What is Data Science?

Contents

Defining Data Science

What is Data Science?

Data Science as one’s attempt to work with data, to find answers to questions that they are exploring.

In a nutshell, it's more about data than it is about science.

If you have data, and you have curiosity, and you working with data, and you're manipulating it, you're exploring it, the very exercise of going through analysing data, trying to get some answers from it is data science.

Also Data Science is something that Data Scientists do.

As per Wikipedia:

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.

For me data science is:

Data is information and Science in its fundamental form is about having a question and searching for its answers using some scientific method.

Data Science is about having questions and then collecting data related to that topic and then manipulating it as per ones need, exploring it using different way to find answers to the questions which will give insights of the data collected.

In simple terms Data Science is working with data using scientific methods.

People do come in Data Science field from various paths.

Advice to succeed as a Data Scientist

1) **Be curious**
2) **Be judgmental**
3) **Be argumentative**

➡ Curiosity is absolute must. If you are not curious you would not know what to do with the data.

➡ Judgmental because if you do not have preconceived notions about things you wouldn't know where to begin with.

➡ Argumentative because if you can argument and if you can plead a case, at least you can start somewhere and then you learn from data and then you modify your assumptions and hypotheses and your data would help you learn. And you might start at the wrong point. You may say that I thought I believed this, but now with data I know this. So, this allows you a learning process.

➡ You should be able to tell a great story by your findings from data

➡ See, what is your competitive advantage. Do you want to be a data scientist in any field or a specific field? Because different fields require different skill sets. So figure out first what you're interested, and what is your competitive advantage. Your competitive advantage is not necessarily going to be your analytical skills. Your competitive advantage is your understanding of some aspect of life where you exceed beyond others in understanding that.

Structured and Unstructured Data

Structured Data:

So structured data is more like tabular data things that you're familiar with in Microsoft Excel format you've got rows and columns and that's called structured data.

Unstructured Data:

Unstructured data is basically data that is coming from mostly from web where it's not tabular, it’s not in rows and columns, it's text it's sometimes in video and audio so you would have to deploy more sophisticated algorithms to extract data and in fact a lot of times we take unstructured data and spend a great deal of time and effort to get some structure out of it and then analyse it.

So if you have something which fits nicely into tables and columns and rows go ahead that's your structured data but if you see if it's a weblog or if you're

trying to get information out of webpages and you've got a gazillion web pages that's unstructured data that would require a little bit more effort to get information out of it.

Regression

➡ If you have ever taken a cab ride a taxi ride you understand regression. Here's how it works. The moment you sit in a cab ride in a cab you see that there's a fixed amount there it's is $2.50(suppose). Rather the cab moves or you get off this is what you owe to the driver the moment you step into a cab that's a constant you have to pay that amount if you have stepped into a cab.

➡ Then as it starts moving for every meter or hundred meters the fare increases by certain amount so there's a there's a fraction there's a relationship between distance and the amount you would pay above and beyond that constant.

➡ And if you're not moving and you're stuck in traffic then every additional minute you have to pay more so as the minutes increase your fare increases as the distance increases your fare increases and while all this is

happening you've already paid a base fare which is the constant

➡ This is what regression is regression tells you what the base fare is and what is the relationship between time and the fare you have paid and the distance you have traveled and the fear you've paid because in the absence of knowing those relationships and just knowing how much people traveled for and how much they paid regression allows you to compute that constant that you didn't know it was $2.50 and it would compute the relationship between the fare and and the distance and the fare and the time.

➡ This is what is regression

Cloud

➡ It allows you to bypass the physical limitations of the computers and the systems you're using, and it allows you to deploy the analytics and storage capacities of advanced machines that do not necessarily have to be your machine or your company's machine.

➡ Cloud allows you not just to store large amounts of data on servers somewhere in California or in Nevada, but it also allows you to deploy very advanced computing algorithms and the ability to do high performance computing using machines that are not yours.

➡ The other thing Cloud is beautiful for is that it allows multiple entities to work with same data at the same time. So, you can be working with the same data that your colleagues in, say, Germany, and another team in India, and another team in Ghana, they are collectively working and they're

able to do so because the information, and the algorithms, and the tools, and the answers, and the results, whatever they needed is available at a central place which we call Cloud.

Defining Data Scientists

As per Wikipedia:

A Data scientist is someone who creates programming code, and combines it with statistical knowledge to create insights on business data.

For me Data Scientists is:

Someone who uses data to find solutions to certain problems using various tools and methods, who can make use of results to make better predictions and take wise decisions.

Also Data Scientists is a great storyteller, who can create great story with the insights he/she had gained in the process of exploring the data.

Difference between Data Science and Statistics

Data science combines multi-disciplinary fields and computing to interpret data for decision making whereas **statistics** refers to mathematical analysis which use quantified models to represent a given set of data.

What is Big Data?

As per Wikipedia:

Big data is a field that treats ways to analyse, systematically extract information from, or otherwise deal with data sets that are too large or complex to be deal with by traditional data-processing application software.

Data with many cases(rows) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate.

Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualisation, querying, updating, information privacy and data source.

Big data was originally associated with three key concepts: *volume*, *variety*, and *velocity*.

When we handle big data, we may not sample but simply observe and track what

happens. Therefore, big data often includes data with sizes that exceed the capacity of traditional software to process within an acceptable time and *value.*

Current usage of the term *big data* tends to refer to the use of predictive analytics, user behaviour analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set.

As per Norman White:

Big data is data that is large enough and has enough volume and velocity that you cannot handle it with traditional database systems.

Big data, was started by Google. When Larry Page and Sergey Brin wanted to, basically, figure out how to solve their page rank algorithm, there was nothing out there. They were trying to store all of the web pages in the world, and there was no technology, there was no way to do this, and so they went out and

developed this approach. Hadoop has copied.

But big data has now also expanded into, how do you analyse? There are new analytical techniques and statistical techniques for handling these really, really, really large data sets.

What is Hadoop?

Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.

As per Norman White:

Traditionally in computation and processing data we would bring the data to the computer. You'd wanna program and you'd bring the data into the program.

In a big data cluster what Larry Page and Sergey Brin came up with is very simple is they took the data and they sliced it into pieces and they distributed each and they replicated each piece or triplicated each piece and they would send it the pieces of these files to thousands of computers first it was hundreds but then now it's thousands now it's tens of thousands. And then

they would send the same program to all these computers in the cluster.

And each computer would run the program on its little piece of the file and send the results back. The results would then be sorted and those results would then be redistributed back to another process.

The first process is called a map or a mapper process and the second one was called a reduce process.

Fairly simple concepts but turned out that you could do lots and lots of different kinds of handle lots and lots of different kinds of problems and very, very, very large data sets.

So the one thing that's nice about these big data clusters is they scale linearly.

You had twice as many servers and you get twice the performance and you can handle twice the amount of data.

So this was just broke a bottleneck for all the major social media companies. Yahoo then got on board. Yahoo hired someone named Doug Cutting who had been working on a clone or a copy of the Google big data architecture and now that's called Hadoop. And if you google Hadoop you'll see that it's now a very popular term and there are many, many, many if you look at the big data ecology there are hundreds of thousands of companies out there that have some kind of footprint in the big data world.

Now we have the computational capabilities to apply some new techniques - machine learning. Where now we can take really large data sets and instead of taking a sample and trying to test some hypothesis we can take really, really large data sets and look for patterns.

And so back off one level from hypothesis testing to finding patterns that maybe will generate hypotheses.

Now this can bother some very traditional statisticians and gets them

really annoyed sometimes that you know you're supposed to have a hypothesis that is not independent of the data and then you test it.

Data Mining?

What is Data Mining?

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use.

Steps in Data Mining?

1) Establishing Data Mining Goal:

Set up goals that you want to achieve through data mining. Also look at the cost and benefits from this. You should also be concerned about level of accuracy and usefulness of the result obtained from data mining.

2) Selecting Data:

The output of the data mining highly depends on the quality of data being used.

3) Preprocessing Data

It is an important step in Data Mining. Data maybe messy, having irrelevant data, even relevant data may be missing. Identifying the wrong aspects of data set is essential. You should develop a method of dealing with missing data and determine whether the data are missing randomly or systematically.

If data missing randomly then simple set of solutions will be enough

But I data is missing systematically then you must determine impact of missing data on the results.

4) Transforming Data

Data transformation is the process of converting **data** from one format or

structure into another format or structure.

An important consideration in data mining is to reduce the number of attributes needed to explain the phenomena. This may require transforming data. Data reduction algorithms can reduce number of attributes.

You often need to transform variable from one type to another type.

5) Storing Data

The transformed data must be stored in a format that makes it conducive for data mining. Ease of reading and writing to database is important. It is also important to store data on servers or storage media to keep data secure and also prevent data mining algorithms from unnecessarily searching for pieces of data scattered on different servers or storage media.

6) Mining Data

This step covers data analysis methods, including parametric and non-parametric methods, machine-learning algorithms.

A good starting point for data mining is Data Visualisation.

7) Evaluating Mining Results

Evaluation could include testing and predictive capabilities of the models on observed data to see how effective and efficient the algorithms have been in reproducing data.

This is know as *“in-sample forecast”*.

This result can be shared to others for feedbacks.

Machine Learning

Machine learning (**ML**) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead.

It is seen as a subset of artificial intelligence.

Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task

Neural Networks
and
Deep Learning

➡ Neural Networks:

Computer Science's attempt to mimic a real, the neurons and how our brain actually functions. So 20, 30 years ago a neural network would have some inputs that would come in they would be fed into different processing nodes that would then do some transformation on them and aggregate them or something and then maybe go to another level of nodes and finally some output would come out. And I can remember training a neural network to recognise digits, handwritten digits and stuff.

So a neural network is trying to use a computer program that will mimic how neurons, how our brains use neurons to process things, brains to synapse, neurons to synapses and building these complex networks that can be trained. So a neural network starts out with some

inputs and some outputs and you keep feeding these inputs in to try to see what kinds of transformations will get to these outputs, and you keep doing this over and over and over again in a way that this network should converge so these input, the transformations will eventually get these outputs.

The problem with neural networks was that even though the theory was there and they did work on small problems like recognising handwritten digits and things like that, they were computationally very intensive, and so they went out of favour.

Then Deep Learning came:

What they did was they just had more multiple layers of neural networks and they use lots and lots and lots of computing power to solve them.

➡ Deep Learning:

Deep learning (also known as deep structured learning or hierarchical learning) is part of a broader family of

machine learning methods based on artificial neural networks. Learning can be supervised, semi-supervised or unsupervised.

Deep learning architectures such as deep neural networks, deep belief networks, recurrent neural networks and convolutional neural networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game programs, where they have produced results comparable to and in some cases superior to human experts.

➡ To get started with Neural Networks

You need to learn some linear algebra. A lot of this stuff is based on matrix and linear algebra, so you need to know how to do, use linear algebra and do transformations. Now, on the

other hand, there's now lots of packages out there that will do deep learning and they'll do all the linear algebra for you, but you should have some idea of what is happening underneath.

Deep learning, in particular, needs really high powered computational power, so it's not something that you're going to go out and do on your notebook for, you could play with it, but if you really want to do it seriously you have to have some special computational resources.

Basic Data Science Skills

If your are going to Data Science team basic skills you should have is:

1) Knowing some algebra and some calculus.
2) Knowing how to program, at least have some computational thinking.
3) Understanding relational databases.
4) Knowing basic probability and some basic statistics.

Internet Of Things (IOT)

In simple terms:

The interconnection via the Internet of computing devices embedded in everyday objects, enabling them to send and receive data.

Applications of Machine Learning

Recommender systems are certainly one of the major applications.

Classifications, cluster analysis, trying to find some of the marketing questions from 20 years ago, market basket analysis, what goods tend to be bought together. That was computationally a very difficult problem, I mean we're now doing that all the time with machine learning.

So predictive analytics is another area of machine learning. We're using new techniques to predict things that statisticians don't particularly like. Decision trees, Bayesian Analysis, naive Bayes, lots of different techniques. The nice thing about them is that in packages like R now, you really have to understand how these techniques can be used and you don't have to know exactly how to do them but you have to understand what their meanings are.

Precision versus recall and the problems of over sampling and over fitting so you can, someone who knows a little about data science can apply these techniques but they really need to know, maybe not the details of the technique as much as how, what the trade-offs are. And I'll give a plug for Foster Provost's book where he's actually written a whole book basically telling practitioners how to use all of these new machine learning techniques.

Applications of Data Science

➡ I think one of the good new applications of data science is in the medical field. Like in drug delivery or cancer treatment.

➡ I think a very interesting one is how now companies can use all the information they're gathering from their customers to actually develop new products that respond to the needs of the customers.

➡ A good new application of data science was the high trending news of Pokémon Go. So they used Ingress. They used data of the Ingress app. The last app of the same company and they choose the locations for Pokémons and gyms according to data from the last app. So they learned with their errors. - Google Search is an application of data science. The Google Search, whenever we want to search anything. So I think it’s all because of data science. Whatever

Google is now, it's all because of data science.

➡ Augmented reality is another application of Data Science

How should companies get started in Data Science?

At the end of the day, for businesses, they know one thing, that if they are unable to measure something, they are unable to improve it. And if they are unable to measure their costs, they are unable to reduce them. If they're unable to measure their profits, they are unable to increase them.

So the first thing a company has to do is to start recording information, start capturing data, data about costs. And the differentiate it by labor costs and material cost, the cost to how much it cost to sell one product and the total cost. And then you look at the revenue, where's your revenue coming from? Is 80% of your revenue coming from 20% of your customers? Or is it the other way around? So first thing first, start capturing data. Once you have data, then you can apply algorithms and analytics to it. So the first thing to do would be to capture data.

If you're not capturing it, start capturing it. If you're capturing it, archive it.

Do not overwrite on your old data thinking you don't need it anymore. Data never gets old. Data is always relevant, even if it's 100 years old, 200 years old. It is relevant to you and and your firm and your success. So keep data, capture it, archive it, make sure nothing goes to waste. Make sure there's a consistency. So someone 20 years later trying to understand that data should be able to do so, so have proper documentation. Do it now.

Put the best practices for data archiving in place the moment you start a business. And if you're already in business and you haven't done it, do it now.

Recruiting Data Scientists

When the companies are hiring people for a data science team, maybe a data scientist or an analyst, or a chief data scientist. The tendency would be to find the person who has all the skills, that they know the domain-specific knowledge. They're excellent in analyzing structured and unstructured data. And they're great at presenting and they've got great storytelling skills. So if you put all this together, you will realize you're looking for a unicorn. And your odds of finding a unicorn are pretty rare.

I think what you need to do to is to see, given the pool of applicants you have, who has the most resonance with your firm's DNA. Because you can teach analytics skills, anyone can learn analytics skills if they dedicate time and effort to it. But what really matters is who's passionate about the kind of business that you do. Someone could be a great data scientist in the retail environment, but they may not be

that excited about working in IT related firms or working with gigabytes of weblogs. But if someone is excited about those weblogs, if someone is excited about health-related data then they would be able to contribute to your productivity much more so.

And I would say if I'm looking for someone, if I have to put together a data science team, I would first look for curiosity. Is that person curious about things not just for data science but anything like, are they curious about why this room is painted a certain way, why do the bookshelves have books, and what kinds of books? They have to have a certain degree of curiosity about everything that is in their vision, that they look at.

The second thing is do they have a sense of humor because, you see, you have to have a lighthearted about it. If someone is too serious about it, they probably would take it too seriously, and would not be able to look at the lighter elements.

The third thing I think, and I think the last thing that I would look for if I had to have a hierarchy, the last thing I would look for are technical skills. I would go through the social skills, curiosity and sense of humor. The ability to tell a story, the ability to know that there is a story there. And then once all is there then I would say, well, can you do the technical side of it? And if there is some hope or some sign of some technical skills, I would take them because I can train them in whatever skills they need. But I cannot teach curiosity. I cannot teach storytelling. I cannot certainly, install sense of humour in anyone.

I think there's no hard and fast rule for hiring data scientists. I think it's going to be a case by case thing. I would say there has to be some sort of technical component, somebody should be able to work with and manipulate the data. They should be able to communicate what they find in the data. I find quite often nobody really cares about the r-square or the confidence interval. So you have to be able to introduce those

things and explain something in a compelling way. And they also have to find somebody who is relatable, because data science it been typically new means that the person in that role has to make relationships and they have to work across different departments.

If these data scientist has a good mathematics and statistics background. They have to consider like problem solving abilities and analysis, the scientist needs to be good in analyzing problems.

The persons they are hiring, they should love to play with data. And then they know how to play with the data visualization. They have analytical thinking. When a company is hiring anyone to work on a data science team, they need to think about what role that person is going to take.

Before a company begins, they need to understand what they want out of their data science team. And then they need to hire to begin it. As they grow a data science team, they need to understand

whether they need engineers, architects, designers to work on visualisation. Or whether they just need more people who can multiply large matrices.

From a skills point of view, let's focus on the technical skills and in that case, first thing would be what kind of a technical platform would you like to adopt? Let's say you want to work in a structured data environment and let's say you want to work in market research. Then the type of skills you need are slightly different than someone who would like to work in big data environments. If you want to work in the traditional market research data, structure data environment, your skills should be some statistical knowledge and some knowledge of basic statistical algorithms, maybe some machine learning algorithms. And these are the tools that you would like to develop. If you want to work in big data, then there's the other aspect of it and that is to be able to store data. So you start with the expertise in storing large amounts of data. And then you look into platforms that allow you to do that. The

next step would be to be able to manipulate large amounts of data, and the final step would be to apply algorithms to those large sets of data.

So it's a three-step process. But most likely it starts, most importantly, it starts with where you would like to be, in what field, in what domain. In terms of platforms, let's you want to be in the traditional predictive analytics environment, and you're not working with big data, then R or Stata, or Python would be your tools. If you're working mostly with unstructured data, then Python is most suitable than R. If you're working with big data, then Hadoop and Spark are the environments that you will be working with. So it all depends upon where you would like to be and what kind of work excites you and then you pick your tools.

In addition to technical skills, the second aspect of the data science is to have the ability to communicate. The communication skills or presentation skills. I call them story telling skills, that is that you have your

analysis done, now can you tell a great story from it? If you have a very large table, can you synthesise this and make it more appealing that when it goes on the screen, or is it part of a document that it just speaks? It sings the findings and the reader just gets it right there. So the ability to present your findings, either verbally, or in a presentation, or in a document. So those communication and presentation skills are equally important as the technical skills are. When you have a grading side, when you're presenting your results, imagine you're driving on a mountain and then there's a sharp turn. And you can't see what's beyond the turn. And then you make that turn and then suddenly, you see a tremendous valley in front of you. And this great sense of awe, that I didn't know that, right? So when you present your findings and you have this great finding and you communicate it well, this is what people feel because they were not expecting it. They were not aware of it and then this great sense of happiness that now I know. And I didn't know this, now I know. And then it empowers them, it

gives them ideas what they can do with this knowledge, this new insight. It's a great sense of joy. And you are able as a data scientist, you are able to share with your clients because you enabled it.

How to use data correctly?

So what has happened is that now the tools are available and datasets are available, people are applying them with not much diligence and I think one of the strange cases which got reported in the newspapers is about the story of a father walking into a Target store in the US and complaining about the fact that the Target was sending mails to his teenage daughter about diapers and milk, baby formula.

He was angry with them. He said, "Why would you like "for my teenage daughter to have a baby?" And he was obviously disturbed by this mail or the ad campaign. And they obviously apologized but then the father returned two weeks later and he apologized to them saying he didn't know his daughter was pregnant.

Now the question is, how did Target know this thing before the father knew. And what has happened is that they would

look at the purchasing behaviour of individuals.

So if you're buying some sort of supplements or vitamins then you know that this is the first trimester of pregnancy. So they know what products to send to you assuming that the person who bought those supplements were pregnant. Now this is a great story about data science and how data science can forecast and predict these consumer behaviours even before the family would find out. And I find it disturbing and strange and odd for a variety of reasons.

First of all, for every correct prediction, you have hundreds of incorrect predictions which we call the false positives and no data scientist actually advertises his or her false positives. We only advertise and promote what we got it right. But when we got it wrong hundreds of times we don't tell it. Second thing is, that's an abuse of data. That's basically not really not giving you much insight.

You've just found a correlation but someone could be purchasing the same material for someone else. So, and then the odds of getting it wrong and the odds of getting false positives is much higher. So I find it strange and I think it gives a false sense of our ability to predict the future.

The reality is about data science and the most important thing for the budding data scientist to know that all forecasts are wrong. They're useful but they're wrong. And so one should not put their faith into the fact that now that we can do predictive analytics that we can solve all problems.

I think a good example is the Google Search. Google published a paper saying they can predict flu epidemics before the Center for Disease Control. And what they did was they were looking at what people were searching on Google so flu symptoms. So Google saw the flu symptom searches before anybody else and they were able to predict it. The thing is these searches are good and they are

correlated with some outcomes but not necessarily all the time.

So at that time, when Google announced, it was a big thing and everybody really like it and well that's a new era of predictive analytics. Only that a few years later they realized that Google started to predict false positives.

That they were predicting things that were not really there or the predictions were not that accurate for a variety or reasons. They changed probably their algorithms and the datasets were not really correlated with the outcomes.

So what's the lesson to learn here? One has to avoid what we call the data hubris. That you should not believe in your models too much because they can lead you astray.

Data science has tremendous potential to bring change in parts of the world, in parts of our society that have been disenfranchised for years. One sees great examples of data science

especially in the developing countries where they are targeting relief efforts. They're targeting food and other aid to individuals, to places that have not been targeted in the past. And the reason it is happening now is the greater availability of data and models and analytics to be able to pinpoint where the greatest needs are.

The ability to design and conduct experiments to see if one were to give micro-credits, small loans to very poor households in developing parts of the world, to see how they affect the individual household's ability to get out a poverty and also the local community's ability to collectively improve their economic well-being by just very small infusions of cash or credit.

So these experiments happening all over the world are allowing that is a direct result of our ability to analyze data and be able to design experiments and then roll out humongous efforts in providing relief, providing credit, providing an opportunity to those who

have been disenfranchised in the past an opportunity to join the rest of the world in prosperity and happiness and health.

The Final Deliverable

➡ The main purpose of analytics is to communicate findings to the concerned who might use these insights to formulate policy or strategy.

➡ Analytics summarise findings in tables and plots.

➡ The data scientist should then use the insights to build a narrative to communicate the findings.

➡ In academia, the final deliverable is in the form of essays and reports (length ranging from 1000 to 7000 words)

➡ In consulting and business, the final deliverable take several forms.

➡ It can be small document of fewer than 1500 words illustrated with tables and plots or it could be a comprehensive document comprising several hundred pages.

NOTE:

While writing a final deliverable narrative is very important because this will help to communicate the message to others.

➡ Before you start your analysis you should plan the scope of the final deliverable. First decide the key message that you want to deliver and then look for the data and analytics to make those cases.

➡ The initial planning and conceptualising of the final deliverable is extremely important for producing a compelling document.

The Report Structure

Before starting the analysis, think on the structure of the report. Will it be brief or detailed report.

A brief report is more to point and presents a summary of key findings.

A detailed report incrementally builds the arguments and contains detail about other relevant works, research methodology, data sources and intermediate findings along with the main result.

The deliverable should follow a prescribed format including:

1) Cover Page
2) Table of contents
3) Executive summary
4) Detailed contents
5) Acknowledgements
6) References
7) Appendices (if needed)

1) Cover Page:

A cover page should include

- ➡ Title of the report
- ➡ Name of the authors
- ➡ Their affiliations and contacts
- ➡ Name of the institutional publisher (if any)
- ➡ Their date of publication

2) Table of contents:

- ➡ Main heading
- ➡ Lists of Tables
- ➡ Figures

This gives a glimpse of what lies ahead in the document

3) Abstract or Executive Summary:

- ➡ It should be included in the document
- ➡ Explain the important points in the arguments in 3 or less paragraphs
- ➡ For larger documents this summary can be longer

4) Introductory Section:

It is helpful for setting up the problem for the readers.

5) A good follow up to the introductory section is a review of available relevant research on the subject matter. The length of literature review section depends on how contested the subject matter is.

6) Methodology Section:

Here you introduce the research methods and data sources you used for the analysis.

If you have collected new data then explain the data collection exercise in some detail. You will refer to the literature review to bolster your choice for variables, data and methods and how they will help you to answer your research questions.

7) Result Section:

This is where we present our empirical findings. Starting with descriptive statistics and illustrative graphics and then testing of your hypothesis. If in case you need to run statistical models, you might turn to regression model or categorical

analysis. If you are working with time series data then explain it.

The business sector holds back on the statistical details and rely on illustrative graphics to summarise the results.

8) Discussion Section:

- ➡ Craft your main argument by buildings on the result you have presented earlier
- ➡ This is the place to rely on narrative to communicate your thesis to the reader
- ➡ You refer reader to the research question and knowledge gaps you identified earlier. You highlight how your findings provide the ultimate missing piece to the puzzle.

9) Conclusion Section:

Generalise your specific findings and take on a rather marketing approach to promote your findings so that reader does not remain stuck in caveats you created earlier.

You might also identify further possible developments in research and applications that could result from your research.

10) Acknowledgements, References and Appendices (if needed)

You should answer the following questions while making reports

1) Have you told readers, at the outset, what they might gain by reading your paper?

2) Have you made the aim of your work clear?

3) Have you explained the significance of your contribution?

4) Have you set your work in the appropriate context by giving sufficient background (including a complete set of relevant references) of your work?

5) Have you addressed the question of practicality and usefulness?

6) Have you identified future developments that might result from your work?

7) Have you structured your paper in a clear and logical fashion?

My Definitions

Data Science:

- Data is information and Science in its fundamental form is about having a question and searching for its answer using scientific methods.

Owning to that

- Data Science is about having questions and then collecting data related to that topic and manipulating it as per ones need, exploring it using different ways to find answers to the questions which will give insights of the data collected.

- In simple terms if I would say then “Data Science is working with data using scientific methods”.

Data Scientist:

Someone who uses data to find solutions to certain problems using various tools and methods, who can make use of results to make better predictions and take wise decisions.

Also Data Scientists is a great storyteller, who can create great story with the insights he/her had gained in the process of exploring the data.