What do Data Scientists do?

1. **Overview**

* A Day in the Life of a Data Scientist
* Data Science Skills and Big Data
* Working on Different File Formats
* Data Science Topics and Algorithms
* What Makes Someone a Data Scientist?
* Lesson Glossary

1. **A Day in the Life of a Data Scientist**

[Link](https://www.coursera.org/learn/what-is-datascience/lecture/qp9yD/a-day-in-the-life-of-a-data-scientist)

1. **Data Science Skills and Big Data**

Aspiring data scientists need to learn many tools such as Python, UNIX commands, pandas, and Jupyter notebook.

1. **Working on Different File Formats**

Standard file formats:

* Delimited text file formats, or .CSV:
  + Files used to store data as text. Each value is seperated be a delimiter.
  + Delimiter – A sequence of one or more characters for specifying the boundary between independent entities or values. (Comma, Tab, Colon, Vertical Bar, Space)
  + Popular file formats: Comma-separated values (.CSV), Tab-seperated values (.TSV)
* Microsoft Excel Open .XML Spreadsheet, or .XLSX
  + Is a Microsoft Excel Open XML file format that falls under the spreadsheet file format. It is an XML-based file format created by Microsoft.
  + Open file format, acessible to most other applications.
  + Can use and save all functions avaibale in Excel
  + Is a secure file format as it cannot save malicious code
* Extensible Markup Language, or .XML
  + Is a markup language with set rules for encoding data.
  + Readable by both humans and machines
  + Similar to .HTML in some respects
  + Does not use predefined tags like .HTML does
  + Platform independent
  + Programming language independent
  + Makes it simplier to share data betweem systems
* Portable Document Format, or .PDF
  + Is a file format develop by Adobe to present documents independent of application software, hardware and operating systems.
  + Can be viewed the same way on any device
  + Is frequently used in legal and financial documents
  + Can also be used to fill in data for forms
* JavaScript Object Notation, or .JSON
  + Is a text-based open standard designed for transmitting structured data over the web.
  + Language-independent data format
  + Can be read in any programming language
  + Easy to use
  + Compatible with a wide range of browsers
  + Considered as one of the best tools for sharing data

1. **Data Science Topics and Algorithms**

* Regression
* Data visualization
* Artificial Neural Network
* Nearest Neighbor
* Turn Unstructured Data to Structured Data

Using complicated machine learning algorithms does not always guarantee achieving a better performance. Occasionally, a simple algorithm such as k-nearest neighbor can yield a satisfactory performance comparable to the one achieved using a complicated algorithm. It all depends on the data.

1. **What Makes Someone a Data Scientist?**

[Link](https://author-ide.skills.network/render?token=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJtZF9pbnN0cnVjdGlvbnNfdXJsIjoiaHR0cHM6Ly9jZi1jb3Vyc2VzLWRhdGEuczMudXMuY2xvdWQtb2JqZWN0LXN0b3JhZ2UuYXBwZG9tYWluLmNsb3VkL0lCTURldmVsb3BlclNraWxsc05ldHdvcmstRFMwMTAxRU4tU2tpbGxzTmV0d29yay9sYWJzL01vZHVsZSUyMDEvUmVhZGluZ19XaGF0X01ha2VzX1NvbWVvbmVfRGF0YVNjaWVudGlzdC5tZCIsInRvb2xfdHlwZSI6Imluc3RydWN0aW9uYWwtbGFiIiwiYWRtaW4iOmZhbHNlLCJpYXQiOjE3MTE1NjI2NDB9.6OVc4ax6zhLriIOeHaSWARDYpZa9G9uBnuVYVfv9SPc)

Notable tools:

* Hadoop
* Stata

1. **Summary: what do data scientists do?**

* Problem investigation
* Environmental Problem investigation
* Simplify Problems
* Education
* Mathematical sciences
* Models and algorithms
* Ability to analyze data
* Storytelling with data
* Data scientist =
  + Curiosity +
  + Computer scientist +
  + Software engineer +
  + Statistician