비즈니스 애널리틱스 PBL 모듈 1

고양시 집값 예측 보고서

지도교수 : 김정욱 교수님

4 조

19010027 신예진

19010069 오지원

19010269 박다인

20011821 박지인

목차

1. 가설 설정

1.1 선정이유 및 분석배경

부동산 R114 에서 발표한 자료에 따르면 최근 2016 년부터 2021 년 11 월까지 서울의 평균 아파트 매매 가격은 108.6% 상승했다.[1] 전국 각지의 도시 집값 추이는 상승 곡선을 그려오고 있다. 특히나 서울특별시 집값은 예외 없이 폭발적으로 상승해 2016 년부터 5 년간 108.6%의 상승률을 보였다. 폭등하는 집값에도 불구하고 수도권을 중심으로 한 대도시 인구 과밀화는 지속되고 있으며 여전히 사람들의 '내 집 마련' 욕구는 수도권을 향한다. 수도권 인구 집중을 위한 대안으로 위성도시를 택하는 경우도 생겨나고 있으며 현실적인 내 집 마련 방안을 생각해보았다. 수도권에 위치하고 있어 서울 생활권을 최대한 누릴 수 있는 거주지를 선정해 집값 예측을 시도하고자 한다.

수도권 위성 도시에 내 집 마련을 계획하는 만큼 '살기 좋은' 도시를 거주지로 선정하려 한다. 그리고 살기 좋을수록 집값이 높을 것이라고 예상했다. 서울과 가까운 이점을 근간으로 가지고 있다면 이왕이면 삶의 질이 높은 도시에 거주할 것이라고 보았기 때문이다. 이에 살기 좋은 도시의 결정 요인을 문화, 환경, 교통 세 가지 측면으로 정하였다. 문화생활 향유가 쉽고, 녹지환경 조성이 잘 되어 있으며, 교통 인프라가 잘 갖추어져 있는 세 가지 관점을 윤택한 주거 선택 요인으로 결정했다. 물론 이 요인들이 좋은 거주지의 유일한 변수가 아니라는 것을 알기에 선정한 주요 3 요인 외에도 유의미한 결정변수를 발굴하고자 한다.

집값 예측을 위한 타겟 도시는 실제 인구수 100 만 이상의 서울 근교 고양시로 잡았다. 구체적인 데이터로 확인하기 이전에 고양시는 위에서 언급한 세 가지 관점에서 살기 좋은 도시라는 가정이 가능하다. 첫째, 고양시는 문화적으로는 고양어울림누리와 같은 복합문화예술공간을 충분히 가지고 있다. 둘째, 공원 면적과 녹지 면적이 주거 환경과 어우러져 자연친화적, 인간친화적이다. 셋째, 경의중앙선, 3 호선 두 개의 지하철 노선이 지나며 KTX 역이 지나가는 '역세권' 교통 인프라를 누린다. 서울과의 연결은 물론 시 내부의 이동 환경도 편리하다는 이점이 뚜렷하다.

---

[1] 뉴스 https://www.mk.co.kr/news/realestate/view/2021/12/1116510/ OR 부동산 R114 https://m.r114.com/

주택가격 상승에 영향을 미치는 요인의 중요도 분석 = A importance analysis of factors affecting the rise in housing price-건국대학교, 홍유경, 학위논문(석사)

수도권 내 집 마련이라는 목표를 검토하기 위해 '고양시는 살기 좋은 도시이다'와 '살기 좋은 동네는 집값이 높다'라는 두 가지 가설을 세웠다. 본 프로젝트에서는 변수 설정과 수집을 통해 두 가설을 검토하고자 한다. 나아가 고양시의 집값을 결정하는 요소를 탐색해 실제 변수들이 집값 결정에 얼마나 영향을 미치는지 알아보려 한다.

2. 변수 설정

2.1 변수 선정배경

집값에 영향을 미치는 요인은 매우 다양하다. 선행연구에 따르면 과거에는 **투자목적**이 가장 강하게 나타난다. 거주지로서의 '집' 보다는 투자를 위한 수단으로 작용하는 경향이 짙었다. 주택 구매 시 중요한 요인은 유동적으로 현금화가 가능한지의 유무였으며 주변지역 개발가능성이 있어 부동산 투자 수익률을 얼마나 도출해낼 수 있는지가 더 중요했다. 다양한 국가 정책들은 금융제도 변화와 더불어 소비자들의 투자방식에 변화를 유도해왔다.

한편, 현대에 들어 집에 대한 인식은 보금자리의 성격을 더 강력하게 갖게 되었다. 투자수익률을 중시하며 자산으로서 강력한 의미를 갖던 집이 주변입지환경, 경제성, 아파트단지특성, 개별주택특성과, 소유 자산 대비 아파트 가격, 주차장, 교통편리성, 교육환경, 단지환경, 전용면적 비율 등 한층 다양한 요소에 의해 선택되고 있다. 또한 실제 집값 형성에 이와 같은 다양한 요소들이 관여되고 있다. 경기변동, 정책변화, 층수, 부대시설, 단지디자인, 단지규모, 브랜드, 대출조건, 자연환경, 생활시설과 같은 외부요인들도 주택가격에 많은 영향을 끼치게 되었다. 사회 문화 욕구가 높아진 사람들에 의해 집값 결정 요인은 다양해졌다. 이러한 요인들 중 고양시 집값 결정에 큰 영향을 미치는 변수들이 무엇인지 고심했고 도시 특성에 맞는 변수를 사용하고자 했다.

고양시 도시계획정책관에서 발간한 **2035 년 고양도시기본계획 열람 자료**[2]를 참고한다.

기정 도시기본계획(2020)	
실천과제	전략
녹색전원도시	- 녹색문화도시계획추진 - 녹도 및 자전거도로의 지속적인 확충 - 한강둔치를 포함한 자연형 하천정비
문화복지도시	- 여성과 노인의 고용증대 - 시민참여형의 전통문화행사 생활화

2

	- 한강과 호수공원을 연계한 시민문화공간 조성
정보교류도시	- 고부가가치의 지식정보산업벨트 구축 - E-bussiness 및 통상전문인력 육성 중 - 유비쿼터스의 시범도시 조성 - 대외, 대북교류의 장 조성

변경 도시기본계획(2030 년)	
실천과제	전략
서울,경기 서북부권의 중심기능 강화 및 MICE 기반 국제고류 도시	- mice 복합단지, 한류월드 완성, 배후지원물류단지 조성 등 국제교류 거점기능의 강화. - 문화생활배후지원을 추으로 수도권 서북부 각 도시와 상생발전 도모
자연과 공존하는 시민행복도시	- 자연과 공존하는 친환경 녹색도시 지향 - 녹색교통 체계 구축, 첨단도시 조성 - 지역특성을 고려한 쾌적한 도심 및 주거환경 조성
문화예술기반의 창조문화산업 도시	- 문화예술기반 확충 - 문화예술 중심의 관광 거점 형성
시민참여의 공동체 도시	- 시민참여 기반의 맞춤형 보건복지체계 구축
통일한국을 선도하는 평화도시	- 평화인권도시의 기반 구축 - 남북교류협력의 배후거점

참고자료를 통해 고양시는 실제 **녹색전원도시, 문화복지, 문화예술** 분야에 힘쓰고 있다는 사실을 파악할 수 있다. 설정한 두 가지 가설 '고양시는 살기 좋은 도시이다'와 '살기 좋은 동네는 집값이 높을 것이다' 라는 가설을 증명하고자 교육환경, 교통인프라, 문화시설, 생활시설과 같은 요인의 변수 설정에 타당성을 획득했다.

2.2 변수 결정

사전조사를 바탕으로 **아파트 실거래가(2021 년 3 월 1 일~2022 년 3 월 1 일)**을 종속변수로 삼았다. 독립변수는 주택 특성과 주택외 특성으로 두 가지 관점 모두를 취했다.

- ✓ **주택 특성**
  - ✓ 동
  - ✓ 전용면적(㎡)
  - ✓ 계약년
  - ✓ 계약월
  - ✓ 건축년도
  - ✓ 거래금액(만원)
- ✓ **주택외 특성**
  - ✓ 학원개수
  - ✓ 대형마트 개
  - ✓ 병원개수
  - ✓ 스타벅스개수
  - ✓ 배스킨

- ✓ 써브웨이
- ✓ 편의점개수
- ✓ 층
- ✓ 집 유형
- ✓ 어린이집/유치원,공원면적
- ✓ 도서관
- ✓ 지하철역개수
- ✓ 초/중/고
- ✓ 반려동물등록수

3. 변수 설명

3.1 변수 설명

- **Column1** : 레코드별 고유 코드로 행별 ID 로 기능
- **동** : 고양시 법정동 정보
- **전용면적(㎡)** : 거래된 아파트, 연립, 주택의 면적 정보
- **계약년** : 거래된 아파트, 연립, 주택의 계약 연도 정보
- **건축년도** : 거래된 아파트, 연립, 주택의 건축 연도 정보
- **거래금액(만원)** : 아파트, 연립, 주택이 거래된 금액
- **학원개수** : : 해당 법정동의 사설학원의 수. 학교교과교습학원(입시검정 및 보습, 국제화, 예능, 특수교육, 종합, 기타), 평생직업교육학원(직업기술, 국제화, 인문사회, 기예, 종합)의 수를 포함
- **대형마트 개수** : 해당 법정동 내에 있는 대형마트의 개수로 쇼핑 센터를 포함.
- **병원개수** : 해당 법정동의 병원과 의원 개수
- **스타벅스개수** : 해당 법정동 내에 있는 스타벅스 개수
- **배스킨** : 해당 법정동 내에 있는 배스킨라빈스 개수
- **써브웨이** : 해당 법정동 내에 있는 써브웨이 개수
- **편의점개수** : 해당 법정동 내에 있는 편의점 개수
- **층** : 거래된 해당 아파트, 연립, 주택의 층 (주택의 경우 모두 1 층)
- **집 유형** : 아파트, 연립, 주택 중 가구가 속하는 주택 유형
- **어린이집/유치원** : 해당 법정동 내에 있는 미취학아동 교육 및 보육 시설 개수 (어린이집과 유치원 개수)
- **공원면적(㎡)** : 해당 법정동에 있는 공원 부지의 면적
- **도서관** : 해당 법정동의 도서관 수. 시도 도서관, 교육청 도서관, 사립 도서관의 수를 포함
- **지하철역개수** : 해당 법정동에 위치한 지하철 역의 개수
- **초/중/고** : 해당 법정동에 있는 초등학교, 중학교, 고등학교 개수의 합
- **반려동물등록수** : 해당 법정동에 등록된 반려동물등록 수

3.2 데이터 설명

2021 년 3 월 1 일부터 ~ 2021 년 12 월 31 일까지 10 개월 간 경기도 고양시의 데이터 12,387 개를 조사했다.

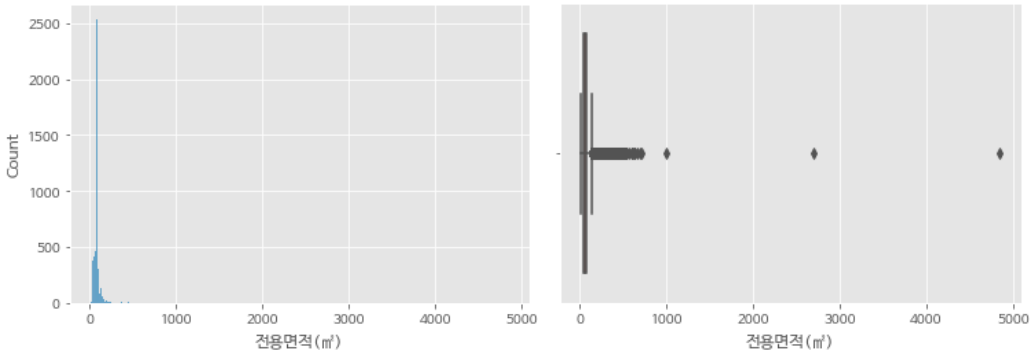
```
RangeIndex: 12386 entries, 0 to 12385
Data columns (total 21 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   동          12386 non-null  int64
 1   전용면적(㎡)   12386 non-null  float64
 2   계약년        12386 non-null  int64
 3   계약월        12386 non-null  int64
 4   건축년도       12386 non-null  int64
 5   거래금액(만원)  12386 non-null  int64
 6   학원개수       12386 non-null  int64
 7   대형마트 개수    12386 non-null  int64
 8   병원개수       12386 non-null  int64
 9   스타벅스개수     12386 non-null  int64
 10  배스킨        12386 non-null  int64
 11  써브웨이       12386 non-null  int64
 12  편의점개수      12386 non-null  int64
 13  층          12386 non-null  int64
 14  집 유형       12386 non-null  int64
 15  어린이집/유치원   12386 non-null  int64
 16  공원면적       12386 non-null  float64
 17  도서관        12386 non-null  int64
 18  지하철역개수     12386 non-null  int64
 19  초/중/고      12386 non-null  float64
 20  반려동물등록수    12386 non-null  int64
dtypes: float64(3), int64(18)
```

* 12386 개의 행, 21 개의 열로 이루어진 데이터.

- ✓ **전용면적**, **계약년, 거래금액**, **층**, **건축년도**, **법적동, 집 유형,동** 변수는 국토교통부 실거래가 공개시스템을 이용했다.
- ✓ **편의점 개수, 대형마트 개수**, **스타벅스 개수, 배스킨 개수, 서브웨이 개수, 학원 개수, 도서관 개수, 병원 개수, 반려동물 등록 수**는 공공데이터포털 사이트에서 소상공인진흥공단 데이터 소스로 활용했다.
- ✓ **어린이집/유치원, 초중고 학교**는 경기데이터드림사이트에서 데이터를 가져왔다.
- ✓ **근린공원현황은** 고양시청 생활정보 사이트에서 가져왔다.
- ✓ **지하철 역 개수**는 네이버 지도를 웹 스크래핑했다. .

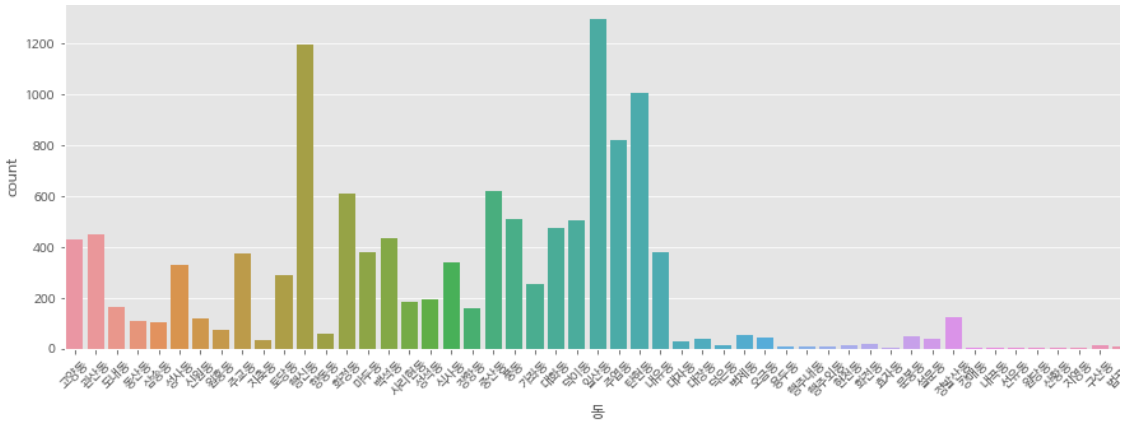
각종 데이터 관련 사이트와 스크래핑을 통해 수집된 데이터를 파이썬을 활용해 연결한다. 본격 예측에 앞서 데이터 분포를 살펴보고자 했고 주요변수를 중심으로 EDA 를 진행한다.

- **전용면적(히스토그램,박스플롯)**



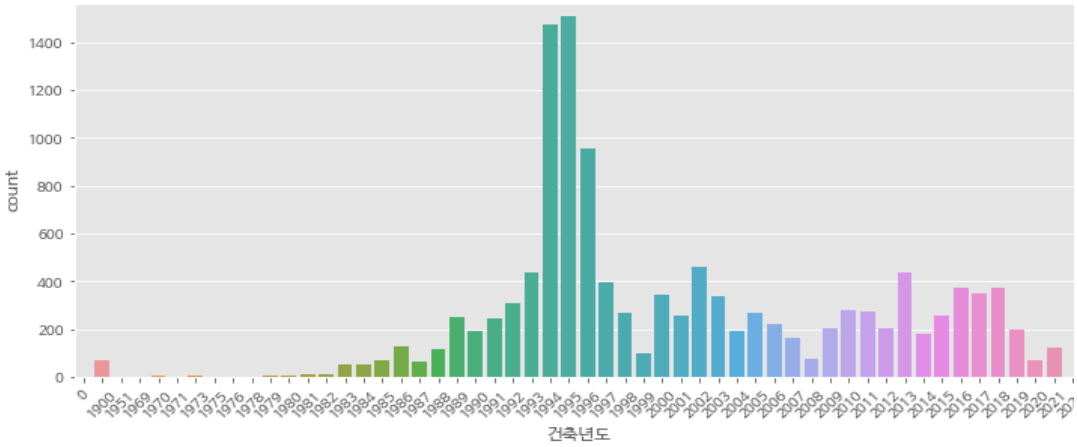
전용면적의 분포가 매우 넓어서 꼬리가 긴 형태의 히스토그램 그래프를 볼 수 있다. 거의 100 평 이내에 값들이 모여 있다.

- **동(막대그래프)**



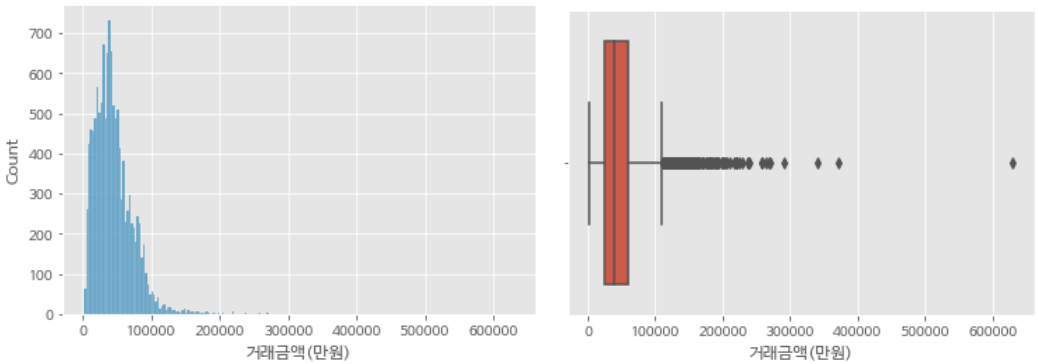
일산동, 행신동, 탄현동, 주엽동에 거래 주택 수가 많은 것을 볼 수 있다.

- **건축년도(막대그래프)**



거래된 주택 중 1995, 1994, 1996 년도에 지어진 건물이 가장 많은 것을 확인할 수 있다.

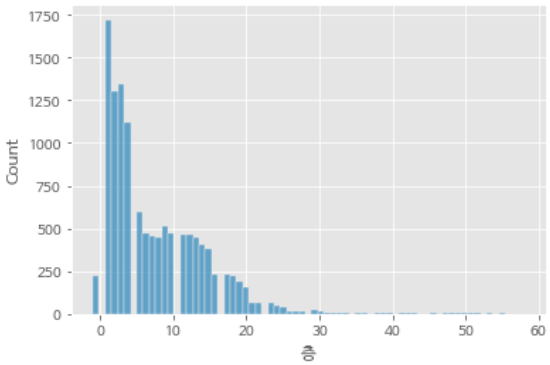
- **거래금액(히스토그램)**



-

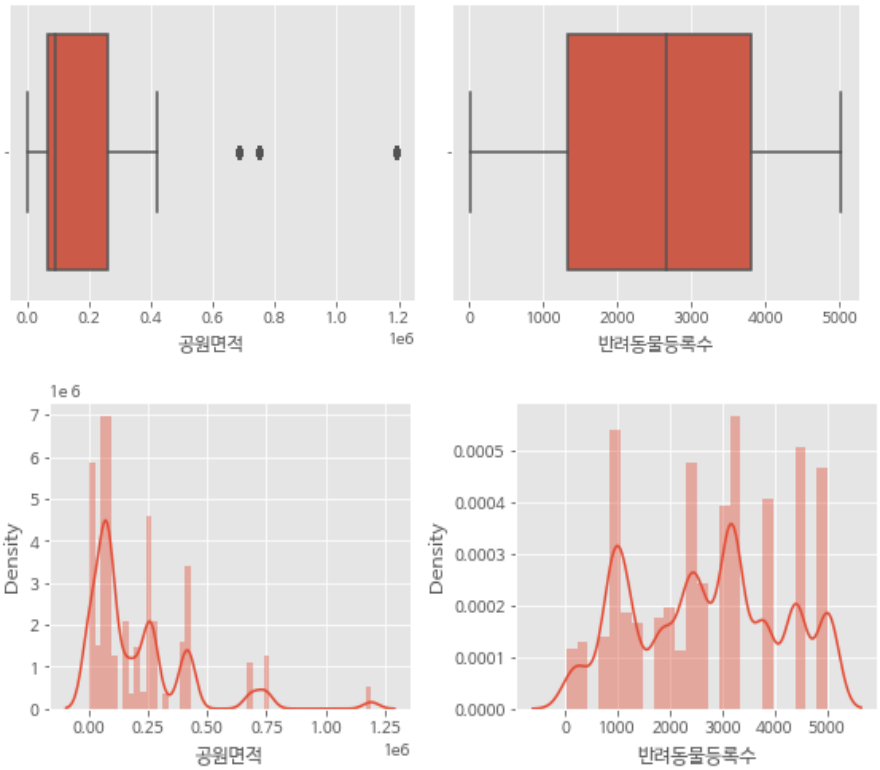
거래금액은 거의 2 억 내로, 거래그램 히스토그램 또한 꼬리가 긴 형태를 보였다.

- **층(히스토그램)**



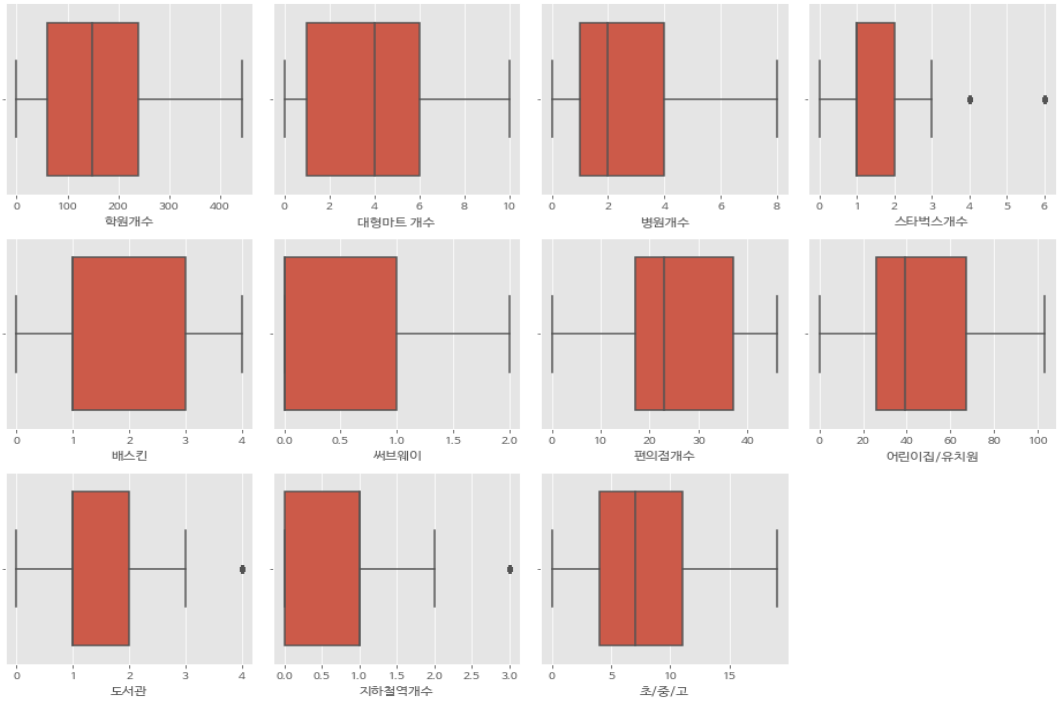
보통 10 층 내외이다.

- **공원면적, 반려동물등록수(박스플롯, 히스토그램)**

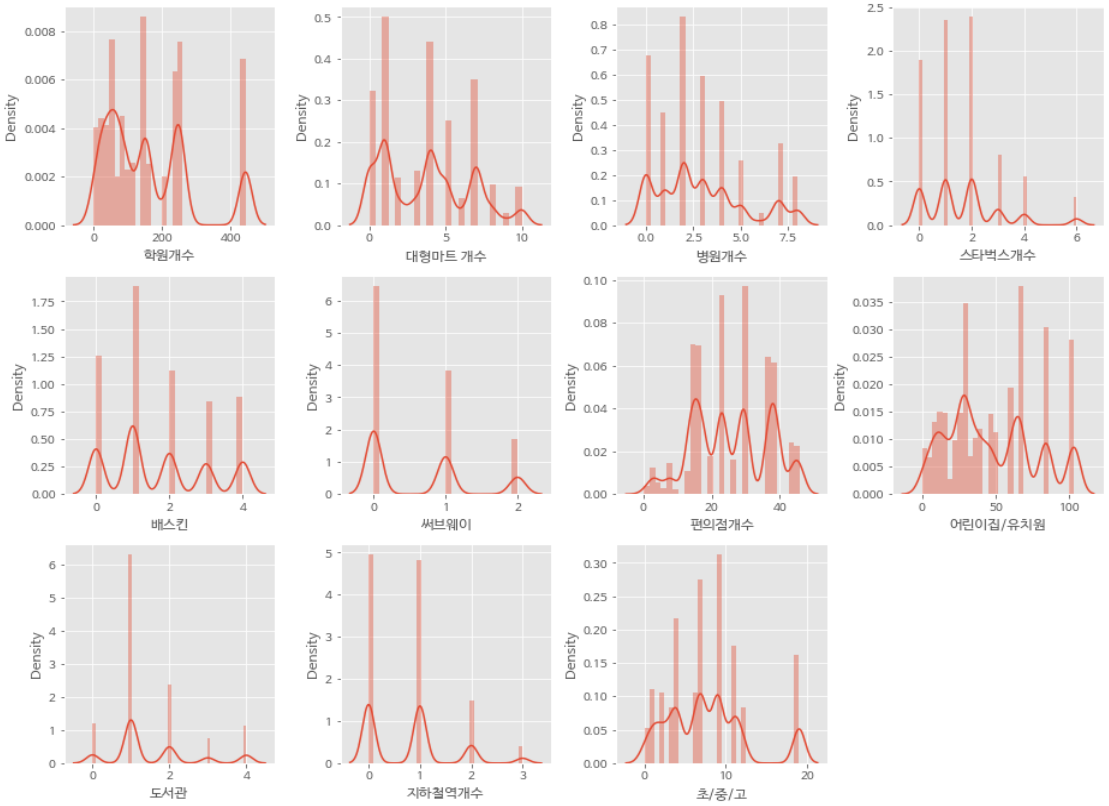


- 공원면적 : 0.2 내외(단위:제곱미터)
- 반려동물 등록 수 : 3000 내외.

- 학원, 대형마트, 병원, 스타벅스, 배스킨, 써브웨이, 편의점, 어린이집/유치원, 도서관, 지하철역, 초/중/고 개수(박스플롯, 히스토그램)



-



- 학원 : 법정동당 학원은 200 개 내외
- 대형마트 : 법정동당 대형마트는 보통 1 개씩 위치.
- 병원 : 법정동당 병원은 2 개 내외.
- 스타벅스 : 법정동당 스타벅스는 1 개 내외.
- 배스킨라빈스 : 법정동당 배스킨라빈스는 1 개 내외.
- 써브웨이 : 써브웨이는 없는 동이 많음.
- 어린이집/유치원 : 법정동당 어린이집/유치원 20 개 내외.
- 도서관 : 법정동당 1 개 내외.
- 지하철역 : 지하철역이 없는 곳이 많지만 보통 1 개씩 있는 편.
- 초/중/고 : 법정동당 초등학교, 중학교, 고등학교의 합이 5 개 내외.

4. 데이터 전처리

4.1 결측치 처리

결측치가 존재하지 않았다. 다만 편의시설의 개수 중 법정동별로 없는 곳들이 존재하여 0 으로 대체 처리한 것 외에는 없다.

4.2 라벨 인코딩

법정동명(일산동, 주엽동 등), 집 유형(아파트, 연립, 주택)과 같이 범주형 변수들은 원-핫 인코딩을 이용해 더미화를 시킬 수도 있지만, 법정동이 너무 많은 탓에 원-핫 인코딩을 사용하기 보다는 숫자로 **라벨인코딩**을 진행했다.

- **법정동명 라벨인코딩**

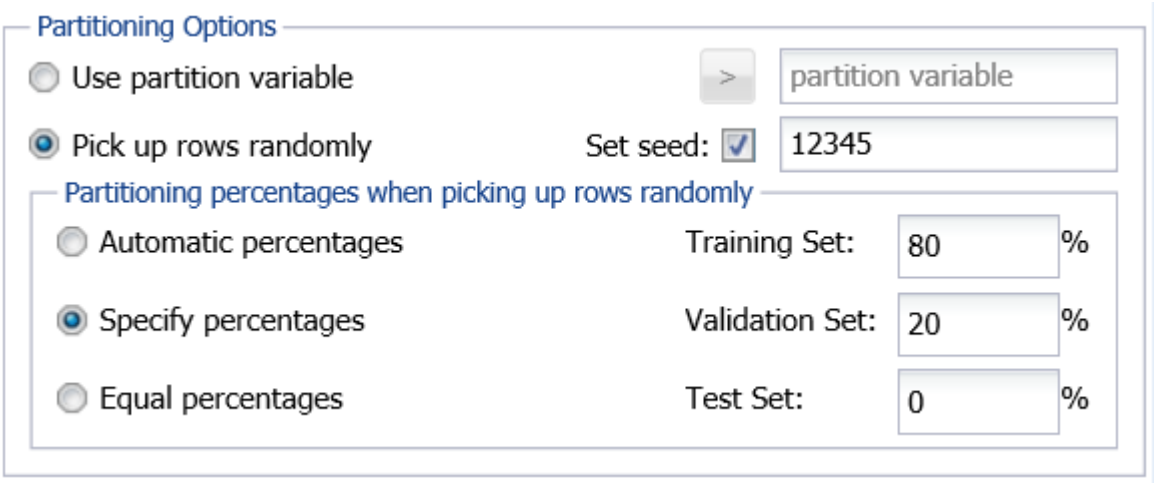
  {일산동:32, 행신동:43, 탄현동: 40, 주엽동:36, 중산동:37, 화정동:49, 풍동:42, 덕이동:11, 대화동:9, 관산동:3, 백석동:16, 고양동:2, 내유동:6, 마두동:14, 주교동:35, 식사동:26, 성사동:24, 토당동:41, 가좌동:0, 성석동:25, 사리현동:19, 도내동:12, 장항동:33, 정발산동:34, 신원동:27, 동산동:13, 삼송동:21, 원흥동:31, 향동동:46, 벽제동:18, 문봉동:15, 오금동:28, 대장동:8, 설문동:23, 지축동:39, 대자동:7, 화전동:48, 덕은동:10, 구산동:4, 현천동:47, 용두동:29, 행주내동:44, 법곳동:17, 행주외동:45, 지영동:38, 내곡동:5, 효자동:50, 강매동:1, 선유동:22, 산황동:20, 원당동:30}

- **집유형 라벨인코딩**

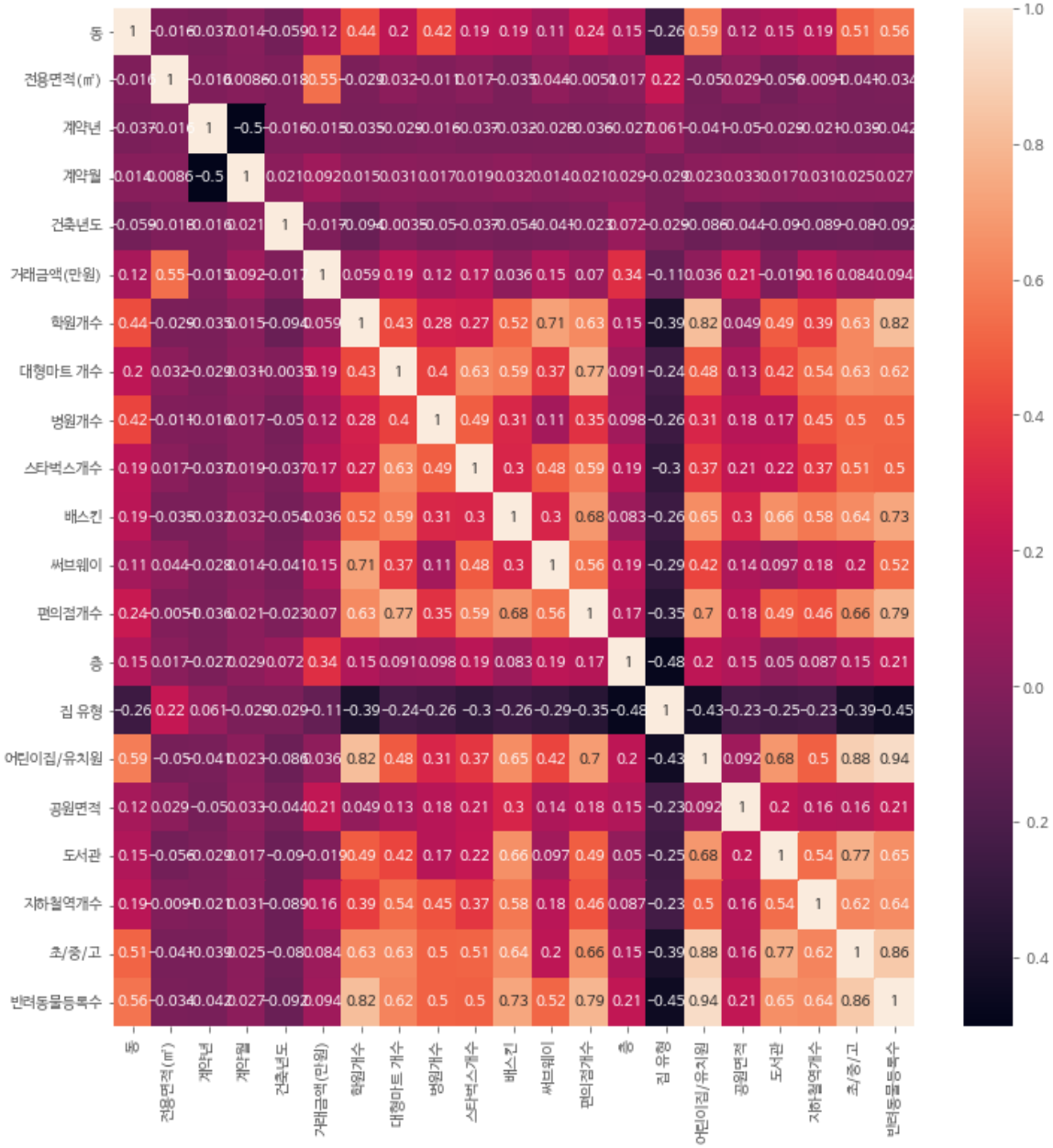
  {아파트:0, 연립:1, 주택:520}

4.3 데이터 분할

Xlminer 의 Standard Partition 기능을 통해 훈련데이터와 테스트데이터 비율을 8:2 로 설정하였다. 데이터의 수가 12000 여 건으로 많았기에 평소 작은 데이터를 분할할 때의 6:4 비율보다 학습력을 더 높이고자 했다. 필요한 검증 수준은 확보하되, 학습 데이터의 비중을 늘려 좋은 모델을 발견하기 위한 결정이다.



5. 상관관계분석



각 변수들 간의 상관관계를 분석해보았다. 밝을 수록 높은 양의 상관관계를 의미하며 어두울수록 높은 음의 상관관계를 보인다. 피어슨 상관계수를 기준으로 강한 상관관계의 기준인 0.7 을 잡고 높은 상관관계를 보이는 변수를 골라 정리한 것은 다음 페이지와 같다.

<높은 양의 상관관계를 가지는 변수 조합>

- ✧ 학원개수 - 써브웨이 : 0.71
- ✧ 학원개수 - 어린이집/유치원 : 0.82
- ✧ 학원개수 – 반려동물등록수 : 0.82
- ✧ 대형마트개수 - 편의점개수 : 0.77
- ✧ 배스킨라빈스 – 반려동물등록수 : 0.73
- ✧ 편의점개수 – 반려동물등록수 : 0.79
- ✧ 어린이집/유치원 - 편의점개수 : 0.7
- ✧ 어린이집/유치원 - 초중고 : 0.88
- ✧ 어린이집/유치원 – 반려동물등록수 : 0.94
- ✧ 도서관 - 초/중/고 : 0.77
- ✧ 초/중/고 – 반려동물등록수 : 0.86

정리하면 다음과 같은 결론을 얻을 수 있다.

- ➢ **학원개수**는 써브웨이, 어린이집/유치원, 반려동물등록수와 상관관계가 높다.
- ➢ **편의점의 개수**는 대형마트, 어린이집/유치원,반려동물등록수의 개수와 상관관계가 높다.
- ➢ **반려동물등록수**는 배스킨라빈스, 어린이집/유치원, 초/중/고의 개수와 상관관계가 높다.
- ➢ **초/중/고 개수**는 도서관, 어린이집/유치원의 개수와 상관관계가 높다.

다중공선성 문제를 해결하기 위해 거래금액과의 상관관계가 낮은 항목들은 삭제한다.

- 병원개수 & 서브웨이 :0.71 -> **서브웨이** 삭제 (0.15)
- 병원개수 & 어린이집/유치원 :0.82 -> **병원개수** 삭제 (0.13)
- 병원개수 & 반려동물등록수 :0.82 -> **병원개수** 삭제 (0.13)
- 대형마트 개수 & 편의점개수 :0.72 -> **대형마트** 삭제 (0.19)
- 배스킨 & 반려동물 등록수 :0.73 -> **반려동물 등록 수** 삭제 (0.094)
- 서브웨이 & 학원개수 :0.71 -> **서브웨이** 삭제 (0.15)
- 편의점개수 & 반려동물 등록수 :0.79 -> **반려동물 등록 수** 삭제 (0.094)
- 어린이집/유치원 & 초/중/고 :0.88 -> **초중고** 삭제 (0.084)
- 어린이집/유치원 & 반려동물 등록수 :0.94 -> **반려동물 등록 수** 삭제 (0.094)

- 도서관 & 초/중/고 :0.77 -> **초중고** 삭제 (0.084)
- 초/중/고 & 반려동물 등록수 :0.86 -> **반려동물 등록 수** 삭제 (0.094)

총 6 개 (병원 개수, 어린이집/유치원, 편의점 개수, 학원개수, 도서관, 초/중/고, 반려동물 등록 수)의 변수를 삭제하는 결론을 도출했다.

6. 다중 선형회귀분석

다중 선형회귀분석은 설명변수가 둘 이상인 회귀분석이다. 이는 결과값을 '예측'하는 머신러닝 방법으로서 가장 많이 사용되는 것이다. 본 모듈에서도 분석의 목표인 '고양시 집값 예측'을 위해 다중선형회귀분석을 사용하고자 한다. 이를 통해 집값 예측뿐만 아니라 집값 결정에 큰 영향을 미치는 변수들이 무엇인지 확인할 수 있고, 미래의 집값 예측을 위해 고려해야할 변수들의 목록을 알 수 있다. 또한 집값을 예측하는 식을 알아내는 것이 가능하다.

6.1 전역탐색(Best Subsets)을 통한 상위 모델 3 가지 선정

위에서 삭제한 6 개의 변수를 제외하고 Best Subsets 을 통해 전역탐색을 실시했다.

Best Subsets Details						
Subset	#Coeffi	RSS	Mallow	R2	Adjuste	Probabi
Subset 1	1	8.3E+12	9175.553	5.55E-16	5.55E-16	0
Subset 2	2	7.34E+12	6968.604	0.115758	0.115668	0
Subset 3	2	5.94E+12	3758.496	0.28398	0.283907	0
Subset 4	3	5.49E+12	2724.011	0.338296	0.338162	0
Subset 5	3	5.02E+12	1643.697	0.394908	0.394786	0
Subset 6	4	4.85E+12	1254.032	0.415433	0.415256	8.2E-248
Subset 7	4	4.85E+12	1253.725	0.415449	0.415272	9.4E-248
Subset 8	5	4.75E+12	1015.747	0.428025	0.427794	2.2E-202
Subset 9	5	4.72E+12	948.5207	0.431548	0.431318	3.2E-189
Subset 10	6	4.67E+12	838.4478	0.437421	0.437137	4.3E-168
Subset 11	6	4.61E+12	702.1661	0.444562	0.444282	6.9E-141
Subset 12	7	4.57E+12	605.9928	0.449707	0.449374	4.7E-122
Subset 13	7	4.52E+12	506.3344	0.45493	0.454599	8.7E-102
Subset 14	8	4.48E+12	414.3579	0.459854	0.459472	1.96E-83
Subset 15	8	4.47E+12	383.3989	0.461477	0.461096	4.73E-77
Subset 16	9	4.43E+12	307.3407	0.465567	0.465135	8.97E-62
Subset 17	9	4.43E+12	299.5049	0.465978	0.465546	3.82E-60
Subset 18	10	4.41E+12	259.6555	0.468171	0.467687	2.66E-52
Subset 19	10	4.4E+12	223.6651	0.470057	0.469575	9.14E-45
Subset 20	11	4.39E+12	206.6482	0.471054	0.470519	1.18E-41
Subset 21	11	4.38E+12	179.6438	0.472469	0.471936	5.86E-36
Subset 22	12	4.37E+12	164.4546	0.473369	0.472784	3.18E-33
Subset 23	12	4.37E+12	161.64	0.473517	0.472932	1.26E-32
Subset 24	13	4.35E+12	127.8261	0.475394	0.474758	0
Subset 25	13	4.34E+12	88.14922	0.477473	0.476839	0
Subset 26	14	4.32E+12	60.36901	0.479034	0.478349	6.24E-12
Subset 27	14	4.32E+12	46.43471	0.479764	0.47908	7.59E-09
Subset 28	15	4.3E+12	15	0.481516	0.480782	N/A

Subset	#Coeffi	RSS	Mallow	R2	Adjuste	Probabi
Subset 28	15	4.3E+12	15	0.481516	0.480782	N/A
Subset 27	14	4.32E+12	46.43471	0.479764	0.47908	7.59E-09
Subset 26	14	4.32E+12	60.36901	0.479034	0.478349	6.24E-12

adjusted R 제곱 값을 기준으로 최상위 모델 3 개를 선정한다.

○ **subset26** 에 포함된 변수 : *전용면적 / 계약월 / 건축년도 / 거래금액 / 대형마트 개수 / 병원 개수 / 스타벅스 개수 / 배스킨 개수 / 서브웨이 개수 / 층 / 집 유형 / 공원면적 / 지하철 역 개수*

○ **subset27** 에 포함된 변수 : *전용면적 / 계약년 / 계약월 / 건축년도 / 거래금액 / 대형마트 개수 / 병원 개수 / 스타벅스 개수 / 배스킨 개수 / 서브웨이 개수 / 층 / 공원면적 / 지하철 역 개수*

○ **subset28** 에 포함된 변수 : *전용면적 / 계약년 / 계약월 / 건축년도 / 거래금액 / 대형마트 개수 / 병원 개수 / 스타벅스 개수 / 배스킨 개수 / 서브웨이 개수 / 층 / 집 유형 / 공원면적 / 지하철 역 개수*

이 세 가지 모델과 함께, 상관관계분석에서 삭제하기로 한 6 가지 변수가 삭제 되지 않고 모든 변수가 포함된 모델 하나를 상위 네 개의 모델로 선정했다.

6.2 상위 4 개 모델 비교

✧ 모델 1 :*동 / 전용면적(m^2) / 계약년 / 계약월 / 건축년도 / 학원개수 / 대형마트개수 / 병원개수 / 스타벅스개수 / 배스킨개수 / 서브웨이개수 / 편의점개수 / 층 / 집유형 / 어린이집 / 유치원 / 공원면적 / 도서관 / 지하철역개수 / 초/중/고 / 반려동물등록수*

✧ 모델 2 : *전용면적 / 계약월 / 건축년도 / 거래금액 / 대형마트 개수 / 병원 개수 / 스타벅스 개수 / 배스킨 개수 / 서브웨이 개수 / 층 / 집 유형 / 공원면적 / 지하철 역 개수*

✧ 모델 3 : *전용면적 / 계약년 / 계약월 / 건축년도 / 거래금액 / 대형마트 개수 / 병원 개수 / 스타벅스 개수 / 배스킨 개수 / 서브웨이 개수 / 층 / 공원면적 / 지하철 역 개수*

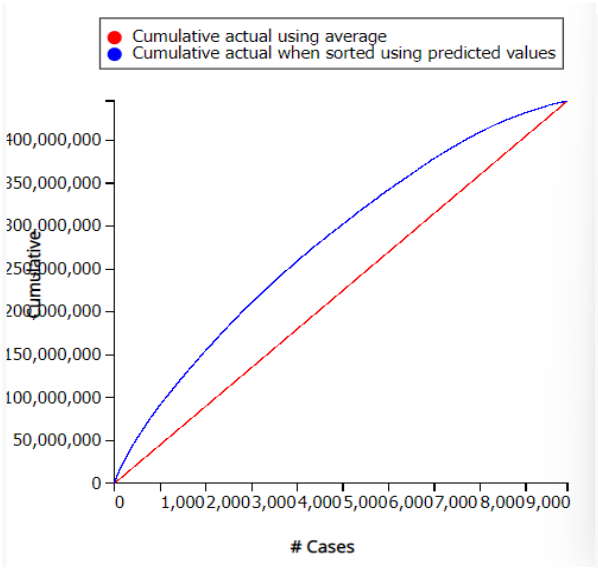
✧ 모델 4 : *전용면적 / 계약년 / 계약월 / 건축년도 / 거래금액 / 대형마트 개수 / 병원 개수 / 스타벅스 개수 / 배스킨 개수 / 서브웨이 개수 / 층 / 집 유형 / 공원면적 / 지하철 역 개수*

	모델 1(모든 변수)	모델 2(Subset26)	모델 3(Subset27)	모델 4(Subset28)
R^2	0.50703	0.479034	0.479764	0.481516
adjusted R^2	0.506033	0.478349	0.47908	0.480782
train_MSE	412800239.86	436243791.56	435632331.45	434165160.77
val_MSE	561178257.85	572041058.91	576299694.72	565465616.56
train_RMSE	20317.49	20886.45	20871.81	20836.63
val_RMSE	23689.20	23917.38	24006.24	23779.52
SED_ERROR	20339.05	20901.22	20886.57	20852.42
P-value 0.05 이상	0 개	0 개	0 개	0 개

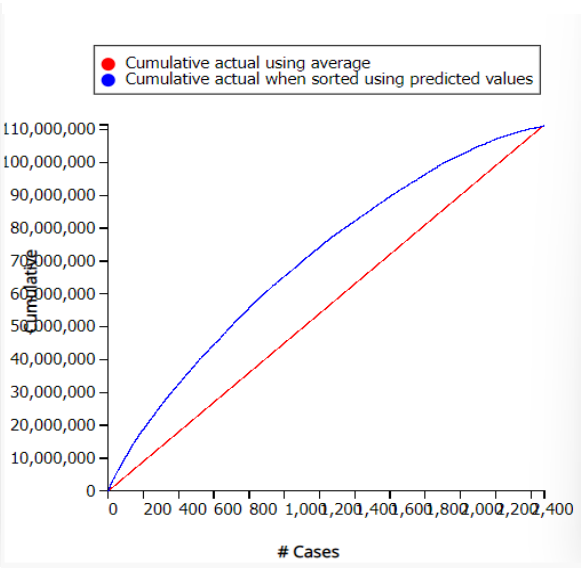
- **R2** : 모형의 적합도를 평가하는 평가지표. 0 부터 1 사이의 값을 가지고 0 에 가까울수록 상관관계의 정도가 없다고 하고, 1 에 가까울수록 상관관계의 정도가 크다고 할 수 있다.
- **RMSE** : MSE 에 루트를 씌운 값. MSE 보다 이상치에 덜 민감하다. 루트를 씌웠기 때문에 큰 오류값에 있어 크게 패널티를 주는 이점이 있으며 극단적이지 않다는 장점이 있다.
- **MSE** : 예측값과 정답의 차이를 제곱한 값.

6.3 최상위 모델 선정 및 해석

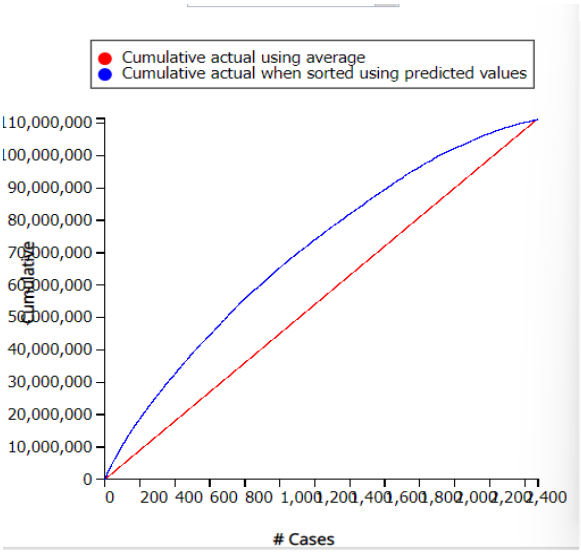
	모델 1(모든 변수)	모델 2(Subset26)	모델 3(Subset27)	모델 4(Subset28)
adjusted R^2	1 위	4 위	3 위	2 위
MSE	1 위	3 위	4 위	2 위
RMSE	1 위	3 위	4 위	2 위
SED_ERROR	1 위	4 위	3 위	2 위



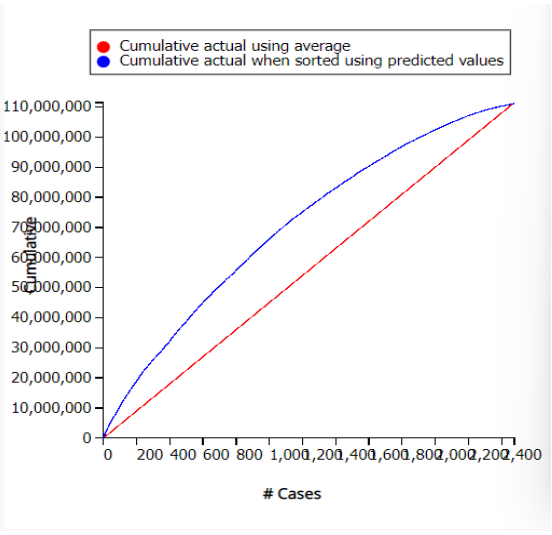
모델 1



모델 2



모델 3



모델 4

모델의 적합성을 adjusted R^2 로 계산해 판단하고, 오류정도는 MSE, RMSE, STD_ERROR 값을 종합적으로 판정했다. 네 모델의 리프트차트를 비교하였으나 육안으로 보기에 큰 차이가 나지 않아 엑셀마이너가 제공하는 리프트 차트만으로는 좋은 모델을 선정에 한계가 있다. 모든 변수를 포함한 모델 1 이 네 가지 중에서 성능이 가장 좋게 나타났다.

하지만 다중공선성 이슈를 고려해야 하므로 상관관계분석을 통해 변수 6 개가 제거된 나머지 3 가지 모델 중 최상위 모델을 선택했다. 그 결과 최종적으로 선택된 모델은 모델 4(Subset28)이다. 이하는 모델 4 의 Coefficients 표다.

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	-13107311	-16738740.34	-9475880.702	1852577.702	-7.0751745	1.59E-12
동	239.59551	203.4071573	275.7838669	18.46152686	12.9780984	3.34E-38
전용면적(㎡)	194.06155	188.533851	199.589253	2.819962431	68.81707	0
계약년	6307.7	4511.208745	8104.191261	916.4818854	6.88251465	6.24E-12
계약월	970.26563	802.3927658	1138.138496	85.64051689	11.3295163	1.43E-29
건축년도	182.56334	145.4688988	219.6577829	18.92376823	9.64730378	6.3E-22
학원개수	-39.02955	-46.46559764	-31.5935089	3.793505776	-10.288518	1.06E-24
대형마트 개	1962.6005	1728.276463	2196.924493	119.5406402	16.4178515	8.77E-60
스타벅스개수	-2336.196	-2793.066247	-1879.32666	233.0726008	-10.023471	1.56E-23
배스킨	-3827.96	-4298.328257	-3357.592166	239.958748	-15.952576	1.38E-56
써브웨이	6557.1639	5453.061513	7661.266336	563.2589946	11.6414722	4.04E-31
층	1004.6818	937.5487757	1071.8149	34.24800159	29.3354879	1.8E-181
집 유형	-2897.876	-3880.26191	-1915.489719	501.1652892	-5.7822756	7.59E-09
공원면적	0.0194856	0.017300307	0.021670809	0.001114808	17.4788481	2.13E-67
지하철역개수	6260.3876	5566.657802	6954.117426	353.9069859	17.6893587	5.8E-69

- 모델 4 에서 목표변수인 집값을 예측하는 데 사용된 변수는 총 14 가지다. 변수 14 개 모두 P-value 가 0.5 를 넘지 않아 통계적으로 유의함을 확인했으며 목표변수인 집값을 예측하는 데 영향을 준다고 할 수 있다.
- estimate 값은 '서브웨이'가 6557.16 으로 가장 높았다. 반면 해당 수치가 가장 낮은 변수는 '배스킨'으로 -3827.96 의 값을 가진다. 이는 '서브웨이'는 가장 높은 양의 영향을 미치고, '배스킨'은 가장 높은 음의 영향을 미친다고 볼 수 있다. Estimate 값을 통해 도출한 집값 예측 식은 다음과 같다.

- ***집값 = 서브웨이*6557.16 + 계약년*6307.7 + 지하철역개수*6260.39 + 대형마트개수*1962.6 + 층*1004.68 + 계약월*970.26 + 동*239.59 + 전용면적*194.06 + 건축년도*182.56 + 공원면적*0.02 - 학원개수*39.03 - 스타벅스개수*2336.20 - 집유형*2897.88 - 배스킨*3827.96***

7. K-최근접 이웃 기법

K-최근접 이웃(K-Nearest Neighbor)은 머신러닝의 대표적인 분류 알고리즘이다. 본 모듈에서는 집값 예측이 목표였으나 knn 을 활용해 집값 예측에 활용되는 데이터를 군집화를 시도했다. 데이터 군집화를 통해 군집화의 결정요인이 되는 변수가 무엇인지 확인할 수 있고 집값 평균에 관한 인사이트를 도출할 수 있다.

K-NN 은 거리를 사용해 가까운 데이터를 동일 군집으로 묶는 알고리즘이다. 거리를 활용해 계산이 이루어지므로 예측변수는 모두 수치형 데이터여야만 한다.

변수 설명에서 제시된 21 개의 데이터를 해당 알고리즘에 적용해본 결과, knn 알고리즘에 있어 Xlminer 가 다룰 수 있는 한계를 초과했다. 이에 집값 분석과의 관계에서 유의했던 변수 일부를 선정해 전체 변수를 축소했다.

축소한 변수는 다음과 같다.

- 거래금액(만원)
- 전용면적(㎡)
- 대형마트개수
- 병원개수
- 스벅개수
- 써브웨이
- 층
- 공원면적(㎡)
- 지하철역개수

위의 9 개 변수는 변수간 상관분석에서 거래금액과의 상관관계에서 0.1 이상을 보인 요소들이다. 이외의 변수는 0.1 에도 달하지 못하므로 제외한다. knn 역시나 훈련세트 80%, 검증세트 20%로 설정해 데이터를 분할했다. 집값과 가장 상관성이 높았던 전용면적을 취하여 예측변수를 '거래금액(만원)'과 '전용면적(㎡)'로 두고 이외 변수를 결과변수로 설정해 분류했다.

첫 번째 모델로 클래스가 4 인 지하철역개수를 결과변수로 설정한다. 거래금액과 전용면적이 예측변수로 작용했을때 지하철역개수를 분류의 기준으로 두고 k=1 로 설정하면 다음과 같은 결과가 도출된다.

- 예측변수 : 거래금액(만원), 전용면적(㎡)
- 결과변수 : 지하철역개수
- k=1

Confusion Matrix				
Actual₩Predicted	0	1	2	3
0	4154	58	6	1
1	139	3909	40	4
2	30	115	1109	3
3	12	29	16	284

Error Report			
Class	# Cases	# Errors	% Error
0	4219	65	1.540649443
1	4092	183	4.472140762
2	1257	148	11.77406523
3	341	57	16.71554252
Overall	9909	453	4.571601574

Metrics	
Metric	Value
Accuracy (#correct)	9456
Accuracy (%correct)	95.42839843

훈련세트의 오분류율은 위과 같이 4.57%에 달하며

Confusion Matrix				
Actual₩Predicted	0	1	2	3
0	762	222	43	10
1	248	646	102	31
2	40	106	167	6
3	13	35	6	40

Error Report			
Class	# Cases	# Errors	% Error
0	1037	275	26.51880424
1	1027	381	37.09834469
2	319	152	47.64890282
3	94	54	57.44680851
Overall	2477	862	34.80016149

Metrics	
Metric	Value
Accuracy (#correct)	1615
Accuracy (%correct)	65.19983851

검증세트의 오분류율은 다음과 같이 33.8%에 달한다. 이어서 k=2 일 때 결과를 보자.

- 예측변수 : 거래금액(만원), 전용면적(㎡)
- 결과변수 : 지하철역개수
- k=2

Confusion Matrix				
Actual₩Predicted	0	1	2	3
0	4029	170	16	4
1	881	3104	86	21
2	172	375	705	5
3	51	116	46	128

Error Report			
Class	# Cases	# Errors	% Error
0	4219	190	4.503436833
1	4092	988	24.14467253
2	1257	552	43.91408115
3	341	213	62.46334311
Overall	9909	1943	19.60843677

Metrics	
Metric	Value
Accuracy (#correct)	7966
Accuracy (%correct)	80.39156323

훈련세트의 오분류율은 19.6%로 k=1 대비 대폭 상승했으며

Confusion Matrix				
Actual₩Predicted	0	1	2	3
0	843	170	21	3
1	340	601	67	19
2	57	134	128	0
3	23	49	8	14

Error Report			
Class	# Cases	# Errors	% Error
0	1037	194	18.70781099
1	1027	426	41.48003895
2	319	191	59.87460815
3	94	80	85.10638298
Overall	2477	891	35.97093258

Metrics	
Metric	Value
Accuracy (#correct)	1586
Accuracy (%correct)	64.02906742

검증세트의 오분류율도 늘어났다. 최적의 k 를 찾기 위해 예측변수와 결과변수를 동일하게 하고 k=10 일 때 Search Log 를 확인해본 결과

Search Log

K	% Misclassification
1	34.88090432
2	36.011304
3	35.93056116
4	36.33427533
5	36.61687525
6	36.21316108
7	36.2535325
8	35.76907549
9	36.29390392
10	36.45538958

Note: Scoring will be done using K=1

k=1 일때 최적의 분류를 하는 것으로 확인되었다.

두 번째 모델로 클래스가 6 인 스타벅스개수를 결과변수로 설정한다. 첫 번째 모델과 동일하게 거래금액과 전용면적을 예측변수로 잡고 결과변수만을 상이하게 스타벅스개수로 둔다. 우선적으로 k=1 로 설정한뒤 도출된 결과는 다음과 같다.

- 예측변수 : 거래금액(만원), 전용면적(㎡)
- 결과변수 : 스타벅스개수
- k=1

Confusion Matrix

Actual₩Predicted	0	1	2	3	4	6
0	2140	42	44	14	3	9
1	45	2588	109	43	3	8
2	5	60	2776	19	2	5
3	3	20	72	845	4	2
4	7	13	18	18	593	7
6	4	7	28	5	2	346

Error Report

Class	# Cases	# Errors	% Error
0	2252	112	4.973357016
1	2796	208	7.439198856
2	2867	91	3.174049529
3	946	101	10.67653277
4	656	63	9.603658537
6	392	46	11.73469388
Overall	9909	621	6.267029973

Metrics

Metric	Value
Accuracy (#correct)	9288
Accuracy (%correct)	93.73297003

훈련세트의 오분류율은 6.27% 정도이며

Confusion Matrix

ActualWPredicted	0	1	2	3	4	6
0	279	150	85	24	8	18
1	121	380	137	39	16	18
2	69	116	432	43	19	15
3	20	28	52	136	9	12
4	17	18	31	27	70	4
6	8	10	23	5	4	34

Error Report

Class	# Cases	# Errors	% Error
0	564	285	50.53191489
1	711	331	46.55414909
2	694	262	37.75216138
3	257	121	47.08171206
4	167	97	58.08383234
6	84	50	59.52380952
Overall	2477	1146	46.26564392

Metrics

Metric	Value
Accuracy (#correct)	1331
Accuracy (%correct)	53.73435608

검증세트의 오분류율은 46.27% 수준이다.

뒤이어 동일한 조건에서 k=2 일때 오분류율을 확인한다.

- 예측변수 : 거래금액(만원), 전용면적(㎡)
- 결과변수 : 스타벅스개수
- k=2

Confusion Matrix

ActualWPredicted	0	1	2	3	4	6
0	1716	94	302	79	7	54
1	471	1644	480	130	16	55
2	18	102	2686	31	15	15
3	9	30	183	705	15	4
4	43	83	125	68	319	18
6	5	9	84	23	6	265

Error Report

Class	# Cases	# Errors	% Error
0	2252	536	23.80106572
1	2796	1152	41.20171674
2	2867	181	6.313219393
3	946	241	25.4756871
4	656	337	51.37195122
6	392	127	32.39795918
Overall	9909	2574	25.9763851

Metrics

Metric	Value
Accuracy (#correct)	7335
Accuracy (%correct)	74.0236149

훈련세트 오분류율은 약 26% 정도로 k=1 일때보다 증가했으며

Confusion Matrix

Actual₩Predicted	0	1	2	3	4	6
0	292	85	132	29	8	18
1	149	272	204	54	9	23
2	61	67	493	46	10	17
3	14	18	74	133	3	15
4	22	18	56	27	44	0
6	7	5	34	5	2	31

Error Report

Class	# Cases	# Errors	% Error
0	564	272	48.22695035
1	711	439	61.7440225
2	694	201	28.96253602
3	257	124	48.24902724
4	167	123	73.65269461
6	84	53	63.0952381
Overall	2477	1212	48.93015745

Metrics

Metric	Value
Accuracy (#correct)	1265
Accuracy (%correct)	51.06984255

검증세트 오분류율은 약 49%까지 상승했다. 스타벅스개수가 분류의 기준으로 작용될 때 최적의 k 를 살펴보자. 역시나 k=10 에서 Search Log 를 사용하면

Search Log

K	% Misclassification
1	46.50787243
2	49.33387162
3	48.68792895
4	49.61647154
5	50.62575696
6	50.90835688
7	50.42389988
8	50.78724263
9	51.79652806
10	51.15058539

Note: Scoring will be done using K=1

k=1 일때 오분류율이 최소가 된다.

8. 신경망 분석 기법

인공신경망으로도 불리는 신경망 모형은 분류와 예측을 위해 사용되는 모형으로, 뇌의 뉴런들이 상호작용하고 경험을 통해 배우는 생물학적 활동을 모형화한 것이다. 인간의 학습과 기억의 특성을 닮았고, 특정사건으로부터 일반화하는 능력 또한 갖고 있다. 신경망을 통해서 높은 예측 성과를 달성할 수 있고, 입력변수와 출력변수 사이의 매우 복잡한 관계도 파악할 수 있다는 장점이 있다.

실제 우리의 데이터를 이용하여 신경망 분석을 진행해보기로 했다. 상관관계가 높았던 변수들 중, '전용면적'과 '대형마트 개수' 변수를 골라 정규화를 진행하였다. 그리고 Training Set 을 80%, Validation Set 를 20%로 설정하여 Partition 을 해주었다. 이후 신경망 분석을 위해 Selected Vatiables 에는 전용면적과 대형마트 개수의 정규화된 값을 선택했고, Output Variable 에는 거래금액의 정규화된 값을 선택했다. 하지만 Version limit violated: #Distinct Classes<=30 창과 함께 신경망 분석은 실행되지 않았다.

고민 끝에 랜덤으로 데이터를 30 개 추출하여 신경망 분석을 진행하기로 했다. 엑셀의 RAND() 함수를 이용하여 데이터 30 개를 앞에서부터 얻었다. 데이터가 크지 않았기에 파티션을 나누지 않았다.

Architecture Search Error Log

NetID	# Hidden Layers	# Neurons (Layer	# Neurons (Layer	Training # Errors	Training % Error
Net 1	1	1	0	29	96.6667
Net 2	1	2	0	29	96.6667
Net 3	1	3	0	29	96.6667
Net 4	1	4	0	28	93.3333
Net 5	1	5	0	29	96.6667
Net 6	1	6	0	30	100
Net 7	1	7	0	28	93.3333
Net 8	1	8	0	29	96.6667
Net 9	1	9	0	28	93.3333
Net 10	1	10	0	28	93.3333
Net 11	1	11	0	28	93.3333
Net 12	1	12	0	28	93.3333
Net 13	1	13	0	29	96.6667

Automatic Neural Network 실행 결과, Net 4 이후부터 에러가 다시 높아졌다. 그래서 Hidden Layers 를 1, Neuron 을 4 로 설정하기로 했다.

Neuron Weights

Neuron Weights: Input Layer - Hidden Layer 1			
Neuron	전용면적 정규화	대형마트 개수 정규화	Bias
Neuron 1	-0.149295719	0.450110018	-0.026462
Neuron 2	-1.634843753	0.819535496	0.47862
Neuron 3	0.671781545	0.92962339	0.919004
Neuron 4	-0.363975751	1.09066027	0.912498

Neuron Weights: Hidden Layer 1 - Output Layer					
Neuron	Neuron 1	Neuron 2	Neuron	Neuron	Bias
5300	-0.411523455	-0.043251236	-1.549442	0.061609	-0.615
10000	0.002127352	-0.11127338	-0.854608	-0.07209	-0.9327
11000	-0.504726745	0.355490033	-0.860425	-0.742158	-0.6742
14950	-0.205401657	-0.413529347	-0.860002	-0.534706	-0.5813
19500	-0.900945453	0.197927996	-0.497697	-0.740955	-0.6492
19700	-0.265380318	-0.233157524	-0.904141	-0.300834	-0.6934
20800	-0.269494401	-0.834053758	-0.614487	-0.117578	-0.6741
23000	0.479185961	-0.130161191	-1.151625	-0.453821	-0.7817
25700	0.548357085	-0.218103487	-0.697346	-0.488234	-0.9419
28600	-0.216143795	-0.738575472	-0.803424	-0.696294	-0.4481
29900	-0.027465306	-1.073942699	-0.840993	0.083348	-0.6467
30500	-0.383852606	-0.433932553	-1.218637	-0.279085	-0.3791
30800	-0.555307342	-1.366669817	0.478769	-0.437689	-0.7044
37500	0.532394149	-0.303979296	-0.279743	-0.724903	-0.979
38000	-1.59663998	-1.52443736	0.460663	-0.112118	-0.4942
40500	0.160010919	-0.955442174	-0.142204	-0.42055	-0.817
43700	-0.709546058	-0.568632821	-0.959705	-0.125828	-0.5112
44900	-0.487980245	-1.064438927	0.117311	-0.257578	-0.762
45000	-0.237412688	-0.840294136	-0.109233	-0.591394	-0.6841
48000	-0.763173383	0.418692258	-0.47412	-0.390452	-0.8061
49000	-0.190956076	0.144991829	-0.122835	-0.819075	-0.9497
55000	0.214372805	-1.128283224	-0.645777	-0.873634	-0.4373
60000	0.175254889	-0.457632664	-0.893671	-0.193605	-0.7921
64500	-0.02804222	-0.171994209	-0.502097	-0.360855	-0.9187
70000	-0.741715008	-0.665972217	-0.263846	-0.465619	-0.5969
77000	-0.434917855	0.160449493	0.286391	-1.155788	-0.9174
82500	-1.175097195	0.666105958	-0.882982	-0.662408	-0.6162
155000	0.646534242	0.071062702	-0.044496	-1.710995	-0.8384

Manual Neural Network 실행 결과, 다음과 같은 Neuron Weights 결과를 얻을 수 있다.

Training Log

Epoch	Training: Network Error (Cross Entrop	Training: Data Error (Misclassificatio
Epoch 1	19.49081117	0.966666667
Epoch 2	18.59787155	0.966666667
Epoch 3	17.57900735	0.966666667
Epoch 4	16.53987421	0.966666667
Epoch 5	15.536676	0.966666667
Epoch 6	14.59567998	0.966666667
Epoch 7	13.72634577	0.966666667
Epoch 8	12.9294091	0.966666667
Epoch 9	12.20151901	0.966666667
Epoch 10	11.53772638	0.966666667
Epoch 11	10.93272168	0.966666667
Epoch 12	10.38138753	0.966666667
Epoch 13	9.878999502	0.966666667
Epoch 14	9.421261349	0.966666667
Epoch 15	9.004273241	0.966666667
Epoch 16	8.624481329	0.966666667
Epoch 17	8.278629946	0.966666667
Epoch 18	7.963723588	1
Epoch 19	7.676999238	0.933333333
Epoch 20	7.415906928	0.933333333
Epoch 21	7.178095861	0.933333333
Epoch 22	6.961403704	0.933333333
Epoch 23	6.763847301	0.933333333
Epoch 24	6.583613725	0.933333333
Epoch 25	6.419051064	0.933333333
Epoch 26	6.26865876	0.933333333
Epoch 27	6.131077517	0.933333333
Epoch 28	6.005078964	0.933333333
Epoch 29	5.889555253	0.933333333
Epoch 30	5.783508813	0.933333333

Epoch 30 까지의 결과이다. Error 는 0.93 보다 더 작아지지 않았다.

Error Report			
Class	# Cases	# Errors	% Error
10000	1	1	100
11000	1	1	100
14950	1	1	100
155000	1	1	100
19500	1	1	100
19700	1	1	100
20800	1	1	100
23000	1	1	100
25700	1	1	100
28600	1	1	100
29900	1	1	100
30500	2	2	100
30800	1	1	100
37500	1	1	100
38000	1	1	100
40500	1	1	100
43700	1	1	100
44900	1	1	100
45000	1	1	100
48000	2	0	0
49000	1	1	100
5300	1	1	100
55000	1	1	100
60000	1	1	100
64500	1	1	100
70000	1	1	100
77000	1	1	100
82500	1	1	100
Overall	30	28	93.33333333

Error Report 에서도 Error 가 93.3%임을 확인할 수 있다. 매우 높은 Error 를 보이고 있음을 알 수 있다. 데이터의 크기가 너무 작았기 때문에, 집 값 예측을 위한 인공신경망의 성능이 좋지 못한 것이라고 사료된다.

9. 최종 변수 선정 및 부동산 예측

*집값 = 서브웨이*6557.16 + 계약년*6307.7 + 지하철역개수*6260.39 +*
*대형마트개수*1962.6 + 층*1004.68 + 계약월*970.26 +*
*동*239.59 + 전용면적*194.06 + 건축년도*182.56 +*
*공원면적*0.02 - 학원개수*39.03 - 스타벅스개수*2336.20 -*
*집유형*2897.88 - 배스킨*3827.96*

이와 같이 다중선형회귀 분석을 통해 나온 결론은 부동산 예측 결과다.

10. 결론 및 제언

10.1 한계점

본 프로젝트를 통해 집값 예측을 위한 변수를 선정하고, 실제 유의한 영향을 미치는 변수를 선별하며 집값 예측을 위한 다중선형 회귀모델을 발굴했다는 점에서 유의미하다. 그러나 결론적으로 네 가지 정도의 한계를 보이며 이는 이후 보완하고자 한다.

첫째는 유의미한 변수 활용의 어려움이다. 여러 데이터 사이트와 스크래핑을 통해서 유의미한 변수를 구했으나 우리가 구하고 싶었던 *'고양시 법정동별 인구수', '등록차량수', '소득수준', '지하철역까지의 거리'*와 같은 데이터들이 공개되지 않거나 존재하지 않아서 분석에 사용하지 못한 것이 아쉽다.

둘째는 거리 데이터 확보의 어려움이 있었다. 사용된 데이터의 변수들은 모두 정확한 수치이지만, 주택에서 유명 프랜차이즈까지의 거리, 지하철역까지의 거리 등 정확한 수치데이터를 사용했더라면 더 좋은 성능을 낼 수 있을 것이다. 부동산 좌표 정보가 존재하면 이를 지도 API 와 연결해 거리 정보를 도출해내는 코드 및 결과를 얻을 수 있다.

셋째는 집값 결정 요인이 매우 다양하다는 것이다. 우리는 나름의 요소를 선정해 집값 결정요인을 살펴보았다. 준비한 변수들도 물론 집값에 영향을 미치는 중요한 요인들이지만, 집값을 결정하는 요인들은 실제 더 다양하다. 그 요인들 중 정책과 같이 정권이 바뀜에 따라 영향을 주는 변수는 데이터를 구할 수 없기도 하고 분석에 반영하기가 어렵다. 특히 집값에 영향을 미치는 것으로 대표적인 변수 중 하나는 금리다. 시간이 지남에 따라 데이터가 상이한데, 미숙한 분석 능력으로는 금리 변화를 반영하지 못했다.

마지막은 인공신경망 예측 활용의 한계였다. 인공신경망의 경우 엑셀 데이터마이닝에서는 Version limit violated: #Distinct Classes<=30 오류가 발생했고, 이로 인해 전체 데이터를 사용하지 못했다. 랜덤 함수를 이용하여 30 개의 데이터만 추출해 사용하다 보니 정확도가 떨어지다 못해 무려 93.3%의 오차율을 보여주었다. 전체 데이터 사용이 어려웠던 점이 인공신경망을 만드는 데에 있어 하나의 한계로 작용하였다.

10.2 결론

고양시는 앞서 설정했던 **'살기 좋은 도시'**의 환경 세 가지를 가지고 있었다. '살기 좋은 도시일수록 집값이 높다'는 가설의 타당성을 검증하기 위해 세 가지 분석 기법을 사용했다. 다중선형회귀분석에 따르면 주택 가격을 예측하는 변수가 많을수록 더 좋은 성능을 띄었다. 또한 최상위모델로 선정된 모형에서 도출한 집값 예측 식은 다음과 같았다.

집값 =
*서브웨이*6557.16+계약년*6307.7+지하철역개수*6260.39+대형마트개수*1962.6+층*1004.68+계약월*970.26+동*239.59+전용면적*194.06+건축년도*182.56+공원면적*0.02-학원개수*39.03-스타벅스개수*2336.20-집유형*2897.88-배스킨*3827.96*

집값에 가장 큰 양의 영향을 주는 것은 **서브웨이**였고, 가장 큰 음의 영향을 주는 것은 **배스킨(배스킨라빈스)**이다. **스타벅스** 개수 또한 -2336.20 의 높은 값을 가져 **고양시 집값에는 프랜차이즈 점포 수가 큰 영향을 주는 것을 알 수 있었다.** 집값이 올라가는데 영향을 주는 변수는 서브웨이, 계약 년도, 지하철 역 개수, 대형마트 개수와 같았다.

적어도 활용한 변수 안에서는 주택 내부 환경과 변수보다 주거지를 둘러싸고 있는 외부 환경이 집값 결정에 큰 영향을 미친다는 인사이트를 발견했다. 이는 귀납적으로 '살기 좋은 도시일수록 집값이 높다'라는 가설을 채택에 힘을 실을 수 있다.

11. 데이터 출처

한국부동산원_아파트 거래규모별 거래현황(연도별,면적)

https://www.data.go.kr/data/15068102/fileData.do

근린공원현황:고양시청

http://www.goyang.go.kr/www/www03/www03_9/www03_9_3.jsp

경기도 고양시_도서관 자료현황

https://www.data.go.kr/data/3078218/fileData.do

경기도_병원 현황

https://www.data.go.kr/data/15058086/openapi.do

소상공인시장진흥공단_상가정보

https://www.data.go.kr/data/15083033/fileData.do

경기도 고양시_반려동물등록현황

https://www.data.go.kr/data/15084285/fileData.do