Answer the questions with proper explanations and examples. You can read any research

paper, online article, book, or watch any video from YouTube to answer the questions.

Must have to submit .pdf file. Your pdf File rename will your\_name\_ds\_assignment\_01

(ex: suman\_ds\_assignment\_01).

**Short Questions:**

1. What is Data Science?

Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. Data science uses complex [machine learning algorithms](https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article) to build predictive models.

The data used for analysis can come from many different sources and presented in various formats.

Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.

2. What is Artificial Intelligence?

Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions.

3. What is Machine Learning?

Machine learning is a branch of [artificial intelligence (AI)](https://www.ibm.com/cloud/learn/what-is-artificial-intelligence) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

4. Future of Data Science and Real-life applications.

Data Science is a buzzword in the technology world right now and for good reason, it represents a major step forward in how computers can learn. The need for Data Scientists are high in demand and this surge is due to evolving technology and generation of huge amounts of data aka Big Data.

## **Future Scope of Data Science :**

Let’s have a look at a few factors that point out to data science’s future, demonstrating compelling reasons why it is crucial to today’s business needs.

* **Companies’ Inability to handle data**

Data is being regularly collected by businesses and companies for transactions and through website interactions. Many companies face a common challenge – to analyze and categorize the data that is collected and stored. A data scientist becomes the savior in a situation of mayhem like this. Companies can progress a lot with proper and efficient handling of data, which results in productivity.

* **Revised Data Privacy Regulations**

Countries of the European Union witnessed the passing of the General Data Protection Regulation (GDPR) in May 2018. A similar regulation for data protection will be passed by California in 2020. This will create co-dependency between companies and data scientists for the need of storing data adequately and responsibly. In today’s times, people are generally more cautious and alert about sharing data to businesses and giving up a certain amount of control to them, as there is rising awareness about data breaches and their malefic consequences. Companies can no longer afford to be careless and irresponsible about their data. The GDPR will ensure some amount of data privacy in the coming future.

* **Data Science is constantly evolving**

Career areas that do not carry any growth potential in them run the risk of stagnating. This indicates that the respective fields need to constantly evolve and undergo a change for opportunities to arise and flourish in the industry. Data science is a broad career path that is undergoing developments and thus promises abundant opportunities in the future. Data science job roles are likely to get more specific, which in turn will lead to specializations in the field. People inclined towards this stream can exploit their opportunities and pursue what suits them best through these specifications and specializations.

* **An astonishing incline in data growth**

Data is generated by everyone on a daily basis with and without our notice. The interaction we have with data daily will only keep increasing as time passes. In addition, the amount of data existing in the world will increase at lightning speed. As data production will be on the rise, the demand for data scientists will be crucial to help enterprises use and manage it well.

Real-life applications:

1. **Healthcare**

* Image Analysis in Medicine
* Genetics and Genomics
* Drug Development
* Virtual Assistants and Health bots

### ****Targeted Advertising****

### ****Website Recommendations****

### ****E-Commerce****

### ****5. Transport****

### ****Text and Advanced Image Recognization****

### ****Gaming****

### ****Security****

### ****Customer Insights****

### ****Augmented Reality****

5. Why is Python so popular? Future of Python.

It uses a simplified syntax with an emphasis on natural language, for a much easier learning curve for beginners. And, because Python is free to use and is supported by an extremely large ecosystem of libraries and packages, it’s often the first-choice language for new developers.

Therefore, let’s have a look at some of the most popular features of Python.

* Python is a multi-paradigm programming language including features like object-oriented, imperative, procedural, functional, reflective, etc.
* Have an intensive set of in-built libraries and tools that will enhance the functions of the language.
* Python has vast community support.
* Python is designed for better code readability as compared to other languages.
* It contains fewer lines of code as compared to other programming languages. Therefore, it is easy to understand.

Future of Python:

The future of Python is thus set in stone as the go-to language for developers engaging in deep machine learning projects, companies looking to scale up by using big data analytics or efforts aimed at achieving social applications of automation!

6. Why Python for data analytics?

Python is essential for data analytics so that,

### It’s Flexible

 If you want to try something creative that’s never done before; then Python is perfect for you. It’s ideal for developers who want to script applications and websites.

### It’s Easy to Learn

Thanks to Python’s focus on simplicity and readability, it boasts a gradual and relatively low learning curve. This ease of learning makes Python an ideal tool for beginning programmers. Python offers programmers the advantage of using fewer lines of code to accomplish tasks than one needs when using older languages. In other words, you spend more time playing with it and less time dealing with code.

### It’s Open Source

Python is open-source, which means it’s free and uses a community-based model for development. Python is designed to run on Windows and Linux environments. Also, it can easily be ported to multiple platforms. There are many open-source Python libraries such as Data manipulation, [Data Visualization](https://www.simplilearn.com/data-visualization-article), Statistics, Mathematics, [Machine Learning](https://www.simplilearn.com/tutorials/machine-learning-tutorial/what-is-machine-learning), and [Natural Language Processing](https://www.simplilearn.com/tutorials/artificial-intelligence-tutorial/what-is-natural-language-processing-nlp), to name just a few (though see below for more about this).

### It’s Well-Supported

Anything that can go wrong will go wrong, and if you’re using something that you didn’t need to pay for, getting help can be quite a challenge. Fortunately, Python has a large following and is heavily used in academic and industrial circles, which means that there are plenty of useful analytics libraries available. Python users needing help can always turn to Stack Overflow, mailing lists, and user-contributed code and documentation. And the more popular Python becomes, the more users will contribute information on their user experience, and that means more support material is available at no cost. This creates a self-perpetuating spiral of acceptance by a growing number of data analysts and data scientists. No wonder Python’s popularity is increasing!

7. Importance of Python in machine learning.

Here are the main reasons **why python is used for machine learning**–

## **The Rich Ecosystem of**[Python Libraries](https://www.naukri.com/learning/articles/powerful-python-libraries-for-data-science-and-machine-learning/)

Python is a high-level scripting language. It has a growing ecosystem of libraries, frameworks, and tools. These tools and libraries are equipped with pre-written codes, that help users to perform a myriad of functions while saving an adequate amount of time spent in code generation.

Some common Python libraries used for machine learning are –

**Matplotlib** – Fast processing and high-quality graphics generation, used in data visualization and graphical plotting

**NumPy** – Creation of a universal data structure helpful in analysis and exchange of algorithms; advanced mathematical operations on huge data sets

**Pandas** – Data manipulation, [data analysis](https://www.naukri.com/learning/what-is-data-analysis-st583-tg1135), data alignment, data set restructuring, and segmentation

**Scikit-Learn** – Data analysis, [data mining](https://www.naukri.com/learning/what-is-data-mining-st549-tg477), statistical modeling

**TensorFlow** – Build and train neural networks; Pattern detection; Numerical computing

[**PyTorch**](https://www.naukri.com/learning/pytorch-courses-certification-training-st553-tg1418)– [Artificial intelligence](https://www.naukri.com/learning/what-is-artificial-intelligence-st551-tg1405), machine learning, and [deep learning](https://www.naukri.com/learning/what-is-deep-learning-st551) applications

These Python libraries make the implementation of AI and [ML algorithms](https://www.naukri.com/learning/articles/top-machine-learning-algorithms-for-beginners/) very easy. This helps faster product development as it enables the developer to solve complex problems without rewriting codes.

## **Flexibility**

No one likes to work on complicated things and Python addresses the same challenge. It is a flexible and super-efficient programming language and can be used along with other languages. It offers the developers an option to choose between OOP or Scripting. In addition, it does not require the recompilation of the source code, which makes it easier to see the results. Thanks to its ease of operations, developers can enjoy working in a safe environment with a reduced probability of errors.

**Simplicity and Consistency**

Python offers readable and concise codes. Since machine learning and artificial intelligence involve complex algorithms, the simplicity of Python adds value and enables the creation of reliable systems. This helps the developer to remain focused on the machine learning problem without worrying about the technical details of the language.

A big reason that adds to the success of Python is its simplicity to learn. It is easier to understand and this helps to easily create machine learning models. Another feature of Python is that it is intuitive and is perfect for a collaborative deployment. It allows faster prototyping and product testing as it is a general-purpose language.

## **Platform Independence**

Python is a binary platform-independent programming language, which means it can run on a range of platforms and software architectures. The developer can write the code, compile, and run on multiple platforms.

Python is popular for its high versatility and can run on any platform, be it Windows, Macintosh, Linux, Solaris MacOS, Unix, and more. Integrating Python with other languages like Java, .NET, C/C++, Perl, PHP, R, etc. is easy.

## **Great Community Support**

Being an open-source language, Python has strong community support. Python is free, has a number of useful libraries and tools to offer, and its documentation can be accessed online. Programmers can discuss their problem statements in forums and chat with others to find solutions.

8. Importance of Python in deep learning & NLP.

* **It’s easy to understand and allows for quick data validation**

The role of [machine learning is to recognize patterns in data](https://asperbrothers.com/blog/ai-vs-machine-learning-vs-deep-learning/). A machine learning engineer is responsible for extracting, processing, refining, cleaning, arranging and making sense of data to develop intelligent algorithms. Python is easy to understand. While linear algebra or calculus concepts can be so complex, they take the maximum amount of effort. Python can be implemented quickly, which helps machine learning engineers to validate an idea promptly.

* **It has a great library ecosystem**

One of the main reasons why Python is the preferred language for machine learning is its access to many libraries. A library is a collection of functions and routines that a programming language can use. Having access to various libraries allows developers to perform complex tasks without the need to rewrite many code lines. Since machine learning heavily relies on mathematical optimization, probability and statistics, Python libraries help [data scientists](https://asperbrothers.com/data-science-services/) perform various studies easily. Here are some of the libraries you can use with Python:

* Pandas – for high-level data structures and analysis.
* Keras – for deep learning.
* Matplotlib – to create 2D plots, histograms, charts, etc.
* StatsModels – for statistical algorithms and data exploration and many others.
* **A low entry barrier**

There is a worldwide deficit of programmers. Python is easy to learn a language – the entry barrier is low. What does it mean? More data scientists can master it quickly, and as a result, they can get involved in [machine learning projects](https://asperbrothers.com/blog/image-recognition-in-python/). Believe it or not, but Python is very similar to the English language, making learning it easier. Thanks to its simple phrase structure, you can confidently work with complex systems.

* **It’s very flexible.**

Why choose Python for machine learning? Because it offers great flexibility. [Developers](https://asperbrothers.com/django-development/) can use Python along with another programming language to achieve their goals. They don’t have to recompile the source code. Any changes can be made instantly, which makes viewing results fast. Thanks to Python’s flexibility, the probability of bugs occurring is minimal.

* **It’s versatile**

An important part of [software development](https://asperbrothers.com/custom-software-development/) is testing. Python for machine learning can run literally on any platform, including Windows, macOS, Linux, Unix, and many, many more! Why it matters? It makes testing a piece of cake, as you can run tests on any platform you like. All your developers have to do is use PyInstaller, for example, to prepare their code to run on different platforms. Using Python for machine learning will save you a lot of time and money!

* **It’s easy to read**

It’s easy to read Python, so any [Python developers](https://asperbrothers.com/blog/python-development-company/) can easily implement it, copy or share it whenever a change in the code is required.

Using Python eliminates confusion, mistakes, and conflicting paradigms, which increases the efficiency of algorithm exchange, sharing of ideas, and tools between AI and machine learning professionals.

Python also provide the best librarys for NLP, The simple syntax and transparent semantics of this language make it an excellent choice for projects that include Natural Language Processing tasks:

* 1. Natural Language Toolkit (NLTK)
  2. TextBlob
  3. CoreNLP
  4. Gensim
  5. spaCy
  6. polyglot
  7. scikit-learn
  8. pattern

9. Supervised Learning with examples.

Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.

Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to **find a mapping function to map the input variable(x) with the output variable(y)**.

In the real-world, supervised learning can be used for **Risk Assessment, Image classification, Fraud Detection, spam filtering**, etc.

Example:

In supervised learning, models are trained using labelled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output.

The working of Supervised learning can be easily understood by the below example and diagram:



Suppose we have a dataset of different types of shapes which includes square, rectangle, triangle, and Polygon. Now the first step is that we need to train the model for each shape.

* If the given shape has four sides, and all the sides are equal, then it will be labelled as a **Square**.
* If the given shape has three sides, then it will be labelled as a **triangle**.
* If the given shape has six equal sides then it will be labelled as **hexagon**.

Now, after training, we test our model using the test set, and the task of the model is to identify the shape.

The machine is already trained on all types of shapes, and when it finds a new shape, it classifies the shape on the bases of a number of sides, and predicts the output.

10. Unsupervised learning with examples.

unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead, models itself find the hidden patterns and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things. It can be defined as:

*Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.*

Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to **find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format**.

**Jobs:**

11. Role and responsibilities of a Database Engineer.

Role and responsibilities :

* Administer and maintain database security, integrity, and optimization
* Develop and run periodic and requested reports
* Develop new databases and applications
* Identify and resolve performance and capacity issues
* Establish and maintain backup and recovery processes
* Maintain associated non-referential data
* Manage user rights and permissions

12. Role and responsibilities of a Data Analyst.

A data analyst is responsible for organizing data related to sales numbers, market research, logistics, linguistics, or other behaviours. They utilize technical expertise to ensure data is accurate and high-quality. Data is then analyzed, designed, and presented in a way that assists individuals, businesses, and organizations make better decisions.

* Using automated tools to extract data from primary and secondary sources
* Removing corrupted data and fixing coding errors and related problems
* Developing and [maintaining databases](https://www.simplilearn.com/what-is-data-management-article), data systems – reorganizing data in a readable format
* Performing analysis to assess quality and meaning of data
* Filter Data by reviewing reports and performance indicators to identify and correct code problems
* Using statistical tools to identify, analyze, and interpret patterns and trends in complex data sets that could be helpful for the diagnosis and prediction
* Assigning numerical value to essential business functions so that business performance can be assessed and compared over periods of time.
* Analyzing local, national, and global trends that impact both the organization and the industry
* Preparing reports for the management stating trends, patterns, and predictions using relevant data
* Working with programmers, engineers, and management heads to identify process improvement opportunities, propose system modifications, and devise [data governance](https://www.simplilearn.com/what-is-data-governance-article) strategies.
* Preparing final analysis reports for the stakeholders to understand the data-analysis steps, enabling them to take important decisions based on various facts and trends.

13. Role and responsibilities of a Data Engineer.

Some of the most common responsibilities for a data engineer include:

* Develop, construct, test, and maintain architectures
* Align architecture with business requirements
* Data acquisition
* Develop data set processes
* Use programming language and tools
* Identify ways to improve data reliability, efficiency, and quality
* Conduct research for industry and business questions
* Use large data sets to address business issues
* Deploy sophisticated analytics programs, machine learning, and statistical methods
* Prepare data for predictive and prescriptive modeling
* Find hidden patterns using data
* Use data to discover tasks that can be automated
* Deliver updates to stakeholders based on analytics

14. Role and responsibilities of a Data Scientist.

Data scientists work closely with business stakeholders to understand their goals and determine how data can be used to achieve those goals. They design data modeling processes, create algorithms and predictive models to extract the data the business needs, and help analyze the data and share insights with peers. While each project is different, [the process for gathering and analyzing data](https://www.northeastern.edu/graduate/blog/data-analysis-project-lifecycle/) generally follows the below path:

1. Ask the right questions to begin the discovery process

2. Acquire data

3. Process and clean the data

4. Integrate and store data

5. Initial data investigation and exploratory data analysis

6. Choose one or more potential models and algorithms

7. Apply data science techniques, such as machine learning, statistical modeling, and artificial intelligence

8. Measure and improve results

9. Present final result to stakeholders

10. Make adjustments based on feedback

11. Repeat the process to solve a new problem

15. Role and responsibilities of a Machine Learning Engineer.

An ML engineer's primary goals are the creation of machine learning models and retraining systems when needed. Responsibilities vary, depending on the organization, but some common responsibilities for this role include:

* Designing ML systems.
* Researching and implementing [ML algorithms](https://www.techtarget.com/whatis/definition/machine-learning-algorithm) and tools.
* Selecting appropriate data sets.
* Picking appropriate data representation methods.
* Identifying differences in data distribution that affects model performance.
* Verifying [data quality](https://www.techtarget.com/searchdatamanagement/definition/data-quality).
* Transforming and converting data science prototypes.
* Performing [statistical analysis](https://www.techtarget.com/whatis/definition/statistical-analysis).
* Running machine learning tests.
* Using results to improve models.
* Training and retraining systems when needed.
* Extending machine learning libraries.
* Developing machine learning apps according to client requirements.