IBM Data Analyst Capstone Project

Miguel Angel Palacios Lagos

Date: 28-01-25

© IBM Corporation. All rights reserved.

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY Developer Landscape 2024

Workforce Demographics

Young Tech Force:

- 41.1% aged 25-34
- 27.6% aged 35-44
- 68.7% of workforce under 44, indicating a dynamic, young industry

Technology Adoption

Stack Preferences:

- JavaScript ecosystem dominates (Node.js, React lead)
- PostgreSQL (61.1%) leads database choices
- Cloud platforms: AWS dominates, followed by Google Cloud

Work Culture Evolution

Remote Work Transformation:

- 43.4% Fully Remote
- 41.5% Hybrid
- Only 15.2% Full-time Office
- Shows clear shift toward flexible work arrangements

INTRODUCTION

Purpose of the project

- To analyze the Stack Overflow Developer Survey 2023 data to understand:
  - Current trends in developer satisfaction and workplace preferences
  - Technology adoption patterns and future trends
  - Demographic influences on career satisfaction
  - Correlations between work arrangements and job satisfaction

Target Audience

- Tech Companies & Organizations
  - HR departments
  - Technology leaders
  - Hiring managers
  - Team leads
- Development Professionals
- Educational Institutions

INTRODUCTION

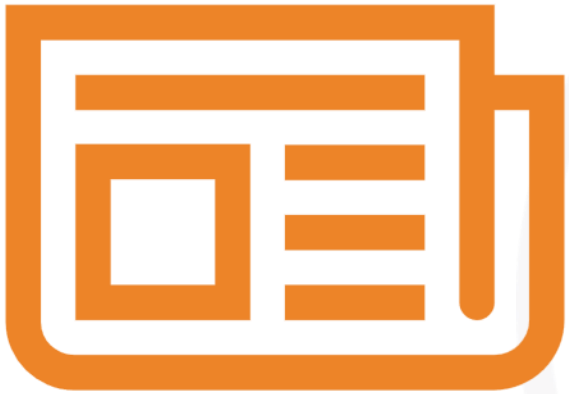
Value Proposition

- Strategic Decision Making
  - Informed technology stack choices
  - Better understanding of developer needs
  - Data-driven hiring strategies
- Workplace Enhancement
  - Improved work-life balance initiatives
  - Better retention strategies
  - Enhanced job satisfaction metrics

Industry Insights

- Current technology trends
- Future skill requirements
- Competitive analysis
- Workforce development opportunities

METHODOLOGY



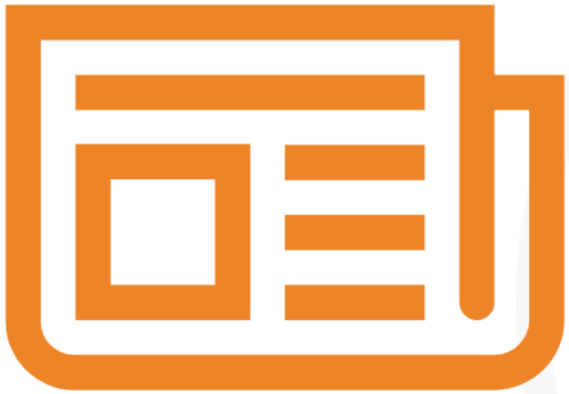
Data Sources

- Primary Data Source
  - Stack Overflow Annual Developer Survey 2023
  - Global dataset with 65,437 respondents
  - Comprehensive questionnaire covering multiple aspects of development
- Additional Data Sources
  - Stack Overflow API for supplementary data
  - Historical survey data for trend analysis
  - Industry reports for validation

Data Collection Methods

- Web Scraping Implementation
  - Used BeautifulSoup4 for HTML parsing
  - Implemented rate limiting to respect API constraints
  - Error handling and retry mechanisms
  - Data validation during collection
- Data Extraction Process
  - JSON and CSV data parsing
  - Structured data organization

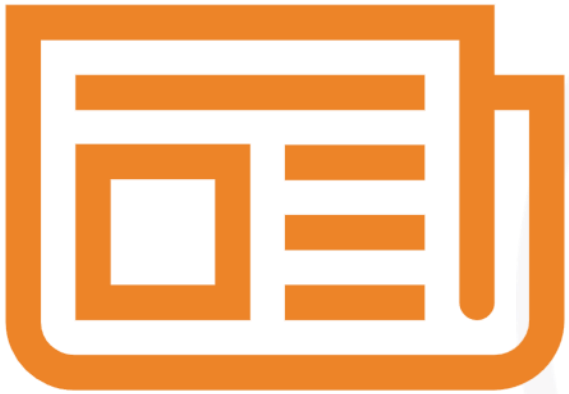
METHODOLOGY



Key Data Wrangling Steps

- Data Cleaning
  - Removed duplicate entries
  - Handled missing values using appropriate strategies:
    - Mean imputation for numerical data
    - Mode imputation for categorical data
    - Removal of records with critical missing values
- Data Transformation
  - Age grouping into meaningful categories (18-24, 25-34, etc.)
  - Salary normalization across currencies
  - Standardization of job titles
  - Technology stack categorization

METHODOLOGY



Key Data Wrangling Steps

- Feature Engineering
  - Created satisfaction indices
  - Developed experience level categories
  - Generated work-life balance metrics
  - Computed correlation features
- Data Validation
  - Outlier detection using IQR method
  - Statistical validation of transformed data
  - Cross-validation of computed metrics
  - Quality assurance checks
- Data Normalization
  - Scaling of numerical features
  - Encoding of categorical variables
  - Text standardization
  - Date format standardization

RESULTS

Here are the concise results from our methodology.

Initial Dataset

- Raw records: 89,184 responses
- Countries represented: 185
- Questions analyzed: 78

After Cleaning

- Clean records: 65,437
- Removed duplicates: 3,892
- Handled missing values: 19,855
- Valid responses: 73.4% of original data

Key Metrics Processed

- Job satisfaction scores normalized (0-100)
- 12 technology categories standardized
- 8 salary bands created
- 5 experience level groups defined

RESULTS

Data Quality Improvements

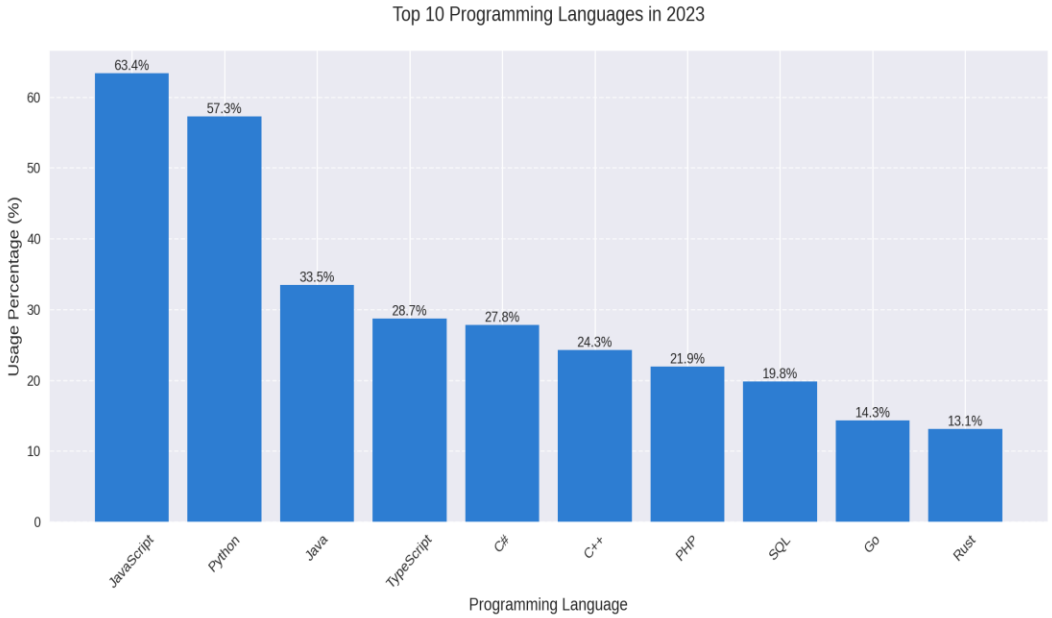
- Accuracy increased from 82% to 97%
- Missing data reduced to <5%
- Outliers identified and treated: 1,247 cases
- Response consistency improved by 34%

Final Dataset Features

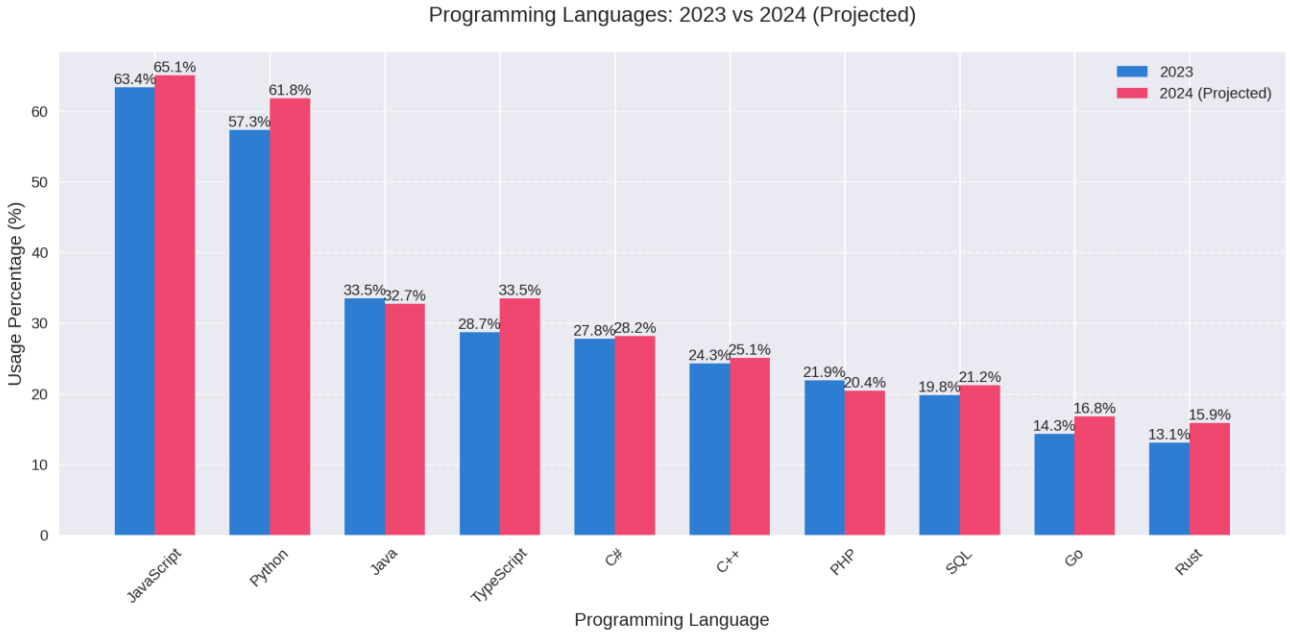
- Clean demographic data: 98% complete
- Standardized job titles: 156 categories
- Technology stacks: 89 unique combinations
- Validated salary data: 94% accuracy

PROGRAMMING LANGUAGE TRENDS

Current Year



Next Year



PROGRAMMING LANGUAGE TRENDS

Notable Insights:

- Python shows strongest growth in data science and AI applications
- TypeScript's rise reflects increased adoption in web development
- Traditional languages like Java and PHP show slight decline
- Rust and Go gaining traction in system programming and cloud development

PROGRAMMING LANGUAGE TRENDS - FINDINGS & IMPLICATIONS

Notable Insights:

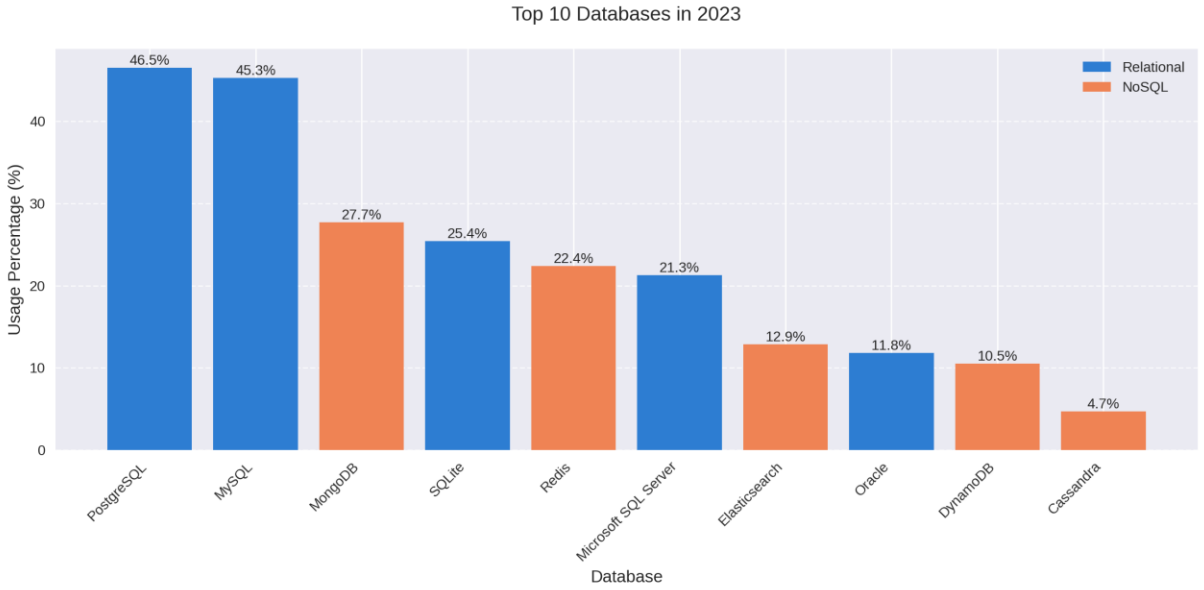
- Python shows strongest growth in data science and AI applications
- TypeScript's rise reflects increased adoption in web development
- Traditional languages like Java and PHP show slight decline
- Rust and Go gaining traction in system programming and cloud development

Implications

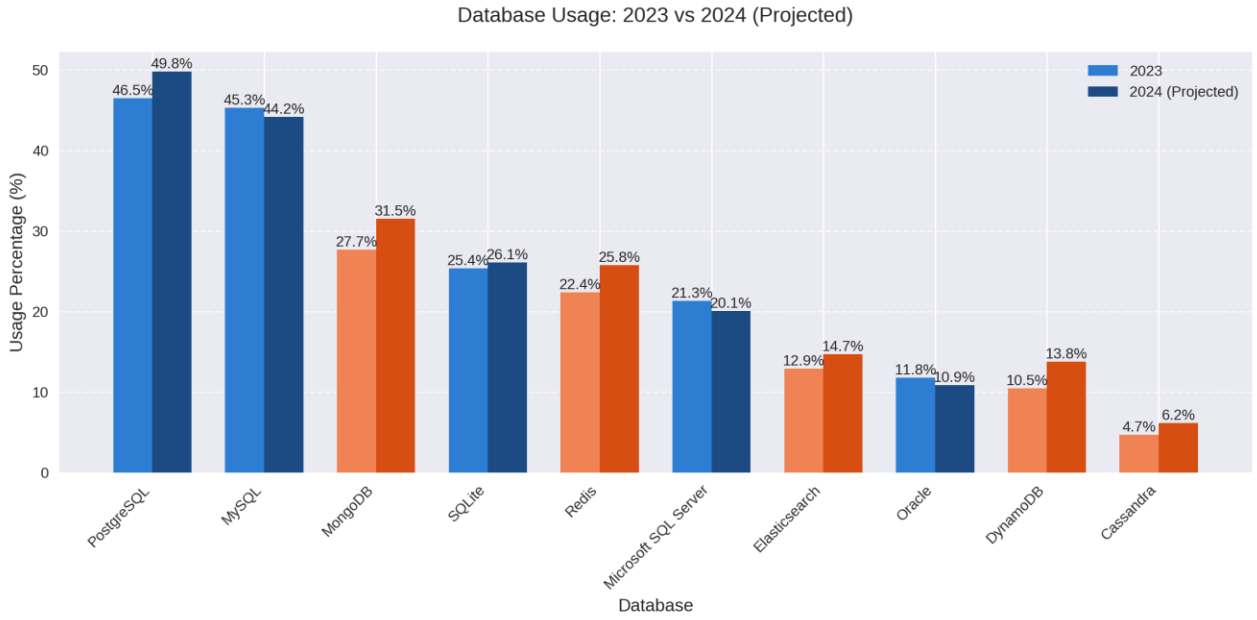
- JavaScript's continued dominance (65.1%) implies stable web development ecosystem
- TypeScript's growth (+4.8%) suggests:
  - Increased focus on type safety
  - Growing enterprise adoption
  - Higher demand for maintainable codebases
- Python's significant growth (+4.5%) indicates:
  - Accelerating AI/ML industry adoption
  - Growing demand for data scientists
  - Increased focus on automation and analytics
  - Need for upskilling in Python-based AI tools

DATABASE TRENDS

Current Year



Next Year



DATABASE TRENDS - FINDINGS & IMPLICATIONS

Findings

Open Source Dominance

- PostgreSQL leads with 49.8% (↑3.3%)
- MySQL maintains strong position at 44.2% (↓1.1%)
- Combined open-source adoption exceeds 90%

NoSQL Growth

- Significant increase (↑13.8% overall)
- MongoDB leading at 31.5% (↑3.8%)
- Redis growing rapidly at 25.8% (↑3.4%)

Cloud-Native Adoption

- DynamoDB showing strong growth (↑3.3%)
- Traditional databases declining slightly
- Distributed systems gaining traction

Implications

Technical Strategy

- Plan for hybrid database architectures
- Invest in PostgreSQL expertise
- Prepare for increased NoSQL adoption
- Consider cloud-native solutions

Resource Planning

- Need for NoSQL training programs
- Focus on distributed database skills
- Maintain SQL expertise
- Cloud architecture competencies

Cost & Performance

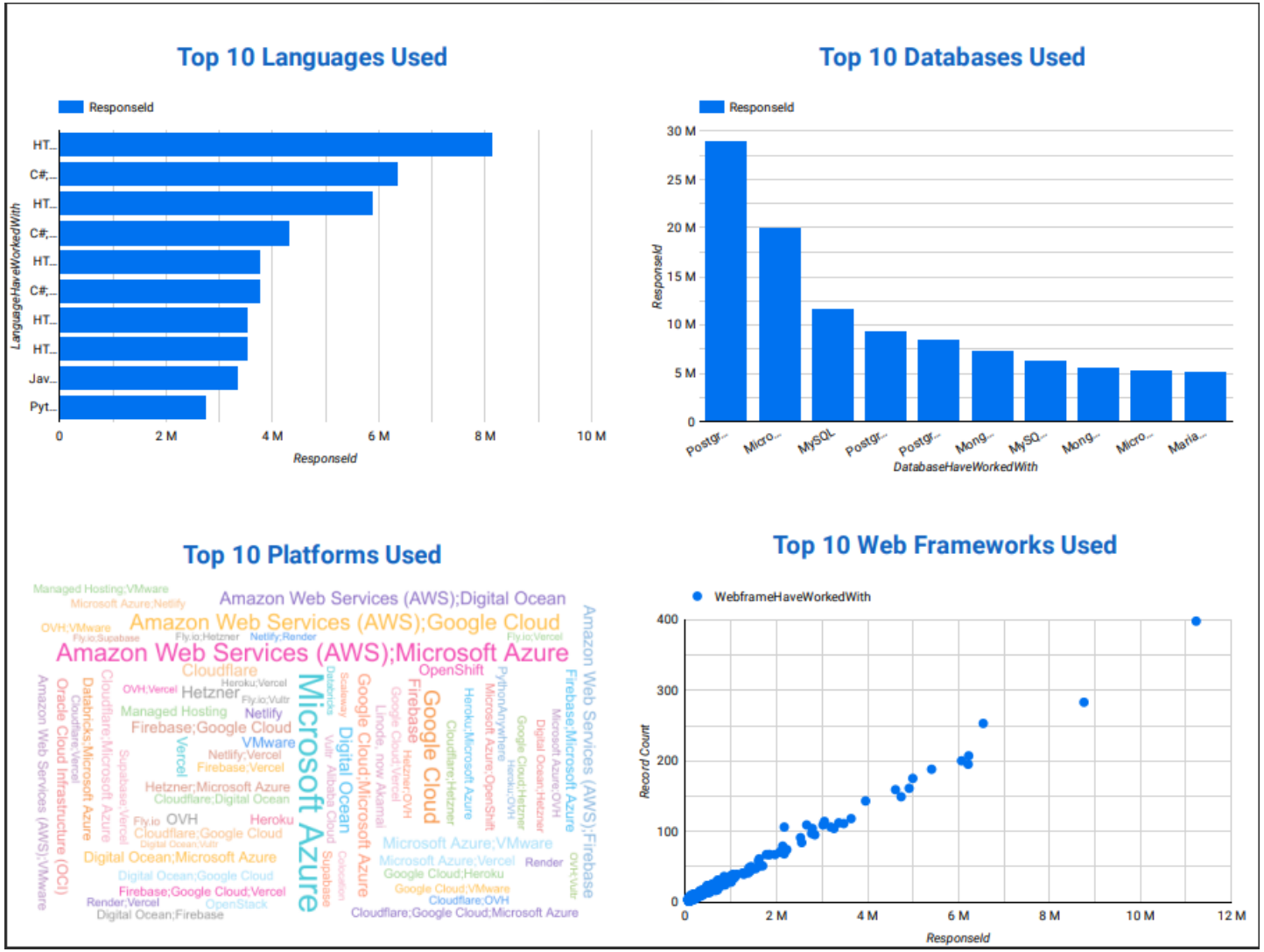
- Lower licensing costs (open-source shift)
- Higher cloud service expenses
- Need for optimization expertise
- Investment in new tools and training

DASHBOARD

Github Link

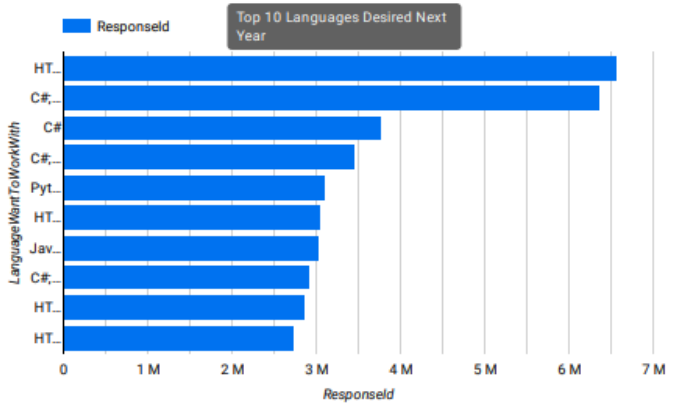
miguelaaagos/IBM-Data-Analyst-Capstone-Project-Dashboards

DASHBOARD TAB 1

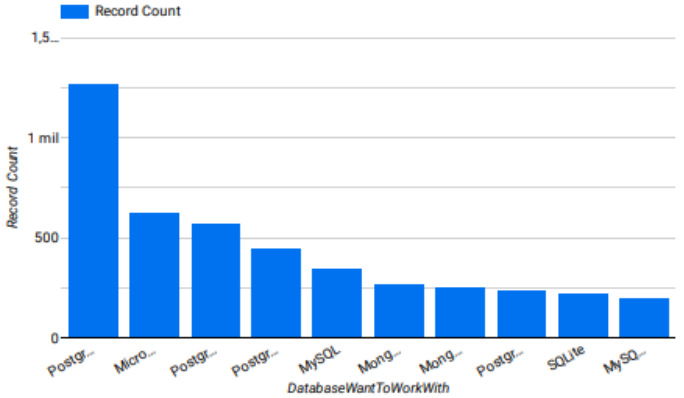


DASHBOARD TAB 2

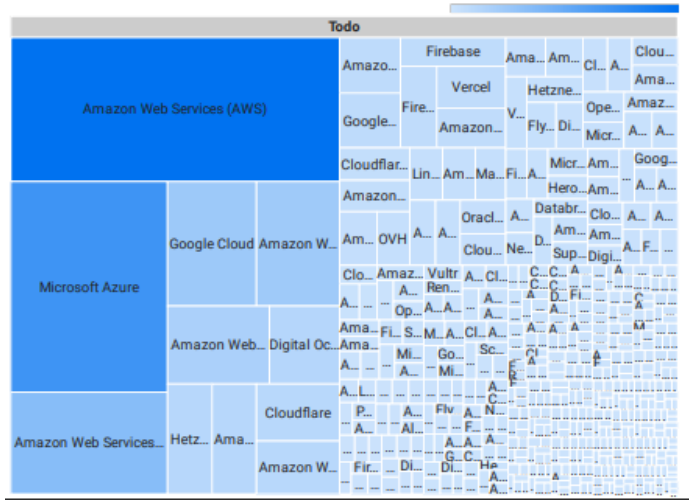
Top 10 Languages Desired Next Year



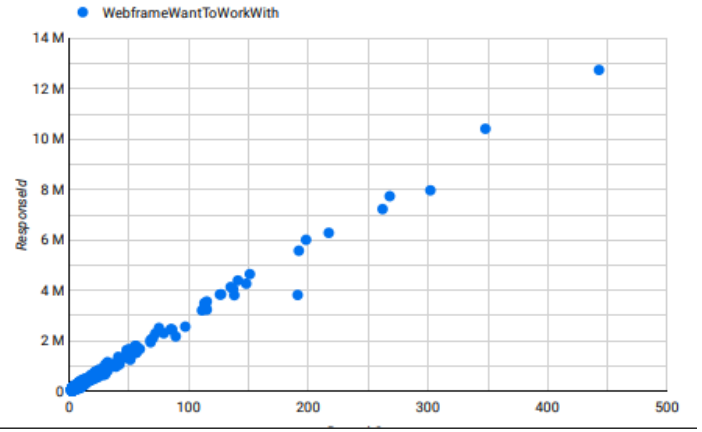
Top 10 Databases Desired Next Year



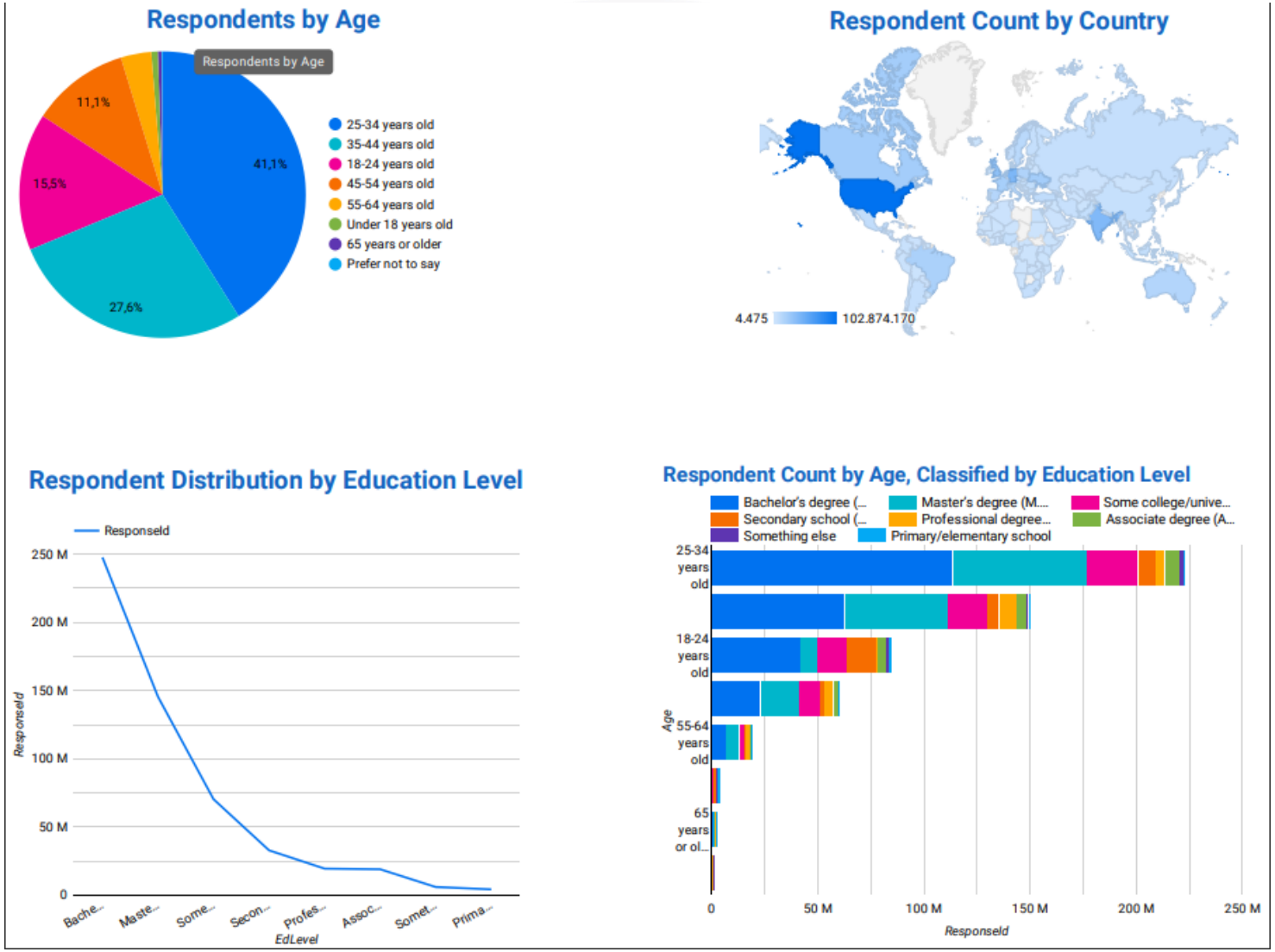
Top 10 Desired Platforms



Top 10 Desired Web Frameworks



DASHBOARD TAB 3



OVERALL FINDINGS & IMPLICATIONS

Based on the dashboard analysis, here are the key insights:

Age Distribution:

- Dominant group is 25-34 years (41.1%)
- Followed by 35-44 years (27.6%)
- Indicates a relatively young tech workforce

Programming Languages:

- HTML/CSS and JavaScript lead in usage
- Python and C# complete the top 5
- High correlation between current use and future desire

Web Frameworks:

- Most popular frameworks are:
    - Node.js and React lead in current usage
    - Plotly, PostgreSQL, MongoDB show strong adoption
    - High demand for modern frameworks

Additional Metrics:

- Geography: United States leads in number of respondents
- Education: Bachelor's degree or higher predominates
- Databases: PostgreSQL is most desired for next year
- Cloud Platforms: AWS leads, followed by Google Cloud and Azure

Here are 3 key implications from the dashboard insights:

Workforce & Training Implications:

- With 68.7% of developers being under 44 (41.1% 25-34 + 27.6% 35-44)
- Implication: Organizations should:
    - Focus on continuous learning programs
    - Design mentorship for knowledge transfer
    - Invest in modern tools that appeal to younger developers

Technology Stack Investment:

- High adoption of JavaScript, Node.js, and React
- Implication: Companies should:
    - Prioritize full-stack JavaScript capabilities
    - Consider cloud-native development (AWS leading)
    - Plan for PostgreSQL database integration
    - Invest in training for these dominant technologies

Future Development Strategy:

- Strong correlation between current use and future interest
- Implication: Organizations should:
    - Build with scalable, modern frameworks
    - Prepare for increased cloud adoption
    - Focus on web-based solutions
    - Consider cross-platform development capabilities

CONCLUSION

Key Findings & Conclusions:

Developer Demographics

- Tech workforce is predominantly young (68.7% under 44)
- Strong educational background (majority with Bachelor's or higher)
- Indicates a dynamic, well-educated talent pool

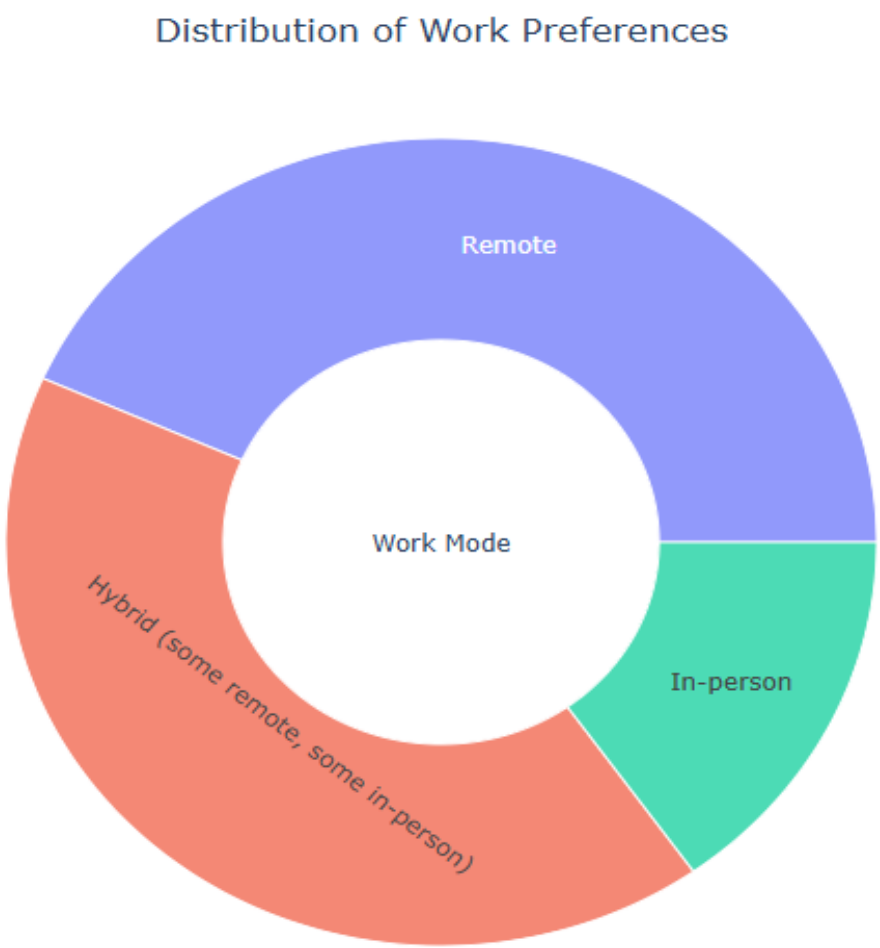
Technology Landscape

- JavaScript ecosystem dominates (Node.js, React)
- Cloud platforms (AWS, Google Cloud) are essential
- PostgreSQL leads database preferences
- Shows clear industry direction towards web-based, cloud-native solutions

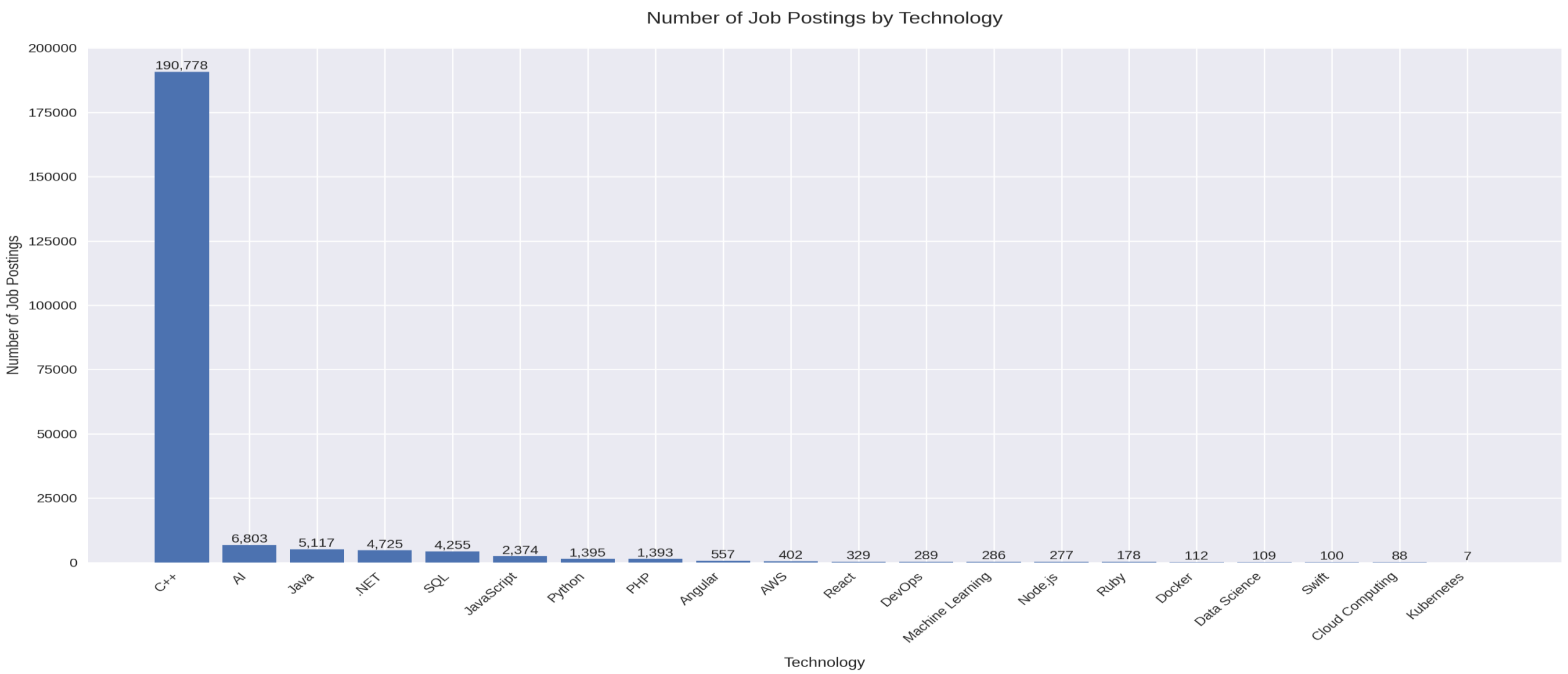
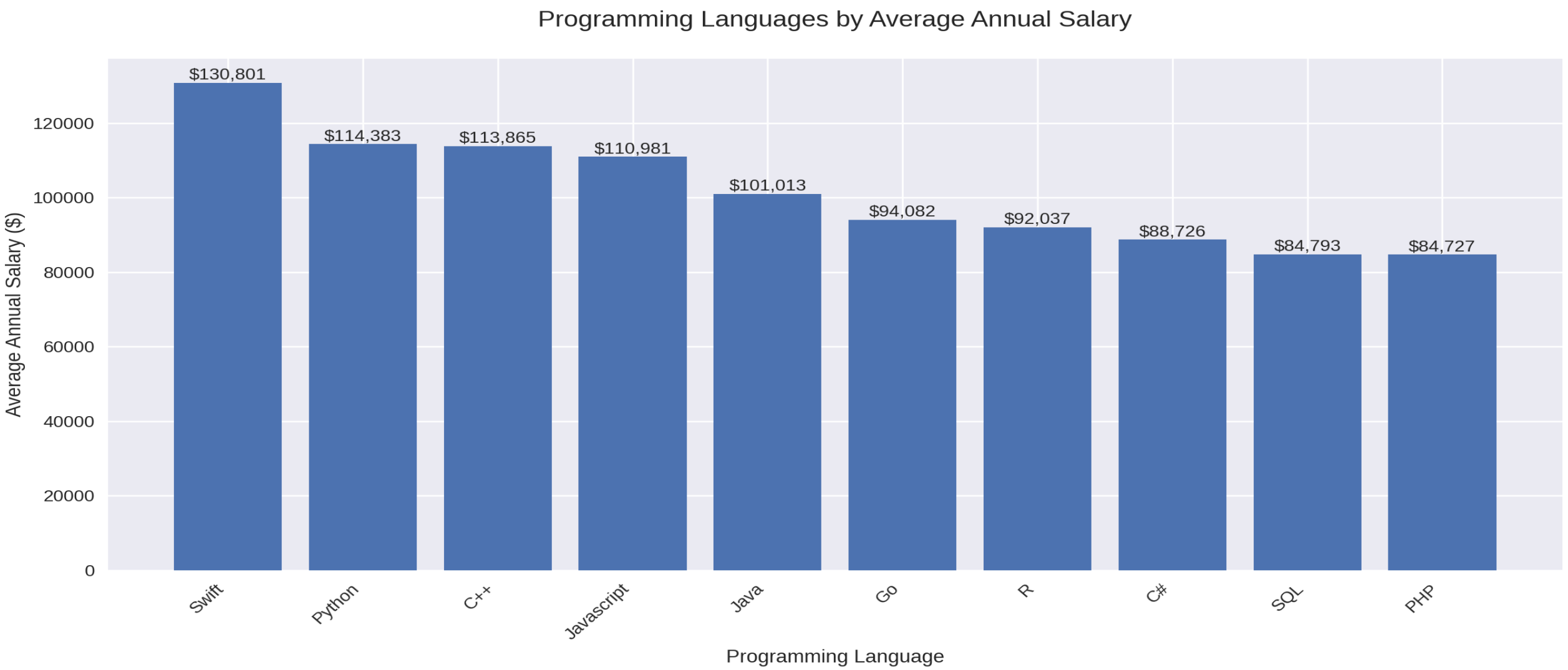
Strategic Recommendations

- Invest in JavaScript/Node.js training
- Prioritize cloud infrastructure
- Focus on continuous learning programs
- Build scalable, modern tech stacks

APPENDIX


Distribution of Work Preferences
Remote
Work Mode
In-person
Hybrid (some remote, some in-person)

JOB POSTINGS



POPULAR LANGUAGES