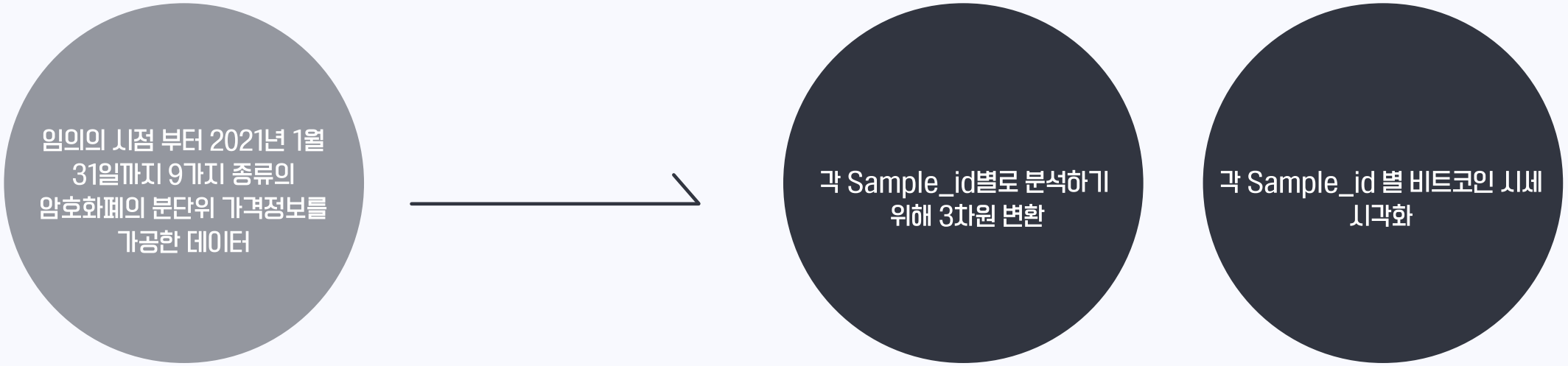
Dacon

인공지능 비트 Trader 경진대회 1주차

2016125005 김강윤
2016125041 옥현진
2016125027 백경환
2016125033 송재섭

INDEX

1-1 진행상황



대회 측 제공 데이터

group_by, group_split 사용

ggplot2 사용

1-1 진행상황

데이터 구조

train_x,y의 default 데이터

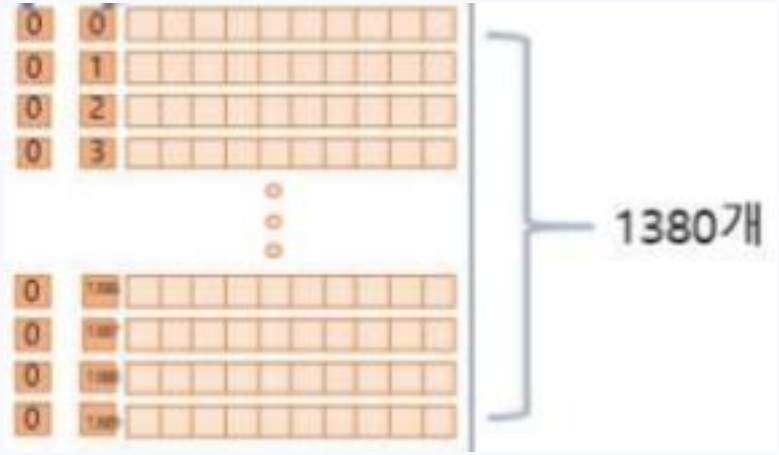
```
'data.frame'

Rows: 10,159,560
Columns: 12
$ sample_id   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ time        <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1...
$ coin_index  <int> 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7...
$ open        <dbl> 1.010004, 1.009808, 1.009808, 1.010200, 1.010985, 1.010...
$ high        <dbl> 1.010004, 1.009808, 1.010200, 1.011181, 1.010985, 1.011...
$ low         <dbl> 1.009612, 1.009808, 1.009808, 1.010200, 1.010200, 1.010...
$ close       <dbl> 1.010004, 1.009808, 1.010200, 1.011181, 1.010200, 1.011...
$ volume      <dbl> 838287.50, 162242.05, 16649.67, 2586971.25, 1129996.00,...
$ quote_av    <dbl> 43160.6328, 8352.2207, 857.3778, 133310.3438, 58216.867...
$ trades      <dbl> 451.15729, 39.23107, 58.84660, 431.54178, 176.53981, 25...
$ tb_base_av  <dbl> 732683.375, 0.000, 16649.666, 2189146.750, 0.000, 12266...
$ tb_quote_av <dbl> 37725.1836, 0.0000, 857.3778, 112811.0469, 0.0000, 6321...
```

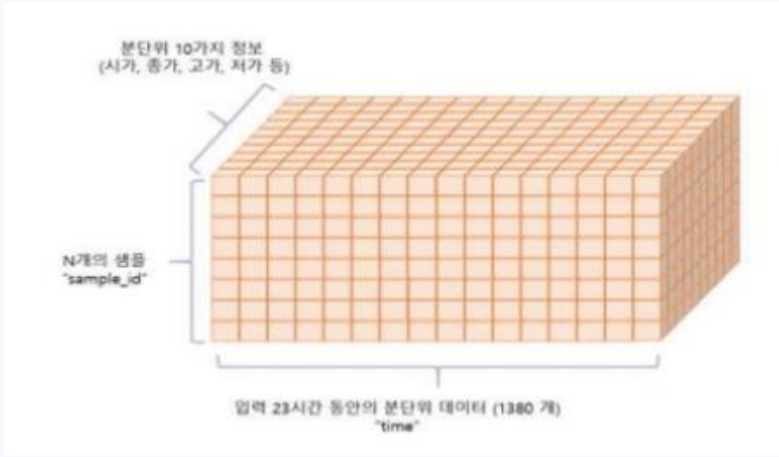
Column	의미
Sample_id	개별 샘플의 인덱스
time	동일 샘플내 시간정보
Coin_index	코인 종류 비식별 인덱스
Open	시가
High	고가
low	저가
close	종가
volume	거래량
Quote_av	견적 통화 거래량
trades	거래 건 수
tb_base_av	매수자 거래량
tb_quote_av	매수자 견적 통화 거래량

1-1 진행상황

Csv 파일은 현재 Sample_id별로 분류 되어있지 않은 2차원 dataframe 이므로 이를 그룹화 하는 작업을 진행

```
df2d_to_3d <- function(df_2d) {
  temp <- df_2d %>%
group_by(sample_id)
  df_3d = group_split(temp)
  return(df_3d)
}
```





train_x의 관측치: 1380(=23시간 *60분)
train_y의 관측치: 120(=2시간*60분)

1-1 진행상황

Before

A data.frame: 6 × 12

	sample_id	time	coin_index	open	high	low	close	volume	quote_av	trades	tb_base_av	tb_quote_av
	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0	0	7	1.010004	1.010004	1.009612	1.010004	838287.50	43160.6328	451.15729	732683.38	37725.1836
2	0	1	7	1.009808	1.009808	1.009808	1.009808	162242.05	8352.2207	39.23107	0.00	0.0000
3	0	2	7	1.009808	1.010200	1.009808	1.010200	16649.67	857.3778	58.84660	16649.67	857.3778
4	0	3	7	1.010200	1.011181	1.010200	1.011181	2586971.25	133310.3438	431.54178	2189146.75	112811.0469
5	0	4	7	1.010985	1.010985	1.010200	1.010200	1129996.00	58216.8672	176.53981	0.00	0.0000
6	0	5	7	1.010396	1.011377	1.010396	1.011377	1226671.25	63211.7227	255.00195	1226671.25	63211.7227

sample_id별로 그룹화

After

A tibble: 1380 × 12

1.	sample_id	time	coin_index	open	high	low	close	volume	quote_av	trades	tb_base_av	tb_quote_av
	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
	0	0	7	1.010004	1.010004	1.009612	1.010004	838287.50	43160.6328	451.15729	732683.375	37725.1836
	0	1	7	1.009808	1.009808	1.009808	1.009808	162242.05	8352.2207	39.23107	0.000	0.0000
	0	2	7	1.009808	1.010200	1.009808	1.010200	16649.67	857.3778	58.84660	16649.666	857.3778
	0	3	7	1.010200	1.011181	1.010200	1.011181	2586971.25	133310.3438	431.54178	2189146.750	112811.0469
	0	4	7	1.010985	1.010985	1.010200	1.010200	1129996.00	58216.8672	176.53981	0.000	0.0000
	0	5	7	1.010396	1.011377	1.010396	1.011377	1226671.25	63211.7227	255.00195	1226671.250	63211.7227
	0	6	7	1.011377	1.011769	1.011377	1.011769	165829.73	8552.8252	78.46214	156767.359	8085.5684

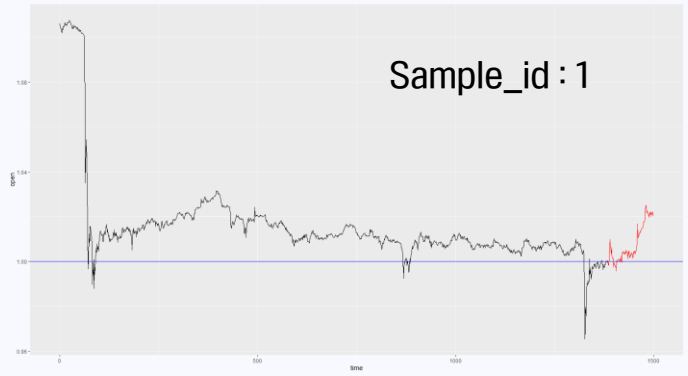
Tibble 1380 * 12 로 변환 됨. 즉 Sample_id (7362개) 마다 1380*12의 데이터프레임이 만들어짐

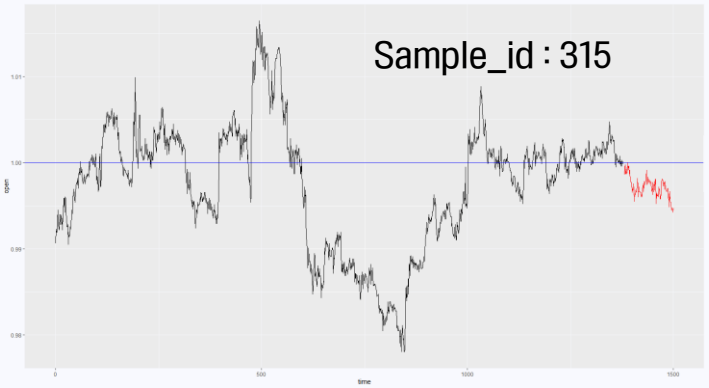
1-1 **진행상황**

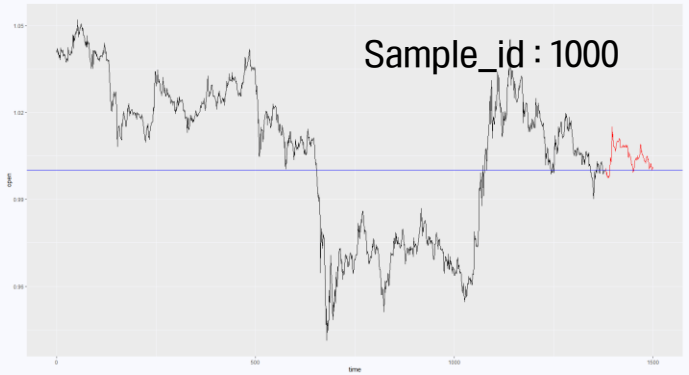
Sample_id별 시각화

특정 Sample_id의 1380개의 분단위 데이터 + 120개의 2시간 동안의 분단위 데이터, train_x와 train_y의 데이터를 합하여 시각화 진행

```
make_graph<-function(idx){
  ggplot(data = train_x_list[[idx+1]],
         aes(x = time, y = open)) + geom_line() + geom_line(data =
train_y_list[[idx+1]],
aes(x=time+1381,y=open),color="red")+geom_hline(yintercept=1, color ="blue")
}
```

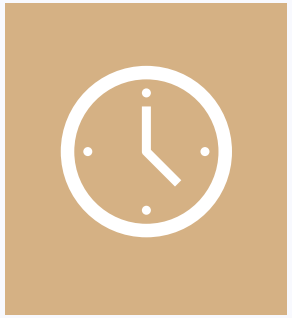






2

문제분석 및 토론



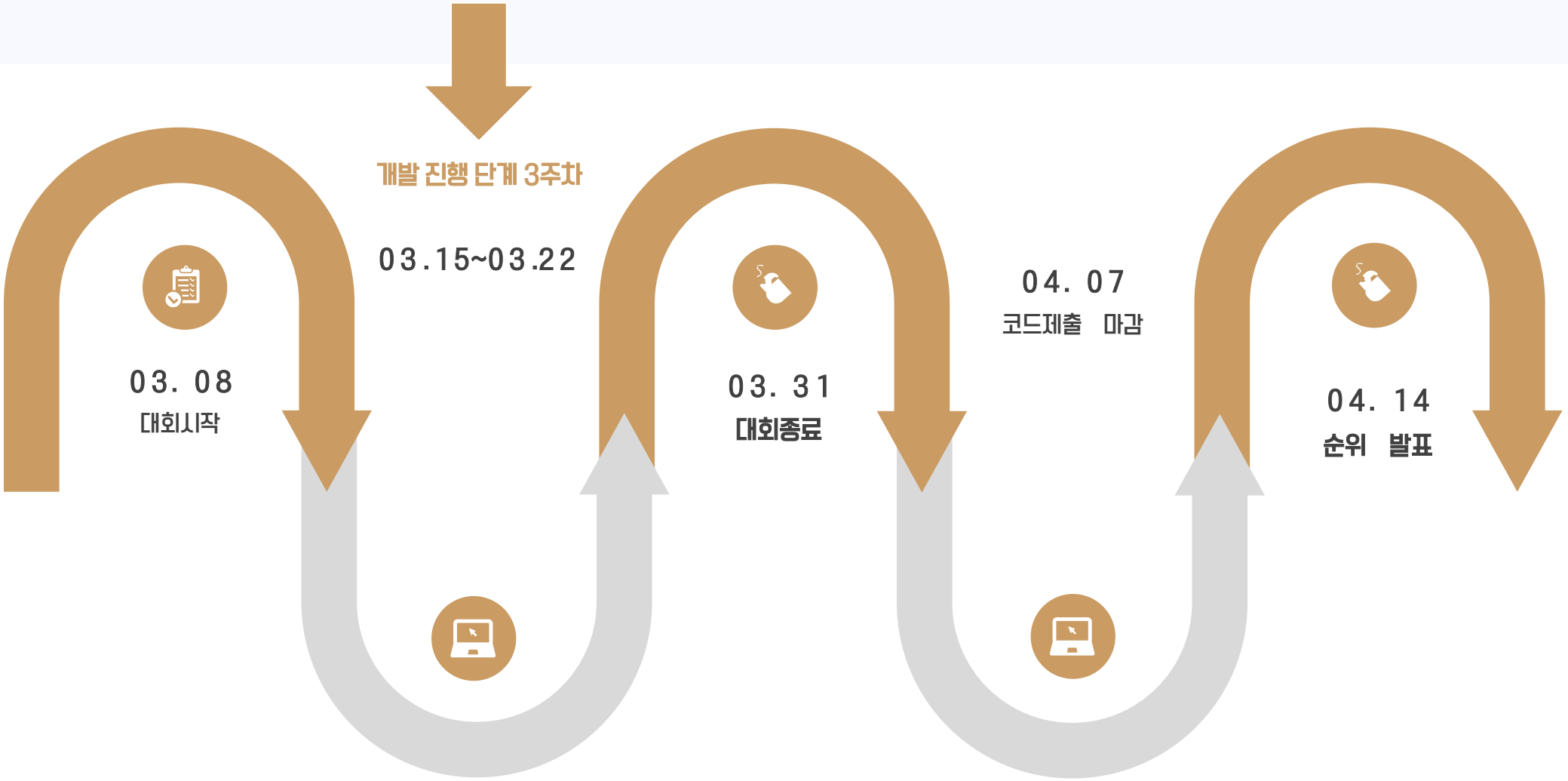
시계열 모델 선정

-AR, MA, ARMA, ARIMA 에서 선정하여 규칙성을 부여 하더라도, 결과가 좋지 않으면 LSTM같은 딥러닝 기법을 이용해야 한다. 모델에 대해 지속적으로 공부를 하면서, 조언을 구하는 쪽으로 방향을 잡았다.

데이터 접근 방식 선정

-529가지 sample_id에 대해 매수량, 매도 시점을 결정해 주어야 한다. 대회에서 요구하는 결과를 보여야 하는데, 어떤 근거로 결정할지에 대한 특정 지표 선정이 필요하다.

03 대회 현황


개발 진행 단계 3주차
03.15~03.22
03. 08
대회시작
03. 31
대회종료
04. 07
코드제출 마감
04. 14
순위 발표

04 4주차 계획

입력 23시간 동안의 분 단위 데이터(총 529가지 samples)

예측 모델 공부 및 선정

기간 3.15 ~ 3.22

데이터 분석 기반	AR, MA, ARMA, ARIMA 에서 선정하여 규칙성을 부여 시계열 데이터 분석 라이브러리인 fbprophet을 활용
딥러닝 기반	다양한 딥러닝 모델 중 적합한 것을 선정 -LSTM -LSTM Bidirectional -LSTM 2-Path -GRU -GRU Bidirectional -GRU 2-Path

Sample_id에 따른 매수량(buy quantity), 매도시점(sell_time) 결정

기간 : 3.15~ 3.22

대회 요구사항	개별 샘플마다 2시간 이내 모두 매도하여 하며 수익률을 남김에 있어 좋은 결과 요구
매수량 Buy_quantity	매수량은 0~1사이 정수이며 가장 좋은 결과를 낼 수 있을 수량을 결정해야한다
매도시점 Sell_time	매도 시간은 0~119사이 정수이며 가장 좋은 결과를 낼 시점을 정해야 한다.

THANK YOU