***2023 Annual Stack Overflow Survey***

**A PROJECT REPORT**

**Submitted by**

**PRATHYUSH MENON 1RVU23CSE352**

**SAANVI VENKATESH KULKARNI 1RVU23CSE391**

**SAMARTH SANJAY PYATI 1RVU23VSE400**

**SHASHWATH JAIN HP 1RVU23CSE426**

*in partial fulfillment for the award of the degree of*

**B.Tech (Hons) – Computer Science & Engineering**

*in* **Data Analysis with Python**

**to Dr. Sahana Prasad**



**School of Computer Science and Engineering**

**RV University**

**RV Vidyaniketan, 8th Mile, Mysuru Road, Bengaluru, Karnataka, India – 562112.**

**December 2023**

# DAP PROJECT:-

# Analysing 2023 Annual Stack Overflow Survey Results

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
|  | **TITLE** |  |
| **1.0** | **INTRODUCTION** | 4 |
|  |  |  |
| **2.0** | **METHODOLOGY** | 5 |
|  | 2.1 Sub section  2.2 Sub section  2.3 Sub section | 6  7  8 |
|  |  |  |
|  |  |  |
| **3.0** | **RESULT AND DISCUSSION** | 9 |
|  | 4.1 Sub section | 10 |
|  |  |  |
| **4.0** | **CONCLUSION** | 11 |
|  |  |  |

1. **INTRODUCTION**

Embarking on an exploration of the 2023 Stack Overflow Developer Survey dataset, this analysis offers a comprehensive journey through the intricate tapestry of the contemporary software development landscape. As we navigate the expansive realm of global developer demographics, we uncover the geographical distribution of survey participants, throwing light on the key contributors shaping this vibrant community. The educational front is scrutinized, revealing insights into the diverse academic backgrounds that developers bring to the table, influencing their participation in the survey.

Moving forward, our investigation dives into the age dynamics of developers, unveiling patterns that highlight the significance of specific age groups within the technology sector. An examination of programming language preferences follows, unveiling the languages that command prominence in the developer ecosystem in 2023.

Shifting the focus to compensation, our analysis dissects median salary distributions across various job roles, intertwining the narratives of work experience and earnings. Venturing into the dynamic arena of artificial intelligence, we explore the adoption trends of AI search and development tools, offering a glimpse into the technological preferences that define the forefront of software engineering.

In the backdrop of the 2023 Annual Stack Overflow survey, this multifaceted analysis aims to extract actionable insights and trends resonating with developers, employers, and technology enthusiasts alike, fostering a deeper understanding of the prevailing dynamics within the global developer community

|  |  |  |
| --- | --- | --- |
|  |  | **.** |
|  |  |  |
|  |  |  |
|  |  |  |

1. **METHODOLOGY**

**This is an explanation of the code used in the project along with its function**

CODE:-

import pandas as pd

from pandas import read\_csv

import seaborn as sns

import matplotlib.pyplot as plt

import operator

import re

import plotly.express as px

Importing necessary libraries such as Pandas is done for data manipulation, Seaborn and Matplotlib for visualization and statistical graphs, Operator for certain operations, Regular Expressions (re) for pattern matching, and Plotly Express for interactive plots.

CODE:-

pd.set\_option('display.max\_columns', 90)

Since there were a lot of columns, this code was used to display only the first 90.

CODE:-

# DATA FILE SETTING UP AND CONFIGURATION

schema = read\_csv('survey\_results\_schema.csv')

data = read\_csv('survey\_results\_public.csv')

data.set\_index('ResponseId', inplace=True)

data.rename(columns={'CompTotal': 'Salary', 'ConvertedCompYearly': 'Total\_Salary'}, inplace=True)

This code was written to read the survey schema and public survey data CSV files. Was also written to set the index of the data DataFrame to the 'ResponseId' column and rename certain columns for consistency.

Data Information: Display the shape and information about the DataFrame, including non-null counts and data types.

CODE:-

# Custom Aggregate functions:

# Function to populate the dictionary

import pandas as pd

import operator

def populateDict(data: pd.core.series.Series, delimiter=';', dropna=True, sortbyVals=False):

\_dict = {}

if dropna:

data = data.dropna()

for i in data:

splitted = i.split(delimiter)

for j in splitted:

\_dict[j] = \_dict.get(j, 0) + 1

if sortbyVals:

\_dict = dict(sorted(\_dict.items(), key=operator.itemgetter(1), reverse=True))

return \_dict

Custom Function: Define a custom function populateDict to populate a dictionary based on the occurrences of values in a Pandas Series.

CODE:-

# Distributions of survey participants according to their country

# took only the first 20 values for better visualization

consideration = 30

countryDistributions = dict(data['Country'].head(consideration).value\_counts(normalize=True))

countries, counts = map(list, zip(\*countryDistributions.items()))

Survey Participants Distribution by Country: Calculate and plot the distribution of survey participants according to their country using a pie chart.

CODE:-

# Wedge out the first country

myExplode = [0.05 if i == 0 else 0 for i in range(len(countries))]

fig, ax = plt.subplots()

ax.pie(counts, labels=countries, autopct='%1.1f%%', startangle=90, colors=sns.color\_palette('viridis'), explode=myExplode)

# Equal aspect ratio ensures that the pie chart is drawn as a circle.

ax.axis('equal')

plt.show()

Pie Chart Plotting: Create and display a pie chart showing the distribution of survey participants by country.

From above observation, we can see that most of the participants who took the survey were from the United States, followed by Germany, the UK, and Spain.

We can also observe that countries from the Philippines to Italy have participated in approximately the same proportions (3.4%).

Conclusion: Provide a conclusion based on the observation of the pie chart.

CODE:-

# Education level of candidates who participated in the survey

data['EdLevel'].fillna("NA", inplace=True)

educationData = data['EdLevel']

# educationCounts = {}

# for i in data['EdLevel']:

# educationCounts[i] = educationCounts.get(i, 0) + 1

# # Sorting Dictionary

# educationCounts = sorted(educationCounts.items(), key=operator.itemgetter(1), reverse=True)

educationCounts = populateDict(educationData, ';', sortbyVals=True)

levels, counts = map(list, zip(\*educationCounts.items()))

# Custom labels for better readability

custom\_labels = ['Bachelor`s', 'Master`s', 'No degree\n(but studied)', 'Secondary School', 'Professional Degree',

'Associate Degree', 'Primary School', 'Something Else', 'NA']

# Create a Seaborn bar plot

plt.figure(figsize=(12, 7))

sns.barplot(x=custom\_labels, y=counts, hue=custom\_labels, palette="viridis", dodge=False, legend=True)

# Adding data labels on top of the bars

for i, count in enumerate(counts):

plt.text(i, count + 0.1, str(count), ha='center', va='bottom')

plt.xlabel("Education Level", fontsize=15)

plt.ylabel("Number of Candidates", fontsize=15)

plt.title("Distribution of Education Levels", fontsize=20)

plt.xticks(rotation=45, ha="right")

plt.tight\_layout()

plt.grid(axis='y', linestyle='--', alpha=0.5)

plt.show()

Education Level Distribution Plotting: Plot the distribution of education levels among survey participants using a bar plot. Candidates with a Bachelor's degree were the ones who participated the most in the survey, along with Master's degree holders.

They are more dominant in the survey and seem to be using the platform to a larger extent.

There were other basic codes used to plot graphs.

Here are the imported library functions in the code:

1. `import pandas as pd`: Imports the Pandas library and assigns it the alias 'pd'. It is used for data manipulation and analysis.

2. `from pandas import read\_csv`: Specifically imports the `read\_csv` function from Pandas to read CSV files.

3. `import seaborn as sns`: Imports the Seaborn library, commonly used for statistical data visualization.

4. `import matplotlib.pyplot as plt`: Imports the `pyplot` module from the Matplotlib library for creating static, animated, and interactive visualizations.

5. `import operator`: Imports the `operator` module, providing a set of efficient functions corresponding to the intrinsic operators of Python.

6. `import re`: Imports the regular expression (regex) module for pattern matching operations.

7. `import plotly.express as px`: Imports the Plotly Express module, a high-level interface for creating a variety of interactive plots.

These libraries and functions are utilized throughout the code for tasks such as data manipulation, visualization, sorting, and pattern matching.

**3.RESULT AND DISCUSSION**

The 2023 Stack Overflow Annual Survey provides valuable insights into the global developer community. A comprehensive analysis of 89,184 survey responses reveals significant patterns. In terms of participant distribution, the majority hail from the United States, followed by Germany, the UK, and Spain. The educational background of participants indicates that individuals with Bachelor's and Master's degrees are predominant, with a notable inclination towards active participation from these cohorts.

Examining age distributions shows that the survey attracts a significant number of participants aged 25-34, followed by those in the 18-24 and 35-44 age groups. The analysis of programming languages underscores the popularity of web development languages like HTML, CSS, and JavaScript, while multipurpose languages such as Python gain prominence in data science and machine learning.

In the context of India, the salary distribution exhibits a median of approximately 1.5 Lakhs per annum, showcasing the specific financial landscape within the country. Notably, there are no outliers, suggesting a relatively uniform distribution. The salary spectrum ranges from a minimum of 0, indicating unpaid interns, to a maximum of 3 Lakhs per annum. Regarding educational qualifications, participants with Bachelor's degrees are the most actively engaged in the survey, closely followed by those with Master's degrees. This prevalence suggests a significant participation trend among individuals with higher education qualifications, indicating that the platform is extensively utilized by those seeking or possessing advanced degrees. The insights derived contribute to a nuanced understanding of salary patterns and educational demographics specific to the Indian context.

We can see that most percentage of people who participated in the survey were from United States, followed by Germany, UK and Spain. We can also observe that countries like Philippines and Italy have participated in approximately same proportions (3.4%). The analysis of survey participation indicates that individuals in the age groups of 25-34 years, 18-24 years, and 35-44 years are the most prominent participants. This suggests that individuals in their mid-20s to mid-40s are present in the jobsOn the other hand, there is a noticeable decline in participation among older individuals and those below 18 years old. The lower engagement from these age groups may imply that individuals in the older age brackets might be less present , possibly due to factors such as retirement or reduced involvement in the professional tech community. Similarly, those below 18 might face restrictions or have less exposure to the professional aspects of the jobs.

In the salary analysis, the United States generally offers higher compensation for various designations compared to Germany. Marketing and Sales Managers in Germany receive notably higher pay than other tech roles. System Administrators consistently earn the least in both countries (excluding students), while positions related to Blockchain, Cloud, Data Engineering, and data analysis or engineering demonstrate higher earning potential in both regions. These trends highlight the influence of location and job specialization on salary disparities in the tech industry.

The most popular AI Search tool are ChatGTP with it having around 50,000 plus users followed by Bing AI and Wolframalpha while the most popular AI developer tools are Github copilot dominating the competition with more than 50% of the market share followed by Tabnine.

We can observe that most salaries are clustered around 100,000 to 200,000 with two prominent peaks at approximately 100,000 and 175,000. There are fewer people earning salaries between 200,000 and 250,000 and another increase in frequency at around 300,000. The curve indicates that the salary distribution has heavy tails and sharp peaks which signifies high kurtosis - meaning there are more frequent large deviations from the mean than what is expected in a normal distribution 12.

We can infer that the majority of the population earns a salary between 100,000 and 200,000. The two peaks at 100,000 and 175,000 suggest that there are two groups of people with different salaries. The group with a salary of 100,000 is more significant than the group with a salary of 175,000. The drop in frequency for salaries between 200,000 and 250,000 indicates that there are fewer people earning salaries in this range. The increase in frequency at around 300,000 suggests that there is another group of people with a higher salary. The heavy tails of the curve indicate that there are more frequent large deviations from the mean than what is expected in a normal distribution 12. This could be due to a variety of reasons such as a few people earning very high salaries or a few people earning very low salaries.

The majority of people have less than 10 years of work experience. The frequency of work experience increases gradually from 0 to around 10 years and then decreases gradually. The curve line overlaid on the bars indicates that the distribution has a medium kurtosis 1. This means that the distribution has a moderate number of outliers compared to a normal distribution.

We can infer that most people have less than 10 years of work experience. The frequency of work experience decreases as the number of years of experience increases. This could be due to a variety of reasons such as people changing jobs or retiring. The medium kurtosis of the distribution indicates that there are a moderate number of outliers compared to a normal distribution 1. This could be due to a variety of reasons such as a few people having very high or low work experience.

Further insights into language proficiency, particularly with TypeScript, reveal varying adoption rates across different countries, with the United States, Germany, and Canada leading in expertise. Salary analysis, focusing on the top three paying countries, emphasizes the substantial earning potential in the United States across various job designations.

Exploring the relationship between work experience and salary highlights nuanced trends. The utilization of AI tools in search and development indicates a preference for tools like ChatGPT and GitHub Copilot. Overall, the survey provides a rich source of information for understanding the dynamics of the global developer landscape in 2023.

**CRITICAL ANALYSIS OF THE STUDY:-**

The study utilizes a large dataset of 89,184 survey responses, providing a comprehensive view of the global developer landscape.

The analysis covers various aspects, including participant demographics, education levels, age distributions, programming language popularity, salary trends, and AI tool usage. This diversity enhances the overall understanding of the developer community.

The use of visualizations such as pie charts, bar plots, and graphs aids in conveying complex information in an accessible manner.

**Scope of improvement**

The 2023 Stack Overflow Annual Survey offers valuable insights into the developer community, yet there are areas for improvement to bolster the study's credibility and applicability. One notable enhancement would be to increase transparency in survey methodology, question formulation, and data-cleaning processes.

A clearer understanding of these aspects would contribute to a more informed interpretation of the findings. Moreover, the study could benefit from a richer narrative by incorporating qualitative insights or exploring correlations between variables. This qualitative layer could provide a deeper understanding of the motivations and nuances behind the observed trends, offering a more comprehensive view of the developer landscape.

To enhance the study's relevance, conducting analyses tailored to specific regions or industries would be invaluable. This approach would yield targeted insights, making the findings more actionable for distinct audiences and providing essential context to the observed trends.

In conclusion, while the 2023 Stack Overflow Annual Survey paints a broad picture of the developer community, incorporating the suggested improvements would elevate the study. Increased transparency, qualitative insights, and targeted analyses would strengthen the reliability of the findings, making them more applicable and insightful for a diverse range of stakeholders in the tech industry.

**4. CONCLUSION**

Our motive coming into the analysis was to dive into the 2023 Stack Overflow Annual Survey data from a perspective that resonates with us, college students. We want to uncover meaningful insights into the world of developers, exploring aspects like global demographics, education journeys, favorite programming languages, salary trends, and how artificial intelligence tools are making their mark. By doing this, we hope to paint a detailed picture of what's happening in the tech world, offering insights that are directly relevant to us as college students diving into the exciting field of computer science and programming.

The exploration of the 2023 Stack Overflow Annual Survey delves into the diverse realm of developers, unraveling details about their demographics, skills, and work practices. Spanning a colossal 89,184 responses, this survey serves as a comprehensive snapshot of the global developer community. It scrutinizes various facets, including age distribution, educational backgrounds, the popularity of programming languages, income trends, and the integration of artificial intelligence tools.

There are several reasons why computer science students like us would find this analysis to be very valuable. Examining the data from the 2023 Stack Overflow Annual Survey will provide students with information that they can use in the future to learn and apply for jobs:

**Demographic Insights:**

Understanding global developer demographics provides students with awareness of the diverse community they are entering. Insights into age, gender, and geographic distribution foster a sense of inclusivity and community.

**Educational Guidance:**

Examining the educational backgrounds of developers can guide computer science students in making informed decisions about their academic journeys. Insights into the types of degrees and educational paths pursued by successful developers can offer valuable direction.

**Programming Language Trends:**

Knowledge of popular programming languages and their usage trends helps students focus on acquiring skills that are in high demand in the industry. This information aids in curriculum planning and skill development.

**Salary Expectations:**

Understanding salary trends based on experience, location, and skill set allows computer science students to set realistic expectations for their future careers. It provides insights into potential earning trajectories and industry benchmarks.

**AI Adoption and Emerging Technologies:**

Exploring the adoption of artificial intelligence tools and emerging technologies in the developer community prepares students for the evolving landscape of the tech industry. This insight helps them align their skill development with the current and future needs of the field.

In essence, this analysis serves as a compass for computer science students like us, it offers guidance on educational choices, skill development, and career expectations within the dynamic realm of software development and technology. It empowers students to make informed decisions and navigate their academic and professional journeys effectively.