Unlocking the Code of Developer Compensation

Ivan Bogachev

Abstract

This study utilized the Stack Overflow Annual Developer Survey 2023 dataset, which comprises responses from approximately 40,000 developers after preprocessing. The research aimed to discern the relative impact of geographical location, professional experience, technical skills, and work mode on software developer compensation across various countries. Employing a CatBoostRegressor model for multivariate regression analysis as well as general statical methods, the study found that location was the predominant factor influencing salaries, followed by professional experience and technical expertise. Work mode held a secondary influence, with fully remote developers reporting higher median earnings, potentially due to opportunities for additional freelance work.

Motivation

Understanding the key drivers of software developer compensation is critical for professionals aiming to **maximize their earning potential**.

This research seeks to demystify the weight of factors like location, experience, skills, and work mode on salary variations across borders.

Empowered with this knowledge, **developers** can make strategic career decisions, while **employers** can establish fair and competitive compensation frameworks, fostering a more equitable and optimized global tech marketplace.

Dataset

Stack Overflow Annual Developer Survey 2023:

Data volume: Responses from about 90,000 developers worldwide, compiled into **survey_results_public.csv** with 89,184 entries and 83 features (it was cut down to 36,833 at the preprocessing stage)

Content Breakdown: Features include, among other things, demographic details, educational background, professional experience, programming preferences, and tool usage.

Schema Guide: Detailed explanations of each feature available in survey_results_schema.csv.

Data Source: Survey conducted by Stack Overflow in May 2023, data available https://cdn.stackoverflow.co/files/jo7n4k8s/production/49915bfd46d0902c3564fd9a06b509d08a20488c.zip/stack-overflow-developer-survey-2023.zip.

Applications: Ideal for research on tech industry trends, developer education, and employment and compensation patterns.

Data Preparation and Cleaning

Dropping NaN Values:

Rows with NaN values in the total compensation column were removed to ensure that the analysis only considered complete cases where the target variable is available.

Feature Selection:

Irrelevant features were dropped to maintain a focus on demographic details, professional experience, programming preferences, tool usage, and compensation, which are pertinent to the research question.

Highly correlated features were also removed to reduce multicollinearity, which could affect the performance of many machine learning models and the interpretability of their outputs.

Data Preparation and Cleaning

Encoding:

Categorical features were one-hot encoded, converting them into a binary matrix that is suitable for machine learning algorithms.

Ordinal features were mapped to numerical values to preserve their order in a format that algorithms can process.

Outlier Handling:

Distribution analysis was conducted to examine the spread of the total compensation data.

Entries above the 95th percentile and below the 5th percentile for each country with above 100 responses were excluded to mitigate the effect of outliers that could skew the results.

Research Question

What is the relative importance of geographical location, professional experience, technical skills and work mode in predicting developer compensation across different countries?

Methods

A strategic combination of methods:

Descriptive Statistics: For a preliminary overview, exploration and data sanity checks.

Correlation Analysis: To eliminate redundant features by removing highly correlated variables.

Regression Models: iteratively several models were tested—Linear Regression, RandomForest Regressor, XGBoost, and CatBoost—with CatBoost being selected for its superior handling of categorical data and predictive accuracy.

These methods were chosen for their strengths in dealing with multicollinearity, dimensionality, and categorical data, ensuring a balanced and accurate prediction model.

Findings

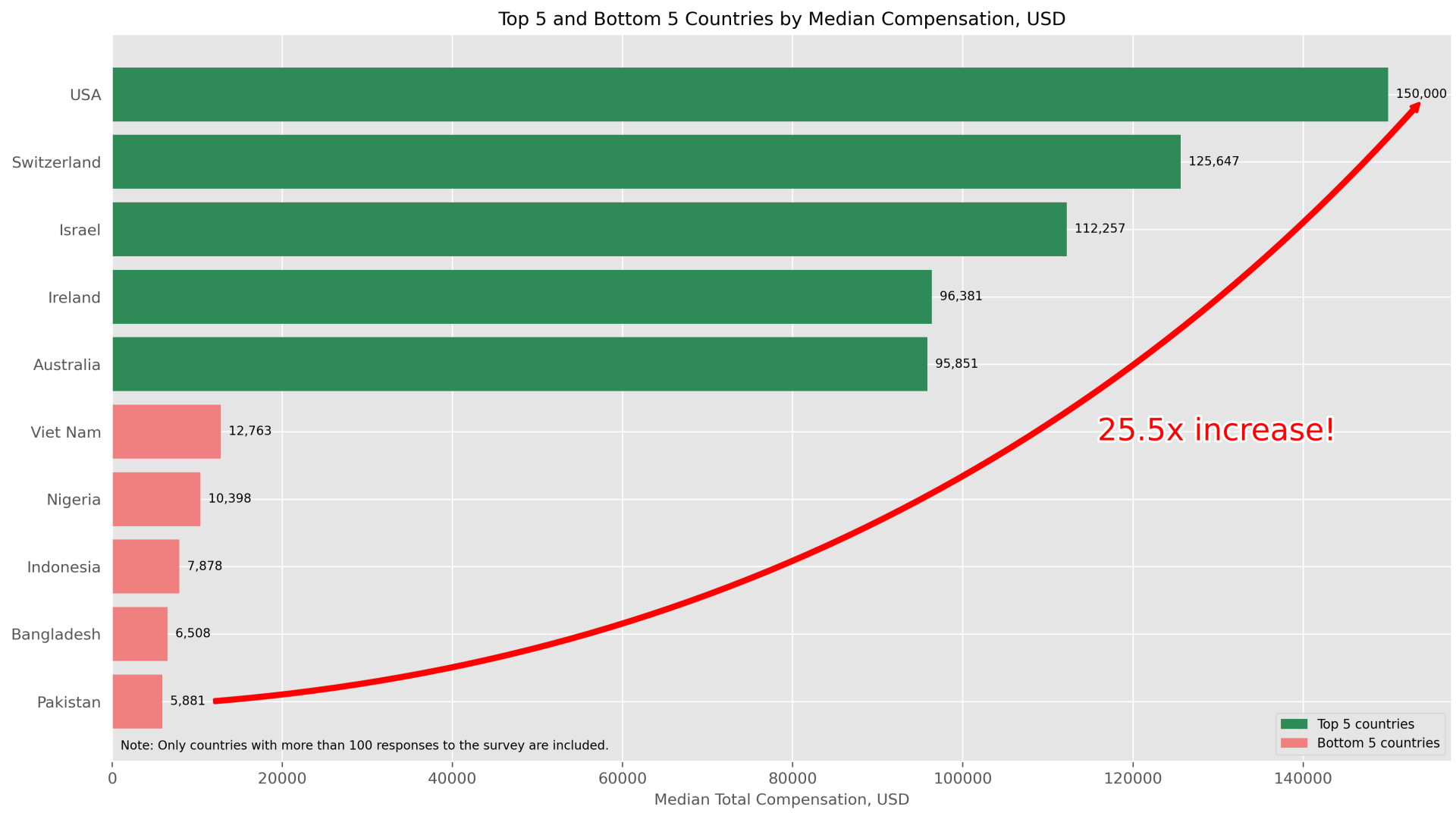
Geographical Location: A Prime Factor by a Considerable Margin Significance:

Major influence on compensation, reflecting cost of living and economic variability across clusters of counries.

Data Insight:

Removal of country data caused a significant drop in a regression model quality (from RMSE: 24939.94, R2: 0.79 to RMSE: 44351.62, R2: 0.32)

25,5 x difference between the lowest and the highest median value for a country (only countries with more than 100 responses were considered)


Top 5 and Bottom 5 Countries by Median Compensation, USD
USA
Switzerland
Israel
Ireland
Australia
Viet Nam
Nigeria
Indonesia
Bangladesh
Pakistan
150,000
125,647
112,257
96,381
95,851
12,763
10,398
7,878
6,508
5,881
25.5x increase!
Note: Only countries with more than 100 responses to the survey are included.
Top 5 countries
Bottom 5 countries
0
20000
40000
60000
80000
100000
120000
140000
Median Total Compensation, USD

Findings

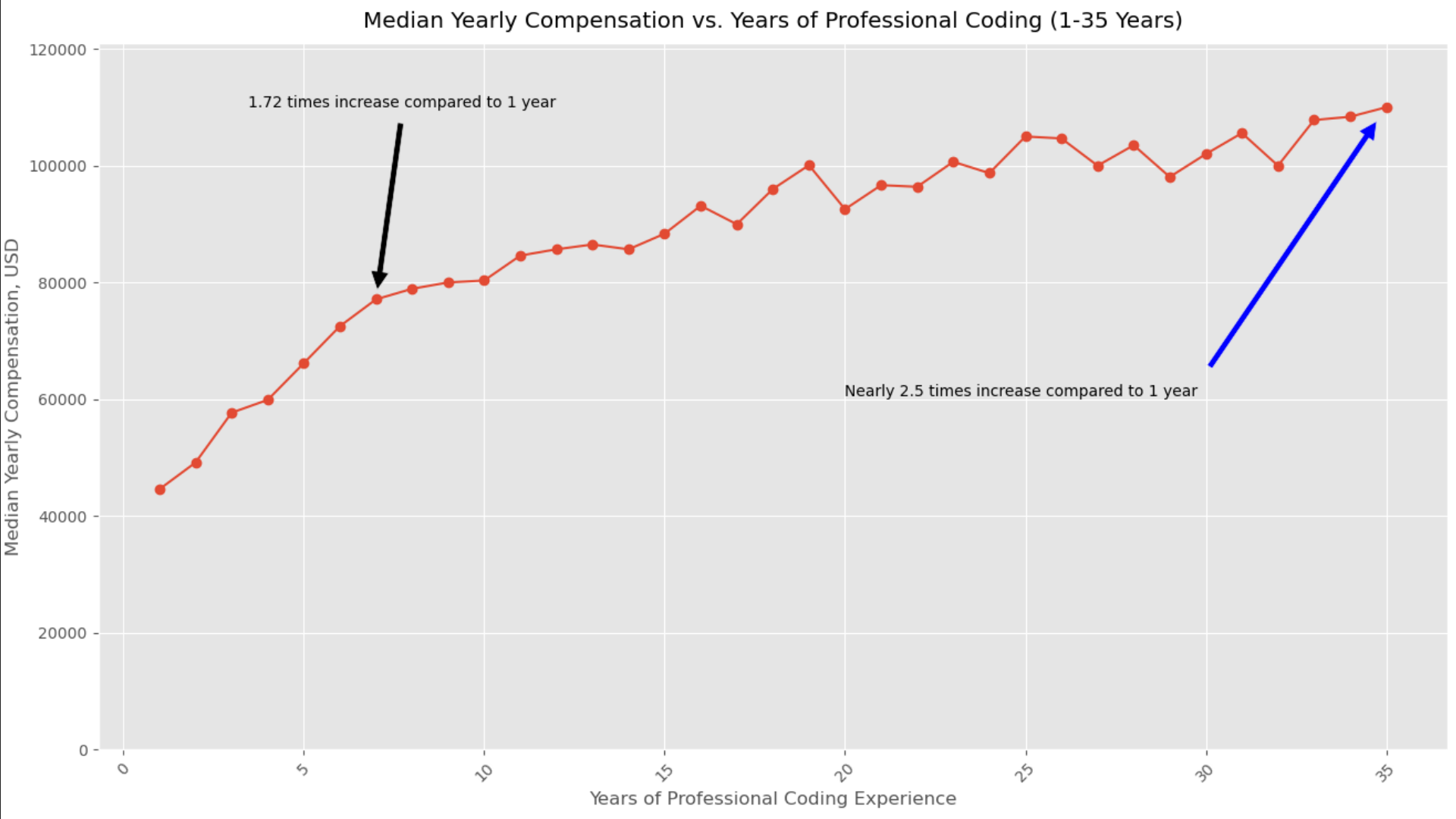
Experience : A Key Predictor of Compensation

Notable Impact: Excluding years of professional experience from our model resulted in a noticeable decrease in accuracy, with RMSE increasing to 27746.03 and R2 decreasing to 0.73.

Data Insight:

Early Career Growth: The initial seven years in a developer's career show the most significant relative salary growth—nearly a 1.72-fold increase.

Long-Term Value: A 2.5-fold salary difference is observed between developers at the start of their careers and those with 35 years of experience, highlighting experience as a valuable asset in the tech industry.


Median Yearly Compensation vs. Years of Professional Coding (1-35 Years)
1.72 times increase compared to 1 year
Nearly 2.5 times increase compared to 1 year
Median Yearly Compensation, USD
Years of Professional Coding Experience
120000
100000
80000
60000
40000
20000
0
0
5
10
15
20
25
30
35

Findings

Technical Skills: Collective Influence on Earnings

Combined Weight: The aggregate removal of programming languages, frameworks, databases, and platforms significantly lowers model performance (RMSE rises to 26992.21, R2 drops to 0.75).

Synergistic Effect: Individual technical skill categories have a less pronounced impact when removed separately. Together, they form a composite index of technical proficiency that is strongly linked to compensation.

Findings

Technical Skills: Average Differences & Notable Extremes

General Observation: Across most technologies, compensation differences are mild (with average standard deviation of about USD10,000), with significant deviations (of about 1,5-2,2x) primarily at the extremes.

Programming Languages:

Ruby tops the median compensation chart, nearly 1.5 times higher than **Dart**, the least compensated.

Cloud Platforms:

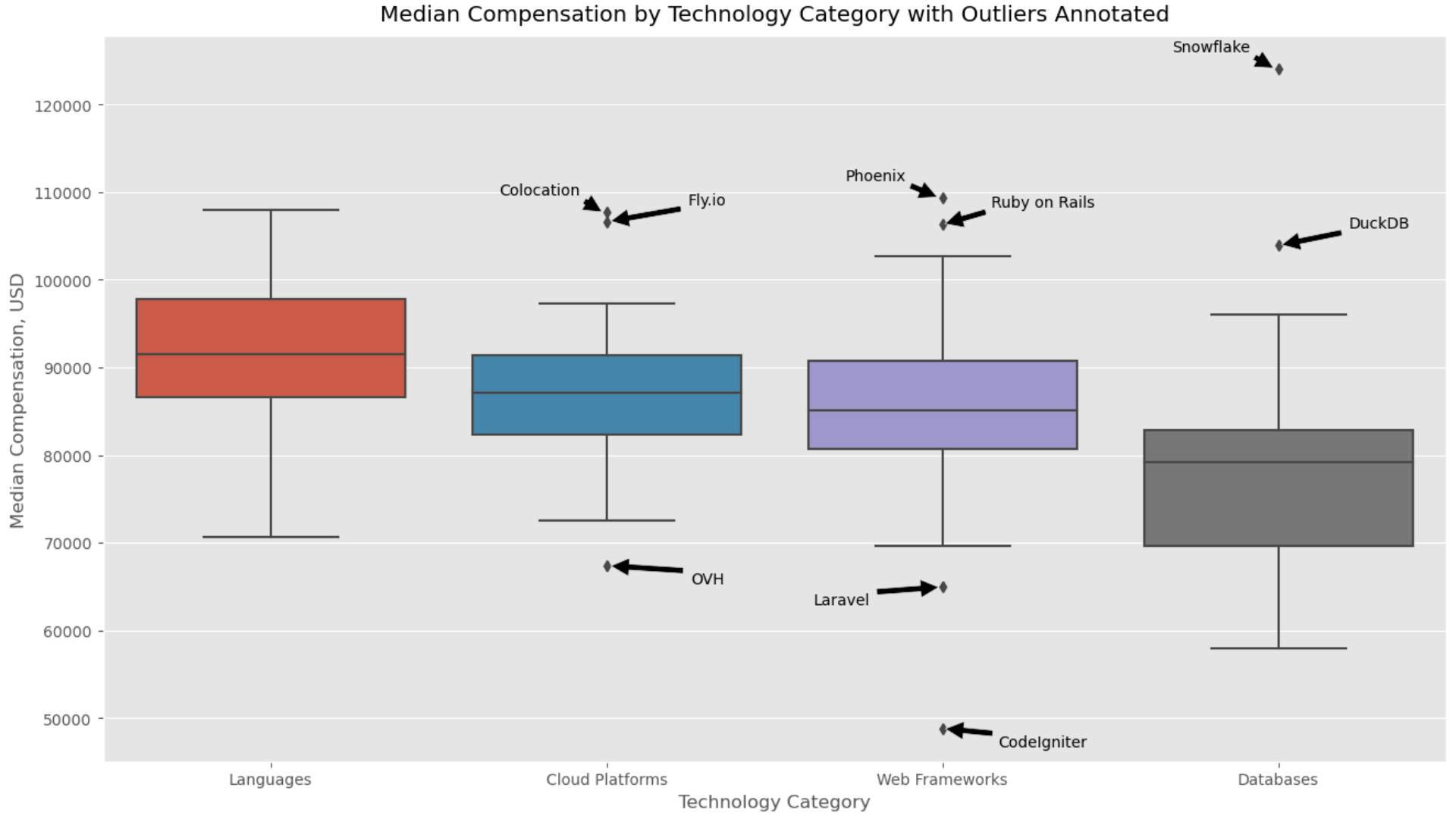
Colocation services lead with earnings approximately 1.5 times that of **OVH**, but without **Colocation** and **Fly.io** on the top end and **OVH** on the bottom end, other platforms are very tight.

Web Frameworks:

Phoenix sits at the highest compensation, with a stark 2.24x difference with **CodeIgniter**. But without **Phoenix** and **Ruby on Ralls** on the top and **CodeIgniter** and **Laravel** on the bottom, the distribution is equally tight.

Databases:

While **Snowflake's** compensation is 2.14x that of **Firebird**, without **Snowflake** and **DuckDB** on the top end, the distribution is also tight.


Median Compensation by Technology Category with Outliers Annotated
Median Compensation, USD
120000
110000
100000
90000
80000
70000
60000
50000
Snowflake
Colocation
Fly.io
Phoenix
Ruby on Rails
DuckDB
OVH
Laravel
CodeIgniter
Languages
Cloud Platforms
Web Frameworks
Databases
Technology Category

Findings

Work Mode: A Secondary Influence on Earnings

Removal of work mode data slightly decreases model accuracy (RMSE: 25152.58, R2: 0.78).

Indicates work mode's influence is secondary to location, experience, and skills.

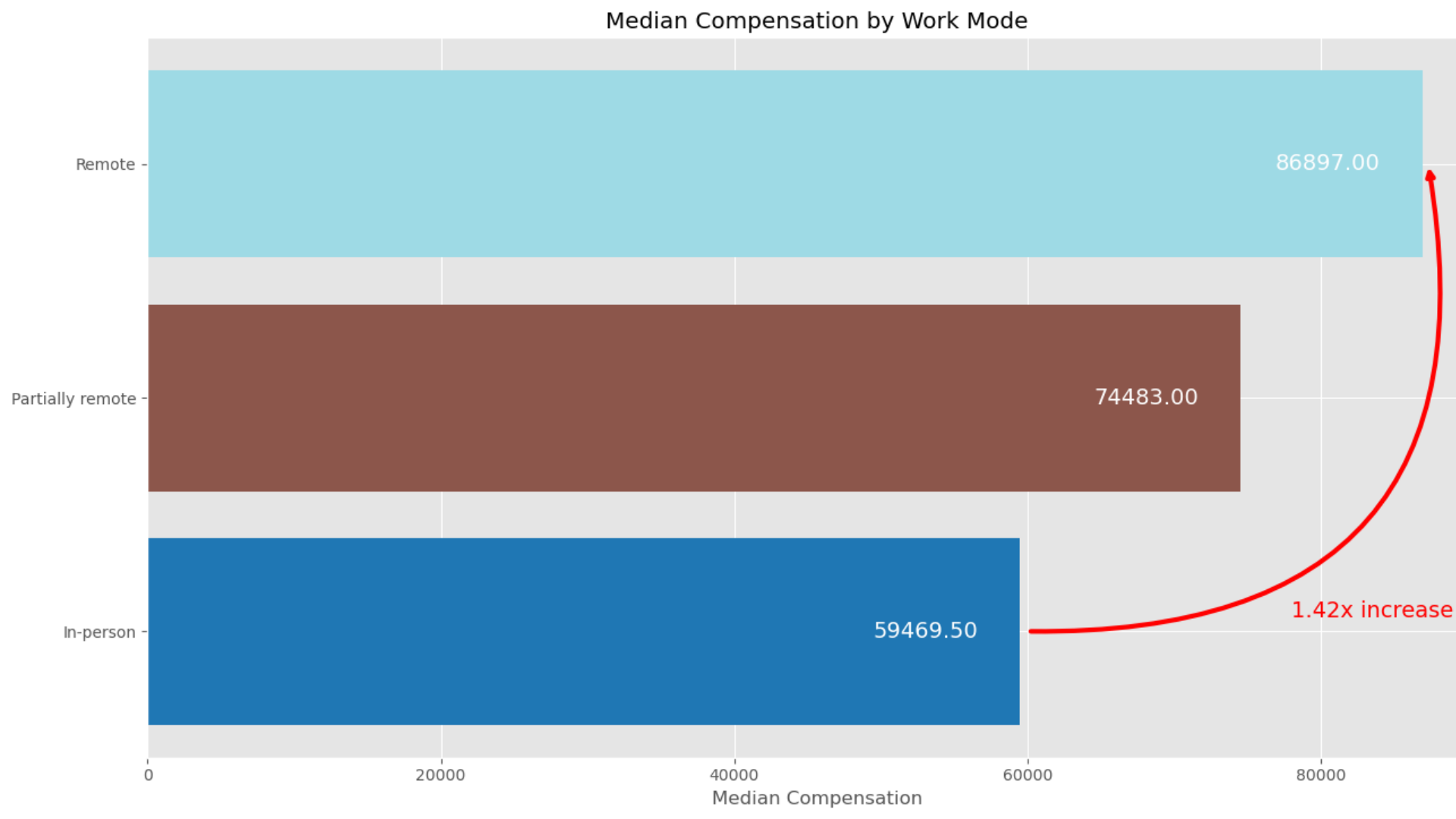
Data Insight:

Developers working remotely report a 1.46x higher median compensation.

Possible contributing factor: remote work may facilitate multiple income sources.

Consideration:

The survey's total compensation measure may include earnings beyond the primary employer, potentially inflating the perceived impact of remote work.


Median Compensation by Work Mode
Remote
Partially remote
In-person
86897.00
74483.00
59469.50
1.42x increase
0
20000
40000
60000
80000
Median Compensation

Findings

Work Mode: A Secondary Influence on Earnings

Removal of work mode data slightly decreases model accuracy (RMSE: 25152.58, R2: 0.78).

Indicates work mode's influence is secondary to location, experience, and skills.

Data Insight:

Developers working remotely report a 1.46x higher median compensation.

Possible contributing factor: remote work may facilitate multiple income sources.

Consideration:

The survey's total compensation measure may include earnings beyond the primary employer, potentially inflating the perceived impact of remote work.

Limitations

Self-Reported Data Bias: The analysis is derived from self-reported data in the Stack Overflow Developer Survey, which can introduce biases. Not all developers participate in the survey, and those who do may not represent the full spectrum of the developer community. Additionally, self-reported compensation figures are not verified and could be inaccurately reported.

Temporal Constraints: The dataset reflects a specific period (May 2023) and may not capture the dynamic nature of the tech industry, where compensation and technology preferences change rapidly.

Work Mode Nuances: The categorization of work mode lacks granularity. It does not differentiate between various forms of remote work or employment types, such as contract versus full-time roles. The possibility of additional income from freelancing for remote workers is also not isolated, which could skew compensation data.

Reduced Sample Size: Initial data included over 90,000 responses, but the need to exclude entries without compensation data reduced the sample size to approximately 45,000. Further data cleaning to remove outliers resulted in a final count of 36,833, which may limit the generalizability of findings.

Sparse Data for Some Features: Certain feature values may have a low number of observations, making it challenging to draw robust conclusions in those cases.

Conclusions

Geographical Location: The analysis indicates that geographical location is the most significant predictor of a developer's compensation. Developers in countries with a higher cost of living and greater demand for tech skills tend to earn significantly more.

Professional Experience: Experience comes in as a strong second in terms of importance. The data shows a clear correlation between years of professional experience and salary, with noticeable increments in compensation particularly evident in the first seven years of one's career.

Technical Skills: Technical skills, when considered as a collective group (including languages, databases, platforms, and frameworks), have a notable impact on earnings. However, within this group, the influence varies, with certain skills commanding higher pay, likely due to their demand or complexity.

Work Mode: Work mode has a discernible, yet smaller, influence on compensation compared to location, experience, and technical skills. Fully remote workers report higher median earnings, potentially due to the flexibility to take on additional projects or the ability to work for employers in higher-paying regions.

Acknowledgements

The dataset was acquired from the Stack Overflow Annual Developer Survey 2023, which is publicly available for analysis. The survey data represents a comprehensive view of developer preferences, tools, and compensation. No informal analysis or external feedback was involved in this study; the data was used as provided. The analysis was conducted independently, without peer review or external inputs.

.

References

All the work and analysis were conducted independently without direct references to external research papers or additional materials. The primary source of data was the Stack Overflow Annual Developer Survey 2023, and the analysis was guided by standard data science practices and methodologies.