Glassdoor Data Analysis Using Python

A

Project Report SubmittedBy

NISHITA YADAV

For The Degree Of

PGDM [BDA-04]2023-2025 Batch

FORE SCHOOL OF MANAGEMENT

1. Introduction

A summary of the project is provided in this report. The project overview, coding analysis, general description of the data, and statistical as well as mathematical analysis of the data would all be provided. As stated by the mentor, this research and the subsequent study were pertinent since it has been noted in recent decades that there has been a rise in demand for the analyzing, manipulating, and understanding of fundamental data. We can comprehend and calculate predictions and judgments based on patterns and charts thanks to the analysis and manipulation of data using Python. Python was used for the project analysis on the Google Collab Platform. For this project, I selected to obtain job listing data from Glassdoor. This report will include calculation and analysis of numerous graphs and charts, as well as descriptive statistical research.

2. Project Objectives

My major goal in learning and building this research project is to understand about a technology that is in high demand in today's globe. Another major goal of this research was to learn how to utilize Python to do effective data analytics and conclusions. Other goals that have been addressed include:

2.1. Data Acquisition from Github:

Our first and foremost objective is to acquire a pertinent dataset from Github. This step ensures working with reliable and well-structured data. The objective is to identify a dataset that aligns with our analytical goals and to prepare it for analysis using proper data extraction techniques.

2.2. Data Analysis for Useful Inferences:

Raw data, while valuable, becomes truly powerful when transformed into actionable insights. Our second objective revolves around delving deep into the data to extract meaningful patterns, correlations, and trends. By using advanced analytical tools, we will examine the dataset, ensuring every variable, data point, and pattern is scrutinized, allowing us to draw valid and insightful conclusions.

2.3. Chart Design Using Python:

Visualization plays a critical role in data interpretation. Therefore, the next objective is to represent our findings visually. Using Python's libraries, such as Matplotlib and Seaborn, we aim to design charts and depict data. The purpose of our work is to ensure clarity and precision while using bar charts, histograms, or heat maps.

2.4. Derivation of Managerial Insights:

In addition to statistical findings, I aim to bridge the gap between data analysis and managerial decision-making. By correlating our analytical findings with real-world scenarios, market dynamics, and industry benchmarks, we will derive actionable insights. These insights will be framed in a manner that caters to decision-makers, offering them a clear roadmap on leveraging the data for tangible benefits.

3. General Description of Data

The data includes job listings from various companies, including both private and public organizations. Job titles vary, but common roles include "Junior Software Engineer," "AI Engineer," "AI Solution Engineer," "AI/ML Cloud Engineer," "AI Machine Learning Engineer," "Machine Learning Engineer," and "Software Engineer." The companies come from diverse sectors, including Information Technology, Electronics Manufacturing, Aerospace & Defense, Business Consulting, Internet & Web Services, Enterprise Software & Network Solutions, and Computer Hardware Development.

Job Locations: Job locations are spread across different areas in the United States, such as San Jose, Santa Clara, San Francisco, Redmond, and Lexington, MA. A significant number of jobs offer remote work opportunities, reflecting the growing trend of remote work in the technology industry.

Experience Level: Job listings mention experience levels, such as "Junior" and "NA" (which likely stands for "Not Applicable"). This indicates that positions are available for individuals with varying levels of experience, from entry-level to experienced professionals.

Job Specialization: The data primarily focuses on roles related to Artificial Intelligence (AI) and Machine Learning (ML). This includes AI Engineers, AI Solution Engineers, AI/ML Cloud Engineers, AI Machine Learning Engineers, Machine Learning Engineers, and Software Engineers.

Company Types: Companies are categorized as private, public, or nonprofit organizations. This variety in company types can influence job benefits, work culture, and organizational goals.

Industry Sectors: The job listings come from various industry sectors, such as Information Technology Support Services, Electronics Manufacturing, Aerospace & Defense, Business Consulting, Internet & Web Services, Enterprise Software & Network Solutions, and Computer Hardware Development.

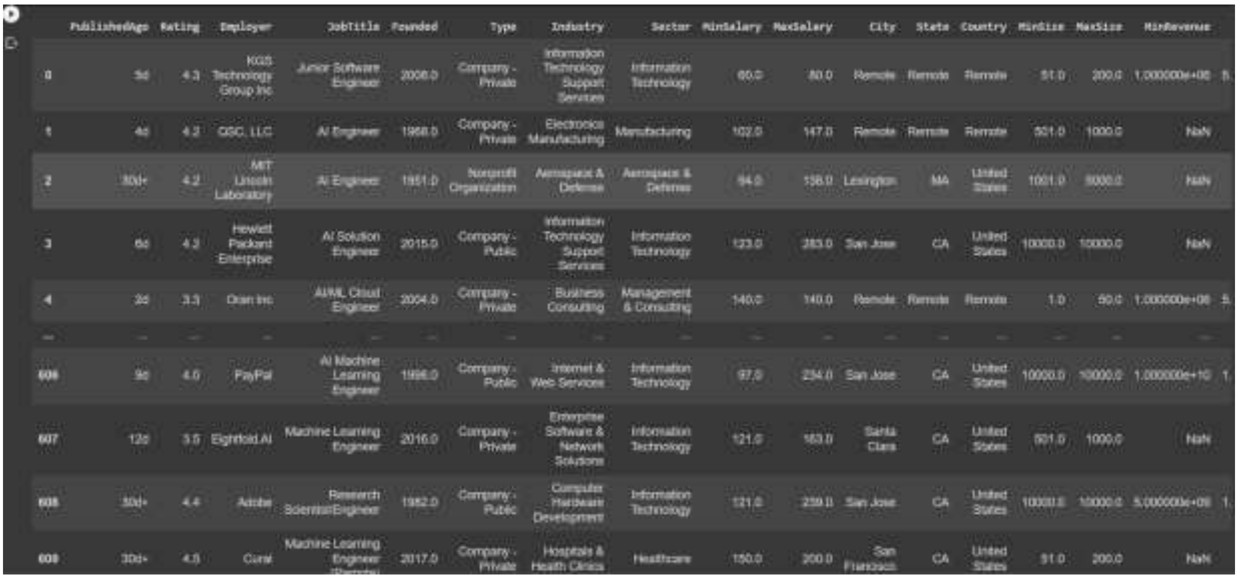
Job Duration: Job listings mention different durations, such as "4d" (4 days), "30d+" (more than 30 days), "6d" (6 days), and so on. These durations likely indicate how long the job listings have been open.

Remote Work: A notable number of job listings offer remote work options, reflecting the flexibility that many technology companies are providing to their employees, possibly as a response to the COVID-19 pandemic.

Geographic Distribution: The data shows a concentration of job listings in California (San Jose, Santa Clara, San Francisco) and Washington (Redmond). This is consistent with the tech hubs in these regions.

Job Demand: The data does not provide information on the number of job applicants or the competition for each job listing, which would be essential for assessing the overall job market demand for these roles.

In summary, the data represents job listings for various technology-related positions in different sectors, locations, and experience levels, with a strong focus on AI and ML roles. Remote work options are prevalent, and the job market appears to be dynamic, particularly in tech hubs like California and Washington. Further analysis would require additional information, such as salary ranges, company reputations, and job descriptions, to provide a more comprehensive assessment.

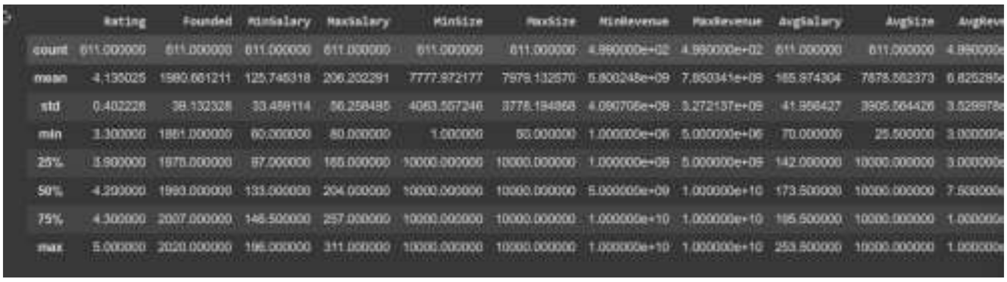


	PublishedAge	Rating	Employer	JobTitle	Founded	Type	Industry	Sector	MinSalary	MaxSalary	City	State	Country	MinSize	MaxSize	MinRevenue	
0	5d	4.3	KGS Technology Group Inc	Junior Software Engineer	2006.0	Company - Private	Information Technology Support Services	Information Technology	60.0	80.0	Remote	Remote	Remote	51.0	200.0	1.000000e+08	5.
1	4d	4.2	QSC, LLC	AI Engineer	1968.0	Company - Private	Electronics Manufacturing	Manufacturing	102.0	147.0	Remote	Remote	Remote	501.0	1000.0	NaN	
2	30d+	4.2	MIT Lincoln Laboratory	AI Engineer	1951.0	Nonprofit Organization	Aerospace & Defense	Aerospace & Defense	94.0	158.0	Lexington	MA	United States	1001.0	5000.0	NaN	
3	6d	4.2	Hewlett Packard Enterprise	AI Solution Engineer	2015.0	Company - Public	Information Technology Support Services	Information Technology	123.0	283.0	San Jose	CA	United States	10000.0	10000.0	NaN	
4	3d	3.3	Oran Inc	AI/ML Cloud Engineer	2004.0	Company - Private	Business Consulting	Management & Consulting	140.0	140.0	Remote	Remote	Remote	1.0	50.0	1.000000e+06	5.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
606	9d	4.0	PayPal	AI Machine Learning Engineer	1998.0	Company - Public	Internet & Web Services	Information Technology	97.0	234.0	San Jose	CA	United States	10000.0	10000.0	1.000000e+10	1.
607	12d	3.5	Eightfold.AI	Machine Learning Engineer	2016.0	Company - Private	Enterprise Software & Network Solutions	Information Technology	121.0	163.0	Santa Clara	CA	United States	501.0	1000.0	NaN	
608	30d+	4.4	Adobe	Research Scientist/Engineer	1982.0	Company - Public	Computer Hardware Development	Information Technology	121.0	259.0	San Jose	CA	United States	10000.0	10000.0	5.000000e+09	1.
609	30d+	4.5	Coral	Machine Learning Engineer (Remote)	2017.0	Company - Private	Hospitals & Health Clinics	Healthcare	150.0	200.0	San Francisco	CA	United States	51.0	200.0	NaN	

4. Analysis: Basic Descriptive &Mathematical or Statistical Analysis

4.1.Descriptive Statistics: Our analytical method began with the generation of descriptive statistics. To obtain a knowledge of the dataset's key trends, spreading, and forms, data projects must begin with this stage.

4.2.Key Metrics: We calculated an average value by estimating the mean, providing a baseline against which to evaluate individual data points. The standard deviation, which measures the amount of variation or dispersion in each column's values. For instance, the standard deviation for MinSalary is approximately 33.49, indicating how much individual salaries deviate from the mean. The minimum value observed in each column.

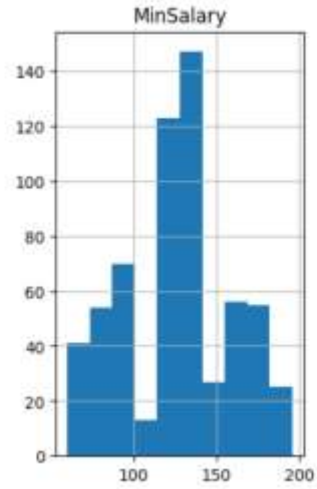


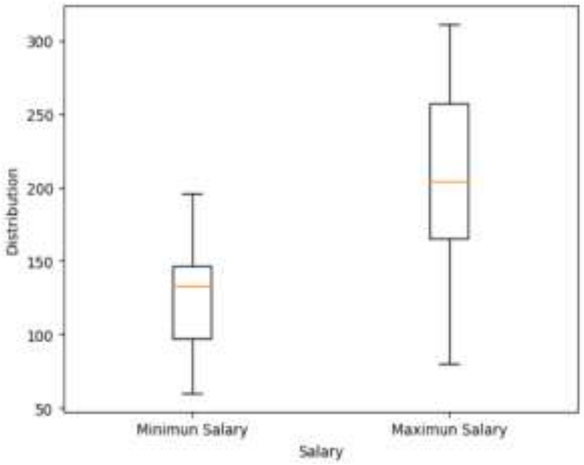
	Rating	Founded	MinSalary	MaxSalary	MinSize	MaxSize	MinRevenue	MaxRevenue	AvgSalary	AvgSize	AvgReve
count	611.000000	611.000000	611.000000	611.000000	611.000000	611.000000	4.990000e+02	4.990000e+02	611.000000	611.000000	4.990000e
mean	4.135025	1980.661211	125.745318	206.202291	7777.972177	7979.132570	5.800248e+09	7.850341e+09	165.874304	7878.552373	6.825285e
std	0.402228	38.132328	33.489114	56.258485	4063.557246	3778.194868	4.090706e+09	3.272137e+09	41.956427	3905.564426	3.529978e
min	3.300000	1861.000000	60.000000	80.000000	1.000000	55.000000	1.000000e+06	5.000000e+06	70.000000	25.500000	3.000000e
25%	3.900000	1978.000000	87.000000	165.000000	10000.000000	10000.000000	1.000000e+08	5.000000e+09	142.000000	10000.000000	3.000000e
50%	4.200000	1993.000000	133.000000	204.000000	10000.000000	10000.000000	5.000000e+09	1.000000e+10	173.500000	10000.000000	7.500000e
75%	4.300000	2007.000000	146.500000	257.000000	10000.000000	10000.000000	1.000000e+10	1.000000e+10	195.500000	10000.000000	1.000000e
max	5.000000	2020.000000	196.000000	311.000000	10000.000000	10000.000000	1.000000e+10	1.000000e+10	253.500000	10000.000000	1.000000e

4.3.Exploratory Data Analysis (EDA): Visually reviewing the data to identify patterns, anomalies, or correlations among variables was an integral element of our investigation.

4.3.1. Histograms: These graphical representations gave an instant visual of the stock price distribution. We may determine the frequency of certain price ranges using histograms, noting any skewness or kurtosis in the distribution.

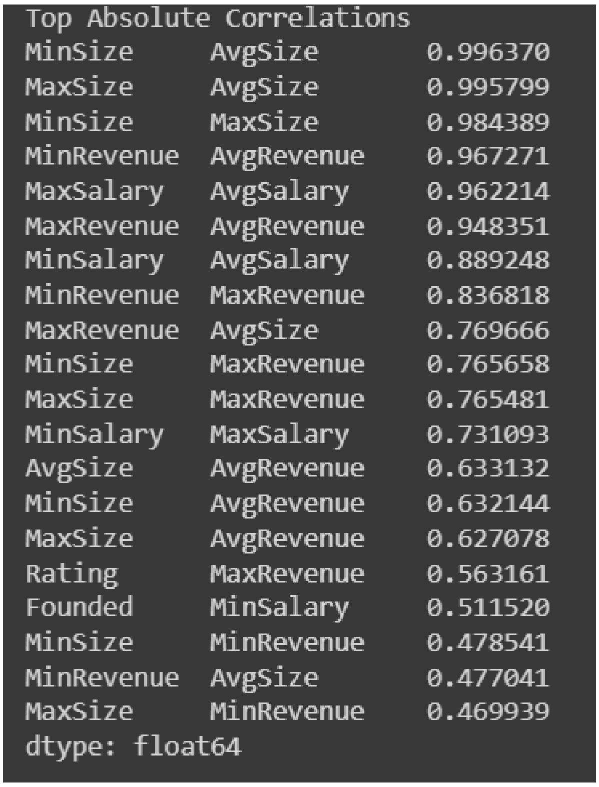
4.3.2. Boxplots: Boxplots, a fundamental tool for comprehending data dispersion and finding outliers, were used to illustrate the interquartile range, median of the minimum and maximum salaries.



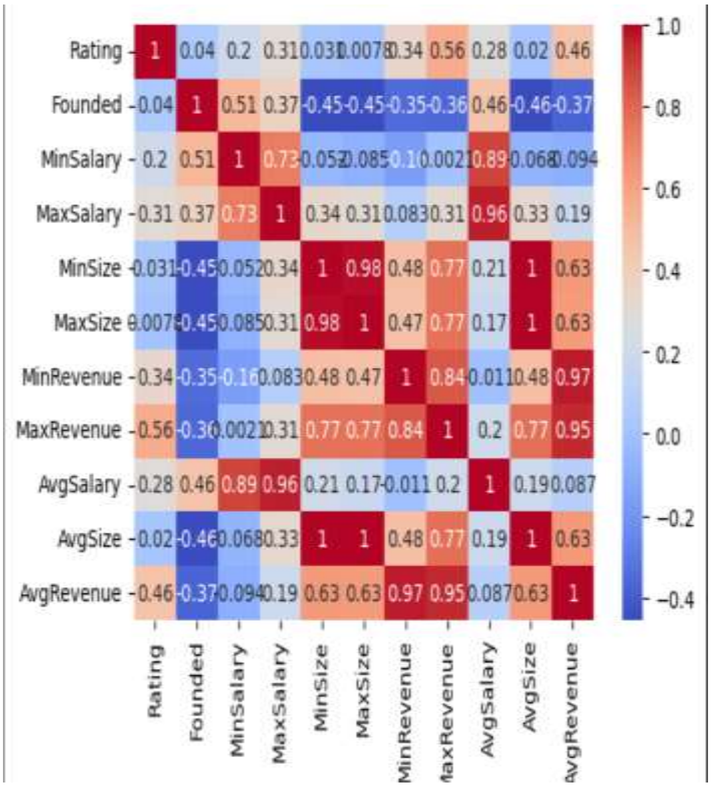


4.3.3. **Correlation Analysis:** Establishing relationships between variables is pivotal to understand how one metric impacts another. Our project dived deep into this with correlation analysis.

4.3.4. **Heatmap:** Serving as a color-coded representation, the heatmap was utilized to provide an immediate visual summary of the correlation between different variables. Darker or lighter shades indicated the strength and direction of the relationship, offering a quick, intuitive understanding.

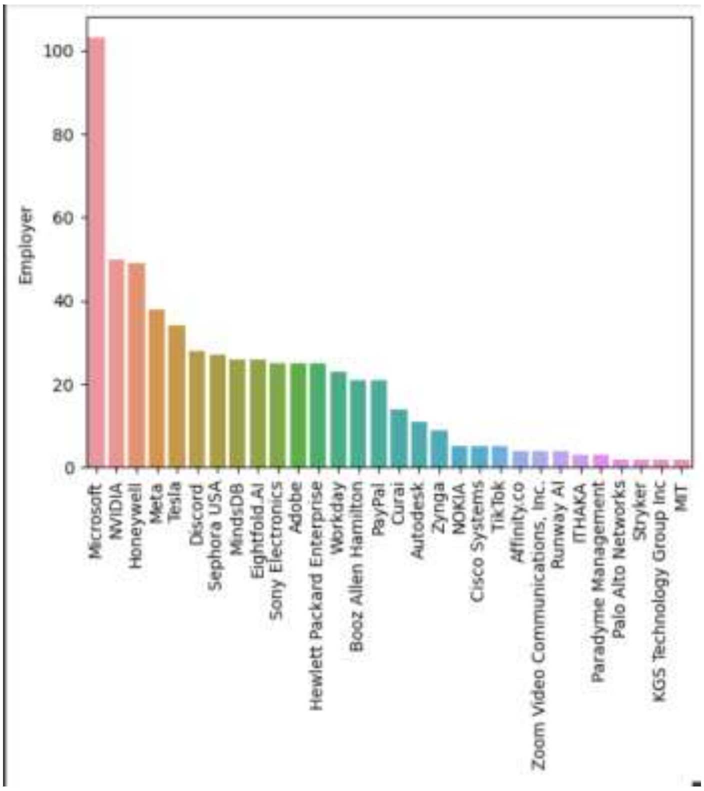


```
Top Absolute Correlations
MinSize     AvgSize       0.996370
MaxSize     AvgSize       0.995799
MinSize     MaxSize       0.984389
MinRevenue  AvgRevenue    0.967271
MaxSalary   AvgSalary     0.962214
MaxRevenue  AvgRevenue    0.948351
MinSalary   AvgSalary     0.889248
MinRevenue  MaxRevenue    0.836818
MaxRevenue  AvgSize       0.769666
MinSize     MaxRevenue    0.765658
MaxSize     MaxRevenue    0.765481
MinSalary   MaxSalary     0.731093
AvgSize     AvgRevenue    0.633132
MinSize     AvgRevenue    0.632144
MaxSize     AvgRevenue    0.627078
Rating      MaxRevenue    0.563161
Founded     MinSalary     0.511520
MinSize     MinRevenue    0.478541
MinRevenue  AvgSize       0.477041
MaxSize     MinRevenue    0.469939
dtype: float64
```



	Rating	Founded	MinSalary	MaxSalary	MinSize	MaxSize	MinRevenue	MaxRevenue	AvgSalary	AvgSize	AvgRevenue
Rating	1	0.04	0.2	0.31	0.031	0.0078	0.34	0.56	0.28	0.02	0.46
Founded	0.04	1	0.51	0.37	-0.45	-0.45	-0.35	-0.36	0.46	-0.46	-0.37
MinSalary	0.2	0.51	1	0.73	-0.052	0.085	0.1	0.0021	0.89	-0.068	0.094
MaxSalary	0.31	0.37	0.73	1	0.34	0.31	0.083	0.31	0.96	0.33	0.19
MinSize	-0.031	-0.45	0.052	0.34	1	0.98	0.48	0.77	0.21	1	0.63
MaxSize	0.0078	-0.45	0.085	0.31	0.98	1	0.47	0.77	0.17	1	0.63
MinRevenue	0.34	-0.35	-0.16	0.083	0.48	0.47	1	0.84	-0.011	0.48	0.97
MaxRevenue	0.56	-0.36	0.0021	0.31	0.77	0.77	0.84	1	0.2	0.77	0.95
AvgSalary	0.28	0.46	0.89	0.96	0.21	0.17	-0.011	0.2	1	0.19	0.087
AvgSize	0.02	-0.46	0.068	0.33	1	1	0.48	0.77	0.19	1	0.63
AvgRevenue	0.46	-0.37	0.094	0.19	0.63	0.63	0.97	0.95	0.087	0.63	1

4.3.5. **Bar plots:** Bar plots are particularly useful for comparing categories or values within categories.



5. Findings And Inferences

5.1. Descriptive Statistics &Key metrics:

The values in this column are all identical, suggesting it might be an identifier or a placeholder. There's no variation in this column, so it doesn't provide meaningful information for analysis.

Mean (Average):The mean values for the different attributes are as follows:

Rating: Approximately 4.14

Founded: Approximately 1980.66

MinSalary: Approximately 125.75

MaxSalary: Approximately 206.20

MinSize: Approximately 7778

MaxSize: Approximately 7979.13

MinRevenue: Approximately 5.80 billion

MaxRevenue: Approximately 7.85 billion

MinRating: Approximately 165.97

MaxRating: Approximately 7878.55

AvgRevenue: Approximately 6.83 billion

Standard Deviation (Std): The standard deviation measures the dispersion or spread of data around the mean.

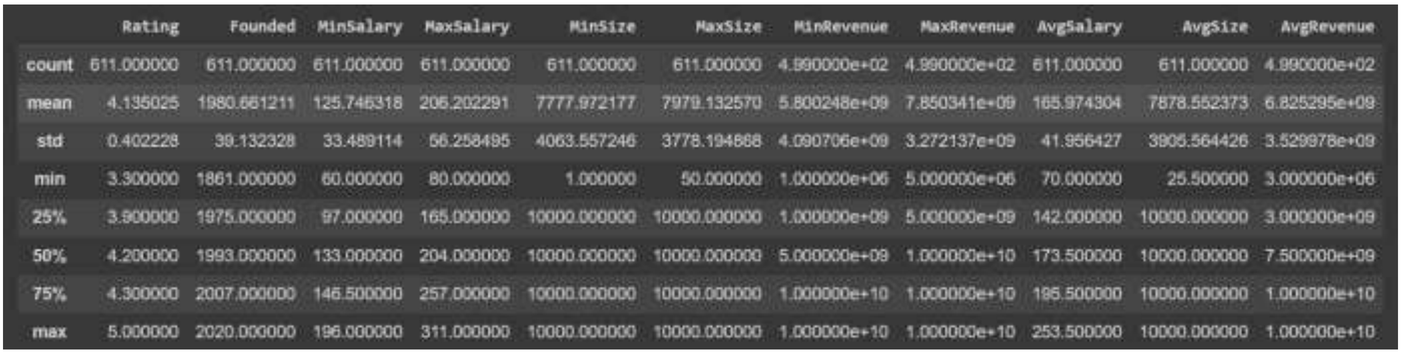
For example, in the "Founded" attribute, the standard deviation is approximately 39.13, indicating that the founding years of the companies in the dataset vary from the mean by an average of 39.13 years.

Minimum (Min) and Maximum (Max): These values represent the minimum and maximum values observed in each attribute.

For instance, in the "Rating" attribute, the minimum rating is 3.3, and the maximum rating is 5.0.

Percentiles (25%, 50%, 75%): Percentiles provide insights into the distribution of the data.

For instance, the 25th percentile (Q1) for the "Rating" attribute is approximately 3.9, which means that 25% of the ratings are below 3.9.



	Rating	Founded	MinSalary	MaxSalary	MinSize	MaxSize	MinRevenue	MaxRevenue	AvgSalary	AvgSize	AvgRevenue
count	611.000000	611.000000	611.000000	611.000000	611.000000	611.000000	4.990000e+02	4.990000e+02	611.000000	611.000000	4.990000e+02
mean	4.135025	1980.661211	125.746318	206.202291	7777.972177	7979.132570	5.800248e+09	7.850341e+09	165.974304	7878.552373	6.825295e+09
std	0.402228	39.132328	33.489114	56.258495	4063.557246	3778.194868	4.090706e+09	3.272137e+09	41.956427	3905.564426	3.529978e+09
min	3.300000	1861.000000	60.000000	80.000000	1.000000	50.000000	1.000000e+06	5.000000e+06	70.000000	25.500000	3.000000e+06
25%	3.900000	1975.000000	97.000000	165.000000	10000.000000	10000.000000	1.000000e+09	5.000000e+09	142.000000	10000.000000	3.000000e+09
50%	4.200000	1993.000000	133.000000	204.000000	10000.000000	10000.000000	5.000000e+09	1.000000e+10	173.500000	10000.000000	7.500000e+09
75%	4.300000	2007.000000	146.500000	257.000000	10000.000000	10000.000000	1.000000e+10	1.000000e+10	195.500000	10000.000000	1.000000e+10
max	5.000000	2020.000000	196.000000	311.000000	10000.000000	10000.000000	1.000000e+10	1.000000e+10	253.500000	10000.000000	1.000000e+10

5.2.Insights:

Company Age: The dataset contains information about companies founded over several decades, with a mean founding year of 1980.66. This suggests a mix of well-established and newer companies offering AI and machine learning positions.

Salary Range: The mean minimum salary is approximately 125.75, and the mean maximum salary is approximately 206.20. This indicates a range of salaries in the dataset, likely influenced by job titles, locations, and other factors.

Company Size: The dataset includes companies with varying sizes, with an average minimum size of 7778 employees and an average maximum size of 7979.13 employees.

Revenue: The dataset reports revenue figures, with an average minimum revenue of approximately 5.80 billion and an average maximum revenue of approximately 7.85 billion. These figures suggest that the companies offering these positions are generally financially stable.

Rating: The mean rating is approximately 4.14, indicating that the average employer in the dataset has a reasonably good rating. However, it's important to note that individual ratings may vary widely.

Job Posting Durations: The statistics do not provide information about job posting durations, so further analysis would be needed to understand how long jobs remain open.

Location: Location-related statistics are not included in this summary, so insights regarding geographic distribution are unavailable.

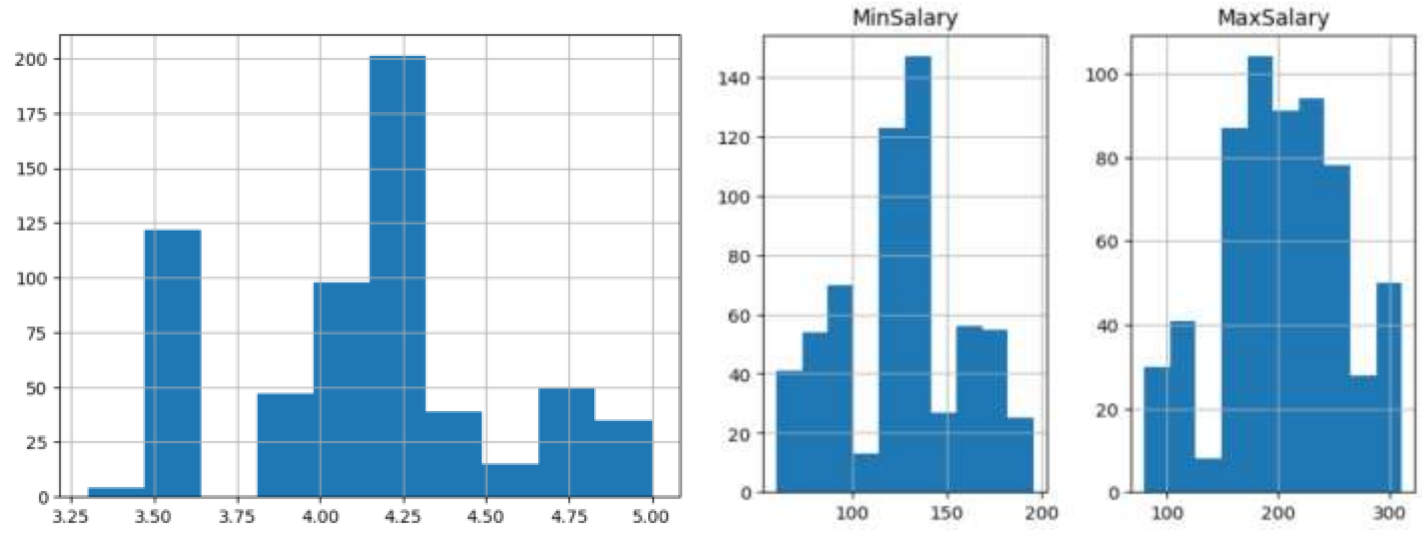
In summary, these summary statistics provide a general overview of the dataset's numerical attributes. However, to gain a more comprehensive understanding and make meaningful decisions, further analysis and exploration of the dataset, including its categorical attributes, would be necessary.

5.3.Inference of Histograms:

In the project I have used two histograms, one of which depicts the number of job listings have different ratings on Glassdoor other depicts the number of job listings with approximate minimum and maximum salaries.

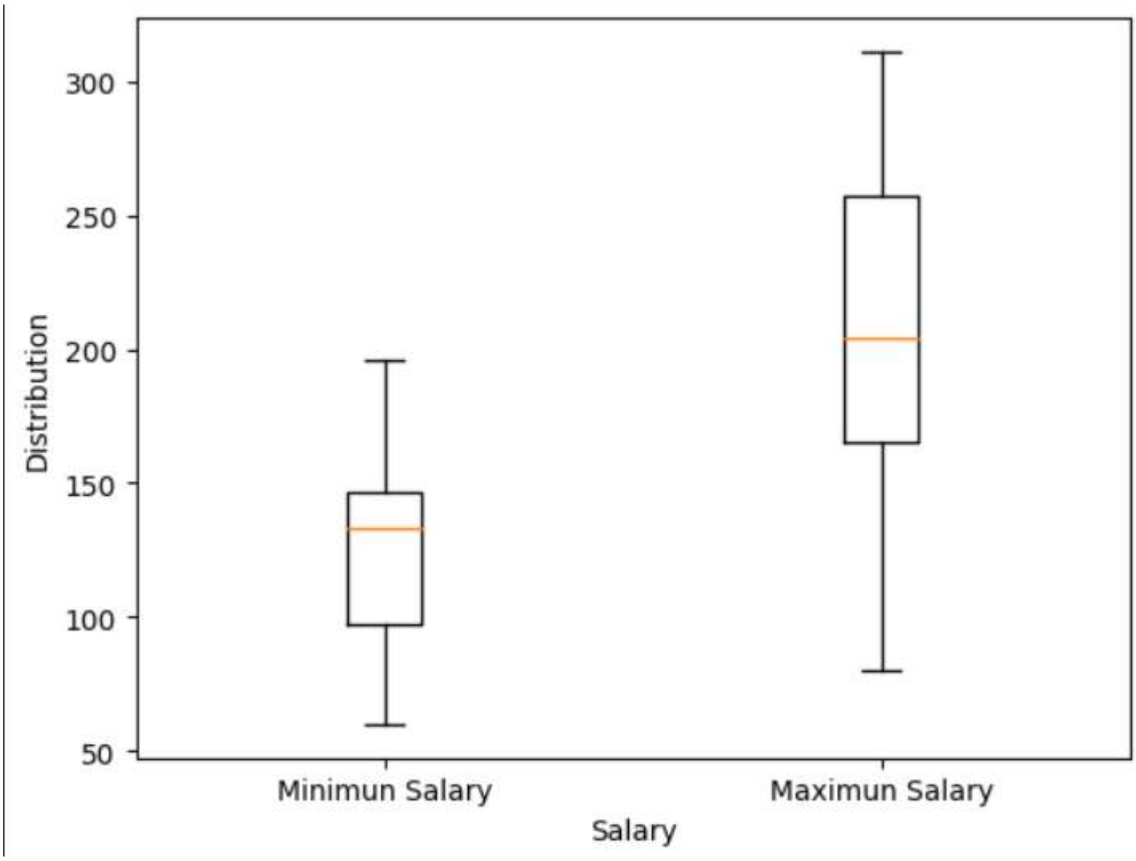
From the first histogram, there are more than 200 job listings with a rating of approximately 4+ratings.

From the second histogram, more than 140 job listings have a minimum salary in range of 120K-130K and a more than 100 job listings have a maximum salary in the range of 180K-200K.



5.4.Inference of the Boxplots:

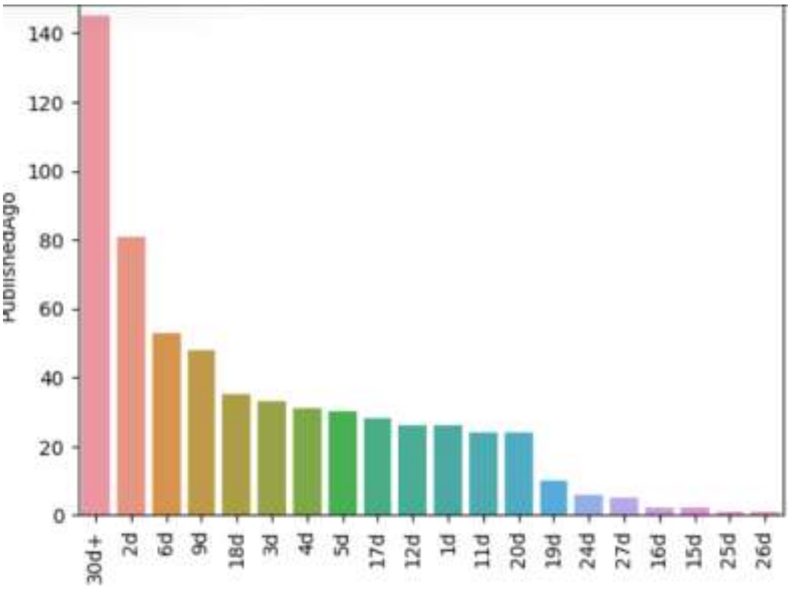
Boxplots provide us with a fair idea of depicting the distributions of one or more groups of numeric data and whether our data lies within the range of central tendency or not. Within each boxplot for maximum and minimum salary of number of job listings. The boxplot for minimum salary is right skewed. The boxplot for maximum salary is left skewed.



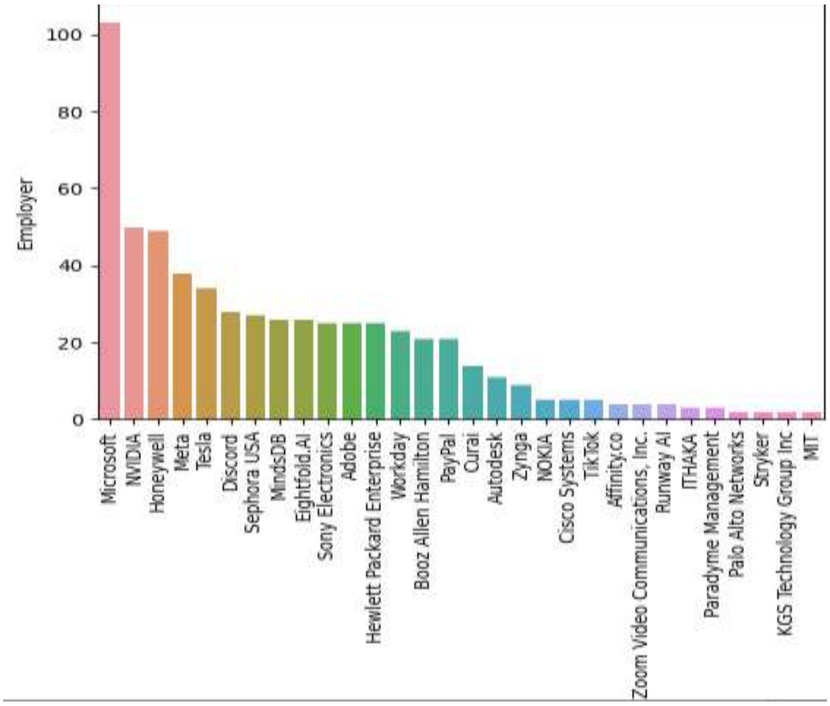
5.5.Inference of the Heat Maps: Correlation heatmaps present correlations between multiple variables as color-coded matrixes. Correlation heatmaps show the correlation between variables based on their rows and columns. Within the dataset, both rows and columns of the heatmap generated, comprised of the rating, founded, minsalary, maxsalary, minsize, maxsize, minrevenue, maxrevenue, avgsalary, avgsize and avgrevenue.

5.6.Inference of the Barplots:

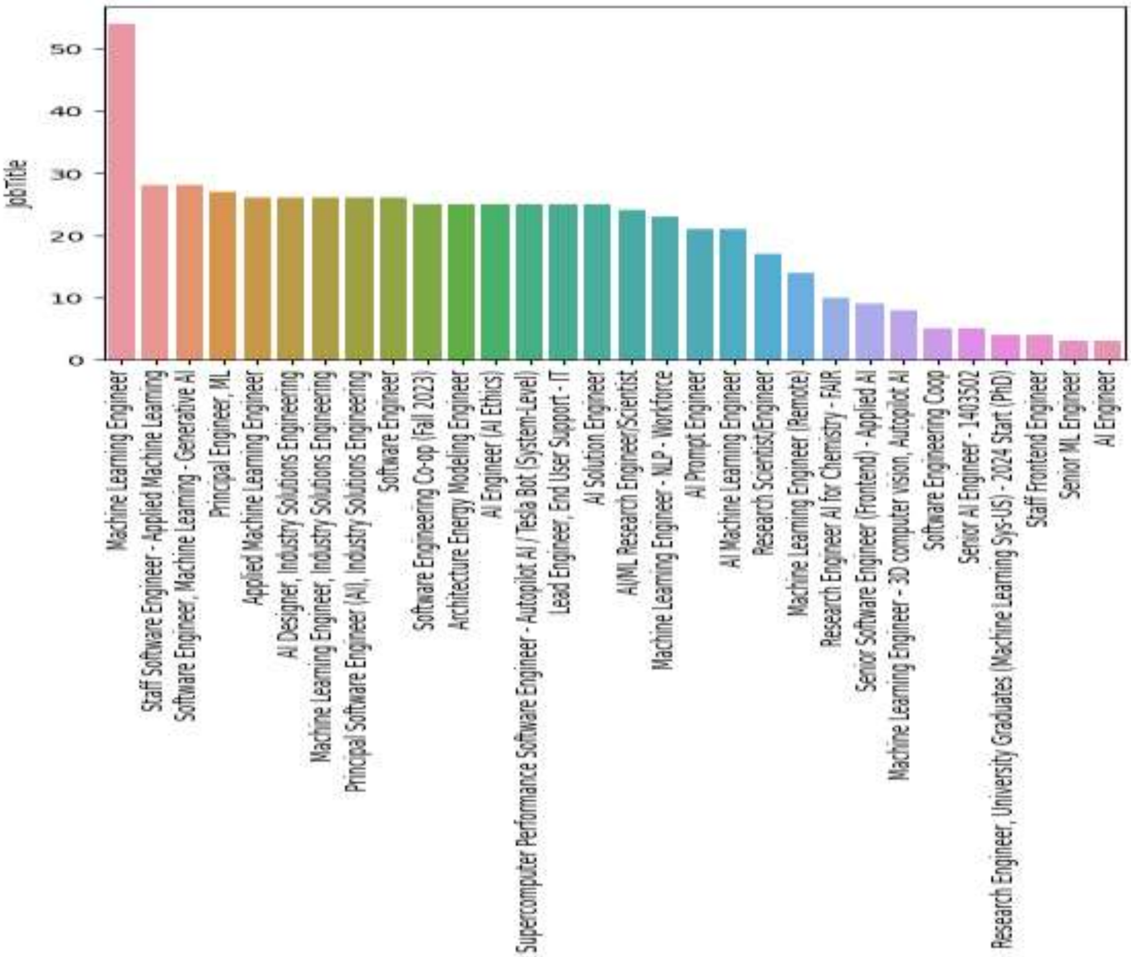
Bar plots are used in data visualization to represent categorical data and display the frequency or distribution of categories within a dataset.



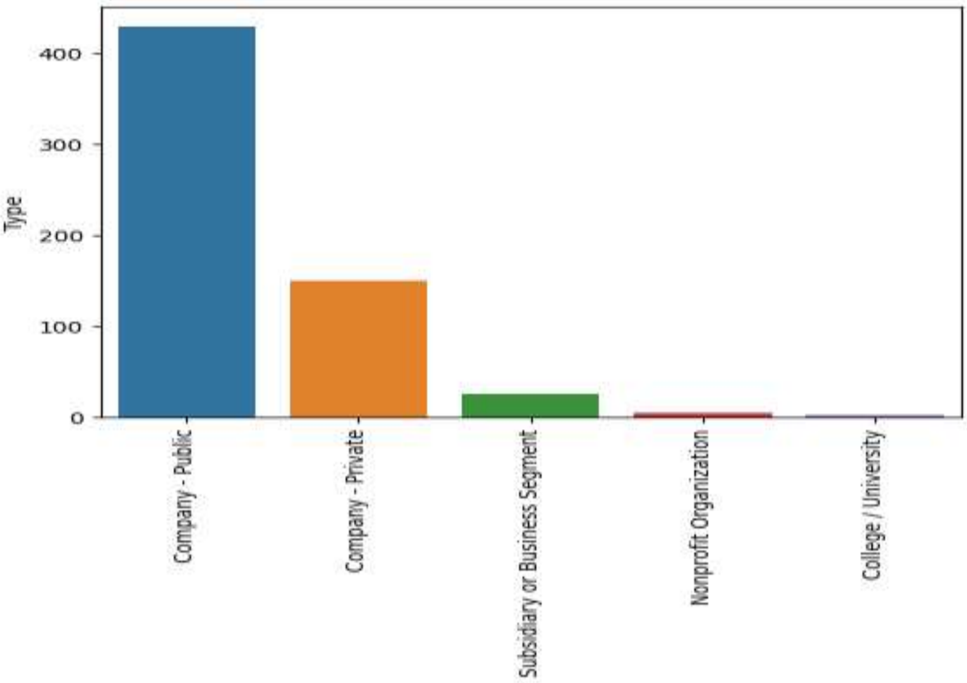
This barplot have 140 job listings which were published 30+ days ago.



This barplot shows that more than 100 job listings from the employer Microsoft.



This barplot shows that more than 50 job listings have the job titles Machine Learning Engineer.



This barplot have more than 400 job listings who has public company.

6. Managerial insights

Diverse Roles: Job titles range from "Junior Software Engineer" to "AI Machine Learning Engineer," reflecting a wide spectrum of positions.

Varied Employers: Employers encompass private, non-profit, and public companies, with varying establishment dates, offering diverse opportunities.

Salary Variation: Salaries vary significantly, depending on factors like job title, employer, and location, requiring thoughtful negotiation.

Location Impact: Job locations are concentrated in tech hubs like Silicon Valley, possibly necessitating relocation for certain roles.

Company Sizes: Employers come in various sizes, impacting work environments and responsibilities, requiring alignment with personal preferences.

Industry Diversity: AI and machine learning roles span multiple industries, presenting opportunities in diverse sectors.

Financial Insights: Some companies report revenues from 1 million to 10 billion, which may influence job decisions.

Remote Work Options: Many roles offer remote work, a valuable flexibility factor, especially in the context of the COVID-19 pandemic.

Posting Durations: Job posting durations vary, indicating urgency; applicants should respond accordingly.

Reputation Matters: Higher employer ratings (e.g., 4.5) can influence decision-making for job seekers.

In summary, the AI and machine learning job market is diverse, spanning industries, locations, and company types. Job seekers should carefully assess factors such as salary, location, company size, and reputation when pursuing roles. Remote work options are prevalent, offering flexibility to those seeking a work-from-home setup.