

**How do you
Value our
Movie?**



CONTENTS

CONTENT 1

Choosing Topic

CONTENT 2

Data Crawling

CONTENT 3

Data Preprocessing & EDA

CONTENT 4

Modeling

CONTENT 5

Result

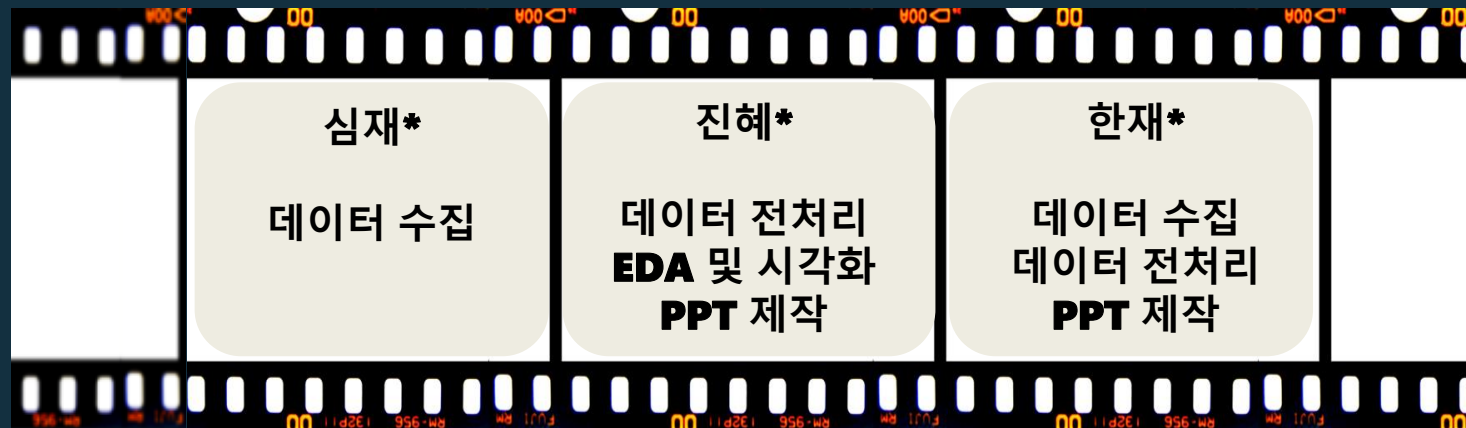
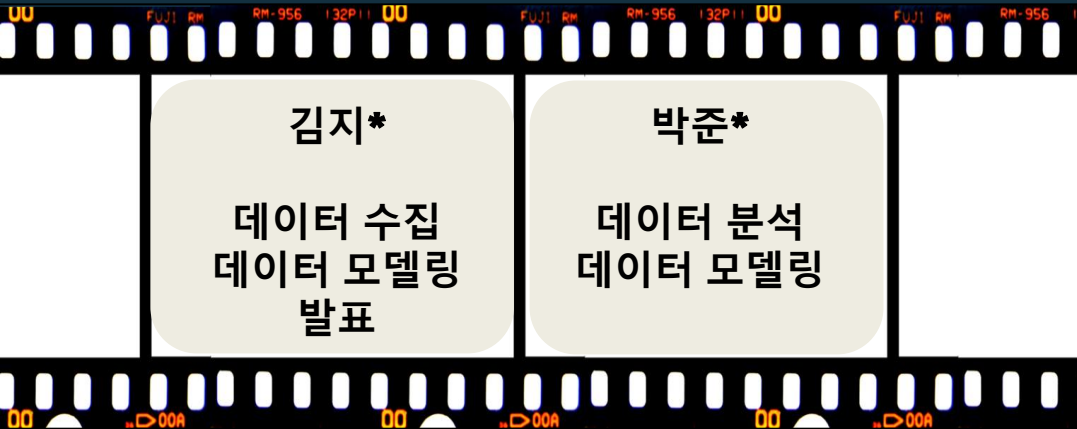
CONTENT 6

Evaluate



Member Roles

CONTENT 0





Choosing Topic

CONTENT

1



**최근 한국 미디어 시장이 주목받고, 성장함에 따라 매출액 극대화, 안정화를 위하여
배급사에서는 수요를 예측하고 그에 따른 전략 수립 필요
→ 객관적인 데이터를 이용하여 적정 제작비용 예측 후 영화 산업 발전에 기여**



Data Collecting

CONTENT 2



- 영화진흥위원회의 2010 ~ 2020년 영화관 입장권 데이터 수집

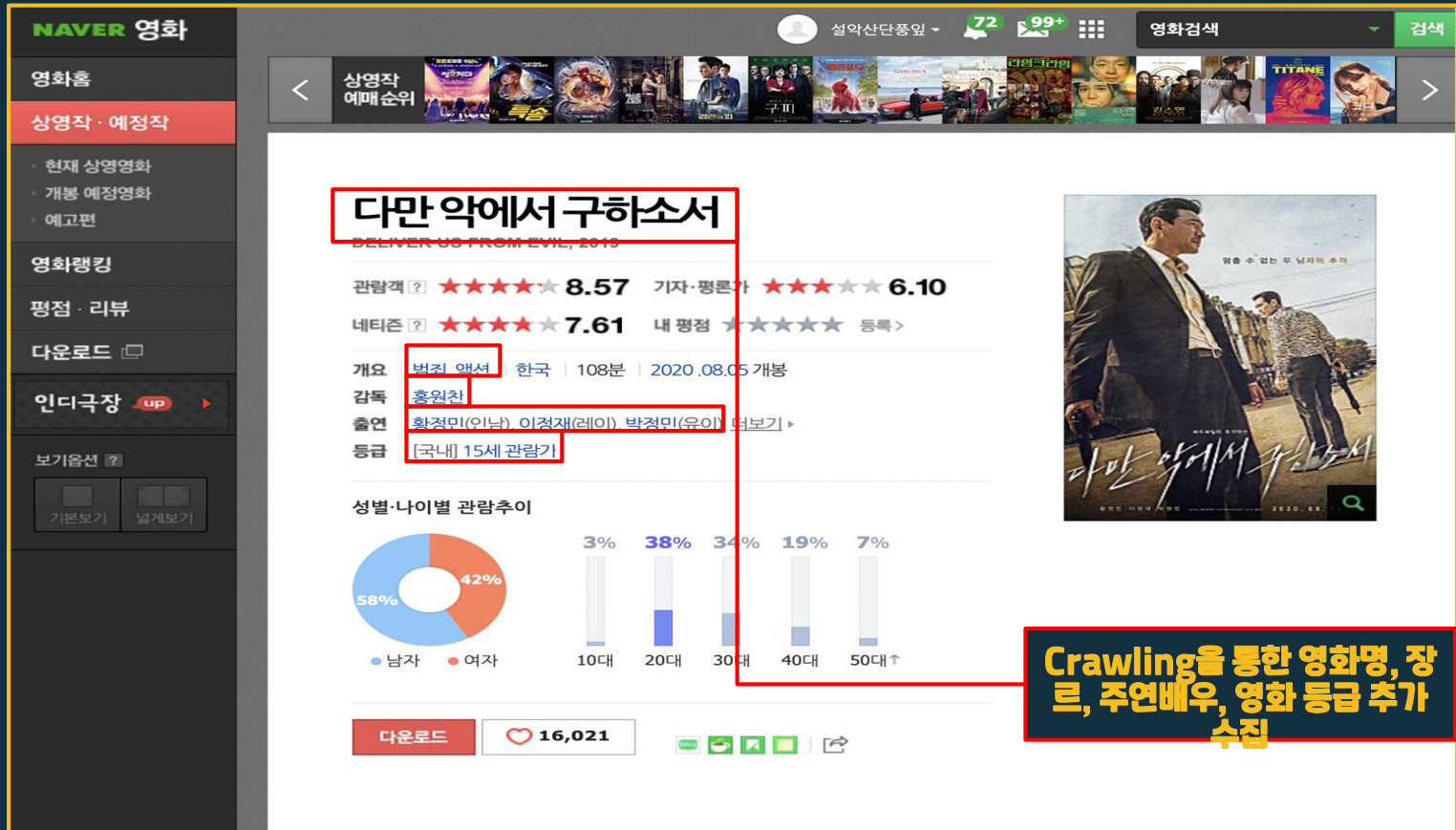


- Crawling을 통한 추가 데이터 수집

Data Collecting (Crawling)

CONTENT

2





Data Collecting (Merge)

CONTENT 2

<Kofic 영화 Data>

	영화명	개봉일	매출액	관객수	스크린수	상영횟수	국적	배급사
0	극한직업	2019.1.23	1.396518e+11	16265618.0	2003.0	292584.0	한국	(주)씨제이이엔엠
1	명량	2014.7.30	1.357484e+11	17613682.0	1587.0	188611.0	한국	(주)씨제이이엔엠
2	어벤져스: 엔드게임	2019.4.24	1.221827e+11	13934592.0	2835.0	242001.0	미국	월트디즈니컴퍼니코리아 유한책임회사
3	베테랑	2015.8.5	1.051682e+11	13414009.0	1115.0	199231.0	한국	(주)씨제이이엔엠
4	신과함께-인과 연	2018.8.1	1.026661e+11	12274996.0	2235.0	179981.0	한국	롯데컬처웍스(주)롯데엔터테인먼트
...	...	...	...	...	...	...	...	...
1042	더 프린세스 : 도둑맞은 공주	2020.8.12	4.574497e+08	55828.0	423.0	4192.0	우크라이나	(주)팝엔터테인먼트

+

<Crawling 영화 Data>

	director	genre	actor				filmrank
0	이병헌	코미디	류승룡(고반장), 이하늬(장형사), 진선규(마형사)				[국내] 15세 관람가
1	정세교	다큐멘터리	김한민, 오타니 료헤이, 이해영				[국내] 12세 관람가
2	안소니 루소	액션	로버트 다우니 주니어(토니 스타크 / 아이언맨), 크리스 에반스(스티브 로저스 / ...				[국내] 12세 관람가
3	류승완	액션	황정민(서도철), 유아인(조태오), 유해진(최상무)				[국내] 15세 관람가
4	김용화	판타지	하정우(강림), 주지훈(해원맥), 김향기(덕춘)				[국내] 12세 관람가
...	...	...	...				...



Data Collecting (Crawling)

CONTENT 2

<Merge Data>

	moviename	openingdate	sales	attendance	screen	playcount	country	distributor	director	genre	actor	filmrank
0	극한직업	2019.1.23	1.396518e+11	16265618.0	2003.0	292584.0	한국	(주)씨제이이엔엠	이병헌	코미디	류승룡(고반장), 이하늬(장형사), 진선규(마형사)	[국내] 15세 관람가
1	명량	2014.7.30	1.357484e+11	17613682.0	1587.0	188611.0	한국	(주)씨제이이엔엠	정세교	다큐멘터리	김한민, 오타니 료헤이, 이혜영	[국내] 12세 관람가
2	어벤져스: 엔드게임	2019.4.24	1.221827e+11	13934592.0	2835.0	242001.0	미국	월트디즈니컴퍼니코리아 유한책임회사	안소니 루소	액션	로버트 다우니 주니어(토니 스타크 / 아이언맨), 크리스 에반스(스티브 로저스 / ...	[국내] 12세 관람가
3	베테랑	2015.8.5	1.051682e+11	13414009.0	1115.0	199231.0	한국	(주)씨제이이엔엠	류승환	액션	황정민(서도철), 유아인(조태오), 유해진(최상무)	[국내] 15세 관람가
4	신과함께-인과 연	2018.8.1	1.026661e+11	12274996.0	2235.0	179981.0	한국	롯데컬처웍스(주)롯데엔터테인먼트	김용화	판타지	하정우(강림), 주지훈(해원맥), 김향기(덕춘)	[국내] 12세 관람가
...	...	...	...	...	...	...	...	...	...	...	...	...
1042	더 프린세스 : 도둑맞은 공주	2020.8.12	4.574497e+08	55828.0	423.0	4192.0	우크라이나	(주)팝엔터테인먼트	올레 말라르	애니메이션	사문영(밀라 목소리), 남도형(루슬란 목소리), 서반석(레스터 목소리)	[국내] 전체 관람가
1043	프리키 데스데이	2020.11.25	4.512257e+08	51812.0	631.0	12882.0	미국	유니버설픽처스인터내셔널 코리아(유)	크리스토퍼 랜던	코미디	빈스 본, 캐서린 뉴튼	[국내] 15세 관람가
1044	날씨의 아이	2019.10.30	4.449562e+08	58968.0	194.0	5921.0	일본	메가박스중앙(주)플러스엠, (주)미디어캐슬, 워너홀컴퍼니(주)	신카이 마코토	애니메이션	심규혁(호다카), 김유림(히나), 최한(스가)	[국내] 15세 관람가
1045	눈의 여왕4	2019.12.24	4.446496e+08	67159.0	271.0	2239.0	러시아	(주)넥스트엔터테인먼트월드(NEW)	로버트 렌스	애니메이션	양정화(눈의 여왕 한국어 목소리), 박지윤(겔다 한국어 목소리), 임채현(헤럴드 한...)	[국내] 전체 관람가
1046	다크 나이트	2008.8.6	4.071936e+08	51681.0	378.0	4445.0	미국	주식회사 해리슨엔컴퍼니, 워너브러더스 코리아(주)	크리스토퍼 놀란	액션	크리스찬 베일(브루스 웨인/배트맨), 히스 레저(조커), 아론 에크하트(하비 덴트)/...	[국내] 15세 관람가 [해외] PG-13

12column X 1047row

영화명, 개봉일, 누적매출액, 누적 관객수, 스크린 수, 상영횟수...



Data Preprocessing & EDA

CONTENT

3

	sales	attendance	screen	playcount	openingmonth
count	1047.0	1047.0	1047.0	1047.0	1047.0
mean	13516930095.5	1702492.7	693.2	44465.4	6.6
std	17335850996.1	2158083.1	388.6	39056.9	3.6
min	407193640.0	51681.0	33.0	2179.0	1.0
25%	3338669977.5	436801.5	416.5	17927.5	3.0
50%	7013842000.0	898945.0	592.0	30736.0	7.0
75%	16707136952.5	2091533.0	883.0	57129.5	10.0
max	139651845516.0	17613682.0	2835.0	292584.0	12.0

#	Column	Non-Null Count	Dtype
0	moviename	1047 non-null	object
1	openingdate	1047 non-null	datetime64[ns]
2	sales	1047 non-null	float64
3	attendance	1047 non-null	float64
4	screen	1047 non-null	float64
5	playcount	1047 non-null	float64
6	country	1047 non-null	object
7	distributor	1047 non-null	object
8	director	1047 non-null	object
9	genre	1047 non-null	object
10	actor	1044 non-null	object
11	filmrnk	1022 non-null	object

dtypes: datetime64[ns](1), float64(4), object(7)

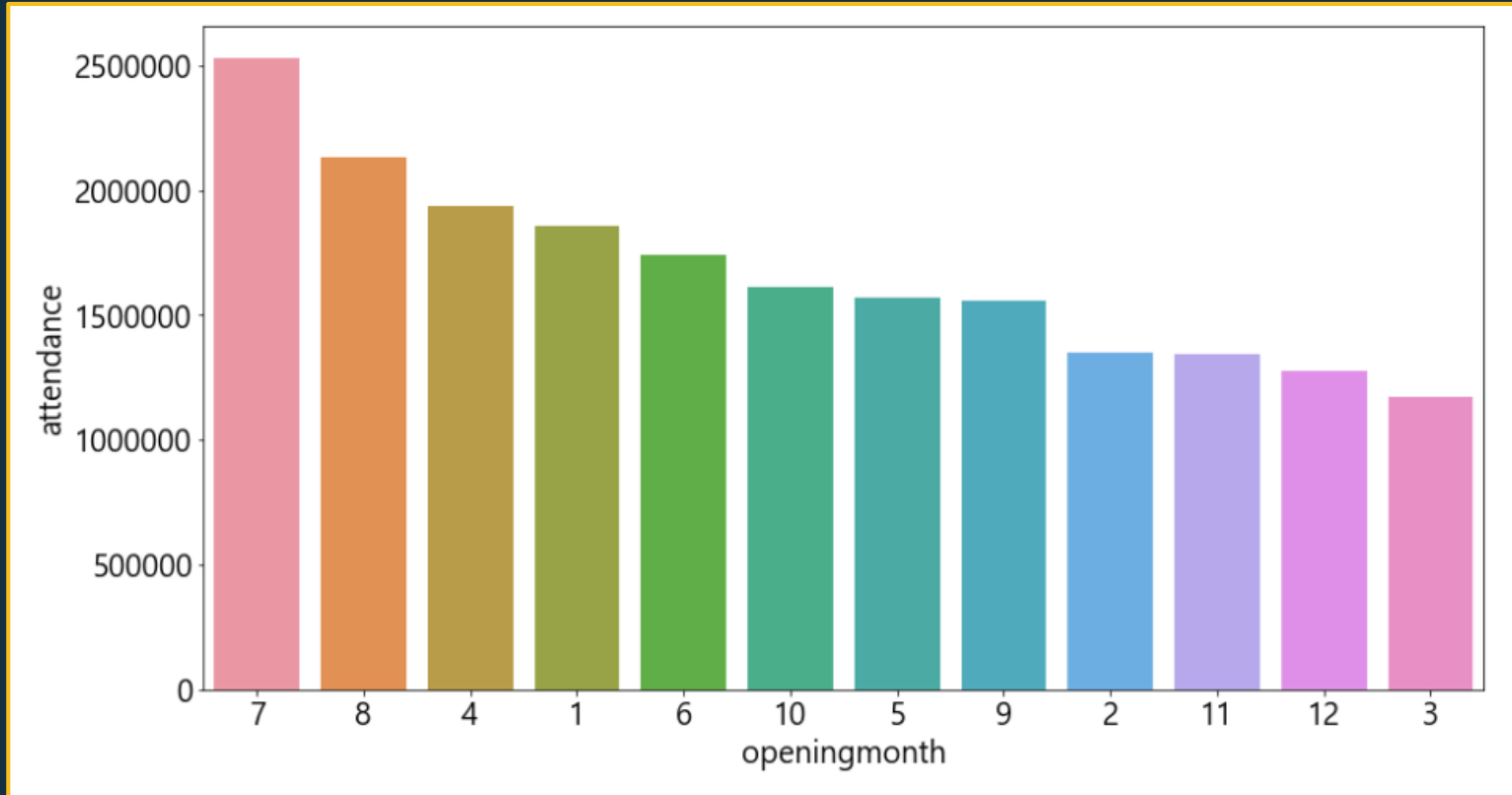
기초통계량 및 Null값 확인

→ Null값 제거



Data Preprocessing & EDA

CONTENT
3



월별 관람객 수



Data Preprocessing & EDA

CONTENT

3



장르별 관람객 수



Data Preprocessing & EDA

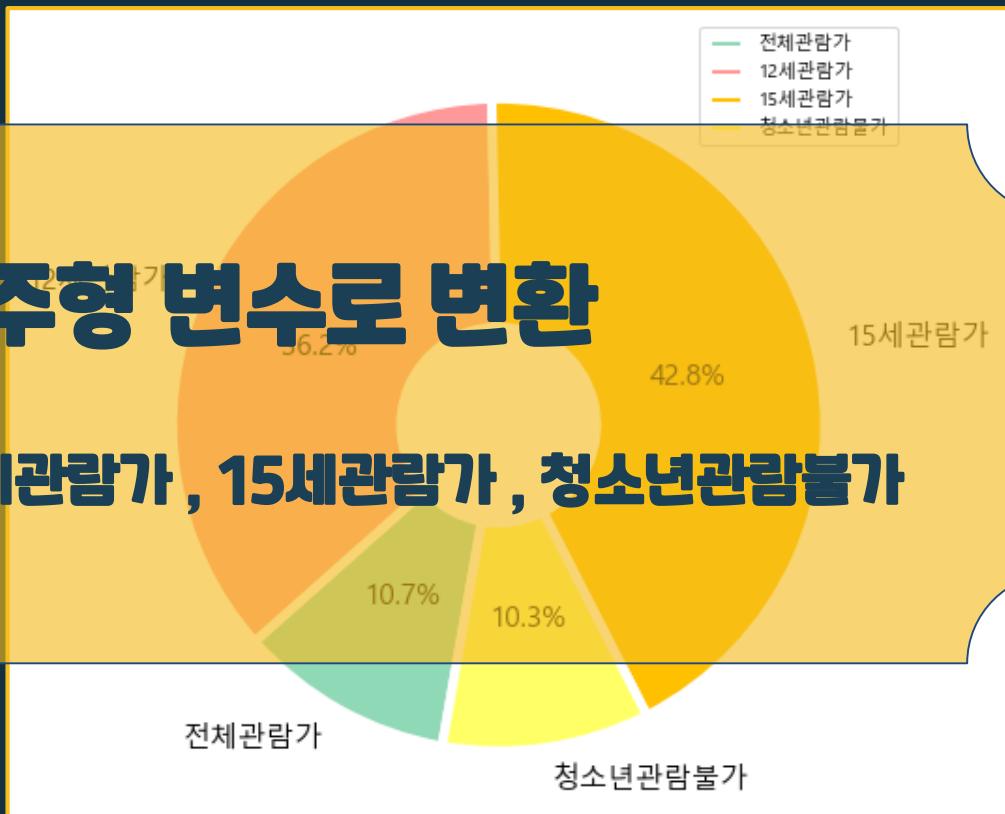
CONTENT

3

	attendance
filmrank	
[국내] 12세 관람가	451008340.0
[국내] 12세 관람가 [해외] NR	3321443.0
[국내] 12세 관람가 [해외] PG	3891654.0
[국내] 12세 관람가 [해외] PG-13	162266971.0
[국내] 15세 관람가	62414628.0
[국내] 15세 관람가 [해외] NR	4610452.0
[국내] 15세 관람가 [해외] PG-13	82567338.0
[국내] 15세 관람가 [해외] R	22470057.0
[국내] 전체 관람가	81834645.0
[국내] 전체 관람가 [해외] G	6925957.0
[국내] 전체 관람가 [해외] PG	80275934.0
[국내] 전체 관람가 [해외] PG-13	6797258.0
[국내] 청소년 관람불가	158968778.0
[국내] 청소년 관람불가 [해외] PG-13	2308596.0
[국내] 청소년 관람불가 [해외] R	15030757.0
[해외] PG	306012.0

범주형 변수로 변환

전체관람가, 12세관람가, 15세관람가, 청소년관람불가



관람등급별 관람객 수



Data Preprocessing & EDA

CONTENT

3

범주형 변수로 변환

한국 → domestic,
국외 → abroad

미국

45.7%

45.3%

3.2%

미국 일본 중국 영국 프랑스



제작국가 전처리



Data Preprocessing & EDA

CONTENT

3

attendance

distributor

(주)씨제이엔엠 364859383

(주)쇼박스 202214255

국내 메이저 영화사 : CJ E & M, 롯데, 쇼박스, 넥스트엔터테인먼트월드 (NEW)

(주)넥스트엔터테인먼트월드(NEW) 159148717

워너브러더스 코리아(주) 128319211

해외 메이저 영화사 : 월트디즈니, 유니버설픽처스, 워너스브러더스

유니버설픽처스인터내셔널 코리아(유) 92935081

월트디즈니컴퍼니코리아 유한책임회사 78981887

롯데컬처웍스(주)롯데엔터테인먼트 48092885

소니픽처스릴리징월트디즈니스튜디오스코리아(주) 45874563

범주형 변수로 변환

→ major_domestic

→ major_abroad

그외 배급사 → etc_distributor

major_domestic
etc
major_abroad

etc

23.4%

major_abroad

배급사별 관람객 수



Data Preprocessing & EDA

CONTENT

3

유야마 쿠니히코 6

우민호 6

시즈노 코분 6

폴 앤더슨 6

크리스토퍼 놀란 6

수치로 변환

..

윌트 도른 1

샘 테일러-존슨 **이전 참여 영화 (총 관객 수 / 영화 수)**

우선호 1

최호 1

강일수 1

Name: director, Length: 692, dtype: int64

감독 변수



Data Preprocessing & EDA

CONTENT

3

actor	수치로 변환		actor
다니엘 래드클리프 (해리 포터), 엠마 왓슨(헤르미온느), 루퍼트 그린트(론 위즐리)	→		3
빌리 웅거(새미), 팻 캐롤, 카를로스 맥 클러스 2세(레이)			3
임수정(연정인), 이 선균(이두현), 류승 룡(장성기)			3
이선균(정배), 최강 화(다림)			2
위하준(하준), 박지 현(지현), 오아연(아 연)	주연 배우 수		3

배우 변수



Data Preprocessing & EDA

CONTENT

3

Dummy Variable 변환

범주형 변수를 Dummy 변수로 변환

country_abroad ...	genre_noir	genre_romance	genre_sf	genre_thriller	genre_war	filmrank_G_rated	filmrank_PG 12	filmrank_PG 15	filmrank_X_rated	filmra
0 ...	0	0	0	0	0	1	0	0	0	
0 ...	0	0	0	0	0	1	0	0	0	
1 ...	0	1	0	0	0	0	0	1	0	
0 ...	0	0	0	0	0	0	0	0	1	
0 ...	0	0	0	0	0	0	0	1	0	
...	...	...	...	...	...	...	...	...	...	
1 ...	0	0	0	0	0	0	0	0	0	1
1 ...	0	0	1	0	0	0	1	0	0	
1 ...	0	0	0	0	0	1	0	0	0	
1 ...	0	0	0	0	0	0	0	1	0	
1 ...	0	0	0	0	0	0	1	0	0	

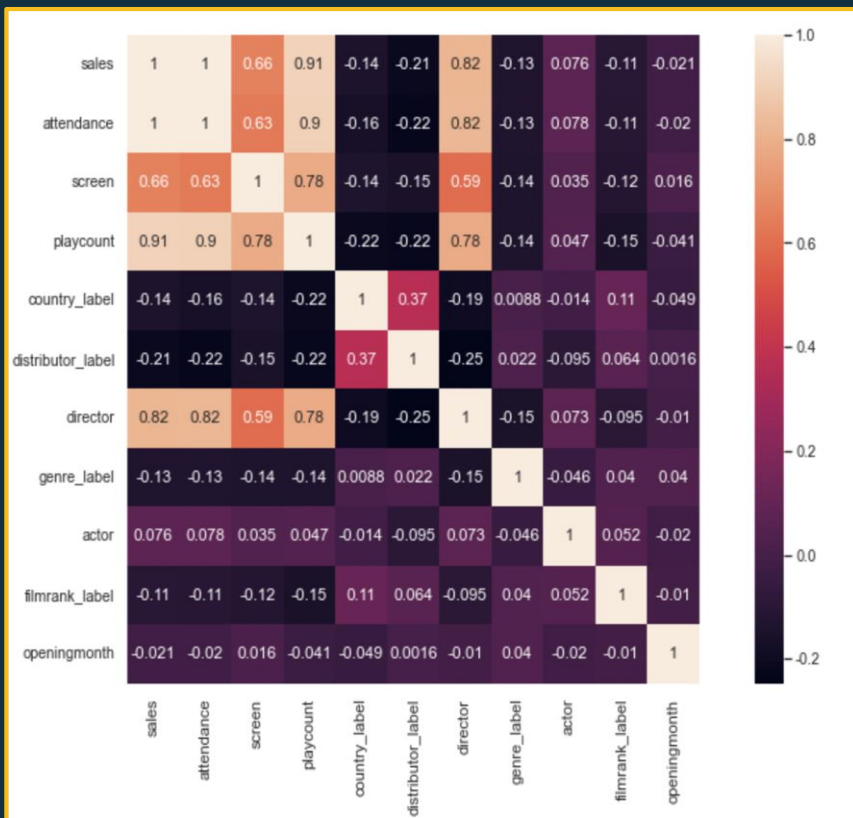
Dummy data



Modeling

CONTENT

4



높은 상관도를 보이는 요소
→ 관객수, 상영 스크린 수
매출 계산에 이미 사용된 요
소 임으로 예측/분석 모델에
사용하기 적합하지 않음

Heat Map



Modeling

CONTENT

4

모든 요소에 대한
상관성과 영향력을
자체적으로 학습하도록 구
성

```
col = ['sales', 'attendance', 'screen', 'playcount', 'country',  
       'distributor', 'director', 'genre', 'actor', 'openingmonth']  
one_hot_col = ['country', 'distributor', 'genre']  
filmrank_one_hot = []  
for filmrank in movie_filmranks:  
    filmrank = filmrank[col]  
    one_hot = pd.get_dummies(filmrank, columns=one_hot_col)  
    filmrank_one_hot.append(one_hot)  
  
for rank in filmrank_one_hot:  
    print(rank.shape)  
  
(186, 19)  
(398, 25)  
(120, 22)  
(317, 25)  
  
x = []  
y = []  
for rank in filmrank_one_hot:  
    x.append(rank.iloc[:, 1:])  
    y.append(rank.iloc[:, 0].to_numpy())  
  
for i in range(len(x)):  
    print(x[i].head())  
    print(y[i][:5])
```



Modeling

CONTENT 4

Model: "sequential_2"

Layer (type)	Output Shape	Param #
dense_10 (Dense)	(None, 256)	5632
dense_11 (Dense)	(None, 256)	65792
dense_12 (Dense)	(None, 256)	65792
dense_13 (Dense)	(None, 256)	
dense_14 (Dense)	(None, 1)	

=====
Total params: 203,265
Trainable params: 203,265
Non-trainable params: 0

18

Model: "sequential_3"

Layer (type)	Output Shape	Param #
dense_15 (Dense)	(None, 256)	6400
dense_16 (Dense)	(None, 256)	65792
dense_17 (Dense)	(None, 256)	65792
dense_18 (Dense)	(None, 256)	
dense_19 (Dense)	(None, 1)	

=====
Total params: 204,033
Trainable params: 204,033
Non-trainable params: 0

12

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 256)	4864
dense_1 (Dense)	(None, 256)	65792
dense_2 (Dense)	(None, 256)	65792
dense_3 (Dense)	(None, 256)	
dense_4 (Dense)	(None, 1)	

=====
Total params: 202,497
Trainable params: 202,497
Non-trainable params: 0

ALL

Model: "sequential_1"

Layer (type)	Output Shape	Param #
dense_5 (Dense)	(None, 256)	6400
dense_6 (Dense)	(None, 256)	65792
dense_7 (Dense)	(None, 256)	65792
dense_8 (Dense)	(None, 256)	
dense_9 (Dense)	(None, 1)	

=====
Total params: 204,033
Trainable params: 204,033
Non-trainable params: 0

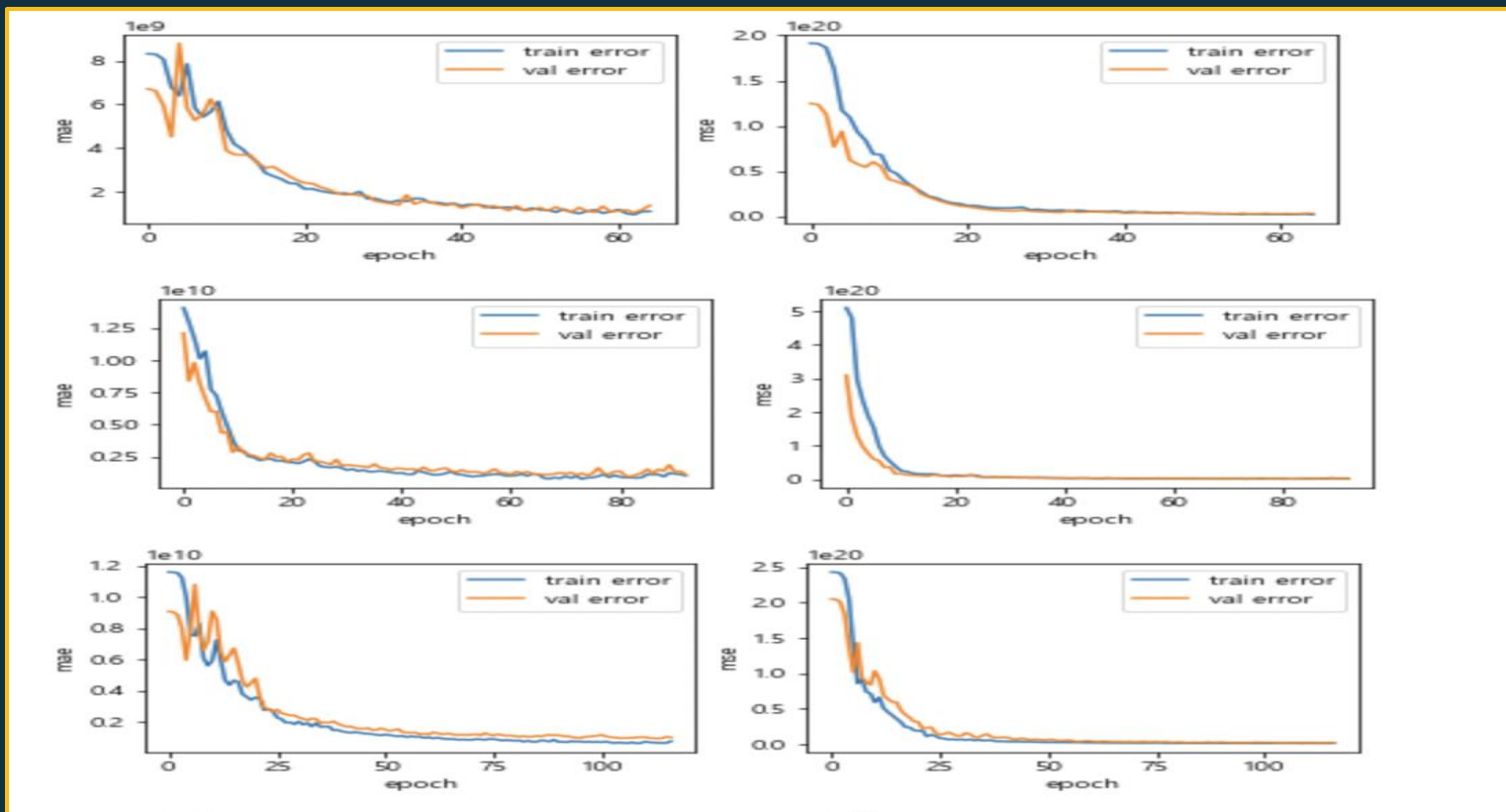
15

Film_rating 별 모델 구축



Modeling

CONTENT 4



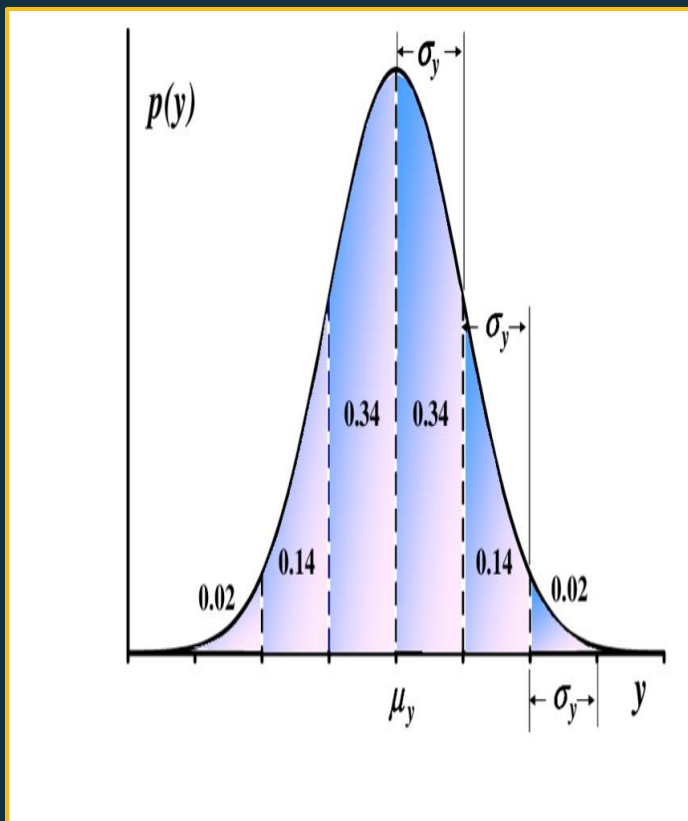
Epoch에 따른 오류 조정



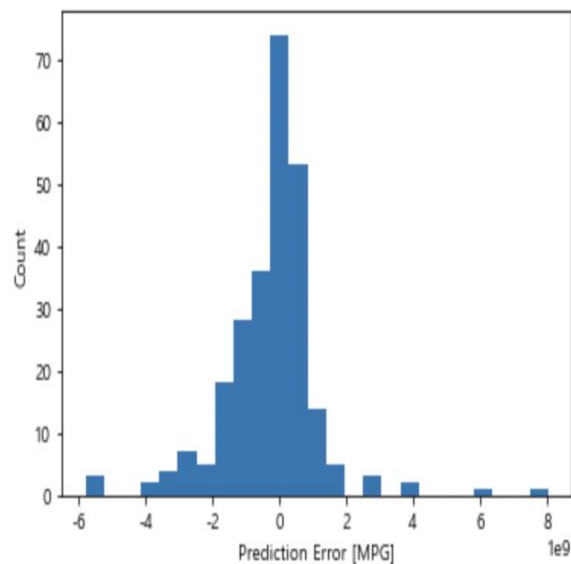
Modeling

CONTENT

4



```
error = x_pred - y_test
plt.hist(error, bins = 25)
plt.xlabel("Prediction Error [MPG]")
_ = plt.ylabel("Count")
```



오차 분포 = 정규 분포 -> 모델과 적용 데이터 신뢰성 확인



Modeling

CONTENT

4

```
def build_model():
    model = keras.Sequential([
        layers.Dense(256, activation='relu', input_shape=(1,)),
        layers.Dense(256, activation='relu'),
        layers.Dense(1)
    ])

    optimizer = tf.keras.optimizers.Adam(0.1)

    model.compile(loss='mse',
                  optimizer=optimizer,
                  metrics=['mae', 'mse'])

    return model
```

```
sales_model = build_model()
sales_model.summary()
```

Model: "sequential_12"

Layer (type)	Output Shape	Param #
dense_40 (Dense)	(None, 256)	512
dense_41 (Dense)	(None, 256)	65792
dense_42 (Dense)	(None, 1)	257
Total params: 66,561		
Trainable params: 66,561		
Non-trainable params: 0		

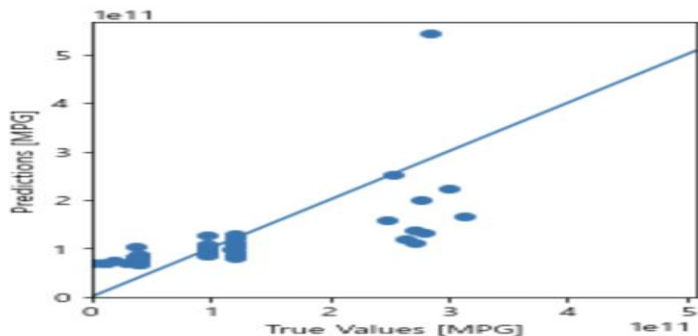
제작비별 매출액 예측 모델 구축



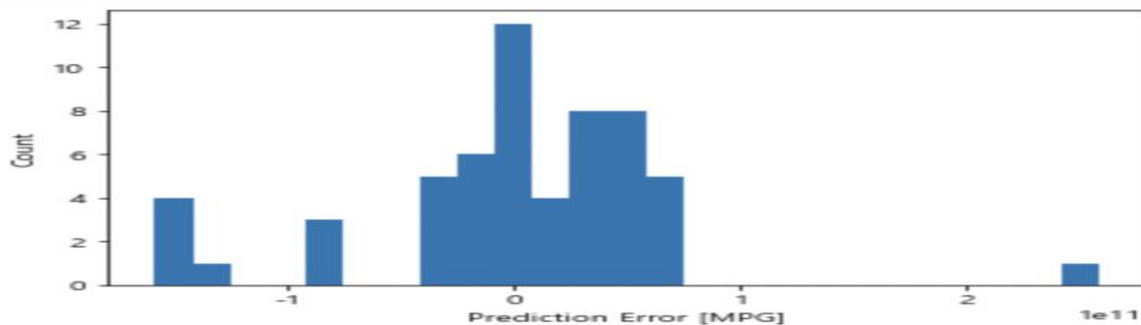
Modeling

CONTENT

4



```
= error = sales_x_pred - sales_y_test  
plt.hist(error, bins = 25)  
plt.xlabel("Prediction Error [MPG]")  
_ = plt.ylabel("Count")
```



오차 분포 상이 → 모델과 적용 데이터 한계 발견



Result

CONTENT

5

	loss	mae	mse	val_loss	val_mae	val_mse	epoch
49	4.415050e+18	1.420023e+09	4.415050e+18	3.541416e+18	1.219342e+09	3.541416e+18	49
50	4.478963e+18	1.278365e+09	4.478963e+18	3.321833e+18	1.181504e+09	3.321833e+18	50
51	4.113902e+18	1.244218e+09	4.113902e+18	4.638447e+18	1.435575e+09	4.638447e+18	51
52	4.365898e+18	1.332125e+09	4.365898e+18	3.440880e+18	1.209643e+09	3.440880e+18	52
53	3.632890e+18	1.212754e+09	3.632890e+18	3.314157e+18	1.203845e+09	3.314157e+18	53

```
In [35]: histes[0]['loss'].tolist()[-1]/(data[0][2].mean())**2
```

0.05685895263136701 → 5%

매출액 예측 모델의 오차



Result

CONTENT
5

$$MSE = \frac{1}{n} \sum \left(y - \hat{y} \right)^2$$

The square of the difference
between actual and
predicted

$\sqrt{(0.0568589)}$
0.2384510432

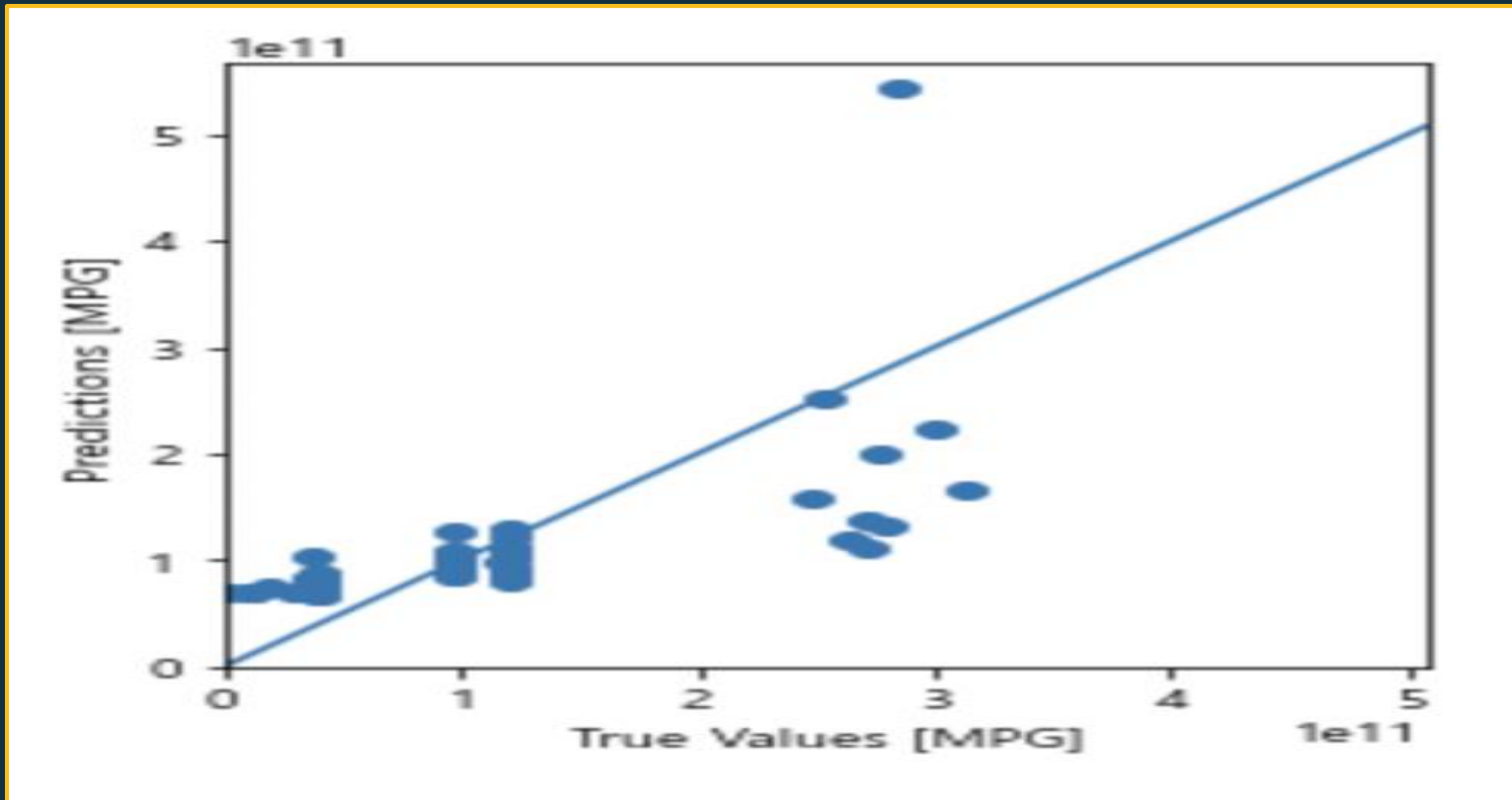
sin	cos	tan	☒ Deg ☐ Rad	
$\sin^{-1}$	$\cos^{-1}$	$\tan^{-1}$	π	e
x^y	x^3	x^2	e^x	10^x
$y\sqrt{x}$	$3\sqrt{x}$	$\sqrt{x}$	ln	log
(	)	1/x	%	n!
7	8	9	+	Back
4	5	6	-	Ans
1	2	3	$\times$	M+
0	.	EXP	$\div$	M-
$\pm$	RND	AC	=	MR

23.85%의 매출 예측 오류



Result

CONTENT
5

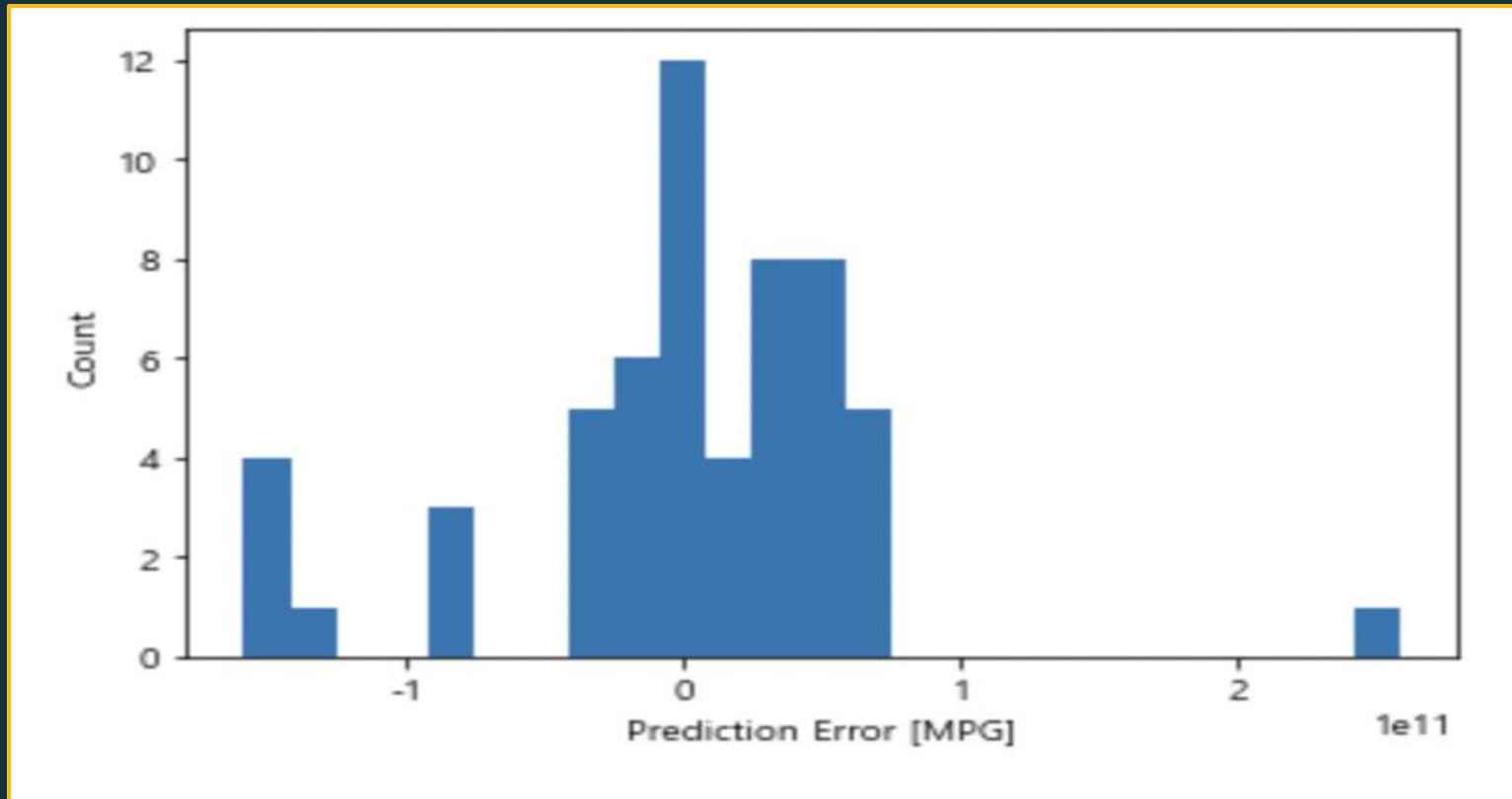


제작비별 매출액 예측 모델 오차



Result

CONTENT
5



제작비별 매출액 예측 모델 오차분포



Evaluate

CONTENT
6

1. 데이터의 다양성
2. 정성적 데이터 분석의 부재
3. 모델의 불확정성
4. 변인의 영향해석
5. 새로운 모델이 필요
6. 커스터마이징과 UI/UX 적용 부재



***Thank you
For Listening!***