EDA

Beibei Du, Wafiakmal Miftah, Suzanna Thompson, Alisa Tian

2022-10-20

IDS702 Final Project EDA

1. Data Overview

The data used in this project comes from Kaggle with 969 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)

Our outcome variable is 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

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)

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.

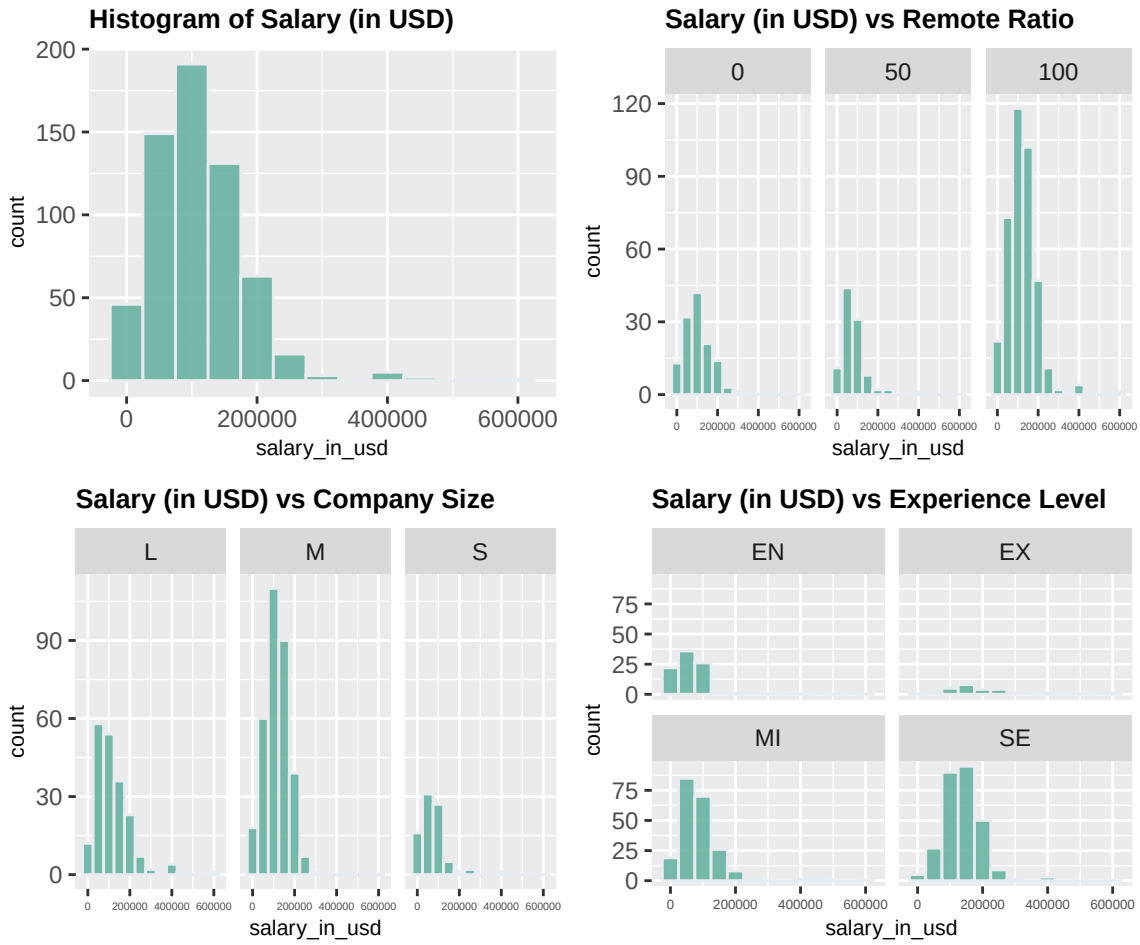
Table 3: Summary of the DS Salary Dataset

	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

Briefly describe other variables in the data. If there are many, do not list them all. Rather, describe the types of variables that are present (e.g., "demographic information").

4. Potential challenges

Describe aspects of the data that may present challenges in the modeling stage. For example, might certain categorical variables need to be collapsed? Is there a lot of missingness? Could the size of the dataset present model selection challenges?