

|  |
| --- |
| **팀 프로젝트 보고서** |
| **집값 예측** |

텍스트, 로고, 상징, 폰트이(가) 표시된 사진

자동 생성된 설명

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **제출일** | **2024. 05. 28** |  | **전공** | **융합시스템공학과** |
| **과목** | **융합시스템창의설계** |  | **팀명** | **5조** |
| **담당교수** | **민현정 교수님** |  | **팀원** | **김범수,이연주,하현성** |

목차

1. 서론

1) 동기 및 문제 설명

2) 참고문헌 및 관련 내용 기술

2. 본론

1) 데이터 수집

2) 데이터 전처리

3) 데이터 분석

4) 문제 및 해결 방법

3. 결과

1) 실험 결과 (중간 보고용)

4. 토의 및 스케줄

1) 토의

2) 학습 내용

3) 스케줄

5. 참고문헌

서론

동기 : 동탄의 집값은 거품이 많이 끼었다고 평가 받으며 시세의 등락도 자주 변동 된다. 집을 사야할 떄 이 위치의 이정도 평수면 얼마가 적정가이고? 집값에 실제 얼마나 영향일 미치는지 알아보고자 주제를 선정하였다.

문제 설명 : 우선 우리조는 처음부터 머신러닝을 통해 집값을 예측하고자 목표를 수립하였다. 집값 예측과 관련된 예제들을 찾아보니 해외에 예시 코드들이 있었으나 데이터셋의 칼럼이 한국의 집 옵션들과는 맞지 않는 문제가 있었다.

본론

1. 데이터 수집

- 아파트 매매 실거래 상세 api

- 국토교통부\_건축물대장정보 서비스

2가지 데이터를 수집하였으나 실제 데이터 열람 결과 “국토교통부\_건축물대장정보 서비스”는 불필요한 데이터로 판단되어 사용하지 않았다. (아파트 매매 실거래 상세 api와 주요 데이터가 겹치고, 집 구매시 주로 고려할 옵션 데이터들이 없음)

수집 과정에서 화성시의 데이터가 수집되지 않아, 강남구로 대체하여 진행중이다.

(다른 지역구도 안 되는 경우가 있어 공공데이터 포털에 문의중)

아파트 매매 실거래 상세 api에서 실제 수집한 코드는 아래와 같다.



노란색 표기한 데이터를 실제로 수집하였으며, 나머지 데이터들은 주로 주소 관련된 데이터라 ‘법정동’ 하나만 주소의 정보로 가져왔다.

코드는 길이 문제로 링크를 별첨하겠다.

<https://github.com/Beomsudev/data_analysis_programming/blob/master/dataCrawling.py>

수집 중 문제 상황은 아래와 같았으며 해결방법도 같이 기술하였다.

1) 처음에 xml데이터가 파징이 되어 있지 않았다.

> xml.etree.ElementTree를 사용하여 파징 코드를 추가하였다.

2) 목표로 했던 동탄의 데이터가 추출되지 않았다.

> 다른 지역구도 안 되는 곳이 많아서 공공데이터 포털에 문의 중이다.

3) api 과부하를 예전에 경험한 적이 있었다.

> 한 달 단위로 가져오면서 1초의 딜레이를 주어 과부하가 걸리지 않게 하였다.

2. 데이터 전처리

- 해외의 데이터 분석 예시를 보니 집값에 영향을 미치는 데이터는 평수, 위치, 건축년도, 지하실, 욕실, 부지 로 평수, 위치, 건축년도는 한국의 요인과 맞지만 나머지는 다르다고 판단하였다. 고로 한국 아파트 데이터중 평수, 건축년도, 위치, 층을 중요 요인으로 판단했고, 분석을 통해 실제 연관이 있는지 확인해 보았다. 원본데이터의 전처리 과정은 아래와 같다.

1) 결측치는 다행히 없어 별도의 처리는 없었다.

2) 거래금액이 “265,000”처럼 회계 표시되어 쉼표를 제거하고 float로 형변환 하였다.

3) 올해인 2024 – “건축년도” 하여 “년식”이라는 신규 열을 추가하였다.

4) 필요없는 (거래)월, (거래)일, 건축년도는 drop 하였다.

5) 법정동, 아파트는 범주형 자료이기에 encoding 과정이 필요하여 Target Encoding을 하였다. (후에 다른 encoding도 적용하여 비교 예정)

코드 링크

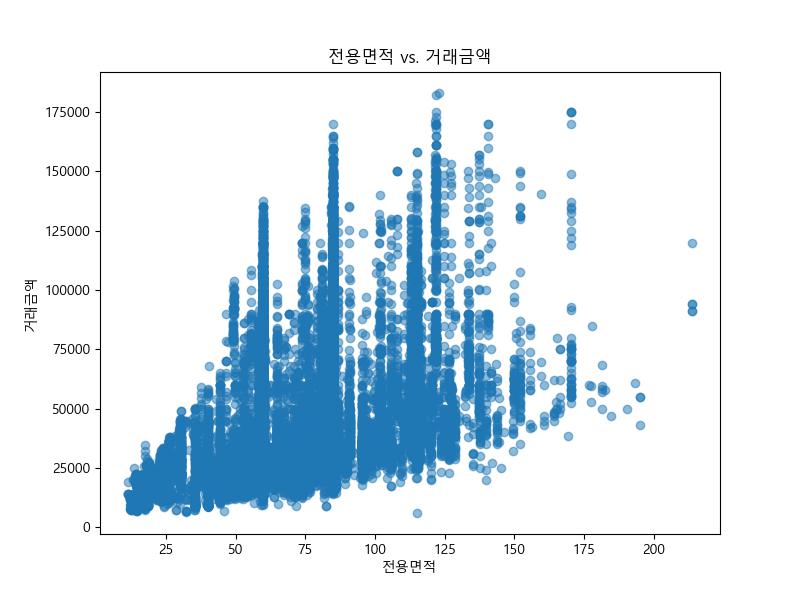
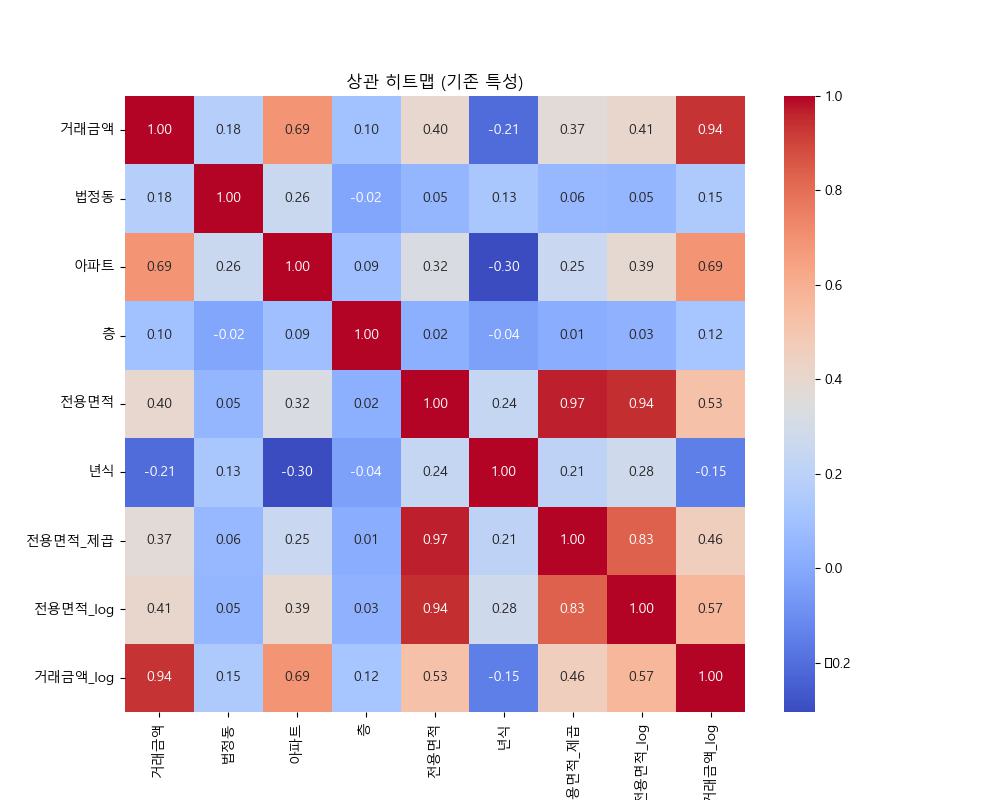
<https://github.com/Beomsudev/data_analysis_programming/blob/master/dataPreprocessing.py>

3. 데이터 시각화

우선 가장 중요할 것으로 예측되는 전용면적을 산점도로 출력하여 데이터를 확인하였다.

그후 히트맵으로 전체데이터를 표현하여 확인하였다.

현재는 확인 과정이라 시각화까지 전처리 코드에 기입하였다.



3. 결과

1) 중간보고용 실험결과

|  |  |  |  |
| --- | --- | --- | --- |
|  | 가격에 미치는 영향 예상 | 실제 결과 | 추후 방향 |
| 년식 | 구축 아파트의 가격도 높기에 상관계수가 그리 높지는 않을 것으로 예측 | +0.21로 많은 영향을 미치지는 못함 | 데이터의 추가 수집으로 계수가 올라가는지 확인 필요 |
| 법정동 | 위치에 따른 가격의 차이가 심하기에 상관계수가 높을 것으로 예측 | +0.18로 예상 보다 많이 낮게 나옴 | 현재 강남구만 데이터를 수집하여 다양한 지역의 데이터 수집필요 |
| 아파트 | 한국은 아파트 시공사의 Name Value가 중요하기에 높을 것으로 예측 | +0.67로 예상 대로 높게 나옴 | 데이터의 추가 수집으로 계수가 올라가는지 확인 필요 |
| 전용면적 | 크기이므로 높은 영향도 예측 | +0.4로 예상 보다는 낮게 나옴 | 예상 보다 낮아 추가 데이터 확보와 특성변환이 필요 |
| 층 | 가격에 큰 영향은 없으나 유의미 할것으로 예측 | +0.1로 예상 수치 만큼 나옴 | 추가 데이터 확보시 계수가 올라가는지 확인 필요 |

4. 토의 및 스케줄

1) 결과에 따른 멤버들 각자의 의견

김범수 : 상관계수가 낮게 나오는 데이터들은 다양한 특성변환을 통해 상관계수가 올라가는지 확인하여 실제로 유의미한 데이터인지 확인해야 합니다.

이연주 : 강남구의 데이터만 확보하여 데이터의 지역성이 다양하지 않으므로 추가 적인 데이터 확보가 필요합니다.

하현성 : 더 다양한 비교 분석을 통해 데이터에 따른 알맞은 시각화 방법이 있는지 검토해 보아야 합니다.

2) 프로젝트 진행 과정

*아래 부분 간단하게만 써놔서 늘여서 기입해주세요~*

*토론을 통해 주제를 정하고 각자 역할을 나누어 맡은 부분을 담당하였다. 아래와 같이 역할을 분담 하여 진행하였다.*

*- 데이터 크롤링 >> 연주*

*- 데이터 분석 >> 범수*

*- 지리분석 >> 현성*

*1차 분석을 통해 위 실험결과의 표와 같은 분석을 하게 되었고 추후에 아래와 같은 추가 작업을 진행 할 예정이다.*

*토론 결과 앞으로 추가할 작업은 아래와 같다.*

* *상관계수치 높이기 (log화 등)*
* *범주형 데이터 encoding 다양하게 적용해 보기*
* *수업시간에 배운 folium으로 지리 정보를 시각화 하기*
* *더 많은 가격 데이터 확보(공공데이터 포탈 회신 필요)*

3) 학습내용

*연주님 각자 맡은 코드 부분에 맞춰서 학습 내용 대강 적어주세요~*

*- 데이터 크롤링 >> 연주 (코드는 다 짜여져 있어서 추가로 할 작업은 없을 것 같아요.)*

*- 데이터 분석 >> 범수 (회귀 분석 추가, 상관계수치 높이는 작업)*

*- 지리분석 >> 현성 (\* 집 값을 지도에 folium 써서 시각화 > 기존에 부탁드린 UI 부분이 필요없어져서.. 이 부분을 추후에 작업 부탁드립니다.)*

4) 스케줄

|  |  |  |  |
| --- | --- | --- | --- |
| **주차** | **기간** | **내용** | **비고** |
| 1주차 | 2024.05.01 ~ 2024.05.07 | 주제 및 데이터 선정 |  |
| 2주차 | 2024.05.08 ~ 2024.05.14 | 데이터 분석 및 처리 코드 개발 |  |
| 3주차 | 2024.05.15 ~ 2024.05.21 | 시각화 코드 개발 |  |
| 4주차 | 2024.05.22 ~ 2024.05.28 | 코드 개발 및 중간 보고서 작성 | 중간 보고서 (5.28) |
| 5주차 | 2024.05.29 ~ 2024.06.04 | 검토 및 디버깅 |  |
| 6주차 | 2024.06.05 ~ 2024.06.11 | 발표 자료 준비 | 최종 발표 (6.11) |
| 7주차 | 2024.06.12 ~ 2024.06.18 | 코드 정리 및 최종 보고서 작성 | 최종 보고서 (6.18) |

5. 참조 문헌

1) Kaggle 데이터 분석 예시 <https://www.kaggle.com/code/samahfetouh/house-price-prediction>

2) *책 대강 찾아서 추가하기 데이터 분석 책*