USO ETF: ANALYSIS AND PREDICTIVE MODELING

PRESENTED BY KATERINA KOVALEVA

DATA Analytics 2023

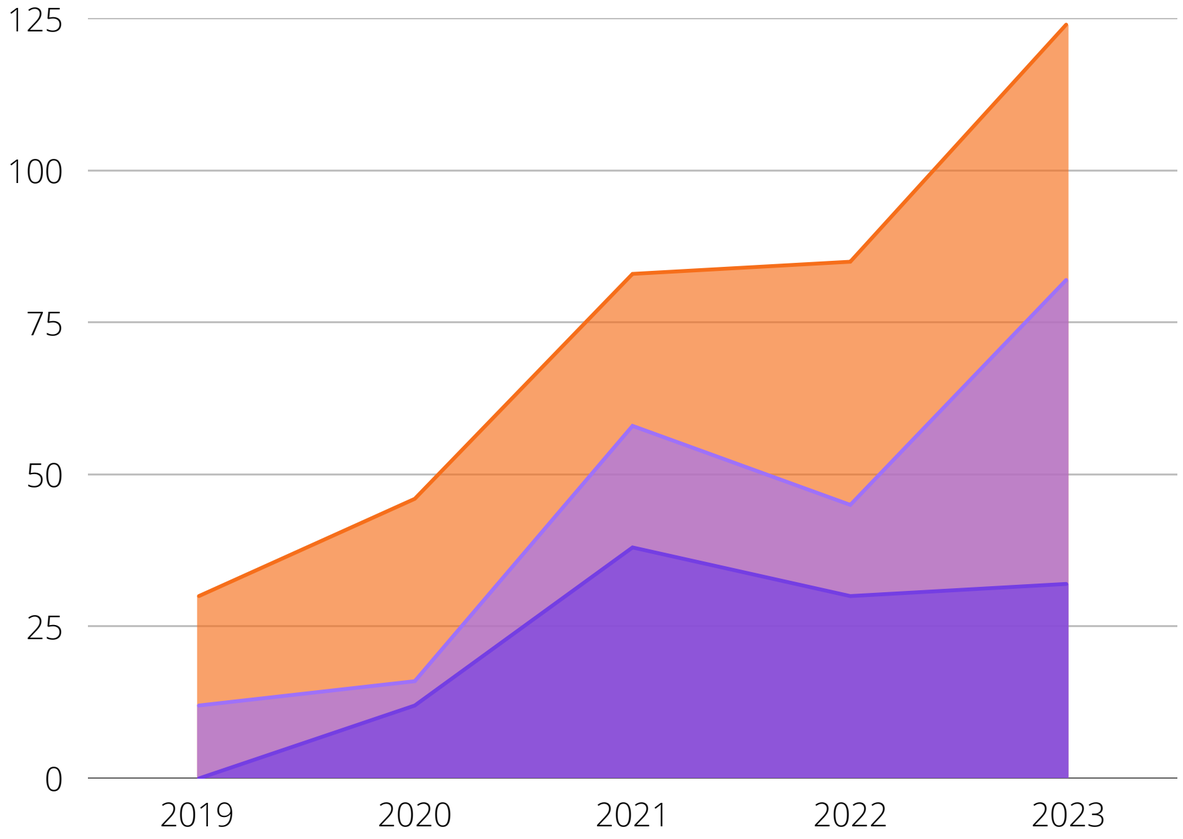
March, 18 | PARIS

Contents

The United States Oil Fund

- The United States Oil Fund (USO) is an exchange-traded fund (ETF) that invests in futures contracts for crude oil, heating oil, gasoline, and other petroleum-based fuels.
- Aims to track the daily price movements of West Texas Intermediate (WTI) light sweet crude oil.
- Designed to provide investors with exposure to crude oil without requiring them to purchase and store barrels of oil physically

What can impact the performance of the USO ETF?



Stocks:

- Energy Stocks
- Transportation Stocks
- Technology Stocks
- Consumer Goods Stocks
- Financial Stocks
- Automakers

What can impact the performance of USO ETF ?

Energy Stocks

Energy stocks are the most obvious stocks that can impact USO ETF performance, as they are directly related to the production and distribution of oil. When the stock prices of energy companies rise, it can indicate an increase in demand for oil, which can lead to higher oil prices

Transportation Stocks

Transportation stocks, particularly those in the airline industry, can impact oil prices. When the stock prices of airline companies rise, it can indicate an increase in demand for air travel, which in turn can lead to increased demand for jet fuel and therefore, higher oil prices

Technology Stocks

Technology Stocks: Technology stocks can impact oil prices indirectly. Advances in technology, such as the development of electric cars or renewable energy sources, can decrease the demand for oil, which can lead to lower oil prices

Consumer Goods Stocks

Consumer Goods Stocks: Consumer goods stocks, particularly those in the retail and automotive industries, can also impact oil prices. When the stock prices of these companies rise, it can indicate an increase in consumer spending, which can lead to increased demand for gasoline and therefore, higher oil prices

What can impact the performance of USO ETF ?

Financial Stocks

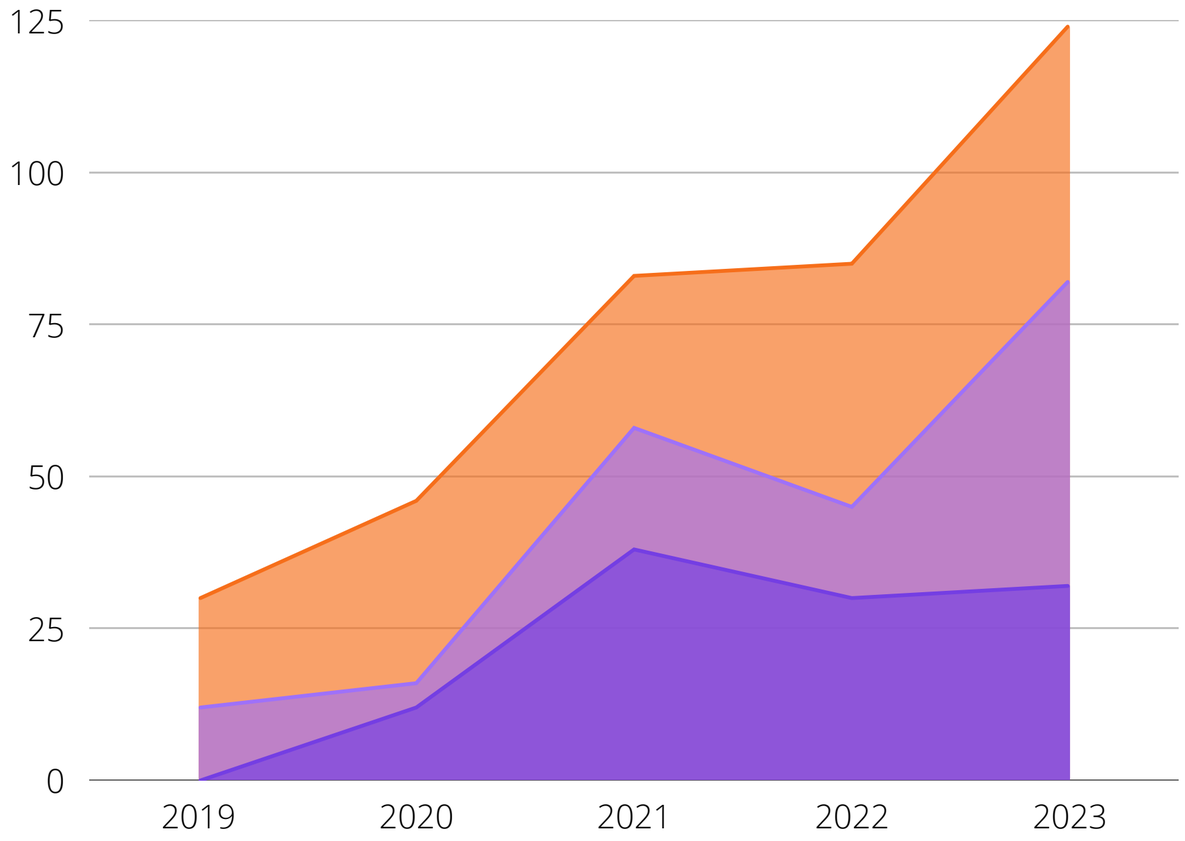
Financial stocks, particularly those in the banking and investment sectors, can also impact oil prices. When these stocks rise, it can indicate increased investor confidence in the economy, which can lead to increased demand for oil and higher oil prices

Automakers

Automakers are sensitive to changes in oil prices since they rely on petroleum products to power their vehicles. If the price of oil rises, the cost of producing and operating cars increases, which can lead to a decline in automaker stock prices

STUDIO SHODWE

What else can impact the performance of the USO ETF?



- Global oil supply and demand
- Geopolitical tensions
- OPEC decisions
- Economic indicators
- Currency fluctuations
- Speculation

I focused more deeply the impact of the following factors on the the USO ETF Performance:

- *Major Energy Stock Prices*
- *Market Indexes (DJ and S&P 500)*
- *WTI Crude Oil Prices*
- *US Dollar Index fluctuations*

Progress Plan

- Finding Data Sources
- Data Collection
- Data Cleaning
- Building ERD
- Exploratory Data Analysis
- Insight and Visualisation
- Building of ML Models
- Evaluation ML Models

DATA SOURCE

Historical Prices from:

- **Yahoo Finance** https://finance.yahoo.com/
- **Nasdaq** https://www.nasdaq.com/

STUDIO SHODWE

Data Collection

- Dataframe: 11/03/2013 - 20/09/2022
- Historical prices of Crude Oil, Standard and Poor's (S&P) 500 Index, Dow Jones Index, US Dollar Index and 5 Major Energy Stocks (Petroleo Brasileiro Petrobras SA, Exxon Mobil Corp, Total Energies SE, Chevron Corp, Shell PLC)
- In total 12 csv files with historical prices downloaded. These files have similar structure
- 2401 rows x 63 columns

STUDIO SHODWE

Data Collection

The datasets contain following columns:

- USO ETF : USO_Open, USO_High, USO_Low, USO_Close, USO_ Adj Close, USO_Volume;
- S&P 500 Index : 'SPX_open', 'SPX_high', 'SPX_low', 'SPX_close';
- Dow Jones Index : "DJIA_open','DJIA_high', 'DJIA_low', 'DJIA_close';
- EURO - USD Exchange Rate : 'USDEUR_Open', 'USDEUR_High', 'USDEUR_Low', 'USDEUR_Close', 'USDEUR_Adj_Close";
- Brent Crude Oil : 'BZ_Open', 'BZ_High', 'BZ_Low', 'BZ_Close', 'BZ_Volume';
- WTI Crude Oil : 'CL_Open', 'CL_High', 'CL_Low', 'CL_Close', 'CL_Volume';
- US Dollar Index : 'USDI_Price', 'USDI_Open', 'USDI_High','USDI_Low';
- Petroleo Brasileiro Petrobras SA, : PBR_Open, PBR_High, PBR_Low, PBR_Close, PBR_ Adj Close, PBR_Volume;
- Exxon Mobil Corp : XOM_Open, XOM_High, XOM_Low, XOM_Close, XOM_ Adj Close, XOM_Volume;
- Total Energies SE : TTE_Open, TTE_High, TTE_Low, TTE_Close, TTE_ Adj Close, TTE_Volume;
- Chevron Corp : CVX_Open, CVX_High, CVX_Low, CVX_Close, CVX_ Adj Close, CVX_Volume;
- Shell PLC : SHEL_Open, SHEL_High, SHEL_Low, SHEL_Close, SHEL_ Adj Close, SHEL_Volume;

STUDIO SHODWE

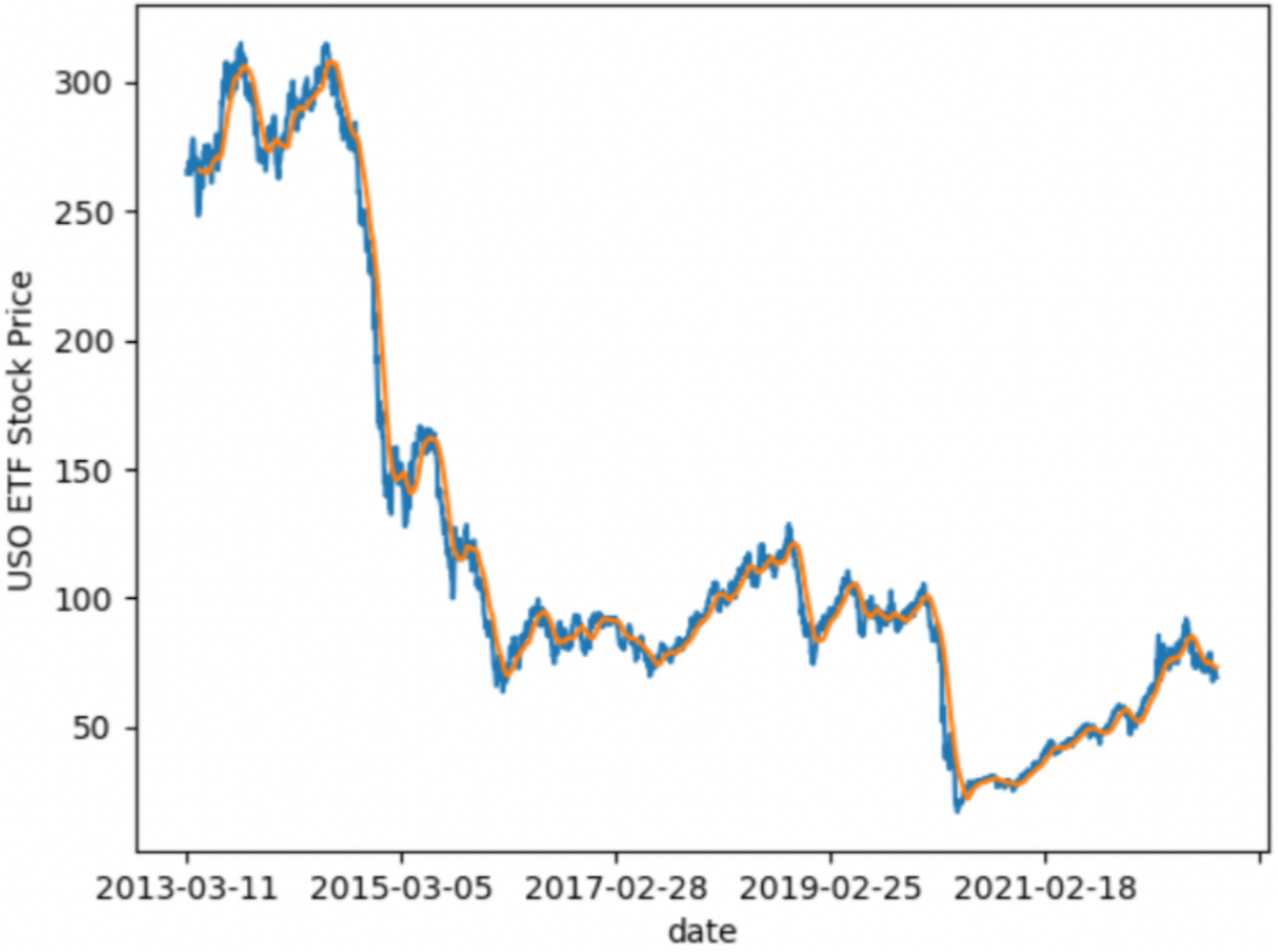
Data Cleaning

- I had 29 missing values in the column 'Volume' in files 'BZ_Crude_Oil' and 'CL_Crude_Oil' so I replaced it by the average of the column
- All the values in the column 'Volume' in files 'SPX' and 'DJIA were missing so I dropped the entire column for these files.
- I changed type of the column 'Date' into datetime

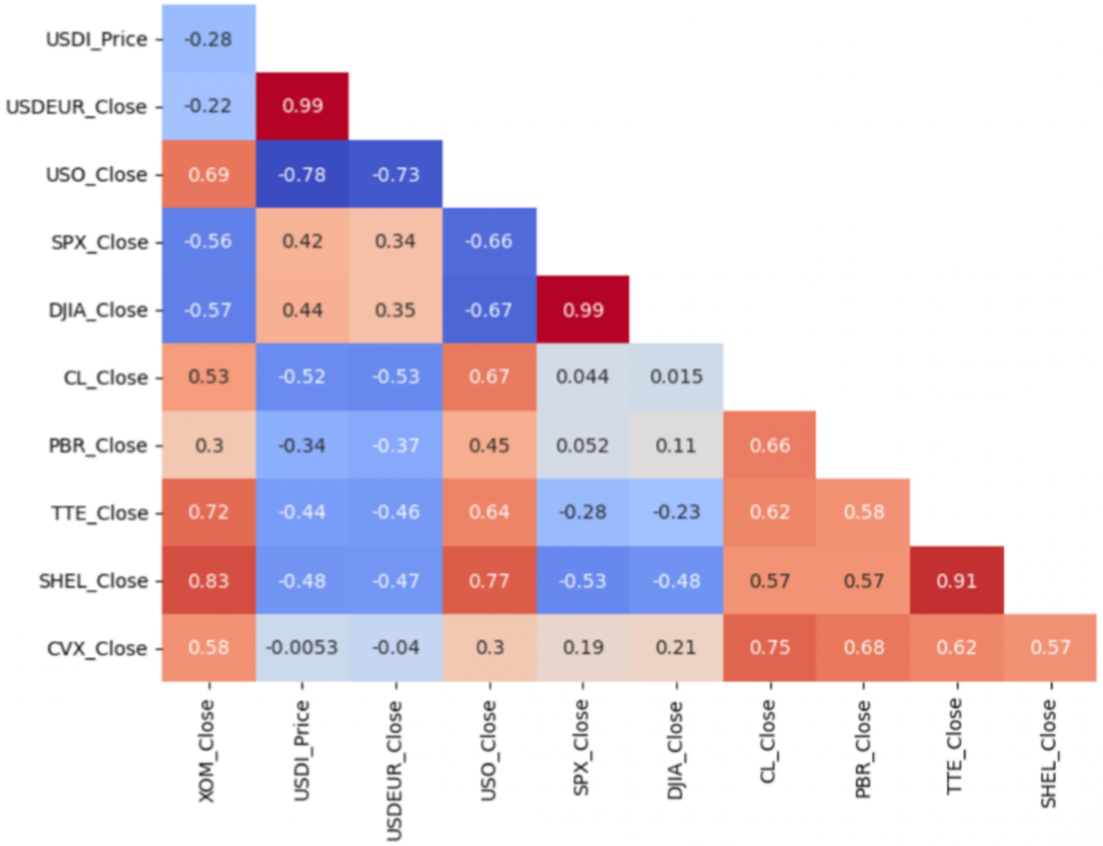
```
missing_values = usdeur.isna().sum()
print("Number of missing values in each column:\n", missing_values)

# Check for NaN values
nan_values = usdeur.isnull().sum()
print("Number of NaN values in each column:\n", nan_values)
```

```
data.drop('Volume', axis=1, inplace=True)
data
```

Historical prices of USO ETF with 30-day rolling mean

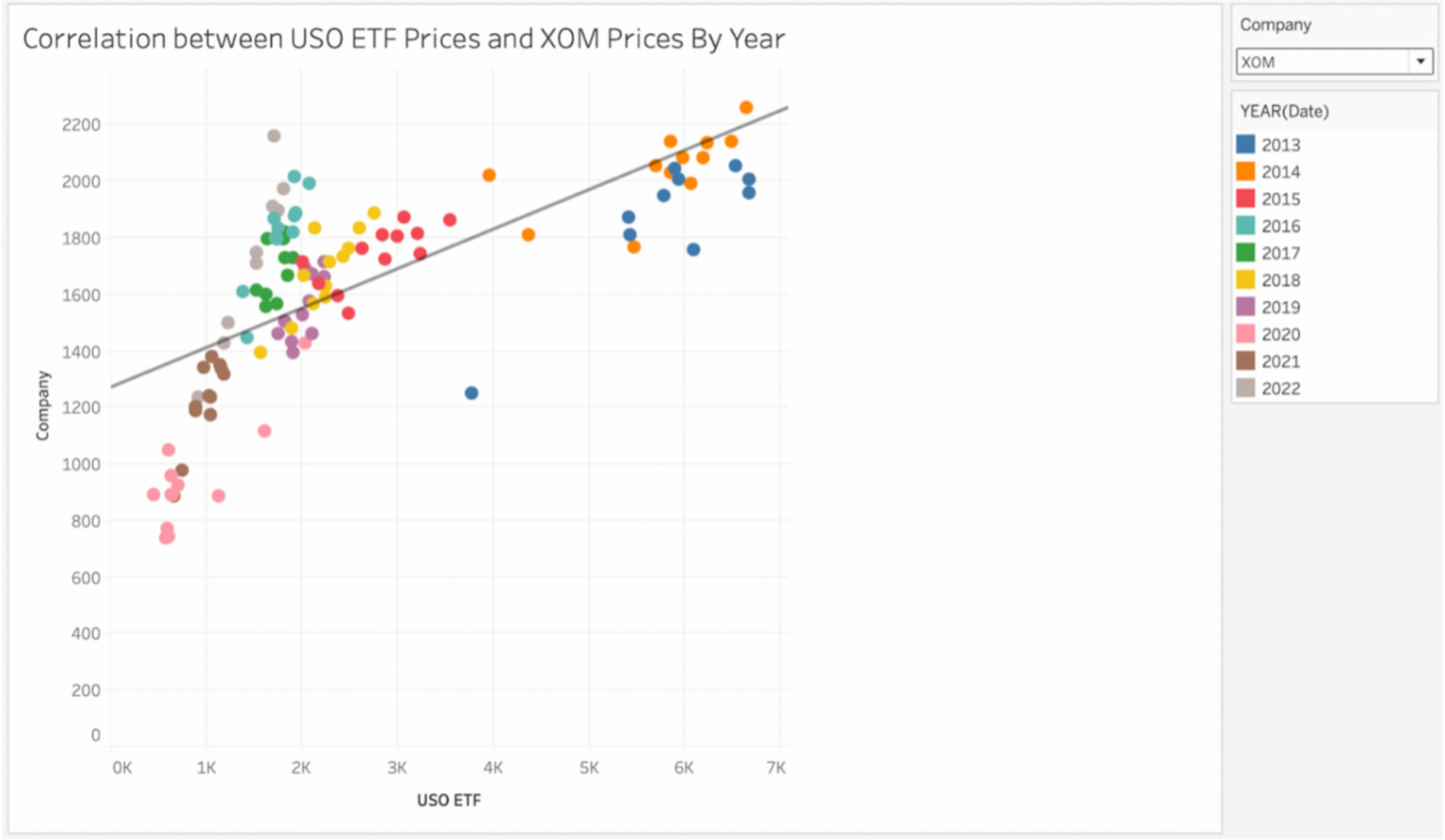


Correlation Matrix



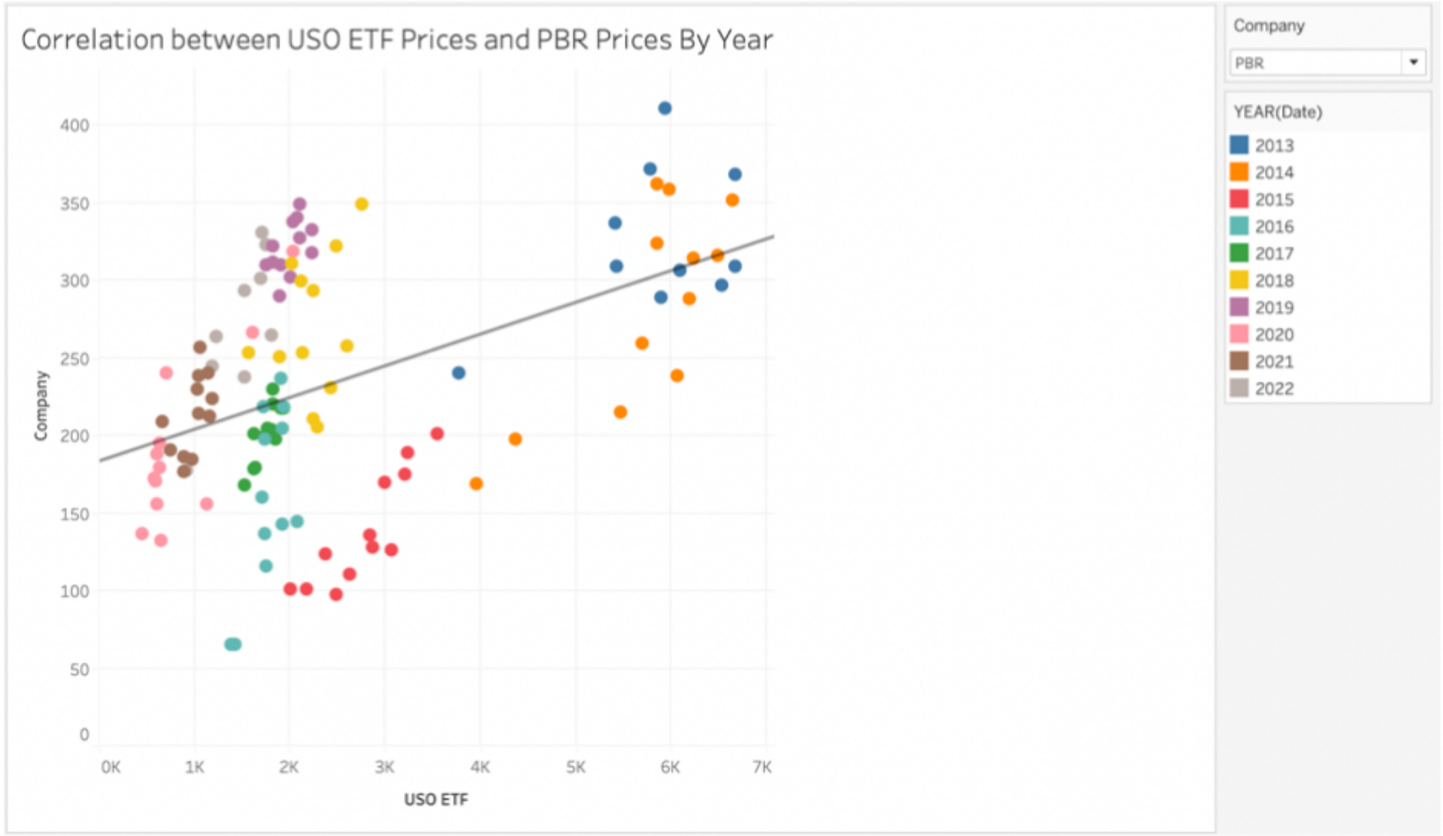
- Negative correlation between the price of USO ETF and US Dollar Index as well as between the price of USO ETF and USD to EUR ratio. It could be linked to the fact that USO ETF tracks the price of oil, so when oil prices increases, the price of the USO ETF goes up.
- However, an increase in oil prices may lead to a decrease in the value of the US dollar because oil is priced in dollars. The US is a major importer of oil, which means that when oil prices rise, the cost of imported oil increases. This can widen the US current account deficit, which is the difference between the value of goods and services that the US imports and exports. A larger current account deficit can lead to a weaker US dollar. This could explain the negative correlation between the USO ETF and the US dollar.

Correlation between USO ETF Prices and XOM prices By Year



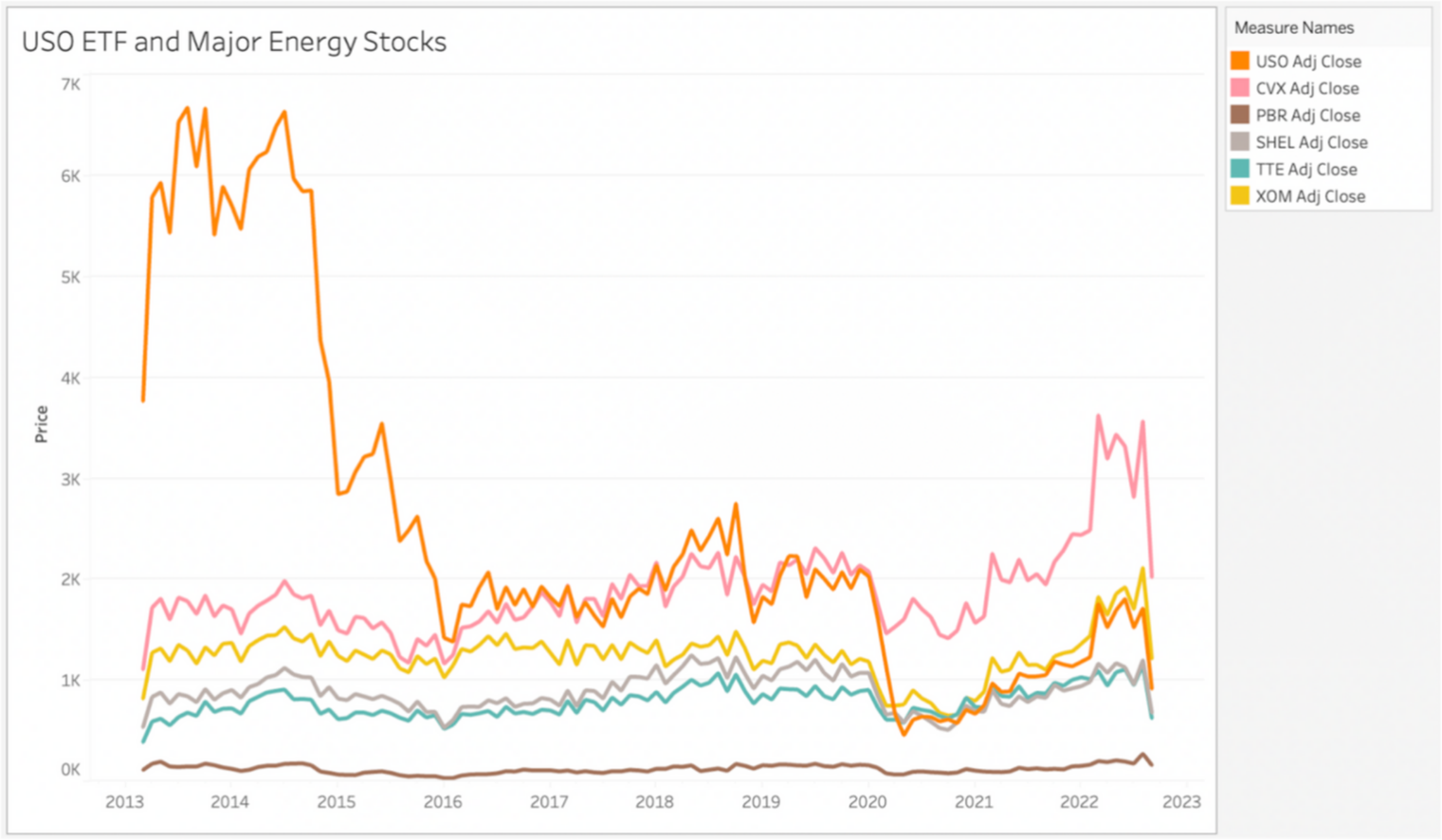
- There is a positive correlation between the close prices of USO ETF, XOM (Exxon Mobil Corp) and SHEL (SHELL PLC) which makes sense because these companies operate in the same industry. This means that the two stocks tend to move in the same direction, either up or down, in response to market conditions or other factors that affect the industry.

Correlation between USO ETF Prices and Major Stock Prices By Year



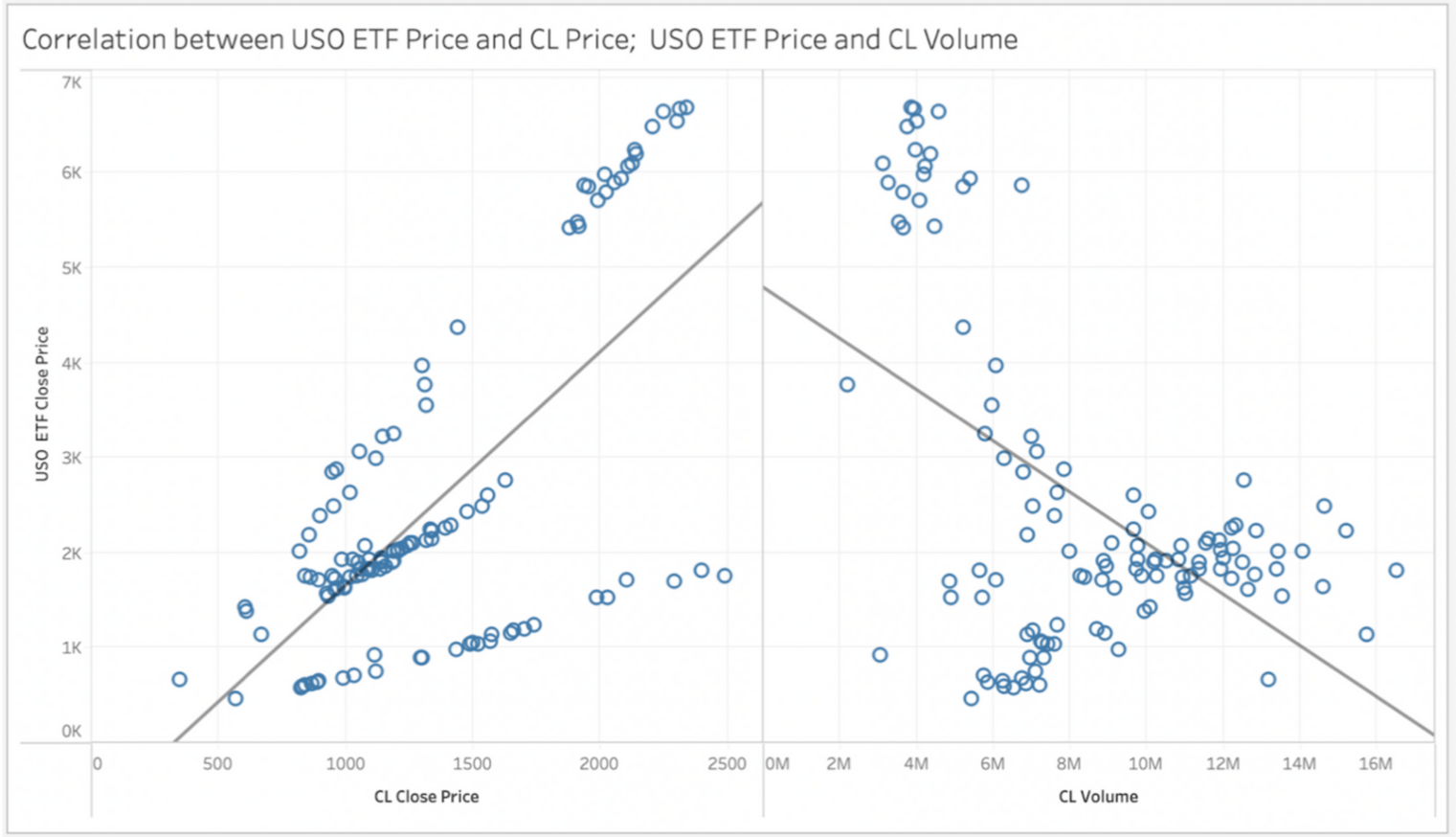
- However the close prices of USO ETF, PBR (Petrobras SA), TTE (Total Energies SE) and CVX (Chevron Corp) are much less correlated

USO ETF and Major Energy Stocks



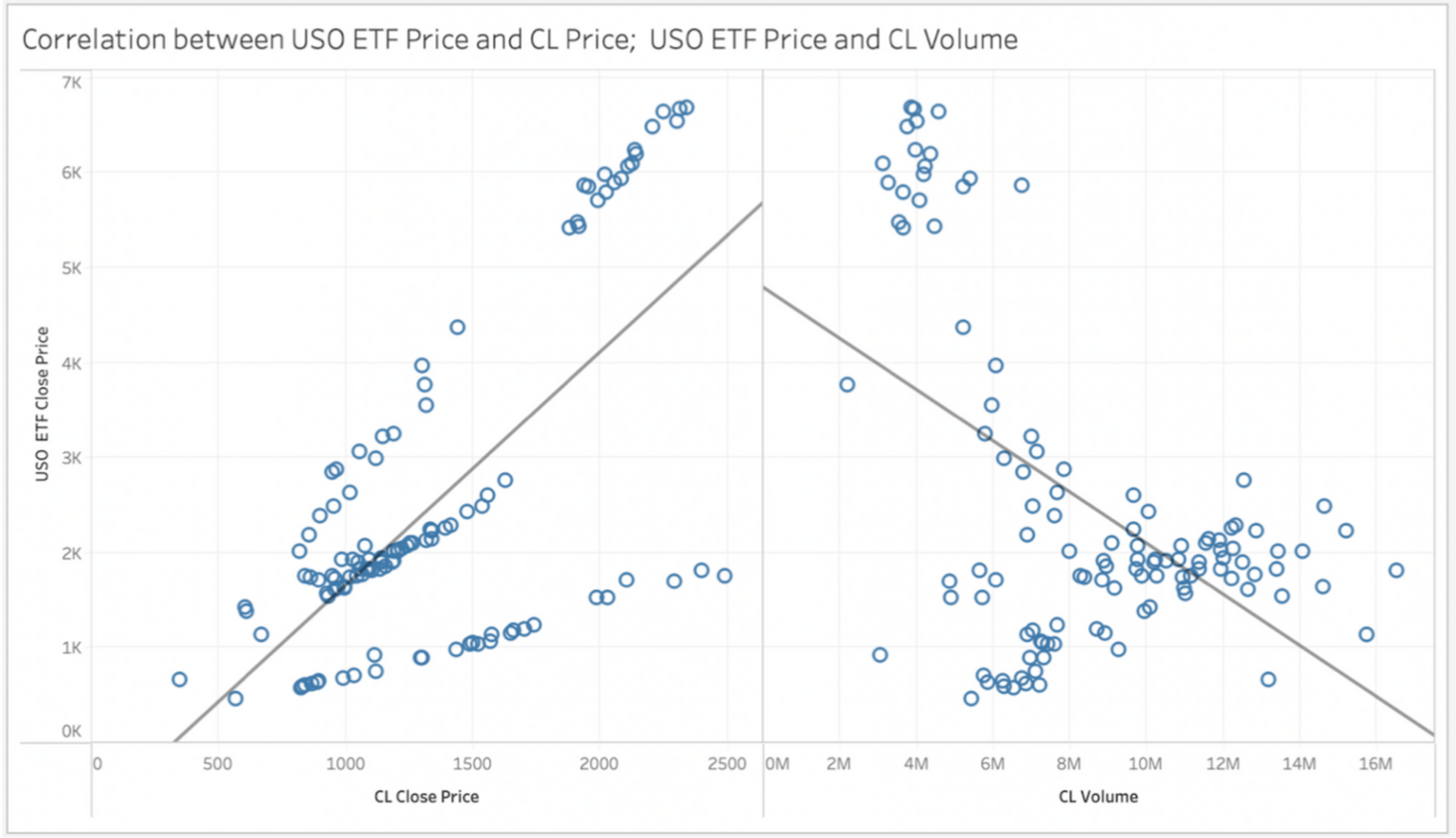
- It is interesting to see that stock prices of USO ETF, Chevron Corp, Shell PLC and Total Energies SE have similar trend during the period February 2016 - March 2020

Correlation between USO ETF Price and CL Price



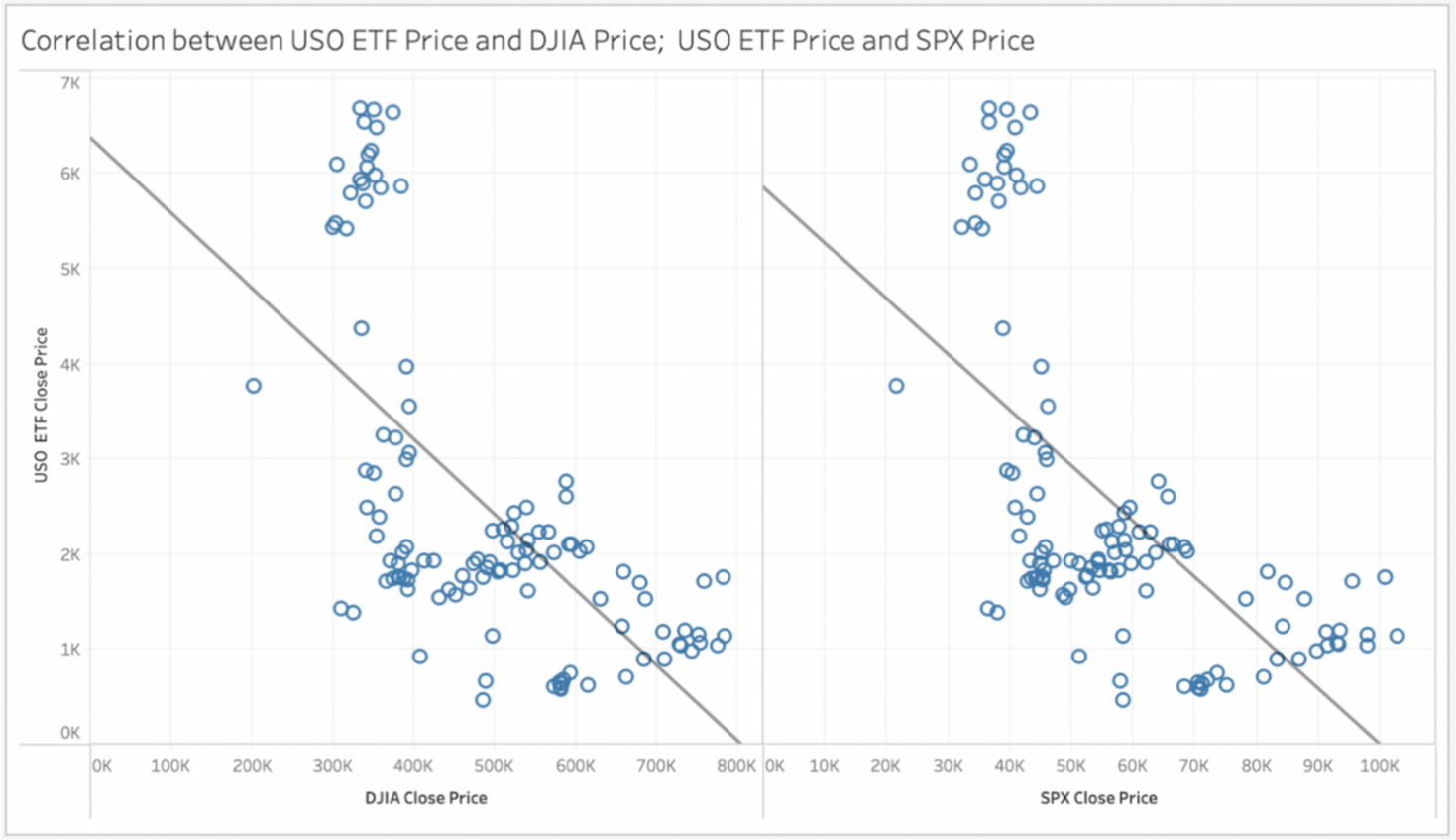
- We can see from the chart below that there is a positive correlation between the close prices of USO ETF and the close prices of WTI Crude Oil (CL) which is obvious because USO ETF is designed to track the performance of the spot price of West Texas Intermediate (WTI) crude oil.

Correlation between USO ETF Price and CL Volume



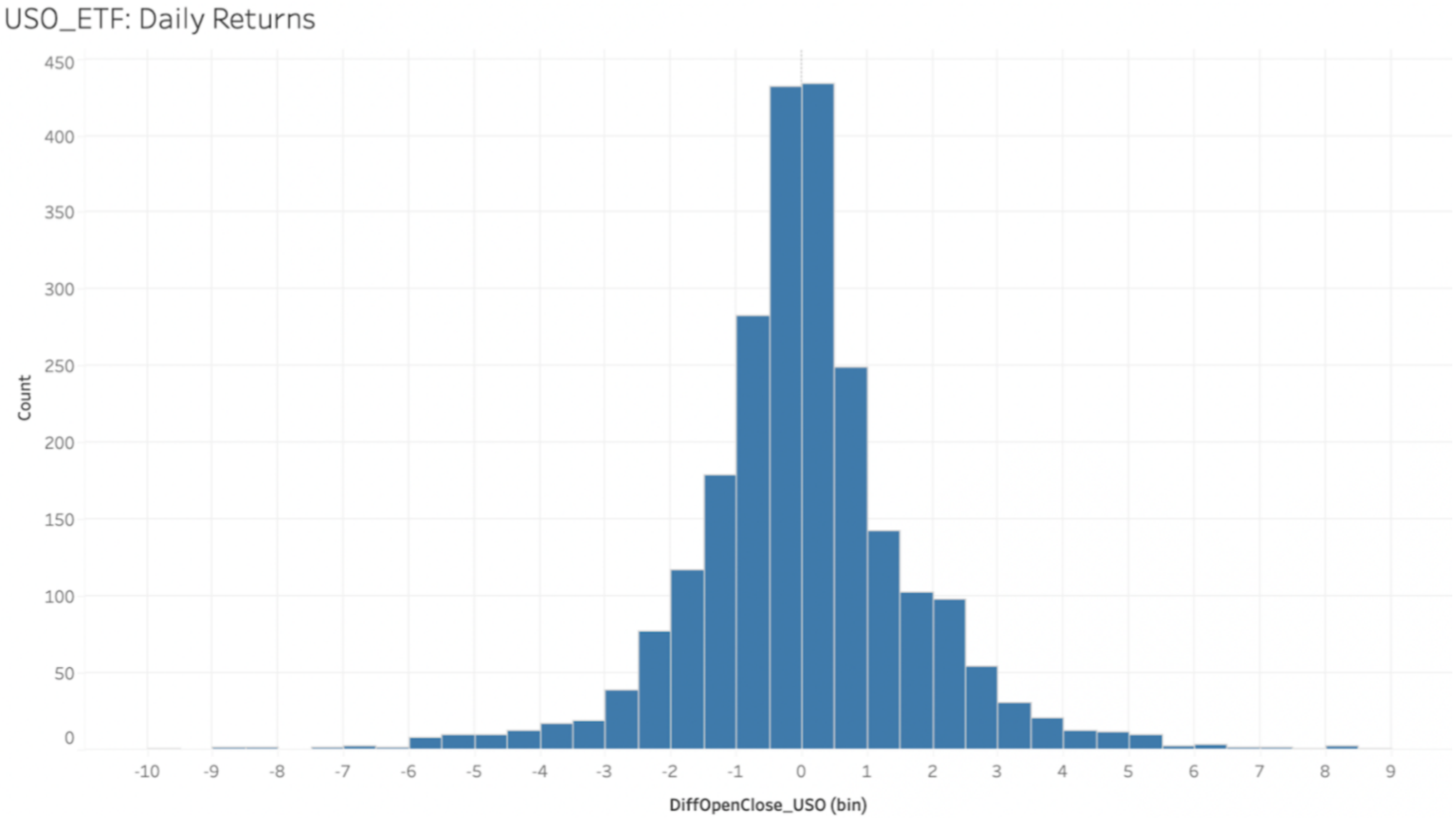
- The negative correlation between the USO ETF and WTI Crude Oil trading volume may be due to the relationship between trading volume and market volatility. Typically, higher trading volume is associated with higher market volatility, which means that prices may fluctuate more rapidly in response to changes in supply and demand or other market events. Conversely, lower trading volume is associated with lower market volatility, which means that prices may be more stable

Correlation between USO ETF Price and Prices of Market Indexes

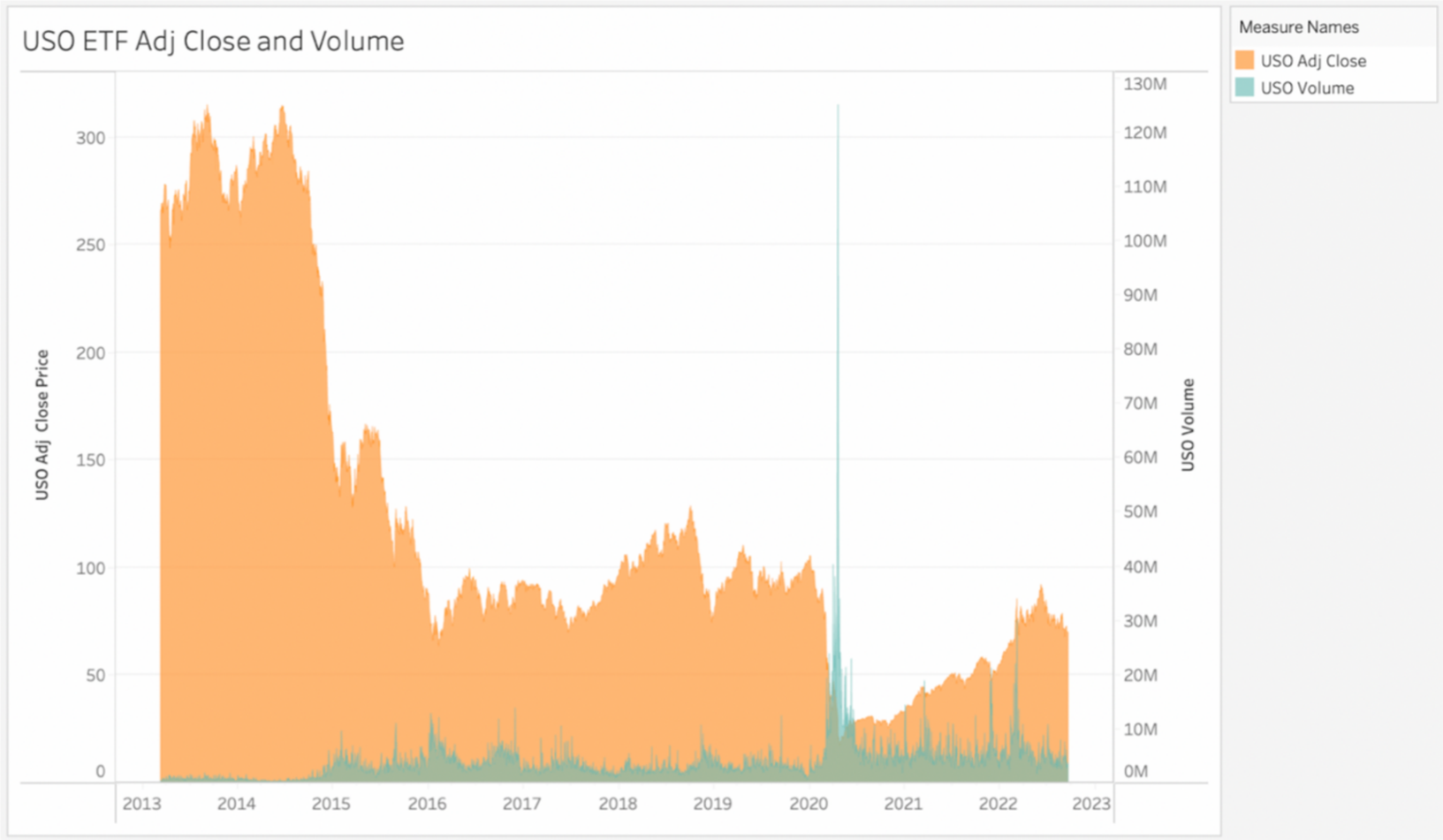


- There is a negative correlation between the close prices of USO ETF and market indexes (S&P 500 and Dow Jones Industrial Average) performance.
- It can be explained by the nature of the oil industry and its relationship with the overall economy. When the economy is performing well, demand for oil increases, and this leads to higher oil prices. However, higher oil prices can have a negative impact on other sectors of the economy, such as manufacturing and transportation, which can cause a decline in stock prices.

Daily returns of USO ETF

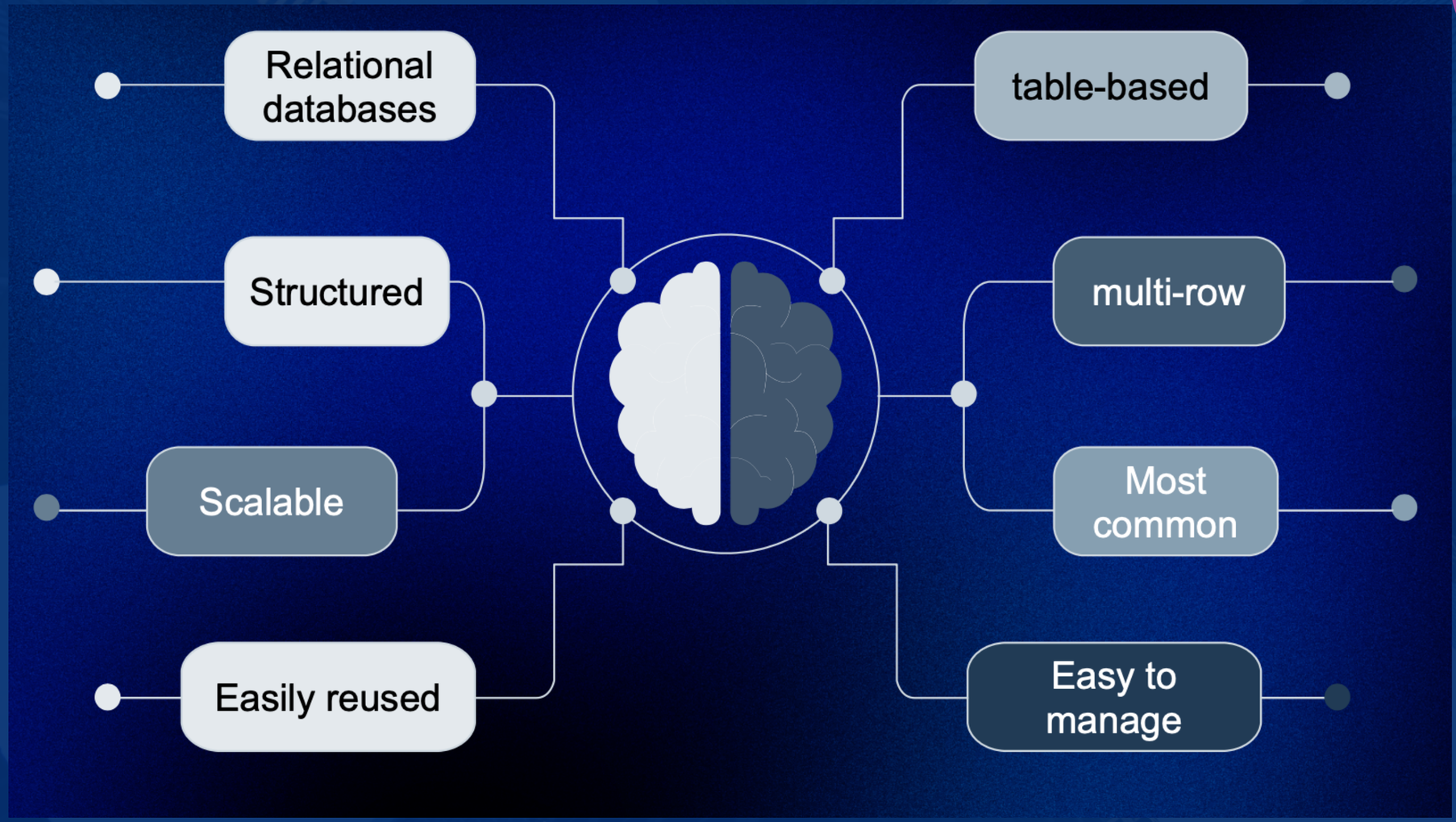


USO EFT Adj Close and Volume

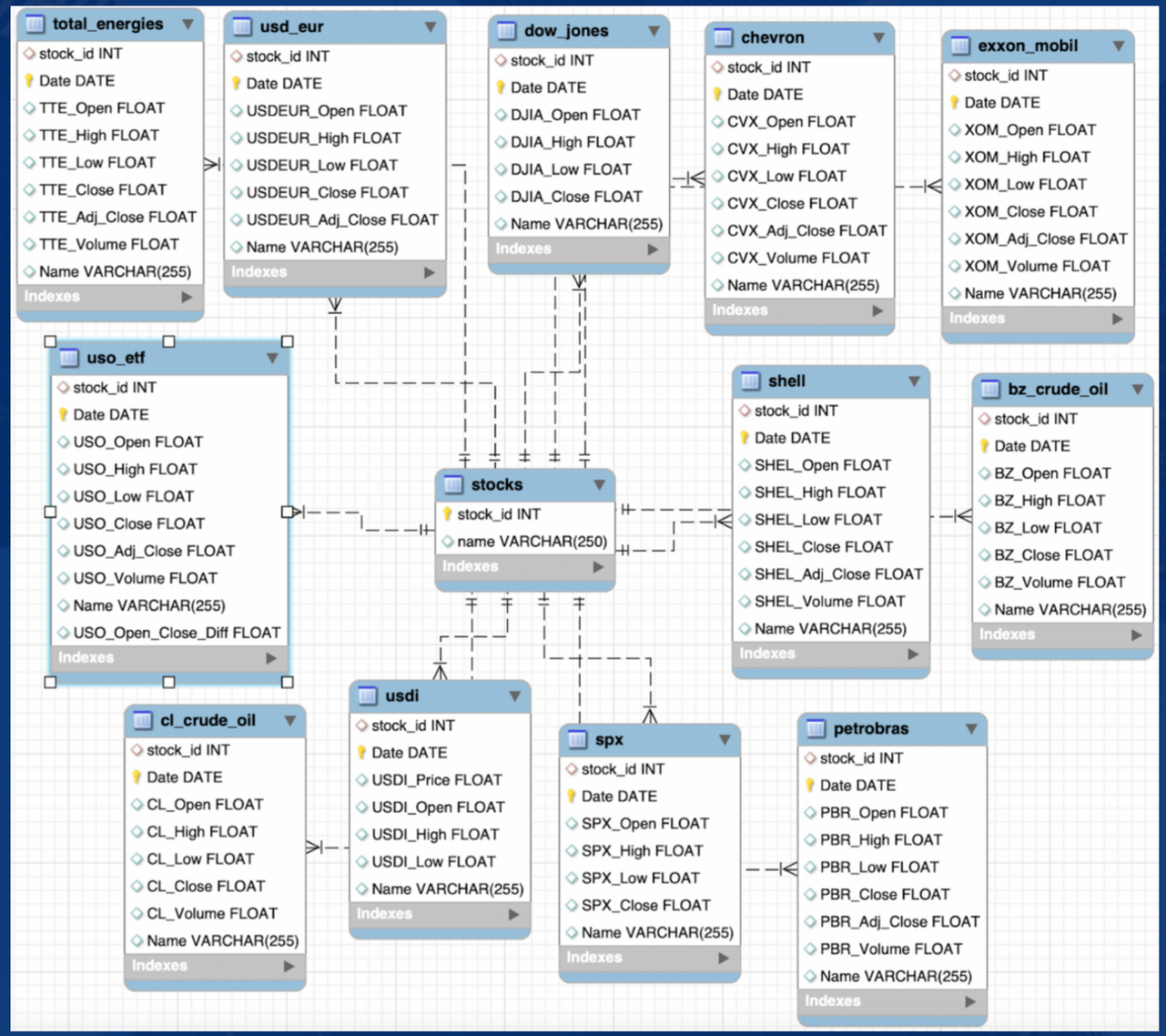


- We can see the downtrend of USO ETF prices since mid-2014
- We can also note that in April 2020 USO ETF trading volume increased sharply. If a stock with a high trading volume is rising, it usually means there is a strong buying pressure as investors demand pushes the stock to higher and higher prices. However in the case of USO ETF the stock with a high trading volume is falling. It suggests that there is a lot of selling pressure which could be the effect of Covid-19

Database selection: Why SQL?



Entity Relationship Diagram (ERD)



MySQL

```sql
CREATE DATABASE USO_ETF;
USE USO_ETF;

CREATE TABLE
    stocks(
        stock_id INT AUTO_INCREMENT,
        name VARCHAR(250) UNIQUE,
        PRIMARY KEY (stock_id)
    );
```

```sql
CREATE TABLE
    IF NOT EXISTS uso_etf(
        stock_id INT,
        Date DATE NOT NULL,
        USO_Open FLOAT,
        USO_High FLOAT,
        USO_Low FLOAT,
        USO_Close FLOAT,
        USO_Adj_Close FLOAT,
        USO_Volume FLOAT,
        Name VARCHAR(255),
        PRIMARY KEY (Date),
        FOREIGN KEY(stock_id) REFERENCES stocks(stock_id)
    );
```

wizard process to import data

```sql
UPDATE exxon_mobil e, stocks s
SET e.stock_id = s.stock_id
WHERE e.Name = s.name;
```

MySQL

```sql
#6. Find the average adjusted closed price for USO for each year:
SELECT
    YEAR(Date) AS Year,
    AVG(USO_Adj_Close) AS Avg_USO_Adj_Close
FROM uso_etf
GROUP BY YEAR(Date);
```

Year	Avg_USO_Adj_Close
2006	497.4778241696565
2007	453.1869324459973
2008	642.6447427621472
2009	274.26539732917905
2010	291.91523900107734
2011	301.4034916105725
2012	283.3529591674805
2013	280.92952274140856
2014	273.01523832290894
2015	132.79238131689647
2016	84.2219044821603
2017	84.01019904148056
2018	106.78183278239581
2019	95.20888876536536
2020	40.593201490258984
2021	47.01202380467975
2022	72.84952192572483

```sql
#5.Find the dates when USO ETF had the highest closing price:
SELECT
    Date,
    USO_Close
FROM uso_etf
WHERE USO_Close = (SELECT MAX(USO_Close) FROM uso_etf);
```

Date	USO_Close
2008-07-14	939.84

Supervised ML models

Target feature : "USO ETF adjusted close"

My goal was to predict what the closing price of USO ETF will be given predictors of the oil market itself as well as other stock market indexes. I have a dataset with around 60 features to predict the adjusted close price of USO ETF. The most important features are: Energy Stocks, S&P prices, DJ prices, NASDAQ prices.

Supervised ML models

I implemented 3 ML models:

Linear Regression with PCA

Gradient Boosting Regressor

Random Forest Regressor

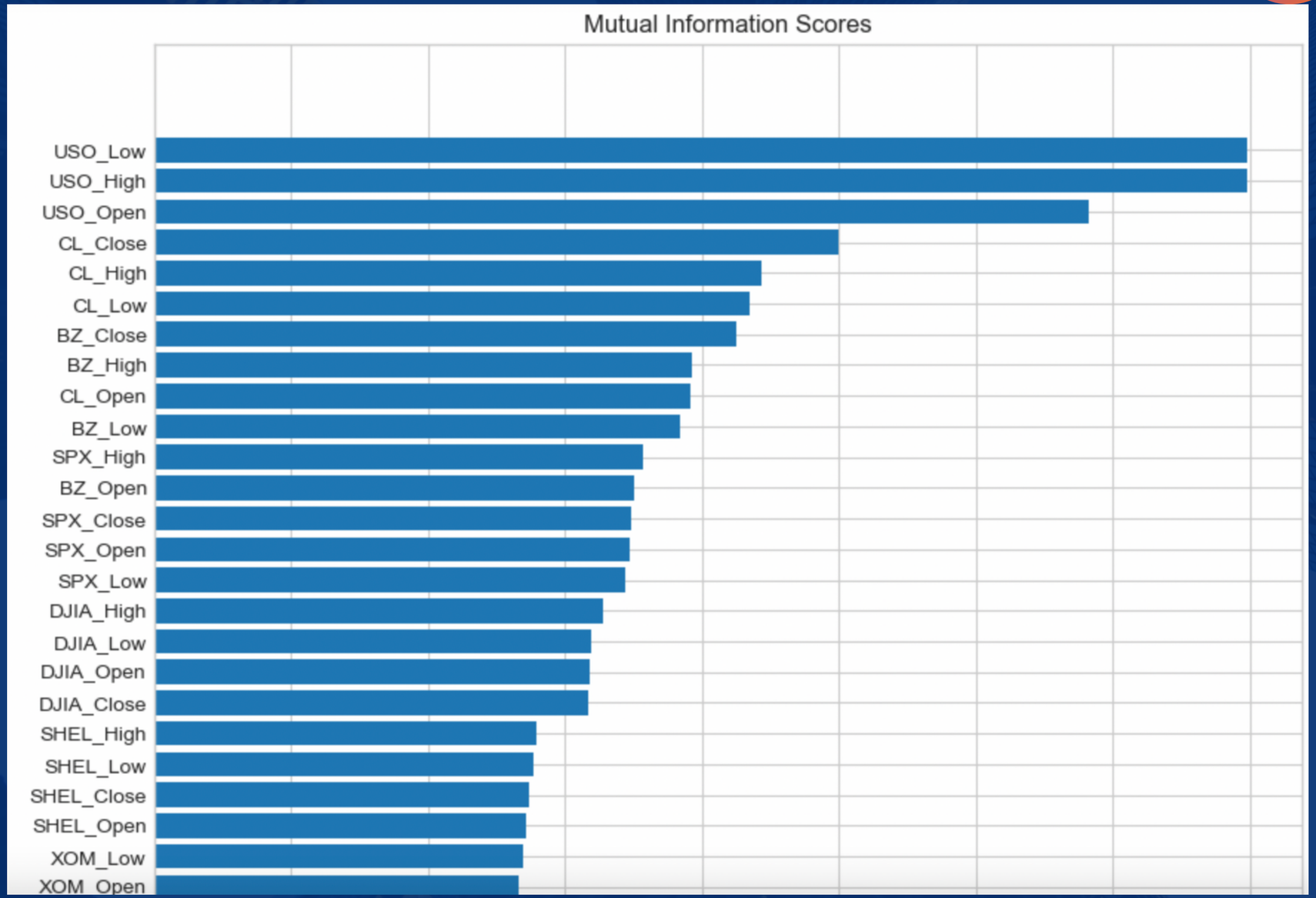
Linear Regression

```python
# Create mutual info scores

def make_mi_scores (X, y):
    mi_scores = mutual_info_regression(X, y)
    mi_scores = pd.Series(mi_scores, name="MI Scores", index=X.columns)
    mi_scores = mi_scores.sort_values(ascending=False)
    return mi_scores

mi_scores = make_mi_scores(X, y)
```

```python
# Construct a bar plot to show each feature's score.

def plot_mi_scores(scores):
    scores = scores.sort_values(ascending=True)
    width = np.arange(len(scores))
    ticks = list(scores.index)
    plt.barh(width, scores)
    plt.yticks(width, ticks)
    plt.title("Mutual Information Scores")

plt.figure(dpi=100, figsize=(10,18))
plot_mi_scores(mi_scores)
```

- I used features engineering to improve the model performance;
- I started by ranking features with mutual information;

Linear Regression



Linear Regression

- I selected the top 12 features except from Crude Oil Prices And USO ETF Open, Low, High Prices because they are very correlated with the target variable
- These 12 features are Prices of Shell PLC and Prices of Market Indexes DJ and S&P 500 (lagged values)
- I've applied PCA

```
X = df.copy()
y = X.pop('USO_Adj_Close')
date = X.pop('Date')
X.pop('USO_Close')
X = X.loc[:, features]

# Standardize the new df. PCA is sensitive to scale.
X_scaled = (X - X.mean(axis=0)) / X.std(axis=0)
```

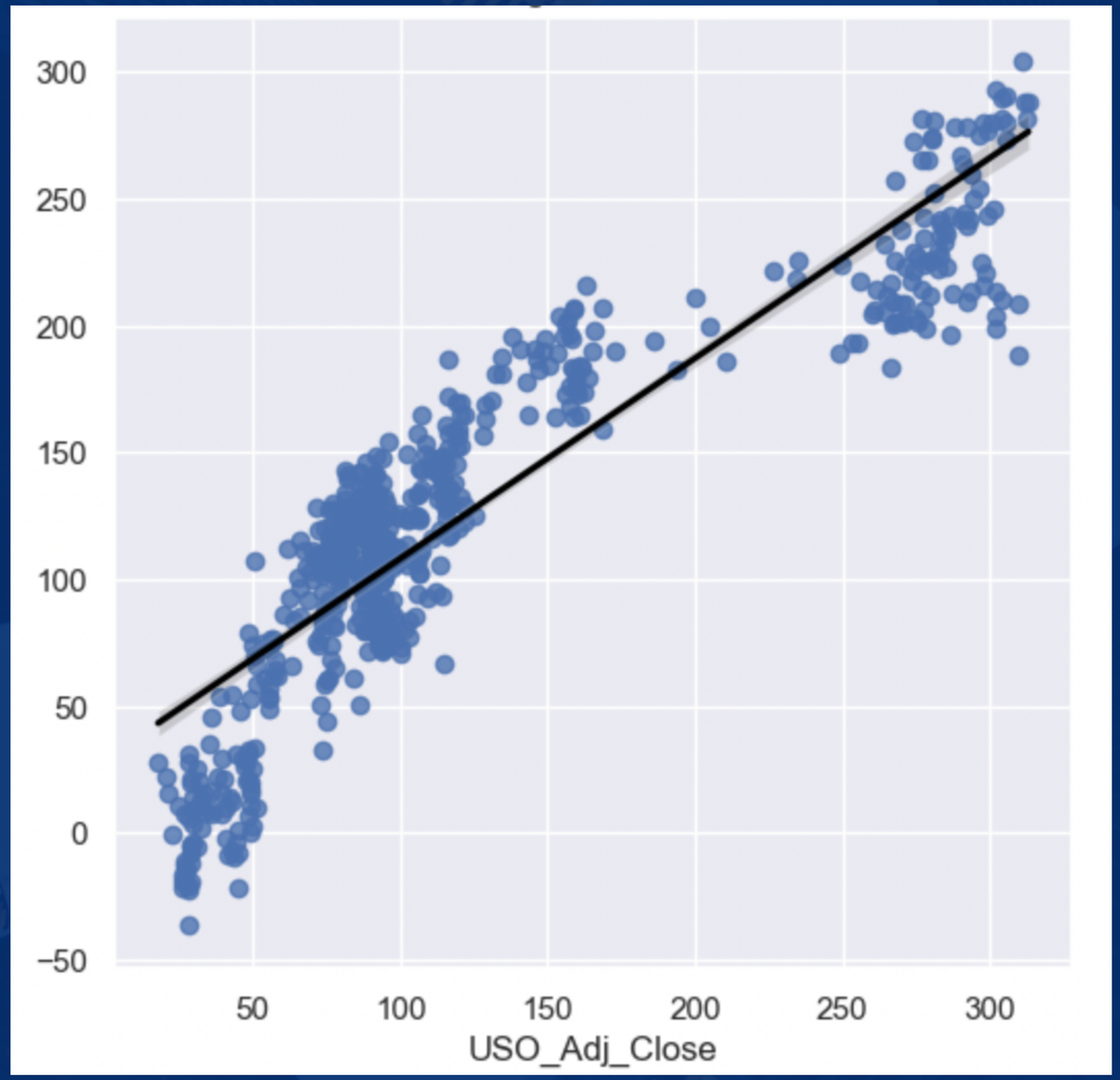
```
# Create principal components
pca = PCA(n_components=10)
X_pca = pca.fit_transform(X_scaled)

# Convert to dataframe
component_names = [f"PC{i+1}" for i in range (X_pca.shape[1])]
X_pca = pd.DataFrame(X_pca, columns=component_names)

X_pca.head()
```

Linear Regression

- I built a new pipeline with PCA

```
# Partition the PCA dataframe into training and validation groups
train_X, val_X, train_y, val_y = train_test_split(X_pca, y, random_state = 0)

lr_model = LinearRegression()

# Bundle preporcessing and modeling code in a pipeline
my_pipeline = Pipeline(steps=[('lr_model', lr_model)])
# Preprocessing of training data, fit model
my_pipeline.fit(train_X, train_y)

# Preprocessing of validation data, get predictions
preds = my_pipeline.predict(val_X)

# Evaluate the model
mae_score = mean_absolute_error(val_y, preds)
print('MAE:', mae_score)

# Display Model
sns.set(rc={"figure.figsize":(6,6)})
sns.regplot(x=val_y, y=preds, line_kws={"color":"black"}).set(title="Linear Regression with PCA")
```

Linear Regression with PCA



Average MAE score: 46.54
RMSE: 35.72
r2: score is 0.80

I run this model through ten-fold cross validation but it showed that PCA didn't significally improve the model performance

Three Models Performance Results

Random Forest Regressor

Average MAE score: 4.54
RMSE: 9.13
r2: score: 0.98

Gradient Boosting Regressor

Average MAE score: 6.10
RMSE: 10.08
r2: score: 0.98

Linear Regression with PCA

Average MAE score: 46.54
RMSE: 35.61
r2: score is 0.80

Random Forest Regressor showed the best results

Auto Regression

```
#AutoRegression

from statsmodels.tsa.api import AutoReg

# Split the data into training and testing sets
train_data = uso["USO_Adj_Close"].loc[:'2021-01-01']
test_data = uso["USO_Adj_Close"].loc['2021-01-01':]

# Fit an autoregressive model with lag 365
model = AutoReg(train_data, lags=365)
result = model.fit()

# Predict the USO ETF prices for the next 21 months
pred = result.predict(start=len(train_data), end=len(train_data)+431)
#print(result.summary)

pred.index = test_data.index

# Plot the actual and predicted USO ETF prices
plt.plot(train_data.index, train_data, label='Training Data')
plt.plot(test_data.index, test_data, label='Testing Data')
plt.plot(pred.index, pred.values, label='Predicted Data')
plt.legend()
plt.show()
```

Prediction of USO ETF price based only on its historical prices

Auto Regression

Prediction of USO ETF price based only its historical prices

Average MAE score: 6.27
RMSE: 7.99
r2: score is 0.72

ARIMA

Prediction of USO ETF price based only its historical prices

Average MAE score: 6.26
RMSE: 7.99
r2: score is 0.72

Challenges and improvements

Database Structure

Improve the database structure. It will be more easy to manage it and to create visualisations in Tableau

Datasets

Include in my dataset more variables to see the impact of major geopolitical events, Technology Stocks and economic conditions on USO EFT prices

Modelling

Create more sophisticated and performant ML predictive models; apply hyper parameter tuning for better model performance

Thank You

Katerina Kovaleva
kd.kovaleva01@gmail.com

IRONHACK DATA ANALYST BOOTCAMP - PARIS MARCH 2023