Stack Over Flow developer Survey 2024 Analysis in Python

BY MUHAMMAD HUSSAIN

Objectives

- To compare the Questions asked in the survey with developers responses
- Find About where most of the developers are from
- Find the age group of developers who respondent to this survey
- What is the highest level of Education maximum developers have
- What is the most common source of developers for learning programming
- For how many years they have coded with and without education
- What are the top 10 most used programming languages used by developers over the past year
- What online course or Certificates they used for learning
- Which are the most common Ai tools used by developers
- Which Ai developer tools they have used over the past year
- How many of the developers Use Ai tools in their Development and many more

Steps and Libraries I used in this Project

- Libraries:
  - Pandas
  - Numpy
  - Matplotlib
  - Matplotlib.pyplot
  - Seaborn
  - Counter from collections

Steps:

- Data preprocessing:
  - Import both datasets i-e survey_results_schema and survey_results_public
  - Convert both to pandas dataFrame
  - Made index to qname column for schema dataframe
  - Then created two lists for selected columns from both dataframes
  - Created two new dataframes from those selected columns data
  - In survey dataframe the age columns has string values likes ‘18-24 years old’ I convert it to numeric by taking its mean and then by MAP functions I change it to numeric
  - For ‘YearsCode’ and ‘YearsCodePro’ Columns which were object at first I convert it to numerice by pandas to_numeric function

Data Analysis

- Set some Customs style and params for Charts
- For Countries I used the values_counts() function and then took the top 10 Countries and plot it's graph
- For age, Education level,YearsCode I just simply plot there graphs using sea born and matplotlib
- For Columns like LearnCode, LanguageHaveWorkedWith, LearnCodeOnline, LearnCodeCoursesCert which had more then one options to select and each option were separated by a colon(;) so I split each option with split() function and then convert in into a list after that by using list comprehension I converted the nested lists into a single list and then applied a counter function from Counter libraries which gives us the count of distinct values in a column in a form or dictionary after that I just plotted it's Graph

Insights

- Most of the Survey Respondents are From United states with the value of 17500 and then from Germany and India with 7000 and 5000 values respectively
- Most of the respondent developers age is 30 with value greater than 30000 but a huge group of 20 and 40 age people also responded with the value of 18500 and 21000
- Most of the respondents highest level of education is Bachelors but a high group of Master graduated also responded then people without a degree were also in a high number
- Most of the developers learn coding through Online courses then book and Online Certifications courses and many more are following in the list
- The most common Ai tool used by developers over the past Year is chatgpt with the number of 50000 users then bing Ai and bard are following the list on 2nd and 3rd positions and then others are most less used
- Github copilot is the most used Ai developer tool by developers with the numbers of 20000
- About 44.4% of the respondent developers Use ai tools in the development and 29.8% want to use it soon while 25.8% have no plans to use it any time soon
- Most of the developers are in the favour of Using Ai in their development then second are very favorable and then unfavour and unsure respondents are very less
- Most of the developers wants to use the Ai tools for testing and then most want it also for Learning about Code base and commiting and debugging the Code

Ending Words From me About this Project

- These are some few Insights there is a lot more to get from these datasets as these are very large datasets and also I have selected only 18 columns while there were a lot more columns but I have only selected 3 types from datasets which are about demographics, Programming and Artificial Intelligence
- I did this for showcasing some of my Python and Data analysis Skills I know there is a lot more in these datasets and can be analyzed more in dept but at the end as it is only for my resume so I analyzed it to here only Hope you will like it
- GitHub Link

Thank you