**발표 시, 이야기해야하는 부분**

1. 각 항목을 추가한 이유
2. 경향성
3. 목표
4. 원하는 결과
5. 어떤 모델을 이용할지 -> ml
6. 시계열 을 이용한 모델에는 어떤 분석이 필요한가.
7. 주제 - 주제선정이유(닥터코퍼)
8. 왜 그 데이터를 사용하는가 - 컬럼 선별 및 추가 이유

-> 모든 컬럼 마다 -> 군집화 별

로 이유를 만들자(에너지, 식품, 금속 등 각 선정 이유)

1. 목표 - 생각하는 결과

- 원자재가 얼마나 영향을 미치는 가에 대한 고찰 -> 영향을 미친다!! (우리의 가설)

- 영향을 받는 대상을 무엇으로 할 것이냐 -> ( 코스피 코스닥 데이터 )= Y

- 연구 결과에 따라 가장 연관성(상관성)이 높은 컬럼을 베타 값이나 가중치로 금융시장에 사용될 가능성 내포

- 저 분석을 바탕으로 10월달(30일->2.5%(1200중의))의 코스피 코스닥 주가 예측

- 예측 -> 성능 평가 -> log loss, mse , cross validation->sliding window

1. 얼마나 했는지 ?

-> 전처리 했고, 상관계수 확인했고 시계열 처리를 위한 스케일링을 해볼 예정이다.

1. 예측을 위한 모델 뭐 쓸건가 -> ~~ 모델을 사용해볼 예정이다.

- xgboost : 트리모델 분류 -> 트리 모델이지만 정확도가 좋고 과적합을 안함.

-> 시계열에 대한 스케일링 해야한다.

- 모델을 좀 찾아봐야함. ?? 2 개 정도 더

- ARIMA: 시계열을 위한 모델

- RF: 랜덤 포레스트

- LSTM: 딥러닝

1. + 고려해야할점 -> 시계열이라는 점 + 공란 처리에 대한 고찰 -> 4번 설명 시

오늘할 것 -> 모델 찾기, 전처리 데이터: 시각화

ppt 템플릿 파랑색이나 초록색 계열로

추가: 날짜 미루고 당기기

병훈

## **시계열 데이터의 특성** [#](http://www.incodom.kr/%EC%8B%9C%EA%B3%84%EC%97%B4%EB%B6%84%EC%84%9D#h_45c0f2e6e59c07a349827f33ec9024ec)

## **정상성(Stationary) vs 비정상성(Non-Stationary)** [#](http://www.incodom.kr/%EC%8B%9C%EA%B3%84%EC%97%B4%EB%B6%84%EC%84%9D#h_a7b608daf8e753d825dae790a08bb4e8)

시계열 데이터는 시간의 흐름에 따라 평균이나 분산 등의 통계적 특성이 변하지 않고, 일정한 추세가 없는 정상성(Stationary) 데이터와 시간에 따라 통계적 특성이 변화하는 비정상성(Non-Stationary) 데이터로 나눌 수 있습니다.

* 자기상관계수 + 편 ~

#####

**스케일링(Scaling) :** pc메모리를 고려하여 오버플로우나 언더플로우를 방지하고 독립 변수의 공분산 행렬 조건수를 감소시킴

* **Standard Scaler :** 평균을 제외하고 표준편차로 나누어 변환(정규 분포를 가정시 사용)
* **Min-Max Scaler :** 가장 많이 활요되며 최소~최대 값이 0~1 또는 -1~1 사이로 변환(정규 분포 가정을 안함)
* **Robust Scaler** : 최소-최대 스케일러와 유사하지만 최소/최대 대신 IQR중 1분위수와 3분위수를 사용하여 변환
* **Normakuzer** : 각 변수들 전체 n개 모든 변수들의 크기들로 나누어서 변환, 모든 변수들의 값은 원점으로 부터 반지름 1 이내

#####

지훈

원자재 가격 변화가 코스피,코스닥에서는 하루 후(or 몇일 후)에 반영될 수도 있지 않나?

그거는 뒤에가서 날짜를 미루거나 당기는 방식으로 시계열 데이터를 다룰 예정입니다

재욱

LSTM

GBRT : 두 모델이 시계열 자료에서 성능이 좋다고 함