Forecasting of S&P500 Index

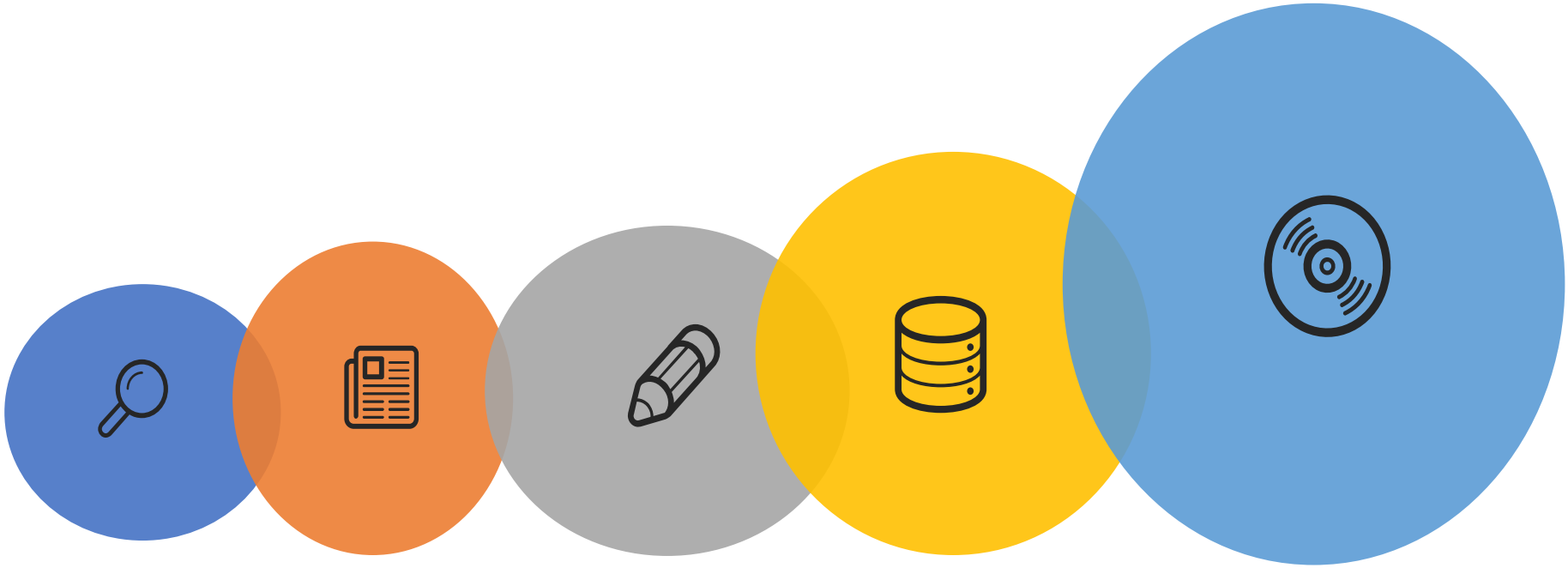
UP

With all forecasting methods

DOWN

Success is not guaranteed!

Logical Diagram of the Research



Literature Research

Godfrey (1964)

Stock price prediction is considered impossible according to the random walk hypothesis, which states the stock market prices moves just like a random walk

Pilinkus (2010)

The relation between macroeconomic indicators and stock prices is confirmed in the most academic works, although there is a lack of comprehensive assessment of causality and dependence of macroeconomic indicators and stock market regarding the time and changing macroeconomic processes

Prazak's (2018)

General macroeconomic indicators provide a statistically significant impact on stock prices in the long run, though strength of the impact may substantially
vary among economic sectors

Wirajaya (2019)

Exchange rates can affect domestic investments such as stocks. This situation will cause a decrease in demand for shares so that stock prices decline.

Data Collection

Research periods:

A 01.01.2010-28.02.2022 (Monthly Basis)

B 01.01.2019-28.02.2022 (Daily Basis)

Price of the S&P 500 index

YAHOO! FINANCE

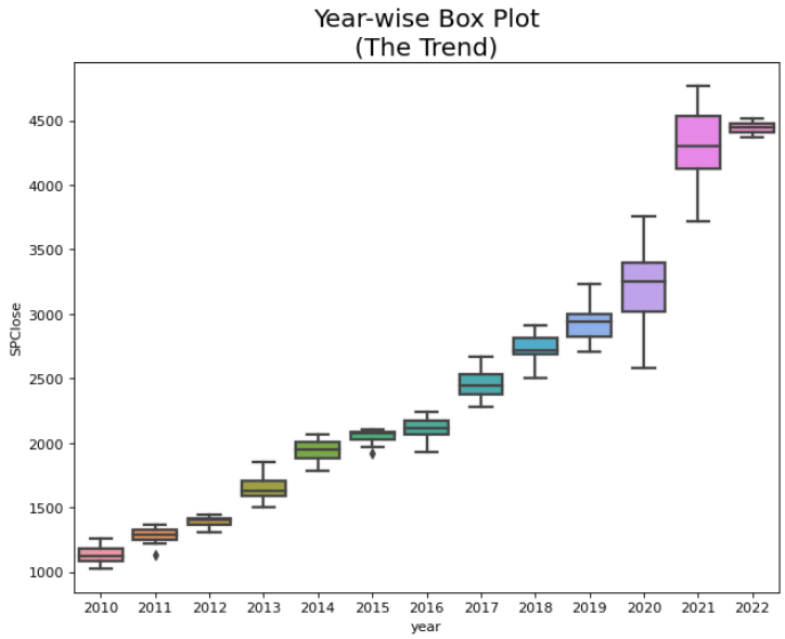
Data of economic indicators

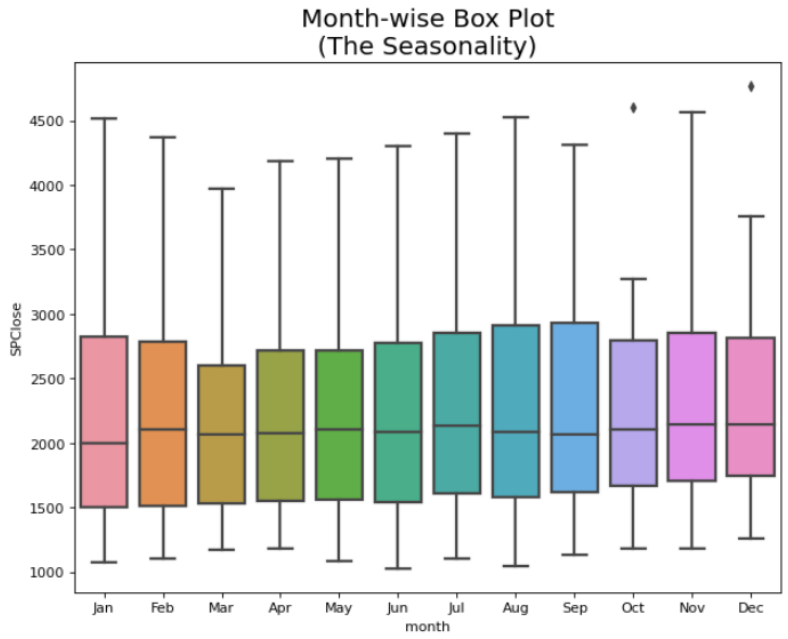
FRED® ECONOMIC DATA | ST. LOUIS FED

Data Exploration

Trend & Seasonality

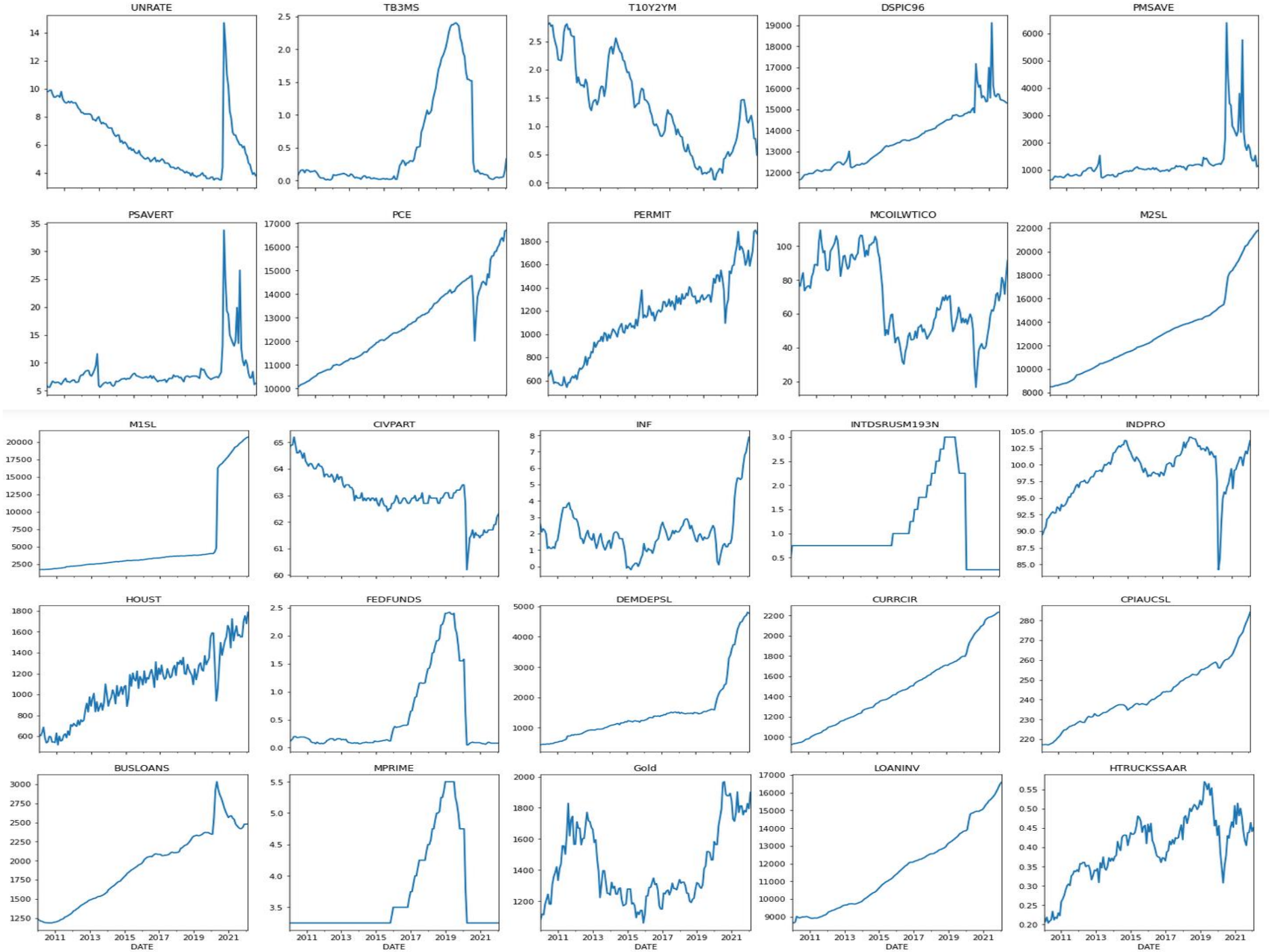
- **Trend**: The linear increasing or decreasing behavior of the series over time.
- **Seasonality**: The repeating patterns or cycles of behavior over time.





Economic Indicators

Macroeconomics	Labour Market	Real Estate Market	Credit Market	Monetary Supply	Consumer financial behaviour	Commodity Marekt
Consumer Price Index	Unemployment rate	Housing starts	Intrest rate	M1	Real disposable personal income	Oil price
Industrial Production Index	Labour force participation rate	New private housing building permits	Treasure Bill	M2	Personal saving	Motor vehicle retail sales
Inflation rate			Comercial & industrial loans	Funds Rate	Demand deposits	Gold price
			Bank prime loan rate	Currency in cerculation	Personal savings rate	
			Bank credit	10 Year treasury	Personal consumption expenditures	


UNRATE
TB3MS
T10Y2YM
DSPIC96
PMSAVE
PSAVERT
PCE
PERMIT
MCOILWTICO
M2SL
M1SL
CIVPART
INF
INTDSRUSM193N
INDPRO
HOUST
FEDFUNDS
DEMDEPSL
CURRCIR
CPIAUCSL
BUSLOANS
MPRIME
Gold
LOANINV
HTRUCKSSAAR
DATE

Data Preparation

D

Normalization

C

Feature Selection

A

Test Stationarity

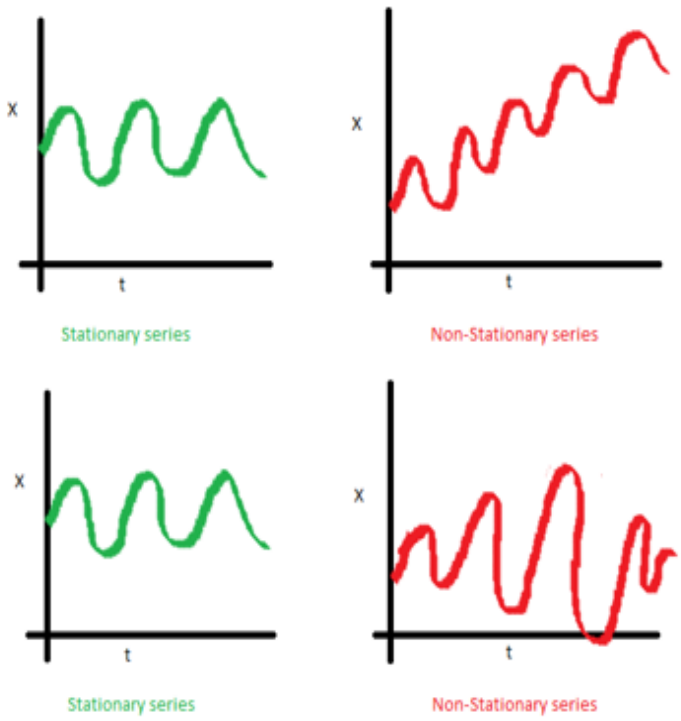
B

Achieve Stationarity

Stationarity

- **Stationary Series:** where the values of the series is not a function of time.
  - Mean, variance are constant over time.

- **How to test stationarity?**



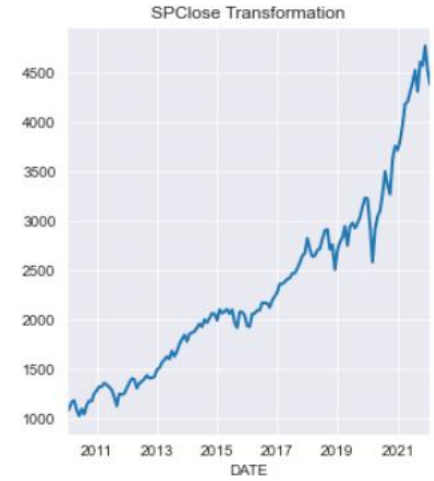
- **Summary Statistics:** Split the time series into two contiguous sequences, then calculate the mean and variance of each group.

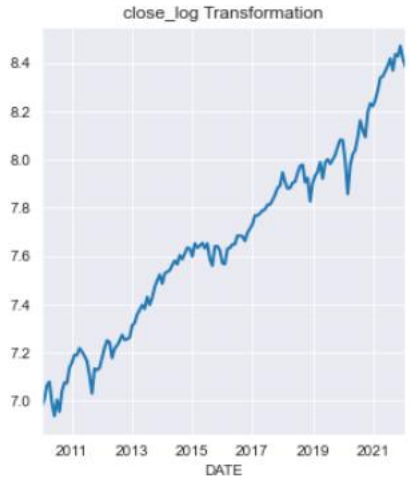
```
mean1=1579.441096, mean2=3011.868767
variance1=120129.438974, variance2=549534.952293
```

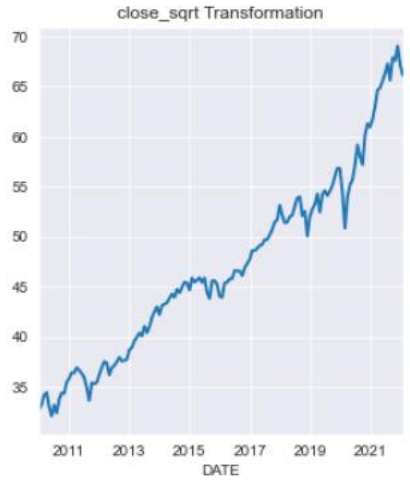
- **Statistical Tests:** ex. ADF test. If p-value > 0.05: Fail to reject the null hypothesis (H0) : the data is non-stationary.

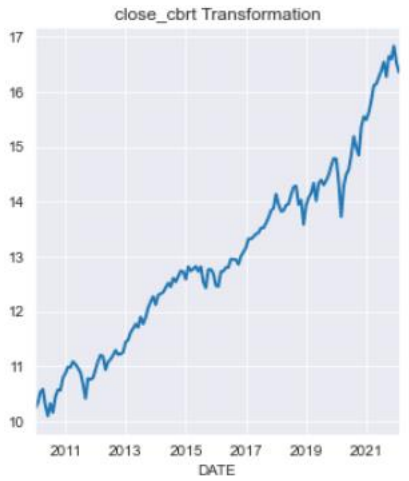
```
ADF Statistic: 1.786475
p-value: 0.998322
Critical Values:
        1%: -3.477
        5%: -2.882
        10%: -2.578
```

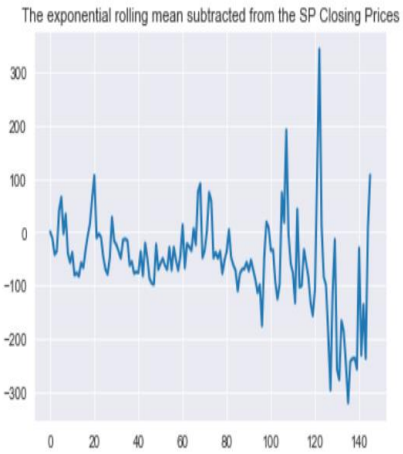
Achieving Stationarity

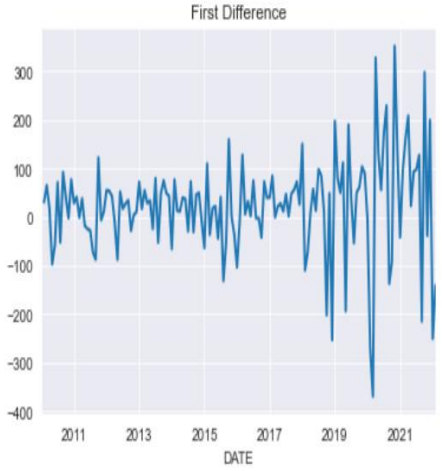


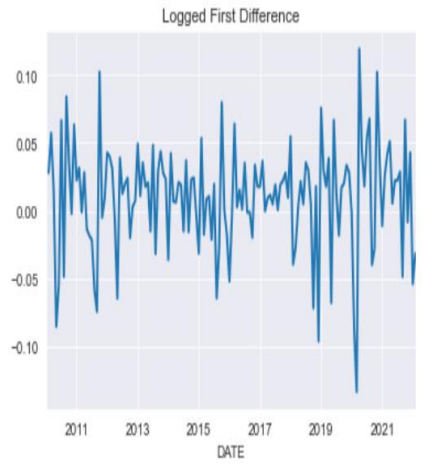


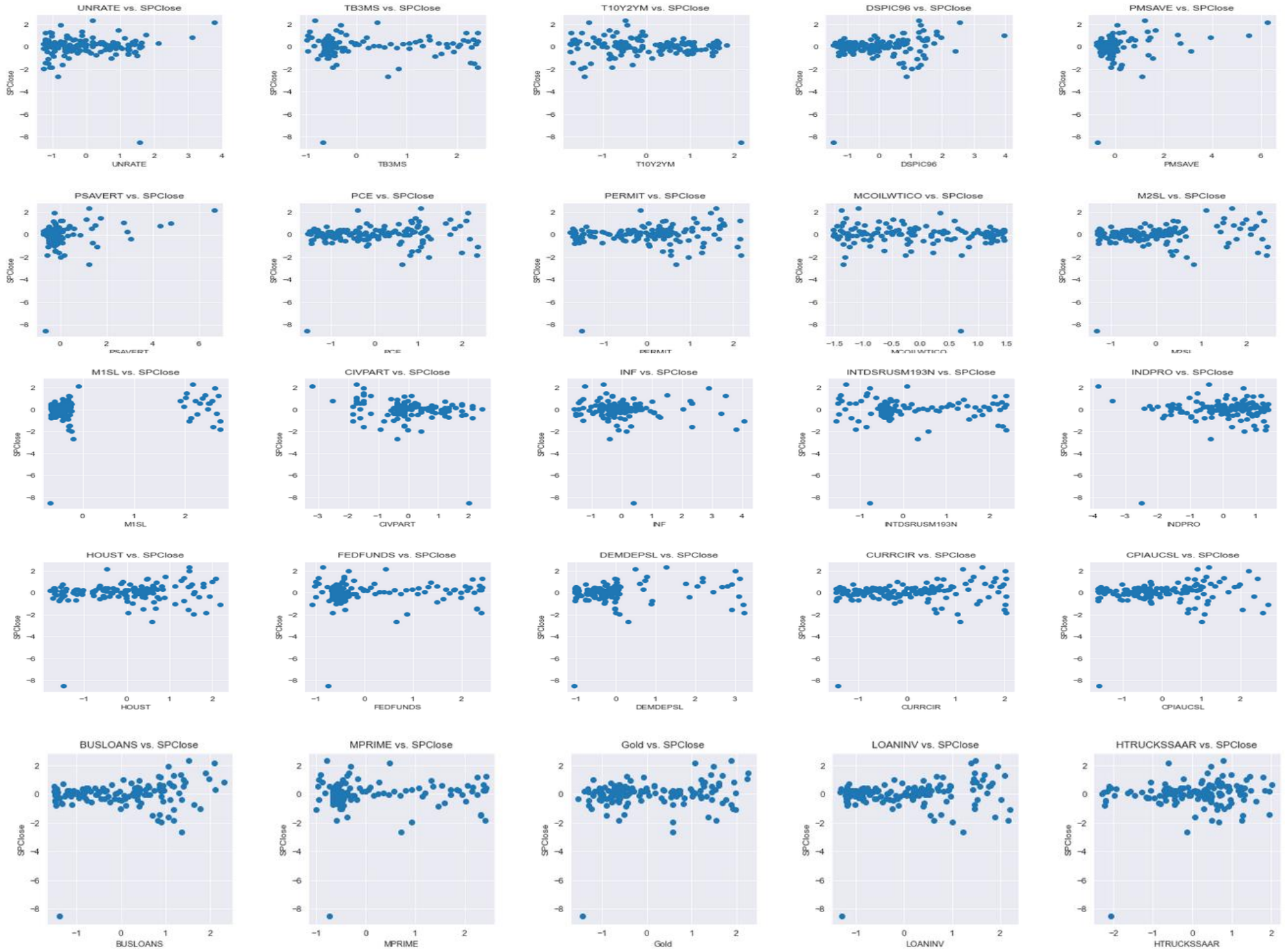










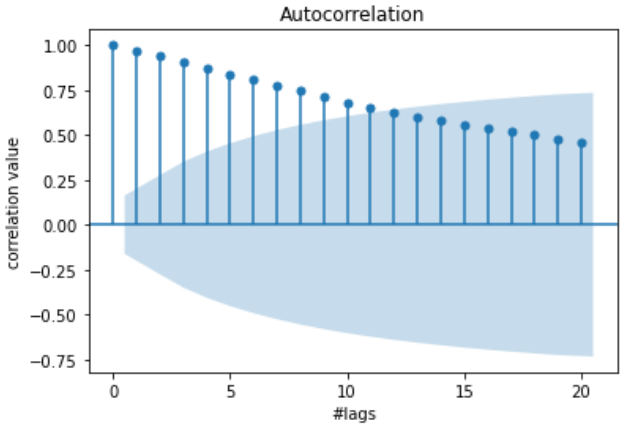

UNRATE vs. SPClose
TB3MS vs. SPClose
T10Y2YM vs. SPClose
DSPIC96 vs. SPClose
PMSAVE vs. SPClose
PSAVERT vs. SPClose
PCE vs. SPClose
PERMIT vs. SPClose
MCOILWTICO vs. SPClose
M2SL vs. SPClose
M1SL vs. SPClose
CIVPART vs. SPClose
INF vs. SPClose
INTDSRUSM193N vs. SPClose
INDPRO vs. SPClose
HOUST vs. SPClose
FEDFUNDS vs. SPClose
DEMDEPSL vs. SPClose
CURRCIR vs. SPClose
CPIAUCSL vs. SPClose
BUSLOANS vs. SPClose
MPRIME vs. SPClose
Gold vs. SPClose
LOANINV vs. SPClose
HTRUCKSSAAR vs. SPClose
SPClose
UNRATE
TB3MS
T10Y2YM
DSPIC96
PMSAVE
PSAVERT
PCE
PERMIT
MCOILWTICO
M2SL
M1SL
CIVPART
INF
INTDSRUSM193N
INDPRO
HOUST
FEDFUNDS
DEMDEPSL
CURRCIR
CPIAUCSL
BUSLOANS
MPRIME
Gold
LOANINV
HTRUCKSSAAR

Random Walk?

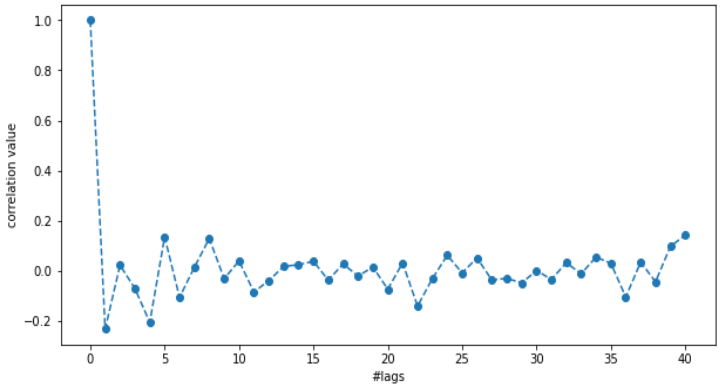
$$X(t) = X(t-1) + Er(t)$$

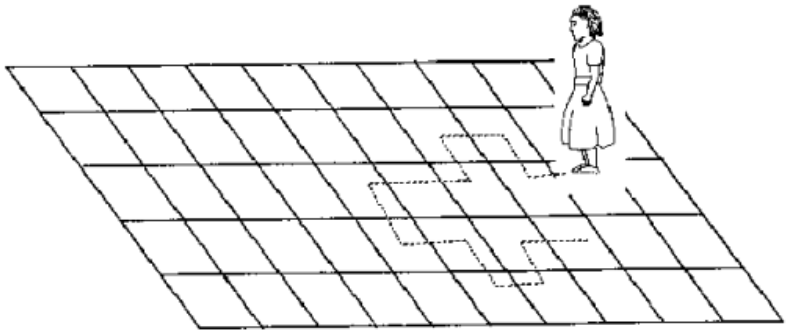
The current observation is a random step from the previous observation.

No obviously learnable structure in the data.



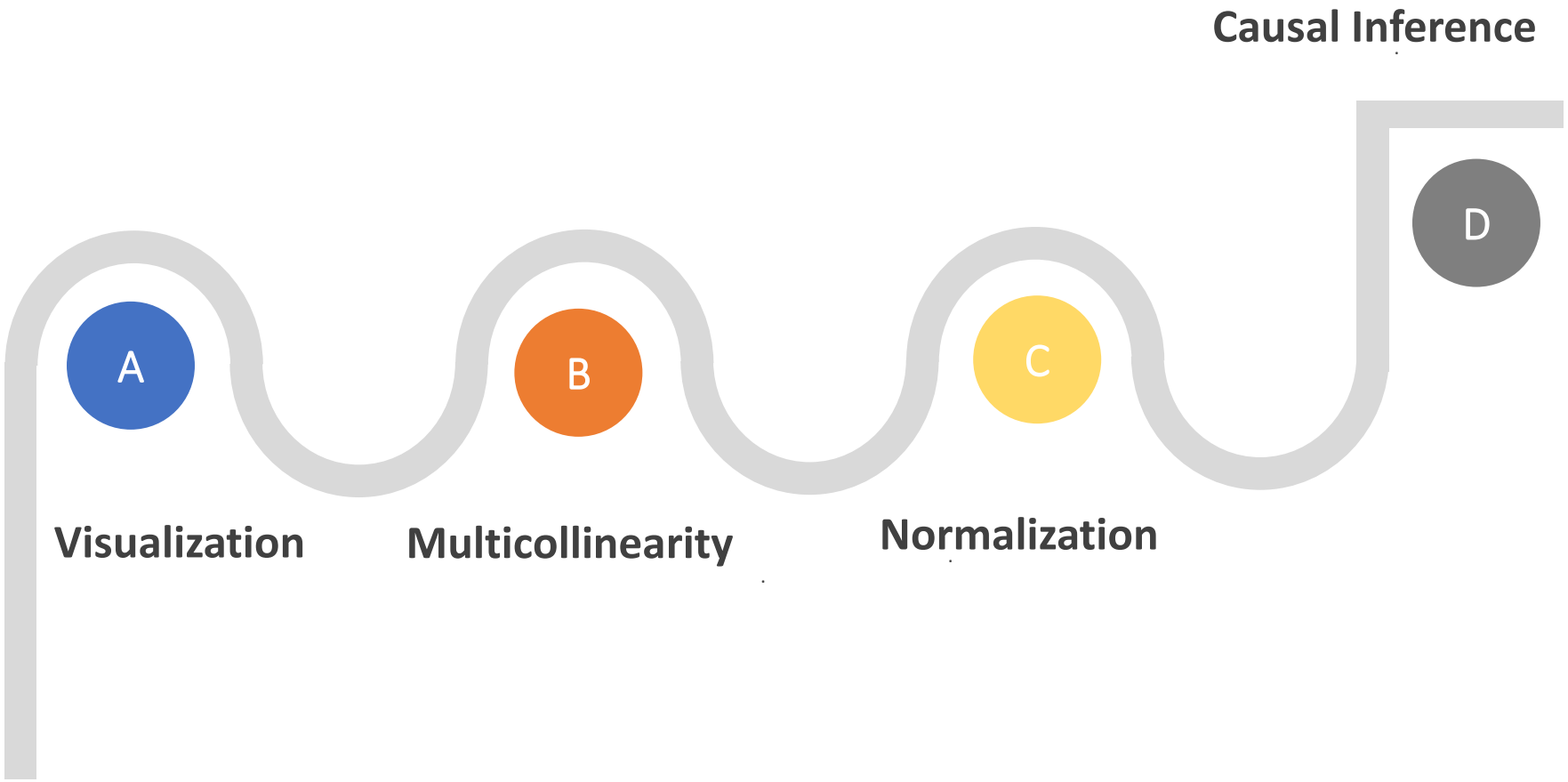
naive forecast



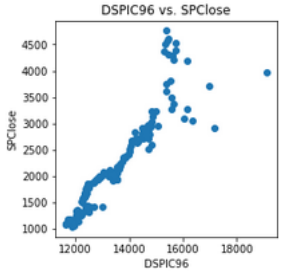


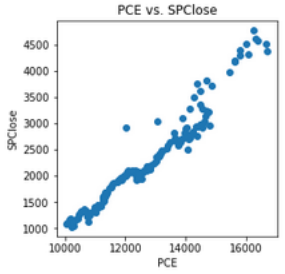
(Source: http://scifun.chem.wisc.edu/WOP/RandomWalk.html)

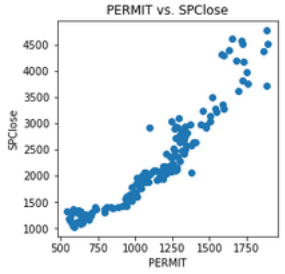
Features Selection

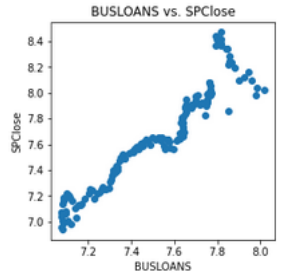


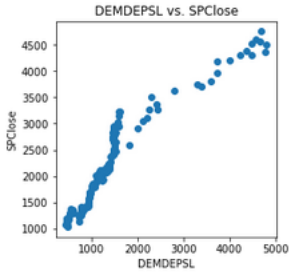
Visualization

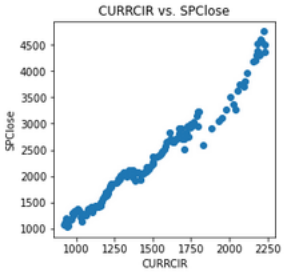


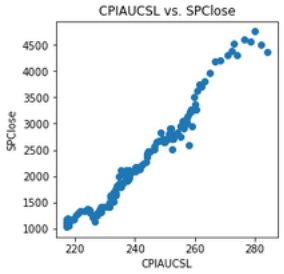


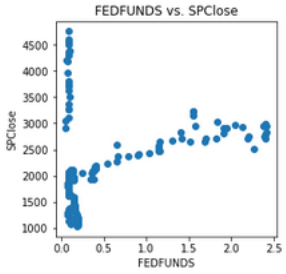


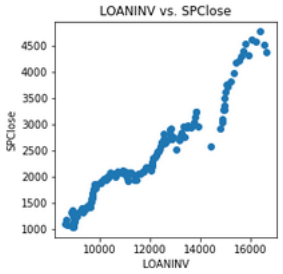


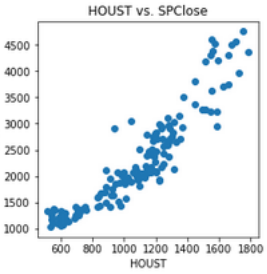


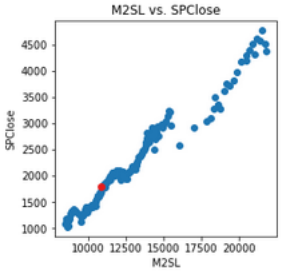


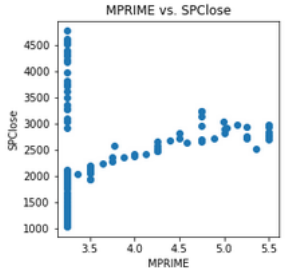


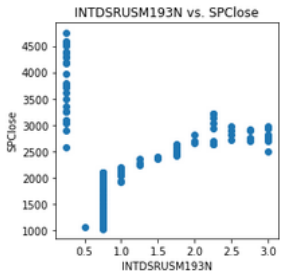


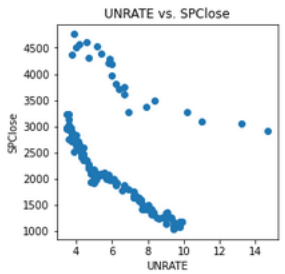


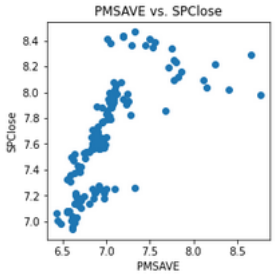


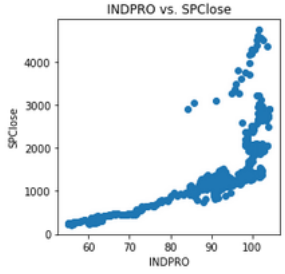












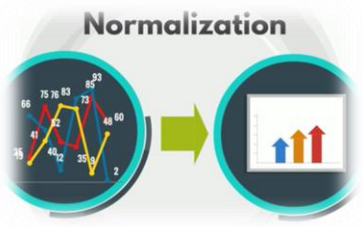
Multicollinearity

when there's correlation between independent variables in a model

VIF test > 10 indicates high correlation.

feature	VIF
UNRATE	255.892800
DSPIC96	33550.079889
PMSAVE	326.433194
PCE	37148.518121
PERMIT	504.339678
M2SL	31755.056956
INTDSRUSM193N	283.485981
INDPRO	11749.370666
HOUST	347.255708
FEDFUNDS	1238.547083
DEMDEPSL	477.342405
CURRCIR	21963.251870
CPIAUCSL	44514.104690
BUSLOANS	2098.610756
MPRIME	24297.726718
LOANINV	14677.348644

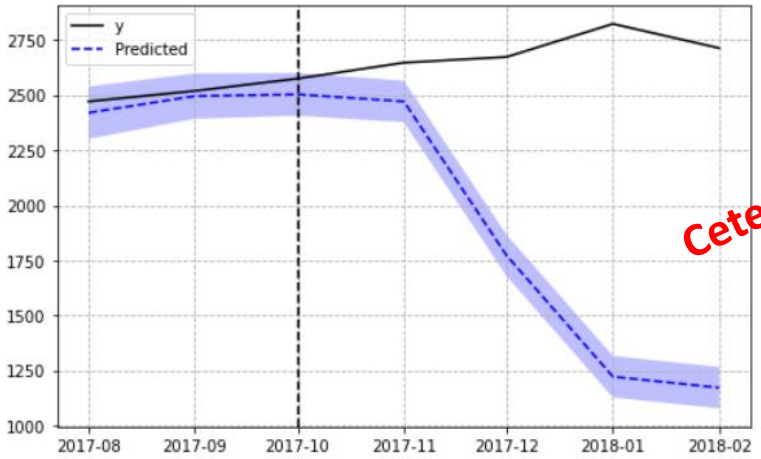
feature	VIF	Name
PMSAVE	9.372721	personal saving
FEDFUNDS	2.175213	Funds Rate
UNRATE	16.165065	Unemployment rate
CURRCIR	17.799241	currency in cerculation
MCOILWTICO	17.253308	Oil Price



Causal Inference

A Bayesian structural time-series models

SPClose vs FEDFUNDS



Note: The first 1 observations were removed due to approximate diffuse initialization.

Ceteris paribus

```
Posterior Inference {Causal Impact}
                          Average            Cumulative
Actual                    2714.71            10858.83
Prediction (s.d.)         1659.4 (30.91)     6637.61 (123.65)
95% CI                    [1599.89, 1721.06] [6399.56, 6884.26]

Absolute effect (s.d.)    1055.3 (30.91)     4221.22 (123.65)
95% CI                    [993.64, 1114.82]  [3974.57, 4459.27]

Relative effect (s.d.)    63.6% (1.86%)      63.6% (1.86%)
95% CI                    [59.88%, 67.18%]   [59.88%, 67.18%]

Posterior tail-area probability p: 0.0
Posterior prob. of a causal effect: 100.0%
```

Statistically Significant:

p-value <= 5%

Causal Inference

B Granger Causality in Time Series

SPClose vs FEDFUNDS

```
res = grangercausalitytests(df_transformed[['FEDFUNDS', 'SPClose']], maxlag=4)
```

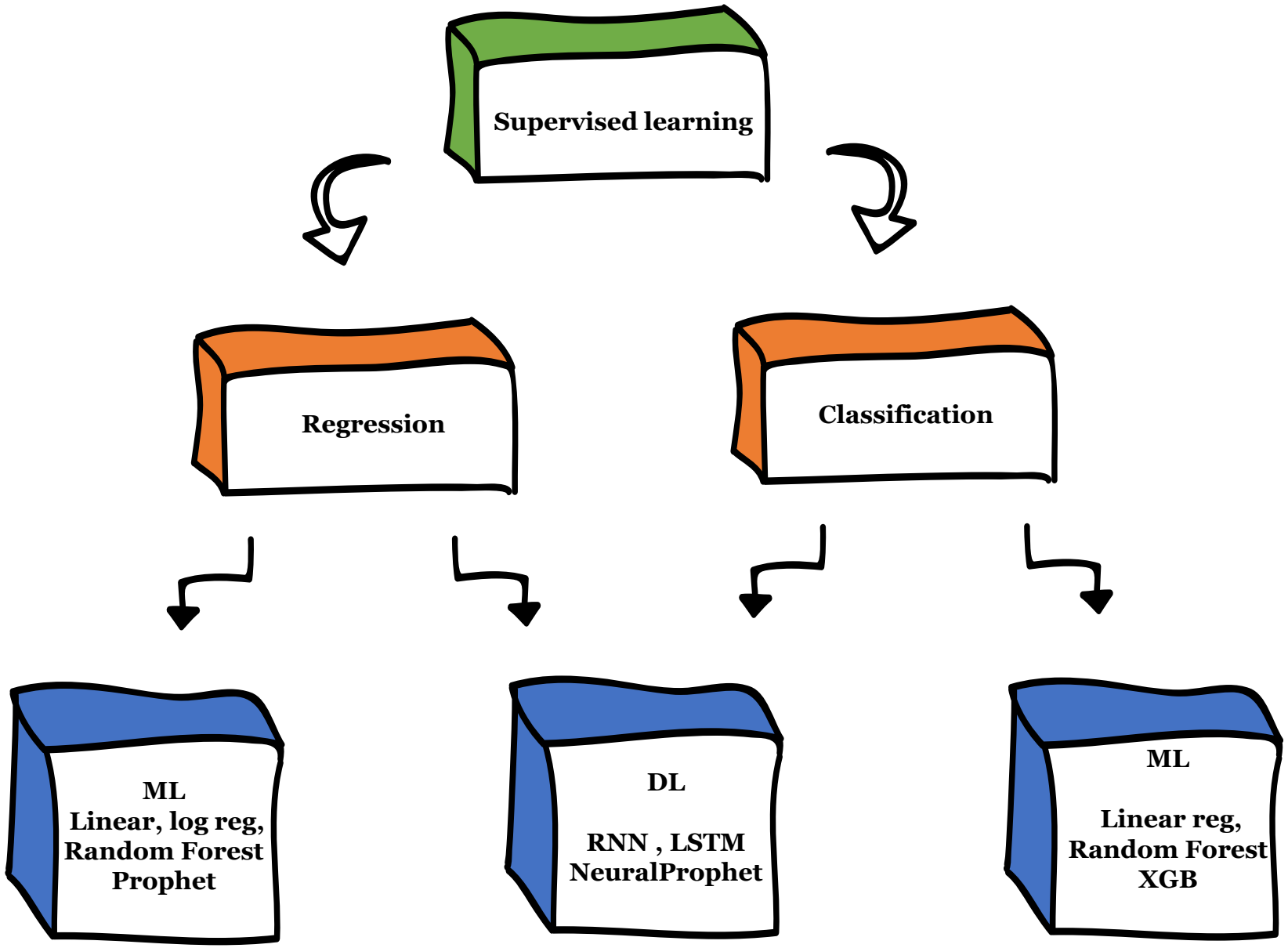
```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=12.5633 , p=0.0005  , df_denom=141, df_num=1
ssr based chi2 test:   chi2=12.8306 , p=0.0003  , df=1
likelihood ratio test: chi2=12.2908 , p=0.0005  , df=1
parameter F test:         F=12.5633 , p=0.0005  , df_denom=141, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=5.0476  , p=0.0077  , df_denom=138, df_num=2
ssr based chi2 test:   chi2=10.4610 , p=0.0054  , df=2
likelihood ratio test: chi2=10.0961 , p=0.0064  , df=2
parameter F test:         F=5.0476  , p=0.0077  , df_denom=138, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=3.7379  , p=0.0128  , df_denom=135, df_num=3
ssr based chi2 test:   chi2=11.7951 , p=0.0081  , df=3
likelihood ratio test: chi2=11.3308 , p=0.0101  , df=3
parameter F test:         F=3.7379  , p=0.0128  , df_denom=135, df_num=3

Granger Causality
number of lags (no zero) 4
ssr based F test:         F=4.0216  , p=0.0041  , df_denom=132, df_num=4
ssr based chi2 test:   chi2=17.1833 , p=0.0018  , df=4
likelihood ratio test: chi2=16.2142 , p=0.0027  , df=4
parameter F test:         F=4.0216  , p=0.0041  , df_denom=132, df_num=4
```

Stationary Data

Statistically Significant:

p-value <= 5%

Modelling


Supervised learning
Regression
Classification
ML
Linear, log reg,
Random Forest
Prophet
DL
RNN , LSTM
NeuralProphet
ML
Linear reg,
Random Forest
XGB

Linear Regression Model

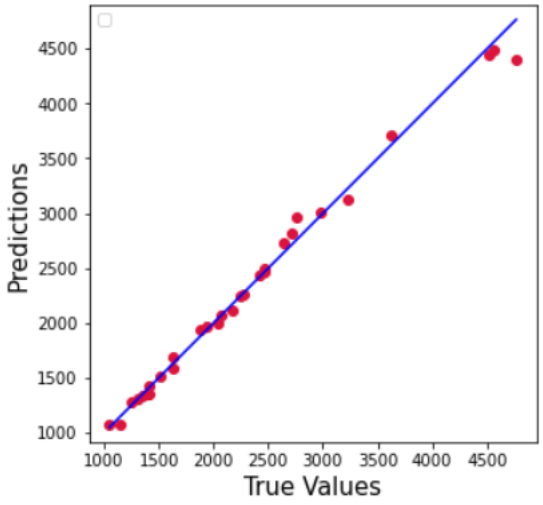
SPClose = β**0** + β**1**UNRATE+ β**2** FEDFUNDS + β**3** MCOILWTICO+ β**54**CURRCIR+ β**5** PMSAVE+ ε

OLS Regression Results

Dep. Variable:	SPClose	R-squared:	0.983
Model:	OLS	Adj. R-squared:	0.982
Method:	Least Squares	F-statistic:	1582.
Date:	Sat, 07 May 2022	Prob (F-statistic):	2.96e-121
Time:	12:49:36	Log-Likelihood:	-907.87
No. Observations:	146	AIC:	1828.
Df Residuals:	140	BIC:	1846.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1897.5719	120.014	-15.811	0.000	-2134.847	-1660.297
PMSAVE	-0.0889	0.025	-3.559	0.001	-0.138	-0.039
FEDFUNDS	-55.9169	18.464	-3.028	0.003	-92.421	-19.413
CURRCIR	2.7390	0.058	47.406	0.000	2.625	2.853
UNRATE	-0.9920	9.616	-0.103	0.918	-20.004	18.020
MCOILWTICO	4.6879	0.609	7.702	0.000	3.484	5.891

Omnibus:	4.559	Durbin-Watson:	0.848
Prob(Omnibus):	0.102	Jarque-Bera (JB):	6.148
Skew:	-0.013	Prob(JB):	0.0462
Kurtosis:	4.005	Cond. No.	2.42e+04



MSE: 0.020044598666524402
RMSE: 0.010022299333262201
r2 score for Random Forest model is 0.9829019374032483

Linear Regression Model

all the economic indicators

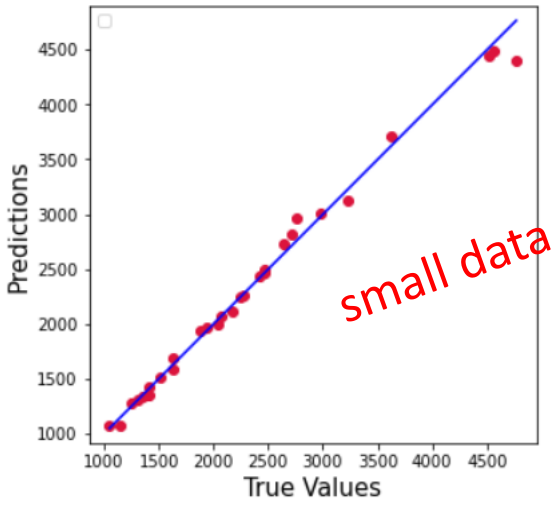
SPClose = β**0** + β**1**UNRATE+ β**2** FEDFUNDS + β**3** MCOILWTICO+ β**54**CURRCIR+ β**5** PMSAVE+……. +ε

OLS Regression Results

Dep. Variable:	SPClose	R-squared:	0.991
Model:	OLS	Adj. R-squared:	0.990
Method:	Least Squares	F-statistic:	856.5
Date:	Sat, 07 May 2022	Prob (F-statistic):	1.46e-122
Time:	13:07:46	Log-Likelihood:	134.61
No. Observations:	146	AIC:	-235.2
Df Residuals:	129	BIC:	-184.5
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.073e-14	0.008	1.27e-12	1.000	-0.017	0.017
UNRATE	0.1607	0.057	2.823	0.006	0.048	0.273
DSPIC96	-1.6341	0.713	-2.290	0.024	-3.046	-0.222
PMSAVE	0.8257	0.382	2.161	0.033	0.070	1.582
PCE	2.0709	0.769	2.692	0.008	0.549	3.593
PERMIT	0.0449	0.056	0.798	0.426	-0.066	0.156
M2SL	-1.7241	0.414	-4.164	0.000	-2.543	-0.905
INTDSRUSM193N	0.2334	0.095	2.464	0.015	0.046	0.421
INDPRO	0.0472	0.040	1.174	0.243	-0.032	0.127
HOUST	0.0091	0.045	0.204	0.838	-0.079	0.097
FEDFUNDS	-0.5826	0.257	-2.269	0.025	-1.091	-0.075
DEMDEPSL	0.4247	0.106	3.992	0.000	0.214	0.635
CURRCIR	1.9326	0.351	5.511	0.000	1.239	2.626
CPIAUCSL	-0.7889	0.287	-2.748	0.007	-1.357	-0.221
BUSLOANS	-0.1569	0.101	-1.557	0.122	-0.356	0.042
MPRIME	0.2935	0.290	1.011	0.314	-0.281	0.868
LOANINV	0.4206	0.200	2.104	0.037	0.025	0.816

Omnibus:	8.564	Durbin-Watson:	1.156
Prob(Omnibus):	0.014	Jarque-Bera (JB):	13.523
Skew:	-0.260	Prob(JB):	0.00116
Kurtosis:	4.397	Cond. No.	450.



MSE: 0.015547185446874464
RMSE: 0.007773592723437232
r2 score for Random Forest model is 0.9867382353522541