1. What is difference between Exploratory Data Analysis and Predictive Data Analysis.

* **Data Exploration** is essentially the first step in any data analysis. It involves summarising the main characteristic of any database or dataset. It’s mostly done in statistical software varying in advancement levels, depending upon the complexity of the dataset. It can be conducted using visual analytics tools as well; this is more preferred, since it enables users to have a quick and simplified look at the most important featured and variables of the dataset. It helps the user quickly determine if the variables on screen are good enough for further analysis.
* **Predictive Analysis** falls under advanced analytics, and is used to make predictions about unknown events that might unfold in the future. In the simplest manner possible, Predictive analysis uses a host of different software, pairing them with many techniques ranging from artificial intelligence (AI), and machine learning, to analyze already existing data and make predictions regarding its course in the future.
* **Data Exploration** uncovers the complex, intricate, and oftentimes invisible relationships **between** measurable variables. **Predictive analysis** on the other hand offers outcomes and possibilities in the future of the variables, from the variables.

2. How would you define the role of a Data Scientist in Product Development Team .

* **Understand your customers:** A Data Scientist helps to gather, manage, and make sense of the data obtained.
* **Drive product vision:** A Data Scientist helps to gather input from customers, company's senior leadership, and competitors, and prioritize them accordingly to build product road map. The competitor part would be worthwhile to gather some competitive intelligence to see how other companies are solving problems, and what are the latest tools available, and so on.
* **Be an interface between data scientists and other teams:** An interface between the data scientists and the rest of the company is needed, so that not every small query about the tech needs an engineer's time to get resolved, that would be a big productivity boost.
* **Defining the business case:** In fields such as Search and Recommendations, good data scientists can make convincing data-driven arguments to engineering about building new data products, articulate problems with existing products, create a backlog, prioritize and get buy-in for their ideas.
* **UX:** Data science products have a front-end component. Depending on organisation structure, data scientist have to collaborate with ‘front-end’ product development team to integrate their products.
* **Experimentation:** Most data science products will be tested against a baseline to demonstrate incremental benefit. You will drive online experimentation, interpret results & make recommendations to business & engineering teams for future iterations of the product.
* **Monitoring**: Post roll-out of a data science product, you should put monitoring programs in place to keep track of data pipelines, data quality & financial, behavioral metrics.

3. Outline the various phases of a typical data science methodology.

* Business understanding: Every project, regardless of its size, starts with business understanding, which lays the foundation for successful resolution of the business problem. The business sponsors needing the analytic solution play the critical role in this stage by defining the problem, project objectives and solution requirements from a business perspective.
* Analytic approach: After clearly stating a business problem, the data scientist can define the analytic approach to solving it. Doing so involves expressing the problem in the context of statistical and machine learning techniques so that the data scientist can identify techniques suitable for achieving the desired outcome.
* Data requirements: Choice of analytic approach determines the data requirements, for the analytic methods to be used require particular data content, formats and representations, guided by domain knowledge.
* Data collection: The data scientist identifies and gathers data resources (structured, unstructured and semi-structured) that are relevant to the problem domain. On encountering gaps in data collection, the data scientist might need to revise the data requirements and collect more data.
* Data understanding: Descriptive statistics and visualization techniques can help a data scientist understand data content, assess data quality and discover initial insights into the data. A revisiting of the previous step, data collection, might be necessary to close gaps in understanding.
* Data preparation: The data preparation stage comprises all activities used to construct the data set that will be used in the modeling stage. These include data cleaning, combining data from multiple sources and transforming data into more useful variables. Moreover, feature engineering and text analytics may be used to derive new structured variables, enriching the set of predictors and enhancing the model’s accuracy.
* Modeling: Starting with the first version of the prepared data set, data scientists use a training set (historical data in which the outcome of interest is known) to develop predictive or descriptive models using the analytic approach already described. The modeling process is highly iterative.
* Evaluation: The data scientist evaluates the model’s quality and checks whether it addresses the business problem fully and appropriately. Doing so requires the computing of various diagnostic measures (as well as other outputs, such as tables and graphs) using a testing set for a predictive model.
* Deployment: After a satisfactory model has been developed that has been approved by the business sponsors, it is deployed into the production environment or a comparable test environment. Such a deployment is often limited initially to allow evaluation of its performance. Deploying a model into an operational business process usually involves multiple groups, skills and technologies.
* Feedback: By collecting results from the implemented model, the organization gets feedback on the model’s performance and observes how it affects its deployment environment. Analyzing this feedback enables the data scientist to refine the model, increasing its accuracy and thus its usefulness. This often overlooked stage can yield substantial additional benefits if undertaken as part of the overall process.

4. Mention 4 tools that a data scientist can rely on to effectively deliver his/her work

* **R.**
* **Python.**
* **Microsoft Power Bi.**
* **Tableau.**