서울 강남구 집값 예측 추이

201844093 임현섭

1. 데이터 수집

DATA 공공데이터포털 .GO.KR

Sheet | Open API

필드명: 자치구명

검색명: 강남

내려받기(CSV)

접수연도	자치구코드	자치구명	법정동코드	법정동명	
2023	11680	강남구	11800	도곡동	1
2023	11680	강남구	10300	개포동	1
2023	11680	강남구	10100	역삼동	1
2023	11680	강남구	10300	개포동	1

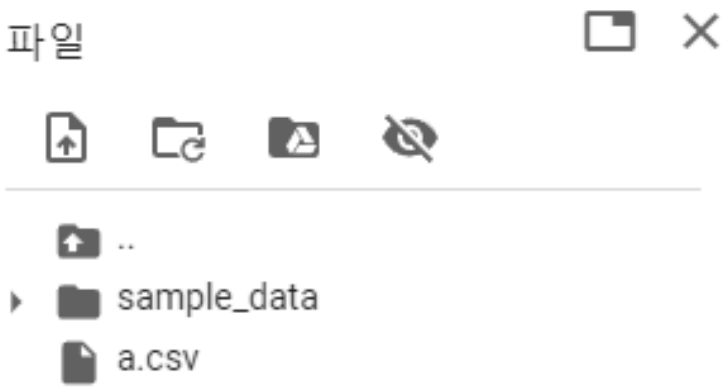
공공데이터포털에서 서울 강남구 부동산의 실거래가를 csv파일로 다운합니다.

1. 데이터 수집

```
# 단계 1: 폰트 설치
import matplotlib.font_manager as fm
import matplotlib.pyplot as plt

!apt-get -qq -y install fonts-nanum > /dev/null
#fontpath = '/usr/share/fonts/truetype/nanum/NanumBarunGothic.ttf'

#font = fm.FontProperties(fname=fontpath, size=9)

#fm._rebuild()



fe = fm.FontEntry(
    fname=r'/usr/share/fonts/truetype/nanum/NanumBarunGothic.ttf', # ttf 파일이 저장되어 있는 경로
    name='NanumGothic')                            # 이 폰트의 원하는 이름 설정
fm.fontManager.ttflist.insert(0, fe)               # Matplotlib에 폰트 추가
plt.rcParams.update({'font.size': 18, 'font.family': 'NanumGothic'}) # 폰트 설치
```

시각화를 이용한 그래프의 한글이 깨짐으로써 폰트를 설치합니다.

1. 데이터 수집




```python
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression

# 데이터 불러오기
file_name = '/content/a.csv'
df = pd.read_csv(file_name, encoding='cp949')
```

공공데이터에서 받아온 csv파일을 a.csv파일로 지정하고 colab에 파일을 업로드하여 데이터를 불러옵니다.

Read_csv 를 하면서 utf-8 오류가 남으로써 encoding을 해줍니다.

2. 데이터 전처리

```
# 결측값 대체
df.fillna(df.mean(), inplace=True)  # 결측값을 평균값으로 대체
```

위의 코드를 통해 결측값을 평균값으로 대체합니다.

3. 변환 과정

```
# 사용할 특성 선택
X = df[['건물면적(㎡)', '토지면적(㎡)']]
y = df['물건금액(만원)']

# 선형 회귀 모델 초기화 및 학습 (2020년부터 2023년까지의 데이터로 학습)
model = LinearRegression()
model.fit(X[df['접수연도'] <= 2023], y[df['접수연도'] <= 2023])
```

집값 예측을 위해 필요한 건물면적과 토지면적을 특성으로 지정합니다.

집값 예측에 메인 타겟인 물건금액(만원) 즉, 집값을 y변수에 지정합니다.

LinearRegression() 으로 모델을 초기화 하고, fit을 이용하여
2020년 ~ 2023년까지의 데이터를 사용하여 모델을 학습시킵니다.

3. 변환 과정

```
# 시각화를 위한 데이터 생성
years = range(2020, 2026)  # 2020년부터 2025년까지 연도 범위
predicted_values = []  # 연도별 예측값을 담을 리스트

# 2020년부터 2023년까지의 예측값 계산
for year in years:
    # 예측을 위한 데이터 구성
    data = df[df['접수연도'] == year][['건물면적(㎡)', '토지면적(㎡)']]
    if not data.empty:
        predicted_value = model.predict(data).mean()
        predicted_values.append(predicted_value)
    else:
        # 해당 연도의 데이터가 없을 경우, 2000년부터 2023년까지의 학습 데이터로 예측
        predicted_value = model.predict(X[df['접수연도'] <= 2023]).mean()
        predicted_values.append(predicted_value)
```

2020년 ~ 2023년 까지의 데이터를 학습하여 2025년까지의 집값을 예측하기 위해 연도범위를 설정합니다.

건물면적과 토지면적을 이용하여 2020년 ~ 2023년은 평균값으로 구성하고

그 이후 데이터들은 학습된 데이터로 예측을 합니다.

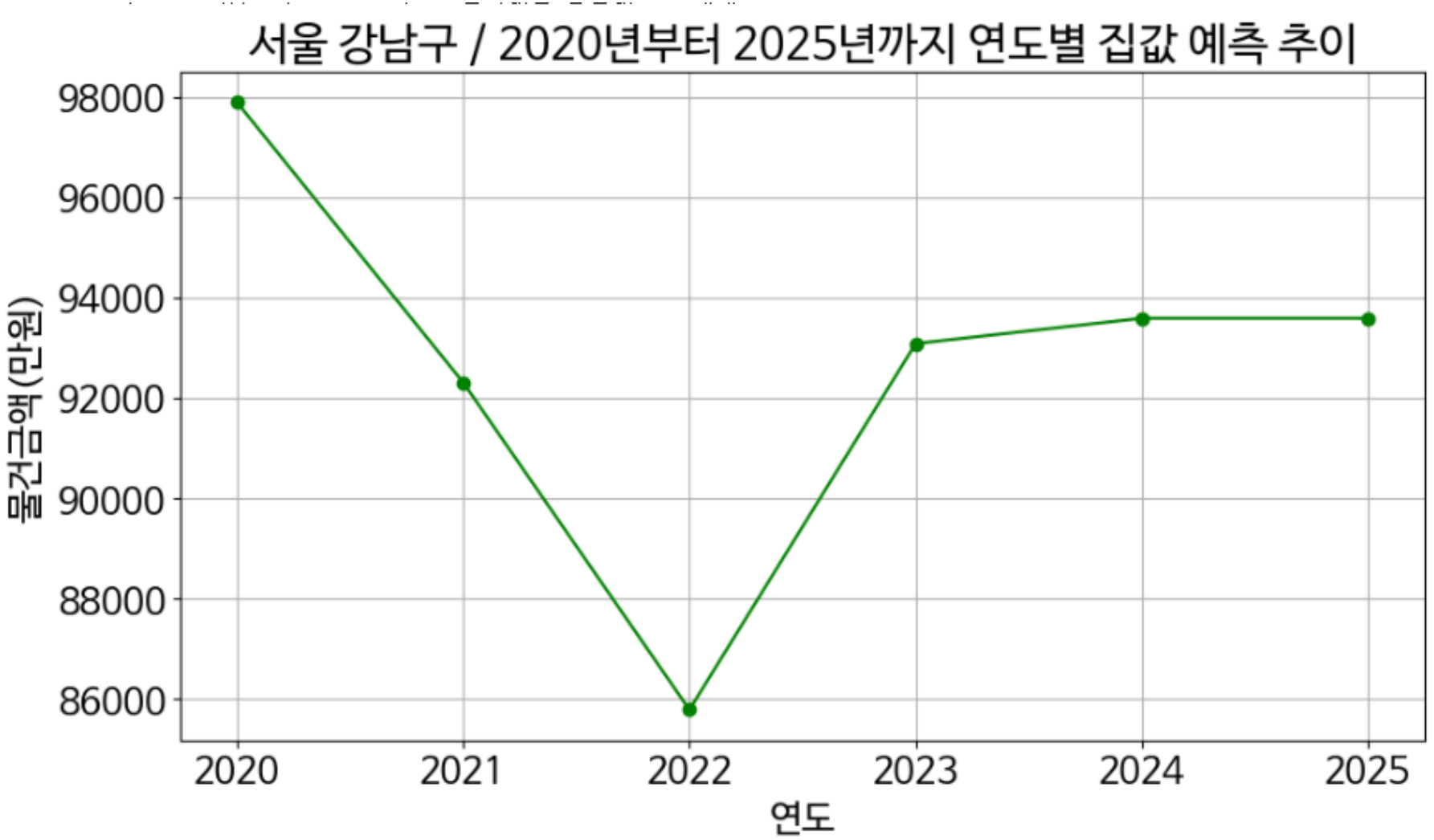
4. 시각화

```
# 선 그래프로 시각화
plt.figure(figsize=(10, 6))
plt.plot(years, predicted_values, marker='o', linestyle='-', color='green')
plt.xlabel('연도')
plt.ylabel('물건금액(만원)')
plt.title('서울 강남구 / 2020년부터 2025년까지 연도별 집값 예측 추이')
plt.grid(True)
plt.tight_layout()
plt.show()
```

X축을 연도별, y축을 집값으로 설정하여 2020년 ~ 2023년까지의 집값 과

그 이후 2025년 까지의 예측된 집값을 그래프로 시각화 합니다.

4. 시각화


서울 강남구 / 2020년부터 2025년까지 연도별 집값 예측 추이
98000
96000
94000
92000
90000
88000
86000
물건금액(만원)
2020
2021
2022
2023
2024
2025
연도

5. 결과

이 모델은 입력된 건물 면적과 토지 면적에 기반하여 물건금액(집값)의 변동을 예측하고, 이를 연도별로 시각화하여 집값의 추이를 보여줍니다.
또한, 2020년부터 2023년까지의 데이터를 학습하여 2023년 이후의 집값을 예측하는 능력을 갖추고 있습니다.
이 모델을 통해 미래 집값의 추이를 예측할 수 있지만, 모델의 정확성은 사용된 데이터와 모델의 복잡성에 영향을 받을 것입니다.