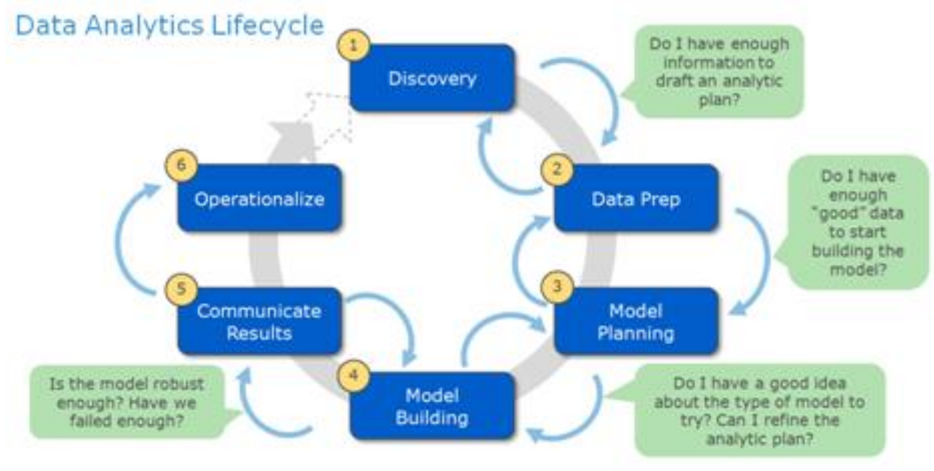
Requirements and importance of data analytics

Data Analytics refers to qualitative and quantitative techniques and processes used to enhance productivity and business gain. Data is extracted, acknowledged and bifurcated to identify and analyse behavioural data, techniques and patterns can be dynamic according to a particular business's need or requirement. Data Analytics is a broader term that has analysis as a subhead and analytics is basically the concepts used to do the analysis.


Data Analytics Lifecycle
1 Discovery
Do I have enough information to draft an analytic plan?
2 Data Prep
Do I have enough "good" data to start building the model?
3 Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4 Model Building
Is the model robust enough? Have we failed enough?
5 Communicate Results
6 Operationalize

Why Data Analytics

Data Analytics is needed in Business to Consumer applications (B2C). Organisations collect data that they have gathered from customers, businesses, economy and practical experience. Data is then processed after gathering and is categorised as per the requirement and analysis is done to study purchase patterns and etc.

The idea is to make sense of the data you have, to analyse it and share better business prospects in the near future and how you're going to do it, is with the concepts of analytics. Data Science involves extraction of trends, patterns and useful information from a set of existing data which will be of no use if not analysed. It is a kind of business intelligence that is now used for gaining profits and making better use of resources. If not analysed this data is going to get wasted whereas if analysed properly this data can help us in finding information that is powerful to bring in a change in the patterns of how business is already working or going. This is going to extract insights that will allow an advantage to a business or an organisation in an economy.

Modelling and visualising is one of the major aspects of analytics and so to get an up gear from this, you really need to understand the intricacies of it as a whole. Earlier data needed a number of skilled analysts to process data whereas we now have tools that are used in running high-speed data analytics on massive amounts of data, and this gives an opportunity to the entrepreneurs to incorporate data analytics when making decisions.

Different decisions can be made as far as your target audience is concerned, your audience can change on the basis of the analysis you have done with the help of data analytics. Social media is another example that has increased the growth of the data and your organisation can make changes based on that too.

There are a variety of ways to analyse data. Some of these methods use algorithms, predictive analytics, machine learning or artificial intelligence (AI). Others require advanced analytics or data science to make sense of unstructured data.

Here are 4 general ways data analysts and data scientists extract patterns and trends from complex data:

1. Data Mining

Most simply stated, data mining is a process used to extract usable data from a large dataset. Data mining involves data collection, warehousing and computer processing. In order to segment and evaluate the data, data mining uses advanced algorithms.

Real-Life Scenario: Data mining is often used in the health care industry during patient clinical trials. The algorithms can evaluate behavioural patterns of large amounts of data for interpretation, knowledge building and decision making.

2. Text Analytics

Text analytics is the process of drawing meaning out of written communication. Usually, text analytics software relies on text mining and natural language processing (NLP) algorithms to find patterns and meaning.

Real-Life Scenario: Text analytics is used to build the auto-correct function on your mobile device. It will not only correct your spelling, but also predict what you’re going to type next based on linguistic analysis and data pattern recognition.

3. Data Visualization

Data visualization presents a clear picture of what the data actually means. Using bar graphs, pie charts, tables and other visuals, data visualization makes the data easier for those making business decisions to comprehend.

Real-Life Scenario: Exercise rings on smart-watch, the energy-use trends from your smart thermostat and the weekly screen time charts on your phone.

4. Business Intelligence

Business intelligence (BI) is the end game. It leverages analytics tools to convert data to actionable insights. Often paired with data visualization techniques, BI provides decision makers with detailed Intel about the state of the business.

Real-Life Scenario: Retailers use BI technology to capitalize on customer trends and extend customized offers in real-time.

Key Technologies in Data Analytics

Data mining, text analytics, data visualization and business intelligence are different ways we can analyse data. There is a plethora of analytics tools available to help data analysts and data scientists do this.

Some of these key technologies are;

Data Analytics Tool	**How It's Used**
Artificial Intelligence	Makes decisions that can provide a plausible likelihood in achieving a goal
NoSQL Database	Delivers a method for accumulation and retrieval of data
R Programming	Assists data scientists in designing statistical software
Data Lakes	Accumulates data without transforming it into structured data
Predictive Analytics	Predicts future behaviour via prior data
Apache Spark	Generates big data transformation via Python, R, Scala and Java
Prescriptive Analytics	Provides guidance about what to do to achieve a desired outcome
In-Memory Database	Saves time by omitting the requirements to access hard drives
Hadoop Ecosystem	Ingests, stores, analyses and maintains large data sets
Blockchain	Distributed ledger technologies have proven valuable in managing data challenges
Microsoft Excel	Aggregates data to create reports and easy-to-use dashboards

Quality issues for data analysis

Ensuring overall data quality is paramount while preparing data for analysis and managing it on an on-going basis. There are 7 general key causes of data quality issues. In preparing your data for analysis and managing your data on an on-going basis, you should continually monitor your data to ensure these issues do not occur.

1. Duplicate data

Duplicate data is when the same data is entered multiple times, but in slightly different ways. Duplicate data is often created when extracting data from multiple siloed systems and merged together in a data warehouse, creating 'copies' of the same record. Duplication may produce skewed or incorrect insights when they go undetected.

For example: A Customer Name spelt slightly differently or Address Data with different syntax and abbreviations.

2. Inconsistent formats

Storing the same type of data in inconsistent formats is a common quality issue.

For example: Storing dates in mixed format, such as (US Date) MMM DD YYYY, (European Date) DD MMM YYYY and (Japan Date) YYYY MMM DD – all are valid, but if these formats are used in the same field, it will make effective analysis near impossible.

3. Incomplete information

This data quality issue occurs when crucial pieces of information are missing, either as a result of failure to input it at the source system, or as a result of ETL processes.

For example: An address details where ZIP/Post Code data is sparse. Without a ZIP code, the ability to conduct location analysis is significantly compromised.

4. Data inconsistency

Data inconsistency is the result of storing data in the same field that is either in a different language or in different units.

For example: Storing Volume in Metric (Litres) as well as Empirical units (Gallons) – the result of this would be that any aggregate analysis would be incorrect.

5. Inaccurate data

This is one of the most difficult data quality issues to spot, and occurs when the format is correct and every value is complete, but potential mis-spellings exist or the data is simply inaccurate.

For example: If sales opportunity data entered into your CRM system by your sales people is incorrect, then your ability to conduct forecast analysis is severely compromised.

6. Invalid data

Data Invalidity is when your data can't possibly correct based on simple rules or logic.

For example: Having values for inventory level with a negative number.

7. Data imprecision

Data imprecision or lack of precision is when data has been stored at a summarized level, as a result of an ETL process, that does not enable users to get to the level of detail they need for analysis.

Where to address data quality issues:

There are three areas in which you can address data quality issues.

1. Address the issue in the source system

The best place to address data quality issues is at the originating data source. This means addressing systems and processes involved in data capture. The challenge with addressing issues at this layer is the high level of intervention needed at a business process layer, or if the data is provided via a third party where you have no control.

2. Fix during the ETL process

If you cannot fix data quality at its source, then you can attempt to fix the issues via your ETL processes. This is often the pragmatic approach taken by many businesses. Using defined rules and smart algorithms, it is possible to fix many issues in this way – ensuring you have a clean data set to report from.

3. Fix at the meta-data layer

Lastly, if you do not have control of the ETL processes and need to analyse a data set 'as is', then you can use rules and logic within a metadata layer to fix (or mask) your data quality issues. In this way, you can apply some of the rules and logic you would have applied in an ETL process, but in this case, the underlying data is not updated. Instead, the rules are applied at run time to the query, and fixes are applied on the fly.

Business risks associated with poor data quality:

To estimate the impact of poor data quality on a business, you need to identify the role data plays in various business processes. This will help you highlight which processes are bound to mess up and cause delays if the data had any of the issues mentioned above. Below, I have listed the most common business risks associated with poor data quality.

1. Missed opportunities
A business is prone to miss opportunities on multiple fronts if they have poor data quality across disparate datasets. For example, with poor lead data, you can miss an opportunity to identify potential prospects.

2. Lost revenue
This is definitely one of the biggest risks that your business can experience due to poor data quality. Incomplete or incorrect data (either it is customer contact information, product information, or ambiguity in financial dataset) can cause you to lose potential clients and incur losses in revenue as a result.

3. Reduced operational efficiency and productivity
When an organization's workforce manually corrects data quality issues before using the data, it can put a strain on their efficiency and productivity rates. Many data analysts and data scientists feel that they spend more time preparing and cleaning data – as compared to performing analysis and forecasting reliable predictions about the business's future. For this reason, your business needs an end-to-end system that utilizes technology to automate data quality validation and implement data quality processes in time. This can transform your data into making it usable at every stage of its lifecycle – without putting in any extra effort in runtime.

4. Customer dissatisfaction
To achieve this, businesses use a ton of customer-generated data to understand their behaviour and preferences. If this data has serious defects, you will obviously end up inferring wrong details about your customers or potential buyers. This can lead to reduced customer satisfaction and brand loyalty.

5. Misanalysis
There are two ways to predict future market, demand, and need. One is to follow your instinct. The second is to look at past data to identify patterns and forecast the probable future. It is obvious that the second way is more reliable. But when it comes to business intelligence or market analysis, your insights are going to be as good as the input data. If the data fed to your analysis algorithm has multiple data quality issues, the identified patterns are going to be inaccurate – leading you to build an incorrect perception about the market's future.

6. Reputational damage

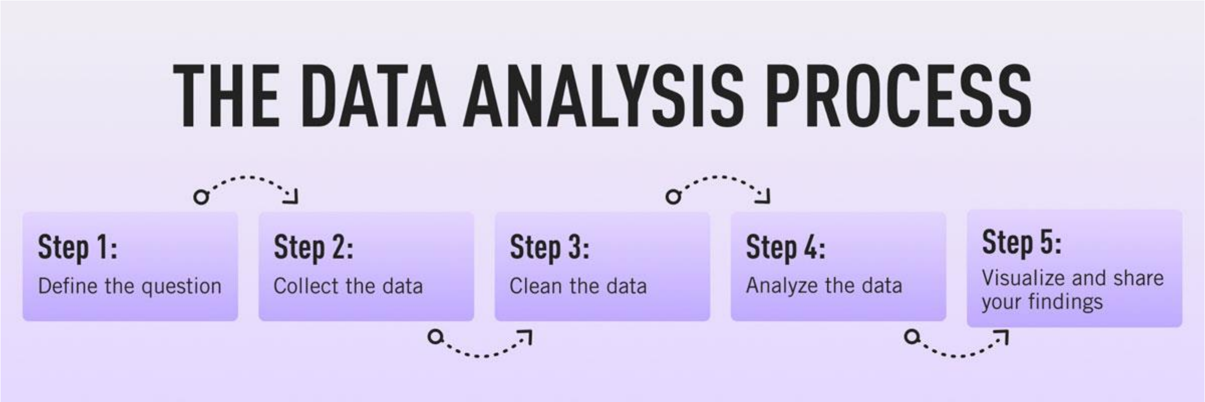
7. Lack of compliance

8. Increased financial costs

Data Analysis Task

Here are the steps involved in data analysis:

1. Defining the question
2. Collecting the data
3. Cleaning the data
4. Analysing the data


THE DATA ANALYSIS PROCESS
Step 1: Define the question
Step 2: Collect the data
Step 3: Clean the data
Step 4: Analyze the data
Step 5: Visualize and share your findings

1. Defining the question

The first step in any data analysis process is to define your objective. In data analytics jargon, this is sometimes called the 'problem statement'. Defining your objective means coming up with a hypothesis and figuring how to test it. Now you've defined a problem, you need to determine which sources of data will best help you solve it.

2. Collecting the data

Once you've established your objective, you'll need to create a strategy for collecting and aggregating the appropriate data. A key part of this is determining which data you need. This might be quantitative (numeric) data, e.g. sales figures, or qualitative (descriptive) data, such as customer reviews.

All data fit into one of three categories: first-party, second-party, and third-party data. Let's explore each one.

First-party data:
First-party data are data that you, or your company, have directly collected from customers. It might come in the form of transactional tracking data or information from your company's customer relationship management (CRM) system. Whatever its source, first-party data is usually structured and organized in a clear, defined way. Other sources of first-party data might include customer satisfaction surveys, focus groups, interviews, or direct observation.

Second-party data:
To enrich your analysis, you might want to secure a secondary data source. Second-party data is the first-party data of other organizations. This might be available directly from the company or through a private marketplace. The main benefit of second-party data is that they are usually structured, and although they will be less relevant than first-party data, they also tend to be quite reliable. Examples of second-party data include website, app or social media activity, like online purchase histories, or shipping data.

Third-party data:
Third-party data is data that has been collected and aggregated from numerous sources by a third-party organization. Often third-party data contains a vast amount of unstructured data points (big data). Many organizations collect big data to create industry reports or to conduct market research. Open data repositories and government portals are also sources of third-party data.

3. Cleaning the data

Once you've collected your data, the next step is to get it ready for analysis. This means cleaning, or 'scrubbing' it, and is crucial in making sure that you're working with high-quality data. Key data cleaning tasks include:

- Removing major errors, duplicates, and outliers—all of which are inevitable problems when aggregating data from numerous sources.
- Removing unwanted data points—extracting irrelevant observations that have no bearing on your intended analysis.
- Bringing structure to your data—general 'housekeeping', i.e. fixing typos or layout issues, which will help you map and manipulate your data more easily.

Many data analysts do (alongside cleaning data) is to carry out an exploratory analysis. This helps identify initial trends and characteristics, and can even refine your hypothesis.

4. Analysing the data:

After cleaned the data, we can now analyse it. The type of data analysis you carry out largely depends on what your goal is. But there are many techniques available. Univariate or bivariate analysis, time-series analysis, and regression analysis etc.

Types of Data Analysis

Descriptive Analysis

Descriptive analysis identifies what has already happened. It is a common first step that companies carry out before proceeding with deeper explorations. The descriptive analysis provides a response to the question "what happened" by presenting historical data in the form of dashboards. Any learning platform might use descriptive analytics to analyse course completion rates for their customers. Or they might identify how many users access their products during a particular period. Perhaps they'll use it to measure sales figures over the last five years. While the company might not draw firm conclusions from any of these insights, summarizing and describing the data will help them to determine how to proceed.

Diagnostic Analysis

Diagnostic analytics focuses on understanding why something has happened. It is literally the diagnosis of a problem. Diagnostic analysis digs deeper into the descriptive analytics data to discover the root causes of the outcomes. This form of analytics is used by businesses because it builds more connections between data and finds patterns of activity. Creating comprehensive information is an important part of diagnostic analysis.

Diagnostic analysis has a variety of business applications like a freight firm is looking into the reason behind delayed delivery in a certain area.

Predictive Analysis

Predictive Analysis is a step up from descriptive and diagnostic investigations. This form of analytics makes predictions about future events based on prior data. The predictive analysis employs the information we've gathered to generate reasonable predictions about what will happen next.

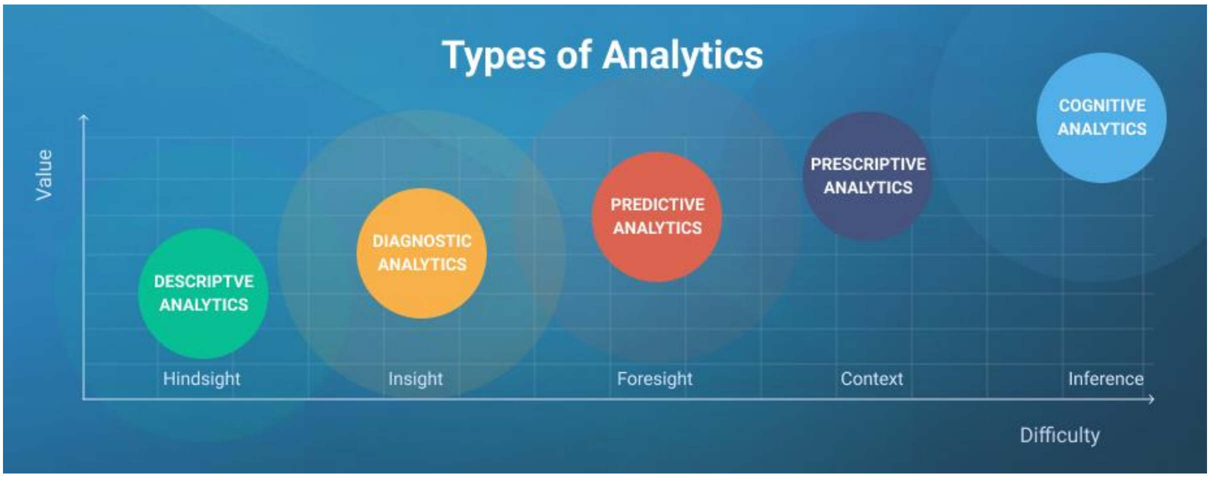
Prescriptive Analysis

Prescriptive analysis allows you to make recommendations for the future. A great example of prescriptive analytics is the algorithms that guide Google's self-driving cars. Every second, these algorithms make countless decisions based on past and present data, ensuring a smooth, safe ride.

Cognitive Analysis

Cognitive analytics is a smart technology that combines a variety of analytical approaches to evaluate big data sets and organize unstructured data. Cognitive analytics system looks through the data in its knowledge base for answers to queries that make sense.

- Analytics having human-like intelligence is referred to as cognitive analytics.
- Artificial intelligence algorithms and machine learning are frequently used in cognitive analytics, allowing a cognitive application to develop over time.
- Simple analytics cannot show some patterns and correlations, but cognitive analytics can.
- Cognitive analytics might be used by a company to track consumer behaviour patterns and developing trends. This allows the company to forecast future outcomes and adjust its goals to improve its performance.



Exploratory Data Analysis

Exploratory data analysis (EDA) is used by data scientists to analyse and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modelling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate.

Importance of exploratory data analysis

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modelling, including machine learning.

Exploratory data analysis tools

Specific statistical functions and techniques you can perform with EDA tools include:

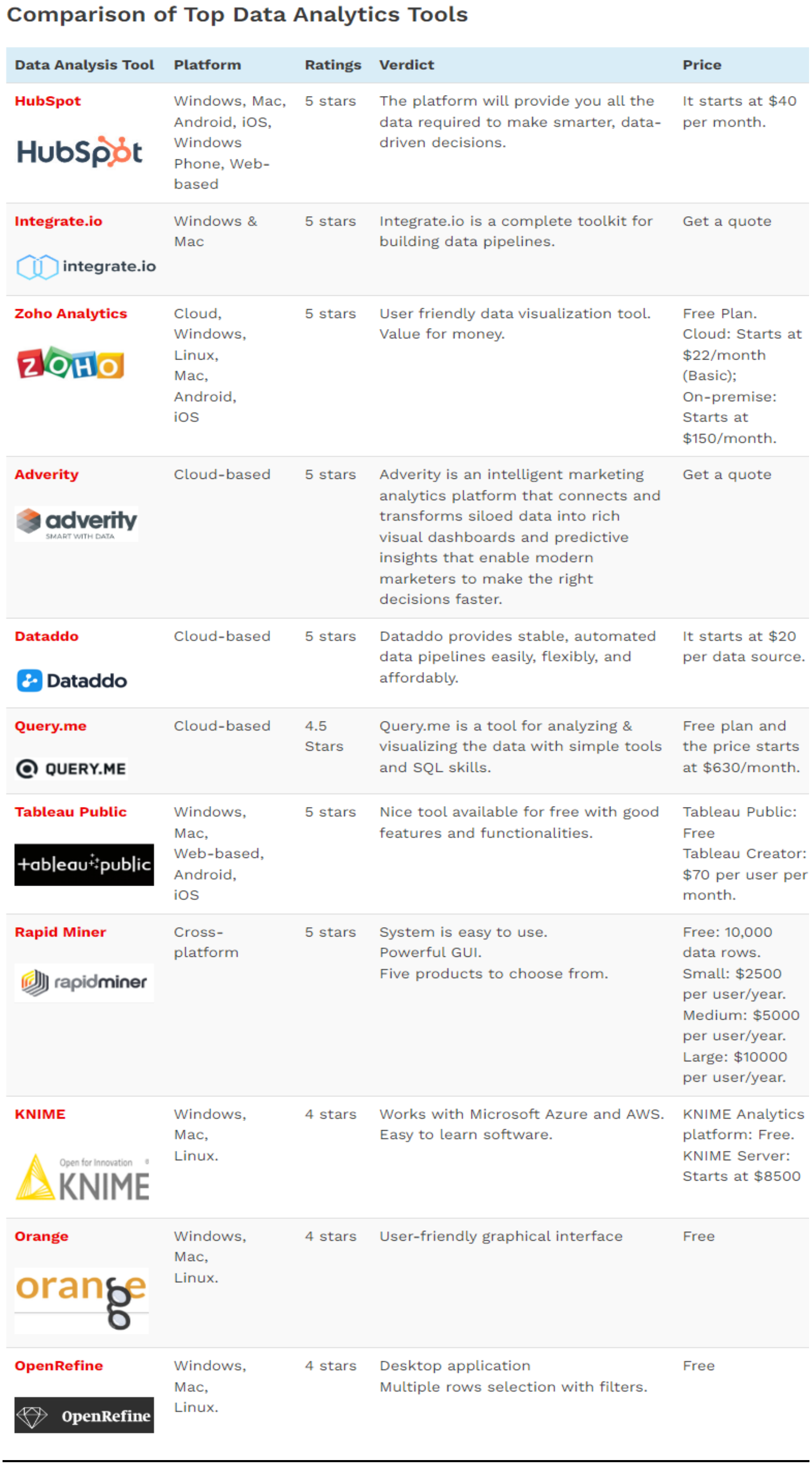
- Clustering and dimension reduction techniques, which help create graphical displays of high-dimensional data containing many variables.
- Univariate visualization of each field in the raw dataset, with summary statistics.
- Bivariate visualizations and summary statistics that allow you to assess the relationship between each variable in the dataset and the target variable you're looking at.
- Multivariate visualizations, for mapping and understanding interactions between different fields in the data.
- K-means Clustering is a clustering method in unsupervised learning where data points are assigned into K groups, i.e. the number of clusters, based on the distance from each group's centroid. The data points closest to a particular centroid will be clustered under the same category. K-means Clustering is commonly used in market segmentation, pattern recognition, and image compression.
- Predictive models, such as linear regression, use statistics and data to predict outcomes.
- Python: An interpreted, object-oriented programming language with dynamic semantics. Its high-level, built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for rapid application development, as well as for use as a scripting or glue language to connect existing components together. Python and EDA can be used together to identify missing values in a data set, which is important so you can decide how to handle missing values for machine learning.
- R: An open-source programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians in data science in developing statistical observations and data analysis.

Types of exploratory data analysis

There are four primary types of EDA:

- **Univariate non-graphical:** This is simplest form of data analysis, where the data being analyzed consists of just one variable. Since it's a single variable, it doesn't deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.
- **Univariate graphical:** Non-graphical methods don't provide a full picture of the data. Graphical methods are therefore required. Common types of univariate graphics include: Stem-and-leaf plots, which show all data values and the shape of the distribution. Histograms, a bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values. Box plots, which graphically depict the five-number summary of minimum, first quartile, median, third quartile, and maximum.
- **Multivariate non-graphical:** Multivariate data arises from more than one variable. Multivariate non-graphical EDA techniques generally show the relationship between two or more variables of the data through cross-tabulation or statistics.
- **Multivariate graphical:** Multivariate data uses graphics to display relationships between two or more sets of data. The most used graphic is a grouped bar plot or bar chart with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable.

Data Analytics Tools



Comparison of Top Data Analytics Tools

Data Analysis Tool	Platform	Ratings	Verdict	Price
HubSpot	Windows, Mac, Android, iOS, Windows Phone, Web-based	5 stars	The platform will provide you all the data required to make smarter, data-driven decisions.	It starts at $40 per month.
Integrate.io	Windows & Mac	5 stars	Integrate.io is a complete toolkit for building data pipelines.	Get a quote
Zoho Analytics	Cloud, Windows, Linux, Mac, Android, iOS	5 stars	User friendly data visualization tool. Value for money.	Free Plan. Cloud: Starts at $22/month (Basic); On-premise: Starts at $150/month.
Adverity	Cloud-based	5 stars	Adverity is an intelligent marketing analytics platform that connects and transforms siloed data into rich visual dashboards and predictive insights that enable modern marketers to make the right decisions faster.	Get a quote
Dataddo	Cloud-based	5 stars	Dataddo provides stable, automated data pipelines easily, flexibly, and affordably.	It starts at $20 per data source.
Query.me	Cloud-based	4.5 Stars	Query.me is a tool for analyzing & visualizing the data with simple tools and SQL skills.	Free plan and the price starts at $630/month.
Tableau Public	Windows, Mac, Web-based, Android, iOS	5 stars	Nice tool available for free with good features and functionalities.	Tableau Public: Free Tableau Creator: $70 per user per month.
Rapid Miner	Cross-platform	5 stars	System is easy to use. Powerful GUI. Five products to choose from.	Free: 10,000 data rows. Small: $2500 per user/year. Medium: $5000 per user/year. Large: $10000 per user/year.
KNIME	Windows, Mac, Linux.	4 stars	Works with Microsoft Azure and AWS. Easy to learn software.	KNIME Analytics platform: Free. KNIME Server: Starts at $8500
Orange	Windows, Mac, Linux.	4 stars	User-friendly graphical interface	Free
OpenRefine	Windows, Mac, Linux.	4 stars	Desktop application Multiple rows selection with filters.	Free