Predicting Software Developer Salaries in the United States

Joao Ferreira, Julia Costa, Akshaj Kabthiyal

Introduction

- The Bureau Labor of Statistics (BLS) projects that employment in the software development field will grow by about 26% by 2031, while the average growth for all occupations is 3%.
- Understanding compensation trends is crucial for talent acquisition and individual career development.
- Famous salary estimation platforms like Salary.com rely solely on variables such as:
  - Age
  - Gender
  - Total Years of Experience

Research Questions

Influence

What are some non-conventional factors influencing software developer salaries in the U.S.?

Effectiveness

How effectively can these factors be in predicting software developers' salary in the US?

Methodology Summary

Step 1

Data Selection

Getting a relevant Dataset that can be used to answer our research questions.

Step 2

Preliminary Data Cleaning

Taking a quick look at the dataset and only keeping data relevant to the target population.

Step 3

Exploratory Data Analysis

Understanding our variables and their relationship with each other.

Step 4

Feature Engineering

Transforming and creating new variables to fit the model.

Step 5

Model Building

Building different kinds of machine learning models.

Step 6

Evaluation

Using metrics to evaluate the performance and validity of models

Data Source

Data was taken from the Stack Overflow Developer Survey 2023.

Variable Name	Description
Age	Age Range of the respondent
EmploymentStatus	Employment Status of respondent
WorkSetting	Work environment (Remote, Hybrid or In-person)
EdLevel	Highest level of formal education
YearsCodeExp	Years of coding experience including any education
YearsCodeProExp	Years of coding experience not including any education
YearsWorkExp	Years of professional work experience
JobRole	Current Job Role
OrgSize	Number of people employed by the Organization
Country	Country of residence
Currency	Currency used day-to-day
CompTotal	Current total Compensation
LanguageHaveWorkedWith	Languages used by respondent for extensive development work.
DatabaseHaveWorkedWith	Database environments used by respondent for extensive development work.
Industry	Industry the respondent works in

Preliminary Data Cleaning

- Kept U.S Developers Only.
- Restricted ‘Compensation Total’ to a Realistic Range ($60K - $400K).
- Converted both ‘Age Ranges’ and ‘Organization Size’ to Numerical Type using the mid-point method.
- Imputed Missing Values for Years of Experience variables.
- Transformed ‘Coding Work Experience’ to Numeric Values.
- Categorized Job Roles into broader categories such as ‘Developer’, ‘Manager’, ‘Academia’, etc.
  - We kept only ‘Developer’ job roles.
- Filled N/As for industry data with the most common industry in the dataset, IT.

Observations Before Cleaning		Observations After Cleaning
89,184	→	**7,103**

Exploratory Data Analysis

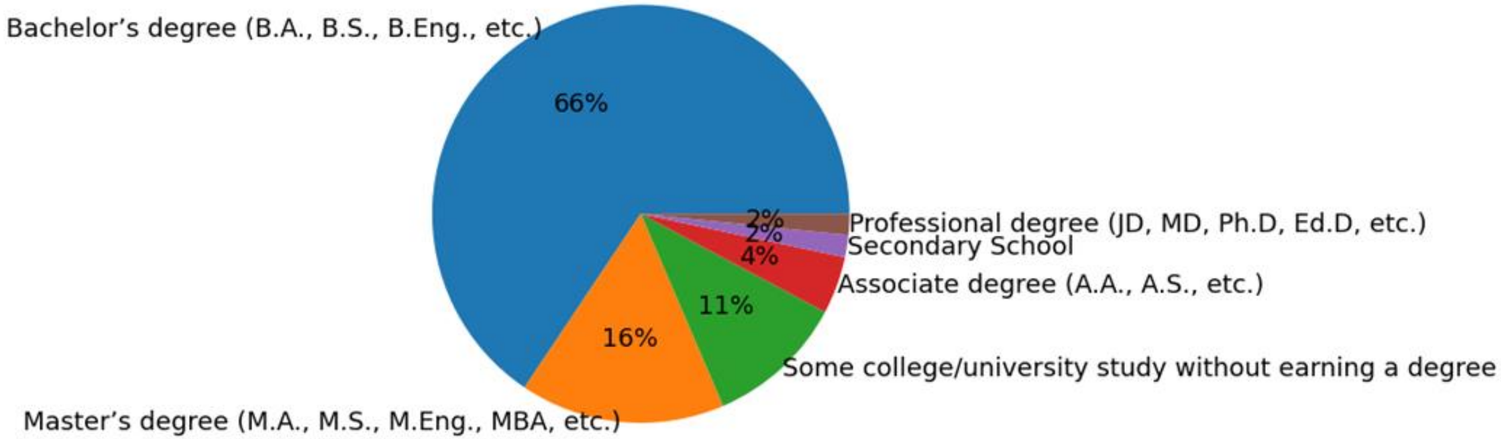
- Univariate Analysis
- Bivariate Analysis

Univariate Analysis

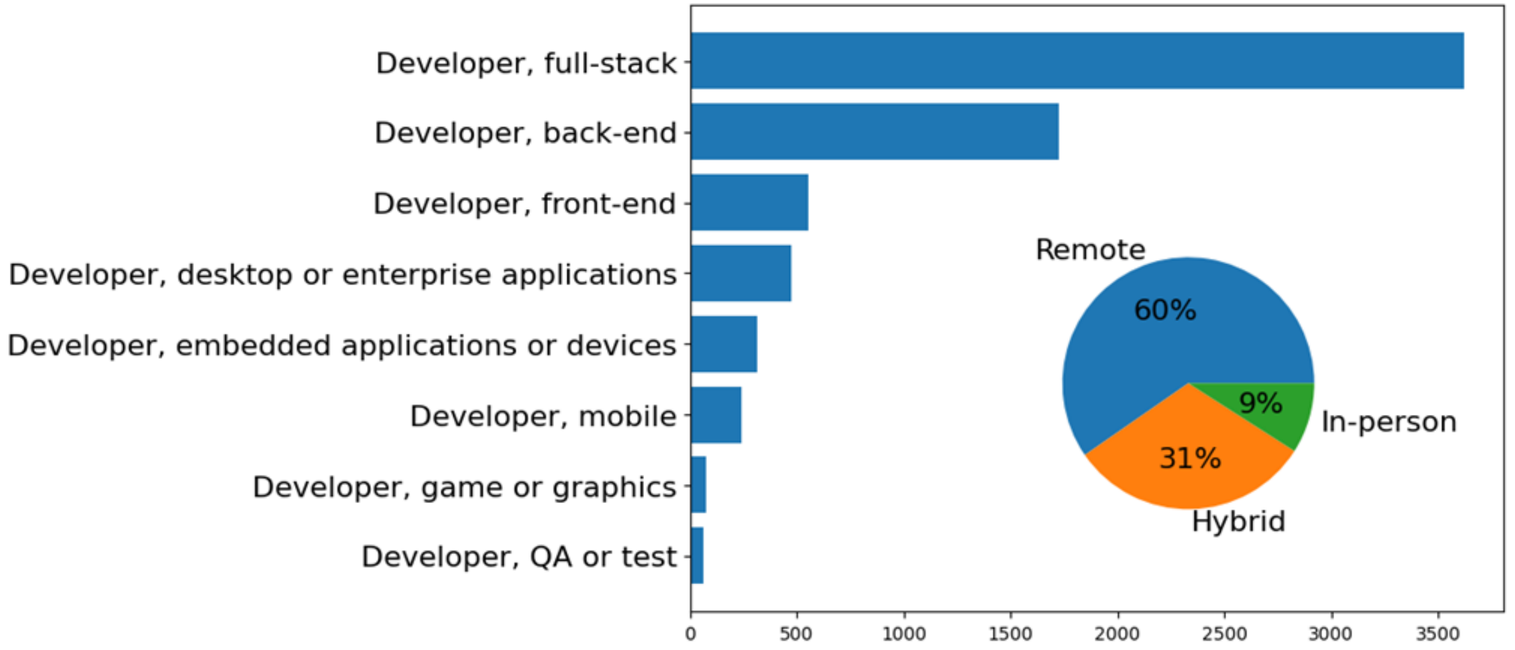
Frequency distribution for the years of different experiences

Univariate Analysis

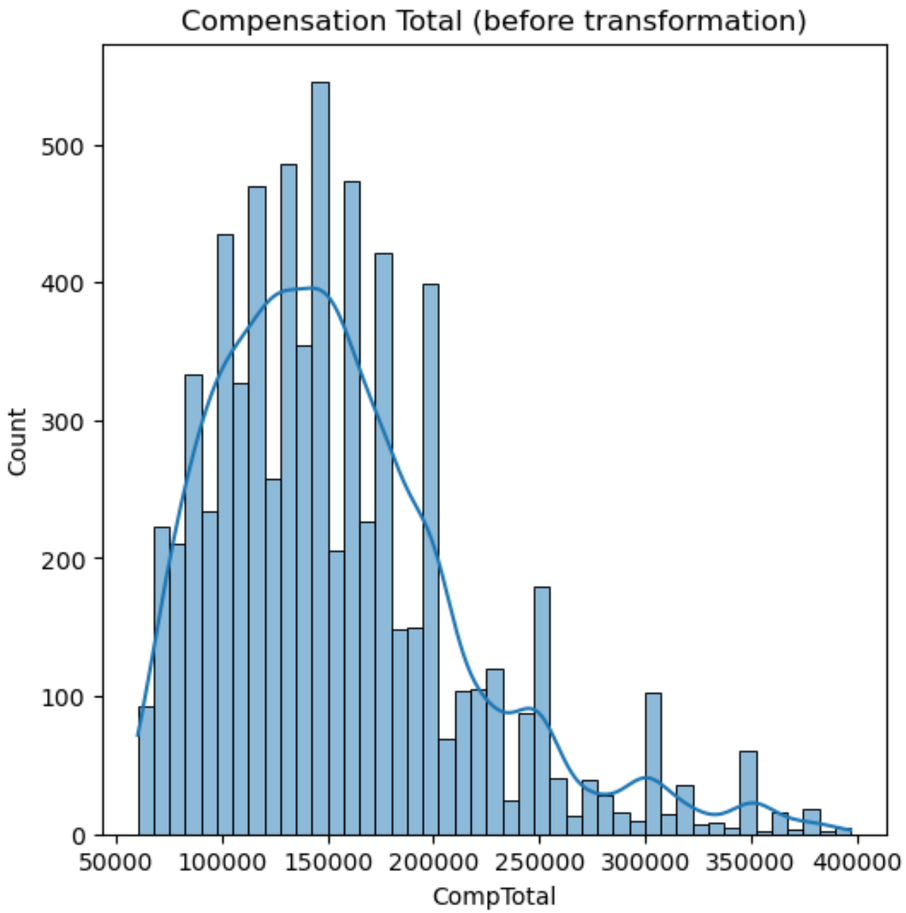
Educational Background of the Respondants



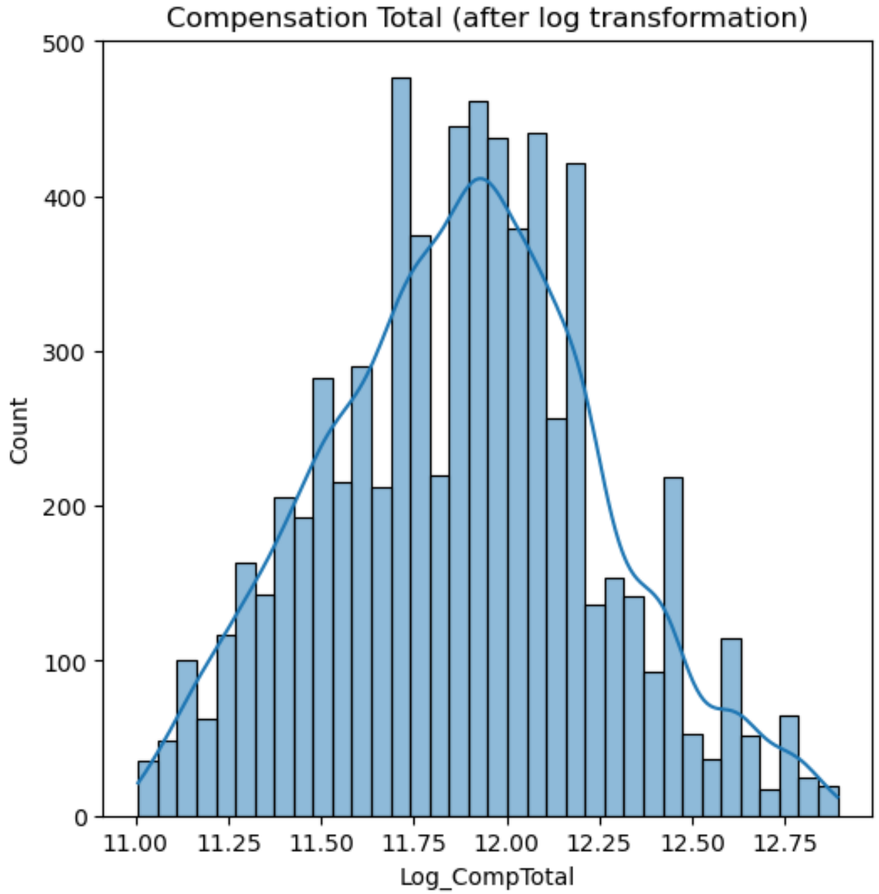
Distribution of developers



Univariate Analysis

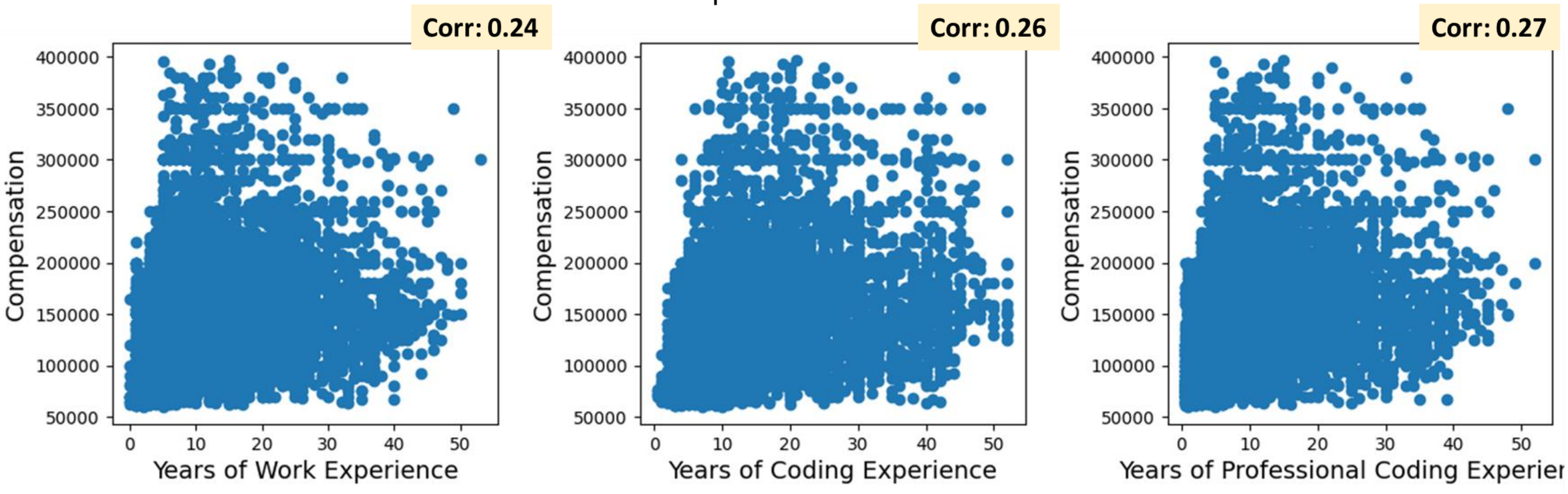


Log
Transformation



Bivariate Analysis

Scatterplot of Compensation vs Different Experiences



Feature Engineering

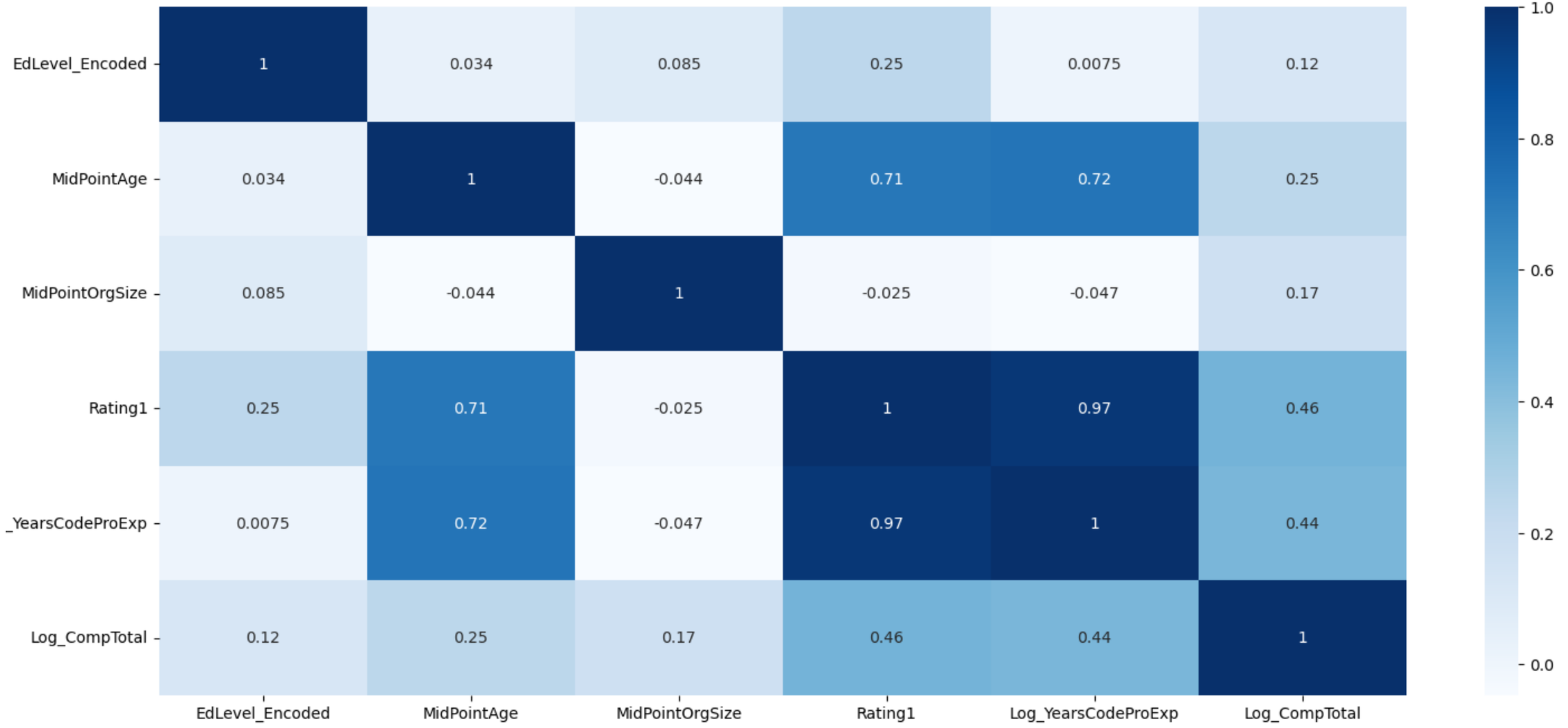
Feature Creation

- **Experience Squared:** Variable created by squaring the years of experience.
- **Rating:** Variable that merges two critical factors—Education Level and Years of Professional Coding Experience. More nuanced indicator of a professional's market worth.

Encoding

- We used One-Hot Encoding to transform the following categorical variables
  - Programming Languages
  - Databases
  - Industry
  - Job Role

Multivariate Analysis



Model Building

Split The Dataset

- 80% for training
- 20% for testing

Model Training

- OLS Regression
- Ridge Regression
- Lasso Regression
- Elastic Net
- Decision Trees

Hyperparameter Tuning

Tuned the model to find the best combinations of parameters.

Model Diagnostics

Performed several tests to check the reliability of the selected model.

Significance Tests

Performed detailed analysis on the statistical significance of our model.

Model Selection

Linear Regression

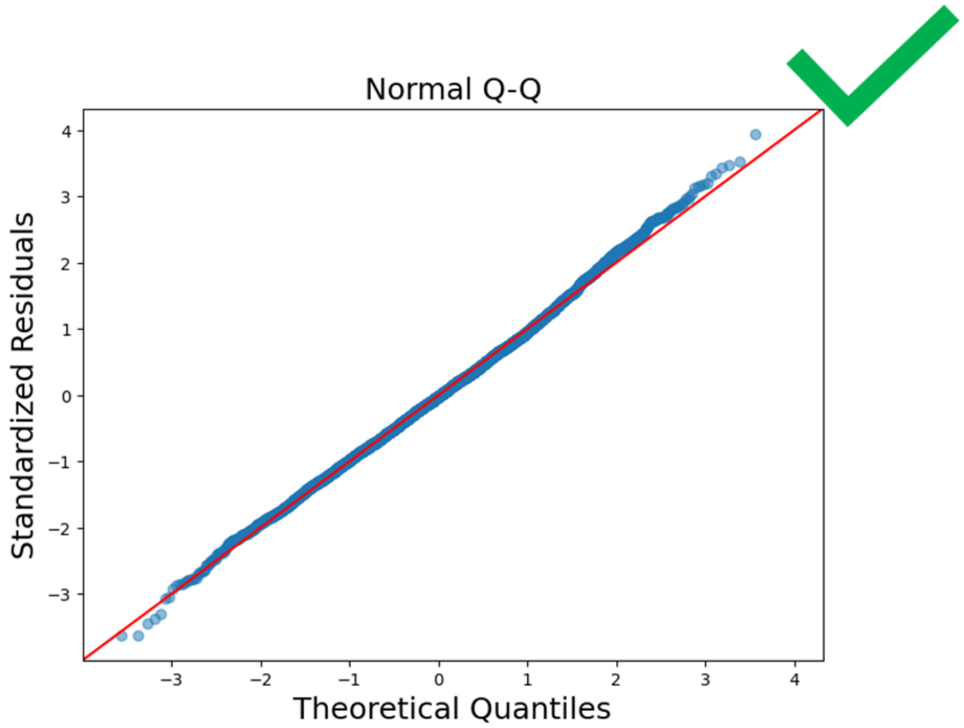
Linear Regression offers a more straightforward explanation for the relationship between variables.

	Model	Best Params	MSE	R^2	Adjusted R^2
0	**Linear Regression**	**{}**	**0.08**	**0.42**	**0.37**
1	**Ridge**	**{'alpha': 10.0}**	**0.08**	**0.42**	**0.37**
2	Lasso	{'alpha': 0.1}	0.13	0.08	0.00
3	ElasticNet	{'alpha': 0.1, 'l1_ratio': 0.2}	0.11	0.24	0.18
4	Decision Tree Regressor	{'max_depth': 5}	0.10	0.32	0.26

Model Diagnostics

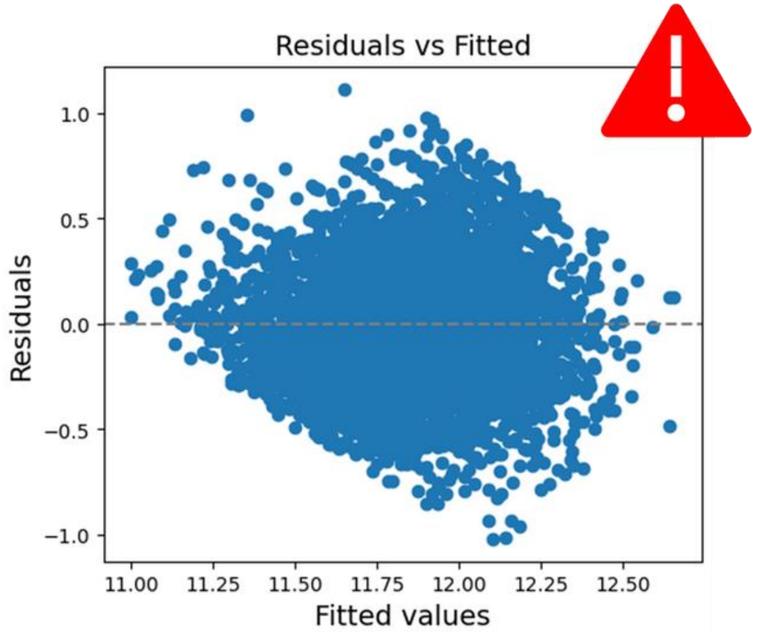
QQ Plot

- Residuals display a normal distribution meeting the assumption for normality within the residuals.



Residual Plot

- The pattern in the plot shows that the variance of the residuals gets bigger and then smaller as the fitted values increases. This may violate the constant variance assumption.



Significance Tests

OLS Regression Results

Dep. Variable:	Log_CompTotal	R-squared:	0.409
Model:	OLS	Adj. R-squared:	0.397
Method:	Least Squares	F-statistic:	36.17
Date:	Wed, 06 Dec 2023	Prob (F-statistic):	0.00
Time:	16:43:19	Log-Likelihood:	-830.74
No. Observations:	5600	AIC:	1873.
Df Residuals:	5494	BIC:	2576.
Df Model:	105		
Covariance Type:	nonrobust		

Variables	Coefficients	t-statistic
Mid-Point-OrgSize	1.554e-05***	15.510
EdLevel_Encoded	-0.0135***	-3.056
Rating1	0.2300***	42.173
Python	0.0245***	2.824
Objective-C	0.0958***	3.447
PostgreSQL	0.0314***	3.475
Snowflake	0.0866***	4.356
Note:		p<0.1;**p<0.05; ***p<0.0

Conclusion

Non-conventional variables can contribute to more accurate Salary Predictions.

- As indicated by the Adjusted R2, 39.7% of the variance in Salary can be explained by our model that combines conventional and non-conventional variables such as Programming Languages known and Databases known .

Salary Estimation websites may benefit by combining Education Level and Professional Coding Experience for a stronger variable for predicting salary. The following are the linear correlations with Salary.

- Education Level: **13%**
- Professional Coding Experience: **41%**
- Education Level + Professional Coding Experience: **43%**

Experience with certain Programming Languages and Databases can lead to higher Salaries.

- Python: **+2.5%**
- Objective-C: **+9.6%**
- PostgreSQL: **+3.1%**
- Snowflake: **+8.7%**

Limitations

Exclusion of Key Variables:

The chosen dataset omits critical variables commonly used in salary prediction, such as Gender, Location, and exact Age.

Impact on Non-Conventional Variable Analysis:

The absence of these variables hinders our ability to test how non-conventional variables work when coupled with conventional variables.

Gender Omission

The study cannot address potential gender-based wage gaps.

Location Omission

The study cannot factor in cost of living and market rate differences across U.S. regions, which significantly impacts accuracy.

FUTURE SCOPE

Since the Constant Variance Assumption is violated, trying non-linear models might be feasible.

VIOLATION OF CONSTANT VARIANCE

Incorporate both conventional and non-conventional variables to more in-depth understanding of the impact that non-conventional variables have on influencing software developer salaries.

EXPAND THE DATASET

Explore how knowledge and application of artificial intelligence (AI) tools can affect salary.

AI KNOWLEDGE AND APPLICATION

Do you have any questions?

Send it to us! We hope you learned something new.

Jdealvarengaferreir@clarku.edu
Jcostasevero@clarku.edu
Akabthiyal@clarku.edu