**Data Cleaning with Pandas**

# **Dataset Overview**

The data was scrapped from Glassdoor's website. There are two versions of the data one is uncleaned and another one is cleaned.

# **Uncleaned Data Description**

|  |  |  |  |
| --- | --- | --- | --- |
| Column Name | Data Type | Data Type Description | Potential Problems/Considerations |
| index | int | Nominal | None |
| Job Title | object | Nominal | None |
| Salary Estimate | object | Ordinal | Inconsistent format, may need data cleaning and conversion to numerical values |
| Job Description | object | Nominal | None |
| Rating | float | Ordinal | Missing values, potential data quality issues |
| Company Name | object | Nominal | None |
| Location | object | Nominal | None |
| Headquarters | object | Nominal | None |
| Size | object | Ratio | Inconsistent format, may need data cleaning and conversion to numerical values |
| Founded | object | Ratio | Missing values, potential data quality issues, may need data cleaning and conversion to numerical values |
| Type of ownership | object | Nominal | None |
| Industry | object | Nominal | None |
| Sector | object | Nominal | None |
| Revenue | object | Ordinal | Inconsistent format, may need data cleaning and conversion to numerical values |
| Competitors | object | Nominal | None |

# **Cleaned Data Description**

In the cleaned version the columns explanation are as follows,

* **Job** **Title**: Title of the job posting
* **Salary** **Estimation**: Salary range for that particular job
* **Job** **Description**: This contains the full description of that job
* **Rating**: Rating of that post
* **Company**: Name of company
* **Location**: Location of the company
* **Headquarter**: Location of the headquater
* **Size**: Total employee in that company
* **Type of ownership**: Describes the company type i.e non-profit/public/private farm etc
* **Industry, Sector**: Field applicant will work in
* **Revenue**: Total revenue of the company
* **min\_salary,max\_salary,avg**\_**salary**: Refers to the minimum, maximum and average salary for that post
* **job\_state**: State where the applicant will work
* **same\_state**: Same state as headquarter or not(Boolean)
* **company\_age**: Age of company
* **python,excel,hadoop,spark,aws,tableau,big\_data:** Some most appeared skills in boolean columns form
* **job\_simp**: Job type
* **seniority**: if job type is senior or not (Boolean)

|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | Data Type Description |
| Job Title | String | Nominal |
| Salary Estimate | String | Ordinal |
| Job Description | String | Nominal |
| Rating | Float | Ordinal |
| Company Name | String | Nominal |
| Location | String | Nominal |
| Headquarters | String | Nominal |
| Size | String | Ratio |
| Type of ownership | String | Nominal |
| Industry | String | Nominal |
| Sector | String | Nominal |
| Revenue | String | Ordinal |
| min\_salary | Integer | Continuous |
| max\_salary | Integer | Continuous |
| avg\_salary | Float | Continuous |
| job\_state | String | Nominal |
| same\_state | Integer | Binary |
| company\_age | Integer | Continuous |
| python | Integer | Binary |
| excel | Integer | Binary |
| hadoop | Integer | Binary |
| spark | Integer | Binary |
| aws | Integer | Binary |
| tableau | Integer | Binary |
| big\_data | Integer | Binary |
| Seniority | String | Ordinal |
| job\_simp | String | Nominal |

# **Dataset Cleaning Process**

Documented in the Notebook [here](Assignment.ipynb).

Done by:

1. Ahmed Maher Ali Ahmed Al-Maqtari
2. Ali Ahmed Saleh Ahmed Yousef