다양한 요소가 집값에 미치는 영향 분석

202244073 김지연

1. 데이터 수집 및 저장

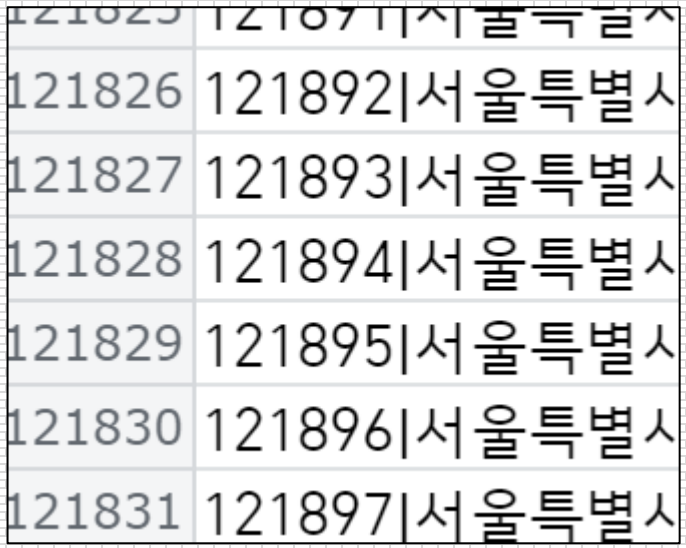
부동산 빅데이터 플랫폼에서 무료 데이터 구매.

- 공동주택 전월세 가격 대중교통 인프라 연계 정보(20241H)
- 공동주택 전월세 가격 대형 유통시설 입지 연계 정보(20241H)
- 공동주택 전월세 가격 교육시설 입지 연계 정보(20241H)
- 공동주택 전월세 가격 공원녹지공간 연계 정보(20241H)
- 초등학교 통학구역 배정 공동주택 전월세가격 정보(20241H)
- 대형의료시설 정보 및 공동주택 전월세가격 정보(20241H)

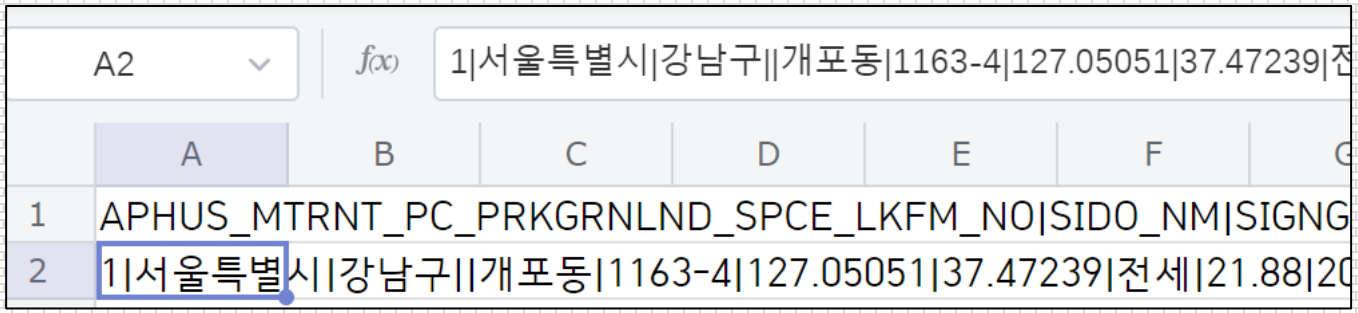
2. 데이터 가공 및 정제

데이터의 문제점:

1. 서울특별시 전체의 데이터로 n만개에 해당되는 정보가 들어있음.
2. 데이터가 구분별로 제대로 정리되어 있지 않음.



▲12만개의 데이터 존재



▲구분이 '|'로만 되어있음

2. 데이터 가공 및 정제

```python
import pandas as pd

# 데이터 정의
data = """
2406076236|서울특별시|서초구||반포동|10|126.99645|37.50389|월세|59.81|20240626|9000|220|10|2000|계성초등학교|126.99675|37.50463|0.08614
2406076237|서울특별시|서초구||반포동|10|126.99645|37.50389|전세|56.16|20240513|84000|0|12|2000|계성초등학교|126.99675|37.50463|0.08614
"""
#이미지상 생략

# 데이터 문자열을 리스트로 변환
rows = [row.split('|') for row in data.strip().split('Wn')]

# 데이터프레임 컬럼 정의
columns = [
    "ID", "시도", "구군", "읍면동", "리", "번지", "경도", "위도", "계약구분", "전용면적",
    "계약일", "보증금", "월세", "층", "건축년도", "초등학교", "초등학교 경도", "초등학교 위도", "초등학교 거리"
]

# 데이터프레임 생성
df = pd.DataFrame(rows, columns=columns)

# 데이터 타입 변환 (숫자 데이터는 float 또는 int로)
convert_columns = ["경도", "위도", "전용면적", "보증금", "월세", "층", "건축년도", "초등학교 경도", "초등학교 위도", "초등학교 거리"]
for col in convert_columns:
    df[col] = pd.to_numeric(df[col], errors='coerce')

# 데이터프레임 출력
print(df.head())

# 파일로 저장 (옵션)
df.to_csv("초등학교재정리본.csv", index=False, encoding="utf-8-sig")
```

▲ 초등학교 통학구역 배정 공동주택 전월세가격 정보(20241H)를 정제하기 위한 코드

서울특별시 서초구 반포동으로 50개의 데이터를 골라 split을 사용하여 칼럼을 나눠 데이터 프레임을 생성

반포동 선택 이유: 서초구가 서울특별시에서 집값이 가장 비싸고 공원의 면적도 가장 커서 둘의 연관관계가 두드러질 것이라는 예상과 함께 선택.

이것을 각각 csv파일로 저장하여 기초적으로 사용할 데이터를 정제

공동부분을 기준으로 데이터들을 합쳐 하나의 csv파일로 만들었다.

2. 데이터 가공 및 정제

```
import pandas as pd
import seaborn as sns

df= pd.read_csv("/content/총정리본.csv")
rdf = df.drop(['ID','시도','구군','리', '녹지공간/공원 거리', '대중교통', '교육시설', '대형유통시설', '대형의료시설', '초등학교'], axis=1)

dfq_mth = rdf.query('계약구분=="월세"')
dfq_mth.to_csv("/content/월세정리본.csv", index=False, encoding="utf-8-sig")

dfq_cht = rdf.query('계약구분=="전세"')
dfq_cht.to_csv("/content/전세정리본.csv", index=False, encoding="utf-8-sig")
```

▲ Drop과 쿼리문을 이용해 삭제 및 정제

합친 파일(총정리본)에서 분석에 불필요한 칼럼을 삭제

- 대중교통 등의 경우 대중교통과 해당 집의 거리만 필요하며 문자열은 집값과의 분석에 불필요하므로 삭제

정리한 파일을 월세와 전세 부분으로 나누어 정리

- 월세와 전세는 보증금과 월세라는 부분에서 큰 차이가 나기 때문에 보다 정확한 분석에 방해가 될 것이라 판단하였다.

3. 데이터 분석

```
import pandas as pd
import seaborn as sns

#전세는 보증금 60000이상 시 비싼 것으로 간주
df= pd.read_csv("/content/전세비쌈정리본.csv")
rdf = df.drop(['계약구분'], axis=1)
rdf.to_csv("/content/최종전세정리본.csv", index=False, encoding="utf-8-sig")

#월세는 보증금 30000이상 혹은 월세 400이상 시 비싼 것으로 간주
df= pd.read_csv("/content/월세비쌈정리본.csv")
rdf = df.drop(['계약구분'], axis=1)
rdf.to_csv("/content/최종월세정리본.csv", index=False, encoding="utf-8-sig")
```

M	N	O
대형의료시	초등학교 거리	expesiveornot
0.885	0.08614	1
2.885	2.08614	1
4.885	4.08614	1
6.885	6.08614	1

▲expensiveornot의 1 요소

나눈 파일에서 각각 집값이 비싼 것인가 아닌가를 판단하는 칼럼을 추가해 분석

전세의 경우 보증금 60000(천 원)이상일 시에 비싼 것으로 간주
월세의 경우 보증금 30000(천 원)이상 혹은 월세 400(만 원)이상일 시에 비싼 것으로 간주

boolean형식으로 추가하여 1과 0으로 판단-> 1을 만드는 데에 큰 기여를 한 요소를 찾아내는 방향으로 진행

3. 데이터 분석

```
import pandas as pd
import seaborn as sns

df= pd.read_csv("/content/최종전세정리본.csv")
rdf = df.drop(['번지', '계약일', '건축년도', '보증금', '월세'], axis=1)
rdf.to_csv("/content/최최종전세정리본.csv", index=False, encoding="utf-8-sig")

df= pd.read_csv("/content/최종월세정리본.csv")
rdf = df.drop(['번지', '계약일' ,'건축년도', '보증금', '월세'], axis=1)
rdf.to_csv("/content/최최종월세정리본.csv", index=False, encoding="utf-8-sig")
```

시각화 및 추가 분석 전 불필요한 칼럼을 제거.

시각화에서는 어떤 요소가 집값에 큰 영향을 미쳤는지를 보여주는 것이므로, 보증금과 월세 칼럼도 삭제.

계약일과 건축년도 또한 날짜 형식이 아닌 정수형으로 포함되어 시각화와 추가 분석(코랩)에 혼돈을 줌으로 삭제.

3-4. 데이터 분석 및 시각화

```
import seaborn as sns
%matplotlib inline
import pandas as pd
from xgboost import XGBClassifier
from sklearn.preprocessing import LabelEncoder
import plotly.express as px


df = pd.read_csv('/content/최최종전세정리본.csv')

X = df.iloc[:,0:7]

y = df.iloc[:,7]
y_labels = LabelEncoder().fit_transform(y)

xgb_cls = XGBClassifier(n_estimators=1000)

xgb_cls.fit(X, y_labels)

feature_series = pd.Series(data=xgb_cls.feature_importances_, index=X.columns )

sns.barplot(x= feature_series, y=feature_series.index)

# plotly로 그래프 생성
fig = px.bar(feature_series, x=feature_series.values, y=feature_series.index, orientation='h', title="반포구 전세 집값의 주요 요소")
fig.write_html("/content/전세_feature_importance.html")
```

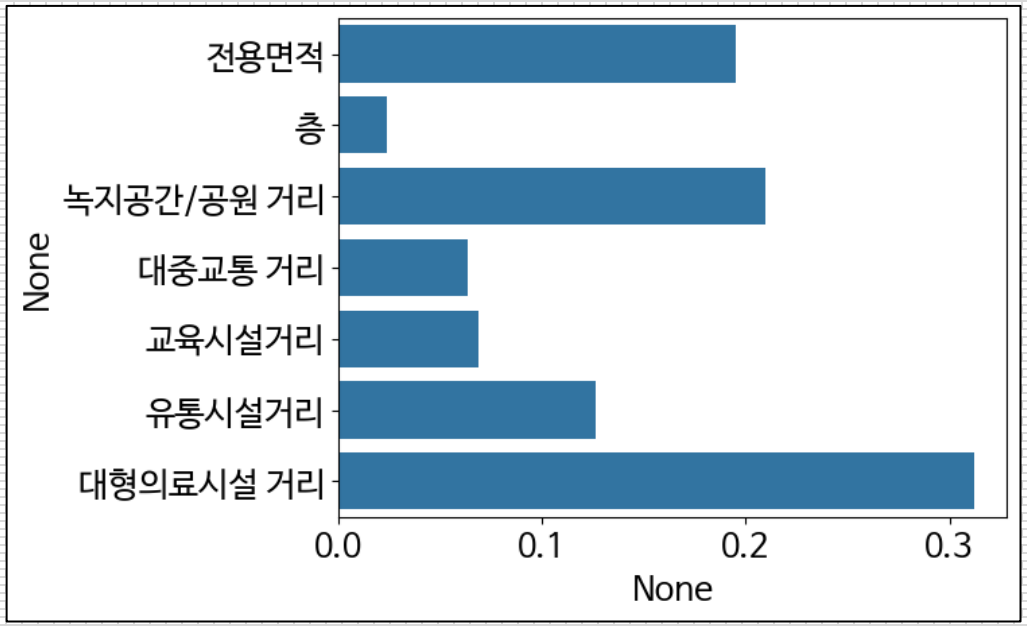
▲ barplot을 이용해 그래프 만들기.
Classifier를 이용해 사전 분류를 진행하였다.

집값이 비싼 것인가 아닌 것인가가 판단되었으므로, 그것에 기여한 정도를 표현해주는 그래프 생성.

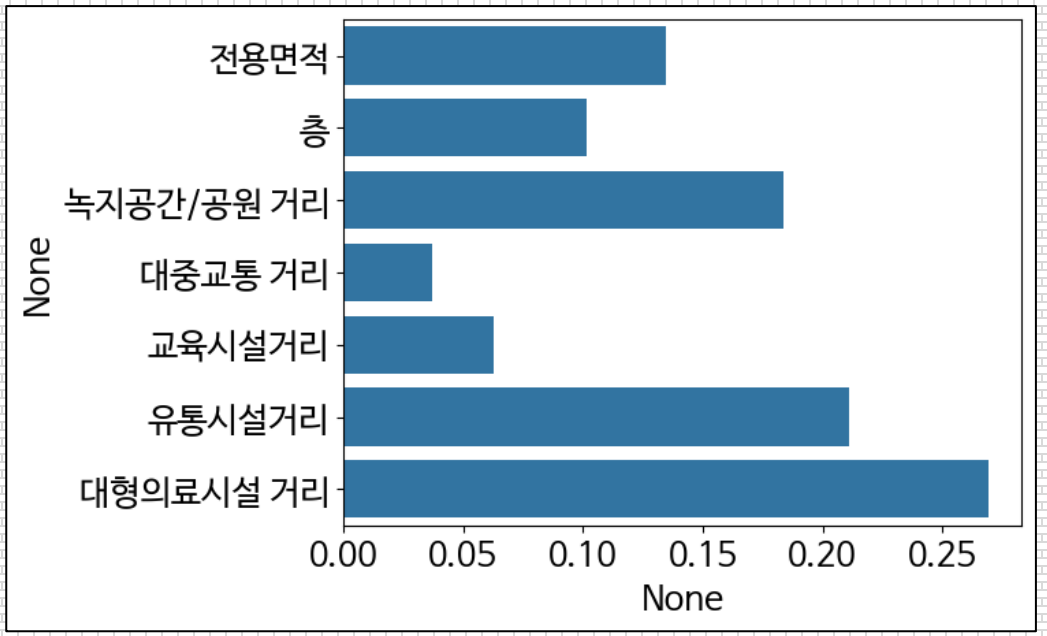
Feature-importance로 각 요소가 집값에 얼마나 영향을 미쳤는지 확인 가능.

전세와 월세의 분석 코드는 동일하며 각 그래프는 html형식으로도 생성되어 유동적인 그래프의 확인이 가능하다.

3-4. 데이터 분석 및 시각화



▲ 전세



▲ 월세

각각 월세와 전세에 기여한 정도를 나타낸 그래프

단, 숫자가 큰 것이 많이 기여했음을 나타내므로 전용면적과 층을 제외한 칼럼들은 반대로 생각해야 한다.
각각 집과의 거리를 나타내는데, 집과 해당 요소들이 가까울수록 집값이 높아지는데 기여하기 때문이다.

따라서 전세와 월세 모두 전용면적, 즉 집의 평수(제곱미터 기준)가 가장 많은 영향을 미친다고 볼 수 있다.

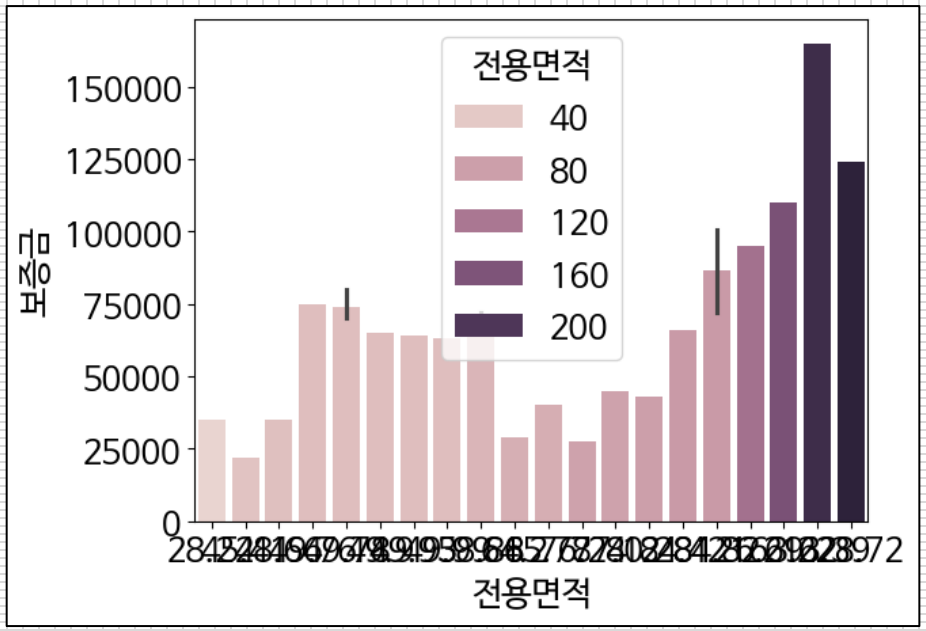
3-4. 데이터 분석 및 시각화

```
import pandas as pd
import seaborn as sns
import plotly.express as px
import matplotlib.pyplot as plt


df = pd.read_csv('/content/최종전세정리본.csv')
df1 = df.drop(['번지', '계약일', '건축년도'], axis=1)
#cdf= df.countplot(data=df, x='보증금', hue='전용면적')
#print(cdf)
#fig = px.bar(df, x='전용면적', y='월세', color='전용면적', color_continuous_scale=px.colors.diverging.Spectral)
#fig.update_layout(width=600)

sns.barplot(data=df, x='전용면적', y='보증금', hue='전용면적')

# plotly로 그래프 생성
fig = px.bar(df, x='전용면적', y='보증금', orientation='h', title="전세와 전용면적의 연관성")
fig.write_html("/content/전세_전용면적.html")
```

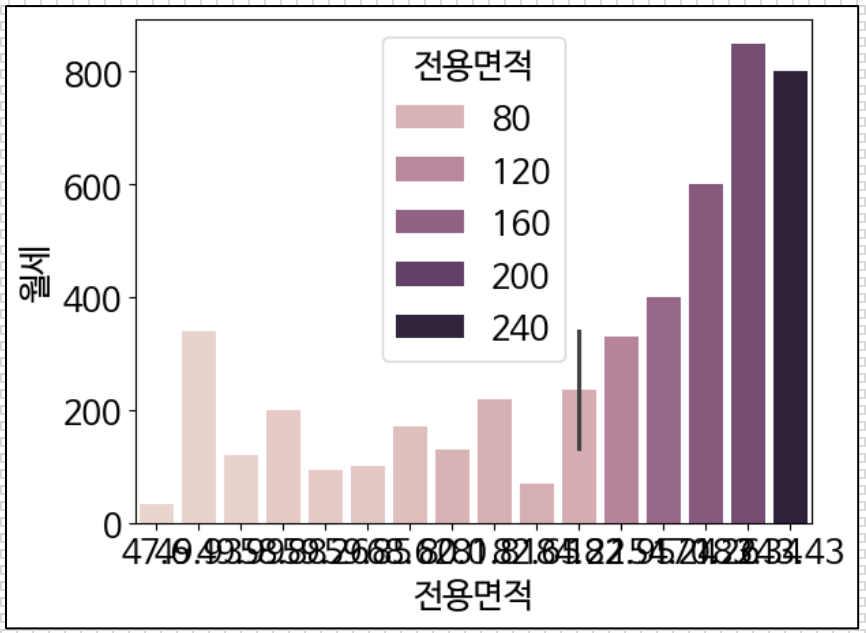
전용면적이 집값에 가장 큰 영향을 미친다는 것을 알아으므로, 집값과 전용면적의 연관성도 그래프로 알아보았다.

Snsbarplot을 사용하여, 각 요소들이 증가할 때 어떤 모습이 되는지를 확인하고자 하였다. 해당 코드 역시 전세와 월세의 코드는 동일하며, html로도 생성되었다.

3-4. 데이터 분석 및 시각화



▲ 전세



▲ 월세

두 그래프 모두 전용면적이 증가할수록 집값이 증가한다는 것을 확인할 수 있다.

전용면적이 증가할 때 무조건적으로 집값이 증가하는 것은 아니나, 일부일 뿐이며 대부분의 수치가 증가하고 있다.

5. 분석 결과

1. 서울특별시 서초구 반포동의 집값에 가장 큰 영향을 미치는 것은 전용면적, 즉 집의 평수이다.

2. 전용면적이 가장 큰 영향을 미치고 있으므로, 전용면적이 증가할 때 집값이 증가한다.