WIE3010 – Data Visualization

Individual Assignment

Kaggle DS & ML Survey: Community, Skills and Industry

Amir Firdaus Bin Abdul Hadi

17204620/2

Dr. Riyaz Ahamed

Table of Contents

1. Introduction

Every data holds a story and every story comes from the people and the community. And they all can have different interpretation and ways of storytelling. This Kaggle survey provides a dataset that holds the story for data science (DS) and machine learning (ML) community and its up to us on how we want to visualize it. This dataset consists of 43 multiple answer questions with 23997 respondents and I will divide it into 3 segments; demographics, learning medium and industry to try turn it into an easy-to-follow story.

The story starts with knowing our community; who are in it and where they are coming from. Things like gender, age group, country and students can be used to understand them better whether it have some roles in shaping this community. We can also relate them to the current issues in tech field. Gender issues such as looking into women in tech and how they are doing in this male-dominated environment. Talent issues such as how the younger people perceive this field and do most countries are able to produce more students in data science.

Then, we journey into the learning route most people ventured upon that might help future data scientist. With access to the Internet, there are a lot of platforms nowadays we can choose from to learn instead of just mainstream university courses. Debating online is enough or formal education is still required. Choosing platform is one thing, but choosing what to learn is another story; what language to learn and what IDEs to use. Moreover, you might doubt whether you have what it takes to break into this field, whether if I am too young or too old for this. Then, keep on reading to find out.

Last but not least, getting into the industry is the end destination for most students and to get paid big money. Hence, knowing what the industry wants surely helps you in achieving that goal. To find out which sector you find interesting, what job title you are aiming, is the salary good enough and the tools they used in making all those things work. This report will provide all those visualizations and you can interpret as you like it.

2. Visualization

2.1 Knowing the demographics

2.1.1 Gender: Is it still a male-dominated community?

The term “women in tech” have been trending and focused on in this day and age. As the years goes by, women have been encouraged to pursue technology as a career. But today, there’s still a wide gap between men and women in tech (Statista, 2021). According to Statista, even though women make up about 47% of the U.S workforce, only 26.7% have jobs in the tech industry. Do that numbers also reflect in our community?

Well, based on figure 1, only 22% of our respondents are female while male is dominating by representing 76% of the community. Problems such as gender inequality, discrimination, or sexual harassment in male-dominated environments are most likely contributing to this data. This shows that more efforts need to be taken together in making sure that more female are more interested and encouraged to be a part of this thriving and innovating field as different demographics bring different ideas, thinking and approach on how we pioneer technology.

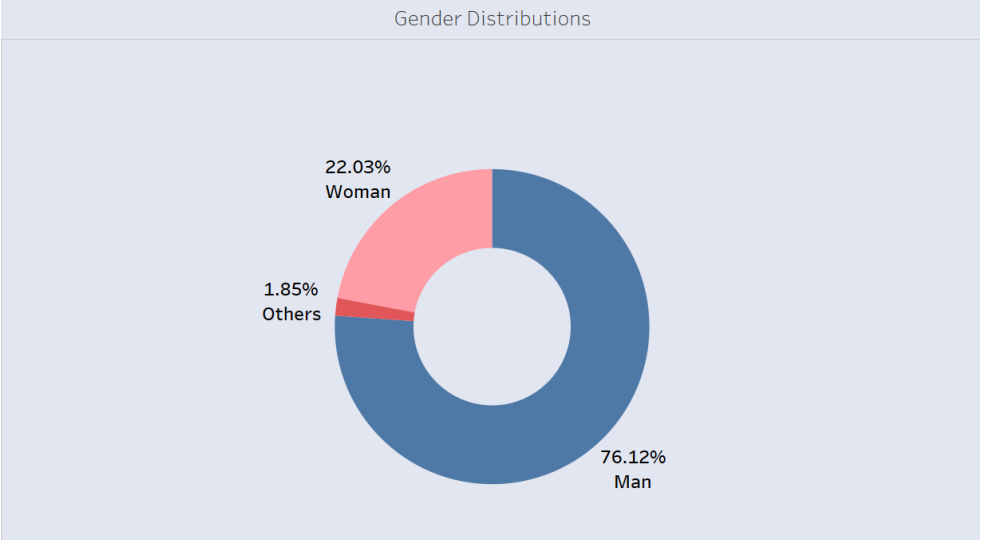


Figure 1: Gender Distributions

2.1.2 Age: Are the youth finds Data Science and Machine Learning worth as a career?

With the evolution of technology going at fast pace, the youth should be more adaptable and easier to learn new things as they developing their skills and tech-savvy compared to the generation above them. This is even more true with data science and machine learning as the new things in tech.

According to figure 2 below, people in range of 18 to 21 years old comprised the most in the community. More generally, people under 30 years old who makes up the young part of this community that represents a lot of us. This serves us a bright future to create more professionals into this community. After the age 30, the numbers went down as the age increases showing fewer old people trying to get this new community. To note the gender aspect as we said before, male is still dominated the charts in all age groups by 72% to 89%. All in all, younger people do find this field interesting and worthwhile to pursue.

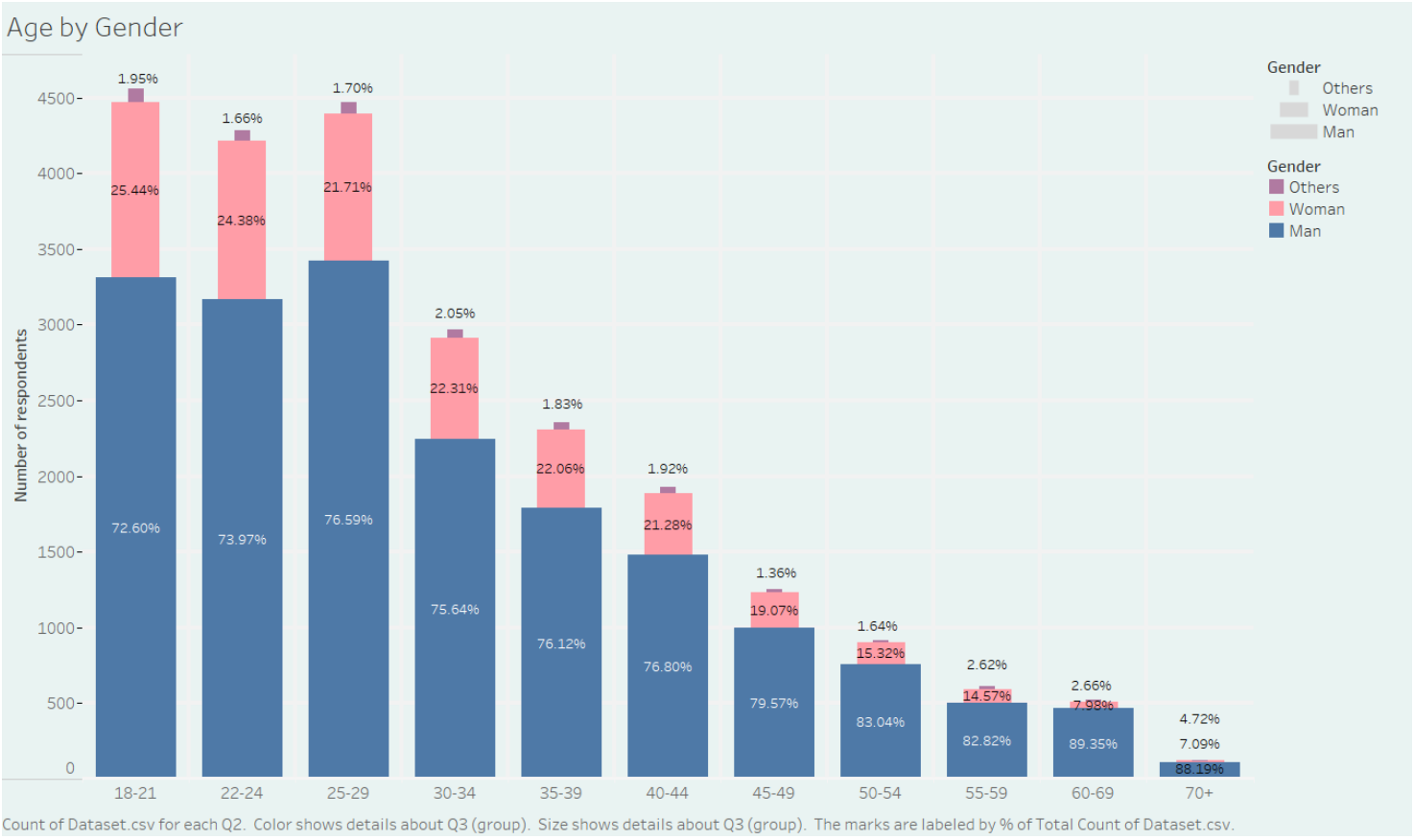
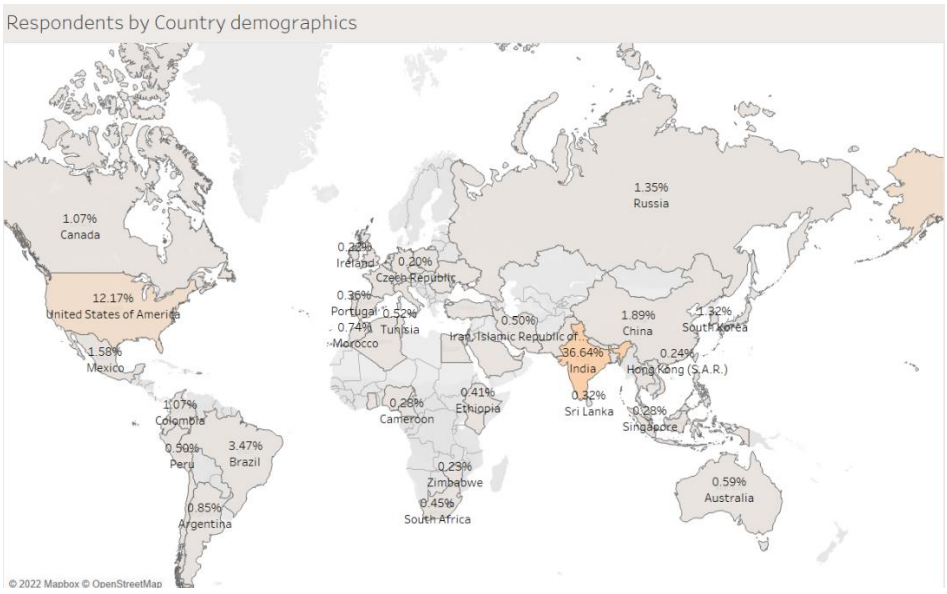


Figure 2: Age distributions by gender

2.1.3 Country: Where are the talents coming from?

Finding out where the talents mostly come from is a valuable information as it can provides the baseline on how they manage to provide the resources to develop them. From the heat map below, India comprises 36.6% of the survey followed by USA with 12.2% which is a big gap between the top and the runner-up. The education system, STEM and social acceptability and high demand and opportunities (Krishnan, 2019) in the country might be the big reasons on why they are so far ahead in producing talents. Figure 4 also display the same country which are India (41.5%) and USA (7.7%) as the main producer of young talents into Data Science community.



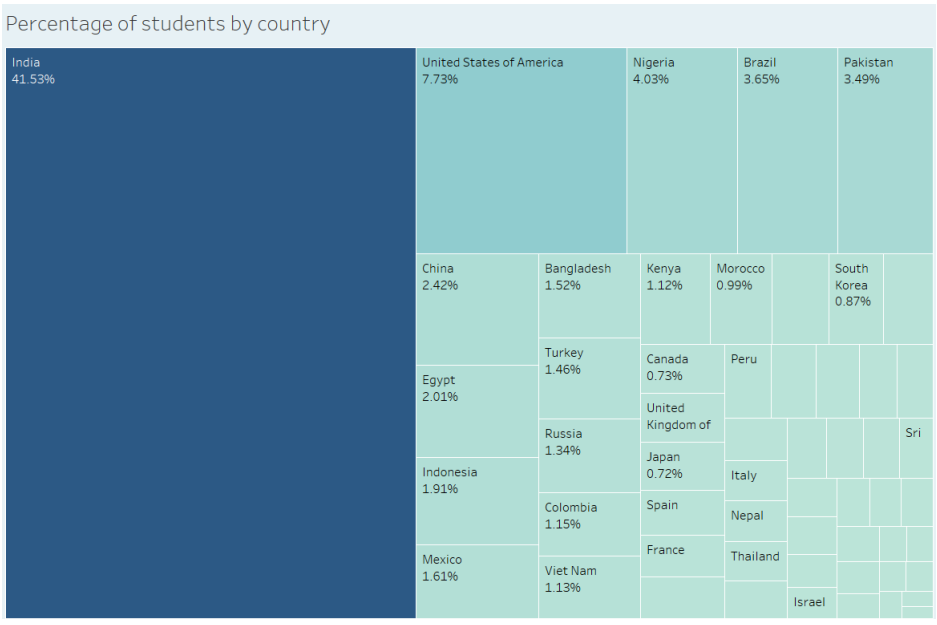


Figure 3 and 4: Country and students demographics

2.2 Knowing the Learning Medium

2.2.1 Platform: Online Education or Formal Education

Talking about professionals in general, most of them requires higher level of formal education which is through university. While this is true for most cases, it is different for technology domain. Why? The simple answer is this; organizations are only interested in what candidates have to offer, degree or not. The skills are what matters the most, wherever you gain them. This is also due to tech talent shortage and industry-wide demand for them making organisation to rethink their priorities (Mersudin, 2022).

As you can see in figure 5, online platform provides a lot of medium to learn your skills. Online courses such as Coursera and EdX takes the cake for the most used platform for learning beating university course. This are able to happen because of free and accessible education which are out of reach for students. They are able to prepare for their career without needing to go for the expensive route.

In another note, they even found YouTube and Kaggle more helpful than syllabus-coded learning in university which are both surprising and not surprising at the same time. This speaks that old-fashioned way of teaching by lecturers should be re-evaluate and changes to make learning more interesting and helpful to students.

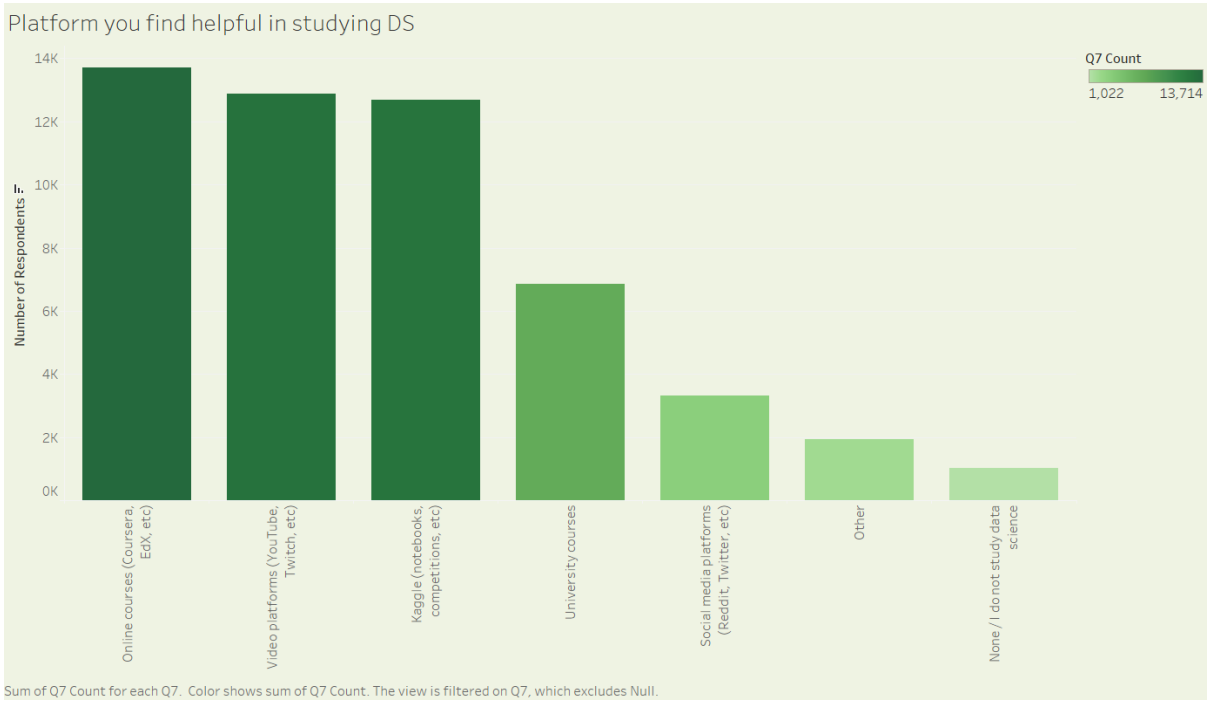


Figure 5: Learning platform

2.2.2 Programming Language: What are people learning and using currently?

The basic foundation of the technology domain is programming language. From figure 6, for data science and machine learning, the most used language is Python by a large margin followed by SQL and R. While from figure 7, the most used IDE is Jupyter Notebook followed by VSCode and PyCharm. From the charts, we can expect which ones are in demands and sought-after skills.

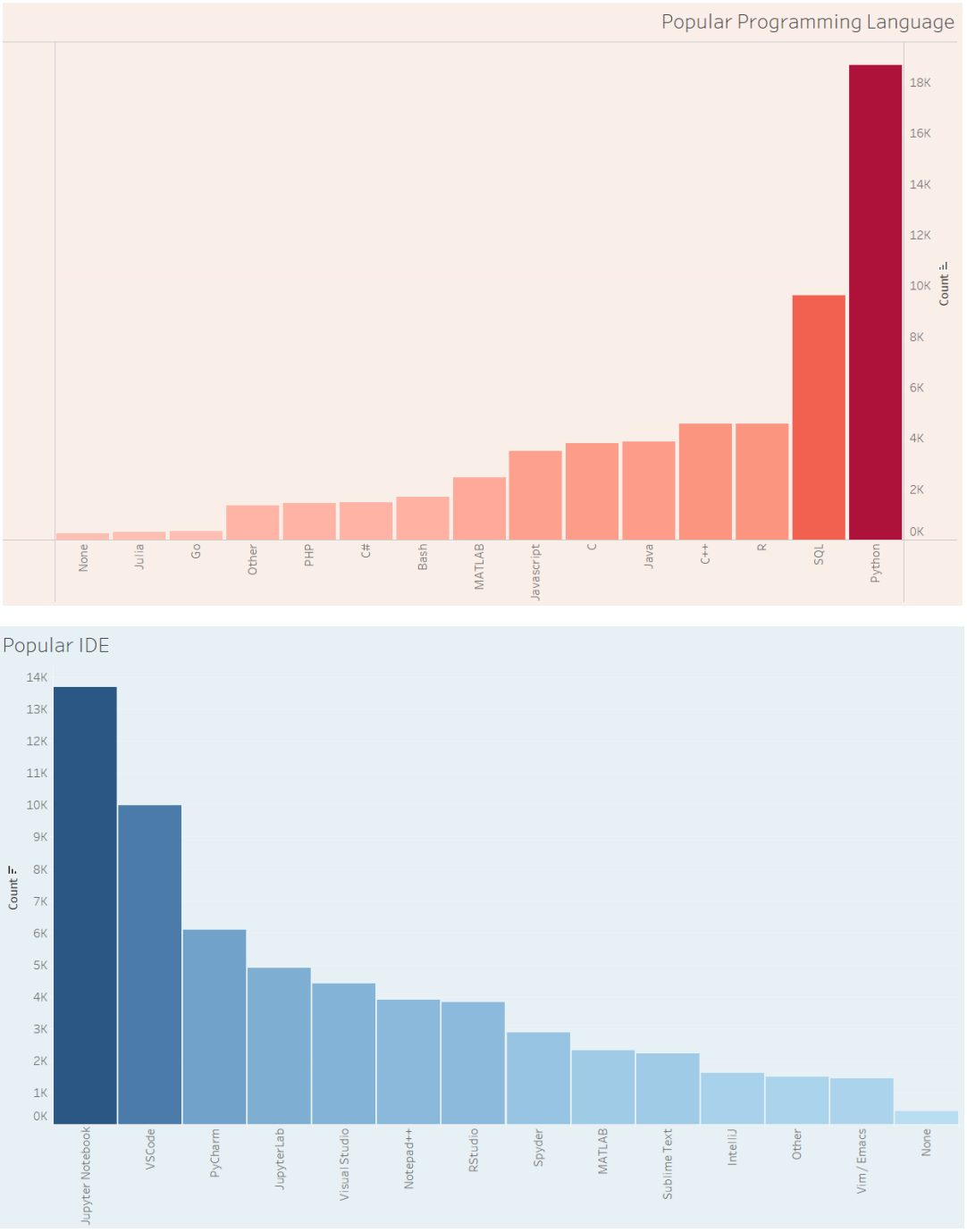


Figure 6 and 7: Popular programming language and IDE

2.2.3 Years of Programming and Age: Young does not mean inexperienced, old does not mean veteran

"Do not judge the book by its cover", there are older people are just trying to learn new skills and there are experienced younger people that started way earlier than the rest. Based on figure 8, we can see that most of the people with less than 5 years of experience is young people and there are none of them in 20+ years as they might have not born or learn to run properly yet. As for old, they are the highest at 20+ years showing there are veterans in this field which are probably consists of higher positions in organizations. Moreover, there are old people in all years of experience category which emphasizes that you are never too old to learn programming.

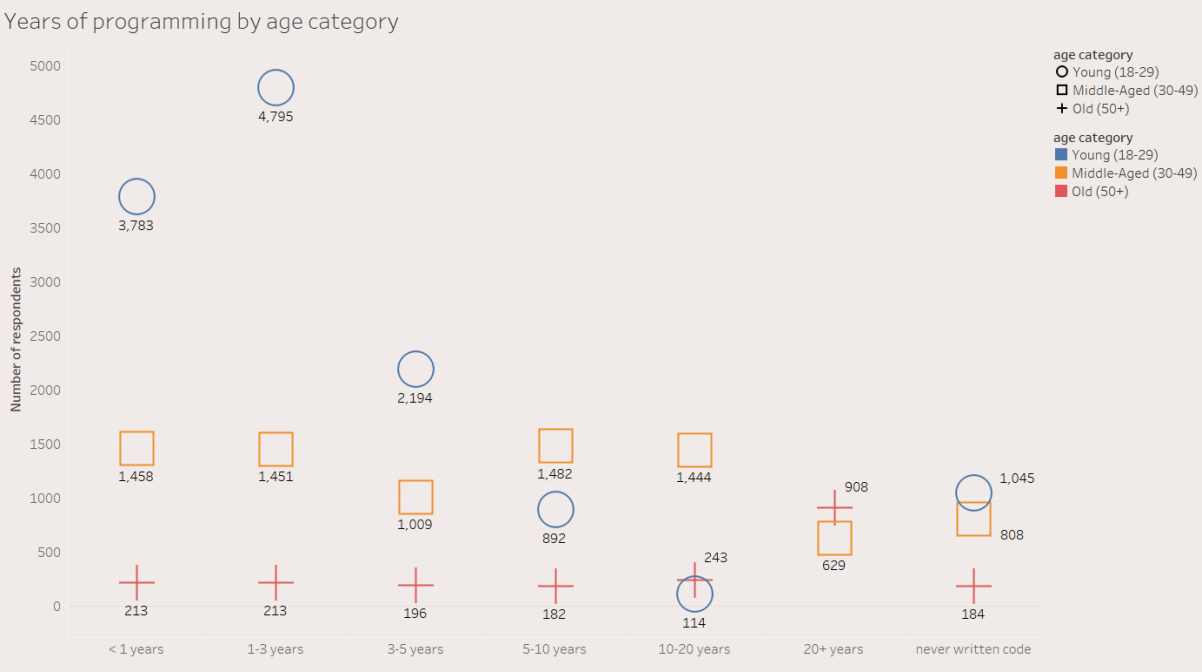


Figure 8: Years of programming by age

2.2.4 Years of Programming and Gender: Are they equally experienced?

As we have declared earlier, there are more males' than females' talent but that does not mean that they are not equally experienced. From figure 9, we can see that there is every gender in each year of programming categories meaning that we have experienced talents from every gender, in the following order; male, female, others. As long you have the skills, no matter what gender you are, you are needed by companies.

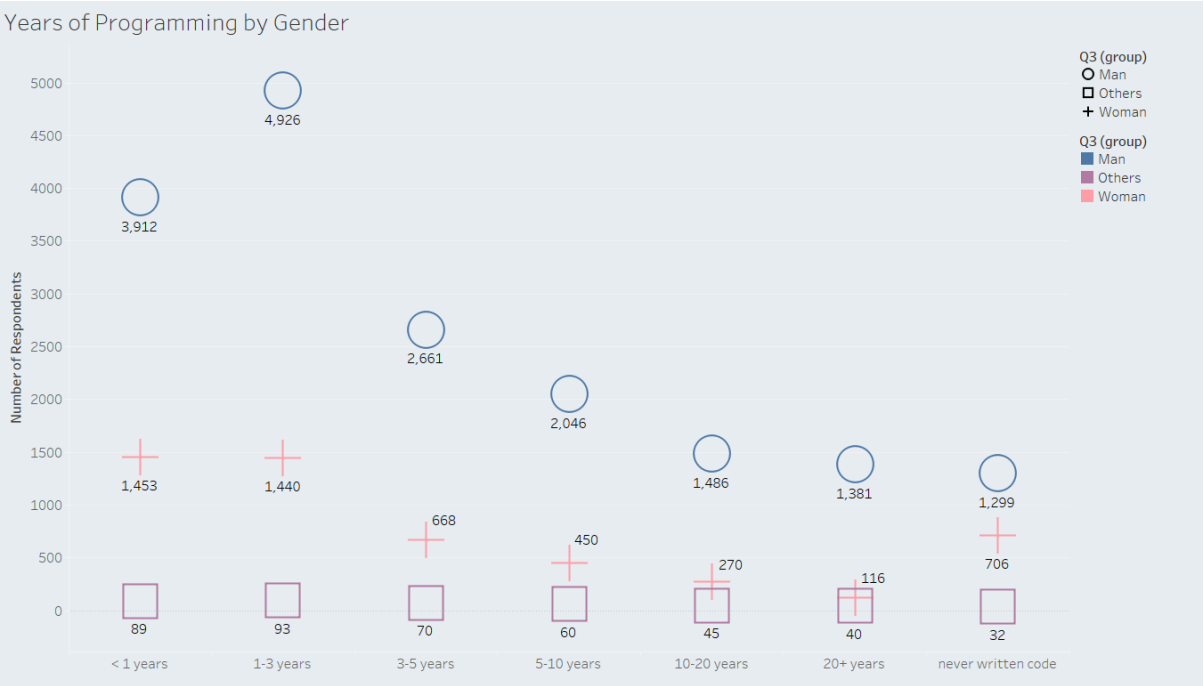


Figure 9: Years of programming by gender

2.2.5 Favourite Media Sources

Other than online courses or languages, media sources also play a vital role in spreading data science related knowledge. New stuffs, techniques and technologies are discovered every day. Missing out on the news, you might fall behind with the rest of the world. From the word cloud, Kaggle are the survey's favourite followed by newsletters and Twitter.

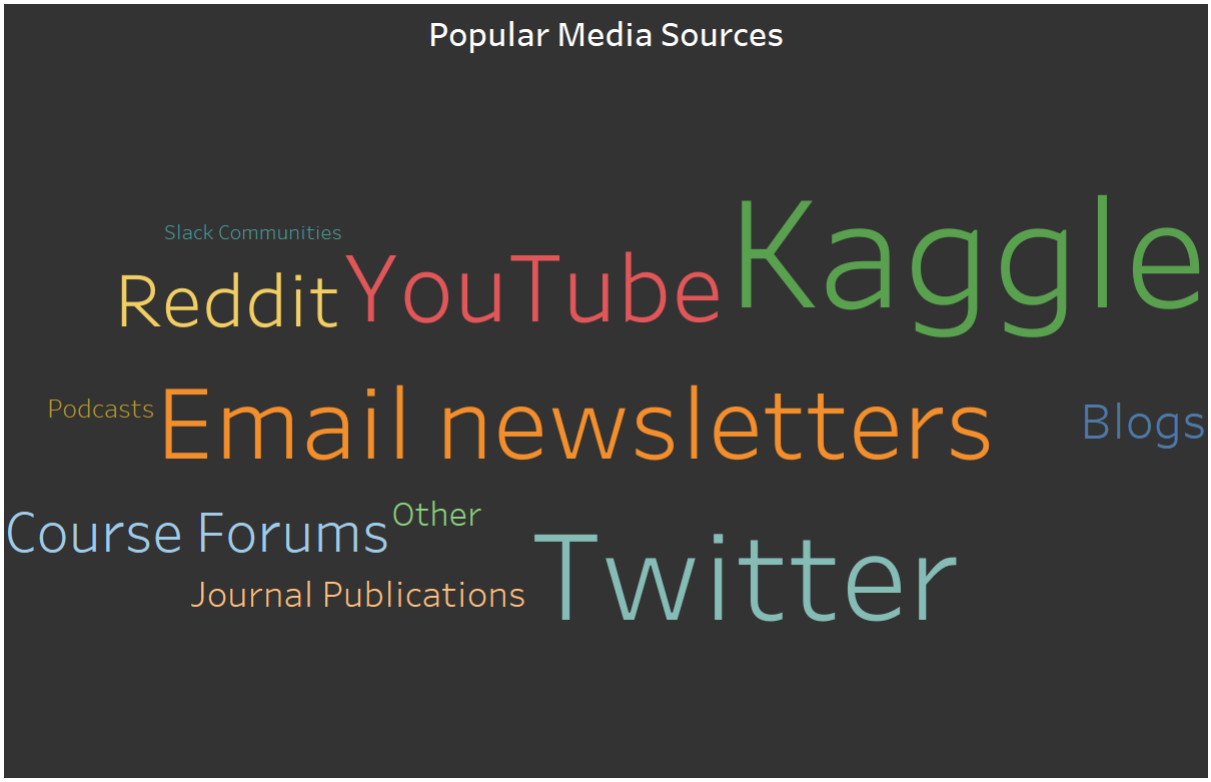


Figure 10: Favourite media sources

2.3 Knowing the Industry

2.3.1 Data Science and Machine Learning: Application in different Industries

The reason why Data Science is a growing field is the applicability of the knowledge in various sectors. The availability and volume of data we produced today, it makes sense that every sector and organization want to have data scientists to turn raw data into profitable insights. Figure 11 shows some of the top industries that uses DS and ML. Apart from the obvious computers/technology sector at the top of the list, education (15.9%), finance (8.8%) and medical (5.6%) are in high demand for data scientist for reasons such as research, stock market analysis, and detecting diseases respectively.

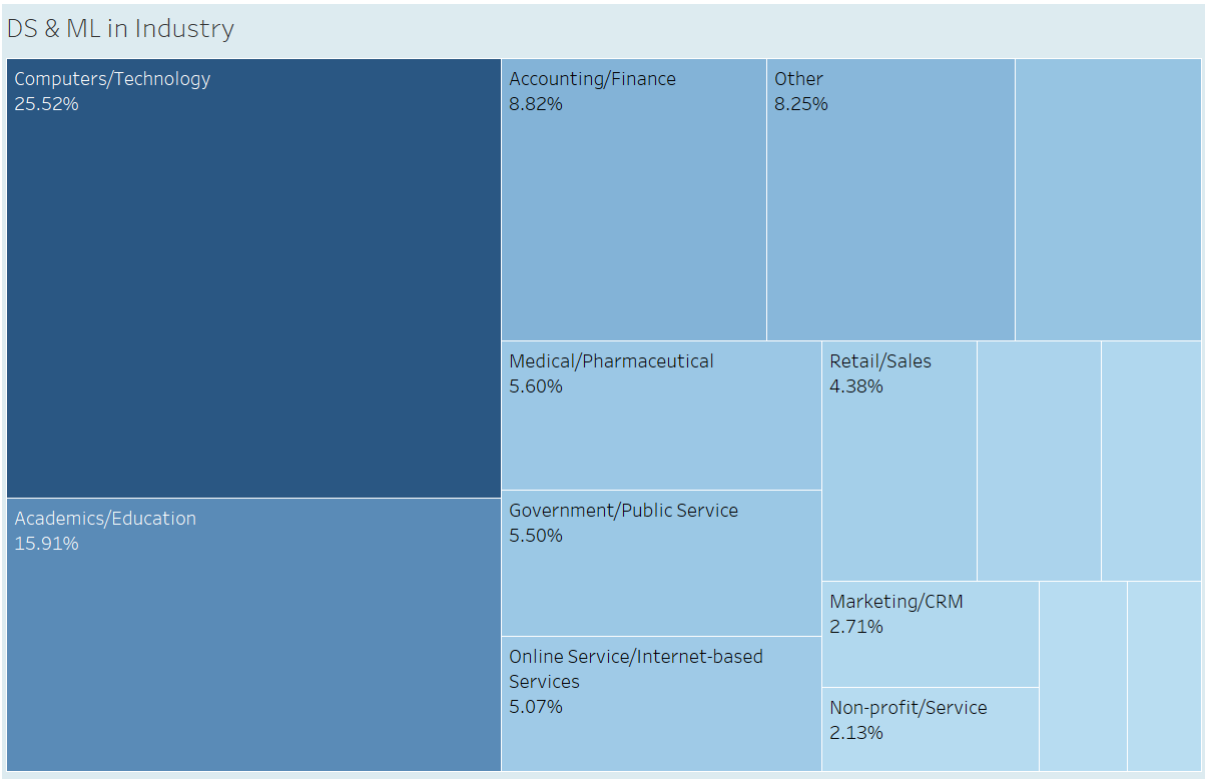


Figure 11: DS and ML in different industries

2.3.2 Data Scientist: One-man army or team player?

Is one data scientist enough? Technically, they can do whole end-to-end data science process on their own, but it does not mean they should. Well, it depends on the company size. Figure 12 below shows that for smaller company (0-49 and 50-249 employees), it is normal to have less than 10 employees for a team. And generally, 5 to 9 employees following the median for all size categories. For bigger company (1000 employees and above), most of them have more than 20 employees to help them in data related works. Depends on your company's business requirements, sometimes it is good to have a small team and sometime it is better to invest more money hiring more talents working together.

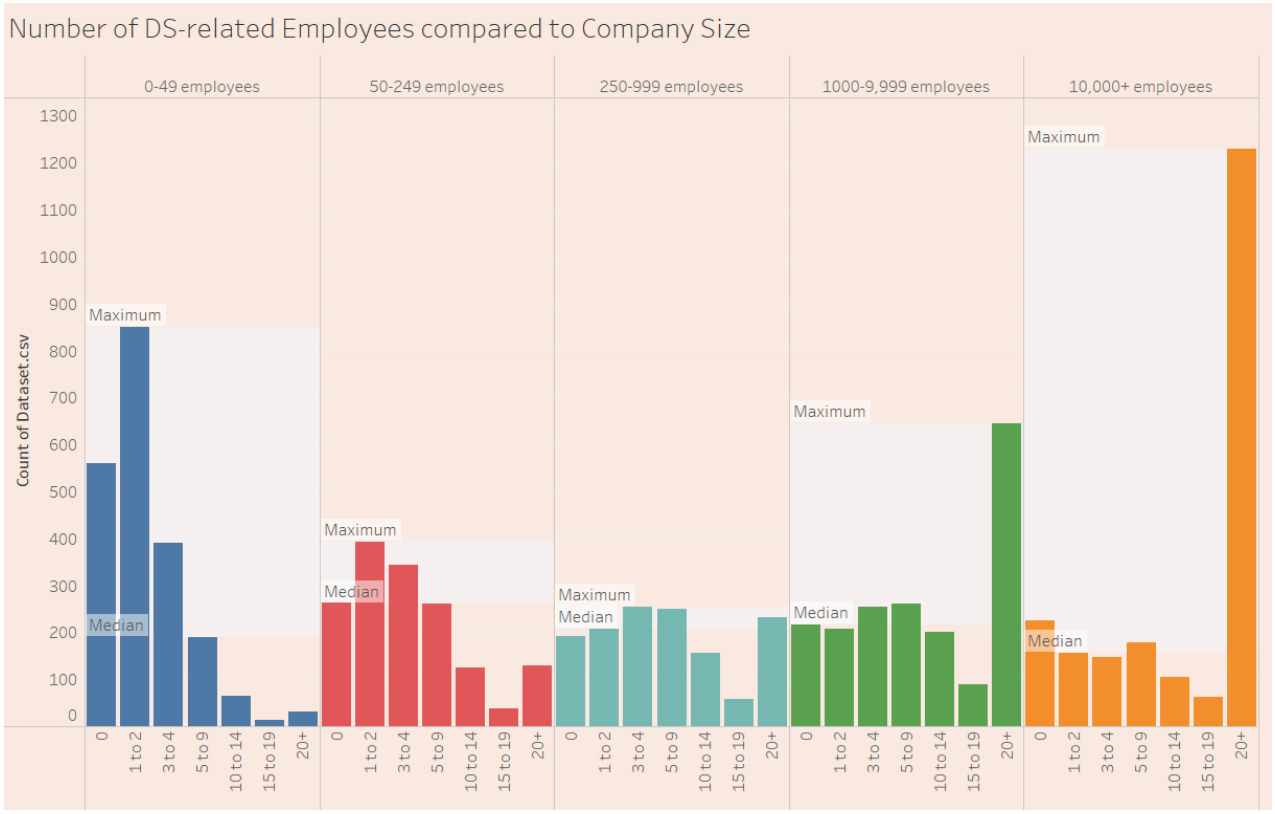


Figure 12: DS-related employees to company size

2.3.3 Job Title: Which title are popular on demand?

Finding out which one is popular is easy with the data we have, but finding out which suits you better is for you to decide. Figure 13 below says that data scientist and data analyst are the top 2 popular job titles for both genders with female preferring data analyst more than data scientist. Coming close to data scientist, with 18%, female have more workforce in education as a teacher/professor meaning they are better in teaching and prefer them instead of other technical jobs.

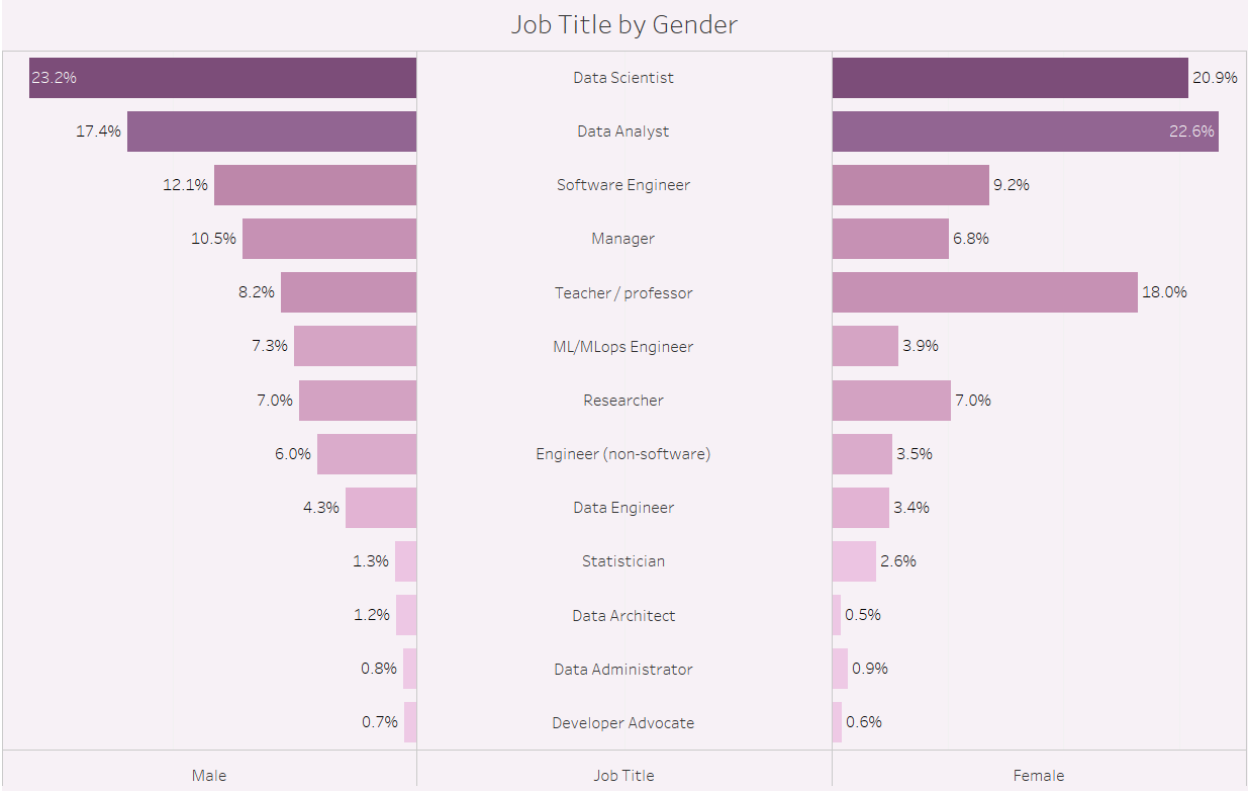


Figure 13: Job Title

2.3.4 Salary Distribution: Equal wage or inequalities?

Focusing on the top 2 popular jobs, while there are more male, there are female representatives on every spectrum of the range including the high-end of the salary of USD 200K and above. This tells us that skilled female talents do get paid based on their skills instead of their gender. It safe to say that the charts will be more even if there are more female getting into data science.

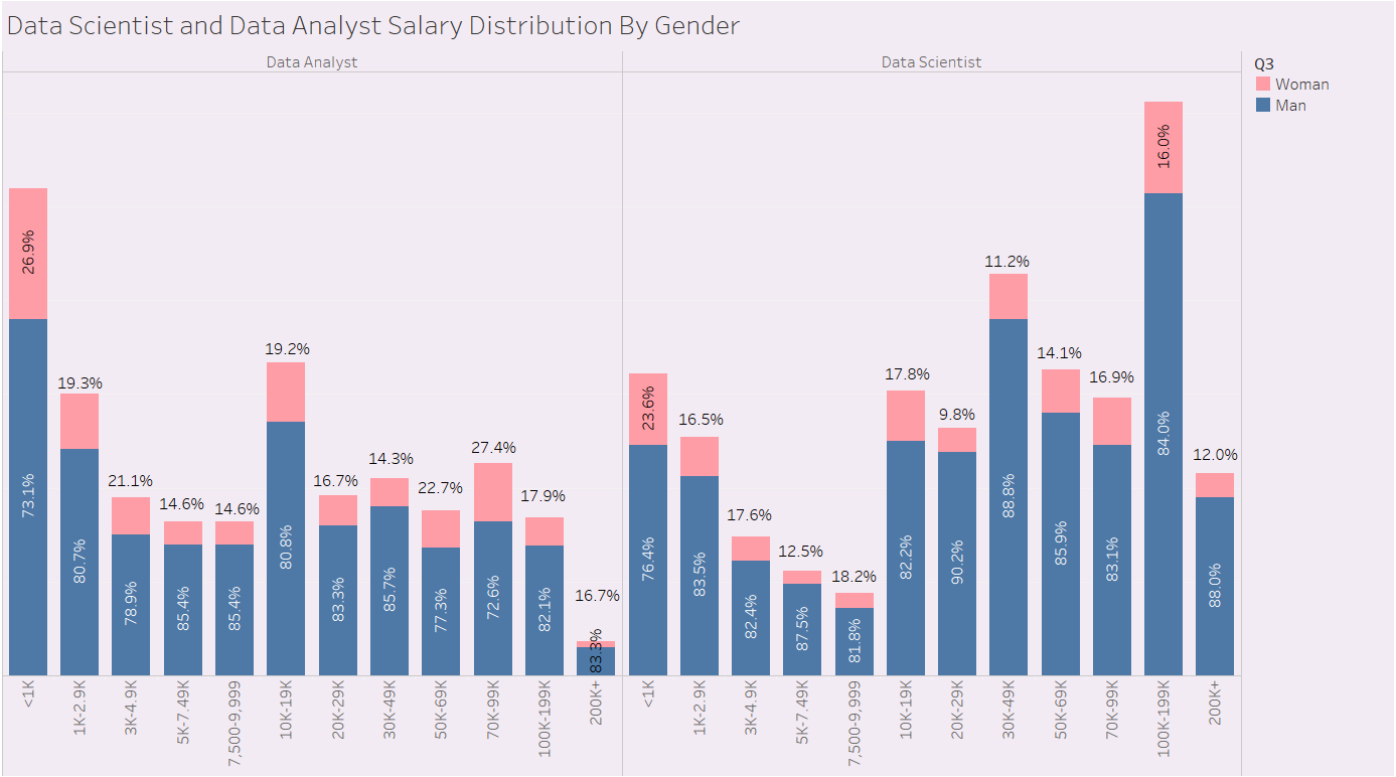


Figure 14: Salary by gender

2.3.5 Salary Distribution: Experienced brings more money?

Experienced do brings in more money. Figure 15 shows that middle-aged have higher percentages in salary more than USD30K and on average of USD100k to USD199K yearly salary compared to younger people with mostly of 10K to 19K. On average, data scientist has higher salary than data analyst and showing large difference when entering USD100K salary range between both jobs meaning data scientist is the more lucrative job.

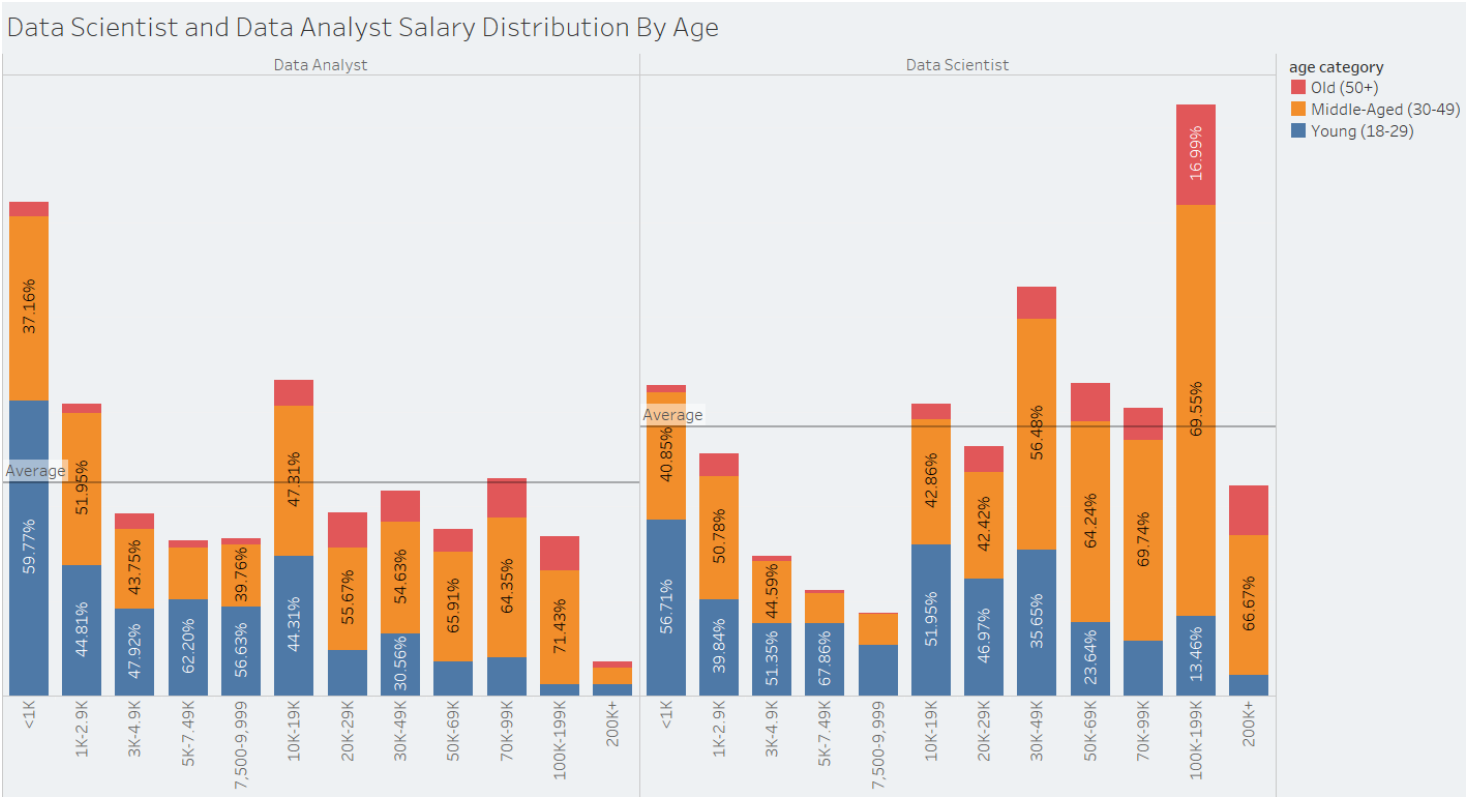


Figure 15: Salary by age category

2.3.6 Popular Tools in the Industry

2.3.6.1 Cloud Platform

Figure 16 and 17: Cloud Platform and experience

2.3.6.2 Data Storage Tools

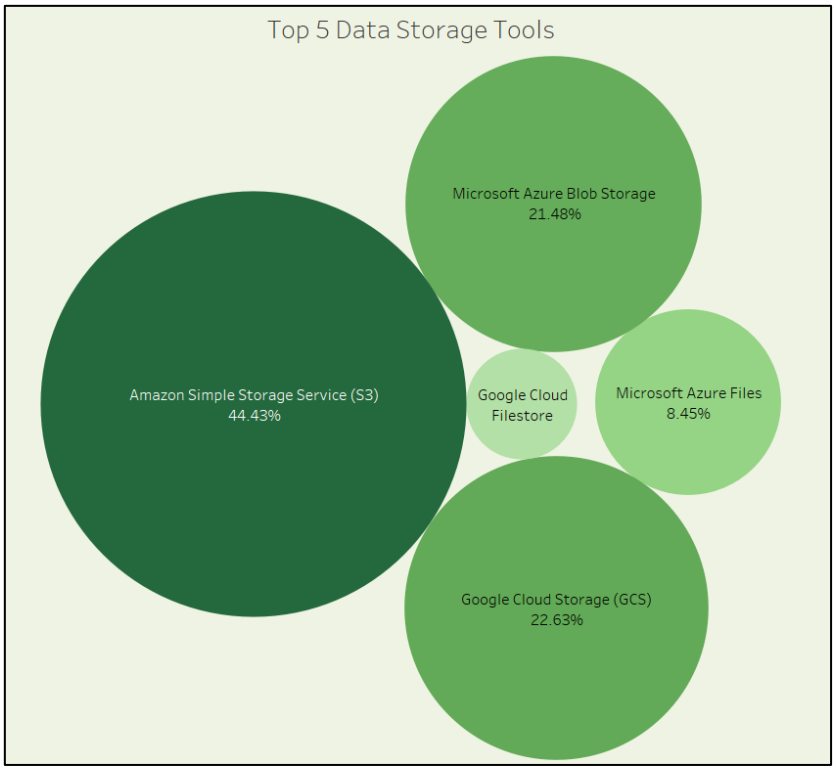


Figure 18: Popular data storage tools

2.3.6.3 Business Intelligence Tools

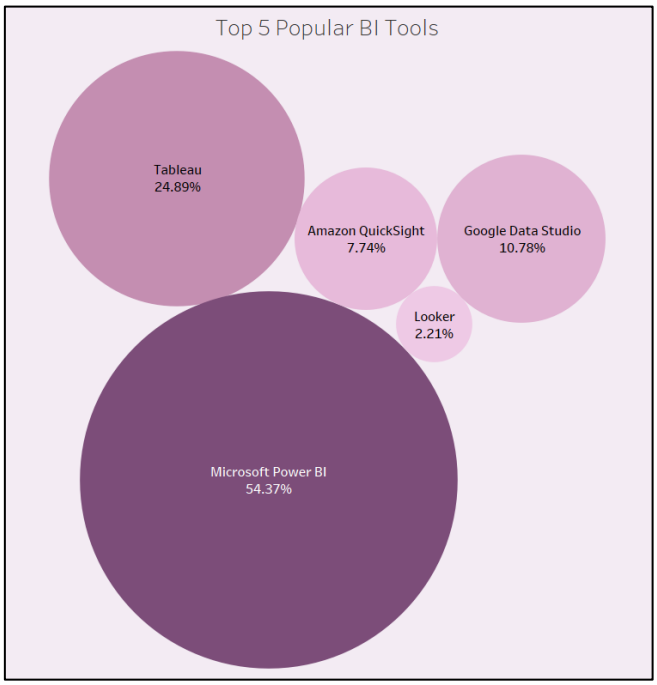


Figure 19: Popular BI tools

2.3.6.4 Machine Learning Tools

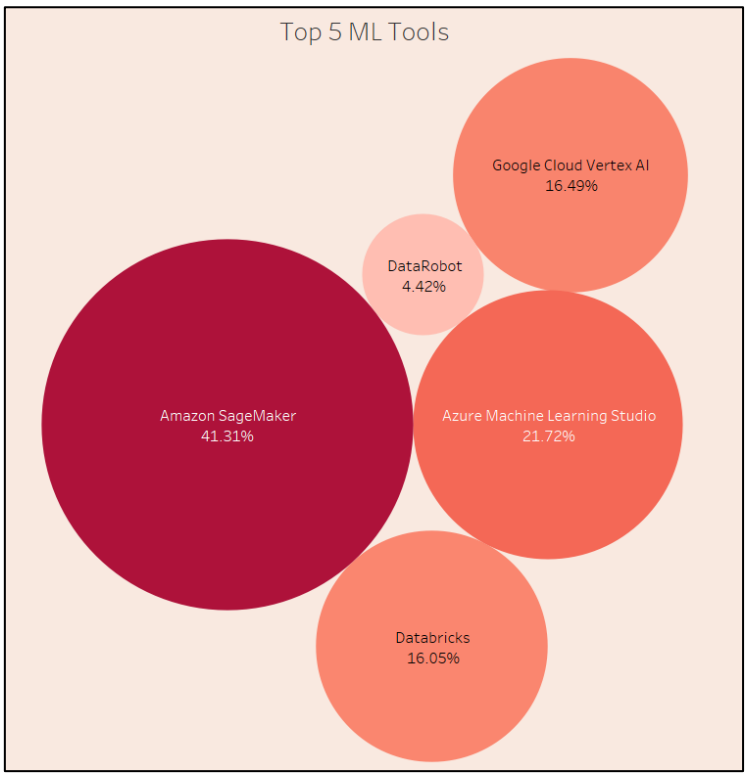


Figure 20: Popular ML tools

3. Conclusion

Going through this report, we have known what are our community comprises of, what learning medium they gone through and the situation over the industries. To summarise things up, the community consists of people from every gender while keeping the male-dominated status and from all age with younger people coming in much more. Other than that, online education is preferred although most of them have formal education and you can learn them no matter your age and gender.

In terms of industry, knowing the tools the industries are using is important so you can plan your learning journey on what you are going to learn. It is highly inefficient to spend your valuable time and efforts to learn certain tool that have little to no value in industry. Hence, for new learners, studying those languages, IDEs and tools might help you in building your technical portfolio later on.

All in all, although looking into others are helpful, it would be meaningless without your own efforts. Stop worrying all these tools and start taking baby steps learning one-by-one. Slowly but surely, you will be able to build your skills needed by the industry. Keep yourself curious, keep study and keep visualise your data journey.

4. Documentation

Pre-processing:

- Remove row for question title
- Remove null and nan values
- Fixing Q26 values because excel detects and convert the value as date
- Transform every question with multiple answers column into a single column through merging mismatched fields in Tableau

Producing each figure:

- Go through this drive for snapshot for each sheet on how it was made: https://drive.google.com/drive/folders/1yJrZAoq3NZcW_9-MskDeRXW1fGT_VawQ?usp=sharing
- Link to each figure: https://drive.google.com/drive/folders/1sQh7wUUXxTACpr4EOW1eEY0kxNbtRfuZ?usp=sharing

5. References

74+ shocking women in tech statistics (2022). (2022, August 15). Retrieved from https://explodingtopics.com/blog/women-in-tech

Azad. (2022, October 9). How to design butterfly chart in tableau in easy way. Retrieved from https://analyticsplanets.com/how-to-design-butterfly-graph-in-tableau-in-easy-way/

Donut chart tableau | How to create a donut chart in tableau. (2021, February 23). Retrieved from https://www.analyticsvidhya.com/blog/2021/02/how-to-create-donut-chart-tableau/

Gender wage gap in tech industry 2021. (2022, June 15). Retrieved from https://www.statista.com/statistics/1254602/tech-gender-wage-gap-for-same-job/

Getting started – Amazon simple storage service (S3) – AWS. (2021). Retrieved from https://aws.amazon.com/s3/getting-started/

Krishnan, A. (2019, November 12). Why Indians are great at tech jobs. Retrieved from https://medium.com/@ajaykrishnan25/why-indians-are-great-at-tech-jobs-786709af497

Pandey, P. (2019, February 20). Word clouds in tableau: Quick & easy. Retrieved from https://towardsdatascience.com/word-clouds-in-tableau-quick-easy-e71519cf507a

Unwin, A. (2020). Why is data visualization important? What is important in data visualization? *2.1*. doi:10.1162/99608f92.8ae4d525

What is Amazon SageMaker? (2021). Retrieved from https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html

Why is it important to have women in tech | Ironhack blog. (2022, September 15). Retrieved from https://www.ironhack.com/us/en/blog/why-is-it-important-to-have-women-in-tech

Why you don't need a degree to work at a top tech company in 2022. (2021, December 30). Retrieved from https://woz-u.com/blog/degree-work-top-tech-company-2022/