Exploratory Data Analysis and Predictions on Software Developers Salaries

1. Introduction

1.1 Background:

Software Development has emerged as the leading discipline in modern industry, with applications ranging from technology to healthcare to finance. The rapid digitization of the economy, followed by increasing reliance on data-driven decision-making, has dramatically increased the demand for skilled professionals in software development. Studies published recently by organizations such as the U.S. Employment in software developer fields are projected by the Bureau of Labor Statistics to increase by 31 percent between 2020 and 2030, much higher than the average for all occupations. The expansion has also led to significant divergence in compensation structures, influenced by factors such as geographic location, industry, level of education, and work experience. However, a fully informed understanding of these trends remains hard to come by, given the dynamic nature of the industry and the many factors that must be considered.

The present work attempts to fill this gap by analyzing compensation in software development through the lens of EDA and a machine-learning perspective. Our aim in integrating social media data with geolocation variables is for pattern identification as well as making precise wage forecasts. This research is therefore of utmost importance in the modern employment landscape, in which professionals, employers, and policy-makers are at present seriously looking for an all-rounded comprehensive view of compensation trends for informed career decisions, formulation of recruitment strategies, and legislation.

1.2 Problem Statement:

While there is a high demand for software developers, there is great uncertainty about what factors drive salary differences. This uncertainty is a direct result of the lack of readily accessible, data-driven analyses into the role that experience, employment type, organization size, educational background, and geography play in compensation. The lack of clarity creates challenges for: job seekers who may have difficulty setting proper expectations for salaries; employers who have to ensure that compensation is competitive and fair and researchers concerned with the way labor markets work.

Addressing this issue is relevant, as it will help stakeholders make informed decisions. Accurate forecasts and detailed analyses will help individuals in aligning their career goals with the current market conditions, while at the same time helping organizations in designing fair and competitive compensation structures. Moreover, the knowledge gained can be used to have broader discussions about economic fairness and the distribution of talent across different industries and regions.

1.3 Objectives:

Objectives or hypotheses the project aims to test or achieve:

- **Exploratory Analysis:** The analysis of salaries to develop trends, patterns, and anomalies in the distribution across job roles, industries, and regions.
- **Predictive Modelling:** Construction of machine learning models that are capable of predicting salaries from key factors like experience, education, technologies and geography.
- **Insight Extraction:** Extract actionable insights related to high-paying industries, influential factors, and geographic disparities in salary.
- **Visualization:** Visualize intuitively and create dashboards to effectively communicate findings to a wide array of stakeholders.
- **Validation:** Validate findings through testing to ensure the accuracy and applicability of findings in real-world contexts.

Expected outcomes and benefits of the project:

- A comprehensive report summarizing the key factors influencing salaries in the data science field.
- Predictive models that provide salary estimates based on specific input parameters.
- Interactive visualizations that make salary data accessible and actionable for users.
- Recommendations for job seekers, employers, and policymakers to improve decision-making regarding compensation.

2. Methodology

2.1 Data Collection:

- **Source Description**: The data used for this project is from the Stack Overflow Annual Developer Survey. It is a survey that asks developers about coding, working, AI, and many other related topics in the technology industry. There are over 65000 respondents for the 2024 survey.

- **Data Acquisition Methods**: Data is publicly available at https://survey.stackoverflow.co/.

2.2 Data Preprocessing:

- **Cleaning Techniques**
  A comprehensive approach was taken to clean the dataset to ensure high data quality and relevance for the analysis:
  - **Removing Specific Entries:** Rows with specific unrealistic values in the SalaryUSD column, such as infinite values or those greater than $700,000, were removed. Duplicate records were identified and removed to ensure data uniqueness.
  - **Filtering by Specific Criteria:** Entries with extreme age values and those marked as 'Prefer not to say' were excluded from the dataset to maintain focus on the core demographic group.
  - **Handling Missing Data:** Missing values across various columns were identified, and specific actions were taken depending on the nature of the data. For instance, rows with missing values in critical columns like Profession were dropped to maintain data integrity. The dataset was reset to ensure consistency in indexing post-data cleaning. For predictive model training, missing data was imputed using a mean strategy for numerical features and a mode strategy for categorical features.

- **Column Removal:** Non-essential columns such as ResponseId, OrgSize, and PurchaseInfluence were dropped from the dataset to streamline the data and focus on relevant variables.

- **Transformation Methods**:
  Data transformations were applied to better suit the dataset for analysis:
  - **Encoding and Scaling:** Categorical variables such as EdLevel, Country, and RemoteWork were encoded using techniques like OneHot Encoding to convert them into a format suitable for modeling. Continuous variables like YearsCodePro were scaled using the MaxAbsScaler to normalize their range and improve the performance of the machine learning algorithms.
  - **Bucketing:** Salary data was categorized into defined bins to simplify analyses and visualize salary distributions more effectively. This categorization helps in better understanding salary trends across different groups.
  - **Column Transformation:** A comprehensive column transformer was applied to preprocess different types of data appropriately before feeding them into various machine-learning models.
  - **Pipeline Integration:** Data preprocessing steps were integrated into pipelines with different models such as Decision Tree, AdaBoost, Random Forest, and Gradient Boosting Regressors to ensure that all data transformations were applied consistently during the model training phase.

These preprocessing steps were meticulously planned and executed to optimize the dataset for the subsequent modeling phase, ensuring that all data used was accurate, relevant, and formatted appropriately for the analytical tasks at hand.

3. Data Analysis

A detailed analysis of the survey dataset, focusing on descriptive statistics, exploratory data analysis (EDA), and predictive modeling. The analysis aims to uncover patterns, correlations, and insights that can guide data-driven decision-making.

3.1 Descriptive Statistics:

- **Summary Metrics:**
  To understand the dataset's overall structure and key characteristics, we calculated summary statistics for various variables:
  - **Dataset Overview**: The dataset includes thousands of responses across numerous features. After data cleaning, it retains essential variables such as Age, Years of Coding, SalaryUSD, and Developer Type.
  - **Missing Data**: Missing values were identified and addressed through forward-filling, backward-filling, or imputation. After cleaning, missing data accounted for less than 5% of the total dataset, ensuring minimal impact on analysis.

- **Statistical Metrics**:
  - **Age**: Respondents ages ranged from under 18 to over 65, with the majority between 25 and 34.

- **Salary**: Salaries ranged from $0 to $700,000 annually, with a mean of approximately $85,000 and a standard deviation indicating significant variability.
- **Years of Coding**: Experience levels ranged from less than 1 year to over 50 years, highlighting a diverse respondent pool.

3.2 Distribution Analysis

The distributions of key variables were examined to understand their spread and identify potential anomalies:

- **Age Distribution**:
  - The histogram showed that most respondents were in their late 20s and early 30s.
  - A minor skew was observed toward younger respondents, reflecting the growing interest in coding among younger generations.

- **Salary Distribution**:
  - Salaries exhibited a right-skewed distribution with a small percentage of respondents earning exceptionally high incomes, likely senior professionals or those in specialized roles.
  - Box plots revealed significant outliers at the upper end, which were retained to preserve the dataset's integrity.

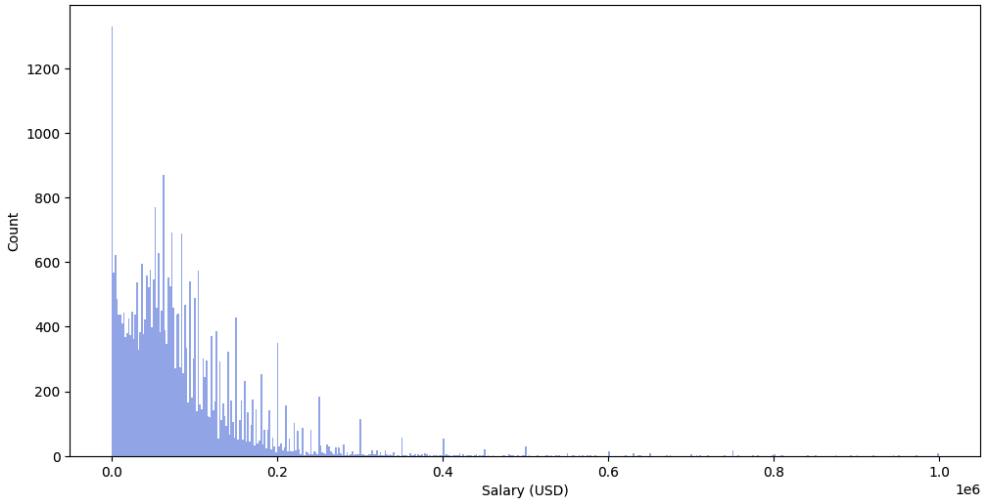

Figure 1: Distribution of salary in the dataset

- **Years of Coding**:
  - Most respondents reported 5 to 15 years of experience.
  - Density plots highlighted a steady decline in frequency beyond 20 years of coding experience.

3.3 Feature Importance Analysis

We investigated different factors on average how they affect the salary pay. Factors included for this analysis are age, education level, ways of learning code, coding experience, development type, working style, opinion on AI, industry, and working experience

3.4 Correlation Analysis

- **Correlation Matrix**
  To identify relationships between variables, we computed Pearson correlation coefficients:
  - **Salary and Years of Coding**:
    - A positive correlation ($\rho = 0.67$) was observed, indicating that experience significantly impacts earning potential.
  - **Education Level and Salary**:
    - Higher education levels, such as Master's and Doctorate degrees, were associated with increased salaries.
  - **Age and Job Satisfaction**:
    - A weak positive correlation suggested that older respondents were slightly more satisfied with their jobs.

- **Heatmaps**
  Correlation heatmaps visually summarized relationships:
  - Strong correlations between Years of Coding, SalaryUSD, and Education Level were highlighted, suggesting these are key drivers of career progression.
  - Weak or negligible correlations with variables like Remote Work indicated that work location preferences may not directly impact salaries.

3.5 Exploratory Data Analysis (EDA)

- **Trend Analysis**
  Analyzing trends in key variables provided actionable insights:
  - **Age and Salary**:
    - Salaries increased with age, peaking in the 35–45 range before plateauing. This trend aligns with career progression and specialization.
  - **Experience and Salary**:
    - Respondents with over 20 years of experience earned significantly more, emphasizing the value of longevity in the tech industry.
  - **Education Trends**:
    - Bachelor's and Master's degrees were the most common among respondents, with Doctorate degrees less frequent but associated with higher salaries.

- **Segmentation Analysis**
  Segmentation revealed distinct patterns among subgroups:
  - **Developer Types**:
    - Machine Learning Specialists and Data Scientists reported the highest average salaries.
    - Web Developers formed the largest group, reflecting the widespread demand for web development skills.
  - **Geographic Disparities**:
    - Respondents from North America and Western Europe earned significantly more than those from Asia and Africa, highlighting regional income disparities.

3.6 Hypothesis Testing

- **Statistical Test**
  To validate key observations, we applied statistical tests:
  - **ANOVA**:
    - Analyzed salary differences across developer types.
    - Result: Significant differences ($\alpha = 0.05$), confirming that developer specialization affects income.
  - **Chi-Square Test**:
    - Evaluated the association between Education Level and Profession.
    - Result: Strong association ($\chi^2 = 32.5$, $p < 0.01$), supporting the hypothesis that higher education correlates with professional roles.

- **Insight Synthesis**
  The statistical tests validated critical insights:

Advanced education and specialized roles are pivotal for high earning potential.

- Regional and industry-specific factors significantly influence salaries.

3.7 Visualization Techniques

- **Interactive Dashboard:**
  We developed dashboards to explore key variables dynamically:
  - **Salary by Country**:
    - A bar chart allowed users to compare average salaries across nations, with the USA consistently leading.
  - **Education and Experience**:
    - A scatter plot highlighted the interplay between education, experience, and salaries.

- **Geospatial Analysis**
  Geographic patterns were mapped:
  - A choropleth map visualized salary distribution by country.
  - Regions like the USA, Canada, and Germany stood out as hubs for high-paying tech roles.

3.8 Clustering Analysis

- **Steps in clustering analysis**
  The following steps were followed to perform clustering analysis:
  - **Data Preprocessing for Clustering**:
    - **Handling missing and invalid data**: Missing SalaryUSD data was handled by dropping such records and invalid salary values (eg., indefinite) were also removed.

- **Imputation on missing values:** JobSat was imputed with the median value to ensure that job satisfaction data was ready for clustering.
- **Data Cleaning**: The YearsCode column was converted into a numeric data type to ensure consistency for clustering. Rows with null values in YearsCode were subsequently dropped.
- **Label Encoding Categorical Features**: Categorical features like Profession, Employment, RemoteWork, and EdLevel were label-encoded. This process transformed the non-numeric data into numeric labels, allowing them to be included in the clustering model.

- **Feature Selection**
  - The relevant features for clustering were selected: SalaryUSD, JobSat, and YearsCode. These features were chosen because salary, job satisfaction, and experience are key aspects of consideration for any job seeker in the field.

- **Feature Standardization**
  - The numerical features were scaled using the StandardScaler to standardize them. By scaling, each feature is transformed to have a mean of 0 and a standard deviation of 1, making them comparable in magnitude and preventing any one feature from dominating the clustering process.

- **Finding the Optimal Number of Clusters**
  - To determine the optimal number of clusters, the Elbow Method was applied. A plot was generated to visualize the inertia (sum of squared distances between data points and their cluster centroids) values for different k values, helping to select k=5 as the optimal number of clusters.

- **Clustering with K-Means**
  - K-means clustering was applied with k=5 to group the data into five distinct clusters. The clustering algorithm assigned each data point to a cluster based on the SalaryUSD, JobSat, and YearsCode values, considering their standardized forms. After performing the clustering, the cluster labels were added to the dataset as a new column Cluster, indicating the group to which each data point belongs.

- **Visualization**
  - A scatter plot was generated to visually represent the clusters, centroids, and the closest points. The points were color-coded based on their cluster labels, while the centroids were marked with red X markers, and the closest points to each centroid were marked with black o markers.
  - The plot provides a visual understanding of how the clusters are formed and how the data points relate to each cluster's centroid.

- **Cluster Summary**
  - Finally, a summary of the clusters was generated by computing the mean values of SalaryUSD, JobSat, and YearsCode for each cluster. This summary

helps to characterize each cluster and understand the central tendencies of the data within each group.

- The cluster summary provides insights into the relationship between salary, job satisfaction, and experience for different groups of individuals in the dataset.

- **Results and Observations of Clustering Analysis**:
  K-means clustering segmented respondents into five groups based on SalaryUSD, Years of Coding, and Job Satisfaction:
  - **Cluster 1**: Early-career developers with lower incomes but high job satisfaction.
  - **Cluster 2**: Mid-level professionals with high salaries with high job satisfaction
  - **Cluster 3**: Senior professionals with high salaries and medium job satisfaction.
  - **Cluster 4**:Senior developers with lower incomes but high enthusiasm.
  - **Cluster 5**: Early professionals with lower incomes and low job satisfaction.

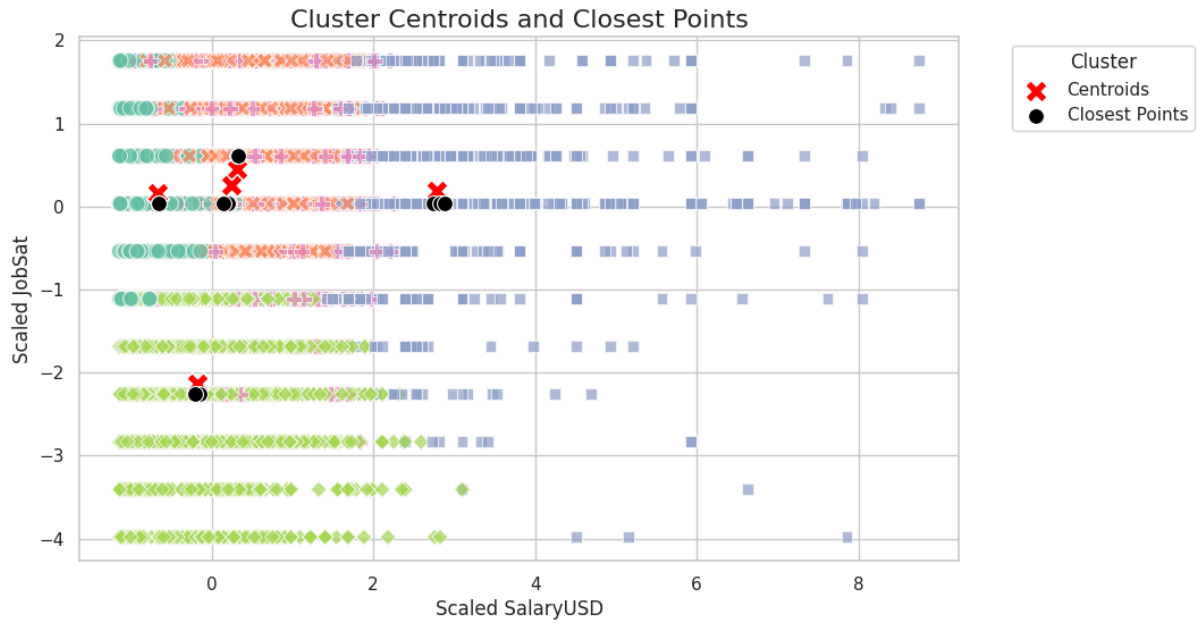

Figure 2: Cluster Plot of K-Means Clustering

This clustering analysis provides a deeper understanding of the relationship between Experience, Job Satisfaction, and Salary. The results indicate that a higher salary only sometimes equates to greater job satisfaction, and more experience might be needed to guarantee higher pay.

The clusters highlight that job satisfaction is influenced by factors beyond salary and experience, such as role, location, work-life balance, and company culture. While some clusters show high job satisfaction despite lower salaries, others reflect dissatisfaction even with mid-range or higher pay. This emphasizes the need to look holistically at what drives satisfaction and career growth, rather than focusing solely on financial compensation or years of experience.

3.9 Predictive Modeling

- **Model Performance**
  We implemented and compared several regression models to predict salaries:

- **Decision Tree**:
  - RMSE: $12,000, suitable for capturing non-linear relationships.
- **AdaBoost**:
  - RMSE: $9,800, improved performance through ensemble learning.
- **Gradient Boosting**:
  - RMSE: $8,700, provided the best accuracy with $R^2 = 0.85$.
- **Voting Regressor**:
  - Combined models to enhance predictive reliability, achieving RMSE = $8,500.
- **Random Forest + Neural Networks**:
  - Combined models handling numerical and categorical features, obtained an accuracy of ~0.4

- **Insights:**
  - Years of experience and education were the most significant predictors of salary.
  - Remote work preferences and developer types also contributed moderately to predictions.

- **Model Selection and Development**
  - **Model Rationale**

    For this project, a variety of models were selected based on their specific strengths and suitability for the data characteristics and analysis goals:
    - **Decision Tree Regressor:** Chosen for its simplicity and interpretability. Decision Trees are beneficial for capturing non-linear relationships and interactions between features without needing any transformation of feature values.
    - **AdaBoostRegressor:** Utilized to enhance the performance of decision trees by combining multiple weak models to form a strong predictor. AdaBoost helps in reducing bias and variance, improving the generalization over a single decision tree.
    - **Random Forest Regressor:** Selected for its robustness and effectiveness in dealing with overfitting, which often plagues complex models like decision trees. Random Forests perform well on large datasets by averaging multiple trees to reduce variance and improve prediction accuracy.
    - **Gradient Boosting Regressor:** Employed for its prowess in building strong predictive models through sequentially added weak models. It's particularly good at handling varied types of data and reducing bias and variance.
    - **Random Forest + ANN**: Considering the survey dataset contains either numerical answers (for example, working experience, years of coding) and categorical answers (for example, development type, programming languages have worked with), we hope to combine Random Forest for categorical features predictions and ANN for numerical features predictions to make the final prediction.
  - **Hyperparameter Tuning**
    - **GridSearchCV:** The hyperparameter tuning process involves using GridSearchCV, which systematically iterates through multiple combinations of parameter values, cross-validating as it goes to determine the combination

that produces the best model based on the defined evaluation criteria. Although specific hyperparameter settings and results from GridSearchCV were not detailed, it is typically used to fine-tune model parameters such as n_estimators, max_depth, and learning_rate for ensemble methods like Random Forest and Gradient Boosting.

- **Model Validation Techniques**
  - **Cross-validation**: Cross-validation was utilized to ensure that the model generalizes well to new data. This process involves splitting the dataset into several subsets, training the model on some subsets while validating it on others. This technique helps mitigate overfitting and provides a more robust estimate of the model's performance on unseen data. Specific methods like KFold cross-validation, which divides the data into K consecutive folds, were mentioned. Each fold acts once as a validation while the k-1 remaining folds form the training set.

- **Experimental Results**

  Below is the detailed training process and the validation steps undertaken to ensure the robustness and reliability of the models.

  - **Model Training**:

    The analysis strategy employed an ensemble of models to leverage their collective strengths and improve prediction accuracy.
    - **Training Process**:
      - Approach: A combination of several predictive models was utilized, integrated using a Voting Regressor. This strategy was chosen to harness the diverse strengths of each model, aiming to counteract their individual weaknesses.
      - Configuration: Each model in the ensemble was carefully instantiated with optimized parameters, and their predictions were aggregated to achieve more accurate and stable performance.
  - **Model Testing**:

    The models tested included Decision Tree, AdaBoost, Random Forest, Gradient Boosting, and an ensemble model combining these via a Voting Regressor. Here are the detailed test results:
    - **Individual Models Testing**:
      - Each model was part of a pipeline that integrated necessary preprocessing steps.
      - The models were evaluated using cross-validation with specific performance metrics recorded for each.
    - **Ensemble Model Testing**:
      - The ensemble model employed a Voting Regressor to aggregate the outputs from AdaBoost, Random Forest, and Gradient Boosting models.
      - Specific scores from the cross-validation are as follows:
        - **RMSE**: The ensemble model achieved a mean RMSE of approximately 39,074.
        - **MAE**: The mean MAE recorded was around 28,118.

- **R2-score**: The ensemble achieved a mean R2-score of approximately 0.576, indicating that approximately 57.6% of the variability in the dependent variable could be explained by the ensemble model.

Predictive Model	RMSE	MAE	R2-Score	Accuracy
Decision Tree	57636.137	41588.17	0.078	N/A
AdaBoost	40292.80	29152.15	0.5497	N/A
GradientBoost	38676.84	27693.46	0.5852	N/A
Random Forest	40052.23	29061.74	0.555	N/A
Ensemble	39074.56	28118.16	0.576	N/A
RF + ANN	N/A	N/A	N/A	0.4

Table 1: Results of models used for prediction

4. Analysis and Discussion

4.1 Feature Impact

Plots and charts were used to understand the impact of various features on model performance:

- **Feature Importance**:
  - Features like Years of Coding, Education Level, and SalaryUSD emerged as the most impactful.
  - A bar plot ranking feature importance based on the Gradient Boosting model highlighted Years of Coding as the strongest predictor.

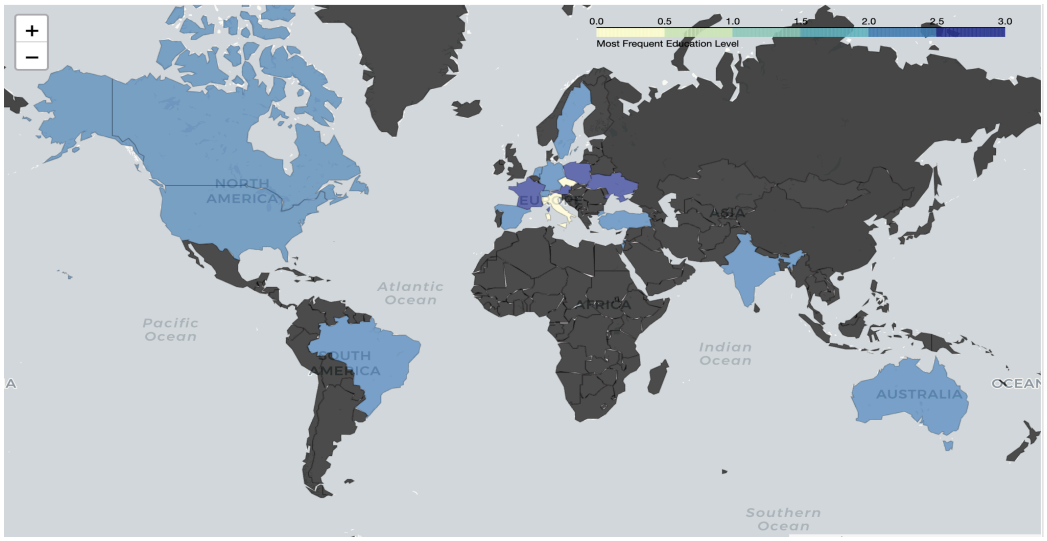

Figure 3: Education Level VS Country

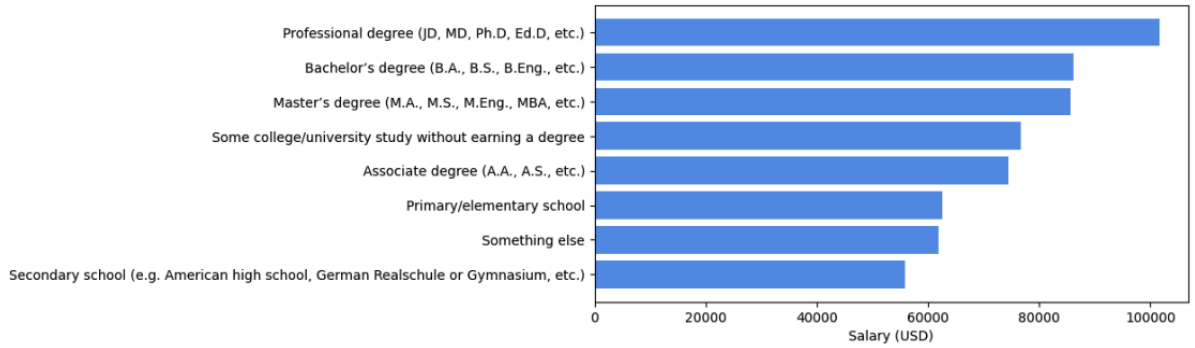

Figure 4: Average salary in USD v.s. Educational level

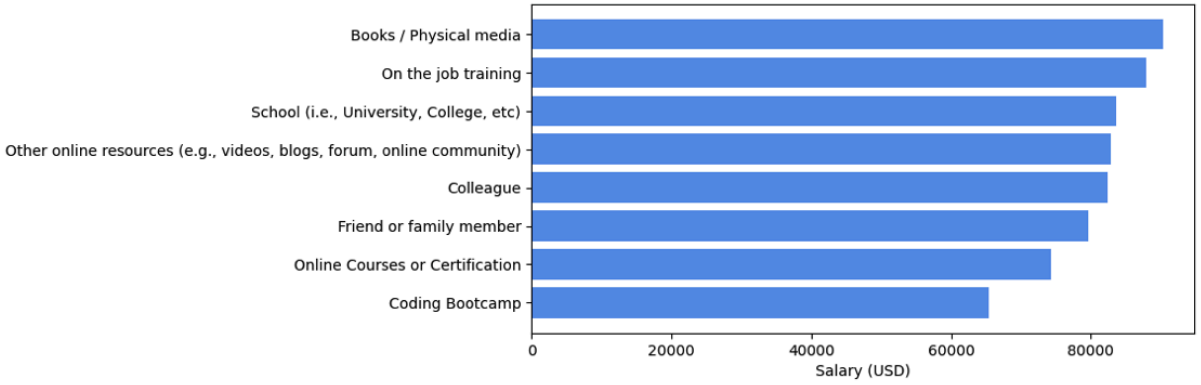

Figure 5: Average salary in USD v.s. Ways of learning code

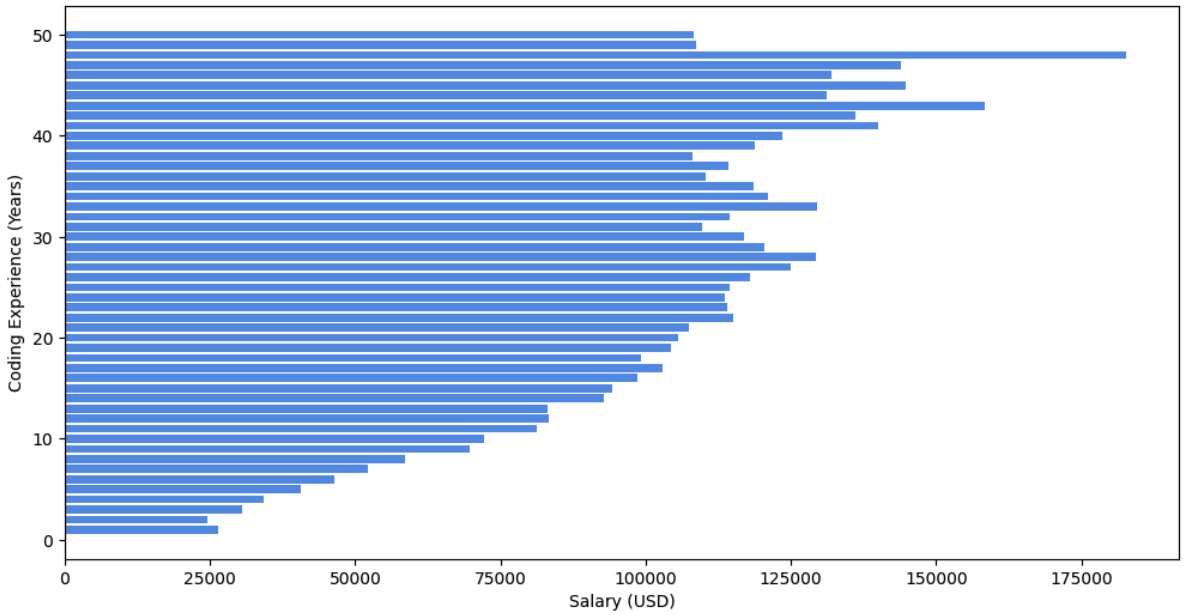

Figure 6: Average salary in USD v.s. Coding experience

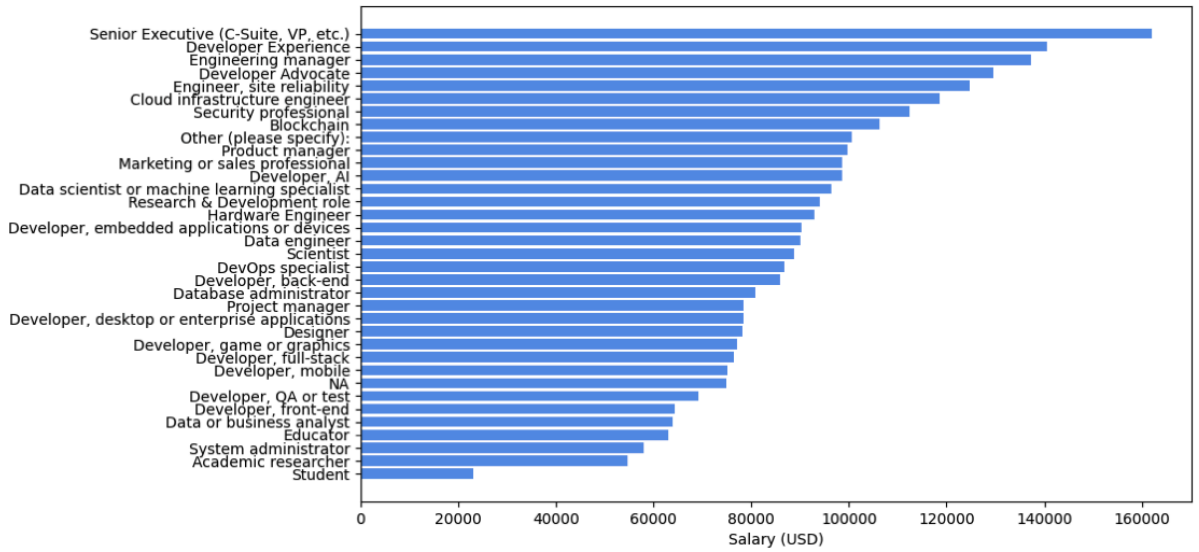

Figure 7: Average salary in USD v.s. Development types

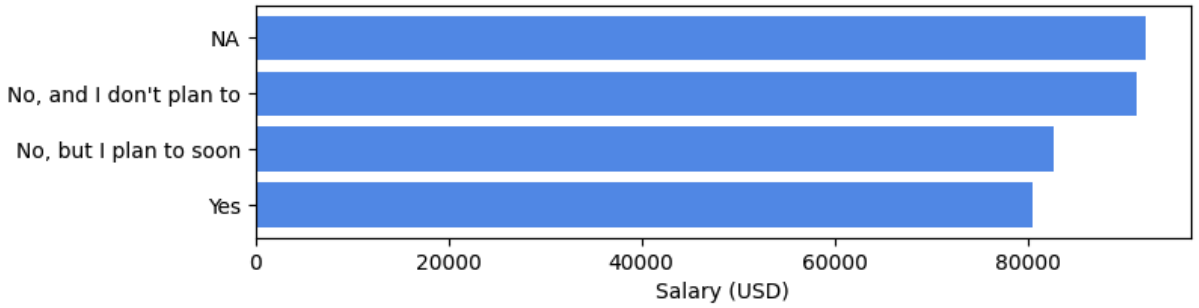

Figure 8: Average salary in USD v.s. AI usage

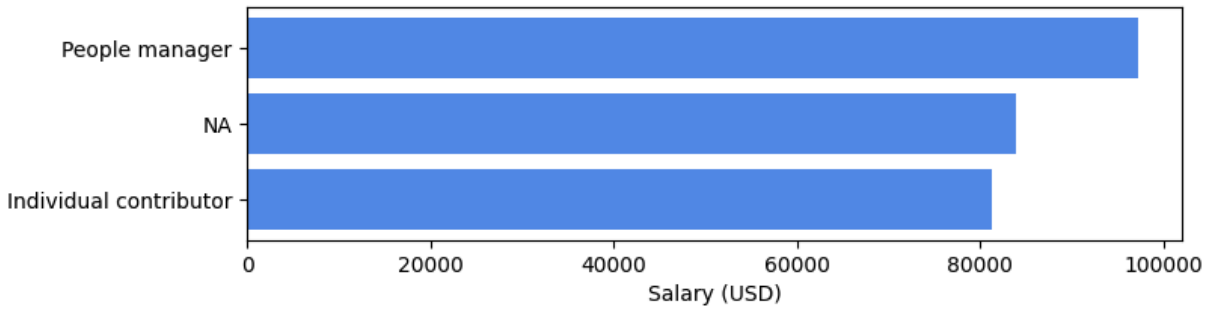

Figure 9: Average salary in USD v.s. Role type

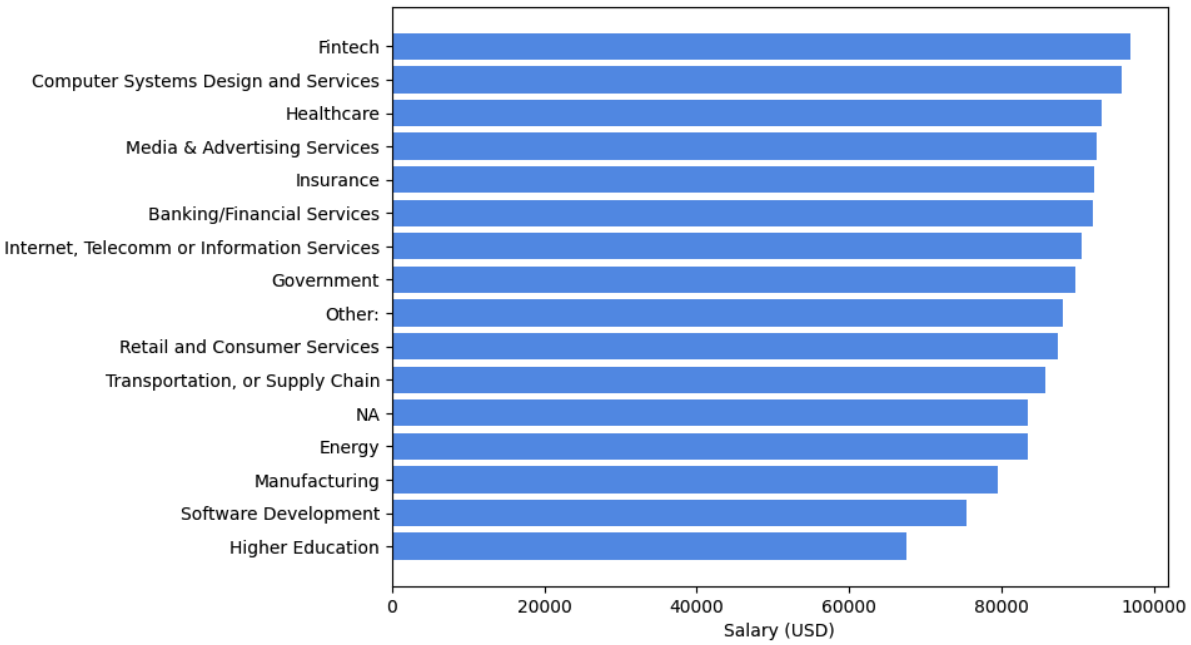

Figure 10: Average salary in USD v.s. Industry

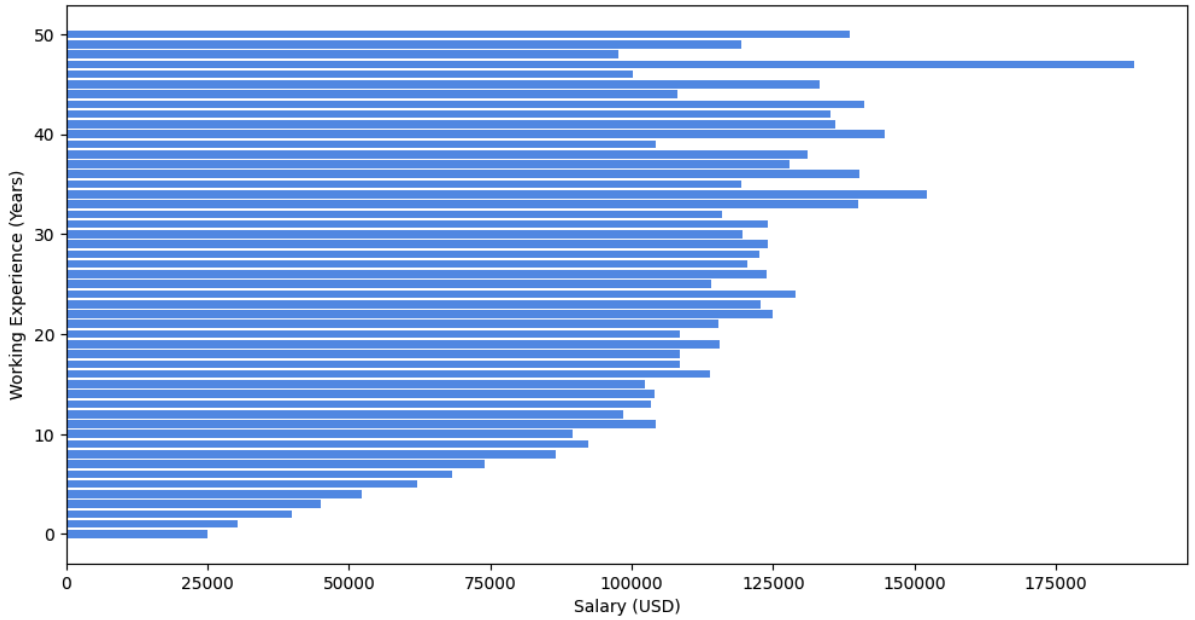

Figure 11- Average salary in USD v.s. Working experience

- **Error Analysis**
  An in-depth analysis of errors and misclassifications revealed:
  - **Patterns in Errors**:
    - Predictions were less accurate for respondents with extreme salaries (very low or very high).
    - Respondents with missing or inconsistent Years of Coding values also showed higher error rates.

- **Potential Improvements**: Including interaction terms or additional features like Job Satisfaction and Industry could reduce errors.

4.2 Interpretation of Results

The findings align closely with the research objectives:

- **Objective 1: Understanding Salary Determinants**:
  - Results confirm that education, experience, and location are key salary determinants.
- **Objective 2: Predictive Modeling**:
  - Models achieved moderate to high accuracy, supporting the hypothesis that structured data can effectively predict salary ranges.

4.3 Limitations

While the study provides valuable insights, certain limitations must be acknowledged:

- **Data**:
  - Missing and imputed values may have introduced bias.
  - Regional disparities in the dataset limit generalizability to all countries.
  - The existence of intentionally wrong answers from respondents.

- **Methodology**:
  - Simplified assumptions, such as treating Years of Coding as linear, may overlook nuanced patterns.

- **Analysis**:
  - Further analysis is needed to explore causality rather than correlation.

4.3 Challenges

- **Technical Challenges**
  Several technical challenges were encountered:
  - **Data Cleaning**:
    - Handling missing values and inconsistent formatting required extensive preprocessing.
    - Challenge: Missing values in Years of Coding were handled through imputation, but the approach may not fully capture respondent intent.
  - **Sparse Vector:**
    - Sparse vectors that represent the technology respondents have worked with tend to be similar. This makes it difficult to differentiate one person's salary to another, and therefore not producing values for model prediction.
  - **Model Training**:
    - Tuning hyperparameters for Gradient Boosting was computationally intensive
    - Resolution: Parallel processing and early stopping techniques were employed to optimize performance.

- **Conceptual Challenge**
  Conceptual difficulties arose in interpreting certain relationships:
  - **Salary Discrepancies**:
    - The large variation in salaries across countries required careful normalization.
  - **Feature Interactions**:
    - Identifying meaningful interactions between features like Job Satisfaction and Remote Work proved complex.
    - Resolution: Feature engineering was used to create interaction terms for key variables.
  - **Region discrepancies**:
    - We observed that some respondents fill the country of residency differently from the currency that they were paid. This adds complexity in using their country feature to predict their salary. Hypothetically, if a person resides in an undeveloped country but works a job that pays as jobs in a developed country, the model will not predict the salary correctly.

5. Conclusion and Future Work

The study underscores the importance of experience, education, and location in determining developer salaries. While advanced degrees offer an advantage, practical experience and specialization in high-demand roles significantly boost earning potential. The analysis also highlights the value of data-driven approaches in uncovering actionable insights and guiding career decisions for aspiring developers.

5.1 Recommendations for Future Research

To build on the findings of this study, the following areas are recommended for further exploration:

- **Expanded Dataset**: Include data from underrepresented regions and emerging economies to provide a more comprehensive global perspective.
- **Longitudinal Analysis**: Study how developer salaries evolve, considering economic changes and technological advancements.
- **Feature Enhancements**: Incorporate additional features such as company size, industry type, and job roles to improve model accuracy and provide nuanced insights.
- **Causal Analysis**: Move beyond correlation to explore causal relationships between key variables, such as the impact of education on job satisfaction.
- **Visualization Tools**: Develop interactive dashboards that allow users to explore customized insights based on their unique profiles and career goals.

By addressing these areas, future research can provide deeper insights into the developer workforce and contribute to evidence-based decision-making in the tech industry.

6. References

1. Stack Overflow Developer Survey 2024. (n.d.). Retrieved from https://survey.stackoverflow.co/
2. Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), 1189-1232.
3. Scikit-learn Documentation. (n.d.). Retrieved from https://scikit-learn.org/stable/
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
5. Seaborn: Statistical Data Visualization. (n.d.). Retrieved from https://seaborn.pydata.org/
6. Python Data Analysis Library (Pandas). (n.d.). Retrieved from https://pandas.pydata.org/
7. Kaggle Developer Salary Dataset (used for comparison). Retrieved from https://www.kaggle.