MDSC-106 Project

# Stack Overflow Developer Survey 2021 Analysis

Mrinal Kanti Saha

21234

I MSc Data Science and Computing

# Contents

## Problem Statement

Stack Overflow is a question-and-answer website for professional and enthusiast programmers. It is the flagship site of the Stack Exchange Network, created in 2008 by Jeff Atwood and Joel Spolsky. It features questions and answers on a wide range of topics in computer programming.

Every year, Stack Overflow conducts a Developer Survey which gets tens of thousands of responses from the user base. It is the longest running survey of software developers (and anyone else who codes!) on Earth. The Stack Overflow Developer Survey 2021 has 39 questions. This survey is the source data of our project with nearly 80,000 responses fielded from over 180 countries and dependent territories.

Our goal is to find some interesting inferences from the dataset of the survey. Most of the analysis has been done on India specific data. Let's see what we can find.

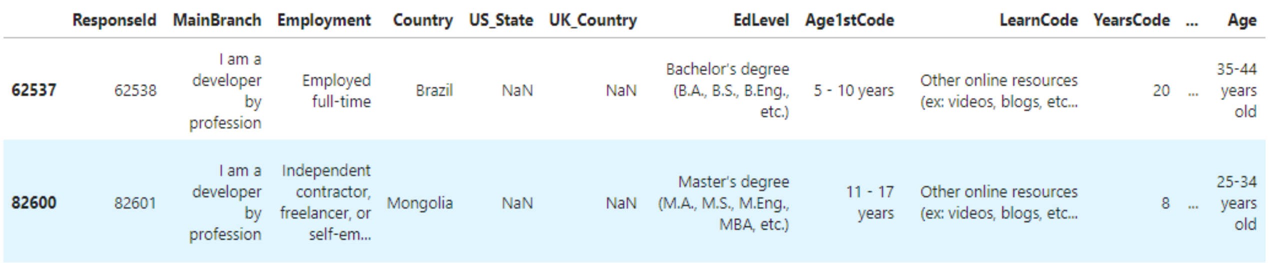
## Data Description

The dataset can be downloaded from here. The link takes you to the page that contains the datasets of all the surveys conducted over past 10 years (and guess what, also the analysis of those data25).

The downloaded zip file contains 4 files :-

- survey_results_public.csv - The main dataset that contains responses to the survey questions. One respondent per row and one feature per answer
- survey_results_schema.csv - The schema set that maps the questions to each feature name.
- so_survey_2021.pdf - PDF file of the survey instrument.
- README_2021.txt - providing other meta information.

The data frame has 83439 rows and 48 features. Here's a snippet of the dataset.



| | ResponseId | MainBranch | Employment | Country | US_State | UK_Country | EdLevel | Age1stCode | LearnCode | YearsCode | ... | Age |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **62537** | 62538 | I am a developer by profession | Employed full-time | Brazil | NaN | NaN | Bachelor's degree (B.A., B.S., B.Eng., etc.) | 5 - 10 years | Other online resources (ex: videos, blogs, etc... | 20 | ... | 35-44 years old |
| **82600** | 82601 | I am a developer by profession | Independent contractor, freelancer, or self-em... | Mongolia | NaN | NaN | Master's degree (M.A., M.S., M.Eng., MBA, etc.) | 11 - 17 years | Other online resources (ex: videos, blogs, etc... | 8 | ... | 25-34 years old |

There are far too many features to be investigated and so, we will choose the columns that we are interested in.

Namely the following :-

- MainBranch - Are you a professional, hobbyist, student?
- Employment - Are you employed? Full time, part time, etc.
- Country - Which country do you live in?
- EdLevel - What is the maximum education you have completed?
- Age1stCode - At what age did you write your first piece of code?
- LearnCode - How did you learn to code?
- YearsCode - How many years have you coded (even as part of your education)?
- YearsCodePro - How many years have you coded professionally?
- DevType - What is your job role (if any)?
- LanguageHaveWorkedWith - What languages have you worked with this year?
- LanguageWantToWorkWith - What languages are you looking forward to working with?
- DatabaseHaveWorkedWith - What DBMS have you worked with this year?
- DatabaseWantToWorkWith - What DBMS are you looking forward to working with?
- PlatformHaveWorkedWith - What cloud platforms have you worked with this year?
- PlatformWantToWorkWith - What cloud platforms are you looking forward to working with?
- OpSys - What operating system do you use?
- NEWStuck - What do you do when you get stuck?
- SOVisitFreq - How frequently do you visit SO?
- Age - What is your Age?
- Gender - What is your gender?
- MentalHealth - How is your mental health?
- ConvertedCompYearly - How much do you earn yearly (in dollars)?

We will be dropping features such as 'Currency', 'CompTotal', 'CompFreq', 'US_State', 'UK_Country', etc as they do not offer information of our interest.

We have reduced the number of columns to 22 from 47 dropping a few features.

Let’s head on to the cleaning part.

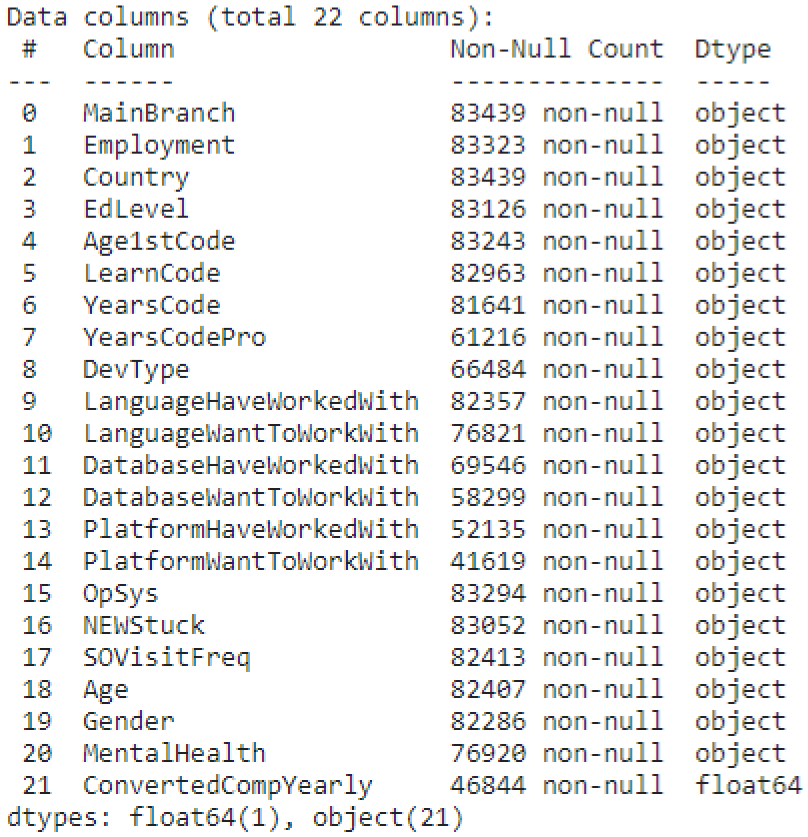
## Data Cleaning and Preparation

So, what is data cleaning?

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset which may lead to problems in analysis.

The notion behind this kind of cleaning is that if any of the fields are incorrect then we assume that the whole response is incorrect.

Now, we'll see some basic stats about our current dataframe.



```
Data columns (total 22 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   MainBranch              83439 non-null  object 
 1   Employment              83323 non-null  object 
 2   Country                 83439 non-null  object 
 3   EdLevel                 83126 non-null  object 
 4   Age1stCode              83243 non-null  object 
 5   LearnCode               82963 non-null  object 
 6   YearsCode               81641 non-null  object 
 7   YearsCodePro            61216 non-null  object 
 8   DevType                 66484 non-null  object 
 9   LanguageHaveWorkedWith  82357 non-null  object 
 10  LanguageWantToWorkWith  76821 non-null  object 
 11  DatabaseHaveWorkedWith  69546 non-null  object 
 12  DatabaseWantToWorkWith  58299 non-null  object 
 13  PlatformHaveWorkedWith  52135 non-null  object 
 14  PlatformWantToWorkWith  41619 non-null  object 
 15  OpSys                   83294 non-null  object 
 16  NEWStuck                83052 non-null  object 
 17  SOVisitFreq             82413 non-null  object 
 18  Age                     82407 non-null  object 
 19  Gender                  82286 non-null  object 
 20  MentalHealth            76920 non-null  object 
 21  ConvertedCompYearly     46844 non-null  float64
dtypes: float64(1), object(21)
```

Most of the features in the data frame have the data type of object.

We will try to convert some of the features into numeric types. Features like YearsCode, YearsCodePro can be converted to numeric types.

Now, let's remove the null rows (if any). Although, as it is a survey, I don't feel there would be such responses.

Now, let's account for the wrong values entered by mistake or intentionally.

- In Gender, we have people who have chosen multiple options (like Man; Woman) that do not make sense. Hence, we will drop these rows.

Now let's clean the rows with the wrong Age. And what do we mean by wrong age?

- We will remove all those rows where Age is less than the age when a person first coded. Because it makes no sense.
- We will remove all those rows where Age is less than the number of years coded.
- We will remove all those rows where Age is less than the number of years professionally.

Finally, we have come to the end of the data preparation and cleaning. The number of responses has been reduced to 82566 from 83439.

Also, we are doing EDA on India specific data. Henc, we will prepare an India specific Dataframe which contains about 10 thousand rows.

Although there can be many more metrics to clean data, we have cleaned our data sufficiently well to head on to the visualization and analysis part.

## Exploratory Data Analysis and Data Visualization

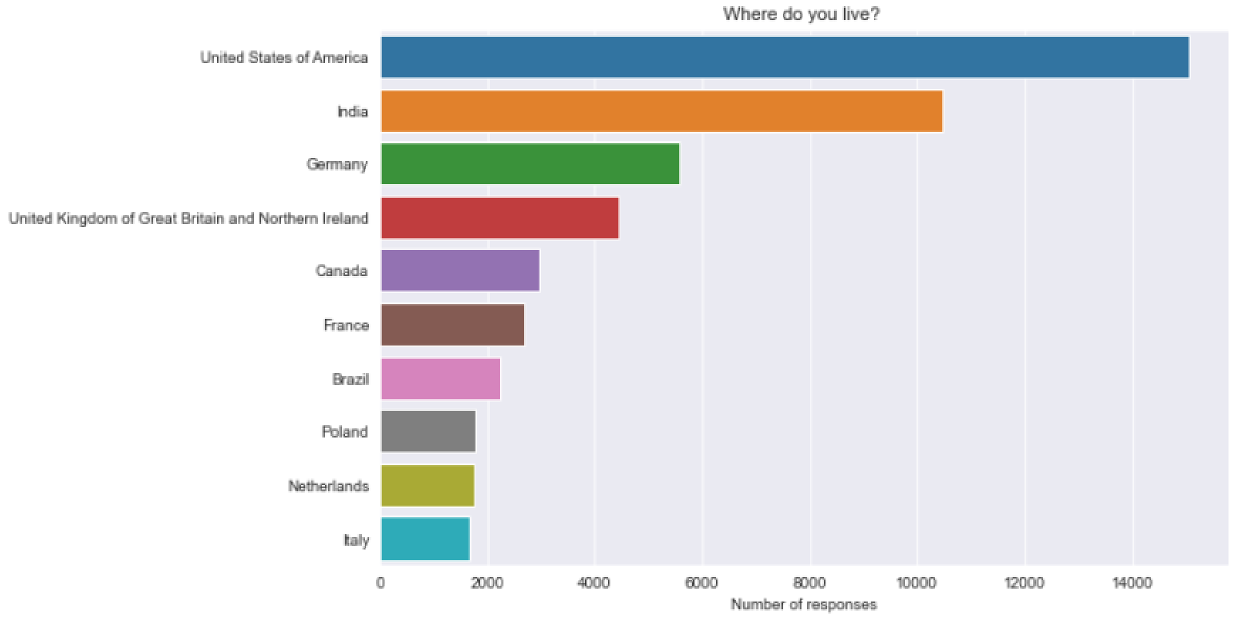
The Official Stack Overflow Developer Survey 2021 Analysis is linked here.

Here, in this section, we will try to see some charts and plots to better understand the data we have here and also we would like to ask some questions to understand the data more comprehensively.

---

### Top 10 Countries of the World

Let's have a look at the demographics of the respondents.



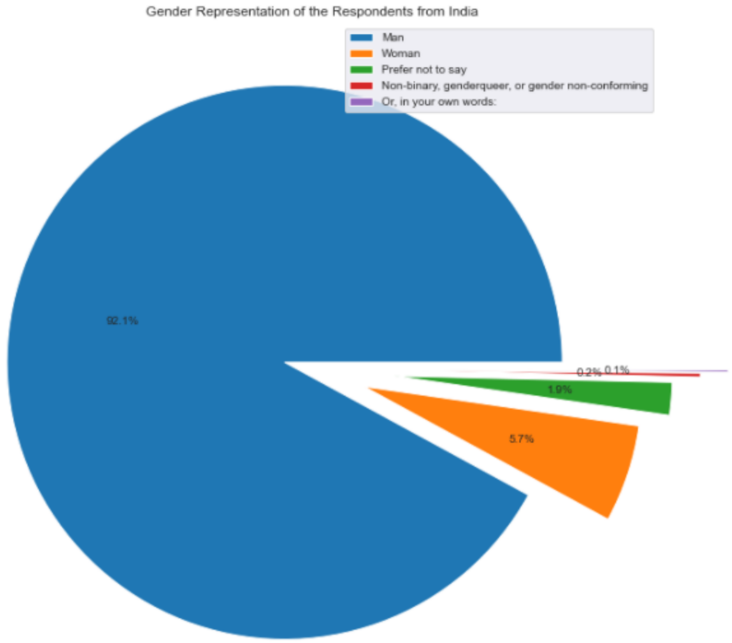
About 15000 responses are from the USA and about 10000 are from India (as we had seen earlier). These two top countries alone make up about 25% of the total responses.

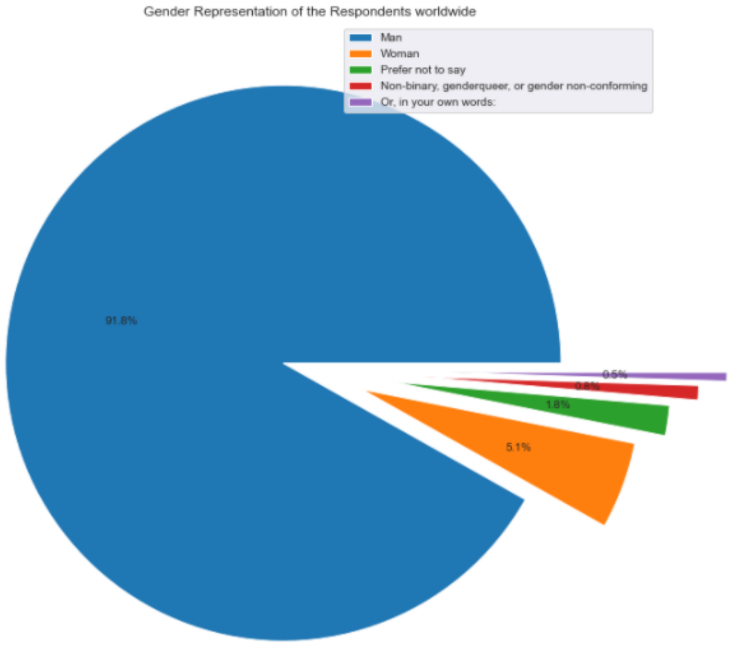
We can infer that the SO community has a very high representation from the USA and India.

Now, we can also say that people from the USA, UK and India are more likely to answer surveys than other people because the survey is available only in English.

---

## Gender Demographics in India vs The World


Gender Representation of the Respondents from India

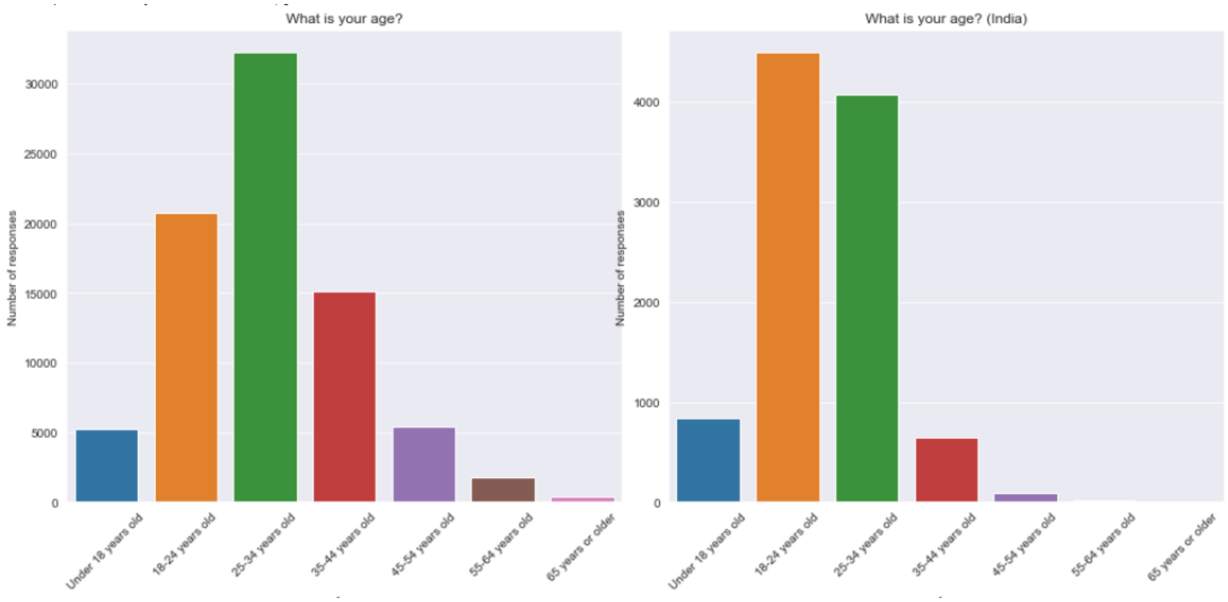

Gender Representation of the Respondents worldwide

Sadly enough, the gender divide is pretty huge. There is very less representation from the female and the non-binary, genderqueer and gender non-conforming genders.

Huge imbalance with about 92% of the respondents being male. Though the field is male dominated, the representation of other genders should be a bit better. Or, we can say that men are more likely to respond to a survey than other genders.

Another interesting inference is that in India, there are very few non-binary gender people according to the survey and a lot of people who are not comfortable to share their gender (as compared to the world). Hence, we can safely infer that the third gender concept is still quite alien in India and because people from other genders do not accept them easily, they choose to hide it and "Prefer not to say".

## Age Demographics in India

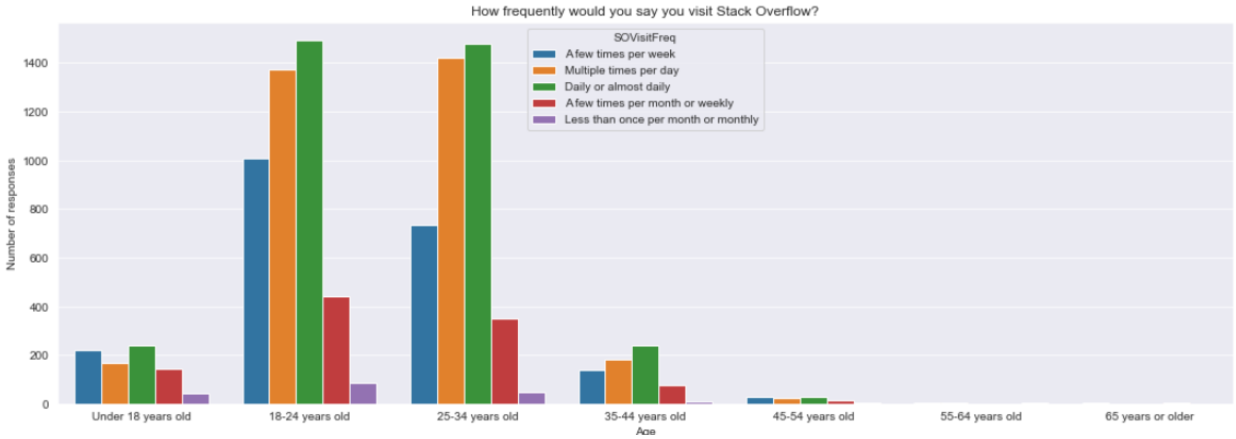


The majority of the stack overflow users can be found to be within the age bracket of 18 to 34 with around 55000 of the responses. We can safely assume that college students and job freshers and trainees visit Stack Overflow for solving their problems (or, that young people are more likely to take surveys than middle aged and old people).

There aren't many respondents from India in the age bracket 45 and older. So, it can be said that compared to other places, people in India had a later exposure to the Internet and related stuff. And, so the newer generation has more representation.

## Age VS SO visiting frequency (India)

Does Age and SO visiting frequency have a relation?

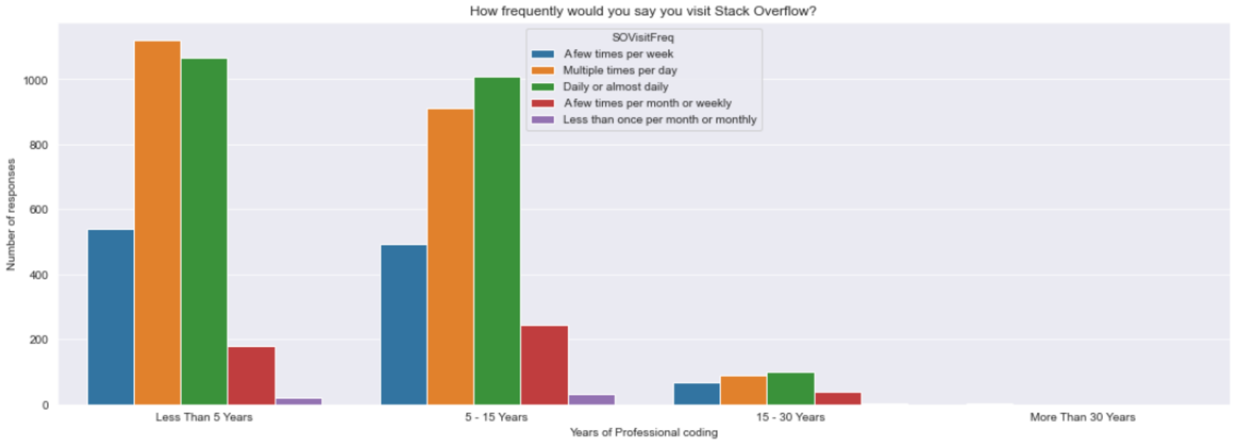


In most age groups, people usually use stack overflow daily or almost daily or a few times a week. As Age increases so does experience, and so people do not need to search for things on stack overflow very often. Also, in the "45 years or older" bracket, the popular usage is a few times a week, which strengthens our assumptions.

---

## Years of Professional Coding vs SO Visiting Frequency (India)

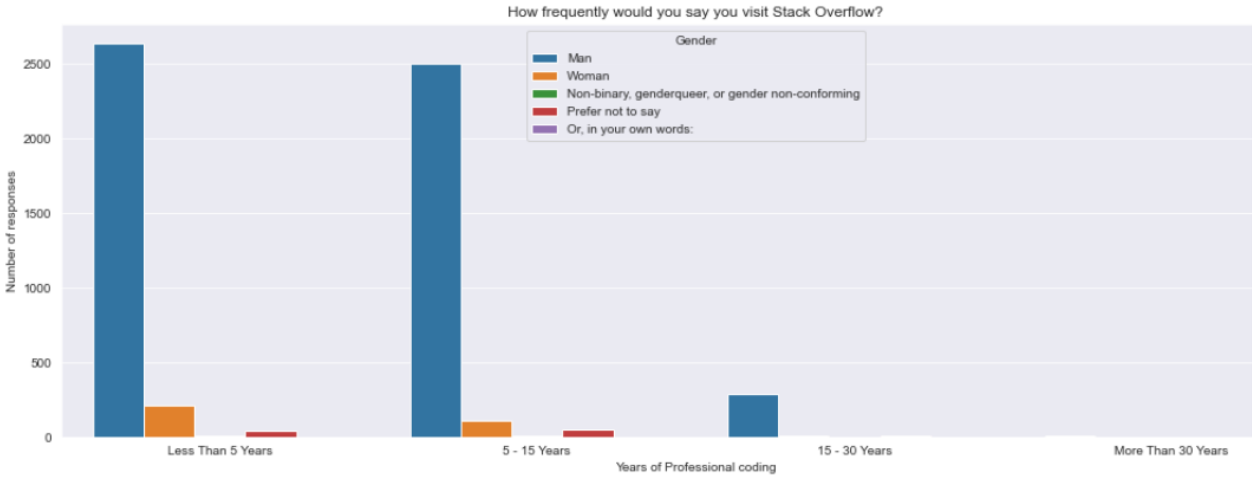
But, the above metric was rather crude. Let's find the effect of Years of professional coding on the frequency of visiting Stack Overflow.



There, we go. Now that we are comparing years of professional coding, we can clearly see that people with more coding experience visit SO less frequently.

## Gender representation in the industry over years

Here, we want to see how the genders are represented now and how they were before.

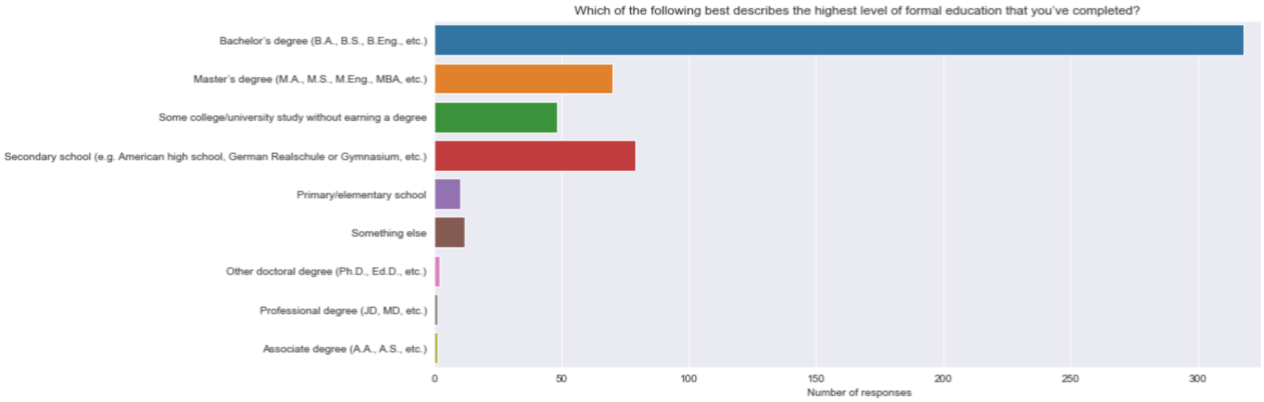


From the graph, we can infer that a few decades back, there weren't many women or non binary gender in the community. It used to be a male dominated community. But the silver lining is that slowly the representation of the other genders are getting better, especially in the last 2 decades. But, we still have a long way to go.

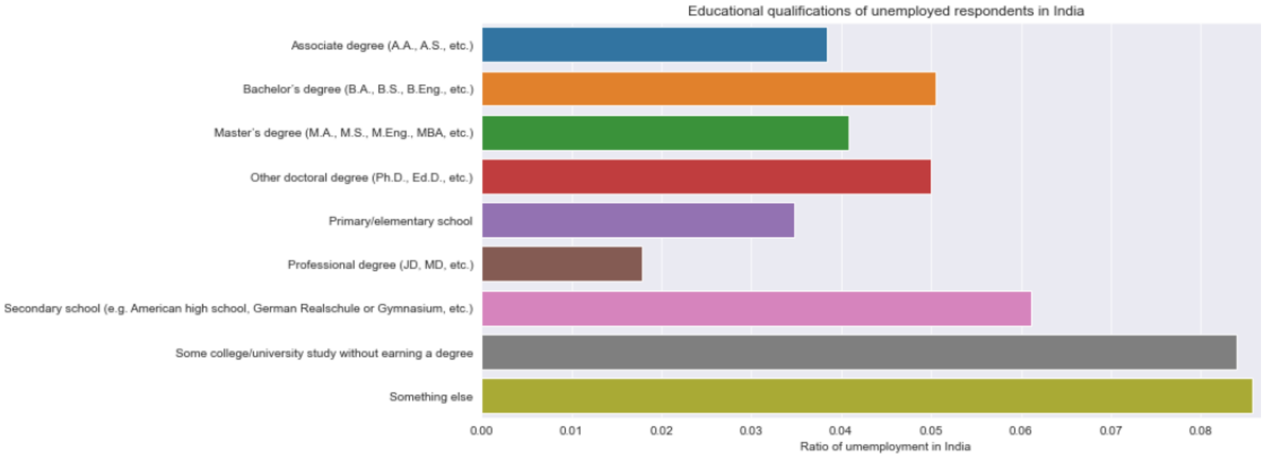
Again, here we find that the non binary gender is inappropriately represented till date. And simultaneously, "Prefer not to say" is in good amount, inferring the same that we found out before.

## Educational Qualifications and Employment Status in India

Let's visualize the Education Levels of the Unemployed people, "Not employed, but looking for work", in India again.



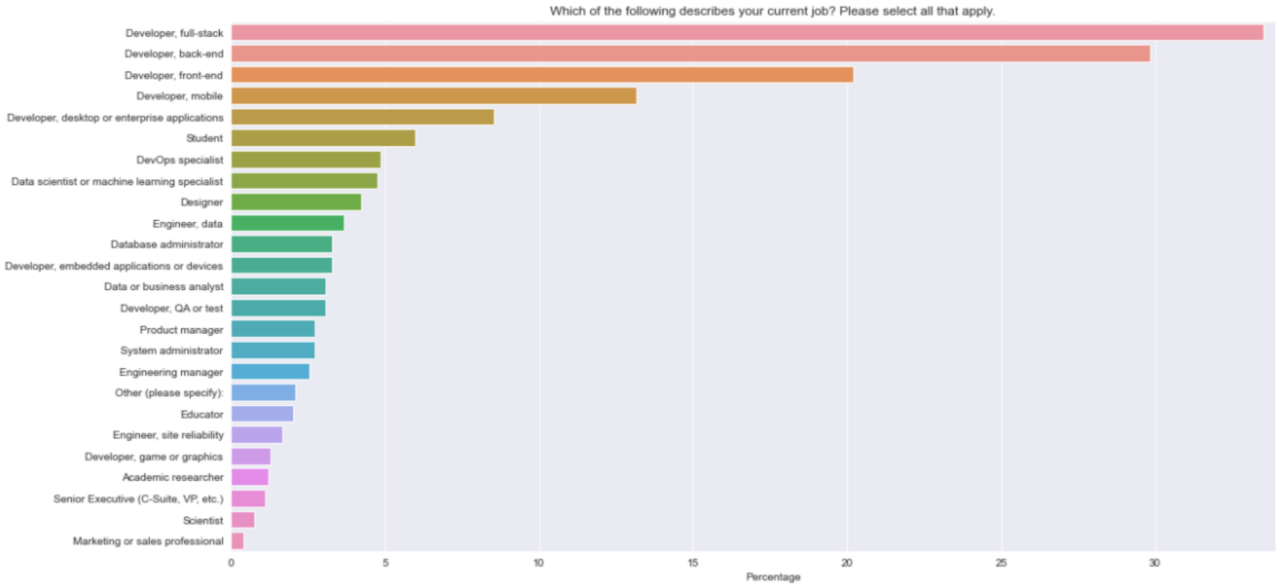
The unemployed people in India are majorly Bachelor degree holders or those who have completed secondary school. Even after a Masters degree, it seems people are still struggling to get into jobs. People with Professional degrees or Doctoral degrees, however, seem to have been better off.



Comparing the ratio of Unemployment of India, it is observed that the number of unemployed professional degree holders to the number of people with professional degrees is the lowest. Comparing the ratios of the master degree holder and the professional degree, we can infer that it is better to go for a professional degree (if possible) than masters or associate degree if a person is looking for a job after the degree's completion.

## Current Job of the respondents (India)

Let's see what all jobs people of India are doing.

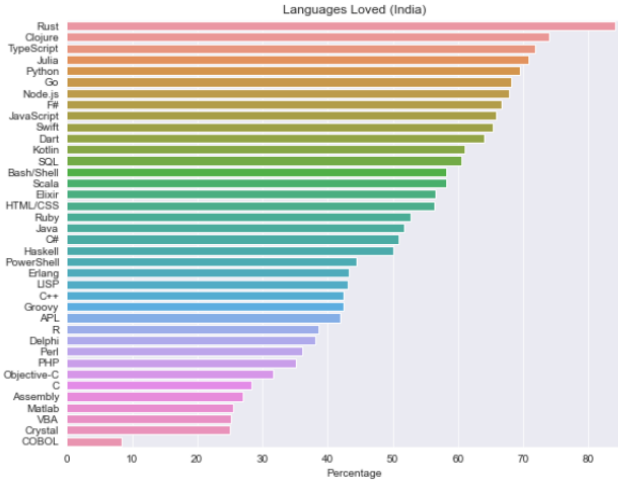


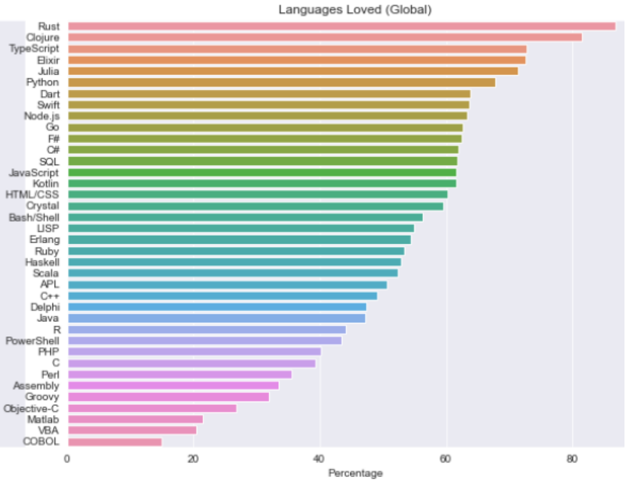
As SO is mainly a forum for developer related stuff, we expected to see a higher percentage of developers (full-stack, front-end and back-end) than others.

---

## Programming, Scripting and Markup Languages Loved (Indian Edition)

Let's see the languages that the respondents have loved working with this year.

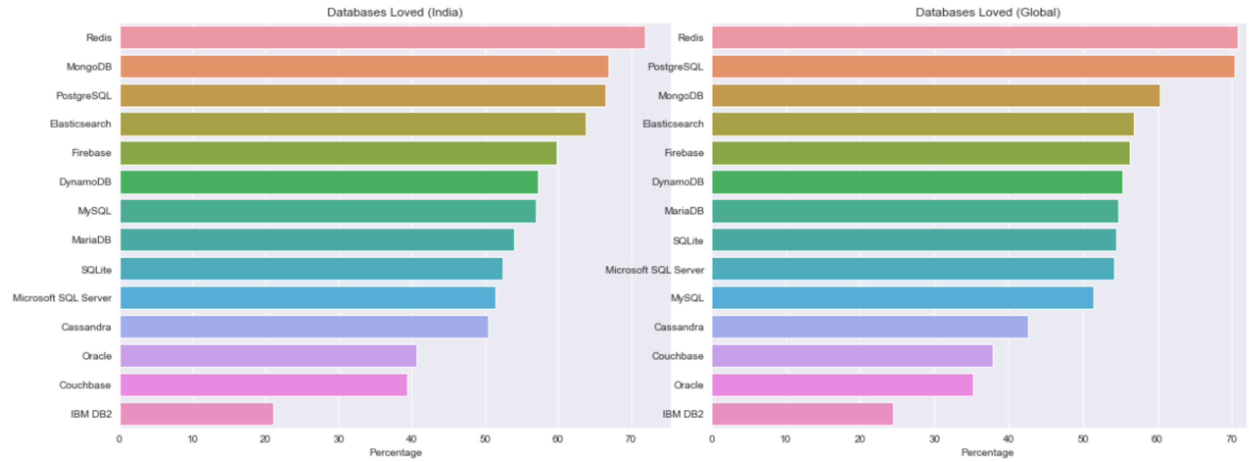




Rust is the most loved language globally and also in India. This is totally unexpected. But it is said that people working with Rust tend to fall in love with it, i.e. Rust retains its user more than any other language. These kinds of results are a feast to an analyst.

## DataBases Loved (Indian Edition)

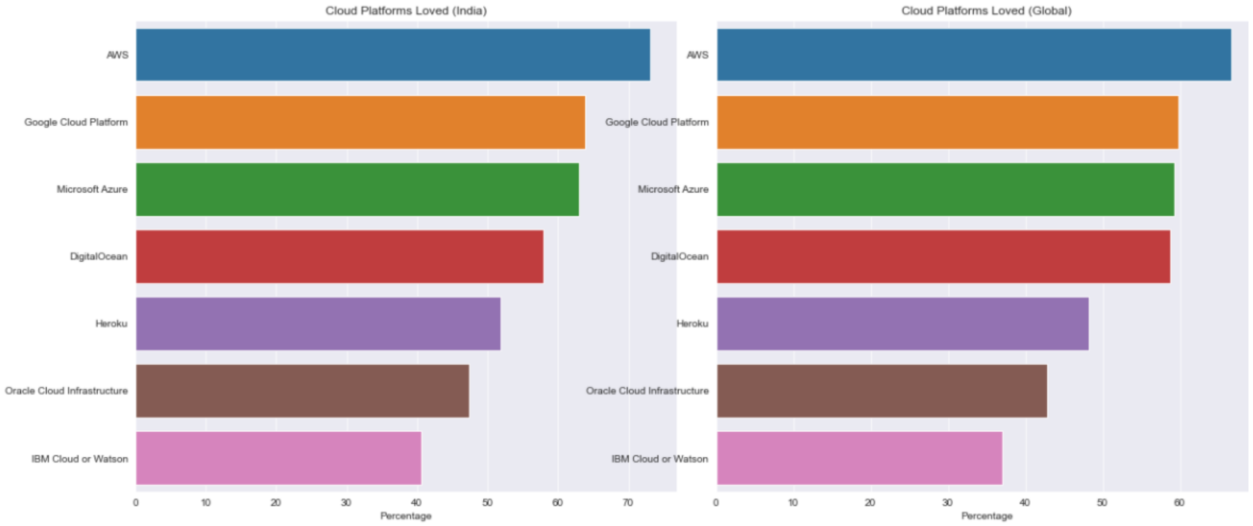
Let's do a similar type of analysis with DataBases.



In India and globally, people are loving Redis, i.e. Redis retains users more than the rest of its competition, followed by MongoDB, two NoSQL databases on the top.

## Cloud Platforms Loved (Indian Edition)

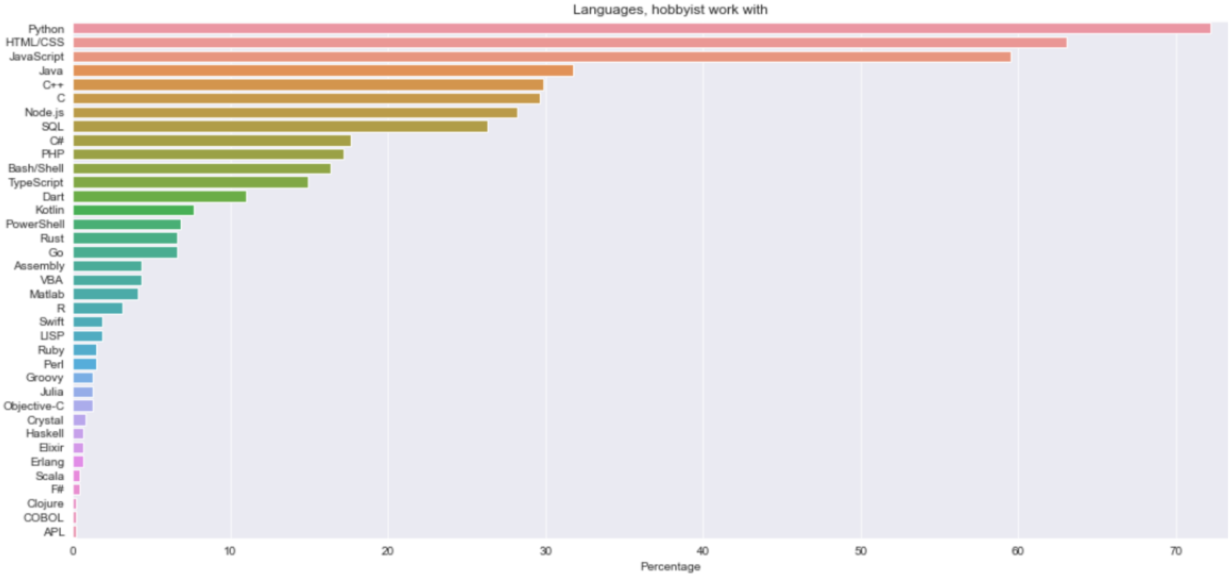
Lets see the most loved cloud platforms used in India.



AWS is the most loved Cloud Platforms which is quite expected, IBM Cloud or Watson is the most dreaded Cloud platform, globally as well as across India.

## Language of the Hobbyists in India

Let's see which languages hobbyists find interesting and easy to learn.

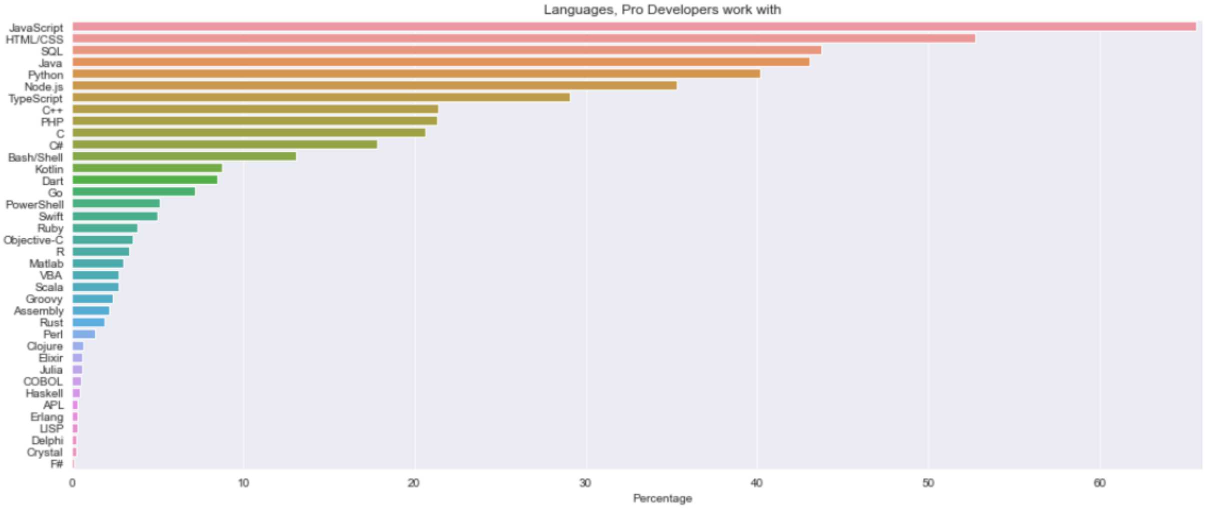


This instates the fact that Python is the most easy-to-learn language. And why not, Python is a general purpose language which suits the needs of the hobbyist, without any trade offs as it is a very powerful and versatile language.

---

## Language of the Developers in India

Let's find the language that professional developers use, i.e. which languages have a good market demand in India in 2020.

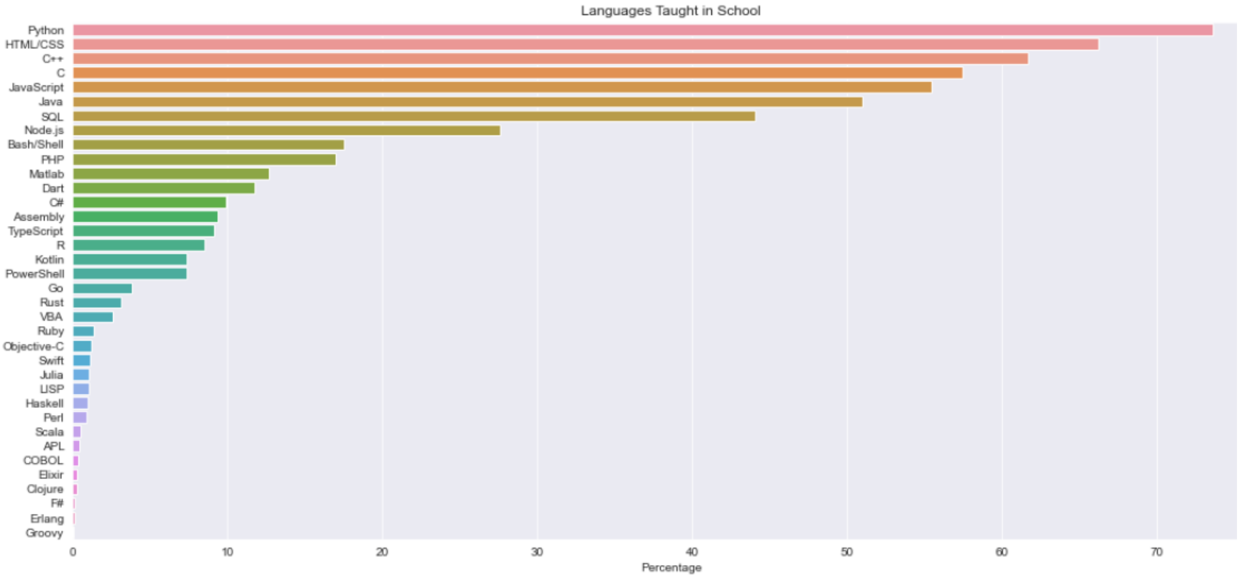


JS, HTML/CSS, SQL form the top three languages used by professional developers in the year 2020. Python isn't used as extensively as we may have assumed, having seen its popularity, still makes it to the top five. Java is still in use and more than Python. So, if you are looking forward to hopping onto such careers, you have a fair idea what to go for.

---

## Languages Taught in Indian Schools

Let's now see if the languages taught in schools are up to date with the requirements of the industry.

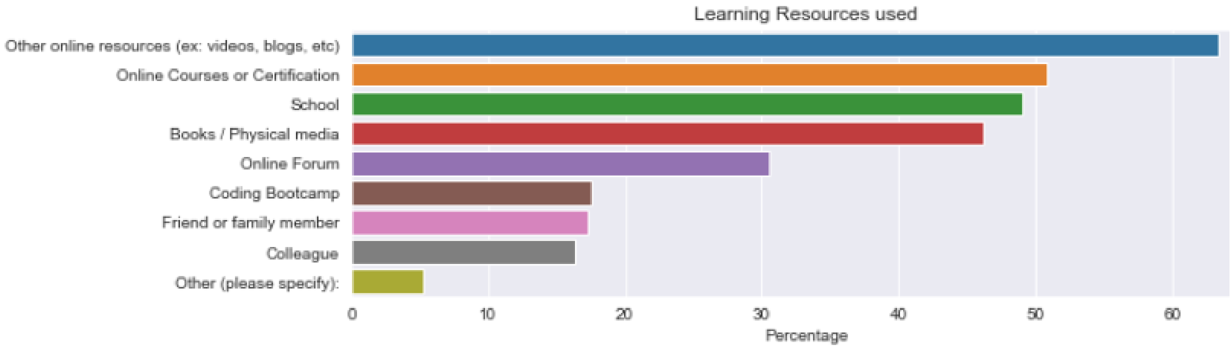


Well, as per the survey, the Indian education system, in terms of coding, is actually quite good. JavaScript lags behind a bit. C++ and C are way up in the bar graph, though they are not extremely popular among developers here. But, these languages, though not used by developers, are like the basic languages of coding. So, it gets a green flag here.

---

## Learning Resources

Let's see what is the most used learning resources for coding.

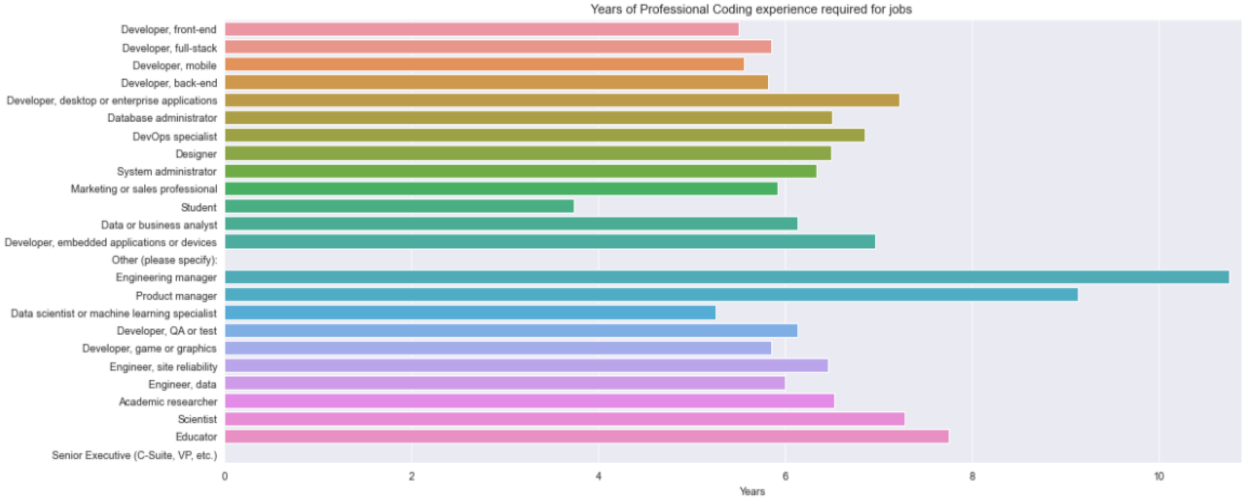


We can see that people prefer Blogs and Tutorials video to Coding Bootcamps and Certification Courses. Probably, that's because they can learn the language and tech of their choice (not much of a choice is schools) and also that courses and bootcamps are often paid. Also, Schools are a major learning resource as per survey.

---

## Years of Professional Coding Required for Job Roles in India

How many years of professional coding are needed for various jobs?

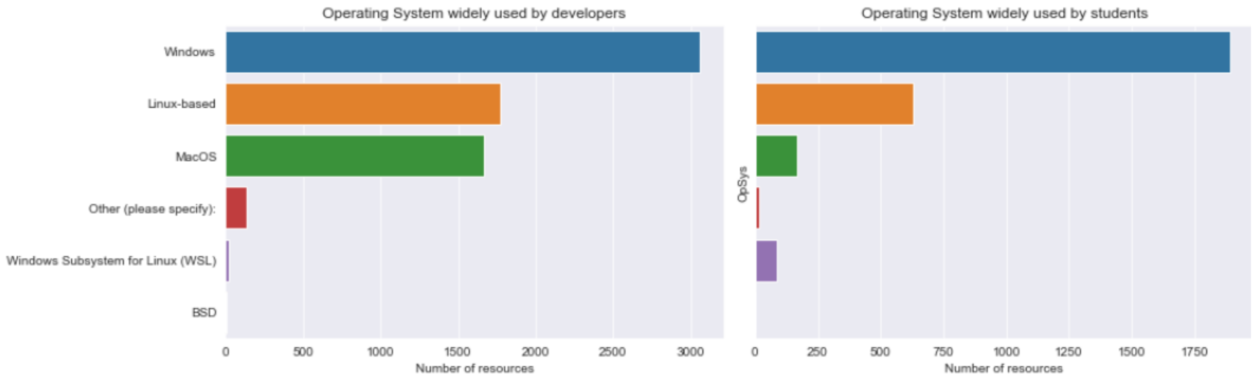


We can infer that Managerial positions (like Engineering Manager, Product Manager) require the most amount of professional coding experience, followed by Educators. Senior Executive could have been the most demanding of professional coding experience, but sadly we do not have any Senior Executive respondents from India.

---

## Operation System widely used by developers and students

Let’s see the comparison of Operating Systems used by students and pro developers.

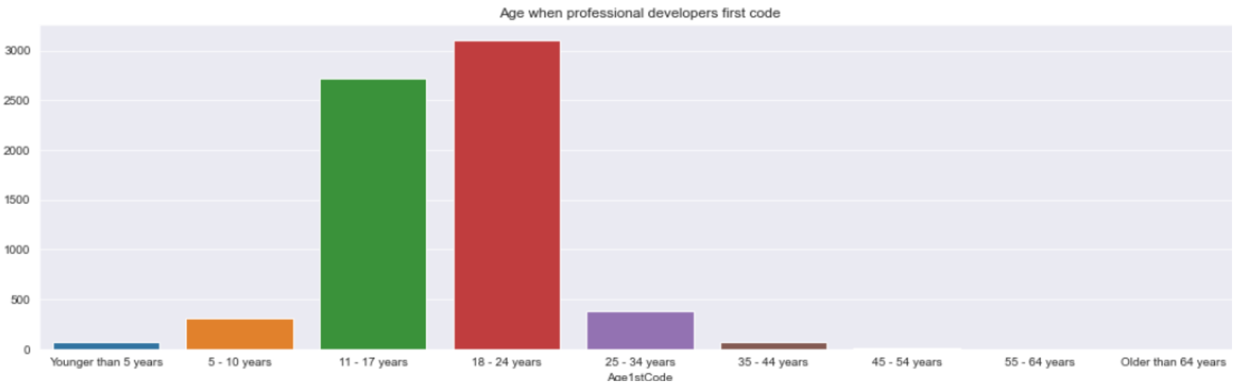


Windows is the most used operating system among both students and professional developers. But if we sum up all the Linux based OS such as Linux, BSD, MacOS, also WSL, we can see that Linux-based Operating Systems are more used than Windows in case of developers. Hence, developers prefer Linux based OS over Windows.

---

## When should you start coding to be a professional developer?

Let's see how age of 1st coding and developer are related.

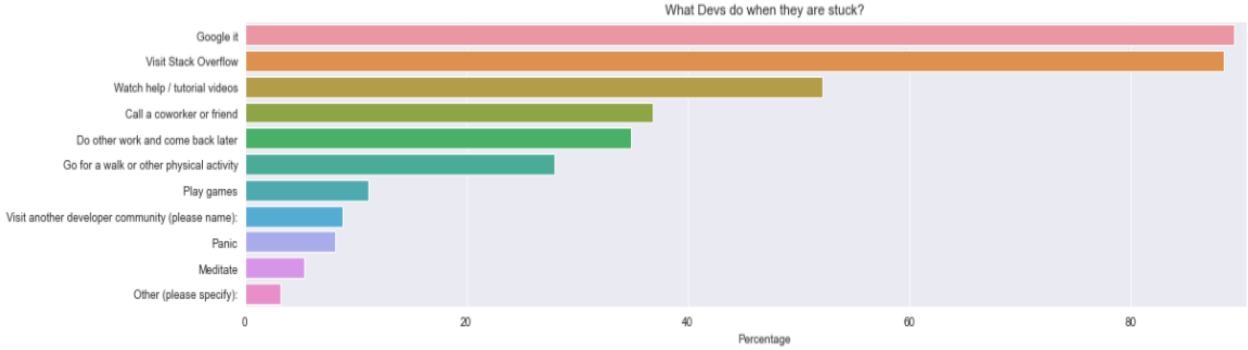


Most of the professional developers start coding at the age of 11 to 24, i.e. they get introduced to coding at schools and colleges. However, there are also professional developers who started their coding journey at the age of 25 - 34 and even some from 35 - 44 years, stating that coding is something that you can start learning even in your late 20s and still build a good career out of it.

Younger than 5 years sounds too good to be true, but because the SO developer survey has this option, we aren't removing these responses.

## What do professional developers do when they are stuck?

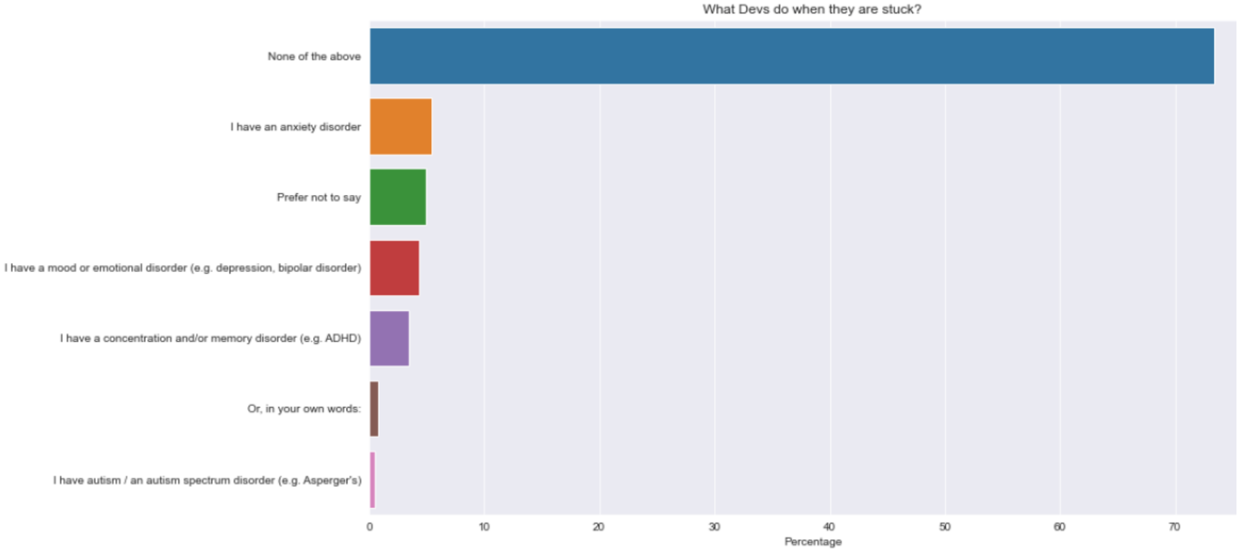
Let's see what professional developers resort to when they are stuck and can't get over it.



Googling the problem and visiting stack overflow are the most widely used methods of finding a solution. Devs when stuck on something, they tend to solve it asap rather than taking a break, going for a walk, playing games or meditating.

## Mental Health Issues

Let's see the mental health problems faced by people in India.



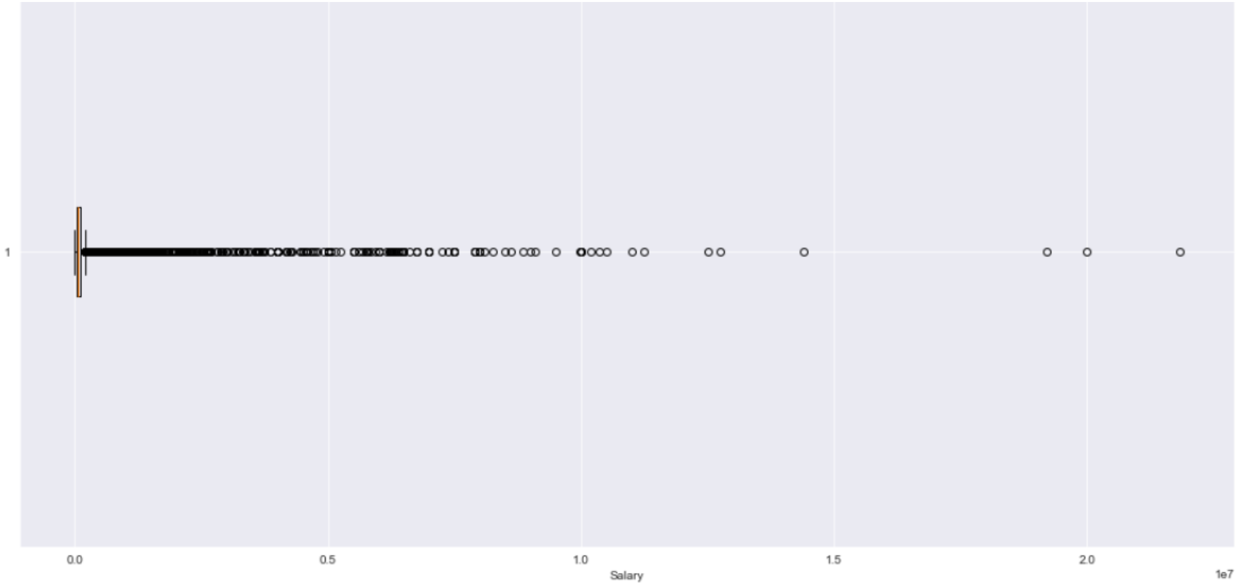
Most of the respondents are well off having no issues of mental health. Doing this analysis on students and developers does not yield very different results.

## Data Modelling and PDA

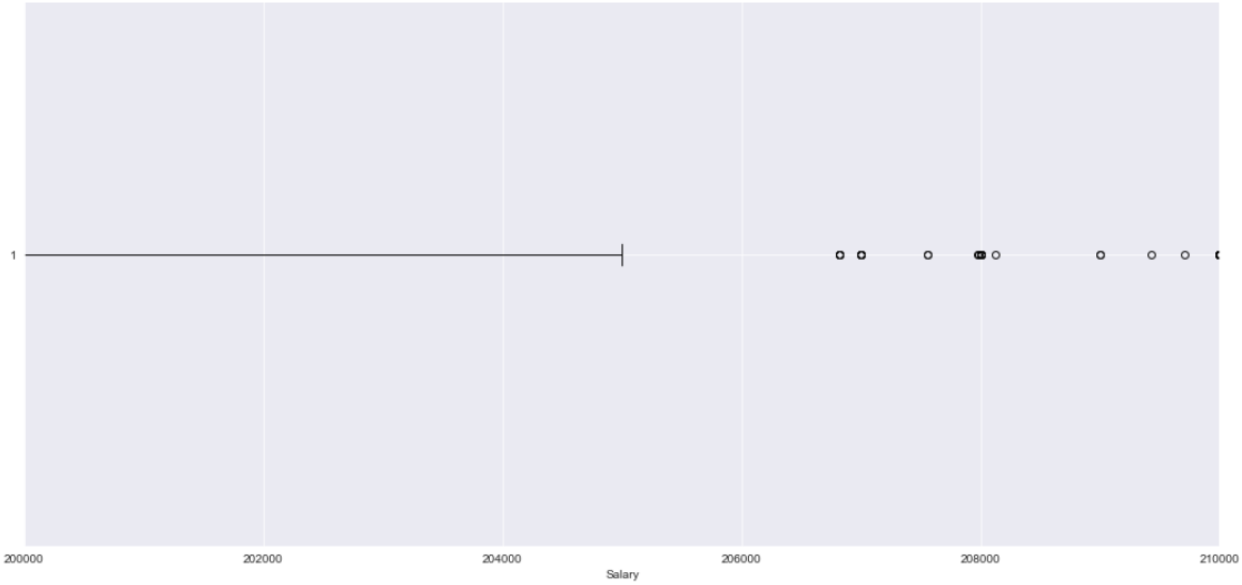
Let's now try to fit a linear model that predicts the Salary or Compensation yearly based on Job Designation, Educational Qualifications, Age, Gender, Country and Years of professional coding.

First, cleaning of these features should be done so that we have a better model.

Next, let us eliminate the outliers, which we will find by looking at the boxplot.

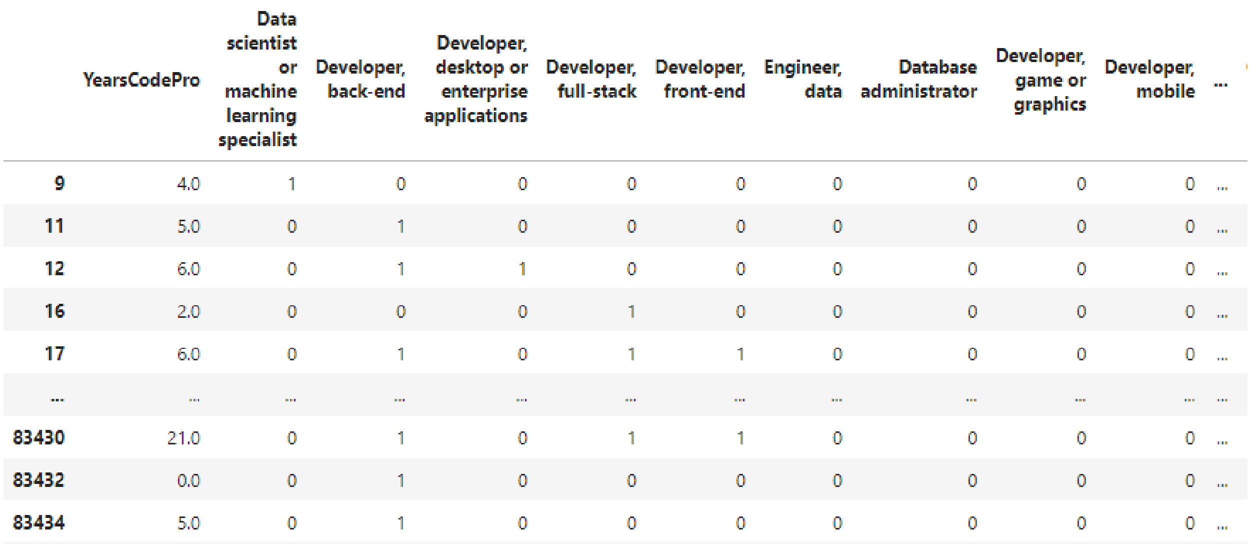


Let's zoom in to find the lower extreme of the box plot and remove all those data points are beyond the lower extreme.



Finally, we have got the outliers and now we will eliminate these values from our dataset and get 33 thousand rows for modelling.

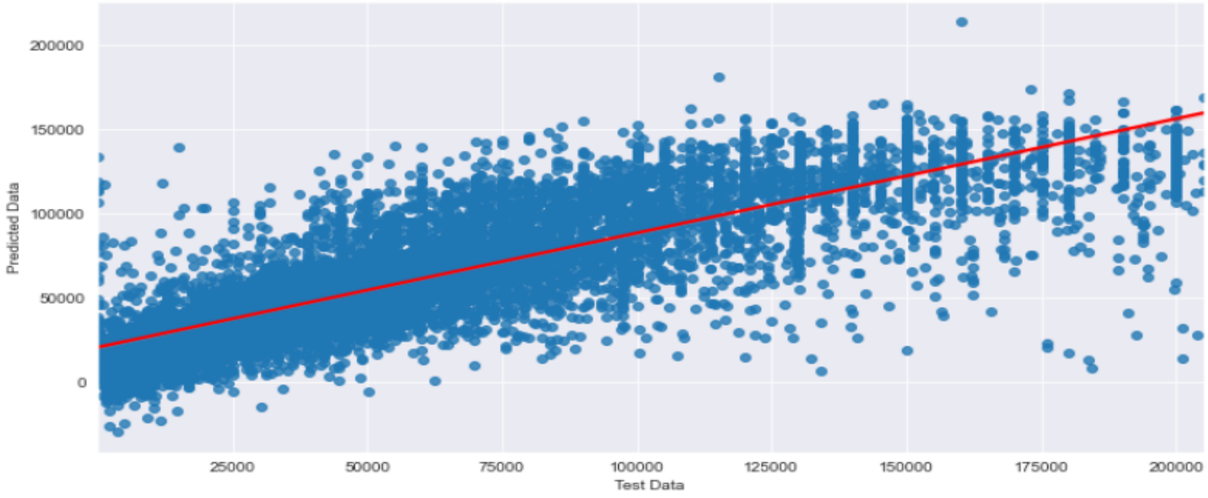
A dummy variable is a numeric variable that represents categorical data, such as gender, race, etc. As Gender, Job Designation, Education level, Age and Country are categorical data, we will have to introduce dummy variables for each category. This is how the independent variables data frame looks after converting categorical variables into dummies.



| | YearsCodePro | Data scientist or machine learning specialist | Developer, back-end | Developer, desktop or enterprise applications | Developer, full-stack | Developer, front-end | Engineer, data | Database administrator | Developer, game or graphics | Developer, mobile | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **9** | 4.0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| **11** | 5.0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| **12** | 6.0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| **16** | 2.0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | ... |
| **17** | 6.0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | ... |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **83430** | 21.0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | ... |
| **83432** | 0.0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| **83434** | 5.0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |

Now, we will divide the cleaned dataframe into training and testing data in the ratio 7:3. After splitting the dataset into a test and train data, we will be fitting a Linear Regression model to it.

The Line of Regression : -



The current model is only 68.6% accurate, which is a good enough model.

## Inferences from Stack Overflow Developer Survey 2021

1. India and USA respondents from around one-fourth of the total survey respondents. People from India and USA are also more likely to respond to the survey as the survey is available only in English. There are a total 180 countries in the survey responses.

Now, let's see some India specific stats :-

2. There is very less representation from the females and the other (non-binary, genderqueer and gender non-conforming) genders. The field is male dominated.

3. In India, there are very few non-binary gender people according to the survey and a lot of people who are not comfortable to share their gender (as compared to the world). Hence, the third gender concept is still quite alien in India and because people from other genders do not accept them easily.

4. We can safely assume that college students and job freshers and trainees visit Stack Overflow for solving their problems (or that young people are more likely to take surveys than middle aged and old people).

5. People in India had a later exposure to the internet and related stuff.

6. With the years of professional coding increasing, we can clearly see that people with more coding experience visit SO less frequently.

7. Few decades back, there weren't many women or non binary gender in the community. It used to be a male dominated community. But slowly the representation of the other genders are getting better, especially in the last 2 decades.

8. The unemployed people in India are majorly Bachelor degree holder or those who have completed secondary school. Even after a Masters degree, it seems people are still struggling to get into jobs.

9. It is better to go for a professional degree than a masters or associate degree if a person is looking for a job after the degree's completion.

10. As SO is mainly a forum for developer related stuff, we expected to see a high percentage of developers.

11. JavaScript, HTML/CSS and Python are the most worked with languages of 2020 in India and as well as worldwide. COBOL is the least worked with language worldwide, but across India, it is more used than languages like F#, Delphi, etc. JS, Python and HTML/CSS are again the languages people wanna work with.

12. Rust is the most loved language globally and also in India.

13. The most worked with database of the year 2020 is MySQL, both across India and Worldwide. The three out of the top four database suites are SQL-based. Only the second is a

NoSQL, i.e. MongoDB. Globally, PostgreSQL is the database people wanna work with the most. But, in India, people are looking forward to working with MongoDB, which is third on the list globally. NoSQL is gaining attention in India.

14. Globally and in India, people are loving Redis.

15. AWS is the most loved Cloud Platforms which is quite expected, IBM Cloud or Watson is the most dreaded Cloud platform, globally as well as across India.

16. Python is a general purpose language which suits the needs of the hobbyist, without any trade offs as it is a very powerful and versatile language.

17. JS, HTML/CSS, SQL form the top three languages used by professional developers in the year 2020. Python isn't used as extensively as we may have assumed having seen its popularity, still it makes it to the top five. Java is still in use and more than Python

18. Students learn or are taught Python, HTML/CSS, JS, C++, C, Java and SQl.

19. We can see that people prefer Blogs and Tutorials video to Coding Bootcamps and Certification Courses.

20. The Indian education system ,in terms of coding, is actually quite good. JavaScript lags behind a bit. A bit more attention to JS might help students just out of schools and colleges to get into good jobs.

21. Managerial positions (like Engineering Manager, Product Manager) require the most amount of professional coding experience, followed by Educators.

22. Developers prefer Linux based OS over Windows.

23. Most professional developers start coding at their teen. However, there are also professional developers who started their coding journey at an older age. Therefore, coding is something that you can start learning even in your late 20s and still build a good career out of it.

24. Googling the problem and visiting stack overflow are the most widely used methods of finding a solution when they are stuck.

25. Most of the respondents are well off having no issues of mental health.

## The Way Forward

- Web Frameworks, Collaboration tools, etc can be analysed for Web Frameworks loved, Collaboration tools Loved, etc.
- The Language of the Hobbists, Professional Developers and Students can be repeated on Cloud Platforms, Web Frameworks, Collaboration Tools, DBMS, etc, to get more results.
- Compensation can be compared for every country and every profession.

## References

https://stackoverflow.com/

https://medium.com/analytics-vidhya/

https://jovian.ai/

https://pandas.pydata.org/

https://matplotlib.org/

https://seaborn.pydata.org/