

# OBJECTIVE






To challenge the Random Walk Hypothesis by building a predictive model that leverages high frequency data from actively traded commodity futures.



# DATA PREPROCESSING

## .MAT TO .CSV

### WORKFLOW:

-  Use glob.glob to retrieve .mat files from 'Train'/'Test' directory
-  Load .mat files and extract dataframes
-  Rename columns for clarity
-  Append date from .mat filename to the dataframe as a column
-  Export each dataframe to a corresponding .csv file



# FEATURES CREATED



**TRADE  
MOMENTUM**

**PRICE  
VOLATILITY**



**SIZE  
PRESSURE**

**CUM VOLUME  
DIFF**

**PRICE  
DIFFERENCE**

**RELATIVE  
STRENGTH INDEX**



**VOLUME  
IMBALANCE**

**WEIGHTED  
AVERAGE PRICE**



**LIQUIDITY RATIO**

**ROLLING  
AVERAGE SPEED**



# DATA CLEANING



## Removing Redundant Fields

Removed columns: 'hour', 'minute', 'second', ... , 'Date', 'sizepressure1', ... , 'MidPrice', 'weighted\_avg\_price', 'time', 'day\_of\_week\_number', 'RS'



## Column Renaming & Data Consistency Check

Verified column names between train and test datasets to ensure consistency



## Identifying Data Types

Identified numerical and categorical columns in the dataset for later usage



## Handling Missing Values and Stationarity

Removed rows with missing and inf values

Shifted columns by 40 & taken difference to ensure stationarity for prediction of 'Y'

# DATA TRANSFORMATION



## Poly Transformation

Performs polynomial transformation on numerical features to capture non-linear relationships

- Computes square and cubic terms for each numerical feature
- Concatenation: Combines original and polynomial features to form transformed batches



## Standardization

Transform numerical features to have a mean of 0 and a standard deviation of 1

- Benefits: Scale Consistency, Convergence Speed, Interpretability, Handling Outliers



## Augmented Dickey-Fuller (ADF) Test for Stationarity

Determines if a time series is stationary or non-stationary

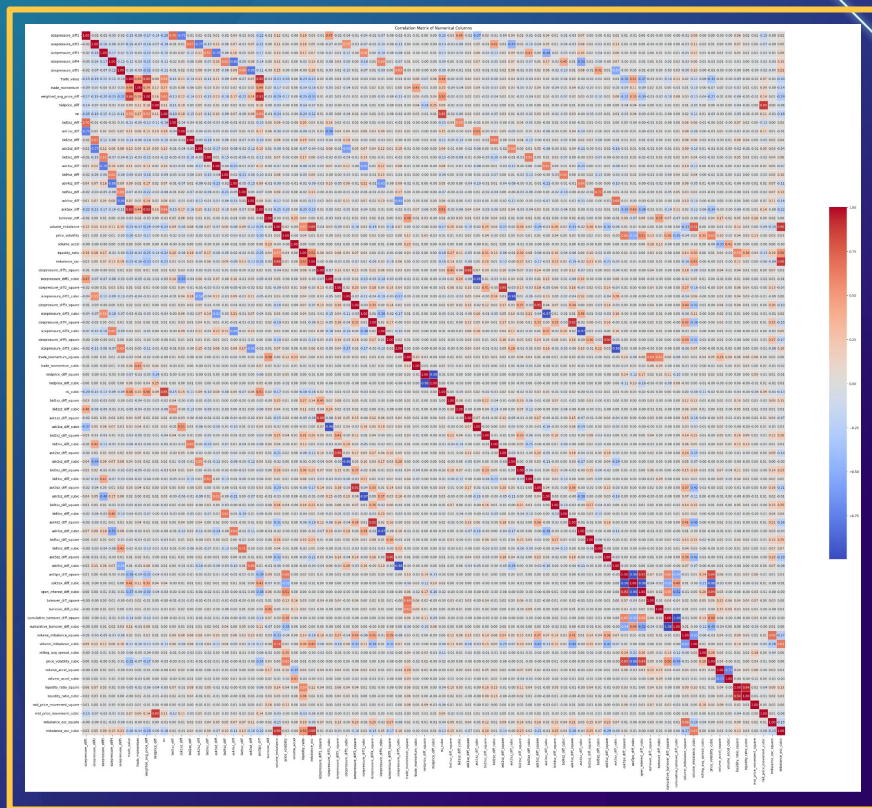
- Stationary Columns: Names of columns with  $p\text{-value} < 0.05$ .
- Non-stationary Columns: Names of columns with  $p\text{-value} \geq 0.05$

## 1 2 Numerical Data

- **Identify Relationships**
- **Multicollinearity Detection**
- **Pattern Recognition**
- **Feature Insights**

- **Pearson Correlation Coefficient**
- **Measures linear relationship between two datasets**

- Removed all the fields with order pair correlation of  $> 0.95$



# CORRELATION MATRIX



## Categorical Data

### Purpose:

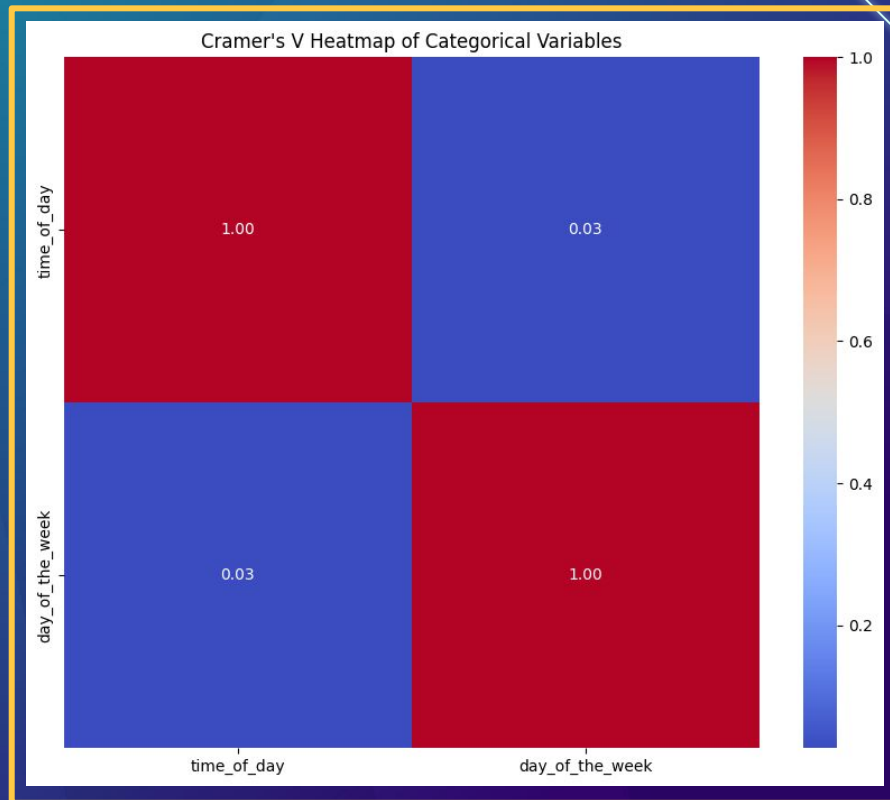
- To measure the association strength between pairs of categorical variables

### Method Used:

- Utilized the `cramers_v` function to calculate Cramér's V statistic
- Computed the confusion matrix for each pair of categorical variables

### Output:

- Removed all the fields with order pair correlation of  $> 0.95$



# VARIATION INFLATION FACTOR



## Categorical Data

### Purpose:

- To measure the association strength between pairs of categorical variables

### Method:

- Utilized the `cramers_v` function to calculate Cramér's V statistic
- Computed the confusion matrix for each pair of categorical variables
- Applied correction factors to the phi-square value



# ENCODING

**Conversion of Data Format:** We converted data from MATLAB's MAT format to CSV for more accessible data manipulation and analysis in Python, using libraries like pandas and scipy.io.

**Feature Engineering:** Developed new predictive features such as rolling averages, weighted average prices, and price volatility from high-frequency trading data. These features aim to capture market dynamics that simple price movements do not reveal.

**Handling of Time Series Data:** Extracted time-related features such as the time of day and the day of the week from timestamps. These features help capture cyclical trends in trading patterns, which are crucial for predicting price movements.

**Data Cleaning and Transformation:** Performed standardization on numerical features to normalize data, ensuring that the magnitude of the variables does not bias the model. Also applied transformations like polynomial expansion to capture non-linear relationships and interactions between features.

**Regularization Techniques:** Utilized Ridge and Lasso regression to enhance model generalization. This step helps in reducing overfitting by penalizing large coefficients in the regression model.

# ORDINARY LEAST SQUARES MODELS

## TRAIN DATA

In -Sample  $R^2$ : 0.019  
Durbin-Watson: 0.091

## TEST DATA

### OLS Regression Results

```
=====
Dep. Variable:          Y      R-squared:                0.019
Model:                  OLS    Adj. R-squared:            0.019
Method:                 Least Squares  F-statistic:         57.40
Date:                  Fri, 19 Apr 2024  Prob (F-statistic):    0.00
Time:                  00:23:23  Log-Likelihood:       -1.6550e+05
No. Observations:      117438   AIC:                 3.311e+05
Df Residuals:          117397   BIC:                 3.315e+05
Df Model:              40
Covariance Type:       nonrobust
```

```
=====
Omnibus:                25274.717  Durbin-Watson:        0.091
Prob(Omnibus):          0.000      Jarque-Bera (JB):     512401.861
Skew:                  -0.519      Prob(JB):             0.00
Kurtosis:              13.180      Cond. No.             8.17
=====
```

**Mean Squared Error: 0.9191311222791497**  
**R-squared: 0.008253292463085926**

# FEATURE CREATION



## Price Difference

Represents the change in mid-price of a stock or commodity between two points in time over a specified period

$$\text{MidPrice}_t - \text{MidPrice}_{t-40}$$

## Trade Momentum

Reflects the velocity of trading activity which helps forecast price direction based on recent trading intensity.

## Size Pressure

Quantifies the imbalance between buy and sell orders, capturing market supply and demand dynamics to predict future price movements.

## Cumulative volume difference

Measures the change in total trading volume over a specified period, used to analyze market activity trends and predict price movements.

# FEATURE CREATION

## Weighted Avg Price

assesses the average price of a security, adjusted for volume, offering a comprehensive view of price trends by accounting for the intensity of trading activity at different price levels.

## Volume Imbalance

Tracks the disparity between buy and sell volumes, indicating potential price trends based on market demand and supply dynamics.

## Relative Strength Index

A momentum oscillator that measures the speed and change of price movements by comparing the magnitude of recent gains to recent losses, aiming to indicate overbought or oversold conditions in the market.

## Rolling avg speed

Measures the average rate of change for a data point over a set period, offering insights into trend consistency and momentum.

# LIMITATIONS:

**Underfitting and Non-Linearity:** The linear regression model may not adequately capture the complex, non-linear relationships between independent variables (IVs). Non-linear or time-series models might provide more accurate predictions by better handling these dynamics.

**Lack of Auto-Regressive Features:** Our model currently does not include auto-regressive components, which are crucial for capturing time-dependent changes and trends in financial time series data.

**Feature Limitation Due to Initial Data Handling:** The initial data preparation and transformation steps may limit the variety and depth of features we can generate, potentially missing out on significant predictors.

**Data Consistency Issues:** There are missing data points and inconsistencies between the train and test datasets, which can lead to biased model training and evaluation.

**Transformation Techniques:** Current transformations are limited; exploring other methods like Yeo-Johnson, power, or Box-Cox transformations could better normalize the features and potentially enhance model performance.

**Feature Selection Methodology:** The feature selection process relies on simpler techniques. More robust methods such as decision trees or advanced regularization techniques could improve feature selection by identifying more predictive and less correlated variables.