텍스트이(가) 표시된 사진

자동 생성된 설명

**예측 모델 학습을 통한 향후 5년 서울시 동별 부동산 매매 최적효율 예측 데이터 모델링**

**201721212 이용준**

최근 5년동안 수도권 아파트 거래금액이 가파른 상승폭을 보이고 있다. 여러 부동산 정책의 변화로 상승한 아파트 가격의 변화로 인해, 2020년도부터 시작해서, 현재 2021년까지 이어지고 있는 주식투자 유행을 넘어 부동산 투자의 유행으로 이어지고 있다. 앞으로의 2030세대가 수도권, 서울 지역에서 아파트를 마련하기는 사실 매우 힘든 상황이다. 하지만, 힘든 상황이라 해도 서울 아파트 매입을 포기할 수는 없는 현실이다. 따라서 필자는 그나마 적은 자금으로 서울의 어느 지역의 아파트에 투자해야 하는지 알아보고 또한 어떤 평수의 아파트 매입이 더 효율적인 투자를 위한 것인지 데이터과학 프로세스를 통해 파악해보도록 하겠다.

데이터과학 프로세스는 데이터수집 - 전처리 및 가공 – 탐색 – 분석 – 예측모델 - 시각화의 과정을 수행한다. 이를 순서대로 수행하며 적합한 투자 지역을 선출해보자.

1. **데이터 수집**

먼저 서울의 아파트 거래금액을 파악할 수 있는 자료를 선정해야 한다. 필요한 자료를 공공데이터 포털과 서울 열린 데이터 광장에서 적절한지 판단하여 수집한다. 필자는 서울 열린 데이터 광장에서 ‘아파트’키워드를 통해 ‘서울시 동별 아파트 매매거래 현황.csv’ 엑셀파일을 수집할 수 있었다. 이 파일로 데이터 과학 프로세스를 수행한다.

자료 출처:

<서울 열린 데이터 광장>

🡪 https://data.seoul.go.kr/

<서울시 동별 아파트 매매거래 현황>

🡪 https://data.seoul.go.kr/dataList/OA-20578/S/1/datasetView.do

<데이터 정보>

공개일자: 2021.04.29 최신수정일자: 2021.04.29 저작권자: 서울특별시

제공기관: 서울특별시 제공부서: 스마트도시정책관 공간정보담당관

관련태그: 아파트, APT, 부동산, 동별 매매

(사업특성상 조사에서 빠진 연도가 있습니다. (예산 등의 사유) 🡪 누락연도: 2016, 2017

1. **데이터 전처리 및 가공**

‘서울 동별 아파트 매매거래 현황.csv’ 파일을 읽어드린 다음, apartment변수에 저장하였다. 이 데이터에 결측값 존재여부를 확인한다. 이미 파일자체가 어느정도 서울시에 의해 어느정도 정리된 자료라 결측값은 존재하지 않는 것으로 나타났다.

apartment에는 거래년도, 행정동.코드, 자치구.코드, 자치구.명, 행정동.명, 공간좌표, 소형.60평방미터.이하..거래건수, 소형.거래금액.평균., 중소형.80…85평방미터..거래건수, 중소형.거래금액.평균., 중형.85…102평방미터..거래건수, 중형.거래금액.평균., 중대형.102…134평방미터..거래건수, 중대형.거래금액.평균., 대형.134평방미터.이상..거래건수, 대형.거래금액.평균. 총 16개의 변수가 존재한다. 16개의 변수 중 행정동.코드, 자치구.코드, 공간좌표는 데이터 탐색에 있어서 필요 없는 열들이므로 제거해준다. 또한 변수 중 자치구.명과 행정동.명 변수가 문자형 데이터로 저장되어 있어 이들을 범주형 데이터로 변환시켜 준다. 이를 통해 자치구, 행정동 별 금액차이와 거래건수 차이를 보기 좋게 파악할 수 있다. 마지막으로 가장 중요한 거래금액.평균. 변수들이 모두 문자형으로 저장되어 있으므로 이들을 모두 숫자형 데이터로 변환시켜준다. 이때 결측값이 발생하였다는 경고문이 나왔으므로, 이를 다시 table(is.na())함수로 파악하고 결측값을 제거한 다음, clean\_apartment 변수에 새로 저장하였다.

Clean\_apartment를 확인해본 결과 거래건수가 0인 경우 거래금액.평균.이 0으로 저장되어 있었다. 따라서 이 0값들은 이후 모든 평균 거래금액을 계산할 때, 평균을 낮추는 역할을 하게 되므로, 소형, 중소형, 중형, 중대형, 대형을 각각 apts, aptms, aptm, aptmb, aptb 변수에 저장해주었다.

아래 그림을 통해 clean\_apartment데이터의 문제를 apts등의 변수로 재가공해 저장함에 따라 어떤 문제가 해결되었는지를 한눈에 파악할 수 있다.

텍스트, 영수증이(가) 표시된 사진

자동 생성된 설명텍스트, 영수증, 스크린샷이(가) 표시된 사진

자동 생성된 설명

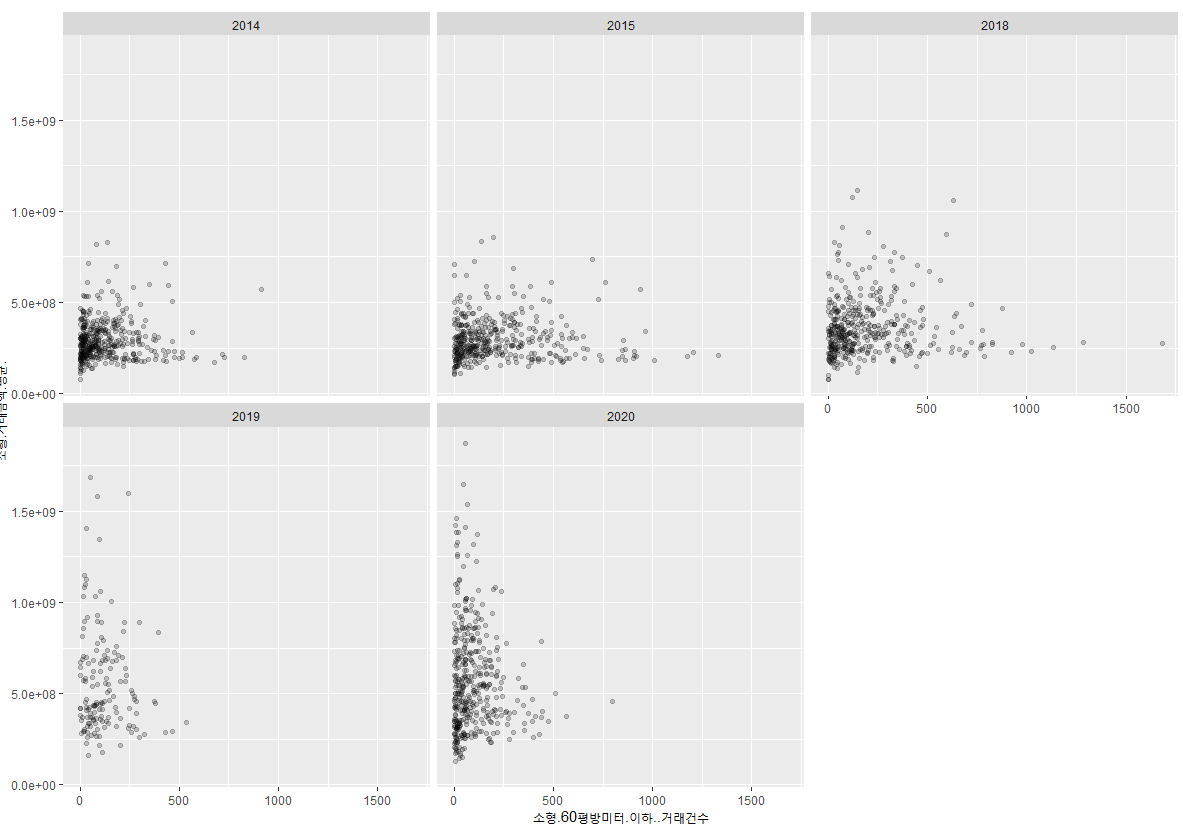
1. **데이터 탐색과 분석**

효과적인 데이터 탐색과 분석을 위해서 plot()함수와 ggplot2를 이용해 데이터를 시각화하여 탐색을 수행한다. 먼저 각 평방미터별 어느 금액의 아파트들이 거래가 활발하게 일어나고 있는지 파악하기 위해, 각 평방미터들의 거래건수와 거래금액.평균.의 상관관계를 거래년도 별로 구분해 시각화하여 분석한다.

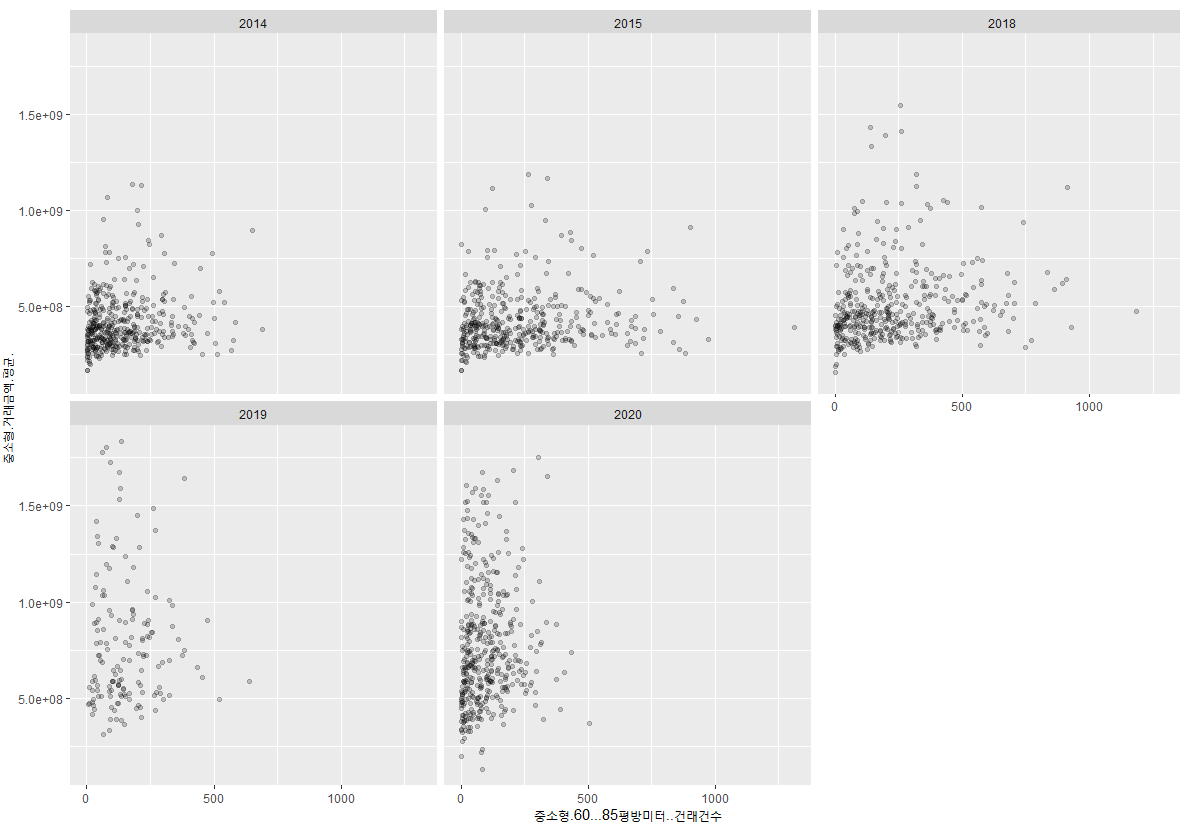
필자의 예상대로 대형아파트의 거래건수가 가장 적은 것을 시각화를 통해 나타났다. 하지만, 소형아파트의 거래건수가 가장 많을 것이라고 예상한 필자의 생각과는 달리 중소형아파트의 거래건수가 훨씬 스펙트럼이 넓고, 거래가 활발하게 일어났다는 것을 시각화를 통해 확인할 수 있다.

또한 거래건수가 2014년부터 증가하다가, 2018년에 최대를 기록하였고, 2018년을 기점으로 거래건수가 감소하기 시작했으며, 2019년도에 거래건수가 위축된 것을 확인할 수 있다. 이후 2020년도에는 작년인 2019년도보다 거래건수가 소폭 상승했다는 사실 또한 확인할 수 있다.

**소형**



**중소형**

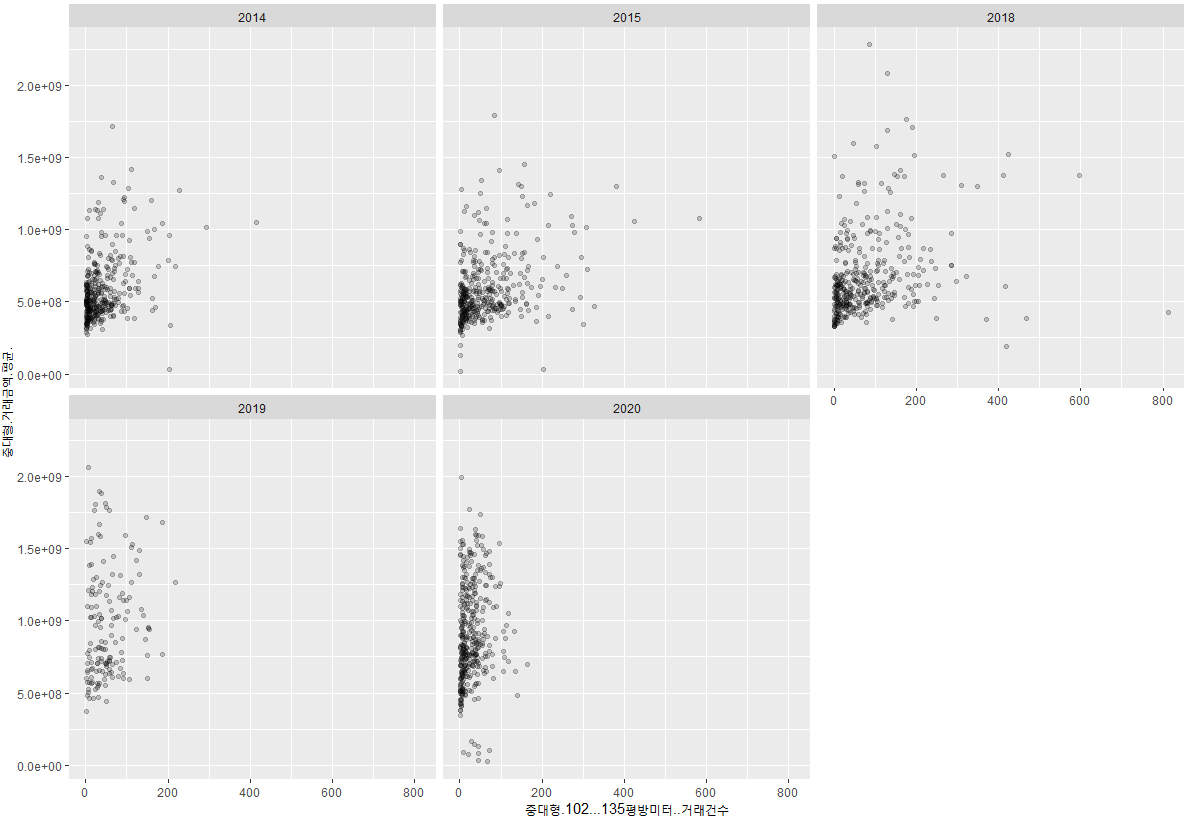


**중형**

텍스트, 하얀색이(가) 표시된 사진

자동 생성된 설명

**중대형**



**대형**

텍스트, 하얀색이(가) 표시된 사진

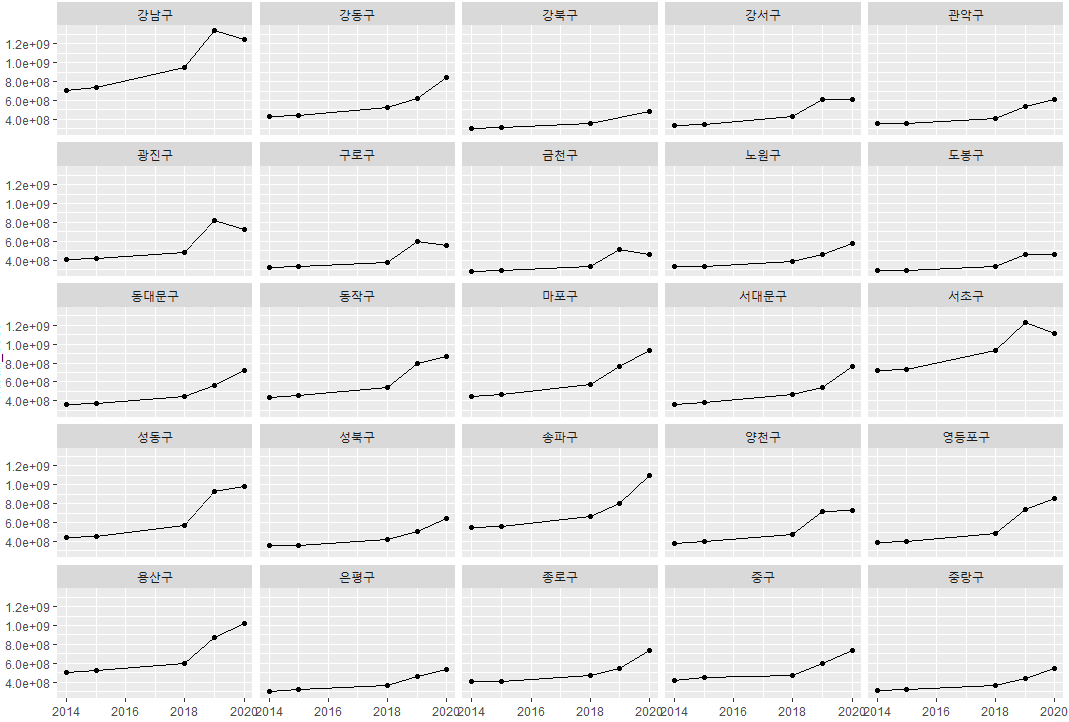
자동 생성된 설명

이제 각 자치구별, 연도에 따른 거래금액.평균.의 증감을 시각화를 통해 나타낸다.

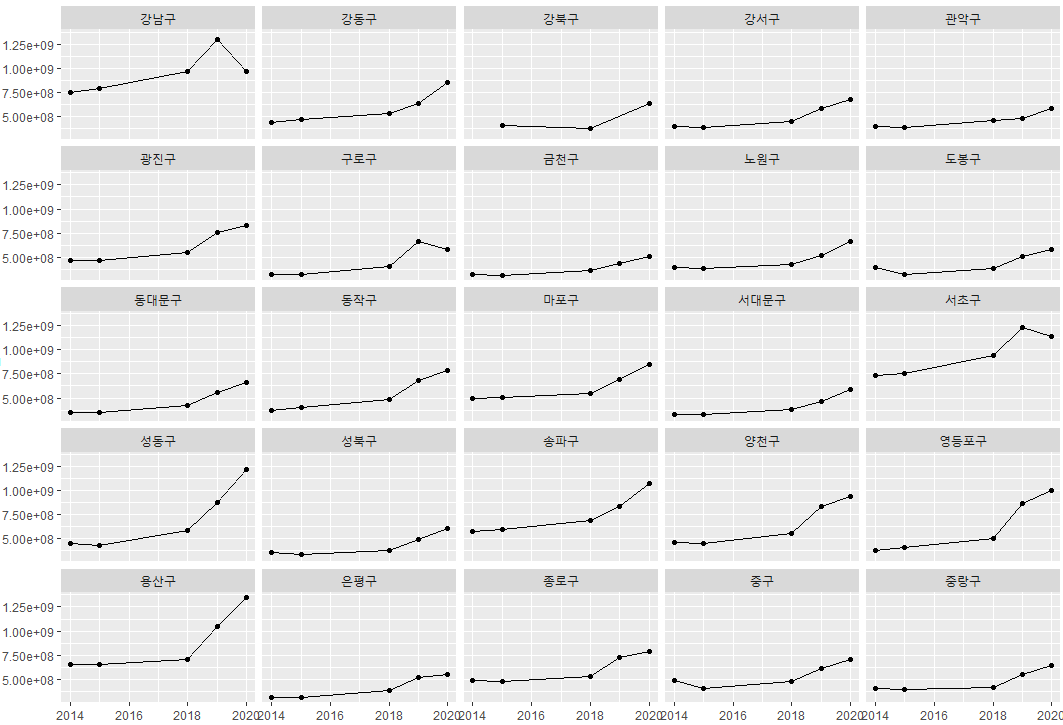
**소형**

텍스트, 지도, 실내이(가) 표시된 사진

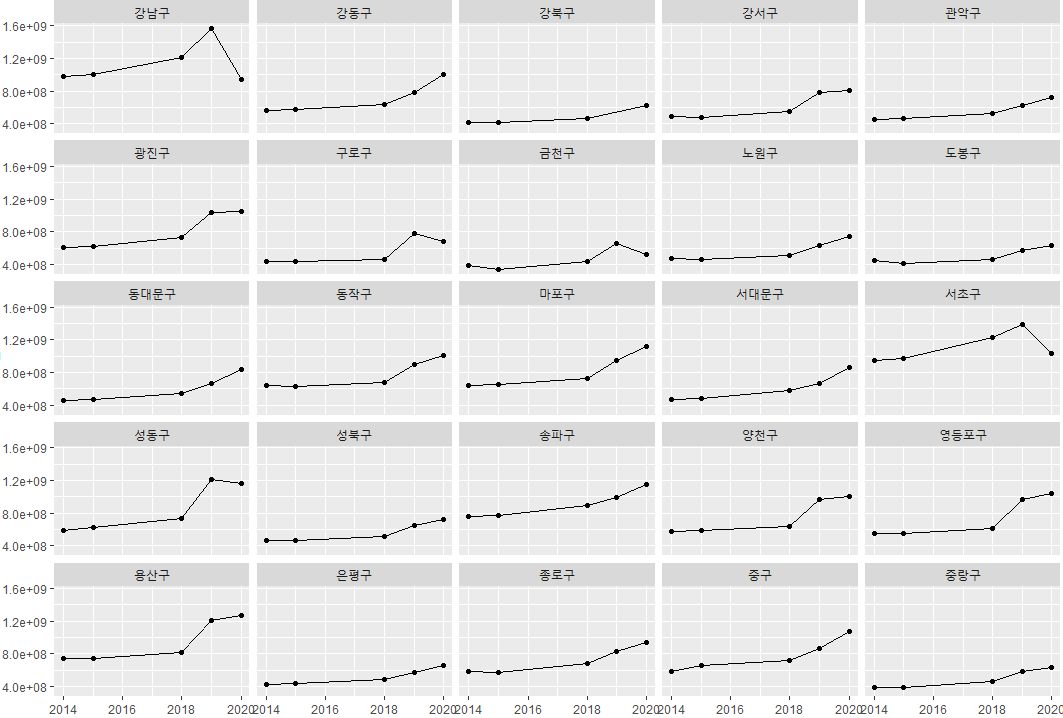
자동 생성된 설명

**중소형**  


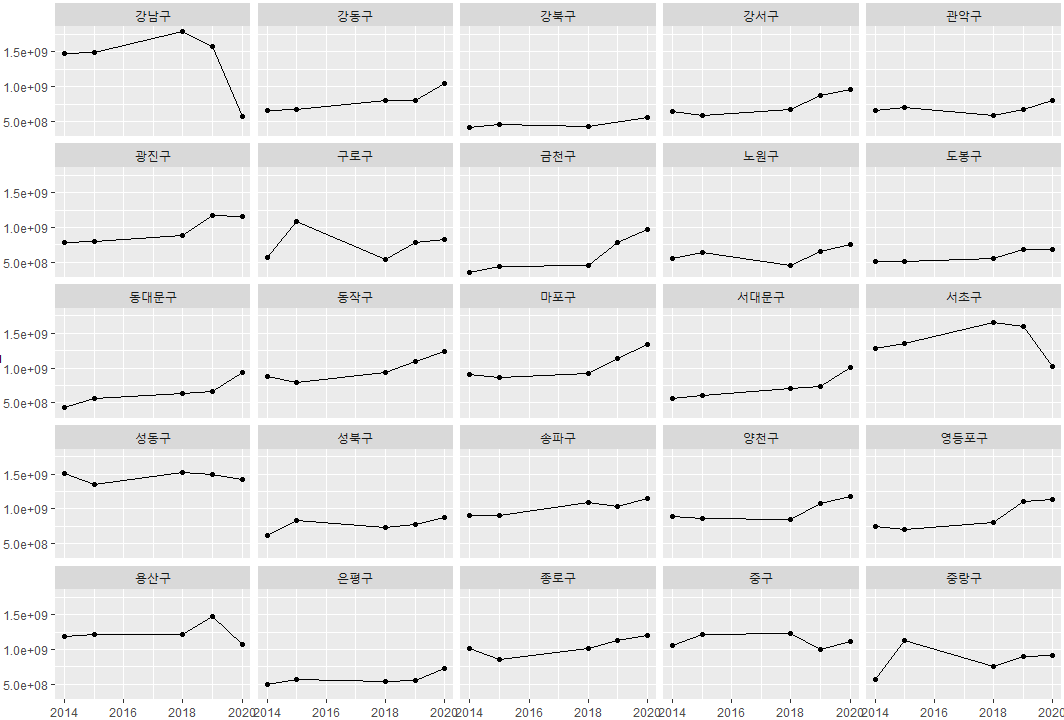
**중형**



**중대형**



**대형**



소형에서는 성동구가 최대 상승폭을 보여주었다. 두 자치구는 거래금액.평균이 2014년 이후 약 2배 상승했다.

중소형에서는 다른 평수들에 비해 전체적으로 모든 자치구가 평균적으로 약 2배의 큰 상승폭을 나타냈다. 이들 중, 소형과 동일하게 성동구 최대 상승폭을 보여주었으며, 2014년 이후 거래금액.평균이 약 2배 이상 상승했다

중형에서는 용산구의 상승폭이 다른 자치구에 비해 압도적인 상승폭을 보여주었다. 특히 성동구는 거래금액.평균이 2014년 이후 최소 3배이상 상승했는데, 이는 모든 평수들의 상승폭 중 최대 상승폭이다.

중대형에서는 중형과 마찬가지로 용산구의 상승폭이 2014년이후 거래금액.평균이 약 2배 상승으로 중대형 중 최대 상승폭이다.

대형에서는 그래프가 이전 그래프들과는 달리 기형적인 모습을 보인다. 파일을 다시 확인해본 결과 자치구.명이 강남구와 서초구인 행에서 거래금액.평균이 이상값이라고 판단되는 데이터들이 있었으나, 서울시가 제공한 실제로 거래가 체결된 금액들의 평균 데이터라 따로 판단할 수 없는 상황이다. 대형에서는 동대문구의 거래금액.평균이 2014년 이후, 2배 상승으로 대형 중 최대 상승폭으로 나타났다.

1. **예측모델**

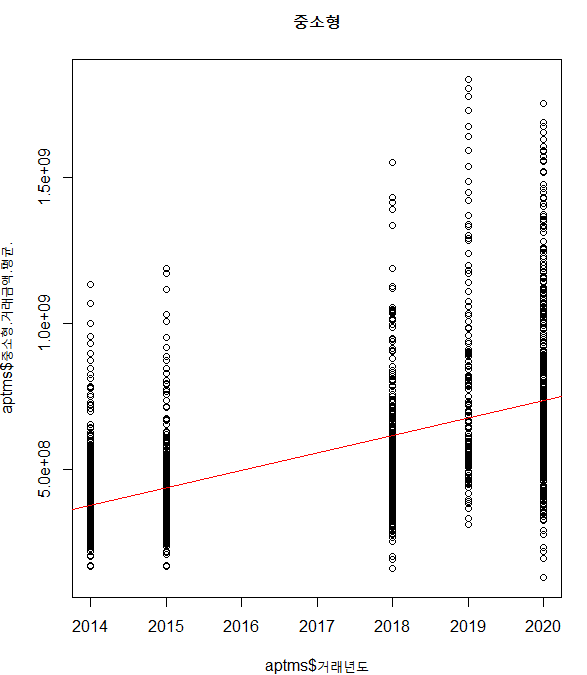
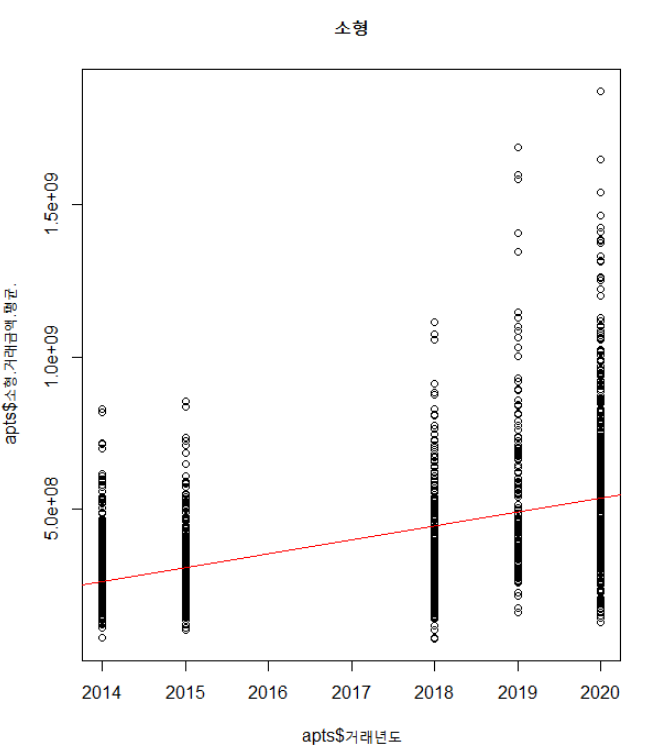
연도에 따른 거래금액.평균의 상승폭을 예측할 수 있는 모델이 필요하므로, 분류모델이 아닌 회귀모델을 만들어야 한다. 따라서 결정 트리나 랜덤 포리스트 모델을 만드는데 사용하는 rpart()나 randomForest()가 아닌 선형회귀함수인 lm()을 이용하여 회귀모델을 만들고 예측에 활용한다. 모든 평수의 아파트들의 거래금액.평균을 예측하기 위해서 apts, aptms, aptm, aptmb, aptb에 모두 lm()함수를 적용하여 apts\_model, aptms\_model, aptm\_model, aptmb\_model, aptb\_model로 모델링한다. 5개의 모델에 대해서 유의미한 모델링인지를 판단하기 위해, summary()함수를 사용한다. apts\_model, aptms\_model, aptm\_model, aptmb\_model 데이터 모두 모델링한 결과, (Intercept)절편과 거래년도 계수 p-값이 2e-16으로 0.05보다 매우 작은 값으로 나타났다. 따라서 이 4개의 모델 전부 통계적으로 유의미한 모델임을 알 수 있으며, 정확하게 에측하는 모델이라 평가할 수 있다.

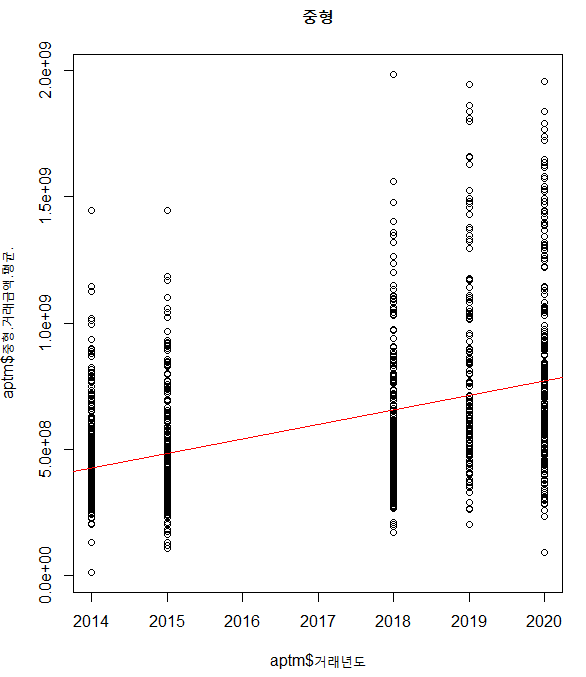
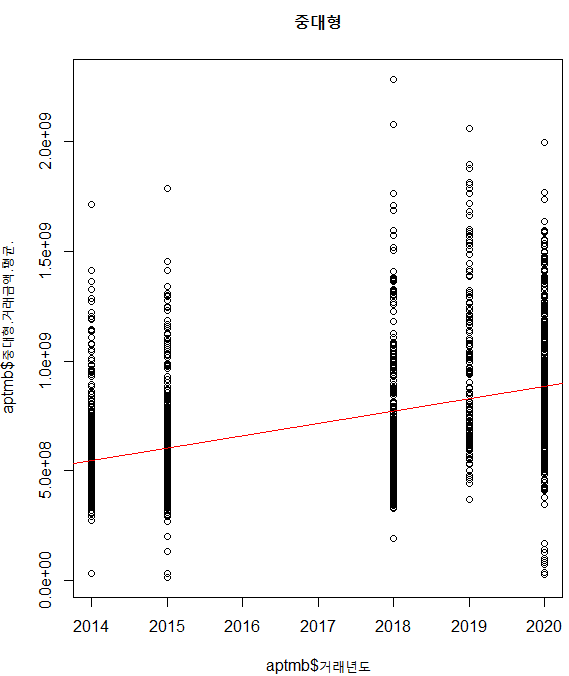
하지만, aptb\_model 데이터의 경우 (Intercept)절편의 p-값이 0.00011이고, 거래년도의 p-값이 8.2e-05로 작은 값이 나왔으나, 다른 모델들에 비해 상대적으로 덜 정확한 모델로 예측할 것이라는 결과가 도출되었다.

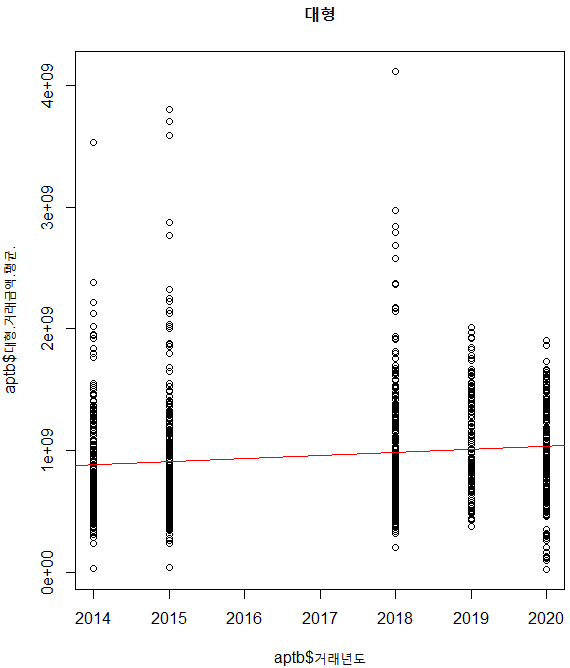
1. **예측모델 시각화**

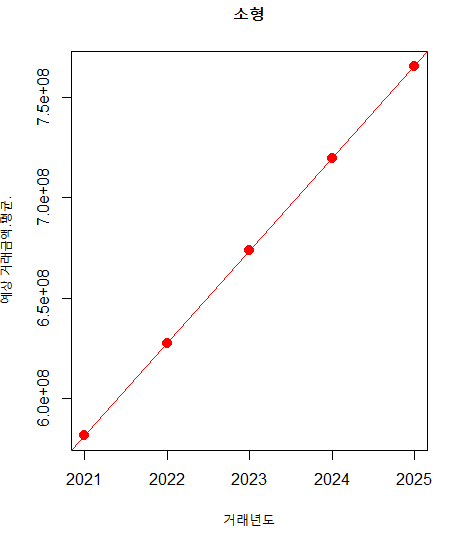
만들어진 예측모델을 기존의 데이터에 적용한 것을 시각화하고, 앞으로 2021년도부터 2025년도까지의 거래금액.평균의 변화를 예측모델을 통해 시각화한다.

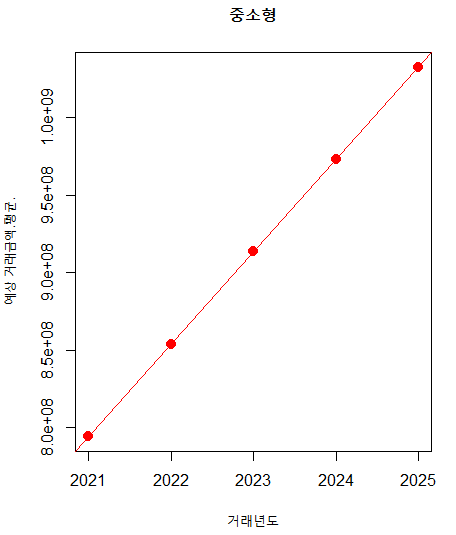
* 1. **: 기존 데이터에 예측모델 시각화**

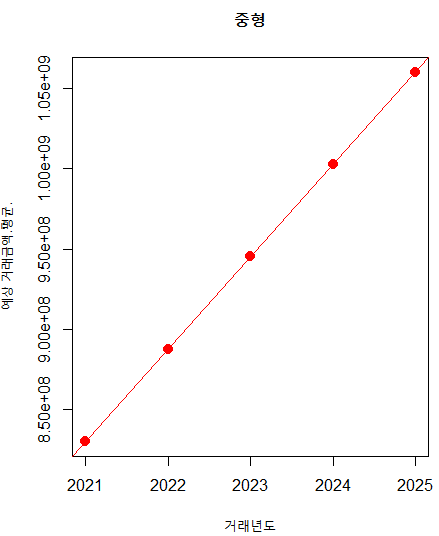
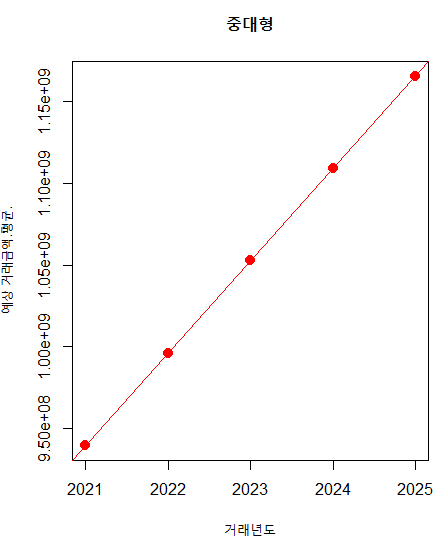


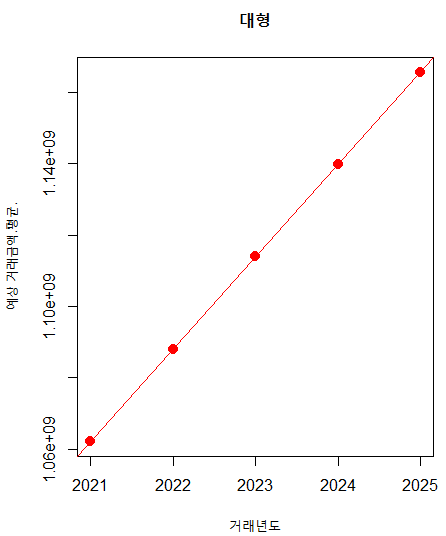




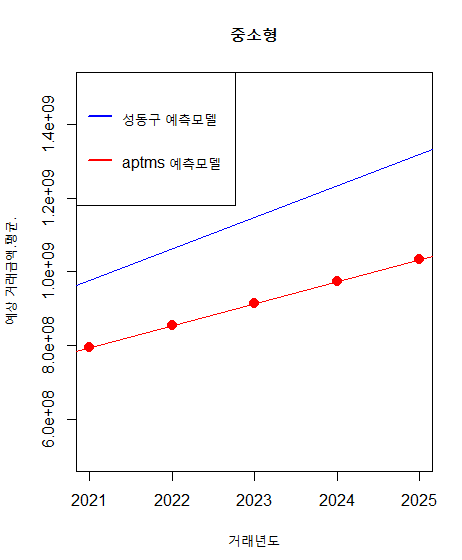
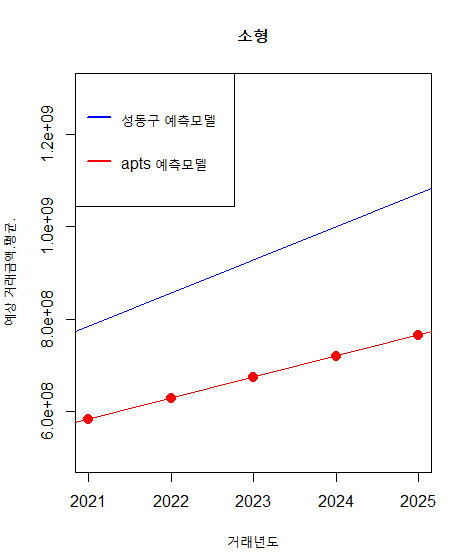
**5.2 : 2021년부터 2025년까지의 거래금액.평균 예측 결과와 학습된 모델 시각화**

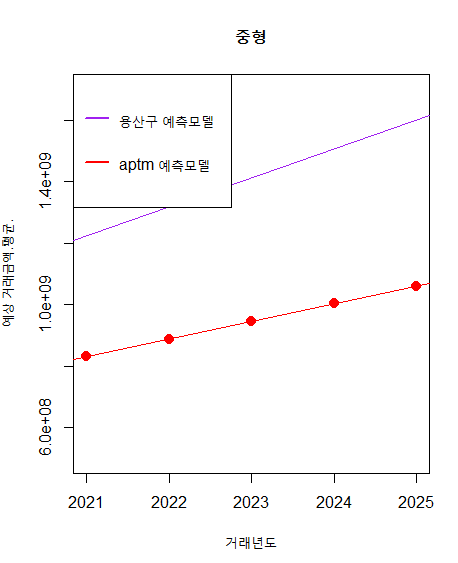
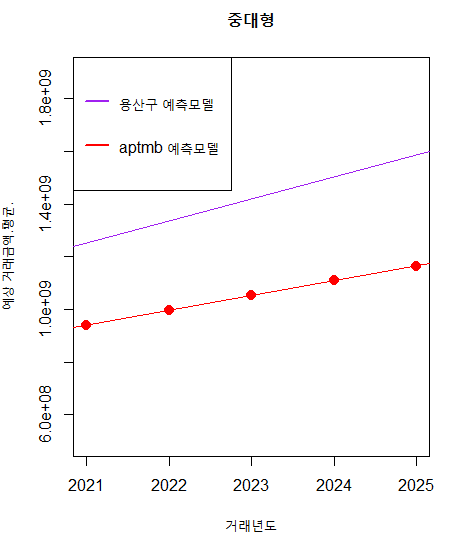


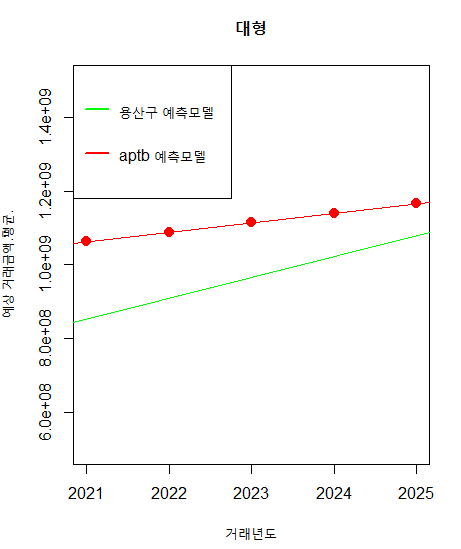




* 1. **: 최대 상승폭을 기록한 자치구의 예측모델 시각화 후 비교**







1. **예측 결론 도출**

데이터 ‘서울시 동별 아파트 매매거래 현황’을 기준으로 조사해본 결과, 전반적으로 모든 자치구의 아파트 거래금액의 평균이 증가한다는 예측 결과가 나타났으며, 소형아파트를 제외한 모든 평수의 아파트의 평균 거래금액이 2025년이후에는 약 10억에 육박하거나 그 이상이 결과가 나타났다. 각 아파트의 평수별로 수익률을 예측해본 결과, 서울의 아파트를 2021년도에 매입한 다음, 2025년도에 매각한다고 가정했을 때, 소형 아파트는 평균 약2.2억의 이익을, 중소형 아파트는 약 2.4억의 이익을, 중형 아파트는 약 3.3억의 이익을, 중대형 아파트는 약 2억의 이익을, 마지막으로 대형 아파트는 약 1억의 이익이 예상되었다. 또한 거래금액의 평균 상승폭이 가장 큰 지역은 용산구로 측정이 되었으며, 이를 통해 투자의 비용 효율적인 측면에서 분석했을 때, 가장 이윤이 많이 남을 것이라 최종 예측되는 최적의 아파트 투자 매물은 용산구의 중형아파트이다. 용산구의 중형아파트를 2021년도에 매입했다고 가정했을 때, 2025년도에 매각한다면, 평균 4억원 정도의 이윤이 창출된다는 결론이 도출되었다.

결론적으로 부동산 투자를 계획 중인 투자자라면, 최대한 빠른 시일 내에, 서울의 어느 아파트를 매입한다 해도 수익을 창출할 수 있으며, 그중 최적의 효율은 용산구 중형 아파트이다. 또한 2025년 이후에 아파트 매입을 계획 중이라면, 투자자는 소형 아파트에 투자할 때에도 10억에 가까운 자금이 필요하게 될 것이다.