Explorative Study of Data Science Model and Its Tools Set Comparison

**Jibran Rasheed Khan1, 2, Idrees Riaz2, Muhammad Waqar2**

1Research Officer at HEJ, ICCBS, University of Karachi

2Department of Computer Science Information & Technology – CSIT, NED University of Engineering and Technology.

# ABSTRACT

Data Science (DS) is a scientific technique to unfold the hidden mystery of data and extract the valuable insight that can boom the business needs, beware of what probably be happened in near future and most importantly make human proactive to act than react on events. Nowadays, data is the key element that use to classify and aids in to foretell each and every aspect of human; behavior, nature, standards, lifestyle, passion, desire, business needs and lot more uncountable. But, data is not so simple nor resides in any specified form. This usually be available in any of 4Vs; Volume, Velocity, Verity and Veracity. Due to complex nature of data, many organization, scientist and practitioners suffer trouble to incorporate them in a suitable platform, desirable tools, libraries and other supportive algorithms or calculations. Also, data science itself not a separate field of study, but it’s a composition of several incorporating fields – Information Technology, Statistics and Science. The core propose of this explorative study is focused on the learning DS generic model and comparison of its tools set in the market. In this work, we will explore each aspect and compositor of data Science, drill down to identify individual field hierarchy, tools, and platform available. Furthermore, it covers the concise comparison study renewed and most preferable element in each filed. The study, will beneficial in identifying actual requirements and composite tool that will ease the practitioners to learn more, building better and flexible system and more compatible integrating features.

# INTRODUCTION

**Data Science (DS)** is a fancy buzzy term that most people unable to understand what is really deal with. DSis a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. Data science is the same concept as data mining and big data: "use the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problems". More clear understanding gain by know what data and scientific methodologies actually make it to call DS.

## Data

It is something that could be measured, collected, reported, analyzed, stored, and whereupon it can be visualized using graphs, images or other analysis tools. By another and simple definition state it as “Data, in general concept refers to the facts that has information or knowledge”.

***Data Types***: The core form of data classified as Number, Text, Image, Audio and Video data. Furthermore, these five core data forms can be achieved from Observational, Simulation, Experimental, Derived or Compiled, Machine, Time-stamped, Spatiotemporal, Open, Dark, Real time and Operational data. The data umbrella is illustrated in the figure 1.

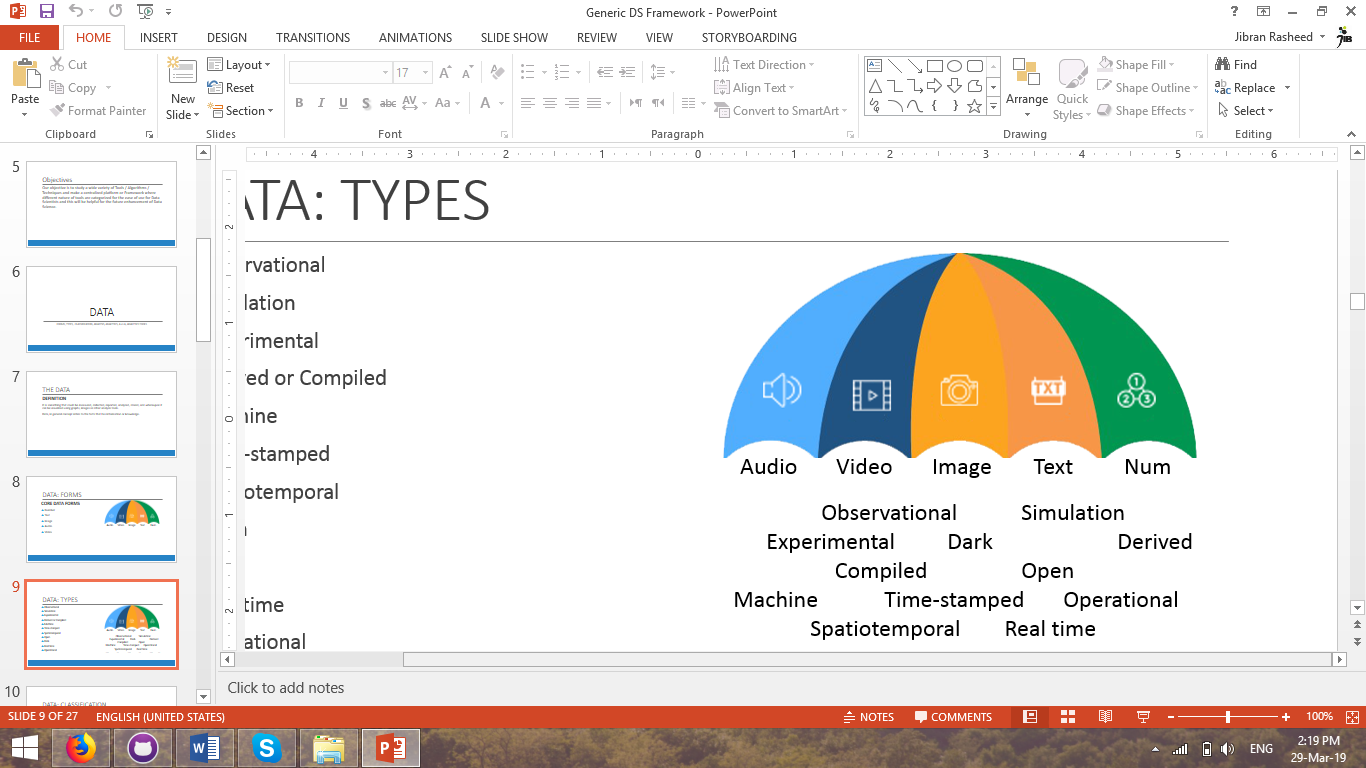


Figure . Data umbrella

*Data Classification*: Data classification is broadly defined as the process of organizing data by relevant categories so that it may be used and protected more efficiently. On a basic level, the classification process makes data easier to locate and retrieve. On the basis of definition, the above said data forms can be classified into different level

* **First Level**
* Structured
* Semi-structured
* Unstructured
* **Second Level**
* Qualitative
* Quantitative
* Geographical Base
* Chronological or Temporal

***Data Analysis*:** It is the process of examining information, especially using a computer, in order to find out something, or to help with making decisions out of it.

**Example Tools**

* OpenRefine, KNIME,
* RapidMiner, Google Fusion Tables,
* Tableau Public, NodeXL,
* WolframAlpha tools are used.

***Data Analytics***: It is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software.

**Example Tools**

* R, Tableau Public,
* Python, SAS,
* Apache Spark,
* Excel, etc

## Analysis vs Analytics

There is a very fine difference between the data analysis and data analytics. It is important to understand what you are going to do with data. Table 1 presents the comparative difference between analysis and analytics

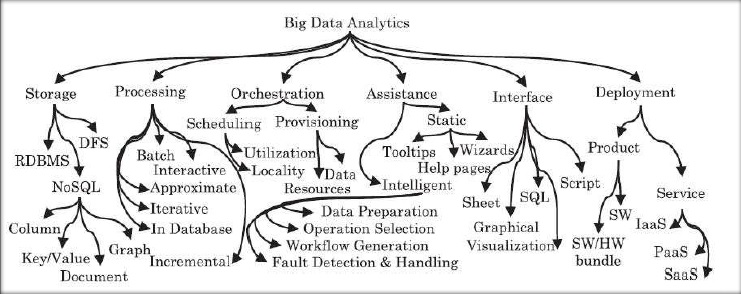
|  |  |  |
| --- | --- | --- |
| **Basis** | **Data Analytics** | **Data Analysis** |
| Form | Used in businesses to make decisions from data which are data-driven | used in businesses to analyze data and take some insights of it. |
| Structure | Consist of data collection and inspect in general and it has one or more users. | Consisted of defining a data, investigation, cleaning, transforming the data to give a meaningful outcome. |
| Tools | Example tools in a market but mainly R, Tableau Public, Python, SAS, Apache Spark, Excel are used. | Example tools OpenRefine, KNIME, RapidMiner, Google Fusion Tables, Tableau Public, NodeXL, WolframAlpha. |
| Process Sequence | Business Case Evaluation, Data Identification, Acquisition & Filtering, Extraction, Validation & Cleansing, Aggregation & Representation, Data Analysis, Visualization, Utilization of Analysis Results. | Data gathering, scrubbing, analysis and interpret precisely so that one can understand what data want to say. |
| Usage | It can be used to find masked patterns, anonymous correlations, customer preferences, market trends that can aid in to notify decisions for business purpose. | It may be descriptive analysis, exploratory analysis, inferential analysis, predictive analysis and take useful insights from the data. |
| Example | 1GB customer purchase data of past 1 year, one has to find that what our customers next possible purchases. | 1GB customer purchase data of past 1 year, one trying to find what happened so far, having look into past. |

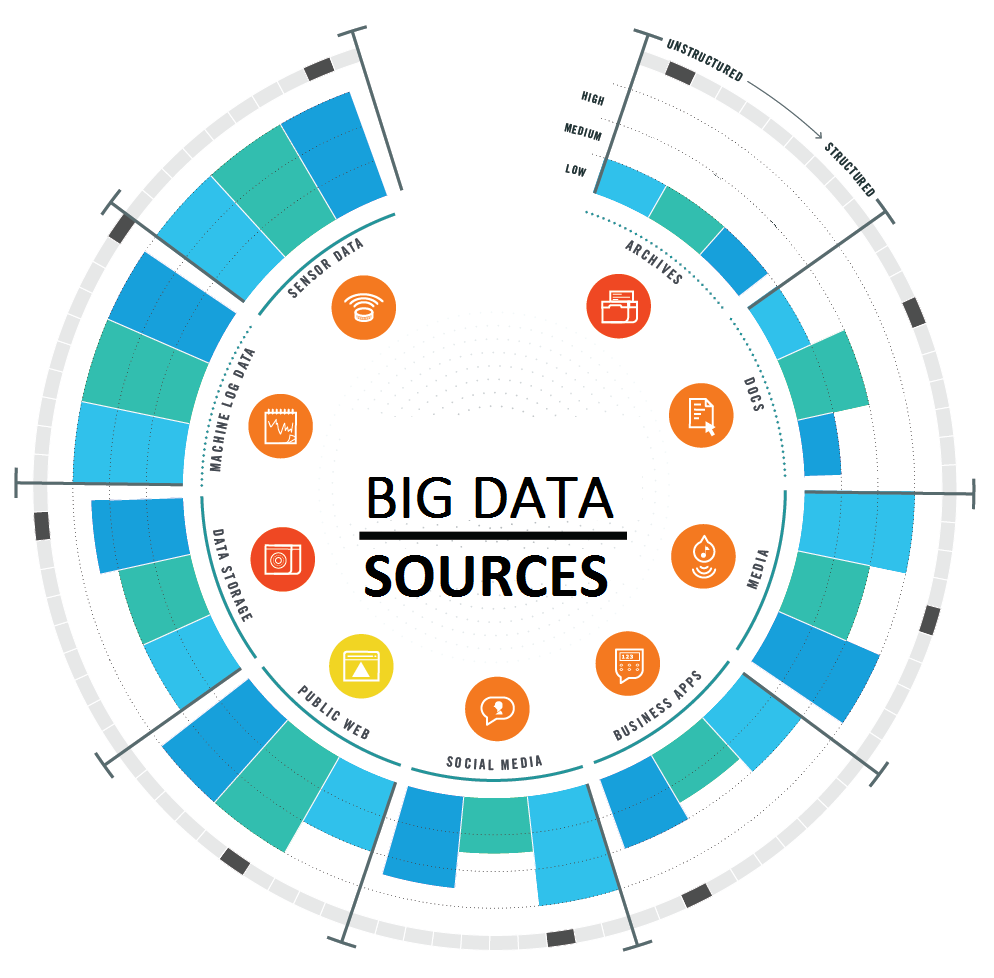
## Big Data

The term big data refers to extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.

Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it’s not the amount of data that’s important. It’s what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.

The combination of six key component fulfill the big data technology stake, these six are also known as big data pillars; Storage, Processing, Orchestration, Assistance, Interface and Deployment as shown in figure 2 below. As discussed above, data could be in many form the potential data sources in today are illustrated in the figure 3.





## Algorithms

In mathematics and computer science, an algorithm is an unambiguous specification of how to solve a class of problems. By another definition, it is a procedure or formula for solving a problem, based on sequence of specified actions. Before driving into data science, it is important to understand what does algorithm do and what is used for.

**WHAT IT DOES?**

* Follow instructions code
* Calculation (Logical/Mathematical)
* Data Manipulation
* Data Processing

**IT IS USED FOR**

* Searching
* Sorting
* Scheduling
* Automated Reasoning
* Automate Operations/Tasks

Algorithms have widespread range, almost in every expect of study and life. Here only concern with the Statistical/Computational statistics, Data mining, Machine learning algorithms. Based on the operation and used, algorithms can be classified into four types are listed below with their renowned algorithm examples.

1. Supervised Learning  
   **e.g:** Nearest Neighbor, Naive Bayes, Decision Trees, Linear Regression, Support Vector Machines (SVM), Neural Networks
2. Unsupervised Learning  
   **e.g:** k-means clustering, Association Rules, mixture models, Self-organizing map, Expectation–maximization algorithm (EM)
3. Semi-Supervised Learning  
   **e.g:** Generative method, Low-density separation method, Graph-based method, Heuristic method
4. Reinforcement Learning  
   **e.g:** Monte Carlo, Q-learning, SARSA, DQN, A3C, NAF, PPO, TRPO

The other way to classify them is based on the algorithms learning approaches as presented in figure 4.

# DATA ANALYSIS & ANALYTICS TOOLS

Although, the data science is not much aged, but it captured attention of everyone across the world. This is the reason lots of companies take part in building handy tools to do the complicated tasks comfortably by offering wide range of tools. It is impossible to list every tool, we focus on some most prominent and high rated tools in the following section.

## **Microsoft HDInsight**

Azure HDInsight is a Spark and Hadoop service in the cloud. It provides big data cloud offerings in two categories, Standard and Premium. It provides an enterprise-scale cluster for the organization to run their big data workloads. Azure HDInsight is an easy, cost-effective, enterprise-grade service for open source analytics that enables customers to easily run popular Apache open source frameworks including Apache Hadoop, Spark, Kafka, and others. The service is available in 27 public regions and Azure Government Clouds in the US and Germany.

Azure HDInsight powers mission critical applications in a wide variety of sectors and enables a wide range of use cases including ETL, streaming, and interactive querying.

# Features

* Reliable analytics with an industry-leading SLA
* It offers enterprise-grade security and monitoring
* Protect data assets and extend on-premises security and governance controls to the cloud
* High-productivity platform for developers and scientists
* Integration with leading productivity applications
* Deploy Hadoop in the cloud without purchasing new hardware or paying other up-front costs

## **SkyTree**

**Skytree, Inc** is a San Jose, California-based startup company that develops machine learning software for enterprise use. Skytree came out of stealth mode in February 2012, announced SkyTree Server. It is a machine learning system to discover patterns and make predictions from complex, massive data. Skytree Server is an enterprise-grade software platform for developing, testing, and deploying advanced analytics solutions for big data. It is designed from the ground up to work on massive, fast changing, and varied datasets with high degree of efficiency and accuracy. Skytree Server’s scalable architecture performs state-of-the-art machine learning methods that were previously not possible on very large data sets. The fundamental design of Skytree Server incorporates advanced algorithms from the latest machine learning research to achieve speeds that are tens of thousands of times faster than existing approaches.

# Features

* Highly Scalable Algorithms
* Artificial Intelligence for Data Scientists
* It allows data scientists to visualize and understand the logic behind ML decisions
* SkyTree via the easy-to-adopt GUI or programmatically in Java
* Model Interpretability
* It is designed to solve robust predictive problems with data preparation capabilities
* Programmatic and GUI Access

## **Talend**

Talend is a big data tool that simplifies and automates big data integration. Its graphical wizard generates native code. It also allows big data integration, master data management and checks data quality. In today’s Big data and the cloud-centric world, it becomes very important for the organizations to harness their enterprise information. Talend is an open source software integration platform helps you in effortlessly turning this data into business insights. The ever-growing demand for [Talend Certification](https://www.edureka.co/talend-for-big-data) today is proof of its worth in the market.

# Features of Talend

* Accelerate time to value for big data projects
* Simplify ETL & ELT for big data
* Talend Big Data Platform simplifies using MapReduce and Spark by generating native code
* Smarter data quality with machine learning and natural language processing
* Agile DevOps to speed up big data projects
* Streamline all the DevOps processes

## **Splice Machine**

Splice Machine is a big data analytic tool. It is designed to scale real-time applications using commodity hardware without application rewrites. The Splice Machine database is a modern, scale-out alternative to traditional RDBMSs, such as Oracle®, MySQL™, IBM DB2® and Microsoft SQL Server®, that can deliver over a 10x improvement in price/performance. As a full-featured SQL-on-Hadoop RDBMS with ACID transactions, the Splice Machine database helps customers power real-time applications and operational analytics, especially as they approach big data scale. Their architecture is portable across public clouds such as AWS, Azure, and Google. It is designed from the ground up to be portable. Leveraging a technology stack of DC/OS, Marathon, ELK and more, applications and storage are containerized, secured and monitored with guaranteed availability. This architecture is portable across public clouds such as AWS, Azure and Google, as well as on premise infrastructure.

Splice Machine is currently available for on premise deployments and as a database service on AWS and Azure. Other cloud platforms will be added over the next few months.

# Features of Splice Machine

* It can dynamically scale from a few to thousands of nodes to enable applications at every scale
* The Splice Machine optimizer automatically evaluates every query to the distributed HBase regions
* Reduce management, deploy faster, and reduce risk
* Consume fast streaming data, develop, test and deploy machine learning models

## **Spark**

Apache Spark is a powerful open source big data analytics tool. It offers over 80 high-level operators that make it easy to build parallel apps. It is used at a wide range of organizations to process large datasets. Apache Spark has become one of the key big data distributed processing frameworks in the world. Spark can be deployed in a variety of ways, provides native bindings for the Java, Scala, Python, and R programming languages, and supports SQL, streaming data, machine learning, and graph processing. You’ll find it used by banks, telecommunications companies, games companies, governments, and all of the major tech giants such as Apple, Facebook, IBM, and Microsoft. Out of the box, Spark can run in a standalone cluster mode that simply requires the Apache Spark framework and a JVM on each machine in your cluster.

**Features**

* It helps to run an application in Hadoop cluster, up to 100 times faster in memory, and ten times faster on disk
* It offers lighting Fast Processing
* Support for Sophisticated Analytics
* Ability to Integrate with Hadoop and Existing Hadoop Data
* It provides built-in APIs in Java, Scala, or Python

## **Plotly**

Plotly was founded by Alex Johnson, Jack Parmer, Chris Parmer, and Matthew Sundquist. it is an analytics tool that lets users create charts and dashboards to share online. Plotly provides online graphing, analytics, and statistics tools for individuals and collaboration, as well as scientific graphing libraries for Python, R, MATLAB, Perl, Julia, Arduino, and REST.

**Features:**

* Easily turn any data into eye-catching and informative graphics
* It provides audited industries with fine-grained information on data provenance
* Plotly offers unlimited public file hosting through its free community plan

## **Lumify**

Lumify is an open source project big data fusion, analysis, and visualization platform designed for anyone. Its intuitive web-based interface helps users discover connections and explore relationships in their data via a suite of analytic options, including 2D and 3D graph visualizations, full-text faceted search, dynamic histograms, interactive geographic maps, and collaborative workspaces shared in real-time. Lumify is a big data fusion, analysis, and visualization platform. It helps users to discover connections and explore relationships in their data via a suite of analytic options.

**Features:**

* It provides both 2D and 3D graph visualizations with a variety of automatic layouts
* It provides a variety of options for analyzing the links between entities on the graph
* It comes with specific ingest processing and interface elements for textual content, images, and videos
* It spaces feature allows you to organize work into a set of projects, or workspaces
* It is built on proven, scalable big data technologies

## **Elasticsearch**

Elasticsearch is the main product of a company called ‘Elastic’. It is used for web search, log analysis, and big data analytics. It is more popular because it is easy to install, scales out to hundreds of nodes with no additional software, and is easy to work with due to its built-in REST API. It is a JSON-based Big data search and analytics engine. It is a distributed, RESTful search and analytics engine for solving numbers of use cases. It offers horizontal scalability, maximum reliability, and easy management. It gives an enhanced experience with security, monitoring, reporting, and machine learning features

**Features**

* It allows combine many types of searches such as structured, unstructured, geo, metric, etc
* Intuitive APIs for monitoring and management give complete visibility and control
* It uses standard RESTful APIs and JSON. It also builds and maintains clients in many languages like Java, Python, NET, and Groovy
* Real-time search and analytics feature to work big data by using the Elasticsearch-Hadoop

## **R-Programming:**

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. It also used for big data analysis. It provides a wide variety of statistical tests. Being an open source software and having a variety of built in statistical commands, R is the most widely used tool by statisticians and Data analysts. There are a lot of new packages being developed and old packages constantly updated in R to support and handle Big Data. This is an advantage mainly for R users as they don’t have to migrate to other platforms to handle their huge volumes of data.

**Features**

* Effective data handling and storage facility,
* It provides a suite of operators for calculations on arrays, in particular, matrices,
* It provides coherent, integrated collection of big data tools for data analysis
* It provides graphical facilities for data analysis which display either on-screen or on hardcopy

## **Pentaho**

Pentaho is a business intelligence (BI) software that provides data integration, OLAP services, reporting, information dashboards, data mining and extract, transform, load (ETL) capabilities. It is headquartered in Orlando, Florida. Pentaho was acquired by Hitachi Data Systems in 2015. On September 19, 2017, Pentaho became part of Hitachi Vantara, a new company that unifies the operations of Pentaho, Hitachi Data Systems and Hitachi Insight Group. Pentaho provides big data tools to extract, prepare and blend data. It offers visualizations and analytics that change the way to run any business. This Big data tool allows turning big data into big insights.

**Features:**

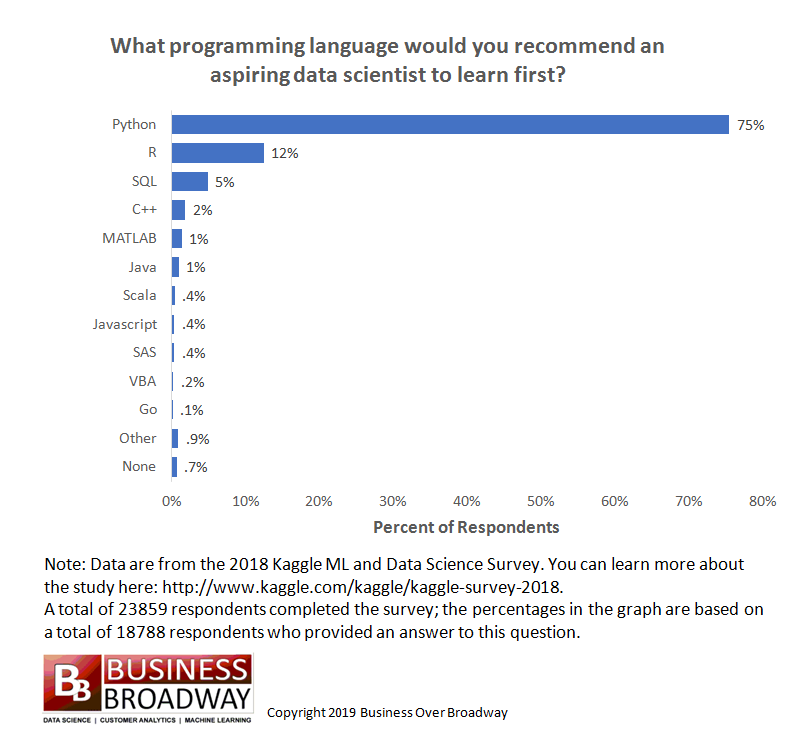
* Data access and integration for effective data visualization
* It empowers users to architect big data at the source and stream them for accurate analytics
* Seamlessly switch or combine data processing with in-cluster execution to get maximum processing
* Allow checking data with easy access to analytics, including charts, visualizations, and reporting
* Supports wide spectrum of big data sources by offering unique capabilities

# BEST DS PROGRAMMING LANGUAGES

The practice of data science requires the use of analytics tools, technologies and programming languages to help data professionals extract insights and value from data. A recent survey of nearly 24,000 data professionals by Kaggle revealed that Python, SQL and R are the most popular programming languages. The most popular, by far, was Python (83% used). Additionally, 3 out of 4 data professionals recommended that aspiring data scientists learn Python first.

Kaggle conducted a worldwide survey in October 2018 of 23,859 data professionals (2018 Machine Learning and Data Science Survey). Their survey included a variety of questions about data science, machine learning, education and more. Kaggle released the raw survey data and many of their members have analyzed the data (see link above). I will be exploring their survey data over the next couple of months. When I find something interesting, I’ll be sure to post it here on my blog. Today’s post is about the data science and machine learning programming languages data professionals used in 2018.

The survey also asked respondents what programming language they would recommend an aspiring data scientist to learn first Results showed that 3 out of 4 data professionals would recommend Python as the programming language aspiring data scientists to learn first. The remaining programming languages are recommended at a significantly lower rate (R recommended by 12% of respondents; SQL recommended by 5% of respondents. When looking at data professionals who identified as a data scientist, we find similar recommendations for aspiring data scientists: Python (78%), R (13%) and SQL (5%).

****

# BEST DS LIBRARIES

# TOP 10 DS FRAMEWORK

# REFERENCES