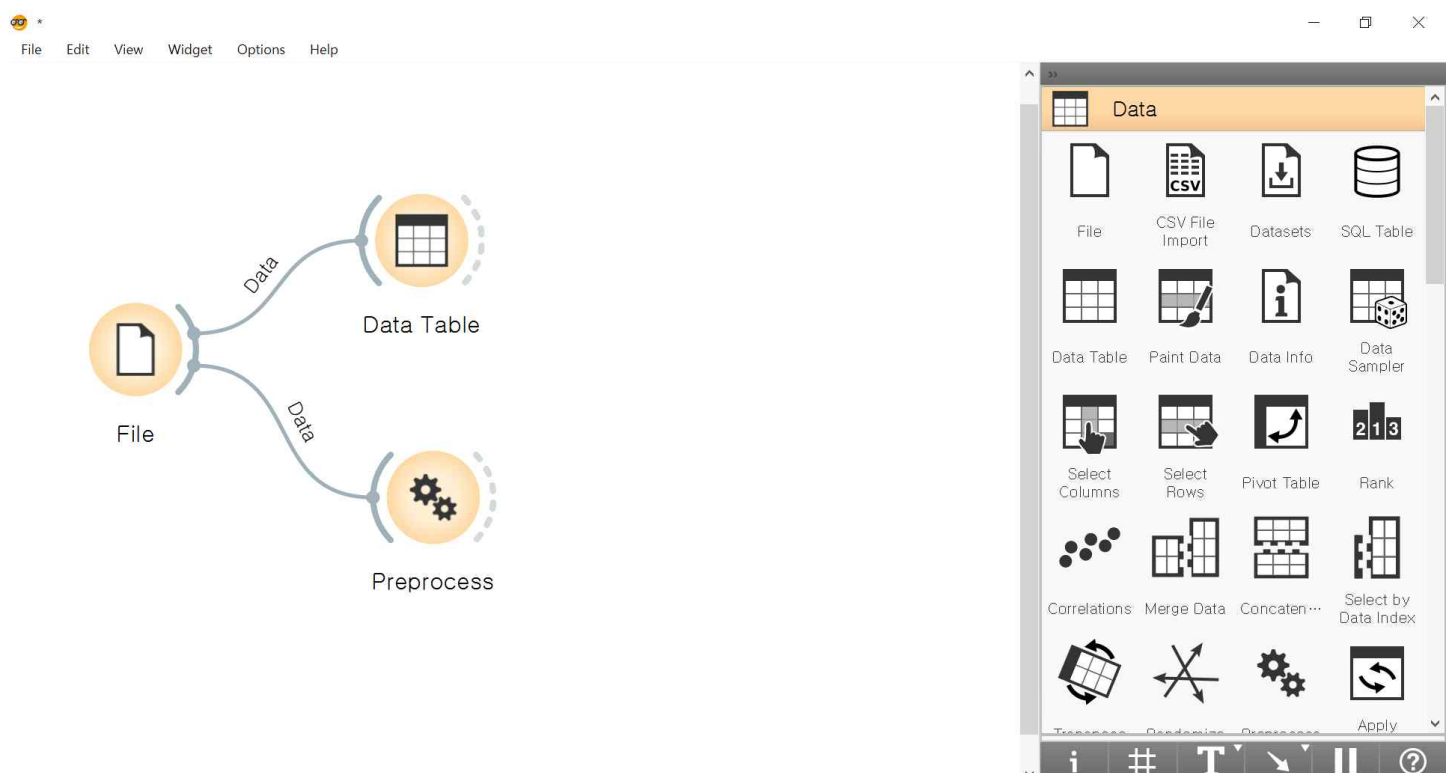
The screenshot shows the 'Data Table' widget interface. The left sidebar contains the following information:

- Info:** 1309 instances, 9 features (19.3 % missing data), Target with 2 values, 4 meta attributes (30.1 % missing data)
- Variables:**
  - ☐ Show variable labels (if present)
  - ☒ Visualize numeric values
  - ☐ Color by instance classes
- Selection:**
  - ☐ Select full rows
- Buttons:** Restore Original Order, Send Automatically (checked)

The main data table displays the following data:

	pclass	gender	age	sibsp	parch	target
7	1	female	63.0000	1	0	77.9
8	1	male	39.0000	0	0	0.00
9	1	female	53.0000	2	0	51.4
10	1	male	71.0000	0	0	49.5
11	1	male	47.0000	1	0	227.
12	1	female	18.0000	1	0	227.
13	1	female	24.0000	0	0	69.3
14	1	female	26.0000	0	0	78.8
15	1	male	80.0000	0	0	30.0
16	1	male	?	0	0	25.9
17	1	male	24.0000	0	1	247.
18	1	female	50.0000	0	1	247.
19	1	female	32.0000	0	0	76.2
20	1	male	36.0000	0	0	75.2
21	1	male	37.0000	1	1	52.5
22	1	female	47.0000	1	1	52.5

2) [Data] - [Preprocess=전처리] 위젯을 선택합니다. 그리고 해당 위젯을 더블클릭합니다.



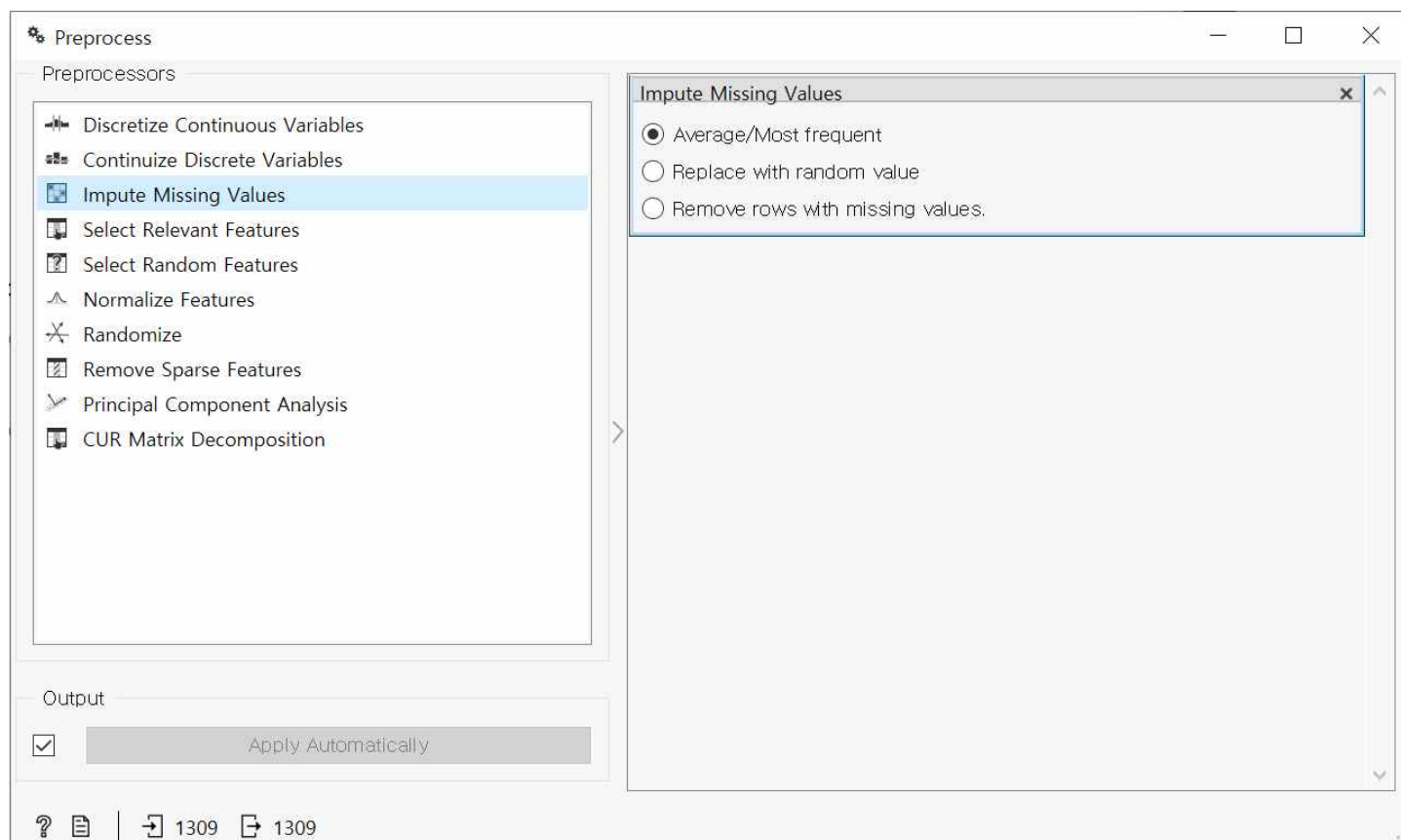
The screenshot shows the 'Data' widget interface. The left sidebar contains the following information:

- Info:** 1309 instances, 9 features (19.3 % missing data), Target with 2 values, 4 meta attributes (30.1 % missing data)
- Variables:**
  - ☐ Show variable labels (if present)
  - ☒ Visualize numeric values
  - ☐ Color by instance classes
- Selection:**
  - ☐ Select full rows
- Buttons:** Restore Original Order, Send Automatically (checked)

The main data table displays the following data:

	pclass	gender	age	sibsp	parch	target
7	1	female	63.0000	1	0	77.9
8	1	male	39.0000	0	0	0.00
9	1	female	53.0000	2	0	51.4
10	1	male	71.0000	0	0	49.5
11	1	male	47.0000	1	0	227.
12	1	female	18.0000	1	0	227.
13	1	female	24.0000	0	0	69.3
14	1	female	26.0000	0	0	78.8
15	1	male	80.0000	0	0	30.0
16	1	male	?	0	0	25.9
17	1	male	24.0000	0	1	247.
18	1	female	50.0000	0	1	247.
19	1	female	32.0000	0	0	76.2
20	1	male	36.0000	0	0	75.2
21	1	male	37.0000	1	1	52.5
22	1	female	47.0000	1	1	52.5

3) 데이터 전처리 기능 중 [Impute Missing Values]를 더블클릭합니다. 그러면 오른쪽에 해당 항목이 표시됩니다. 그 중 이 모델에서는 “Average/Most frequent”로 설정하여 빈 값을 채우도록 하겠습니다. 선택이 다 완료됐다면 해당 창을 닫고, 다시 캔버스로 돌아옵니다.



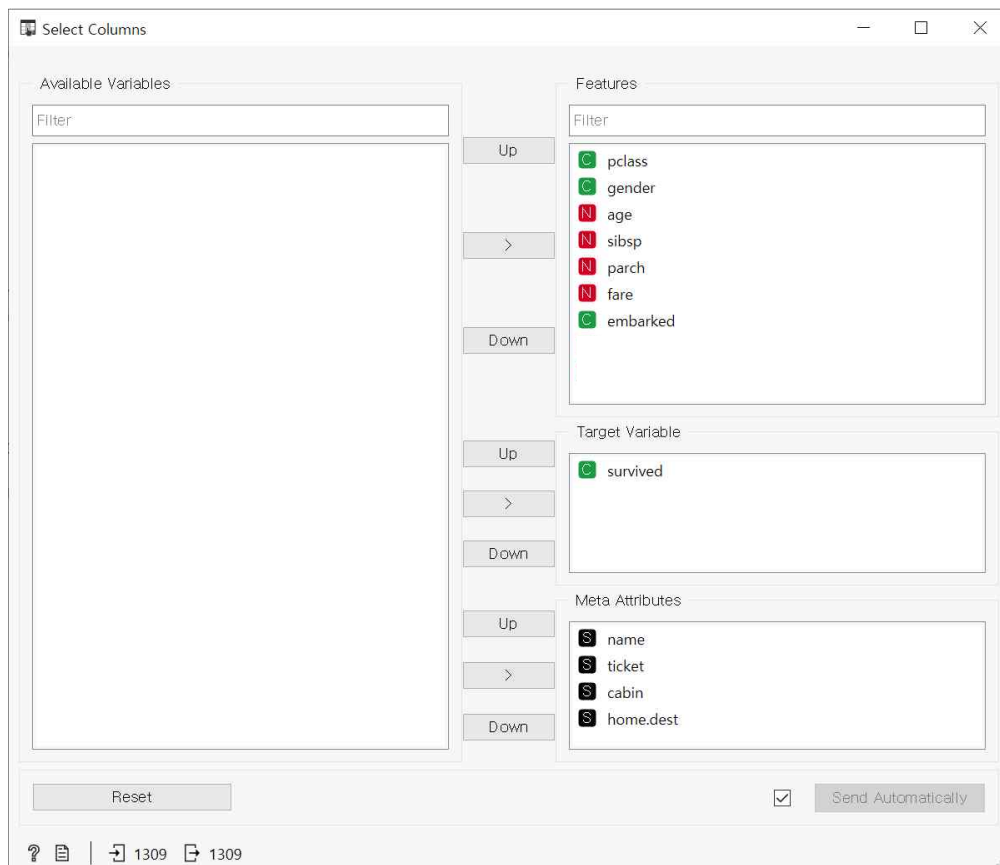
### 3. 열 선택하기

- 전처리한 데이터에서 모델 학습에 사용할 데이터를 선택해야 합니다. [Data] - [Select Columns] 위젯을 선택합니다.



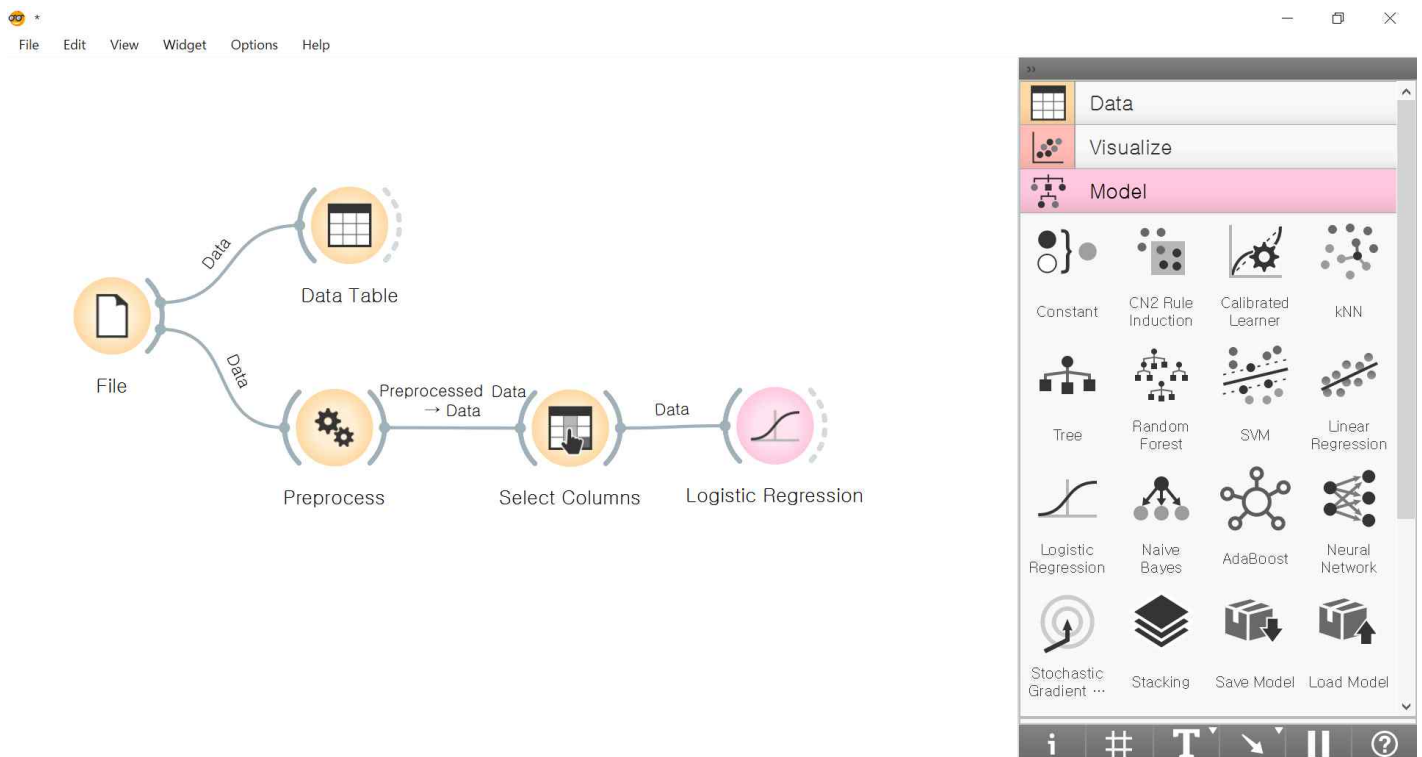


- 그리고 해당 위젯을 더블클릭하여 모델 학습에 사용할 열과 예측하고 싶은 feature를 아래와 같이 선택합니다.



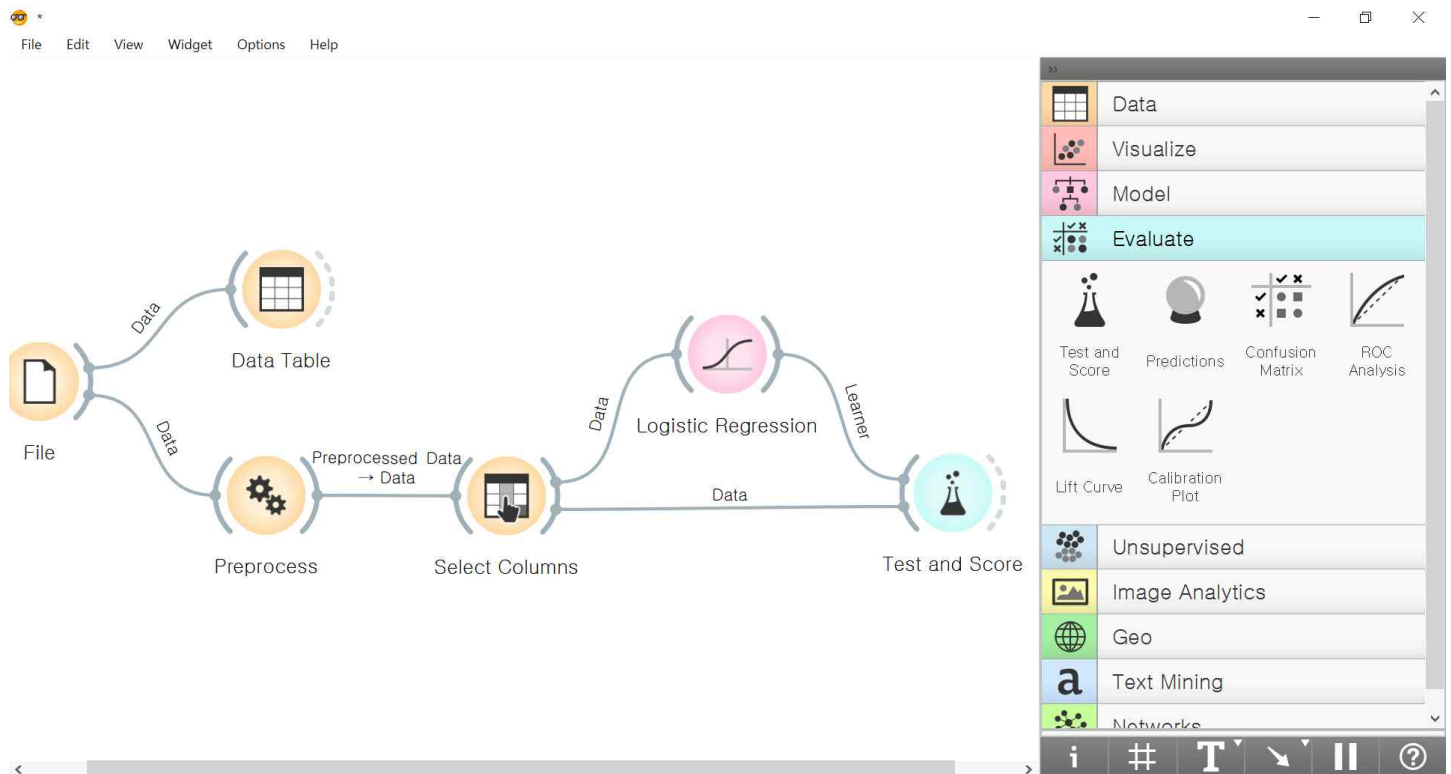
#### 4. 알고리즘 선택하기

- [Model] - [Logistic Regression] 위젯을 선택하고, "Select Columns"위젯과 알고리즘 위젯을 연결합니다.



## 5. 모델 평가하기

- [Evaluate] - [Test and Score] 위젯을 선택하고 로지스틱 회귀 위젯과 연결합니다.
- 그리고 모델 평가에 필요한 데이터 입력을 위해 [Select columns] 위젯과 연결합니다.

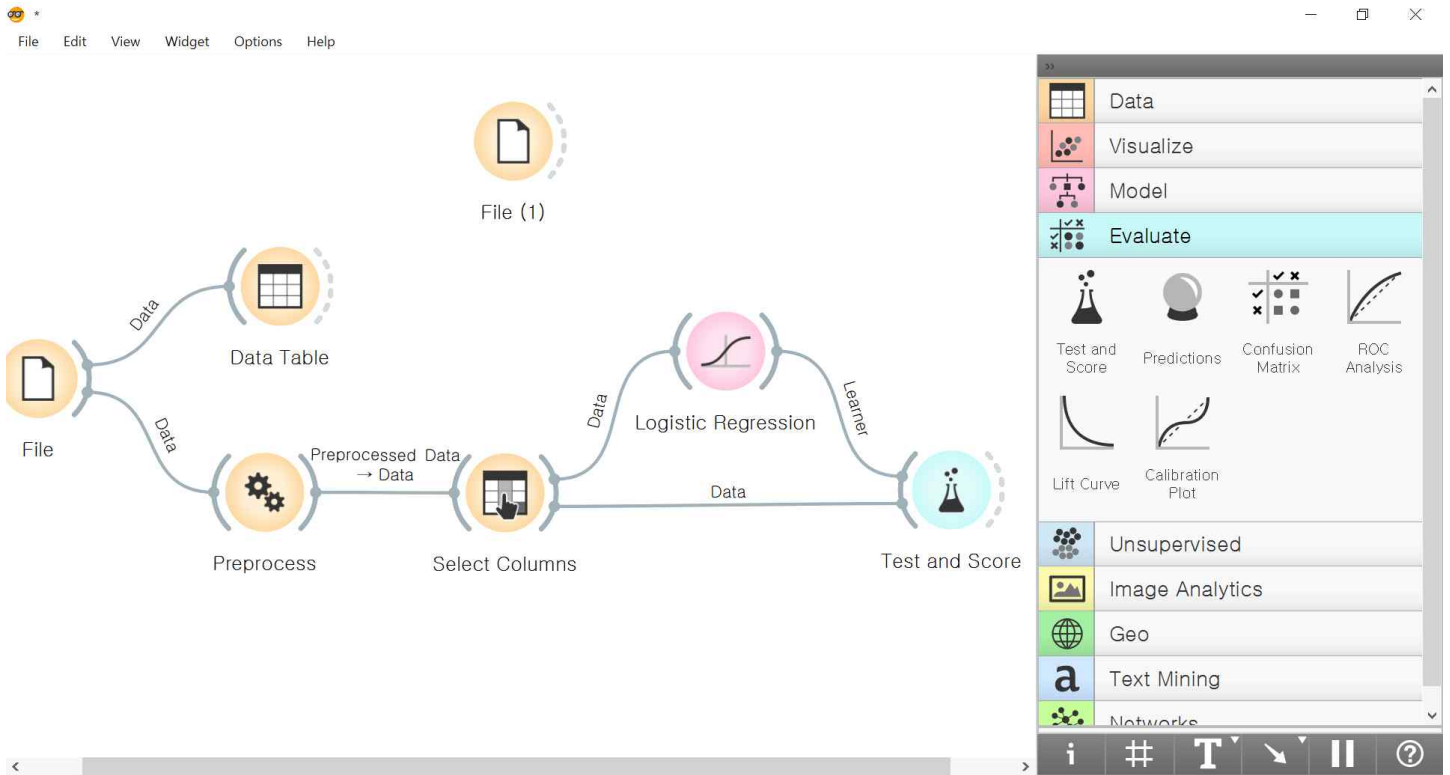


- [Test and Score] 위젯을 더블클릭하여 모델 평가 결과를 확인할 수 있습니다.

Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.974	0.939	0.939	0.942	0.939

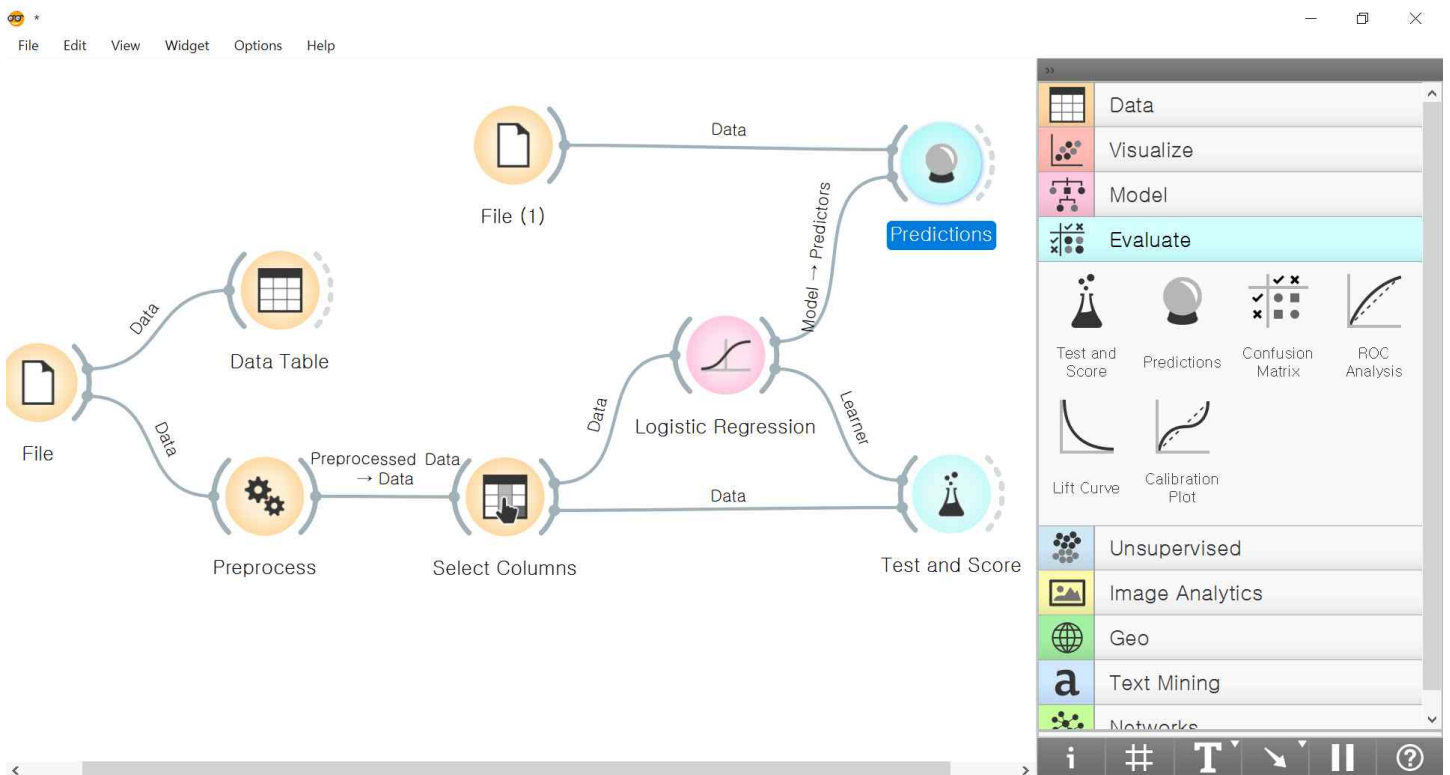
### - 새로운 데이터 예측해보기

- 1) 예측하고 싶은 데이터인 "타이타닉 테스트 데이터.csv"를 [File]위젯으로 불러옵니다.



2) 새로운 데이터의 예측 결과를 확인하기 위해 [Evaluate] - [Predictions] 위젯을 선택합니다. 업로드된 데이터 파일 위젯과 [Predictions] 위젯을 연결하고, 훈련된 모델인 [Logistic Regression] 위젯과 [Predictions] 위젯을 연결합니다.

- [Predictions] 위젯을 더블클릭하여 새로운 데이터에 입력한 사람들의 생존 여부를 확인할 수 있습니다.



# 키, 몸무게에 따른 체형 분류하기

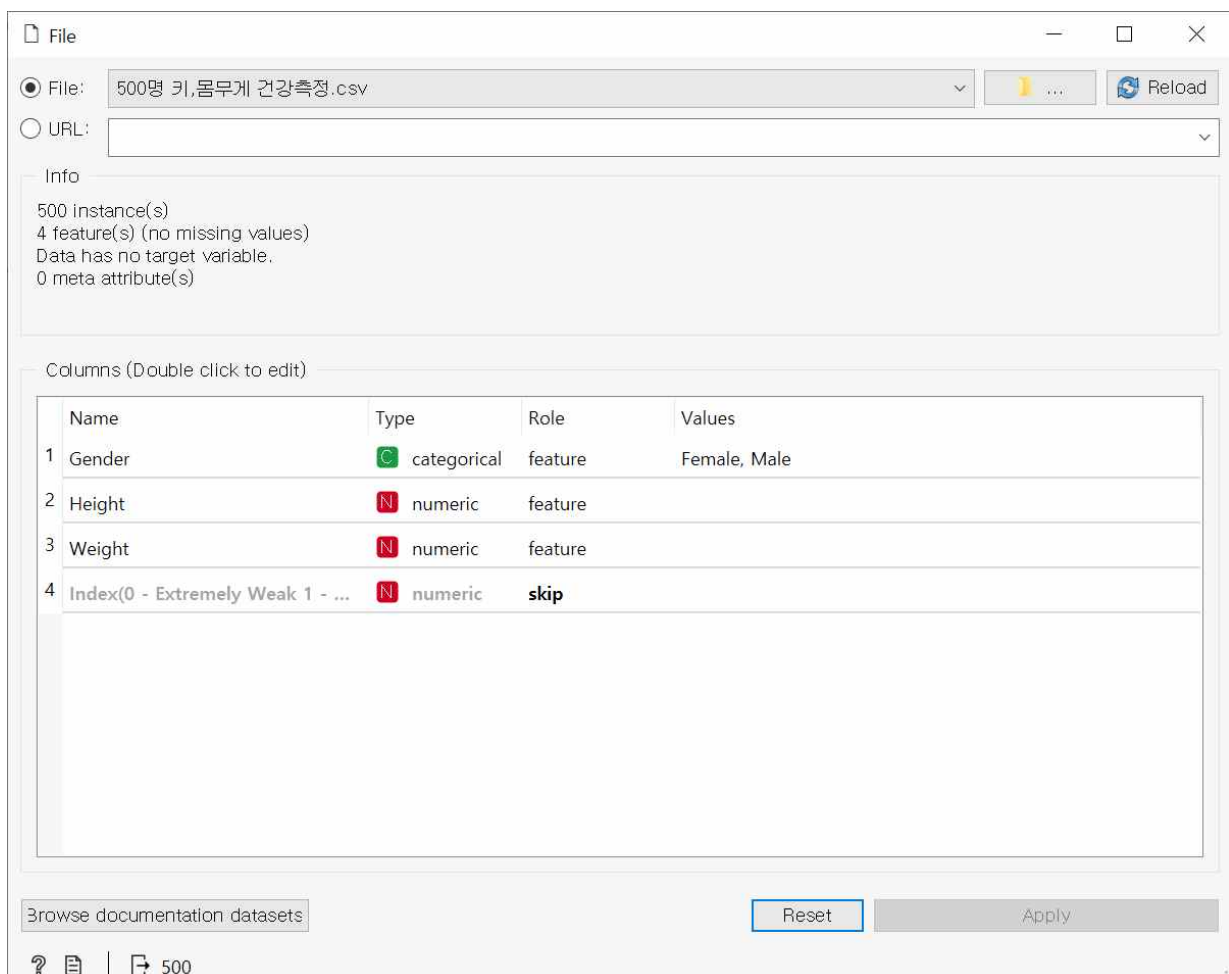
## [K-means]

### 1. 데이터 가져오기

- [Data] 카테고리 - [File] 위젯을 선택합니다.
- 파일 위젯을 더블클릭합니다.
- “500명 키, 몸무게 건강측정.csv” 파일을 불러옵니다.

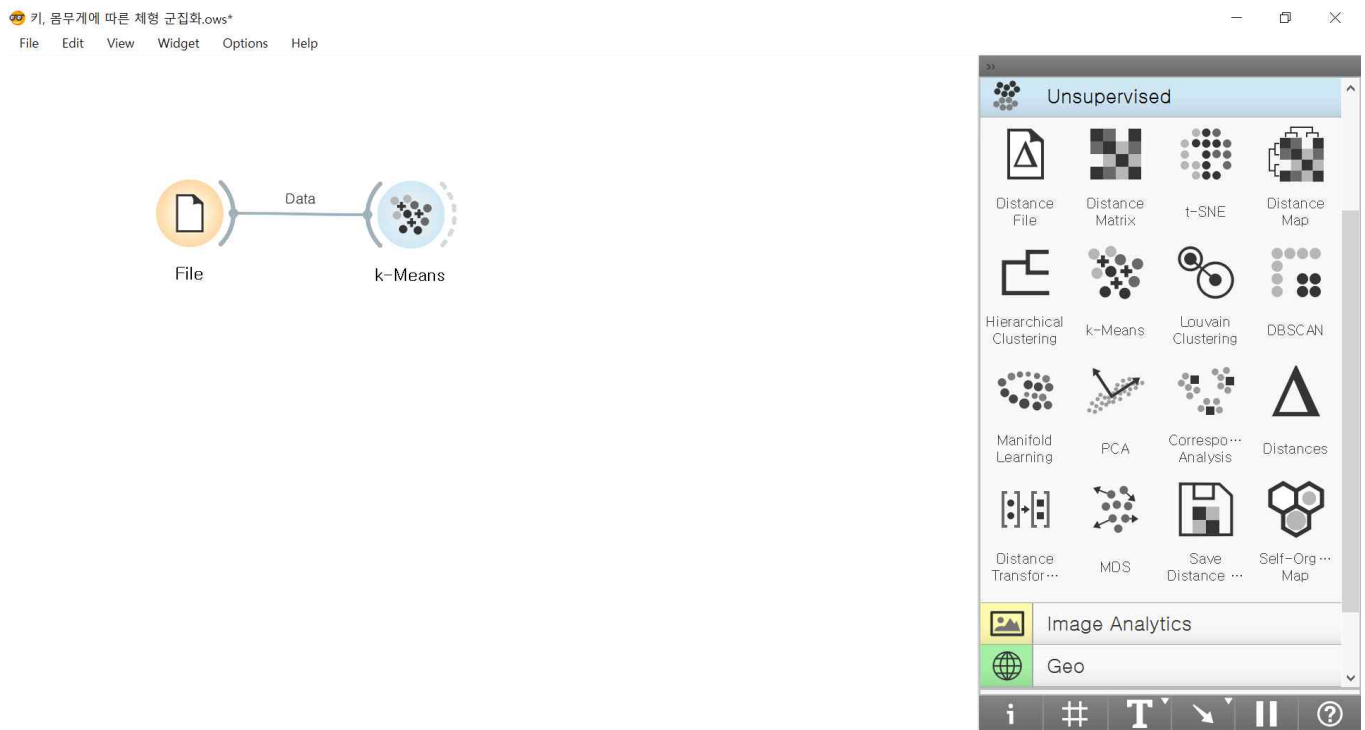
### 2. 데이터 전처리하기

- 각 데이터(feature)의 데이터 유형(Type)을 바꿔줍니다.
  - 1) Gender: categorical
  - 2) Height: numerical
  - 3) Weight: numerical
  - 4) Index: skip(사용하지 않음)
- 유형을 다 바꾼 후, Apply를 클릭하면 파일 업로드 완료! 데이터의 유형도 다 설정 완료했습니다.



### 3. 알고리즘 선택하기

- [Unsupervised] - [K-means] 위젯을 클릭합니다.



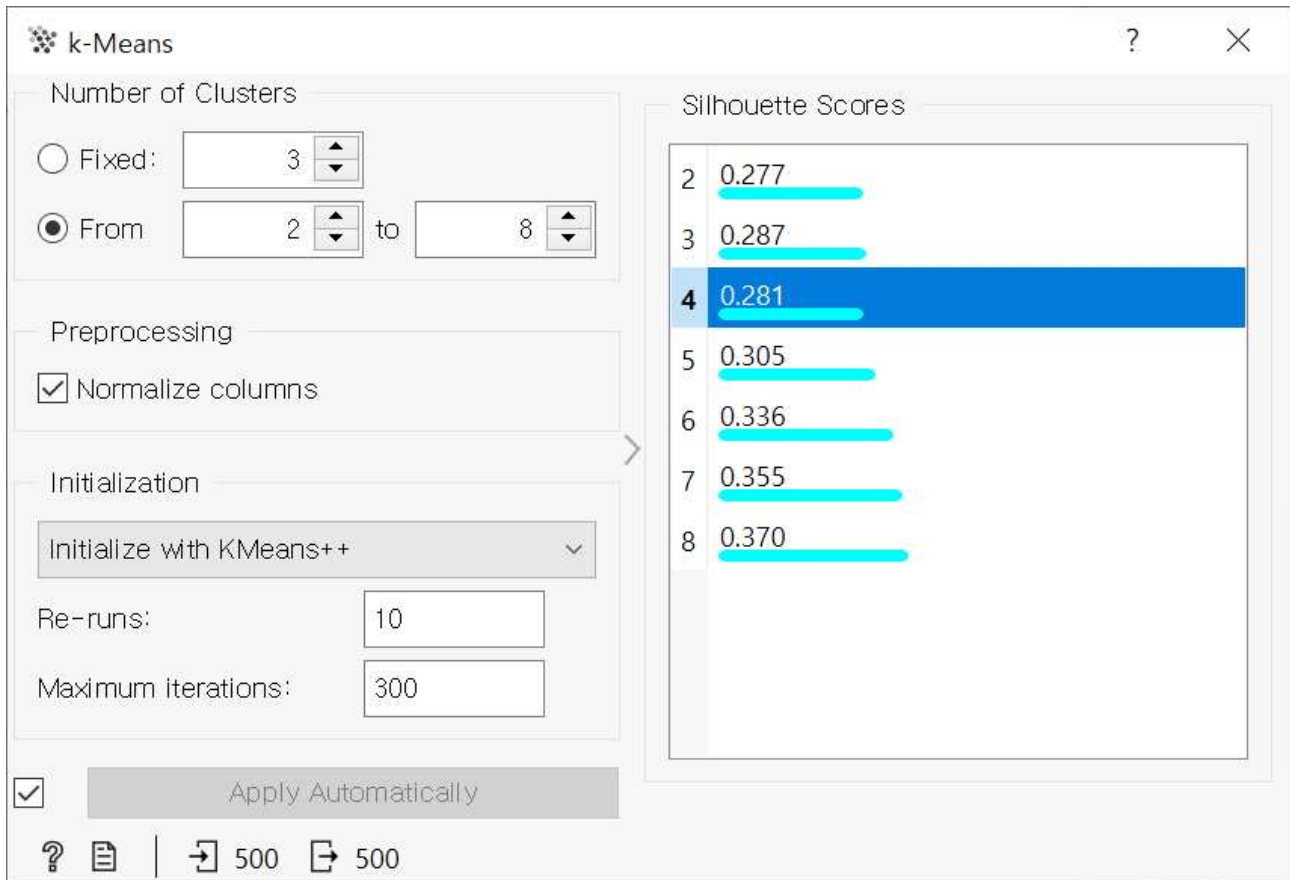
- [k-means]를 더블클릭합니다. k-means 위젯에서 클러스터 수를 정해야 합니다.

#### 1) Number of Clusters

- From 2 to 8 선택 후, 오른쪽 실루엣 점수에서 3이나 4를 선택합니다. (티셔츠 사이즈를 보통 s, m, l, xl로 나누는 것으로 생각하고 3이나 4를 선택했습니다. 데이터의 유형이나 문제상황에 따라 실루엣 점수는 분석하는 사람에 따라 다르게 선택될 수 있습니다.)

#### 2) Initialization : Initialize with K Means++선택

- k-means위젯에서 설정을 완료한 후, 닫기를 클릭하고 다시 캔버스로 돌아옵니다.

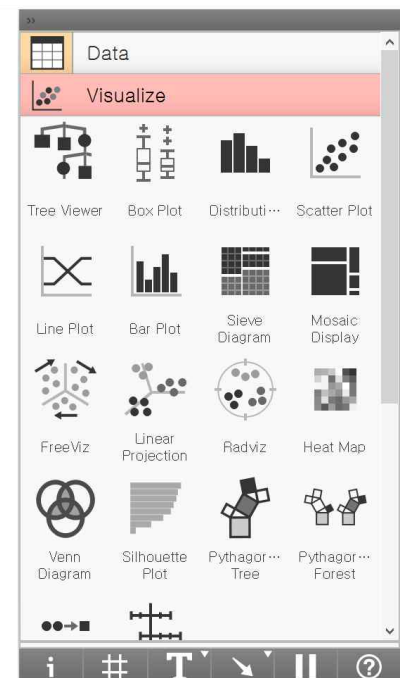
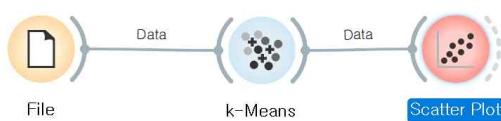


#### 4. 모델 평가하기

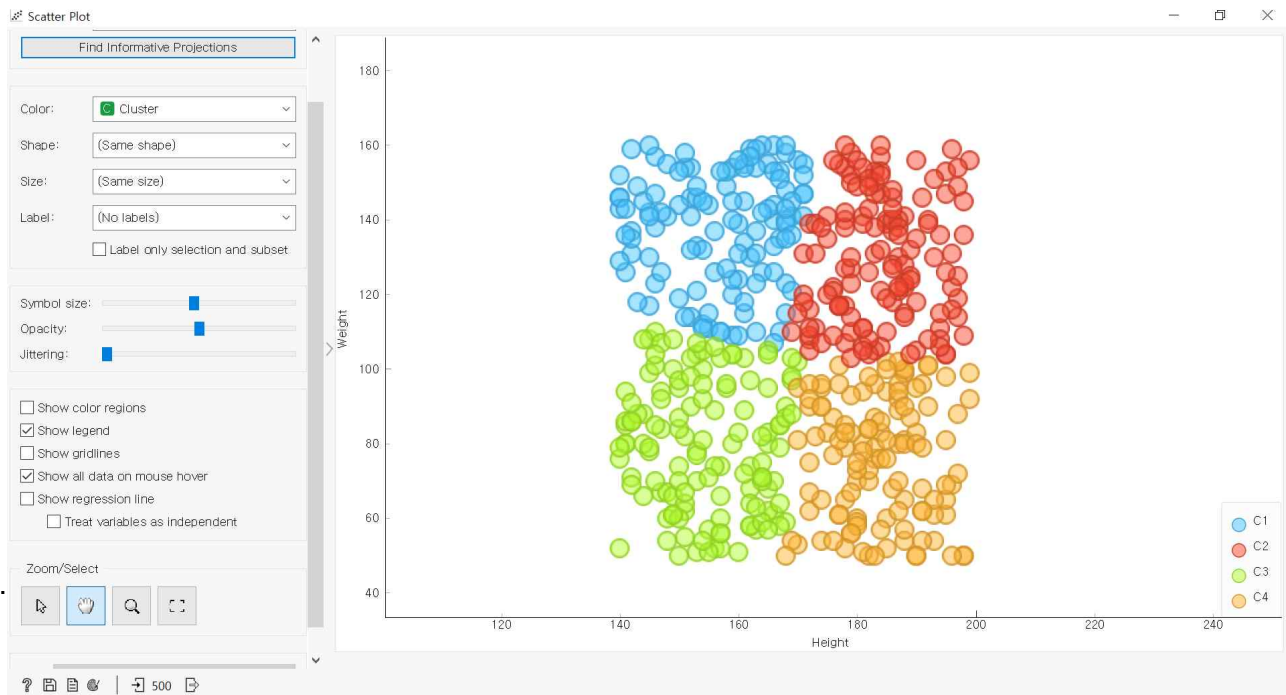
- [Visualize] - [Scatter Plot] 위젯을 선택합니다. Scatter plot을 더블클릭하면 k-means 알고리즘으로 훈련된 데이터셋을 시각화해서 볼 수 있습니다.

키, 몸무게에 따른 체형 군집화.ows\*

File Edit View Widget Options Help



- Scatter plot위젯을 더블클릭하여 머신러닝 모델이 어떻게 데이터를 나눴는지 확인합니다



- 클러스터의 수를 4개로 선택했을 경우, 모델이 4개의 군집(Cluster)으로 데이터를 묶은 것을 확인할 수 있습니다.

**\*\*더 구체적인 설명은 연수에서...!! 파이팅**