AI
organizational
scalability
-
a sample data book

Released on leanpub.com/ai-organizational-scalability

ABOUT THIS BOOK

This e-book is derived from a Jupyter Notebook that I released on 2021-01-06 as part of a “challenge” on Kaggle.com – *you can access the original file there*[1].

If you do not know what I am referring to: a Jupyter Notebok is a way to have (online, on the cloud; or offline, on your computer) a tool that allows to “take notes” that contain executable code (e.g. in Python or R), usually to carry out some computation (e.g. Machine Learning creation and execution of a model) and produce visualizations.

So, why an e-book about what, in the end, is a data-based e-book already online?

The specific “challenge” was to analyze some data from a survey (the annual Kaggle survery between its own community), while providing both a data storytelling, i.e. telling a story via data and, of course, visualizations.

This book extracts the storytelling part (i.e. data, charts) without the programming code, but adding a deeper explanation of the rationale of choices.

As I wrote above, it is “derived from”, not just an e-book version of the Jupyter Notebook.

Also because, if anything, the reverse is true: I had the idea of creating template notebooks based on my experience since the 1980s in data-driven decision making, cultural and organizational change, and supporting (both from the customer and the vendor side) commercial negotiations.

Also, as I had started learning these tools only at the start of the first COVID-19 lockdown, in March 2020, a secondary aim was to create an array of potentially reusable and potentially “skills showcase” notebooks, after the datasets that I started sharing when I joined Kaggle because I needed a place where I could share permanently some datasets that I was creating (initially used R, a statistical package, for the visualizations).

[1] https://www.kaggle.com/robertolofaro/ai-organizational-scalability-and-kaggle-survey

Both the Jupyter Notebook and this e-book are based on two elements of my data-related (decisions support for management) experience since the late 1980s:

1) *use only relevant data, not just pile up data*[2]
2) whatever you do, document choices
3) last but not least, document lessons learned.

In most cases, then I shared in my activities with others these lessons, e.g. via workshops.

This first data-based book (or “databook”) also leverages on the initial reason why in the early 1980s I approached Artificial Intelligence for the first time (specifically, PROLOG).

Back then, it was to help structure and access knowledge using an intuitive way to “ask questions”, but I was not surprised by the “AI Winter” that followed: extracting knowledge from humans was time-consuming and cumbersome, a librarian job that required also knowledge of the specific domain.

To compare: AI 1980s it was akin to creating the list of links on Yahoo (I remember registering my first website with them: if was akin to filling in a bibliographical card in a library), vs. AI 2010s (Machine Learning e.g.) filling in information for Google so that it creates its own results.

Right now, beside Machine Learning in its various forms, my main interested is in blending different disciplines to use AI both for automated discovery, monitoring, alert, audit, and also automation of some tasks.

But mainly as a kind of virtual “*cobot*[3]” (collaborative robot), i.e. an autonomous physical/software combination that is designed to interact with and augment their own “human colleagues”.

As I wrote above, this book is actually a “template”, and therefore the specific case is more to share a dataset that has been discussed on Kaggle by many others, so that you can compare the approaches.

Within the first chapter, I will discuss more my approach to Artificial Intelligence within a business environment, and why I selected the title of this book as title for my Jupyter Notebook submission on Kaggle.

[2] https://robertlofaro.com/relevantdata

[3] https://en.wikipedia.org/wiki/Cobot

CONTENTS

1 THE RATIONALE

Yes, there is a rationale for each and every choice represented within the Jupyter Notebook.

Before that, the choice was: why was interesting to carry out the analysis.

Somebody would say: it is a contest with prizes.

Useful, but not the prime motivator (albeit receiving funding to support further analysis activities would be welcome).

There are few elements:

1) I worked on the introduction of new technologies, processes, organizational structures since mid-1980s
2) since I started using in spring 2020 (COVID-19 lockdown) Kaggle as part of my Machine Learning journey and to network with people I could learn from or eventually work with (if I will have projects with a budget), I wanted to add “real” projects that could end up on my CV
3) in my past activities, I got used both in political activities and with my first employer to prepare summaries, presentations, and other “outline material”
4) as I described within the “About this book”, I see current Artificial Intelligence as a continuation and evolution of my past data-centric activities.

Therefore, my approach could be summarized into what I did in the past: blending qualitative and quantitative.

A decade ago shared in an article my approach to convert “qualitative” into “quantitative”, and how that helped in activities ranging from the obvious “software/vendor selection”, to auditing and summarizing multiple activities to identify behavioral patterns, to redesigning organizational structures, processes, or identifying motivation underlying mere economic side-effects.

Sometime your start from quantitative information and identify qualitative patterns, sometimes you identify a set of target behavioral patterns, define a clustering on each (the “levels”, such as the ubiquitous 1-to-5 scale), then use the clusters to identify, against the different behavioral patterns, the “profile”.

And, in change, identify also the transition patterns, overlapping of patterns, etc.

Seems either really ranging from hocus-pocus to too much quantitative, but it is neither.

You have a toolset, negotiate a mandate, study the situation (sometimes you are entitled also to a preliminary assessment, a “feasibility study”, before committing to a mandate), and decide which tools are to be used, and which are either irrelevant or “bells and whistles” (so common in both ICT and management consulting).

Enter modern Artificial Intelligence.

All the above in the 1990s and 2000s was done manually, from the early 2000s using e.g. radar charts in Excel.

What you will find in the next few chapters is an example, a narrative through data and guided by both the “mandate” (a fictional one, in this case) and the data.

In reality, the report I posted would have been just a first phase, to ask confirmation on which areas to focus on, before committing to e.g. building models to “cluster” automatically, or using other tools, or even adding further investigation or data collection.

In this case, I decided to do what I did few times between the late 1990s and mid-2000s, i.e. identify a visual tool that could be used across this first phase to identify further areas of analysis, informative but that allowed comparison across multiple dimensions of analysis without having to spend 10 minutes to study each chart design.

I wrote on purpose “each chart design”.

Both on the customer and supplier side, I saw reports where it seemed as if the consultants had decided that what they had been asked was how cool they were in using a completely different representation every few pages.

Not because either the mandate or the data required it, but because it was less boring.

Well, if you have to revise balance sheets, budgets, etc, creativity is welcome- in the design phase.

But in the collection and preliminary analysis phase, you need an ability to quickly highlight “pain points”, and then spend time where it seems promising.

So, less is better.

In this report, as a provocation, I decided to do as I did in 2002 for a customer: selected just two simple tools, radar charts and tables with some KPI computation, to highlight potential issues, and let then readers decide.

Yes, I added also the compulsory histogram and heatmap- but not as a showcase of my skills (also if I spent some time years ago studying and trying most visualizations available with a library/framework/paradigm called “ggplot”, but using R).

This is the main design choice: keep it simple, do it fast, make it repeatable as much as possible.

When I liked a chart, i.e. decided that was fit for my purposes, and clear enough to understand intuitively, by just browsing through, I converted the code generating that chart it into a function, to allow adding more elements of analysis while keeping a consistent visual approach.

The initial dataset had 355 columns, i.e. questions (or choices, for multiple choice questions), by choice for the analysis selected a smaller number.

Anyway, as any such activity, I did not choose “at will”, I used the “Kaggle Executive Summary” as if it were a preliminary analysis.

The concept is simple: somebody did a study, represented by the "Kaggle Executive Summary", covering all the questions, all the options, all the information- horizontally, i.e. without choosing a specific area of analysis.

In my case, I wanted to understand better the community, considering that in any corporate software initiative in the future is to be expected to have to consider also some data-based "intelligent tools".

I wanted to understand if, after few years, this new Artificial Intelligence trend started showing signs of what I could recognize as "success factors" in other cases, or we could soon be heading into another "AI Winter".

I will be frank: my first forays in Artificial Intelligence was on PROLOG, in the 1980s, before I started working, trying also LISP, but I was deep enough into PROLOG that I was in contact with an Italian association called GULP (Gruppo Utenti Logic Programming).

Then, worked creating models on a Decision Support System (henceforth DSS) for customers (number crunching for senior managers so that they themselves, or their assistants, could do what-if analysis and also goal seeking across multiple dimensions).

I considered creating a PROLOG Expert system to "explain" models to business users, but then switched job and, as it was something done in my spare time (and no colleagues used PROLOG), it ended there.

So, "explainability of Artificial Intelligence" is a theme that I consider critical- and this requires a mix of skills that extends beyond number crunching or wizardry on specific tools.

At the same time, I saw with DSS models, as well a decade later with Data Warehousing and Business Intelligence how success in using a technology implied, again, having it into the hands of those who had a communication line with those making decision.

Or even decision makes (in the 1980s, in Italy, was unusual to have managers using personally a computer, imagine a DSS).

In way too many companies there is too much "shelfware", investment in software packages or custom development that was maybe even deployed, but never went past the "Proof of Concept" phase.

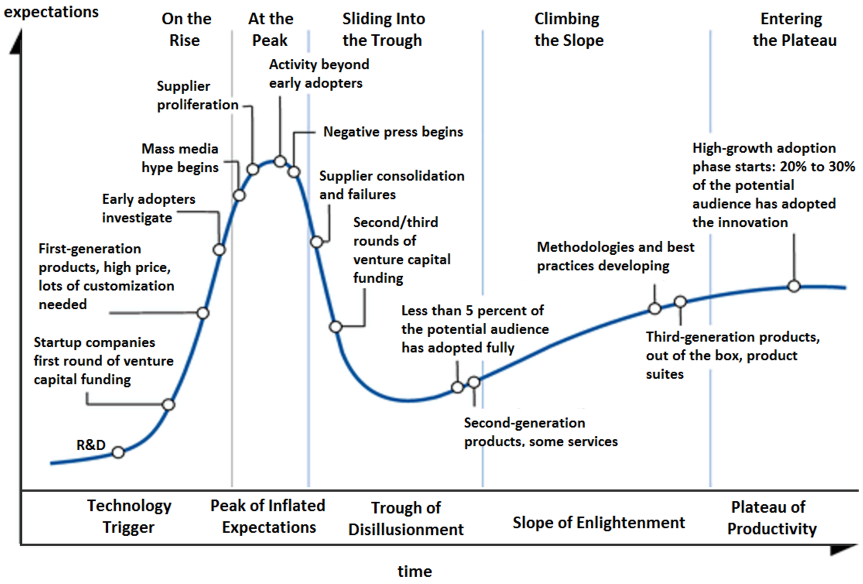
Or, as I heard at a German corporate software user group a manager report a CIO had said “our company has more pilots than Lufthansa”.

Pilot projects are fine- but since the 1980s I saw many that could be more aptly described as a “pilot done because you had to jump on the bandwagon to be trendy”.

And this is one of the reasons why I also witnessed various “winters”- in various technologies (the “predictive models” I read about today were part of my 1980s DSS models, then abandoned).

In the early 1980s, my approach to PROLOG actually derived from my prior interest in designing languages, and trying to apply to natural languages what was known as BNF- as I wrote in previous pages, representation and access to knowledge.

You probably saw *the curve about technology and hype*[4]


expectations
On the Rise
At the Peak
Sliding Into the Trough
Climbing the Slope
Entering the Plateau
Activity beyond early adopters
Supplier proliferation
Negative press begins
Mass media hype begins
Supplier consolidation and failures
High-growth adoption phase starts: 20% to 30% of the potential audience has adopted the innovation
Early adopters investigate
Second/third rounds of venture capital funding
First-generation products, high price, lots of customization needed
Methodologies and best practices developing
Less than 5 percent of the potential audience has adopted fully
Third-generation products, out of the box, product suites
Startup companies first round of venture capital funding
Second-generation products, some services
R&D
Technology Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time

I toyed then in the 1990s with some examples of “neural networks”, as, prior to “electronic brains” (as in the late 1970s computers were still called in Italy), I had been interested in the way worked human brains- electrical activity, etc.

[4] See https://en.wikipedia.org/wiki/File:Hype-Cycle-General.png

Jump to the mid-2000s, I purchased a Playstation 2 while living in Brussels, after seeing in London few years before a Sony presentation on the architecture of the Emotion Engine- along with the Linux software development kit available to develop software on it.

Jump forward to the late 2010s, when I had decided to create again a consulting company and planned to blend my cultural and organizational experience and past AI toying with more modern approaches.

So, I purchased from a distributor an Intel “Neural Network on USB”, a Movidius- but then had to work on other projects.

Therefore, I had a chance to “play” hands-on with various technologies, but always looking at their capability to support data-driven decision making.

So, if you ask people on the technical side of information technology, they would tell that I am more on the “process/business side”.

And if you ask people on the business side, they would say that I am an “information technology person”.

In my experience, a key element in spreading acceptance of a new method, process, and technology is to extend it past the “technology experts”, and closer to business.

So, to move past the “showcase pilot” and closer to those who can actually influence its adoption.

There is another famous curve that I used e.g. with start-ups, showing “early adopters” and other “cultural roles” both in consumer and corporate environments.

This book is an e-book about a Jupyter Notebook, so I will skip a long discussion about that “influencing” part.

If you are interested, e.g. have a look at Nir Eyal’s “Hooked”.

The title of this is about the future: it is not just a matter of “adoption”, but a matter of building (and retaining) momentum.

Just few paragraph about the title of the Jupyter Notebook and this e-book.

If you look at webinars about Artificial Intelligence, there is a significant quota of presentations from small teams.

Might be start-ups, academic labs, competence centres- but, overall, also in corporate environments, the size of most teams talks about craftsmen, not large pipelines.

Expanding, as with any corporate technology, would require having both small and “scalable” teams.

Not necessarily larger teams- also aggregation of “cells” that enable something more complex, replicable, and that can be aggregated when what is needed is a more systemic approach.

If you think e.g. at autonomous vehicles, in urban centres you probably should consider their interaction with other devices, vehicles, and, yes, the town itself (and its inhabitants).

As I wrote in the introduction to this book, I see more a blending of humans and “virtual/ritual robots”, i.e. “cobots”.

From this perspective, already decades ago in some countries started digital transformation: I do remember the 1999 sponsoring from the OECD of “e-government”, what eventually gave us “open data”.

The key element was, as is now, comparability.

Since 2015 I have been sharing articles about “sustainability”, as I have seen since the 1980s, first really in the Army, then in business, how the “supply chain of information” is what defines sustainability of organizations, processes, and also single transactions.

Almost a decade ago in Milan, Italy, attended a workshop on IoT that clearly highlighted the trend: more data, more frequently, everywhere.

First side-effect: as shown also by NASA collection of exploration missions, or even the observation satellites of a generation before, we had always a backlog of data analysis.

While learning since March 2020, using Machine Learning I tried to redo also something that in mid-2000s required complex, expensive, and even custom software: piece of cake.

With some humble learning, in minutes or hours was able to do so.

Since 2015, after attending the Expo 2015 in Milan, started drafting a book on business and sustainability from a number crunching perspective, as the announce of the UN SDGs showed that we were converging toward what I had seen decades before in banking in Italy.

Or: converging toward a degree of transparency and data harmonization that enabled then to carry out analyses without having to re-invent the wheel.

If you look at Machine Learning, you will see that many models “recycle” pre-existing analyses, e.g. “pre-trained models”.

This delivers both a time- and resources-saving element, and enables easier replicability.

There is a risk: if you “inherit” also the biases in the models that you use as you starting point, as shown by that joke that you can spot some image analysis models as having inherited from a famous model and dataset, as, whenever presented something new…

… they start listing breeds of dogs that might be within the image.

The *UN SDGs website*[5] contains information about each one of the 17 goals decided in 2015.

I could share the image, but, formally, would first need a written permission- hence, I can talk about it, write about it, describe it- but cannot use the images[6].

Jokes apart, the key concept is simple: if you have a look at the UN website, it is clear that each goal is about data- collection, monitoring, audit.

And if you look at the details of any of the 17 goals, e.g. *11 on “sustainable cities and communities”*[7], you will see that COVID-19 has had a significant impact.

To keep our complex societies working, more data-based management, monitoring, execution have been activated in 2020.

[5] https://sdgs.un.org/goals – contains details on each goal, its target, associate publications and actions

[6] “Any use of SDG branding for fundraising and commercial use requires prior written permission.”

[7] https://sdgs.un.org/goals/goal11

On 2021-01-21 attended a webinar on the impacts of COVID-19 on supply chain risk: each and every presentation from businesses explained, in reality, how they had to rely more on data, on transparency on who was doing what and how, in order to restructure their activities.

And most said just one thing: yes, COVID-19 in 2020 accelerated the trend, but now that most have changed their way of working, there is no way back.

Or: with or without Artificial Intelligence, there will be more data-based business management.

Which implies more managers, or those working with them, able to work with more data, faster, and with tools that enable them not just to do the usual “ex-post” (monthly, weekly, quarterly, etc) reporting, but also identify potential trends.

From risks of disruption of supply chain, to preventive maintenance of vehicles (i.e. replacing components before they break down, using data from the vehicles to identify metal fatigue and other parameters).

To the obvious more advanced, continuous tests on traffic on roads, optimizing the logistics of shops deliveries, etc.

All this requires data- and the integration of data analysis with business, which, in turn, requires a mutual understanding.

Right now most presentations are concerned with “interaction” or “integration in a known environment”.

In the future, the aggregation might be temporary and transactional, and even in corporate environments could “emerge” new aggregations that solve specific temporary needs, not having been pre-planned before.

If you read the previous paragraph: it is what happens when humans interact- you have processes, rules, etc, but generally adapt and can augment performance when you aggregate your adaptation with the adaptation of others that interact with you.

Until you reach a state where your aggregated performance exceeds the performance of the components interacting- then, suddenly, what was working before stops.

This is what I call “organizational scalability”.

It is not simply “being able to act small and large”, but of generating more value than just the sum.

AI, right now, in most applications, is not really looking at aggregating on a transactional basis.

So, it can deliver value in specific cases, domains, etc, even learn from data- but to justify something more than mere re-using of what has already been used elsewhere, we need something else.

Therefore, as in other technologies, I see as needed something more than a mere “knowledge transfer”- it is a matter of generating awareness, and then expanding the ability to have new AI-integrating solutions “emerge”.

To this end, as in other cases, I see critical the involvement of people whose job title typically requires interacting with "business domain specialists" and decision-makers, i.e. Business Analyst and Product/Project Manager.

And that was my selection within the data provided by the 2020 Kaggle survey.

As I wrote in the introduction, I considered as if the “Kaggle Executive Summary” were a “preliminary assessment” on the data, and had been asked to then move on.

Actually, along with the survey, I used as a source McKinsey's *"Global Survey - The State of AI in 2020"*, carried out between 2020-06-09 and 2020-06-19, containing 2,395 responses, pag. 13: "representing the full range of regions, industries, company sizes, functional specialties, and tenures. Of those respondents, 1,151 said their organizations had adopted AI in at least one function and were asked questions about their organizations’ initiatives." (retrieved and read 2021-01-02).

Unfortunately, data for the latter were not available, so I just enclosed within the Jupyter Notebook my reading notes from both.

I worked also with virtual and outsourced teams since the 1980s, and therefore I have been exposed to different IT service cultures, in different phases.

The Kaggle survey covered the whole geographic reach of the Kaggle community.

This choice is to identify if there are any cultural and market differences that emerge from the data, as this could be useful e.g. to identify in the future different team-building approaches in different countries.

I selected three countries, that would allow the typical "follow the sun" (24/24) that I was used to for decades while working with multinationals:

- United States
- India
- Europe.

A qualification on the latter.

These are, in alphabetical order, the countries considered within Europe: Belarus, Belgium, France, Germany, Greece, Ireland, Italy, Netherlands, Poland, Portugal, Romania, Russia, Spain, Sweden, Switzerland, Turkey, Ukraine, United Kingdom of Great Britain and Northern Ireland.

I considered as "Europe" not only the EU Member states that have members (probably many more are also in "other"), but a geographical Europe, from the Atlantic to the Urals.

Again, will be discussed in the future more in detail- right now , time to talk about the structure of this book, where each chapter is equivalent to a section of the Jupyter Notebook.

Each chapter has the same structure:

- an introduction to the "why" this chapter is therefore
- as an example, the section from the Jupyter Notebook
- anyway, the section will be without code- just data and charts.

Whenever needed, a box will also show further commentary:

commentary

A brief outline of the structure of this book is within the next page.

This is the structure of this book:

Chapter 2 Executive Summary

- Context
- Focus adopted
- Preliminary results
- Structure of the report
- References

Chapter 3 Assumptions and Choices

- Assumptions (Themes selected, Population selection)
- Choices (Structure, Data sources, Publication release)

Chapter 4 Dimensions of Analysis

- Questions
- Values distribution
- Job titles
- Geographic coverage

Chapter 5 Data Preparation

- Data by country
- Create subset(s) focusing on Europe, India, USA

Chapter 6 Data Visualization

- Check BA/PM frequency by country
- Confirm comparison countries
- Check decision influencers
- Compare demographic distribution
- Compare companies characteristics
- Compare technical characteristics
- Assess results and analysis
    - Focus on the questions to use to compare countries
    - Identify correlations
- Further investigations planned

Chapter 7 Appendix

- Notes from the Kaggle Executive Summary
- Notes for the McKinsey Survey

Chapter 8 Conclusions

2 EXECUTIVE SUMMARY

In my view, the key element of this first section of any report is remembering that you are writing to convey to an audience the results of your analysis- here, the audience of the Survey Executive Summary

Despite what many books say, whenever preparing a report, I first draft the executive summary, covering the starting point and information about the structure of the report.

Then, it is completed step by step after each section is completed, and amended as needed, before a final review: a roadmap, not a plan.

My approach is to try to make it as short as possible, and coherent (up to reusing phrases from the final version of each section of the report), ensuring that has a coherent storyline.

To design your storyline and approach to this section, e.g. you can read *Aristotle's "Poetics"*[8] , or use the (free) tools associated to a book on script writing that I read years ago, *Syd Fields' "Screenplay"*[9], that contained what he called *"Paradigm"*[10], i.e. a visual structure of the narrative, along with an approach (of course, you have to adapt to your audience and purposes- the link contains also an audio lesson).

Then, the other sections go further down into details, but, as in the final version of the Executive summary, each section is "connected" to the next, and connects from the previous one.

[8] https://en.wikipedia.org/wiki/Poetics_(Aristotle)

[9] https://sydfield.com/

[10] https://sydfield.com/syd_resources/the-paradigm-worksheet/

In this case, the “Executive Summary” section was structured as follows (it is a personal choice- not a standard):

- **Context**
  where you set the tone, define the boundaries of your ignorance, and share pointers for your audience- if they disagree something that is not withiin your “context”, then probably that is what has to be redesigned (in other domains is called also “scope”, but I prefer to use that for the more detailed subsection that follows)
- **Focus adopted**
  where you actually “visualize” with words your boundaries but also further constraints deriving from your choices on both boundaries, and what you consider within these boundaries, providing e.g. also the list of your sources of reference
- **Preliminary results**
  in my view, if you are preparing a report for somebody else, it is up to your audience to make choices- including disregarding your results; as a consultant/analyst, you can just share what, based upon the data and focus as well as your experience, you identified- in my business domain, also when dealing with data (e.g. checking the invoices associated with a contract), there is a difference between what the data say, what you can derive, the scenarios of how that “can derive” turns into “what is means”, and “choices”: the latter is not up to you- and I saw many reports when instead it was blatant that the consultant/analyst had an axe to grind; an element of “influence” is present in any analysis report, but should be clearly separated from what is derived from the data and focus
- **Structure of the report**
  the structure of the report is a narrative in and by itself, but its content should be coherent with what was presented within the executive summary: the structure per se is presented as a logical evolution, but probably its development contains many twists, turns, blind alleys- you could share those as material for reference should in the future an evolution of the analysis is needed, but do not get carried away with the “storytelling” part- you are not writing the next Clarke, Crichton, Grisham besteseller: you are sharing the results of your analysis in a way that is transparent, traceable (to data and “why” of your analysis mission)- if you feel the urge to share you “travelogue through data, write a separate narrative book
- **References**
  material that you want to share.

A. EXECUTIVE SUMMARY

Context

As outlined *within the Kaggle introduction to the 2020 survey Executive Summary*[11]: “the survey was live for 3.5 weeks in October, and after cleaning the data we finished with 20,036 responses”.

Membership of the community implies either having already some data science skills, or at least a willingness to observe, learn from, network with those who have.

[11] https://www.kaggle.com/c/kaggle-survey-2020/overview

Focus adopted

My use of *the Kaggle survey results dataset*[12] is focused on identifying some potential trends on the bridging between data science and business, notably on the management side.

To that end, selected people whose job title typically requires interacting with "business domain specialists" and decision-makers, ***Business Analyst*** and ***Product/Project Manager***.

This is because, since the 1980s, in successive rounds of "new data-based decision-making" in business organizations, observed that only when awareness is widespread, and use eventually (at different degrees of expertise) is within the direct circle of decision-makers, and not just in a closed self-referential technical group, a technology is adopted.

As an example, in the 1980s Decision Support Systems were probably too technical and too complex, and embraced by managers in their 30s, at a time when more senior managers did not even use personally a PC.

In the 1990s, business intelligence tools (now widespread) lowered the "knowledge bar", by enabling also intuitive data investigation, and expanding on visualizations that were both easy to obtain, and business-relevant.

But it all starts with people, not technology.

Therefore, this report is a comparative analysis between three countries (or aggregations thereof) that:

- represent different market concepts
- have a different mix of respondents (age, education, etc)
- have number of respondents in the same order of magnitude.

[12] https://www.kaggle.com/c/kaggle-survey-2020/data

After an initial check across all the countries contained within the dataset, therefore three countries have been selected:

- Europe
- India
- United States of America (henceforth USA).

Europe is an aggregation of countries geographically in Europe, from the Atlantic to the Urals, not just the European Union, as anyway not all the European Union countries are explicitly represented within the survey dataset.

Preliminary results

The analysis started by reading the Executive Summary released by Kaggle, and then reviewing other material, to confirm the focus of the report.

In early January 2021, while preparing the report, decided to replace as source of reference other documents with a new report produced in June 2020 by McKinsey on the "State of AI", and focused on the business impacts side, that confirmed some elements within the analysis.

Unfortunately, those data are not available, just the Executive Summary, and therefore I added the link under *"References"* later in this section.

As discussed within the report, this notebook is just a first phase, as more advanced analyses will be carried out in the future.

The purpose was to share a feed-back on what is inside the dataset, and suggest which other data (or integration with other data sources) might be useful, within the focus shown above.

The key section of this report, to understand what follows, is the (mainly non-technical) ***Section B. ASSUMPTIONS AND CHOICES***, where choices both about data and structure of the report are presented.

From the analysis of the data, some first results (see ***Section E. DATA ANALYSIS*** for more details):

- the three countries cover 11,992 respondents out of 20,036, i.e. 59.50% of the respondents to the survey
- across the job titles of Business Analyst and Product/Project Manager, the three countries cover 898 respondents out of 1,490, i.e. 60.3% of the respondents to the survey
- by focusing on Business Analysts and Product/Project Managers (898 out of 11,992, i.e. 7.5%), the gender gap is even greater than on the Executive Summary by Kaggle, i.e. the higher up the decision-making chain, the more "gender equality" is lacking- also in Europe, despite all the "positive bias" initiatives adopted since the beginning of the XXI century
- the audience (members of Kaggle) is obviously biased, showing at least interest into the machine learning domain, but nonetheless the differences between the three countries selected in demographics and company they work for are a confirmation of what is observed routinely also in other surveys on the general IT population
- nowhere there is a strong correlation (positive or negative) between the roles selected and the other questions, albeit, after identifying the average correlation for the three countries, there are differences that might be worth further investigation.

Areas of further investigation that, based upon the results, identified:

- explore the differences, and try to identify patterns at least for the three questions that show the higher level of difference (Q32, Q24, Q20) for the roles selected
- verify if those patterns apply also to the community of respondents at large, or are just an indication, due to the job titles selected, applying just to the three countries
- for Europe, as the subdivision in countries is present, identify if the above mentioned patterns show differences
- also, would like to see if these differences are matched by differences vs. UN SDG for both the three "macro-countries" and the individual European countries.

Structure of the report

The report is composed by just one long notebook, and some files reproducing its contents, charts, results for comparison and reproducibility purposes.

By the end of January 2021, a PDF version will be added to the files.

The structure of this report:

- A. Executive Summary
- B. Assumptions and Choices
- C. Dimensions of Analysis (data selection)
- D. Data Preparation
- E. Data Analysis
- F. Appendix

Section B contains the focus adopted and the production process, while Section C contains the data selection criteria.

Section D prepares the data for further use, creating different aggregations.

Section E is the core of the report:

- first discussing each one of the features selected (16 columns out of 355)
- then identifying why some are not relevant to the focus of this report
- on those selected to remain, a correlation analysis
- finally, based upon the results, listing the other potential areas of future investigation that I would like to follow, e.g. checking if any pattern is also consistent with patterns within the UN SDGs, notably for individual European countries (already posted on Kaggle datasets on UN SDGs, as part of another ongoing publication project, on *my kaggle profile*[13].

Within Section F, I enclosed my notes/quotes from the two reports listed within the references.

[13] https://kaggle.com/robertolofaro

References

Out of many more documents, I considered as a reference for the analysis two reports:

- *Kaggle's executive summary PDF*[14], containing the above mentioned 20,036 responses (retrieved and read 2019-11-20)
- *McKinsey's "Global Survey - The State of AI in 2020"*[15], carried out between 2020-06-09 and 2020-06-19, containing 2,395 responses, pag. 13: "representing the full range of regions, industries, company sizes, functional specialties, and tenures. Of those respondents, 1,151 said their organizations had adopted AI in at least one function and were asked questions about their organizations' initiatives." (retrieved and read 2021-01-02).

Only the data from the Kaggle survey have been used in report, albeit the McKinsey survey executive summary (13 pages) can be read online for free (and, upon registration for a free account, the PDF version can be downloaded).

[14] https://www.kaggle.com/kaggle-survey-2020

[15] https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2020

3 ASSUMPTIONS AND CHOICES

This section is not about the results of assumptions and choices, as those are presented:

- as an outline, within the “Executive Summary” section
- as details, within the sections that follow.

Instead, this section, that should be as short as possible, and should further explain what within the “Context” and “Focus selected” subsections of the “Executive Summary” were highlighted.

As I wrote before, this is a first experiment on reusing Jupyter Notebooks to deliver a “live” version of reports I used to produce in the past not only for “quantitative” analyses (e.g. checking invoicing, quality, KPIs, etc), but also “qualitative turned quantitative” (e.g. from organizational change, to process- and activities-auditing, and also for vendor selection, representing through numeric values clustering in categories or rankings).

For this report, considered as part of the “**Assumptions**”:

- the main thesis, described directly under the headline as a bullet list
- **Themes selected**, the subsetting of the original dataset features (the columns)
- **Population selection**, the subsetting of the original dataset data points (the rows selections: in this case, by job title and by country)

In other cases, this section could actually contain other elements to share the constraints to the report.

As an example, this is typically where the boundaries of the agreed mission to produce the report are detailed, constraints of a “contractual” value.

Generally, the “Assumptions” part defines also what can and cannot be expected from the report, and how far the analysis went.

The “contractual” part, for a consultant, is more critical than it seems, as often the production of a report includes also involving people from the organization that commissioned the report, and it is not unusual to be exposed to what could be defined under at least two banners:

- **bandwagon-effect**: using the analysis team assigned to the report to “settle organizational scores”, e.g. by feeding into the analysis influences of choices, either directly, or through biased information (more often by omitting than by falsifying information)
- **resistance to change**: similar to the prior point, but slightly different in execution, and often working through “playing with deadlines”, by stretching time to carry out actions needed to support the production of the report.

Depending on the nature of the report, there might be other influences from the organizational structures, willingly or not- but you can find more discussion about that e.g. by reading the books I shared (can be read for free) on my *issuu.com profile*[16]*.*

For this report, considered as part of the “**Choices**”:

- **Structure**, explaining how the report is partitioned in sections
- **Data sources**, which data have been used within the report
- **Publication release**, even for a Jupyter Notebook, generally there is some supporting information that is released along with the report (e.g. copies of the charts, references, etc).

For other data-based reports part of this publication experiment, I will keep using this same structure and post additional “reports” attached to the datasets that I published on my *Kaggle profile*[17], all using “open data” (I.e neither proprietary nor confidential information has been used).

In some cases, the existing dataset that I shared on Kaggle is a selection of a dataset that I had created for other projects.

[16] https://issuu.com/robertolofaro

[17] https://kaggle.com/robertolofaro

The notes about the publication release and structure are a little bit more "technical" than what I would write for a report released without a data interaction component.

Eventually, some of the technical remarks will not be needed, as management-level users will routinely receive reports that actually allow to interact with data beyond what current Business Intelligence and Analytics allow.

Probably reports eventually will be akin to Datamarts in the 1990s, i.e. the audience will be able to "spawn" analyses based on the report, by using the report to "seed" such new analyses, with options to either retain a linkup with the original data evolutions, or "freeze" the point of reference.

In this report, where "radar charts" usually show information across just three countries on multiple options, often the static "radar chart" would be enough.

During presentations for projects, generally I was asked questions about specific combinations of values (what is represented by the shaded area), and therefore the feature of Plotly that enables to export charts as HTML files that then do not need to access the original dataset to enable interaction with the chart, is quite useful.

Imagine sending around an intracompany chart attached as an HTML page, or post it on an intranet (even on a private cloud): would be better than sharing Excel or Powerpoint files, or even PDF files printed out of spreadsheets (as there are plenty of tools to extract e.g. Excel and Word from an Acrobat file).

The only issue: fine if you use "open data", i.e. information that is public, but if you are working on restricted, confidential, or proprietary data, using a dynamic chart that connects to an external server to present the data might require some authorization.

B. ASSUMPTIONS AND CHOICES

In the appendix, after the data analysis, I will share more analysis, but this preamble is to explain the logic in my use of the *dataset*[18]*.*

ASSUMPTIONS

1) Artificial Intelligence to avoid facing yet another "*AI Winter*"[19] has to become part of "the new normal" that the convergence of "*Digital Transformation*"[20], accelerated by the COVID-19 pandemic side-effects, is delivering.
2) The transformation will affect both society and businesses, but I will focus on the business side.
3) What I observed since the 1980s is that technology, to become embedded in businesses, needs:
    1) a **mandate**, or at least (for initiatives started from the bottom of an organization, as spontaneous experiments, and not as initiatives, programmes, etc from the top) clear sponsorship from decision-makers

    2) adequate **resources** (budget, but also infrastructure and integration within the ordinary budget of e.g. business units or corporate ICT)

    3) last but not least, **talent** , i.e. an environment that attracts, retains, develops human resources, and spreads awareness across the whole organization.

[18] https://www.kaggle.com/c/kaggle-survey-2020/data

[19] https://en.wikipedia.org/wiki/AI_winter

[20] https://en.wikipedia.org/wiki/Digital_transformation

THEMES SELECTED

After reading the Survey Executive Summary report and reviewing the data, the following selection criterias have been applied to choose which data (questions) should be used in the analysis:

- identifying people
- defining the organizations they work for
- identifying the level of adoption of data-based decision-making (not just AI- also Business Intelligence, as e.g. Microsoft PowerBI allows integration also with R and Python)

POPULATION SELECTION

These are the **job titles present within the dataset**:

Data Scientist', 'Research Scientist', 'Data Analyst', 'Statistician', 'Data Engineer', 'Product/Project Manager', 'Currently not employed', 'Student', 'Software Engineer', 'Machine Learning Engineer', 'DBA/Database Engineer', 'Business Analyst', 'Other'

My use of the *Kaggle survey results dataset*[21] is focused on identifying some potential trends on the bridging between data science and business, notably on the management side.

To that end, selected people whose job title typically requires interacting with "business domain specialists" and decision-makers, ***Business Analyst*** and ***Product/Project Manager***.

This report is a comparative analysis between three countries (or aggregations thereof) that:

- represent different market concepts
- have a different mix of respondents (age, educational background, etc)
- have number of respondents in the same order of magnitude.

[21] https://www.kaggle.com/c/kaggle-survey-2020/data

The three countries selected:

- ***Europe*** (as aggregation of countries geographically in Europe, from the Atlantic to the Urals, not just the European Union, see in sections C-D)
- ***India***
- ***United States of America*** (referenced within the commentary as ***USA***).

CHOICES

STRUCTURE

Key points:

1. This report is built using just one long Jupyter Notebook
2. Added output exporting logic, to ease producing a report for non-Jupyter users
3. For the Plotly charts, the export is in HTML format, to allow interaction from any browser
4. To enhance readability for the textual part of the export, each “print” statement is preceded by the printing of a “contextual positioning” of what is going to be printed

More notebooks have been prepared to identify the focus of the analysis, and investigate other options.

From those notebooks, only the items that could contribute to the narrative chosen for this notebook have been included.

Whenever there was a group of lines that was used more than once, tried to create functions, to enforce consistency in data presentation and visualization, and enable future expansion of the analysis while keeping a consistent format.

Except for key questions, where used different visualizations, after identifying the target discussed above, selected just two charts, using Plotly (to allow interacting with data points):

- a radar chart with multiple (Europe, India, USA) cases
- a static radar chart separating each country, to better highlight patterns relevant to each country, without any scaling issues that e.g. a single country with values out of range might force

Future releases might be restructured to add more functions and different visualizations.

DATA SOURCES

This first report uses exclusively the Kaggle 2020 survery dataset to identify what the data tell, and highlight further items to discuss.

PUBLICATION RELEASE

Along with the notebook, released also:

- a ZIP file to enable to read the results and view individually each chart (for the plotly charts, each HTML file is the chart "live", as interfaces with the Plotly server to reproduce the chart, enabling to visualize details)
- an HTML file containing the whole execution of the notebook as released, to enable reproducibility and verify data, generated by the following command:

```
jupyter nbconvert --execute --to html NOTEBOOKNAME.ipynb
```

As this report was an example, and anyway was delivered as a single file, did not consider expanding this section to actually add the information that is referenced.

In the sections **C Dimension of Analysis** and **D Data Preparation**, respectively chapter 4 and 5 in this book, already present the details about data selections.

Anyway, in this case the report is, despite its apparent length, quite short and straightforward.

In other cases, it might be useful to replicate the information:
- within the “Executive Summary”
- within this section “Assumptions and Choices”
- within each section or subsection where relevant.

Beside clarity, there is also another issue: if a report is a “single run”, i.e. not to evolve into a yearly, quarterly, monthly, weekly report (as most data-based reports on business, “panel” surveys, KPIs, eventually become), it is fine to replicate information around the report whenever useful.

If the report has to be re-released, it is better to automate as much as possible.

As this first Jupyter Notebook was an example and potential template, I made the latter choice.

While visual formatting could have been easier if I had chosen to write as “markdown” lists of questions, values, etc, I selected instead to create (and call) “printing” functions using the data at hand to ensure that I needed, in most case, to update just once if, e.g. visual or questions/data selections had to be changed.

Even on a small report as this one, it actually helped to quickly fix visualizations and headings.

4 DIMENSIONS OF ANALYSIS

A first caveat: as I workedin the 1980s with tools that allowed to create "multidimensional storage", and in the 1990s with "star schema" databases as well as more traditional relational databases and other types not much more common, and proprietary databases built for decision support such as Arbor's Essbase, actually "selling" projects to the business side of companies, or even just as project manager and business analyst, I have something in common with business analysts, product managers, and project managers.

Remember that the purpose of this dataset was to allow an analysis to be carried out: not as fancy as building predictive models, but often in business you need to understand where you are and where you are heading to- and both require business understanding (albeit in some cases actually in the past would have found useful using unsupervised learning- more about this later).

Consider a car:

- you have a **model**
- but you have also **plants** where it is assembled from components
- **markets** where it is sold
- **time** period to track production/sales
- and then **variables** (in current data science parlance, I should say "features") such as manufacturing cost, distribution cost, overheads (for those costs that aren't directly related to the product), plus many others, down to (hopefully positive) net result.

Now, “aggregate” (sum up vertically and across) on the various dimensions, and you get a picture of the net results in each period.

In another industry, such as banking, you could talk about branches, products, customer segments, etc.

In retail, you might talk about “food” and “non-food”, and then talk about different subdivisions down to the single packaging.

In the end, be it a financial or a chemical product, you have elements composing the product, distribution channels, customer segments, and many datapoints that you collect to monitor, control, manage.

All these different perspectives are “dimensions of analysis”[22].

You can add more dimensions (such as the model and plants) or have less (such as considering only model, no matter where it is produced).

It all depends on your purposes.

If you want to track just sales, you do not need the plant dimension.

If you want to track logistics costs or quality, probably you do.

In this case, as I wrote in previous chapters considered that:

- the audience of my fictional analysis report is the same audience that already received the Survey Executive Summary
- between the 20036 rows across 355 columns (the questions and multiple choices), I had to select those relevant.

The analysis discussed within the previous section, ***“B. Assumptions and Choices”*** focused on:

- job title: Business Analyst and Product/Project Managers
- country: Europe, India, USA.

This, within the context of understanding who they were, within which corporate environment operated, and with which level of technology.

As I wrote before, each section is step within the storytelling, and each step, in this case, focuses the next.

Therefore, in this case, the resulting choices were relatively simple.

[22] For a conceptual description, see https://en.wikipedia.org/wiki/Product_structure_modeling

The structure of this section the report mirrors the end result of the choices, as it has to be a “narrative”, but of course getting to that point implies exploring different options.

In a real business environment, doing what I did (sharing only the material related to the result) is common- albeit, in my view, a mistake.

Choices made under the current level of knowledge and constraints might involve excluding choices and information.

In business, when it was time e.g. to redo an organizational design or an evaluation of choices for a budget, often what as used was…

...the report resulting from the previous “design” exercise.

And this, even if the conditions that existed within the context of the previous round had changed.

Therefore, in a real business case, in this phase would actually:

- do as I did for this report, i.e. share the material relevant
- separately, keep track of the other choices that have been excluded, and the reasons.

In this sample report, I produced also temporary Jupyter Notebooks that helped filter out what was not relevant, while keeping track not only of positive choices (what I retained), but also of negative choices (what I removed, and why).

As an example of the process, I left within the report few questions that, at an initial analysis, seemed relevant, but then, in reality, either had no impact (e.g. gender gap was unfortunately more or less consistent across the three countries), or gave results that made sense only if further contextualized by other questions that were not part of the survey (e.g. the “influencer” element).

If, for example, the 2021 version of the Kaggle Survey were to include different questions, questions that were excluded in my report on the 2020 Kaggle Survey (e.g. the “influencer”, but also others) might again be added to the report.

To this end, it was was also useful, as I wrote in the previous chapter, to barter some flexibility in formatting the textual part (tables, lists, etc) in exchange of an easier ability to add and remove questions from the report with minimal manual intervention (to ensure consistency).

C. DIMENSIONS OF ANALYSIS

QUESTIONS SELECTED

PERSON	Q1 What is your age (# years)?
	Q2 What is your gender?
	Q3 In which country do you currently reside?
	Q4 What is the highest level of formal education that you have attained or plan to attain within the next 2 years?
	Q5 Select the title most similar to your current role (or most recent title if retired):
COMPANY	Q20 What is the size of the company where you are employed?
	Q21 Approximately how many individuals are responsible for data science workloads at your place of business?
	Q22 Does your current employer incorporate machine learning methods into their business?
	Q24 What is your current yearly compensation (approximate $USD)?
	Q25 Approximately how much money have you (or your team) spent on machine learning and/or cloud computing services at home (or at work) in the past 5 years (approximate $USD)?
	Q23_Part_1 Select any activities that make up an important part of your role at work: (Select all that apply) - Analyze and understand data to influence product or business decisions
TECH	Q6 For how many years have you been writing code and/or programming?
	Q11 What type of computing platform do you use most often for your data science projects?
	Q15 For how many years have you used machine learning methods?
	Q32 Which of the following business intelligence tools do you use most often?
	Q38 What is the primary tool that you use at work or school to analyze data? (Include text response)

DATASET DIMENSIONS

The original dataset has 20037 rows (the first one being the actual question).

It contains 355 columns (representing the questions and, for multiple choices, each option).

The selected dataset used to seed the analysis contains 20036 rows over 16 columns.

VALUES DISTRIBUTION

In the next few pages, the set of values available for each question within the selected subset, from most to least frequent within each question.

The questions selected have been divided in three groups, as shown in the table within the previous page: person, company, technology.

QUESTIONS ABOUT THE PERSON

Table ordered by number of answers, descending

question: Q1 *What is your age (# years)?*		*question: Q2* *What is your gender?*	
25-29	4011	Man	15789
22-24	3786	Woman	3878
18-21	3469	Prefer not to say	263
30-34	2811	Prefer to self-describe	54
35-39	1991	Nonbinary	52
40-44	1397		
45-49	988		
50-54	698		
55-59	411		
60-69	398		
70+	76		

Table ordered by number of answers, left-to-right, top-to-bottom

question: Q3 *In which country do you currently reside?*			
India	5851	USA	2237
Other	1388	Brazil	694
Japan	638	Russia	582
UK	489	Nigeria	476
China	474	Germany	404
Turkey	344	Spain	336
France	330	Canada	301
Indonesia	290	Pakistan	283
Italy	267	Taiwan	267
Australia	231	Mexico	227
South Korea	190	Egypt	179
Colombia	177	Ukraine	170
Iran, Islamic Republic of...	162	Kenya	153
Netherlands	151	Singapore	149
Poland	148	Viet Nam	147
Bangladesh	143	South Africa	141
Argentina	134	Morocco	133
Malaysia	133	Thailand	132
Portugal	122	Greece	111
Tunisia	99	Philippines	99
Israel	97	Peru	95
Chile	85	Sweden	78
Republic of Korea	76	Saudi Arabia	76
Sri Lanka	72	Switzerland	68
Nepal	62	Romania	61
Belgium	60	United Arab Emirates	59
Belarus	59	Ireland	54
Ghana	52		

Table ordered by number of answers, descending

question: Q4 *What is the highest level of formal education that you have attained or plan to attain within the next 2 years?*			
Master’s degree	7859	Bachelor’s degree	6978
Doctoral degree	2302	Some college/university study without earning a bachelor’s degree	1092
Professional degree	699	I prefer not to answer	399
No formal education past high school	240		

Table ordered by number of answers, descending

question: Q5 *Select the title most similar to your current role (or most recent title if retired)*			
Student	5171	Data Scientist	2676
Software Engineer	1968	Other	1737
Currently not employed	1652	Data Analyst	1475
Research Scientist	1174	Machine Learning Engineer	1082
Business Analyst	798	Product/Project Manager	692
Data Engineer	437	Statistician	290
DBA/Database Engineer	125		

QUESTIONS ABOUT THE COMPANY

Tables ordered by number of answers, descending

question: Q20 What is the size of the company where you are employed?			
0-49 employees	4208	10,000 or more employees	2238
1000-9,999 employees	1934	50-249 employees	1671
250-999 employees	1352		

question: Q21 Approximately how many individuals are responsible for data science workloads at your place of business?			
1-2	2645	0	2291
20+	2247	3-4	1783
5-9	1324	10-14	692
15-19	300		

question: Q22 Does your current employer incorporate machine learning methods into their business?	
We are exploring ML methods (and may one day put a model into production)	2353
No (we do not use ML methods)	2222
We have well established ML methods (i.e., models in production for more than 2 years)	1915
We recently started using ML methods (i.e., models in production for less than 2 years)	1802
I do not know	1588
We use ML methods for generating insights (but do not put working models into production)	1250

question: Q24 What is your current yearly compensation (approximate $USD)?			
$0-999	2128	10,000-14,999	665
1,000-1,999	581	100,000-124,999	573
40,000-49,999	552	30,000-39,999	540
50,000-59,999	510	5,000-7,499	488
15,000-19,999	449	60,000-69,999	408
20,000-24,999	404	70,000-79,999	394
7,500-9,999	371	150,000-199,999	347
2,000-2,999	330	125,000-149,999	315
25,000-29,999	310	90,000-99,999	280
4,000-4,999	279	80,000-89,999	273
3,000-3,999	264	200,000-249,999	115
300,000-500,000	55	> $500,000	50
250,000-299,999	48		

question: Q25 *Approximately how much money have you (or your team) spent on machine learning and/or cloud computing services at home (or at work) in the past 5 years*			
$0 ($USD)	3856	$1000-$9,999	1829
$100-$999	1764	$1-$99	1317
$10,000-$99,999	1075	$100,000 or more ($USD)	729

question: Q23_Part_1 *Select any activities that make up an important part of your role at work: - Analyze and understand data to influence product or business decisions*	
Analyze and understand data to influence product or business decisions	6421

QUESTIONS ABOUT TECHNOLOGY

The questions in section are focused on the actual use of technologies.

Tables ordered by number of answers, descending

question: Q6 *For how many years have you been writing code and/or programming?*			
3-5 years	4546	1-2 years	4505
< 1 years	3313	5-10 years	2552
10-20 years	1751	20+ years	1329
I have never written code	1124		

question: Q11 *What type of computing platform do you use most often for your data science projects?*	
A personal computer or laptop	13348
A cloud computing platform (AWS, Azure, GCP, hosted notebooks, etc)	2358
A deep learning workstation (NVIDIA GTX, LambdaLabs, etc)	834
None	292
Other	197

question: Q15 *For how many years have you used machine learning methods?*			
Under 1 year	6312	1-2 years	3459
I do not use machine learning methods	2075	2-3 years	1631
3-4 years	893	5-10 years	801
4-5 years	784	10-20 years	244
20 or more years	175		

question: Q32 *Which of the following business intelligence tools do you use most often?*			
Tableau	540	Microsoft Power BI	462
Google Data Studio	167	Qlik	69
Other	57	Salesforce	51
Amazon QuickSight	36	SAP Analytics Cloud	31
Alteryx	27	TIBCO Spotfire	22
Looker	19	Einstein Analytics	6
Sisense	6	Domo	5

question: Q38 *What is the primary tool that you use at work or school to analyze data? (Include text response)*	
Local development environments (RStudio, JupyterLab, etc.)	6107
Basic statistical software (Microsoft Excel, Google Sheets, etc.)	4223
Business intelligence software (Salesforce, Tableau, Spotfire, etc.)	798
Advanced statistical software (SPSS, SAS, etc.)	781
Other	695
Cloud-based data software & APIs (AWS, GCP, Azure, etc.)	686

Table ordered by number of answers, descending

DIMENSION: JOB TITLES

These are the **job titles that have been considered in this analysis**:

- 'Business Analyst'
- 'Product/Project Manager'

DIMENSION: GEOGRAPHIC COVERAGE

While the initial analysis considers all the countries covered within the Kaggle survey data, to compare the current status, three geographical areas with roughly the same sample size across the job titles selected:

- USA
- India
- Europe (from the Atlantic to the Urals, i.e. including EU, UK, former USSR European countries)

5 DATA PREPARATION

This section is really short within the book, as it presents just the results of the data preparation, i.e. the values selected.

Within the Jupyter Notebook, instead there is the data selection logic.

Also in this case the choice was to enable easier maintenance, and therefore this section builds on the data selections and structuring done from the section ***B. Assumption and Choices*** and section ***C. Dimensions of Analysis***.

In a real business project, as this is where data are prepared for the section E. Data Visualization that follows, I would also add any data restructuring, KPI creation, KPI analysis, etc.

Again, the concept is “layering” the report, to reduce the number of changes.

As for the previous section, defining data transformations and KPIs is almost never a straightforward process.

Often it is a matter of trial-and-error.

In some cases, only at the end of a first round of analysis you might discover that a KPI (Key Performance Indicator) does not differentiate enough within your data selection.

It is normal sometimes to have to get back to revising previous sections- with the agreement of those involved.

D. DATA PREPARATION

As discussed in the previous chapter, the analysis is focused on two job titles (Business Analyst and Product/Project Manager) and three countries (Europe as defined in the previous chapter, India, USA).

DATA BY COUNTRY

The column "BA & PM" contains the number of respondents with job title *Business Analyst* or *Product/Project Manager.*

The column percent is

column percent = column ba_pm / column answers

Table ordered by number of answers, descending

Country	*Answers*	*BA & PM*	*%*
India	5851	327	5.59
USA	2237	224	10.01
Other	1388	98	7.06
Brazil	694	70	10.09
Japan	638	62	9.72
Russia	582	61	10.48
UK	489	45	9.20
Nigeria	476	16	3.36
China	474	23	4.85
Germany	404	33	8.17
Turkey	344	17	4.94
Spain	336	27	8.04
France	330	38	11.52
Canada	301	26	8.64
Indonesia	290	15	5.17
Pakistan	283	9	3.18
Taiwan	267	22	8.24
Italy	267	36	13.48
Australia	231	21	9.09

Country	*Answers*	*BA & PM*	*%*
Mexico	227	26	11.45
South Korea	190	9	4.74
Egypt	179	4	2.23
Colombia	177	21	11.86
Ukraine	170	11	6.47
Iran	162	6	3.70
Kenya	153	7	4.58
Netherlands	151	17	11.26
Singapore	149	12	8.05
Poland	148	11	7.43
Viet Nam	147	11	7.48
Bangladesh	143	2	1.40
South Africa	141	6	4.26
Argentina	134	19	14.18
Malaysia	133	9	6.77
Morocco	133	2	1.50
Thailand	132	13	9.85
Portugal	122	13	10.66
Greece	111	9	8.11
Tunisia	99	5	5.05
Philippines	99	11	11.11
Israel	97	4	4.12
Peru	95	13	13.68
Chile	85	13	15.29
Sweden	78	9	11.54
Saudi Arabia	76	9	11.84
Republic of Korea	76	10	13.16
Sri Lanka	72	2	2.78
Switzerland	68	4	5.88
Nepal	62	1	1.61
Romania	61	---	---
Belgium	60	4	6.67
United Arab Emirates	59	11	18.64
Belarus	59	4	6.78
Ireland	54	8	14.81
Ghana	52	4	7.69

FOCUSING ON EUROPE, INDIA, USA

Country	*Answers*	*BA & PM*	*%*
India	5851	327	5.59
Europe	3834	347	9.05
USA	2237	224	10.01

For the purpose of this analysis, the geographical coverage considered for Europe is the following:

Country	*Answers*	*BA & PM*	*%*
Russia	582	61	10.48
UK	489	45	9.20
Germany	404	33	8.17
Turkey	344	17	4.94
Spain	336	27	8.04
France	330	38	11.52
Italy	267	36	13.48
Ukraine	170	11	6.47
Netherlands	151	17	11.26
Poland	148	11	7.43
Portugal	122	13	10.66
Greece	111	9	8.11
Sweden	78	9	11.54
Switzerland	68	4	5.88
Romania	61	---	---
Belgium	60	4	6.67
Belarus	59	4	6.78
Ireland	54	8	14.81

Table ordered by number of answers, descending

6 DATA VISUALIZATION

For this sample report, the data visualization section was important, but not really that complex.

What was more complex was to identify a single, unique way of representing different dimensions of analysis ((in this case, job titles and countries) across the subset of questions ("features") selected.

The approach used was to simply try different options, and then see which one seemed more promising.

As the purpose of this report was to be a fictional preliminary report following the delivery of the Kaggle Survey Executive Summary, for the same audience, I did apply some constraints that seem unusual, but are realistic in business:

- **time constraint**- I will not say how much time "cost" the whole report, but I would just say that thinking about the data took more then preparing the report- on purposes
- **intuitive access to information**- except for an histogram, a bar chart, and some heat maps (that are anyway common nowadays in business for risk management and assessing impacts during a crisis or to see the level of compliance available with different scenarios), I disposed of all the usual paraphernalia, and used the humble radar chart, always presenting the three countries as an aggregate, and then each individual country.

Within the Jupyter Notebook, the process is automated: call a function and will produce the different radar charts (courtesy of the formatting work done in section ***D. Data Preparation***).

Within this book, for the sake of keeping it as short as possible, only where really needed I shared both the “aggregate” radar chart and those for each country.

When a section is composed anyway of steps, I prefer to “guide through the development process”, i.e. to share also the steps followed, and the status of each step.

In this case, the section was divided in 7 further “subsections” (the last one being always a summary).

For this first example, decided also to leave inside a couple of examples of “negative choices”, i.e. preliminary checks through visualization that resulted in the removal of some questions.

In each step, I also added a textual summary (“Results”) followed by an indication of what was the next step.

I used visualization as a way to deliver this preliminary analysis, as writing what was clearly shown by the radar chart as a descriptive detailed analysis would have just added pages, but not that much in terms of analysis.

As described in previous chapters, decided to use Plotly for some of the radar charts, mainly to show how interactive charts would allow to do what, in the past, usually delivered with a static radar charts and few Powerpoint slides.

The only caveat, that I also shared online, is that I used Plotly with “open data”, i.e. information that is neither confidential nor restricted.

In this case, having just 1-to-3 “areas” to visualize, across multiple options, even a static radar chart would be fine.

Personally, in the past saw that up to 7-8 different areas can be visualized and explained, if the presentation approach is consistent and there is not too much “information overload”.

The latter is when you try to present too much in a single chart.

Might be funny and a showcase of your own skills (actually, with most Jupyter Notebook charts, the “skills” really amount to googling to find something close to what you need, and then tweaking it).

In a business presentation, having a collection of slides where each slide requires few minutes to explain how to read a single chart is disruptive, and defies the purpose of moving from tables to moving to charts.

In my biased view, a report containing visualizations should be something that anybody with knowledge on the business domain should be able to browse and read with minimal reference to legenda etc.

Litmus test: if you explain acronyms in column titles, might make sense if you share the same explanation across, as in business report usually should not be expected a sequential reading- decision-makers, notably when they have deep business domain expertise, are inclined to read following their our “storyline”, not your own.

This is a lesson that you learned quickly when worked in the 1980s on Executive Information Systems, that often were composed of dozens of screens that defined a specific reading sequence.

A decision-maker went through, and asked you a question whose answer was few charts down the road, and twisted in another way.

Other types of reports are more “sequential”, but I am not focusing on those- as I wrote since the beginning of this book, my purpose is to use Jupyter Notebook files to produce report that could be equivalent to those I produced in the past.

Thanks to the integration of data and logic, anybody could actually modify a Jupyter Notebook report to add data, add KPIs, modify logic.

Provided that all is:

- intuitive whenever possible
- using the same approach for the same results, and hopefully having a single place to align if you want to replicate an existing analysis
- last but not least: documented: every choice, every option should have an explanation, if it is not self-explanatory.

As for the explanatory (i.e. textual) part, delivered via markdown cells: if it contains information extracted from data, should be as much as possible generated from data, not copied by hand from data, to ensure easier update, and avoid the risk of leaving behind obsolete data.

E. DATA VISUALIZATION

Data visualization is a storyline within the storyline.

It is a series of steps to increase understanding and see how far I could go in my research with the available data.

Generally, E1-E2 are about overall distribution, while E3-E6 are about comparatively profiling countries.

For the former, I selected the usual (tables, histograms), for the latter a comparison tool that I used since the early 2000s in organizational change activities, the radar chart.

While many dislike it, I found it useful to compare patterns between organizations and also in purchasing and quality/audit activities.

Anyway, the structure and functions (as well as the dataframes created above with self-explanatory names and comments) should enable altering the logic, e.g. to selectively adopt different visualizations for different questions.

My target is an hypothetical request from an hypothetical business customer asking questions about presence/absence of people with those job titles, and the mix of corporate and technological environments they work in, as well as demographic differences between the three countries (Europe, India, USA).

The initial idea was to have a radar chart comparing the three countries, to ease a visual inspection of the report,

As even the first question showed significant differences in demographics, whenever appropriate, decided to create, in each question, both the aggregate radar (as significant differences are an information per se), as well as individual radar charts by country.

The data visualization / analysis part is divided in sever “steps” (subsections), each one focused on a specific assumptions to verify, and leading the continuation on the others.

Within each “step”, one or more questions, or even the whole dataset, as considered.

As I wrote in the previous page, the main visualization adopted is a radar chart, as it allows to compare multiple elements across the three countries, but in some other cases (e.g. correlation analysis) a more traditional visualization was used (histogram, bar chart, heat map).

This book is “pie-free”: I decided on purpose to refrain from using the ubiquitous pie chart, as the purpose of this analysis is to compare the three countries across multiple elements of analysis, not as parts of a whole.

This table contains the steps within the data analysis, the column **OK/ NOK** states if the results of that step supported my original question (*identifying some potential trends on the bridging between data science and business, notably on the management side).*

Step	*Purpose*	*OK/ NOK*
E1. Check BA/PM frequency by country	verify from percentages if could make sense to check Europe as an aggregate, vs. India and USA	OK
E2. Confirm comparison countries	verify if, be aggregating countries in Europe, there is a unit comparable in size to India and USA	OK
E3. Check decision influencers	check if the distribution of E2. matches that of influencers	NOK
E4. Compare demographic distribution	check, for BAs/PMs, how Europe, India, USA compare	OK
E5. Compare companies characteristics	check, for BAs/PMs, how Europe, India, USA compare	OK
E6. Compare technical characteristics	check, for BAs/PMs, how Europe, India, USA compare	OK
E7. Assess results and analysis	summarize results from the commentary in E1-E6	OK
E8. Further investigations planned	share ideas about other areas of analysis, or new questions	OK

E1. Check BA/PM frequency by country

Lacking some further information (e.g. since how long a data science team has been created in organizations in each country), some answers cannot be directly obtained from data.

It is anyway more interesting to consider the distribution of those titles across countries (ordered by frequency, left-to-right, top-to-bottom):

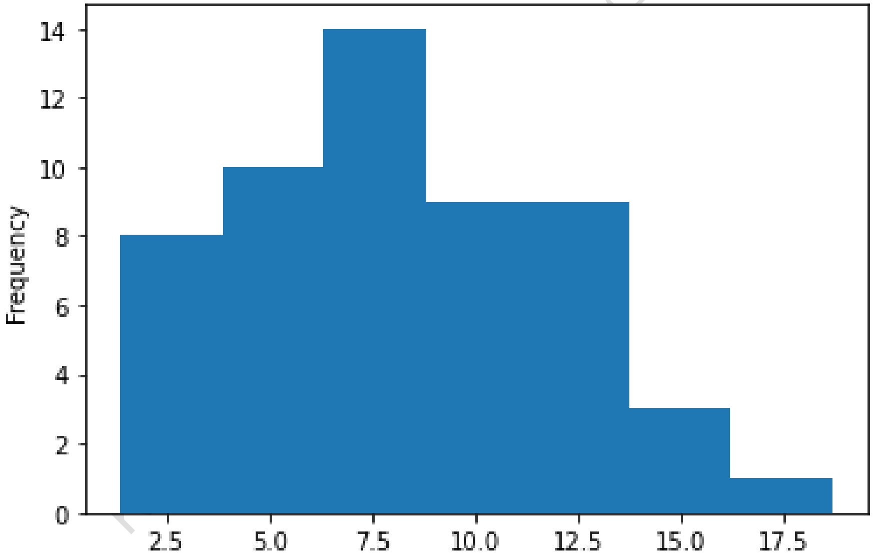
Country	*%*	*Country*	*%*
United Arab Emirates	18.64	Chile	15.29
Ireland	14.81	Argentina	14.18
Peru	13.68	Italy	13.48
Republic of Korea	13.16	Colombia	11.86
Saudi Arabia	11.84	Sweden	11.54
France	11.52	Mexico	11.45
Netherlands	11.26	Philippines	11.11
Portugal	10.66	Russia	10.48
Brazil	10.09	USA	10.01
Thailand	9.85	Japan	9.72
UK	9.20	Australia	9.09
Canada	8.64	Taiwan	8.24
Germany	8.17	Greece	8.11
Singapore	8.05	Spain	8.04
Ghana	7.69	Viet Nam	7.48
Poland	7.43	Other	7.06
Belarus	6.78	Malaysia	6.77
Belgium	6.67	Ukraine	6.47
Switzerland	5.88	India	5.59
Indonesia	5.17	Tunisia	5.05
Turkey	4.94	China	4.85
South Korea	4.74	Kenya	4.58
South Africa	4.26	Israel	4.12
Iran, Islamic Republic of of...	3.70	Nigeria	3.36
Pakistan	3.18	Sri Lanka	2.78
Egypt	2.23	Nepal	1.61
Morocco	1.50	Bangladesh	1.40
Romania	---		

The percentage of Business Analysts and Product/Project Managers in each country could be an indicator of the focus and team size.

Also, could be an indicator of how much those answering are involved in bringing new products or services on the market.

A lower number of project managers might both imply smaller teams, e.g. delivering outsourced data science services on-demand, or even just experimenting.

CLUSTERING OF COUNTRIES BY PERCENTAGE OF PM-BA ON NUMBER OF RESPONDENTS



Results:

- there is a distribution with differences between countries
- multiple countries differ by level of presence of BAs and Product/Project Managers between respondents
- it could make sense to check other dimensions

Next step:

- checking if the aggregations Europe, India, USA have a comparable size

E2. Confirm comparison countries

As stated above, the intent is to focus the analysis on three continental-level countries, to identify is there are different patterns vs. the other questions selected for this report.

These are the three candidates:

- Europe
- India
- United States of America

The question in this step is if the resulting size is comparable.

As the survey results do not cover all the Member States of the European Union, selected instead a concept of Europe closer to "from Atlantic to Urals", and including also countries that are not part of the European Union (yet).

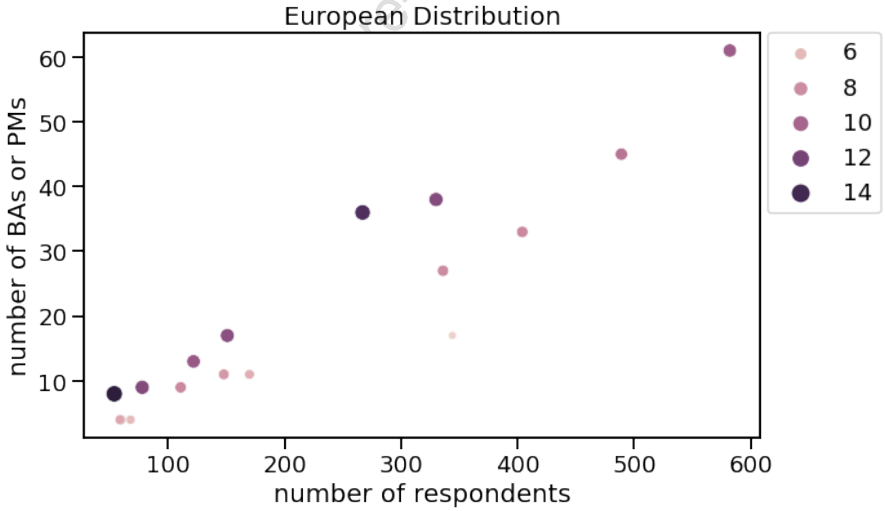
Which countries are considered to be Europe, for the purpose of this report, between those individually available in the dataset?

As you can see, considered geographical and business integration, as e.g. Turkey is integrated within the European supply chains, and the same applies (in other industries, e.g. energy) with Russia and Ukraine.

Note: as I have contacts on Kaggle in other European countries that are missing from the list (and from the data), I assume that either they are within "Other", or that have not answered to the survey.

Country	*Answers*	*BA & PM*	*%*
Belarus	59	4	6.78
Belgium	60	4	6.67
France	330	38	11.52
Germany	404	33	8.17
Greece	111	9	8.11
Ireland	54	8	14.81
Italy	267	36	13.48
Netherlands	151	17	11.26
Poland	148	11	7.43
Portugal	122	13	10.66
Romania	61	---	---
Russia	582	61	10.48
Spain	336	27	8.04
Sweden	78	9	11.54
Switzerland	68	4	5.88
Turkey	344	17	4.94
Ukraine	170	11	6.47
UK	489	45	9.2

Table alphabetically ordered by country



The chart shows that there is not direct correlation between the number of respondents in each European country and the percentage of Business Analysts and Product/Project Managers (represented by the size and hue of each dot).

Country	*Answers*	*BA & PM*	*%*
India	5851	327	5.59
Europe	3834	347	9.05
USA	2237	224	10.01
Europe components			
Russia	582	61	10.48
UK.	489	45	9.20
Germany	404	33	8.17
Turkey	344	17	4.94
Spain	336	27	8.04
France	330	38	11.52
Italy	267	36	13.48
Ukraine	170	11	6.47
Netherlands	151	17	11.26
Poland	148	11	7.43
Portugal	122	13	10.66
Greece	111	9	8.11
Sweden	78	9	11.54
Switzerland	68	4	5.88
Romania	61	---	---
Belgium	60	4	6.67
Belarus	59	4	6.78
Ireland	54	8	14.81

Table ordered by number of answers, descending

Results:

- within European countries, there is no direct correlation between number of respondents and presence of BAs/PMs
- aggregating Europe as defined above, it is comparable in size and percentage of BAs/PM with India and USA
- it could make sense to check other dimensions

Next step:

- working on the influencers distribution

E3. Check decision influencers

For the purpose of this report, a single part of Question 23, the one focused on a potential role by the respondent in influencing product and business choices, was selected.

The table under ***E2. Confirm comparison countries*** summarizes the distribution of the job titles target of this report, country by country.

There was a question within the survey (*Question 23: "Select any activities that make up an important part of your role at work"*), that was a multiple choice.

The first option, of interest here, was within the question Q23_Part_1:

> Q23 Part 1: Select any activities that make up an important part of your role at work: (Select all that apply) - Analyze and understand data to influence product or business decisions

From the distribution above, you could assume that countries with a lower distribution of the roles of targe, business analysts and product/project managers, could have a lower number of "influencers", but this is not the case:

- respondents to the survey: 20036
- number of influencers (as per answers to the question): 6421

So, it would seem as if a whopping 32% of Kaggle members consider themselves influences on product or business decisions.

In the next few pages, a table with more details, for all the countries.

Due to the number of columns, I had to shorten the title, while removed also the decimals from the column “% on BA & PM”, that contains the percentage of reported influencers over the number of Business Analysts and Product/Project Managers: they are all 3-digits percentages, therefore decimals are irrelevant (rounded up to next unit if above 0.50).

Just in case:

- “Answ.” = “Answers”
- “Infl.” = “Influencers”
- “% Infl. on Total” = “% Influencers on Total”

Table ordered by number of answers, descending

Country	*Answ.*	*BA & PM*	*%*	*Infl.*	*% on BA & PM*	*% Infl. on Total*
India	5851	327	5.59	1347	412	23.02
USA	2237	224	10.01	1022	456	45.69
Other	1388	98	7.06	500	510	36.02
Brazil	694	70	10.09	286	409	41.21
Japan	638	62	9.72	183	295	28.68
Russia	582	61	10.48	182	298	31.27
UK	489	45	9.20	240	533	49.08
Nigeria	476	16	3.36	166	1038	34.87
China	474	23	4.85	95	413	20.04
Germany	404	33	8.17	166	503	41.09
Turkey	344	17	4.94	77	453	22.38
Spain	336	27	8.04	127	470	37.80
France	330	38	11.52	109	287	33.03
Canada	301	26	8.64	138	531	45.85
Indonesia	290	15	5.17	80	533	27.59
Pakistan	283	9	3.18	56	622	19.79
Taiwan	267	22	8.24	59	268	22.10
Italy	267	36	13.48	94	261	35.21
Australia	231	21	9.09	92	438	39.83
Mexico	227	26	11.45	96	369	42.29

Table ordered by number of answers, descending

South Korea	190	9	4.74	45	500	23.68
Egypt	179	4	2.23	49	1225	27.37
Colombia	177	21	11.86	80	381	45.20
Ukraine	170	11	6.47	60	545	35.29
Iran	162	6	3.70	30	500	18.52
Kenya	153	7	4.58	61	871	39.87
Netherlands	151	17	11.26	84	494	55.63
Singapore	149	12	8.05	50	417	33.56
Poland	148	11	7.43	57	518	38.51
Viet Nam	147	11	7.48	54	491	36.73
Bangladesh	143	2	1.40	28	1400	19.58
South Africa	141	6	4.26	58	967	41.13
Argentina	134	19	14.18	44	232	32.84
Malaysia	133	9	6.77	33	367	24.81
Morocco	133	2	1.50	38	1900	28.57
Thailand	132	13	9.85	37	285	28.03
Portugal	122	13	10.66	52	400	42.62
Greece	111	9	8.11	36	400	32.43
Tunisia	99	5	5.05	17	340	17.17
Philippines	99	11	11.11	36	327	36.36
Israel	97	4	4.12	36	900	37.11
Peru	95	13	13.68	38	292	40.00
Chile	85	13	15.29	36	277	42.35
Sweden	78	9	11.54	33	367	42.31
Saudi Arabia	76	9	11.84	26	289	34.21

Table ordered by number of answers, descending

Country	*Answ.*	*BA & PM*	*%*	*Infl.*	*% on BA & PM*	*% Infl. on Total*
Republic of Korea	76	10	13.16	17	170	22.37
Sri Lanka	72	2	2.78	21	1050	29.17
Switzerland	68	4	5.88	29	725	42.65
Nepal	62	1	1.61	12	1200	19.35

Romania	61	---	---	18	---	29.51
Belgium	60	4	6.67	16	400	26.67
United Arab Emirates	59	11	18.64	25	227	42.37
Belarus	59	4	6.78	12	300	20.34
Ireland	54	8	14.81	23	288	42.59
Ghana	52	4	7.69	15	375	28.85

Results:

- the purpose of using this question was to complement the job titles (Q5)
- from the results, there is no direct relationship between being a Business Analyst or Product/Project Manager and influencing product or business decision
- this is a theme that is worth investigating, but would require further data
- therefore, the **question is excluded from the report**, as probably its results are obfuscated by both the interpretations of the question by respondents, and other information

Next step:

- working on the demographic distribution

E4. Compare demographic distribution

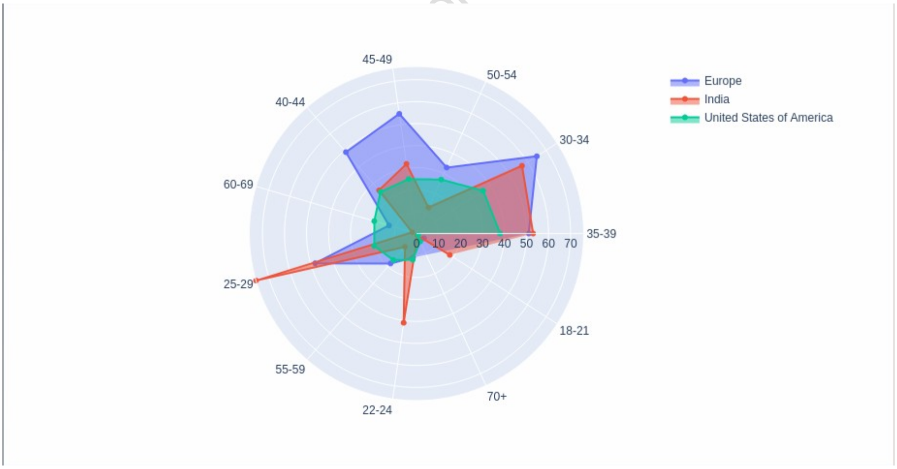
In this section, just looked at how the three "countries" (Europe, India, USA) compared in terms of demographics, across five variables:

PERSON	Q1	What is your age (# years)?
	Q2	What is your gender?
	Q3	In which country do you currently reside?
	Q4	What is the highest level of formal education that you have attained or plan to attain within the next 2 years?
	Q5	Select the title most similar to your current role (or most recent title if retired):

Question: Q1 What is your age (# years)?

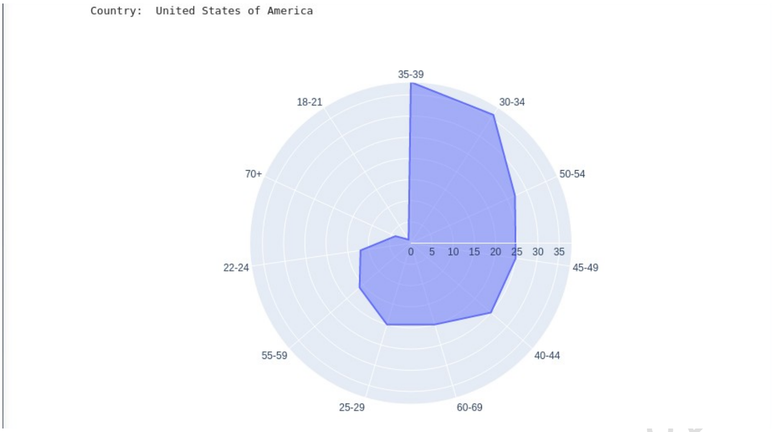
Table ordered by categorical order (age range)

Age range	*Europe*	*India*	*USA*
18-21	4	18	1
22-24	11	41	12
25-29	48	76	20
30-34	65	57	36
35-39	51	53	38
40-44	49	26	25
45-49	55	32	25
50-54	33	13	27
55-59	18	8	16
60-69	13	2	20
70+	---	1	4

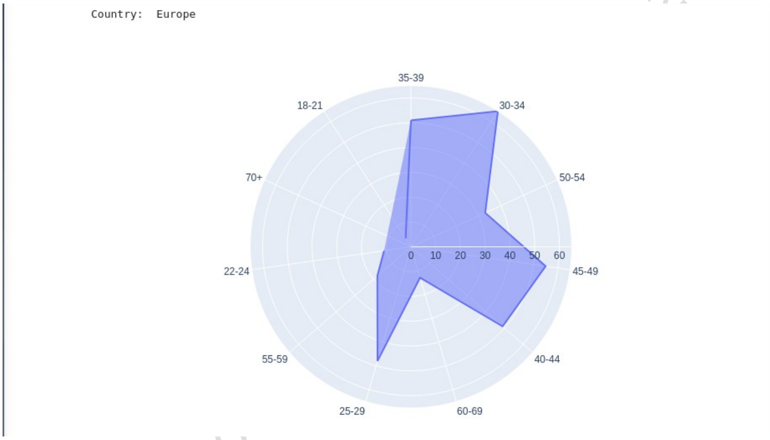


The radar chars shows that each country has a different patterns.

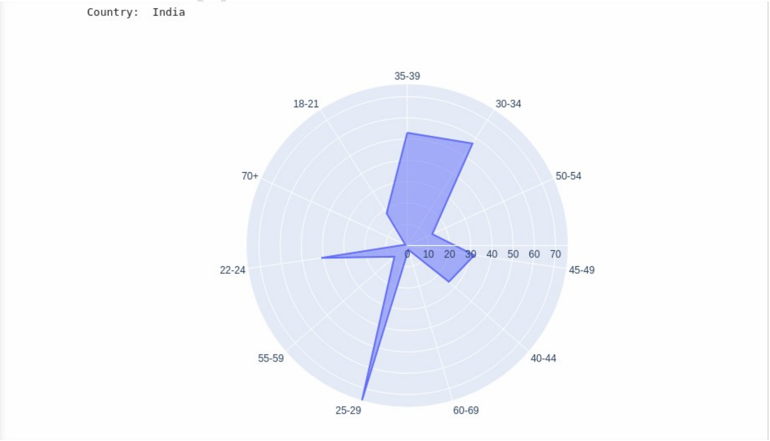
Considering the number of options, the next pages shows the radar chart for each country, that can be visually compared.



```
Country:  United States of America
```



```
Country:  Europe
```

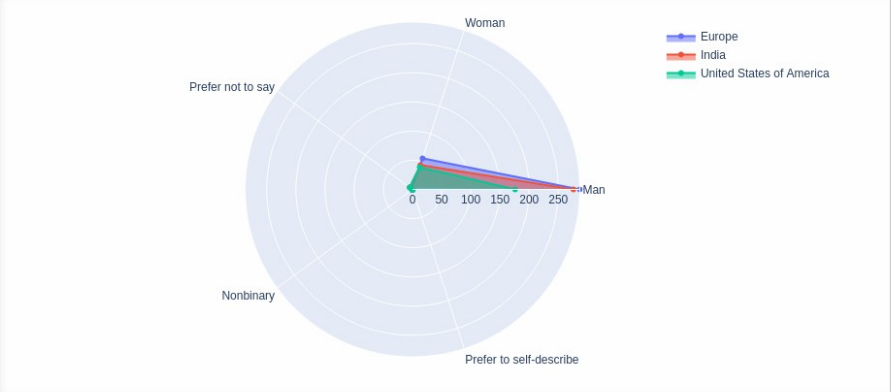


```
Country:  India
```

Question: Q2 What is your gender?

Table ordered by number of answers, descending

Gender	*Europe*	*India*	*USA*
Man	287	276	176
Woman	56	44	40
Prefer not to say	2	6	6
Nonbinary	1	---	1
Prefer to self-describe	1	1	1



The focus of this report is just on roles that usually are associated with higher frequency of interaction with decision makers.

The Executive Summary report published by Kaggle on the whole survey about gender differences overall states (page 5): *"Data science is still suffering from a large gender gap in the workplace, as 82% of users identify as men. This is only a slight change from last year's results, where 84% of users identified as males. This is the first year we've differentiated between "Nonbinary" and "Prefer to self-describe," with each answer coming in around a third of a percent."*

As you can see from both the table above and the radar chart, also for the Kaggle community, the higher you go within organizations, the higher the gender gap.

And this difference shows little variation between the three countries.

Question: Q3 In which country do you currently reside?

Table ordered by number of answers, descending

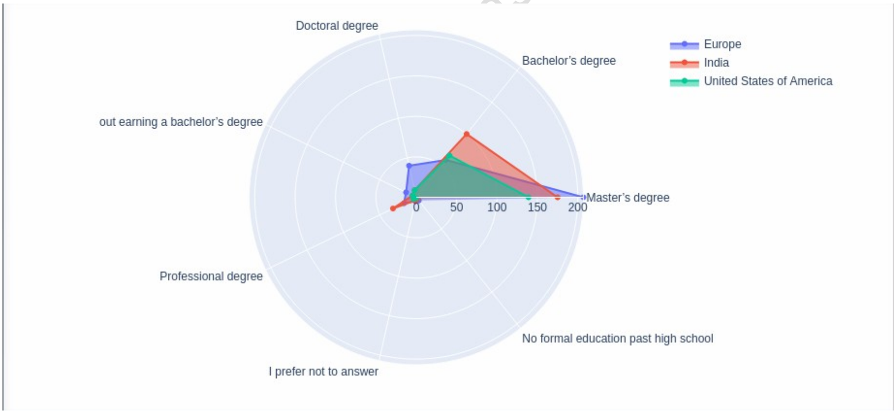
Country	*Answers*	*BA & PM*	*%*
Europe	3834	347	9.05
India	5851	327	5.59
USA	2237	224	10.01

The table above contains an extract of the same data already shown in sections *E1. Check BA/PM frequency by country* and *E2. Confirm comparison countries*.

Question: Q4 What is the highest level of formal education that you have attained or plan to attain within the next 2 years?

Table ordered by categorical order (level of formal education, descending)

Level of formal education	*Europe*	*India*	*USA*
Doctoral degree	40	9	9
Master’s degree	207	175	139
Bachelor’s degree	59	100	66
Some college/university study without earning a degree	14	5	5
Professional degree	17	32	4
I prefer not to answer	5	4	1
No formal education past high school	5	2	---



In terms of formal education, there is little difference between the three countries: for the job titles selected, a master's degree (and with lesser but still significant frequency, a bachelor's degree) in both India and USA are common, while a PhD has a non negligible presence in Europe.

Question: Q5 Select the title most similar to your current role (or most recent title if retired)

For the question about jobtitles, Q5, having just two values, a radar chart would be useless:

Job title	*Europe*	*India*	*USA*
Business Analyst	149	200	107
Product/Project Manager	198	127	117

As can be seen from the table above, both Europe and United States of America have slightly more Product/Project Managers within the respondents than Business Analysts, while in India there is a larger proportion of Business Analysts.

Results:

- the gender gap shown within the Kaggle Executive Summary of the whole survey is even stronger when focusing on the Business Analyst and Product/Project Manager job titles
- it would be worth investigating why in India the profile seems to closer to a standard IT project, i.e. more business analysts than product/project managers, while in Europe is just the opposite: might be due to a gap in knowledge (hence, more interest in joining Kaggle) or other factors related to market structure
- it is anyway interesting to note how in Europe a PhD seems to be almost as common as a Bachelor's degree, while in all the three countries anyway it a Master's degree the most common educational level

Next step:

- working on companies characteristics

E5. Compare companies characteristics

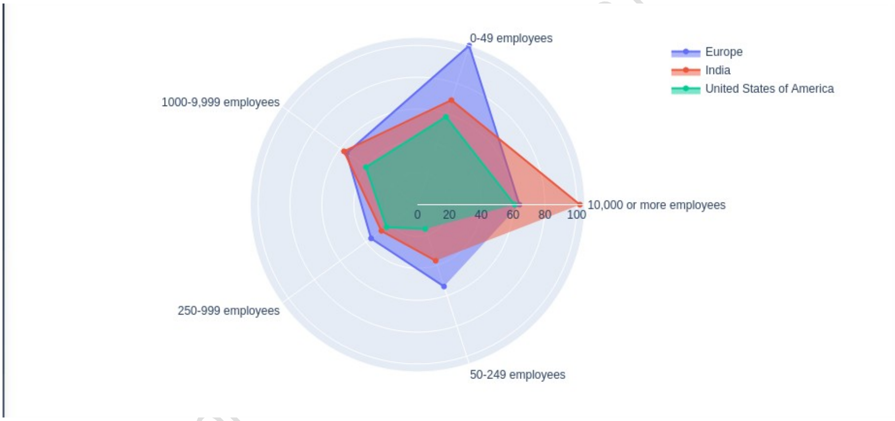
In this section, just looked at how the three "countries" (Europe, India, USA) compared in terms of company characteristics, across five variables:

COMPANY	Q20 What is the size of the company where you are employed?
	Q21 Approximately how many individuals are responsible for data science workloads at your place of business?
	Q22 Does your current employer incorporate machine learning methods into their business?
	Q24 What is your current yearly compensation (approximate $USD)?
	Q25 Approximately how much money have you (or your team) spent on machine learning and/or cloud computing services at home (or at work) in the past 5 years (approximate $USD)?
	Q23_Part_1 Select any activities that make up an important part of your role at work: (Select all that apply) - Analyze and understand data to influence product or business decisions

Question: Q20 What is the size of the company where you are employed?

Table ordered by categorical order (company size)

Company size	*Europe*	*India*	*USA*
10000 or more employees	64	102	61
1000-9999 employees	55	57	40
250-999 employees	36	28	24
50-249 employees	54	37	16
0-49 employees	105	69	58



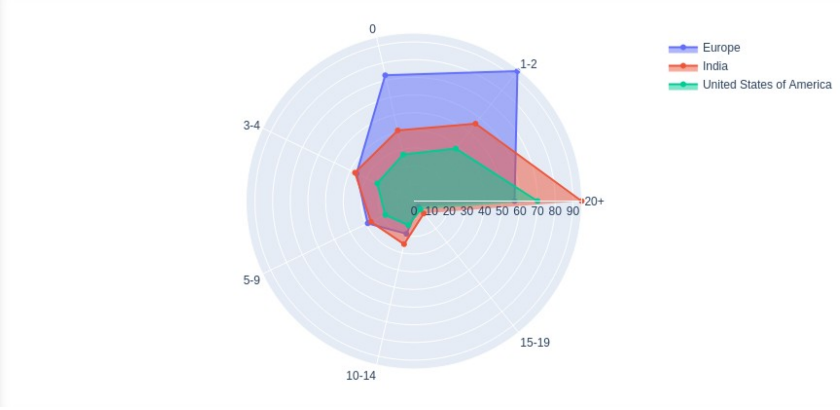
More than the table, it is the radar chart that clearly shows the difference between the responders.

While those from USA are spread across company sizes, in Europe the focus of those on Kaggle answering is predominanty the typical small company (0-49), while India is predominantly represented (again, just for these two job titles, ***Business Analyst*** and ***Product/Project Manager***) by employees of very large companies.

Question: Q21 Approximately how many individuals are responsible for data science workloads at your place of business?

Table ordered by categorical order (number of individuals)

Team size	*Europe*	*India*	*USA*
20+	57	95	70
15-19	4	9	6
10-14	19	25	14
5-9	29	27	18
3-4	36	37	23
1-2	94	56	38
0	73	41	27

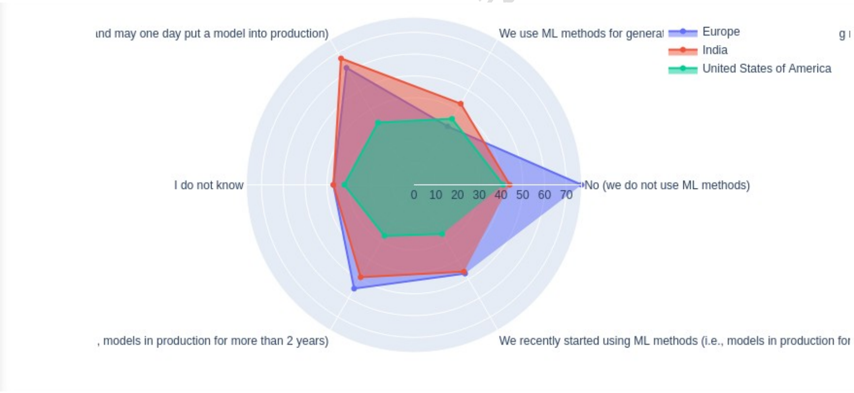


On how many individuals are directly responsible for data science workloads, both India and USA are oriented toward 20 or more, while Europe, confirming what was shown by the previous question, is geared toward smaller number- predominantly 1 or 2 people per company.

Question: Q22 Does your current employer incorporate machine learning methods into their business?

Table ordered by categorical order (level of ML methods business use)

Use of ML methods	*Europe*	*India*	*USA*
We have well established ML methods	55	49	27
We use ML methods for generating insights	31	43	35
We recently started using ML methods	47	46	26
We are exploring ML methods	62	67	33
No (we do not use ML methods)	77	44	41
I do not know	37	37	32



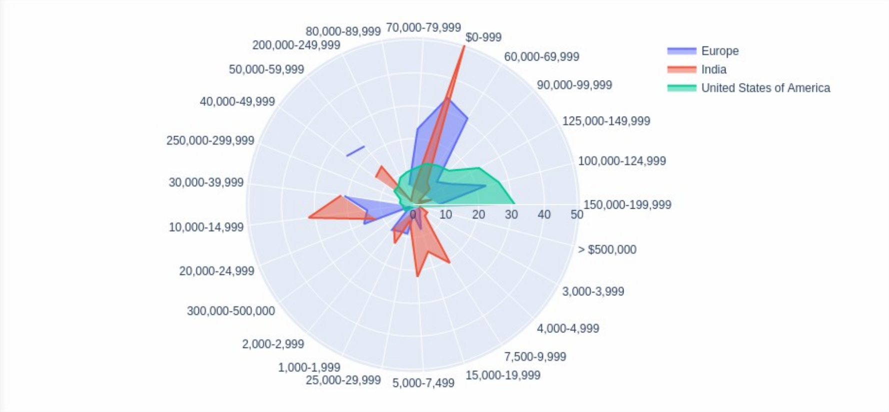
In terms of actual use within business of Machine Learning methods, USA is more evenly spread, while in both India and Europe the focus is on sperimentation ("recently started" and "exploring").

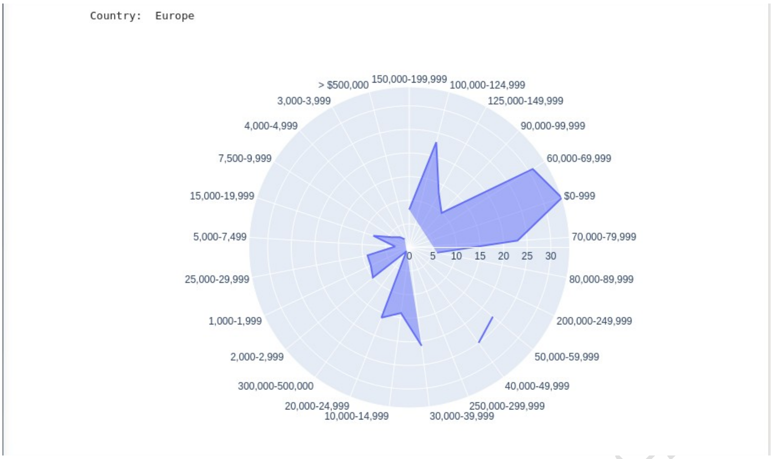
Also, Europe confirms its lag in actual integration of machine learning in business, having almost the double of "No (we do not use ML methods)" than both India and USA.

Question: Q24 What is your current yearly compensation (approximate $USD)?

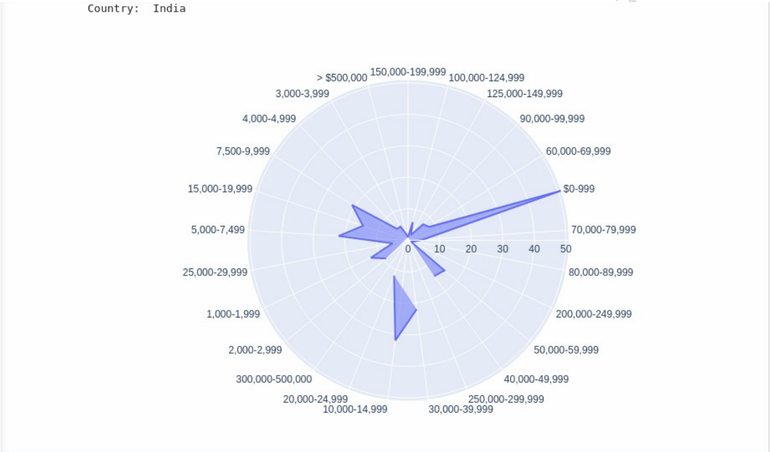
Table ordered by categorical order (yearly compensation, descending)

Yearly compensation	*Europe*	*India*	*USA*
> $500000	---	2.0	---
$300000-500000	1.0	---	3
$250000-299999	---	---	4
$200000-249999	---	1.0	9
$150000-199999	8.0	1.0	31
$125000-149999	13.0	2.0	23
$100000-124999	23.0	6.0	27
$90000-99999	10.0	7.0	15
$80000-89999	6.0	2.0	10
$70000-79999	23.0	5.0	11
$60000-69999	31.0	8.0	14
$50000-59999	23.0	15.0	7
$40000-49999	25.0	14.0	7
$30000-39999	21.0	22.0	4
$25000-29999	9.0	5.0	1
$20000-24999	16.0	12.0	3
$15000-19999	8.0	15.0	---
$10000-14999	14.0	32.0	3
$7500-9999	4.0	21.0	---
$5000-7499	3.0	22.0	1
$4000-4999	3.0	5.0	---
$3000-3999	2.0	5.0	---
$2000-2999	10.0	9.0	2
$1000-1999	9.0	13.0	1
$0-999	34.0	51.0	13

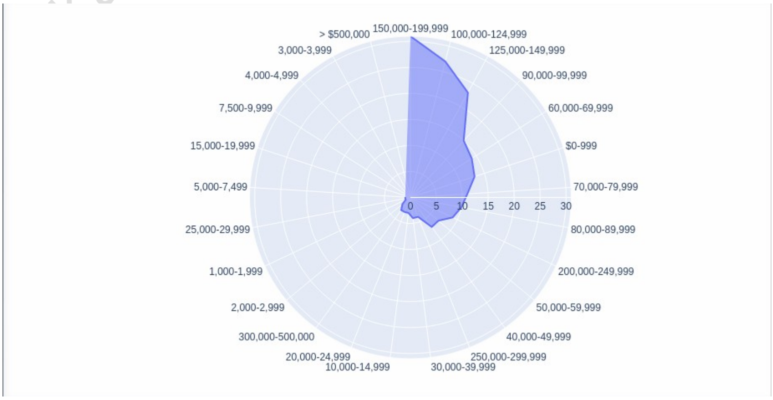




```
Country:  Europe
```



```
Country:  India
```

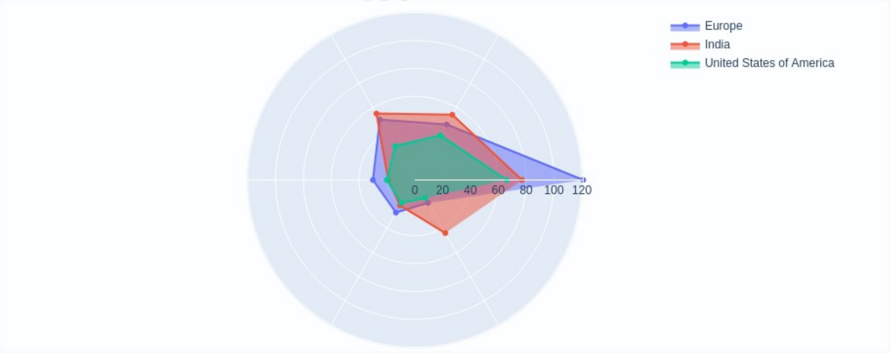


As expected, the differences in yearly compensation would require integrating with further information, e.g. if the activities are internal or directed toward external customers.

Question: Q25 Approximately how much money have you (or your team) spent on machine learning and/or cloud computing services at home (or at work) in the past 5 years (approximate $USD)?

Table ordered by categorical order (budget level, descending)

Budget level	*Europe*	*India*	*USA*
$100000 or more	27	21	19
$10000-99999	30	19	20
$1000-9999	46	54	37
$100-999	50	55	28
$1-99	19	44	15
$0 ($USD)	121	77	66



This question actually merges multiple dimensions:

- home and work
- machine learning and cloud computing

which makes difficult to consider or identify "why", i.e. the actual allocation of costs toward business ends (e.g. technological alignment, digital transformation, migration to the cloud, new services, etc.)

Results:

- the questions in this group are actually those that would benefit more of a merger of the database of this survey and the survery from McKinsey
- what the data show, is that, within the Business Analyst or Product/Project Manager responders, there is a difference between countries, both in terms of structure and, probably, aim (external, i.e. for customers, vs. internal, i.e. business development)
- blended with the answers within the previous sections, could actually influence some considerations (discussed later in this report)
- the information about expenditure on machine learning and/or cloud services as well as the one on remuneration probably would require additional questions to clarify what they really represent

Next step:

- working on the technical characteristics

E6. Compare technical characteristics

In this section, just looked at how the three "countries" (Europe, India, USA) compared in terms of technical characteristics, across five variables:

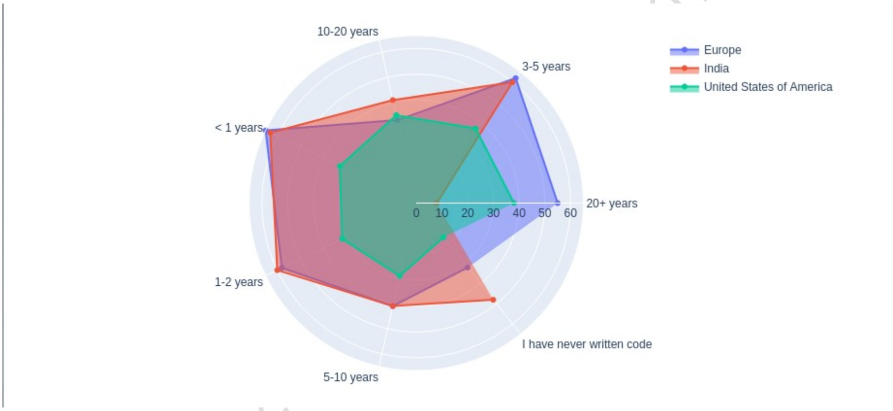
TECH	Q6 For how many years have you been writing code and/or programming?
	Q11 What type of computing platform do you use most often for your data science projects?
	Q15 For how many years have you used machine learning methods?
	Q32 Which of the following business intelligence tools do you use most often?
	Q38 What is the primary tool that you use at work or school to analyze data? (Include text response)

The questions selected for this section are mainly to cross-reference potential doubts deriving from the previous sections, e.g. on demographic, distribution of the two roles selected (***Business Analyst*** and ***Product/Project Manager***), and information about the business the respondents are working for.

Question: Q6 For how many years have you been writing code and/or programming?

Table ordered by categorical order (years of experience, descending)

Years of experience	*Europe*	*India*	*USA*
20+ years	55	8	38
10-20 years	33	41	35
5-10 years	41	41	29
3-5 years	62	60	37
1-2 years	58	60	32
< 1 years	65	63	33
I have never written code	32	48	17



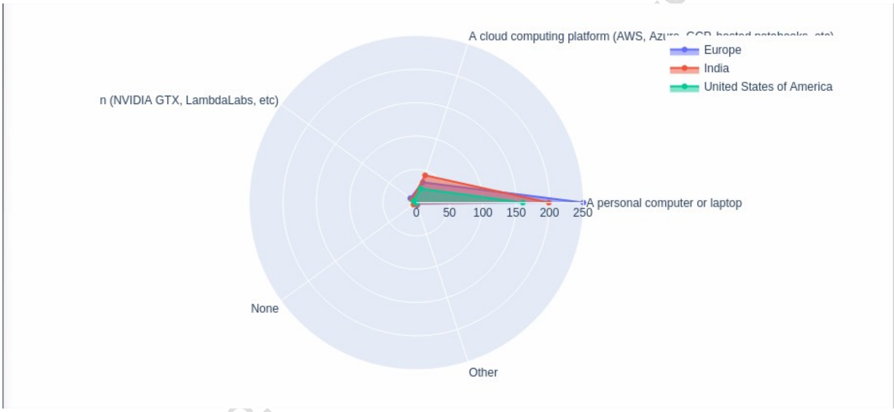
The radar chart highlights the difference between the three countries, with India having both a younger (in terms of number of years of experience) or newcomer base.

USA has fewer respondents that never programmed, while both USA and Europe have a larger group of members with 10 or more years of experience than India, but it is on the oldest range (20+ years) that Europe and USA numbers highlight probably a different market structure.

Question: Q11 What type of computing platform do you use most often for your data science projects?

Table ordered by categorical order (descending by complexity)

Computing platform	*Europe*	*India*	*USA*
A deep learning workstation (NVIDIA GTX etc)	11	8	5
A cloud computing platform (AWS Azure GCP etc)	32	43	22
A personal computer or laptop	251	199	160
Other	4	3	1
None	3	5	1



This question has no divergence: still to be confirmed (through other information) how much those numbers represent the business users or students.

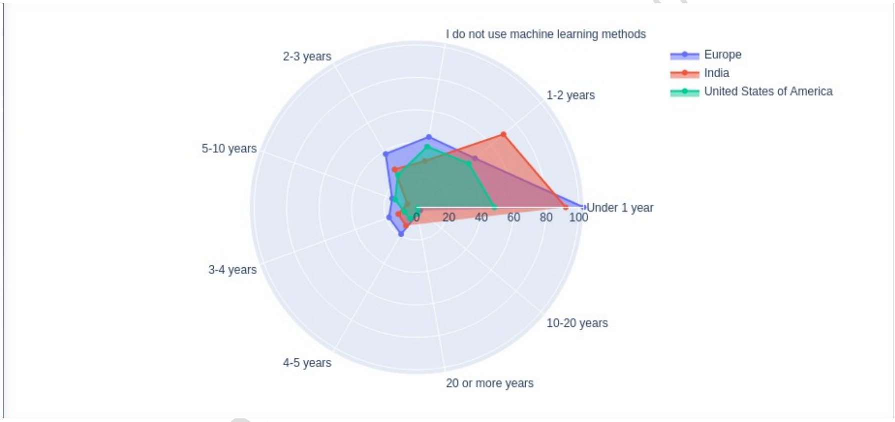
Using a laptop or desktop is common also in business, from observation, for at least two reasons:

- using cloud services to develop data analysis still receives, for data analysis activities, lukewarm acceptance (if any) from the business side
- purchasing ad hoc hardware (e.g. a machine learning workstation) is still rare for "line" business uses, and more the domain of (various forms of) research

Question: Q15 For how many years have you used machine learning methods?

Table ordered by categorical order (number of years used ML, descending)

ML methods since	*Europe*	*India*	*USA*
20 or more years	4	---	4
10-20 years	3	1.0	1
5-10 years	16	6.0	14
4-5 years	19	13.0	8
3-4 years	18	12.0	8
2-3 years	38	27.0	23
1-2 years	47	70.0	42
Under 1 year	103	92.0	48
I do not use machine learning methods	44	29.0	38

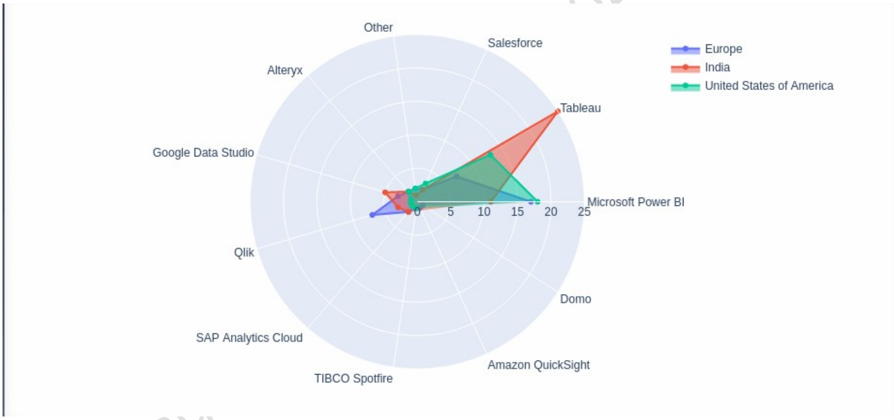


The main difference is shown from 5 years or more of experience in using machine learning methods, three times more common in Europe and USA than in India, albeit both Europe and India show significantly more "newcomers" than USA.

Question: Q32 Which of the following business intelligence tools do you use most often?

Table ordered by alphabetical order

Business Intelligence	*Europe*	*India*	*USA*
Alteryx	2	2	2
Amazon QuickSight.1	1	---	---
Domo	1	---	---
Google Data Studio	3	5	1
Microsoft Power BI	17	11	18
Other	2	1	2
Qlik	7	3	1
Salesforce	2	2	3
SAP Analytics Cloud	2	2	1
Tableau	7	25	13
TIBCO Spotfire	1	---	1



In both USA and Europe Microsoft Power BI (thanks probably to the diffusion in corporate environments of Office 365) is more common in Europe and USA, while Tableau is more common in India.

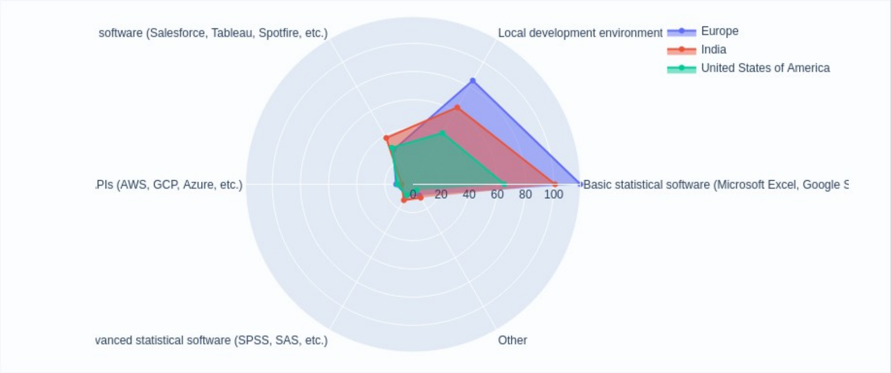
In Europe, along with Power BI, also Qlik has a not so small presence.

The absence of Salesforce and SAP Analytics Cloud are probably due respectively to the different audience (Kaggle is anyway targeting those who are focused more on data than on business), and the relatively recent introduction of the SAP offer (even in Europe, also in SAP customer environments is not so common, as it is associated mainly to the transition to the more recent offer, S/4 Hana)

Question: Q38 What is the primary tool that you use at work or school to analyze data? (Include text response)

Table ordered by categorical order (complexity level, descending)

Analysis tool	*Europe*	*India*	*USA*
Cloud-based data software & APIs (AWS GCP etc)	12	8	11
Advanced statistical software (SAS SPSS etc)	9	13	9
Local development environments (Jupyter Rstudio etc)	85	63	42
Basic statistical software (Microsoft Excel etc)	119	101	65
Business intelligence software (Salesforce Tableau etc)	28	38	30
Other	9	11	5



Remembering, again, that this representation focuses only on those whose job title (***Business Analyst*** or ***Product/Project Manager***) usually implies more contact with business.

Therefore, "to analyze data" usually implies "sharing analyses".

Until when also business users will be used to e.g. Jupyter Notebooks or other "live data" tools that are closer to data analysis and statistics, probably the humble spreadsheet will still be more common, at least in Europe and India.

Results:

- for the purpose of this report, three questions deliver interesting highlights: number of years of experience in programming, tools used for analysis, and tools used for business intelligence
- the key point is the difference in demographic represented by the years of experience
- as for the tools, could be interesting to see also how those tools are used on data-related activities, i.e. if Power BI is used also integrated with Python and R, i.e. integrated with the data science pipeline

Next step:

- summarize results

E7. Assess results and analysis

E7.1. Focus on the questions to use to compare countries

To summarize the preceding sections, this was the list of questions initially considered:

PERSON	Q1 What is your age (# years)?
	Q3 In which country do you currently reside?
	Q4 What is the highest level of formal education that you have attained or plan to attain within the next 2 years?
	Q5 Select the title most similar to your current role (or most recent title if retired):
COMPANY	Q20 What is the size of the company where you are employed?
	Q21 Approximately how many individuals are responsible for data science workloads at your place of business?
	Q22 Does your current employer incorporate machine learning methods into their business?
	Q24 What is your current yearly compensation (approximate $USD)?
TECH	Q6 For how many years have you been writing code and/or programming?
	Q15 For how many years have you used machine learning methods?
	Q32 Which of the following business intelligence tools do you use most often?
	Q38 What is the primary tool that you use at work or school to analyze data? (Include text response)

Data used in this section:

- number of rows (i.e. people): 898
- number of columns (i.e. questions, see table above): 12

Following the results of the analysis and visualization of each individual question, the following questions will not be considered, for the reason listed within the section shown in the last column:

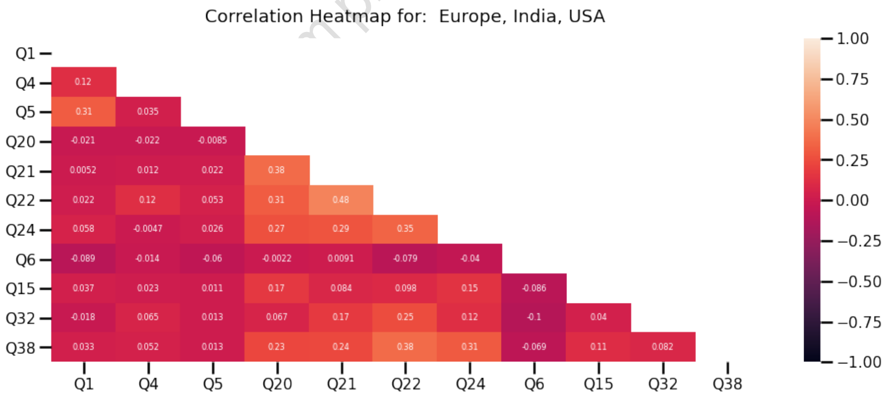
Question	*Reason*	*See*
Q2	gender gap is a constant	E4
Q23_ Part_1	"influencer" status would require further data	E3
Q25	question definition obfuscates purpose	E5
Q11	question definition obfuscates purpose	E6

Across section E7, the following dimensions will be used for the correlation analysis:

- Country: Europe, India, USA (question Q3)
- Job title: Business Analyst, Product/Project Manager (question Q5)

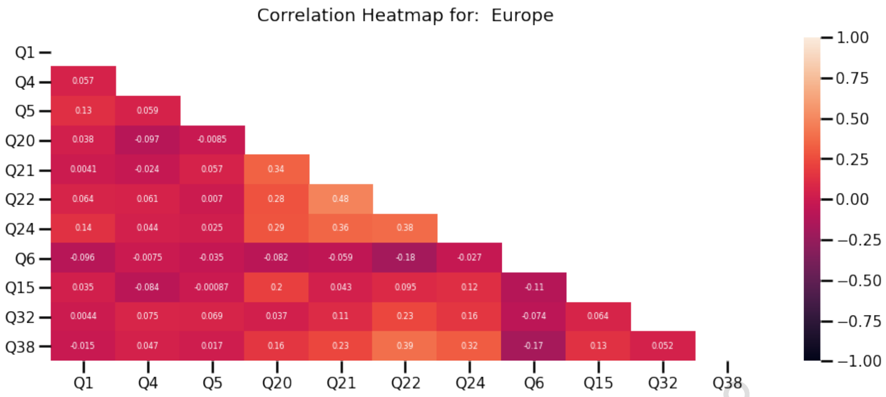
E7.2. Identify correlations

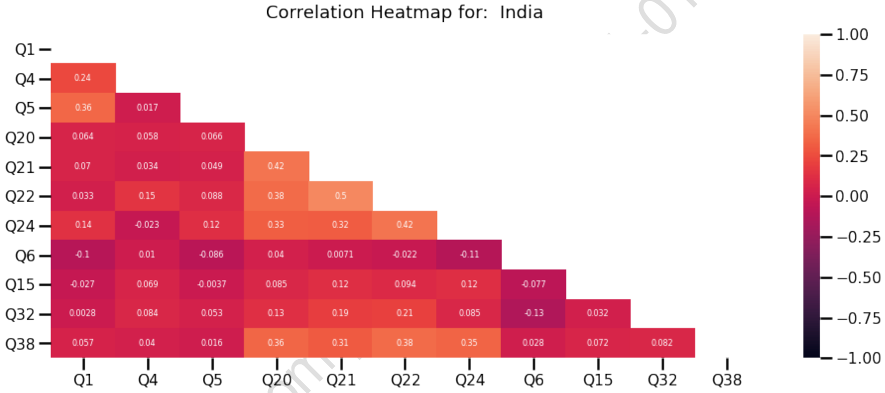
Before comparing countries, identify if the level of correlations between the variables is significantly different between the three countries.

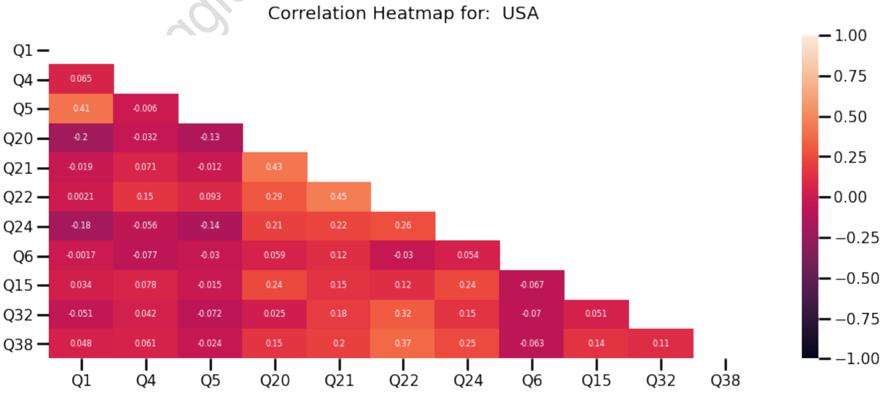


The heatmap showns the level of positive (number > 0) or negative (number < 0) correlation between the questions.

While correlation is not causation, it is interesting look at the first heatmap before moving on the next page, where the correlation is assessed country-by-country.



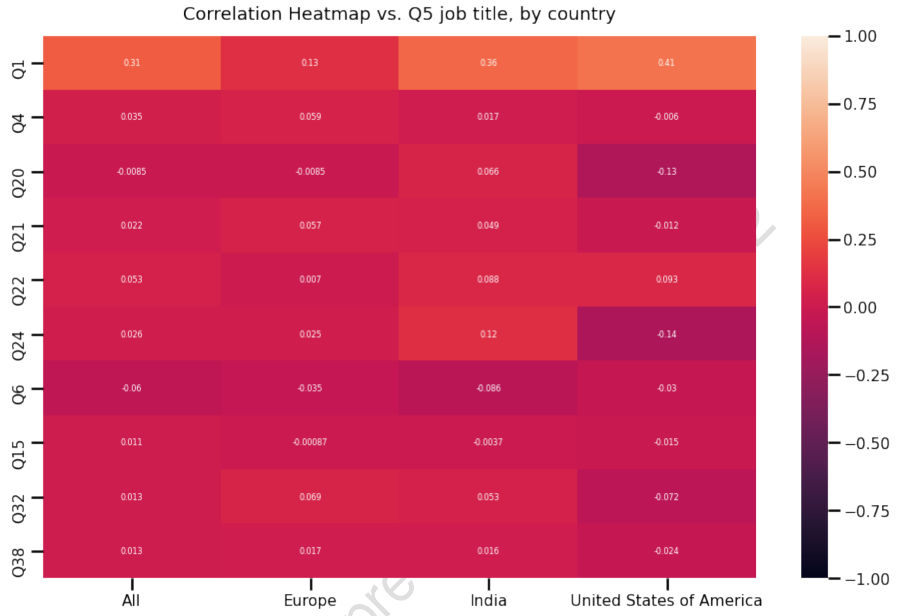




A simple visual inspection, i.e. on the first column, shows some differences between countries.

Anyway, nowhere there is a strong correlation (positive or negative).

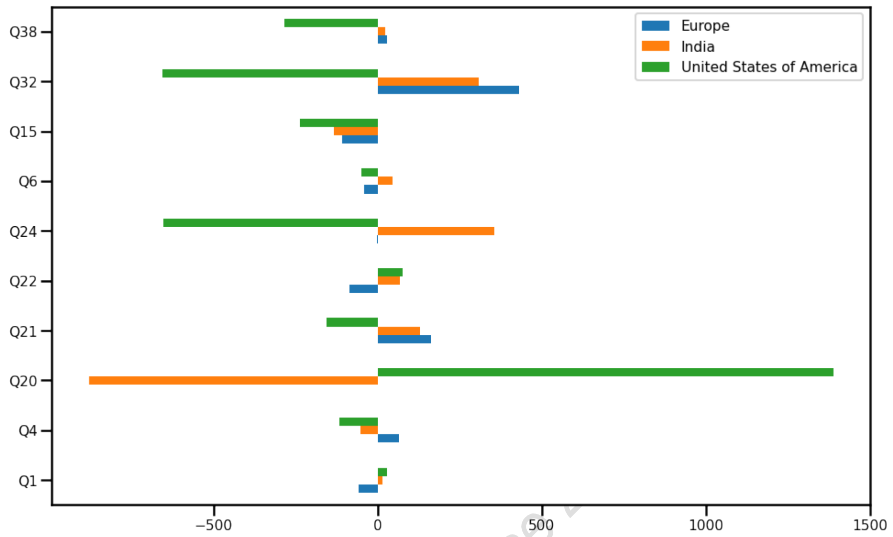
As this analysis is focused on three countries and two job titles (Business Analyst and Product/Project Manager), I decided to look at the correlation in each country for each question vs. the job title, Q5.



For your convenience, this is the table

Question	*All*	*Europe*	*India*	*USA*
Q1	0.313395	0.130021	0.362423	0.407846
Q4	0.035229	0.058566	0.017045	-0.005971
Q20	-0.008469	-0.008514	0.065998	-0.125937
Q21	0.021500	0.056761	0.048955	-0.011880
Q22	0.052836	0.006985	0.088373	0.093309
Q24	0.026063	0.025376	0.118834	-0.144002
Q6	-0.059702	-0.035166	-0.085810	-0.030329
Q15	0.010828	-0.000873	-0.003716	-0.014856
Q32	0.013004	0.069108	0.052920	-0.072178
Q38	0.012927	0.016585	0.015968	-0.023864

The following bar chart shows how much, on each question, the three countries differ from the average:



Question	
Q38	Primary tool for analysis
Q32	Business intelligence tools
Q15	For how many years have you used machine learning
Q6	For how many years have you been writing code
Q24	What is your current yearly compensation
Q22	Does your current employer incorporate machine learning
Q21	Approximately how many individuals are responsible
Q20	What is the size of the company
Q4	What is the highest level of formal education
Q1	What is your age (# years)?

As you can see, if you look at e.g. Q38 (Primary tool used for analysis), Q32 (Business intelligence tools) differentiate India from Europe and USA, as do Q24 (yearly compensation), Q21 (size of the team assigned to the AI workload), and Q20 (size of the company).

The details showed in sections E1 to E6 provide the details.

E8. Further investigations planned

Having identified some potential differences between the three countries, there could be an obvious path.

First, exploring the differences, and trying to identify patterns at least for the three questions that show the higher level of difference (Q32, Q24, Q20) for the roles selected.

Then, verify if those patterns apply also to the community of respondents at large, or are just an indication, due to the roles selected, of how in each of the countries there are different approaches.

At the same time, for Europe, as the subdivision in countries is present, identify if the above mentioned patterns show differences.

Also, would like to see if these differences are matched by differences vs. UN SDG for both the three "macro-countries" and the individual European countries.

While the preliminary analysis contained in this report was data-driven, the actual choices would depend on the purposes of the analysis, e.g. to identify different approaches to use in different countries when involving a third party to complement with data science / Machine Learning skills business and IT skills available “in house”.

7 READING NOTES

In this sample report, it was easy to add the "reading notes", as I had just two references.

In other cases, along with the Jupyter Notebooks on "negative choices" (i.e. directions of analysis that have been tested and abandoned or ruled out), a "reading notes" would be worth a book.

When preparing more complex reports, e.g. to assess which new project between many should be selected, often hundreds of pages and dozens of files are reviewed.

Also, sometimes the preliminary analysis, packaged within the final report, involves creating databases (e.g. to create a document catalogue, as I did for some audit projects), creating datasets (e.g. to evaluate alternatives), SQL scripts, etc.

All of that usually is part of separate volumes, to avoid overloading the main report.

In this case, as it was a sample Jupyter Notebook that by design had to be shared as a single file, as I wrote in previous chapters I left inside some "negative choices" documented, as examples, and added this last section to share my reading notes that, after reading the reports, used to prepare the report.

So, consider the concept, that I will replicate also in future same Jupyter Notebooks, but a more realistic business application would probably be represented by a collection of files and Notebooks.

F. APPENDIX

F1. Notes from the Kaggle Executive Summary

page 2

Based on responses from 20,036 Kaggle members, we've created this report focused on the 13% (2,675 respondents) who are currently employed as data scientists

=> see note on page 2

page 3

Data science continues to have a heavy gender imbalance, with most identifying as male

The vast majority of data scientists are under 35 years old

Over half of data scientists have graduate degrees

More than half of data scientists have less than three years of experience with machine learning

Scikit-learn is the most popular machine learning tool in 2020, with over four in five data scientists using it

Tableau and PowerBI are the most popular business intelligence tools

page 5

Similar to 2019 results, data scientists tend to be in their late 20s or early 30s, with about 60% between 22 and 34. Only one in five professional data scientists are 40 or older. There are signs of the numbers skewing even younger, as generation Z gets more involved. Nearly 7% of data scientists are aged 18-21, an increase from last year's 5%. Though not included in this chart, responses from students have also increased each year (26.8% in 2020, 21% in 2019, 22.9% in 2018). As these students graduate into the workforce, we may see future surveys with even younger data scientists.

page 6

Two countries have far more representation in the Kaggle community. India makes up almost 22% of Kaggle data scientists, while 14.5% reside in the United States. Brazil is a distant third, at under 5%.

page 7

Graduate degrees continue to be the norm for data scientists, with over 68% having obtained either a Master's or doctoral degree. Fewer than 5% of data scientists have no degree beyond a high school diploma.

page 8

Data science and machine learning are quickly changing, so it's no surprise over 90% of Kaggle data scientists maintain ongoing education. While about 30% take traditional higher education courses, many more learn through online materials.

Coursera, Udemy, and Kaggle Learn top the most common mediums in our survey. Unsurprisingly, many Kaggle data scientists chose multiple resources in the survey, with an average of 2.8 mediums selected.

page 9

Most Kaggle data scientists have at least a few years of experience under their belt. Just over 8% of data scientists have been programming since the 20th century! That's not to say there aren't newcomers, however. Over 9% have taken up programming in the last year. Just under 2% of data scientists claim to have never written code at all.

Compared to the global audience, United States data scientists have significantly greater programming experience. In the US, 37% have been programming 10 or more years, versus 22% worldwide.

page 11-12

US data scientists salaries 18.6% 100k-125k 18% 125k-150k 21.3% 150k-200k 8.9% 200k-250k 1.4% 250k-300k 3.9% 300k-500k 0.8% >500k

globally: 6.8% 4.5% 4.7% 1.6% 0.3% 0.7% 0.5%

page 14

median salary by contry: USA 125k-150k Germany 70-80k others: lower => check other countries

page 15-16-17

on companies employee data science, data science teams, enterprise machine learning adoption: biased by the absence of turnover information to see the weight

page 18

interactive development environments

>74.1% Jupyter lab 33.2% Visual Studio code 31.9% PyCharm >31.5% RStudio 21.8% Spyder >19.4% Notepad++

page 19-20

methods and algorithms usage

>83.7% linear or logistic regression => i.e. the same I used in the 1980s on DSS >78.1% decision trees or random forests >61.4% gradient boosting machines (xgboost, lightgbm, etc) >43.2% CNN >31.4% Bayesian approaches 30.2% RNN 28.2 Neural networks (MLPs, etc) 14.8% Transformer networks (BERT, gpt-3, etc) >7.3% GAN >6.5% evolutionary approaches 4.5% other 1.7% none

Python-based tools continue to dominate the machine learning frameworks. Scikit-learn, a swiss army knife applicable to most projects, is the top with four in five data scientists using it. TensorFlow and Keras, notably used in combination for deep learning, were each selected on about half of the data scientist surveys. Gradient boosting library xgboost is fourth, with about the same usage as 2019.

The fifth place tool, PyTorch, climbed above 30%, up from about 26% in 2019.

The most popular of the tools added to the survey this year is R-based Tidymodels, reaching over 7 percent.

machine learning framework usage >82.8% scikit-learn >50.5% tensorflow >50.5% keras >48.4% xgboost >30.9% pytorch 26.1% lightgbm 14.1% caret 13.7% catboost 10% prophet 7.5% fast.ai 7.2% tidymodels 6% h2o 3 2.1% mxnet 3.7% other 3.2% none 0.7% JAX

page 21

enterprise cloud computing

There are clearly three big players in cloud computing, and it's no surprise who: Amazon Web Services, Google Cloud Platform, and Microsoft Azure. Notably, more data scientists are using the cloud overall. In 2019, about 25% had not adopted cloud computing, which decreased to 17% in this year's survey.

48.2% AWS 35.3% GCP 29.4% Microsoft Azure 17.1% none 5.6% IBM Cloud / Red Hat 4.1% other 3% Oracle Cloud 2.9% VMWare cloud 1.9% Salesforce cloud 1.8% SAP Cloud 0.9% Alibaba cloud 0.7% Tencent cloud

page 23-24-25-26

enterprise machine learning product usage

Those who use AWS, Google Cloud Platform, or Microsoft Azure were asked about machine learning (ML) tools in particular. Over half of these data scientists do not use ML in the cloud.

Of those with ML usage, Amazon SageMaker was the most popular answer, followed closely by Google Cloud AI and ML.

55.2% no/none 16.5% amazon sagemaker 14.8% google cloud ai platform / google cloud ml engine 12.9% azure machine learning studio 8% google cloud vision ai 7.8% google cloud natural language 6.4% azure cognitive services 4.3% amazon rekognition 4.3% google cloud video ai 3.7% amazon forecast 2.9% other

enterprise big data

Business Intelligence tools help data scientists visualize their data, but four in 10 do not use one. The majority do employ BI, with Tableau as the most popular tool. Microsoft Power BI and Google Data Studio round out the top three.

data scientist usage of business intelligence tools

>38.8% none 33.3% tableau \27% microsoft power bi 9.1% google data studio 6.4% other 5% qlik 2.9% amazon quicksight 2.8% salesforce 2.5% looker 2.1% alteryx 2% SAP analytics cloud 1.4% tibco spotfire 1.2% sisense 0.9% einstein analytics 0.7% domo

Regarding databases, there isn't a clear favorite among data scientists. MySQL was mentioned most often (35.6%), followed by PostgreSQL (28.86%) SQL Server (24.93%).

database usage by data scientists

35.6% mysql 28.9% postgresql 24.9% microsoft sql server 18.7% mongodb 16.5% sqlite 15.4% none 13.5% google cloud bigquery 12.9% oracle database 9.3% amazon redshift 9.1% microsoft azure data lake storage 7.9% other 6.7% amazon athena 5.9% goole cloud sql 5.6% snowflake 5.1% amazon dynamodb 4.2% microsoft access 3.5% ibm db2 2.8% google cloud firestore

As with machine learning overall, many data scientists (33%) do not use auto ML tools. Google Cloud AutoML saw gains from last year’s survey, nearly 14% versus 6% in 2019.

13.9% google cloud automl 9.5% h2o driverless ai 8.4% datarobot automl 6.5% databricks automl

page 27

Among data scientists who use tools to manage machine learning experiments, TensorBoard is a clear favorite (over 21%). The closest competitor is Weights & Biases, with 6%. However, the vast majority (68%) of data scientists do not use special tools to keep track of and manage their ML experiments.

Usage of machine learning experiment tools

681.% no/none 21.6% tensorboard 6% weights&biases 5.4% other 3.1% trains 2.3% neptune 1% domino model monitor 0.9% polyaxon 0.8% guild.ai 0.7% comet.ml 0.6% sacred+omniboard

F2. Notes for the McKinsey Survey

cover

"Since our 2019 survey, artificial intelligence has become more of a revenue driver. Companies earning the most from AI plan to invest even more in response to COVID-19—and perhaps widen the gap with others."

page 2

"Overall, half of respondents say their organizations have adopted AI in at least one function

AI adoption was about equal across regions last year, this year's respondents working for companies with headquarters in Latin American countries and in other developing countries are much less likely than those elsewhere to report that their companies have embedded AI into a process or product in at least one function or business unit. By industry, respondents in the high-tech and telecom sectors 2 are again the most likely to report AI adoption, with the automotive and assembly sector falling just behind them (down from sharing the lead last year).

The business functions in which organizations adopt AI remain largely unchanged from the 2019 survey, with service operations, product or service development, and marketing and sales again taking the top spots"

page 2

"The use cases that most commonly led to cost decreases are optimization of talent management, contact-center automation, and warehouse automation"

page 4

"This year we asked about adoption of deep learning—a type of machine learning that uses neural networks and can sometimes deliver superior results—for the first time. Just 16 percent of respondents say their companies have taken deep learning beyond the piloting stage. Once again, high- tech and telecom companies are leading the charge, with 30 percent of respondents from those sectors saying their companies have embedded deep- learning capabilities."

from the comment from Micheael Chui, partner, McKinsey Global Institute, San Francisco " However, there was a bit of a decrease in bullishness this year, perhaps reflecting the passing of AI's hype phase. We do think AI is worth the investment, but it requires effective execution to generate significant value, particularly at enterprise scale."

Beside the usual concepts about performance and leadership attuned (up to championing IA initiatives directly from the C-suite), as in other past technological waves I witnessed since the 1980s, it is not just a matter of budgets, or share of the ICT budget, but it is a matter of internalizing skills and mindset.

page 5

"High performers also tend to have the ability to develop AI solutions in-house—as opposed to purchasing solutions— and they typically employ more AI-related talent, such as data engineers, data architects, and translators, than do their counterparts. They also are much more likely than others to say their companies have built a standardized end-to- end platform for AI-related data science, data engineering, and application development."

from commentary by Bryce Hall, associate partner, Washington DC: "Many executives now realize that AI solutions typically need to be developed or adapted in close collaboration with busi- ness users to address real business needs and enable adoption, scale, and real value creation. As a result, we see companies increasingly developing a bench of AI talent and launching training programs to raise the overall analytics acumen across their organizations."

Other dimensions are discussed in the report, e.g. the risks associated with AI.

Which cover reputational but also business risks, such as misuse of recommendation systems in business decision-making that is base: which converges with the current drive toward ensuring transparency via explainability in terms understandable to business users

unfortunately, as noted by Roger Burkhardt, partner, new york on page 10: "Overall, however, the results are concerning. While some risks, such as physical safety, apply to only particular industries, it's difficult to understand why universal risks aren't recognized by a much higher proportion of respondents. "

and adds a curious side-effect, probably linked to the paradigms used, e.g. learning from the past, in a time (the COVID-19 pandemic and its associated phases of universal business lockdown) when the drivers of past performance of a business are not relevant: "Generally, respondents from companies that have adopted more AI capabilities are more likely to report seeing AI models misperform amid the COVID-19 pandemic than others are. Responses indicate that high-performing organizations, which tend to have adopted more AI capabilities than others, are witnessing more misperformance than companies seeing less value from AI. These high-performing organizations' models were particularly vulnerable within marketing and sales, product development, and service operations (Exhibit 6)—the areas where AI adoption is most commonly reported."

data reported on page 12

" Respondents from AI high performers most often say their models have misperformed within the business functions where AI is used the most. 32% Marketing and sales 21% Product and/or service development 19% Service operations

page 13, methodology

" The online survey was in the field from June 9 to June 19, 2020, and garnered responses from 2,395 participants representing the full range of regions, industries, company sizes, functional specialties, and tenures. Of those respondents, 1,151 said their organizations had adopted AI in at least one function and were asked questions about their organizations' AI use. To adjust for differences in response rates, the data are weighted by the contribution of each respondent's nation to global GDP. McKinsey also conducted interviews with executives between May and August 2020 about their companies' use of AI. All quotations from executives were gathered during those interviews. "

8 NEXT STEPS

If you read the chapters so far, and visited the online *Jupyter Notebook*[23], you saw that this book contains:

- a discussion of the rationale behind this book and my Jupyter Notebook, an experiment in reusing concepts that I developed in the past to deliver reports turning qualitative into quantitative, to support decision-making
- an application to Jupyter Notebook of software and document development approaches that belong to a different era, pre-Internet, and pre-PC
- the use of material from the Jupyter Notebook linked above, to show how could be converted into a more traditional report.

The overall concept was to use “open data” (i.e. publicly available, with no restriction), free tools, and web-based platforms to both develop and deliver a report, as well as offer various access options.

Since 2012, I wrote my books on change etc[24] using Microsoft Office, including Excel, and sometimes also R for some charts.

In this case, I used: Debian 10, Python 3.7.3, Jupyter Notebook, Libre Office, and, of course, Kaggle.com, as well as various Python analysis and visualization libraries (from Numpy and Pandas, to Matplotlib, Plotly, Seaborn).

The next step is quite straightforward.

[23] https://www.kaggle.com/robertolofaro/ai-organizational-scalability-and-kaggle-survey

[24] See examples on https://issuu.com/robertolofaro

Since March 2020 followed online and offline various courses to add hands-on Machine Learning understanding, more than competencies.

Meaning: being able to read and understand the design, structure, and, when possible, results of various algorithms.

As written at the beginning of this book, my Artificial Intelligence experience was first in the 1980s- another era, and other technologies.

Nowadays, my humble PC, as well as free online computational resources such as those delivered by Kaggle.com allow to test and validate concepts at a fraction of the costs and human resources needed in the past.

Actually, instead of creating algorithms, you can benefit from the work done by others, shared e.g. in libraries such as Sci-Kit and PyTorch.

I used also TensorFlow and others, but I decided to keep developing skills on Sci-Kit and PyTorch, while keeping using Numpy and Pandas; so, I still have more learning work to do.

The Jupyter Notebook that I delivered in early January 2021 was both an experiment and a first step.

The concept really was to see if I could challenge myself in doing in few days first the Jupyter Notebook, then this book, simply to replicate what, for business, had to do to deliver business reports.

The next step on this sample Jupyter Notebook will actually be to work in parallel on two activities:

- further applications of the same logic and template (with evolutions, if needed) to other datasets that I already shared on my Kaggle profile since 2019, focused mainly on a an online sustainability publication project started in 2015
- continuing with this “fictional business project”, to use other tools and paradigms (e.g. Machine Learning) as if had received further report-writing assignments.

In the mid-1980s developed training curricula on Information Technology first as an experiment to escape boredom while serving (compulsory service) in the Italian Army, then for business purposes.

Along with training curricula, developed also analysis methodologies, first in the late 1980s to help replicate my approach to Decision Support System models building, then, from the early 1990s, as Head of Training and Methodologies for the Italian branch of a French multinational.

If you want to see an idea of how the Jupyter Notebook that I designed as a first experiment cold evolve into publications, you can have a look at a “stack” of mini-books that I released between 2015 and 2018 on Issuu.com.

It was a *fictional change business case*[25] on the introduction of the cultural, organizational, systems changes associated with compliance (over 200 pages across few “episodes” covering all the lifecycle of a 6-months programme built on a deadline).

So, you have now my “roadmap”: evolving the “template” Jupyter Notebook with real cases using “open data”, while also converting the storytelling in a single “episode” into a data-centered narrative using incidentally the various tools, a narrative that will span few “episodes”.

It is my way to do what I did in the past with other skills update initiatives that used free resources- sharing online the path.

Now, the plan.

First, by end February 2021 I will release a “business case” using data about visits to heritage sites in Italy, courtesy of data publicly available on a Government Ministry since few years ago.

You can see a sample *within a dataset that I posted on Kaggle a while ago*[26].

Then, will add further “business cases” using datasets similar to those, concerning sustainability, that I uploaded since January 2020 on Kaggle.com, after selecting data from various sources.
My approach is generally to build datasets, as I used to do in the past

[25] https://issuu.com/robertolofaro/stacks/86e5ef3fcb0c4f4dac7cd1b2cfef33a4

[26] https://www.kaggle.com/robertolofaro/italian-cultural-heritage-sample-data-20152018

in my Decision Support Systems (1980s) and Business Intelligence/Data Warehousing (1990s-2000s) activities, to ensure a degree of “functional stability”.

What I consider “functional stability”?

Having a point of reference so that, if you filter changes in information by e.g. complementing data (when this makes sense), you do not need to change what is not directly related.

It is a concept that was common decades ago, when any software or data definition change might impact tens of thousands of people all connected to a shared system.

It is apparently less and less common in modern environments.

I am skeptical of the business sustainability of software environments where changes to A have dozens of impacts totally unrelated, simply because by consensus a change was introduced in a library by those who did not consider the impacts (maybe because never used the impacted functionality in their environment, and did not bother asking for feed-back).

You could of course release a configured environment- but that is acceptable only if you assume that the “owners” of the configuration will keep evolving their “live reports” (from data preparation to models to training environment).

But the current evolution model of Python and its libraries works only if your customers do not actually evolve anything- they just “consume” your reports, as otherwise they would need to have internal staff able to juggle with all the different combinations delivered by all the different suppliers that focus just on what they need to have their own “live report” work.

So far, had limited impacts on my own experiment (e.g. changes in Pandas and Seaborn, as well an incompatibility with Numpy in Windows 10- the reason to switch to Linux as a test on “portability”).

Stay tuned!

ABOUT THE AUTHOR

If you spent time reading (or at least browsing through) this book, you are welcome to join me on https://linkedin.com/in/robertolofaro