YBIGTA 14기 교육세션

# FEATURE ENGINEERING

13기 디자인팀
김민정 류재원

# Feature Engineering in Data Science Process

- Data Collection
- EDA
- Data Preprocessing
- **Feature Engineering**
- Modeling

INTRO

# FEATURE ENGINEERING?

모델에 데이터를 넣기 직전의 단계

모델의 성능을 높이기 위해 주어진 초기 데이터로부터 특징을 찾아 가공하고 생성하는 과정
(모델에 입력할 데이터를 만드는 과정의 일부)

머신러닝의 성능은 데이터의 양과 질에 굉장히 의존적이기 때문에, 모델 성능에 미치는 영향이 크다

시간이 많이 소요되는 과정

# CONTENTS



참고

모델 돌리기 전까지
4~7번 반복해야함

# CONTENTS

# Variable Identification

| id | date_account_created | gender | age | signup_method | first_device_type | first_browser | country_destination |
|---|---|---|---|---|---|---|---|
| 4ft3gnwmtx | 2010-09-28 | FEMALE | 56.0 | basic | Windows Desktop | IE | US |
| bjjt8pjhuk | 2011-12-05 | FEMALE | 42.0 | facebook | Mac Desktop | Firefox | other |
| 87mebub9p4 | 2010-09-14 | -unknown- | 41.0 | basic | Mac Desktop | Chrome | US |
| lsw9q7uk0j | 2010-01-02 | FEMALE | 46.0 | basic | Mac Desktop | Safari | US |
| 0d01nltbrs | 2010-01-03 | FEMALE | 47.0 | basic | Mac Desktop | Safari | US |

# Variable Identification

| id | date_account_created | gender | age | signup_method | first_device_type | first_browser | country_destination |
|---|---|---|---|---|---|---|---|
| 4ft3gnwmtx | 2010-09-28 | FEMALE | 56.0 | basic | Windows Desktop | IE | US |
| bjjt8pjhuk | 2011-12-05 | FEMALE | 42.0 | facebook | Mac Desktop | Firefox | other |
| 87mebub9p4 | 2010-09-14 | -unknown- | 41.0 | basic | Mac Desktop | Chrome | US |
| lsw9q7uk0j | 2010-01-02 | FEMALE | 46.0 | basic | Mac Desktop | Safari | US |
| 0d01nltbrs | 2010-01-03 | FEMALE | 47.0 | basic | Mac Desktop | Safari | US |

- Predictor

  → date_account_created ~ first_browser

- Target

  → country_destination

# Variable Identification

- Continuous

  연속형 변수의 경우에는, **Central tendency**와 변수의 **분산**에 대해 알아야 합니다

  → Mean, median, mode, Min, Max, Range, Quantile …

- Categorical

  각 카테고리의 분포를 이해하기 위해서 **frequency table**을 이용합니다

  → Count와 Count%를 나타내는 bar chart 그려보기 등

# Univariate Analysis

- Continuous

  연속형 변수의 경우에는, Central tendency와 변수의 분산에 대해 알아야 합니다

  → Mean, median, mode, Min, Max, Range, Quantile …

- Categorical

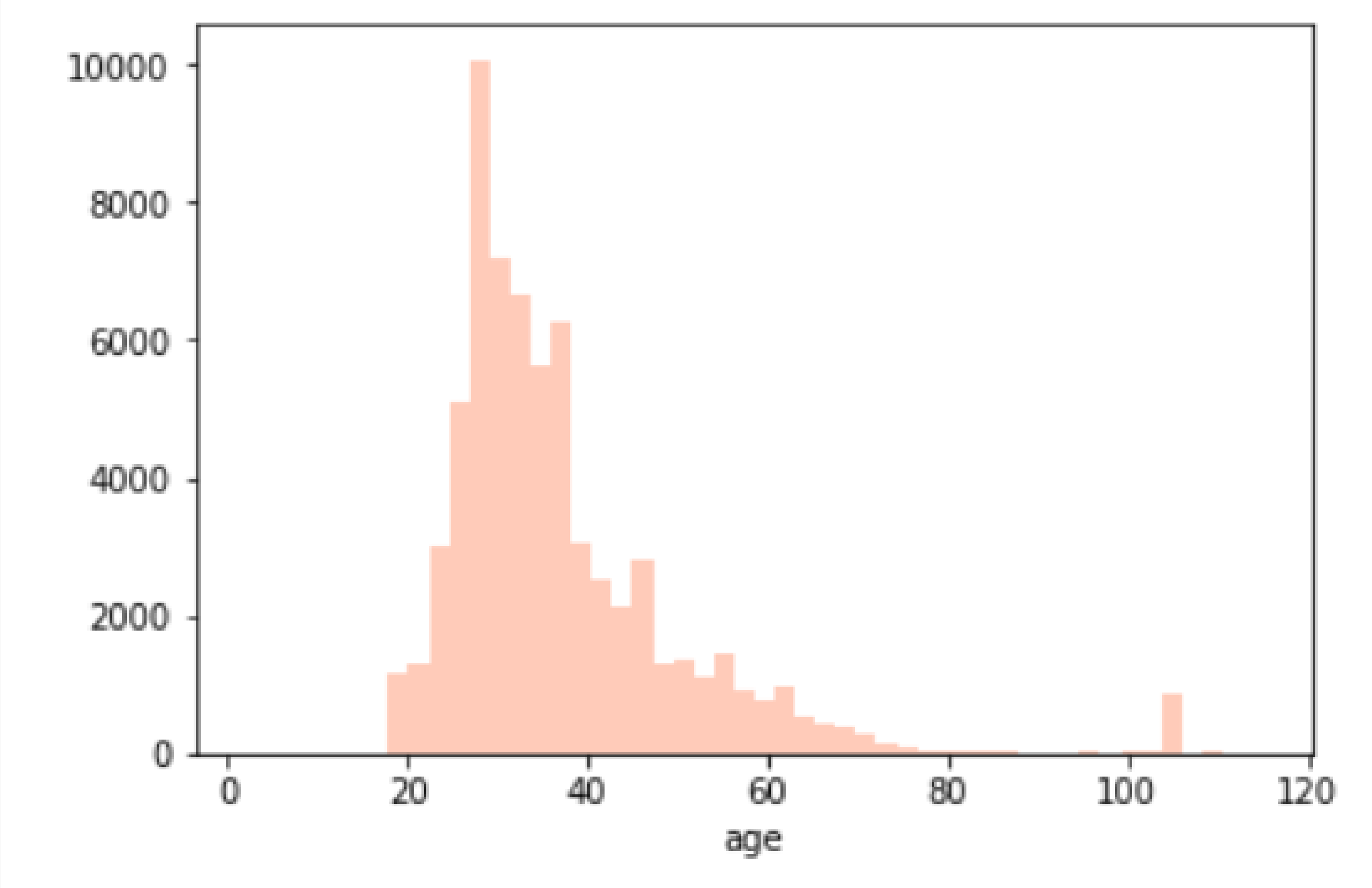
  각 카테고리의 분포를 이해하기 위해서 frequency table을 이용합니다

  → Count와 Count%를 나타내는 bar chart 그려보기 등

실습 시간에 Pandas Profiling 이용해서 파악해보세요!

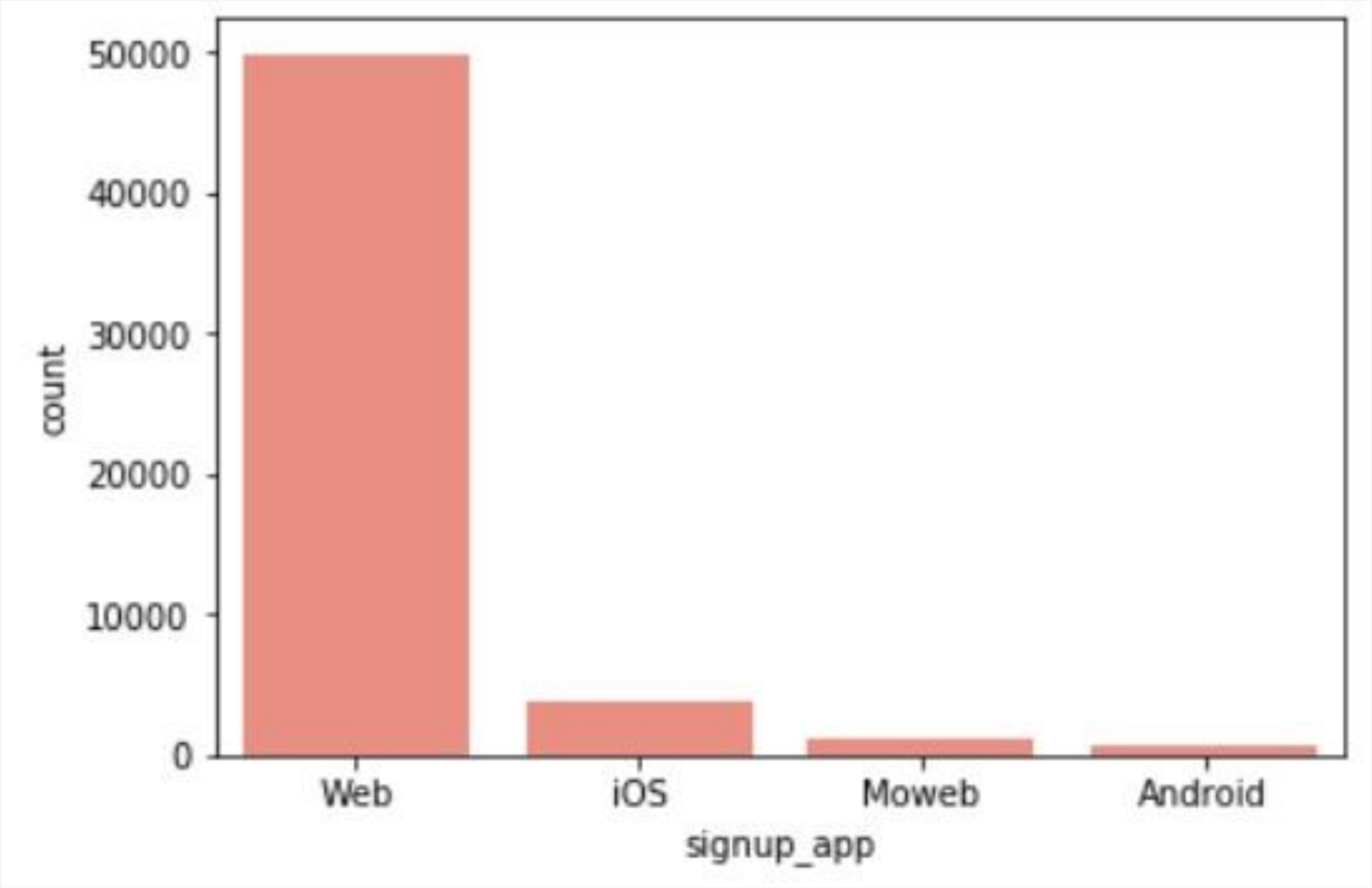
# Univariate Analysis

- Continuous Variable



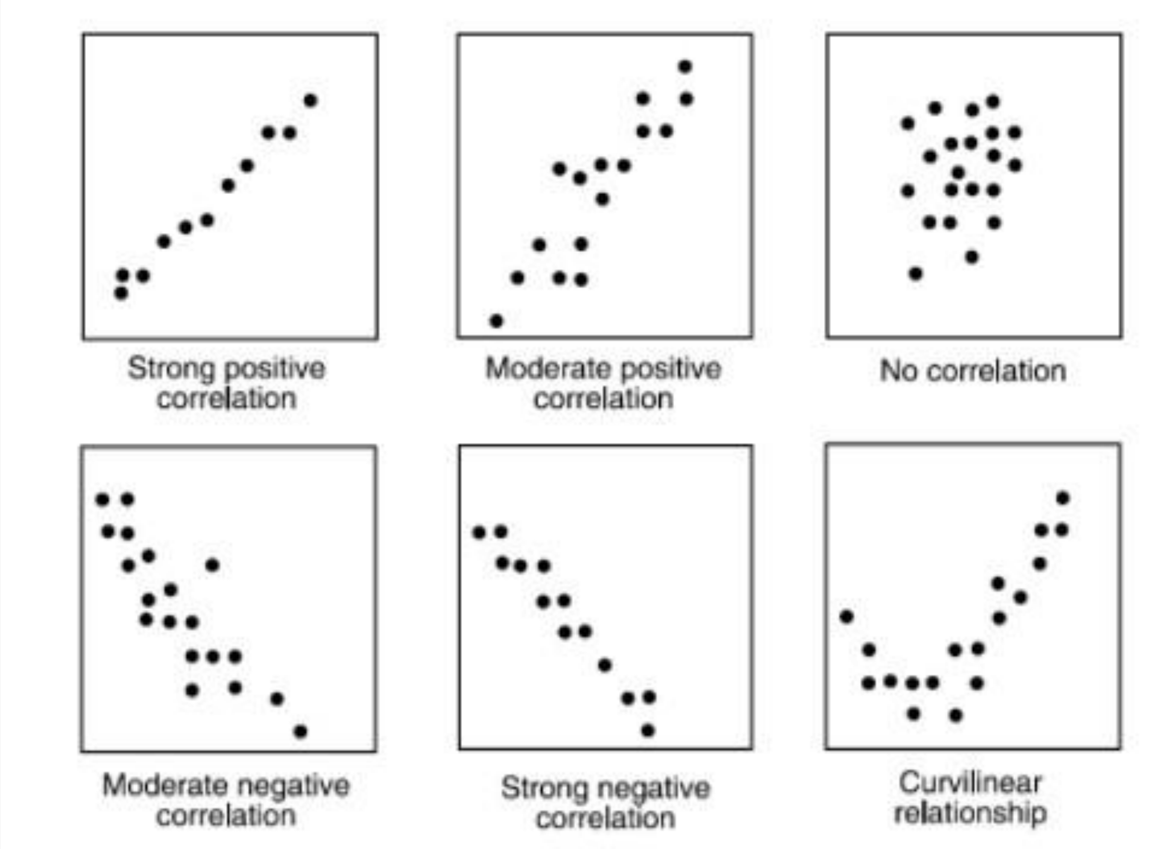
# Univariate Analysis

- Categorical Variable



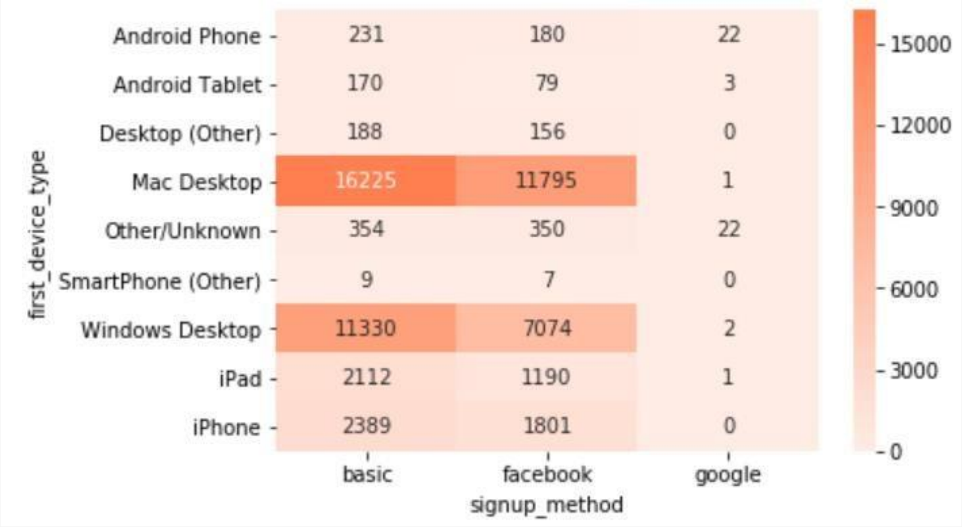
# Bi-variate Analysis

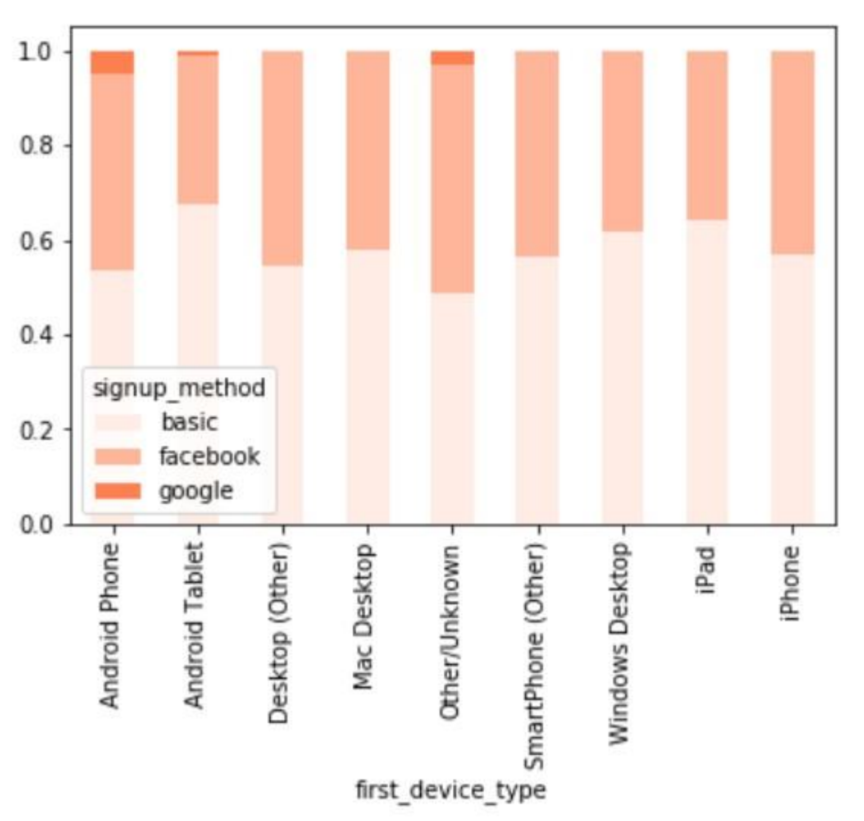
- Continuous & Continuous



# Bi-variate Analysis

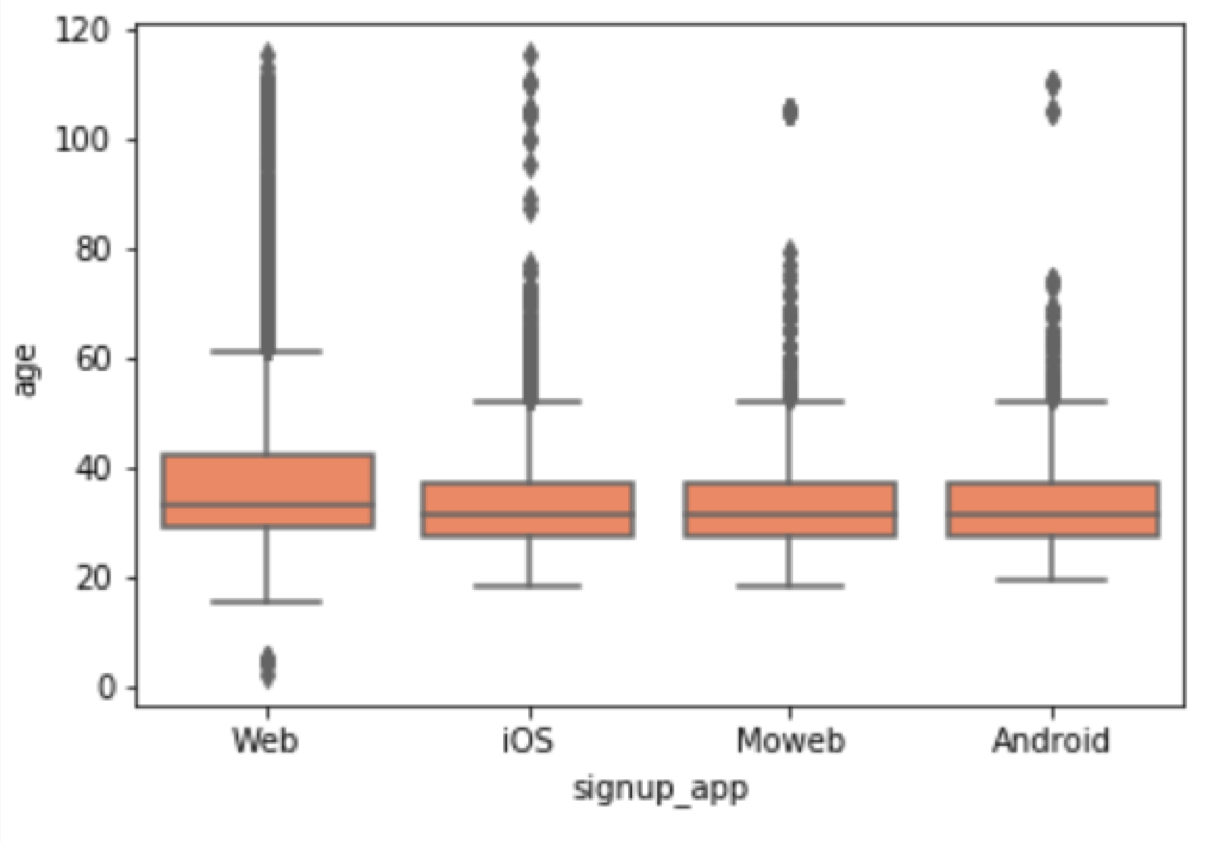
- Categorical & Categorical



# Bi-variate Analysis

- Continuous & Categorical



# Why Data Have Missing Values?

1. **Missing Completely at Random (MCAR)**

   어떤 변수의 결측치가 완전히 무작위로 발생한 경우

2. **Missing at Random (MAR)**

   어떤 변수의 결측의 여부가 자료 내의 다른 변수와 관련이 있는 경우
   (예: 학업 점수의 결측 여부가 소득 수준과 관련 있을 때)

3. **Missing not at Random, Non-Ignorable (MNAR)**

   어떤 변수의 결측의 여부가 해당 변수와 관련이 있는 경우
   (예: 학업 점수가 낮은 학생들이 학업 점수에 응답하지 않음)

4. **Structually Missing Data**

Missing Values Treatment

# How To Handle Missing Values?

1 Deletion

| date_first_booking | gender | age |
| --- | --- | --- |
| ~~2010-08-03~~ | ~~-unknown-~~ | ~~22~~ |
| 2010-08-03 | MALE | 38 |
| 2010-08-02 | FEMALE | 56 |
| 2012-09-08 | FEMALE | 42 |
| ~~2010-02-18~~ | ~~-unknown-~~ | ~~41~~ |
| ~~2010-01-02~~ | ~~-unknown-~~ | ~~22~~ |
| 2010-01-05 | FEMALE | 46 |
| 2010-01-13 | FEMALE | 47 |
| 2010-07-29 | FEMALE | 50 |

| date_first_booking | gender | age |
| --- | --- | --- |
| 2010-08-03 | -unknown- | 22 |
| 2010-08-03 | MALE | 38 |
| 2010-08-02 | FEMALE | 56 |
| 2012-09-08 | FEMALE | 42 |
| 2010-02-18 | -unknown- | 41 |
| 2010-01-02 | -unknown- | 22 |
| 2010-01-05 | FEMALE | 46 |
| 2010-01-13 | FEMALE | 47 |
| 2010-07-29 | FEMALE | 50 |

# How To Handle Missing Values?

## 2 Heuristic Imputation

| Name | Sex | Survived |
|---|---|---|
| Mr. Owen | Male | F |
| **Mrs. Bradley** | | T |
| Miss. Laina | Female | T |
| Mrs. Jarques | Female | F |
| **Mr. William** | | F |
| Mr. James | Male | T |

| Name | Sex | Survived |
|---|---|---|
| Mr. Owen | Male | F |
| Mrs. Bradley | Female | T |
| Miss. Laina | Female | T |
| Mrs. Jarques | Female | F |
| Mr. William | Male | F |
| Mr. James | Male | T |

# How To Handle Missing Values?

3 Mean/Median/Mode

Generalized Imputation

모든 사람의 나이 평균: 35

Case Imputation

Female 나이 평균: 30
Male 나이 평균 : 40

# How To Handle Missing Values?

4 Prediction Model

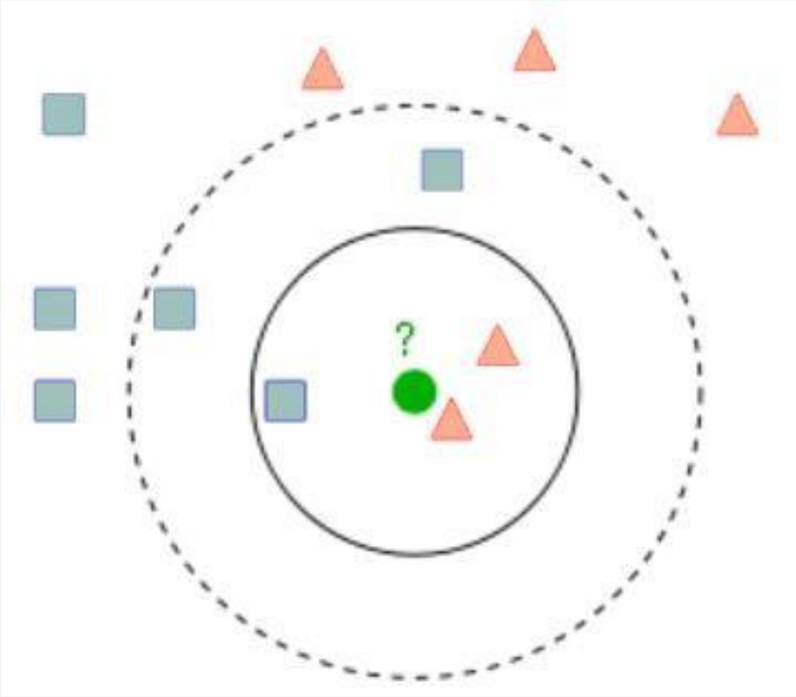
| timestamp_first_active | date_first_booking | gender | age |
|---|---|---|---|
| 20090319043255 | 2010-08-03 | -unknown- | 22 |
| 20090523174809 | 2010-08-03 | MALE | 38 |
| 20090609231247 | 2010-08-02 | FEMALE | 56 |
| 20091031060129 | 2012-09-08 | FEMALE | 42 |
| 20091208061105 | 2010-02-18 | -unknown- | 41 |
| 20100101215619 | 2010-01-02 | -unknown- | 22 |

train

test

# How To Handle Missing Values?

## 5 KNN Imputation




결측치가 있는 변수별로 모델을 짤 필요가 없음
데이터간의 연관성을 고려함

시간이 오래 걸림
K 정하기 어려움

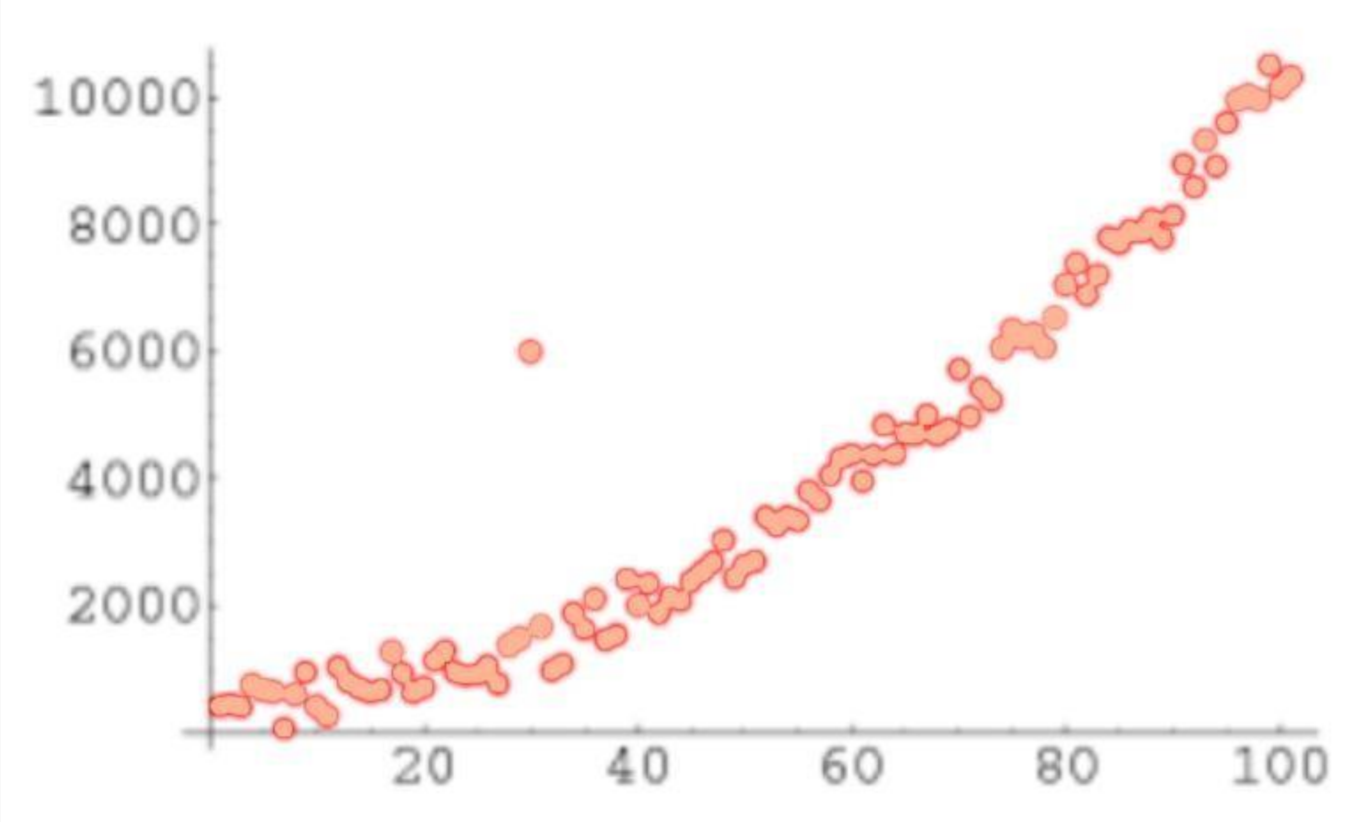
# How To Handle Missing Values?

6 Model Itself!

## Example : Tree Based Models

https://stats.stackexchange.com/questions/96025/how-do- decision-treelearning-algorithms-deal-with-missing- values-under-the-hoo

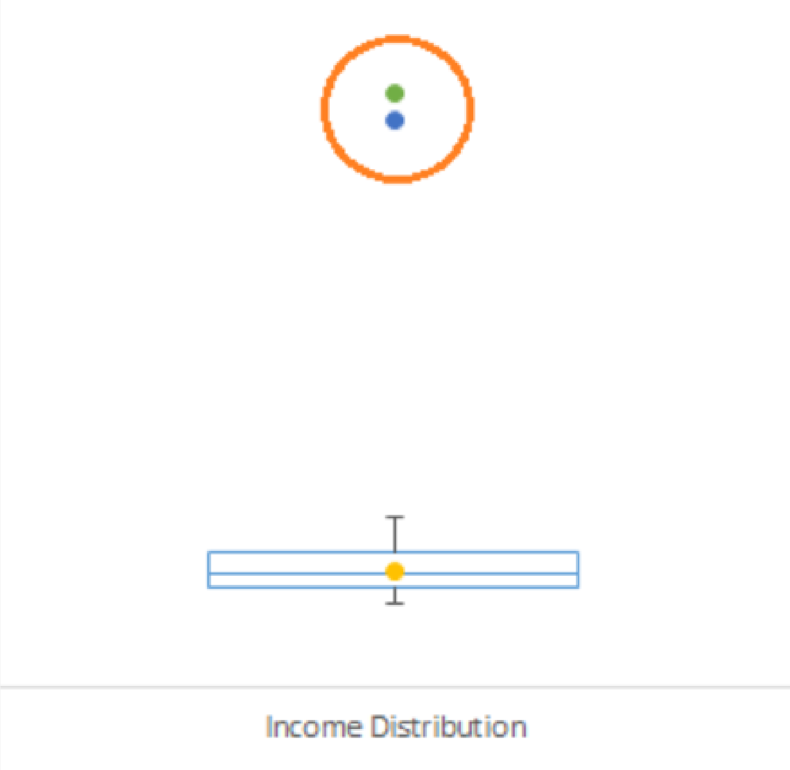
# What is an Outlier?

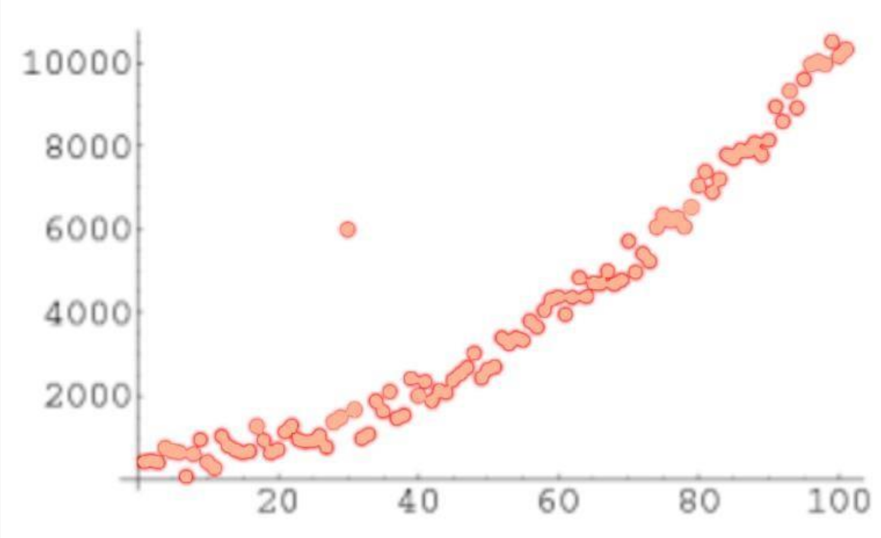


Outlier Treatment

# What is an Outlier?

| Univariate | Bivariate |
| --- | --- |



# Check Outliers

| | |
|---|---|
| 사분위수(IQR) | 하한선: Q1 – 1.5 * 사분위수 범위<br>상한선: Q3 + 1.5 * 사분위수 범위 |
| Capping Method | 백분위 수에서 5% ~ 95% 를 벗어나 있는 값 |
| Data Points | 평균으로부터 3 표준편차 이상 벗어난 경우 |

## Causes of Outliers

- Data Entry Error
- Measurement Error
- Experimental Error
- Intentional Error
- Data Processing Error
- Sampling Error
- Natural Outliers

Outlier Treatment

# How to Handle Outlier

1 Deletion

| Height(cm) |
|---|
| 180 |
| 165 |
| 159 |
| 191 |
| ~~62~~ |
| 173 |
| 175 |
| ~~302~~ |

2 Capping

| 연봉(만원) |
|---|
| 3,000 |
| 4,000 |
| 5,000 |
| 4,500 |
| 5,500 |
| 7,000 |
| 50,000 |
| ~~250,000~~ |

→

| 연봉(만원) |
|---|
| 3,000 |
| 4,000 |
| 5,000 |
| 4,500 |
| 5,500 |
| 7,000 |
| 10,000 |
| 10,000 |

# How to Handle Outlier

## 3 Assign New Value

| Height(cm) |
| --- |
| 180 |
| 165 |
| 159 |
| 191 |
| ~~62~~ |
| 173 |
| 175 |
| ~~302~~ |

→

| Height(cm) |
| --- |
| 180 |
| 165 |
| 159 |
| 191 |
| NA |
| 173 |
| 175 |
| NA |

Mean, median, mode…

# How to Handle Outlier

## 4 Transformation

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 100 |

Log10

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.3 | 0.47 | 0.6 | 0.69 | 0.7 | 0.84 | 0.9 | 0.95 | 1 | 2 |

# Model Itself!

## Example : Tree Based Models

https://www.quora.com/Why-are-tree-based-models-robust-to-outliers

# Feature Selection vs. Feature Extraction

## 어떤 Feature가 유용한가?

## 차원 축소의 효과

| Selection | Extraction |
|---|---|
| • 전체 특징의 부분집합을 선택해서 간결하게 만드는 것<br>• Domain Knowledge에 의한 직접 선택<br>• 자동 특징 선택 | • 고차원의 원본 feature공간을 저차원의 새 feature 공간으로 투영<br>• 원본 feature 공간의 선형 or 비선형 결합<br>• (예) PCA |

출처: http://terryum.io/korean/2016/05/05/FeatureSelection_KOR/

Feature Engineering

# Variable Transformation & Variable Creation

## Variable Transformation

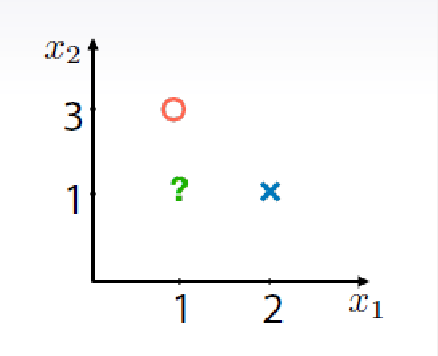
## Variable Creation

| ML model | Assumptions | Advantages | Disadvantages | Feature Scaling | Missing Data | Outliers | Suitable for | Learning | Example Use |
|---|---|---|---|---|---|---|---|---|---|
| Naïve Bayes Classifier | Features are independent | • Performs well with categorical variables<br>• Converges faster: less training time<br>• Good with moderate to large training data sets<br>• Good when dataset contains several features | • Correlated features affect performance | No | Can handle missing data (it ignores missing data) | Robust to outliers | • Classification<br>• Multiclass classification | Supervised | • Sentiment Analysis<br>• Document categorisation<br>• Email Spam Filtering |
| Support Vector Machine (SVM) | None | • Good for datasets with more variables than observations<br>• Good performance<br>• Good of-the-shelf model in general for several scenarios<br>• Can approximate complex non-linear functions | • Long training time required<br>• Tuning is required to determine which kernel is optimal for non-linear SVMs | Yes | Sensitive | Robust to outliers | • Classification<br>• Regression | Supervised | • Stock market forecasting<br>• Value at risk determination |
| Linear Regression | Linear relation between features and target | • Interpretability<br>• Little tuning | • Correlated features may affect performance<br>• Extensive feature engineering required | Yes | Sensitive | Sensitive | Regression | Supervised | • Sales forecasting<br>• House pricing |
| Logistic Regression | Linear relation between features and the log odds | • Interpretability<br>• Little tuning | • Correlated features may affect performance<br>• Extensive feature engineering required | Yes | Sensitive | Potentially sensitive | Classification | Supervised | • Risk Assessment<br>• Fraud Prevention |
| Classification and Regression Trees | None | • Interpretability<br>• Render feature importance<br>• Saves on data preparation | • Do not fit well to continuous variables<br>• It does not predict beyond the range of the response values in the training data.<br>• Not very accurate<br>• Overfits | No | No | Robust to outliers | • Classification<br>• Regression | Supervised | • Risk Assessment<br>• Fraud Prevention |
| Random Forests | None | • Interpretability<br>• Render feature importance<br>• Saves on data preparation<br>• Does not overfit<br>• Good performance /accuracy<br>• Robust to noise<br>• Little if any parameter tuning required<br>• Apt at almost any machine learning problem | • It does not predict beyond the range of the response values in the training data<br>• Biased towards categorical variables with several categories<br>• Biased in multiclass problems toward more frequent classes | No | No | Robust to outliers | • Classification<br>• Regression | Supervised | • Credit Risk Assessment<br>• Predict breakdown of a mechanical parts (automobile industry).<br>• Assess probability of developing a chronic disease (healthcare)<br>• Predicting the average number of social media shares |
| Gradient Boosted Trees | None | • Great performance<br>• Apt at almost any machine learning problem<br>• It can approximate most non-linear function | • Prone to overfit<br>• Needs some parameter tuning | No | No | Robust to outliers | • Classification<br>• Regression | Supervised | |
| K-nearest neighbours | None | • Good performance | • Slow when predicting<br>• Susceptible to high dimension (lots of features) | Yes | Sensitive | Robust to outliers | • Classification<br>• Regression | Supervised | • Gene expression<br>• Protein-protein interaction<br>• Content retrieval (of webpages for example) |
| AdaBoost | None | • It doesn't overfit easily<br>• Few parameters to tune | | No | Can handle | Sensitive | • Classification<br>• Regression | Supervised | |
| Neural Networks | None | • Can approximate any function<br>• Great Performance | • Long training time<br>• Several parameters to tune, including neuronal architecture<br>• Prone to overfit<br>• Little interpretability | Yes | Sensitive | Can handle outliers, and it affects performance if they are too many | • Classification<br>• Regression | Supervised | |
| K-Means Clustering | • clusters are spherical<br>• clusters are of similar size | • Fast training | • Need to determine k, the number of clusters<br>• Sensitive to initial points and local optima | Yes | | Sensitive | • Segmentation | Unsupervised | |
| Hierarchical clustering | | • No a priori information about the number of clusters required | • Final number of clusters to be decided by the scientist<br>• Slow training | Yes | Sensitive | Sensitive | • Segmentation | Unsupervised | |
| PCA | • Correlation among features | | | Yes | Sensitive | Sensitive | | | |

ML Model Cheat Sheet
첨부한 자료를 참고해주세요

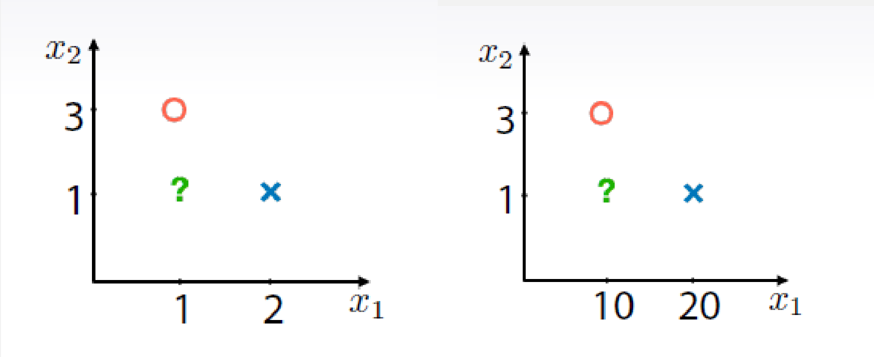
# Variable Transformation

## Centralizing & Scaling



# Variable Transformation

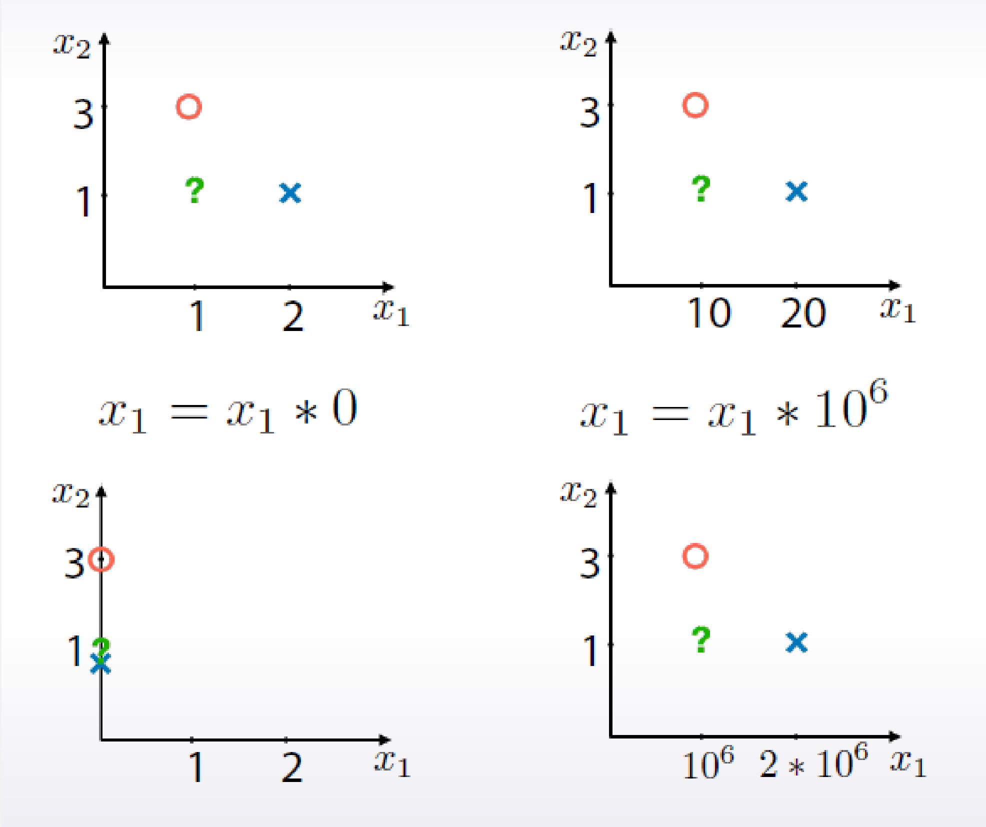
## Centralizing & Scaling



Feature Engineering

# Variable Transformation

## Centralizing & Scaling



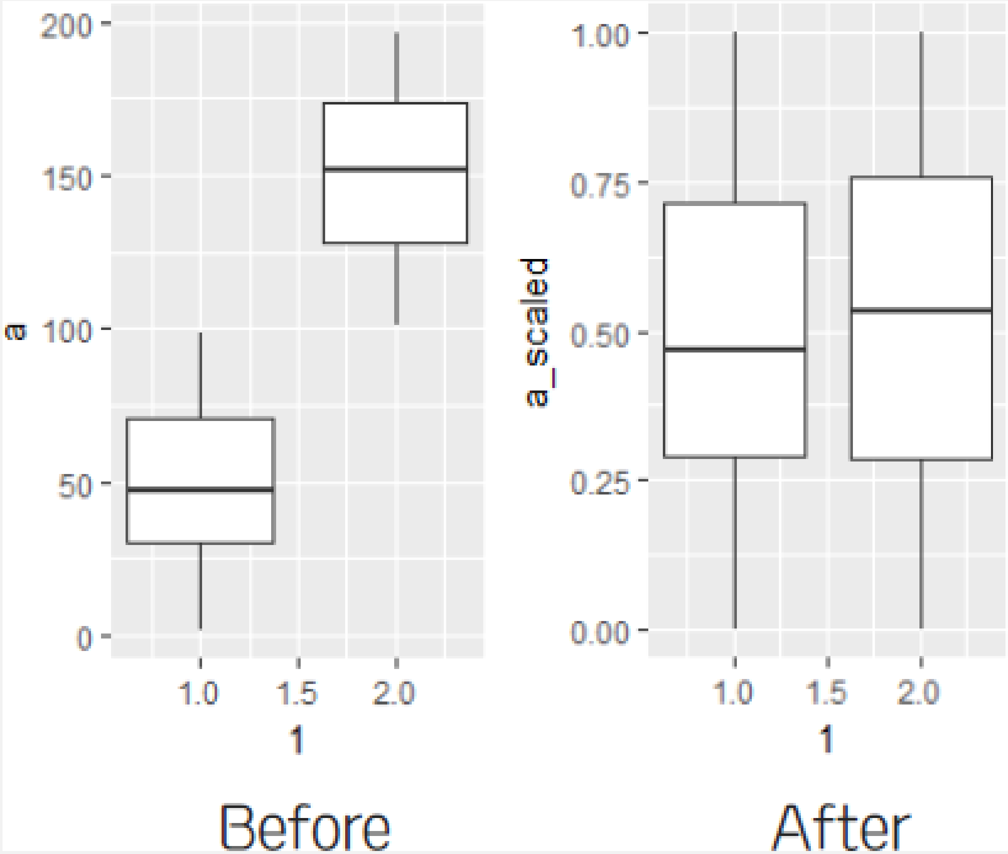
# Variable Transformation

Centralizing & Scaling

## Min-Max Scaling

{X-min(X)} / {max(X)-min(X)}

To [0,1]



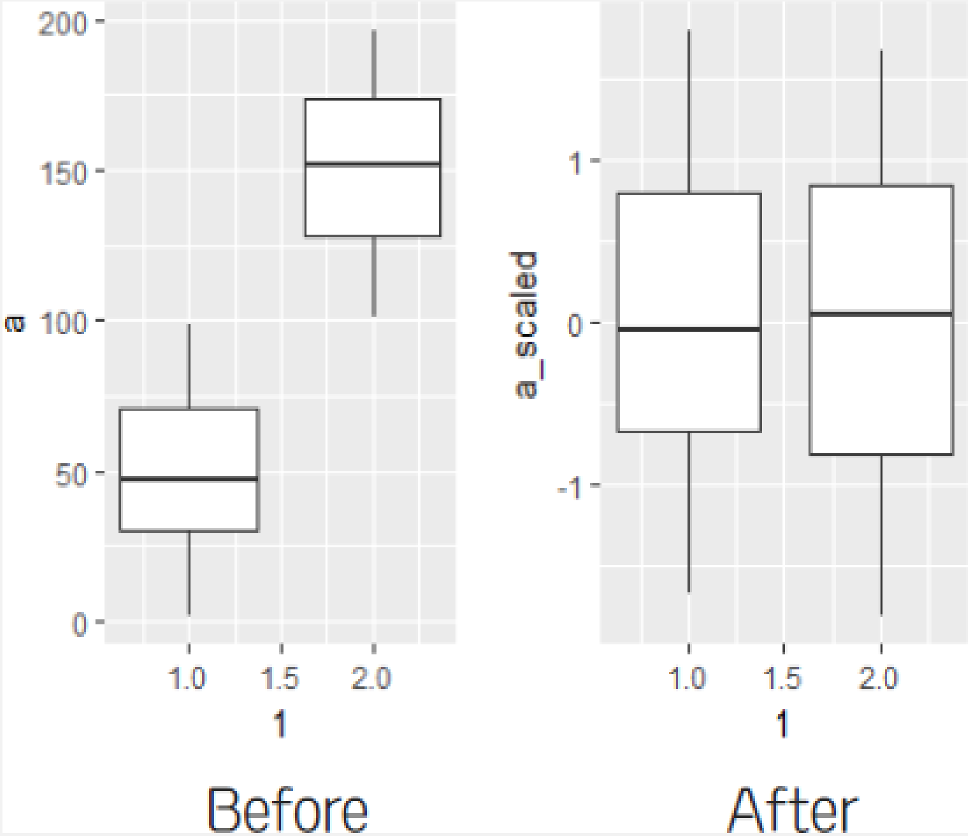
# Variable Transformation

Centralizing & Scaling

## Standardization

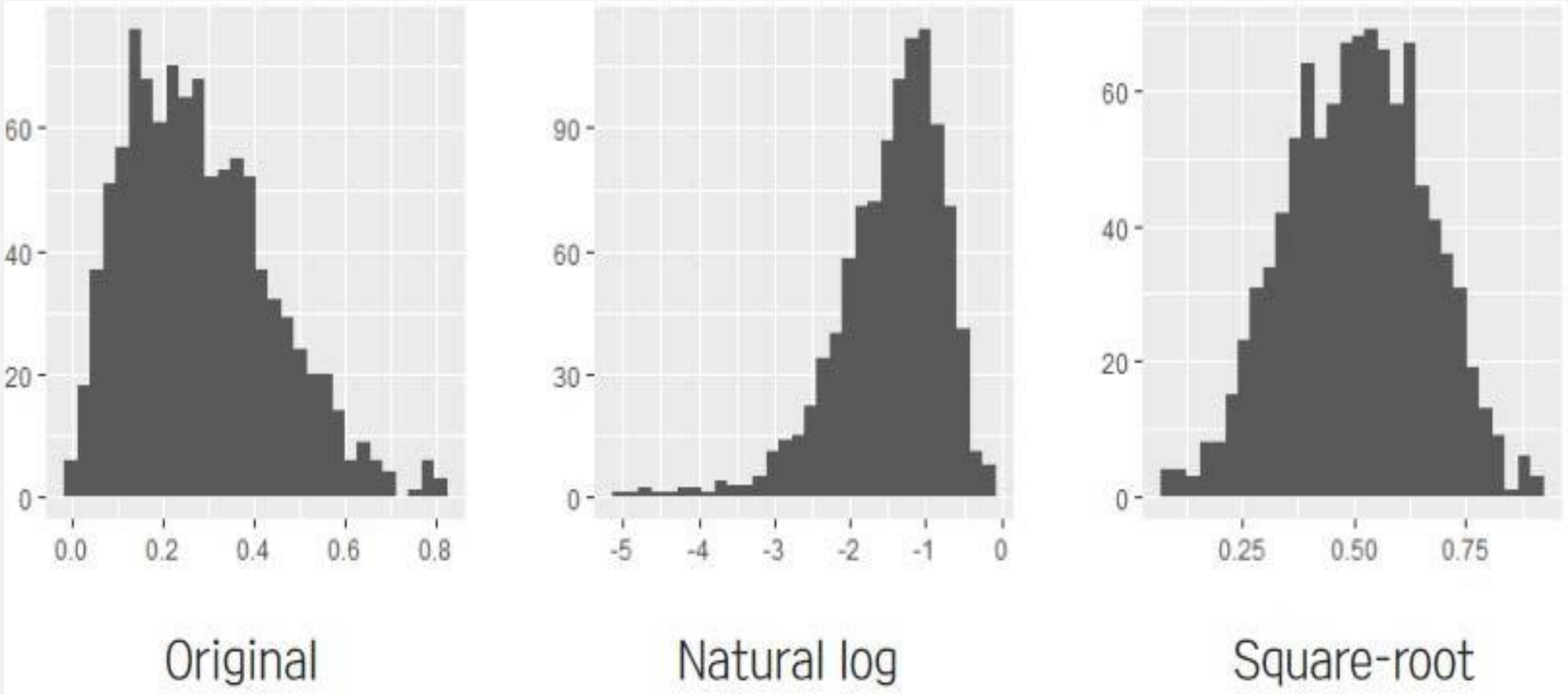
{X-mean(X)} / std(X)

Mean = 0, std = 1



# Variable Transformation

Symmetrizing



# Variable Transformation

## Binning

| Age |
| --- |
| 56 |
| 42 |
| 41 |
| 46 |
| 37 |
| 50 |
| 46 |

| Age (Transformed) |
| --- |
| 50대 |
| 40대 |
| 40대 |
| 40대 |
| 30대 |
| 50대 |
| 40대 |

# Variable Transformation

## Top/Bottom/Zero coding

| Age | | Age (Transformed) |
|---|---|---|
| 56 | → | 50세 이상 64세 이하 |
| 17 | | 25세 미만 |
| 41 | | 35세 이상 49세 이하 |
| 29 | | 25세 이상 34세 이하 |
| 18 | | 25세 미만 |
| 68 | | 65세 이상 |
| 34 | | 25세 이상 34세 이하 |

# Variable Creation

| Date | sales |
| --- | --- |
| 2019.01.01 | 1744 |
| 2019.01.02 | 1332 |
| 2019.01.03 | 922 |
| 2019.01.04 | 2448 |
| 2019.01.05 | 1864 |
| 2019.01.06 | 1760 |

# Variable Creation

| Date | Weekday | sales |
|---|---|---|
| 2019.01.01 | Tue | 1744 |
| 2019.01.02 | Wed | 1332 |
| 2019.01.03 | Thu | 922 |
| 2019.01.04 | Fri | 2448 |
| 2019.01.05 | Sat | 1864 |
| 2019.01.06 | Sun | 1760 |

Feature Engineering

# Variable Creation

| Date | Weekday | Daynumber_since_year_2019 | is_Holiday | Days_till_holidays | sales |
|---|---|---|---|---|---|
| 2019.01.01 | Tue | 0 | T | 0 | 1744 |
| 2019.01.02 | Wed | 1 | F | 4 | 1332 |
| 2019.01.03 | Thu | 2 | F | 3 | 922 |
| 2019.01.04 | Fri | 3 | F | 2 | 2448 |
| 2019.01.05 | Sat | 4 | F | 1 | 1864 |
| 2019.01.06 | Sun | 5 | T | 0 | 1760 |

# Variable Creation

## One-hot Encoding

| Device |
|---|
| Windows |
| Mac |
| Mac |
| Mac |
| Windows |

| Windows | Mac |
|---|---|
| 1 | 0 |
| 0 | 1 |
| 0 | 1 |
| 0 | 1 |
| 1 | 0 |

**문자형 Categorical Variable의 경우, 컴퓨터가 인식할 수 있도록 숫자형으로 바꿔주는 방법**
**가장 많이 사용되는 방법 중 하나입니다.**

# Variable Creation

## One-hot Encoding

- 차원 증가의 문제
- 유사도 표현 불가

| 리트리버 | 웰시코기 | 냉장고 | 독수리 |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 |

# Variable Creation

## Feature Interaction

The integration of two features would modify the behavior of one or both features

| shot_id | Min_remaining | Sec_remaining | Time_remaining |
|---|---|---|---|
| 0 | 10 | 27 | 627 |
| 1 | 10 | 22 | 622 |
| 2 | 7 | 45 | 465 |
| 3 | 6 | 52 | 412 |
| 4 | 6 | 19 | 379 |

# Variable Creation

같은 성격의 Indicator Variable이 많을 때?

| MedicalKeyword_1 | MedicalKeyword_2 | MedicalKeyword_3 | MedicalKeyword_4 | MedicalKeyword_5 | MedicalKeyword_6 | | MedicalKeyword_48 | MedicalKeyword_sum |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 1 | 0 | | 0 | 6 |
| 0 | 0 | 1 | 1 | 0 | 0 | | 0 | 8 |
| 0 | 0 | 1 | 1 | 0 | 0 | | 1 | 5 |
| 0 | 0 | 0 | 0 | 0 | 1 | … | 0 | 4 |
| 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 1 |
| 0 | 0 | 1 | 0 | 0 | 1 | | 1 | 6 |
| 0 | 1 | 1 | 0 | 0 | 0 | | 0 | 7 |
| 1 | 0 | 1 | 0 | 0 | 1 | | 1 | 12 |

Feature Engineering

# Variable Creation

| Age |
|---|
| 56 |
| 42 |
| 41 |
| NA |
| 37 |
| NA |
| 46 |

| Age | Age_is_null |
|---|---|
| 56 | 0 |
| 42 | 0 |
| 41 | 0 |
| NA | 1 |
| 37 | 0 |
| NA | 1 |
| 46 | 0 |

**결측 여부가 중요한 경우 / 그러나 차원 증가!**

# 그 밖의 FEATURE ENGINEERING?

**External Data**

기존 주어진 데이터 외의 다른 외부데이터를 활용해 성능을 높입니다

**Error Analysis**

모델을 통해 나온 결과를 바탕으로 특징을 만드는 방법

- Start with Lagrer Errors: 모든 값을 확인하기 보다 에러가 큰 feature부터 확인합니다
- Segment by classes: 평균 에러 값을 기준으로 segment를 나누어 비교, 분석합니다
- Unsupervised clustering: 패턴 발견에 어려움이 있을 경우, 비지도학습인 clustering 알고리즘을 사용하여 분류되지 않은 값을 확인

FINISH?

이제 데이터셋 하나가 완성됐습니다!

다시 처음으로 돌아가

다른 데이터셋도 만들어 봅시다

# Q&A

이제는 실습시간!

Jupyter Notebook을 이용해서 실습해봅시다