교육용 게임 이용패턴 분석을 통한 평가 정답률 예측

JYSHM

신현민, 오용석, 정세화, 최종엽

목차

공모전 소개

Featured Code Competition

2019 Data Science Bowl

Uncover the factors to help measure how young children learn

DATA SCIENCE BOWL
Passion. Curiosity. Purpose.

Presented by Booz Allen | Hamilton & kaggle

$160,000
Prize Money

Booz Allen Hamilton · 2,786 teams · 20 days to go (13 days to go until merger deadline)

Overview | Data | Notebooks | Discussion | Leaderboard | Rules | Team | My Submissions | Submit Predictions

Overview

- Description
- Evaluation
- Timeline
- Prizes
- Notebook Requirements
- About The DSB
- DSB Hosts & Partners

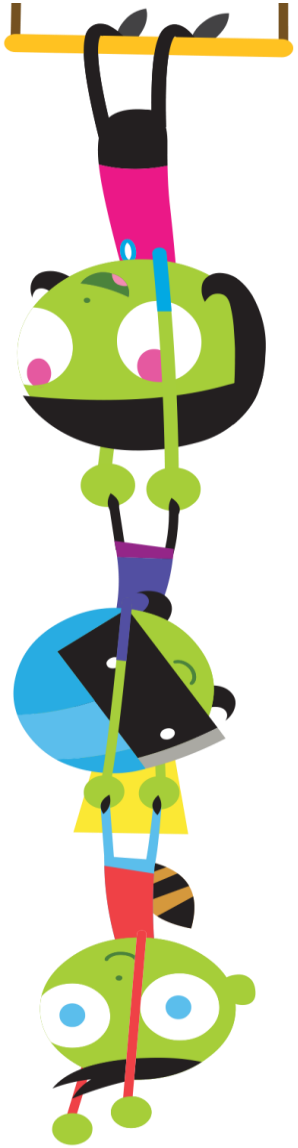
Illuminate Learning. Ignite Possibilities.

Uncover new insights in early childhood education and how media can support learning outcomes. Participate in our fifth annual Data Science Bowl, presented by Booz Allen Hamilton and Kaggle.

PBS KIDS, a trusted name in early childhood education for decades, aims to gain insights into how media can help children learn important skills for success in school and life. In this challenge, you'll use anonymous gameplay data, including knowledge of videos watched and games played, from the PBS KIDS Measure Up! app, a game-based learning tool developed as a part of the CPB-PBS Ready To Learn Initiative with funding from the U.S. Department of Education. Competitors will be challenged to predict scores on in-game assessments and create an algorithm that will lead to better-designed games and improved learning outcomes. Your solutions will aid in discovering important relationships between engagement with high-quality educational media and learning processes.

Data Science Bowl is the world's largest data science competition focused on social good. Each year, this competition gives Kagglers a chance to use their passion to change the world. Over the last four years, more than 50,000+ Kagglers have submitted over 114,000+ submissions, to improve everything from lung cancer and heart disease detection to ocean health.

For more information on the Data Science Bowl, please visit DataScienceBowl.com



연구목표

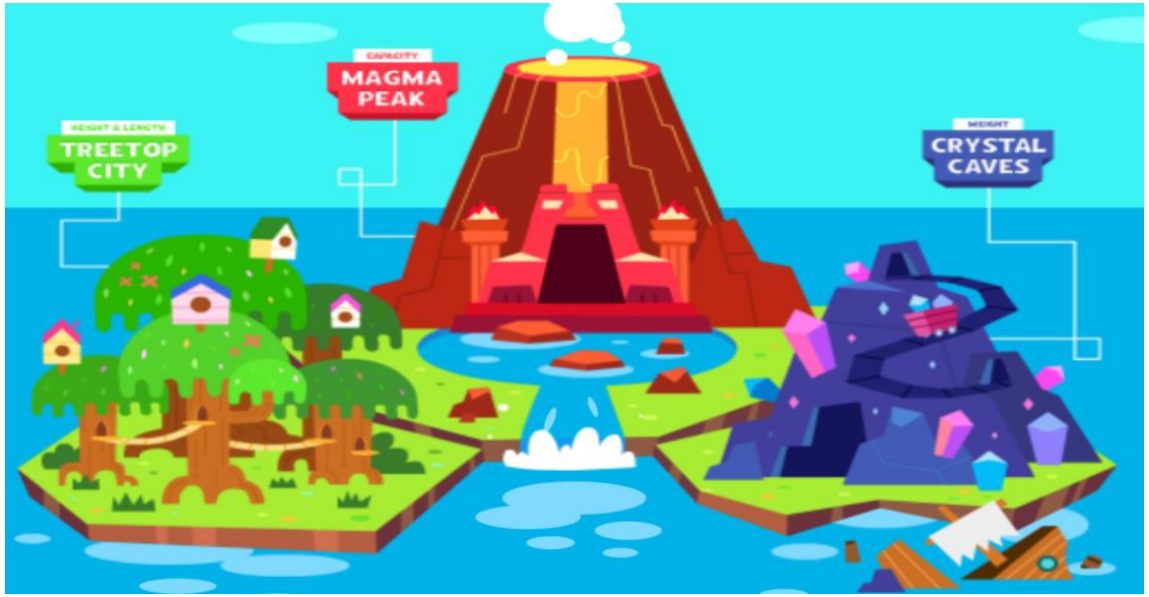
- 아이가 게임을 이용한 데이터를 사용하여 평가를 몇번만에 통과하는지 예측

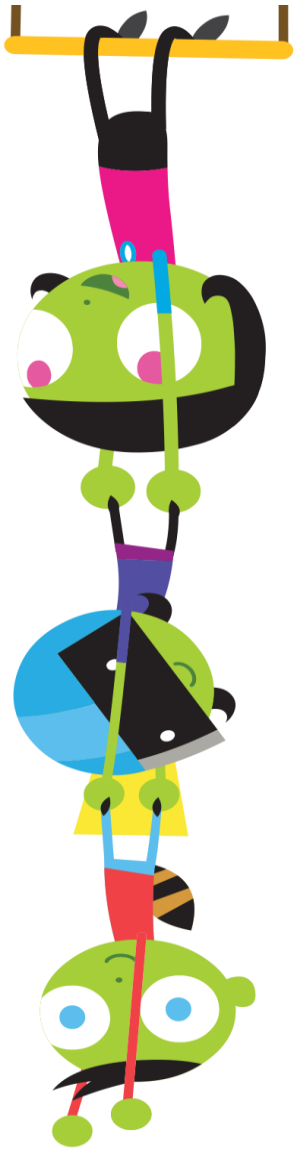
- 아래 4개의 점수 그룹으로 범주화(accuracy_group 데이터에 표시)

점수	내용
3점	첫 번째 시도에서 평가 해결
2점	두 번째 시도에서 평가 해결
1점	3회 이상의 시도 후 평가 해결
0점	평가 미해결

PBS KIDS Measure UP!

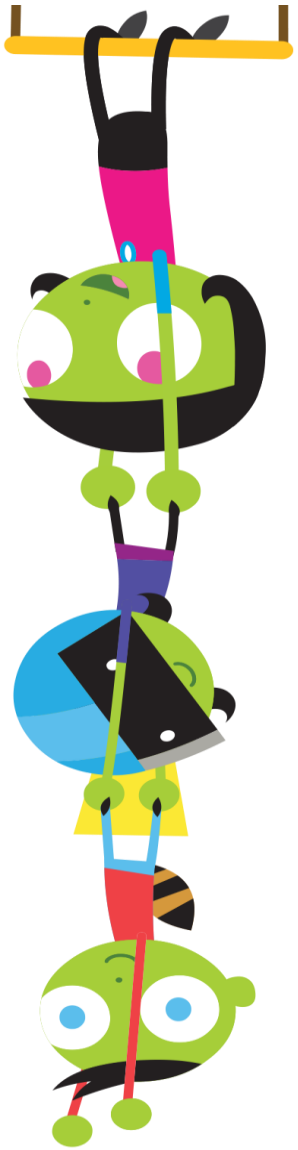
- 3~5세 아동이 게임을 통하여 길이, 너비, 용량 및 무게에 중점을 둔 초기 **STEM**(**S**cience, **T**echnology, **E**ngineering, **M**athmatics) 개념을 학습
- 총 3가지 맵이 존재 : **Tree Top**(길이와 높이),
  Crystal Caves(무게),
  Magma Peak(크기와 용량)
- 해당 애플리케이션은 각 맵의 주제와 관련된 **비디오**(Clip), **게임**(Game), **활동**(Activity), **평가**(Assessment)로 구성





PBS KIDS Measure UP!

Measure UP!



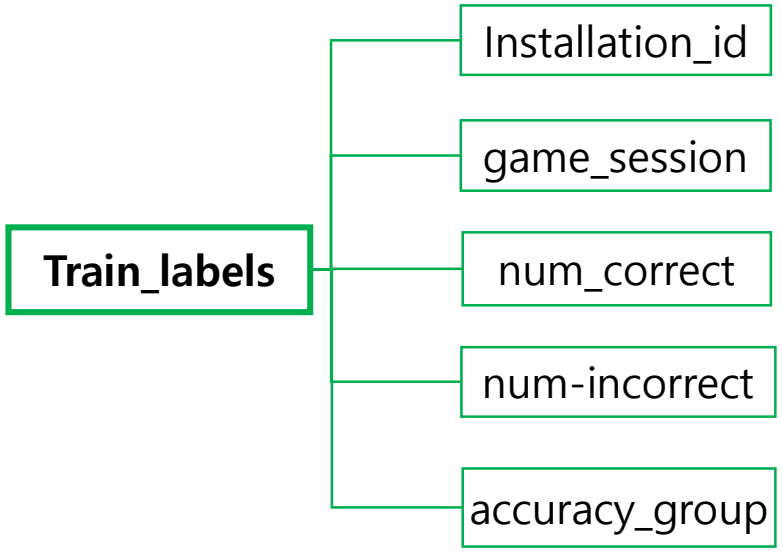
MAP

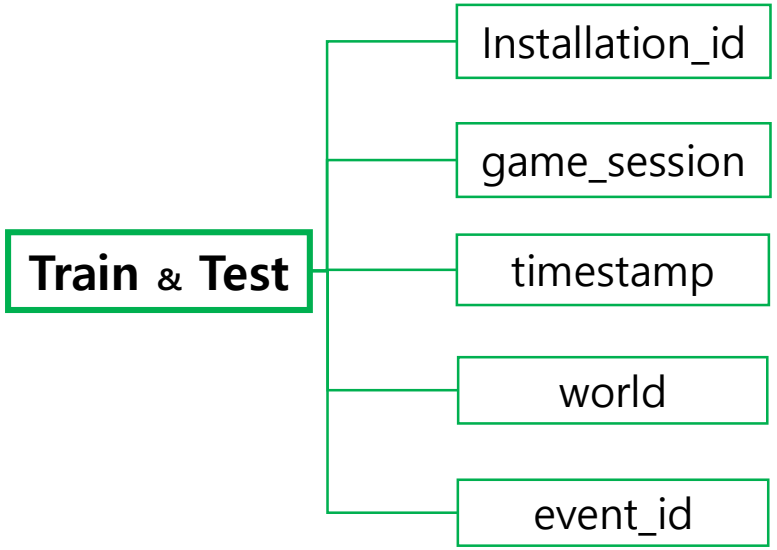
TREETOP CITY

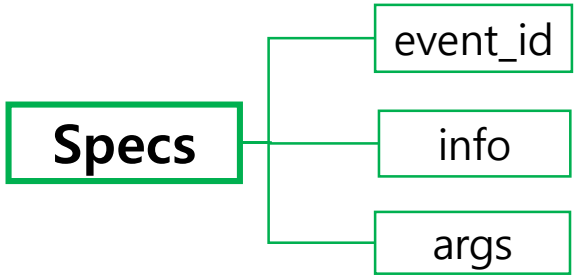
EDA

Data

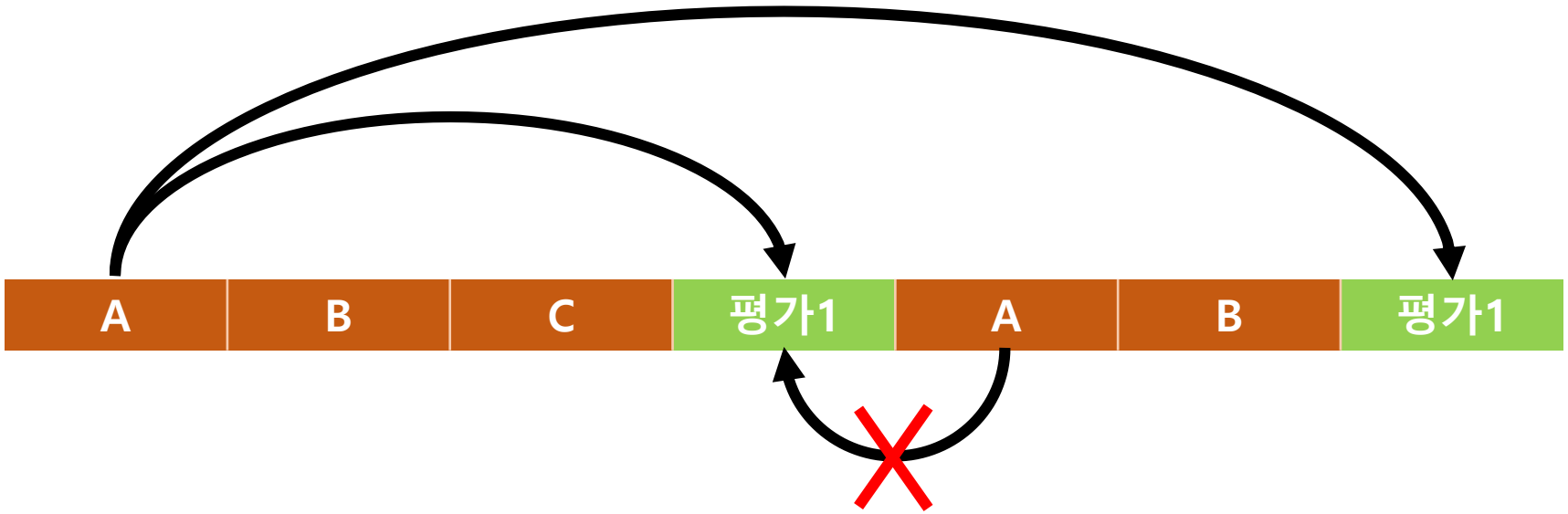
Data	내용
Train_labels	훈련세트의 평가에서 계산되는 방법을 보여주기 위한 데이터 파일
Specs	다양한 이벤트 유형의 스펙을 제공
Train	게임 플레이 이벤트를 포함하는 주요 데이터 파일
Test	게임 플레이 이벤트를 포함하는 주요 데이터 파일





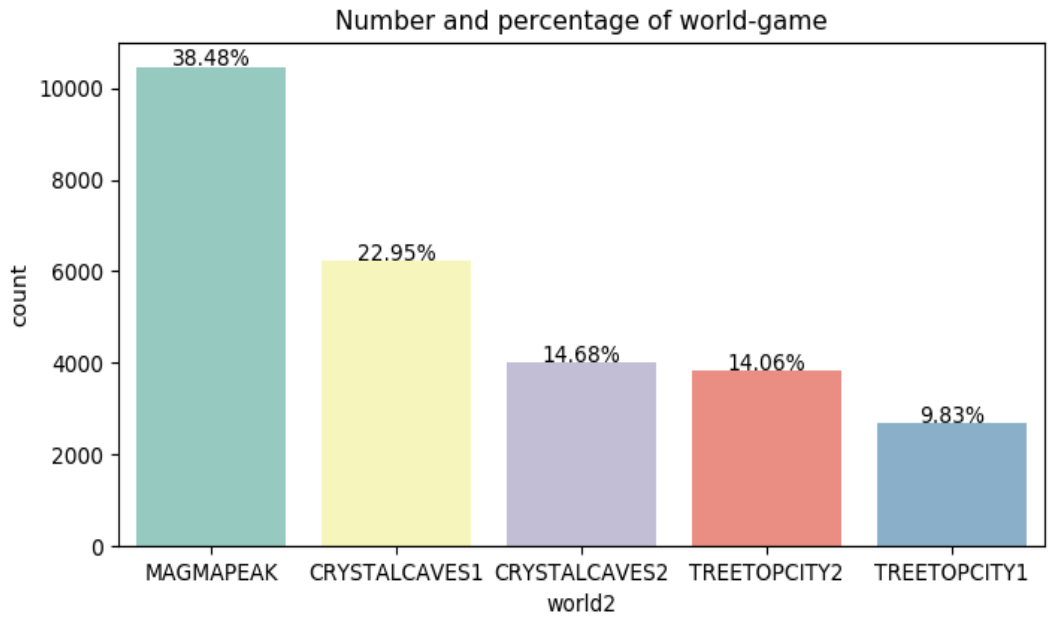


변수 생성 핵심 개념

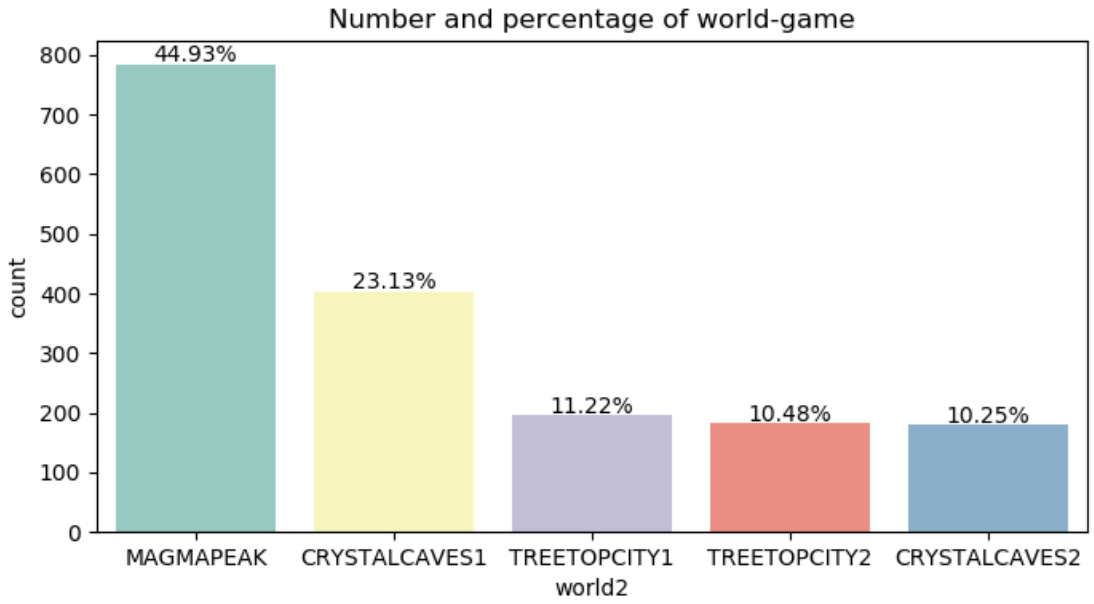


평가 이전의 컨텐츠만 영향을 준다.

해당 평가 별 Game 시행 횟수

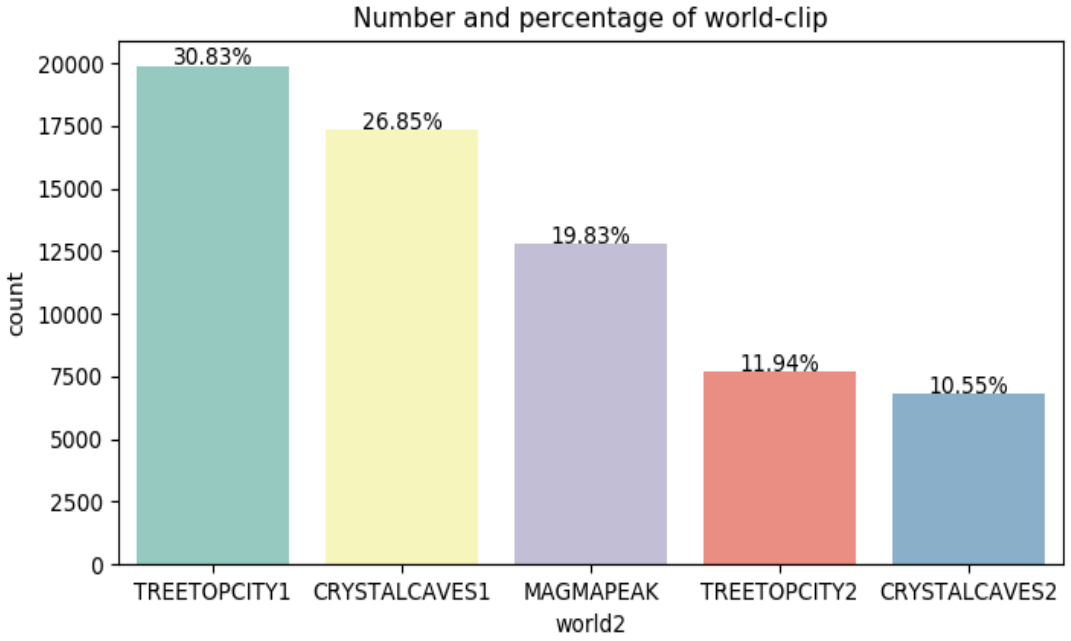


Train Data

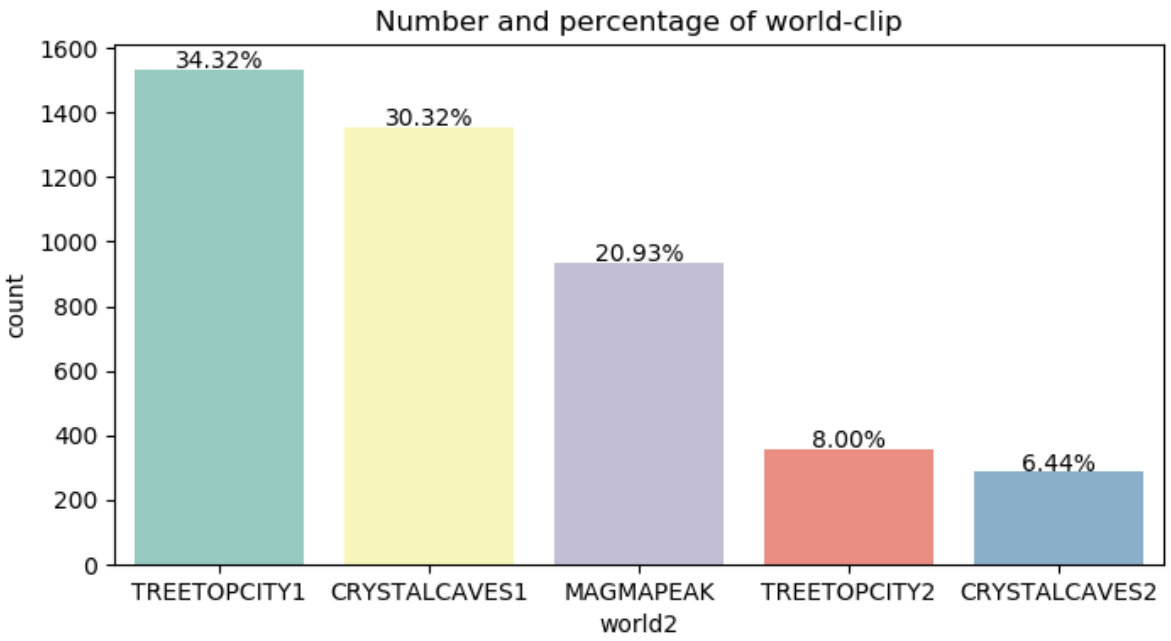


Test Data

해당 평가 별 Clip 시행 횟수

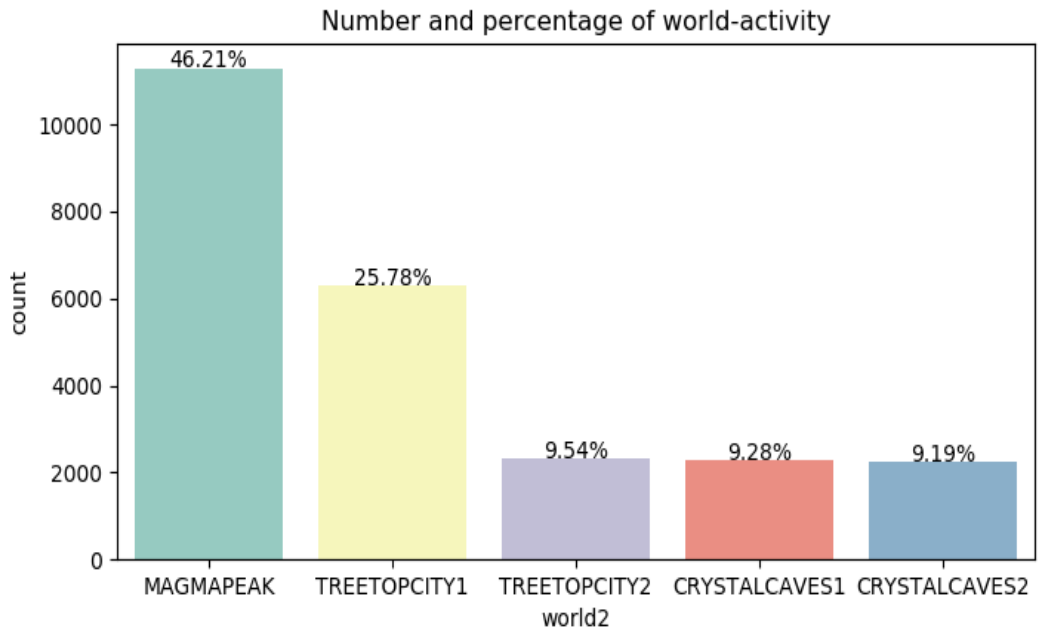


Train Data

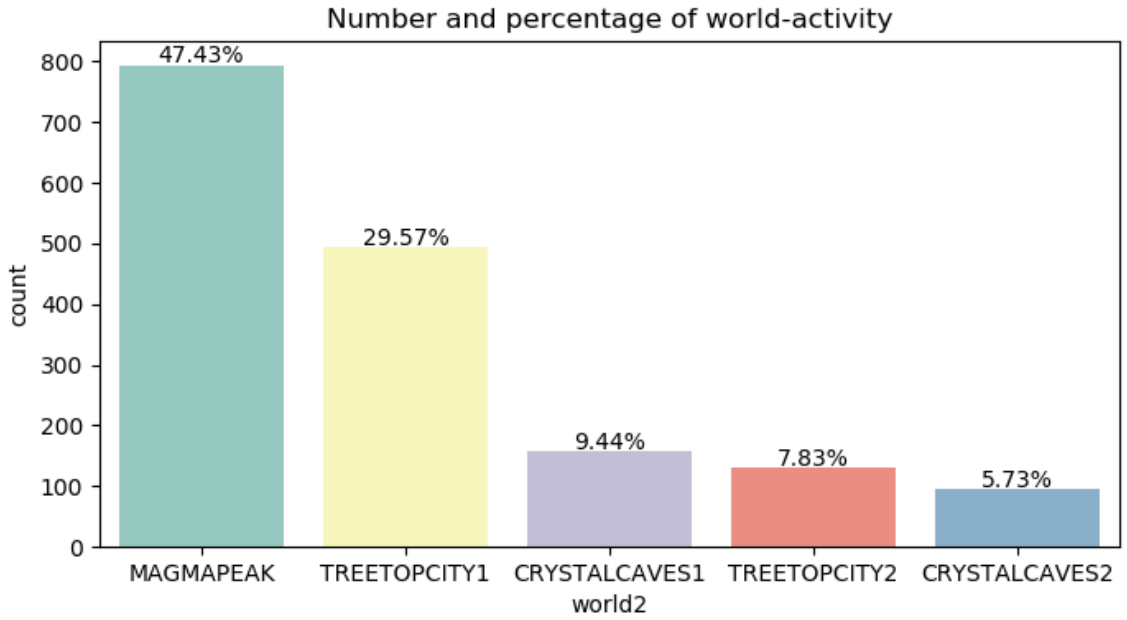


Test Data

해당 평가 별 Activity 시행 횟수

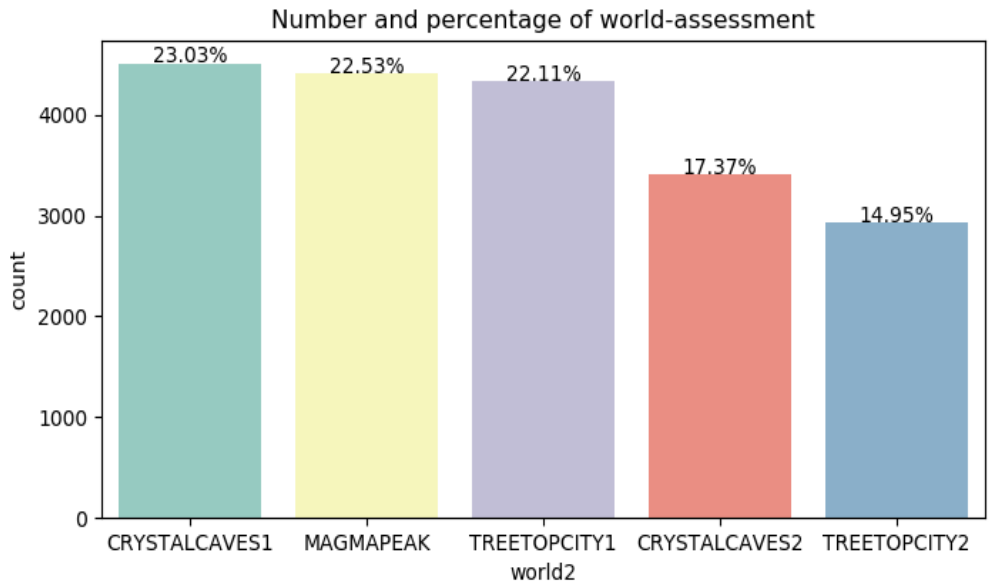


Train Data

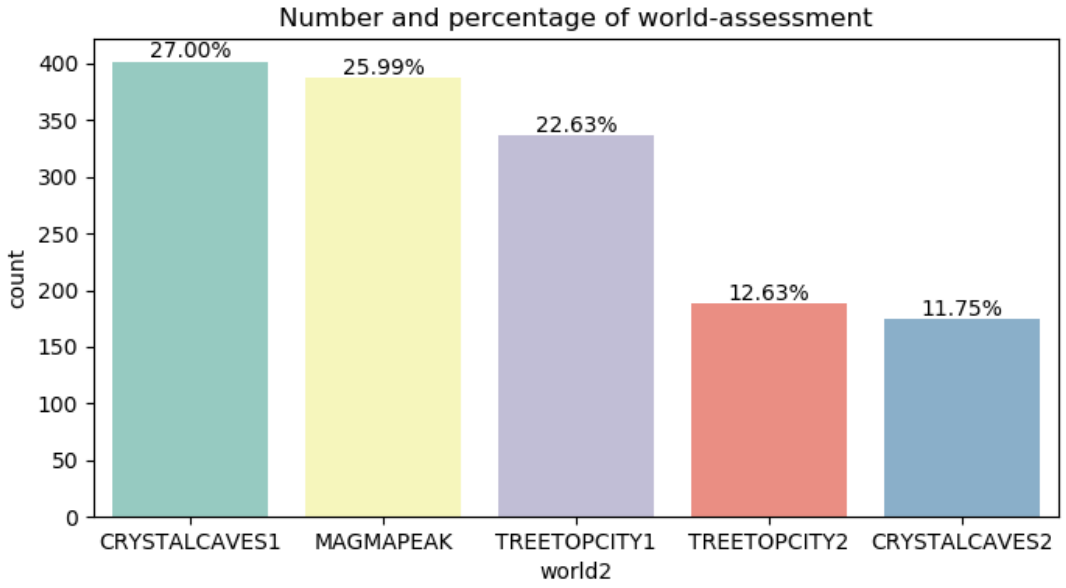


Test Data

World별 평가시도 횟수

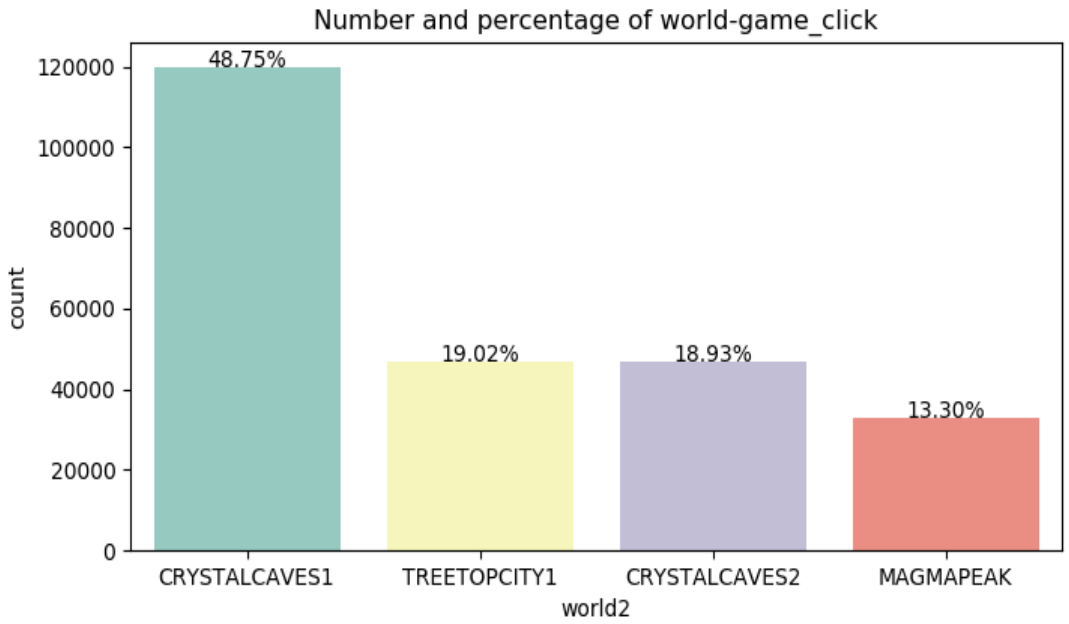


Train Data

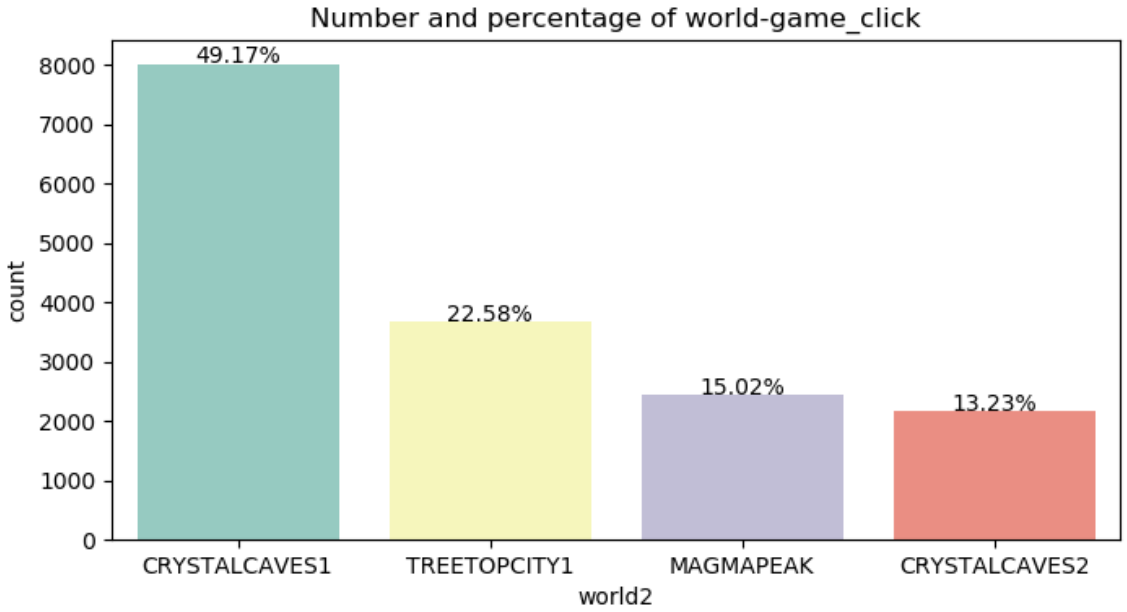


Test Data

World별 Game 정답클릭 횟수

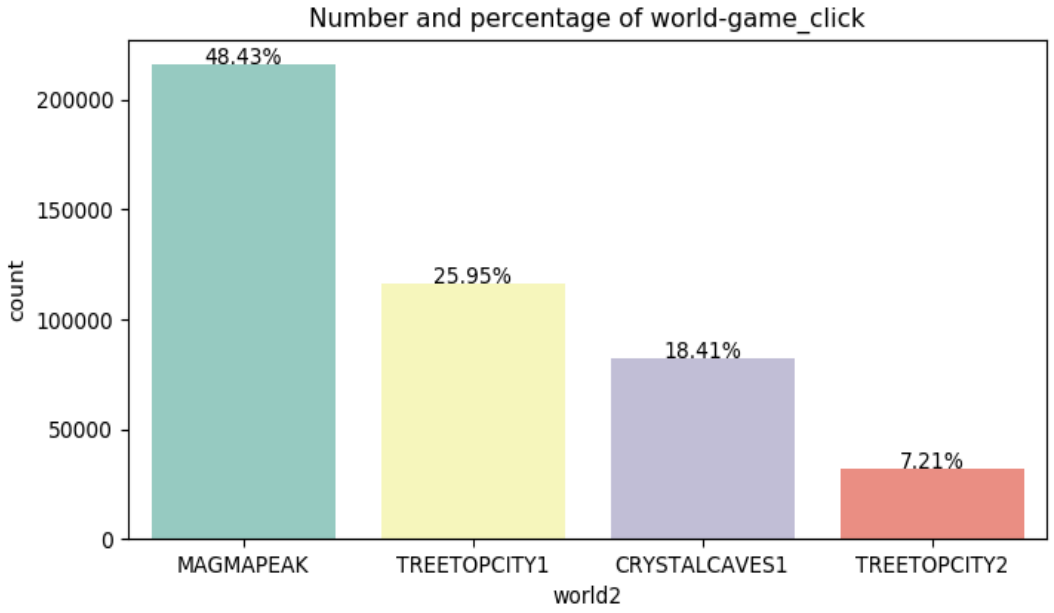


Train Data

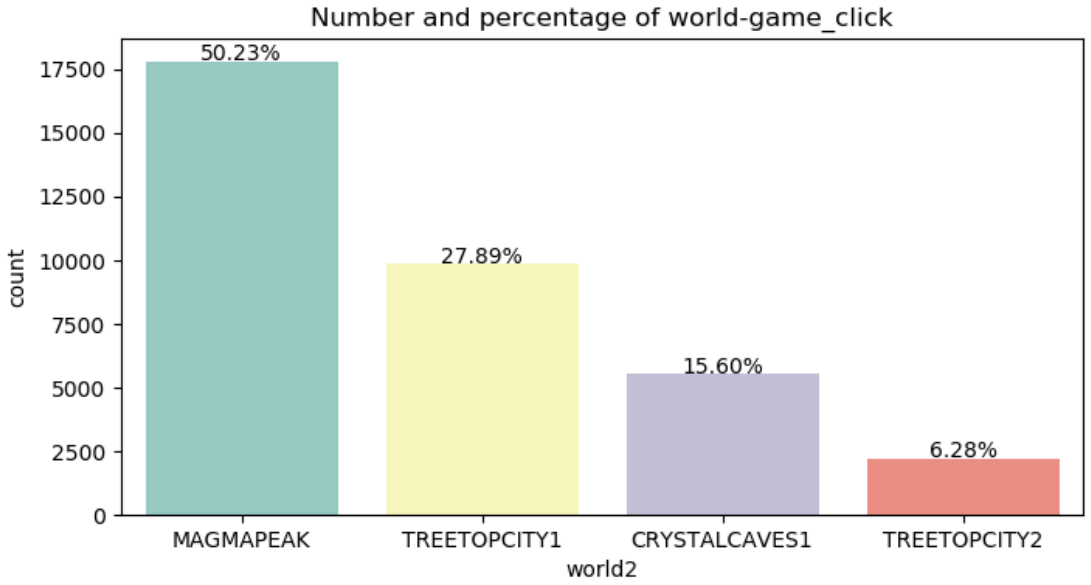


Test Data

World별 Activity 정답클릭 횟수

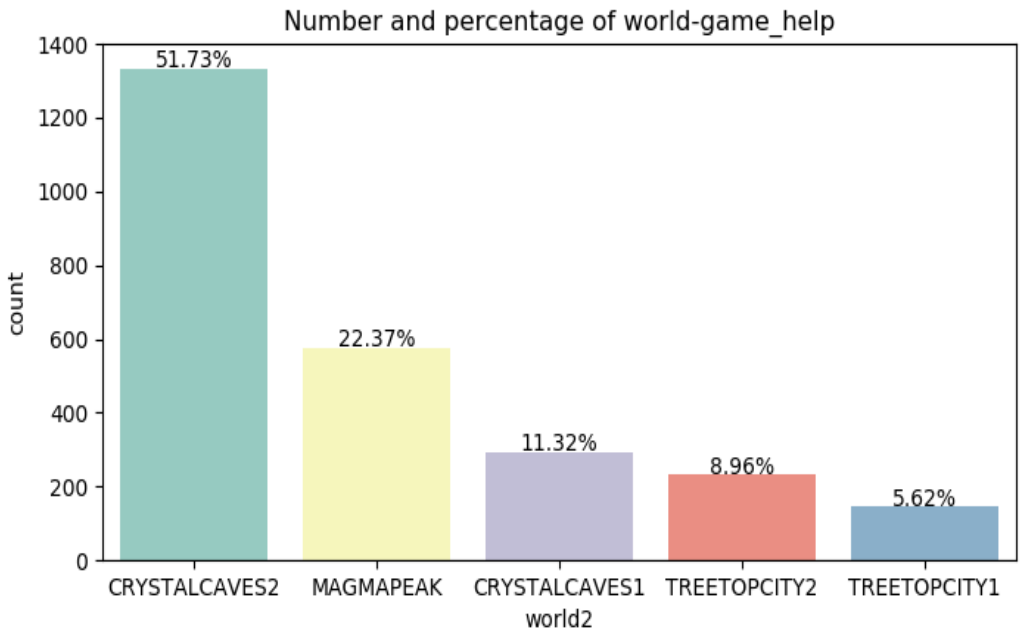


Train Data

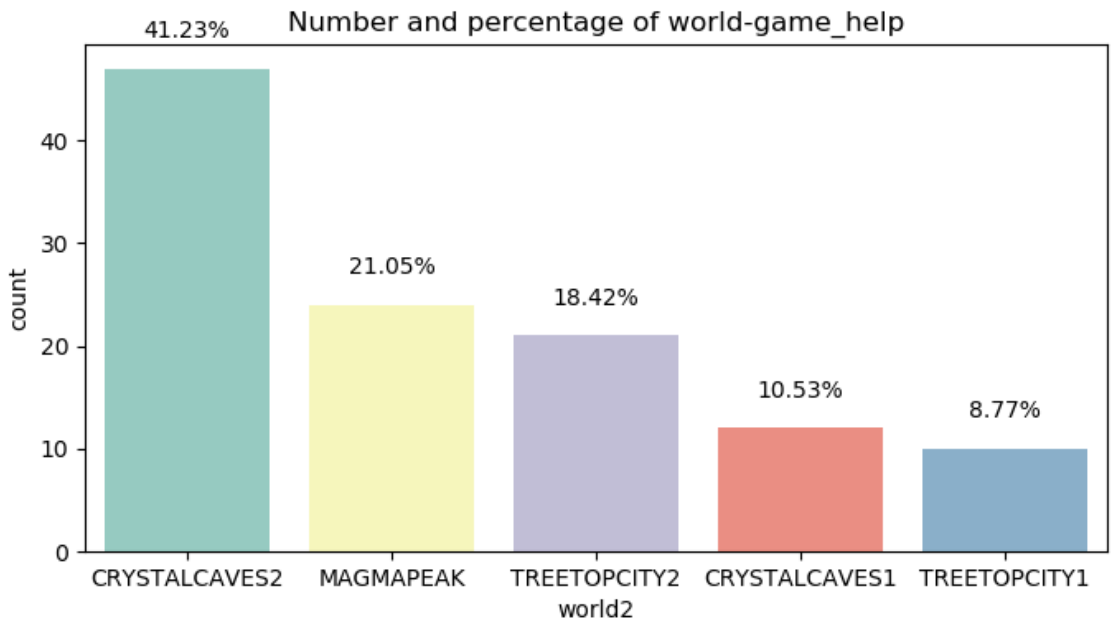


Test Data

Game 내 도움말 클릭 수

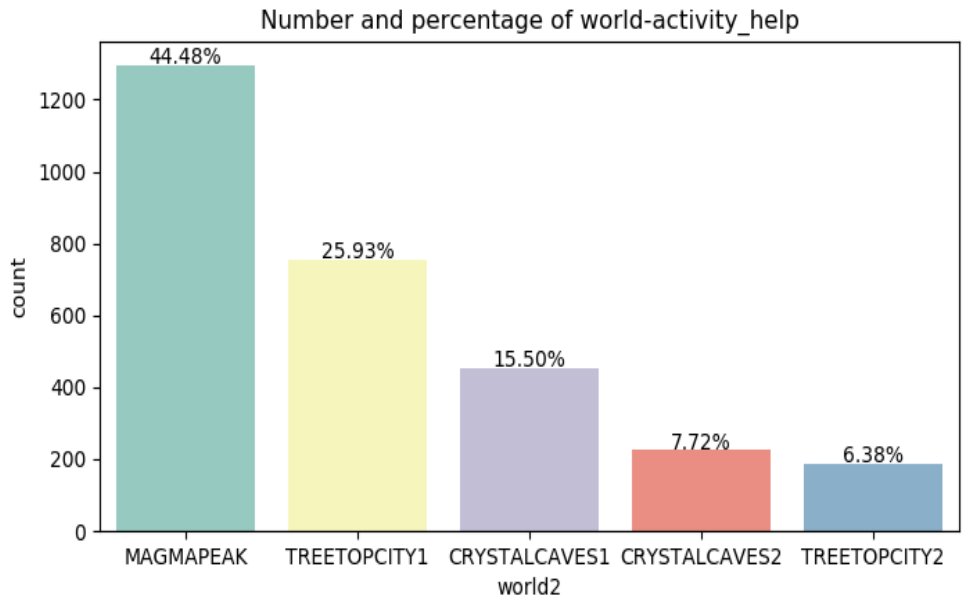


Train Data

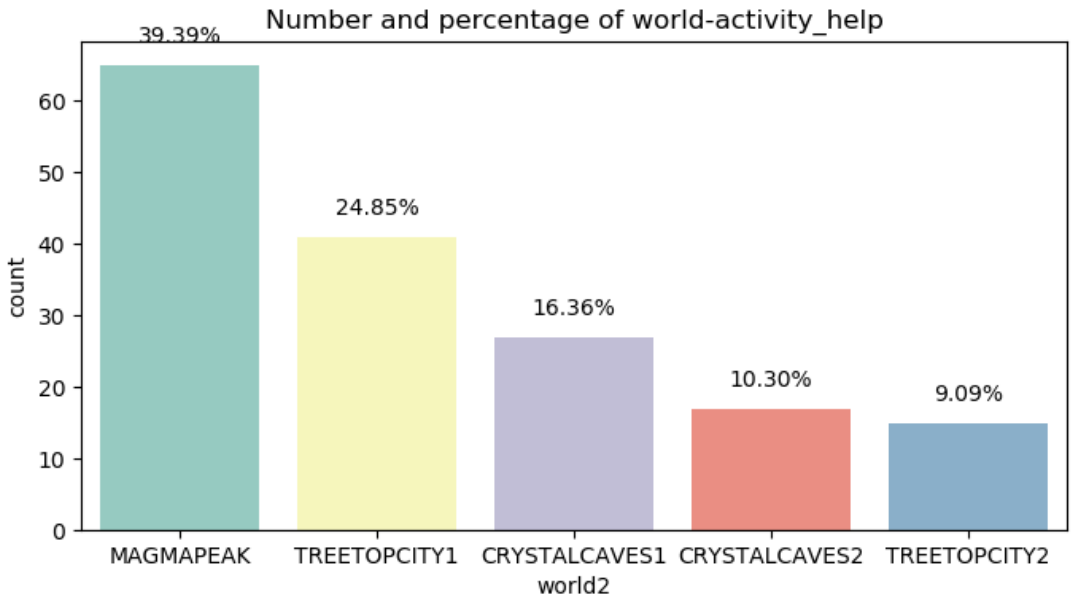


Test Data

Activity 내 도움말 클릭 수



Train Data



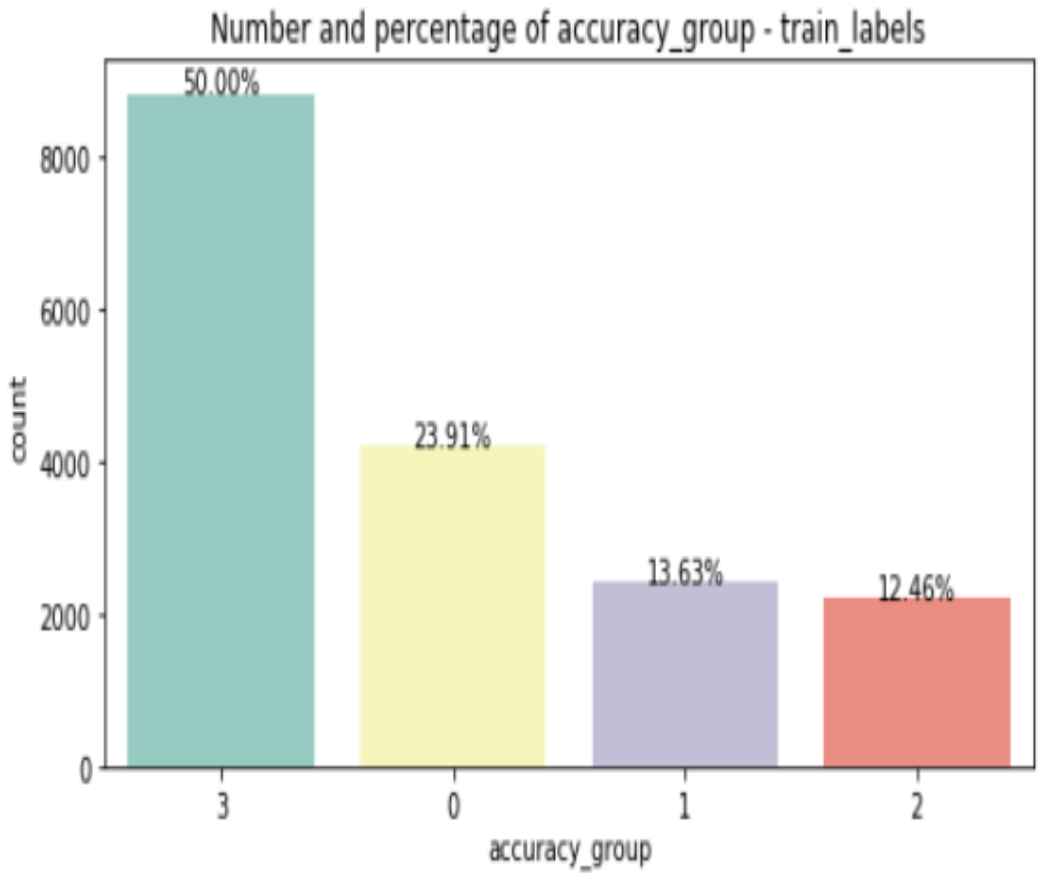
Test Data

변수설명

X	변수내용
X1	해당 평가 수행 전 Game 실행 횟수
X2	해당 평가 수행 전 Clip 실행 수
X3	해당 평가 수행 전 Activity 실행 수
X4	Clip 시청시간 점수화
X5	Game별 game_session에서 최초 정클릭까지의 횟수
X6	Activity별 game_session에서 최초 정클릭까지의 횟수
X7	사람별 평가별 시도 차수
X8	Game 내 도움말 클릭 수
X9	Activity 내 도움말 클릭 수

모델링

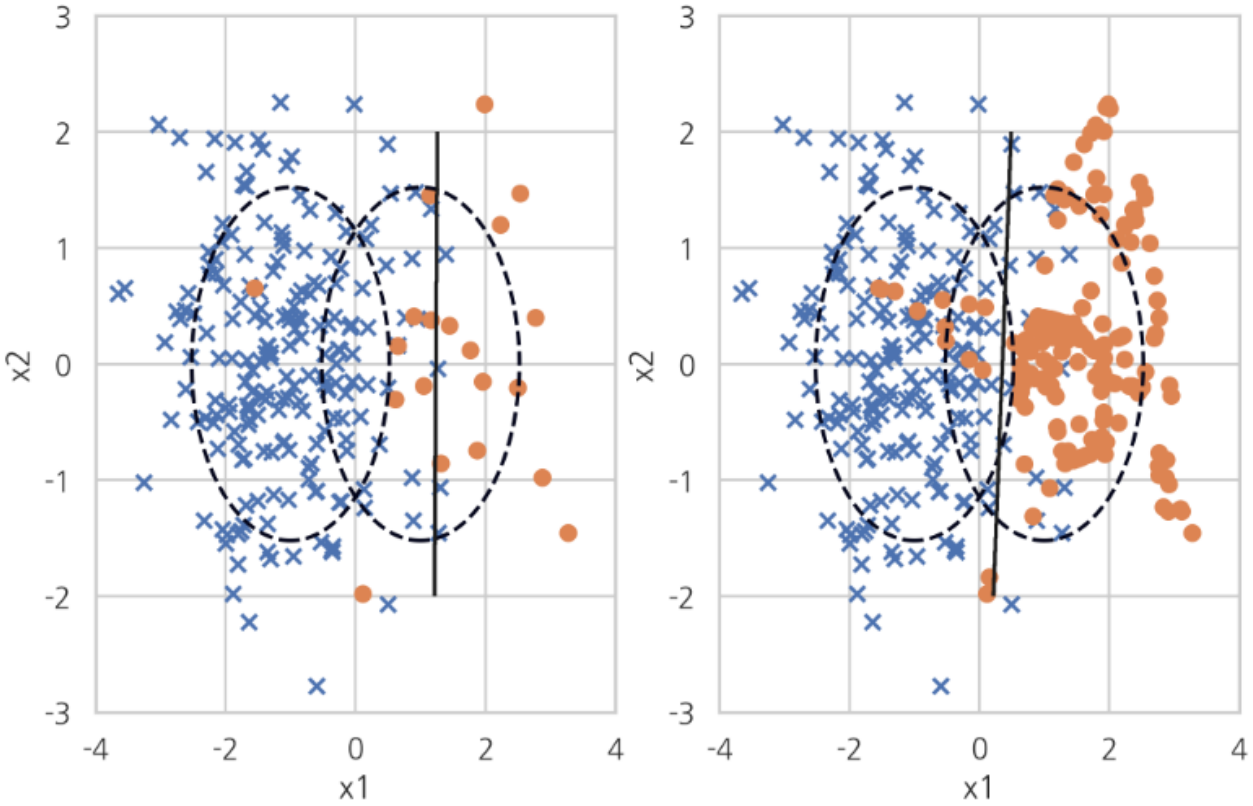
데이터 불균형

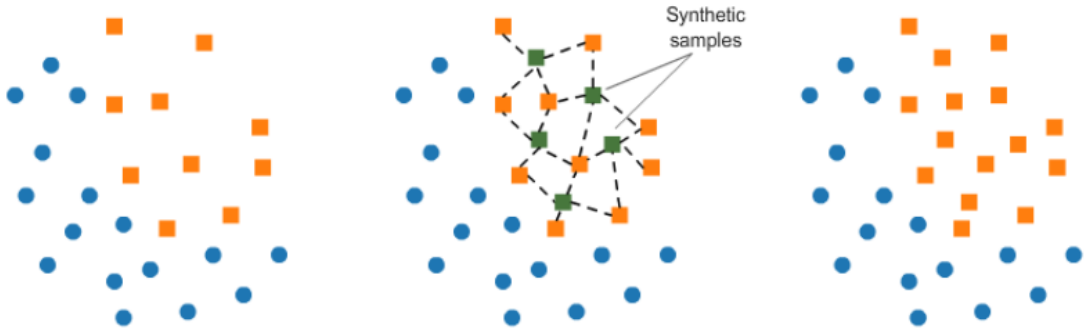


SMOTE

```
train = feed_train.loc[feed_train['accuracy_group'].dropna().index,:]

x_train = train.iloc[:,np.arange(5,14)]
x_train  = np.array(x_train)

y_train = train['accuracy_group']
y_train = np.array(y_train)

x_test= feed_test.iloc[:,np.arange(5,14)]
x_test = np.array(x_test)

from imblearn.over_sampling import SMOTE
sm = SMOTE(sampling_strategy='auto')
x_resampled, y_resampled = sm.fit_sample(x_train,y_train)
```

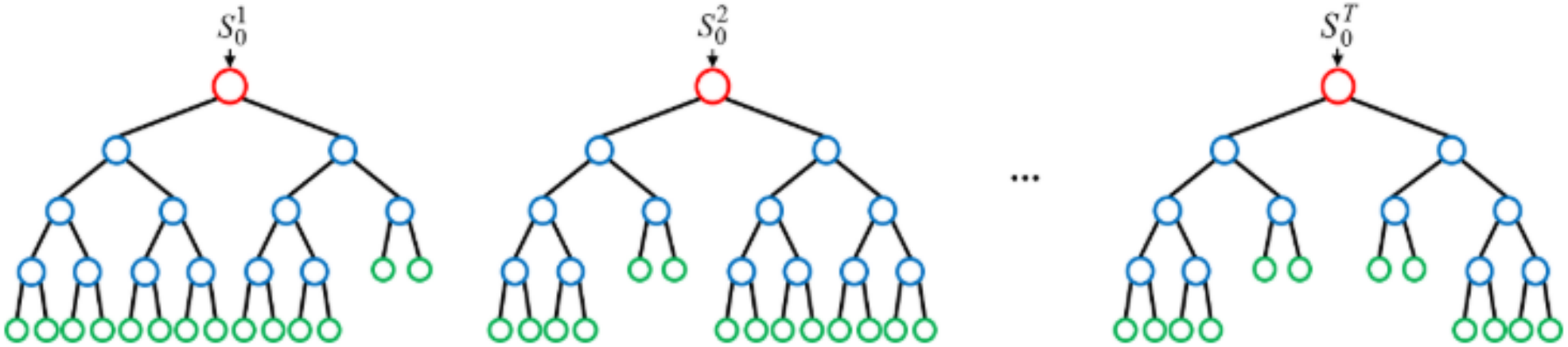
SMOTE





모델선정

랜덤포레스트(Random Forest)



랜덤포레스트는 수많은 의사결정 트리가 모여 만들어진 숲으로 표현한다.

예측을 해야하는 Y값이 범주형 데이터이므로 분류모델인 **Random Forest** 채택

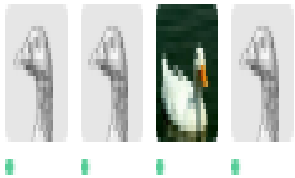
모델평가

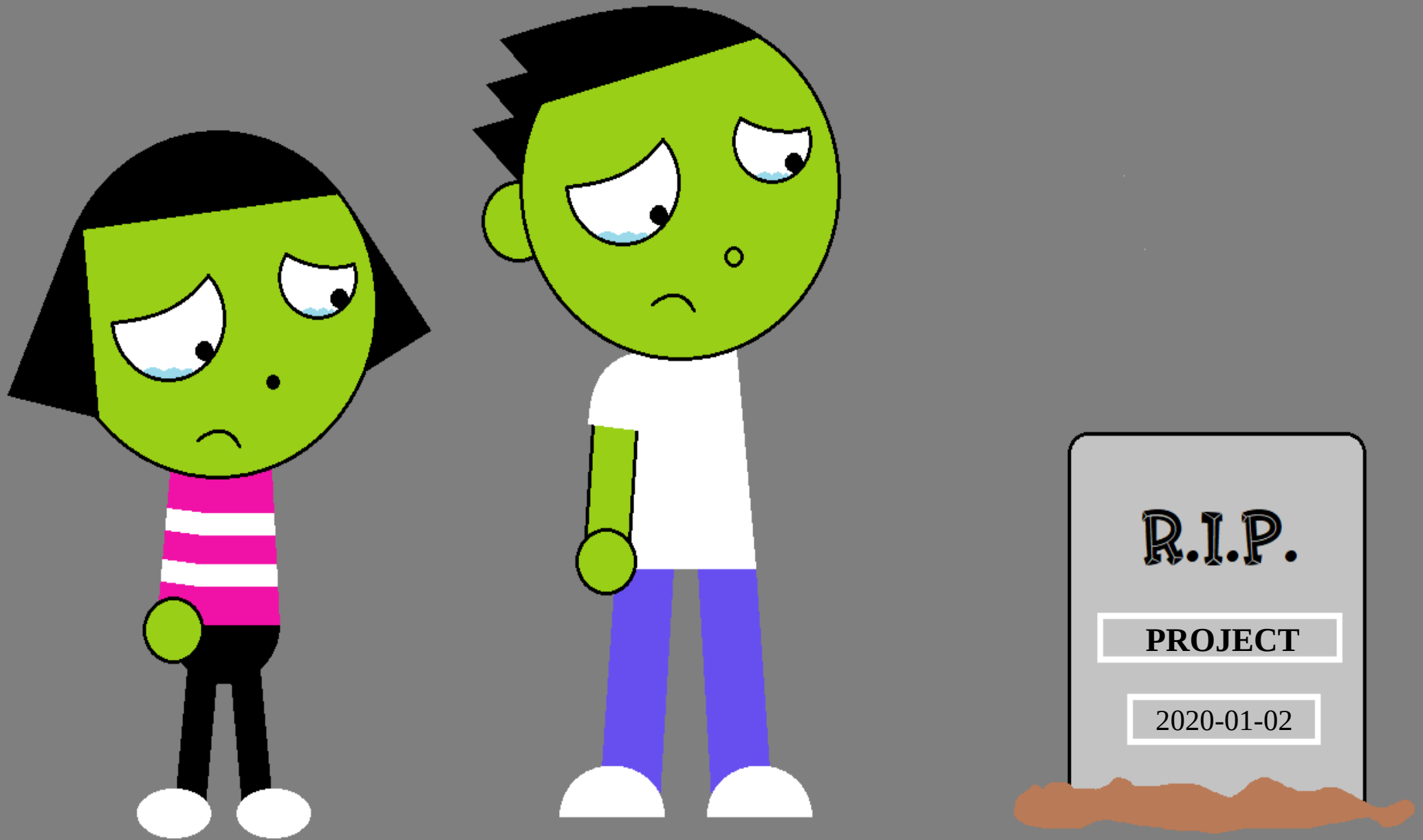
```
In [1041]: from sklearn.ensemble import RandomForestClassifier as rf

In [1042]: from sklearn.model_selection import GridSearchCV

In [1043]: from sklearn.model_selection import KFold

In [1044]: parm_grid = {'max_depth' : list(np.arange(1,10)),
      ...:                'max_features' : list(np.arange(1,9))}

In [1045]: cv1 = KFold(n_splits=5, shuffle= True, random_state=0)

In [1046]: grid = GridSearchCV(rf(n_estimators=100), parm_grid, cv=cv1)

In [1047]: from sklearn.model_selection import train_test_split

In [1048]: train_x, test_x, train_y, test_y = train_test_split(x_resampled, y_resampled,
random_state=10)

In [1049]: grid.fit(train_x, train_y)
Out[1049]:
GridSearchCV(cv=KFold(n_splits=5, random_state=0, shuffle=True),
             error_score=nan,
             estimator=RandomForestClassifier(bootstrap=True, ccp_alpha=0.0,
                                              class_weight=None,
                                              criterion='gini', max_depth=None,
                                              max_features='auto',
                                              max_leaf_nodes=None,
                                              max_samples=None,
                                              min_impurity_decrease=0.0,
                                              min_impurity_split=None,
                                              min_samples_leaf=1,
                                              min_samples_split=2,
                                              min_weight_fraction_leaf=0.0,
                                              n_estimators=100, n_jobs=None,
                                              oob_score=False,
                                              random_state=None, verbose=0,
                                              warm_start=False),
             iid='deprecated', n_jobs=None,
             param_grid={'max_depth': [1, 2, 3, 4, 5, 6, 7, 8, 9],
                         'max_features': [1, 2, 3, 4, 5, 6, 7, 8]},
             pre_dispatch='2*n_jobs', refit=True, return_train_score=False,
             scoring=None, verbose=0)
```

```
In [1050]: grid.score(train_x, train_y)
Out[1050]: 0.5403429432824571

In [1051]: grid.score(test_x,test_y)
Out[1051]: 0.44635387224420575
```


```
In [1050]: grid.score(train_x, train_y)
Out[1050]: 0.5403429432824571

In [1051]: grid.score(test_x,test_y)
Out[1051]: 0.44635387224420575
```

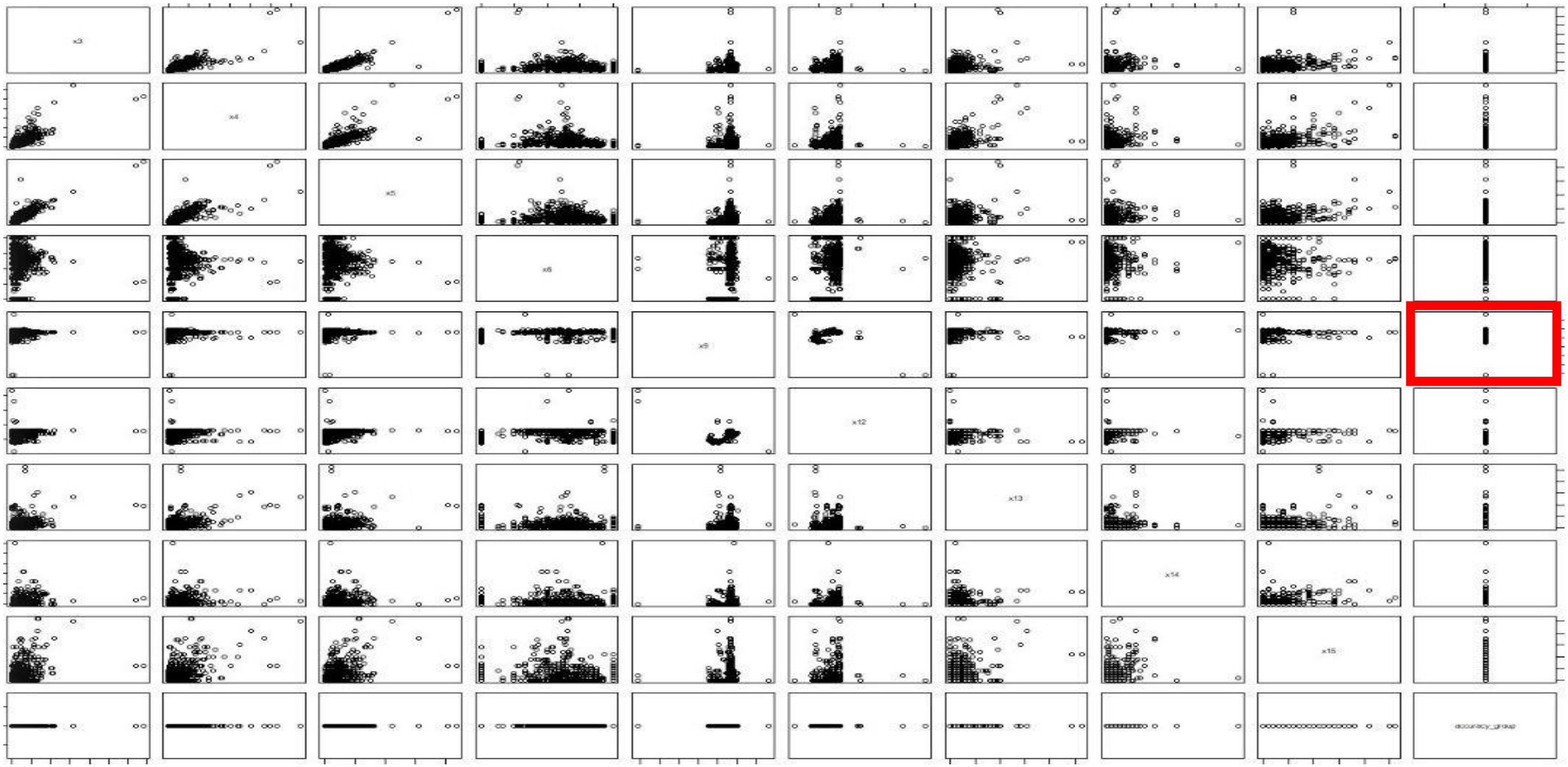
결과

2520	KIC		0.142	5	1h

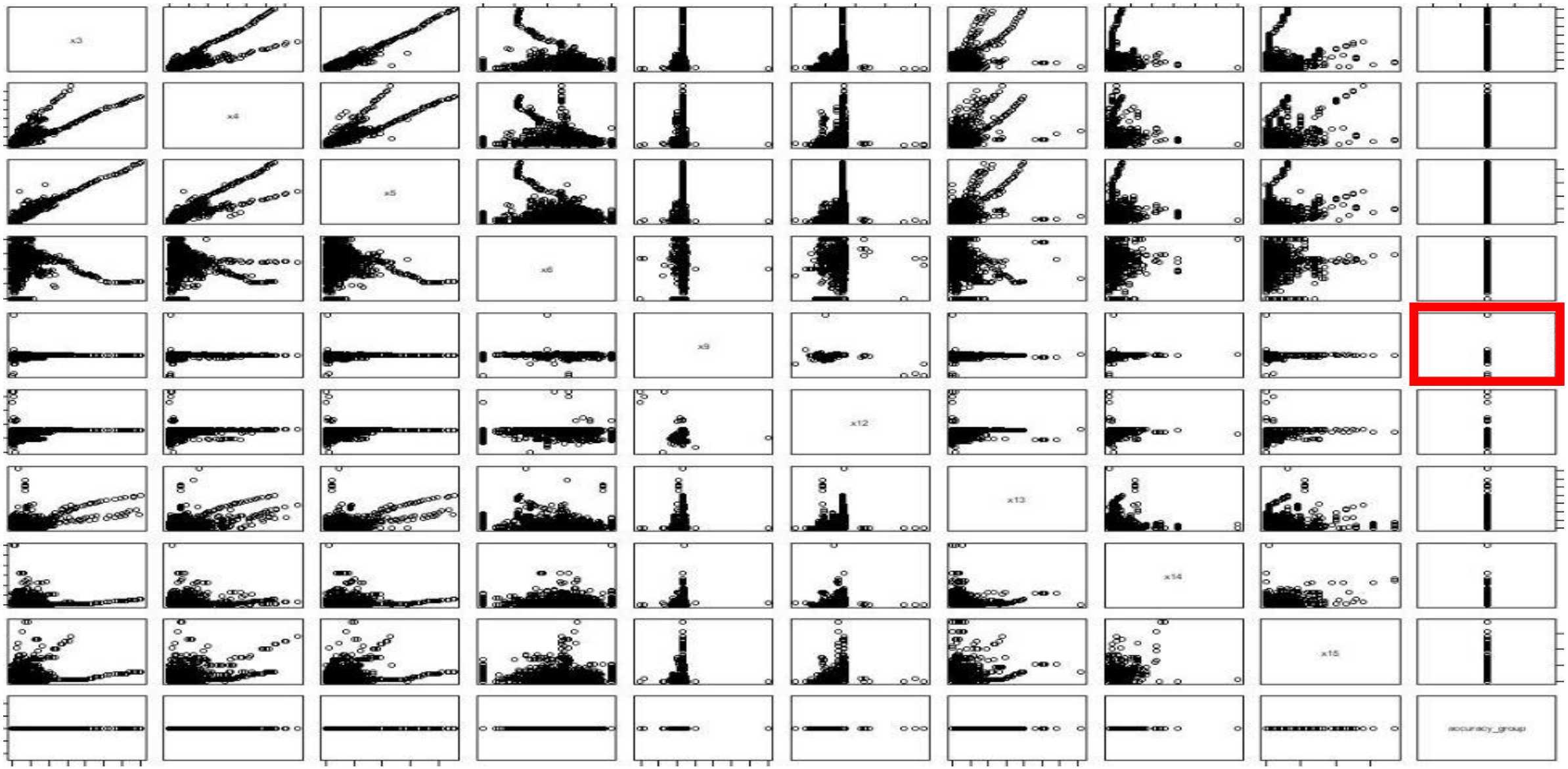

R.I.P.
PROJECT
2020-01-02

Why?

종속변수 값이 2인 그룹의 설명변수 산점도

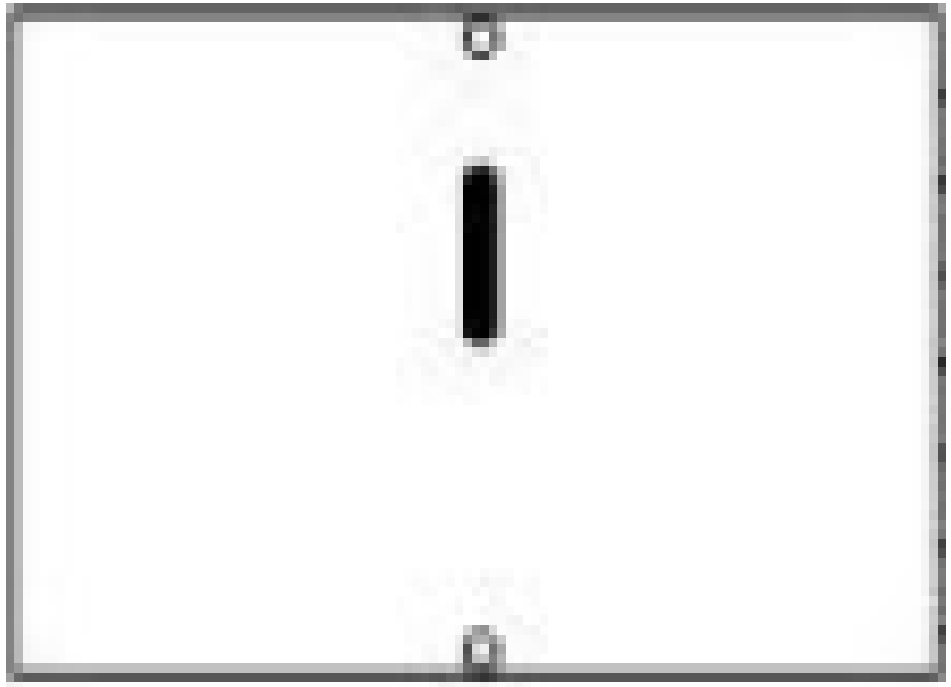


종속변수 값이 3인 그룹의 설명변수 산점도

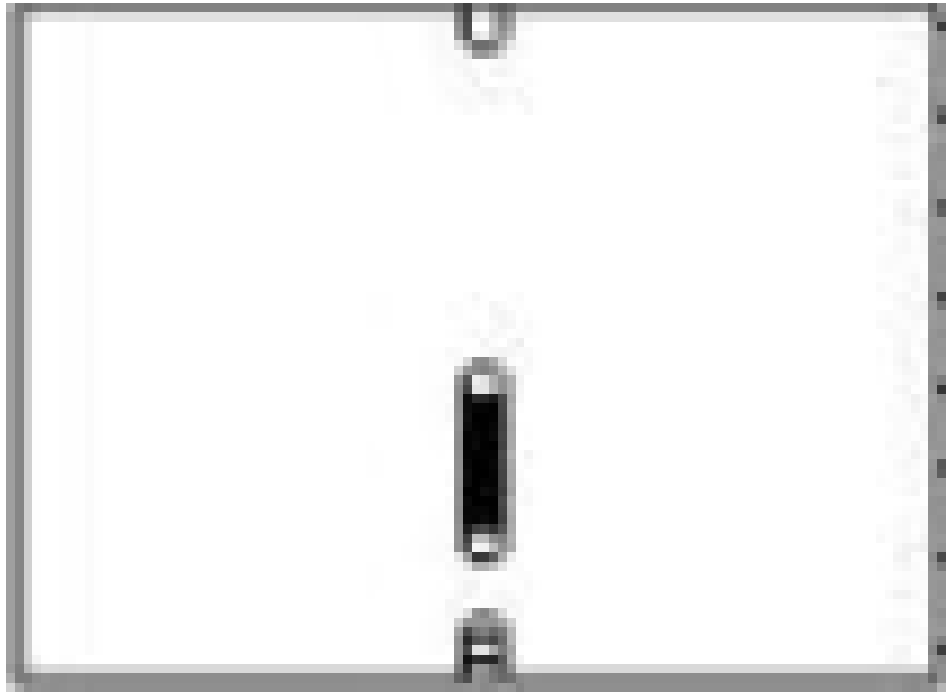


설명변수와 종속변수의 관계

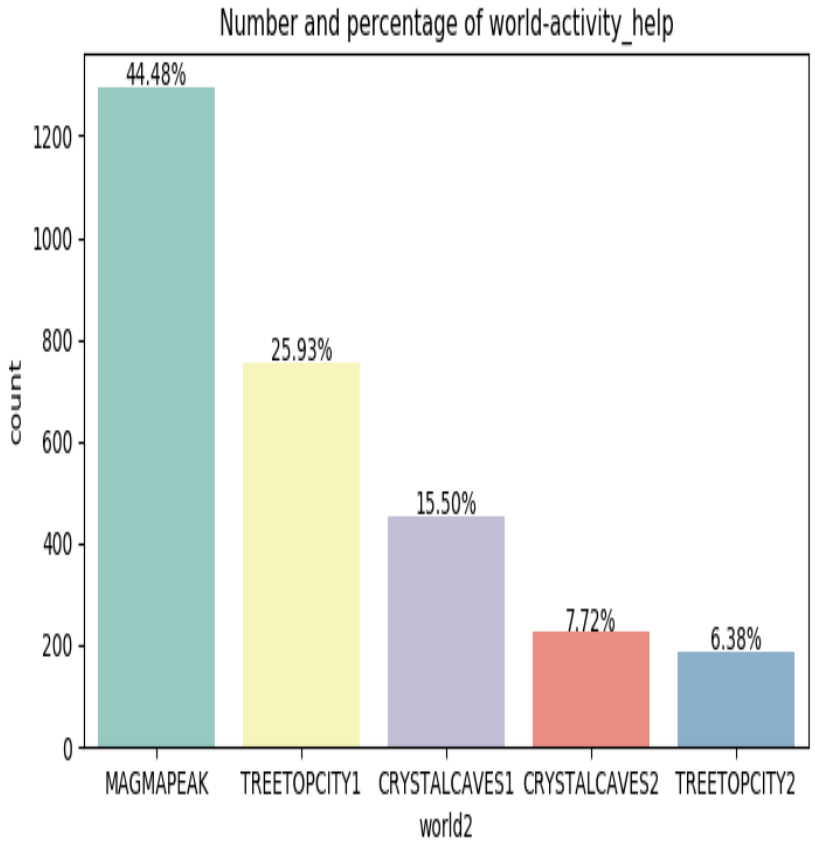
2점



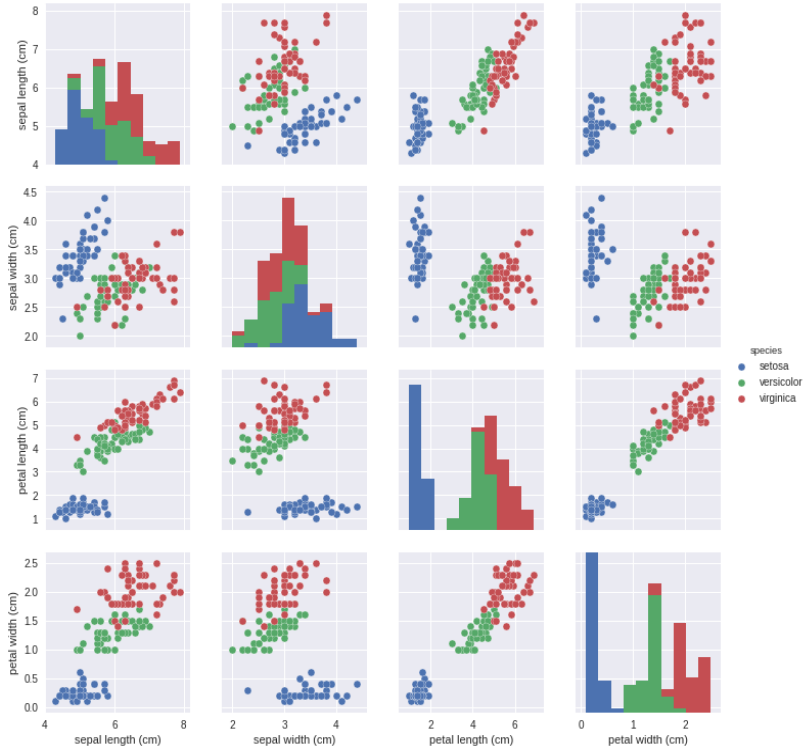
3점



실패요인

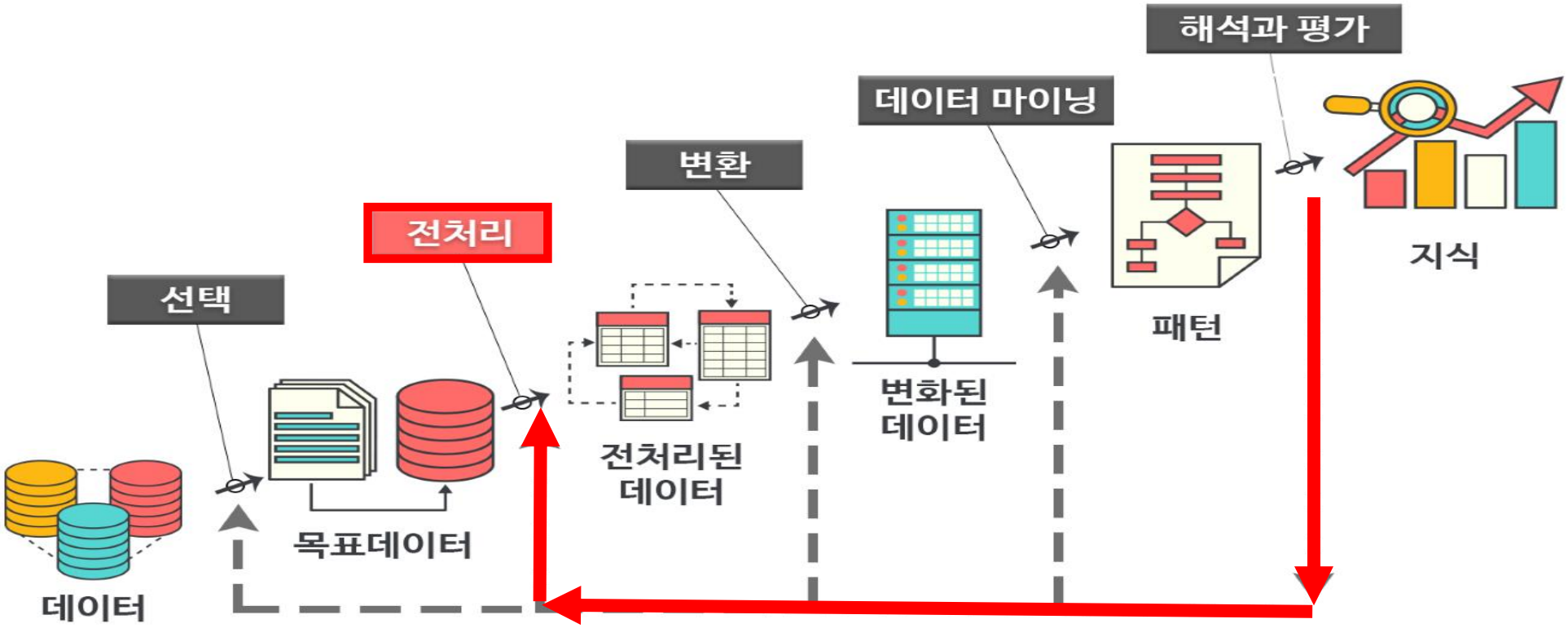


변수간 영향이 있는지 고려



변수가 데이터를 잘 분류하는지 고려

현재상황



Q & A

출처 및 참고자료

PPT템플릿 - http://pptbizcam.co.kr/

데이터 - https://www.kaggle.com/c/data-science-bowl-2019