


상품 시세 분석 및 예측

2019-2



20140557 강하림

DIRECTORY 목차

주제 및 목표

데이터 정제

데이터 분석

데이터 예측

주제 및 목표

**온라인 상품 가격 데이터를 분석하여 현재의
특정 상품 가격을 예측해보자!**



**‘청바지’ 가격 데이터를 분석하여 미래의
‘청바지’ 가격을 예측해보자!**

데이터 선정 이유

**시간에 비해 많은 데이터양.. 시도해 보았지만,
기간내에 불가능할 것이라는 판단!
한 품목에 대해서만 해보자!
휴대전화 , TV → 과거 데이터 부족**

사계절 내내 입을 수 있는 청바지는 어떨까?

데이터 선정 및 소개

온라인 수집가격 정보

설명 : 온라인에서 수집된 가격정보, 수집일자, 품목명, 판매가격 등 총 8개 항목, 각 항목은 ,(콤마) 문자로 구분

데이터 구조

수집일자	상품ID	품목ID	품목명	상품명	판매가격	할인가격	혜택가격
COLLECT_DAY	GOOD_ID	PUM_ID	PUM_NAME	GOOD_NAME	SALES_PRICE	DISCOUNT_PRICE	BENEFIT

데이터 수집 기간

2014년 01월 ~ 2019년 10월

분석에서 사용된 데이터

2015년 01월 ~ 2019년 10월 (하루 평균 5만개의 데이터 : 약 $50,000 \times 365 \times 4.5$ = 약 1억 개의 데이터)

데이터 정제

	A	B	C	D	E	F	G	H
1	collect_day	good_id	pum_id	pum_name	good_name	sales_price	discount_p	benefit
2	#####	5.27E+08	C013010	점퍼	여성 니팅	46550	46550	0
3	#####	6.41E+08	E061010	세탁세제	[LG생활	9500	9500	0
4	#####	1.98E+09	E031040	전기레인지	Q아람/퀵	64700	64700	0
5	#####	1.7E+09	A018060	스낵과자	3000 120g	17000	16440	-560
6	#####	5.73E+08	C013010	점퍼	[NII][NII]	42000	39900	-2100
7	#####	2.07E+09	E061030	전구	led 형광등	14600	14600	0
8	#####	1.18E+09	C011020	남자상의	[루이까또	99000	94050	-4950
9	#####	1.18E+09	C011020	남자상의	[루이까또	99000	94050	-4950
10	#####	1.42E+09	C014010	아동복	와이드시	28640	28640	0
11	#####	1.13E+09	C013010	점퍼	[지오지아	20000	16000	-4000

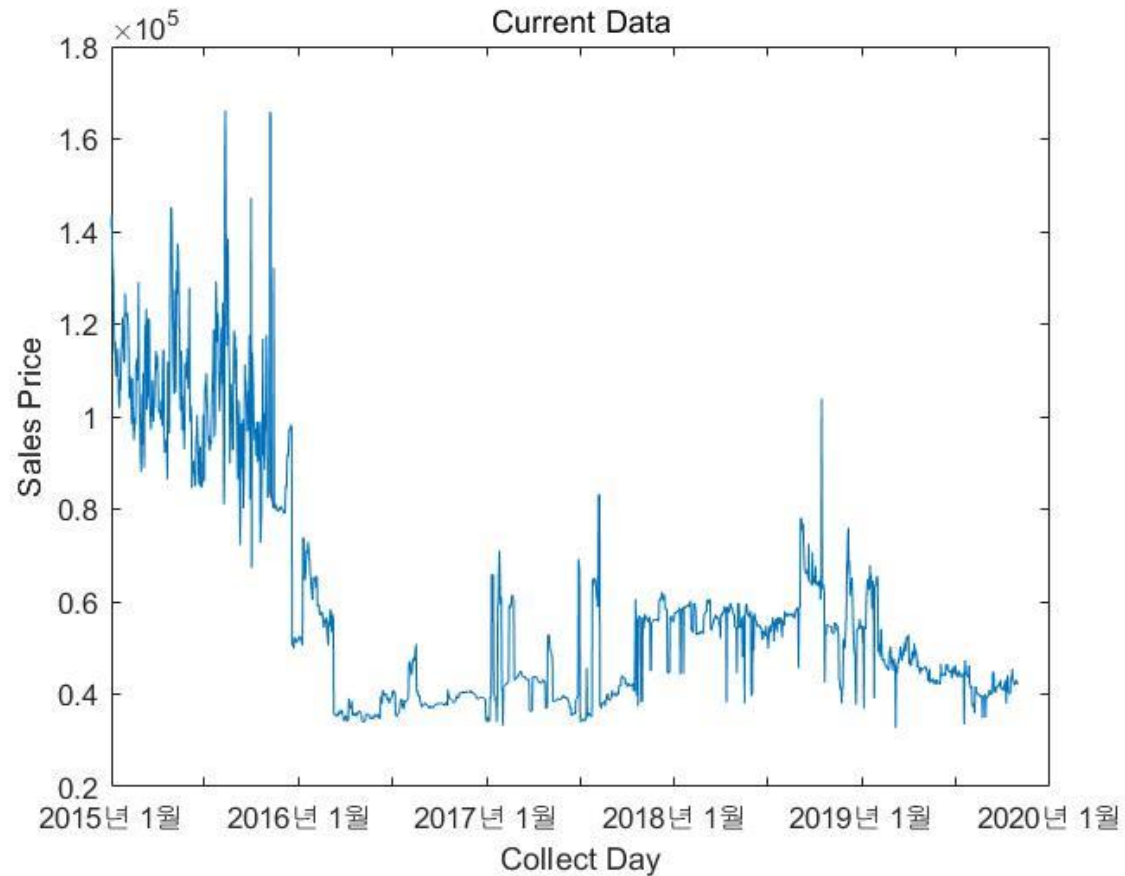
1	2	3
collect_day	pum_name	sales_price
2015-11-01	'청바지'	8.2715e+04
2015-11-02	'청바지'	8.2501e+04
2015-11-03	'청바지'	8.3175e+04
2015-11-04	'청바지'	8.3405e+04
2015-11-05	'청바지'	8.3308e+04
2015-11-06	'청바지'	1.6586e+05
2015-11-07	'청바지'	NaN
2015-11-08	'청바지'	NaN
2015-11-09	'청바지'	8.2274e+04
2015-11-10	'청바지'	8.1737e+04
2015-11-11	'청바지'	8.1215e+04
2015-11-12	'청바지'	8.0184e+04
2015-11-13	'청바지'	1.3217e+05
2015-11-14	'청바지'	8.0207e+04
2015-11-15	'청바지'	8.0133e+04
2015-11-16	'청바지'	8.0138e+04

1일 평균
청바지 판매 가격

```

year = '2018';
month = ['01'; '02'; '03'; '04'; '05'; '06'; '07'; '08'; '09'; '10'; '11'; '12'];
dataset = table();
for i=1 : length(month)
    path = sprintf('%s\\%s%s\\*.csv', year, year, month(i,:));
    datafiles = dir(path);
    numoffiles = length(datafiles);
    for j=1 : numoffiles
        datafilename = [datafiles(j).folder '\\' datafiles(j).name];
        T = readtable(datafilename);
        if isempty(T)
            continue
        else
            try
                T = T(:, {'collect_day', 'pum_name', 'sales_price'});
            catch
                fprintf('Error : %s - %d', month(i,2), j);
                continue
            end
            T.collect_day = datetime(T.collect_day);
            today = T.collect_day(1);
            T.pum_name = categorical(T.pum_name);
            n_T = table();
            c_index = T.pum_name == '청바지';
            A = T(c_index,:);
            cell T = {today, '청바지', mean(A.sales_price)};
            n_T = [n_T; cell T];
            n_T.Properties.VariableNames = T.Properties.VariableNames;
            if isempty(dataset)
                dataset = n_T;
            else
                dataset = [dataset; n_T];
            end
        end
        fprintf('%d', j);
        clearvars -except dataset datafiles numoffiles year month i j
    end
    fprintf('%d\n', i);
end
    
```

데이터 분석



기간 : 2015년 1월 1일 ~ 2019년 10월 31일
총 일 수 : 1765일

품목 : '청바지'

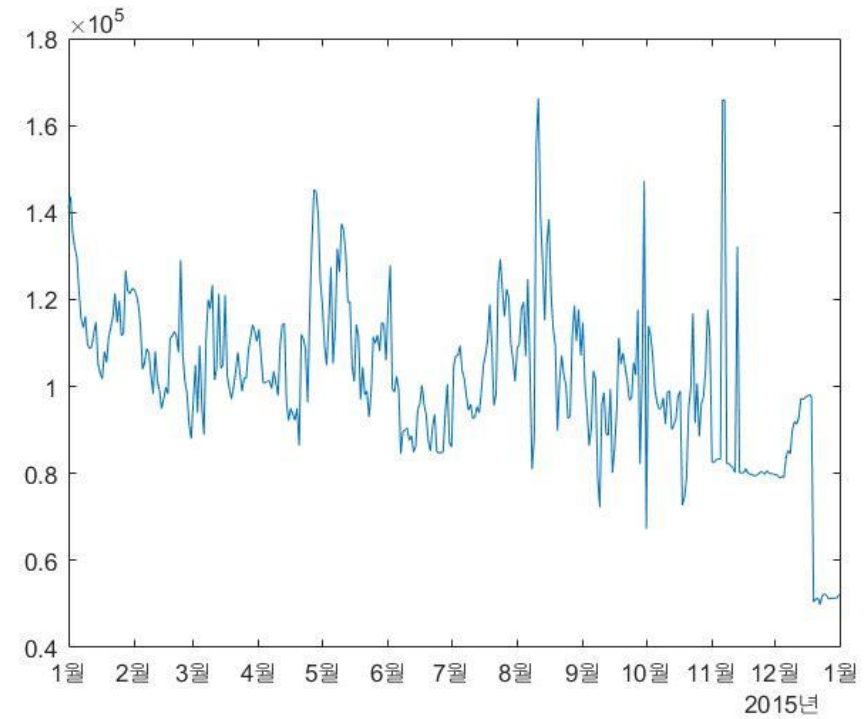
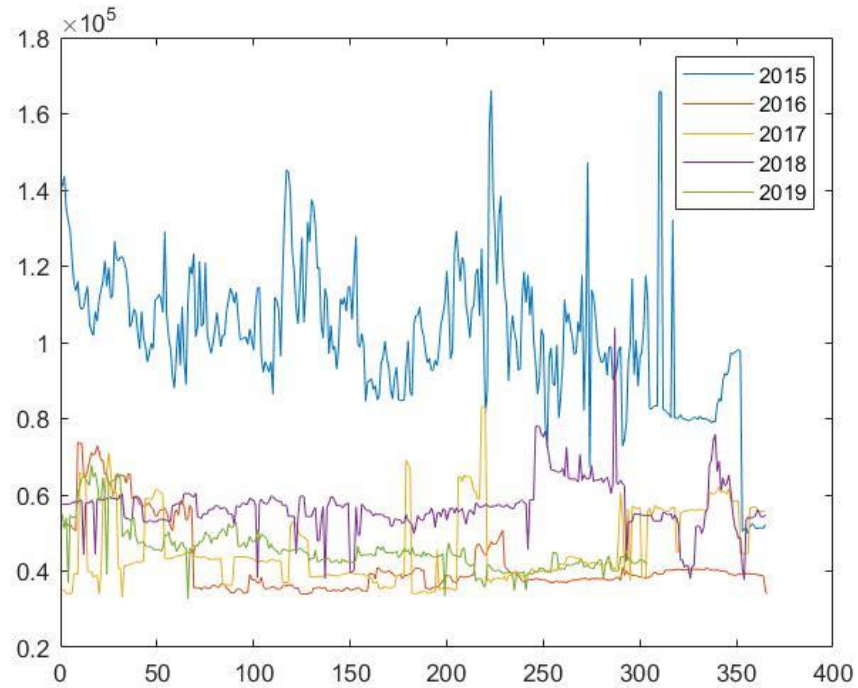
최소값 : 32732 원
평균값 : 51998 원
최대값 : 166220 원

`sales_price: 1765×1 double`

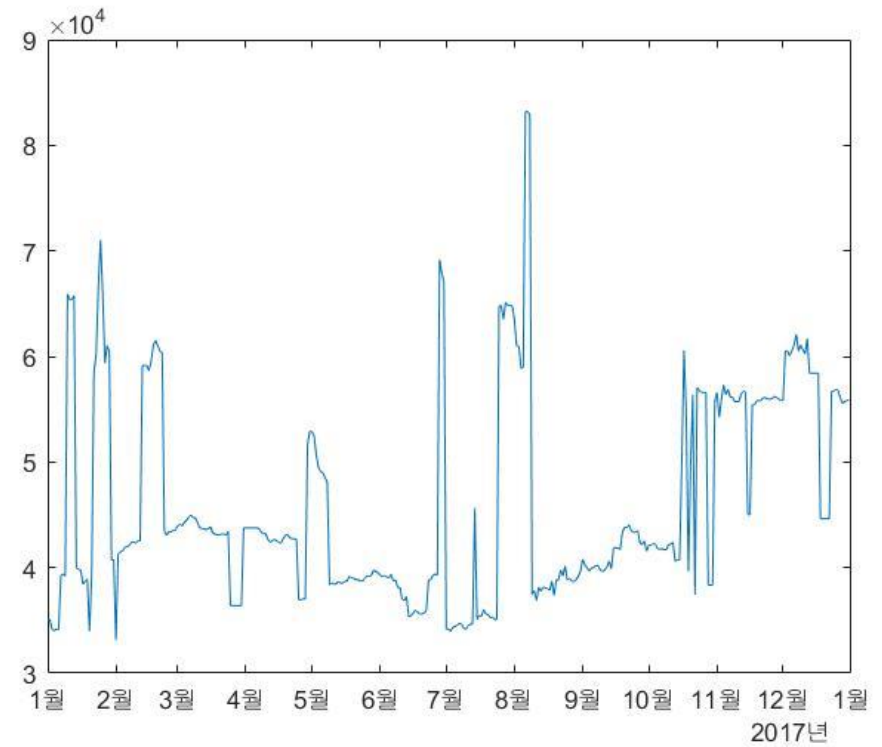
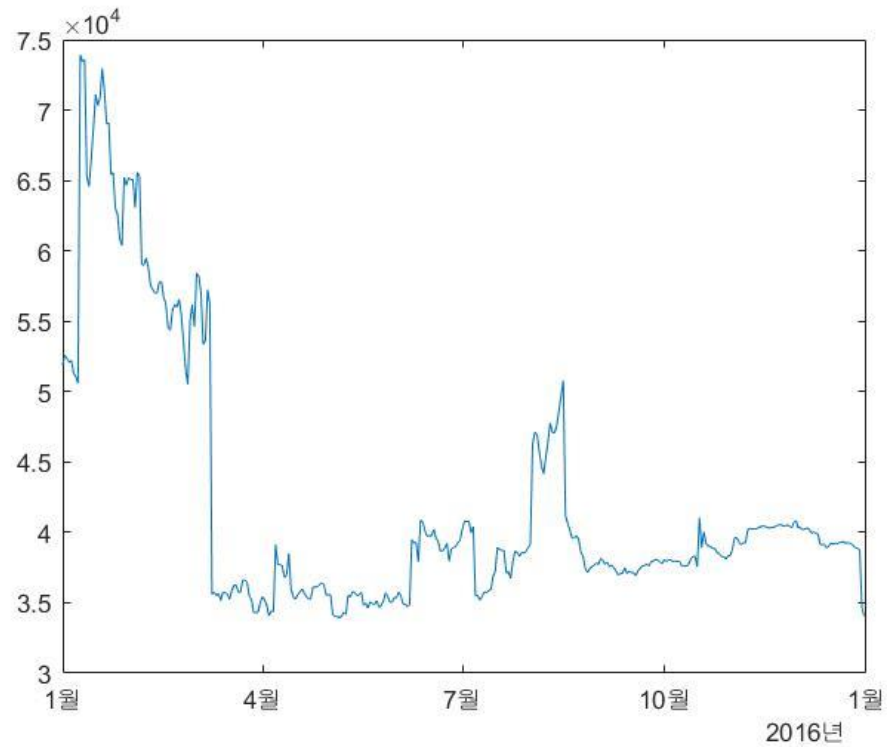
Values:

Min	32732
Median	51998
Max	1.6622e+05

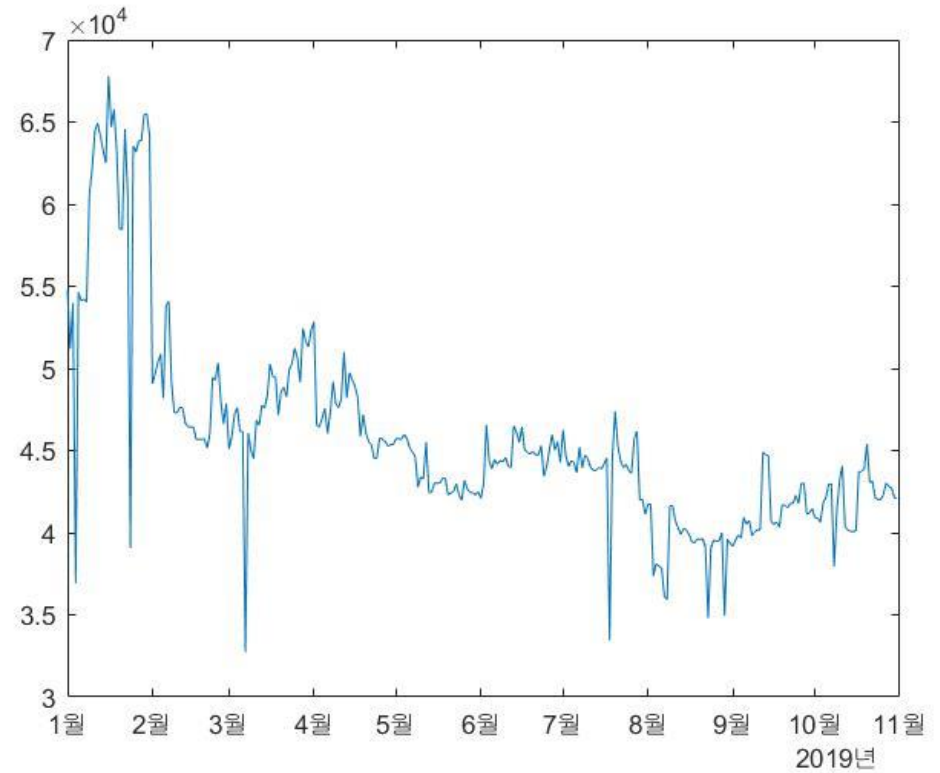
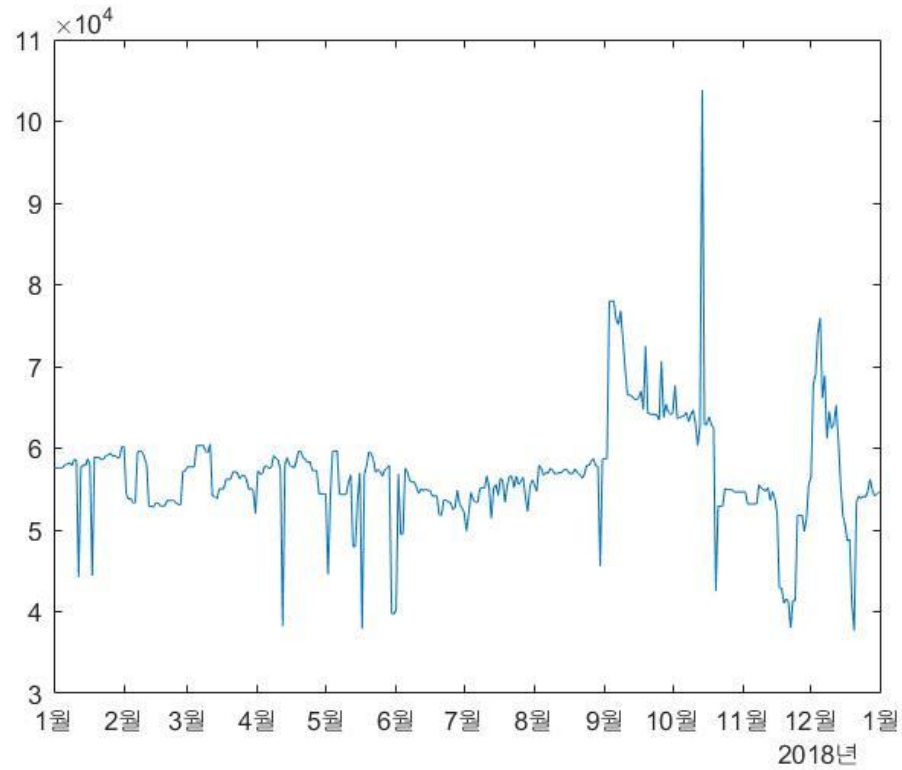
데이터 분석



데이터 분석



데이터 분석



데이터 정제 (예측을 위한)

```
% 청바지 평균값 데이터 정제
jean_data = readtable('mean\tt.xlsx');
jean_data.sales_price = fillmissing(jean_data.sales_price, 'nearest');
lenofdata = length(jean_data.sales_price);

for i=1 : length(jean_data.collect_day)
    jean_data.collect_day(i) = strip(jean_data.collect_day(i), '');
end

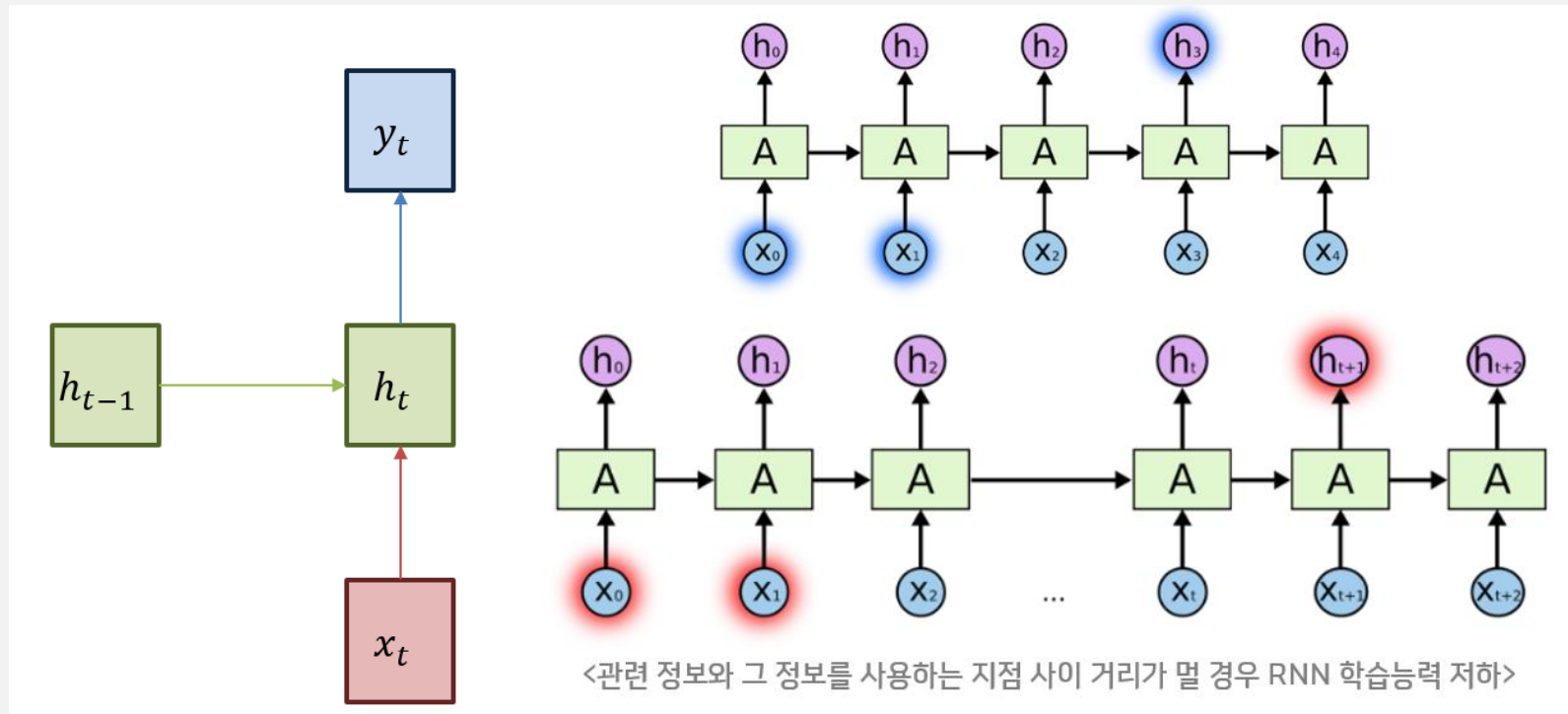
X = 1:1:lenofdata;
Y = jean_data.sales_price;
data = Y';
```

```
% 2015년 1월 1일 ~ 2019년 5월 6일 (90%) : 훈련데이터 세트
% 2019년 5월 7일 ~ 2019년 10월 31일 (10%) : 평가데이터 세트
numTimeStepsTrain = floor(0.9*numel(data));
dataTrain = data(1:numTimeStepsTrain+1);
dataTest = data(numTimeStepsTrain+1:end);

% Sales_price 값을 0과 1사이의 값으로 표준화
mu = mean(dataTrain);
sig = std(dataTrain);
dataTrainStandardized = (dataTrain - mu) / sig;
XTrain = dataTrainStandardized(1:end-1);
YTrain = dataTrainStandardized(2:end);
```

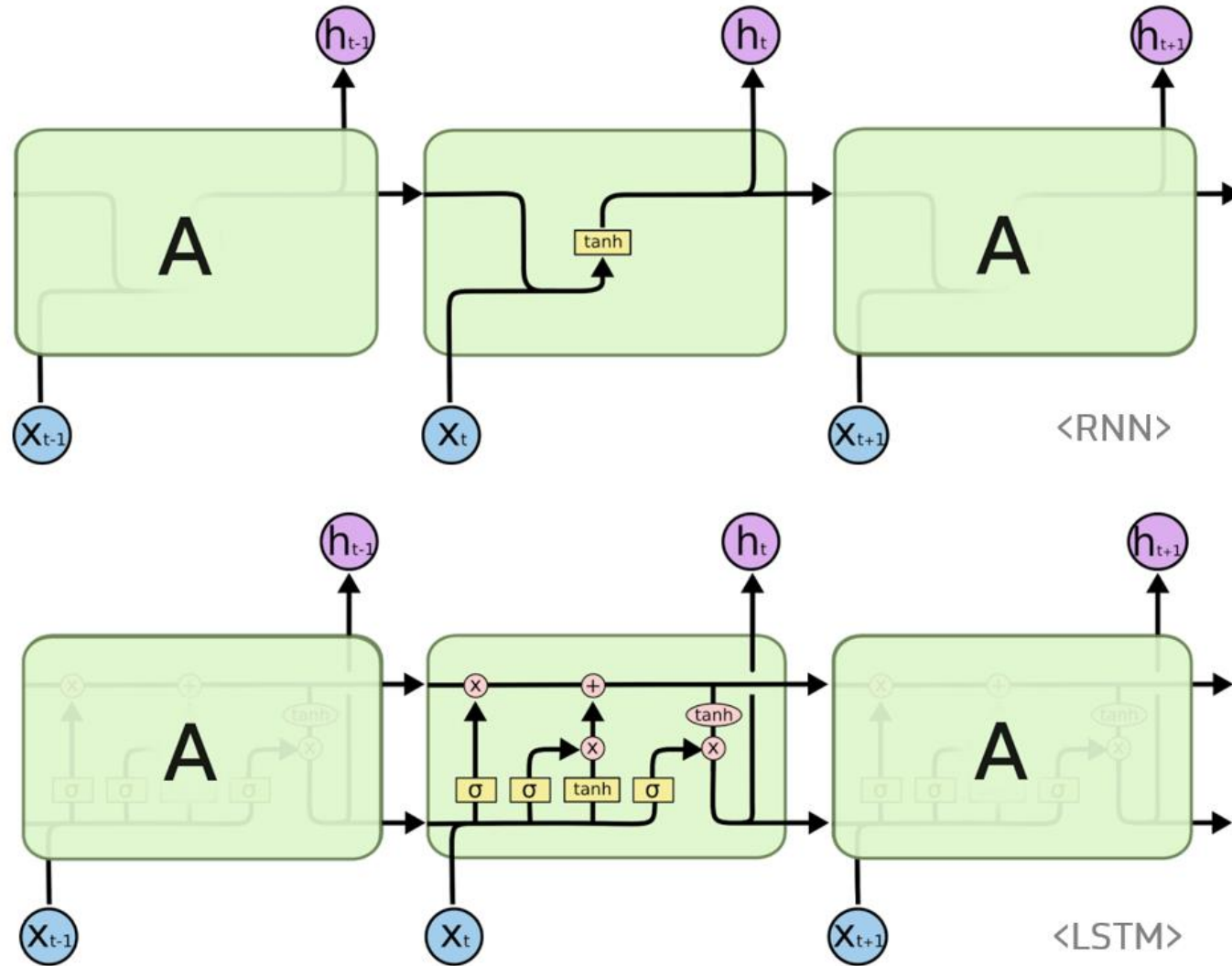
예측을 위한 선정모델 : LSTM

LSTM : Long Short-Term Memory Models



예측을 위한 선정모델 : LSTM

설명



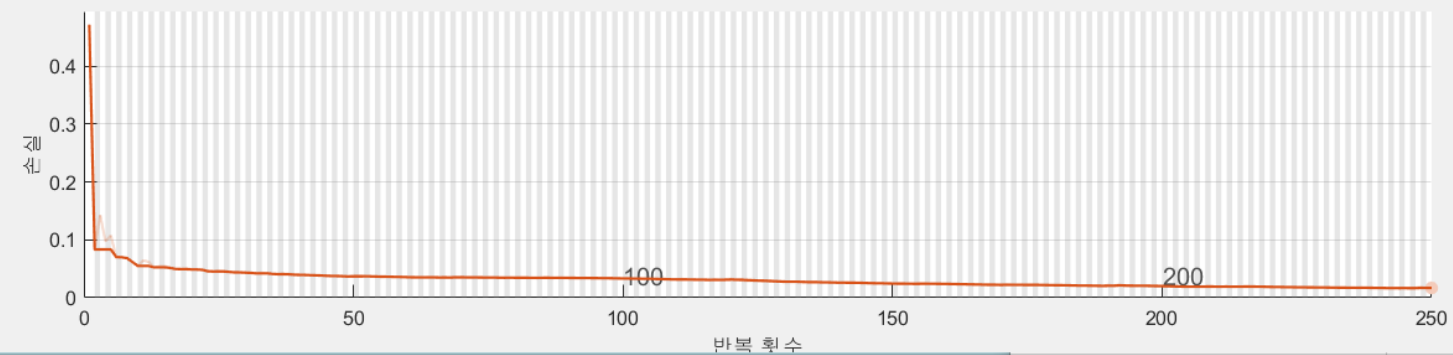
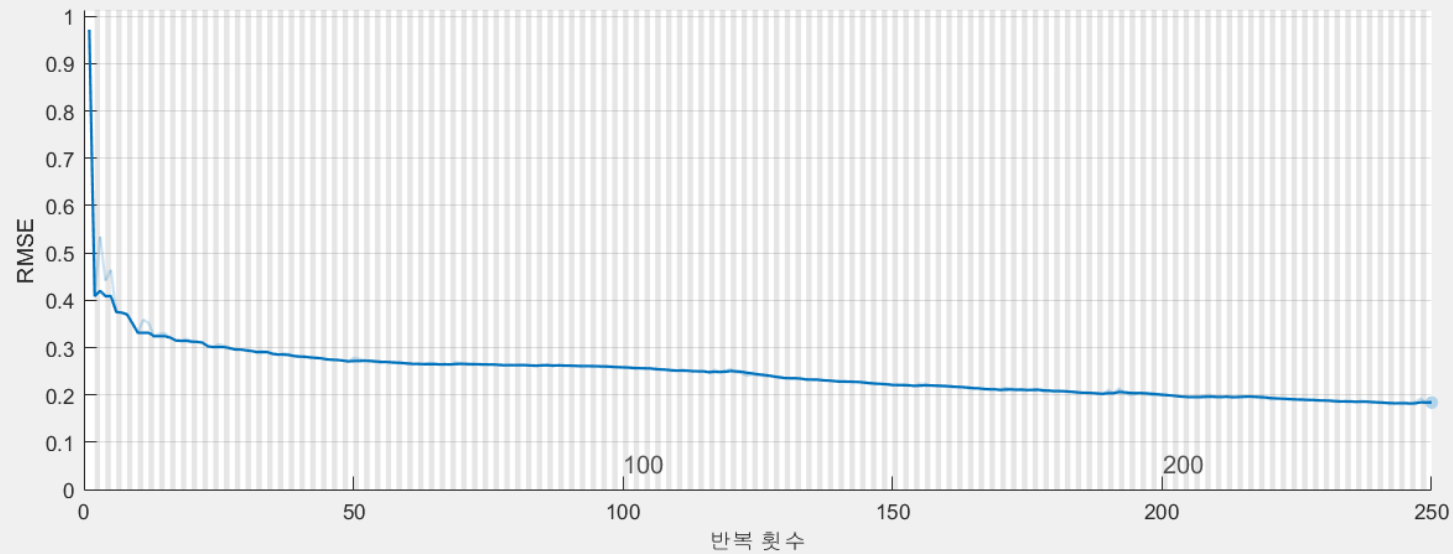
데이터 예측

LSTM Net Architecture 모델 생성 및 모델 학습

```
%LSTM Net Architecture 정의
numFeatures = 1;
numResponses = 1;
numHiddenUnits = 200;
layers = [ ...
    sequenceInputLayer(numFeatures)
    lstmLayer(numHiddenUnits)
    fullyConnectedLayer(numResponses)
    regressionLayer];
options = trainingOptions('adam', ...
    'MaxEpochs',250, ...
    'GradientThreshold',1, ...
    'InitialLearnRate',0.005, ...
    'LearnRateSchedule','piecewise', ...
    'LearnRateDropPeriod',125, ...
    'LearnRateDropFactor',0.2, ...
    'Verbose',0, ...
    'Plots','training-progress');
%학습
net = trainNetwork(XTrain,YTrain,layers,options);
```

데이터 예측

LSTM Net Architecture 모델 학습 Progress (Epoch :250)



데이터 예측

```
dataTestStandardized = (dataTest - mu) / sig;  
XTest = dataTestStandardized(1:end-1);  
net = predictAndUpdateState(net,XTrain);  
[net,YPred] = predictAndUpdateState(net,YTrain(end));  
  
%   테스트 기간 만큼 예측  
numTimeStepsTest = numel(XTest);  
for i = 2:numTimeStepsTest  
    [net,YPred(:,i)] = predictAndUpdateState(net,YPred(:,i-1),'ExecutionEnvironment','  
end  
  
%   데이터 표준화 해제  
YPred = sig*YPred + mu;  
YTest = dataTest(2:end);  
rmse = sqrt(mean((YPred-YTest).^2))
```

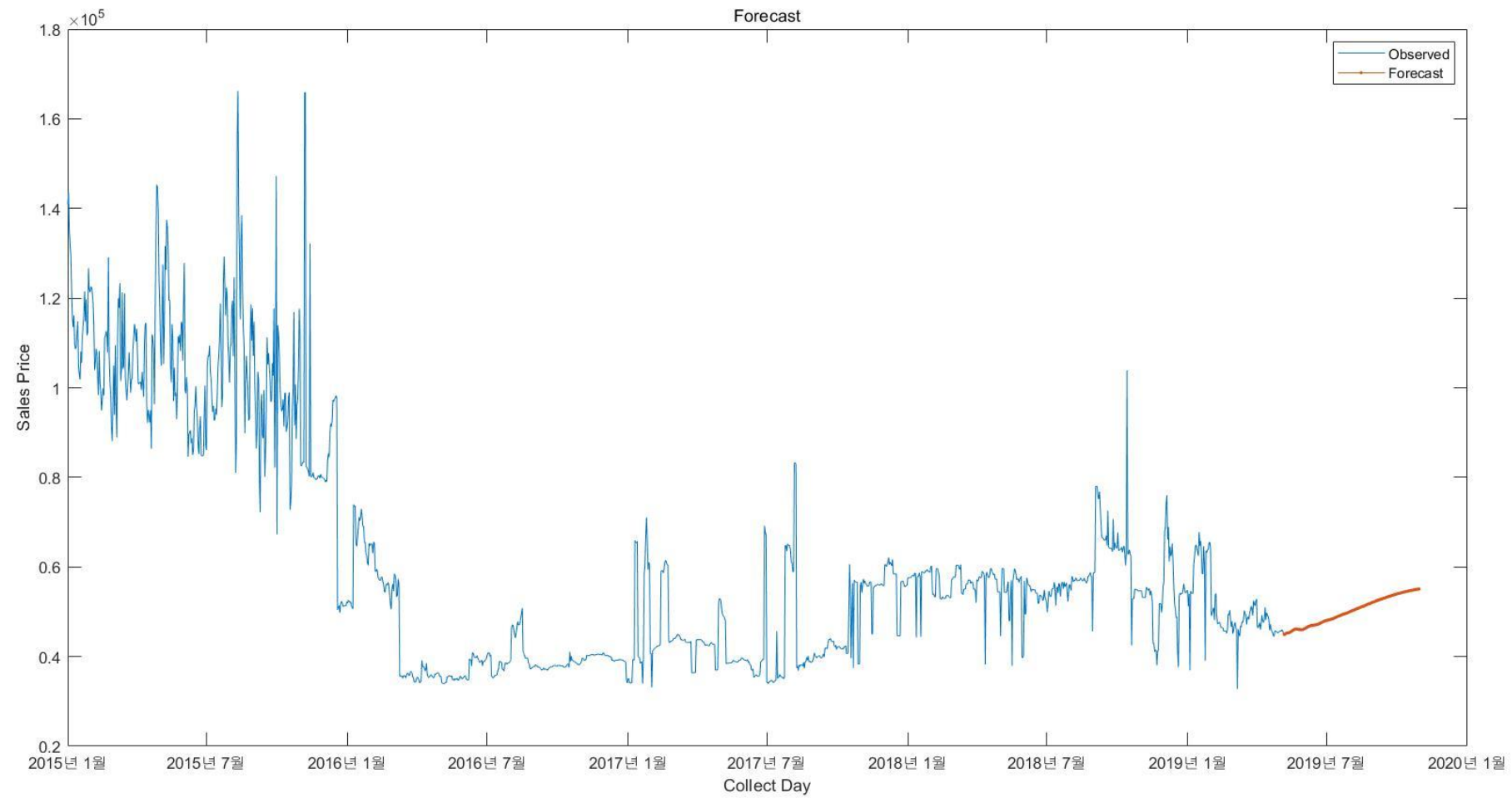
학습 후 모델 초기화 및 예측

데이터 예측

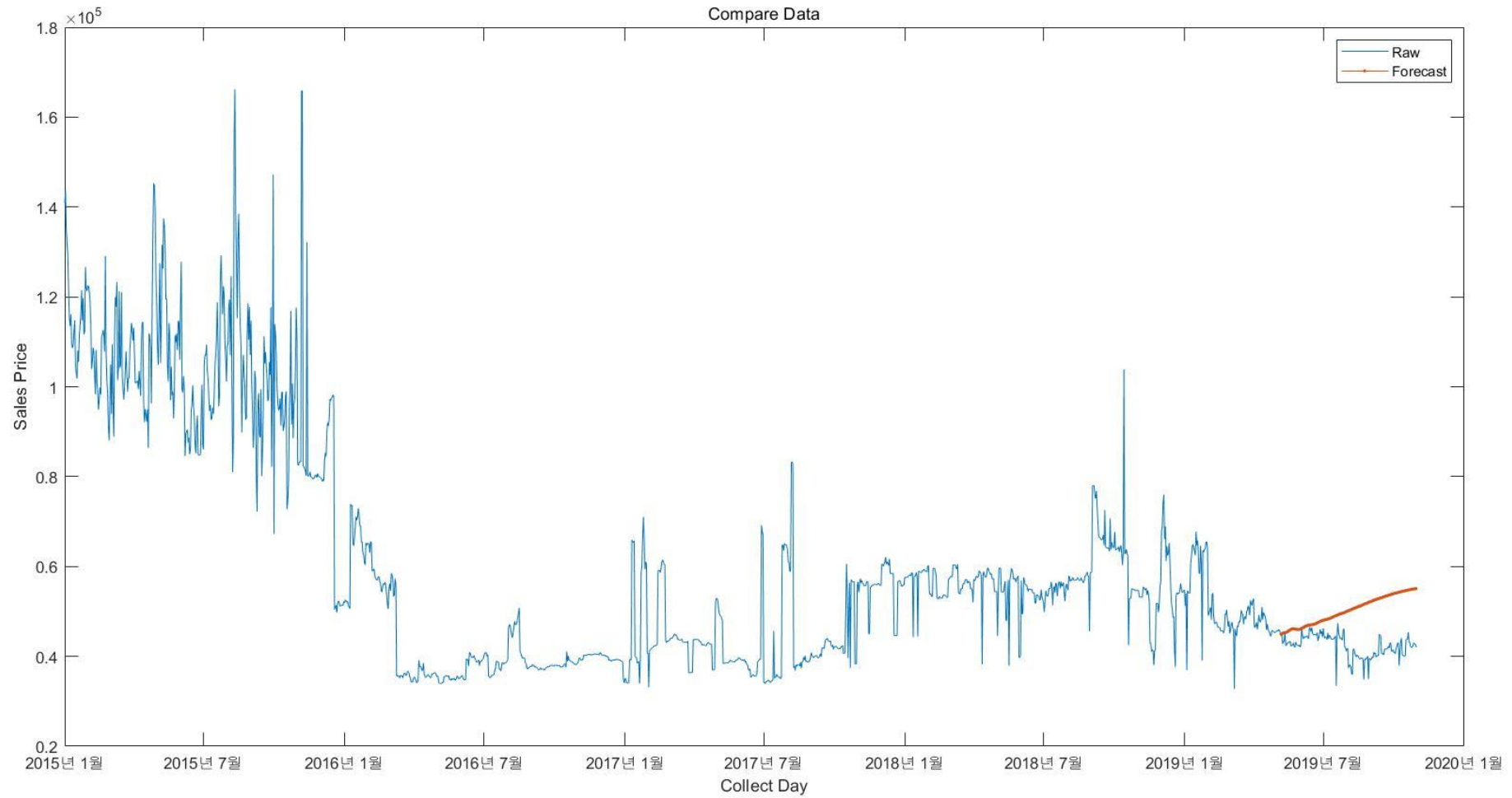
```
% X Label : Collect Day 준비
x_data = datetime(jean_data.collect_day);
x_train = x_data(1:numTimeStepsTrain+1);
x_train = x_train';
x_pred = x_data(numTimeStepsTrain:numTimeStepsTrain+numTimeStepsTest);

% 훈련 + 예측 그래프 그리기
figure
plot(x_train(1:end-1),dataTrain(1:end-1))
hold on
plot(x_pred,[data(numTimeStepsTrain) YPred],'.-')
hold off
xlabel("Collect Day")
ylabel("Sales Price")
title("Forecast")
legend(["Observed" "Forecast"])
```

데이터 예측



데이터 예측

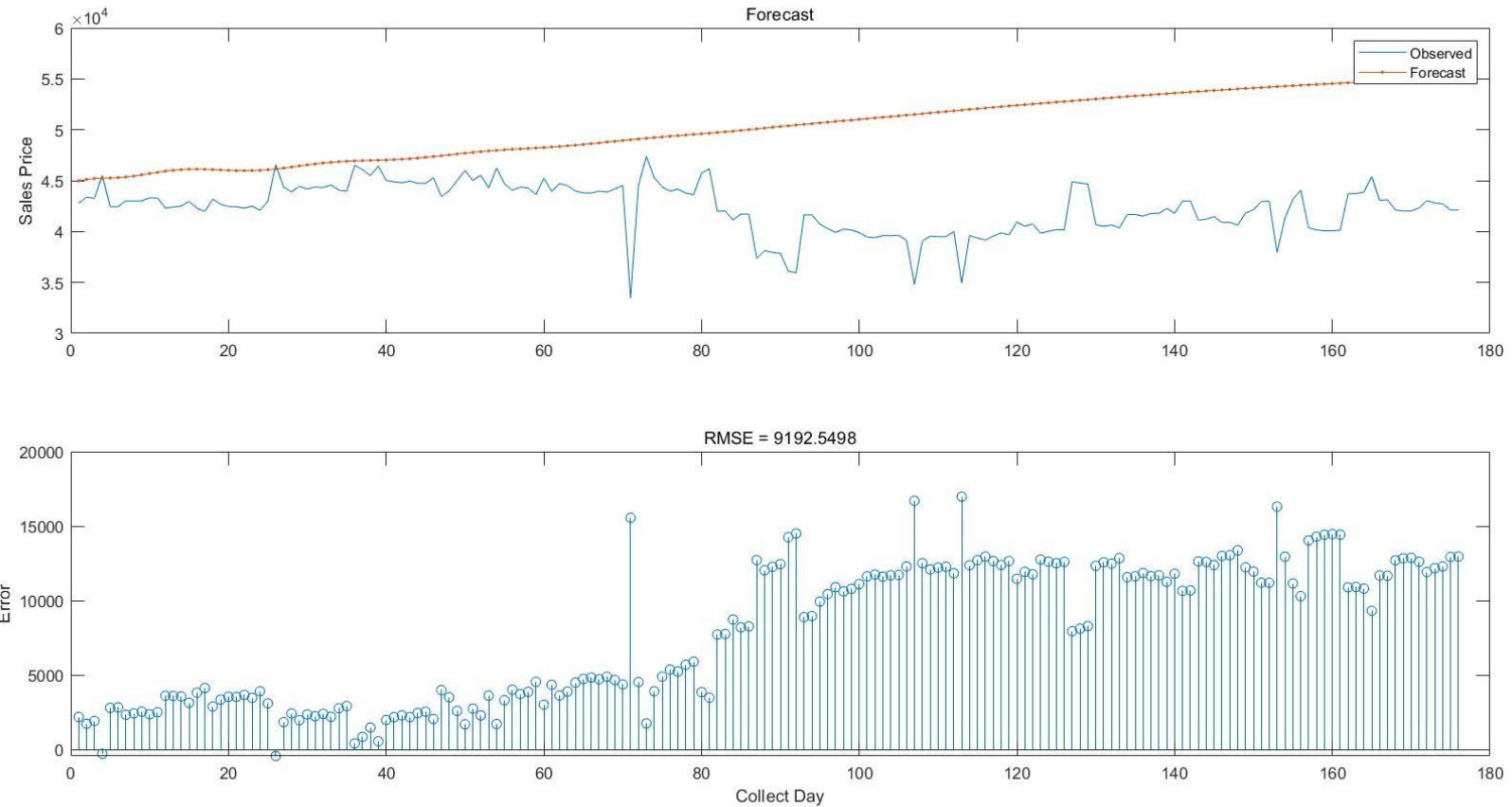


데이터 예측

% 예측과 실제 값 비교 및 RMSE 평가

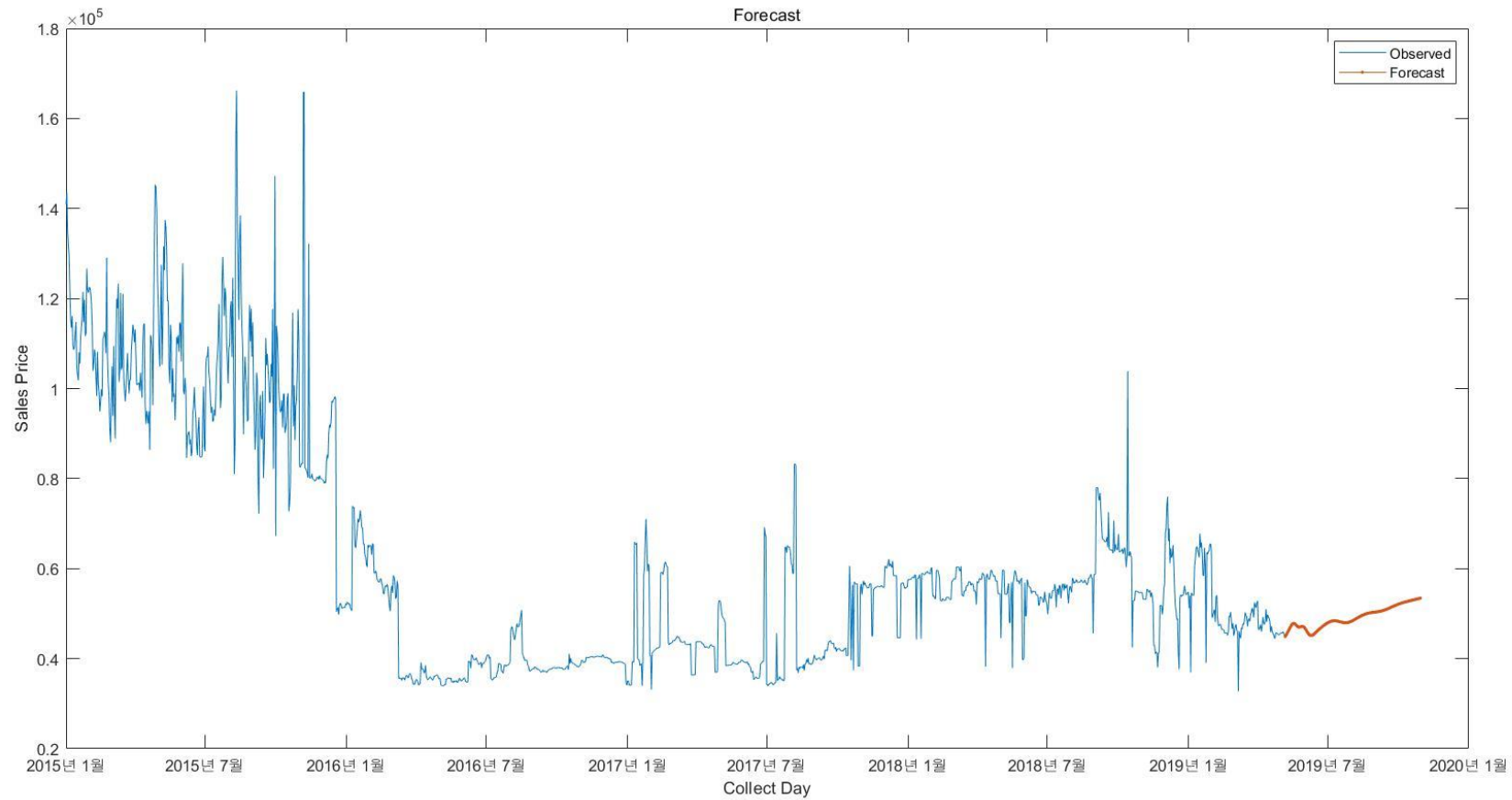
```
figure
subplot(2,1,1)
plot(YTest)
hold on
plot(YPred, 'r')
hold off
legend(['Obs'
ylabel("Sales Price")
title("Forecast")

subplot(2,1,2)
stem(YPred - YTest)
xlabel("Collect Day")
ylabel("Error")
title("RMSE")
```

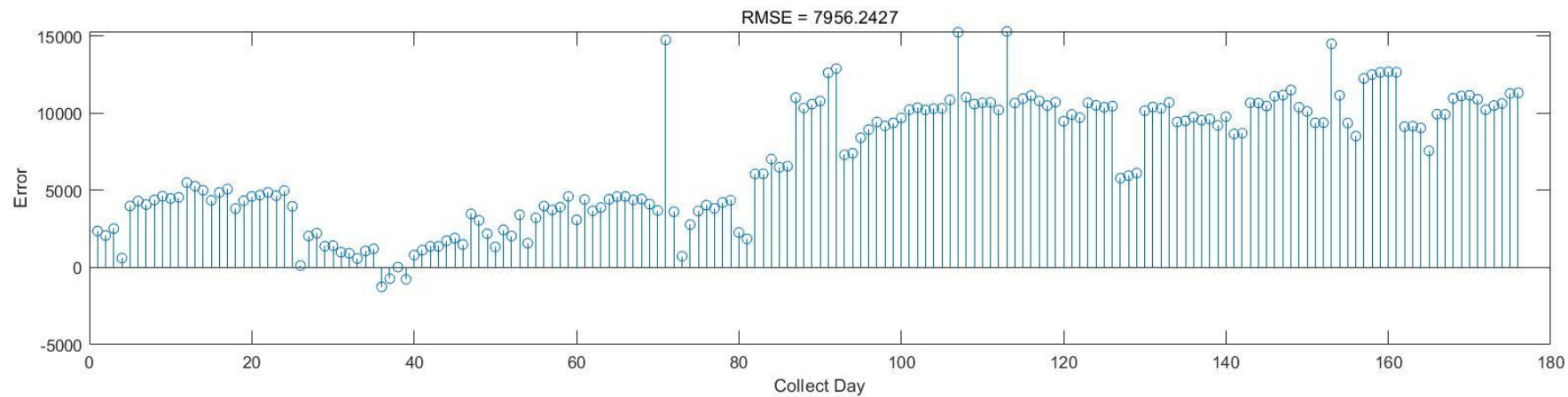
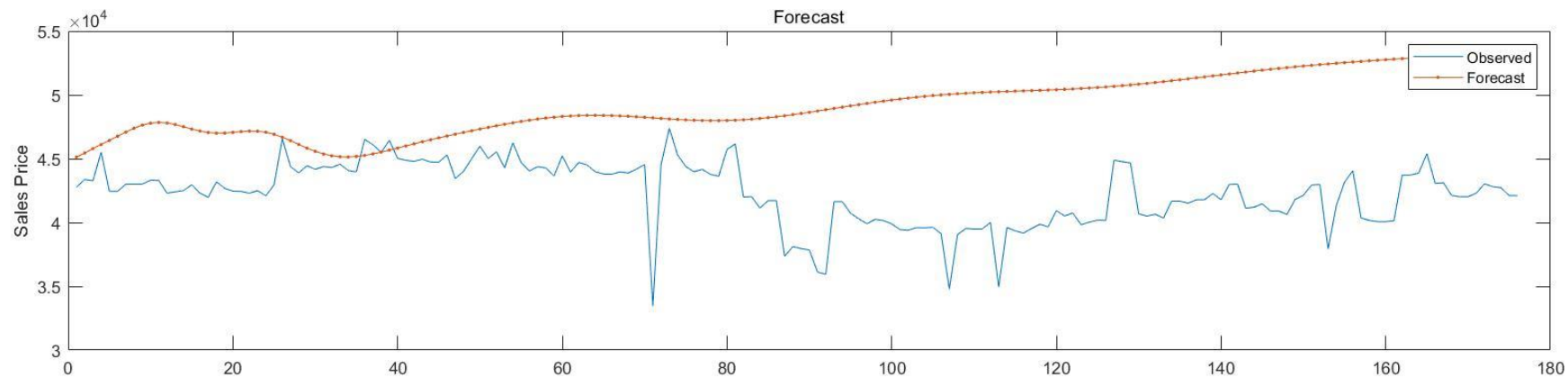


데이터 예측

```
options = trainingOptions('adam', ...  
    'MaxEpochs', 500, ...
```



데이터 예측



데이터 예측의 의미

하나의 품목인 '청바지' 가격만을 예측하였지만,
이를 더 시간을 투자하여 모든 품목, 모든 데이터로 확대
한다면, 주식 예측과 같이 상품 시세도 충분히 예측을 할
수 있다. 라고 생각하였습니다.

이를 통해 사고 싶은 물건을 지금 사면 싸게 살 수 있
는지, 현재 가격이 싼 편인지 정도를 확인할 수 있습니다.

시간 투자

1주차 : 전체 데이터 정제 시도

2주차 : '청바지 데이터' 정제 (코드 + 결과)

**3주차 : 다양한 회귀분석 및 딥 러닝을 통해 데이터 분석
+ LSTM Net 선정 및 코드 작성, 결과 도출**

이상, 청바지 가격 예측 분석

이었습니다.

감사합니다.