What makes a developer happy and how can we predict their salary?

I. Abstract

This final report presents an analysis of the 2023 Stack Overflow survey data involving 89,184 software engineers from 185 countries. The report aims to predict developer compensation and understand the factors influencing job satisfaction. The analysis includes multilinear regression models for compensation prediction and ordinal regression for job satisfaction, exploring variables like age, education, work experience, coding languages, remote work, and AI usage. Exploratory data analysis showcases relationships between these variables and compensation, job satisfaction, and demographics. The multiple linear model regression results with RMSE of 1.097 and R^2 of 0.14 successfully predicts a developer's salary. For job satisfaction, organizational size emerged as a significant determinant, revealing that employees in smaller organizations exhibited higher job satisfaction levels compared to those in larger establishments. For instance, individuals within organizations comprising 2 to 9 employees had 2.35 times higher odds of experiencing job satisfaction (i.e., being satisfied or neutral versus dissatisfied) than their counterparts in larger organizations, while controlling for other variables. These findings emphasize the influential role of organizational size in determining levels of job satisfaction among employees.

II. Introduction

For over a decade, Stack Overflow has been a pivotal source of invaluable insights into the dynamic landscape of the developer community. Renowned as the authoritative voice among analysts, IT leaders, and reporters, Stack Overflow consistently provides cutting-edge perspectives on the ever-evolving developer experience. Developers worldwide turn to this comprehensive report not only to stay abreast of the latest trends but also to gain profound insights into the trajectory of emerging technologies. It also 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 from May 8 to May 19, attracting participation from approximately 91,000 software engineers across 185 countries. This year's survey sought to capture how AI/ML influences developers thinking and their workflows. 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. Embarking on a new career is a significant decision that demands a thoughtful and informed approach. Recognizing the pivotal role that a chosen profession plays in shaping one's life, it becomes imperative to meticulously explore career options and assess key factors such as work-life balance and compensation, particularly in the dynamic field of development. Our research endeavors will focus on illuminating these critical aspects through a series of targeted questions:

- *How can we predict a developer's yearly compensation?*

With this question, we want to understand labor market trends and wage disparities. We want to uncover the variables that influence compensation, such as experience, education, geographic location, and technology proficiency, shedding light on the dynamics of the tech job market.

- *What elements shape a developer's job satisfaction?*

With this question, we want to understand what brings happiness to developers at work. Developers can identify the aspects of their job that contribute to their happiness, enabling them to make informed career decisions and prioritize what matters most and employers can leverage this insight to foster a work environment that ensures the happiness of their developers, ultimately resulting in increased productivity and higher employee retention.

III. Methods

III. Data: Salary

To answer the question "How can we predict a developer's yearly compensation?" We first conducted data analysis and relevant data cleaning. We observed non-numeric entries like 'less than 1 year' in the 'professional programmer experience' variable and we solve this by converting it to a numerical value of .5 year. Additionally, variable selection challenges arise due to few numeric variables, exemplified by age being categorized (e.g., 'under 18') rather than numerical, in that sense we have decided to get rid of those under 18 because our focus is on those who have already entered the workforce and are engaged in the industry. Furthermore, 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.

The initial research inquiry pertains to a prediction model, specifically aiming to forecast a developer's annual compensation based on certain independent variables. For this question we aim to build a predictive multilinear regression model using the following variables:

- *Age:* 18-24 years old, 25-34 years old, 35-44 years old, 45-54 years old, 55-64 years old, 65 years or older, prefer not to say, under 18 years old.
- *Type of work:* Hybrid, In-person, Remote.
- *Years of coding experience:* Less than 1 year, numbers from 1 to 50 years, more than 50 years.
- *Education level:* Associate degree, bachelor's degree, master's degree, elementary school, professional degree, secondary school, some college, something else.
- *Years of working experience:* from 0 to 50 years.
- *Number of programming languages each developer knew and worked with:*from 0 to 51 languages.

III. Model : Salary

This multiple linear regression model is related to a prediction problem where we aim to forecast a developer's annual compensation based on certain independent variables. In terms of variable selection, we will be comparing the model that includes the interaction term with another model that excludes it.

Interaction Terms: We are interested in how the relationship between years of coding experience and a developer's yearly compensation varies across different levels of education, so for this research question we will include an interaction term with educational level and years of coding experience.

For each of the two models, we will perform the following model fitting procedures:

Multicolinearity: We will use Variance Inflation Factor (VIF) to assess multicolinearity among all the predictor variables in this model. High VIF values indicate that it's hard to distinguish the individual effects of predictor variables on the outcome because they are too closely linked.

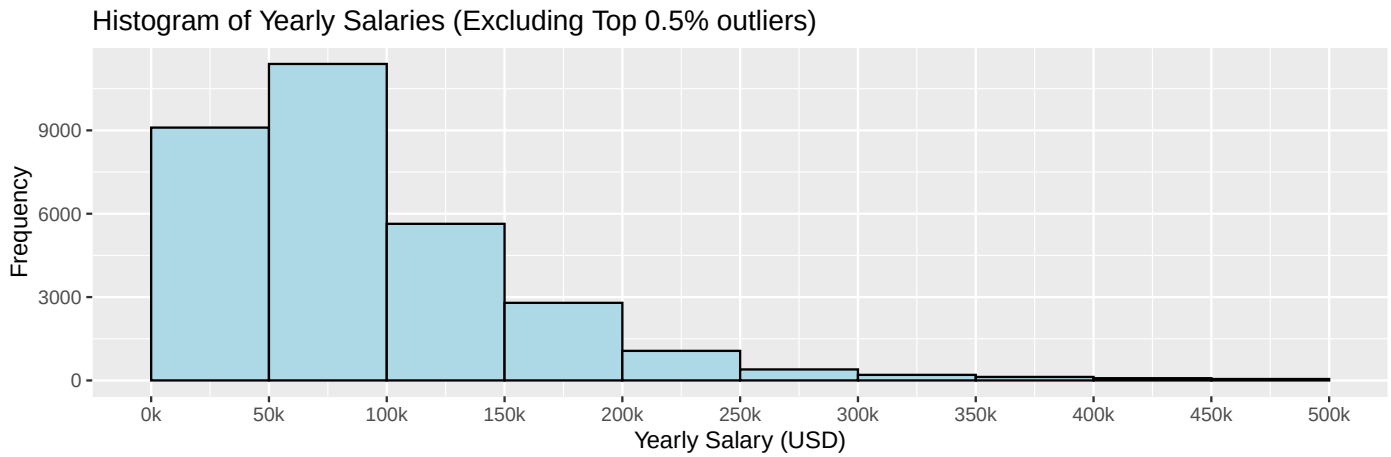
Multiple Linear Regression Model Assumptions: *Linearity*, *Equal variance of errors*, *Normality of errors*, *Independence of errors*. We will assess the assumptions with residual plots and transform predictors or the outcome as needed.

Influential Points: After running the main regression with potential variable transformations to address the multilinear model assumptions, we will observe the cook's distance and leverage in the last two diagnostic plots to check if the model have influential points, which are individual observations that can have a large impact on the model as a whole, that need to be handled.

III. Model Assessment: Salary

We will employ cross-validation on both models and compare their performance using Root Mean Square Error (RMSE). A lower RMSE indicates a more accurate model with predictions closely mirroring actual outcomes. The model with the lowest RMSE will be chosen to predict a developer's annual compensation.

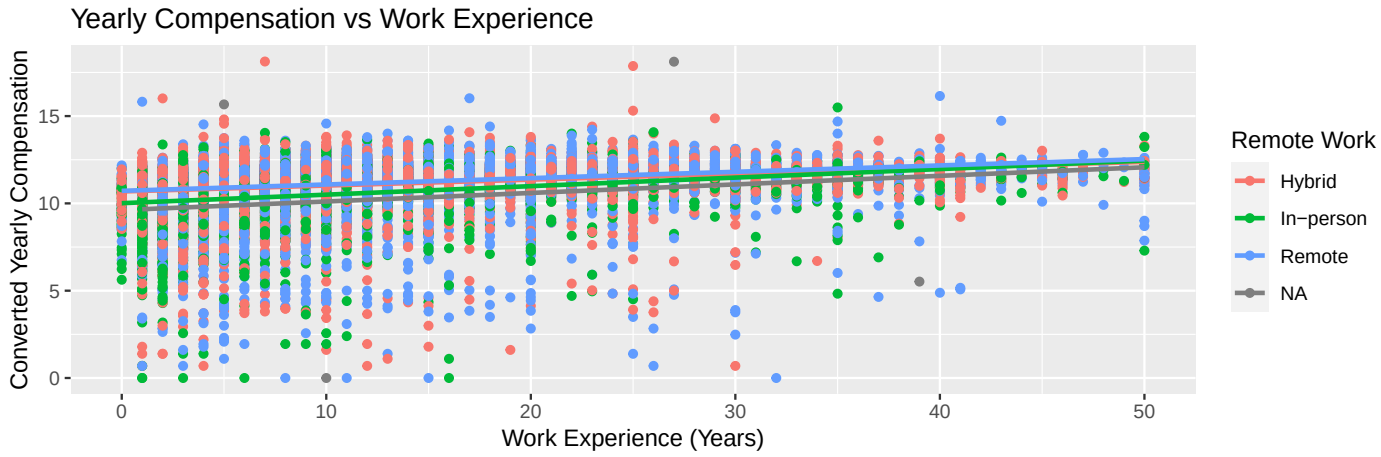
IV. Exploratory Data Analysis Results: Salary



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).

Relationship between Work Experience and Converted Yearly Compensation

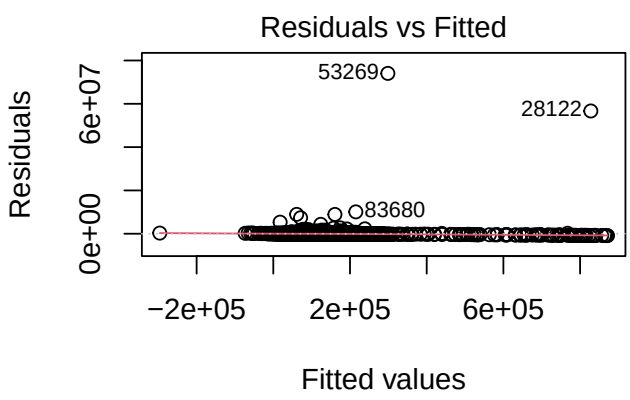
If we analyze the distribution of remote work across different levels of work experience and compensation, we can observe that it is well-balanced between hybrid, in-person, and remote arrangements. It appears that the mode of work does not significantly influence the salary that individuals are earning.

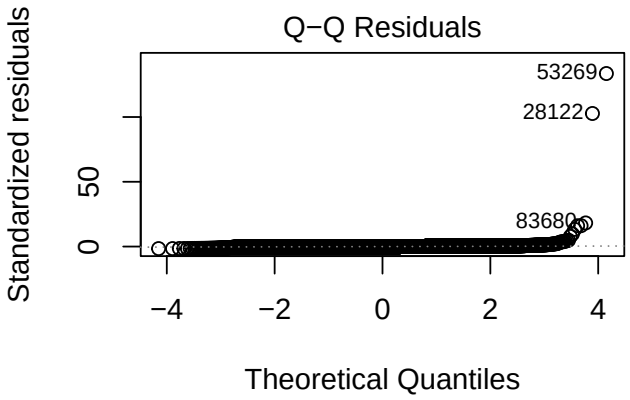


IV. Model Results: Salary

Multicolinearity: The VIF for the preliminary main regression (converted yearly compensation ~ the rest of the variables) shows that most of the variables have low multicolinearity, with the adjusted GVIF close to 1, suggesting that they are not correlated with each other. While work experience and years of coding experience do show moderate multicolinearity, they are within generally acceptable levels. Therefore, all these variables were included in the model, since the overall low levels of multicolinearity should not significantly impact the model.

Variable Transformation: After running our main regression with added interaction term, we observed the first two diagnostic plots to see if the model violate any of the linear regression assumptions.

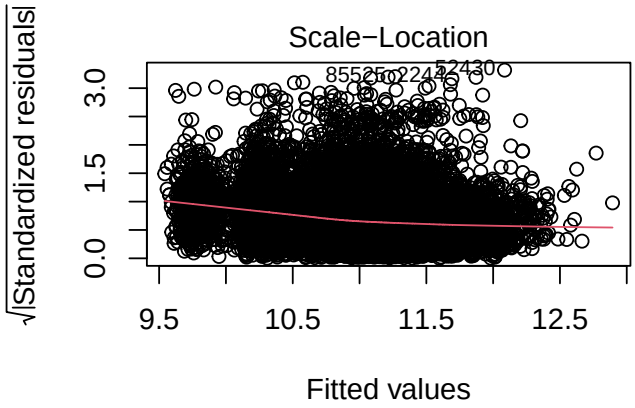


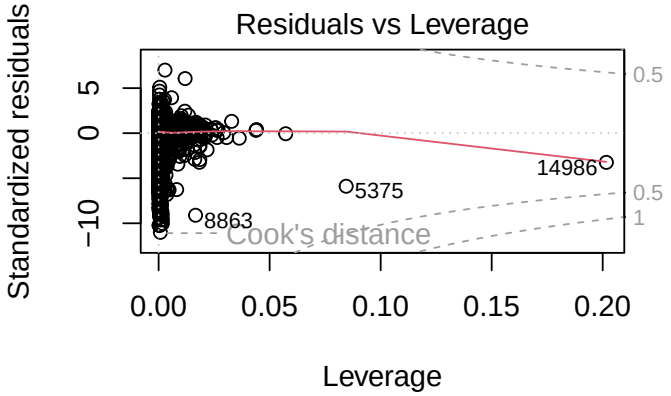


- *Linearity*: From the residual vs fitted plot, it shows the linear relationship between predictor variables and the mean of outcome variable is not violated.
- *Equal variance of errors*: Form the residual vs fitted plot, it shows the equal variance of errors is violated because it does not demonstrate a cloud shape equally spread around 0. To address the assumption of equal variance of errors, we transformed the outcome variable log(converted yearly compensation). The plots after transformation can be found in *Appendix 2*.
- *Normality of errors*: From the qq plot, it shows the normality of errors is violated because the tailing points are not clustered around the 45 degree line. After we address the equal variance of errors assumption through outcome variable transformation, this assumption should be addressed.
- *Independence of errors*: Observations are independent of each other, which is addressed by the study design.

Influential Points:

After running our main regression with variable transformations, we observed the last two diagnostic plots to check if the model has influential points that need to be handled.





From the Residual vs Leverage plot, we observed that there is no influential point lies on Cook's distance. Then, we began to inspect points with either high leverage or high standardized residuals. we decided to drop point 14986, which had high leverage. Upon removing point 14986 and observing my regression outcome, we concluded that point 14986 is indeed influential as the p-value of the interaction term between primary/elementary school educational level and years of coding experience becomes insignificant (from 0.04882* to 0.66865) after its removal, indicating that this point has a substantial impact on the model as a whole. Subsequently, we removed point 1893, which also had high leverage, and revisited the model. However, the significance level of the model remained unchanged, suggesting that point 1893 is not an influential point as it doesn't have substantial impact on the model as a whole. We decided to keep point 1893 and stop my influential points checking at this point. In the end, we removed only one influential point 14986. The plots with outliers being removed can be found in *Appendix 3*.

Our final multiple linear regression model has log-transformed yearly compensation as the outcome variable and several predictor variables: age, number of coding languages, remote work status, work experience, education level, and years of coding experience. Additionally, it includes an interaction term between education level and years of coding experience to assess their joint impact on the compensation. This model is based on an updated dataset, where influential point 14986 has been excluded for more accurate results. The regression result table can be found in *Appendix 4*.

IV. Model Assessment: Salary

We developed two models for comparison: the main model, which includes an interaction term, and a secondary model without this interaction term. To evaluate the performance of these models, we employed the Root Mean Square Error (RMSE) as our primary metric. Through cross-validation, we found that the RMSE of the main model was approximately 1.097. In contrast, the secondary model exhibited a slightly higher RMSE of approximately 1.099. Given that a lower RMSE indicates a model with predictions more closely aligned with actual outcomes, we determined that the main model, with its marginally lower RMSE, is the more accurate of the two. A low R-squared value of 0.14 indicates that the 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.

Table 1: Model Metrics

Metric	Value
RMSE	1.0980769
R^2	0.1436584
MAE	0.6978156

Model Prediction:

Here is an example illustrating the utility of the model: For an individual aged between 25 to 34 years old, proficient in 6 coding languages, working in a hybrid environment (some remote, some in-person), with 5 years of work experience, holding a master's degree (M.A., M.S., M.Eng., MBA, etc.), and having 8 years of professional coding experience, the model predicts a yearly compensation of approximately $62,782.53 USD.

V. Data: Job Satisfaction

To answer the question "What elements shape a developer's job satisfaction?" We created a variable from a questionnaire that ranges from 1 to 5, considering 5 as the number that represented the highest satisfaction. For its construction, please refer to Appendix 1.

To answer this question it is necessary to consider variables that are logically associated with job satisfaction, in that sense we decided to use the following independent variables:

- *Age:* 18-24 years old, 25-34 years old, 35-44 years old, 45-54 years old, 55-64 years old, 65 years or older, prefer not to say, under 18 years old.
- *Type of work:* Hybrid, In-person, Remote.
- *Organization size:* 2 to 9 employees,10 to 19 employees, 20 to 99 employees,100 to 499 employees, 500 to 999 employees,1,000 to 4,999 employees, 5,000 to 9,999 employees, 10,000 or more employees, I don't know, just me
- *Yearly compensation in USD:* From 1 to 74,351,432 USD dollars.
- *Usage and incorporation of Artificial Intelligence to the job:* No and I don't plan to, No but I plan to soon, Yes

V. Model : Job Satisfaction

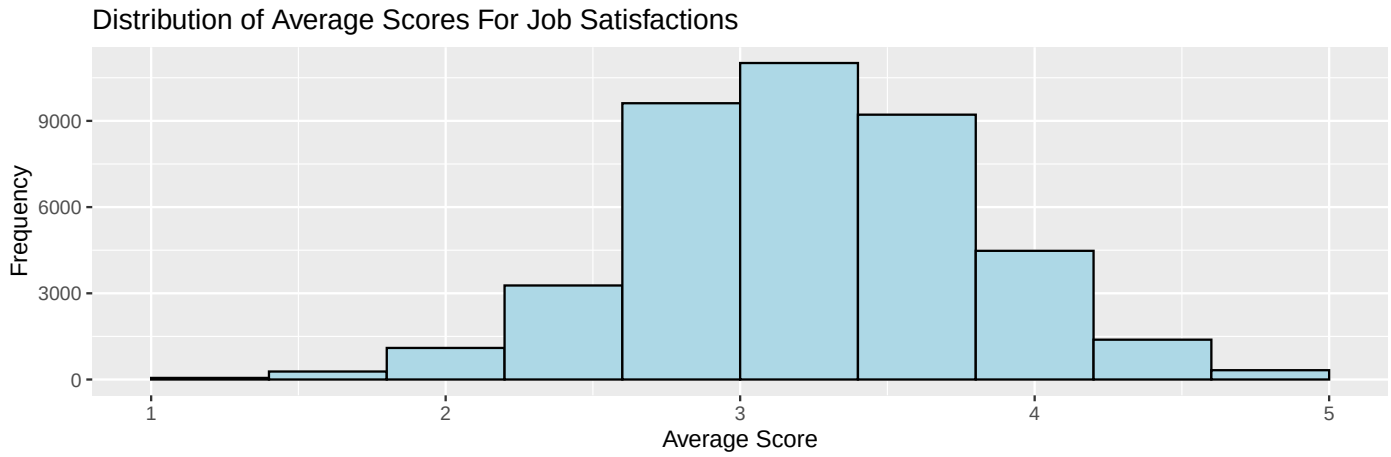
This question is related to an inference problem where we want to understand which variables impact job satisfaction. We want to measure the job satisfaction metric on an ordinal scale ("Satisfied", "Neutral", "Dissatisfied"). Specifically, we want to comprehend how predictors affect the likelihood of transitioning from one category to another and for this reason we will use an ordinal regression.

For the selection of the dependent variable, we will convert the survey results for Job Satisfaction. The data was classified using quartiles. Data from the first quartile downwards will correspond to the category 'Dissatisfied', from the 1st quartile up to the 3rd quartile will correspond to the category 'Neutral', and data above the 3rd quartile will represent the category 'Satisfied'.

We are interested in how the relationship between a developer's yearly compensation and job satisfaction varies across different categories of age so we will include an interaction term with age and yearly compensation. In summary, including this interaction term with age and yearly compensation in our model will allow us to investigate how the relationship between yearly compensation and job satisfaction is modified or varies across different categories of age.

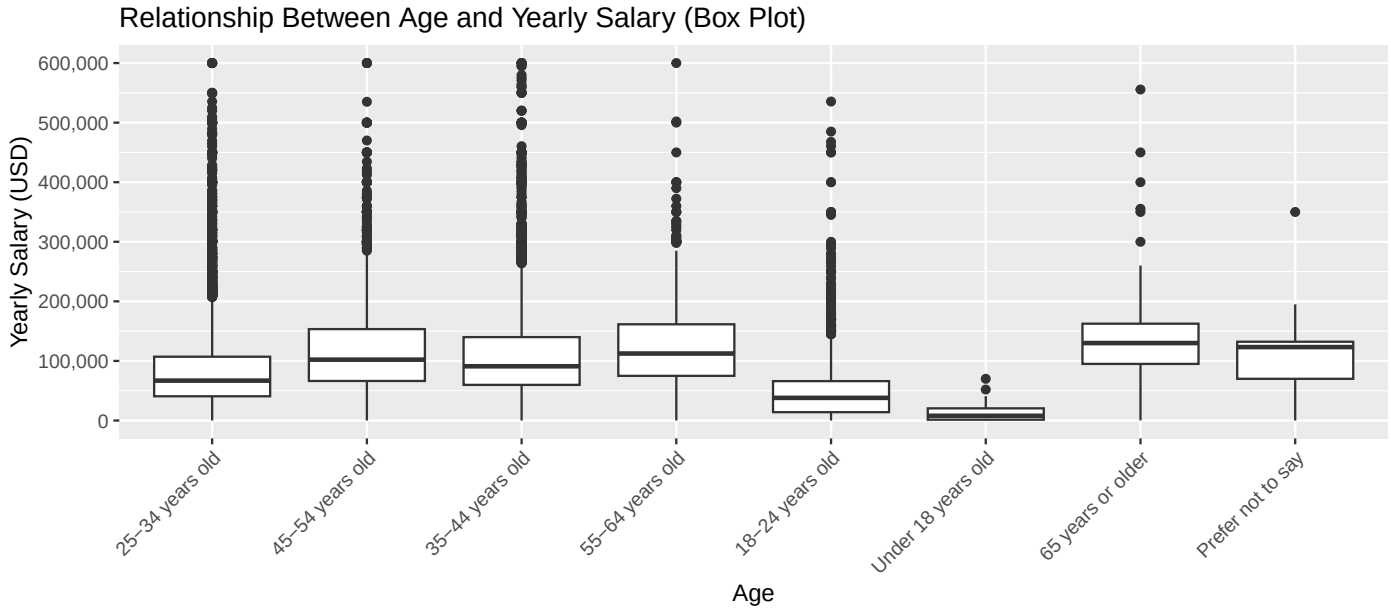
VI. Exploratory Data Analysis Results: Job Satisfaction

Now we are going to explore the distribution of our variable of interest: Job Satisfaction.



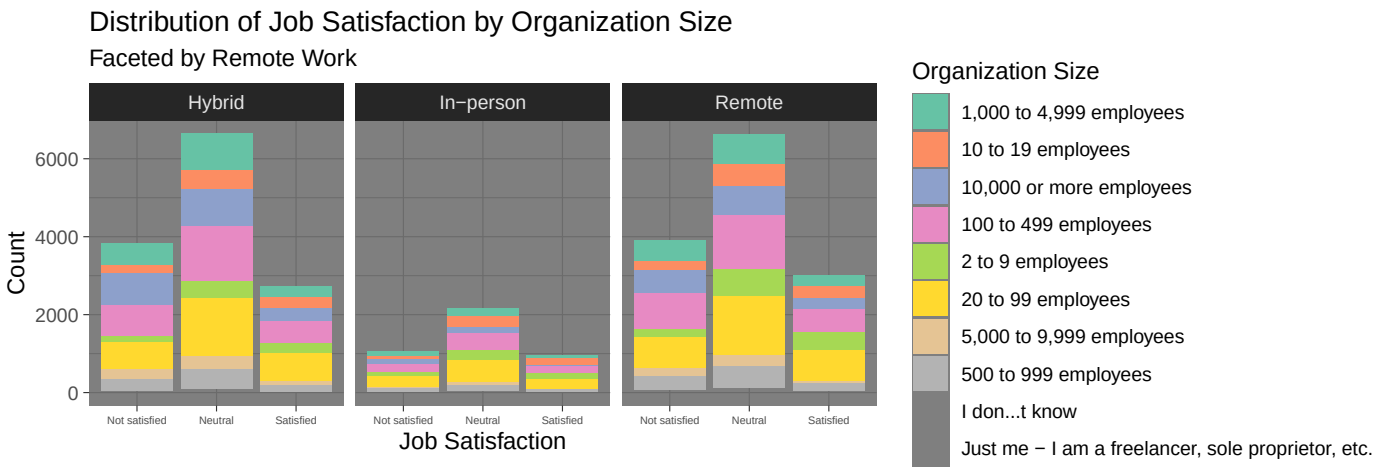
The histogram illustrates the distribution of the average scores for the Job Satisfaction variable. The scores tend to follow a near-normal distribution, with the bulk of observations falling within a standard deviation of the mean (approximately 3). We then categorized the Job Satisfaction variable into three groups: Not Satisfied (1.0 - 2.875), Neutral(2.875 - 3.625), Satisfied (3.625 - 5.0).

Relationship between Age, compensation and Job satisfaction



Our aim was to investigate the interplay between job satisfaction, compensation, and age graphically. 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.

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.

VI. Model Results: Job Satisfaction

The coefficients from the ordinal model can be somewhat difficult to interpret because they are scaled in terms of logs. For interpret this ordinal model we will convert the coefficients into odds ratios. We will exponentiate the estimates and confidence intervals. These coefficients are called proportional odds ratios:

		JobSatisfaction	
Predictors	Odds Ratios	CI	p
Not satisfied|Neutral	0.59	0.38 – 0.94	**0.026**
Neutral|Satisfied	5.68	3.58 – 8.99	**<0.001**
Age25-34 years old	0.83	0.49 – 1.42	0.501
Age35-44 years old	0.77	0.41 – 1.44	0.412
Age45-54 years old	0.37	0.16 – 0.88	**0.025**
Age55-64 years old	1.79	0.30 – 10.56	0.519
Age [65 years or older]	0.15	0.00 – 154.56	0.591
Converted Yearly Compensation	1.03	0.98 – 1.07	0.244
Remote Work Status [In-person]	1.04	0.97 – 1.11	0.271
Remote Work Status [Remote]	0.99	0.94 – 1.03	0.546
Organization Size [10 to 19 employees]	1.87	1.70 – 2.06	**<0.001**
Organization Size [10,000 or more employees]	0.81	0.74 – 0.88	**<0.001**
Organization Size [100 to 499 employees]	1.18	1.09 – 1.27	**<0.001**
Organization Size [2 to 9 employees]	2.35	2.14 – 2.59	**<0.001**
Organization Size [20 to 99 employees]	1.53	1.41 – 1.65	**<0.001**
Organization Size [5,000 to 9,999 employees]	0.83	0.74 – 0.93	**0.002**
Organization Size [500 to 999 employees]	1.03	0.93 – 1.14	0.543

Organization Size [Just me - I am a freelancer, sole proprietor, etc.]	1.48	1.08 – 2.03	**0.016**
AI Usage [No, but I plan to soon]	0.98	0.93 – 1.04	0.532
AI Usage [Yes]	0.94	0.90 – 0.99	**0.027**
Age25-34 years old:Converted Yearly Compensation	1.01	0.96 – 1.06	0.673
Age35-44 years old:Converted Yearly Compensation	1.01	0.96 – 1.08	0.624
Age45-54 years old:Converted Yearly Compensation	1.09	1.01 – 1.17	**0.037**
Age55-64 years old:Converted Yearly Compensation	0.96	0.82 – 1.12	0.581
Age [65 years or older] × Converted Yearly Compensation	1.17	0.65 – 2.11	0.611
Observations	30298		
R^2 Nagelkerke	0.031		

Here we are also computing Confidence intervals (CIs) for parameter estimates, this CIs provide a range of likely values for a population parameter. If a 95% CI does not cross 0, it means the parameter estimate is statistically significant.

It is interesting to observe that most of our variables are not significant if we consider that they do not have a p-value < .05. In principle, we would think that our selected variables such as age, compensation, or work style influence people's job satisfaction, but this is not observed in our model. However, we can observe that the only significant variable is the one related to the size of the organization, mainly within companies with few employees.

Let's interpret 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 (i.e., satisfied or neutral versus dissatisfied) in the job is 2.35 times that of workers who work in a larger organization, holding constant all other variables.

Another variable that we see is significant is for those people who use Artificial Intelligence to execute their work. In this case, we see that for the case of these employees, the odds of being more satisfied (i.e., satisfied or neutral versus dissatisfied) in the job is 5.60% lower [i.e., (1 - 0.944) x 100%] than those who do not plan to use it, holding constant all other variables.

VI. Model Assessment: Job Satisfaction

To evaluate our ordinal model, we will follow the steps below:

We will evaluate predicted probabilities of being in each category and then generate a confusion matrix to assess the accuracy of the predictions. Then we will assess the proportional odds assumption. To do this, we will compare the predicted probabilities using the multinomial model, which is a more precise model, to the predicted probabilities with the ordinal model. For this purpose, we will create a new data frame with different combinations of predictor values, keeping the variable "Yearly compensation in USD" constant using the average of our data. Then, we will compare the predicted probabilities from the ordinal model and the multinomial model, as well as the confusion matrices for both models.

To assess whether the proportional odds assumption holds in this model, we will compare the predicted probabilities obtained from the more precise multinomial model to those obtained from the ordinal model. However, it's crucial to acknowledge that this evaluation involves a degree of subjectivity, as determining a threshold for significant discrepancies in predicted probabilities indicating a violation of the assumption can be challenging.

To generate predictions, we will create a new data frame with all the possible combinations of predictor values. It's generally easier to utilize different combinations of categorical variables while keeping continuous predictors constant, thas why in this case, we keep the salary variable constant using the mean of the data.

Because in this model we have more than 400 different combinations of unique values, we are going to analyze this assumption by comparing 8 predicted probabilities from the ordinal model and the multinomial model, considering age of 25-34 years old, the mean of log compensation, a hybrid type of work, and each size of the organization, and a yes in AI usage:

Table 3: Multiple Linear Regression Model Predictions

Not satisfied	Neutral	Satisfied
0.1698265	0.5025650	0.3276086
0.2094514	0.5117166	0.2788321
0.2494657	0.5084968	0.2420375
0.2991085	0.5039218	0.1969697
0.3164537	0.5184325	0.1651138
0.3246110	0.5121118	0.1632772
0.3625045	0.5033645	0.1341310
0.3831006	0.4665789	0.1503205

Table 4: Ordinal Model Predictions

Not satisfied	Neutral	Satisfied
0.1759441	0.4949392	0.3291167
0.2117215	0.5077164	0.2805621
0.2477013	0.5109583	0.2413404
0.2991005	0.5038234	0.1970761
0.3276650	0.4954345	0.1769005
0.3343727	0.4930943	0.1725329
0.3773152	0.4753042	0.1473806
0.3827815	0.4727296	0.1444890

Notice that the predicted probabilities for the Neutral category are very similar for both models, and we would conclude that we do not have strong evidence that the proportional odds assumption is violated.

We will now compare the accuracy of the predictions between the two models by using a confusion matrix for each of them:

Table 5: Confusion Matrix Ordinal Model

	Not satisfied	Neutral
Not satisfied	13	8593
Neutral	10	15103
Satisfied	3	6576

As we can see in our ordinal model, we only obtained predictions for 2 classes (Not satisfied and Neutral) while we didn't have any for Satisfied, despite the original sample being well-balanced with 8,606 individuals in the Not satisfied group, 15,113 in the Neutral group, and 6,579 in the Satisfied group. This model shows an accuracy of 49.9%, suggesting that the model predicts job satisfaction as well as a coin flip. Surprisingly, despite having various factors, they don't aid us in predicting people's satisfaction levels.

Table 6: Confusion Matrix Multinomial Model

	Not satisfied	Neutral	Satisfied
Not satisfied	11	8592	3
Neutral	22	15084	7
Satisfied	10	6562	7

On the contrary, when we run the multinomial model, we can see that it is predicting the Satisfied class, albeit with very few values. The remarkable aspect is that despite this, this model has nearly the same accuracy (49.8%) as the ordinal model.

VII. Conclusion

Our predictive model aimed to estimate a developer's yearly compensation demonstrates satisfactory accuracy in estimating developer compensation, as indicated by a commendable RMSE value of 1.097654. The model's strength lies in its inclusion of a wide range

of variables, such as age, coding experience, and educational background, which collectively add depth to its predictive capacity. Notably, it features an interaction term between education level and years of coding experience, providing insights into how these factors jointly influence compensation.

However, it also reveals areas of limitations requiring refinement before deployment. The model's use of categorical age brackets (e.g., 18-24, 25-34 years) rather than continuous numerical data limits its precision and could introduce biases. Additionally, the data is unbalanced across these categories, with fewer samples in older age groups, potentially skewing predictions. Another limitation is the handling of non-numeric data. The conversion of non-numeric entries in the 'professional programmer experience' variable to numerical values (e.g., 'less than 1 year' to 0.5 years) involves assumptions that may not accurately reflect the reality. Without clarification on the context behind these entries, there's a risk of introducing bias into the model. 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.

To enhance the accuracy and reliability of our developer compensation prediction model, future work should concentrate on improving data collection and processing. Key steps include adding more relevant variables and using a different type of model such as polynomial regressions or GAMs that captures the relationship between variables more effectively to better inform compensation. We should obtain numerical age data and ensure balanced age representation to address data skewness from underrepresented age groups. If only unbalanced categorical age data is available, we should consider methods like weighted regression or SMOTE for dataset balancing. Additionally, with clearer and more comprehensive data definition, we could revise 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.

Our ordinal model observations indicated that 'Neutral' job satisfaction consistently had the highest predicted probability across diverse scenarios, irrespective of age or organization size.

Although most of our results didn't yield significance in inferring their influence on an individual's job satisfaction, what we discovered is a discernible pattern: as organizational size increases, there is a corresponding increase in the likelihood of predicting job dissatisfaction. This observed trend suggests an inverse correlation between an organization's size and the levels of job satisfaction among its employees. These findings provide valuable insights into the potential impact of organizational size on job satisfaction, indicating the necessity for further investigation to gain a better understanding of the dynamics and underlying factors contributing to job satisfaction across various organizational scales.

Another noteworthy aspect is that we can conclude that our model performs as well as a multiple linear regression model, thanks to the verification of the proportional odds assumption.

The study faced various limitations, the models' inability to predict Job Satisfaction adequately indicated their limited predictive power despite incorporating multiple factors. Moving forward, there are opportunities for use more 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.

VIII. Appendix

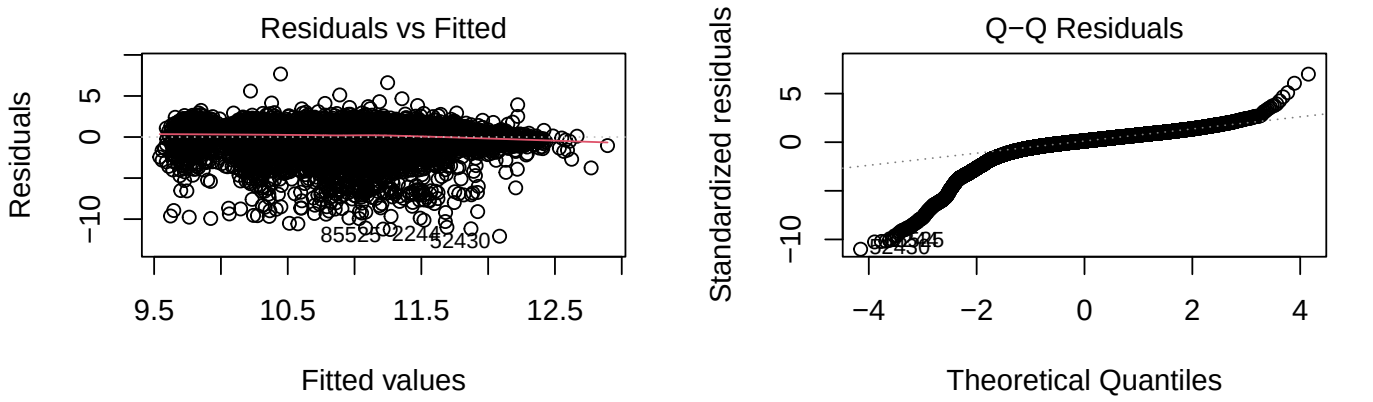
Appendix 1

Within the survey, there were 8 questions related to Job Satisfaction that measured how each person felt in their job.To each of the questions, we assigned a value from 1 to 5, considering 5 as the number that represented the highest satisfaction. 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:

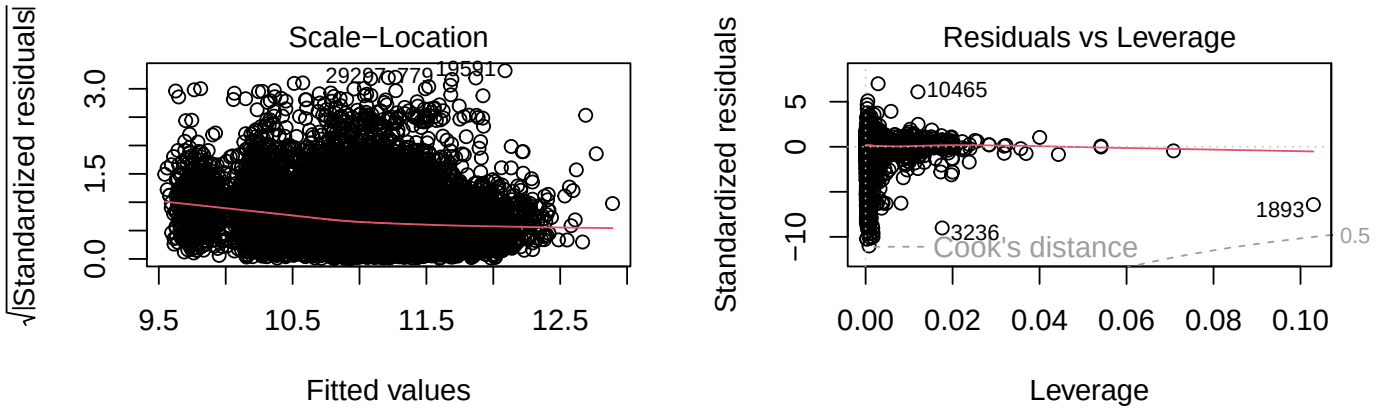
	Job Satisfaction Survey				
Questions	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
1) I have interactions with people outside of my immediate team.	1	2	3	4	5
2) Knowledge silos prevent me from getting ideas across the organization (i.e., one individual or team has information 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 this 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 presume that this information is missing because this section of the survey was optional and some people chose not to answer it.

Appendix 2



Appendix 3



Appendix 4

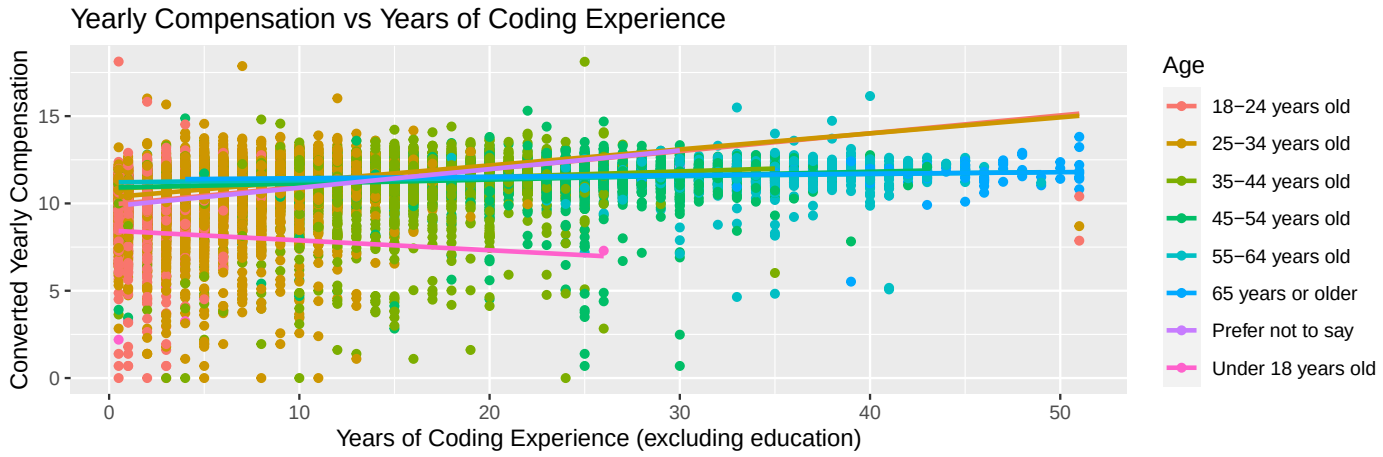
		log(‘Converted Yearly Compensation‘)	
Predictors	Estimates	CI	p
(Intercept)	10.16	10.04 – 10.28	**<0.001**
Age25-34 years old	0.50	0.45 – 0.54	**<0.001**
Age35-44 years old	0.49	0.43 – 0.55	**<0.001**
Age45-54 years old	0.18	0.10 – 0.27	**<0.001**
Age55-64 years old	-0.07	-0.19 – 0.06	0.286
Age [65 years or older]	-0.27	-0.50 – -0.04	**0.022**
Number of Coding Languages	0.02	0.01 – 0.02	**<0.001**
Remote Work Status [In-person]	-0.45	-0.49 – -0.42	**<0.001**
Remote Work Status [Remote]	0.06	0.03 – 0.08	**<0.001**
Years of Work Experience	0.02	0.01 – 0.02	**<0.001**
Education Level [Bachelor’s degree (B.A., B.S., B.Eng., etc.)]	-0.12	-0.23 – -0.00	**0.043**
Education Level [Master’s degree (M.A., M.S., M.Eng., MBA, etc.)]	0.01	-0.11 – 0.13	0.861
Education Level [Primary/elementary school]	-0.01	-0.36 – 0.35	0.973
Education Level [Professional degree (JD, MD, Ph.D, Ed.D, etc.)]	0.07	-0.09 – 0.22	0.399
Education Level [Secondary school (e.g. American high school, German Realschule or Gymnasium, etc.)]	-0.11	-0.26 – 0.03	0.127
Education Level [Some college/university study without earning a degree]	-0.19	-0.32 – -0.07	**0.002**
Years of Professional Coding	0.02	0.01 – 0.03	**<0.001**
Education Level [Bachelor’s degree (B.A., B.S., B.Eng., etc.)] × Years of Professional Coding	0.02	0.01 – 0.02	**<0.001**
Education Level [Master’s degree (M.A., M.S., M.Eng., MBA, etc.)] × Years of Professional Coding	0.00	-0.00 – 0.01	0.382
Education Level [Primary/elementary school] × Years of Professional Coding	-0.01	-0.03 – 0.02	0.685
Education Level [Professional degree (JD, MD, Ph.D, Ed.D, etc.)] × Years of Professional Coding	0.00	-0.01 – 0.02	0.368

Education Level [Secondary school (e.g. American high school, German Realschule or Gymnasium, etc.)] × Years of Professional Coding	-0.00	-0.01 – 0.01	0.893
Education Level [Some college/university study without earning a degree] × Years of Professional Coding	0.01	-0.00 – 0.02	0.092
Observations	30297		
R^2 / R^2 adjusted	0.144 / 0.144		

Appendix 5

Relationship between Years of Coding Experience and Converted Yearly Compensation

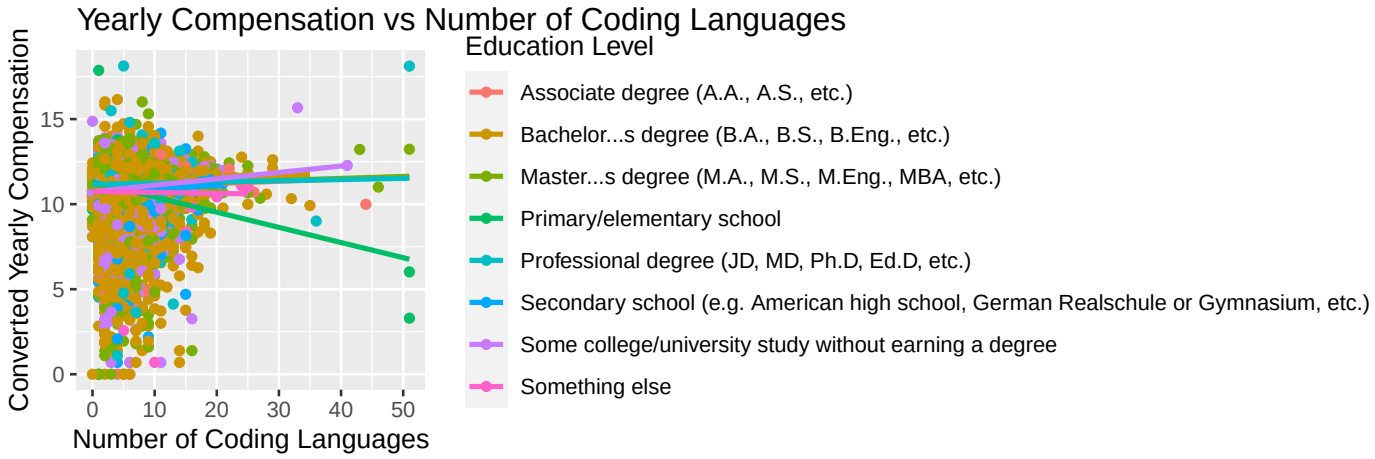
Generally, as coding experience increases, converted yearly compensation also rises across all age groups. However, this trend is not observed in individuals under 18 years old; For every category of work type, there's a noticeable increase in the converted yearly compensation as years of coding experience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



Appendix 6

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.



IX. References

Stack Overflow. (2023). Stack Overflow Developer Survey 2023. Retrieved from https://insights.stackoverflow.com/survey