



Open in app

Get started



Published in Towards Data Science



Rashi Desai

Follow

Jan 3, 2020 · 9 min read · ✨ · 🎧 Listen



Save



Top 10 Skills for a Data Scientist

That will get you hired in data.

TOP 10 SKILLS FOR DATA SCIENCE

With gallons of coffee to clear up the inbox, welcome back to the grind! 😊



[Open in app](#)[Get started](#)

Per my understanding, Data Science has always been about combining the tools best suited to get the job done. It is about the extraction of knowledge from data to answer a particular question. For me, putting it simply, data science is a power that allows businesses and stakeholders to make informed decisions and solve problems with data.

Now, not every technologist is passionate about every other skill, but she would be excited about skills from her area of work. So are some of the skills for a Data Scientist. As we gear up for new technology trends and more significant challenges to solve in the new year, it is essential that we set our base strong.

In no particular order, let's get to know the *Top 10 Skills for a Data Scientist in 2020!*

1. Probability & Statistics

Data Science is about using capital processes, algorithms, or systems to extract knowledge, insights, and make informed decisions from data. In that case, making inferences, estimating, or predicting form an important part of Data Science.

Probability with the help of statistical methods helps make estimates for further analysis. Statistics is mostly dependent on the theory of probability. Putting it simply, both are intertwined.

What can you do with Probability and Statistics for Data Science?

1. Explore and understand more about the data
2. Identify the underlying relationships or dependencies that may exist between two variables
3. Predict future trend or forecast a drift based on the previous data trends
4. Determine patterns or motive of the data
5. Uncover anomalies in data

Especially for data-driven companies where stakeholders depend on data for decision





Open in app

Get started

2. Multivariate Calculus & Linear Algebra

Most machine learning, invariably data science models, are built with several predictors or unknown variables. A knowledge of multivariate calculus is significant for building a machine learning model. Here are some of the topics of math you can be familiar with to work in Data Science:

1. Derivatives and gradients
2. Step function, Sigmoid function, Logit function, ReLU (Rectified Linear Unit) function
3. Cost function (most important)
4. Plotting of functions
5. Minimum and Maximum values of a function
6. Scalar, vector, matrix and tensor functions

Summary

Linear Algebra for Data Science: Matrix algebra and eigenvalues

Calculus for Data Science: Derivatives and gradients

Gradient Descent from Scratch: Implement a neural network from scratch

3. Programming, Packages and Softwares

Of course! Data Science essentially is about programming. Programming Skills for Data Science brings together all the fundamental skills needed to transform raw data into actionable insights. While there is no specific rule about the selection of programming language, Python and R are the most favored ones.





Open in app

Get started

hand. Python, however, seems to have become the closest thing to a lingua franca for data science.

Read more about the *Top 10 Python Libraries for Data Science* [here](#).

In no particular order, here's a list of programming languages and some packages for Data Science to choose from:

1. Python
2. R
3. SQL
4. Java
5. Julia
6. Scala
7. MATLAB
8. TensorFlow (great for Data Science in Python)

And, I am not writing *What can you do with programming skills in Data Science* 😊

Everything below down from here is about coding. Data Science, without familiarity with coding experience or knowledge, can be a bit difficult. I, therefore, prefer to brush up my Python skills first, read literature about the project I'd be working and then start building up the code.

4. Data Wrangling

Often the data a business acquires or receives is not ready for modeling. It is, therefore,





Open in app

Get started

insights. For data wrangling, you basically acquire data, combine relevant fields, and then cleanse the data.

What can you do with Data Wrangling for Data Science?

1. Reveal a deep-lying intelligence within your data by gathering data from multiple channels
2. Provide a very accurate representation of actionable data in the hands of business and data analysts in a timely matter
3. Reduce processing time, response time, and the time spent to collect and organize unruly data before it can be utilized
4. Enable data scientists to focus more on the analysis of data, rather than the cleaning part
5. Lead the data-driven decision-making process in a direction supported by accurate data

5. Database Management

For me, data scientists are different people, master of all jacks. They have to know math, statistics, programming, data management, visualization, and what not to be a “full-stack” data scientist.

As I mentioned earlier, 80% of the work goes into preparing the data for processing in an industry setting. With heaps and large chunks of data to work on, it is quintessential that a data scientist knows how to manage that data.

Database Management quintessentially consists of a group of programs that can edit, index, and manipulate the database. The DBMS accepts a request made for data from an application and instructs the OS to provide specific required data. In large systems, a





Open in app

Get started

2. Manipulate the data itself, the data format, field names, record structure, and file structure
3. Defines rules to write, validate and test data
4. Operate on record-level of database
5. Support multi-user environment to access and manipulate data in parallel

Some of the popular DBMS include: MySQL, SQL Server, Oracle, IBM DB2, PostgreSQL and NoSQL databases (MongoDB, CouchDB, DynamoDB, HBase, Neo4j, Cassandra, Redis)

6. Data Visualization

What does data visualization necessarily mean? For me, it is a graphical representation of the findings from the data under consideration. Visualizations effectively communicating and lead the exploration to the conclusion.

I am a Data Visualization person at core. It gives me the power to craft a story from data and create a comprehensive presentation. Data Visualization is one of the more essential skills because it is not just about representing the final results, but also understand and learn the data and its vulnerability.

It is always better to portray things visually; the real value is well-established and understood. When I create a visualization, I am sure to get meaningful information, which can be surprising out it holds power to influence the system.

Histograms, Bar charts, Pie charts, Scatter plots, Line plots, Time series, Relationship maps, Heat maps, Geo Maps, 3-D Plots, and a long list of visualizations you can use for your data. For a more detailed list, visit [here](#).

What can you do with Data Visualization for Data Science?



[Open in app](#)[Get started](#)

4. Identify factors that influence customer behavior
5. Understand which products to place where
6. Display trends from news, connections, websites, social media
7. Visualize volume of information
8. Client reporting, employee performance, quarter sales mapping
9. Devise marketing strategy targeted to user segments

Some of the popular Data Visualization tools include: Tableau, PowerBI, QlikView, Google Analytics (For Web), MS Excel, Plotly, Fusion Charts, SAS

7. Machine Learning / Deep Learning

If you work with a company that manages and operates on vast amounts of data, where the decision-making process is data-centric, it may be the case that a demanded skill is Machine Learning. ML is a subset of the Data Science ecosystem, just like Statistics or Probability that contributes to the modeling of data and obtaining results.

Machine Learning for Data Science includes algorithms that are central to ML; K-nearest neighbors, Random Forests, Naive Bayes, Regression Models. PyTorch, TensorFlow, Keras also find its usability in Machine Learning for Data Science

What can you do with Machine Learning for Data Science?

1. Fraud and Risk Detection and Management
2. Healthcare (one of the booming Data Science fields! Genetics, Genomics, Image analysis)
3. Airline route planning





Open in app

Get started

6. Improved Interactive Voice Response (IVR)

7. Comprehensive language and document recognition and translation

8. Cloud Computing

The practice of data science often includes the use of cloud computing products and services to help data professionals access the resources needed to manage and process data. [customerthink.com] An everyday role of a Data Scientist generally includes analyzing and visualizing data that are stored in the cloud.

You may have read that data science and cloud computing go hand in hand, typically because Cloud computing gives a hand to data scientists to use platforms such as AWS, Azure, Google Cloud that provides access to databases, frameworks, programming languages, and operational tools.

Familiar with the fact that data science includes interaction with large volumes of data, given the size and the availability of tools and platforms, understanding the concept of cloud and cloud computing is not just a pertinent but critical skill for a data scientist.

What can you do with Cloud Computing for Data Science?

1. Data Acquisition
2. Parsing, munging, wrangling, transforming, analyzing and sanitizing data
3. Data mining [Exploratory Data Analysis (EDA), summary statistics, ...]
4. Validate and test predictive models, recommender systems, and such models
5. Tune the data variables and optimize model performance

Some popular cloud platforms for Data Science include Amazon Web Services, Windows Azure, Google Cloud, or IBM Cloud. I also read sometime back that people are now



[Open in app](#)[Get started](#)

9. Microsoft Excel

We know MS Excel as probably one of the best and most popular tools to work with data. We might be hearing, “*Hey, did you receive the Excel boss sent?* Wait, aren’t we discussing skills for Data Science? Excel? I always wondered there must be some easy way to manage data. Over time, exploring Excel for data management, I realized, Excel is:

1. Best editor for 2D data
2. A fundamental platform for advanced data analytics
3. Get a live connection to a running Excel sheet in Python
4. You can do whatever you want, whenever you want and save as many versions as you prefer
5. Data manipulation is relatively easy

Most non-technical people today often use Excel as a database replacement. It may be a wrong usage because it lacks version control, accuracy, reproductivity, or maintainability to some extent. However, what Excel can do is somewhat surprising as well!

What can you do with Excel for Data Science?

1. Naming and creating ranges
2. Filter, sort, merge, trim data
3. Create Pivot tables and charts
4. Visual Basic for Applications (VBA) [Google it if you don’t know already. It’s an MS Excel superpower, and this space won’t do justice to its explanation. VBA is the programming language of Excel which allows you to run loops, macros, if..else]
5. Clean data: remove duplicate values, change references between absolute, mixed and relative
6. Look-up required data among thousands of records





Open in app

Get started

10. DevOps

I've always heard and believed that Data Science is for someone who knows mathematics, statistics, algorithms, and data management. Now, some time back, I met someone with 6+ years of experience in core DevOps looking for a career change to Data Science. A curious me looked in if and how DevOps can be a part of the Data Science. I don't know much (actually, anything) about DevOps, but one thing was for sure: The growing significance of DevOps for Data Science.

DevOps is a set of methods that combines software development and IT operations that aims to shorten the development life cycle and provide uninterrupted delivery with high software quality.

DevOps teams closely work with the development teams to manage the lifecycle of applications effectively. Data transformation demands close collaboration of data science teams with DevOps. DevOps team is expected to provide highly available clusters of Apache Hadoop, Apache Kafka, Apache Spark, and Apache Airflow to tackle data extraction and transformation.

What can be done with DevOps for Data Science?

1. Provision, configure, scale and manage data clusters
2. Manage information infrastructure by continuous integration, deployment, and monitoring of data
3. Create scripts to automate the provisioning and configuration of the foundation for a variety of environments.

Thank you for reading! I hope you enjoyed the article. Do let me know what skill are you looking forward to learning or exploring in your Data Science journey?

Happy Data Tenting!



[Open in app](#)[Get started](#)

Know your author

Rashi is a graduate student at the University of Illinois, Chicago. She loves to visualize data and create insightful stories. She is a User Experience Analyst and Consultant, a Tech Speaker, and a Blogger.

Sign up for The Variable

By Towards Data Science

Every Thursday, the Variable delivers the very best of Towards Data Science: from hands-on tutorials and cutting-edge research to original features you don't want to miss. [Take a look.](#)

By signing up, you will create a Medium account if you don't already have one. Review our [Privacy Policy](#) for more information about our privacy practices.



Get this newsletter

[About](#) [Help](#) [Terms](#) [Privacy](#)

Get the Medium app

