A

Industrial Project Report

(Project Semester Jan-June 2021)

"(EDA- Stack Overflow Developer Survey 2020)"

Submitted by

(Yashraj Singh Rawat)

Student ID: 19MCAL004

GIP Providers Pvt. Ltd.

Under the Guidance of

Faculty Internship Guide:

Name: Mr. Harshit Sharma

Designation: Assistant Professor

Industry Guide:

Name: Mr. Rajat Goyal

Designation: Director

Department of Information Technology and Computer Application

JECRC UNIVERSITY, JAIPUR

June 2021

Preface

The present report is the outcome of the Internship Program of **Jaipur Engineering College and Research Center, JECRC University.** The objective of the internship was to familiarize the student with the implementation of the knowledge he earned on the campus and apply it on real world applications. The practical knowledge is far different from the bookish knowledge that a student achieves in an institution. The major problem that I faced during my internship was that there were not sufficient free API's to extract data, as almost all API's require some credit to provide their data.
The report focuses on the scrapping the data and reviews of 'Grras Solutions' and performing analysis on the data extracted. An important thing that I feel important to mention that in some cases, some practices are performed which are not accepted theoretically.

The present is not free of limitations. There might have problems regarding lack of limitation in some aspects and also some minor mistakes such as typing mistakes. These few drawbacks have occurred merely due to time limitation and lack of secondary sources of information.

Though I have tried my best to keep the report free from errors, I apologize if any error is found which was not deliberately made. If the report can help any person in providing information, I will feel that the purpose of the report has been fulfilled. Please feel free to contact me if any question arises.

Yashraj Singh Rawat
19MCAL004

Acknowledgement

The satisfaction that accompanies with the successful completion of any task would be incomplete without the mention of people whose cooperation made it possible, whose constant guidance and encouragement crown all efforts with success. I am grateful to the Project Guide Mr. Rajat Goyal and Mr. Sachin Yadav for the guidance, inspiration and constructive suggestions that helped me in the preparation of this project. I would also like to thank my fellow Classmates who helped in successful completion of the project by giving their Input and Views on the Project.

Yashraj Singh Rawat

Declaration

I hereby declare that the project work entitled “Data Science Projects” is an authentic record of my own work carried out at “Global It Providers” as requirements of six months project semester for the award of degree of Master of Computer Applications, JECRC University, under the guidance of “Rajat Goyal” and “Harshit Sharma”, during January to June, 2021.

Yashraj Singh Rawat
19MCAL004

Date: 24/06/2021

Certified that the above statement made by the student is correct to the best of our knowledge.

Mr. Harshit Sharma
(Assistant Professor)

Mr. Rajat Goyal
(Director)

Abstract

Global IT Providers was founded in 2014 since then it has emerged as a niche managed hosting, infrastructure management and server Management Company.
GIP specializes in managing servers and core IT infrastructure for small and medium sized organizations.

Exploratory Data Analysis - an approach / philosophy for data analysis which employs a variety of graphical techniques to maximize insight into a data set, uncover underlying structure, extract important factors, detect outliers & anomalies, test underlying assumptions, develop parsimonious models and determine optimal factor settings.

General problem areas consist of Uni variate, Multi-factor, Regression an Multivariate. For each of these 4 problem areas, heavy emphasis will be placed not only on the selection of appropriate EDA techniques, but also on the interpretation of output from such techniques so as to form a full and complete set of valid scientific/engineering conclusions. In short, the analyses will be conclusions-driven, and EDA will be the primary tool to develop such conclusions.
EDA techniques to be discussed include standard commonly-used tools such as histograms, probability plots, box plots, residual plots, and less commonly-used (but powerful) tools such as 4-plots, lag plots, PPCC plots, bi-histograms, block plots, GANOVA plots, interaction plots, transformation plots, spectral plots, Youden plots, a variety of "multi-plots", etc.
EDA principles, of course, serve as the link between data set and EDA technique. These principles are the "guidance system" to choose the appropriate EDA technique from the collection of possible EDA techniques. Such principles will be discussed along the way in conjunction with each data set.

The data sets will be drawn primarily from science and engineering applications, but I additionally include a few non-scientific data sets (e.g., Product reviews), and a few CSV data sets.

Final Certificate

Global IT Providers

Hosting & Server Management

Date: 20th June 2021

TO WHOMSOEVER IT MAY CONCERN

This is to certify that **Yashraj Singh Rawat** has done his internship in **Data Science** at **GIP technologies Pvt. Ltd. , from January 20th 2021 to June 20, 2021**

He has worked on a project titled EDA-Stack Overflow Survey 2020

During his internship he has demonstrated his skills with self-motivation to learn new skills. His performance exceeded our expectations and he was able to complete the project on time.

We wish him all the best for his upcoming career.

For GIP Technologies Private Limited

Authorized Signatory

Rajat Goyal
Director, GIP technologies Pvt. Ltd.

GIP TECHNOLOGIES PVT. LTD.
B-4, Opp. Dainik Bhaskar, Bhaskar Flyover, JLN Marg, Jaipur, Rajasthan, 302015
GST No. 08AAGCG5142A1Z1

accounts@globalitproviders.com www.globalitproviders.com call : 8448444608

Offer Letter

Global IT Providers

Hosting & Server Management

To,

YashRaj Singh Rawat,

JECRC University Jaipur.

Sub: Regarding your internship in our Company's Technical Department.

Dear YashRaj,

We would like to confirm that your application for internship in Technical Department has been accepted. Here are the terms of internship while working with the Company:

1. Duration of internship will be from **January 20th, 2021 to June 15th, 2021.**
2. You will be designated as "Trainee" and will be entitled for a **stipend of Rs 0 p.m.** as per Company's Policy.
3. You will not be entitled or any other benefits from the company during this tenure.
4. From time to time, your performance will be evaluated and based on this, your incentives will be decided.
5. During internship, you are expected to abide Code of Conduct prescribed by the Company for all the employees.

Please feel free to contact us in case of further details. Wishing you good luck for your future endeavors.

For GIP Technologies Private Limited

Sincerely,

Rajat Goyal Authorized Signatory

Director

GIP Technologies Pvt. Ltd.

GIP TECHNOLOGIES PVT. LTD.
B-4, Opp. Dainik Bhaskar, Bhaskar Flyover, JLN Marg, Jaipur, Rajasthan, 302015
GST No. 08AAGCG5142A1Z1

accounts@globalitproviders.com www.globalitproviders.com call : 8448444608

Joining Report

Annexure II

Joining Report (MCA VI Semester 2021)

(To be sent by student within a week of joining by Email/ Registered Post to Head of the Concerned Department, JECRC University, Jaipur)

Student Details

Student Name: Yashraj Singh Rawat Student ID: 19MCAL004

Residential Address of Student: 145 Mahaveer Nagar-Y, 80 feet road, Sanganer, Jaipur
Mobile Number: 9461070460 Email ID: yashrajrawat733@gmail.com

Project Details

Title of Project: _____ Project Type: Core/ Non Core

Organization Name: GIT Technologies Pvt Ltd
Site Address: b-4 opp. Dainik Bhaskar, Bhaskar Flyover, JLN Marg, Jaipur
Phone Number: 8448444608 Email ID: rajat0377@gmail.com
Head Office Address: B4, Third Floor, Vivek Vihar Opp. Dainik Bhaskar office Jaipur,Rajasthan
Phone Number: 8448444608 Email ID: rajat0377@gmail.com

I hereby inform that I have joined the organization on 20th January, 2021 for the VI semester Industrial Project

Dated: 20-01-2021 Signature of Student

Certificate by the Industry Guide

Certified that the above mentioned student has joined our organization for the VI semester Industrial Project

For GIP Technologies Private Limited

Authorized Signatory

Dated: 20-01-2021 Signature of Industry Guide

(with company seal)

Name: Rajat Goyal Designation: Director

Company Profile

Global IT Providers was founded in 2014 since then it has emerged as a niche managed hosting, infrastructure management and server Management Company.

GIP specializes in managing servers and core IT infrastructure for small and medium sized organizations. We provide security, server administration and disaster prevention which facilitates management of Intranets, Online Customer Support Solutions, Web-based and regular Email messaging solutions, and business-critical IT infrastructure. The company provides a wide range of web services to more than 100 clients including Corporate, Government organizations, Online Media and individual entities.

GIP offers Web Hosting, Network Management, Server Management and Solutions that include Web servers, Application servers, Database server, and Hosted Exchange, SharePoint,Cloud Infrastructure, Virtualization and Email Servers in stand-alone or multi-tiered or clustered architecture basis.

Our business portfolio is designed to deliver cost effective and end-to-end business solutions right from conceptualization to implementation with a focus on enhancing productivity and maximizing business performance.

Our mission is simple to consistently deliver the highest quality hosting services to a worldwide audience while maintaining our honesty and integrity in how we do business. We seek to cultivate an environment where our business and our clients can achieve mutual success. We live, sleep and breathe hosting as individuals and as a team, we absolutely love what we do.

Table of Contents

Data Analysis and Interpretation

- Exploratory Data Analysis

  i. Stack Overflow Developer Survey 2020

Inference and Conclusion

Recommendations

References and Future Work

Data Science

- Statistics
- Collection of Data
- Presentation of Data
- Analysis of Data
- Interpretation of Data

Statistics

“Statistics is the mathematical science involving the collection, analysis and interpretation of data”

“Statistics are the classified facts representing the conditions of people in a state. In particular they are the facts, which can be stated in numbers or in tables of numbers or in any tabular or classified arrangement”

Descriptive Statistics: Collection, Organization, summarization and presentation of data.

Inferential Statistics: Generalizing from sample to population, performing estimations and hypothesis testing, and making predictions.

Data Collection

For statistical analysis, whether it is business, economics, social sciences, science, or other fields, the basic problem is to collect facts and figures relating to particular phenomenon under study. It is the first step and this is the foundation upon which the entire data set. Careful planning is essential before collecting the data. There are different methods of collection of data such as census, sampling, primary, secondary, etc., and the investigator should make use of correct method.

Primary Data: Primary data is the one, which is collected by the investigator himself for the purpose of a specific inquiry or study. Such data is original in character and is generated by survey conducted by individuals or research institution or any organization.

Secondary Data: Secondary data are those data which have been already collected and analyzed by some earlier agency for its own use; and later the same data are used by a different agency.

Presentation of data

The mass data collected should be presented in a suitable, concise form for further analysis. The collected data may be presented in the form of tabular or diagrammatic or graphic form.

Analysis of data

The data presented should be carefully analyzed for making inference from the presented data such as measures of central tendencies, dispersion, correlation, regression etc.

Interpretation of data

The final step is drawing conclusion from the data collected. A valid conclusion must be drawn on the basis of analysis. A high degree of skill and experience is necessary for the interpretation.

Projects

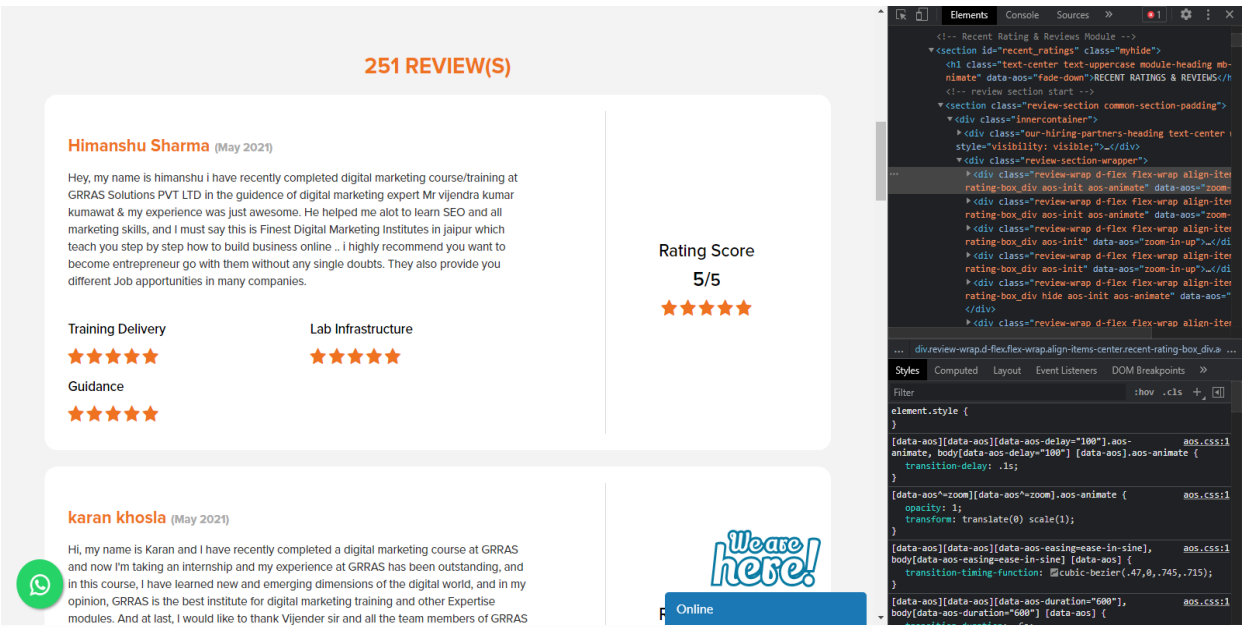
- Web Scraping
  i. Grras Reviews
  ii. Amazon Product Reviews

- Exploratory Data Analysis
  i. Stack Overflow Developer Survey 2020

GRRAS REVIEWS
(Web Scrapping)

GRRAS Solutions specializes in the domain of Red Hat Linux training, AWS Cloud Computing, Digital Marketing, Python, Website Design & Development, Big-data Hadoop for In-house training, Industrial/ Internship training, Online Learning and Corporate Training. Being an authorized and renowned partner of Red Hat since 2008, from last 12 years we hold special badge of honor for providing excellent business and learning facility across India. It also has our own Pearson VUE examination center, Red-hat Authorized Centre & Criterion Authorized Testing Center.

To get started with scrapping the data we here use python libraries such as requests and bs4. The requests library is the DE facto standard for making HTTP requests in Python. It abstracts the complexities of making requests behind a beautiful, simple API so that you can focus on interacting with services and consuming data in your application. Beautiful Soup is a library that makes it easy to scrape information from web pages. It sits atop an HTML or XML parser, providing Python idioms for iterating, searching, and modifying the parse tree.



Here is a list of reviews we are going to extract and based on those reviews we are going to analyze how good is Grras Solutions.

```
In [156]: import requests
          import bs4

In [157]: url='https://grras.com/page/ratings-reviews'

In [158]: resp=requests.get(url)
```

On making a request from the server we get status code 200. Any code in 2xx is a code which tells that you have successfully connected to the server.

```
In [159]: resp.status_code
Out[159]: 200

In [5]: resp.headers
Out[5]: {'Date': 'Tue, 08 Jun 2021 09:32:10 GMT', 'Server': 'Apache/2.4.18 (Ubuntu)', 'Cache-Control': 'no-cache, public', 'Set-Cooki
        e': 'XSRF-TOKEN=eyJpdiI6InBHUHM4ZWJrcmVVbUJub255eTA4U2c9PSIsInZhbHVlIjoidzk4cWFJTVZIcFJKVytkbE1XQmlqNWtOSnhSYWRzeHlOQ2kzYmhWaWZ
        EcXpuQ2ZVWlNXSjB2XC9oaWRuWHBIZWlab0c4V3htMUFEbThabmVTTTRBR2NBPT0iLCJtYWMiOiJmOTM4YTMwMzIwNjIyODY2YjE5ZTIwNmU1ZGMyN2JlOTQwMTYwMW
        Q4MmZlMWJlYTRiN2Q1MGI1MzFhY2NjMmFmIn0%3D; expires=Tue, 08-Jun-2021 11:32:11 GMT; Max-Age=7200; path=/, laravel_session=eyJpdiI6
        InNDOWwxUjFLK3pVMkg5YllSK1AwWmc9PSIsInZhbHVlIjoicXh6aHVGWVZGUHpReFZOQlFoTjhYeUM4aDdrc080T1ZqWlFXVjFac0JDSDF0cnNZbEJIM1hOTldrblN
        jb1RkU1cwc3BuUDhUYU1KNUM1Nlh2VHBwVHc9PSIsIm1hYyI6IjU4ZmRjMmI2OTBhZGMxOGM3YmU5ZmFjOGFmMzM2NDc3MmU0OGRhMDgxM2YwYjM0OTdiNjcwYzBmMT
        ZlYWZjMzEifQ%3D%3D; expires=Tue, 08-Jun-2021 11:32:11 GMT; Max-Age=7200; path=/; HttpOnly', 'Vary': 'Accept-Encoding', 'Content
        -Encoding': 'gzip', 'Referrer-Policy': '', 'Content-Length': '46322', 'Keep-Alive': 'timeout=5, max=100', 'Connection': 'Keep-A
        live', 'Content-Type': 'text/html; charset=UTF-8'}

In [6]: resp.headers['content-type']
Out[6]: 'text/html; charset=UTF-8'
```

Now we use soup to convert text/HTML format to beautiful soup object format.

```
In [7]: soup=bs4.BeautifulSoup(resp.content,'html5')

In [8]: divs=soup.find_all('div',attrs={'class':'review-wrap-content'})
```

On converting HTML into beautiful soup objects we get the functionality to access tags. Here we have select division tag of class ‘review-wrap-content’. Using this divs we can have access to all the div tags in the ‘review-wrap-content’. Also we can access other tags such as h3, which here tells us the name of the person who posted review on Grras.

```
In [9]: divs[1].h3.text
Out[9]: 'karan khosla (May 2021)'

In [10]: divs[1].p.text
Out[10]: "Hi, my name is Karan and I  have recently completed a digital marketing course at GRRAS and now I'm taking an internship and m
y experience at GRRAS has been outstanding, and in this course, I have learned new and emerging dimensions of the digital worl
d, and in my opinion, GRRAS is the best institute for digital marketing training and other Expertise modules. And at last, I wo
uld like to thank Vijender sir and all the team members of GRRAS for bringing out the best in me."

In [11]: name,time=divs[1].h3.text.split('(')

In [12]: name
Out[12]: 'karan khosla '

In [13]: time
Out[13]: 'May 2021)'
```

Now we created a dictionary Grras which contains the name, month, reviews and other information posted by the users on Grras website.

So that we can have access to all the reviews and ratings all at once.

```
In [14]: Grras={
             'Reviews':[],
             'Month':[],
             'Users':[],
             'Training_Delivery':[],
             'Lab_Infra':[],
             'Guidance':[],
         }

         for div in divs:
             review=div.p.text
             Grras['Reviews'].append(review)

         for div in divs:
             name,time=div.h3.text.split('(')
             Grras['Users'].append(name)
             Grras['Month'].append(time[:-1])

In [15]: Grras['Reviews'][:3]
Out[15]: ['Hey, my name is himanshu i have recently completed digital marketing course/training at GRRAS Solutions PVT LTD in the guiden
ce of digital marketing expert Mr vijendra kumar kumawat & my experience was just awesome. He helped me alot to learn SEO and a
ll marketing skills, and I must say this is Finest Digital Marketing Institutes in jaipur which teach you step by step how to b
uild business online .. i highly recommend you want to become entrepreneur go with them without any single doubts.\nThey also p
rovide you different Job apportunities in many companies.',
 "Hi, my name is Karan and I  have recently completed a digital marketing course at GRRAS and now I'm taking an internship and
my experience at GRRAS has been outstanding, and in this course, I have learned new and emerging dimensions of the digital worl
d, and in my opinion, GRRAS is the best institute for digital marketing training and other Expertise modules. And at last, I wo
uld like to thank Vijender sir and all the team members of GRRAS for bringing out the best in me.",
 'I have a good experience at grras solutions institute, even here everyone is supporting.\nAlso, the workshops they give are f
abulous. In one line , I only say " it is one of the best institute " in every aspect.']
```

Here we extract name,time and reviews of the people on Grras.com.

```
In [16]: # Name, Time, Reviews Scrapped

         for name,time,review in zip(Grras['Users'],Grras['Month'],Grras['Reviews']):
             print(name.upper())
             print(time)
             print(review)
             print('_'*120)
             print('\n')
```

And here are some of the top reviews.

HIMANSHU SHARMA
May 2021
Hey, my name is himanshu i have recently completed digital marketing course/training at GRRAS Solutions PVT LTD in the guidence of digital marketing expert Mr vijendra kumar kumawat & my experience was just awesome. He helped me alot to learn SEO and all marketing skills, and I must say this is Finest Digital Marketing Institutes in jaipur which teach you step by step how to buil d business online .. i highly recommend you want to become entrepreneur go with them without any single doubts.
They also provide you different Job apportunities in many companies.

KARAN KHOSLA
May 2021
Hi, my name is Karan and I have recently completed a digital marketing course at GRRAS and now I'm taking an internship and my experience at GRRAS has been outstanding, and in this course, I have learned new and emerging dimensions of the digital world, and in my opinion, GRRAS is the best institute for digital marketing training and other Expertise modules. And at last, I would like to thank Vijender sir and all the team members of GRRAS for bringing out the best in me.

RITIK SHARMA
Apr 2020
I have a good experience at grras solutions institute, even here everyone is supporting.
Also, the workshops they give are fabulous. In one line , I only say " it is one of the best institute " in every aspect.

NIHARIKA CHAUHAN
Apr 2020
It have been wonderful experience with Grras ". I learnt Data Science with python here ,Trainers were always able to address ev ery question we had and every problem immediately and adequately. It’s clear that our success and career is their top priority.
I am thinking of persuing machine learning and big data hadoop also from Grras.

MEGHA SHARMA
Apr 2020
I did the whole course online. It was the first time I tried e-learning and I am very satisfied with the outcome. My feedback w

Similarly we have to find the ratings of the users.

For that we first have to inspect Grras.com/reviews and find that where the ratings lies for infrastructure, labs and guidance.

```
In [17]: divs[0].div.find_all('div',attrs={'class':"col-lg-3 col-md-6 review-feature"})

Out[17]: [<div class="col-lg-3 col-md-6 review-feature">
                    <h5>Training Delivery</h5>
                    <div class="star-rating">
                                <span class="fa fa-star"></span>
                              <span class="fa fa-star"></span>
                              <span class="fa fa-star"></span>
                              <span class="fa fa-star"></span>
                              <span class="fa fa-star"></span>
                            </div>
                  </div>,
          <div class="col-lg-3 col-md-6 review-feature">
                    <h5>Lab Infrastructure</h5>
                    <div class="star-rating">
                                  <span class="fa fa-star"></span>
                                  <span class="fa fa-star"></span>
                                  <span class="fa fa-star"></span>
                                  <span class="fa fa-star"></span>
                                  <span class="fa fa-star"></span>
                              </div>
                  </div>,
          <div class="col-lg-3 col-md-6 review-feature">
                    <h5>Guidance</h5>
                    <div class="star-rating">
                                <span class="fa fa-star"></span>
                              <span class="fa fa-star"></span>
                              <span class="fa fa-star"></span>
                              <span class="fa fa-star"></span>
                              <span class="fa fa-star"></span>
                            </div>
                  </div>]
```

We have found the class and division where the infrastructure lies. But this if for user 1st only as you can see divs[0] at the starting of code. Now to access each user we have to loop through our code.

```
In [20]: divs[0].div.find_all('div',attrs={'class':"col-lg-3 col-md-6 review-feature"})[0].div.find_all('span',attrs={'class':"fa fa-star"

Out[20]: [<span class="fa fa-star"></span>,
          <span class="fa fa-star"></span>,
          <span class="fa fa-star"></span>,
          <span class="fa fa-star"></span>,
          <span class="fa fa-star"></span>]

In [21]: #rating
         len(divs[0].div.find_all('div',attrs={'class':"col-lg-3 col-md-6 review-feature"})[0].div.find_all('span',attrs={'class':"fa fa-s

Out[21]: 5

In [22]: #Lab_Infra
         len(divs[0].div.find_all('div',attrs={'class':"col-lg-3 col-md-6 review-feature"})[1].div.find_all('span',attrs={'class':"fa fa-s

Out[22]: 5

In [23]: #Guidance
         len(divs[0].div.find_all('div',attrs={'class':"col-lg-3 col-md-6 review-feature"})[2].div.find_all('span',attrs={'class':"fa fa-s

Out[23]: 5
```

This group of code iterate through all the division and save users index numbers and their ratings.

```
In [24]: c=1
         for div in divs:
             l=[]
             for d in div.find('div',attrs={"class":"row review-feature-wrapper"}).find_all('div',attrs={'class':"col-lg-3 col-md-6 revie
                 l.append(len(d.div.find_all('span',attrs={'class':"fa fa-star"})))
             print(c,l)
             Grras['Training_Delivery'].append(l[0])
             Grras['Lab_Infra'].append(l[1])
             Grras['Guidance'].append(l[2])
             print('_'*70)
             c+=1
```

Here you can see how it's happening.

```
1 [5, 5, 5]
______________________________________________________________________
2 [5, 5, 5]
______________________________________________________________________
3 [5, 5, 5]
______________________________________________________________________
4 [5, 5, 5]
______________________________________________________________________
5 [5, 5, 5]
______________________________________________________________________
6 [5, 5, 5]
______________________________________________________________________
7 [5, 5, 5]
______________________________________________________________________
8 [5, 5, 5]
______________________________________________________________________
9 [5, 4, 5]
______________________________________________________________________
10 [4, 4, 4]
```

Now we merge the code to find name, month and reviews with ratings.

```
In [26]: def get_content(soup):
             Grras={
             'Reviews':[],
             'Month':[],
             'Users':[],
             'Training_Delivery':[],
             'Lab_Infra':[],
             'Guidance':[],
         }

             for div in divs:
                 review=div.p.text
                 Grras['Reviews'].append(review)
                 #_____________________________________________________________
                 name,time=div.h3.text.split('(')
                 Grras['Users'].append(name)
                 Grras['Month'].append(time[:-1])
                 #_____________________________________________________________
                 l=[]
                 for d in div.find('div',attrs={"class":"row review-feature-wrapper"}).find_all('div',attrs={'class':"col-lg-3 col-md-6 re
                     l.append(len(d.div.find_all('span',attrs={'class':"fa fa-star"})))
                 Grras['Training_Delivery'].append(l[0])
                 Grras['Lab_Infra'].append(l[1])
                 Grras['Guidance'].append(l[2])
                 c+=1
             return data
```

Here I simply print the output on my notebook. It contains name, month, reviews and ratings of the users.

```
In [27]: for name,time,review,t,l,g in zip(Grras['Users'],Grras['Month'],Grras['Reviews'],Grras['Training_Delivery'],Grras['Lab_Infra'],Gr
             print("NAME : ",name.upper())
             print("TIME : ",time)
             print("REVIEWS : ",review)
             print('\n')
             print("---------------------------")
             print("Ratings".center(25))
             print("---------------------------")
             print("TRAINING DELIVERY : ",t)
             print("LABS : ",l)
             print("GUIDANCE : ",g)
             print('_'*120)
             print('\n')
```

Output: As you can see we have extracted the data successfully from Grras.com to our notebook.

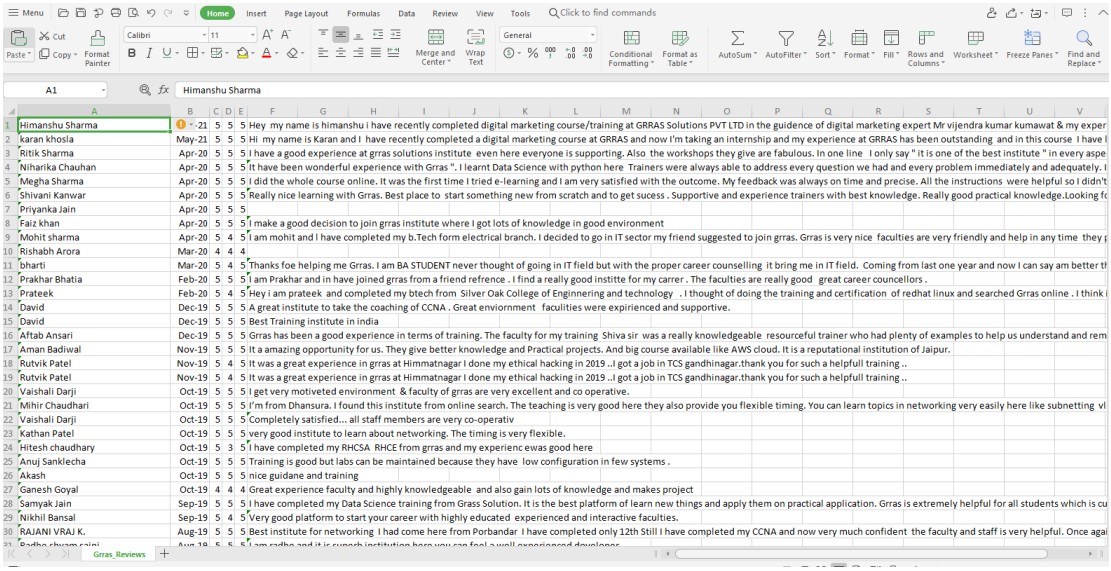
```
NAME :  HIMANSHU SHARMA
TIME :  May 2021
REVIEWS :  Hey, my name is himanshu i have recently completed digital marketing course/training at GRRAS Solutions PVT LTD in t
he guidence of digital marketing expert Mr vijendra kumar kumawat & my experience was just awesome. He helped me alot to learn
SEO and all marketing skills, and I must say this is Finest Digital Marketing Institutes in jaipur which teach you step by step
how to build business online .. i highly recommend you want to become entrepreneur go with them without any single doubts.
They also provide you different Job apportunities in many companies.


---------------------------
         Ratings
---------------------------
TRAINING DELIVERY :  5
LABS :  5
GUIDANCE :  5
____________________________________________________________________________________________________________________


NAME :  KARAN KHOSLA
TIME :  May 2021
REVIEWS :  Hi, my name is Karan and I  have recently completed a digital marketing course at GRRAS and now I'm taking an intern
ship and my experience at GRRAS has been outstanding, and in this course, I have learned new and emerging dimensions of the dig
ital world, and in my opinion, GRRAS is the best institute for digital marketing training and other Expertise modules. And at l
ast, I would like to thank Vijender sir and all the team members of GRRAS for bringing out the best in me.


---------------------------
         Ratings
---------------------------
TRAINING DELIVERY :  5
LABS :  5
GUIDANCE :  5
____________________________________________________________________________________________________________________
```

We now store the extracted data in .CVS format so that it can be accessible using ms-excel as well.

```
In [28]: def write_csv(data):
             fp=open('Grras_Reviews.csv','w')
             for name,time,review,t,l,g in zip(Grras['Users'],Grras['Month'],Grras['Reviews'],Grras['Training_Delivery'],Grras['Lab_Infra'
                 review=review.replace('\n',' ')
                 review=review.replace(',',' ')
                 line=f'{name},{time},{t},{l},{g},{review}\n'
                 fp.write(line)
             fp.close()
             print('Data Written Successfully')

In [29]: write_csv(Grras)

         Data Written Successfully
```

Here is the .CVS file.

As you can see data is arranged in well ordered and can be accessible easily as well.



1	Himanshu Sharma	-21	5	5	5	Hey my name is himanshu i have recently completed digital marketing course/training at GRRAS Solutions PVT LTD in the guidence of digital marketing expert Mr vijendra kumar kumawat & my exper
2	karan khosla	May-21	5	5	5	Hi my name is Karan and I have recently completed a digital marketing course at GRRAS and now I'm taking an internship and my experience at GRRAS has been outstanding and in this course I have l
3	Ritik Sharma	Apr-20	5	5	5	I have a good experience at grras solutions institute even here everyone is supporting. Also the workshops they give are fabulous. In one line I only say " it is one of the best institute " in every aspe
4	Niharika Chauhan	Apr-20	5	5	5	It have been wonderful experience with Grras ". I learnt Data Science with python here Trainers were always able to address every question we had and every problem immediately and adequately. I
5	Megha Sharma	Apr-20	5	5	5	I did the whole course online. It was the first time I tried e-learning and I am very satisfied with the outcome. My feedback was always on time and precise. All the instructions were helpful so I didn't
6	Shivani Kanwar	Apr-20	5	5	5	Really nice learning with Grras. Best place to start something new from scratch and to get sucess . Supportive and experience trainers with best knowledge. Really good practical knowledge.Looking fo
7	Priyanka Jain	Apr-20	5	5	5	
8	Faiz khan	Apr-20	5	5	5	I make a good decision to join grras institute where I got lots of knowledge in good environment
9	Mohit sharma	Apr-20	5	4	5	I am mohit and I have completed my b.Tech form electrical branch. I decided to go in IT sector my friend suggested to join grras. Grras is very nice faculties are very friendly and help in any time they p
10	Rishabh Arora	Mar-20	4	4	4	
11	bharti	Mar-20	5	4	5	Thanks foe helping me Grras. I am BA STUDENT never thought of going in IT field but with the proper career counselling it bring me in IT field. Coming from last one year and now I can say am better th
12	Prakhar Bhatia	Feb-20	5	5	5	I am Prakhar and in have joined grras from a friend refrence . I find a really good institte for my carrer . The faculties are really good great career councellors .
13	Prateek	Feb-20	5	4	5	Hey i am prateek and completed my btech from Silver Oak College of Enginnering and technology . I thought of doing the training and certification of redhat linux and searched Grras online . I think i
14	David	Dec-19	5	5	5	A great institute to take the coaching of CCNA . Great enviornment faculities were expirienced and supportive.
15	David	Dec-19	5	5	5	Best Training institute in india
16	Aftab Ansari	Dec-19	5	5	5	Grras has been a good experience in terms of training. The faculty for my training Shiva sir was a really knowledgeable resourceful trainer who had plenty of examples to help us understand and rem
17	Aman Badiwal	Nov-19	5	5	5	It a amazing opportunity for us. They give better knowledge and Practical projects. And big course available like AWS cloud. It is a reputational institution of Jaipur.
18	Rutvik Patel	Nov-19	5	4	5	It was a great experience in grras at Himmatnagar I done my ethical hacking in 2019 ..I got a job in TCS gandhinagar.thank you for such a helpfull training ..
19	Rutvik Patel	Nov-19	5	4	5	It was a great experience in grras at Himmatnagar I done my ethical hacking in 2019 ..I got a job in TCS gandhinagar.thank you for such a helpfull training ..
20	Vaishali Darji	Oct-19	5	5	5	I get very motiveted environment & faculty of grras are very excellent and co operative.
21	Mihir Chaudhari	Oct-19	5	5	5	I'm from Dhansura. I found this institute from online search. The teaching is very good here they also provide you flexible timing. You can learn topics in networking very easily here like subnetting vl
22	Vaishali Darji	Oct-19	5	5	5	Completely satisfied... all staff members are very co-operativ
23	Kathan Patel	Oct-19	5	5	5	very good institute to learn about networking. The timing is very flexible.
24	Hitesh chaudhary	Oct-19	5	3	5	I have completed my RHCSA RHCE from grras and my experienc ewas good here
25	Anuj Sanklecha	Oct-19	5	5	5	Training is good but labs can be maintained because they have low configuration in few systems .
26	Akash	Oct-19	5	5	5	nice guidane and training
27	Ganesh Goyal	Oct-19	4	4	4	Great experience faculty and highly knowledgeable and also gain lots of knowledge and makes project
28	Samyak Jain	Sep-19	5	5	5	I have completed my Data Science training from Grass Solution. It is the best platform of learn new things and apply them on practical application. Grras is extremely helpful for all students which is cu
29	Nikhil Bansal	Sep-19	5	4	5	Very good platform to start your career with highly educated experienced and interactive faculties.
30	RAJANI VRAJ K.	Aug-19	5	5	5	Best institute for networking I had come here from Porbandar I have completed only 12th Still I have completed my CCNA and now very much confident the faculty and staff is very helpful. Once agai

Now to perform analysis on the data-set we first have to import libraries like numpy, pandas and matplotlib. Which helps in data analysis and data visualization.

```
In [31]: import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt

In [32]: Grras.keys()

Out[32]: dict_keys(['Reviews', 'Month', 'Users', 'Training_Delivery', 'Lab_Infra', 'Guidance'])

In [33]: print(len(Grras['Reviews']),len(Grras['Month']),len(Grras['Users']),len(Grras['Training_Delivery']),

         251 251 251 251 251 251

In [34]: df=pd.DataFrame(Grras)

In [35]: df=df[['Users','Month','Reviews','Training_Delivery','Lab_Infra','Guidance']]

In [36]: df.info()

         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 251 entries, 0 to 250
         Data columns (total 6 columns):
          #   Column             Non-Null Count  Dtype
         ---  ------             --------------  -----
          0   Users              251 non-null    object
          1   Month              251 non-null    object
          2   Reviews            251 non-null    object
          3   Training_Delivery  251 non-null    int64
          4   Lab_Infra          251 non-null    int64
          5   Guidance           251 non-null    int64
         dtypes: int64(3), object(3)
         memory usage: 11.9+ KB
```

We then find ratings on training delivery, labs and guidance to student in the organization.

```
In [38]: df.describe()
```

Out[38]:

	Training_Delivery	Lab_Infra	Guidance
count	251.000000	251.000000	251.000000
mean	4.888446	4.701195	4.872510
std	0.394342	0.595280	0.379053
min	3.000000	2.000000	3.000000
25%	5.000000	5.000000	5.000000
50%	5.000000	5.000000	5.000000
75%	5.000000	5.000000	5.000000
max	5.000000	5.000000	5.000000

```
In [39]: df['Training_Delivery'].value_counts()
```

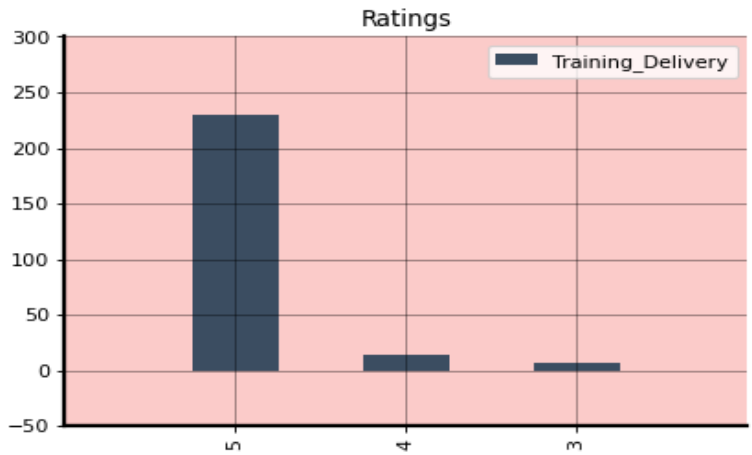
```
Out[39]: 5    230
         4     14
         3      7
         Name: Training_Delivery, dtype: int64
```

```
In [40]: train=df['Training_Delivery'].value_counts()
         train
```

```
Out[40]: 5    230
         4     14
         3      7
         Name: Training_Delivery, dtype: int64
```

According to the analysis 90% of the students are satisfied by the training given to them.

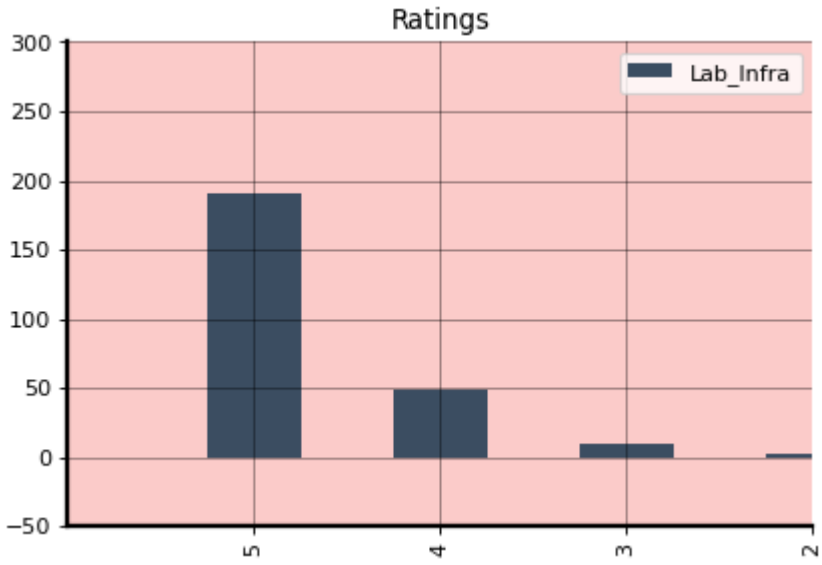
In [41]:
```
plt.figure(dpi=80)
ax=plt.gca()
ax.set_facecolor('#fbcbc9')
plt.title('Ratings')
ax.spines['top'].set_visible(False)
ax.spines['right'].set_visible(False)
ax.spines['left'].set_color('black')
ax.spines['bottom'].set_color('black')
train.plot(kind='bar',color='#3b4d61')
plt.legend()
ax.spines['left'].set_lw(2)
ax.spines['bottom'].set_lw(2)
plt.grid(alpha=.4,color='black')
plt.axis([-1,3,-50,300])
plt.show()
```



In [42]:
```
train2=df['Lab_Infra'].value_counts()
```

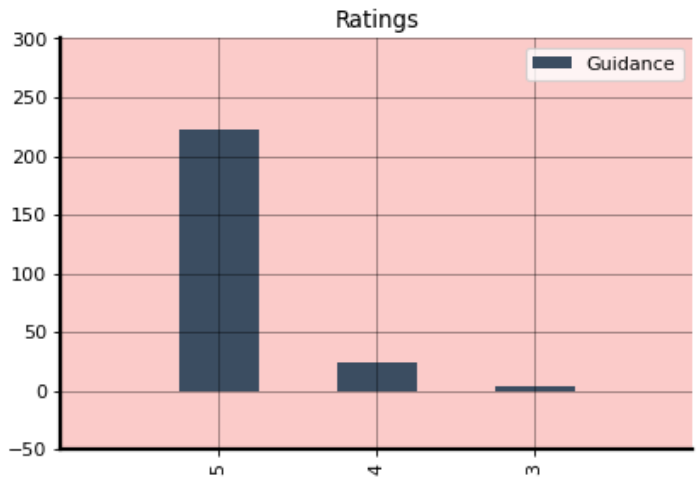
In [43]:
```
plt.figure(dpi=80)
ax=plt.gca()
ax.set_facecolor('#fbcbc9')
plt.title('Ratings')
ax.spines['top'].set_visible(False)
ax.spines['right'].set_visible(False)
ax.spines['left'].set_color('black')
ax.spines['bottom'].set_color('black')
train2.plot(kind='bar',color='#3b4d61')
plt.legend()
ax.spines['left'].set_lw(2)
ax.spines['bottom'].set_lw(2)
plt.grid(alpha=.4,color='black')
plt.axis([-1,3,-50,300])
plt.show()
```

Here lab infrastructure need to be maintain properly as 20% students are not satisfied with the labs.



90% of the students are satisfied by the teaching given to them.

```
In [45]: plt.figure(dpi=80)
         ax=plt.gca()
         ax.set_facecolor('#fbcbc9')
         plt.title('Ratings')
         ax.spines['top'].set_visible(False)
         ax.spines['right'].set_visible(False)
         ax.spines['left'].set_color('black')
         ax.spines['bottom'].set_color('black')
         train3.plot(kind='bar',color='#3b4d61')
         plt.legend()
         ax.spines['left'].set_lw(2)
         ax.spines['bottom'].set_lw(2)
         plt.grid(alpha=.4,color='black')
         plt.axis([-1,3,-50,300])
         plt.show()
```



Conclusion:

So by extracting and analyzing the data we got to know that lab infrastructure should be maintained as fair number of student are not happy with the labs, which might be a negative aspect for the organization.

Also providing good infrastructure in labs also helps the student to stay motivated on the work without any disturbance and good infrastructure also helps in attracting new audience.

Amazon Product Reviews
(Web Scraping)

Here I scraped reviews from Amazon of the product One Plus 9R 5G. And according to the reviews and the ratings, I have decided whether to buy this product or not.

At first, we have to import libraries like bs4 and request. Beautiful Soup is a library that makes it easy to scrape information from web pages. It sits atop an HTML or XML parser, providing Pythonic idioms for iterating, searching, and modifying the parse tree. The requests module allows you to send HTTP requests using Python. The HTTP request returns a Response Object with all the response data (content, encoding, status, etc).

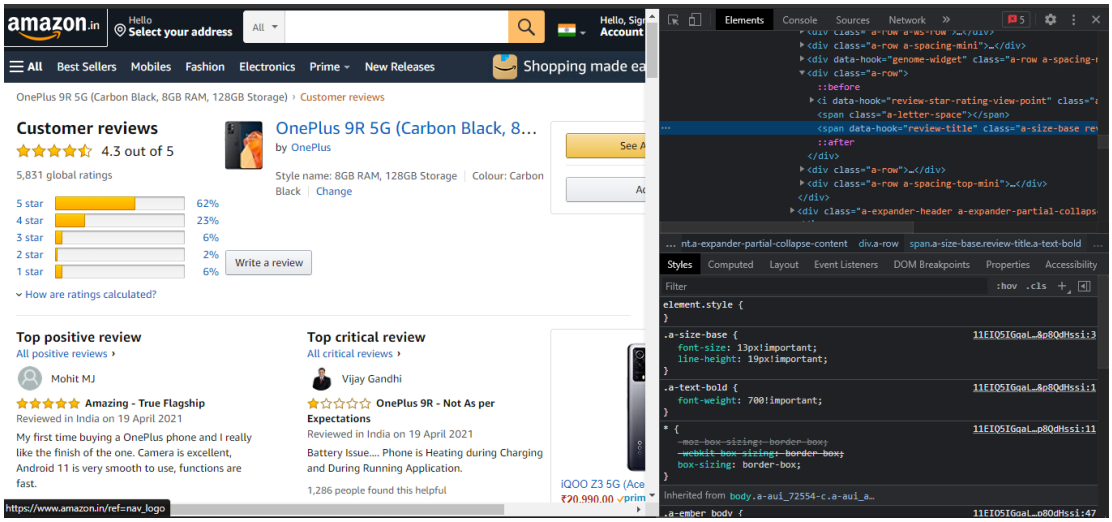
```
In [33]: url="https://www.amazon.in/Test-Exclusive_2020_1178-Multi-3GB-Storage/product-reviews/B089MTJVLD/ref=cm_cr_arp_d_paging_btm_next_

In [34]: import requests
         import bs4
         print(url)

         https://www.amazon.in/Test-Exclusive_2020_1178-Multi-3GB-Storage/product-reviews/B089MTJVLD/ref=cm_cr_arp_d_paging_btm_next_1?i
         e=UTF8&reviewerType=all_reviews&pageNumber=1
```

Now we check that the request we have made to the URL has successfully received by the server or not. On getting data in the range of 200-300 signifies that the request has been made and received successfully.

```
In [3]: resp=requests.get(url)

In [4]: resp.status_code
Out[4]: 200

In [5]: resp.headers['content-type']
Out[5]: 'text/html;charset=UTF-8'
```

We will also check the type of data we have received. And further extract the whole text/HTML data from the URL in the latest HTML5 format.

```
In [6]: soup=bs4.BeautifulSoup(resp.content,'html5')
```



Here now we have to find the tags and classes that the reviews are using. So that we can extract only the data we need and not the remaining data.

```
In [7]: soup.find(attrs={"class":"a-row a-spacing-small review-data"})

Out[7]: <div class="a-row a-spacing-small review-data"><span class="a-size-base review-text review-text-content" data-hook="review-bod
        y">

          <span>
         Battery Issue.... Phone is Heating during Charging and During Running Application.
        </span>

        </span></div>
```

Here is how we can do it .

```
In [8]: soup.find('div',attrs={"class":"a-row a-spacing-small review-data"}).text.strip()

Out[8]: 'Battery Issue.... Phone is Heating during Charging and During Running Application.'

In [9]: reviews = [tag.text.strip() for tag in soup.find_all(attrs={"class":"a-row a-spacing-small review-data"})]
```

Now from the remaining data, we have to clean the data so that no additional spaces are used. For eg. Removing the extra spaces from front and back of the text using the strip method.

```
In [10]: len(reviews)

Out[10]: 10
```

After making sure that the data is cleaned, now we have to do this for all the reviews. But a single page contain only 10 reviews. So we first have to loop through those 10 reviews and than move to the next page. Also we have to make sure that the reviews we are extracting also be stored.

```
In [11]: print(*reviews,sep='\n\n')

Battery Issue.... Phone is Heating during Charging and During Running Application.

Facing heating issue while using camera app and general usage in 60hz refresh rate if I use 120hz it gradually heating issue in
crease please solve this problem

My first time buying a OnePlus phone and I really like the finish of the one. Camera is excellent, Android 11 is very smooth to
use, functions are fast.Just my first day of usage so hard to be critical of anything. The downside for me is that there is no
place expand the memory.

Within 10 minutes of usage. It’s felt like over heat. I can’t experienced any mobile like this.

Iam OnePlus user since 4 years, I exchange my one Plus 7 pro mobile with OnePlus 9 R Not expected From One Plus, quality is not
good, Look like cheep mobile, camera quality also not good, one Plus 7 pro superb mobile, totally iam disappointed with 9R.

I don't why no reviewer is speaking about it. It was heating with a normal usage.

After 5 days of usage writing this review.1. Best camera quality for the price2. Snapdragon 870 is doing its best in speed with
120Hz display. - Best3. 65W charger takes only 35 min to charge 100% from 15%. - Best4.  Fluid amoled display is awesome on its
smoothness. - best5. Lake blue colour is simply awesome 💙Issues1. Facing very lite heating issue.2. Found a bug and informed O
nePlus customer service which is if we turn on call recording on call it indicates opponent also that we are recording call. Th
ey told they will correct it in next update. Hope this will be cleared but not sure.I give 95/100 for this mobile.

I am writing down this after using for a couple of days. I got the carbon black one with 12GB RAM.  Overall the device is good,
and honestly is the reskin of one plus 8T with lower price 😐.  I will list down the pros and cons -pros -1. With Oxygen OS 11,
the overall user experience is quite smooth. You will get an update as soon as you finish setup.2. The Warp 65 fast charging is
fine, takes around 35 mins for one full charge.3. The screen to body ratio is good.4. The matt finish on the back side is aweso
me Cons -1. Battery drains a little fast, not sure if some future updates will fix it 2. Rear Camera is ok, can be better 3. I
```

As you can see, we have extracted all the reviews from the first page and you can also check that all the reviews are different. There may be chances that a few might get repeated.

Similarly we will extract the data of ratings.

```
In [12]: ratings=[tag.text.split()[0] for tag in soup.find_all('i',attrs={"data-hook":"review-star-rating"})]
```

```
100%|██████████████████████████████████████████████████████████| 99/99 [01:33<00:00,  1.06it/s]
```

100% means that the entire data in extracted successfully.

```
In [14]: def get_all_reviews(start,end):
             url="https://www.amazon.in/Test-Exclusive_2020_1178-Multi-3GB-Storage/product-reviews/B089MTJVLD/ref=cm_cr_arp_d_paging_btm_r
             page=0
             err_pages=[]
             data={
                 'reviews':[],
                  'ratings':[]
                  }
             for c in tqdm.tqdm(range(start,end)):
                 resp=requests.get(url.format(c,c))
                 if resp.status_code==200:
                     soup=bs4.BeautifulSoup(resp.content,'html5')
                     reviews = [tag.text.strip() for tag in soup.find_all(attrs={"class":"a-row a-spacing-small review-data"})]
                     ratings=[tag.text.split()[0] for tag in soup.find_all('i',attrs={"data-hook":"review-star-rating"})]
                     data['reviews'].extend(reviews)
                     data['ratings'].extend(ratings)
                     if len(reviews)<5:
                         print('-'*100)
                         print('successful'.center(50))
                         print('-'*100)
                         break
                 elif resp.status_code==503:
                     err_pages.append(c)
                 else:
                     print(f'Error!! {resp.status_code} {resp.reason}')
                     break

                 page+=1
             return data,page,err_pages
```

The above code is used for looping through the pages from the start to the end, storing ratings and reviews, checking the number of errors pages occurred and returning the data, no of page and error pages.

```
In [15]: data,page,err_pages=get_all_reviews(1,100)
         100%|██████████████████████████████████████████████████████████| 99/99 [01:33<00:00,  1.06it/s]

In [16]: page
Out[16]: 99

In [17]: len(data['reviews'])
Out[17]: 840

In [18]: len(err_pages)
Out[18]: 15
```

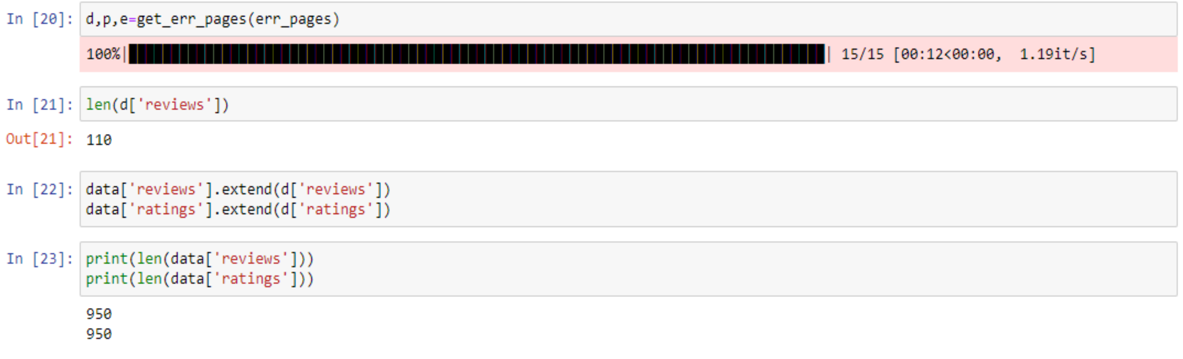
As you can see that 100% data has been extracted out of which no of pages are 99, because the range is from 1-100 (excluding 100). A single page contains 10 reviews and total no. Of reviews we got is 840, which means 84 pages are extracted successfully and remaining wont get loaded because of some issues.

So now we try to extract the data that are not loaded successfully, means the error pages.

```
In [19]: def get_err_pages(nums):
             url="https://www.amazon.in/Test-Exclusive_2020_1178-Multi-3GB-Storage/product-reviews/B089MTJVLD/ref=cm_cr_arp_d_paging_btm_r
             page=0
             err_pages=[]
             data={
                 'reviews':[],
                  'ratings':[]
                  }
             for c in tqdm.tqdm(nums):
                 resp=requests.get(url.format(c,c))
                 if resp.status_code==200:
                     soup=bs4.BeautifulSoup(resp.content,'html5')
                     reviews = [tag.text.strip() for tag in soup.find_all(attrs={"class":"a-row a-spacing-small review-data"})]
                     ratings=[tag.text.split()[0] for tag in soup.find_all('i',attrs={"data-hook":"review-star-rating"})]
                     data['reviews'].extend(reviews)
                     data['ratings'].extend(ratings)
                     if len(reviews)<5:
                         print('-'*100)
                         print('successful'.center(50))
                         print('-'*100)
                         break
                 elif resp.status_code==503:
                     err_pages.append(c)
                 else:
                     print(f'Error!! {resp.status_code} {resp.reason}')
                     break

                 page+=1
             return data,page,err_pages
```

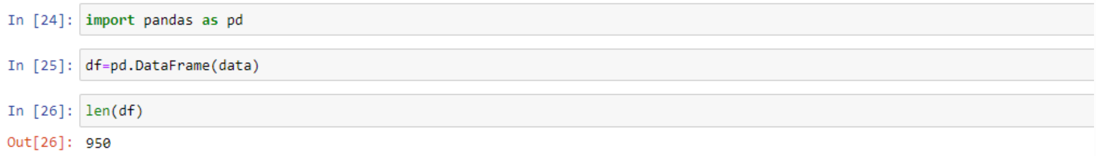
Here we loop through the error pages and tried to get more and more data, so that we can predict right outcome.

Even after extracting the remaining 15 pages we only get data from 11 pages.



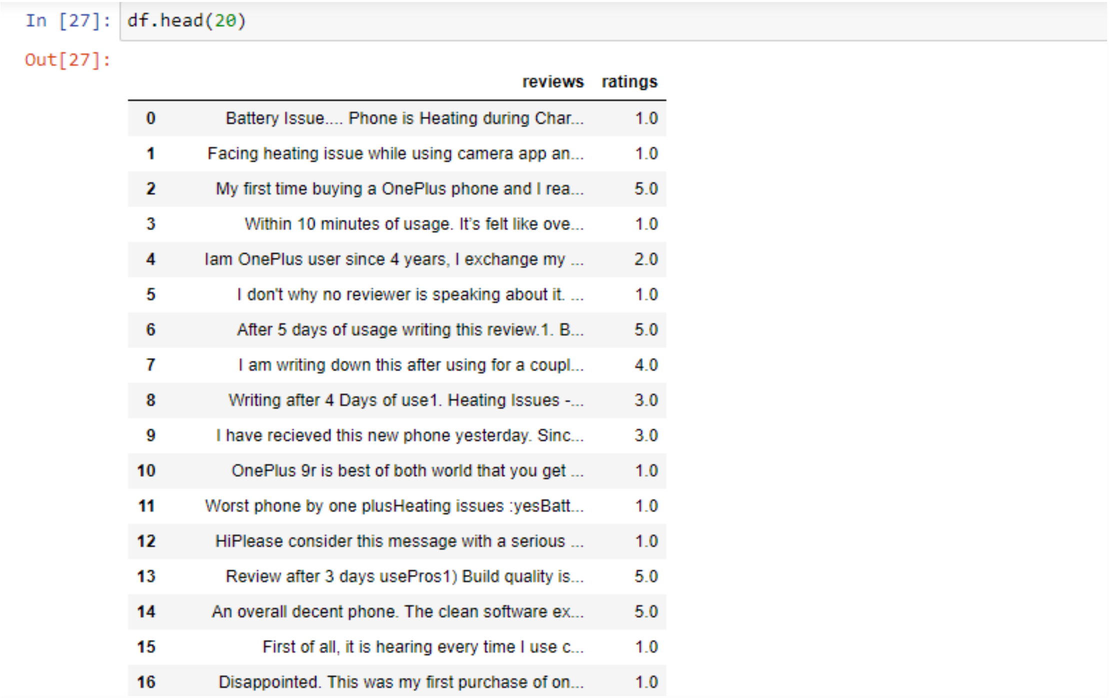
```
In [20]: d,p,e=get_err_pages(err_pages)

         100%|██████████████████████████████████████████████████| 15/15 [00:12<00:00,  1.19it/s]

In [21]: len(d['reviews'])

Out[21]: 110

In [22]: data['reviews'].extend(d['reviews'])
         data['ratings'].extend(d['ratings'])

In [23]: print(len(data['reviews']))
         print(len(data['ratings']))

         950
         950
```

Total number of reviews we get from this data is 950 out of 990. We then added the remaining 110 reviews and ratings to the earlier 840 rating and reviews.



```
In [24]: import pandas as pd

In [25]: df=pd.DataFrame(data)

In [26]: len(df)

Out[26]: 950
```

We can now easily convert this extracted data which is in form of list to the data frame. A data frame is a spreadsheet like structure, which contains heterogeneous values. Before converting data in data frame we must notice the size of data. The data must not be inadequate, i.e. the data you are converting to data frame, theirs columns must be of same size otherwise error occur.



```
In [27]: df.head(20)

Out[27]:
```

	reviews	ratings
0	Battery Issue.... Phone is Heating during Char...	1.0
1	Facing heating issue while using camera app an...	1.0
2	My first time buying a OnePlus phone and I rea...	5.0
3	Within 10 minutes of usage. It’s felt like ove...	1.0
4	Iam OnePlus user since 4 years, I exchange my ...	2.0
5	I don't why no reviewer is speaking about it. ...	1.0
6	After 5 days of usage writing this review.1. B...	5.0
7	I am writing down this after using for a coupl...	4.0
8	Writing after 4 Days of use1. Heating Issues -...	3.0
9	I have recieved this new phone yesterday. Sinc...	3.0
10	OnePlus 9r is best of both world that you get ...	1.0
11	Worst phone by one plusHeating issues :yesBatt...	1.0
12	HiPlease consider this message with a serious ...	1.0
13	Review after 3 days usePros1) Build quality is...	5.0
14	An overall decent phone. The clean software ex...	5.0
15	First of all, it is hearing every time I use c...	1.0
16	Disappointed. This was my first purchase of on...	1.0

```
In [28]: df.tail(20)
```

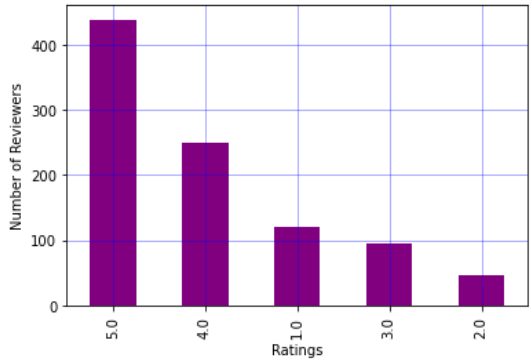
Out[28]:

	reviews	ratings
930	Mobile battery very puru	4.0
931	Value for money, best option for oneplus exper...	5.0
932	Excellent mobile. And best performanceBut 48 m...	5.0
933	Finger print is not quick as expected.Battery ...	4.0
934	Battery drains fast....charges very fast....pr...	4.0
935	Good phone - works as promised!	5.0
936	One plus quality superb	5.0
937	Fast charging supported	4.0
938	It was unacceptable... major heating issue.	1.0
939	Amazons pqckage was not up to the mark but one...	4.0
940	Awesome buy for this price, if you love perfor...	5.0
941	Really It's Awesome 😎❤️	5.0
942	Superbbbb mobile	5.0
943	excellent mobile	5.0
944	I will like and would recommend the phone to e...	5.0
945	I m facing vibration sense issue..it is not wo...	4.0
946	Excellent mobile with Excellent performance	4.0

As you can see that the ratings and reviews are not stored in data frames and now we can perform analysis on the data frame.

```
In [29]: len(df['reviews'].unique())
Out[29]: 935

In [30]: df.shape
Out[30]: (950, 2)

In [31]: df['ratings'].value_counts()
Out[31]: 5.0    438
         4.0    249
         1.0    121
         3.0     95
         2.0     47
         Name: ratings, dtype: int64
```

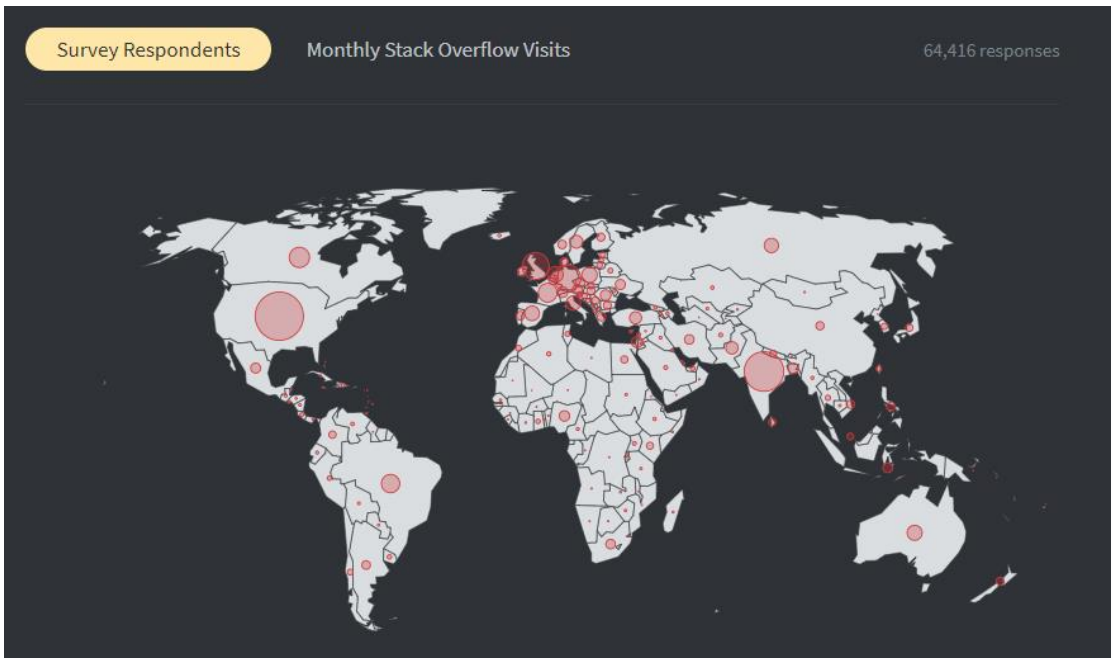
```
In [45]: import matplotlib.pyplot as plt
         df['ratings'].value_counts().plot(kind='bar',color='Purple')
         plt.xlabel('Ratings')
         plt.ylabel('Number of Reviewers')
         plt.grid(color='b',alpha=0.4)
```



You can see total no. Of reviews, ratings given by the users and then can analyze that 80% of the people have given 4/5 star rating, from which you can say that one can buy this phone.

Exploratory Data Analysis
(Stack Overflow Developer Survey)

For almost a decade, Stack Overflow's annual Developer Survey held the honor of being the largest survey of people who code around the world. This year, rather than aiming to be the biggest, it set out to make our survey more representative of the diversity of programmers worldwide. That said, the survey is still big. This year's survey was taken by nearly 65,000 people. In efforts to reach beyond the Stack Overflow network and seek representation from a greater diversity of coders, they advertised the survey less on our own channels than in previous years and sought ways to earn responses from those who may not frequent our sites.



Now its time to load the data set and perform analysis on the same.

```
In [1]: import pandas as pd

In [2]: #Loading Dataset
        survey_raw_df=pd.read_csv('survey_results_public.csv')
```

After loading the data set we first see what columns are there in the data set. So that we can have an idea of what type of data we are working with.

```
In [3]: survey_raw_df
Out[3]:
```

	Respondent	MainBranch	Hobbyist	Age	Age1stCode	CompFreq	CompTotal	ConvertedComp	Country	CurrencyDesc	...	SurveyEase	SurveyLength
0	1	I am a developer by profession	Yes	NaN	13	Monthly	NaN	NaN	Germany	European Euro	...	Neither easy nor difficult	Appropriate in length
1	2	I am a developer by profession	No	NaN	19	NaN	NaN	NaN	United Kingdom	Pound sterling	...	NaN	NaN
2	3	I code primarily as a hobby	Yes	NaN	15	NaN	NaN	NaN	Russian Federation	NaN	...	Neither easy nor difficult	Appropriate in length
64456	64858	NaN	Yes	NaN	16	NaN	NaN	NaN	United States	NaN	...	NaN	NaN
64457	64867	NaN	Yes	NaN	NaN	NaN	NaN	NaN	Morocco	NaN	...	NaN	NaN
64458	64898	NaN	Yes	NaN	NaN	NaN	NaN	NaN	Viet Nam	NaN	...	NaN	NaN
64459	64925	NaN	Yes	NaN	NaN	NaN	NaN	NaN	Poland	NaN	...	NaN	NaN
64460	65112	NaN	Yes	NaN	NaN	NaN	NaN	NaN	Spain	NaN	...	NaN	NaN

64461 rows × 61 columns

So as you can see we have 64461 rows and 61 columns, which itself make it a huge data of developers around the world.

```
In [4]: survey_raw_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64461 entries, 0 to 64460
Data columns (total 61 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Respondent                    64461 non-null  int64
 1   MainBranch                    64162 non-null  object
 2   Hobbyist                      64416 non-null  object
 3   Age                           45446 non-null  float64
 4   Age1stCode                    57900 non-null  object
 5   CompFreq                      40069 non-null  object
 6   CompTotal                     34826 non-null  float64
 7   ConvertedComp                 34756 non-null  float64
 8   Country                       64072 non-null  object
 9   CurrencyDesc                  45472 non-null  object
 10  CurrencySymbol                45472 non-null  object
 11  DatabaseDesireNextYear        44070 non-null  object
 12  DatabaseWorkedWith            49537 non-null  object
 13  DevType                       49370 non-null  object
 14  EdLevel                       57431 non-null  object
 15  Employment                    63854 non-null  object
 16  Ethnicity                     45948 non-null  object
 17  Gender                        50557 non-null  object
 18  JobFactors                    49349 non-null  object
 19  JobSat                        45194 non-null  object
 20  JobSeek                       51727 non-null  object
 21  LanguageDesireNextYear        54113 non-null  object
```

```
 34  NEWOnboardGood                42623 non-null  object
 35  NEWOtherComms                 57205 non-null  object
 36  NEWOvertime                   43231 non-null  object
 37  NEWPurchaseResearch           37321 non-null  object
 38  NEWPurpleLink                 54803 non-null  object
 39  NEWSOSites                    58275 non-null  object
 40  NEWStuck                      54983 non-null  object
 41  OpSys                         56228 non-null  object
 42  OrgSize                       44334 non-null  object
 43  PlatformDesireNextYear        50605 non-null  object
 44  PlatformWorkedWith            53843 non-null  object
 45  PurchaseWhat                  39364 non-null  object
 46  Sexuality                     43992 non-null  object
 47  SOAccount                     56805 non-null  object
 48  SOComm                        56476 non-null  object
 49  SOPartFreq                    46792 non-null  object
 50  SOVisitFreq                   56970 non-null  object
 51  SurveyEase                    51802 non-null  object
 52  SurveyLength                  51701 non-null  object
 53  Trans                         49345 non-null  object
 54  UndergradMajor                50995 non-null  object
 55  WebframeDesireNextYear        40024 non-null  object
 56  WebframeWorkedWith            42279 non-null  object
 57  WelcomeChange                 52683 non-null  object
 58  WorkWeekHrs                   41151 non-null  float64
 59  YearsCode                     57684 non-null  object
 60  YearsCodePro                  46349 non-null  object
dtypes: float64(4), int64(1), object(56)
memory usage: 30.0+ MB
```

As there are a total of 64461 people who have participated in this survey. But as you can see there are numbers in 46k, 57k even 30k which means that there are many missing values in the data set. And we can either remove fix these values or remove the missing values.

```
In [5]: survey_raw_df.columns

Out[5]: Index(['Respondent', 'MainBranch', 'Hobbyist', 'Age', 'Age1stCode', 'CompFreq',
       'CompTotal', 'ConvertedComp', 'Country', 'CurrencyDesc',
       'CurrencySymbol', 'DatabaseDesireNextYear', 'DatabaseWorkedWith',
       'DevType', 'EdLevel', 'Employment', 'Ethnicity', 'Gender', 'JobFactors',
       'JobSat', 'JobSeek', 'LanguageDesireNextYear', 'LanguageWorkedWith',
       'MiscTechDesireNextYear', 'MiscTechWorkedWith',
       'NEWCollabToolsDesireNextYear', 'NEWCollabToolsWorkedWith', 'NEWDevOps',
       'NEWDevOpsImpt', 'NEWEdImpt', 'NEWJobHunt', 'NEWJobHuntResearch',
       'NEWLearn', 'NEWOffTopic', 'NEWOnboardGood', 'NEWOtherComms',
       'NEWOvertime', 'NEWPurchaseResearch', 'NEWPurpleLink', 'NEWSOSites',
       'NEWStuck', 'OpSys', 'OrgSize', 'PlatformDesireNextYear',
       'PlatformWorkedWith', 'PurchaseWhat', 'Sexuality', 'SOAccount',
       'SOComm', 'SOPartFreq', 'SOVisitFreq', 'SurveyEase', 'SurveyLength',
       'Trans', 'UndergradMajor', 'WebframeDesireNextYear',
       'WebframeWorkedWith', 'WelcomeChange', 'WorkWeekHrs', 'YearsCode',
       'YearsCodePro'],
      dtype='object')
```

These are the number of columns in the survey_raw_df, the length of which is 61.

We have given another CSV file named ‘survey_results_schema’ which contains the questions regarding the data. These question are based on the columns, and each column has a question. There are total of 61 questions.

We try to find out answers to these questions as much as possible.

```
In [7]: schema_fname=pd.read_csv('survey_results_schema.csv',index_col='Column')

In [8]: schema_raw=schema_fname['QuestionText']

In [9]: schema_raw

Out[9]: Column
        Respondent            Randomized respondent ID number (not in order ...
        MainBranch            Which of the following options best describes ...
        Hobbyist                                        Do you code as a hobby?
        Age                   What is your age (in years)? If you prefer not...
        Age1stCode            At what age did you write your first line of c...
                                                   ...
        WebframeWorkedWith    Which web frameworks have you done extensive d...
        WelcomeChange         Compared to last year, how welcome do you feel...
        WorkWeekHrs           On average, how many hours per week do you wor...
        YearsCode             Including any education, how many years have y...
        YearsCodePro          NOT including education, how many years have y...
        Name: QuestionText, Length: 61, dtype: object
```

Data Preparation & Cleaning

While the survey responses contain a wealth of information, we'll limit our analysis to the following areas:

- Demographics of the survey respondents & the global programming community
- Distribution of programming skills, experiences and preferences
- Employment related information, preferences and opinions

```
In [10]: selected_columns=[
             #Demographics
             'Country','Age','Gender','EdLevel', 'UndergradMajor',
             #Programming Experience
             'YearsCodePro', 'Hobbyist', 'Age1stCode', 'LanguageDesireNextYear', 'YearsCode','LanguageWorkedWith','NEWLearn','NEWStuck',
             #Employment
             'Employment','DevType','WorkWeekHrs','JobSat','JobFactors', 'NEWOvertime','NEWEdImpt'
         ]

In [11]: len(selected_columns)

Out[11]: 20
```

Now I extract copy of the data from these columns into a new data frame survey_df, which we can continue to modify without affecting the original data frame.

```
In [12]: survey_df=survey_raw_df[selected_columns].copy()

In [13]: schema=schema_raw[selected_columns]
```

Here we have a new data frame named ‘survey_df ’ which have only 20 column.

```
In [14]: survey_df
```

Out[14]:

	Country	Age	Gender	EdLevel	UndergradMajor	YearsCodePro	Hobbyist	Age1stCode	LanguageDesireNextYear	Yea
0	Germany	NaN	Man	Master's degree (M.A., M.S., M.Eng., MBA, etc.)	Computer science, computer engineering, or sof...	27	Yes	13	C#;HTML/CSS;JavaScript	
1	United Kingdom	NaN	NaN	Bachelor's degree (B.A., B.S., B.Eng., etc.)	Computer science, computer engineering, or sof...	4	No	19	Python;Swift	
2	Russian Federation	NaN	NaN	NaN	NaN	NaN	Yes	15	Objective-C;Python;Swift	
3	Albania	25.0	Man	Master's degree (M.A., M.S., M.Eng., MBA, etc.)	Computer science, computer engineering, or sof...	4	Yes	18	NaN	
4	United States	31.0	Man	Bachelor's degree (B.A., B.S., B.Eng., etc.)	Computer science, computer engineering, or sof...	8	Yes	16	Java;Ruby;Scala	
...	...	...	...	...	...	...	...	...	...	
	United			Master's degree	Computer science,	Less than 1				

And based on these 20 columns we have questions accordingly.

```
In [15]: schema
```

```
Out[15]: Column
Country                                                 Where do you live?
Age                       What is your age (in years)? If you prefer not...
Gender                    Which of the following describe you, if any? P...
EdLevel                   Which of the following best describes the high...
UndergradMajor                       What was your primary field of study?
YearsCodePro              NOT including education, how many years have y...
Hobbyist                                            Do you code as a hobby?
Age1stCode                At what age did you write your first line of c...
LanguageDesireNextYear    Which programming, scripting, and markup langu...
YearsCode                 Including any education, how many years have y...
LanguageWorkedWith        Which programming, scripting, and markup langu...
NEWLearn                  How frequently do you learn a new language or ...
NEWStuck                  What do you do when you get stuck on a problem...
Employment                Which of the following best describes your cur...
DevType                   Which of the following describe you? Please se...
WorkWeekHrs               On average, how many hours per week do you wor...
JobSat                    How satisfied are you with your current job? (...
JobFactors                Imagine that you are deciding between two job ...
NEWOvertime               How often do you work overtime or beyond the f...
NEWEdImpt                 How important is a formal education, such as a...
Name: QuestionText, dtype: object
```

```
In [16]: len(schema)

Out[16]: 20
```

```
In [17]: survey_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64461 entries, 0 to 64460
Data columns (total 20 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Country                 64072 non-null  object
 1   Age                     45446 non-null  float64
 2   Gender                  50557 non-null  object
 3   EdLevel                 57431 non-null  object
 4   UndergradMajor          50995 non-null  object
 5   YearsCodePro            46349 non-null  object
 6   Hobbyist                64416 non-null  object
 7   Age1stCode              57900 non-null  object
 8   LanguageDesireNextYear  54113 non-null  object
 9   YearsCode               57684 non-null  object
 10  LanguageWorkedWith      57378 non-null  object
 11  NEWLearn                56156 non-null  object
 12  NEWStuck                54983 non-null  object
 13  Employment              63854 non-null  object
 14  DevType                 49370 non-null  object
 15  WorkWeekHrs             41151 non-null  float64
 16  JobSat                  45194 non-null  object
 17  JobFactors              49349 non-null  object
 18  NEWOvertime             43231 non-null  object
 19  NEWEdImpt               48465 non-null  object
dtypes: float64(2), object(18)
memory usage: 9.8+ MB
```

Most columns have the data type object, either because they contain values of different types, or they contain empty values, which are represented by Nan. Only two of the columns were detected as numeric columns (Age and WorkWeekHrs), even though there are a few other columns which have mostly numeric values. To make my analysis easier, I will convert some other columns into numeric data types.

```
In [18]: schema.Age1stCode

Out[18]: 'At what age did you write your first line of code or program? (e.g., webpage, Hello World, Scratch project)'

In [19]: survey_df.Age1stCode.unique()

Out[19]: array(['13', '19', '15', '18', '16', '14', '12', '20', '42', '8', '25',
       '22', '30', '17', '21', '10', '46', '9', '7', '11', '6', nan, '31',
       '29', '5', 'Younger than 5 years', '28', '38', '23', '27', '41',
       '24', '53', '26', '35', '32', '40', '33', '36', '54', '48', '56',
       '45', '44', '34', 'Older than 85', '39', '51', '68', '50', '37',
       '47', '43', '52', '85', '64', '55', '58', '49', '76', '72', '73',
       '83', '63'], dtype=object)
```

To help analyze our data easily and to perform computations I have converted object data types to numeric.

```
In [24]: #converting into numeric values
         survey_df['Age1stCode']=pd.to_numeric(survey_df.Age1stCode, errors='coerce')
         survey_df['YearsCode']=pd.to_numeric(survey_df.YearsCode, errors='coerce')
         survey_df['YearsCodePro']=pd.to_numeric(survey_df.YearsCodePro, errors='coerce')
```

Basic Statistics

```
In [25]: survey_df.describe()
```

Out[25]:

	Age	YearsCodePro	Age1stCode	YearsCode	WorkWeekHrs
count	45446.000000	44133.000000	57473.000000	56784.000000	41151.000000
mean	30.834111	8.869667	15.476572	12.782051	40.782174
std	9.585392	7.759961	5.114081	9.490657	17.816383
min	1.000000	1.000000	5.000000	1.000000	1.000000
25%	24.000000	3.000000	12.000000	6.000000	40.000000
50%	29.000000	6.000000	15.000000	10.000000	40.000000
75%	35.000000	12.000000	18.000000	17.000000	44.000000
max	279.000000	50.000000	85.000000	50.000000	475.000000

There seems to be a problem with the age column , as the minimum value is 1 and max value is 279. This is a common issues with surveys: responses may contain invalid values due to accidental or intentional errors while responding. A simple fix would be ignore the rows where the values in the age column is higher than 100 years or lower than 10 years as invalid survey responses.

Finding and removing outliers:

```
In [26]: #outlier
         survey_df[survey_df['Age']>100].index
Out[26]: Int64Index([14375], dtype='int64')
```

```
In [27]: #outliers
         survey_df[survey_df['Age']<10].index
Out[27]: Int64Index([8793, 11600, 12271, 20042, 25061, 26952, 54687, 58292, 64383], dtype='int64')
```

```
In [28]: #Removing Outliers
         survey_df.drop(survey_df[survey_df['Age']>100].index,inplace=True)
         survey_df.drop(survey_df[survey_df['Age']<10].index,inplace=True)
```

The same hold true for WorkWeekHrs. Lets ignore entries where the value for the column is higher than 140 hours (~20 hours per day).

```
In [29]: survey_df.drop(survey_df[survey_df['WorkWeekHrs']>140].index,inplace=True)
```

The gender column also allows picking multiple options, but to simplify our analysis, we'll remove values containing multiple options.

```
In [30]: schema.Gender
Out[30]: 'Which of the following describe you, if any? Please check all that apply. If you prefer not to answer, you may leave this ques
         tion blank.'
```

```
In [31]: survey_df['Gender'].value_counts()

Out[31]: Man                                                        45895
         Woman                                                       3835
         Non-binary, genderqueer, or gender non-conforming            385
         Man;Non-binary, genderqueer, or gender non-conforming        121
         Woman;Non-binary, genderqueer, or gender non-conforming       92
         Woman;Man                                                     73
         Woman;Man;Non-binary, genderqueer, or gender non-conforming   25
         Name: Gender, dtype: int64
```

```
In [32]: import numpy as np

In [33]: survey_df.where(~(survey_df['Gender'].str.contains(';',na=False)),np.nan,inplace=True)

In [34]: survey_df['Gender'].value_counts()

Out[34]: Man                                            45895
         Woman                                           3835
         Non-binary, genderqueer, or gender non-conforming  385
         Name: Gender, dtype: int64
```

I've now cleaned up and prepared the data set for analysis. Let's now look at the sample of rows from the data frame.

```
In [35]: survey_df.sample(10)

Out[35]:
```

	Country	Age	Gender	EdLevel	UndergradMajor	YearsCodePro	Hobbyist	Age1stCode	LanguageDesireNextYear	Ye
8033	Iran	30.0	Man	Master's degree (M.A., M.S., M.Eng., MBA, etc.)	Another engineering discipline (such as civil,...	4.0	Yes	15.0	JavaScript	
52899	United States	32.0	Woman	Master's degree (M.A., M.S., M.Eng., MBA, etc.)	Another engineering discipline (such as civil,...	5.0	No	15.0	Bash/Shell/PowerShell;Python;SQL	
2326	Canada	27.0	Man	Bachelor's degree (B.A., B.S., B.Eng., etc.)	Computer science, computer engineering, or sof...	5.0	No	10.0	Go;Rust	
18252	Malta	27.0	Man	Master's degree (M.A., M.S., M.Eng., MBA,	Computer science, computer engineering, or	7.0	Yes	7.0	Bash/Shell/PowerShell;Dart;Go;HTML/CSS;Java;Ja...	

Exploratory Analysis and Visualization

Before we ask interesting questions about the survey responses, it would help to understand what the demographics i.e. country, age, gender, education level, employment level etc. of the respondents looks like. Its important to explore these variables in order to understand how representative the survey is of the worldwide programming community, as a survey of this scale generally tends to have some selection bias.

```
In [36]: import seaborn as sns
         import matplotlib
         import matplotlib.pyplot as plt
         %matplotlib inline

         sns.set_style('darkgrid')
         matplotlib.rcParams['font.size']=14
         matplotlib.rcParams['figure.figsize']=(9,5)
         matplotlib.rcParams['figure.facecolor']='#00000000'
```

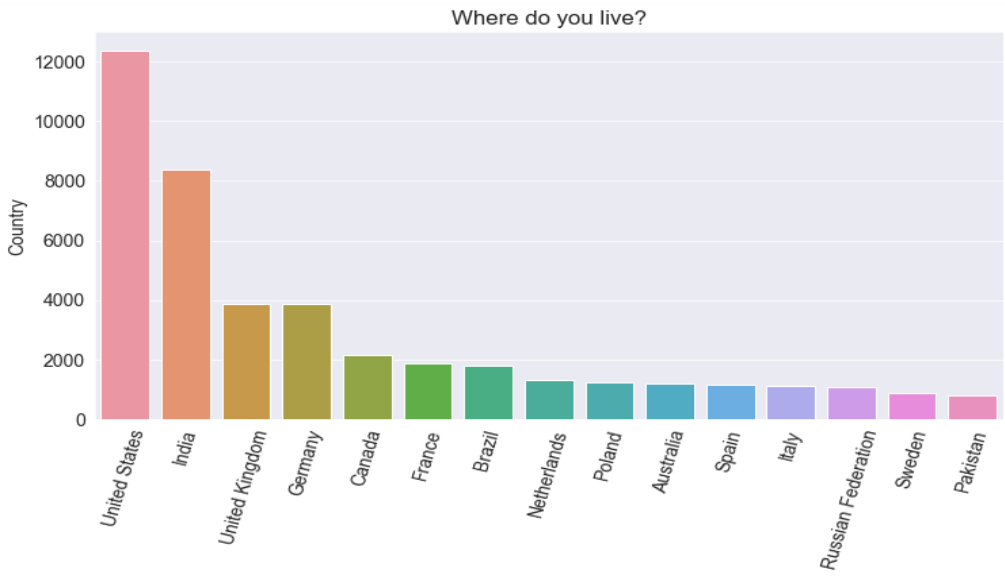
Country

Counting total number of countries from which there are responses in the survey and plotting top 15 countries with highest response.

```
In [37]: schema['Country']
Out[37]: 'Where do you live?'

In [38]: survey_df['Country'].nunique()
Out[38]: 183

In [39]: top_countries=survey_df['Country'].value_counts().head(15)
         print(top_countries)

United States         12371
India                  8364
United Kingdom         3881
Germany                3864
Canada                 2175
France                 1884
Brazil                 1804
Netherlands            1332
Poland                 1259
Australia              1199
Spain                  1157
Italy                  1115
Russian Federation     1085
Sweden                  879
Pakistan                802
Name: Country, dtype: int64
```

We can visualize this information using bar charts.

```
In [40]: plt.figure(figsize=(12,6))
         plt.xticks(rotation=75)
         plt.title(schema.Country)
         sns.barplot(top_countries.index,top_countries)
         plt.show()
```



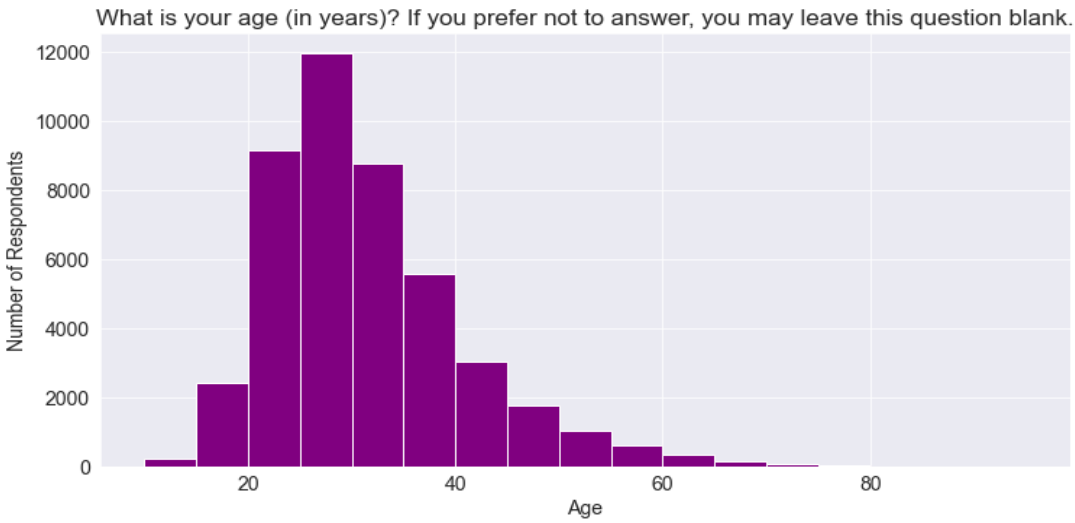
It appears that a high number of respondents are from USA and India - which one might expect since these countries have the highest population (apart from China), and since the survey is in English, which is the common language used by professionals in US, India & UK. We can already see that the survey may not be representative of the entire programming community - especially from non-English speaking countries.

Age

The distribution of the age of respondents is another important factor to look at, and we can use a histogram to visualize it.

```
In [41]: schema.Age
Out[41]: 'What is your age (in years)? If you prefer not to answer, you may leave this question blank.'
```

```
In [42]: plt.figure(figsize=(12,6))
         plt.title(schema.Age)
         plt.xlabel('Age')
         plt.ylabel('Number of Respondents')
         plt.hist(survey_df.Age,bins=np.arange(10,100,5),color='purple')
         plt.show()
```



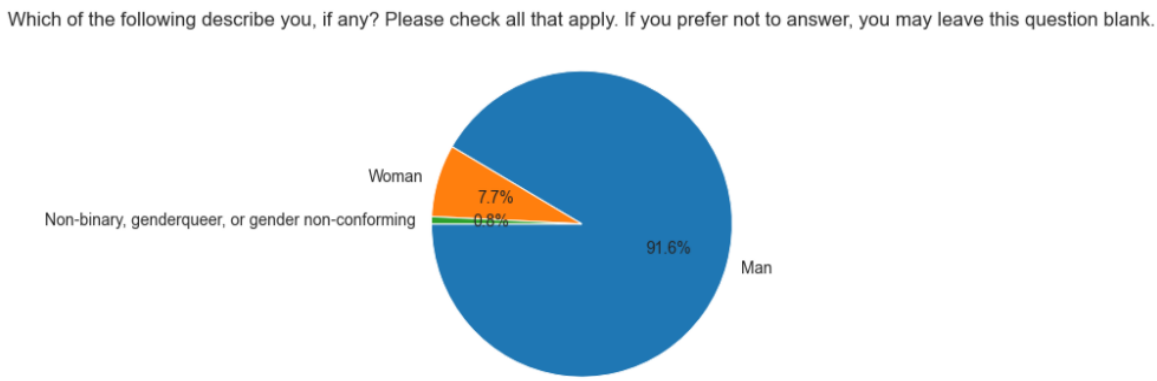
It appears that a large percentage of respondents are in the age range of 20-45, which is somewhat representative of the programming community in general, as a lot of people has taken up computer as a field of study or profession in last 20 years.

Gender

Lets look at the distribution of responses of gender . It is a well known fact that women and non-binary gender are under representative in the programming community, so we might expect to see a skewed distribution here.

```
In [43]: schema.Gender

Out[43]: 'Which of the following describe you, if any? Please check all that apply. If you prefer not to answer, you may leave this ques
         tion blank.'

In [44]: gender_counts=survey_df['Gender'].value_counts()
```

```
In [45]: gender_counts

Out[45]: Man                                                45895
         Woman                                               3835
         Non-binary, genderqueer, or gender non-conforming    385
         Name: Gender, dtype: int64

In [46]: plt.figure(figsize=(12,6))
         plt.title(schema.Gender)
         plt.pie(gender_counts,labels=gender_counts.index,autopct='%1.1f%%',startangle=180)
         plt.show()
```



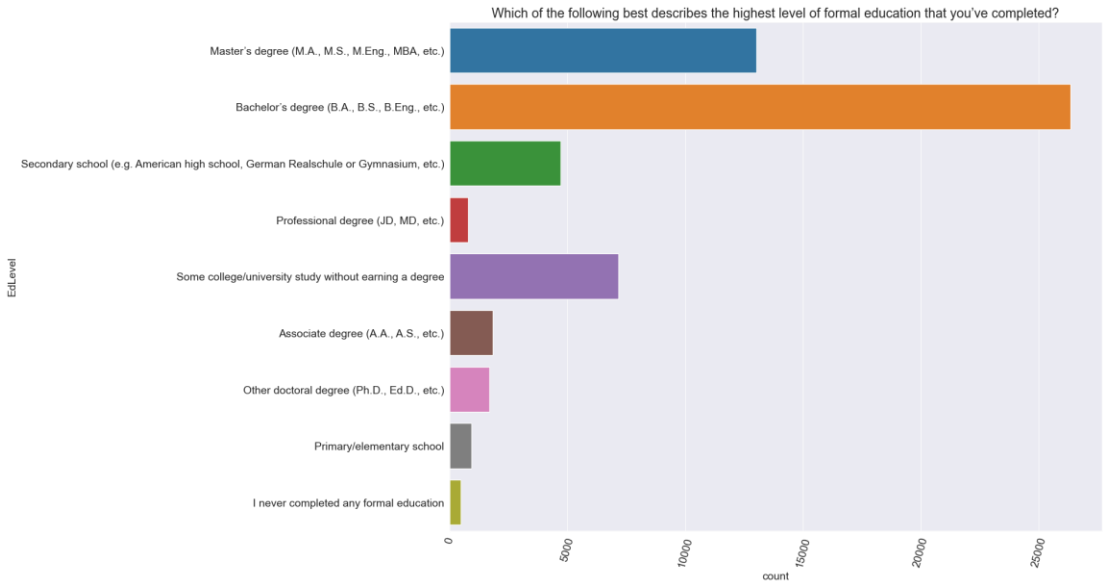
Only about 8% of survey respondents who have answered the question identify as women or non-binary genders in the programming community - which is estimated to be around 12%.

Education Level

Formal education in computer science is often considered an important requirement of becoming a programmer. Lets see if this indeed the case, especially since there are many free resources & tutorials available online to learn programming. We will use a horizontal bar plot to compare education levels of respondents.

```
In [47]: schema['EdLevel']
Out[47]: 'Which of the following best describes the highest level of formal education that you've completed?'
```

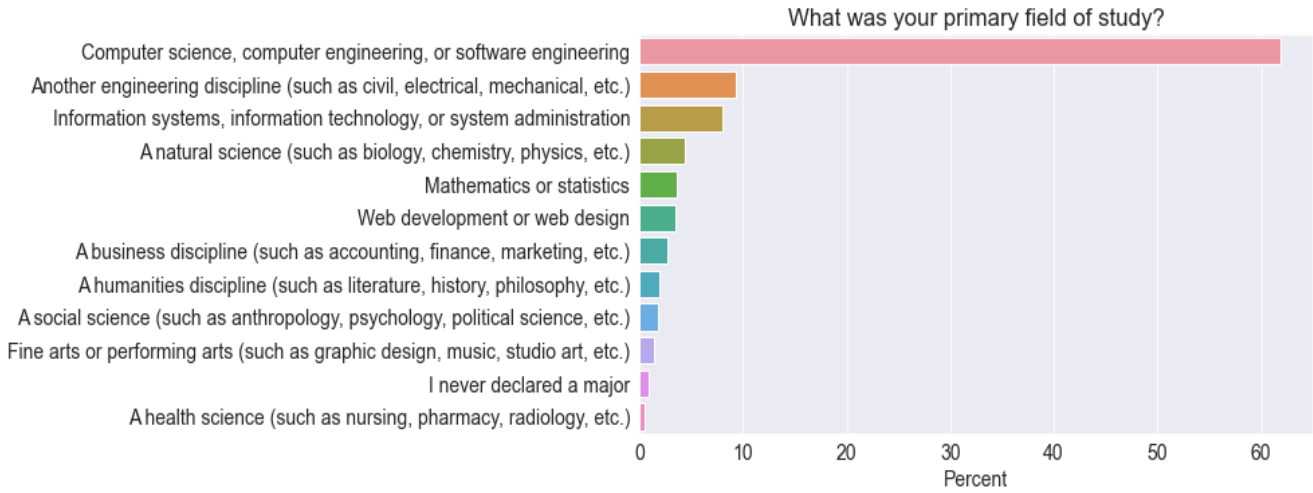
```
In [48]: plt.figure(figsize=(15,12),dpi=100)
         sns.countplot(y=survey_df['EdLevel'])
         plt.xticks(rotation=75)
         plt.title(schema['EdLevel'])
         plt.show()
```



It appears that well over half of the respondents holds a bachelor's or master's degree, so most programmers definitely seem to have some college education, although it's not clear from this graph alone if they hold a degree in computer science.

Let's also plot undergraduate major's, but this time we'll convert the numbers into percentages, and sort it by percentage values to make it easier to visualize the order.

```
In [49]: schema.UndergradMajor
Out[49]: 'What was your primary field of study?'

In [50]: Undergrad_pct=survey_df.UndergradMajor.value_counts()*100/survey_df.UndergradMajor.count()

In [51]: sns.barplot(Undergrad_pct,Undergrad_pct.index)
         plt.title(schema.UndergradMajor)
         plt.xlabel('Percent')
         plt.show()
```



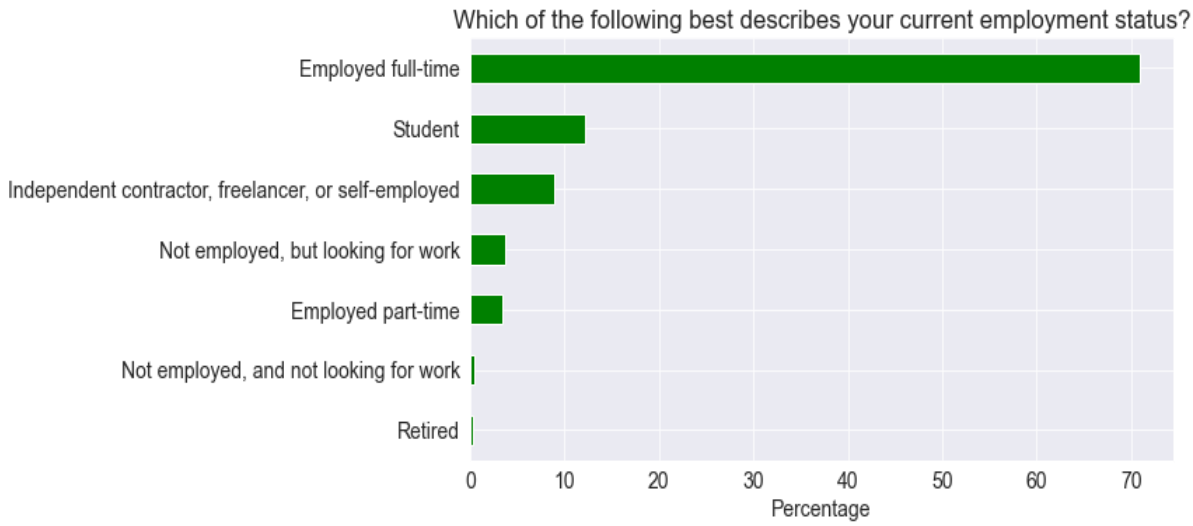
It turns that 40% of programmers holding a college degree have a field of study other than computer science - which is very encouraging. This seems to suggest that while college education is helpful in general, you do not need to pursue a major in computer science to become a successful programmer.

Employment¶

Freelancing or contract work is a common choice among programmers, so it would be interesting to compare the breakdown between full time, part time & freelance work. Let's visualize the data from Employment column.

```
In [52]: schema.Employment
Out[52]: 'Which of the following best describes your current employment status?'
```

```
In [53]: (survey_df.Employment.value_counts(normalize=True,ascending=True)*100).plot(kind='barh',color='g')
         plt.title(schema.Employment)
         plt.xlabel('Percentage')
Out[53]: Text(0.5, 0, 'Percentage')
```



It appears that close to 10% of respondents are employed part time or as freelancers.

Developer Type

DevType contains the information about the roles held by respondents.

```
In [54]: schema.DevType
Out[54]: 'Which of the following describe you? Please select all that apply.'
```

```
In [55]: survey_df['DevType'].value_counts()
Out[55]: Developer, full-stack
         4396
         Developer, back-end
         3056
         Developer, back-end;Developer, front-end;Developer, full-stack
         2214
         Developer, back-end;Developer, full-stack
         1465
         Developer, front-end
         1390

         ...
         Academic researcher;Database administrator;Developer, back-end;Developer, desktop or enterprise applications;Developer, front-e
         nd;Developer, full-stack;Developer, mobile;DevOps specialist                                                          1
         Data or business analyst;Data scientist or machine learning specialist;Database administrator;Developer, back-end;Developer, de
         sktop or enterprise applications;Developer, full-stack;Engineer, data;Product manager;Scientist       1
         Database administrator;Developer, back-end;Developer, full-stack;Engineer, data;Engineering manager;Product manager
         1
         Data or business analyst;Database administrator;Designer;Developer, desktop or enterprise applications;DevOps specialist;Engine
         er, data                                                                                                              1
         Data or business analyst;Developer, full-stack;DevOps specialist;Engineer, site reliability;Engineering manager
         1
         Name: DevType, Length: 8213, dtype: int64
```

Lets define a helper function which turns a column containing lists of values into a data frame with one column for each possible option.

```
In [56]: def split_multicolumn(col_series):
             result_df=col_series.to_frame()
             options=[]
             #Iterate over columns
             for idx, value in col_series[col_series.notnull()].iteritems():
                 #Break each value into list of options
                 for option in value.split(';'):
                     #Add the option as a column to result
                     if not option in result_df.columns:
                         options.append(option)
                         result_df[option]=False
                     #Mark the value in the option column as True
                     result_df.at[idx,option]=True
             return result_df[options]
```

```
In [57]: dev_type_df=split_multicolumn(survey_df.DevType)
```

```
In [58]: dev_type_df
Out[58]:
```

	Developer, desktop or enterprise applications	Developer, full-stack	Developer, mobile	Designer	Developer, front-end	Developer, back-end	Developer, QA or test	DevOps specialist	Developer, game or graphics	Database administrator	...	System administrator	Engineering manager	F m
0	True	True	False	False	False	False	False	False	False	False	...	False	False	
1	False	True	True	False	False	False	False	False	False	False	...	False	False	
2	False	False	False	False	False	False	False	False	False	False	...	False	False	
3	False	False	False	False	False	False	False	False	False	False	...	False	False	
4	False	False	False	False	False	False	False	False	False	False	...	False	False	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
64456	False	False	False	False	False	False	False	False	False	False	...	False	False	
64457	False	False	False	False	False	False	False	False	False	False	...	False	False	
64458	False	False	False	False	False	False	False	False	False	False	...	False	False	
64459	False	False	False	False	False	False	False	False	False	False	...	False	False	
64460	False	False	False	False	False	False	False	False	False	False	...	False	False	

64306 rows × 23 columns

The dev_type_df has one column for each option that can be selected as a respondent. If a responded has selected the option, the value in the column is True, otherwise it is False. We can now use the column wise totals to identify the most common roles.

```
In [59]: dev_type_totals=dev_type_df.sum().sort_values(ascending=False)
```

```
In [60]: dev_type_totals
Out[60]: Developer, back-end                              26996
         Developer, full-stack                            26915
         Developer, front-end                             18128
         Developer, desktop or enterprise applications    11687
         Developer, mobile                                 9406
         DevOps specialist                                 5915
         Database administrator                            5658
         Designer                                          5262
         System administrator                              5185
         Developer, embedded applications or devices       4701
         Data or business analyst                          3970
         Data scientist or machine learning specialist     3939
         Developer, QA or test                             3893
         Engineer, data                                    3700
         Academic researcher                               3502
         Educator                                          2895
         Developer, game or graphics                       2751
         Engineering manager                               2699
         Product manager                                   2471
         Scientist                                         2060
         Engineer, site reliability                        1921
         Senior executive/VP                               1292
         Marketing or sales professional                    625
         dtype: int64
```

As one might expect, the most common roles include 'Developer' in the name.

Asking and Answering Questions

We have already gained several insights about the respondents and the programming community in general, simply by exploring individual columns of the data set. Let's ask some specific questions, and try to answer them using data frame operations and interesting visualizations.

Q. What were the most popular languages in 2020?

To answer this, I use LanguageWorkedWith column.

```
In [61]: schema.LanguageWorkedWith

Out[61]: 'Which programming, scripting, and markup languages have you done extensive development work in over the past year, and which d
         o you want to work in over the next year? (If you both worked with the language and want to continue to do so, please check bot
         h boxes in that row.)'
```

```
In [63]: survey_df.LanguageWorkedWith

Out[63]: 0                                  C#;HTML/CSS;JavaScript
         1                                        JavaScript;Swift
         2                                 Objective-C;Python;Swift
         3                                                      NaN
         4                                      HTML/CSS;Ruby;SQL
                                       ...
         64456                                                  NaN
         64457    Assembly;Bash/Shell/PowerShell;C;C#;C++;Dart;G...
         64458                                                  NaN
         64459                                             HTML/CSS
         64460                        C#;HTML/CSS;Java;JavaScript;SQL
         Name: LanguageWorkedWith, Length: 64306, dtype: object
```

First, We'll split this column into a data frame containing a column of each languages listed in the options.

```
In [64]: languages_worked_df=split_multicolumn(survey_df.LanguageWorkedWith)
```

```
In [65]: languages_worked_df

Out[65]:
```

	C#	HTML/CSS	JavaScript	Swift	Objective-C	Python	Ruby	SQL	Java	PHP	...	VBA	Perl	Scala	C++	Go	Haskell	Rust	Dart	Julia	As
0	True	True	True	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	
1	False	False	True	True	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	
2	False	False	False	True	True	True	False	False	False	False	...	False	False	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	
4	False	True	False	False	False	False	True	True	False	False	...	False	False	False	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
64456	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	
64457	True	True	True	True	True	True	True	True	True	True	...	True	True	True	True	True	True	True	True	True	
64458	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	
64459	False	True	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	
64460	True	True	True	False	False	False	False	True	True	False	...	False	False	False	False	False	False	False	False	False	

64306 rows × 25 columns

It appears that a total of 25 languages were included among the options. Lets aggregate these to identify the percentage of respondents who selected each language.

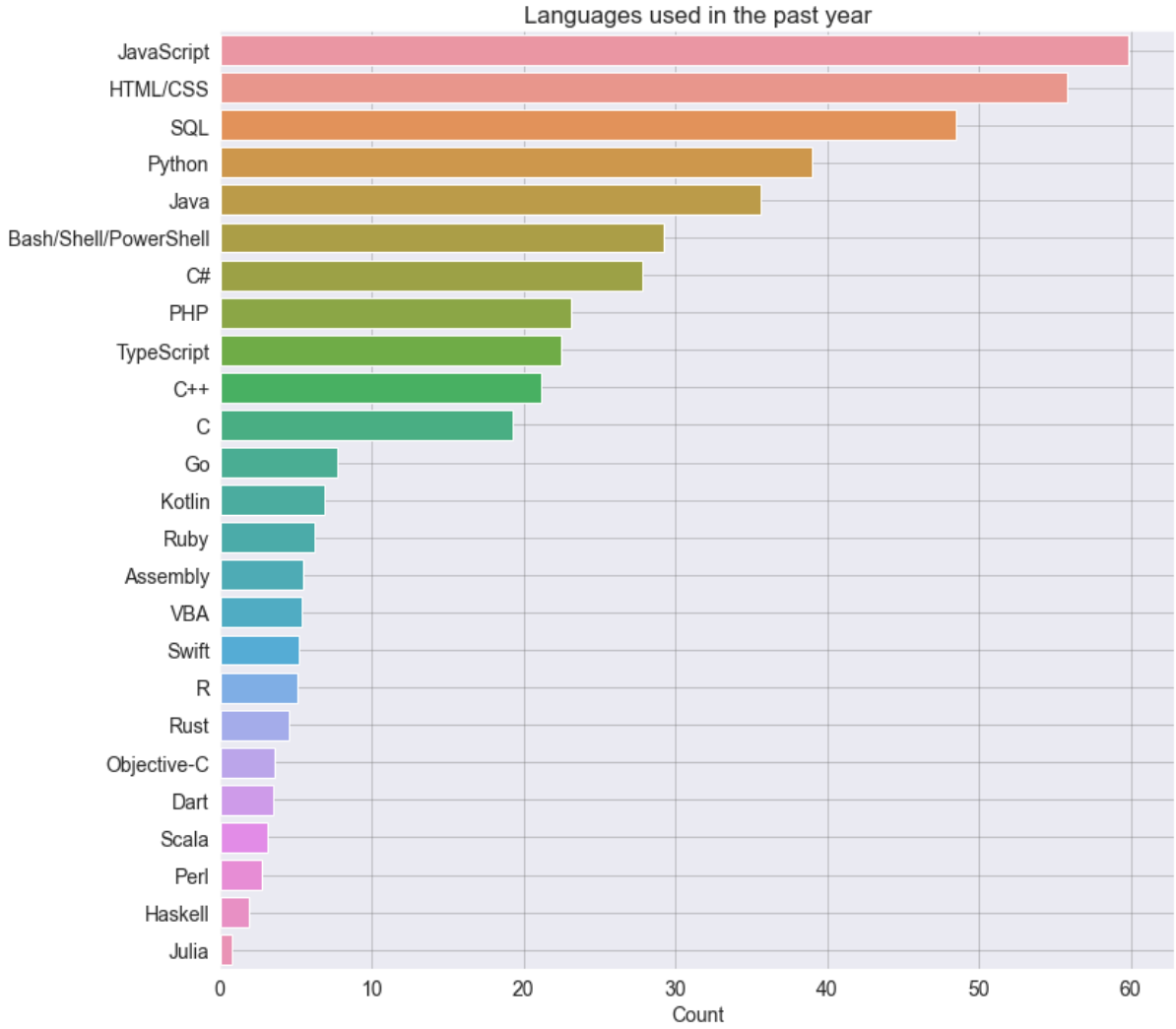
```
In [66]: languages_worked_percentages=languages_worked_df.mean().sort_values(ascending=False)*100
```

Languages worked percentage

```
In [67]: languages_worked_percentages

Out[67]: JavaScript               59.893323
         HTML/CSS                 55.801947
         SQL                      48.444935
         Python                   39.001026
         Java                     35.618760
         Bash/Shell/PowerShell    29.239884
         C#                       27.803004
         PHP                      23.130035
         TypeScript               22.461357
         C++                      21.114670
         C                        19.236152
         Go                        7.758219
         Kotlin                    6.887382
         Ruby                      6.229590
         Assembly                  5.447392
         VBA                       5.394520
         Swift                     5.226573
         R                         5.064846
         Rust                      4.498803
         Objective-C               3.603085
         Dart                      3.517557
         Scala                     3.150561
         Perl                      2.757130
         Haskell                   1.861413
         Julia                     0.782198
         dtype: float64
```

Plotting:

```
In [68]: plt.figure(figsize=(12,12))
         sns.barplot(languages_worked_percentages,languages_worked_percentages.index)
         plt.title('Languages used in the past year')
         plt.xlabel('Count')
         plt.grid(color='gray',alpha=0.5)
         plt.show()
```



Perhaps not surprisingly, Java script & HTML/CSS comes out at the top as web development is one of the most sought skills today and it also happens to be one of the easiest to get started with. SQL in necessary for working with relational databases, so its no surprise that most programmers work with SQL on regular basis. For other form of development, Python seems to be a popular choice, beating out java, which was the industry standard for server & application development for over 2 decades.

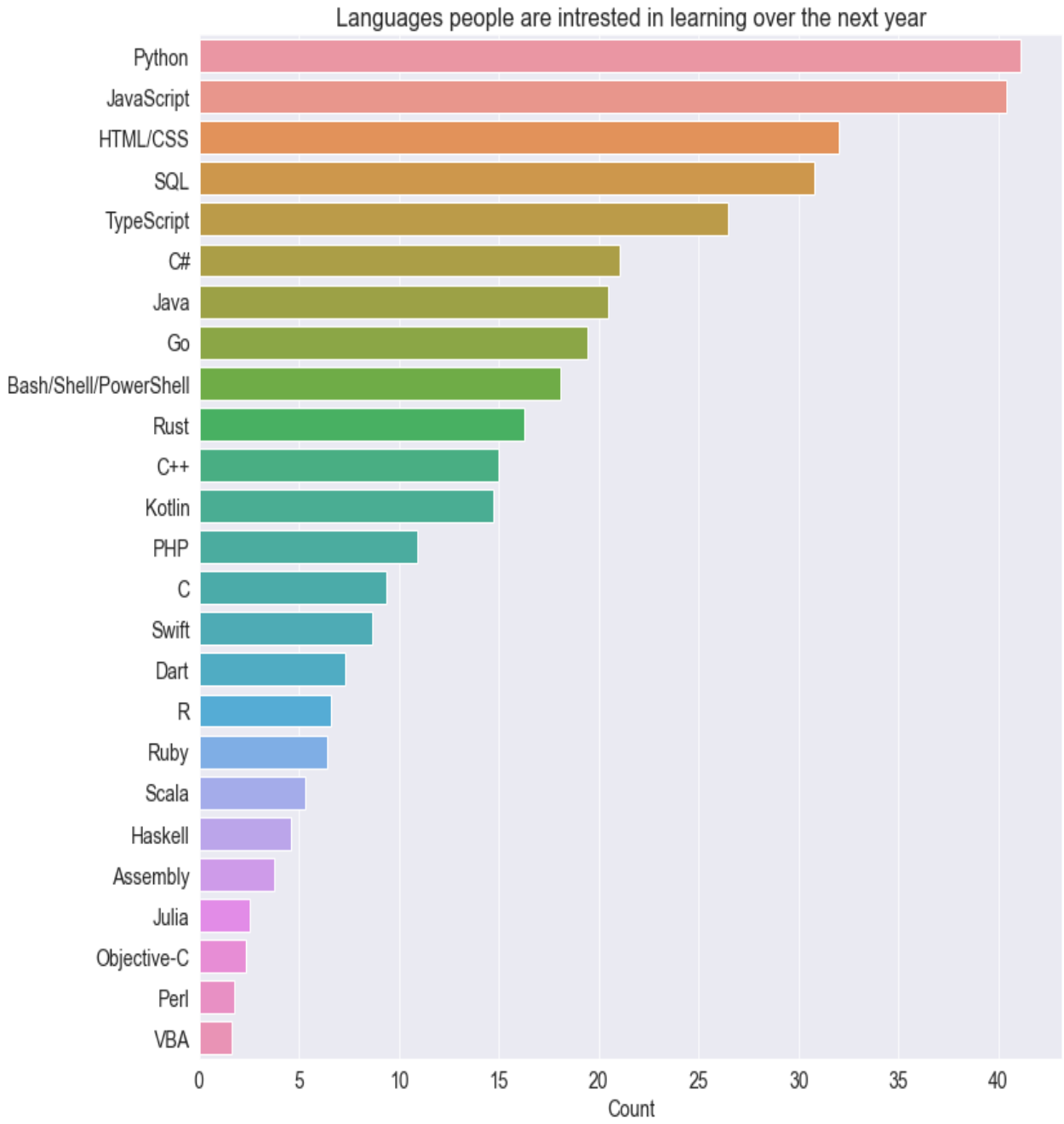
Q. Which languages are the most people interested to learn over the next year?

For this we use the LanguagesDesireNextYear column, with similar processing as the previous one.

```
In [69]: languages_intrested_df=split_multicolumn(survey_df.LanguageDesireNextYear)
         languages_intrested_percentages=languages_intrested_df.mean().sort_values(ascending=False)*100
         languages_intrested_percentages

Out[69]: Python                   41.143906
         JavaScript               40.425466
         HTML/CSS                 32.028116
         SQL                      30.799614
         TypeScript               26.451653
         C#                       21.058688
         Java                     20.464653
         Go                       19.432090
         Bash/Shell/PowerShell    18.057413
         Rust                     16.270643
         C++                      15.014151
         Kotlin                   14.760676
         PHP                      10.947657
         C                         9.359935
         Swift                     8.692812
         Dart                      7.308805
         R                         6.571704
         Ruby                      6.425528
         Scala                     5.326097
         Haskell                   4.593662
         Assembly                  3.766367
         Julia                     2.540976
         Objective-C               2.338818
         Perl                      1.761888
         VBA                       1.611047
         dtype: float64
```

```
In [70]: plt.figure(figsize=(12,12))
         sns.barplot(languages_intrested_percentages,languages_intrested_percentages.index)
         plt.title('Languages people are intrested in learning over the next year')
         plt.xlabel('Count')
         plt.show()
```

Once again, its not surprising that python is the language most people are intrested in learning- since it is an easy-to-learn general purpose programming language well suited for a variety of domains: application development, numeric computing, data analysis, machine learning etc. I am using python for this very analysis.


Languages people are intrested in learning over the next year
Python
JavaScript
HTML/CSS
SQL
TypeScript
C#
Java
Go
Bash/Shell/PowerShell
Rust
C++
Kotlin
PHP
C
Swift
Dart
R
Ruby
Scala
Haskell
Assembly
Julia
Objective-C
Perl
VBA
0
5
10
15
20
25
30
35
40
Count

Q. Which are the most loved languages i.e. a high percentage of people who have used the language want to continue learning & using it over the next year?

We can here use pandas array operation which will make it easy to work on the problem. Here what we can do:

- Create a new data frame language_loved_df which contains a True value for a language only if the corresponding values in languages_worked_df and languages_interested_df are both true.
- Take the column-wise sum of languages_loved_df and divide it by the columns-wise sum of languages_worked_df to get the percentage of respondents.
- Sort the result into descending order and plot a horizontal bar graph.

```
In [71]: languages_loved_df=languages_worked_df & languages_intrested_df
```

```
In [72]: languages_loved_df
```

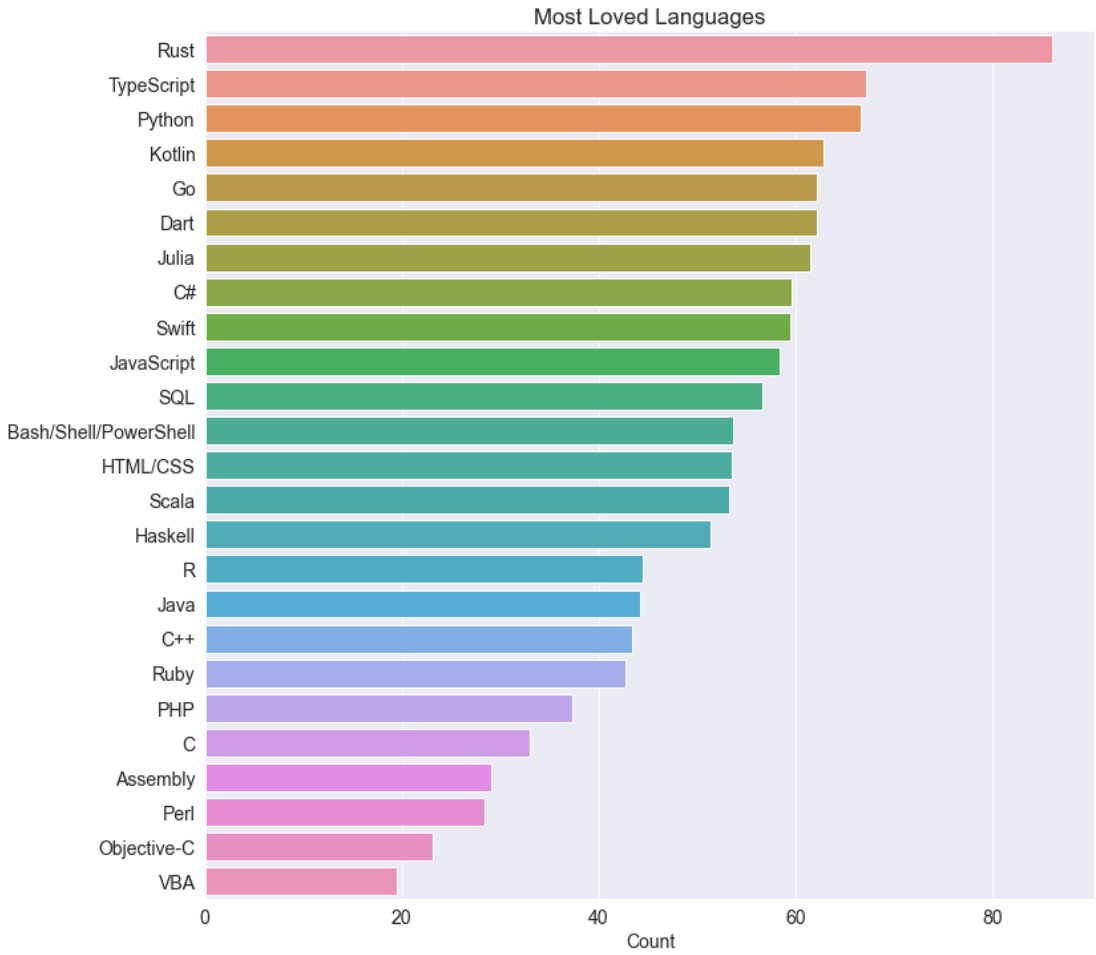
	Assembly	Bash/Shell/PowerShell	C	C#	C++	Dart	Go	HTML/CSS	Haskell	Java	...	Perl	Python	R	Ruby	Rust	SQL	Scala	Swift	TypeScript	VB/
0	False	False	False	True	False	False	False	True	False	False	...	False	False	False	False	False	False	False	False	False	Fals
1	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	True	False	Fals
2	False	False	False	False	False	False	False	False	False	False	...	False	True	False	False	False	False	False	True	False	Fals
3	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	Fals
4	False	False	False	False	False	False	False	False	False	False	...	False	False	False	True	False	False	False	False	False	Fals
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
456	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	Fals
457	True	True	True	True	True	True	True	True	True	True	...	True	True	True	True	True	True	True	True	True	Tru
458	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	Fals
459	False	False	False	False	False	False	False	True	False	False	...	False	False	False	False	False	False	False	False	False	Fals
460	False	False	False	True	False	False	False	True	False	True		False	False	False	False	False	True	False	False	False	Fals

Finding Percentage:

```
In [73]: languages_loved_percentages=(languages_loved_df.sum()*100/ languages_worked_df.sum()).sort_values(ascending=False)
```

Rust has been Stack Flow's most loved languages for 4 years in a row, followed by typescript which has gained a lot of popularity in the past few years as a good alternative to JavaScript for web development. Python features number 3, despite being the one of the most widely used language in world

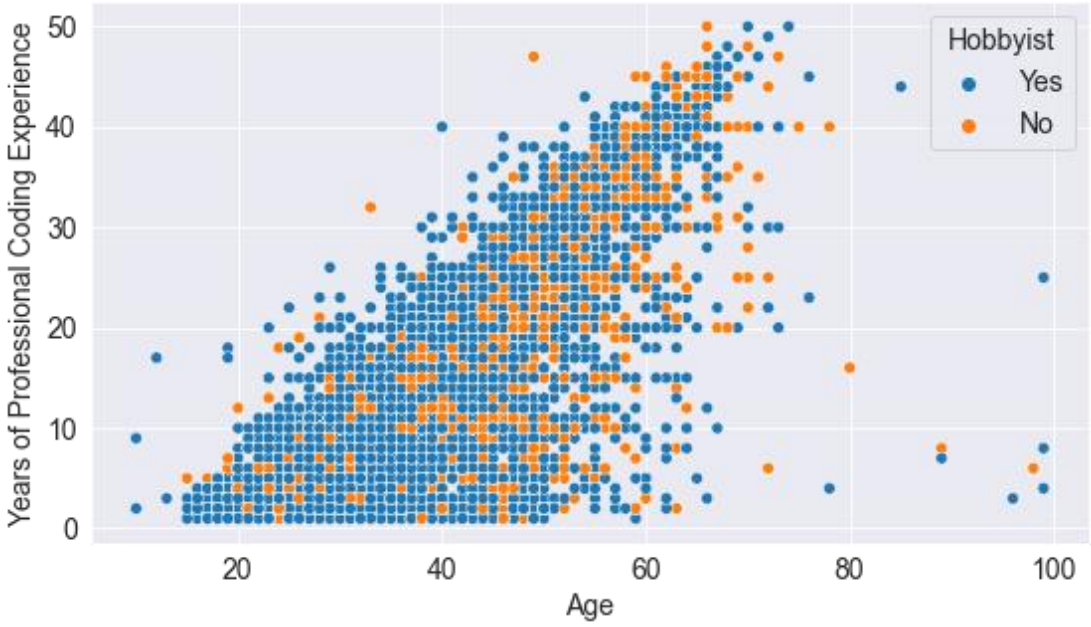
```
In [74]: languages_loved_percentages

Out[74]: Rust                    86.069824
         TypeScript              67.114373
         Python                  66.598884
         Kotlin                  62.813276
         Go                      62.176789
         Dart                    62.068966
         Julia                   61.431412
         C#                      59.623021
         Swift                   59.476346
         JavaScript              58.353888
         SQL                     56.607710
         Bash/Shell/PowerShell   53.688241
         HTML/CSS                53.494594
         Scala                   53.257651
         Haskell                 51.378446
         R                       44.427387
         Java                    44.108273
         C++                     43.415820
         Ruby                    42.735896
         PHP                     37.232755
         C                       32.983023
         Assembly                29.089352
         Perl                    28.369994
         Objective-C             23.133362
         VBA                     19.458057
         dtype: float64
```



Q. How important is it to start young to build a career in programming?

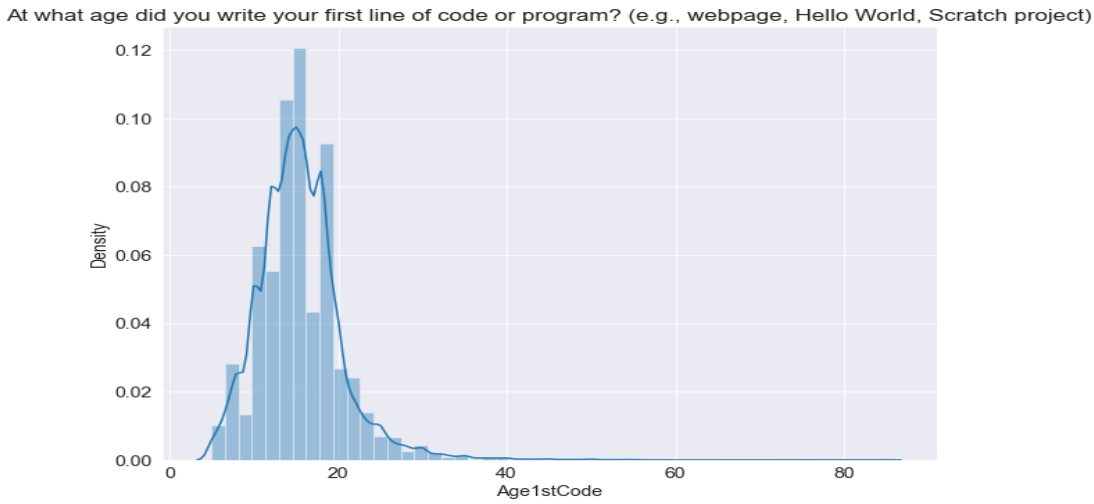
Here I create a scatter plot Age vs YearsCodePro to answer this question.

```
In [81]: schema.YearsCodePro
Out[81]: 'NOT including education, how many years have you coded professionally (as a part of your work)?'
```



You can see points all over the graph, which seems to indicate that you can start programming at any age. Also, many people have been coding for several decades professionally also seems to enjoy it as a hobby.

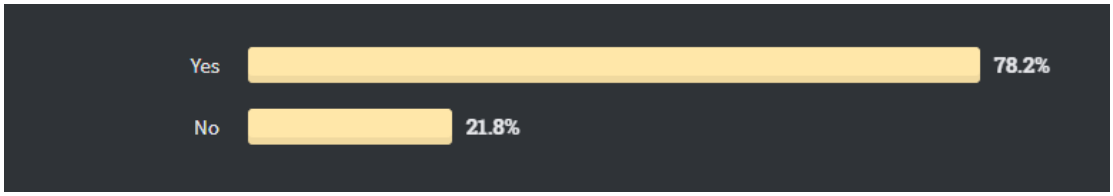
We can also view the distribution of Age1stCode column to see when the respondents tried programming for the first time.



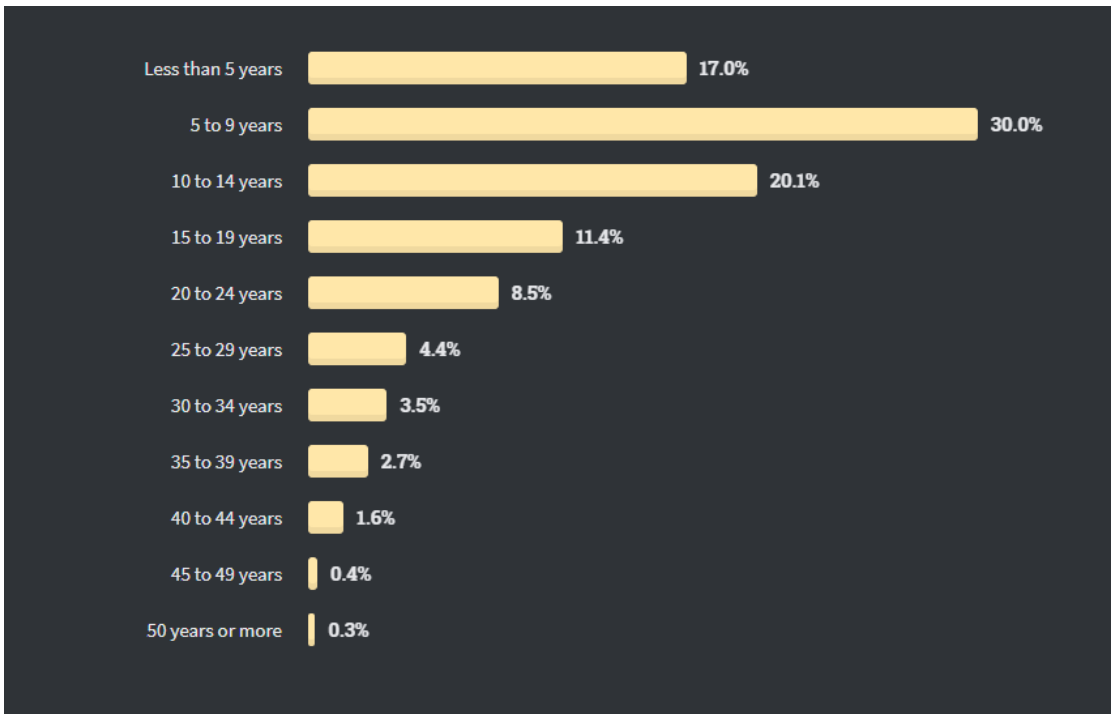
As you might expect, most people seem to have some exposure to programming before the age of 40, but there are people of all ages and walks of life who are learning to code.

Some Other Analysis

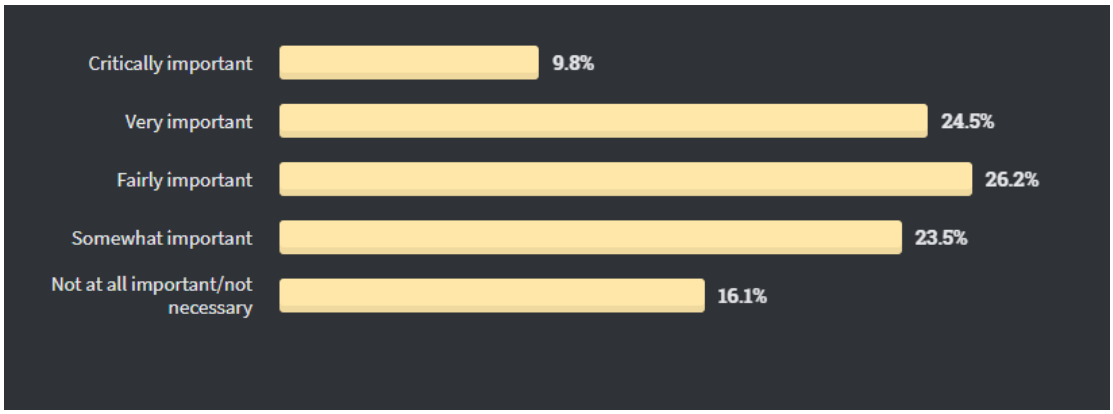
Coding As a Hobby



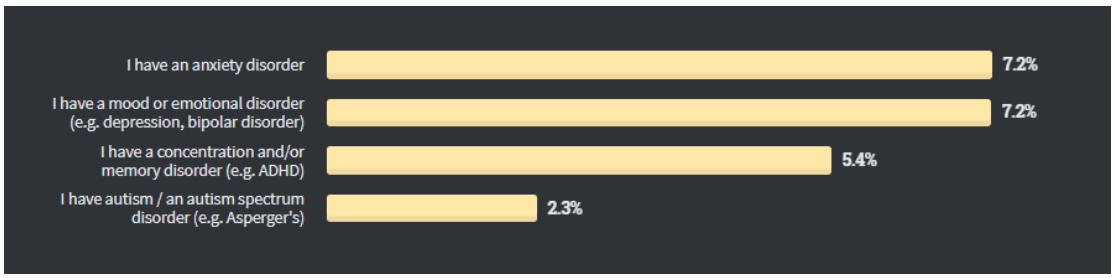
Experience



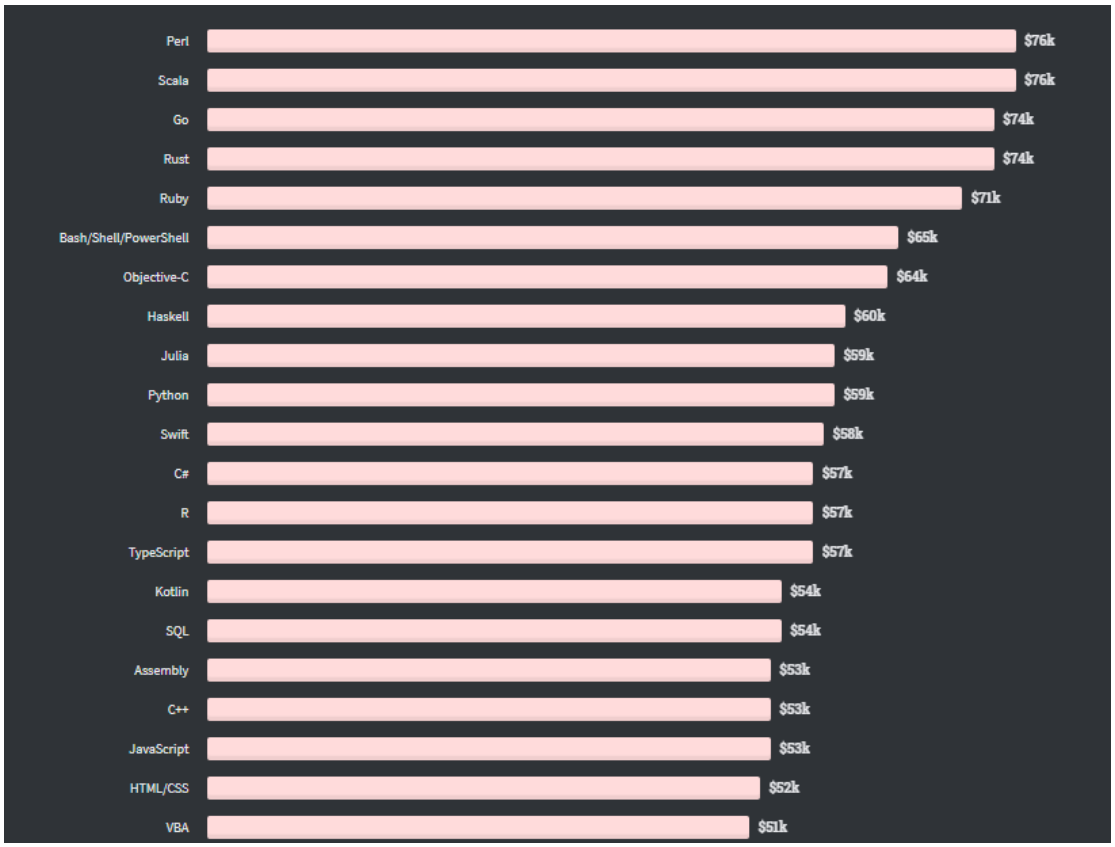
Formal Education Importance



Disability Status

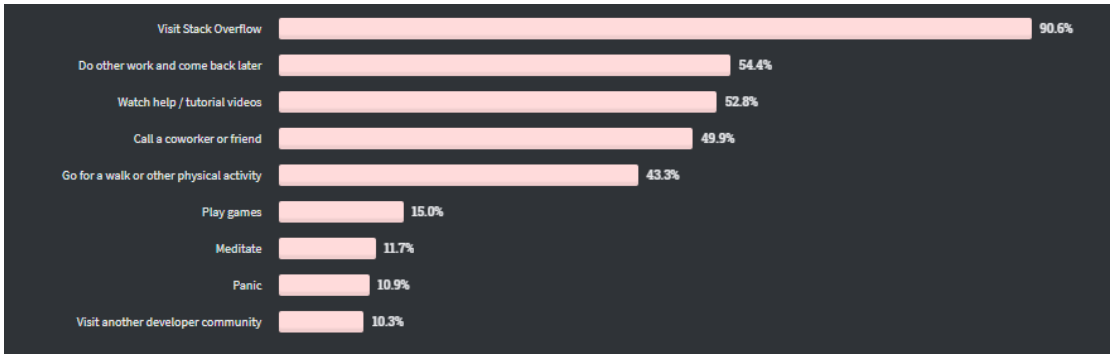


Top Paying Technologies

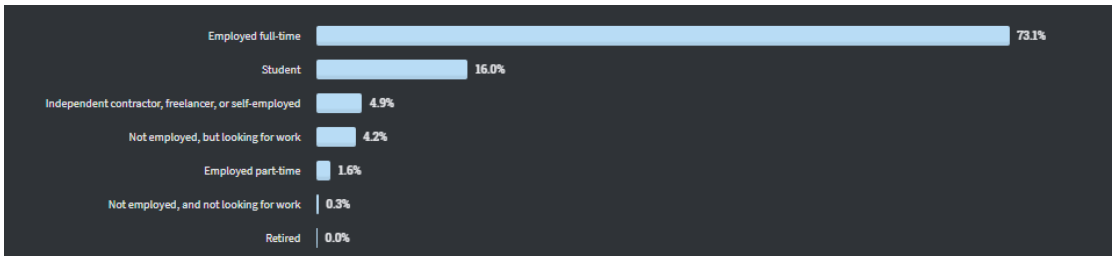


Learning New Tech

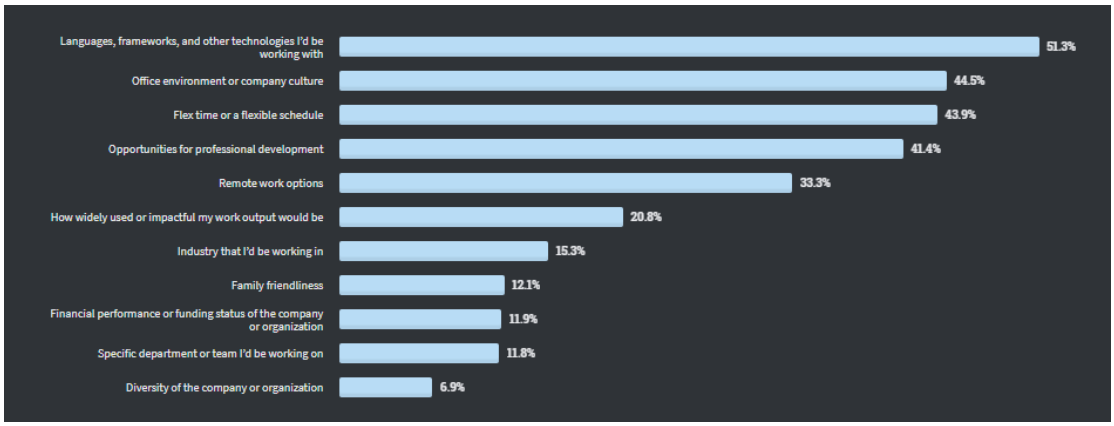
What do you do when get stuck?



Employment Status (India)



Most Important Job Factors



Inferences and Conclusions

I have drawn many interesting inferences from the survey, here's a summary of the few of them:

- Based on the demographics of the survey respondents, we can infer that the survey is somewhat representative of the overall programming community, although its definitely has fewer responses from programmers in non-English speaking countries and from women and non binary gender.
- The programming community is not as diverse as it can be, and although things are improving, we should take more efforts to support & encourage members of underrepresented communities - whether it is in terms of age, country, race, gender or otherwise.
- Most programmers hold a college degree, although a fairly large percentile did not have computer science as a major in college, so a computer science degree isn't compulsory for learning to code or building a career in programming.
- A significant percentage of programmers either work part time or as a freelancer, and this can be a great way to break into the field, especially when you are just getting started.
- JavaScript and HTML/CSS are the most used programming languages in 2020, closely followed by SQL and Python.
- Python is the language most people are interested in learning - since it is easy to learn general purpose programming language well suited for variety of domains.
- Rust and Type Script are the most loved languages in 2020, both of which have small but fast growing communities. Python is closed third, despite already being widely used language.
- Programmers around the world seems to be working for around 40 hours a week on average, with slight variations by country.
- You can learn and start programming professionally at any age, and you are likely to have a long and full filling career if you also enjoy programming as a hobby.

Recommendations

To the Organization:

i. Facilitation; The management of A.C should buy more facilities such as computers, vehicles so as to enable smooth running of the company's activities and respond to the dynamic competition environment. This technological advancement will enable the organization to change from manual to computerized methods of processing documents and proper record keeping.

ii. More opportunities to students to do intern in their organization; The intern would also recommend the Organization to continue giving internship placements to as many students as they can because some miss this experience which is also important requirement of the University due to the fact that they failed to get placements.

iii. Allowances, the organization should give allowances to interns most especially transport allowances to cater for transport cost most students stay far from the internship places hence increasing my expenses. Therefore the intern recommends the Organization to put that in to action in order to motivate interns and boost their productivity levels in performing their tasks during the field attachment.

iv. Serious supervision to the workers and students, the organization should increase and ensure more supervision over the employees in order to work effectively and also eliminate workers who relax, work lazily and perform actively after seeing their supervisors.

v. Job enlargement. The management of Amnesty Commission should also carry out job enlargement and enrichment such that it mitigate the conflict amongst employees for roles and tasks .This will ensure good industrial relations between the supervisors and subordinates at the organization.

To the university:

vi. Constant supervision of students, The intern recommends the university to carry out constant supervision and monitoring of students during the internship training so as to encourage them to perform the duties fully and also accurately. This will also put a close link between the academic supervisors and the field supervisors so as to foster appropriate assessment of what the interns are doing in the field.

vii. Secure Internship placements for students. The University should help students to secure internship positions according to their respective programs undertaken at the University through giving students recommendations in order to ease their training periods and also avoid the ache gotten by students in search of internship placements.

viii. Should continue with internship program, this is because it helps to prepare the students for their careers in future and also enable the students to practice the theoretical knowledge obtained during class be exercised practically. It also helps to develop students understanding of work ethics, employment demands, responsibilities and opportunities.

References and Future Work

There's a wealth of information to be discovered using the survey, and I've barely scratched the surface. Here are some ideas for future exploration:

- Repeat the analysis for different age groups & genders, and compare the results.
- Choose a different set of columns (I have chosen 20 out of 65) to analyze other facts of data.
- Prepare an analysis focuses on diversity - and identify areas where underrepresented communities are at par with the majority (eg. education) and where they aren't (eg. salary).
- Compare the result of this year's survey with the previous years and identify interesting trends.

References:

- Stack Overflow Developer Survey: https://insights.stackoverflow.com/survey/2020
- Pandas User Guide: https://pandas.pydata.org/pandas-docs/stable/user_guide/index.html
- Matplotlib User Guide: https://matplotlib.org/stable/users/index.html
- Seaborn User Guide: https://seaborn.pydata.org/tutorial.html
- opendatasets Python Library: https://pypi.org/project/opendatasets/