IDS702_Final Project

Beibei Du, Wafiakmal Miftah, Alisa Tian, Suzanna Thompson

2022-10-02

IDS702 Final Project Proposal

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

1.1 Selected Dataset

We have selected the dataset from Kaggle. [https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries].

1.1.1 Background

As data scientists, we are curious about the salaries of data scientists in the industry or research field. We decide to do Exploratory Data Analysis (EDA) and statistical analysis to find pattern and trend in the salaries of data scientists. We want to find what variable(s) has the most effect in the salaries.

1.1.2 Data

There are 11 columns/variables in this dataset. The variables are: `work_year`, `experience_level`, `employment_type`, `job_title`, `salary`, `salary_currency`, `salaryinusd`, `employee_residence`, `remote_ratio`, `company_location`, `company_size`. We have a total of 607 observations and each observation belongs to an individual works in the data industry.

We will describe each of the variables in detail below: - `work_year`:The year that the individuals was paid

- `experience_level`: Experience level of the individual: EN Entry-level / Junior MI Mid-level / Intermediate SE Senior-level / Expert EX Executive-level / Director [Discrete]
- `employment_type`: Type of employment of the individual: PT –> Part-time; FT –> Full-time; CT –> Contract; FL –> Freelance [Discrete]
- `job_title`: The Job title fo the individual. [Discrete]
- `salary`: The total gross salary amount the individual has received. [Continuous]
- `salary_currency`: The currency of the salary paid as an ISO 4217 currency code. [Discrete]
- `salaryinusd`: The salary in USD (FX rate divided by avg. USD rate for the respective year via fxdata.foorilla.com). [Continuous]
- `employee_residence`: Employee's primary country of residence in during the work year as an ISO 3166 country code. [Discrete]

- `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%). [Discrete]
- `company_location`: The country of the employer's main office or contracting branch as an ISO 3166 country code. [Discrete]
- `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). [Discrete]

1.2 Two proposed research questions

1. What are the most effective factors that increasing the salaries of individuals who worked full time in the data industry? [The outcome is continuous]
2. In what ways do factors such as company size and salaries, and etc impact whether the individual worked remotely or not? [The outcome is discrete]