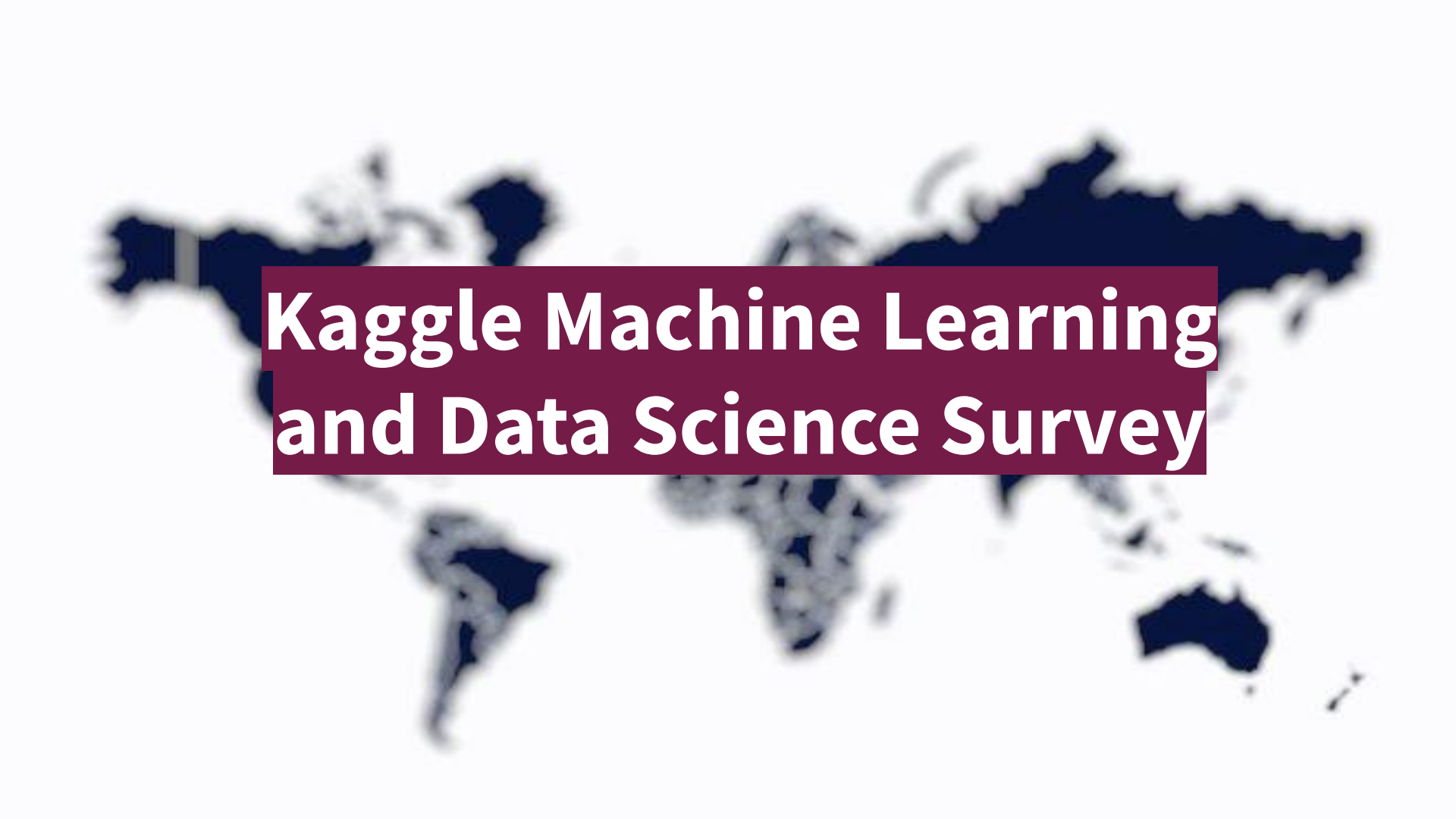
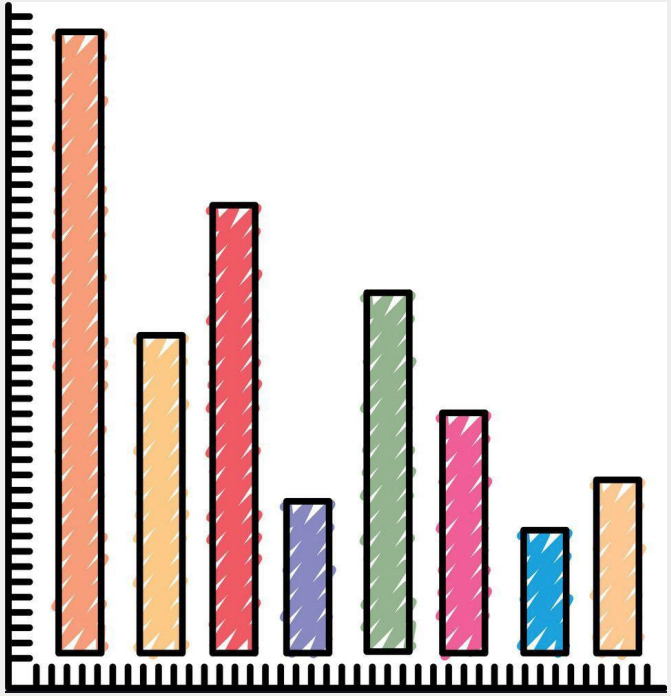

Kaggle Machine Learning
and Data Science Survey

About the Survey

- Kaggle has conducted an industry-wide survey for the years 2018,2019,2020,2021
- The survey presents a comprehensive view of the state of data science and machine learning
- The survey includes raw numbers, regional comparisons and technological aspects required for various job profiles

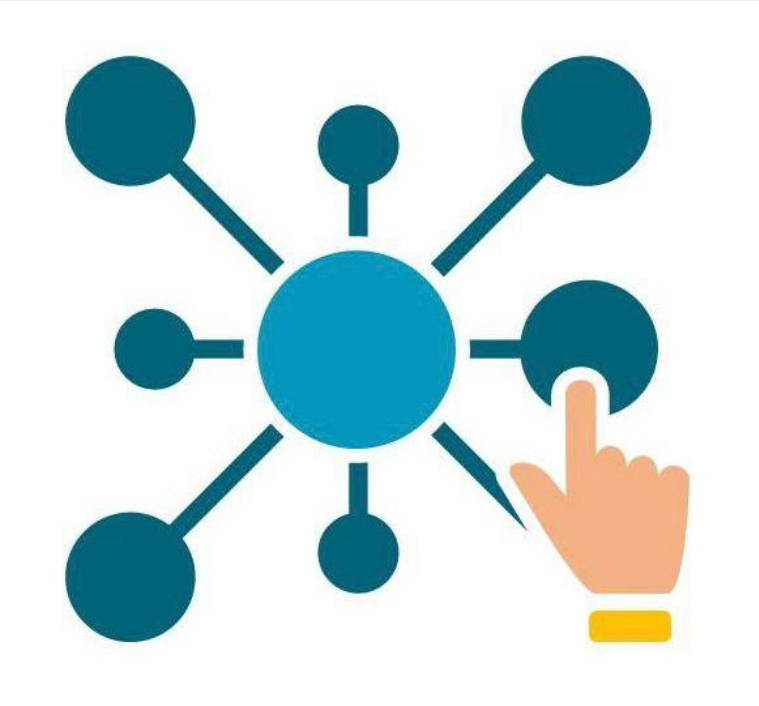
Dataset



- There are four data sets each for year 2021,2019 and 18
- Extracted the rows for the role of data scientist to understand the underlying patterns

Year	Original Rows	After Extracting
2021	25975	3616
2020	20038	2676
2019	19719	4085
2018	23861	4137

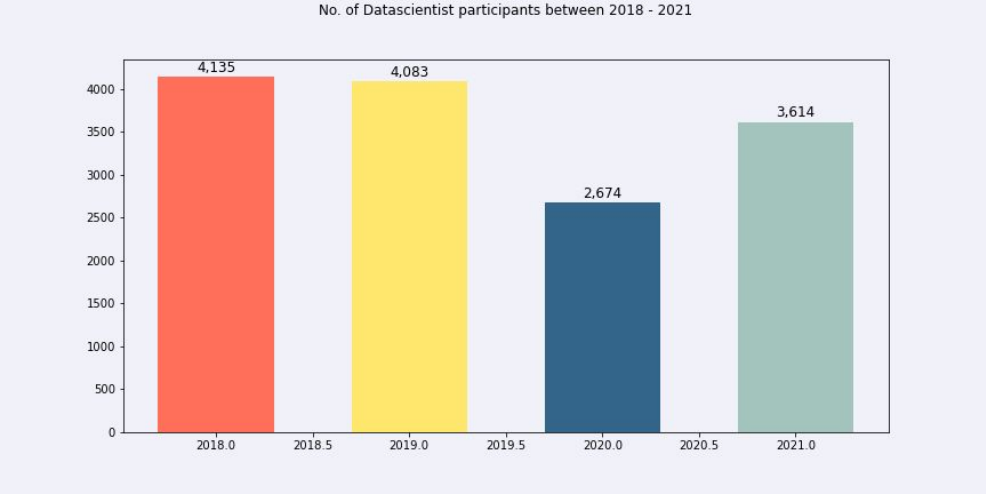
Dataset



Features:

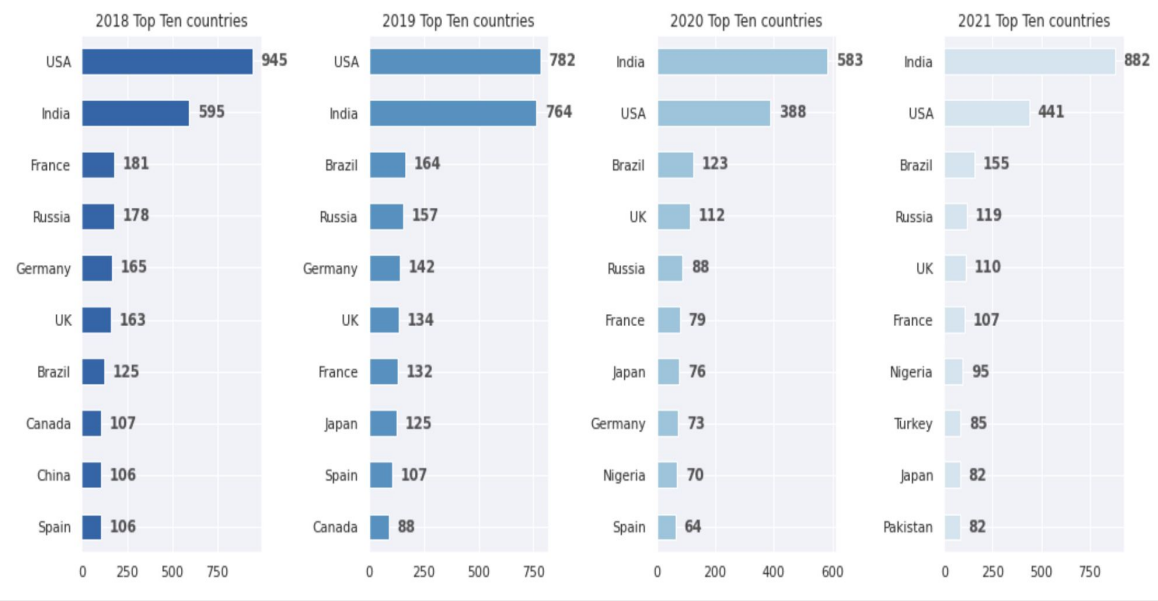
1. Age
2. Gender
3. Country
4. Job Role
5. Salary
6. Coding Experience
7. Programming Languages Used
8. Visualization Libraries Used
9. ML Methods
10. Cloud Platforms Regularly used etc.

Data Scientists Participated in Survey



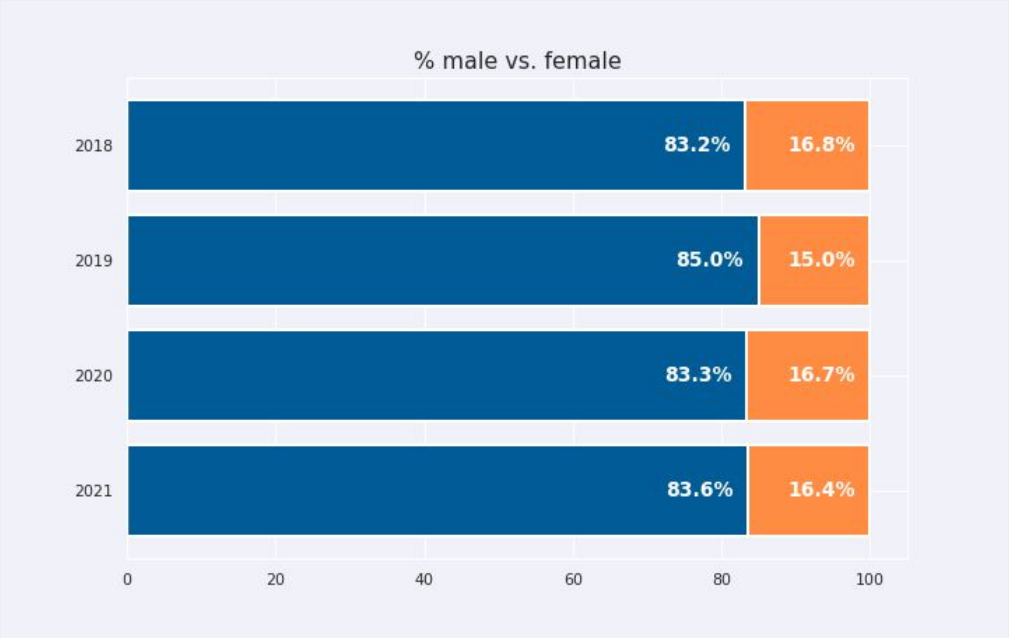
- The survey consistently receives high responses with 2,400 to 4,100 identifying as data scientists out of 17,000 to 24,000 participants yearly.
- Though there was a dip in the responses in the year 2020, the number immediately picked up in the following year

Top 10 Countries that Participated in the Survey



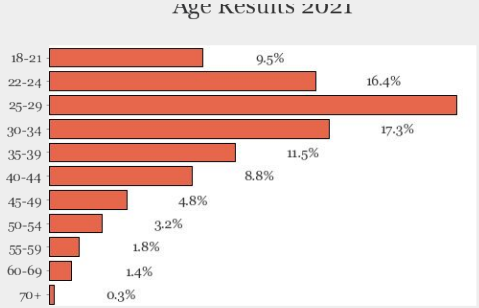
- India continues to top the list for the third year in a row, with 882 participants in 2021. This number has almost doubled from four years ago while the U.S continues to remain in the top 2 of the list, the number of participants has dropped by 40% when compared to the peak in 2019
- Apart from India and U.S, countries that remain at top 10 of the list for the past 5 years include Russia, Brazil, UK and Germany
- African countries were absent from the top 10 list four years ago, but Nigeria broke the pattern by making it into the top 10 in 2020.

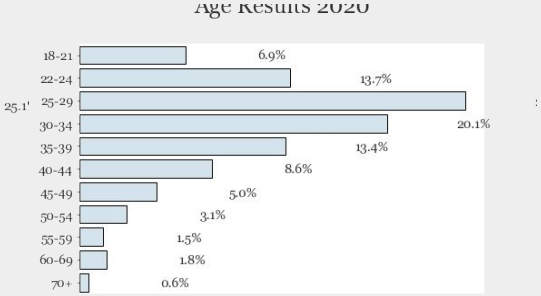
Gender Disparity in Survey Participation

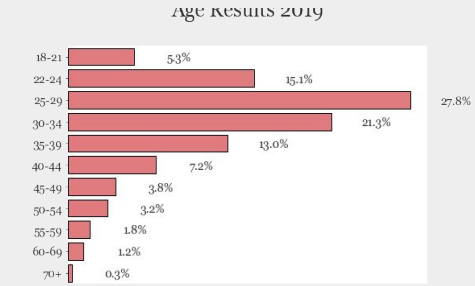


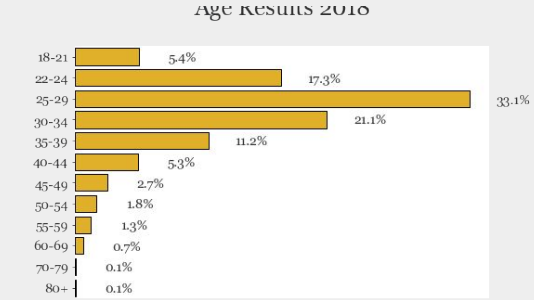
- The number of male and female respondents has definitely increased as the number of people taking part in the survey has increased.
- The rise in the number of male respondents is more pronounced than the rise in female respondents.
- Clearly, males outnumbered females in the survey.

Data Science: A Field for all Ages?



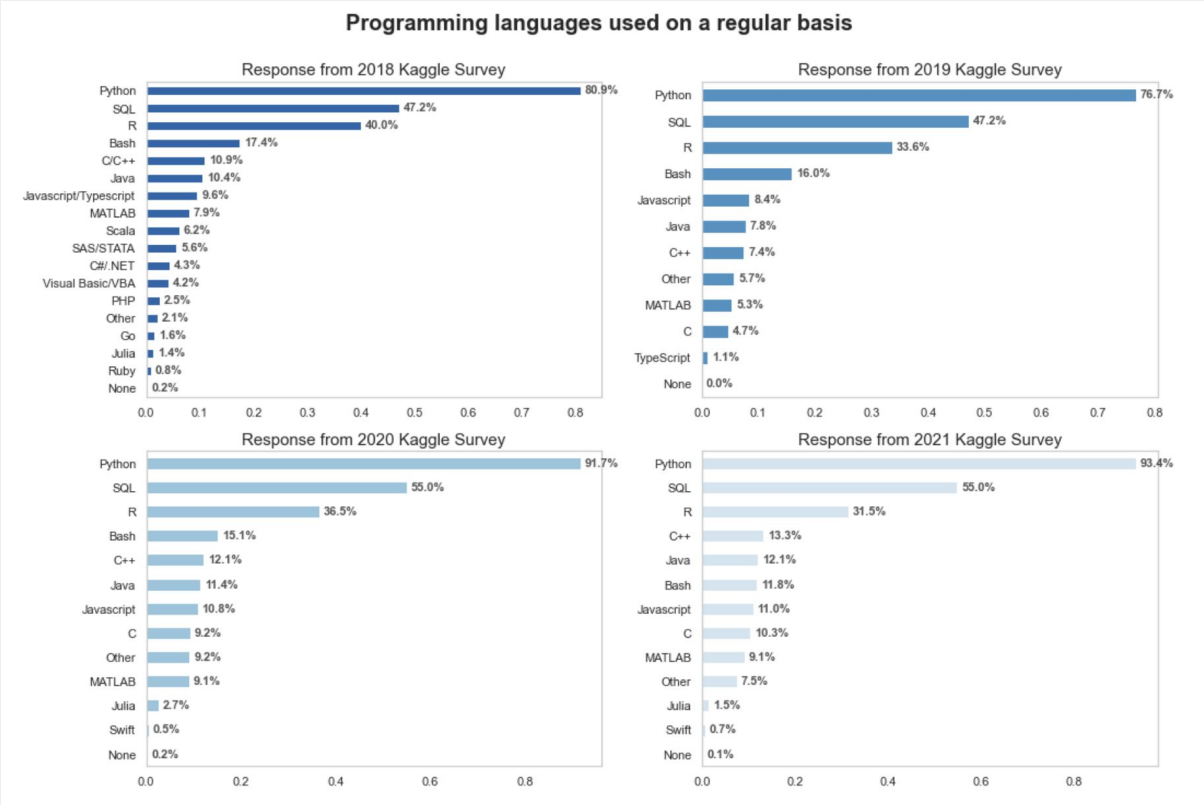






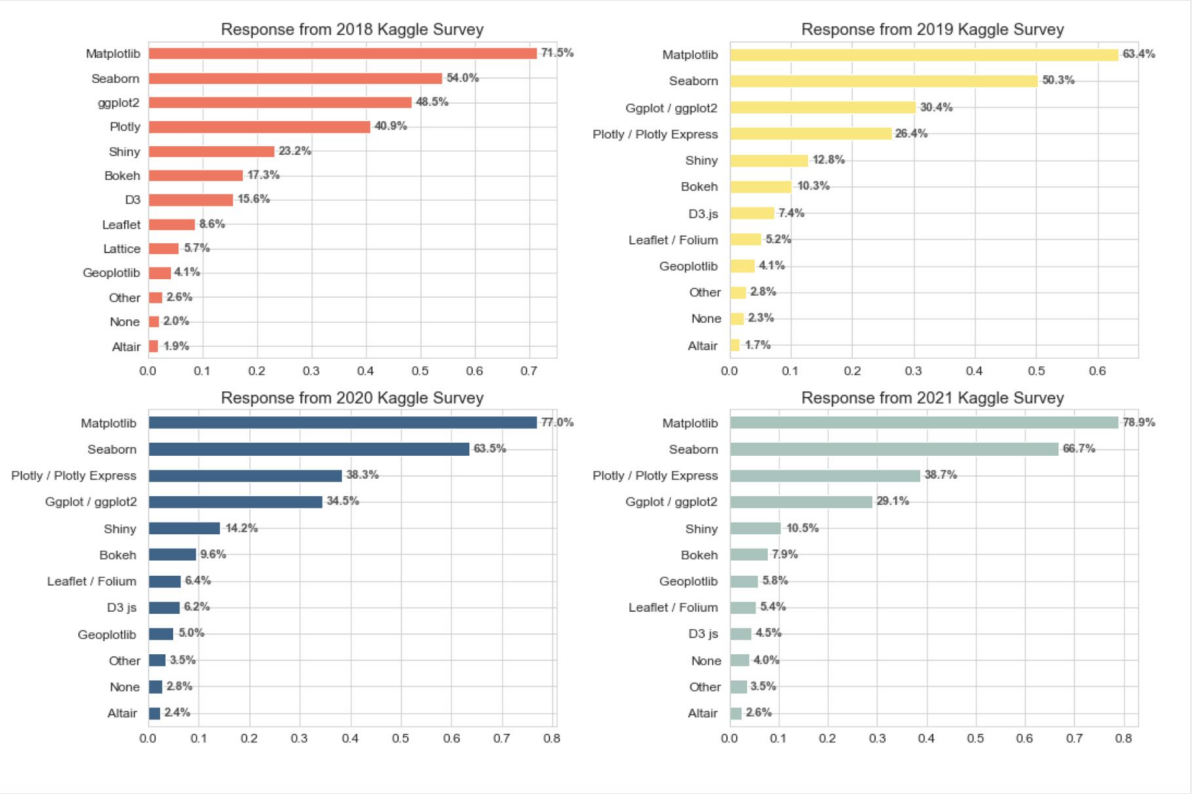
- The rise in data science popularity and Online courses has led to a significant increase in the number of people practicing analytics, especially for those below 21 years old
- The number of responses has remained more or less the same for age bracket 22-24 years
- Looking at the opposite side of the chart, we see either very experienced users or people who became interested in data science later in their careers; believe me, it's never too late! It's nice to see a few hundred of them around

Programming Languages Used by Data Scientists



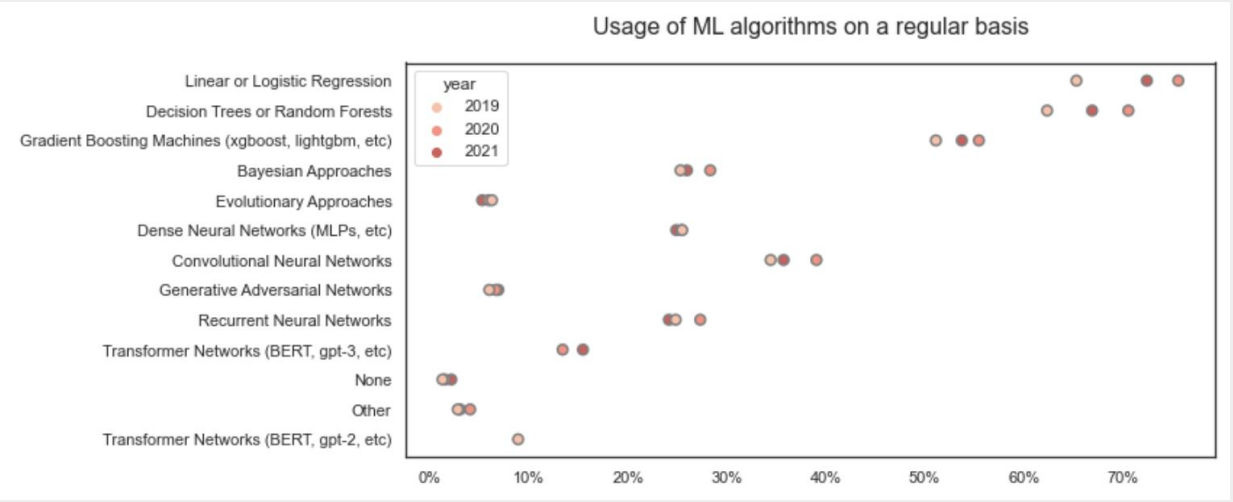
- Python usage at work has significantly increased from 80% to 91% over four years, with over 91% reporting its usage
- The usage of R among data scientists in the industry is rapidly decreasing
- Despite the decline in R usage, the percentage of data scientists suggesting SQL as a first learning priority for aspiring data scientists has slightly increased

Do Data Scientists Prefer Matplotlib and Seaborn?



- Whatever your job or role is, you've probably used Matplotlib once or twice. According to the table above, roughly 60% of the participants visit the library on a regular basis
- Seaborn is the second most popular visualization library
- GGplot and Plotly have remained relatively unchanged over the years

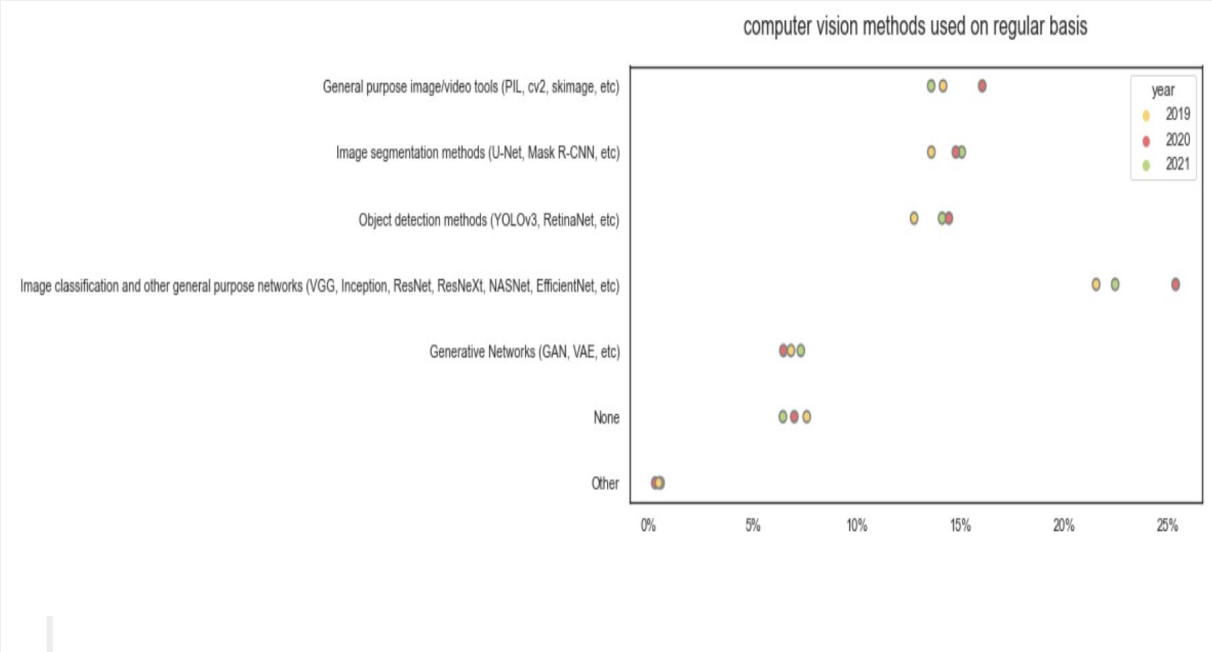
ML Algorithms used by Data Scientists



- Decision trees and random forests are the second most commonly used machine learning algorithms, and their popularity has remained consistent even with the rise of gradient boosting machines like XGBoost
- In recent years, there has been a significant surge in the adoption of specialized neural networks, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), while the utilization of traditional Dense Neural Networks (DNNs) has observed a decline.

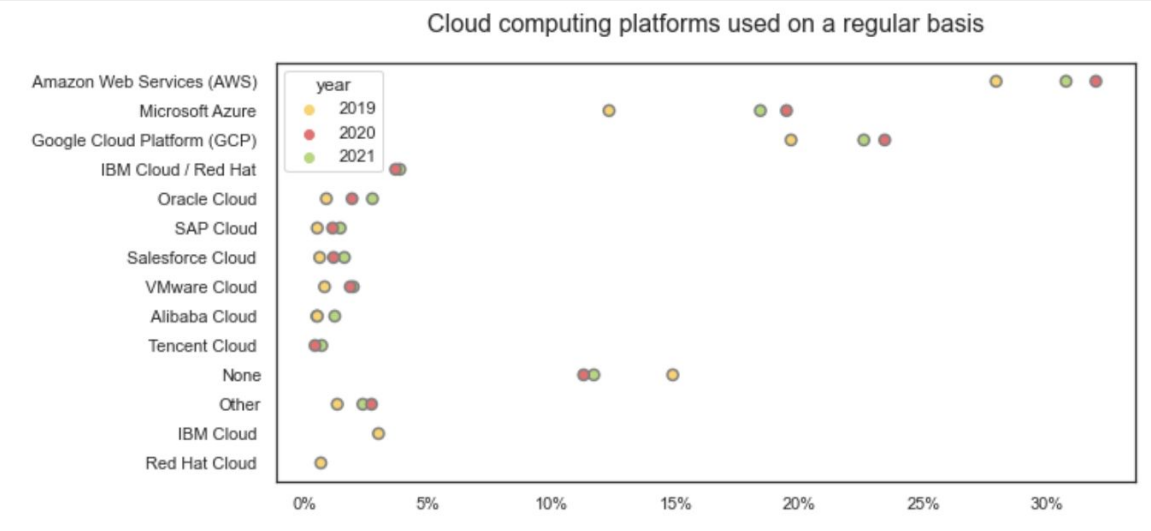
- The usage of recurrent neural networks (RNNs) has experienced a significant rise, even though not as stunning as that of CNNs, owing to their application in processing time series and sequential data such as word sequences.

Stagnation in Computer Vision Development



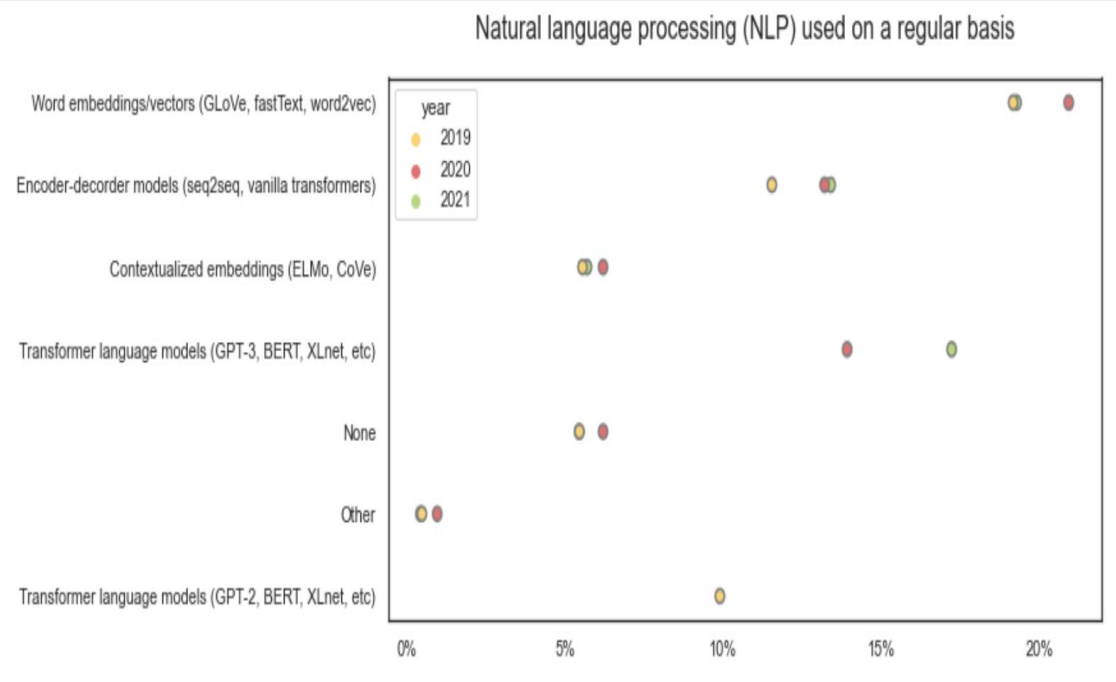
- Computer vision methods have remained relatively constant over the last three years in terms of their development and advancement. There have been no significant changes or breakthroughs in the field during this time
- Usage of image classification with different networks have been increased over time

AWS Dominance Among Top Cloud Service Providers



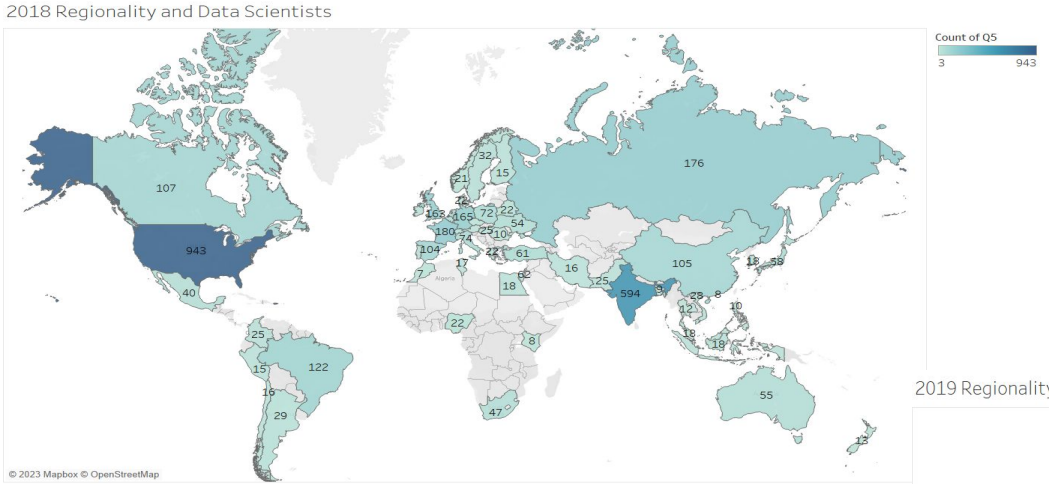
- According to Kaggle Survey, among top cloud leaders, AWS Cloud Platform users increased every year by at least 2.5 - 5 %, while Azure and GCP users increased by at least 3.04% among their niche players
- A small percentage of respondents do not use cloud services at all

Underlying Patterns in NLP



- Word Embeddings emerged as a winner in the year 2021, suddenly out of the blue
- It is noteworthy that all natural language processing models are increasingly being utilized on a frequent basis over the years

Regionality and Data Scientists 2018 and 2019

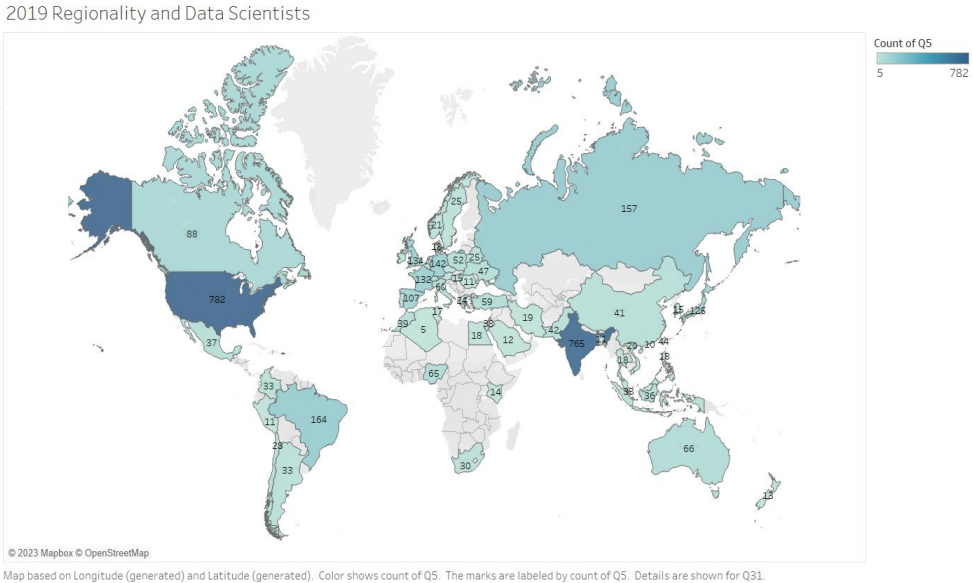


2018:

- Top five countries with the highest number of participants: USA, India, France, Russia, and Germany.
- Implications of the presence of European countries in the top five.
- Emerging interest in data science and machine learning in India.

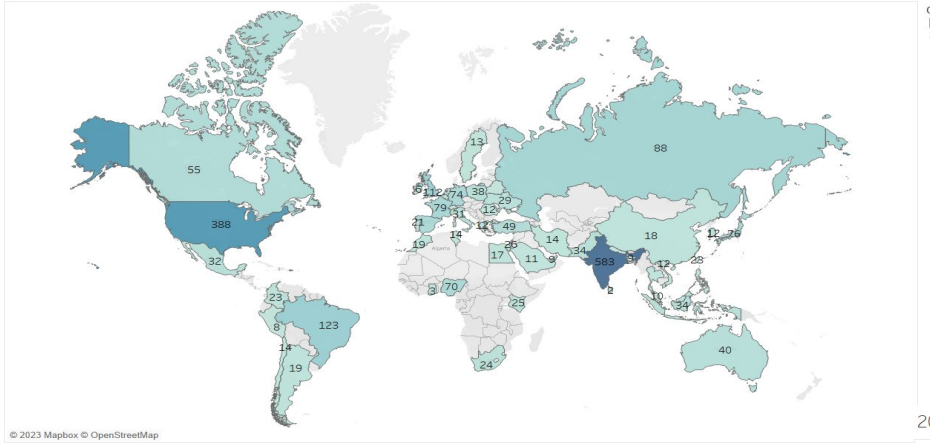
2019:

- Shift in the rankings, with India surpassing the US to become the country with the highest number of participants.
- Emergence of Brazil as a new player in the top five.
- Maintained positions of the Germany and Russia in the top five.
- Decrease in the number of participants from the US.



Regional Differences in 2020 and 2021

2020 Regionality and Data Scientists

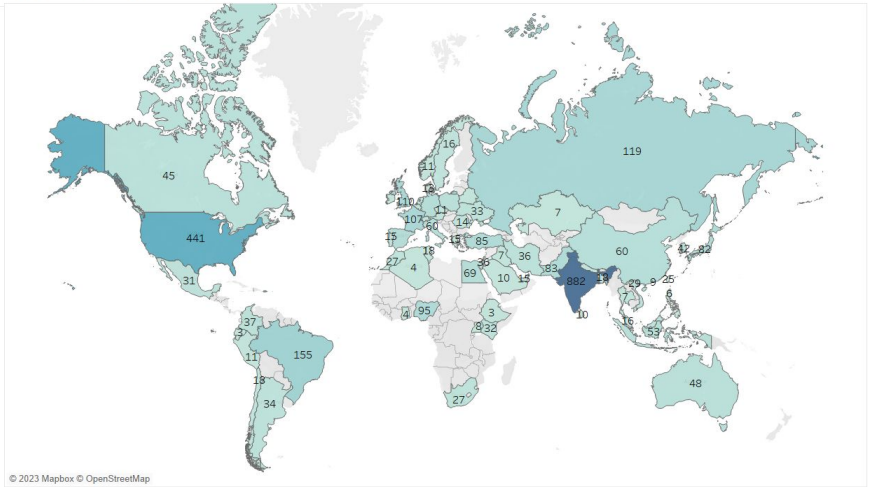


Map based on Longitude (generated) and Latitude (generated). Color shows count of Q5. The marks are labeled by count of Q5. Details are shown for Q3.

2020:

- India had the highest number of participants with 583 respondents, comprising around 20% of total respondents.
- The USA ranked second with 338 respondents, representing around 12% of total respondents.
- Other top countries included Brazil, the UK, and Russia.

2021 Regionality and Data Scientists

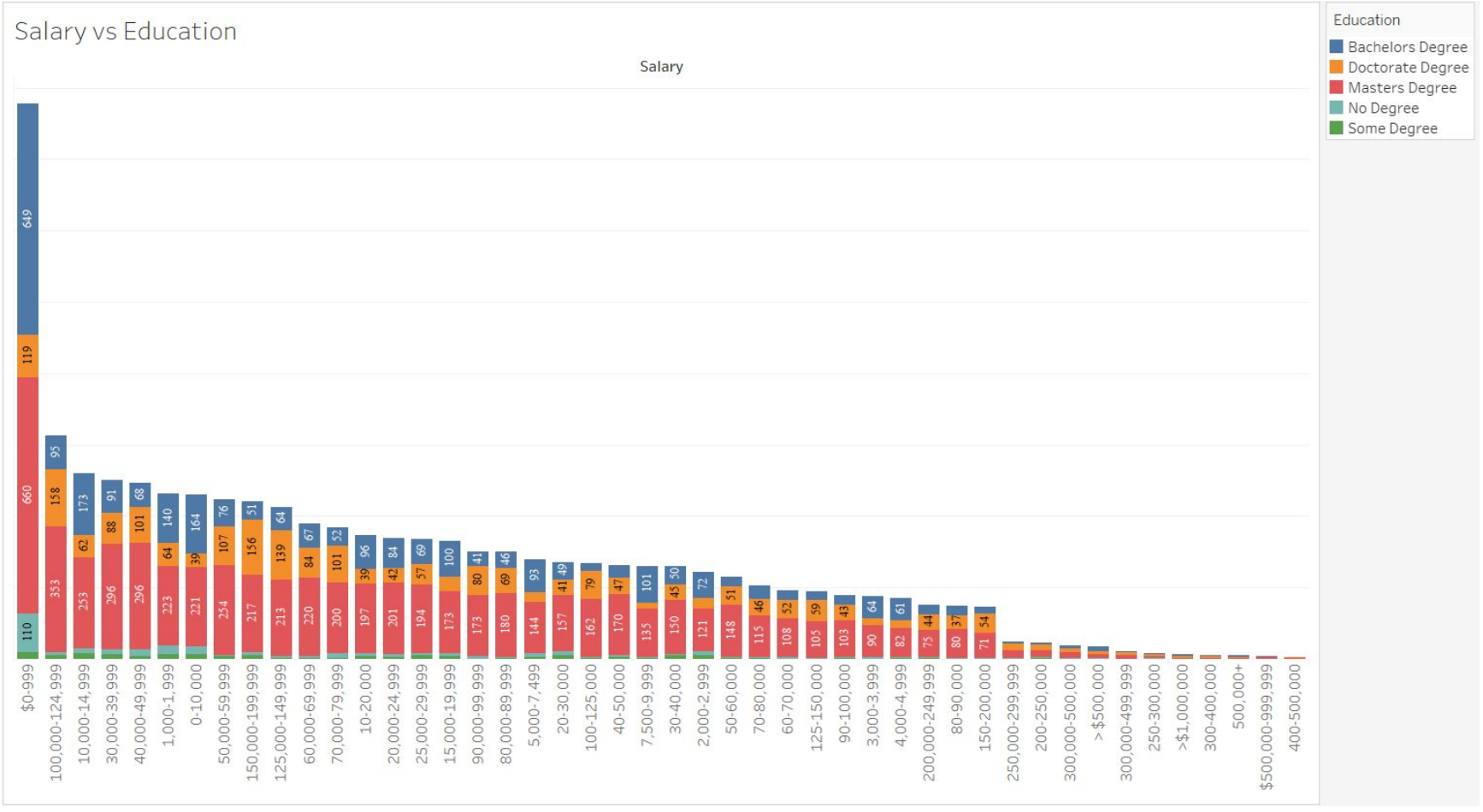


Map based on Longitude (generated) and Latitude (generated). Color shows count of Q5. The marks are labeled by count of Q5. Details are shown for Q3.

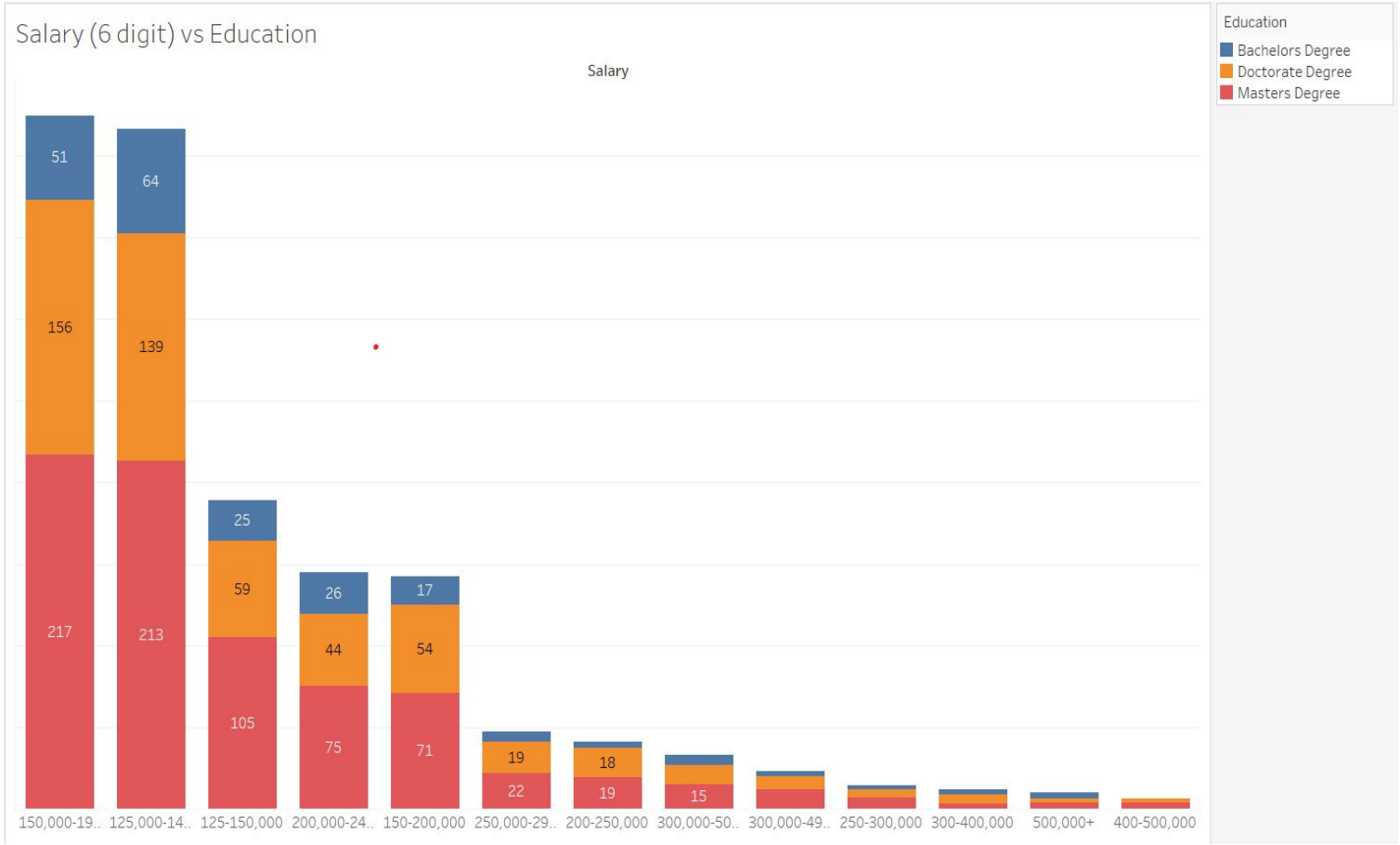
2021:

- India continued to have the highest number of participants with 882 respondents, comprising around 24% of total respondents.
- The USA saw an increase in the number of participants and moved up in the rankings, with 441 respondents representing around 12% of total respondents.
- Other top countries in 2021 included Brazil, Russia, and France.

Education, Salary and Data Scientists

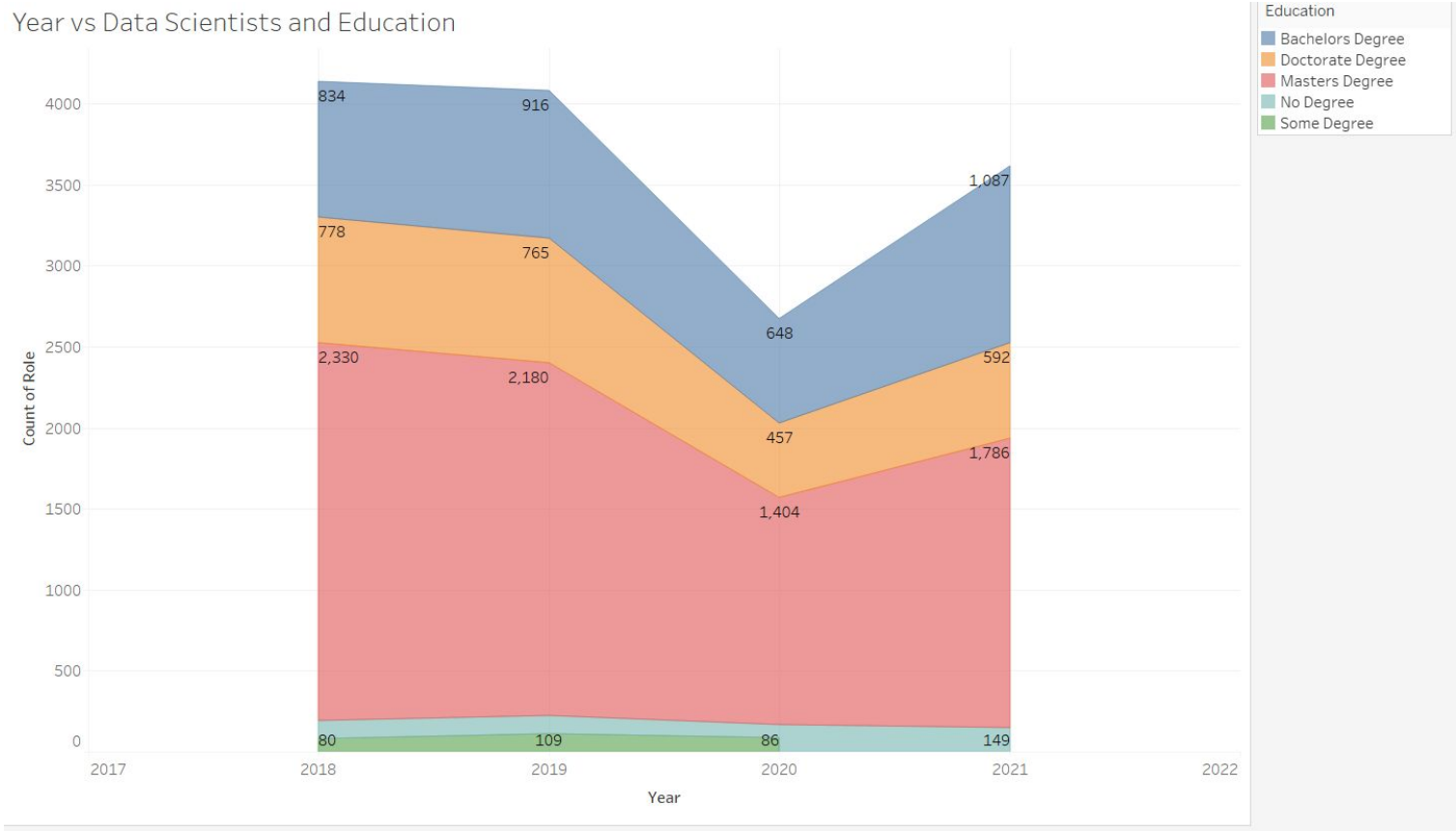


Education, Salary and Data Scientists



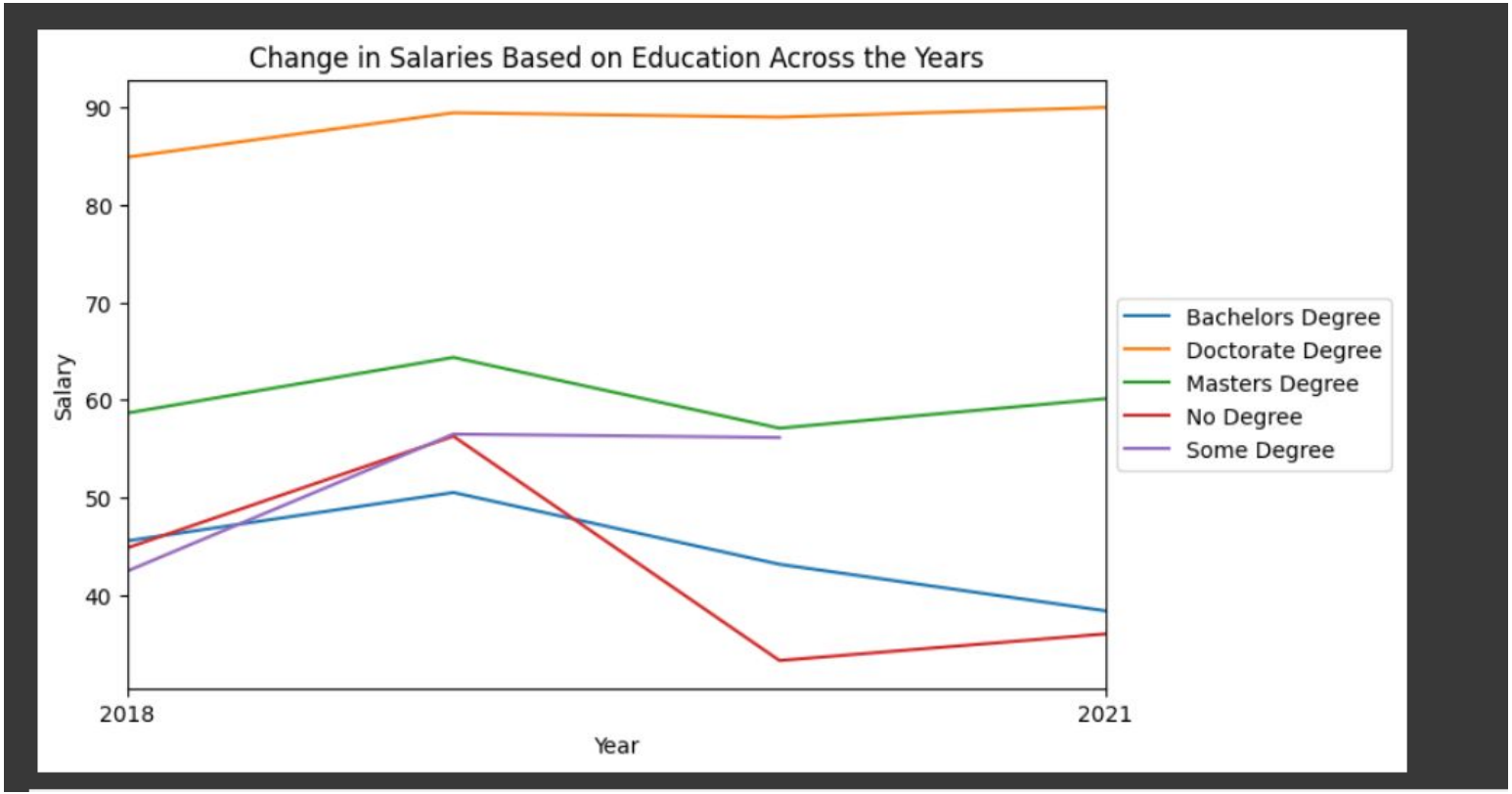
- If someone want to be a data scientist and want to get a 6 digit salary, pursuing masters might be a good option. We can see that most of the 6 digit salaries in Data Science field are from the Masters student graduates.
-

Education, Salary and Data Scientists



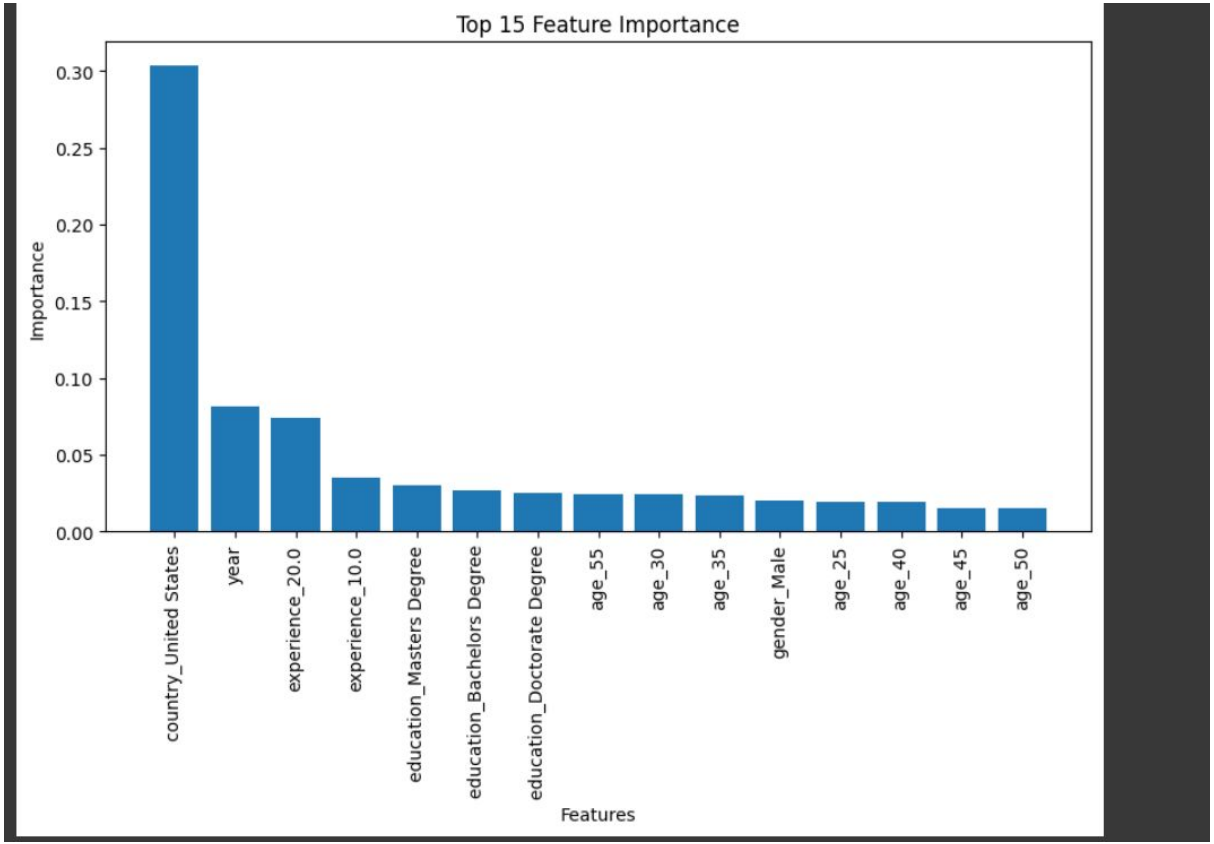
- We can see that there is a decrease in no of data scientists from across different education backgrounds in 2020, might be due to the pandemic.
- Also, there are more data scientists coming from Masters degree.

Education, Salary and Data Scientists



- We can see in this plot that Doctorate graduates salaries across years have remained almost constant and that too at a higher salary.
- Salary for bachelors grad data scientists have been decreasing over the years
- Masters grad data scientist also have some ups and downs but their salary level is in a particular range.
- Earning a salary without a degree has become relatively difficult.

Random Forest - Feature Importance



We have trained a Random Forest model and other models to determine the factors that have a greater impact on salary.

- Now, for certainly we know why there are more data scientists in the USA.
- More experience means more Salary.

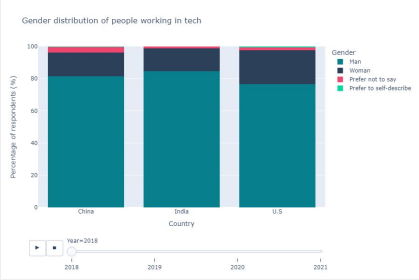
The Story of India and China-At a Glance

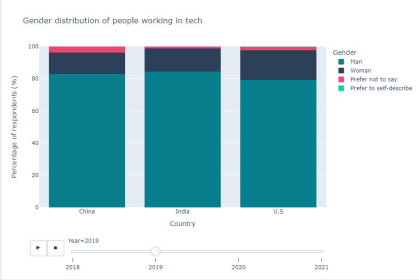
Motivation-Interesting to see how two manufacturing intensive nations moved towards technological revolution and how well or not are they performing in the recent years

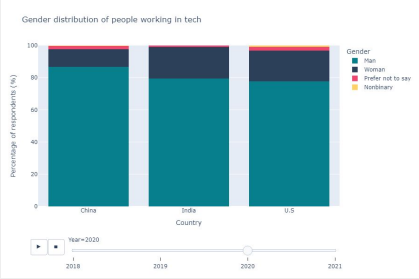
Factors taken into consideration-

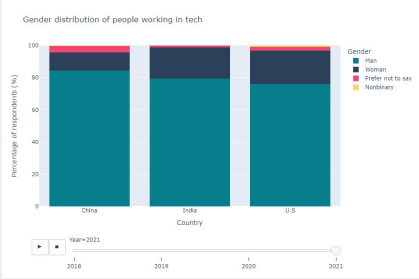
1. Gender
2. Age
3. Salary
4. Level of Education
5. Machine Learning usage
6. Popular Programming Languages

Gender Distribution in Technology Industry



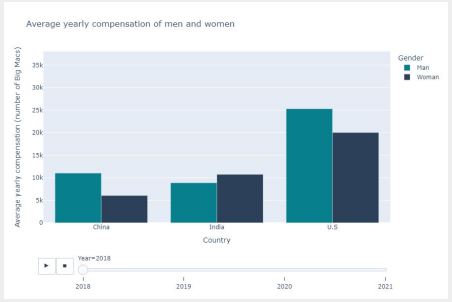


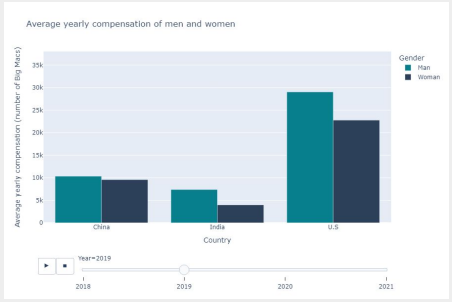


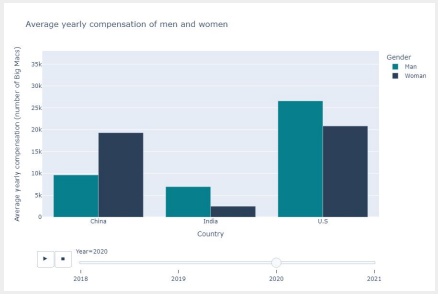


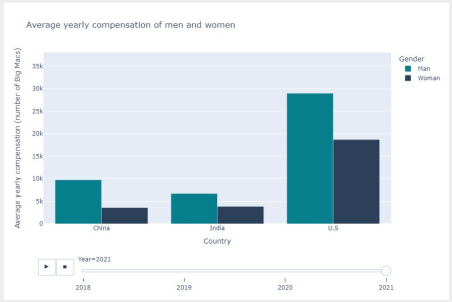
- Proportion of females working in technology industry increased over the years in both India and US
- Reason- Covid(Work from Home)
- In China the figure of proportion of females decreased slightly in 2020 and then increased by 2% in 2021
- Reason-Further analysis required
- **Of the three countries, China has higher gender imbalance in the Technology Industry**

Gender Pay Gap



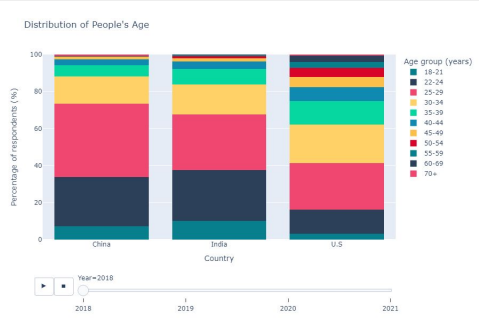


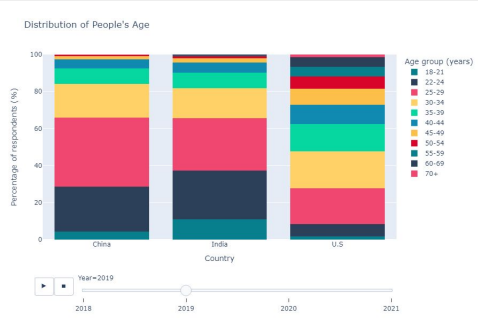


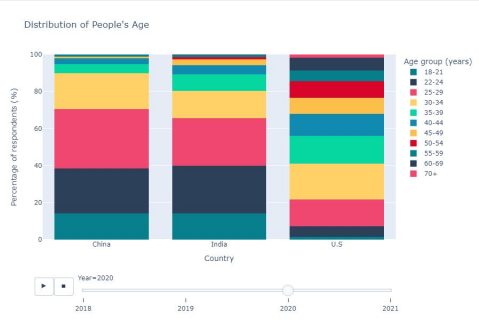


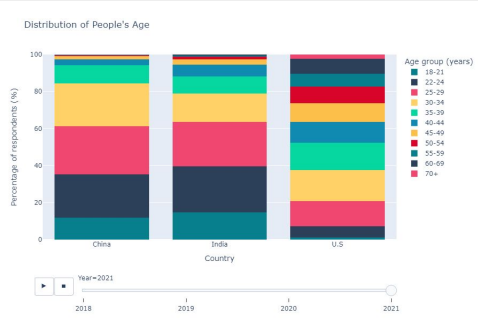
- India has smallest pay gap
- Interesting to see women got paid significantly higher than men in the year 2020 in China-50% more
- In the next year the compensation of women in China fell flat
- U.S continued to have increasing trend of gender pay gap over the years-highest in 2021 where men got paid 5% more than women

Distribution of the Age of the Working Population



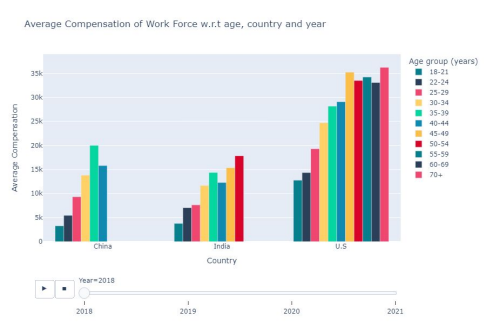


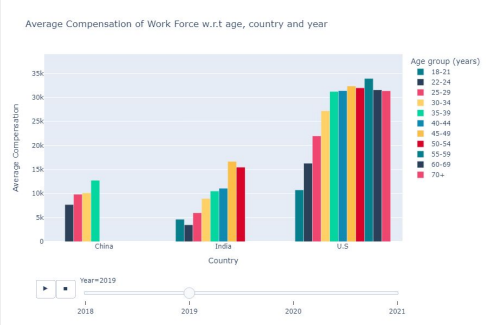


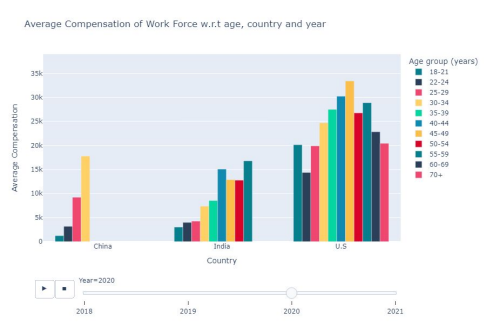


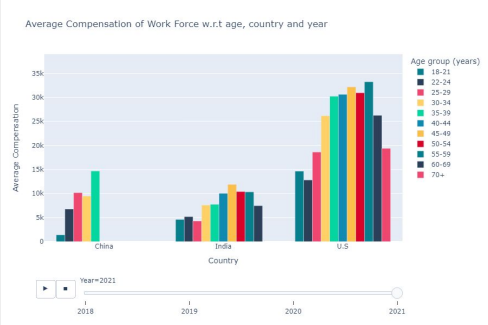
- The proportion of young population(18-23 years) shooted up in 2020 probably due to innumerable opportunities that the period has presented
- In all the four years, the workers till the age of 44 years make up more than three fourths of the population
- In the U.S, we see a balanced distribution of work force in the U.S

Average Compensation of Workforce w.r.t Age, Country & Year



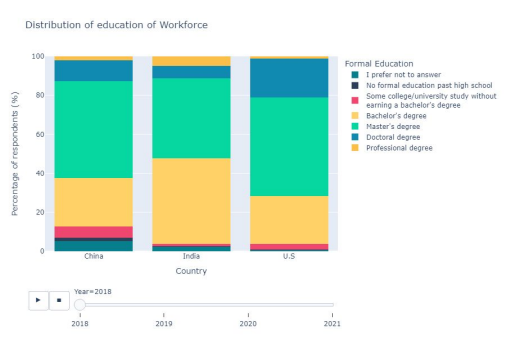


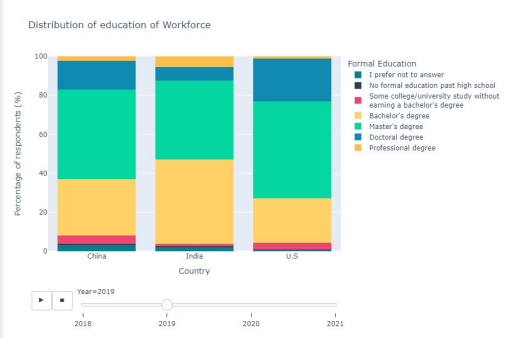


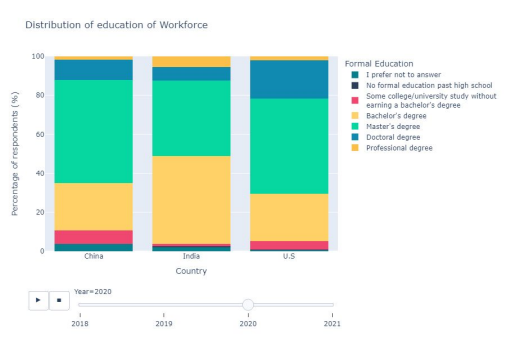


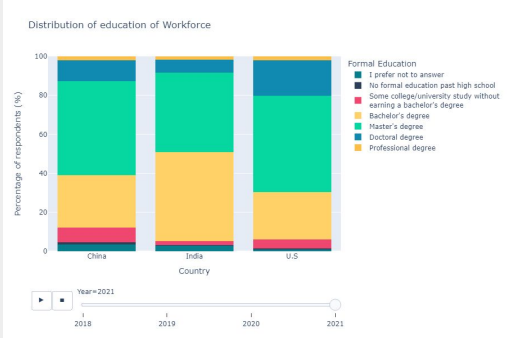
- In India, from 2018-2021 we see that compensation is moving towards the scenario of evening out(leveraging demographic dividend)
- In China, the absence of older age groups in the technology industry shows that it is a relatively new industry
- The U.S showed similar results in the four years, where middle aged groups earned relatively more(Exception:2020, where younger population also got higher salaries-Covid)

The Level of Education of the Workforce



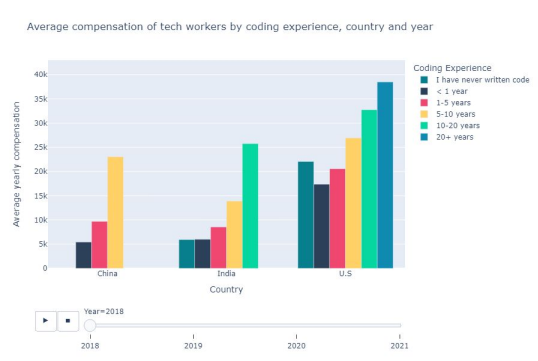


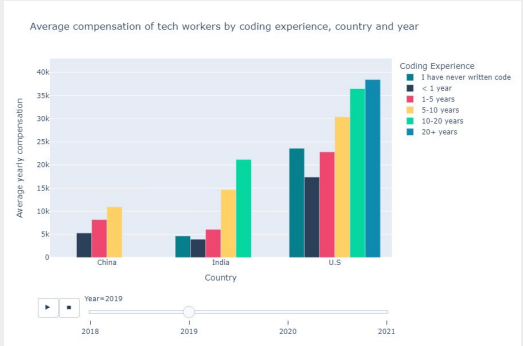


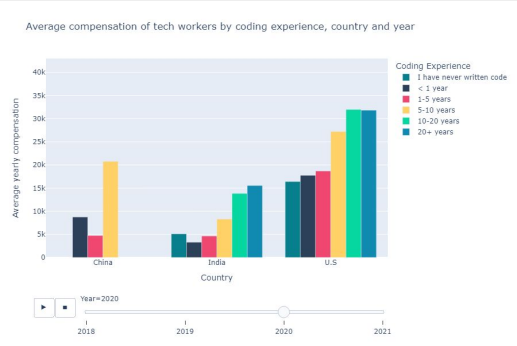


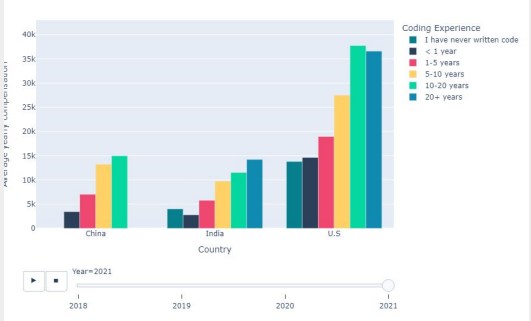
- The percentage of people pursuing higher levels of education has been decreasing in all the three countries since 2019
- In the same direction, proportion of people with no formal degree also increased in all the three countries since 2018
- Possible case of a certain section of people becoming ambitious and a certain section in a scenario of not being able to complete their degree

Average Compensation w.r.t Coding Experience



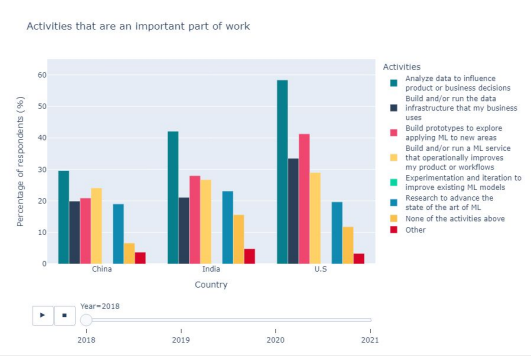


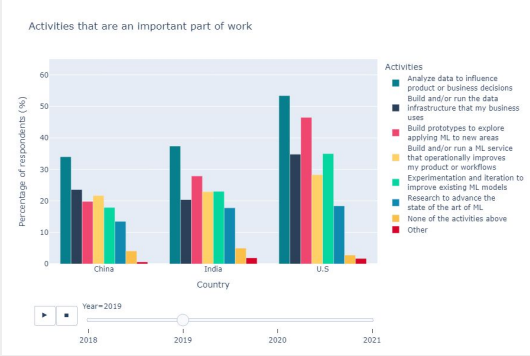


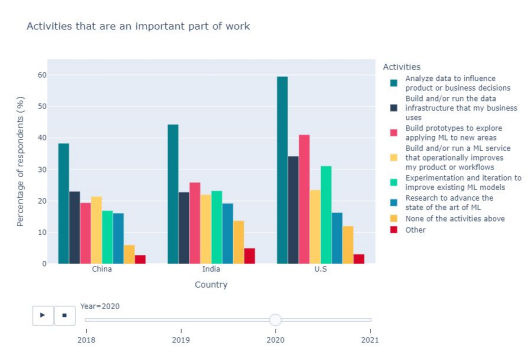


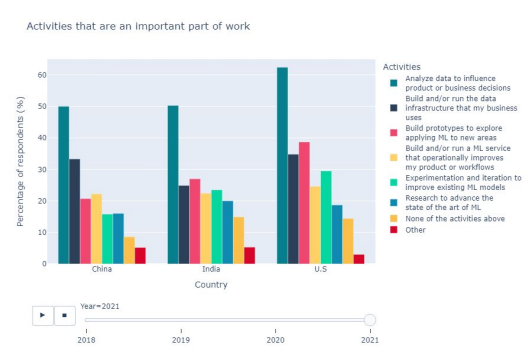
- Usually Coding experience leads to higher salaries
- Anomalies: China(2020), India(2019-21), U.S(2019)
- China has no workforce that had experience of writing code past 10 years(Technology Service Sector is relatively in a nascent yet blooming state

Is Machine Learning Important at Work?



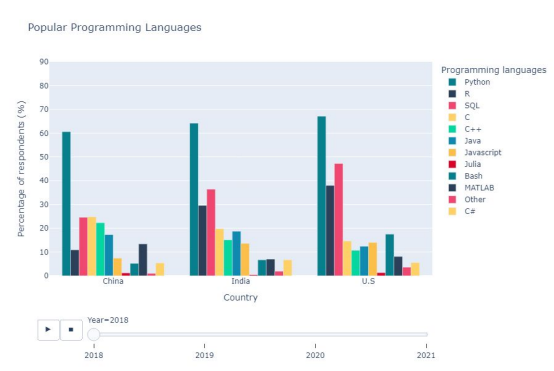


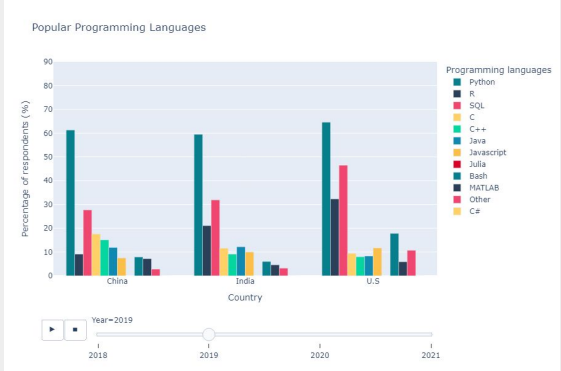


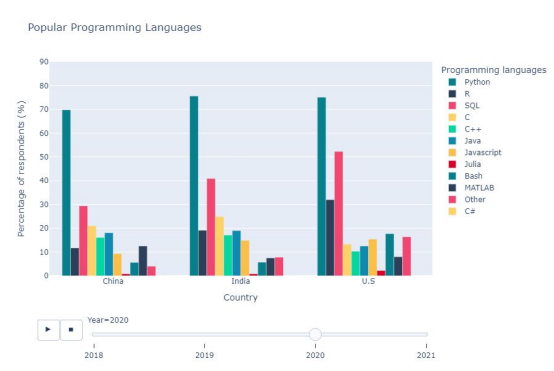


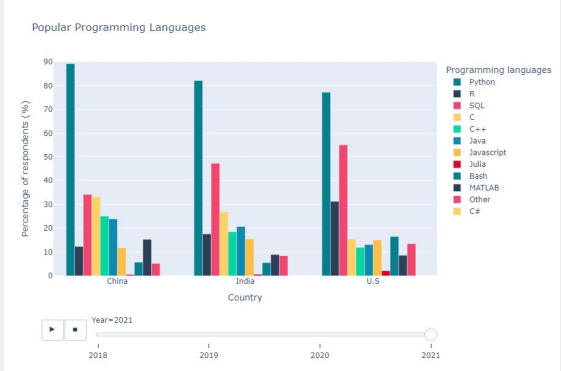
- Proportion of people using Machine Learning at work increasing in India and China(expansion of highly skilled workforce) since 2020
- The opposite is observed in the tech market of U.S since 2018
- Analyzing data to influence day-to-day business is an essential task at work in all the countries

Popular Programming Languages









- Popularity of Python and SQL skyrocketed in all the three countries since 2018(tech boom-covid)
- Increase in usage of C and C++ suggests an inclination to hardware industry in India(to become self-reliant) and China(leading producer)

Observations

- China has higher gender imbalance in the Technology Industry
- Gender pay gap in India is the smallest
- India and China are on the right track of leveraging demographic dividend
- Though specializations aren’t required, formal education does make difference in both India and China
- Coding Experience drives salaries in India and China(not in U.S)
- Analyzing Data >>>Building ML Models in both the countries

THANK YOU

Suhas Aitham

SaiRam Gajavalli

Spurthy Vupputuri

Sreevishakkanth

Kiran Kumar Mudududla Ravindernath