Data Mining 7조 최종 보고서

인천아파트 평균매매가 예측과 분석

12132591 정동호

12131820 이건도

12161890 하나영

< 목 차 >

3. 모델링 적용

A. Data Set 구분

B. 모델링 – 의사결정학습법(Decision Tree)

C. 모델링(추가) – 의사결정학습법(Dection Tree)

D. 모델링 – 회귀분석(Regression)

4. 결론

A. 모형평가

B. 기대효과

C. 한계점

1. 기획 단계

A. 주제 및 선정 이유

i. 주제

인천광역시 아파트 단지 데이터를 이용한 인천아파트 제곱미터당 평균 매매 가격 예측 및 영향을 주는 변수 분석

ii. 분석 단위

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

iii. 선정 이유

이 주제의 선택 이유는 두가지입니다. 먼저, 정부의 부동산 정책에도 불구하고 아파트 가격은 눈에 띄는 감소세를 보이지 않고 오히려 증가세를 보이기도 하면서 아파트 가격에 영향을 주는 요인이 무엇인지 궁금하게 되었습니다. 그리고 저희가 선정한 변수와 아파트 가격 간의 관계 분석을 통해서 각 변수 별로 아파트 가격에 얼마나 영향을 주는지, 해당 변수를 통제한다면 무슨 변화가 일어나는지, 그리고 현재 정부의 부동산대책의 방향성을 판단하거나 새로운 대책이 무엇이 될 수 있는지 이러한 분석을 통해 인천 아파트 가격 동향 예측을 통해 알고 싶었습니다.

B. 분석방법

분석 방법은 다음과 같습니다.

- 선정한 변수에 대한 EDA 진행
- 의사결정학습법(Decision tree)을 이용한 인천아파트 제곱미터당 평균 매매 가격 분석
- 인천아파트 제곱미터당 평균 매매 가격과 관련된 모든 독립변수들에 대한 관계를 분석할 수 있는 중회귀분석

C. 데이터 조사

국토교통부>실거래가 공개시스템 : 아파트 단지별 거래가격, 면적, 건축연도(노후도)

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

그 외 구글링 등 인터넷 서치를 통해 정보 수집

D. 변수 설명

분류	변수	변수 설명
종속변수 (목적변수)	평균 매매 가격	실제 인천에 있는 아파트의 제곱미터당 평균 매매 가격과 아파트 값에 영향을 주는 변수들을 이용하여 제곱미터당 평균매매가격을 추정(단위 : 만원/ m^2)
독립변수 (내부요인)	1.평균 전세가 2.분양면적 3.아파트 노후도 4.단지세대수 5.개별세대수	아파트의 월별 제곱미터당 평균 전세가(단위 : 만원/m^2) 아파트의 전용면적+주거면적. 소위 집안 면적(단위 : m^2) 18 년 12 월 기준 아파트가 지어진 연도의 차(연식) 아파트 전체 단지 내 입주하고 있는 세대의 수 아파트 분양면적에 따라 입주하고 있는 세대의 수
(외부요인)	6.역세권 점수 7.교육시설 수 8.문화시설 수 9.쇼핑시설 수 10.토지면적 11.개발호재* **(범주형)** 12.평균 인구수 13.평균 세대수 14.평균 혼인건수 15.스타벅스	해당 읍면동을 지나는 지하철역의 개수 해당 읍면동 초,중,고 합계 해당 읍면동의 영화관, 박물관, 공원, 테마파크 등 문화시설 수 해당 읍면동의 백화점, 마트, 편의점 등 쇼핑시설 수 해당 읍면동의 토지 면적(단위 : km^2) 해당 읍면동의 택지개발사업, 교통시설 예정지역 등의 존재 유무 읍면동 인구수의 3년 평균 (단위 : 건수) 읍면동 세대수의 3년 평균 (단위 : 건수) 구별 혼인건수의 3년 평균 (단위 : 건수) 해당 읍면동 안의 스타벅스 매장 개수 (단위 : 건수) * 제외하고 연속형 변수
시계열분석	. 1.인천 전체 평균 매매가격 2.인천 구군별 매매가격 3.인천 거래규모별 매매가격 4.인천광역시 인구수 5.인천광역시 세대수 6.강남아파트 매매가격 7.공시지가변동률* 8.기준금리	인천광역시 전체 제곱미터당 평균 매매가격(단위 : 만원/m^2) 인천광역시 구군별 제곱미터당 평균 매매가격(단위 : 만원/m^2) 인천광역시 거래규모별 제곱미터당 평균 매매가격(단위 : 만원/m^2) 인천광역시 인구수 변화(단위 : 명) 인천광역시 세대수 변화(단위 : 원) 강남아파트 제곱미터당 평균 매매가격(단위 : 만원/m^2) 1) 월별 : 읍면동의 매월 지가변동 상황 2) 누계 : 〔(당해월 지가지수 / 전년말 지가지수) - 1〕* 100 금리 체계의 기준이 되는 금리

초기 변수	내부요인 -아파트 거래건수 -실거래 가격 -실매매가격 변동률 -실전세가격 변동률 -아파트 면적 -아파트 단지규모 -아파트 노후도 -아파트 미분양규모 -난방방식 -층수(저층, 중층, 고층)	외부요인 -공시지가 변동률 -기준금리 -기준시가 -부동산 정책 (청약제도, 전매제한 등) -개발호재 유무 -역세권 -교육시설(학군, 거리)	외부요인 -건설사 브랜드 -공원 및 문화시설 -도심까지의 거리 -서울집값 -인구수 -세대수 -유동인구 -혼인건수
최종선택한 입력 변수	1. 분양면적 2. 단지 세대수 3. 개별 세대수 4. 노후도 5. 역세권 점수 6. 평균전세가 7. 토지면적 8. 문화시설수	9. 쇼핑시설수 10. 스타벅스 11. 평균인구수 12. 평균세대수 13. 교육시설수 14. 개발호재 15. 평균혼인수	
최종 변수 (회귀분석)	1. 분양면적 2. 단지 세대수 3. 노후도 4. 역세권 점수 5. 평균전세가 6. 토지면적	7. 쇼핑시설수 8. 스타벅스 9. 평균인구수 10. 교육시설수 11. 개발호재 12. 평균혼인수	

E. 예상 기대효과 및 한계

i. 예상 기대효과

분석을 통해 인천 아파트 매매가에 변동을 주는 요인들을 파악하여, 정부의 부동산 대책의 방향성 또는 새로운 정책 대안 등을 제시할 수 있을 것이다. 과제를 하며 집값에 대한 우리가 몰랐던 부분을 배울 수 있을 것 같다.

ii. 예상 한계

한계점으로는 세세한 아파트 단지별로 분석이나, 조사된 변수들의 자료 부족으로 완벽한 예측이 힘들 것이다. 전문가도 예측하기 힘든 집값 예측 조건들의 다양한 외적 변수와 예상치 못한 변수들까지는 전부 포함시키지 못하는 한계점이 있다.

모델링 적용에 있어 학사 수준 분석 방법이 실제 적용하기에는 다소 무리가 있을 수 있다는 한계가 있을 수 있다.

실거래가 경우, 시차를 두고 거래를 하는데 거래건수마다 각각 다르기 때문에 수집하기 힘들다.

최근 3년간 거래된 데이터이기 때문에 예상치 못한 미래의 사건, 사고에 대해서는 예측의 정확도의 한계가 존재할 수 있다.

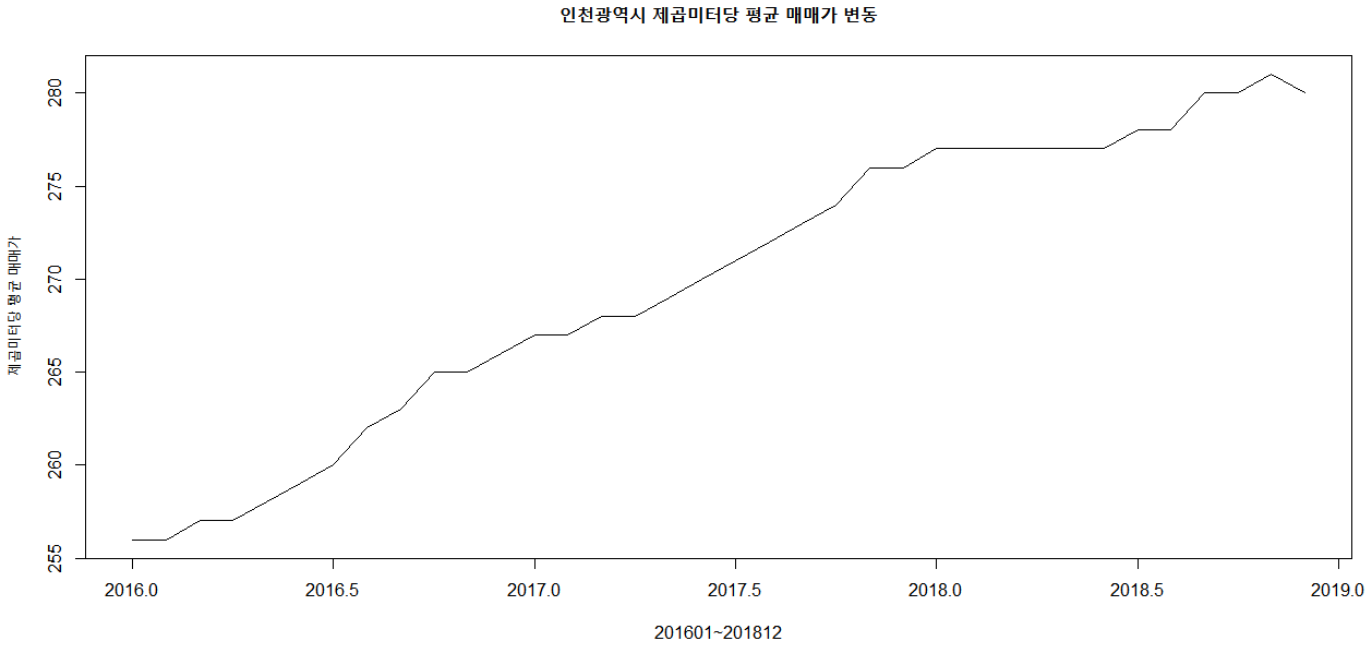
단지 내 아파트에도 층 별, 방향 별 등 세부적인 사항보다는 단지를 기준으로 예측하는 또 다른 변수들이 존재한다.

2. EDA 설명

A. 인천광역시 아파트 제곱미터당 평균 매매 가격 변동

여기서 인천광역시 아파트 제곱미터당 평균 매매 가격은 인천광역시에 있는 모든 평균 매매가 데이터를 바탕으로 한 시계열 데이터이다.

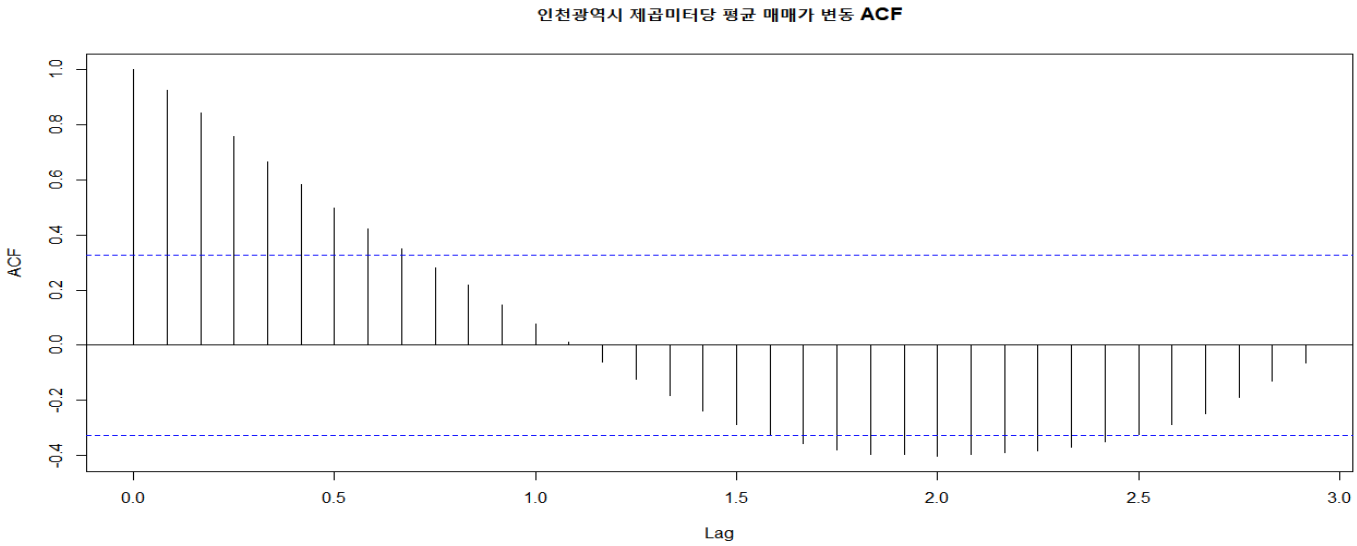
기간은 2016년 1월부터 2018년 12월의 자료를 이용해서 시계열 그래프와 분석을 진행했다. 인천광역시 전체 아파트 제곱미터당 평균 매매가격 변화는 다음과 같다.



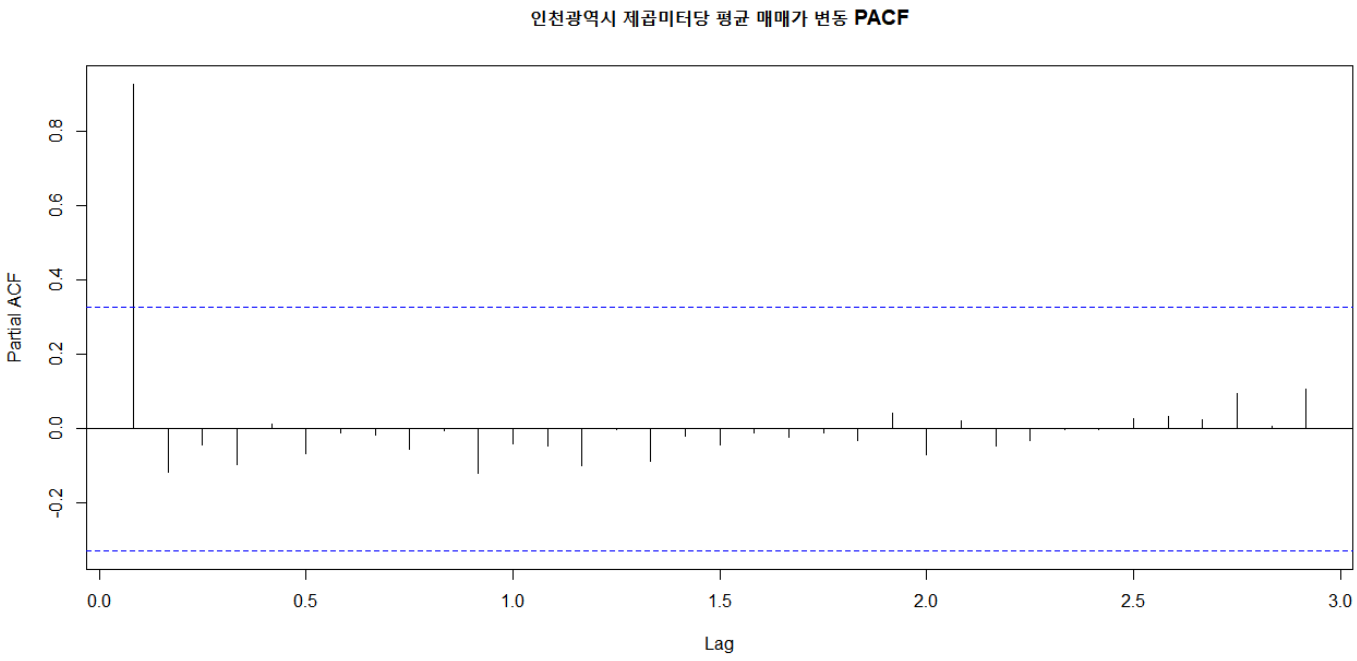
전반적으로 평균 매매 가격은 상승한다는 것을 확인할 수 있다. 2016년 1월부터 2018년 12월까지 제곱미터당 25만원 정도 상승한 것을 확인할 수 있다.

이 변동이 향후 미래에는 어떻게 변화할 지에 대해 시계열 분석을 이용하여 예측과 실제 2019년 1월부터 3월까지의 변동을 비교해보도록 한다.

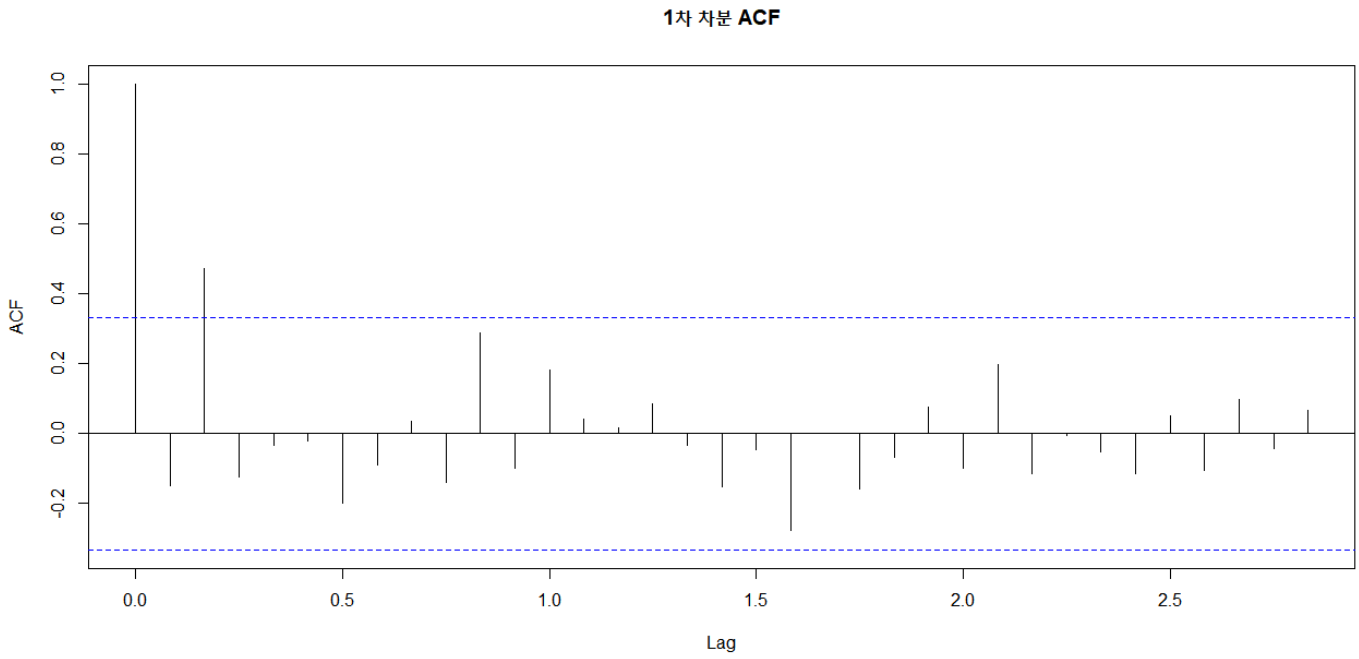
정상성의 만족을 하는지를 확인하기 위해 ACF와 PACF 그래프를 통해 확인했다.



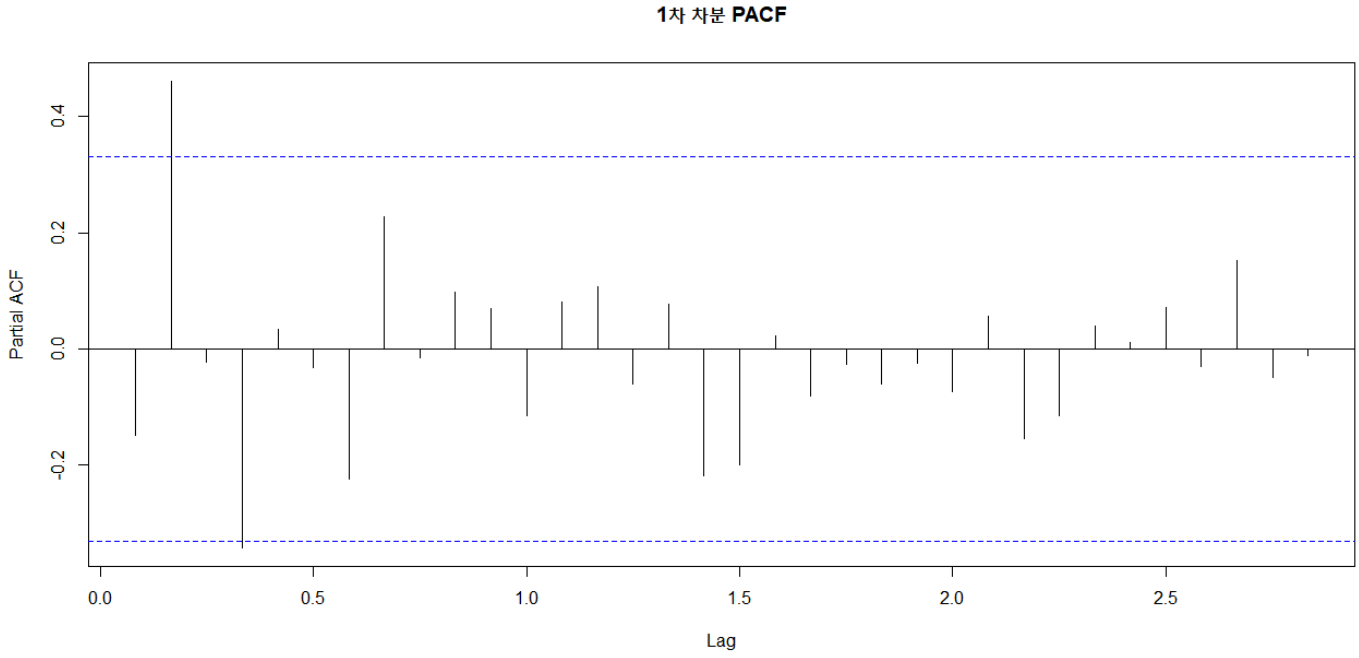
ACF의 경우, 지수적으로 급격하게 감소하지 않고, 일정한 수준으로 감소하는 모양세이다. 정상성을 만족한다고 볼 수 없다.



PACF를 보면, lag 1에서 절단값을 보이지만, ACF가 정상성을 만족하지 않으므로 1차 차분을 진행한다.



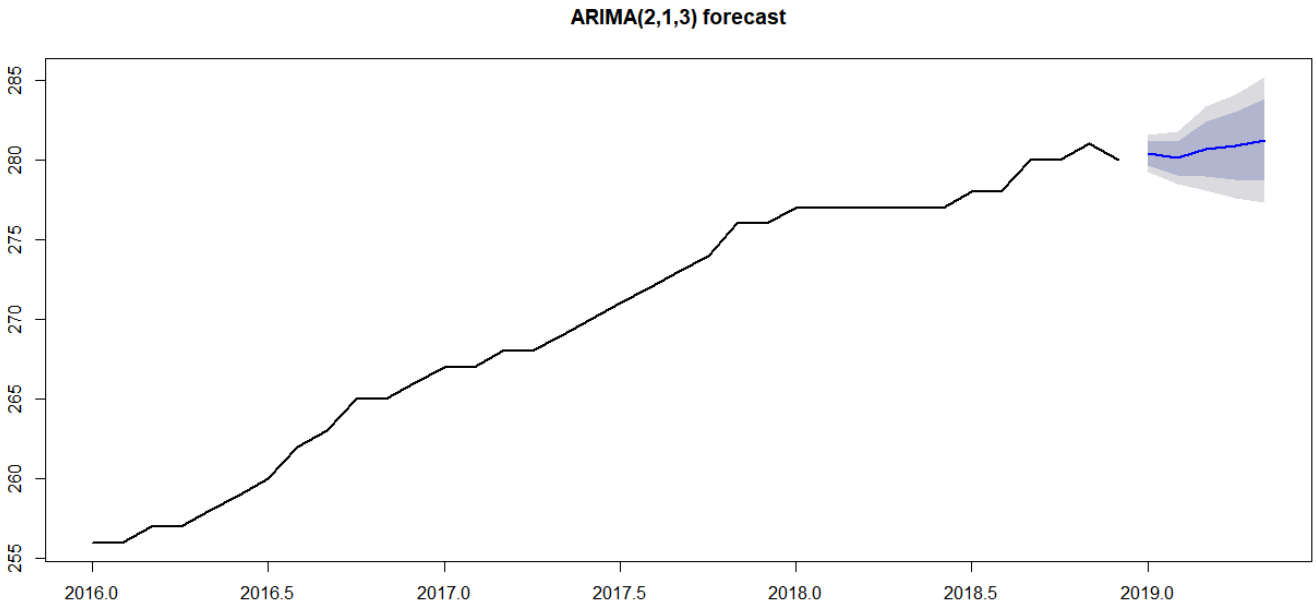
1차 차분 진행 후, ACF를 보면 lag 3이후로 0에 가깝다고 볼 수 있으므로 lag 3 이후로 절단값을 갖고, 또한 정상성을 만족하는 것으로 보인다. lag 3이후로 절단값을 가지므로 MA(3)이라 판단할 수 있다.



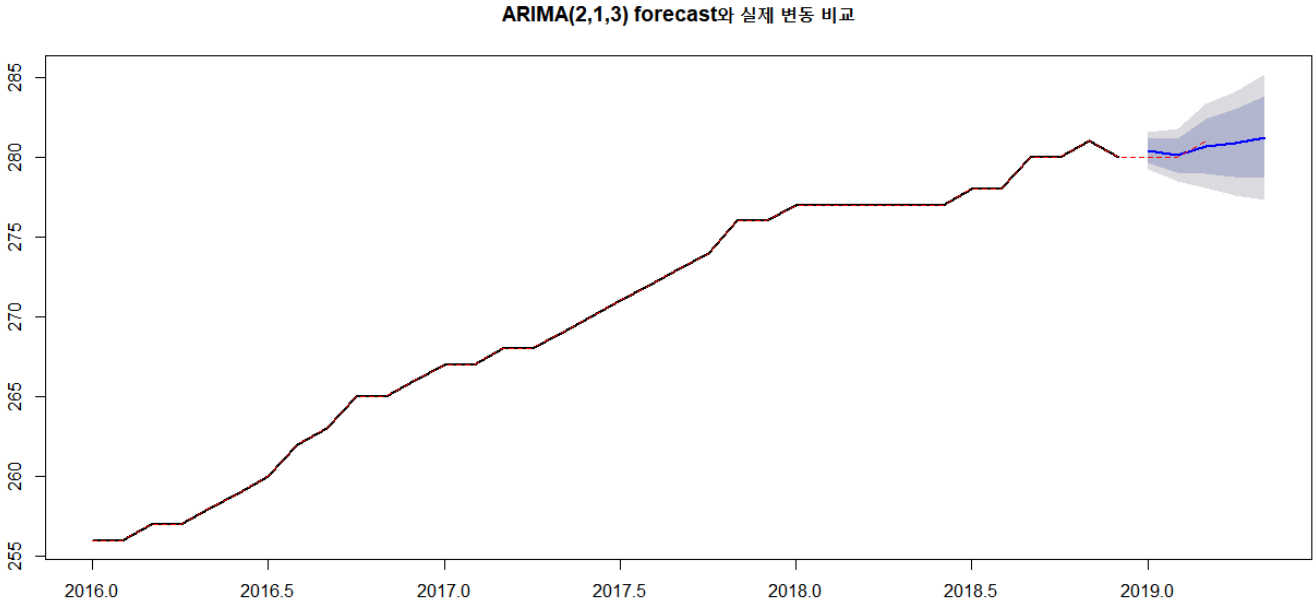
또한 1차 차분 PACF를 보면 lag 2에서 절단값을 가지므로 정상성도 만족하고 AR(2)로 판단할 수 있다.

이러한 결과를 토대로 ARIMA 모형을 생각하면, AR(2), 1차차분 진행, MA(3) 이므로 ARIMA(2,1,3)으로 생각해볼 수 있다.

ARIMA(2,1,3)으로 적합한 결과를 보면 다음과 같다.

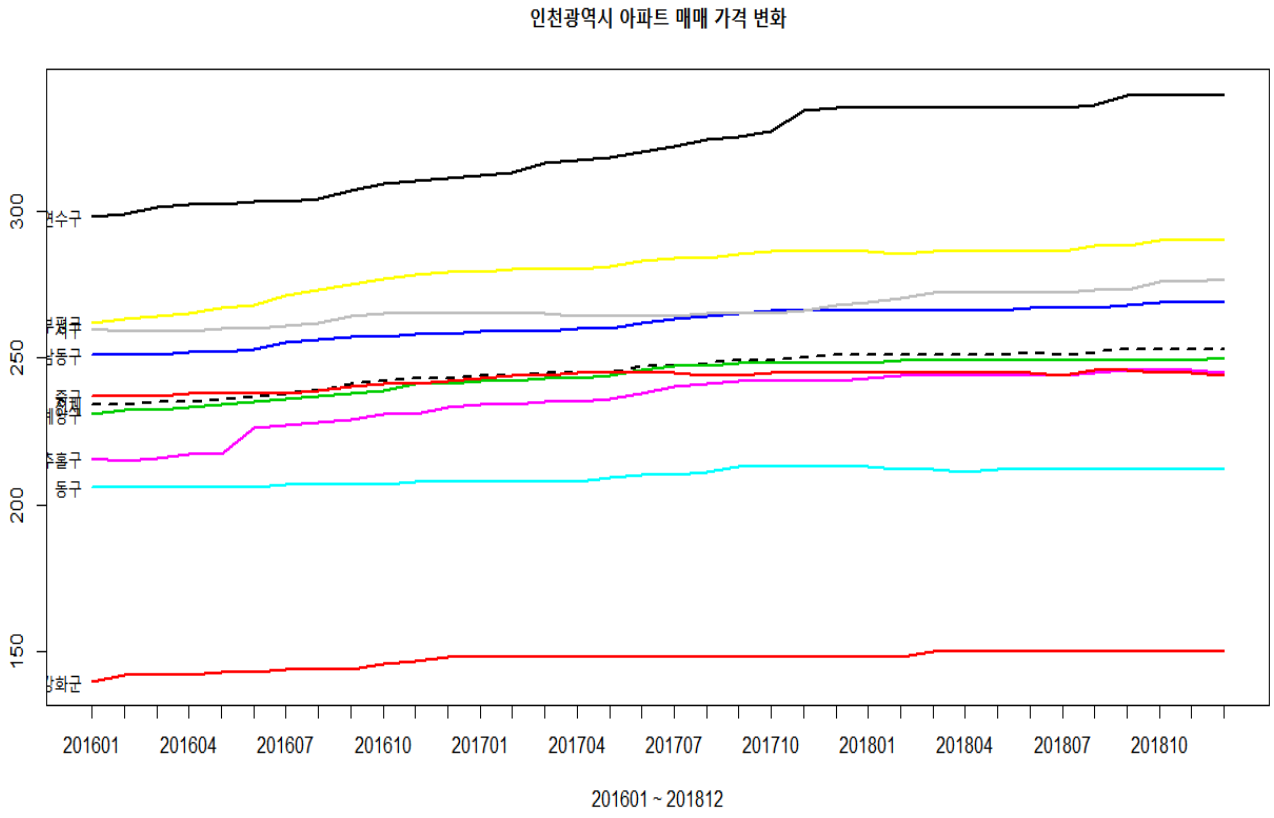


상승세를 보인다는 것을 예측할 수 있고, 실제 2019년 1월부터 2019년 3월의 변동 데이터와 비교를 해보면



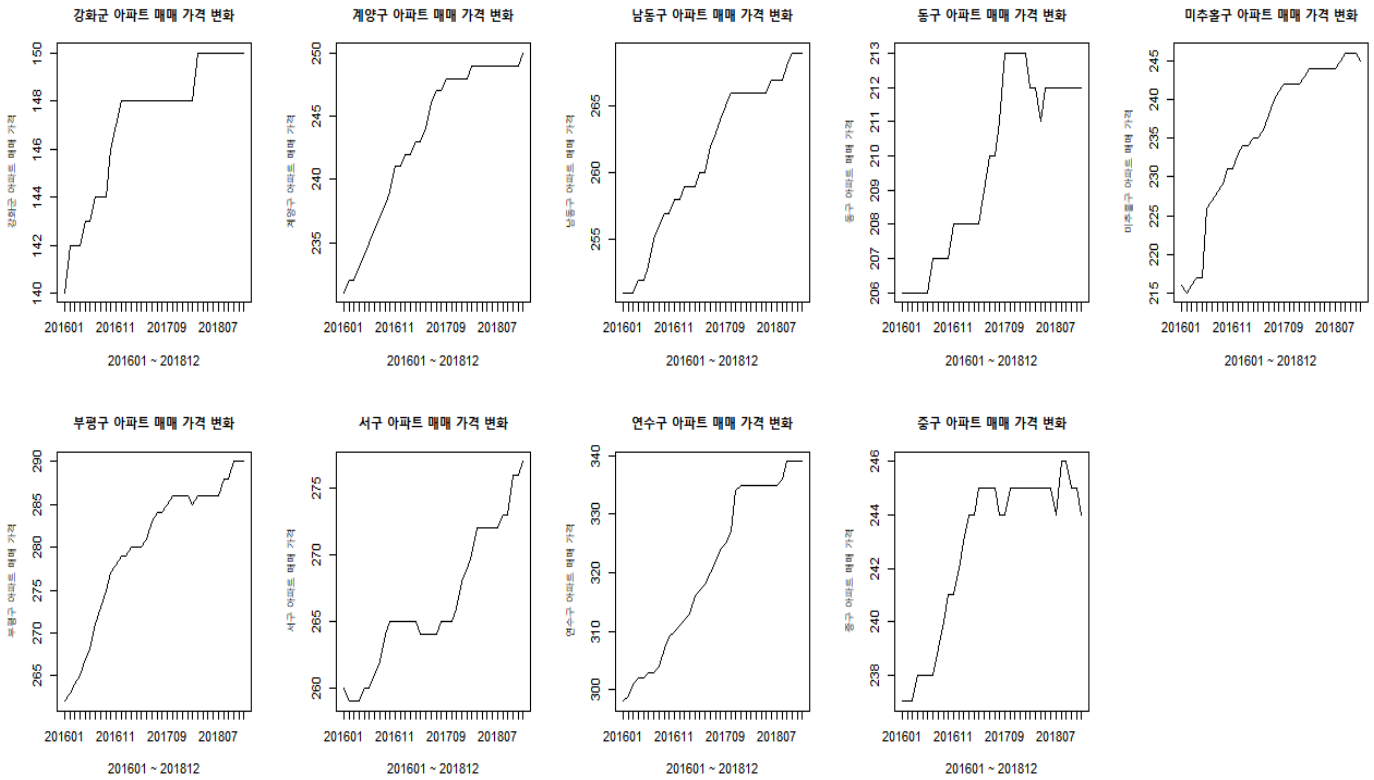
빨간색 점선이 실제 데이터 변동이고, 그림상 어느정도 비슷하다는 것을 확인할 수 있다. 앞으로 지속적으로 평균 매매가는 상승한다고 예측할 수 있다.

B. 인천광역시 구군별 제곱미터당 평균 매매 가격 변화



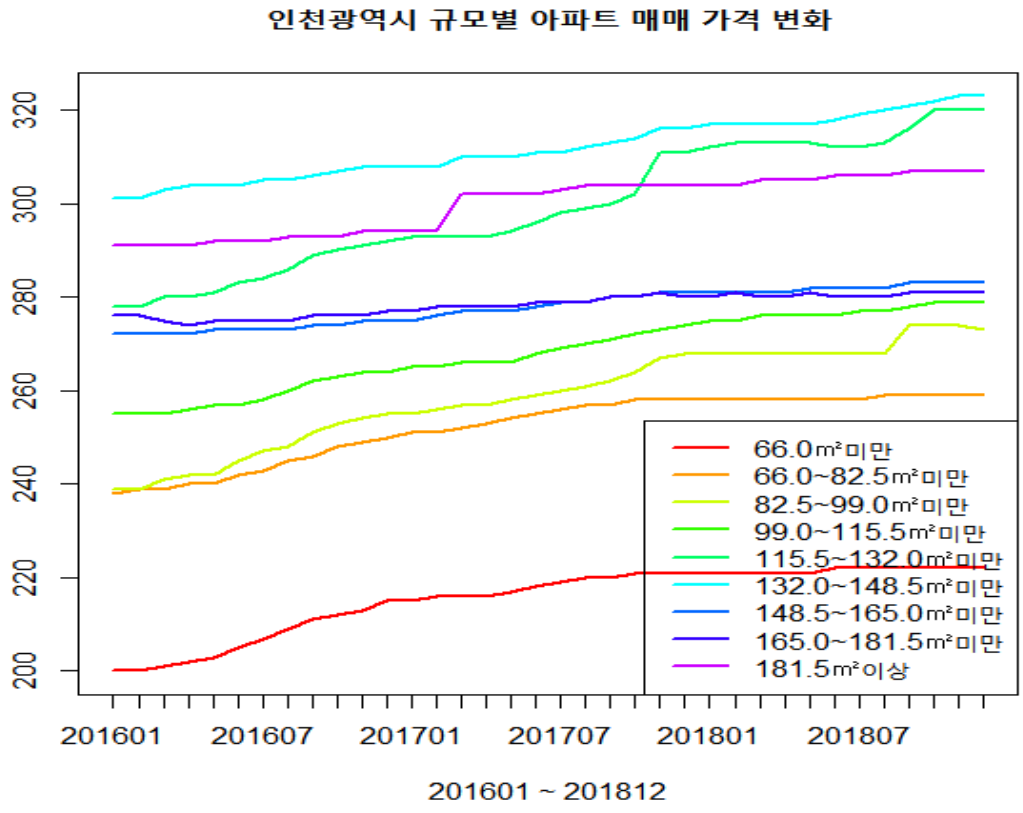
연수구가 독보적으로 인천광역시 중에서 제일 높고, 강화군이 제일 낮다는 것을 확인할 수 있다. 미추홀구의 경우, 2016년 1월에는 7위로 하위권이었지만, 꾸준히 증가하여 중구와 계양구와 비슷한 수준으로 올라간 것을 확인할 수 있다. 꾸준히 증가할 수 있었던 요인은 인하대학교 주변 아파트 단지가 집값 상승에 영향을 주었다고 짐작할 수 있다. 전체적으로는 모든 구군에서 조금씩 상승하는 것을 확인할 수 있고, 독보적으로 연수구는 크게 상승한다는 것을 확인할 수 있다. 인천국제도시라는 송도의 프리미엄 입지와 대규모 쇼핑시설, 아파트 단지 등으로 인해 짐작할 수 있다.

각 행정구역별로 세부적인 제곱미터당 평균 매매가 변동을 보면



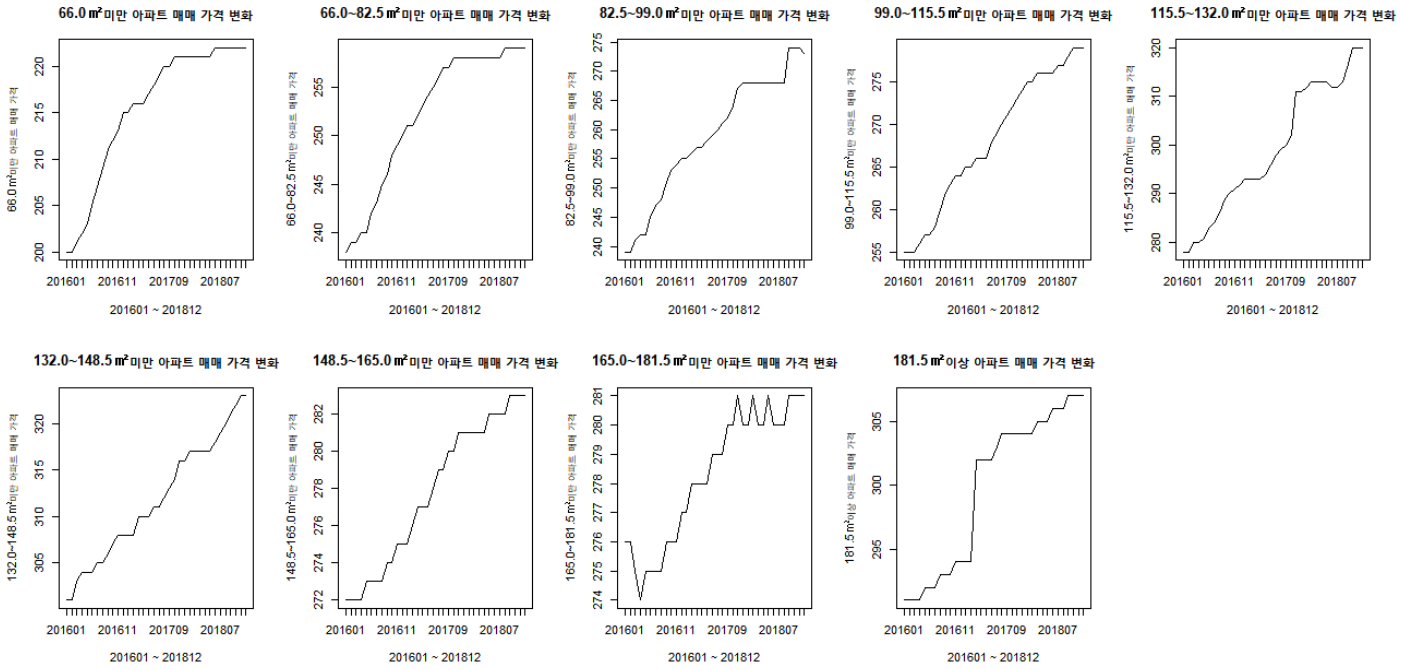
모든 행정구역이 전체적으로 증가한 것을 확인할 수 있다. 다만 동구의 경우, 최고점 이후 감소했다는 것을 확인할 수 있다.

C. 인천광역시 거래규모별 아파트 제곱미터당 평균 매매 가격 변동



전반적으로 증가 추세를 보이고 있고, 최상위권은 132.0~148.5m^2이고, 최하위권은 66.0m^2미만이다.

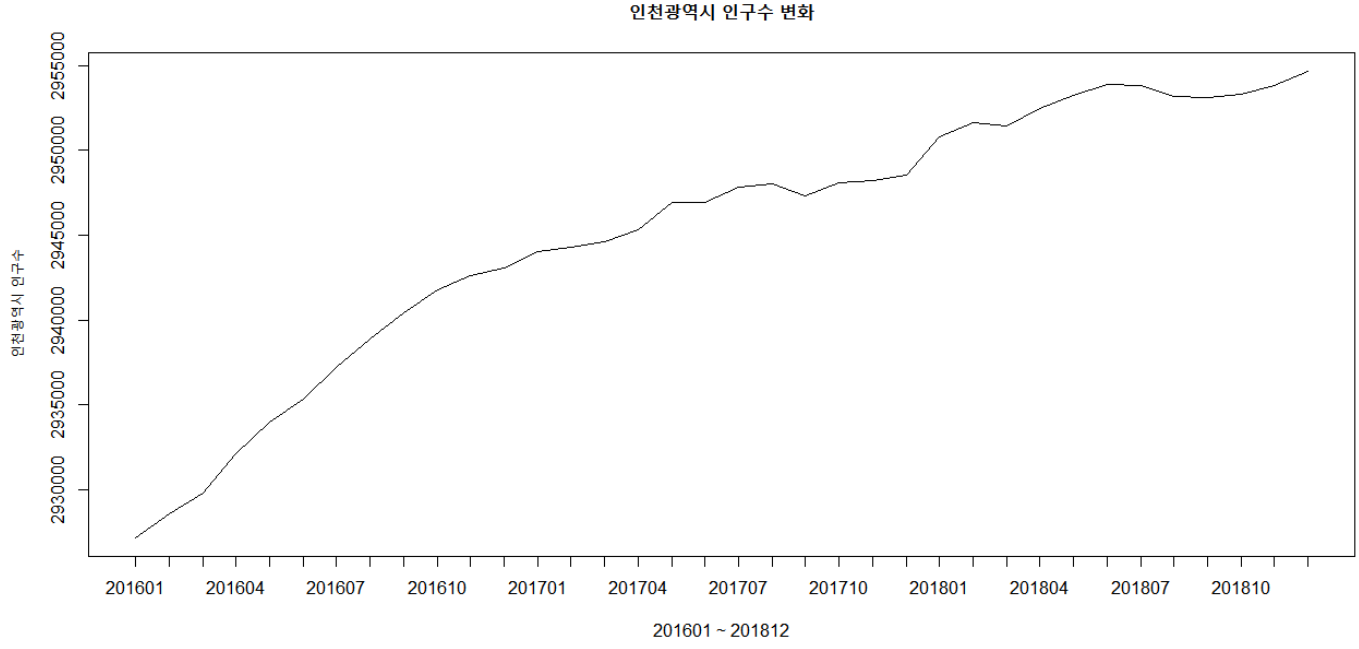
무조건 규모가 클수록 아파트의 가격이 높아지는 것이 아닌 것을 확인할 수 있으며, 중간 규모 수준의 아파트 가격대가 더 높은 경우도 있다는 것을 확인할 수 있다.



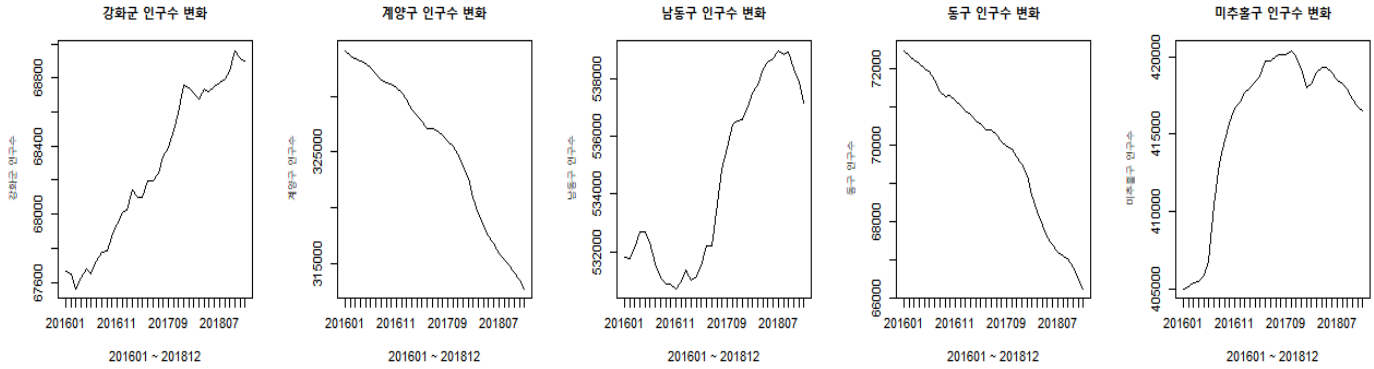
다음의 개별의 그래프를 보면, 모든 그래프가 상승세를 보이는 것을 확인할 수 있다.

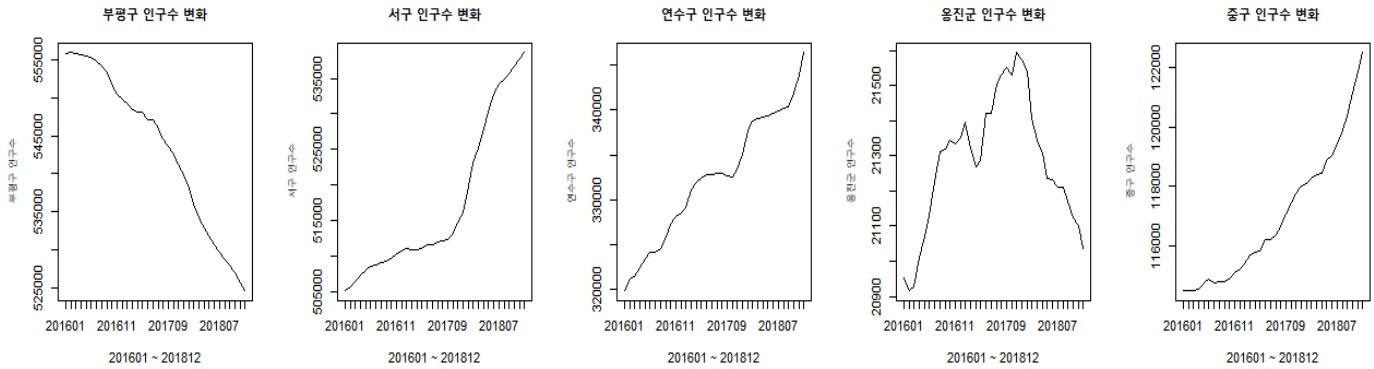
D. 인천광역시 인구수 변화

2016년 1월부터 2018년 12월까지의 인천광역시 인구수 변화는 다음과 같다.



우선 인천광역시 인구수는 전체적으로 증가한 것을 알 수 있다. 또한 이러한 상승세를 유지한다면 3백만 명 돌파는 시간 문제이다. 또한 A의 인천광역시 아파트 제곱미터당 평균 매매가의 시계열 그래프와 같은 추세를 보이므로, 두 변수간 양의 상관관계를 생각할 수 있다..

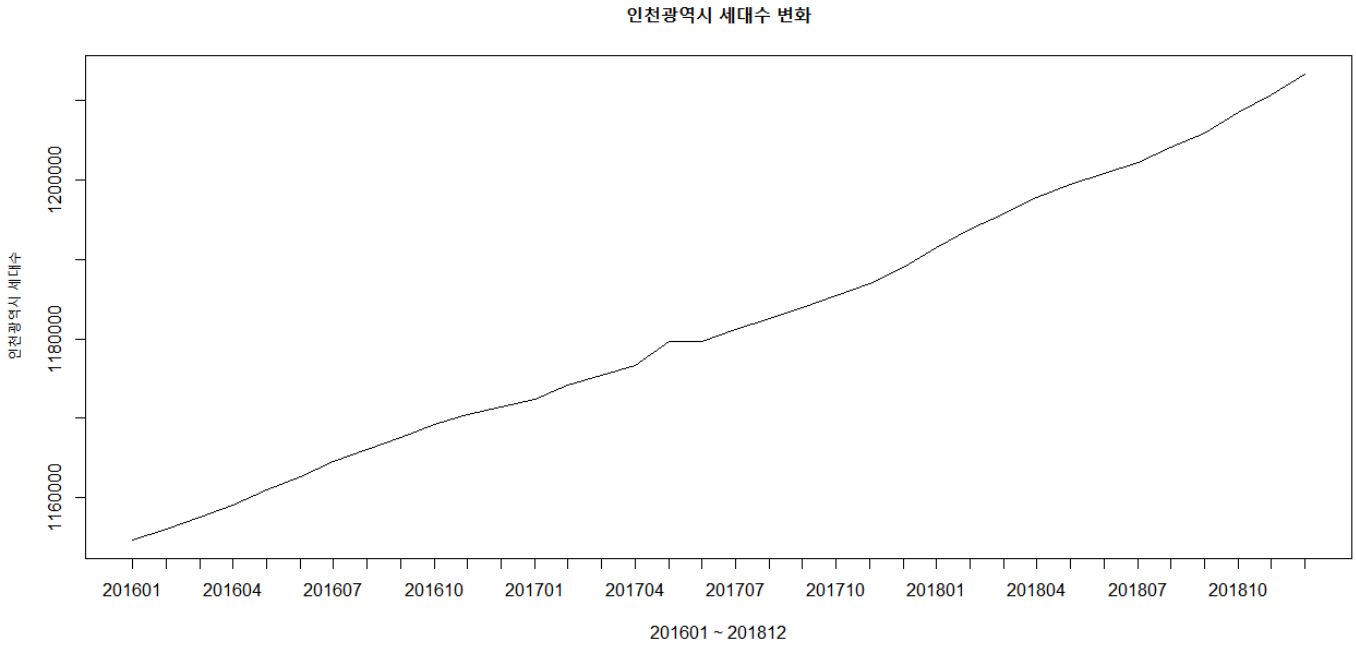


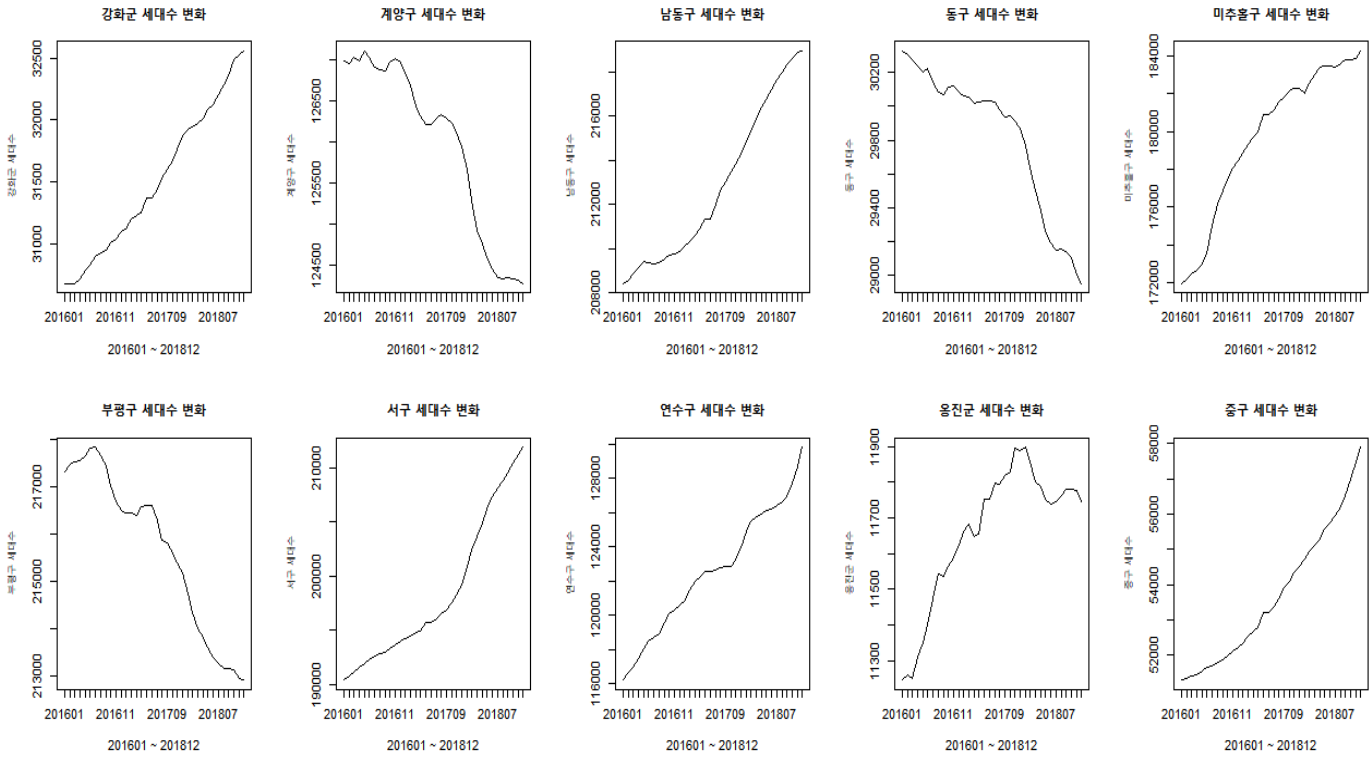


세부적으로 본다면, 강화군, 서구, 연수구, 중구는 전체적으로 증가하는 것을 알 수 있다. 계양구, 동구, 부평구의 경우는 지속적으로 감소한 것을 확인할 수 있다. 남동구의 경우, 16년에는 증가하다 감소하고 이후 다시 증가를 하다가 다시 감소하는 형세를 보인다. 미추홀구의 경우, 급격하게 증가하다가 최고점 이후 감소세를 보이고 있다. 옹진군은 전체적으로 증가했다가 감소세를 보이고 있다.

E. 인천광역시 세대수 변화

2016년 1월부터 2018년 12월 세대수는 인구수 변화와 유사하다

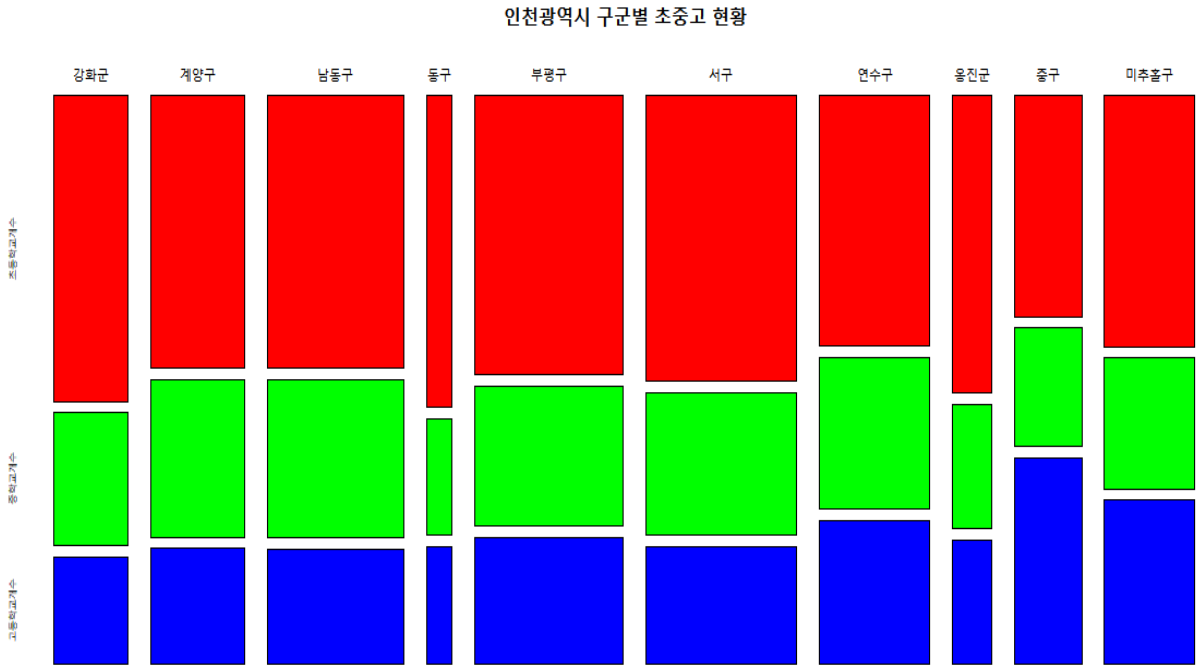




개별로 보아도 인구수와 유사한 추세를 보인다는 것을 확인할 수 있다.

F. 인천광역시 교육시설 수 현황

인천광역시에 있는 초등학교, 중학교, 고등학교 현황을 각 행정구역별로 구분한 그래프이다.



빨간색(초등학교 수), 녹색(중학교 수), 파란색(고등학교 수)이고 모든 행정구역에서 초등학교 수의 비중이 크다는 것을 확인할 수 있다. 중구의 경우, 초등학교와 고등학교 비중이 비슷하다. 중학교와 고등학교 비중은 거의 비슷하거나 중학교 비중이 약간 큰 것을 확인할 수 있다.

분할표를 보면

	초등학교개수	중학교개수	고등학교개수	Row Total
강화군	23	10	8	41
	0.315	0.041	0.356	
계양구	26	15	11	52
	0.000	0.161	0.185	
남동구	38	22	16	76
	0.000	0.252	0.290	
동구	8	3	3	14
	0.147	0.113	0.041	
부평구	42	21	19	82
	0.028	0.005	0.028	
서구	44	22	18	84
	0.103	0.001	0.246	
연수구	28	17	16	61
	0.196	0.081	0.117	
옹진군	12	5	5	22
	0.095	0.091	0.017	
중구	15	8	14	37
	0.650	0.274	2.906	
미추홀구	23	12	15	50
	0.153	0.078	0.726	
Column Total	259	135	125	519

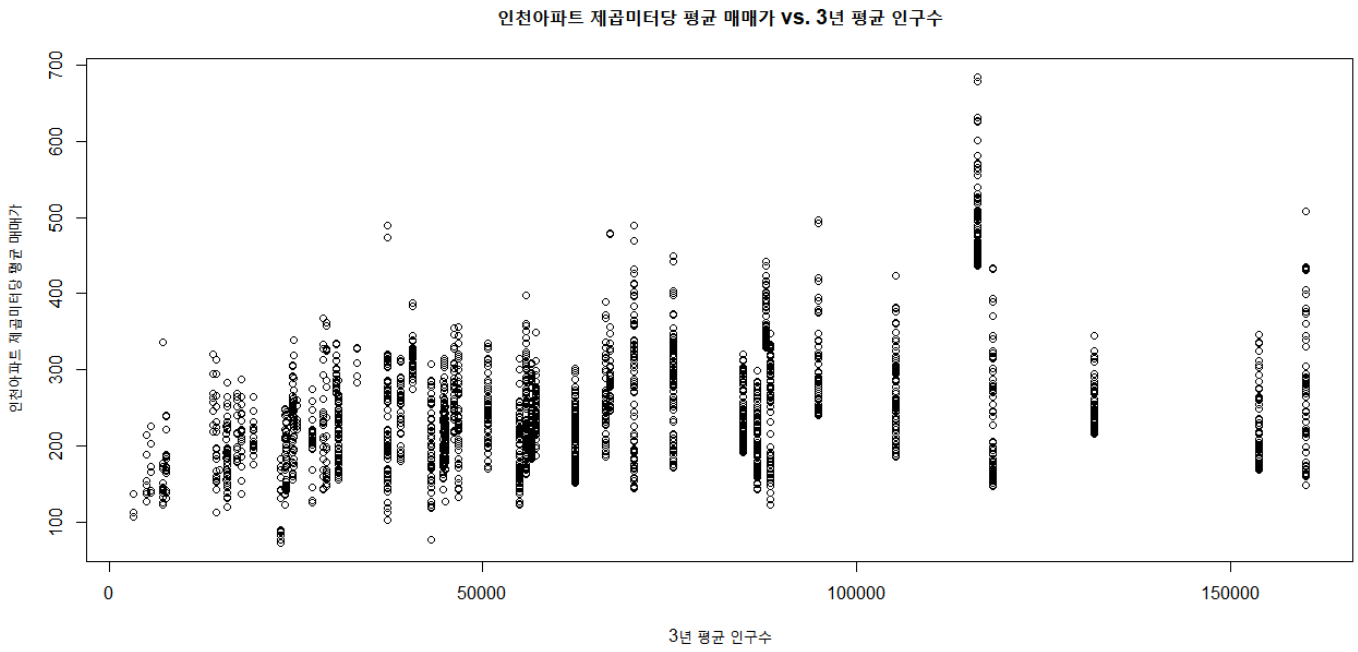
Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 7.697246 d.f. = 18 p = 0.9828108

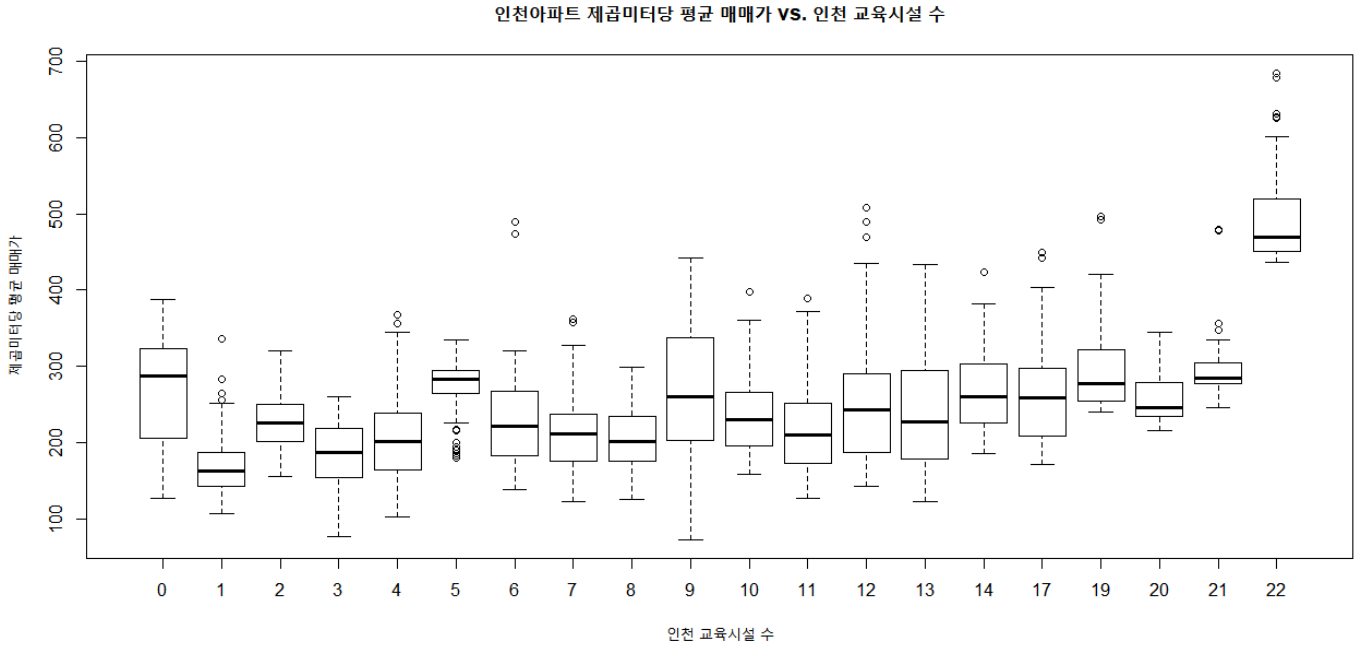
P-value가 0.05보다 크므로 교육시설 수와 인천광역시 구군 변수 간의 연관성이 없다고 볼 수 있다.

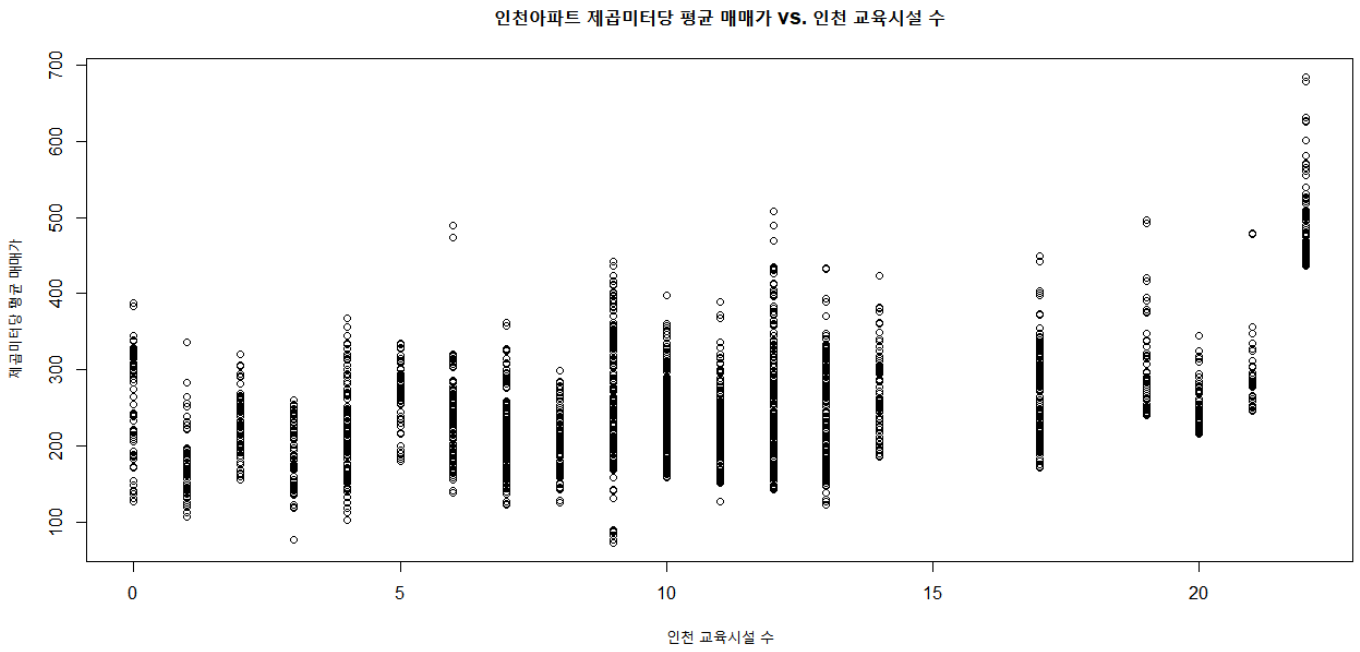
G. 인천아파트 제곱미터당 평균 매매가 vs. 인천 3년 평균인구수



전체적으로 3년 평균 인구수가 클수록 인천아파트 제곱미터당 평균 매매가도 높아지는 추세를 보인다고 할 수 있다. Spearman 상관관계를 통해 분석하면, $\rho = 0.3203403$ 정도로 어느정도 양의 상관관계를 보인다고 할 수 있다. 특정 3년 평균 인구수에서 집값의 가격대가 높은 것을 확인할 수 있는데 앞서 보인 송도동의 특징을 고려한다면 송도동의 가격대일 수 있다.

H. 인천아파트 제곱미터당 평균 매매가 vs. 인천 교육시설 수

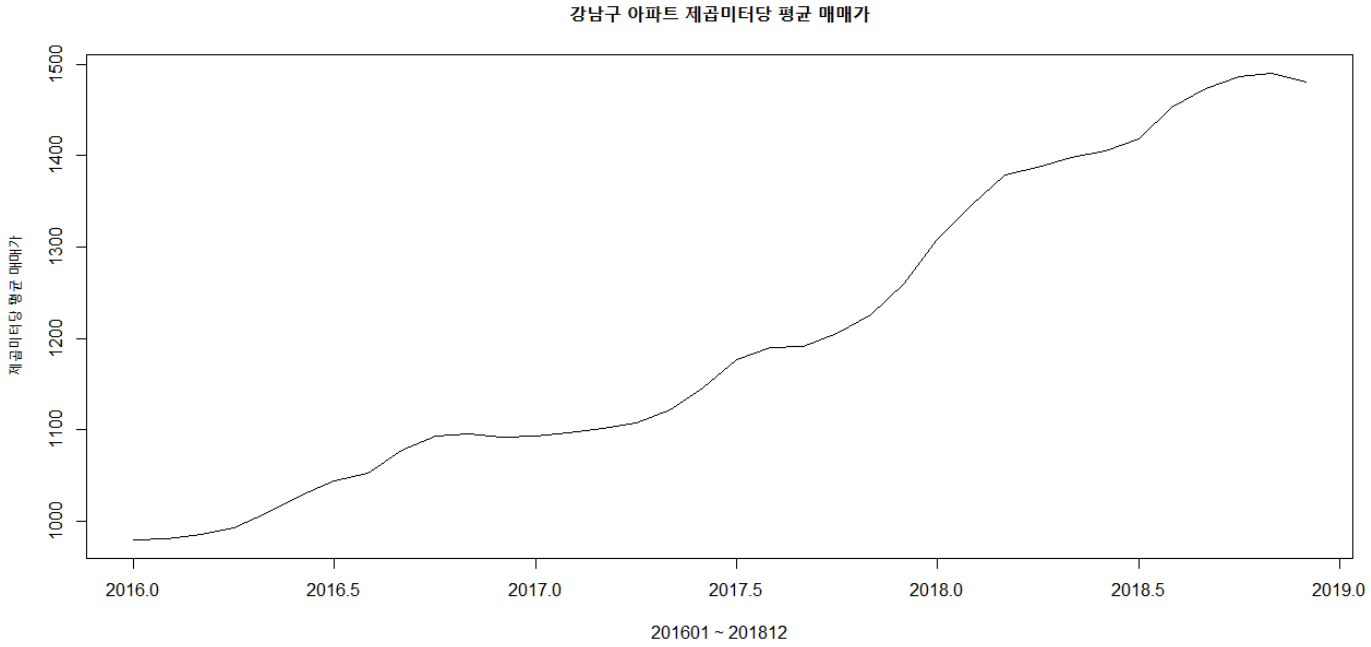




Spearman 상관관계 분석 결과, $\rho = 0.3233$로 어느정도 양의 상관관계를 보인다고 볼 수 있다. 인천아파트 제곱미터당 평균 매매가와 교육시설 수 관계로 알 수 있는 점은 교육시설 수가 무작정 많다고 해서 집값이 크게 상승하지 않는다는 점이다. 다만 교육시설 수가 9에서 13 사이의 경우, 같은 교육시설 수내에서 가격대가 넓게 분포되어 있는 것을 확인할 수 있다. 또한 교육시설 수가 22인 경우, 다른 경우에 비해 훨씬 높은 가격대가 형성되어 있다는 것을 확인할 수 있다. 22개의 교육시설 수를 가지고 있는 동은 송도동이다

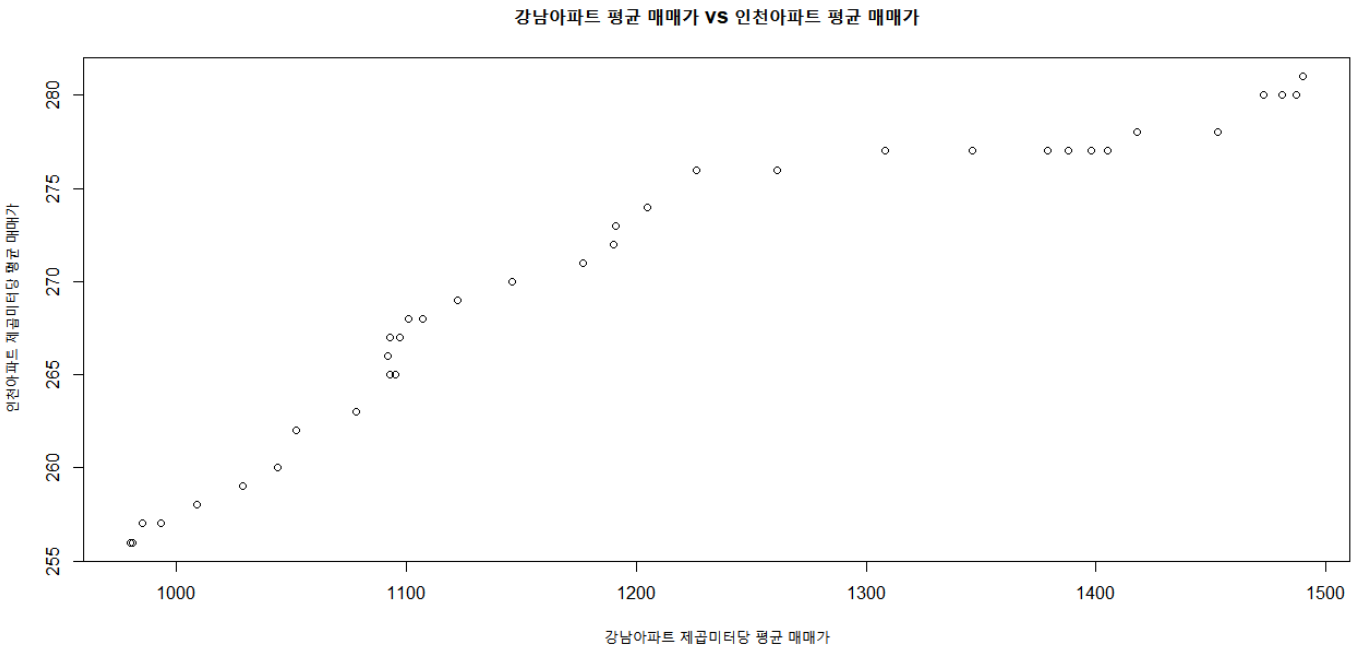
I. 강남구아파트 제곱미터당 평균 매매가격과 인천아파트 전체 제곱미터당 평균 매매가격 관계

인천광역시 전체 아파트 제곱미터당 평균 매매가와 서울에 대표적인 집값을 상징하는 강남구 아파트 제곱미터당 평균 매매가격과의 관계를 파악한다.



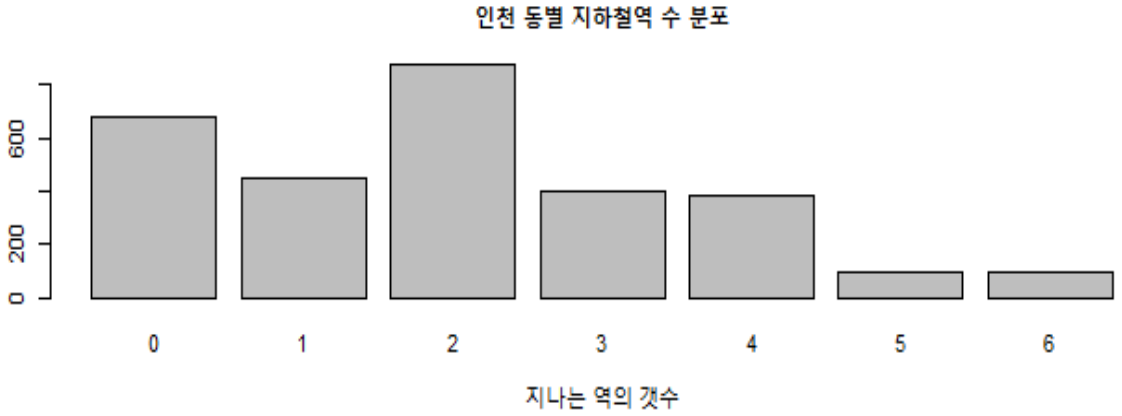
2016년 1월부터 2018년 12월까지의 강남구 전체 아파트 제곱미터당 평균 매매 가격의 시계열 그래프이

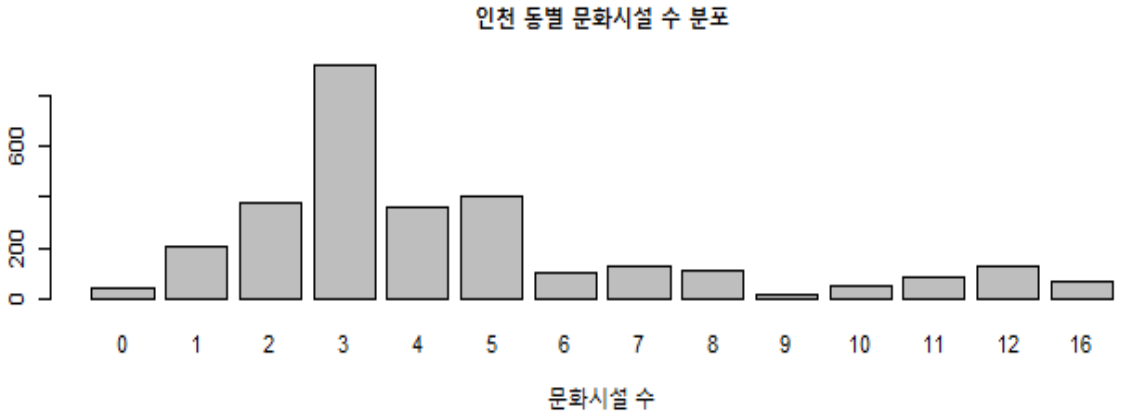
다. 전체적으로 가파른 성장세를 보이고 있다. 3년 동안 제곱미터당 약 500만원정도가 상승했다. 2018년 말에 들어서 약간 상승세가 감소세를 보이는 데 그것은 정부의 부동산 관련 규제로 인한 효과라는 것을 짐작할 수 있다.

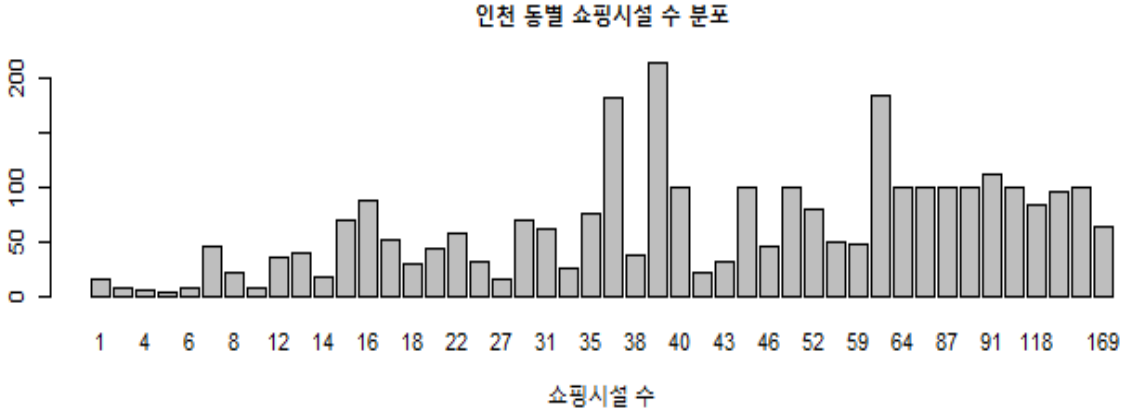


강남아파트 집값의 변동이 얼마나 인천 전체 아파트 집값과 상관이 있는지 파악하기 위해 산점도를 그렸다. Correlation은 0.9439로 매우 강하다. 또한 spearman correlation에서 $\rho = 0.9951633$이므로 매우 강한 양의 상관관계를 보인다고 볼 수 있다.

J. 인천 동별 지하철역, 문화시설 수, 쇼핑 시설 수

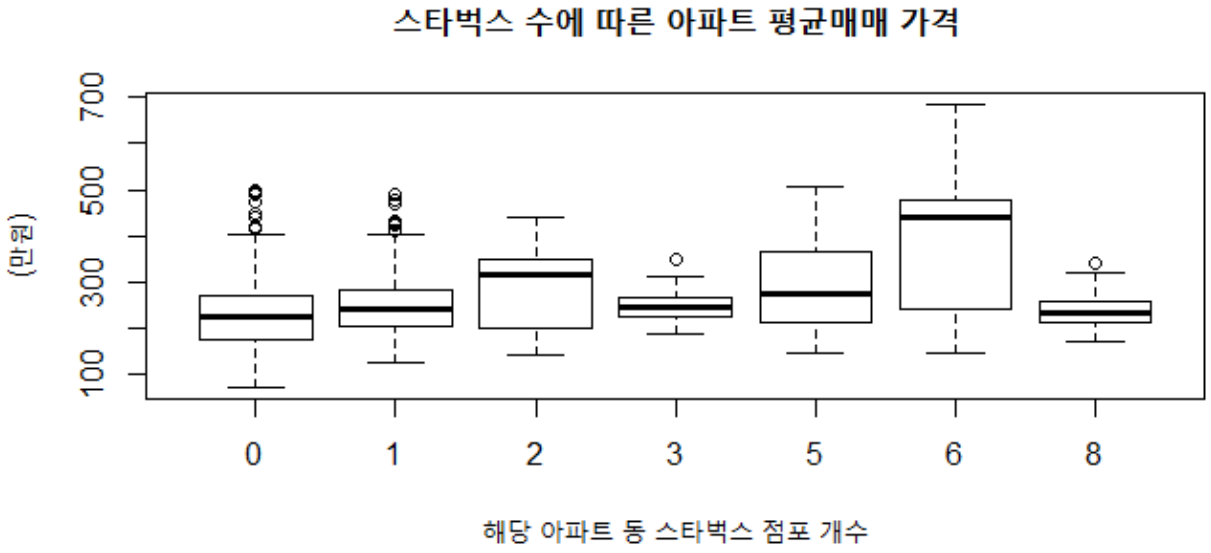


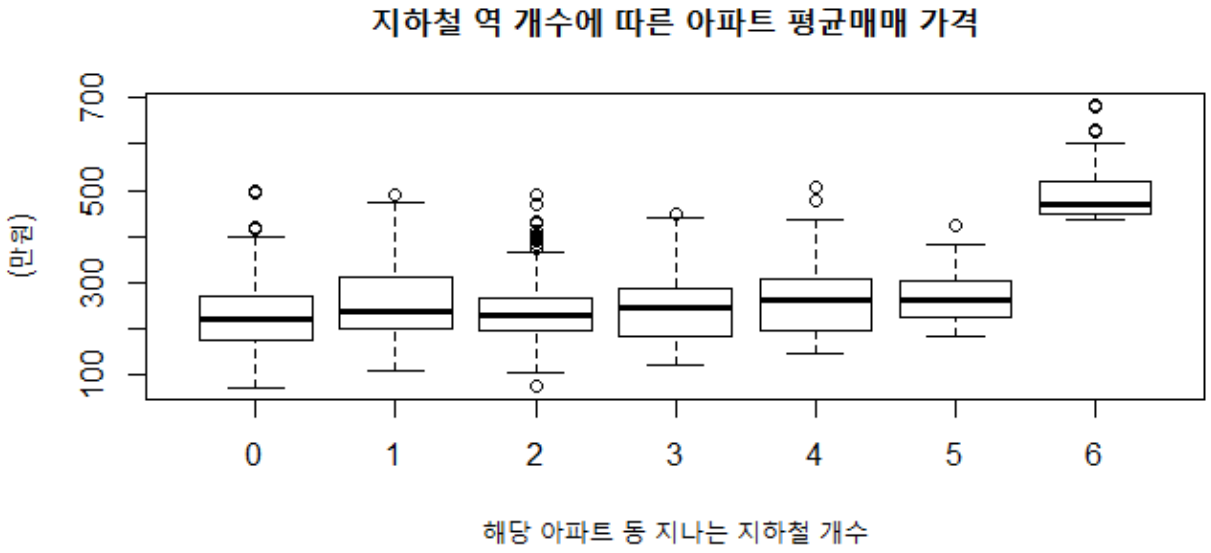




인천광역시 동별로 지하철 역, 문화시설 수, 쇼핑시설 수의 경우, 산정방법은 해당 동 범위에 지나가는 지하철 역의 수, 포함된 문화시설 수, 쇼핑시설 수를 계산한 것이다. 대개 지하철 역은 2개가 가장 많은 분포를 보인다. 문화시설 수는 3개로 가장 많은 분포를 보인다. 쇼핑 시설 수는 39개에서 가장 많은 것을 확인할 수 있다.

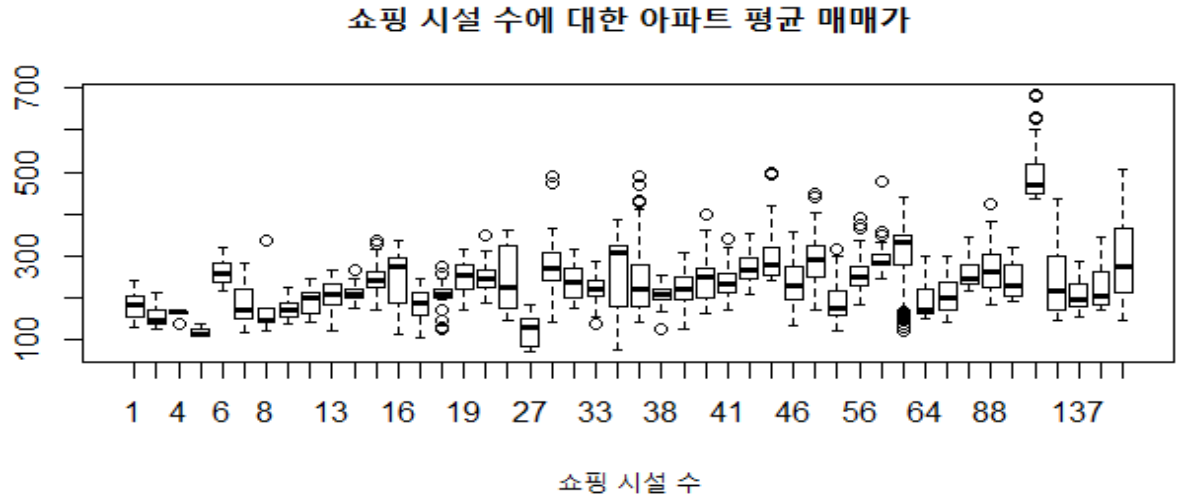
K. 인천아파트 제곱미터당 평균매매가 vs 스타벅스 수, 지하철 역 개수

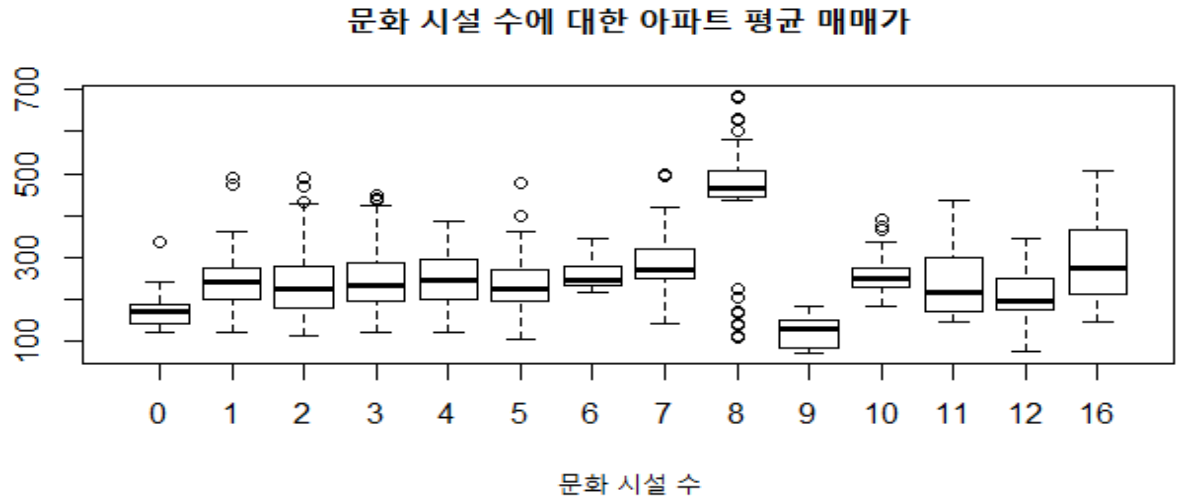




첫번째 그림에서 스타벅스 8개점포의 동은 인천공항이 있는 중구 운서동이다. 대체적으로 스타벅스 수가 많다고 해서 인천아파트 제곱미터당 평균매매가가 급격하게 증가하는 것으로 보이지는 않는다. 지하철 역 개수에 따라 인천아파트 제곱미터당 평균매매가는 0개에서 5개까지는 수준이 비슷하지만, 6개부터는 많이 증가한 것으로 보인다. 6개가 지나는 동은 모두 송도동이므로 단순히 지하철 역의 개수가 많다고 높은 것이 아니라 다른 요인이 있다는 것을 생각할 수 있다.

L. 인천아파트 제곱미터당 평균매매가격 vs 쇼핑시설 수, 문화시설 수

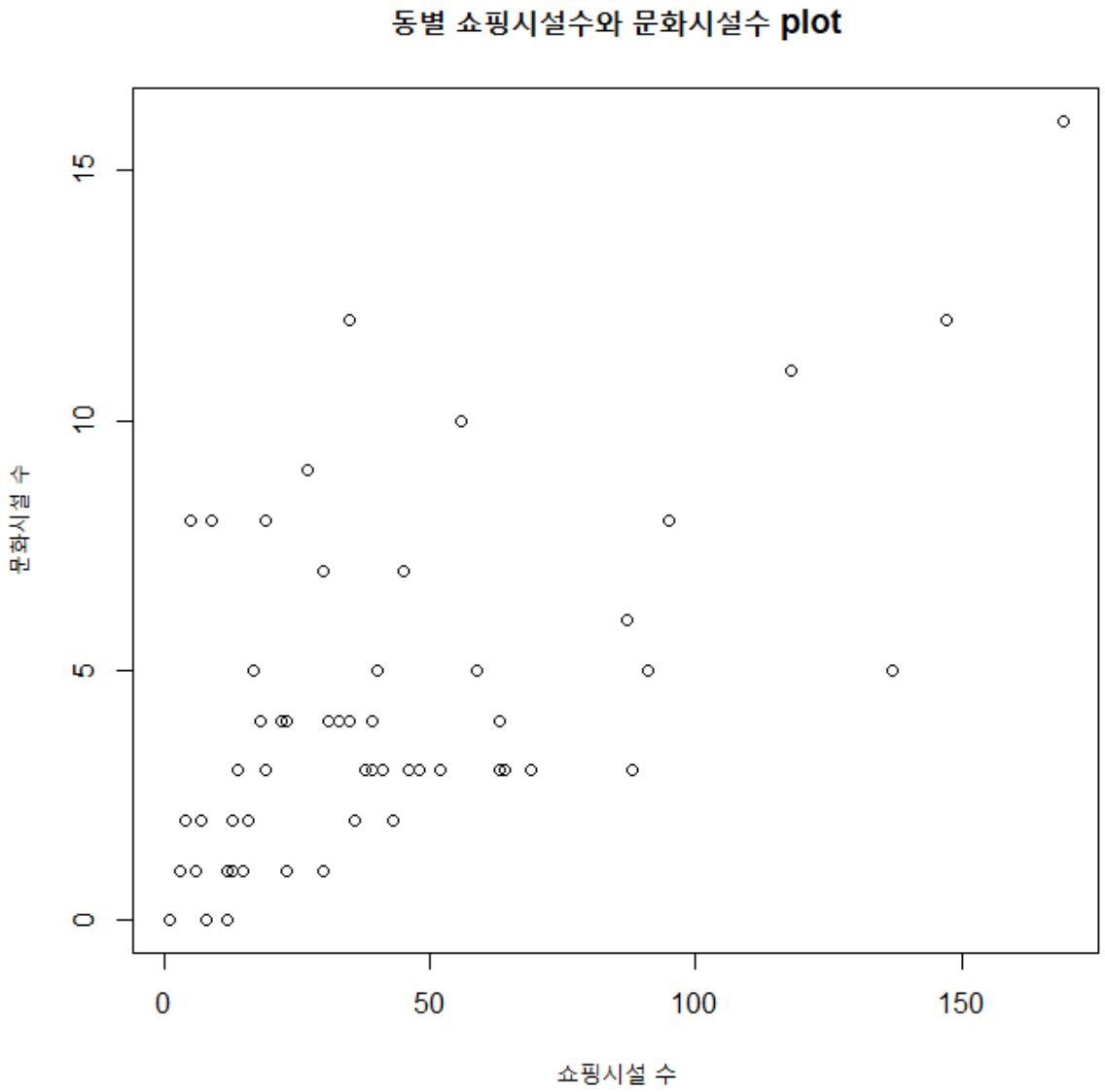




쇼핑 시설 수는 해당 동에 분포하고 있는 대형마트, 백화점, 편의점, 마켓 등을 합산한 값이다. 쇼핑 시설 수가 많을수록 인천아파트 제곱미터당 평균 매매가격도 상승하는 추세를 보인다. 하지만 급격한 변화는 보인다고 볼 수 없으며, 쇼핑 시설이 95개인 경우, 다른 곳에 비해 가격대가 높게 형성되어 있는 것을 확인할 수 있다. 그곳은 바로 송도동임을 짐작할 수 있다.

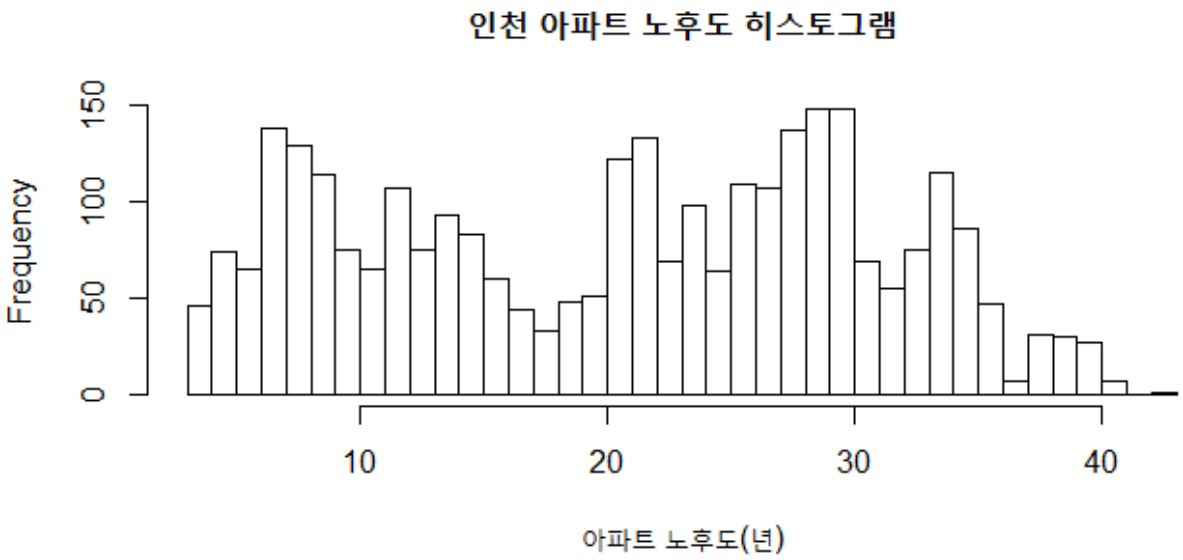
문화 시설 수는 해당 동에 분포하고 있는 영화관, 박물관, 공원 등을 합산한 값이다. 문화 시설 수가 많을수록 인천아파트 제곱미터당 평균 매매 가격이 증가한다고 볼 수 없다. 다만, 문화 시설 수가 8인 경우 가장 높은 가격대를 형성했다. 문학동, 송도동 등이 있는데, 주로 송도동이 많이 포함되어 있다.

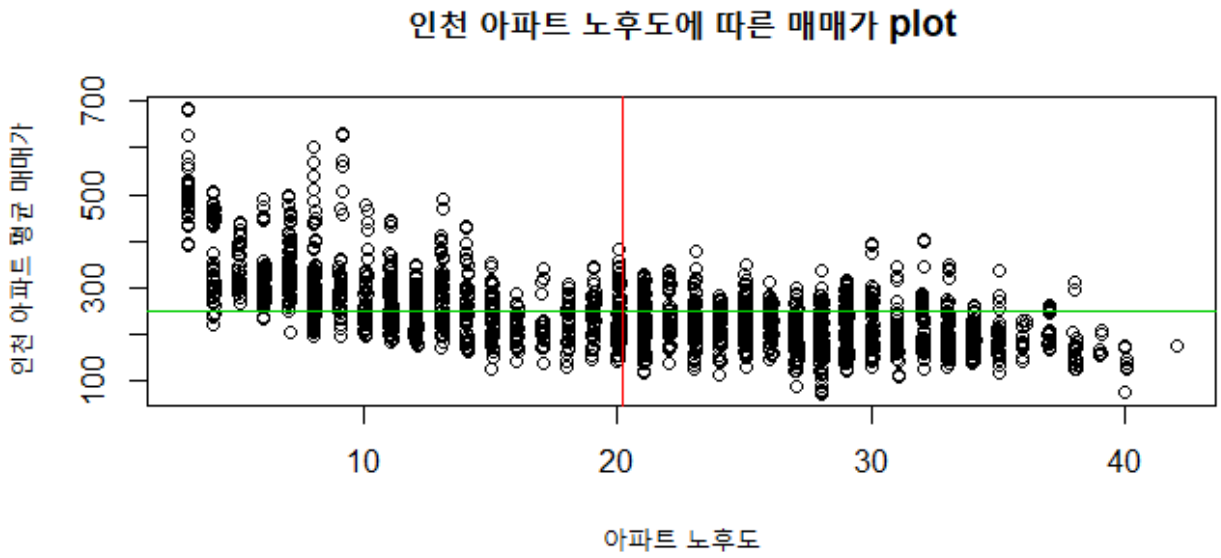
M. 쇼핑시설 수 vs 문화시설 수



인천 동별 쇼핑 시설 수에 대한 문화 시설 수 산점도이다. 대체적으로 우상향하는 분포를 보이고 있으며, 쇼핑 시설 수가 0에서 50사이에서 문화 시설 수가 많이 분포되어 있다는 것을 확인할 수 있다.

N. 인천 아파트 노후도 & 인천아파트 제곱미터당 평균 매매가 vs 인천아파트 노후도

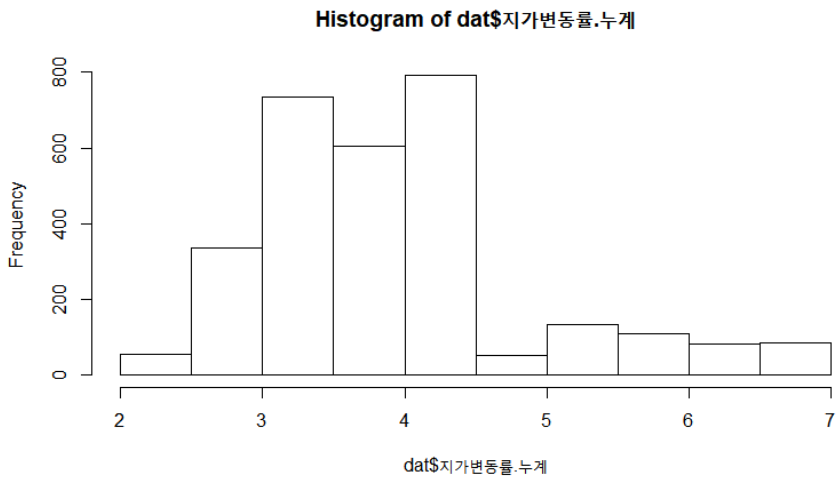




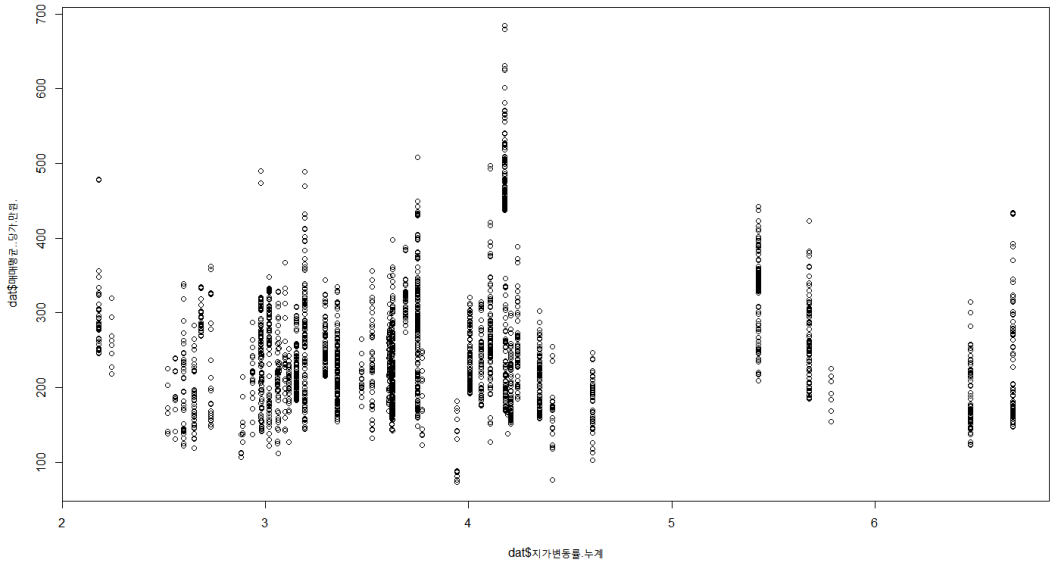
두번째 산점도에서 빨간색 선은 노후도 평균(약 20년)을 의미하고 녹색 선은 인천아파트 최근 3년 제곱미터당 평균 매매가의 평균값(약 249만원)이다.

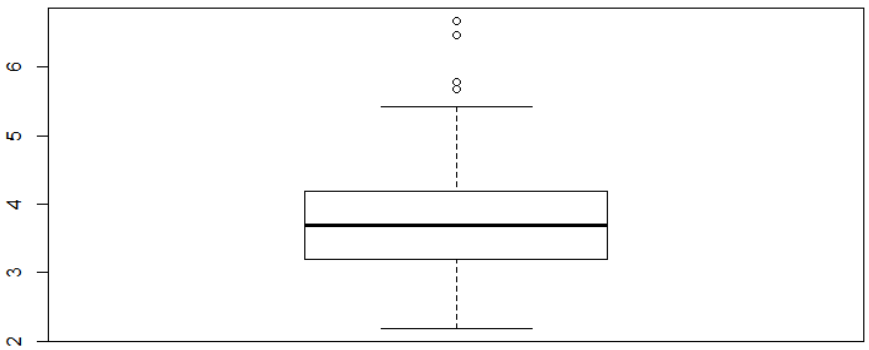
아파트 노후도에 따라 노후도가 높을수록 인천아파트 제곱미터당 평균 매매가는 낮아지는 추세를 확인할 수 있다.

O. 지가변동률과 인천아파트 제곱미터당 평균매매가 분석



지가변동률~매매평균가 (산점도, boxplot)

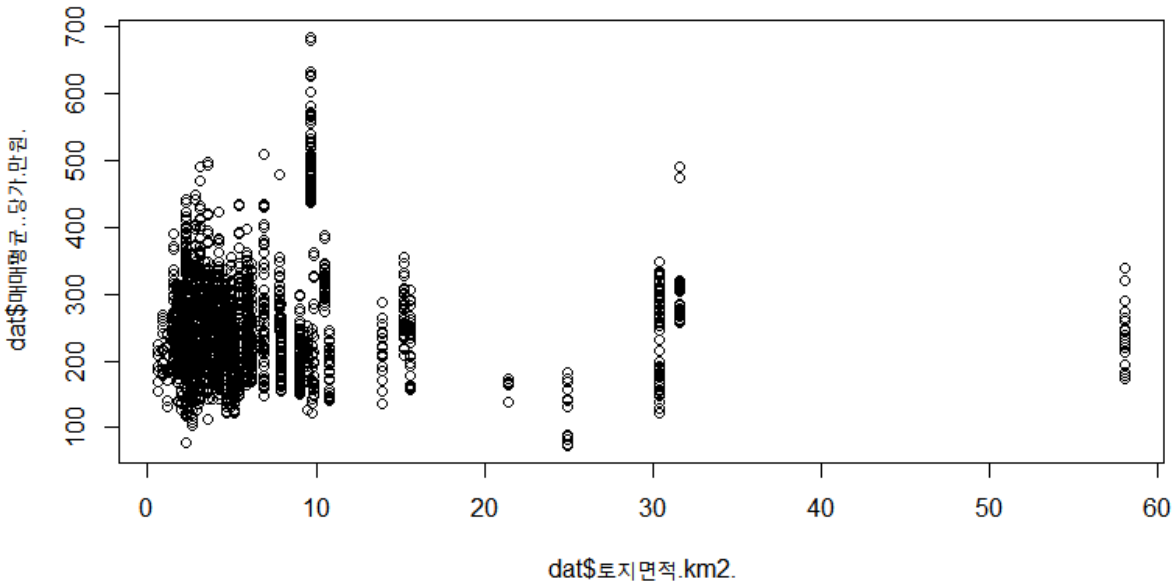


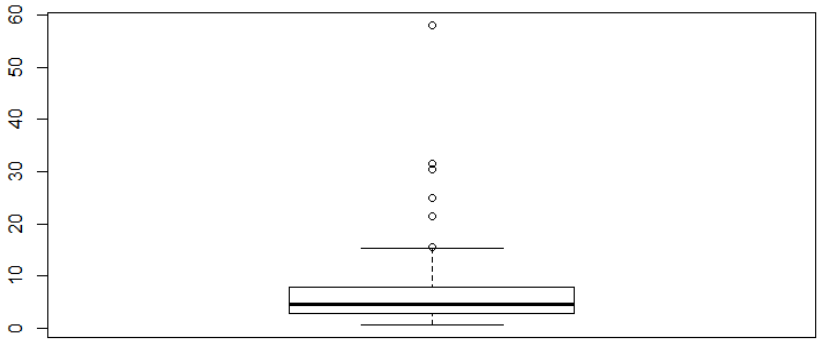


제1사분위수가 3.194, 중앙값이 3.689, 제3사분위수가 4.182이다. 사분위수에서 알 수 있듯이 대체로 지가변동률 누계값 3~4사이에 많이 분포하는 것으로 보인다. 5.6이상 되는 outlier가 4개 정도의 동에서 나

타나는 것으로 보인다. 의외로 지가변동률이 가장 높은 동에서의 제곱미터당 평균 매매가는 그리 높지 않은 것을 알 수 있다. 또한 지가변동률이 가장 낮은 동에서 제곱미터당 평균 매매가격의 최저가 200후반으로 꽤 높은 것으로 보아 안정된 입지를 가지고 있는 것으로 예측할 수 있다. 산점도의 지가변동률 4.3에서 유독 최고가의 제곱미터당 평균매매가를 가진 동이 나타난다.

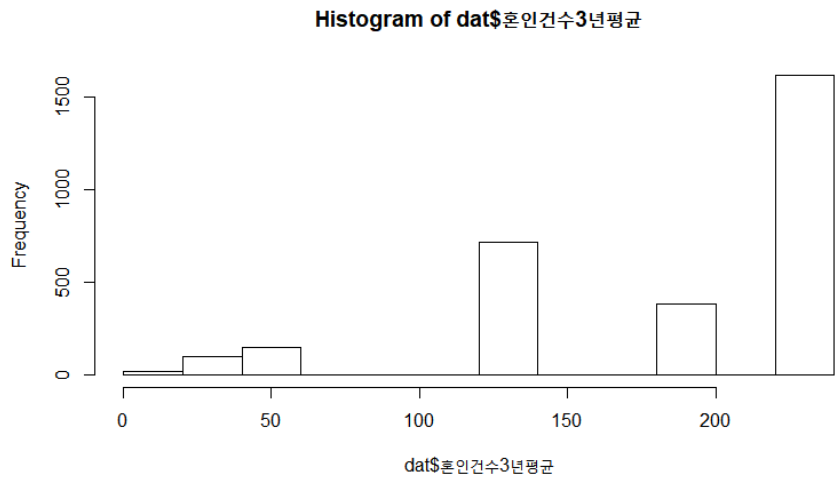
P. 인천아파트 제곱미터당 평균 매매가와 토지면적 분석



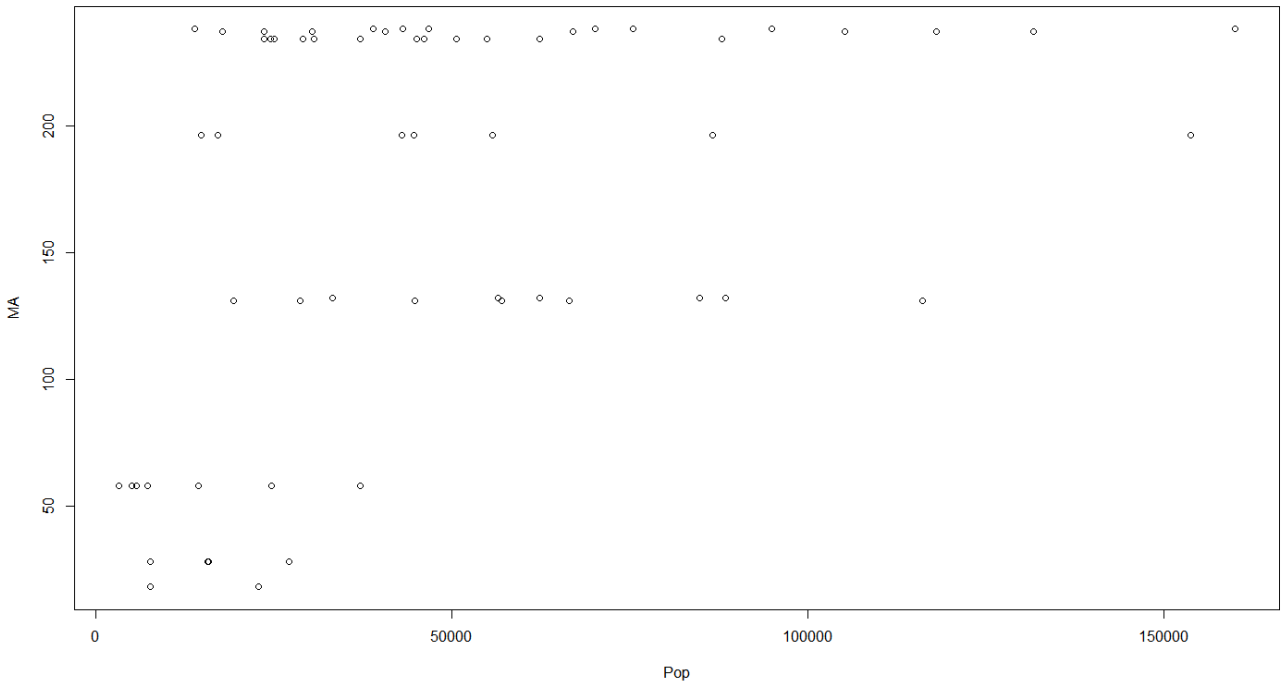


토지면적의 제1사분위수는 2.79, 중앙값은 2.79,제3사분위수는 4.55이고 범위내 최대 관측값 15.23다. 왼쪽으로 굉장히 많이 쏠린 것으로 보아 최대치와의 차이는 굉장히 크다고 볼 수 있다. 위 지가변동률에서 본 것과 같이 토지면적 10에서 다른 것과는 비교될 만큼 높은 가격대를 보이는 동이 존재함을 알 수 있다.

Q. 3년 평균 혼인건수와 인구수 관련 분석

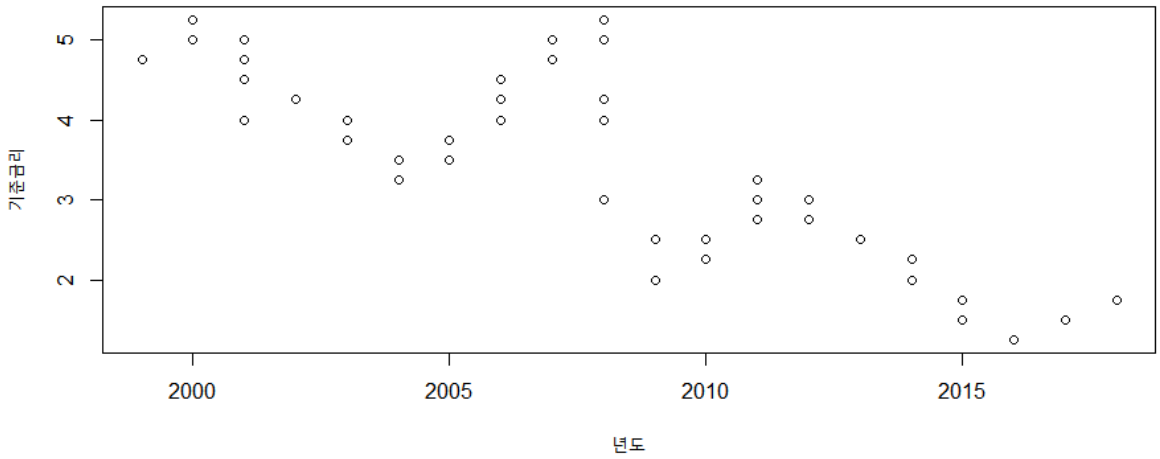


구별 평균혼인건수~인구수(산점도)



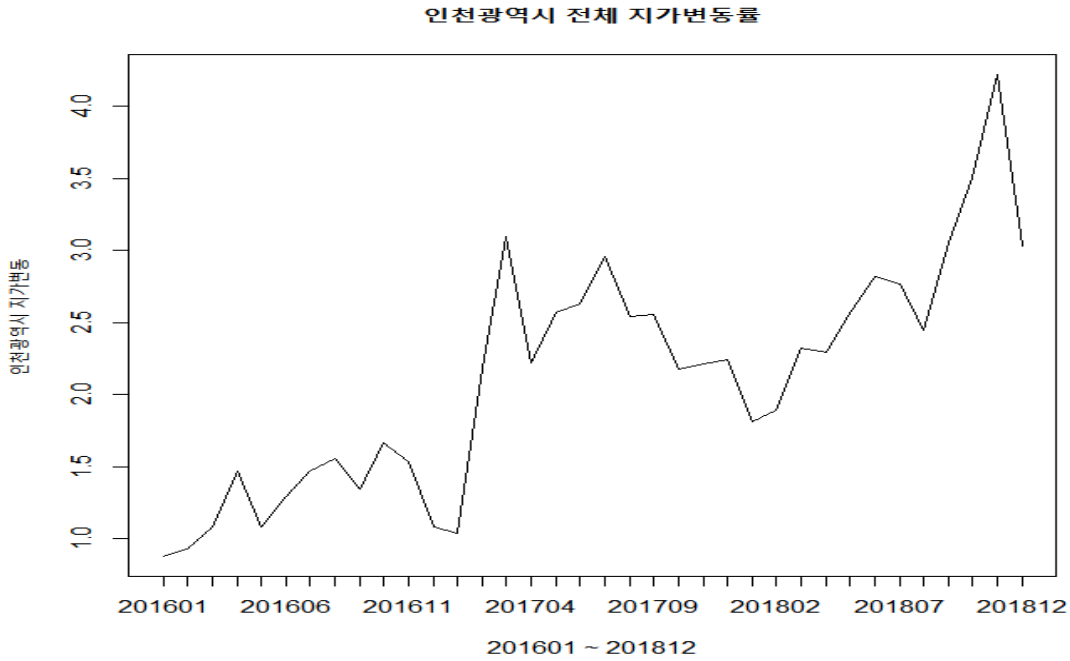
인구수가 적은 곳에서 혼인건수(100건 이하)가 적은 것을 알 수 있다.

R. 기준금리 흐름 (1999년도~2018년도)

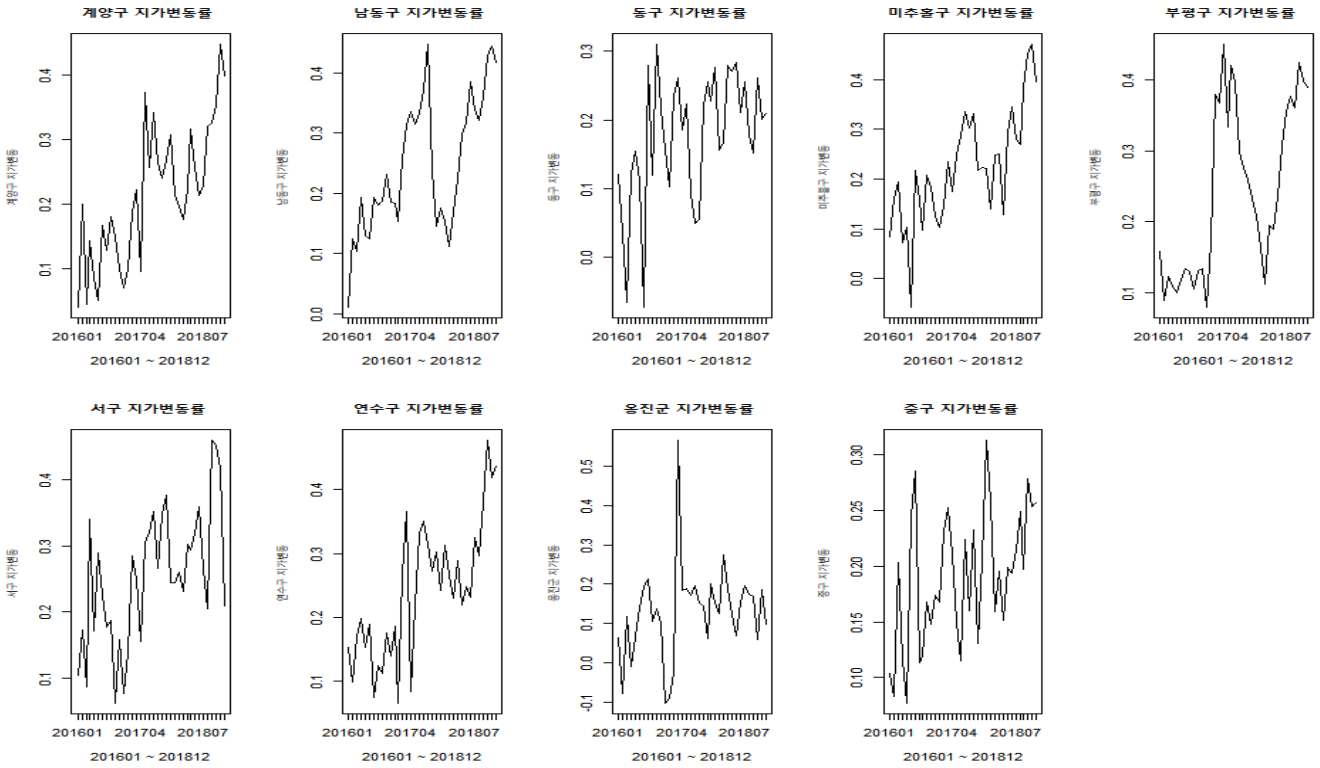


전체적인 추세가 감소하는 형태임을 확인할 수 있다.

S. 지가변동률 시계열 분석



3년간 인천 전체 지가변동률은 전반적으로 증가하는 추세이다.



3년간의 지가변동률을 차례대로 살펴보면 계양구, 미추홀구, 서구, 연수구, 중구는 전반적으로 증가하는 추세이다. 남동구는 2017년에 하락하다가 다시 증가하고 있다. 부평구는 2016년은 낮은 추세를 보이다가 2017년와 2018년 사이에 큰 변동이 있었다. 동구는 2016년 후반기이후로 비슷한 추세를 보인다. 옹진군은 전체적으로 낮은 지가변동률을 보인다.

3. 모델링 적용

A. Data Set 구분

모델링을 학습, 검증, 테스트하기 위해서 Train, Validation, Test set을 4 : 3 : 3의 비율로 랜덤추출을 통해 구분했다.

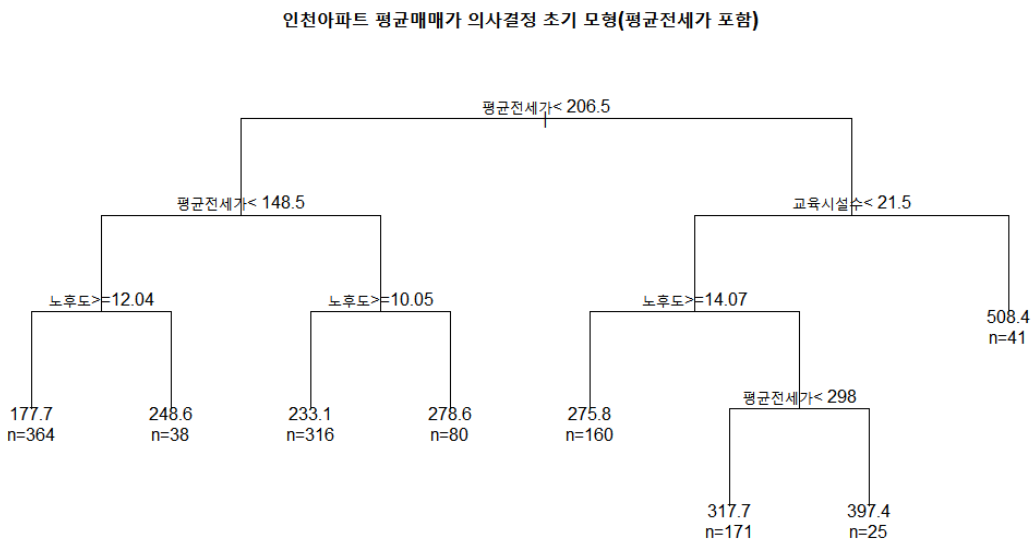
Train set을 이용해 의사결정학습법(Decision tree)과 회귀모형(Regression) 학습하고 Validation set을 통해 최적의 모형 선택 후 Test set을 통해 모형의 설명력을 판단한다.

Data Set 종류	Data 수
Train set	1195
Validation set	895
Test set	895

B. 모델링 – 의사결정학습법(Decision Tree)

R에서 많이 쓰이는 의사결정학습법 package 중 rpart package 이용하여 모델링 진행하고 목적변수가 연속형이라 회귀 의사결정학습법을 적용한다.

i. 초기모형 - Train set을 통한 의사결정학습법 결과

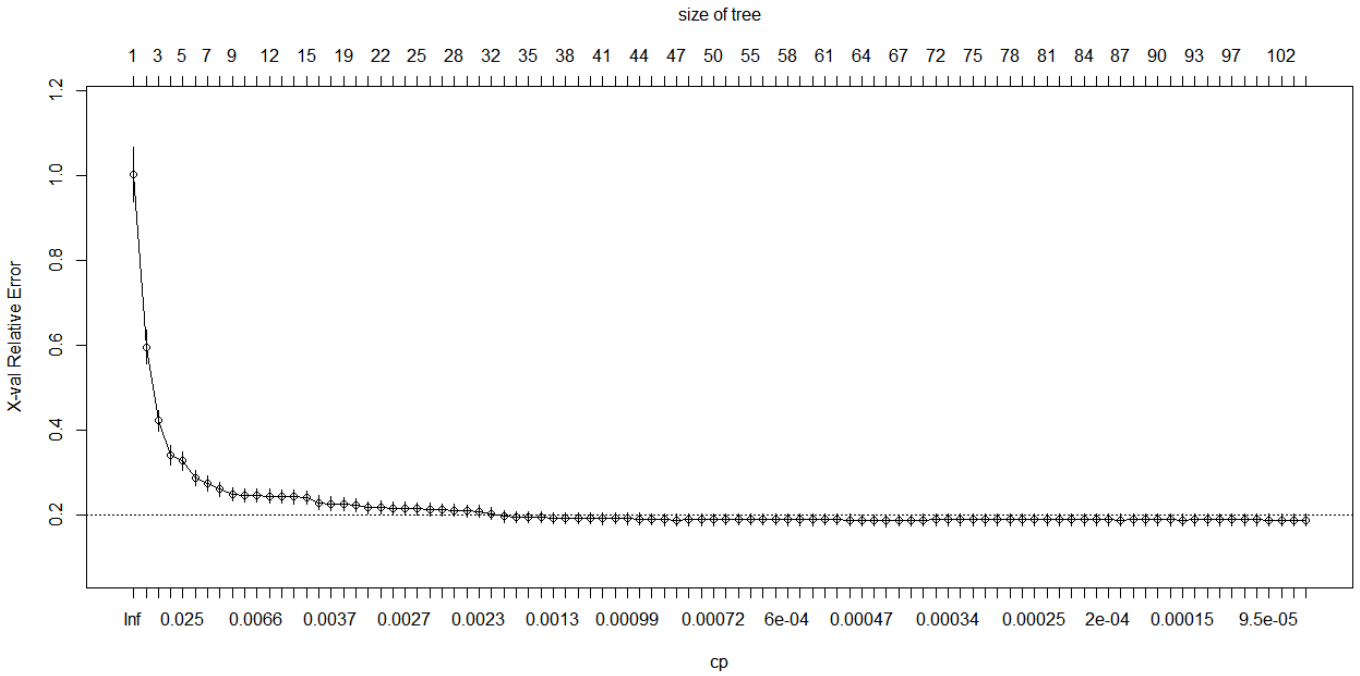


<그림 1>

초기 모형의 경우, Terminal node의 수는 8개이고, 노드 분리에 있어 가장 영향을 주는 변수가 "평균전세가"임을 알 수 있습니다. 그 다음으로는 "노후도"임을 알 수 있다. 앞서 EDA 시 교육시설 수가 22개인 동이 송도동임을 고려하면 508.4는 송도동 관련 데이터가 많이 포함되어 있다고 짐작할 수 있다.

ii. Pruning 적용

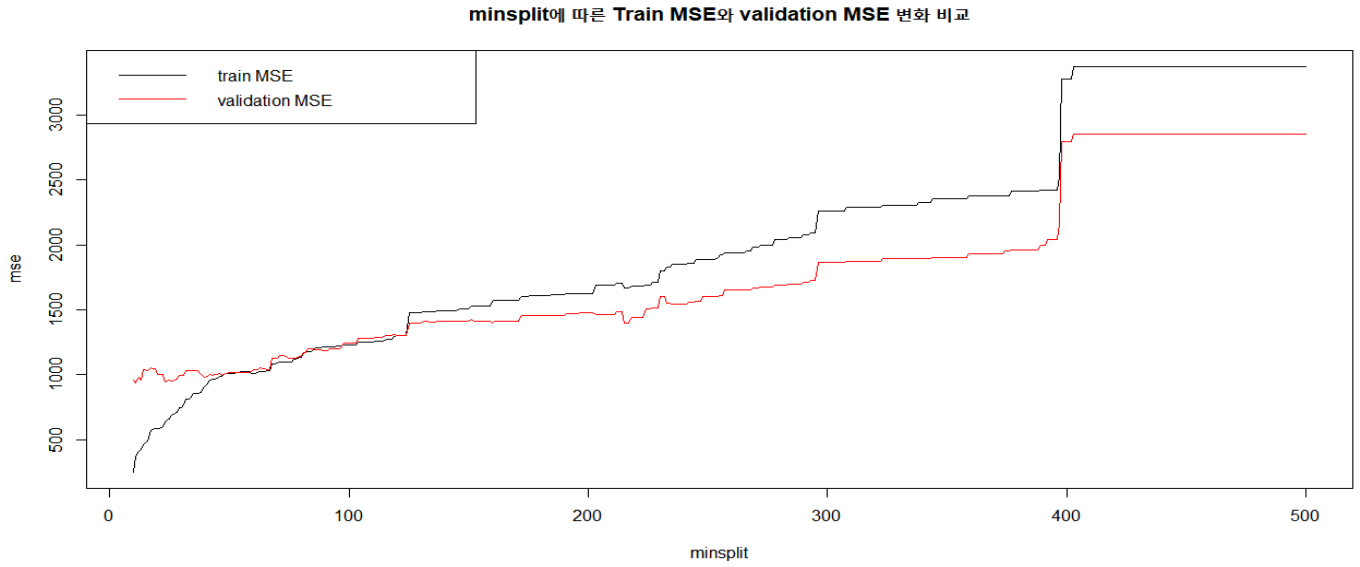
초기모형 그림1의 경우 rpart 함수의 적용 결과만을 판단한 것이기에 정확한 Pruning을 위해 Complexity parameter(오분류값)이 0이라는 조건을 두어 가장 낮은 오분율에서 적절한 terminal node 수가 몇 개인지 판단하기로 했다.



<그림 2>

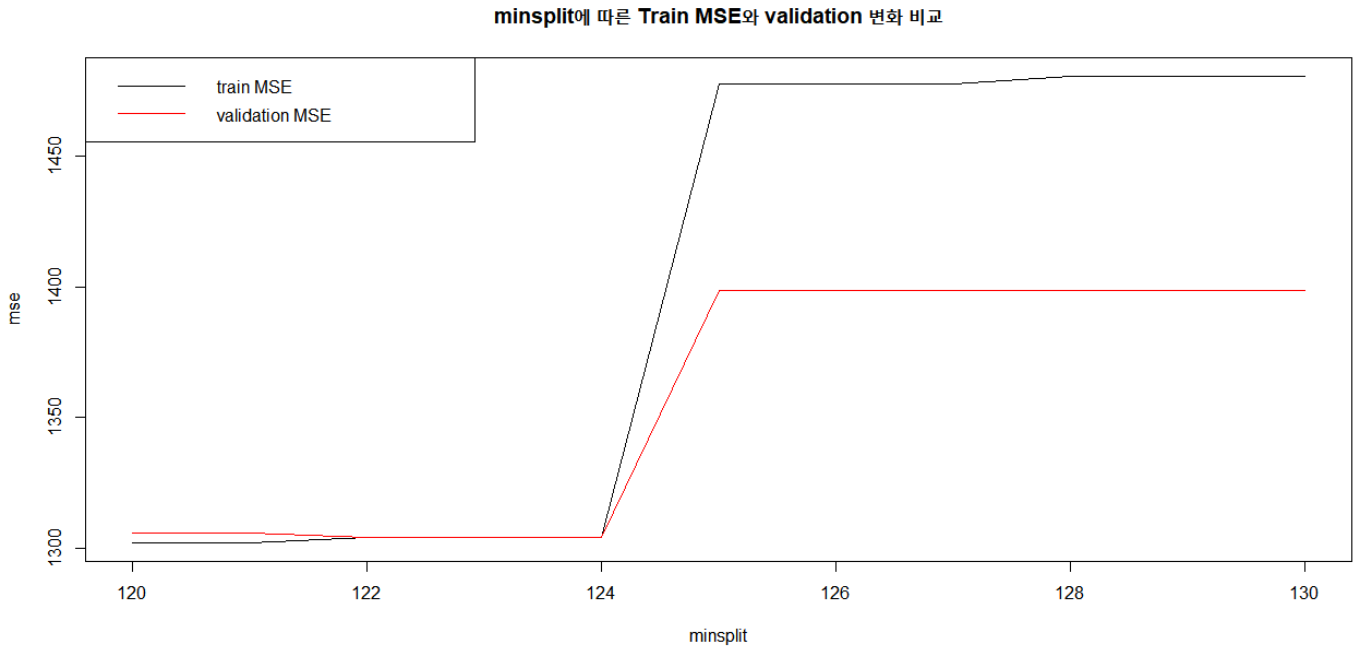
그림2를 보시면, terminal node의 수가 증가할수록 오분율 값이 계속 낮아지는 것을 확인할 수 있다.

그러므로 적절한 terminal node의 수를 정하기 위해 Train set과 Validation set을 통해 terminal node수에 따라 각 MSE의 변동을 비교하여 최적의 terminal node수를 정하기로 했다. 먼저 minsplit(노드 분할하기 위해 필요한 최소 데이터의 개수)변동은 결국 terminal node 수와 연관 있기 때문에 minsplit에 따라 Train set과 Validation set의 MSE을 측정했다.



<그림 3>

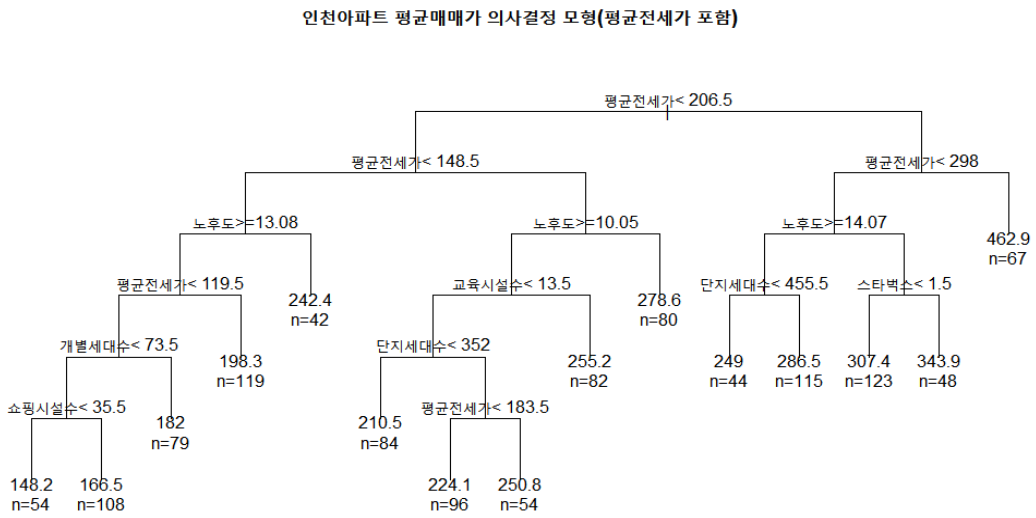
Minsplit이 낮을수록(terminal node 수가 많을수록) Train set과 Validation set의 MSE 또한 낮아지는 것을 확인할 수 있다. 다만, 어느 시점부터는 Validation set의 MSE가 Train set의 MSE보다 높아지는 추세를 확인할 수 있다.



<그림 4>

<그림3>과 <그림4>를 보면 minsplit이 500부터 125까지 validation MSE가 train MSE보다 계속 낮다는 것을 확인할 수 있습니다. 그러므로, <그림2>의 결과와 <그림4>의 분석에 따라 minsplit이 125일 때 지점을 선택하여 Pruning을 진행했다.

iii. 최종 모형 회귀 의사결정 학습법 결과



<그림 5>

<그림 5>에 대해 Train MSE, Validation MSE, Test MSE을 비교한 결과

Train MSE	1477.903
Validation MSE	1398.396
Test MSE	1303.861

세 개의 MSE 값 중에서 Test MSE가 가장 낮은 것을 확인할 수 있습니다.

종합적인 판단 결과 terminal node수가 15개를 가진 회귀 의사결정 모형으로 선택했다. Root node을 나누는 결정적인 변수가 "평균전세가"인 것을 알 수 있으며, 또한 이후 노드 분할에 있어 가장 많이 기준이 되는 변수는 평균전세가"임을 알 수 있다.

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

위의 평균전세가, 노후도, 단지세대수가 인천아파트 평균매매가 형성에 많은 영향을 주는 변수라는 것을 의미한다고 볼 수 있습니다.

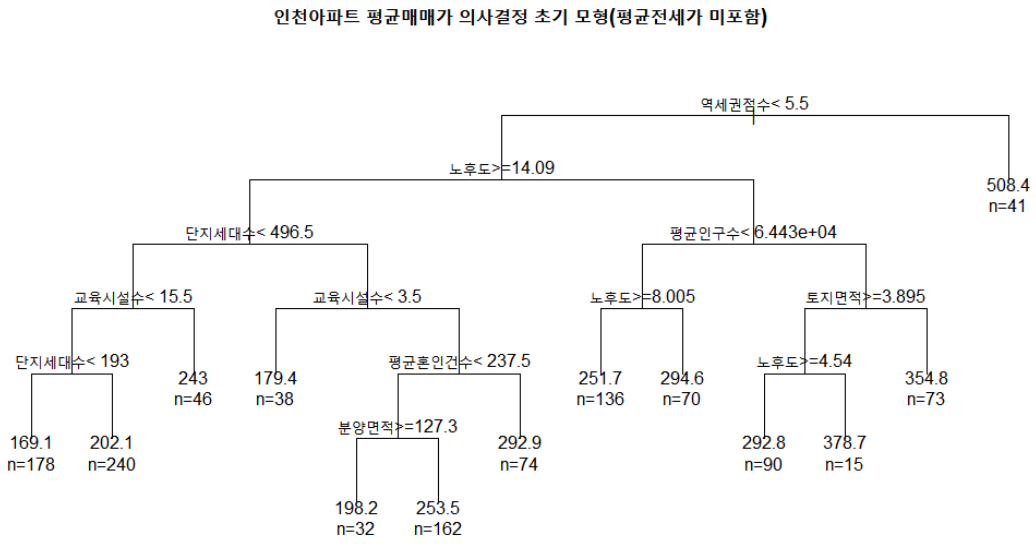
iv. 결론

회귀 의사결정학습법을 통해서 인천아파트 평균매매가를 분석한 결과, 아파트와 관련된 내부변수, 즉 평균전세가, 노후도, 단지세대수가 다른 외부변수에 비해 인천 아파트 평균매매가에 많은 영향을 준다는 것을 확인할 수 있다. 또한 초기모형(그림1)에서 가장 높은 값을 형성하는데 영향을 주는 변수가 "평균전세가"와 "교육시설 수" 변수였지만, 최종모형(그림5)에는 모두 "평균전세가"임을 확인할 수 있다. 그 이유로는 평균전세가도 평균매매가처럼 다른 변수에 영향을 받은 것을 가격에 반영하기 때문이다. 그래서 평균전세가가 다른 변수들에 비해 평균매매가에 많은 영향을 주는 것으로 짐작할 수 있다.

다음에 의사결정학습법은 "평균전세가"를 제외한다면, 어떤 변수들이 평균매매가에 영향을 주는지 파악하기 위해 진행한다.

C. 모델링(추가) – 의사결정학습법(Dection Tree) "평균전세가" 제외한 의사결정 학습법

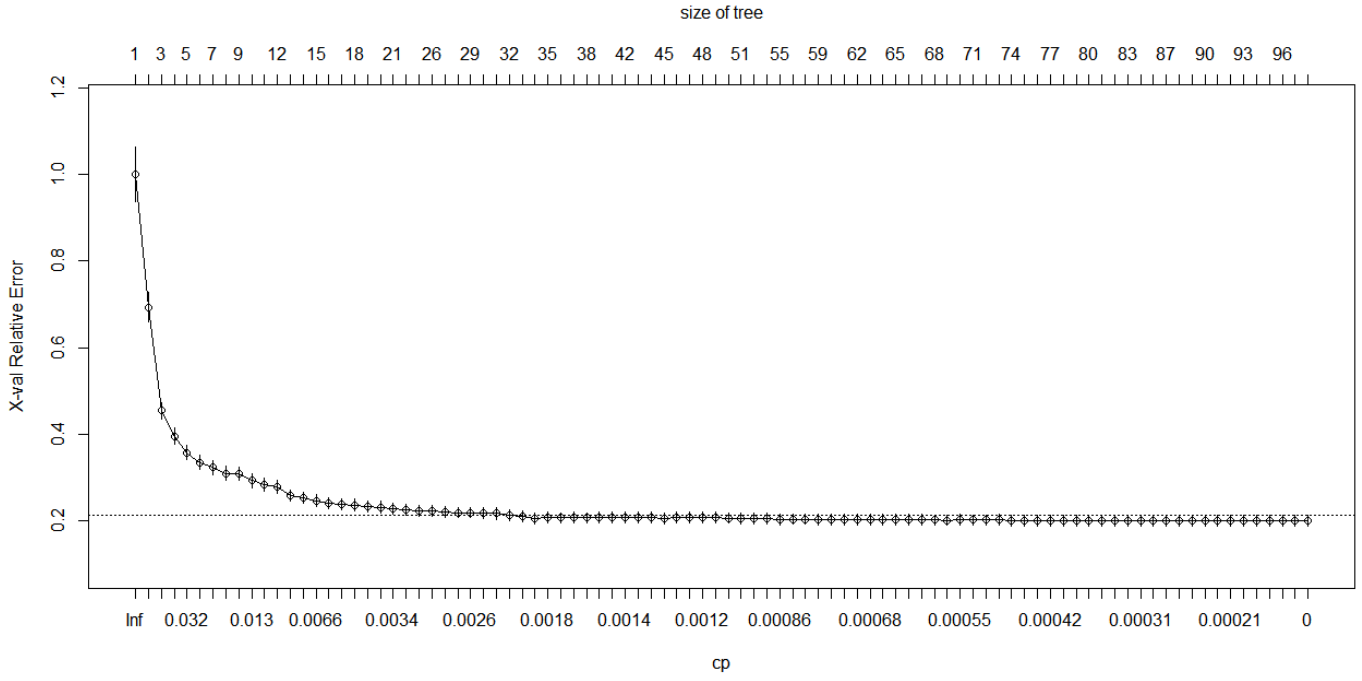
i. 초기모형(평균전세가 제외)



<그림 6>

초기모형 그림6에 따르면, root node 분기 기준이 "역세권점수"임을 파악할 수 있다. 노후도, 평균인구수, 단지세대수, 교육시설수, 토지면적, 평균혼인건수, 분양면적 변수가 분기기준으로 등장했다. 508.4의 경우, 역세권점수가 가장 높은 동이 송도동임을 짐작할 수 있다.

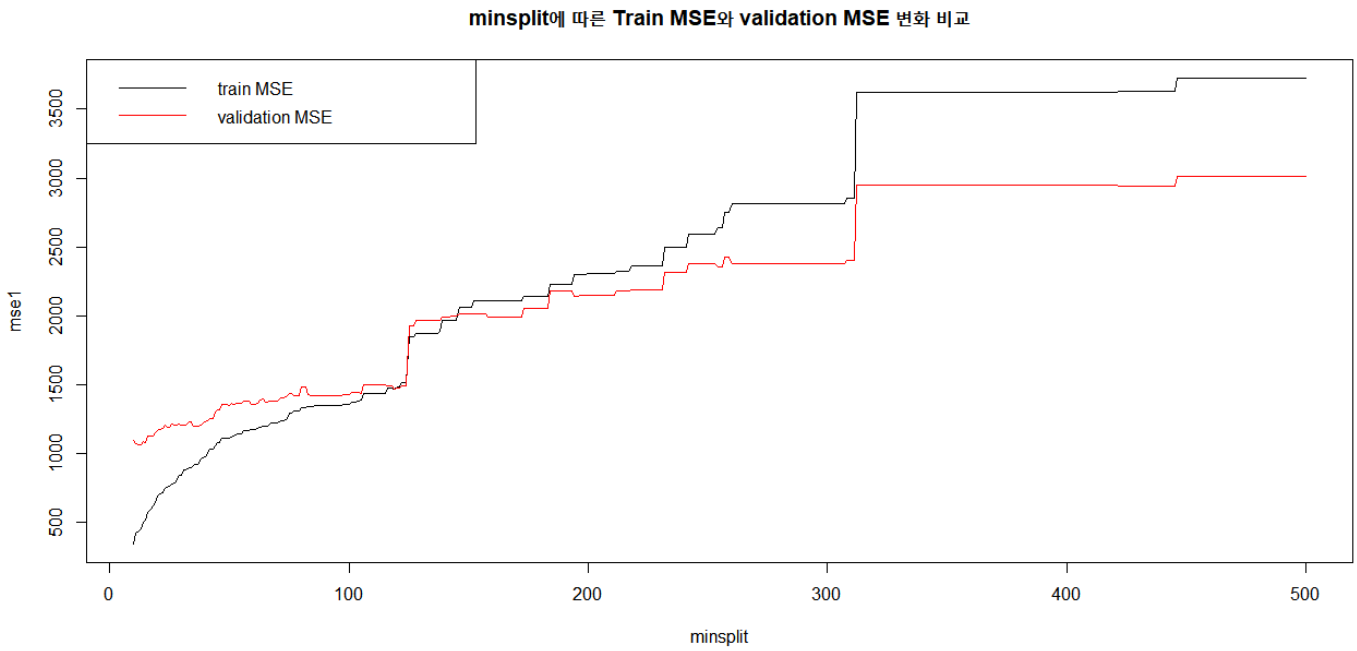
ii. Pruning 적용



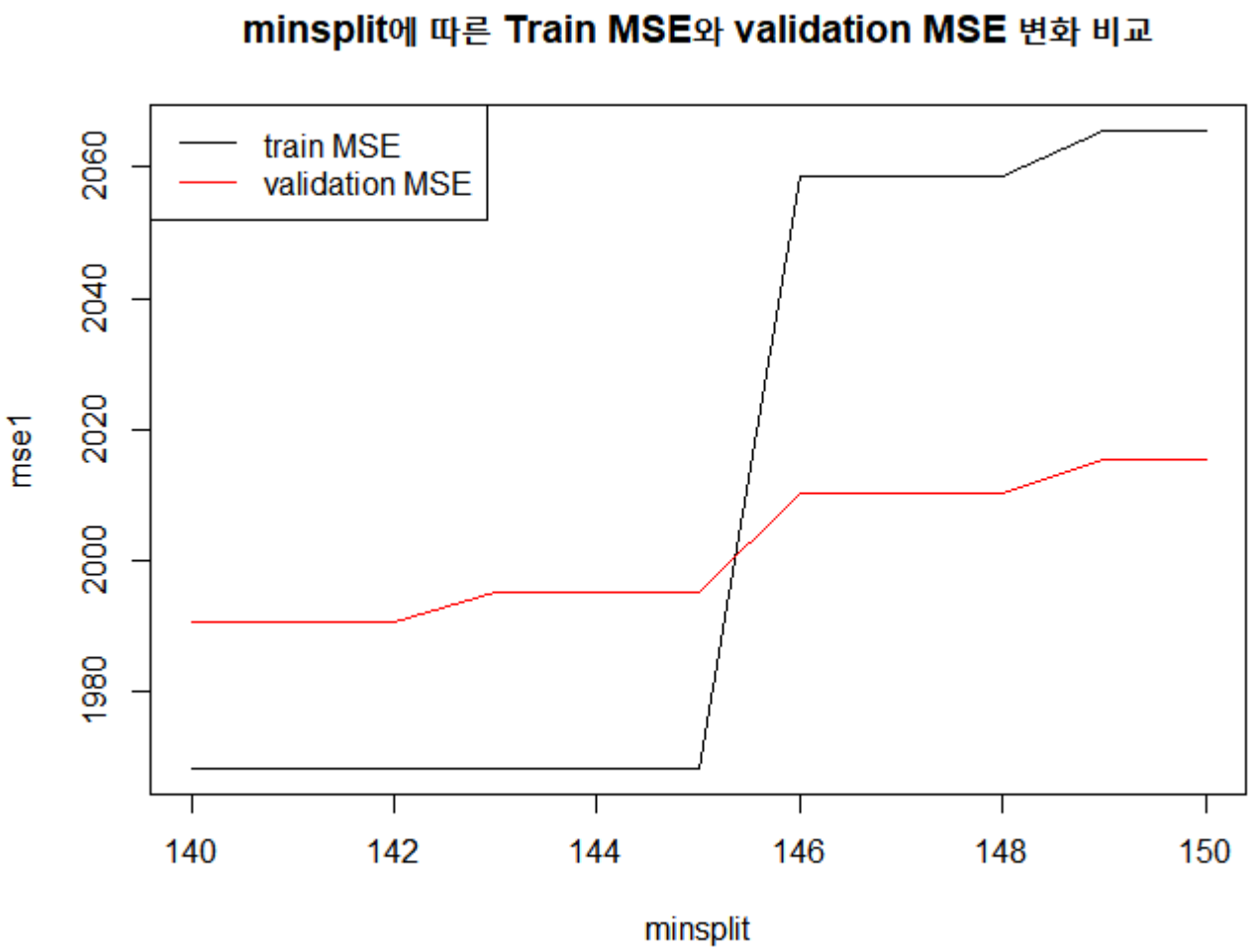
<그림 7>

그림7을 보면 terminal node 수 증가에 따라 오분율이 계속 낮아지는 것을 확인할 수 있다. Train set MSE

와 Validation set의 MSE 비교를 통해 최적의 노드 수를 결정하기로 했다.



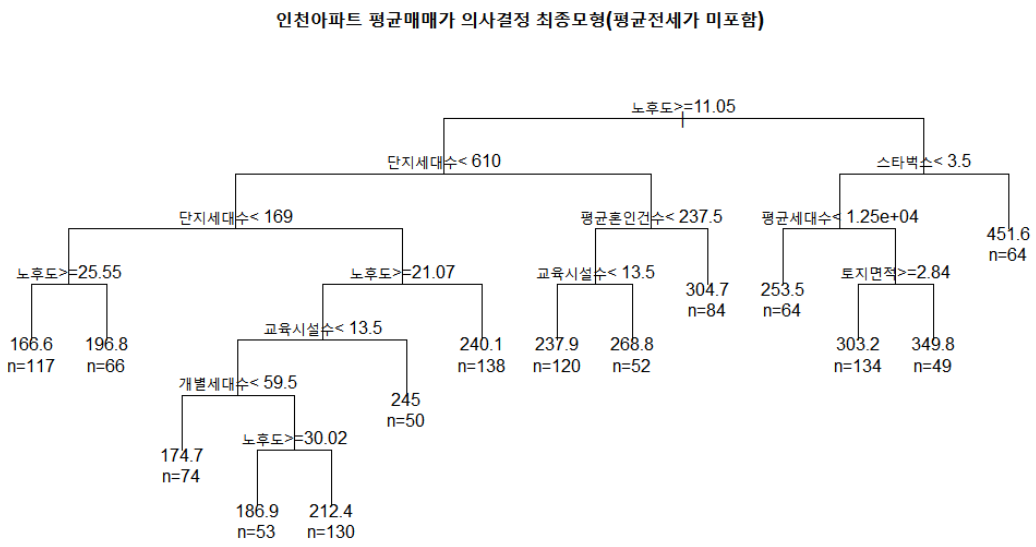
<그림 8>



<그림 9>

그림8과 그림9을 종합하여 판단하면, minsplit이 146에서 선택하는 것이 최적이라는 것을 고려할 수 있다. (terminal node 수는 14개)

iii. 최종모형 결과(평균전세가 미포함 경우)



<그림 10>

최종모형 그림10에 따라 Train MSE, Validation MSE, Test MSE을 비교해보면,

Train MSE	2058.659
Validation MSE	2010.161
Test MSE	2244.963

오히려 Train MSE, Validation MSE보다 Test MSE가 높다는 결과를 볼 수 있다.

변수에 따른 분기 노드의 수를 파악한다면

변수	분기 노드 수
노후도	4
단지세대수, 교육시설수	2
스타벅스, 평균세대수, 개별세대수, 평균혼인건수, 토지면적	1

가장 영향을 주는 변수는 "노후도"로 내부 변수에 해당한다.

iv. 결론(평균전세가 미포함)

평균전세가 미포함한 초기모형과 최종모형을 비교하면, 초기모형과는 다르게 최종모형에서의 root node가 역세권점수에서 노후도로 바뀐 것을 확인할 수 있다.

평균전세가를 포함한 결과와 미포함한 결과를 분석하면, 전반적인 data set의 MSE의 변화를 통해 평균전세가가 인천아파트 평균매매가에 많은 영향을 준다는 것을 유추할 수 있습니다. 노후도 변수 또한 인천 아파트 평균매매가 형성에 상당하게 영향을 주는 것을 유추할 수 있습니다.

D. 모델링 – 회귀분석(Regression)

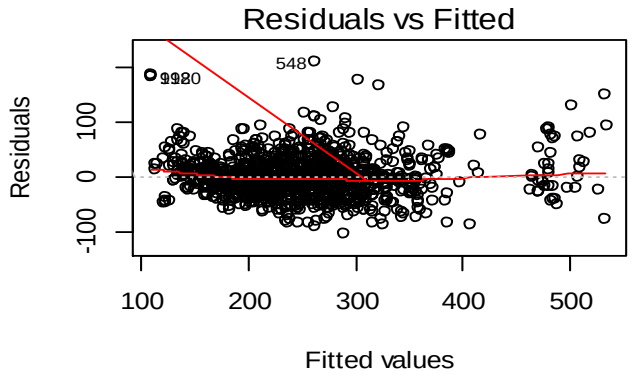
- 독립성(자기상관성), 선형성, 등분산성, 정규성 만족하는지 잔차 분석
- 다중공선성 검정 후 3가지 모형선택법을 이용해서 설명변수 선택하기
- (1)Forward selection , (2)Backward selection, (3) Stepwise selection

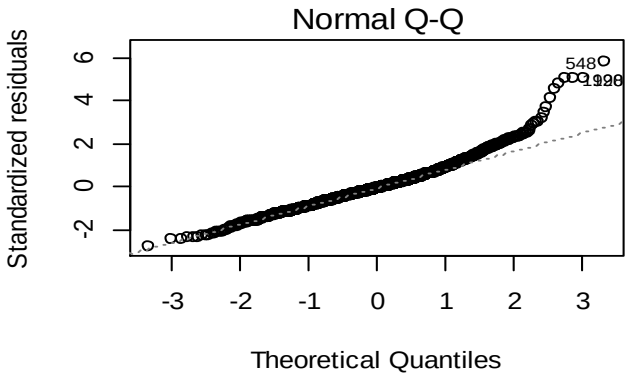
i. 초기모형 설정 및 모형진단

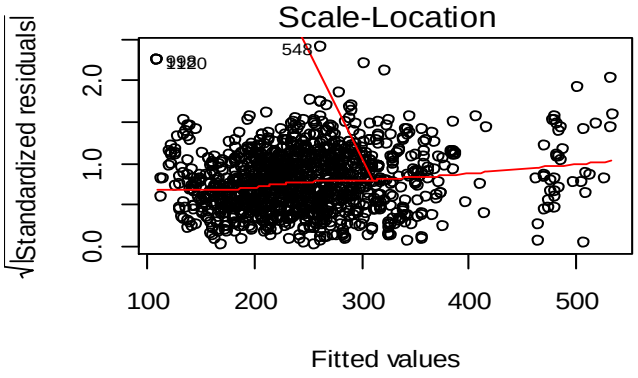
Y : 평균매매가, 종속변수(목적)

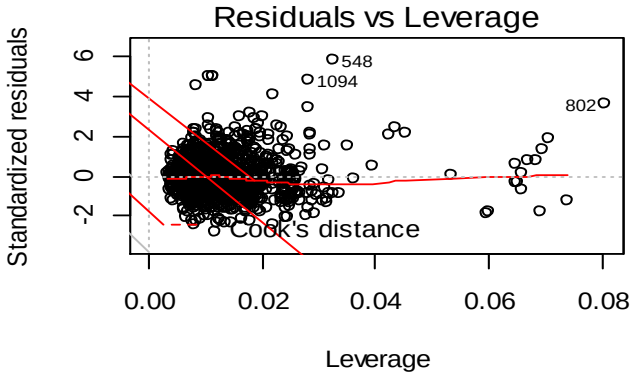
독립변수	설명	독립변수	설명	독립변수	설명
X_1	분양면적	X_6	평균전세가	X_{11}	평균인구수
X_2	단지세대수	X_7	토지면적	X_{12}	평균세대수
X_3	개별세대수	X_8	문화시설수	X_{13}	교육시설수
X_4	노후도	X_9	쇼핑시설수	X_{14}	개발호재
X_5	역세권점수	X_{10}	스타벅스	X_{15}	평균혼인건수

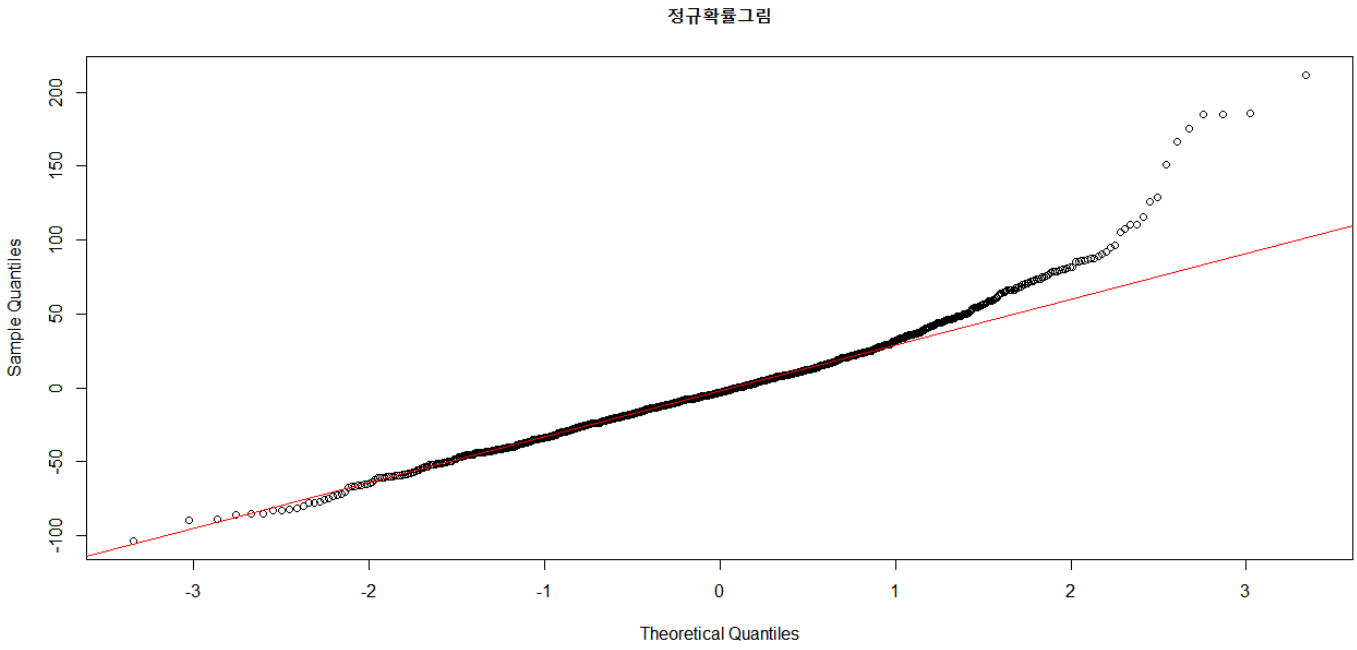
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} + \beta_{15} X_{15} + \varepsilon$$





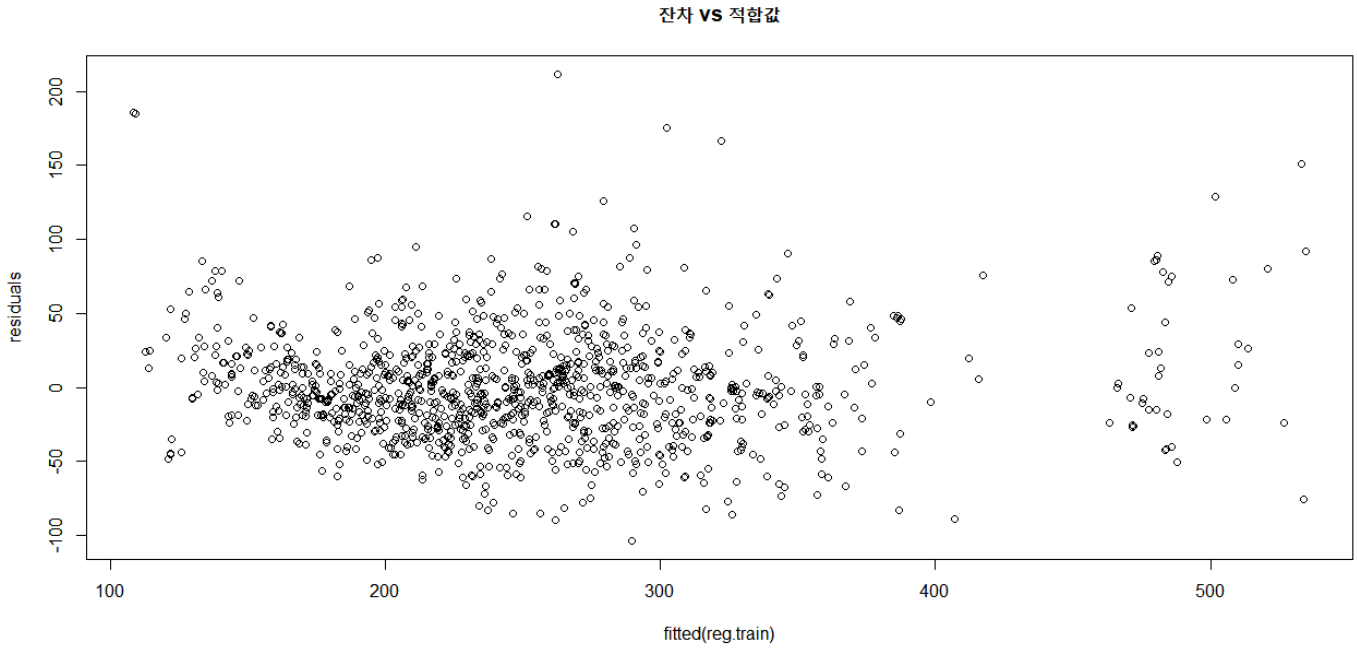




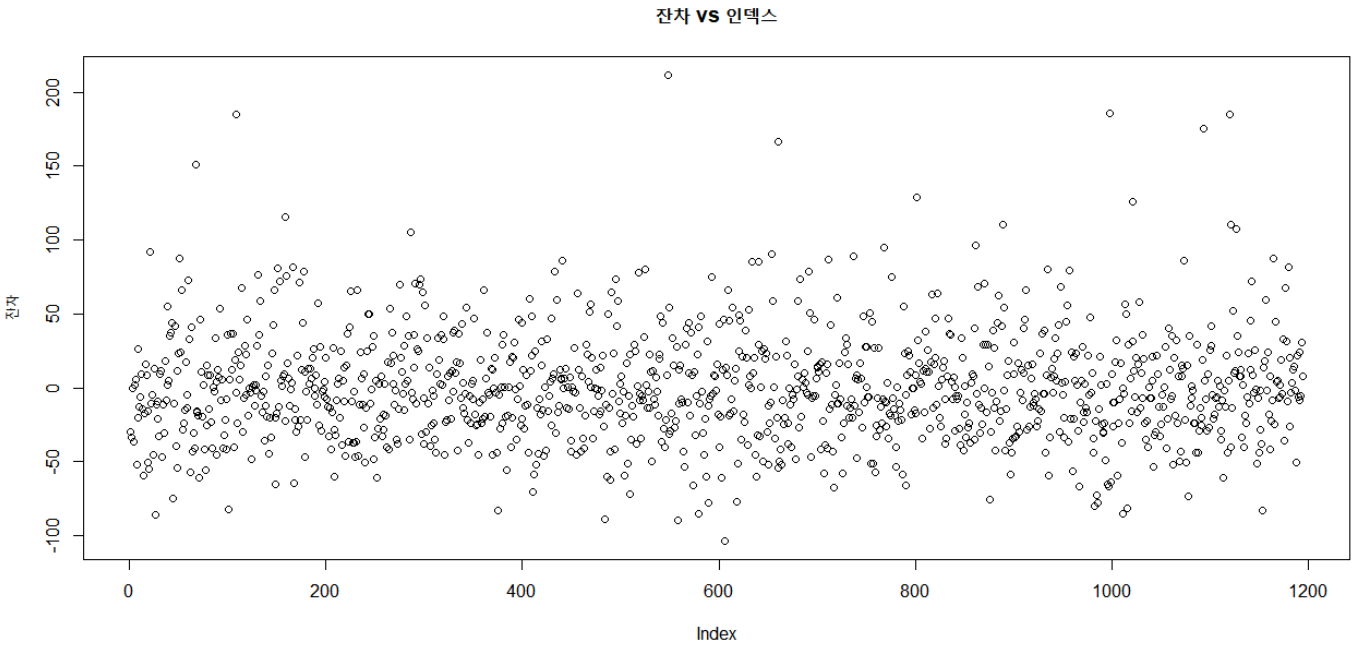


1) 정규성 : 정규확률그림에 따르면 직선을 보이다가 점점 직선을 벗어나는 것처럼 보인다. 하지만 표본수가 굉장히 크기 때문에 중심극한정리에 의해 큰 랜덤표본은 근사적으로 정규분포를 따른다.

2) 독립성 : 더빈왓슨 통계량에 의하면 2.045로 2보다 크므로 자기상관성이 없다고 볼 수 있다.

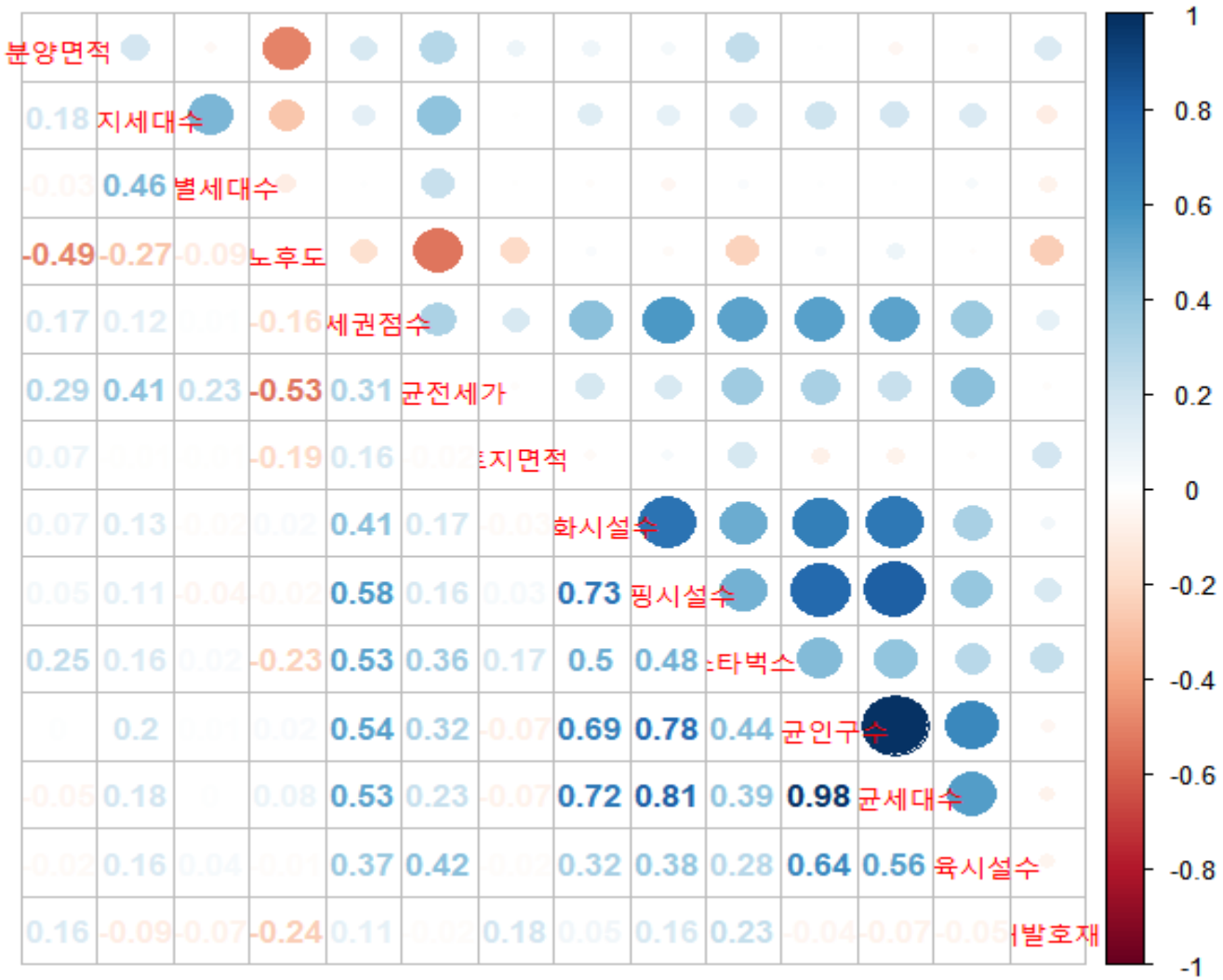


3) 선형성 : 잔차와 예측치 사이에 특별한 형태가 보이지 않으므로 만족한다.



4) 등분산성 : 잔차 대 인덱스 플롯에 따르면 등분산성을 만족하는 것으로 볼 수 있다.

ii. 다중 공선성 확인



다중 공선성 VIF가 10보다 크면 다중 공선성이 존재한다고 볼 수 있는데, 평균인구수(X_{11})와 평균세대수(X_{12}) 사이에서 나타난다.

두 변수 간의 상관관계를 분석했을 때 0.9824로 매우 큰 상관관계가 있음을 알 수 있다.

다중 공선성은 추정된 회귀계수의 불안정성과 관련되어 있으므로 이를 해결하기 위해 변수선택법을 활용하겠다.

iii. 변수 선택법

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 stepwise의 결과가 동일하게 나타나고 결정계수가 forward보다 크므로

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

을 채택한다. 다만 다중 공선성을 고려해서 최종 모형을 고려해야한다.

iv. Validation set으로 초기모형과 stepwise 모형 검증

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

을 Train set을 통해 적합 후 Validation set을 통해 검증을 했다.

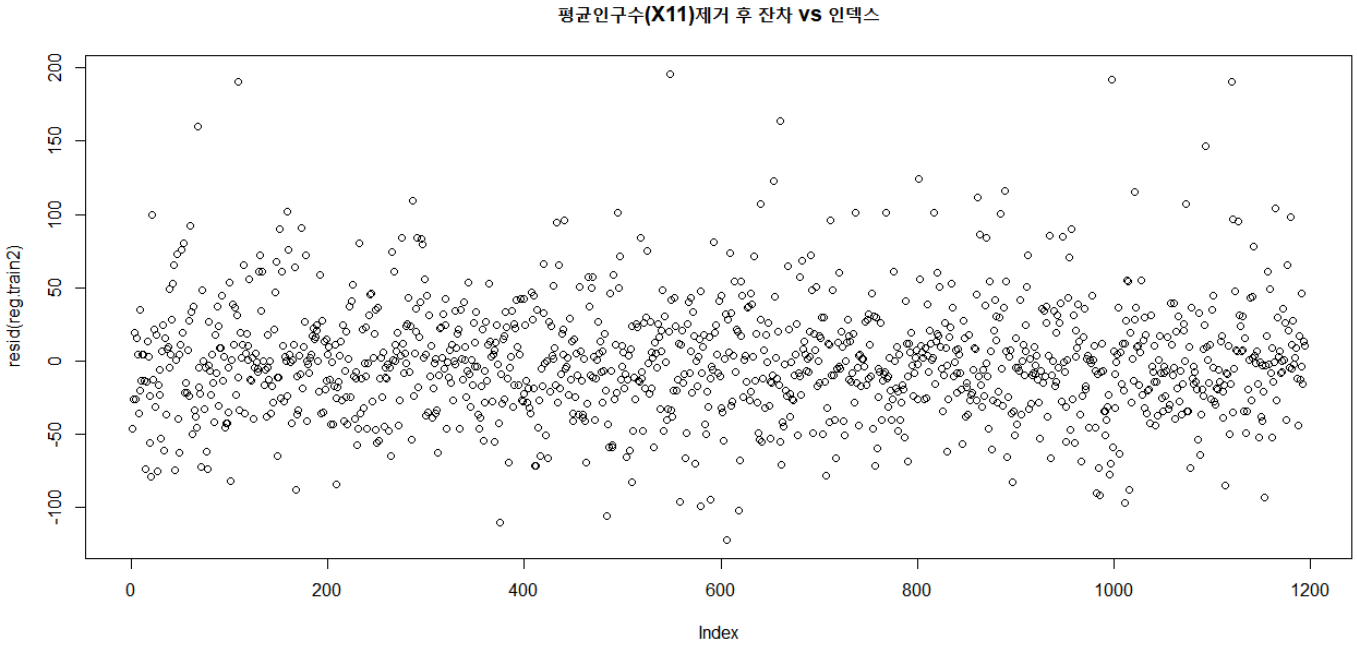
	초기모형(독립변수 15개)	초기 Stepwise 모형(독립변수 13개)
Train MSE	1369.551	1367.26
Validation MSE	1220.949	1217.055

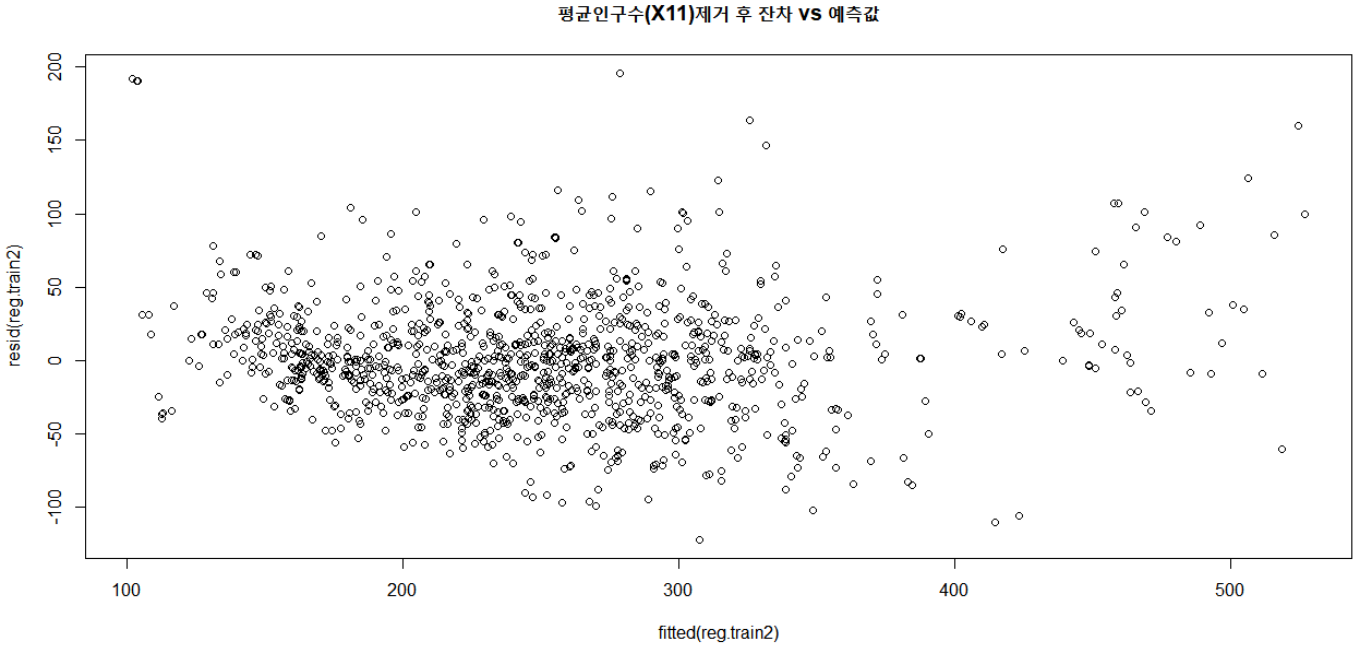
기존의 Validation MSE가 1220.949인데 1217.055로 감소했다. 결정계수는 0.7888로 증가했으므로 설명력이 증가했음을 알 수 있다. 하지만 여전히 평균인구수(X_{11})와 평균세대수(X_{12}) 사이에 공선성이 존재한다. 두 변수를 각각 제거하여 비교해보겠다.

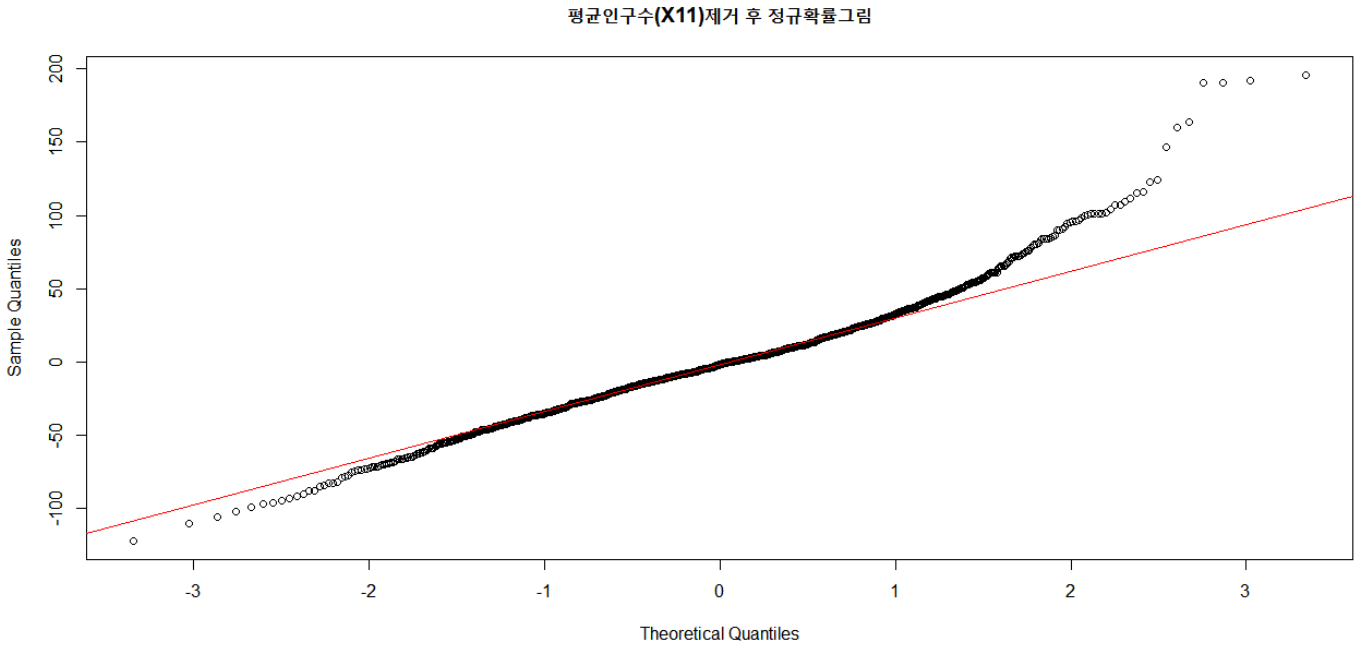
v. 평균인구수(X_{11})를 제거 후 회귀분석 진행

1) 평균인구수(X_{11})를 제거 후 초기모형

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







평균인구수(X_{11})를 제거 후 초기모형에 대한 잔차 진단을 보면, 등분산성, 선형성, 정규성을 만족하는 것을 확인할 수 있다. 독립성의 경우 더빈왓슨 검정에 따라 더빈왓슨 통계량이 2.044992이므로 독립성 또한 만족한다.

2) 평균인구수(X_{11})를 제거 후 Stepwise 변수 선택

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

을 Train set을 통해 적합 후 Validation set을 통해 검증을 했다.

	X_{11}제거 모형(독립변수 14개)	X_{11}제거 Stepwise 모형(독립변수 12개)
Train MSE	1527.122	1525.889
Validation MSE	1347.754	1348.721

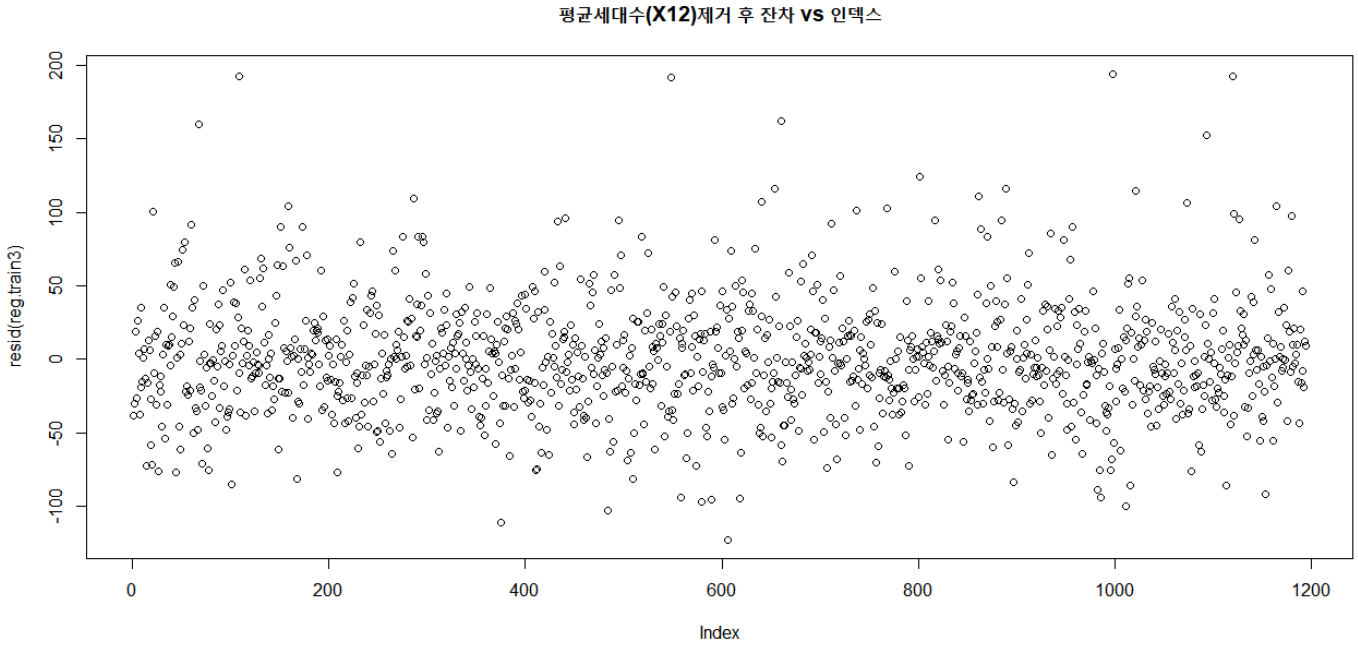
Tranin MSE는 낮아졌지만, Validation MSE는 약간 증가했다.

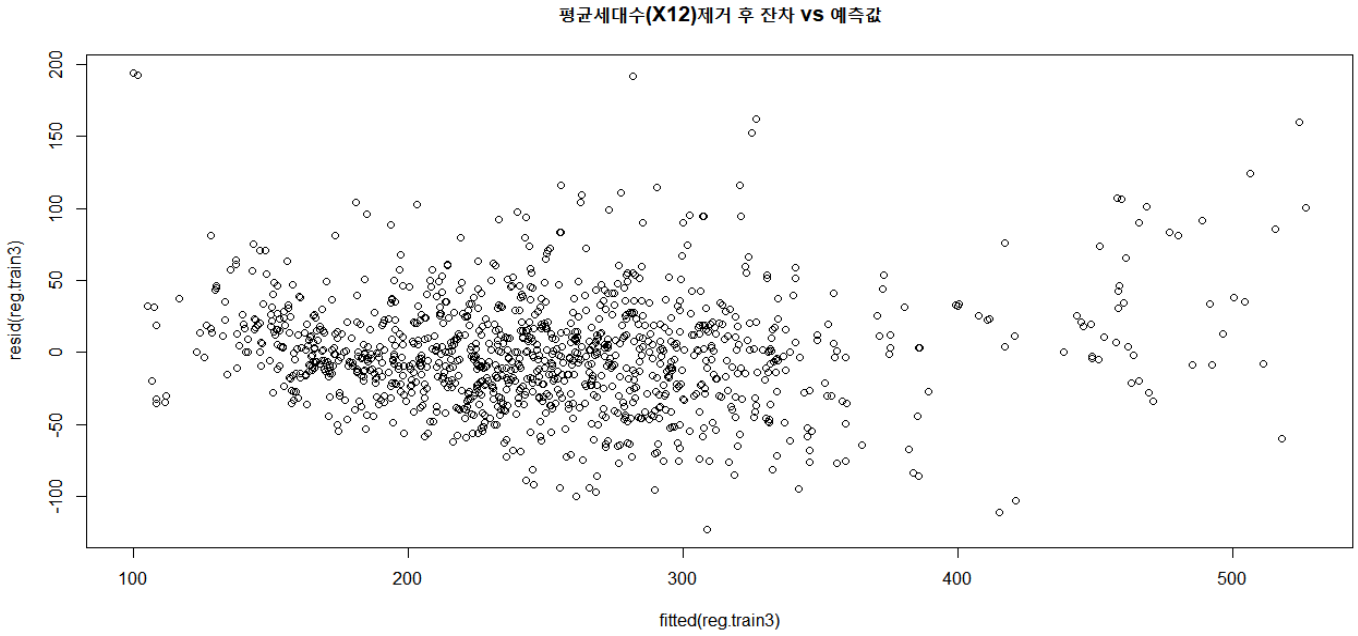
반대로 평균세대수(X_{12})도 제거하여 회귀분석을 진행하도록 한다.

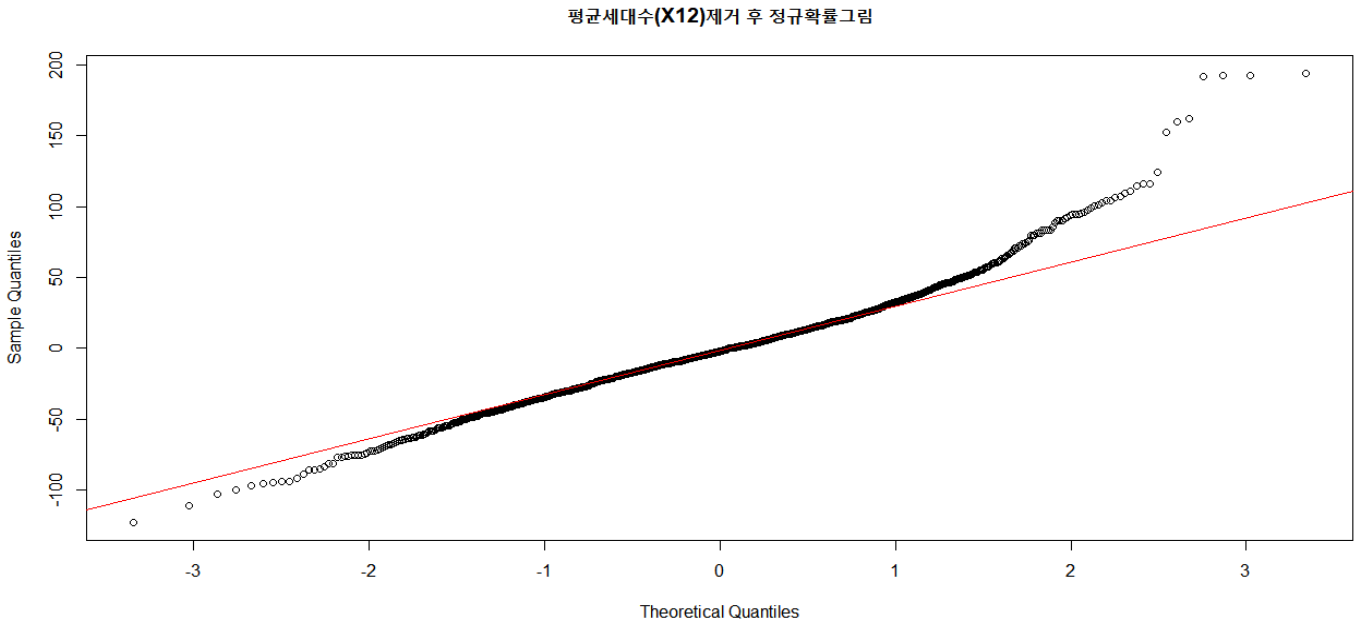
vi. 평균세대수(X_{12})를 제거 후 회귀분석 진행

1) 평균세대수(X_{12})를 제거 후 초기모형

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







평균세대수(X_{12})를 제거 후 초기모형에 대한 잔차 진단을 보면, 등분산성, 선형성, 정규성을 만족하는 것을 확인할 수 있다. 독립성의 경우 더빈왓슨 검정에 따라 더빈왓슨 통계량이 2.048581이므로 독립성 또한 만족한다.

2) 평균세대수(X_{12})를 제거 후 Stepwise 변수 선택

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

을 Train set을 통해 적합 후 Validation set을 통해 검증을 했다.

	X_{12}제거 모형(독립변수 14개)	X_{12}제거 Stepwise 모형(독립변수 12개)
Train MSE	1509.502	1511.268
Validation MSE	1328.958	1334.331

Train MSE와 Validation MSE는 약간 증가했다.

vii. 최종 회귀모형 선택

모든 모형의 MSE를 비교하면

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

최종 회귀모형 선택 시, 다중 공선성을 해결하고, 변수 선택한 모형 중에서 Validaton MSE가 낮은 모형은 선택한다. 그래서 최종모형은 평균세대수(X_{12}) 제거 stepwise 적용한 모형이다.

최종 모형을 통한 Test MSE값은 1474.591로 1334.331인 Validation MSE보다 큰 것을 확인할 수 있다.

viii. 최종 회귀모형 분석

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

에서 추정회귀계수는 다음과 같다.

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

추정회귀계수를 내림차순으로 정렬한다면

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

개발호재, 역세권점수, 스타벅스, 교육시설수, 평균전세가 순으로 확인 수 있고, 집값에 많은 영향을 준다고 잘 알려진 변수들이 상위권에 존재한다.

개발호재가 인천아파트 평균매매가 형성에 많은 영향을 주는 것을 알 수 있다. 개발호재 경우 범주형 변수이므로, 개발호재가 예상되거나, 알려진다면, 제곱미터당 약 19만원이 상승할 것이라고 유추할 수 있다. 스타벅스의 경우 3번째로 인천아파트 평균매매 추정값에 영향을 주는 변수인데, 스타벅스가 주변 상권을 고려해서 입지를 한다는 점과 스타벅스가 입지를 한 위치는 유동 인구 증가와 브랜드 이미지로 인해 주변 상권의 가격이 올라간다는 점을 생각해보면, 아파트 집값에도 영향을 준다고 볼 수 있다. 의외로 평균인구수가 인천아파트 평균 매매가에 영향력이 적다는 것은 앞서 EDA를 보면, 전체 인천 인구수는 지속적으로 증가했지만, 세부적으로는 몇몇 구의 경우 인구수가 감소하는 것을 확인할 수 있다. 하지만 인천 구군별 아파트 평균 매매가는 지속적으로 상승했다는 것과 같이 생각해본다면 상쇄되는 영향으로 인해 상대적으로 추정회귀계수가 낮다는 것을 추론할 수 있다.

반대로 평균혼인건수, 쇼핑시설수, 토지면적, 노후도의 경우 인천 아파트 집값을 떨어뜨리는 변수라는 것을 확인할 수 있다. 평균혼인건수의 경우, 우리나라 젊은 층 사이 비혼 증가와 1인가구수 비중이 증가 등을 짐작할 수 있다. 토지면적의 경우, 인천아파트 평균매매가격이 0과 $10km^2$사이에 대부분 집중 분포되어 있기 때문이란 것을 추론할 수 있다. 노후도는 예상되는 결과이다. 신축이 비싸고 노후도가 1년 증가할수록 집값은 제곱미터당 2만원씩 떨어진다는 것을 확인할 수 있다.

4. 결론

A. 모형평가

i. Decision Tree

초기모형은 노드 8개로 만들어졌고 의미 있는 변수가 '평균전세가'와 '교육시설수' 였다면, Validation 과 Train set의 MSE를 기준으로 적정 terminal node를 15개로 의사결정 모형을 만든 결과로는, root node인 '평균전세가' 부터 '노후도', '단지세대수', '교육시설수' 등등의 순으로 의미 있는 변수임을 알 수 있었습니다. '평균전세가'를 포함한 결과와 미포함한 결과를 분석하면, 전반적인 Data set의 MSE의 변화를 통해 '평균전세가'가 인천아파트 '평균매매가'에 많은 영향을 준다는 것을 유추할 수 있습니다.

ii. Regression

회귀모형의 최종모형 선택은 다중공선성의 제거를 위해 평균세대수(X_{12})변수를 제거하고 Stepwise을 통해 변수 선택된 모형이고, 그 결과 validation MSE도 train MSE에 비해 낮아짐을 알 수 있었습니다. 의사결정모형과는 다르게 '개발호재' 변수의 coefficient가 약 18.85로 가장 높았고 '역세권점수', '스타벅스', '교육시설수' 등의 순으로 유의한 결과가 나타났습니다.

B. 기대효과

대부분 수요가 안정적인 도시는 전세가에 따라 매매가가 연동되는 성향이 짙다고 하는데, 지역마다 다르지만 이는 보통 아파트의 매매가에서 전세가가 차지하는 비율인 전세가율이 75% 가 넘어가면 아파트 매매가도 보통 상승한다 라는 법칙에 따라 두 변수가 굉장히 상관관계가 높다고 보여집니다. 인천의 아파트의 경우 재건축 계획, 현황이 서울보다는 적은 편이라 고령화된 아파트일수록 집값도 대체로 떨어지는 추세를 가지고 있습니다. 게다가 신축아파트들도 송도, 청라 같은 신도시에 많이 지어져 아파트 노후도가 매매가에 꽤 큰 영향을 주는 것을 알 수 있습니다.

역세권의 인기는 교통 환경이 열악한 지역일수록 높습니다. 특히 서울과 가까운 인천이나 경기도지역은 역세권이냐 아니냐 에 따라 집값이 크게 양분되는데, 화성시 청계동의 경우 동탄역을 기준으로 바로 인접해 있는 아파트와 도보로 사용하기 힘든 아파트의 가격이 $3.3m^2$당 800만원이나 차이가 나는 만큼 역세권이 아파트값에 큰 영향을 끼치는 것을 알 수 있습니다. 스타벅스의 경우 모든 매장은 본사에서 직접 운영, 관리하는 직영점으로만 운영되고 있어, 본사에서 건물 인근의 상권 등 조건을 따진 뒤 입점 가불가를 결정하기 때문에, 아파트 가격에 영향을 주기 전 영향을 받는 경우가 커 보입니다. Stepwise selection을 통해 제거된 변수 중 하나인 '개별세대수' 변수는 지역에 상관없이 보통 균일한 편으로 분양 면적당 50~150세대 정도 되기 때문에 아파트값에 큰 영향을 끼치지 않아 보이고 '문화시설수' 변수에 포함된 공원의 개수 또한 해당 동에 많다 해서 매매가에 큰 상관이 없어 보입니다. 이러한 의미 있는 여러 변수들을 통해 인천의 아파트 매매가격을 2019년 이후로도 어느정도 잘 예측할 수 있을 것이라고 전망이 됩니다.

C. 한계점

i. 분석 전 한계

세부적인 아파트 단지별로 분석이나, 조사된 변수들의 자료 부족으로 완벽한 예측이 힘들 것입니다. 전문가도 예측하기 힘든 집값 조건들의 다양한 외적 변수와 예상치 못한 변수들까지는 전부 포함시키지 못하는 점과, 모델링 적용에 있어 학사 수준 분석 방법이 실제 적용하기에는 다소 무리가 있을 수 있다는 한계가 있을 수 있습니다. 또한 실거래가 경우, 시차를 두고 거래를 하는데 거래건수마다 각각 다르기 때문에 수집하기 힘들다는 부분과 최근 3년간 거래된 데이터이기 때문에 예상치 못한 미래의 사건, 사고에 대해서는 예측의 정확도의 한계가 존재합니다. 단지 내 아파트에도 층 별, 방향 별 등 세부적인 사항보다는 단지를 기준으로 예측하였습니다.

ii. 분석 후 한계

의사결정모형의 Pruning 단계에서 terminal node수가 많을수록 오분율이 계속 낮아지는 추세를 갖기에 적정한 수준의 terminal node를 선정하는데 어려움이 있었고, 회귀모형의 경우 다중 공선성을 해결하기위해 변수를 제거 및 선택을 했음에도 불구하고 test set의 MSE가 validation set의 것보다 크게 나온 것이 해석의 어려움이 있었습니다. 그 외에도 인천공항이 있는 중구의 운서동의 스타벅스가 가장 많은데 아파트값이 낮은 편이거나, 송도 같은 특정동만 집값이 높은 곳의 해석, 그리고 회귀분석시 cook's distance에서 outlier들이 보이는 변수들 해석의 어려움이 있었습니다.