상황 설정

- 내년 상반기 **해외 진출**을 계획 하고 있는 영화 제작사 '에이스 메이커 무비 웍스'

- 흥행이 보장된 작품을 제작하고 싶어 수익 예측 모델을 의뢰해 옴.

- 지난 2022년도에 제작한 작품:
  영화 3편, 시리즈물 1편(12부작)

- 이번 2023년도 하반기에 작품 선정 후 2024년도에 **최대 2편 제작**을 계획 중

Sai

Ace Maker

Movie Works

sai_

목차

Sai_

문제 정의 및 가설 수립

어떤 작품이 흥행할까?

어떻게 제작해야 수익을 극대화할 수 있을까?

목표설정 및 데이터 준비:

- 가설1: 장르에 따른 선호도가 존재 할 것이다
- 가설2: 상영시간이 적당해야 수익이 클 것이다
- 가설3: 예산이 크면 수익도 클 것이다
- 가설4: 봄, 가을에 수익이 클 것이다

목표설정 및 데이터 준비:

데이터셋 선정	"TMDB 5000 Movie Dataset"
THE MOVIE DB	TMDB(The Movie Database)에서 수집한 약 5000개의 영화 정보가 포함된 데이터셋 영화 제목, 개봉일, 배우, 예산, 수익 등

EDA 및 가설 확인

사용한 컬럼	내용
budget	영화의 예산
genres	영화의 장르
release_date	영화의 개봉일
revenue	영화의 수익
runtime	영화의 상영시간
title	영화의 제목
cast	영화의 출연진

목표설정 및 데이터 준비:

컬럼	예시
budget	245000000(int)
genres	[{"id": 28, "name": "Action"}, {"id": 12, "nam...
release_date	2009-12-10(str)
revenue	2787965087
runtime	162.000(float)
title	Avatar
cast	[{"cast_id": 242, "character": "Jake Sully", "...

EDA 및 가설 확인: 결측치 처리

목표설정 및 데이터 준비:

컬럼	결측치 형태
revenue, budget, runtime	0: 대체된 결측치
genres, cast	[]: 빈 리스트 형태의 결측지
runtime	2개

목표설정 및 데이터 준비:

EDA 및 가설 확인: 이상치 처리

컬럼	이상치 형태
수치형 변수	Million 단위로 기입된 수치
범주형 변수	딕셔너리 형태의 복잡한 Value

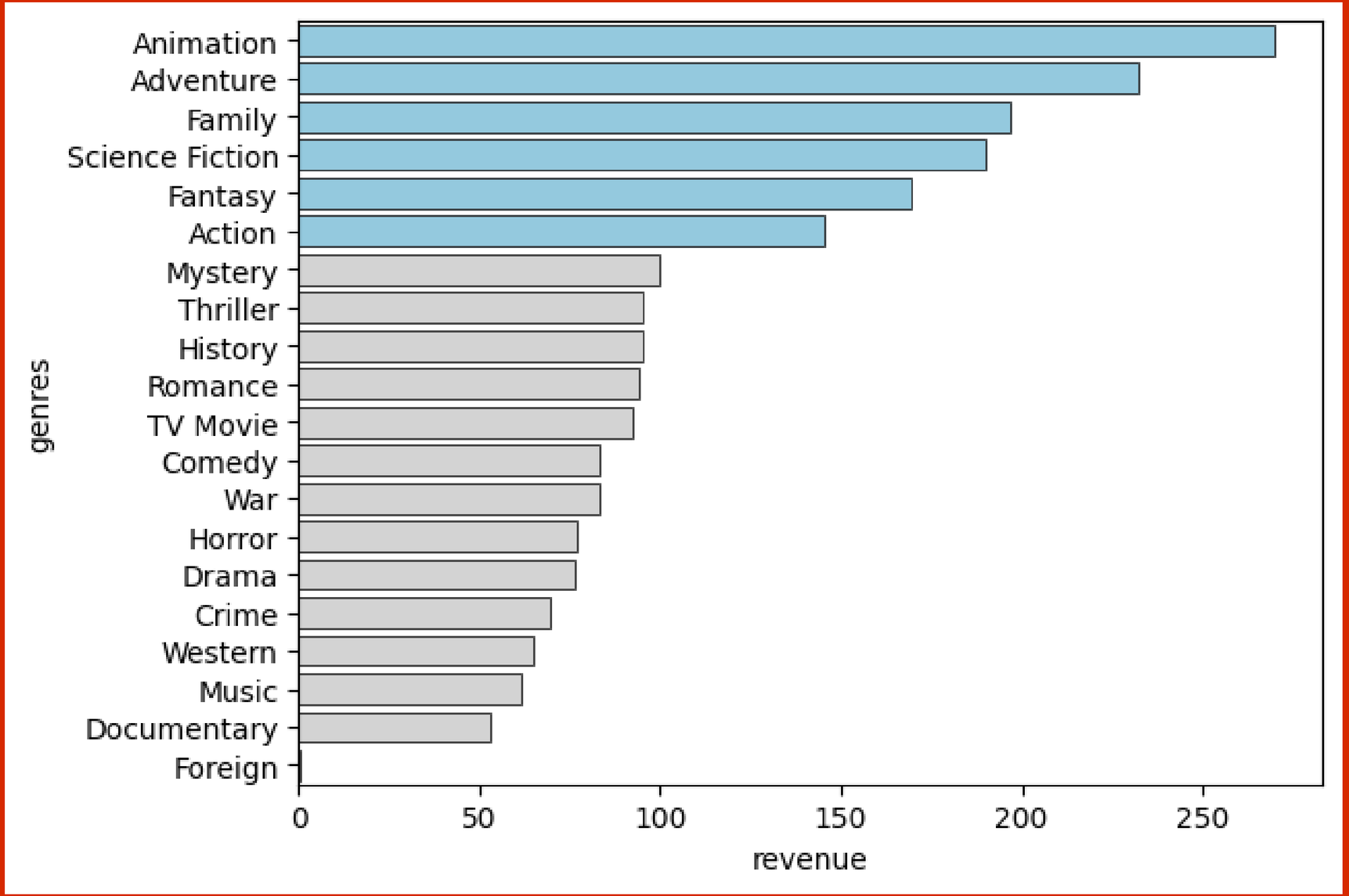
EDA 및 가설 확인: Feature Engineering

목표설정 및 데이터 준비:

컬럼	결측치 형태
title	제목의 단어 수(split(' '))
cast	출연자의 출연작 개수에 따라 등급 부여 1-5개:0, 6-10개:1, 11-20개:2, 20-개:3
month	release_date 컬럼에서 월 정보만 추출 2019-09-17 : 9
genre_rank	평균 수익을 기준으로 수익이 제일 낮은 장르부터 0-18번 등급 부여

EDA 및 가설 확인: 가설 확인

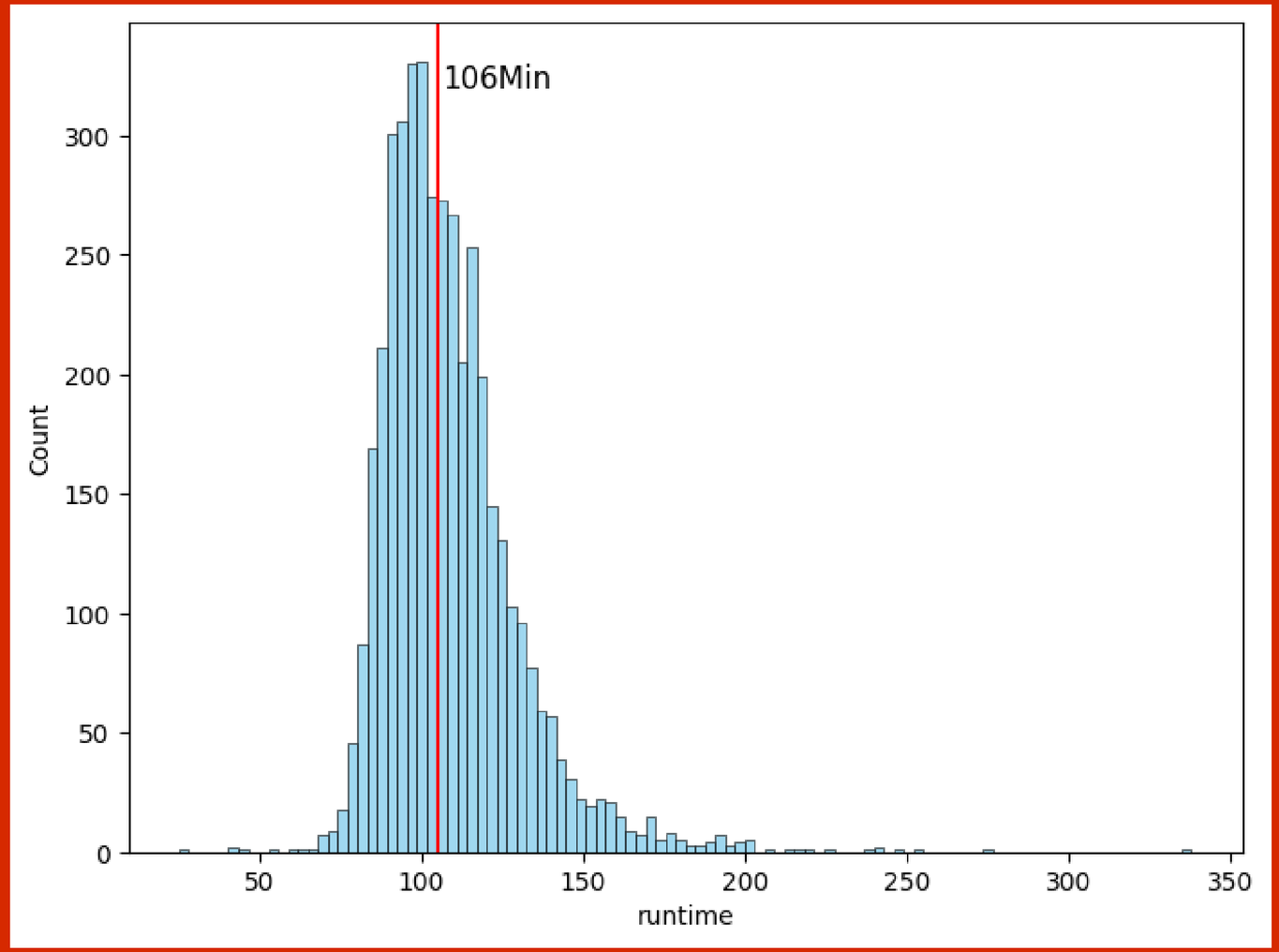
- 가설1: 장르에 따른 선호도가 존재 할 것이다



EDA 및 가설 확인: 가설 확인

O

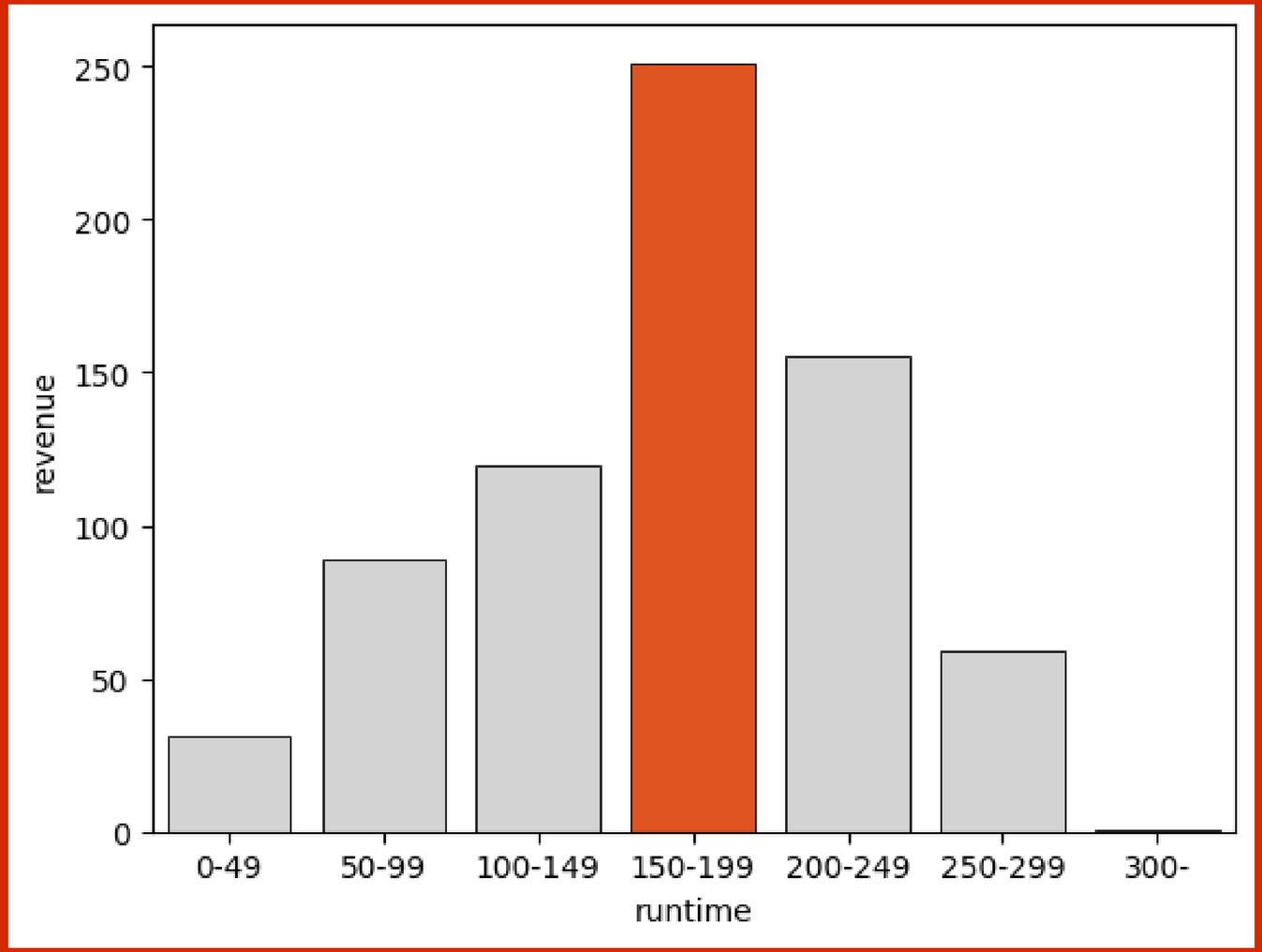
가설2: 상영시간이 적당해야 수익이 클 것이다



EDA 및 가설 확인: 가설 확인

O

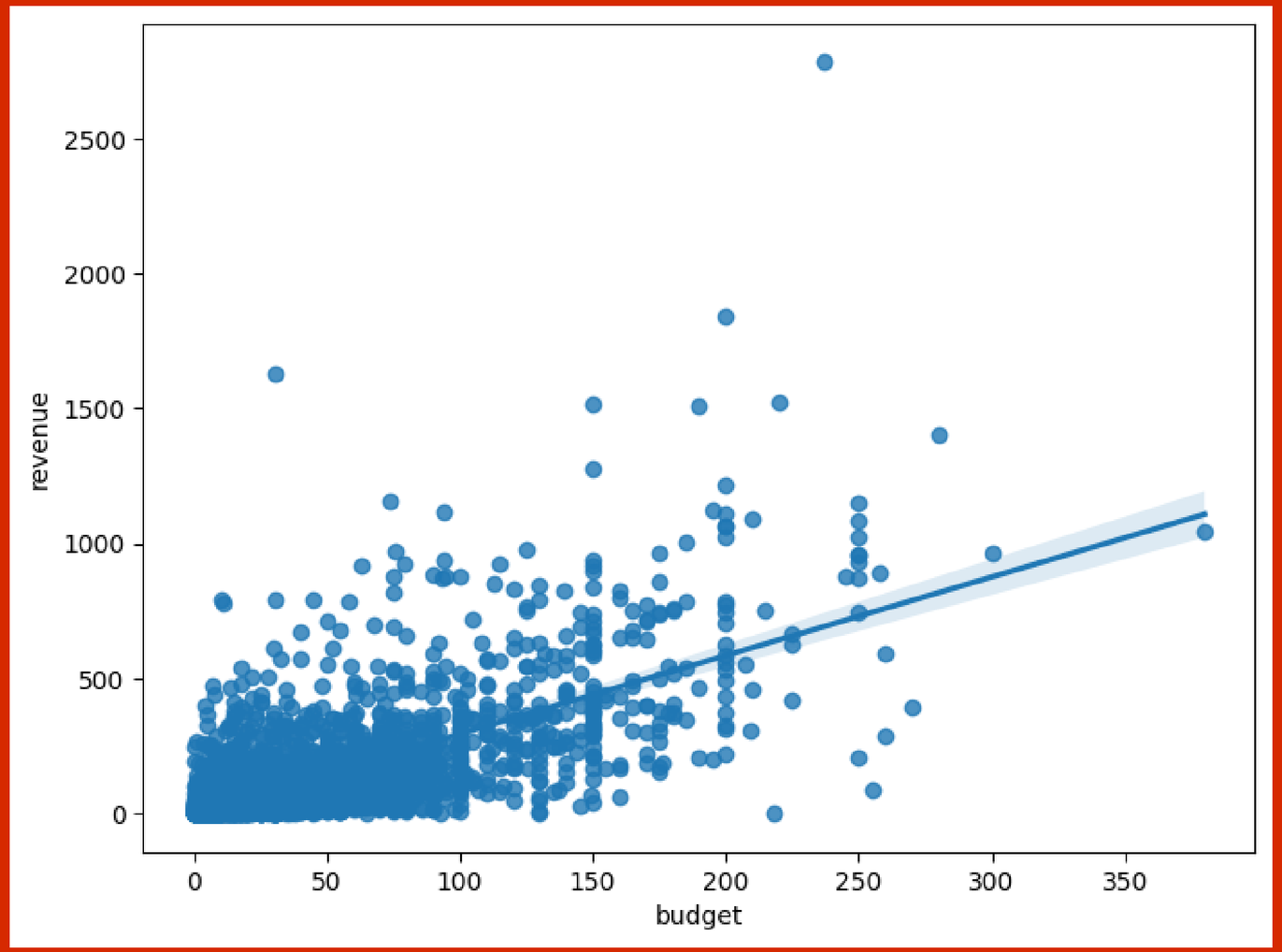
가설2: 상영시간이 적당해야 수익이 클 것이다



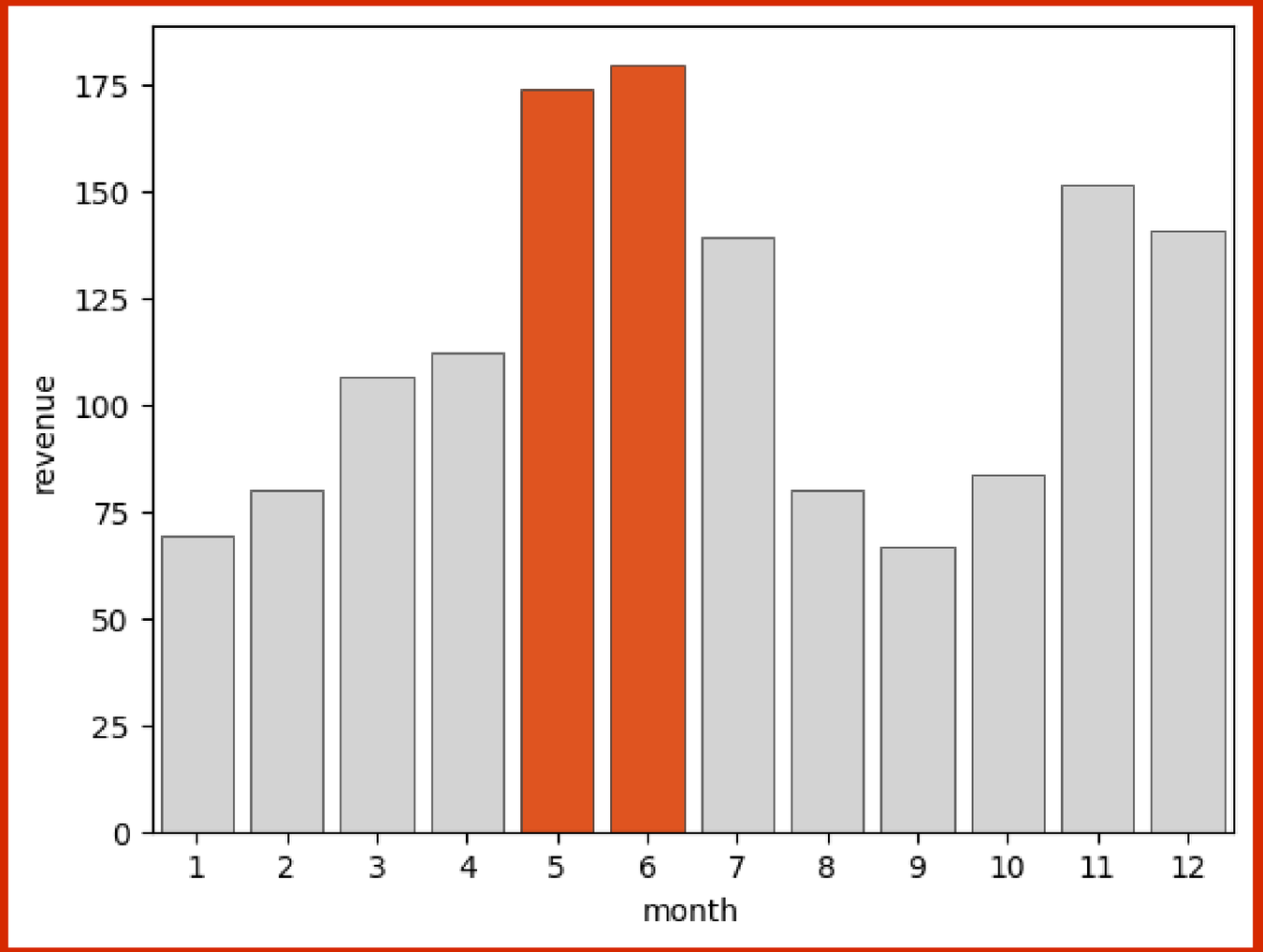
EDA 및 가설 확인: 가설 확인

O

가설3: 예산이 크면 수익도 클 것이다



EDA 및 가설 확인: 가설 확인



△

- 가설4: 봄, 가을에 수익이 클 것이다

기준모델: 단순 선형 회귀 모델

- 타겟과 변수간의 스피어만 상관계수

	budget	runtime	title	cast	month	genre_rank
revenue	0.693	0.212	0.093	0.213	0.035	0.307

- 평가지표: MAE, R2

(0 .xxx)이하 반올림	학습 데이터	검증 데이터
MAE	59.073	62.708
R2	0.456	0.507

모델 선택 및 튜닝: 랜덤 포레스트 회귀 모델
(RandomForest Regressor)

(0 .xxx)이하 반올림	다중 선형 회귀	다항 선형 회귀	Lasso 회귀	랜덤 포레스트 회귀	xgboost 회귀
MAE (학습/검증)	59.018 / 62.649	60.277 / 65.633	61.981 / 65.848	50.823 / 60.926	51.754 / 66.790
R2 (학습/검증)	0.456 / 0.508	0.508 / 0.553	0.466 / 0.515	0.501 / 0.494	0.665 / 0.525

- 기준모델보다 낮은 MAE: **(50.823 / 60.926)** : (59.073 / 62.708)
- 학습 / 검증의 차이가 작은 R2: **(0.501 / 0.494)** : (0.456 / 0.507)

모델 선택 및 튜닝: 하이퍼 파라미터 튜닝(scoring=MAE)
GridSearchCV

- max_depth: 트리의 최대 깊이
- min_samples_leaf: 말단 노드(더이상 확장하지 노드)가 되기 위한 최소 샘플 수
- min_samples_split: 노드를 분할하기 위한 최소 샘플 수
- n_estimators: 기본 모델의 수 (weak learner의 수)

모델 선택 및 튜닝: 하이퍼 파라미터 튜닝(scoring=MAE)

GridSearchCV

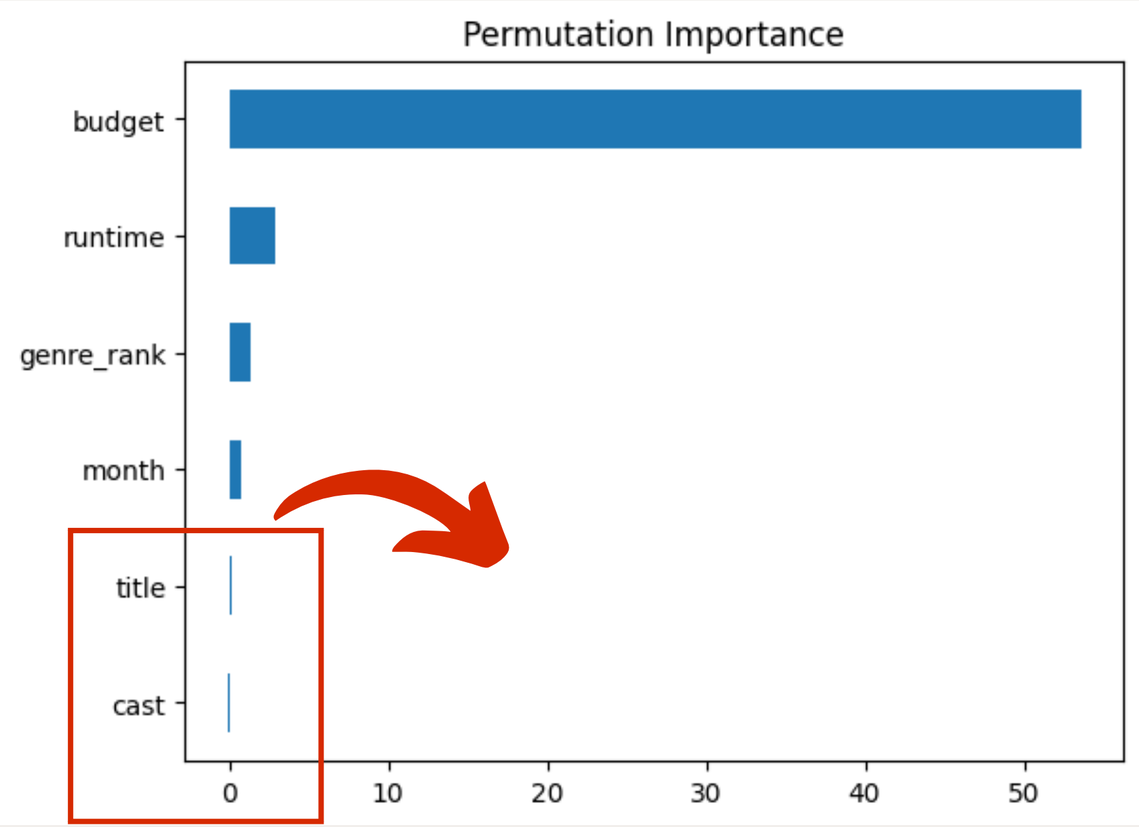
	최적 파라미터
max_depth	9
min_samples_leaf	9
min_samples_split	2
n_estimators	200

- max_depth: 트리의 최대 깊이
- min_samples_leaf: 말단 노드(더이상 확장하지 노드)가 되기 위한 최소 샘플 수
- min_samples_split: 노드를 분할하기 위한 최소 샘플 수
- n_estimators: 기본 모델의 수 (weak learner의 수)

	☒튜닝 전	튜닝 후
MAE	50.823 / 60.926	54.135 / 60.765
R2	0.501 / 0.494	0.503 / 0.493

모델 선택 및 튜닝: 순열중요도 확인

Permutation Importances



모델 선택 및 튜닝: 컬럼 선택 후 다시 튜닝

GridSearchCV

	최적 파라미터
max_depth	13
min_samples_leaf	8
min_samples_split	2
n_estimators	200

- max_depth: 트리의 최대 깊이
- min_samples_leaf: 말단 노드(더이상 확장하지 노드)가 되기 위한 최소 샘플 수
- min_samples_split: 노드를 분할하기 위한 최소 샘플 수
- n_estimators: 기본 모델의 수 (weak learner의 수)

	기준 모델	최종 모델
MAE	59.073 / 62.708	54.127 / 59.658
R2	0.456 / 0.507	0.550 / 0.522

일반화 성능 확인

(0 .xxx) 이하 반올림	학습 데이터	평가 데이터
MAE	54.127	54.681
R2	0.550	0.505

수익 예측

작품 리스트

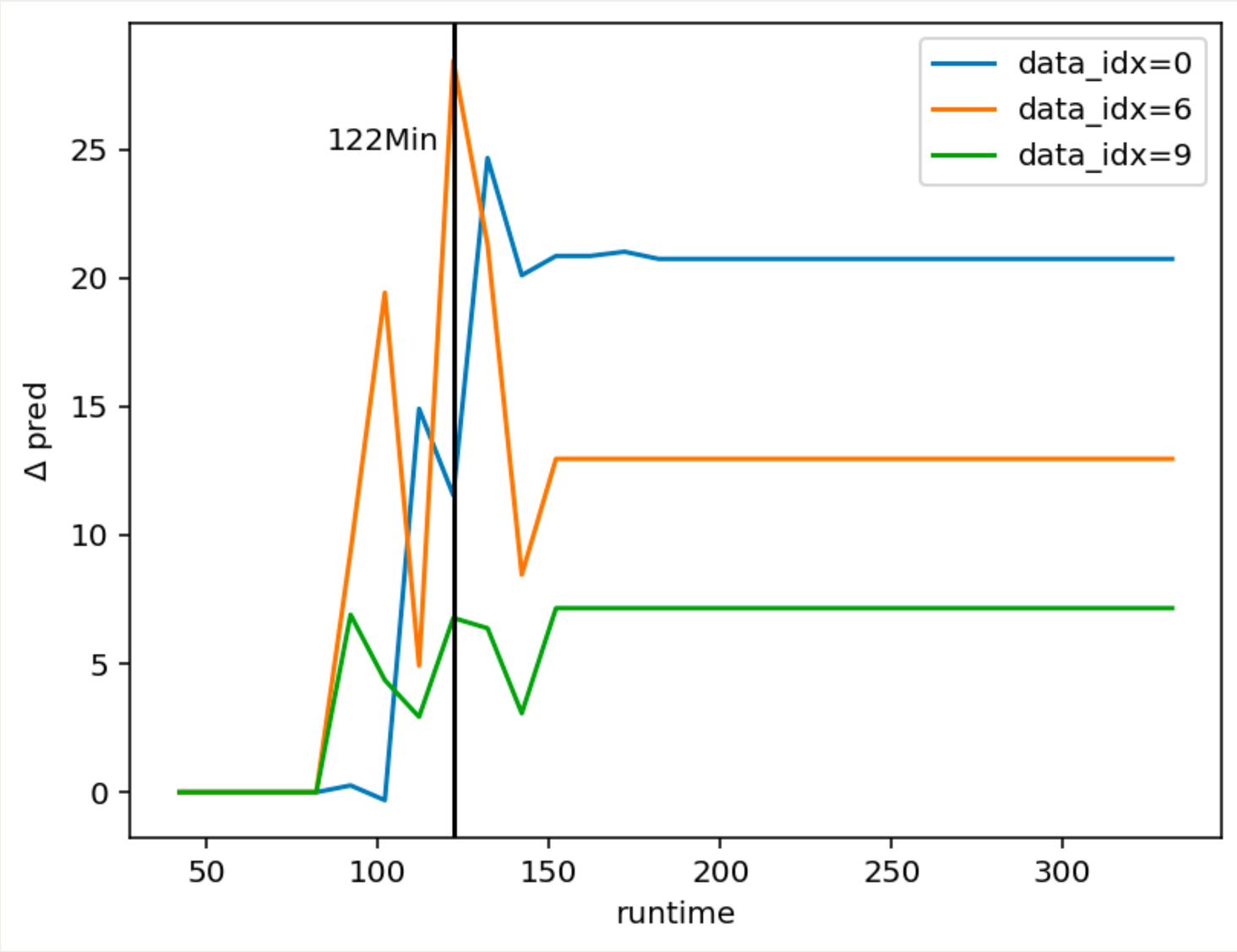
실전 대입

작품 번호	예산(M)	☒예측한 수익 (M)
0	55.000	132.816 ✓
1	9.000	20.583
2	8.000	23.038
3	0.250	3.339
4	1.900	12.222
5	0.600	15.739
6	15.600	45.170 ✓
7	20.000	33.564
8	1.596	18.218
9	26.000	62.550 ✓

수익 극대화를 위한 방향 제시

실전 대입

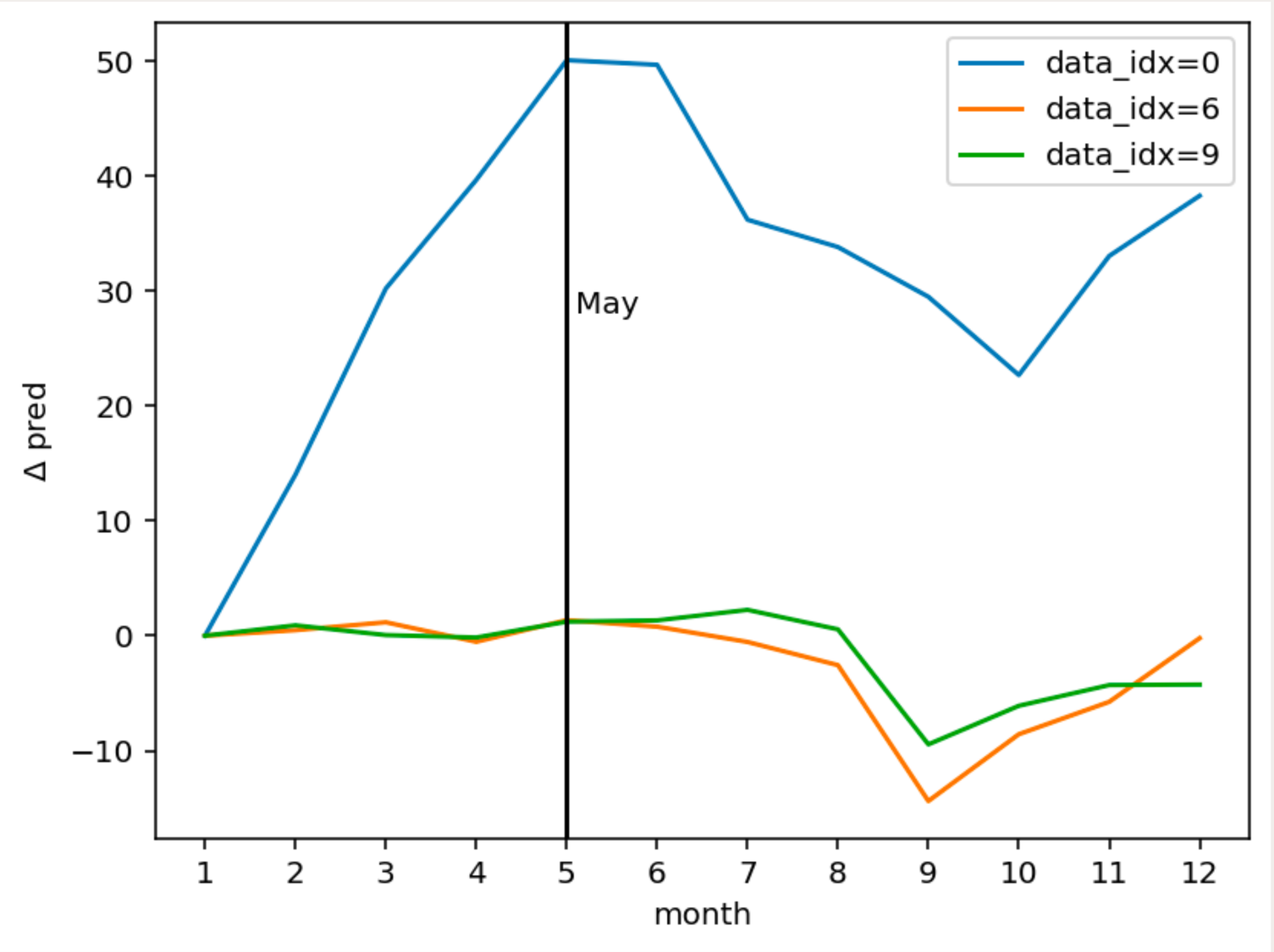
- 0,6 번 작품의 영상 길이 조정



수익 극대화를 위한 방향 제시

실전 대입

- 0,6 번 작품의 영상 길이 조정
- 0번 작품의 개봉일 조정



Ace Maker

Movie Works

sai_