EDA

Beibei Du, Wafiakmal Miftah, Suzanna Thompson, Alisa Tian

2022-10-19

IDS702 Final Project EDA

Team member: Beibei(Bella) Du, Wafiakmal Miftah, Suzanna Thompson, Wenjing Tian

1. Data Overview

The data used in this project comes from Kaggle with 969 observation with 11 variables. 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
USD	190000	190000
USD	135000	135000
USD	135000	135000
USD	80000	80000
INR	633000	8191
USD	160000	160000

This table

Table 3: Summary of the DS Salary Dataset

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
work_year	1	969	2021.61	0.63	2022	2021.74	0.00	2020	2022	2	-1.38	0.73	0.02
experience_level*	2	969	3.24	1.02	4	3.42	0.00	1	4	3	-1.22	0.24	0.03
employment_type*	3	969	2.99	0.22	3	3.00	0.00	1	4	3	-4.95	58.07	0.01
job_title*	4	969	25.44	12.03	21	24.90	11.86	1	57	56	0.51	0.14	0.39
salary	5	969	270650.76	1256778.24	123000	124644.42	71164.80	2324	30400000	30397676	16.61	354.61	40373.53
salary_currency*	6	969	14.58	4.10	17	15.30	0.00	1	17	16	-1.34	0.43	0.13
salary_in_usd	7	969	119022.10	68378.55	113000	115044.01	66383.41	2324	600000	597676	1.15	4.05	2196.64
employee_residence*	8	969	48.96	19.06	62	52.12	0.00	1	63	62	-1.01	-0.59	0.61
remote_ratio	9	969	67.49	43.31	100	71.81	0.00	0	100	100	-0.74	-1.26	1.39
company_location*	10	969	46.51	17.49	58	49.50	0.00	1	59	58	-1.07	-0.49	0.56
company_size*	11	969	1.83	0.60	2	1.78	0.00	1	3	2	0.08	-0.37	0.02

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

Present descriptive statistics and exploratory plots in whichever format you think is best (tables, figures) for your primary relationship of interest (dependent variable and primary independent variable, if applicable). Describe your findings.

2.1. Descriptive stats and plots Answering question 1

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?