Exploratory Data Analysis Report

Tina Yi, Shaila Guereca, Faraz Jawed, Keon Nartey

Overview

Stack Overflow serves as a central hub for computer programmers worldwide, offering a platform for both questions and answers as well as acting as a comprehensive resource for programming knowledge (Resource - https://insights.stackoverflow.com/survey) Each year, Stack Overflow conducts an extensive online survey. The 2023 edition ran from May 8 to May 19, attracting participation from approximately 91,000 software engineers across 185 countries. Following rigorous privacy and data consent checks, 89,184 responses were deemed fit for analysis. This survey encapsulated 84 diverse variables spanning seven primary sections, including Basic Information, Education, Work and Career, Technology and Tech Culture, Stack Overflow Usage, Artificial Intelligence, and Professional Developer Insights. For the purposes of our study, we narrowed down to a dataset of 40,734 entries based on variable selection.

Research Questions:

1. **Compensation Prediction**: How can we estimate a developer's yearly compensation? We aim to build a predictive model factoring in age, work mode, years of coding experience outside formal education, educational background, total work experience, and proficiency across different programming languages.

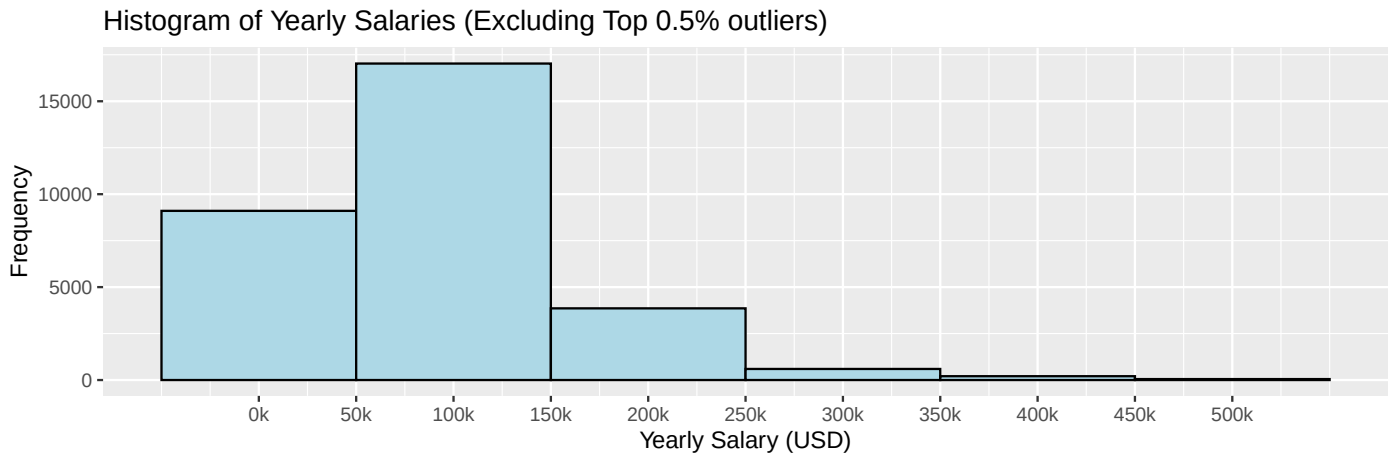
2. **Job Satisfaction Determinants**: What elements shape a developer's job satisfaction? We are investigating the influence of variables like age, organization size, yearly compensation, work mode, and the degree of artificial intelligence integration in their professional tasks.

Characteristic	**N = 40,734**[1]
Converted Yearly Compensation	74,963 (44,304, 120,000)
Unknown	9,598
Age	
18-24 years old	4,913 (12%)
25-34 years old	18,769 (46%)
35-44 years old	11,462 (28%)
45-54 years old	4,053 (9.9%)
55-64 years old	1,250 (3.1%)
65 years or older	184 (0.5%)
Prefer not to say	52 (0.1%)
Under 18 years old	51 (0.1%)
Remote Work	
Hybrid (some remote, some in-person)	17,574 (43%)
In-person	6,287 (15%)
Remote	16,823 (41%)
Unknown	50
Years of Coding Experience Not Including Education	8 (4, 15)
Unknown	2,101
Education	
Associate degree (A.A., A.S., etc.)	1,306 (3.2%)
Bachelor's degree (B.A., B.S., B.Eng., etc.)	19,275 (47%)
Master's degree (M.A., M.S., M.Eng., MBA, etc.)	11,034 (27%)
Primary/elementary school	180 (0.4%)
Professional degree (JD, MD, Ph.D, Ed.D, etc.)	1,821 (4.5%)
Secondary school (e.g. American high school, German Realschule or Gymnasium, etc.)	1,868 (4.6%)
Some college/university study without earning a degree	4,801 (12%)
Something else	449 (1.1%)
Work Experience	9 (5, 16)
Unknown	146
Number of Coding Languages	5.0 (3.0, 7.0)
Organization Size	
1,000 to 4,999 employees	4,550 (12%)

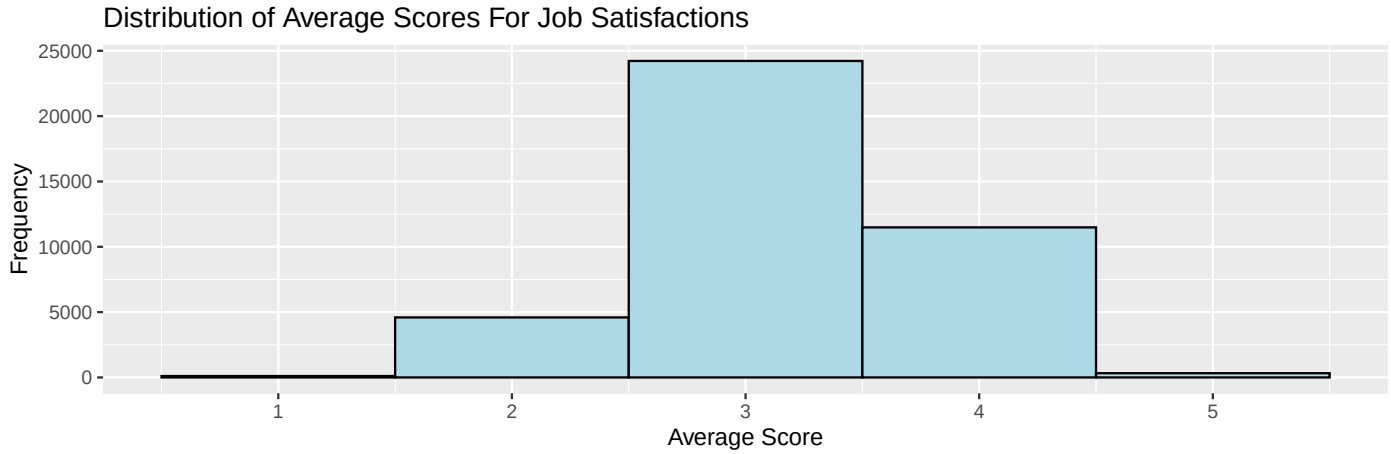
10 to 19 employees	3,271 (8.6%)
10,000 or more employees	4,969 (13%)
100 to 499 employees	7,756 (20%)
2 to 9 employees	3,680 (9.6%)
20 to 99 employees	8,675 (23%)
5,000 to 9,999 employees	1,659 (4.3%)
500 to 999 employees	2,844 (7.4%)
I don't know	553 (1.4%)
Just me - I am a freelancer, sole proprietor, etc.	228 (0.6%)
Unknown	2,549
AI Usage	
No, and I don't plan to	12,143 (30%)
No, but I plan to soon	11,044 (27%)
Yes	17,547 (43%)
Job Satisfaction	
Not satisfied	908 (2.2%)
Neutral	13,409 (33%)
Satisfied	26,417 (65%)

[1]Median (IQR); n (%)

Ouput Graphs For Outcome Variables



The histogram displays the distribution of the Converted Yearly Compensation variable, excluding potential outliers. The majority of respondents indicated salaries in the 50k-150k range, though some reported earnings as high as 500-600k, and a few even in the millions (not shown in the histogram as outliers).



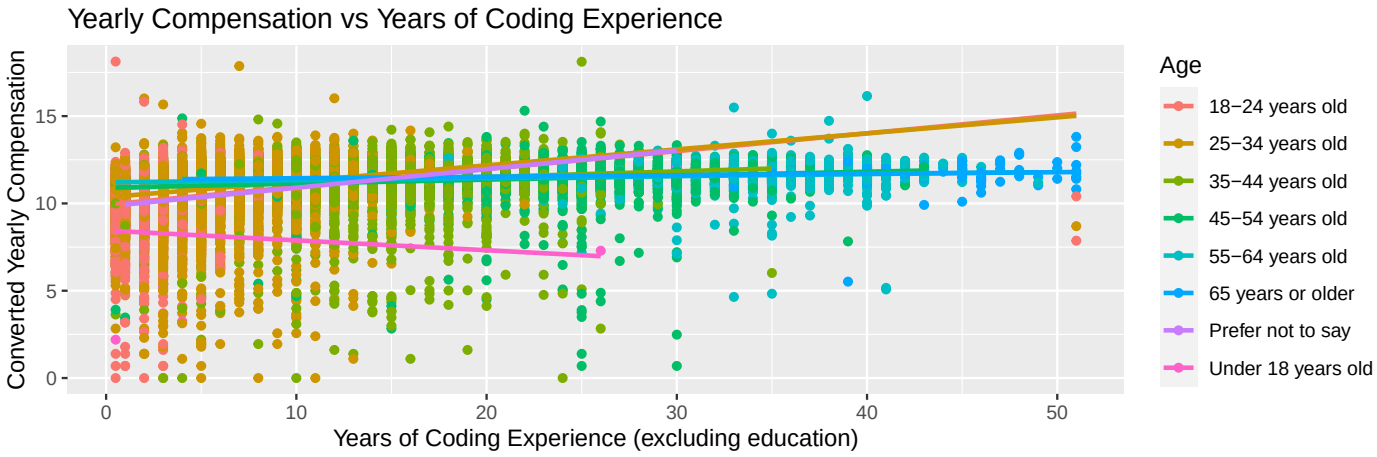
The histogram illustrates the distribution of the average scores for the Job Satisfaction variable. The scores tend to follow a near-normal distribution, with the bulk of observations falling within a standard deviation of the mean (approximately 3). Based on these average scores, we then categorized the Job Satisfaction variable into three groups: Satisfied, Not Satisfied, and Neutral.

Compensation Prediction - Multilinear Regression

We are examining the relationship between continuous/discrete predictor variables and the outcome variable for the compensation prediction multi-linear regression, using categorical variables to color-code the lines in the plot.

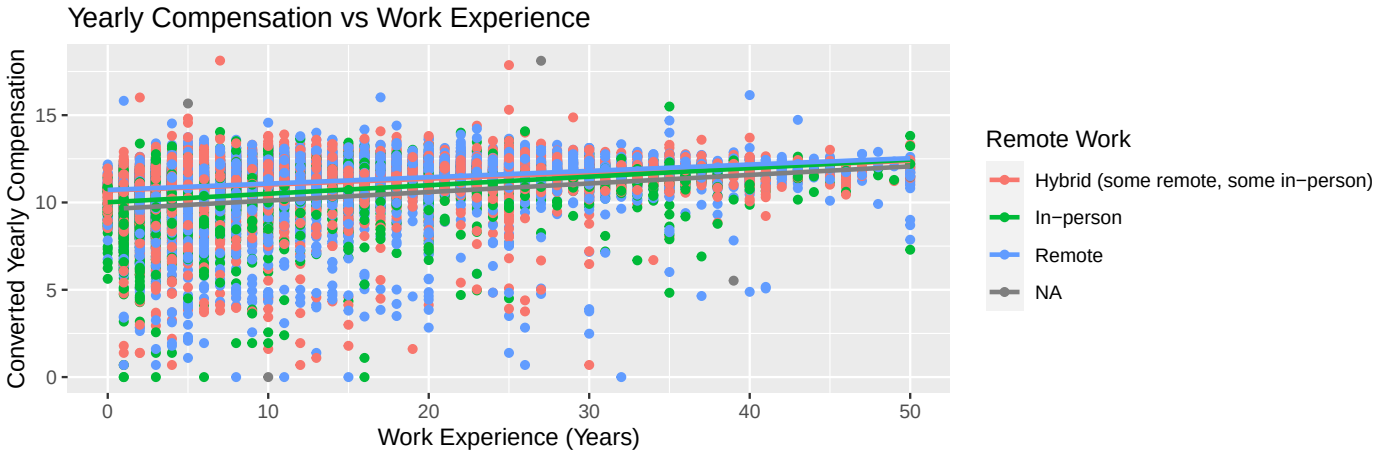
Relationship between Years of Coding Experience and Converted Yearly Compensation

Generally, as coding experience increases, converted yearly compensation also rises across all age groups. However, this trend is not observed in individuals under 18 years old; For every category of work type, there's a noticeable increase in the converted yearly compensation as years of coding experience grow; A consistent rise in converted yearly compensation is seen across all education levels as coding experience advances.



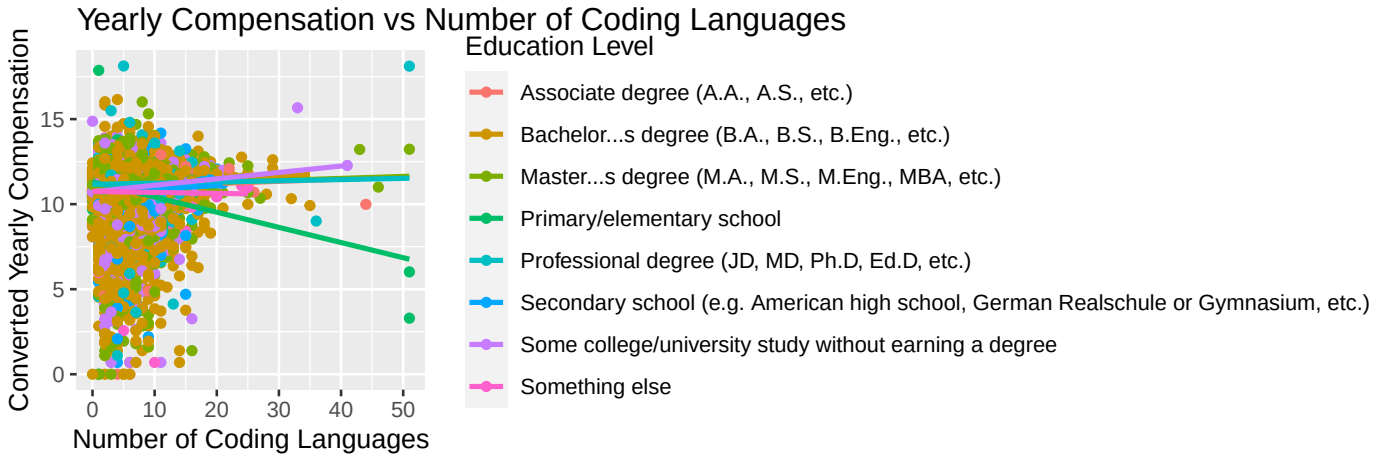
Relationship between Work Experience and Converted Yearly Compensation

If we analyze the distribution of remote work across different levels of work experience and compensation, we can observe that it is well-balanced between hybrid, in-person, and remote arrangements. It appears that the mode of work does not significantly influence the salary that individuals are earning.



Relationship between Number of Coding Languages and Converted Yearly Compensation

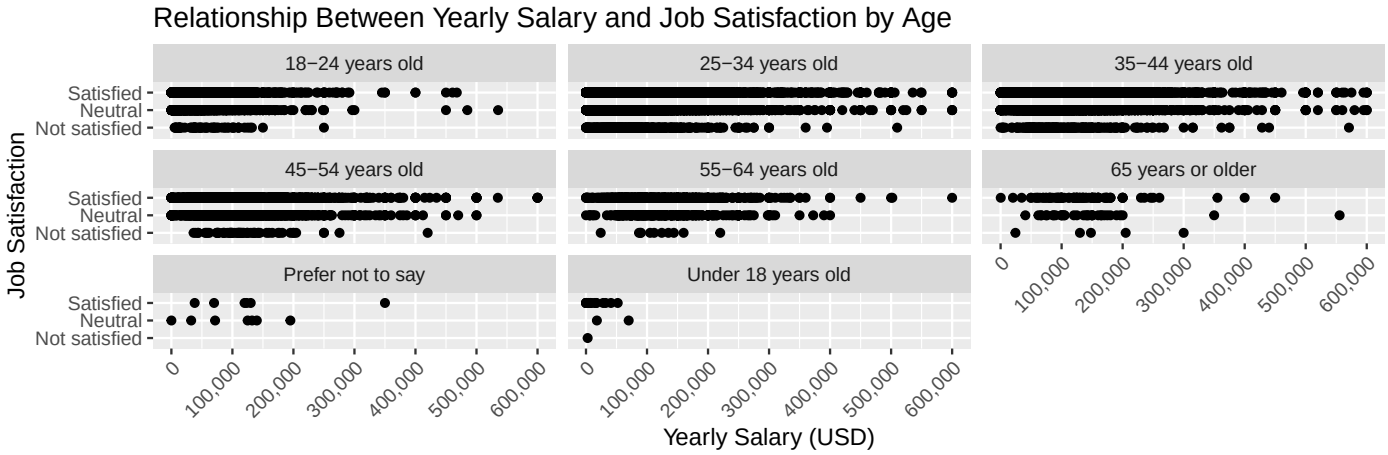
Generally, as the number of coding languages known increases, converted yearly compensation also rises across most age groups. However, this trend is not observed in individuals under 18 years old and for individual's who prefer not to say; For every category of work type, there's a noticeable increase in the converted yearly compensation as the number of coding languages known grows; A consistent rise in converted yearly compensation is seen across most education levels as the number of coding languages known increases. However, this trend is not observed in individuals who have only primary/elementary school degree and those individuals with something else degree.



Job Satisfaction Determinants - Multinomial Logistic regression

In this section, we examine the relationship between the predictor variables and the outcome variable intended for the job satisfaction determinants multinomial logistic regression. The subsequent plots aim to elucidate how these predictor variables interact with the outcome variable, offering insight into the distribution and interrelation of our variables.

Relationship between Age, compensation and Job satisfaction

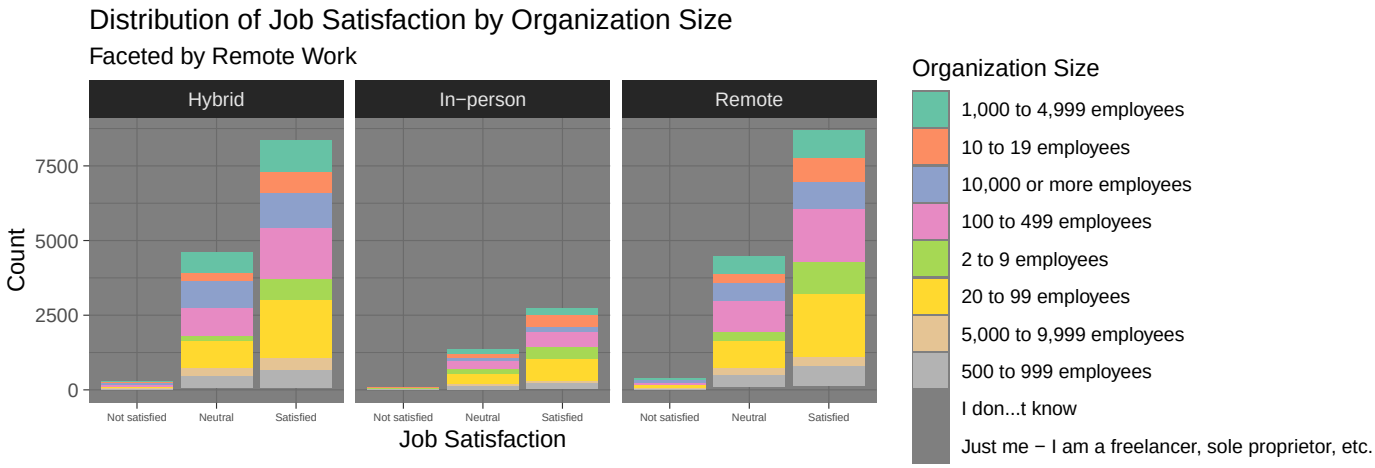


The graph above helps explore the relationship between job satisfaction, compensation and age. It can be seen that for most age-groups, the general trend is that the higher an individual's salary, the more likely are they generally satisfied. For the purpose of better visualization, and to remove the affect of outliers, we decided to plot the relationship between these variables where compensation was capped at 600k USD. Another interesting finding: for ages 65 and older, compensation is not correlated with a person's job satisfaction and there is not a clear trend to be seen.

Relationship between Job satisfaction, Remote Work and AI Usage

This chart visualizes the Job satisfaction with the AI Usage of an individual (whether they use AI already, plan to use it, or don't plan to use it at all). It can be seen that the distribution is fairly even in terms of AI usage, with it being more spread out for people who're satisfied with their job. However for individuals satisfied with their job, their log of yearly compensation tends to be higher, which confirms the notion discussed in the previous plot. Another important thing to see is that the distribution is a lot varied for users who have a clear stance on AI (either they don't use it and don't plan to use it at all or they have already starting using it).

Relationship between Job satisfaction, RemoteWork and Organization size



This plot is to show the relationship between Job Satisfaction and two of our predictor variables: Organization Size and Remote Work. An interesting trend in our data was that the number of In-person workers is less frequent. In terms of organization size, the most common answers were either 20-99 employees or 100-499 employees, and are fairly distributed among all three categories of Job Satisfaction, with the majority of individuals in these organization sizes being generally Satisfied.

Other characteristics

The Stack Overflow Annual Developer Survey encompasses diverse variables spanning developer demographics, experience, tools, practices, and more. Broadly categorized, these variables cover:

Demographics: Details like age, gender, residence country, education, job title, salary, etc; **Experience**: Programming languages, technologies, and methodologies adopted; **Tools**: Programming instruments, cloud platforms, and database systems utilized; **Practices**: Approaches to learning, open-source contributions, and community participation; **Stack Overflow Usage**: Usage frequency, account status, and community involvement; **Artificial Intelligence**: Engagement with AI tools, their perceived benefits, and trust in their accuracy.

Potential challenges

In our modeling, a challenge is consolidating 8 related questions on job satisfaction into one variable. Each question is scored from 1 (lowest satisfaction) to 5 (highest). For negatively phrased questions, the scale is reversed. We then computed the average across these questions for a unified job satisfaction score. The questions and scale are detailed below:

	Job Satisfaction Survey				
Questions	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
1) I have interactions with people outside of my immediate team.	1	2	3	4	5
2) Knowledge silos prevent me from getting ideas across the organization (i.e., one individual or team has information that isn't shared with others).	5	4	3	2	1
3) I can find up-to-date information within my organization to help me do my job.	1	2	3	4	5
4) I am able to quickly find answers to my questions with existing tools and resources.	1	2	3	4	5
5) I know which system or resource to use to find information and answers to questions I have.	1	2	3	4	5
6) I often find myself answering questions that I've already answered before.	5	4	3	2	1
7) Waiting on answers to questions often causes interruptions and disrupts my workflow.	5	4	3	2	1
8) I feel like I have the tools and/or resources to quickly understand and work on any area of my company's code/system/platform.	1	2	3	4	5

For the specified variable, an average of 47,386 missing values exists per question, often because when a person did not answer one of the questions, they also did not answer the remaining 7 questions. We also observed non-numeric entries like 'less than 1 year' in the 'professional programmer experience' variable. Additionally, variable selection challenges arise due to few numeric variables, exemplified by age being categorized (e.g., 'prefer not to say', 'under 18', '65 or older') rather than numerical.

Also to perform our analysis, we created an additional variable that was not originally present in the database. The purpose was to determine the number of programming languages that each person knew. We counted the number of languages that each individual listed to assess whether knowing more programming languages has an impact on compensation.