PROJECT REPORT - DATA SCIENCE

CURRENT TRENDS IN THE DATA SCIENCE AND MACHINE LEARNING INDUSTRY

Problem Statement

- Comparing the current trends in machine learning and data science and comparing it with the previous year trends.
- Analyzing the learning(student) community in data science and various distributions.
- Inferring on how could a student get well equipped to get into industry

Introduction

The pandemic has necessitated online education and remote work and has led to students as well as professionals to explore more in the tech field. Machine learning and Data Science have become the “buzz” words. But almost everyone might have confusion on which platform to begin with, which algorithm to learn in short where to start. Hence we have visualised the data from previous year and compared the data with 2021 to get an in-depth knowledge on various tools and algorithms that might have seen a rise or fall in its popularity in 2021. This summarizes the skill set required for a person who is getting into data science.

Data overview

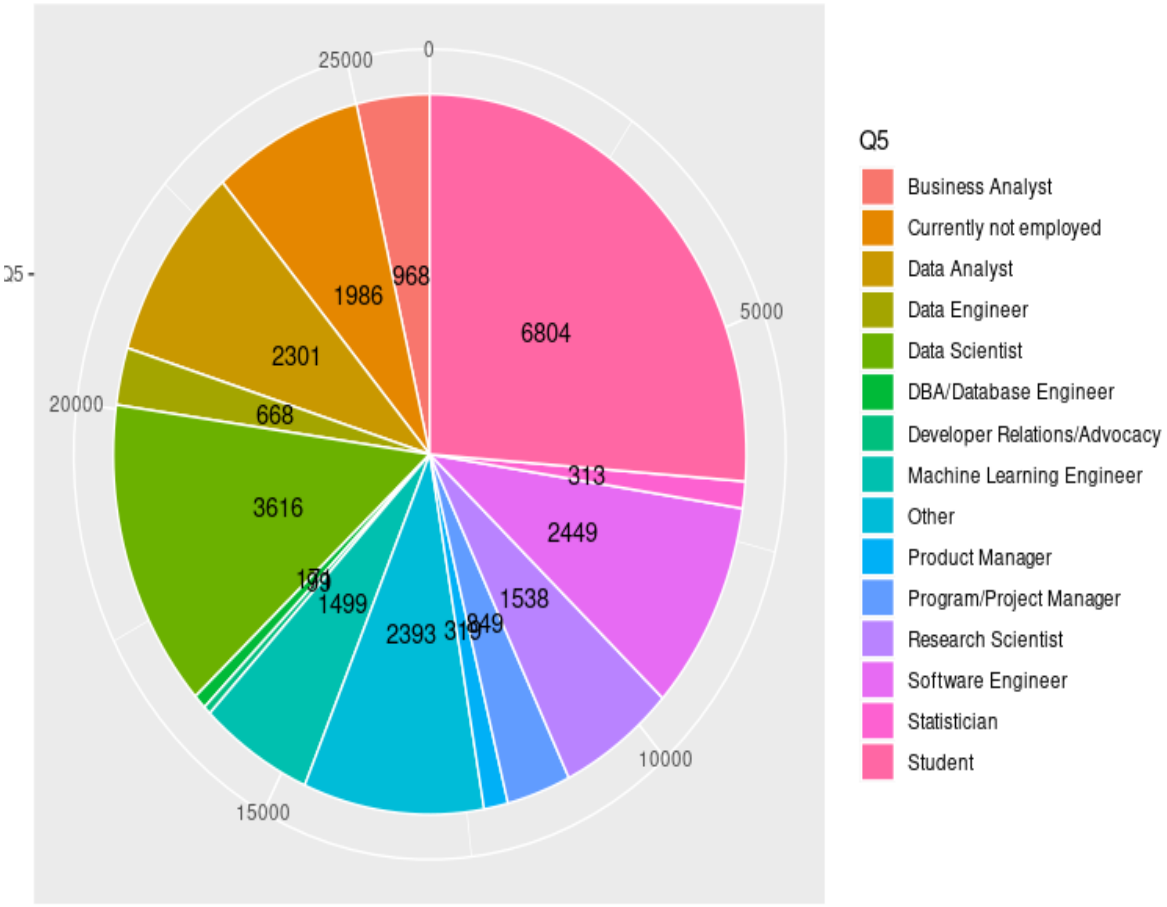
This dataset has been formed by using a kaggle survey. Students as well as working professionals were asked to participate in the survey . This dataset contains information on the gender and age of the average kaggle user, their position in the tech world, and techniques and tools they use the most. Information on preference has also been considered.

Methodology

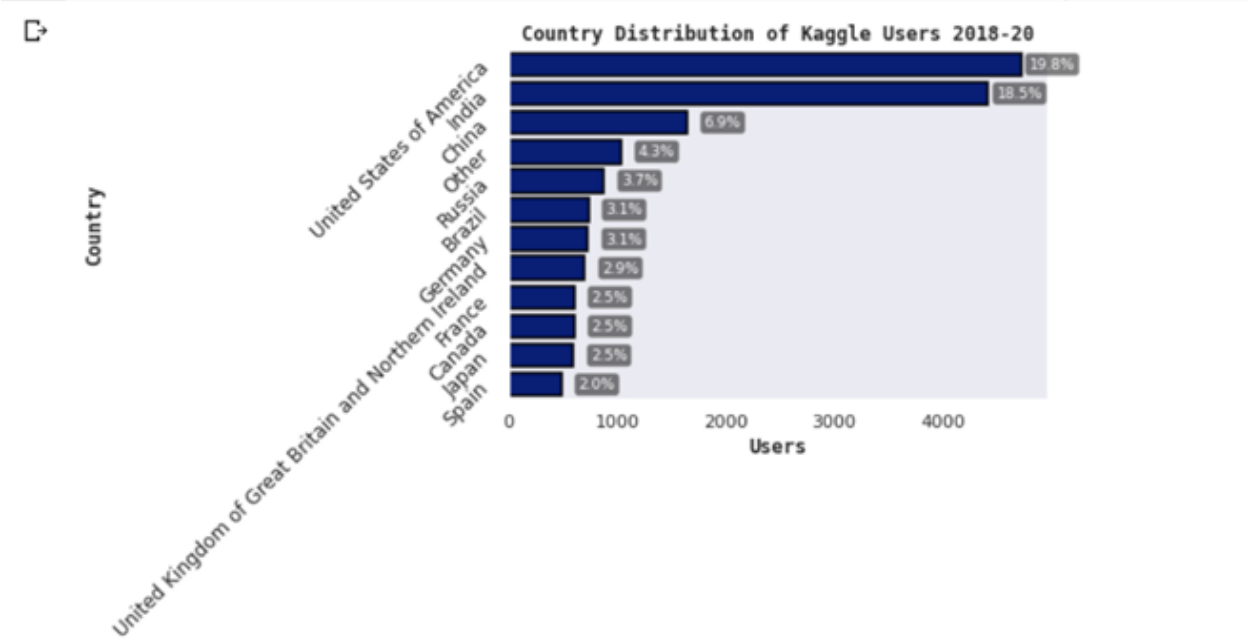
Visualisation of data using various types of graphs such as bar graphs, pie charts get the maximum, get the range and average and do comparisons on the results obtained.For this purpose we have used various libraries like Seaborn, GGplot and matplotlib.The programming languages used are R and python

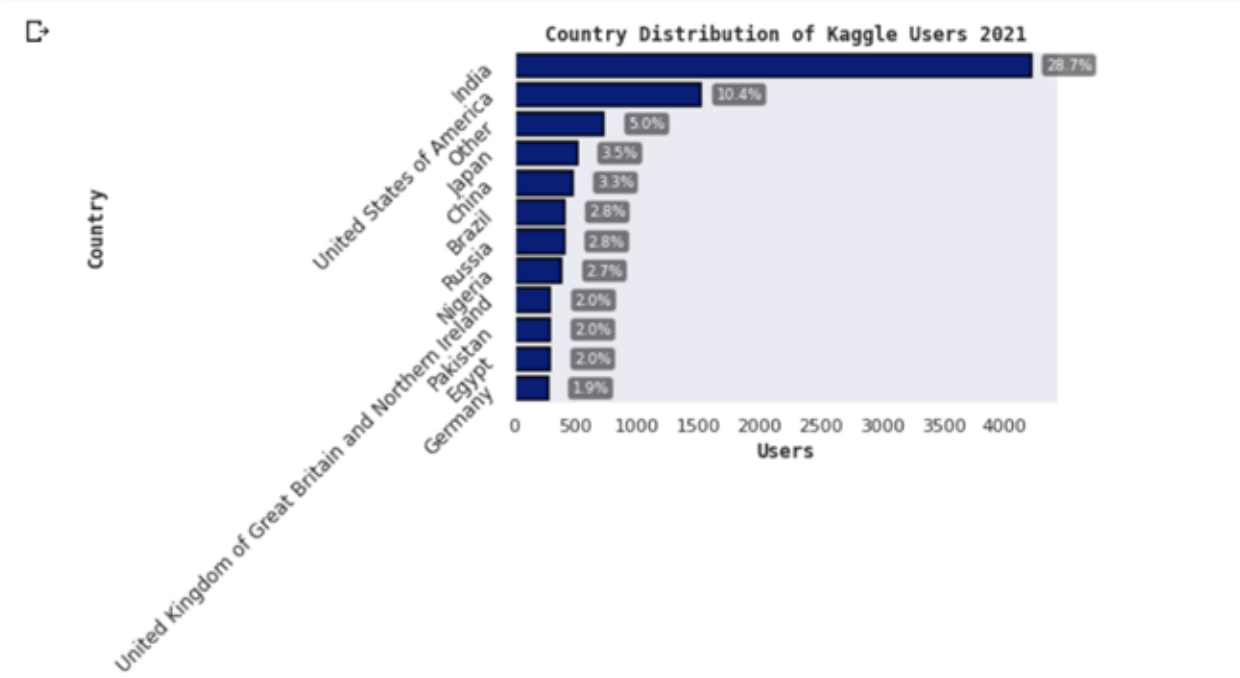
Survey distribution

- Students community are the major users of kaggle/the data provided(the survey) has the highest chunk of students
- It is followed by data scientists,software engineers,machine learning engineers
- It is also used by people from various fields like advocacy,project managers, database engineers

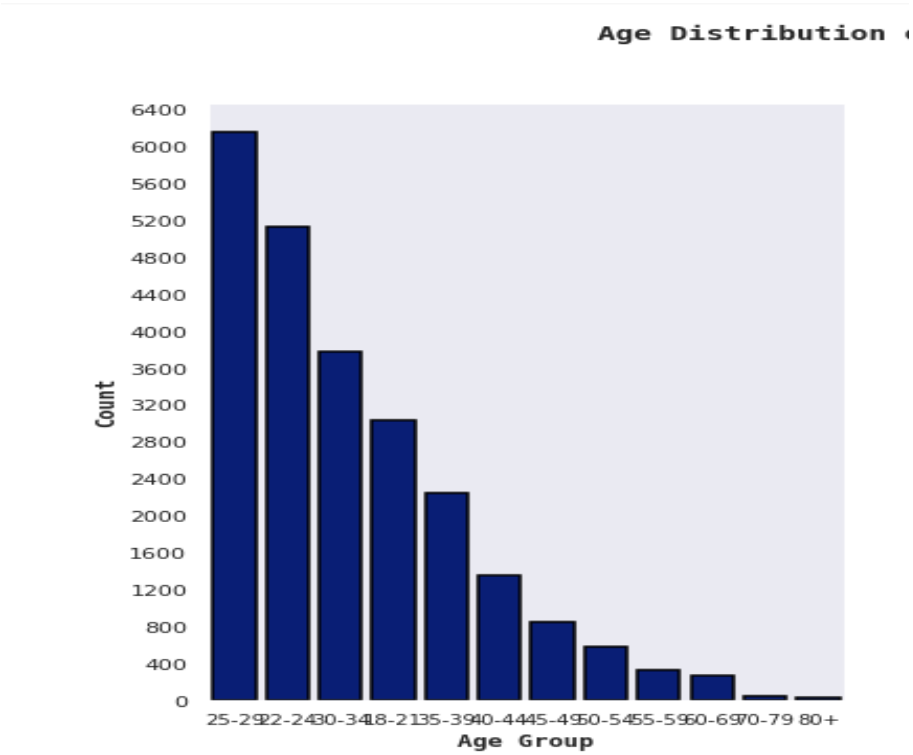

0
5000
10000
15000
20000
25000
968
1986
2301
668
3616
199
1499
2393
319
849
1538
2449
313
6804
Q5
Business Analyst
Currently not employed
Data Analyst
Data Engineer
Data Scientist
DBA/Database Engineer
Developer Relations/Advocacy
Machine Learning Engineer
Other
Product Manager
Program/Project Manager
Research Scientist
Software Engineer
Statistician
Student

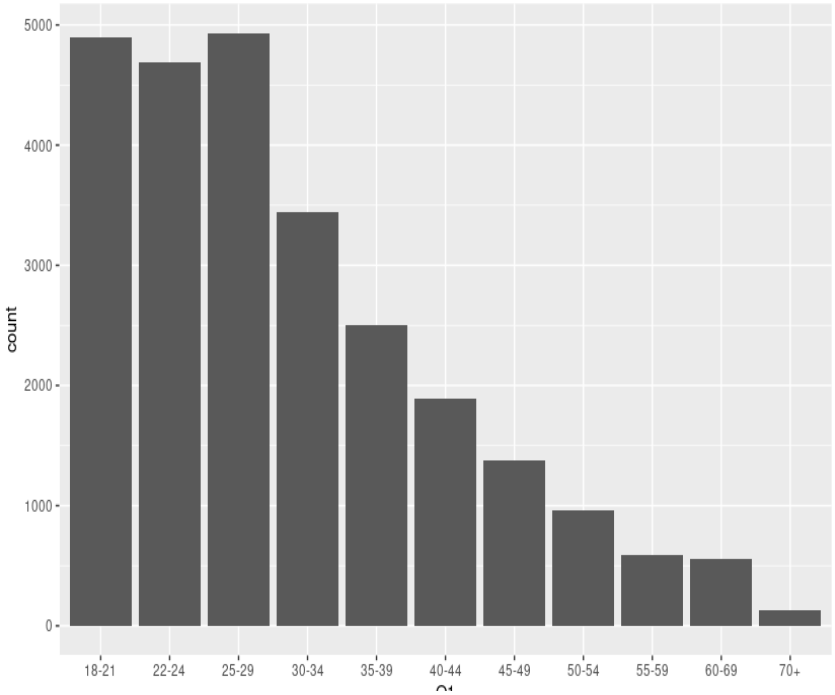
Country-wise kaggle users



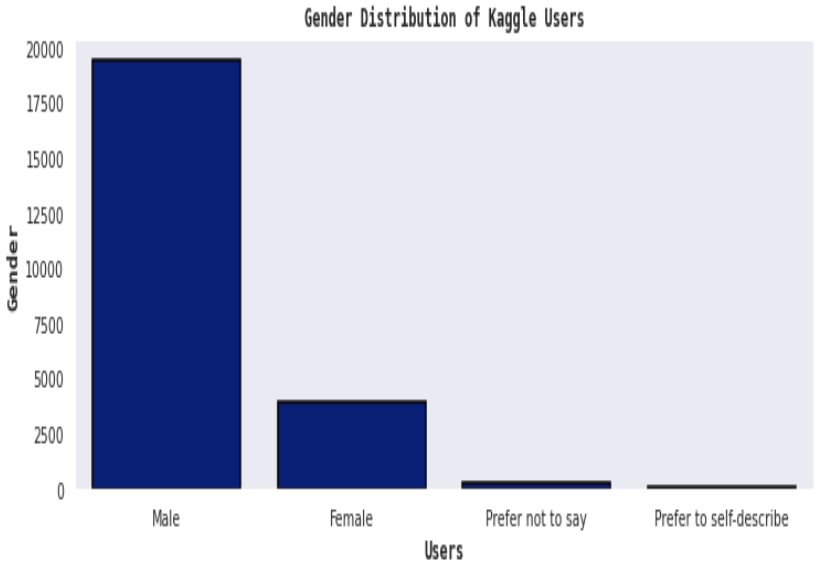


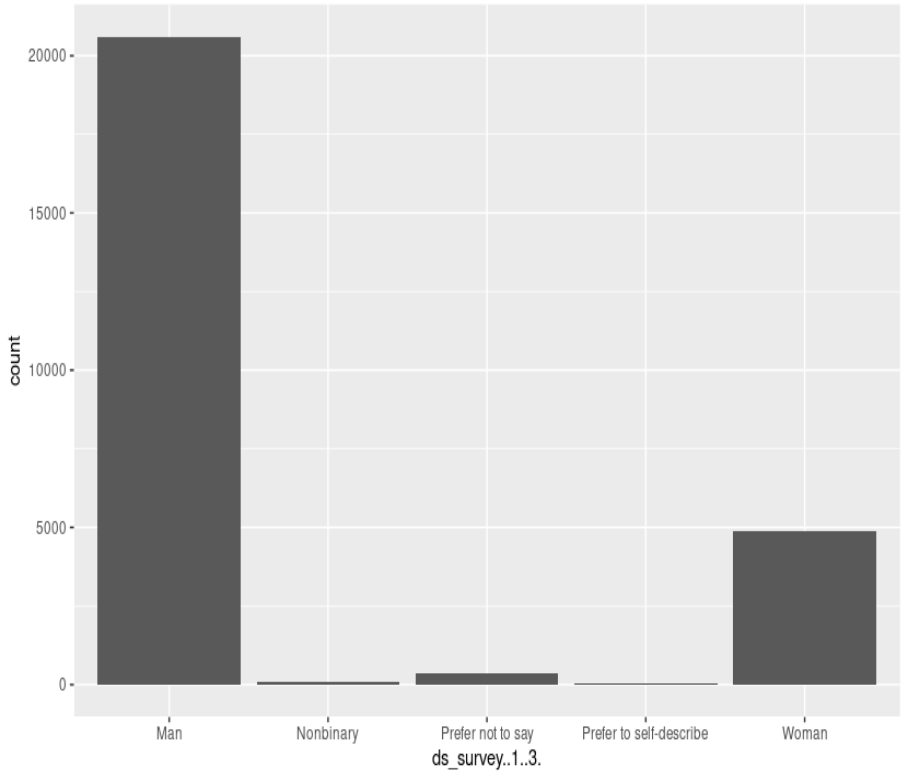
Age distribution





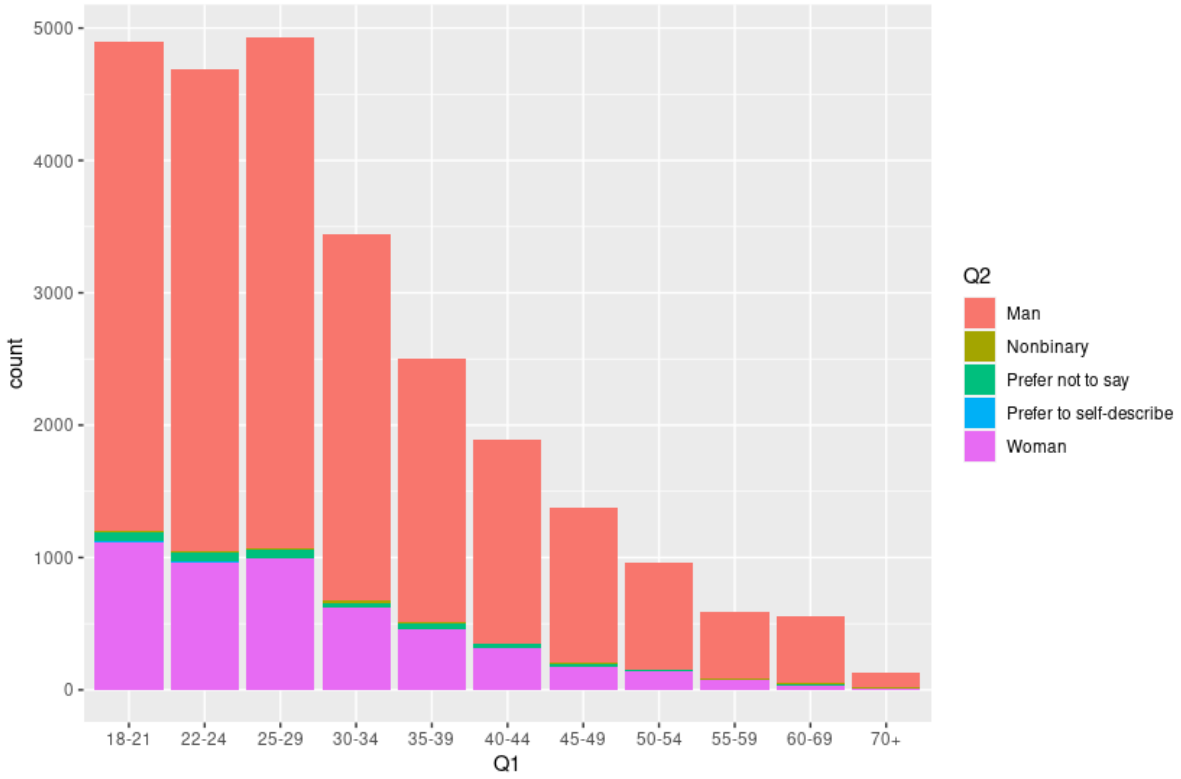
Gender distribution





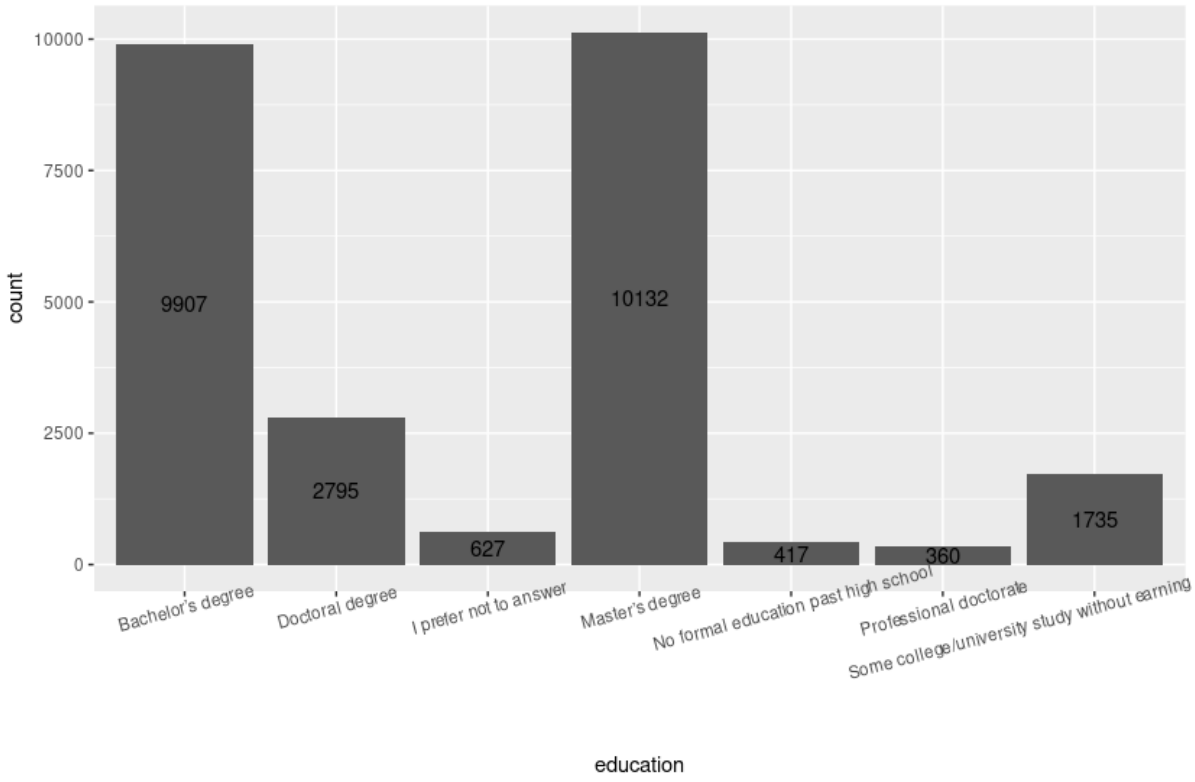
Age-gender

- Age-group of 18-21 has most women users(could be students).
- The highest users in total are in 25-29 however highest women users in 18-21.
- The difference in the number of women in ds/ml is less than half the number of men.



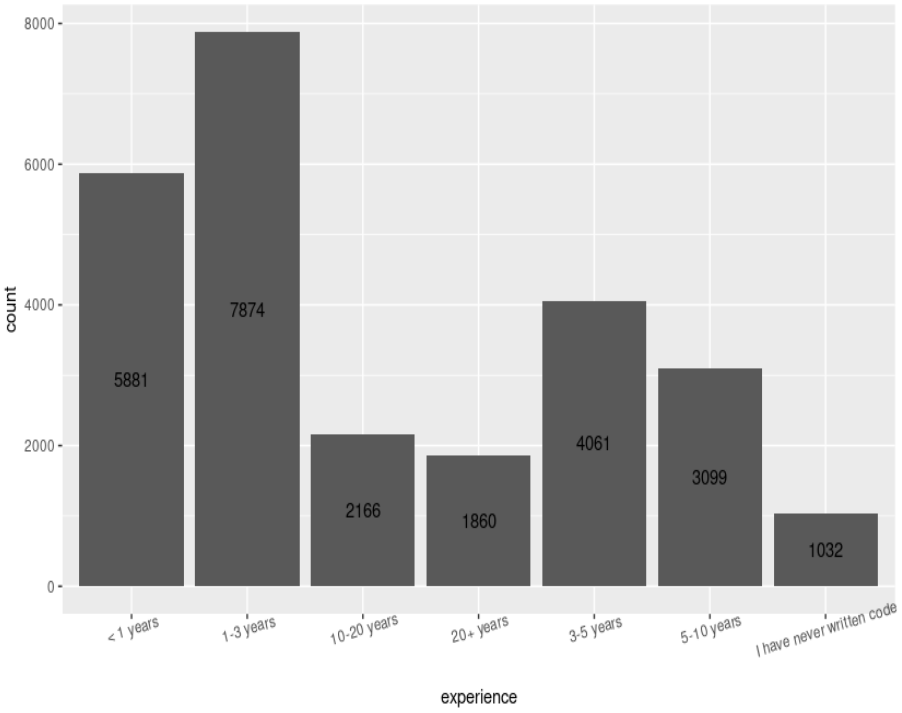
Education

- Most of the population into ml or ds are master degree holders followed by bachelors.
- We see that people who have no formal education are also into data science.

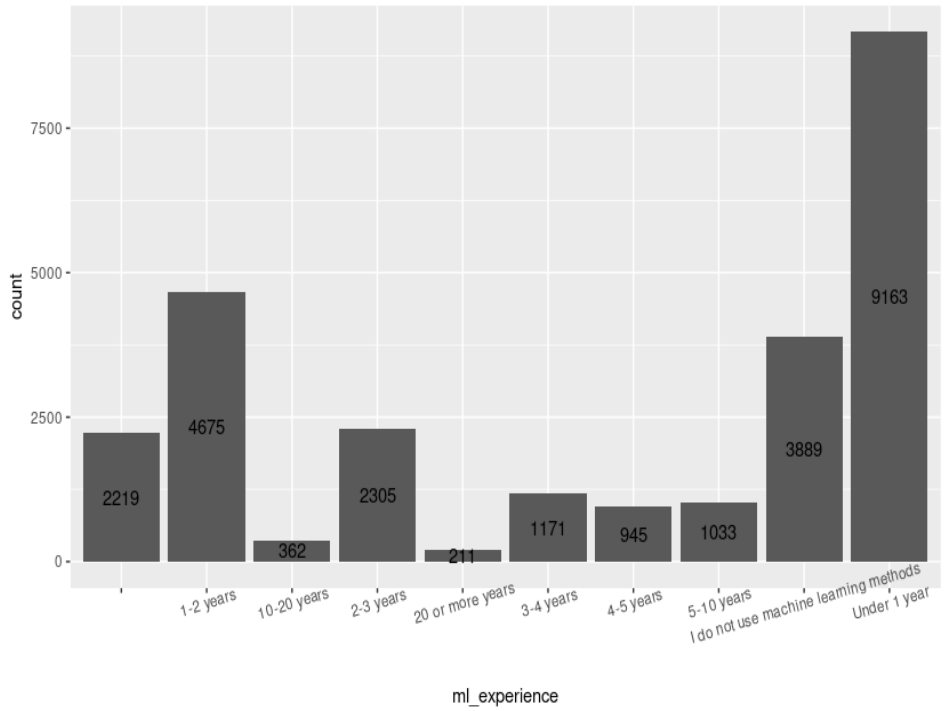


Experience

Experience in coding

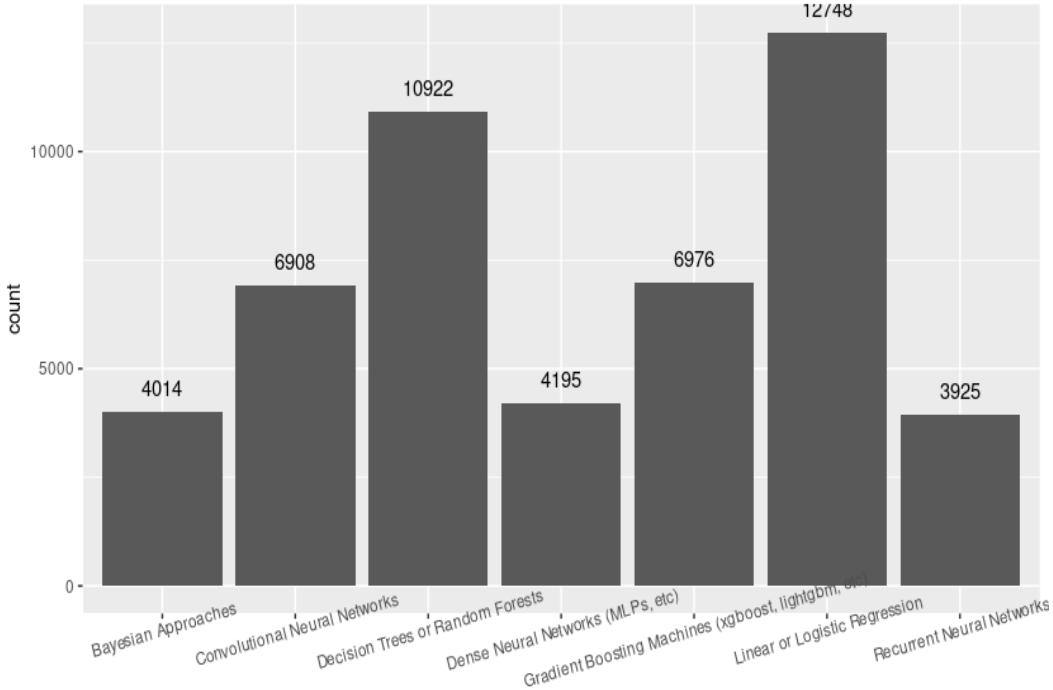


Ml experience



Machine Learning algo stats

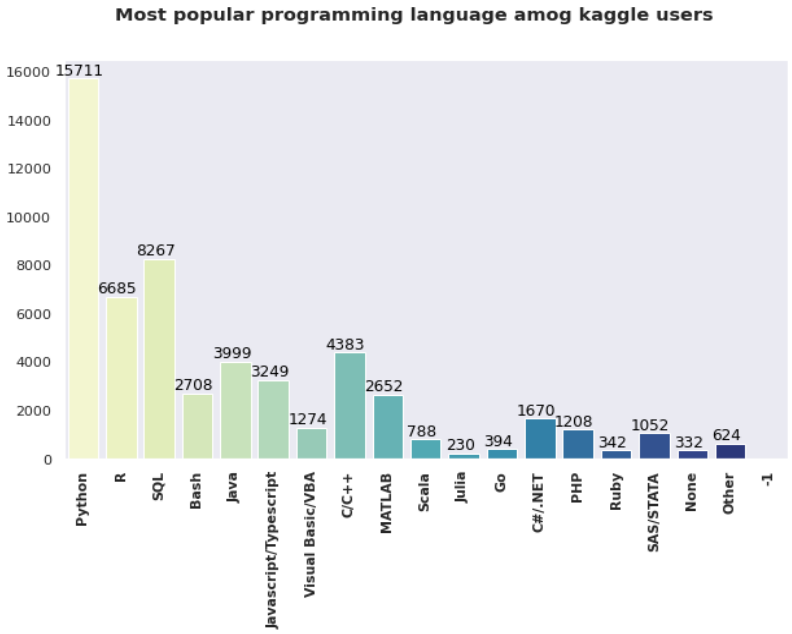
- Most popular ml algorithm is Linear/Logistic regression followed by decision tree and Xgboost
- All these are in-demand, highly used ml/dl algorithms

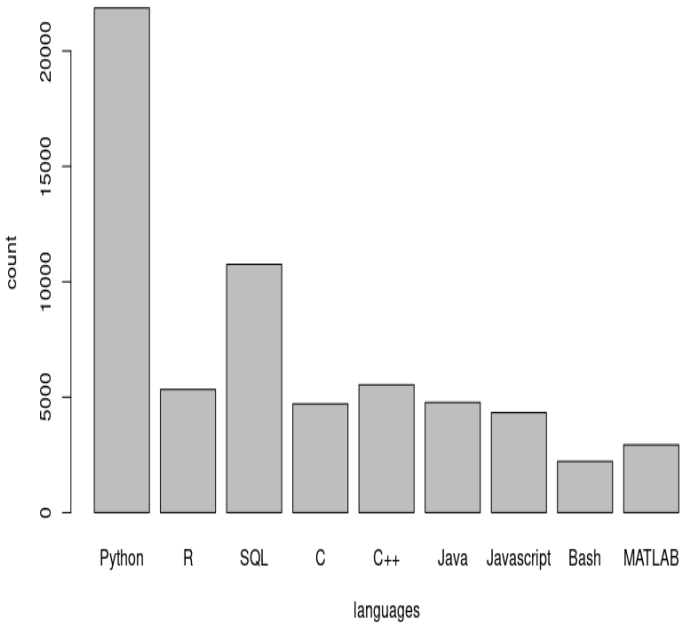


Programming Languages
2020

Most popular language: Python(15711)

Least Popular language: Julia(230)





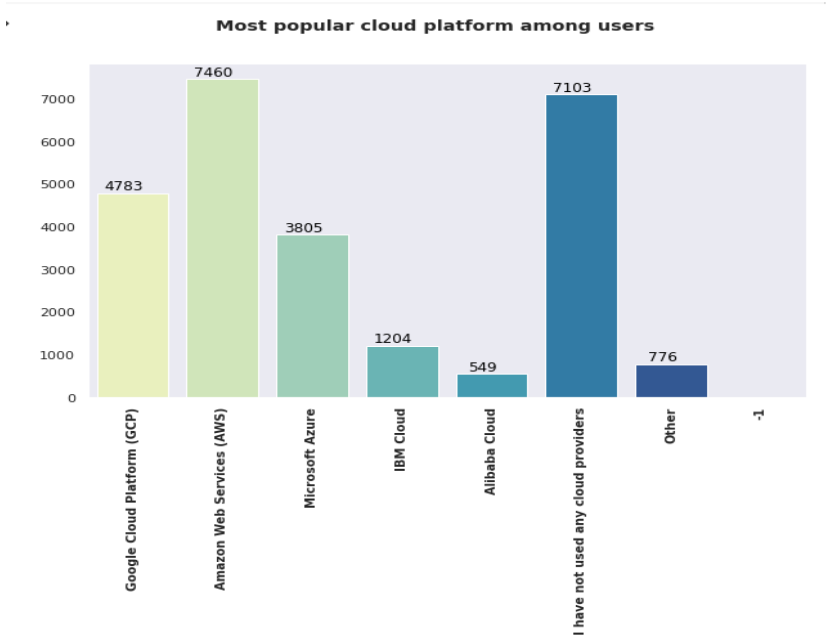
2021

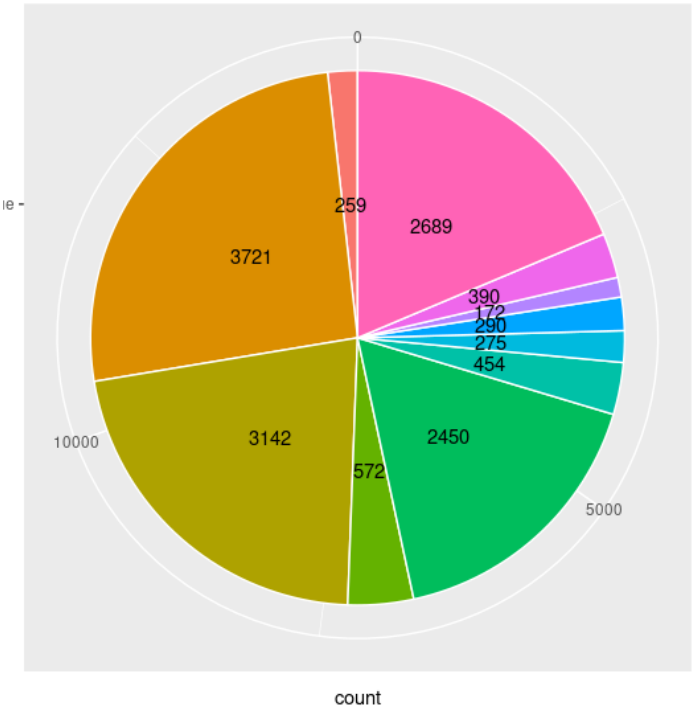
We observe that the trends remain the same and Python is the most used language followed by SQL in data science

Cloud Platform
2020

Most popular : Amazon web services(7460)

Least popular: Alibaba cloud(549)





2021

AWS, GCP and Azure remain to be the most used cloud platforms. The overall users of cloud platforms has increased by around 5000.

Popular job roles
2018- 2020

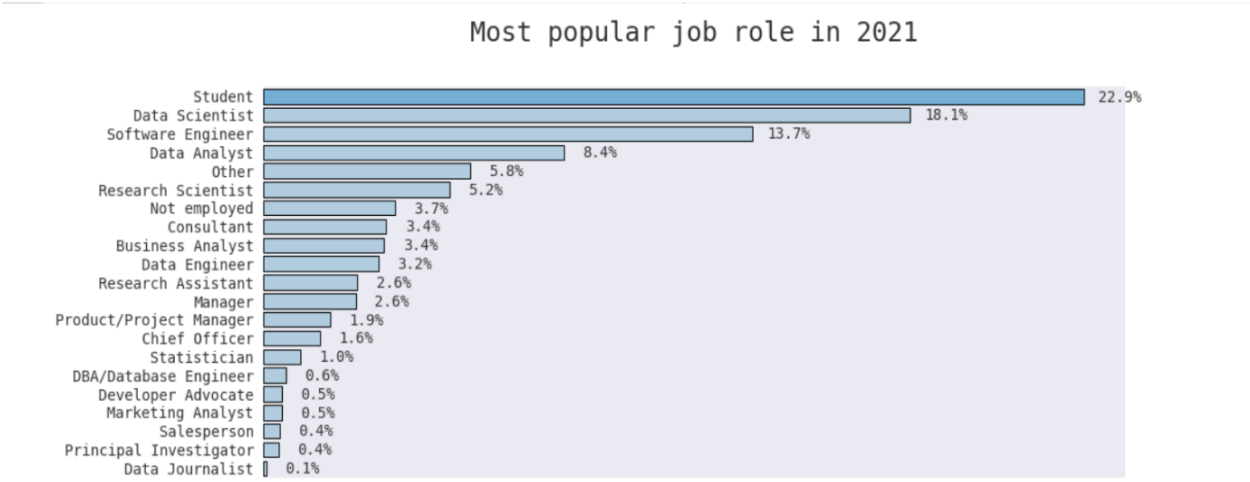
We can observe the most popular job role is computer/technology and least popular is

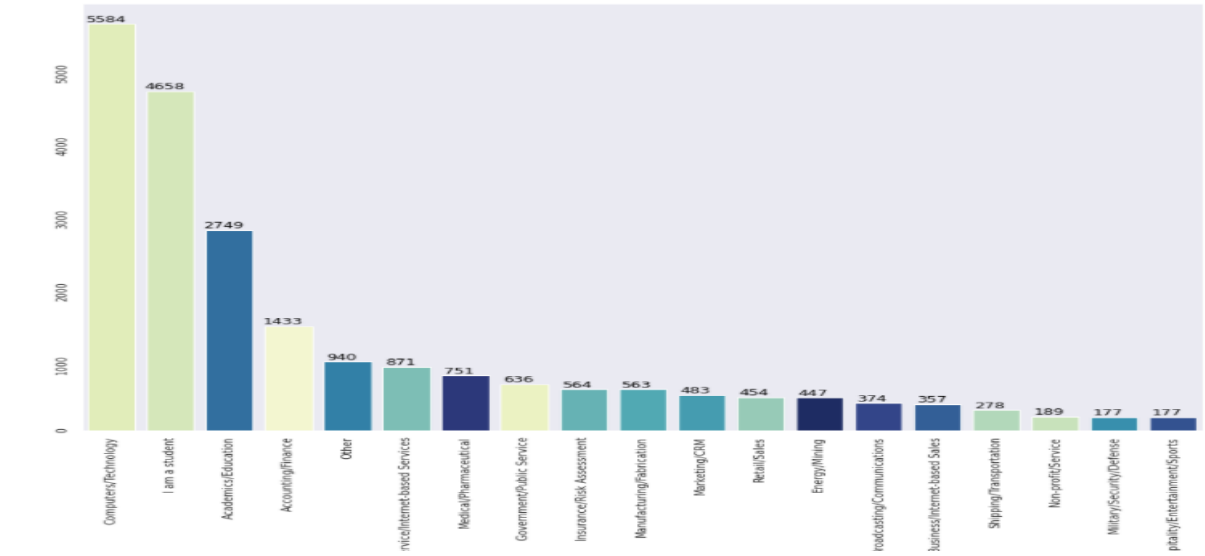
Hospitality entertainment/ sports.

Popular Job roles in 2021
2021

Most popular job role: Student

Least popular job role: Data Journalist

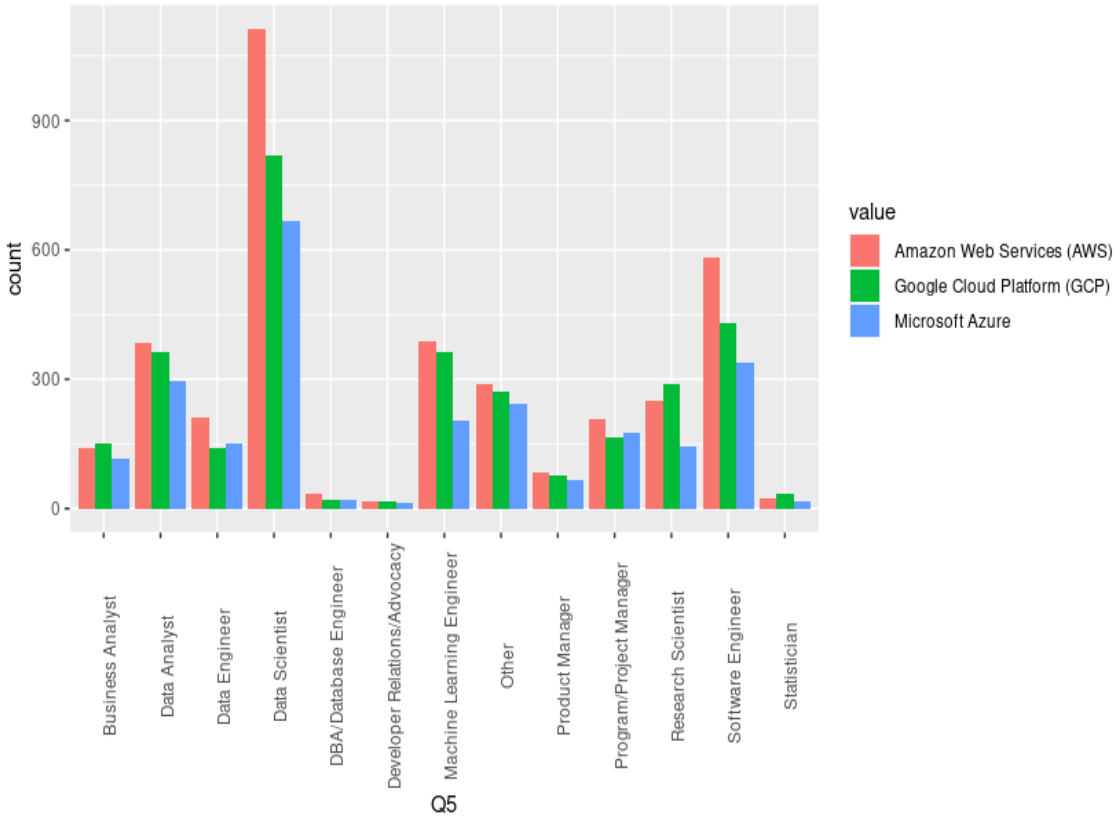




cloud-role

This is a graph that tells the usage of cloud by each role

- AWS is most used by most of the professions.
- We see that research scientist,statistician,business analyst,student use GCP more.
- But we see that the difference in users is less where gcp is used more
- Microsoft azure is almost used as much as gcp

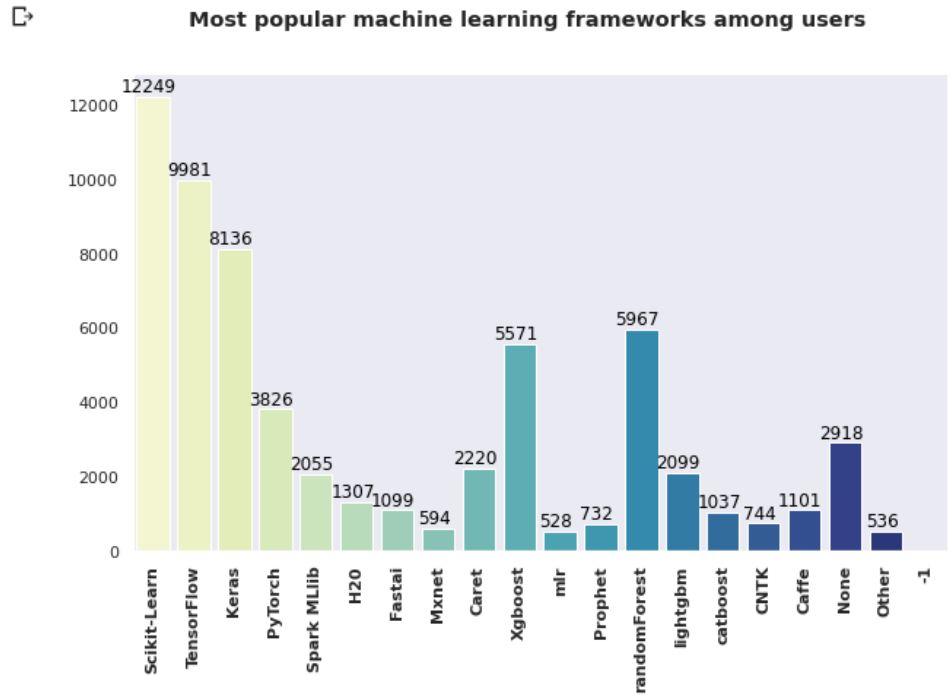


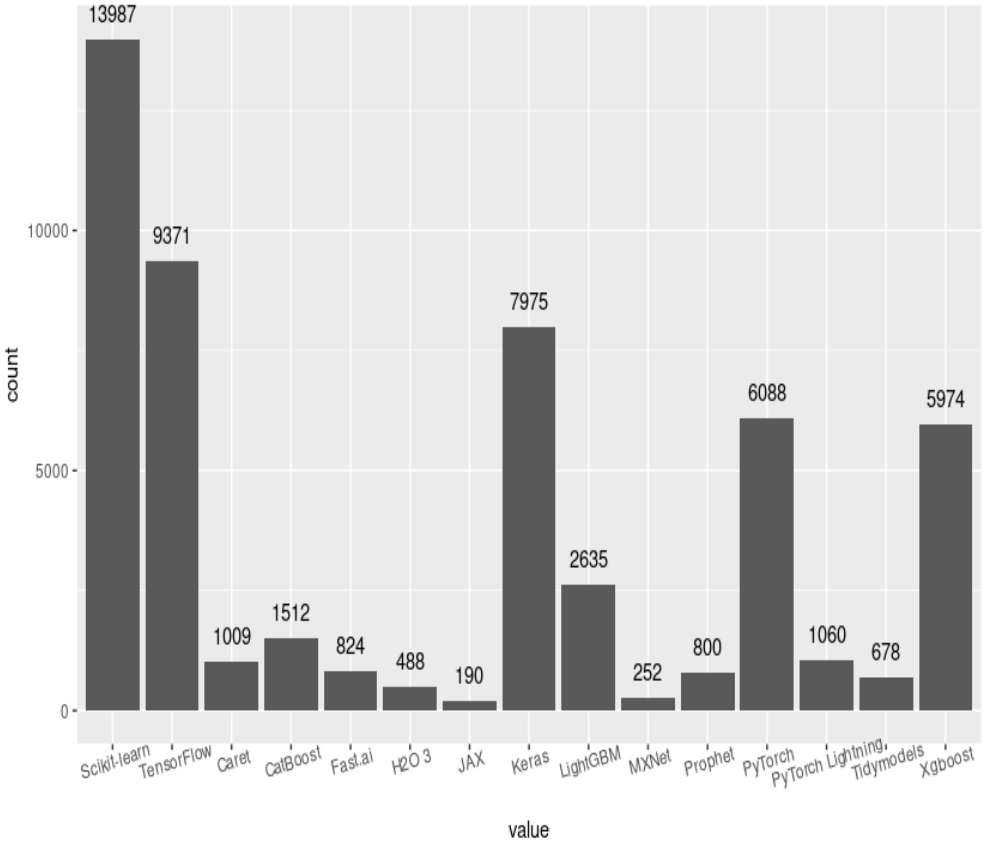
framework

2020

Most popular: Scikit-learn

Least popular: mlr


Most popular machine learning frameworks among users
12249
9981
8136
3826
2055
1307
1099
594
2220
5571
528
732
5967
2099
1037
744
1101
2918
536
0
2000
4000
6000
8000
10000
12000
Scikit-Learn
TensorFlow
Keras
PyTorch
Spark MLlib
H20
Fastai
Mxnet
Caret
Xgboost
mlr
Prophet
randomForest
lightgbm
catboost
CNTK
Caffe
None
Other
-1

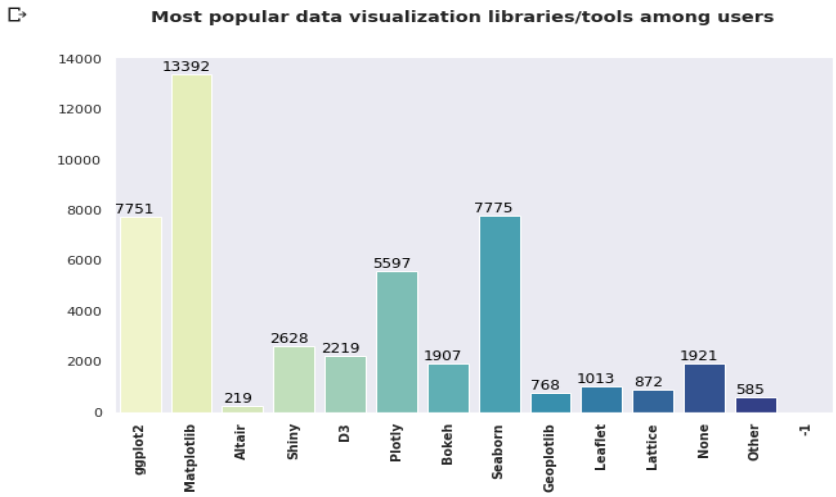

13987
9371
1009
1512
824
488
190
7975
2635
252
800
6088
1060
678
5974
0
5000
10000
count
Scikit-learn
TensorFlow
Caret
CatBoost
Fast.ai
H2O 3
JAX
Keras
LightGBM
MXNet
Prophet
PyTorch
PyTorch Lightning
Tidymodels
Xgboost
value

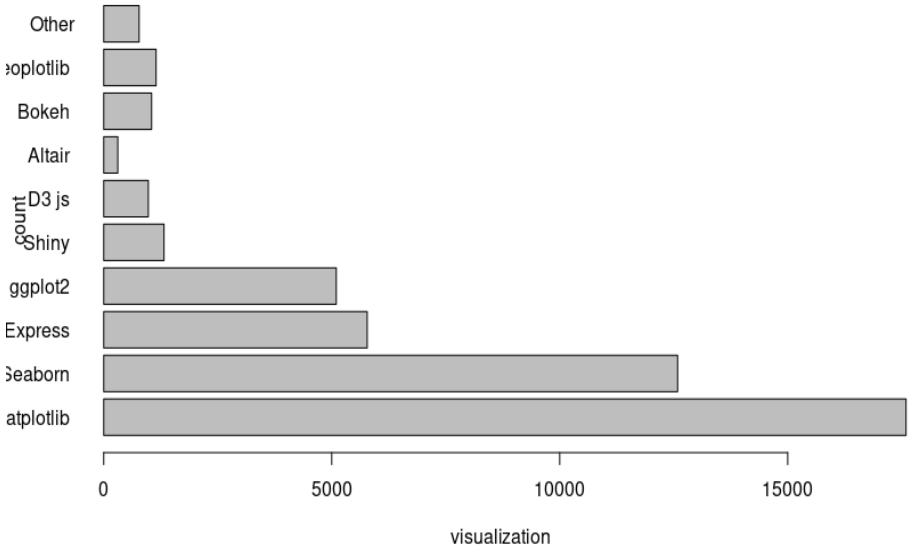
Visualization Tool/library

2020

Most popular: matplotlib(13392)

Least popular: altair(219)



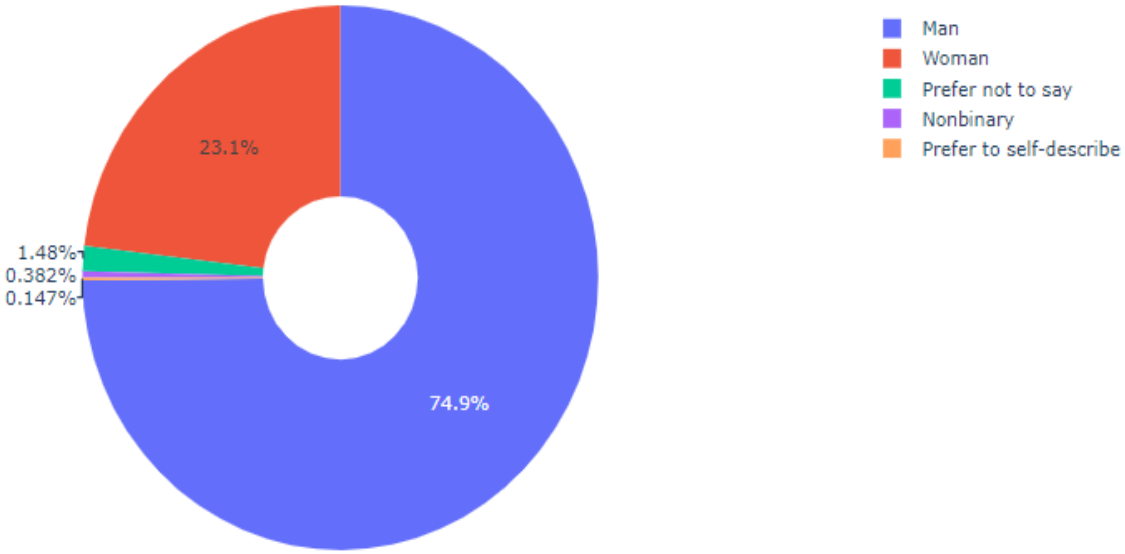


2021: Matplotlib remains to be most used visualization tool increasing its popularity followed by seaborn and ggplot2

Role : Student

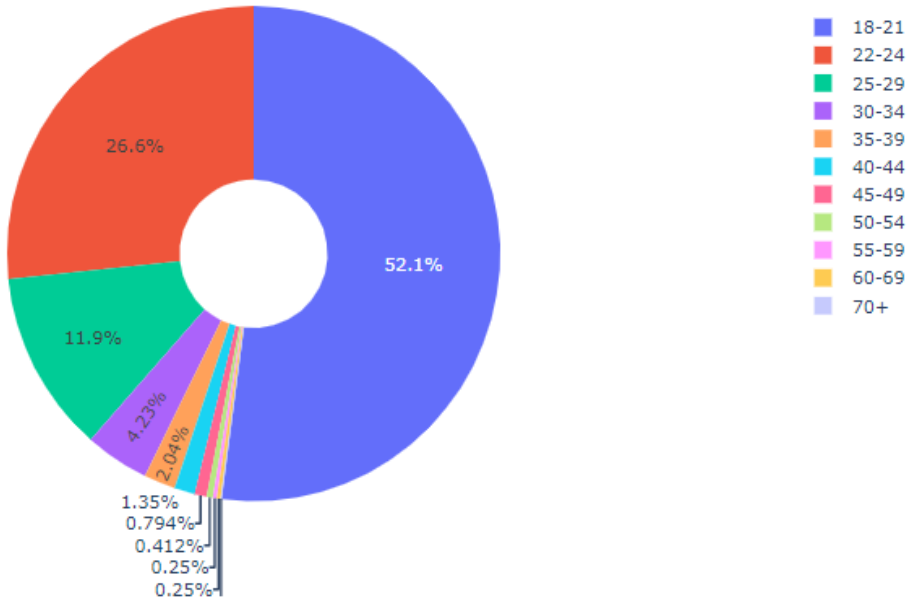
Gender

- We can see the gender gap in the student community.
- About 75% of the students are men.
- And only 23.1% are women.



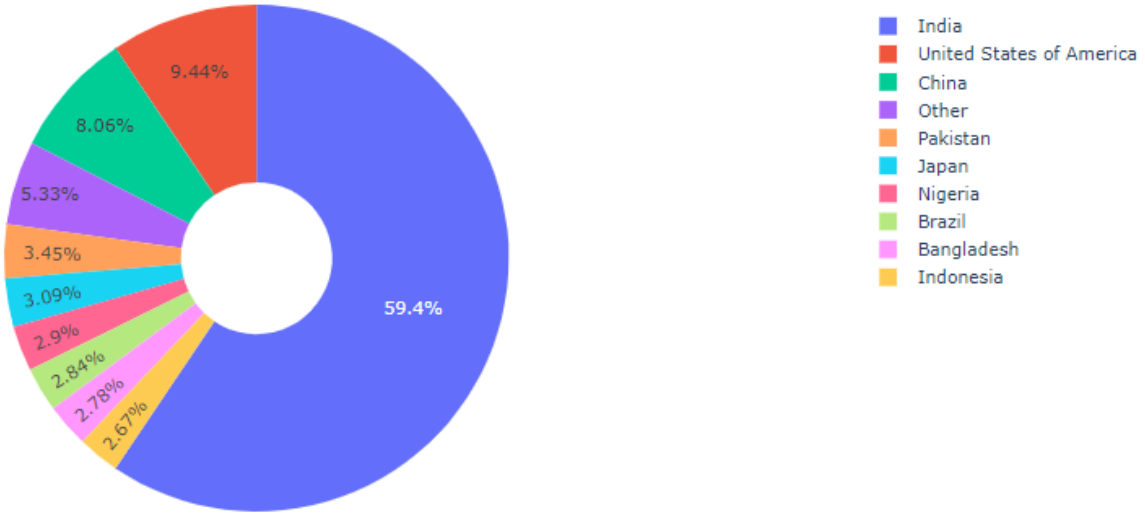
Age

- More than half of the student community aged less than 25



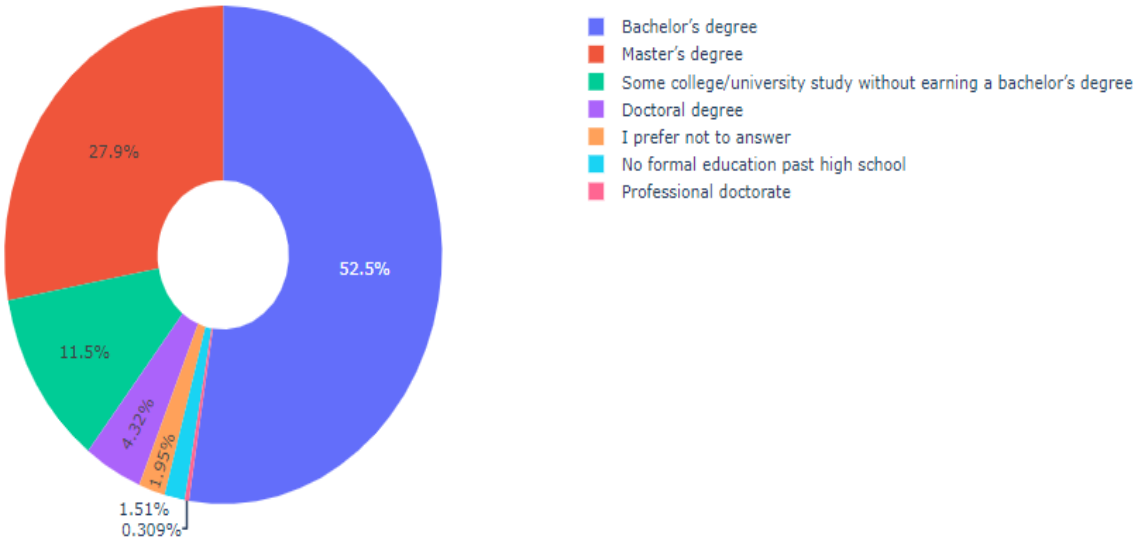
Top 10 countries

- The majority of students that actively took part in the survey are from India.



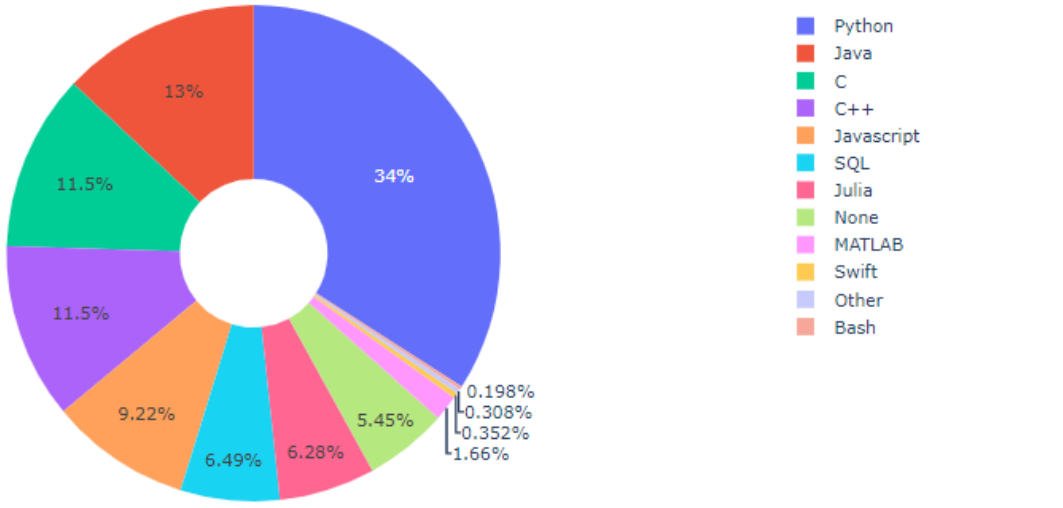
Educational Level

- Most of the students have the bachelor's degree or are planning to get bachelor's degree in next 2 years
- And a very good percent are having Master's degree in their hands.



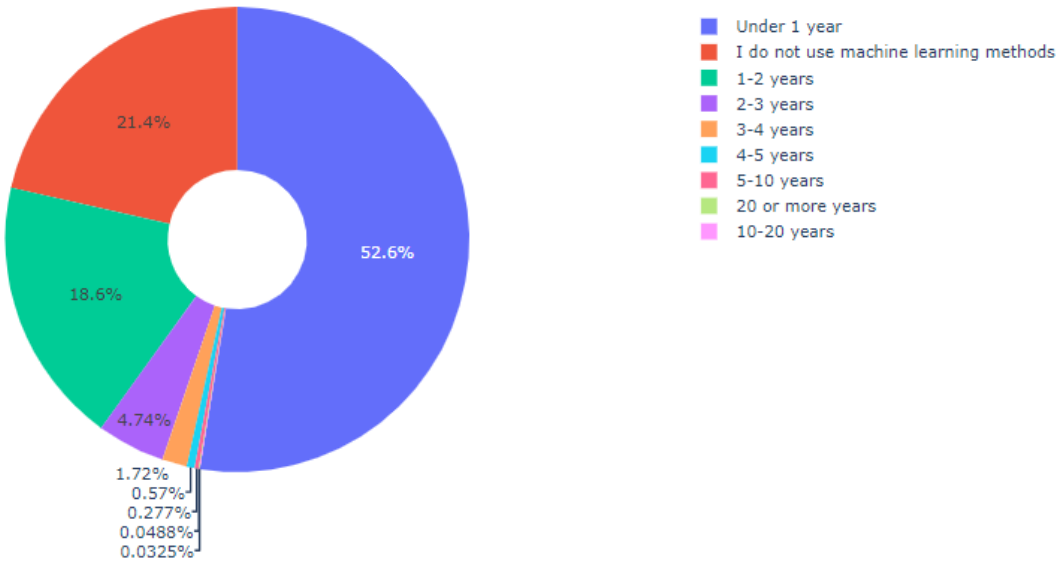
Popular Programming Languages

- Most popular language among students is Python,followed by Java
- C and C++ are equally popular among the students



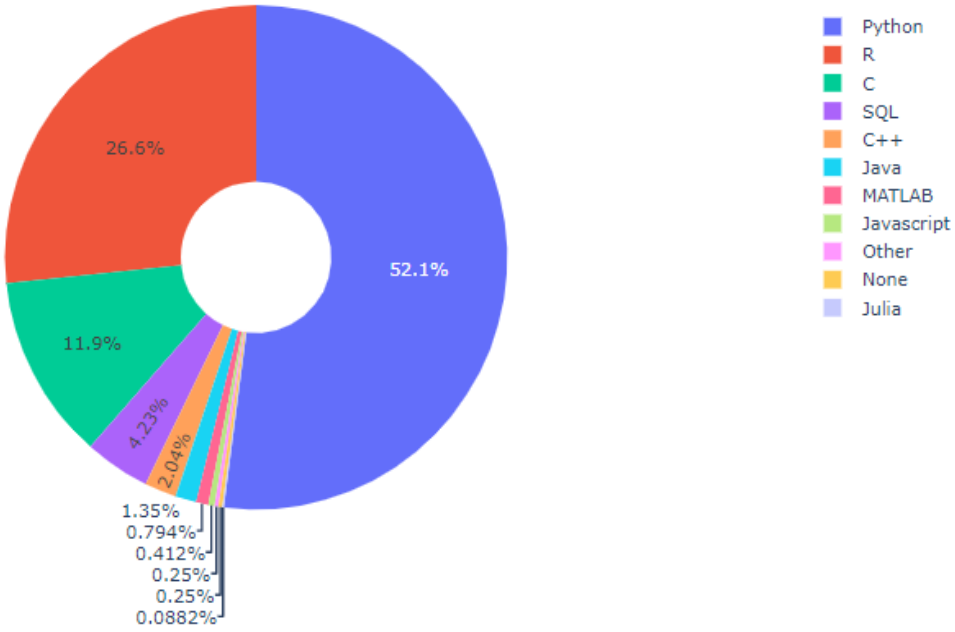
For how many years have you used machine learning methods?

- Most of them are using ML methods for under 1 Year.
- We can relate it to why ML has became popular nowadays.



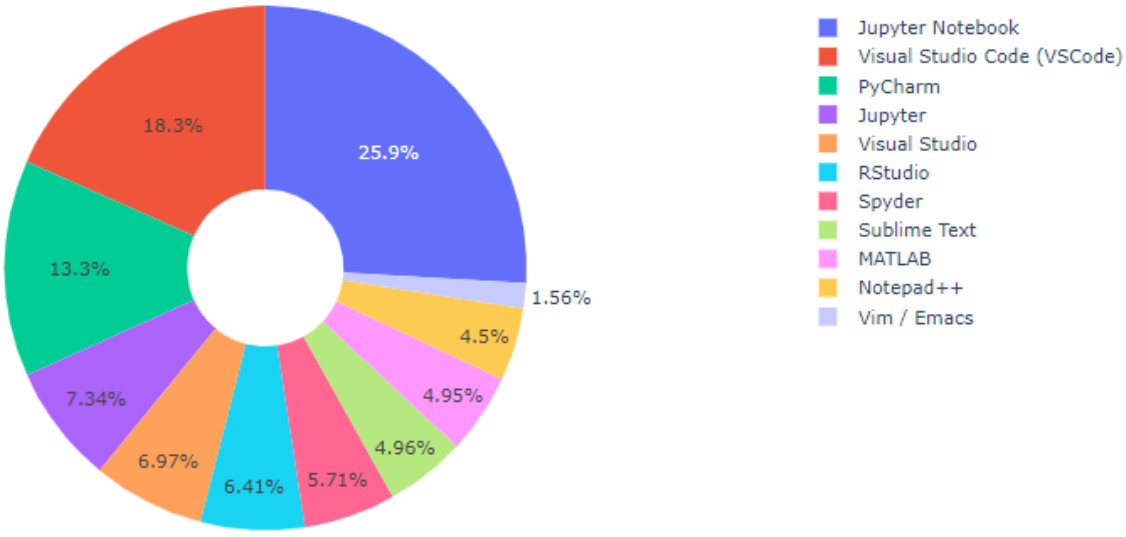
Recommended Language

- More than 50% of students recommended Python



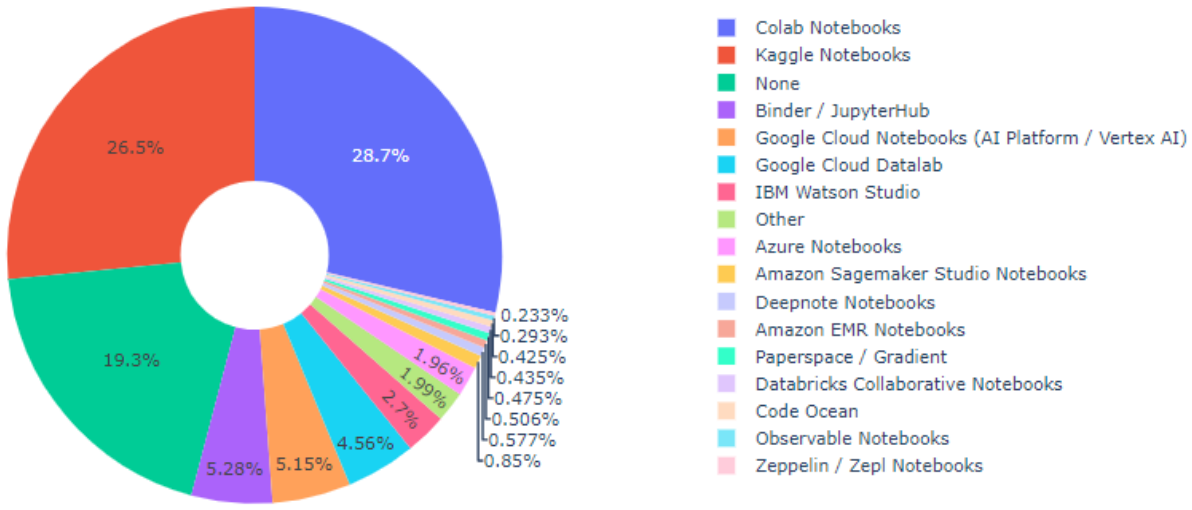
IDE

- The most popular IDE among students is Jupyter Notebook, which is followed by Vscode and pycharm.



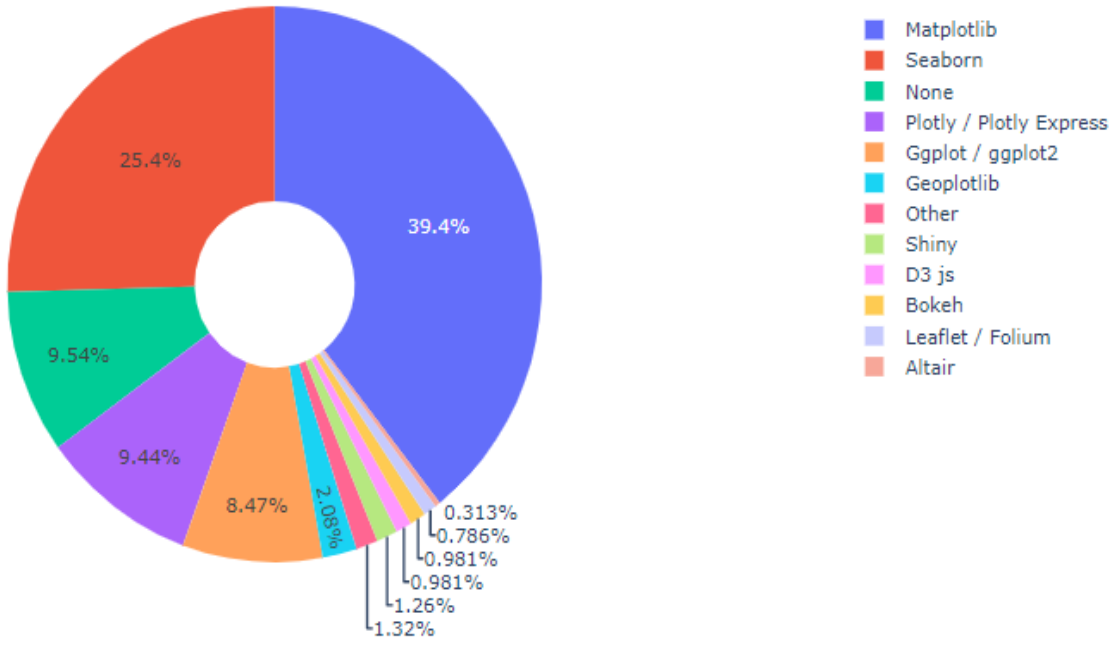
Hosted Notebook Products

- Most of them are using Colab and Kaggle Notebooks.
- We can't agree more as these are really popular among students.
- And we have to clearly notice that about 19% are not using the notebook products.



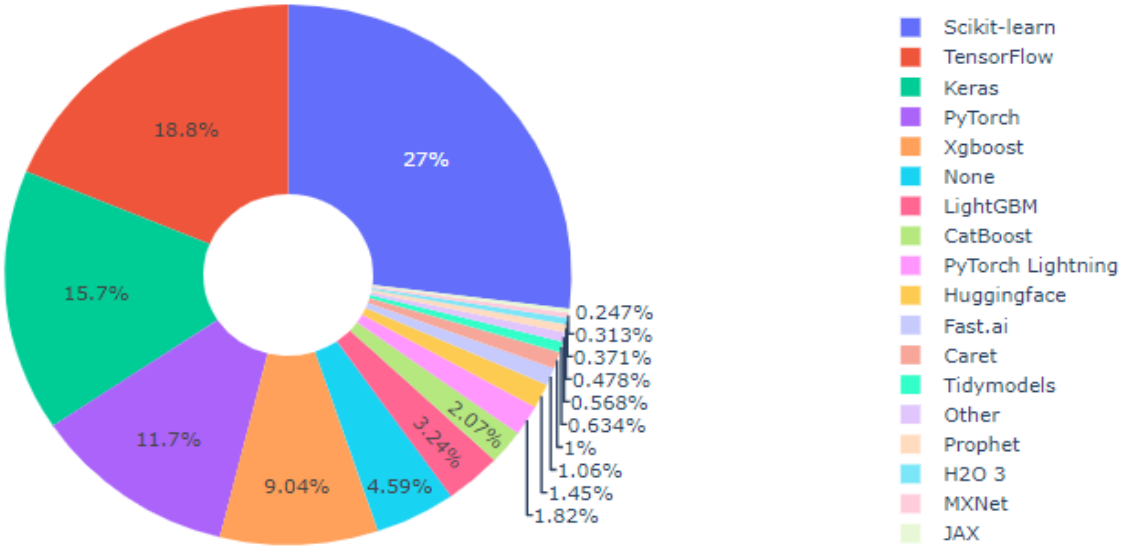
Data Visualization Libraries

- Matplotlib Tops the chart among all the other Visualization libraries which is followed by Seaborn. .



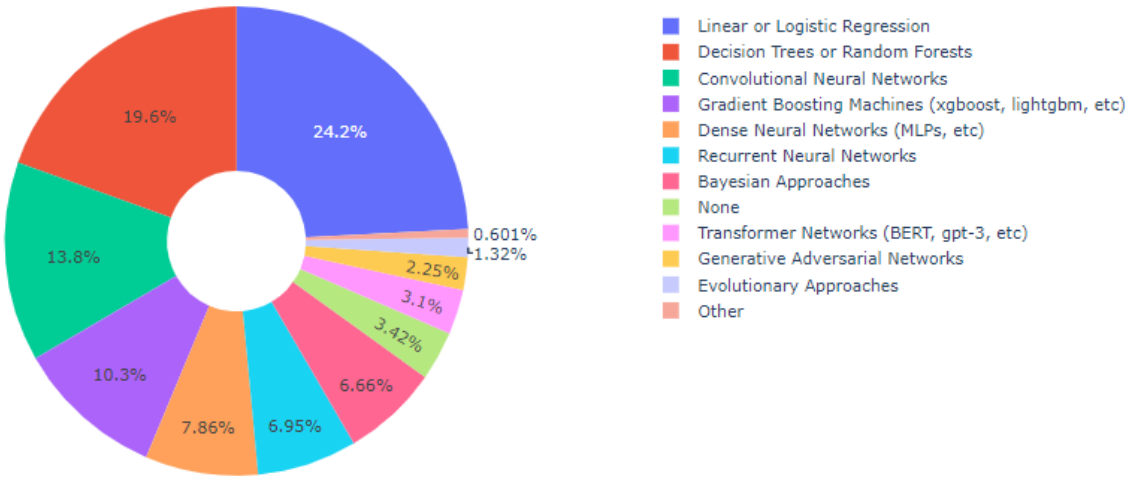
ML Frameworks

- Scikit-learn is the most favourite ML Framework among students followed by TensorFlow,keras and PyTorch.



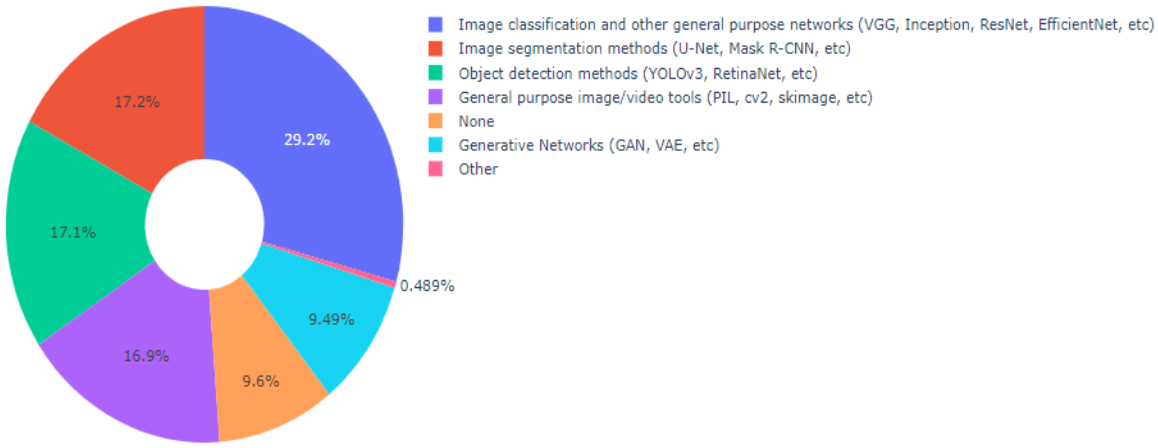
ML Algorithms

- Linear and Logistic Regression are the most commonly implemented ML algorithms by Students Followed by Decision Trees and CNN.



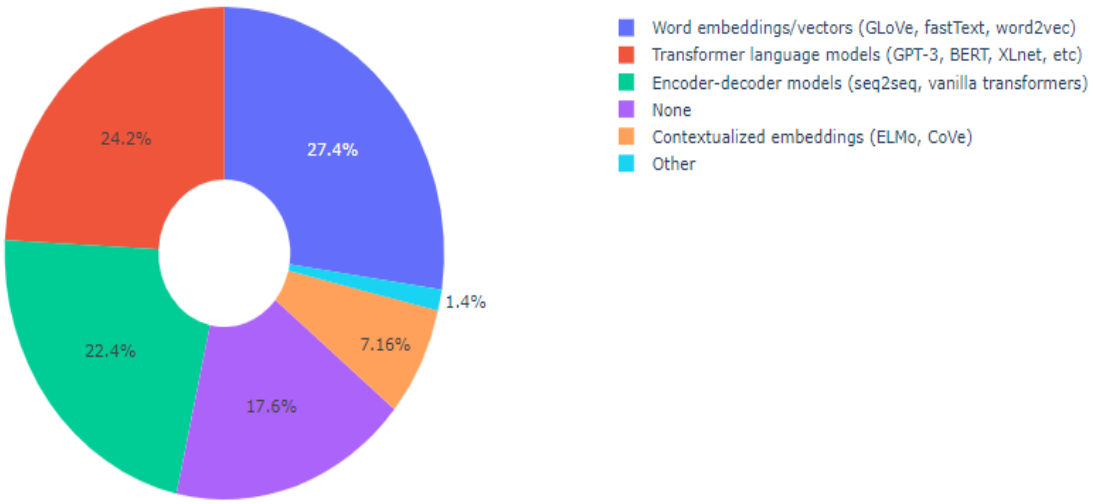
Computer Vision Methods

- Coming to Computer Vision Methods Image classification and other general purpose networks are more sought out methods among students.



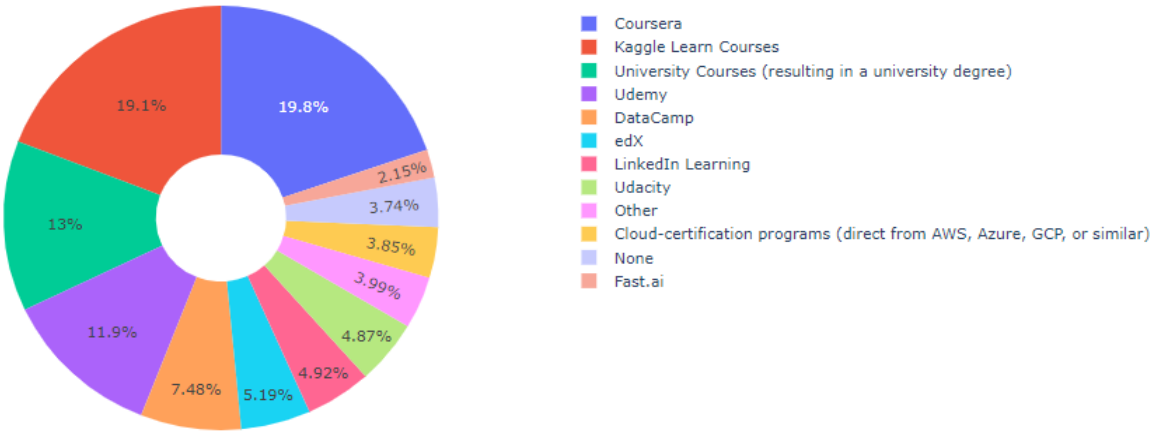
NLP Methods

- In NLP Methods we can see very less number of students are working on contextualized embeddings and others.



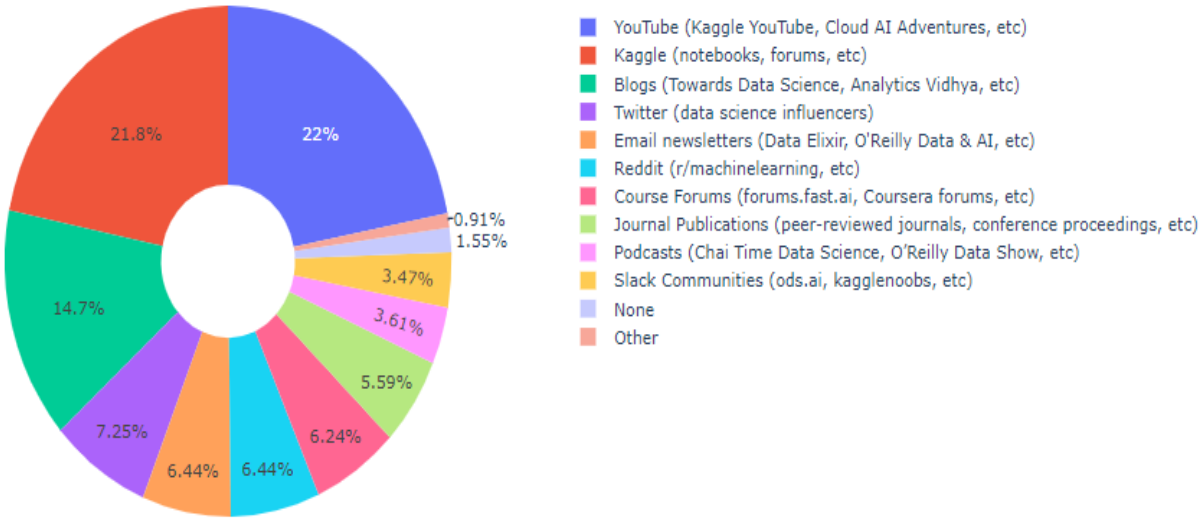
Data Science Courses

- Most of the students are learning Data Science Courses from Coursera and Kaggle Learn Courses.



Media

- Majority of the students utilise YouTube and the Kaggle website to catch up with current trends and develop new skills.



Inference

- Today we see ML and DS usage in various fields ranging from Data Scientist to a developer relations
- We saw the various skills(cloud, programming languages,ML algos, frameworks etc) that are in use in Data Science.
    - ➔ Languages: Python,R, SQL
    - ➔ Cloud: AWS/Azure/GCP
    - ➔ Frameworks: scikit-learn/Keras/tensorFlow
    - ➔ Visualization: matplotlib/seaborn/ggplot
- We also observed that the student community is well equipped with the demanded skills and the highest degree is bachelor's(from the survey).
- We can observe that most of the industry working community holds a master's degree
- We observe that both in the students and industry, the major composition is of men and this gender disparity is really alarming.
- The Internet (online platforms such as Coursera,YouTube,Kaggle), especially in today's world, is the biggest tool and is playing a major role in education.

Colab notebook links

https://colab.research.google.com/drive/1iKT-tk6YoLRAuXLt9gelGjXHAcpA43Uv?usp=sharing

https://colab.research.google.com/drive/1UIDlSAcxSZ8fvNf6J9PyUNkgAZIqonkK

https://amritauniv-my.sharepoint.com/:u:/g/personal/upsreya_am_students_amrita_edu/EWMFAhEKNd1Fu4SqSl6Xk-4Bl2Zl4LI8e9TCCeu10rboxg?e=vuQsej

TEAM MEMBERS

Appasani Deepthi (AM.EN.U4CSE19208)
Arya R N (AM.EN.U4CSE19209)
Siddhi Menon (AM.EN.U4CSE19251)
Srilekha Somanchi (AM.EN.U4CSE19254)
Ukkalam Pannaga Sreya (AM.EN.U4CSE19258)