지표를 이용한 삼성전자 주가 예측 및 상관관계분석

주식4조
김윤지, 남슬아, 박준배, 조해원

[목 차]

1. 실습 목표 

- 무위험 이자율(국채금리), 원자재(금, 원유3종), 환율(달러, 엔), 주요 주가지수(나스닥, 항셍지수, 코스피, 니케이 225, 필라델피아 반도체지수)와 삼성전자 주가 간의 상관관계를 분석한다.

- 분석 결과 최종 선택된 변수에 다양한 모형(회귀 모형, 시계열 모형)을 적용하여 최종 예측 모형을 도출한다.

- 도출된 모형에 2022년 1월~4월까지의 데이터셋을 적용하여 성능을 검증한다.

2. 실습 계획

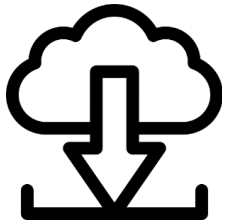
- 데이터 수집(04.21)
- 데이터 전처리(04.22~04.25)
- 데이터 확인, 변수 선택(04.25~04.26)
- 분석 및 결과 해석(04.27~05.02)
- 최종 정리(05.03)

3. 데이터 수집(04.21)

- 기간: 2019년 ~ 2021년 3개년 자료 수집

	항목	수집 이유
주요 주가지수	- 나스닥 - 코스피 - 항셍 - 니케이225 - 필라델피아반도체지수	- 한국은 무역의존도(전체수출입/GDP)가 63.51%(통계청, 2019)를 차지하는 국가로, 국외 주가지수도 주가에 영향을 미칠 것으로 사료됨. - 삼성전자의 경우 반도체 사업의 비중이 크기 때문에 반도체지수도 추가로 수집
무위험 이자율	- 국고채 3년물 - 미국채 10년물	- T-Note 중에서 특히 10년물은 은행같은 큰 기관들이 선호하기 때문에 수집
원자재	- 금 선물 - 유가 3종(두바이유, 브렌트유, WTI)	- 안전자산인 금, 대표적인 증시변수 유가3종과의 관계를 파악하기 위해 수집
환율	- USD/KRW - JPY/KRW	- 환율은 여러 선행 연구에서 영향력이 검증되었음. 또한 한국 주식 시장은 외국인투자자의 영향력이 크기 때문에 환율이 주요 변수일 것으로 사료됨. - 위안화는 재정환율로만 구해지기 때문에 제외.

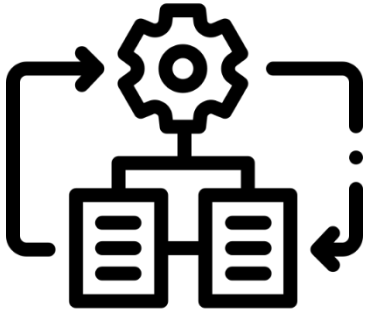
3. 데이터 수집(04.21)



- 데이터 수집 경로

Yahoo Finance	한국은행 100대 지표	한국석유공사	Finance Data Reader
- 주요 주가지수 - 환율(USD,JPY) - 금 선물	- 한국 국채(3년물)	- 유가 3종(두바이유, 브렌트유, WTI)	- 미국 국채(10년물)

4. 데이터 전처리(04.22~04.25)

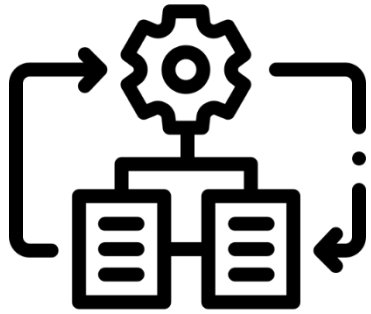


(1) 데이터타입 가공 – 각 지표의 “date” 타입을 일치

(2) 결측치 처리 (보간법 사용)

(3) 데이터 프레임 병합

4. 데이터 전처리(04.22~04.25)



(1) 데이터타입 가공 – 각 지표의 “date” 타입을 ‘datetime’으로 일치시켜준다.

[주요 주가지수 데이터]

Date	SAMSUNG	NDAQ	HSI	KOSPI	N225	SOX
2018-12-24	NaN	72.924530	25651.380859	NaN	NaN	1069.390015
2018-12-25	NaN	NaN	NaN	NaN	19155.740234	NaN
2018-12-26	34644.144531	75.318916	NaN	2028.010010	19327.060547	1131.099976
2018-12-27	34875.742188	76.668159	25478.880859	2028.439941	20077.619141	1139.489990
2018-12-28	35286.039062	76.516144	25504.199219	2041.040039	20014.769531	1147.369995
2018-12-31	NaN	77.504288	25845.699219	NaN	NaN	1155.170044
2019-01-02	35331.636719	76.962700	25130.349609	2010.000000	NaN	1165.300049
2019-01-03	34283.078125	74.891365	25064.359375	1993.699951	NaN	1096.030029
2019-01-04	34146.312500	77.865356	25626.029297	2010.250000	19561.960938	1143.959961
2019-01-07	35331.636719	75.832008	25835.699219	2037.099976	20038.970703	1166.239990

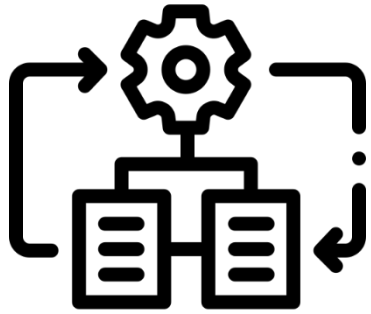
dtype='datetime64[ns]

[유가 3종 데이터]

	Date	Dubai	Brent	WTI
0	01월?02일	51.86	54.91	46.54
1	01월?03일	53.2	55.95	47.09
2	01월?04일	55.59	57.06	47.96
3	01월?07일	56.79	57.33	48.52
4	01월?08일	56.18	58.72	49.78
5	01월?09일	58.07	61.44	52.36
6	01월?10일	59.6	61.68	52.59
7	01월?11일	61.16	60.48	51.59
8	01월?14일	58.92	58.99	50.51
9	01월?15일	58.63	60.64	52.11

object

4. 데이터 전처리(04.22~04.25)



(1) 데이터타입 가공 – 각 지표의 “date” 타입을 ‘datetime’으로 일치시켜준다.

- 타 데이터는 to_datetime 함수만으로 변환이 용이했으나 유가 데이터의 경우 ‘year’이 원 데이터에 존재하지 않아 별도의 함수를 사용하여 자료 변경 후 to_datetime 적용.

[유가 3종 데이터]

	Date	Dubai	Brent	WTI
0	01월?02일	51.86	54.91	46.54
1	01월?03일	53.2	55.95	47.09
2	01월?04일	55.59	57.06	47.96
3	01월?07일	56.79	57.33	48.52
4	01월?08일	56.18	58.72	49.78
5	01월?09일	58.07	61.44	52.36
6	01월?10일	59.6	61.68	52.59
7	01월?11일	61.16	60.48	51.59
8	01월?14일	58.92	58.99	50.51
9	01월?15일	58.63	60.64	52.11

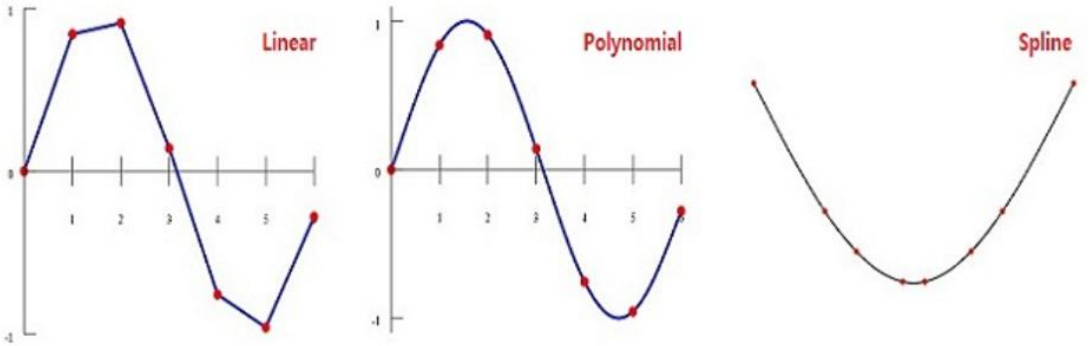
object

```
oil1['Date']=oil1['Date'].apply(lambda x: x.replace('?',''))
oil1['Date']=oil1['Date'].apply(lambda x: x.replace('월','-'))
oil1['Date']=oil1['Date'].apply(lambda x: x.replace('일',''))
oil1['Date']=oil1['Date'].apply(lambda x: '2019-'+x)
oil1['Date']=pd.to_datetime(oil1['Date'],infer_datetime_format=True)
```

4. 데이터 전처리(04.22~04.25)

(2) 결측치 처리

- 보간법: 실변수 x의 함수 f(x)의 모양은 미지이나, 어떤 간격을 가지는 2개 이상인 변수의 값에 대한 함수 값이 알려져 있을 경우, 그 사이의 임의의 x에 대한 함수값을 추정하는 방법
선형 보간법. 다항 보간법. 스플라인 보간법. 지수 보간법 등이 있다.

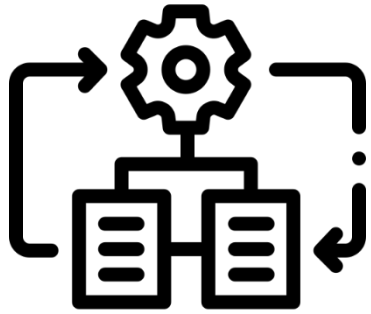


- 선형 보간법(linear interporate)을 사용하여 결측치를 대체

$$f(x) = f(x0) + \frac{f(x1) - f(x0)}{x1 - x0}(x - x0)$$

이미지 출처: https://m.blog.naver.com/PostList.naver?blogId=ceoyangsj

4. 데이터 전처리(04.22~04.25)



(2) 결측치 처리

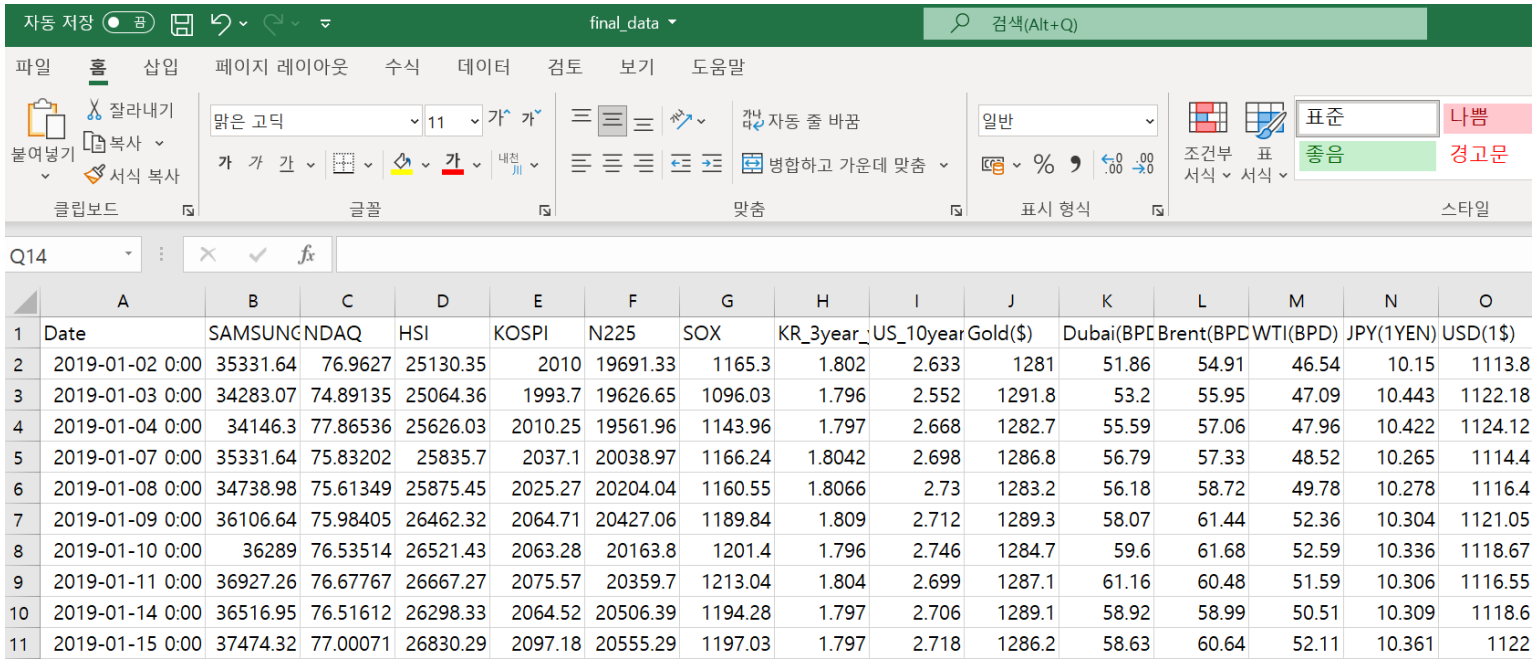
- 결측치 날짜 기준 전일과 후일 데이터를 기준으로 선형 보간법을 적용함
- 니케이 지수의 경우 2019.01.02 , 01.03일이 모두 결측치로 전일 값을 적용할 수 없었기 때문에 2018년 12월의 데이터를 사용하여 선형 보간법 적용 후 2018년의 데이터 삭제.

Date	SAMSUNG	NDAQ	HSI	KOSPI	N225	SOX
2018-12-24	NaN	72.924538	25651.380859	NaN	NaN	1069.390015
2018-12-25	NaN	NaN	NaN	NaN	19155.740234	NaN
2018-12-26	34644.140625	75.318924	NaN	2028.010010	19327.060547	1131.099976
2018-12-27	34875.738281	76.668159	25478.880859	2028.439941	20077.619141	1139.489990
2018-12-28	35286.039062	76.516129	25504.199219	2041.040039	20014.769531	1147.369995
2018-12-31	NaN	77.504280	25845.699219	NaN	NaN	1155.170044
2019-01-02	35331.628906	76.962700	25130.349609	2010.000000	NaN	1165.300049
2019-01-03	34283.074219	74.891365	25064.359375	1993.699951	NaN	1096.030029
2019-01-04	34146.316406	77.865349	25626.029297	2010.250000	19561.960938	1143.959961
2019-01-07	35331.628906	75.832001	25835.699219	2037.099976	20038.970703	1166.239990

4. 데이터 전처리(04.22~04.25)

(3) 데이터 병합

- 데이터 타입과 결측치 처리가 완료된 주가지수, 유가3종, 금 선물, 환율, 무위험 이자율 데이터를 1개 파일로 병합하여 csv 저장



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Date	SAMSUNG	NDAQ	HSI	KOSPI	N225	SOX	KR_3year_	US_10year	Gold($)	Dubai(BPD	Brent(BPD	WTI(BPD)	JPY(1YEN)	USD(1$)
2	2019-01-02 0:00	35331.64	76.9627	25130.35	2010	19691.33	1165.3	1.802	2.633	1281	51.86	54.91	46.54	10.15	1113.8
3	2019-01-03 0:00	34283.07	74.89135	25064.36	1993.7	19626.65	1096.03	1.796	2.552	1291.8	53.2	55.95	47.09	10.443	1122.18
4	2019-01-04 0:00	34146.3	77.86536	25626.03	2010.25	19561.96	1143.96	1.797	2.668	1282.7	55.59	57.06	47.96	10.422	1124.12
5	2019-01-07 0:00	35331.64	75.83202	25835.7	2037.1	20038.97	1166.24	1.8042	2.698	1286.8	56.79	57.33	48.52	10.265	1114.4
6	2019-01-08 0:00	34738.98	75.61349	25875.45	2025.27	20204.04	1160.55	1.8066	2.73	1283.2	56.18	58.72	49.78	10.278	1116.4
7	2019-01-09 0:00	36106.64	75.98405	26462.32	2064.71	20427.06	1189.84	1.809	2.712	1289.3	58.07	61.44	52.36	10.304	1121.05
8	2019-01-10 0:00	36289	76.53514	26521.43	2063.28	20163.8	1201.4	1.796	2.746	1284.7	59.6	61.68	52.59	10.336	1118.67
9	2019-01-11 0:00	36927.26	76.67767	26667.27	2075.57	20359.7	1213.04	1.804	2.699	1287.1	61.16	60.48	51.59	10.306	1116.55
10	2019-01-14 0:00	36516.95	76.51612	26298.33	2064.52	20506.39	1194.28	1.797	2.706	1289.1	58.92	58.99	50.51	10.309	1118.6
11	2019-01-15 0:00	37474.32	77.00071	26830.29	2097.18	20555.29	1197.03	1.797	2.718	1286.2	58.63	60.64	52.11	10.361	1122

5. 데이터 확인 및 최종 변수 선택 (04.25~04.26)

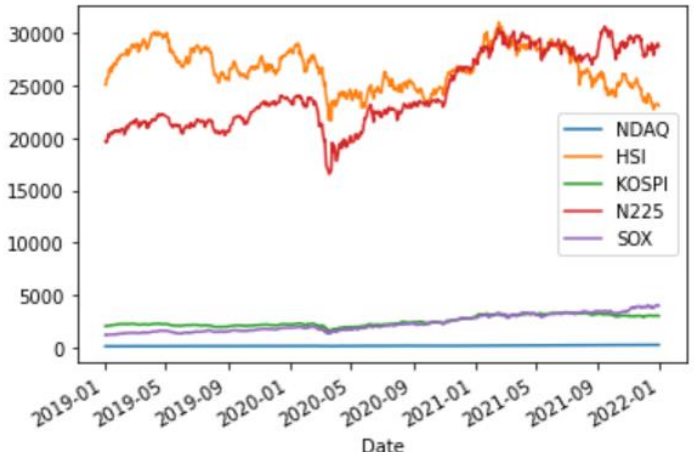
- 데이터 확인
  (1) 데이터 분포 파악
  (2) 데이터 스케일링
  (3) 상관계수 확인

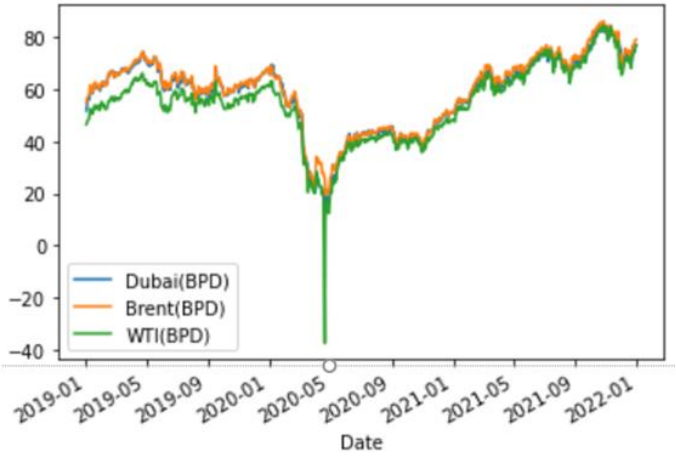
- 최종 변수 선택
  (1) Lasso
  (2) Ridge
  (3) RandomForest
  (4) PCA

5. 데이터 확인 및 최종 변수 선택 (04.25~04.26)

- 데이터 확인

  (1) 데이터 분포 확인

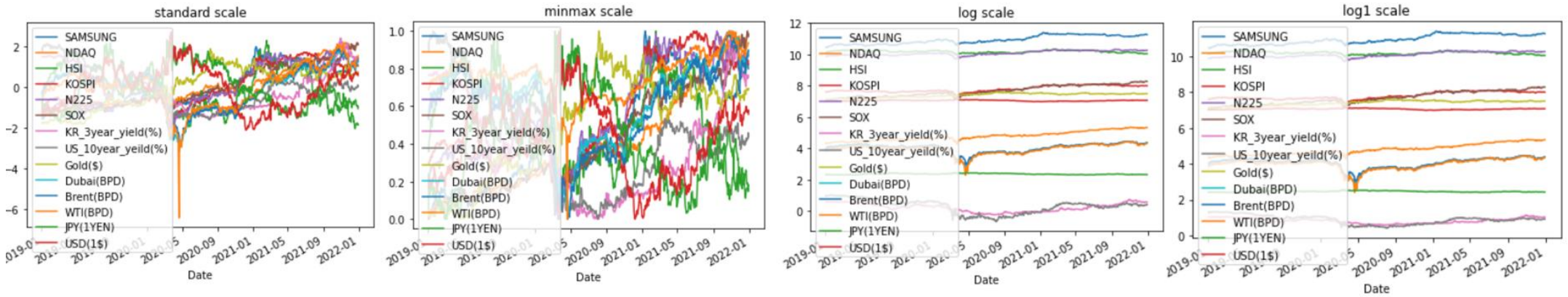




5. 데이터 확인 및 최종 변수 선택 (04.25~04.26)

- 데이터 확인

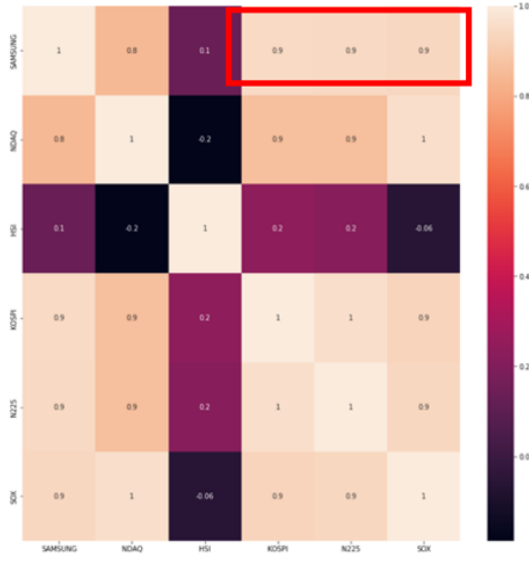
  (2) 데이터 스케일링



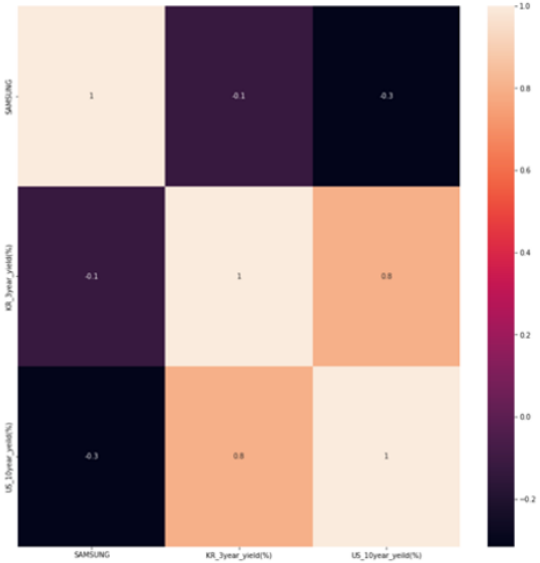
5. 데이터 확인 및 최종 변수 선택 (04.25~04.26)

- 데이터 확인

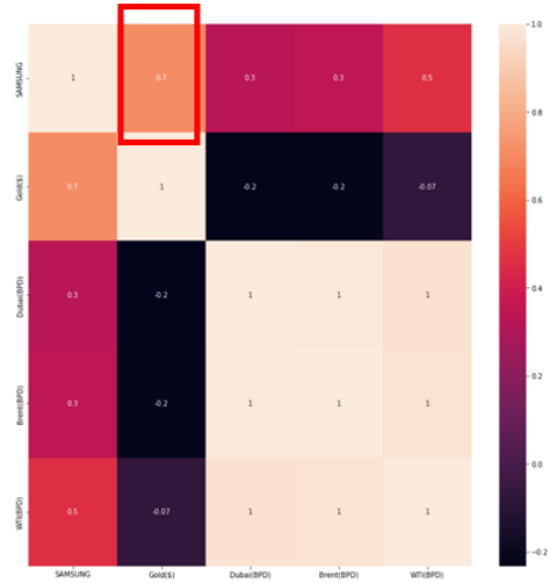
  (3) 상관계수 확인(log1 scaled 자료 기준)



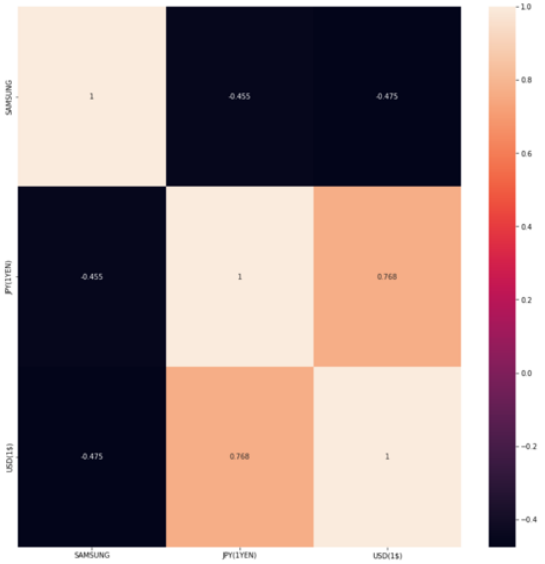
[주가지수 - 삼성전자 주가]



[무위험이자율 - 삼성전자 주가]



[원자재 - 삼성전자 주가]



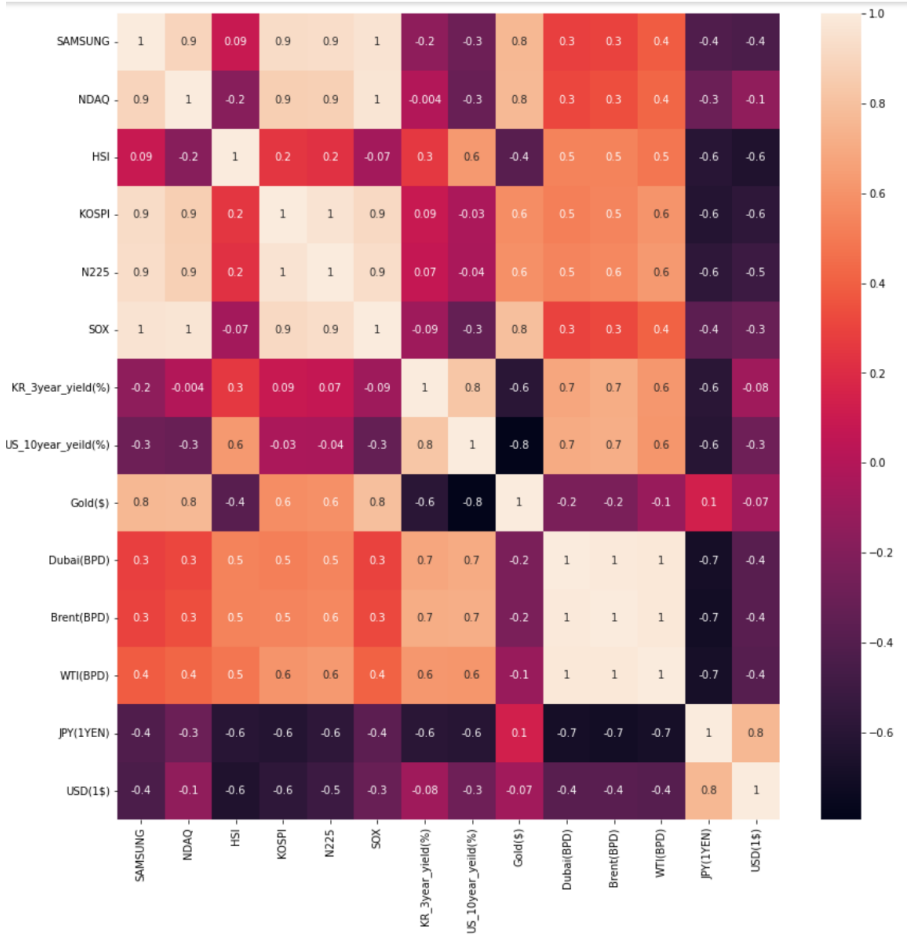
[환율 – 삼성전자 주가]

5. 데이터 확인 및 최종 변수 선택 (04.25~04.26)

- 데이터 확인

(3) 상관계수 확인(log1 scaled 자료 기준)

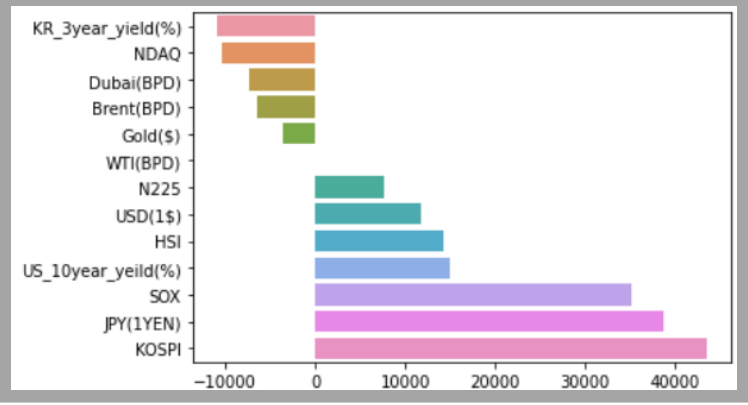
	NDAQ	HSI	KOSPI	N225	SOX	KR_3year_yield(%)	US_10year_yeild(%)
NDAQ	1.000000	-0.182301	0.864740	0.870670	0.965895	-0.008687	-0.259116
HSI	-0.182301	1.000000	0.246782	0.226753	-0.068341	0.272663	0.633434
KOSPI	0.864740	0.246782	1.000000	0.964487	0.912995	0.089553	0.021579
N225	0.870670	0.226753	0.964487	1.000000	0.930141	0.077058	0.018657
SOX	0.965895	-0.068341	0.912995	0.930141	1.000000	-0.091684	-0.268626
KR_3year_yield(%)	-0.008687	0.272663	0.089553	0.077058	-0.091684	1.000000	0.835448
US_10year_yeild(%)	-0.259116	0.633434	0.021579	0.018657	-0.268626	0.835448	1.000000
Gold($)	0.762720	-0.383327	0.579362	0.591668	0.791810	-0.600460	-0.755002
Dubai(BPD)	0.310472	0.533522	0.521323	0.539247	0.295718	0.687859	0.740459
Brent(BPD)	0.334682	0.535127	0.538393	0.551763	0.312929	0.716855	0.751930
WTI(BPD)	0.418680	0.484472	0.602743	0.623042	0.405240	0.621372	0.654722
JPY(1YEN)	-0.296513	-0.604263	-0.624788	-0.575919	-0.364864	-0.595541	-0.621986
USD(1$)	-0.146675	-0.644965	-0.575676	-0.510509	-0.312293	-0.684852	-0.299350



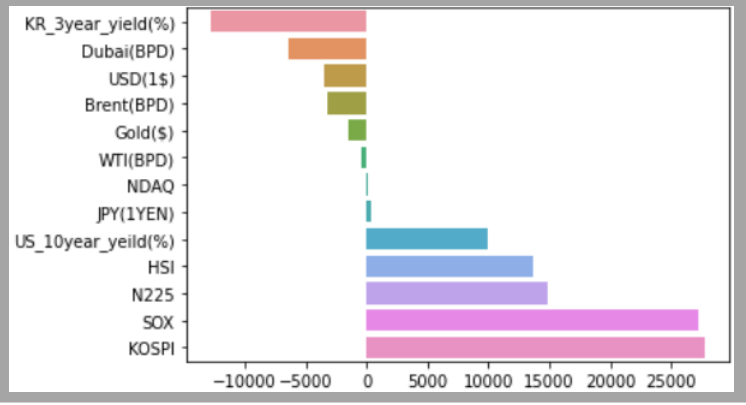
5. 데이터 확인 및 최종 변수 선택 (04.25~04.26)

- 최종변수 선택 (log1 Scaled 자료 기준)

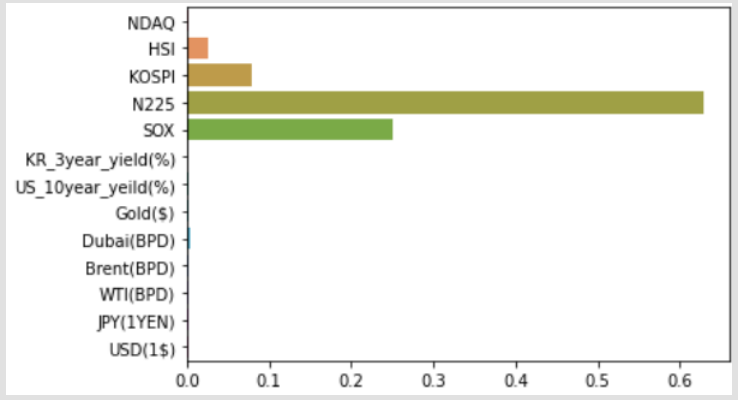
<Lasso>



<Ridge>



<Random Forest>

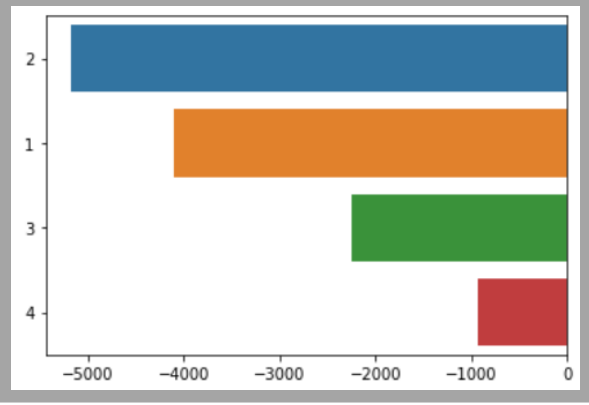


- 방법에 따라 상관관계가 높은 변수에 차이가 존재함

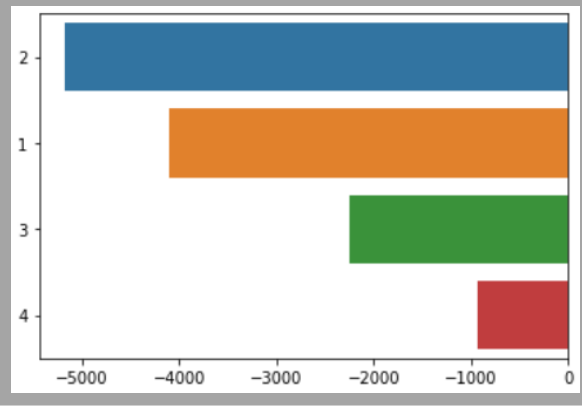
5. 데이터 확인 및 최종 변수 선택 (04.25~04.26)

- 최종변수 선택 (log1p Scaled 자료 기준)

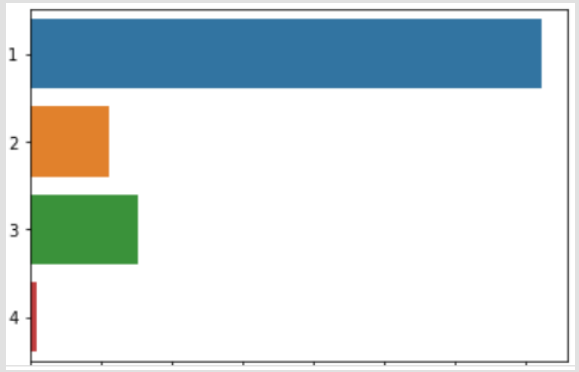
<PCA_Lasso>



<PCA_Ridge>



<PCA_Random Forest>



- 어떤 방법이 주요 feature 도출에 효과적일지 토의 중에 있음

감사합니다