KUBIG ML GROUP ASSIGNMENT

PREDICTION OF MOVIE AUDIENCE

영화 관객수 예측

Group 2

15기 김지호
15기 염윤석
15기 우명진
15기 이제윤

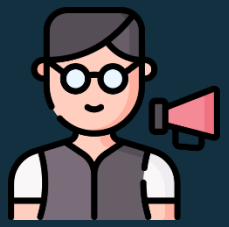
CONTENTS

CONTENT
1

과제 목표

CONTENT
1

과제 목표



봉준호 감독

관객은 몇 명?

송강호 이선균 조여정...

2019-05-30 개봉

스릴러 장르

CONTENT
2

변수 설명

- title: 영화의 제목
- distributor: 배급사 종류*(롯데, CJ, ...)*
- genre: 영화의 장르*(드라마, 코미디, ...)*
- release_time: 개봉한 "년도-월-일" str
- time: 영화 러닝타임[분]
- screening_rat: 상영 관람 등급

	title	distributor	genre	release_time	time	screening_rat	director	dir_prev_bfnum	dir_prev_num	num_staff	num_actor	box_off_num
0	개들의 전쟁	롯데엔터테인먼트	액션	2012-11-22	96	청소년 관람불가	조병옥	NaN	0	91	2	23398
1	내부자들	(주)쇼박스	느와르	2015-11-19	130	청소년 관람불가	우민호	1161602.50	2	387	3	7072501
2	은밀하게 위대하게	(주)쇼박스	액션	2013-06-05	123	15세 관람가	장철수	220775.25	4	343	4	6959083
3	나는 공무원이다	(주)NEW	코미디	2012-07-12	101	전체 관람가	구자홍	23894.00	2	20	6	217866
4	불량남녀	쇼박스(주)미디어플렉스	코미디	2010-11-04	108	15세 관람가	신근호	1.00	1	251	2	483387

CONTENT
2

변수 설명

director: 감독명

dir_prev_bfnum
the director's previous box-office number
감독별 이전 관객수 평균

dir_prev_num
the director's previous (movies) number
감독별 이전 출품 영화 수

num_staff

num_actor: 주연 배우수

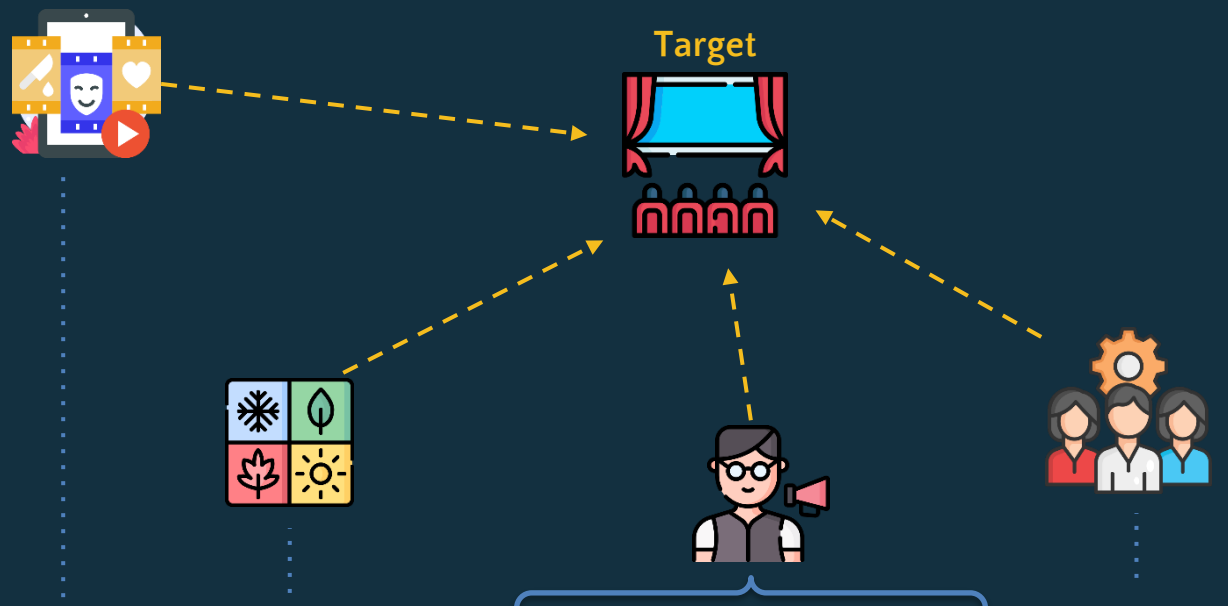
box_off_num: 관객수

감독 관련 **columns**

	title	distributor	genre	release_time	time	screening_rat	director	dir_prev_bfnum	dir_prev_num	num_staff	num_actor	box_off_num
0	개들의 전쟁	롯데엔터테인먼트	액션	2012-11-22	96	청소년 관람불가	조병옥	NaN	0	91	2	23398
1	내부자들	(주)쇼박스	느와르	2015-11-19	130	청소년 관람불가	우민호	1161602.50	2	387	3	7072501
2	은밀하게 위대하게	(주)쇼박스	액션	2013-06-05	123	15세 관람가	장철수	220775.25	4	343	4	6959083
3	나는 공무원이다	(주)NEW	코미디	2012-07-12	101	전체 관람가	구자홍	23894.00	2	20	6	217866
4	불량남녀	쇼박스(주)미디어플렉스	코미디	2010-11-04	108	15세 관람가	신근호	1.00	1	251	2	483387

CONTENT
2

변수 설명



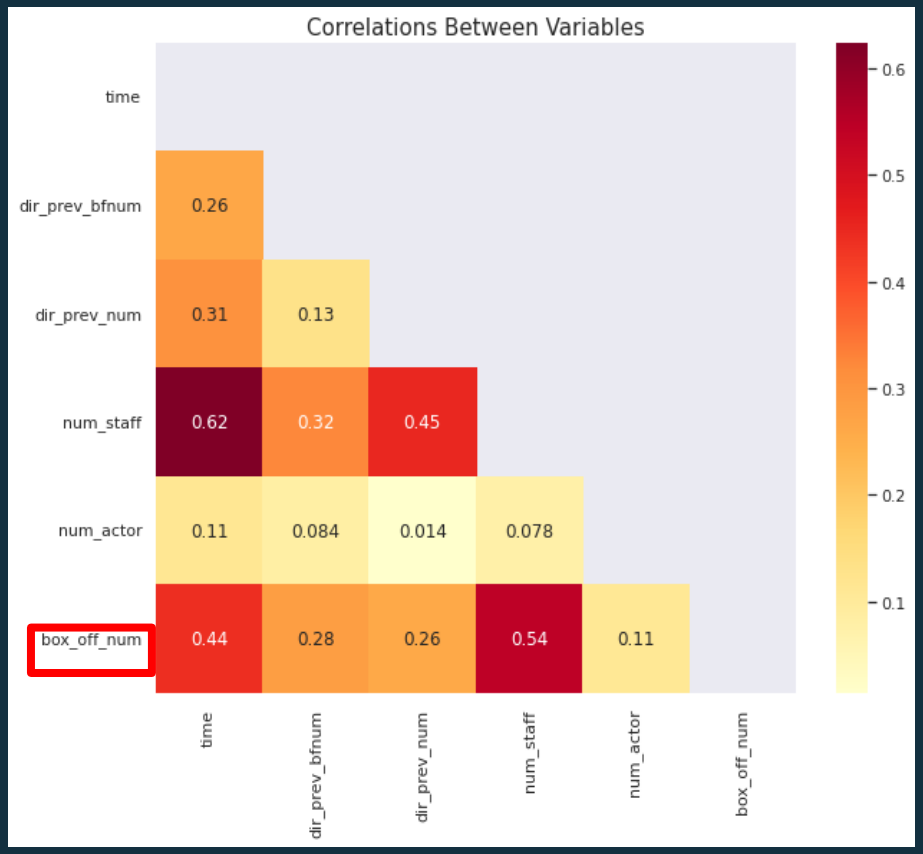
Target

	title	distributor	genre	release_time	time	screening_rat	director	dir_prev_bfnum	dir_prev_num	num_staff	num_actor	box_off_num
0	개들의 전쟁	롯데엔터테인먼트	액션	2012-11-22	96	청소년 관람불가	조병옥	NaN	0	91	2	23398
1	내부자들	(주)쇼박스	느와르	2015-11-19	130	청소년 관람불가	우민호	1161602.50	2	387	3	7072501
2	은밀하게 위대하게	(주)쇼박스	액션	2013-06-05	123	15세 관람가	장철수	220775.25	4	343	4	6959083
3	나는 공무원이다	(주)NEW	코미디	2012-07-12	101	전체 관람가	구자홍	23894.00	2	20	6	217866
4	불량남녀	쇼박스(주)미디어플렉스	코미디	2010-11-04	108	15세 관람가	신근호	1.00	1	251	2	483387

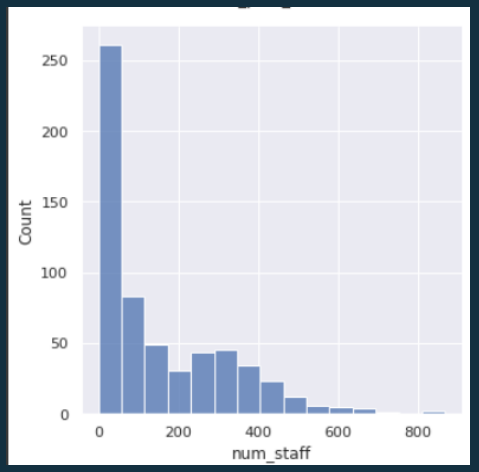
CONTENT 3

EDA + Preprocessing - Numerical Data

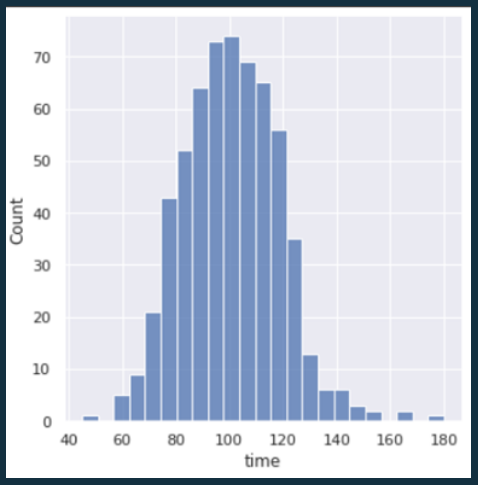
#영화 제목, 영화 감독 → drop



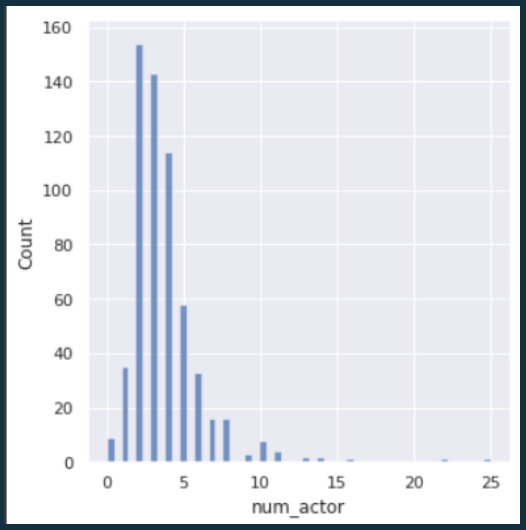
#스텝 수(0.54)



#상영 시간(0.44)



#주연 배우 수(0.11)



CONTENT
3

EDA + Preprocessing
- Numerical Data

#감독이 이전에 참여한 영화의 관객 수

```
df.isnull().sum()
```

title	0
distributor	0
genre	0
release_time	0
time	0
screening_rat	0
director	0
dir_prev_bfnum	330
dir_prev_num	0
num_staff	0
num_actor	0
box_off_num	0

dtype: int64

55%

```
hss_idx=set(df.index[df.director == "홍상수"])
df.loc[df.director == "홍상수","dir_prev_bfnum":"dir_prev_num"]
```

	dir_prev_bfnum	dir_prev_num
15	NaN	0
19	NaN	0
115	NaN	0
164	39317.0	1
331	NaN	0
506	NaN	0
523	NaN	0

NaN ≠ 데뷔작

(시도1) 직접 계산해서 채우기 → target 변수 정보가 필요

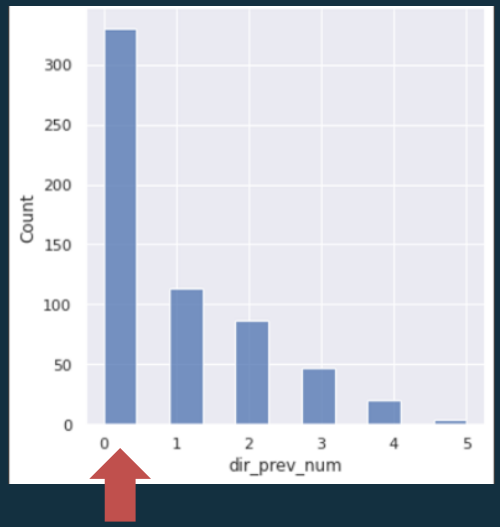
(시도2) median / 0로 단순 대체

➡ 결측치 0으로 대체

CONTENT 3

EDA + Preprocessing - Numerical Data

#감독이 이전에 참여한 영화의 개수




```
hss_idx=set(df.index[df.director == "홍상수"])
df.loc[df.director == "홍상수","dir_prev_bfnum":"dir_prev_num"]
```

	dir_prev_bfnum	dir_prev_num
15	NaN	0
19	NaN	0
115	NaN	0
164	39317.0	1
331	NaN	0
506	NaN	0
523	NaN	0

상관계수 : 0.26 0.2 ☹

Why?

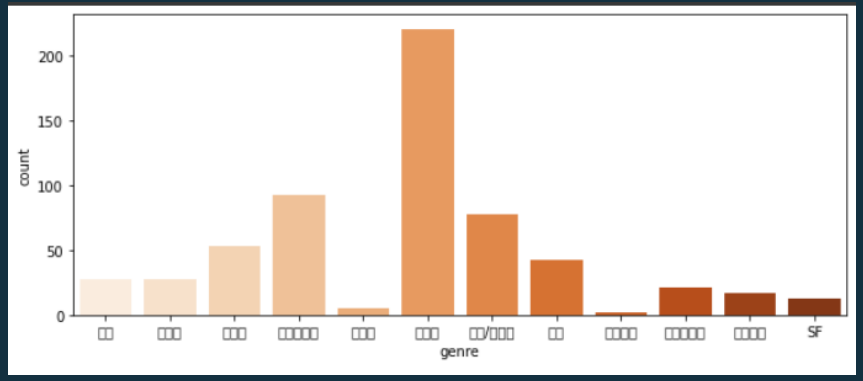
Data에 한 감독의 **모든** 작품이 반영된 것이 아니므로 이렇게 결측치를 채우면 정보 왜곡임

➡ **그대로 반영**

CONTENT 3

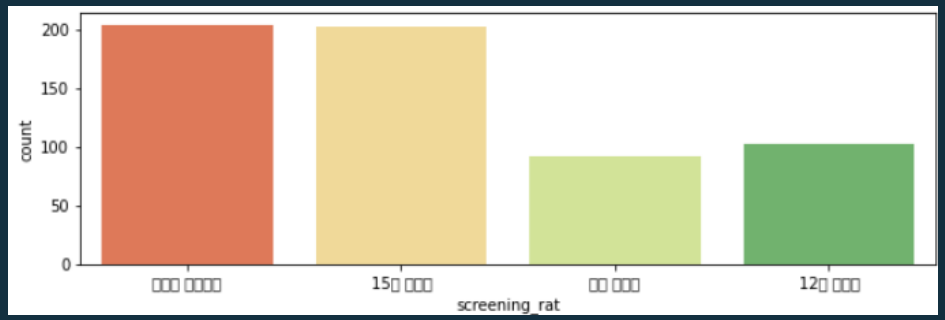
EDA + Preprocessing - Categorical Data

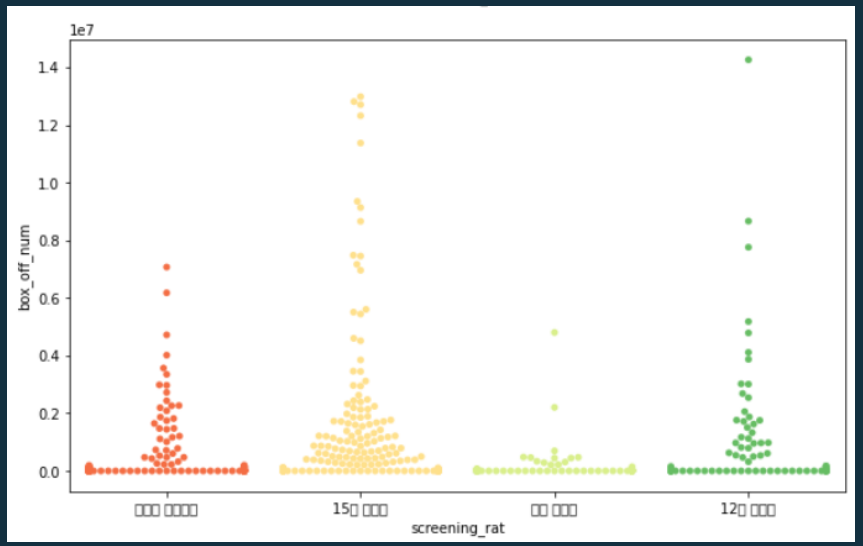
#장르



장르별 영화 관객 수

#상영 등급





상영 등급별 영화 관객 수

CONTENT 3

EDA + Preprocessing - Categorical Data

#배급사

```
CJ 엔터테인먼트          54
롯데엔터테인먼트          52
(주)NEW              30
(주)마운틴픽쳐스          29
(주)쇼박스             26
                    ..
OAL(올)               1
(주)에이원 엔터테인먼트      1
(주)콘텐츠 윙             1
위더스필름               1
퍼스트런                 1
Name: distributor, Length: 169, dtype: int64
```

→ [독점 배급사 5개 + 기타] 로 범주화

```
CJ 엔터테인먼트          54
롯데엔터테인먼트          52
(주)NEW              30
(주)마운틴픽쳐스          29
(주)쇼박스             26
인디스토리              26
골든타이드픽처스          15
(주) 케이알씨지          14
(주) 영화사조제          10
영화사 진진             10
시네마달                10
어뮤즈                 10
(주)키노아이            10
시너지                 10
인디플러그               9
NEW                  8
인벤트 디               8
KT&G 상상마당           8
```

A

B

C

D

→ 작품 수 5단위로 묶어서 범주화

독점 배급사들을 한 그룹으로 묶는 건 비효율적

CONTENT
3

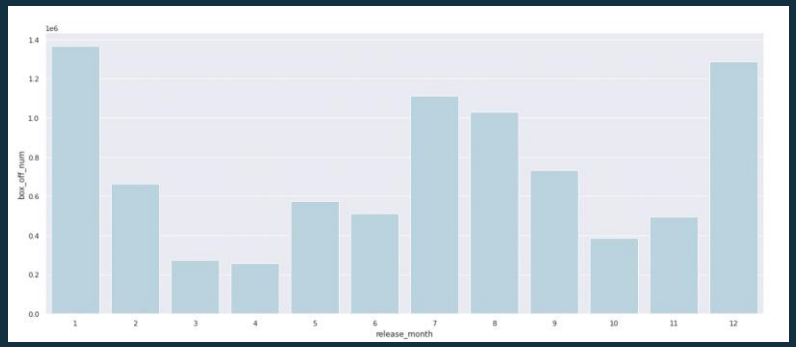
EDA + Preprocessing
– Categorical Data

#개봉일

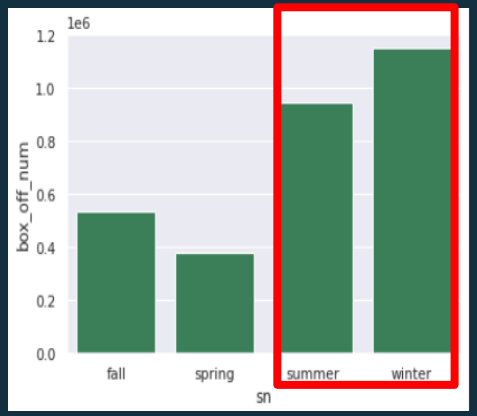
release_time
2012-11-22
2013-05-23
2014-09-18
2012-03-15
2015-07-27
...
2013-09-12
2014-03-20
2010-09-30
2015-05-14
2013-01-30

개봉 연도 / 월 / 일자

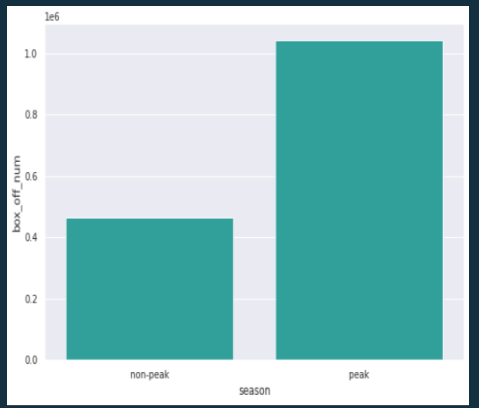
개봉월



계절별 영화 관객 수



성수기/비수기 영화 관객 수



- 관객 수는 여름/겨울에 집중 되어있음
- 사계절 / peak non-peak 두 가지 전처리 시도함

➡ 성수기/비수기로 범주화

CONTENT
3

Preprocessing

최종 전처리 함수

사용자 정의 함수

```
# 개봉일자에서 연월일 분리
def release_time(x):
  return str(x).split("-")


# 월별로 peak/ non-peak 두 부류로 구분
def season(x):
  if x in [3,4,5,9,10,11]:
    return "non-peak"
  else:
    return "peak"

# 배급사를 5개사 + 기타로 그룹 짓기
def distributor(x, top5):
  if x in top5:
    return x
  else:
    return "기타"
```

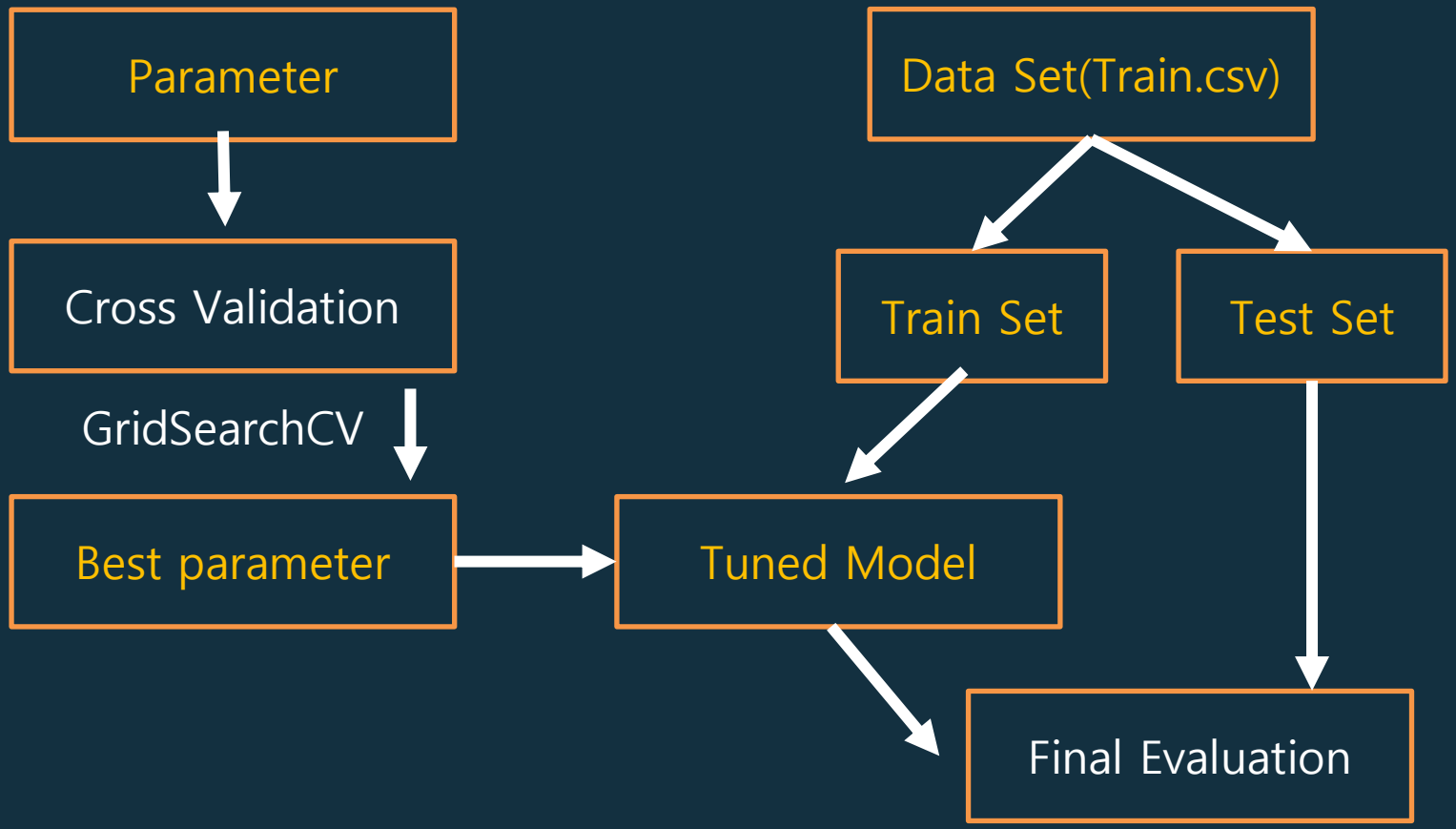
최종 전처리 함수

```
def preprocessing_data(data):
  df = data.copy()

  #release time--> season(peak vs non peak)
  df["release_month"] = df["release_time"].map(lambda x : int(release_time(x)[1]
  df["season"] = df["release_month"].map(lambda x : season(x))

  #distributor
  top5 = ['CJ 엔터테인먼트', '롯데엔터테인먼트', '(주)NEW', '(주)마운틴픽쳐스', '(주)쇼박스']
  df["distributor"] = df["distributor"].map(lambda x: distributor(x, top5))

  #dir_prev_bfnum
  df["dir_prev_bfnum"].fillna(0, inplace=  True)

  #dropping
  df.drop(["title"], axis = 1, inplace = True)
  df.drop(["release_time","release_month"], axis =1, inplace = True)
  df.drop("director", axis =1, inplace=  True)

  df_num = df.select_dtypes(include = ["float64", "int64"]).copy()
  df_cat = df.select_dtypes(include = ["object", "category"]).copy()
  col_num = list(df_num.columns)
  col_cat = list(df_cat.columns)

  print("raw data column : ", list(data.columns))
  print("numerical columns : ", col_num)
  print("categorical columns : ", col_cat)

  df_dummies = pd.get_dummies(df_cat, drop_first = True)



  return df_num, df_dummies
```

CONTENT
4

4. Modeling & Evaluation

[Dataset]
1. train.csv : target variable 포함
2. test.csv : target variable 미포함 (Prediction 대상 데이터)



CONTENT
4

4. Modeling & Evaluation

1. Lasso Regression

Best estimator : Lasso(alpha = 10, random_state = 42)
RMSE(Test) : 1172308
R2(Test) : 0.223

2. Ridge Regression

Best estimator : Ridge(alpha = 10, random_state = 42)
RMSE(Test) : 1160797
R2(Test) : 0.238

3. Random Forest Regressor(Bagging)

Best estimator : RandomForestRegressor(max_depth = 5, max_features=8, n_estimators = 500)
RMSE(Test) : 1139180
R2(Test) : 0.266

4. Extra Trees Regressor

Best estimator : ExtraTreesRegressor(max_depth = 5, max_features = 8, n_estimators = 700)
RMSE(Test) : 1136197
R2(Test) : 0.2702

CONTENT 4

4. Modeling & Evaluation

1. Lasso Regression
2. Ridge Regression
3. Random Forest Regressor
4. Extra Tree Regressor

[Ensemble]
Voting & Stacking

5. Voting

RMSE(Test) : 1106627
R2(Test) : 0.3077

5. Stacking

RMSE(Test) : 1142044
R2(Test) : 0.26268

CONTENT
4

4. Modeling & Evaluation

[DACON 제출 결과]

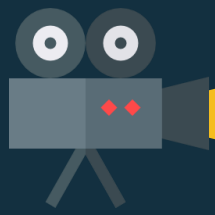
"Linear regressor 기반 학습보다는 Decision Tree 기반의 Ensemble들"

<RMSE>
RandomForestRegressor : 1337678

ExtraTreeRegressor : 1364790

[conclusion]

- numerical data vs categorical data
  - 수치형 데이터와 범주형 데이터 개수 차이
  - 선형 회귀에 맞는 데이터 형태이기보다는 Tree 기반의 모델이 더 좋다.



Thank you For Listing!