EDA Project

INT – 353

Submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology
IN
Computer Science and Engineering

School of Computer Science & Engineering
Lovely Professional University, Jalandhar

Done By :-
Aryan Jain
(12019711)

Section :-
K20MP

Candidate Declaration and Certificate

We hereby certify that the work, which is being presented in this project report entitled, **Finding out the insights from any dataset according to your choice**, in partial fulfillment of the requirements for the degree of **Bachelor of Technology**, submitted in the **Computer Science and Engineering** , Lovely Professional University, Jalandhar, Punjab; by **Aryan Jain(12019711**) is the authentic record of our own work carried out under the supervision of **Mr. Abhijeet Dutta**, Computer Science and Engineering, Lovely Professional University, Punjab.

We further declare that the matter embodied in this project report has not been submitted by us for the award of any other degree

INDEX PAGE

Introduction

We have to choose any dataset according to our choice and take out various insights from the dataset after performing various tasks like data cleaning and handling, univariate analysis, bivariate analysis and multivariate analysis.

Dataset Chosen

I choose Stack Overflow Annual Developer Survey Dataset 2022

Domain Of The Dataset

Stack Overflow

About The Domain

Stack overflow is the very popular website which connects the developers together and, on this website, different developers/students from different countries post their problems and different developers/students post the answers of the problem and resolve their issues. Basically, it's a question-and-answer website which helps the developers and students a lot.

This website releases the survey form for developers every year and collects the information of different developers which helps to the owner of the website to find best developers to hire and these survey form also helps website to improve.

About The Dataset

This dataset contains about 56 columns which represents the questions asked from the user and 73269 rows which represents the user responded to the survey.

Data Cleaning

Original dataset contains various columns which are of no use so, I not modify my original data but I copy the original dataset into the new dataset because data is very costly so I can not simply remove the columns and make changes in the dataset and other benefit to copy the dataset in the new dataset that If by mistake I loose any information from the new dataset than I can retrieve this information from original dataset.

Columns in My New Dataset

- MainBranch :- Which of the following options best describes you today? Here, by "developer" we mean "someone who writes code."
- Employment :- Which of the following best describes your current employment status?
- RemoteWork :- Which best describes your current work situation?
- EdLevel :- Which of the following best describes the highest level of formal education that you've completed?
- LearnCode :- How did you learn to code? Select all that apply.
- YearsCode :- Including any education, how many years have you been coding in total?
- YearsCodePro :- NOT including education, how many years have you coded professionally (as a part of your work)?

- DevType :- Which of the following describes your current job? Please select all that apply.
- BuyNewTool :- When buying a new tool or software, how do you discover and research available solutions? Select all that apply.
- Country :- Where do you live?
- Currency :- Which currency do you use day-to-day? If your answer is complicated, please pick the one you're most comfortable estimating in.
- LanguageHaveWorkedWith :- Which programming, scripting, and markup languages have you done extensive development work in over the past year.
- Database :- Which database environments have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the database and want to continue to do so, please check both boxes in that row.)
- Platform :- Which cloud platforms have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the platform and want to continue to do so, please check both boxes in that row.)
- Webframe :- Which web frameworks and web technologies have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the framework and want to continue to do so, please check both boxes in that row.)
- MiscTech :- Which other frameworks and libraries have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the framework and want to continue to do so, please check both boxes in that row.)
- ToolsTech :- Which developer tools have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the technology and want to continue to do so, please check both boxes in that row.)
- NEWCollabTools :- Which development environments did you use regularly over the past year, and which do you want to work with over the next year? Please check all that apply.
- OpSys :- What is the primary operating system in which you work?
- VCInteraction :- How do you interact with your version control system? Select all that apply.
- OfficeStackAsync :- Which collaborative work management tools did you use regularly over the past year, and which do you want to work with over the next year? Select all that apply
- OfficeStackSync :- Which communication tools did you use regularly over the past year, and which do you want to work with over the next year? Select all that apply
- Blockchain :- How favorable are you about blockchain, crypto, and decentralization?
- NEWSOSites :- Which of the following Stack Overflow sites have you visited? Select all that apply.
- SOVisitFreq :- How frequently would you say you visit Stack Overflow?
- SOAccount :- Do you have a Stack Overflow account?
- SOPartFreq :- How frequently would you say you participate in Q & A on Stack Overflow? By participate we mean ask, answer, vote for, or comment on questions.
- SOComm :- Do you consider yourself a member of the Stack Overflow community?
- Age :- What is your age?
- Gender :- Which of the following describe you, if any? Please check all that apply.
- Tbranch :- Would you like to participate in the Professional Developer Series?
- ICorPM :- Are you an independent contributor or people manager?
- WorkExp :- How many years of working experience do you have?
- TimeSearching :- On an average day, how much time do you typically spend searching for answers or solutions to problems you encounter at work? (This includes time spent searching on your own, asking a colleague, and waiting for a response).
- TimeAnswering :- On an average day, how much time do you typically spend answering questions you get asked at work?
- SurveyLength :- How do you feel about the length of the survey this year?
- SurveyEase :- How easy or difficult was this survey to complete?

Approach to visualize the dataset

My Dataset contains various columns and these columns may have some correlation b/w them and there may be very null values. So, I move column by column according to my insights and handle the data and visualize it.

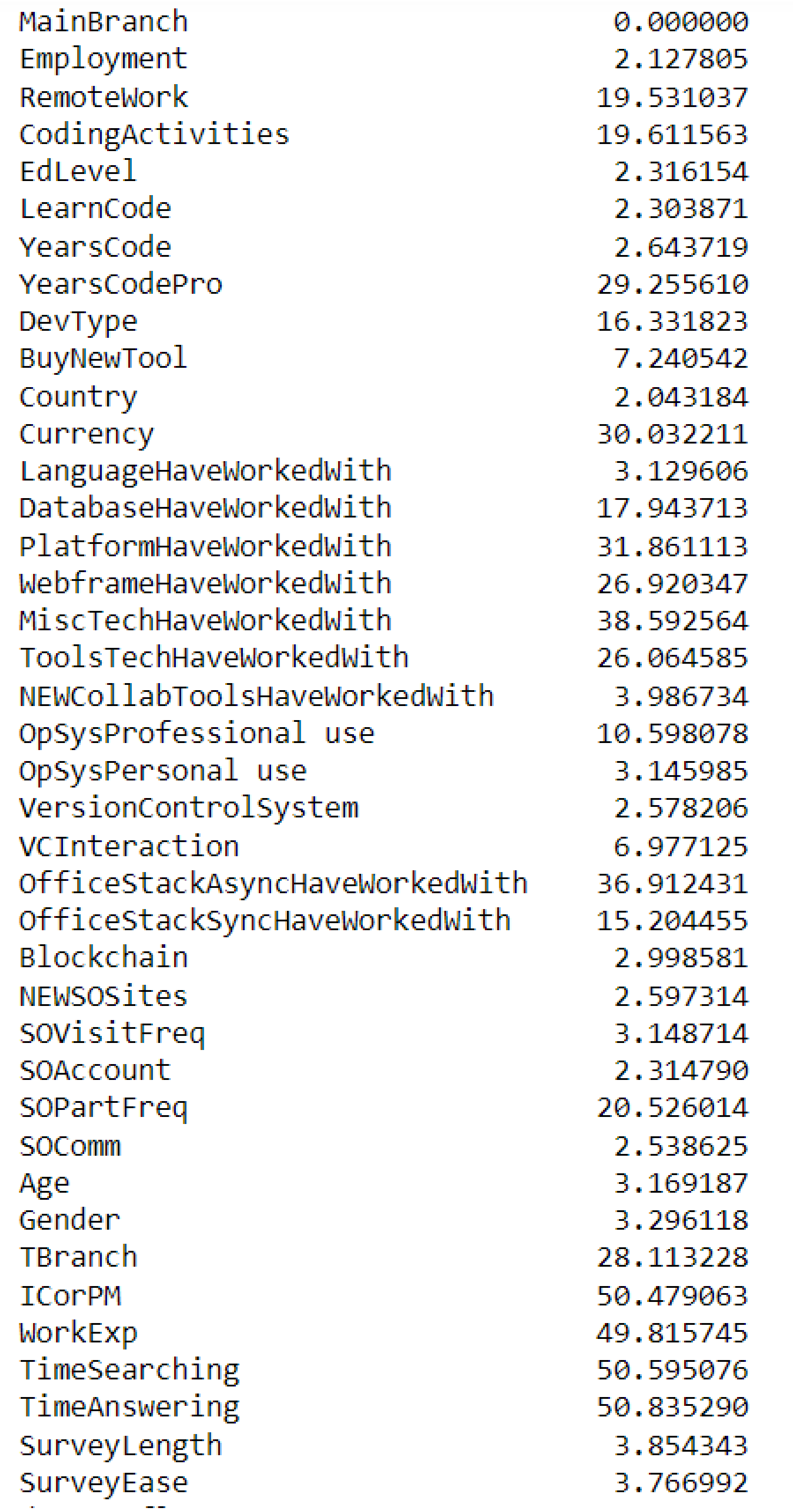
Analysis with insights

1. Common type of survey questions used by these type of websites

   Survey Questions depends on domain so according to this domain we visualize the null values in columns the columns have higher null values should not ask.

   The analysis done below contains various columns which have null values more than 20 percent but these questions are compulsory in this domain and these null values also depends on main branch.

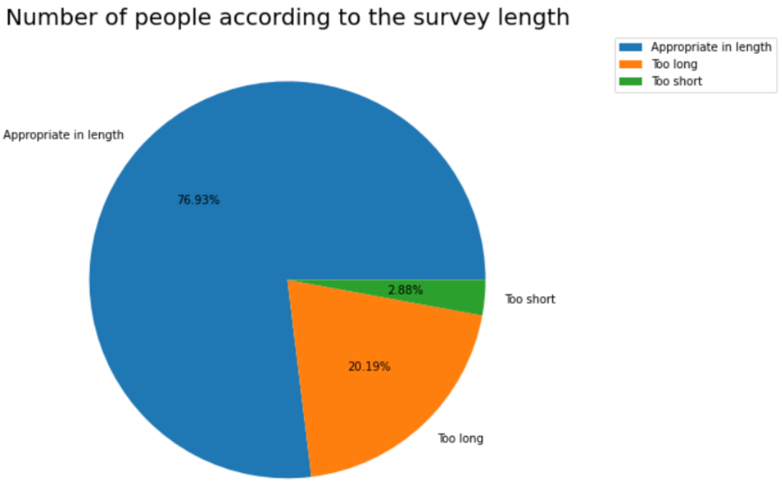
   So, the below questions are enough to ask and visualize.

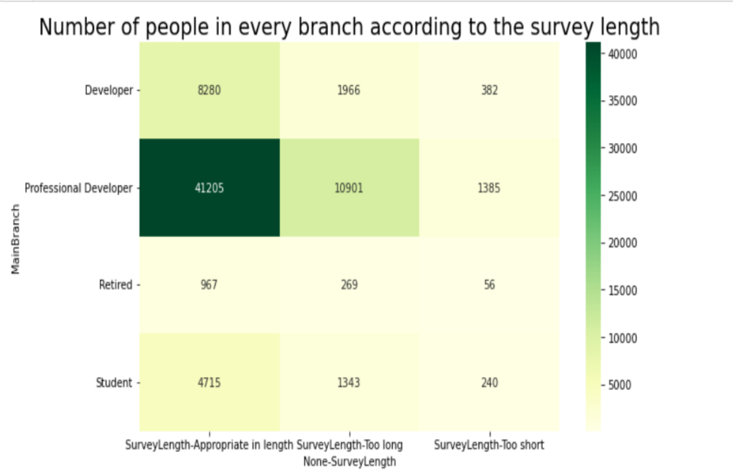


```
MainBranch                           0.000000
Employment                           2.127805
RemoteWork                          19.531037
CodingActivities                    19.611563
EdLevel                              2.316154
LearnCode                            2.303871
YearsCode                            2.643719
YearsCodePro                        29.255610
DevType                             16.331823
BuyNewTool                           7.240542
Country                              2.043184
Currency                            30.032211
LanguageHaveWorkedWith               3.129606
DatabaseHaveWorkedWith              17.943713
PlatformHaveWorkedWith              31.861113
WebframeHaveWorkedWith              26.920347
MiscTechHaveWorkedWith              38.592564
ToolsTechHaveWorkedWith             26.064585
NEWCollabToolsHaveWorkedWith         3.986734
OpSysProfessional use               10.598078
OpSysPersonal use                    3.145985
VersionControlSystem                 2.578206
VCInteraction                        6.977125
OfficeStackAsyncHaveWorkedWith      36.912431
OfficeStackSyncHaveWorkedWith       15.204455
Blockchain                           2.998581
NEWSOSites                           2.597314
SOVisitFreq                          3.148714
SOAccount                            2.314790
SOPartFreq                          20.526014
SOComm                               2.538625
Age                                  3.169187
Gender                               3.296118
TBranch                             28.113228
ICorPM                              50.479063
WorkExp                             49.815745
TimeSearching                       50.595076
TimeAnswering                       50.835290
SurveyLength                         3.854343
SurveyEase                           3.766992
```

2. What should be the survey length

My survey length column contains approx. 3.85 percent null values so I replace these null values with mode after comparison with main branch column



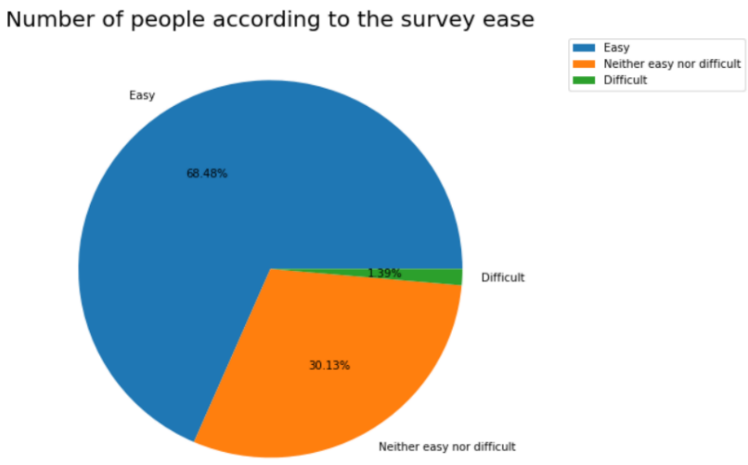


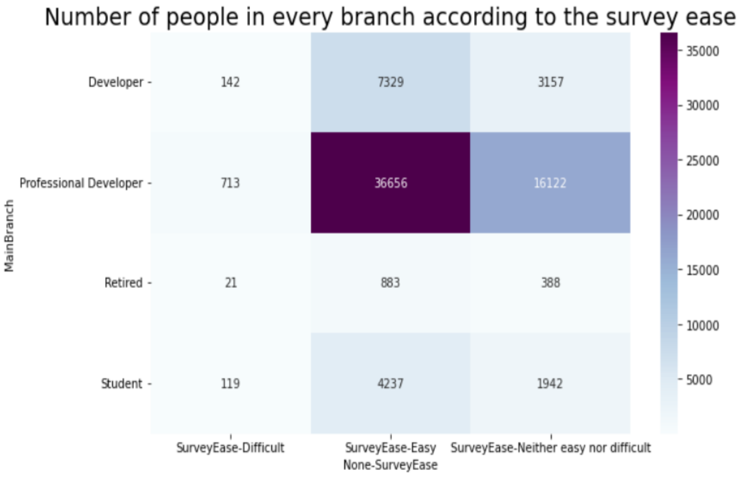
Above Analysis shows that there are more number of people who said that survey length is ok.

So, this survey length is enough.

3. Difficulty level of the survey form

This column contains approx. 3.76 percent null values so I can replace these null values with mode after comparison with every branch



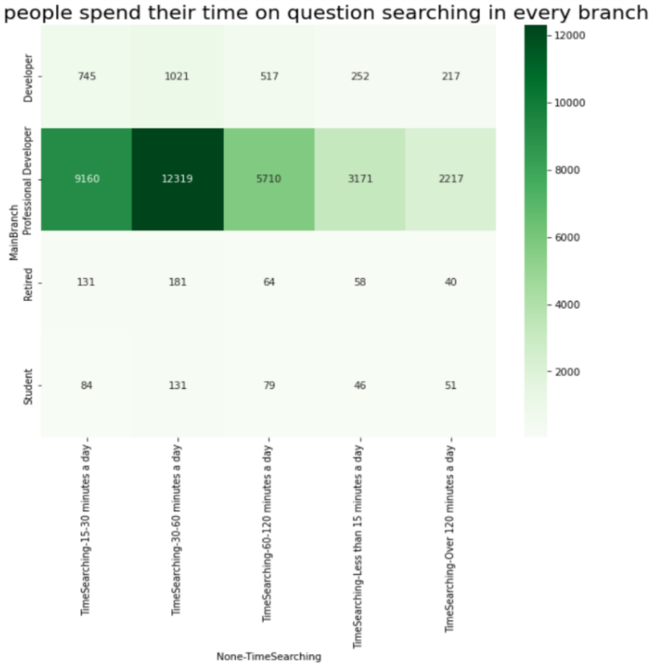


Analysis shows that survey is easy according to the every branch.

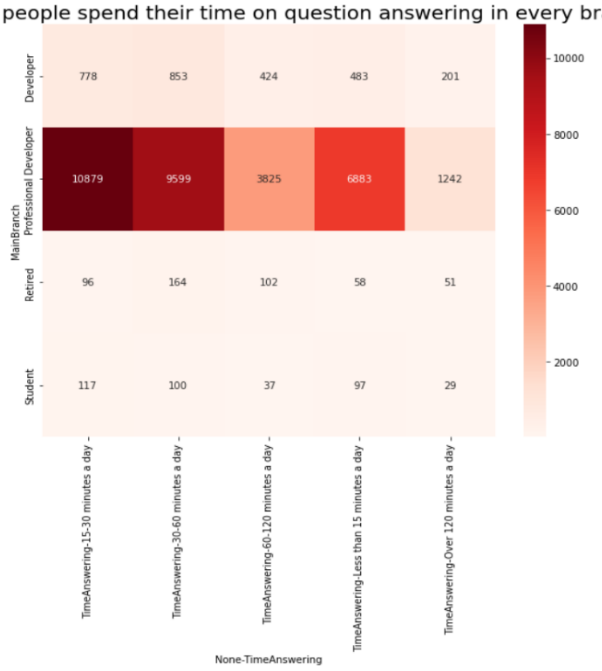
4. Average time to answer/search the question

These columns contains approx 50 percent null values but I can not replace and remove because I need this column for our visualization and it gives us exact information.

Number of people spend their time on question searching in every branch



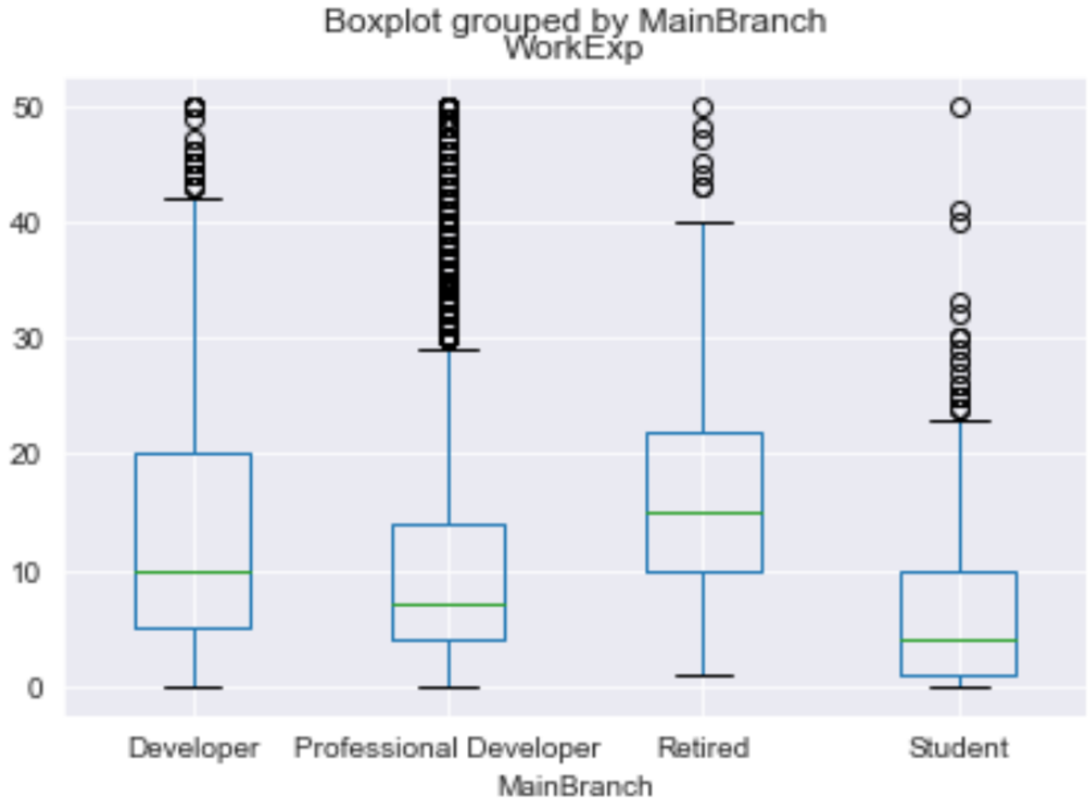
Number of people spend their time on question answering in every branch



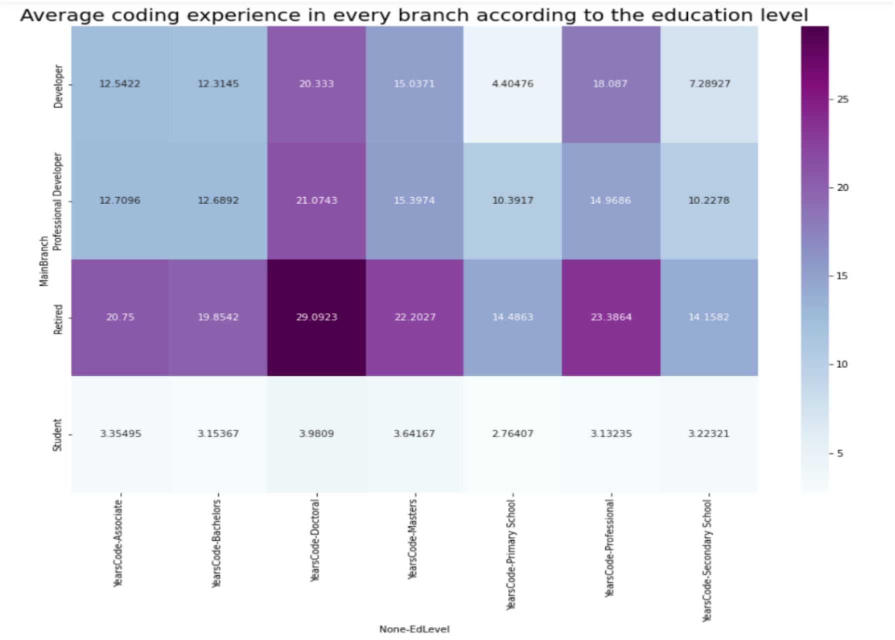
Above analysis shows that there are professional developers who contribute more toward answering and searching the questions they spend 30-60 minutes a day for searching the questions and spend 15-30 minutes a day for answering questions.

5. Average work experience according to the level of the knowledge in every type of field

Work Experience column contains about 50 percent null values and also it contains various outliers which is not good for our visualisation if I visualise this column then I will not get accurate information.



So, I will calculate experience based on YearsCode column

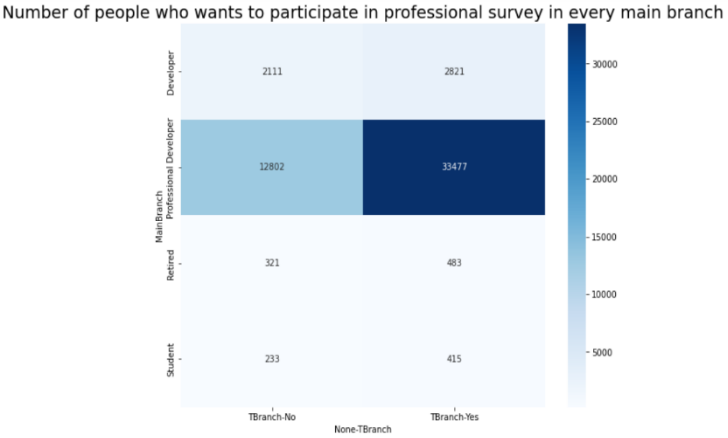


Analysis shows that in the student branch who done masters and doctoral degree have highest experience, in the branch of retired who done doctoral degree have highest experience, in the branch of professional developer who done doctoral degree and masters degree have highest experience and in the branch of developers who done doctoral degree and are professionals have highest experience.

Overall in every branch who did doctoral or masters or who are professionals have highest experience.

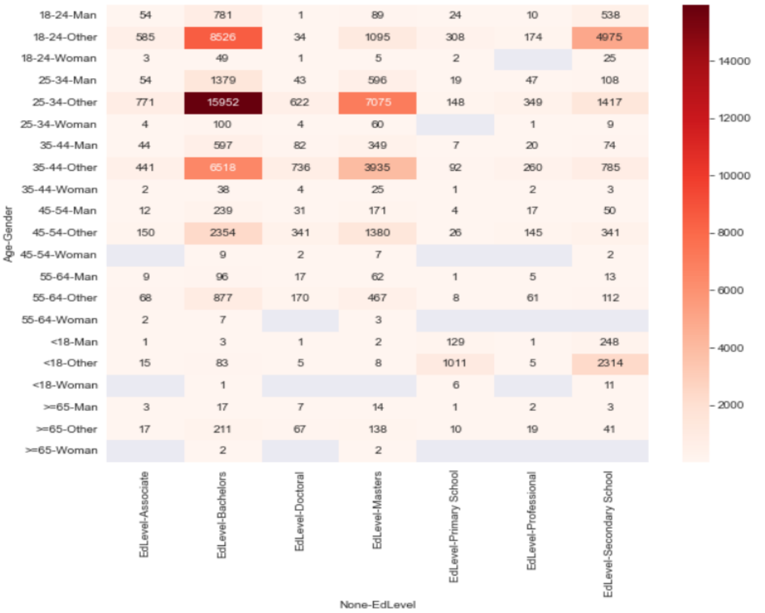
6. Number of people wants to become professional developers

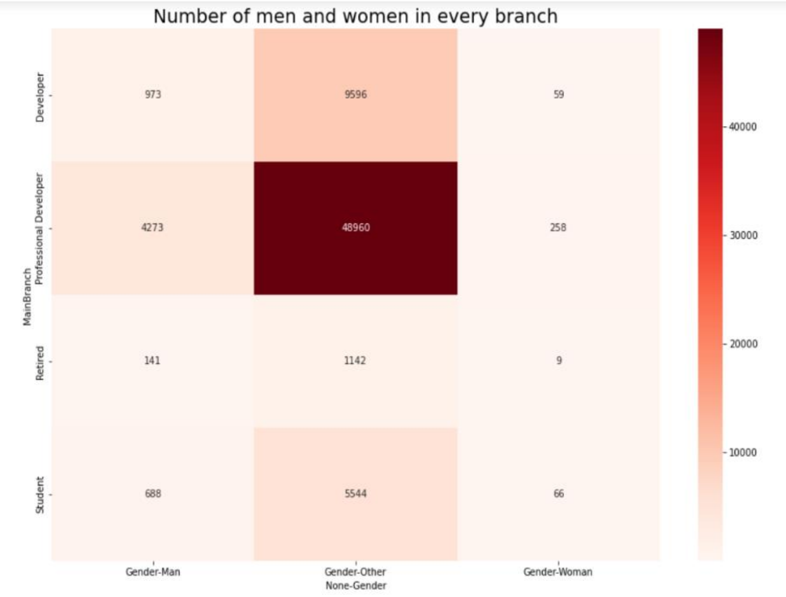
For this insight we visualize TBranch column which contains about 20 percent null values but I can not replace or remove because I need exact information



This analysis shows that there are students , developers and professional developers who wants to participate in professional series means that they wants to become professional developers.

7. Level of Education based on their ages and gender

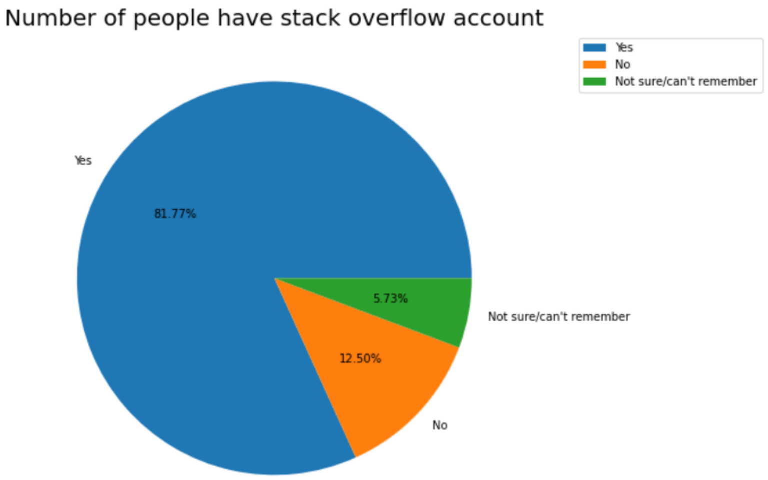


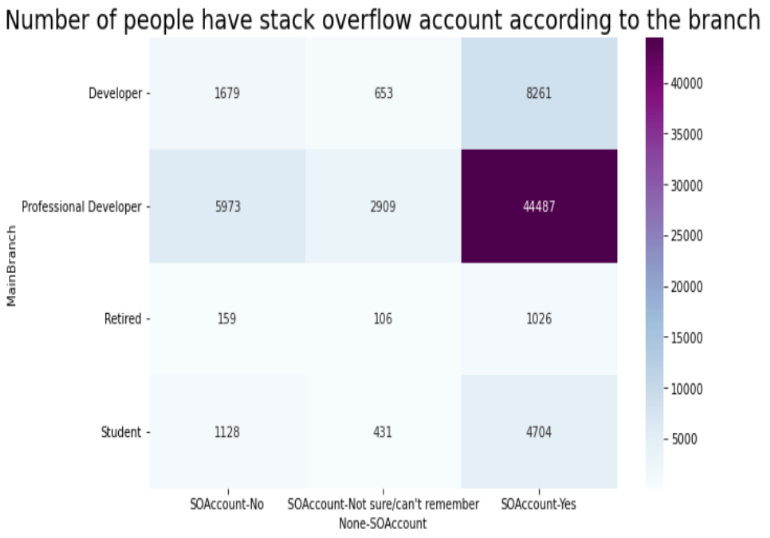


Based on the above analysis people who not specify their gender or belongs to other gender have highest qualification in the age range 18-45.

8. Number of people use stack overflow

People who have stack overflow account means that they are using stack overflow

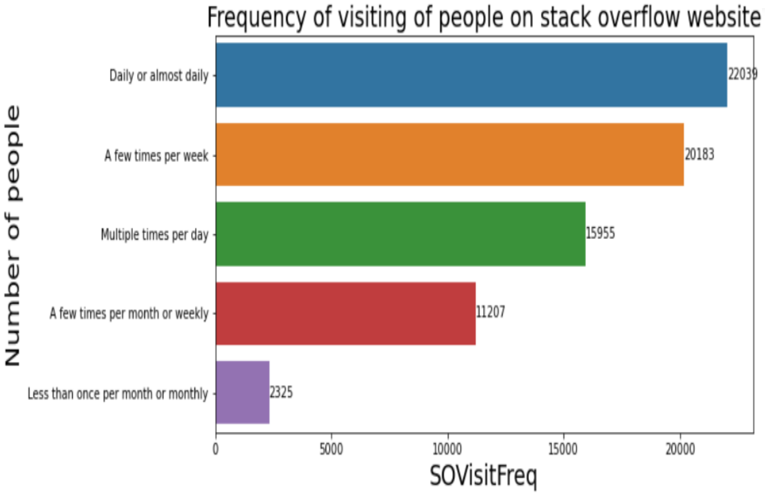


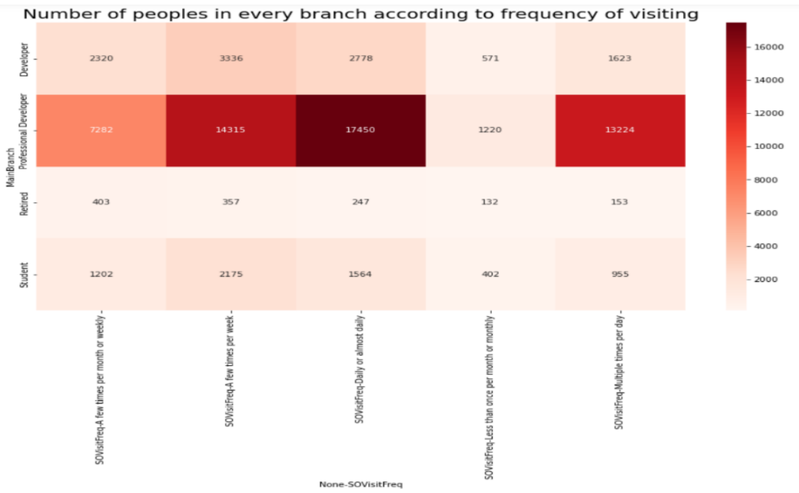


In every branch huge number of people use stack overflow

9. Number of people use stack overflow regularly

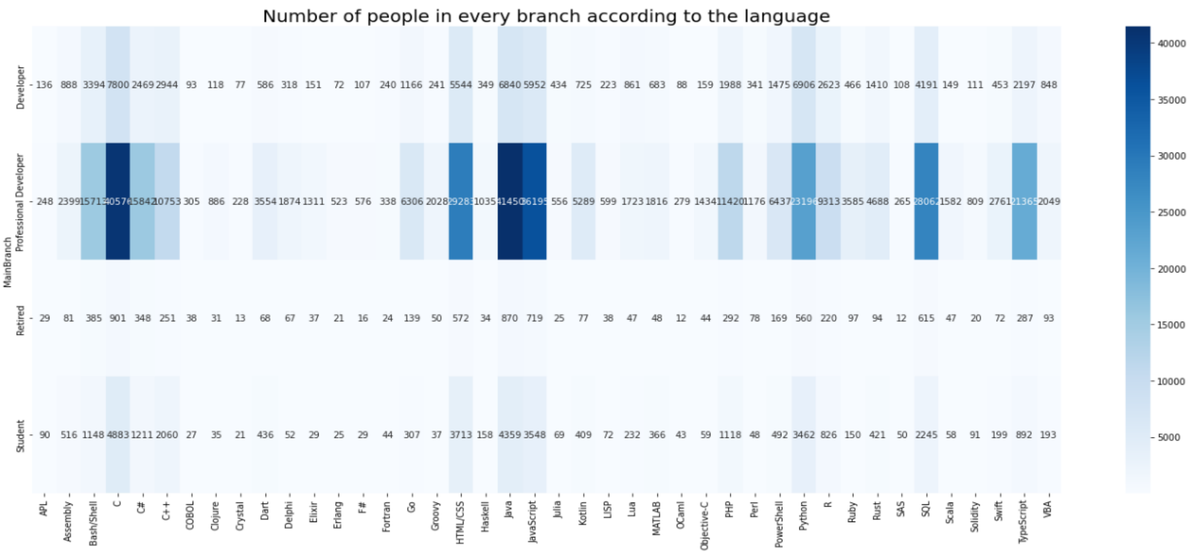
For this insight I have to visualise SOVisitFreq column which contains approx. 3.15 percent null values I replace these null values with mode after comparing with main branch





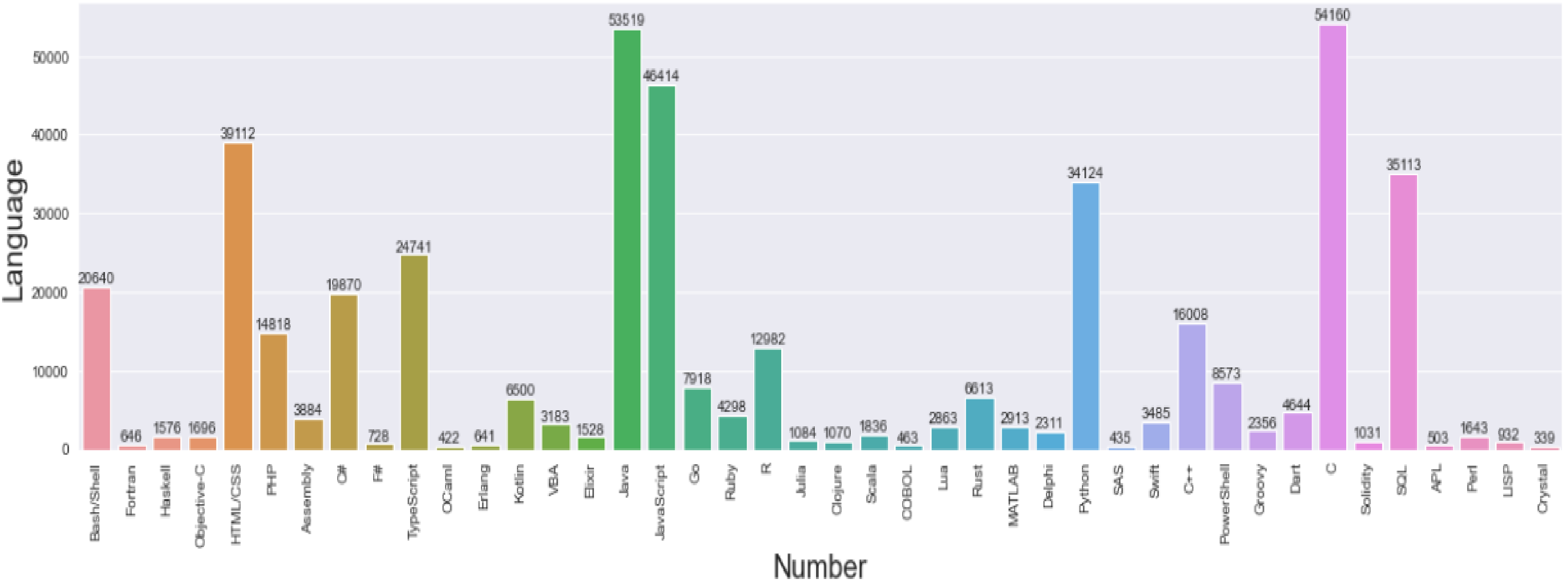
In every branch mostly people use almost daily or few times per week.

10. Which type of coding language users mostly used



Students use C language, Retired person also use C language , professional developers use Java , C , Javascript language , developers use C , Java , Python , Javascript language.

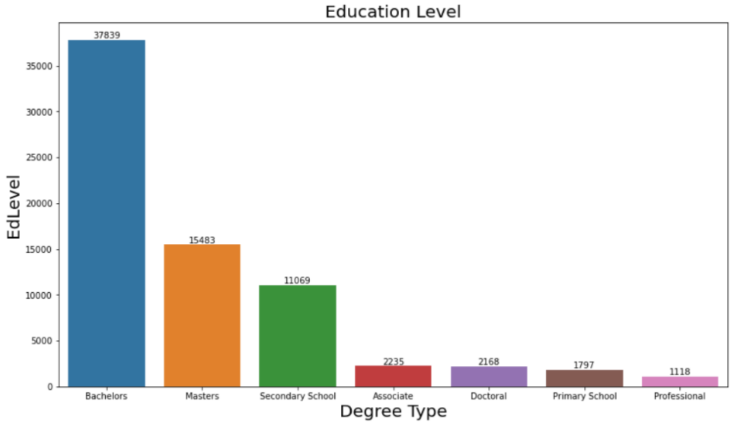
11. Popular coding language

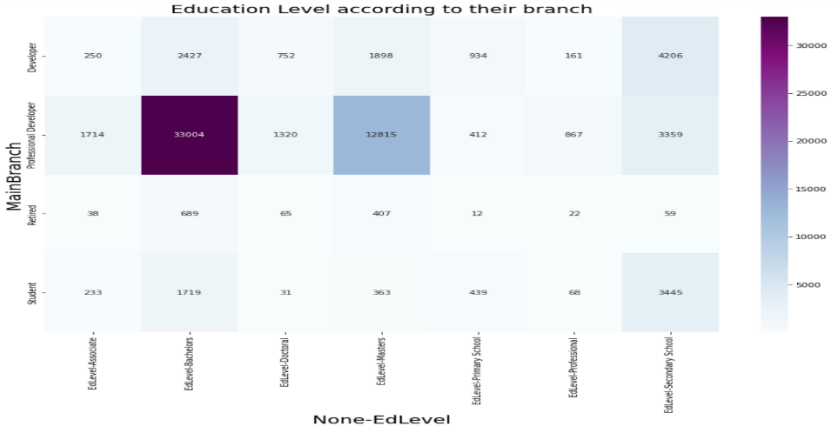


C , Java , Javascript , SQL , Python , HTML/CSS are the popular languages

12. Number of people belongs to different type of education level

I need to visualise EdLevel column for this which contains null values and I replace these null values with mode after comparing with main branch

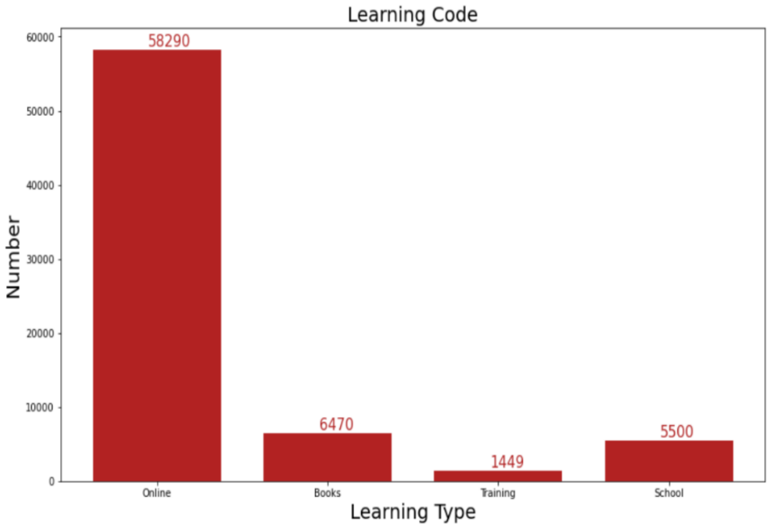


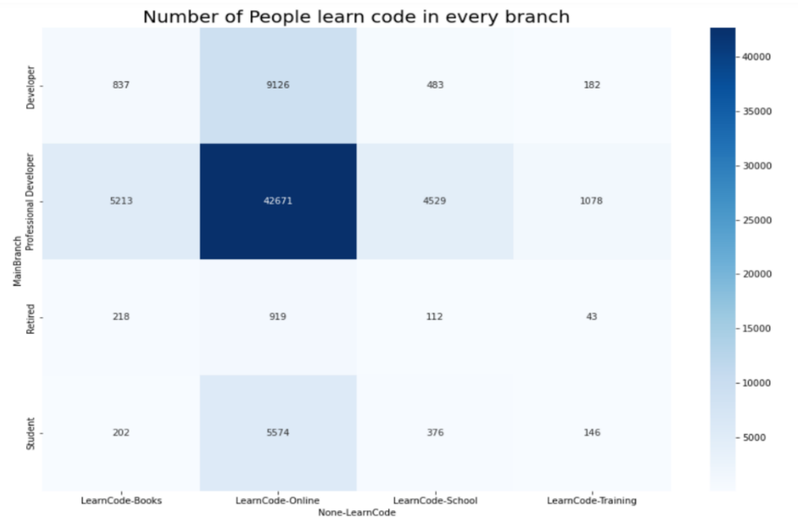


	EdLevel-Associate	EdLevel-Bachelors	EdLevel-Doctoral	EdLevel-Masters	EdLevel-Primary School	EdLevel-Professional	EdLevel-Secondary School
Developer	250	2427	752	1898	934	161	4206
Professional Developer	1714	33004	1320	12815	412	867	3359
Retired	38	689	65	407	12	22	59
Student	233	1719	31	363	439	68	3445

Most of the people did bachelors degree , professional developers did bachelors and masters degree , students did secondary school , developers also did secondary school and bachelors degree.

13. Number of people learn online

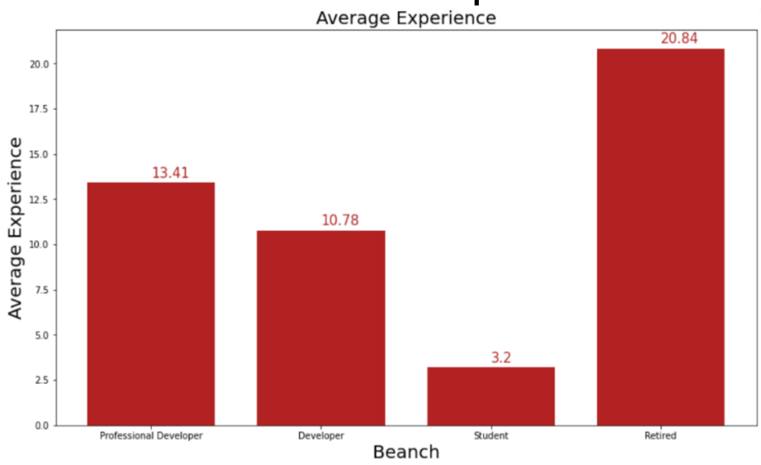


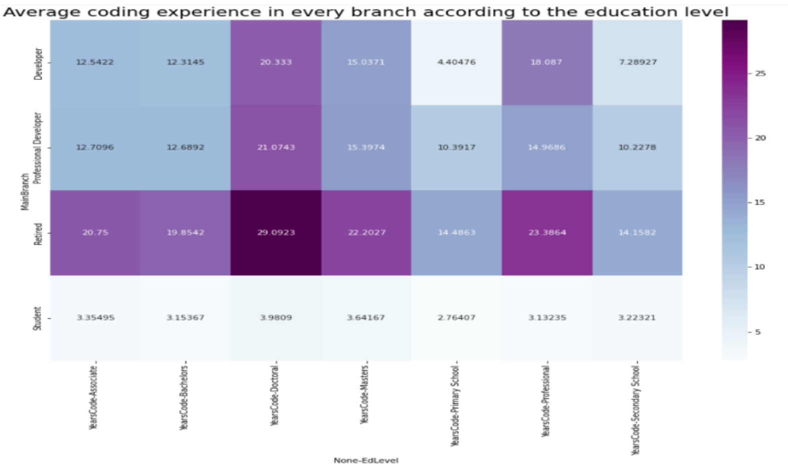


Huge number of people in every branch learn online. This information is very helpful for the company they can launch different types of online courses after providing free trial which increases the sales of the company.

14. Finding best developer in every branch according to the education level

Those persons who have highest experience are the best developers





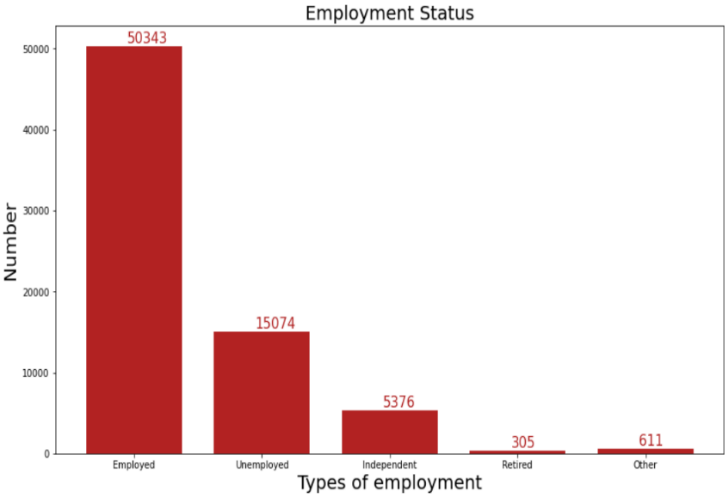
There are retired and professional developers who have highest coding experience.

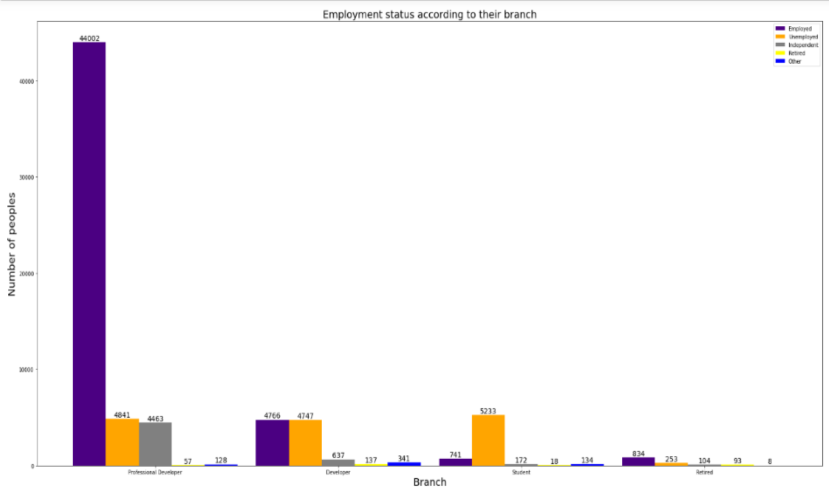
In every branch who did doctoral, masters, professional and bachelors degree have highest experience.

This information is very useful for the company now company will contact to the retired, professional developers or developers for the hiring according to education level.

In terms of education level doctoral, masters or bachelors degree or professionals have highest experience

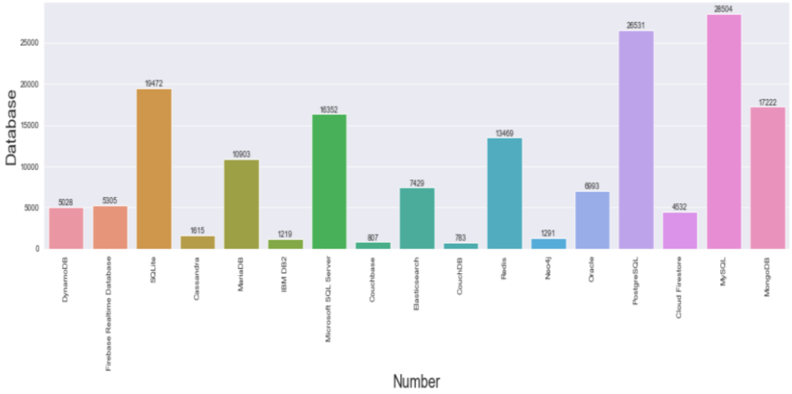
15. Finding best developer who are currently not doing job

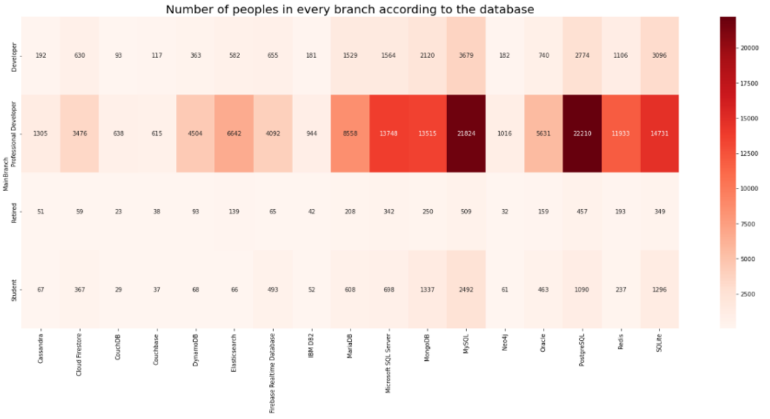




Most of the persons are employed. If we see unemployed people, then most developers and students are not employed and according to the above best developer visualisation company will connect to student and developers for hiring.

16. Types of Database mostly used

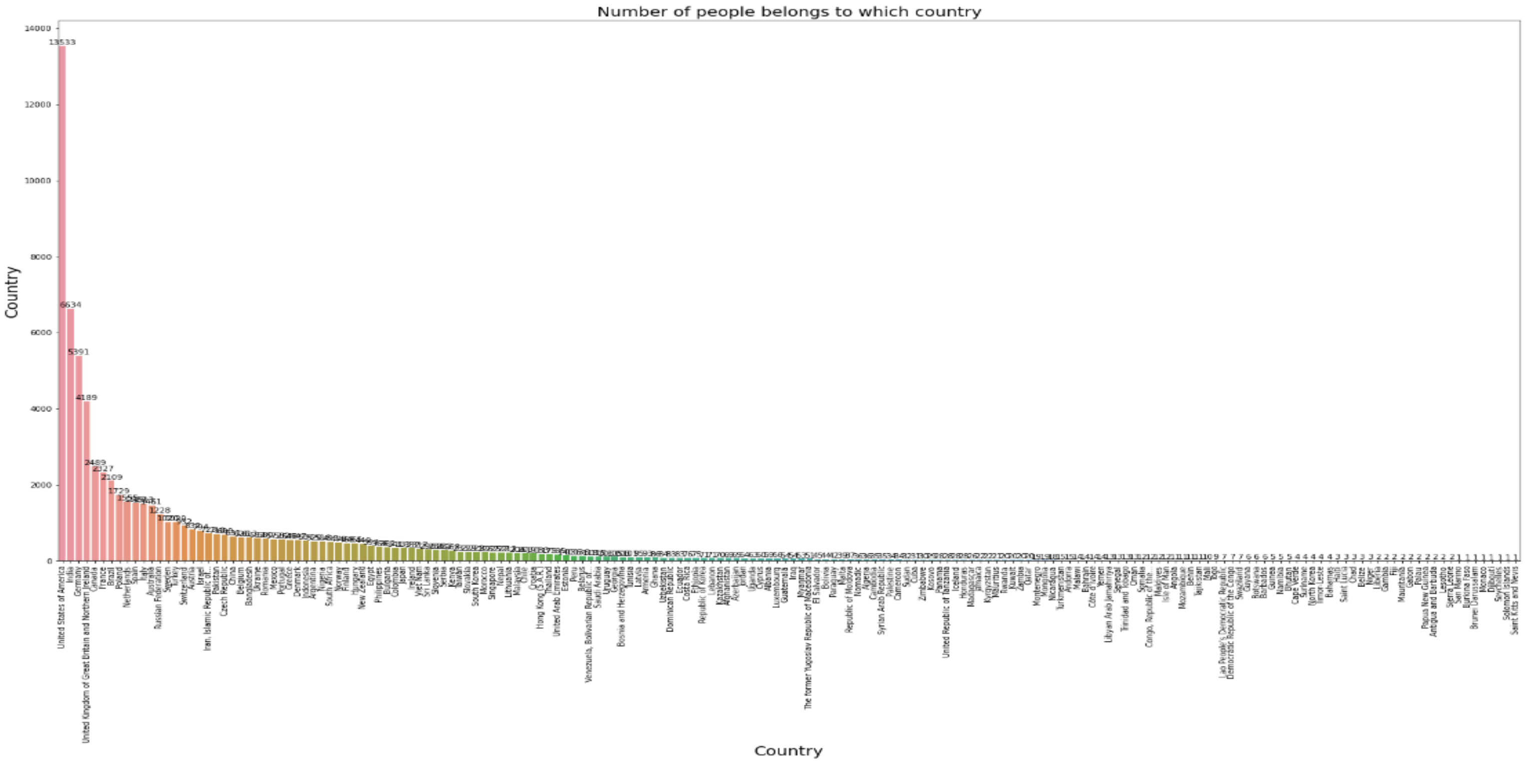




MySQL, PostgreSQL, MongoDB, SQLite, Microsoft SQL Server are most used.

Mostly students use MySQL, MongoDB, SQLite , PostgreSQL , professional developers use MySQL, PostgreSQL and developers use MySQL and SQLite.

17. Number of developers belongs to which country



Developers belong to United States of America, India, Germany, United Kingdoms and Canada.

This information is very helpful for the company because company can launch their trials software first to these countries because developers use them and review their products.

18. Maximum number of developers belongs to which country

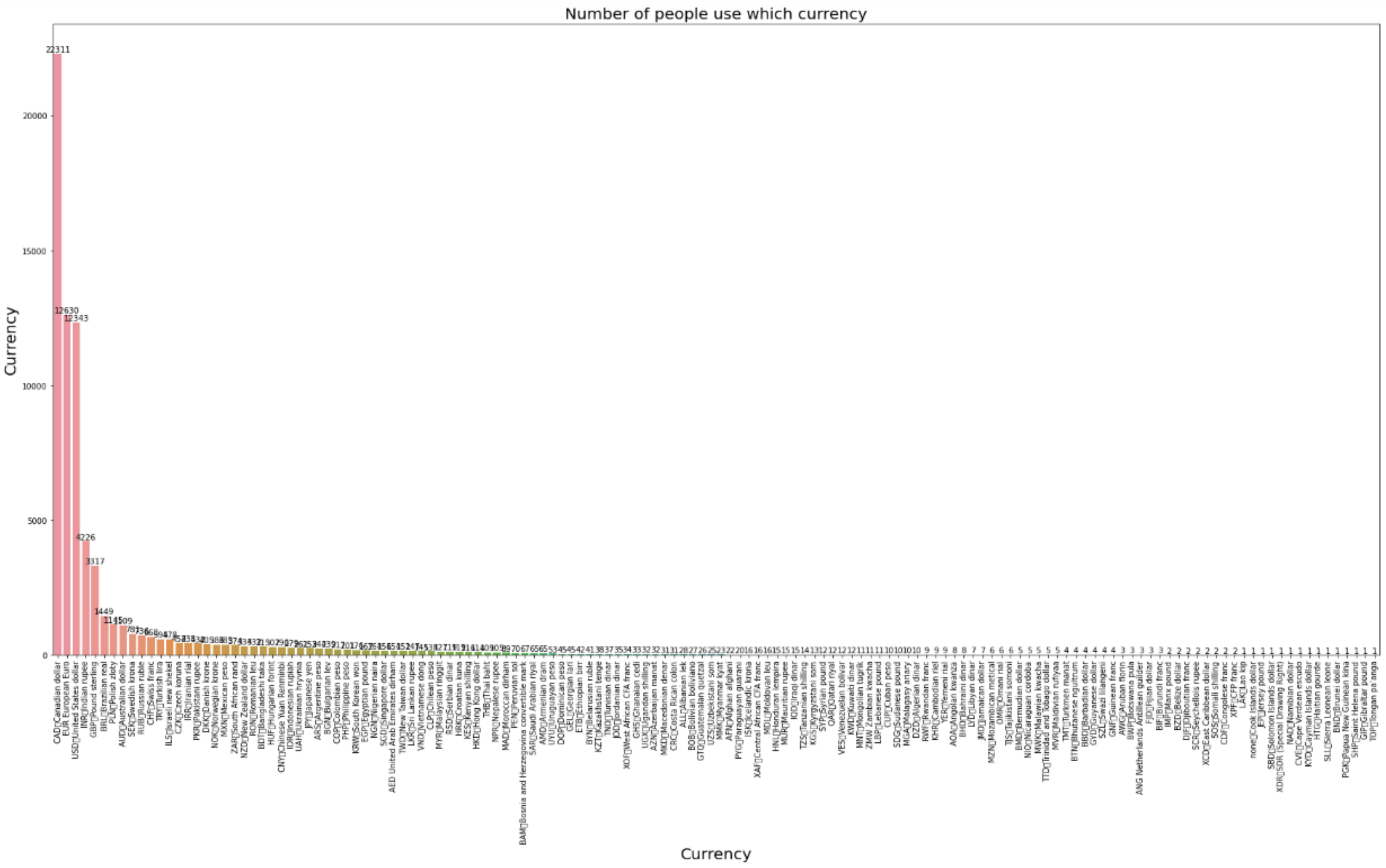
From the above analysis maximum number of developers belongs to United States of America and India.

Company will launch their trial software in USA and India and more developers will use these products and company will collect the data which is helpful for the company to modify their products which will increase the sales of the company.

19. What type of currency people used

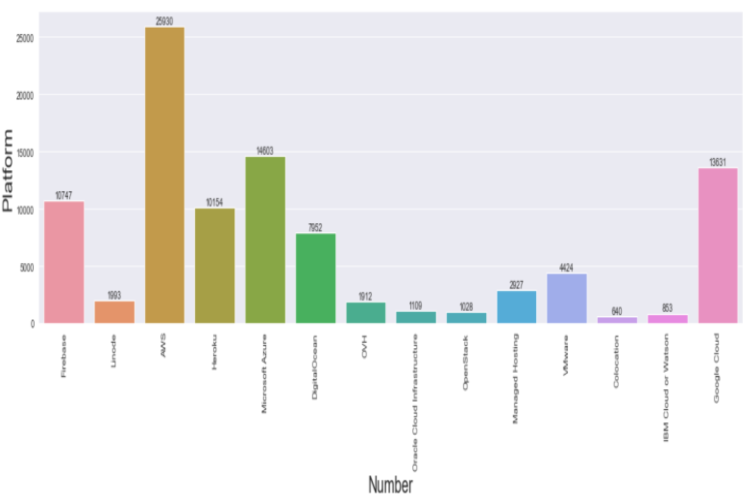
Basically, currency depends on the country in which you are living but, this website take payment in online mode and in online mode people can pay with any currency.

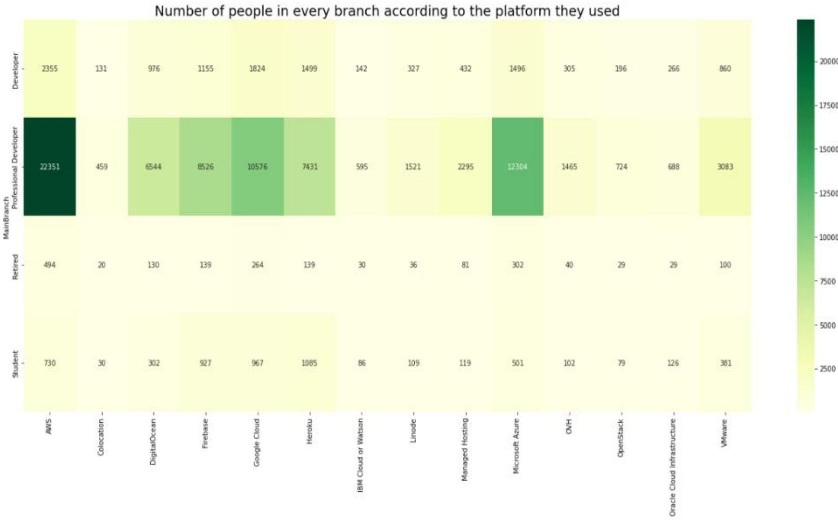
This information is very helpful for the company because now company can add new payment method to their website which is very helpful for the users and also the company sales will increase.



According to the above analysis, most of the people use Canadian dollar, Euro and United States dollar for their day-to-day life transaction.

20. What type of cloud platform mostly used



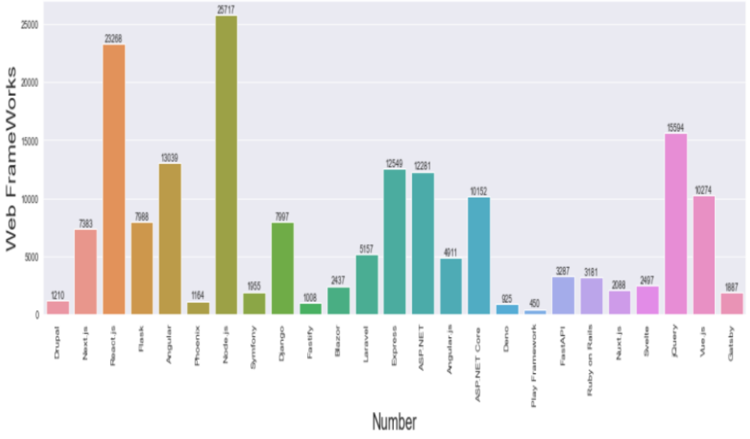


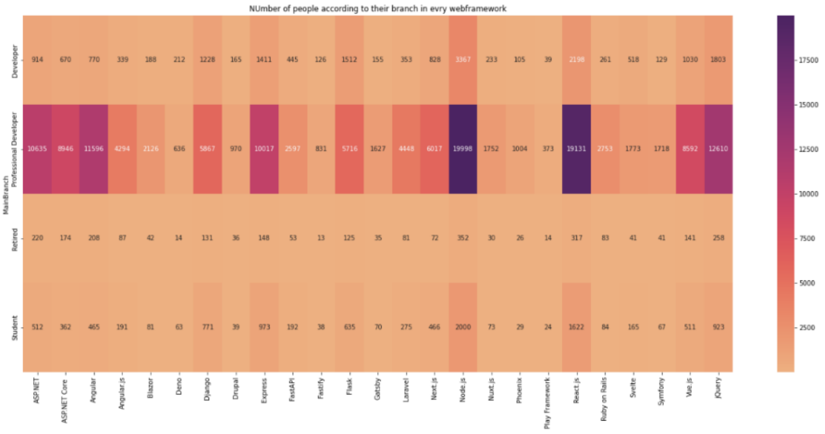
According to above visualisation AWS, Azure and google cloud are mostly used.

And in every branch huge number of the people use AWS , Azure , Google cloud.

And most of the students use Heroku.

21. What type of web framework mostly used



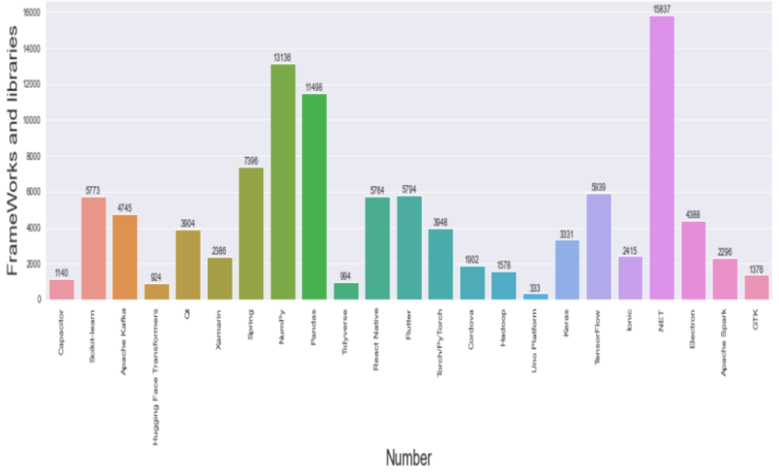


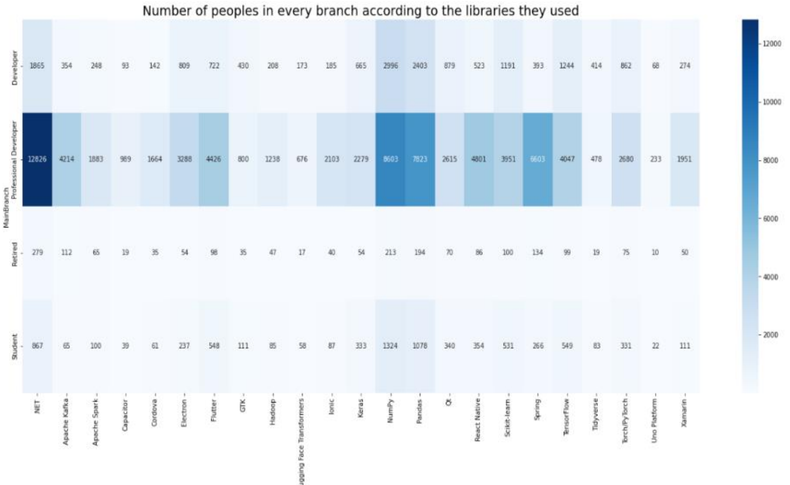
Above analysis shows that Node.js and React.js are majorly used webframeworks.

In every branch Node.js mostly used.

Professional developers use Node.js , React.js , Angular , JQuery , ASP.NET webframeworks.

22. Which other frameworks and libraries used





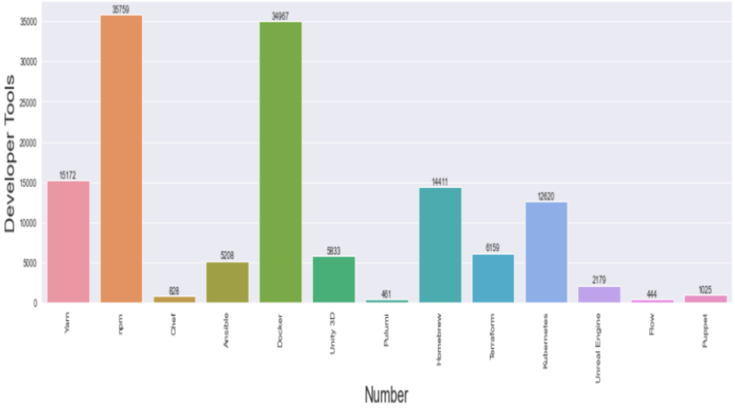
Above analysis shows that NET, NumPy and Pandas are mostly used.

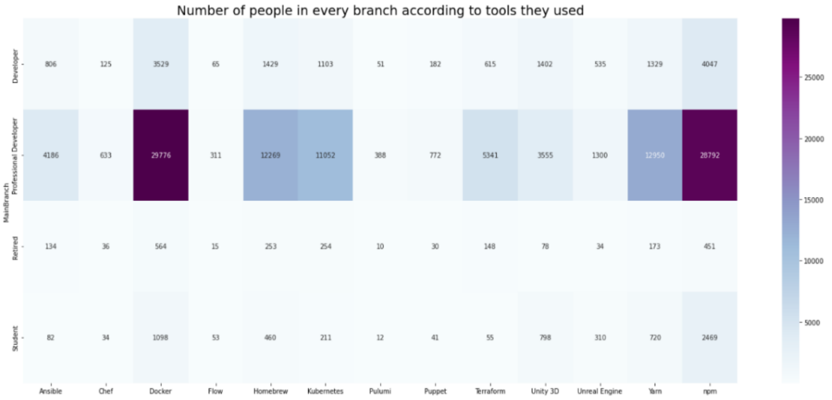
All branches mostly used NET, NumPy and Pandas.

Students mostly used NumPy and Pandas

Professional Developers mostly use NET, NumPy, Pandas, Spring and Flutter.

23. What type of developer tools are use



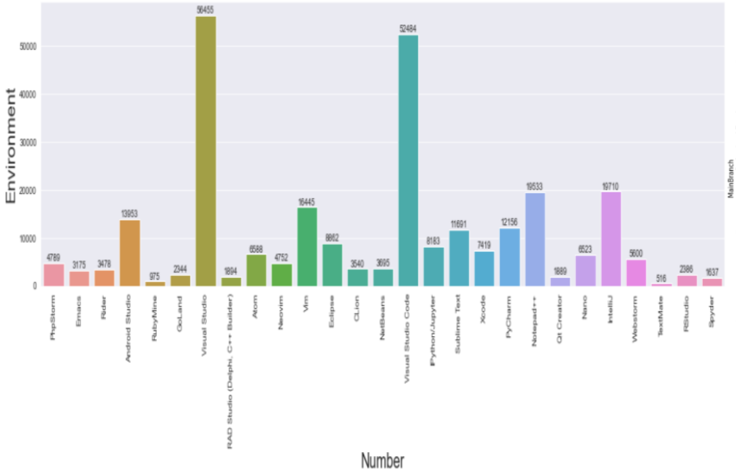


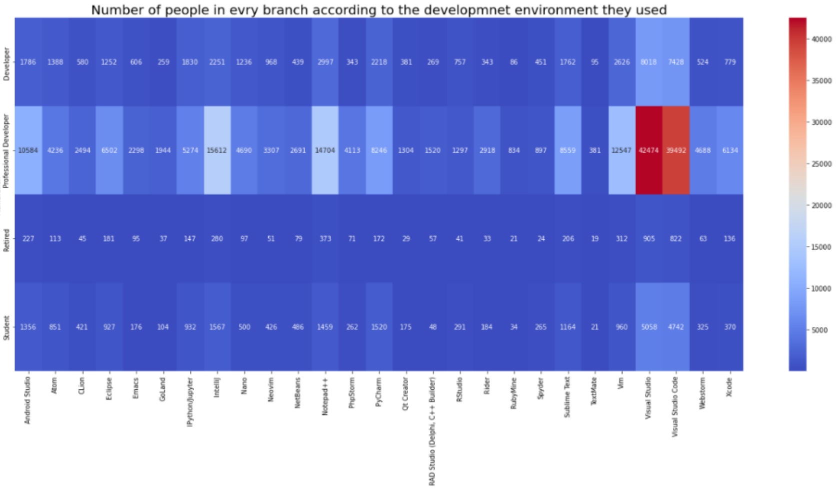
Above analysis shows that npm and Docker mostly used

All branches mostly use npm and Docker.

Professional developers mostly use npm, Docker, Homebrew, Kubernetes and Yam.

24. Which type of development environment use

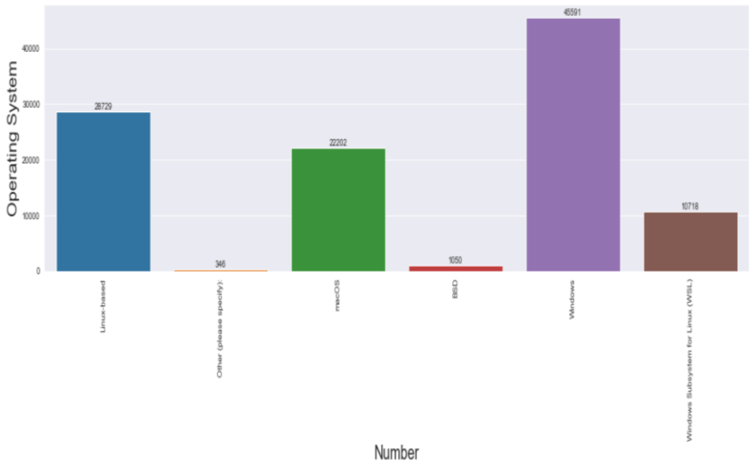


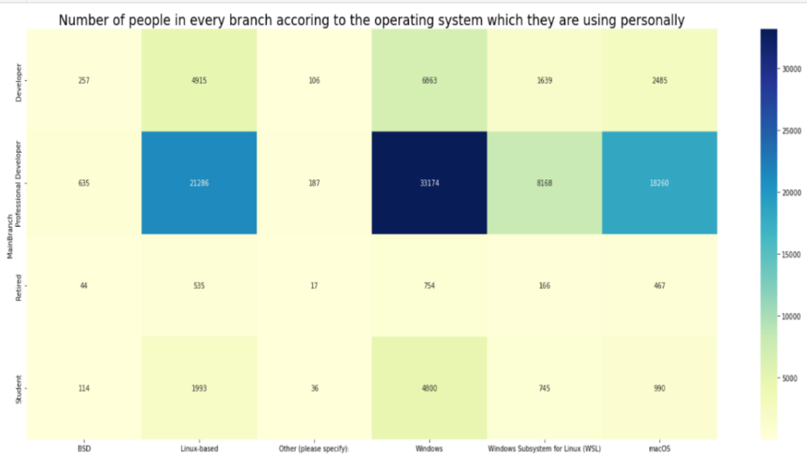


Above analysis shows that Visual Studio and Visual Studio Code mostly use.

In all branches Visual Studio and Visual Studio Code mostly use.

25. Which operating system mostly use





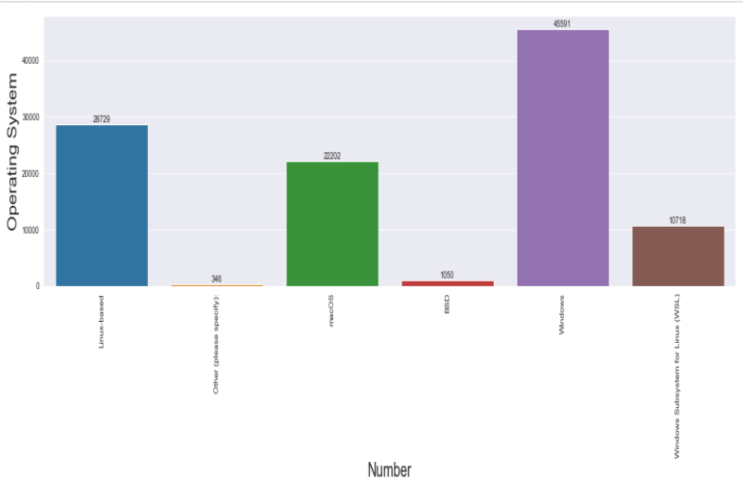
Above analysis shows that Windows is the primary operating system which users use

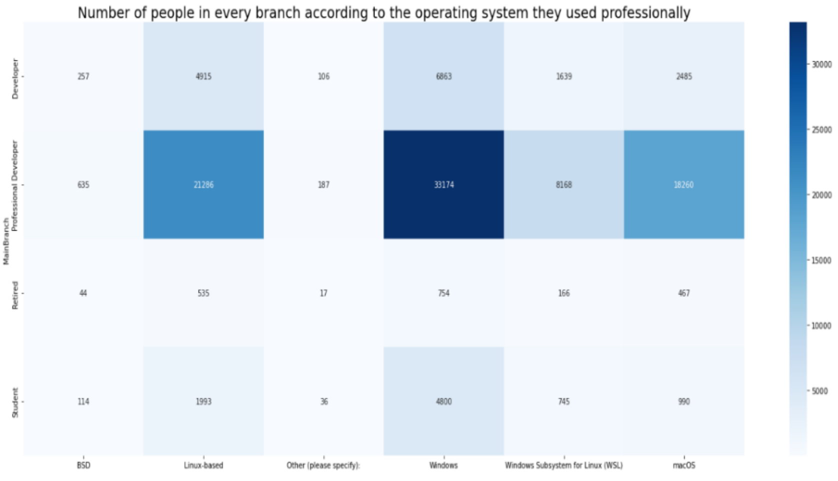
All branches mostly use Windows operating system.

Professional Developers mostly use Windows, Linux, MacOS.

Develops mostly use Windows and Linux.

26. Operating System mostly use professionally





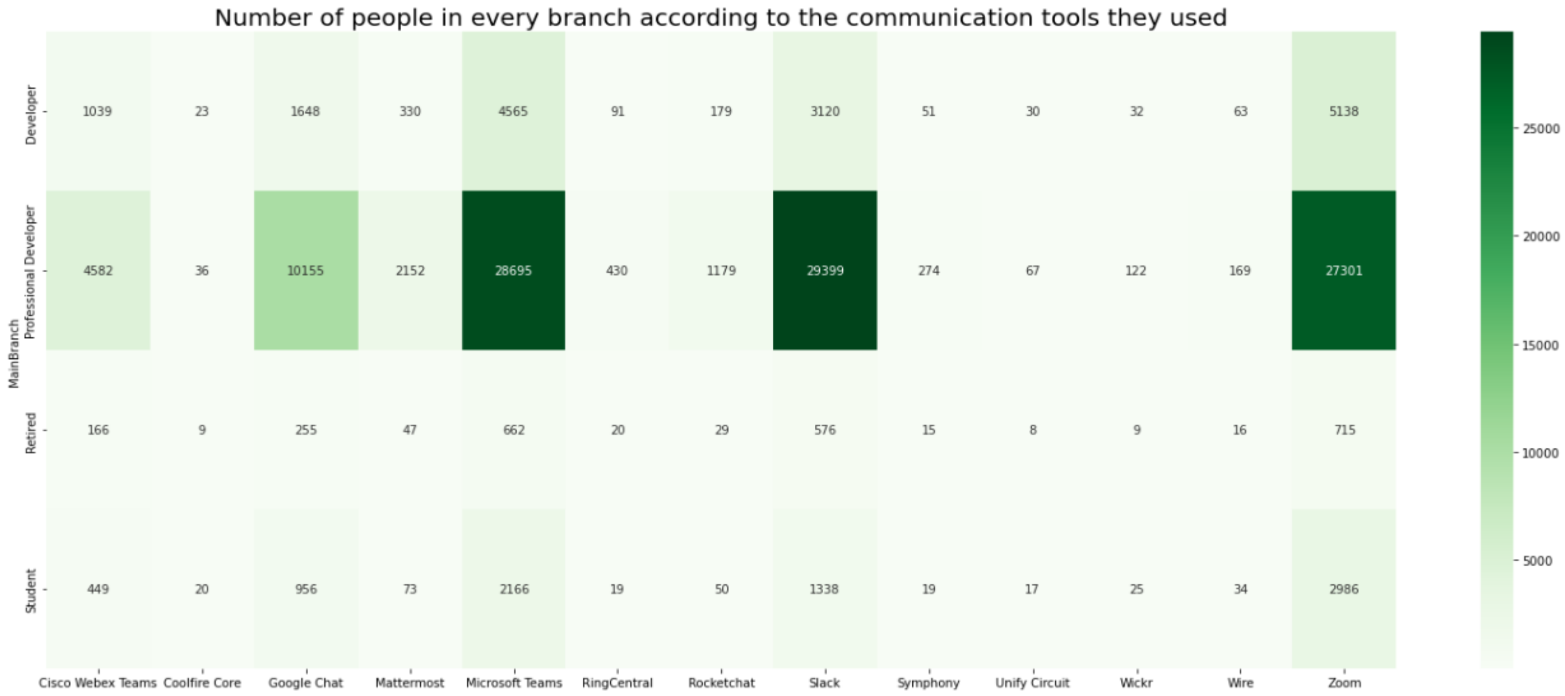
Above analysis shows that Windows is the primary operating system which users use

All branches mostly use Windows operating system.

Professional Developers mostly use Windows, Linux, MacOS.

Develops mostly use Windows and Linux.

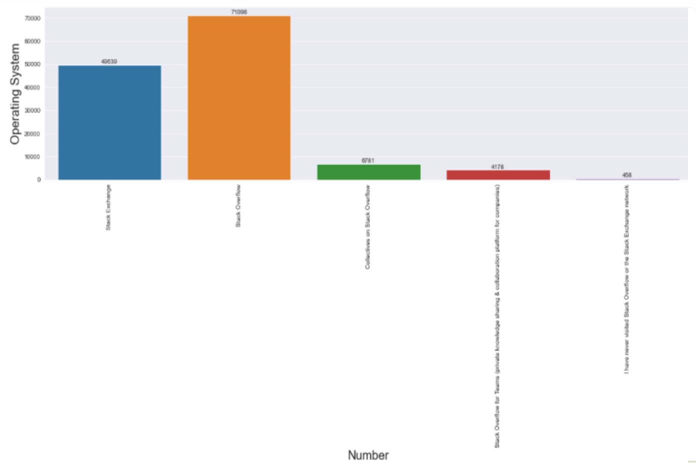
27. What type of communication mostly use

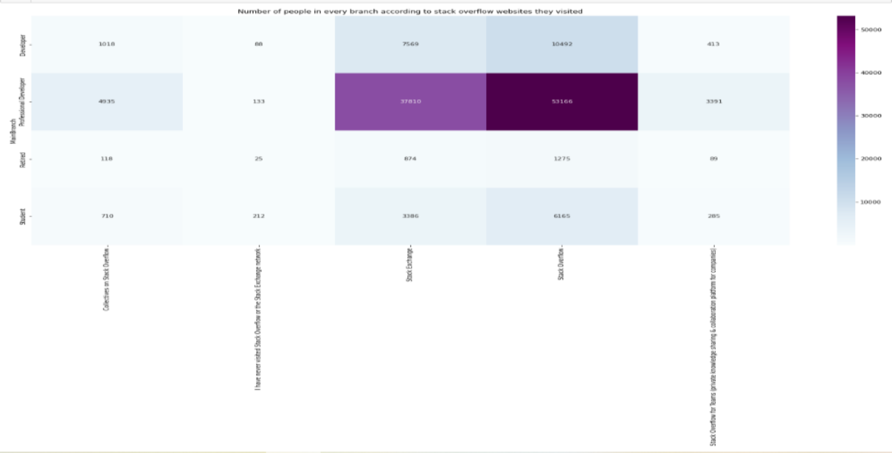


Above analysis shows that in every branch huge number of people use slack, Microsoft Teams, Zoom.

Professional developers are also use Google Chat.

28. Websites frequently visits



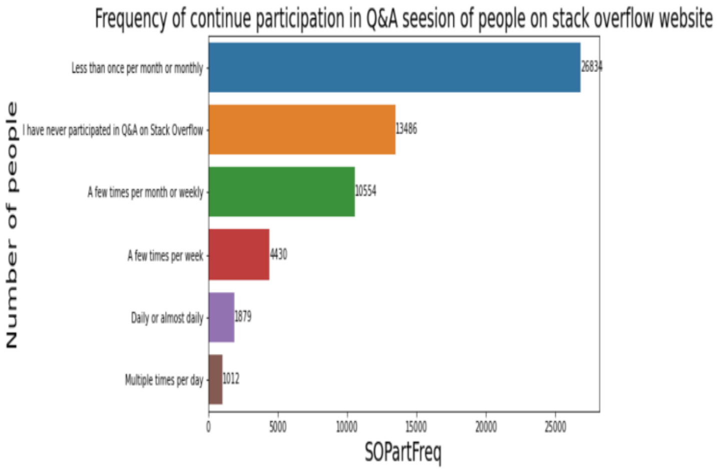


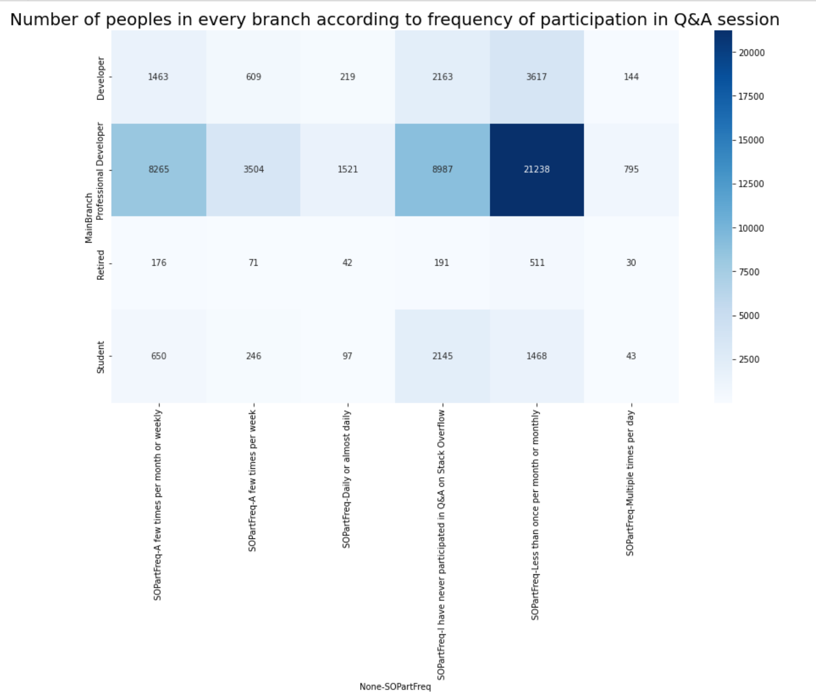
Above analysis shows that in all branch people visit Stack Overflow and Stack Exchange Websites.

Professional Developers also visit collectives on Stack Overflow.

This information is very useful for the company because based on this information company will work on websites on which users visit.

29. Number of people frequently participate in Q&A sessions



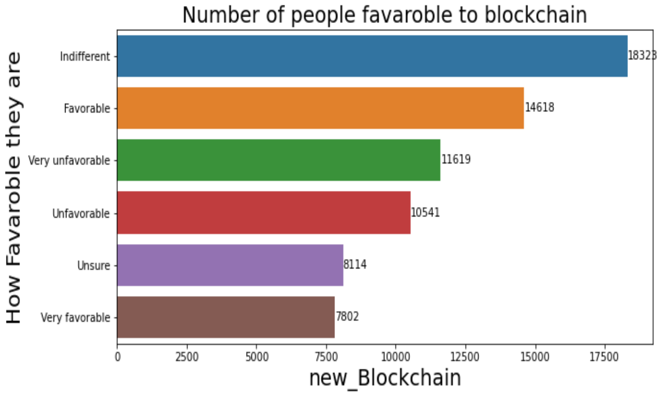


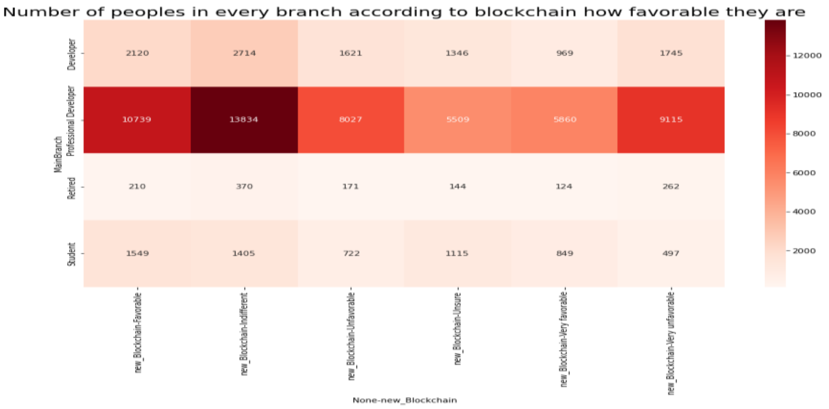
Above analysis shows that in all branches people participate in Q&A sessions are less than once per month or monthly.

Most of the students never participated in Q&A session.

Most of the professional developers participate in Q&A session a few times per month or weekly.

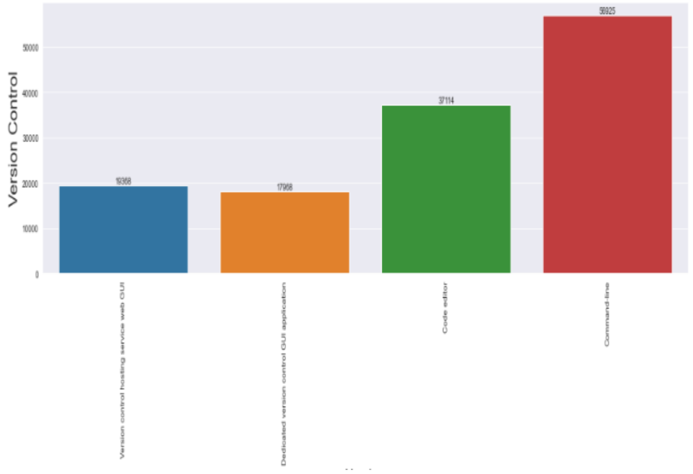
30. How users are favourable about blockchain

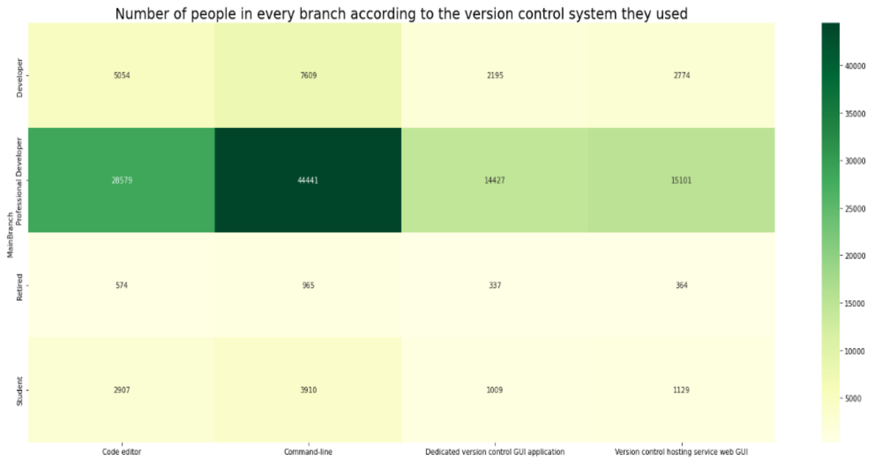




Analysis shows that in all branches more number of people are indifferent about blockchain while students are favourable about blockchain and some of the professional developers are favourable about blockchain.

31. Which version control system mostly use

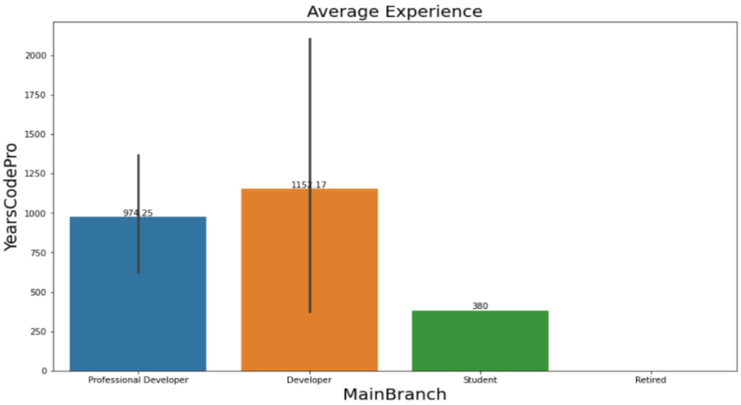




Above analysis shows that in every branch people use command line.

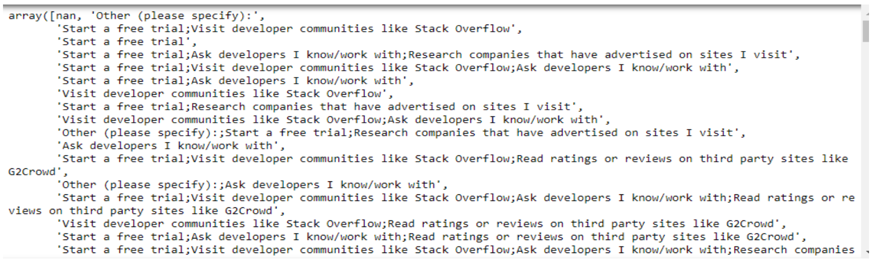
Professional developer also use code editor.

32. Professional coding experience



Developers have the highest professional coding experience.

33. How many users buy new tools



```
array([nan, 'Other (please specify):',
       'Start a free trial;Visit developer communities like Stack Overflow',
       'Start a free trial',
       'Start a free trial;Ask developers I know/work with;Research companies that have advertised on sites I visit',
       'Start a free trial;Visit developer communities like Stack Overflow;Ask developers I know/work with',
       'Start a free trial;Ask developers I know/work with',
       'Visit developer communities like Stack Overflow',
       'Start a free trial;Research companies that have advertised on sites I visit',
       'Visit developer communities like Stack Overflow;Ask developers I know/work with',
       'Other (please specify):;Start a free trial;Research companies that have advertised on sites I visit',
       'Ask developers I know/work with',
       'Start a free trial;Visit developer communities like Stack Overflow;Read ratings or reviews on third party sites like
G2Crowd',
       'Other (please specify):;Ask developers I know/work with',
       'Start a free trial;Visit developer communities like Stack Overflow;Ask developers I know/work with;Read ratings or re
views on third party sites like G2Crowd',
       'Visit developer communities like Stack Overflow;Read ratings or reviews on third party sites like G2Crowd',
       'Start a free trial;Ask developers I know/work with;Read ratings or reviews on third party sites like G2Crowd',
       'Start a free trial;Visit developer communities like Stack Overflow;Ask developers I know/work with;Research companies
```

This shows that every user start a free trial or use the tool that provides by company.

Summary

The summary of my whole project is every people learn online code and professional developer have the highest experience in every field and professional developer also knows the many technologies that’s why company prefer professional developers with more experience.

Bachelor’s degree is in the trend most of the people did bachelors degree and get job after completion of bachelor's degree.

I also realize that who have highest coding experience they also know many technologies whether they belong to student, developer, retired or professional developers’ category.

Nobody buy any kind of developer tools all are using free trials or use the tools which provides by company in which they work.

Survey form is very important for every company because they collect the thoughts and data of the people who filled survey and with this data, they improve their technologies and increase their sales.

One more important thing is that while analysing your data you must be careful because data is very important, we should look every possibility before removing and replacing the null values.

References

Dataset Kaggle Link: -

https://www.kaggle.com/datasets/uzairrehman/stackoverflowannualdevelopersurvey2022?resource=download

To Download Dataset: -

https://drive.google.com/file/d/1h0ONBV8zd5Xf77eqVv5HPSwhVkUusX2e/view

Seaborn Reference Link: -

https://seaborn.pydata.org/

GitHub Link; -

https://github.com/aryanjain1908/Stack_Overflow_Survey_Analysis