데이터 분석 개요

이 한 준 교수

데이터 분석

- 데이터의 정보를 인식가능한 수준으로 요약하는 과정
- 데이터 속에서 가치 있는 인사이트를 찾는 과정

형태에 따른 데이터 분류

- **정형(structured) 데이터**

  일반적인 관측치-변수 형태의 데이터

  새로운 관측치가 들어올 때마다 미리 설정된 특성(변수)를 기록

  - [예시]
    - 카드사 카드매출 데이터
    - 건강보험공단 진료내역 데이터

- **비정형(unstructured) 데이터**

  텍스트, 이미지, 음성 등 정형 데이터의 형태로 표현하기 힘든 데이터

  데이터 정형화 작업을 거쳐 분석에 활용

  - [예시]
    - 소셜 미디어의 사진, 텍스트
    - 웹/앱 로그 데이터

관계형 데이터베이스의 활용

- 관계형 데이터베이스(RDB : Relational Database)
  정형데이터를 효율적으로 저장하는 체계
  키(key)와 값(value)들의 관계를 구조화하여 데이터를 구성
  SQL 클라이언트를 통해 데이터 결합/추출

- [예제] 카드사 데이터베이스

거래 데이터

거래ID	회원ID	가맹점ID	결제일시	결제금액

회원 데이터

회원ID	성별	연령	거주지역

가맹점 데이터

가맹점ID	업종코드	주소	수수료율

시점에 따른 데이터 분류

- 횡단(cross-sectional) 데이터

  고정된 특정 시점의 데이터

  일반적으로 관측치 간에는 관련이 없고, 관측치 간 비교를 활용

  - [예시]
    - 2018년 상품별 매출 데이터 : 어떤 상품이 많이 팔렸나?
    - 지난 한달 간 신규 유입고객 데이터 : 고객들이 어떤 특성을 가지고 있나?

- 시계열(time series) 데이터

  고정된 관측치에 대한 복수 시점의 데이터

  특정 시점의 값은 더 과거 시점의 값에 영향을 받음

  - [예시]
    - 2018년 삼성전자 주가 : 내일 주가는 어떻게 될까?

주가 데이터의 시점에 따른 활용

- **횡단 데이터**

  시점을 고정시킨 후 여러 종목의 등락을 비교

- **시계열 데이터**

  관심 종목을 정하고 여러 시점에 대한 정보를 활용

	1일	2일	3일	4일	5일
삼성전자	2,551,000	2,581,000	2,554,000	2,606,000	2,601,000
SK텔레콤	265,500	262,000	259,500	267,500	266,000
셀트리온	225,900	249,700	250,000	266,900	302,500
카카오	146,500	149,000	156,000	156,000	159,500

셀트리온 주가 예측

다른 종목 등락률 예측

시계열 데이터 분석의 어려움

- 시계열 모형의 장단점
  관심 대상의 과거의 패턴은 높은 수준으로 설명 가능
  미래 예측력의 문제 발생

데이터와 데이터 공간

데이터 공간

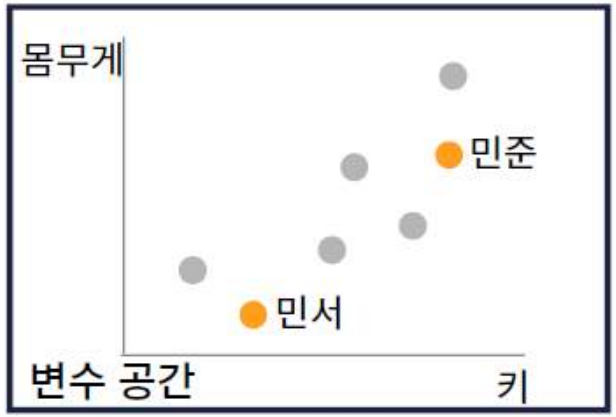
변수 개수만큼의 p차원이 형성되고 관측치는 n개의 점으로 표현됨

- [예시] 키, 몸무게 데이터
  - 2차원 공간이 형성되고 관측치의 위치는 점으로 표현 가능

	A	B	C
1	이름	키	몸무게
2	민서	170	60
3	민준	180	70

데이터

=



데이터 결합의 필요성과 다차원의 제약

- 새로운 관계로부터 새로운 인사이트 도출 가능
  - [예제] 매출 데이터와 날씨 데이터의 결합

주문일자	주문시간	주문상품	주문금액	연령대	날씨	기온
9월 27일	09:00	아메리카노	4,100	40대	비	6 ℃
⋮	⋮	⋮	⋮	⋮	⋮	⋮
9월 28일	15:07	ICED 아메리카노	4,100	30대	맑음	28 ℃
⋮	⋮	⋮	⋮	⋮	⋮	⋮

- 변수가 많아지면 분석의 난이도와 알고리즘의 복잡도가 증가
  - [예제] 차원의 저주(Curse of dimensionality)
    - 변수가 너무 많아 특정 조합에 관측치가 충분하지 않은 상황

데이터 분석 과정

■ 일반적인 데이터 분석 과정(연구/실험)

분석 목표 설정 → 데이터 수집/가공 → EDA/요약 → 모형/시각화

■ 현실적인 기업 데이터 분석 과정

데이터 추출/EDA → 분석 주제 설정 → 데이터 가공/요약 → 모형/시각화

데이터 분석과 숫자 그리고 스토리

- 데이터 분석을 통한 숫자(인사이트)의 확인
  상식적으로 납득하기 어려운 정보 도출 가능성이 높음
  - [예시] 가족 카드 사용, 가족 ID 사용 등

30대 남성 옥션베스트

하기스 하기스 에어솔솔썸머팬티 4단계 대형 여아용 48PX2팩 /

22% **38,500**원 ~~49,900원~~

무료배송

★★★★★ 상품평 17 · 구매 274

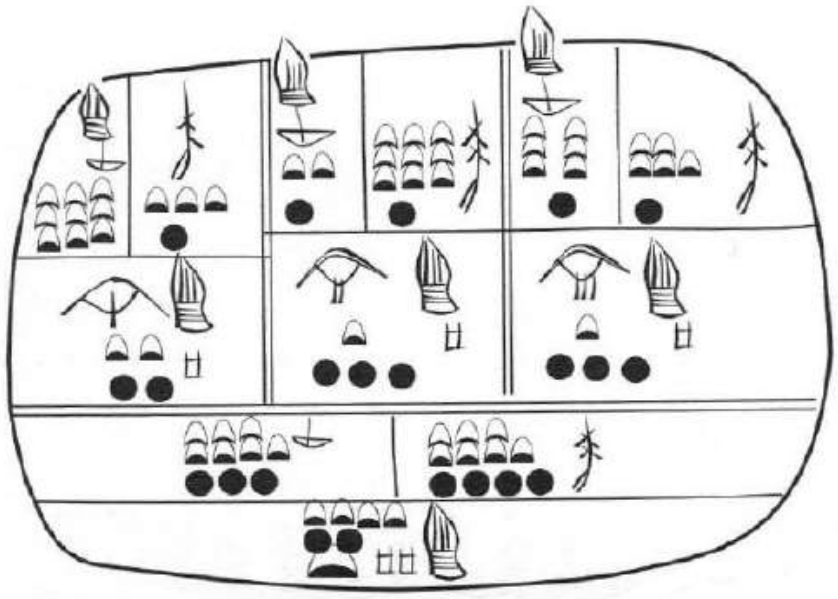
- 분석에 결과에 대한 이해와 해석
  업무에 대한 이해와 상상력이 필요

데이터 요약의 역사

- 요약은 데이터 분석의 기본

  더 많은 양을 효율적으로 빠르고 손쉽게 처리하는 것이 중요

  - [예제] 기원전 3,000년 전 수메르인의 점토판



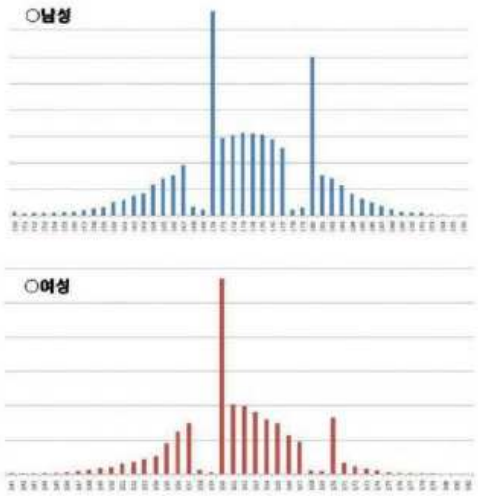
연도	3,000	2,999	2,998	합계
작물 A	9	12	16	37
작물 B	13	19	15	47
합계	22	31	31	84

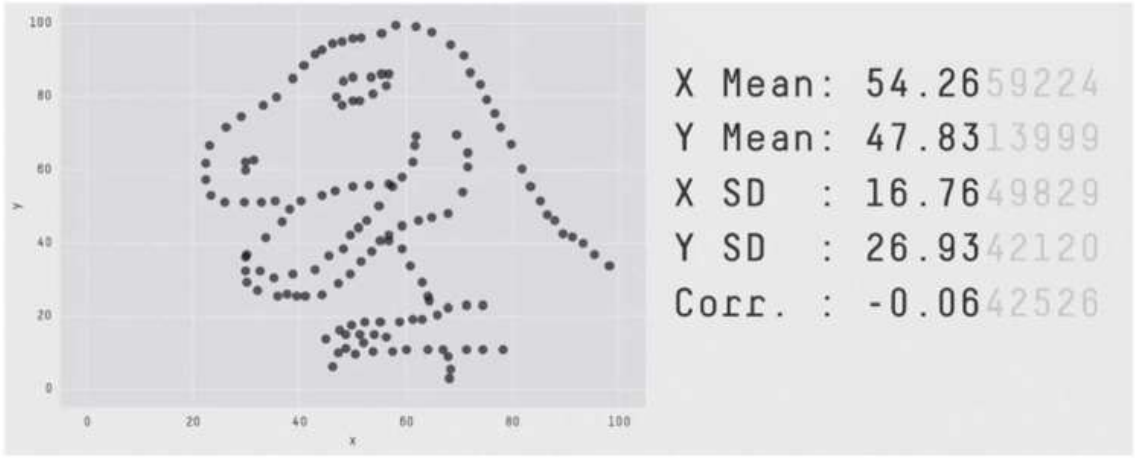
데이터 시각화의 필요성

- **탐색적 데이터 분석에 활용**
  변수의 특성과 변수 간의 관계를 그래프로 확인

- **분석 결과의 공유**
  숫자 대신 시각 요소를 활용한 그래프로 분석 결과를 효과적으로 전달

  - [예시]설문조사 키 분포(왼쪽)와 같은 숫자, 다른 분포(오른쪽)





시각화의 장점

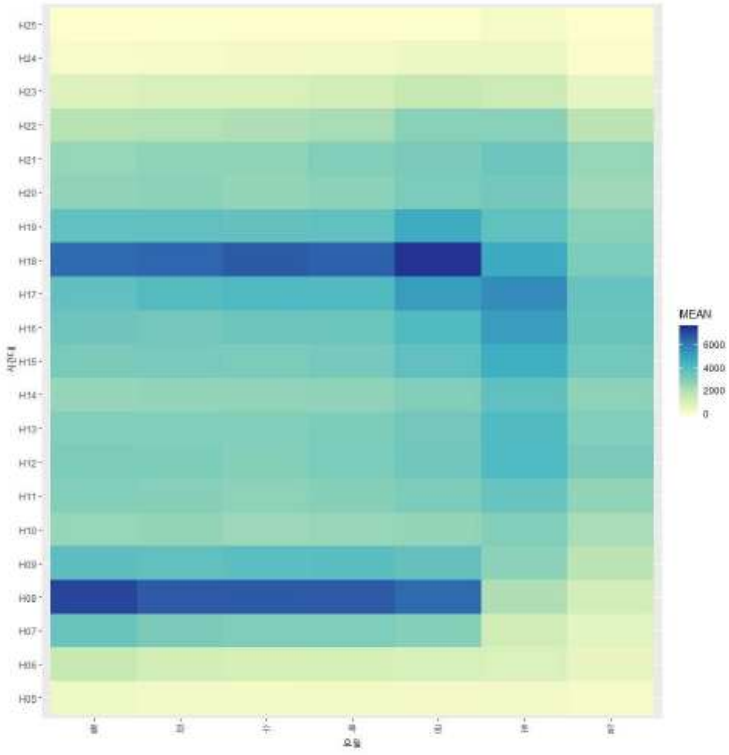
- 직관적인 인식 가능

  숫자에 비해 더 쉽게 정보를 확인하거나 더 많은 정보를 표현 가능

  - [예시] 시간대별 지하철 이용객수 요약 – 교차표(왼쪽)와 열지도(오른쪽)

	시간대	월	화	수	목	금	토	일
1	H05	423.375	368.875	343.375	346.6	336.5	367.6	250.250
2	H06	1455.375	1169.875	1136.375	1122.7	1056.1	965.1	604.000
3	H07	3626.500	3149.625	3012.375	3015.8	2860.3	1224.3	785.500
4	H08	7194.750	6792.750	6749.750	6787.8	6417.2	1976.9	1183.500
5	H09	3899.125	3830.875	3946.750	3979.6	3709.4	2742.3	1768.000
6	H10	2556.750	2622.125	2451.000	2506.4	2639.3	2937.6	2167.625
7	H11	2928.250	2828.125	2739.375	2876.0	3039.4	3623.8	2635.500
8	H12	3061.750	3039.000	2893.625	3091.8	3432.9	4257.2	3128.750
9	H13	2980.125	2973.000	2966.125	3043.4	3372.3	4201.7	2960.125
10	H14	2570.625	2607.125	2646.375	2698.7	2947.0	3857.1	2736.750
11	H15	3155.750	3194.500	3077.250	3279.1	3882.9	4670.9	3359.375
12	H16	3464.250	3346.625	3508.750	3551.4	4205.9	5223.0	3650.250
13	H17	3856.000	4093.250	4267.500	4202.6	5167.5	5626.5	3684.375
14	H18	6399.750	6505.750	6736.000	6607.2	7531.0	4801.1	3072.625
15	H19	3798.625	3848.000	3718.875	3862.0	4788.1	3818.8	2815.500
16	H20	2679.250	2744.625	2620.750	2744.4	3062.5	3310.6	2382.000
17	H21	2573.375	2700.250	2700.750	2949.5	3159.4	3504.4	2583.000
18	H22	1818.500	1912.000	2008.875	2172.1	2803.5	2800.2	1757.250
19	H23	910.125	1021.000	1037.250	1151.9	1487.0	1354.7	678.375
20	H24	220.500	261.000	313.750	338.7	472.4	459.3	112.000
21	H25	10.500	13.625	15.625	15.0	26.9	240.4	0.000

→



다양한 데이터 분석

- 데이터 요약(aggregation)
  숫자와 그래프를 활용하여 데이터의 특성을 확인
- 검정(test)
  관심있는 차이에 대한 통계적 유의성을 확인
- 모의실험(simulation)
  확률 분포 등을 활용하여 관심 대상에 대한 기댓값을 확률적으로 예측
  사건의 빈도와 심도에 대한 모의실험을 활용한 손해율 계산 등에 활용

다양한 데이터 분석(2)

- 데이터 기반 의사결정(data-driven decision)
  단순 요약이 아닌 데이터를 활용한 예측값 등을 토대로 의사결정을 내리는 방법
  규칙기반 시스템이나 기계학습 알고리즘을 주로 활용
  - [예시] 은행 대출, 카드 발급 심사

- 규칙기반 시스템(rule-based system)
  확인된 사실, 가설, 분석 결과를 바탕으로 조건이나 규칙을 설정
  금융 상품 가입 심사 등 정보가 제한적일 때 주로 활용

- 기계학습 알고리즘(machine learning algorithm)
  다양한 변수의 관계 속에서 의미 있는 정보와 패턴을 파악, 활용
  신용등급 등 활용가능한 정보가 많을 때 활용

EDA 실습

이 한 준 교수

EDA(Exploratory Data Analytics, 탐색적 자료 분석) 실습

- EDA란
  - 수집한 데이터가 들어왔을 때, 이를 다양한 각도에서 관찰하고 이해하는 과정

- 필요성
  - 데이터가 표현하는 현상을 더 잘 이해하고, 데이터에 대한 잠재적인 문제를 발견할 수 있음
  - 문제 정의 단계에서 미쳐 발생하지 못했을 다양한 패턴을 발견하고, 이를 바탕으로 기존의 가설을 수정하거나 새로운 가설을 세울 수 있음

EDA(Exploratory Data Analytics) 실습

- 이상치(Outlier) 발견 기법
  - 개별 데이터 관찰
  - 통계값 활용
  - 시각화 활용
  - 클러스터링(머신러닝 기법) 활용

EDA(Exploratory Data Analytics) 실습

- 속성에 따른 데이터 분류

Categorical Variable (Qualitative) 범주형 변수(정성)	Nominal Data 명목형 자료	원칙적으로 숫자로 표시할 수 없으나, 편의상 숫자화. (순위의 개념이 없음) 예시) 남자-0, 여자-1
	Ordinal Data 순서형 자료	원칙적으로 숫자로 표시할 수 없으나, 편의상 숫자화. (순위의 개념이 있음) 예시) 소득분위 10분위 > 9 분위 > 8 분위
Numeric Variable (Quantitative) 수치형 변수(정량)	Continuous Data 연속형 자료	데이터가 연속량으로서 셀 수 있는 형태. 예시) 키 - 166.1cm
	Discrete Data 이산형 자료	데이터가 비연 속량으로서 셀 수 있는 형태 예시) 자식 수 5명

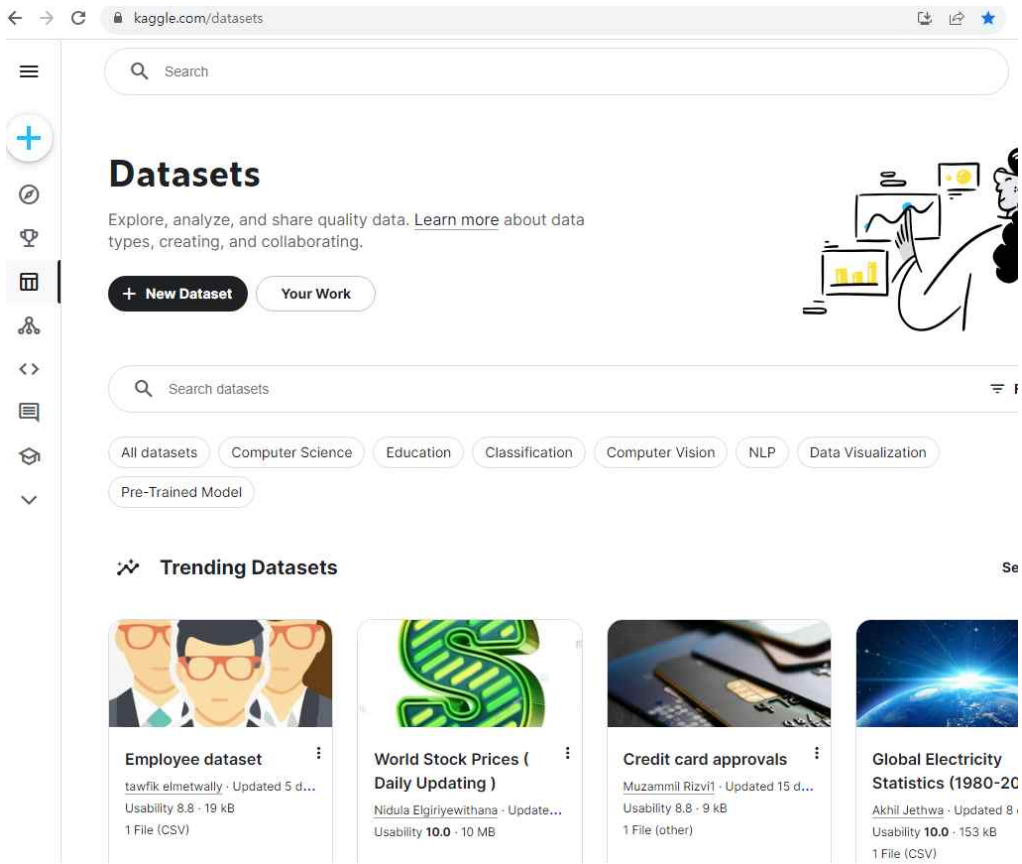
EDA(Exploratory Data Analytics) 실습

- 속성 간 관계 분석하기

데이터 조합	요약 통계	시각화
Categorical - Categorical	교차 테이블	모자이크 플롯 Ex. 성별과 당뇨병 유무
Numeric - Categorical	카테고리별 통계 값	박스 플롯 Ex. 체중과 당뇨병 유무
Numeric - Numeric	상관계수	산점도 Ex. 체중과 신장

EDA(Exploratory Data Analytics) 실습

- Kaggle 사이트 활
  - https://www.kaggle.com/
  - 혹시 가입 안된 경우 가입!

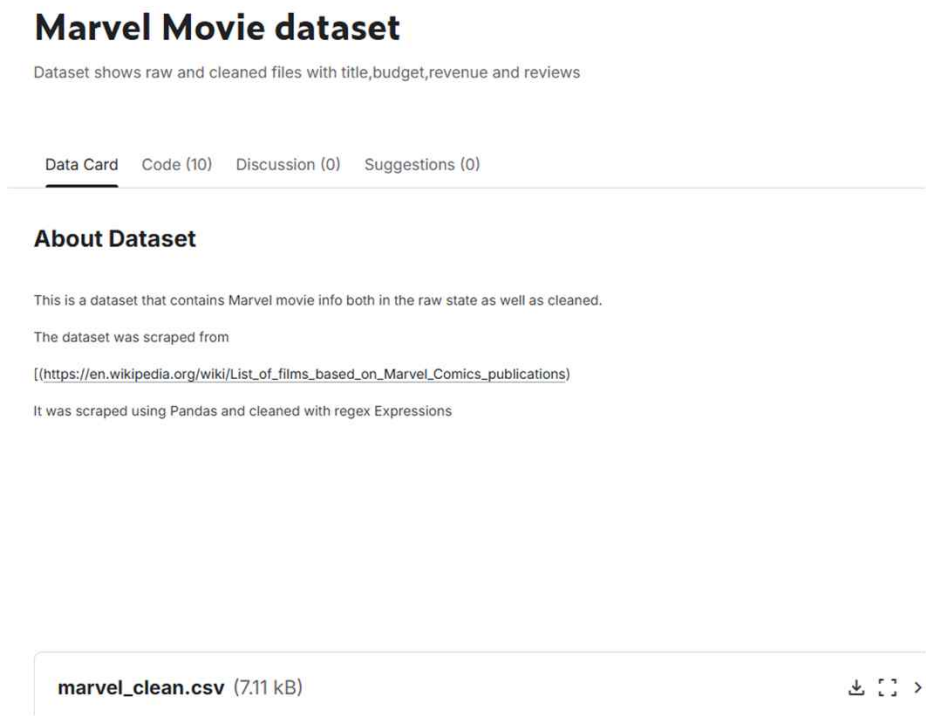


EDA 예시

- 데이터부터 준비해보자

  https://www.kaggle.com/datasets/minisam/marvel-movie-dataset?select=marvel_clean.csv

  - 마블 무비 데이터셋
  - **marvel_clean.csv** 다운받기 (marvel.csv는 전처리 전)
    - "data"폴더 아래에 두기



EDA 예시

- 데이터 설명
  - 칼럼

1. Title : 영화의 제목
2. Distributor : 영화 배급업체
3. ReleaseDateUS : 미국에서의 영화 개봉 시기
4. Budget : 영화예산(단위 : 백만달러)
5. OpeningWeekendNorthAmerica : 주말동안 미국에서 벌어들인 수익
6. NorthAmerica : 북미에서 벌어들인 총 수익
7. OtherTerritories : 다른 지역에서 벌어들인 총 수익
8. Worldwide : 전 세계에서 벌어들인 총 수익

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

데이터 설명

- 마블 영화들의 배급사, 배급일, 예산, 매출 관련 지표 포함

[44]:
```
df = pd.read_csv('data/marvel_clean.csv', index_col='Title')
df[:5]
```

[44]:

Title	Distributor	ReleaseDateUS	Budget	OpeningWeekendNorthAmerica	NorthAmerica	OtherTerritories	Worldwide
Howard the Duck	Universal Pictures	1986-08-01 00:00:00	37000000	5070136	16295774	21667000	37962774
Blade	New Line Cinema	1998-08-21 00:00:00	45000000	17073856	70087718	61095812	131183530
X-Men	20th Century Fox	2000-07-14 00:00:00	75000000	54471475	157299717	139039810	296339527
Blade II	New Line Cinema	2002-03-22 00:00:00	54000000	32528016	82348319	72661713	155010032
Spider-Man	Sony Pictures	2002-05-03 00:00:00	139000000	114844116	403706375	418002176	821708551

[46]:
```
df.columns
```

[46]:
```
Index(['Distributor', 'ReleaseDateUS', 'Budget', 'OpeningWeekendNorthAmerica',
       'NorthAmerica', 'OtherTerritories', 'Worldwide'],
      dtype='object')
```

살펴보고 싶은 주제

1. 배급사별로 사용한 예산과 평균 세계 매출, 그리고 예산 대비 매출 칼럼을 만들어보자
2. 가장 높은 예산, 매출을 가진 두 배급사의 영화 종류, 전세계 매출액을 비교해보자

[50]:
```
# 배급사 종류
df['Distributor'].unique()
```

[50]:
```
array(['Universal Pictures', 'New Line Cinema', '20th Century Fox',
       'Sony Pictures', 'Lionsgate Films', 'Paramount Pictures',
       'Walt Disney Studios Motion Pictures', 'IMAX Entertainment',
       '20th Century Studios'], dtype=object)
```

[52]:
```
# 배급사 종류 갯수
df['Distributor'].nunique()
```

[52]:
```
9
```

[56]:
```
# 필요없는 칼럼 제거(ReleaseDateUS)
df = df.drop('ReleaseDateUS', axis=1)
df[:5]
```

[56]:

Title	Distributor	Budget	OpeningWeekendNorthAmerica	NorthAmerica	OtherTerritories	Worldwide
Howard the Duck	Universal Pictures	37000000	5070136	16295774	21667000	37962774
Blade	New Line Cinema	45000000	17073856	70087718	61095812	131183530
X-Men	20th Century Fox	75000000	54471475	157299717	139039810	296339527
Blade II	New Line Cinema	54000000	32528016	82348319	72661713	155010032
Spider-Man	Sony Pictures	139000000	114844116	403706375	418002176	821708551

[62]:
```
# 영화사별로 그룹화한 뒤 지표 계산(단위: 백만)
df_company = df.groupby('Distributor').mean()/1000000
```

[64]:
```
df_company
```

[64]:

Distributor	Budget	OpeningWeekendNorthAmerica	NorthAmerica	OtherTerritories	Worldwide
20th Century Fox	125.823529	68.514330	170.840730	245.062349	415.903425
20th Century Studios	67.000000	7.037017	23.855569	24.819497	48.675066
IMAX Entertainment	0.000000	1.500000	1.521787	1.330495	2.852282
Lionsgate Films	34.000000	9.052989	20.930583	11.469487	32.400070
New Line Cinema	54.666667	21.887714	68.282648	70.083662	138.366309
Paramount Pictures	157.500000	89.380752	247.132640	260.118346	507.250986
Sony Pictures	160.307692	96.202437	297.595613	466.011377	763.606990
Universal Pictures	108.000000	40.870869	94.426640	87.823628	182.250268
Walt Disney Studios Motion Pictures	200.300000	139.083504	374.640128	633.036067	1007.752502

[68]:
```
# 예산(Budget) 대비 매출(Worldwide) 칼럼 추가
df_company['Budget_Profit_ratio'] = (df_company['Worldwide']/df_company['Budget']) * 100
```

[72]:
```
df_company
```

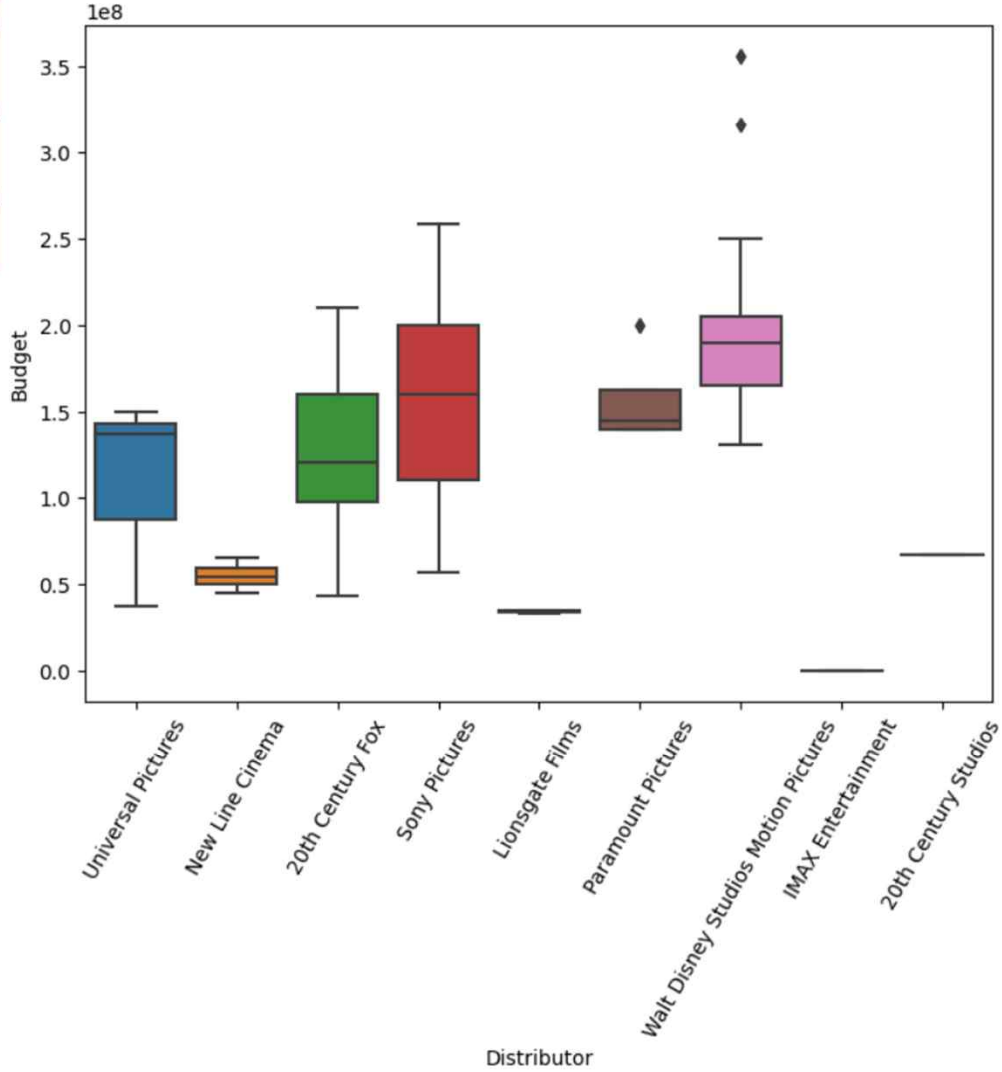
[72]:

Distributor	Budget	OpeningWeekendNorthAmerica	NorthAmerica	OtherTerritories	Worldwide	Budget_Profit_ratio
20th Century Fox	125.823529	68.514330	170.840730	245.062349	415.903425	330.545032
20th Century Studios	67.000000	7.037017	23.855569	24.819497	48.675066	72.649352
IMAX Entertainment	0.000000	1.500000	1.521787	1.330495	2.852282	inf
Lionsgate Films	34.000000	9.052989	20.930583	11.469487	32.400070	95.294325
New Line Cinema	54.666667	21.887714	68.282648	70.083662	138.366309	253.109102
Paramount Pictures	157.500000	89.380752	247.132640	260.118346	507.250986	322.064118
Sony Pictures	160.307692	96.202437	297.595613	466.011377	763.606990	476.338333
Universal Pictures	108.000000	40.870869	94.426640	87.823628	182.250268	168.750248
Walt Disney Studios Motion Pictures	200.300000	139.083504	374.640128	633.036067	1007.752502	503.121569

[78]:
```
# 배급사별 예산(Budget)을 그래프로 비교해보자 (박스플롯 예시)
data = pd.concat([df['Budget'], df['Distributor']], axis=1)
f, ax = plt.subplots(figsize=(8,6))
fig = sns.boxplot(x = 'Distributor', y='Budget', data = data)
plt.xticks(rotation=60)
```

```
C:\Users\Lee\anaconda3\lib\site-packages\seaborn\_core.py:1225: FutureWarning: is_categorical
n. Use isinstance(dtype, CategoricalDtype) instead
  if pd.api.types.is_categorical_dtype(vector):
C:\Users\Lee\anaconda3\lib\site-packages\seaborn\_core.py:1225: FutureWarning: is_categorical
n. Use isinstance(dtype, CategoricalDtype) instead
  if pd.api.types.is_categorical_dtype(vector):
C:\Users\Lee\anaconda3\lib\site-packages\seaborn\_core.py:1225: FutureWarning: is_categorical
n. Use isinstance(dtype, CategoricalDtype) instead
  if pd.api.types.is_categorical_dtype(vector):
```

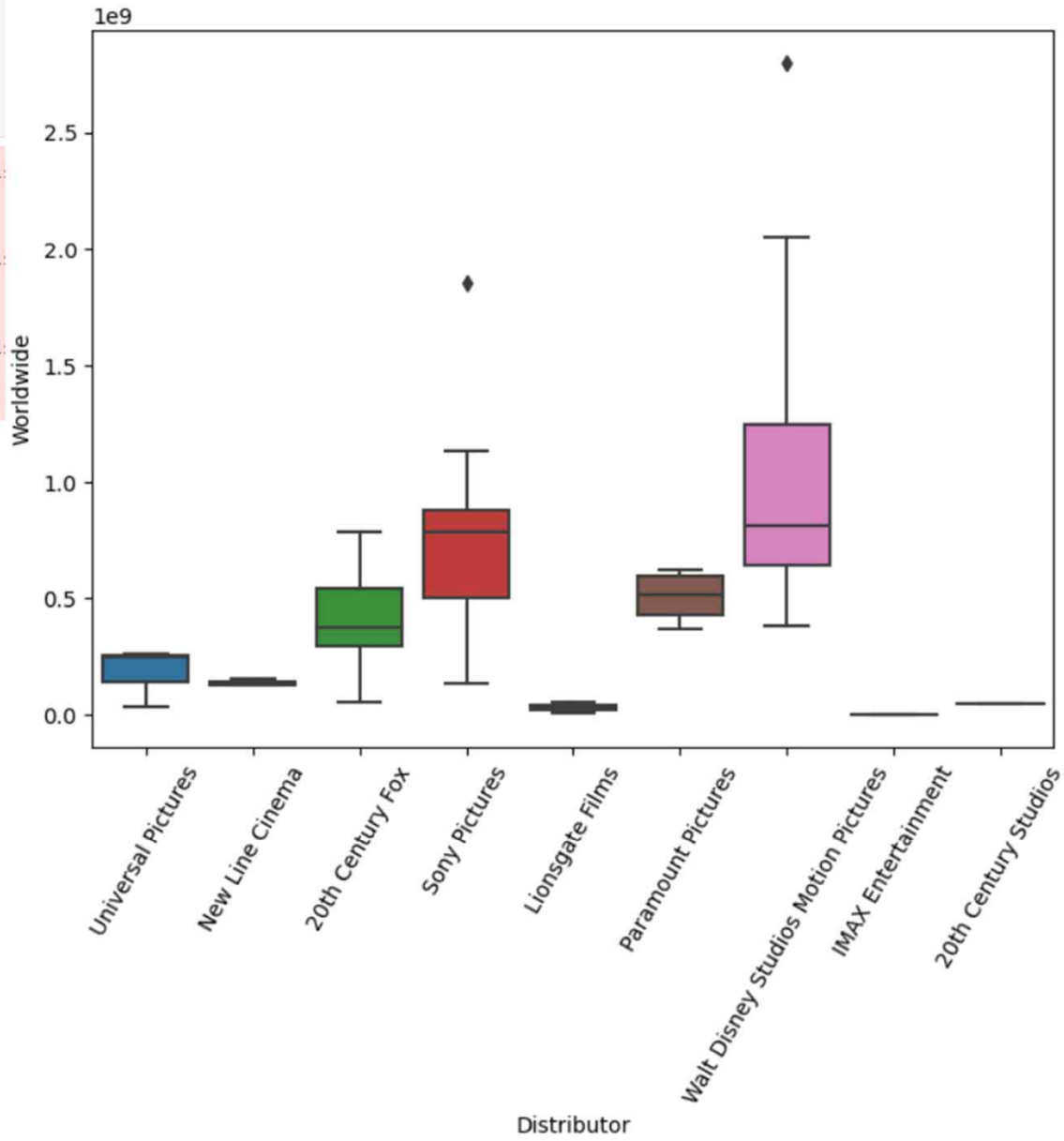
[78]:
```
(array([0, 1, 2, 3, 4, 5, 6, 7, 8]),
 [Text(0, 0, 'Universal Pictures'),
  Text(1, 0, 'New Line Cinema'),
  Text(2, 0, '20th Century Fox'),
  Text(3, 0, 'Sony Pictures'),
  Text(4, 0, 'Lionsgate Films'),
  Text(5, 0, 'Paramount Pictures'),
  Text(6, 0, 'Walt Disney Studios Motion Pictures'),
  Text(7, 0, 'IMAX Entertainment'),
  Text(8, 0, '20th Century Studios')])
```



```python
# 배급사별 전세계매출(Worldwide)을 그래프로 비교해보자 (박스플롯 예시)
data = pd.concat([df['Worldwide'], df['Distributor']], axis=1)
f, ax = plt.subplots(figsize=(8,6))
fig = sns.boxplot(x = 'Distributor', y='Worldwide', data = data)
plt.xticks(rotation=60)
```

```
C:\Users\Lee\anaconda3\lib\site-packages\seaborn\_core.py:1225: FutureWarning: i
n. Use isinstance(dtype, CategoricalDtype) instead
  if pd.api.types.is_categorical_dtype(vector):
C:\Users\Lee\anaconda3\lib\site-packages\seaborn\_core.py:1225: FutureWarning: i
n. Use isinstance(dtype, CategoricalDtype) instead
  if pd.api.types.is_categorical_dtype(vector):
C:\Users\Lee\anaconda3\lib\site-packages\seaborn\_core.py:1225: FutureWarning: i
n. Use isinstance(dtype, CategoricalDtype) instead
  if pd.api.types.is_categorical_dtype(vector):
```

```
(array([0, 1, 2, 3, 4, 5, 6, 7, 8]),
 [Text(0, 0, 'Universal Pictures'),
  Text(1, 0, 'New Line Cinema'),
  Text(2, 0, '20th Century Fox'),
  Text(3, 0, 'Sony Pictures'),
  Text(4, 0, 'Lionsgate Films'),
  Text(5, 0, 'Paramount Pictures'),
  Text(6, 0, 'Walt Disney Studios Motion Pictures'),
  Text(7, 0, 'IMAX Entertainment'),
  Text(8, 0, '20th Century Studios')])
```



Top 2 배급사 비교

- Sony Pictures와 Walt Disney Studios 비교

[89]:
```
# 두 배급사 상영작 비교
sony = df[df['Distributor']=='Sony Pictures']
sony[:5]
```

[89]:

Title	Distributor	Budget	OpeningWeekendNorthAmerica	NorthAmerica	OtherTerritories	Worldwide
Spider-Man	Sony Pictures	139000000	114844116	403706375	418002176	821708551
Spider-Man 2	Sony Pictures	200000000	88156227	373585825	415390628	788976453
Ghost Rider	Sony Pictures	110000000	45388836	115802596	112935797	228738393
Spider-Man 3	Sony Pictures	258000000	151116516	336530303	554341323	890871626
Ghost Rider: Spirit of Vengeance	Sony Pictures	57000000	22115334	51774002	80789928	132563930

[93]:
```
disney = df[df['Distributor']=='Walt Disney Studios Motion Pictures']
disney[:5]
```

[93]:

Title	Distributor	Budget	OpeningWeekendNorthAmerica	NorthAmerica	OtherTerritories	Worldwide
The Avengers	Walt Disney Studios Motion Pictures	220000000	207438708	623357910	895455078	1518812988
Iron Man 3	Walt Disney Studios Motion Pictures	200000000	174144585	409013994	805797258	1214811252
Thor: The Dark World	Walt Disney Studios Motion Pictures	170000000	85737841	206362140	438209262	644571402
Captain America: The Winter Soldier	Walt Disney Studios Motion Pictures	170000000	95023721	259766572	454497695	714264267
Guardians of the Galaxy	Walt Disney Studios Motion Pictures	170000000	94320883	333176600	440152029	773328629

[95]:
```
# 가장 높은 매출 기록 영화 비교
sony.sort_values(by='Worldwide', ascending=False)[:5]
```

[95]:

Title	Distributor	Budget	OpeningWeekendNorthAmerica	NorthAmerica	OtherTerritories	Worldwide
Spider-Man: No Way Home	Sony Pictures	200000000	260138569	780418859	1072000000	1852418859
Spider-Man: Far From Home	Sony Pictures	160000000	92579212	390532085	741395911	1131927996
Spider-Man 3	Sony Pictures	258000000	151116516	336530303	554341323	890871626
Spider-Man: Homecoming	Sony Pictures	175000000	117027503	334201140	545965784	880166924
Venom	Sony Pictures	100000000	80255756	213515506	641498448	855013954

[97]:
```
disney.sort_values(by='Worldwide', ascending=False)[:5]
```

[97]:

Title	Distributor	Budget	OpeningWeekendNorthAmerica	NorthAmerica	OtherTerritories	Worldwide
Avengers: Endgame	Walt Disney Studios Motion Pictures	356000000	357115007	858373000	1937901401	2797800564
Avengers: Infinity War	Walt Disney Studios Motion Pictures	316000000	257698183	678815482	1369544272	2048359754
The Avengers	Walt Disney Studios Motion Pictures	220000000	207438708	623357910	895455078	1518812988
Avengers: Age of Ultron	Walt Disney Studios Motion Pictures	250000000	191271109	459005868	946397826	1405403694
Black Panther	Walt Disney Studios Motion Pictures	200000000	202003951	700059566	646853595	1346913161

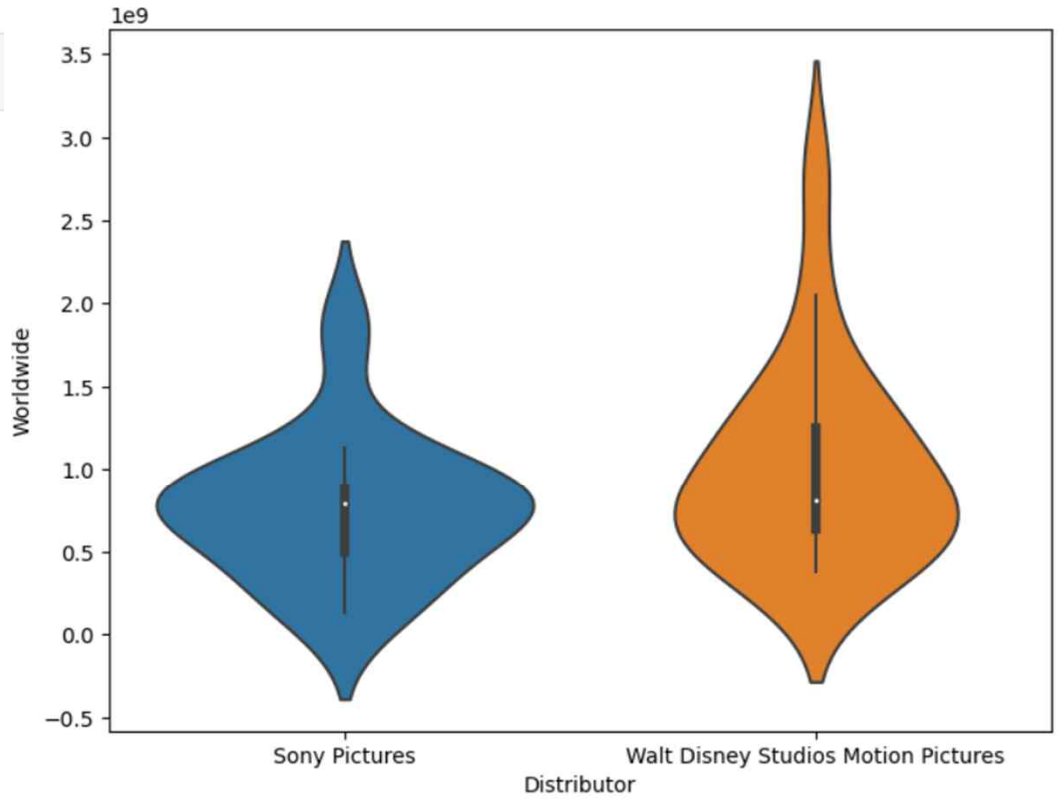
두 배급사의 매출 비교 그래프를 그려보자(바이올린 플롯 예시)

[108]:
```
# sony와 disney 데이터 병합
s_vs_d = pd.concat([sony, disney])
s_vs_d.head(3)
```

[108]:

Title	Distributor	Budget	OpeningWeekendNorthAmerica	NorthAmerica	OtherTerritories	Worldwide
Spider-Man	Sony Pictures	139000000	114844116	403706375	418002176	821708551
Spider-Man 2	Sony Pictures	200000000	88156227	373585825	415390628	788976453
Ghost Rider	Sony Pictures	110000000	45388836	115802596	112935797	228738393

[110]:
```
fig, ax = plt.subplots(figsize=(8,6))
sns.violinplot(x='Distributor', y='Worldwide', data=s_vs_d)
```



Paper Review

이 한 준 교수

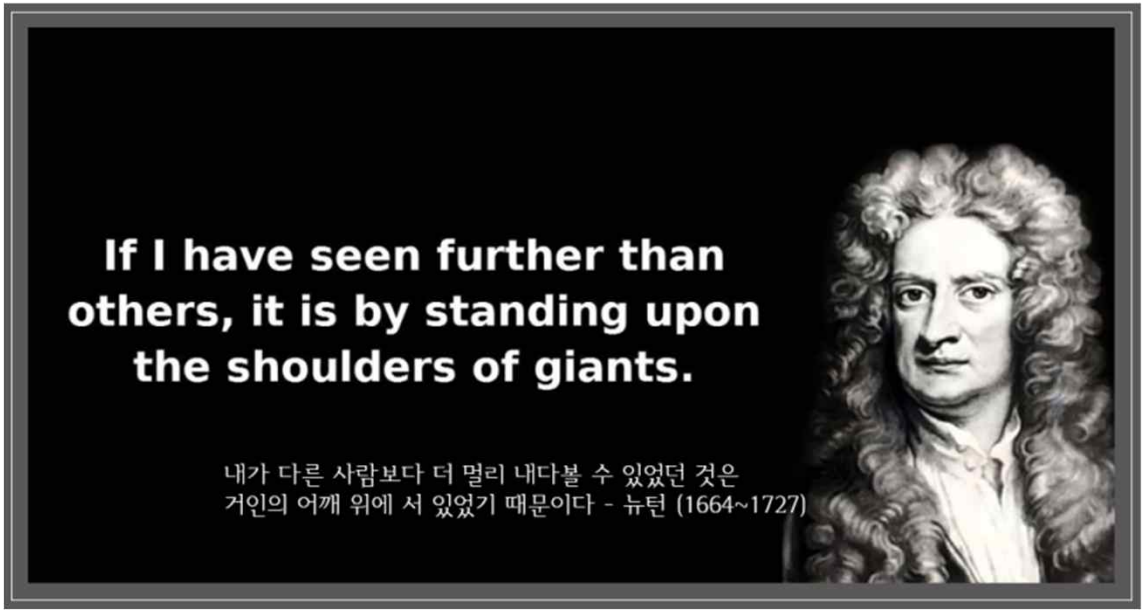
논문 리뷰

- 논문 리뷰의 목적
  - 연구 분야에 대한 이해 심화
    - 기존 연구자들의 연구를 파악하고 연구의 흐름과 최신 동향 이해할 수 있음
      - 데이터 분석 분야, 머신러닝 분야, 경영정보 분야에 어떤 연구들이 이루어지고 있고, 연구의 동향이 어떻게 되는지, 어떤 연구가 더 필요한지 등

  - 비판적 사고력과 분석 능력 제고
    - 논문의 연구 질문, 방법론, 데이터, 분석 과정, 결론을 평가하며 논리적, 비판적 사고력을 키울 수 있음
      - 논문은 저자의 주장을 논리적 근거를 기반으로 기술

논문 리뷰

- 논문 리뷰의 목적
  - 연구 아이디어 도출에 도움
    - 새로운 연구주제는 아무런 인풋 없이 생겨나지 않음



논문 리뷰

- 논문 리뷰의 목적
  - 연구 방법론 습득 및 적용 능력 제고
    - 이론적으로 배운 연구 방법을 실제 연구에 활용한 사례를 통해 학습할 수 있음
    - 새로운 연구 방법을 익힐 수 있음
  - 논문 작성 능력 제고
    - 논리적 글쓰기 훈련에 큰 도움

논문 리뷰

- 연구 역량 함양의 필수코스
  - 논문 리뷰는 “연구”에 대한 학습 과정

- 그래서!
  이번 25-1학기에는 데이터 분석 분야의 다양한 논문을 리뷰해 보자
  - 자신이 직접 리뷰해보고, 의문점을 가져보아야 도움이 됨
  - 쉬운 논문부터 조금씩 시작

논문 리뷰

- 논문 리뷰는 처음이라…
  - 모르는 용어, 개념에 대한 설명은 구글링이나 ChatGPT, 혹은 참고한 논문을 다시 참조하여 이해
  - 논문의 성격에 따라 새로운 기술을 제시하는 연구가 있고, 새로운 분야에 기술을 적용하는 연구가 있음
    - 이번 학기 리뷰할 논문들은 SOTA(State of the art) 모델은 아님
      - SOTA: 데이터 분석 분야에서 사용 가능한 현존 최고 성능의 모델
    - 경영정보 분야 도메인에 새로운 기법을 적용하여 시사점을 도출하는데 의의를 둔 연구 위주로 리뷰할 예정

논문 리뷰

- 이번 학기, 연구력을 키워보자
  - 연구력(연구 역량, 연구 경험, 논문 리뷰 경험)은 대학원 진학하지 않는 이상 필요없을까?
  - 다른 사회생활, 직업에는 별 쓸모 없을까?

논문 #1 기업가치 예측

- Paper #1. 머신러닝 기반 기업가치 예측 모형
  - Research Question: 기업가치(Q)를 예측하는 모델 구축
  - Dataset: 잡플래닛 리뷰
  - Methodology: 머신러닝

논문 #1 기업가치 예측

- 기업의 가치를 어떻게 측정할까?

논문 #1 기업가치 예측

📌 1. 회계적 지표 (Accounting-Based Metrics)

기업의 재무제표를 기반으로 가치를 평가하는 방식

🔷 1) 매출액 (Revenue)

- 일정 기간 동안 기업이 창출한 총 수익
- 기업의 규모와 성장성을 보여주는 기본적인 지표

🔷 2) 영업이익 (Operating Profit) & 순이익 (Net Profit)

- **영업이익**: 매출에서 영업 비용을 제외한 이익
- **순이익**: 영업이익에서 세금과 금융비용 등을 제외한 최종 이익
- 기업의 수익성과 효율성을 평가하는 데 중요

🔷 3) 자산총액 (Total Assets) & 부채총액 (Total Liabilities)

- 기업이 보유한 총 자산과 부채 규모를 나타냄
- **자기자본 (Equity) = 자산 - 부채**

🔷 4) ROA (총자산이익률, Return on Assets)

$ROA = (\text{순이익} / \text{총자산}) \times 100$

- 기업이 자산을 얼마나 효율적으로 활용하여 이익을 창출하는지 측정

🔷 5) ROE (자기자본이익률, Return on Equity)

$ROE = (\text{순이익} / \text{자기자본}) \times 100$

- 투자자 입장에서 기업이 자기자본을 활용해 얼마나 수익을 냈는지 평가

📌 2. 시장적 지표 (Market-Based Metrics)

주식시장 데이터를 활용하여 기업의 가치를 평가하는 방식

🔷 6) 시가총액 (Market Capitalization, Market Cap)

$\text{시가총액} = \text{주가} \times \text{발행주식수}$

- 기업이 주식시장에서 평가받는 가치를 나타냄
- 대형주(Large Cap), 중형주(Mid Cap), 소형주(Small Cap)로 구분

🔷 7) PER (주가수익비율, Price-to-Earnings Ratio)

$PER = \text{주가} / \text{주당순이익}(EPS)$

- 기업이 벌어들이는 이익 대비 주가 수준을 나타냄
- PER이 낮으면 저평가, 높으면 고평가로 해석 가능

🔷 8) PBR (주가순자산비율, Price-to-Book Ratio)

$PBR = \text{주가} / \text{주당순자산}(BPS)$

- 기업의 장부가치(Book Value) 대비 주가가 얼마나 높은지 측정

🔷 9) EV/EBITDA (기업가치 대비 영업이익 배수)

$EV/EBITDA = (\text{시가총액} + \text{순부채})/EBITDA$

- 기업의 전체 가치를 영업이익과 비교하여 평가
- 기업 간 비교 시 유용

📌 3. 경제적 가치 지표 (Economic Value-Based Metrics)

기업이 창출하는 경제적 부가가치를 반영하여 평가

🔷 10) EVA (경제적 부가가치, Economic Value Added)

$EVA = NOPAT - (\text{투하자본} \times \text{가중평균자본비용}, WACC)$

- 기업이 투자 비용 대비 초과 이익을 창출했는지 평가

🔷 11) MVA (시장 부가가치, Market Value Added)

$MVA = \text{시가총액} - \text{투하자본}(TotalInvestedCapital)$

- 기업이 투자자들에게 제공한 부가가치

🔷 12) FCF (자유현금흐름, Free Cash Flow)

$FCF = \text{영업현금흐름} - \text{자본지출}(CAPEX)$

- 기업이 운영과 투자를 지속할 수 있는 현금 여유

논문 #1 기업가치 예측

기업가치 Q (Tobin's Q)란?

Tobin's Q (토빈의 Q)는 경제학자 **제임스 토빈(James Tobin)**이 제안한 지표로, 기업의 **시장 가치(Market Value)**와 **자산의 대체 비용(Replacement Cost)** 간의 비율을 나타냅니다.

📌 공식:

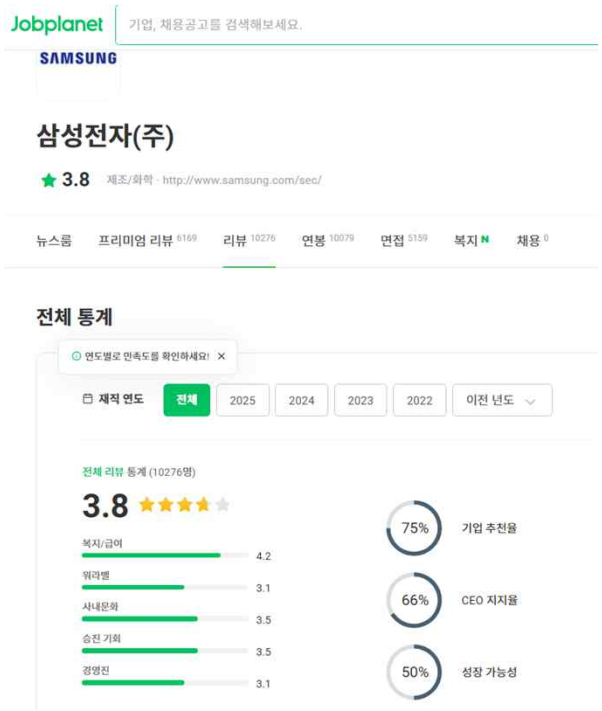
$$Q = \frac{\text{기업의 시장 가치 (Market Value)}}{\text{기업의 자산 대체 비용 (Replacement Cost)}}$$

Q 지표의 의미 🔍

- **Q > 1** :
  - 기업의 시장 가치가 자산의 재조달 비용보다 높음
  - 투자자들이 기업의 미래 성장 가능성을 높게 평가 (고평가 가능성)
  - 기업이 새로운 투자를 진행할 유인이 있음
- **Q = 1** :
  - 시장 가치와 자산 대체 비용이 동일
  - 기업이 보유한 자산 가치와 시장에서 평가받는 가치가 일치
- **Q < 1** :
  - 기업의 시장 가치가 자산의 대체 비용보다 낮음
  - 기업이 저평가되었거나 투자 매력이 낮음
  - 기업이 신규 투자를 진행할 유인이 적음

논문 #1 기업가치 예측

- 데이터셋
  - 잡플래닛

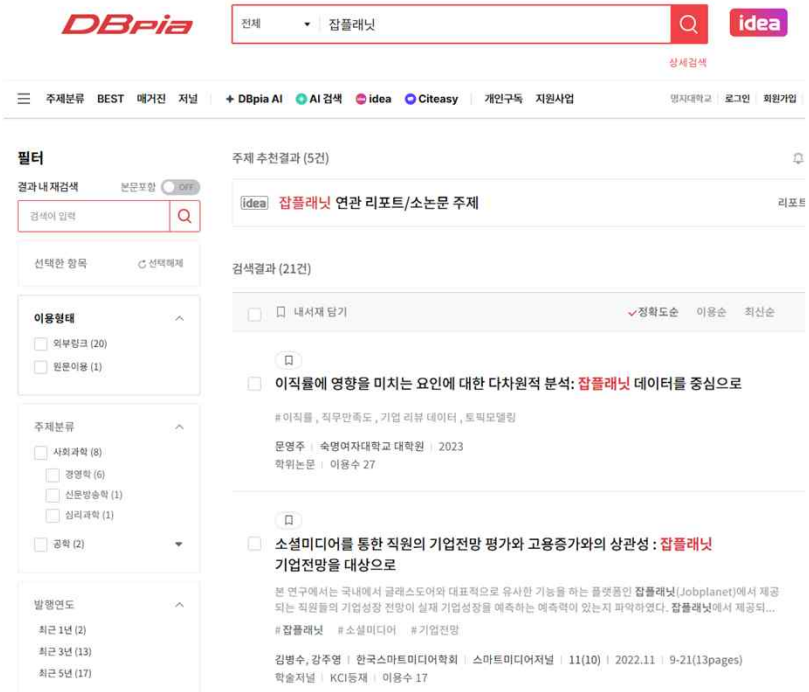


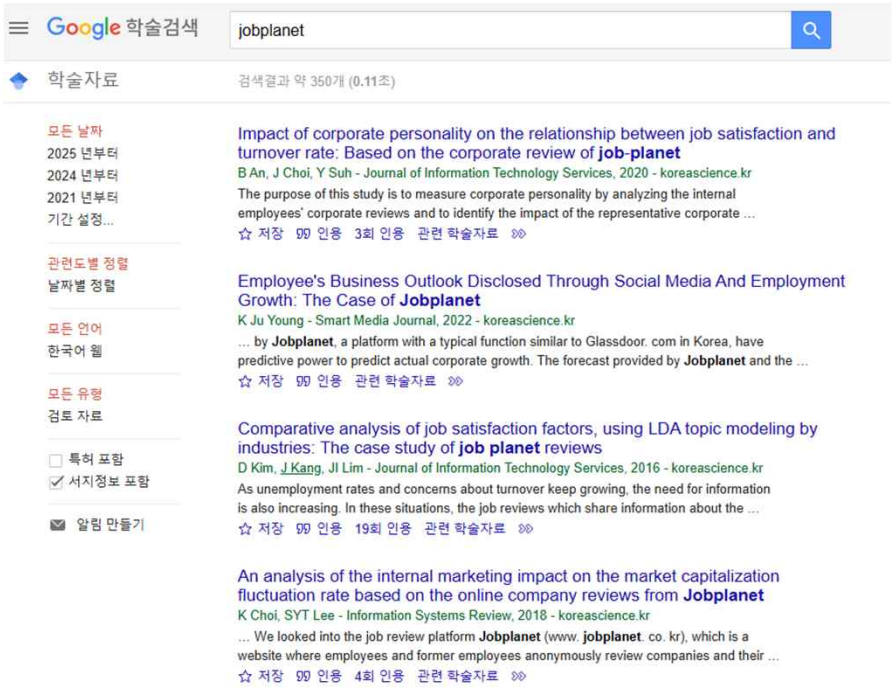
(그림 2) 잡플래닛에 게시된 기업리뷰 예시

(Figure 2) Example of Firm Review posted on the JobPlanet

논문 #1 기업가치 예측

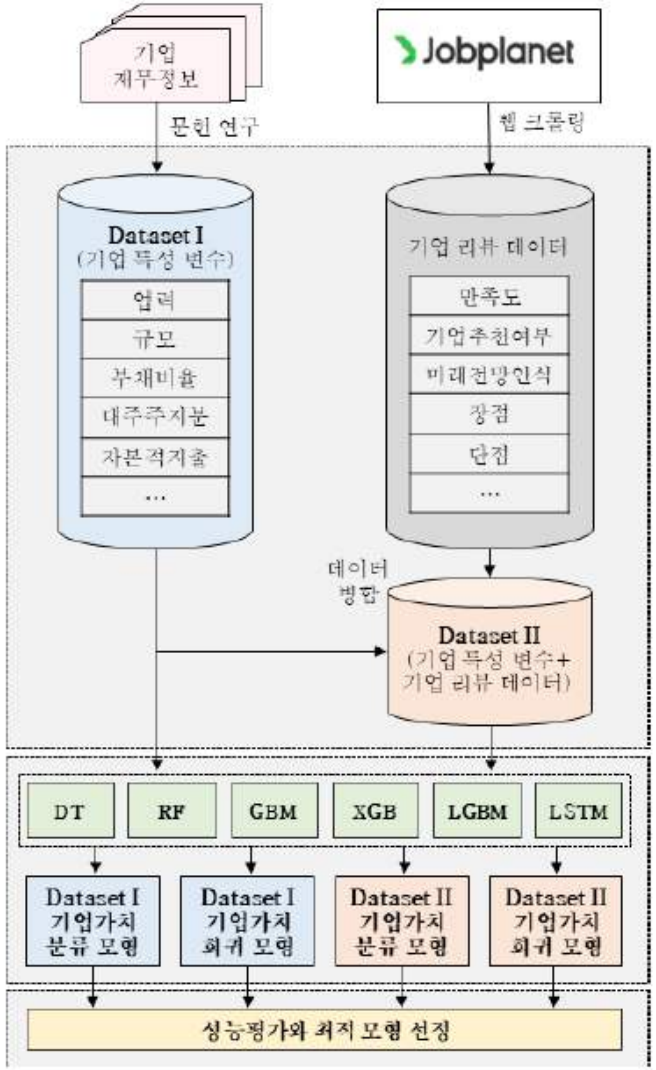
- 데이터셋
  - 잡플래닛 데이터로 어떤 연구들을 했을까? 구글스칼라, Dbpia 검색





논문 #1 기업가치 예측

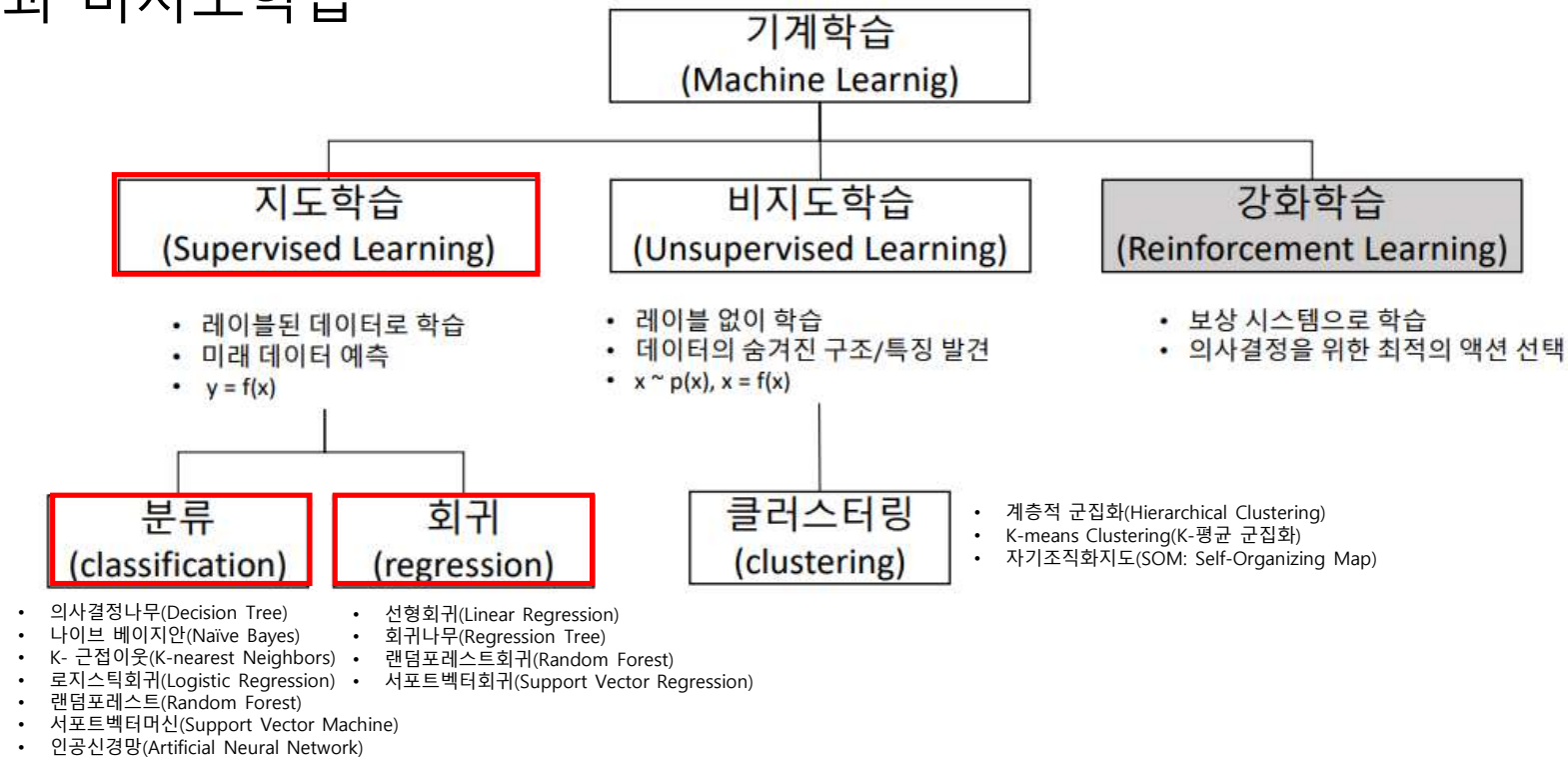
- Dataset II와 Dataset I 두 가지를 만든 이유?
- 지난 3년간의 데이터로 예측한 이유?
- 여러 머신러닝 알고리즘을 쓴 이유?



(그림 1) 연구흐름도
(Figure 1) Research Procedure

논문 #1 기업가치 예측

- 머신러닝 분류
  - 지도학습과 비지도학습



논문 #1 기업가치 예측

(표 1) 회귀모형 성능 비교
(Table 1) Performance Comparison among Regression Predictive Models

모형	Dataset I		Dataset II	
	MAE	RMSE	MAE	RMSE
DT	1.7591	1.3262	1.0375	1.0134
RF	0.4471	0.9878	0.3945	0.7102
GBM	0.5008	0.8802	0.4490	0.7739
XGB	0.4141	0.8256	0.4054	0.6702
LGBM	0.3993	0.7624	0.3760	0.6603
LSTM	0.3882	0.6545	**0.3587**	**0.4654**

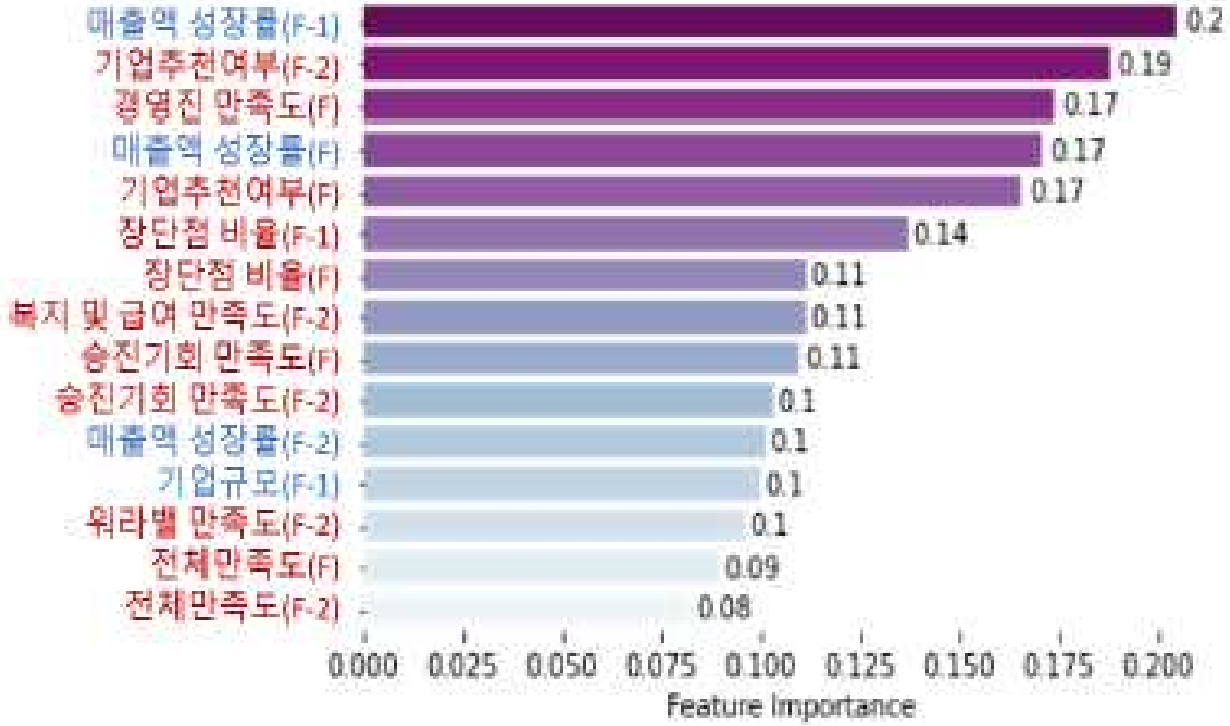
MAE(Mean Absolute Error)는 예측값과 실제값의 절대 오차 평균, RMSE(Root Mean Squared Error)는 제곱 오차의 평균을 루트 씌운 값

(표 2) 분류모형 예측 정확도 비교
(Table 2) Performance Comparison among Classification Predictive Models

모형	Dataset I				Dataset II			
	정확도	정밀도	재현율	F1점수	정확도	정밀도	재현율	F1점수
DT	60.0	62.2	49.4	55.0	61.4	64.1	51.8	57.3
RF	58.0	59.0	52.1	55.4	57.5	59.8	45.5	51.7
GBM	63.5	66.0	55.4	60.2	70.7	74.8	62.5	68.0
XGB	66.1	69.9	56.7	62.6	68.5	72.7	59.2	65.3
LGBM	58.4	61.8	44.1	51.5	61.3	68.4	42.3	52.3
LSTM	55.1	70.7	61.0	65.5	**73.2**	**75.8**	**85.1**	**80.2**

정밀도(Precision)는 모델이 긍정 클래스로 예측한 것 중 실제로 맞은 비율, 재현율(Recall)**은 실제 긍정 클래스 중에서 모델이 맞춘 비율

논문 #1 기업가치 예측



(그림 3) LSTM 분류모형의 변수중요도

(Figure 3) Feature Importance of LSTM Classification Predictive Model

팀별 활동

이 한 준 교수

팀별 활동

- 대략의 관심 분야 정해보기
  - Ex. 금융, 온라인 리뷰, … 등등
  - 추후 변경 가능

- 팀명 정하기
  - 단톡방에 캡처 올리기 – 팀명과 관심분야