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.
- **Guiding Industry Development:** Provide valuable insights to shape education and training programs for current and future tech professionals.
- **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.

- **Software Developers:** Interested in comparing their salaries with industry benchmarks and understanding factors that could enhance their earning potential.
- **Academic and Industry Researchers:** Focused on labor market trends and the economic factors influencing tech professionals.
- **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.

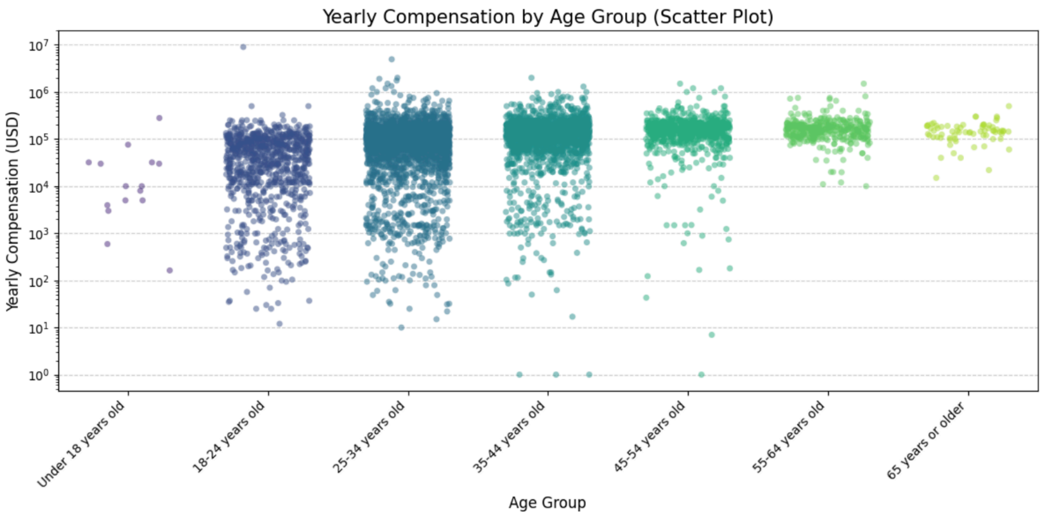


Fig. 1 Before the data cleaning and balancing

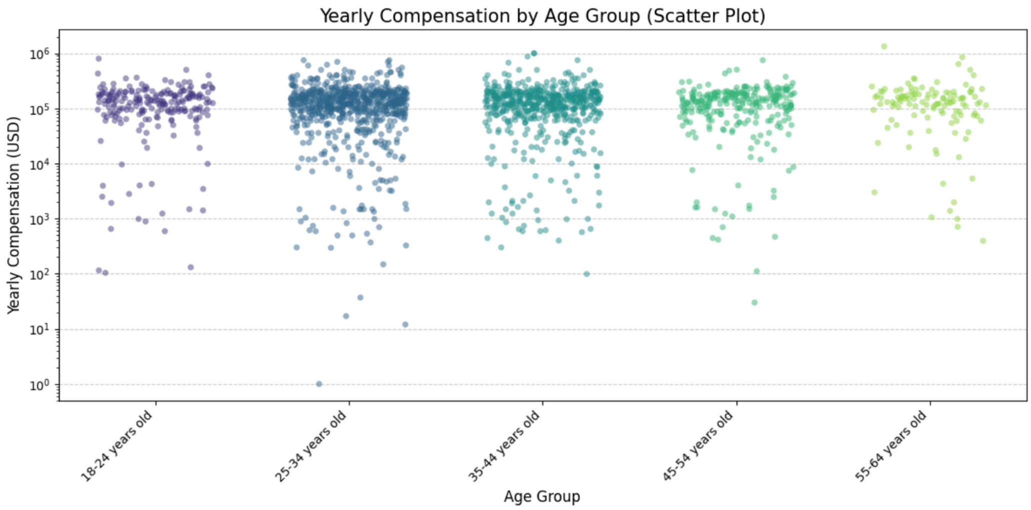
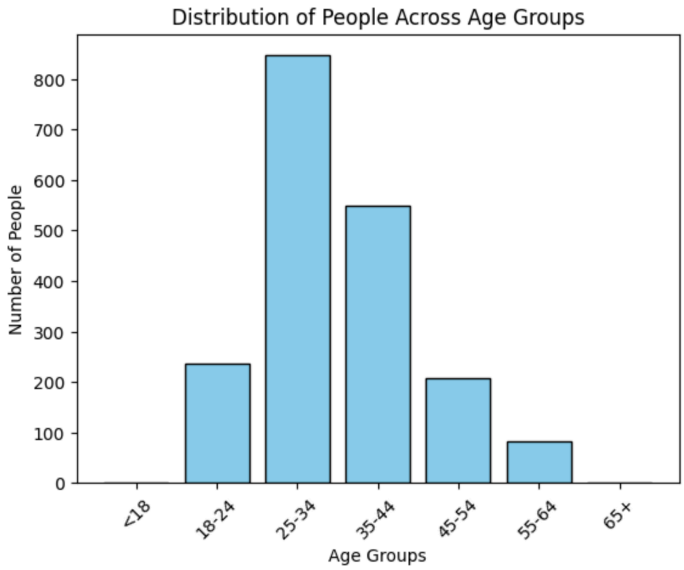


Fig. 2 After the data cleaning and balancing

Our original data consisted of many separate columns for each of the questions from the survey a lot of which were multiple choice or just non-numerical ordinal scale variables such as age described as gaps (i.e under 18 or 18-24). So to run a linear regression model we either had to introduce a lot of new dummy variables for each of the possible responses or neglect the fact that we deal with a non-numerical variable and set each of the options equal to some number.

We attempted to use the first method as it at least preserves the necessary linearity and ended up with a data frame of 150+ columns, most of them were actually useless since the dataset contained few data for some classes, for example there was only one person younger then 18.

So to make the model more meaningful we removed all the insignificant columns and found ourselves with values for age, job, experience, education, AI usage, Remote working, areas of interest etc.
The full list with the meaning of it is provided in the Appendix.



One of the most challenging part in dealing with data from surveys: since some people can give absolutely untruthful answers (like earning 2000000$ per year and having experience of 3 years) and it creates outliers which skew the statistics.

That is why we removed outliers not only using the 3-sigma rule to make tails of plots less stretched out, but also did manual cleaning to make data more meaningful and more convenient to work with.

We have also **deleted the rows with part-time workers and freelancers,** since it would complicate everything. We do not have concrete data about work hours of people, thus we consider that everybody who is not a part-time worker or a freelancers works full-time. Our “story” of analysis about investigating people, who are “fully” involved into their job.

Another approach which was essential for our dataset is to group jobs in subjective but meaningful as much as it is possible way. At least to avoid 30+ dummy variables indicating only the position. You can have a look on the result on Fig. 3 and Fig. 4 below.

It is important to mention that we are investigating people who are not only involved in IT industry directly, but people who are more associated with managing and technical roles. So we’ve decided to remove group of “Non-Technical roles” which consists of Developer Advocate, Marketing or sales professional , Educator position etc.

By reducing dummy variables to a core set of well-populated predictors (age groups, experience measures, AI use, learning methods, remote status, education, and broad role categories), treating implausible salary–experience pairs via a 3-sigma rule plus manual review, excluding part-timers and freelancers to focus on full-time professionals, and rebalancing classes to correct original imbalances, we achieved a cleaner, more stable dataset whose resulting regression coefficients will be both more reliable and easier to interpret.

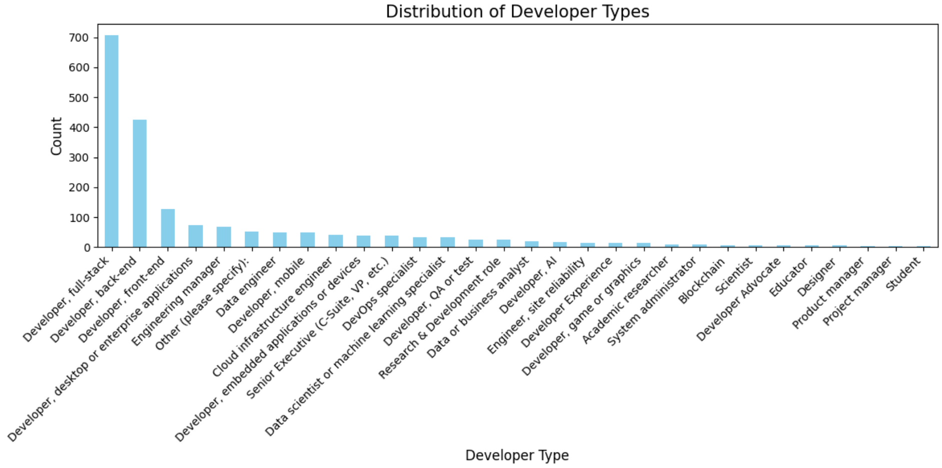


Fig. 3 Before grouping jobs

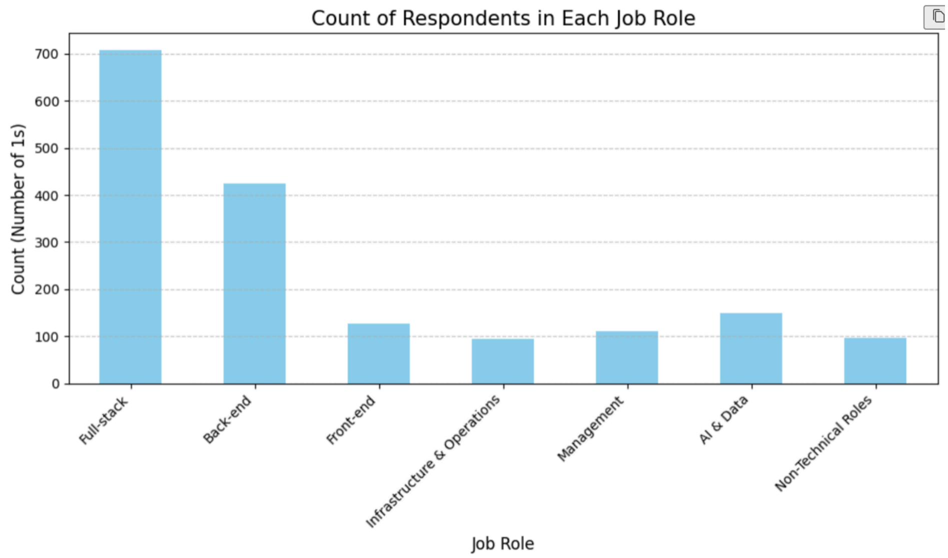


Fig. 4 After grouping jobs

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.

Professional Experience and Income

Based on our survey, although the data shows an extra year of work experience is linked with about $779 more in annual salary, the evidence isn't strong enough to confirm that this effect is reliable.

Analysis:

Null Hypothesis (H_0): Years of professional experience do not affect annual income.

Alternative Hypothesis (H_1): Years of professional experience positively affect annual income.

	coef	std err	t	P>|t|	[0.025	0.975]
WorkExp	1144.1376	307.653	3.719	0.000	540.975	1747.300

Interpretation:

The coefficient suggests that, holding other factors fixed, each additional year of professional experience is associated with an average increase of approximately **$1144.13** in annual income.

The **p-value** of 0.000 indicates that this relationship is not statistically significant at conventional levels (e.g., $\alpha = 0.05$). This means we do not have sufficient evidence to conclude that work experience has a definitive impact on annual income in this model.

Education and Salary Impact

Due to the results, people with more formal education earn considerably more each year — ranging from about $24,410 extra for a Bachelor's to $41,760 extra for a Professional degree—and this finding is robust enough that it's very unlikely to be due to random chance.

Analysis:

Null Hypothesis (H_0): Higher education levels do not influence annual income.

Alternative Hypothesis (H_1): Higher education levels lead to an increase in annual income.

	coef	std err	t	P>|t|	[0.025	0.975]
Bachelor Degree	1.123e+04	2914.712	3.852	0.000	5513.535	1.69e+04
Master Degree	1.637e+04	3359.586	4.872	0.000	9780.422	2.3e+04
Professional Degree	2.755e+04	5767.316	4.778	0.000	1.62e+04	3.89e+04

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).

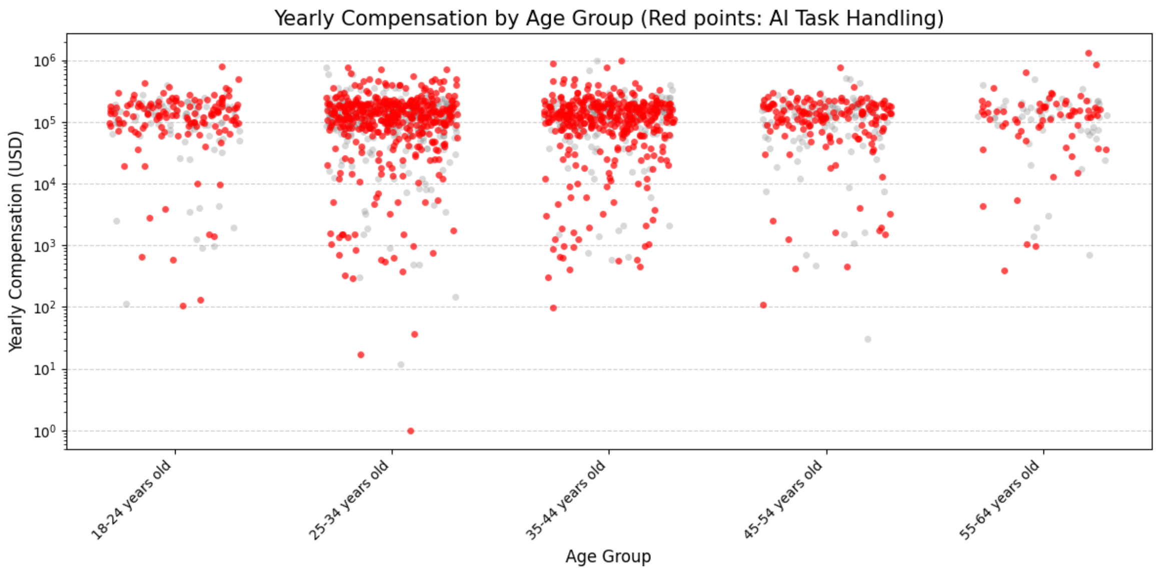
Interpretation:

The regression analysis indicates that higher education levels have a statistically significant positive effect on annual income:

Bachelor's Degree: Associated with an average increase of $11,230 in annual income ($p = 0.000$).
Master's Degree: Associated with an average increase of $16,370 in annual income ($p < 0.000$).
Professional Degree: Associated with an average increase of $27,550 in annual income ($p = 0.000$).

These results confirm that formal education is strongly associated with higher earnings, even after controlling for other variables. The p-values for all education levels are below the conventional threshold of 0.05, indicating that these relationships are statistically significant.

AI Adoption and Earnings



According to the answers, people who invest in paid AI learning tools see an additional boost of about $11,780 in annual salary, whereas free AI tools doesn't significantly affect earnings—suggesting that investing in premium AI resources may enhance skills that lead to higher income.

Analysis:

Null Hypothesis (H_0): Utilising free/paid AI tools in professional tasks does not affect annual income.

Alternative Hypothesis (H_1): Utilising free/paid AI tools in professional tasks influences annual income.

	coef	std err	t	P>|t|	[0.025	0.975]
AI-powered search (free)	-6429.5535	4560.275	-1.410	0.159	-1.54e+04	2514.308
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:

AI-powered search (free): The coefficient is **-6,430** with a **p-value of 0.159**, suggesting a negative association with annual income; however, this relationship is not statistically significant.

AI-powered search (paid): The coefficient is **11,770** with a **p-value of 0.013**, indicating a statistically significant positive impact on annual income for individuals investing in paid AI tools.

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. Here are what particularly AI tools we are talking about (from most frequent answers):

Country Differences

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

Analysis:

Null Hypothesis (H_0): Individuals in the IT industry in Ukraine earn the same as their counterparts in other countries.

Alternative Hypothesis (H_1): Individuals in the IT industry in Ukraine earn differently than their counterparts in other countries.

	coef	std err	t	P>|t|	[0.025	0.975]
InUkraine	-9.397e+04	3223.886	-29.147	0.000	-1e+05	-8.76e+04

Note: The coefficient for 'InUkraine' is statistically significant at the p < 0.001 level.

Interpretation:

The coefficient for the 'InUkraine', which indicates whether the human works in Ukraine variable is **-93,970**, with a **p-value < 0.000**, indicating that, all else being equal, IT professionals in Ukraine earn significantly less than their counterparts in other countries. This substantial wage disparity may be attributed to factors such as currency fluctuations and varying economic conditions.

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.

$H_0 : \beta_\mathrm{i} = 0,$

$H_1 : \beta_\mathrm{i} \neq 0,$

$$t \, : \, \frac{\hat{\beta}_\mathrm{i}}{\mathrm{SE}\left(\hat{\beta}_\mathrm{i}\right)} \; \sim \; t_{n-K-1}$$

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 (β=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:

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

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 listed in the Appendix enter the model simultaneously.

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                 Remote   No. Observations:                 6825
Model:                          Logit   Df Residuals:                     6811
Method:                           MLE   Df Model:                           13
Date:                Wed, 30 Apr 2025   Pseudo R-squ.:                 0.07456
Time:                        18:23:51   Log-Likelihood:                -2239.0
converged:                       True   LL-Null:                       -2419.3
Covariance Type:            nonrobust   LLR p-value:                 4.194e-69
========================================================================================
                           coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                   -0.2868      0.199     -1.440      0.150      -0.677       0.104
YearsCodePro             0.0407      0.005      7.636      0.000       0.030       0.051
JobSat                   0.0383      0.018      2.139      0.032       0.003       0.073
Developer                0.5794      0.133      4.343      0.000       0.318       0.841
25-34 years old          0.8156      0.104      7.814      0.000       0.611       1.020
```

```
45-54 years old              0.6916      0.159      4.347      0.000      0.380      1.004
35-44 years old              0.9345      0.116      8.061      0.000      0.707      1.162
Academic                    -0.3048      0.133     -2.294      0.022     -0.565     -0.044
Online Courses               0.3581      0.080      4.460      0.000      0.201      0.515
AI-powered search (paid)     0.4312      0.167      2.584      0.010      0.104      0.758
InUkraine                    1.4743      0.171      8.609      0.000      1.139      1.810
Management                  -0.2713      0.163     -1.665      0.096     -0.591      0.048
Back-end                     0.6992      0.126      5.558      0.000      0.453      0.946
Front-end                    0.5510      0.213      2.590      0.010      0.134      0.968
============================================================================================
```

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 = 0.0857, p = .001), indicating more satisfied employees are likelier to work remotely.
- **Age Groups**: Compared to 18–24 yrs, all older cohorts (25–34, 35–44, 45–54, 55–64) have significantly higher remote-work odds (e.g. 55–64 yrs coef = 2.8497, p < .001), suggesting experience and seniority boost remote eligibility.
- **Geographic Location**: Being based in Ukraine strongly increases remote-work probability (coef = 1.7198, p < .001), reflecting cross-border outsourcing trends.
- **Company Size**: Employees at medium-sized firms are less likely remote than those at small firms (coef = –0.4360, p = .030), implying smaller organizations may offer greater flexibility.
- **Role Category**: Management (coef = 0.9629, p = .041), Full-stack (0.5466, p = .013), Back-end (0.6354, p = .018), and Front-end (0.8789, p = .033) roles significantly increase remote-work odds, highlighting core technical and leadership positions as most adaptable to remote arrangements

Conclusions

Our findings indicate that **formal education has a strong and statistically significant impact on earnings**, with higher degrees leading to higher salaries. **AI adoption in professional tasks is also positively associated with increased income, particularly for developers using paid AI tools.** However, **work experience alone does not show a significant effect** when controlling for other factors. Additionally, **individuals from Ukraine earning significantly less than their counterparts elsewhere.**

These results highlight the evolving landscape of IT careers, where traditional factors like experience are becoming less predictive, while AI skills and formal education play an increasingly important role.

What we have now is a set of significant variables (according to p-value) and aside from this, the model includes some constants, which, in our opinion, better to leave, even when they are not significant to highlight what “our story” of the research is about.

The logistic-regression model applied a logit model to estimate the probability of remote work via maximum-likelihood estimation. With a highly significant likelihood-ratio test (p < .0001), the analysis demonstrates that both individual attributes (e.g., job satisfaction, age, location) and organizational factors (company size, role category) jointly influence remote-work probability.

Appendix

You can find shortened version of code with data preprocessing and models build on GitHub.

As we've mentioned, this 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 under it the meaning is provided:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     ConvertedCompYearly   R-squared:                       0.404
Model:                             OLS   Adj. R-squared:                  0.401
Method:                  Least Squares   F-statistic:                     135.6
Date:                 Wed, 30 Apr 2025   Prob (F-statistic):               0.00
Time:                         18:06:40   Log-Likelihood:                -52964.
No. Observations:                 4226   AIC:                         1.060e+05
Df Residuals:                     4204   BIC:                         1.061e+05
                    Df Model:                          21
                    Covariance Type:            nonrobust
=============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
const                        8.18e+04   6902.447     11.851      0.000    6.83e+04    9.53e+04
YearsCodePro                2525.2856    302.657      8.344      0.000    1931.919    3118.652
WorkExp                     1144.1376    307.653      3.719      0.000     540.975    1747.300
JobSat                      2049.8738    493.326      4.155      0.000    1082.695    3017.052
Developer                   1.543e+04   4639.352      3.326      0.001    6336.274    2.45e+04
45-54 years old            -2.757e+04   5070.884     -5.437      0.000   -3.75e+04   -1.76e+04
35-44 years old             -852.4923   2992.405     -0.285      0.776   -6719.188    5014.203
55-64 years old            -5.959e+04   7471.690     -7.976      0.000   -7.42e+04   -4.49e+04
Remote                      1.453e+04   3432.190      4.234      0.000    7801.366    2.13e+04
Academic                   -1.359e+04   3947.230     -3.443      0.001   -2.13e+04   -5851.193
Job Training                -865.2351   2159.998     -0.401      0.689   -5099.972    3369.502
Online Courses             -1.852e+04   2175.699     -8.514      0.000   -2.28e+04   -1.43e+04
AI-powered search (free) -6429.5535   4560.275     -1.410      0.159   -1.54e+04    2514.308
AI-powered search (paid)  1.177e+04   4752.763      2.477      0.013    2449.377    2.11e+04
Small                      -3.895e+04   2520.935    -15.450      0.000   -4.39e+04    -3.4e+04
Medium                     -2.409e+04   2588.762     -9.306      0.000   -2.92e+04    -1.9e+04
InUkraine                  -9.397e+04   3223.886    -29.147      0.000      -1e+05   -8.76e+04
Bachelor Degree             1.123e+04   2914.712      3.852      0.000    5513.535    1.69e+04
Master Degree               1.637e+04   3359.586      4.872      0.000    9780.422     2.3e+04
Professional Degree         2.755e+04   5767.316      4.778      0.000    1.62e+04    3.89e+04
Management                  3.891e+04   4761.919      8.172      0.000    2.96e+04    4.82e+04
Back-end                    1.385e+04   2637.689      5.249      0.000    8674.842     1.9e+04
Front-end                   1545.4747   4555.610      0.339      0.734   -7385.929    1.05e+04
==============================================================================
Omnibus:                     1486.732   Durbin-Watson:                   2.004
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             8107.593
Skew:                           1.585   Prob(JB):                         0.00
Kurtosis:                       9.000   Cond. No.                         184.
==============================================================================
```

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).