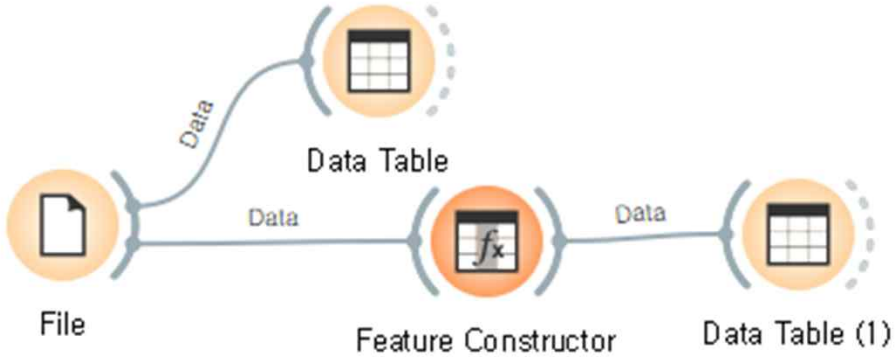
orange 활용 데이터 분석 및 머신 러닝

14차시

데이터의 전처리

Feature Constructor

새로운 특성 만들기



Data Table (1) - Orange

	사망자수	중상자수	경상자수	부상신고자수	사상자수
1	520	6417	6617	517	14071
2	195	1423	2022	209	3849
3	40	591	1371	160	2162
4	26	707	1260	129	2122
5	288	5195	9290	1272	16045
6	181	3657	8468	677	12983
7	441	19893	88838	6945	116117
8	4	297	4003	154	4458
9	370	7924	48345	3172	59811
10	303	11000	50973	4764	67040
11	90	412	466	281	1249
12	26	81	98	32	237
13	340	1376	1745	668	4129
14	4	8	8	1	21
15	92	185	203	46	526
16	30	113	113	24	280
17	130	1285	2213	543	4171
18	1	0	3	0	4

Feature Constructor - Orange

Variable Definitions

New | Name... | Expression...

Select Feat | Select Func

Numeric
Categorical
Text
Date/Time
Duplicate Selected Variable

Send

31 31

Feature Constructor - Orange

Variable Definitions

New | 사상자수 | 사망자수+중상자수+부상신고자수+경상자수

Remove

Select Feature | Select Function

사상자수 := 사망자수+중상자수+부상신고자수+경상자수

전국 전통시장 정보를 활용하여 전처리과정을 알아보자.

위도, 경도를 만들기 위해 주소 채우기

Data Table (2) - Orange

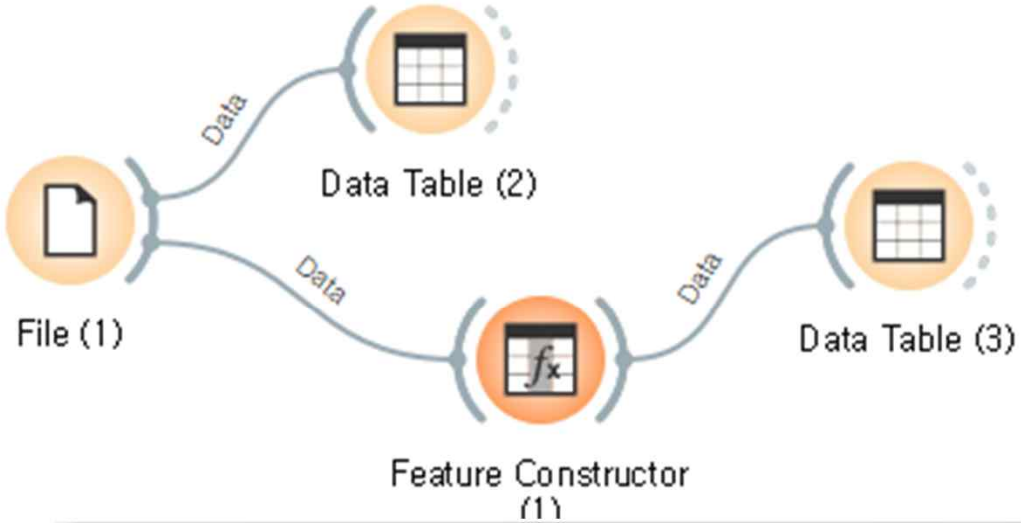
	시장명	소재지도로명주소	소재지지번주소	취급품목
1	중앙시장	강원도 강릉시 금성로21	강원도 강릉시 성남동 50	농산물,축산물,...
2	강릉성남시장	강원도 강릉시 중앙시장길 ...	강원도 강릉시 성남동 53-8	농산물,축산물,...
3	강릉동부시장	강원도 강릉시 수문길19번...	-	농산물,축산물,...
4	강릉서부시장	강원도 강릉시 임영로 155...	강원도 강릉시 용강동 29	농산물,축산물,...
5	주문진종합시장	강원도 강릉시 주문진읍 구...	-	
6	주문진건어물시...	강원도 강릉시 주문진읍 해...	-	
7	주문진수산시장	강원도 강릉시주문진읍 시...	-	
8	간성시장	-	강원도 고성군 간성읍	
9	거진시장	강원도 고성군 거진읍	강원도 고성군 거진읍	
10	동해남부재래시...	강원도 동해시 청운1길 29	강원도 동해시 동회동	
11	동쪽바다중앙시...	강원도 동해시 중앙시장길 ...	강원도 동해시 발한동	
12	북평민속시장	강원도 동해시 오일장길 32	강원도 동해시 북평동	
13	묵호시장	강원도 동해시 일출로 59-1	-	
14	삼척중앙시장	강원도 삼척시진주로 12-21	강원도 삼척시 남양동	
15	도계중앙시장	-	강원도 삼척시 도계읍	
16	도계전두시장	-	강원도 삼척시 도계읍	
17	삼척번개시장	강원도 삼척시 중앙로 14-27	강원도 삼척시 사직동	

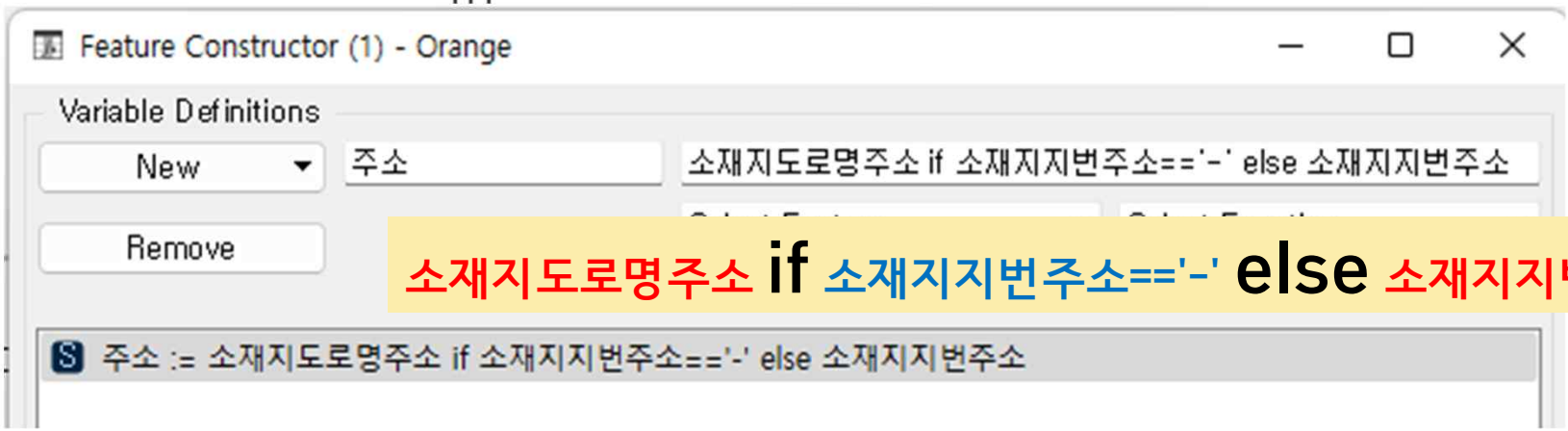
1413 | 1 | 1413

Data Table (3) - Orange

	시장명	소재지도로명주소	소재지지번주소	취급품목	주소
1	중앙시장	강원도 강릉시 ...	강원도 강릉시 성남동 50	농산물,축산물,수...	강원도 강릉시 성남동 50
2	강릉성남시장	강원도 강릉시 ...	강원도 강릉시 성남동 53-8	농산물,축산물,수...	강원도 강릉시 성남동 53-8
3	강릉동부시장	강원도 강릉시 ...	-	농산물,축산물,수...	강원도 강릉시 수문길19번옆길 12-1
4	강릉서부시장	강원도 강릉시 ...	강원도 강릉시 용강동 29	농산물,축산물,수...	강원도 강릉시 용강동 29
5	주문진종합시장	강원도 강릉시 ...	-	농산물,수산물,가...	강원도 강릉시 주문진읍 구시장길 8...
6	주문진건어물시...	강원도 강릉시 ...	-	수산물,가공식품,...	강원도 강릉시 주문진읍 해안로1748
7	주문진수산시장	강원도 강릉시...	-	수산물,음식점업	강원도 강릉시주문진읍 시장1길 4 -...
8	간성시장	-	강원도 고성군 간성읍 신안...	농산물,축산물,수...	강원도 고성군 간성읍 신안리 311-10
9	거진시장	강원도 고성군 ...	강원도 고성군 거진읍 거진...	농산물,축산물,수...	강원도 고성군 거진읍 거진리 272
10	동해남부재래시...	강원도 동해시 ...	강원도 동해시 동회동 442	농산물,축산물,수...	강원도 동해시 동회동 442
11	동쪽바다중앙시...	강원도 동해시 ...	강원도 동해시 발한동 7-1	농산물,축산물,수...	강원도 동해시 발한동 7-1
12	북평민속시장	강원도 동해시 ...	강원도 동해시 북평동 15-5	농산물,축산물,수...	강원도 동해시 북평동 15-5
13	묵호시장	강원도 동해시 ...	-	농산물,축산물,수...	강원도 동해시 일출로 59-1
14	삼척중앙시장	강원도 삼척시...	강원도 삼척시 남양동 55-4	농산물,축산물,수...	강원도 삼척시 남양동 55-4
15	도계중앙시장	-	강원도 삼척시 도계읍 도계...	농산물,가공식품,...	강원도 삼척시 도계읍 도계리 390-1
16	도계전두시장	-	강원도 삼척시 도계읍 전두...	농산물,축산물,수...	강원도 삼척시 도계읍 전두리 81-20

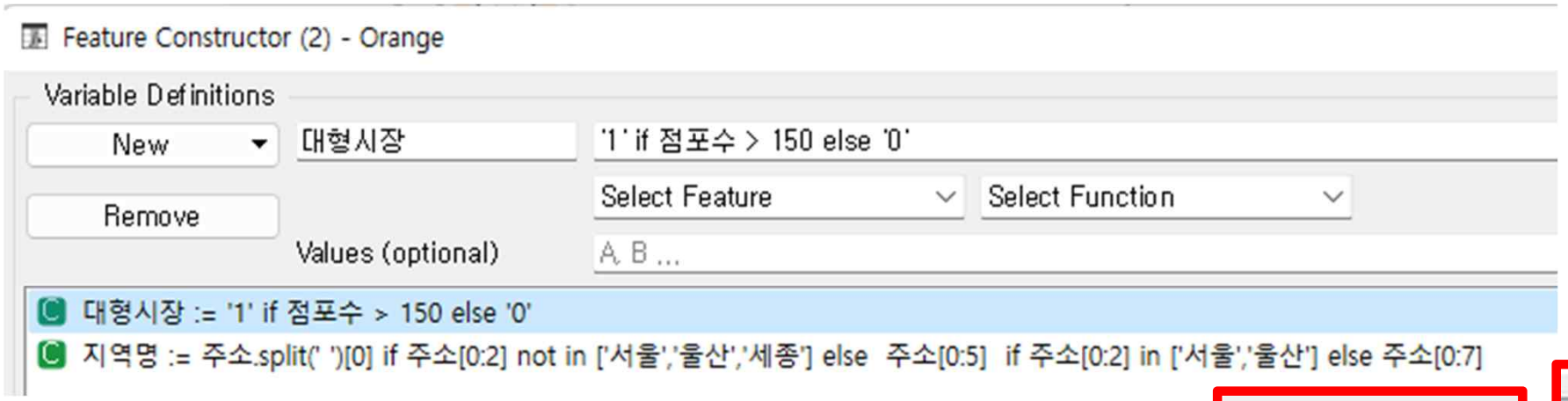
위도, 경도를 만들기 위해 주소 채우기





소재지도로명주소 if 소재지지번주소=='-' else 소재지지번주소

새로운 특성(feature) 만들기



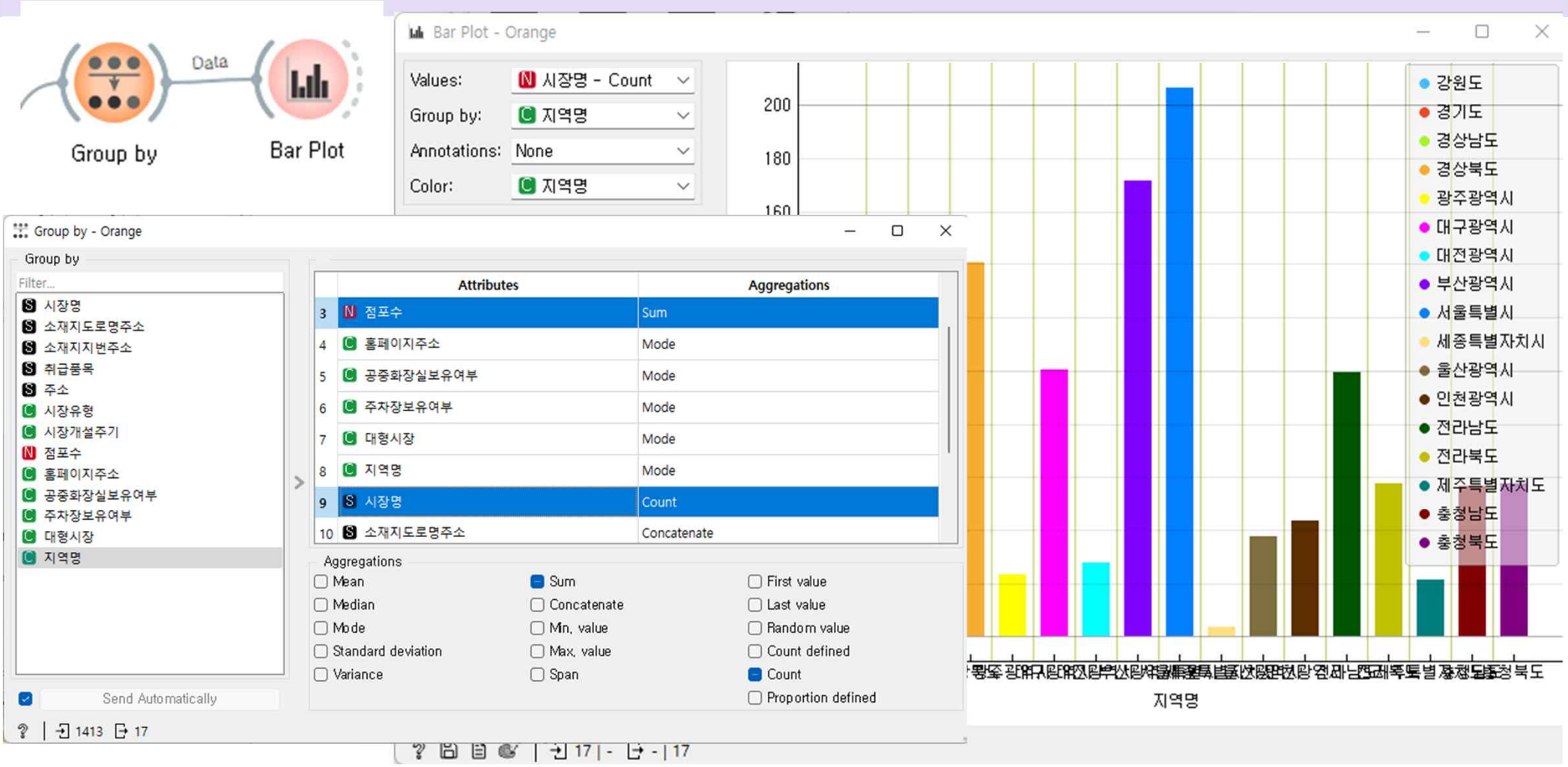
'1' if 점포수 > 150 else '0'

주소.split(' ')[0] if 주소[0:2] not in ['서울','울산','세종'] else 주소[0:5] if 주소[0:2] in ['서울','울산'] else 주소[0:7]

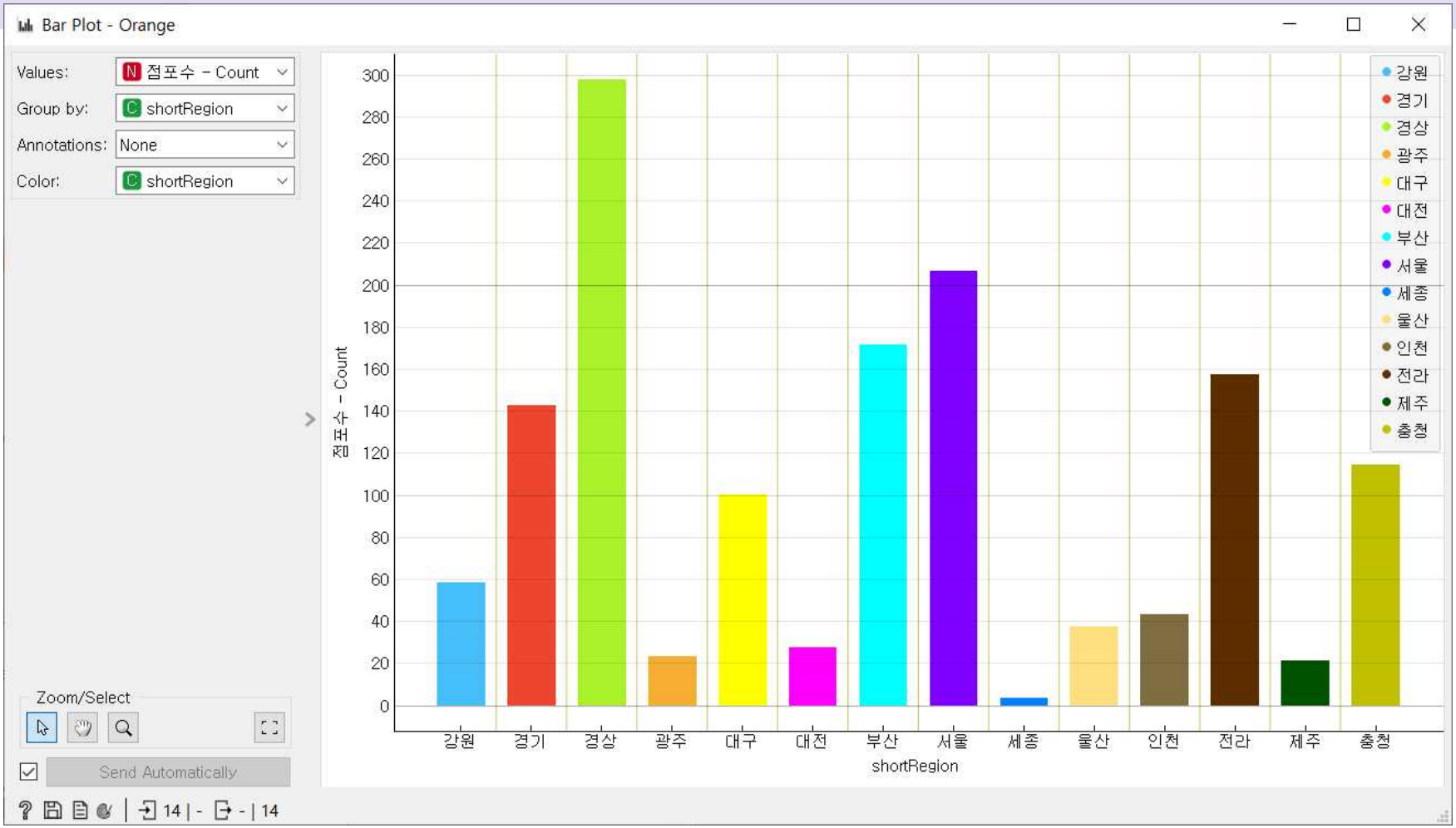
대형시장
1
0
1
0

지역명
강원도
강원도
강원도
강원도
강원도
강원도

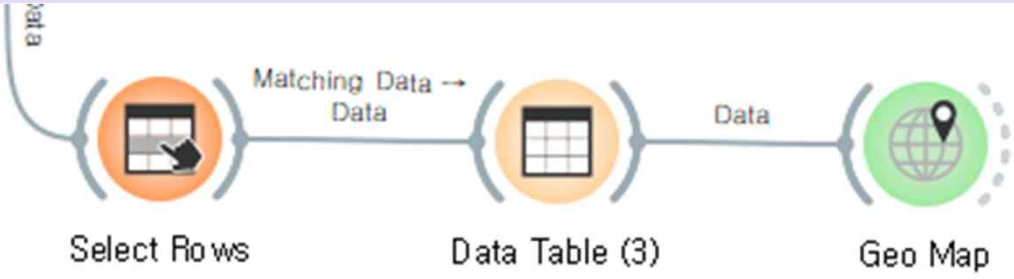
새로 만든 특성을 시각화해서 살펴보기

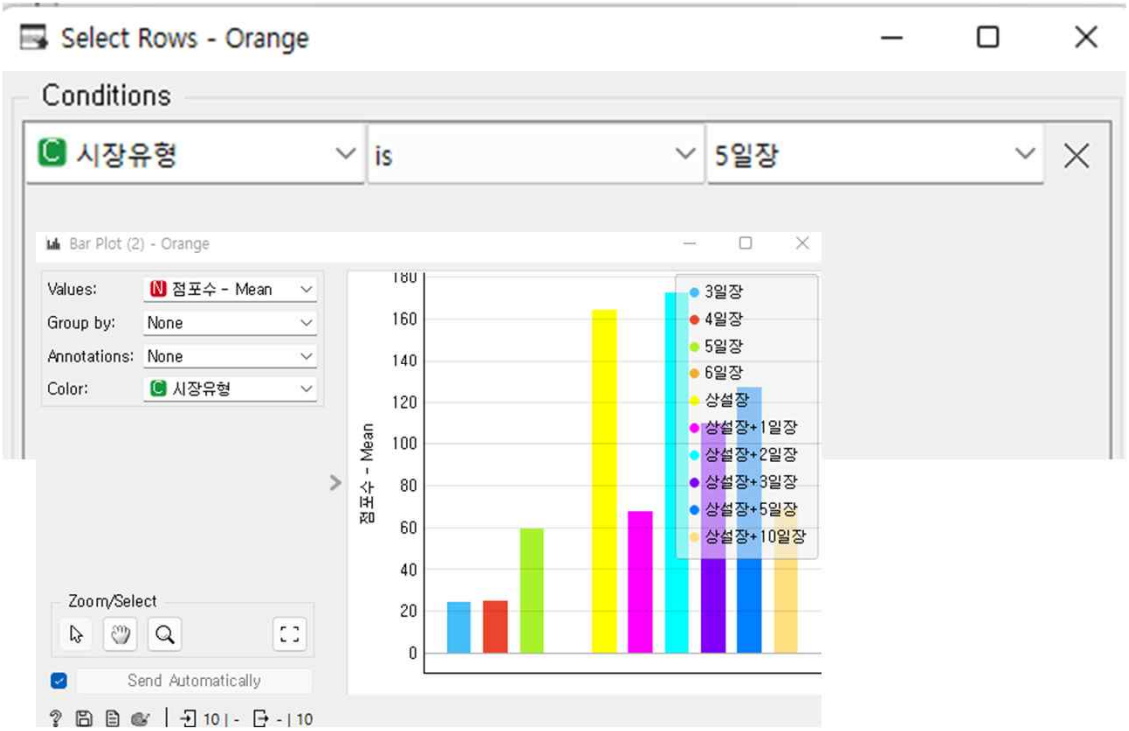


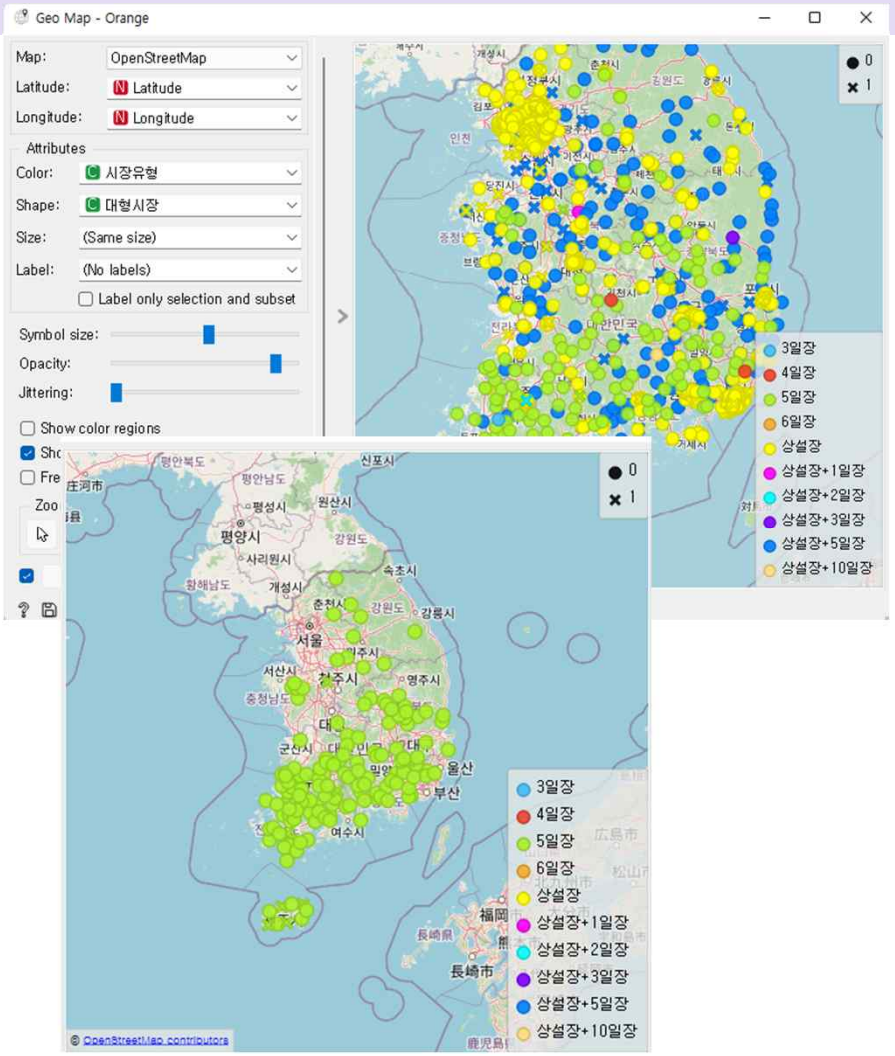
전통시장현황


Bar Plot - Orange
Values: 점포수 – Count
Group by: shortRegion
Annotations: None
Color: shortRegion
Zoom/Select
Send Automatically
14 | - - | 14
점포수 – Count
300
280
260
240
220
200
180
160
140
120
100
80
60
40
20
0
강원
경기
경상
광주
대구
대전
부산
서울
세종
울산
인천
전라
제주
충청
shortRegion

전통시장 시장 유형별 정보 및 시각화







데이터 유형별 전처리

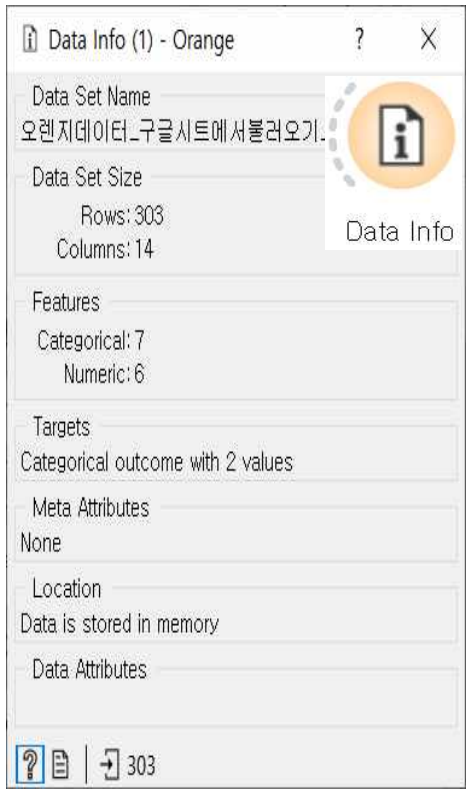
데이터의 종류

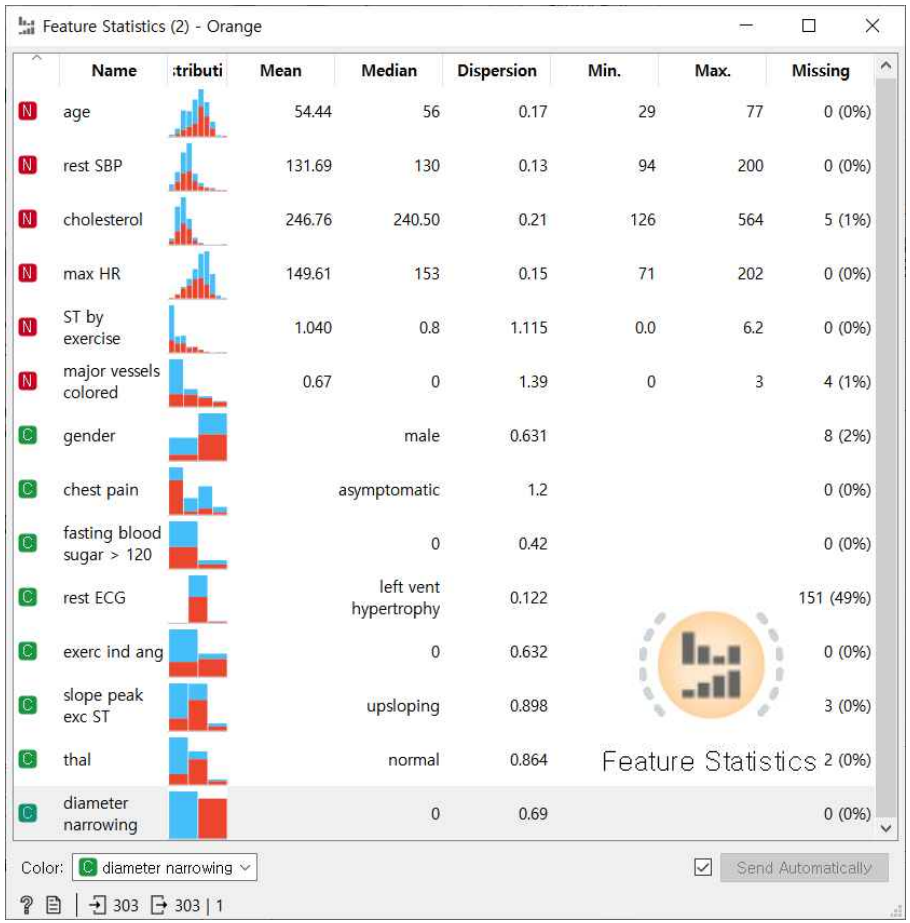
			비교	최빈값	순서	중앙값	평균값	덧셈 뺄셈	곱셈 나눗셈
Categorical Data (범주형)	**Nominal**	순서를 매길 수 없음 • 긍정/부정, 개/고양이	○	○					
	Ordinal	순서가 있으나 항목별 차이가 일정하지 않음 • 만족도, 별점	○	○	○	○			
Numeric Data (연속형)	**Interval**	0점이 존재하지 않음 • 기온, 연도, IQ	○	○	○	○	○	○	
	Ratio	0점이 존재함(Real Zero) • 길이, 무게, 나이, 개수	○	○	○	○	○	○	○

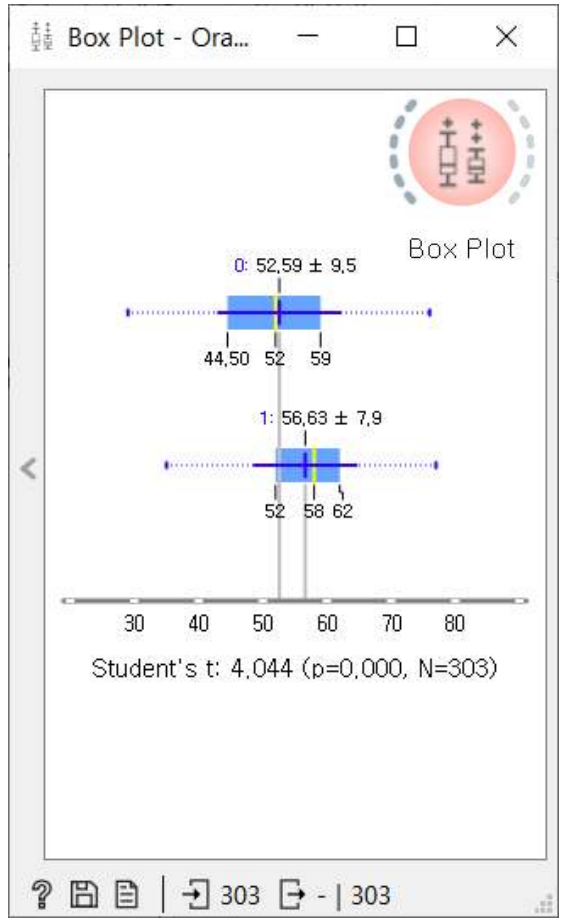
심장병 데이터

속성	의미	속성	의미
age	나이	**max HR**	최대 심장 박동수
gender	성별(male, female)	**exerc ind ang**	협심증 유발 운동(0, 1))
chest pain	가슴 통증 유형 (4가지)	**ST by exercise**	비교적 안정되기까지 운동에 의해 유발되는 ST 분절 하강 정도
rest SBP	입원 시 안정 혈압(mmHg)	**slope peak exc ST**	최대 운동 ST 분절 기울기(0~3)
cholesterol	혈청 콜레스테롤(mg/dl)	**major vessels colored**	형광 투시된 주요 혈관 수(0~3)
fast blood sugar > 120	공복 혈당(0, 1)	**thal**	탈륨 스트레스 검사 결과(3가지)
rest ECG	안정 심전도 결과(3가지)	**diameter narrowing**	심장병 진단(0, 1)

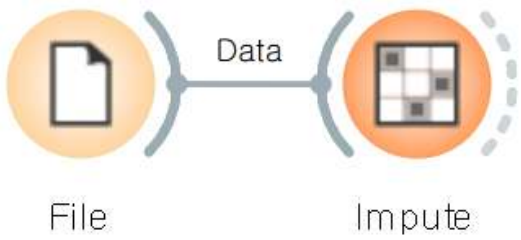
데이터 살펴보기


Data Info (1) - Orange
Data Set Name
오렌지데이터_구글시트에서불러오기.
Data Set Size
Rows: 303
Columns: 14
Features
Categorical: 7
Numeric: 6
Targets
Categorical outcome with 2 values
Meta Attributes
None
Location
Data is stored in memory
Data Attributes
303
Data Info


Feature Statistics (2) - Orange
Name
:tributi
Mean
Median
Dispersion
Min.
Max.
Missing
age 54.44 56 0.17 29 77 0 (0%)
rest SBP 131.69 130 0.13 94 200 0 (0%)
cholesterol 246.76 240.50 0.21 126 564 5 (1%)
max HR 149.61 153 0.15 71 202 0 (0%)
ST by exercise 1.040 0.8 1.115 0.0 6.2 0 (0%)
major vessels colored 0.67 0 1.39 0 3 4 (1%)
gender male 0.631 8 (2%)
chest pain asymptomatic 1.2 0 (0%)
fasting blood sugar > 120 0 0.42 0 (0%)
rest ECG left vent hypertrophy 0.122 151 (49%)
exerc ind ang 0 0.632 0 (0%)
slope peak exc ST upsloping 0.898 3 (0%)
thal normal 0.864 2 (0%)
diameter narrowing 0 0.69 0 (0%)
Color: diameter narrowing
Send Automatically
303
303 | 1
Feature Statistics


Box Plot - Ora...
Box Plot
0: 52,59 ± 9,5
44,50 52 59
1: 56,63 ± 7,9
52 58 62
30 40 50 60 70 80
Student's t: 4,044 (p=0,000, N=303)
303
- | 303

데이터 전처리



Impute (2) - Orange

Default Method

- (•) Don't impute
- ( ) Average/Most frequent
- ( ) As a distinct value
- ( ) Fixed values; numeric variables: 0 , time: 1970-01-01 09:00:00
- ( ) Model-based imputer (simple tree)
- ( ) Random values
- ( ) Remove instances with unknown values

Individual Attribute Settings

Filter...

- N age
- C gender
- C chest pain
- N rest SBP
- N cholesterol
- C fasting blood sugar > 120
- C rest ECG
- N max HR
- C exerc ind ang
- N ST by exercise

- ( ) Default (above)
- ( ) Don't impute
- ( ) Average/Most frequent
- ( ) As a distinct value
- ( ) Model-based imputer (simple tree)
- ( ) Random values
- ( ) Remove instances with unknown values
- ( ) Fixed value

Restore All to Default

Apply Automatically

303 | - 303

Feature Statistics (2) - Orange

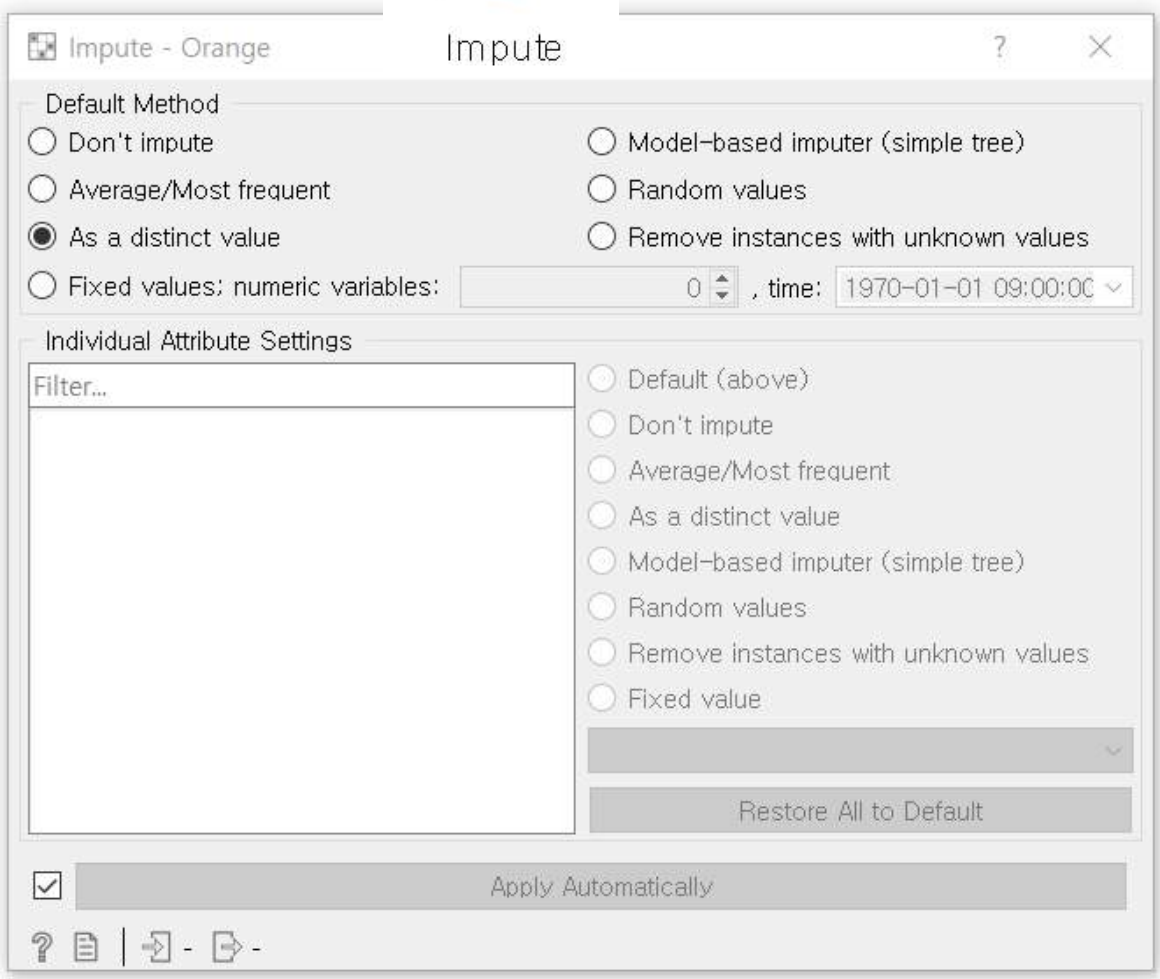
	Name	Distribution	Mean	Median	Dispersion	Min.	Max.	Missing
N	max HR		149.61	153	0.15	71	202	0 (0%)
N	ST by exercise		1.040	0.8	1.115	0.0	6.2	0 (0%)
N	major vessels colored		0.67	0	1.39	0	3	4 (1%)
C	gender			male	0.631			8 (2%)
C	chest pain			asymptomatic	1.2			0 (0%)
C	fasting blood sugar > 120			0	0.42			0 (0%)
C	rest ECG			left vent hypertrophy	0.122			151 (49%)
C	exerc ind ang			0	0.632			0 (0%)
C	slope peak exc ST			upsloping	0.898			3 (0%)

Color: C diameter narrowing

Send Automatically

303 303 | 1

impute



- Average/Most frequent:
: 평균값이나 최빈값으로 채움

- As a distinct value:
: 채워져있음/비어있음을 나타내는 새로운 컬럼 추가. 숫자데이터에만 적용 가능.

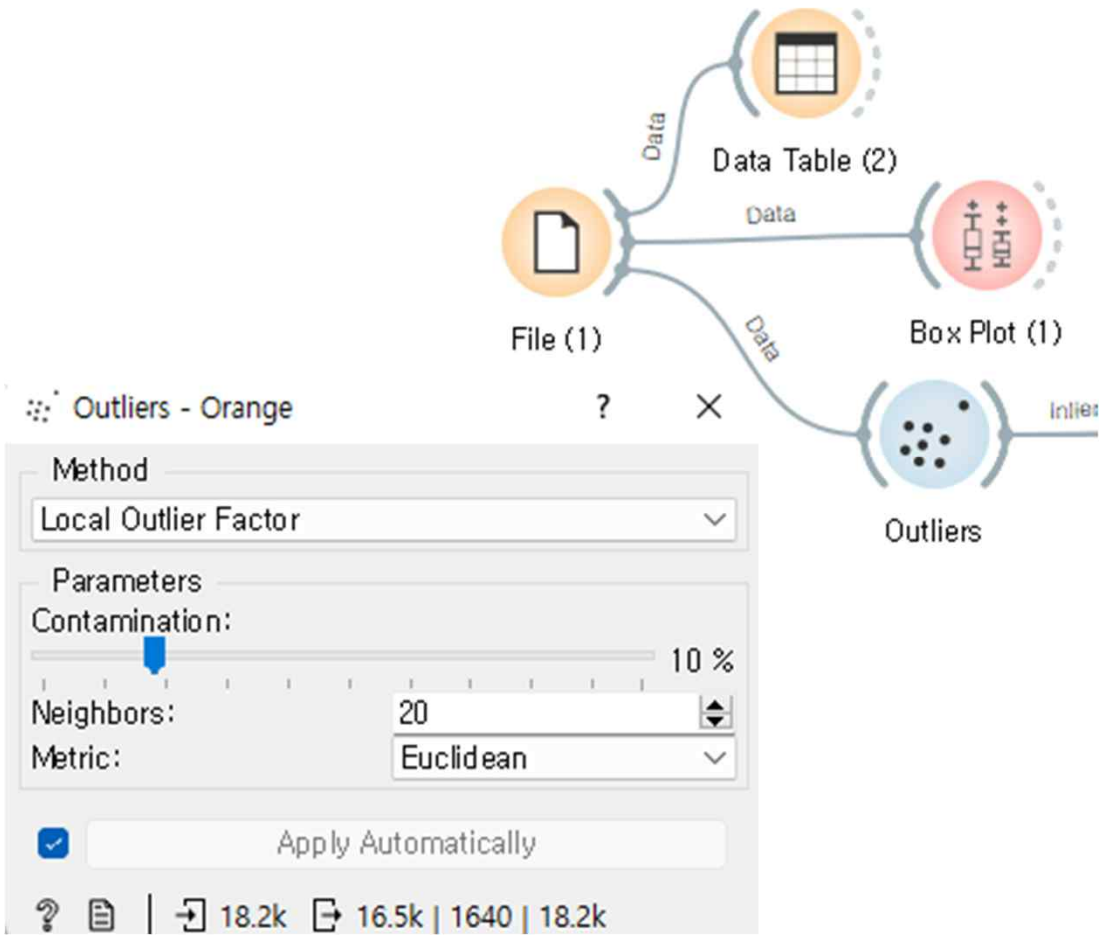
- Fixed values:
: 고정된 값으로 채움

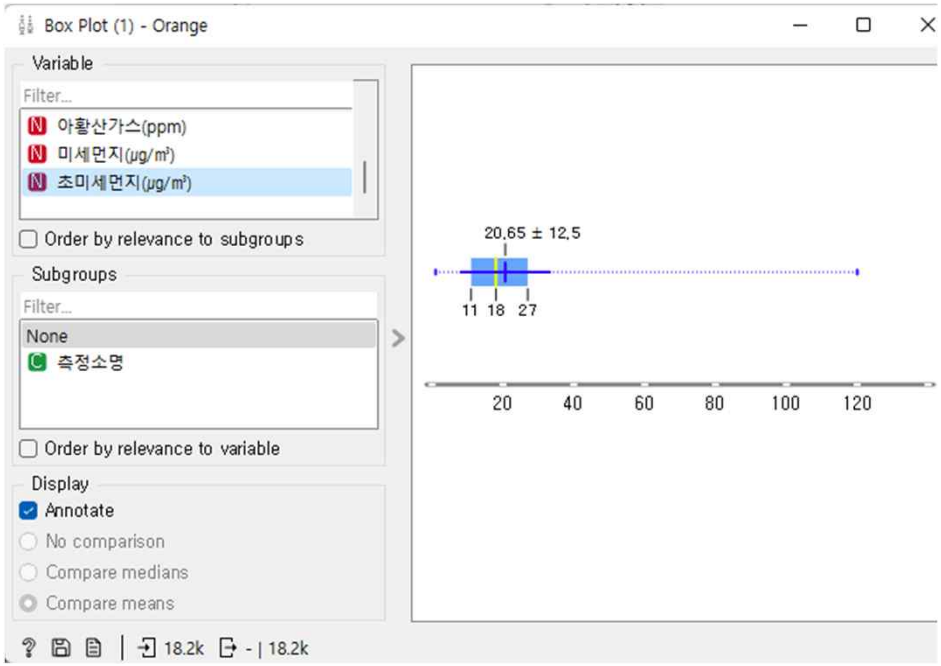
- Model-based imputer (simple tree):
: 다른 특성값들의 분포를 보고 어떤 값을 채워 넣을지 결정

- Random values:
: 랜덤값으로 채움

- Remove instances with unknown values:
: 결측치가 있는 인스턴스를 삭제함

Outlier 제거하기

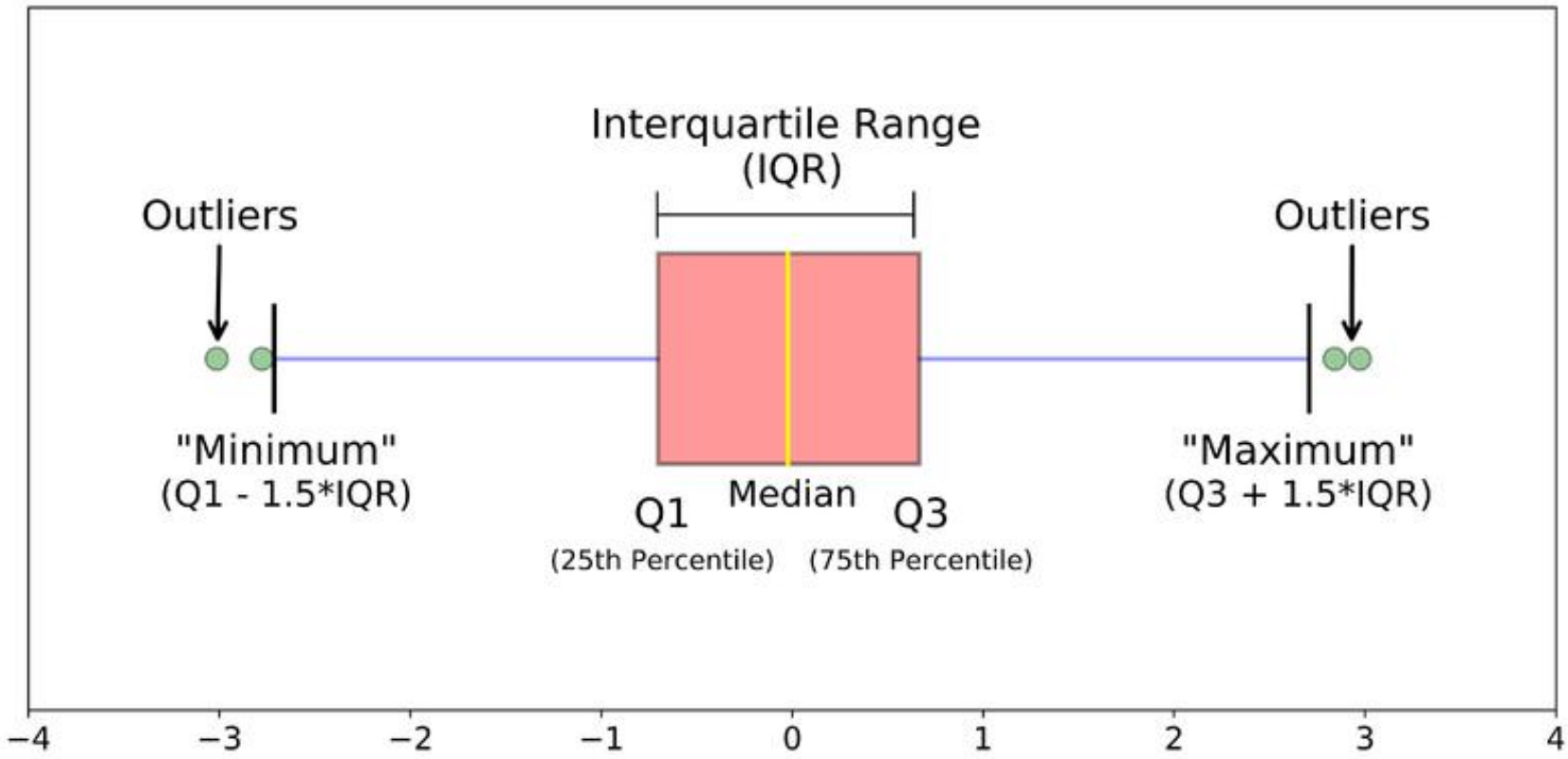




이상치 판정

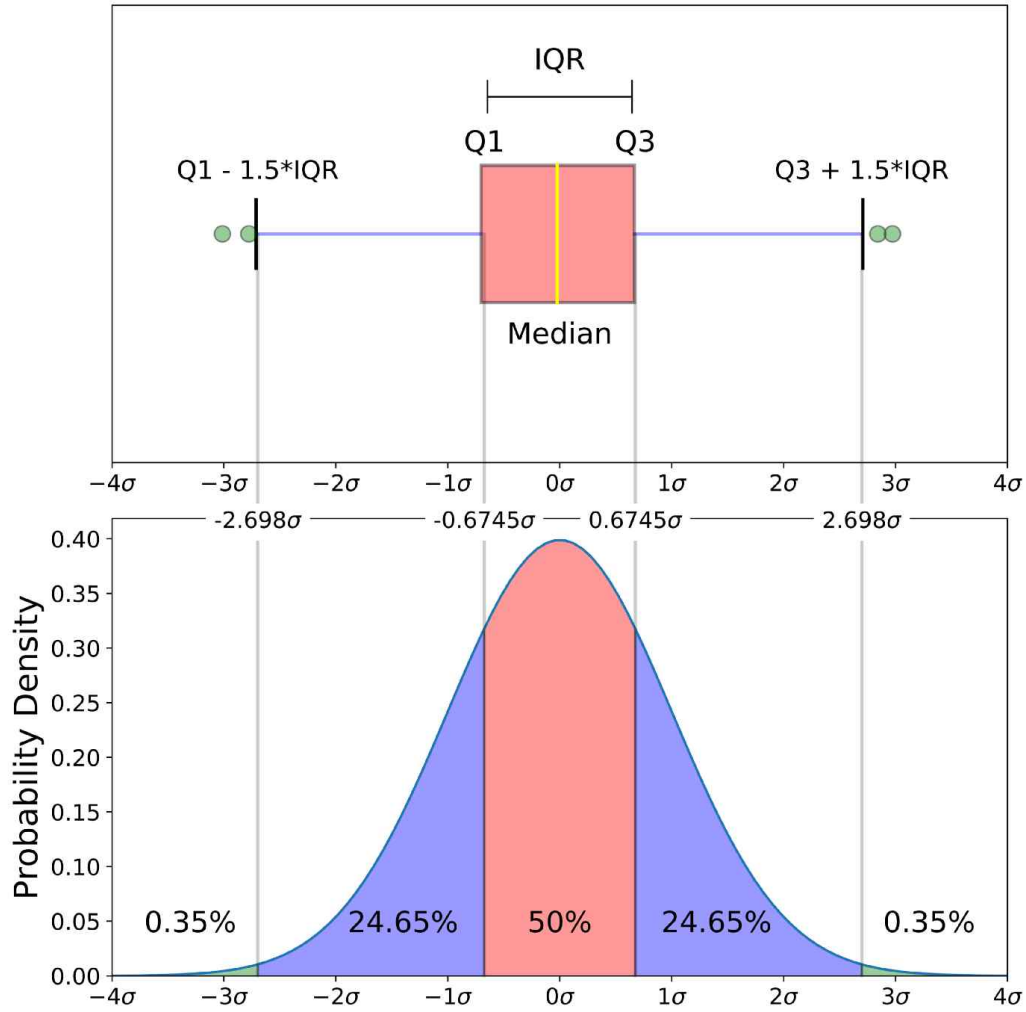
Box Plot

Outliers



https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

이상치 판별



https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

이상치 제거 후


Box Plot (1) - Orange
Variable
Filter...
아황산가스(ppm)
미세먼지(μg/m³)
초미세먼지(μg/m³)
Order by relevance to subgroups
Subgroups
Filter...
None
측정소명
Order by relevance to variable
Display
Annotate
No comparison
Compare medians
Compare means
18.2k
- | 18.2k
20,65 ± 12,5
11 18 27
20 40 60 80 100 120
Inliers → Data
Outliers
Box Plot (2)
20,71 ± 12,4
12 18 27
10 20 30 40 50 60 70 80 90 100
Annotate
No comparison
Compare medians
Compare means
16.5k
- | 16.5k

데이터분석 연습

- 세계 사회경제 지표를 활용한 데이터 분석 및 기대수명 예측