**Introduction**

Company XYZ manages IT infrastructure across multiple locations globally. They are interested in predicting network performance based on various factors, such as server response time, network bandwidth, and data transfer rate. Hence, they generated synthetic dataset to simulate a scenario in which various factors influence network performance.

**Dataset features.**

1. **Numeric features**
   * **Server response time (server\_response\_time):** Represents the time taken by servers to respond to network requests. Values are normally distributed around a mean of 100 with a standard deviation of 20.
   * **Network bandwidth (network\_bandwidth):** Signifies the amount of data that can be transmitted over the network per unit of time. Values are uniformly distributed between 1 and 100.
   * **Data transfer rate (data\_transfer\_rate):** Reflects the speed at which data is transferred across the network. Values are normally distributed around a mean of 50 with a standard deviation of 10.
2. **Categorical features**

* **Operating system type (OS\_type):** Denotes the operating system used, with options including "Windows," "Linux," and "Mac."
* **Server location (server\_location):** Indicates the geographic location of the server, with options including "US," "Europe," and "Asia."
* **Connection type (connection\_type):** Represents the type of network connection, with options including "DSL," "Fiber," and "Cable."

1. **Target Variable**

* **Response variable (response\_variable):** This is the variable they aim to predict.

**Possible hypothetical research questions**

The following hypothetical research questions aim to guide both exploratory data analysis to understand the dataset better and data modeling to build predictive models for network performance based on the provided synthetic dataset features.

**Data analysis research questions**

1. **Exploratory Data Analysis (EDA)**

* How does the distribution of server response time vary across different operating systems?
* Is there a correlation between network bandwidth and data transfer rate, and does this relationship differ based on server location?

1. **Categorical feature analysis**

* What is the predominant operating system type across different server locations, and does it have an impact on the response variable (network performance)?
* Can we identify any significant patterns or trends in network performance based on the type of network connection (DSL, Fiber, Cable)?

**Data modeling research questions**

1. **Predictive modeling**

* Given the features (server response time, network bandwidth, data transfer rate, OS type, server location, connection type), how accurately can we predict network performance (response variable)?
* What is the relative importance of each feature in predicting network performance, and are there interactions between certain features that significantly impact the model?

1. **Optimization and generalization**

* Can we develop a model that generalizes well across different server locations, or does the impact of certain features vary significantly in different geographic regions?
* How robust is the predictive model when applied to different types of network connections, and are there specific factors that significantly contribute to performance variations?