# Exploratory Data Analysis & Feature Engineering

Exploratory Data Analysis and Visualization of DJI Data from Dec 31 2009 to Nov 15 2017. Libraries used were numpy, pandas, seaborn. Matplotlib, sklearn.metrics, math, datetime. Data is present in Github repo.

Missing data was filled using forward fill (Since it is Timeseries data, so previous day’s data will be close). Of all the features, Closing Values, Volume, Oil Prices, Gold Prices and Euro exchange rates were chosen as representative to the dataset as these cover the major capital market influences. These 5 features were plotted against date.

Then Moving Averages of 3 days, 15 days, 30 days, 45 days and 60 days were plotted and analyzed for Daily Index Closing Values, Oil Prices and Euro exchange rates. These plots clearly show the correlation between historical price data and stock prices. Specifically, 15 days and 30 days moving averages are most useful. These moving averages also tell that the seasonality is not stationary for Closing Values, Oil Prices and Euro exchange rates. Thus, Unit Root Tests, Augmented Dickey-Fuller was conducted. Although it is clearly visible graphs, the data is non-stationary. As expected, the test gives statistically significant results which was visualized using a custom function “*visualize\_adfuller\_results”.* P-Value came out to be 0.000 for each of the 5-features.

The data was then transformed using First Order Differencing and Log Transform which made it significantly stationary, with a p-value of 0.925. This allowed further plotting of autocorrelation and partial autocorrelation. This also opens the scope to implementation of Autoregression models as ARIMA, SARIMA etc. Autocorrelation for First Order Differencing yields promising results with datapoints distributed within 0.00 +/- 0.05. Although Log Transform Auto correlation was disappointing. Autocorrelation with a lag of 100 days for log transformed Daily Closing Value was between 1.0 and 0.8

Based on this we can comment that moving averages, log transformations and commodities etc other market data are powerful predictor features

Github Link to Notebook: <https://github.com/mandalnilabja/soc2022/blob/main/Week7Assignment.ipynb>

Reference: <https://www.kaggle.com/code/andreshg/timeseries-analysis-a-complete-guide/notebook>