CRASH COURSE IN DATA SCIENCE

Outline

A. Introduction: How Data Science Became a Prime Business and Research Tool
B. Course Objectives and Limitations
C. Course Proper
a. What is Data Science?
b. Data Science Versus Other Commonly Misused Terms
i. Big Data
ii. Machine Learning
iii. Artificial Intelligence
c. What Do You Need to Know: the Very Basics
i. Programming Languages
1. Python
2. R
ii. Math (Statistics)
d. Processes of Data Science
i. Data Acquisition
ii. Data Exploration
iii. Data Modelling and Visualization
e. How to Become a Data Scientist if You're….
i. An Undergrad
ii. An Experienced Professional or a Career-shifter
D. References

INTRODUCTION

How Data Science Became a Prime Business and Research Tool

Data has always been around, but it has never been more prevalent, illustrative, and predictive of events, trends, and even human behavior than it is at present. It is now a prime business and research tool as it enables individuals, groups, enterprises, and even governments make sound and well-informed decisions not based on a hunch, intuition, probability, or previous experiences, but based on data that depicts current movements or behavior, or can even predict or forecast trends or events based on current data.

Uses of
Data Science
Today

Video Clip 1: Uses of Data Science Today.
Data Science YouTube Channel. November 24, 2017
https://www.youtube.com/watch?v=e242eYPq0nE

With businesses from all sectors or industries utilizing data to leverage or move their businesses forward, this created a huge demand in the job market for data scientists or professionals who have experience in statistics, programming, analytics, communications, and even project management. In 2020, job portal Glassdoor ranked the job post data scientist as the third best job in the U.S. Furthermore, the Bureau of Labor Statistics estimates the job growth of the field by three-fold. Besides data scientist work being a lucrative, multi-specialty career that pays a hefty 100k a year, what else does this tell us?

Anyone can be a data scientist, and you don't have to be an I.T. major to gather, handle, process, and interpret data. Don't let the word "data" throw you off. Data science is a vast field that encompasses various fields:

- I.T. – part and parcel of data science is programming, and the common programming languages it uses to harvest and analyze data are Python and R, both of which will be briefly discussed in the succeeding sections. Software engineering and cloud computing are also advantageous skills in this field
- Mathematics – also part and parcel of data science is mathematics, specifically statistics. A 2017 paper featured in the International Journal of Data Science and Analytics noted how statistics play an integral role in almost every data processing cycle step. Knowledge of algebra and algorithms are also essential.
- Analytics – this is more of a skill than a field, a skill that is vital in any field that deals with quantitative data
- Any non-computing field – why is this relevant? It is important to realize that data exists in any industry, whether it be financial, I.T., medical, military, sports, or any other field, and knowing the

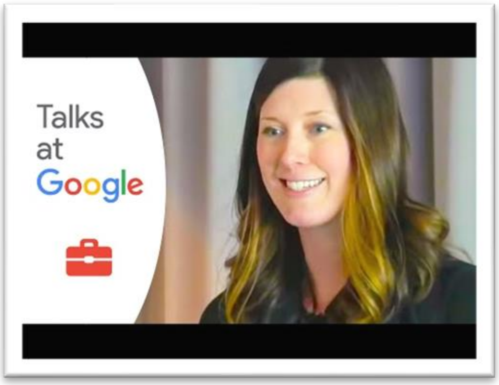
specific business processes involved in any of these industries is equally essential in putting the dots of data together to piece the whole story and make viable and relevant decisions from it.

So, in simplest terms, what is the objective of data science? Why is it an indispensable business, research, and decision-making tool?

If you missed it in the last bullet, here's why.

It pieces the whole story. The various methodologies involved in the whole data science cycle helps connect the dots of data to stitch the bigger picture. With effective storytelling and visualization, a picture can reach even the least technical people or the novice conversant in the room.

That's the power of data. What makes it science is the methodical processes involved from the collection phase, processing, interpretation, and visualization. In this crash course, we will present the basics of data science that should give any novice reader on the topic enough to go on or a good starting point to further their reading and self-learning.



Video Clip 2: Storytelling with Data by Cole Nussbaumer Knaflic
Talks at Google YouTube Channel, November 12, 2015
https://youtu.be/8EMW7io4rSI

Points to Ponder

Can you imagine how business decisions were made before data science come to prominence?

OBJECTIVES AND LIMITATIONS

Course Objectives and Limitations

To say that this course's objective is to impart fundamental or even just an idea of what data science is and how essential it is in many enterprises today is a given.

Think of this section as a foreword instead.

This crash course aims to introduce novice readers like you to the topic, whether you're an incoming undergrad who's still on the fence on what to pursue in college or what courses to take or an experienced professional looking to take the plunge on the data science bandwagon or looking to make a definite career move into this ever-growing multi-specialty field.

Yes, data science is a multi-specialty field, so it is sometimes difficult for beginners to choose a starting point. Do you learn Python or R programming first? Do you learn Python programming only or R only? Will learning statistics be enough? This crash course will not go into the details of Python or R programming, nor statistics and others. It will, however, provide basic information on these topics and other topics related to data science.

The basic information will be limited to its definitions, its uses or purpose, and how it is utilized in the context of data science and its cyclical processes.

COURSE PROPER

What is Data Science?

As an interdisciplinary field, data science can be difficult to define as it is ever-growing and ever-evolving. As we said in the previous sections, data is nothing new. It had always existed even before computers came about. Data is the elemental form of information. Numbers, quantities, metrics, units of measure, and other descriptors, are all data. Without any methodological processes, these data would be just that – data.

But behind those numbers and other descriptors is a story. That's where the "science" in data science comes in. if you think about it, data science is the product of years, or even decades, of a collective effort by various professionals or data specialists to apply scientific computing methods to refine the processes of collection, manipulation, analysis, and visualization of large volumes of data.

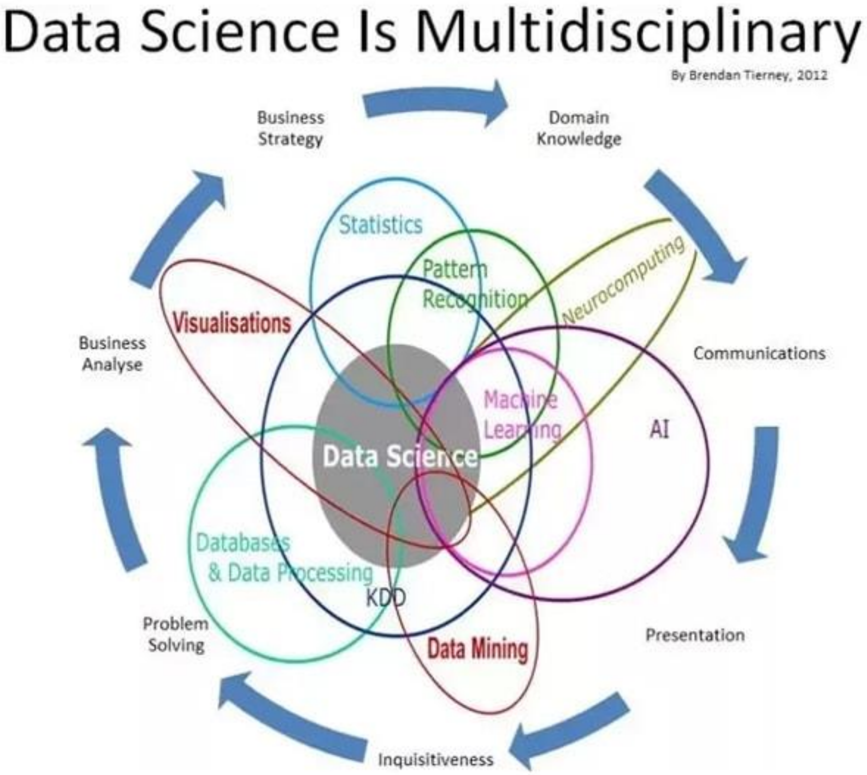


Image from https://www.simplilearn.com/data-science-vs-data-analytics-vs-machine-learning-article#:~:text=Because%20data%20science%20is%20a,machine%20or%20a%20mechanical%20process.

Data Science Versus Other Commonly Misused Terms

Data Science versus Big Data

These two terms are usually equally mentioned in one sentence. Other times, they are mentioned synonymously. Whereas data science is an interdisciplinary cyclical process, big data, as the word implies, refers to volumes and volumes of data and the processes involved in gathering these, how to store them (which involves building infrastructures for scalable processing and storage) and how to query them for retrieval purposes.

Data Science versus Machine Learning

These two are not used synonymously but are usually used in a sentence in conjunction or next to each other, i.e., complementary. That being said, instead of putting these two terms at odds against each other, let's see how these fit together.

Data Science, being the interdisciplinary field that it is, takes on a broader perspective. To put it in realistic terms, those in this field are either called data scientists or architects. They deal with larger projects or problems, or the bigger picture, such as raising market share or increasing market exposure of a product. In other words, it usually deals with intangible problems and solutions. On the other hand, Machine Learning, as the name suggests, is more focused on creating tangible solutions through algorithms. The expected outcome of machine learning is based on formulated requirements based on the analysis of data.

In short, and using a normal Google query as an example, data science is responsible for populating search results (and its rankings), while machine learning is responsible for smart search features like voice search, autocorrect, and predictive text.

Data Science versus Artificial Intelligence

The simplest way to differentiate these two concepts is by looking at how they treat data. Data science tries to make sense of what the data is trying to say through different tools like statistics, while A.I. uses data to implement machine learning algorithms thus, allowing machines to respond and behave based on expected outcomes. Those expected outcomes are usually based on known human responses and behaviors.

In short, data science tries to understand data while A.I. implements data.

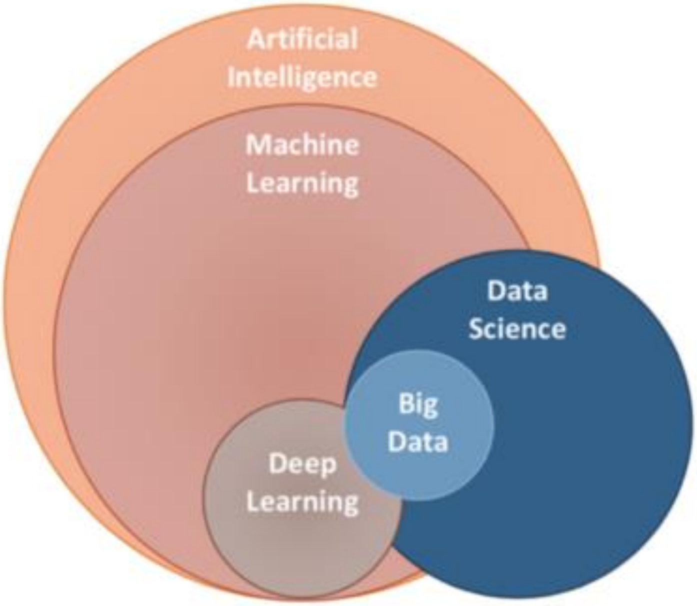


Image from https://medium.com/ai-in-plain-english/data-science-vs-artificial-intelligence-vs-machine-learning-vs-deep-learning-50d3718d51e5

The image above is a perfect yet simple depiction of the bigger scheme of things in the data world. A.I. encompasses machine learning and deep learning, and big data, and all these have a symbiotic relationship with data science. And although data science can and is a stand-alone discipline, its purpose is greatly defined by how A.I., ML, big data, and other related concepts such as deep learning utilize its results through various implementations. Without A.I., etc., data science seems like a concept that's only good on paper, or, like a law or bill that exists but is not implemented at all, rendering it useless.

Differentiate and relate with data science these two other terms:

- ***Deep Learning***
- ***Statistics***

While we know, based on the diagram, that Deep Learning has the same symbiotic relationship with Data Science, what about Statistics? Is Statistics really a part of Data Science or is it the other way around? How are these two related?

What Do You Need to Know: the Very Basics

Like we have mentioned countless times in this piece, Data Science is an interdisciplinary field. But among the sea of skills, a data scientist must possess the indispensable or essential ones? Let's briefly go through each:

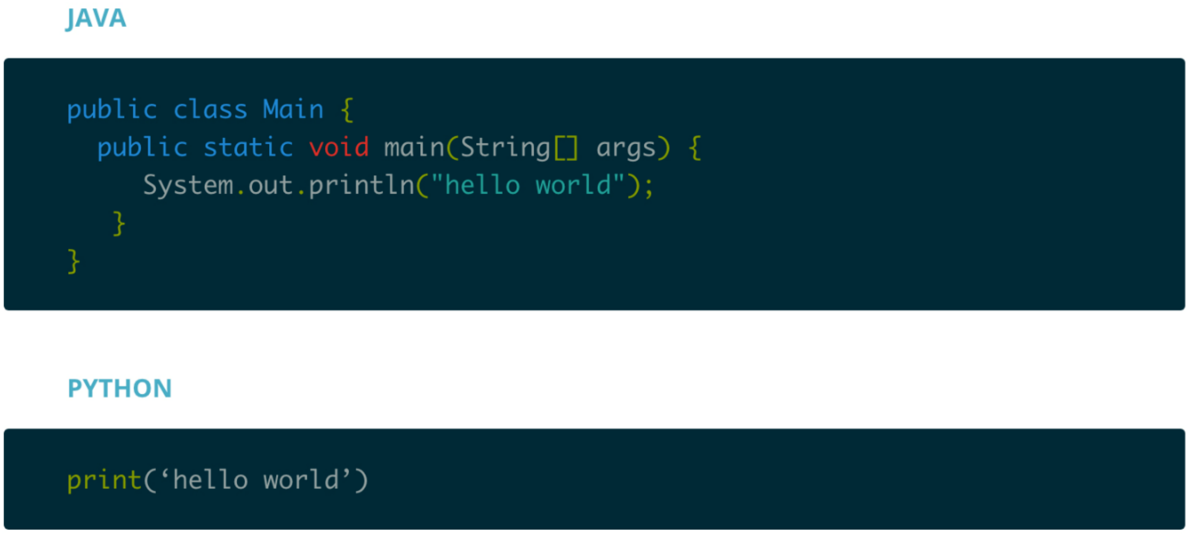
1. Python Programming

Python has been around since the early '90s when Guido van Rossum created the language with influences from the ABC language. What's appealing with Python is its ease of use, simplicity, flexibility, and generality.

Python is not an original language per se. Think of it as a combination of syntaxes for various purposes like internet protocols, text processing services, binary data, various modules like mathematical, numerical, programming, O.S. interfaces, data processing, compression and archiving, and many more. Many of the codes are written in C, which means it is not self-hosting, but basic data types like strings, dictionaries, numbers, and lists are enabled and are extensible to C and C++.

Is it advisable to learn Python right away for beginners? Are there any pre-requisite languages necessary?

Learning fundamental programming languages like C, C++, or even Java is a definite plus, but remember when we said that Python is simple? Here's how simple it is compared to Java:



JAVA

```
public class Main {
  public static void main(String[] args) {
     System.out.println("hello world");
   }
}
```

PYTHON

```
print(‘hello world’)
```

Image from https://www.pluralsight.com/blog/software-development/why-python#:~:text=Python%20is%20easy%20to%20use,waste%20time%20with%20confusing%20syntax.

Python gets to the point compared to syntax-heavy languages like Java. For beginners, this is a great advantage and a great motivator because, as we said, it gets straight to the point. It allows the beginner programmer to develop the mindset of a programmer. Some professionals may say that this – simplistic learning - puts the beginner at a disadvantage, especially in the long run, so again, it's a matter of preference for the novice learner. But if the end game is to learn about data science in an expedited fashion, then there's no harm in learning Python as your first programming language.

Companies and applications that use Python include Dropbox, Google, Spotify, and Netflix. At Google, Python is employed along with C++. It gave Google the perfect combination of low maintenance with high efficiency (Python) and low latency with efficient memory (C++). Other apps like Spotify and Netflix use Python mainly for data analytics.

And finally, not only is Python simple, but it is also stable. How stable? The companies mentioned above use Python for their web-based and or mobile apps and for desktop-based clients, which speaks to the cross-platform usability and stability of the language. Cross-usability of a language adds to its power because developers can use it to create almost anything in various media forms to be implemented across platforms.

These are the reasons why Python is also widely used in data science, with about 2 out of 3 data scientists using Python for their daily data tasks.

Video Clip 3: *How Is Python Used In Data Science? [For Beginners!] by Kim Fessel Course Report on YouTube Channel. September 15, 2020. https://www.youtube.com/watch?v=_FUc7QML8dU&feature=youtu.be*

2. R Programming

R Programming was launched by New Zealand statisticians Ross Ihaka and Robert Gentleman (which is probably where the "R" in R programming came from). Part of its language is written in R. It is self-hosting. It is also partly written in Fortran and C.

Whereas Python is a multi-platform, multi-purpose programming language, R is written for statistics and the subsequent computations and visualizations needed.

Both R and Python are used in data science. Each programming language has its strengths in the context of data science. For example, Python is more used with machine and deep learning and is easier to deploy or implement across production platforms. On the other hand, R programming is frequently employed in statistical computations, researches, or models. Some scientists use one language over the other, but there's no reason to choose. Both languages are complementary in the context of data science.

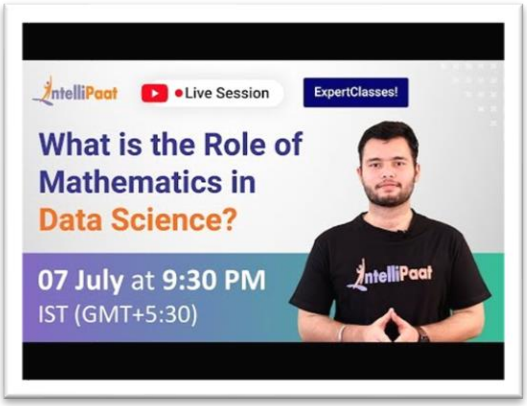
Can you enumerate more strengths of Python and R and how these strengths complement each other?

3. Statistics

Statistics is a mathematical field that, like data science, is concerned with collecting and processing quantitative data, but that's where the similarities stop. While a big part of data science is the statistical part, this is only about half of the input and output, because as you might know by now, data science encompasses a lot of disciplines or areas of interest.

Another striking difference between the two, aside from the breadth of information data science has over statistics, is that statistics quantify not only finite possibilities but, well, possibilities, whether certain or uncertain. Data science, on the other hand, thanks to other fields that focus on improving prediction like machine learning and A.I., can make more definite quantifiable interpretations because of these.

It is still vital to learn statistics as one needs to understand fundamental statistical concepts like regression, hypothesis testing, conditional probability (Bayesian Thinking), probability theory, and distributions, among many others.



Video Clip 4: *What is the Role of Mathematics in Data Science?*
Intellipaat YouTube Channel. July 7, 2020
https://www.youtube.com/watch?v=9ZOwkidcFQ8

Processes of Data Science

1. Data Acquisition

   This is more commonly known as Data Collection. This is where data sets in the form of comma-separated value (CSV) formats, SQL tables, Excel files, or text files come in. Python allows data scientists to create their own or call datasets from the web, while newer packages of R like Rvest or Magrittr also enable one to gather, clean, and divvy up the information contained within the data sets.

2. Data Exploration

   Of course, once you've collected data, the next logical step is to explore what you've got, skim through the data, if you will. Exploring data is like looking and studying data that has been transferred to a spreadsheet or a table. But instead of using apps like Excel, which, by the way, is also memory-intensive, Python and R both have their functionalities that help optimize this intensive step. Python has its Panda library, while R has its deeply rooted foundation in statistics and offers statistical tools and analysis to your data.

3. Data Modelling and Visualization

   Both these steps aim to represent data with minor differences visually. Think of Data Modelling to classify and store data and visually represent the relationships between the data groups or types. On the other hand, data visualization is exactly what the words imply – visualization – illustrating or explaining the results through diagrams, photos, tables, and other visual aids.

Further expand the steps of data science while also expanding on the role of Python and R. You can use items 1 and 2 as examples.

How to Become a Data Scientist if You're…

1. An Undergraduate

   A bachelor's degree in data science is the most straightforward answer here. But you can also major in I.T., mathematics, physics, computer science, or even computer engineering. These can give you a solid background in data science. You can learn the rest of the hard skills not taught in your majors through other means like online classes, MOOCs or short courses, or on the job itself.

2. An Experienced Professional or a Career-Shifter

If you're an experienced professional in I.T., programming, statistics, mathematics, software engineering, and the like, there's not much transitioning needed. Learning materials for Python, R, or statistics for data science (whichever competency you need to know) are abundant online. Just make sure you learn from reputable sources like known MOOC sites. Read reviews before enrolling.

If you're a career-shifter, no problem! The same advice goes out to you guys, but of course, the road will be longer and probably harder depending on your level of background knowledge on vital hard skills such as programming, coding, or statistics.

Points to Ponder

Whether you're an undergrad earning your bachelor's or an experienced professional looking to try your hands on data science, a self-inventory of your hard and soft skills are necessary. Do you have any of the hard skills we just tackled here – Python programming, R programming and or statistics? What is your level of knowledge and experience with any of these? How can you leverage your knowledge into, say, a job application for a data science post?

Final Words

The field of Data Science is all about what you can bring to the table and collaboration. Of course, you need to be equipped with any of the basic skills like Python, R, and statistics, or all of them, but having experience working within a team and in an industry that could greatly benefit from data science aside from the business and financial world like healthcare or the military sciences, are sure assets. At the core of this interdisciplinary field is teamwork and collaboration – professionals with different backgrounds and levels of expertise in data science working together in scavenging for bits and pieces of data to put them together to paint a picture of what's going on and what to do best. This is what data science is in a nutshell.

Points to Ponder

Teamwork and collaboration are often undervalued soft skills in any industry of job post. Research for any use cases where collaboration was exhibited in the different processes of data science.

REFERENCES

1. Pierson, Lillian (2017). *Data Science for Dummies*. John Wiley & Sons, Inc., Hoboken, New Jersey.
2. Data Science (2017). *Uses of Data Science Today*. Data Science YouTube Channel. November 24, 2017.
3. Glassdoor (2020). *50 Best Jobs in America for 2020*. https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm
4. U.S. Bureau of Labor Statistics (2020). *Occupational Employment and Wages, May 2019, 15-2098 Data Scientists and Mathematical Science Occupations, All Other.* https://www.bls.gov/oes/current/oes152098.htm#(1)
5. Nussbaumer Knaflic, Cole (2015). *Storytelling with Data*. Talks at Google YouTube Channel. November 12, 2015.
6. Przybyla, Matt (2020). *Data Science vs. Machine Learning. Here's the Difference.* https://towardsdatascience.com/data-science-vs-machine-learning-heres-the-difference-530883d6de3a
7. Upadrashta, Pradyumna S.(2019). *What is the Difference Between a Data Scientist and a Machine Learning Engineer? (Answer)* https://www.quora.com/What-is-the-difference-between-a-data-scientist-and-a-machine-learning-engineer
8. Sasikumar, Srihari (2020). *Data Science vs. Data Analytics vs. Machine Learning: Expert Talk*. https://www.simplilearn.com/data-science-vs-data-analytics-vs-machine-learning-article#:~:text=Because%20data%20science%20is%20a,machine%20or%20a%20mechanical%20process.
9. Simplilearn (2017). *Data Science vs. Big Data vs. Data Analytics.* Simplilearn YouTube Channel.
10. Thakur, Naresh (2020). *The Differences Between Data Science, Artificial Intelligence, Machine Learning, and Deep Learning*. https://medium.com/ai-in-plain-english/data-science-vs-artificial-intelligence-vs-machine-learning-vs-deep-learning-50d3718d51e5
11. PluralSight (2016). *5 Reasons You Should Learn Python Now. https://www.pluralsight.com/blog/software-development/why-python#:~:text=Python%20is%20easy%20to%20use,waste%20time%20with%20confusing%20syntax.*
12. Hardson-Hurley, Krishelle (2018). *11 Beginner Tips for Learning Python Programming*. https://realpython.com/python-beginner-tips/
13. Python.org (2021). *General Python FAQ. https://docs.python.org/3/faq/general.html#what-is-python*
14. Python.org (2021). *The Python Standard Library*. https://docs.python.org/3/library/index.html#library-index
15. Eggleston, Liz (2020). *How is Python Used for Data Science?* https://www.coursereport.com/blog/how-is-python-used-for-data-science-metis#:~:text=Python%20is%20a%20general%20purpose,academic%20research%20and%20statistical%20models.
16. Wrathematics (2011). *How Much of R is Written in R?* https://librestats.com/2011/08/27/how-much-of-r-is-written-in-r/
17. Cotton, Richie (2020). *Python vs. R for Data Science: What's the Difference?* https://www.datacamp.com/community/blog/when-to-use-python-or-r
18. Hornik, Kurt (2020). *Frequently Asked Questions on R. https://cran.r-project.org/doc/FAQ/R-FAQ.html#What-is-R_003f*

19. Data-Driven Science (2018). *Python vs. R for Data Science: And the winner is..* https://medium.com/@datadrivenscience/python-vs-r-for-data-science-and-the-winner-is-3ebb1a968197
20. Elite Data Science (No Publishing Date). *How to Learn Statistics for Data Science, The Self-Starter Way. https://elitedatascience.com/learn-statistics-for-data-science*
21. Bock, Tim (No Publishing Data). *Statistics vs. Data Science: What's the Difference?* https://www.displayr.com/statistics-vs-data-science-whats-the-difference/#:~:text=Statistics%20is%20a%20mathematically%2Dbased,in%20a%20range%20of%20forms.
22. Intellipaat (No Publishing Date). *Data Science vs. Artificial Intelligence. https://intellipaat.com/blog/data-science-vs-artificial-intelligence-difference/*
23. Mittal, Vartul (2018). *Ten Steps for Analyzing Unstructured Data*. https://vartulmittalspeaker.medium.com/10-steps-for-analyzing-unstructured-data-1b4f48544c9a
24. Michigan State University (2019). *Actionable Tips to Analyze Unstructured Data*. https://www.michiganstateuniversityonline.com/resources/business-analytics/actionable-tips-to-analyze-unstructured-data/
25. IBM Cloud Education (2020). *Data Modeling*. https://www.ibm.com/cloud/learn/data-modeling
26. Data Natives (2019). *Getting Into Data Modeling and Visualisation -Discover What's Hidden in Data. https://datanatives.io/getting-into-data-modeling-and-visualisation-discover-whats-hidden-in-data/#:~:text=This%20is%20called%20data%20visualization,data%20objects%2C%20associations%20and%20rules.*
27. Welhs, Claus, and Ickstadt, Katja (2018). *Data Science: the Impact of Statistics.* International Journal of Data Science and Analytics (2018) 6:189–194. https://doi.org/10.1007/s41060-018-0102-5