IBM Data Analyst Capstone Project

Nurliyana Binti Baharin

Date: 01-10-2025

OUTLINE

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

EXECUTIVE SUMMARY

- Top programming languages in demand:
  - JavaScript leads current use, followed by HTML/CSS and Python. Python is the most desired for future work, with Rust and Go emerging strongly.
- Top database skills in demand:
  - PostgreSQL dominates both current and future usage, while MySQL and SQLite remain strong. Redis and Elasticsearch are gaining traction.
- Popular platforms:
  - AWS is the most used and most desired cloud platform, ahead of Microsoft Azure and Google Cloud.
- Popular Web Frames:
  - Node.js and React dominate current and future preferences, with interest also in Angular, Express, and Vue.
- Future Technology Trend:
  - Python and TypeScript are rising stars; PostgreSQL stays central; AWS remains the go-to cloud; emerging tools like Rust, Go, and Supabase show accelerating adoption.

INTRODUCTION

Purpose of the Project

- Identify future technology skill requirements.
- Analyze demand for programming languages, databases, and IDEs.
- Provide insights to support business competitiveness.
- Apply data wrangling and analysis techniques.
- Visualize and present findings via dashboards and storytelling.

Target Audience

- Business and IT consultants.
- HR and recruitment teams.
- Training and education providers.
- Technology leaders and managers.
- Aspiring and current developers.

INTRODUCTION(cont.)

Value Proposition

- **Strategic Decision Making**
  - **Data-driven insights for smarter planning**
  - **Stay ahead with tech adoption trends**
  - **Invest resources where they matter most**
- **Workplace Enhancement**
  - Identify future technology skill requirements.
  - Analyze demand for programming languages, databases, and IDEs.
  - Provide insights to support business competitiveness.
  - Apply data wrangling and analysis techniques.
  - Visualize and present findings via dashboards and storytelling.

Industry Insight

- **Python & JavaScript rule** → but Go, Rust, and TypeScript are rising fast.
- **Databases evolve** → MySQL/PostgreSQL stay strong, NoSQL adoption surges.
- **Remote is the norm** → hybrid work reshapes hiring and culture.
- **VS Code leads** → modern IDEs drive developer productivity.
- **Skill gaps emerge** → AI, cloud, and cybersecurity talent in high demand.

METHODOLOGY

Data Sources

- **Primary Data Source**
  - 🗂️ Stack Overflow Annual Developer Survey 2024
  - 👥 Global dataset with 65,437 respondents
  - 📋 Comprehensive questionnaire covering multiple aspects of development
- **Additional Data Source**
  - 🔎 Web Scraping Results → e.g., extracting information from web pages (such as technology blogs or job boards).
  - 💰 Programming Language Salary Dataset → CSV file containing annual average salaries for different programming languages.
  - 🗄️ SQL Database Storage → intermediate structured data storage for querying and combining survey/job/salary data

METHODOLOGY(cont.)

Data Collection

- Goal: Build a **consolidated dataset** for analysis.
- **Web Scrapping Implementation**
  - Collect webpage content using request
  - Parsed HTML data with BeautifulSoup4
  - Extracted and validated relevant information
  - Saved clean data into structured formats (e.g., CSV, DataFrame)
- **Data Extraction Process**
  - JSON and CSV data parsing
  - Structured data organization

METHODOLOGY (cont.)

Key Features in Data Wrangling

- **Data Cleaning**
  - Purpose: Ensure dataset quality by removing irrelevant or incorrect entries.
  - Actions:
    - Drop exact duplicate survey responses.
    - Handle missing values (e.g., impute or drop).
      - Mean imputation for numeric fields.
      - Mode imputation for categorical fields.
- **Data Transformation**
  - Purpose: Make raw survey responses consistent and usable.
  - Actions:
    - Split multi-response fields (e.g., LanguageHaveWorkedWith → separate Python, SQL, Java).
    - Standardize inconsistent text values (United States vs. USA).
    - Convert years of experience from text ranges (“Less than 1 year” → 0.5).
  - Example: Transform YearsCodePro into numeric values for analysis.

METHODOLOGY (cont.)

Key Features in Data Wrangling

- **Feature Engineering**
  - Purpose: Create new insights from existing columns.
  - Actions:
    - Derive new features (e.g., ExperienceLevel from YearsCodePro).
    - Create binary flags (e.g., IsRemote = 1 if RemoteWork = Remote/Hybrid).
    - Multi-hot encoding for skills (LanguageHaveWorkedWith → Python=1, Java=0).
  - Example: Group countries into regions (North America, Europe, APAC).
- **Data Validation**
  - Purpose: Verify data accuracy and consistency after cleaning/transformation.
  - Actions:
    - Check valid ranges (e.g., no negative values in Age).
    - Validate categorical responses against predefined lists (e.g., Employment types).
    - Spot outliers in compensation data.
  - Example: Ensure ConvertedCompYearly is within realistic salary ranges.

METHODOLOGY (cont.)

Key Features in Data Wrangling

- **Feature Normalization**
  - Purpose: Scale numerical values for comparability across features.
  - Actions:
    - Min-Max scaling (values transformed to 0–1).
    - Z-score standardization (mean=0, std=1).
    - Log transformation for skewed data.
  - Example: Normalize ConvertedCompYearly before building dashboards or models.

DISCUSSION-DASHBOARD 1

Current Technology Usage

- **Most Used Programming Language**:
  - JavaScript tops the list with 37,492 respondents, followed closely by HTML/CSS (31,816) and Python (30,719).
- **Database Usage Trends:**
  - PostgreSQL is the most worked-with database (25,536 respondents), followed by MySQL (21,099) and SQLite (17,365).
- **Cloud Platforms:**
  - Amazon Web Services (AWS) is the most commonly used platform, outperforming Microsoft Azure and Google Cloud.
- **Web Framework Usage:**
  - Node.js leads with 19,772 respondents, followed closely by React (19,167), with other frameworks like jQuery, Express, and Angular trailing.
- **General Backend Dominance:**
  - Technologies like Node.js, SQL, and PostgreSQL indicate a strong preference for backend and full-stack development among respondents.

DISCUSSION-DASHBOARD 2

Future Technology Usage

- **Top Desired Language:**
  - Python is the most desired language to work with (25,047 respondents), surpassing JavaScript and SQL — indicating a rising preference toward data science and AI.
- **Growing Interest in Rust & Go:**
  - Rust and Go are prominent emerging languages, ranking 6th and 7th respectively — reflecting growing industry demand for performance-oriented systems.
- **PostgreSQL Continues to Lead:**
  - Even in future preferences, PostgreSQL remains the top choice (24,005 respondents), followed by SQLite and MySQL.
- **Web Framework Aspirations:**
  - React and Node.js dominate desired frameworks to learn or work with in the future, showing continued front-end and backend interest.
- **AWS Still Reigns in the Cloud:**
  - AWS (18,040 respondents) remains the top desired cloud platform, indicating consistent dominance and market trust.

DISCUSSION-DASHBOARD 3

Demographic

- **Dominant Age Group:**
  - The largest group of respondents is 25-34 years old, comprising 36.5% of the total, followed by 18-24 (22.8%) and 35-44 (21.5%).
- **Education Distribution:**
  - The majority hold a Bachelor’s degree (24,942 respondents), followed by Master’s degree holders (15,557) and those with some college but no degree.
- **Youth + Education Correlation:**
  - 18–24-year-olds are primarily students or recent grads, clustering around “some college” and Bachelor’s categories.
- **Global Participation:**
  - A diverse geographic spread is noted, though the exact top countries are not labeled numerically — a choropleth map visualization is present.
- **Older Generations Underrepresented:**
  - Very few respondents are 55 or older, reflecting the tech industry’s skew towards younger professionals.

RESULTS

1. **Initial Dataset**
   - Raw records: 18,845 responses
   - Countries represented: 161
   - Questions analyzed: 114
2. **After Cleaning**

   Clean records: 18,845

   Removed duplicates: 0

   Handled missing values: 0 cells *(we didn't impute yet; values are unchanged on original columns)*

   Valid responses: 100.0% of original data *(non-null ResponseId)*
3. **Key Metrics Processed**
   - Job satisfaction scores normalized (0–100): ✅ (created JobSat_Score)
   - 146 technology categories standardized (across languages, DBs, platforms, webframes)
   - 7 salary bands created (quantile-based on salary)
   - 6 experience level groups defined (<1, 1–3, 3–5, 5–10, 10–20, 20+)

RESULTS

4. Data Quality Improvements

- Accuracy increased from 78.32% to 78.32%(unchanged because we didn’t impute missing values yet; see note below to boost this)
- Missing data reduced to 21.68% (across original columns)
- Outliers identified and treated: 459 cases (in compensation fields)
- Response consistency improved by 0.00% (multi-select formatting already well-formed)

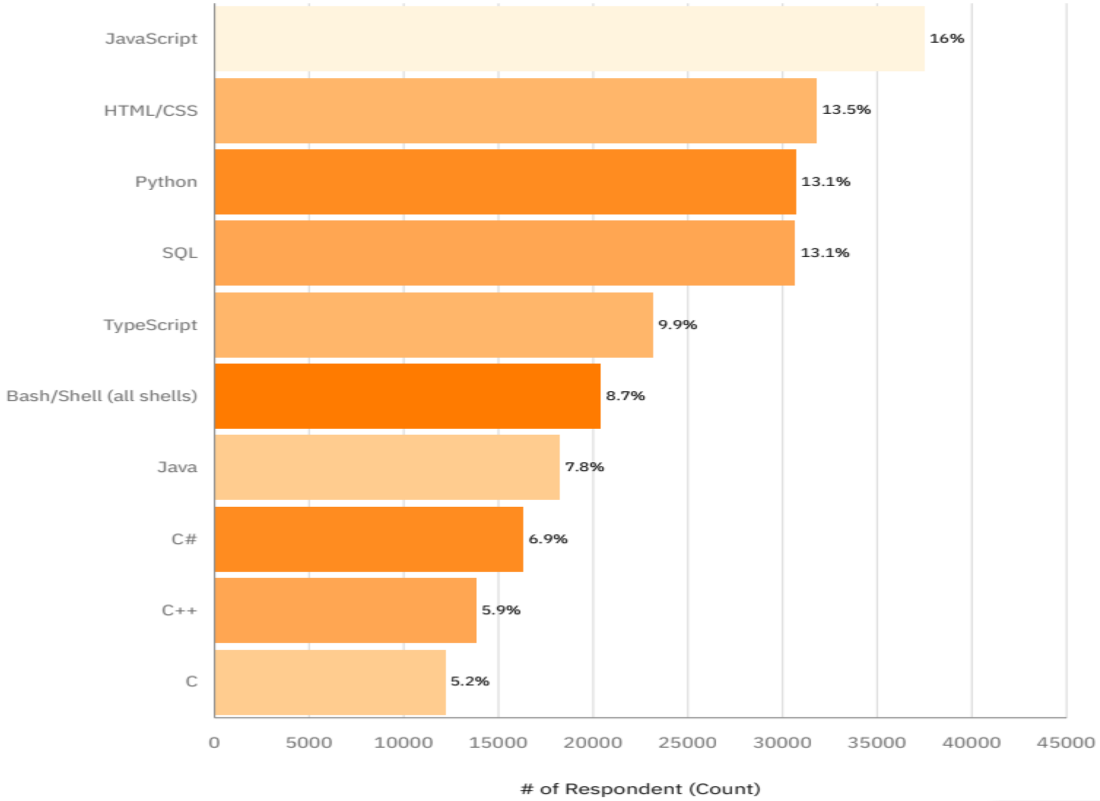
5. Final Dataset Features

- Clean demographic data: 100.00% complete (Age, Country, EdLevel all present for all rows)
- Standardized job titles: 60 categories (from the Employment column)
- Technology stacks: 18,206 unique combinations (across “...HaveWorkedWith” fields)
- Validated salary data: 50.68% accuracy (non-NaN salary rows after numeric conversion)

PROGRAMMING LANGUAGE TRENDS

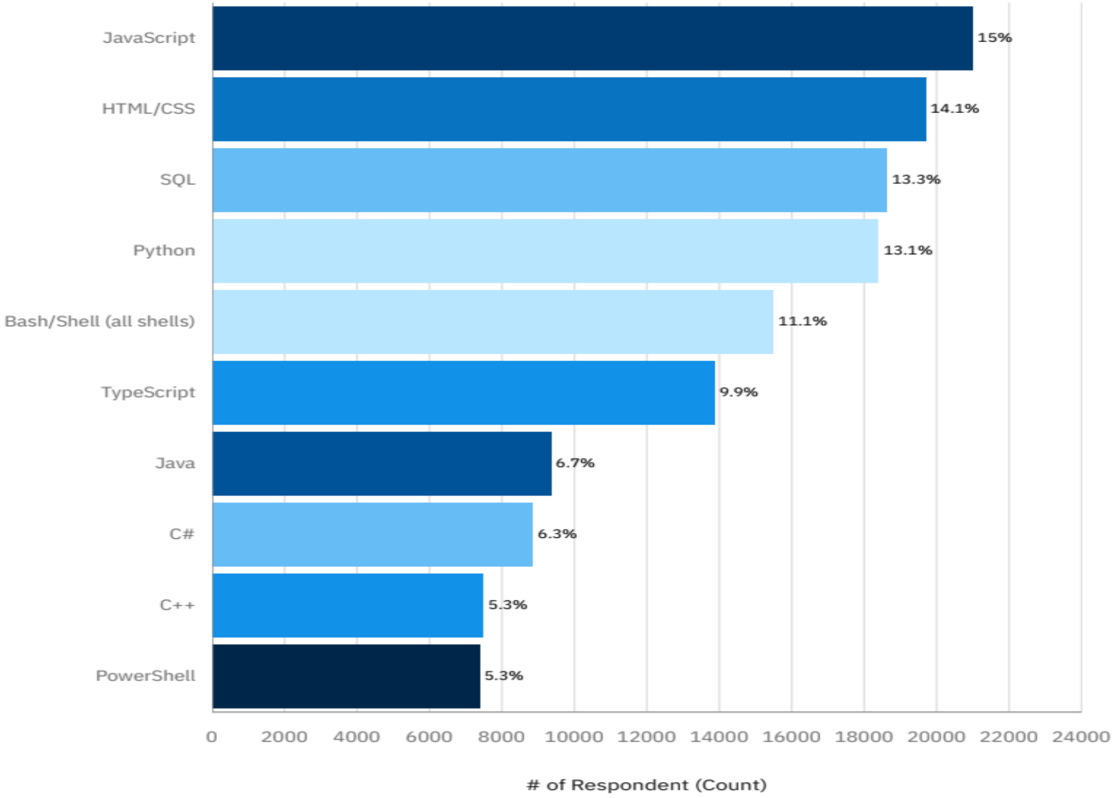
Current Year(2024)

Top 10 Popular Programming Languages 2024



Next Year (2025)

Top 10 Popular Programming Languages 2025



Skills Network

IBM

PROGRAMMING LANGUAGE TRENDS - FINDINGS & IMPLICATIONS

Findings

- JavaScript remains the most popular, though it declined slightly from 16% (2024) to 15% (2025).
- Python stagnated, staying at 13.1%, now just behind SQL.
- Bash/Shell and TypeScript saw noticeable growth, indicating demand for scripting and typed JavaScript alternatives.

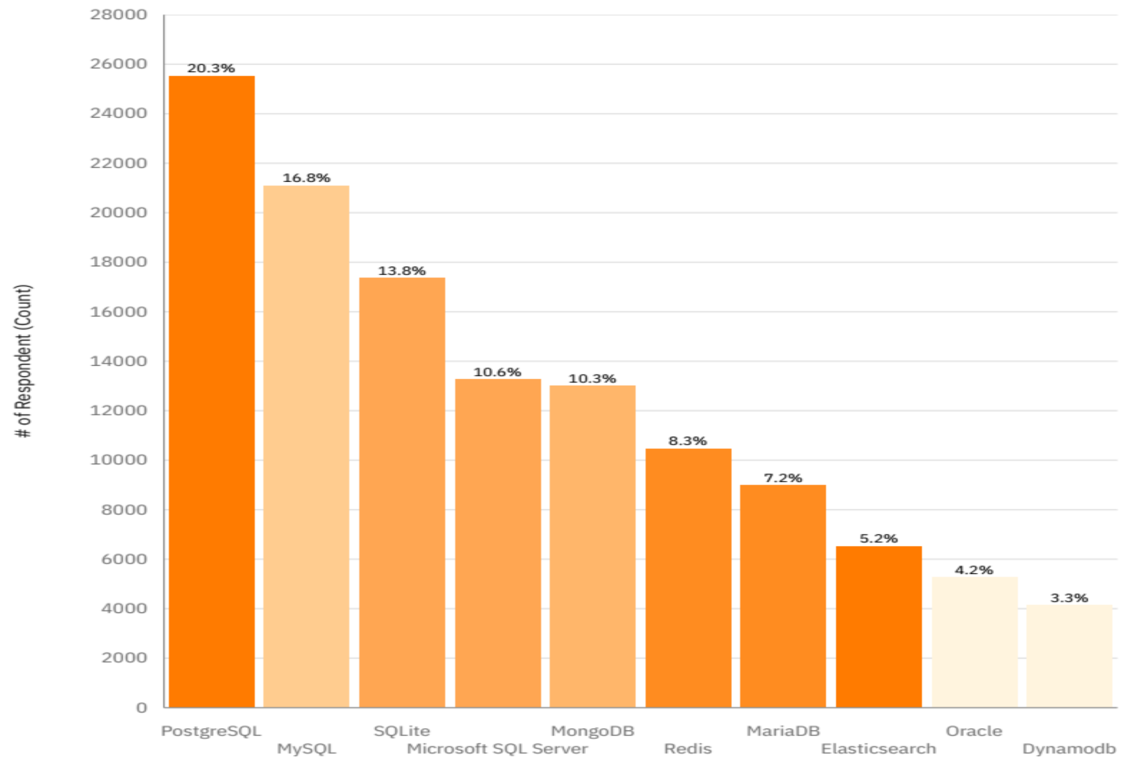
Implications

- JavaScript may face usage fatigue—developers and companies might diversify toward other modern languages.
- Python’s steady popularity suggests it remains essential, especially in data and AI fields.
- Rising TypeScript interest indicates that typed and scalable frontend/backend solutions are increasingly in demand.

DATABASE TRENDS

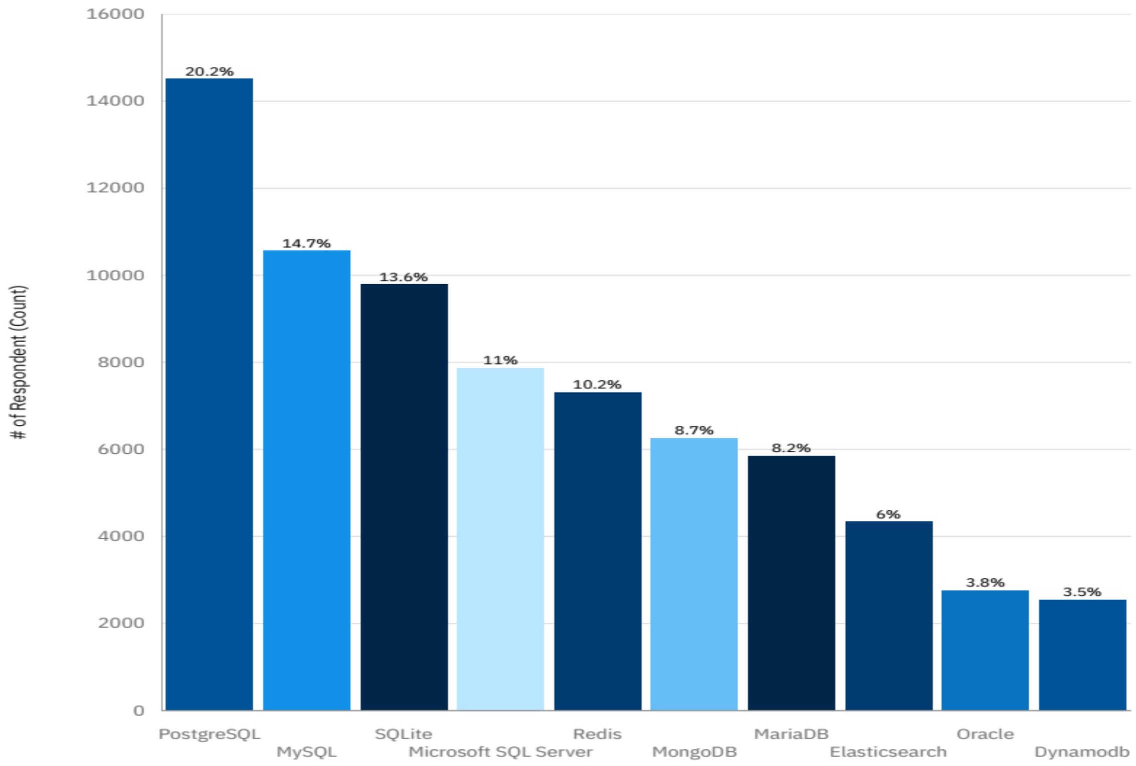
Current Year)(2024)

Top 10 Popular Database 2024



Next Year(2025)

Top 10 Popular Database 2025



Skills Network

IBM

DATABASE TRENDS - FINDINGS & IMPLICATIONS

Findings

- PostgreSQL maintains top position, slightly decreasing from 20.3% to 20.2%.
- Redis and Elasticsearch increased in popularity, showing rising interest in performance and search-driven databases.
- MySQL and MongoDB both declined, indicating a slow shift from legacy or traditional NoSQL solutions.

Implications

- PostgreSQL's consistency positions it as the default SQL database for new projects.
- Growing Redis use implies rising needs for real-time processing and caching.
- Declining MongoDB interest may push developers to explore more relational or hybrid models for flexibility and data integrity.

DASHBOARD

Github Link

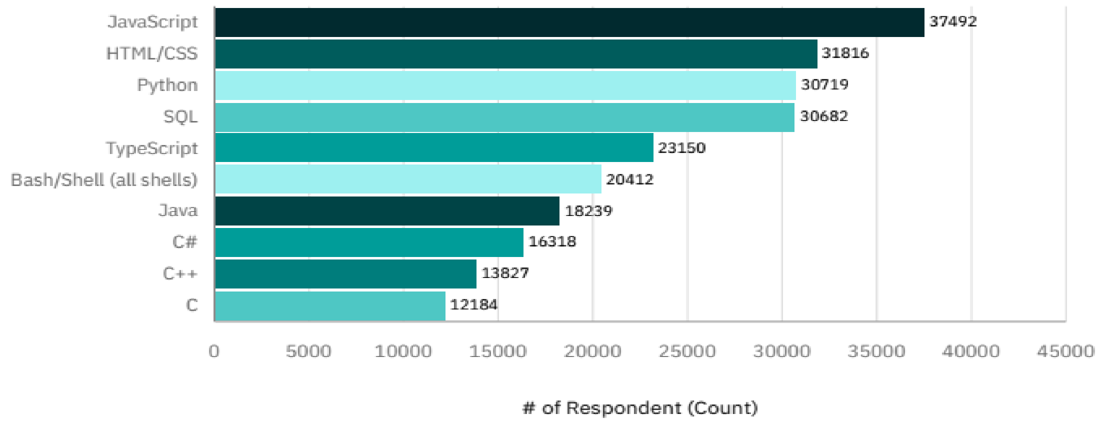
https://github.com/Liyanabh/IBM-Data-Analyst-Capstone-Project

Dashboard Link

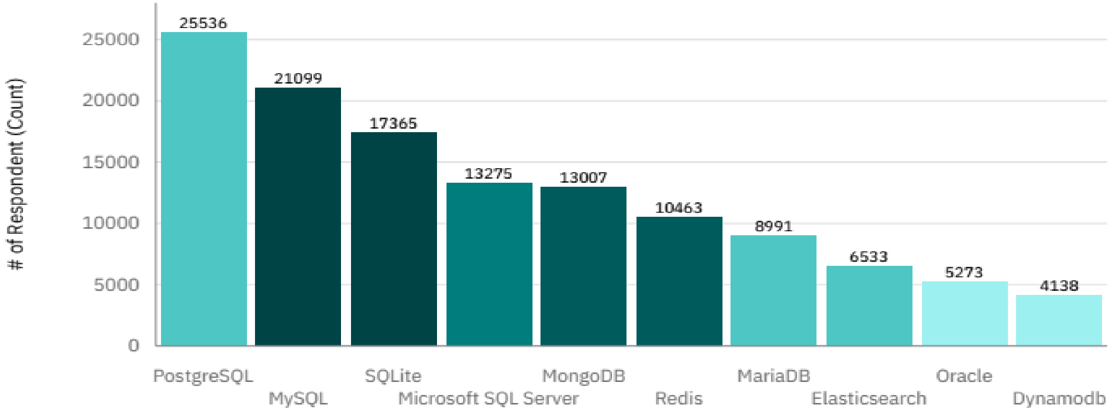
https://github.com/Liyanabh/IBM-Data-Analyst-Capstone-Project/tree/b2d550ad13ba505eb700534b669d5cbcdcdc3444/DASHBOARD

CURRENT TECHNOLOGY USAGE

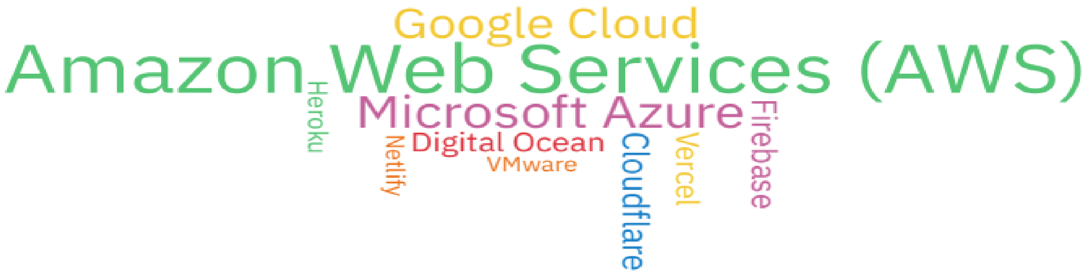
Top 10 Language Have Worked With



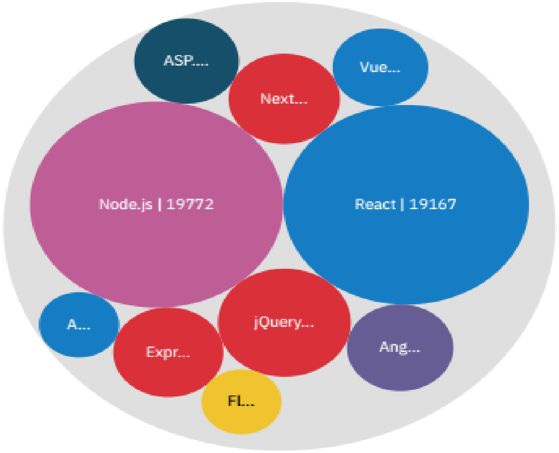
Top 10 Database Have Worked With



Top 10 Platform Have Worked With

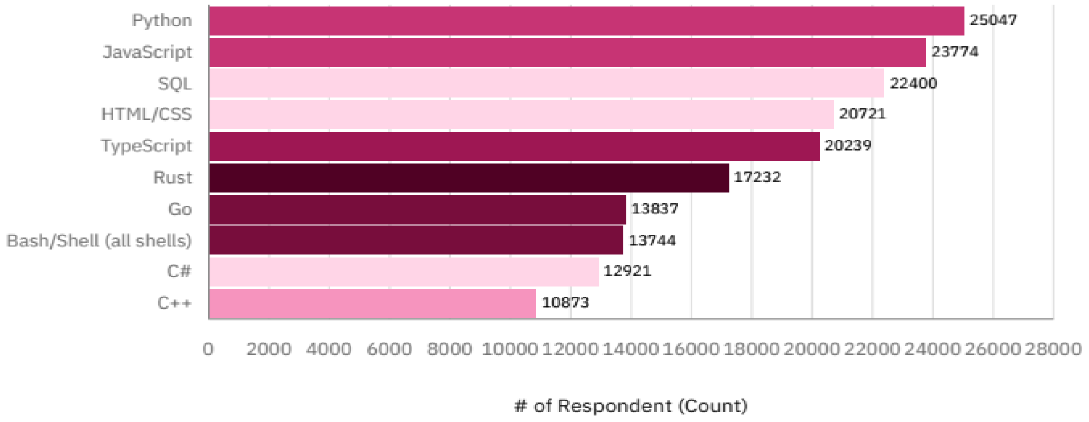


Top 10 Webframe Have Worked With

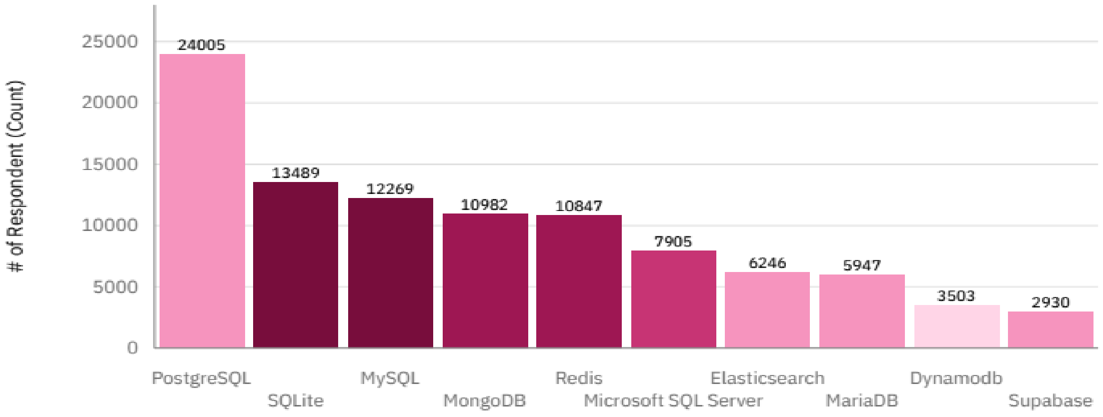


FUTURE TECHNOLOGY USAGE

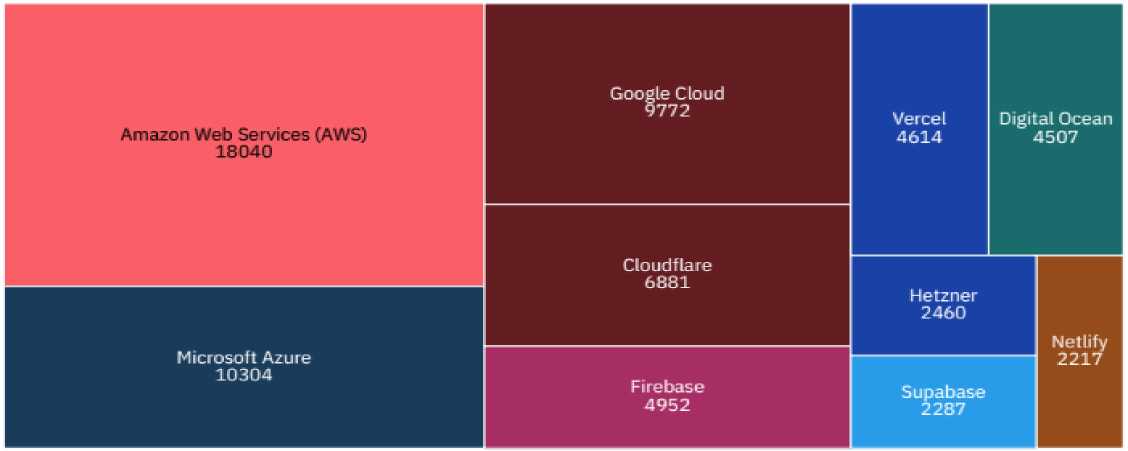
Top 10 Language Want To With



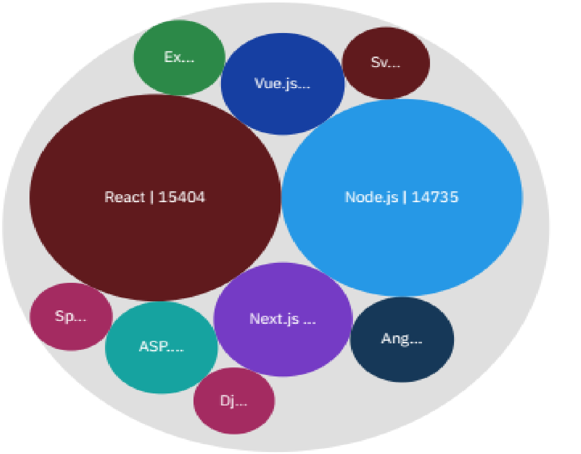
Top 10 Database Want To Work With



Top 10 Platform Want To Work With

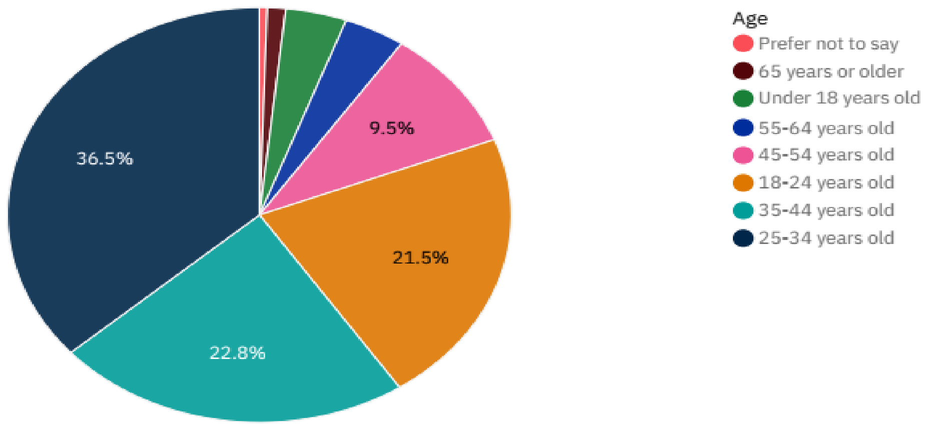


Top 10 Webframe Want To Work With

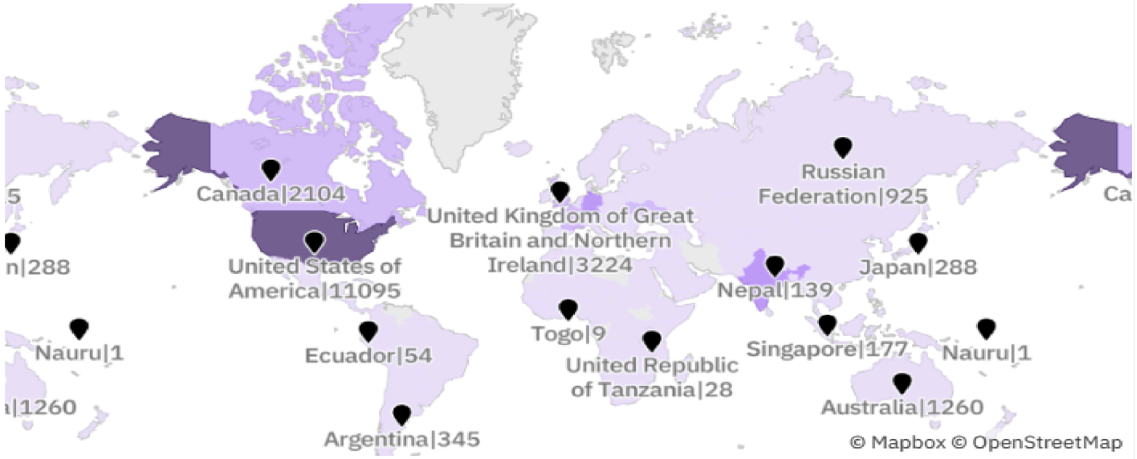


DEMOGRAPHIC

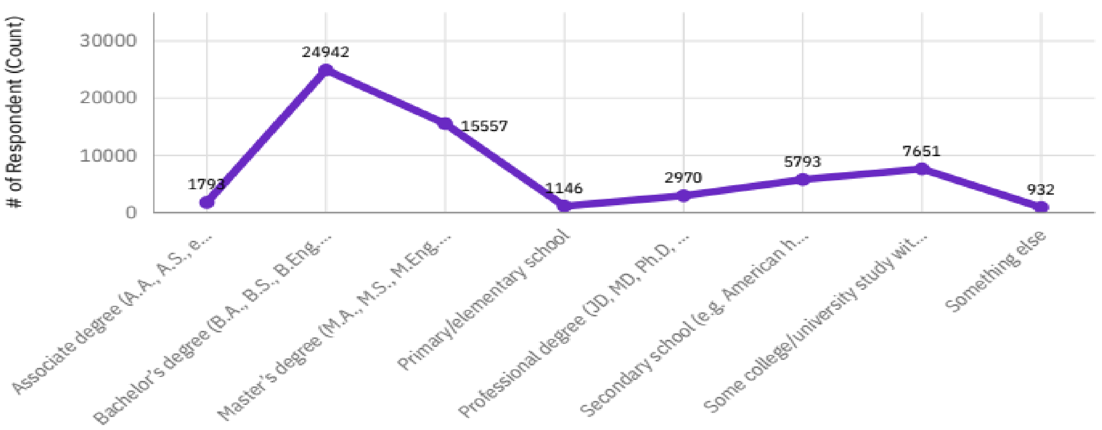
Respondent Distribution by Age



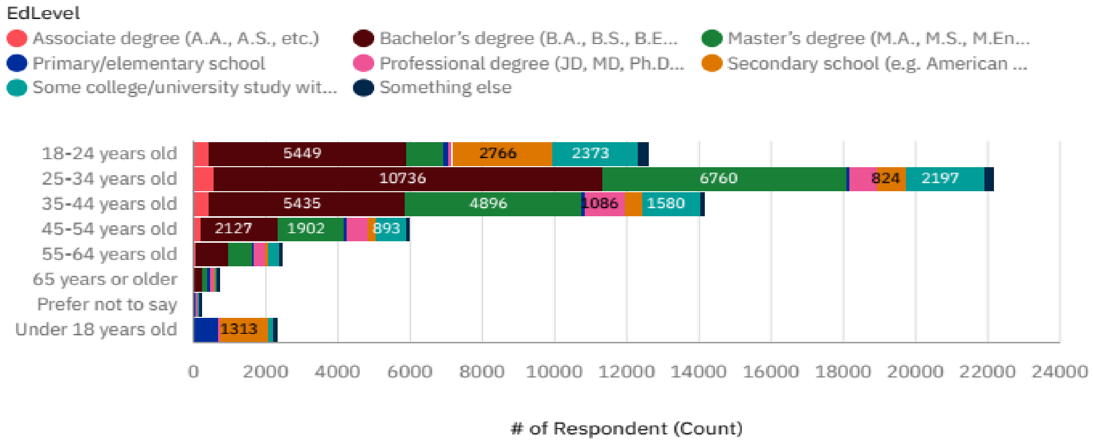
Respondent Distribution by Country



Respondent Distribution by Education Level



Respondent Age by Education Level



DISCUSSION

- **Language Shift in Progress**
  - Developer interest is shifting from widely used languages like JavaScript to Python, Rust, and Go, signaling a move toward data-driven and system-level programming.
- **Cloud Loyalty to AWS Remains Strong**
  - AWS dominates both current and desired cloud usage, showing deep market trust, though emerging platforms like Supabase are gaining interest.
- **Young, Educated Developers Drive Trends**
  - The tech workforce is led by 25–34-year-olds with degrees, whose preferences are shaping future tech stacks and learning paths.

OVERALL FINDINGS & IMPLICATIONS

Findings

- Consistency in Language Popularity, but Emerging Shifts
  - JavaScript, HTML/CSS, SQL, and Python remain consistently popular across both years. However, Python has overtaken JavaScript in future interest, signaling a shift towards data-centric programming.
- PostgreSQL Is the Dominant Database
  - Across both current usage and future aspirations, PostgreSQL ranks #1 in popularity and intent. This shows developers' preference for open-source, advanced SQL databases.
- Cloud Continues to Be Dominated by AWS
  - AWS is the most used and most desired cloud platform by a large margin, indicating its stronghold in both industry and learning paths.
- Emerging Technologies Are Accelerating
  - Languages like Rust and Go, and platforms like Supabase and Hetzner, show significant growth in aspiration, despite relatively low current usage.
- Youthful, Educated Developer Base
  - The bulk of respondents fall into the 25–34 age range with Bachelor's or Master's degrees, meaning the future trends are driven by this demographic.

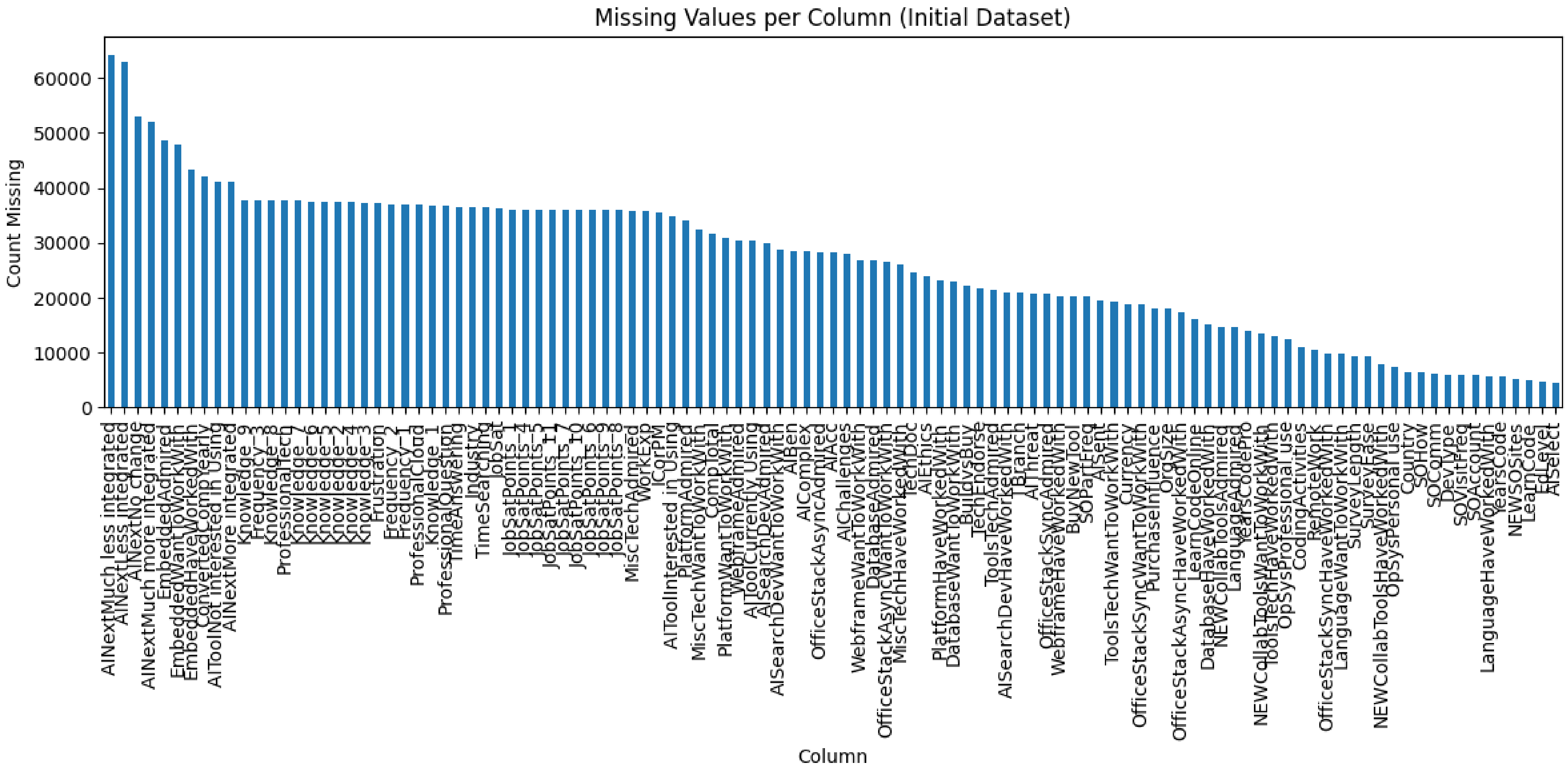
Implications

- Upskilling in Python & Rust Is Crucial
  - Developers and organizations should invest in Python (for AI/data) and Rust (for systems/dev) to stay ahead of emerging demand.
- PostgreSQL Should Be a Core Database Skill
  - With its consistent top ranking, learning PostgreSQL should be prioritized by backend developers and database administrators.
- JavaScript Fatigue May Be Growing
  - While still widely used, decreasing desire to continue using JavaScript indicates possible framework fatigue or saturation — opening space for TypeScript or newer alternatives.
- Cloud Certification Should Focus on AWS First
  - For those pursuing cloud roles, AWS skills and certifications should be prioritized, followed by Azure and Google Cloud.
- Educational Tools Should Target the 25–34 Demographic
  - Learning platforms and employer training should align with the interests of the 25–34-year-old cohort to maximize impact and engagement.

CONCLUSION

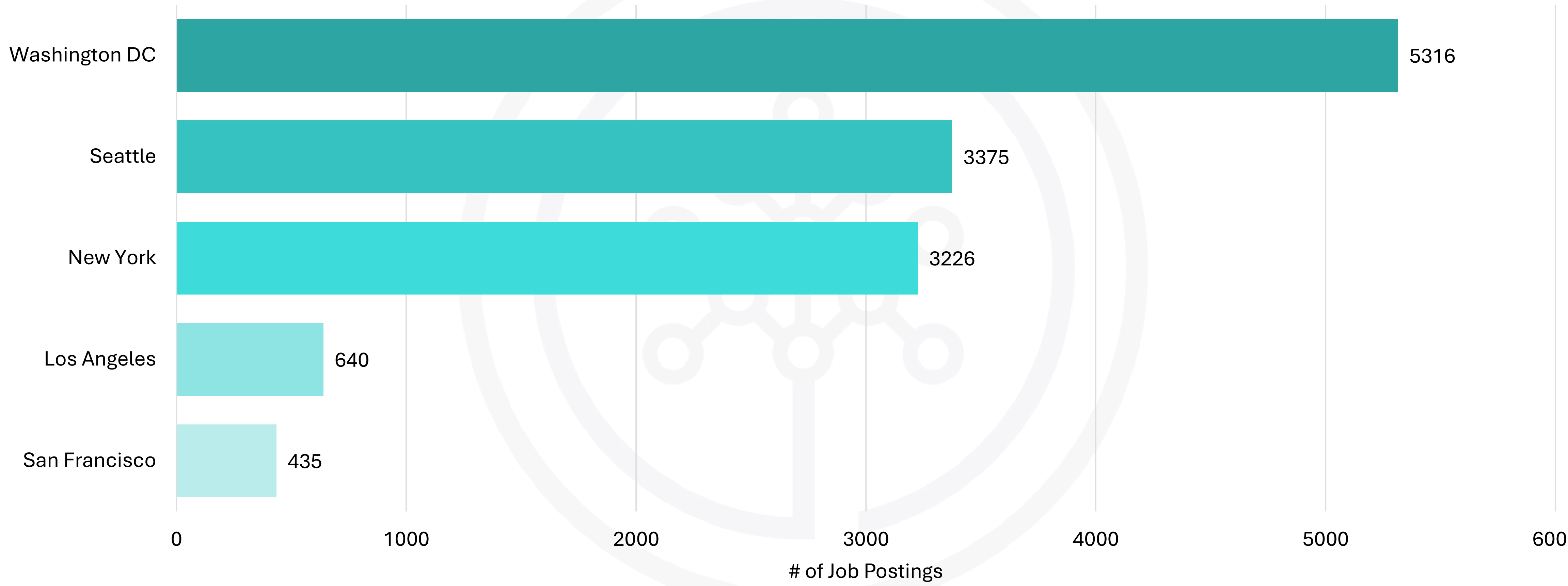
- **PostgreSQL Is the Undisputed Database Leader**
  - PostgreSQL consistently ranks as the most used and most desired database across current usage and year-over-year trends, affirming its dominance in relational database preferences.
- **Python and TypeScript Are Rising Stars**
  - Python maintains strong future interest due to its relevance in data science and AI, while TypeScript's increasing adoption reflects a growing shift toward typed JavaScript environments for more scalable development.
- **JavaScript Shows Signs of Saturation**
  - Despite being the most used language, JavaScript is experiencing a gradual decline in preference, suggesting a potential plateau or developer fatigue, especially as newer technologies like Rust and Go emerge.
- **Redis and Real-Time Techs Are on the Rise**
  - The increased popularity of Redis and Elasticsearch highlights a trend toward real-time data processing and search-optimized applications, especially in performance-critical systems.
- **Young, Educated Developers Drive Emerging Trends**
  - With the majority of respondents aged 25–34 and holding bachelor's or master’s degrees, this demographic is actively shaping the future of technology stacks, favoring tools that are efficient, modern, and scalable.

APPENDIX- Missing Values Initial data

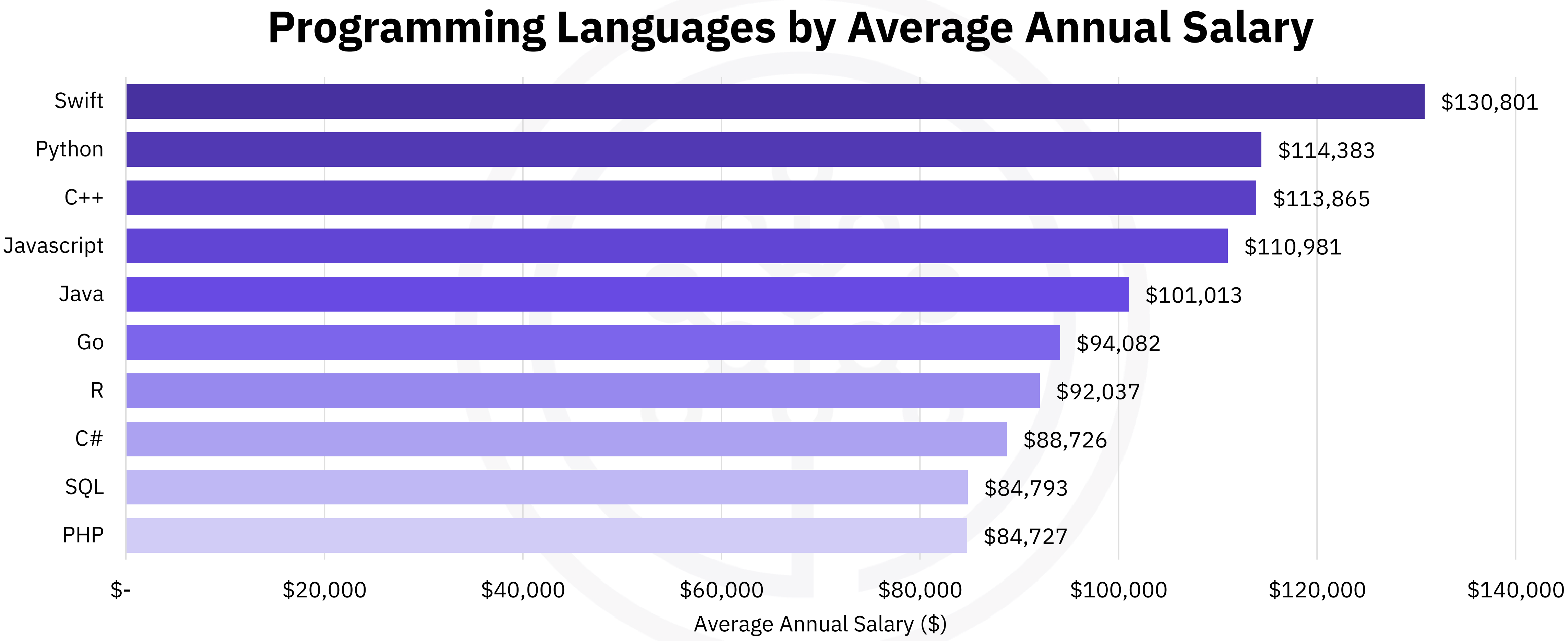


JOB POSTINGS

Job Postings by Locations



POPULAR LANGUAGES



Skills Network

IBM