# JOB APPLICATION ANALYSIS

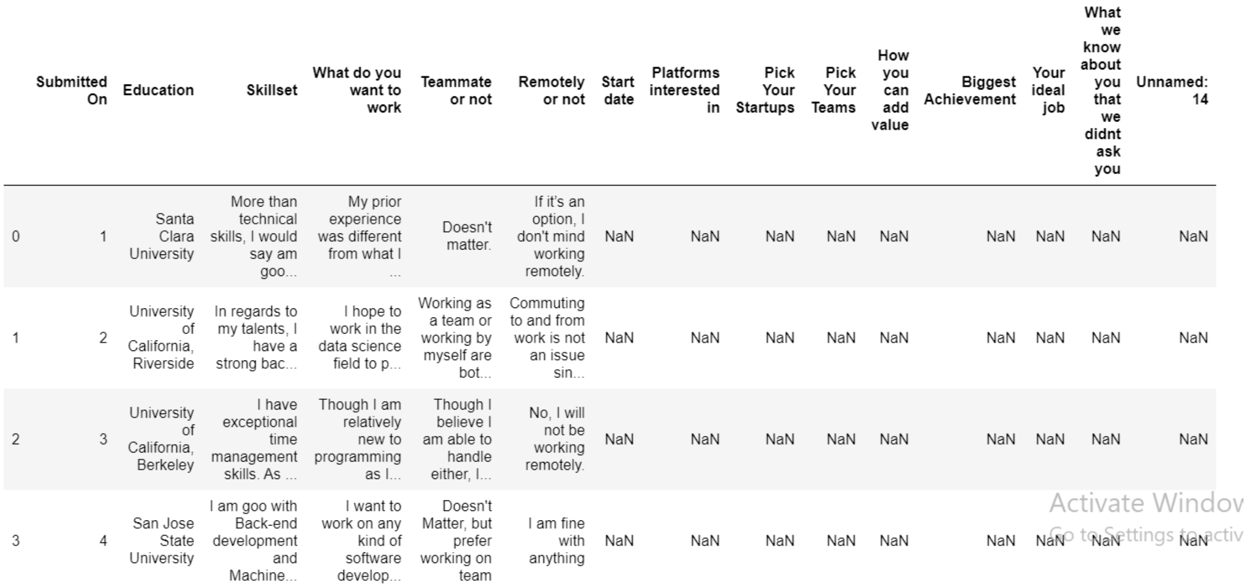
## I. Introduction

The original file: https://docs.google.com/spreadsheets/d/1eGPU-awXQFqB4__AC4lmCt4vjFy1rt3m-6zZ85_siGE/edit?usp=sharing

The cleaned file: https://docs.google.com/spreadsheets/d/1idtPjBjZ-qqJJrFvWDRVC_mH2V7Dx2sEPIeOhB9AYdo/edit?usp=sharing

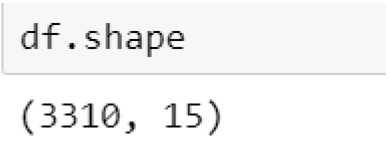
Things were cleaned:

- Arrange values into right columns
- Rename the header of columns



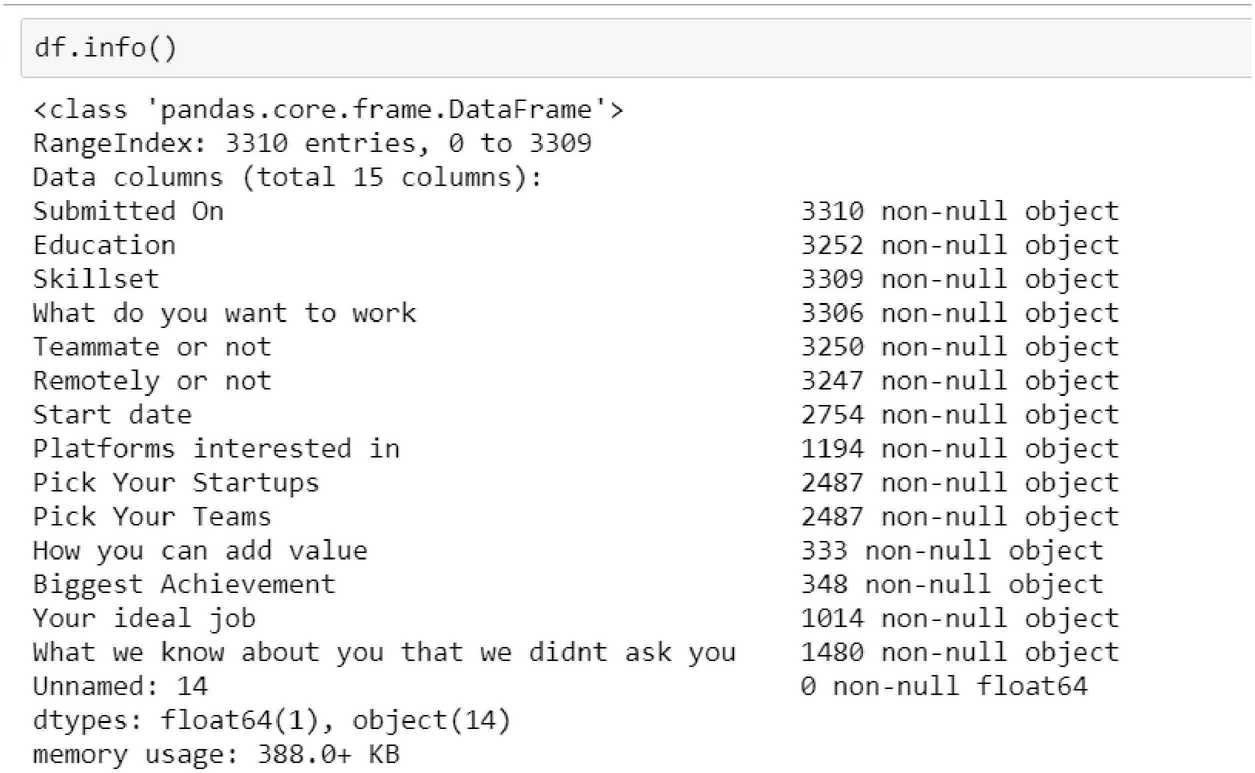
| | Submitted On | Education | Skillset | What do you want to work | Teammate or not | Remotely or not | Start date | Platforms interested in | Pick Your Startups | Pick Your Teams | How you can add value | Biggest Achievement | Your ideal job | What we know about you that we didnt ask you | Unnamed: 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | Santa Clara University | More than technical skills, I would say am goo... | My prior experience was different from what I ... | Doesn't matter. | If it’s an option, I don't mind working remotely. | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 1 | 2 | University of California, Riverside | In regards to my talents, I have a strong bac... | I hope to work in the data science field to p... | Working as a team or working by myself are bot... | Commuting to and from work is not an issue sin... | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 2 | 3 | University of California, Berkeley | I have exceptional time management skills. As ... | Though I am relatively new to programming as I... | Though I believe I am able to handle either, I... | No, I will not be working remotely. | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 3 | 4 | San Jose State University | I am goo with Back-end development and Machine... | I want to work on any kind of software develop... | Doesn't Matter, but prefer working on team | I am fine with anything | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |

Overview of the dataset



```
df.shape
(3310, 15)
```

There are 3310 observation and 15 columns



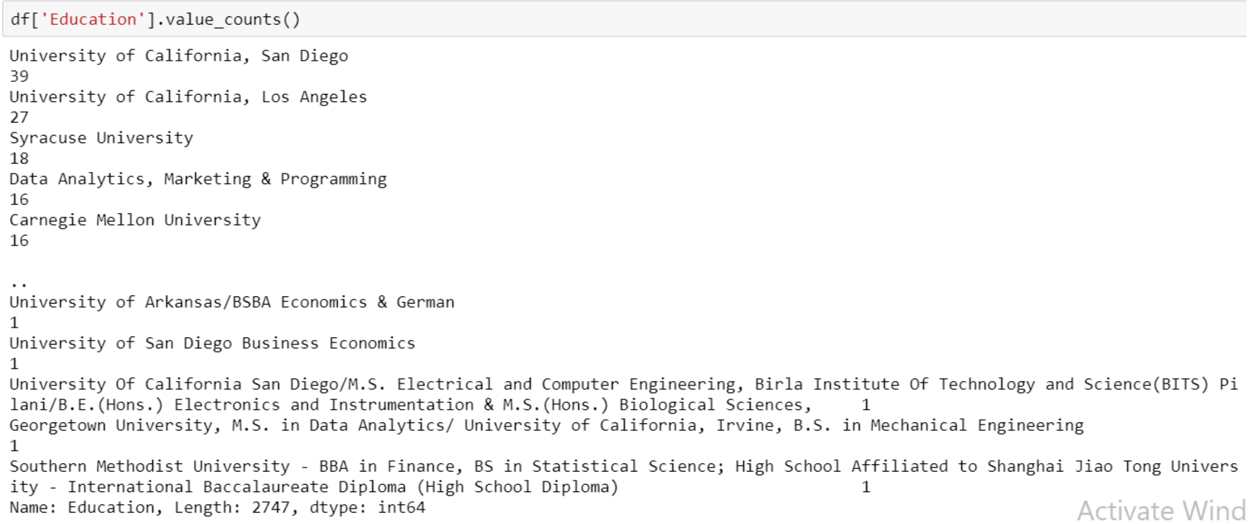
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3310 entries, 0 to 3309
Data columns (total 15 columns):
Submitted On                                    3310 non-null object
Education                                       3252 non-null object
Skillset                                        3309 non-null object
What do you want to work                        3306 non-null object
Teammate or not                                 3250 non-null object
Remotely or not                                 3247 non-null object
Start date                                      2754 non-null object
Platforms interested in                         1194 non-null object
Pick Your Startups                              2487 non-null object
Pick Your Teams                                 2487 non-null object
How you can add value                           333 non-null object
Biggest Achievement                             348 non-null object
Your ideal job                                  1014 non-null object
What we know about you that we didnt ask you    1480 non-null object
Unnamed: 14                                     0 non-null float64
dtypes: float64(1), object(14)
memory usage: 388.0+ KB
```

Picture above gave us more details of each column. Although there are 3310 observations, there are a lot null values in each column. This means not every candidate finished all of the questions.

## II. Part 1 - Talent Analysis

### 1. Education

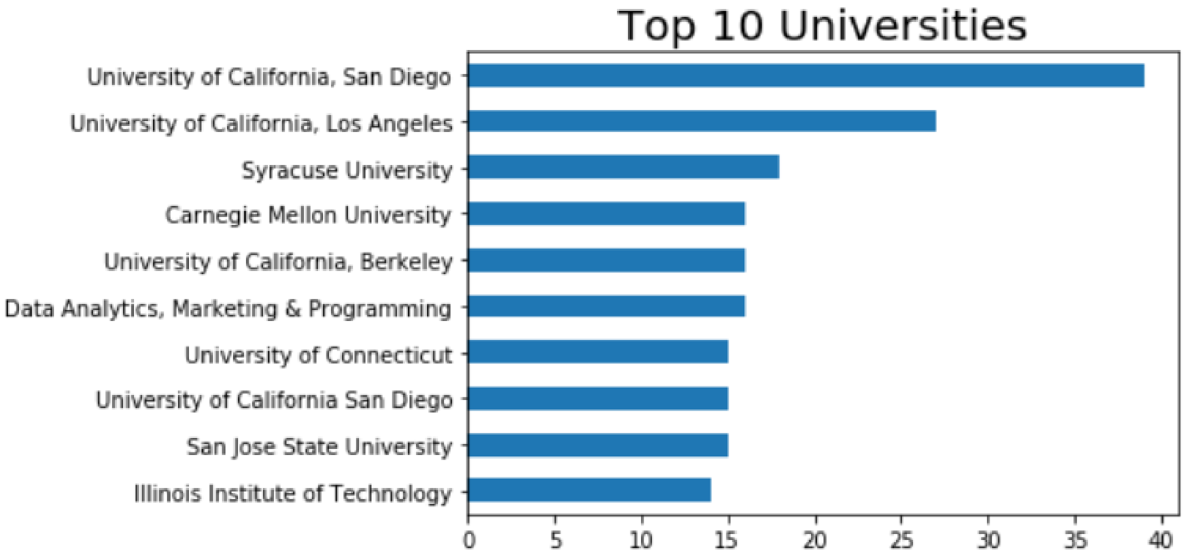


```
df['Education'].value_counts()
```

```
University of California, San Diego
39
University of California, Los Angeles
27
Syracuse University
18
Data Analytics, Marketing & Programming
16
Carnegie Mellon University
16

..
University of Arkansas/BSBA Economics & German
1
University of San Diego Business Economics
1
University Of California San Diego/M.S. Electrical and Computer Engineering, Birla Institute Of Technology and Science(BITS) Pi
lani/B.E.(Hons.) Electronics and Instrumentation & M.S.(Hons.) Biological Sciences,     1
Georgetown University, M.S. in Data Analytics/ University of California, Irvine, B.S. in Mechanical Engineering
1
Southern Methodist University - BBA in Finance, BS in Statistical Science; High School Affiliated to Shanghai Jiao Tong Univers
ity - International Baccalaureate Diploma (High School Diploma)                      1
Name: Education, Length: 2747, dtype: int64
```

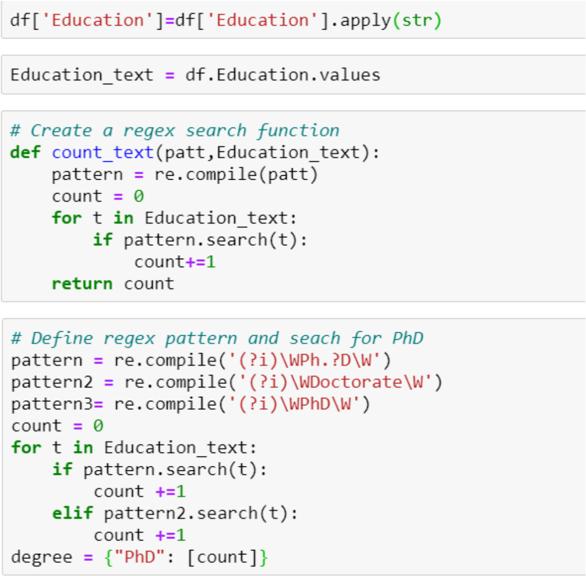
Based on the information that candidates filled in, we can see top universities they are attending. However, different students filled in the information in different ways. For example, some wrote

“UC” instead of “University of California”. However, in some perspectives, we can know that **Forkaia are attracting a lot of students from Top Universities around the country.**



In this part, I would like to pay attention to find out how many students have completed or are pursuing bachelor degree, master degree and PhD programs.

- For PhD degree, I would like to find the frequency of keywords “PhD”, “Ph.D” or “Doctorate”.
- For Master degree, the frequency of keywords “Master”, “Masters”, “MS”, or “M.S” are counted.
- For MBA degree, the frequency of keywords “MBA” are counted
- For Bachelor degree, I would like to find the frequency of keywords “Bachelor”, “Bachelors”, “BS”, “BA”, “B.S” or “B.A.



```
df['Education']=df['Education'].apply(str)
```

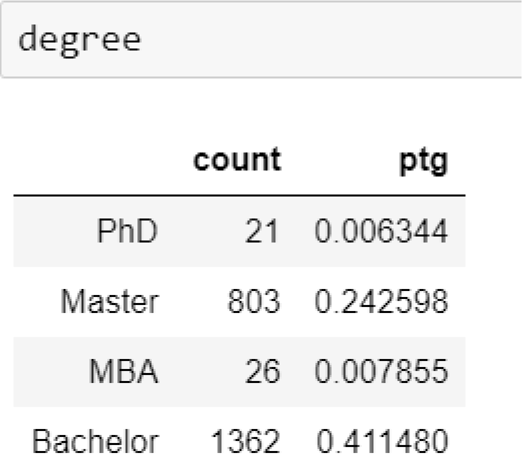
```
Education_text = df.Education.values
```

```
# Create a regex search function
def count_text(patt,Education_text):
    pattern = re.compile(patt)
    count = 0
    for t in Education_text:
        if pattern.search(t):
            count+=1
    return count
```

```
# Define regex pattern and seach for PhD
pattern = re.compile('(?i)\WPh.?D\W')
pattern2 = re.compile('(?i)\WDoctorate\W')
pattern3= re.compile('(?i)\WPhD\W')
count = 0
for t in Education_text:
    if pattern.search(t):
        count +=1
    elif pattern2.search(t):
        count +=1
degree = {"PhD": [count]}
```

*Illustration of codes*

Pictures below is what I found

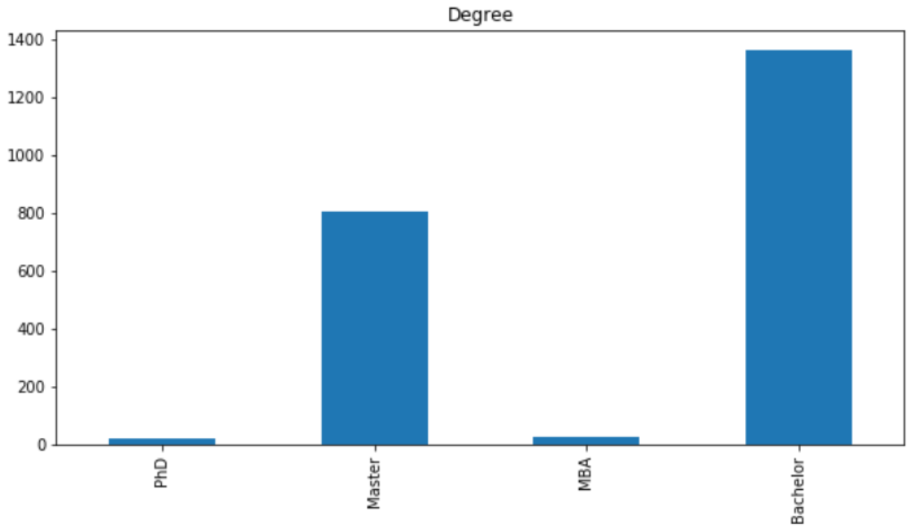


```
degree
```

| | count | ptg |
|---|---|---|
| PhD | 21 | 0.006344 |
| Master | 803 | 0.242598 |
| MBA | 26 | 0.007855 |
| Bachelor | 1362 | 0.411480 |

Forkaia should be proud to say that there are 21 PhD, 803 Master, 26 MBA and more than 1300 Bachelor candidates have applied to the company.

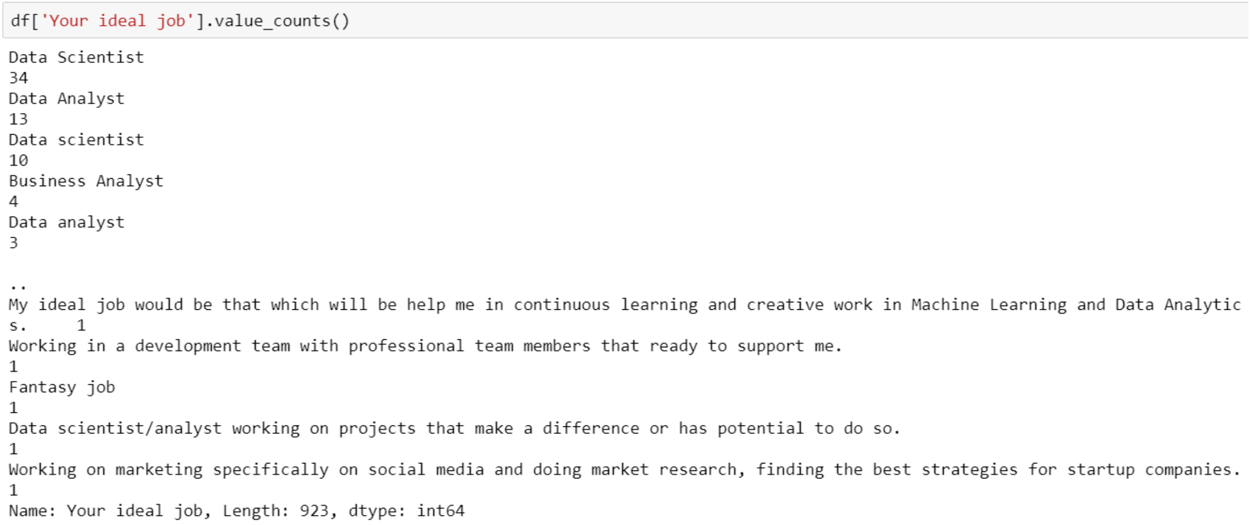
Regarding percentage, 24% of students would like to work for Forkaia are currently pursuing Master of Science degree. Also, more than 40% have completed or are currently studying Bachelors.



2. Job Title

Computer Scientist, Data Scientist, Software Engineer are those titles that the company would like to showcase. Job Title can be retrieved from different columns including “Your ideal job”, “Pick your Team” or major in “Education”.

### Regarding “Your Ideal Job” column



```
df['Your ideal job'].value_counts()
```

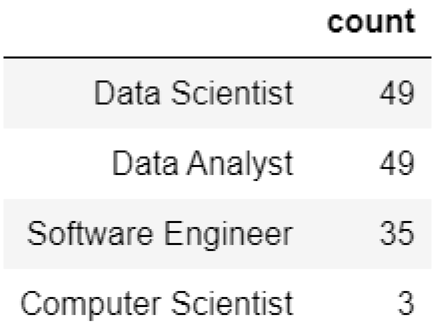
```
Data Scientist
34
Data Analyst
13
Data scientist
10
Business Analyst
4
Data analyst
3

..
My ideal job would be that which will be help me in continuous learning and creative work in Machine Learning and Data Analytic
s.     1
Working in a development team with professional team members that ready to support me.
1
Fantasy job
1
Data scientist/analyst working on projects that make a difference or has potential to do so.
1
Working on marketing specifically on social media and doing market research, finding the best strategies for startup companies.
1
Name: Your ideal job, Length: 923, dtype: int64
```

There are 1014 observations in 923 different values.

- For Data Scientist, I would like to find the frequency of these words: Data Scientist, Data Science.
- For Data Analytics, I would like to find the frequency of these words: Data Analytics, Data Analyst, Business Analyst, Business Analytics.

- For Software Engineer, I would like to find the frequency of these words: Software Engineer, Software Development, Developer.
- For Computer Scientist, I would like to find the frequency of these words: Computer Science, Computer Scientist.

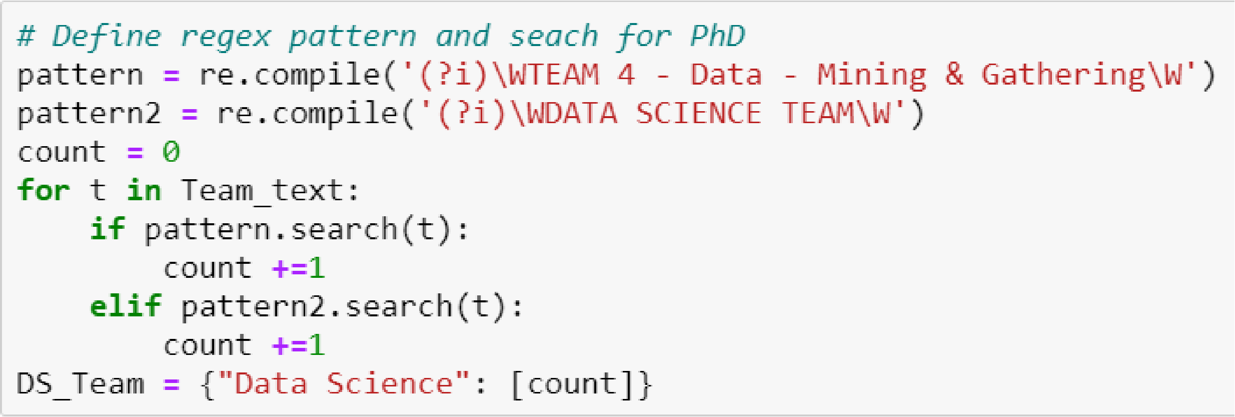
Here's the result of Job Title in column Your Ideal Job



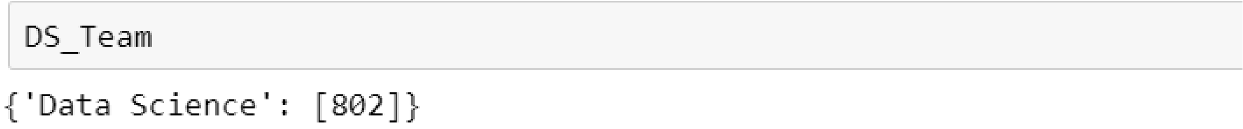
| | count |
|---|---|
| Data Scientist | 49 |
| Data Analyst | 49 |
| Software Engineer | 35 |
| Computer Scientist | 3 |

**Regarding "Pick Your Team" columns**

As you know, there are 5 teams: Development, Data, Business & Marketing, Creative and Project Management. This means we can only retrieve the number candidates picking Development and Data team from this columns. We can assume that for those picking Development Team, they would like to work as Software Engineer, and for those picking Data Team, they would like to work as Data Scientist or Analyst.

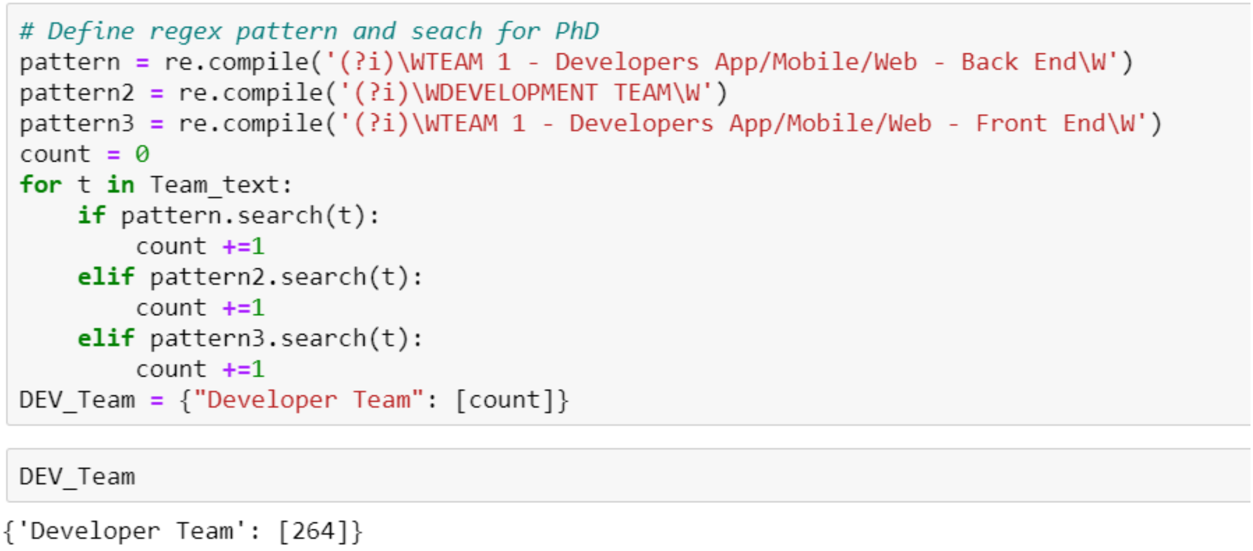


```
# Define regex pattern and seach for PhD
pattern = re.compile('(?i)\WTEAM 4 - Data - Mining & Gathering\W')
pattern2 = re.compile('(?i)\WDATA SCIENCE TEAM\W')
count = 0
for t in Team_text:
    if pattern.search(t):
        count +=1
    elif pattern2.search(t):
        count +=1
DS_Team = {"Data Science": [count]}
```



```
DS_Team
```

```
{'Data Science': [802]}
```



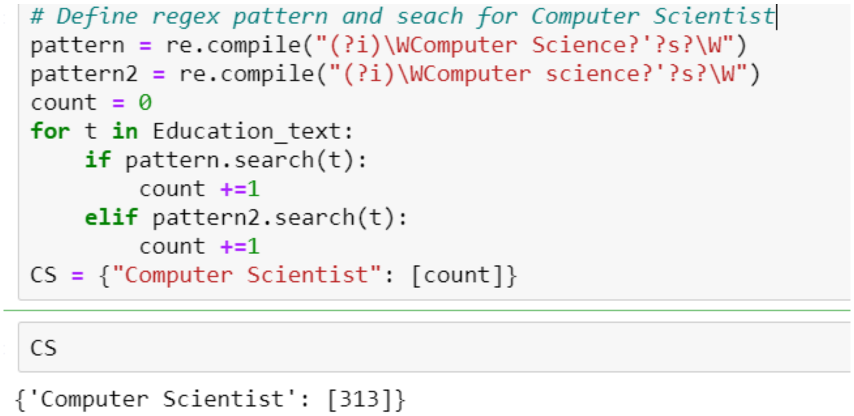
```
# Define regex pattern and seach for PhD
pattern = re.compile('(?i)\WTEAM 1 - Developers App/Mobile/Web - Back End\W')
pattern2 = re.compile('(?i)\WDEVELOPMENT TEAM\W')
pattern3 = re.compile('(?i)\WTEAM 1 - Developers App/Mobile/Web - Front End\W')
count = 0
for t in Team_text:
    if pattern.search(t):
        count +=1
    elif pattern2.search(t):
        count +=1
    elif pattern3.search(t):
        count +=1
DEV_Team = {"Developer Team": [count]}
```

```
DEV_Team
```

```
{'Developer Team': [264]}
```

Although there is a little bit overlapped as many candidates would like to work in different teams, Forkaia can be confident to say that it has 802 candidates would like to work as data scientist and analyst. There are 264 candidates that would like to work as Software Engineer.

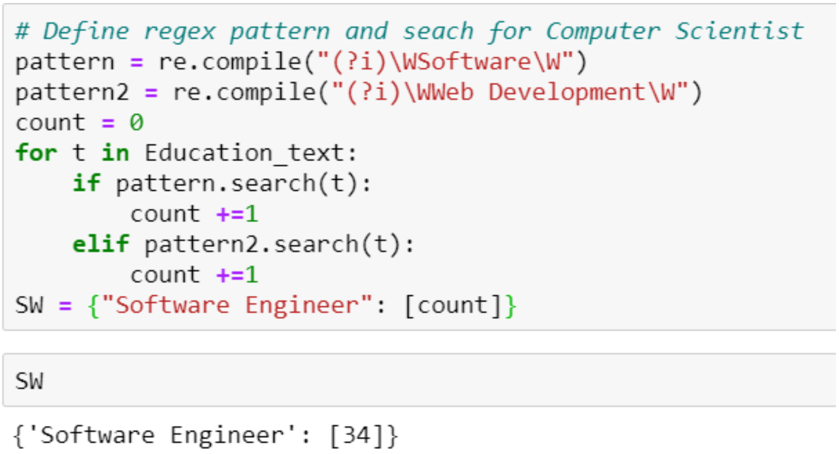
**Regarding major in "Education" column**



```
# Define regex pattern and seach for Computer Scientist
pattern = re.compile("(?i)\WComputer Science?'?s?\W")
pattern2 = re.compile("(?i)\WComputer science?'?s?\W")
count = 0
for t in Education_text:
    if pattern.search(t):
        count +=1
    elif pattern2.search(t):
        count +=1
CS = {"Computer Scientist": [count]}
```

```
CS
```

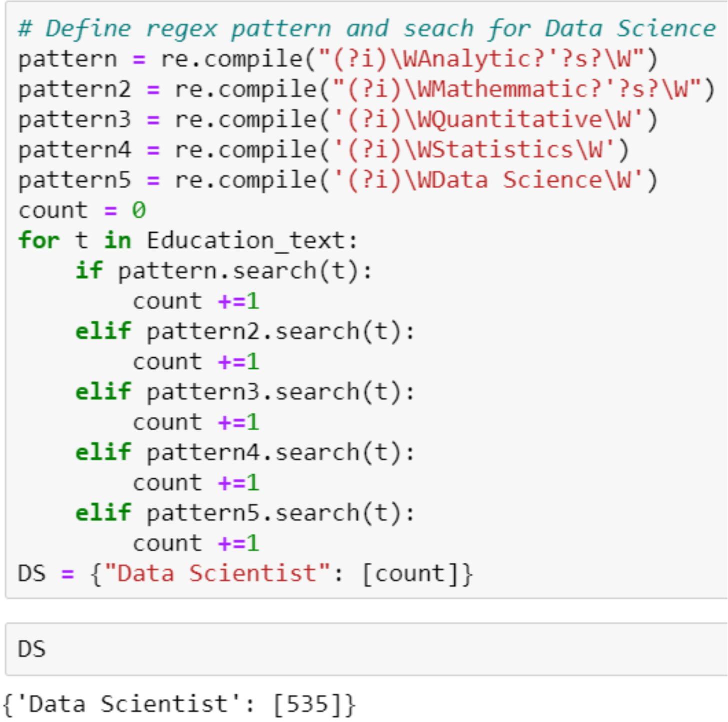
```
{'Computer Scientist': [313]}
```



```
# Define regex pattern and seach for Computer Scientist
pattern = re.compile("(?i)\WSoftware\W")
pattern2 = re.compile("(?i)\WWeb Development\W")
count = 0
for t in Education_text:
    if pattern.search(t):
        count +=1
    elif pattern2.search(t):
        count +=1
SW = {"Software Engineer": [count]}
```

```
SW
```

```
{'Software Engineer': [34]}
```



```
# Define regex pattern and seach for Data Science
pattern = re.compile("(?i)\WAnalytic?'?s?\W")
pattern2 = re.compile("(?i)\WMathemmatic?'?s?\W")
pattern3 = re.compile('(?i)\WQuantitative\W')
pattern4 = re.compile('(?i)\WStatistics\W')
pattern5 = re.compile('(?i)\WData Science\W')
count = 0
for t in Education_text:
    if pattern.search(t):
        count +=1
    elif pattern2.search(t):
        count +=1
    elif pattern3.search(t):
        count +=1
    elif pattern4.search(t):
        count +=1
    elif pattern5.search(t):
        count +=1
DS = {"Data Scientist": [count]}
```

```
DS
```

```
{'Data Scientist': [535]}
```

According to the majors the candidates have completed or are pursuing, there are:

- 313 computer scientist
- 535 data scientist/ analyst
- 34 software engineer

**Conclusion**

To answer the question that how many data scientists, software engineers and computer scientists Forkaia have, I would like to say:

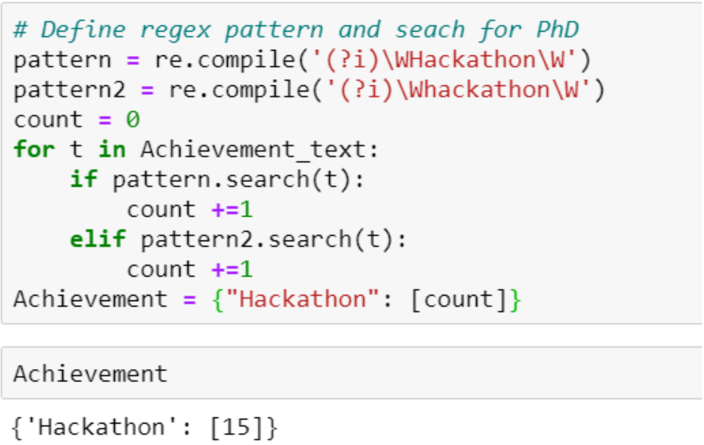
- Get the number of candidates picking Data Team for the number of Data Scientist **(802 Data Scientists)**
- Get the number of candidates picking Development Teams for the number of Software Engineer **(264 Software Engineers)**
- Get the number of candidates studying Computer Science (in "Education") for the number of Computer Scientist **(313 Computer Scientists)**

Reasons for choosing this approach is:

- Although the column Ideal Job has the closest meaning to the Job title, it is has least observation (just 1014 observation). A lot of values in this columns are not meaningful enough for counting number of job titles.
- Title "computer scientist" can be only retrieved from major Computer Science.

- Number of people picking the team Data and Development is the closest reflection of counting Data Scientists and Software Development because current or applied position is a better metrics to analyze than majors.

### 3. Awards – Achievement



```
# Define regex pattern and seach for PhD
pattern = re.compile('(?i)\WHackathon\W')
pattern2 = re.compile('(?i)\Whackathon\W')
count = 0
for t in Achievement_text:
    if pattern.search(t):
        count +=1
    elif pattern2.search(t):
        count +=1
Achievement = {"Hackathon": [count]}
```

```
Achievement
```

```
{'Hackathon': [15]}
```

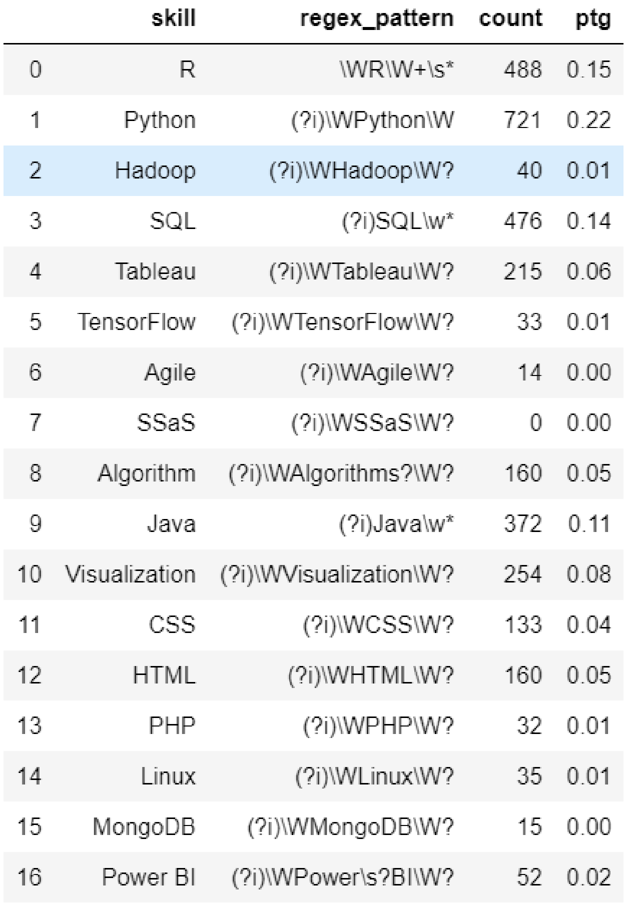
Hackathon are mention 15 times in 348 observations of Biggest Achievement. However, after checking manually the meaning of 15 results mentioned Hackathon, there are only 11 people said that they have participated, won finalist rounds or prizes of Hackathon.

Given the complexity of text “Analysis” Hackathon, the analyst decided not to go deeper in this parts regarding key word “Honor”or “Dean’s List”.

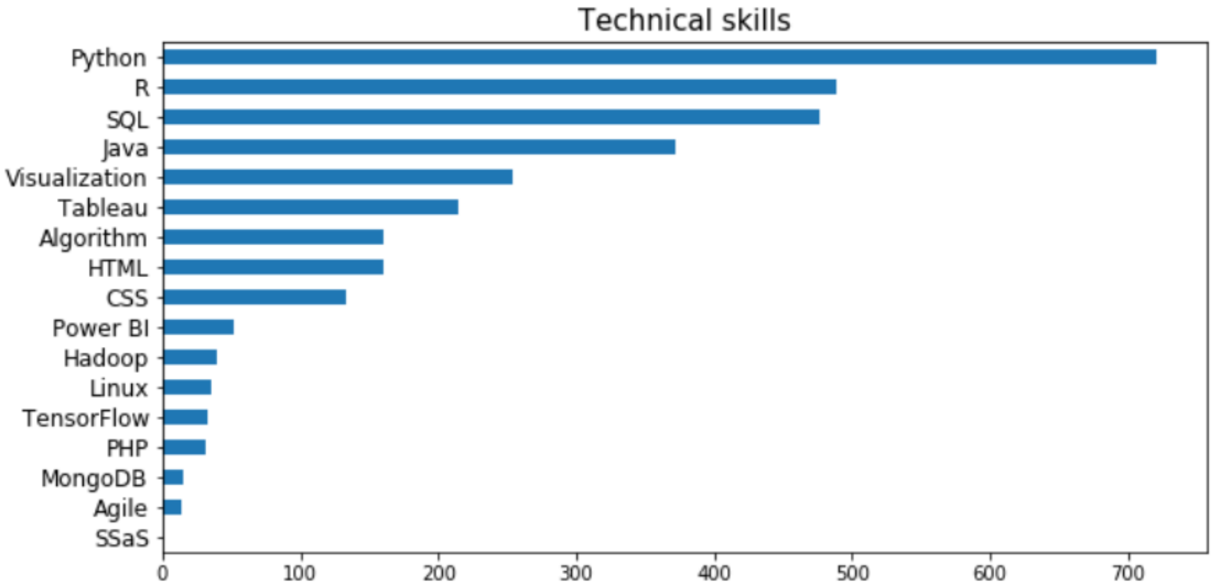
## III. Part Two

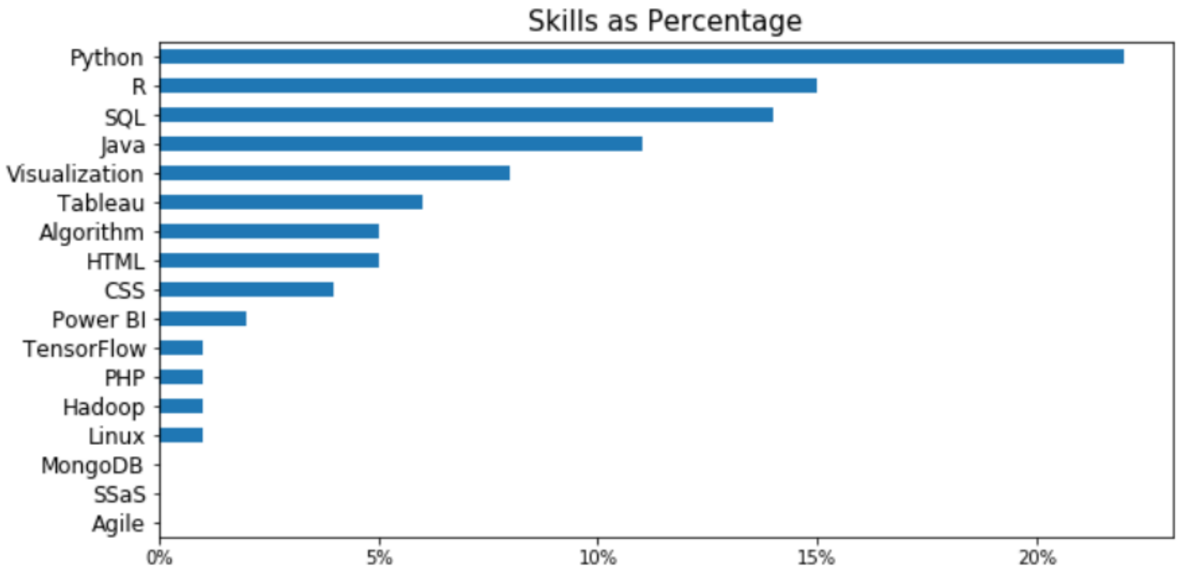
### 1. Skillset

Regarding the questions of skillsets that candidate can add value to the company, pictures below are the results of counting the frequency of technical skills that are mentioned.



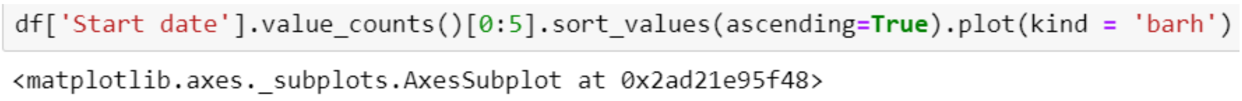
| | skill | regex_pattern | count | ptg |
|---|---|---|---|---|
| 0 | R | \WR\W+\s* | 488 | 0.15 |
| 1 | Python | (?i)\WPython\W | 721 | 0.22 |
| 2 | Hadoop | (?i)\WHadoop\W? | 40 | 0.01 |
| 3 | SQL | (?i)SQL\w* | 476 | 0.14 |
| 4 | Tableau | (?i)\WTableau\W? | 215 | 0.06 |
| 5 | TensorFlow | (?i)\WTensorFlow\W? | 33 | 0.01 |
| 6 | Agile | (?i)\WAgile\W? | 14 | 0.00 |
| 7 | SSaS | (?i)\WSSaS\W? | 0 | 0.00 |
| 8 | Algorithm | (?i)\WAlgorithms?\W? | 160 | 0.05 |
| 9 | Java | (?i)Java\w* | 372 | 0.11 |
| 10 | Visualization | (?i)\WVisualization\W? | 254 | 0.08 |
| 11 | CSS | (?i)\WCSS\W? | 133 | 0.04 |
| 12 | HTML | (?i)\WHTML\W? | 160 | 0.05 |
| 13 | PHP | (?i)\WPHP\W? | 32 | 0.01 |
| 14 | Linux | (?i)\WLinux\W? | 35 | 0.01 |
| 15 | MongoDB | (?i)\WMongoDB\W? | 15 | 0.00 |
| 16 | Power BI | (?i)\WPower\s?BI\W? | 52 | 0.02 |





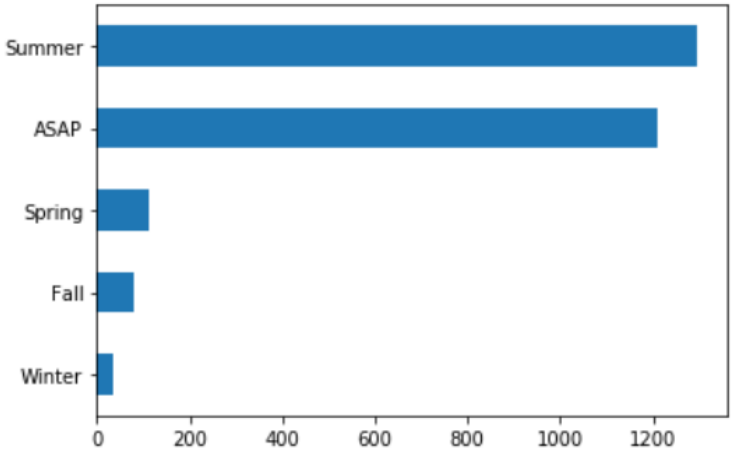
Python, R, SQL and Java are the technical skills that are acquired most by candidates.

### 2. Start date



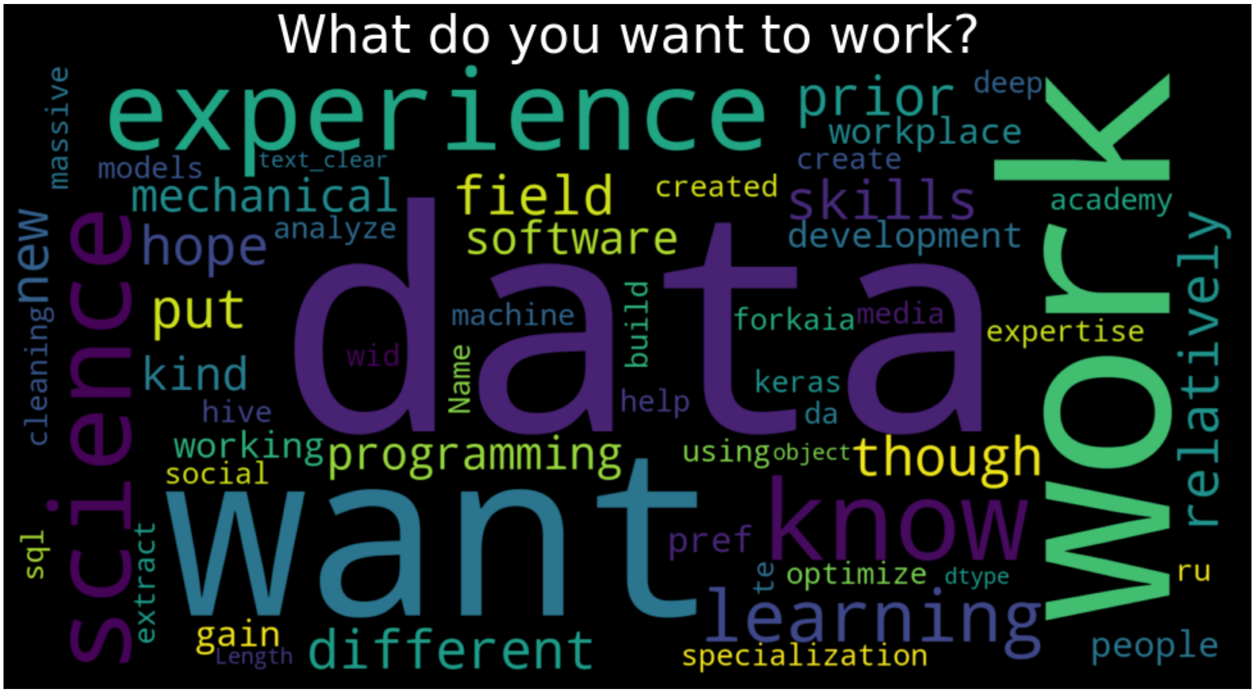
```
df['Start date'].value_counts()[0:5].sort_values(ascending=True).plot(kind = 'barh')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2ad21e95f48>
```



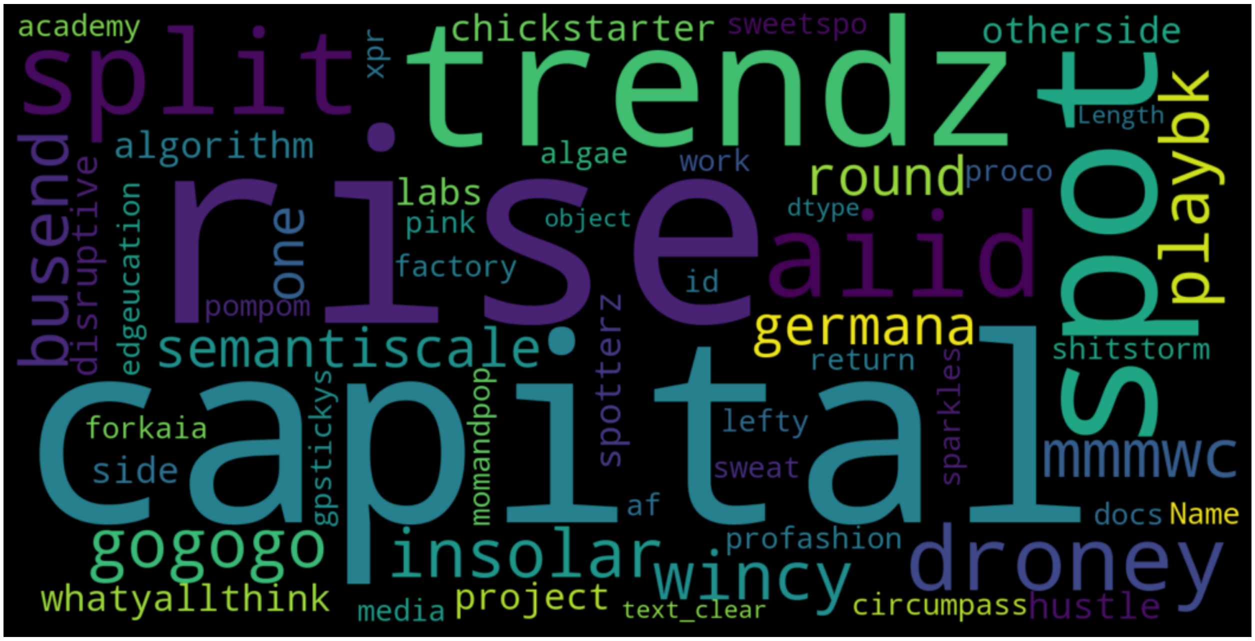
### 3. Text analysis

**What candidates would like to work on?**

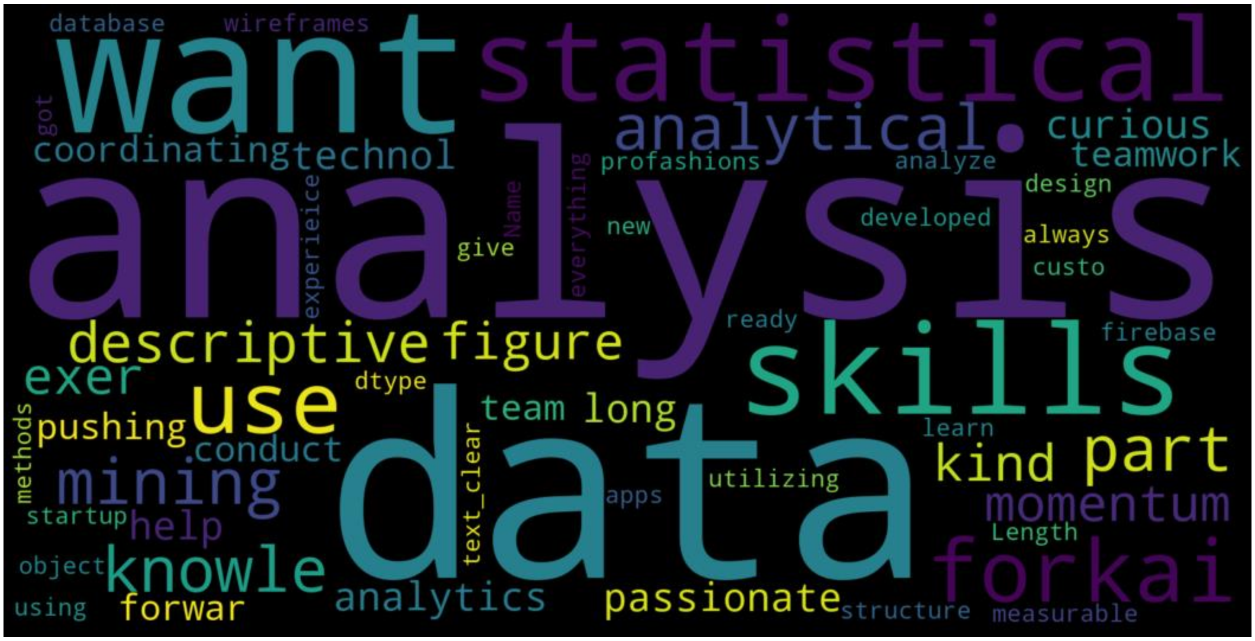


Candidates would like to work on data science and analysis as well as programming, software development and SQL, hive, deep learning, mechanical engineer.

**Most picked start-ups**

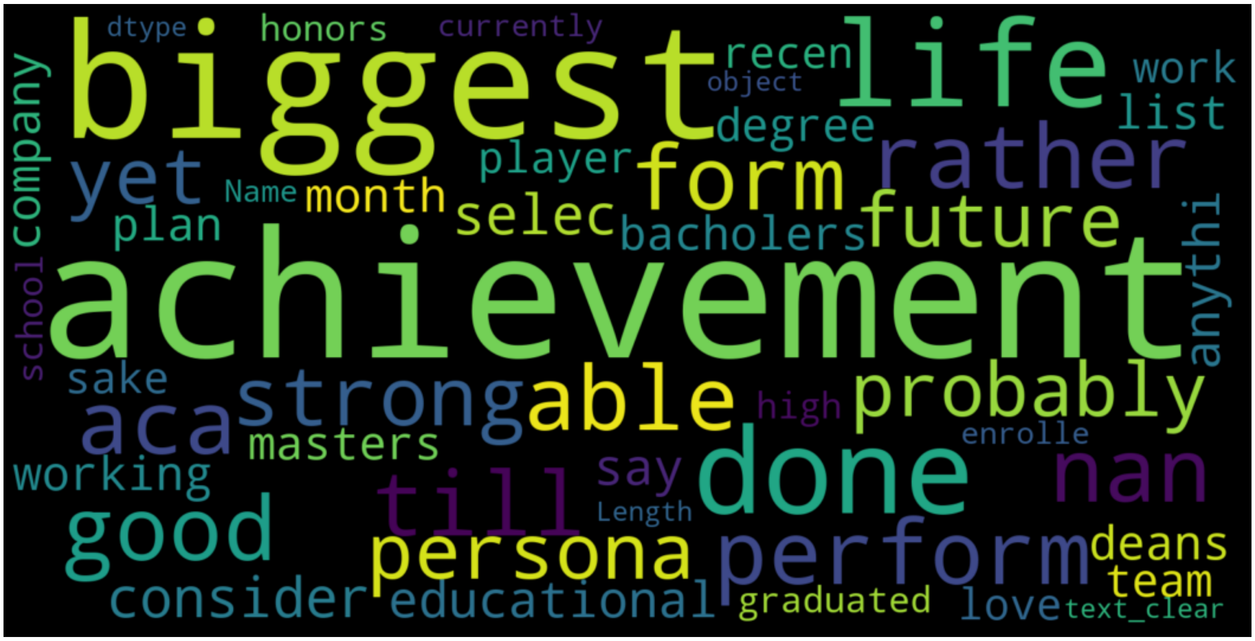


**How candidates would like to add values**



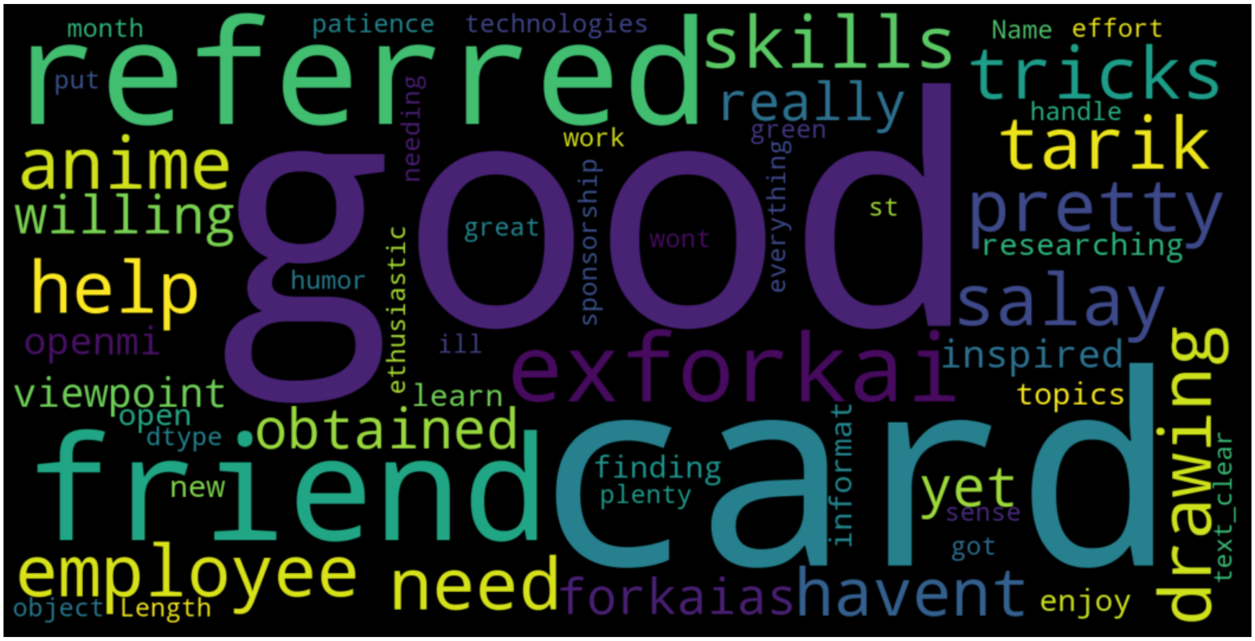
Candidates would like to contribute to the company by their skillsets in statistics, data analysis, data mining. They also would like to show their passion and teamwork as soft skills.

**Biggest Achievement**



Regarding biggest achievement, candidates shared a lot about their educational achievement such as master degree, bachelors degree, dean list, honors awards. They also had good performance in their past experience, projects or some moments in life.

**What else candidates would like to share**



Many candidates were referred by their friends to apply for Forkaia. They also show their willingness to help and enthusiasm.