

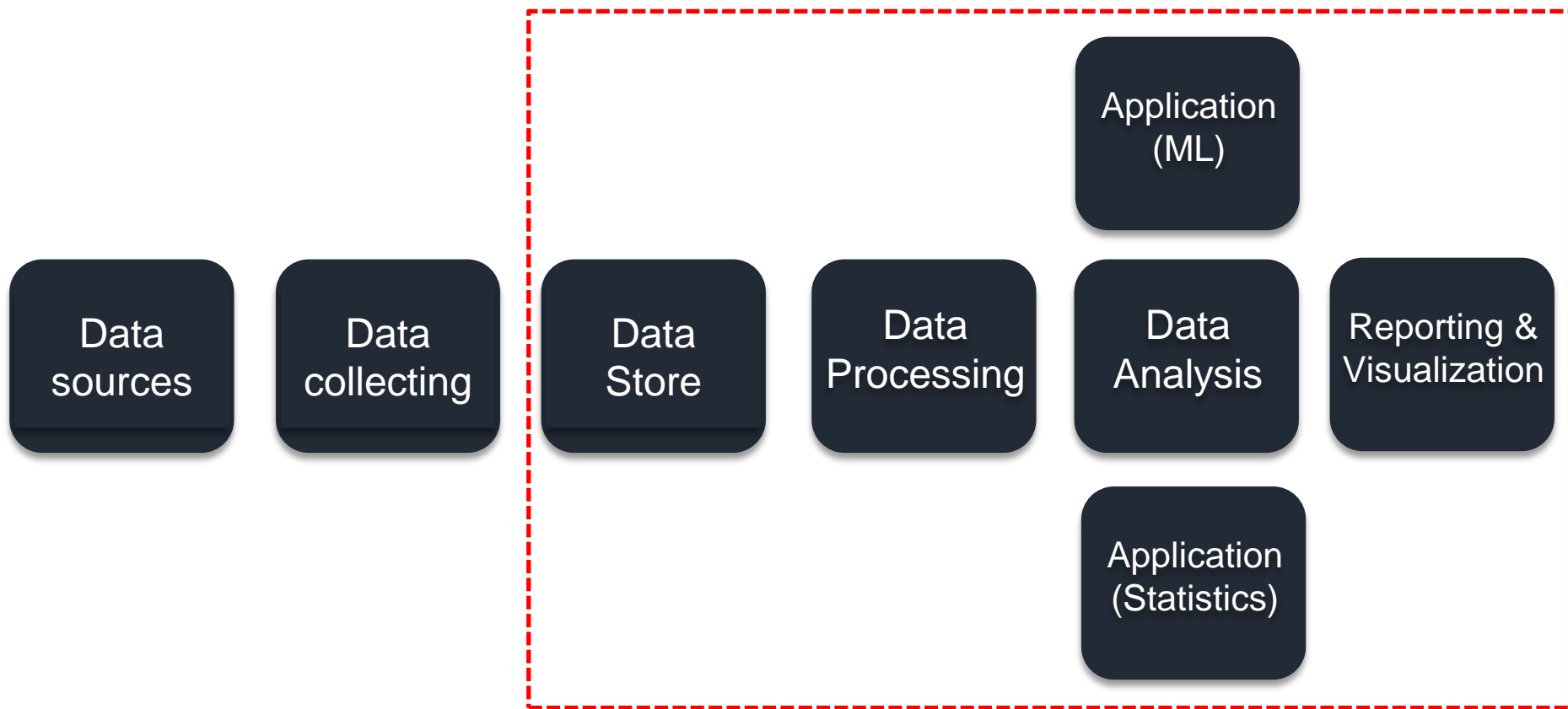
탐색적 데이터 분석

홍익 대학교
Hyun-Sun Ryu

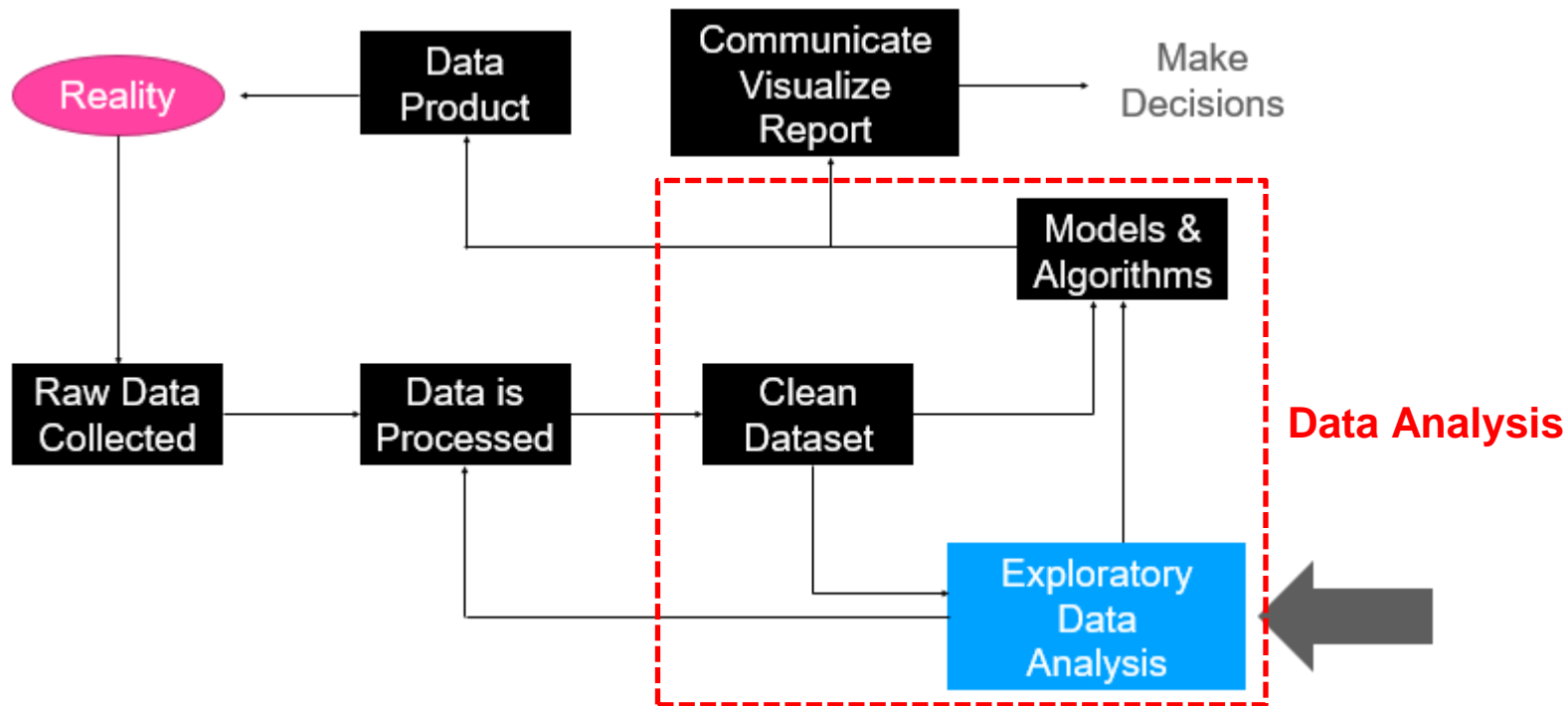
목차

- 데이터의 내용 파악
- 변수의 유형 파악
- 기본 통계 정보 관찰
- 수치 요약

빅데이터 처리 프로세스



데이터 사이언스 프로세스



확증적 데이터 분석 vs. 탐색적 데이터 분석

- **확증적 데이터 분석(CDA: Confirmatory Data Analysis)**
 - 가설을 설정한 후, 수집한 데이터로 가설을 평가하고 추정하는 전통적인 분석
 - **추론통계**
- **탐색적 데이터 분석(EDA: Exploratory Data Analysis)**
 - 원 데이터(Raw data)를 가지고 유연하게 데이터를 탐색하고, 데이터의 특징과 구조로부터 얻은 정보를 바탕으로 통계모형을 만드는 분석방법
 - 주로 빅데이터 분석에 사용
 - **기술통계**

확증적 데이터 분석 vs. 탐색적 데이터 분석

1. 확증적 데이터 분석(CDA)



2. 탐색적 데이터 분석(EDA)



탐색적 데이터 분석 과정

1. 데이터에 대한 질문을 만든다.
2. 데이터를 시각화, 변형 및 모델링하여 질문에 대한 답을 찾는다.
3. 질문을 개선하거나 새로운 질문을 만들기 위해 학습한 방법을 사용한다.

질문하기



질문하기: 정량적(Quantitative) 질문 유형

- 서술형 질문

- 문제의 개념 또는 주제를 설명
- 빈도, 하루 중 시간, 사용 목적 등과 같은 제품의 사용을 이해하는 것

- 비교 질문

- 두 그룹, 개념 또는 기타 변수 간의 차이를 분석하는데 사용
- 두 제품 간의 사용 빈도 비교, 남성 대 여성의 브랜드 선호도 등

- 관계 기반 질문

- 인과관계에 기반한 질문들은 한 변수가 다른 변수에 어떻게 영향을 미치는지 이해하는데 효과가 있음
- 색상이 특정 제품을 구매하려는 욕구에 어떻게 영향을 미치는지

질문하기: 질적(Qualitative) 질문 유형

- 탐색적 질문

- 정량적 질문의 서술형 문제와 유사하게 선입견으로 결과에 주지 않고 무엇인가를 이해하는 것으로 보임
- 제품이 어떻게 사용되는지 또는 특정 주제에 대한 인식을 묻는 것

- 예측 질문

- 주제나 행동을 둘러싼 의도나 미래의 결과를 이해하려고 함
- 소비자가 왜 특정한 상황에서 행동하는지

- 해석적 질문

- 결과에 영향을 주지 않고 특정 주제 또는 개념에 대한 피드백 수집
- 새로운 제품 개념을 테스트하고 전달 요청이 어떻게 해석되는지 이해하는 것

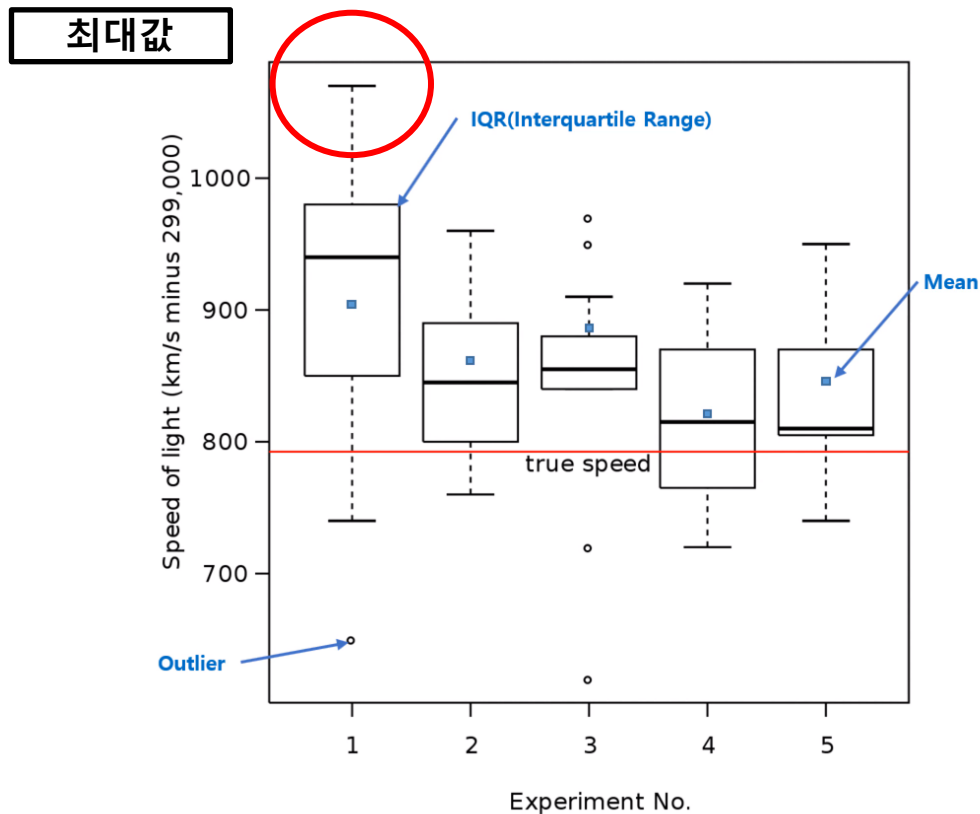
이상값(outlier)발견 기법

- 개별 데이터 관찰: 데이터 값을 눈으로 보며 전체적인 추세와 특이사항 관찰
 - 통계값 활용: 요약 통계지표
 - 시각화 활용: 확률밀도함수, 히스토그램, 점플롯, 워드클라우드, 시계열 차트, 지도
 - 머신러닝 기법 활용: 클러스터링 등을 통해서 이상치 확인
-
- 통계기반탐지
 - 편차 기반 탐지
 - 거리 기반 탐지

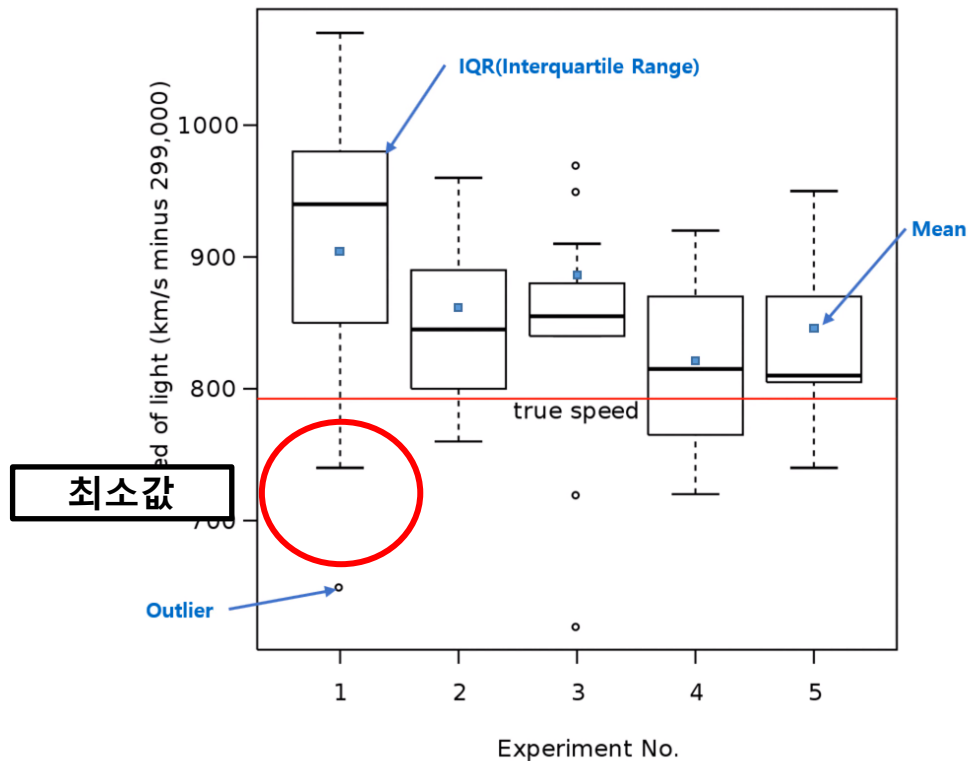
5가지 숫자 요약(five-number summary): Box plot

- 데이터 집합에 대한 정보를 제공하는 통계량으로 가장 중요한 표본 백분위수 5가지로 구성
 - **최대값(maximum)**
 - **상위 사분위수(upper quartile) 또는 제3사분위수(Q3):**
중앙값 기준으로 상위 50%중의 중앙값, 전체 데이터 중 상위 25%에 해당
 - **중앙값(median) :** 데이터의 가운데 순위에 해당 하는 값
 - **하위 사분위수(lower quartile) 또는 제1사분위수(Q1):**
중앙값 기준으로 하위 50%중의 중앙값, 전체 데이터 중 하위 25%에 해당
 - **최소값(minimum)**

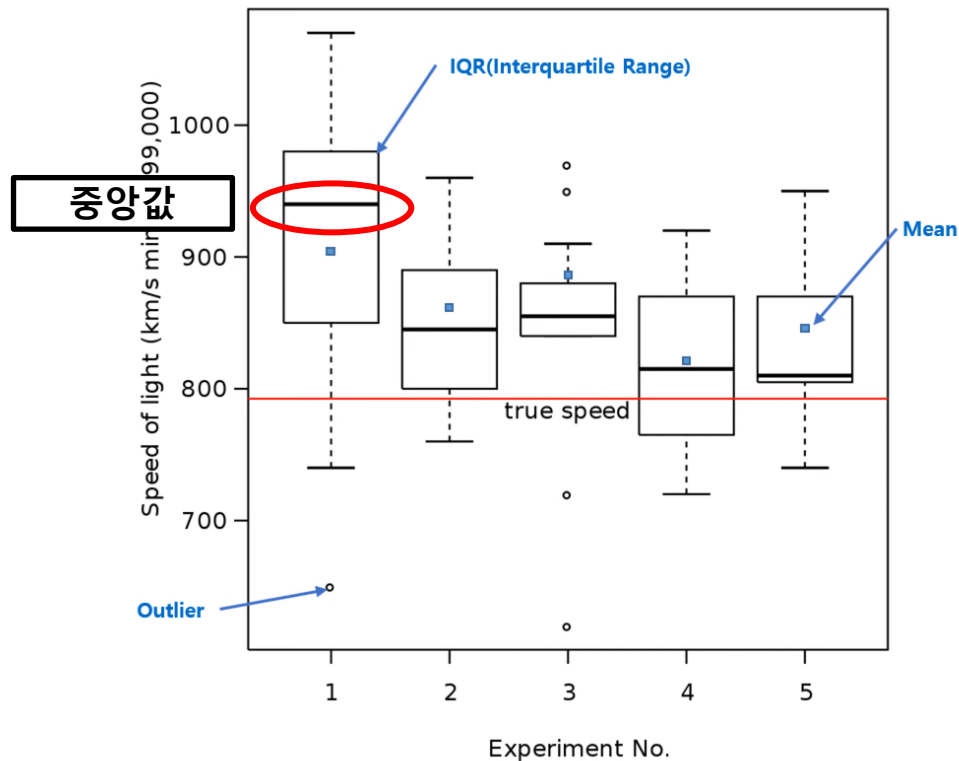
5가지 숫자 요약(five-number summary): Box plot



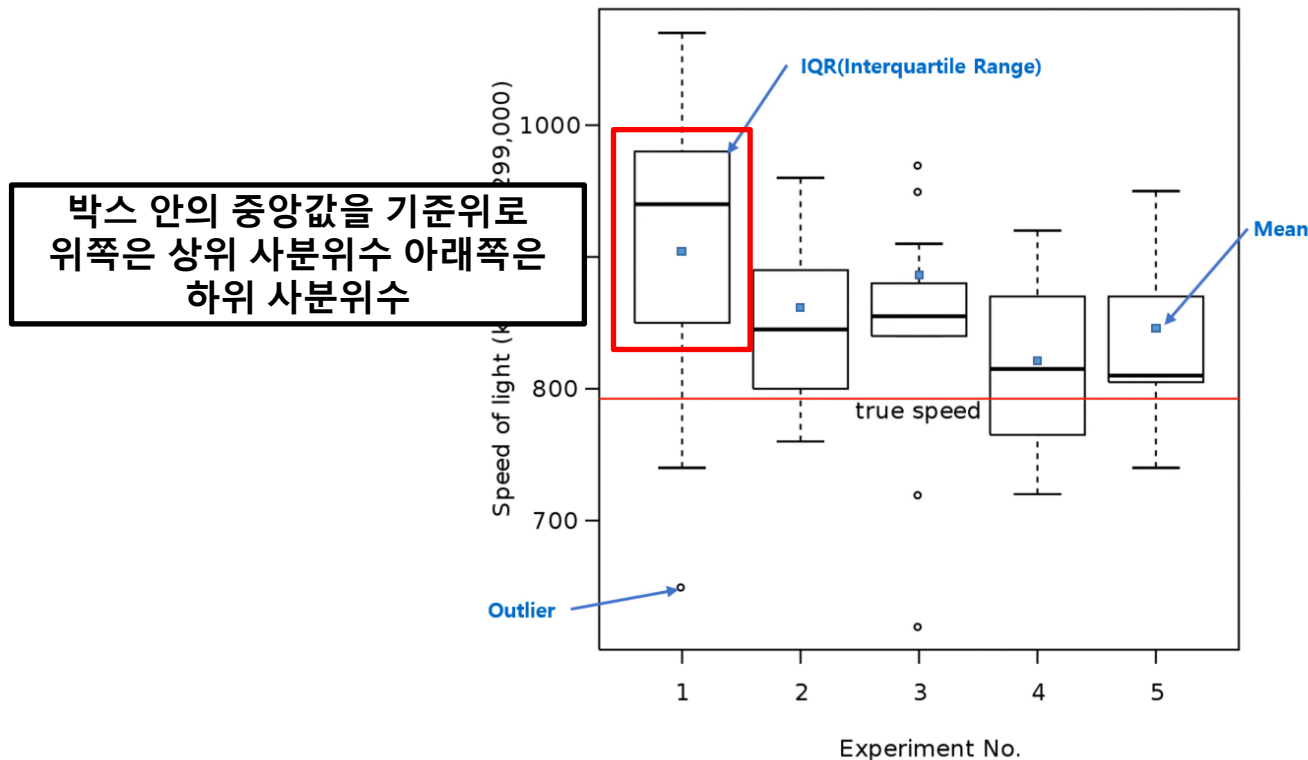
5가지 숫자 요약(five-number summary): Box plot



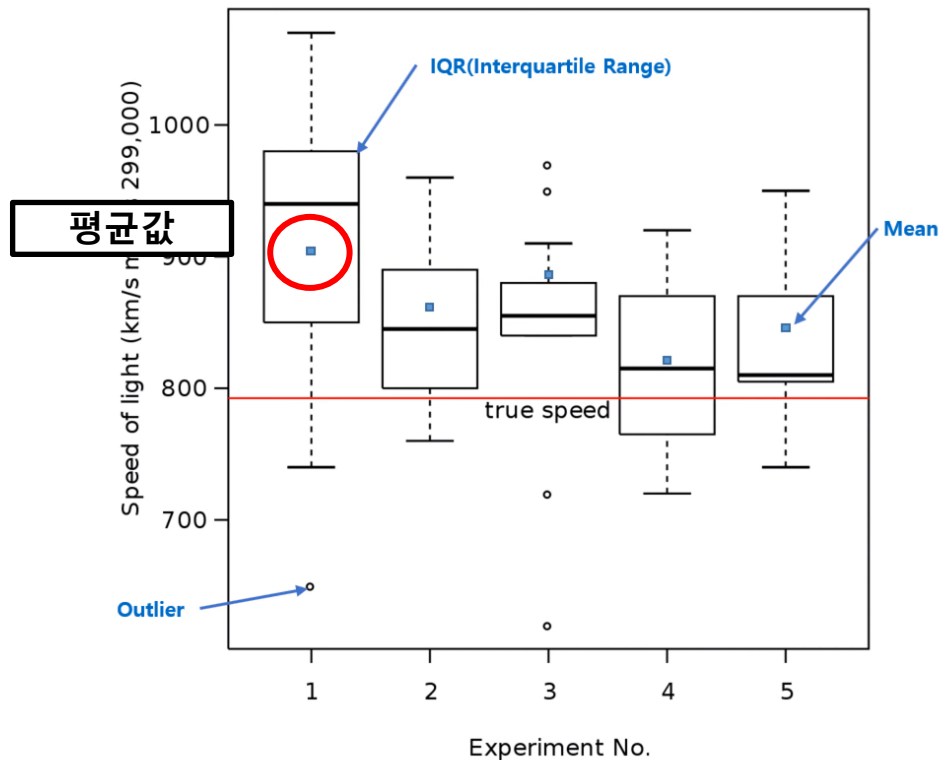
5가지 숫자 요약(five-number summary): Box plot



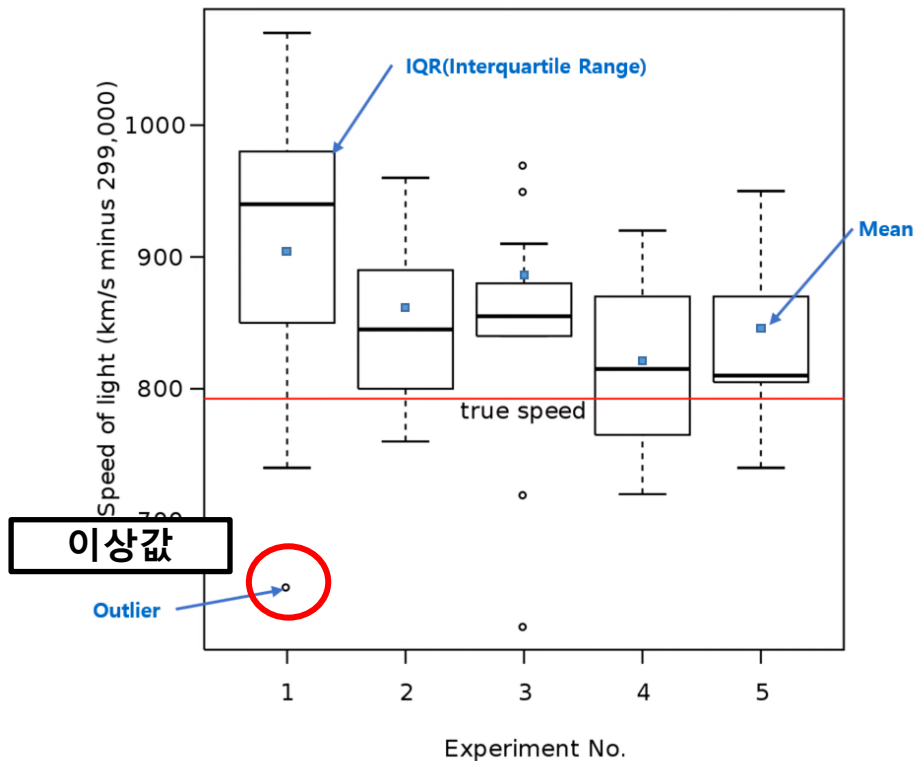
5가지 숫자 요약(five-number summary): Box plot



5가지 숫자 요약(five-number summary): Box plot



5가지 숫자 요약(five-number summary): Box plot



Descriptive Statistics(기술 통계)

- 시각화를 통한 데이터 관찰



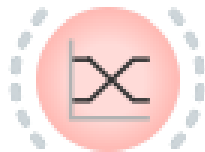
Box Plot



Distributions



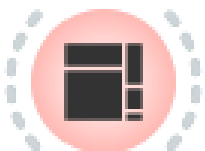
Scatter Plot



Line Plot



Bar Plot



Mosaic Display



Pythagorean Tree



Venn Diagram

속성 간의 관계 분석



▪ Categorical Variable (Qualitative) (범주형 변수)

Nominal Data

- 원칙적으로 숫자로 표시할 수 없으나 편의상 숫자화
- 남자-0, 여자-1

Ordinal Data

- 원칙적으로 숫자로 표시할 수 없으나 편의상 숫자화
- 순위의 개념이 있음
- 소득분위 10분위, 9분위

속성 간의 관계 분석

- Numeric Variable(Quantitative) (수치형 변수)



Continuous Data

- 데이터가 연속량 으로서 셀 수 있는 형태
- 키 – 167.2cm

Discrete Data

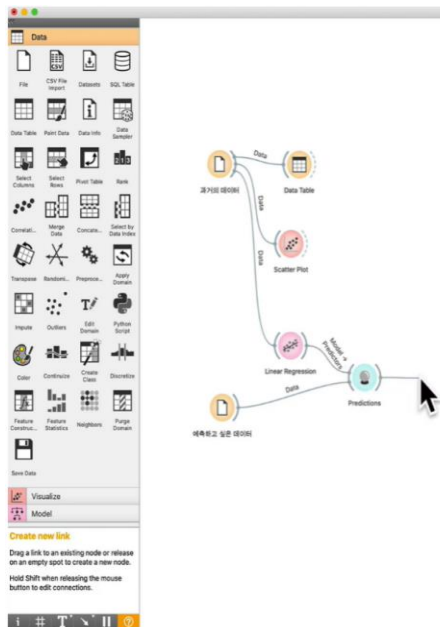
- 데이터가 비연속량 으로서 셀 수 있는 형태
- 자녀 수 4명

속성 간의 관계 분석

데이터 조합	요약 통계	시각화
Categorical – Categorical	교차 테이블	모자이크 플롯
Numeric – Categorical	카테고리별 통계 값	박스 플롯
Numeric – Numeric	상관계수	<u>산점도</u>

오렌지3(Orange3)

- Orange3는 코드 없이 드래그 앤 드롭으로 데이터를 분석할 수 있는 도구
- 시각화 뿐 아니라 머신러닝에 사용되는 다양한 모델도 제공하는 강력한 도구
 - 표(수정, 처리) (입력)
 - 시각화 (입력, 출력)
 - 머신러닝 (처리)



오렌지

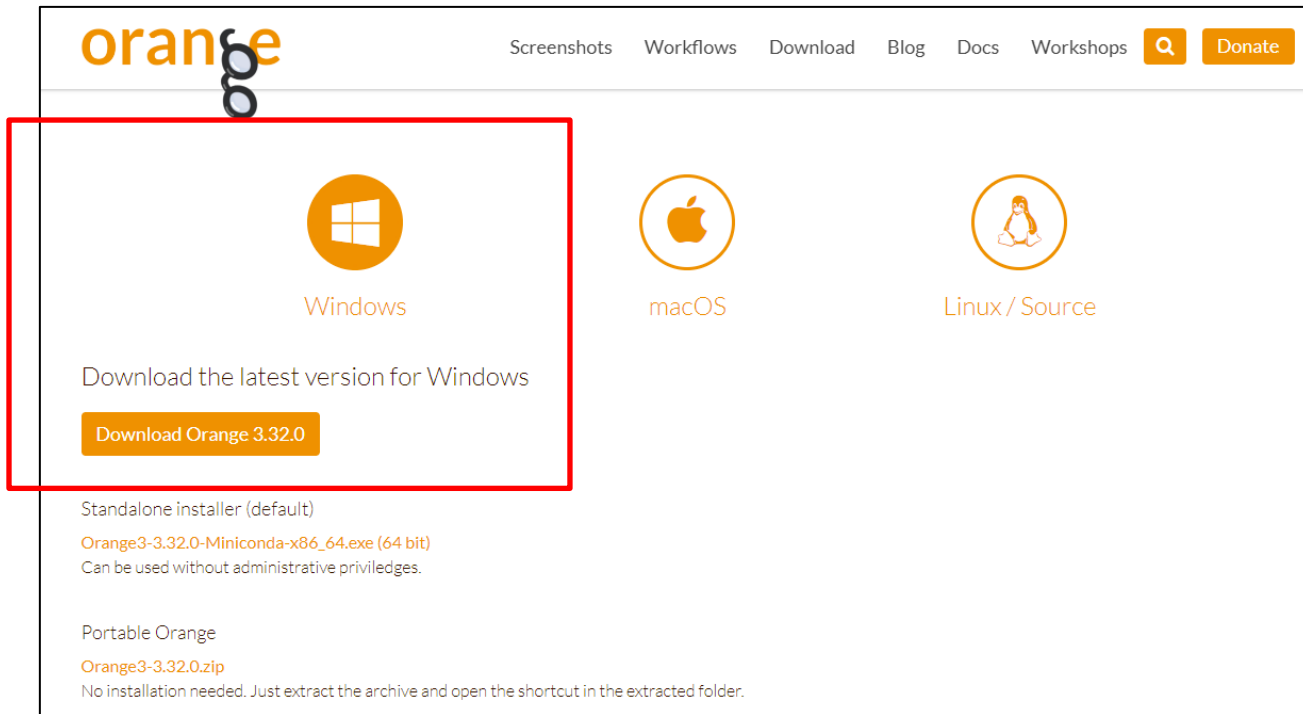
코드 없이
드래그 앤 드롭

표의 분석
시각화
머신러닝

통계
데이터 마이닝
데이터 과학

오렌지3 다운로드

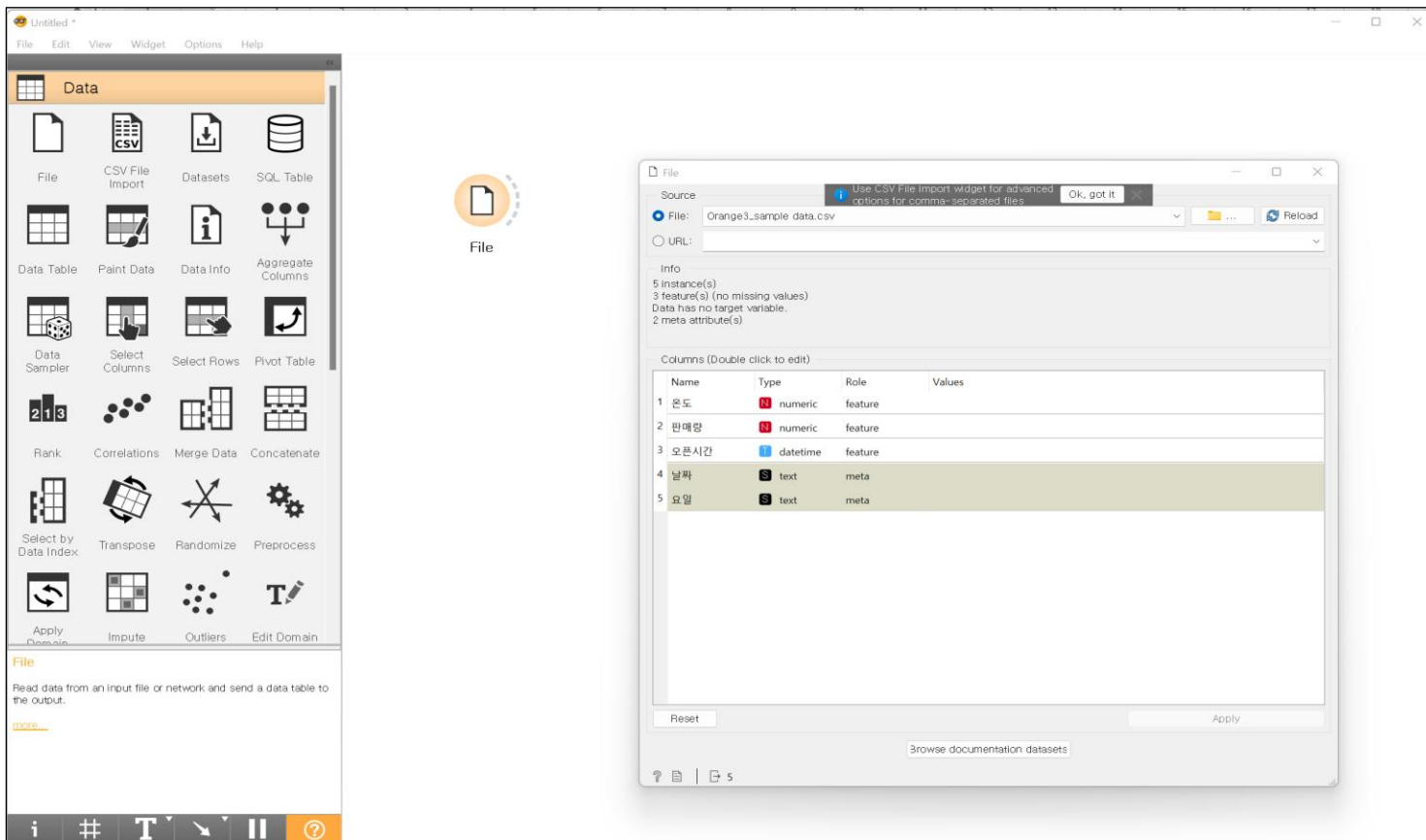
- <https://orangedatamining.com/download/#windows>



오렌지3 기본 사용법

- 데이터 파일: Orange3_sample data.csv
- File → Role
 - **Skip** : 무시해도 되는 데이터
 - **Meta** : 분석을 할 것은 아니지만 참고할 데이터, 예를 들어 날짜.
 - **Target** : 예측하고자 하는 값, 종속변수
 - **Feature** : 원인에 해당하는 값, 독립변수

오렌지3 기본 사용법(File)



The screenshot displays the Orange3 File widget configuration window. The main window shows the 'Data' widget palette on the left, a central workspace with a 'File' widget icon, and the 'File' widget configuration panel on the right.

File Widget Configuration Panel:

- Source:** File: Orange3_sample data.csv (with a dropdown arrow and a 'Reload' button).
- Info:** 5 instance(s), 3 feature(s) (no missing values), Data has no target variable, 2 meta attribute(s).
- Columns (Double click to edit):**

	Name	Type	Role	Values
1	온도	numeric	feature	
2	판매량	numeric	feature	
3	오픈시간	datetime	feature	
4	날짜	text	meta	
5	요일	text	meta	

Buttons: Reset, Apply, Browse documentation datasets.

오렌지3 기본 사용법(Data Table)

The screenshot displays the Orange3 software interface. On the left is the 'Data' widget palette, and on the right is the main workspace showing a workflow and the 'Data Table' widget configuration window.

Workflow: A 'File' widget is connected to a 'Data Table' widget.

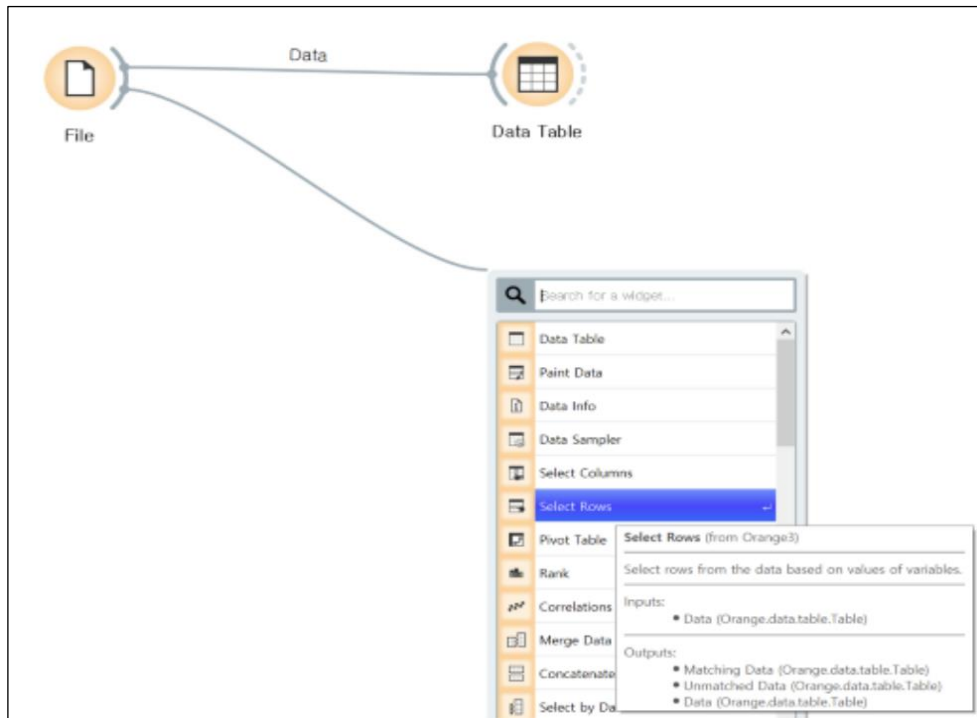
Data Table Widget Configuration:

- Info:** 5 instances (no missing data), 3 features, No target variable, 2 meta attributes.
- Variables:**
 - ☒ Show variable labels (if present)
 - ☐ Visualize numeric values
 - ☒ Color by instance classes
- Selection:**
 - ☒ Select full rows
- Buttons:** Restore Original Order, Send Automatically (checked).

Data Table View:

	날짜	요일	온도	판매량	오픈시간
1	2020-06-01 0:00	금	20	40	10:00:00
2	2020-06-02 0:00	토	21	42	10:01:00
3	2020-06-03 0:00	일	22	44	10:02:00
4	2020-06-04 0:00	월	23	46	10:01:00
5	2020-06-05 0:00	화	24	48	10:10:00

오렌지3 기본 사용법(select rows)



오렌지3 기본 사용법(filtering)

The screenshot displays the Orange3 data mining software interface. On the left is a widget palette titled "Data" containing various data processing widgets. The main workspace shows a workflow with the following components and connections:

- File** widget connected to **Data Table** widget via a connection labeled "Data".
- File** widget connected to **Select Rows** widget via a connection labeled "Data".
- Select Rows** widget connected to **Data Table (1)** widget via a connection labeled "Matching Data".

The **Select Rows** widget's configuration window is open, showing the following settings:

- Conditions:** A table with one condition:

Variable	Operator	Value
판매량	is below	44
- Remove unused features:** ☐
- Remove unused classes:** ☐
- Send Automatically:** ☒

At the bottom of the **Select Rows** window, there are buttons for "Add Condition", "Add All Variables", and "Remove All". The status bar at the bottom of the window shows icons for help, data, and a progress indicator.

오렌지3 기본 사용법(filtering)

The screenshot displays the Orange3 software interface. On the left is a widget palette with various data processing tools. The main workspace shows a workflow: a 'File' widget connects to a 'Data Table' widget, and another 'File' widget connects to a 'Select Rows' widget, which then connects to a 'Data Table (1)' widget. A 'Data' label is placed between the two 'File' widgets. A 'Data Table (1)' widget is also connected to the 'Data Table' widget via a 'Matching Data' connection.

The 'Data Table (1)' widget is open, showing the following information:

Info
2 instances (no missing data)
3 features
No target variable.
2 meta attributes

Variables
☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

Restore Original Order

☒ Send Automatically

The data table contains the following data:

	날짜	요일	온도	판매량	오픈시간
1	2020-06-01 0:00	금	20	40	10:00:00
2	2020-06-02 0:00	토	21	42	10:01:00

오렌지3 기본 사용법(filtering)

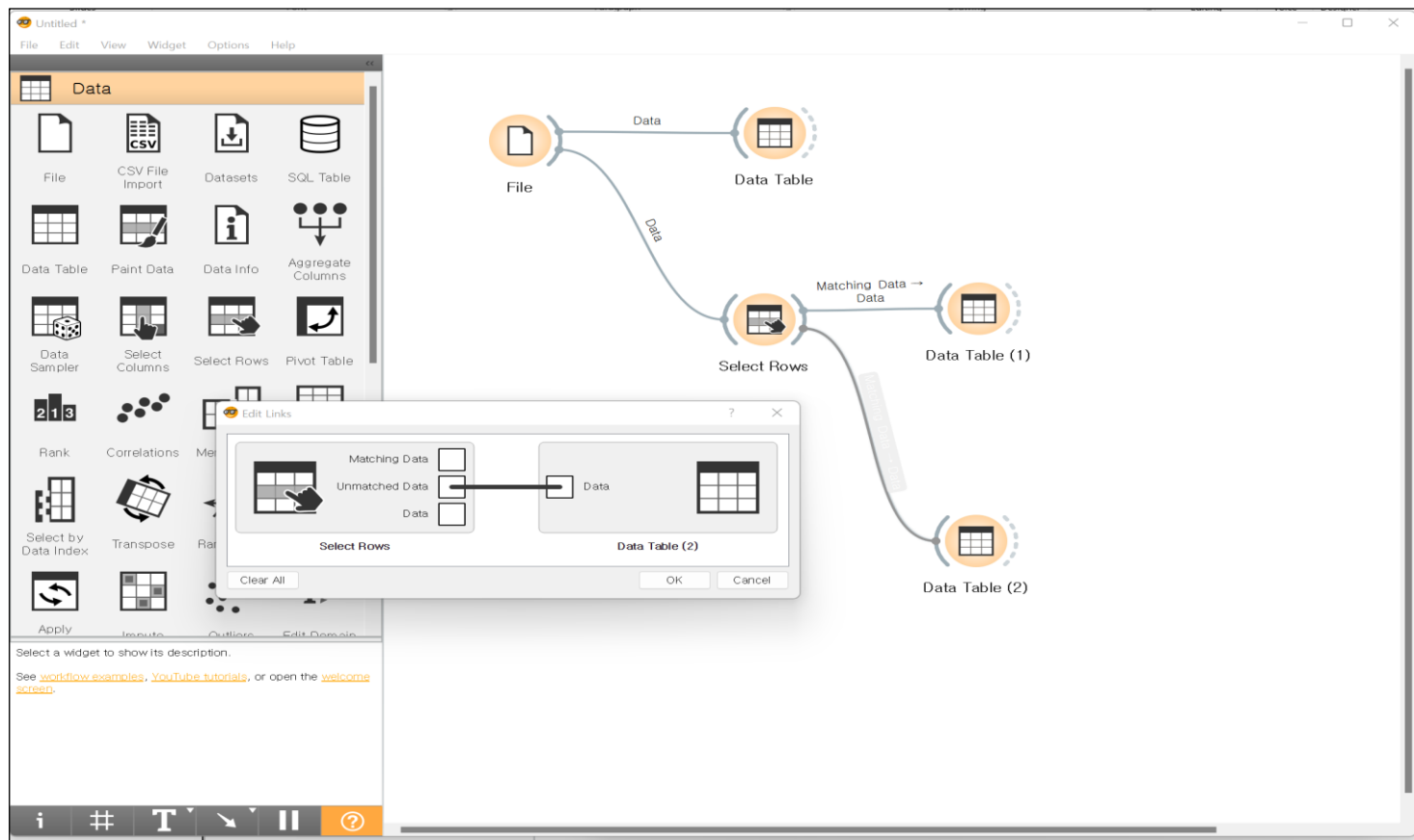


표 조작하기

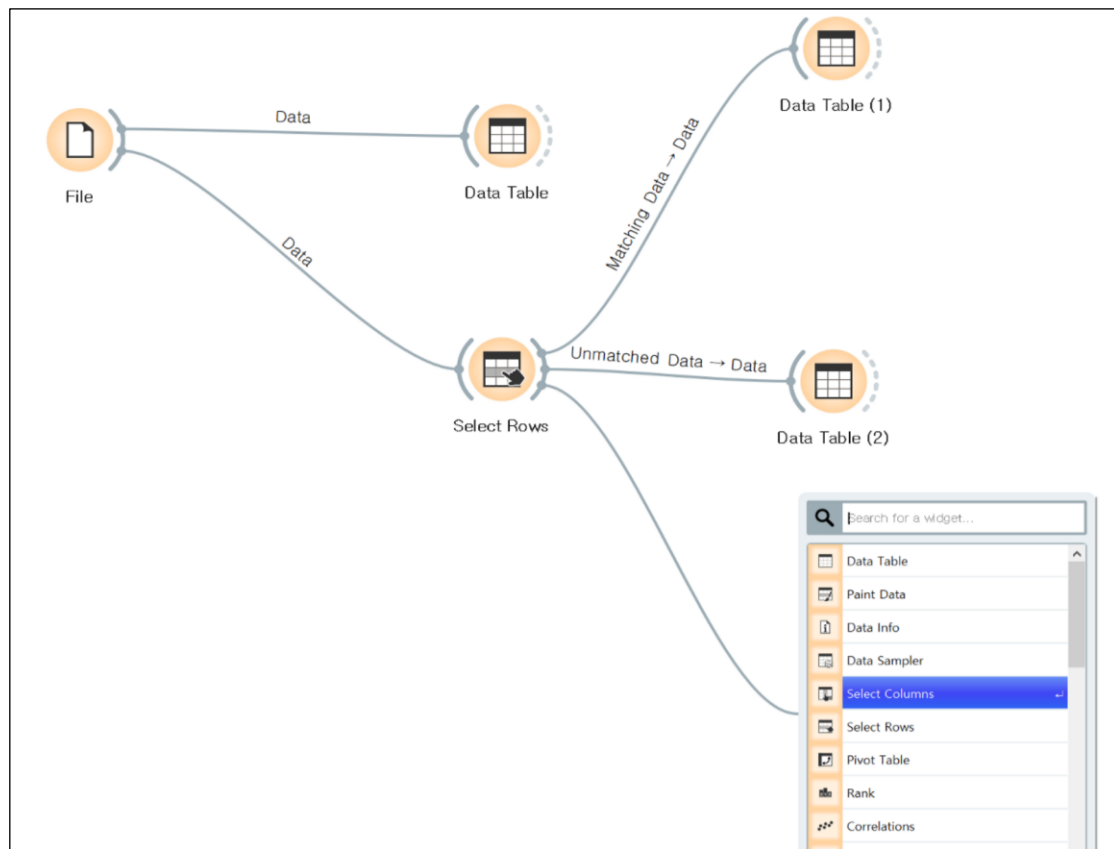


표 조작하기

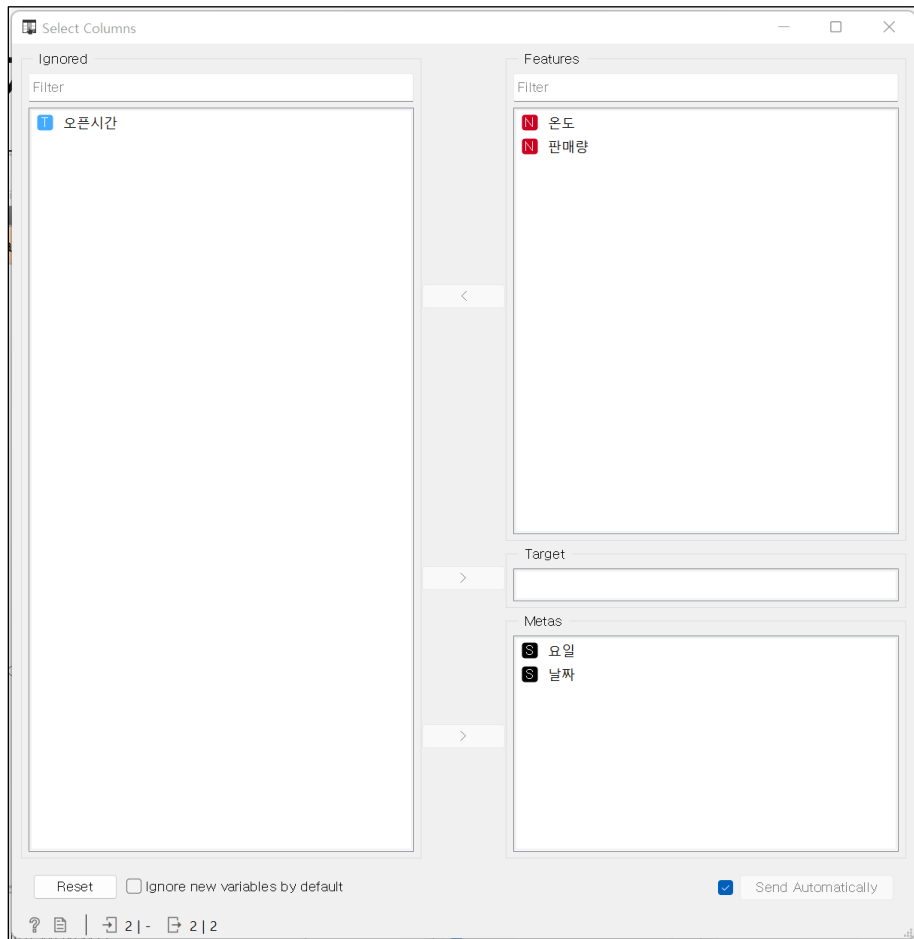


표 조작하기

The screenshot displays the Orange3 data mining software interface. On the left is a widget toolbox with various data processing widgets. The main workspace shows a workflow diagram where data from a 'File' widget is split into two paths. One path goes directly to a 'Data Table' widget. The other path goes through a 'Select Rows' widget, which then branches into two 'Data Table' widgets labeled 'Data Table (1)' and 'Data Table (2)'. The 'Select Rows' widget is configured to select full rows based on matching data. A third 'Data Table' widget, labeled 'Data Table (3)', is shown in a separate window, displaying a preview of the data and its settings.

Widget Toolbox:

- Data: File, CSV File Import, Datasets, SQL Table, Data Table, Paint Data, Data Info, Aggregate Columns, Data Sampler, Select Columns, Select Rows, Pivot Table, Rank, Correlations, Merge Data, Concatenate, Select by Data Index, Transpose, Randomize, Preprocess.

Workflow Diagram:

```
graph LR; File[File] -- Data --> DT[Data Table]; File -- Data --> SR[Select Rows]; SR -- "Matching Data -> Data" --> DT1[Data Table (1)]; SR -- "Unmatched Data -> Data" --> DT2[Data Table (2)]; SR -- "Matching Data -> Data" --> DT3[Data Table (3)]
```

Data Table (3) Preview:

	요일	날짜	온도	판매량
1	금	2020-06-01 0:00	20	40
2	토	2020-06-02 0:00	21	42

Data Table (3) Settings:

- Info: 2 instances (no missing data), 2 features, No target variable, 2 meta attributes.
- Variables: ☒ Show variable labels (if present), ☐ Visualize numeric values, ☒ Color by instance classes.
- Selection: ☒ Select full rows.
- Buttons: Restore Original Order, Send Automatically.

표 조작하기

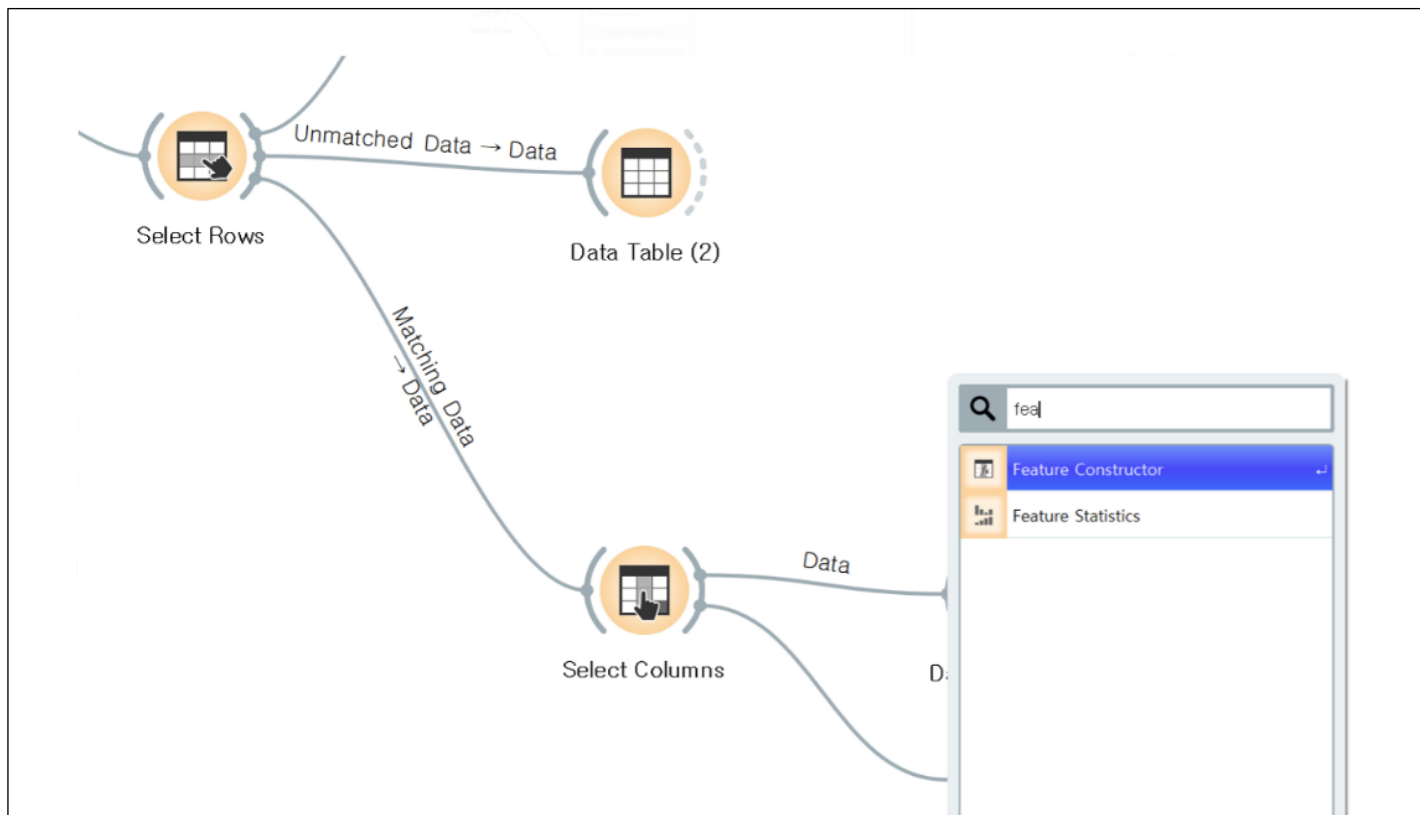


표 조작하기

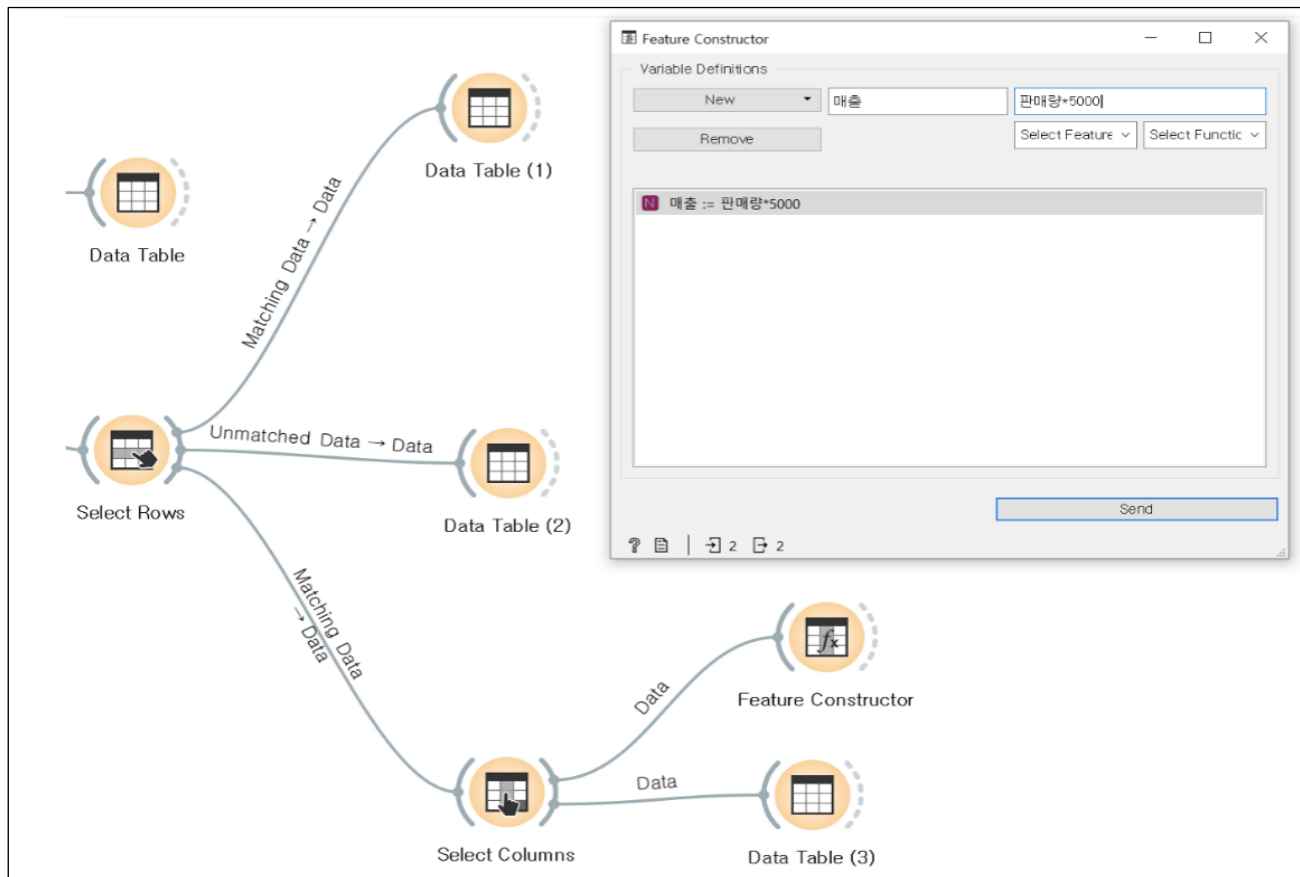


표 조작하기

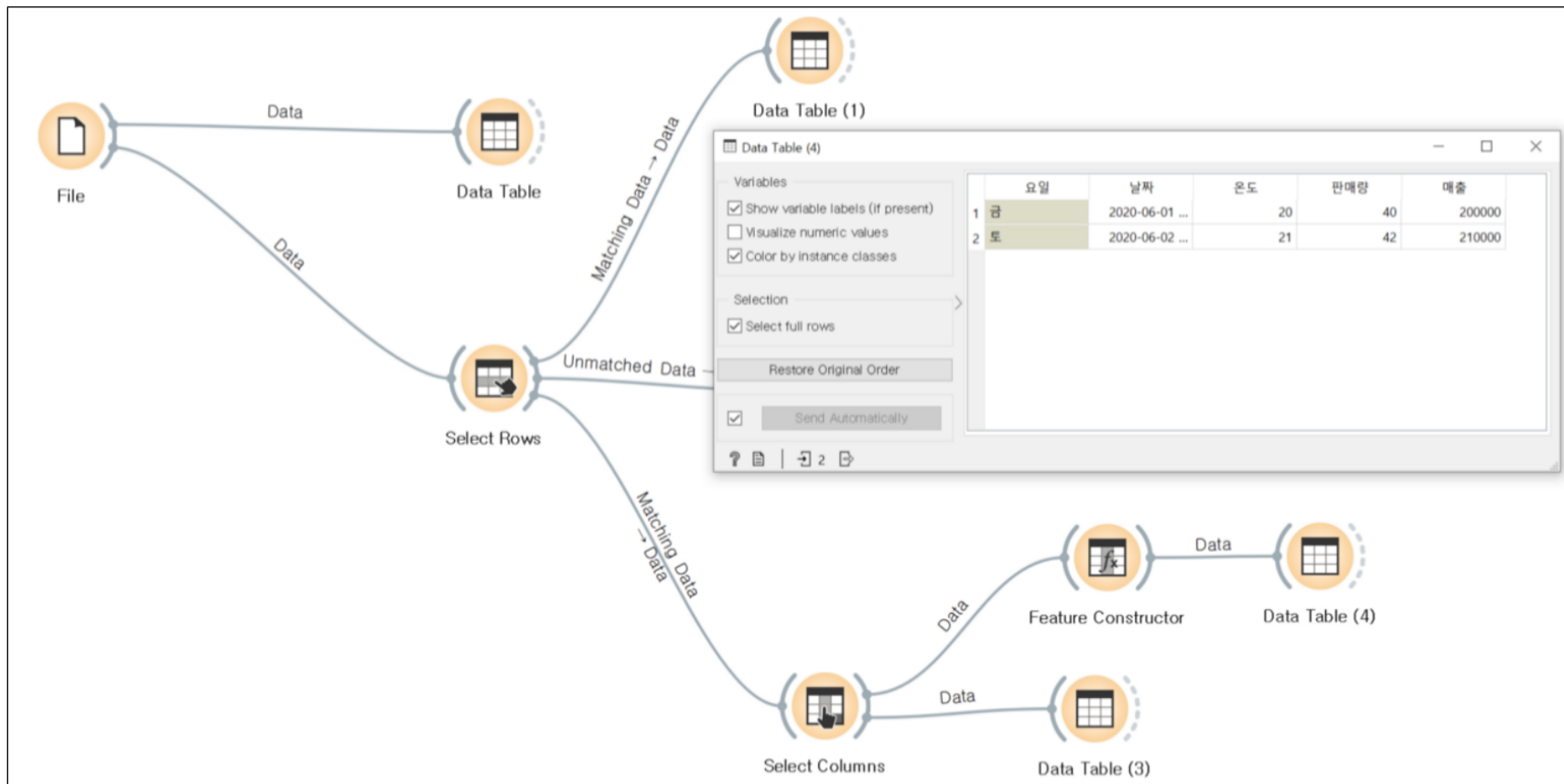
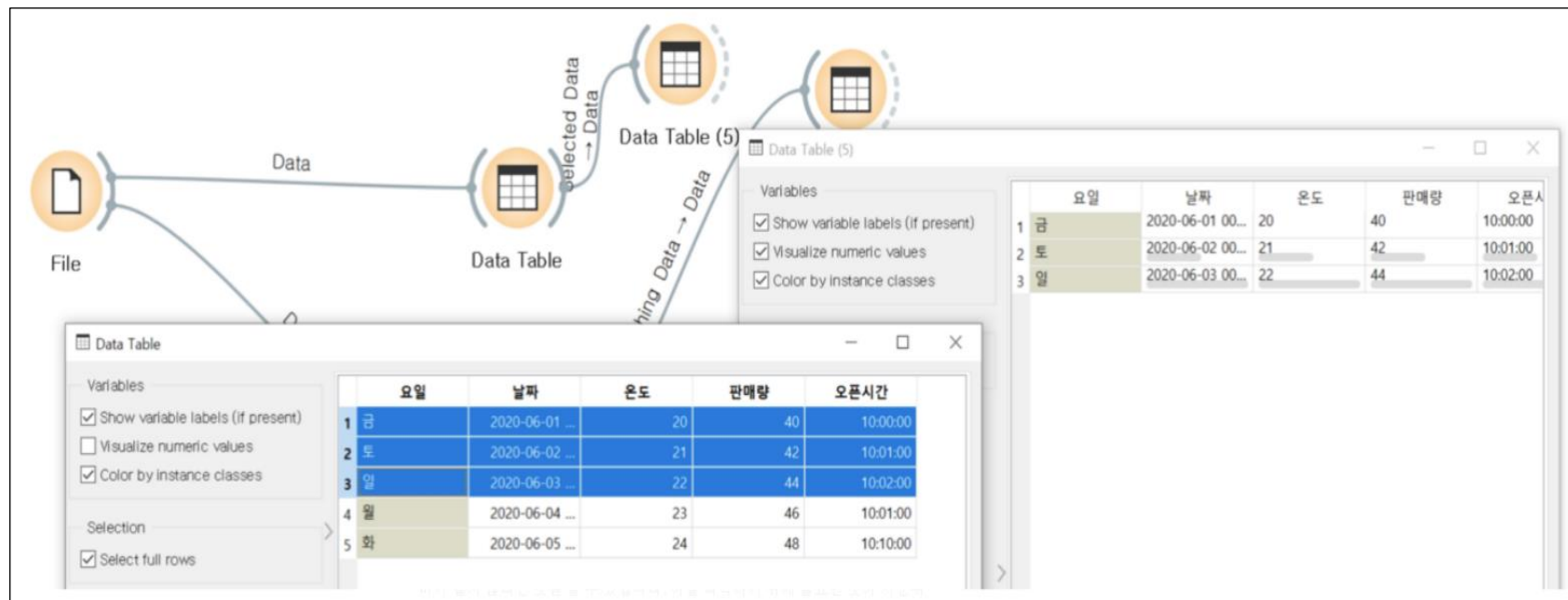


표 조작하기



통계의 시각화

- 데이터 파일: Orange3_sample data.csv, Orange3_sample data(2).csv

The image displays two Orange3 workflow diagrams and two screenshots of the Data Table widget.

Workflow 1: A 'File' widget is connected to a 'Data Table (1)' widget via a 'Data' link.

Workflow 2: A 'File (1)' widget is connected to a 'Data Table' widget via a 'Data' link.

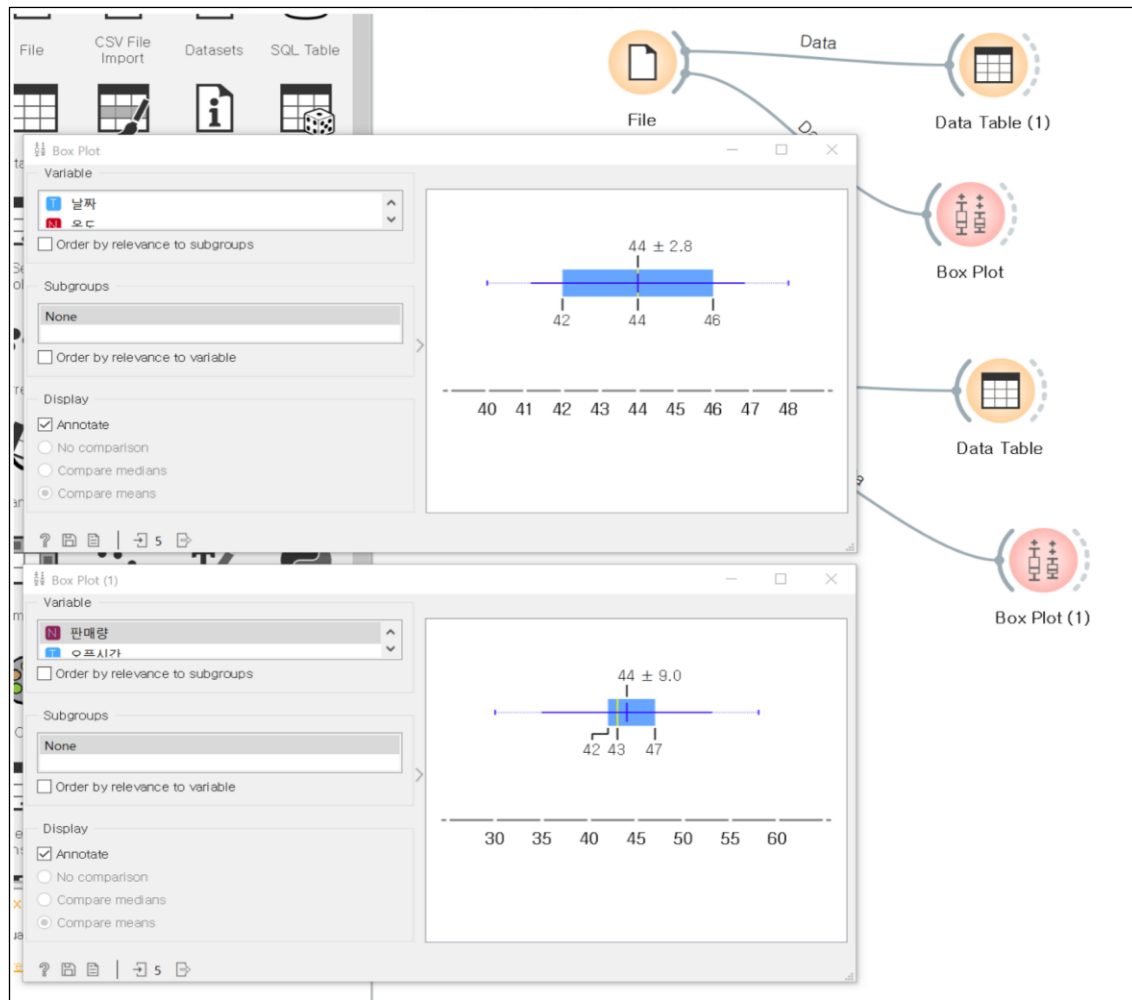
Data Table (1) Screenshot: The widget shows a table with 5 rows and 5 columns. The columns are labeled '요일' (Day of Week), '날짜' (Date), '온도' (Temperature), '판매량' (Sales), and '오픈시' (Opening Time). The data is as follows:

	요일	날짜	온도	판매량	오픈시
1	금	2020-06-01 00...	20	40	10:00:00
2	토	2020-06-02 00...	21	42	10:01:00
3	일	2020-06-03 00...	22	44	10:02:00
4	월	2020-06-04 00...	23	46	10:01:00
5	화	2020-06-05 00...	24	48	10:10:00

Data Table Screenshot: The widget shows a table with 5 rows and 5 columns. The columns are labeled '요일' (Day of Week), '날짜' (Date), '온도' (Temperature), '판매량' (Sales), and '오픈시' (Opening Time). The data is as follows:

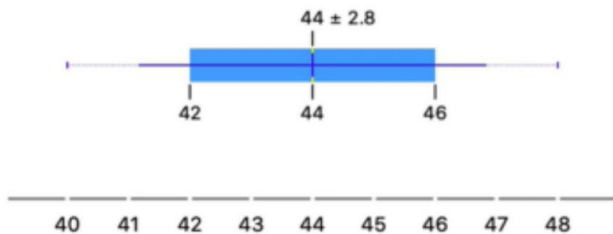
	요일	날짜	온도	판매량	오픈시
1	금	2020-06-01 00...	20	30	10:00:00
2	토	2020-06-02 00...	21	42	10:01:00
3	일	2020-06-03 00...	22	43	10:02:00
4	월	2020-06-04 00...	23	47	10:01:00
5	화	2020-06-05 00...	24	58	10:10:00

통계의 시각화

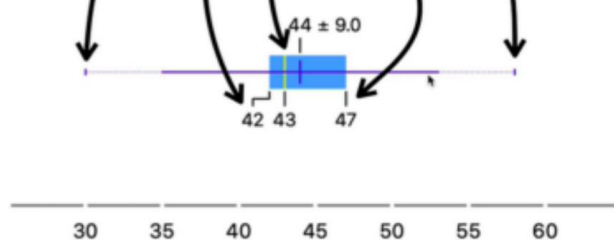


통계의 시각화

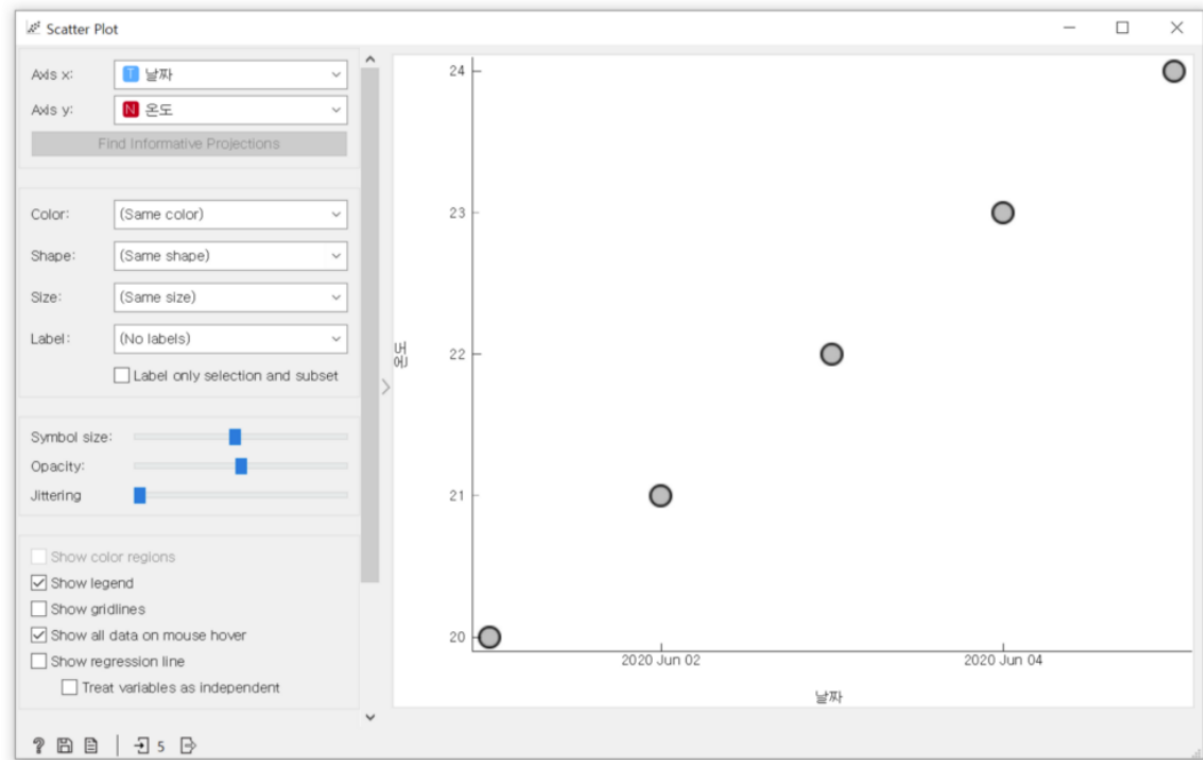
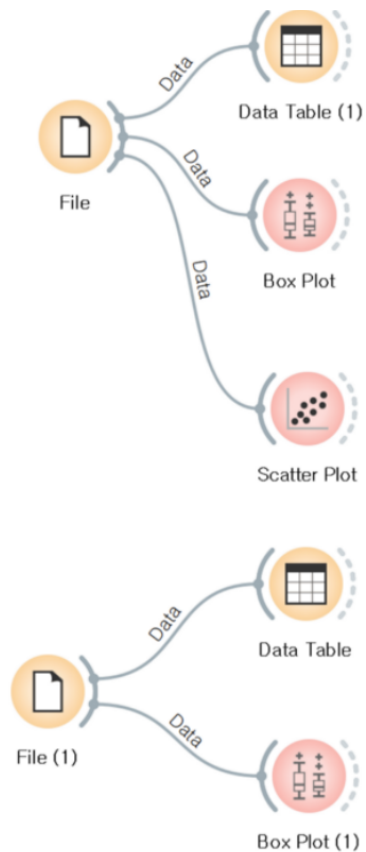
날짜	판매량
2020-06-01	40
2020-06-02	42
2020-06-03	44
2020-06-04	46
2020-06-05	48



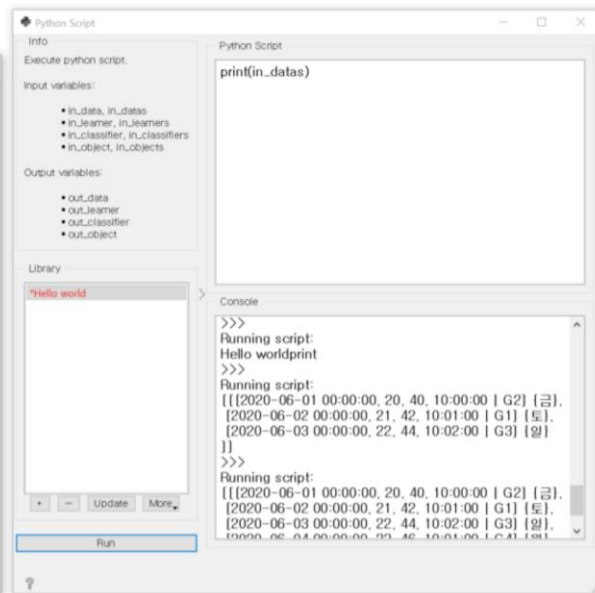
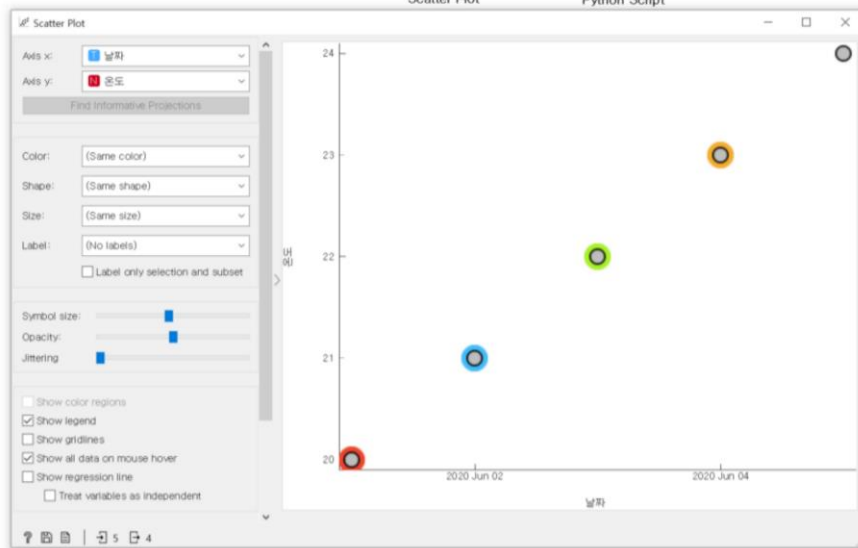
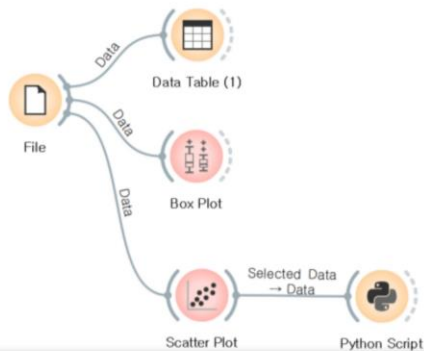
날짜	판매량
2020-06-01	30
2020-06-02	42
2020-06-03	43
2020-06-04	47
2020-06-05	58



통계의 시각화



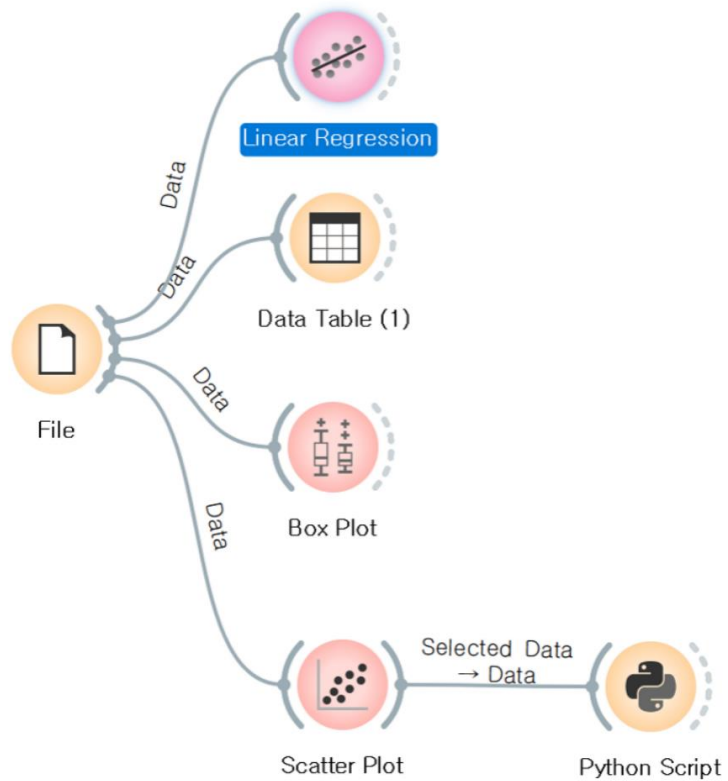
통계의 시각화



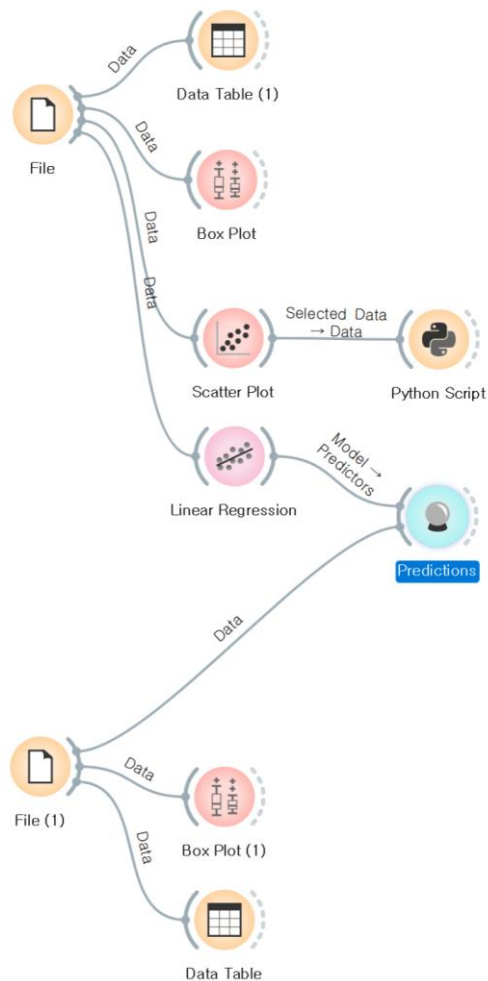
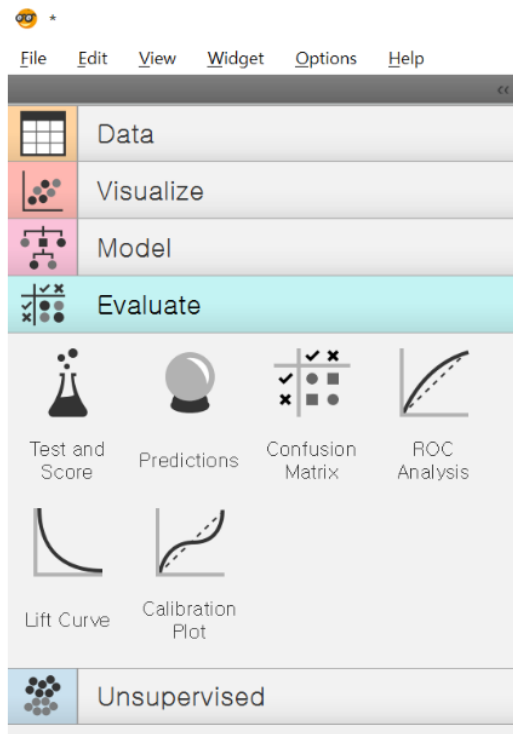
통계의 시각화








통계의 시각화



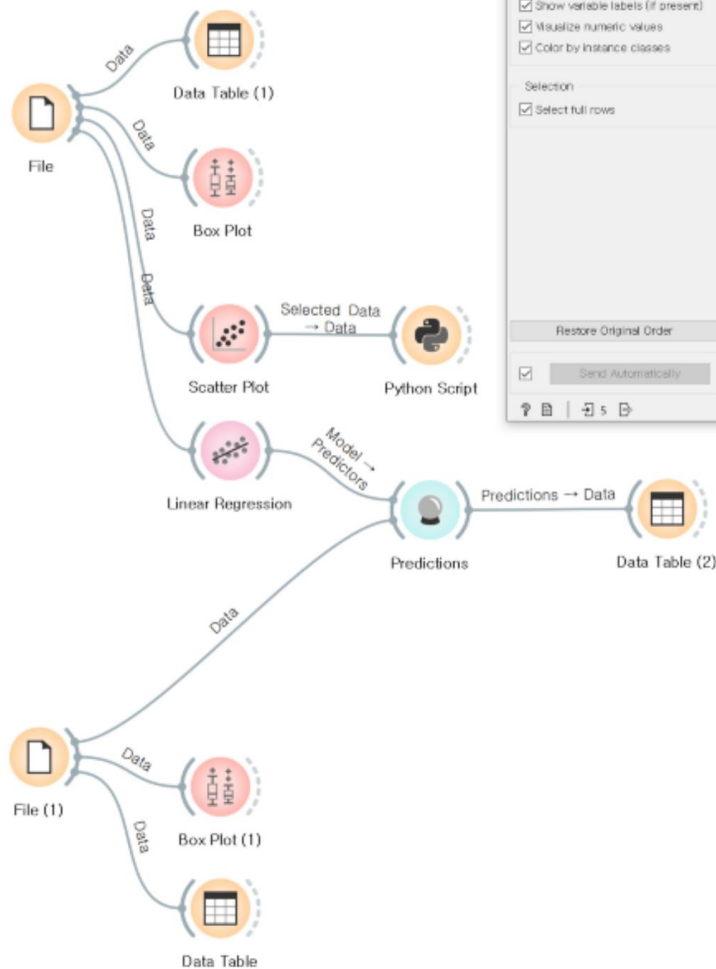
통계의 시각화



통계의 시각화

	Name	Type	Role	Values
1	날짜	 datetime	meta	
2	온도	 numeric	feature	
3	판매량	 numeric	skip	
4	오픈시간	 datetime	skip	
5	요일	 text	skip	

통계의 시각화



Data Table (2)

	날짜	Linear Regression	온도
1	2020-06-01 00:00	40	20
2	2020-06-02 00:00	42	21
3	2020-06-03 00:00	44	22
4	2020-06-04 00:00	46	23
5	2020-06-05 00:00	48	24

Variables

- ☒ Show variable labels (if present)
- ☒ Visualize numeric values
- ☒ Color by instance classes

Selection

- ☒ Select full rows

Restore Original Order

☒ Send Automatically

질문 있나요?

hsryu13@hongik.ac.kr

