Decoding Developer Salaries: Econometric Analysis of Experience, AI Adoption, and Regional Trends

Econometrics. Final report

Mariia Onyshchuk
Mykhailo Ponomarenko

Introduction

This study utilises data from the **2024 Stack Overflow Developer Survey** to conduct an econometric analysis, aiming to identify and quantify the factors that impact annual compensation among IT professionals and to see the difference between different groups and workers.

Aim

In this changing environment, it's important to understand what factors affect how much IT professionals earn. Traditional factors like education, years of experience, and job roles have always played a part in determining salaries. However, the rise of AI tools introduces new elements that could influence income.
The main aim of our research is to explore and quantify the determinants of people in IT industry salaries using data from the Stack Overflow Survey. By econometric analysis, we seek to identify how such impact incomes, with a specific focus on differences revealed by salaries reported in UAH and USD (which are only currencies, do not determine whether the country is Ukraine only or USA only).

On top of that we want to estimate what factors lead people to work **remotely** in their everyday work by using methods of logistic regression on the same data and to test the hypotheses about the significance of different factors.

Motivation

Our research is driven by several key objectives:

- **Understanding Salary Drivers:** Explore the impact of factors—from years of coding experience to the adoption of AI technologies—on developer earnings.
- **Informing Compensation Strategies:** Assist companies in designing competitive compensation packages that reflect observed market trends.
- **Exploring Tech Trends:** Analyse broader trends in the tech industry, with the important caveat that the self-reported survey data may not fully correspond to real-world figures.
- **Understanding reasons for working remotely:** Estimate what influence that people choose to work remotely.

Target Audience

Our findings will be valuable to a diverse range of stakeholders, including:

- **HR Departments and Recruitment Agencies:** Seeking data to design competitive compensation packages and understand evolving market trends.

- **Investors and Financial Analysts:** Using salary trends as one indicator of the tech industry's growth potential.

Data analysis

Unprocessed data can look overwhelming from the first glance. On Fig. 1 noticeable how the data is unbalances, which can lead to the biasedness of the model. Fig. 2 represents the result of data analysis, cleaning of the dataset and balancing the groups in order to make our research qualitative.

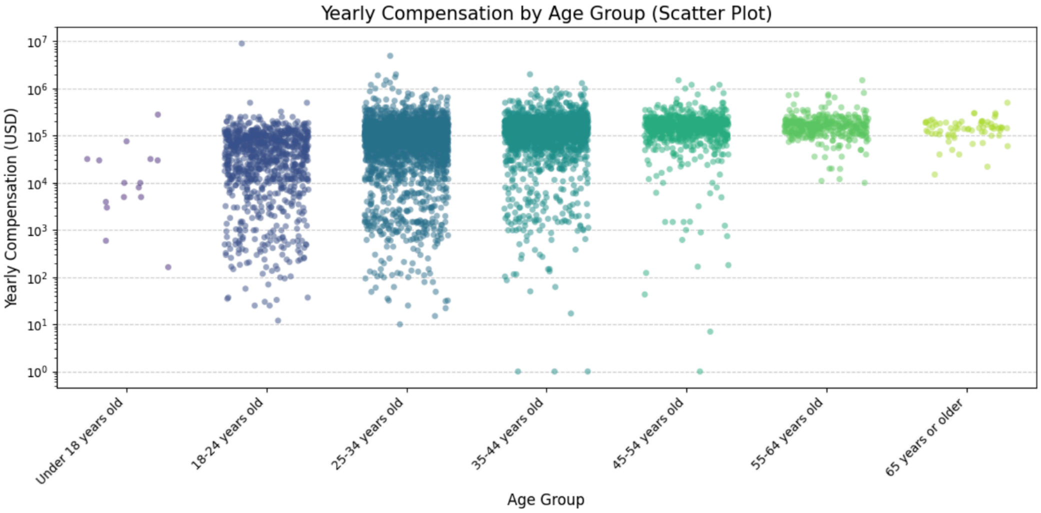


Fig. 1 Before the data cleaning and balancing

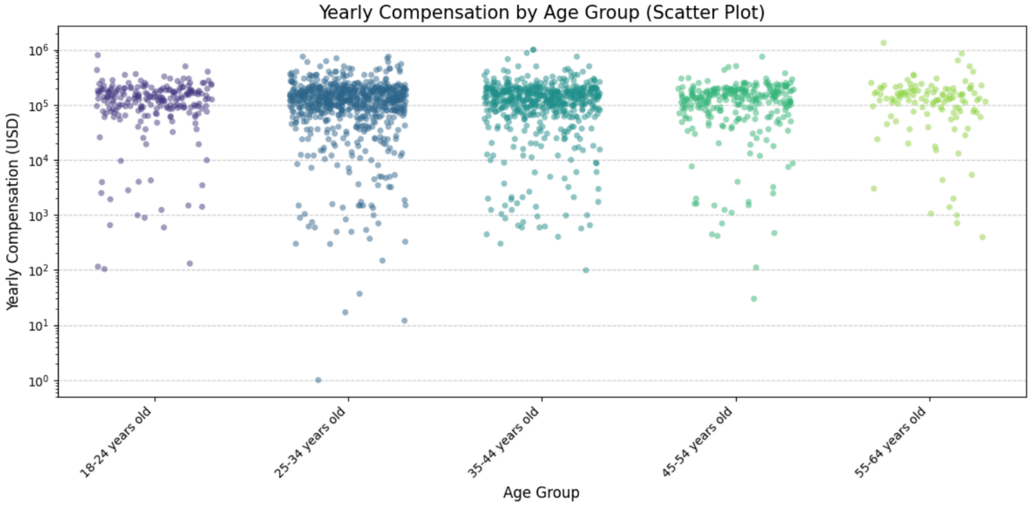


Fig. 2 After the data cleaning and balancing

Raw survey data were highly unbalanced and split into dozens of separate columns—many representing multiple-choice or ordinal responses (e.g. age ranges). To run a linear regression, we either had to create hundreds of dummy variables or assign arbitrary numeric codes to non-numeric answers. The dummy-variable approach ballooned our frame to over 150 columns, many of which were nearly empty (for example, only one respondent was under 18). We therefore cleaned and balanced the dataset, reducing dimensionality and ensuring more reliable model estimates.

We distilled our dataset by dropping insignificant columns, leaving only key predictors such as age, job role, years of experience, education level, AI usage, remote-work status, and areas of interest (full definitions in the Appendix).

To guard against extreme outliers—like implausible salary or experience entries—we applied the 3-sigma rule to trim the tails and then performed targeted manual cleaning. This two-step approach ensured a more balanced sample and more reliable regression results.

We focused our sample on full-time professionals by removing part-time workers and freelancers—roles for which we lack consistent data on hours worked—and framed our analysis around individuals "fully" engaged in their jobs.

To avoid an unwieldy set of position-specific dummies, we consolidated job titles into a handful of meaningful categories and excluded "Non-Technical" roles (e.g. Developer Advocate, Marketing/Sales, Educator) to hone in on core technical and managerial functions. Figures 3 and 4 illustrate how these broad groupings yield a more balanced distribution (see Fig. 3, 4 in Annex).

Finally, we addressed data quality and imbalance by applying a 3-sigma rule plus manual review to trim implausible salary–experience outliers, dropping under-populated dummy variables, and rebalancing class sizes. The result is a leaner, more stable dataset whose regression estimates will be both reliable and interpretable.

Linear assumptions checking

Multicollinearity

Here is our correlation matrix heatmap. It is visible that for obvious reasons we removed the WorkExp column as it is almost the same as YearsCodePro (the meaning of the columns are at the Annex of the work). On top of that we decided to drop age columns as they are higly correlated with working experience. And also the educations dummy variables were replaced with 'Years of education' equal to either 10, 14, 16 or 18 based on person's degrees.

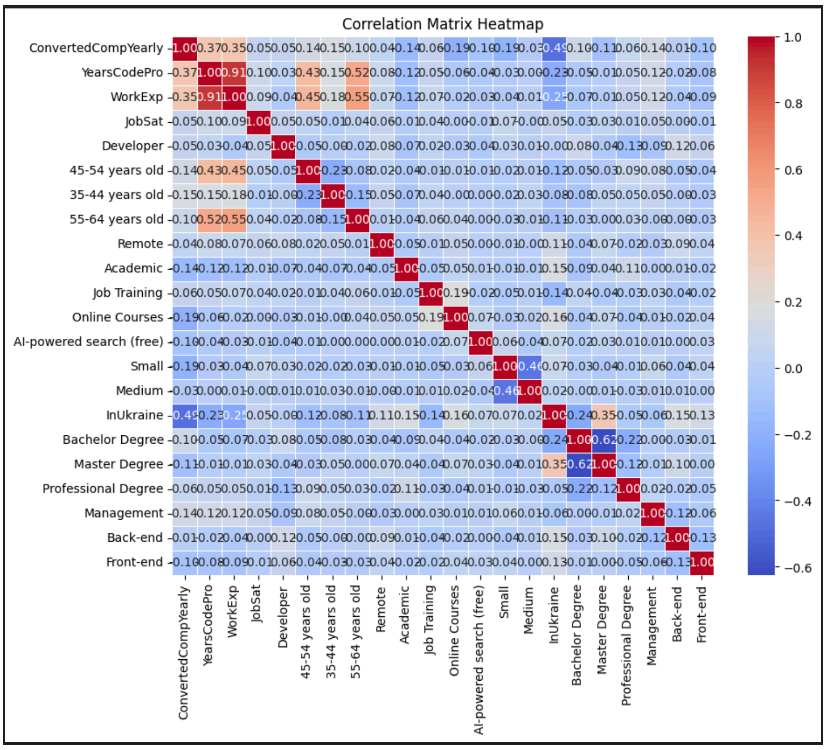


Fig. 3: Correlations

Normality

The p-value of Kolmogorov-Smirnov test on our data is 7.805914704509913e-28 which means our data is definetiely not normal, partially because of a large right skew as can be seen on the plot, so we are going to clean the outliears using standard deviation gaps from the mean. After doing so, the p-value is 0.002 which is not reallly great, but all right for some of the usual confidence levels.

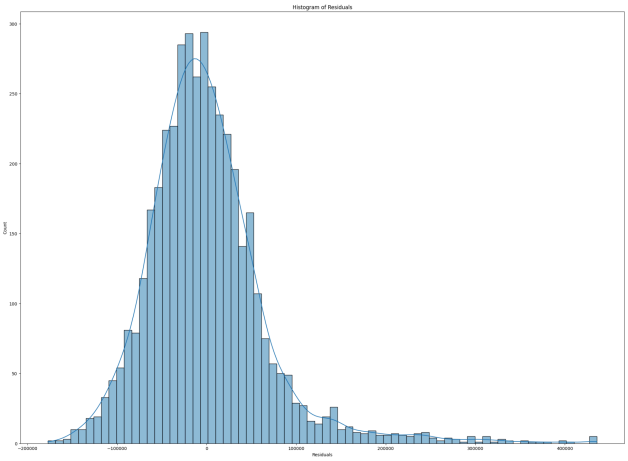


Fig. 4: Distribution before

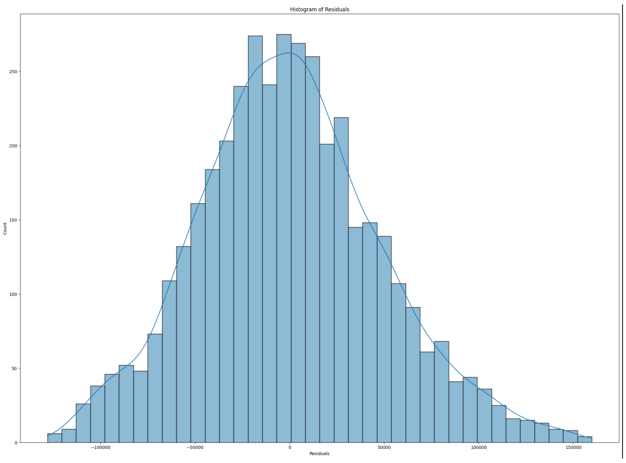


Fig. 5: Distribution after

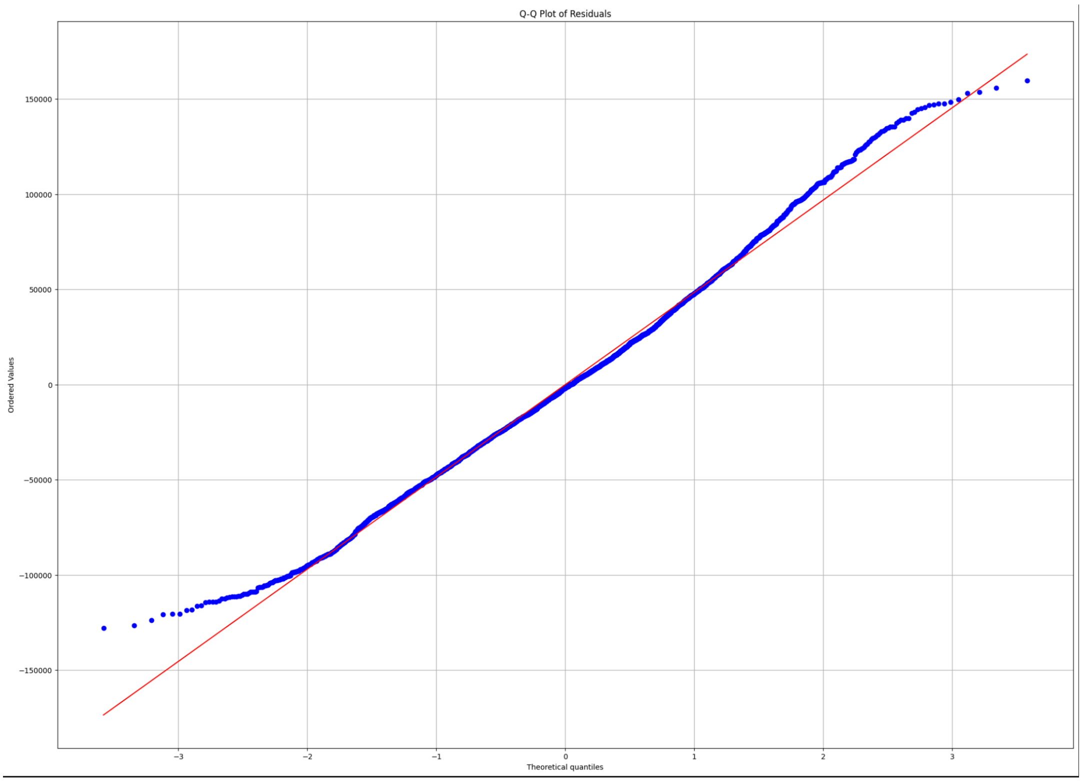


Fig. 6: Q-Q plot

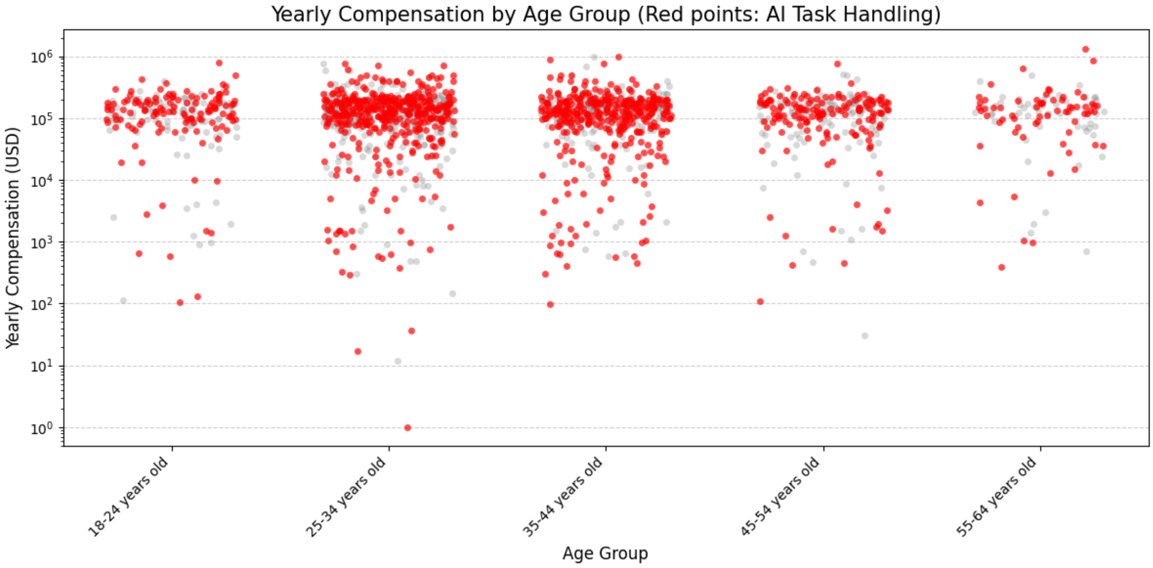


Fig. 7: People who use the AI tool according to the age group

● Prefer not to use AI tool
● AI Task Handling

Homoscedasticity

We applied a White test and obtained an LM p-value of 4.816449297707804e-25, thus heteroscedasticity is defibetely present. So we made a decision to use a homoscedasticity-robust model HC2.

All other assumptions were not violated from the beginning

Hypotheses and objectives description

Full model attached in the appendix of the report

Since we have all data to implement the idea of our research, lets plot the dependencies and make hypotheses analysis.

Education and Salary Impact

We tested whether higher education levels raise annual income:

H_0: Education has no effect on earnings

H_1: More education нуфкiincreases earnings

```
                              coef    std err          t      P>|t|      [0.025      0.975]
Years of Education       2772.1190    521.959      5.311      0.000    1749.098    3795.140
```

Note: The coefficients represent the estimated increase in annual income (in USD) associated with each education level, relative to the baseline category (e.g., high school diploma or equivalent).

Interpretation:

Due to the results, on avarage, the additional year of education increase the yearly compensation on around 2800$. This results strongly support that higher formal education is associated with higher annual income.

AI Adoption and Earnings

We tested whether using AI-powered search tools affects income:

H_0: Using free or paid AI tools has no impact on annual earnings

H_1: Using free or paid AI tools influences annual earnings

```
                            coef    std err          t      P>|t|      [0.025      0.975]
AI-powered search (free)  -1.85e+04   3963.371     -4.667      0.000   -2.63e+04   -1.07e+04
AI-powered search (paid)  1.177e+04   4752.763      2.477      0.013    2449.377    2.11e+04
```

Note: The coefficients represent the change in annual income (in USD) associated with each variable.

Interpretation:

Although paid AI tools show a positive coefficient, it isn't as significant ($p = 0.013$) as free AI tools, which yield more significant negative effect ($p < 0.001$), suggesting that investing in paid AI resources is associated with higher income.

Straight away answering on a question about **perfect collinearity**: using paid AI tool does not always mean that the person do not use also the free version of another tool. The answer was derived from the question with more than just these 2 options as an answer. People could choose both options. What particularly AI tools we are talking about (from most frequent answers) are visualised on the Fig 8.

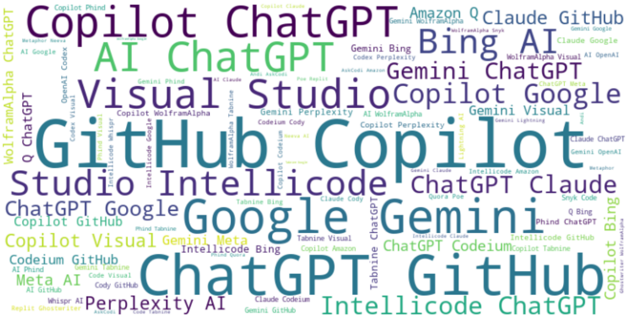


Fig. 8: the most frequently used AI tools

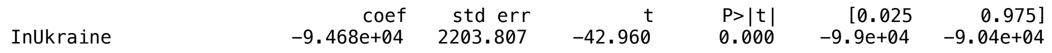
Country Differences

Developers in Ukraine tend to earn significantly less than their counterparts in other countries. We suppose the reasons intuitively obvious for Ukrainians.

Hypothesis Test

H_0: Ukrainian IT salaries equal those in other countries

H_1: Ukrainian IT salaries differ from those abroad



```
                               coef    std err          t      P>|t|      [0.025      0.975]
InUkraine                -9.468e+04   2203.807    -42.960      0.000    -9.9e+04   -9.04e+04
```

Interpretation:

IT professionals in Ukraine earn markedly less than peers elsewhere, with our model estimating a salary gap of about $94,680 ($p < 0.001$). This disparity likely reflects broader economic factors such as exchange-rate differences, local market conditions, cost-of-living variances etc.

Methods

For the hypotheses we've used two-sided t-tests, which determine whether the estimated coefficient is **significantly different from zero in either direction**—that is, whether the predictor has any effect, positive or negative, on annual income.

$$H_0 : \beta_{\mathrm{i}} = 0,$$
$$H_1 : \beta_{\mathrm{i}} \neq 0,$$
$$t : \frac{\hat{\beta}_{\mathrm{i}}}{\mathrm{SE}(\hat{\beta}_{\mathrm{i}})} \sim t_{n-K-1}$$

Note that all coefficient significances reported above were computed directly from the OLS estimation output—specifically, the t-statistics and p-values produced by statsmodels' `.summary()` method—ensuring each null hypothesis ($\beta=0$) is evaluated under the Student's t-distribution. We model total compensation for an i-th individual as a linear function of K predictors (years of coding, AI usage, experience, job satisfaction, dummy variables for demographics, learning methods, roles, etc.), plus an intercept and error term:

$$\text{Compensation} = \beta_0 + \beta_1 \text{WorkExp} + \beta_2 \text{AI} + \beta_3 \text{Remote} + \beta_4 \text{JobSat} + \ldots + \varepsilon$$

Logistic Regression Analysis of Remote Work Determinants

In this section, we estimate a logistic regression where the dependent variable is **Remote** (1 = remote work, 0 = on-site), using maximum likelihood via a logit link function. Meaning of predictors and the model are located in the Appendix enter the model simultaneously.

Overall, the logit model (6825 observations, Pseudo $R^2 \approx 0.074$, LLR $p < .0001$) shows modest explanatory power typical of survey data and a highly significant improvement over the null model.

- **Job Satisfaction**: Each one-unit increase in JobSat raises the log-odds of remote work (coef = 0.0479, p = 0.038,), indicating more satisfied employees are likelier to work remotely.
- **Geographic Location**: Being based in Ukraine strongly increases remote-work probability (coef = 1.8390, p < .001), reflecting cross-border outsourcing trends.
- **Company Size**: Employees at medium-sized firms are less likely remote(coef = –0.03) than those at small firms (coef = 0.0140), implying smaller organizations may offer greater flexibility.
- **Role Category**: Back-end (0.6934, p = .000), and Front-end (0.6081, p = 0.029) roles significantly increase remote-work odds, highlighting core technical positions as most adaptable to remote arrangements. Conversely, according to the coefficient of Management position (coef = -0.4088, p = 0.053), leadership roles decrease remote-work odds.

Conclusions

In our analysis, we find that each additional year of professional experience reliably boosts annual income—by roughly $2377 —allowing us to reject the notion that experience doesn't matter. At the same time, formal education, as it was expected, positively influence the income.

When it comes to AI adoption, the quality of tools makes a real difference. Simply using free AI-powered search offers no clear earnings benefit, whereas investing in paid solutions increase the average yearly income due to the answers in the survey.

Finally, geography and work context shape both income and working arrangements. Our model shows Ukrainian IT professionals earn nearly $95 000 less than their international peers, highlighting stark economic disparities. In parallel, a logistic regression on remote-work status confirms that both individual traits (age, job satisfaction) and company factors (size, role category) jointly determine who works from home. Together, these findings underscore that while experience and education set the baseline for earnings, strategic investment in paid AI tools and broader economic and organizational environments critically influence IT career outcomes.

Appendix

You can find a shortened version of the code with data preprocessing and models built on GitHub. Here is our OLS model, where you can check our results, which we've obtained. Also, there are some variables, which we haven't mentioned, but the meaning is provided in the table below.

```
                            OLS Regression Results

       ==============================================================================
       Dep. Variable:     ConvertedCompYearly   R-squared:                       0.390
       Model:                             OLS   Adj. R-squared:                  0.388
       Method:                  Least Squares   F-statistic:                     310.8
       Date:                 Mon, 05 May 2025   Prob (F-statistic):               0.00
       Time:                         20:25:31   Log-Likelihood:                -53011.
       No. Observations:                 4226   AIC:                         1.061e+05
       Df Residuals:                     4212   BIC:                         1.061e+05
                    Df Model:                         13
                    Covariance Type:                 HC2
==============================================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                   6.193e+04   9205.330      6.727      0.000    4.39e+04       8e+04
YearsCodePro            2377.5439    126.579     18.783      0.000    2129.454    2625.634
JobSat                  2050.9190    505.771      4.055      0.000    1059.625    3042.213
Developer               1.613e+04   4008.748      4.025      0.000    8277.099      2.4e+04
Remote                  1.633e+04   3748.592      4.355      0.000    8979.398     2.37e+04
Academic               -1.435e+04   3685.102     -3.895      0.000   -2.16e+04   -7130.998
Online Courses          -2.02e+04   2143.613     -9.422      0.000   -2.44e+04     -1.6e+04
AI-powered search (free) -1.85e+04  3963.371     -4.667      0.000   -2.63e+04   -1.07e+04
Small                  -3.874e+04   2572.382    -15.058      0.000   -4.38e+04   -3.37e+04
Medium                  -2.38e+04   2636.940     -9.025      0.000    -2.9e+04   -1.86e+04
InUkraine              -9.468e+04   2203.807    -42.960      0.000    -9.9e+04   -9.04e+04
Management              4.256e+04   6120.996      6.953      0.000    3.06e+04    5.46e+04
Back-end                1.394e+04   2946.564      4.731      0.000    8163.696    1.97e+04
Years of Education      2772.1190    521.959      5.311      0.000    1749.098    3795.140
       ==============================================================================
       Omnibus:                     1549.550   Durbin-Watson:                   2.000
       Prob(Omnibus):                  0.000   Jarque-Bera (JB):             8672.481
       Skew:                           1.653   Prob(JB):                         0.00
       Kurtosis:                       9.191   Cond. No.                         193.
       ==============================================================================

                                         Notes:
                    [1] Standard Errors are heteroscedasticity robust (HC2)
```

Logit model which also was mentioned:

```
                      Optimization terminated successfully.
                             Current function value: 0.313578
                                       Iterations 8
                                          Logit Regression Results
       ==============================================================================
       Dep. Variable:                 Remote   No. Observations:                 4226
       Model:                          Logit   Df Residuals:                     4210
       Method:                           MLE   Df Model:                           15
       Date:                Mon, 05 May 2025   Pseudo R-squ.:                 0.07663
       Time:                        20:25:31   Log-Likelihood:                -1325.2
       converged:                       True   LL-Null:                       -1435.2
       Covariance Type:            nonrobust   LLR p-value:                 1.814e-38
==============================================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                     -0.3254      0.435     -0.748      0.454      -1.178       0.527
ConvertedCompYearly      3.79e-06   8.75e-07      4.332      0.000    2.08e-06      5.5e-06
YearsCodePro               0.0330      0.007      4.530      0.000       0.019       0.047
JobSat                     0.0479      0.023      2.079      0.038       0.003       0.093
Developer                  0.5241      0.181      2.891      0.004       0.169       0.879
Academic                  -0.5448      0.171     -3.193      0.001      -0.879      -0.210
Job Training              -0.0548      0.107     -0.515      0.607      -0.264       0.154
Online Courses             0.3800      0.108      3.531      0.000       0.169       0.591
AI-powered search (free)   0.0532      0.228      0.234      0.815      -0.393       0.499
Small                      0.0140      0.128      0.109      0.913      -0.238       0.266
Medium                    -0.0304      0.129     -0.235      0.814      -0.284       0.223
InUkraine                  1.8390      0.241      7.629      0.000       1.367       2.311
Management                -0.4088      0.211     -1.936      0.053      -0.823       0.005
Back-end                   0.6934      0.163      4.244      0.000       0.373       1.014
Front-end                  0.6081      0.279      2.179      0.029       0.061       1.155
Years of Education         0.0305      0.024      1.250      0.211      -0.017       0.078
==============================================================================================
```

Table 1: Interpretation of the coefficients:

YearsCodePro	Number of years the person has been coding professionally (as part of work).
WorkExp	Number of years of professional work experience.
JobSat	Job satisfaction rating, on a scale from 1 to 10.
Developer	Indicates whether the person has part “Developer” in the name of the position
45-54 years old	Age group category(18-45 y.o. is a base group).
35-44 years old	
55-64 years old	
Remote	Work arrangement type, indicating if the position is remote (1=Yes, 0=No).
Academic	Indicates if the person has an academic background, such as a degree in computer science.
Job Training	Indicates if person received training during employment (1 = Yes, 0 = No).
Online Courses	Indicates if person uses online courses for learning (1 = Yes, 0 = No).
AI-powered search/dev tool (free)	Indicates if person uses free AI-powered search or development tools (1 = Yes, 0 = No).
AI-powered search/dev tool (paid)	Indicates if person uses paid AI-powered search or development tools (1 = Yes, 0 = No).
Small	Indicates wheater the size of the company is small
Medium	Indicates wheater the size of the company is medium
InUkraine	Indicates if person is based in Ukraine (1 = Yes, 0 = No).
Bachelor Degree	Indicates if person has a bachelor's degree (1 = Yes, 0 = No).
Master Degree	Indicates if person has a master's degree (1 = Yes, 0 = No).
Professional Degree	Indicates if person has a professional degree (1 = Yes, 0 = No).
Management	Indicates if person holds a management position (1 = Yes, 0 = No).
Back-end	Indicates if person works on back-end development (1 = Yes, 0 = No).
Front-end	Indicates if person works on front-end development (1 = Yes, 0 = No).

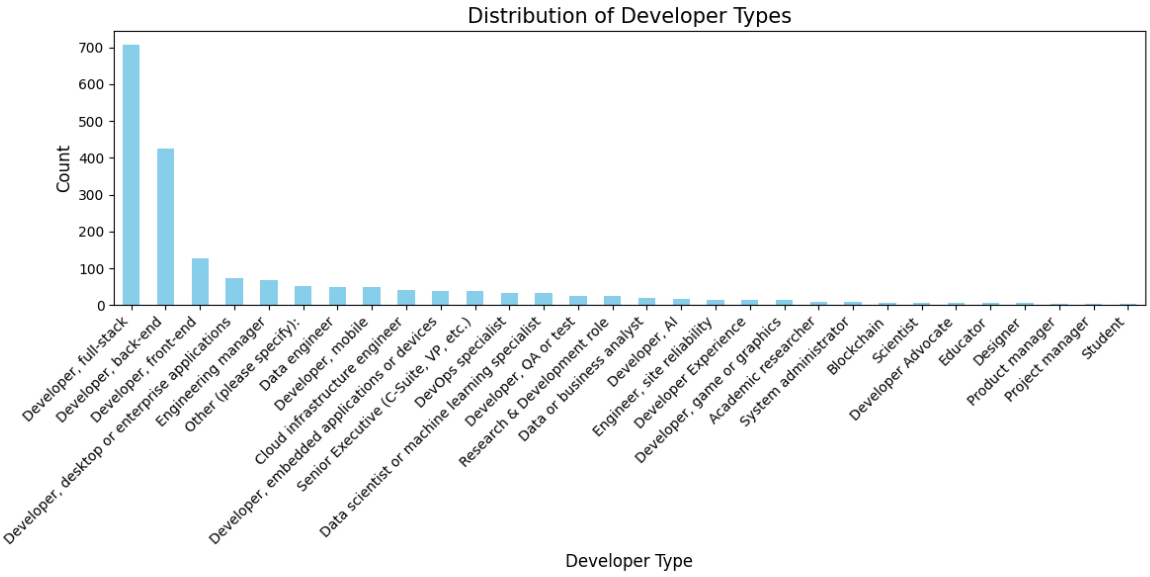


Fig. 9: Before grouping jobs

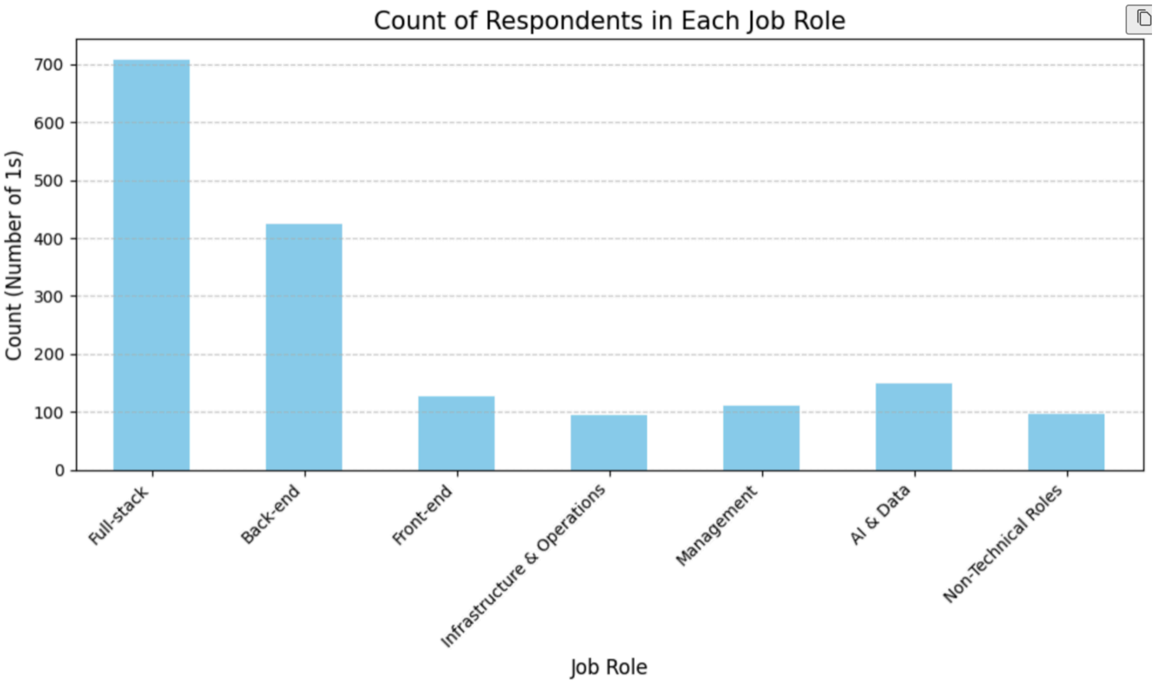


Fig. 10: After grouping jobs