Final Project

Course: Special Topics (5000)- Big Data
Instructor: Rajeev Maharaj
Presenter: Kiko

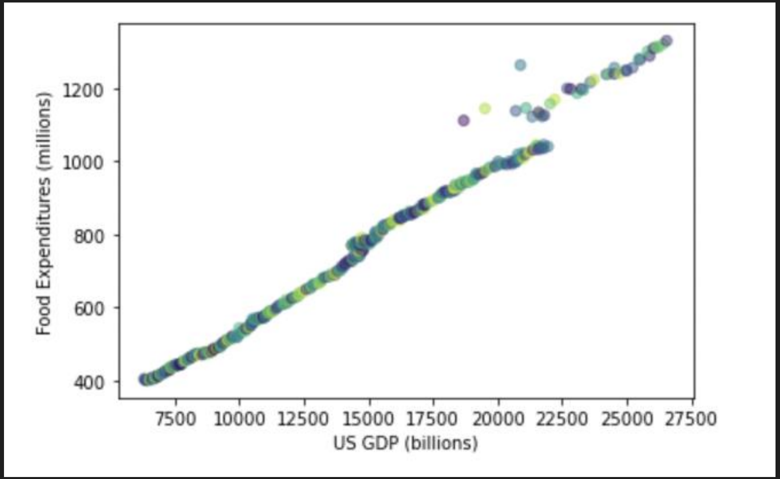
Table of contents

Part 1

Project Improvements

```
In [79]: df.head() # print Head
         df.shape

Out[79]: (374, 9)

In [80]: # Any null variables?
         df.isnull().sum() # There are no missing values in the dataset.

Out[80]: DATE                    0
         Food Expenditures       0
         GDP                     0
         S&P 500                 0
         Home Price Index        0
         Unemployed Rate         0
         Real Personal Income    0
         Retail Sales            0
         CPI                     0
         dtype: int64
```

Check for the Null values data



Scatter between the GDP and Food Expenditures

Standardized/Normalized Coefficients

Standardized Coefficients

```
const                  -2.804359
GDP                     8.639437
S&P 500                 9.840898
Home Price Index        3.802038
Unemployed Rate        15.157339
Real Personal Income   -0.450195
Retail Sales           -1.302611
CPI                     3.686326
dtype: float64
```

Normalized Coefficients

```
GDP                     0.201485
S&P 500                 0.229505
Home Price Index        0.088669
Unemployed Rate         0.353492
Real Personal Income    0.010499
Retail Sales            0.030379
CPI                     0.085971
dtype: float64
```

- Standardized coefficients are expressed in units of standard deviations of the IVs and DV, particularly useful when the IVs are measured on different scales or units, as it allows us to standardize the scale and compare the effect sizes
- Normalized coefficients are scaled to the range [0,1], allow us to evaluate the proportional contribution of each IV to the overall variation in the DV, while holding all other predictors constant and scaling the IVs to the same range.

Compare OLS models

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Food Expenditures   R-squared:                       0.995
Model:                            OLS   Adj. R-squared:                  0.995
Method:                 Least Squares   F-statistic:                 1.126e+04
Date:                Thu, 04 May 2023   Prob (F-statistic):               0.00
Time:                        19:46:17   Log-Likelihood:                -1581.3
No. Observations:                 374   AIC:                             3179.
Df Residuals:                     366   BIC:                             3210.
Df Model:                           7
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                  -89.5902     31.947     -2.804      0.005    -152.412     -26.768
GDP                      0.0296      0.003      8.639      0.000       0.023       0.036
S&P 500                  0.0411      0.004      9.841      0.000       0.033       0.049
Home Price Index         0.2501      0.066      3.802      0.000       0.121       0.379
Unemployed Rate         10.9683      0.724     15.157      0.000       9.545      12.391
Real Personal Income    -0.0008      0.002     -0.450      0.653      -0.004       0.003
Retail Sales            -0.0001      0.000     -1.303      0.194      -0.000    6.79e-05
CPI                      1.4669      0.398      3.686      0.000       0.684       2.249
==============================================================================
Omnibus:                      523.386   Durbin-Watson:                   1.266
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           125182.891
Skew:                           6.705   Prob(JB):                         0.00
Kurtosis:                      91.619   Cond. No.                     1.41e+07
==============================================================================
```

OLS Model with Monthly Data

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Food Expenditures   R-squared:                       0.994
Model:                            OLS   Adj. R-squared:                  0.993
Method:                 Least Squares   F-statistic:                     777.0
Date:                Sat, 25 Mar 2023   Prob (F-statistic):           1.00e-34
Time:                        22:04:09   Log-Likelihood:                -153.62
No. Observations:                  41   AIC:                             323.2
Df Residuals:                      33   BIC:                             336.9
Df Model:                           7
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                 -352.4315    170.300     -2.069      0.046    -698.909      -5.955
GDP                     -0.0074      0.013     -0.580      0.566      -0.033       0.019
S&P 500                  0.0270      0.012      2.241      0.032       0.002       0.052
Home Price Index         0.8567      0.465      1.844      0.074      -0.088       1.802
Unemployed Rate          4.8120      2.955      1.628      0.113      -1.200      10.824
Real Personal Income     0.0197      0.005      3.599      0.001       0.009       0.031
Retail Sales          8.234e-05      0.000      0.435      0.666      -0.000       0.000
CPI                      3.5342      1.355      2.608      0.014       0.777       6.292
==============================================================================
Omnibus:                       18.592   Durbin-Watson:                   1.825
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               29.187
Skew:                           1.261   Prob(JB):                     4.59e-07
Kurtosis:                       6.275   Cond. No.                     4.87e+07
==============================================================================
```

OLS Model with Quarterly Data

Bigger sample size, higher R-squared, better and more accurate prediction

VIF Comparison

	variables	VIF
0	GDP	12197.642069
1	S&P 500	325.994025
2	Home Price Index	1689.876666
3	Unemployed Rate	57.352612
4	Real Personal Income	1411.197878
5	Retail Sales	2677.502748
6	CPI	5548.297449

VIF for Quarterly Data

	variables	VIF
0	GDP	749.895225
1	S&P 500	58.660310
2	Home Price Index	135.346486
3	Unemployed Rate	18.375398
4	Real Personal Income	685.659645
5	Retail Sales	1881.955387
6	CPI	956.176722

VIF for Monthly Data

Bigger sample size, lower VIF, mitigating risks of multicollinearity

Autoregression Model

What does this mean?

AutoReg Model Results

Dep. Variable:	Food Expenditures	No. Observations:	272
Model:	AutoReg(8)	Log Likelihood	-692.995
Method:	Conditional MLE	S.D. of innovations	3.340
Date:	Thu, 04 May 2023	AIC	1405.991
Time:	19:47:32	BIC	1441.750
Sample:	8 272	HQIC	1420.360

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.6801	0.882	0.771	0.441	-1.049	2.409
Food Expenditures.L1	0.5602	0.061	9.225	0.000	0.441	0.679
Food Expenditures.L2	0.3304	0.070	4.741	0.000	0.194	0.467
Food Expenditures.L3	0.1904	0.073	2.620	0.009	0.048	0.333
Food Expenditures.L4	0.0522	0.074	0.709	0.478	-0.092	0.196
Food Expenditures.L5	-0.0175	0.074	-0.238	0.812	-0.162	0.127
Food Expenditures.L6	-0.0199	0.073	-0.274	0.784	-0.162	0.122
Food Expenditures.L7	0.0692	0.070	0.992	0.321	-0.067	0.206
Food Expenditures.L8	-0.1631	0.061	-2.688	0.007	-0.282	-0.044

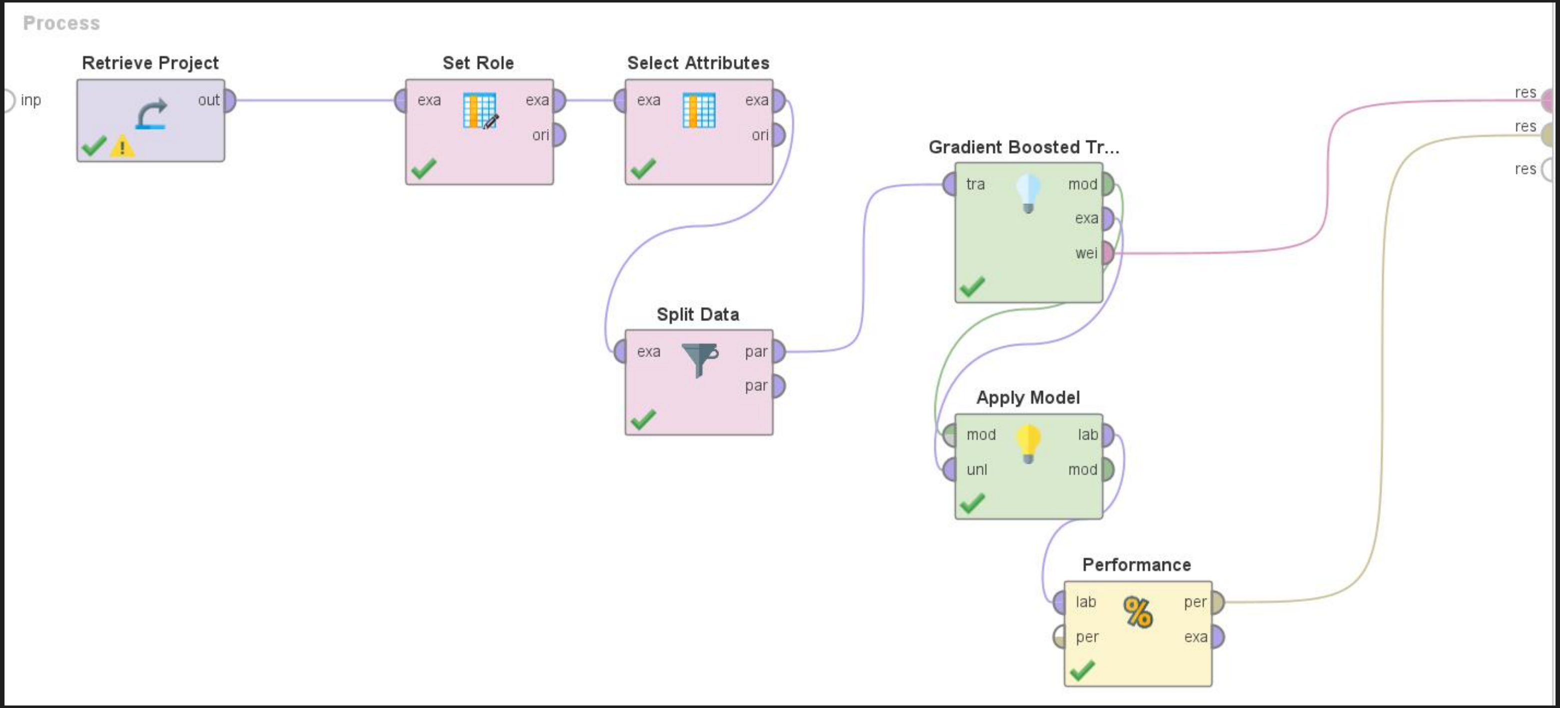
Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-1.1555	-0.5120j	1.2638	-0.4336
AR.2	-1.1555	+0.5120j	1.2638	0.4336
AR.3	-0.4067	-1.2085j	1.2751	-0.3017
AR.4	-0.4067	+1.2085j	1.2751	0.3017
AR.5	0.9981	-0.0000j	0.9981	-0.0000
AR.6	1.2210	-0.0000j	1.2210	-0.0000
AR.7	0.6648	-1.2227j	1.3918	-0.1707
AR.8	0.6648	+1.2227j	1.3918	0.1707

The autoregression model shows that past values of Food Expenditures have a significant impact on current values

With the strongest impact coming from the first and second lagged variables. The model can be used to predict future values of Food Expenditures based on past values.

Rapidminer Comparison



Comparison

Rapidminer

Attribute	Weight
CPI	12,988,241
Home Price Index	1,432,044
S&P 500	106,616
Unemployment Rate	73,757
GDP	35,991
Retail Sales	30,543
Real Personal Income	7,459

Linear Model

Attribute	Weight
Unemployment Rate	15.16
S&P 500	9.84
GDP	8.64
Home Price Index	3.80
CPI	3.69
Retail Sales	1.30
Real Personal Income	0.45

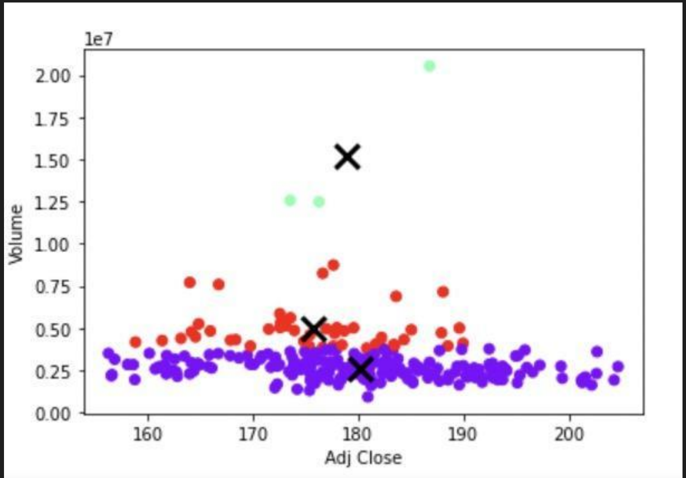
Comparison

Criterion	Value from Rapidminer	Value from Linear Model
Root Mean Squared Error	1.581	12.883
Relative Error	0.16% +/- 0.17%	1.63%
Squared Error	2.501 +/- 6.591	165.961
Squared Correlation	1.000	0.99745

Part 2
Trading Strategy

Cluster & Regression

```
Cluster centers:
 [[1.80161587e+02 2.59651436e+06]
 [1.78874588e+02 1.51909333e+07]
 [1.75745935e+02 4.94874783e+06]]
```



Cluster center with the highest 'Adj Close' value (1.801) and the lowest 'Volume' value (2.59 million) represents a group of stocks with high prices but low trading activity;

Cluster center with the lowest 'Adj Close' value (1.75) and the highest 'Volume' value (4.94 million) represents a group of stocks with lowest price but high trading activity;

A unit increase in the 'Adj Close' feature, the model predicts an increase of 0.9677 in the Close, holding all other features constant.

```
Coefficients: [ 9.67770334e-01 -1.17146491e-07  2.19059972e
Intercept: 8.700571860585171
```

For a unit increase in the 'Volume' feature, the model predicts an increase of 2.1906 in the Close, holding all other features constant.

Regression Model Summary

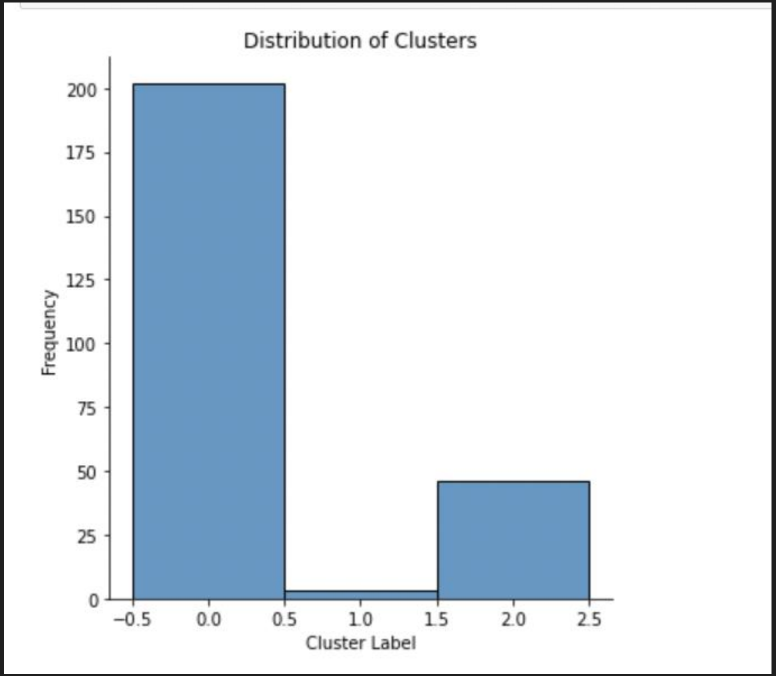
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Close   R-squared:                       0.972
Model:                            OLS   Adj. R-squared:                  0.971
Method:                 Least Squares   F-statistic:                     2817.
Date:                Fri, 05 May 2023   Prob (F-statistic):          1.17e-190
Time:                        14:11:58   Log-Likelihood:                -497.09
No. Observations:                 251   AIC:                             1002.
Df Residuals:                     247   BIC:                             1016.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         8.7006      1.951      4.459      0.000       4.857      12.544
Adj Close     0.9678      0.011     90.555      0.000       0.947       0.989
Volume    -1.171e-07    7.5e-08     -1.562      0.120   -2.65e-07    3.06e-08
cluster       0.2191      0.175      1.250      0.213      -0.126       0.564
==============================================================================
Omnibus:                      113.399   Durbin-Watson:                   0.039
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               15.754
Skew:                           0.170   Prob(JB):                     0.000379
Kurtosis:                       1.821   Cond. No.                     6.39e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.39e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

R-squared value of 0.972 indicates that 97.2% of the variation in the 'Close' stock price is explained by the predictor variables included in the model.

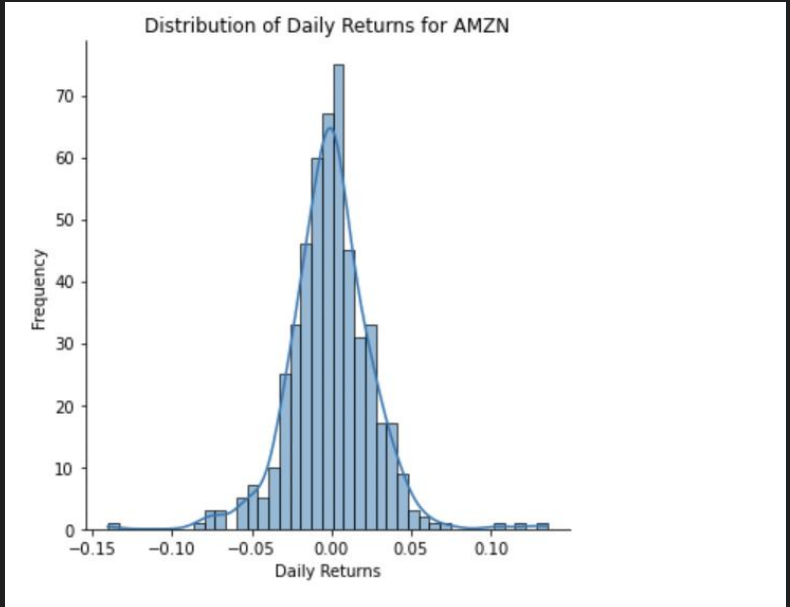
F-statistic of 2817 and its associated probability (p-value) of 1.17e-190 indicate that the overall model is statistically significant and the predictor variables are jointly significant in predicting the 'Close' stock price.

Cluster - Amazon Stock

Data range: 2021-04-27 - 2023-04-26



Positively Skewed (Skewed to the right)

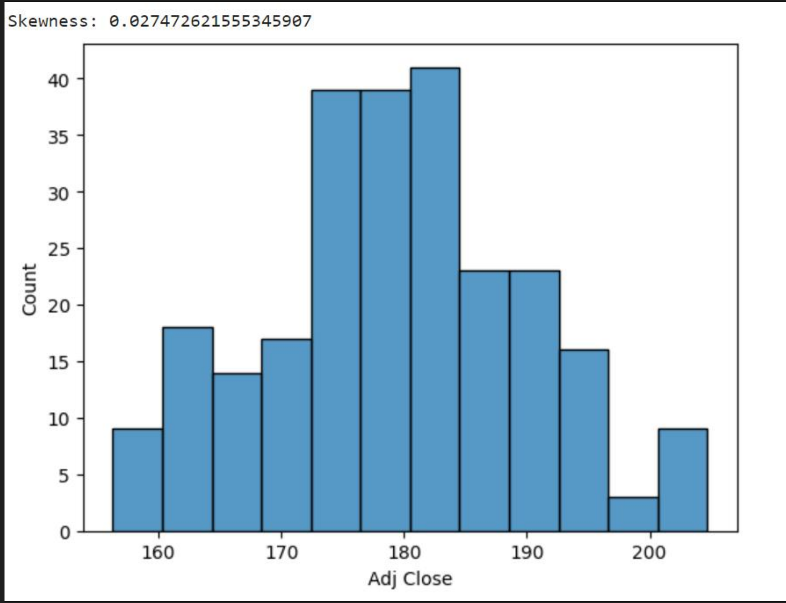


Mean Daily Return: -0.00063726866387762
Standard Deviation of Daily Return: 0.025729389715897808

Normal Distribution

Cluster - Amazon Stock

Data range: 2021-04-27 - 2023-04-26



Positively Skewed (Skewed to the right)

Moving Averages

```
5-day moving average:
 Date
2021-04-27           NaN
2021-04-28           NaN
2021-04-29           NaN
2021-04-30           NaN
2021-05-03    172.011502
                 ...
2023-04-20    103.132001
2023-04-21    104.022000
2023-04-24    104.716000
2023-04-25    104.770000
2023-04-26    104.906000
Name: Adj Close, Length: 504, dtype: float64
10-day moving average:
 Date
2021-04-27           NaN
2021-04-28           NaN
2021-04-29           NaN
2021-04-30           NaN
2021-05-03           NaN
                 ...
2023-04-20    102.004000
2023-04-21    102.494000
2023-04-24    102.898000
2023-04-25    103.163001
2023-04-26    103.878001
Name: Adj Close, Length: 504, dtype: float64
20-day moving average:
 Date
2021-04-27         NaN
2021-04-28         NaN
2021-04-29         NaN
2021-04-30         NaN
2021-05-03         NaN
                ...
2023-04-20    101.2580
2023-04-21    101.6705
2023-04-24    102.0745
2023-04-25    102.3010
2023-04-26    102.6880
Name: Adj Close, Length: 504, dtype: float64
```

```
50-day moving average:
 Date
2021-04-27        NaN
2021-04-28        NaN
2021-04-29        NaN
2021-04-30        NaN
2021-05-03        NaN
                ...
2023-04-20    98.1482
2023-04-21    98.2864
2023-04-24    98.4458
2023-04-25    98.5450
2023-04-26    98.6538
Name: Adj Close, Length: 504, dtype: float64
100-day moving average:
 Date
2021-04-27        NaN
2021-04-28        NaN
2021-04-29        NaN
2021-04-30        NaN
2021-05-03        NaN
                ...
2023-04-20    95.4062
2023-04-21    95.5417
2023-04-24    95.6643
2023-04-25    95.7658
2023-04-26    95.8502
Name: Adj Close, Length: 504, dtype: float64
200-day moving average:
 Date
2021-04-27          NaN
2021-04-28          NaN
2021-04-29          NaN
2021-04-30          NaN
2021-05-03          NaN
                ...
2023-04-20    107.03585
2023-04-21    106.99900
2023-04-24    106.94840
2023-04-25    106.88355
2023-04-26    106.84970
Name: Adj Close, Length: 504, dtype: float64
```

Distance From Moving Average

```
Distance from 5-day moving average:
 Date
2021-04-27          NaN
2021-04-28          NaN
2021-04-29          NaN
2021-04-30          NaN
2021-05-03    -2.687009
                ...
2023-04-20     0.677997
2023-04-21     2.937999
2023-04-24     1.493999
2023-04-25    -2.200000
2023-04-26     0.074003
Name: Adj Close, Length: 504, dtype: float64
Distance from 10-day moving average:
 Date
2021-04-27          NaN
2021-04-28          NaN
2021-04-29          NaN
2021-04-30          NaN
2021-05-03          NaN
                ...
2023-04-20     1.805998
2023-04-21     4.465999
2023-04-24     3.311999
2023-04-25    -0.593001
2023-04-26     1.102002
Name: Adj Close, Length: 504, dtype: float64
Distance from 20-day moving average:
 Date
2021-04-27          NaN
2021-04-28          NaN
2021-04-29          NaN
2021-04-30          NaN
2021-05-03          NaN
                ...
2023-04-20     2.551998
2023-04-21     5.289499
2023-04-24     4.135499
2023-04-25     0.269000
2023-04-26     2.292003
Name: Adj Close, Length: 504, dtype: float64
```

```
Distance from 50-day moving average:
 Date
2021-04-27         NaN
2021-04-28         NaN
2021-04-29         NaN
2021-04-30         NaN
2021-05-03         NaN
                ...
2023-04-20    5.661798
2023-04-21    8.673599
2023-04-24    7.764199
2023-04-25    4.025000
2023-04-26    6.326203
Name: Adj Close, Length: 504, dtype: float64
Distance from 100-day moving average:
 Date
2021-04-27          NaN
2021-04-28          NaN
2021-04-29          NaN
2021-04-30          NaN
2021-05-03          NaN
                ...
2023-04-20     8.403798
2023-04-21    11.418299
2023-04-24    10.545699
2023-04-25     6.804200
2023-04-26     9.129803
Name: Adj Close, Length: 504, dtype: float64
Distance from 200-day moving average:
 Date
2021-04-27         NaN
2021-04-28         NaN
2021-04-29         NaN
2021-04-30         NaN
2021-05-03         NaN
                ...
2023-04-20   -3.225852
2023-04-21   -0.039001
2023-04-24   -0.738401
2023-04-25   -4.313550
2023-04-26   -1.869697
Name: Adj Close, Length: 504, dtype: float64
```

Standard Deviation

```
# Calculate standard deviation of the stock
std = df['Close'].std()
print(std)
```

```
31.408318029079926
```

The closing prices of the AMZN stock are spread out or dispersed on average by 31.4 units away from the mean.

If standard deviation is high, it indicated the stock price is highly volatile and can fluctuate widely from the average price;

If the standard deviation is low, the stock price is less volatile and tends to be stable;

It is a critical tool to assess the degrees of the risk;

Create Signal

```
Signal based on 5-day moving average:
 Date
2021-04-27    sell
2021-04-28    sell
2021-04-29    sell
2021-04-30    sell
2021-05-03    sell
              ...
2023-04-20     buy
2023-04-21     buy
2023-04-24     buy
2023-04-25    sell
2023-04-26     buy
Name: signal_ma5, Length: 504, dtype: object
Signal based on 10-day moving average:
 Date
2021-04-27    sell
2021-04-28    sell
2021-04-29    sell
2021-04-30    sell
2021-05-03    sell
              ...
2023-04-20     buy
2023-04-21     buy
2023-04-24     buy
2023-04-25    sell
2023-04-26     buy
Name: signal_ma10, Length: 504, dtype: object
Signal based on 20-day moving average:
 Date
2021-04-27    sell
2021-04-28    sell
2021-04-29    sell
2021-04-30    sell
2021-05-03    sell
              ...
2023-04-20     buy
2023-04-21     buy
2023-04-24     buy
2023-04-25     buy
2023-04-26     buy
Name: signal_ma20, Length: 504, dtype: object
```

```
Signal based on 50-day moving average:
 Date
2021-04-27    sell
2021-04-28    sell
2021-04-29    sell
2021-04-30    sell
2021-05-03    sell
              ...
2023-04-20     buy
2023-04-21     buy
2023-04-24     buy
2023-04-25     buy
2023-04-26     buy
Name: signal_ma50, Length: 504, dtype: object
Signal based on 100-day moving average:
 Date
2021-04-27    sell
2021-04-28    sell
2021-04-29    sell
2021-04-30    sell
2021-05-03    sell
              ...
2023-04-20     buy
2023-04-21     buy
2023-04-24     buy
2023-04-25     buy
2023-04-26     buy
Name: signal_ma100, Length: 504, dtype: object
Signal based on 200-day moving average:
 Date
2021-04-27    sell
2021-04-28    sell
2021-04-29    sell
2021-04-30    sell
2021-05-03    sell
              ...
2023-04-20    sell
2023-04-21    sell
2023-04-24    sell
2023-04-25    sell
2023-04-26    sell
Name: signal_ma200, Length: 504, dtype: object
```

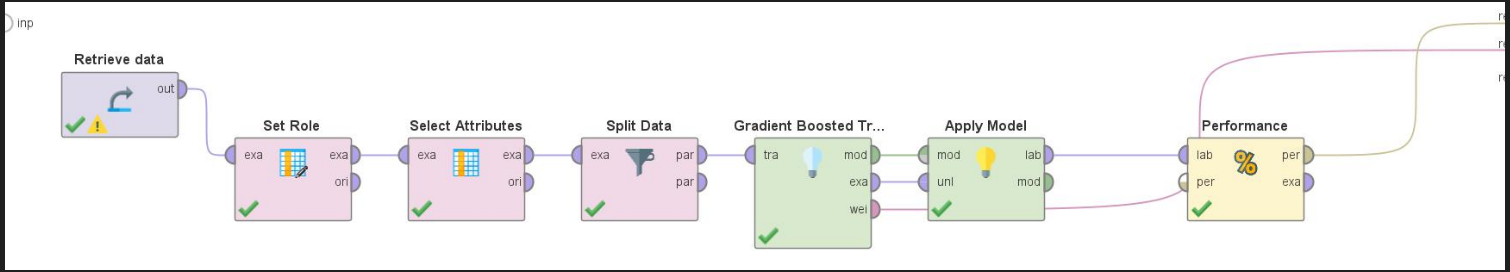
trading above a particular moving average can be a useful strategy

it suggests that the stock is in an uptrend and has upward momentum

When the stock price is consistently above the moving average, it indicates that the trend is up, and traders may consider buying the stock in anticipation of further price increases.

More factors to consider: company fundamental , economic conditions, overall marketing trends, etc.

Rapidminer



Criterion	Value from Rapidminer
Root Mean Squared Error	6.40
Relative Error	2.9% +/- 2.16%
Squared Error	40.959 +/- 52.167
Squared Correlation	0.965

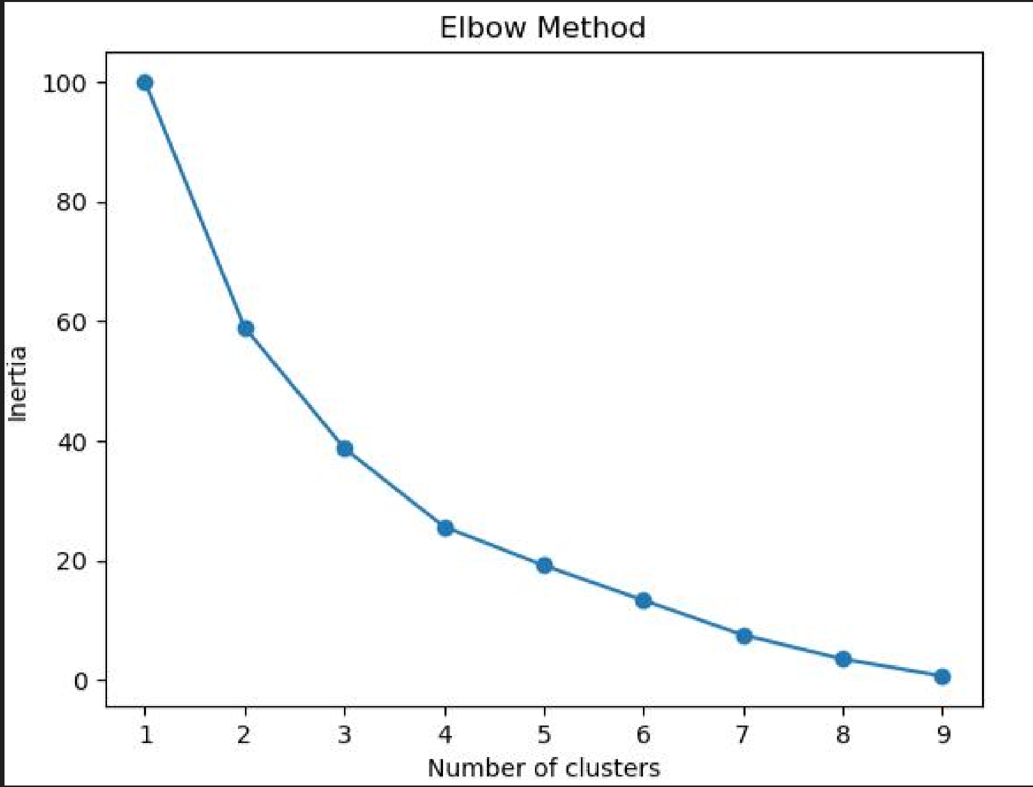
Attribute	Weight
ma_5	492,223
ma_100	31,373
ma_50	7,788
ma_200	1,667
ma_20	753
ma_10	752

New Cluster

Correlation

	AMZN	BAC	CVX	GE	JNJ	NKE	TSLA	UNH	V	XOM
AMZN	1.000000	0.316050	0.207868	0.241136	0.269406	0.476263	0.477203	0.331206	0.446870	0.191731
BAC	0.316050	1.000000	0.683110	0.694206	0.469716	0.522982	0.284419	0.542879	0.653686	0.648667
CVX	0.207868	0.683110	1.000000	0.609137	0.425959	0.423944	0.254490	0.531661	0.564129	0.867385
GE	0.241136	0.694206	0.609137	1.000000	0.380753	0.486613	0.244797	0.436342	0.558591	0.598655
JNJ	0.269406	0.469716	0.425959	0.380753	1.000000	0.390823	0.136380	0.622621	0.521647	0.343339
NKE	0.476263	0.522982	0.423944	0.486613	0.390823	1.000000	0.357895	0.470866	0.608410	0.391003
TSLA	0.477203	0.284419	0.254490	0.244797	0.136380	0.357895	1.000000	0.257951	0.382373	0.189226
UNH	0.331206	0.542879	0.531661	0.436342	0.622621	0.470866	0.257951	1.000000	0.596118	0.429546
V	0.446870	0.653686	0.564129	0.558591	0.521647	0.608410	0.382373	0.596118	1.000000	0.497600
XOM	0.191731	0.648667	0.867385	0.598655	0.343339	0.391003	0.189226	0.429546	0.497600	1.000000

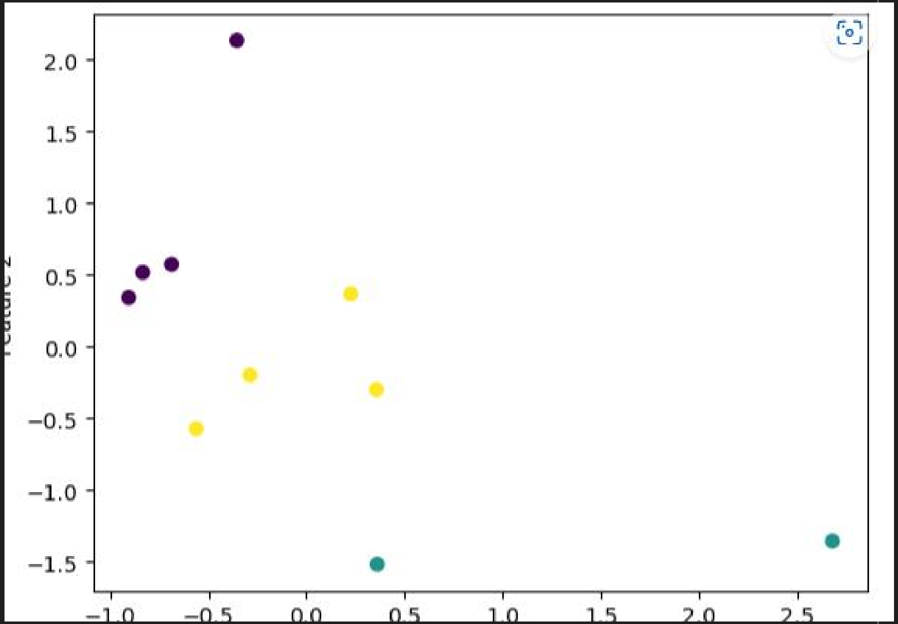
Cluster Choose



According to Elbow Curve, We Believed that cluster of 3 would be optimal

Cluster

	Ticker	Cluster
0	AMZN	1
1	BAC	2
2	CVX	2
3	GE	2
4	JNJ	0
5	NKE	0
6	TSLA	1
7	UNH	0
8	V	0
9	XOM	2



Thanks!