A Report

On

Data Science and MLOps Landscape in Industry

Introduction

As a Data Analyst in the IT sector, I strongly believe that the integration of advanced data analytics and tools like Python, SQL, and R could transform traditional IT operations into more efficient, data-driven systems capable of keeping pace with the demands of modern technology-driven businesses. The adoption of these analytical tools can also benefit other industries by enhancing decision-making and operational efficiency. The findings from the latest McKinsey Global Survey about the state of AI in 2021 indicate that the use of data-driven technologies continues to grow, with significant benefits for organizations that leverage them effectively. A majority of McKinsey survey respondents now report that their organizations are adopting data analytics capabilities, reflecting their increasing impact on business outcomes.

However, effectively utilizing and scaling data analytics to deliver actionable insights can be challenging. My experience has shown that, while many IT teams have begun exploring data analytics, only a fraction of these initiatives fully transition into operational use. Moving from the exploratory phase of data analysis to practical, scalable solutions is complex, as it requires tailoring analytical models to meet specific business needs and ensuring they integrate seamlessly into existing IT systems.

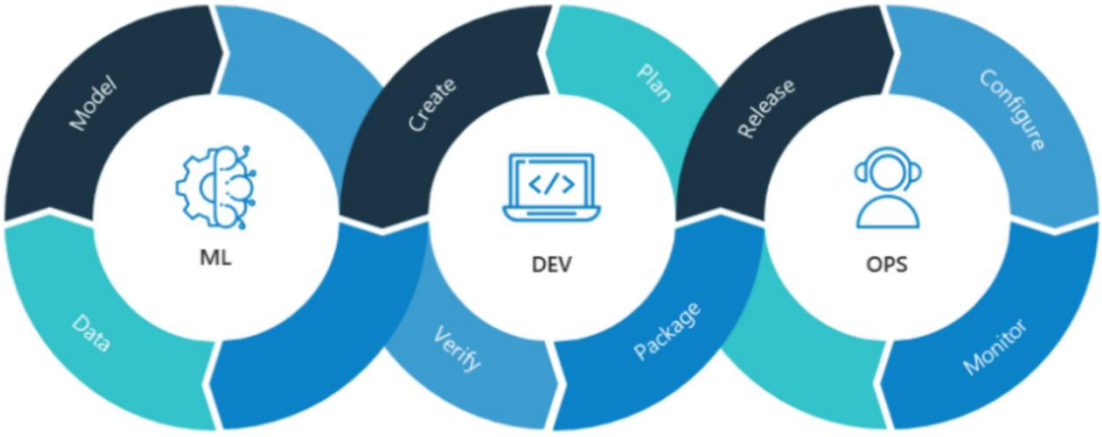

Fig. Overview of MLOps

Models as part of an experiment are good, but models in production are great. MLOps, as the name implies, brings operationalization to the table, providing resources for bringing models from test environments into production.

Analysis's Target

The goal of this analysis is to extract insights from the responses of 2022 Kaggle Machine Learning & Data Science Survey about the state of AI Adoption and ML Operationalization in the industry in 2022 as well as about the Data Science landscape in the market. As I'm curious to see how the MLOps and AI adoption progressing in other organizations and what's the current trends in Data Science I'll try to enlighten the following main topics:

1. **What's the state of Machine Learning adoption in the enterprise today?**
    - What's the percentage of enterprises deploying data science and machine learning in production today?
    - Does the company's size or sector play a role in AI Adoption? Are larger companies more likely than smaller companies to have deployed AI in their organization?
2. **What's the enterprise AI tech stack?** The modern AI stack is a collection of tools, services, and processes imbibed with MLOps practices that allow developers and operations teams to build ML pipelines efficiently in terms of resource utilization, team efforts, end-user experience, and maintenance activities. It would be interesting if, for instance, we would answer the following questions:
    - Are Cloud-native solutions a must-have for business today?
    - What are the most popular tools for Data Storage, Data Management, AutoML, Business Intelligence, etc.?
    - What frameworks and libraries are commonly used in the market for Machine Learning and Data Science?
    - Are transfer learning methods mature enough to be used in the business environment?
    - Do we really work with big data and deep learning methods to such an extent that we need specialized hardware for ML models training?
3. **AI Careers & Job Outlook in 2022:**
    - What are the top AI job positions?
    - What does an AI professional do?
    - What are the professional AI skills in demand for 2022?
4. **AI Salary Overview**

▪ Methodology

In order to have as much as I can a representative dataset for the analysis, I'll keep in the dataset only the professionals, namely the respondents that fulfill the criteria listed below:

- currently are not students (answer **No** the **Q5** question)
- currently are employed (They didn't answer "Currently not employed" to the **Q23** question)
- have answered in what industry they are currently employed (or their most recent employer if retired) - **Q24 question has an answer, not None**

As it can be seen below, ~ **37.9% of the total responses** meet the above criteria and the analysis will be conducted based on these responses.

Key figures

23,997	9094	37.9%
# of respondents in the survey	# of professionals	of the respondents are in the analysis scope

To ensure that the dataset is as representative as possible for the analysis, a rigorous selection process was implemented to filter out only those respondents who meet specific professional criteria. This approach aims to focus on individuals who are actively engaged in their careers and can provide valuable insights relevant to the study's objectives.

The first criterion stipulates that respondents must not be students; therefore, anyone who answered "Yes" to Question 5 regarding their student status was excluded from the dataset. Additionally, to maintain the integrity of the professional sample, only those who indicated they are currently employed by not selecting "Currently not employed" in response to Question 23 were retained. Furthermore, to ensure that the respondents' industries are accurately represented, it was necessary for them to have provided an answer to Question 24, which asks about their current or most recent industry of employment. Respondents who left this question unanswered were also excluded.

After applying these stringent filters, it was found that approximately 37.9% of the total survey respondents met all the specified criteria and were thus included in the final dataset for analysis. Out of the initial 23,997 respondents, 9,094 individuals qualified as professionals based on the outlined parameters. This subset of respondents represents a focused group of active professionals across various industries, providing a robust foundation for the subsequent analysis. By concentrating on this refined dataset, the study aims to derive meaningful conclusions that are more closely aligned with the experiences and perspectives of working professionals, thereby enhancing the relevance and applicability of the findings.

Outlier Analysis

It would be also interesting to examine if there are some **"outlier respondents"** that have marked all the answers for the multiple-choice questions.
For that, I calculated the average number of choices that each respondent selected in the multiple-choice questions. I found out that each respondent selects 1 - 2 options in the multiple-choice questions on average.

Only 2% of the respondents in the scope have an average number of selections greater than 3, which cannot affect the results of the analysis. **Also, it doesn't necessarily mean that we have to address them as outliers. One explanation would be that they might have many years of coding or ML experience, so makes sense to be familiar with many frameworks and work with a variety of libraries.**
As the tables below illustrate, this hypothesis is valid since the biggest percentage of the respondents with more than 3 selections on average, have strong coding and machine learning experience.

So I won't discard these respondents or treat them differently.

	Years of Machine Learning Experience	Nbr of respondents	%
0	2. < 1 years	12	6.320000
1	3. 1-2 years	30	15.790000
2	4. 2-3 years	34	17.890000
3	5. 3-4 years	24	12.630000
4	6. 4-5 years	27	14.210000
5	7. 5-10 years	49	25.790000
6	8. 10-20 years	14	7.370000

Fig. Machine Learning Experience

	Years of Coding Experience	Nbr of respondents	%
0	2. < 1 years	10	5.260000
1	3. 1-3 years	32	16.840000
2	4. 3-5 years	28	14.740000
3	5. 5-10 years	50	26.320000
4	6. 10-20 years	41	21.580000
5	7. 20+ years	29	15.260000

Fig. Coding Experience

In the table below, we can also see the average number of choices that respondents selected for each of the multiple-choice questions and we might be able to conclude the following findings:

- ✓ The professionals who participated in the survey, use on average 2 programming languages on a regular basis, 3 Machine Learning Algorithms, and 2 Machine Learning Frameworks.

In addition, they usually don't use natural language processing (NLP) methods like Word embeddings/vectors (GLoVe, fastText, word2vec), Encoder-decoder models (seq2seq, vanilla transformers), Contextualized embeddings, or Transformer language models

	Question	Question Title	Nbr of available Choices	Average number of selected choices
0	Q6	On which platforms have you begun or completed data science courses?	12	2
1	Q7	What products or platforms did you find to be most helpful when you first started studying data science?	7	2
2	Q10	Did your research make use of machine learning? - Yes, the research made advances related to some novel machine learning method (theoretical research)	3	0
3	Q12	What programming languages do you use on a regular basis?	15	2
4	Q13	Which of the following integrated development environments (IDE's) do you use on a regular basis?	14	3
5	Q14	Do you use any of the following hosted notebook products?	16	1
6	Q15	Do you use any of the following data visualization libraries on a regular basis?	15	2
7	Q17	Which of the following machine learning frameworks do you use on a regular basis?	15	2
8	Q18	Which of the following ML algorithms do you use on a regular basis?	14	3
9	Q19	Which categories of computer vision methods do you use on a regular basis?	8	1
10	Q20	Which of the following natural language processing (NLP) methods do you use on a regular basis?	6	0
11	Q21	Do you download pre-trained model weights from any of the following services?	10	1
12	Q28	Select any activities that make up an important part of your role at work:	8	2
13	Q31	Which of the following cloud computing platforms do you use?	12	1
14	Q33	Do you use any of the following cloud computing products?	5	1
15	Q34	Do you use any of the following data storage products?	8	1
16	Q35	Do you use any of the following data products (relational databases, data warehouses, data lakes, or similar)?	16	1
17	Q36	Do you use any of the following business intelligence tools?	15	1
18	Q37	Do you use any of the following managed machine learning products on a regular basis?	13	1
19	Q38	Do you use any of the following automated machine learning tools?	8	1
20	Q39	Do you use any of the following products to serve your machine learning models?	12	1
21	Q40	Do you use any tools to help monitor your machine learning models and/or experiments?	15	1
22	Q41	Do you use any of the following responsible or ethical AI products in your machine learning practices?	9	1
23	Q42	Do you use any of the following types of specialized hardware when training machine learning models?	9	1
24	Q44	Who/what are your favorite media sources that report on data science topics?	12	3

Fig. Average number of choices by questions

It covers a range of topics from programming languages and IDEs to machine learning frameworks and cloud computing services. The survey includes 25 questions with an average of 1-3 selected choices per question, providing insights into the current trends and preferences in the field. The findings aim to guide future developments and resource allocation in data science.

Index

1. What's the state of Machine Learning adoption in the enterprise today?

The first thing that I want to understand from the survey responses, is the state of ML adoption in different industries today. In the 2022 Kaggle Machine Learning & Data Science survey of 9,094 professionals coming from different industries, as it can be seen in the chart below,

- a percentage of 25.52% working in tech companies,
- a 15.91% in the academic field,
- and the rest distributed from the finance sector to shipping and transportation.

✓ ***Which sector would you bet is a high performer in AI and has made big progress in terms of AI adoption?***

Before I answer that, let's see how AI adoption looks like broadly, across all sectors, in 2022.

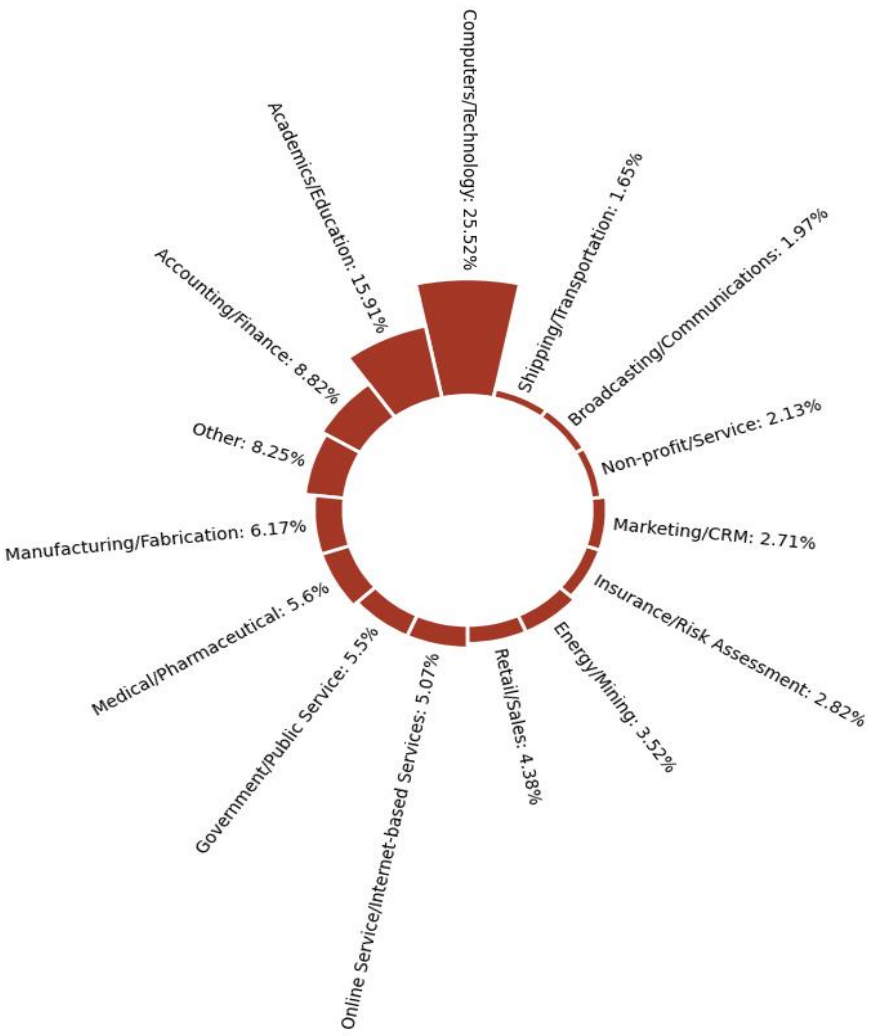

Fig. AI across sectors

The data shows that about ~ **33%** of respondents say that their organizations have Machine Learning models in production, either in an advanced stage or in an intermediate stage (they recently started using ML methods), while a percentage of **10.2%** uses ML methods for generating insights. However, a considerable percentage of the participants, **21.7%**, answered that their companies haven't started yet using AI and ML techniques while **17.1%** of the respondents say that have started exploring the capabilities of this new technology.

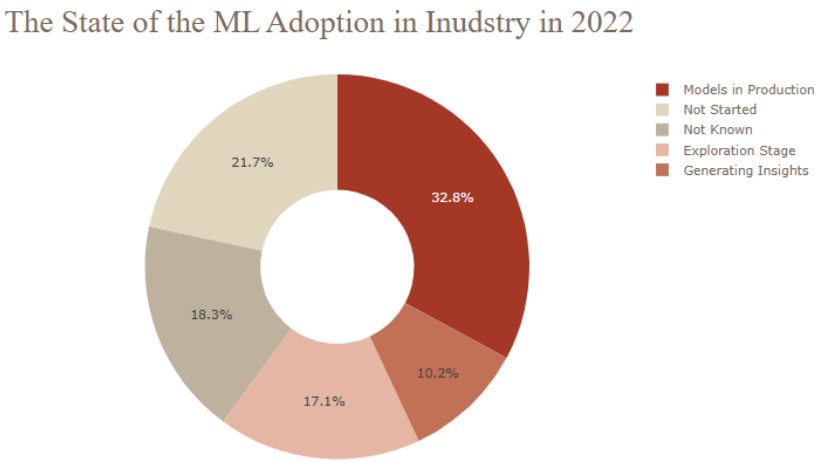

Fig. ML adoption in Industry in 2022

From the above analysis, it reveals that 32.8% of organizations have models in production, while others are at various stages of exploration and insight generation. The data highlights the varying levels of ML integration across different sectors.

Fig. State of ML adoption

The above visual visualizes the progression of machine learning adoption across various stages, from initial exploration to advanced implementation. The chart highlights the

distribution of organizations at different ML maturity levels, emphasizing their journey toward production-ready models.

Now, let's come back to the question above and try to answer it by extracting some insights from the survey results.

It is clear in the following chart that companies providing **Internet-based services** have a **better adoption of Machine Learning and Data Science** followed by **Insurance companies**, whereas **non-profit organizations and the government sector score undoubtedly lower for the adoption of various AI-related technologies.** A key reason for the lower AI adoption among governments and non-profit organizations is the bureaucracy and the established processes that take too long. In these sectors, might be less encouragement for employees to take risks and innovate.

In the private sector, employers tend to put a strong focus on experimentation, innovation, and growth. For instance, **companies providing Internet-based services** could gather many data from the user's online activities and the employees can apply analytics and other innovative ideas in order to improve the services that their company provides. The **insurance sector** is also leveraging AI technologies for insurance advice, underwriting claims processing, fraud prevention, risk management, and direct marketing. Customer behavior and advances in technology have opened the door for AI in the insurance market to create value, reduce costs, increase efficiency and achieve higher customer satisfaction and trust. **Retail** has also embraced AI technologies, with 27% of the professionals working in the retail sector, saying their companies have well-established machine learning methods in production.

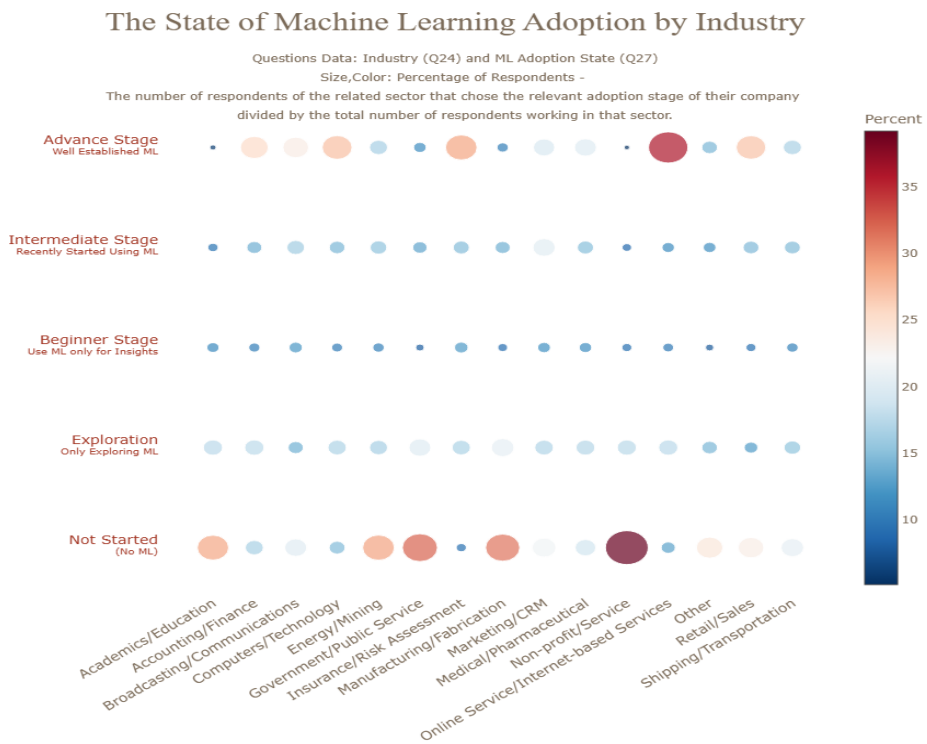

Fig. Progress of ML adoption

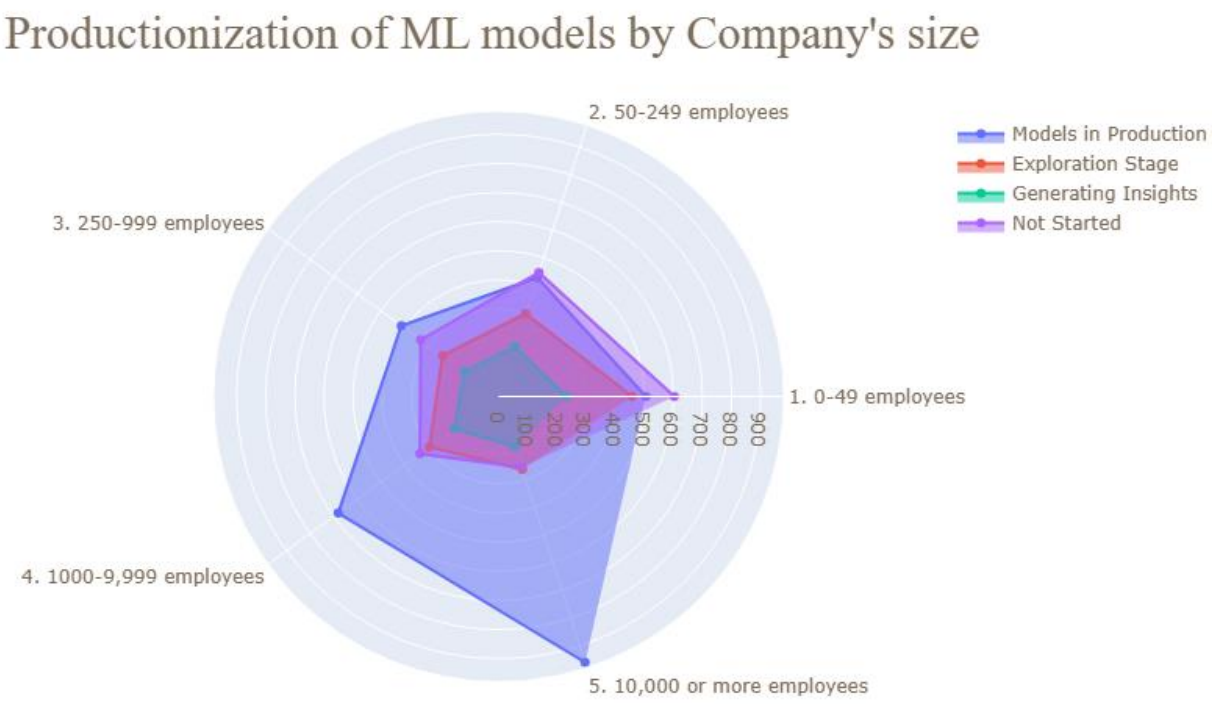

Fig. Productionization of ML models by Company size

Another important insight that comes up from the analysis is that big companies are leading the way in AI adoption.

The survey results show that larger companies, with 1000-9,999 employees or more than 10,000 are the leading AI adopters. There are several reasons that may explain why larger companies outpace smaller ones in AI adoption. For one, because large firms tend to serve large markets, they can better amortize the high fixed costs associated with employing AI production technologies over more sales. In addition to that, larger firms offer higher wages and more benefits, increasing the pool of top AI talent these firms have access to. Finally, because vendors of AI systems benefit from supplying companies with the largest consumer base, vendors may focus on creating relationships and contracts with larger firms, enabling these firms to be more exposed to the value AI systems can bring to their businesses.

In the next section, I'll explore tools and practices used in the market, according to the survey responses to establish an adaptable infrastructure for Machine learning and Data Science projects.

2. Overview of the enterprise AI technology stack

Machine learning was mainly in the experimental stage in the enterprise market not long ago. The Data Science teams always start with a Proof Of Concept (POC) approach and eventually gain traction even with a non-standardized production deployment process because of the business results achieved by the model. In order to scale this solution successfully with re-usability and reliability, the AI stack requires hardware and software optimizations in architectural areas of computing, memory, and networking.

o Usage of Cloud Computing Platforms

According to several reports about the Cloud Computing Market in 2022, the adoption of cloud technologies continues to accelerate. Cloud computing has influenced the rise of machine learning and artificial intelligence. Factors such as affordable storage, availability of GPUs, faster AI training and inferencing performance, lower costs, and protection against attacks made machine learning accessible and affordable to businesses. Most companies lack the infrastructure and expertise to implement AI applications themselves.

As the following radar chart depicts, companies that have models in production use also cloud computing platforms which is reasonable since the cloud makes it easy for enterprises to experiment with machine learning capabilities and scale up as projects go into production and demand increases.

	Usage of Cloud Computing Platforms	Nbr of respondents	%
0	No	4994	54.920000
1	Yes	4100	45.080000

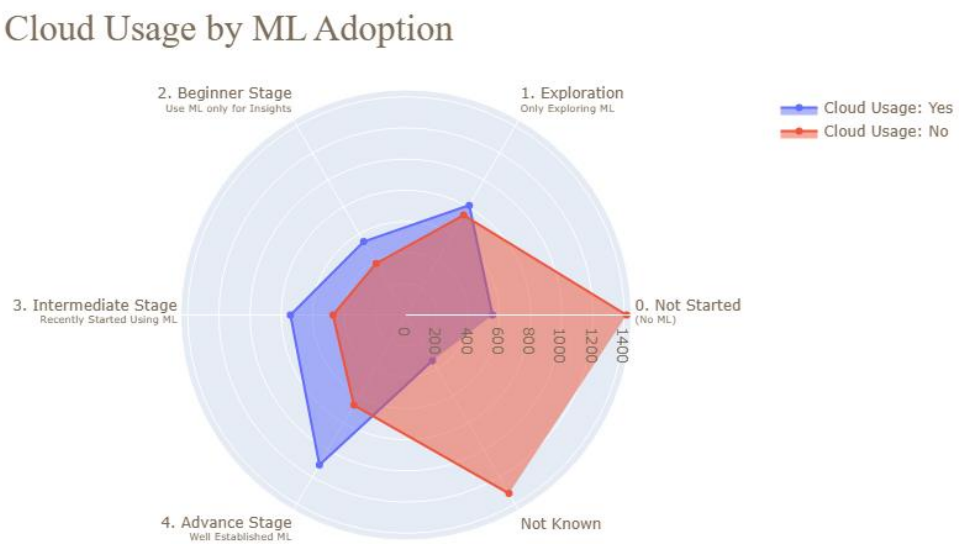

Fig. Cloud Usage by ML Adoption

- **Which cloud computing platforms are used for Machine Learning operations?**

In the following visualizations, we can see the most popular cloud computing platforms by sector as well as by country. It is immediately obvious that Amazon Web Services (AWS) and Google Cloud Platform (GCP) are the dominant ones as well as that Alibaba Cloud is quite famous in Asia.

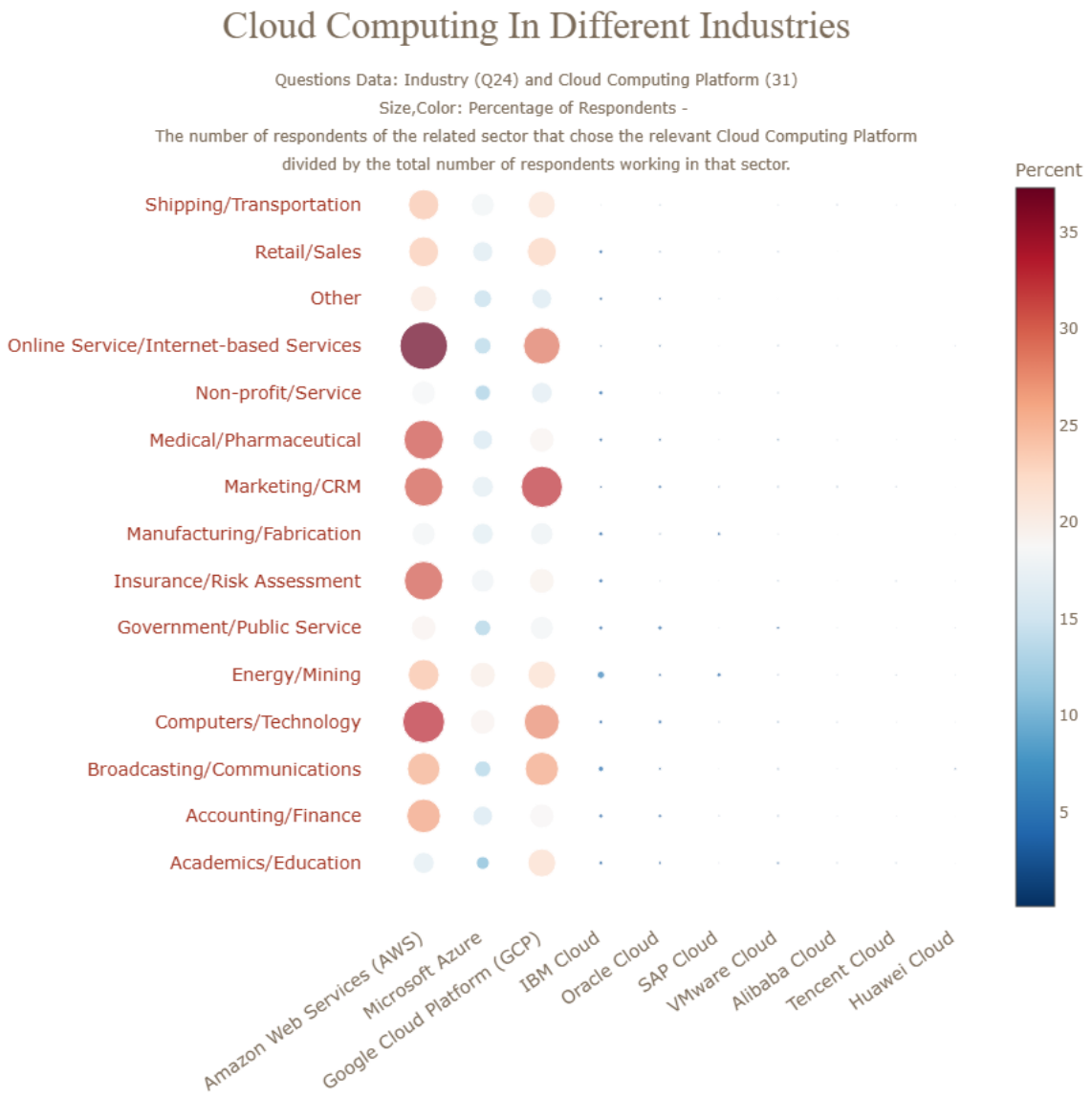

Fig. Cloud Computing Usage in Industries

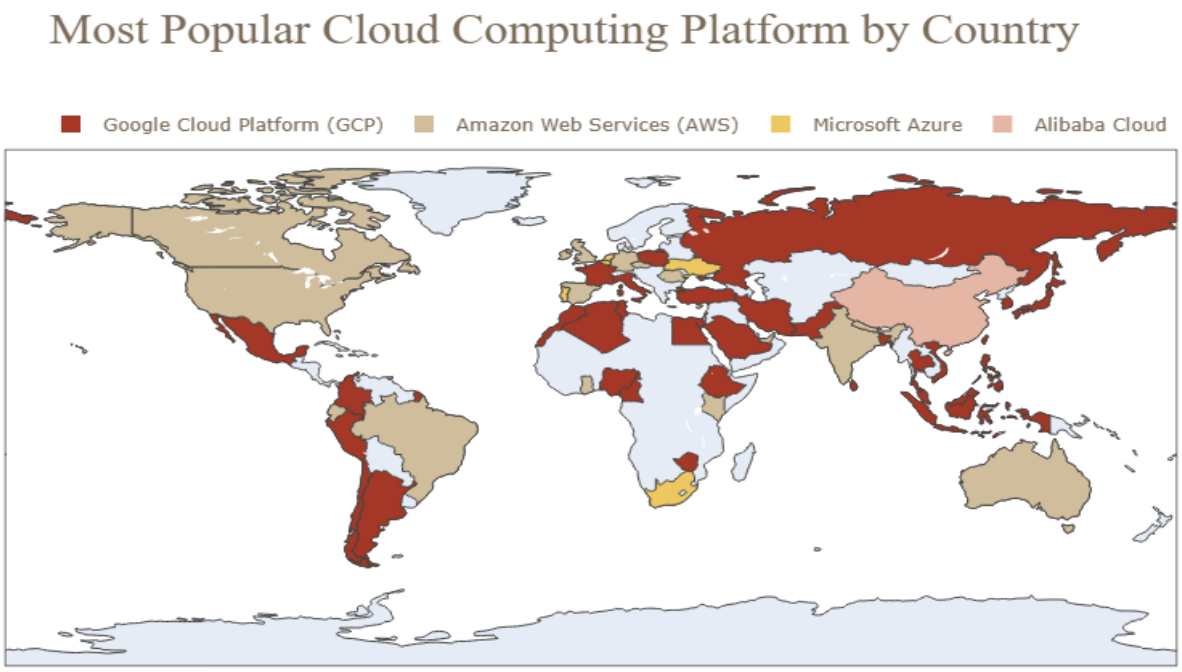

Fig. Popular Cloud Platform in the World

- **Machine Learning tools & products popular in 2022**

The following graphs summarize the usage patterns of other tools, techniques, databases, platforms, and frameworks used by professionals.

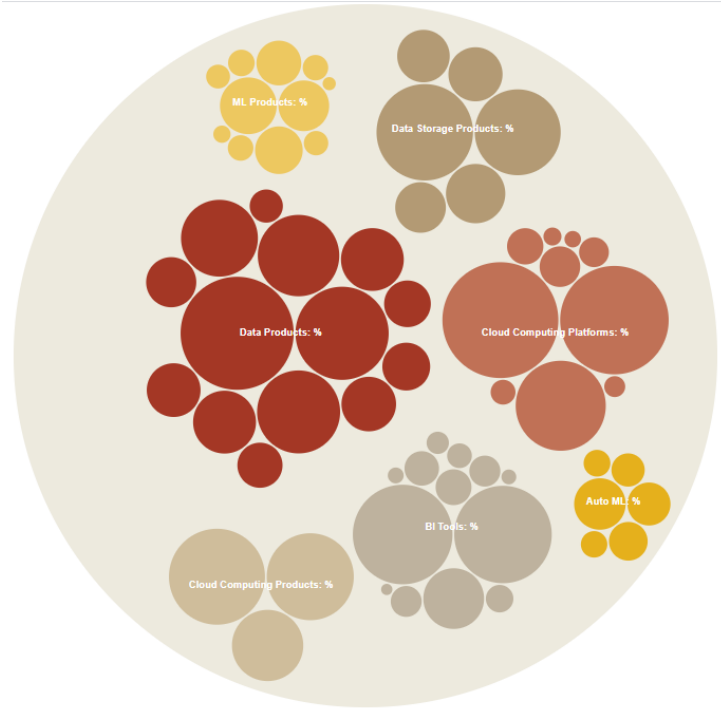

Fig. Machine Learning tools & products popular in 2022

Each company has a unique technology stack with software that they prefer to use with their proprietary data. There are a number of different platforms that go into each category of the stack. These categories include Visualization & Analytics, Computation, Storage Distribution & Data Warehouses. There are too many platforms to count, but in the following illustration, I'll be going over the popular cloud computing services and products that I have seen across the survey responses, offered by the top 4 giant Tech Companies: Amazon, Google, Microsoft & IBM.

- **Amazon top products:**
  - The most commonly used product provided by Amazon is **Amazon Web Services (AWS)** cloud computing platform, as it is used by 2346 respondents out of 9094 (**25.8%** of the professionals).
  - The second most popular is the Amazon Simple Storage Service (S3) as it's used by 17.8% of the respondents in the scope.
- **Google top products:**
  - As above, the most popular product offered by Google is its cloud computing platform, **Google Cloud Platform (GCP)**, used by **22.6%** of the respondents.
  - Secondly comes the Google Cloud Compute Engine which is slightly more popular than the Google Cloud Storage.
- **Microsoft top products:** The Microsoft products that dominate in the market according to the survey respondents' choices are Microsoft Power BI (18.23% of the responses in scope) and **Microsoft Azure** (used by **15.57%** of the respondents), and so it ranks 3rd in the list with the top cloud computing platforms (1st: AWS, 2nd: GCP).
- **IBM top products:** From IBM products, the IBM Watson Studio, followed by the IBM Cloud / Red Hat has gained the most popularity.

NOTES:

- The size of the rectangles in the third level of the treemap indicates the number of respondents using the relevant product/service, while the size of the rectangles and the counts respectively in the second level doesn't correspond to the number of respondents using Amazon, Google, etc. in general. The counts of each of the 4 companies in the second level of the map are just the sum of the respondents that use each of their services/products in the 3rd level. However, if the same user uses two or more products, provided by the same company it will be counted twice in the total sum of the second level. That's why the counts in the second level should not be taken into account as they do not represent the accurate total number of respondents that use them (it's a higher number than expected).

- The color of the rectangles in the third level of the treemap indicates the percentage of the respondents using the relevant product/service and it is applied the same logic as above.

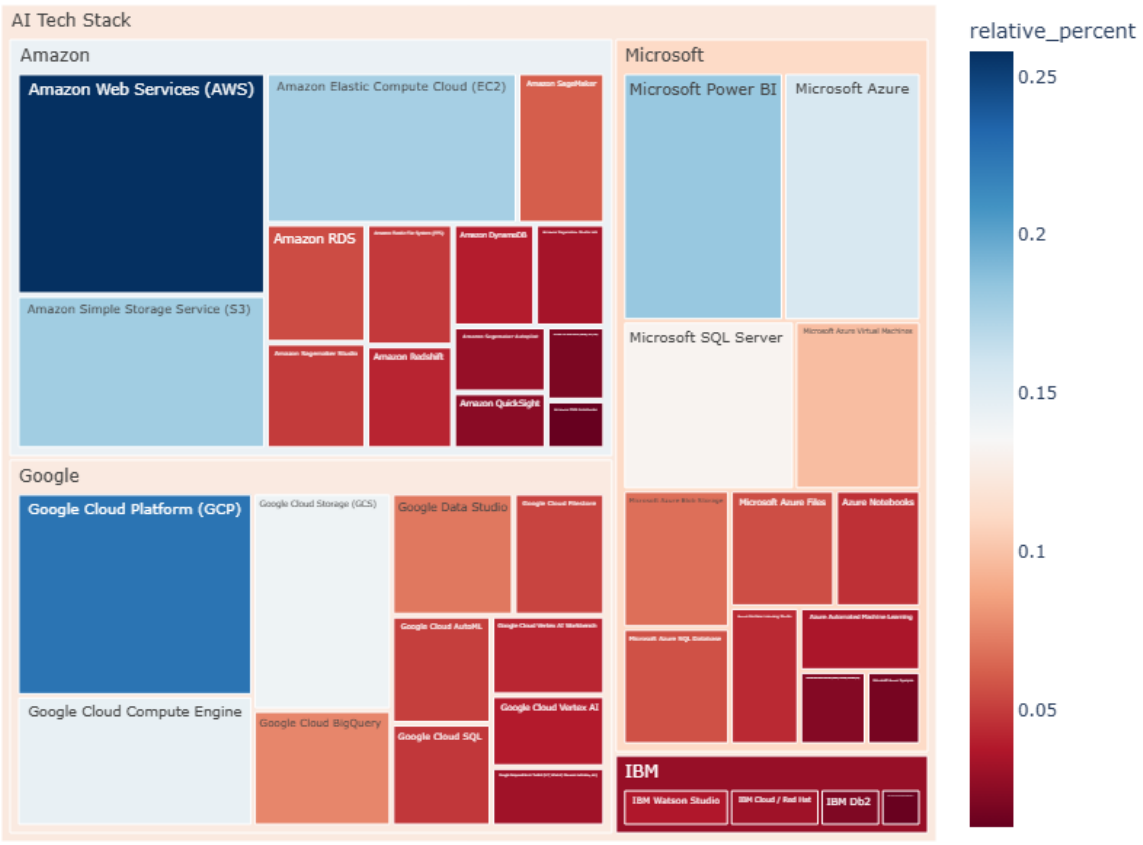

Fig. AI Technology Stack Usage by Cloud Providers: Relative Market Share Analysis

o Frameworks, libraries and languages for Machine Learning & Data Science

Frameworks, libraries, and languages for Machine Learning & Data Science provide the tools necessary to develop, train, and deploy models efficiently. Popular programming languages like **Python** and **R** are widely used due to their extensive ecosystem of libraries. Frameworks such as **TensorFlow**, **PyTorch**, and **Scikit-learn** simplify tasks like data preprocessing, model training, and evaluation. Additionally, libraries like **Pandas**, **NumPy**, and **Matplotlib** help with data manipulation, numerical computations, and visualization, making the entire workflow more seamless.

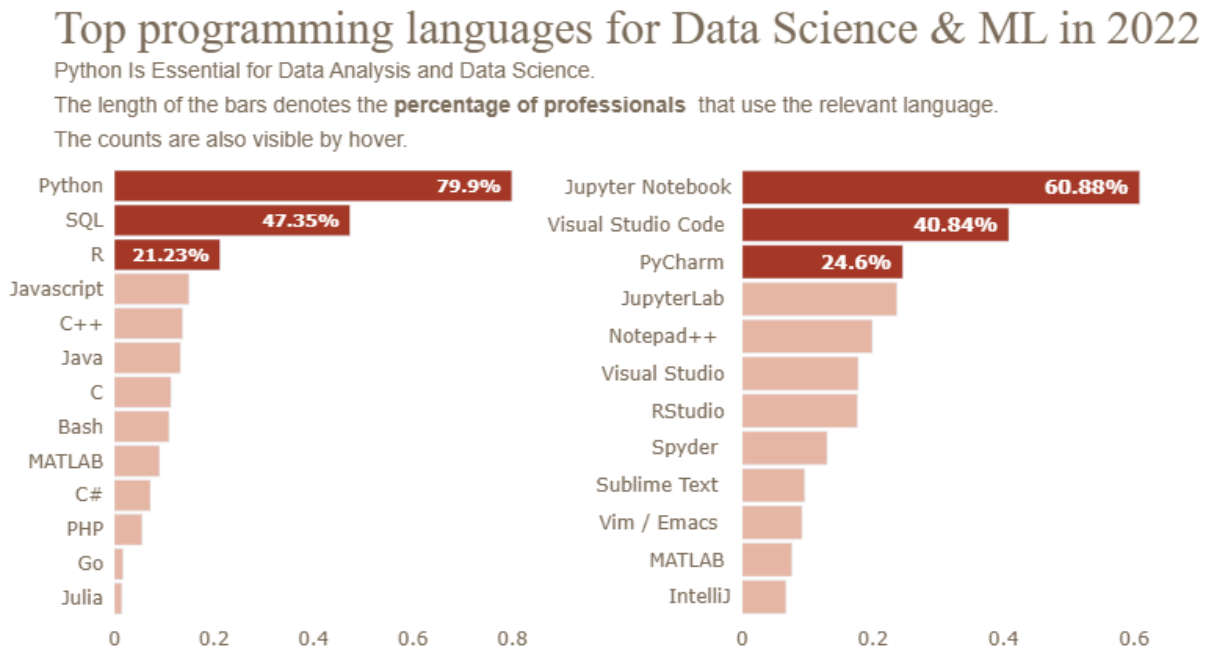

Fig. Top languages for Data Science & ML in 2022

When it comes to the programming languages, the bar plot shows that Python is the most popular language followed by SQL and R.

- **Python** is the dominant language in the Machine Learning and Data Science field with 79.9% of the professionals using it for their daily tasks. Python is widely used in the industry, and it is also by far the language most recommended to beginners.
- **SQL** is necessary required when working with databases. Having at least a basic understanding of SQL and database management would go a long way in your career.
- **R:** a percentage of 21.2% of the respondents working in industry use R. While in most cases Python is the default choice when analyzing data and applying statistical methods, R is preferred as we'll see in a later section by many statisticians.

Together, these languages form a powerful toolkit for data professionals. Python's versatility makes it the leader, especially for machine learning and large-scale data projects. SQL serves as the bridge to accessing and organizing data stored in databases, a prerequisite for most analytical work. R, while less dominant, shines in specialized statistical applications. For someone starting out, learning Python first is the consensus advice, followed by SQL for practical data handling, with R as an optional third skill depending on the focus of their career—whether it leans toward machine learning or deep statistical analysis.

Top Data Visualization Libraries and ML Frameworks

Data Visualization Libraries	% of respondents
Matplotlib	64.19
Seaborn	49.97
Plotly / Plotly Express	27.57
Ggplot / ggplot2	20.95
None	13.46
Shiny	6.39
Geoplotlib	4.96
Bokeh	4.66
D3 js	4.33
Leaflet / Folium	3.36
Other	3.11
Altair	1.68
Pygal	1.14
Highcharter	1
Dygraphs	0.88

ML Frameworks	% of respondents
Scikit-learn	57.52
TensorFlow	37.42
Keras	31.78
Xgboost	26.71
PyTorch	26.08
LightGBM	13.03
Huggingface	8.71
CatBoost	7.13
None	6.32
PyTorch Lightning	5.28
Caret	5.17
Fast.ai	3.6
Other	3.31
Tidymodels	3.28
JAX	1.07

Fig. Top Visualization Libraries and ML Frameworks

An important task in Data Science is representing information in a visual context. **How can you make it easy to understand real-time trends and business insights present in the data?**

The answer is ... **Data Visualizations!!!**

Can you believe that the human brain takes only 13 milliseconds to process an image?

Humans love stories, and visualizations allow us to create one from data. Understanding data requires the use of data visualizations, and this is because visuals are processed 60,000 times faster than text inside the human brain. Using charts or graphs to visualize vast amounts of complex information is more straightforward than digging spreadsheets or reports.

The table above at the left provides the top Data Visualization Libraries that are excellent choices for creating visually appealing and insightful data representations according to the survey respondents, with the top-end respondents mainly preferring and using the originals **Matplotlib**, **Seaborn**, and **Plotly**, with **Ggplot** for R.

Without surprising us, the top Machine Learning Frameworks are **Scikit-learn**, followed by **Tensorflow** and **Keras** which are usually used for productionizing Deep Learning Models. Both frameworks are user-friendly and they provide high-level APIs for building and training models easily.

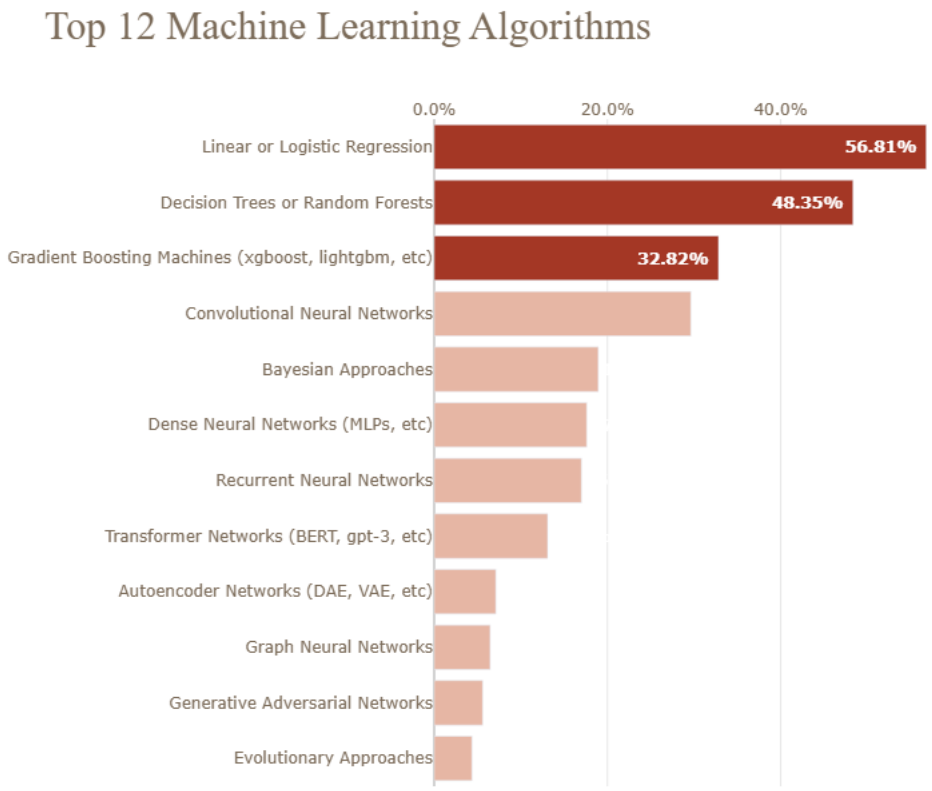

Fig. Top ML Algorithms

In terms of the top commonly used Machine Learning Algorithms we can see first in the list the **Linear or Logistic Regression**, followed by **Decision Trees or Random Forests**. That's neither a surprise for a couple of reasons:

1. These algorithms perform very well and achieve high accuracy in a variety of tasks with structured data,
2. they are easy to implement and they don't require huge hardware resources and time for training and/or inferencing.
3. Another important reason is that these Machine Learning methods offer **interpretability and explainability** that are becoming essential in solutions we build nowadays. Especially in fields such as healthcare or banking, interpretability and explainability could for example help overcome some legal constraints. **In solutions that support a human decision, it is essential to establish a trust relationship and explain the outcome and the internal mechanics of an algorithm. The whole idea behind interpretable and explainable ML is to avoid the black box effect.**

Next on the list is the **Gradient Boosting Machines** which are really powerful methods that usually achieve good accuracy, while later we can see the "Black Boxes algorithms" such as

Convolutional Neural Networks, Transformer networks, Autoencoder, etc. that perform very well when we have unstructured data, such as text and images.

The same insights are also reflected in the second plot below, where it can be seen that Linear or Logistic Regression, and Decision Trees or Random Forests are commonly used across all sectors whereas Convolutional Neural Networks are most popular in tech companies, used by the **37%** of the respondents working in the tech sector. They are also used in the Academic field where research scientists explore new algorithms for processing images, videos or text. These sectors usually don't lack in training resources and interpretability is not a must-have.

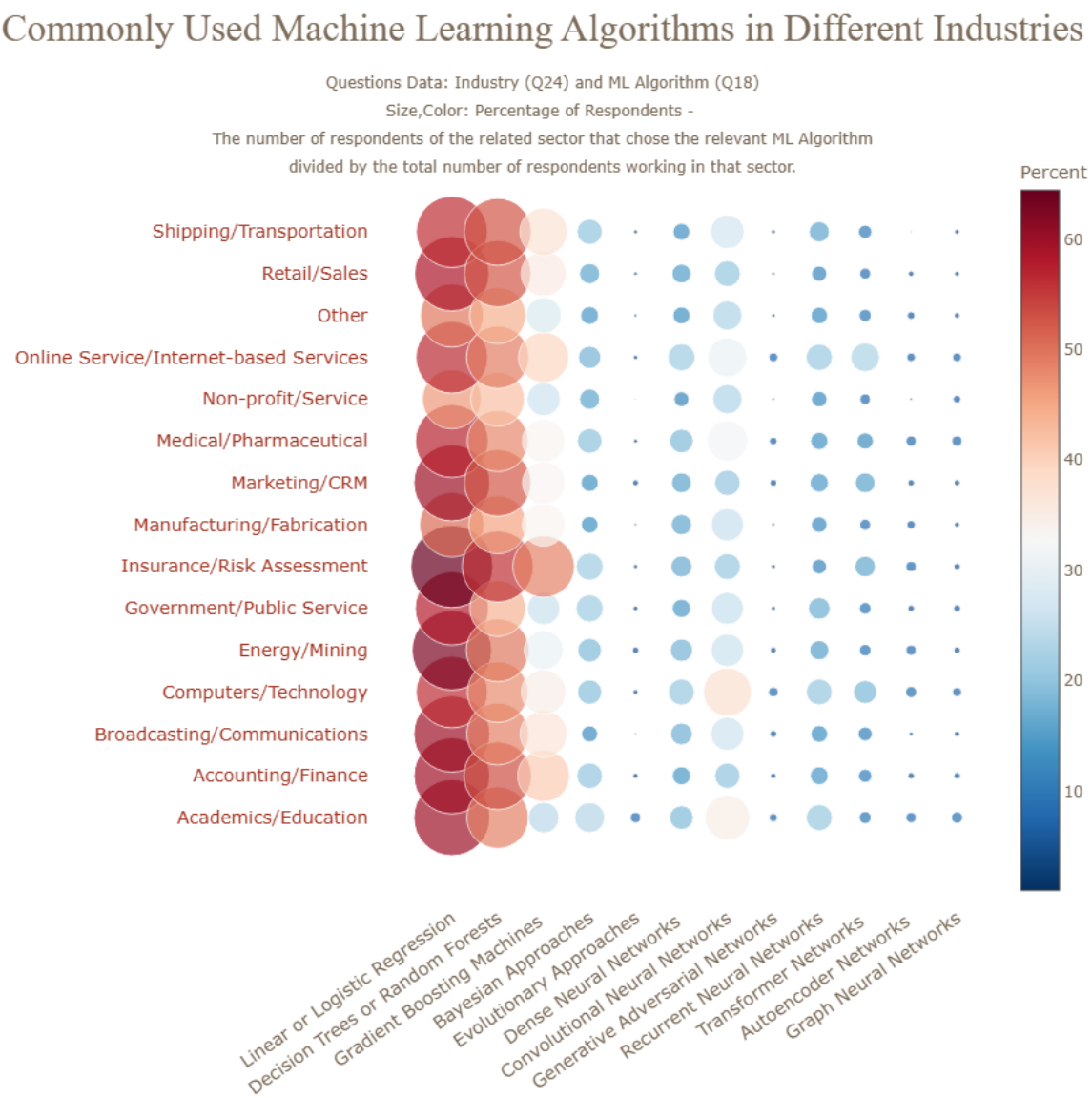

Fig. Common ML Algorithms in Industries

o Transfer learning in the business world

Transfer learning is quite popular nowadays and it aims to save time and effort and provides the advantage of using tested models. This way, companies cut costs by avoiding the need for a high-cost GPU for retraining the model. The goal is to make machine learning as human as possible. Transfer learning is mostly used in **computer vision and natural language processing tasks** due to the huge amount of computational power required.

The following charts represent the percentage of respondents that use pre-trained models, specified below, for Computer Vision and NLP respectively on a regular basis.

It is clear that a higher percentage of respondents use pre-trained image classification models rather than transformer language models which is kinda expected due to **"ImageNet moment"**.

Pretraining entire models to learn both low and high-level features has been practiced for years by the computer vision (CV) community. Most often, this is done by learning to classify images on the large ImageNet dataset. ULMFiT, ELMo, and the BERT model have the last years brought the NLP community an "ImageNet for language"---that is, a task that enables models to learn higher-level nuances of language, similarly to how ImageNet has enabled the training of CV models that learn general-purpose features of images. So, I expect the next years to see also a bigger percentage of professionals in AI use pre-trained models for NLP tasks.

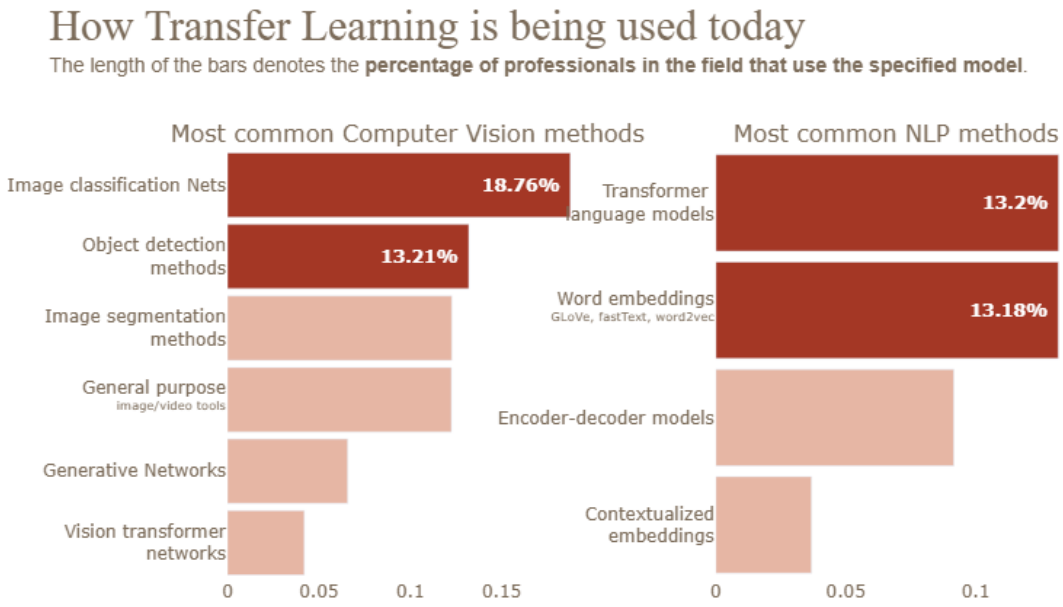

Fig. Usage of Transfer Learning

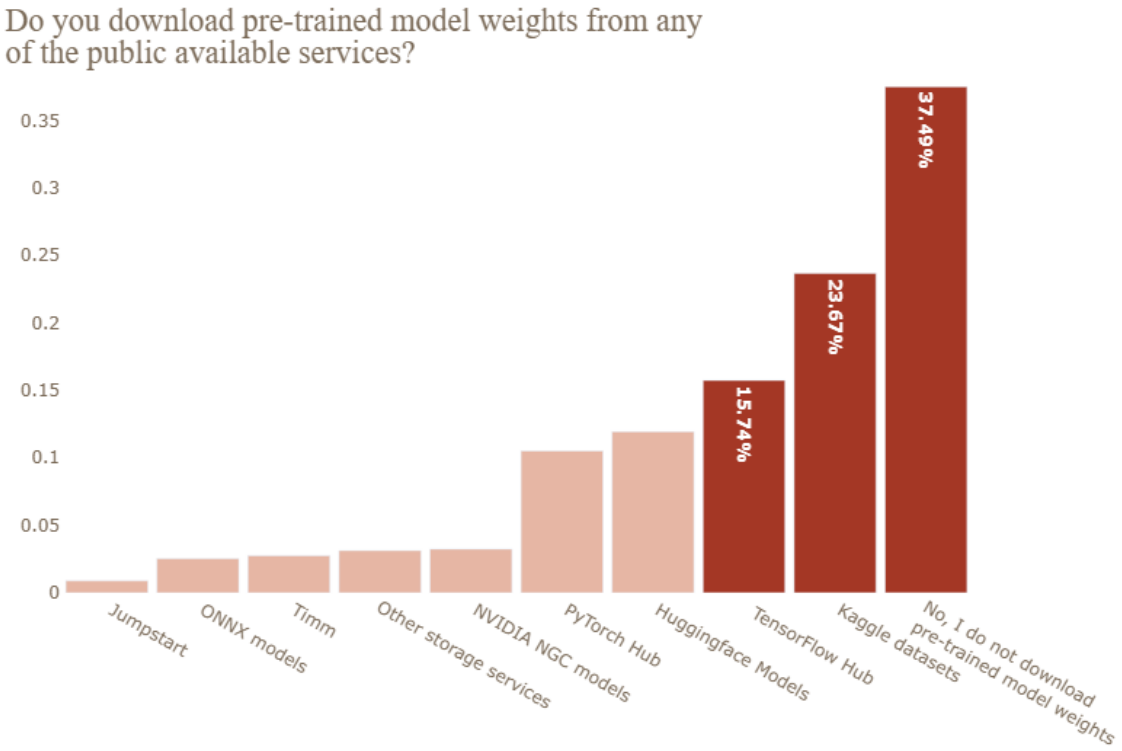

Fig. Downloads of Pre-trained model Weights

NLP Users

In the tables below, we can then see the number of professionals that use pre-trained models and methods for NLP / CV tasks on a regular basis along with the relative percentages. The percentages column has been calculated by dividing the number of professionals in each role that use CV/NLP methods by the total number of respondents that have this job role. The key takeaway is that CV / NLP methods and pre-trained models are used mostly by **Machine Learning Engineers**, **Data Scientists**, **Data Architects**, **Developer Advocate**, and **Research Scientists**.

	Use of NLP Methods and Pre-trained Models	Nbr of respondents	%
0	No	7399	81.360000
1	Yes	1695	18.640000

These roles show the highest adoption rates, as indicated by the calculated percentages, which reflect the proportion of professionals in each role leveraging these advanced techniques. This trend underscores the importance of pre-trained models in streamlining complex tasks like natural language processing and computer vision, particularly in roles that focus on innovation, development, and research. The widespread use among these professionals suggests a growing reliance on pre-trained models to enhance efficiency and accuracy in AI-driven projects.

	Role	Nbr of respondents	%
0	Machine Learning/ MLops Engineer	251	44.660000
1	Data Scientist	582	30.420000
2	Developer Advocate	17	28.810000
3	Research Scientist	143	24.240000
4	Data Architect	20	21.050000
5	Data Engineer	57	16.720000
6	Software Engineer	157	16.170000
7	Manager (Program, Project, Operations, Executive-level, etc)	132	15.980000
8	Teacher / professor	120	14.630000
9	Statistician	12	9.760000
10	Data Analyst (Business, Marketing, Financial, Quantitative, etc)	116	7.670000
11	Data Administrator	5	7.140000
12	Engineer (non-software)	32	6.910000
13	Other	51	6.820000

Fig. Different Roles Response by use of NLP models and Pre-trained models

	Use of CV Methods and Pre-trained Models	Nbr of respondents	%
0	No	6705	73.730000
1	Yes	2389	26.270000

	Role	Nbr of respondents	%
0	Machine Learning/ MLops Engineer	251	44.660000
1	Data Scientist	582	30.420000
2	Developer Advocate	17	28.810000
3	Research Scientist	143	24.240000
4	Data Architect	20	21.050000
5	Data Engineer	57	16.720000
6	Software Engineer	157	16.170000
7	Manager (Program, Project, Operations, Executive-level, etc)	132	15.980000
8	Teacher / professor	120	14.630000
9	Statistician	12	9.760000
10	Data Analyst (Business, Marketing, Financial, Quantitative, etc)	116	7.670000
11	Data Administrator	5	7.140000
12	Engineer (non-software)	32	6.910000
13	Other	51	6.820000

Fig. Different Roles Response by use of CV methods and Pre-trained models

o Usage of specialized hardware for ML models training

There are broadly 2 stages to a Machine Learning project. The first stage is ML **Model Training** and the second stage is the **Model Inference**.

Training an ML model requires more computational power and resource. Especially when working with Neural Networks, it is essential to process huge amounts of data to train the model. This process usually involves some heavy matrix calculations. GPUs are a specialized hardware used for Machine Learning because they can perform multiple, simultaneous computations. This enables the distribution of training processes and can significantly speed up machine learning operations. With GPUs, we can accumulate many cores that use fewer resources without sacrificing efficiency or power. However, GPU is not the only specialized hardware that is used for ML. There are also other types of specialized hardware as we'll see below, but the GPU is the one that is used most commonly.

So, when designing our deep learning architecture we have to consider multiple factors for our decision to use GPUs or any other specialized hardware or not (dataset size, model size, etc.). **As the survey data shows only 31% of the respondents use specialized hardware like GPU for ML model training.**

	Specialized Hardware Usage	Nbr of respondents	%
0	No	6263	68.870000
1	Yes	2831	31.130000

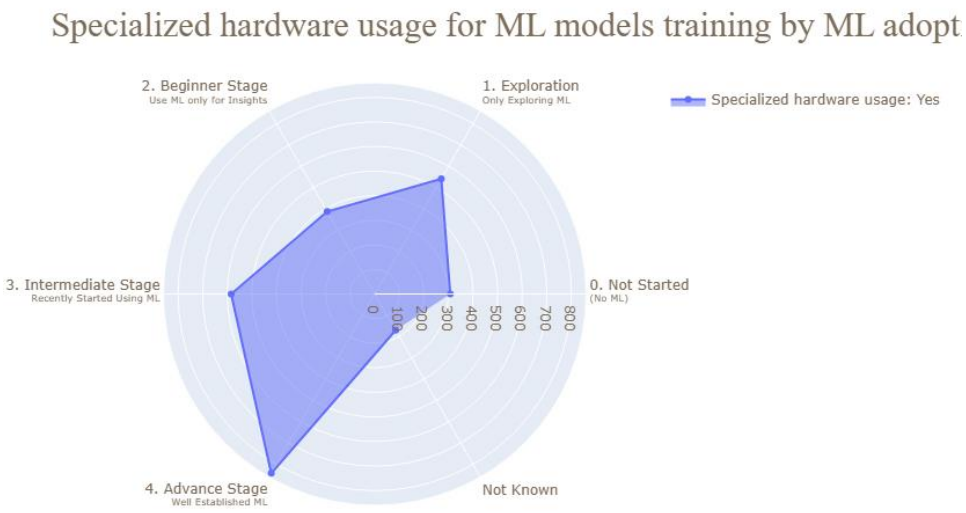

Fig. Special Hardware usage for ML models training

Companies with Machine Learning Models in production either in an advanced or intermediate stage are more likely than the ones that started recently exploring ML capabilities to use GPUs for training their ML Models as it can be seen in the illustration above.

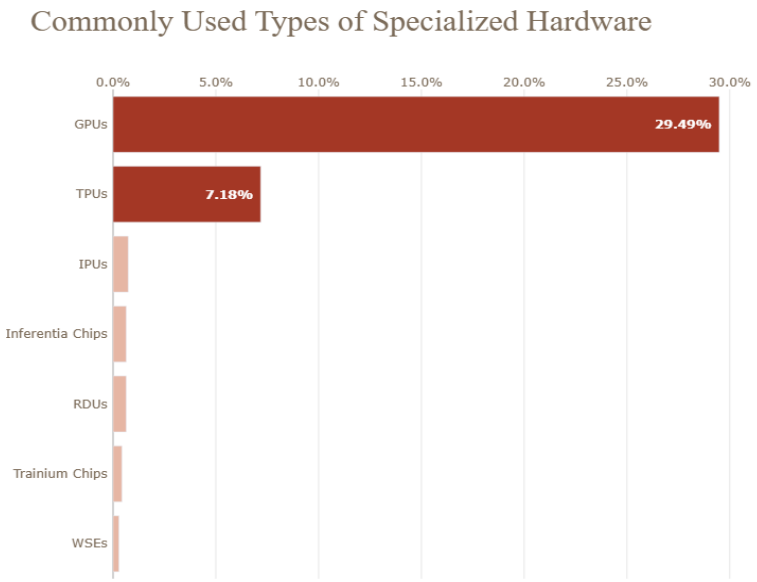

Fig. Common Hardware used for ML

Specialized Hardware Userss

The table below shows the number of professionals that use specialized hardware for ML model training. The percentages column has been calculated by dividing the number of professionals in each role that use GPUs or TPUs, etc. by the total number of respondents that have the same job role.

	Role	Nbr of respondents	%
0	Machine Learning/ MLops Engineer	352	62.630000
1	Data Scientist	811	42.390000
2	Research Scientist	242	41.020000
3	Data Engineer	119	34.900000
4	Data Architect	33	34.740000
5	Manager (Program, Project, Operations, Executive-level, etc)	277	33.540000
6	Software Engineer	291	29.970000
7	Developer Advocate	16	27.120000
8	Teacher / professor	203	24.760000
9	Engineer (non-software)	85	18.360000
10	Data Analyst (Business, Marketing, Financial, Quantitative, etc)	264	17.450000
11	Other	112	14.970000
12	Data Administrator	10	14.290000
13	Statistician	16	13.010000

Fig. Responses by Specialized Hardware Users

3. AI job roles and key skills needed to build a career in AI

Whether the insights from the 2022 Kaggle Machine Learning & Data Science Survey illustrated in this notebook so far or the progress Artificial Intelligence and Machine Learning has made today excite you to get into the AI and Data Science world and build a career in AI, this section is the right place for you. In this part, I'll provide some insights about the different job roles and the top skills required, based on the responses of the professionals who participated in the survey.

As it has been seen in the above sections, a lot of companies across different industries are adopting AI solutions. Enterprises have also recognized the benefits of having an in-house team for data analytics. This has led to the rise of AI-related jobs. However, the different titles present in the market may confuse a newcomer. Different titles also require different specializations, which makes it difficult for an aspirant to choose the role they are equipped for and interested in.

o AI jobs description: roles, responsibilities and skills required

So, let's have first a look at the most in-demand AI jobs according to the survey respondents that already have a job position related to AI.

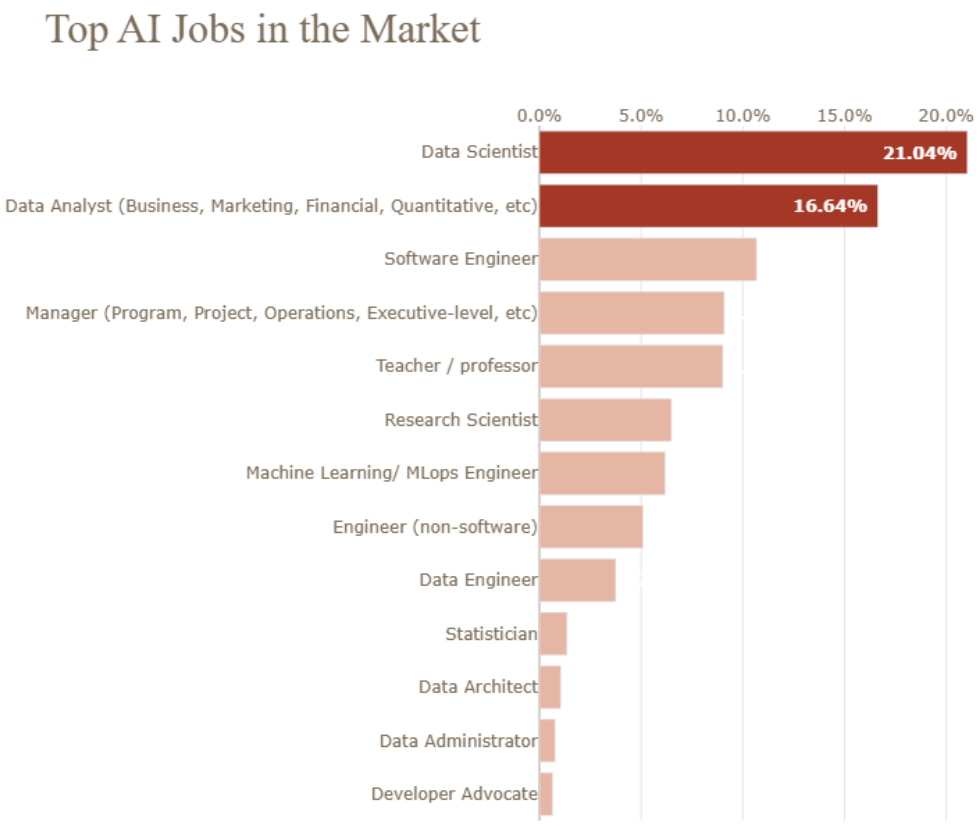

Fig. Top AI Jobs

Unsurprisingly, the **Data Scientists** ranked first in the chart with the most common data-related jobs. With 1,913 respondents they form 21.04% of our data professionals (9,094 in total), considerably ahead of **Data Analysts** in second place with 16.64%, followed by **Software Engineers** with 10.68%.

But what industries are actually hiring AI specialists and what AI roles do they seek??

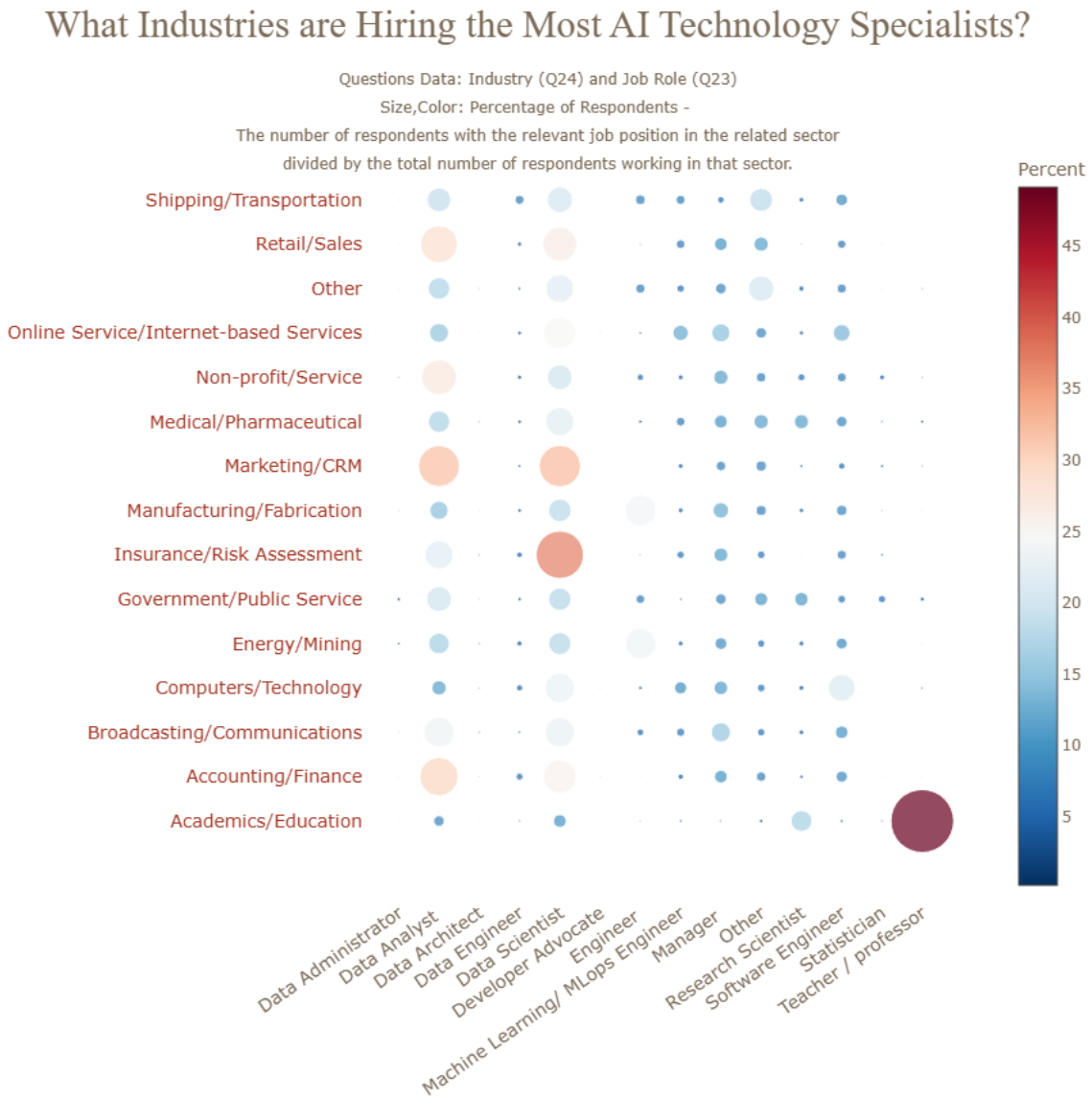

Fig. Industries Hiring AI Specialists

The scatter plot shows that 37.10% of employees in **Insurance companies** are **Data Scientists**, making them top the list of industries hiring Data Scientists. Data science can enable insurers to develop effective strategies to acquire new customers, develop personalized products, analyze risks, assist underwriters, implement fraud detection systems, and much more.

Second in the list with the sectors that occupy the most data scientists proportionally with the total number of respondents working in that sector is the **Marketing** and **CRM** companies, followed by the **Retail/Sales** field and the companies offering **Internet-based services**. A wider range of information is available to these companies, therefore Data science helps them to put these data to efficient use to drive more business and refine their products/services offerings. These sectors as it can be seen also seek Data Analysts.

Now, let's focus on the **Data Scientists** and **Data Analysts** since they are the most popular job roles as well as on the **Machine Learning Engineers** and **Research Scientists** who are core components of the AI & Data Science teams, and see how a typical day at work looks like. Let's see the main tasks and the responsibilities that they have.

Note: In order to create the following chart, for each activity, I counted the number of respondents (Data Scientists, Analysts, ML engineers) who chose it and I calculated the percentages of each activity that you see below based on their total sum.

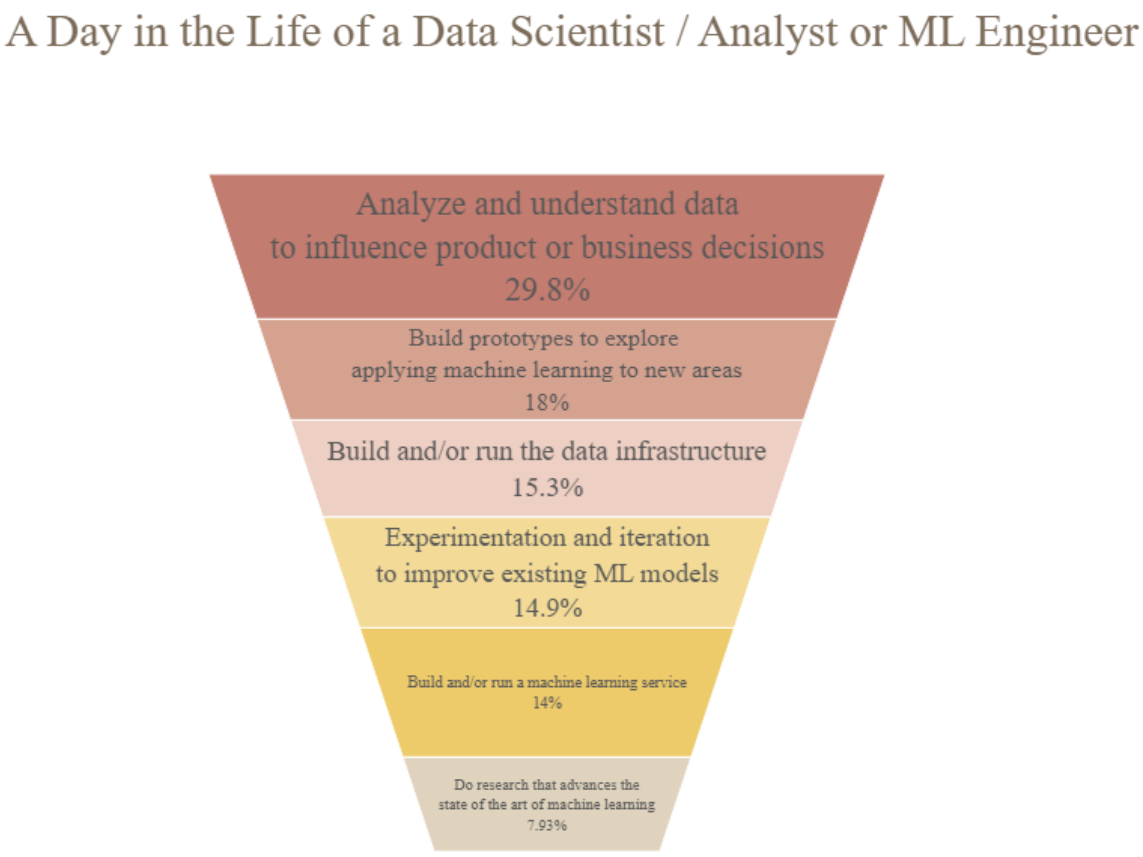

Fig. Lifestyle of Data Scientist/ ML Engineer

The top level of the reversed pyramid represents the most common activity whereas going down we see the tasks, implemented less commonly. In addition to that, you can also see the most relevant activities per role in the illustrations below.

Key insights:

- So, **29.8%** of the total activities that the respondents do is **Analyze and understand data to influence product or business decisions.** Data analysis dominates Data Scientists and Data Analysts' activities as is also illustrated in the following visualizations. The main task of those two roles is to analyze data to identify patterns and trends and extracts actionable insights for driving business decisions.
- The second most common activity is to **implement Machine Learning methods to explore new areas**. In this task Machine Learning Engineers, Data Scientists, and Research Scientists are mainly involved.
- In the third and fourth positions are the **Experimentation and iteration to improve existing ML models** and **Build a machine learning service**. Perhaps is not a surprise that Machine Learning Engineers are mainly responsible for these activities.
- One less common activity is to **Build and run data infrastructure** where all 4 roles contribute almost equally.
- Last but not least, is to **Do research that advances the state of the art of machine learning** which as it's expected undertaken mostly by Research Scientists.

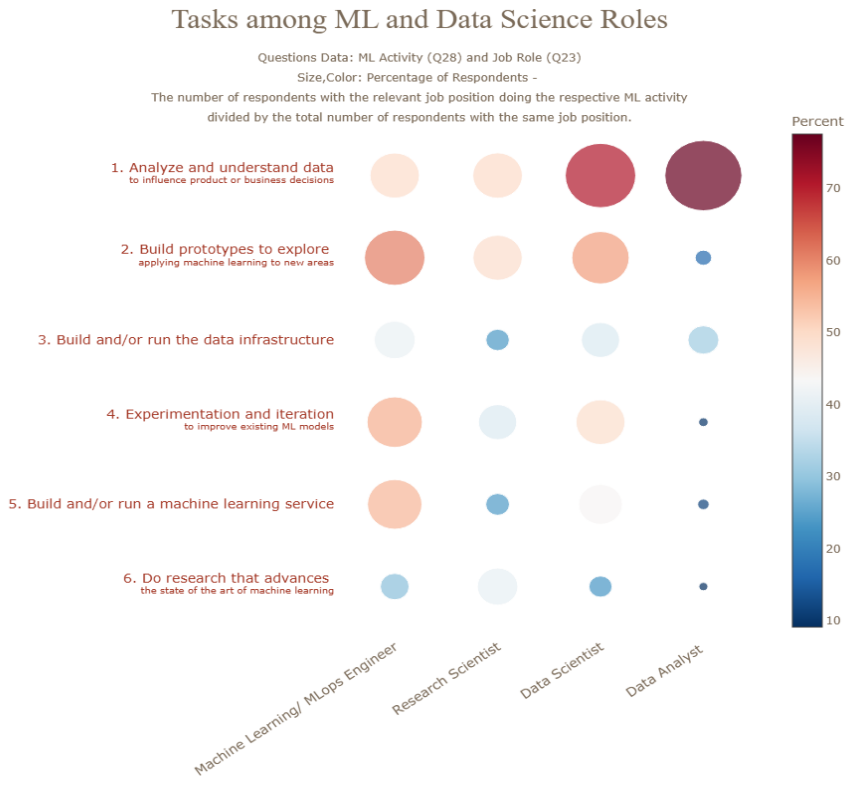

Fig. Task of ML and Ds Roles

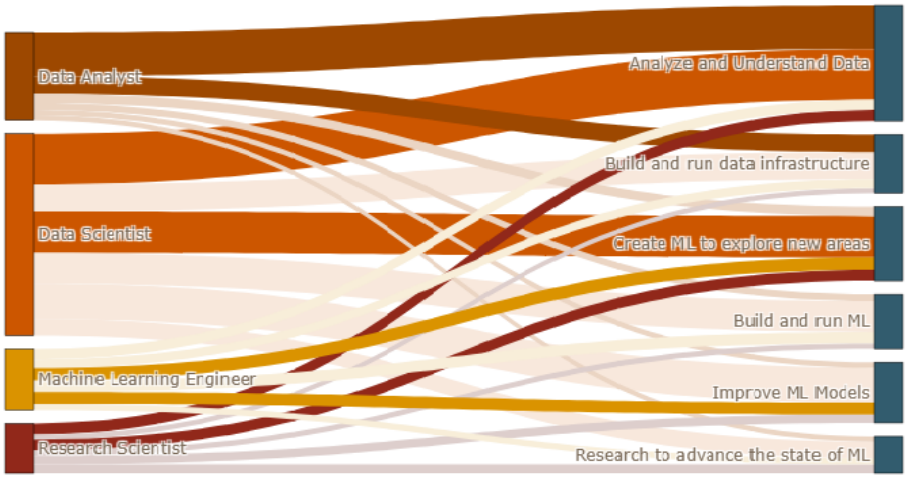

Fig. Responsibilities of different roles

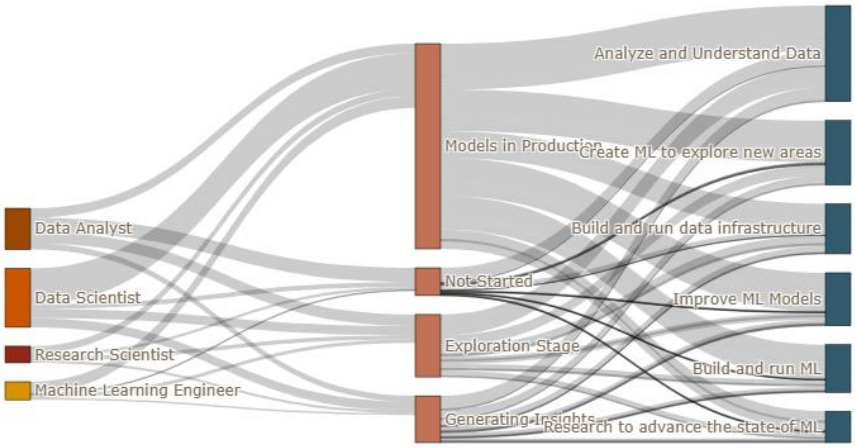

Fig. Responsibilities of different roles by its Subsections

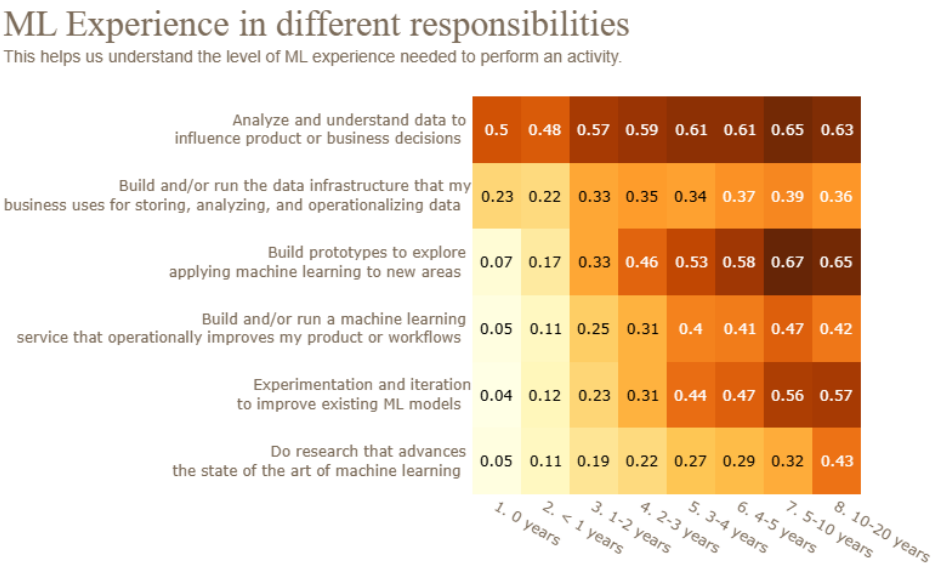

Fig. ML Experience in different responsibilities

The chart above shows the percentage of respondents at a particular Machine Learning experience level for each responsibility. This helps us understand the level of ML expertise needed to perform a task.
The main key takeaways are:

- Data Analysis activities show higher percentages of individuals with ML experience of 2-3 years or more.
- Machine learning-related tasks such as Applying ML methods to new areas and improving existing ML models have greater percentages at the higher experience ranges.

Below you can see the distribution of the years of coding experience and experience using ML methods. While a big group of respondents has many years of coding they don't have many years experience in using Machine Learning methods.

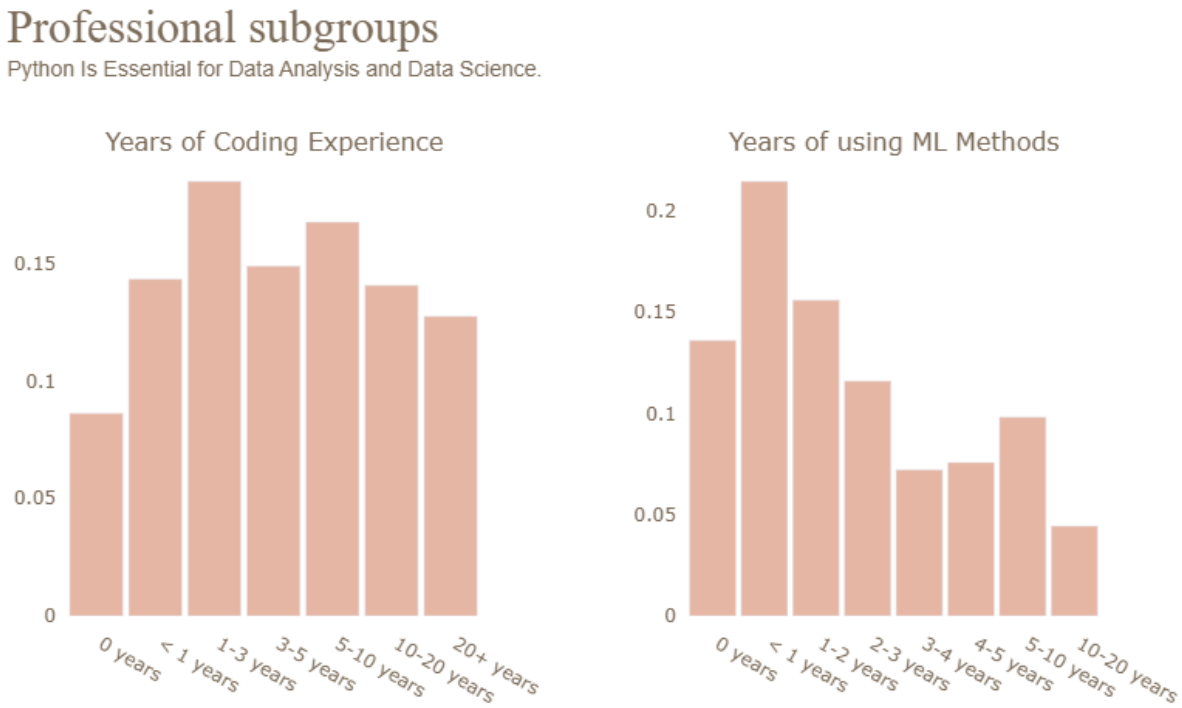

Fig. Coding Experience and use of ML methods

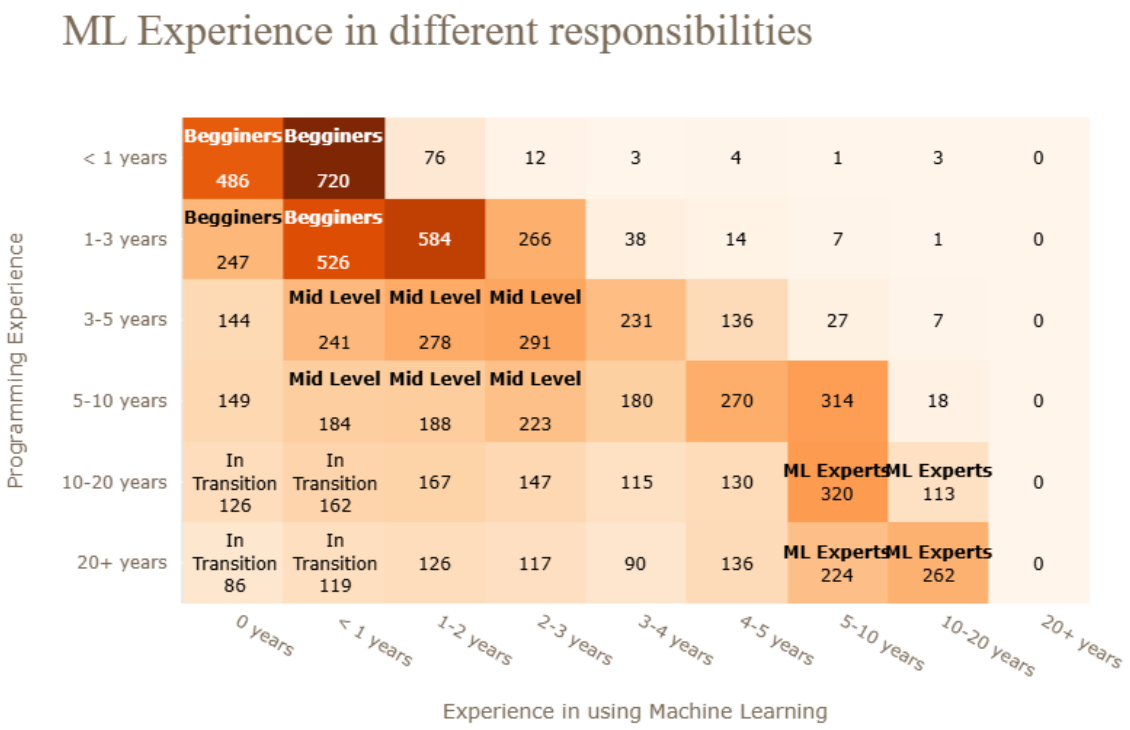

Fig. ML Experience in different responsibilities

In the figure above we can also see a categorization of the professionals:

- The first group is the **Beginners - Juniors**. They have less than 3 years of experience in both coding and ML methods and they make up around **21.8%** of all the professionals who participated in the survey.
- The second group are **Coders in transition** (**5.4%**). Those people have decades-long coding experience for working with data, however, they have started working with machine learning only recently. These may be for example software engineers transitioning into data engineers or Machine Learning Engineers.
- The third category in the lower right corner is the **Machine Learning Experts** (**~10%**). Those people have been coding since long before the current AI revolution - with 10 or even over 20 years of both ML and coding experience, they may have started to specialize in the topic around the 2000s or even late 1990s. These people were doing machine learning before it was hype.
- The last group is the **Mid Level Data Scientists or ML Engineers** (**~15.4%**) with a solid understanding of ML concepts and a strong coding background.

So, to help you get your dream job in the AI and Data Science field, especially if you belong to the Beginners or Coders in Transition group I analyze below the top skills required for working with data and Machine Learning.

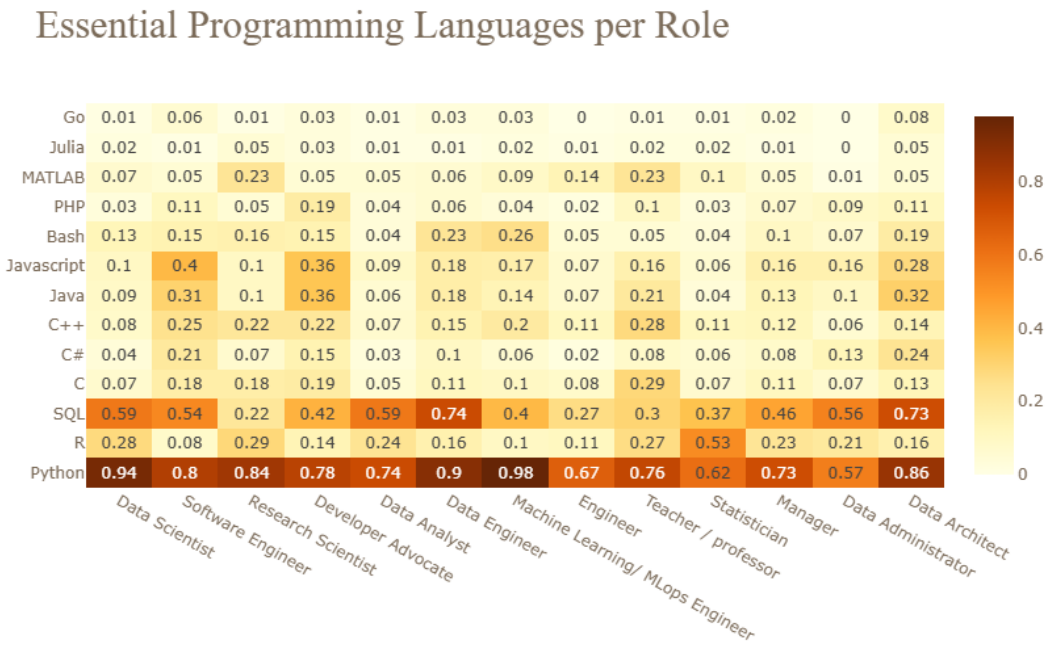

Fig. Essential programming languages per role

Regarding the most important programming language that you need to know, it's pretty obvious that is Python. You can see in the table above that **Python** is required for each role, along with **SQL** most of the time. Statisticians should also have R knowledge while Software Engineers and Developers might also work with Java and Javascript.

If you are thinking to become a Machine Learning Engineer, a Data Architect, or a Data Scientist then it would be beneficial to get familiarized with Cloud technologies since these roles require working with cloud computing platforms and other cloud services.

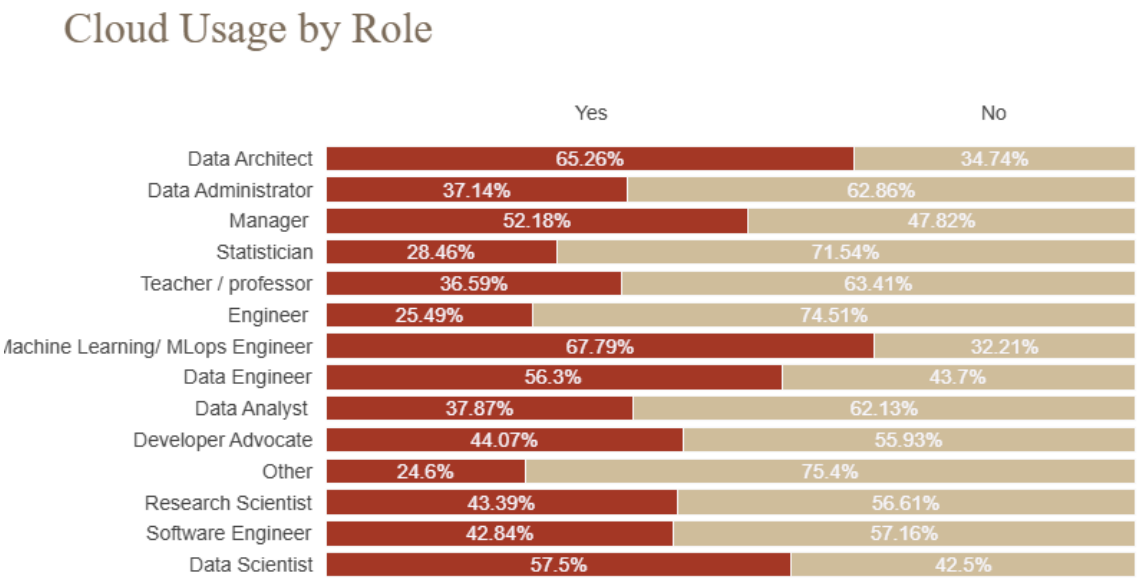

Fig. Use of cloud by Role

Since machine learning and AI jobs entail the development of algorithms, let's have a look at the ML algorithms that an aspiring professional should know. The ones that are common for every role but especially for Data Scientists are **Linear or Logistic Regression** and **Decision Trees or Random Forests.** Data Scientists should also be able to use **Gradient Boosting Machines algorithms** while **Research Scientists** and **Machine Learning Engineers** should have a solid understanding of Deep Neural Networks since they use **Convolutional Neural Networks, MLPs, RNNs, and Transformers** on a regular basis.

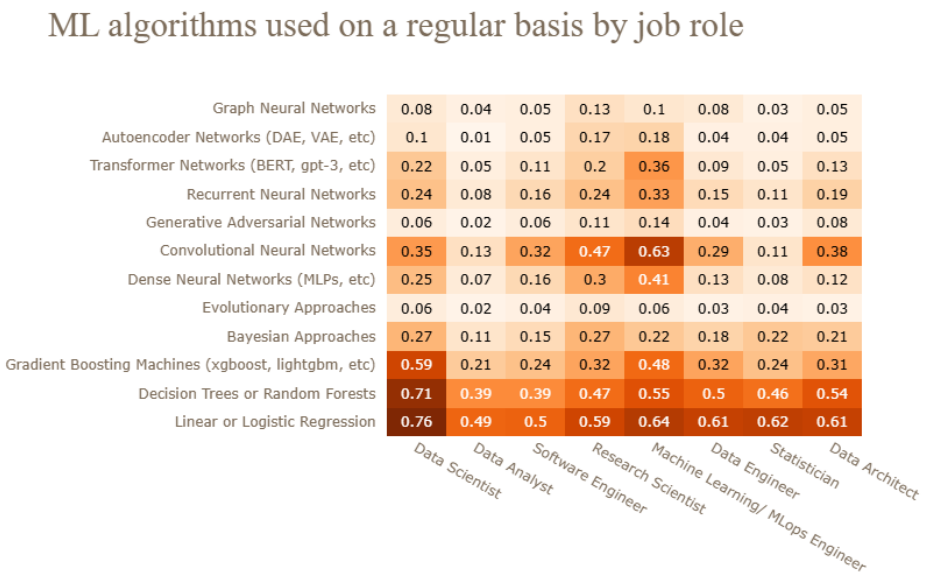

Fig. Regularly used ML Algorithms

When it comes to the Machine Learning Frameworks **Scikit-learn** is a must-have for Data Scientists and Machine Learning Engineers while **PyTorch**, **Tensorflow**, and **Keras** are used a lot by Machine Learning Engineers, Research Scientists, Data Architects, and Data Scientists for research and production needs.

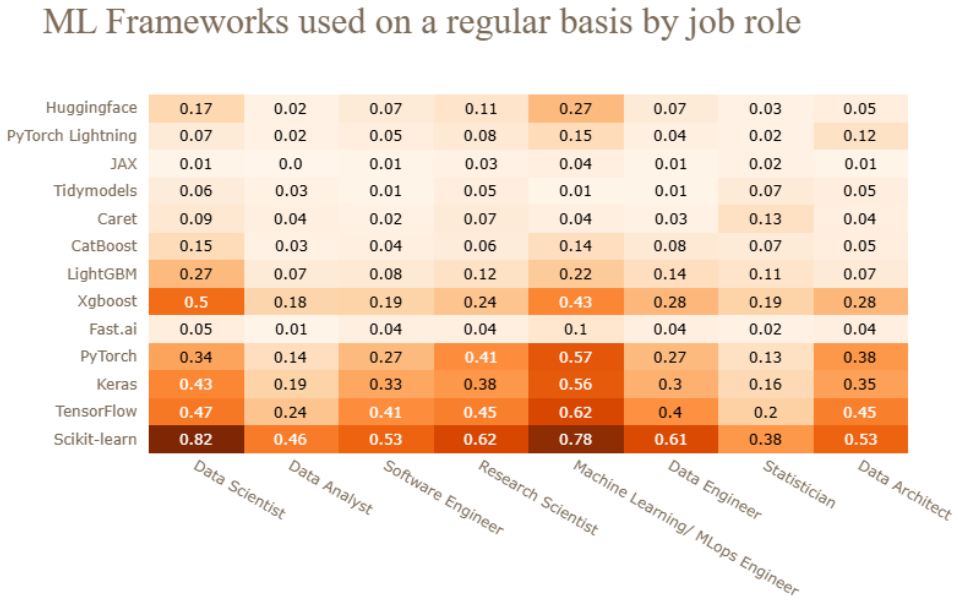

Fig. Regularly used ML Frameworks

o Data science team sizing

Here I look at the relationship between company and Data Science team size. It seems that larger companies have bigger data science teams.

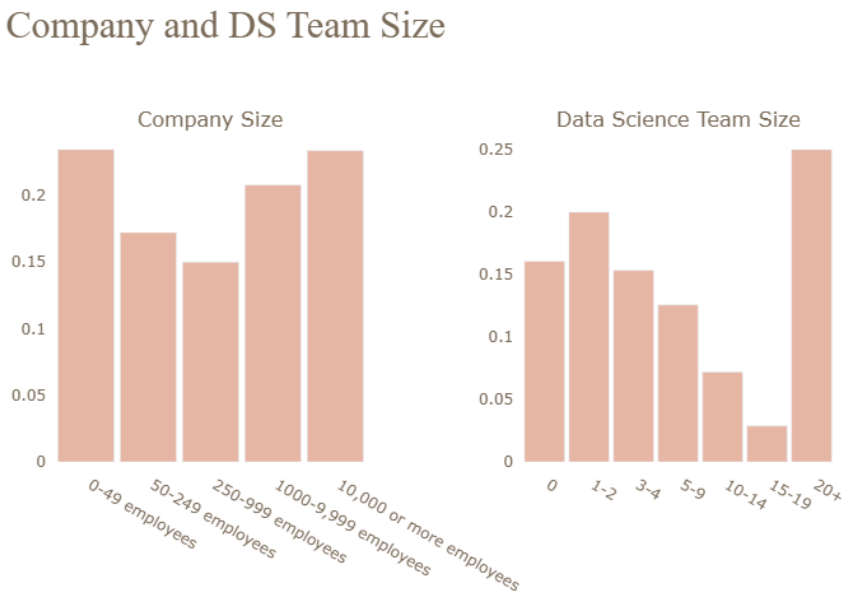

Fig. Company and DS Team Size

Fig. Company Size

From the illustration above we can notice that there is a correlation between the company's size and the Data Science team's size. Smaller companies have mostly Data Science Teams of 1-2 individuals while the larger ones have a much bigger team of 20+ members meaning that each member will have concrete responsibilities and tasks.

o What education do AI specialists need?

Education requirements for data science and machine learning professionals vary by position, employer, and industry. Some data science professionals hold a mix of education levels. For example, someone might earn a bachelor's in computer science and complete a data science bootcamp. Or, they might complete a bachelor's in an unrelated field and then earn a master's in data science.

Let's have a look at the highest level of education that the professionals of the Kaggle Survey have. Almost half of them (**43.51%**) hold a Master's degree while 24.76% have a Bachelor's degree. So, from my point of view, the Master's degree tends to be a must-have for the market.

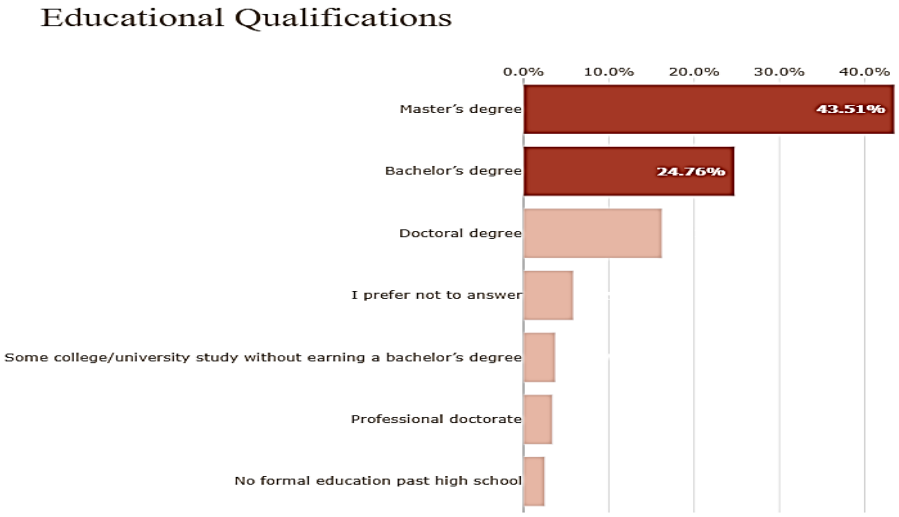

Fig. Educational Qualification

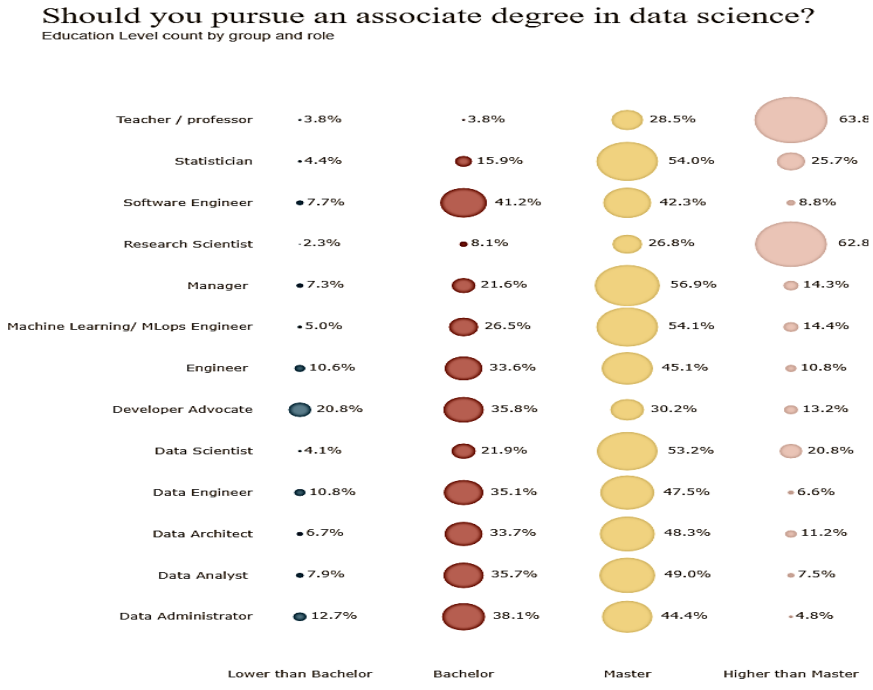

Fig. Response to should pursue an Associate Degree in Ds

Data scientists typically need at least a bachelor's degree in computer science, data science, or a related field. However, many employers in this field prefer a master's degree in data science or a related discipline.
Data analysts and data engineers usually need a bachelor's degree. Becoming a data scientist or computer and information research scientist usually requires a master's.

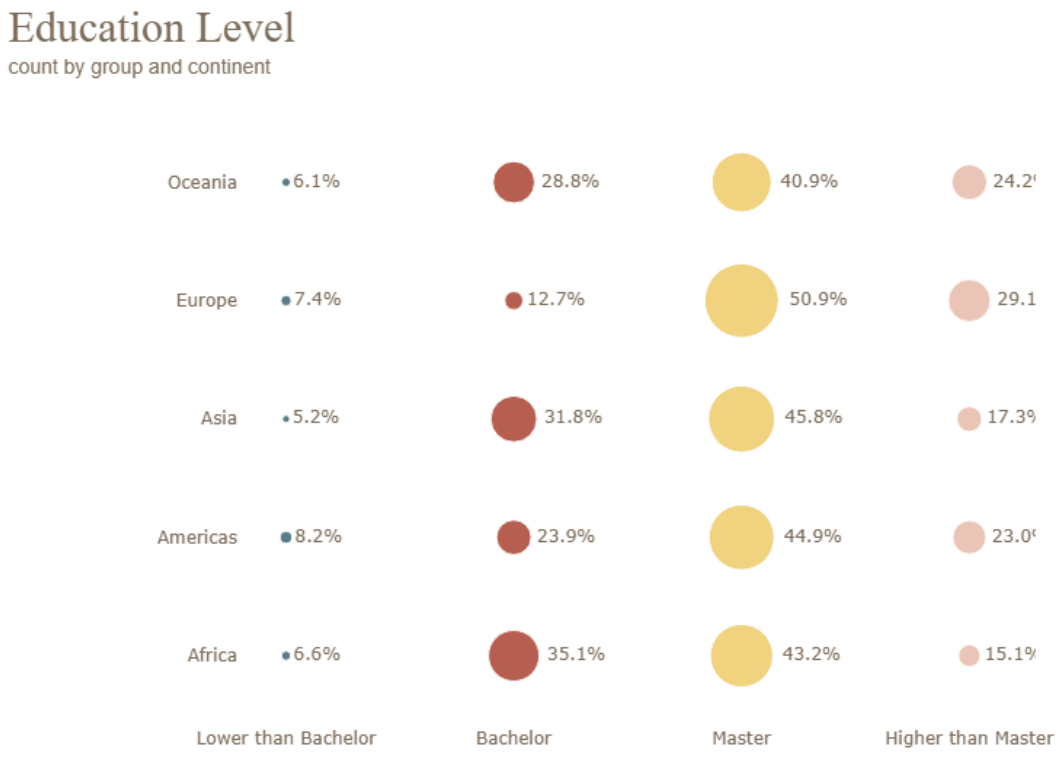

Fig. Education Levels by Country

4. Artificial Intelligence salaries (by role, industry, education & more)

I hope the last part of the analysis to help you in your salary negotiations or when negotiating a job offer.

So, the $100 Dollar Question: How Much Do Artificial Intelligence (AI) and Data Jobs Actually Pay?

Well, the exact numbers of AI salaries depend on many factors, including specific job responsibilities, industry, experience, education level, and geographic location.

Therefore, for the salary benchmarking I'll get each factor separately and do a salary comparison based on that. We would get more representative insights if I would take into account all of them at once, or jointly, for instance examine salaries based on industry and job roles, or based on country, industry, and job roles. However, I want to keep the analysis simple so let's do the deep dives by exploring each factor separately.

Starting with the analysis of the yearly compensation by job role, it is clear that the 1st best-paying salary is for **Data Architects (median at 65,000 US dollars per year)**, followed by **Managers (median at 55,000 US dollars per year)** and **Data Scientists**, earning slightly less **(median at 45,000 US dollars per year)** while **Statisticians** are paid less than any other profession.

Disclaimer: The exact numbers of the salaries might be not fully accurate because we have to take into consideration all the factors mentioned at the beginning of the section for the salary benchmarking instead of examining them one by one. But we can get an overview of the market trends in 2022.

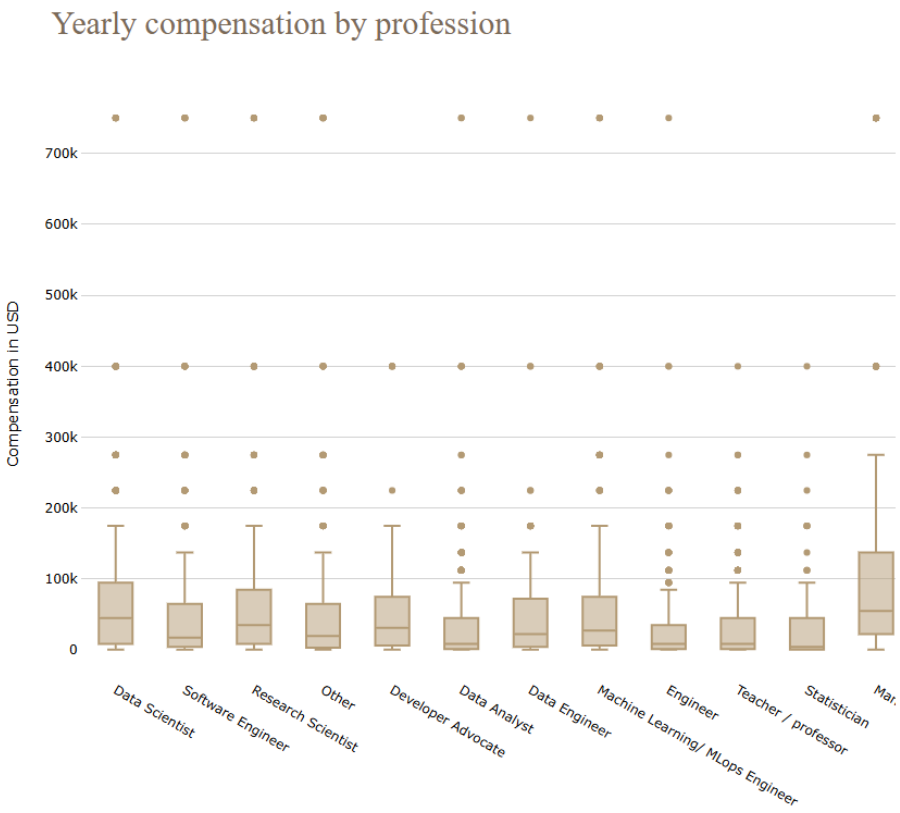

Fig. Yearly Compensation by Profession

Moving on to the comparison by industry in the first place as it can be seen in the chart are the **Medical / Pharmaceutical** and **Insurance companies**, offering 45,000 US dollars yearly compensation on average.

Even if the numbers are not accurate, the trends though look reasonable. The Pharmaceutical and Health Sciences sector played a key role during the COVID-19 pandemic. To deal with the global crisis, traditional competitors teamed up to accelerate research, and this “new normal” mindset triggered organizations to rethink their operational models.

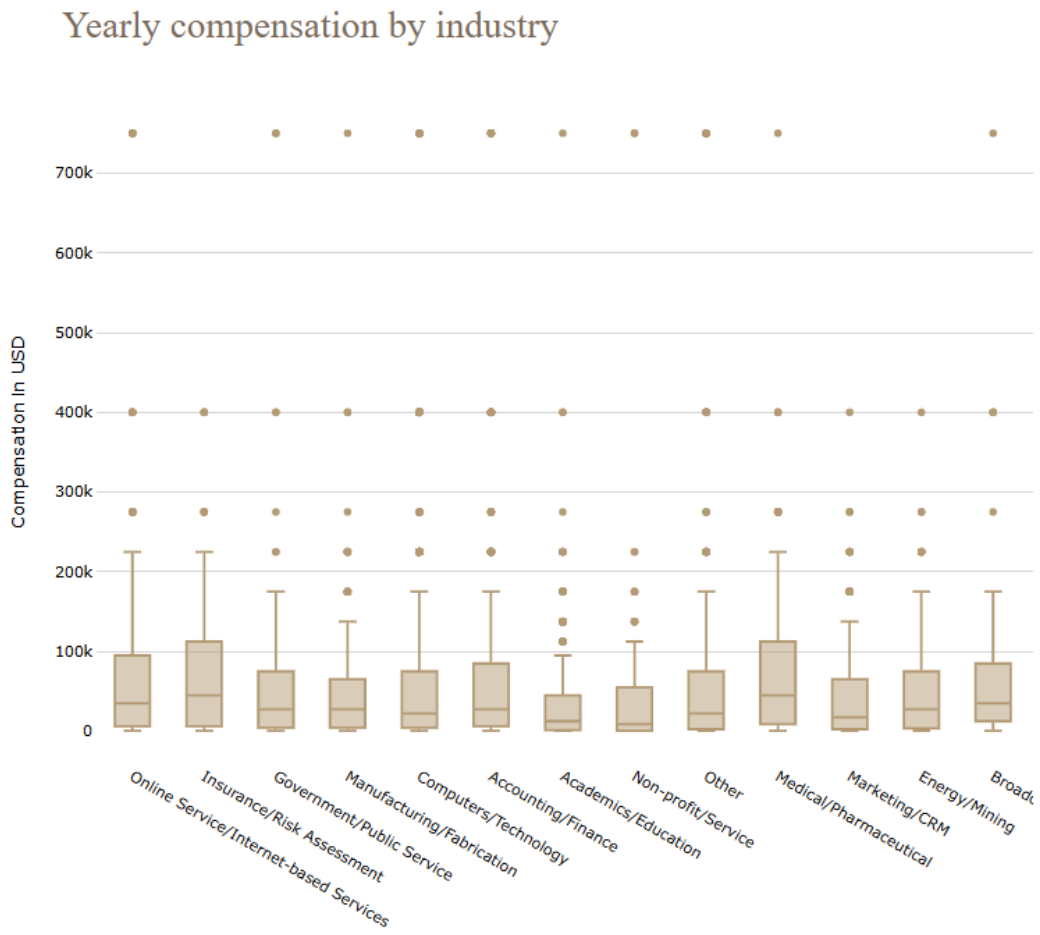

Fig. Yearly compensation by industry

As you might expect there's a clear correlation between education level and salary. Generally, it seems that the more educated you are, the greater your salary becomes.

The same applies to years of coding experience or ML experience.

The below visual shows that the yearly compensation increases with higher levels of education. Individuals with a Master's degree or higher tend to have higher median compensation compared to those with a Bachelor's degree or lower. The box plots indicate a wider range of compensation for those with higher education levels, suggesting greater variability in earnings. Additionally, there are outliers at the higher end of the compensation spectrum for all education levels, but these are more pronounced for those with advanced degrees.

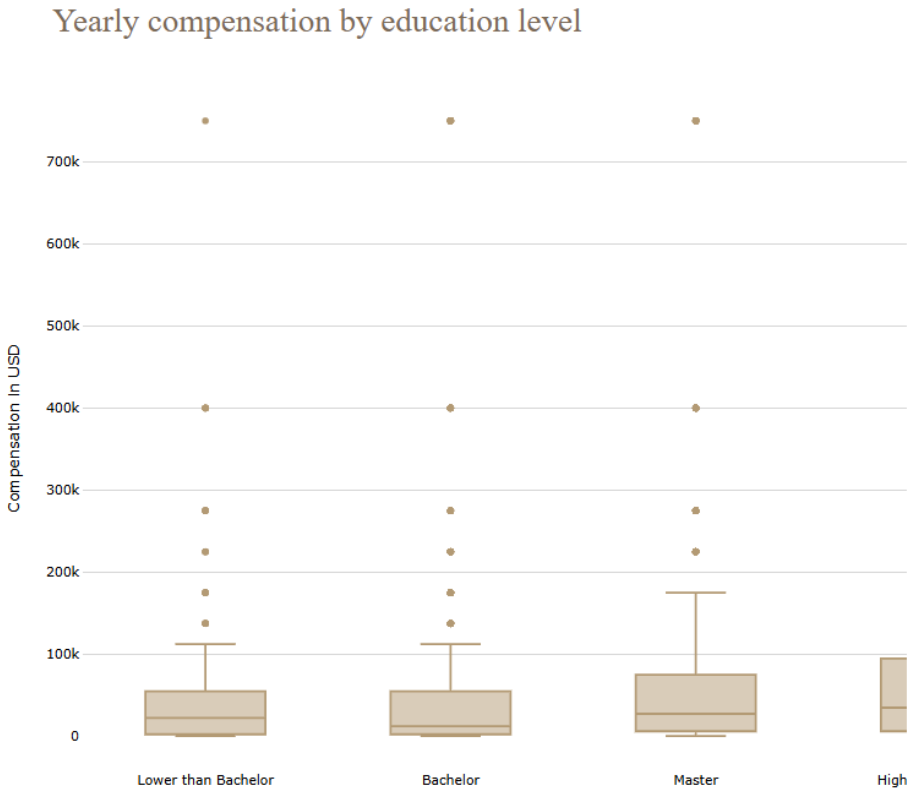

Fig. Yearly compensation by Education level

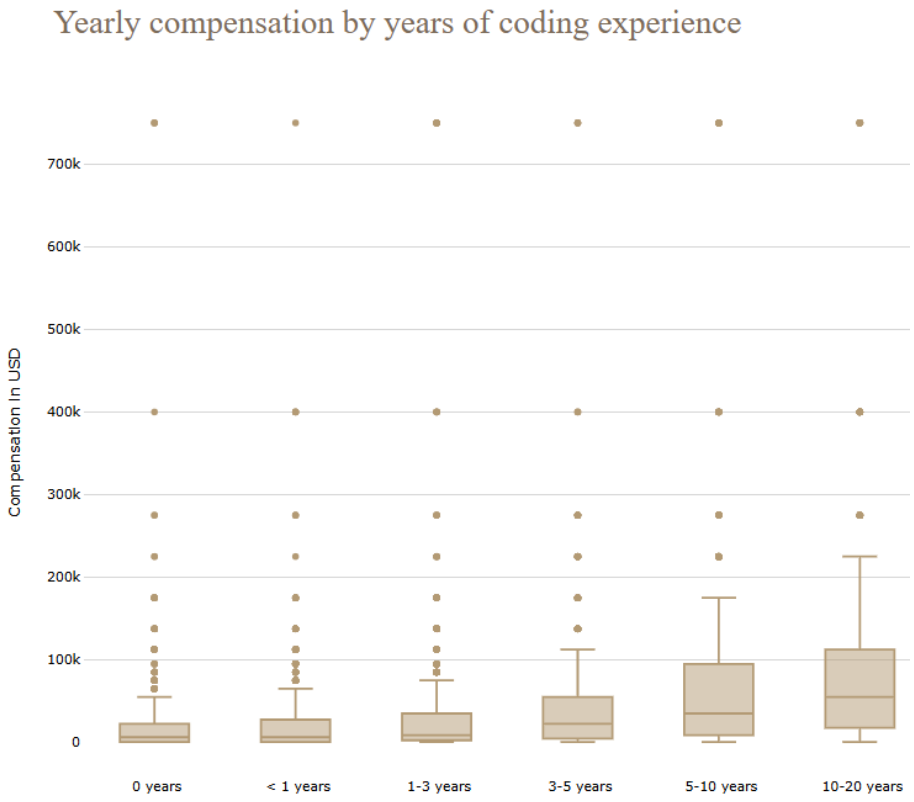

Fig. Yearly compensation by coding experience

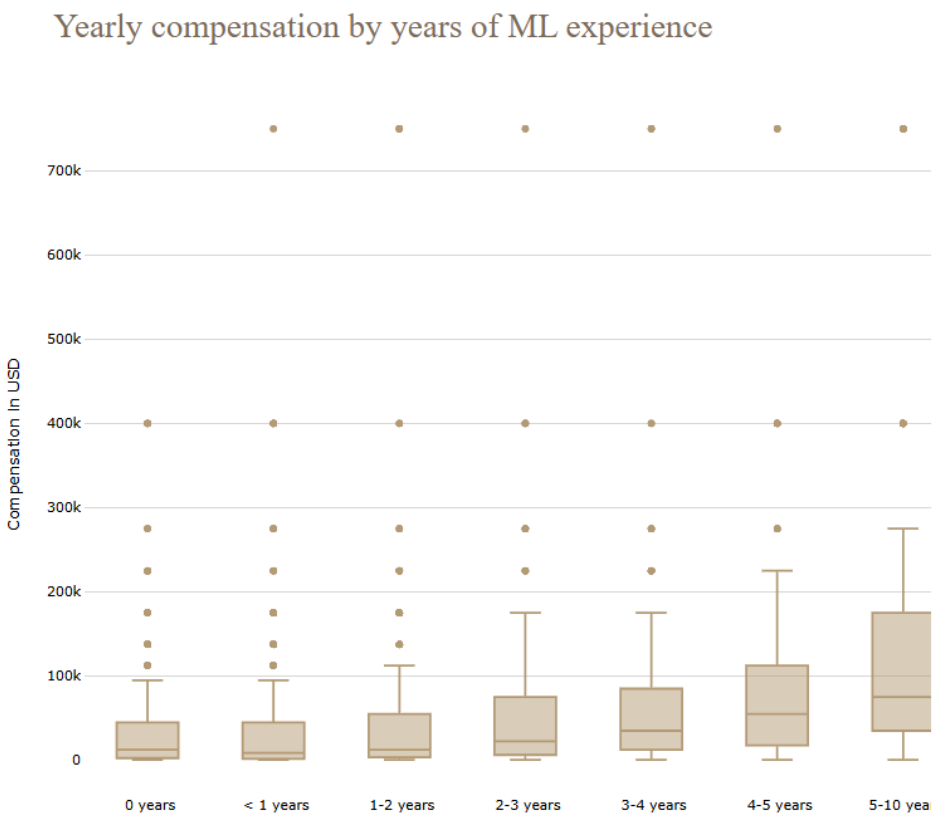

Fig. Yearly compensation by ML experience

In terms of continent, it seems that the **Americas and Oceania pay higher salaries for AI jobs compared to Europe, Asia, and Africa.**

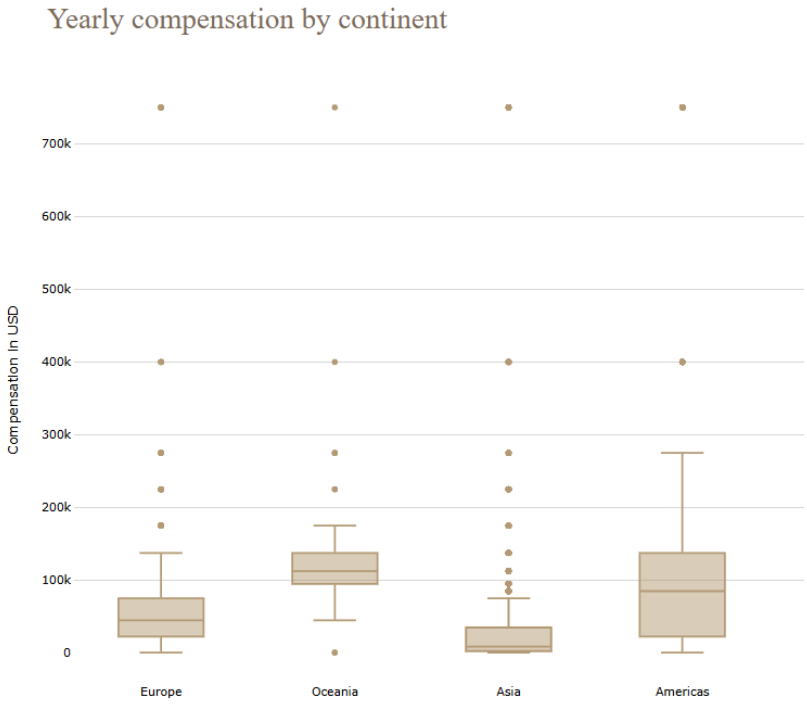

Fig. Yearly compensation by continent

Another clear trend is that large companies pay higher wages. One explanation could be that workers in big firms are more skilled.

Fig. Yearly compensation by company Size

Conclusion

All in all, my goal through this analysis was to provide insights about the state of AI adoption & MLOps in Industry, by examing to what extent enterprises have Machine Learning models in production, what are the main tools that they use for Data Storage, Model training, deployment, and other processes, what are the main frameworks and libraries used on a regular basis as well as what are the most common AI job roles that the companies seek.

Key Takeaways

- ✓ **21.7%** of the professionals in the survey said that their companies **haven't started yet to explore Machine Learning methods** vs **32.8%** of the respondents who stated that their organizations have already **Machine Learning models in production** either in advanced or in an intermediate stage.
- ✓ **Online / Internet-based Services**, **insurances**, and **tech** companies are the leaders in the adoption of Artificial Intelligence.
- ✓ Even if smaller companies might be better candidates for the implementation of AI, due to the absence of legacy systems, the survey results show that big companies are leading at the moment the way in AI adoption.
- ✓ **45%** of the professional that participated in the survey use **Cloud Computing Platforms** with Amazon Web Services (AWS) and Google Cloud Platform (GCP) being the dominant ones in 2022.
- ✓ The most popular AI jobs are Data Scientist and Data Analyst.
- ✓ **Top Skills Required for a Data Scientist / Machine Learning Engineer:**
  - **Programming Languages:** Python, SQL
  - **Machine Learning Frameworks:** Scikit-learn, Tensorflow, Keras
  - **Machine Learning Algorithms:** Linear and Logistic Regression, Decision Trees, Gradient Boosting Machines, CNNs, MLPs, Transformers
  - **Experience using Cloud Computing Platforms**
  - **Data Visualization Libraries:** Matplotlib, Seaborn, Plotly
- ✓ The main responsibilities of a **Data Scientist** are:
  - Analyze and understand data to influence product or business decisions
  - Build prototypes to explore applying machine learning to new areas
  - Experimentation and iteration to improve existing ML models

    while for a **Machine Learning Engineer:**
    - Build prototypes to explore applying machine learning to new areas
    - Experimentation and iteration to improve existing ML models
    - Build and/or run a machine learning service that operationally improves the products or workflows

- ✓ **43.51%** of the professionals hold a Master's degree
- ✓ Transfer Learning methods used mainly in Computer Vision Tasks
- ✓ Only **31.3%** of the respondents **use specialized hardware when training machine learning models** which indicates either that usually we don't deal with big data or deep neural networks that require huge resources for training or that the companies don't invest in specialized hardware and this causes a bottleneck to the productionization of ML models.

References

1. Mckinsey Report: The state of AI in 2021
2. Global Cloud Computing Market Report 2022: Increased Resource, User Mobility, and Ongoing Migration of Applications Over the Cloud Driving Growth - ResearchAndMarkets.com
3. ML Operationalization: Building a path to real-world business success
4. Kaggle Notebook: Spending dollars for MS in Data Science - Worth it ?
5. Kaggle Notebook: A story told through a heatmap
6. Kaggle Notebook: Data Science in 2021 : Adaptation or Adoption?
7. Kaggle Notebook: Head in the Clouds
8. What is Cloud Computing? The Key to Putting Models into Production
9. Kaggle Notebook: Data Science and MLOps Landscape in Industry

Thankyou