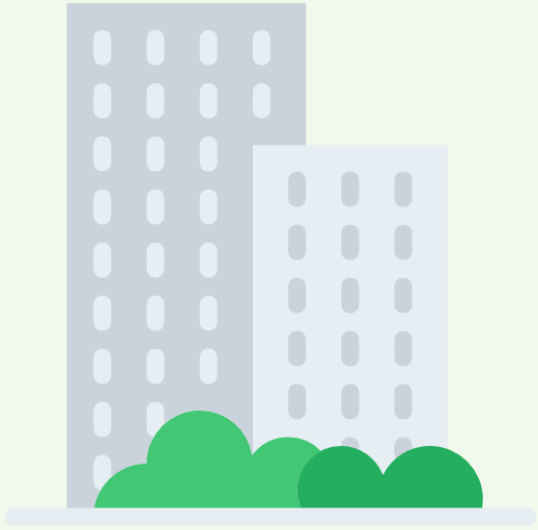
데이터마이닝 프로젝트

인천광역시 아파트 단지 데이터를 이용한 아파트 제곱미터당 평균 매매가격에 영향을 주는 변수 분석

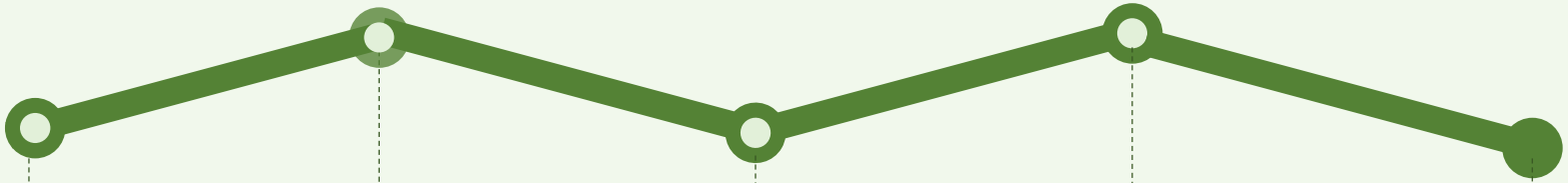


7조

팀장 : 정동호

팀원 : 이건도 , 하나영

목차
CONTENTS



01
서론

주제선정
선정이유

02
Data
handling

데이터 조사
변수설명

03
EDA

EDA 결과

04
Modeling

Data set 구분
의사결정트리
회귀분석

05
결론

분석결과 해석

01. 서론

A) 주제 및 선정 이유

i. 주제

- 인천광역시 아파트 단지 데이터를 이용한 아파트 m^2당 평균 매매가격 예측 및 영향을 주는 변수분석

ii. 분석 단위

- 최근 3년(2016년 1월부터 2018년 12월 까지의 월 기준)
  인천광역시 단지별 아파트 m^2당 평균 매매 가격 기준 (단위 : 만원/ m^2)

iii. 선정 이유

- 부동산 정책에도 불구, 증가하는 아파트 가격 요인이 무엇일까 ?
- 선정 변수들이 아파트가격에 얼마나 영향을 미치고 통제한다면 어떤 변화가 일어날지 ?
- 정부의 부동산 대책의 방향을 제시 및 판단을 분석을 통해 인천 아파트 동향을 알아보고자.

01. 서론

B) 분석 방법

선정한 변수에 대한 EDA 진행

의사결정학습법 이용한 인천 아파트 평균매매 가격 분석

아파트 가격과 관련된 모든 독립변수들에 대한 중회귀분석

02. Data Handling

1) 데이터 조사

- **국토교통부>실거래가 공개시스템** : 아파트 단지별 거래가격, 면적, 건축연도
- **국토교통부>아파트 주거환경 통계** : 인천 대중교통 및 교육시설 인접
- **통계청>국가통계목록** : 미분양 규모, 혼인건수
- **한국감정원** : 실매매가격 변동률
- **부동산114 >REPS 3.0** : 단지비교, 거래건수, 단지규모
- **공공데이터 포털**
- **인천광역시청> 통계정보** : 인천 구 동별 인구수 및 세대수
- **인천교육청> 학교현황** : 초,중,고 현황

02. Data Handling

2) 변수 설명

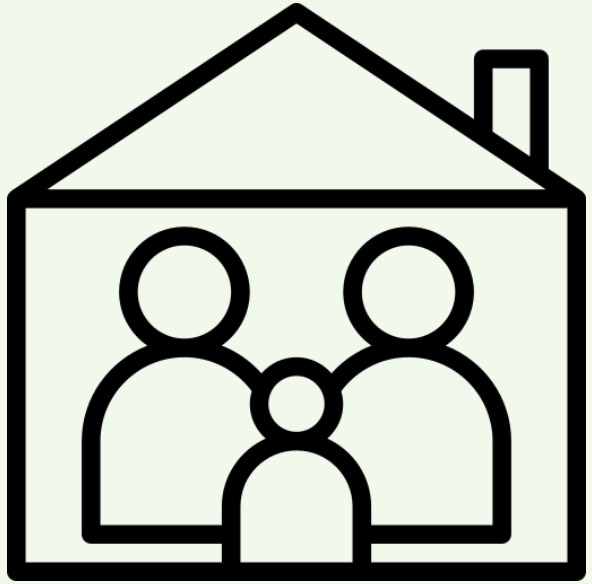
- **종속 변수**

인천 아파트
평균 매매 가격
(단위 : 만원/ m^2)

02. Data Handling

2) 변수 설명

- **독립 변수**
  내부적 요인



평균 전세가

아파트 노후도

분양면적

단지 세대수

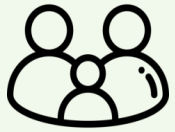
개별 세대수

02. Data Handling

2) 변수 설명

- **독립 변수 - 외부적 요인**

기준금리
공시지가 변동률
부동산 정책

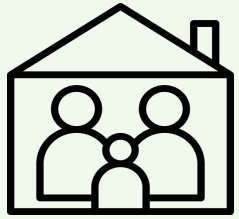


인구수
세대수
혼인수

역세권
개발호재
교육시설
문화,쇼핑,근린시설

02. Data Handling

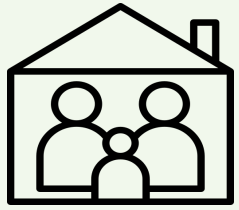
2) 주요 변수 설명



아파트 전세가

- **아파트 전세가격이 상승하면 아파트 매매가격에도 영향을 줄 것이다.**

: 아파트의 월별 제곱미터당 평균 전세가(단위 : 만원/m^2)



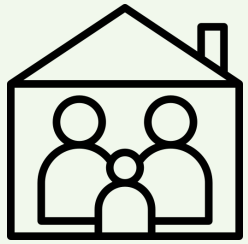
아파트 노후도

- **지어진 지 오래된 아파트일수록 시설이 노후화 되고 낙후화 되어 매매가격이 낮을 것이다.**

:18년12월 기준 아파트가 지어진 연도의 차(연식)

02. Data Handling

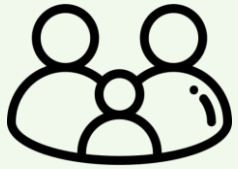
2) 주요변수 설명



분양면적

- **기본적으로 분양면적(평수)가 클수록 아파트 가격이 높을 것이다.**

: 아파트의 전용면적+주거면적 (단위 : m^2)



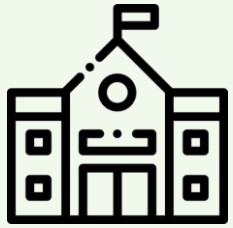
인구수

- **인구 수가 많은 지역은 아파트의 수요가 높을 것이다.**

: 읍면동 인구수의 3년 평균 (단위 : 건수)

02. Data Handling

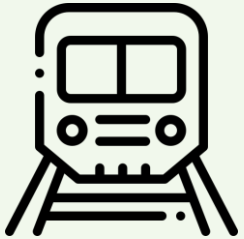
2) 주요변수 설명



교육시설 수

- **학교가 아파트 단지와 가까울수록, 학군이 좋을수록 아파트 가격이 높을 것이다.**

: 읍면동 초,중,고 합계



역세권

- **대중교통의 접근성이 좋은 역세권은 아파트 가격에 중요한 역할을 할 것이다.**

: 해당 읍면동을 지나는 지하철역의 개수

02. Data Handling

2) 주요변수 설명

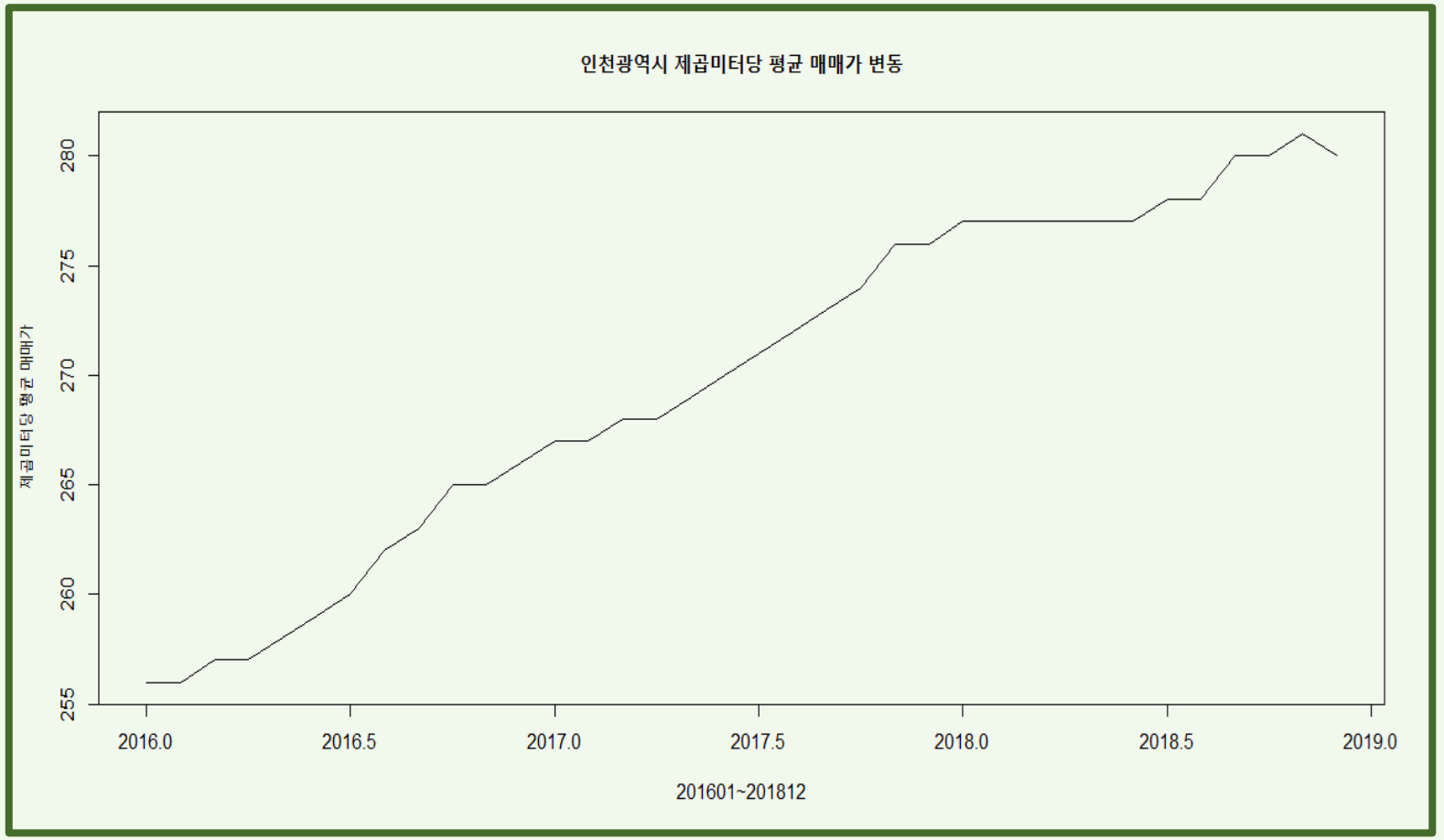
- **신도시개발, 지하철노선 확장, 산업단지 조성 등 개발호재가 아파트 시세에 영향을 줄 것이다.**

개발호재 : 해당 읍면동의 택지개발사업, 교통시설 예정지역 등의 존재 유무

인천광역시 개발계획 : 송도국제신도시, KTX송도역 예정, 영종하늘도시개발, 도화도시개발, 인천2호선과 7호선 연장, 검단신도시개발, 도시재생 뉴딜사업 등

03. EDA

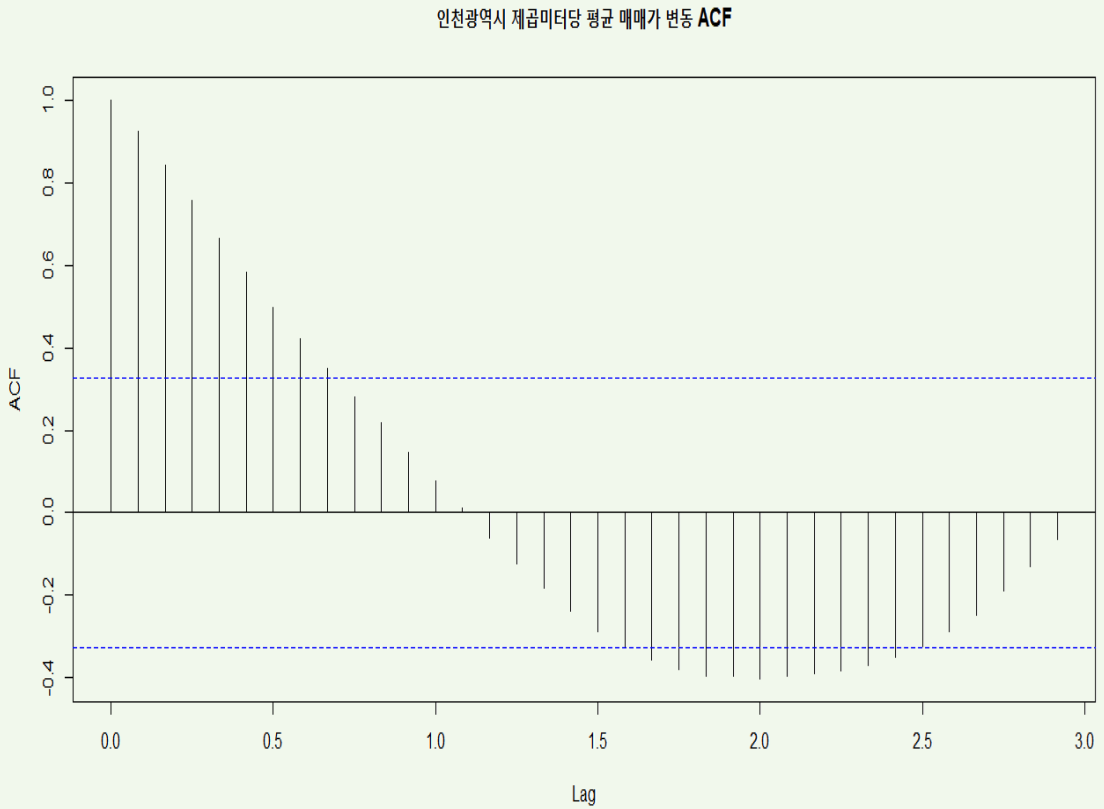
A) 인천광역시 아파트 m^2당 평균매매 가격 변동

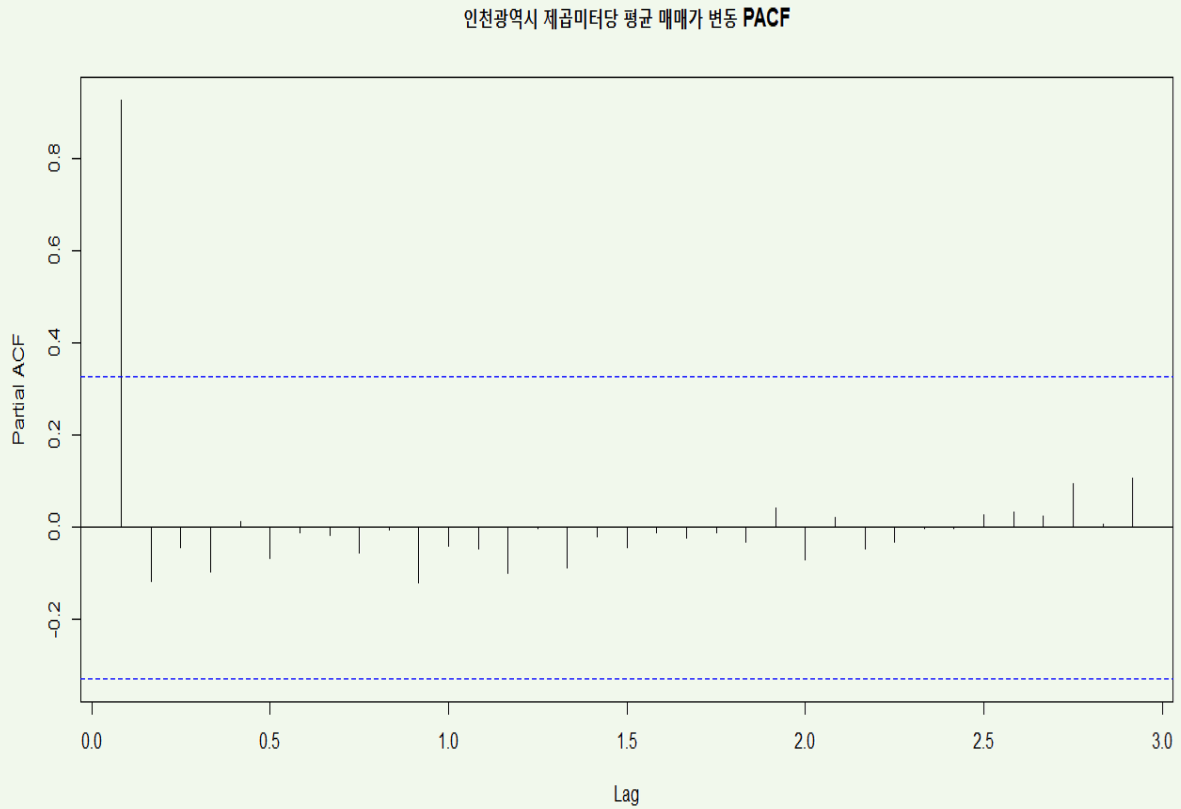


- 전반적인 평균 매매가 상승
- 시계열 분석 예측
- 실제 2019년 1월부터 3월까지의 변동을 비교

03. EDA

A) 인천광역시 아파트 m^2당 평균매매 가격 변동



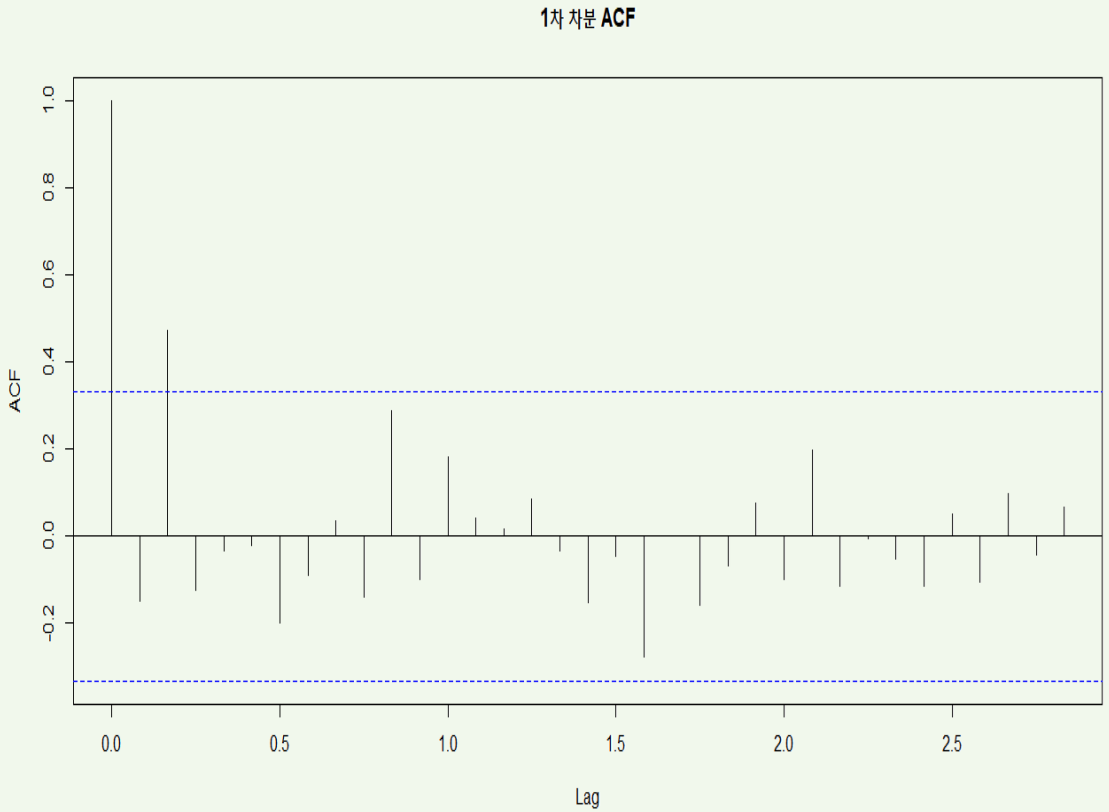


· 정상성을 만족하지 않는 ACF와 lag1 절단값 보이는 PACF.

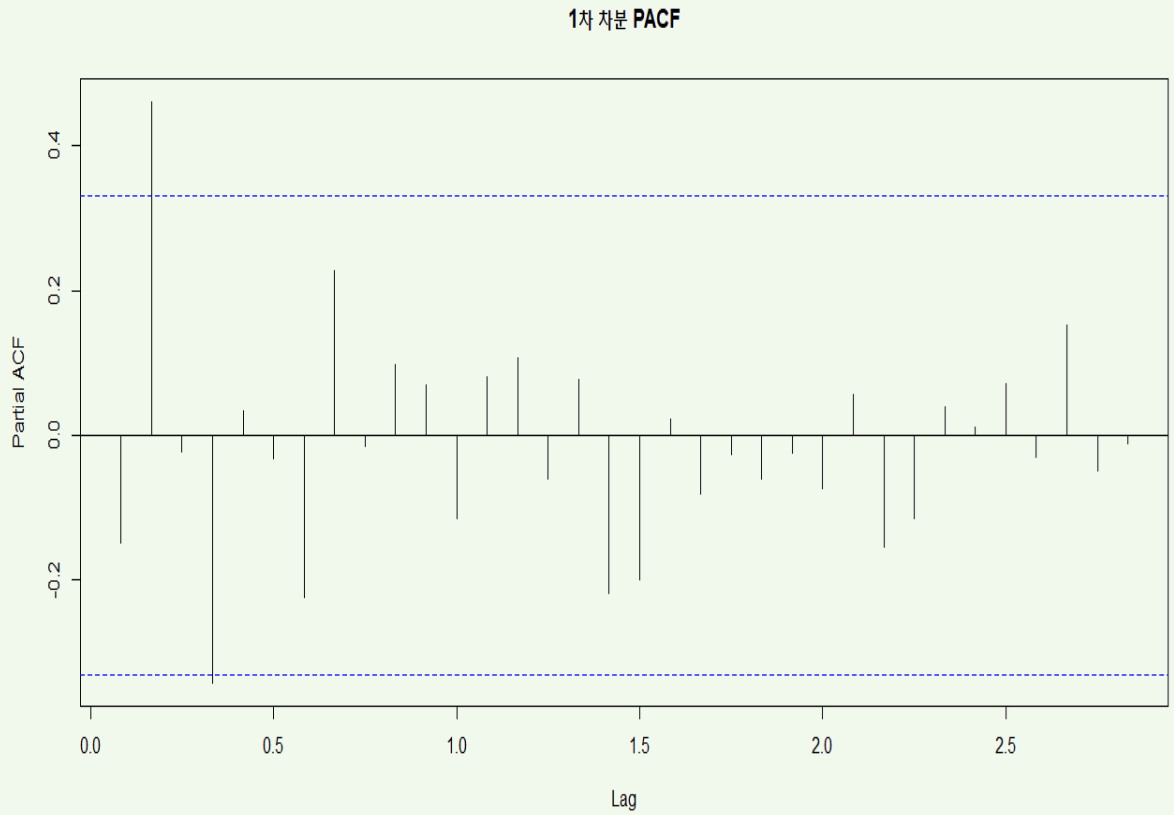
· 1차 차분 진행

03. EDA

A) 인천광역시 아파트 m^2당 평균매매 가격 변동



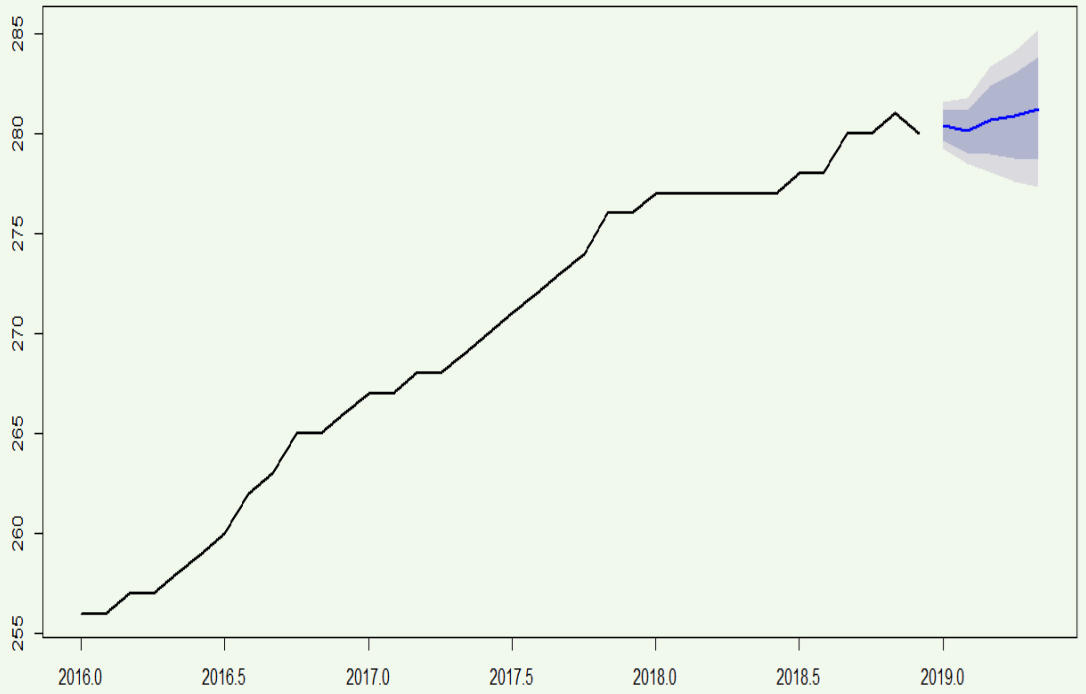
· MA(3) 정상성확인

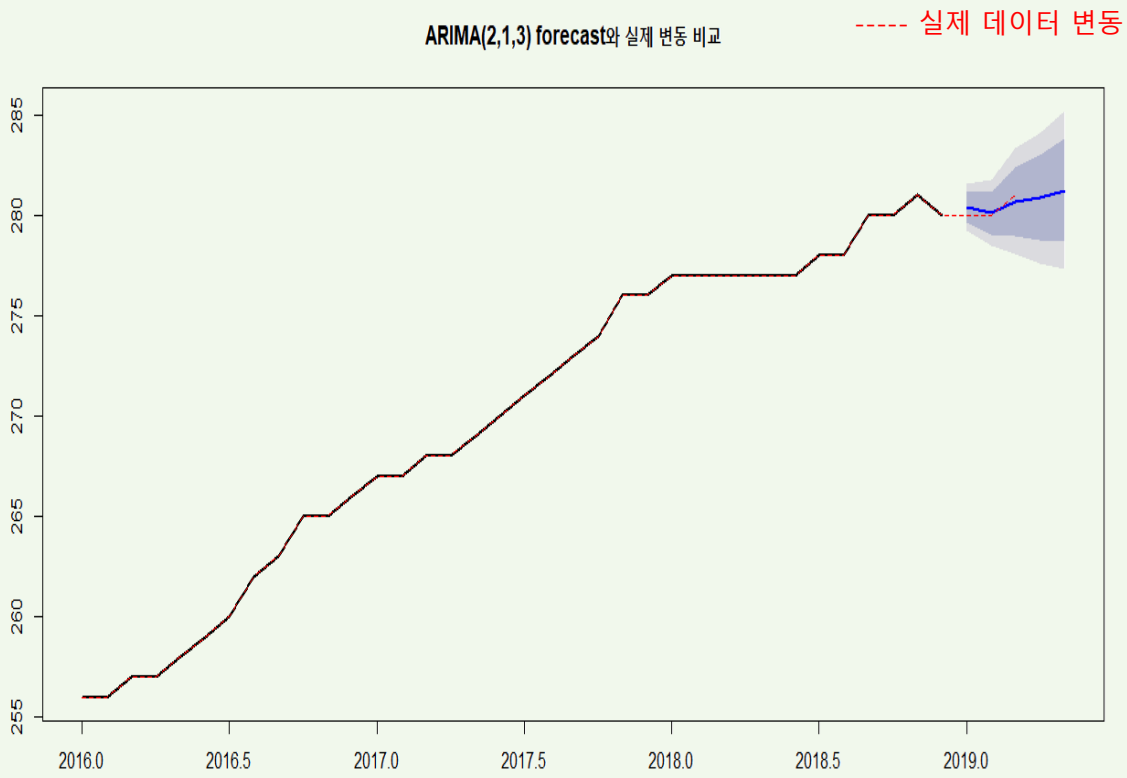


· AR(2) 판단. ARIMA(2,1,3) ?

03. EDA

A) 인천광역시 아파트 m^2당 평균매매 가격 변동



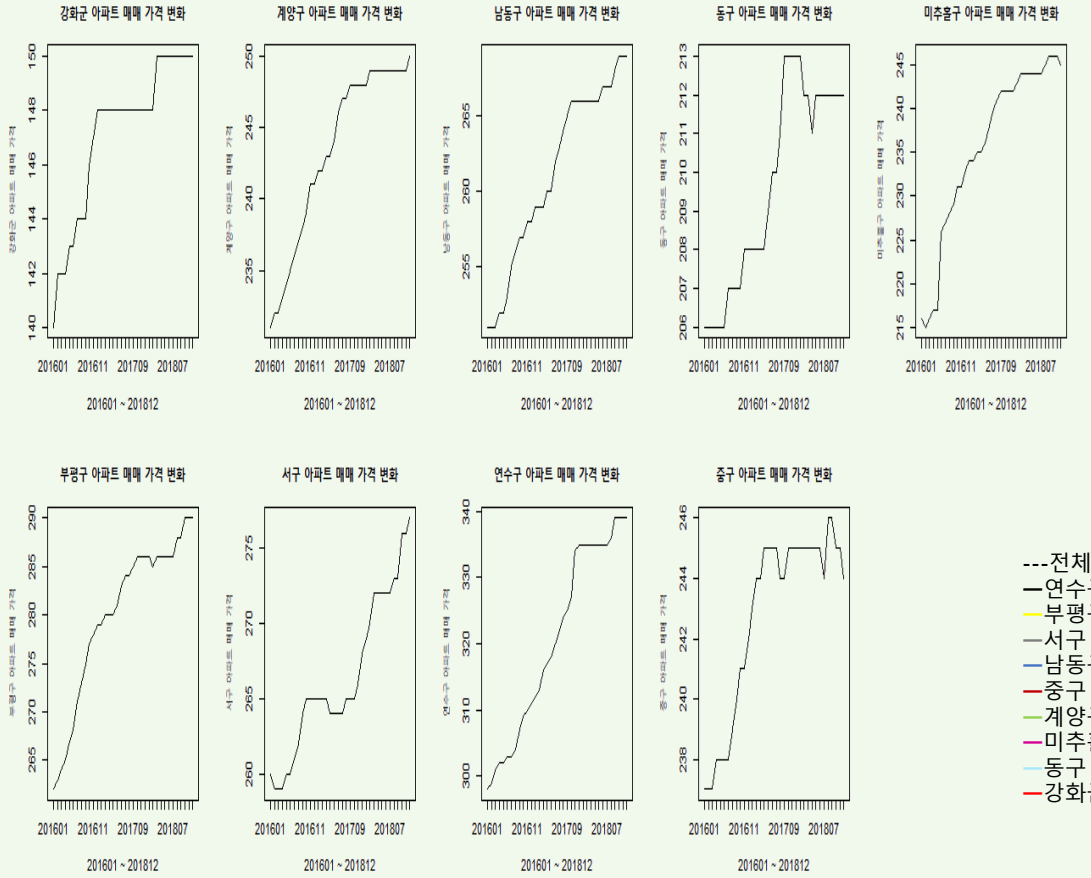


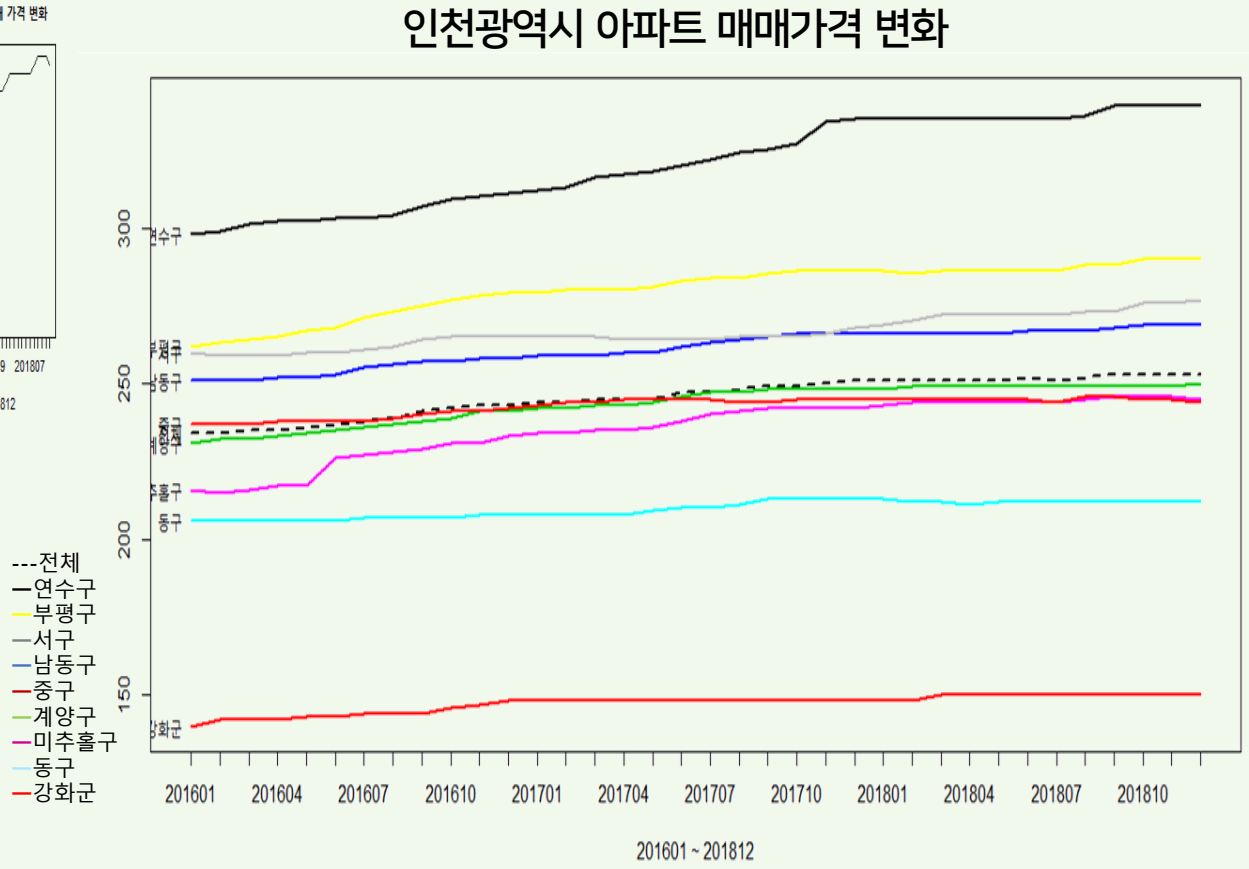
· 2019년에도 지속적인 평균매매가 상승 예상

· 실제 변동과 어느정도 비슷함을 확인

03. EDA

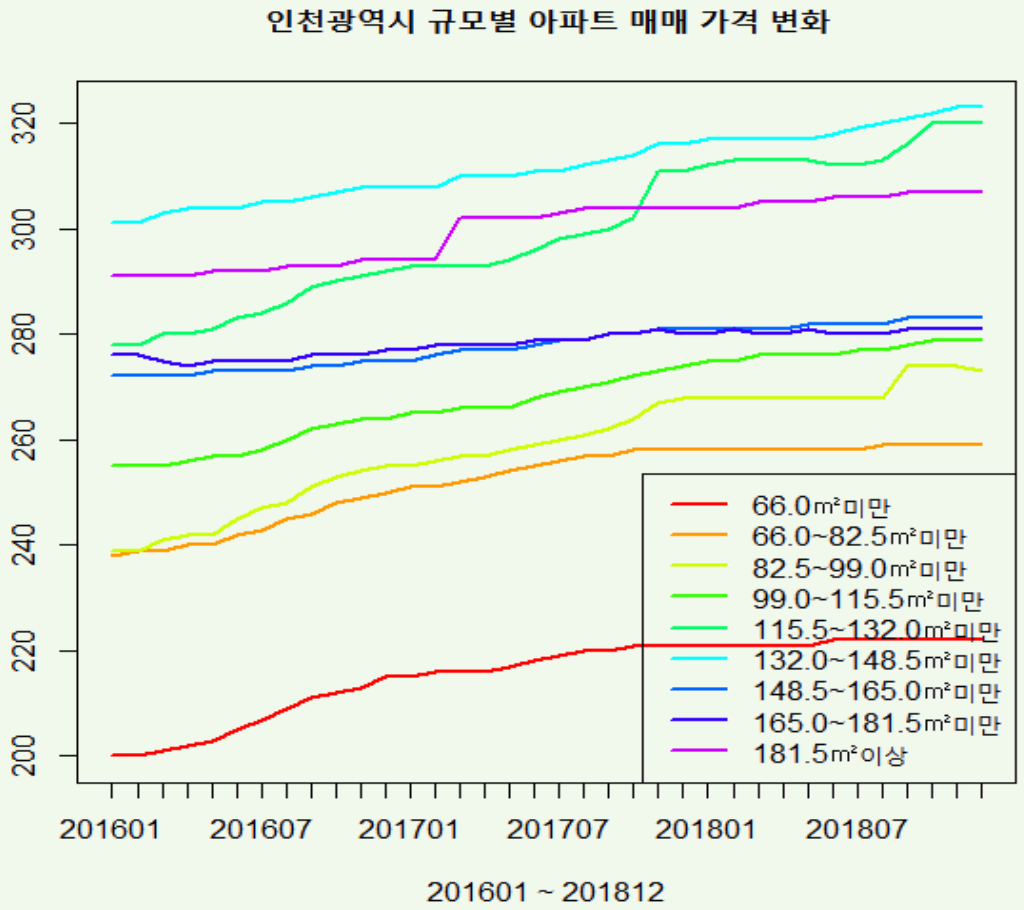
B) 인천광역시 구군별 m^2당 매매 가격 변화

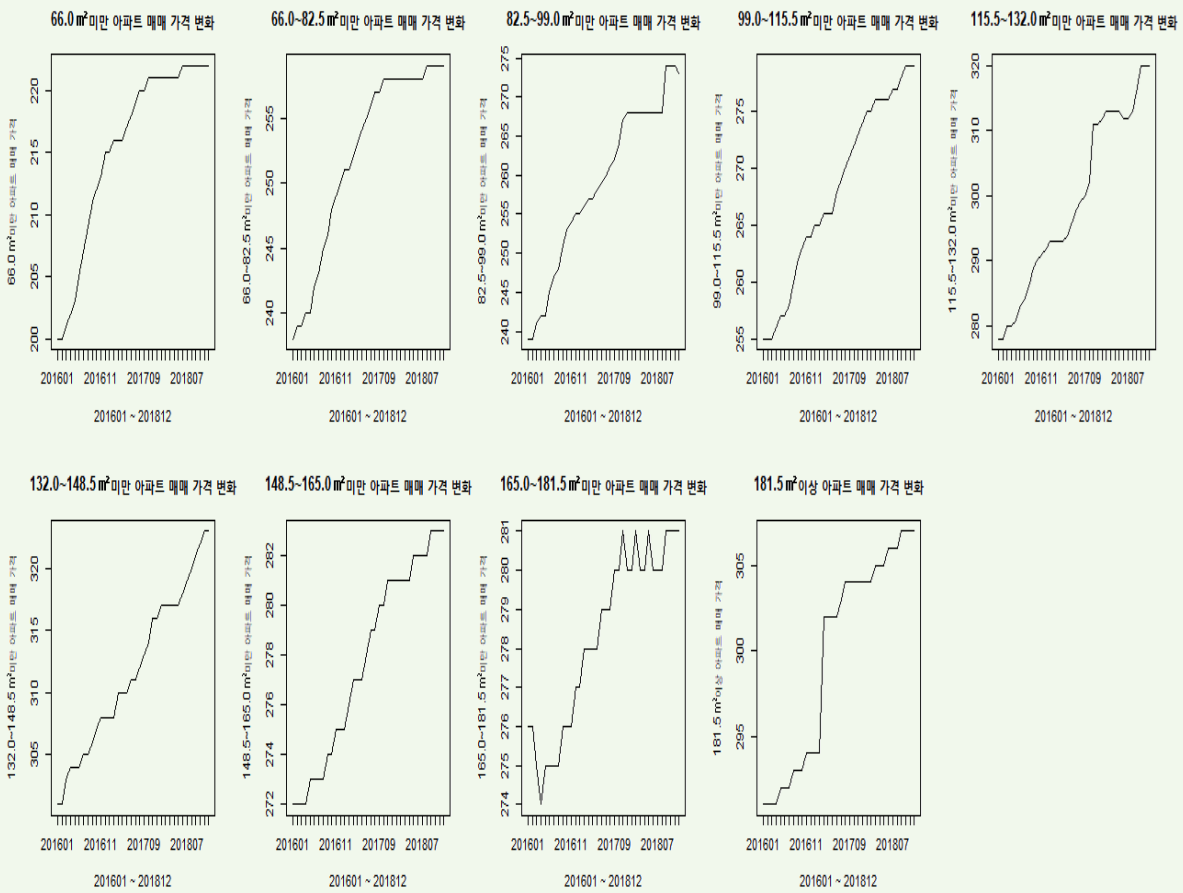




03. EDA

C) 인천광역시 거래규모별 m^2당 평균매매 가격 변동





04. Modeling

1) Data set 구분

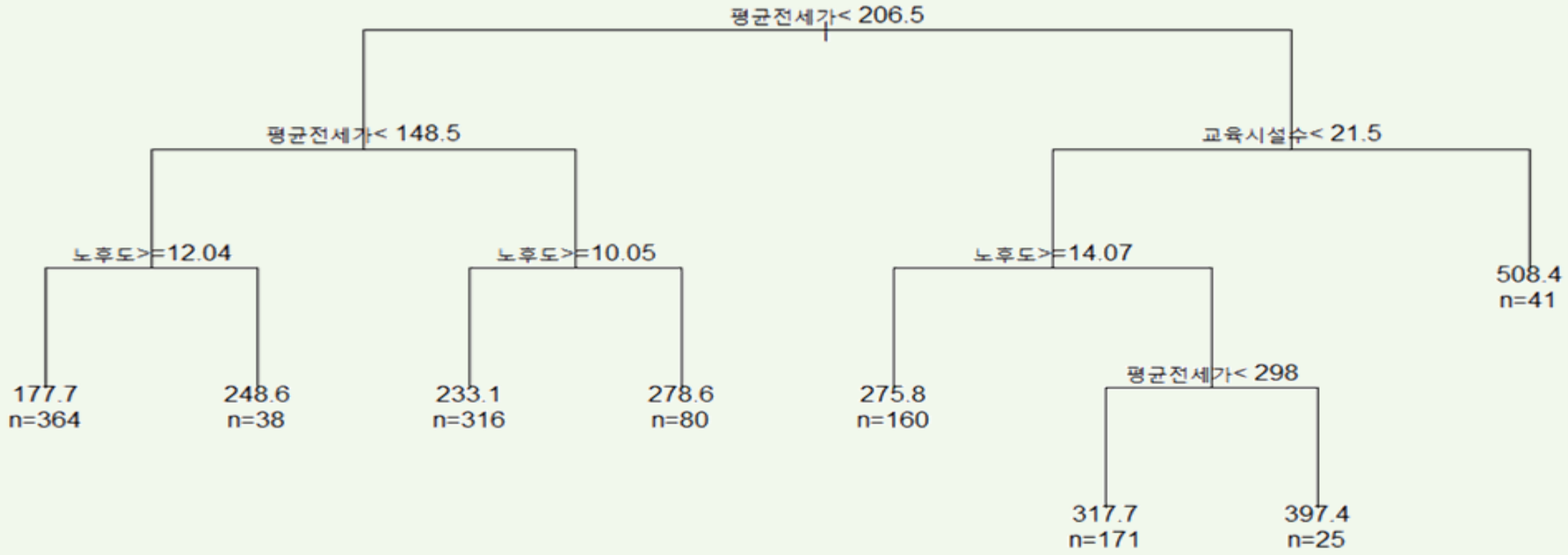
Training Set
1195

Validation Set
895

Test Set
895

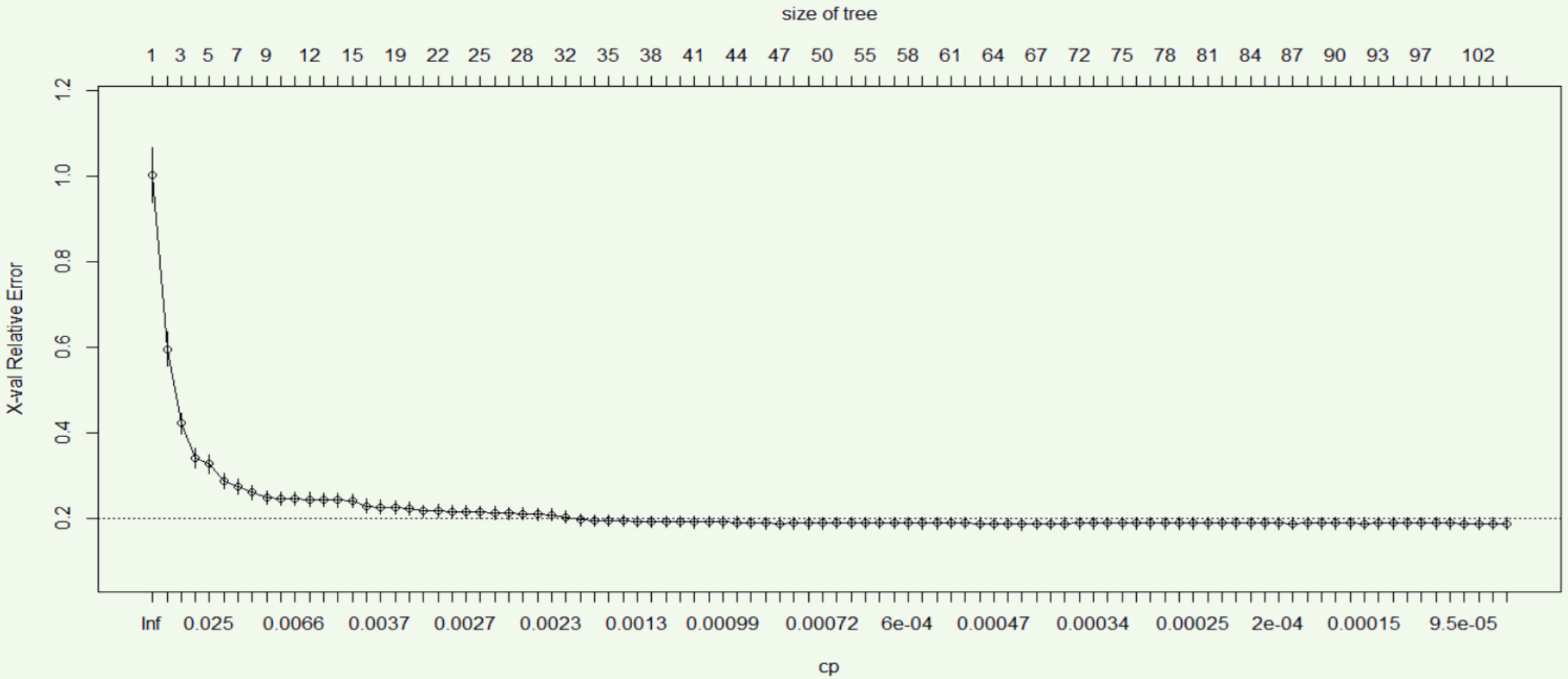
04. Modeling

2) 의사결정트리 - Training



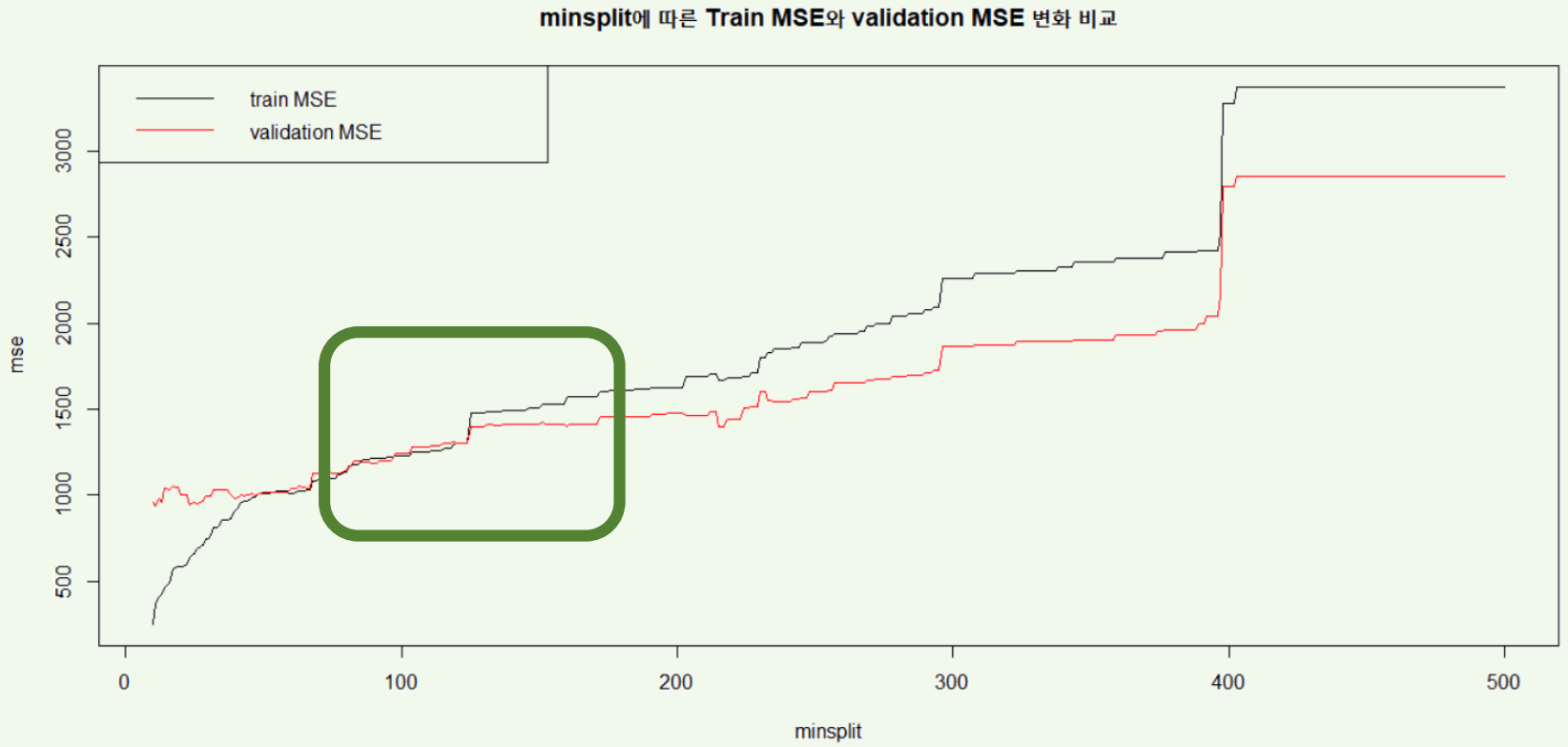
04. Modeling

2) 의사결정트리 – Pruning



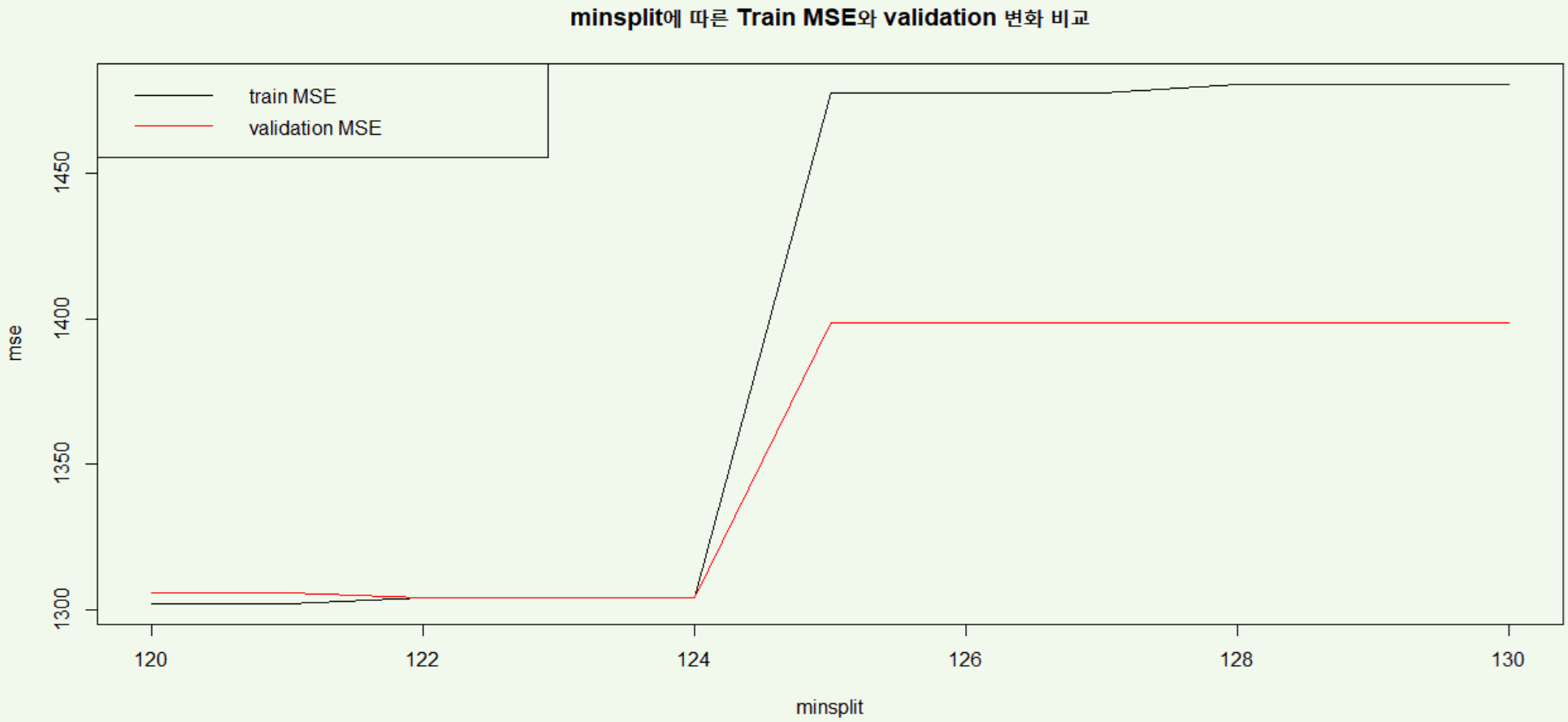
04. Modeling

2) 의사결정트리 – MSE 비교



04. Modeling

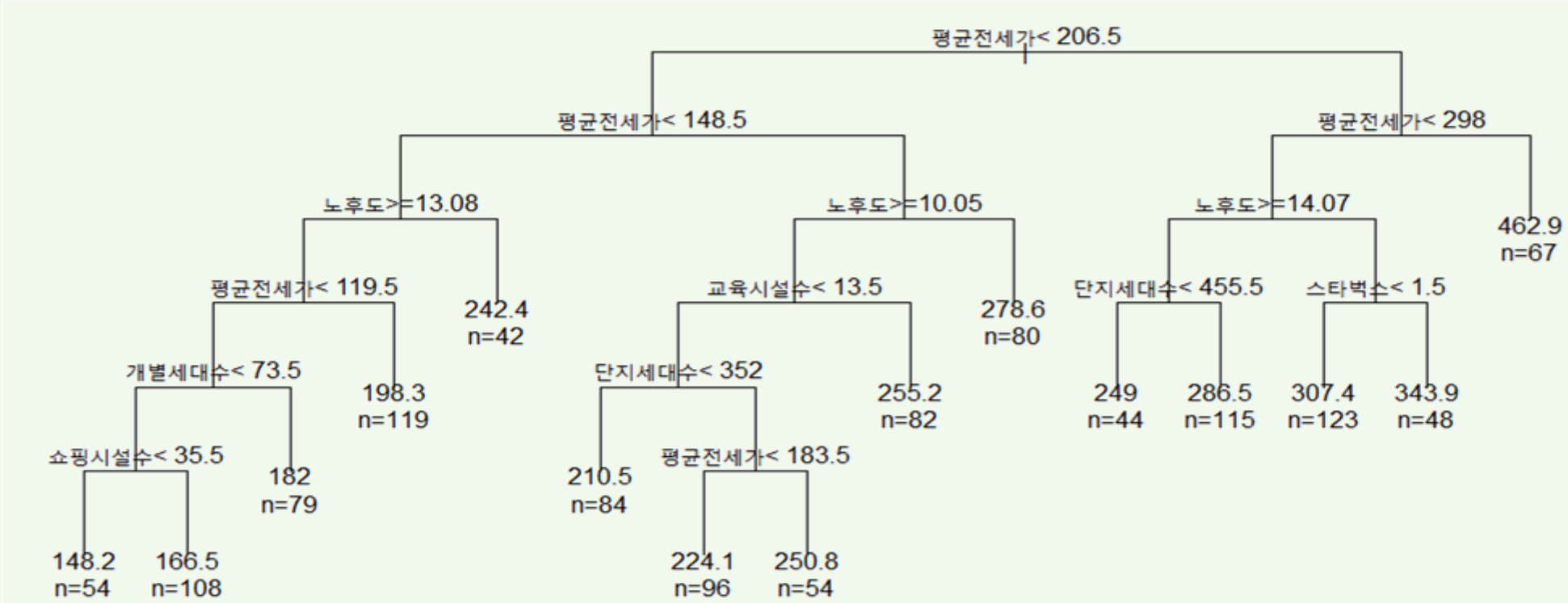
2) 의사결정트리 – MSE 비교



04. Modeling

2) 회귀 의사결정트리 – 최종 모형

인천아파트 평균 매매가 의사결정모형(평균전세가 포함)



04. Modeling

2) 회귀 의사결정트리 – 최종 모형

Train MSE	1477.903
Validation MSE	1398.396
Test MSE	1303.861

변수명	분기 노드 수
평균전세가	5
노후도	3
단지세대수	2
교육시설 수 , 스타벅스, 개별세대수, 쇼핑시설 수	1

04. Modeling

3) 회귀분석 – 초기모형 선택

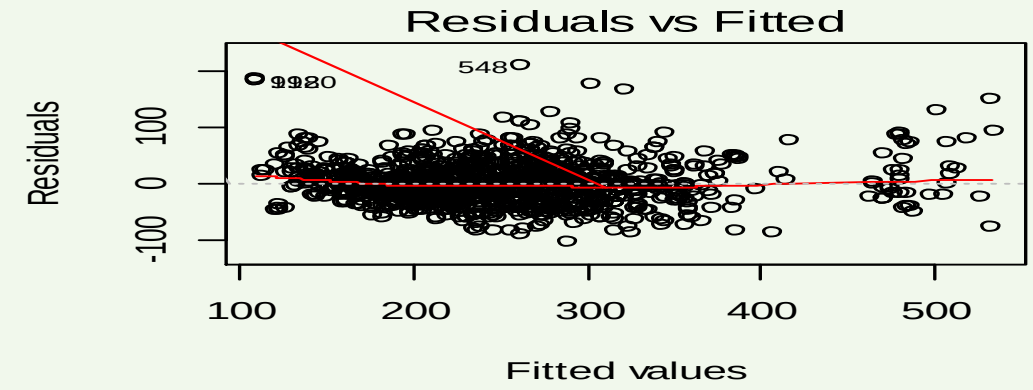
종속변수 Y = 인천아파트 평균 매매가

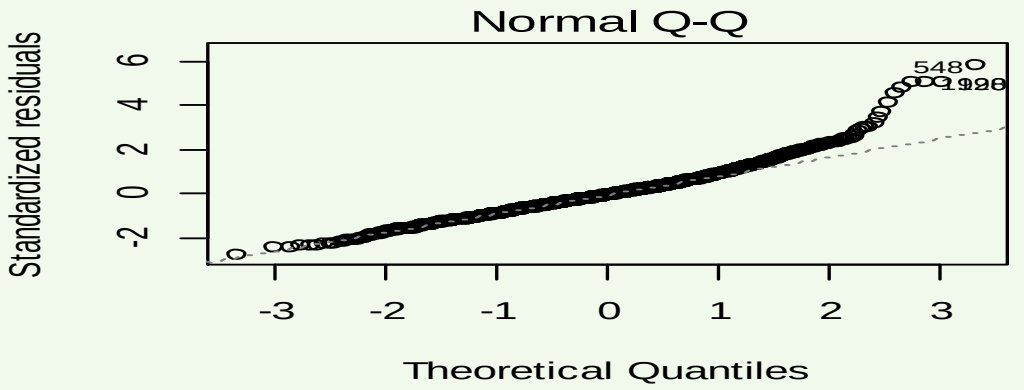
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} + \beta_{15} X_{15} + \varepsilon$$

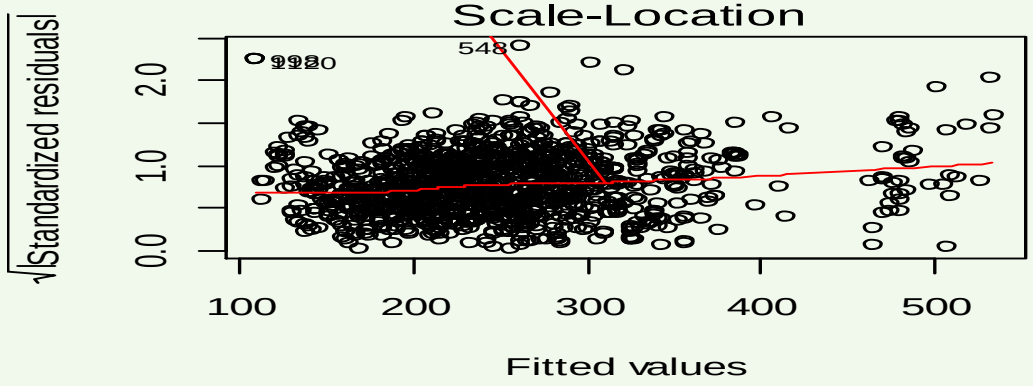
독립변수	설명	독립변수	설명	독립변수	설명
X_1	분양면적	X_6	평균전세가	X_{11}	평균인구수
X_2	단지세대수	X_7	토지면적	X_{12}	평균세대수
X_3	개별세대수	X_8	문화시설수	X_{13}	교육시설수
X_4	노후도	X_9	쇼핑시설수	X_{14}	개발호재
X_5	역세권점수	X_{10}	스타벅스	X_{15}	평균혼인건수

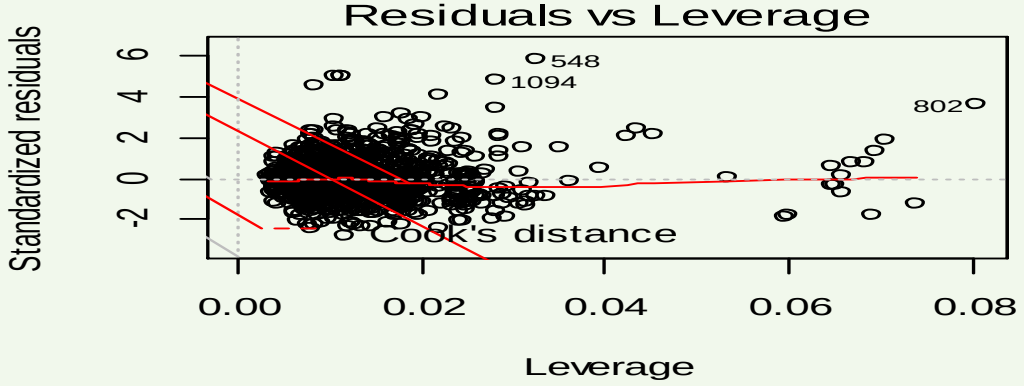
04. Modeling

3) 회귀분석 – 잔차 plot　　정규성, 독립성, 선형성, 등분산성 확인



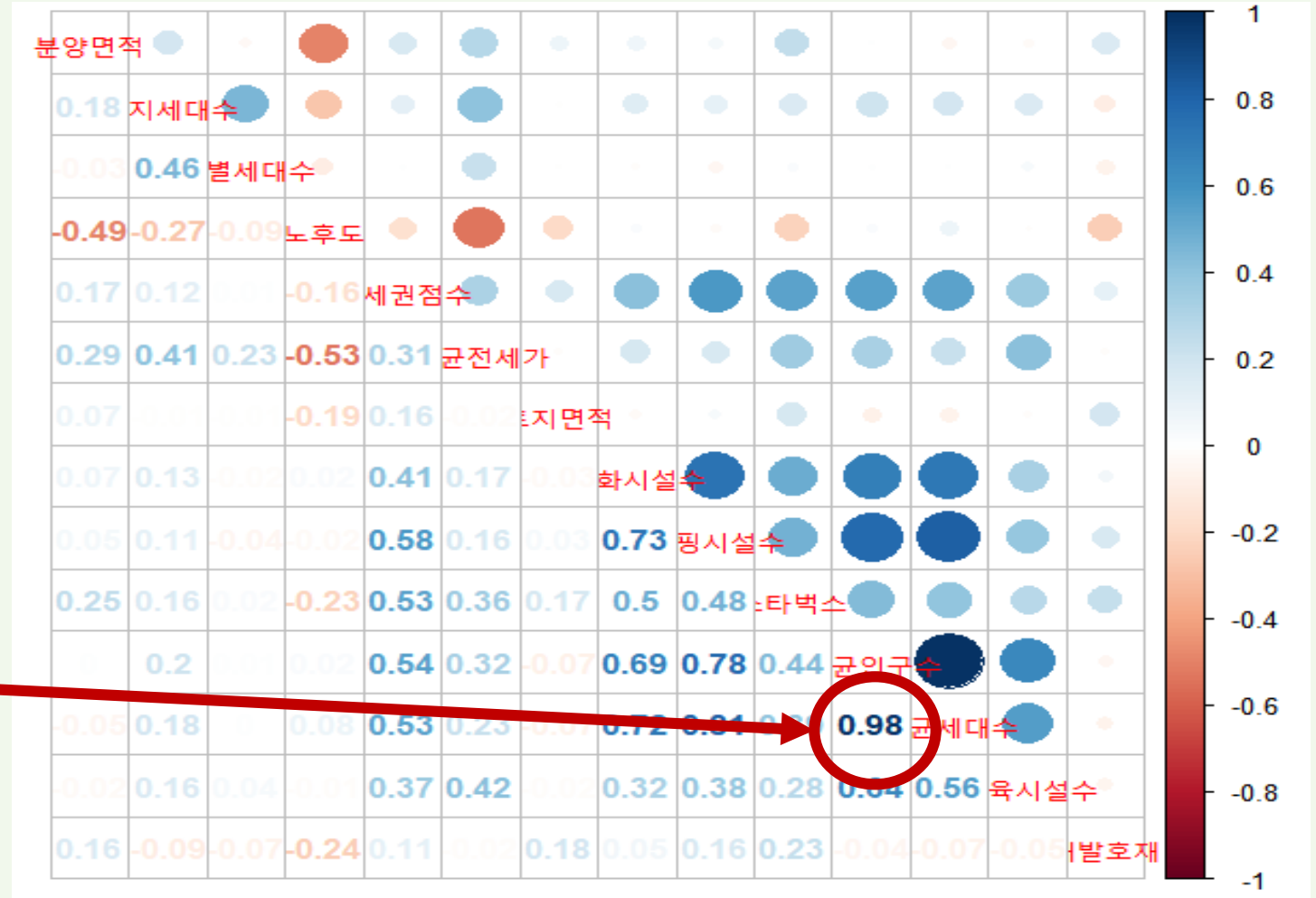






04. Modeling

3) 회귀분석 – 다중공선성 확인



평균 인구수와 평균 세대수에서 다중공선성 나타남!!

04. Modeling

3) 회귀분석 - 변수선택법

1. Forward selection

lm(Y ~ X6 + X4 + X13 + X10 + factor(X14) + X5 + X9 + X11 + X12 + X2 + X15 + X7 + X8, data = train)

Multiple R-squared: 0.7989, Adjusted R-squared: 0.7967

2. Backward selection

lm(Y ~ X2 + X4 + X5 + X6 + X7 + X8 + X9 + X11 + X12 + X13 + factor(X14) + X15, data = train)

Multiple R-squared: 0.7988, Adjusted R-squared: 0.7968

3. Stepwise selection

lm(Y~ X2 + X4 + X5 + X6 + X7 + X8 + X9 + X11 + X12 + X13 + factor(X14) + X15, data=train)

Multiple R-squared: 0.7988, Adjusted R-squared: 0.7968

Backward와 Stepwise 동일한 결과 -> Stepwise모형 채택!
단, X11과 X12의 다중공선성을 고려하여 최종모형 선택하기

04. Modeling

3) 회귀분석 – 최종모형 선택 과정

회귀모형	Train MSE	Validation MSE
초기 모형	1369.551	1220.949
초기 모형 stepwise	1367.26	1217.055
평균인구수(X_{11}) 제거 초기 모형	1527.122	1347.754
평균인구수(X_{11}) 제거 stepwise	1525.889	1348.721
평균세대수(X_{12}) 제거 초기 모형	1509.502	1328.958
평균세대수(X_{12}) 제거 stepwise	1511.268	1334.331

04. Modeling

3) 회귀분석 – **최종모형 선택**

$$\hat{y} = 137.9609 + 0.0964x_1 + 0.0042x_2 + (-1.9706)x_4 + 6.7173x_5 + 0.6467x_6 + (-0.6610)x_7 + (-0.3640)x_9 + 3.7135x_{10} + 0.0002x_{11} + 2.21412x_{13} + 18.8459x_{14} + (-0.0616)x_{15}$$

변수	이름	추정회귀계수
X_{14}	개발호재	18.8459
X_5	역세권점수	6.7173
X_{10}	스타벅스	3.7135
X_{13}	교육시설수	2.21412
X_6	평균전세가	0.6467
X_1	분양면적	0.0964
X_2	단지세대수	0.0042
X_{11}	평균인구수	0.0002
X_{15}	평균혼인건수	-0.0616
X_9	쇼핑시설수	-0.3640
X_7	토지면적	-0.6610
X_4	노후도	-1.9706

05. 결론

A) 모형 평가

i. **Decision Tree**

- 초기모형 8개 노드
- MSE기준 모형
- 최종 모형 15개
- 평균전세가

ii. **Regression**

- Stepwise selection
- 다중공선성 제거 (평균세대수)
- 개발호재, 역세권점수, 스타벅스, 교육시설수

05. 결론

B) 기대 효과

· 안정적인 도시의 전세가와 매매가의 연동 성향과 높은 상관관계

· 서울보다 낮은 인천 아파트 재건축과 신도시 신축 아파트들의 매매가에 미치는 영향

· 교통환경이 열악한 인천이나 경기도 지역 역세권의 따라 큰 아파트 가격 양분

· 가맹점없이 본사 직영점으로만 운영되는 스타벅스의 인근 상권 조건

· 지역마다 큰 차이 없는 '개별세대수' 와 '문화시설수 '

→ 의미 있는 변수들을 통해 인천 아파트 매매가격을
2019년 이후로도 어느정도 예측 가능

05. 결론

C) 한계점

i. 분석 전 한계

- 세부적인 아파트 단지별 완벽한 변수 자료들의 분석
- 전문가도 예측 힘든 아파트 가격의 변수들 포함
- 모델링 적용의 학사 수준 분석 방법의 한계
- 최근 3년간의 데이터로 인한 예상치 못한 미래 사건들
- 단지 내 아파트 층, 방향, 구조 등 세부적인 사항
- 건수마다 달라지는 시차를 두고 거래하는 실거래가

ii. 분석 후 한계

- 계속 감소하는 오분율에 따른 적정 terminal node 수준 선정
- 변수제거에도 test set MSE가 validation set의 것보다 큰 것에 대한 회귀분석 해석
- 인천공항이 있는 운서동 아파트 값이 낮은편 이지만 가장 많은 스타벅스 매장
- 송도 같은 특정 신도시의 높은 집값 해석
- cook's distance에서의 outlier 변수 해석

THANK YOU!