



ML/DS 지식 따라가기

데이터 전처리와 EDA



데이터 전처리와 EDA

데이터 전처리의 절차

Data Exploration

데이터를 어디서 찾지?



Data Cleaning

데이터가 가진 문제를 해결하자!



EDA

데이터가 가진 특성을 확인하자!



Data Transformation

데이터를 분석하기 쉽게 만들자!

데이터 전처리와 EDA

데이터 전처리의 절차

데이터 누락

사람	직급	월급
Jason	과장	\$ 27,000
David	차장	
Wilson	사원	\$ 13,000

Data Cleaning

데이터가 가진 문제를 해결하자!

데이터 전처리와 EDA

데이터 전처리의 절차

Data Cleaning

데이터가 가진 문제를 해결하자!

데이터 누락

누락의 종류

(1) MCAR

- Missing Completely At Random
- 완전 무작위 결측
- ex) 종이에 커피가 묻어 특정 값이 안 보임

(2) MNAR

- Missing Not At Random
- 결측된 이유가 결측된 열과 관련 있음
- ex) 구매가를 안 알려주려고 가격을 누락함

(3) MAR

- Missing At Random
- 결측된 이유가 결측되지 않은 열과 관련 있음
- ex) 어디에 썼는지 안 알려주려고 가격을 누락함

데이터 전처리와 EDA

데이터 전처리의 절차

데이터 누락

Data Cleaning

데이터가 가진 문제를 해결하자!

해결법

(1) 행/열 제거하기

(2) 평균값 넣기

+

(3) K-NN으로 추측하기

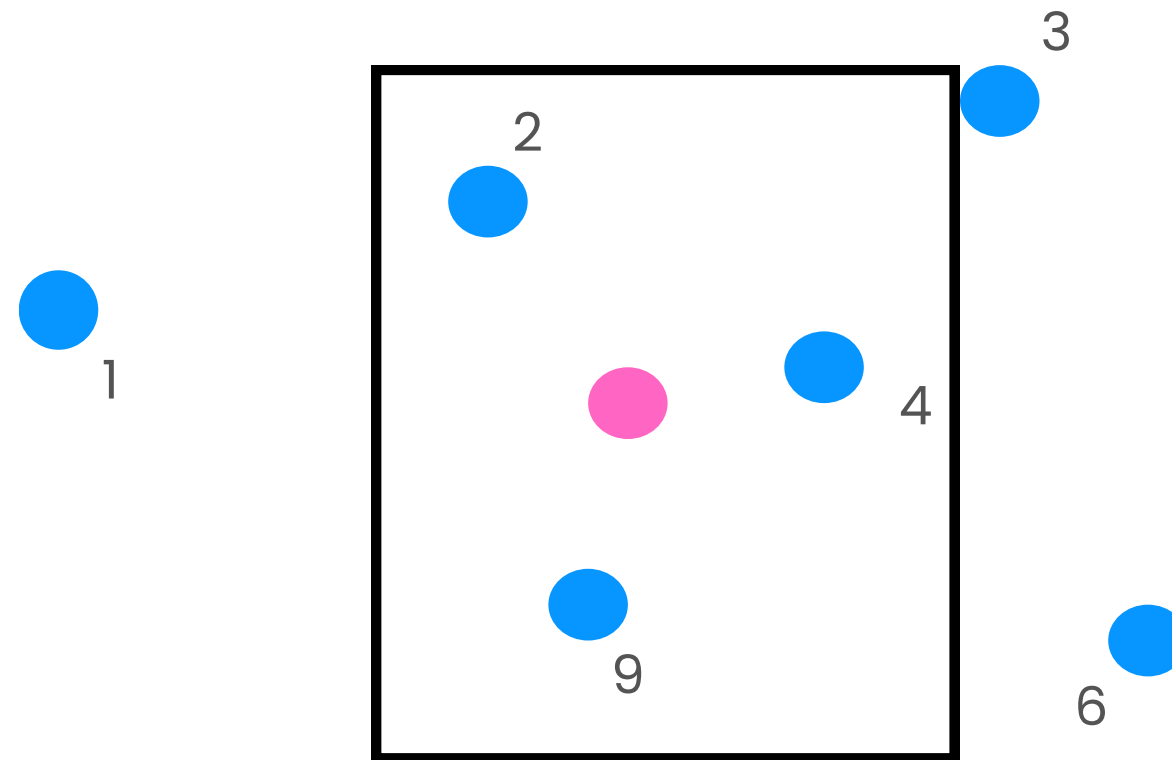
(4) Missforest로 추측하기

데이터 전처리와 EDA

데이터 전처리의 절차

근처에 있는 값들을 바탕으로 누락된 값을 추측하는 방법

**KNN을 이용한
추측**



● 의 값 :
 $(2+4+9) / 3 = 5$

데이터 전처리와 EDA

데이터 전처리의 절차

Random Forest를 바탕으로 누락된 값을 추측하는 방법

**Missforest를
이용한 추측**

독립변수 독립변수 종속변수

사람	직급	월급
Jason	과장	\$ 27,000
David	차장	
Wilson	사원	\$ 13,000

독립변수를 기반으로 종속변수를
추측하는 RF 모델 제작



특정 값이나 특정 횟수를
기준으로 종료함

데이터 전처리와 EDA

데이터 전처리의 절차

데이터 누락

데이터 중복

사람	직급	월급
Jason	과장	\$ 27,000
David	차장	\$ 20,000
Wilson	사원	\$ 13,000
Wilson	사원	\$ 13,000

Data Cleaning

데이터가 가진 문제를 해결하자!

데이터 전처리와 EDA

데이터 전처리의 절차

데이터 누락

데이터 중복

Data Cleaning

데이터가 가진 문제를 해결하자!

해결법

중복은 제거합시다

ex)

```
test_df[~test_df.duplicated()]
```

데이터 전처리와 EDA

데이터 전처리의 절차

Data Cleaning

데이터가 가진 문제를 해결하자!



사람	직급	월급
Jason	과장	\$ 27,000
David	차장	\$ 20,000
Wilson	사랑	\$ 13,000
Wilson	사원	\$ 13,000

데이터 전처리와 EDA

데이터 전처리의 절차

데이터 누락

데이터 중복

데이터 오류

해결법

단순 오타인 경우 :
확인 후 수정

오타가 아닌 경우 :
배경 지식을 가지고 있어야 함

Data Cleaning

데이터가 가진 문제를 해결하자!

데이터 전처리와 EDA

데이터 전처리의 절차

Data Cleaning

데이터가 가진 문제를 해결하자!



사람	직급	월급
Jason	과장	\$ 27,000
David	차장	\$ 20,000
Carlson	사장	\$ 10,000,000
Wilson	사원	\$ 13,000

데이터 전처리와 EDA

데이터 전처리의 절차

해결법

Data Cleaning

데이터가 가진 문제를 해결하자!



(1) 제거하기

(2) 값 변경하기

+

(3) 가중치를 조정하기

데이터 전처리와 EDA

EDA

Exploratory Data Analysis



탐색적인

EDA = 데이터를 탐색하고 이해하는 과정!

데이터 전처리와 EDA

EDA

EDA의 대상

일변량 (Univariate)

- 분석할 변수가 1개
- 데이터를 설명하고, 데이터의 패턴을 확인함

다변량 (Multi-variate)

- 분석할 변수가 여러개
- 변수간의 관계를 확인함

데이터 전처리와 EDA

EDA

시각화 (Graphic)

- 데이터를 한눈에 파악하여 대략적인 형태 파악 가능

비시각화 (Non-Graphic)

- 정확한 값을 파악하기 좋음

EDA의 종류

데이터 전처리와 EDA

EDA

`df.describe()`

	mpg	cyclinders	displacement	weight	accerleration \
count	398.000000	398.000000	398.000000	398.000000	398.000000
mean	23.514573	5.454774	193.425879	2970.424623	15.568090
std	7.815984	1.701004	104.269838	846.841774	2.757689
min	9.000000	3.000000	68.000000	1613.000000	8.000000
25%	17.500000	4.000000	104.250000	2223.750000	13.825000
50%	23.000000	4.000000	148.500000	2803.500000	15.500000
75%	29.000000	8.000000	262.000000	3608.000000	17.175000
max	46.600000	8.000000	455.000000	5140.000000	24.800000
	model year	origin			
count	398.000000	398.000000			
mean	76.010050	1.572864			
std	3.697627	0.802055			
min	70.000000	1.000000			
25%	73.000000	1.000000			
50%	76.000000	1.000000			
75%	79.000000	2.000000			
max	82.000000	3.000000			

데이터 전처리와 EDA

EDA

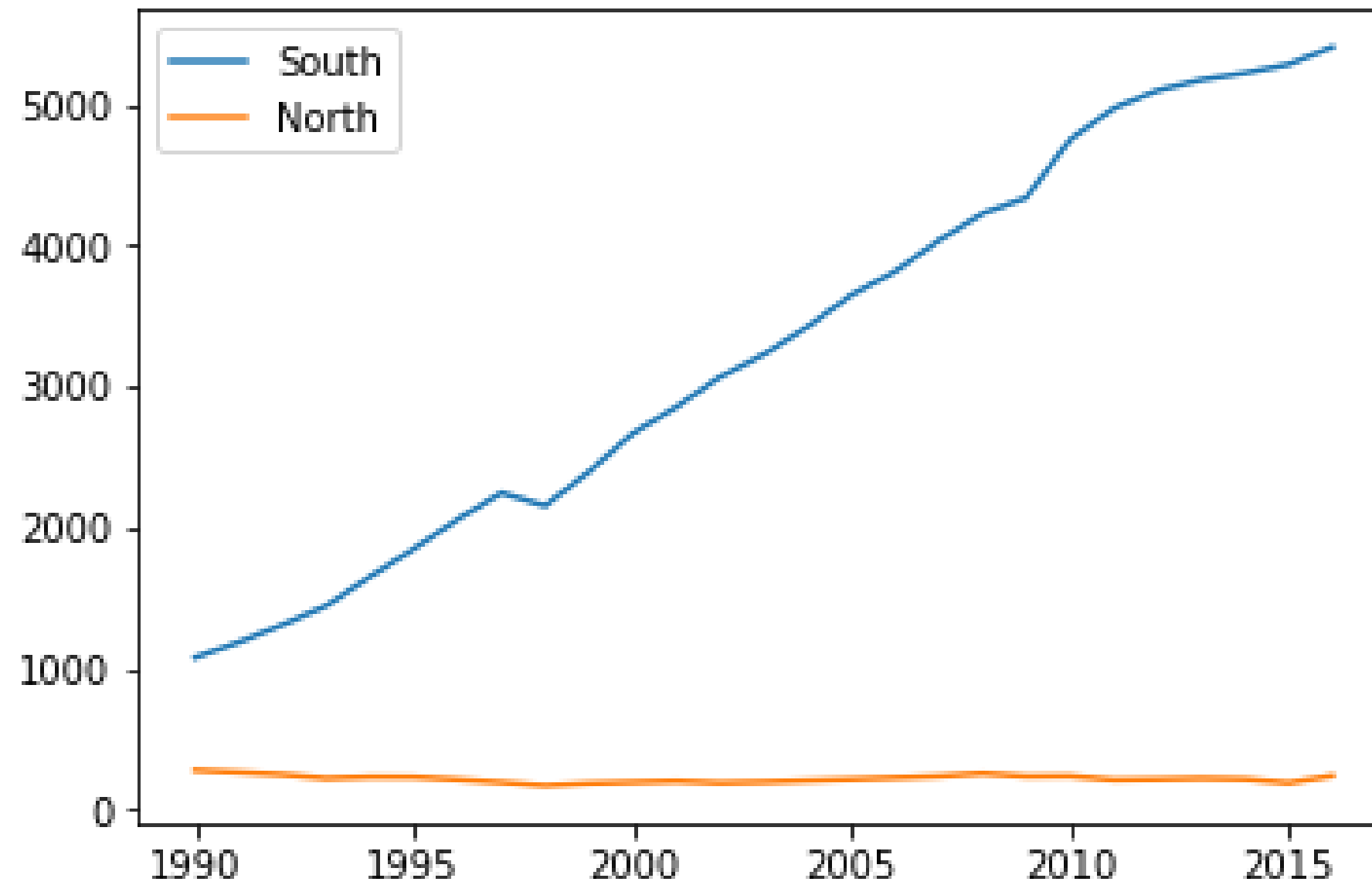
df.corr()

[Output]						
	mpg	cylinders	displacement	weight	accerleration	\
mpg	1.000000	-0.775396	-0.804203	-0.831741	0.420289	
cylinders	-0.775396	1.000000	0.950721	0.896017	-0.505419	
displacement	-0.804203	0.950721	1.000000	0.932824	-0.543684	
weight	-0.831741	0.896017	0.932824	1.000000	-0.417457	
accerleration	0.420289	-0.505419	-0.543684	-0.417457	1.000000	
model year	0.579267	-0.348746	-0.370164	-0.306564	0.288137	
origin	0.563450	-0.562543	-0.609409	-0.581024	0.205873	
	model year	origin				
mpg	0.579267	0.563450				
cylinders	-0.348746	-0.562543				
displacement	-0.370164	-0.609409				
weight	-0.306564	-0.581024				
accerleration	0.288137	0.205873				
model year	1.000000	0.180662				
origin	0.180662	1.000000				

데이터 전처리와 EDA

EDA

df.plot()



데이터 전처리와 EDA

EDA

matplotlib에 많음

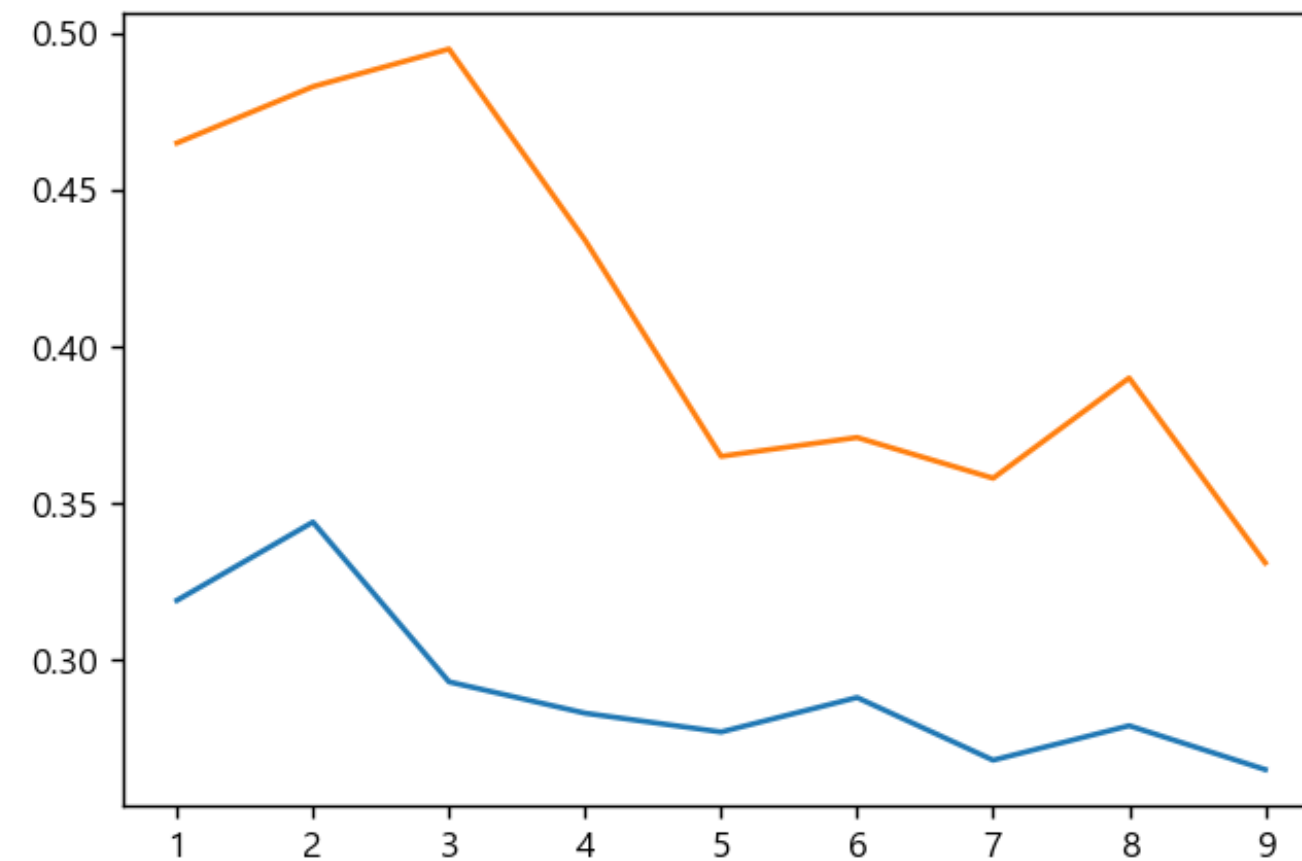
데이터의 분포를 살피는 시각화

데이터의 분포는 다음과 같이 살필 수 있습니다. 대표적인 몇 가지만 살펴보도록 하겠습니다.

- 범주형 : bar
- 수치형
 - 이산형 : bar
 - 연속형 : kdeplot, histogram
- 범주형 + 수치형 : boxplot, violinplot, etc
- 수치형 + 수치형 : scatter

데이터 전처리와 EDA

EDA



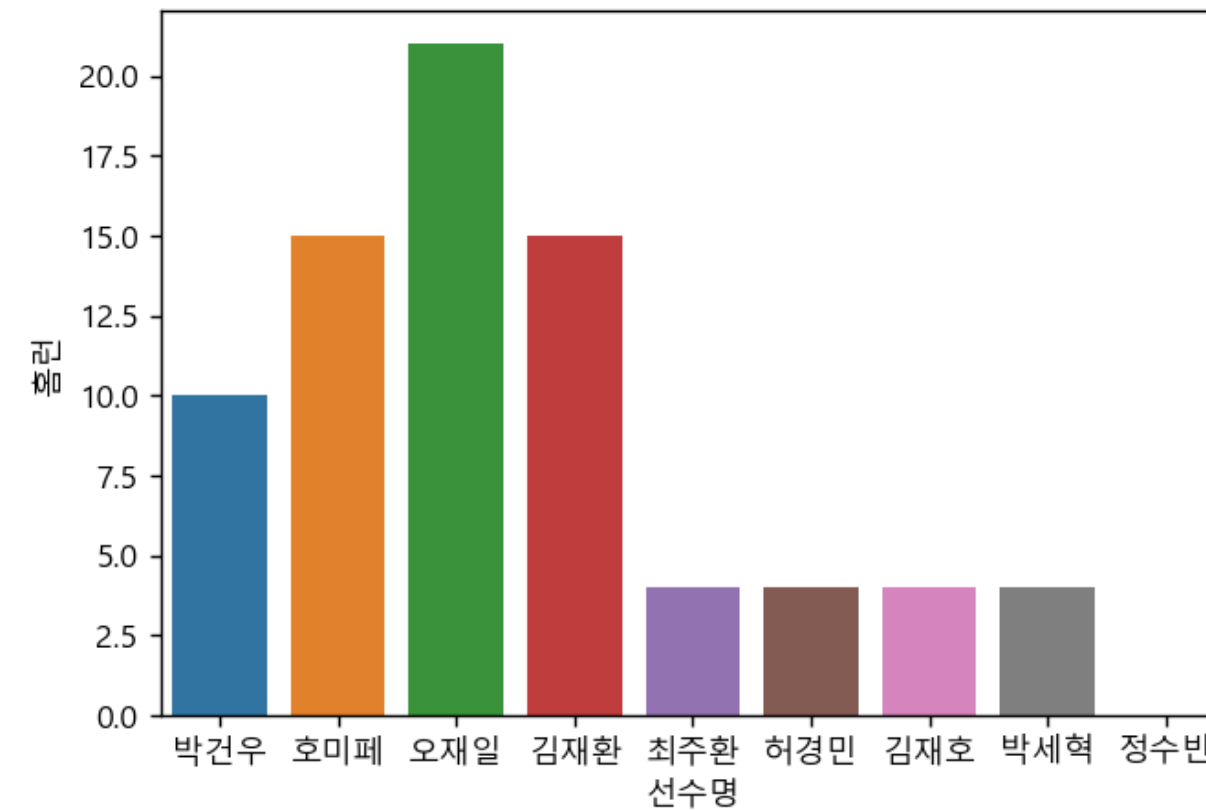
선 그래프
line plot

plt.plot / df.plot / sns.lineplot

데이터 전처리와 EDA

EDA

막대 그래프
bar plot

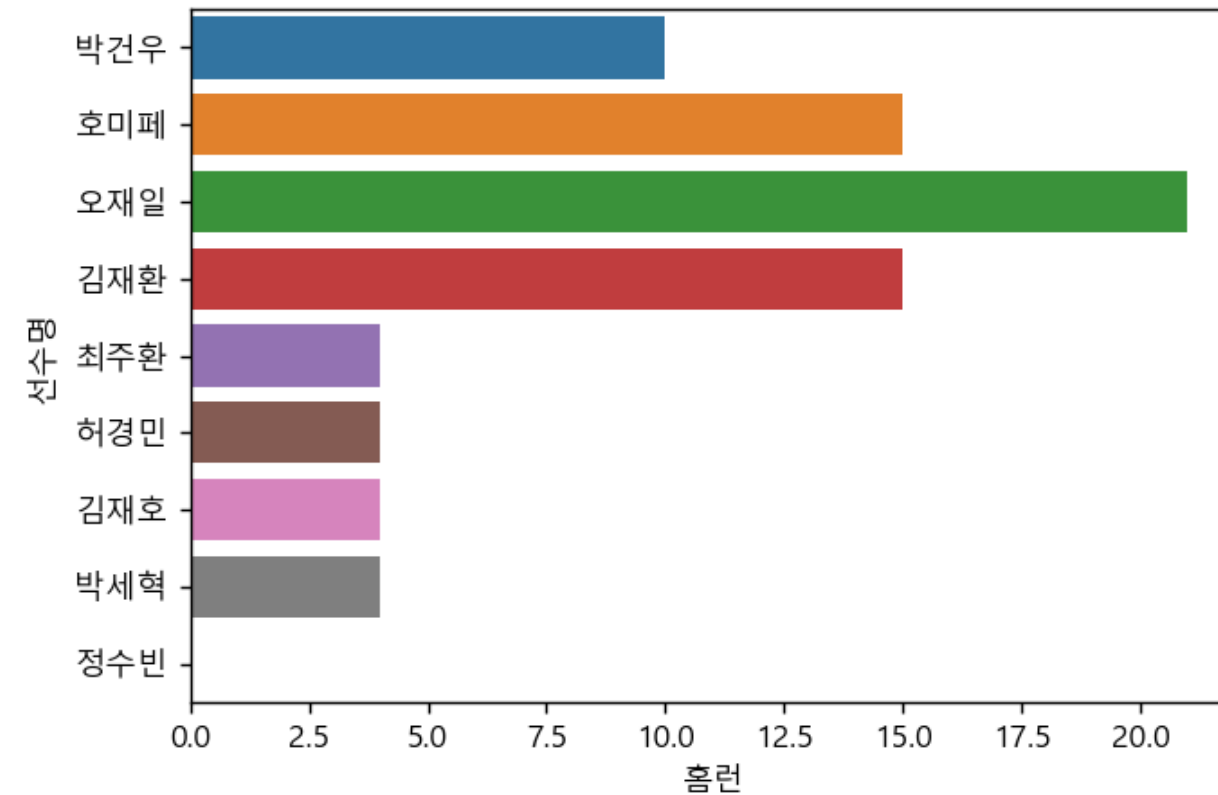


`plt.bar` / `df.plot(kind=bar)` / `sns.barplot`

데이터 전처리와 EDA

EDA

가로막대 그래프
barh plot

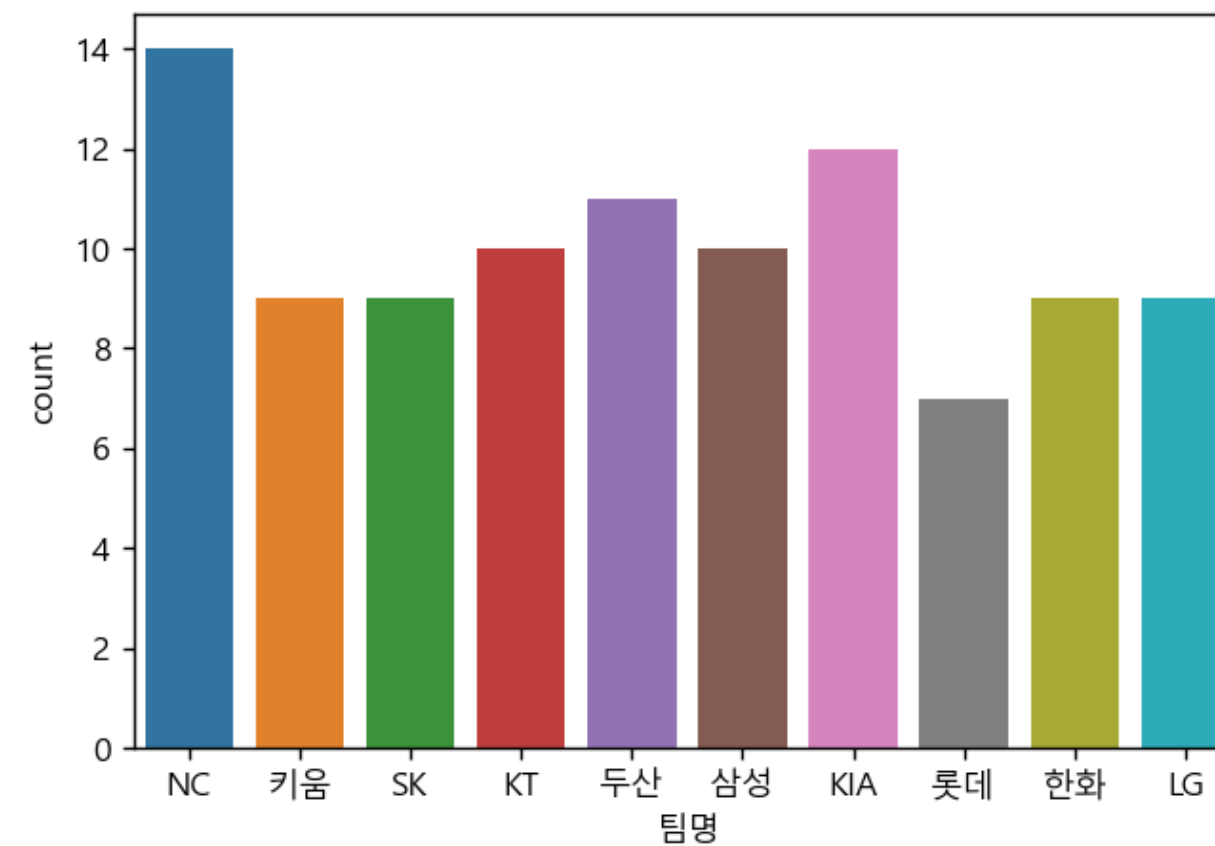


`plt.barh` / `df.plot(kind=barh)` / `sns.barplot`

데이터 전처리와 EDA

EDA

개수 그래프
count plot

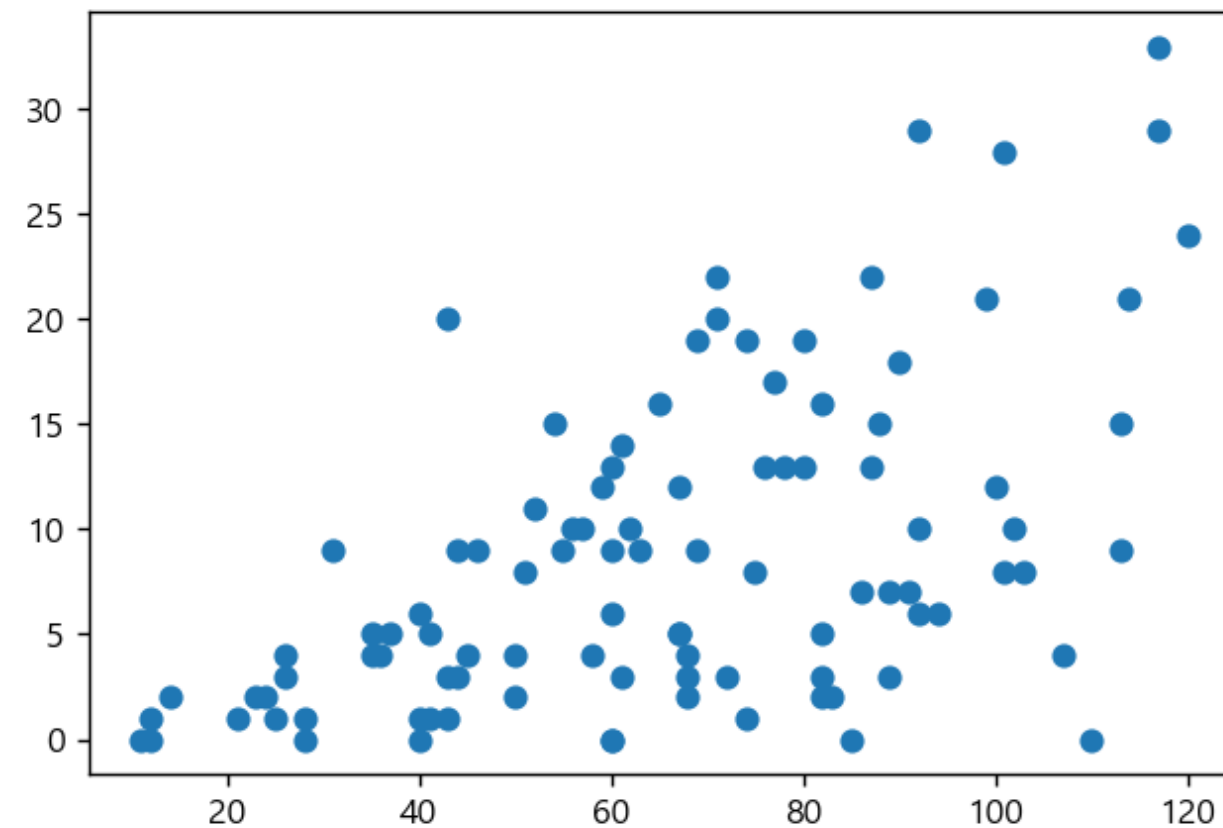


`sns.countplot`

데이터 전처리와 EDA

EDA

산점도
scatter plot

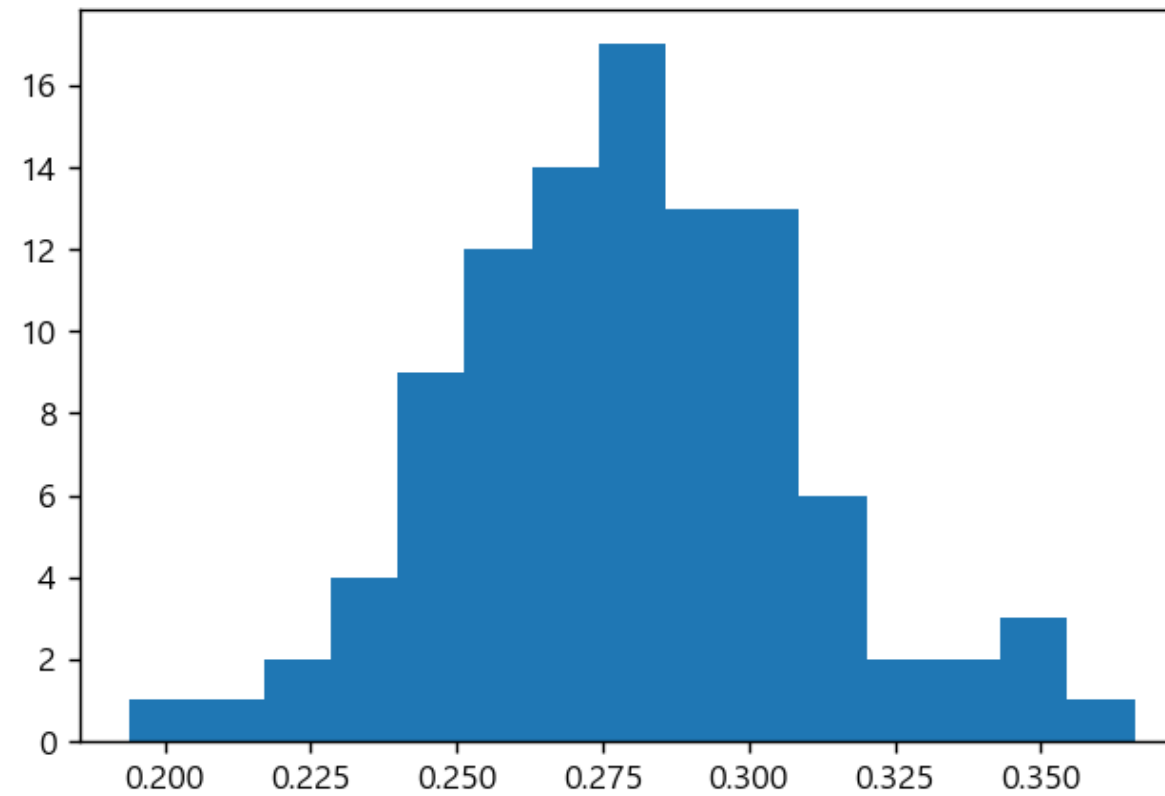


plt.scatter / df.plot(kind=scatter) / sns.scatterplot

데이터 전처리와 EDA

EDA

히스토그램
histogram



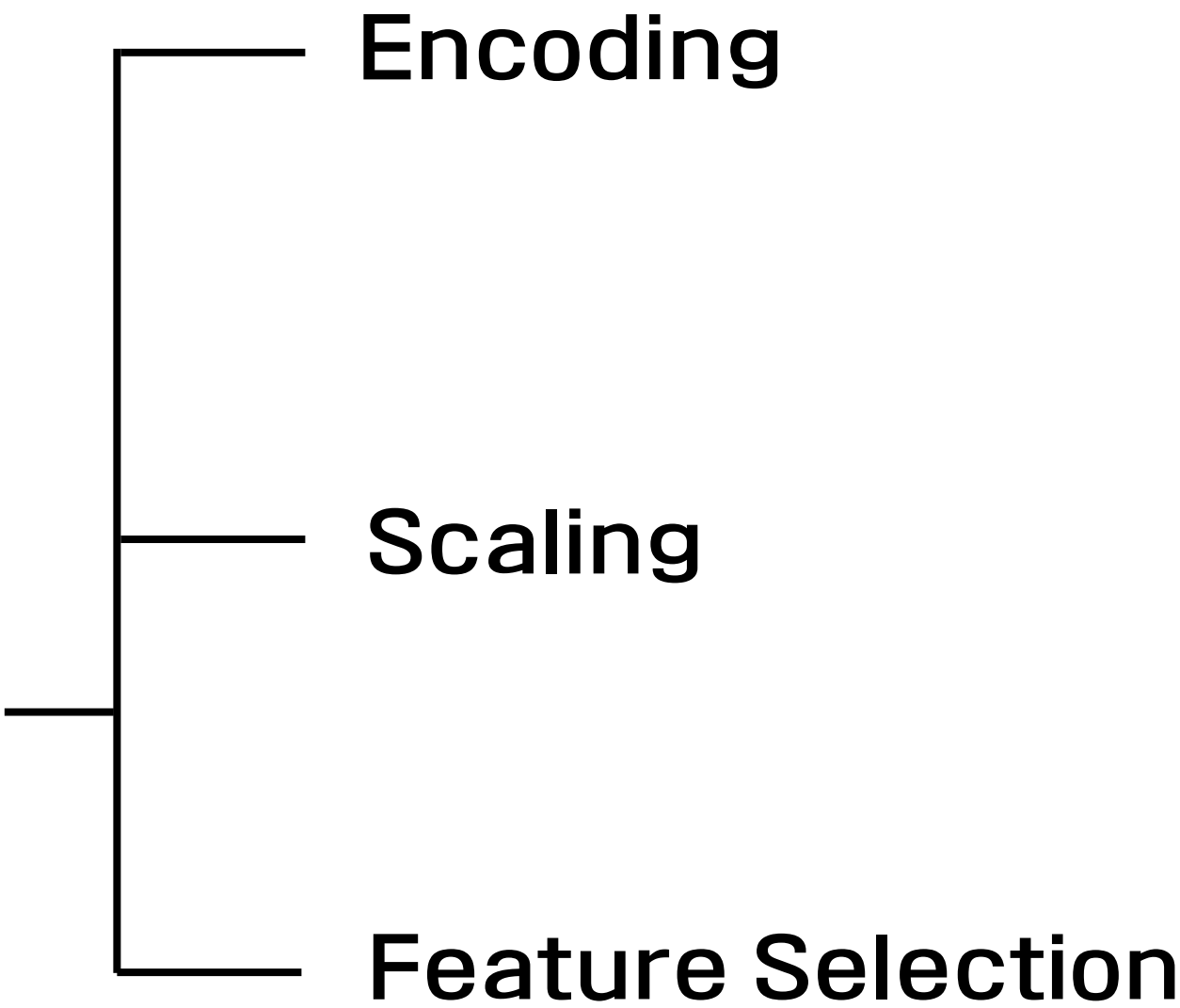
plt.hist / df.plot(kind=hist) / sns.distplot

데이터 전처리와 EDA

Data Transformation

Data Transformation

데이터를 분석하기 쉽게 만들자!



데이터 전처리와 EDA

Data Transformation

Encoding

(1) One hot Encoding

[0, 1, 0] [1, 0, 0] [0, 0, 1]

(2) Label Encoding

Red = 1 Blue = 2 Pink = 3

(3) Ordinal Encoding

초딩 = 1 중딩 = 2 고딩 = 3

데이터 전처리와 EDA

Data Transformation

(1) Min-Max Scaling

Scaling

$$\frac{x - x_{min}}{x_{max} - x_{min}}$$

데이터 전처리와 EDA

Data Transformation

(2) Maximum Absolute Scaling

Scaling

$$x_{scaled} = \frac{x}{\max(|x|)}$$

데이터 전처리와 EDA

Data Transformation

(3) Standard Scaling

Scaling

$$z = \frac{x - \mu}{\sigma}$$

데이터 전처리와 EDA

Data Transformation

(4) Robust Scaling

Scaling

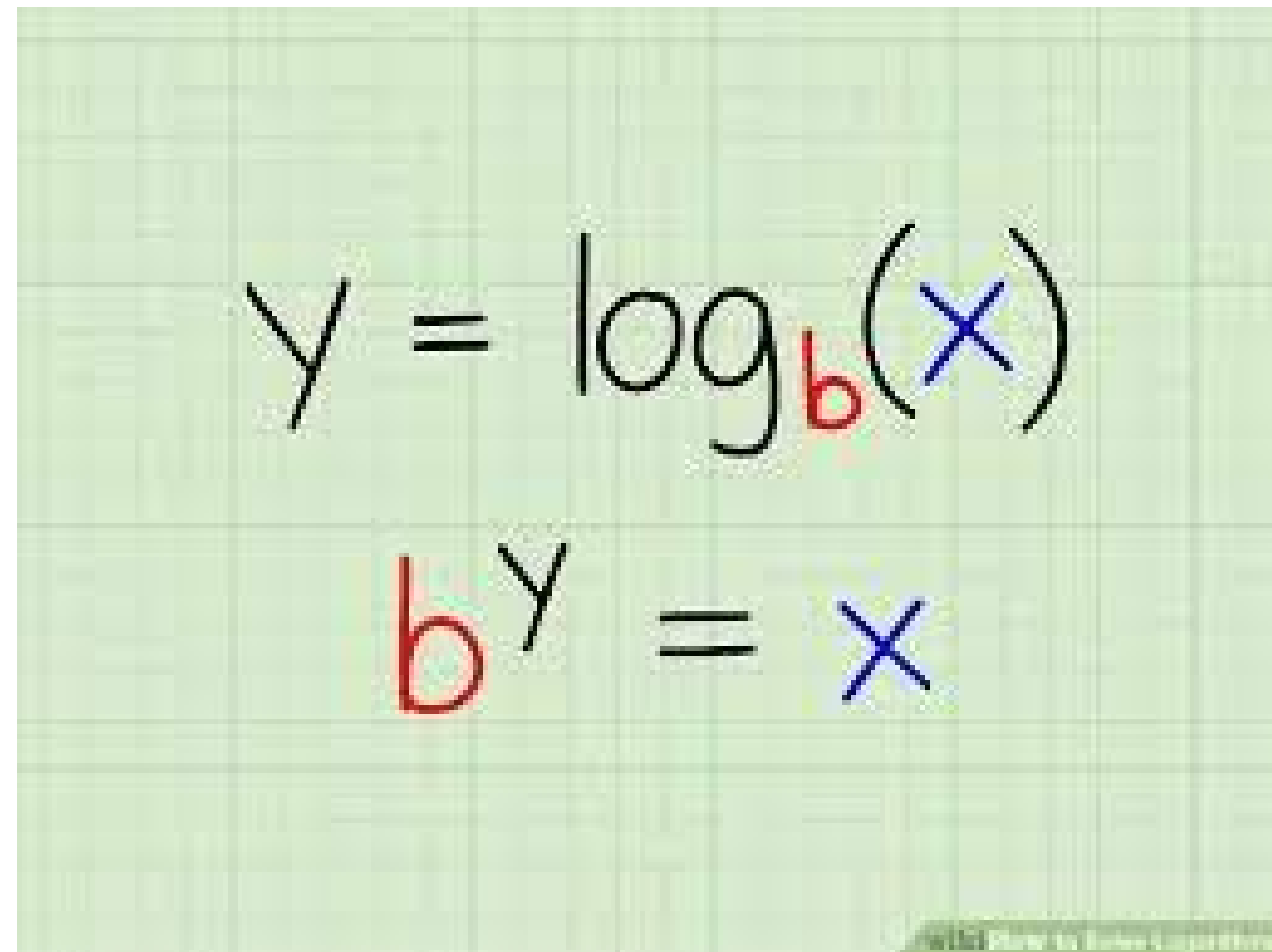
$$X_{\text{scale}} = \frac{x_i - x_{\text{med}}}{x_{75} - x_{25}}$$

데이터 전처리와 EDA

Data Transformation

(5) Log transformation

Scaling



The image shows two mathematical formulas for log transformation written in a cursive style on a green grid background. The first formula is $y = \log_b(x)$, where the base b is red and the argument x is blue. The second formula is $b^y = x$, where the base b is red and the argument x is blue. A small text label 'with Power to Exponential' is visible in the bottom right corner of the grid.

$$y = \log_b(x)$$
$$b^y = x$$

데이터 전처리와 EDA

Data Transformation

예측을 더 잘하기 위해 더 도움이 되는 열만 남기기!

**Feature
Selection**

(1) Correlation Coefficient

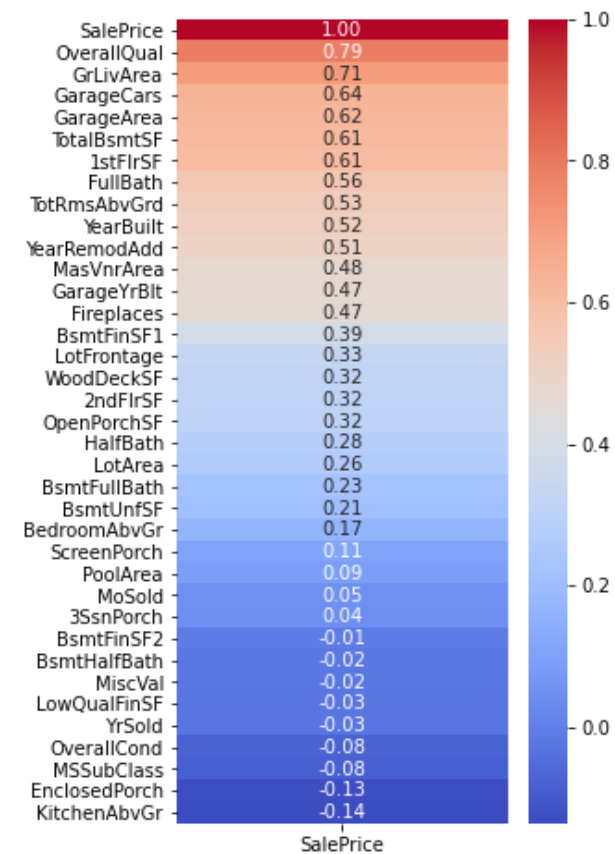
(2) Mutual Information

데이터 전처리와 EDA

Data Transformation

(1) Correlation Coefficient

Feature
Selection



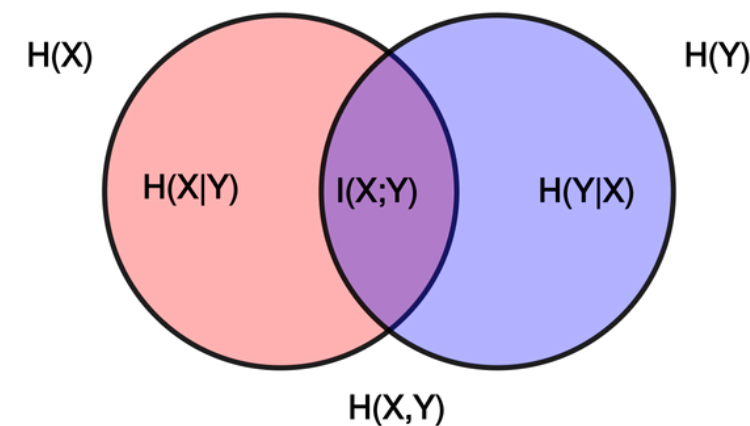
상관 계수를 보고 낮은 열을 제거함

데이터 전처리와 EDA

Data Transformation

(2) Mutual Information

**Feature
Selection**



**두 변수 간의 관계를 보고,
독립성이 강한 열을 제거함**

데이터 전처리와 EDA

Thank you