

탐색적 데이터 분석(실습)

홍익 대학교
Hyun-Sun Ryu

타이타닉 데이터

- 건조 당시 세계 최대의 여객선이었지만, 1912년의 최초이자 최후의 항해 때 빙산과 충돌해 침몰한 여객선.

<https://www.kaggle.com/competitions/titanic/data>



타이타닉 데이터 분석 프로세스

1. 문제 정의하기

Titanic 탑승자 데이터를 활용하여 생존자 예측하기

타이타닉 데이터 분석 프로세스

2. 데이터 읽기

Passenger	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, female	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr.	male	35	0	0	373450	8.05		S
6	0	3	Moran, M	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, male	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, M	male	2	3	1	349909	21.075		S
9	1	3	Johnson, M	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, M	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, M	female	58	0	0	113783	26.55	C103	S
13	0	3	Saunders	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, M	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, M	female	55	0	0	248706	16		S
17	0	3	Rice, M	male	2	4	1	382652	29.125		Q
18	1	2	Williams, I	male		0	0	244373	13		S
19	0	3	Vander Pl	female	31	1	0	345763	18		S
20	1	3	Masselman	female		0	0	2649	7.225		C
21	0	2	Fynney, M	male	35	0	0	239865	26		S
22	1	2	Beesley, M	male	34	0	0	248698	13	D56	S
23	1	3	McGowan	female	15	0	0	330923	8.0292		Q
24	1	1	Sloper, Mr	male	28	0	0	113788	35.5	A6	S
25	0	3	Palsson, M	female	8	3	1	349909	21.075		S
26	1	2	Asplund, M	female	28	1	5	247077	31.2875		C

타이타닉 데이터 분석 프로세스

변수명	내용	입력형식	속성
Passenger	ID	숫자	meta
Survived	생존여부	숫자	Categorical
Pclass	티켓 등급	숫자	Categorical
Name	이름	텍스트	Meta
Sex	성별	텍스트	Categorical
Age	나이	숫자	Numeric
Sib Sp	함께 탑승한 형제 또는 배우자수	숫자	Numeric
Parch	함께 탑승한 부모 또는 자녀수	숫자	Numeric
Ticket	티켓 번호	텍스트, 숫자	Categorical
Fare	운임	숫자	Numeric
Cabin	선실번호	텍스트, 숫자	Categorical
Embarked	승선한 항구	텍스트	Text

타이타닉 데이터 분석 프로세스

3. 데이터 불러오기



File

- Data – File 에서 저장된 train 파일 불러오기

타이타닉 데이터 분석 프로세스

4. 데이터 Type 과 Role 확인

Columns (Double click to edit)				
	Name	Type	Role	Values
1	PassengerId	N numeric	feature	
2	Survived	C categorical	feature	0, 1
3	Pclass	N numeric	feature	
4	Sex	C categorical	feature	female, male
5	Age	N numeric	feature	
6	SibSp	N numeric	feature	
7	Parch	N numeric	feature	
8	Fare	N numeric	feature	
9	Embarked	C categorical	feature	C, Q, S
10	Name	S text	meta	
11	Ticket	S text	meta	
12	Cabin	S text	meta	

- Type과 Role을 확인하고 알맞은 값으로 변경할 수 있음
- PassengerId 는 1번부터 891번까지 나열된 일련번호로 Role을 meta로 지정하여 참고용으로 사용

타이타닉 데이터 분석 프로세스

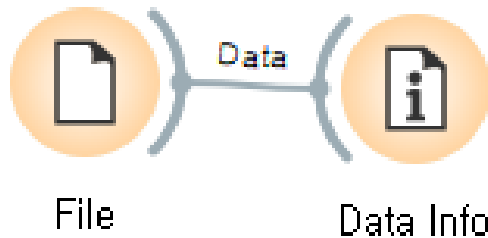
4. 데이터 Type 과 Role 확인

PassengerId	N	numeric	meta
Survived	C	categorical	feature
Pclass	C	categorical	feature
Sex	C	categorical	feature
Age	N	numeric	feature
SibSp	N	numeric	feature
Parch	N	numeric	feature
Fare	N	numeric	feature
Embarked	C	categorical	feature
Name	S	text	meta
Ticket	S	text	meta

- Type과 Role을 확인하고 알맞은 값으로 변경할 수 있음
- PassengerId 는 1번부터 891번까지 나열된 일련번호로 Role을 meta로 지정하여 참고용으로 사용

타이타닉 데이터 분석 프로세스

- 데이터 정보 확인하기
 - Data info 위젯을 연결하여 File의 기본 정보를 확인



타이타닉 데이터 분석 프로세스

- 데이터 정보 확인하기
 - 891개의 Row값이 있으며 12개의 Columns가 있는 것을 알 수 있음,
 - Feature는 Categorical 4항목, Numeric 4항목으로 이루어졌음
 - Targets 값은 File에서 Role을 변경하거나 Select Columns위젯을 이용하는 방법이 있음

Data Info ? X

Data Set Name
train

Data Set Size
Rows: 891
Columns: 12

Features
Categorical: 4
Numeric: 4

Targets
None

Meta Attributes
Categorical: -
Numeric: 1
Text: 3

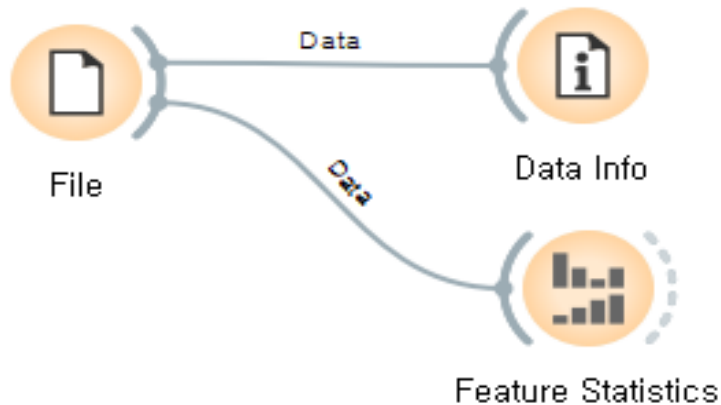
Location
Data is stored in memory

Data Attributes

? | 891

타이타닉 데이터 분석 프로세스

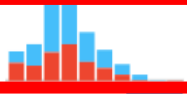






- 데이터 정보 확인하기



- Feature Statistics 위젯을 연결하여
Feature의 통계량을 확인

타이타닉 데이터 분석 프로세스

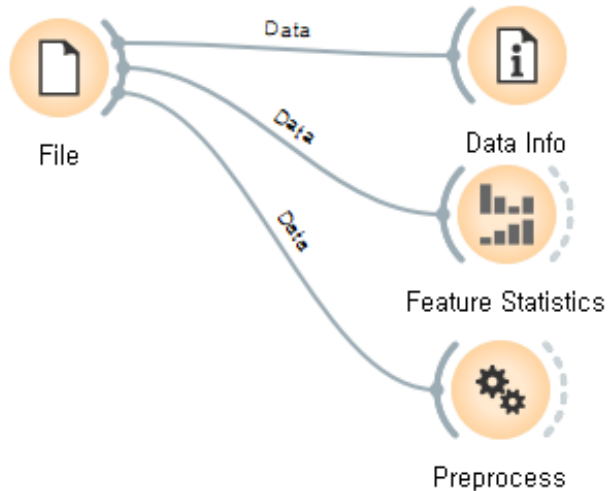
데이터 정보 확인하기

	Name	Distribution	Center	Dispersion	Min.	Max.	Missing
N	Age		29.6991	0.4888	0.42	80.00	177 (19%)
C	Embarked		S	0.76			2 (0%)
C	Survived		0	0.666			0 (0%)
C	Pclass		3	0.998			0 (0%)
C	Sex		male	0.649			0 (0%)
N	SibSp		0.52	2.11	0	8	0 (0%)
N	Spouse		0.00	0.00	0	0	0 (0%)

- Age에 177개(19%), Embarked에는 2개의 결측치
- Age의 결측치가 많으며, 나이에 따라 생존 여부와 관계가 있을 것으로 예상됨

타이타닉 데이터 분석 프로세스

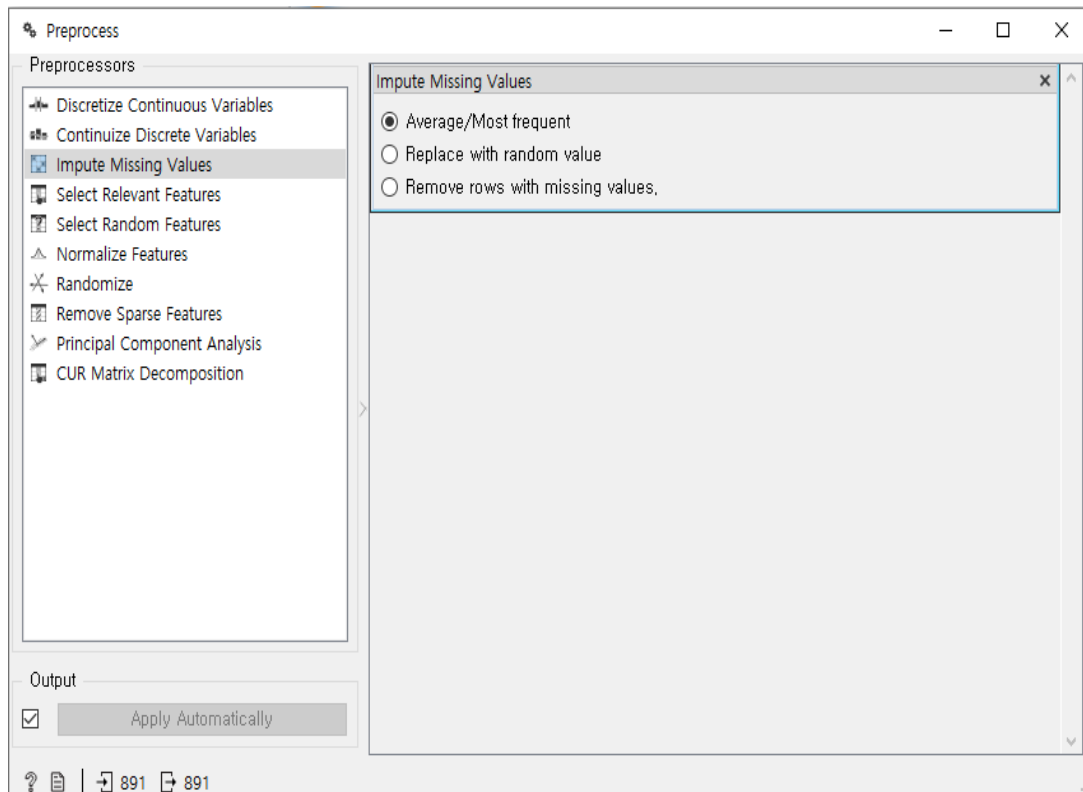
- 데이터 전처리



- Preprocess 위젯을 연결하여 결측치에 대해 데이터 전처리를 실시
- 결측치가 있는 행을 제거하거나 결측값을 임의의 값으로 채우는 방법을 선택

타이타닉 데이터 분석 프로세스

- 데이터 전처리
 - **Average/Most frequent**
결측값을 평균 또는 가장 빈도가 높은 값으로 바꿈

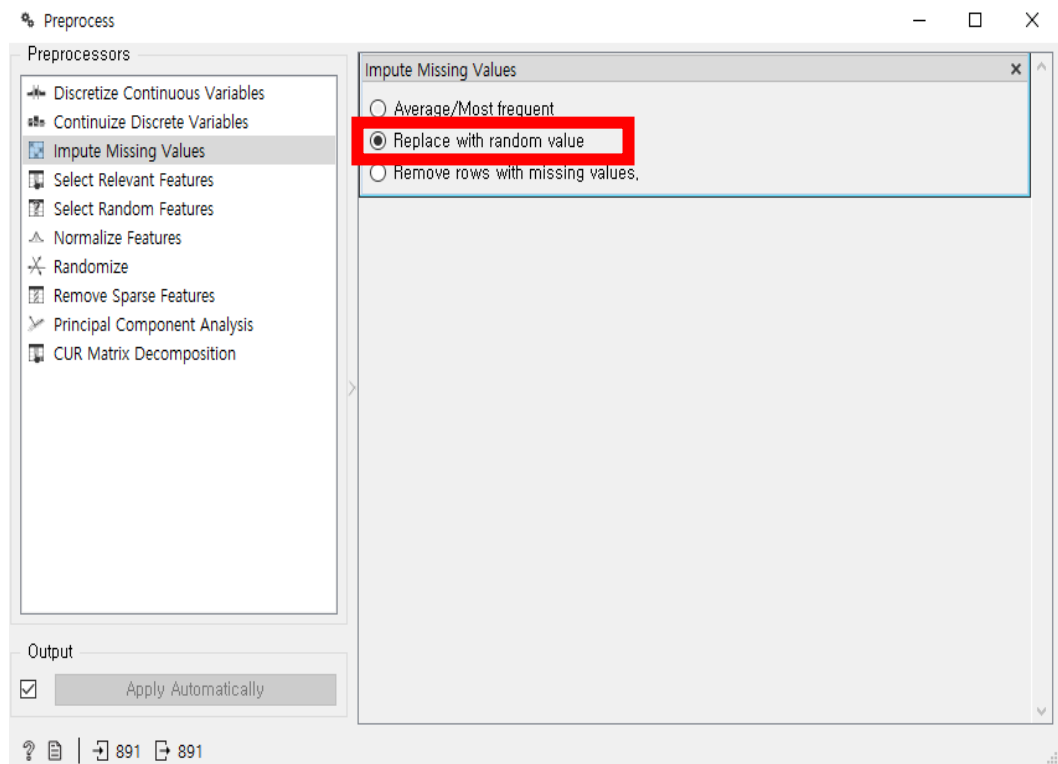


타이타닉 데이터 분석 프로세스

■ 데이터 전처리

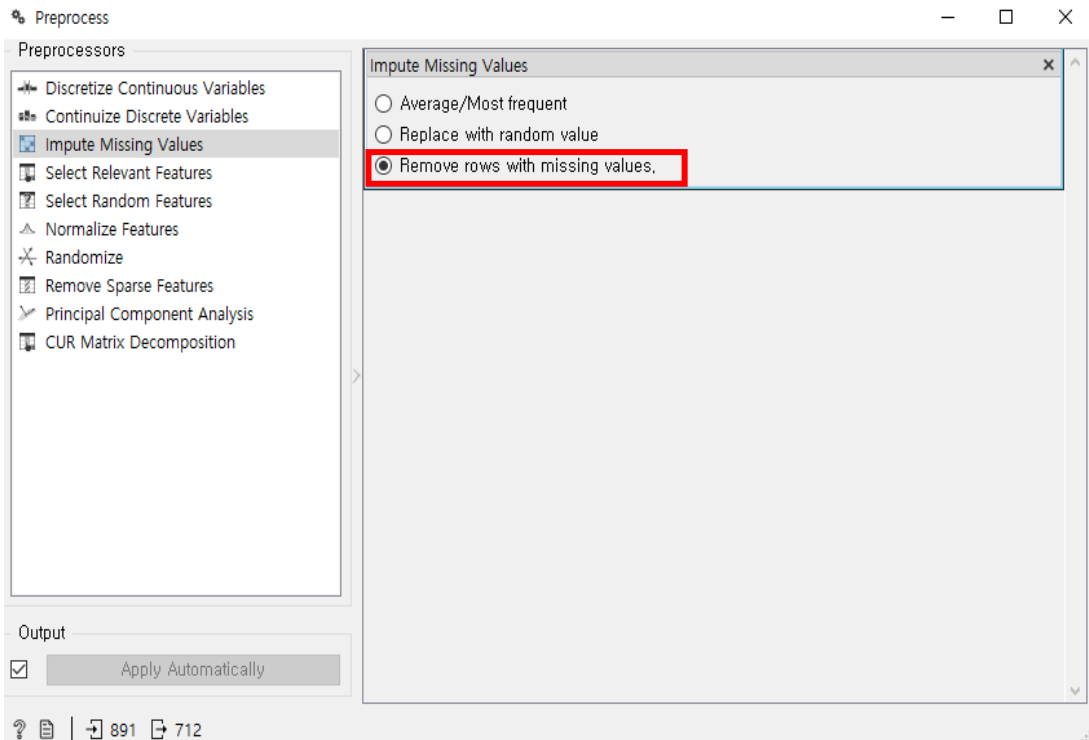
■ Replace with random value

결측값을 각 변수의 범위 내에
있는 임의의 값으로 바꿈



타이타닉 데이터 분석 프로세스

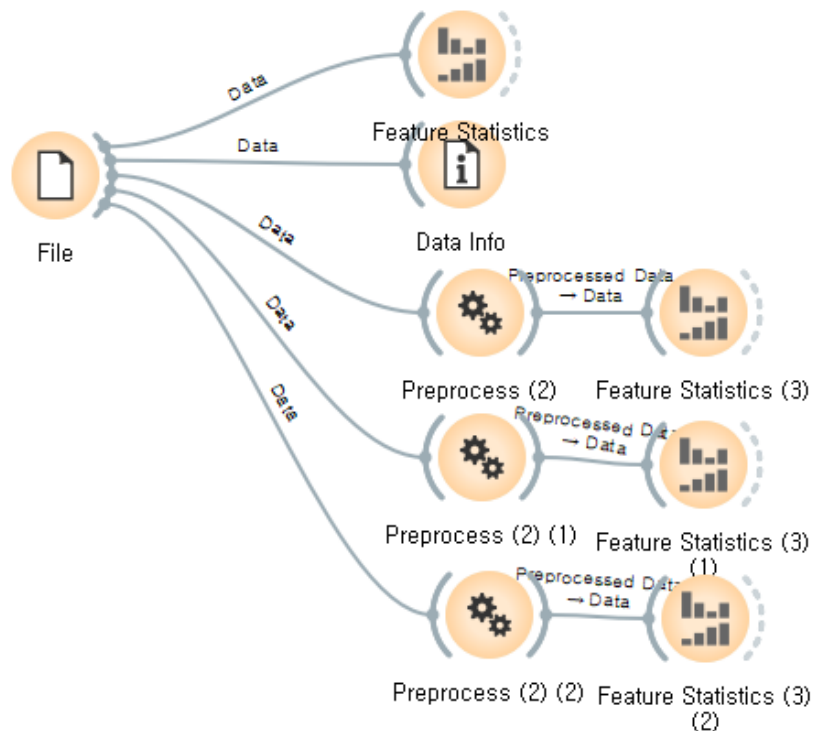
- 데이터 전처리
 - Remove with random value
 - 결측값이 있는 행을 제거
- (본 실습에서 선택)



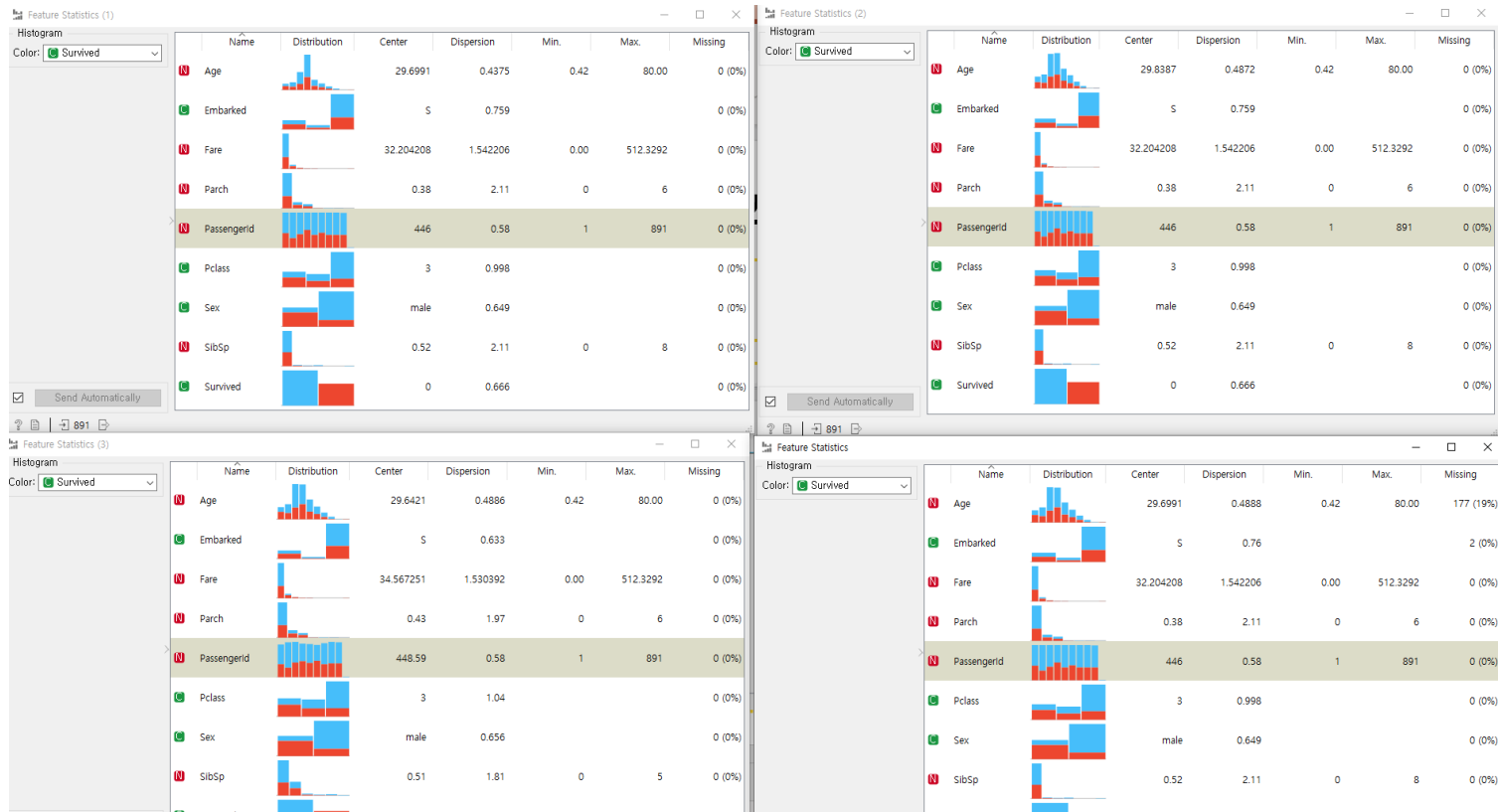
타이타닉 데이터 분석 프로세스

- 데이터 전처리 실습

- 위에 제시된 세가지 방법으로
데이터 전처리를 실행하고 원
데이터의 통계값과 비교

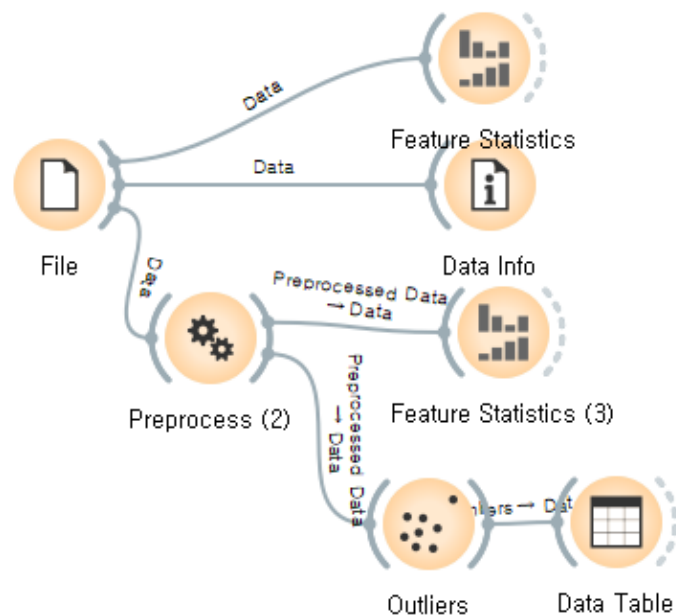


타이타닉 데이터 분석 프로세스



타이타닉 데이터 분석 프로세스

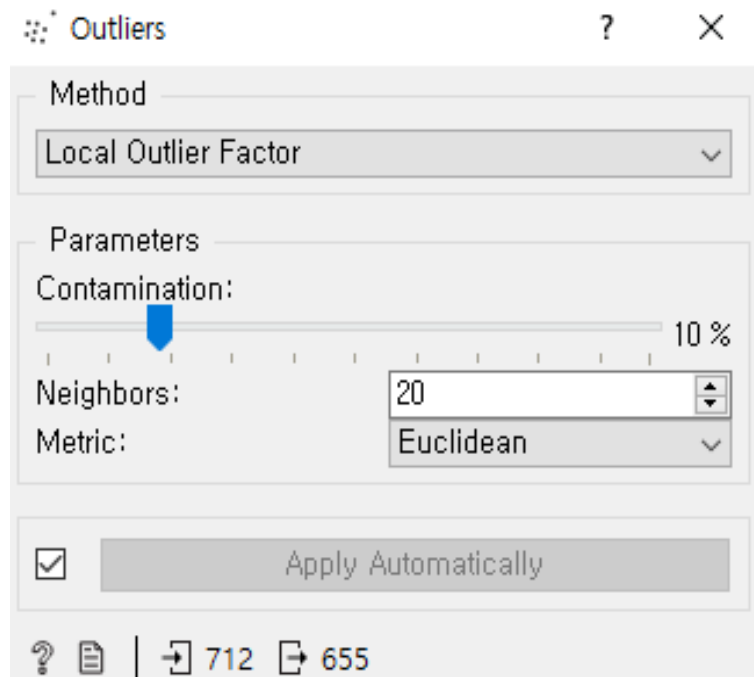
- 이상값 탐지하기



- 이상값 탐지를 위하여 outliers 위젯을 연결
- 위젯을 열고 이상값 탐지를 위한 설정

타이타닉 데이터 분석 프로세스

- 이상값 탐지하기



- Local Outlier Factor는 적당히 높은 차원 데이터 세트에서 이상값 탐지를 수행하는 효율적인 방법 중 하나임
- 이상값 탐지 후 712개의 인스턴스가 655개로 줄어든 것을 확인할 수 있음.

타이타닉 데이터 분석 프로세스

1. 클래스 별로 몇 명이 탑승했을까?
2. 생존한 사람이 많을까? 죽은 사람이 많을까?
3. 생존에 탑승 클래스가 영향을 미칠까?
4. 생존에 성별이 영향을 미칠까?

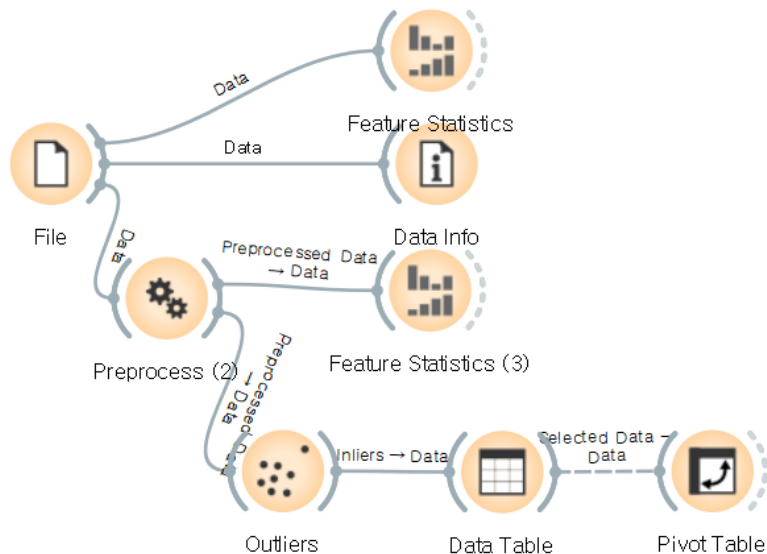
타이타닉 데이터 분석 프로세스

- 질문하기1

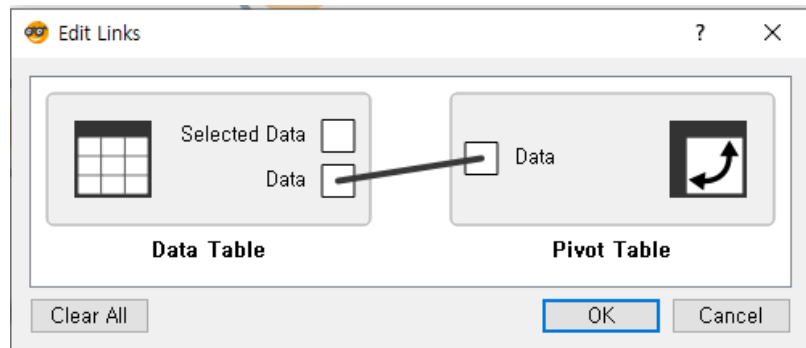
클래스 별로 몇 명이 탑승했을까?

타이타닉 데이터 분석 프로세스

- 질문하기: 데이터 변형 및 시각화 (Pivot Table)



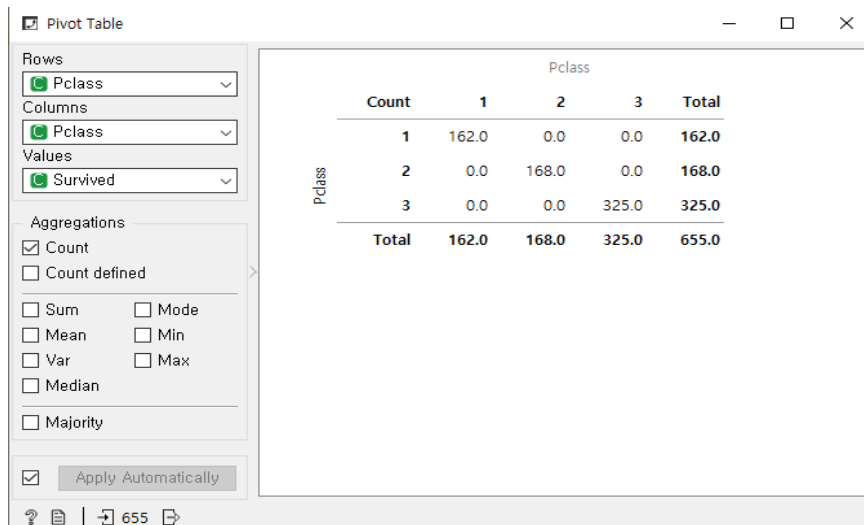
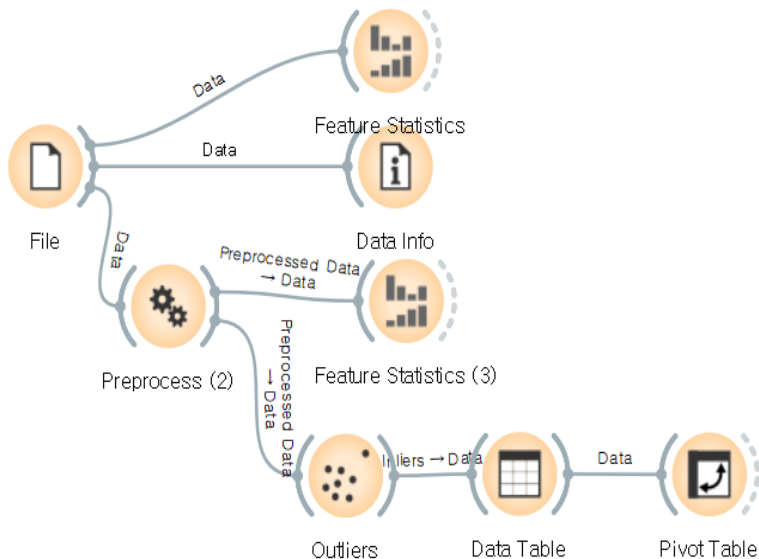
- 클래스 별로 몇 명이 탑승했을까?



타이타닉 데이터 분석 프로세스

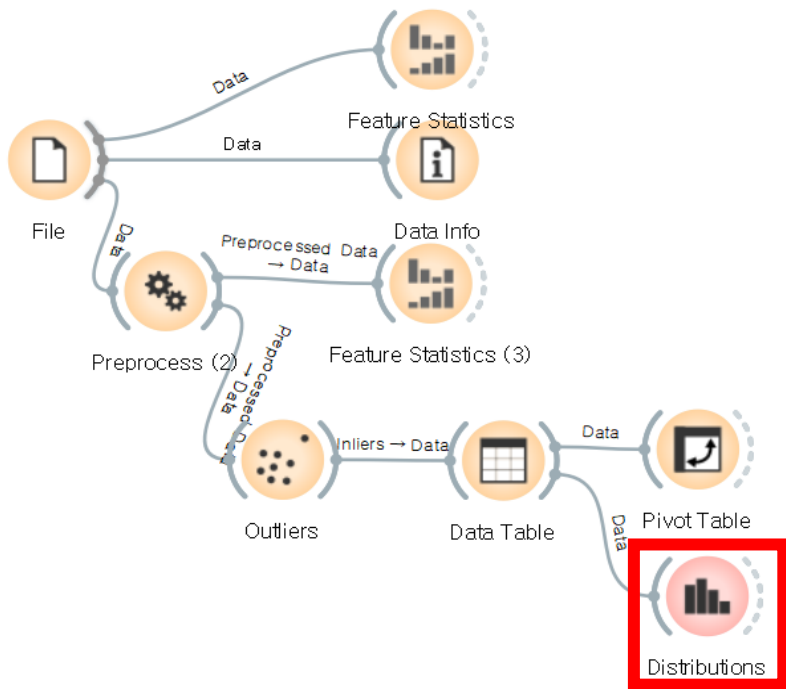
- 질문하기: 데이터 변형 및 시각화(Pivot Table)

- 클래스 별로 몇 명이 탑승했을까?

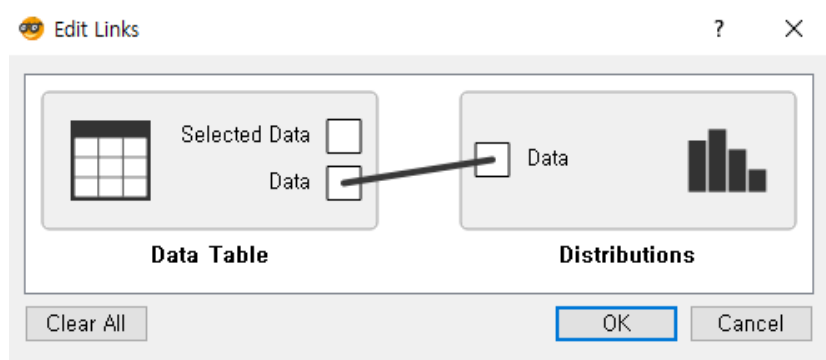


타이타닉 데이터 분석 프로세스

- 질문하기: 데이터 변형 및 시각화 (Distribution)

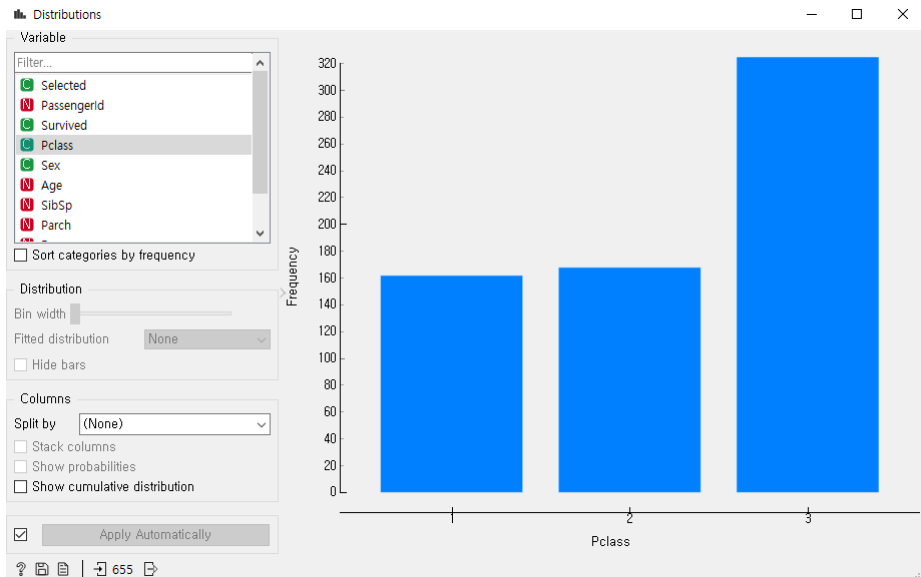


- 클래스 별로 몇 명이 탑승했을까?



타이타닉 데이터 분석 프로세스

- 질문하기: 데이터 변형 및 시각화 (Distribution)



- 클래스 별로 몇 명이 탑승했을까?

타이타닉 데이터 분석 프로세스

- 질문하기: 데이터 변형 및 시각화 → 대답하기

클래스 별로 몇 명이 탑승했을까?

답변: 3등급이 325명으로 가장 많고 1,2등급은 비슷한 인원이 탑승

속성 간의 관계 분석(참고)

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

타이타닉 데이터 분석 프로세스

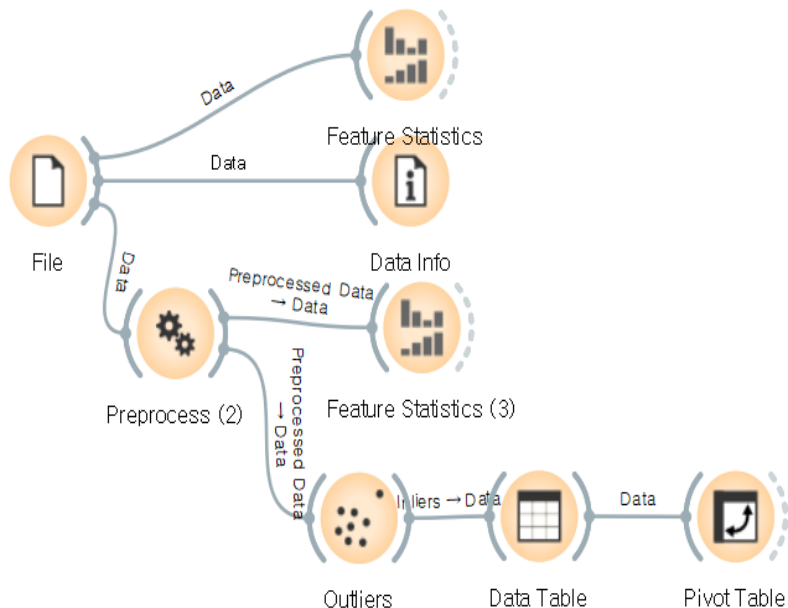
- 질문하기2

생존한 사람이 많을까? 죽은 사람이 많을까?

타이타닉 데이터 분석 프로세스

■ 질문하기2: 데이터 변형 및 시각화 (Pivot Table)

- 생존한 사람이 많을까? 죽은 사람이 많을까?



Pivot Table

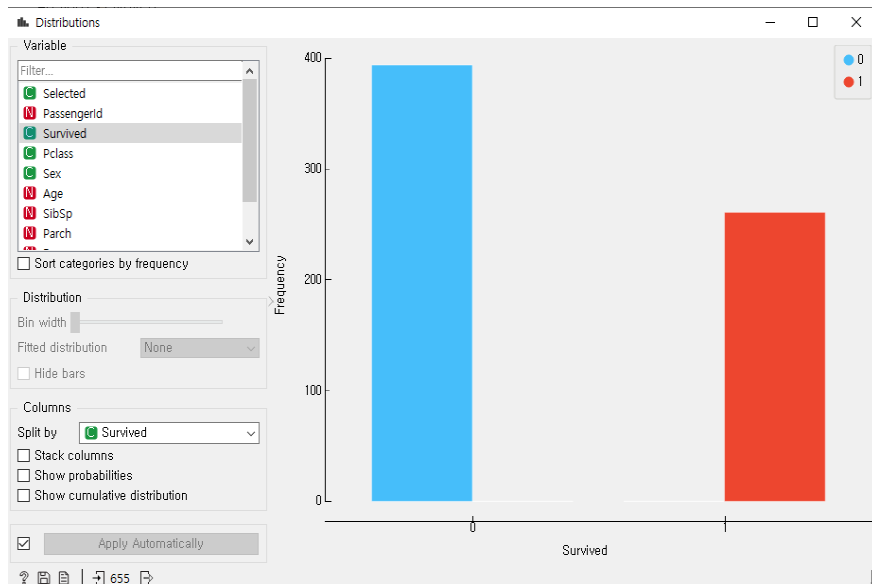
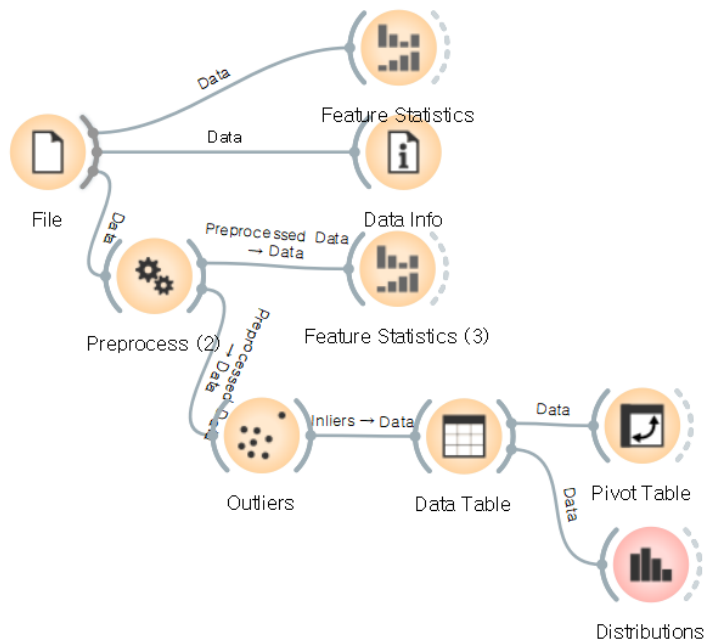
Rows	Survived
Columns	Survived
Values	Survived
Aggregations	<input checked="" type="checkbox"/> Count <input type="checkbox"/> Count defined <input type="checkbox"/> Sum <input type="checkbox"/> Mean <input type="checkbox"/> Mode <input type="checkbox"/> Min

	Survived		
	Count	0	1
Survived	0	394.0	0.0
	1	0.0	261.0
Total		394.0	261.0

타이타닉 데이터 분석 프로세스

■ 질문하기2: 데이터 변형 및 시각화 (Distribution)

- 생존한 사람이 많을까? 죽은 사람이 많을까?



타이타닉 데이터 분석 프로세스

- 질문하기2: 데이터 변형 및 시각화→ 대답하기

생존한 사람이 많을까? 죽은 사람이 많을까?

답변: 사망자가 394명 생존자가 261명으로 사망자가 생존자보다 더 많으며
탑승자의 60%가 사망한 대형사고임을 알 수 있음

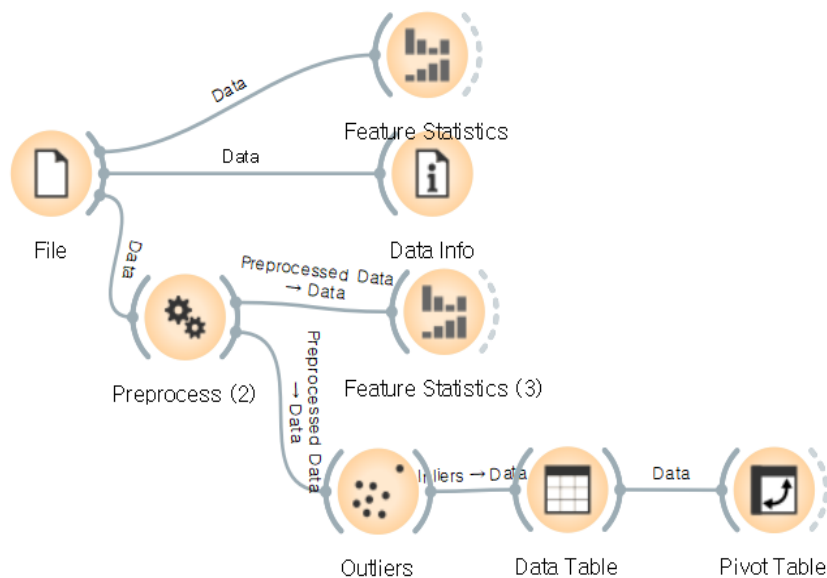
타이타닉 데이터 분석 프로세스

- 질문하기3

생존에 탑승 클래스가 영향을 미칠까?

타이타닉 데이터 분석 프로세스

- 질문하기3: 데이터 변형 및 시각화 (Pivot Table) 생존에 탑승 클래스가 영향을 미칠까?



Pivot Table

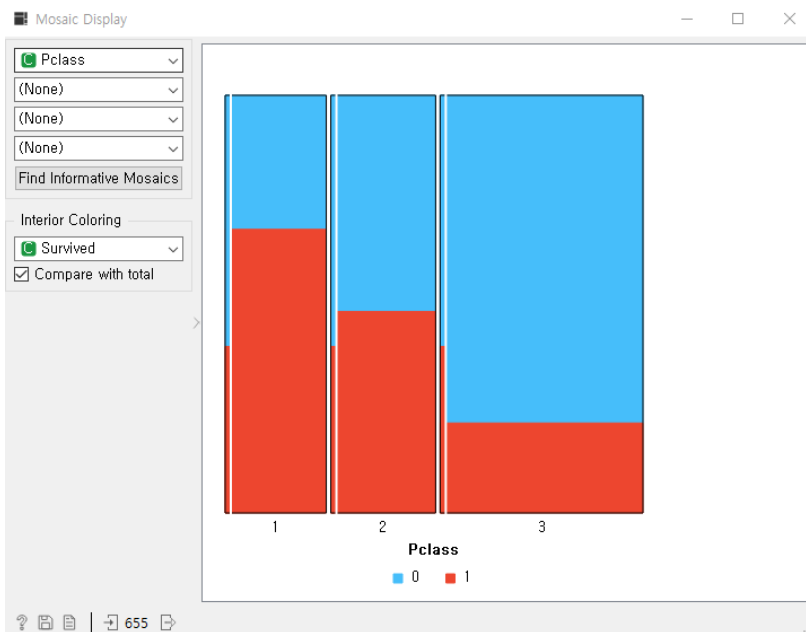
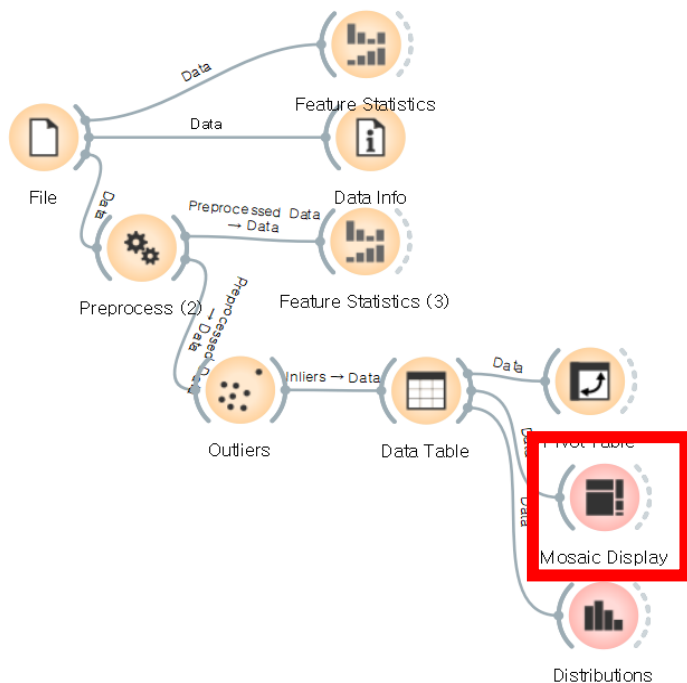
		Survived		
Pclass	Count	0	1	Total
	1	52.0	110.0	162.0
	2	87.0	81.0	168.0
	3	255.0	70.0	325.0
	Total	394.0	261.0	655.0

Rows: Pclass
Columns: Survived
Values: Survived
Aggregations: ☒ Count, ☐ Count defined, ☐ Sum, ☐ Mean, ☐ Var, ☐ Mode, ☐ Min, ☐ Max

타이타닉 데이터 분석 프로세스

■ 질문하기3: 데이터 변형 및 시각화 (Mosaic Display)

생존에 탑승 클래스가 영향을 미칠까?



타이타닉 데이터 분석 프로세스

- 질문하기3: 데이터 변형 및 시각화→ 대답하기

생존에 탑승 클래스가 영향을 미칠까?

답변: 탑승 등급이 높을 수록 생존자가 많은 것을 확인할 수 있음.

따라서 등급이 높을 수록 생존 확률이 높아진다고 예측할 수 있음

타이타닉 데이터 분석 프로세스

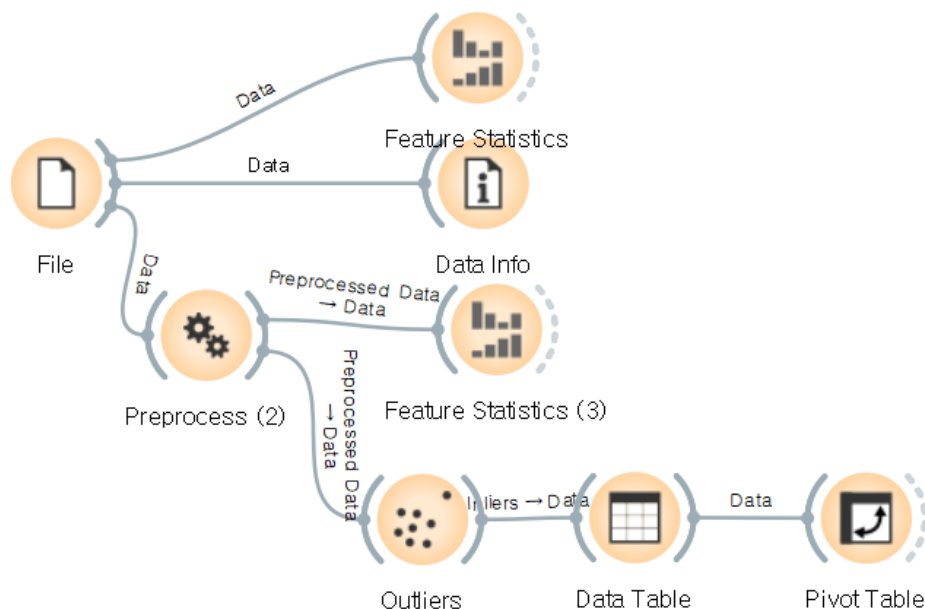
- 질문하기4

생존에 **성별**이 영향을 미칠까?

타이타닉 데이터 분석 프로세스

■ 질문하기4: 데이터 변형 및 시각화 (Pivot Table)

생존에 성별이 영향을 미칠까?

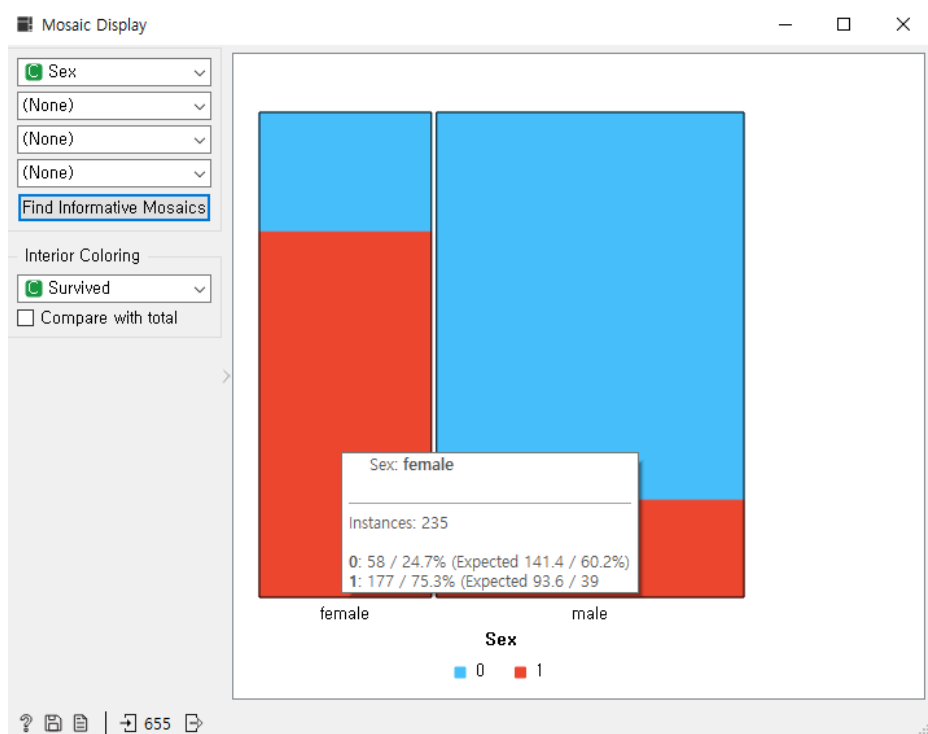


Pivot Table

Survived				
Sex	Count	0	1	Total
	female	58.0	177.0	235.0
	male	336.0	84.0	420.0
Total	394.0	261.0	655.0	

타이타닉 데이터 분석 프로세스

■ 질문하기4: 데이터 변형 및 시각화 (Mosaic Display)



생존에 성별이 영향을 미칠까?

타이타닉 데이터 분석 프로세스

- 질문하기4: 데이터 변형 및 시각화→ 대답하기

생존에 성별이 영향을 미칠까?

답변: 앞에서 우리는 생존자와 사망자의 비율이 4:6임을 확인

이번 질문에서 여성의 경우 생존자는 177명 사망자는 58명으로 사망자와 생존자 비율이 7.5:2.5로 생존 확률이 평균보다 월등히 높음. 반대로 남성의 경우 생존자와 사망자의 비율이 2:8로 전체 평균보다 사망비율이 더 높음.

타이타닉 데이터 분석 프로세스

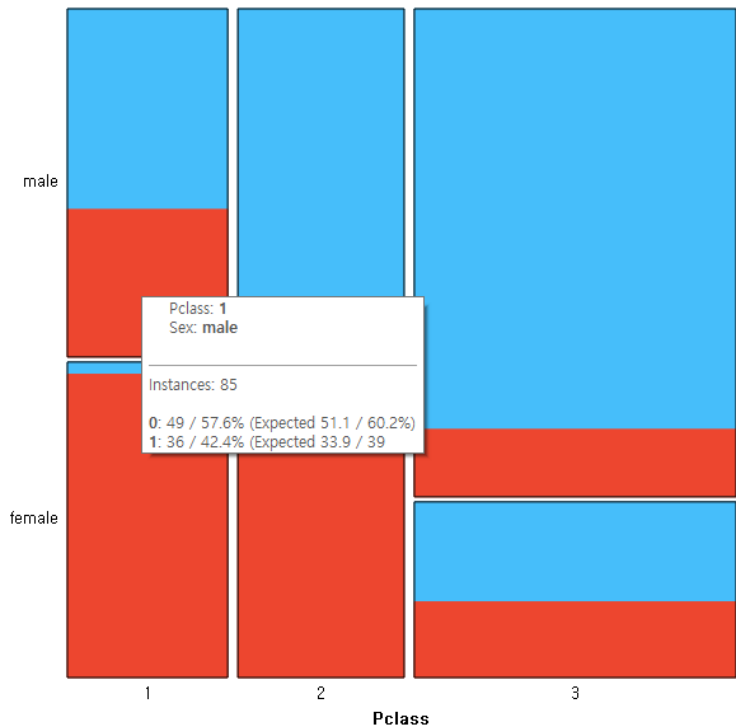
- 질문하기5

우리는 생존에 성별과 탑승 등급이 영향이 있다는 사실을 알 수 있었다.

그렇다면 생존에 성별과 탑승 등급이 복합적으로 영향을 끼치는 것일까?

타이타닉 데이터 분석 프로세스

■ 질문하기5: 데이터 변형 및 시각화 (Mosaic Display)



남자	1등급	2등급	3등급
사망	49(57.6%)	81(84.4%)	206(86.2%)
생존	36(42.4%)	15(15.6%)	33(13.8%)

남자	비율
사망	80%
생존	20%

여자	1등급	2등급	3등급
사망	3(3.9%)	6(8.3%)	49(57%)
생존	74(96.1%)	66(91.7%)	37(43%)

여자	비율
사망	25%
생존	75%

타이타닉 데이터 분석 프로세스

- 질문하기5: 데이터 변형 및 시각화→ 대답하기

생존에 성별과 탑승등급이 복합적으로 영향을 미칠까?

답변:

- 그래프와 표를 살펴봤을 때 여자의 경우 1,2등급에서는 대부분 높은 확률로 생존하였으며 3등급의 경우 사망자가 생존자보다 많이 발생
- 반면 남자의 경우 1등급에 탑승한 경우에도 사망자가 더 많이 발생하였으나 2,3등급 탑승자에 비해서는 높은 생존확률을 보임

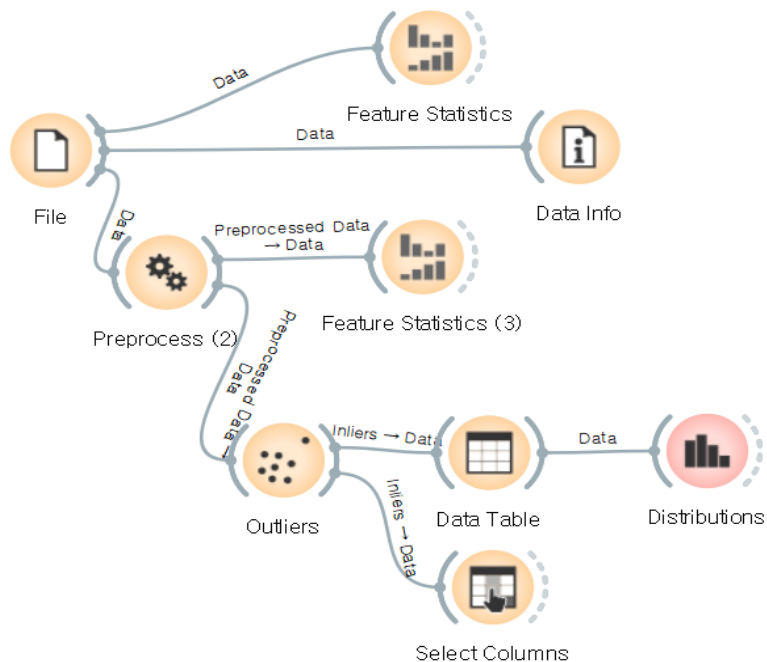
타이타닉 데이터 분석 프로세스

- 질문하기6

나이에 따라 생존확률에 차이가 있을까?

타이타닉 데이터 분석 프로세스

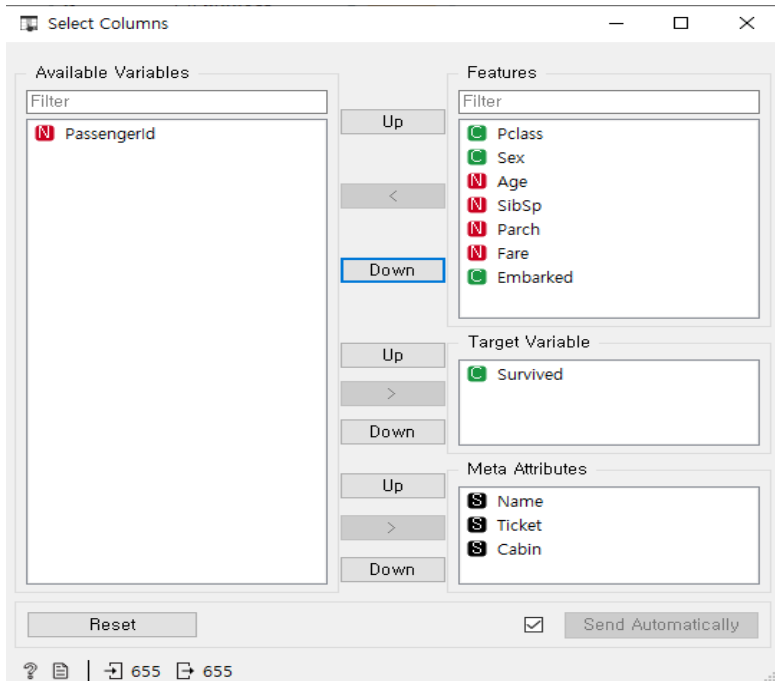
- 질문하기6: 데이터 변형 및 시각화



- Outliers 위젯에 Select Columns 위젯을 연결

타이타닉 데이터 분석 프로세스

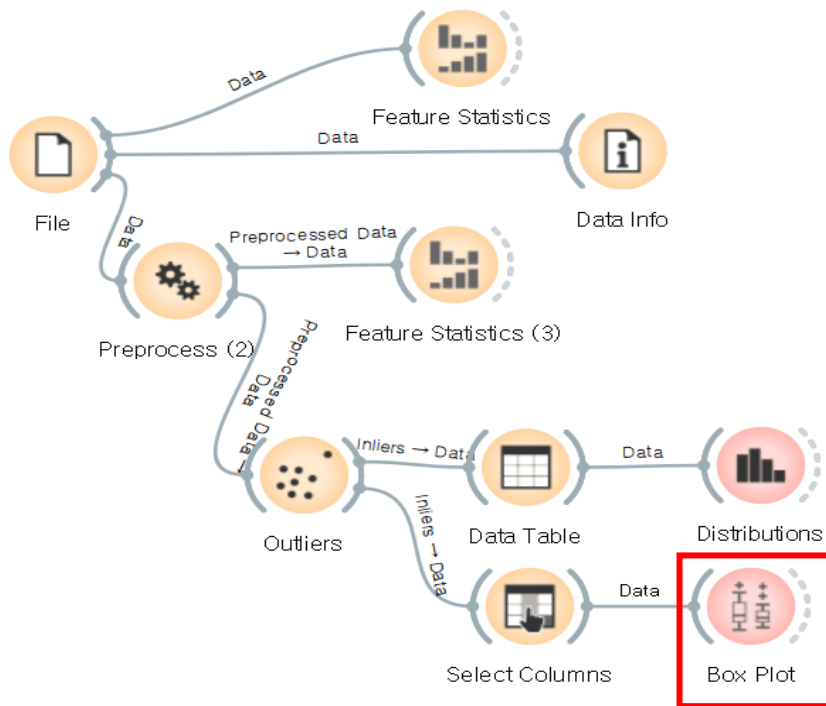
- 질문하기6: 데이터 변형 및 시각화



- Target 값은 Survived로 설정하고
- Passengerid는 변수 제외

타이타닉 데이터 분석 프로세스

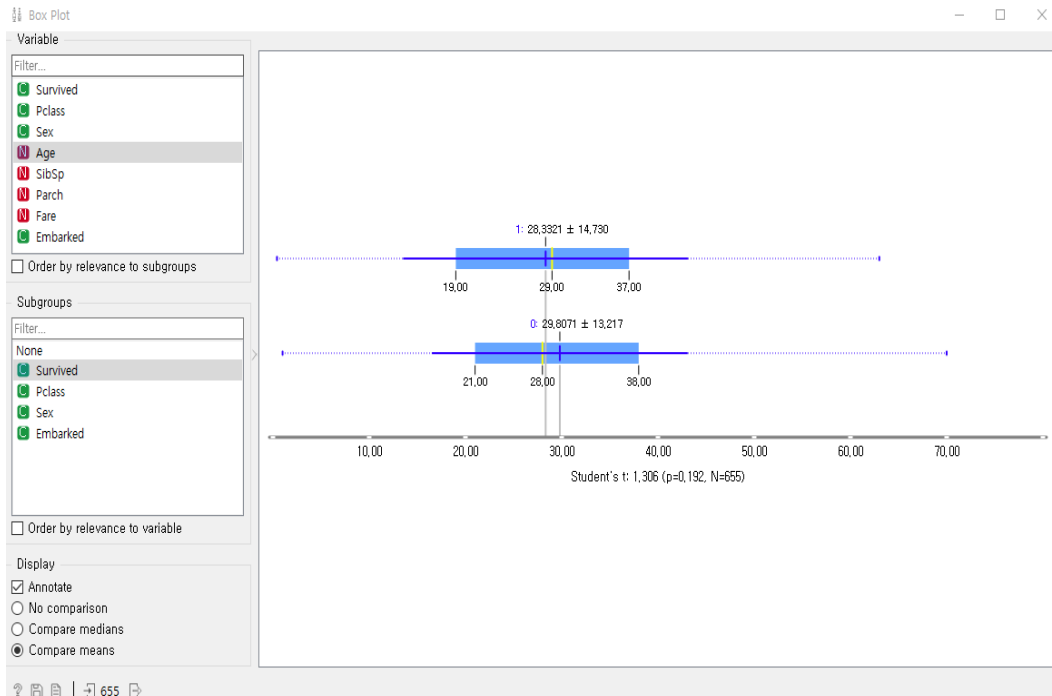
- 질문하기6: 데이터 변형 및 시각화 (Box Plot)



- Box Plot을 연결

타이타닉 데이터 분석 프로세스

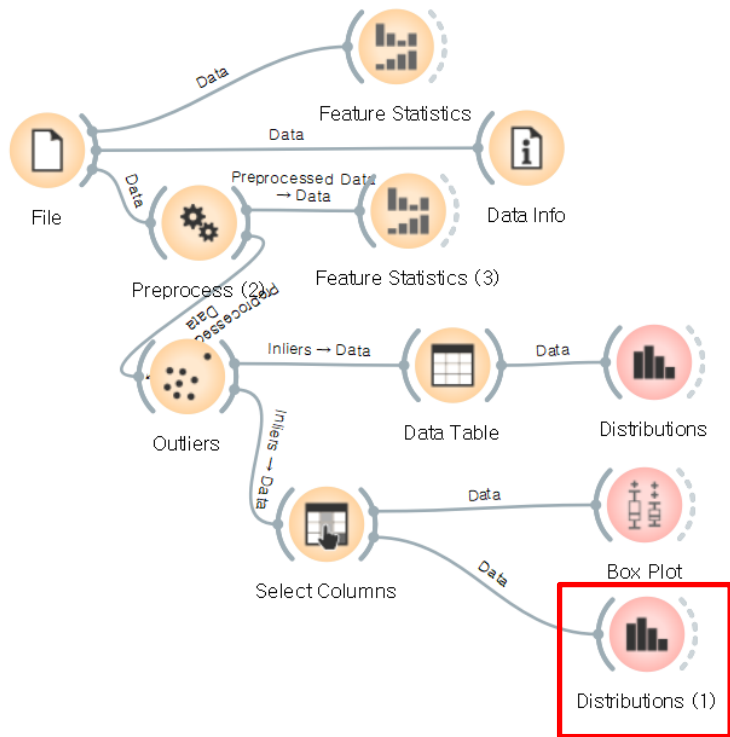
■ 질문하기6: 데이터 변형 및 시각화



- 생존자 박스가 사망자 박스보다 전체적으로 왼쪽으로 치우쳐져 있어 생존자의 나이가 조금 더 어린 것을 알 수 있음.

타이타닉 데이터 분석 프로세스

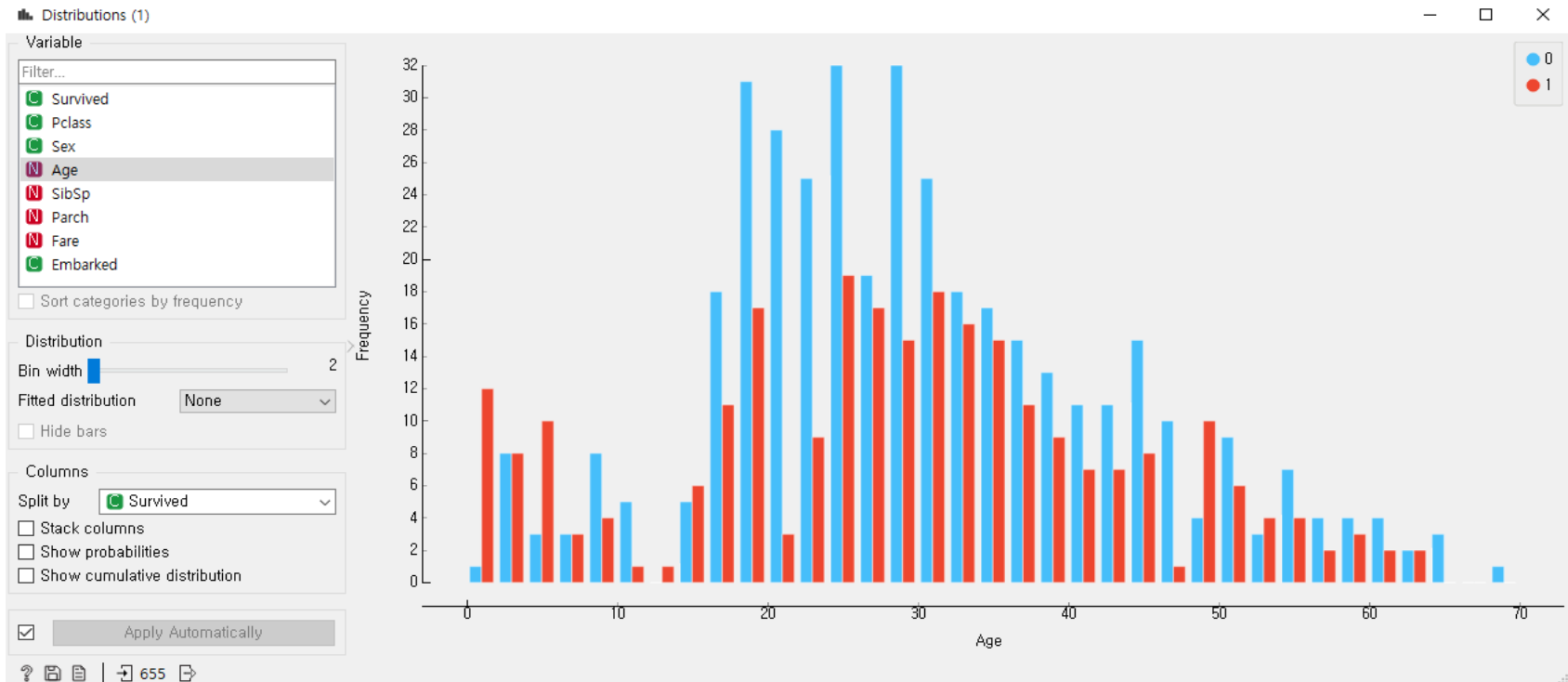
- 질문하기6: 데이터 변형 및 시각화(Distributions)



- Select Columns에 Distributions 위젯을 연결하여 그래프로 확인

타이타닉 데이터 분석 프로세스

■ 질문하기6: 데이터 변형 및 시각화



타이타닉 데이터 분석 프로세스

- 질문하기6: 데이터 변형 및 시각화→ 대답하기

나이에 따라 생존확률에 차이가 있을까?

답변:

- 단순히 생존자의 수를 비교해 보았을 때는 평균적으로 나이가 많은 사람보다 젊은 사람이 많았지만 나이별 사망자 수와 생존자의 비율을 봤을 때는 20~30대의 사망자 비율이 다른 연령대에 비하여 비교적 높음.
- 이러한 통계 결과는 젊은층에 구호활동과 같은 변수가 작용했을 가능성이 높다고 판단

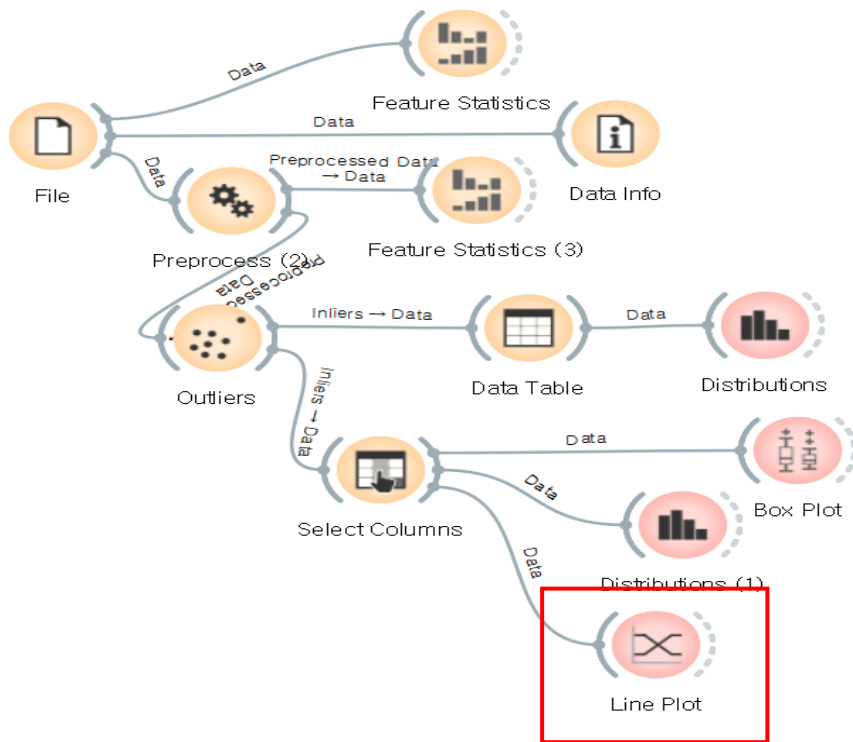
타이타닉 데이터 분석 프로세스

- 질문하기7

타이타닉호에서 생존을 위해 가장 효율적인 방법은 무엇일까?

타이타닉 데이터 분석 프로세스

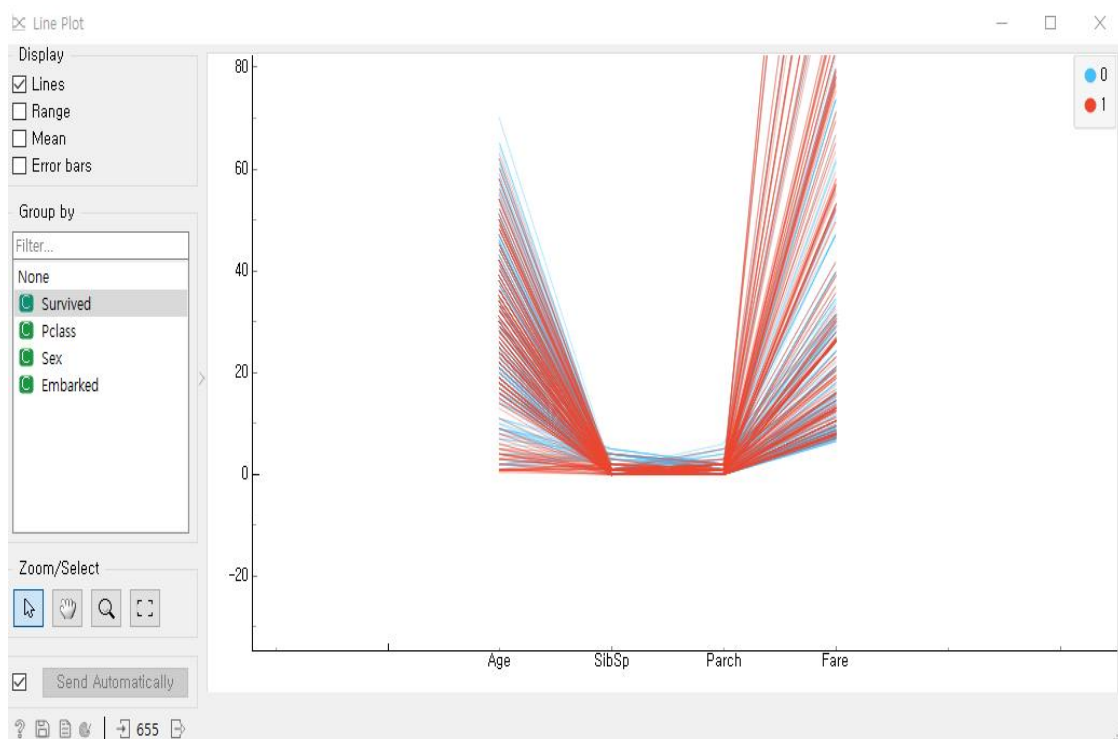
- 질문하기7: 데이터 변형 및 시각화 (Line plot)



- Select Columns에 Line Plot 위젯을 연결하여 그래프로 확인

타이타닉 데이터 분석 프로세스

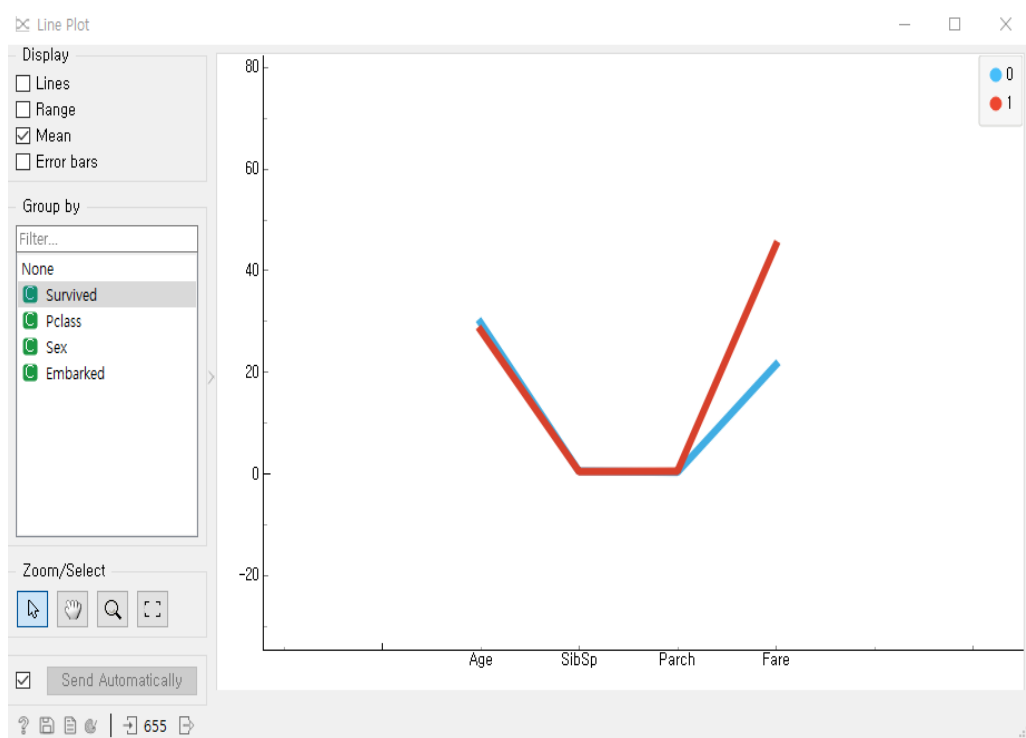
■ 질문하기7: 데이터 변형 및 시각화



- 초기 Display 설정에 Lines로 체크되어 있어 전체 인스턴스 값에 대한 그래프가 모두 표시 되어 매우 복잡함.

타이타닉 데이터 분석 프로세스

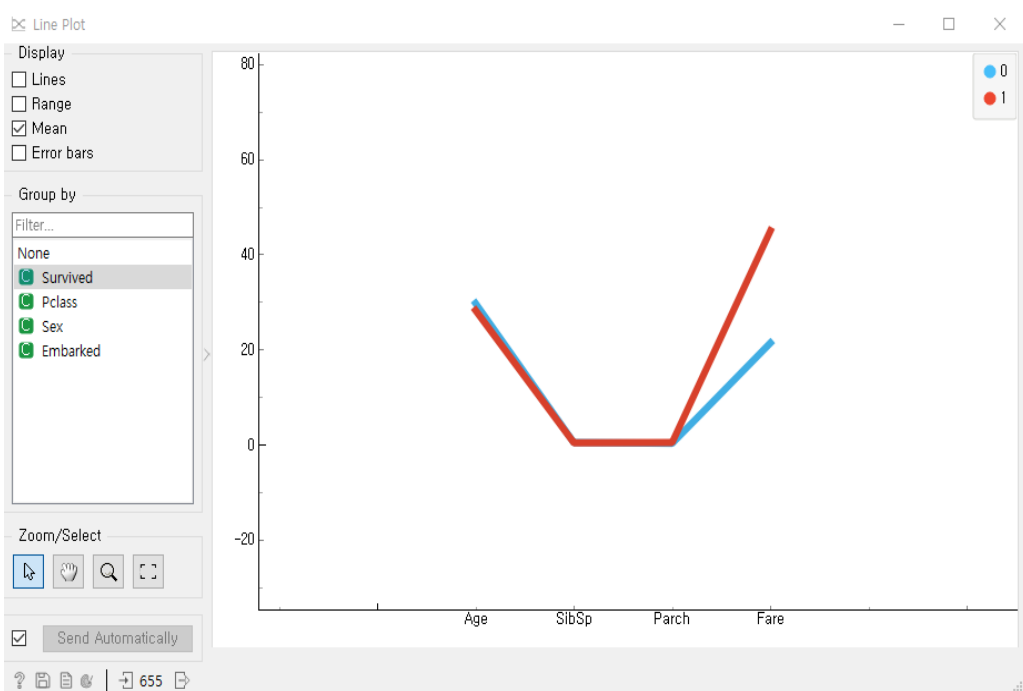
■ 질문하기7: 데이터 변형 및 시각화



- Lines를 해제하고 Mean(평균값)을 선택하면 그래프를 단순화 하여 볼 수 있음.

타이타닉 데이터 분석 프로세스

- 질문하기7: 데이터 변형 및 시각화



- 이 그래프를 해석해 보면 Fare 값이 높을 수록 낮은 값의 Fare보다 생존 확률이 높음

타이타닉 데이터 분석 프로세스

- 질문하기7: 데이터 변형 및 시각화→ 대답하기

타이타닉호에서 생존을 위해 가장 효율적인 방법은 무엇일까?

답변:

- Line Plot 결과 Fare(요금)이 높을 수록 생존 확률이 사망확률과 확실한 차이가 발생하는 것을 확인
- 이 결과는 앞에서 Pclass에 따른 생존 확률 결과와 일맥상통하여 타이타닉호에서 살아남을 수 있는 가장 확실한 방법은 다른 변수를 제외했을 때 높은 등급의 Class에 탑승하는 것임

타이타닉 데이터 분석 프로세스

- 결과

- 데이터 분석 및 시각화를 통하여 알게 된 결과는

1. 여자와 어린아이가 많이 생존했다.
2. 높은 등급(비싼 요금)의 탑승자일 수록 생존을 많이 했다.
3. 남자들의 탑승자 수가 많았으며 남자들의 사망률은 전체 평균에 비해 높다.
4. 신체건강한 젊은 남자들이 사망률이 높은 것은 구호활동과 같은 변수가 작용했을 확률이 높다.
5. 일반적인 생존확률을 예측할 때에는 Pclass와 연관성이 매우 높을 것이다.

데이터 내용 파악

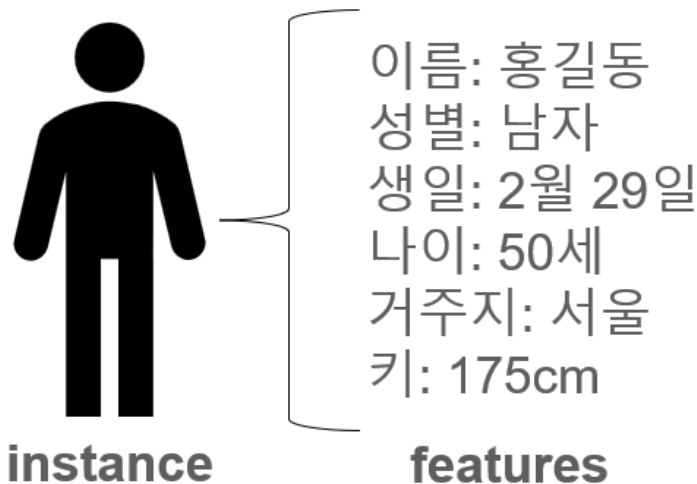
<사람을 예시로 하는 데이터 파악>

- **Instance:**

한 개체를 뜻함. 사람을 예시로 들면 사람
1이라고 표시

- **Features:**

사람 한 명을 나타내는 속성값. 예를 들면
성별, 키, 몸무게 등 instance를 설명하는
속성



변수의 유형 파악

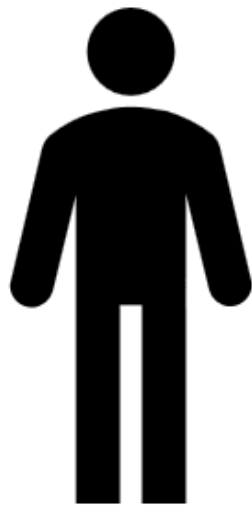
<사람을 예시로 하는 Features의 Type 구분>

- Type에는 Text, Categorical, Numeric, Datetime로 구분



이름: 홍길동(Text – 문자열)
성별: 남자(Categorical – 범주형)
생일: 2월 29일(Datetime – 날짜)
나이: 50세(Numeric – 숫자)
거주지: 서울(Categorical – 범주형)
키: 175cm(Numeric - 숫자)

변수의 유형 파악



범주형: 분류에 많이 사용.

숫자: 연산에 사용.

문자열: 단어 정보 추출.

날짜: 시간 순으로 정렬.

변수의 유형 파악

<사람을 예시로 하는 Features의 Role 구분하기>

- Role에는 Target, Features, Meta, Skip으로 구분
- Type별 Role은 데이터를 어떤 방식으로 활용하는지에 따라 다름

Target: 목표로 하는 값

Features: 목표에 영향을 주는 값

Meta: 목표에 영향을 주진 않으나 참고할 만한 값

Skip: 무시해도 되는 값

기본 통계 정보 관찰

- Distribution: 이산분포
- Center
 - Categorical feature의 경우 최빈값
 - Numeric feature의 경우 평균값
- Dispersion: 분산
- Minimum: 최솟값
- Maximum: 최댓값
- Missing: 결측값 수(빠진 값)

기본 통계 정보 관찰

- 기본 통계 정보 중 이상값들을 관찰할 필요가 있음
 - 이상값들을 관찰함으로써 오류 등을 발견할 수 있기 때문
- 이상값을 탐지하기 위한 위젯으로 'Outliers' 위젯을 활용 가능
- Outliers 위젯은 데이터에서 이상치를 검출하는 역할만 하므로 이를 시각화할 도구가 필요합니다.
- Outliers 위젯에서 이상치를 검출하고 이를 Scatter Plot을 활용해 시각화해보도록 합니다.
- Outliers 위젯은 왼쪽 'Data' 메뉴에서 찾아 클릭하거나 드래그&드랍합니다.

질문 있나요?

hsryu13@hongik.ac.kr

