

Salaries in the Fields of Data Science, Artificial Intelligence, and Machine Learning

Abhishek Anand, Bolun Lu, Siming Xu, Rijin Lnu

2022-10-17

Abstract

[[Add summary of our report when analyses is complete.]]

Index Terms / Keywords

[[Complete list of keywords/acronyms that we reference.]]

AI: Artificial Intelligence DS: Data Science ML: Machine Learning USD: United States Dollar

Introduction

A shift towards data-centered decision-making is taking hold and at the center of this shift are those working in the fields of data science (“DS”), artificial intelligence (“AI”), and machine learning (“ML”). As businesses seek to optimize operations and information using vastly increased computing power, the shape and size of data analytics has increased exponentially over the last two decades, and even more prominently in the last decade with the rise of machine learning, cloud computing, and advances in the field of artificial intelligence.

Exploring the composition and trends of salaries in this sector is equally as intriguing, both from a research perspective and from the perspective of what can be waiting for students of Data Analytics (DS/AI/ML) in academic programs globally.

The data we’ve chosen for this project comes from <https://salaries.ai-jobs.net/>, a website collecting salary data from around the world.

This data set has salary (in USD) for a variety of technical roles, and a variety of information on the characteristics of the employee, employer, the job itself, and the market.

The data set contains the following variables:

Citations and Related Work (> Siming)

[[Add citations we consulted, and a summary of related work!]]

Data

Data Cleaning (> Abhishek)

Follow all cleaning steps outlined in class `_clean.csv`

Summary of Data (> Bolun)

- median, mean, sd, range, IQR
- anything else discussed in class
- maybe include boxplots

work_year	The year the salary was paid.								
experience_level	<p>The experience level in the job during the year with the following possible values:</p> <table> <tr> <td>EN</td><td>Entry-level / Junior</td></tr> <tr> <td>MI</td><td>Mid-level / Intermediate</td></tr> <tr> <td>SE</td><td>Senior-level / Expert</td></tr> <tr> <td>EX</td><td>Executive-level / Director</td></tr> </table>	EN	Entry-level / Junior	MI	Mid-level / Intermediate	SE	Senior-level / Expert	EX	Executive-level / Director
EN	Entry-level / Junior								
MI	Mid-level / Intermediate								
SE	Senior-level / Expert								
EX	Executive-level / Director								
employment_type	<p>The type of employment for the role:</p> <table> <tr> <td>PT</td><td>Part-time</td></tr> <tr> <td>FT</td><td>Full-time</td></tr> <tr> <td>CT</td><td>Contract</td></tr> <tr> <td>FL</td><td>Freelance</td></tr> </table>	PT	Part-time	FT	Full-time	CT	Contract	FL	Freelance
PT	Part-time								
FT	Full-time								
CT	Contract								
FL	Freelance								
job_title	The role worked in during the year.								
salary	The total gross salary amount paid.								
salary_currency	The currency of the salary paid as an ISO 4217 currency code.								
salary_in_usd	The salary in USD (FX rate divided by avg. USD rate for the respective year via fxdata.foorilla.com).								
employee_residence	Employee's primary country of residence in during the work year as an ISO 3166 country code.								
remote_ratio	<p>The overall amount of work done remotely, possible values are as follows:</p> <table> <tr> <td>0</td><td>No remote work (less than 20%)</td></tr> <tr> <td>50</td><td>Partially remote</td></tr> <tr> <td>100</td><td>Fully remote (more than 80%)</td></tr> </table>	0	No remote work (less than 20%)	50	Partially remote	100	Fully remote (more than 80%)		
0	No remote work (less than 20%)								
50	Partially remote								
100	Fully remote (more than 80%)								
company_location	The country of the employer's main office or contracting branch as an ISO 3166 country code.								
company_size	<p>The average number of people that worked for the company during the year:</p> <table> <tr> <td>S</td><td>less than 50 employees (small)</td></tr> <tr> <td>M</td><td>50 to 250 employees (medium)</td></tr> <tr> <td>L</td><td>more than 250 employees (large)</td></tr> </table>	S	less than 50 employees (small)	M	50 to 250 employees (medium)	L	more than 250 employees (large)		
S	less than 50 employees (small)								
M	50 to 250 employees (medium)								
L	more than 250 employees (large)								

Figure 1: A caption

[[Add language describing data.]]

Descriptive Statistics (> Bolun) [[Add language summarizing descriptive statistics.]]

Descriptive Plots

Scatterplots (> Bolun) [[Add language summarizing the relationships shown in descriptive plots.]]

Barplots / Histograms (> Abhishek) [[Add language summarizing the relationships shown in descriptive plots.]]

Multivariate plots (> Abhishek) [[Add language summarizing the relationships shown in descriptive plots.]]

[[Overarching summary of plots.]]

Assumptions Tests

Additivity (Rijin)

Linearity (Rijin)

Homogeneity (Rijin)

Normality (Rijin) [[Summary of the tests.]]

Correlation Check / Comments [[Describe any correlations that were observed. How did we deal with this?]]

Technical Approach (> Siming)

[[Language on our overall approach to the analysis.]]

Modeling (All)

Linear models

t-tests

ANOVA tests

Findings (Siming + All)

Conclusion (Siming + All)

Future Studies (All)