# The Stack Overflow Data Set Analysis –

Dataset Overview –

The dataset comprises of people and their background applying for a job at Stack overflow. Their entire educational and professional background with personal details are shared to evaluate how fit they will be in the organization.

Problem Statement –

We need to perform detailed analysis of the data points shared with us. Which aspect is impacting which ones to determine one’s credibility or the ability to secure a job in the organization needs to be scrutinized.

**Technical Expertise** -

* Determining that what share of applicants posses a hobby of coding. As obviously people having coding as a hobby is more likely to be passionate about it and should turn out as an excellent asset for the organization however not everyone takes their hobbies seriously.
* Checking if the people code as a hobby if also contribute to the open source project. This showcases their thoughts towards the tech community.
* The number boot camps/Hackathons did the applicant appeared in to showcase the eagerness of the applicant in a career on tech industry.
* Language , Frame work , Dev Platforms , Operating systems used till date.
* Professional and overall coding experience.
* If the applicant is self paced and doesn’t always require to be pushed.

**Ethics** –

* Ethical practice followed by the applicants.
* Agreement/disagreement of the applicant with the laid down ethics policy of the organization applying for.

**Personal Dossier –**

* To evaluate that from which background most applicants are from.
* Their educational backgrounds and determining the trend of which educational background people are applying in Stackoverflow.
* US military background , if applicable.
* Gender , sexual preference , Medical status etc of the applicant.

**Expectation –**

* Remuneration expectation of the applicants and if they are in sync with organization’s interest.
* The salary type expected (Monthly /weekly) etc
* Track record of previous organizations and if the expectations are in sync with that.

## Inputs Shared –

A brief detailing of the kind of input shared here in dataset for the evaluation. This inputs are primarily the expectation of the applicants based on which they’ll make a career choice with Stack Overflow

**Job Assessment Based On** **-**

* Financial performance/Funding status of the company.
* The specific department applied on.
* The tech stack to be working with.
* Diversity , perks and benefits , Organizational work environment.
* Self growth scopes/opportunities.

**Other preferences** –

* How should the employer contact the employee.(methods of communication delivered).
* Involvement in stack overflow competitions .

# Workflow –

The flow in which execution of the project will proceed will be described here –

* **Exploring the Dataset –** Reading the dataset on a given variable. Checking it’s rows and columns. The value counts of each. In our given dataset there are 129 columns and 98855 rows with details.
* **Missing Values –** Not all columns and rows will have inputs in them. Finding out these missing values are very important in order to analyze the data properly. The below columns are having the maximum share of missing values. These columns where majority of the values are missing , will not play any significant role in analyzing the dataset , hence they will be removed from the dataset.
* **Time After Bootcamp –** 93.27% values are missing.
* **Hackathon Reasons -** 74% Values are missing.
* **Ergonomic Devices –** 66% Values are missing.
* **Stack Over Flow Jobs Recommended –** 61% Values are missing.
* **AD Blocker Reasons –** 65% Values are missing.
* **MilitaryUS –** 84% Values are missing.
* **Other Missing Values –** Apart from the above stated columns there are other columns as well with missing values but the share of missing values in them are less than the standard benchmark.
* **Handling Missing Values -**  There are several methods using which one can handle the missing values present in a dataset. By analyzing the data it’s evident that the ideal approach for this dataset will be the mode method , where the missing values gets replaced with the most frequent occurred value in that particular column.
* **Research Method –**
* Finding values counts of all major aspects.
* Finding which feature is impacting other ones.
* Evaluation of Co relations.
* Visualization with necessary libraries for better understanding.
* Use describe method to find the mean value of assessment of individual priorities.
* **The Findings -**  The findings of the research work done extensively done on the dataset reflects the result as below –
* The majority of applications processed through USA (20721) & India( 13721)
* Although Linux being the ideal companion for a developer , The ratio of OS using is much higher with windows(60698) . Apple(MacOs-20325) and Linux(17684) stands far behind.
* Visual Studio remains as the best choice for the applicants worldwide.
* Python as Language , MYSQL for database management remains the number 1 choice for the applicants.
* The working applicants generally is associated with mid sized firms (20-100 employees)
* The majority of the people in the dataset are working full time.
* On average most of the applicants are back end and full stack developers having , 3-8 years of experience.
* The majority of workforce in the tech world comprises the age group of 25-34 or at least our data tells us so.
* Just like the real world , male white supremacy reflects on our dataset as well ,out of 98000 odd applicants , 78000 are white male.
* Just like the ideal scenario , majority of the applicants are from computer science background(nearly 71,000)
* Only handful of applicants(14000 only) are contributing on stack overflow by participating.
* The majority of the ongoing generation is ready accept the challenges and new horizons which AI is expanding into.
* The traditional study pattern is changing rapidly , while the previous generation was more into studies of normal graduation , this generation is way more focused on technology and computers while choosing their subjects.

## The Preference Of Applicants while selecting an organization to work for are as given below and is derived from the dataset -

|  |  |  |
| --- | --- | --- |
| The Industry | Technologies Used | Opportunity of growth |
| Funding Status of the company | Compensation and Benefits | Organizational Diversity |
| The Working Department | Work Culture | Product |
| Technologies Used | WFH Options | Salary & Benefits |
| Compensation and Benefits |  | Stock Options |