# 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

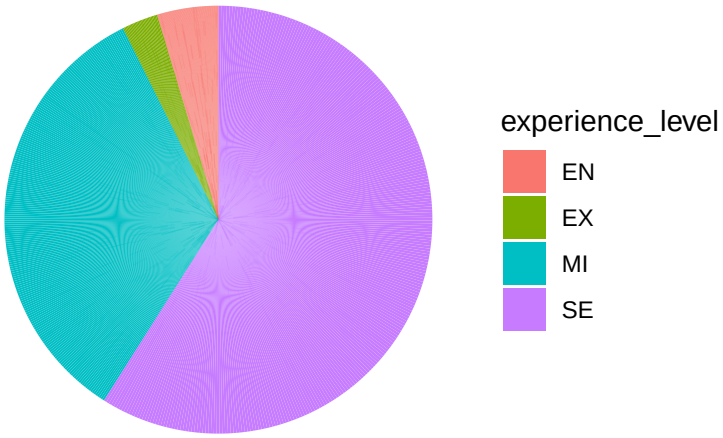
1. Which factors are associated with an increase in salary for data science jobs? (Continuous outcome)
2. How do company size, company location, employment type, employee residence and job title affected the remote work ratio of a data scientist? (Discrete outcome)

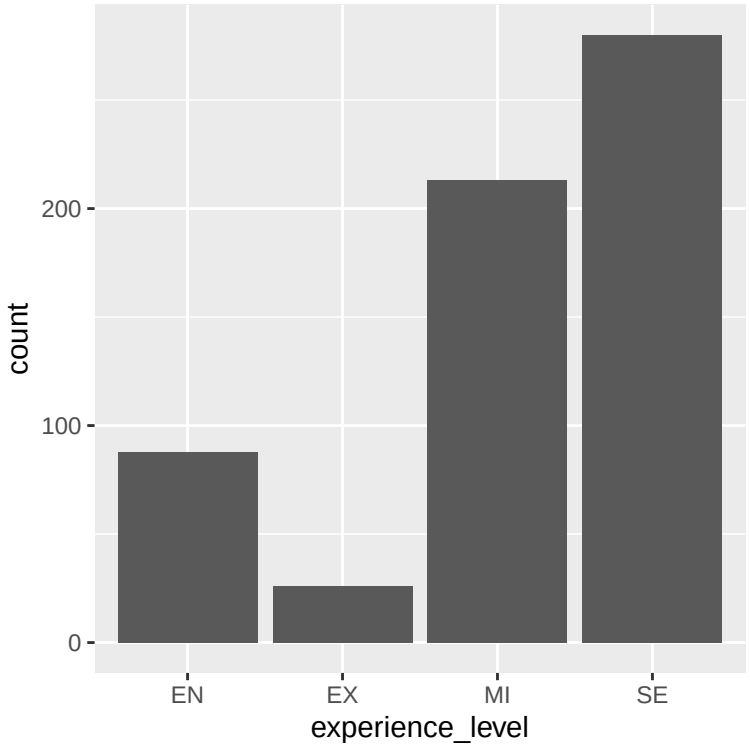
## 2. Primary relationship of interest

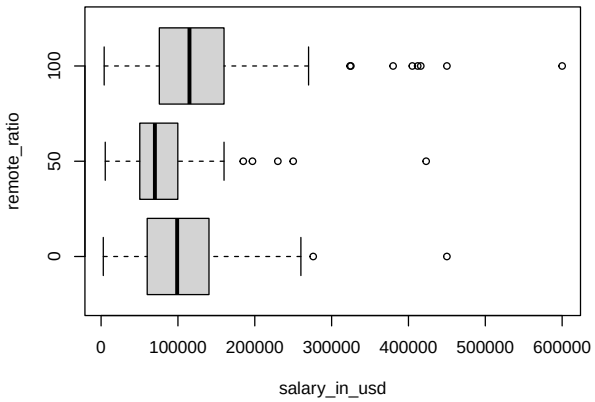
Table 3 below is showing the descriptive statistic for each variable. Variable with asterisks are categorical variable that needs to be look into further in model building.

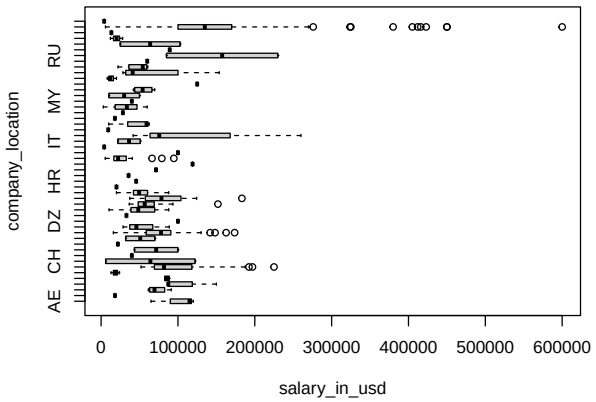
### 2.1. Descriptive stats and plots Answering question 1

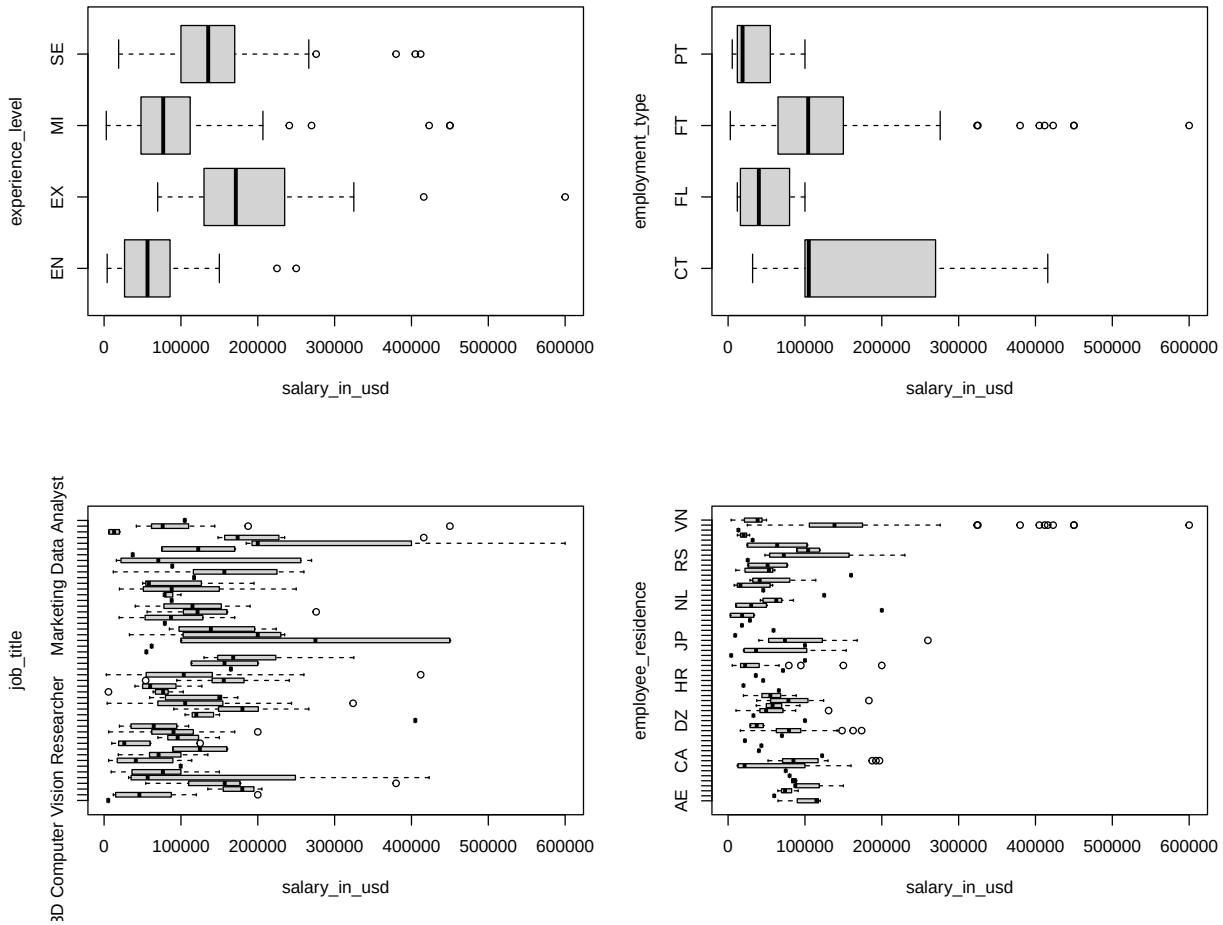
In trying find the variable that potentially has statistical relationship with the outcome variable, we can plot each variable with salary_in_usd:




count
200
100
0
EN
EX
MI
SE
experience_level


remote_ratio
0
50
100
0
100000
200000
300000
400000
500000
600000
salary_in_usd


company_location
AE
CH
DZ
HR
IT
MY
RU
0
100000
200000
300000
400000
500000
600000
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?