

드라마 흥행요인 분석

멀티캠퍼스 파이썬 D반

CONTENTS



Chapter **01** 주제 선정 배경 및 목표

Chapter **02** 데이터 수집

Chapter **03** 데이터 분석 및 결과

Chapter **04** 웹 시연

Chapter 01

주제 선정 배경 및 목표

드라마 사업 부진

- 현재 KBS를 제외한 지상파 드라마들이 수익 실현이 어려워져 모습을 감추고 있고, 2021년 02월 기준 방송 중이거나 예정인 MBC와 SBS 드라마는 각 2편에 불과하다.
- 현재 지상파 드라마들은 꾸준한 시청률 하락과 저조한 화제성으로 입지가 흔들리기 시작했다.
최근에는 메인 드라마방송 시간대에 예능을 편성하면서 방송사 수익을 위해 변화하는 모습을 보이고 있다.

드라마는 유명 배우, 사전 관심도가 흥행 요인?

- 최근 JTBC에서 방영된 드라마 '라이브 온'은 황정민, 정다빈 등 하이틴 스타들이 출연하며 시작 전 화제를 모았지만, 시청률이 2%대를 머물렀다.
- 앞서 JTBC는 웹툰 원작 '이태원 클라쓰' '내 아이디는 강남미인' 등으로 성공신화를 이룬 만큼, 이번 웹툰 원작 드라마에도 기대가 쏠렸지만 최근에는 시청률이 저조한 모습을 보이고 있다.

프로젝트 목표

데이터 시각화	드라마 관련 정보들의 변수별 데이터 시각화와 드라마 시청률과 변수 사이의 데이터 시각화를 해보자
데이터 분석	웹 크롤링으로 얻은 여러가지 데이터로 머신러닝을 통해 드라마 시청률 예측을 구현해보자
	추천시스템 알고리즘을 활용하여 사용자가 본 드라마의 정보를 통해 드라마를 추천해주는 시스템을 구현해보자

일정

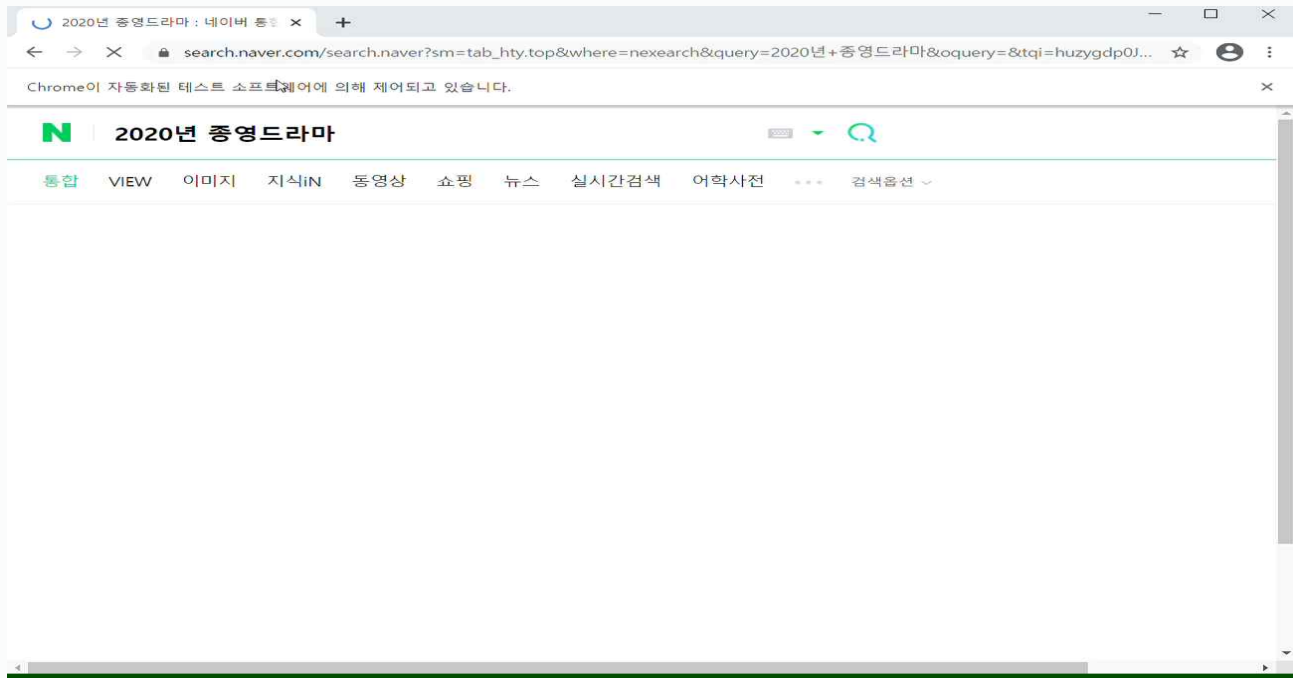
항목	과업	1월				2월										
		27일	28일	29일	30일	3일	4일	5일	8일	9일	10일	11일	12일	15일	16일	17일
주제 선정	브레인스토밍	→														
	주제 확정			→												
데이터 확보 및 분석	웹 크롤링					→										
	데이터 분석										→					
결론 도출 및 PT 자료준비	결론 도출											→				
	PT 자료														→	

Chapter 02

데이터 수집

웹 크롤링

- 네이버에서 제공되는 2010~2020년 종영드라마 정보를 수집하기 위해 Selenium, BeautifulSoup 라이브러리를 통해 웹 크롤링 진행



웹 크롤링

● 1474 X 35 크기의 데이터 수집 (2010~2020 종영드라마)

2019년 드라마_원본.csv [C:\Users\Wksaul\Downloads\W] - Excel

























	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	드라마 이름	방송사	방송기간	총 부작	요일 및 시간	내용	제작사	시청률	등장인물	1방송객수	영화객수	1등장인물	2방송객수	2영화객수	2등장인물	3방송객수	3영화객수	3등장인물
2	환상의타이밍	kbc	광주방송	2019.09.05	16부작	금 오후 08:55	KBC 광주방송	케이엠비디										
3	나쁜사랑	MBC	2019.12.02	129부작	월~금 오전 07:50	거대한 운	기획 장제	{'5.8', '6.3', '6.0', '6.2', '6.3',	신고은	7	0	이선호	19	11	오승아	11	1	윤종현
4	맛 좀 보실래요	SBS	2019.11.12	124부작	월~금 오전 08:35	우리 집안	SBS 미디어	{'9.4', '8.6', '8.1', '9.2', '8.4',	심이영	29	18	서도영	16	1	서하준	8	1	한가람
5	꽃길만 걸어요	KBS1	2019.10.28	123부작	월~금 오전 08:30	전송된 시	연출 박기	{'22.4', '22.2', '21.5', '22.1',	최윤소	21	7	설정환	11	1	심지호	18	4	정유민
6	우아한 모녀	KBS2	2019.11.04	103부작	월~금 오후 07:50	엄마에 의	아이엠비디	{'15.8', '15.7', '14.9', '15.8',	최명길	41	8	차예련	16	12	김홍수	15	8	김명
7	사랑은 뷰티풀 인생은	KBS2	2019.09.28	100부작	토, 일 오후 07:55	원가 되기	HB엔터테	{'23.9', '28.0', '21.3', '26.1',	설인아	9	0	김재영	15	4	조승희	24	9	윤박
8	두 번은 없다	MBC	2019.11.02	72부작	토 오후 09:05	서울 한복	엔터테인먼트	{'8.1', '9.2', '7.9', '10.3', '9.2	윤여정	68	33	박세완	10	5	박승연	26	4	오지
9	사랑의 불시착	tvN	2019.12.14	16부작	토, 일 오후 09:10	어느 날 돌	문화창고	{'6.1', '6.8', '7.4', '8.5', '8.7',	현빈	12	15	손예진	10	20	서지혜	25	5	김정
10	스토브리그	SBS	2019.12.16	16부작	금, 토 오후 10:00	팬들의 눈	김픽처스	{'5.5', '7.8', '9.6', '11.4', '12	남궁민	29	7	박은빈	35	4	오정세	26	73	조병
11	검사내전	JTBC	2019.12.16	16부작	월, 화 오후 09:30	미디어 속	(주)에스피	{'5.0', '5.0', '4.7', '4.2', '3.6',	이선균	18	37	정려원	18	8	이성재	22	16	김광
12	간택 - 여인들의 전쟁	TV조선	2019.12.14	16부작	토, 일 오후 11:00	정통 왕조	아이그라	{'2.6', '2.9', '2.6', '3.6', '3.3',	진세연	18	5	김민규	14	3	도상우	12	0	이열
13	블랙독	tvN	2019.12.16	16부작	월, 화 오후 09:30	기간제 교	스튜디오	{'3.3', '4.4', '4.4', '4.3', '5.5',	서현진	19	12	라미란	29	50	하준	7	6	이창
14	99억의 여자	KBS2	2019.12.04	32부작	수, 목 오후 10:00	우연히 현	빅토리콘	{'7.0', '8.5', '9.4', '11.3', '7.8	조여정	25	8	김강우	19	27	정웅인	37	29	오나
15	초콜릿	JTBC	2019.11.29	16부작	금, 토 오후 10:50	메스저링	드라마하	{'3.5', '4.4', '4.3', '4.6', '4.4',	윤계상	12	16	하지원	15	26	장승조	13	2	민진
16	하차있는 인간들	MBC	2019.11.27	32부작	수, 목 오후 08:55	꽃미남 형	메이스트	{'2.2', '3.0', '2.5', '3.0', '2.3',	오연서	22	10	안재현	21	3	김슬기	22	10	김태
17	싸이코패스 다이어리	tvN	2019.11.20	16부작	수, 목 오후 09:30	어쩌다 목	스튜디오	{'1.8', '1.5', '1.9', '2.2', '2.4',	윤시윤	21	2	정인선	17	11	박성훈	15	5	이한
18	농부사관학교 2	SBS	2019.12.15	4부작	월 오전 12:05	뜨거운 여	SBS모비	{'1.4', '1.3', '0.7', '0.8',										
19	루악인간	JTBC	2019.12.30	2부작		은퇴 위기	드라마하	{'1.8', '1.6',	안내상	81	40	김미수	4	3	장혜진	6	14	최덕
20	연애 기다린 보람 - 내	MBC Dram	2019.12.28	2부작		소문난 오	기획 이현											
21	드라마 스테이지 - 빅	tvN	2019.12.25	1부작		빅데이터	연출 주성	{'0.7',	송재림	19	8	전소민	19	6	이승원	8	5	서윤
22	VIP	SBS	2019.10.28	16부작	월, 화 오후 10:00	백화점 상	디스토리	{'6.8', '7.6', '8.0', '9.1', '7.8',	장나라	23	5	이상윤	20	4	이청아	27	18	박선
23	커넥트	OBS 경인	2019.11.02	8부작	토 오후 05:05	그동안 모	슈퍼퍼멘			8	1							

웹 크롤링

- 멜론 월간차트를 활용하여 순위권에 속하는 드라마 OST를 점수화하여 변수로 사용하기 위해 웹 크롤링을 통해 정보 수집

2010년 01월 ~ 2021년 01월 드라마 OST 순위 정보 수집

[illegible]

	2018년 11월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	1KB
	2018년 12월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	1KB
	2019년 01월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	1KB
	2019년 02월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	1KB
	2019년 03월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	1KB
	2019년 04월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	1KB
	2019년 05월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	1KB
	2019년 06월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	1KB
	2019년 07월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	1KB
	2019년 08월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	1KB
	2019년 09월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	1KB
	2019년 10월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	1KB
	2019년 11월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	1KB
	2019년 12월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	1KB
	2020년 01월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	1KB
	2020년 02월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	2KB
	2020년 03월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	2KB
	2020년 04월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	2KB
	2020년 05월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	2KB
	2020년 06월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	2KB
	2020년 07월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	2KB
	2020년 08월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	2KB
	2020년 09월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	2KB
	2020년 10월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	1KB
	2020년 11월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	1KB
	2020년 12월 월본차트	2021-02-11 오전 1:16	한컴오피스 2018 ...	1KB
	2021년 01월 월본차트	2021-02-11 오전 1:17	한컴오피스 2018 ...	1KB

공공데이터 활용

- 드라마가 방영 되는 기간에 평균 기온, 강수량을 변수로 사용하기 위해 정보 수집

기상청 기상자료개방포털 보도자료

국가기후데이터센터 소개 | [*가-가](#) | 로그인 | 사이트맵 | [☆ 즐겨찾기](#) | [ENG\(info\)](#)

기상자료개방포털이란?

데이터

기후통계분석

간행물

소통과 참여

ALL

기후통계분석

평년값

통계분석

조건별통계

기온분석

강수량분석

다중지점통계

24절기

순위값

Home > 기후통계분석 > 통계분석 > 기온분석

기온분석 - 그래프

그래프

분포도

■ 자료설명

지점별로 기온의 시계열 분석을 확인합니다.
월, 연의 평균기온, 최저기온, 최고기온을 각각 조회할 수 있습니다.

* (그래프) 평균최고(최저)기온: 일최고(최저)기온의 월평균

* '지역/지점'의 '지역'은 전국 및 광역 단위의 평균 제공(1973년~) (전국 및 광역별 평균에 사용된 지점은 전국 평균산출에 사용되는 45개 지점이며, 제주도는 제주시와 서귀포시 자료임)

■ 검색조건

자료구분 일

자료형태 기본

기간 20210118 ~ 20210216

부족한 정보 추가

- 드라마 방영시간대, 요일, 장르 정보는 웹 크롤링으로 수집되지 않아 직접 추가 작업 진행

- 장르는 위키백과 정보를 참고하여 진행

- 드라마 방영시간대, 요일은 라벨링 작업 진행

오전1타임(~12시)=A 오후1타임(오후12시~오후8시)=B 밤1타임(오후8시~)=C

월: 1, 화:2, 월~화: 3, 수:4, 목:5, 수~목:9, 금:10, 토: 11, 월~목:12, 월~금:22, 금~토:21, 일: 13, 토~일:24

Chapter 03

데이터 분석 및 결과

Chapter **03-1**

시청률 예측 모델

데이터 전처리

- 데이터 시각화, 예측 모델 구현 전에 데이터 전처리 작업 진행
- 배우 평균활동 건수, 평균시청률, 더미변수(방송사, 요일 및 시간, 장르) 처리

	평균시청률	배우_방송	배우_영화	연출_드라마	연출_영화	작가_드라마	작가_영화	부작	요일시간_10C	요일시간_11C	...	genre_사극	genre_서스펜스	genre_스릴러	genre_액션	genre_청춘	genre_추리	genre_코메디	genre_판타지	genre_하이틴	genre_휴먼
0	7.26333	17.400000	5.4	13.0	1.0	0.0	0.0	123	0	0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	20.5033	20.800000	3.4	4.0	1.0	0.0	0.0	124	0	0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	39.2467	20.400000	3.8	15.0	0.0	7.0	0.0	106	0	0	...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
3	13.41	41.800000	8.4	5.0	0.0	5.0	0.0	80	0	0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	13.89	21.200000	8.8	12.0	0.0	13.0	0.0	52	0	0	...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...
104	4.00625	24.200000	4.4	7.0	0.0	0.0	0.0	16	0	0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
105	3.7875	18.000000	12.6	0.0	0.0	1.0	3.0	16	0	0	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

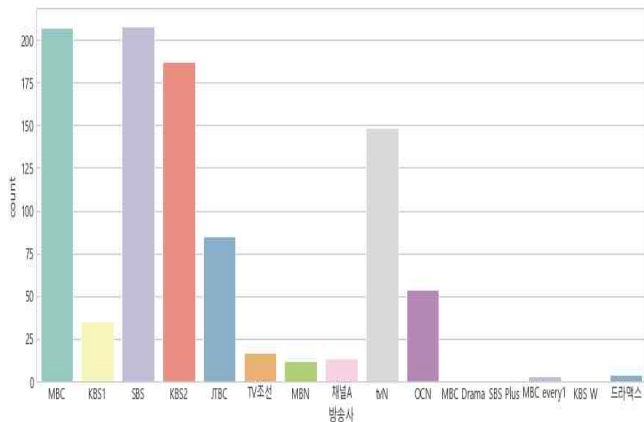
```
df['총 부작'] = df['총 부작'].str.replace('부작', '').astype(int)
```

```
[ ] df['배우_방송'] = np.mean(df[['방송갯수_1','방송갯수_2','방송갯수_3','방송갯수_4','방송갯수_5']],axis=1)
df['배우_영화'] = np.mean(df[['영화갯수_1','영화갯수_2','영화갯수_3','영화갯수_4','영화갯수_5']],axis=1)
```

```
[ ] # 시청률 평균 => 리스트에 있는 값들을 평균내기
df['평균시청률'] = None
df['시청률'] = df['시청률'].str[1:-2]
for i in range(0, len(df)) :
    df['평균시청률'][i] = np.mean(list(map(float, df['시청률'][i].replace(" ", "").split(','))))
df['평균시청률']
```

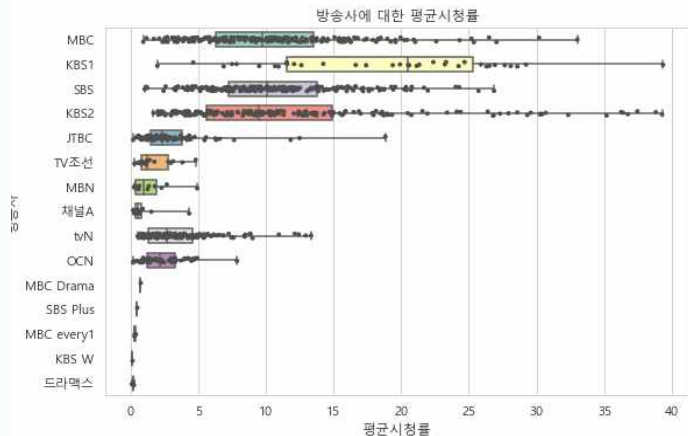
데이터 시각화 - 방송사

방송사별 드라마 갯수



MBC, SBS가 가장 많은 드라마를 방영
케이블 방송사 중 tvN이 가장 많은 드라마 방영

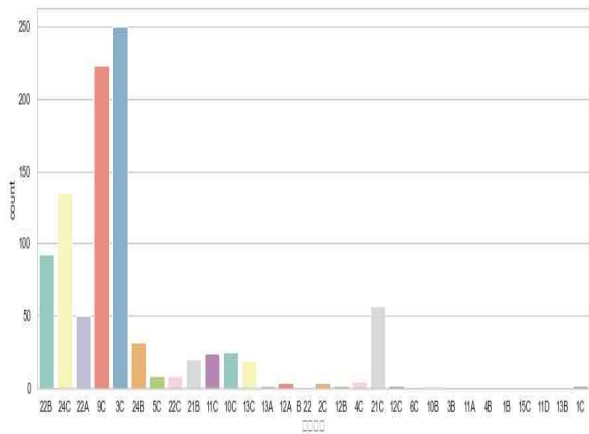
방송사에 대한 평균시청률



지상파 방송국이 케이블 방송국보다 시청률이 높게 나타남
지상파 방송국 중 KBS1이 타 방송국 대비 분산과 평균 모두 크게 나타남

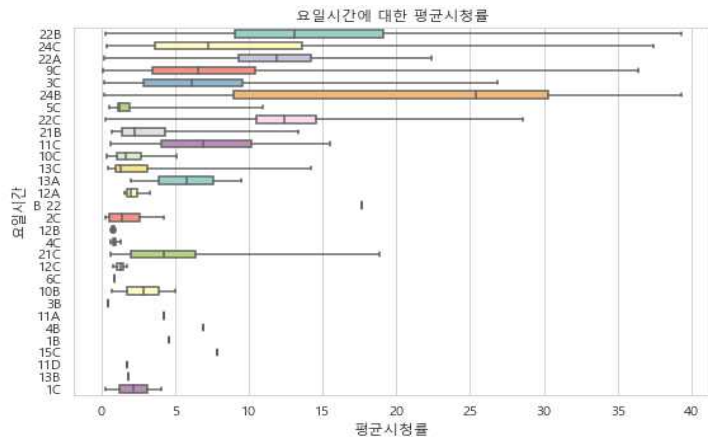
데이터 시각화 - 요일/방송시간대

요일/시간대 드라마 갯수



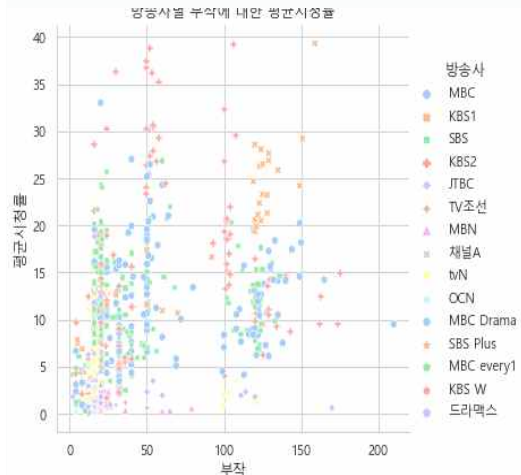
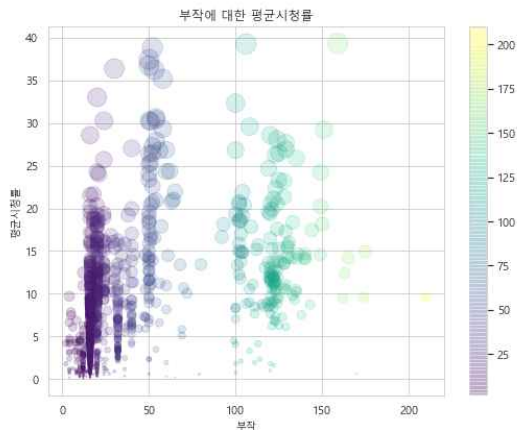
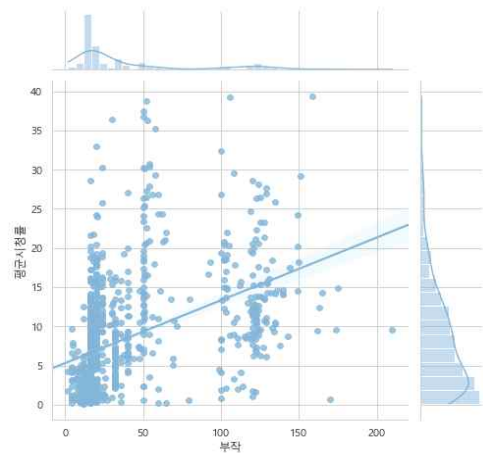
- 월-화20시이후, 수-목 20시이후에 가장 많은 드라마 방영
- 오전타임(08~12시)드라마의 경우 수가 적음
- => 오전드라마의 경우 100부작을 넘는 경우가 많아 시청률 유지가 힘들다고 생각

요일/시간대에 대한 평균시청률



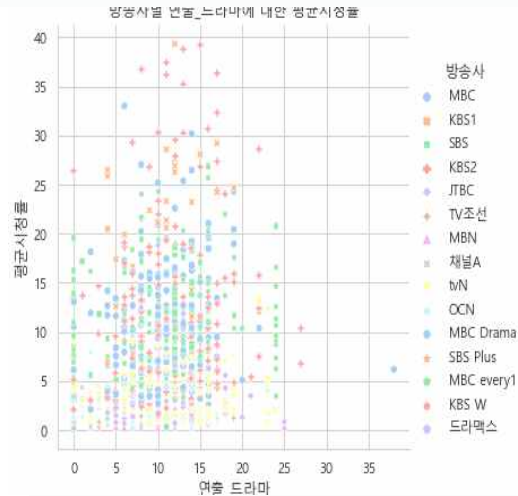
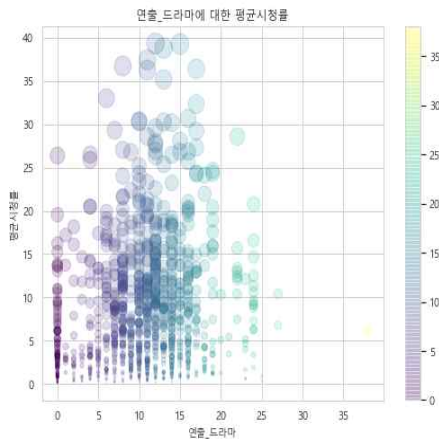
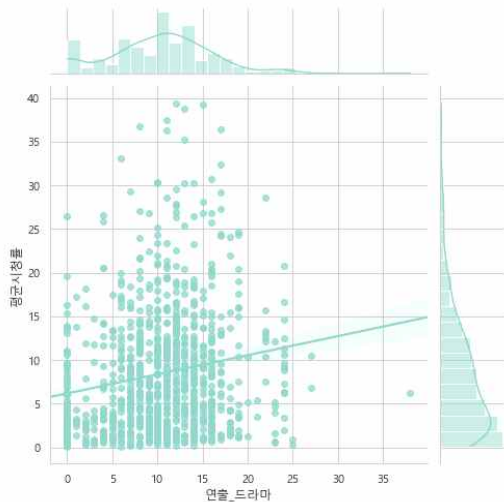
- 밤시간대의 시청률이 대체적으로 높음
- 일일드라마(월~금) 오전타임, 오후(12~20시)타임과 토~일 20시 이후 드라마는 타 시간대 대비 시청률이 높지만 분산이 크게 나타남

데이터 시각화 - 부작에 대한 평균시청률



- 대체로 부작횟수가 많을수록 평균시청률도 높아짐
- 케이블 대비 지상파의 드라마가 부작횟수가 높다는 것을 알 수 있음

데이터 시각화 - 연출_드라마에 대한 평균시청률

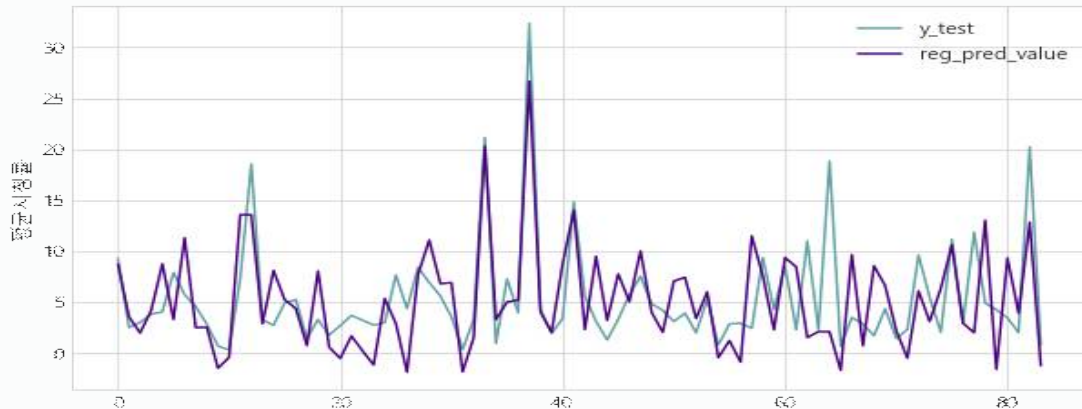


- 전반적으로 연출자가 제작한 드라마 갯수가 많을수록 평균시청률이 높아짐
- 연출자가 제작한 드라마 갯수가 적더라도 지상파에서 제작한 드라마의 평균시청률은 높다는 것을 알 수 있음

시청률 예측 모델 구현

- 사용변수 : 드라마 이름, 방송사, 부작, 요일시간, 배우_방송, 배우_영화, 연출_드라마, 연출_영화, 작가_드라마, 작가_영화, 강수량, 기온, 스포츠 대회 여부, 장르, 평균시청률
- 범주형 변수(방송사, 요일시간, 장르, 스포츠 대회여부) 더미화
 - 장르는 한 셀 당 범주값이 여러개씩 들어있어 다중 더미화 진행
- train data: 2010~2019년 데이터, test data: 2020년 데이터

모델 1 - Regression Model

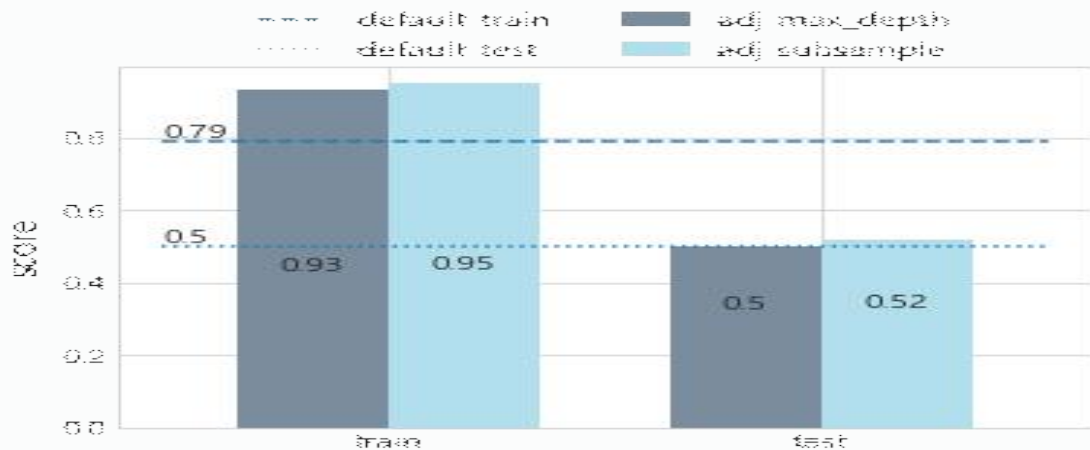


- 하이퍼파라미터가 없어 튜닝하지 않고 분석진행

train data score: 0.625, test data score: 0.346

변수의 계수값이 11.3244, 11.0062, 4.4841, 4.8773로 나타난 방송사_KBS1, 요일시간_24B, genre_막장, 방송사_KBS2가 중요 변수
(요일시간_24B: 토~일 12~20시 방영 드라마)

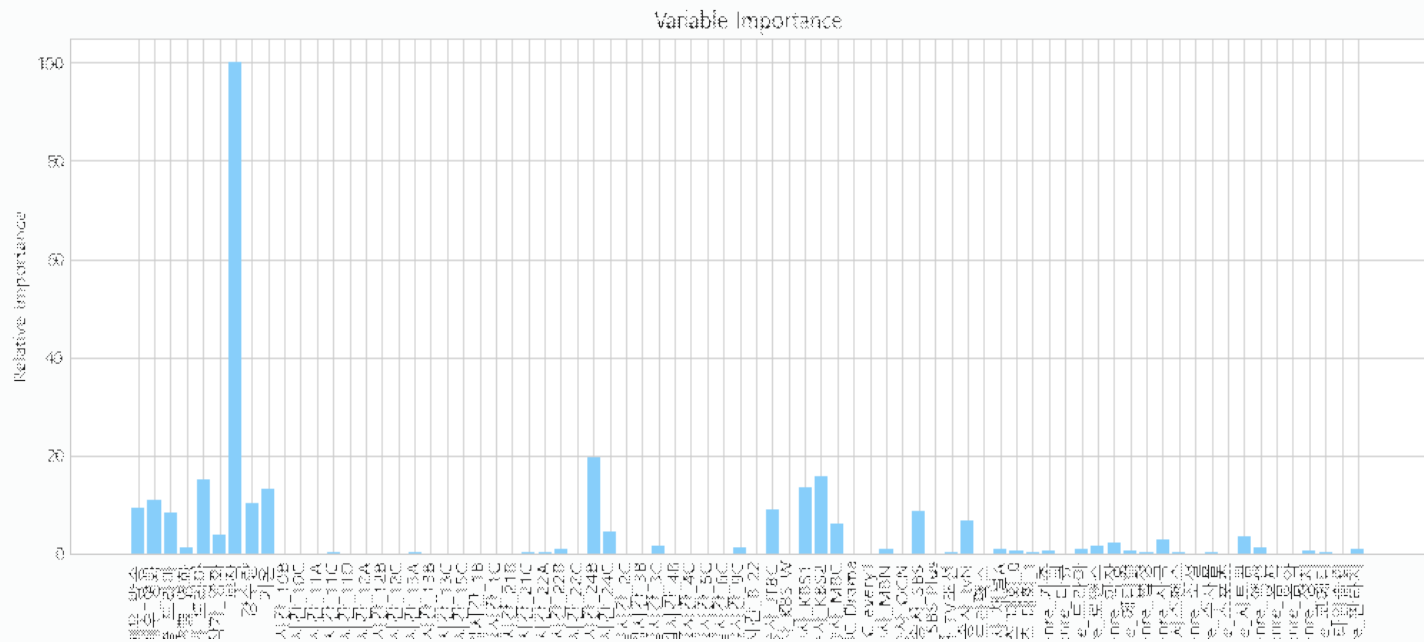
모델 2 - Gradient Boosting



- 하이퍼파라미터 튜닝후 분석 진행

- default test(random_state=0): train score: 0.794, test score: 0.497
- max_depth=5로 설정: train score: 0.931, test score: 0.497 => 훈련 데이터 스코어만 상승
- subsample=0.8로 추가 설정: train score: 0.950, test score: 0.520 => 두개 모두 상승

모델 2 - Gradient Boosting

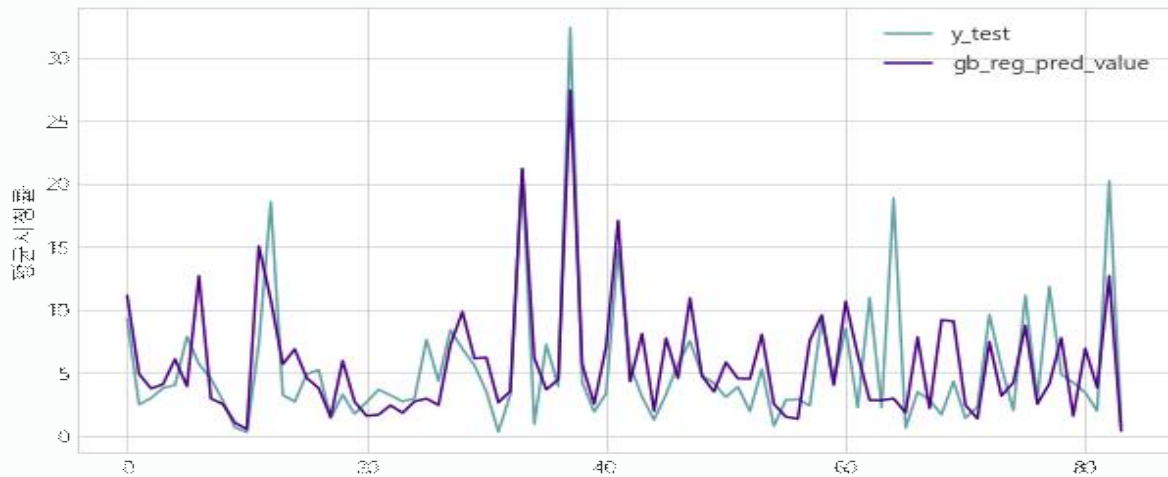


- 중요도 변수를 확인해보면 다음의 그래프와 같음

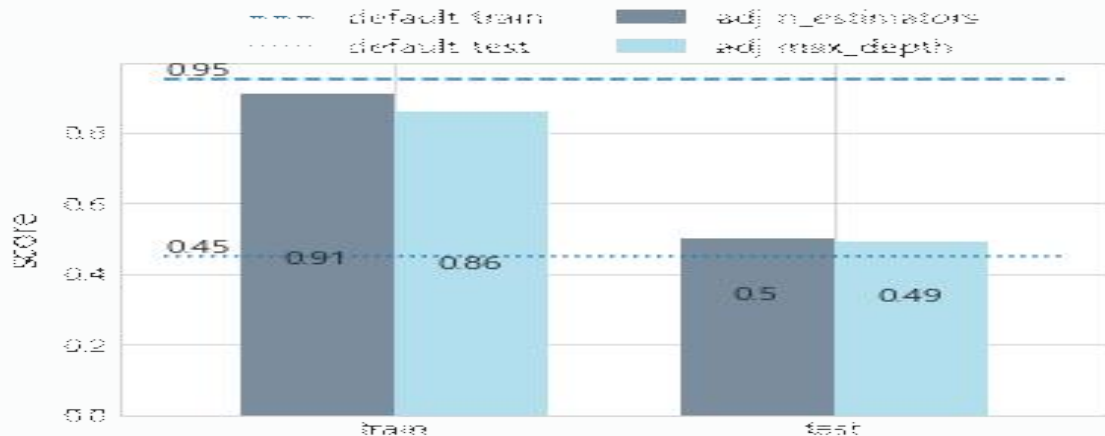
1. 부작, 2. 요일시간_24B, 3. 방송사_KBS2, 4. 작가_드라마 (요일시간_24B: 토~일 12~20시 방영 드라마)

모델 2 - Gradient Boosting

- 2020년도 데이터 test 결과 그래프



모델 3 - Random Forest



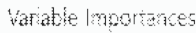
- 하이퍼파라미터 튜닝후 분석 진행

- default_test(random_state=0): train score: 0.952, test score: 0.451

- n_estimators=5(트리갯수를 5로 설정): train score: 0.907, test score: 0.501 => test score 상승

- max_depth=10(노드갯수 최대치 설정): train score: 0.858, test score: 0.488 => 두 개 모두 하락

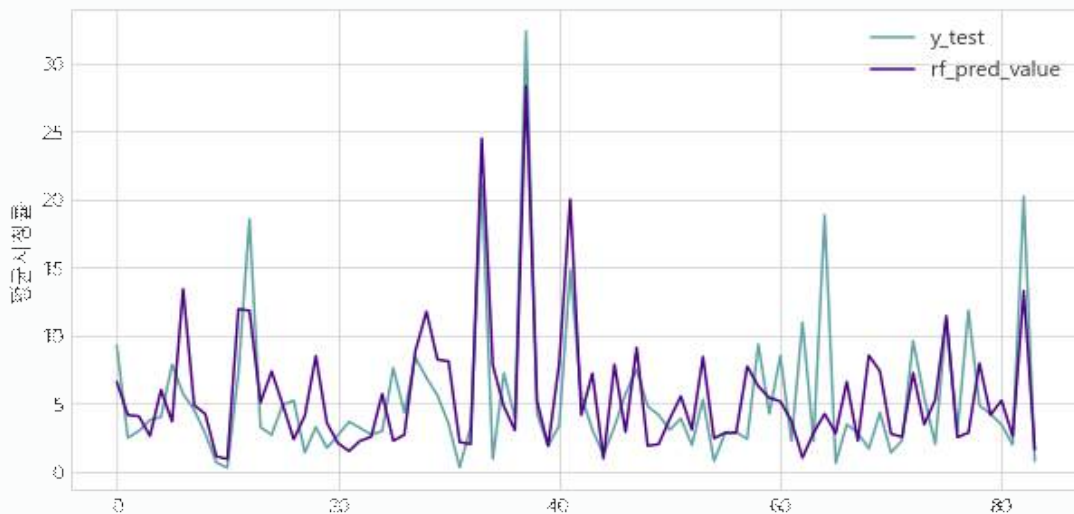
모델 3 - Random Forest



1. 부작, 2. 요일시간 24B, 3. 방송사 KBS2, 4. 방송사 KBS1 (요일시간 24B: 토~일 12~20시 방영 드라마)

모델 3 - Random Forest

- 2020년도 데이터 test 결과 그래프



결과

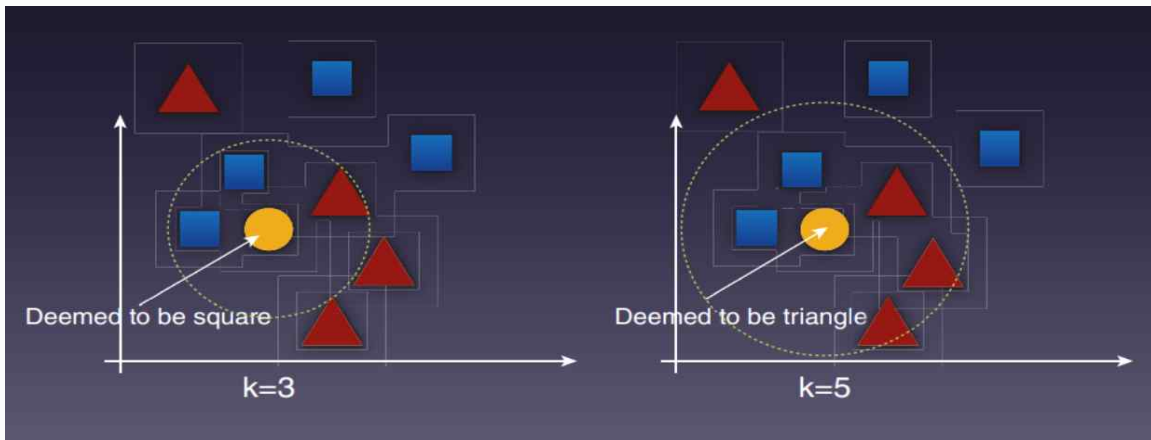
- 모델의 성능 지표를 RMSE로 설정하였고, 결과는 다음과 같음.
- Regression Model: 4.226, Gradient Boosting: 3.624, Random Forest: 3.694
Gradient Boosting, Random Forest, Regression Model 순으로 성능이 좋았음.

Chapter 03-2

드라마 추천 시스템

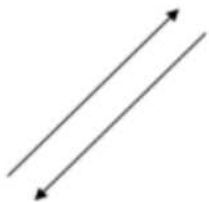
배경 및 원리

- 추천 시스템을 구현하기 위해서는 분류가 진행되어야 함
- K-Nearest Neighbors(KNN) 사용

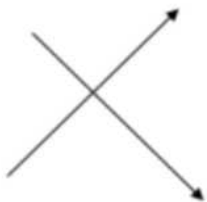


배경 및 원리

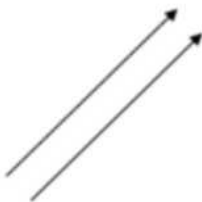
- 협업 필터링
- 드라마간의 유사도를 계산하여 유사도가 가장 가까운 드라마 3가지 선정



코사인 유사도 : -1



코사인 유사도 : 0



코사인 유사도 : 1

$$\text{식: } \text{simil}(x, y) = \cos(\vec{x} \cdot \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \times \|\vec{y}\|}$$

데이터 전처리

- 드라마명, 장르, 시청률 데이터 사용

- 지상파와 케이블 시청률 양상이 달라서 이를 고려하기 위해 z-score를 사용하여 표준화

$$z = \frac{x - \text{mean}}{\text{std}}$$

	A	B	C
1	강남스캔들	드라마	평균시청률
2	비켜라 운명아	드라마	7.28
3	하나뿐인 내편	막장,로맨스,코메디	20.52
4	내 사랑 치유기	드라마	39.28
5	황후의 품격	사극,로맨스,서스펜스,액션	13.42
6	신과의 약속	로맨스,드라마,가족	13.91
7	운명과 분노	로맨스	14.51
8	복수가 돌아왔다	로맨스,코메디,막장	6.11
9	일단 뜨겁게 청소하라	드라마,로맨스	5.35
10	SKY 캐슬	드라마	2.81
11	나쁜형사	수사물,서스펜스,액션	12.50
12	톱스타 유백이	드라마	6.92
13	남자친구	로맨스	2.45
14	대장금이 보고있다	코메디,드라마	8.48
15	알함브라 궁전의 추억	드라마	1.19
16	프리스트	드라마	8.44
17	차탈래 부인의 사랑	드라마,코메디	1.91
18	붉은 달 푸른 해	서스펜스,액션,수사물	8.30
19	비밀과 거짓말	드라마	4.71
20	신의 퀴즈 : 리부트	드라마	11.93
21	끝까지 사랑	드라마	2.28
22	커피야 부탁해	드라마	14.90
23	열두밤	로맨스	0.38
24	죽어도 좋아	코메디	0.35
25	계룡선녀전	코메디,로맨스,판타지	2.61
26	뽀빠이	가족,드라마	3.79

	A	B	C
1	드라마 이름	장르	시청률
2	강남스캔들	드라마	-0.12357
3	비켜라 운명아	드라마	1.895296
4	하나뿐인 내편	막장,로맨스,코메디	4.757312
5	내 사랑 치유기	드라마	0.812632
6	황후의 품격	사극,로맨스,서스펜스,액션	0.886878
7	신과의 약속	로맨스,드라마,가족	0.979431
8	운명과 분노	로맨스	-0.30309
9	복수가 돌아왔다	로맨스,코메디,막장	-0.41903
10	일단 뜨겁게 청소하라	드라마,로맨스	-0.30234
11	SKY 캐슬	드라마	3.118925
12	나쁜형사	수사물,서스펜스,액션	-0.1785
13	톱스타 유백이	드라마	-0.42975
14	남자친구	로맨스	1.701599
15	대장금이 보고있다	코메디,드라마	-1.0526
16	알함브라 궁전의 추억	드라마	1.68615
17	프리스트	드라마	-0.61794
18	차탈래 부인의 사랑	드라마,코메디	0.031019
19	붉은 달 푸른 해	서스펜스,액션,수사물	-0.51616
20	비밀과 거짓말	드라마	0.58481
21	신의 퀴즈 : 리부트	드라마	-0.48994
22	끝까지 사랑	드라마	1.037912
23	커피야 부탁해	드라마	-1.15792
24	열두밤	로맨스	-1.16969
25	죽어도 좋아	코메디	-0.83704
26	계룡선녀전	코메디,로맨스,판타지	0.044153

데이터 전처리

1

	A	B	C
1	강남스캔들	드라마	-0.12357
2	비켜라 운명아	드라마	1.895296
3	하나뿐인 내편	막장, 로맨스, 코메디	4.757312
4	내 사랑 치유기	드라마	0.812632
5	황후의 품격	사극, 로맨스, 서스펜스, 액션	0.886878
6	신과의 약속	로맨스, 드라마, 가족	0.979431
7	운명과 분노	로맨스	-0.30309
8	복수가 돌아왔다	로맨스, 코메디, 막장	-0.41903
9	일단 뜨겁게 청소하라	드라마, 로맨스	-0.30234
10	SKY 캐슬	드라마	3.118925
11	나쁜형사	수사물, 서스펜스, 액션	-0.1785
12	톱스타 유백이	드라마	-0.42975
13	남자친구	로맨스	1.701599
14	대장금이 보고있다	코메디, 드라마	-1.0526
15	알함브라 궁전의 추억	드라마	1.68615
16	프리스트	드라마	-0.61794
17	차탈레 부인의 사랑	드라마, 코메디	0.031019
18	붉은 달 푸른 해	서스펜스, 액션, 수사물	-0.51616
19	비밀과 거짓말	드라마	0.58481
20	신의 퀴즈 : 리부트	드라마	-0.48994
21	끝까지 사랑	드라마	1.037912

2

장르 더미화

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	강남스캔들	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	강남스캔들	0	2	3	0	0	0	0	0	0	0	0	0	0	0	0
3	비켜라 운명아	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
4	하나뿐인 내편	0	0	3	4	0	0	0	0	0	0	0	0	0	0	0
5	내 사랑 치유기	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
6	황후의 품격	0	0	3	0	0	0	0	7	8	0	0	0	12	0	0
7	신과의 약속	0	1	2	3	0	0	0	2	0	0	0	3	0	0	0
8	운명과 분노	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
9	복수가 돌아왔다	0	0	3	4	0	0	0	0	0	0	0	0	0	0	0
10	일단 뜨겁게 청소하라	0	2	3	0	0	0	0	0	0	0	0	0	0	0	0
11	SKY 캐슬	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
12	나쁜형사	0	0	0	0	0	0	0	0	8	9	0	0	12	0	0
13	톱스타 유백이	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
14	남자친구	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
15	대장금이 보고있다	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
16	알함브라 궁전의 추억	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
17	프리스트	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
18	차탈레 부인의 사랑	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
19	붉은 달 푸른 해	0	0	0	0	0	0	0	0	8	9	0	0	12	0	0
20	비밀과 거짓말	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
21	신의 퀴즈 : 리부트	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
22	끝까지 사랑	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
23	커피야 부탁해	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
24	열두밤	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
25	죽어도 좋아	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	계룡선녀전	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0

3

드라마이름, 장르, 시청률순으로
컬럼화

	A	B	C
1	drama_id	genre_id	rate
2	1	2	-0.12357
3	2	2	1.895296
4	3	3	4.757312
5	3	4	4.757312
6	3	14	4.757312
7	4	2	0.812632
8	5	3	0.886878
9	5	7	0.886878
10	5	8	0.886878
11	5	11	0.886878
12	6	1	0.979431
13	6	2	0.979431
14	6	3	0.979431
15	7	3	-0.30309
16	8	3	-0.41903
17	8	4	-0.41903
18	8	14	-0.41903
19	9	2	-0.30234
20	9	3	-0.30234
21	10	2	3.118925
22	11	8	-0.1785
23	11	9	-0.1785
24	11	11	-0.1785
25	12	2	-0.42975
26	13	3	1.701599
27	14	2	-1.0526
28	14	14	-1.0526
29	15	2	1.68615
30	16	2	-0.61794
31	17	2	0.031019
32	17	14	0.031019
33	18	8	-0.51616
34	18	9	-0.51616
35	18	11	-0.51616

모델 적용 및 결과

● 드라마간의 유사도 행렬

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	0	1	-1	1	-1	0	0	0	-1	0.5	0	1	0	-1	0	0.707107	-1
3	1	-1	1	-1	1	0	0	0	1	-0.5	0	-1	0	1	0	-0.70711	1
4	2	1	-1	1	-1	0	0	0	-1	0.5	0	1	0	-1	0	0.707107	-1
5	3	-1	1	-1	1	0	0	0	1	-0.5	0	-1	0	1	0	-0.70711	1
6	4	0	0	0	0	1	0.408248	0	0	0	0	0	0.57735	0	0	0	0
7	5	0	0	0	0	0.408248	1	0	0	0.353553	0.707107	0	0.707107	0	0.5	0	0
8	6	0	0	0	0	0	0	1	0	-0.35355	0	0	0	0	-0.5	-0.5	0
9	7	-1	1	-1	1	0	0	0	1	-0.5	0	-1	0	1	0	-0.70711	1
10	8	0.5	-0.5	0.5	-0.5	0	0.353553	-0.35355	-0.5	1	0.5	0.5	0	-0.5	0.707107	0.353553	-0.5
11	9	0	0	0	0	0	0.707107	0	0	0.5	1	0	0	0	0.707107	0	0
12	10	1	-1	1	-1	0	0	0	-1	0.5	0	1	0	-1	0	0.707107	-1
13	11	0	0	0	0	0.57735	0.707107	0	0	0	0	0	1	0	0	0	0
14	12	-1	1	-1	1	0	0	0	1	-0.5	0	-1	0	1	0	-0.70711	1
15	13	0	0	0	0	0	0.5	-0.5	0	0.707107	0.707107	0	0	0	1	0	0
16	14	0.707107	-0.70711	0.707107	-0.70711	0	0	-0.5	-0.70711	0.353553	0	0.707107	0	-0.70711	0	1	-0.70711
17	15	-1	1	-1	1	0	0	0	1	-0.5	0	-1	0	1	0	-0.70711	1
18	16	0	0	0	0	0	0.707107	0	0	0.5	1	0	0	0	0.707107	0	0
19	17	1	-1	1	-1	0	0	0	-1	0.5	0	1	0	-1	0	0.707107	-1
20	18	0	0	0	0	0.288675	0.353553	-0.35355	0	0.5	0.5	0	0	0	0.707107	0	0
21	19	1	-1	1	-1	0	0	0	-1	0.5	0	1	0	-1	0	0.707107	-1
22	20	0	0	0	0	0	0	-0.5	0	0.353553	0	0	0	0	0.5	0	0
23	21	1	-1	1	-1	0	0	0	-1	0.5	0	1	0	-1	0	0.707107	-1
24	22	0	0	0	0	0.816497	0.5	0	0	0	0	0	0.707107	0	0	0	0
25	23	1	-1	1	-1	0	0	0	-1	0.5	0	1	0	-1	0	0.707107	-1
26	24	0.57735	-0.57735	0.57735	-0.57735	0	0.408248	0	-0.57735	0.866025	0.57735	0.57735	0	-0.57735	0.408248	0.408248	-0.57735
27	25	0	0	0	0	-0.40825	-1	0	0	-0.35355	-0.70711	0	-0.70711	0	-0.5	0	0
28	26	-0.70711	0.707107	-0.70711	0.707107	-0.40825	0	0	0.707107	-0.35355	0	-0.70711	0	0.707107	0	-0.5	0.707107
29	27	0	0	0	0	0	-0.70711	0	0	-0.5	-1	0	0	0	-0.70711	0	0
30	28	-1	1	-1	1	0	0	0	1	-0.5	0	-1	0	1	0	-0.70711	1

● 추천 시스템의 정확도 측정 지표로 RMSE 사용

7: 3으로 training, test set으로 분리

84~88% 정확도

```
In [14]: np.sqrt(get_mse(user_pred, ratings_train))
```

```
Out[14]: 0.8446533710119778
```

```
In [15]: np.sqrt(get_mse(user_pred, ratings_test))
```

```
Out[15]: 0.8811544163817294
```

모델 적용 및 결과

- 각 드라마와 유사도가 가장 가까운 드라마 5개 선정

유사도 가까운 순서

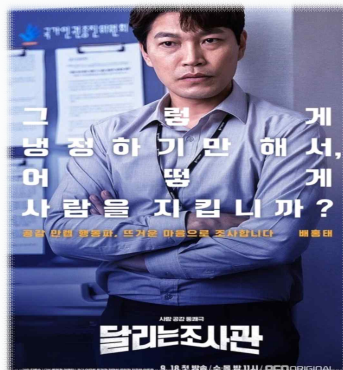
드라마 번호

추천 드라마 번호

	A	B	C	D	E	F
1		1	2	3	4	5
2	1	177	141	175	62	55
3	2	85	124	111	114	179
4	3	38	73	70	142	178
5	4	85	124	111	114	179
6	5	98	31	42	155	49
7	6	74	154	29	58	67
8	7	95	56	34	46	16
9	8	170	192	184	13	109
10	9	185	110	47	24	87
11	10	85	124	111	114	179
12	11	158	127	189	181	118
13	12	177	141	175	62	55
14	13	104	92	84	27	60
15	14	162	66	36	93	90
16	15	85	124	111	114	179
17	16	177	141	175	62	55
18	17	111	15	114	63	61
19	18	118	181	127	186	108
20	19	85	124	111	114	179
21	20	177	141	175	62	55
22	21	85	124	111	114	179
23	22	177	141	175	62	55
24	23	95	56	34	46	16
25	24	197	122	13	192	99
26	25	70	142	73	178	69

모델 적용 및 결과

- 예시: 닥터 프리즈너의 추천 드라마



Chapter 04

웹 시연

Q&A

감사합니다.
