[스파르타코딩클럽] 파이썬 데이터분석 첫걸음 - 4주차

매 주차 강의자료 시작에 PDF파일을 올려두었어요!

▼ PDF 파일

▼ 단축키 모음

▼ 실행(Run)

- `shift` + `enter`

▼ 자동완성

- `Tab`

[수업 목표]

1. 데이터 분석을 기획할 수 있다.
2. 분석을 위한 가설을 세울 수 있다.
3. 데이터 분석 노트를 작성할 수 있다.

[목차]

모든 토글을 열고 닫는 단축키
Windows : `Ctrl` + `alt` + `t`
Mac : `⌘` + `⌥` + `t`

---

01. 오늘 배울 것

▼ 1) 주식 데이터를 분석하기

- 실제 여러분이 관심있는 데이터는 실생활에서 만날 수 있는 재미있는 데이터겠죠? 오늘은 앞으로 무엇을 더 할 수 있을지 알아봅시다.

▼ 2) 분석한 데이터로 인사이트 얻기

- 데이터가 정답을 가져다 주는 것이 아닙니다.
- 정답을 찾는데 데이터를 활용해 봅시다.

▼ 3) 나만의 보고서 만들기

- 노트북은 공유가 가능하기도 하지만 보고용이나 공유용으로 사용하기에 쉽지 않아요
- 노트북으로 분석한 결과를 공유하는 방법을 가르쳐 드릴게요.

02. 주식 데이터 - 전처리하기

> 💡 **일상에서 활용하기 위한 첫걸음! 주식데이터를 분석해봅시다.**

▼ 4) 코드 데이터 불러오기

```
import pandas as pd

code = pd.read_csv('./data/corpgeneral.csv', header=0)
code.head(5)
```

	회사명	종목코드	업종	주요제품	상장일	결산월	대표자명	홈페이지	지역
0	JS전선	5560	절연선 및 케이블 제조업	선박선,고무선,전력선,통신선 제조	2007-11-12	12월	이익희	http://www.jscable.co.kr	충청남도
1	거북선2호	101380	NaN	운송장비(선박) 임대	2008-04-25	12월	신주선	NaN	부산광역시
2	거북선6호	114140	NaN	NaN	2009-10-01	12월	김연신	NaN	제주특별자치도
3	교보메리츠	64900	NaN	부동산 투자,운용	2002-01-30	12월	김 상 진	NaN	서울특별시
4	국제관광공사	28780	NaN	NaN	1966-03-18	12월	NaN	NaN	NaN

▼ 참고) 온라인에서 최신 코드 데이터 가져오기

```
import pandas as pd
code = pd.read_html('http://kind.krx.co.kr/corpgeneral/corpList.do?method=download', header=0)[0]
code.head(5)
```

▼ 5) 필요한 데이터 전처리

```
# 필요한 데이터 자르기
code = code[['회사명', '종목코드']]
code
```

```
# 컬럼명 바꾸기
code_result = code.rename(columns={'회사명': 'corp', '종목코드': 'code'})
code_result
```

03. 주식 데이터 - 종가 그래프 그리기

▼ 6) 종목 이름으로 원하는 종목 코드 가져오기

> 🚨 잠깐! 우리는 코스피에 있는 종목들만 분석하고 있어요.
> 코스닥 주식들을 분석하려면 .KS를 → .KQ로 바꿔주세요!

```
corp_name = "카카오"
condition = "corp=='{}'".format(corp_name)

kakao = code_result.query(condition)
kakao = kakao['code']
kakao_string = kakao.to_string(index=False)
kakao_string = kakao_string.strip()
kakao_string = kakao_string.rjust(6, '0')
kakao_code = kakao_string + '.KS' #코스피 종목코드
kakao_code
```

▼ 7) 종목 코드로 종목 데이터 가져오기

▼ **[코드스니펫] 라이브러리 가져오기**

```
conda install -c anaconda pandas-datareader
```

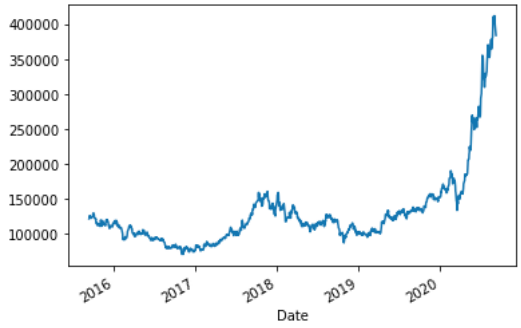
```
import pandas_datareader as pdr
kakao_stock_df = pdr.get_data_yahoo(kakao_code)
kakao_stock_df
```

Date	High	Low	Open	Close	Volume	Adj Close
2020-09-03	416000.0	404000.0	415500.0	410000.0	739215.0	410000.0
2020-09-04	404500.0	389500.0	390000.0	402000.0	1168529.0	402000.0
2020-09-07	401500.0	391000.0	401500.0	392000.0	928939.0	392000.0
2020-09-08	398500.0	380500.0	393500.0	390000.0	1020539.0	390000.0
2020-09-09	388000.0	379000.0	381500.0	384000.0	907069.0	384000.0

- High: 최고가 / Low: 최저가 / Open: 시작가 / Close: 종가
- 여러가지 데이터를 얻을 수 있습니다!

▼ 8) 종가 그래프 그리기

```
kakao_stock_df['Close'].plot()
```



04. 주식 데이터 - 상관관계란?

▼ 9) 상관 분석이란?

👉 두 데이터가 어느 정도의 상관관계를 가지고 있는지를 분석합니다.

▼ 피어슨 상관계수(수식주의!)

$$
\begin{aligned}
r &= \frac{\text{degree to which X and Y vary together}}{\text{degree to which X and Y vary separately}} \\
&= \frac{\text{covariablity of X and Y}}{\text{variability of X and Y separately}} \\
&= \frac{Cov[X,Y]}{\sqrt{Var[X]Var[Y]}} \\
&= \frac{SP_{XY}}{\sqrt{SS_X SS_Y}}
\end{aligned}
$$

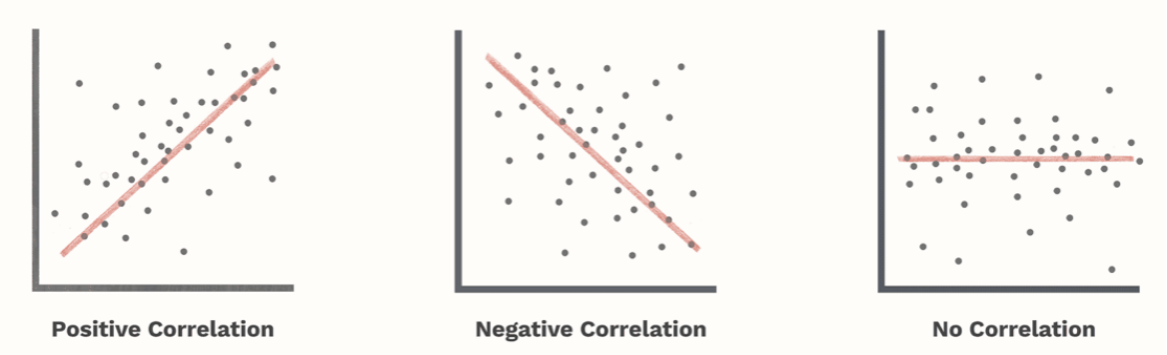
- 하지만 우리는 수학자가 아니기 때문에 파고들어서 공부하지는 않습니다.
  먼저 익히고 그 다음에 파고들어 이해합니다.
- 이론과 내용은 알고 있으면 좋지만 직접 구현하지 않기 때문입니다.

▼ 실제 머리속에 가지고 있는 컨셉

$$
r = \frac{\text{X 와 Y 가 함께 변하는 정도}}{\text{X 와 Y 가 각각 변하는 정도}}
$$

▼ 10) 상관 관계를 보는 법

- 상관 계수는 1부터 -1까지 존재 합니다.
- 기울기이기 때문이죠
- 기울기가 1이면 매우 강한 상관관계가 있습니다
  - 하지만 이 세상에서 데이터로 만날 케이스는 하나입니다. 뒤에 보여드리죠
- 0이면 상관관계를 찾을 수 없는것이죠
- -1이면? 1과 마찬가지로 강한 상관관계가 있습니다. X축이 증가하면 Y축이 감소합니다
- '보통은 0.6 ~ 0.4 사이의 값을 가지면 상관 관계가 있다'로 판단합니다



▼ 실제 여러분이 만날 데이터

	삼성전자	LG전자	카카오	NAVER	CJ	한화	현대자동차	기아자동차
삼성전자	1.000000	0.072774	0.678914	0.760835	-0.370804	-0.357820	-0.261648	0.483834
LG전자	0.072774	1.000000	-0.147646	-0.431879	0.493073	0.317671	0.371329	-0.070075
카카오	0.678914	-0.147646	1.000000	0.682700	-0.788729	-0.828159	0.014257	0.833377
NAVER	0.760835	-0.431879	0.682700	1.000000	-0.543533	-0.464186	-0.539871	0.377423
CJ	-0.370804	0.493073	-0.788729	-0.543533	1.000000	0.924626	-0.068346	-0.725753
한화	-0.357820	0.317671	-0.828159	-0.464186	0.924626	1.000000	-0.099603	-0.779787
현대자동차	-0.261648	0.371329	0.014257	-0.539871	-0.068346	-0.099603	1.000000	0.402249
기아자동차	0.483834	-0.070075	0.833377	0.377423	-0.725753	-0.779787	0.402249	1.000000

05. 주식 데이터 - 준비하기

▼ 11) 라이브러리 불러오기

```
import pandas as pd
import pandas_datareader as pdr
from datetime import datetime
```

▼ 12) 주식 데이터 받아오기 요약

- 전체 코드 정보를 받아옵니다

```
code = pd.read_csv('./data/corpgeneral.csv', header=0)
code = code[['회사명', '종목코드']]
code_result = code.rename(columns={'회사명': 'corp', '종목코드': 'code'})
```

- 코드를 받아오는 함수 만들기

👉 종목명을 입력하면 코드를 가져오는 함수를 만들어서 계속 써먹어 봅시다!

```
# 회사명으로 주식 종목 코드를 획득할 수 있도록 하는 함수
def get_code(code_result, corp_name):
    condition = "corp=='{}'".format(corp_name)
    code = code_result.query(condition)['code'].to_string(index=False)
    # 위와같이 code명을 가져오면 앞에 공백이 붙어있는 상황이 발생하여 앞뒤로 sript() 하여 공백 제거
    code = code.strip()
    code = code.rjust(6, '0')
    code = code + '.KS'
    return code
```

- 잘 되나 확인!

👉 삼성전자의 코드를 함수로 가져와 볼까요?

```
# ex) 삼성전자의 코드를 구해보겠습니다.
samsung_code = get_code(code_result, '삼성전자')
samsung_code
```

06. 주식 데이터 - 상관관계 분석

▼ 13) 주가간의 상관관계 보기

```
companies = ['삼성전자', 'LG전자', '카카오', 'NAVER', 'CJ', '한화', '현대자동차', '기아자동차']
start = datetime(2019,1,1)
end = datetime(2019,12,31)
stocks_of_companies = pd.DataFrame({'Date': pd.date_range(start=start, end=end)})
stocks_of_companies
```

```
for company in companies:
    company_code = get_code(code_result, company)
    stock_df = pdr.get_data_yahoo(company_code, start, end)
    stocks_of_companies = stocks_of_companies.join(pd.DataFrame(stock_df['Close']).rename(columns={'Close':company}), on='Date')
stocks_of_companies.tail(5)
```

```
corr_data = stocks_of_companies.corr()
```

	삼성전자	LG전자	카카오	NAVER	CJ	한화	현대자동차	기아자동차
삼성전자	1.000000	0.072774	0.678914	0.760835	-0.370804	-0.357820	-0.261648	0.483834
LG전자	0.072774	1.000000	-0.147646	-0.431879	0.493073	0.317671	0.371329	-0.070075
카카오	0.678914	-0.147646	1.000000	0.682700	-0.788729	-0.828159	0.014257	0.833377
NAVER	0.760835	-0.431879	0.682700	1.000000	-0.543533	-0.464186	-0.539871	0.377423
CJ	-0.370804	0.493073	-0.788729	-0.543533	1.000000	0.924626	-0.068346	-0.725753
한화	-0.357820	0.317671	-0.828159	-0.464186	0.924626	1.000000	-0.099603	-0.779787
현대자동차	-0.261648	0.371329	0.014257	-0.539871	-0.068346	-0.099603	1.000000	0.402249
기아자동차	0.483834	-0.070075	0.833377	0.377423	-0.725753	-0.779787	0.402249	1.000000

07. 주식 데이터 - 상관관계 그래프 그리기

▼ 14) 상관관계 그래프 그리기 위한 라이브러리 불러오기

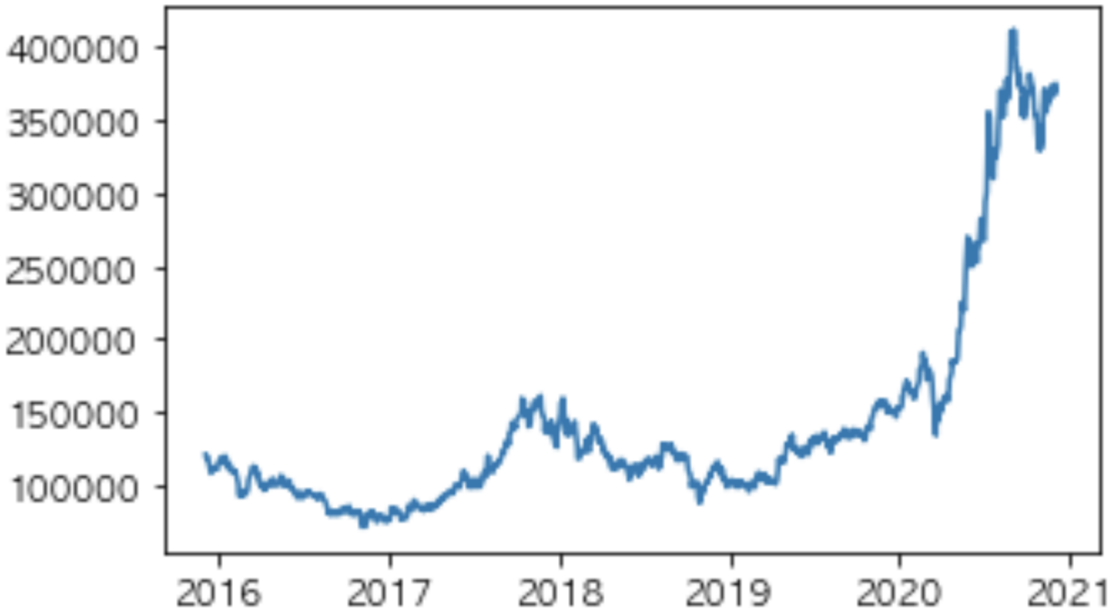
```
import matplotlib.pyplot as plt
import seaborn as sns

# Apple은 'AppleGothic', Windows는 'Malgun Gothic'을 추천
plt.rcParams['font.family'] = "Malgun Gothic"
plt.rcParams['axes.unicode_minus'] = False # 마이너스 기호 깨지는 걸 막아줘요
```

▼ [코드스니펫] seaborn 설치하기

```
conda install -c anaconda seaborn
```

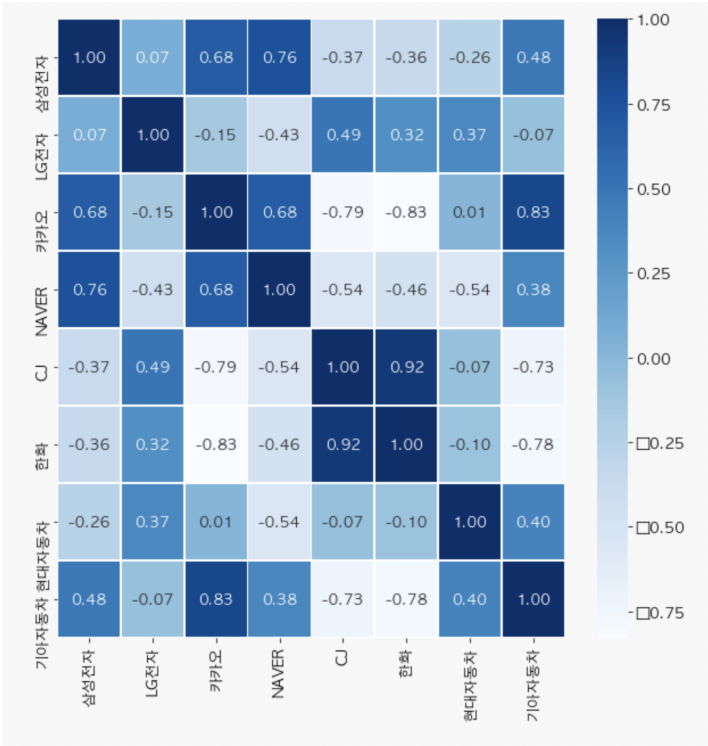
- seaborn으로 그래프를 그려볼게요

```
plt.figure(figsize=(5,3))
sns.lineplot(data=kakao_stock_df['Close'])
```



▼ 15) 상관관계 그래프 그리기

```
plt.figure(figsize=(10,10))
sns.heatmap(data = corr_data, annot=True, fmt = '.2f', linewidths=.5, cmap='Blues')
plt.show()
```



08. 데이터스튜디오 - 준비하기

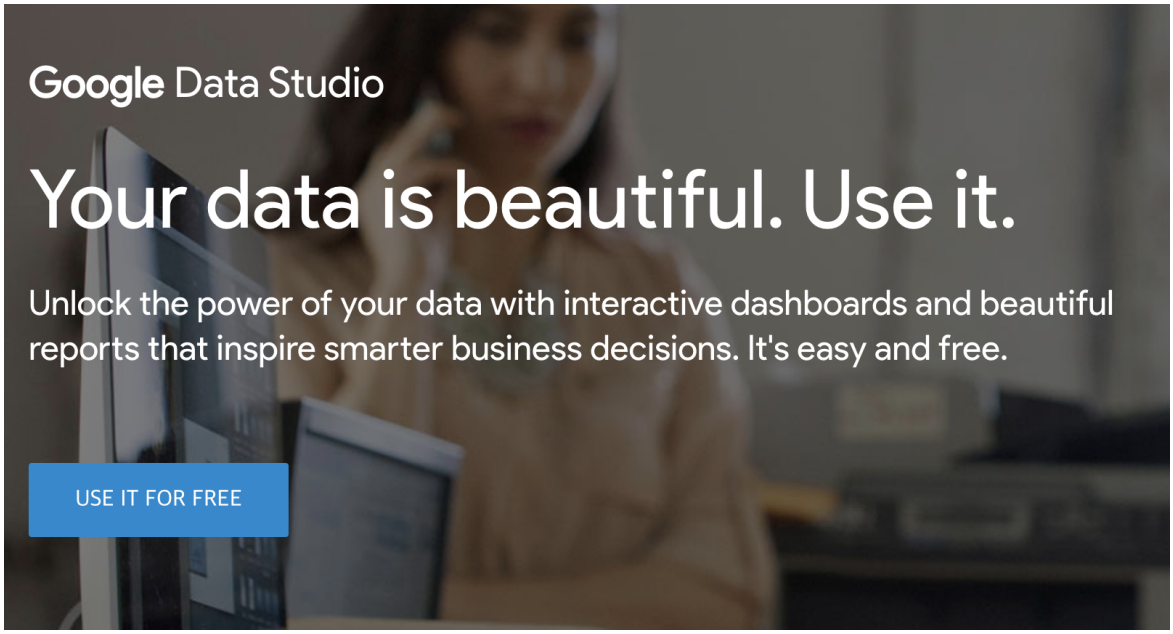
▼ 16) 구글 데이터 스튜디오

파이썬으로 보고서를 만들 수 있습니다. 그래프도 그릴 수 있어요. 하지만 공유하기 쉽지 않습니다. 또 언제 축 이름과 사이즈 색깔 지정하고 있나요? 데이터만 있으면 그래프를 만들어서 공유할 수 있도록 도와주는 도구를 소개 해드릴게요.

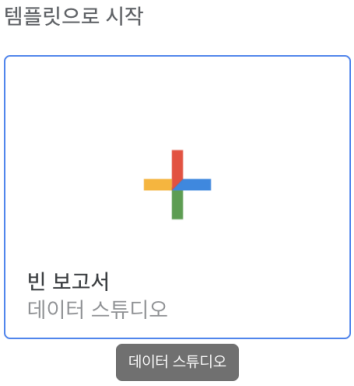
▼ **[코드스니펫] 데이터 스튜디오**

```
https://datastudio.google.com/u/0/navigation/reporting
```

▼ 17) 시작하기



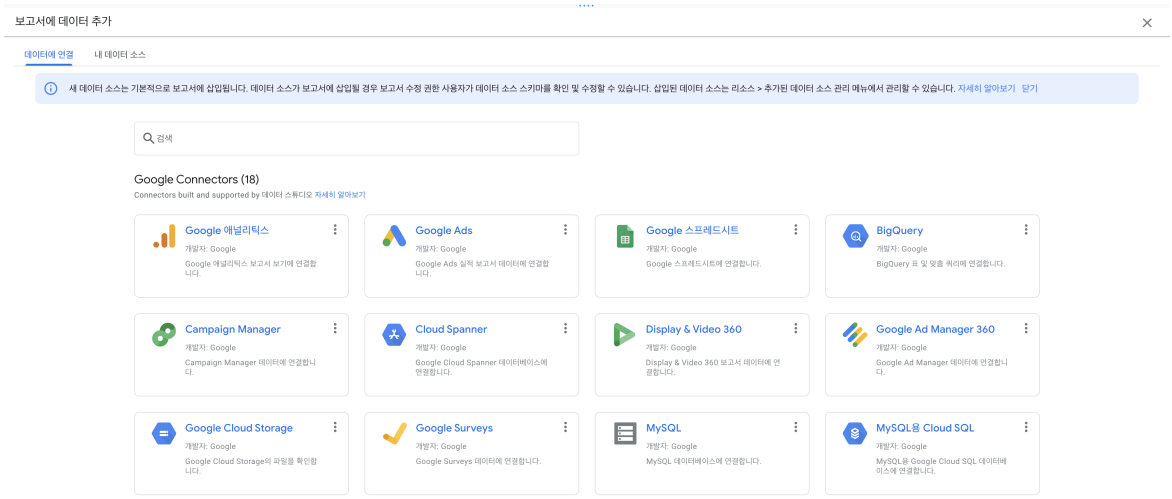
- Use it for free 를 클릭해주세요
- 구글 계정으로 로그인 해주세요.



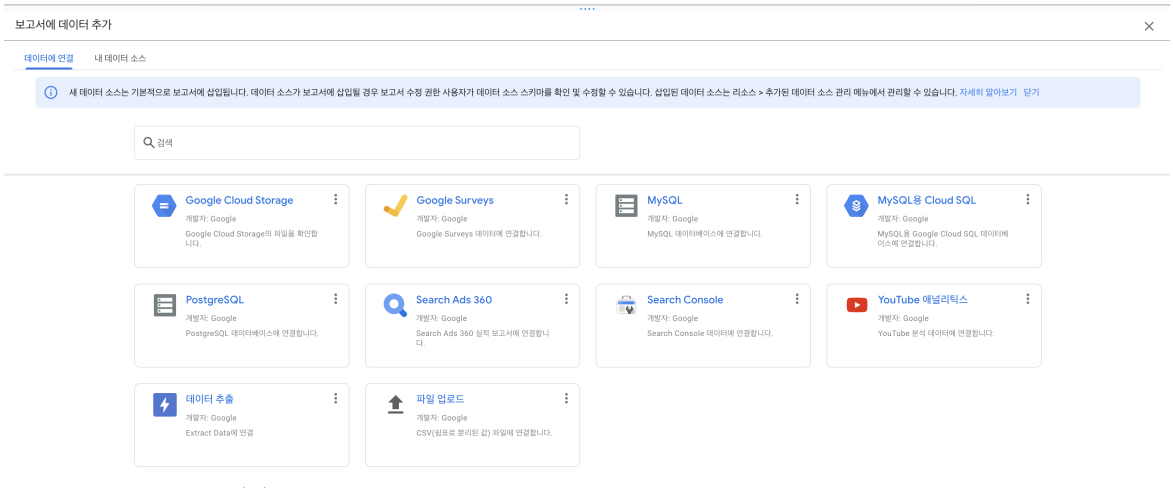
- 빈 보고서를 눌러서 새 보고서를 만들어주세요!

▼ 18) 데이터 입력하기

- 데이터 추가 버튼을 눌러주세요.



- 메뉴하단에 **파일 업로드**를 눌러 업로드 할 파일을 선택합니다.



- 무슨...데이터요?
- 아무 데이터나 좋아요. 예를 들어 우리가 구한 주식들의 등락율 데이터를 넣어볼까요?

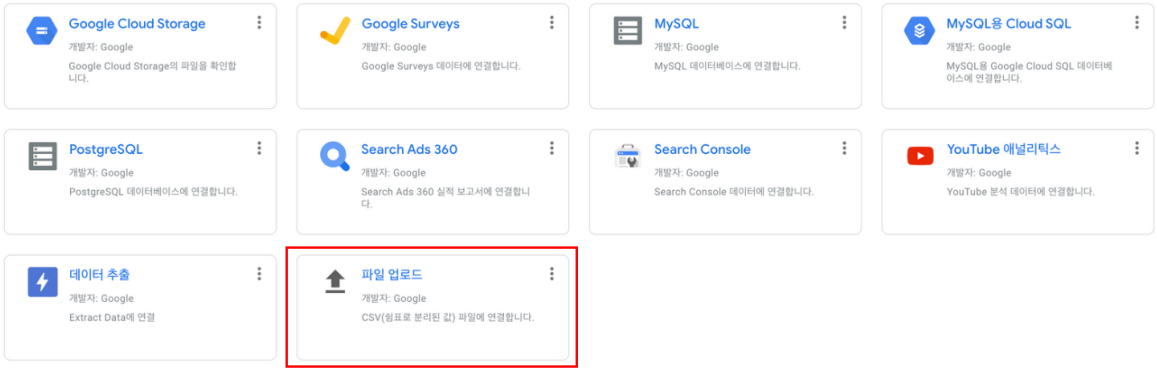
▼ 19) 데이터 추출하기

▼ **[코드스니펫] 데이터 추출하기**

```
import pandas as pd
import pandas_datareader as pdr
from datetime import datetime

code = pd.read_csv('./data/corpgeneral.csv', header=0)
code = code[['회사명', '종목코드']]

# 컬럼명 바꾸기
code_result = code.rename(columns={'회사명': 'corp', '종목코드': 'code'})
# 종목 코드 6자리만들기
code_result.code = code_result.code.map('{:06d}'.format)

def get_code(code_result, corp_name):
    condition = "corp=='{}'".format(corp_name)
    code = code_result.query(condition)['code'].to_string(index=False)
    # 위와같이 code명을 가져오면 앞에 공백이 붙어있는 상황이 발생하여 앞뒤로 sript() 하여 공백 제거
    code = code.strip()
    return code

companies = ['삼성전자', 'LG전자', '카카오', 'NAVER', 'CJ', '한화', '현대자동차', '기아자동차']
start = datetime(2019,1,1)
end = datetime(2020,9,10)
stocks_of_companies = pd.DataFrame({'Date': pd.date_range(start=start, end=end)})

for company in companies:
```

```
    company_code = get_code(code_result, company) + '.KS'
    stock_df = pdr.get_data_yahoo(company_code, start, end)
    stocks_of_companies = stocks_of_companies.join(pd.DataFrame(stock_df['Close'].pct_change()).rename(columns={'Close':compa

stocks_of_companies = stocks_of_companies.dropna()
stocks_of_companies.to_csv('./stock_change.csv',sep=',', na_rep='NaN',index = False)
```

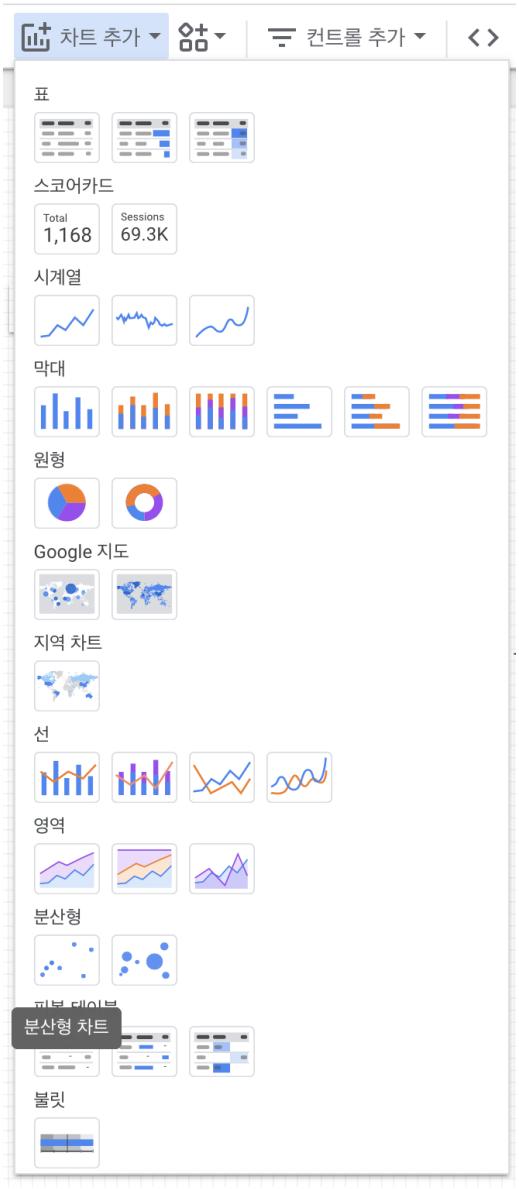
👉 csv파일이 만들어졌다면 파일 업로드를 클릭!



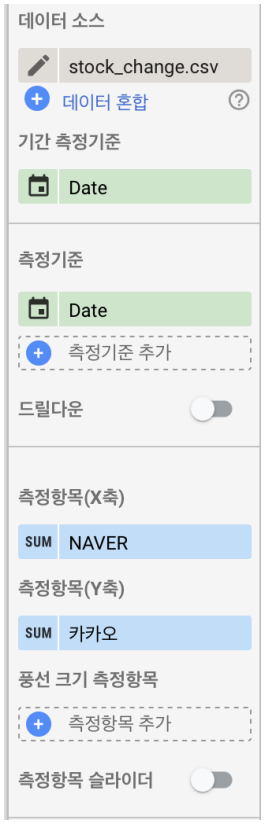
09. 데이터스튜디오 - 그래프 그리기

▼ 20) 그래프 그리기

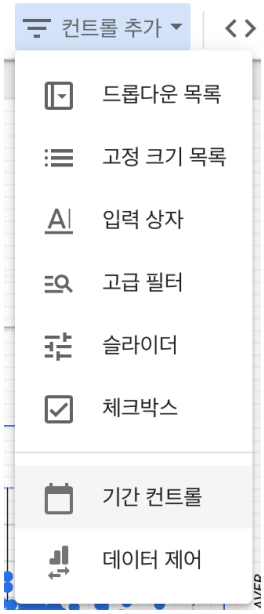
1. 차트 추가를 누르고 분산형 차트를 추가해줍니다



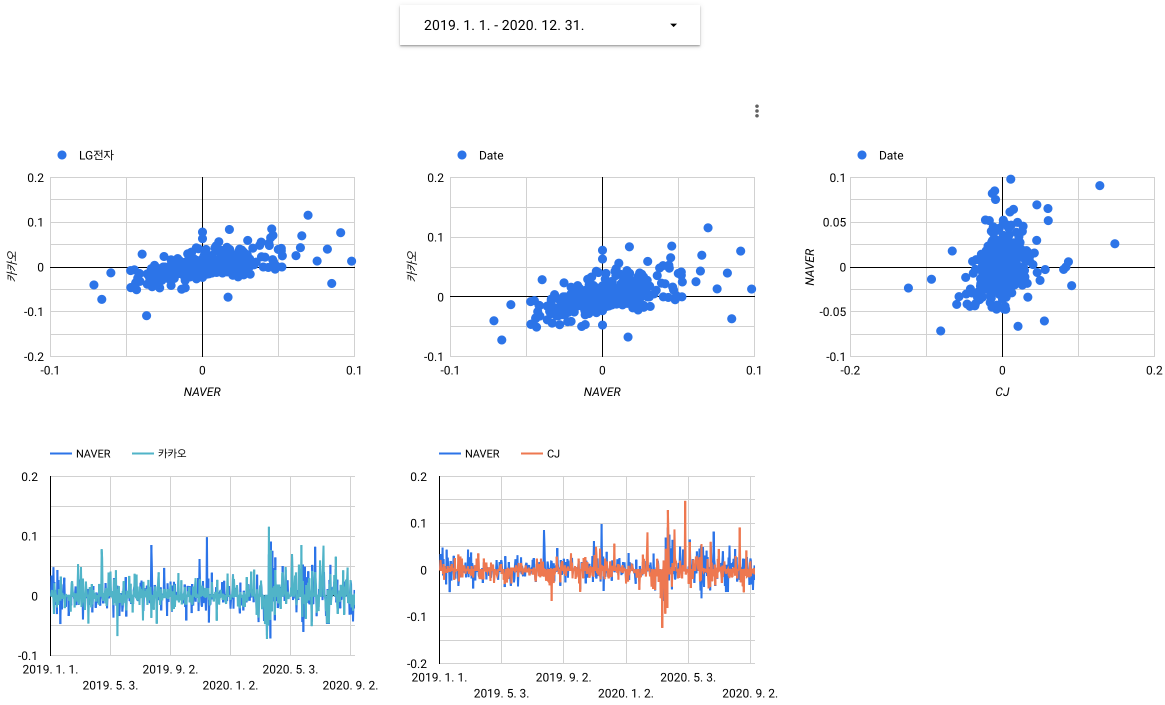
2. 그래프를 클릭한 뒤, 측정 기준, 측정항목을 선택해줍니다



3. 컨트롤 추가를 통해 쉽게 날짜를 바꿀 수 있는 컨트롤러를 추가해줍니다.

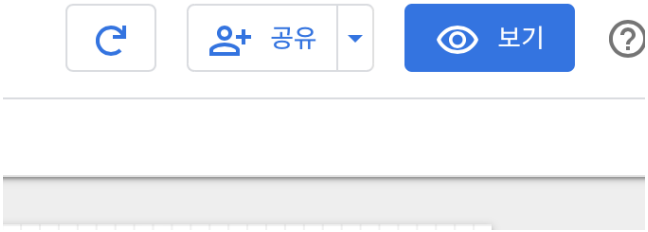


4. 다양한 그래프를 추가 해줍니다.

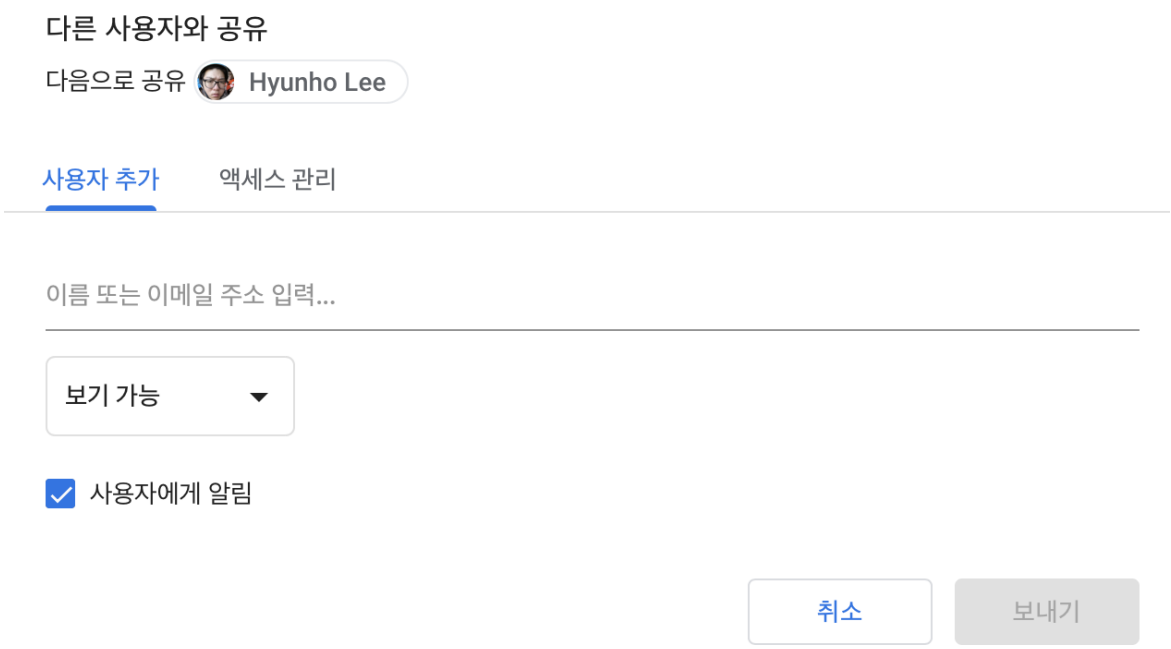


▼ 21) 공유하기

- 이제 내가 만든 보고서를 공유할 시간입니다.

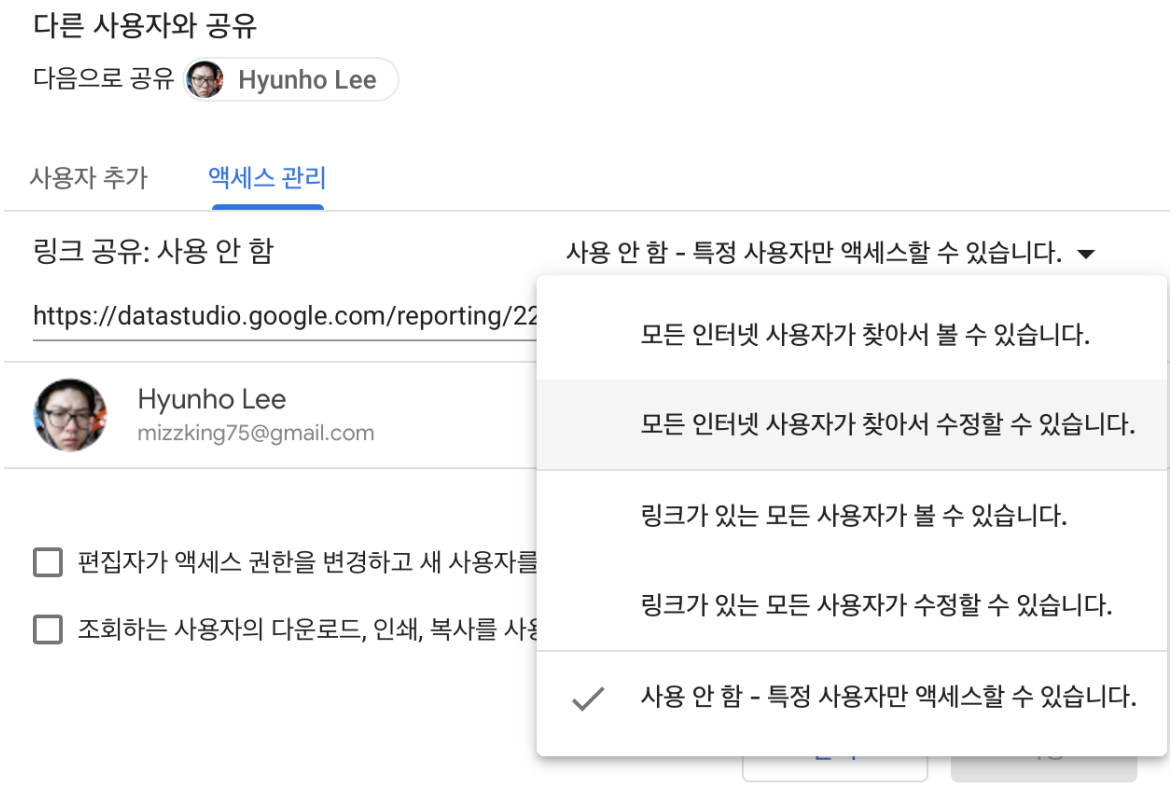


- 오른쪽 상단의 공유 버튼을 눌러 공유 할 사용자들의 이메일을 입력하면 됩니다.



▼ 22) 설정하기

- 매번 보여주고 싶은 사람에게 이메일을 물어보고 보내는 것은 번거로운 일입니다.
- 우리의 보고서의 열람과 수정 권한을 설정해 편리하고 안전하게 공유해보세요.

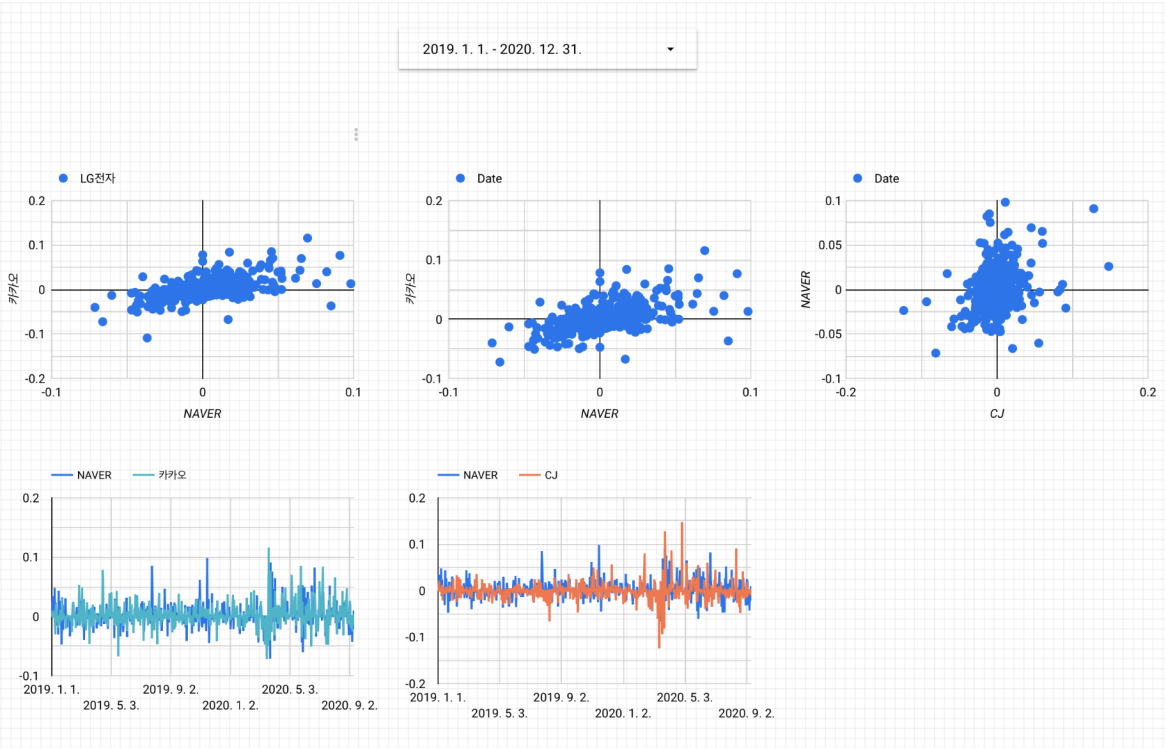


10. 마무리 & 숙제 설명

분석한 보고서의 URL을 제출하기!

- 잘 따라오셨다면 링크만 만들어서 제출하면 끝!✴️

▼ 결과 화면



+ 한걸음 더: 위의 주가 데이터 분석 보고서를 르탄이 또는 친구에게 자랑해보세요!

11. 4주차 숙제 답안 코드

▼ **[코드스니펫] - 4주차 숙제 답안 코드**

전체 코드

```
https://datastudio.google.com/reporting/339a8efe-5770-4d7a-8055-ac8d20816acd
```