EDA

Beibei Du, Wafiakmal Miftah, Suzanna Thompson, Alisa Tian

2022-10-21

IDS702 Final Project EDA

1. Data Overview

The data used in this project comes from Kaggle with 607 observation with 11 variables. This dataset has no null or missing value. This is the link to original post: Data Science Job Salaries

The variables in this dataset are:

Table 1: Variable Names and Description

Column	Description
work_year	The year the salary was paid.
experience_level	The experience level in the job during the year with the following possible values: EN Entry-level / Junior MI Mid-level / Intermediate SE Senior-level / Expert EX Executive-level / Director
employment_type	The type of employement for the role: 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.
salaryinusd	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	The overall amount of work done remotely, possible values are as follows: 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	The average number of people that worked for the company during the year: S less than 50 employees (small) M 50 to 250 employees (medium) L more than 250 employees (large)

For one of our research questions we choose the outcome variable to be salary_in_usd to ensure all the data is converted to the same unit (usd). As shown by the table below, that there are other currency in this dataset.

Table 2: Salary Variable vs Salary in USD Variable

salary_currency	salary	salary_in_usd
EUR	70000	79833
USD	260000	260000
GBP	85000	109024
USD	20000	20000
USD	150000	150000
USD	72000	72000

We choose salary_in_usd over similar columns, salary and salary_currency, in order to standardize our outcome variable and reduce noise and colinearity.

1.2 Two proposed research questions

1. Which factors are associated with an increase in salary for data science jobs? (Continuous outcome)
2. How do company size, company location, employment type, employee residence and job title affected the remote work ratio of a data scientist? (Discrete outcome) For this question, there are three possible values for the remote work ratio, 0, 50, and 100; these signify an in-person job, a hybrid job, and a fully remote job, respectively.

2. Primary relationship of interest

Table 3 below is showing the descriptive statistic for each variable. Variable with asterisks are categorical variable that needs to be look into further in model building.

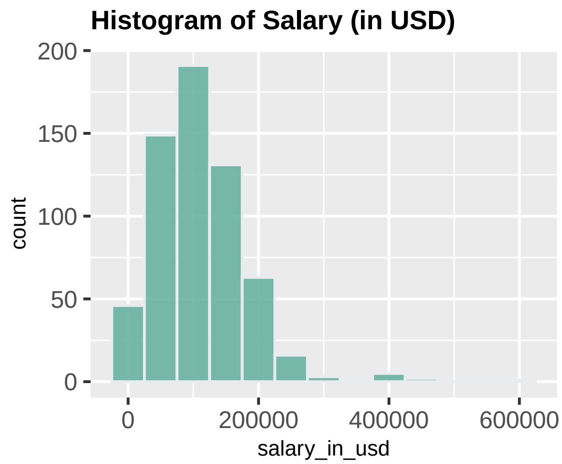
Table 3: Summary of the DS Salary Dataset

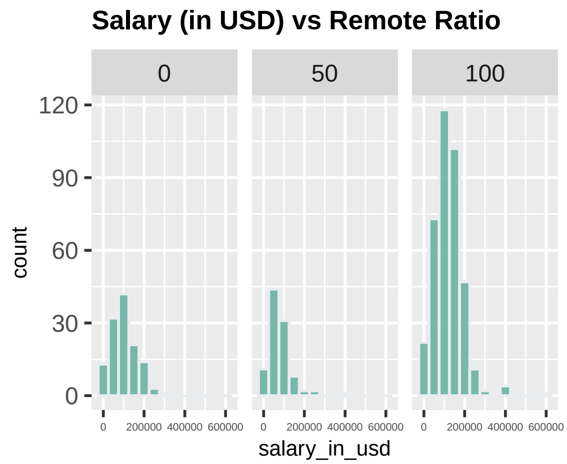
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X	1	607	303.00	175.37	303	303.00	225.36	0	606	606	0.00	-1.21	7.12
work__year	2	607	2021.41	0.69	2022	2021.51	0.00	2020	2022	2	-0.73	-0.66	0.03
experience__level*	3	607	3.13	1.03	3	3.28	1.48	1	4	3	-1.04	-0.10	0.04
employment__type*	4	607	2.99	0.24	3	3.00	0.00	1	4	3	-4.14	45.81	0.01
job__title*	5	607	21.96	10.49	18	21.00	7.41	1	50	49	0.88	0.40	0.43
salary	6	607	324000.06	1544357.49	115000	118919.11	68706.65	4000	30400000	30396000	13.98	244.57	62683.54
salary__currency*	7	607	14.03	4.38	17	14.67	0.00	1	17	16	-1.03	-0.38	0.18
salary__in__usd	8	607	112297.87	70957.26	101570	106157.63	62906.72	2859	600000	597141	1.66	6.26	2880.07
employee__residence*	9	607	41.41	18.27	56	43.66	0.00	1	57	56	-0.67	-1.22	0.74
remote__ratio	10	607	70.92	40.71	100	76.08	0.00	0	100	100	-0.90	-0.90	1.65
company__location*	11	607	36.89	16.03	49	39.07	0.00	1	50	49	-0.77	-1.09	0.65
company__size*	12	607	1.81	0.65	2	1.76	0.00	1	3	2	0.21	-0.73	0.03

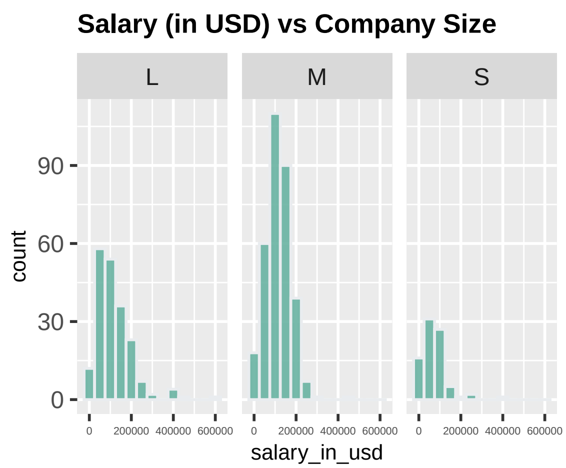
2.1. Descriptive stats and plots Answering question 1

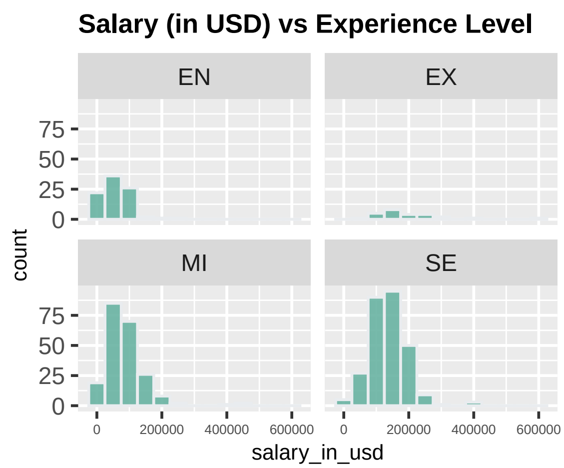
Our first outcome variable, salary in USD, ranged from USD2,859 to USD600,000. Diving into each variable, the majority of the respondents are Senior Level employee (46%), which majority working fully remote (71%), full in office (19%), or hybrid (10%). Most of the Senior Level employee works in medium size companies (66%), while the rest are working in large companies (26%) or small companies (8%). Most of their employment status are Full-Time (99%), while the rest are Contracts or Freelance. None of them are Part-Time employee. The second biggest respondents are Mid/Intermediate Level employee (35%), which majority working fully remote (54%), full in office (26%), or hybrid (20%). Most of the Mid/Intermediate Level employee works in medium size companies (46%), while the rest are working in large companies (40%) or small companies (14%). Most of their employment status are Full-Time (97%), while the rest are Contracts, Freelance, or Part-Time employee. The remaining respondents are Entry/Junior Level employee (15%) and Executive/Director level (4%).

There are 50 different company location, which mostly in the US (58%), followed by Great Britain (8%), Canada (5%), and the rest of the world. While the employee residence data shows that respondents live in 57 different country, probably made possible by the ability to work remotely. Most of the respondents lived in the US (55%), followed by Great Britain (7%), India (5%), and the rest of the world. There are 50 different job titles in this dataset, but all of them are in the field of data science. The job title variable is dominated by data scientist (24%), data engineer (22%), data analyst (16%), while the rest are varied but mostly have “engineer” or “data” in the title. Based on the barplot shown below, there are certain pattern from Remote Ratio, Company Size and Experience level when plotted against salary (in USD).









2.2. Descriptive stats and plots Answering question 2

3. Other characteristics

Our dataset includes work_year, which is the year that the data was collected. This variables contains three unique values, 2020, 2021, and 2022, which contain 11.86%, 35.75%, and 52.39% respectively. Similarly, job_title has 50 unique values. Below is a table showing the unique values in the variable and their respective counts.

From table 4, we can see that the job titles related to data science are not in short supply. Furthermore, company_location has 50 unique values and employee_residence has 57 unique values. They are each listed from table 5 to table 8.

The company_size variable is also of note: it has three potential variables, S for small, M for medium, and L for large. Its distribution of counts is listed in Table 9.

Table 4: Summary of Job Titles

Var1	Freq
3D Computer Vision Researcher	1
AI Scientist	7
Analytics Engineer	4
Applied Data Scientist	5
Applied Machine Learning Scientist	4
BI Data Analyst	6
Big Data Architect	1
Big Data Engineer	8
Business Data Analyst	5
Cloud Data Engineer	2

Var1	Freq
Computer Vision Engineer	6
Computer Vision Software Engineer	3
Data Analyst	97
Data Analytics Engineer	4
Data Analytics Lead	1
Data Analytics Manager	7
Data Architect	11
Data Engineer	132
Data Engineering Manager	5
Data Science Consultant	7

Var1	Freq
Data Science Engineer	3
Data Science Manager	12
Data Scientist	143
Data Specialist	1
Director of Data Engineering	2
Director of Data Science	7
ETL Developer	2
Finance Data Analyst	1
Financial Data Analyst	2
Head of Data	5

Var1	Freq
Head of Data Science	4
Head of Machine Learning	1
Lead Data Analyst	3
Lead Data Engineer	6
Lead Data Scientist	3
Lead Machine Learning Engineer	1
Machine Learning Developer	3
Machine Learning Engineer	41
Machine Learning Infrastructure Engineer	3
Machine Learning Manager	1

Var1	Freq
Machine Learning Scientist	8
Marketing Data Analyst	1
ML Engineer	6
NLP Engineer	1
Principal Data Analyst	2
Principal Data Engineer	3
Principal Data Scientist	7
Product Data Analyst	2
Research Scientist	16
Staff Data Scientist	1

Table 5: Summary of Employee Residence(1)

AE	AR	AT	AU	BE	BG	BO	BR	CA	CH	CL	CN	CO	CZ	DE	DK	DZ	EE	ES	FR	GB	GR	HK	HN	HR
3	1	3	3	2	1	1	6	29	1	1	1	1	1	25	2	1	1	15	18	44	13	1	1	1

Table 6: Summary of Employee Residence(2)

HU	IE	IN	IQ	IR	IT	JE	JP	KE	LU	MD	MT	MX	MY	NG	NL	NZ	PH	PK	PL	PR	PT	RO	RS	RU	SG	SI	TN	TR	UA	US	VN
2	1	30	1	1	4	1	7	1	1	1	1	2	1	2	5	1	1	6	4	1	6	2	1	4	2	2	1	3	1	332	3

Table 7: Summary of Company Location(1)

AE	AS	AT	AU	BE	BR	CA	CH	CL	CN	CO	CZ	DE	DK	DZ	EE	ES	FR	GB	GR	HN	HR	HU	IE	IL
3	1	4	3	2	3	30	2	1	2	1	2	28	3	1	1	14	15	47	11	1	1	1	1	1

Table 8: Summary of Company Location(2)

IN	IQ	IR	IT	JP	KE	LU	MD	MT	MX	MY	NG	NL	NZ	PK	PL	PT	RO	RU	SG	SI	TR	UA	US	VN
24	1	1	2	6	1	3	1	1	3	1	2	4	1	3	4	4	1	2	1	2	3	1	355	1

Table 9: Summary of Company Size

L	M	S
198	326	83

4. Potential challenges

A challenge we'll need to address before modeling is the amount of job titles that are present in our data. To solve this, we will either collaspe certain job titles into relevant categories or exclude job titles based on some metric. In that, if the job title "NLP Engineer" only shows up once, we may decide to drop it and all other job titles with similar counts. This decision will be informed by domain knowledge and field experts. We face a similar problem with employee_residence and company_location. Not only is more than half of our data from the United States, Great Britian, and Canada, but the two variables are also highly correlated. Because we're investigating the relationship between work place attributes and the remote-ratio, we will include both location-based variables. There are various common data issues that need to be addressed before modeling. We consider them below. Messy data is defined to be a dataset where values are unstandardized, unorganized, or bias. Largely, our data is clean. The only potential messiness in our data comes from the large number of job titles as mentioned above. Another common data issue is a lack of data. As our data has 607 entries, we do not face this problem. Finally, confounding variables can cause modeling issues as there may be variables that are related to our questions at hand that are not present in our data. Because our research questions are somewhat poignant in their phrasing, we will not encounter this as a data issue. One might suppose that a level of education may have an effect on a data scientist's salary as it tends to have an effect in other fields. However, because data science is a relatively new field, there are not much data on whether or not a higher-education degree has an effect on the ensuing salary. For this reason, we find no issue in assuming it has marginal effect.