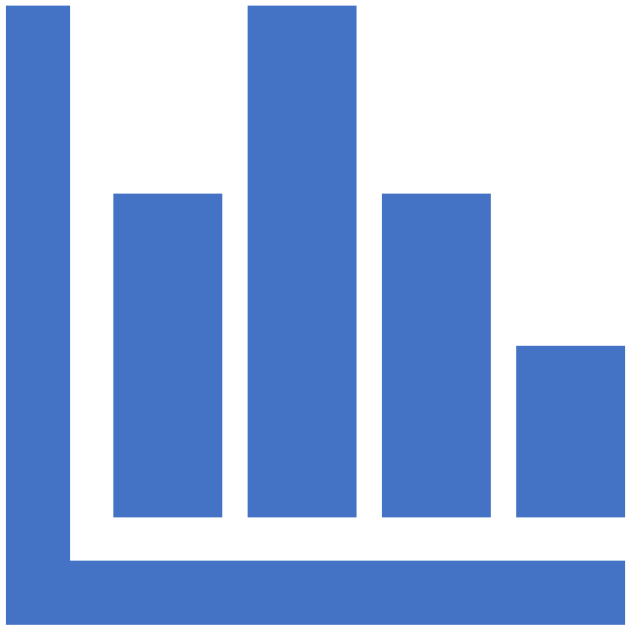


EDA

Exploratory Data Analysis

Aaron Judge

RF | Bats/Throws: R/R | 6' 7" 282LBS | Age: 31
Draft: 2013 | Rd: 1, #32, New York Yankees | Fresno State

	PA	AB	R	H	HR	SB	AVG	OBP	SLG	OPS
2021	633	550	89	158	39	6	.287	.373	.544	.917
2022	696	570	133	177	62	16	.311	.425	.686	1.111
2023	458	367	79	98	37	3	.267	.406	.613	1.019
8 Seasons	3,619	3,005	614	846	257	43	.282	.396	.586	.982

Player Apps

illustrator
Pitch Highlighter
Player Comparison
Player Similarity
Random Video

Shifts
Swing Take Profile
Transaction History
Zone Swing Profile

2023 MLB Percentile Rankings

savant

Value

	POOR — AVERAGE — GREAT	
Batting Run Value	97	42
Baserunning Run Value	15	-2
Fielding Run Value	55	1

Batting

xwOBA	100	.460
xBA	90	.284
xSLG	100	.711
Avg Exit Velocity	100	97.5
Barrel %	100	27.1
Hard-Hit %	100	63.5
Sweet-Spot %	68	35.8
Chase %	93	19.6
Whiff %	2	36.6
K %	16	28.4
BB %	100	19.2

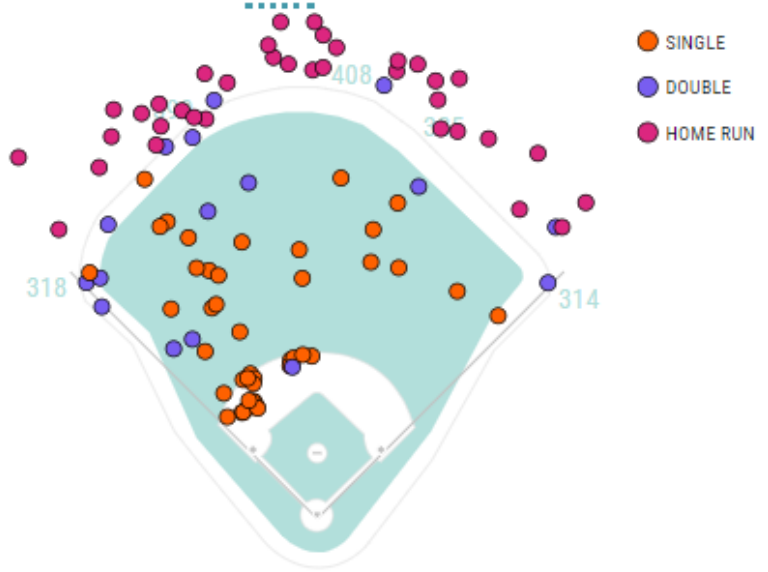
Fielding

Range (OAA)	NOT QUALIFIED	0
Arm Value	75	1
Arm Strength	85	90.4

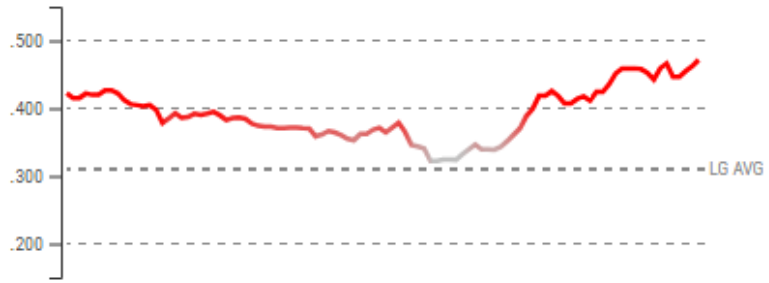
Running

Sprint Speed	32	26.6

2023 Hits Spray Chart



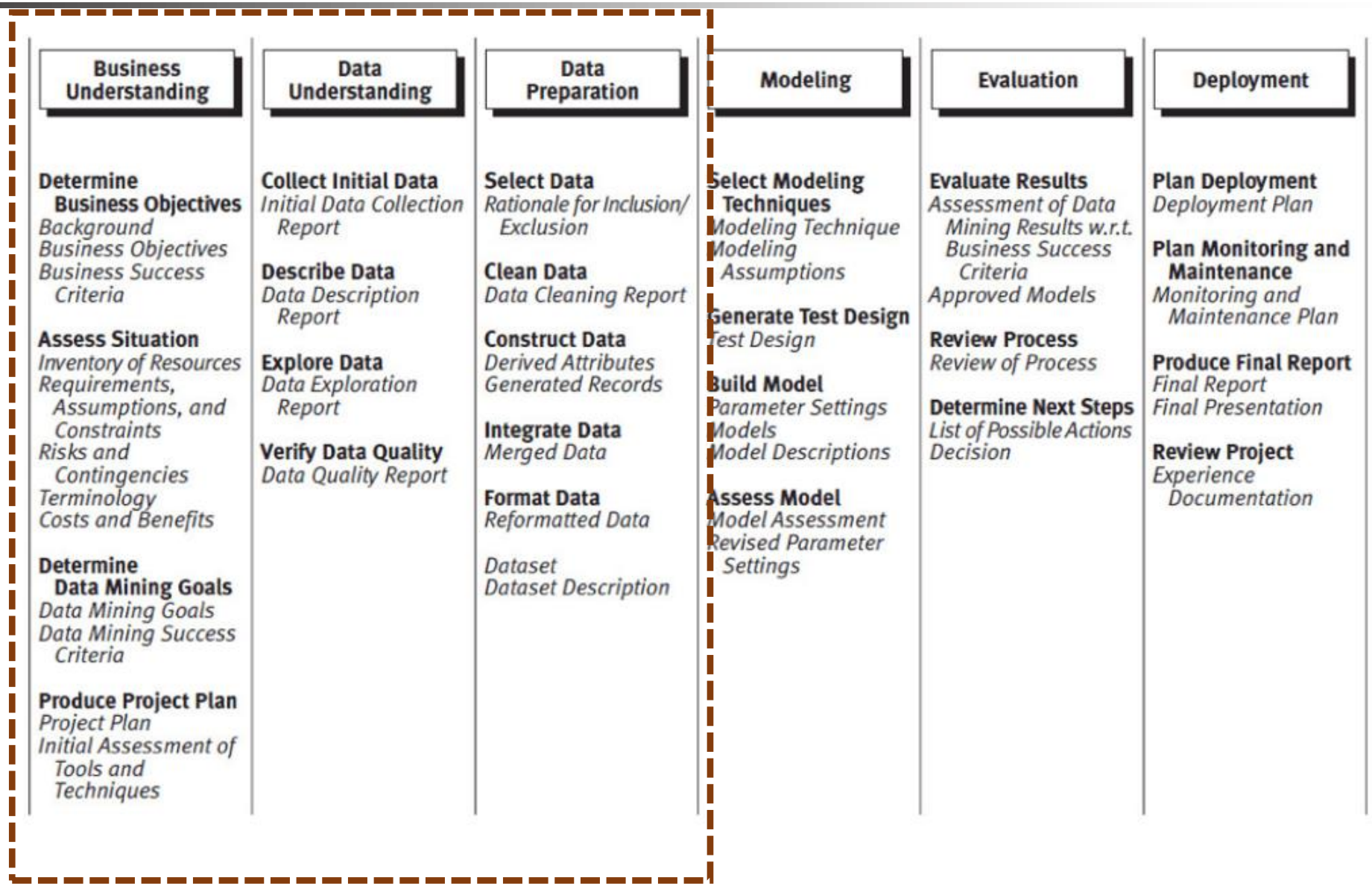
100 PAs Rolling xwOBA



머니볼 세이버 매트릭스

https://www.youtube.com/watch?v=zjPKa922e4w

EDA



Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	**Select Data** *Rationale for Inclusion/ Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data* *Dataset* *Dataset Description*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Descriptions* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

CRISP-DM tasks in **bold**, and outcomes in *italic* (table from CRISP-DM Guide)

EDA

Exploratory Data Analysis: 데이터 탐색적 분석

: 데이터를 시각화하고 통계적으로 분석하여 숨겨진 패턴, 이상치, 관계성을 발견하는 과정

- **데이터에 대해 알아보는 것!!**
  - **시각화 도구를 통해 패턴 발견**
  - **데이터 특이성 확인**
  - **통계 기법을 통한 가설 검정 과정**

EDA

Exploratory Data Analysis: 데이터 탐색적 분석

- **왜?**
  - 데이터 수집 단계에서 실수로 입력한 부분이 있는지?
    (Detection of mistakes)
  - 데이터에 대한 가정이 맞는지?
    (Checking of assumptions)
  - 데이터 유형에 맞는 적절한 모델 선택하기위해서
    (Preliminary selection of appropriate models)
  - 설명변수들 사이의 관계 파악하기위해서
    (determining relationships among explanatory variables)
  - 설명변수와 목적변수 사이의 관계 파악하기위해서
    (assessing the direction and rough size of relationships between explanatory and outcome variables)

EDA

EDA Overview w.Practice

- **실습**
  2.00.EDA.OverView_w.Practice.ipynb

- **기본 통계 분석 : 데이터 크기, 타입, 기초 통계량, 결측값 확인**
- **데이터 시각화 : 히스토그램, 박스플롯, 산점도, 상관관계 분석**
- **이상치 탐지 : 박스플롯, IQR, Z-score 활용**
- **변수 간 관계 분석 : 상관관계 분석**
- **결측값 처리 : 평균/중앙값 대체**
- 데이터 변환 : 정규화, 로그 변환, 범주형 변수 인코딩

.

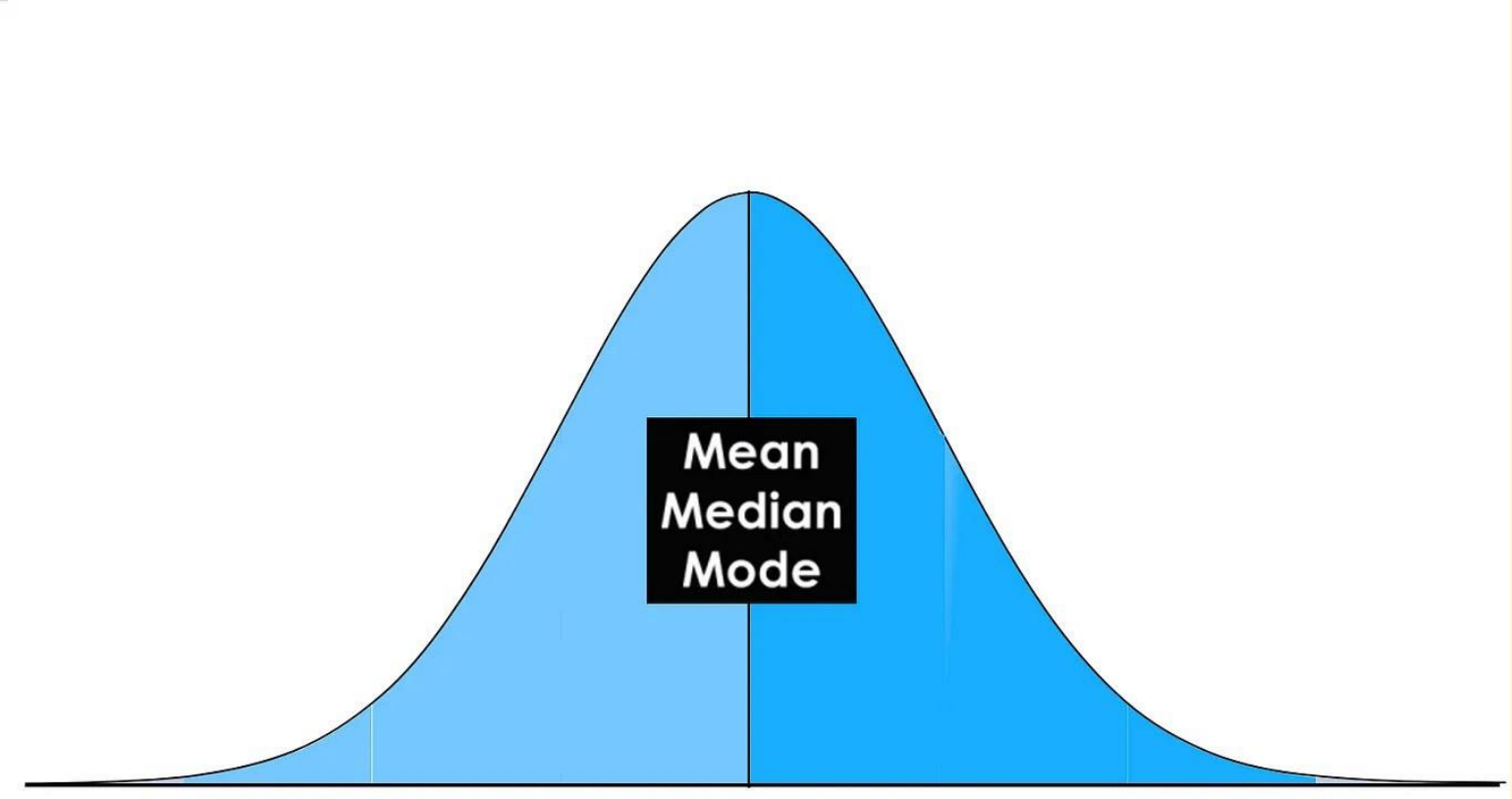
EDA

평균, 표준편차, Z-Score, 정규분포, CLT

EDA

평균, 표준편차, Z-Score, 정규분포, CLT

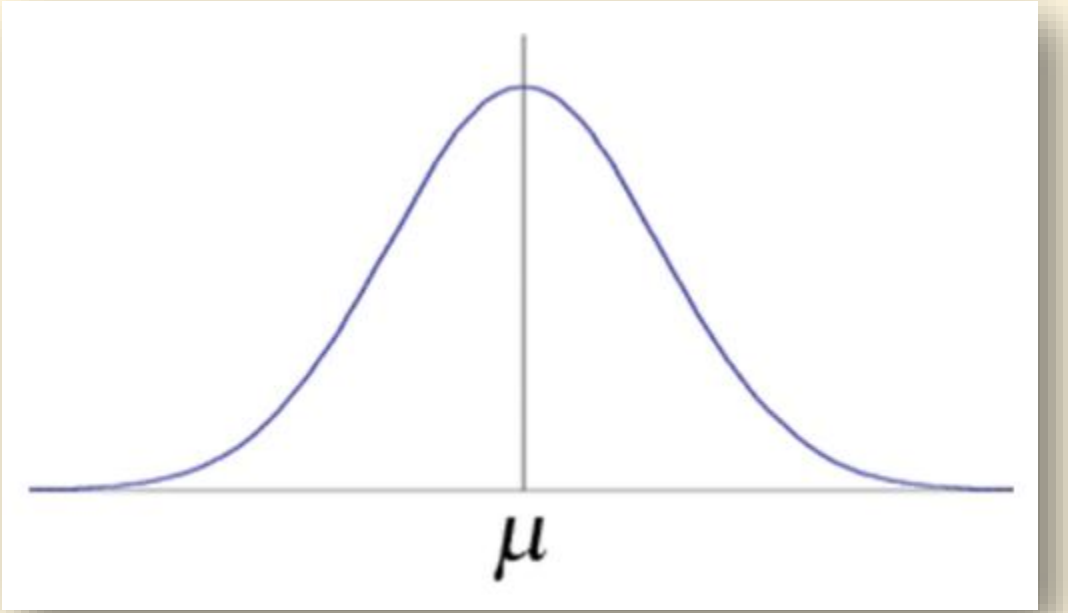
: 표본의 데이터를 모두 더한 후 표본의 데이터 개수 n으로 나눈 것 - wiki



EDA

평균, 표준편차, Z-Score, **정규분포**, CLT

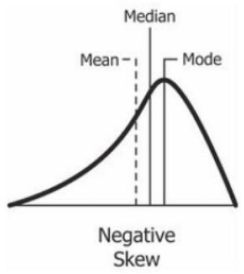
: 평균과 표준편차가 주어져 있을 때 엔트로피를 최대화하는 분포이다.
정규분포곡선은 좌우 대칭이며 하나의 꼭지를 가진다- wiki

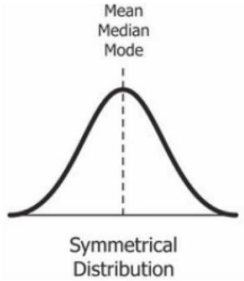


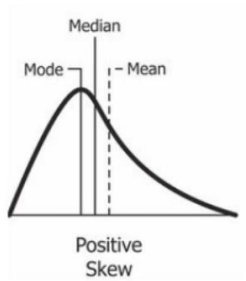
EDA

평균, 표준편차, Z-Score, **정규분포**, CLT

왜도(Skewness) : 분포의 좌우 비대칭성







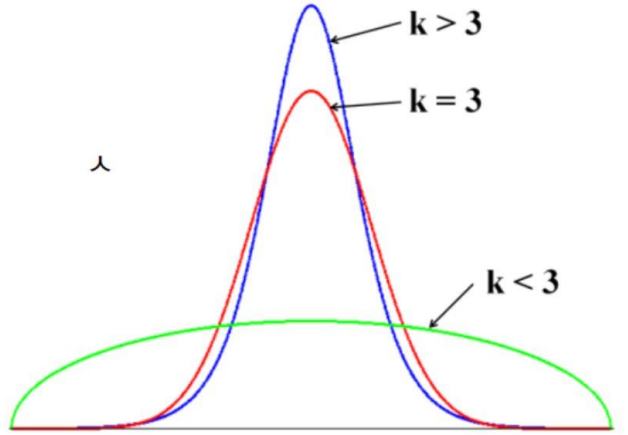
Mean<Median<Mode
Negative Skew

Mean=Median=Mode
Zero Skew

Mean>Median>Mode
Positive Skew

https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa

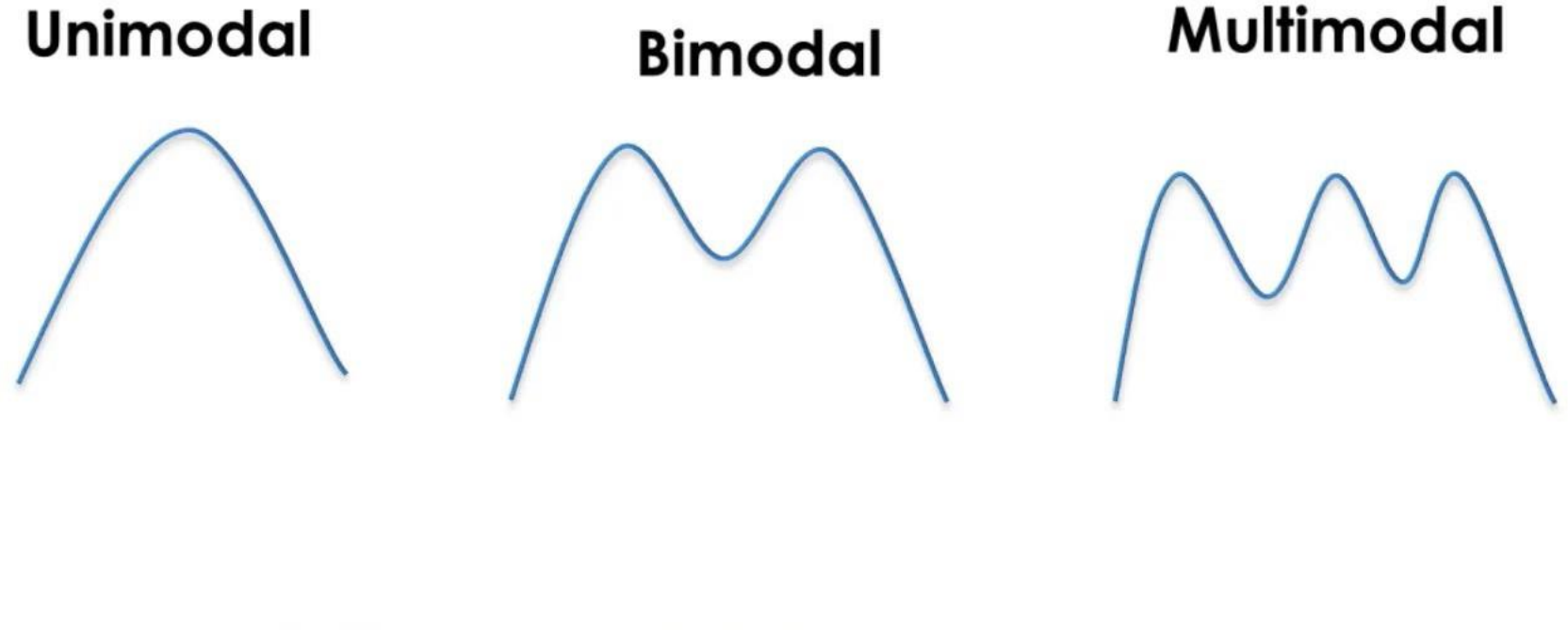
첨도(kurtosis) : 분포의 뾰족한 정도



https://www.researchgate.net/figure/Illustration-of-the-skewness-and-the-kurtosis_fig1_273463074

EDA

평균, 표준편차, Z-Score, **정규분포**, CLT



EDA

$$S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

평균, **표준편차**, Z-Score, 정규분포, CLT

통계집단의 분산의 정도 또는 자료의 산포도를 나타내는 수치로, 분산의 음이 아닌 제곱근 즉,
분산을 제곱근한 것으로 정의 - wiki

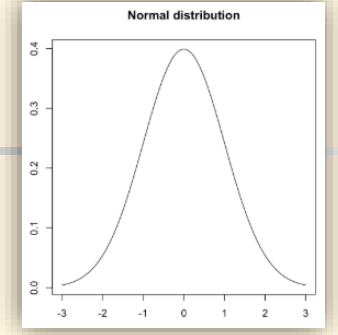
데이터가 평균을 기준으로 얼마나 퍼져 있는지를 나타내는 통계적 지표

반	평균 점수	표준편차
A반	80점	5
B반	80점	15

표준편차 작음 (성적 편차가 적음)

표준편차 큼 (성적 편차가 큼)

EDA

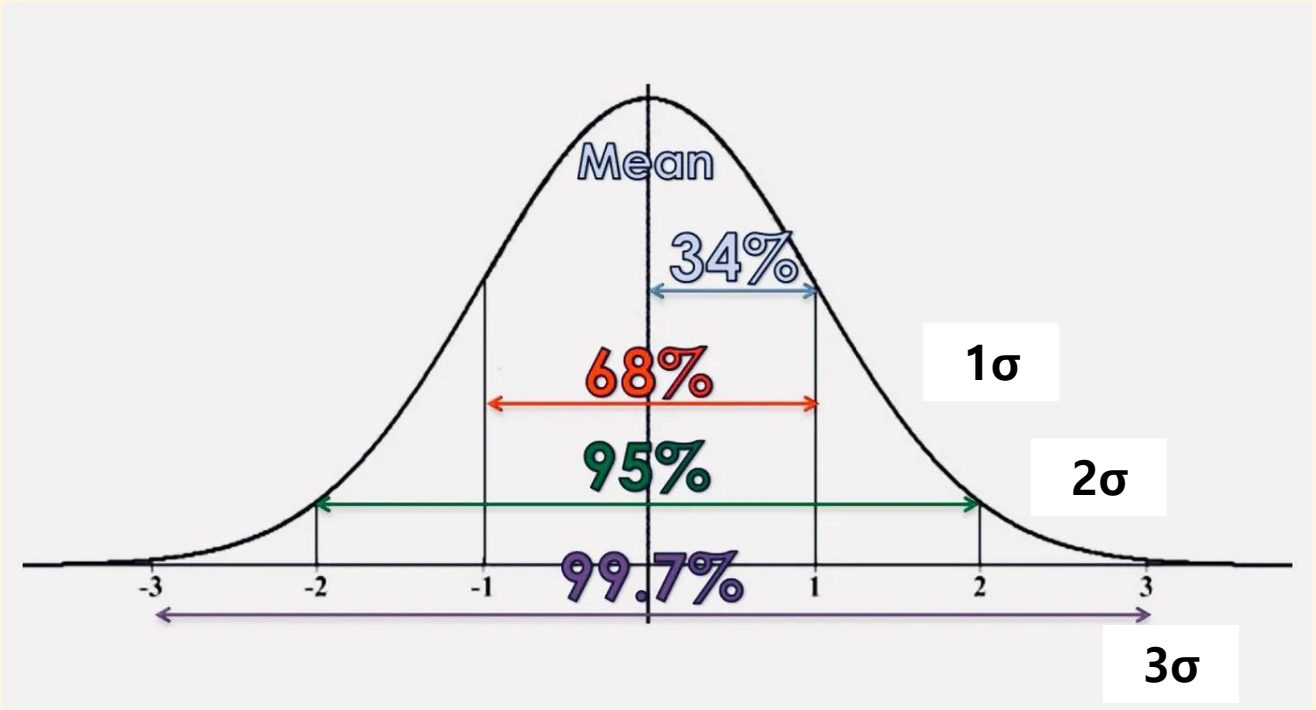


$$S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

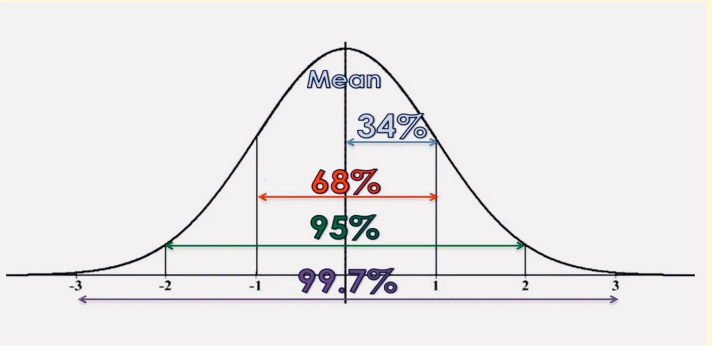
평균, **표준편차**, Z-Score, **정규분포**, CLT

정규분포 안에서 표준편차(σ)에 따라 데이터가 어떻게 분포하는지 확인

```
std_dev = np.std(data) # 표준편차 (모집단)
    = np.std(data, ddof=1) # ddof=1: (표본)
```



EDA



평균, **표준편차**, Z-Score, **정규분포**, CLT

ex, 평균 키 170cm, 표준편차가 5cm

- **Q1. 중심을 기점으로 68% 학생의 키 범위는?** **1σ**
- **Q1. 중심을 기점으로 95% 학생의 키 범위는?** **2σ**
- **Q1. 중심을 기점으로 99.7% 학생의 키 범위는?** **3σ**

EDA

평균, 표준편차, **Z-Score**, 정규분포, CLT

특정 값이 전체 데이터에서 차지하는 백분위 위치 파악 가능

	age	z_score
0	22.0	-0.530005
1	38.0	0.571430
2	26.0	-0.254646
3	35.0	0.364911
4	35.0	0.364911

```
1 # 백분위 수
2 from scipy.stats import norm
3
4 percentile = norm.cdf(-0.53) * 100
5 print(percentile)
```

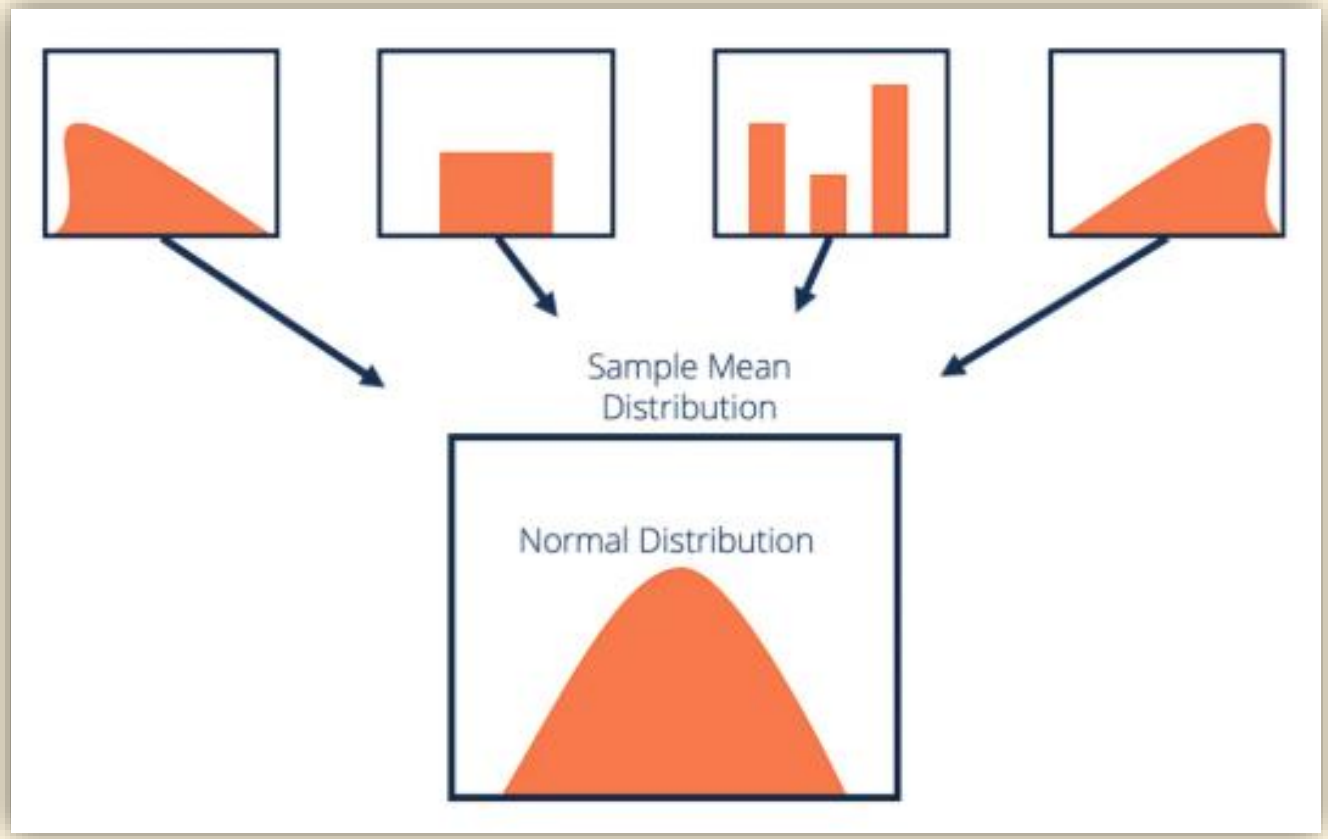
```
29.805596539487645
```

```
# 분석 변수 (age)
num_var = "age"

# 평균, 표준편차
mean_value = df[num_var].mean()
std_value = df[num_var].std()

# Z-스코어
df["z_score"] = (df[num_var] - mean_value) / std_value

# 상위 5개 데이터 출력
from IPython.display import display
display(df[[num_var, "z_score"]].head(5))

# 백분위 수 계산
from scipy.stats import norm
percentile = norm.cdf(-0.53) * 100
```

EDA

평균, 표준편차, Z-Score, 정규분포, **CLT**

중심 극한 정리 : 동일한 확률분포를 가진 독립 확률 변수 n개의 평균의 분포는 n이 적당히 크다면 정규분포에 가까워진다는 정리 - wiki

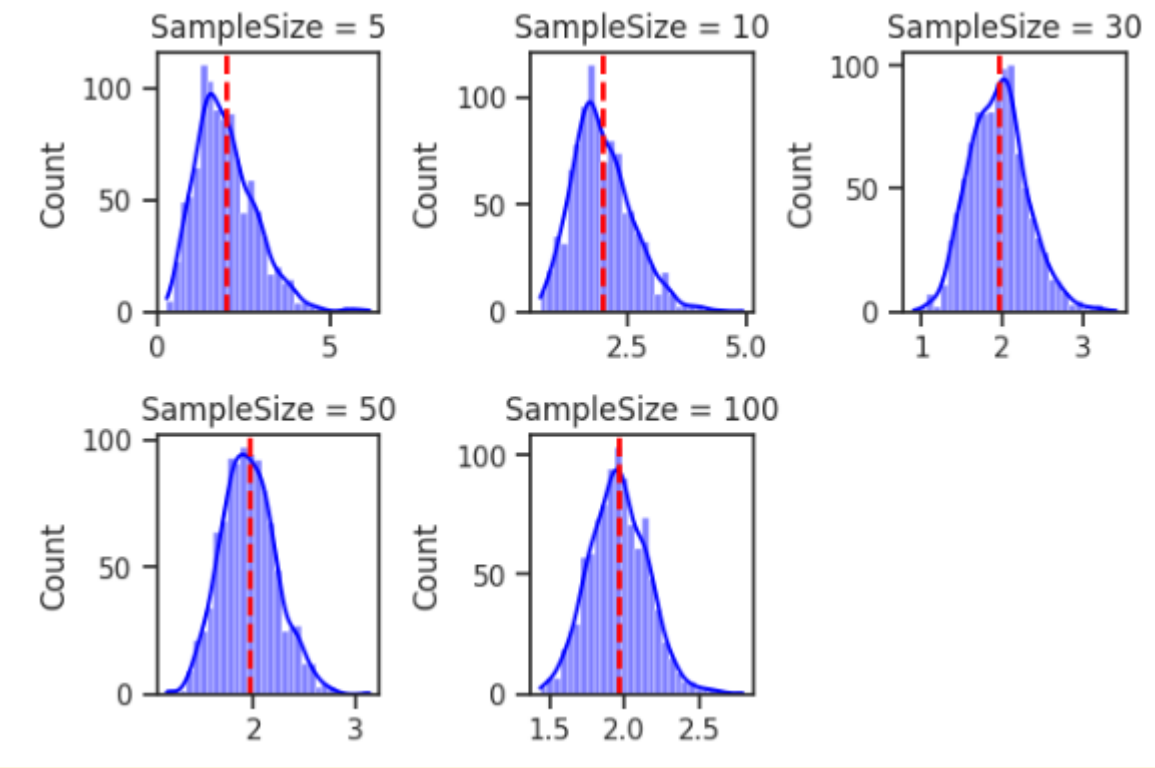


https://corporatefinanceinstitute.com/resources/data-science/central-limit-theorem/

EDA

평균, 표준편차, Z-Score, 정규분포, **CLT**

동일한 확률분포를 가진 독립 확률 변수 n개의 평균의 분포는
n이 적당히 크다면 정규분포에 가까워진다는 정리 - wiki




```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# 랜덤 시드를 고정 (재현 가능성을 위해)
np.random.seed(42)

# 1. 모집단을 비정규분포(지수분포)에서 샘플링
population = np.random.exponential(scale=2, size=10000)  # 지수분포 데이터 생성

# 2. 표본 크기 설정 (작은 값부터 증가시키며 실험)
sample_sizes = [5, 10, 30, 50, 100]  # 다양한 표본 크기

plt.figure(figsize=(12, 8))  # 그래프 크기 설정

for i, n in enumerate(sample_sizes, 1):
    sample_means = []  # 표본 평균 저장할 리스트

    # 여러 개의 표본을 추출하여 평균을 계산
    for _ in range(1000):  # 1000개의 표본 생성
        sample = np.random.choice(population, size=n, replace=False)  # 표본 추출
        sample_means.append(np.mean(sample))  # 표본 평균 계산 후 저장

    # 3. 표본 평균들의 분포 시각화
    plt.subplot(2, 3, i)  # 서브플롯 생성
    sns.histplot(sample_means, bins=30, kde=True, color='blue')  # KDE 곡선 추가
    plt.axvline(x=np.mean(population), color='red', linestyle='dashed', linewidth=2)  # 모집단 평균 표시
    plt.title(f"sample size = {n}")

# 4. 전체 그래프 제목 및 출력
plt.suptitle("Central Limit Theorem: Variation in the distribution of sample means with sample size", fontsize=15)
plt.tight_layout()
plt.show()
```

EDA

Exploratory Data Analysis: 데이터 탐색적 분석

변수 수에 따른 구분

	UNIVARIATE	MUTIVARIATE

Graphical	• **Quantitative Variable**:
	  • Histogram
	  • Boxplots
	  • Normal QQ-plot
	• **Categorical Variable**: Bar Charts
	• Time data – Line Plot
		• **One Categorical and One Quantitative Variable:**
		  • Side-by-side Boxplots
		• **Two or More Categorical Variables:**
		  • Grouped Bar Chart
		• **Two or More Quantitative Variables:**
		  • Scatterplot
		  • Correlation Heatmap
		  • Pairplot
		• Missing Data Detection

Non-Graphical	• **Categorical Variable:** tabular representation of frequency (*or relative frequency*)
	• **Quantitative Variable:**
	  • Location (*mean, median)*
	  • Spread (*IQR, std dev, range*)
	  • Modality (*mode*)
	  • Shape (*skewness, kurtosis*)
	  • Outliers
	• Missing Data Detection
		• **One Categorical and One Quantitative Variable:** ***standard univariate nongraphical statistics for the quantitative variables separately for each level of the categorical variable.***
		  • Mean
		  • Median
		  • Range and Spread measures
		• **Two or More Quantitative Variables:**
		  • Correlation
		  • Covariance
		  • Descriptive stat per
		• Missing Data Detection

EDA

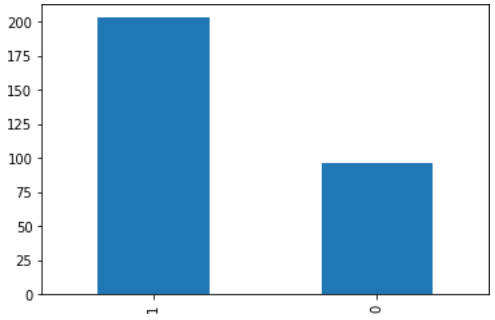
Exploratory Data Analysis: 데이터 탐색적 분석

변수 수에 따른 구분

	단변량	다변량
그래픽 방법	> 정량변수: - 히스토그램 - 상자 그림 - 일반 QQ 플롯 > 범주형 변수: - 막대 차트 > 시간 데이터: - 라인 플롯	> 하나의 범주형 변수와 하나의 양적 변수: - 나란히 상자 그림 > 둘 이상의 범주형 변수: - 그룹화된 막대 차트 > 둘 이상의 정량변수: - 산포도 - 상관관계 히트맵 - 쌍도표 > 누락된 데이터 감지
비그래픽 방법	> 범주형 변수: - 빈도(또는 상대 빈도)를 표로 표현 > 정량변수: - 위치(평균, 중앙값) - 확산(IQR, 표준 개발, 범위) - 양식 (모드) - 모양(왜도, 첨도) - 이상치 > 누락된 데이터 감지	> 하나의 범주형 변수와 하나의 양적 변수: - 범주형 변수의 각 수준(평균, 중앙값, 범위 및 확산 측정값)에 대해 개별적으로 양적 변수에 대한 표준 일변량 비그래픽 통계 > 둘 이상의 정량변수: - 상관관계 - 공분산 - 변수별 설명 통계 - 누락된 데이터 감지

EDA

단변량 Univariate 변수의 특성 파악



	age
count	299.000000
mean	54.521739
std	9.030264
min	29.000000
25%	48.000000
50%	56.000000
75%	61.000000
max	77.000000

- **범주형** 데이터 (Categorical Variable)
  - 전체량 + 빈도 + 구성비율

- **수치형** 데이터 (Quantitative Variable)
  - 값의 범위: 최소? 최대? 대부분 중간?
  - 모양: 오른쪽 or 왼쪽 치우침?
  - Outlier detection: 이상치?
  - Missing data 여부 확인: 결측치?

EDA

단변량 Univariate 변수의 특성 파악

- **수치형 변수:**
  - **그래픽 방법:** 히스토그램, 박스플롯, Q-Q 플롯 등
    분포, 중앙값, 범위를 시각적으로 분석
  - **비그래픽 방법:** 평균, 중앙값, 최빈값, 분산, 표준편차 등
    변수의 특성을 수치적으로 파악
- **범주형 변수**
  - **그래픽 방법:** 막대그래프 등
    각 범주의 빈도를 시각적으로 표현.
  - **비그래픽 방법:** 각 범주의 빈도, 상대 빈도의 테이블 표현
    범주별 특성 분석

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False

EDA

단변량 Univariate 변수의 특성 파악

- **수치형 변수:**
  - **그래픽 방법:** 히스토그램, 박스플롯, Q-Q 플롯 등
    분포, 중앙값, 범위를 시각적으로 분석
  - **비그래픽 방법:** 평균, 중앙값, 최빈값, 분산, 표준편차 등
    변수의 특성을 수치적으로 파악
- **범주형 변수**
  - **그래픽 방법:** 막대그래프 등
    각 범주의 빈도를 시각적으로 표현.
  - **비그래픽 방법:** 각 범주의 빈도, 상대 빈도의 테이블 표현
    범주별 특성 분석

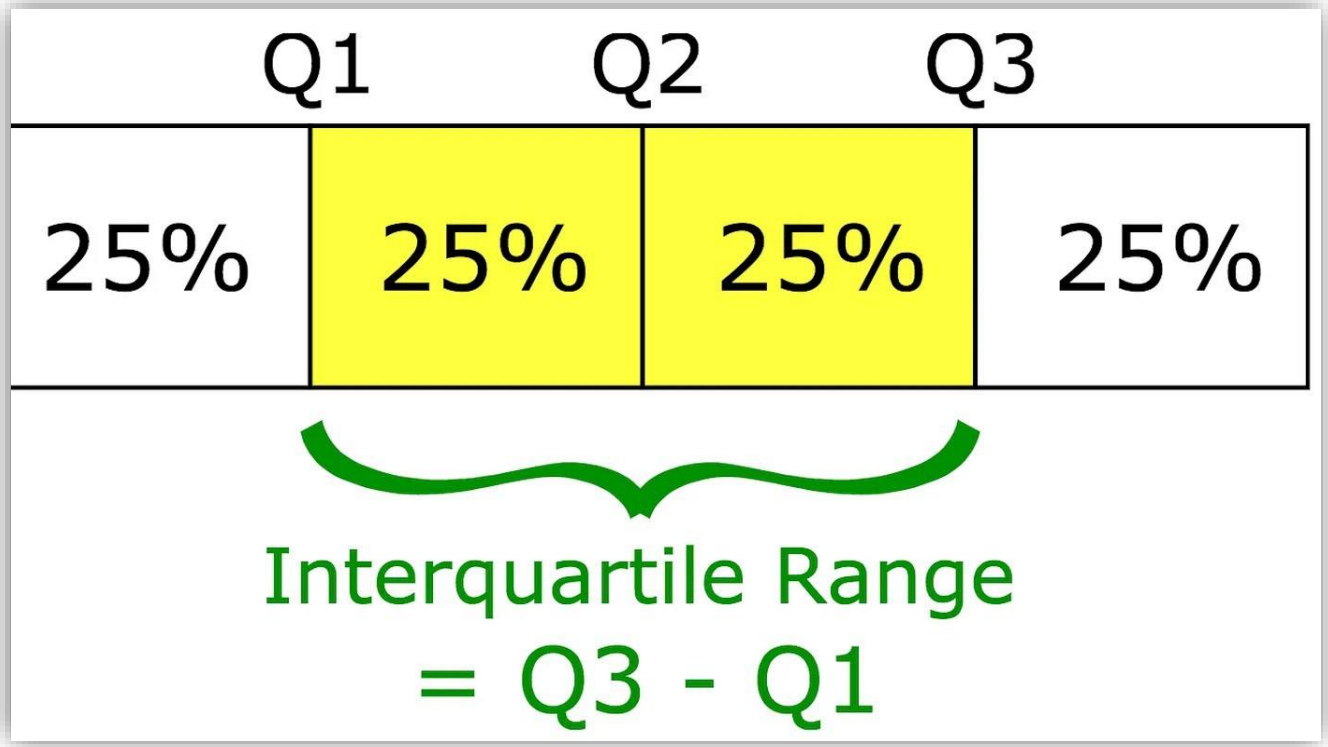
- **실습**

  2.01. 2.01.EDA.Univariate.ipynb

EDA

Univariate non-graphical EDA
단변량 - 그래프 X

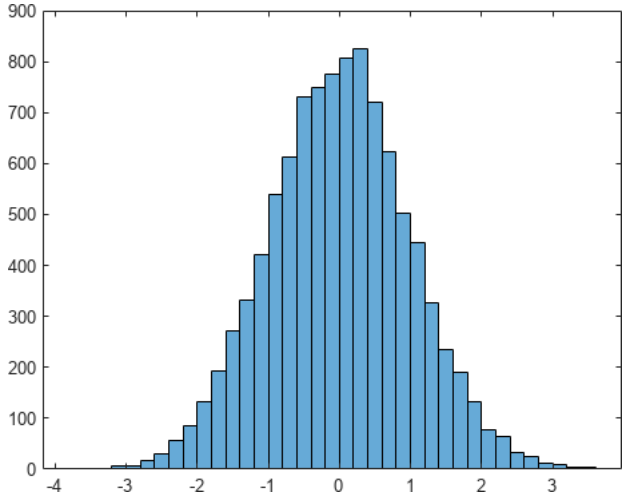
4분위수

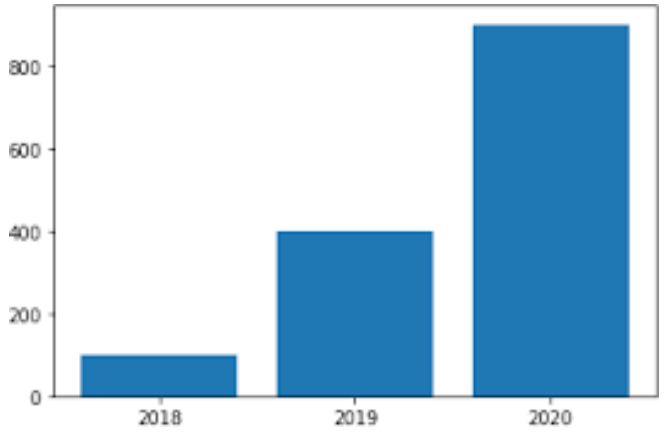


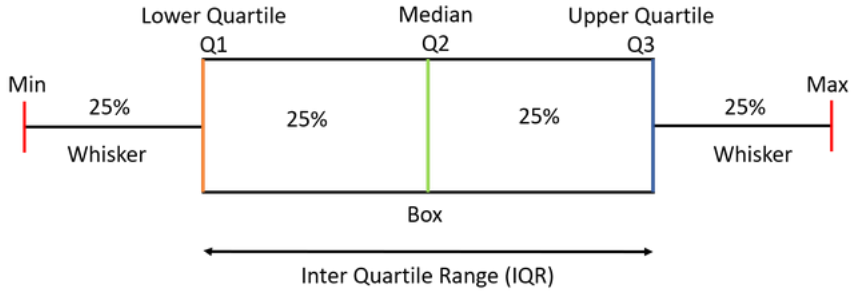
EDA

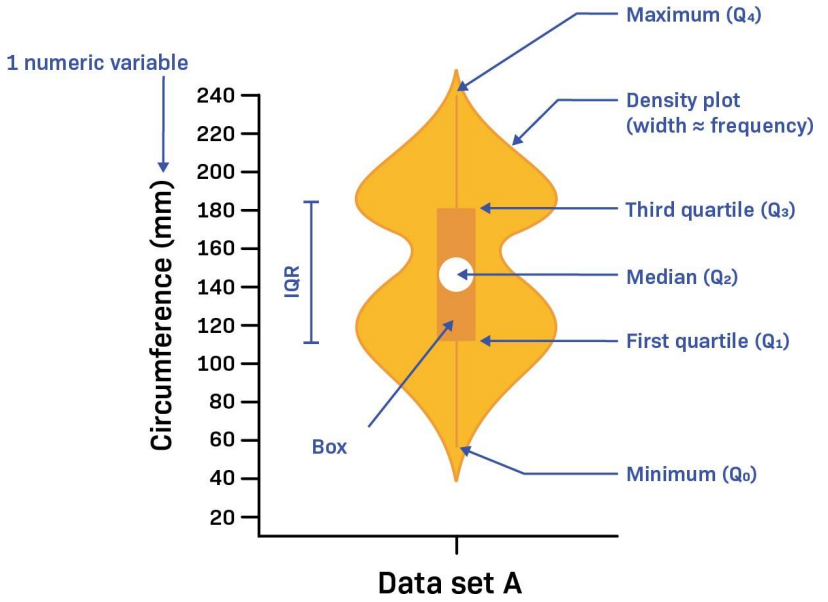
Univariate graphical EDA
단변량 - 그래프 0

- 히스토그램
- 막대그래프
- Box plot
- Violin plot
- QQ-plot





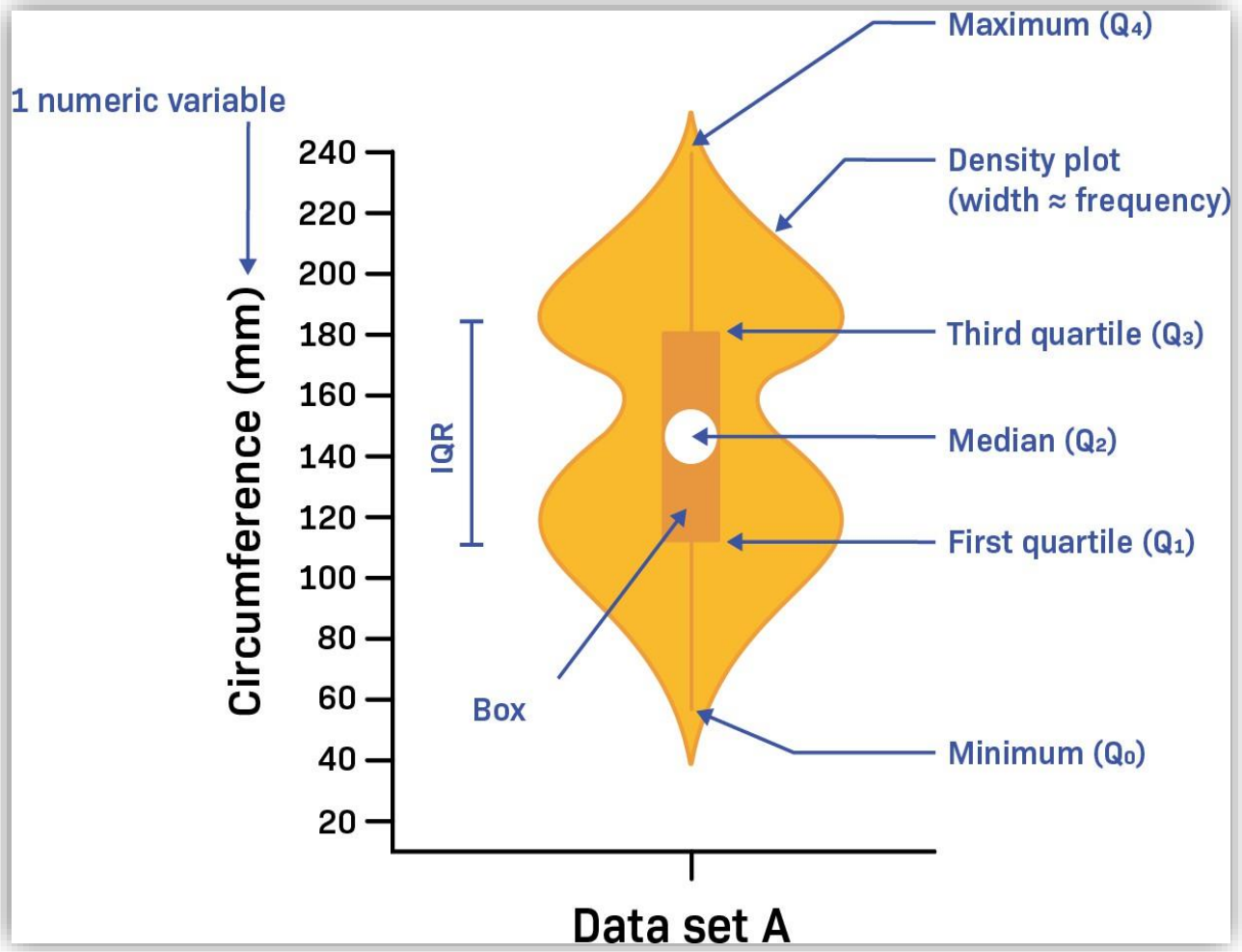




EDA

Univariate graphical EDA
단변량 - 그래프 0

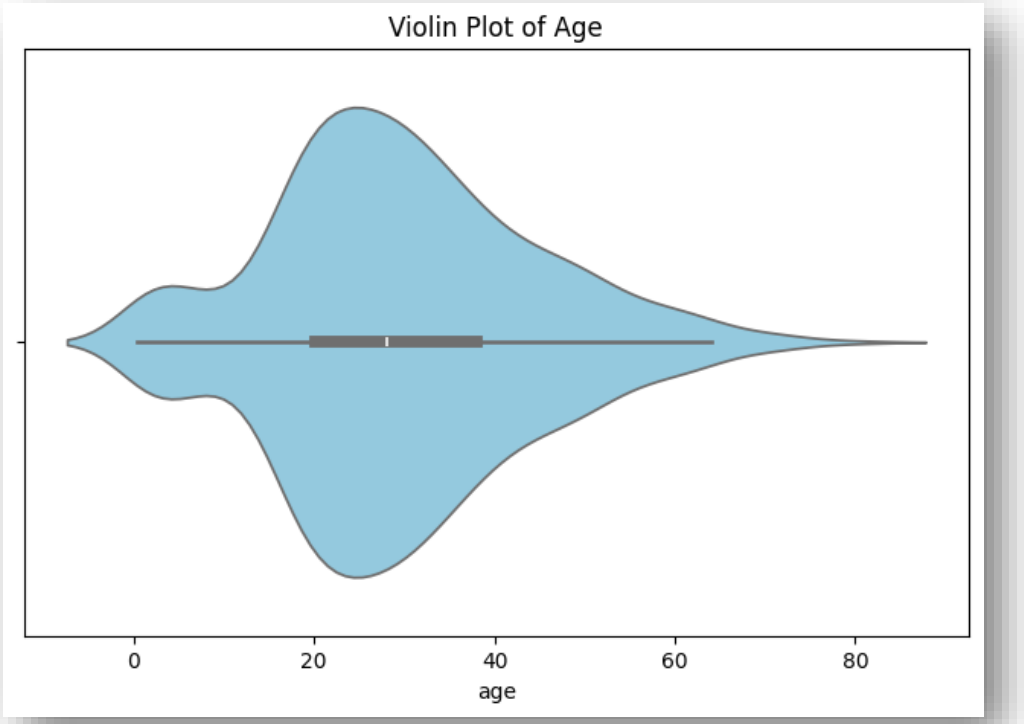
- 히스토그램
- 막대그래프
- Box plot
- **Violin plot**
- QQ-plot



EDA

Univariate graphical EDA
단변량 - 그래프 0

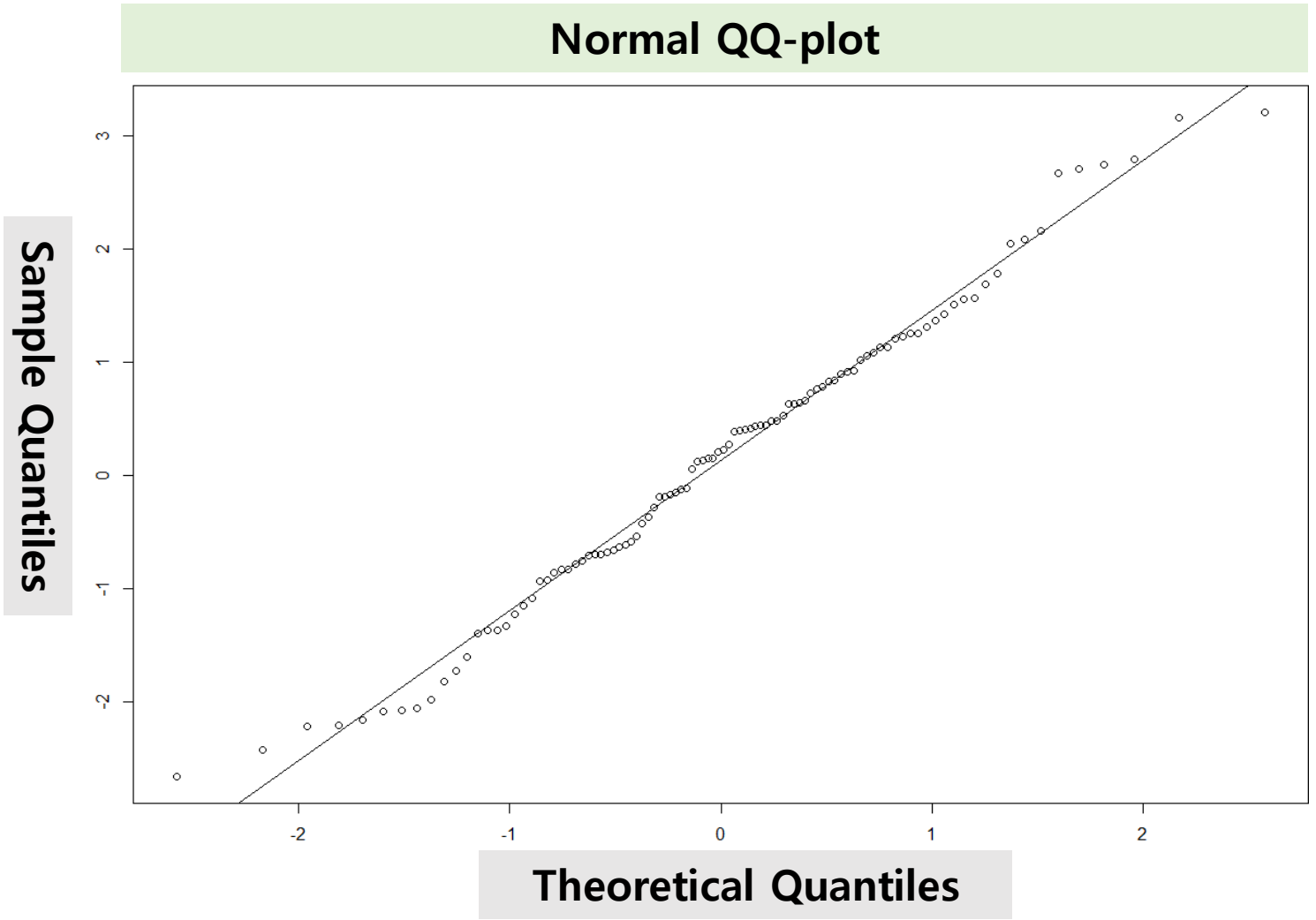
- 히스토그램
- 막대그래프
- Box plot
- **Violin plot**
- QQ-plot




```
# Age - Violin Plot

plt.figure(figsize=(8,5))
sns.violinplot(x=df["age"], color="skyblue")

plt.title("Violin Plot of Age")
plt.show()
```

EDA

Univariate graphical EDA
단변량 - 그래프 0



EDA

Univariate graphical EDA
단변량 - 그래프 0

정규성 검정(Normality Test)

- **Q-Q 플롯 : 데이터**터의 정규성을 평가하기 위한 시각적 도구
  데이터의 실제 분포와 정규분포의 이론적 분포 비교

- **분위수 비교 :** 데이터를 구간으로 나누어 정규분포와 비교 -> 데이터의 분포 특성 이해

- **밀도 플롯 :** 데이터의 전체 분포를 시각적으로 확인 -> 정규분포와 비교, 정규성 평가

- **실습**

  2.01.EDA.NormalityTest.ipynb

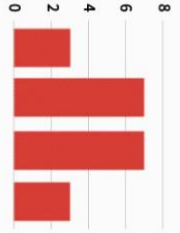
EDA

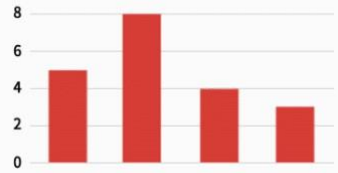
Multi-variate non-graphical EDA
다변량 – 그래프 X

- 교차표

품종	과중(단위: g)	당도(단위: Brix)
설	23.5	12.6
설	23.4	12.5
설	21.7	12.2
설	22.0	11.5
설	22.1	12.0
설	22.8	12.6
설	22.6	12.3
설	22.2	12.4
설	23.2	13.2
설	22.8	12.2
설	22.5	12.2
설	22.2	12.5
설	23.2	12.6
설	22.6	11.9
설	22.5	12.3
설	23.1	12.8
설	22.8	12.7
설	23.5	13.0
설	22.5	11.7
설	24.0	12.1

		과중				
		21.70~22.28	22.28~22.86	22.86~23.44	23.44~24.02	합계
당도	11.50~11.93	1	2	0	0	3
	11.93~12.35	2	4	0	1	7
	12.35~12.78	2	2	2	1	7
	12.78~13.20	0	0	2	1	3
	합계	5	8	4	3	20





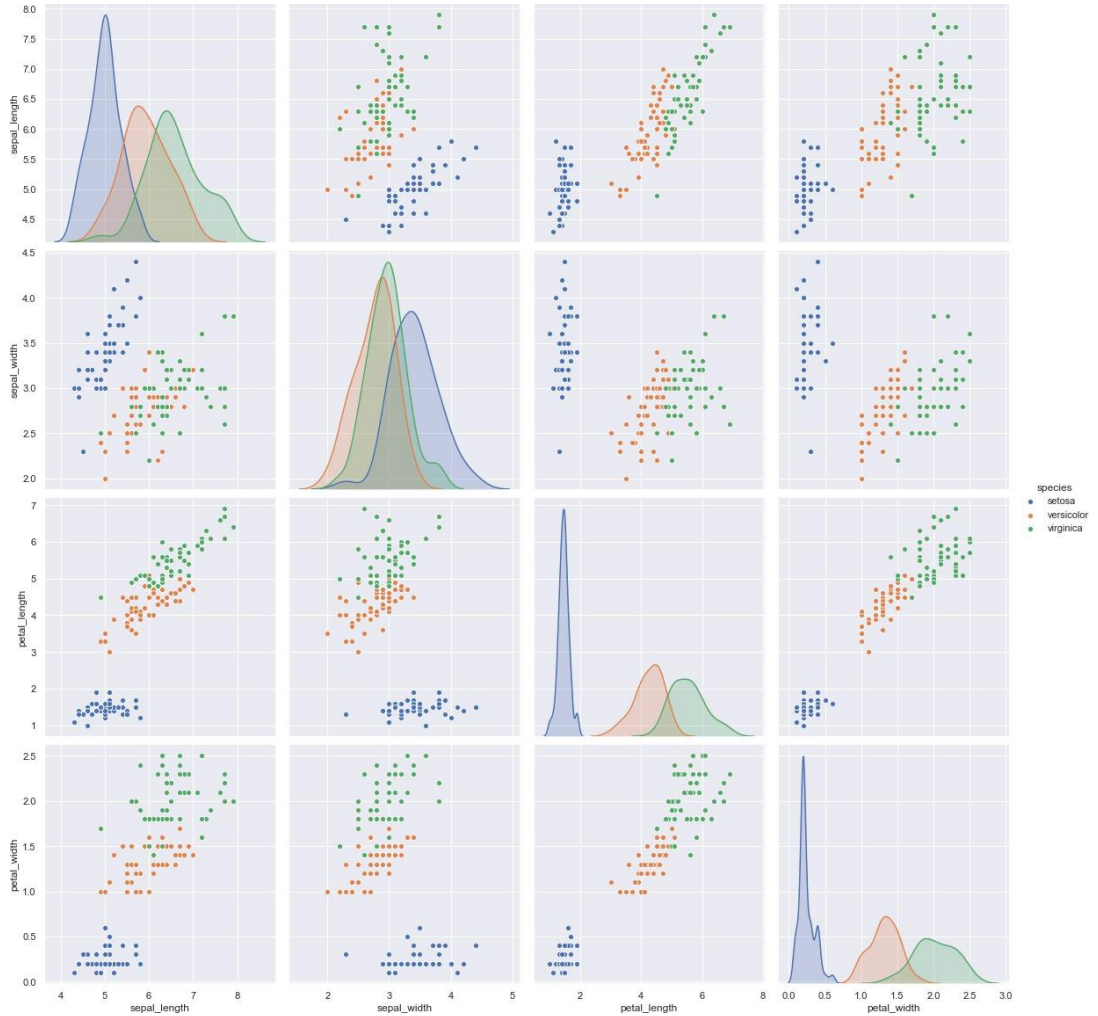
- 상관관계

	Categorical	Continuous
Categorical	Lambda, Corrected Cramer's V	Point Biserial, Logistic Regression
Continuous	Point Biserial, Logistic Regression	Spearman, Kendall, Pearson

EDA

Multi-variate graphical EDA
다변량 – 그래프 O

- Parallel plot
- Pair plot



EDA

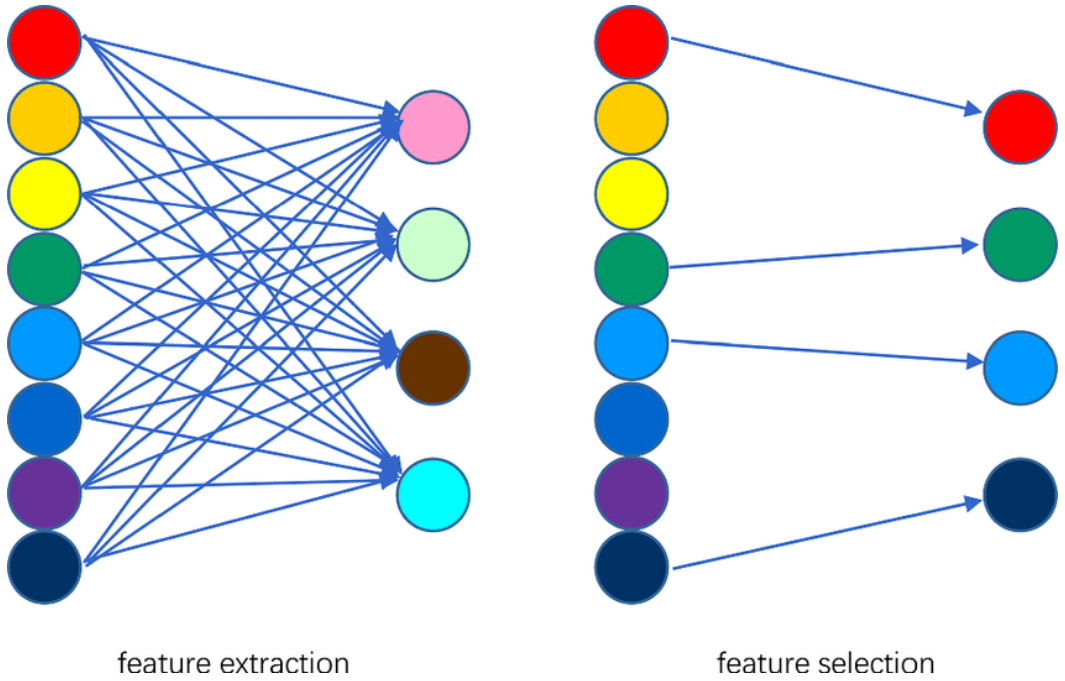
Multi-variate graphical EDA

- 실습

  2.01.EDA.Multi_variate.ipynb

EDA

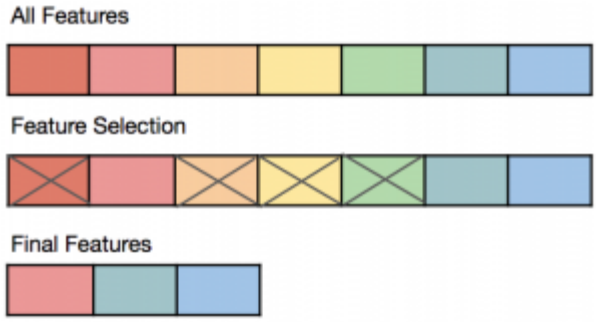
Feature Extraction *vs* Feature Selection



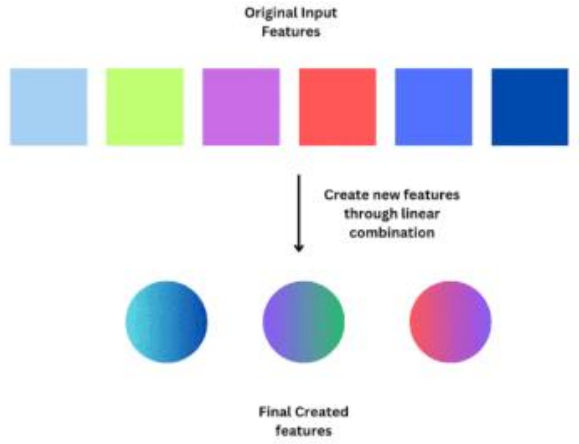
EDA

Feature Selection *vs* Feature Extraction

- **Feature Selection : 변수 선택**



- **Feature Extraction : 변수 추출(변환)**



EDA

Feature Selection *vs* Feature Extraction

- **Feature Selection (특성 선택)**

  - 주어진 데이터에서 **가장 중요한 특성**들을 **선택**하는 과정
  - 원본 데이터의 특성 중 일부를 선택하여 불필요한 특성을 제거 -> 차원 축소
  - 기존 특성 중에서 선택

  - **methodologies**
    - ✓ **전진 선택법(Forward Selection):** 가장 설명력이 높은 특성부터 차례대로 추가
    - ✓ **후진 제거법(Backward Elimination):** 모든 특성으로부터 출발하여 설명력이 낮은 특성부터 제거

  - **모델 해석력 UP** <- 중요한 특성만 남기기 때문에 결과에 대한 이해 UP.
  - **과적합(overfitting) 방지** <- 불필요한 특성 제거

EDA

Feature Selection *vs* Feature Extraction

- **Feature Selection (특성 선택)**

  - **methodologies**
    - ✓ **전진 선택법(Forward Selection):** 가장 설명력이 높은 특성부터 차례대로 추가
    - ✓ **후진 제거법(Backward Elimination):** 모든 특성으로부터 출발하여 설명력이 낮은 특성부터 제거

  - **실습**

    2.01.FeatureSelection.methodologies.ipynb

EDA

Feature Selection *vs* Feature Extraction

▪ Feature Extraction (특성 추출)

- 원본 데이터에서 **새로운 특성**을 **생성**하는 과정
- 기존 특성들로부터의 **수학적 변환** or **조합**을 통해 생성
- 데이터의 차원을 줄이면서도 중요한 정보를 유지하는 데 중점

- **methodologies**
  - ✓ **PCA (주성분 분석, Principal Component Analysis):** 여러 특성의 선형 결합을 통해 새로운 축(주성분) 생성
    -> 데이터를 새로운 저차원 공간으로 변환.
  - ✓ **SIFT (Scale-Invariant Feature Transform):** 이미지 데이터에서 중요한 부분을 추출, 특징 벡터로 변환

- 원래 특성 공간에서 중요한 정보를 유지하면서 **데이터 차원 축소**
- **모델 성능 향상** <- 불필요한 특성 제거

EDA

Feature Selection *vs* Feature Extraction

	Feature Selection (특성 선택)	Feature Extraction (특성 추출)
결과	기존 특성 중에서 중요한 **특성 선택**	기존 특성의 조합이나 변환을 통해 **새로운 특성 생성**
차원 축소	기존 특성 중 불필요한 것을 **제거**	기존 데이터 공간을 새로운 **저차원 공간**으로 **변환**
해석	특성은 기존 특성 선택 -> **해석 용이**	새로운 특성 생성(원래 속성과는 다름) -> **해석 어려움**
주요 기법	Lasso, Decision Trees, SVM 등	PCA, LDA, T-SNE 등

EDA

AI 개발

“Applied machine learning is basically feature engineering”
— Andrew Ng

When working with a paucity of data, or less feature-rich data, which is all too common for data scientists tasked with coming up with predictions based on just a dozen or so features, *feature engineering* is essential to **eke and tease out all the available ‘signal’ that’s present in the limited data**; as well as to **overcome the limitations of popular machine-learning algorithms**, for example, difficulty in separating data based on multiplicative or divisive feature interactions.

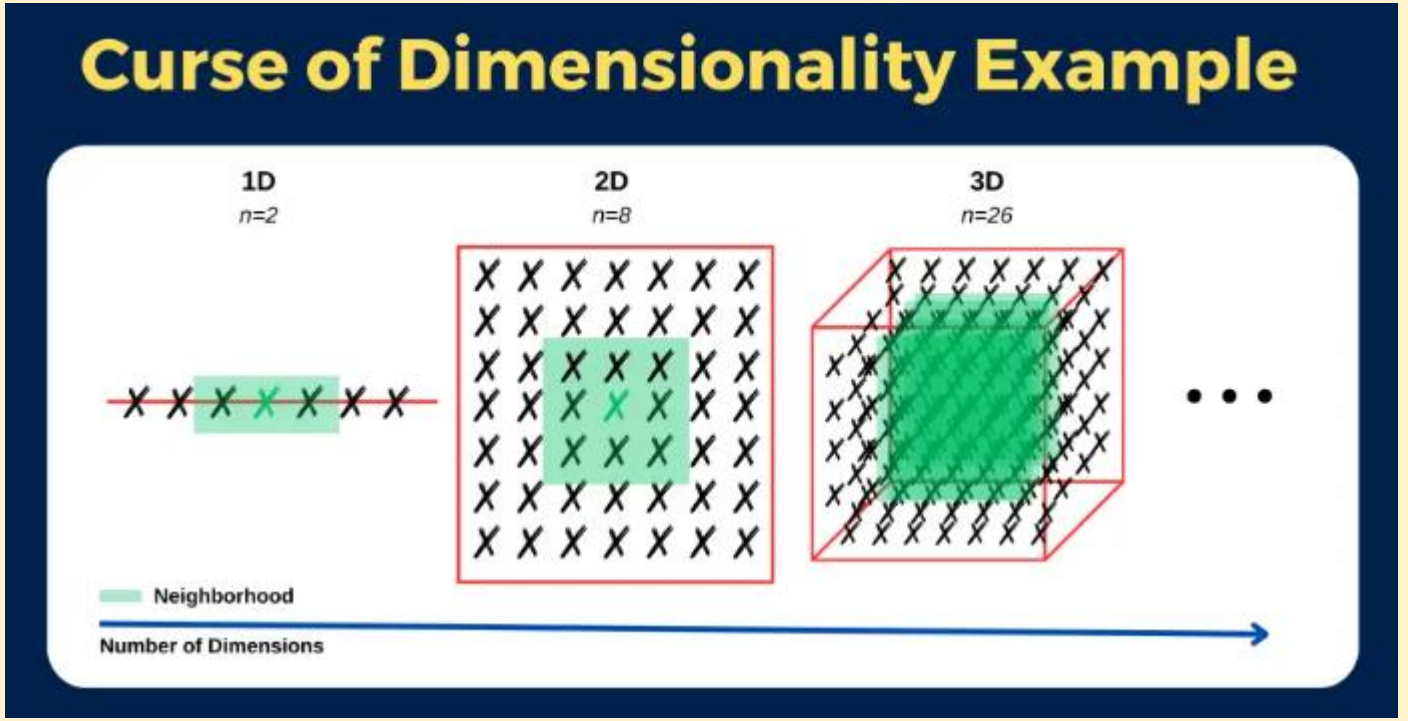
EDA

Curse of dimension

데이터의 차원이 높아질 수록 알고리즘의 실행이 아주 까다로워지는 현상

차원의 저주는 일상 경험의 3차원 물리적 공간과 같은 저차원 환경에서는 발생하지 않는

고차원 공간에서 데이터를 분석하고 정리할 때 발생하는 다양한 현상을 말한다 - wiki



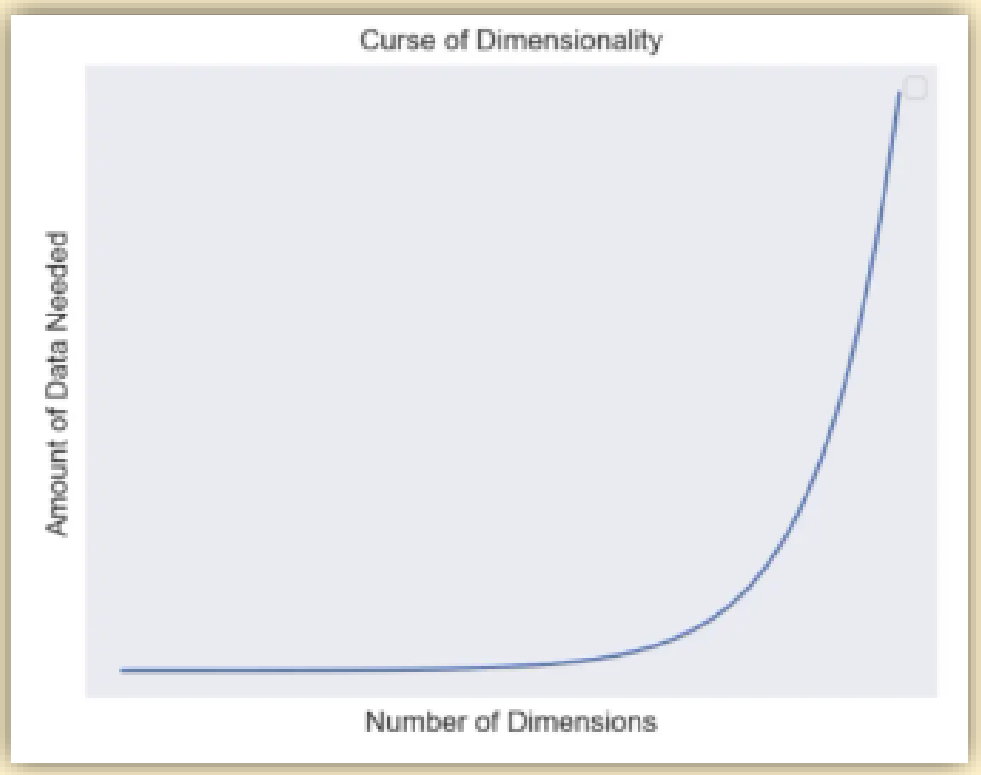
EDA

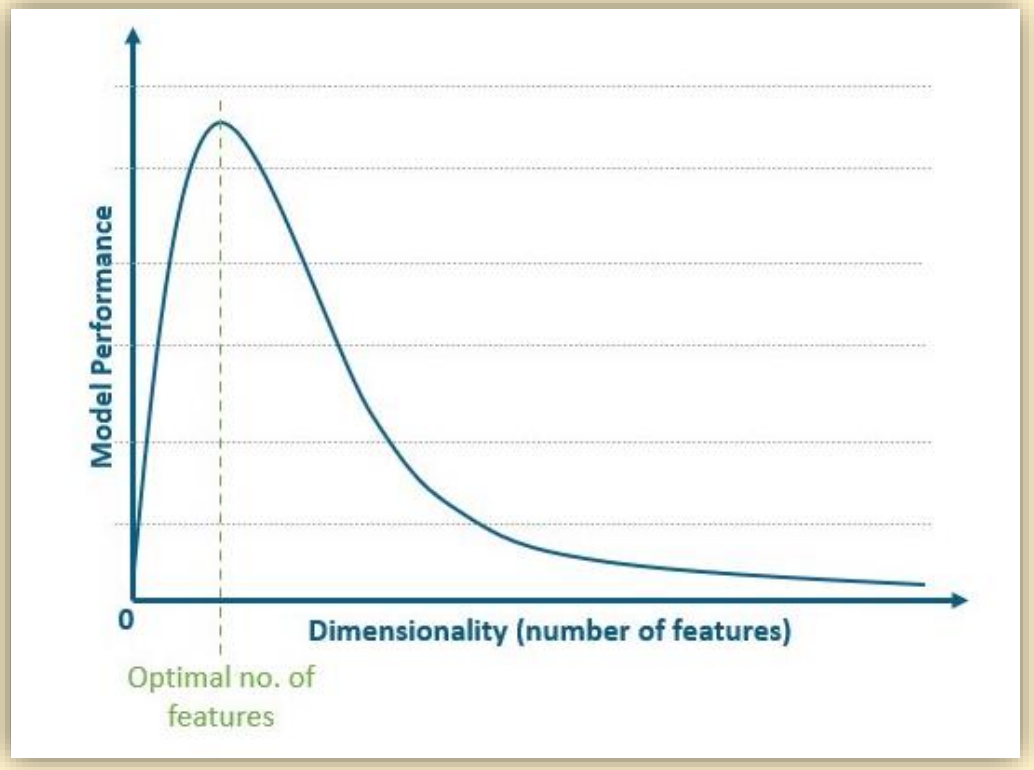
Curse of dimension

데이터의 차원이 높아질 수록 알고리즘의 실행이 아주 까다로워지는 현상

차원의 저주는 일상 경험의 3차원 물리적 공간과 같은 저차원 환경에서는 발생하지 않는

고차원 공간에서 데이터를 분석하고 정리할 때 발생하는 다양한 현상을 말한다 - wiki





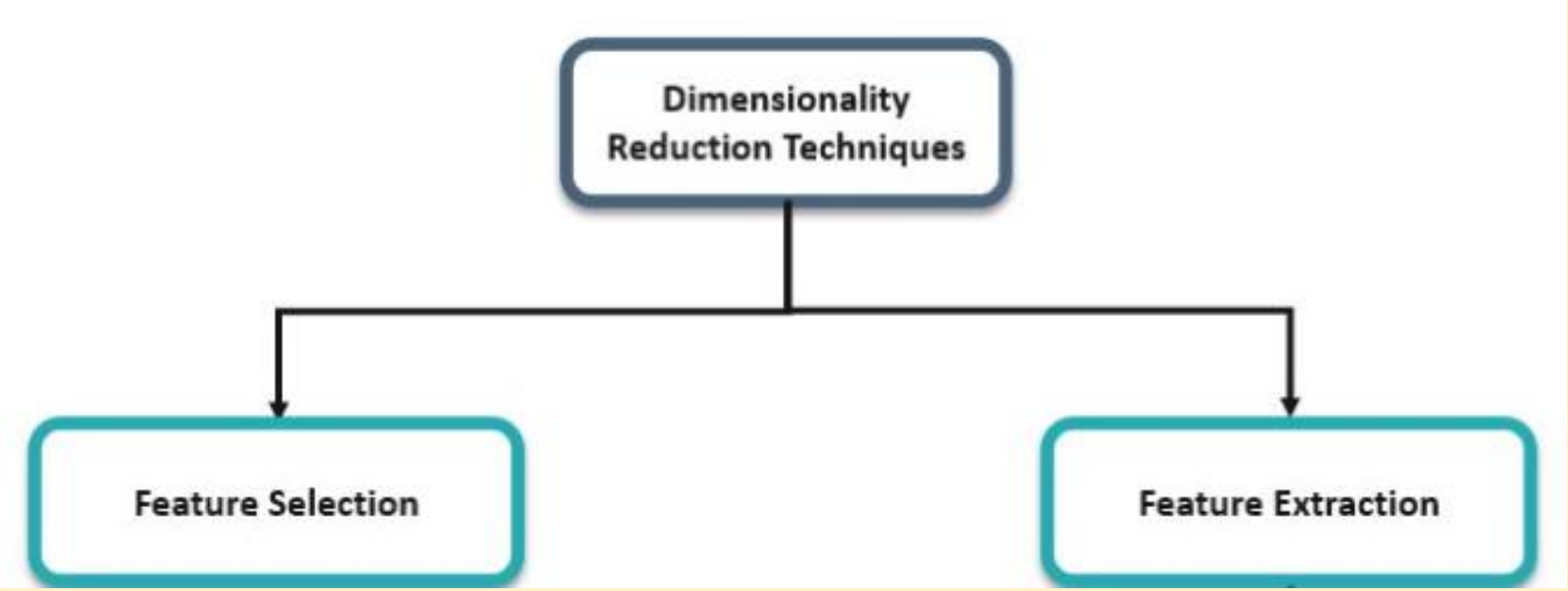
EDA

Curse of dimension

데이터의 차원이 높아질 수록 알고리즘의 실행이 아주 까다로워지는 현상

차원의 저주는 일상 경험의 3차원 물리적 공간과 같은 저차원 환경에서는 발생하지 않는

고차원 공간에서 데이터를 분석하고 정리할 때 발생하는 다양한 현상을 말한다 - wiki



EDA

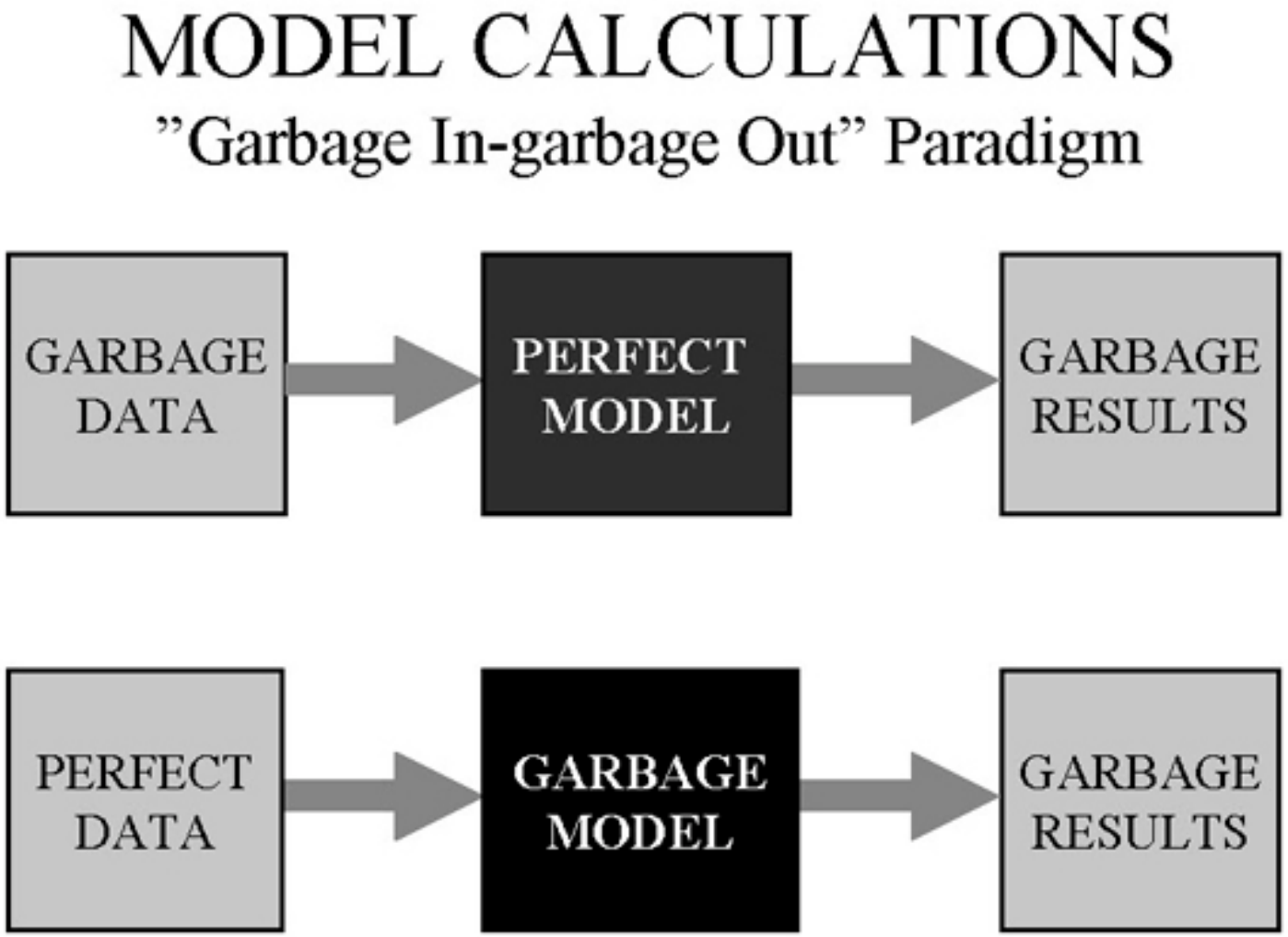
Curse of dimension

- 실습

  2.03.EDA.CD.ipynb

Feature selection

EDA


MODEL CALCULATIONS
"Garbage In-garbage Out" Paradigm
GARBAGE DATA
PERFECT MODEL
GARBAGE RESULTS
PERFECT DATA
GARBAGE MODEL
GARBAGE RESULTS

EDA

Feature selection

:수집 특성 중에서 목적에 가장 유용한 특성 선택

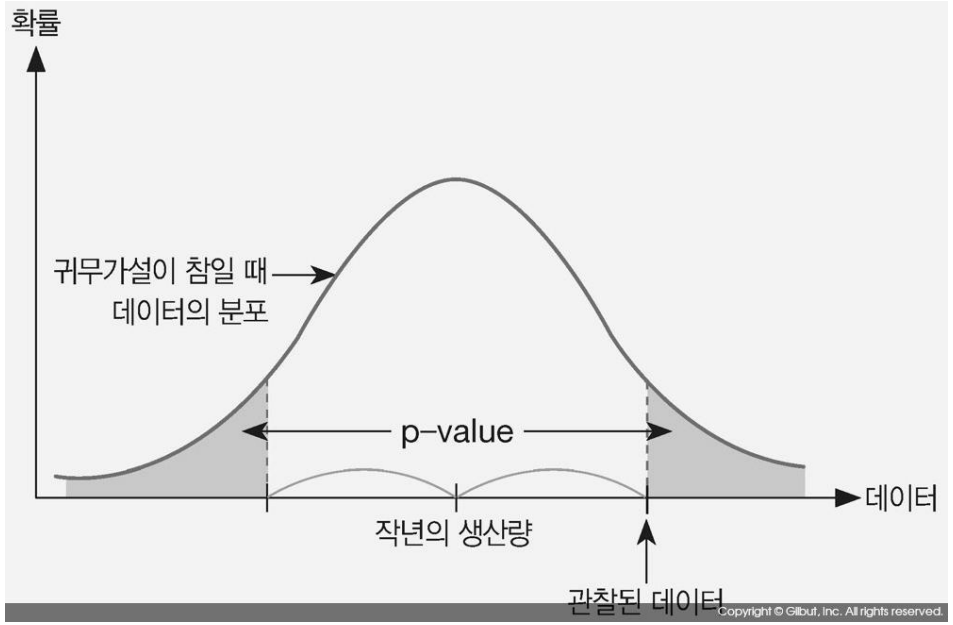
- **Filter method: 통계적 상관관계를 이용하여 전처리에서 선택**
  - information gain 가장 많은 정보를 제공하는 기능이 선택
  - chi-square test 결과가 기능과 독립적인지 확인하는 데 사용. **p-값이 낮은 특징 선택**
  - correlation coefficient 목표변수와 상관관계가 높은 특성을 선택
  - variance threshold: 분산이 더 높은(즉, 데이터의 변형이 더 많은) 기능이 선택
  - **높은 상관계수(영향력)을 가지는 피쳐 활용**

EDA

Feature selection

p-값(p-value)
관찰된 데이터가 귀무가설과 양립하는 정도를 0에서 1 사이의 수치로 표현한 것
p-value가 작을수록 그 정도가 약하다고 보며, 특정 값 (대개 0.05나 0.01 등) 보다 작을 경우 귀무가설 기각

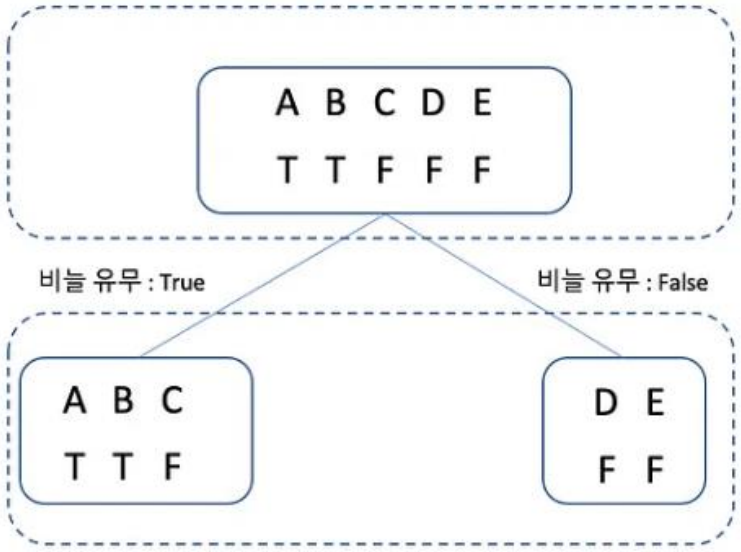
- 데이터 분석에서 특성 선택 시
  p-값을 기준으로 특정 특성이 결과 변수와 상관 관계가 있는지 판단
  -> p-값이 낮은 특성을 선택



EDA

Feature selection

- 엔트로피 : 데이터 세트의 무질서도를 측정
- **정보 이득** : 특성이 결과에 대한 엔트로피를 얼마나 감소시키는지를 나타냄.
- **정보 이득이 큰 특성**은 더 유용한 정보를 제공한다고 판단하여 선택



분할 전 노드

분할 후 노드

분할 전 엔트로피 = $-\left(\frac{2}{5}\log_2\left(\frac{2}{5}\right)+\frac{3}{5}\log_2\left(\frac{3}{5}\right)\right)$

분할 후 엔트로피 = $-\left(\frac{3}{5}\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right)+\frac{2}{3}\log_2\left(\frac{2}{3}\right)\right)+\frac{2}{5}\left(\frac{0}{2}\log_2\left(\frac{0}{2}\right)+\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right)\right)$

정보 획득 = 분할 전 엔트로피 - 분할 후 엔트로피 = 0.67 - 0.38 = 0.29

EDA

Feature selection

- 엔트로피 : 데이터 세트의 무질서도를 측정
- **정보 이득** : 특성이 결과에 대한 엔트로피를 얼마나 감소시키는지를 나타냄.
- **정보 이득이 큰 특성**은 더 유용한 정보를 제공한다고 판단하여 선택

- **실습**

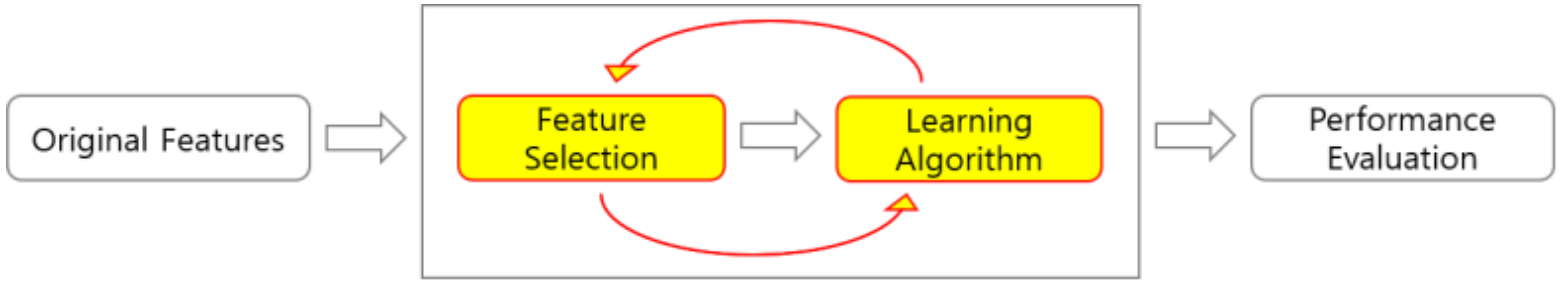
  2.01.FeatureSelection.Entropy.ipynb

- **Task**
  2.01.FeatureSelection.methodologies.ipynb 와 결과 비교

EDA

Feature selection

- 예측 모델의 성능을 직접적으로 평가
- **모델의 성능을 기준**으로 특성들의 부분 집합을 선택 or 제거
- 이 과정을 반복하면서, 최적의 특성 조합을 찾음



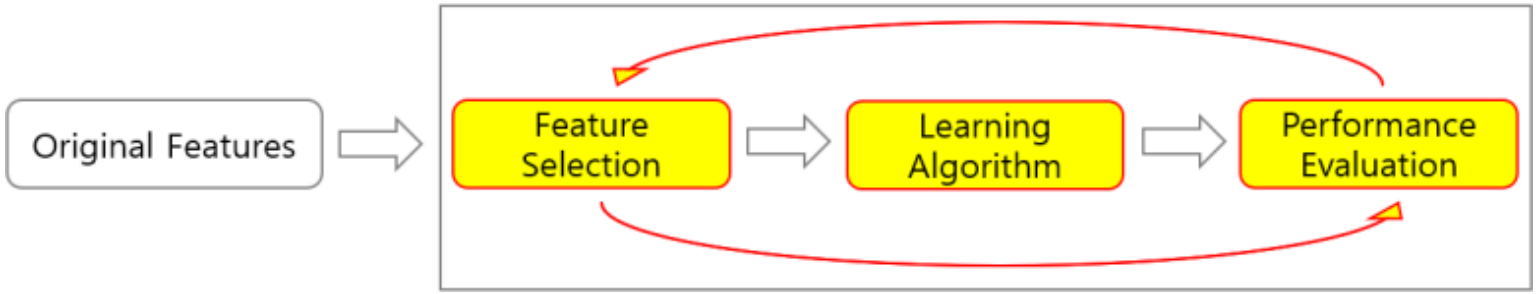
- **Wrapper method: 예측 정확도를 기준으로 Best subset을 찾음**
  - recursive feature elimination(**RFE**): SVM을 사용하여 재귀적으로 제거하는 방법
  - sequential feature selection(**SFS**): greedy 알고리즘으로 빈 subset에서 피처를 하나씩 추가
  - **모델 기반**: cross-validation으로 테스트할 셋 별도 마련

    * 계산 비용이 높고 시간 많이 소요

EDA

Feature selection

- 특성 선택(feature selection)과 **규제화(regularization)**를 결합
- 모델의 복잡성 감소 -> 과적합(overfitting) 방지



- **Embedded method: filter + wrapper**
  - **LASSO** : L1-norm을 통해 제약 주는 방법
  - **Ridge** : L2-norm을 통해 제약을 주는 방법
  - **Elastic Net** : 위 둘을 선형결합한 방법
  - **SelectFromModel**: decision tree 기반 알고리즘을 이용 피처 선택 (RandomForest, LightGBM 등)

EDA

Feature selection

- 특성 선택(feature selection)과 **규제화(regularization)**를 결합
- 모델의 복잡성 감소 -> 과적합(overfitting) 방지

- **Embedded method: filter + wrapper**
  - **SelectFromModel**: decision tree 기반 알고리즘을 이용 피처 선택 .

- **실습**

  2.01.FeatureSelection.SelectFromModel.ipynb

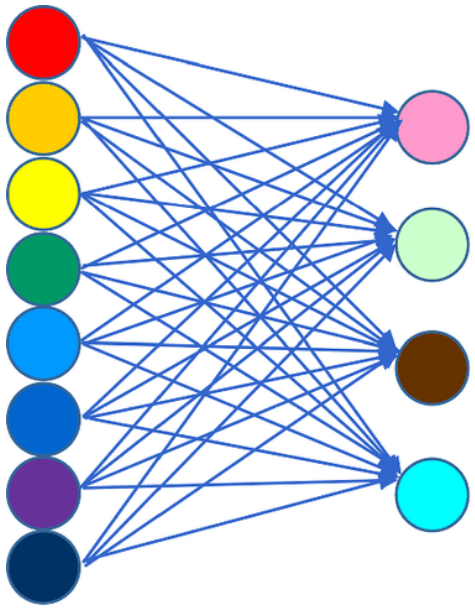
```
선택된 특성들의 중요도: [[-2.51673255 -1.08077345]
 [-0.20662058 -0.94364829]
 [ 2.72335313  2.02442173]]
```

- **Task**
  시각화를 통해 결과의 타당성 확인

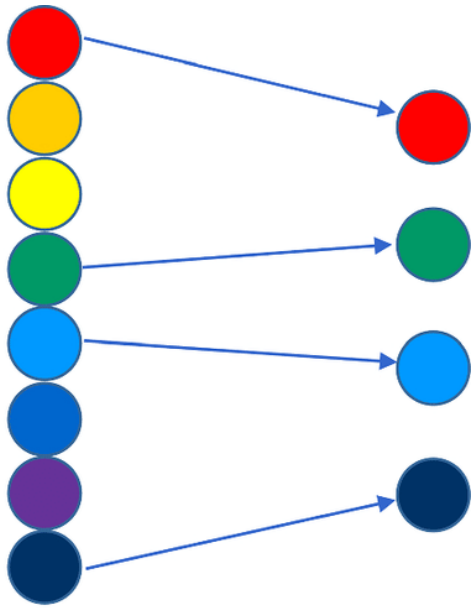
Feature extraction

EDA

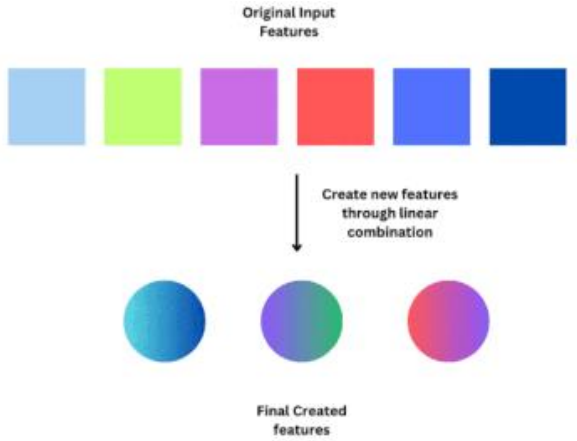
Feature Extraction



feature extraction

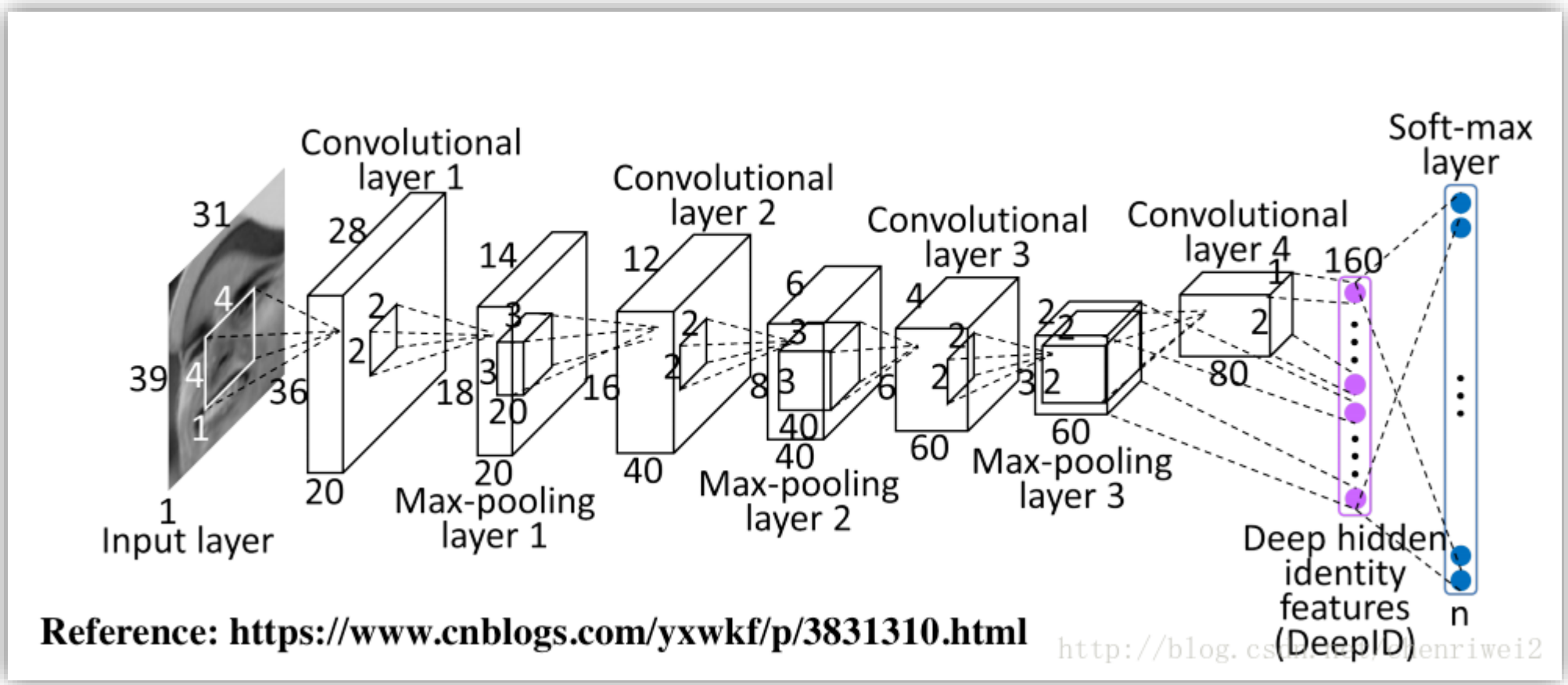


feature selection



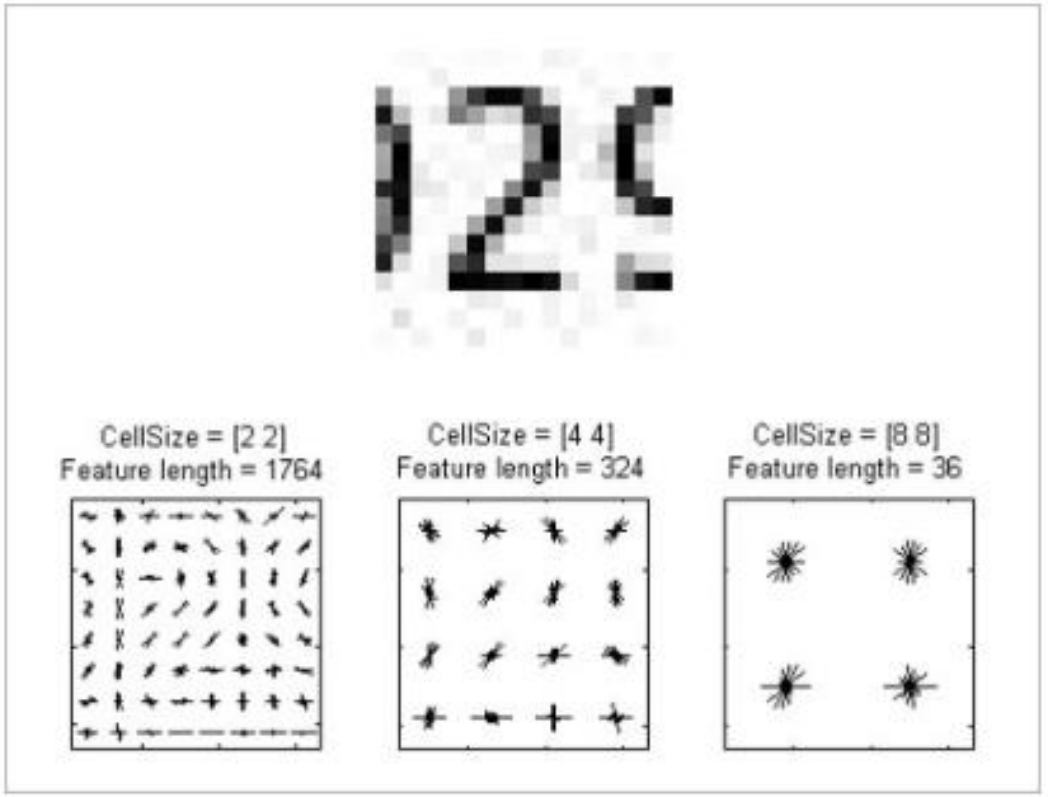
EDA

Feature Extraction - 이미지



EDA

Feature Extraction - 이미지



EDA

Feature Extraction - PCA

Principal Component Analysis

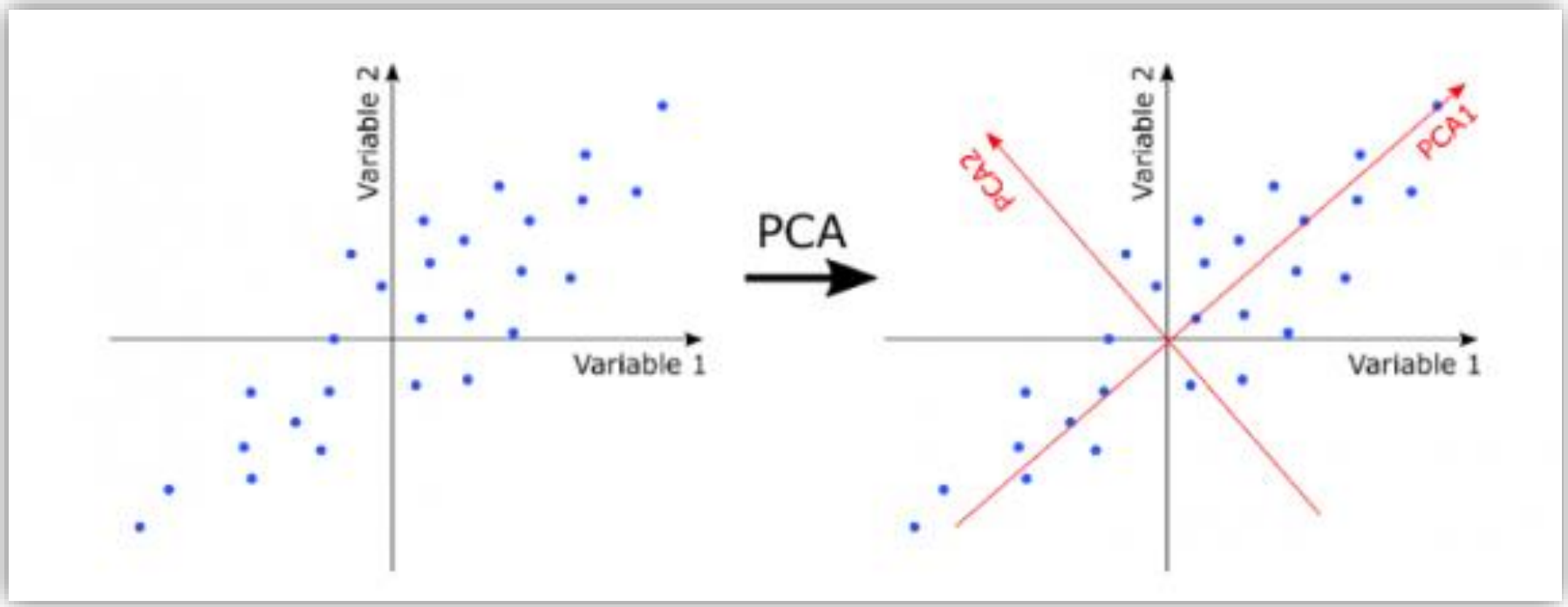
- **고차원 데이터를 저차원으로 축소하면서 데이터의 분산(variance)을 최대한 보존하는 방법**
- **특성 선택 , 데이터의 시각화, 노이즈 제거 등에 사용**

- **Eigenvector, Eigenvalue**
  - 공분산 행렬의 고유벡터와 고유값을 통해 데이터의 주요 방향 탐색
  - 고유값이 큰 고유벡터는 데이터의 분산이 많이 분포된 방향을 나타냄
  - **-> 고유값이 큰 고유벡터를 선택하여 주성분 형성**

EDA

Feature Extraction - PCA

Principal Component Analysis

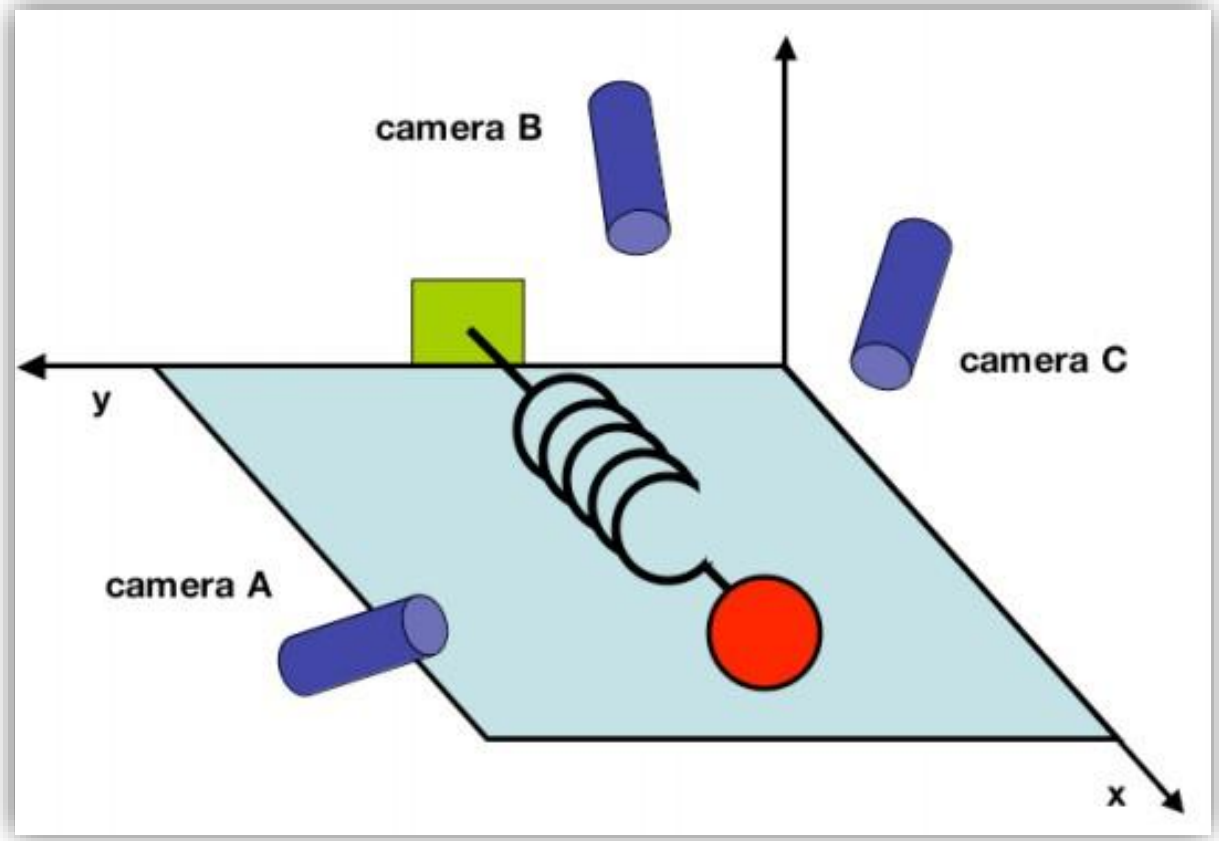


EDA

Feature Extraction - PCA

Principal **C**omponent **A**nalysis

용수철 움직임을 가장 잘 기록할 수 있는 카메라 배치?



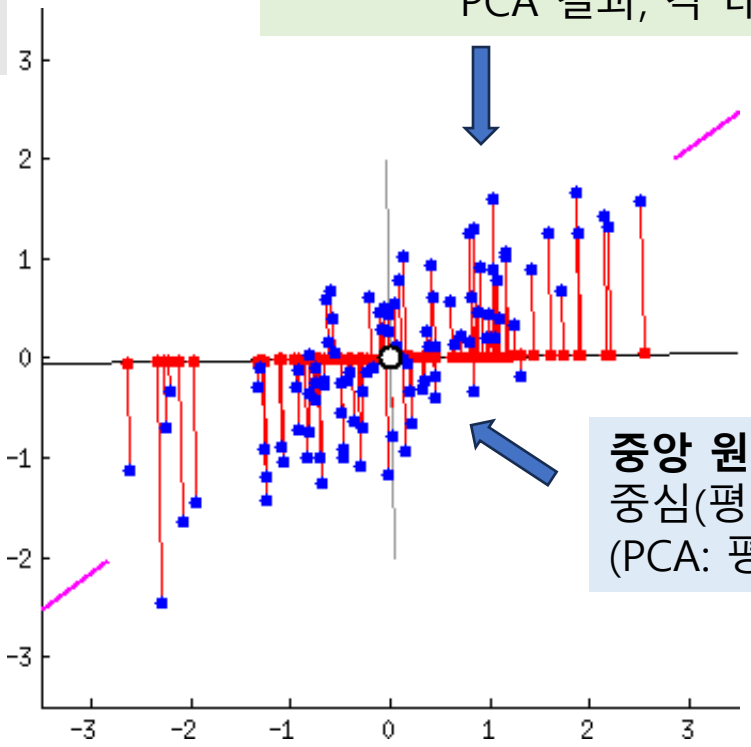
EDA

Feature Extraction - PCA

Principal **C**omponent **A**nalysis

• 각 점 : 원래의 다차원 데이터가 주성분 축으로 투영된 결과
PCA 결과, 각 데이터 포인트 위치

• **Y축** 두 번째 주성분(PC2)



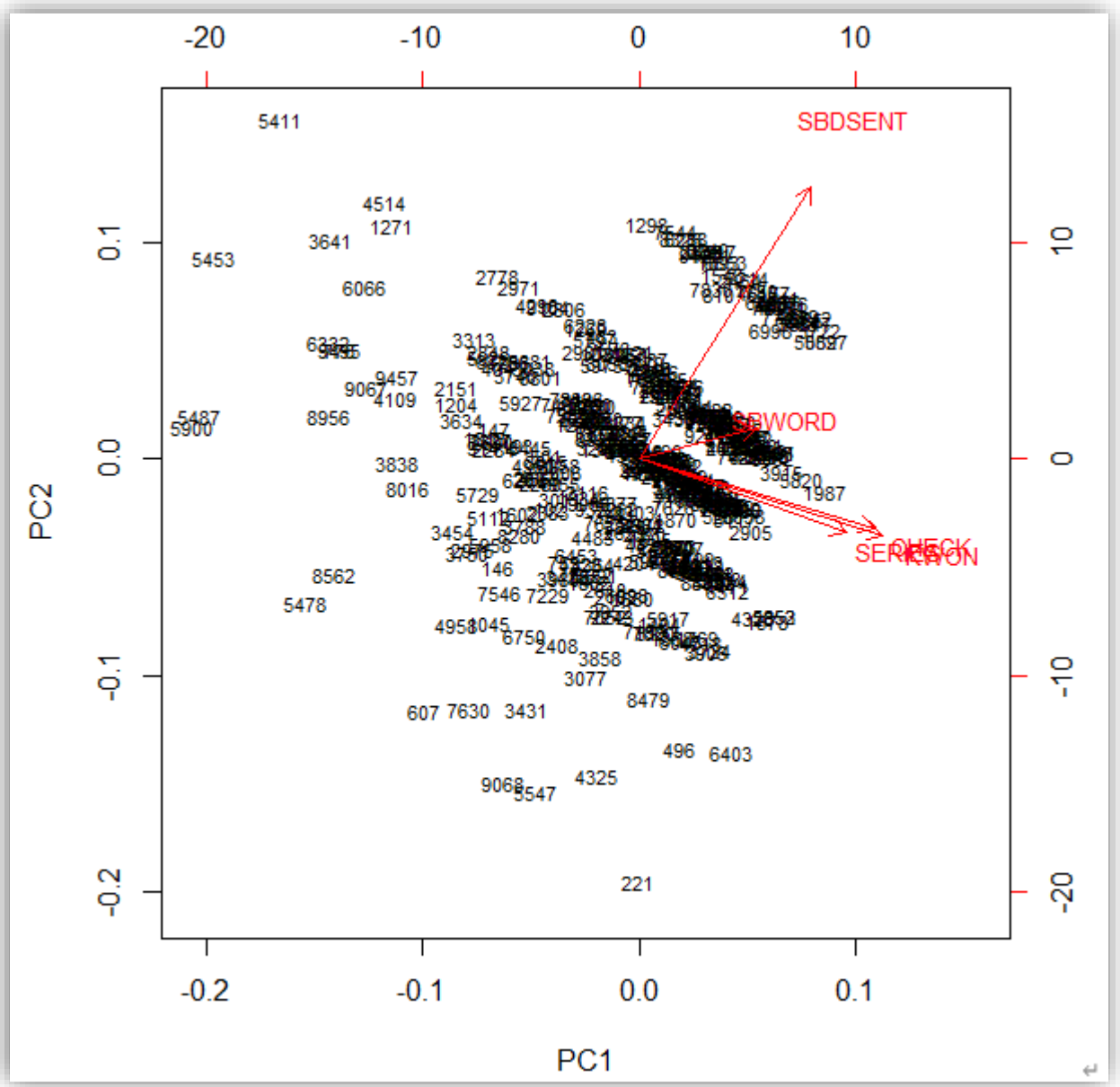
중앙 원 - 기준점
중심(평균점) 을 기준으로 데이터 정렬
(PCA: 평균이 0이 되도록 정규화 한 후 주성분 계산)

• **X축** 첫 번째 주성분(PC1)

EDA

Feature Extraction - PCA

Principal Component Analysis



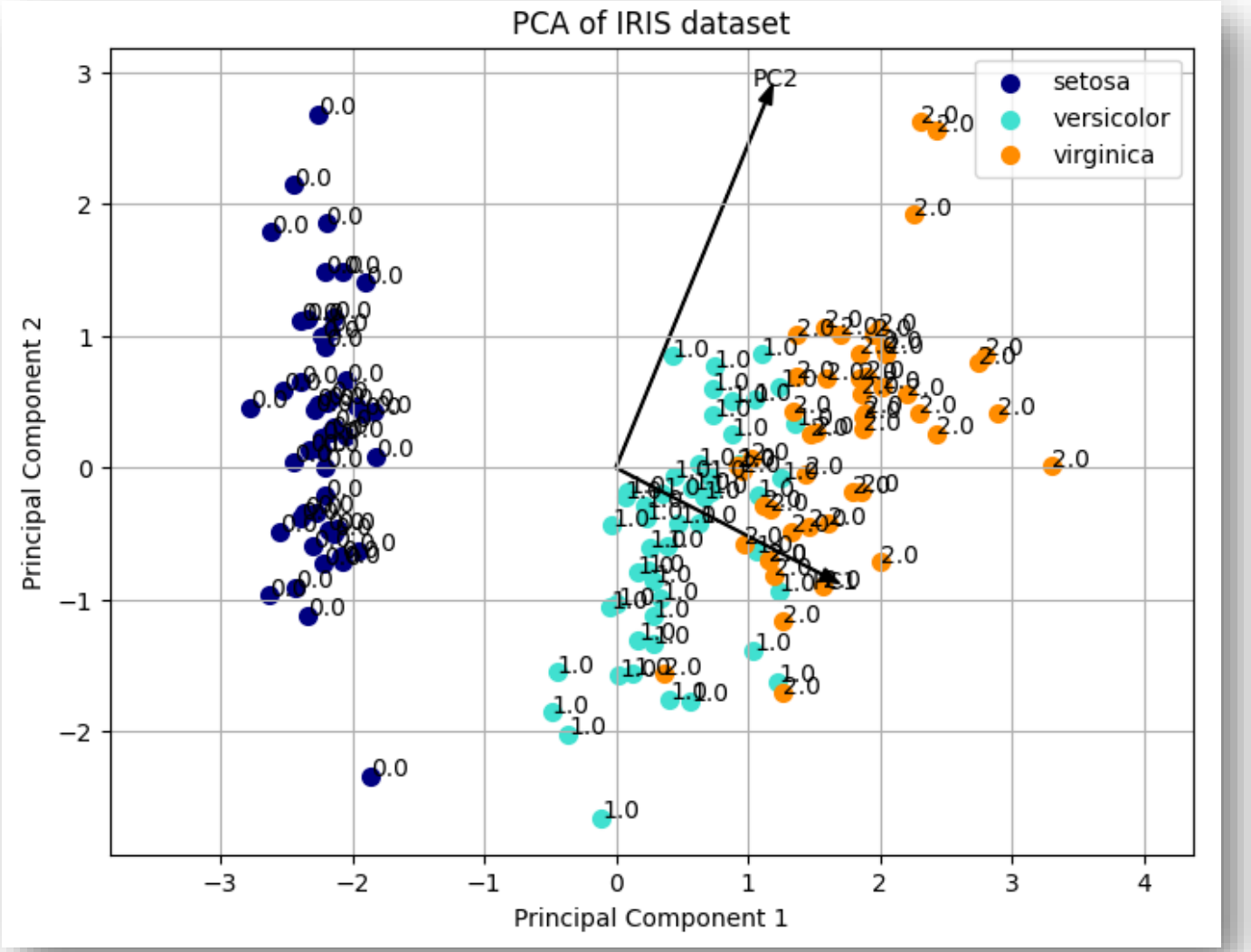
EDA

Feature Extraction - PCA

Principal Component Analysis

- 실습

  2.01.EDA.PCA.IRIS.ipynb

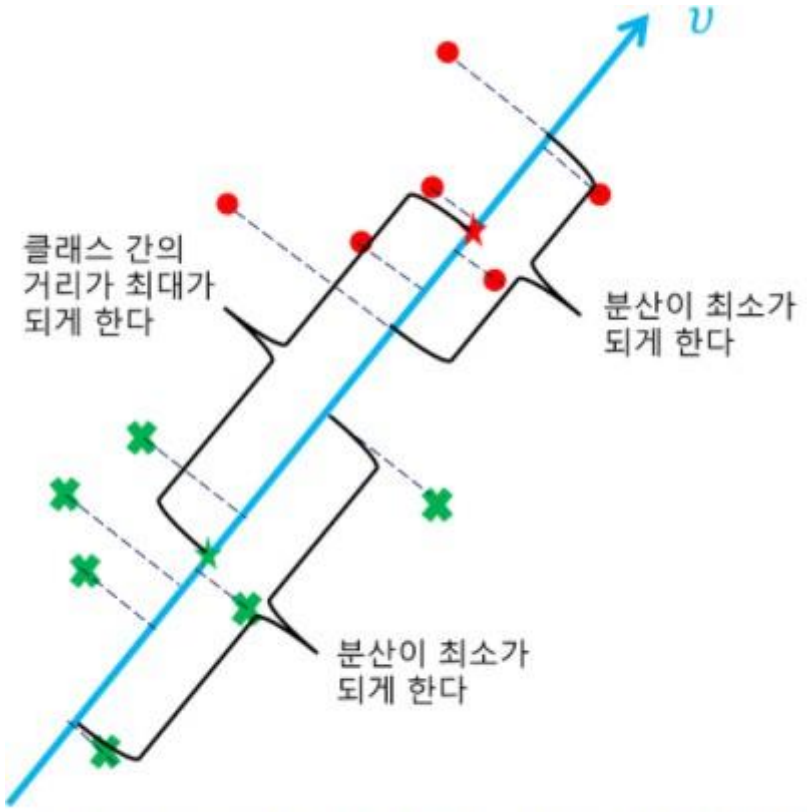


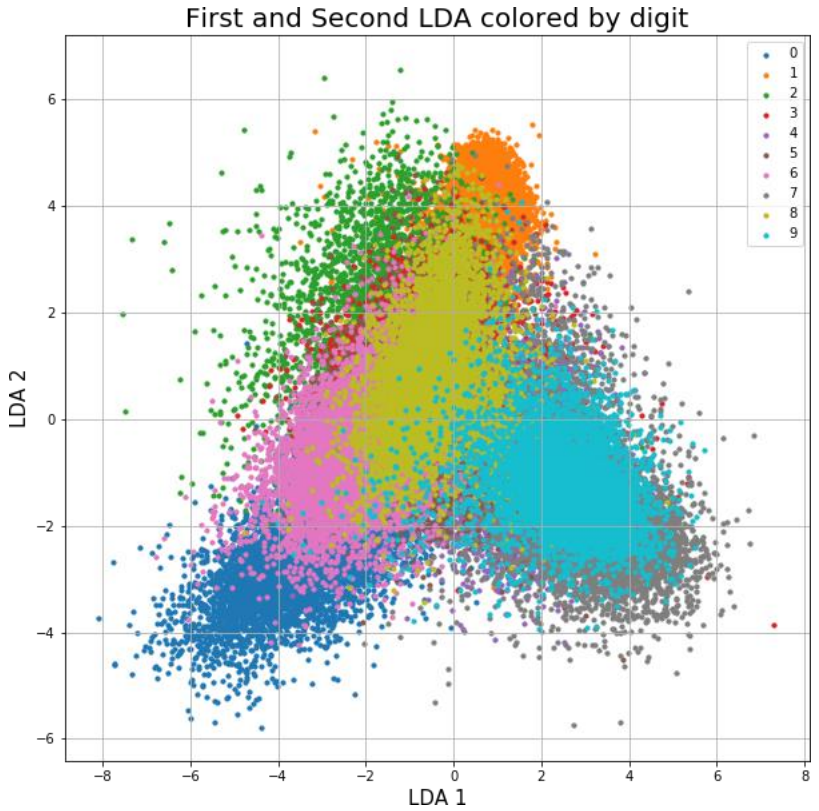
EDA

Feature Extraction - LDA

Linear Discriminant Analysis

- PCA 에서 확장된 차원 축소 기법.
- 지도 학습(supervised - learning)에서 적용하는 차원 축소 기법이자,
- 입력 데이터의 클래스(정답) 를 최대한 분리할 수 있는 축을 찾는 기법

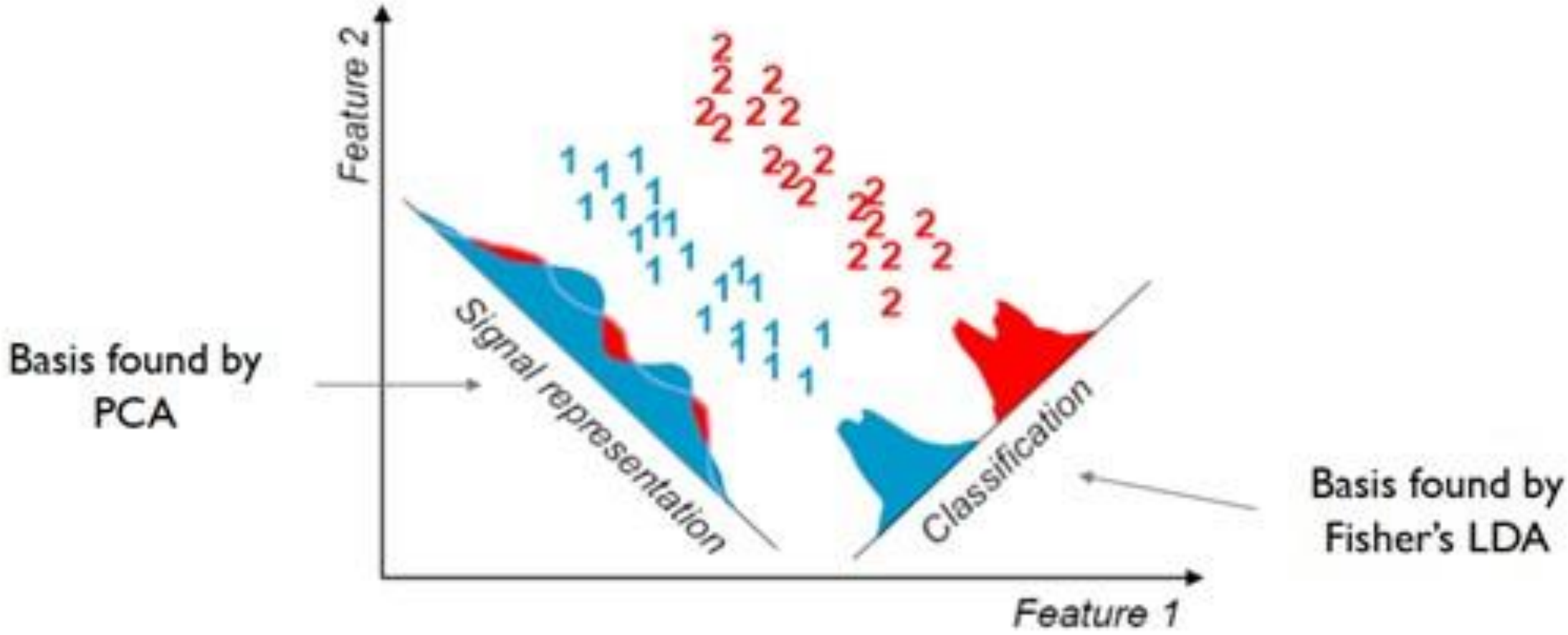




LDA : 클래스 간의 거리가 최대가 되게 하면서,
동시에 클래스의 분산이 최소가 되게 하는 벡터 탐색

EDA

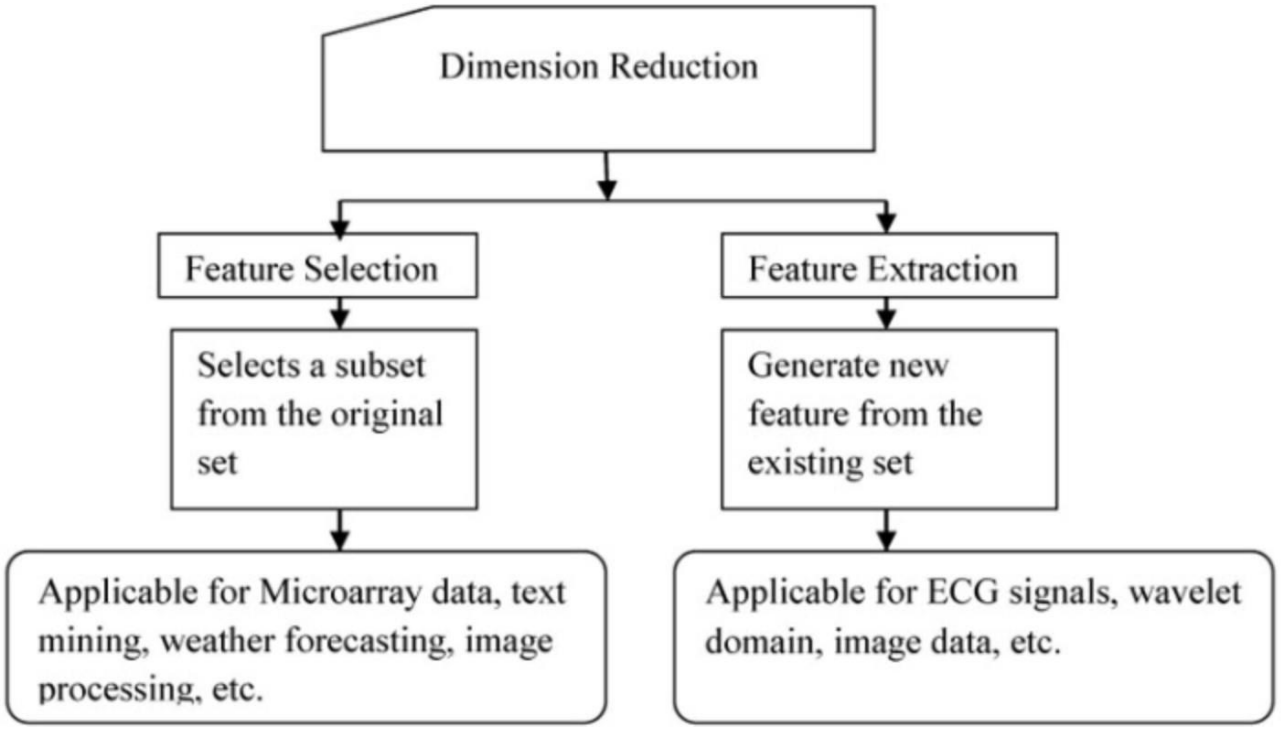
PCA *vs* LDA



https://casa-de-feel.tistory.com/19

EDA

Feature Extraction



EDA

- 실습

  2.01.EDA_iris.ipynb

  Task : 기타 여러 탐색 및 시각화 기법 적용

  2.01.EDA.IRIS_FeatureEngineering.ipynb

EDA

- **실습**

2.03_EDA.Heart_Disease.ipynb

Data: heart_disease.csv

pandas-profiling : EDA에 널리 사용되는 도구

TASK : 결과 분석(report.html)

Data: heart_disease.csv

Column	Meaning	설명
age	Age in years	나이
sex	Sex (1 = male; 0 = female)	성별
cp	Chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)	흉통유형
trestbps	Resting blood pressure (in mm Hg on admission to the hospital)	혈압
chol	Serum cholestoral in mg/dl	콜레스테롤
fbs	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)	공복혈당
restecg	Resting electrocardiographic results (0 = normal; 1 = having ST-T wave abnormality; 2 = probable or definite left ventricular hypertrophy)	심전도(안정시
thalach	Maximum heart rate achieved	심박수
exang	Exercise induced angina (1 = yes; 0 = no)	협심증
oldpeak	ST depression induced by exercise relative to rest	우울중
slope	The slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)	ST 기울기
ca	Number of major vessels (0-3) colored by flourosopy	혈관수
thal	Thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)	빈혈
target	Presence of heart disease (1 = disease; 0 = no disease)	심장병 유무

EDA

- 실습

2.03_EDA.HR.ipynb

[TASK]
HR 데이터, 각 속성별 의미 파악

	birthday	entry_year	department	marital_status	performance_rating	job_satisfaction	working_hours	salary	last_year_salary	num_companies_worked	attrition
0	1980-7-20	2013	sales	single	high	very high	8.33	9431500	8923739	8.0	yes
1	1972-11-8	2011	rnd	married	very high	medium	6.93	5170672	4617495	NaN	no

attrition
yes
no

EDA

[EDA 실습]
타이타닉호 생존율 분석

- **목표 : 타이타닉호 승객 변수를 분석하여 생존율과의 상관관계 찾기**
  (EDA - 상관분석, 상관계수, 시각화 등)
- **데이터 :** 타이타닉 데이터
- **데이터 전처리 :** 결측치 처리, 중앙값 치환 최빈값 치환 등
- **데이터 탐색**
  - 정보확인, 통계분석 : info() 등 - 시각화 : pie().countplot() 등
- **데이터 모델링**
  - 모든 변수 간 상관계수 구하기 - 지정한 두 변수 간 상관계수 구하기
  - 기타 모델링
- **결과 시각화**

EDA

[EDA 실습]
타이타닉호 생존율 분석

- **Process**

  - 데이터 로드 (titanic 데이터)
  - 데이터 확인
  - 기본 정보 탐색
  - 시각화(차트) ex, 등급별 생존자 수

  - 데이터 분석
    - 상관 분석 -> 상관 계수 구하기

```
titanic_corr = titanic.corr(method = 'pearson')
titanic_corr
```

    - 상관 계수 해석하기

  - 특정 변수 사이의 상관 계수 구하기

```
titanic['survived'].corr(titanic['adult_male'])
```

  - 상관분석 시각화
    - 산점도

```
sns.pairplot(titanic, hue = 'survived')
pairplot()
```

    - 히트맵 시각화

EDA

[EDA 실습]
타이타닉호 생존율 분석

■ 분석 수행

- 타이타닉호의 생존자와 관련된 변수의 상관관계 탐색
- 생존과 가장 상관도가 높은 변수는?
- 상관 분석 - 피어슨 상관 계수
- 변수 간의 상관관계 시각화

EDA

[EDA 실습]
타이타닉호 생존율 분석

- **상관 분석**
  - 두 변수의 선형적 관계 분석

  - **상관관계(두 변수의 관계의 강도) 분석**
    - 단순 상관 분석
      : 두 변수
    - 다중 상관 분석
      : 세 개 이상의 변수 간 관계의 강도 측정

상관 계수
0.0 ~ 0.2: 상관관계가 거의 없음
0.2 ~ 0.4: 약한 상관관계가 있음
0.4 ~ 0.6: 상관관계가 있음
0.6 ~ 0.8: 강한 상관관계가 있음
0.8 ~ 1.0: 매우 강한 상관관계가 있음

EDA

[EDA 실습]
타이타닉호 생존율 분석

상관관계.pdf

EDA

유용한 사이트들

1. [KaKr] 탐색적 데이터 분석(EDA) 설명 + 예시 (Kaggle)
Kaggle 커뮤니티에서 제공하는 EDA에 대한 설명과 예시.
https://www.kaggle.com/code/subinium/kakr-eda

2. 탐색적 데이터 분석(EDA) (JMP)
EDA의 기본 개념과 요약 통계, 그래픽 도구를 사용하여 데이터를 이해하는 방법에 대한 설명
https://www.jmp.com/ko_kr/statistics-knowledge-portal/exploratory-data-analysis.html

3. 탐색형 데이터 분석이란? (IBM)
EDA를 수행할 때 사용할 수 있는 통계 기능과 기법 설명
https://www.ibm.com/kr-ko/topics/exploratory-data-analysis

4. 탐색적 데이터 분석(EDA, Exploratory Data Analysis) 이란? (DeepLink)
EDA의 개념, 절차
https://blog.deeplink.kr/?p=2130

THANK YOU