데이터 과학 기초

02

탐색적 데이터 분석

경북대학교 배준현 교수
(joonion@knu.ac.kr)

02. 탐색적 데이터 분석

- **데이터에 대한 두 가지 접근법:** CDA *.vs.* EDA
  - **확증적 데이터 분석:** CDA, *confirmatory* data analysis
    - 가설을 수립하고 데이터를 통해 통계적 유의성을 검정하는 전통적 분석 기법
    - *Ronald Fisher*: 가설검정, 신뢰구간, 유의수준, 유의확률(*p-value*)
  - **탐색적 데이터 분석:** EDA, *exploratory* data analysis
    - 정해진 가설과 모형없이 데이터의 구조와 특성을 통해 통찰을 얻는 분석 기법
    - *John Tukey*: EDA는 우리가 존재한다고 믿는 것들은 물론이고, 존재하지 않는다고 믿는 것들을 발견하려는 태도, 유연성, 그리고 자발성이다.

02. 탐색적 데이터 분석

- **탐색적 데이터 분석**: EDA, Exploratory Data Analysis
  - 데이터에 대한 기본적인 이해를 하기 위한 탐색과 분석 과정
    - 데이터의 기본적인 유형, 구조, 분포, 관계 등을 파악
  - 기술 통계: Descriptive Statistics
    - 데이터의 정리, 요약, 해석, 표현을 통해 자료의 특성을 규명
    - 도수분포표, 평균, 분산, 표준편차, 상관계수
  - 데이터 시각화: Data Visualization
    - 시각적 도구를 이용한 데이터의 이해
    - 산점도, 히스토그램, 선/막대 그래프, 상자 플롯, 파이 차트, 등등.

02. 탐색적 데이터 분석

- **데이터의 유형: Data Types**
  - **숫자형**(연속형, **양적 자료**): Numeric (Continuous, Quantitative)
    - 수치로 나타낼 수 있는 변수. 산술/논리 연산을 적용할 수 있다.
    - 주요 분석 대상: 평균, 분산, 표준편차, 분포 등.
  - **범주형**(명목형, **질적 자료**): Categorical (Nominal, Qualitative)
    - 기호나 이름으로 구분할 수 있는 변수. 산술/논리 연산을 적용할 수 없다.
    - 주요 분석 대상: 빈도, 히스토그램(histogram)

02. 탐색적 데이터 분석

- **변수**: Variables
  - 통계학에서 말하는 변수: 연구, 조사, 관찰하고 싶은 대상의 특징
    - 예) 키, 몸무게, 혈액형, 매출액, 온도, 습도, 미세먼지 농도, 등
  - 단일변수 데이터: Univariate Data
    - 일변량 자료: 하나의 변수로만 구성된 데이터 (벡터)
  - 다중변수 데이터: Multivariate Data
    - 다변량 자료: 두 개 이상의 변수로 구성된 자료 (행렬, 데이터 프레임)

02. 탐색적 데이터 분석

- **변수의 종류:**
  - 목적 변수(종속 변수): Target(Dependent) Variable
    - 어떤 분석을 통해 추정하거나 예측하고자 하는 목적이 되는 데이터
    - 독립 변수의 값의 변화에 따라 영향을 받는 종속 변수
  - 특징 변수(독립 변수): Feature(Independent) Variable
    - 목적 변수의 추정이나 예측을 위해 사용하는 데이터의 특성
    - 종속 변수의 값에 독립적으로 영향을 주는 변수

02. 탐색적 데이터 분석

- **데이터셋:** *dataset*
  - 데이터의 집합: 주로 2차원 테이블(행렬) 형태로 정리된 데이터
    - **변수**: 열(*column*), **관측값**: 행(*row*)
    - 데이터 프레임: R/Pandas에서 데이터셋의 유형
  - 데이터 과학 분야에서 유명한 데이터셋 4개를 탐색해보자:
    - **붓꽃** 데이터셋: *iris* dataset
    - **보스턴 집값** 데이터셋: *housing* dataset
    - **펭귄** 데이터셋: *penguins* dataset

02. 탐색적 데이터 분석

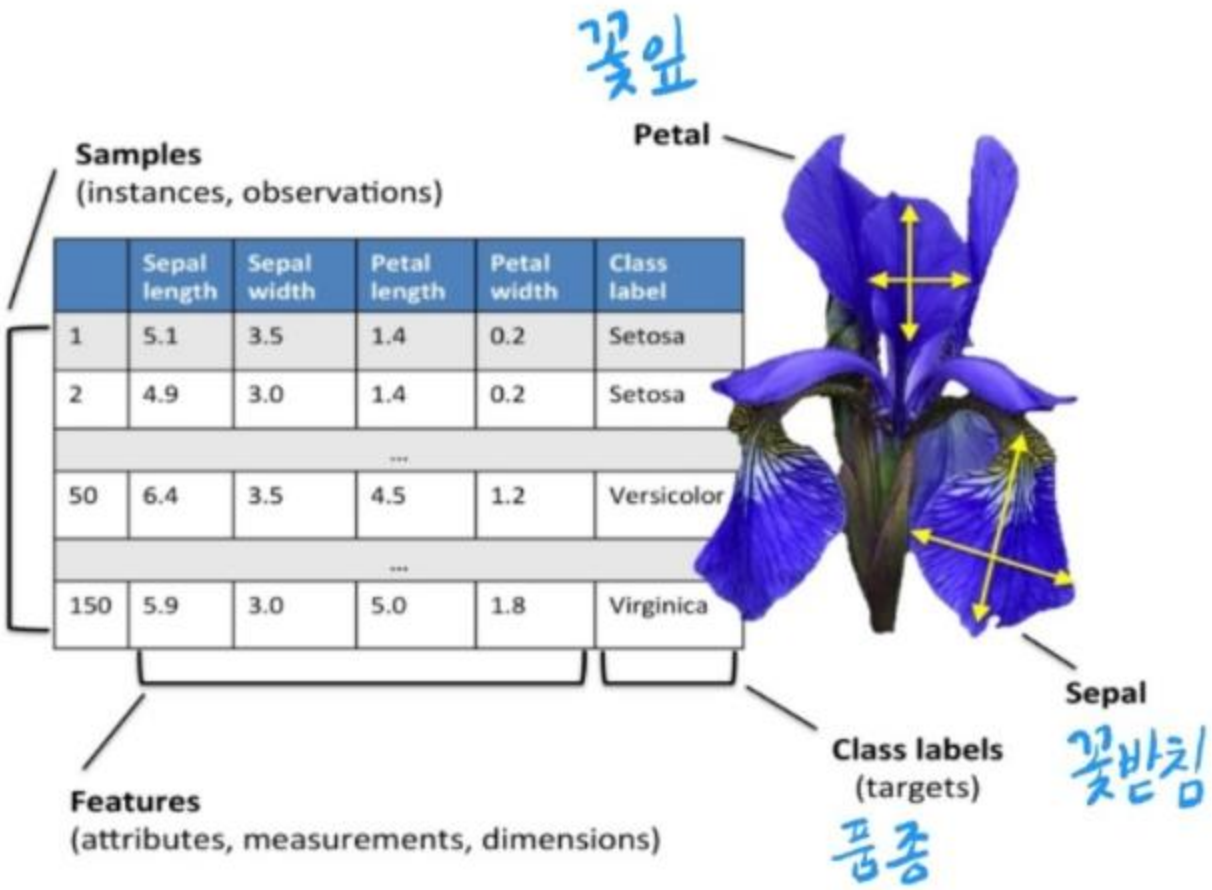
- **붓꽃 데이터셋:** *IRIS* dataset
  - **로널드 피셔의 연구:**
    - **데이터로만 붓꽃(iris)의 품종을 구분할 수 있을까?**

setosa

versicolor

virginica

02. 탐색적 데이터 분석

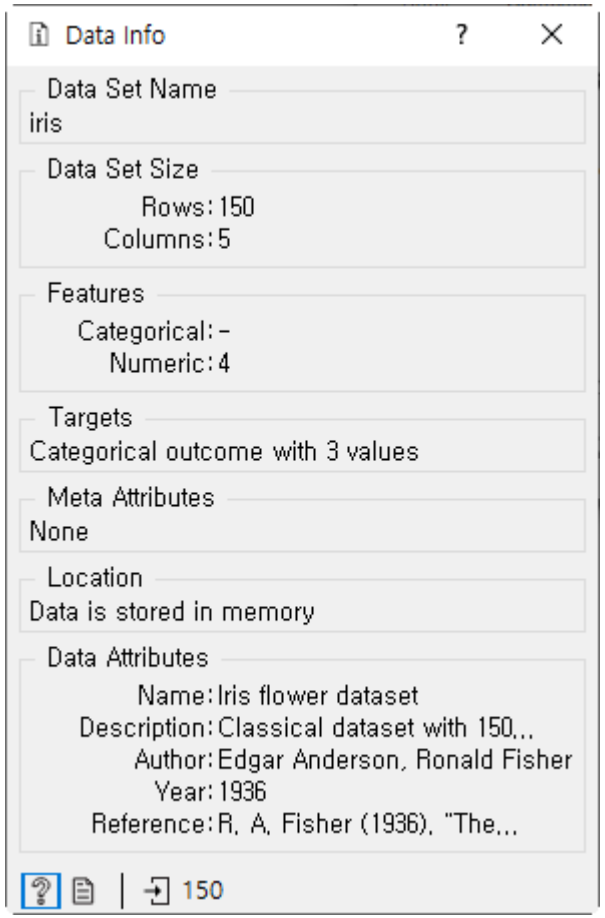


	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

이미지 출처: 브런치, 야사와 만화로 배우는 인공지능
https://brunch.co.kr/@hvnpoet/82

02. 탐색적 데이터 분석

- IRIS dataset: **다변량 자료**
  - Features: 4**개의 숫자형 독립변수**
    - sepal length: **꽃받침 길이**
    - sepal width: **꽃받침 너비**
    - petal length: **꽃잎의 길이**
    - petal width: **꽃잎의 너비**
  - Targets: **범주형 종속변수**
    - iris: **붓꽃의 품종(3종)**



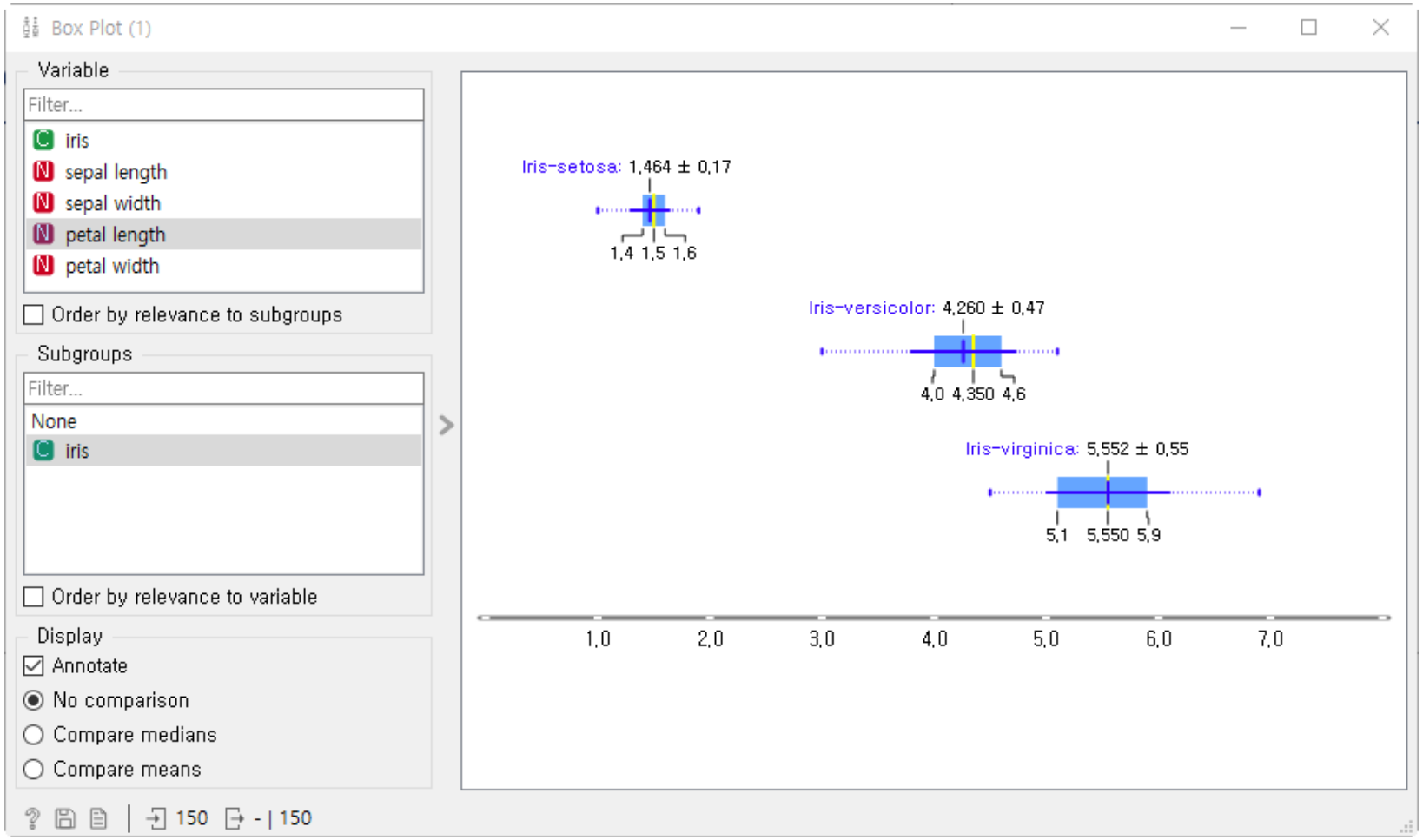
02. 탐색적 데이터 분석

■ 연속형 자료의 탐색과 분석:

- **평균**: 전체 변량의 총합을 변량의 개수로 나눈 값
  - $(\text{평균}) = \frac{(\text{변량})\text{의 총합}}{(\text{변량})\text{의 개수}}$
- **중앙값**: 자료의 변량을 순서대로 나열할 때, 중앙에 위치하는 값
  - 매우 크거나 작은 값이 있을 경우에는 평균보다 더 자료의 특성을 더 잘 반영.
- **분산**: 편차를 제곱한 값의 평균, **표준편차**: 분산의 양의 제곱근
  - $(\text{분산}) = \frac{(\text{편차})^2\text{의 총합}}{(\text{변량})\text{의 개수}}$, $(\text{표준편차}) = \sqrt{(\text{분산})}$
  - 분산(표준편차)의 값이 클수록, 평균을 중심으로 흩어져 있는 정도가 크다.
  - 분산(표준편차)의 값이 작을수록, 평균을 중심으로 흩어져 있는 정도가 작다.

02. 탐색적 데이터 분석

- Orange: Box Plot

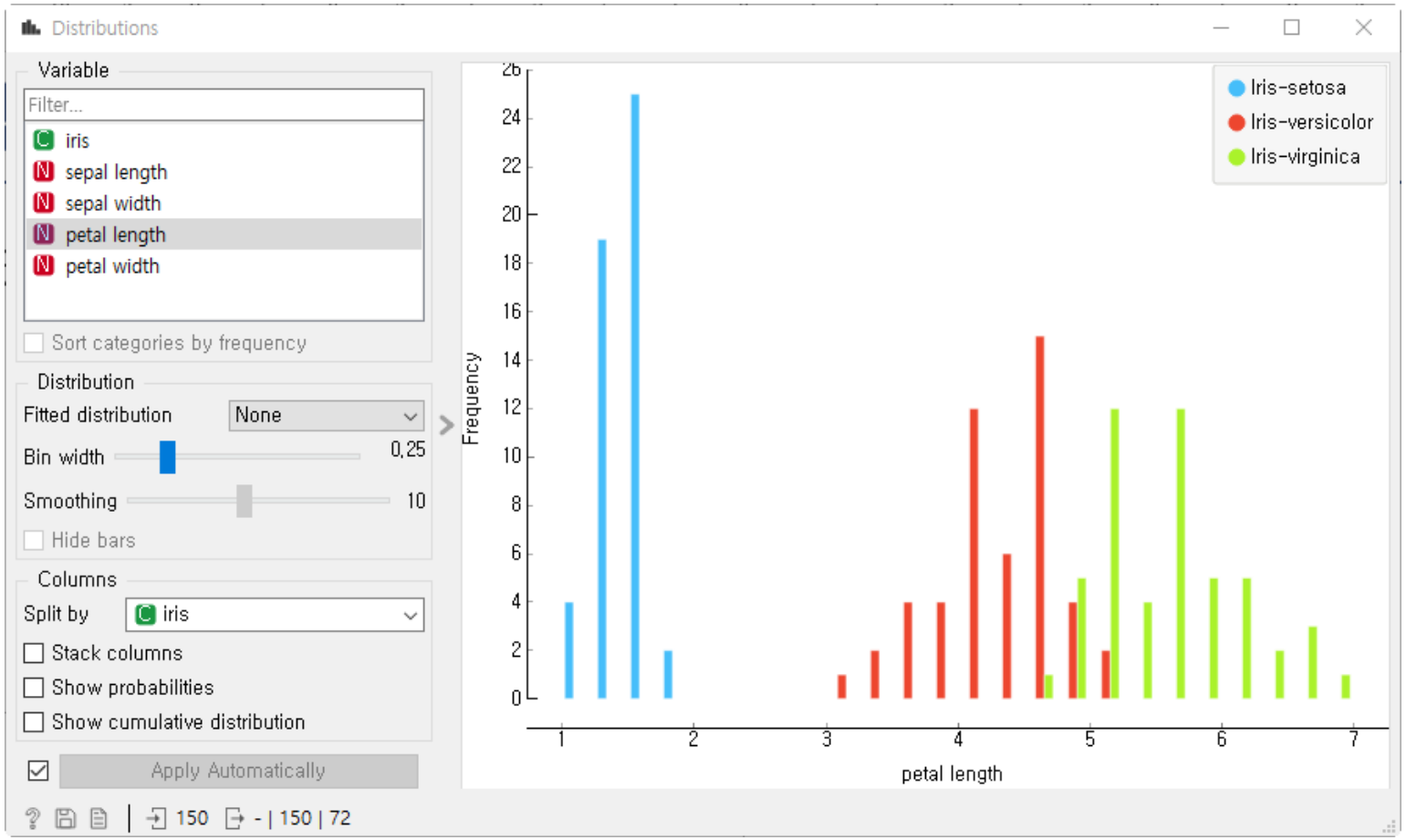


02. 탐색적 데이터 분석

- **범주형 자료의 탐색과 분석:**
  - 평균, 분산, 표준편차 등의 통계적 특성을 가지지 않음
    - 각 변수의 빈도(frequency)를 막대 그래프 등으로 파악
  - 도수분포표: 데이터를 정리하여 도수의 분포를 표로 나타낸 것
    - 도수: 각 구간에 속하는 자료의 수
  - 히스토그램(histogram): 도수분포표를 그래프로 나타낸 것

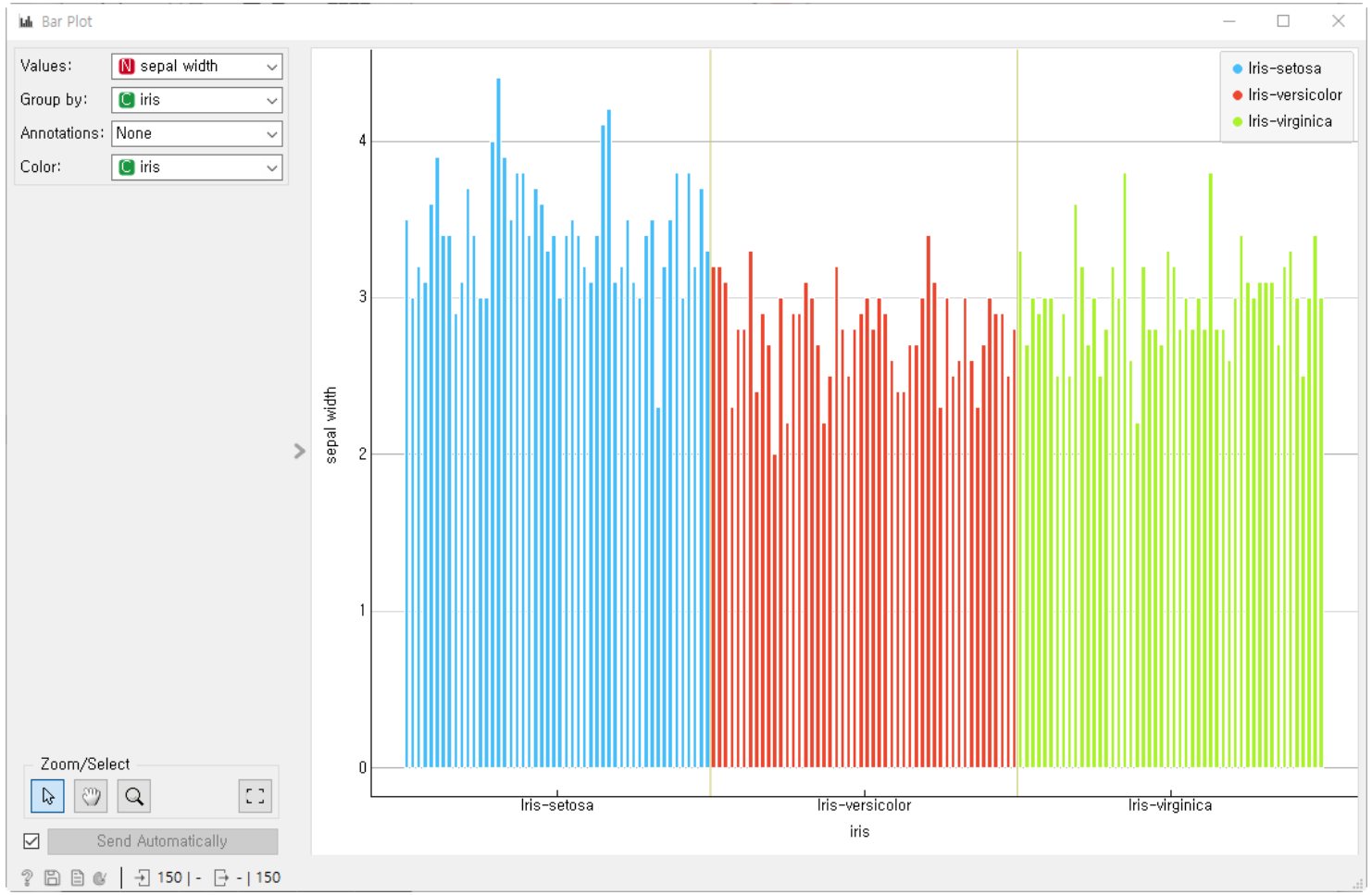
02. 탐색적 데이터 분석

- Orange: Distributions



02. 탐색적 데이터 분석

- Orange: Bar Plot

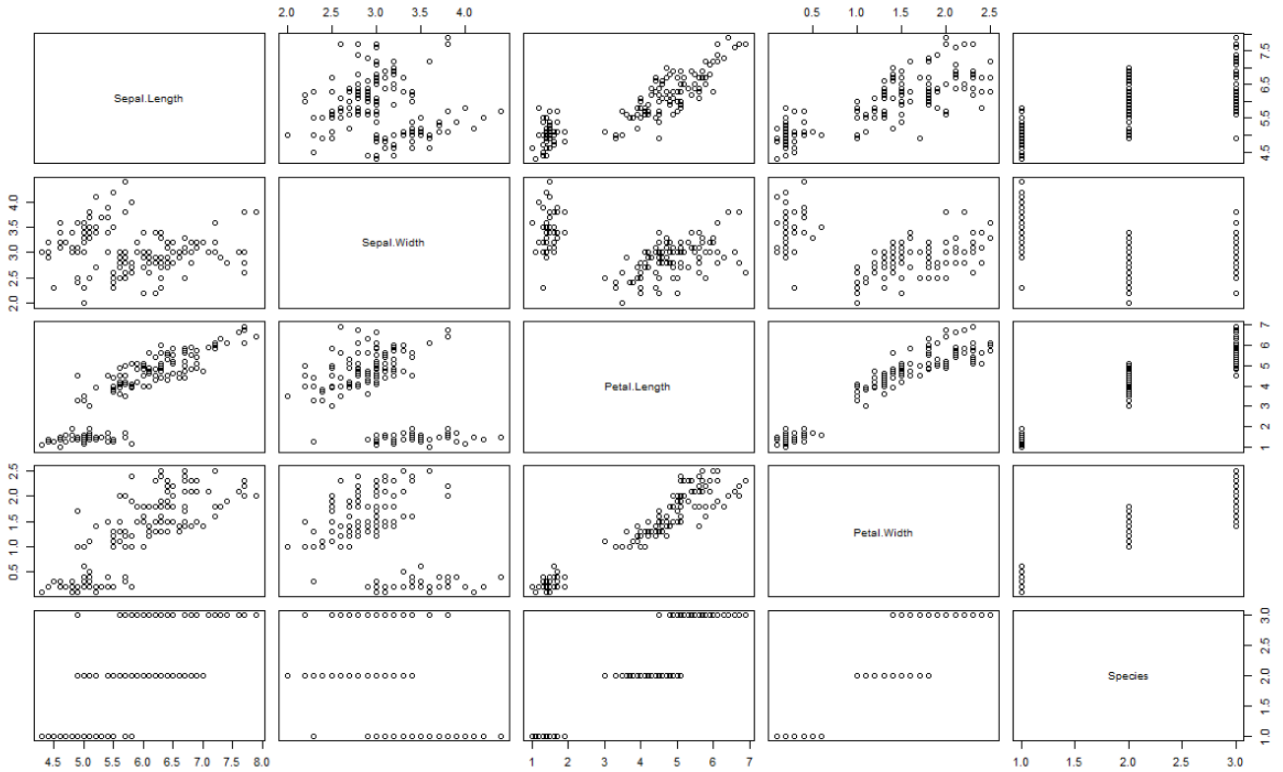


02. 탐색적 데이터 분석

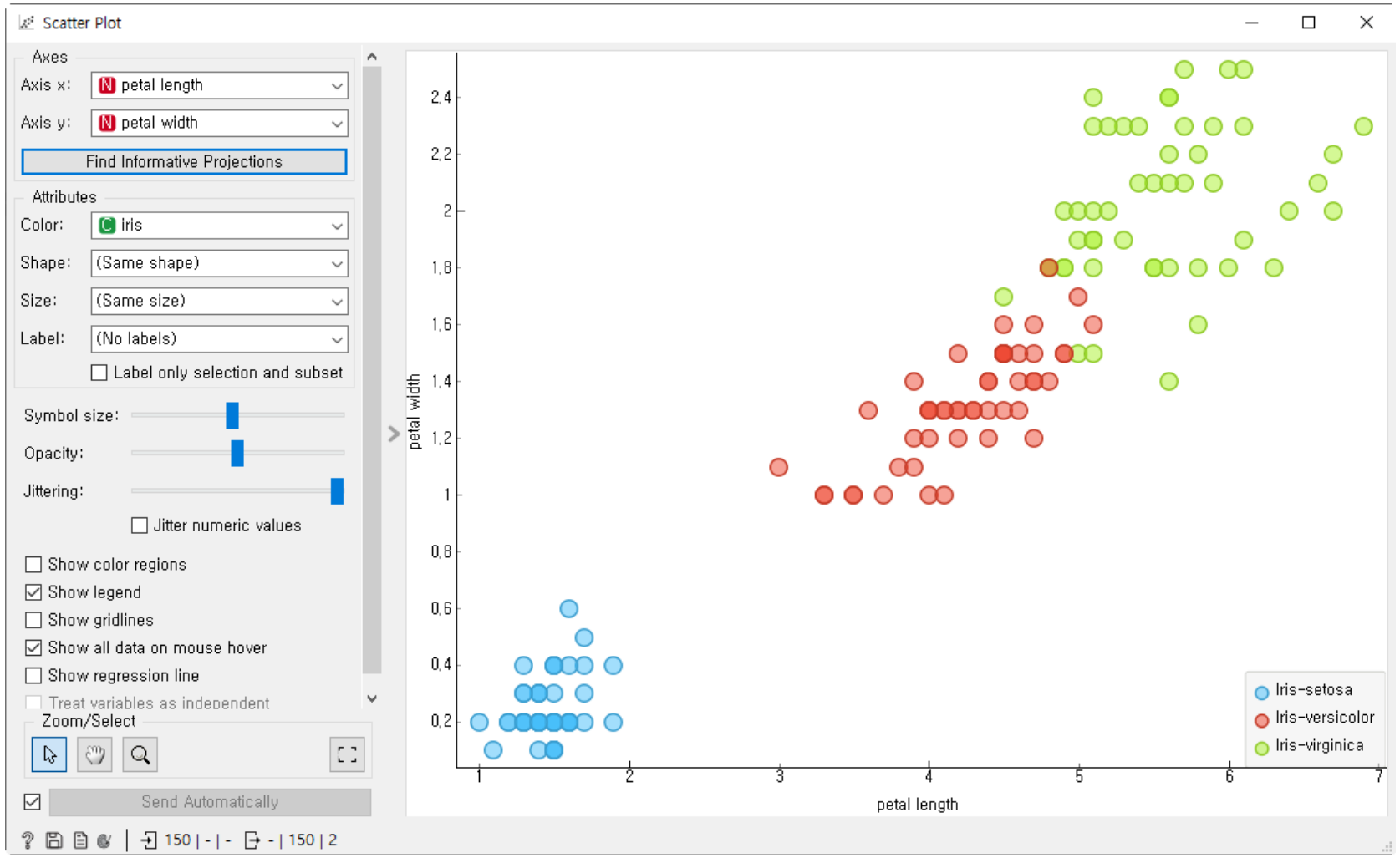
- **데이터 시각화**: Data Visualization
  - 숫자형/범주형 데이터를 그래프나 그림 등의 시각적 형태로 표현하는 것
  - 탐색적 데이터 분석 과정에서 데이터를 파악하는 중요한 기술 중의 하나
  - 주요 시각화 방법:
    - 선 그래프, 막대 그래프, 히스토그램
    - 박스 플롯: Box Plot
    - 산점도: Scatter Plot
    - 모자이크 플롯: Mosaic Display
    - 히트맵: Heat Map

02. 탐색적 데이터 분석

- **산점도**: Scatter Plot
  - 두 개의 변수로 구성된 자료의 분포를 알아보는 그래프.
  - 관측값들의 분포를 통해 두 변수 사이의 관계를 파악할 수 있음.



- Orange: Scatter Plot

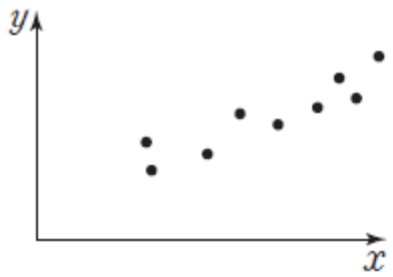


상관 분석: Correlation Analysis

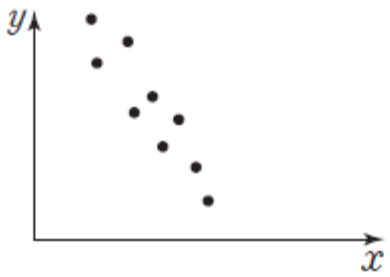
- 두 변수 간에 어느 정도의 선형적 관계가 있는지를 파악하는 방법.
- 상관 계수: Correlation Coefficient
  - 상관 관계의 정도를 나타내는 지수
- 피어슨 상관 계수: *Pearson's Correlation Coefficient*
  - 두 개의 데이터 X, Y에 대해서, $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$,
  - X와 Y가 함께 변하는 정도 / X와 Y가 각각 변하는 정도
  - $r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$

02. 탐색적 데이터 분석

- **피어슨 상관계수의 해석:**
  - $0 < r \leq 1$: 양의 상관관계가 있다. x가 증가하면 y도 증가한다.
  - $-1 \leq r < 0$: 음의 상관관계가 있다. x가 증가하면 y는 감소한다.
  - r의 절대값이 클수록 두 변수 x, y의 선형적인 상관성이 높다.



(a) 상관계수(r)가 1에 가까움.

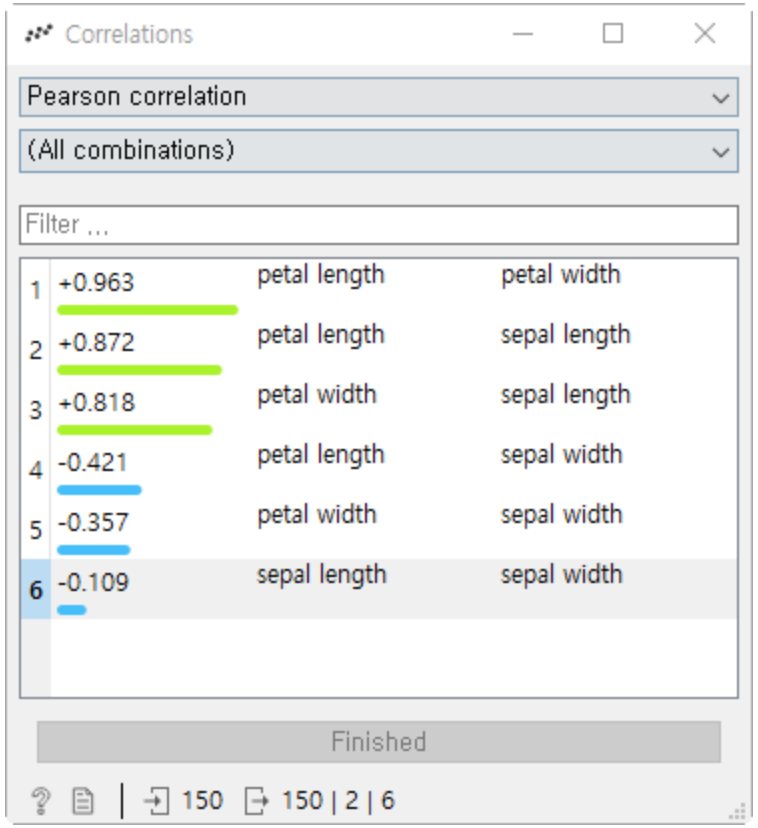


(b) 상관계수(r)가 −1에 가까움.

(c) 상관계수(r)가 0에 가깝고 뚜렷한 상관관계 없음.

(d) 상관계수(r)가 0에 가까우나 비선형적 상관관계

- Orange: Correlations



Correlation does not imply *causation*!

02. 탐색적 데이터 분석

Science › tumour

Why going to university increases risk of getting a brain tumour

Highly educated people are more likely to suffer from brain tumours than those who do not progress as far in their education

SHARE COMMENTS

By Andrew Gregory
23:30, 20 JUN 2016

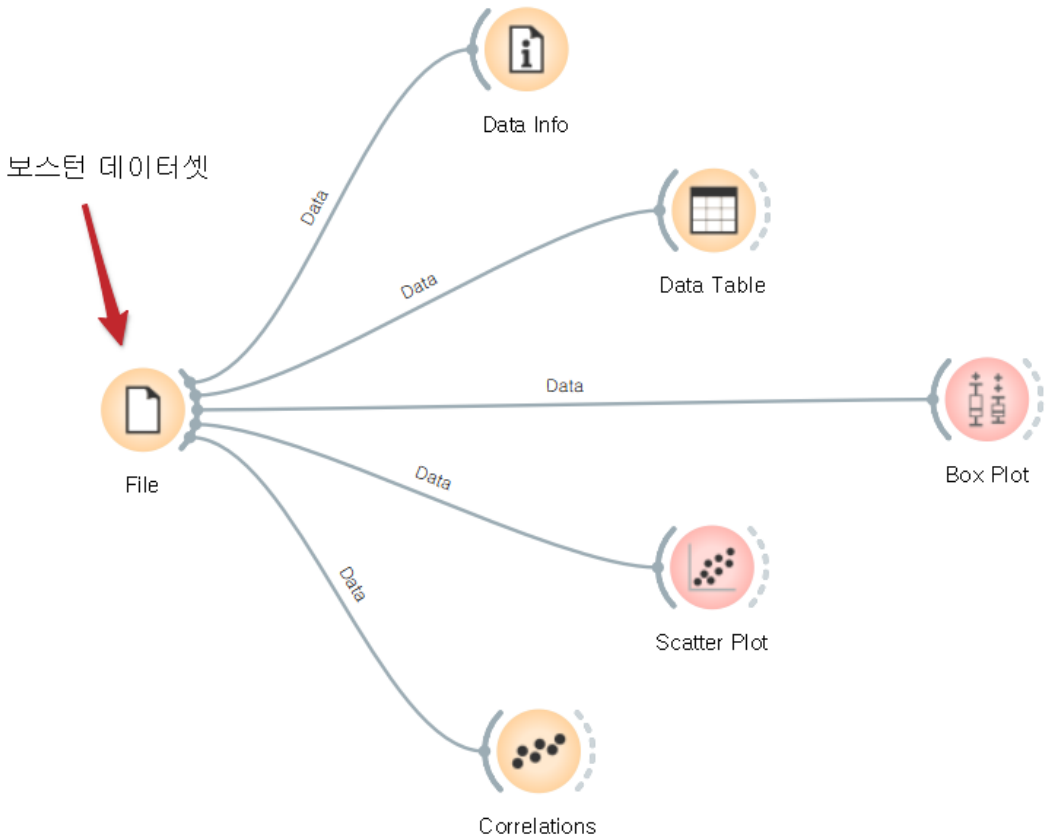
SCIENCE

02. 탐색적 데이터 분석

- **보스턴 하우징 데이터셋**
  - 1978년 미국 보스턴 지역의 주택 가격에 관련된 데이터셋
    - 총 14개의 변수: 13개의 특징 변수, 1개의 목적 변수
    - 다변량 자료: 14개의 변수는 모두 숫자형(numeric)
    - 총 506개의 관측값

02. 탐색적 데이터 분석

- **탐색적 데이터 분석**: housing.tab



Data Info

Data Set Name
housing

Data Set Size
Rows: 506
Columns: 14

Features
Categorical: -
Numeric: 13

Targets
Numeric target variable

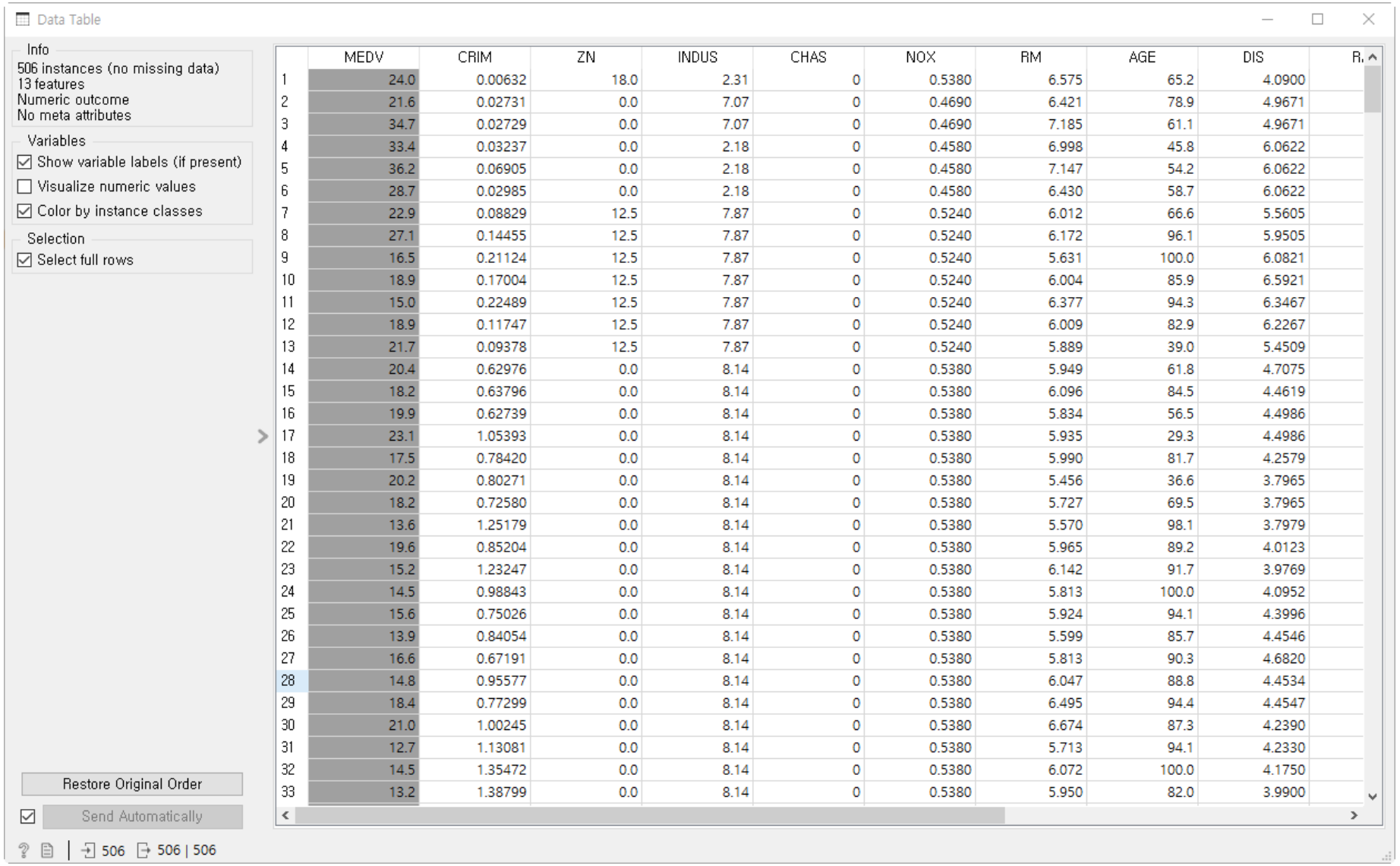
Meta Attributes
None

Location
Data is stored in memory

Data Attributes
Name: Housing dataset
Description: Data collected by the U.S...
Author: U.S Census Service
Year: 1978
Reference: Harrison, D. and Rubinfeld,...

506

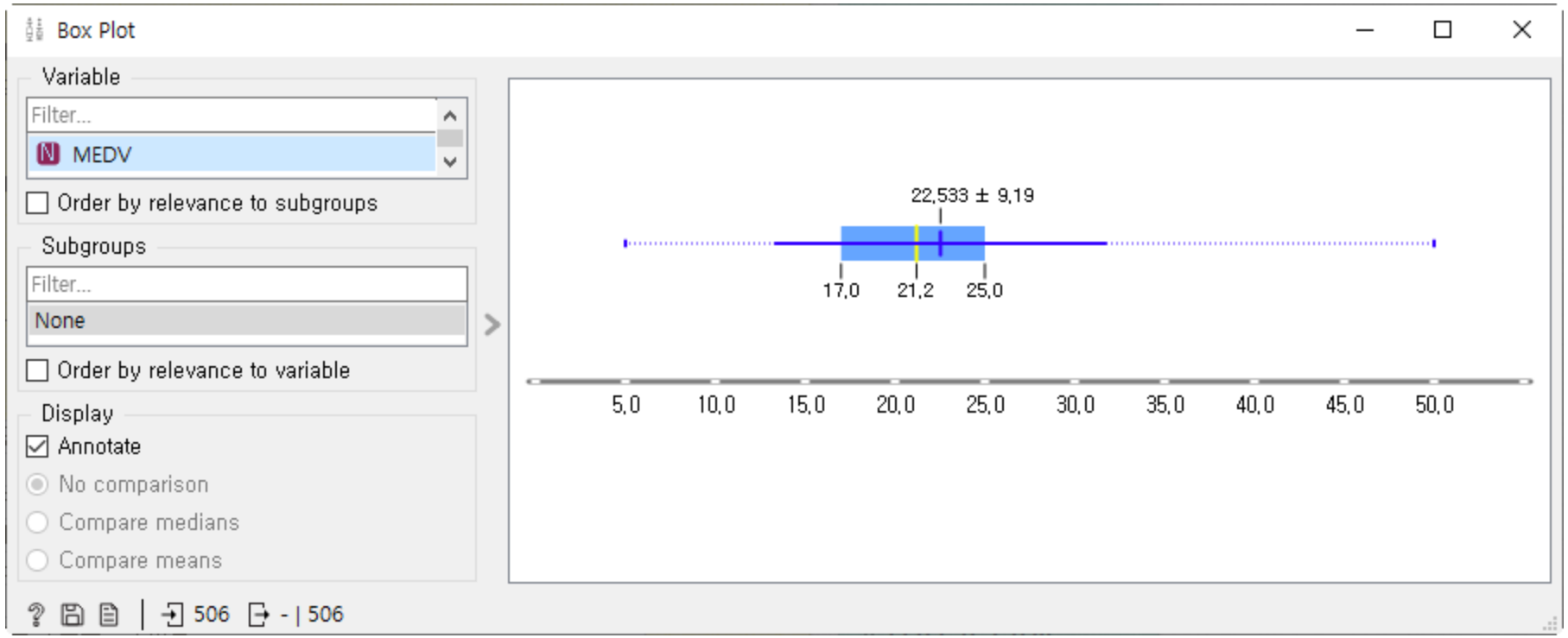
02. 탐색적 데이터 분석



	MEDV	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	R.
1	24.0	0.00632	18.0	2.31	0	0.5380	6.575	65.2	4.0900	
2	21.6	0.02731	0.0	7.07	0	0.4690	6.421	78.9	4.9671	
3	34.7	0.02729	0.0	7.07	0	0.4690	7.185	61.1	4.9671	
4	33.4	0.03237	0.0	2.18	0	0.4580	6.998	45.8	6.0622	
5	36.2	0.06905	0.0	2.18	0	0.4580	7.147	54.2	6.0622	
6	28.7	0.02985	0.0	2.18	0	0.4580	6.430	58.7	6.0622	
7	22.9	0.08829	12.5	7.87	0	0.5240	6.012	66.6	5.5605	
8	27.1	0.14455	12.5	7.87	0	0.5240	6.172	96.1	5.9505	
9	16.5	0.21124	12.5	7.87	0	0.5240	5.631	100.0	6.0821	
10	18.9	0.17004	12.5	7.87	0	0.5240	6.004	85.9	6.5921	
11	15.0	0.22489	12.5	7.87	0	0.5240	6.377	94.3	6.3467	
12	18.9	0.11747	12.5	7.87	0	0.5240	6.009	82.9	6.2267	
13	21.7	0.09378	12.5	7.87	0	0.5240	5.889	39.0	5.4509	
14	20.4	0.62976	0.0	8.14	0	0.5380	5.949	61.8	4.7075	
15	18.2	0.63796	0.0	8.14	0	0.5380	6.096	84.5	4.4619	
16	19.9	0.62739	0.0	8.14	0	0.5380	5.834	56.5	4.4986	
17	23.1	1.05393	0.0	8.14	0	0.5380	5.935	29.3	4.4986	
18	17.5	0.78420	0.0	8.14	0	0.5380	5.990	81.7	4.2579	
19	20.2	0.80271	0.0	8.14	0	0.5380	5.456	36.6	3.7965	
20	18.2	0.72580	0.0	8.14	0	0.5380	5.727	69.5	3.7965	
21	13.6	1.25179	0.0	8.14	0	0.5380	5.570	98.1	3.7979	
22	19.6	0.85204	0.0	8.14	0	0.5380	5.965	89.2	4.0123	
23	15.2	1.23247	0.0	8.14	0	0.5380	6.142	91.7	3.9769	
24	14.5	0.98843	0.0	8.14	0	0.5380	5.813	100.0	4.0952	
25	15.6	0.75026	0.0	8.14	0	0.5380	5.924	94.1	4.3996	
26	13.9	0.84054	0.0	8.14	0	0.5380	5.599	85.7	4.4546	
27	16.6	0.67191	0.0	8.14	0	0.5380	5.813	90.3	4.6820	
28	14.8	0.95577	0.0	8.14	0	0.5380	6.047	88.8	4.4534	
29	18.4	0.77299	0.0	8.14	0	0.5380	6.495	94.4	4.4547	
30	21.0	1.00245	0.0	8.14	0	0.5380	6.674	87.3	4.2390	
31	12.7	1.13081	0.0	8.14	0	0.5380	5.713	94.1	4.2330	
32	14.5	1.35472	0.0	8.14	0	0.5380	6.072	100.0	4.1750	
33	13.2	1.38799	0.0	8.14	0	0.5380	5.950	82.0	3.9900	

02. 탐색적 데이터 분석

- **목적 변수**: Target Variable
  - MEDV: **본인 소유 주택 가격의 중앙값**(단위: $1,000)

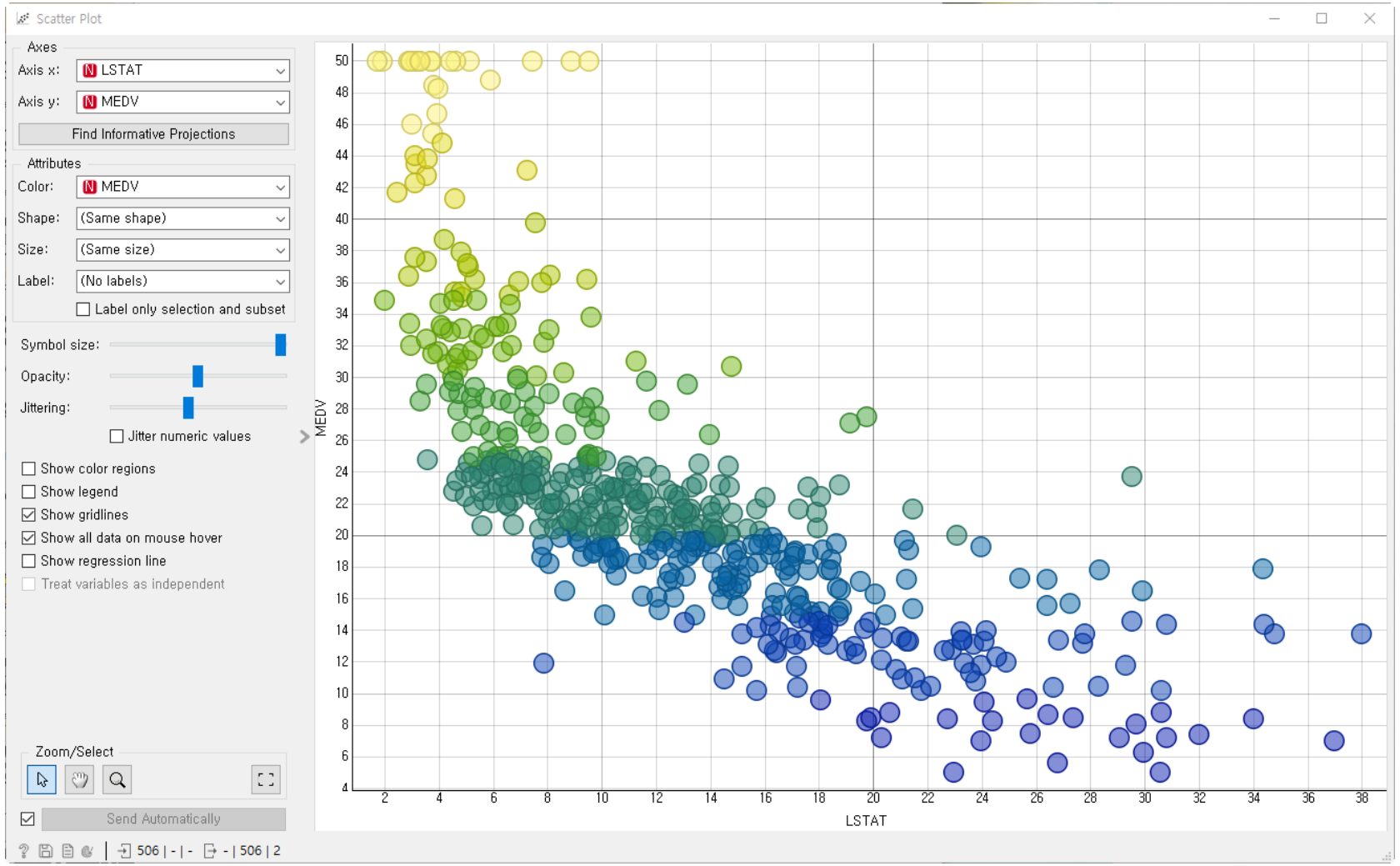


02. 탐색적 데이터 분석

- **특징 변수**: Feature Variables
  - CRIM: 타운별 1인당 범죄율
  - ZN: 25,000 평방피트 초과 거주지 비율
  - INDUS: 비소매 상업지역이 점유하는 토지 비율
  - CHAS: 찰스 강에 인접 여부
  - NOX: 10ppm당 농축 일산화질소
  - RM: 주택 1가구당 평균 방의 수
  - AGE: 1940년 이전 건축 주택 비율
  - DIS: 5개 직업센터와의 거리
  - RAD: 방사형 도로까지의 접근성 지수
  - TAX: 10,000달러 당 재산세율
  - PTRATIO: 타운별 학생/교사 비율
  - B: 타운별 흑인의 비율
  - LSTAT: 모집단의 소득 하위계층 비율

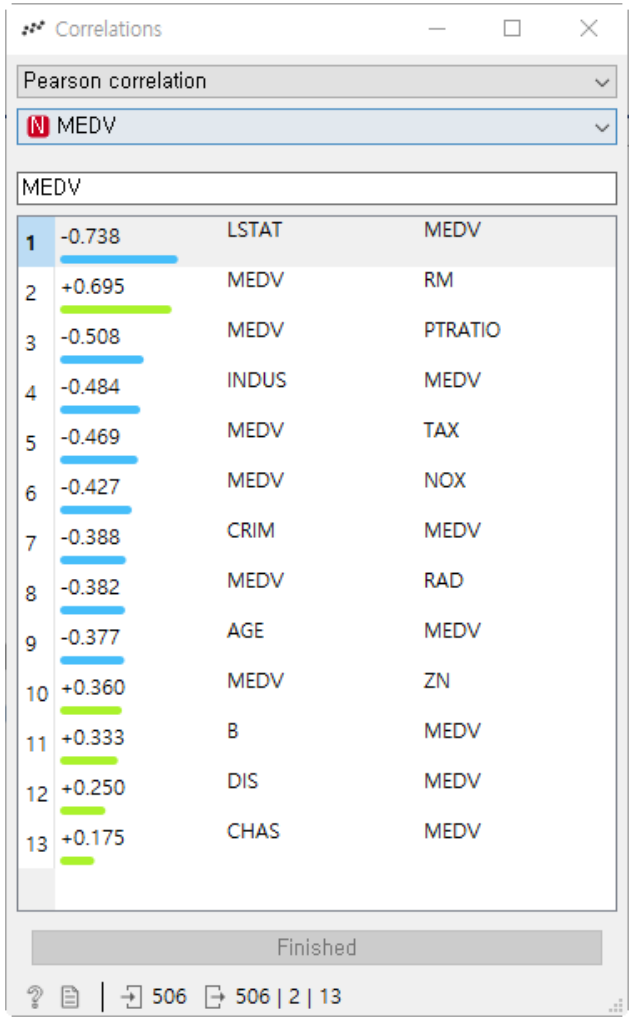
02. 탐색적 데이터 분석

■ 산점도: Scatter Plot



02. 탐색적 데이터 분석

- **상관 분석**: Correlations
  - MEDV-LSTAT: -0.738
  - MEDV-RM: +0.695
  - MEDV-PTRATIO: -0.508
  - MEDV-INDUS: -0.484
  - MEDV-TAX: -0.469



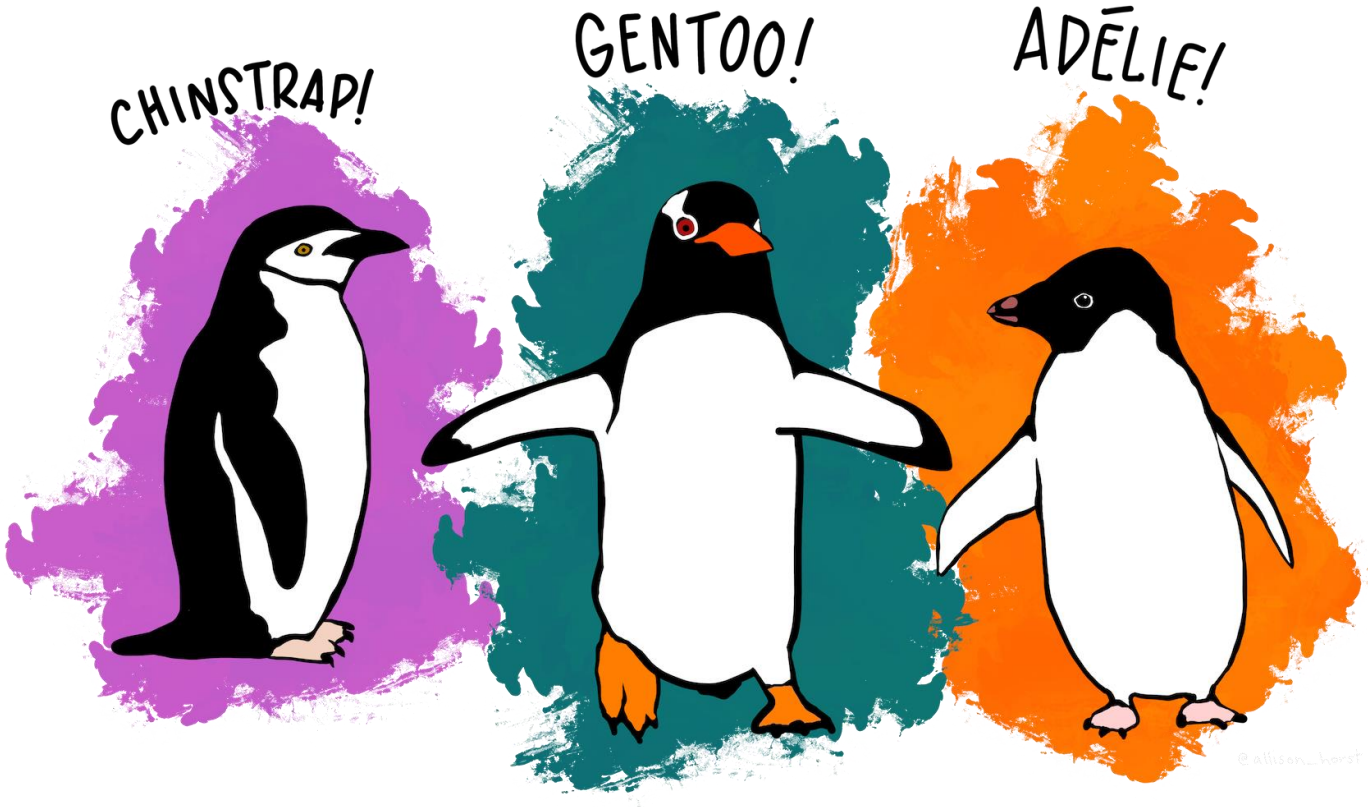
- **팔머펭귄 데이터셋: *palmerpenguins* dataset**
  - 남극의 팔머 군도에 서식하는 3종의 펭귄에 대한 데이터셋
  - 데이터 분석과 시각화 교육용으로 적절한 특성을 가지고 있음

https://bit.ly/36rDgFx

Gorman, Kristen B., Tony D. Williams, and William R. Fraser. "Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (genus Pygoscelis)." PloS one 9.3 (2014): e90081.

02. 데이터 탐색

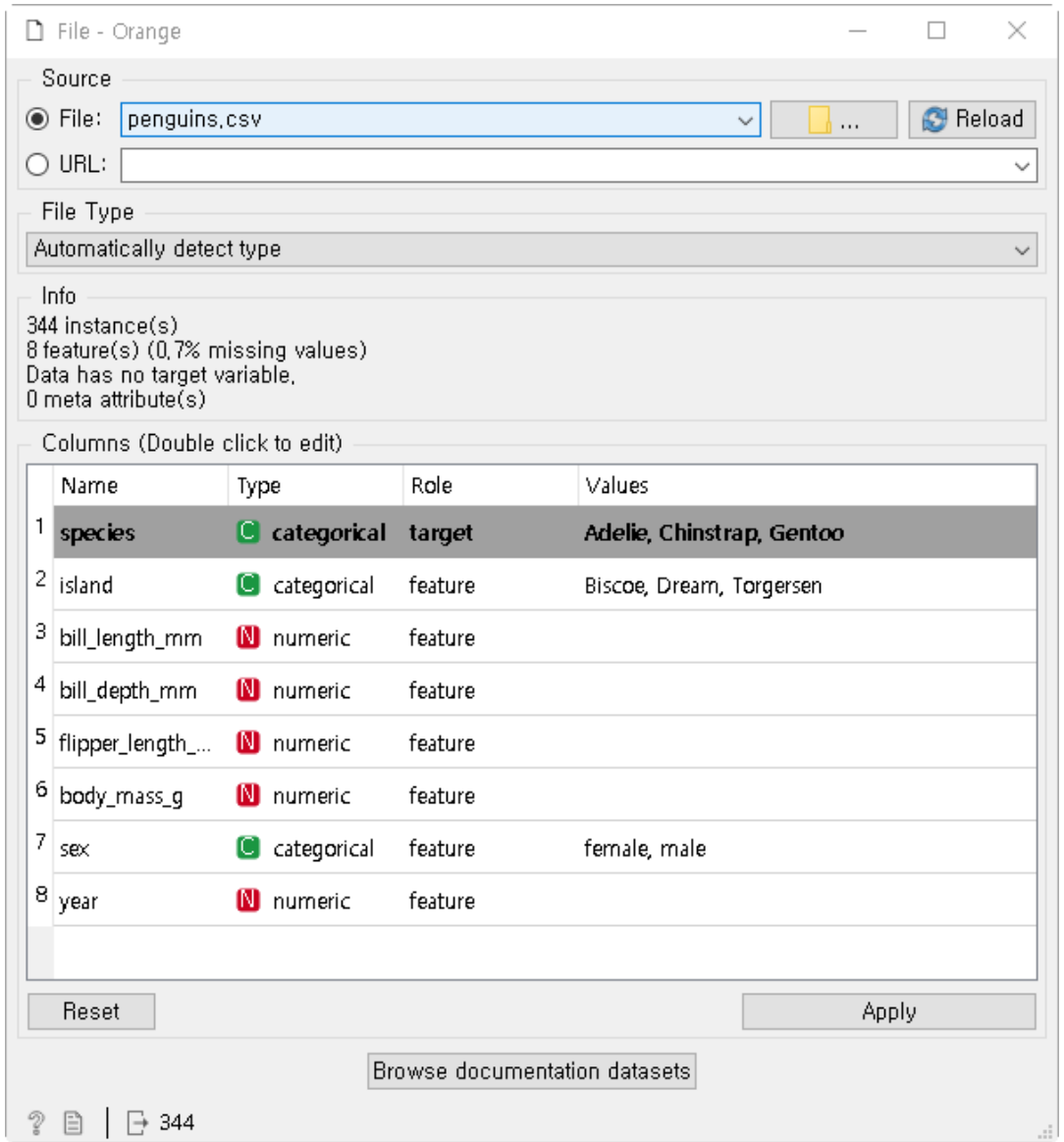
- **팔머펭귄의 종류:**
  - 턱끈: *chinstrap*
  - 젠투: *gentoo*
  - 아델리: *adelie*



Artwork by @allison_horst

02. 데이터 탐색

- **데이터셋 정보:**
  - 관측값: 344개
  - 특징변수: 8개
    - 수치형 변수: 5개
    - 범주형 변수: 3개
    - 종속변수: *species*
    - 독립변수: 7개

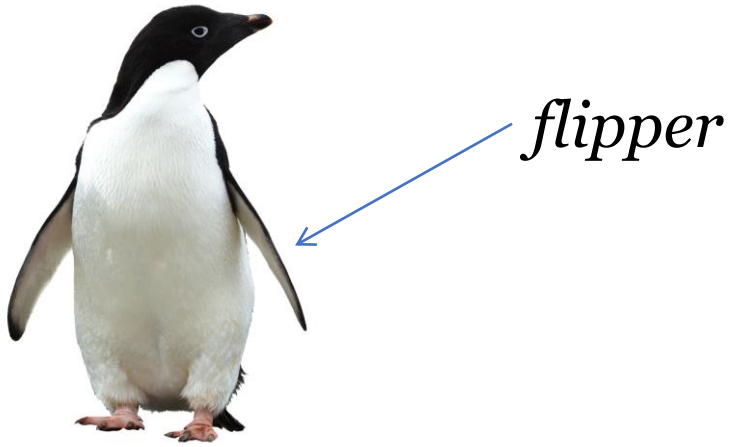


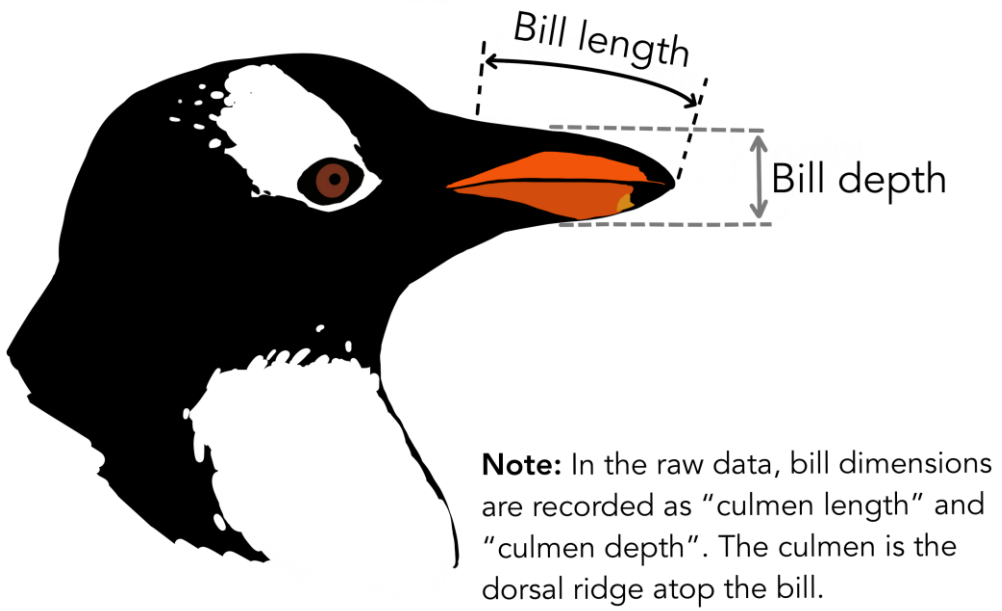
02. 데이터 탐색

- **범주형 변수**: *categorical* variables
  - *species*: **종**
    - Adelie, Chinstrap, Gentoo
  - *island*: **섬(서식지)**
    - Biscoe, Dream, Torgersen
  - *sex*: **성별**
    - female, male

02. 데이터 탐색

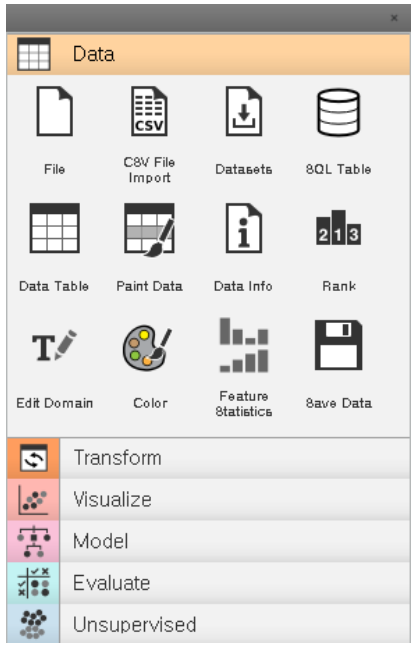
- **수치형 변수**: *numeric* variables
  - *bill_length_mm*: **부리의 길이**
  - *bill_depth_mm*: **부리의 깊이**
  - *flipper_length_mm*: **팔(?)의 길이** (**날개? 지느러미?**)
  - *body_mass_g*: **체중**
  - *year*: **연구년도**(2007, 2008, 2009)

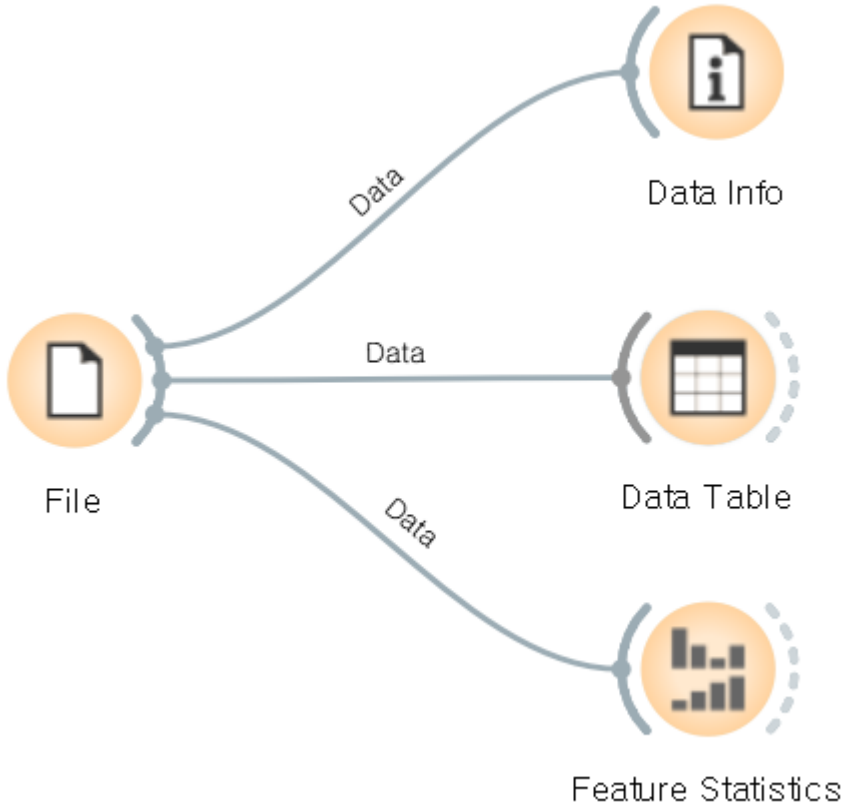




Note: In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal ridge atop the bill.

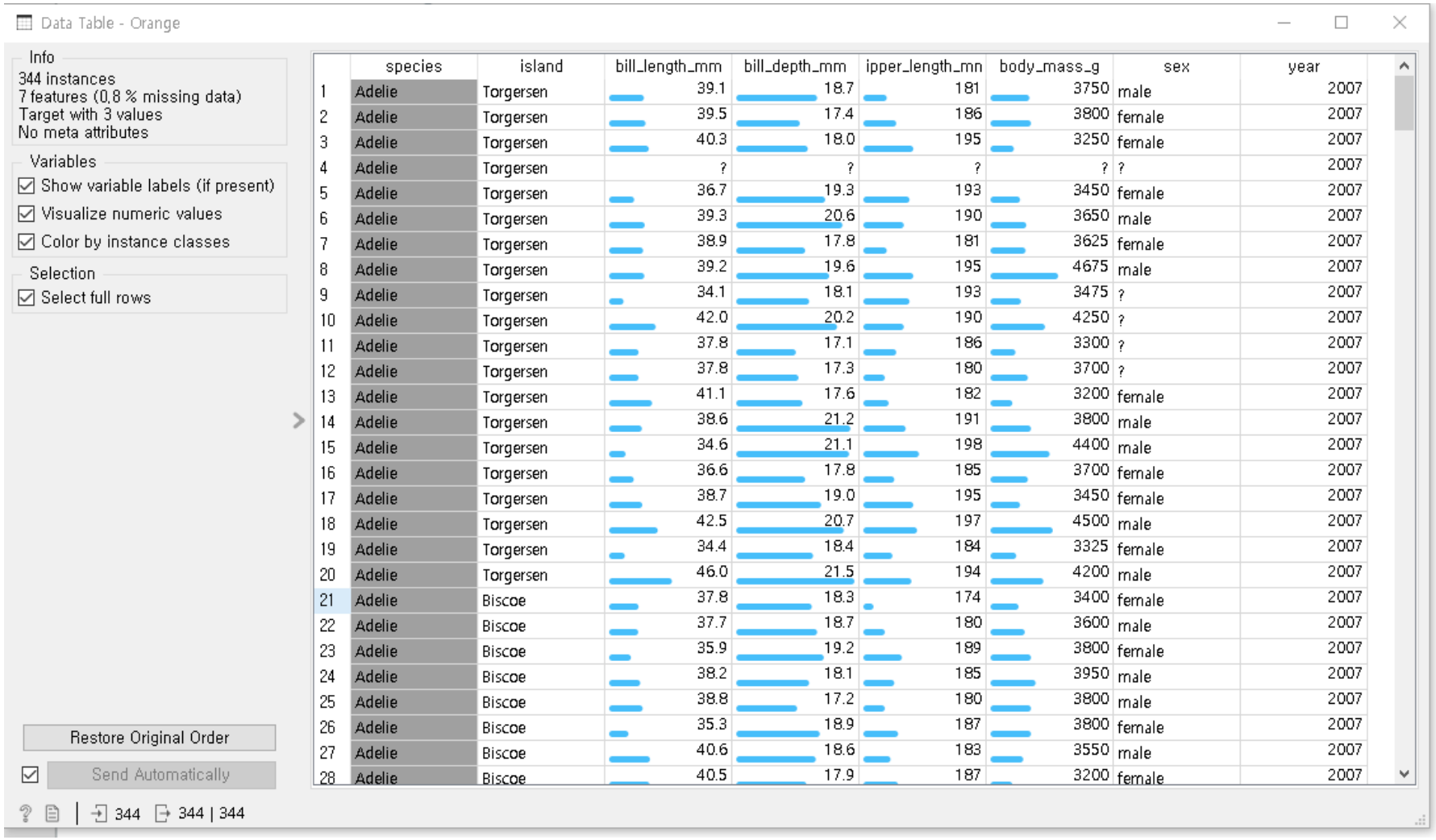
- **데이터 탐색**: Exploring Dataset





02. 데이터 탐색

- Data > Data Table

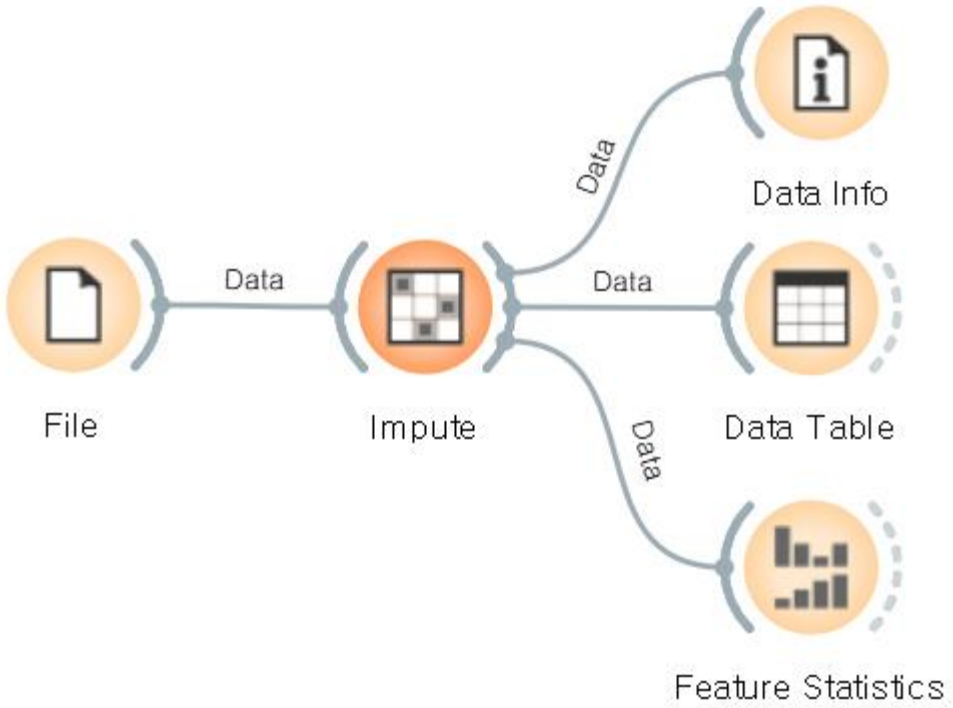


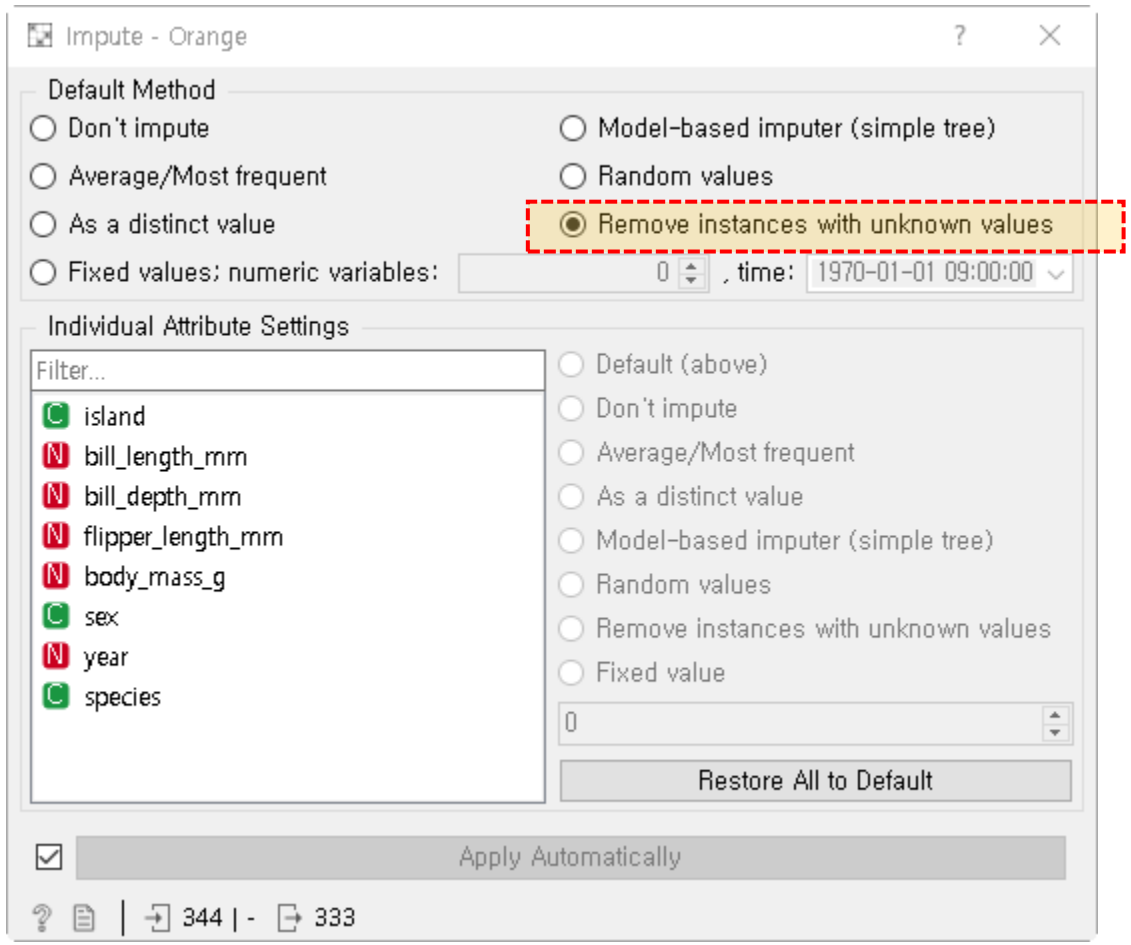
	species	island	bill_length_mm	bill_depth_mm	ipper_length_mn	body_mass_g	sex	year
1	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
2	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
3	Adelie	Torgersen	40.3	18.0	195	3250	female	2007
4	Adelie	Torgersen	?	?	?	?	?	2007
5	Adelie	Torgersen	36.7	19.3	193	3450	female	2007
6	Adelie	Torgersen	39.3	20.6	190	3650	male	2007
7	Adelie	Torgersen	38.9	17.8	181	3625	female	2007
8	Adelie	Torgersen	39.2	19.6	195	4675	male	2007
9	Adelie	Torgersen	34.1	18.1	193	3475	?	2007
10	Adelie	Torgersen	42.0	20.2	190	4250	?	2007
11	Adelie	Torgersen	37.8	17.1	186	3300	?	2007
12	Adelie	Torgersen	37.8	17.3	180	3700	?	2007
13	Adelie	Torgersen	41.1	17.6	182	3200	female	2007
14	Adelie	Torgersen	38.6	21.2	191	3800	male	2007
15	Adelie	Torgersen	34.6	21.1	198	4400	male	2007
16	Adelie	Torgersen	36.6	17.8	185	3700	female	2007
17	Adelie	Torgersen	38.7	19.0	195	3450	female	2007
18	Adelie	Torgersen	42.5	20.7	197	4500	male	2007
19	Adelie	Torgersen	34.4	18.4	184	3325	female	2007
20	Adelie	Torgersen	46.0	21.5	194	4200	male	2007
21	Adelie	Biscoe	37.8	18.3	174	3400	female	2007
22	Adelie	Biscoe	37.7	18.7	180	3600	male	2007
23	Adelie	Biscoe	35.9	19.2	189	3800	female	2007
24	Adelie	Biscoe	38.2	18.1	185	3950	male	2007
25	Adelie	Biscoe	38.8	17.2	180	3800	male	2007
26	Adelie	Biscoe	35.3	18.9	187	3800	female	2007
27	Adelie	Biscoe	40.6	18.6	183	3550	male	2007
28	Adelie	Biscoe	40.5	17.9	187	3200	female	2007

02. 데이터 탐색

- Data > Feature Statistics

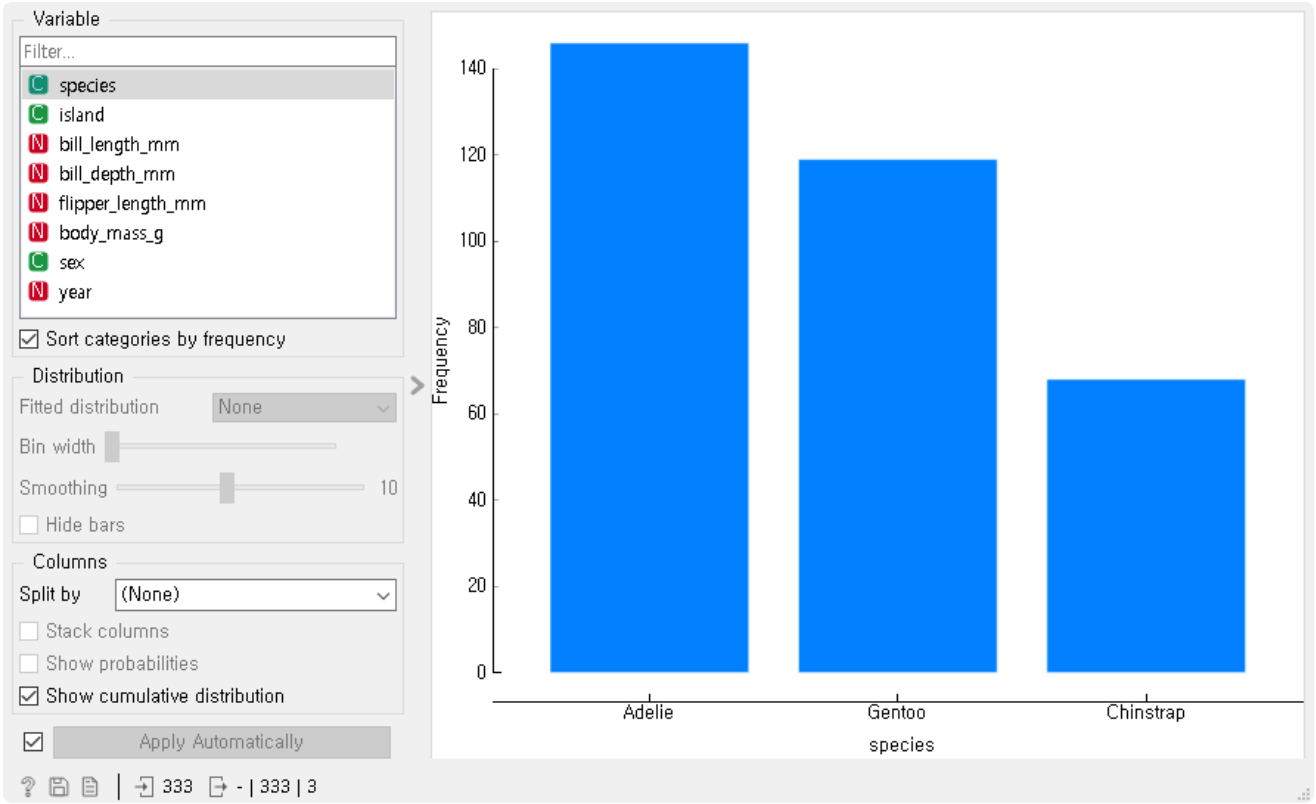
- **결측치 제거: *missing* values**
  - Transform > Impute

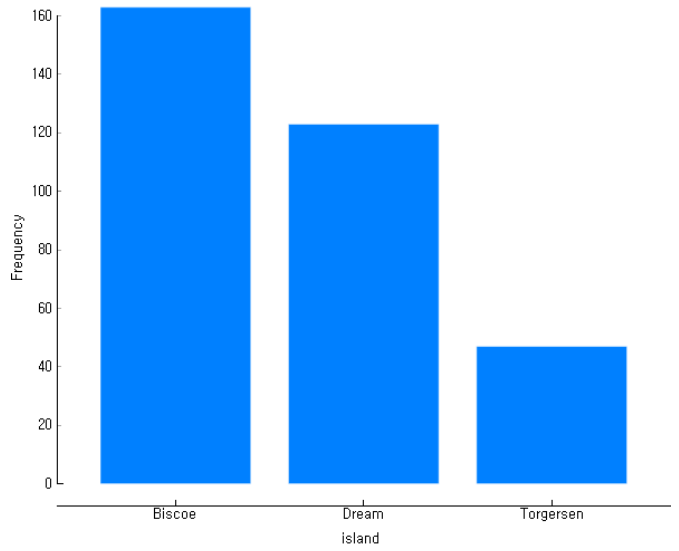


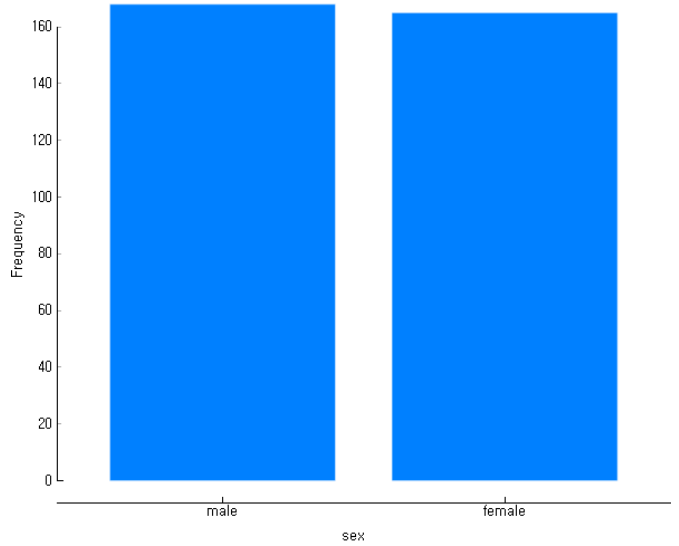


02. 데이터 탐색

- **범주형 데이터의 탐색:** Visualize > *Distributions*
  - **빈도표:** *frequency* table
    - **종별**(species), **섬별**(island), **성별**(sex)

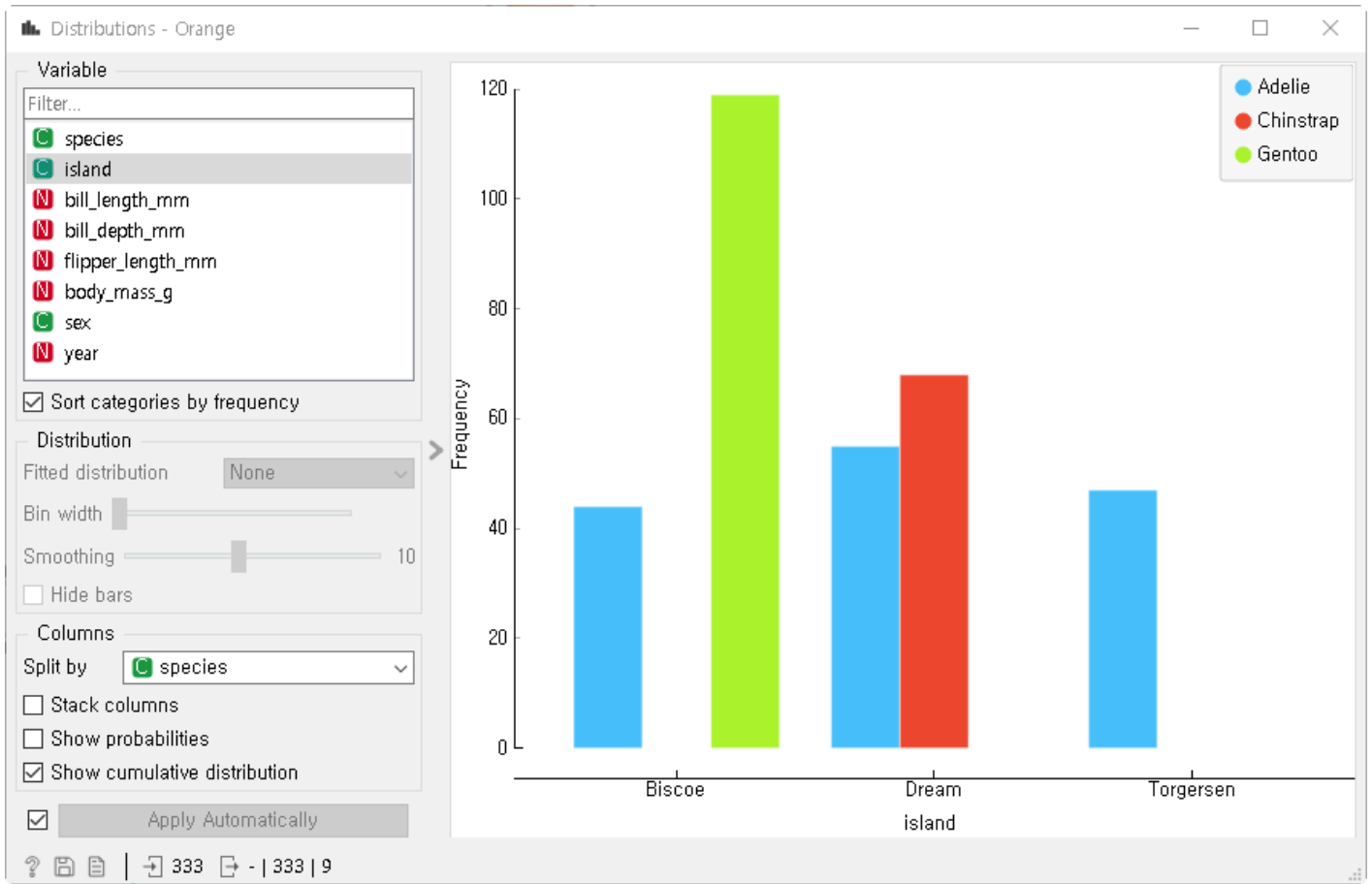


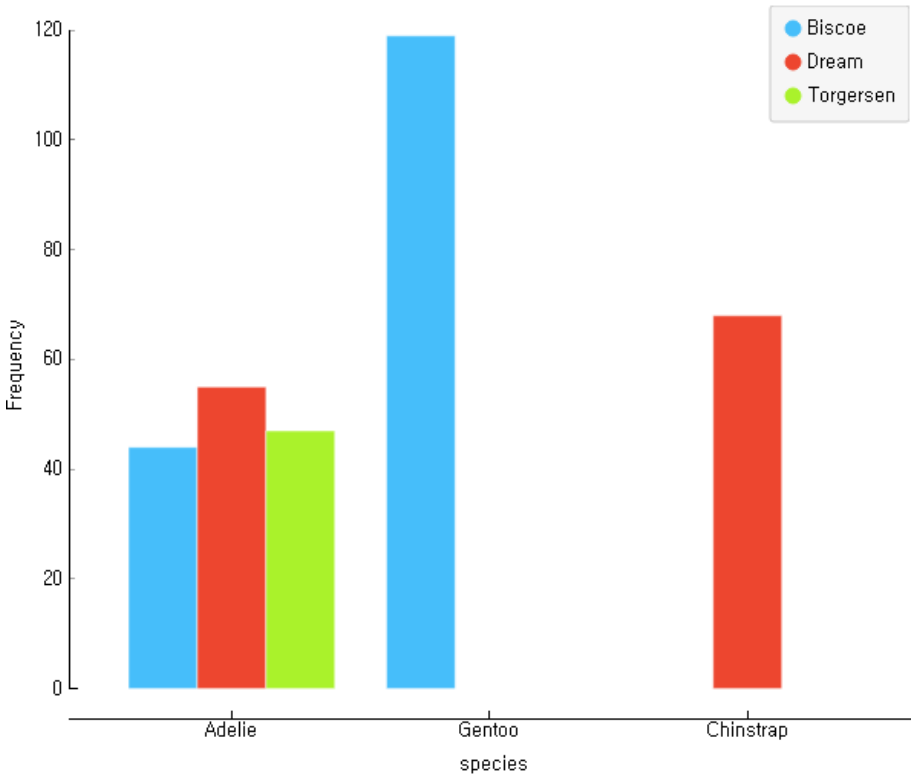




02. 데이터 탐색

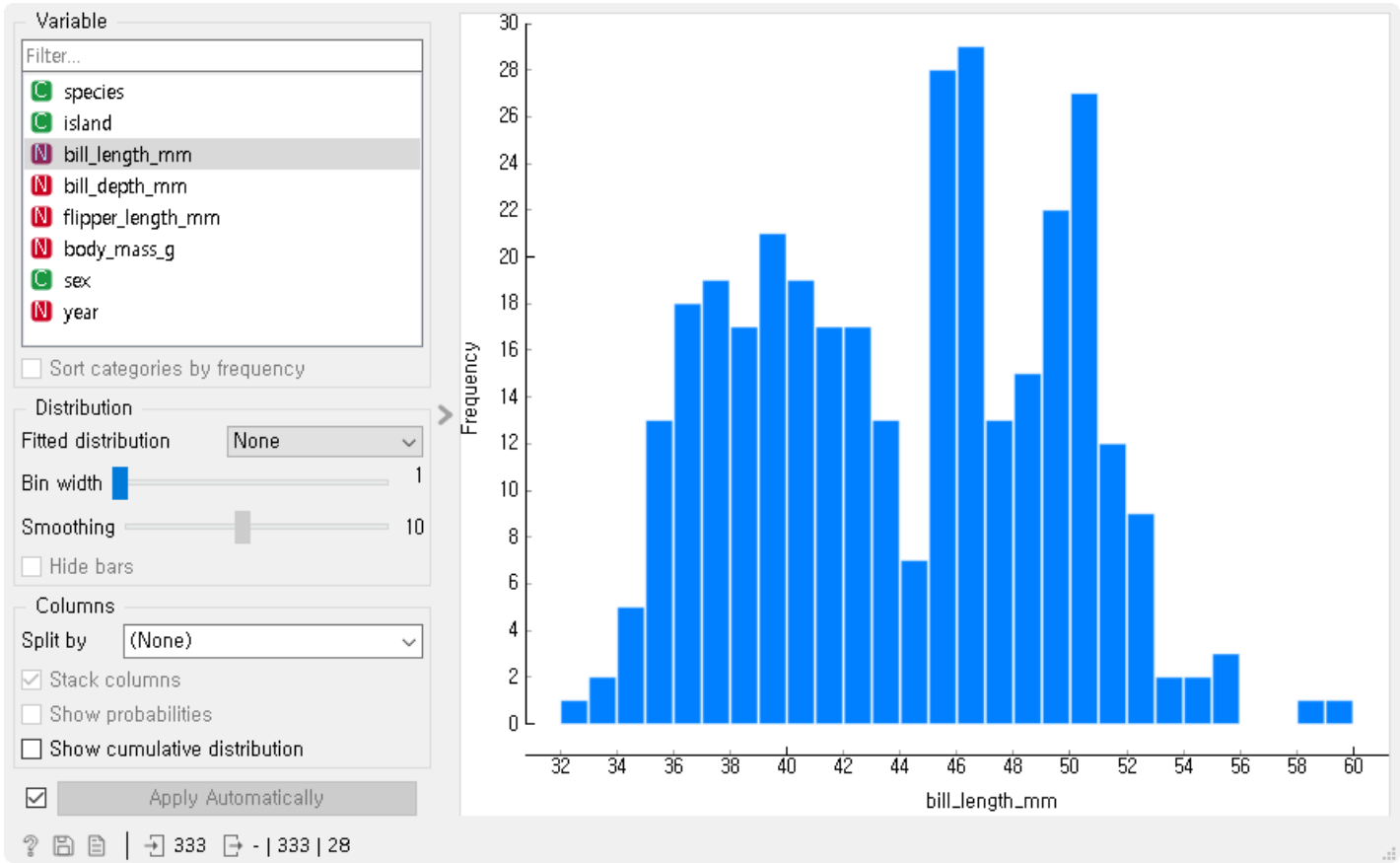
- 섬별로 어떤 종이 서식하는가? 또는, 종별로 어떤 섬에 서식하는가?

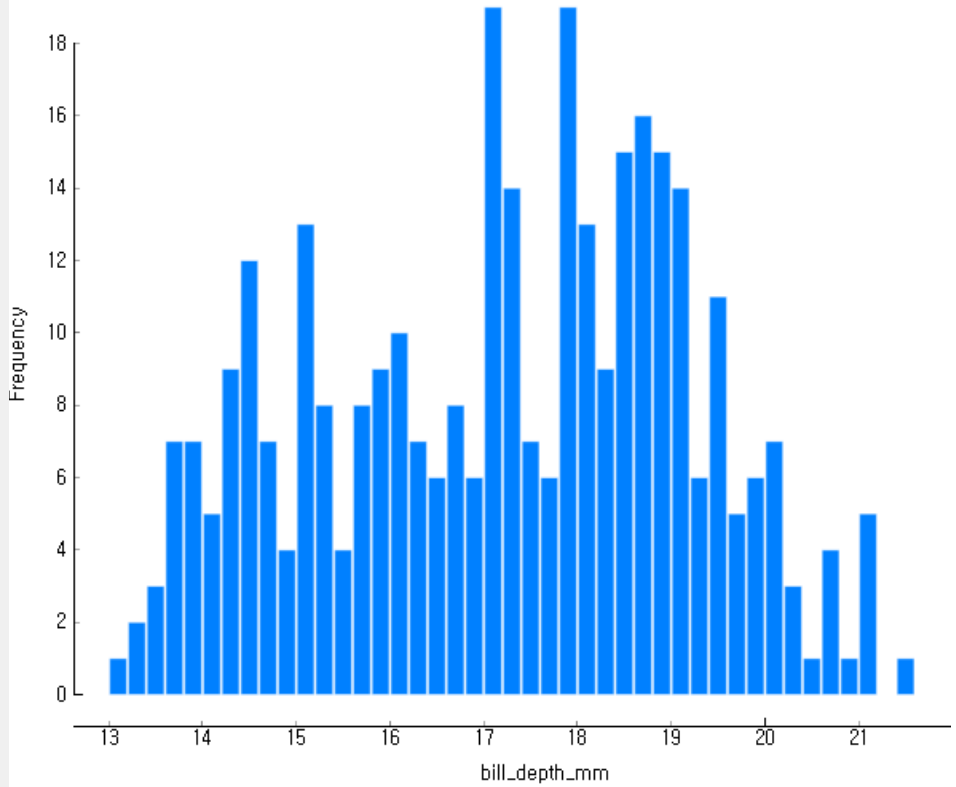




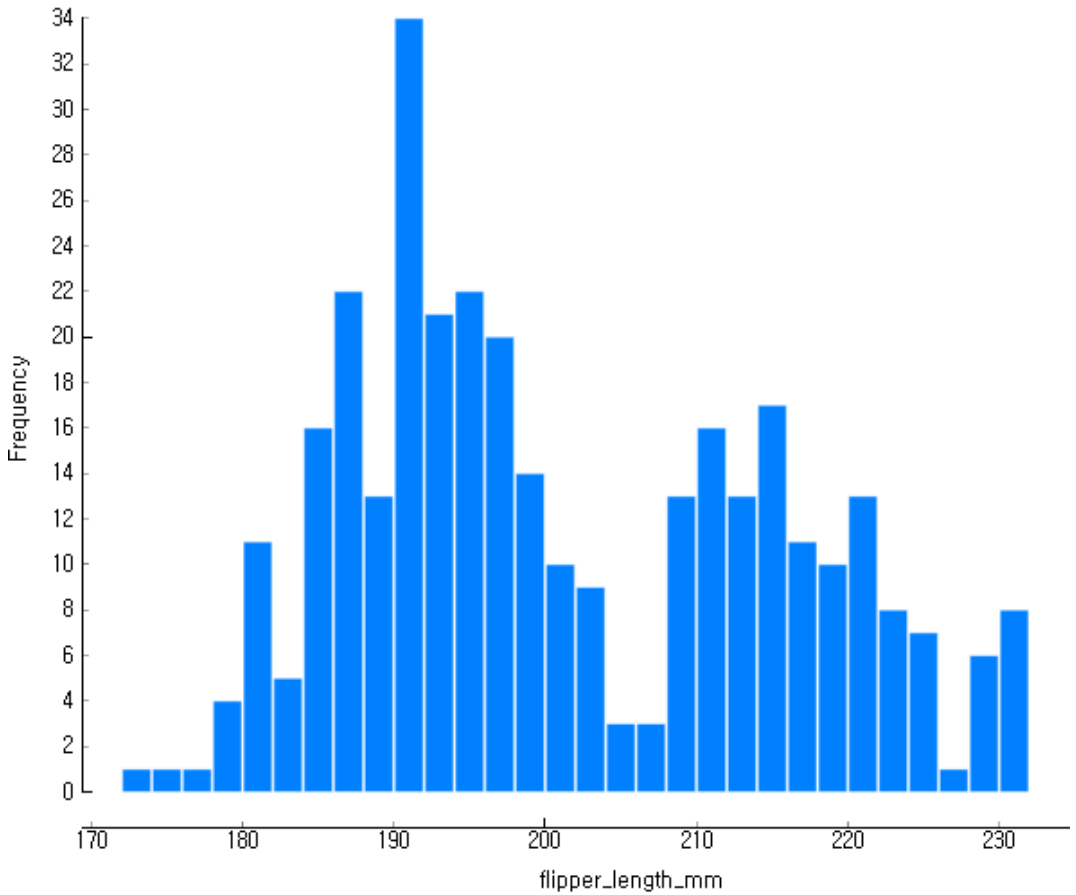
02. 데이터 탐색

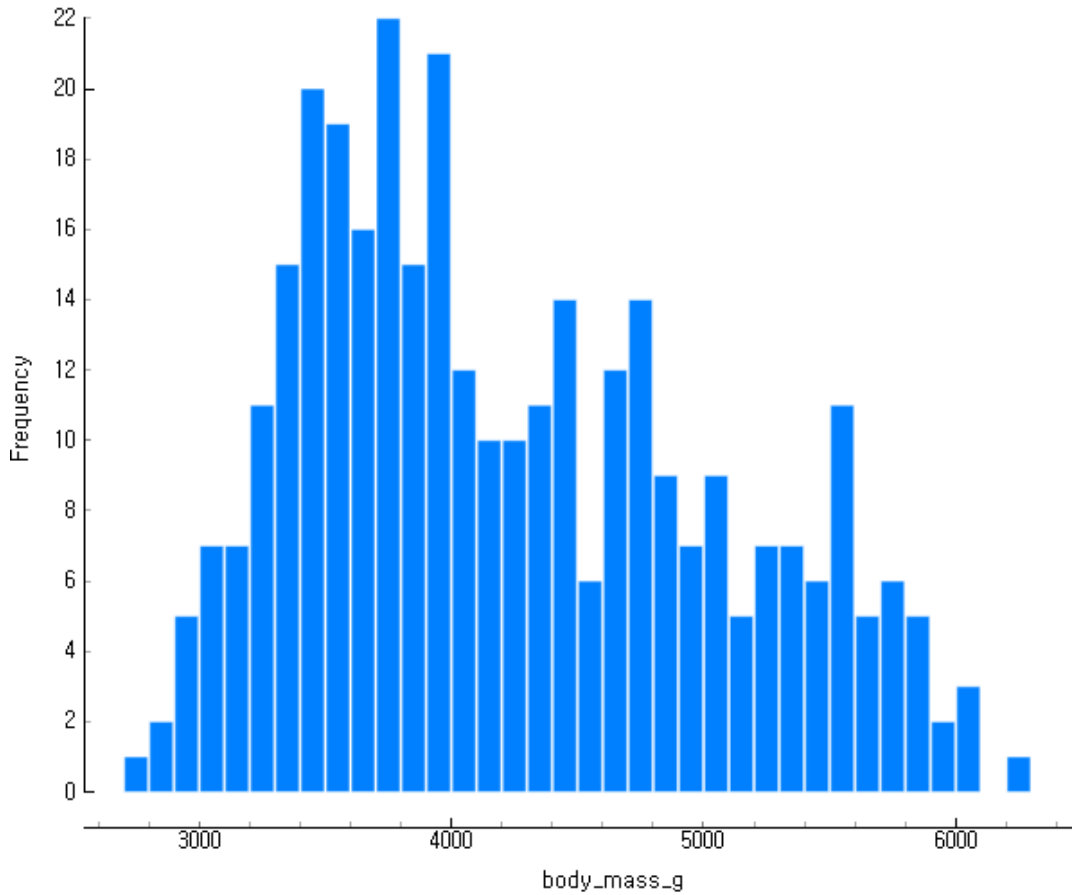
- **수치형 변수의 탐색**: Distributions
  - **히스토그램**: *histogram*





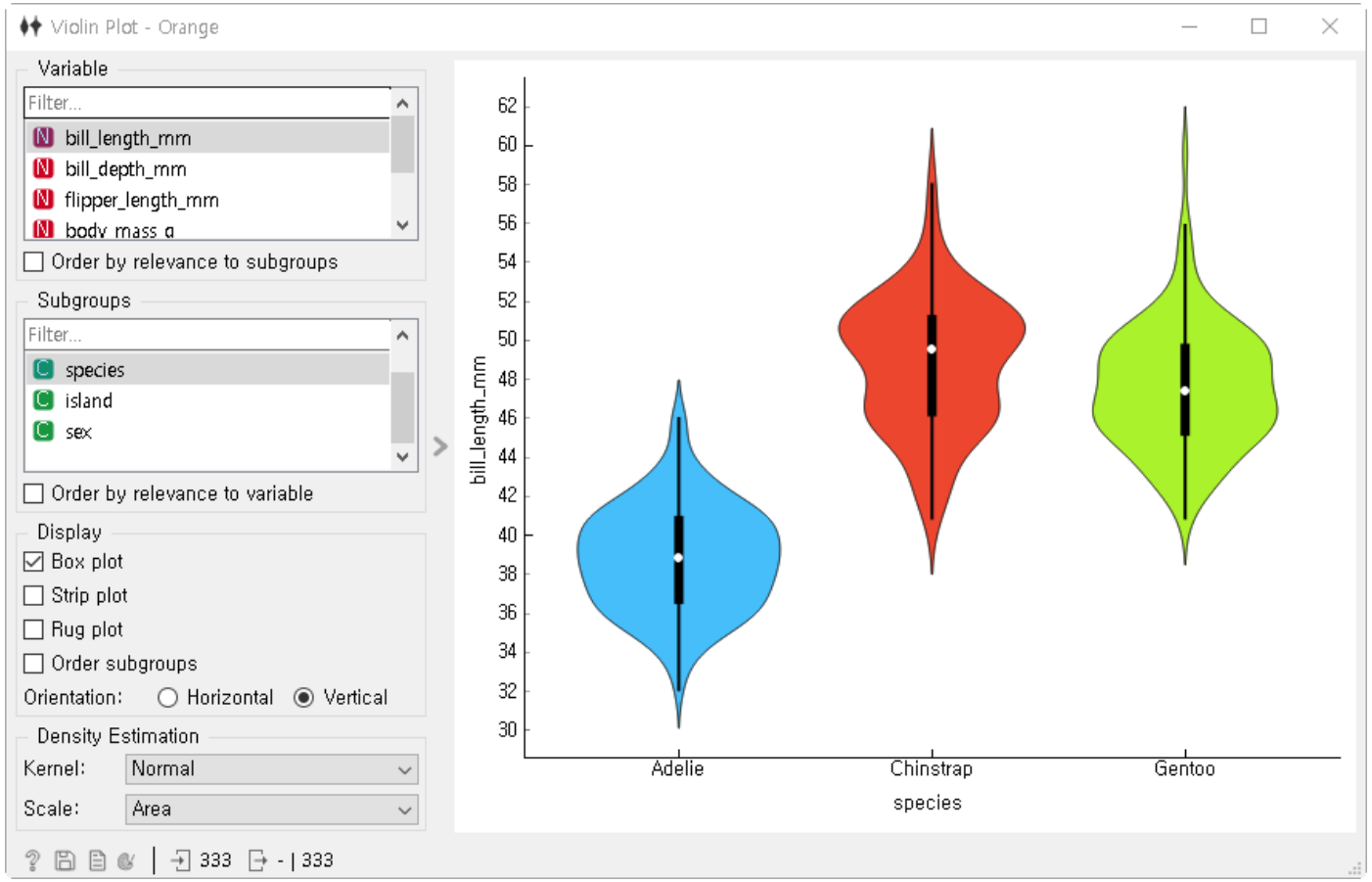
02. 데이터 탐색

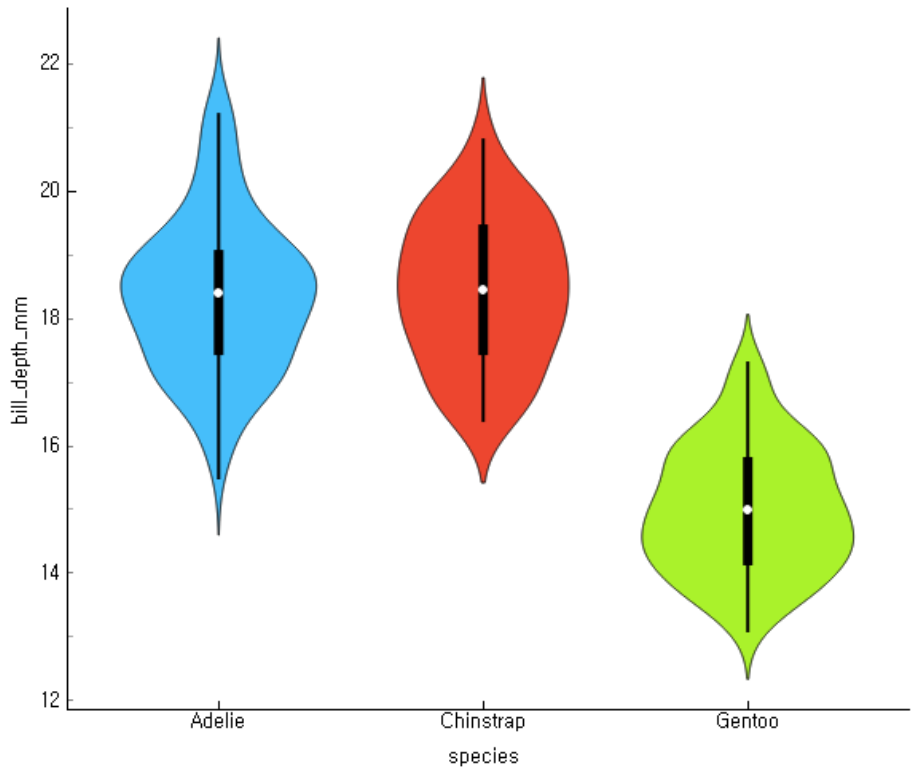

Frequency
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
170
180
190
200
210
220
230
flipper_length_mm


Frequency
22
20
18
16
14
12
10
8
6
4
2
0
3000
4000
5000
6000
body_mass_g

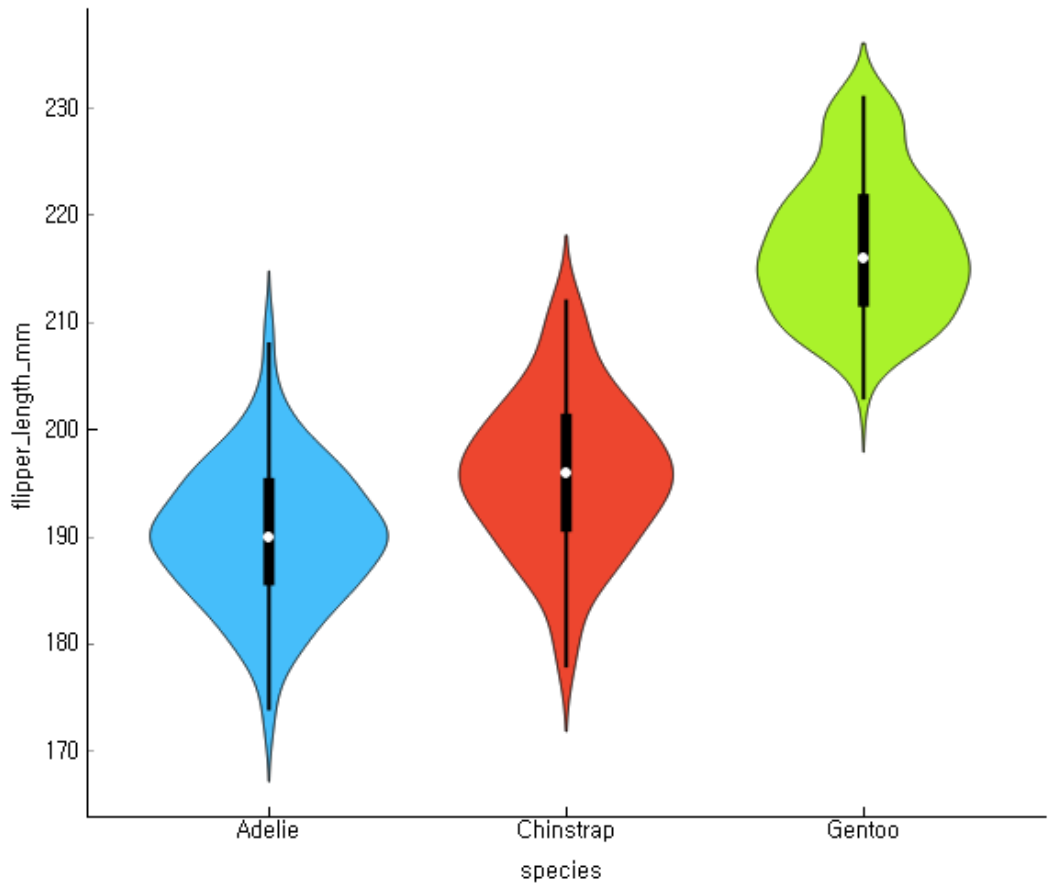
02. 데이터 탐색

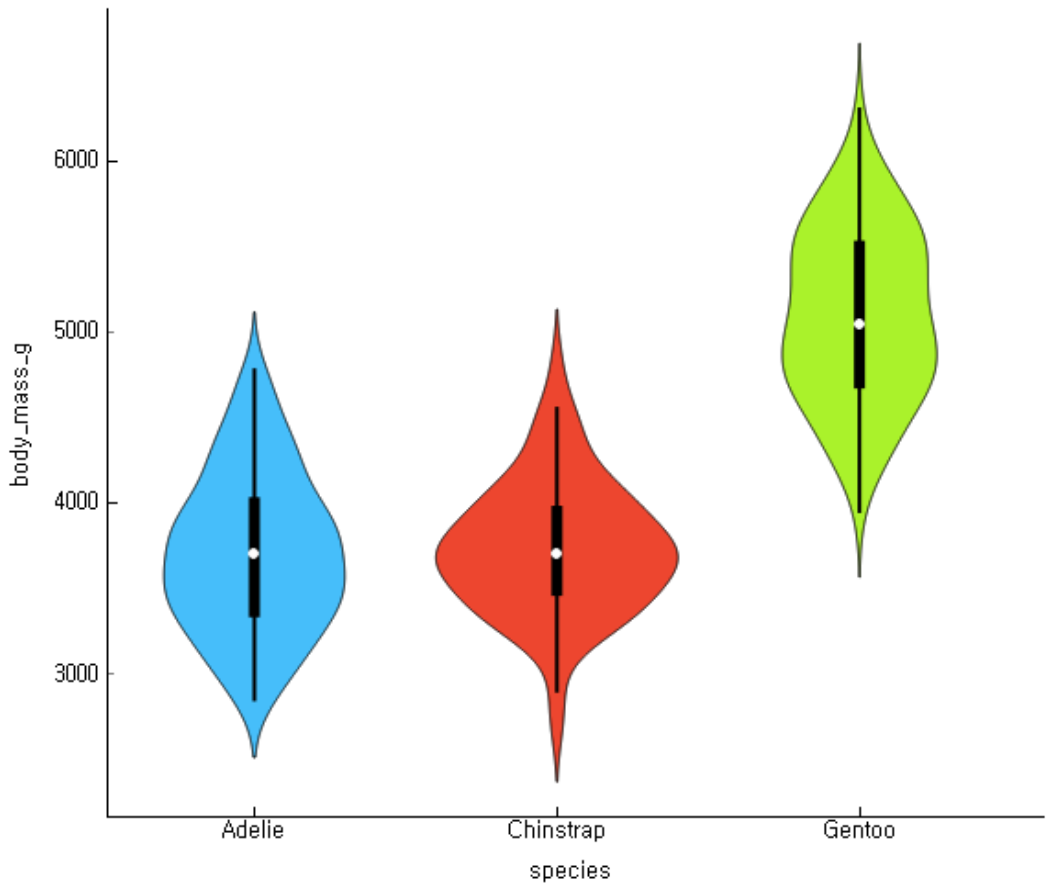
- **종별로 수치형 데이터의 분포 탐색**: Visualize > Violin Plot





02. 데이터 탐색

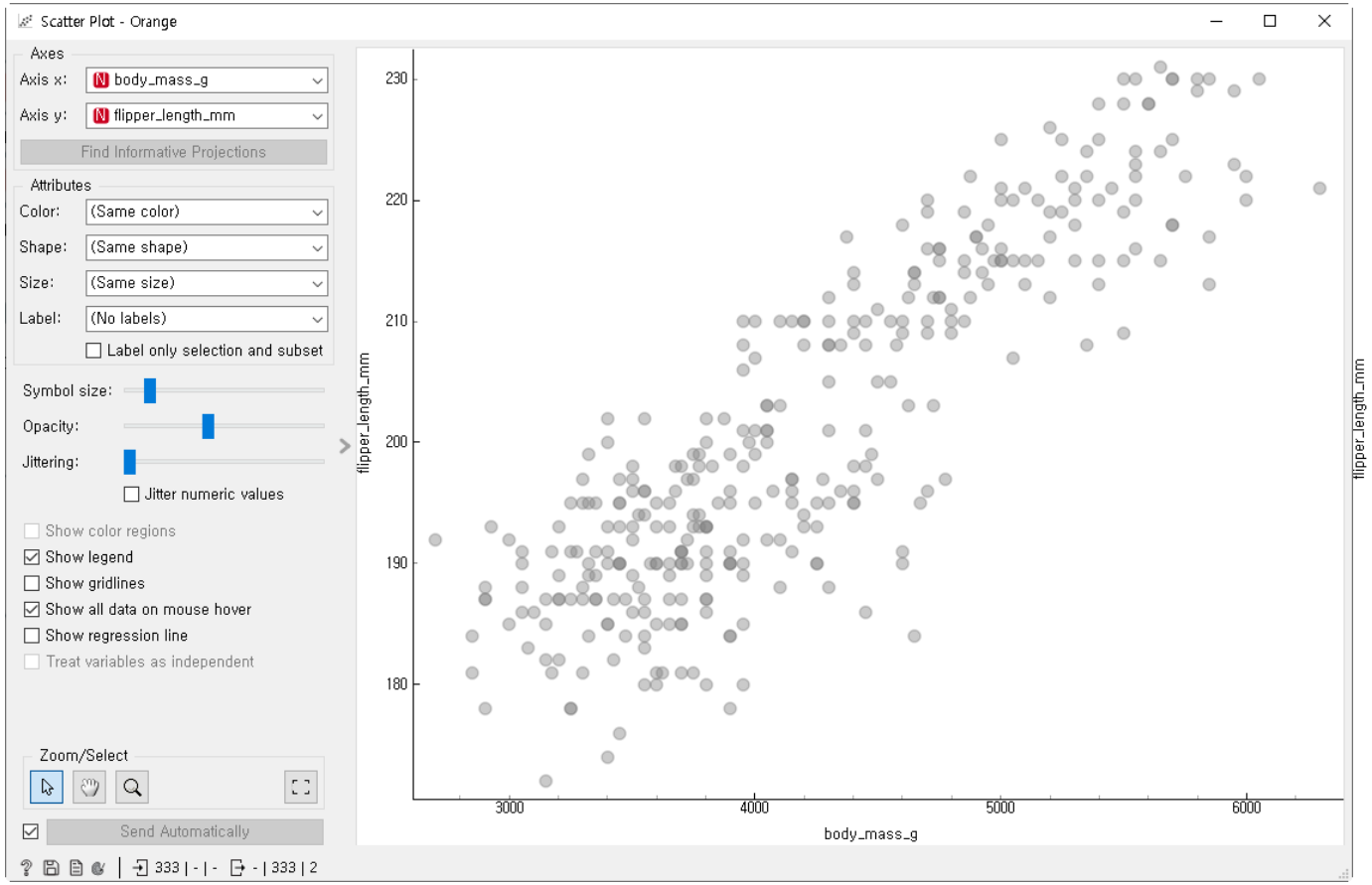

flipper_length_mm
230
220
210
200
190
180
170
Adelie
Chinstrap
Gentoo
species

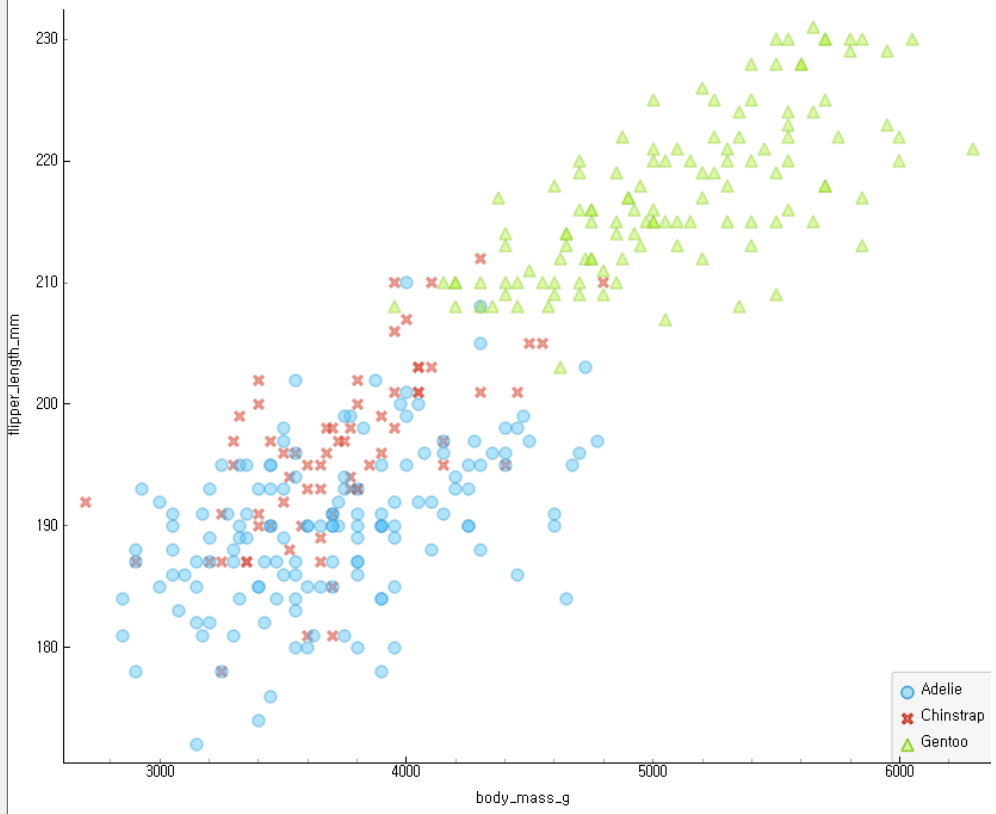

body_mass_g
6000
5000
4000
3000
Adelie
Chinstrap
Gentoo
species

02. 데이터 탐색

■ 수치형 변수 간의 관계 탐색: Scatter Plot

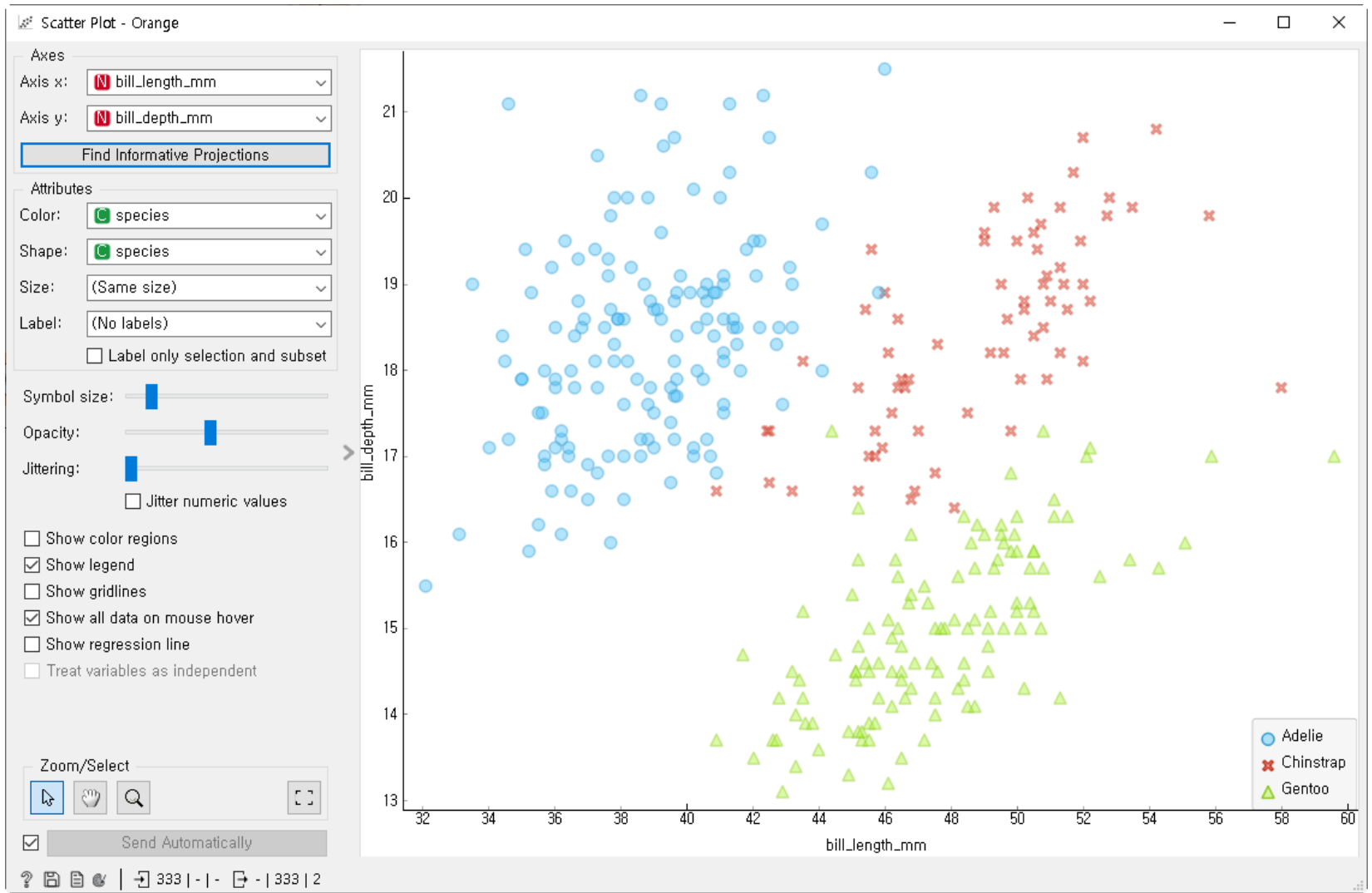
- 체중과 날개의 길이





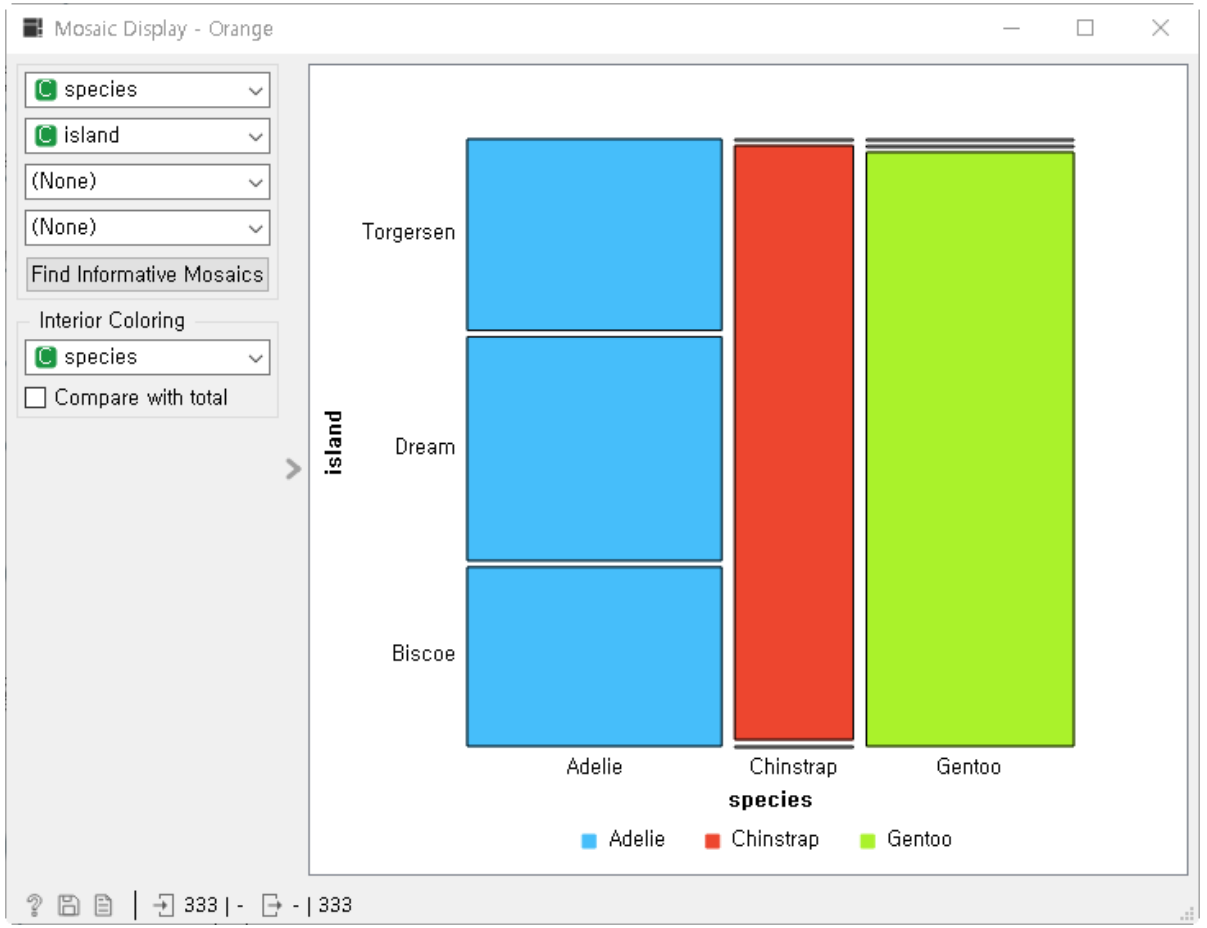
02. 데이터 탐색

- 부리의 길이와 높이

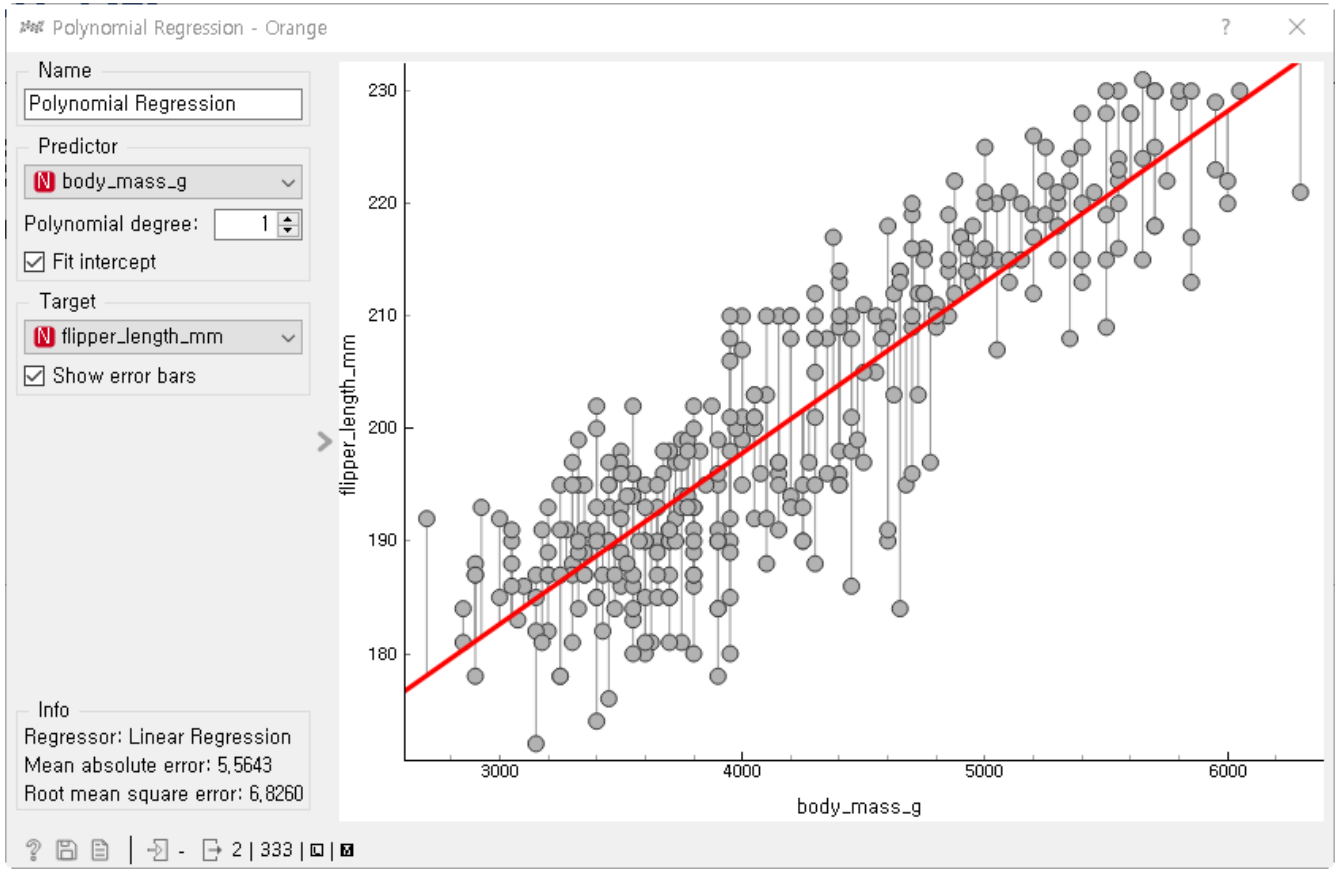


■ 범주형 변수 간의 관계 탐색: Mosaic Plot

- 종과 서식지와의 관계

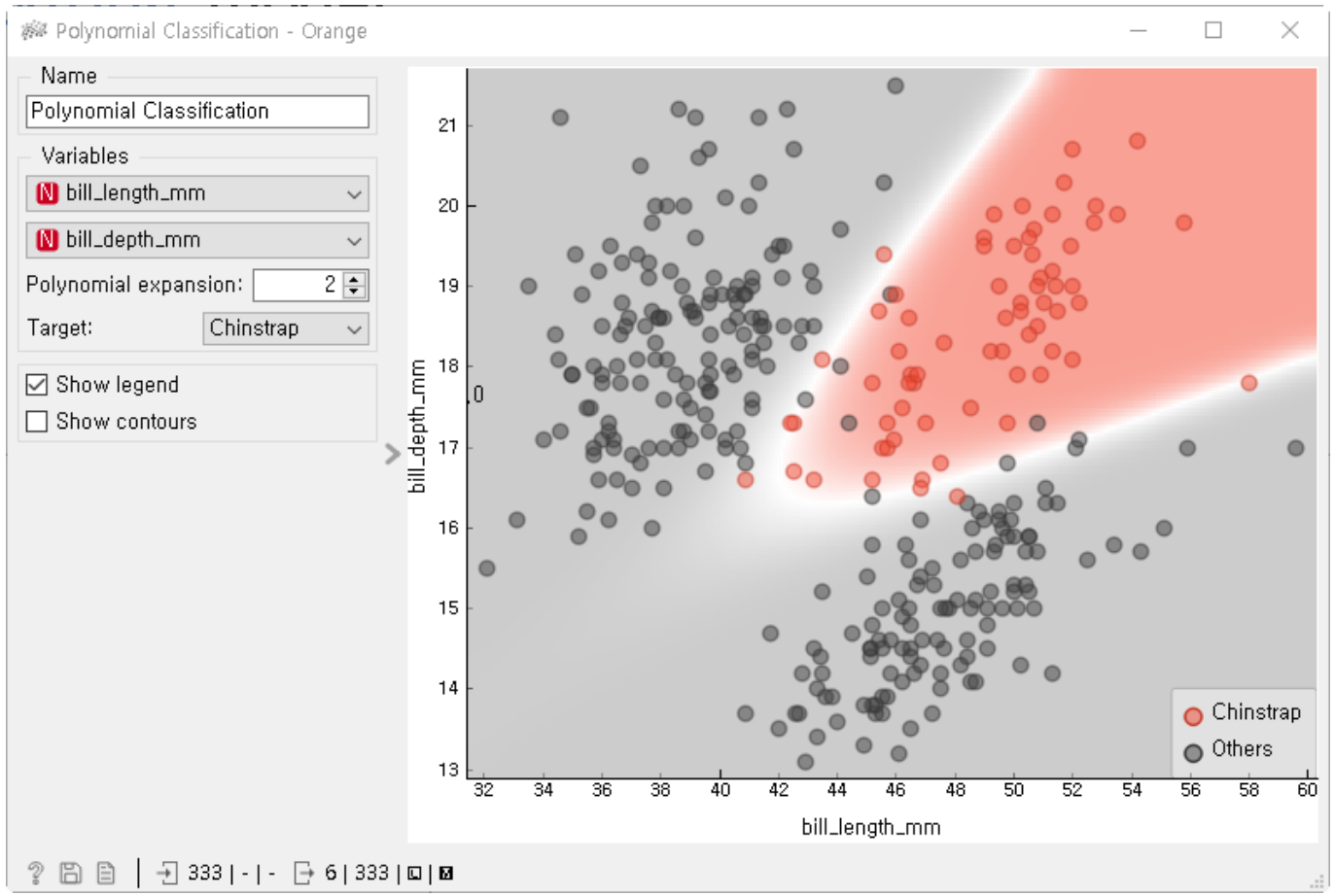


- **예측 모델**: *prediction* model
  - Educational > Polynomial Regression
  - **체중에 따른 날개의 길이 예측 모델**



02. 데이터 탐색

- **분류 모델**: *classification* model
  - Educational > Polynomial Classification
  - **턱끈 펭귄 분류하기**



Any Questions?