29th Nov 2023

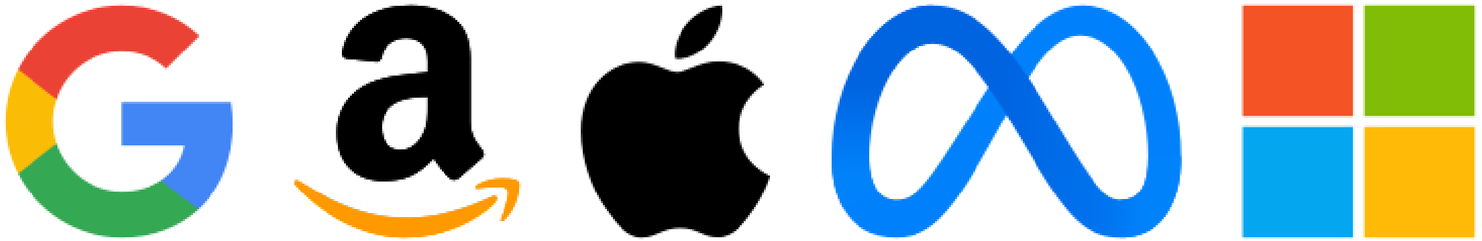
_WHAT MAKES A DEVELOPER HAPPY AND RICH?

Stack Overflow Survey 2023

Keon Nartey, Faraz Jawed, Shaila Güereca, Tina Yi

MOTIVATION

- Most of us will work in the tech industry
- Important to understand work-life balance and expected compensation
- Stack overflow is used by many developers around the world and studying that data can derive meaningful insights about our future work.
- Interesting to know trends among developer responses and inferring meaning from it.



INTRODUCTION

- 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. Each year, Stack Overflow conducts an extensive online survey.
- The 2023 edition ran May 8 - May 19
- **91,000** software engineers from **185** countries
- Following rigorous privacy and data consent checks, **89,184** responses were deemed fit for analysis

Questionnaire

7 primary sections, including **84 variables**

- Basic Information
- Education
- Work and Career
- Technology and Tech Culture
- Stack Overflow Usage
- Artificial Intelligence
- Professional Developer Insights

RESEARCH QUESTIONS

Yearly compensation and **Job satisfaction** were our two outcome variables for the two research questions below

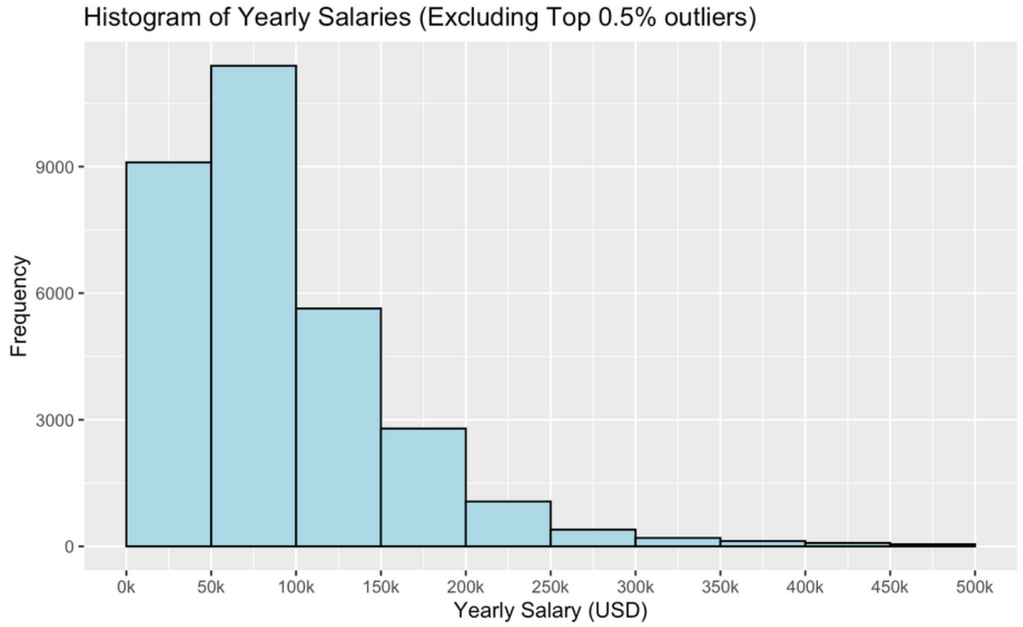
Yearly Compensation

How much would a developer earn based on their answers in the survey?

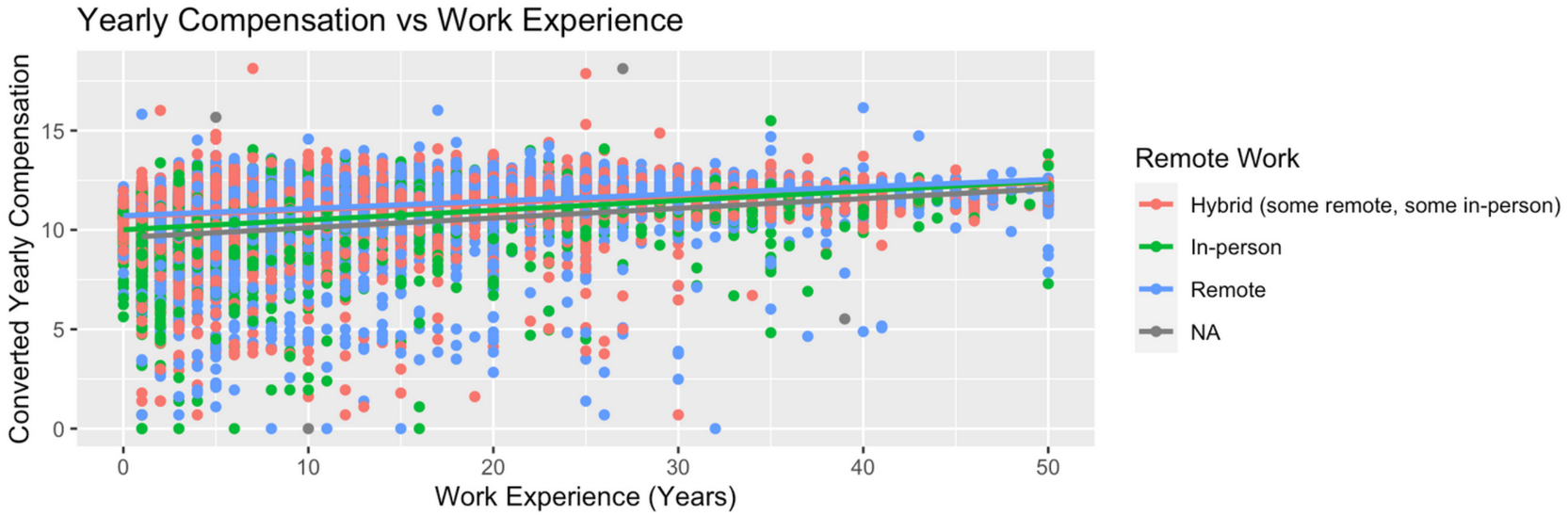
Job Satisfaction

What factors contribute to a developer being satisfied at their job?

EXPLORATORY DATA ANALYSIS(SALARY)

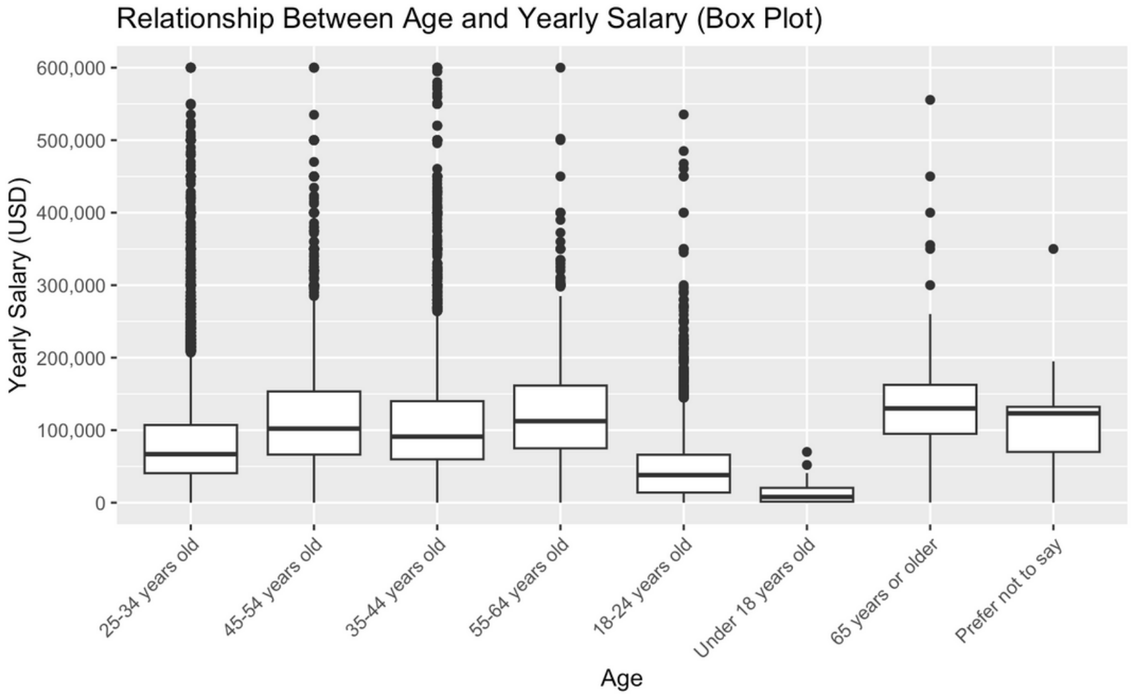


The histogram displays a right-skewness of Converted yearly salary, excluding outliers. The majority of respondents indicated salaries in the 50 - 150K range, with high earners being in 500-600K with a few millions not shown.



The distribution of remote work across different levels of work experience and compensation is well-balanced between hybrid, in-person, and remote arrangements. However, the kind of work does not significantly influence the salary that individuals are earning.

EXPLORATORY DATA ANALYSIS(JOB SATISFACTION)

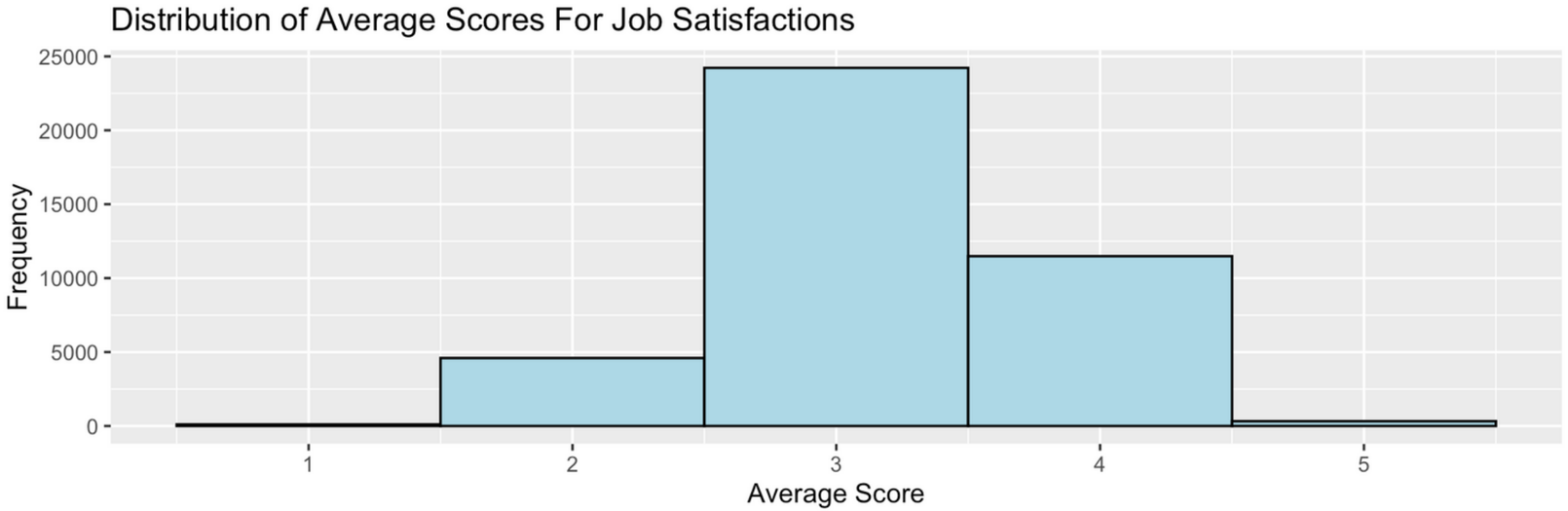


Our aim was to investigate the interplay between job satisfaction, compensation, and age. We observed a consistent trend where higher salaries correlated with greater overall satisfaction. To enhance visualization and mitigate outlier effects, we capped the salary at 600K USD. As anticipated, the data revealed that, on average, as individuals aged, their median salary tended to rise, aligning with an increased likelihood of job satisfaction. Notably, individuals under 18 had the lowest median pay, followed by those aged 18-24, while individuals aged 65 and older boasted the highest median salary.

The distribution of the average scores for job satisfaction tends to follow a near-normal distribution, with bulk of the observations falling within a standard deviation of the mean (approximately 3).

We categorized into 3 groups based on the quantiles.

- Not Satisfied (1.0 - 2.875)
- Neutral (2.875 - 3.625)
- Satisfied(3.625 - 5. 0)



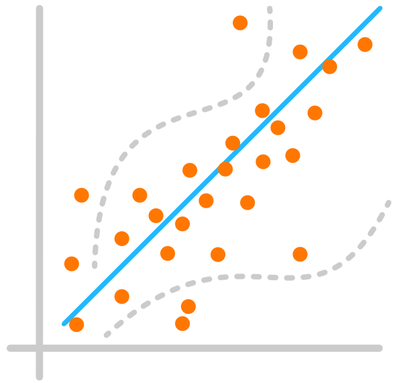
RESEARCH QUESTION 1

MULTIPLE LINEAR REGRESSION

Methods

Prediction: Aiming to forecast a developer's annual compensation (USD)

Approach: Multiple Linear Regression



Variables

Selected Predictor Variables:

- Age
- Type of work (remote, in-person, hybrid)
- Years of coding experience
- Education level
- Years of working experience
- Number of programming languages developers knew and worked with

Interaction Terms : Education level and Years of coding experience

Model Fitting

- Multicollinearity Check (VIF)
- Multiple Linear Regression Model Assumptions: Linearity, Equal variance of errors, Normality of errors, Independence of errors
- Influential Points

Model Assessment

- Cross Validation (Model 1 VS Model 2)
- Root Mean Square Error (RMSE)
- R^2

MODEL FITTING

01. Low Multicollinearity

The VIF for the preliminary main regression (converted yearly compensation ~ the rest of the variables) shows that most of the variables have low multicollinearity, with the adjusted GVIF close to 1, suggesting that they are not correlated with each other.

02. Log-transformation

Our final multilinear regression model has log-transformed yearly compensation as the outcome variable and every predictor variable with an interaction term between education level and years of coding experience.

03. Influential Points

This model is based on an updated dataset with influential points excluded for more accurate results.

MODEL ASSESSMENT

Main Model: With Interaction Term

Metric	Value
RMSE	1.0976539
R^2	0.1437715
MAE	0.6978818

Cross Validation

Secondary Model: Without Interaction Term

Metric	Value
RMSE	1.0991778
R^2	0.1423495
MAE	0.6987709

We choose the main model with a lower RMSE of **1.097** compared to the secondary model , indicating it's a better model with predictions generally closer to the real outcomes.

A low R-squared value of **0.14** indicates that the main model explains only a small portion of the variability in the response variable. However, in our case of dealing with human factors and complex systems influencing salaries, a lower R^2 might still be considered acceptable due to the inherent variability in the data.

PREDICTING FARAZ'S SALARY

Name

Random MIDSter

Age

26 (upon graduating)

Education

Master's Degree

Work Experience

1 year

Professional Coding Experience

2 years

Work Environment

Hybrid (online + in-person)

Number of coding languages

6

How much salary would a random MIDS graduate earn based on our model?

Predicted Yearly Compensation: **$50,816.03 USD**.

Is the market bad? Or is it our model?

RESEARCH QUESTION 2

ORDINAL REGRESSION

Methods

Inference: Elements thats shape a developer's job satisfaction metric on an ordinal scale ("Satisfied", "Neutral", "Dissatisfied")

Approach: Ordinal Regression

Variables

Selected Predictor Variables:

- Age
- Type of work
- Yearly compensation (USD)
- Usage and incorporation of Artificial Intelligence to the job
- Organization Size

Interaction Terms : Age and Yearly compensation.

Model Assessment: Proportional odds assumption

We compared the predicted probabilities obtained from the more precise **multinomial model** to those obtained from the **ordinal model**.

Multiple Linear Regression

Not satisfied	Neutral	Satisfied
0.17	0.50	0.33
0.21	0.51	0.28
0.25	0.51	0.24
0.30	0.50	0.20
0.32	0.52	0.17
0.32	0.51	0.16
0.36	0.50	0.13
0.38	0.47	0.15

Ordinal Model

Not satisfied	Neutral	Satisfied
0.18	0.49	0.33
0.21	0.51	0.28
0.25	0.51	0.24
0.30	0.50	0.20
0.33	0.50	0.18
0.33	0.49	0.17
0.38	0.48	0.15
0.38	0.47	0.14

Confusion Matrix

- This model shows an accuracy of **49.9%**, suggesting that the model predicts job satisfaction as well as a coin flip.

Confusion Matrix Ordinal Model

	Not Satisfied	Neutral
Not Satisfied	13	8,593
Neutral	10	15,103
Satisfied	3	6,576

Confusion Matrix Multinomial Model

	Not Satisfied	Neutral	Satisfied
Not Satisfied	11	8,592	3
Neutral	22	15,084	7
Satisfied	10	6,562	7

ORGANIZATION SIZE MATTERS

Smaller organization

=

More satisfaction

The organizations with the smallest group of employees (from 2 to 9 employees). For workers whose companies are from 2 to 9 employees, the odds of being more satisfied (ie. Satisfied or neutral versus dissatisfied) in the job is 2.35 times that of workers who work in larger organization, holding all other variables constant.

Additionally, we can see that as the organization becomes larger, it is more likely to be predicted as 'Not satisfied"

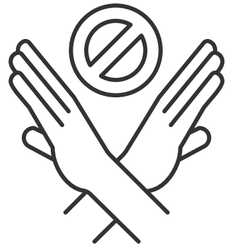
STRENGTHS & LIMITATIONS (SALARY)

Strengths

- Inclusion of a wide range of variables during priori selection which collectively add depth to its predictive capacity.

- Interaction term provides insights into how these factors jointly influence compensation.

Limitations



- Age brackets (e.g., 18-24, 25-34 years) as a categorical variable rather than continuous numerical data limits its precision and could introduce biases, causes unbalance and potentially skewing predictions

- A low R-squared value of 0.14 suggests that the model does not fit the data very well. This could be due to several reasons, such as the model missing important explanatory variables, or the relationship between variables being non-linear while the model is linear.

FUTURE WORK (SALARY)

THE FUTURE IS BRIGHT

Not always tho

Improve data collection and balancing dataset

Improving data collection and processing by obtaining numerical age data and ensuring balanced age representation to address data skewness.

Consider methods like weighted regression or SMOTE for dataset balancing.

Improve conversion of non-numeric variables in data

Revising the conversion of non-numeric data, like professional experience, into numerical values through a detailed scale or alternative techniques to further improve model precision and reduce bias.

Add more variables/explore other advanced models

Add more relevant variables and use a different type of model such as polynomial regressions or GAMs that captures the relationship between variables more effectively.

STRENGTHS, LIMITATIONS AND FUTURE WORK

JOB SATISFACTION

STRENGTHS

Indicated that ‘Neutral’ job satisfaction consistently had the highest predicted probability across diverse scenarios, irrespective of age or organization size. Interestingly, larger organizations showed a trend toward being predicted as ‘Not satisfied’.

LIMITATIONS

Lack of explainability between our predictor variables and outcome (except for Org Size), which makes it hard to infer the actual reasons behind differing satisfaction levels.

FUTURE WORK

Use advanced machine learning models or exploration of new features that might enhance the prediction of Job Satisfaction, collecting more diverse data or additional features could better capture factors that affect job satisfaction.

Thank you!

Only one *easy* question allowed.