**DS04 Activity 1**

**Task 1: List and explain the roles of a Data Science team (1 page)**

**1. Data Scientist**

**Responsibilities:** Data scientists are the team's core, analyzing and interpreting complex data. They develop machine-learning models, statistical analyses, and data-driven solutions for business problems. They also handle data preprocessing, feature engineering, and model evaluation.

**Skills:** Strong programming skills (Python, R), machine learning expertise, statistics, data visualization, and domain knowledge.

**2. Data Engineer**

**Responsibilities:** Data engineers focus on data collection, storage, and preparation. They design and maintain data pipelines, ensuring data is accessible and clean for data scientists. They work with databases, data warehousing, ETL (Extract, Transform, Load) processes, and data integration.

**Skills:** Proficiency in database technologies (SQL, NoSQL), data warehousing, ETL tools, scripting, and knowledge of big data technologies (Hadoop, Spark).

**3. Machine Learning Engineer**

**Responsibilities:** Machine learning engineers specialize in deploying and maintaining machine learning models in production. They work closely with data scientists to implement models, scale them, and ensure their performance and reliability in real-world applications.

**Skills:** Strong programming skills, model deployment, cloud computing, DevOps, and experience with machine learning frameworks (TensorFlow, PyTorch).

**4. Business Analyst**

**Responsibilities:** Business analysts bridge the gap between technical and non-technical stakeholders. They help define project goals, translate business requirements into data tasks, and communicate insights in a way that drives decision-making. They often play a vital role in project management and strategy development.

**Skills:** Strong communication, domain knowledge, project management, and data interpretation skills.

**5. Data Analyst**

**Responsibilities:** Data analysts focus on descriptive analytics. They explore data, create visualizations, and provide insights based on historical data trends. While they may not build machine learning models, they play a crucial role in uncovering actionable insights.

**Skills:** Data visualization tools (e.g., Tableau, Power BI), statistical analysis, data querying, and Excel proficiency.

**Data Scientist Manager**

**Responsibilities:** Data science managers oversee the team's operations, set goals, allocate resources, and ensure projects align with the organization's objectives. They act as a bridge between the data science team and upper management, helping to strategize and prioritize projects.

**Skills:** Leadership, project management, communication, and a strong understanding of data science and business goals.

**7. Data Quality Analyst**

**Responsibilities:** Data quality analysts ensure data accuracy, consistency, and reliability. They identify data issues, implement data quality standards, and collaborate with data engineers to maintain clean and high-quality data.

**Skills:** Data auditing, data quality tools, attention to detail, and data governance.

A well-rounded data science team collaborates to extract valuable insights and drive data-driven decision-making within an organization. Each role plays a unique part in the data science process, ensuring that data is collected, processed, analyzed, and deployed effectively.

**Task 2: Review the Data Cleansing, Exploration and Machine learning activities.**

**1.Data Cleansing**

Data cleansing, also known as data cleaning or data preprocessing, is a critical step in the data science process. It involves preparing raw data for analysis and Modeling. Here's a review of this activity:

**Importance:** Data cleansing is essential because real-world data is often messy and inconsistent. It helps ensure the accuracy and quality of data, which is crucial for reliable results.

**Tasks:** This stage typically involves handling missing values, dealing with outliers, correcting data types, and resolving duplicates. It may also include imputing missing values, normalizing or scaling data, and handling categorical variables.

**Tools:** Common tools and libraries for data cleansing include pandas (Python), Excel, and various data cleaning libraries in programming languages.

**Challenges:** Challenges in data cleansing include making decisions about how to handle missing data and outliers, as well as managing large datasets efficiently.

**2.Data Exploration**

Data exploration, often called exploratory data analysis (EDA), is the process of gaining insights into the dataset. Here's a review of this activity:

**Importance:** Data exploration helps data scientists understand the characteristics of the data, identify patterns, and make informed decisions about the modeling approach.

**Tasks:** During EDA, analysts use summary statistics, data visualizations (histograms, scatter plots, etc.), and hypothesis testing to understand data distributions, correlations, and relationships between variables.

**Tools:** Tools and libraries like matplotlib, seaborn, and Jupyter notebooks are commonly used for data exploration.

**Challenges:** Challenges include dealing with large and complex datasets, selecting the most appropriate visualization methods, and choosing the right statistical tests.

**3. Machine Learning**

Machine learning is the heart of data science, where models are built to make predictions or decisions based on data. Here's a review of this activity:

**Importance:** Machine learning allows data scientists to create predictive models, classify data, and automate decision-making processes, adding significant value to organizations.

**Tasks:** In this phase, data scientists select appropriate algorithms, split data into training and testing sets, train models, tune hyperparameters, and evaluate model performance using metrics like accuracy, precision, recall, and F1 score.

**Tools:** Common tools and libraries for machine learning include scikit-learn, TensorFlow, PyTorch, and various data preprocessing libraries.

**Challenges:** Challenges include selecting the right algorithm for the problem, addressing overfitting or underfitting, and optimizing model performance.

Overall, data cleansing ensures that data is reliable, data exploration helps understand the data's characteristics, and machine learning leverages this information to build models for predictive and prescriptive analysis. These activities are interrelated and essential for the success of data science projects.

**Task 3: Prepare a report of 1-2 pages explaining the output of each activity in 2-3 lines for each activity and include it in the Project Report.**

**1. Data Cleansing**

Data cleansing was a critical initial step in the project. This process involved handling missing values, outliers, and ensuring data consistency. The result is a cleaned dataset that is ready for analysis. It reduced data errors and enhanced data quality, providing a solid foundation for subsequent stages.

**2. Data Exploration**

The data exploration phase allowed us to gain valuable insights into the dataset. It provided a comprehensive view of data characteristics, distributions, and relationships between variables. Through summary statistics and visualizations, we discovered patterns and trends that informed our decision-making for subsequent analysis.

**3. Machine Learning**

Machine learning was the core of the project, where we applied various algorithms to build predictive models. The output of this stage includes trained machine learning models. These models can make predictions and classify data based on the patterns and relationships identified during data exploration. Evaluation metrics, such as accuracy and precision, were used to assess model performance.

**4. Model Evaluation and Selection**

In this phase, we evaluated the machine learning models to select the best-performing one. The output is a selected model that provides the highest predictive accuracy and meets the project's objectives. This model will be used for making predictions on new, unseen data.

**Conclusion**

The data cleansing phase provided us with a clean and reliable dataset, the data exploration phase equipped us with valuable insights, and the machine learning phase produced predictive models. These outputs together form the foundation for data-driven decision-making, contributing to the success of our project.

This report outlines the key outcomes of each project activity, highlighting their significance in the context of your data analysis and prediction project.