

# 01. IBM and Oracle Stock Data (2004-2024) (조유경님)

<https://www.kaggle.com/datasets/ranatalha71/ibm-and-oracle-stock-data-2004-2024>

Data Files (1)

IBM\_ORCL.csv(233.19 kB)

Data Columns (8)

Date / Open / High / Low / Close / Adj Close / Volume / Stock

- Date: The date of the stock price record.
- Open: The opening price of the stock on that date.
- High: The highest price of the stock on that date.
- Low: The lowest price of the stock on that date.
- Close: The closing price of the stock on that date.
- Adj Close: The adjusted closing price of the stock on that date, accounting for corporate actions.
- Volume: The number of shares traded on that date.
- Stock: The ticker symbol (IBM or ORCL) indicating which company's stock data is represented.
- 날짜 : 주가 기록 날짜.
- 시작가 : 해당 날짜의 주식 시작가.
- 최고가 : 해당 날짜에 주식의 가장 높은 가격.
- 최저가 : 해당 날짜의 주식의 가장 낮은 가격.
- 종가 : 해당 날짜의 주식 종가.
- 조정 종가 : 기업 활동을 반영하여 해당 날짜에 주가가 조정된 종가를 말합니다.
- 거래량 : 해당 날짜에 거래된 주식 수.
- 주식 : 어떤 회사의 주가 데이터가 표현되는지를 나타내는 티커 기호(IBM 또는 ORCL).

## Use Cases

This dataset is ideal for:

- Financial Analysis: Conducting historical performance analysis of IBM and Oracle stocks.
- Machine Learning: Training models for stock price prediction.
- Comparative Studies: Comparing the market behavior and trends of IBM and Oracle over two decades.

- Investment Strategies: Backtesting investment strategies using historical data.
- 재무 분석 : IBM과 Oracle 주식의 과거 성과 분석을 실시합니다.
- 머신러닝 : 주가 예측을 위한 모델 학습.
- 비교 연구 : 20년 이상 IBM과 Oracle의 시장 행태와 추세를 비교합니다.
- 투자 전략 : 과거 데이터를 사용하여 투자 전략을 백테스팅합니다.

## 02. Amazon Sales Dataset (조유경님)

<https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset>

Data Files(1)

amazon.csv(4.74 MB)

Data Columns (16)

- product\_id - Product ID
- product\_name - Name of the Product
- category - Category of the Product
- discounted\_price - Discounted Price of the Product
- actual\_price - Actual Price of the Product
- discount\_percentage - Percentage of Discount for the Product
- rating - Rating of the Product
- rating\_count - Number of people who voted for the Amazon rating
- about\_product - Description about the Product
- user\_id - ID of the user who wrote review for the Product
- user\_name - Name of the user who wrote review for the Product
- review\_id - ID of the user review
- review\_title - Short review
- review\_content - Long review
- img\_link - Image Link of the Product
- product\_link - Official Website Link of the Product
- product\_id - 제품 ID
- product\_name - 제품 이름
- 카테고리 - 제품 카테고리
- discounted\_price - 제품의 할인된 가격
- actual\_price - 제품의 실제 가격
- discount\_percentage - 제품에 대한 할인율
- 평가 - 제품 평가
- rating\_count - Amazon 평가에 투표한 사람 수
- about\_product - 제품에 대한 설명
- user\_id - 제품에 대한 리뷰를 작성한 사용자의 ID
- user\_name - 제품에 대한 리뷰를 작성한 사용자의 이름
- review\_id - 사용자 리뷰의 ID

- review\_title - 짧은 리뷰
- review\_content - 긴 리뷰
- img\_link - 제품의 이미지 링크
- product\_link - 제품의 공식 웹사이트 링크

## Use Cases

- Dataset Walkthrough
- Understanding Dataset Hierarchy
- Data Preprocessing
- Exploratory Data Analysis
- Data Visualization
- Making Recommendation System

This is a list of some of that things that you can do on this dataset. It's not definitely limited to the one that is mentioned there but a lot more other things can also be done.

- 데이터 세트 워크스루
- 데이터 세트 계층 이해
- 데이터 전처리
- 탐색적 데이터 분석
- 데이터 시각화
- 추천 시스템 만들기

이것은 이 데이터 세트에서 할 수 있는 몇 가지 일의 목록입니다. 거기에 언급된 것에 확실히 국한되지는 않지만 훨씬 더 많은 다른 일도 할 수 있습니다.

### 03. Adidas Sales Dataset (조유경님)

<https://www.kaggle.com/datasets/heemalichaudhari/adidas-sales-dataset>

Data Files(1)

Adidas US Sales Datasets.xlsx(698.66 kB)

Data Columns(14)

-> 컬럼명 전처리 필수

Retailer / Retailer ID / Invoice Date / Region / State / City / Product /

Price per Unit / Units Sold / Total Sales / Operating Profit / Operating Margin / Sales Method

아래는 xlsx파일 열어본 결과

Adidas Sales Database													
Retailer	Retailer ID	Invoice Date	Region	State	City	Product	Price per Unit	Units Sold	Total Sales	Operating Profit	Operating Margin	Sales Method	
Foot Locker	1185732	2020-01-01	Northeast	New York	New York	Men's Street Foot	\$50.00	1,200	\$600,000	\$300,000	50%	In-store	
Foot Locker	1185732	2020-01-02	Northeast	New York	New York	Men's Athletic Foot	\$50.00	1,000	\$500,000	\$150,000	30%	In-store	
Foot Locker	1185732	2020-01-03	Northeast	New York	New York	Women's Street F	\$40.00	1,000	\$400,000	\$140,000	35%	In-store	
Foot Locker	1185732	2020-01-04	Northeast	New York	New York	Women's Athletic	\$45.00	850	\$382,500	\$133,875	35%	In-store	
Foot Locker	1185732	2020-01-05	Northeast	New York	New York	Men's Apparel	\$60.00	900	\$540,000	\$162,000	30%	In-store	
Foot Locker	1185732	2020-01-06	Northeast	New York	New York	Women's Apparel	\$50.00	1,000	\$500,000	\$125,000	25%	In-store	
Foot Locker	1185732	2020-01-07	Northeast	New York	New York	Men's Street Foot	\$50.00	1,250	\$625,000	\$312,500	50%	In-store	
Foot Locker	1185732	2020-01-08	Northeast	New York	New York	Men's Athletic Foot	\$50.00	900	\$450,000	\$135,000	30%	Outlet	
Foot Locker	1185732	2020-01-21	Northeast	New York	New York	Women's Street F	\$40.00	950	\$380,000	\$133,000	35%	Outlet	
Foot Locker	1185732	2020-01-22	Northeast	New York	New York	Women's Athletic	\$45.00	825	\$371,250	\$129,938	35%	Outlet	
Foot Locker	1185732	2020-01-23	Northeast	New York	New York	Men's Apparel	\$60.00	900	\$540,000	\$162,000	30%	Outlet	
Foot Locker	1185732	2020-01-24	Northeast	New York	New York	Women's Apparel	\$50.00	1,000	\$500,000	\$125,000	25%	Outlet	
Foot Locker	1185732	2020-01-25	Northeast	New York	New York	Men's Street Foot	\$50.00	1,220	\$610,000	\$305,000	50%	Outlet	
Foot Locker	1185732	2020-01-26	Northeast	New York	New York	Men's Athletic Foot	\$50.00	925	\$462,500	\$138,750	30%	Outlet	
Foot Locker	1185732	2020-01-27	Northeast	New York	New York	Women's Street F	\$40.00	950	\$380,000	\$133,000	35%	Outlet	
Foot Locker	1185732	2020-01-28	Northeast	New York	New York	Women's Athletic	\$45.00	800	\$360,000	\$126,000	35%	Outlet	
Foot Locker	1185732	2020-01-29	Northeast	New York	New York	Men's Apparel	\$60.00	850	\$510,000	\$153,000	30%	Outlet	
Foot Locker	1185732	2020-01-30	Northeast	New York	New York	Women's Apparel	\$50.00	950	\$475,000	\$118,750	25%	Outlet	
Foot Locker	1185732	2020-01-31	Northeast	New York	New York	Men's Street Foot	\$50.00	1,200	\$600,000	\$300,000	50%	Outlet	
Foot Locker	1185732	2020-02-01	Northeast	New York	New York	Men's Athletic Foot	\$50.00	900	\$450,000	\$135,000	30%	Outlet	
Foot Locker	1185732	2020-02-02	Northeast	New York	New York	Women's Street F	\$40.00	900	\$360,000	\$126,000	35%	Outlet	
Foot Locker	1185732	2020-02-03	Northeast	New York	New York	Women's Athletic	\$45.00	825	\$371,250	\$129,938	35%	Outlet	
Foot Locker	1185732	2020-02-04	Northeast	New York	New York	Men's Apparel	\$60.00	825	\$495,000	\$148,500	30%	Outlet	
Foot Locker	1185732	2020-02-05	Northeast	New York	New York	Women's Apparel	\$50.00	950	\$475,000	\$118,750	25%	Outlet	
Foot Locker	1185732	2020-02-06	Northeast	New York	New York	Men's Street Foot	\$60.00	1,220	\$732,000	\$366,000	50%	Outlet	
Foot Locker	1185732	2020-02-07	Northeast	New York	New York	Men's Athletic Foot	\$55.00	925	\$508,750	\$152,625	30%	Outlet	
Foot Locker	1185732	2020-02-08	Northeast	New York	New York	Women's Street F	\$50.00	900	\$450,000	\$157,500	35%	Outlet	
Foot Locker	1185732	2020-02-09	Northeast	New York	New York	Women's Athletic	\$50.00	850	\$425,000	\$148,750	35%	Outlet	
Foot Locker	1185732	2020-02-10	Northeast	New York	New York	Men's Apparel	\$60.00	875	\$525,000	\$157,500	30%	Outlet	
Foot Locker	1185732	2020-03-03	Northeast	New York	New York	Women's Apparel	\$65.00	1,000	\$650,000	\$162,500	25%	Outlet	
Foot Locker	1185732	2020-03-04	Northeast	New York	New York	Men's Street Foot	\$60.00	1,250	\$750,000	\$375,000	50%	Outlet	

## 04. e-Commerce (Walmart) Sales Dataset (조유경님)

<https://www.kaggle.com/datasets/devarajv88/walmart-sales-dataset>

Data Files(1)

walmart.csv(23.03 MB)

Data Columns (9)

1. **User\_ID**: User ID
2. **Product\_ID**: Product ID
3. **Gender**: Sex of User
4. **Age**: Age in bins
5. **Occupation**: Occupation(Masked)
6. **City\_Category**: Category of the City (A,B,C)
7. **StayInCurrentCityYears**: Number of years stay in current city
8. **Marital\_Status**: Marital Status
9. **ProductCategory**: Product Category (Masked)
10. **Purchase**: Purchase Amount
11. **User\_ID** : 사용자 ID
12. **Product\_ID** : 제품 ID
13. **성별** : 사용자의 성별
14. **연령** : 빈에 표시된 연령
15. **직업** : 직업(가면)
16. **City\_Category** : 도시의 카테고리 (A,B,C)
17. **StayInCurrentCityYears** : 현재 도시에 머무른 연수
18. **결혼 상태** : 결혼 상태
19. **ProductCategory** : 제품 카테고리(마스크됨)

### Useases

1. **Customer Segmentation**
2. **Market Basket Analysis**
3. **Personalized Marketing**
4. **Demand Forecasting**
5. **Product Recommendation Systems**

6. Customer Lifetime Value (CLV) Analysis
7. Sales and Revenue Analysis
8. Urban vs. Rural Analysis
9. Occupational Influence on Purchases
10. Customer Loyalty Programs
11. Marital Status and Shopping Behavior
12. Price Sensitivity Analysis
13. Optimizing pricing strategies to maximize sales and profitability
14. 고객 세분화
15. 마켓바구니 분석
16. 개인화된 마케팅
17. 수요 예측
18. 제품 추천 시스템
19. 고객 생애 가치(CLV) 분석
20. 판매 및 수익 분석
21. 도시 대 농촌 분석
22. 구매에 대한 직업적 영향
23. 고객 로열티 프로그램
24. 결혼 상태와 쇼핑 행동
25. 가격 민감도 분석
26. 판매와 수익성을 극대화하기 위한 가격 책정 전략 최적화

## 05. IBM Attrition Dataset (지승민님)

<https://www.kaggle.com/datasets/yasserh/ibm-attrition-dataset>

Data Files(1)

IBM.csv(94.18 kB)

Data Columns(13)

Age: Age of employee

Attrition: Employee attrition status

Department: Department of work

DistanceFromHome

Education: 1-Below College; 2- College; 3-Bachelor; 4-Master; 5-Doctor;

EducationField

EnvironmentSatisfaction: 1-Low; 2-Medium; 3-High; 4-Very High;

JobSatisfaction: 1-Low; 2-Medium; 3-High; 4-Very High;

MaritalStatus

MonthlyIncome

NumCompaniesWorked: Number of companies worked prior to IBM

WorkLifeBalance: 1-Bad; 2-Good; 3-Better; 4-Best;

YearsAtCompany: Current years of service in IBM

나이: 직원의 나이

이탈률: 직원 이탈 상태

부서: 업무 부서 집에서의

거리

교육: 1-대학 이하; 2-대학; 3-학사; 4-석사; 5-박사;

교육 현장

환경 만족도: 1-낮음; 2-보통; 3-높음; 4-매우 높음;

직무 만족도: 1-낮음; 2-보통; 3-높음; 4-매우 높음;

결혼 상태

월 소득

근무한 회사 수: IBM 이전에 근무한 회사 수

WorkLifeBalance: 1-나쁨; 2-좋음; 3-중음; 4-최고;

YearsAtCompany: IBM에서 현재 근무한 연수

## UseCases



- Understand the Dataset & cleanup (if required).
- Build classification models to predict the anticipated attrition of employees.
- Also fine-tune the hyperparameters & compare the evaluation metrics of various classification algorithms.
- 데이터 세트를 이해하고 필요한 경우 정리합니다.
- 직원의 예상 이탈을 예측하기 위해 분류 모델을 구축합니다.
- 또한 하이퍼파라미터를 미세 조정하고 다양한 분류 알고리즘의 평가 지표를 비교합니다.

## 06. Trending YouTube Video Statistics (조유경님)

<https://www.kaggle.com/datasets/datasnaek/youtube-new>


Data Files(csv : 10 / json : 10)

- CA\_category\_id.json
- CAvideos.csv
- DE\_category\_id.json
- DEvideos.csv
- FR\_category\_id.json
- FRvideos.csv
- GB\_category\_id.json
- GBvideos.csv
- IN\_category\_id.json
- INvideos.csv
- JP\_category\_id.json
- JPvideos.csv
- KR\_category\_id.json
- KRvideos.csv
- MX\_category\_id.json
- MXvideos.csv
- RU\_category\_id.json
- RUvideos.csv
- US\_category\_id.json
- USvideos.csv

Data Columns

- 10개의 csv 파일 / 10개의 json파일
- json파일 예시

## About this file

 Add Suggestion

This file does not have a description yet.

```
▼ "root" : { 3 items
  "kind" : string "youtube#videoCategoryListResponse"
  "etag" : string ""Id9bINPKjAjjV7EZ4EKKeEGrhao/1v2mrzYSYG6onNLt2qTj13hkQZk""
  ▼ "items" : [ 31 items
    ▼ 0 : { 4 items
      "kind" : string "youtube#videoCategory"
      "etag" : string ""Id9bINPKjAjjV7EZ4EKKeEGrhao/Xy1mB4_yLrHy_BmKmPBggy2mZQ""
      "id" : string "1"
      ▶ "snippet" : { ... } 3 items
    }
    ▼ 1 : { 4 items
      "kind" : string "youtube#videoCategory"
      "etag" : string ""Id9bINPKjAjjV7EZ4EKKeEGrhao/UZ1oLI1z2dx1h045ZTFR3a3NyTA""
      "id" : string "2"
      ▶ "snippet" : { ... } 3 items
    }
    ▼ 2 : { 4 items
      "kind" : string "youtube#videoCategory"
      "etag" : string ""Id9bINPKjAjjV7EZ4EKKeEGrhao/nqRIq97-xe5XRZTxbknKFVe5Lmg""
      "id" : string "10"
      ▶ "snippet" : { ... } 3 items
    }
    ▶ 3 : { ... } 4 items
    ▶ 4 : { ... } 4 items
```

## 07. 서울주택도시공사\_자치구별 임대주택 현황(한동우님)

<https://www.data.go.kr/data/15063237/fileData.do#/tab-layer-file>

Data Files(1)

서울주택도시공사\_자치구별 임대주택 현황.csv

Data Columns(15)

자치구 / 영구임대 / 공공임대 / 국민임대 / 장기전세 / 주거환경 / 외국인임대 / 행복주택 / 재개발임대 / 역세권청년 / 다가구 / 도시형생활주택 / 전세임대 / 장기안심 / 기타임대

Data Rows(26)

서울시 25개 구 + 의정부시 Data

## 08. 대졸자직업이동경로조사(GOMS) (김민진님)

<https://analysis.keis.or.kr/gomsSubject.do#none>

**GOMS 주제별 통계**는 대졸자의 학교생활, 경제활동 상태, 사업체 특성, 구직활동 등 특정 주제에 대한 주요 내용을 연도별로 정리하여 제공하는 기초 통계입니다.

### Data Columns

학교생활
<a href="#">성·학교유형별 졸업자 분포</a>
<a href="#">전공계열·학교유형별 졸업자 분포</a>
<a href="#">성별 대학 선택 이유</a>
<a href="#">학교유형별 대학 선택 이유</a>
<a href="#">전공계열별 대학 선택 이유</a>
<a href="#">성별 전공 선택 이유</a>
<a href="#">학교유형별 전공 선택 이유</a>
<a href="#">전공계열별 전공 선택 이유</a>
<a href="#">성별 대학 및 전공 만족도</a>
<a href="#">학교유형별 대학 및 전공 만족도</a>
<a href="#">전공계열별 대학 및 전공 만족도</a>
경제활동 상태
<a href="#">성별 경제활동 상태</a>
<a href="#">학교유형별 경제활동 상태</a>
<a href="#">전공계열별 경제활동 상태</a>
현재 일자리
<a href="#">취업자의 성별 산업 분포</a>
<a href="#">취업자의 학교유형별 산업 분포</a>
<a href="#">취업자의 전공별 산업 분포</a>
<a href="#">취업자의 성별 직업 분포(~2016)</a>
<a href="#">취업자의 성별 직업 분포(2017~)</a>
<a href="#">취업자의 학교유형별 직업 분포(~2016)</a>
<a href="#">취업자의 학교유형별 직업 분포(2017~)</a>

<a href="#">취업자의 전공별 직업 분포(~2016)</a>
<a href="#">취업자의 전공별 직업 분포(2017~)</a>
<a href="#">취업자의 성별 사업체 규모</a>
<a href="#">취업자의 학교유형별 사업체 규모</a>
<a href="#">취업자의 전공별 사업체 규모</a>
<a href="#">취업자의 성별 사업체 유형</a>
<a href="#">취업자의 학교유형별 사업체 유형</a>
<a href="#">취업자의 전공별 사업체 유형</a>
<a href="#">취업자의 성별 종사상 지위</a>
<a href="#">취업자의 학교유형별 종사상 지위</a>
<a href="#">취업자의 전공별 종사상 지위</a>
근로소득
<a href="#">성별 월평균 근로소득</a>
<a href="#">학교유형별 월평균 근로소득</a>
<a href="#">전공계열별 월평균 근로소득</a>
<a href="#">산업별 월평균 근로소득</a>
<a href="#">직업별 월평균 근로소득(~2016)</a>
<a href="#">직업별 월평균 근로소득(2017~)</a>
<a href="#">사업체 규모별 월평균 근로소득</a>
<a href="#">사업체 유형별 월평균 근로소득</a>
<a href="#">종사상지위별 월평균 근로소득</a>
근로시간
<a href="#">성별 주당 평균 근로시간</a>
<a href="#">학교유형별 주당 평균 근로시간</a>
<a href="#">전공계열별 주당 평균 근로시간</a>
<a href="#">산업별 주당 평균 근로시간</a>
<a href="#">직업별 주당 평균 근로시간(~2016)</a>
<a href="#">직업별 주당 평균 근로시간(2017~)</a>
<a href="#">사업체 규모별 주당 평균 근로시간</a>
<a href="#">사업체 유형별 주당 평균 근로시간</a>
<a href="#">종사상지위별 주당 평균 근로시간</a>
구직활동

<a href="#">성별 평균 구직활동 기간</a>
<a href="#">학교유형별 평균 구직활동 기간</a>
<a href="#">전공계열별 평균 구직활동 기간</a>
<a href="#">성별 일자리 정보 얻는 방법</a>
<a href="#">학교유형별 일자리 정보 얻는 방법</a>
<a href="#">전공계열별 일자리 정보 얻는 방법</a>
<a href="#">성별 희망 사업체 유형</a>
<a href="#">학교유형별 희망 사업체 유형</a>
<a href="#">전공계열별 희망 사업체 유형</a>
<a href="#">성별 구직활동의 어려운 점</a>
<a href="#">학교유형별 구직활동의 어려운 점</a>
<a href="#">전공계열별 구직활동의 어려운 점</a>

## 09. 청년패널조사(YP)(김민진님)

<https://analysis.keis.or.kr/ypSubject.do#none>

**청년패널(YP) 주제별 통계**는 청년층의 학교생활, 취업자 분포, 임금, 구직활동 등 특정 주제에 대한 주요 내용을 연도별로 정리하여 제공하는 기초 통계입니다.

### Data Columns

|학교생활|

|[대학생의 성별 대학교 유형 분포](#)|

|[대학생의 성별 전공계열 분포](#)|

|[대학생의 성별 전반적 학교성적 분포](#)|

|[대학생의 성별 학비부담자 분포](#)|

|[대학생의 성별 아르바이트 종류](#)|

|[대학생의 성별 아르바이트 이유](#)|

|[대학생의 성별 아르바이트 경험 성과](#)|

|[대학생의 성별 휴학 사유](#)|

|취업자 분포|

|[취업자의 성별 분포](#)|

|[취업자의 학력별 분포](#)|

|[취업자의 종사상 지위별 분포](#)|

|[취업자의 산업별 분포](#)|

|[취업자의 직업별 분포](#)|

|[취업자의 지역별 분포](#)|

|[취업자의 회사유형별 분포](#)|

|[취업자의 사업체 규모별 분포](#)|

|[취업자의 성 · 학력별 분포](#)|

|[취업자의 성 · 종사상 지위별 분포](#)|

|[취업자의 학력 · 종사상 지위별 분포](#)|

|[취업자의 성 · 산업별 분포](#)|

|[취업자의 성 · 직업별 분포](#)|

|[취업자의 성 · 지역별 분포](#)|

|[취업자의 성 · 회사유형별 분포](#)|

|[취업자의 성 · 사업체 규모별 분포](#)|

|[취업자의 학력 · 산업별 분포](#)|

|[취업자의 학력 · 직업별 분포](#)|

|[취업자의 학력 · 지역별 분포](#)|

|[취업자의 학력 · 회사유형별 분포](#)|

|[취업자의 학력 · 사업체 규모별 분포](#)|



|임금(소득)|

|[성별 월평균 임금](#)|

|[학력별 월평균 임금](#)|

|[종사상 지위별 월평균 임금](#)|

|[산업별 월평균 임금](#)|

|[직업별 월평균 임금](#)|

|[회사유형별 월평균 임금](#)|

|[사업체규모별 월평균 임금](#)|

|[성 · 학력별 월평균 임금](#)|

|[성 · 종사상 지위별 월평균 임금](#)|

|[성 · 산업별 월평균 임금](#)|

|[성 · 직업별 월평균 임금](#)|

|[성 · 회사유형별 월평균 임금](#)|

|[성 · 사업체규모별 월평균 임금](#)|

|근로시간|

|[성별 주당 평균 총 근로시간](#)|

|[성별 주당 평균 근로시간](#)|

|[종사상지위별 주당 평균 근로시간](#)|

|[회사유형별 주당 평균 근로시간](#)|

|[사업체규모별 주당 평균 근로시간](#)|

|구직활동|

|[미취업자의 성별 평균 구직기간](#)|

|[미취업자의 학력별 평균 구직기간](#)|

|[미취업자의 성별 평균 구직 시도 횟수](#)|

|[미취업자의 학력별 평균 구직 시도 횟수](#)|

|[미취업자의 성별 취업준비 어려운점](#)|

## 10. 구인구직취업현황 (김민진님)

<https://eis.work.go.kr/eisps/rpt/reptDtl.do?menuId=020010010>

Data 양 상당히 많음

# 11. Employee Future Prediction (지승민님)

<https://www.kaggle.com/datasets/tejashvi14/employee-future-prediction>

Data Files(1)

Employee.csv(195.25 kB)

Data Columns(9)

Education / JoiningYear / City / PaymentTier / Age / Gender / EverBenched /  
ExperienceInCurrentDomain / LeaveOrNot

교육 / 가입년도 / 도시 / 결제 단계 / 나이 / 성별 / EverBenched / 현재 도메인에서의 경험 /  
LeaveOrNot

## UseCases

A company's HR department wants to predict whether some customers would leave the company in next 2 years. Your job is to build a predictive model that predicts the prospects of future and present employee.

Perform EDA and bring out insights

Dummy Data Used For A Private Hackathon

한 회사의 HR 부서는 일부 고객이 향후 2년 내에 회사를 떠날지 예측하고자 합니다. 여러분의 업무는 미래와 현재 직원의 전망을 예측하는 예측 모델을 구축하는 것입니다.

EDA를 수행하고 통찰력을 도출하기

개인 해커톤에 사용된 더미 데이터

## 12. UCI SECOM Dataset (조유경님)

<https://www.kaggle.com/datasets/paresh2047/uci-semcom>

Data Files(1)

uci-secom.csv(6.06 MB)

Data Columns(592)

- Data Set Characteristics: Multivariate
- Number of Instances: 1567
- Area: Computer
- Attribute Characteristics: Real
- Number of Attributes: 591
- Date Donated: 2008-11-19
- Associated Tasks: Classification, Causal-Discovery
- Missing Values? Yes
- 데이터 세트 특성: 다변량
- 인스턴스 수: 1567
- 지역: 컴퓨터
- 속성 특성: 실제
- 속성 수: 591
- 기부일: 2008-11-19
- 관련 작업: 분류, 인과 발견
- 누락된 값? 예
- 전처리 필요

# 13. 서울시 1인가구(연령별) 통계(한동우님)

<https://data.seoul.go.kr/dataList/10995/S/2/datasetView.do>

1인가구(연령별)

자료갱신일 : 2023-07-28 / 수록기간 : 년 2010 ~ 2022

출처 : 서울특별시, 서울특별시기본통계

알람설정

등계표조회

항목[1/1]

자치구별[26/26]

성별[3/3]

연령별[16/16]

시점[1/9]

(단위 : 가구)

새창보기

주석보기

주소보기

행열전환

부가기능설정

분석

차트

다운로드

인쇄

도움말

자치구별(1)	자치구별(2)	성별(1)	2022			
			합계			
			소계	20세미만	20~24세	25~29세
합계	소계	계	1,564,187	12,872	129,586	
		남자	730,639	5,074	45,438	
		여자	833,548	7,798	84,148	
	종로구	계	28,424	503	3,318	
		남자	13,561	170	1,201	
		여자	14,863	333	2,117	
	중구	계	25,247	351	1,931	
		남자	11,943	102	608	
		여자	13,304	249	1,323	
	용산구	계	41,437	349	2,641	
		남자	18,916	35	550	
		여자	22,521	314	2,091	
	성동구	계	46,548	665	5,046	
		남자	22,786	276	2,158	
		여자	23,762	389	2,888	
	광진구	계	69,391	475	7,516	
		남자	32,997	172	2,828	
		여자	36,394	303	4,688	
	동대문구	계	68,169	1,368	11,456	
		남자	33,750	570	4,544	
		여자	34,419	798	6,912	
	종랑구	계	64,334	213	2,886	
		남자	31,290	78	926	
		여자	33,044	135	1,960	

사이트 자체에서 분석 / 차트 기능이 있음  
python으로 하기엔 난이도가 있을듯.

## 14. World Bank Youth Unemployment Rates(박창현)

<https://www.kaggle.com/datasets/sovannt/world-bank-youth-unemployment>

Data Files(1)

API\_ILO\_country\_YU.csv(17.15 kB)

Data Columns(7)

Country Name / Country Code / 2010 / 2011 / 2012 / 2013 / 2014

### About Dataset

World Bank - Youth Unemployment rates (IPO) by country, 2010 - 2014

나라 이름별 / 연도 별 IPO 수치

그다지 할 수 있는게 많지 않을듯.

## 15. Spotify Popular East Asian Artists and Tracks(강민우님)

<https://www.kaggle.com/datasets/crxxom/spotify-popular-east-asian-artists-and-tracks>

Data Files(16)

Top 100 artists(7 files)

Top 1000 tracks(7 files)

east\_asia\_top\_artists.csv(208.2 kB)

east\_asia\_top\_tracks.csv(1.36 MB)

Data Columns(10 / 14)

### Features in top track datasets

1. **song\_name**: name of the song
2. **album\_name**: the album that the song is in
3. **album\_link**: external link to the album of the song on Spotify
4. **artist\_name**: name of the artist that produced the song
5. **popularity**: a popularity metric calculated by Spotify, ranging from 0-100 with 100 being the most popular, a song that is popular recently is more likely to have a higher score than a popular song in the past
6. **release\_date**: release date of the song
7. **song\_link**: external link to the song on Spotify
8. **duration\_ms**: duration of song in milliseconds
9. **explicit**: boolean value that indicated if the song is explicit or not
10. **query\_genre**: query\_genre of the track

### Features in top artist datasets

1. **artist\_name**: name of the artist
2. **popularity**: a popularity metric calculated by Spotify, ranging from 0-100 with 100 being the most popular

3. **followers**: number of followers of the artist on Spotify
4. **artist\_link**: external link to artist page on Spotify
5. **genres**: list of genres that the artist is involved in
6. **top\_track**: top track of the artist based on Spotify API
7. **top\_track\_album**: the album that the top track is in
8. **top\_track\_popularity**: the popularity of top track; a popularity metric calculated by Spotify, ranging from 0-100 with 100 being the most popular
9. **top\_track\_release\_date**: release date of the top track
10. **top\_track\_duration\_ms**: the duration of the top track in milliseconds
11. **top\_track\_explicit**: boolean value that indicated if the top track is explicit or not
12. **top\_track\_link**: external link to the top track on Spotify
13. **top\_track\_album\_link**: external link to the album of the top track on Spotify
14. **query\_genre**: query\_genre of the artist

## 상위 트랙 데이터 세트의 기능

1. **song\_name** : 노래의 이름
2. **album\_name** : 노래가 들어 있는 앨범
3. **album\_link** : Spotify의 노래 앨범에 대한 외부 링크
4. **artist\_name** : 노래를 제작한 아티스트의 이름
5. **인기** : Spotify에서 계산한 인기 지표로 0~100점 범위이며 100점이 가장 인기 있는 점수입니다. 최근에 인기 있는 노래는 과거에 인기 있었던 노래보다 점수가 높을 가능성이 더 높습니다.
6. **release\_date** : 노래의 발매일
7. **song\_link** : Spotify의 노래에 대한 외부 링크
8. **duration\_ms** : 노래의 길이(밀리초)
9. **explicit** : 노래가 노골적인지 아닌지를 나타내는 부울 값
10. **query\_genre** : 트랙의 query\_genre

## 상위 아티스트 데이터 세트의 기능

1. **artist\_name** : 아티스트 이름
2. **인기도** : Spotify가 계산한 인기도 지표로 0~100까지이며 100이 가장 인기 있는 수치입니다.
3. **팔로워** : Spotify에서 아티스트를 팔로워하는 수
4. **artist\_link** : Spotify의 아티스트 페이지로의 외부 링크
5. **장르** : 아티스트가 참여하는 장르 목록
6. **top\_track** : Spotify API에 따른 아티스트의 톱 트랙



7. **top\_track\_album** : 상위 트랙이 들어 있는 앨범
8. **top\_track\_popularity** : 인기 트랙의 인기도. Spotify에서 계산한 인기 지표로 0~100까지이며 100이 가장 인기 있는 트랙입니다.
9. **top\_track\_release\_date** : 상위 트랙의 출시 날짜
10. **top\_track\_duration\_ms** : 상단 트랙의 지속 시간(밀리초)
11. **top\_track\_explicit** : 최상위 트랙이 명시적인지 여부를 나타내는 부울 값
12. **top\_track\_link** : Spotify의 상위 트랙에 대한 외부 링크
13. **top\_track\_album\_link** : Spotify의 상위 트랙 앨범에 대한 외부 링크
14. **query\_genre** : 아티스트의 query\_genre

## 16. The Movies Dataset (박창현)

<https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset?select=keywords.csv>

Data Files(7)

**credits.csv**(189.92 MB)

**keywords.csv**(6.23 MB)

**links.csv**(989.11 kB)

**links\_small.csv**(183.37 kB)

**movies\_metadata.csv**(34.45 MB)

**ratings.csv**(709.55 MB)

**ratings\_small.csv**(2.44 MB)

Data Files Content

**movies\_metadata.csv:** The main Movies Metadata file. Contains information on 45,000 movies featured in the Full MovieLens dataset. Features include posters, backdrops, budget, revenue, release dates, languages, production countries and companies.

**keywords.csv:** Contains the movie plot keywords for our MovieLens movies. Available in the form of a stringified JSON Object.

**credits.csv:** Consists of Cast and Crew Information for all our movies. Available in the form of a stringified JSON Object.

**links.csv:** The file that contains the TMDB and IMDB IDs of all the movies featured in the Full MovieLens dataset.

**links\_small.csv:** Contains the TMDB and IMDB IDs of a small subset of 9,000 movies of the Full Dataset.

**ratings\_small.csv:** The subset of 100,000 ratings from 700 users on 9,000 movies.

The Full MovieLens Dataset consisting of 26 million ratings and 750,000 tag applications from 270,000 users on all the 45,000 movies in this dataset can be accessed [here](#)

**movies\_metadata.csv:** 주요 영화 메타데이터 파일입니다. Full MovieLens 데이터 세트에 등장하는 45,000개의 영화에 대한 정보가 들어 있습니다. 포스터, 배경, 예산, 수익, 출시일, 언어, 제작 국가 및 회사가 특징입니다.

**keywords.csv:** MovieLens 영화의 영화 줄거리 키워드를 포함합니다. 문자열화된 JSON 객체 형태로 제공됩니다.

**credits.csv:** 모든 영화의 출연진 및 제작진 정보로 구성되어 있습니다. 문자열화된 JSON 객체 형태로 제공됩니다.

**links.csv:** Full MovieLens 데이터 세트에 포함된 모든 영화의 TMDB 및 IMDB ID가 포함된 파일입니다.

**links\_small.csv:** 전체 데이터 세트의 9,000개 영화 중 작은 하위 세트의 TMDB 및 IMDB ID를 포함합니다.

**ratings\_small.csv:** 9,000개 영화에 대한 700명의 사용자의 100,000개 평점의 하위 집합입니다.

이 데이터 세트에 있는 45,000개 영화에 대한 270,000명의 사용자로부터 얻은 2,600만 개의 평가와 750,000개의 태그 애플리케이션으로 구성된 Full MovieLens 데이터 세트는 [여기에서 액세스할 수 있습니다.](#)

## 17.Youth Employment Average (2014-2023) (조유경님)

<https://www.kaggle.com/datasets/evelyn001k/youth-employment-average-2014-2023>

Data Files(4)

Average version by Country - Education average.csv(15.29 kB)

Average version by Country - Employment average.csv(15.18 kB)

Average version by Country - Parents living average.csv(2.37 kB)

Average version by Country - Youth Population average.csv(2.36 kB)

Data Columns(4/4/3/3)

Observation Country / Sex / isced 11 / Educated levels

Observation Country / Sex / isced 11 / Educated levels

Observation Country / Sex / Ovserved Value

Observation Country / Sex / Ovserved Value

관찰 국가 / 성별 / 11세 / 교육 수준

관찰 국가 / 성별 / 11세 / 교육 수준

관찰 국가 / 성별 / 관측값

관찰 국가 / 성별 / 관측값

## 18. UNEMPLOYMENT, YOUTH AGES 15-24 (조유경님)

<https://www.kaggle.com/datasets/sreelalh/unemployment-youth-ages-1524>

Data Files(1)

**YouthUnemploy.csv(3.17 kB)**

Data Columns(3)

Rank / Country / Unemployment

According to the OECD, unemployment is people above a specified age not being paid employment or self-employment but currently available for work during the reference period. Unemployment is measured by the unemployment rate, which is the number of unemployed people as a percentage of the labour force. This dataset contains youth unemployment rates (ages 15-24).

OECD에 따르면 실업률은 특정 연령 이상의 사람들이 급여를 받지 않고 자영업을 하지 않지만 기준 기간 동안 현재 일할 수 있는 사람을 말합니다. 실업률은 실업률로 측정되며, 이는 노동력의 백분율로 나타낸 실업자 수입입니다. 이 데이터 세트에는 청년 실업률(15~24세)이 포함되어 있습니다.

## 19. Music Genre Classification (박창현)

<https://www.kaggle.com/datasets/purumalgi/music-genre-classification?select=train.csv>

### Data Files(3)

- submission.csv
- test.csv
- train.csv

### Data Columns

- submission.csv (11)
- Acoustic/Folk\_0
- Alt\_Music\_1
- Blues\_2
- Bollywood\_3
- Country\_4
- HipHop\_5
- Indie Alt\_6
- Instrumental\_7
- Metal\_8
- Pop\_9
- Rock\_10
- test.csv(16)
- Artist Name
- Track Name
- Popularity
- danceability
- energy
- key
- loudness
- mode
- speechiness
- acousticness
- instrumentalness
- liveness

- valence
- tempo
- duration\_in min/ms
- time\_signature
- train.csv (17)
- Artist Name
- Track Name
- Popularity
- danceability
- energy
- key
- loudness
- mode
- speechiness
- acousticness
- instrumentalness
- liveness
- valence
- tempo
- duration\_in min/ms
- time\_signature
- Class

## 20. Korea Income and Welfare (박창현)

<https://www.kaggle.com/datasets/hongsean/korea-income-and-welfare/data>

Data Files(2)

Korea Income and Welfare.csv(4.33 MB)

job\_code\_translated.xlsx(19.87 kB)

Data Columns(14)

- id
- year : study conducted
- wave : from wave 1st in 2005 to wave 14th in 2018
- region: 1) Seoul 2) Gyeonggi 3) Gyeongnam 4) Gyeongbuk 5) Chungnam 6) Gangwon & Chungbuk 7) Jeolla & Jeju
- income: yearly income in M KRW (Million Korean Won. 1100 KRW = 1 USD)
- family\_member: no. of family members
- gender: 1) male 2) female
- year\_born
- education\_level: 1) no education (under 7 yrs-old) 2) no education (7 & over 7 yrs-old) 3) elementary 4) middle school 5) high school 6) college 7) university degree 8) MA 9) doctoral degree
- marriage: marital status. 1) not applicable (under 18) 2) married 3) separated by death 4) separated 5) not married yet 6) others
- religion: 1) have religion 2) do not have
- occupation: this will be provided in separated code book
- company\_size
- reason\_none\_worker: 1) no capable 2) in military service 3) studying in school 4) prepare for school 5) prepare to apply job 6) house worker 7) caring kids at home 8) nursing 9) giving-up economic activities 10) no intention to work 11) others
- ID
- 년도 : 연구 수행
- 웨이브 : 2005년 1차 웨이브부터 2018년 14차 웨이브까지
- 지역: 1) 서울 2) 경기 3) 경남 4) 경북 5) 충남 6) 강원 & 충북 7) 전라 & 제주
- 소득 : 연간 소득 (M KRW) (백만 원, 1100 KRW = 1 USD)



- family\_member: 가족 구성원 수
- 성별 : 1) 남자 2) 여자
- 태어난 해
- education\_level: 1) 교육 없음(7세 미만) 2) 교육 없음(7세 이상) 3) 초등학교 4) 중학교 5) 고등학교 6) 대학 7) 대학 학위 8) MA 9) 박사 학위
- 결혼: 결혼 상태. 1) 해당 없음(18세 미만) 2) 기혼 3) 사망으로 인한 별거 4) 별거 5) 아직 결혼하지 않음 6) 기타
- 종교: 1) 종교가 있다 2) 종교가 없다
- 직업: 이는 별도의 코드북에 제공됩니다.
- 회사 규모
- reason\_none\_worker: 1) 능력이 없음 2) 군 복무 중 3) 학교 재학 중 4) 학교 준비 5) 취업 준비 6) 가사 노동자 7) 집에서 아이 돌보기 8) 간호 9) 경제 활동 포기 10) 일할 의향 없음 11) 기타

## 21. 🏰 Top 100 KDrama 2023 (박창현)

<https://www.kaggle.com/datasets/gianinamariapetrascu/top-100-k-drama-2023>

Data Files(1)

top100\_kdrama.csv(100.26 kB)

Data Columns(15)

- ID
- Title
- Genre
- Tags
- Synopsis
- Rank
- Popularity
- Score
- Episodes
- Duration
- Watchers
- Start\_date
- End\_date
- Day\_aired
- Main Role

### Usecases

- Explore the **most popular genres** and determine whether there is a correlation between them and higher ratings.
- Analyze whether **the day on which a drama is aired** has an impact on its rating.
- Use the data to create a **recommendation system** for K-Dramas, providing viewers with personalized suggestions based on their preferences.
- **가장 인기 있는 장르를** 살펴보고 , 그 장르와 높은 시청률 사이에 상관관계가 있는지 확인하세요.
- **드라마가 방영되는 요일이** 시청률에 영향을 미치는지 분석해 보세요 .

- 데이터를 사용하여 K-드라마 **추천 시스템을** 구축하고 , 시청자의 선호도에 따라 개인화된 추천을 제공합니다.

## 22. bike seoul sharing(박창현)

<https://www.kaggle.com/datasets/willianoliveiragibin/bike-seoul-sharing/data>

Data Files(1)

**SeoulBikeData.csv(604.17 kB)**

Data Columns(14)

- Date
- Rented Bike Count
- Hour
- Temperature(℃)
- Humidity(%)
- Wind speed (m/s)
- Visibility (10m)
- Dew point temperature(℃)
- Solar Radiation (MJ/m2)
- Rainfall(mm)
- Snowfall (cm)
- Seasons
- Holiday
- Functioning Day

### About Dataset

my data set is based on bike sharing in Korea developing its mobility movement in all areas of the country where it is possible to walk and all the people who can and should ride share bikes in the open because there are proposals that the companies themselves deliver to the people

내 데이터 세트는 도보로 이동 가능한 모든 지역과 자전거를 타야 하는 모든 사람들이 야외에서 자전거를 공유하도록 하는 한국의 자전거 공유에 기반을 두고 있습니다. 회사 자체에서 사람들에게 제안을 전달하기 때문입니다.

## 23. Bestsellers Unveiled\_Global Top Selling Books (박소연님)

<https://www.kaggle.com/datasets/marianadeem755/bestsellers-unveiled-global-top-selling-books/code>

Data Files(1)

**best\_selling\_books\_2.csv(9.7 kB)**

Data Columns(6)

1. Rank: This column indicates the position of each Top sellingbook relative to the other Books.
2. Title: It contains the Lists of the names of those books that have achieved significant sales.
3. Author: It Provides the names of the authors who wrote the best-selling books.
4. Volume Sales: It Represents the total number of copies sold globally of each book.
5. Publisher: It Indicates the entity responsible for publishing each book.
6. Genre: It Categorizes each book into a specific literary genre or type.
7. 순위: 이 열은 가장 많이 팔린 각 책의 다른 책과 비교한 순위를 나타냅니다.
8. 제목: 상당한 판매를 달성한 책 이름의 목록이 수록되어 있습니다.
9. 저자: 베스트셀러를 쓴 저자의 이름을 제공합니다.
10. 판매량: 각 책이 전 세계적으로 판매된 총 사본 수를 나타냅니다.
11. 출판사: 각 책을 출판하는 책임이 있는 기관을 나타냅니다.
12. 장르: 각 책을 특정한 문학 장르나 유형으로 분류합니다.

## Useases

- Market Insights: This Dataset can be used by the analysts to identify trends in book sales across different genres over time.
- Author Analysis: By Analyzing this Dataset the researchers and enthusiasts can study the success of authors and their impact on the literary market.
- Publisher Trends: By taking Deep Insights about the Dataset we can know that the publishing houses can analyze their competitors and market performance based on the success of their published books.

- Consumer Behavior: By analyzing this Dataset we can understand which genres and authors are most popular can and provide insights into reader preferences and consumption patterns.
- 시장 통찰력: 이 데이터 세트는 분석가가 시간 경과에 따라 다양한 장르의 도서 판매 추세를 파악하는 데 사용할 수 있습니다.
- 작가 분석: 이 데이터 세트를 분석함으로써 연구자와 애호가들은 작가들의 성공과 문학 시장에 미치는 영향을 연구할 수 있습니다.
- 출판사 동향: 데이터 세트에 대한 심층적인 통찰력을 통해 출판사에서 출판한 책의 성공에 따라 경쟁사와 시장 성과를 분석할 수 있다는 것을 알 수 있습니다.
- 소비자 행동: 이 데이터 세트를 분석하면 어떤 장르와 작가가 가장 인기 있는지 파악할 수 있으며, 독자의 선호도와 소비 패턴에 대한 통찰력을 제공할 수 있습니다.

## 24. Early Classification of Diabetes(함현지님)

<https://www.kaggle.com/datasets/andrewmvd/early-diabetes-classification>

Data Files(1)

**diabetes\_data.csv(20.86 kB)**

Data Columns(17)

- age
- gender
- polyuria
- polydipsia
- sudden\_weight\_loss
- weakness
- polyphagia
- genital\_thrush
- visual\_blurring
- itching
- irritability
- delayed\_healing
- partial\_paresis
- muscle\_stiffness
- alopecia
- obesity
- class

### Usecases

- Create a classification model to predict diabetes;
- Explore the most common features associated with diabetic risk.
- 당뇨병을 예측하기 위한 분류 모델을 만듭니다.
- 당뇨병 위험과 관련된 가장 일반적인 특징을 살펴보세요.

## 25. Motor Vehicle Collisions - Crashes (함현지님)

<https://www.kaggle.com/datasets/kirbysasuke/crashes>

Data Files(1)

**MotorVehicle\_Collisions-\_Crashes.csv(444.98 MB)**

Data Columns(29)

- CRASH DATE
- CRASH TIME
- BOROUGH
- ZIP CODE
- LATITUDE
- LONGITUDE
- LOCATION
- ON STREET NAME
- CROSS STREET NAME
- OFF STREET NAME
- NUMBER OF PERSONS INJURED
- NUMBER OF PERSONS KILLED
- NUMBER OF PEDESTRIANS INJURED
- NUMBER OF PEDESTRIANS KILLED
- NUMBER OF CYCLIST INJURED
- NUMBER OF CYCLIST KILLED
- NUMBER OF MOTORIST INJURED
- NUMBER OF MOTORIST KILLED
- CONTRIBUTING FACTOR VEHICLE 1
- CONTRIBUTING FACTOR VEHICLE 2
- CONTRIBUTING FACTOR VEHICLE 3
- CONTRIBUTING FACTOR VEHICLE 4
- CONTRIBUTING FACTOR VEHICLE 5
- COLLISION\_ID
- VEHICLE TYPE CODE 1
- VEHICLE TYPE CODE 2



- VEHICLE TYPE CODE 3
- VEHICLE TYPE CODE 4
- VEHICLE TYPE CODE 5

## 26. 국민유형별걷기분석정보(박소연님)

[https://www.bigdata-culture.kr/bigdata/user/data\\_market/detail.do?id=10eb0f80-2594-11eb-af9a-4b03f0a582d6](https://www.bigdata-culture.kr/bigdata/user/data_market/detail.do?id=10eb0f80-2594-11eb-af9a-4b03f0a582d6)

Data Files(1)

국민유형별걷기분석정보(202303).csv (10.82KB)

Data Columns(6)

MESURE\_AGRDE\_FLAG\_NM

MESURE\_AGRDE\_FLAG\_NM

WEEK\_ODR

DALY\_ODR

MESURE\_NMPR\_CO

AVRG\_PACE\_CO

측정연령대구분명

성별구분코드

주차수

일별차수

측정인원수

평균걸음수

예시

<https://www.kaggle.com/code/eddieyun/bellabeat-case-study-20220808>