Project Report

I. Objective :

a. To analyze data for identifying key trends in the relationship between trader performance and market sentiment.
b. To build an SARIMAX model that predicts the Sentiment value for the future dates.

II. Data Set Overview :

a. **Historical Data** :
   i. Columns : 16
   ii. Rows : 211224
   iii. Null values are not present.
   iv. The Mean of all the numerical columns is appeared to be much greater than Median i.e. 50th Percentile. Based on this dataset appears to be Right Skewed.
   v. The huge difference between Mean and Standard Deviation and that of 25th and 75th Percentiles with extreme values suggests presence of outliers.
   vi. The dataset contains 5 Categorical columns apart from “Timestamp”.

b. **Market Sentiment Data** :
   i. Columns : 4
   ii. Rows : 2644
   iii. The Mean of “Value” column is nearly equal to Median i.e. 50th Percentile. This suggests that data is roughly Normally Distributed.
   iv. The min and max are not extreme as compared to the Quartiles(25th, 50th and 75th Percentile).
   v. 2 columns are Categorical.

III. Exploratory Data Analysis :

a. Merged the two separate datasets on a common “Date” column.
b. After merging, dropped the null values.
c. Key Insights
   - Profitability is not depending on Fees paid and Trade Sizes ($USD).
   - Greed led to highest Loss whereas Extreme Fear led to highest Profit.
   - Certain coins like AVAX, @109, ENA, etc performed better than others.
   - Buying resulted in comparatively more loss than Selling.
   - In short, it's not about timing the market, it's about time in the market.

IV. Data Preprocessing and Transformation :

a. Since the objective is to build a time-forecasting model, I took only the “Date” and “Value” as data for the model training.
b. Setted the “Date” as index as the SARIMAX model expects index to be of Date data type.
c. Time-forecasting models like ARIMA, SARIMAX, etc. expects data to fulfil some conditions :
    i. **Stationarity** : Constant Mean and Standard Deviation.
    ii. **Seasonality** : Repeating or Cyclic pattern in dataset.
d. **Stationarity** Check :
    We use Fuller’s test to check if the data is stationary.
    Since our p-value is less than 0.05, Value is stationary.
    So we aren't required to do Differencing.
e. Deciding the (p, d and q) Order :
    i. **p** : **Auto-Regressive** order
    ii. **d** : **Differencing** order
    iii. **q** : **Moving Average** order
    iv. Since we didn’t required Differencing, d = 0
    v. p is decided using PACF plot. We can see that after 2nd Lag, the rapid decrease starts. hence p = 2.
    vi. Since there is no rapid drop observed in the ACF plot, I took q = 1.
f. **Seasonality** Check :
    i. The repeating pattern is observable after a 30 days period, as displayed in the “Seasonal” component of the plot.

V. Model Building :

a. The SARIMAX model is present in **statsmodels.tsa.statespace.sarimax**.
b. It is trained on the dataset.
c. The performance is measured using R2 Score and Root Mean Squared Error
    i. **R2 Score** : 0.71 (nearer to 1 is better).
    ii. **RMSE** : 8.71 (much less than Mean of the target 46.98)
d. The model is saved using the **pickle** library in the folder “/outputs/models”.