유 퀴즈 온 더 블럭

데이터를 통한 머신러닝 프로젝트

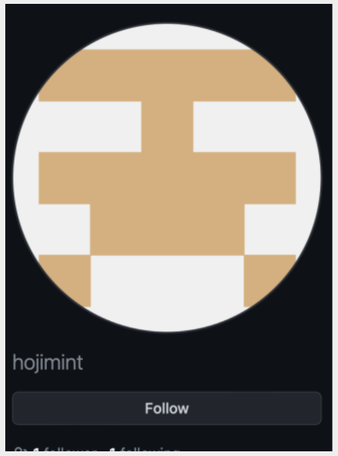
2024.02

목 차

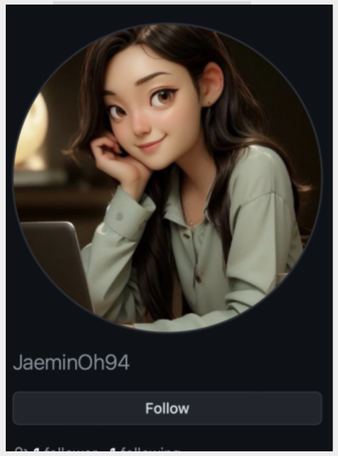
팀 소개

유기적인 구성으로 발표의 흐름을 표현하기 좋습니다.

1



2



3

CHAPTER

02

개요

다양한 출연자 들이 등장하는 '유 키즈 온더 블럭'
! 내가 아는 사람이 나온다면 과연 ?

CHAPTER

03

데이터 수집

계획준비 1단계

계획준비 2단계

계획준비 3단계

나무위키

나무위키 클로링을 통해
'유키즈'데이터를 수집 하였습니다

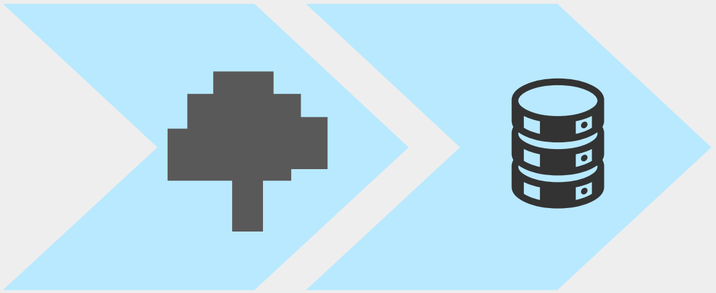
유튜브 API

유튜브 API 를 통해 '유키즈' 데이터
를 수집 하였습니다.

수기

직접 영상을 확인 하여 필요한 데이
터를 수집 하였습니다.

계획준비 1단계



나무위키

```
df_1 = df # 시즌1
df_1
```

	EP.:회차	날짜:날짜	에피소드:주제	출연자:출연자	촬영지:촬영지	시청률:시청률
0	1	2018년 8월 29일	Step by Step	서유림[1], Sm Uzzal & Sompa Sharin[2], 임일빈[3], 김명...	서울시 종로구 계동·옥인동 일대	2.3%
1	2	2018년 9월 5일	만남	김성중[6], 박종훈[7], 정재영 & 김성섭[8], 최재이[9], 길영인 & 양은...	역삼동	1.9%
2	3	2018년 9월 12일	비와 당신[12]	박현준[13], 김정원 & 조아영[14], 신호순 & 토토[15], Emo & Is...	장충동/서울 을지로	1.8%
3	4	2018년 9월 19일	Power Up	김재영[21], 이화영[22], 최혜원 & 정서현[23], 김윤배 & 박시현[24]...	을지로[27]/신촌&연희동[28]	1.9%
4	5	2018년 9월 26일	동네 한 바퀴	김수현[29], 김지현[30], 유광수 & 박현정 & 박누리 & 조가영[31], 고...	연희동[36]	1.5%
5	6	2018년 10월 3일	New Rules	문혜정[37], 박대은[38], 장승희 & 문희진[39], 최재용 & 박형주 & 최...	서울시 용산구&남산	1.6%
6	7	2018년 10월 10일	초심으로	한재호[45], 정해정[46], 정재훈 & 양선호 & 이영훈[47], 문관일 & 김...	대구	1.5%
7	8	2018년 10월 17일	가을이 오면	오재성[53], 이경수[54], 허성행[55], 김세운[56], 채우정[57], 정...	삼청동, 효자동	1.7%
8	9	2018년 10월 24일	압구정 날라리	홍정우[61], 원예나 & 이지현 & 김민하[62], 조규찬 & 장상훈 & 신형원[...	압구정	1.6%
9	10	2018년 10월 31일	잊혀진 계절	이재하[68], 박경철[69], 강영욱[70], 이원경[71], 이승주 & 송수영[...	마포구 망원동&은평구 응암동	1.7%
10	11	2018년 11월 7일	시간과 낙엽	김지혜[77], 최기락[78], 최선남[79], 유성모[80], 이윤경[81], 김...	강남구/송파구 풍납동 올림픽 공원	1.7%
11	12	2018년 11월 14일	YES or YES	나한성 & 임민현[85], 현수용 & 강숙자[86], 서인희[87], 김경열[88]...	서울시 중구	1.5%

회차, 날짜, 에피소드, 출연자, 촬영지,시청률 등을 수집 하였습니다.

계획준비 2단계

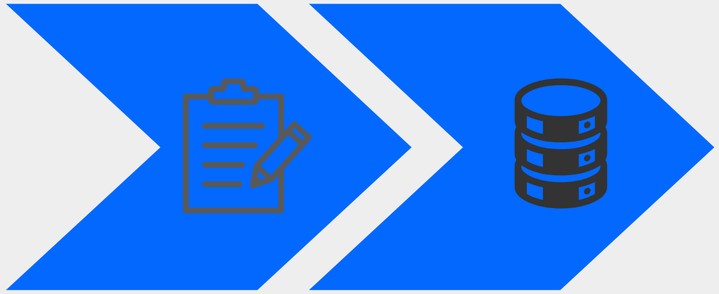
유튜브 API

```
request = youtube.channels().list(part="snippet,contentDetails,statistics", id=id)
response = request.execute()

channel_overview = {
    'title' : response['items'][0]['snippet']['title'],
    'description' : response['items'][0]['snippet']['description'],
    'publishedAt' : response['items'][0]['snippet']['publishedAt'],
    'viewCount' : response['items'][0]['statistics']['viewCount'],
    'subscriberCount' : response['items'][0]['statistics']['subscriberCount'],
    'videoCount' : response['items'][0]['statistics']['videoCount'],
    'uploads' : response['items'][0]['contentDetails']['relatedPlaylists']['uploads']
    }

df_channel_overview = pd.DataFrame([channel_overview])
df_channel_overview
```

	title	description	publishedAt	viewCount	subscriberCount	videoCount	uploads
0	유 퀴즈 온 더 튜브	tvN '유 퀴즈 온 더 블럭' 공식 유튜브 채널	2020-07-24T09:54:31.459547Z	602217815	900000	4772	UU920m3pMPH45qztdhppZhwA

채널명,영상ID, 카테고리ID, 게시일 ,제목, 설명, 출연자_정보, 재생시간
,조회수, 좋아요수, 댓글수, 재생시간(초), 등의 데이터 수집

계획준비 3단계



수기

7	박지선	연예인	W	청년
8	배한욱	전문기술	W	청년
9	김태연	연예인	W	청년
10	최설아	학생	W	청년
11	이진재	학생	M	청년
12	이인봉	기타	M	청년
13	김희숙	기타	W	중년
14	박정훈	운동선수	M	청년
15	박완용	운동선수	M	청년
16	임서준	운동선수	M	청년
17	일마즈	서비스	M	청년
18	메멧	서비스	M	청년
19	안다희	서비스	W	청년

이름에 맞는 직업, 성별, 나이 를 수기로 작성하여 데이터를 수집하였습니다.

CHAPTER

04

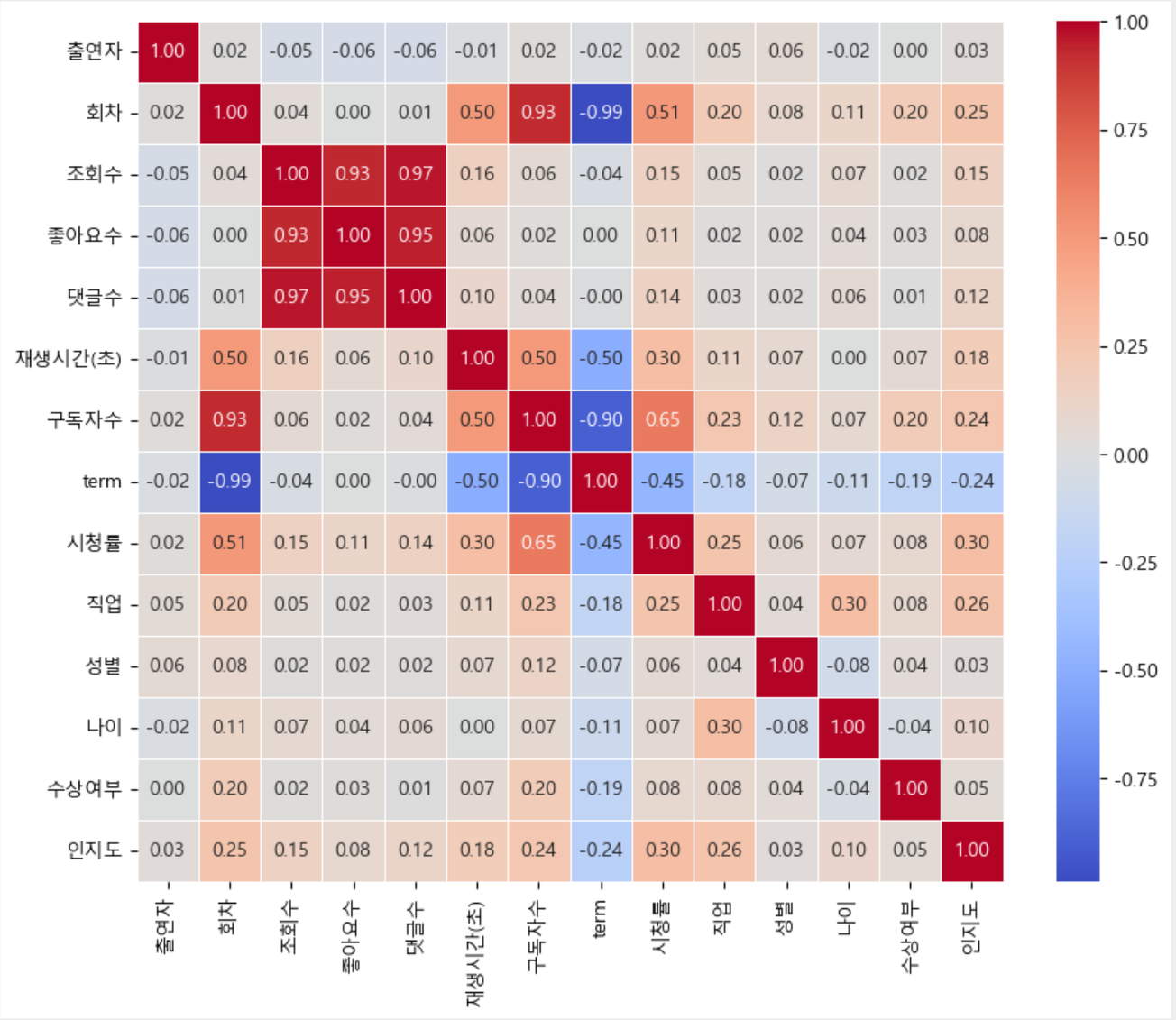
EDA

비트맵

인사이트 도출

1. 유튜브와 시청률의 관계

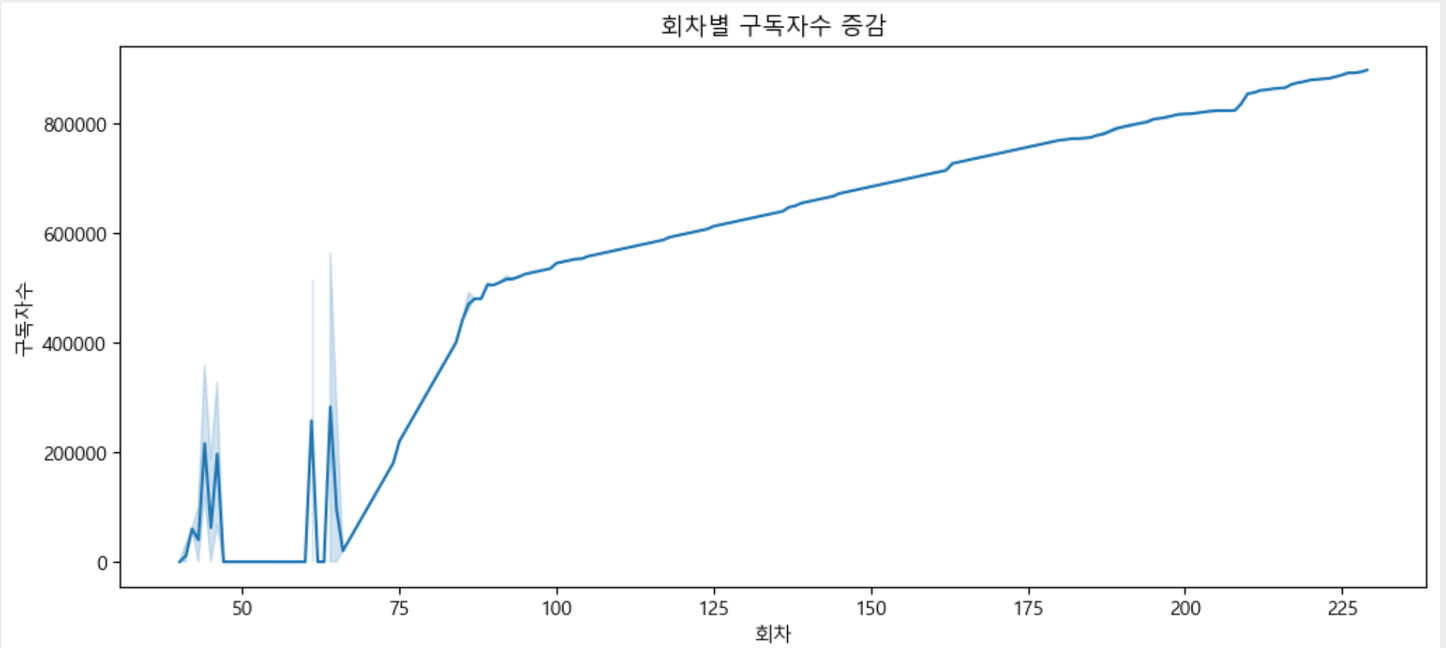
2. 유튜브 데이터의 관계



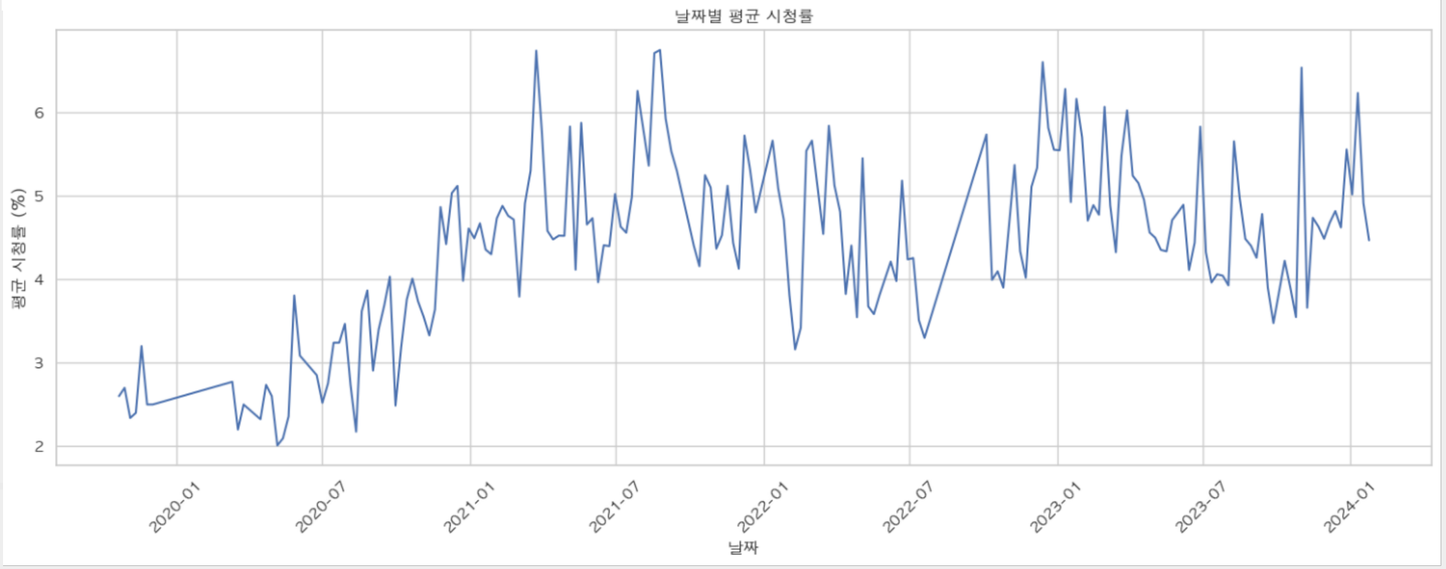
그래프

인사이트 도출

3.
회차별 구독자 수 증가



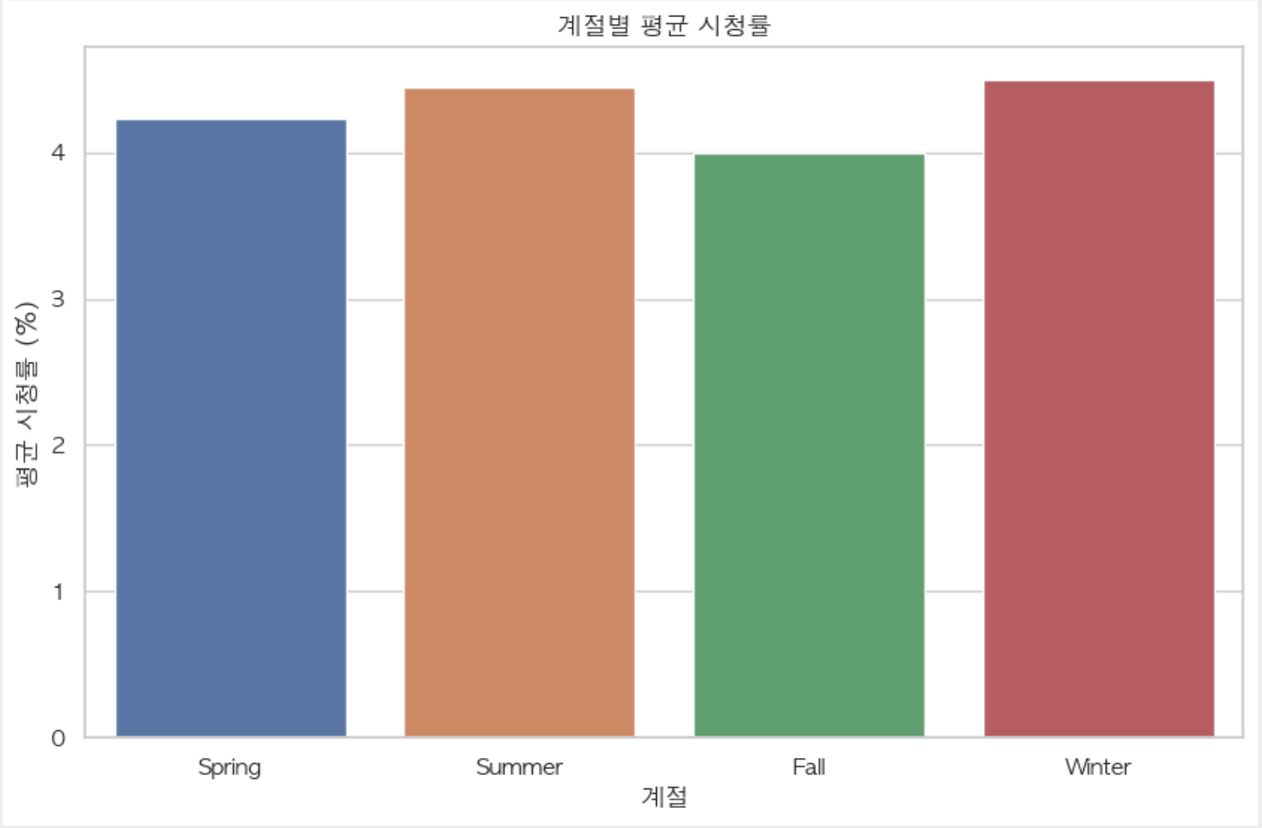
4.
날짜별 시청률 변화



막대 그래프

인사이트 도출

5.
계절별 시청률의 관계

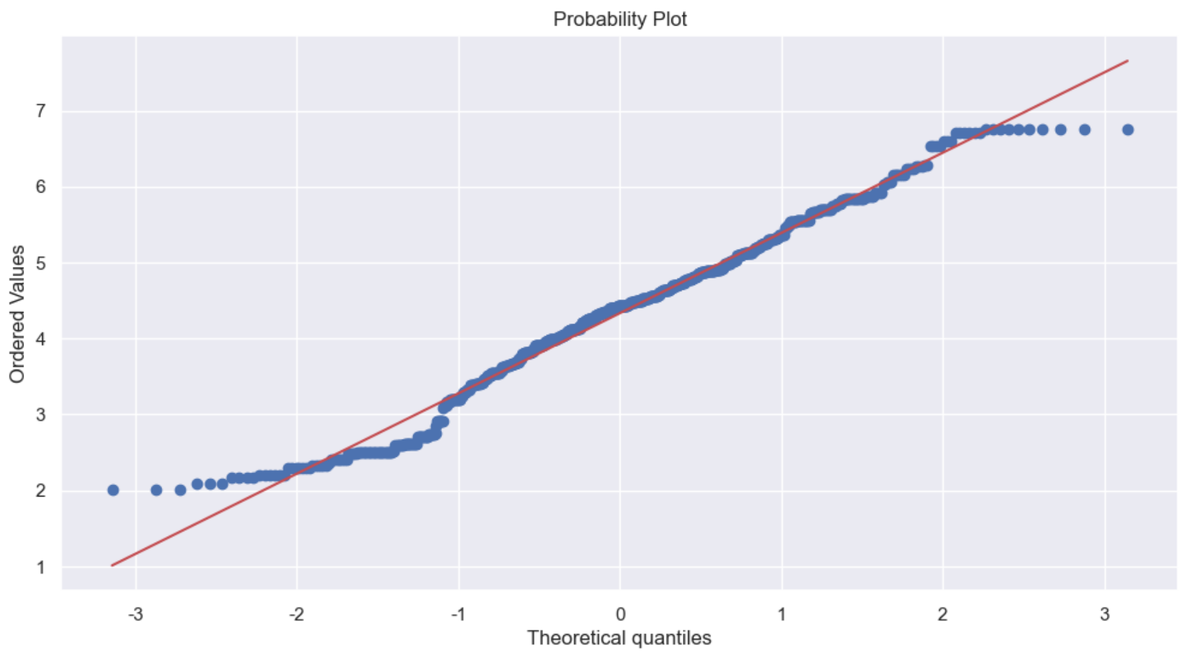


CHAPTER

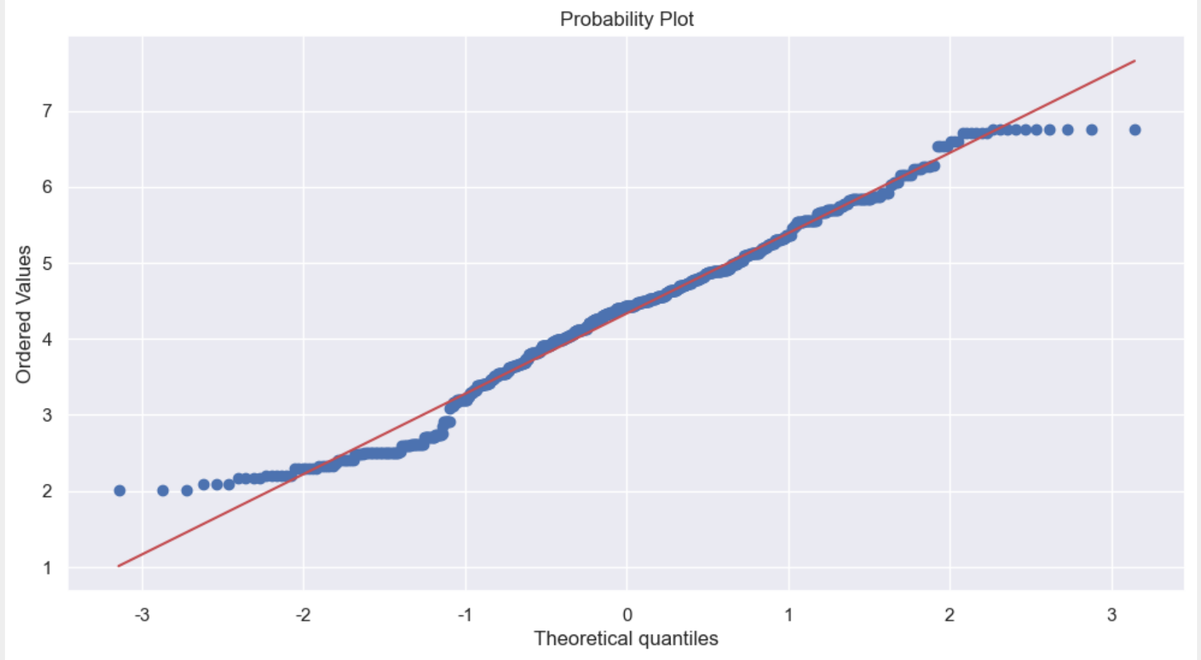
05

머신러닝

log 이용



GradientBoostingRegressor
Gradient Boosting score: 0.5362 (0.0465)
XGBRegressor
Xgboost score: 0.5996 (0.0321)
LGBMRegressor
LGBM score: 0.8170 (0.0449)



GradientBoostingRegressor
평균 제곱 오차 (MSE): 0.15959967792062352
XGBRegressor
평균 제곱 오차 (MSE): 0.1639369727770908

y타겟변경

```
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.preprocessing import StandardScaler

X = view_df[['출연자', '회차', '조회수', '좋아요수', '댓글수', '재생시간(초)', '구독자수', 'term', '직업', '성별
y = view_df['시청률']

scaler = StandardScaler()
```

MSE 시청률: 0.18978968969719767
r2 시청률: 0.8309623167276127

```
# X와 y 나누기
X = cl_labelEn_df[['회차', '직업', '성별', '나이']]  # 인적사항을 넣음
y = cl_labelEn_df[['조회수', '좋아요수', '댓글수', '시청률', '재생시간(초)', '구독자수', '수상여부', '인지도']]

# 데이터 분할
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)  # 8:2

# Gradient Boosting 모델 정의
gb_model = GradientBoostingRegressor()
```

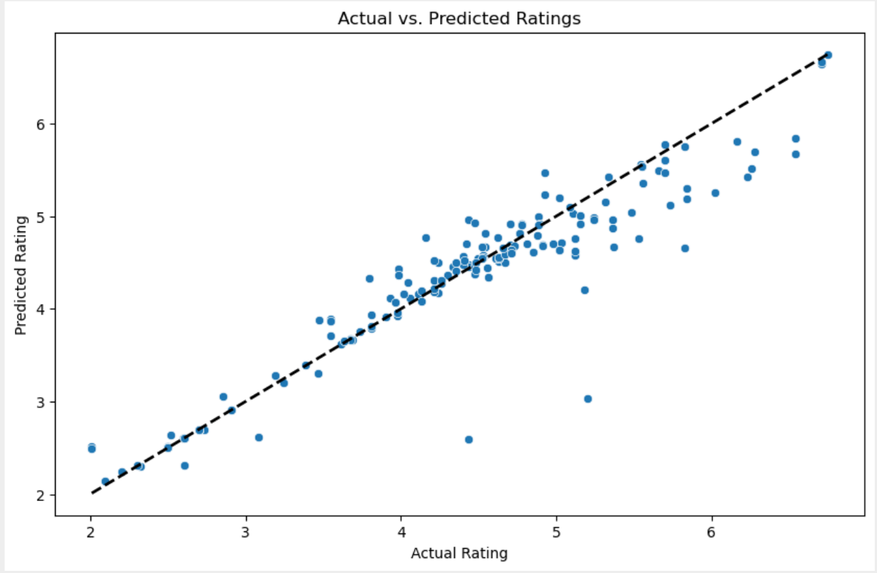
MSE 시청률: 0.18533141630927333
r2 시청률: 0.8335871085620313

컬럼 삭제 및 변경

```
df_data.head()
```

Python

	출연자	회차	구독자수	시청률	직업	성별	나이	수상여부	인지도	날짜
0	나영석	40	0	2.6	서비스	M	중년	Y	Y	2019-10-22
1	이명진	40	0	2.6	서비스	M	청년	N	N	2019-10-22
2	김부연	41	0	2.7	기타	F	중년	N	N	2019-10-29
3	김유자	41	0	2.7	기타	F	청년	N	N	2019-10-29
4	김만순	41	0	2.7	기타	F	노년	N	N	2019-10-29



Random Forest - MSE: 0.1335886601557344 MAE: 0.20838505489021963
Gradient Boosting - MSE: 0.24195357082994837 MAE: 0.35801050954304214
XGBoost - MSE: 0.279934983511718 MAE: 0.37771518187608544

유튜브 컬럼 삭제, 날짜 데이터 추가.

'시청률' 한정 머신러닝

CHAPTER

06

질문

발표를 들어주셔서
감사합니다.