Datascience assignment1

1. what is data science?

Data science combines multiple fields, including statistics, scientific methods, artificial intelligence (AI), and data analysis, to extract value from data. Those who practice data science are called data scientists, and they combine a range of skills to analyse data collected from the web, smartphones, customers, sensors, and other sources to derive actionable insights.

Data science encompasses preparing data for analysis, including cleansing, aggregating, and manipulating the data to perform advanced data analysis. Analytic applications and data scientists can then review the results to uncover patterns and enable business leaders to draw informed insight.

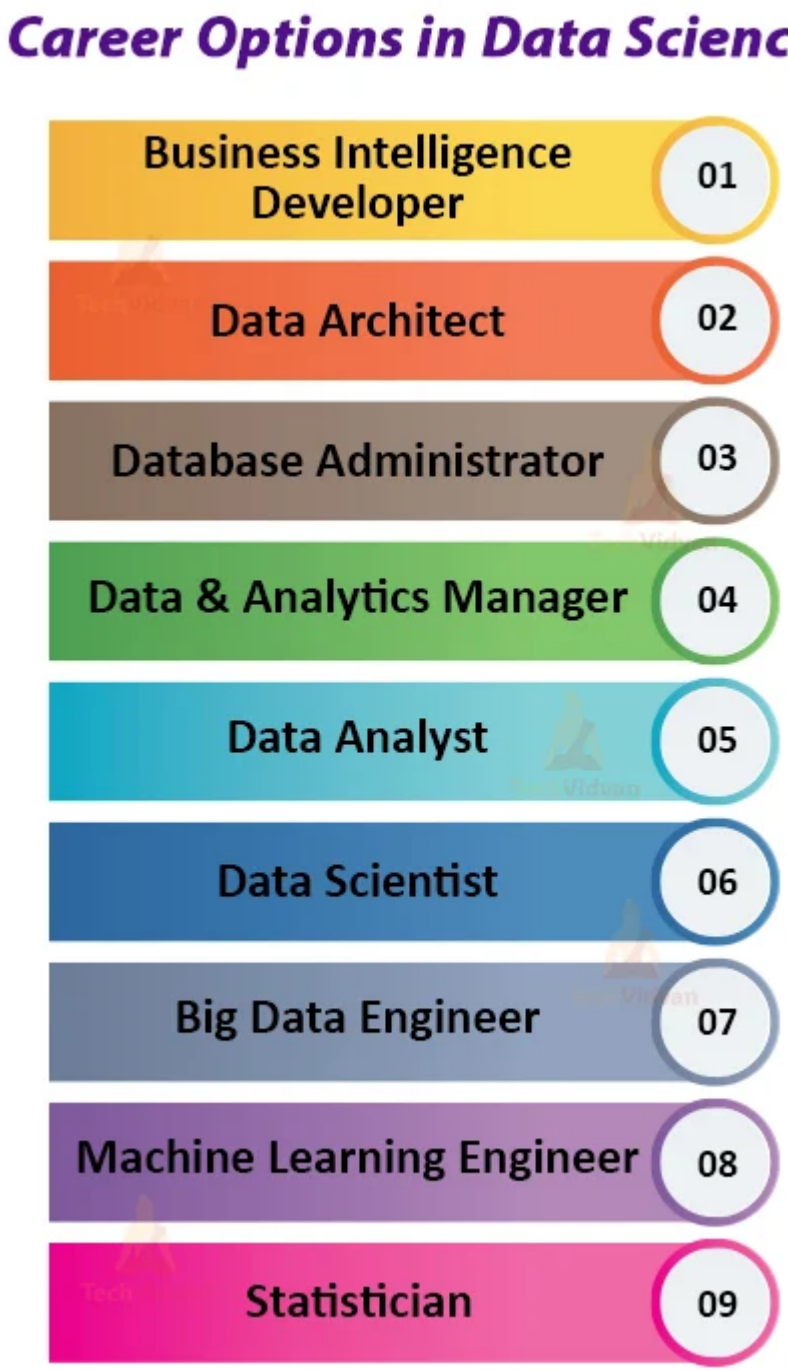
2. What is data?

In computing, data is information that has been translated into a form that is efficient for movement or processing. Relative to today's computers and transmission media, data is information converted into binary digital form. It is acceptable for data to be used as a singular subject or a plural subject. Raw data is a term used to describe data in its most basic digital format.

3. data science vs machine learning

datascience	machine learning
1. It deals with understanding and finding hidden patterns or useful insights from the data, which helps to take smarter business decisions 2. It is used for discovering insights from the data. 3. It is a broad term that includes various steps to create a model for a given problem and deploy the model 4. A data scientist needs to have skills to use big data tools like Hadoop, Hive and Pig, statistics, programming in Python, R, or Scala. 5. It can work with raw, structured, and unstructured data.	1. It is a subfield of data science that enables the machine to learn from the past data and experiences automatically. 2. It is used for making predictions and classifying the result for new data points. 3. It is used in the data modelling step of the data science as a complete process. 4. Machine Learning Engineer needs to have skills such as computer science fundamentals, programming skills in Python or R, statistics and probability concepts, etc 5. It mostly requires structured data to work on

4. what kind of opportunities with data science?



5.data science roles?

Data Science is a multifaceted discipline that combine several areas of technical expertise. Anyone acquiring overall expertise in this domain can be called a Data Scientist. But when it comes to hands-on business execution, specialisation is unavoidable due to the enormity of data-based activities in an organization. Depending on the specific data-tasks handled, several Data Science job roles have evolved across industries. They often appear to be overlapping in terms of skill-set and responsibility – but in reality, each of them has well-defined areas of action within the data framework.

Datascience assignment1

Let us look at the ten top job profiles that Data Scientists currently fill in at the workplace. Although their names are extremely familiar, do we really know the exact requirements for each of these roles?

Data Analyst

This is the fundamental job role in Data Science domain. As the name implies, the major focus in this job is on data analysis and reporting activities. A Data Analyst collects and organizes data, and then sifts out the necessary from the redundant. This cleaned up final data is the input on which they perform the required analysis and derive conclusions based on the findings. These findings could be revealing historical trends, or inferring future tendencies. Business decisions are made based on these findings. Visualizing the data patterns and presenting the results in the most effective communication format are necessary add-on requirements for the Data Analyst.

Business Analyst

A Business Analyst is essentially a Data Analyst whose analytical activities are targeted towards the internal systems and processes of a business organisation. The analysis is aimed at finding solutions to continually improve these business processes and design more effective ways-of-working. Business Analysts are expected to explore possibilities to streamline business operations, lower costs, and refine the decision-making process. Obviously, this will involve a clear knowledge of business processes and project management as well as expertise in software testing.

Business Intelligence Developer

A Business Intelligence (BI) Developer works on creating and maintaining BI interfaces and tools. Such BI solutions are used for data query, data visualisation, data dashboard designing, and data reporting. Being focused on data-querying, BI Developer are usually adept in designing complex applications and statistical models via SQL, Python and R. This is a job role that combines the skills of Data Engineers, Data Analysts, and Software Developers.

Machine Learning Engineer

The Machine Learning (M/L) Engineer designs, deploys and maintains software and algorithms based on Artificial Intelligence (AI) technology. Here, the objective is to automate predictive models such that the system can use data inputs to self-learn “on-the-job”, and keep refining itself to produce more accurate predictions. It is the responsibility of the M/L Engineer to organise and analyse collected data and identify the best input to train and validate the machine learning model.

Data Engineer

A Data Engineer essentially develops, implements and maintains the infrastructure that enables data cleaning, data preparation and manipulation. Data infrastructure is necessary to transform data into an analysable format – on which the Data Analysts can perform their duties. To create such an infrastructure, Data Engineers extract transform, and load the collected data – and continue to maintain and manipulate this data to always keep it updated for ready use.

Data Modeler

It is evident from the job title that a Data Modeler designs maintains and refines data models. These models are part of overall database designing activities, and are deployed for database implementation. The Data Modeler works in tandem with the Data Administrators and Data Architects – looking for opportunities to improve overall data availability and database performance.

Data Architect

A Data Architect takes a high-level view regarding the architecture and infrastructure aspects of data management. This is a role that constantly focuses on the organizations specific business requirements and accordingly designs the end-to-end data management architecture of the company. The Data Architect is not only concerned with databases but also with the overall flow of data in the system – right from the point data enters the company till it leaves the system. That

would span every data activity, like – data collection, data storage, data retrieval, data use, data modelling, and data security

Database Administrator

A Database Administrator coordinates with Data Modelers and Data Architects to implement and maintain database solutions. However, while the modeler and architects deal with the theoretical logic of the database, the administrator handles the practical logistic and technical issues regarding deployment and maintenance. The Database Administrator ensures availability of and access to a database, monitors data backup and restore routines, manages overall database performance, and perform policing tasks to ensure data security and integrity.

Marketing Scientist

This is a role that applies the principles of Data Science to marketing and sales data – with the aim is to solve associated business problems, like, field force sizing or marketing ROI. At a macro level, this job is similar to a general data scientist, but marketing data is the specialization here. A Marketing Scientist supports business decision-making by suitably interpreting data and identifying recognizable patterns in data to unearth latent trends in customer behaviour. Generally, this involves experiments devised through various data models to validate or reject any particular hypothesis.

6.Introduction to R language?

R Programming Language – Introduction

R is an open-source programming language that is widely used as a statistical software and data analysis tool. R generally comes with the Command-line interface. R is available across widely used platforms like Windows, Linux, and macOS. Also, the R programming language is the latest cutting-edge tool.

It was designed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team. R programming language is an implementation of the S programming language. It also combines with lexical scoping semantics inspired by Scheme. Moreover, the project conceives in 1992, with an initial version released in 1995 and a stable beta version in 2000

7.Installation to R and Rstudio new versions?

Installing R and RStudio:-

To get started with R, you need to acquire your own copy. This appendix will show you how to download R as well as RStudio, a software application that makes R easier to use. You'll go from downloading R to opening your first R session.

Both R and RStudio are free and easy to download

How to Download and Install R

R is maintained by an international team of developers who make the language available through the web page of The Comprehensive R Archive Network. The top of

Datascience assignment1

the web page provides three links for downloading R. Follow the link that describes your operating system: Windows, Mac, or Linux.

Windows

To install R on Windows, click the “Download R for Windows” link. Then click the “base” link. Next, click the first link at the top of the new page. This link should say something like “Download R 3.0.3 for Windows,” except the 3.0.3 will be replaced by the most current version of R. The link downloads an installer program, which installs the most up-to-date version of R for Windows. Run this program and step through the installation wizard that appears. The wizard will install R into your program files folders and place a shortcut in your Start menu. Note that you’ll need to have all of the appropriate administration privileges to install new software on your machine.

RStudio:-

RStudio is an application like Microsoft Word—except that instead of helping you write in English, RStudio helps you write in R. I use RStudio throughout the book because it makes using R much easier. Also, the RStudio interface looks the same for Windows, Mac OS, and Linux. That will help me match the book to your personal experience.

You can download RStudio for free. Just click the “Download RStudio” button and follow the simple instructions that follow. Once you’ve installed RStudio, you can open it like any other program on your computer—usually by clicking an icon on your desktop.

8.Basic R language points ?

Introduction to R and RStudio
RStudio interface and basics
Basic arithmetic and variable assignment
Comparison and logical operators
Data types
Vectors
Functions
Loops and conditionals
Probability and statistics
Data frames
Data visualisation

9.Data types and variables in R?

A variable in programming is used to store some data which will be used by the program. Consider it as a container which holds the data. Few rules to define variables in R
Variable names cannot contain spaces
Example - "Bill Amt" is invalid

A variable name can start with a dot but dot should not follow the number. If starting dot is not followed by a number, then it's valid
Example -.1BillAmt is invalid

A variable name should not start with a number
Example - 7Name is invalid

A variable name can contain letters, numbers, underscores and dots
Example - Bill_Name1. is valid Data Types in R

Data is available in various forms. In programming, data types are associated with a variable. A data type describes the type of data a variable can hold. Also, it is important to remember that everything in R is an object.

The basic data types in R are as follows,
Character
Numeric
Integer
Logical
Complex

10.Arithematic operations in R

R Operators
R supports majorly four kinds of binary operators between a set of operands. In this article, we will see various types of operators in R Programming language and their usage.

Types of the operator in R language
Arithmetic Operators
Logical Operators
Relational Operators
Assignment Operators

Datascience assignment1

Miscellaneous Operator

Arithmetic Operators

Arithmetic operations simulate various math operations, like addition, subtraction, multiplication, division, and modulo using the specified operator between operands, which may be either scalar values, complex numbers, or vectors. The operations are performed element-wise at the corresponding positions of the vectors.

Addition operator (+):

The values at the corresponding positions of both the operands are added. Consider the following R snippet to add two vectors:

```
Input : a <- c (1, 0.1)
        b <- c (2.33, 4)
        print (a+b)
Output : 3.33  4.10
```

Subtraction Operator (-):

The second operand values are subtracted from the first. Consider the following R snippet to subtract two variables:

```
Input : a <- 6
        b <- 8.4
        print (a-b)
Output : -2.4
```

Multiplication Operator (*):

The multiplication of corresponding elements of vectors and Integers are multiplied with the use of ‘*’ operator.

```
Input : B= matrix(c(4,6i),nrow=1,ncol=2)
        C= matrix(c(2,2i ),nrow=1, ncol=2)
        print (B*C)
Output : 8+0i  -12+0i
```

The elements at corresponding positions of matrices are multiplied.

Division Operator (/):

The first operand is divided by the second operand with the use of ‘/’ operator.

```
Input : a <- 1
        b <- 0
        print (a/b)
Output : -Inf
```

Power Operator (^):
The first operand is raised to the power of the second operand.

```
Input : list1 <- c(2, 3)
        list2 <- c(2,4)
        print(list1^list2)
Output : 4  81
```

Modulo Operator (%%):
The remainder of the first operand divided by the second operand is returned.

```
Input : list1<- c(2, 3)
        list2<-c(2,4)
        print(list1%%list2)
Output : 0  3
```

11.what are the different facets of data?

Data science is focused on making sense of complex datasets and in building predictive models from those data. As such, it encompasses a wide array of different activities, from the upstream processes of acquiring, cleaning and integrating data to downstream processes of analysis, modeling and prediction. There are many facets of data science, including:

Identifying the structure of data
Cleaning, filtering, reorganizing, augmenting, and aggregating data
Visualizing data
Data analysis, statistics, and modeling
Machine Learning
Assembling data processing pipelines to link these steps
Leveraging high-end computational resources for large-scale problems
Often, different tools address different parts of this process. Therefore, interoperability among tools, based on common data structures and interfaces, is an important element in enabling the construction of complex, multifaceted data analysis pipelines. It is in this sense that we can talk about an ecosystem for data science. For any particular application, you might only be interested in a subset of these operations.

Datascience assignment1

12.Types of data

The data is classified into majorly four categories:

Nominal data
Ordinal data
Discrete data
Continuous data
Further, we can classify these data as follows:

Types of data

Let us discuss the different types of data in Statistics herewith examples.

Qualitative or Categorical Data
Qualitative data, also known as the categorical data, describes the data that fits into the categories. Qualitative data are not numerical. The categorical information involves categorical variables that describe the features such as a person's gender, home town etc. Categorical measures are defined in terms of natural language specifications, but not in terms of numbers.

Sometimes categorical data can hold numerical values (quantitative value), but those values do not have a mathematical sense. Examples of the categorical data are birthdate, favourite sport, school postcode. Here, the birthdate and school postcode hold the quantitative value, but it does not give numerical meaning.

Nominal Data
Nominal data is one of the types of qualitative information which helps to label the variables without providing the numerical value. Nominal data is also called the nominal scale. It cannot be ordered and measured. But sometimes, the data can be qualitative and quantitative. Examples of nominal data are letters, symbols, words, gender etc.

The nominal data are examined using the grouping method. In this method, the data are grouped into categories, and then the frequency or the percentage of the data can be calculated. These data are visually represented using the pie charts.

Ordinal Data

Ordinal data/variable is a type of data that follows a natural order. The significant feature of the nominal data is that the difference between the data values is not determined. This variable is mostly found in surveys, finance, economics, questionnaires, and so on.

The ordinal data is commonly represented using a bar chart. These data are investigated and interpreted through many visualisation tools. The information may be expressed using tables in which each row in the table shows the distinct category.

Quantitative or Numerical Data

Quantitative data is also known as numerical data which represents the numerical value (i.e., how much, how often, how many). Numerical data gives information about the quantities of a specific thing. Some examples of numerical data are height, length, size, weight, and so on. The quantitative data can be classified into two different types based on the data sets. The two different classifications of numerical data are discrete data and continuous data.

Discrete Data

Discrete data can take only discrete values. Discrete information contains only a finite number of possible values. Those values cannot be subdivided meaningfully. Here, things can be counted in whole numbers.

Example: Number of students in the class

Continuous Data

Continuous data is data that can be calculated. It has an infinite number of probable values that can be selected within a given specific role.

13.Qualitative Vs quantitaive data?

1. Quantitative data

Quantitative data seems to be the easiest to explain. It answers key questions such as “how many, “how much” and “how often”.

Datascience assignment1

Quantitative data can be expressed as a number or can be quantified. Simply put, it can be measured by numerical variables.

Quantitative data are easily amenable to statistical manipulation and can be represented by a wide variety of statistical types of graphs and charts such as line, bar graph, scatter plot, and etc.

Examples of quantitative data:

Scores on tests and exams e.g. 85, 67, 90 and etc.
The weight of a person or a subject.
Your shoe size.
The temperature in a room.
There are 2 general types of quantitative data: discrete data and continuous data. We will explain them later in this article.

2. Qualitative data

Qualitative data can’t be expressed as a number and can’t be measured. Qualitative data consist of words, pictures, and symbols, not numbers.

Qualitative data is also called categorical data because the information can be sorted by category, not by number.

Qualitative data can answer questions such as “how this has happened” or and “why this has happened”.

Examples of qualitative data:

Colors e.g. the color of the sea
Your favorite holiday destination such as Hawaii, New Zealand and etc.
Names as John, Patricia,…..
Ethnicity such as American Indian, Asian, etc

14.structured Vs unstructured Vs semi structured data

1)Structured Data
Structured data is generally tabular data that is represented by columns and rows in a database.
Databases that hold tables in this form are called relational databases.
The mathematical term “relation” specify to a formed set of data held as a table.
In structured data, all row in a table has the same set of columns.
SQL (Structured Query Language) programming language used for structured data.
Structured Vs Unstructured Data

2) Semi-structured Data
Semi-structured data is information that doesn’t consist of Structured data (relational database) but still has some structure to it.
Semi-structured data consist of documents held in JavaScript Object Notation (JSON) format. It also includes key-value stores and graph databases.
Structured Vs Unstructured Data
3) Unstructured Data
Unstructured data is information that either does not organize in a pre-defined manner or not have a pre-defined data model.
Unstructured information is a set of text-heavy but may contain data such as numbers, dates, and facts as well.
Videos, audio, and binary data files might not have a specific structure. They’re assigned to as unstructured data.
Structured Vs Unstructured Data

Datascience assignment1

In short, Structured data is stored is predefined format and is highly specific; whereas unstructured data is a collection of many varied data types which are stored in their native formats; while semi structured data that does not follow the tabular data structure models associated with relational databases or other data table forms

15.data science methedology?

10 Steps of Data Science Methodology:-

1. Business Understanding

For any project or problem-solving, the first stage is always understanding the business. This involves defining the problem, project objectives, and requirements of the solutions. This step plays a critical role in defining how the project will develop. A thorough discussion with the clients, understanding how their business works, requirements from the product or service, and clarifying each aspect of the problem can take time and prove to be laborious, but it is a necessity.

2. Analytic Approach

After the problem has been clearly defined, the analytical approach which will be used to solve the problem can be defined. This means expressing the problem in the framework of statistical and machine learning techniques. There are different models that can be used and it depends on the type of outcome needed.

Statistical analysis can be used if it requires summarising, counting, finding trends in the data. To assess the relationships between various elements and the environment and how they affect each other, a descriptive model can be used.

And for predicting the possible outcomes or calculating the probabilities, a predictive model can be used which is a data mining technique. A training set that is a set of historical data that includes its outcomes, is used for predictive modeling.

Must Read: Reasons to Become Data Scientist

3. Data Requirements

The analytical approach chosen in the previous stage defines the kind of data needed to solve the problem. This step identifies the data contents, formats, and the sources for data collection. The data selected should be able to answer all the ‘what’, ‘who’, ‘when’, ‘where’, ‘why’ and ‘how’ questions about the problem.

4. Data Collection

In the fourth stage, the data scientist identifies all the data resources and collects data in all forms such as structured, unstructured, and semi-structured data that is relevant to the problem. Data is available on many websites and there are premade datasets that can also be used.

At times, if there is a requirement for important data that is not accessible freely, certain investments need to be made in order to obtain such datasets. If later there are any gaps identified within the collected data that is hindering the project development, the data scientist has to revise the requirements and collect more data.

The more the data acquired, the better the models will be built that can produce more effective outcomes.

5. Data Understanding

In this stage, the data scientist tries to understand the data collected. This involves applying descriptive analysis and visualization techniques to the data. This will help in a better understanding of the data content and the quality of the data and developing initial insights from the data. If there areany gaps identified in this step, the data scientist can go back to the previous step and gather more data.

6. Data Preparation

This stage comprises all the activities needed to construct the data to make it suitable to be used for the modeling stage. This includes data cleaning i.e. managing missing data, deleting duplicates, changing the data into a uniform format, etc., combining data from various sources, and transforming data into useful variables.

This is one of the most time-consuming steps. However, there are automated methods available today that can accelerate the process of data preparation.

Datascience assignment1

At the end of this stage, only the data needed to solve the problem is retained to make the model run smoothly with minimal errors.

7. Modeling

The dataset prepared in the previous stage is used for creating the modeling stage. Here the type of model to be used is defined by the approach decided upon in the analytical approach stage. Thus, the kind of dataset varies depending on whether it is a descriptive, predictive approach or a statistical analysis.

This is one of the most iterative processes in the methodology as the data scientist will use multiple algorithms to arrive at the best model for the chosen variables. It also involves combining various business insights that are continuously being discovered which leads to refining the prepared data and model.

Read: Data Science Career Path

8. Evaluation

The data scientist evaluates the quality of the model and ensures that it meets all the requirements of the business problem. This involves the model undergoing various diagnostic measures and statistical significance testing. It helps in interpreting the efficacy with which the model arrives at a solution.

9. Deployment

Once the model has been developed and approved by the business clients and other stakeholders involved, it is deployed into the market. It could be deployed to a set of users or into a test environment. Initially, it might be introduced in a limited way, until it is tested completely and been successful in all its aspects.

10. Feedback

The last stage in the methodology is feedback. This includes results collected from the deployment of the model, feedback on the model's performance from the users and clients, and observations from how the model works in the deployed environment.

Data scientists analyze the feedback received, which helps them refine the model. It is also a highly iterative stage as there is a continuous back and forth

between the modeling and feedback stages. This process continues till the model is providing satisfactory and acceptable results