**Data Science**

**Team 9** - 구준서 김상준 박선인 이하은 한웅재

# 1. 데이터셋 개요

<https://www.kaggle.com/datasets/fahadrehman07/data-science-jobs-and-salary-glassdoor/>

미국 구인공고 웹사이트 Glassdoor에서 추출한 데이터과학 직업 관련 데이터입니다.

956행 x 15열로 구성되어 있으며, 직책, 예상 급여, 직무 설명, 회사 평점, 그리고 위치, 규모, 산업 등 주요 회사 정보 등 구직 관련 필수적인 정보가 포함되어 있습니다.

|  |  |
| --- | --- |
| Feature | Description |
| Job Title | 직책명 |
| Salary Estimate | 회사가 제공하는 직무에 대한 예상 급여 |
| Job Description | 직무 설명 |
| Rating | 회사 평가 |
| Company Name | 회사명 |
| Location | 작업 위치 |
| Headquarters | 회사 본사 |
| Size | 회사의 직원 수 |
| Founded | 회사가 설립된 연도 |
| Type of ownership | 민간, 공공, 정부 및 비영리 조직과 같은 소유권 유형 |
| Industry | Aerospace, Energy 등 회사가 서비스를 제공하는 산업 유형 |
| Sector | 산업(에너지), 부문(석유, 가스)와 같이 어떤 유형의 서비스를 제공하는지 |
| Revenue | 회사의 총 수익 |
| Competitors | 회사 경쟁사 |

# 2. 데이터 타입 분포

|  |  |
| --- | --- |
| **데이터 타입** | **개수** |
| object | 12 |
| int64 | 2 |
| float64 | 1 |

# 3. 결측치 현황

명시된 null은 없지만, 수치형 변수에 대해 '-1' 값으로 결측을 표현한 더티 데이터가 존재합니다.

|  |  |  |
| --- | --- | --- |
| **컬럼명** | **결측값 형태** | **결측 수** |
| Rating | -1 | 34건 |
| Founded | -1 | 97건 |
| Competitors | -1 | 634건 |

전처리 실습을 위해, 인위적인 drop을 통해 결측값을 추가로 확보하겠습니다.

# 4. 수치형 변수 통계 요약

|  |  |  |
| --- | --- | --- |
| **통계항목** | **Rating** | **Founded** |
| count | 956 | 956 |
| mean | 3.60 | 1774.61 |
| std | 1.07 | 598.94 |
| min | -1.0 | -1 |
| 25% | 3.30 | 1937 |
| 50% | 3.80 | 1992 |
| 75% | 4.20 | 2008 |
| max | 5.00 | 2019 |

# 5. 고유값(unique value) 개수

|  |  |
| --- | --- |
| **컬럼명** | **고유값 수** |
| Job Title | 53 |
| Salary Estimate | 92 |
| Company Name | 343 |
| Location | 200 |
| Headquarters | 198 |
| Size | 9 |
| Type of ownership | 14 |
| Industry | 60 |
| Sector | 25 |
| Revenue | 14 |
| Competitors | 147 |

텍스트, 스크린샷, 디스플레이, 도표이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

# 6. 고유값 예시 (상위 5개)

## Job Title

|  |  |
| --- | --- |
| **Job Title** | **건수** |
| Data Scientist | 178 |
| Data Engineer | 61 |
| Data Analyst | 60 |
| Senior Data Scientist | 48 |
| Machine Learning Engineer | 20 |

텍스트, 스크린샷, 소프트웨어, 디스플레이이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

## Salary Estimate

|  |  |
| --- | --- |
| **Salary Estimate** | **건수** |
| $80K-$90K | 18 |
| $90K-$120K | 17 |
| $100K-$130K | 16 |
| $110K-$140K | 15 |
| $70K-$90K | 14 |

## Industry

|  |  |
| --- | --- |
| **Industry** | **건수** |
| IT Services | 133 |
| Health Care Services & Hospitals | 83 |
| Colleges & Universities | 51 |
| Insurance Carriers | 36 |
| Biotech & Pharmaceuticals | 35 |

위와 같이, 범주형 데이터들의 경우 차원의 수가 매우 높고, 정제되지 않은 데이터의 형태를 띄고 있는 것을 확인했습니다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

이를 해결하기 위해, Salary Estimate의 경우 구간 평균 값을 활용해 수치형으로 전환하거나, Job Description의 경우 대표 키워드를 통해 재분류하는 등, 다양한 전처리 기법들을 적용할 계획입니다.

# 7. 프로젝트 목표

산업별 또는 평점별 분포 분석, 급여 정규화 및 클러스터링, 회귀모델 등의 후속 분석을 통해 다음과 같은 결과를 얻고자 합니다.

1. 데이터과학 직업군의 급여예측 모델 구현
2. 기업 특성에 따른 직무 환경 분석
3. 기업의 규모, 업종, 소유 형태, 위치 등과 급여의 상관관계 분석
4. 구직자관점에서의 적합한 회사유형 파악, 급여 기대 수준의 현실화