**서울 구별 아파트 값 요인 분석**

17101967 이동준

19101206 김진호

16101388 이병창

목차

[데이터 분석 목적](#_Toc90836790)

[데이터 분석 설계](#_Toc90836791)

[데이터 분석](#_Toc90836792)

[데이터 선정](#_Toc90836793)

[데이터 전처리](#_Toc90836794)

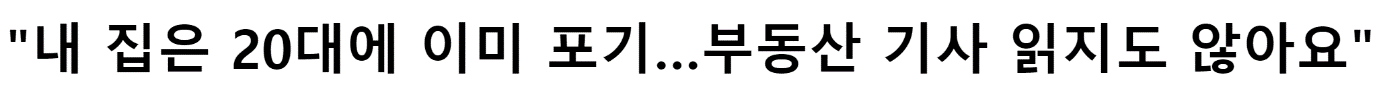
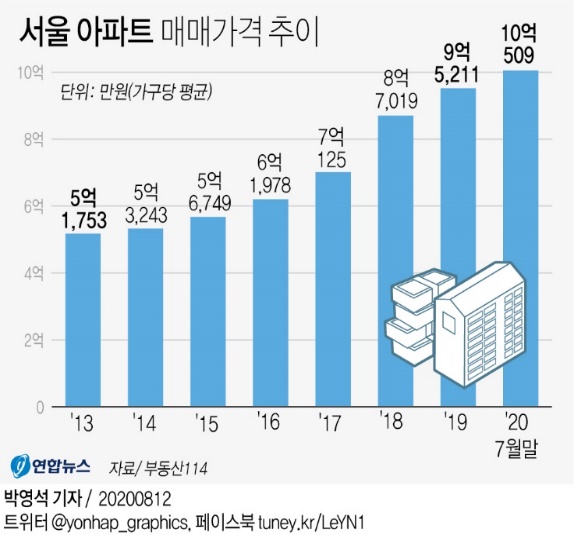
[K-means clustering](#_Toc90836795)

[Decision-Tree Regression & Radom Forest Regression](#_Toc90836796)

[데이터 분석 결과](#_Toc90836797)

[코드 첨부](#_Toc90836798)

**데이터 분석 목적**



위 아파트 매매가격의 가파른 증가폭과 다소 비관적인 기사 제목에서 보듯이 요즘 20-30대는 집을 구하는 데 있어서 상당히 어려움을 겪고 있습니다. 따라서 저희 팀은 어떤 원인이 가파른 아파트 매매가격 증가에 영향을 주는지 알고 싶어 졌습니다.

**데이터 분석 설계**

**<데이터 얻는 방식>**

서울 열린데이터 광장

통계청 국가통계포털

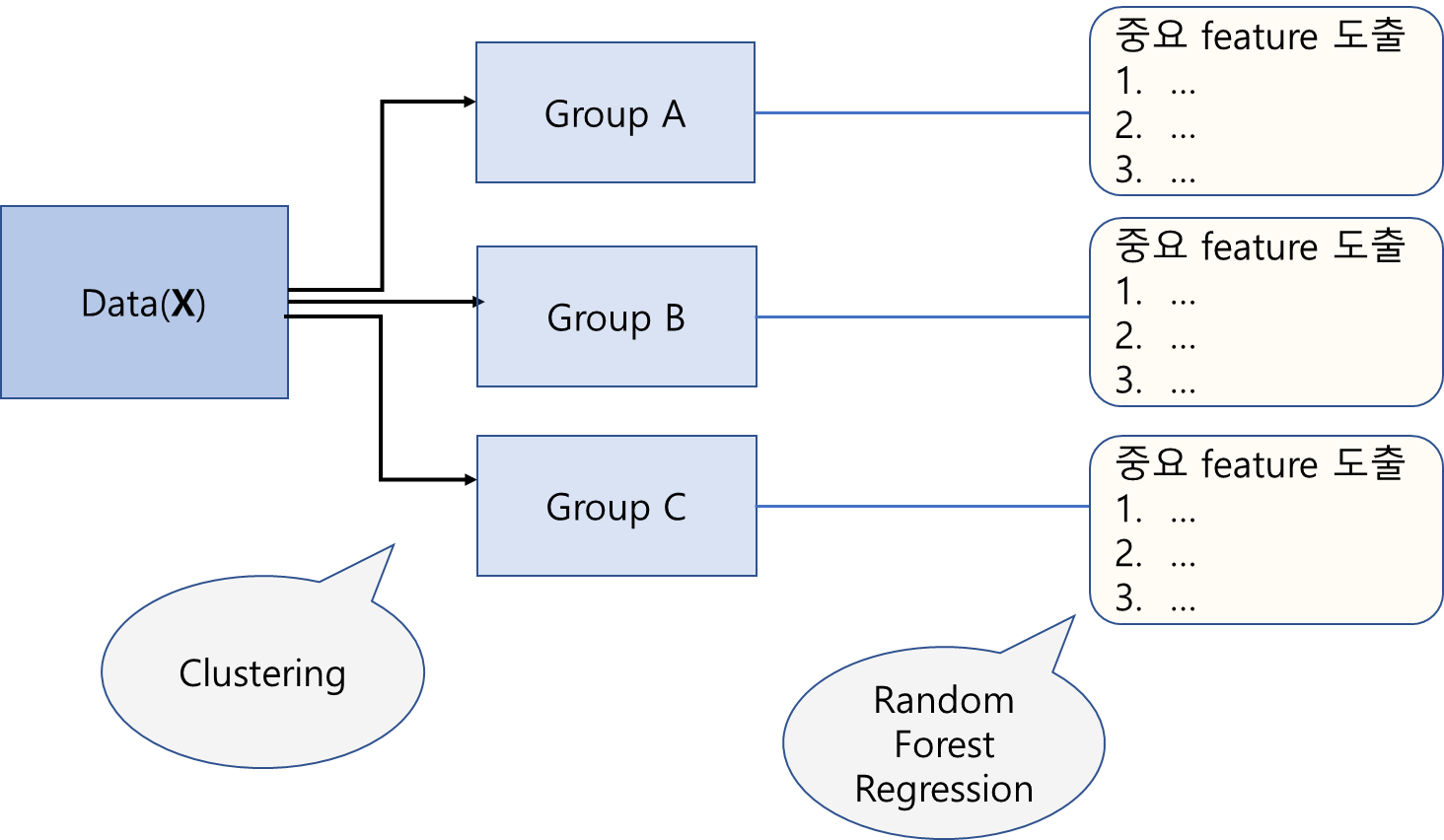
지방 인허가데이터

서울부동산정보광장



저희는 아파트 매매가격에 대한 정보를 서울부동산 정보광장 사이트에서 매매가격에 영향을 줄만한 요소인 X값들을 국가통계포털, 서울 열린 데이터 광장, 지방 인허가 데이터 개방에서 찾아보기로 정했습니다.

**<알고리즘 설계>**



저희는 위와 같이 대략적으로 데이터 분석 과정을 설계하였습니다. 먼저 매매가격에 영향을 줄 만한 요소를 각자 준비해오고 이를 통합해서 통합 데이터에 해당하는 X로 만듭니다. 이 다음에 이 데이터를 Clustering 기법을 통해 k개의 그룹화를 합니다. 그 후 각 그룹에 대해서 Random Forest Regression을 돌려서 각 그룹에 영향을 주는 중요한 Feature의 순서들을 각각 얻어보는 것으로 설계하였습니다.

위와 같이 설계한 이유를 설명 드리겠습니다.

첫번째로 Clustering을 하는 이유는 25개의 구가 다른 특징들을 지닐 것으로 예상이 되었고 그에 따라서 중요한 feature값들이 다르게 나올 것 이라고 직관적으로 생각했기 때문이고 Clustering의 결과로서 각 그룹이 나누어지면 서울 구가 어떤 특징을 바탕으로 나누어지는 알 수 있을 것이라고 생각했기 때문입니다.

두 번째로 Random Forest Regression을 하는 이유는 저희의 최종 목적인 아파트 매매 가격에 가장 영향을 많이 미치는 요소가 무엇인지 분석할 수 있기 때문입니다.

**<기대 효과>**

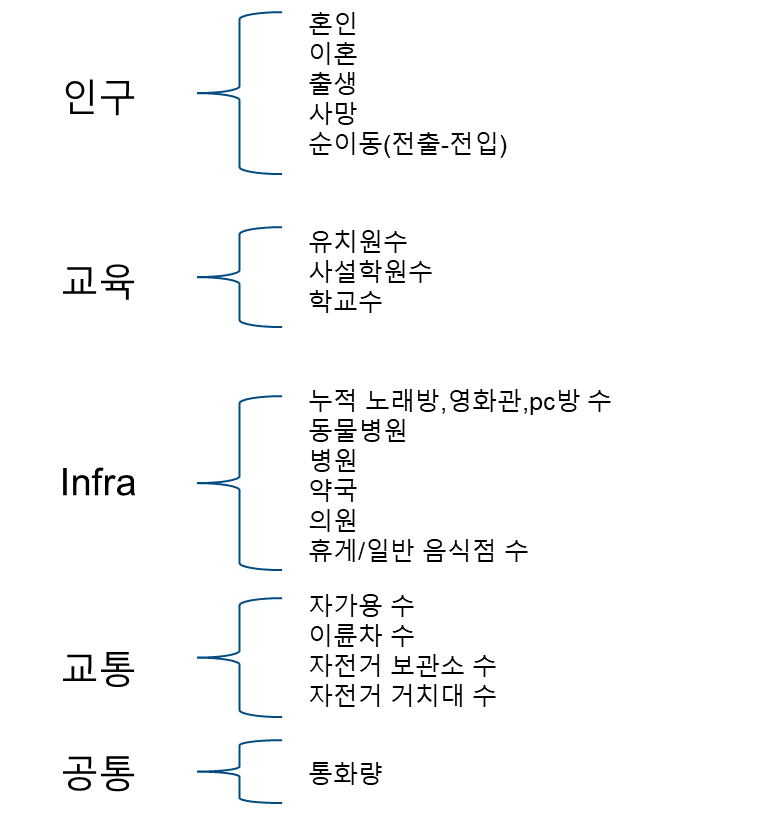
1. Clustering 결과로서 서울 25개 구를 나누어보면서 각 그룹화된 구들의 특징을 알 수 있을 거라 기대했습니다.

2. Clustering 후에 그룹에 따라 Regression을 진행하기 때문에 각 그룹에 공통적으로 나타나는 Feature가 있을 것이고 차이가 나는 Feature가 있을 텐데 이를 바탕으로 유의미한 Insight를 도출할 것이라 기대했습니다.

**데이터 분석**

**<데이터 가져오기>**

먼저 아파트 매매가격에 영향을 주는 요소가 무엇일지 팀원 각자 생각하고 찾아보기로 했습니다.

그 모든 요소를 정리하면 다음과 같습니다.

이 모든 요소를 2014년부터 2020년까지 각 구별(25개)로 모두 구해서 총 2100개의 데이터와 22개의 feature로 구성된 통합 데이터 X를 만들었습니다.

텍스트, 컴퓨터, 실내, 하얀색이(가) 표시된 사진

자동 생성된 설명

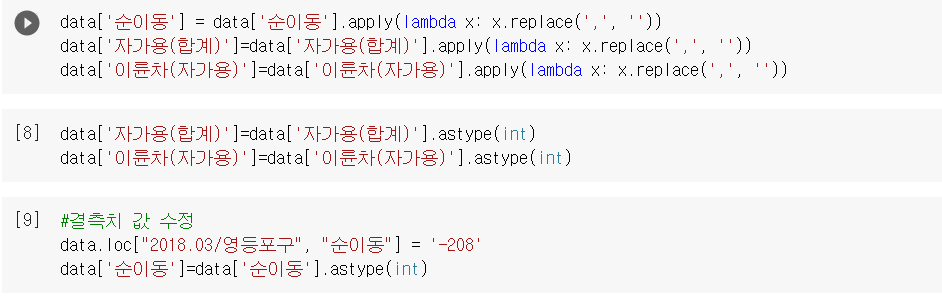
**<데이터 전처리>**

먼저 데이터가 어떻게 구성되어 있는지 살펴보았습니다

테이블이(가) 표시된 사진

자동 생성된 설명

‘순이동’ , ‘자가용(합계)’ , ‘이륜차(자가용)’의 데이터 형식이 object라서 clustering을 하기 위해서 모두 거리 연산이 가능한 int값 또는 float값으로 바꾸어야 함을 알 수 있습니다



먼저 3가지 column에서 ,(comma) 값을 제거하고 그 다음 각 data type을 int로 변경하였습니다.

이 과정에서 ‘순이동’ column 에서 결측치가 발견되어서 수정해주고 역시 int로 변경하였습니다.

테이블이(가) 표시된 사진

자동 생성된 설명

X의 feature별로 그 값의 편차가 크고 Clustering은 거리를 계산하기 때문에 정규화가 필수적이라 판단하고 정규화를 모든 feature별로 시켜주었습니다.

**<K-means Clustering\_1 >**

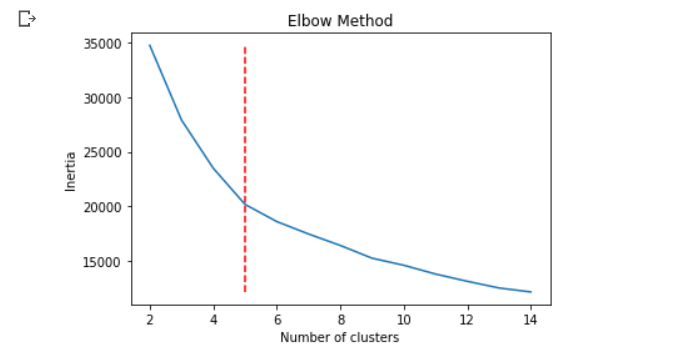
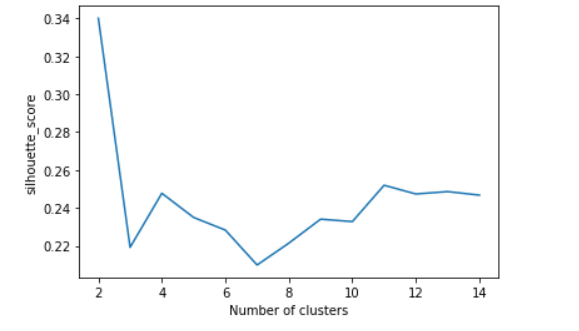
위 정규화된 데이터를 바탕으로 Clustering을 진행해야 하는데 몇 개의 cluster로 하는 것이 적절 한지 판단하기 위해 k에 따른 Inertia 값과 k에 따른 silhouette score 를 각각 구해주었고 silhouette diagram 또한 그려주었습니다.

**텍스트이(가) 표시된 사진

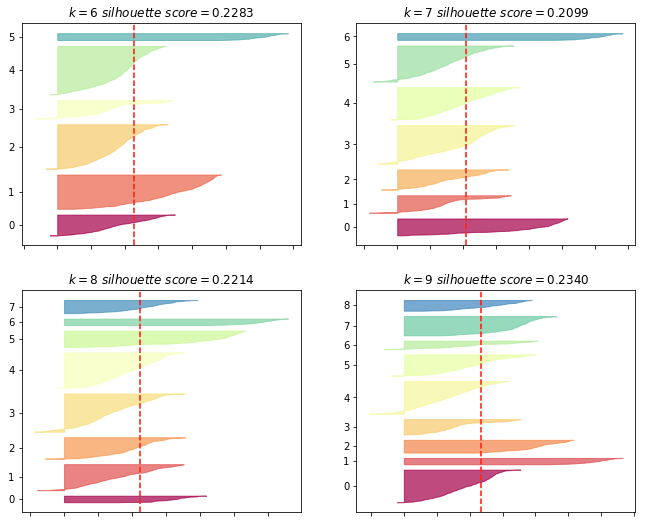
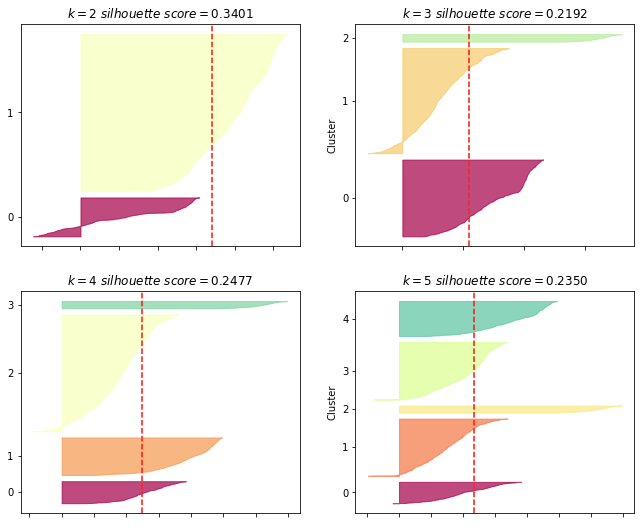
자동 생성된 설명**

먼저 위 코드 내용대로 k의 범위는 2에서 14까지로 잡고 random\_state=300으로 지정함으로써 일관된 결과를 나타내기 위해 초기 중심 위치를 정해주었습니다. 위 코드에 따라서 inertia\_arr에는 k값에 따른 inertia값들이 저장되고 마찬가지로 silhouette\_score\_arr에는 k에 따른 실루엣 스코어 값이 저장되게 하였습니다. (위 코드에서 kmeans\_per\_k는 각 k에따라 학습된 kmeans 모델을 원소로 받는 배열이고 이 배열은 후에 분석에서 쓰입니다)

이 배열들을 바탕으로 각 k에 따른 inertia와 silhouette\_score를 matplotlib를 이용해 그래프로 표현한 결과는 다음과 같습니다.

위 k에 따른 각각의 최적의 지점을 찾아보면 Elbow Method의 경우 꺾이는 지점인 k=5에서 실루엣 스코어의 경우 잠시 반등하는 부분인 k=4에서 최적의 지점임을 알 수 있었습니다.



silhouette diagram의 경우에도 k=4 또는 k=5일 때가 나머지 diagram들에 비해 좋아보임을 알 수 있었습니다.

**<K-means clustering\_2>**

**\* Elbow method 와 silhouette score를 바탕으로 각각 최적의 k인 k=4, k=5로 k-means clustering**

**K=4**

**텍스트이(가) 표시된 사진

자동 생성된 설명**

**K=5**

**텍스트이(가) 표시된 사진

자동 생성된 설명**

**결과**

클러스터링 두 결과에 공통적으로 볼 수 있는 구의 그룹이 있는데 그 그룹은 다음과 같다

**A. 강남구**

**B. 강서구, 노원구, 송파구**

**C. 금천구, 서대문구, 용산구, 종로구, 중구**

**D. 강동구, 강북구, … ,영등포구 (나머지 16개 구)**

또한 두 결과에서 차이점으로 볼 수 있는 점은 k=5 클러스터링의 경우 D 그룹이 2017년도를 기준으로 양분되면서 대략 2018년 이전의 D그룹과 2018년 이후의 D그룹으로 나누어 볼 수 있다는 점이다.

따라서 다음과 같은 추가 분석 지점을 설정하고 분석하였습니다.

**1. A, B, C 그룹이 확실하게 분리되는 이유는 무엇일까?**

**2. K=5일 때, 2018년 기준으로 D그룹이 분리되는 이유는 무엇일까?**

**<K-means clustering 추가 분석>**

**★ A, B, C 그룹이 확실하게 분리되는 이유**

**※** column의 순서는 [혼인, 이혼, 출생, 사망, 통화량, 유치원 , 사설학원 (인구 1000명당), 순 이동, 자가용(합계), 이륜차(자가용), 누적 보관소, 누적 거치대, 누적 노래방 ,누적 영화관 , 누적 pc방 ,동물병원, 병원, 약국, 의원, 휴게음식점, 일반음식점, 학교] 입니다

**A 그룹: 강남구**

강남구만이 독특하게 나누어지는 이유를 알기 위해서 A그룹의 중심을 찍어보고 이 중심값들과 다른 그룹(B,C,D)와의 중심값의 차이를 바탕으로 어떤 column에서 중심값이 가장 차이가 나는지 알아 보았습니다. (k=4일 때의 중심값으로 설정하였습니다)

텍스트이(가) 표시된 사진

자동 생성된 설명

먼저 k=4일 때 A그룹과 B그룹의 중심값 차이입니다. 그 차이값이 확연하게 다른 Column으로는

사설학원 수(인구 1000명당) , 영화관, 병원 , 약국 , 의원 , 휴게음식점, 일반음식점 이 있었습니다.

텍스트이(가) 표시된 사진

자동 생성된 설명

다음은 A그룹과 C그룹 사이의 중심값 차이입니다. 추가적으로 차이가 난 Column은 자가용(합계), 동물병원 수, 학교 수 였습니다.

테이블이(가) 표시된 사진

자동 생성된 설명

A그룹과 D그룹 사이의 차이입니다. 마찬가지로 차이가 난 Column은 A그룹과 C그룹 사이의 차이가 나는 Column과 동일했습니다.

**결론**

강남구만이 다른 그룹으로 분리된 공통적인 이유는 사설학원 수, 영화관, 병원, 약국, 의원, 휴게음식점, 일반음식점이 있었습니다. 추가적으로 차이가 난다고 볼 수 있는 column은 자가용(합계), 동물병원 수, 학교 수 로 볼 수 있었습니다. 또한 이 차이 값이 모두 양수 인 것으로 보아 다른 구들과 비교해서 강남구가 위 column들의 값이 훨씬 큼을 알 수 있습니다.

**B 그룹: 강서구, 노원구, 송파구**

\* A그룹과 비교하는 것은 A그룹 분석과 동일하므로 제외하였습니다.

테이블이(가) 표시된 사진

자동 생성된 설명

B그룹과 C 그룹 사이의 차이입니다. 출생,사망, 혼인, 이혼 의 값이 크게 차이가 났고 그 외로 유치원 수, 자가용(합계) , 노래방 , 학교 에서 큰 차이를 보였습니다.

텍스트이(가) 표시된 사진

자동 생성된 설명

B그룹과 D그룹 사이의 차이입니다. C그룹과의 차이보다는 두드러지게 나타난 부분은 없지만 크게 차이나는 부분을 찾자면 동일하게 유치원, 자가용(합계), 학교 에서 차이를 보였습니다.

**결론**

강서구,노원구,송파구만 따로 그룹으로 묶인 이유의 공통적인 부분은 유치원 수, 자가용 수, 학교 수에서 차이가 났음을 알 수 있습니다. 또한 그 값이 양수인것으로 보아 이 구들이 다른 구들에 비해 column들의 값이 양수임을 알 수 있습니다.

이와 더불어 인구의 측면에 해당하는 혼인,이혼,출생,사망 수가 상대적으로 많은 것을 알 수 있는데 이것을 통해 이 3개의 구가 다른 구들보다 총 인구수가 많은 것을 유추해볼 수 있고 인구 수가 많기에 유치원,학교,자가용 수도 다른 구 들에 비해 많은 것이 아닌지 생각할 수 있었습니다.(실제로 자치구별 인구 수 통계를 봤는데 강서,노원,송파구가 자치구별 1,2,3등에 해당하는 가장 많은 인구 수를 가지고 있었습니다.)

**C 그룹: 금천구, 서대문구, 용산구, 종로구, 중구**

**\*** A그룹, B그룹과의 비교는 위에서 진행했으므로 생략하였습니다.

텍스트이(가) 표시된 사진

자동 생성된 설명

C그룹과 D그룹의 차이입니다. 기본적으로 중심 차이 값이 거의 음수인 것으로 보아 대부분의 column에서 그 값이 다른 구들보다는 작은 것임을 알 수 있습니다. 또한 가장 차이가 나는 부분은 노래방 수였습니다.

**결론**

C 그룹의 경우 다른 그룹과 달리 혼인,이혼,출생,사망수가 상대적으로 다른 구에 비해 적은 것을 알 수 있고 (중심 값 차이가 다 음수) 이를 통해 상대적으로 자치구별 총 인구가 적음을 알 수 있습니다. 따라서 인구와 연관될 다른 변수 값들이 대부분 음수를 가짐을 유추해 볼 수 있습니다.(실제 자치구별 통계에서도 이 구들의 인구수가 25개 전체의 구에서 가장 낮은 편에 속했습니다)

**★ K=5일 때 2018년 기준으로 D그룹이 분리되는 이유**

D그룹 내에서 분리되는 이유를 알기 위해서 K=5일 때의 두 클러스터 사이의 중심값을 구하고 이 차이를 구해보았습니다. (K=1이 2018년 이후 클러스터이고 k=3이 2018년 이전 클러스터입니다.)

텍스트이(가) 표시된 사진

자동 생성된 설명

실제 가장 차이가 나는 column은 통화량, 누적 보관소, 누적 거치대 이었습니다.

먼저 통화량의 경우는 2014년부터 2020년까지 꾸준하게 증가하였기 때문에 시기별 특징이긴 하지만 큰 의미는 없습니다.

누적 보관소, 누적 거치대의 경우 유의미한 차이를 보이는데 2018년 이후에는 그 값이 그 이전보다 상당히 증가하였음을 알 수 있습니다. 이는 실제로 2015년 10월부터 서울시 공용자전거인 따릉이가 설치되기 시작했는데 서울 전체로 확대되기까지 약 2년정도가 걸려서 실질적으로 2018년 이후에는 따릉이가 완전 활성화가 되었습니다.

따라서 2018년을 기준으로 나누게 된 계기는 실질적으로 따릉이의 활성화 입니다.

**<K-means clustering 추가 분석2>**

**\* k=5일 때를 기준으로 분석 , 그룹은 크게 다음과 같이 나누어집니다**

**A. 강서구, 노원구, 송파구**

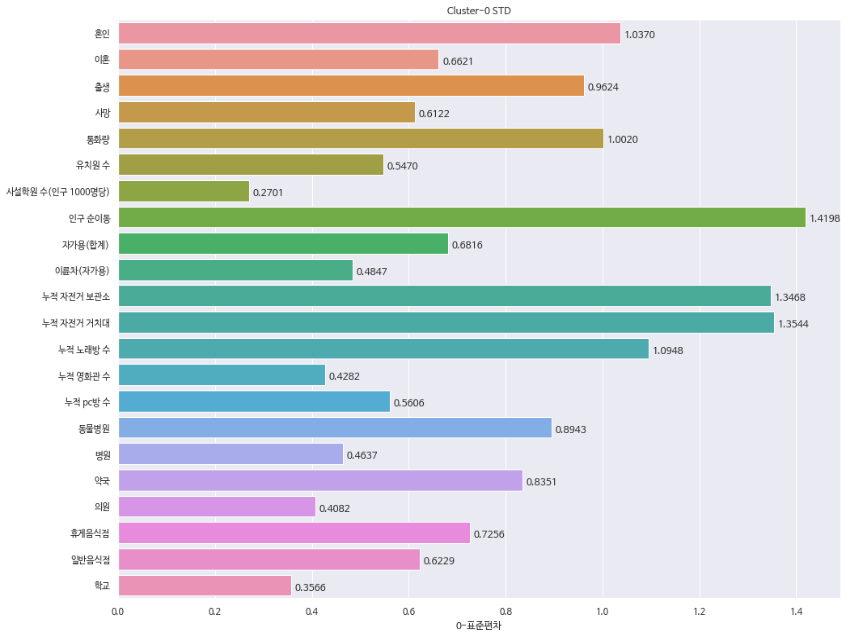
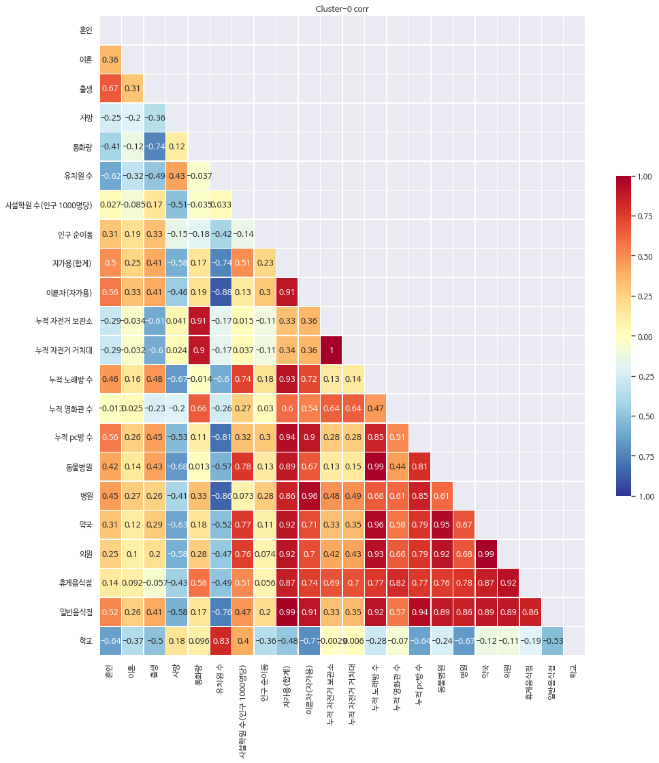
**B. 강동구, 강북구, … ,영등포구 (나머지 16개 구) <약 2018년 이후>**

**C. 강남구**

**D. 강동구, 강북구, … ,영등포구 (나머지 16개 구) <약 2018년 이전>**

**E. 금천구, 서대문구, 용산구, 종로구, 중구**

**A 그룹:** **강서구, 노원구, 송파구**



병원, 약국, 의원이 높은 상관관계를 보입니다. 병원과 의원 근처에 약국이 존재하기 때문으로 보입니다.

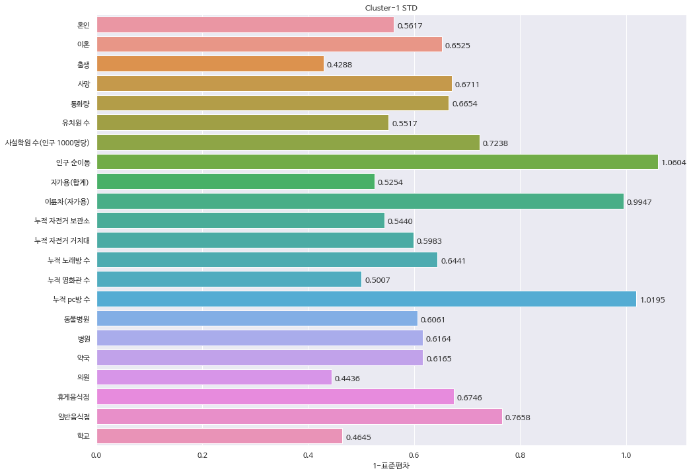
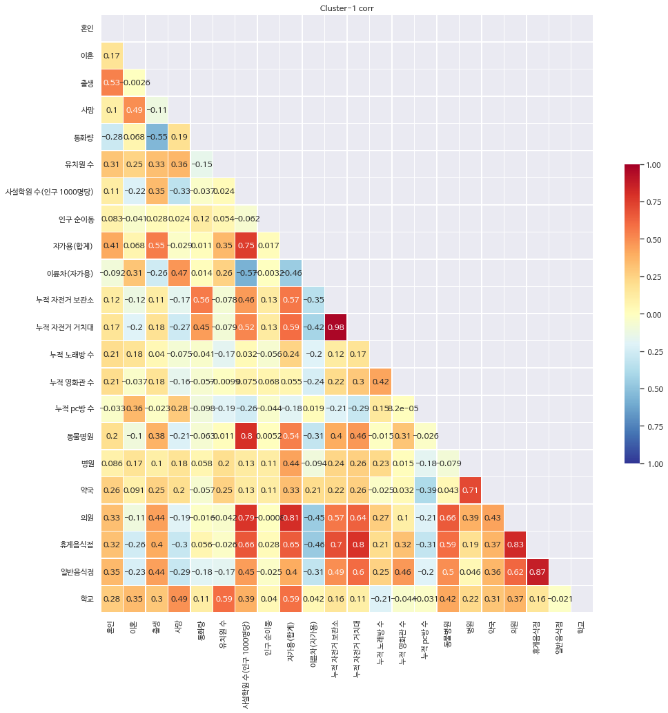
학교와 유치원 수는 다른 feature와 연관성이 없는 것으로 보입니다.

자동차의 수가 일반음식점, 휴게음식점, 의원, 약국, 병원, 동물병원, pc방 수, 노래방 수과 높은 상관관계를 갖는 것을 볼 때 다른 시설을 이용할 때 자동차를 주로 이용한다는 것을 알 수 있습니다.

사설학원 수의 표준편차가 작을 것을 볼 때 클러스터 내의 학원 수는 크게 변하지 않는 것을 알 수 있습니다.

인구 순이동의 표준편차가 큰 것을 볼 때 클러스터 내의 인구 증감폭이 매우 큰 것을 알 수 있습니다.

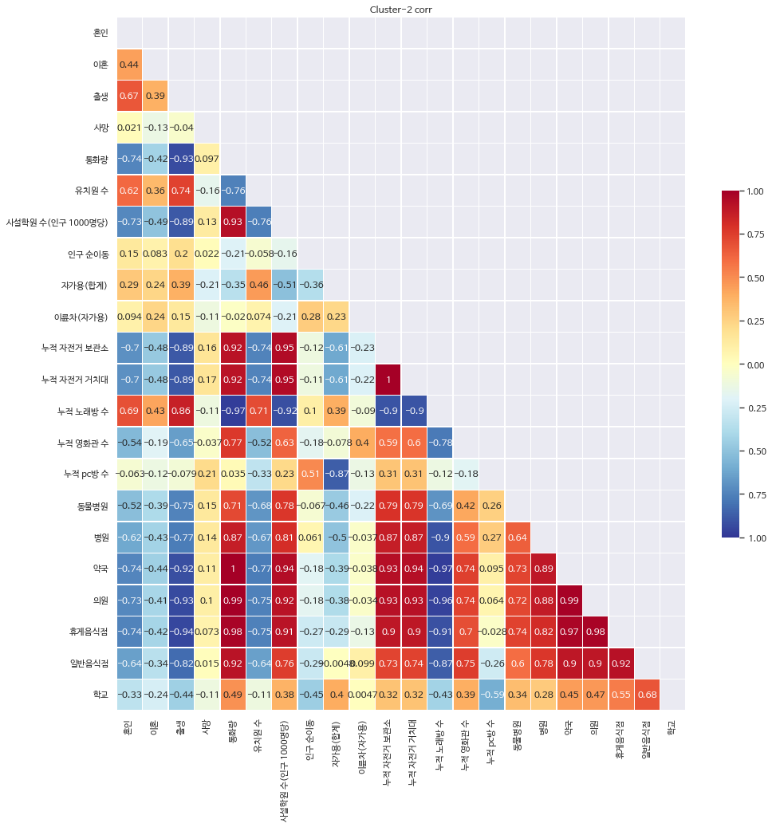
**B 그룹: 강동구, 강북구, … ,영등포구 (나머지 16개 구) <약 2018년 이후>**

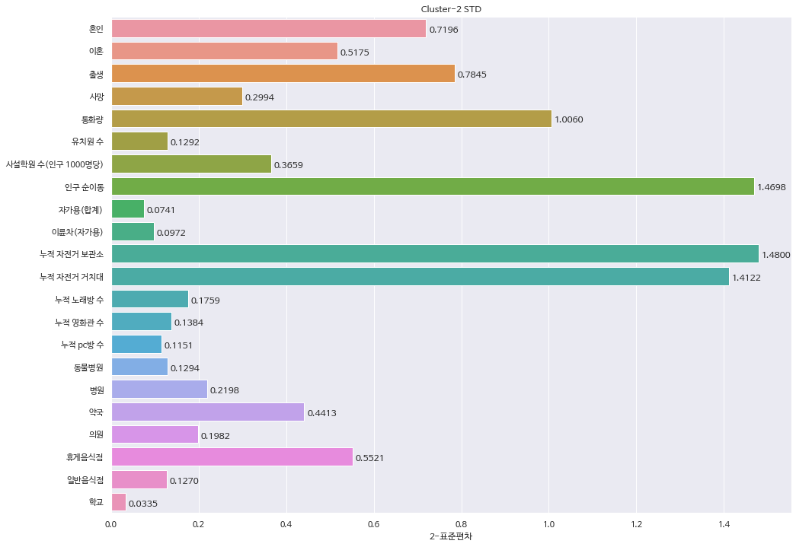


일반음식점과 휴게음식점을 제외하면 각 feature 간의 상관관계가 거의 없는 것으로 보입니다.

feature간 표준편차 역시 인구 순이동, 이륜차, pc방 수를 제외할 경우 비슷한 것을 봤을 때, centroid와 거리는 가깝지만 거리에 영향을 미치는 feature가 각 데이터마다 다르다는 것을 추정할 수 있습니다.

**C 그룹 : 강남구**



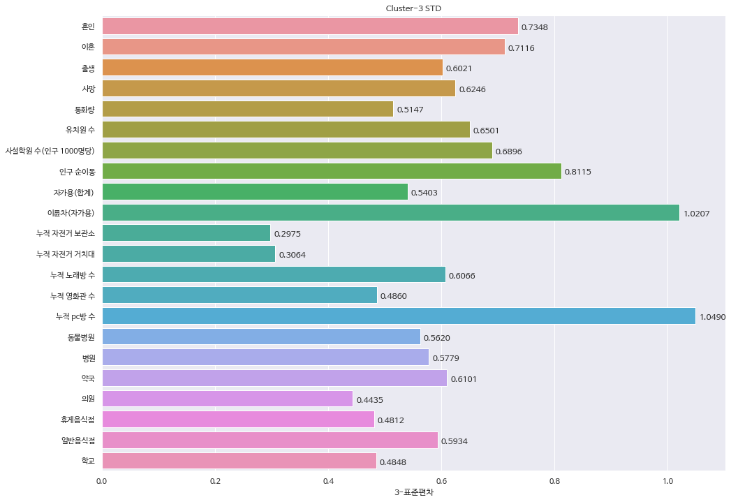
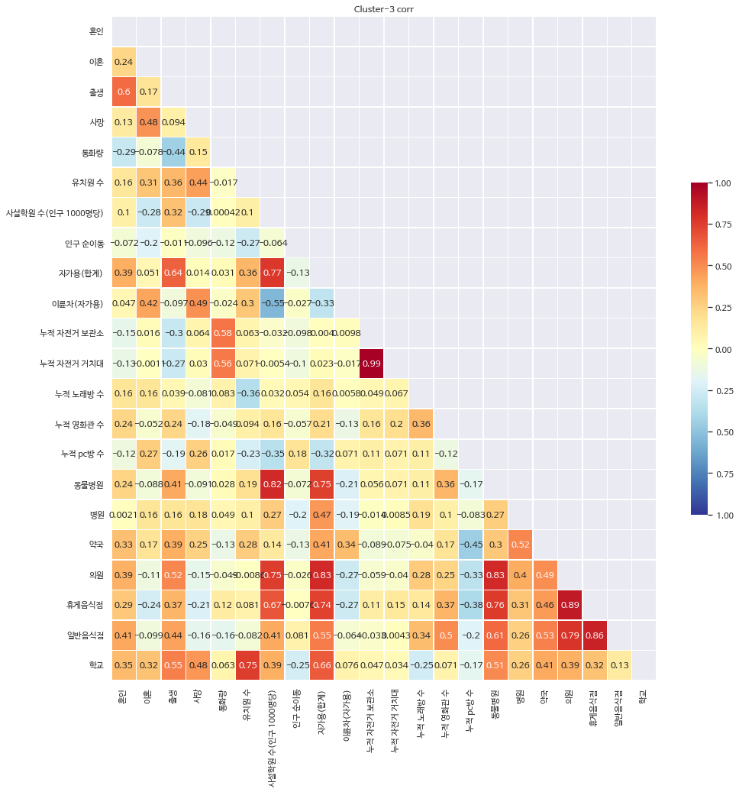


상관관계가 높은 feature과 상관관계가 낮은 feature가 명확하게 분리되고 있습니다.

병원, 음식점 같은 인프라 관련 feature와 사설학원 수, 자전거 거치대 수의 상관관계가 높게 나오고, 나머지 feature들은 상관관계가 거의 없는 것으로 보입니다.

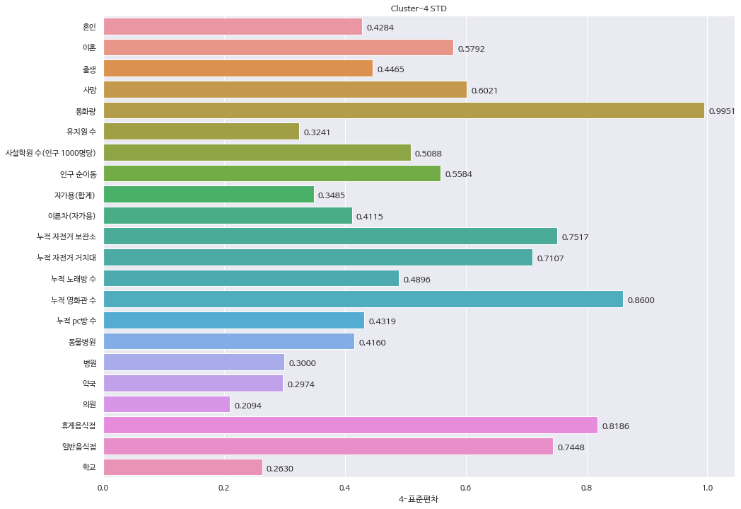
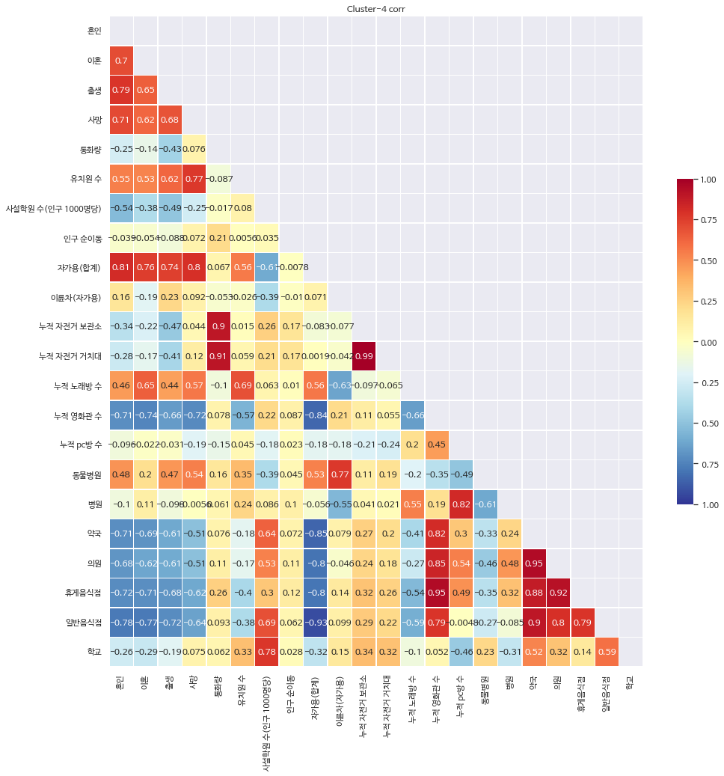
인구 순 이동, 자전거 거치대, 자전거 보관소 수의 표준편차가 큰 것으로 보아 증감 폭이 매우 큰 것을 알 수 있습니다. 특히 자전거 거치대, 자전거 보관소의 경우 데이터가 계속 누적되는 데이터이기 때문에 클러스터 내에서 자전서 거치대와 보관소를 비교적 많이 설치했다는 것을 알 수 있습니다.

**D 그룹 : 강동구, 강북구, … ,영등포구 (나머지 16개 구) <약 2018년 이전>**



각 feature 간의 상관관계가 낮은 것으로 보입니다. 클러스터 2와 클러스터 4는 구는 같고 년도가 나누어지는데 그래서인지 상관관계 그래프와 표준편차 그래프가 비슷한 양상을 보입니다.

**E 그룹 : 금천구, 서대문구, 용산구, 종로구, 중구**



의원의 표준편차가 낮을 것을 볼 때 클러스터 4는 의원 수의 변화가 거의 없었다는 것을 알 수 있습니다.

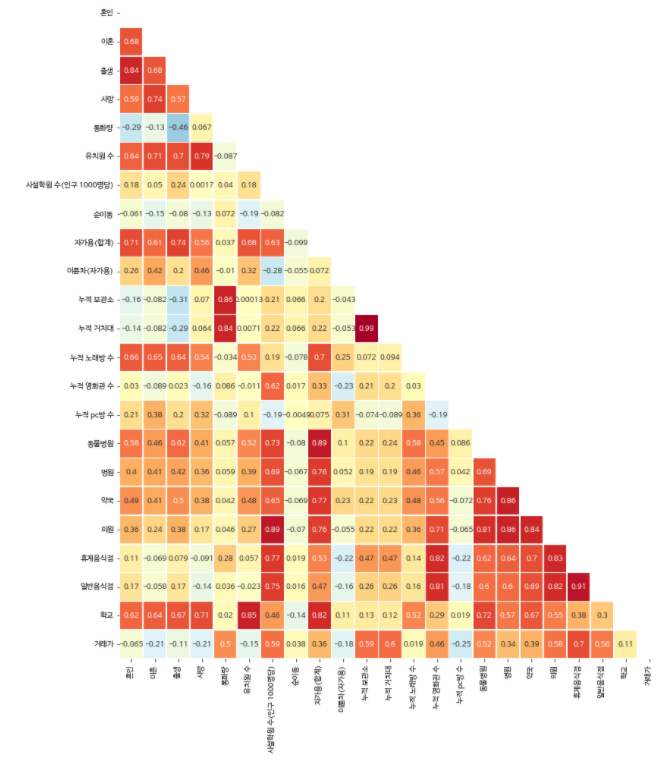
의원과 휴게음식점의 상관관계가 높을 것을 볼 때 증감 양상이 비슷하다는 것을 알 수 있습니다.

**결론**

feature간 상관관계가 높거나 표준편차가 큰 feature가 존재하는 클러스터도 있지만 그렇지 않은 클러스터도 있는 것을 알 수 있습니다.

이는 centroid와 데이터 간의 거리를 최소로 하는 K-means 클러스터링의 특징 때문입니다.

**<Regression>**



* 거래가가 포함된 correlation matrix

텍스트, 영수증이(가) 표시된 사진

자동 생성된 설명

* 거래가에 대한 correlation

집값을 예측하기 전, 거래가에 대한 correlation을 구해봤을 때, 위와 같은 결과를 보여줬습니다. 이전에 예측했던 것과 다르게 통화량이 생각보다 상관관계가 낮았고, 예상대로 음식점의 수나 공공자전거 수, 사설 학원 등 인프라에 대한 상관관계가 높았음을 알 수 있었습니다.

1. 각 클러스터 별 집값 상승 요인 확인 및 예측(Decision Tree)

텍스트이(가) 표시된 사진

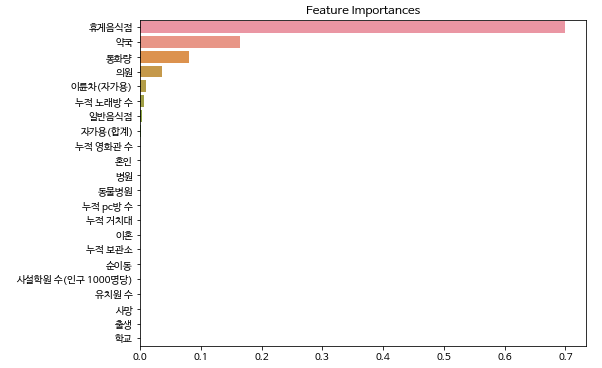
자동 생성된 설명

Decision Tree를 이용하여 각 클러스터에 대해 regression을 진행하여 집값을 예측하고 중요 feature를 알아내었습니다.

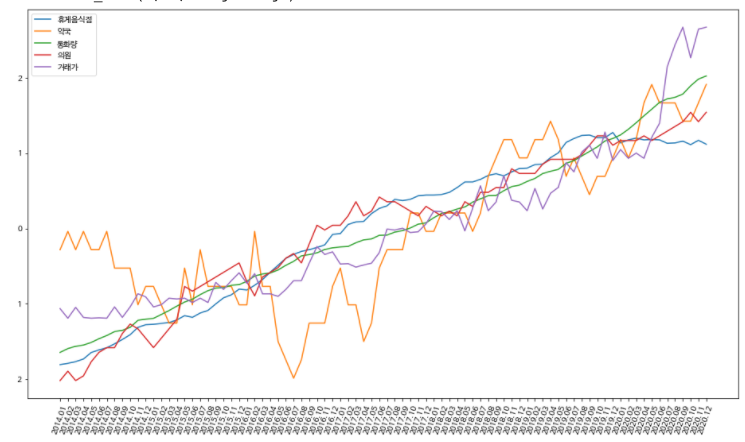
1. K = 0 (노원, 송파, 강서)

텍스트이(가) 표시된 사진

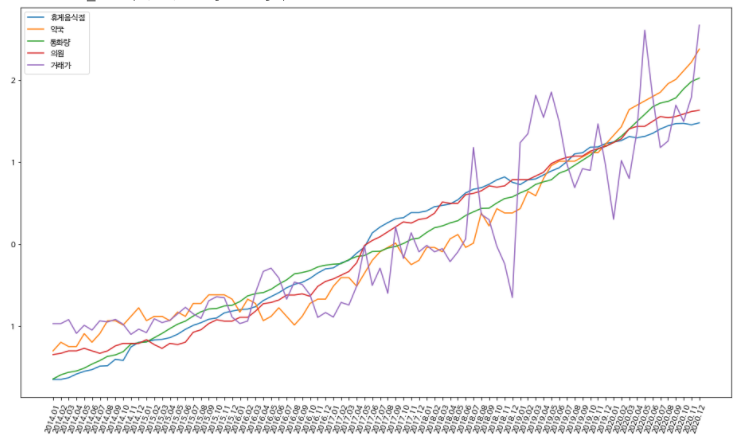
자동 생성된 설명



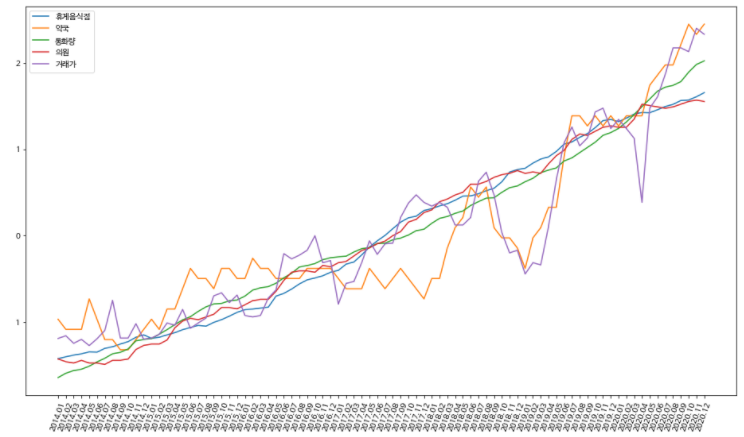
노원 송파 강서 구역에서의 집값 예측 중요 feature로는 휴게음식점, 약국, 통화량, 의원 등이 있었습니다. 이 feature들과 거래가의 변화 추이를 살펴보기 위해 해당 feature들만 추출하였고, 구에 따라 그래프를 각자 그려보았습니다.



* 노원구



* 송파구



* 강서구

모든 feature의 변화량을 한번에 보기 위해 StandardScaling을 한 데이터를 그래프에 그려넣었으며, x축은 기간, y축은 scale된 값을 의미합니다. 학습된 결과에서는 집값을 예측하는데 제일 중요한 feature가 휴게 음식점으로 나와 휴게 음식점의 증가와 거래가의 증가 추이가 제일 비슷할 것으로 예상되었는데, 이와 달리 대부분의 feature가 거래가와 비슷한 증가 추이를 보였습니다.

또한 노원구에서는 약국의 증가와 거래가의 증가가 크게 연관이 없었지만 강서구에서는 상당히 비슷한 추이를 가져 두번째로 예측 중요 feature에 오른 것으로 판단하였습니다.

텍스트이(가) 표시된 사진

자동 생성된 설명텍스트이(가) 표시된 사진

자동 생성된 설명텍스트이(가) 표시된 사진

자동 생성된 설명

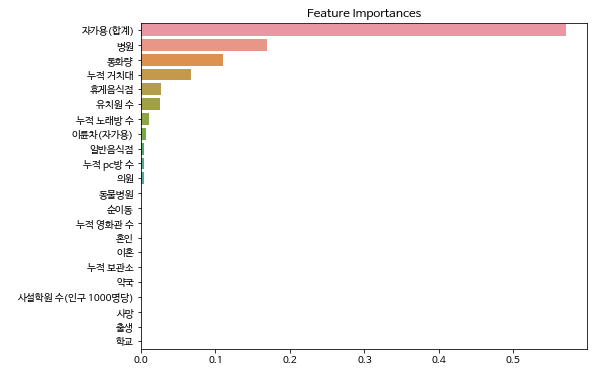
* 왼쪽부터 노원, 송파, 강서의 correlation

각 구의 거래가에 대한 correlation을 살펴보았을 때는 학습을 하기 전 제일 중요할 것이라 생각되었던 통화량이 예상대로 거래가와 높은 상관관계를 가졌음을 알 수 있었습니다.

1. K = 1 (은평구, 강북구, 광진구, 동작구, 영등포구, 성북구, 서초구 등 2017 이후)

텍스트이(가) 표시된 사진

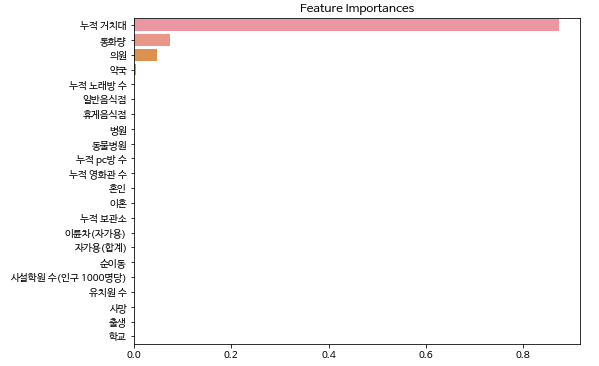
자동 생성된 설명



1. K = 2 (강남구)

텍스트이(가) 표시된 사진

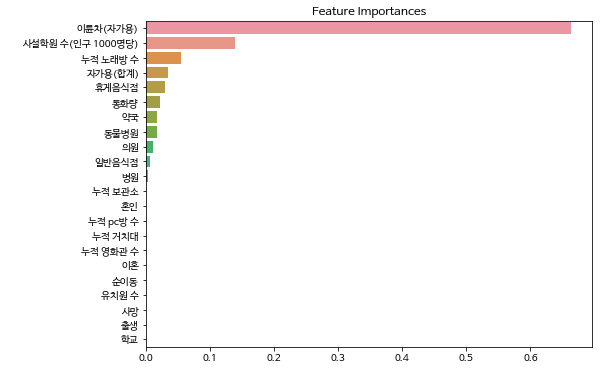
자동 생성된 설명



1. K = 3 (은평구, 강북구, 광진구, 동작구, 영등포구, 성북구, 서초구 등 2017 이전)

텍스트이(가) 표시된 사진

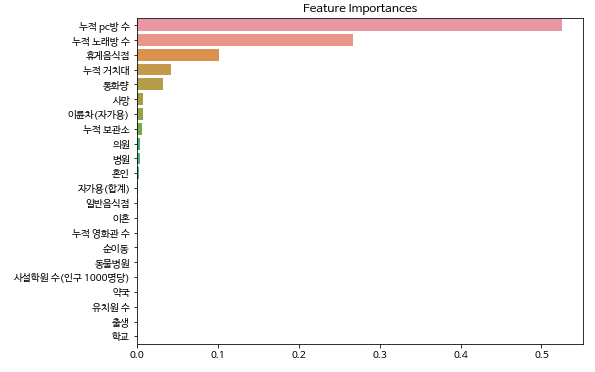
자동 생성된 설명



1. K = 4 (금천구, 용산구, 서대문구, 중구, 종로구)

텍스트이(가) 표시된 사진

자동 생성된 설명



그 외에 다른 클러스터들을 살펴보았을 때 전부 거래가의 변화량에 영향을 주는 feature들이 다르게 나온 것을 알 수 있었습니다. 하지만 소수의 feature가 항상 높은 중요도를 가졌고, 이 feature들이 클러스터마다 다른 결과를 보인 것과 train의 Rmse와 test의 Rmse가 차이가 매우 큰 것을 보아 overfitting이 되었다고 판단하여 Random Forest를 이용하여 좀 더 나은 결과를 얻는 방식을 선택하기로 하였습니다.

1. 각 클러스터 별 집값 상승 요인 확인 및 예측(RandomForest)

텍스트이(가) 표시된 사진

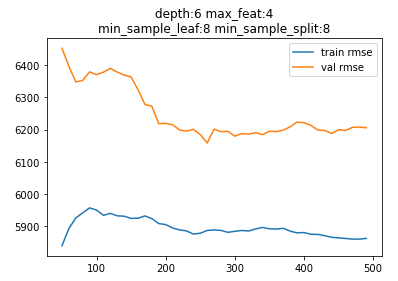
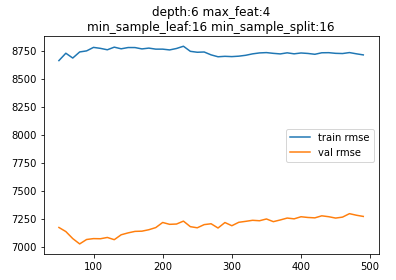
자동 생성된 설명

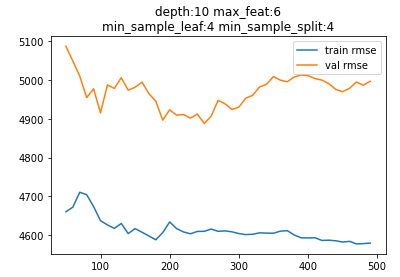
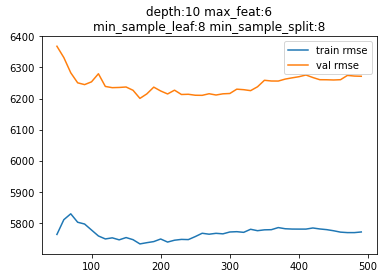
위의 학습과 달리 validation set을 만들고 이를 통해 overfitting이 되었는지를 확인한 후 test를 하기로 하였습니다.

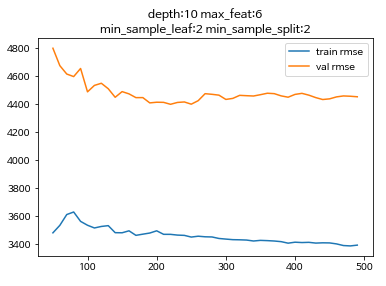
텍스트이(가) 표시된 사진

자동 생성된 설명

Hyperparameter를 변경해가며 학습을 하여 overfitting을 최소화하는 모델을 찾기 위해 노력하였습니다. GridSearchCV를 이용하여





여러 hyperparameter로 학습을 해본 결과 약간의 overfitting을 피할 순 없었지만, Max\_depth=12, max\_features=8, min\_sample\_leaf=2, min\_sample\_split=2, tree=200 정도에서 가장 좋은 결과를 얻었습니다.

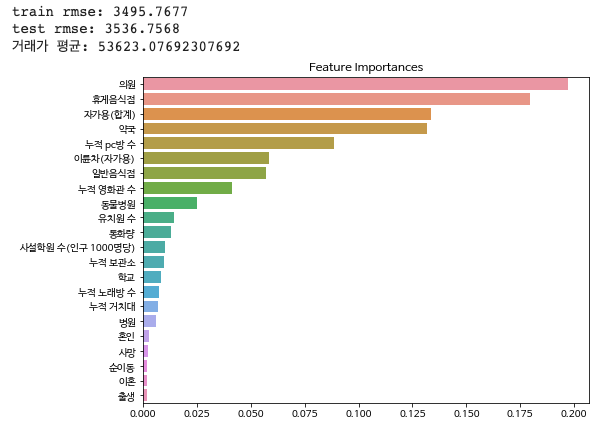
텍스트이(가) 표시된 사진

자동 생성된 설명

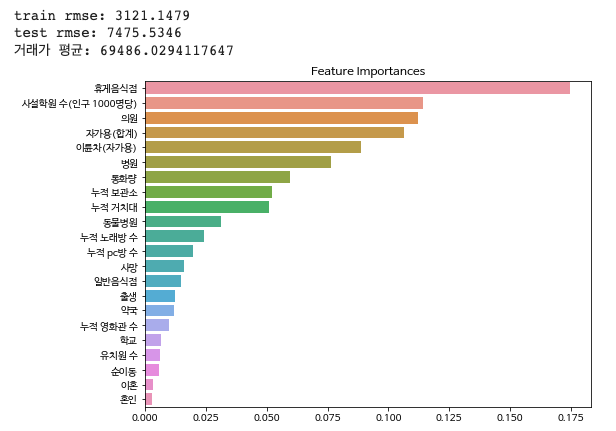
마지막으로 위에서 얻은 hyperparameter를 이용해 test set에 대하여 돌렸을 때 위와 같은 Rmse가 나왔고 Decision tree를 이용하여 학습했을 때보다 훨씬 나은 결과를 얻게 되었습니다.

위와 같은 방식으로 모든 클러스터에 대해 random forest를 돌렸고, 다음과 같은 결과가 나왔습니다.

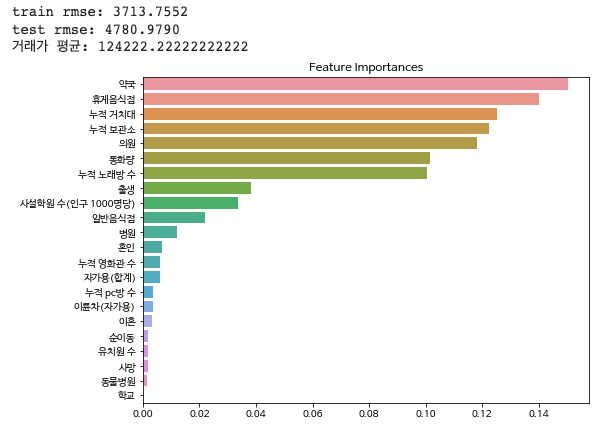
1. K = 0 (노원, 송파, 강서)



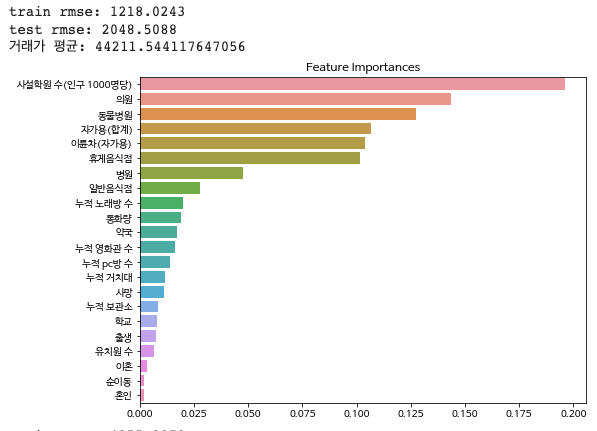
1. K = 1 (은평구, 강북구, 광진구, 동작구, 영등포구, 성북구, 서초구 등 2017 이후)



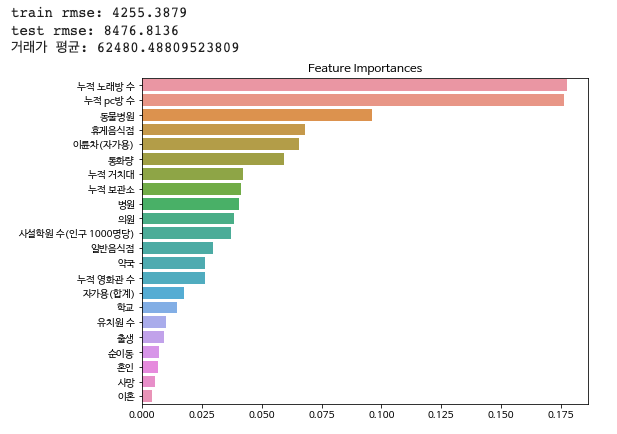
1. K = 2 (강남구)



1. K = 3 (은평구, 강북구, 광진구, 동작구, 영등포구, 성북구, 서초구 등 2017 이전)



1. K = 4 (금천구, 용산구, 서대문구, 중구, 종로구)



대부분의 rmse가 decision tree를 이용하여 예측을 했을 때보다 더 좋은 결과를 얻었습니다. 또한 k=1, 3에서의 중요 feature가 decision tree를 이용했을 때와 많이 바뀌었음을 알 수 있었습니다. 2017년 전에는 영등포구, 성북구, 서초구 등의 자치구에서는 사설 학원의 증가 수가 집값 변동에 큰 영향을 미치는 반면 2017년 이후에는 음식점의 수가 더 크게 영향을 주었음을 알 수 있었습니다.

**<데이터 분석 결과>**

1. K-means clustering 결과

**A. 강남구**

**B. 강서구, 노원구, 송파구**

**C. 금천구, 서대문구, 용산구, 종로구, 중구**

**D. 강동구, 강북구, … ,영등포구 (나머지 16개 구)**

구는 위와 같이 4개의 그룹으로 나누어 볼 수 있었는데 각 그룹의 특징을 정리하면 다음과 같다

**A. 강남구**

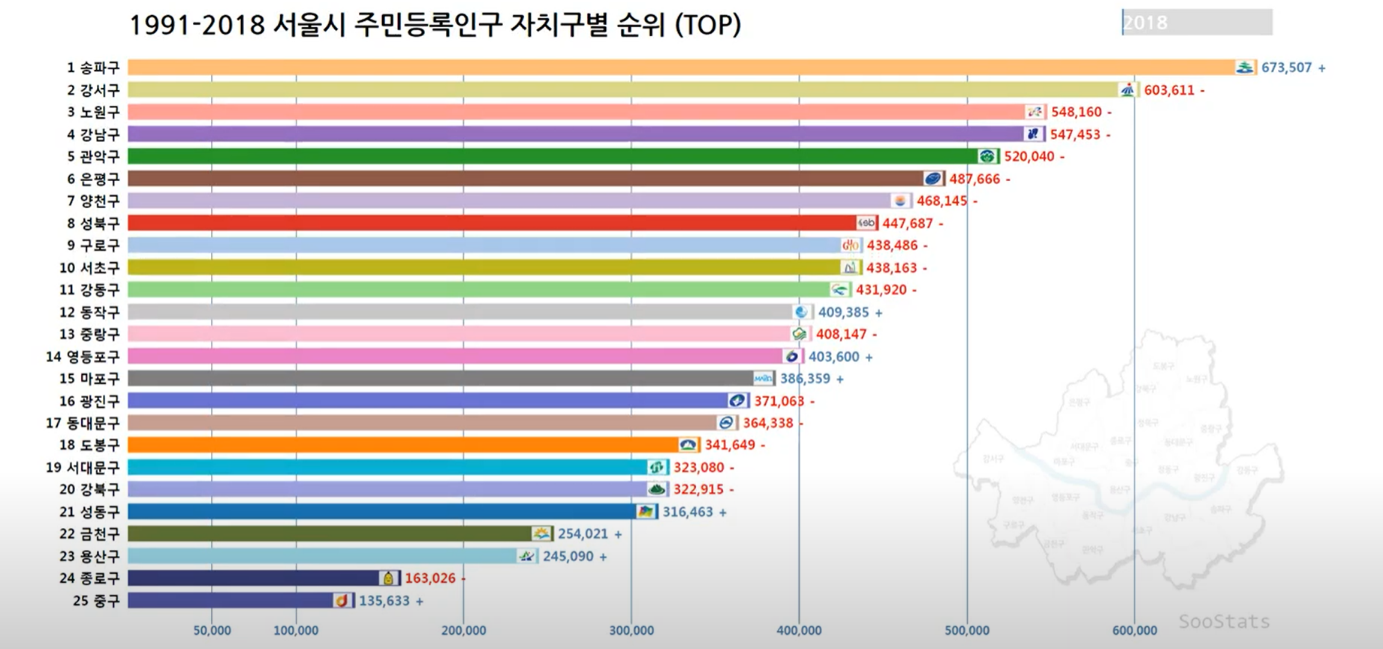
강남구는 **사설학원 수, 영화관, 병원, 약국, 의원, 휴게음식점, 일반음식점** 와 같은 column에서 다른 구들과 차이가 났고 추가적인 column 으로는 **자가용(합계), 동물병원 수, 학교 수**가 있었습니다. 실제 데이터를 보더라도 위와 같은 부분에서 다른 구들과 확실하게 값이 차이가 났습니다. 또한 위 column들의 값들이 강남구만 압도적으로 값이 더 컸습니다

**B. 강서구, 노원구, 송파구**

강서구,노원구,송파구는 **유치원 수, 자가용 수, 학교 수** 면에서 다른 구들과 차이가 났음을 알 수 있었습니다. 또한 실제 자치구별 통계가 의미하듯이 이 3개의 구가 가장 인구수가 많았는데 이를 혼인 ,이혼, 출생,사망 column의 값들이 다른구들에 비해 큰 것을 통해 유추해 볼 수 있었습니다.

**C. 금천구, 서대문구, 용산구, 종로구, 중구**

금천구, 서대문구, 용산구, 종로구, 중구는 **노래방 수** 에서 다른 구들과 가장 차이가 있었습니다. 전반적으로 이 그룹은 중심값들의 차이가 음수인 것을 통해 인구의 수가 적음을 알 수 있었습니다.



**최종 결론**

위 그래프는 자치구별 인구 순위를 나타낸 통계표인데 실제로 저희 팀이 분석한 결과가 이 인구 순위와 비슷하게 맞아 떨어짐을 알 수 있었습니다. 현재 통계는 2018년이고 가장 많은 인구가 있는 송파,강서,노원구 가장 적은 인구가 있는 금천,용산,종로,중구, 서대문구 그리고 그외 나머지 구들이 있고 마지막으로 강남구만 유별나게 그룹되어 있습니다.

아마 저희가 가져온 X값들이 대부분 인구와 밀접한 연관을 가지는 값들이고 인구의 크기에 따라 크게 3가지 그룹으로 구분이 되었음을 알 수 있었고 그 중에서도 강남구는 유독 다른 구들과의 차별성을 보이는 것을 알 수 있었습니다.

다음은 2번째 clustering 분석 최종 결론입니다.

**최종 결론**

feature간 상관관계가 높거나 표준편차가 큰 feature가 존재하는 클러스터도 있지만 그렇지 않은 클러스터도 있는 것을 알 수 있습니다. 이는 centroid와 데이터 간의 거리를 최소로 하는 K-means 클러스터링의 특징 때문입니다. 구가 비슷하고 년도와 개월 수가 분리되어 클러스터링 되는 경우에는 두 개의 클러스터의 표준편차와 상관관계 점수가 비슷하게 나옵니다. 클러스터가 구의 특징을 어느 정도 반영한다고 해석할 수 있습니다. 표준편차가 작다고 해서 그 feature가 클러스터의 주요 feature인 것을 아닌 것으로 보입니다. 이것 역시 centroid와 데이터 사이 거리를 최소로 하는 K-means 클러스터링으로 분류했기 때문으로 보입니다. 또한 y값을 반영하지 않고 클러스터 내부의 feature들로만 비교했기 때문에 regression 결과로 나온 중요 feature와 차이가 있는 것을 알 수 있습니다.

Regression 최종 결론입니다.

**최종 결론**

이 작업을 하기 전에는 각 자치구별 집값 변동 요인이 대부분 비슷할거라 생각을 하였습니다. 하지만 regression을 하고 중요 feature를 도출해본 결과, 각 자치구마다 집값 변동에 대한 여러 특색이 있었음을 알 수 있었습니다.

대부분의 자치구에서 음식점, 약국, 의원 등의 인프라가 중요 요인으로 나왔으나, 강남구에서는 특히 통화량에 영향을 많이 받는 모습을 보였습니다. 또한 k=1, 3에 속하는 서초, 성북, 영등포 등의 지역에서는 2017년을 기준으로 휴게음식점의 중요도가 사설 학원보다 더 변동에 큰 영향을 주는 요인으로 판단하였습니다.

하지만 학습을 시킨 데이터가 시간에 따라 증가하는 데이터이기 때문에 feature의 값들이 증가함에 따라 집값의 요인이 변한다고 하기에는 무리가 있다고 생각하였습니다. 또한 데이터의 부족과 실제 집값을 예측하기에는 부족한 feature수로 인해 overfitting이 발생하였고, 제대로 된 예측을 하지 못했다고 판단을 하였습니다. 그럼에도 각 지역에서 증가하는 feature가 다름을 알 수 있었고 각 지역의 특색에 대한 간단한 insight를 었을 수 있었습니다.

**코드 첨부**

**저희 팀이 각자 분석한 colab 링크입니다. 대부분 비슷하나 clustering분석1, clustering 분석2, Regression 분석을 하였습니다.**

[**클러스터링 분석 1**](https://colab.research.google.com/drive/1vOPf5g1pxQVwyL-S3gjHgJKBoaw7WXpk?usp=sharing)

[**클러스터링 분석 2**](https://colab.research.google.com/drive/1QopayCw6aDo-NL2DawcB--ud8FxYywBU?usp=sharing)

[**Regression 분석**](https://drive.google.com/file/d/1BoKJGiu7fdSUGwwtEcQE6HYwSkKfK29e/view?usp=sharing)