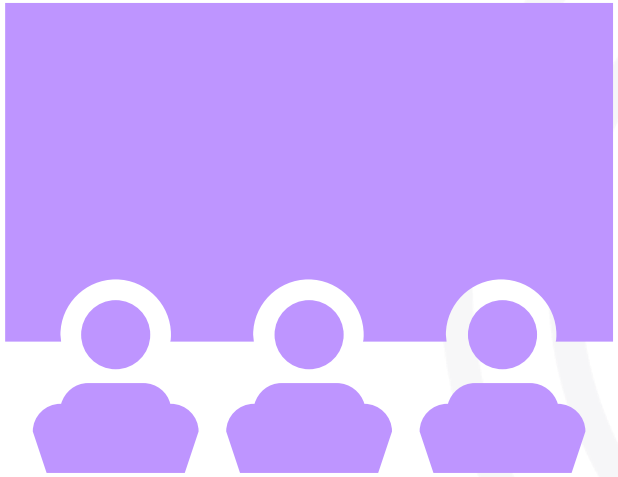
Stack Overflow Survey Data Analytics

Capstone Project

Name: Punyaja Mishra

Date: February 05, 2025

OUTLINE



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

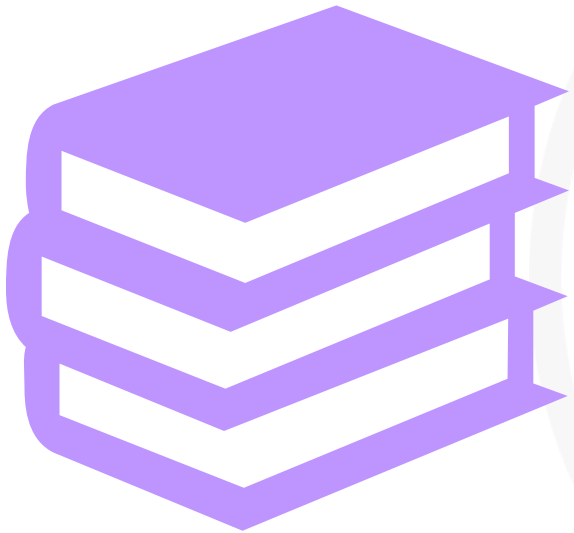
EXECUTIVE SUMMARY

- **Data**:
  - Stack Overflow Survey Worldwide Dataset
- **Objective**:
  - Gain insights into the developer community
  - Explore coding activities, tech trends, demographics and employment distribution
- **Challenges:**
  - Pre-processed and filtered for consistency, with tasks like handling missing values, duplicates, and normalization being crucial for accurate analysis
  - Several data issues encountered such as multiple values per field, missing data and duplicates
- **Stakeholders**:
  - Stack overflow Team & Users
  - Support decision-making and platform enhancements

EXECUTIVE SUMMARY

- **Tools and Methods:**
  - Multiple platforms, tools and programming languages like Jupyter Lab, python, SQL, IBM Cognos and Excel
  - Data Wrangling, Visualization and dashboard creation
- **Key findings:**
  - Identified significant trends in developer preferences, including the most popular programming languages, coding activities, and employment types
- **Outcome**:
  - Interactive dashboards and better insights into today's technological trends

INTRODUCTION



- **Problem**: Complex Stack Overflow Developer Survey Dataset with messy data
- **Objective**: Analyze developer demographics, coding activities, tech preferences, and employment trends.
- **Significance**: Insights will inform decision-making for business, educators and developers
- **Scope**: Responses from global developers covering various technologies and career aspects
- **Approach**: Data cleaning, exploratory analysis, and interactive visualizations for insights

METHODOLOGY

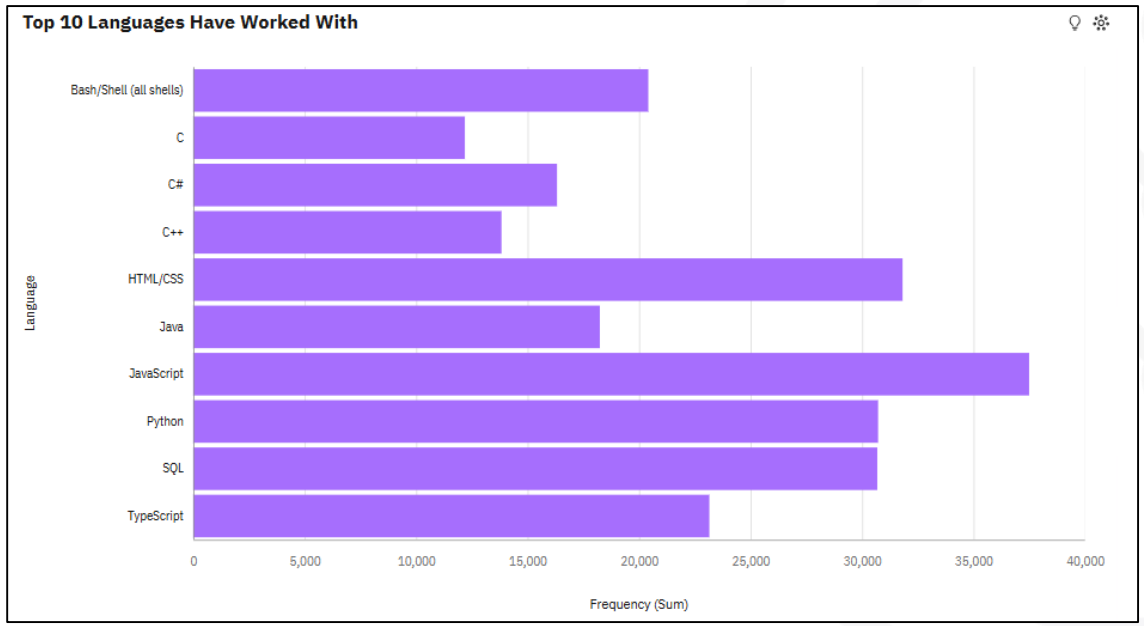
- **Data Collection**:
  - Utilized publicly accessible Stack Overflow API for data gathering
- **Sample Size:**
  - Worldwide data of 60,000+ responses with 100+ attributes
- **Data Cleaning**:
  - **Handling Missing Values**: Replaced missing values with mean (mean imputation method)
  - **Normalization**: Applied techniques to scale numeric data to ensure consistency
  - **Data Splitting**: Addressed multiple values per field by splitting values into separate records

METHODOLOGY

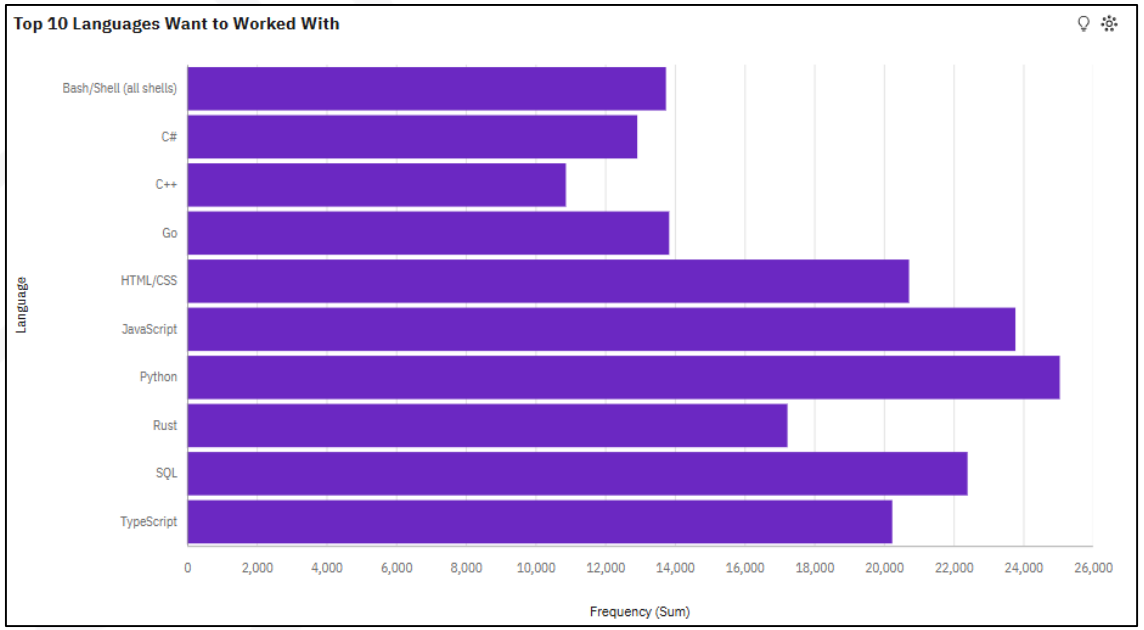
- Analysis Tools:
  - Jupyter Lab: Data exploring and analysis
  - Python: Data Wrangling, analysis and visualization
  - Excel: Used for data cleaning and preprocessing
  - IBM Cognos: For dashboard creation and reporting
- Visualization Techniques:
  - Multiple charts used such as bar charts, column charts, aps, tree maps, hierarchy packed bubble chart and word clouds

PROGRAMMING LANGUAGE TRENDS

Current Year



Next Year

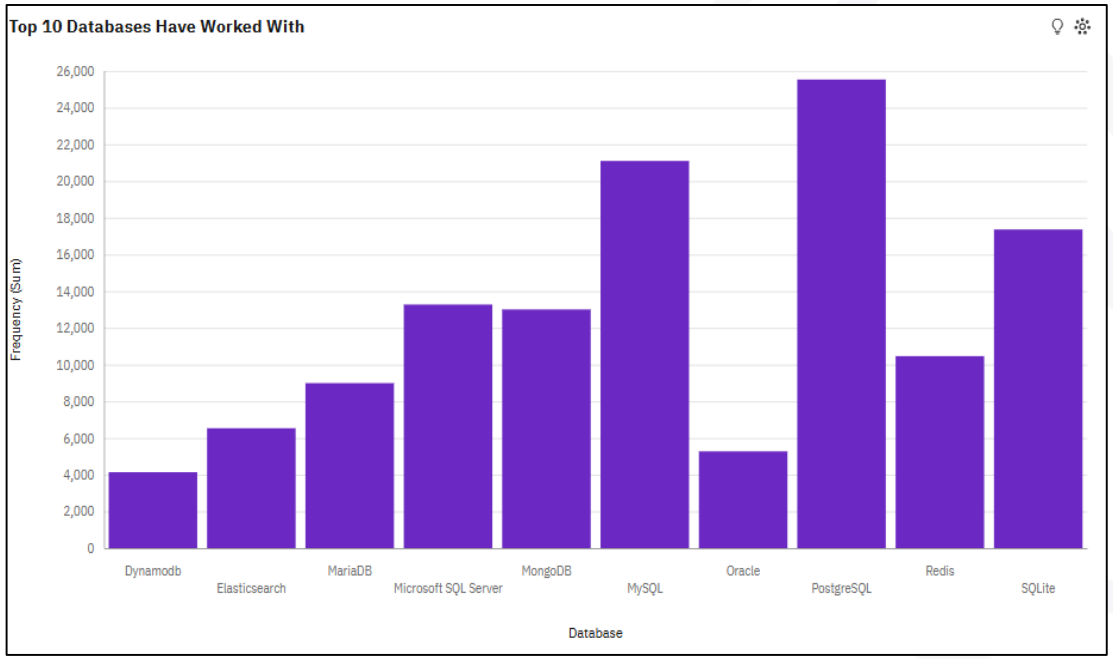


TOP 10 PROGRAMMING LANGUAGE TRENDS - FINDINGS & IMPLICATIONS

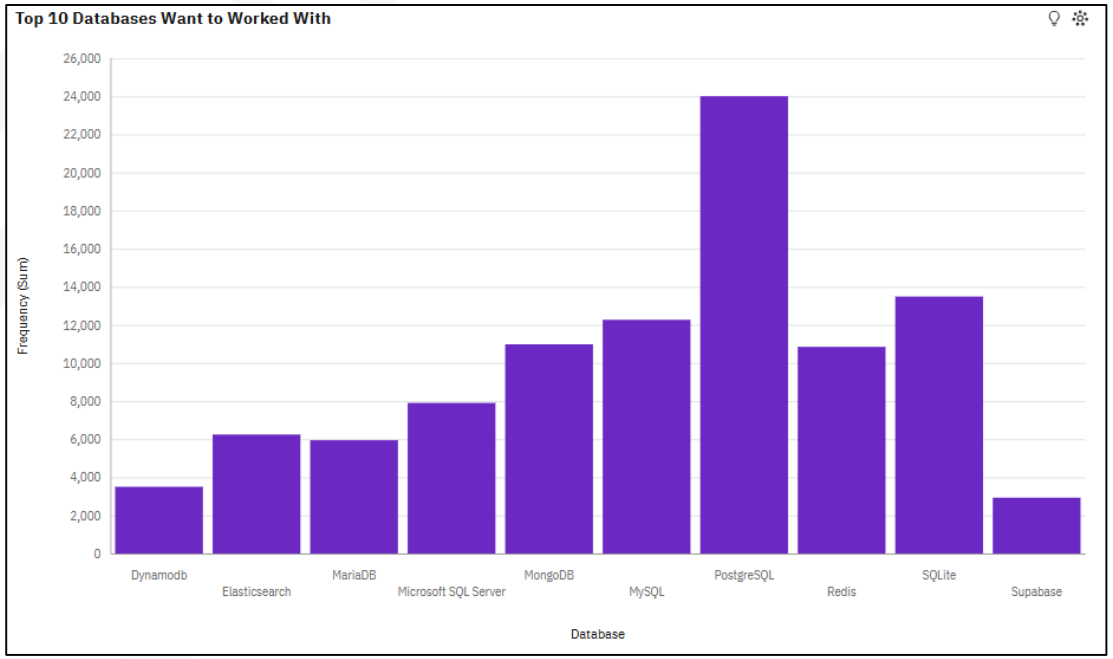
FINDINGS	IMPLICATIONS
Most popular programming language: **PYTHON**	High demand for **Python** indicates a focus on data science and AI in the developer community
Low-frequency languages: **C, Java**	Less popularity in these languages suggests limited developer interest in these languages
Languages gaining traction: **Go, Rust**	The rise of niche languages reflects emerging trends in specialized development fields

DATABASE TRENDS

Current Year



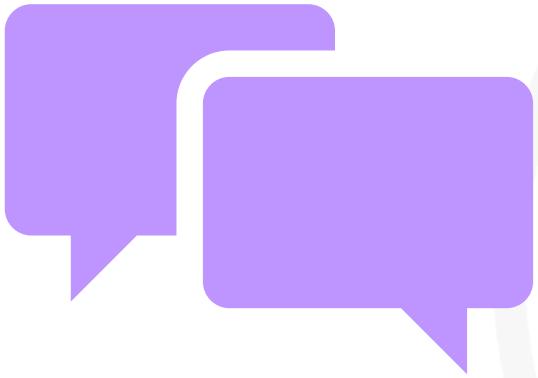
Next Year



TOP 10 DATABASE TRENDS - FINDINGS & IMPLICATIONS

FINDINGS	IMPLICATIONS
Most popular Database: **PostgreSQL**	High demand for **PostgreSQL** both in current and future trend signifies strong demand for open-source, relational databases that offer flexibility and scalability
Low-frequency languages: **Oracle**	Oracle's low adoption reflects a trend away from older, enterprise-driven db solutions in favor of more modern, cost-effective alternatives
Languages gaining traction: **MongoDB, Supabase**	The rise of MongoDB and Supabase highlights the increasing focus on NoSQL and cloud-native db to support web-applications and real time processing

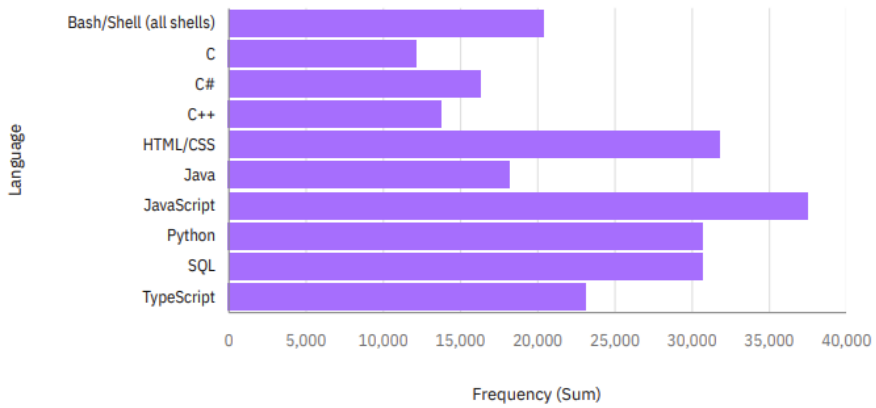
DASHBOARDS



CURRENT TECHNOLOGY USAGE TAB

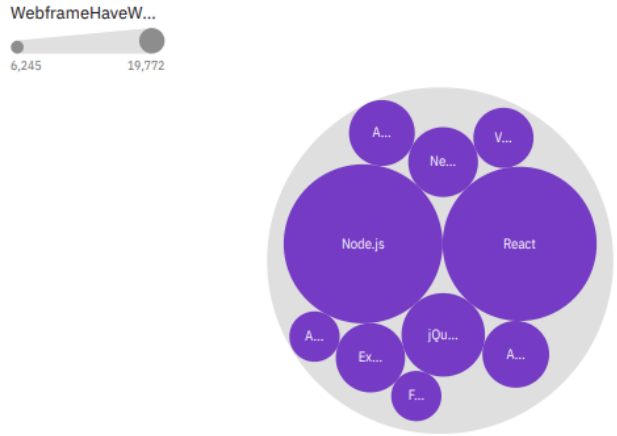
Current technology usage tab

Top 10 Languages Have Worked With

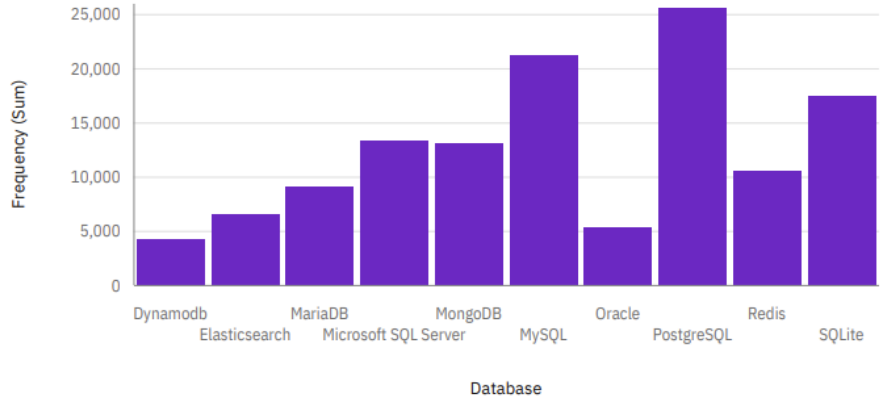


Platforms Have Worked With

Top 10 Web Frames Have Worked With



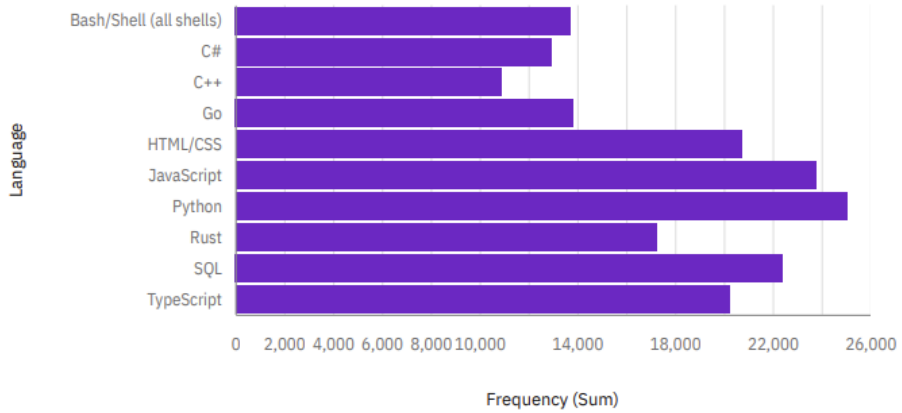
Top 10 Databases Have Worked With



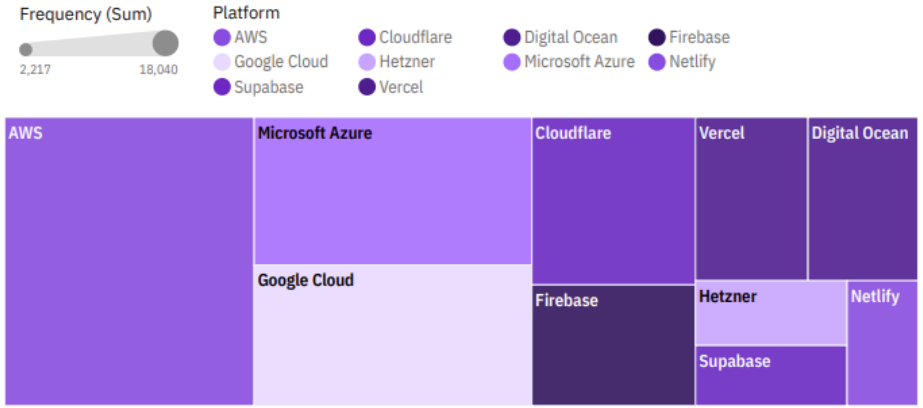
FUTURE TECHNOLOGY USAGE TAB

Future technology trends tab

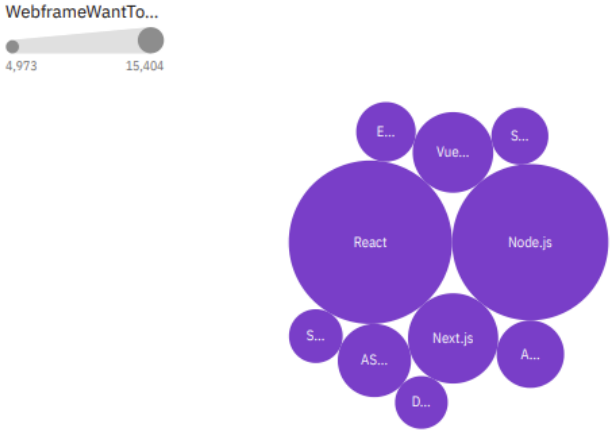
Top 10 Languages Want to Worked With



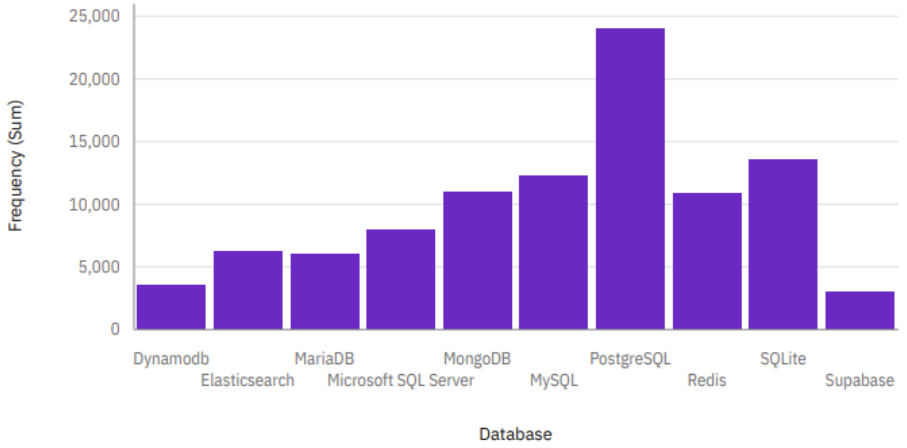
Platforms Want to Work With



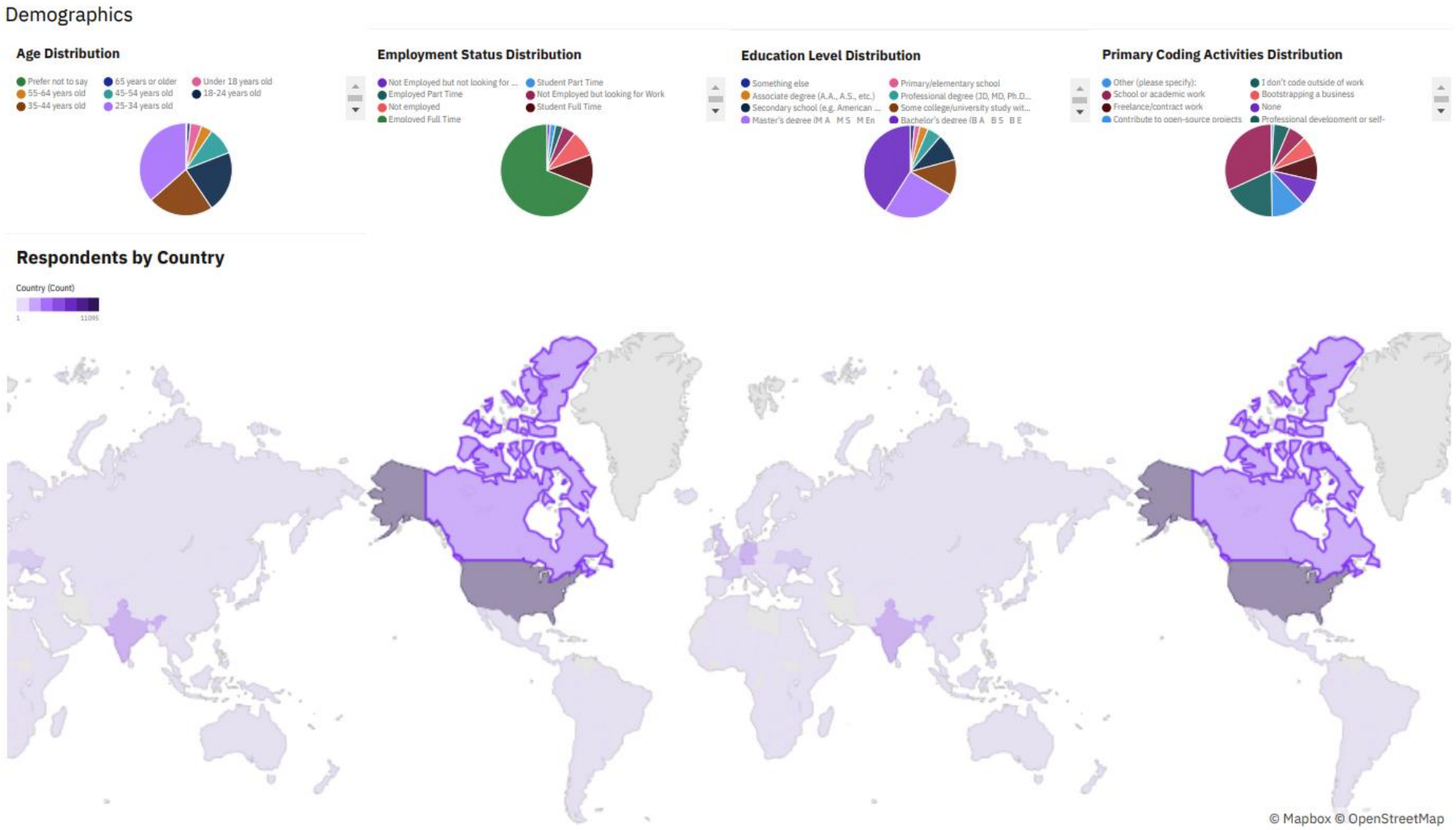
Top 10 Web Frames Want to Work With



Top 10 Databases Want to Worked With



DEMOGRAPHICS VISUALIZATION



DISCUSSIONS – KEY INSIGHTS FROM DASHBOARDS

- Cloud Platforms & Databases
  - AWS, Microsoft Azure, and Google Cloud are the most used platforms and also the most desired ones. Supabase is emerging as a preferred technology in both databases and platforms
- Web Framework Trends
  - React and Node.js dominate both current usage and future interest. Django, Svelte, and Spring Boot are gaining traction, while jQuery and Flask are declining in popularity
- Developer Demographics
  - Most survey respondents are from the US, Canada, India, Germany, and Ukraine, with lower representation from other regions like France and the UK.
- Age & Education
  - The majority of developers fall within the 25-34 age group, followed by 35-44, with most holding a bachelor's degree, then master's, some college, and secondary education.
- Employment & Coding Activities
  - Most respondents are full-time employed developers, followed by students and those not employed. Many engage in coding for hobbies, professional development, and open-source contributions.

OVERALL FINDINGS & IMPLICATIONS

AREA	FINDINGS	IMPLICATIONS
Platforms & Databases	AWS, Microsoft Azure and Google Cloud are the most used and most desired1	Cloud computing remains dominant, with developers focusing on multi-cloud expertise
Emerging Platforms	Supabase is gaining interest in both database and platform trends.	Rising demand for open-source and scalable cloud solutions.
Web Frameworks	React and Node.js are the most used and most wanted frameworks.	Full-stack JavaScript remains the industry standard for modern web development.
Future Web Frameworks	Django, Svelte, and Spring Boot are gaining traction, while jQuery and Flask see less interest.	A shift towards modern, scalable frameworks and reduced reliance on older technologies.

OVERALL FINDINGS & IMPLICATIONS

AREA	FINDINGS	IMPLICATIONS
Geography	Most responses came from the US, followed by Canada, India, Germany, and Ukraine	Developer communities are strongest in North America, with growing participation from India and Europe
Age Groups	Largest group is 25-34, followed by 35-44, then 18-24, and 45-54	Most developers are mid-career professionals, with a steady influx of young developers
Employment	Majority are full-time developers, followed by students and aspiring developers	High engagement from students suggests strong pipeline for future talent

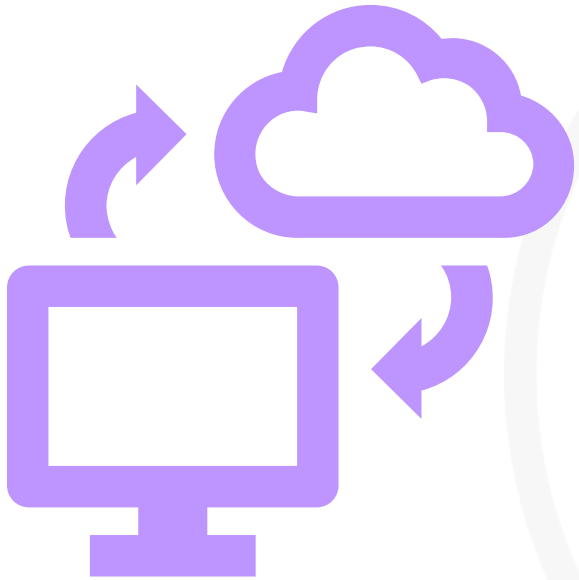
OVERALL FINDINGS & IMPLICATIONS

AREA	FINDINGS	IMPLICATIONS
Education	Most have a bachelor's degree, followed by master's, then some college and secondary education	Traditional education still dominates, but alternative learning paths are growing.
Coding Activities	Many code as a hobby, for professional development, and contribute to open source.	Developers are highly engaged beyond work, showing passion and continuous learning.

CONCLUSION

- Survey highlights strong trends in cloud computing with AWS, Azure, and Google Cloud leading the market
- JavaScript frameworks like React and Node.js dominate, while newer frameworks like Svelte and Django are gaining interest
- Developers are primarily mid-career professionals, with a steady influx of students and aspiring developers
- Open-source contributions and coding as a hobby show high engagement and passion within the developer community

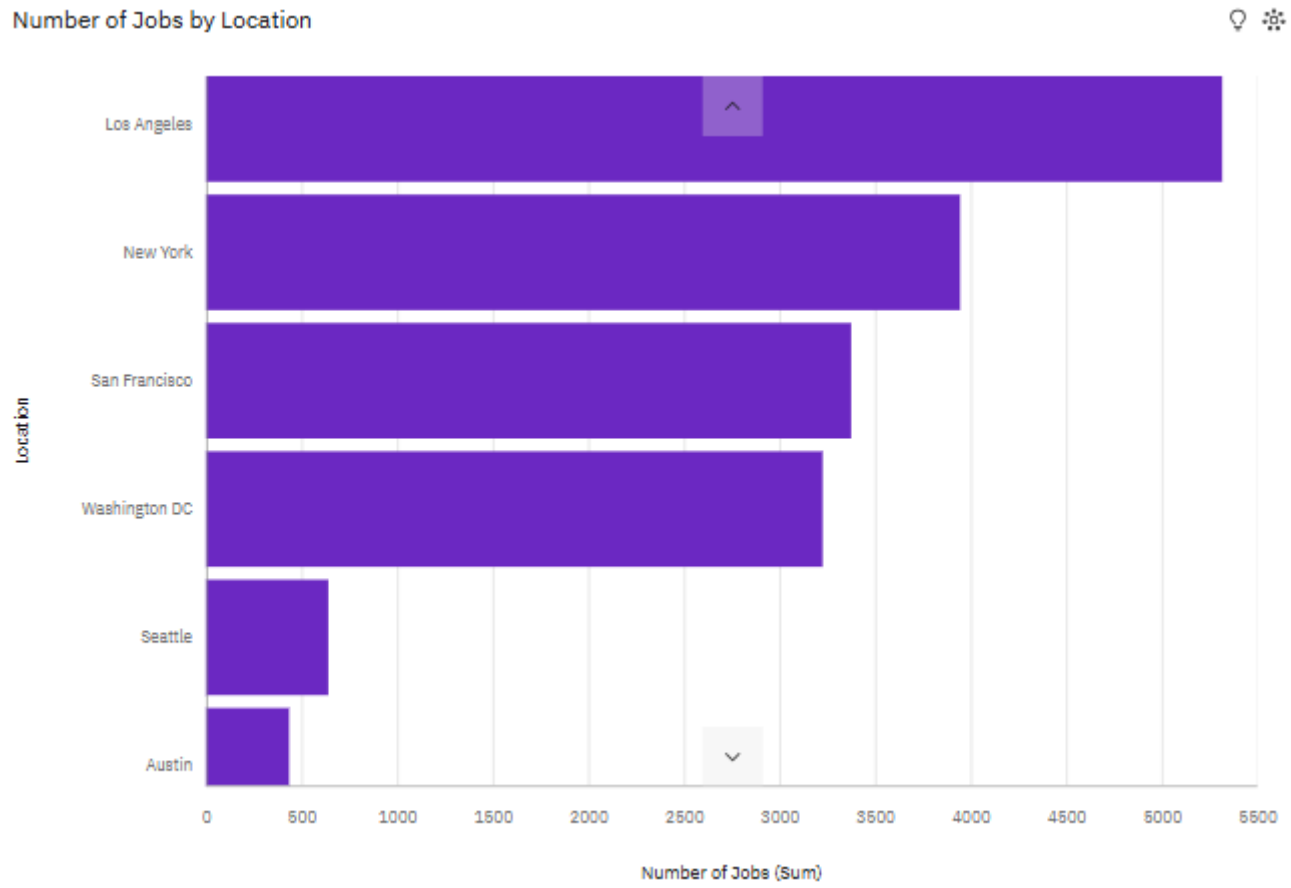
APPENDIX



- Data Source: Stack Overflow Developer Survey 2024
- IBM Cognos Free Trial and Jupyter Lab provided by IBM via Coursera
- Coursera educational platform facilitated access to this project
- Data had a lot of categorical data that had to be converted into a numerical data to be able to visualize in Python
- Columns like Employment, Platforms Worked With, Coding Activities, and Platforms Want to Work With had lot of values separated by a semi-colon

JOB POSTINGS

Job posting data gathered using the Job API



POPULAR LANGUAGES

Job postings data using web scraping presented in the descending order of salary.