# Bitcoin Data Science Project

This project is designed to analyze Bitcoin data, focusing on both numerical and text features. The goal is to gain insights into Bitcoin price movements and the factors influencing these movements, using various data science techniques.

**Project Outline**

**Part 1: Predict Using the Numerical Data**

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

**Objective**: To understand the underlying patterns, trends, and distributions in the Bitcoin dataset.

**Hint:**

a. Visualize the historical price movements using line charts to identify trends.
b. Compute the parametric measures
c. Compute the non-parametric measures
d. Compare b and c

**2. Handling Missing Values in Time Series**

**Objective**: To effectively deal with any missing data points to maintain the integrity of the time series.

**Hint**:

a. Identify the extent and nature of missing values in the dataset.
b. Explore the way to handle missing values in time series

**3. Identifying Outliers in Time Series**

**Objective**: To detect anomalies that may affect the quality of predictions.

**Hint**:

a. Visualize outliers using box plots and time series plots.
b. Explore the way to find outliers in time series

**4. Exploring Moving Averages in Time Series**

**Objective**: To smooth the time series data and identify trends.

**Hint**:

a. Calculate various types of moving averages (simple, exponential) and analyze their effects on the time series.
b. Compare short-term vs. long-term moving averages to highlight different trends.
c. Visualize the moving averages alongside the original data to provide clarity.

**5. Exploring Lag Variables in Time Series**

**Objective**: To assess the impact of previous time steps on the current value.

**Hint**:

a. Use autocorrelation functions (ACF) and partial autocorrelation functions (PACF) to identify significant lags.
b. Understand the importance of lag variables in forecasting.

**6. Metrics for Time Series Prediction**

**Objective**: To evaluate the performance of different forecasting models.

**Hint**:

a. Explore common metrics
b. Compare the pros and cons of each metric in the context of time series data.
c. Select the most appropriate metrics for model evaluation based on the characteristics of the dataset.

**7. Developing a Time Series Model with Facebook Prophet**

**Objective**: To create robust forecasting models using the Facebook Prophet tool.

**Tasks**:

1. **Model A**: Develop the model using only the numerical features available in the dataset.
2. **Model B**: Incorporate lag and moving average variables (and numerical features) into the dataset and develop a new model.

---

**Part 2: Incorporating Numerical and Text Features**

**1. Developing TF-IDF for Text Features**

**Objective**: To quantify textual data related to Bitcoin (such as news articles or tweets) for further analysis.

**Hint**:

a. Clean and preprocess the text data, including tokenization, stop-word removal, and stemming.
b. Implement the TF-IDF (Term Frequency-Inverse Document Frequency) technique to transform text into numerical features.
c. Analyze the TF-IDF results to identify the most significant terms related to Bitcoin price movements.

**2. Integrating TF-IDF into the Best Model from Part 1**

**Objective**: To enhance the predictive power of the time series model by integrating textual features.

**Tasks**:

1. **Model C:** Add the TF-IDF features to the most successful model developed in Part 1.
2. Re-evaluate the model's performance using the chosen metrics from Part 1.
3. Compare the results to see how the inclusion of textual data affects forecast accuracy and overall model performance.