Real Estate Data Cleaning with Python and Excel

Project Description

I scraped data from a Bangladeshi real estate website for web scraping practice. The data was raw and messy, requiring extraction and cleaning to make it structured and ready for analysis. I used two methods to clean the data: one with Python's pandas, NumPy, and regular expression libraries, and the other with Excel Power Query.

Solution

To address this problem, the following steps were undertaken among many others:

1. **Data Import and Transformation**: Imported the dataset into a pandas DataFrame and Power Query, then performed the necessary transformations.

2. **Data Format Analysis**: Reviewed the provided data.

3. **Exploratory Data Analysis (EDA):** Checked data types, identified null values and duplicates, and removed records with mostly null values.

4. **Extraction and Cleaning Functions:** Created functions for extracting and cleaning data based on patterns.

5. **Save the Cleaned Data**: Save the cleaned and structured data into a CSV file.

Challenges

The most challenging part was cleaning the 'price' column, which contained various formats such as actual amounts, values in lakhs, crores, and per square feet. Initially, I tried using built-in functions from various Python libraries, but this was not entirely successful. In the second attempt, I used functions and regular expressions, but it still didn't work perfectly. Finally, in the third attempt, I combined built-in functions with custom functions and regular expressions, which successfully cleaned the data.

Limitations

1. **Floor Column**: Could not figure out how to clean ‘floor’ column of the dataset.

2. **Manual Adjustments**: Some manual transformations were required to achieve the desired cleaning in when using Excel Power Query of the ‘price’ column.

Files

1. **real_estate_raw_data.csv**: Raw file.
2. **real_estate_cleaning.ipynb**: Python cleaning script.
3. **output_final.csv**: Cleaned and structured CSV file generated by the Python script.
4. **Real Estate Cleaned Data.xlsx**: Cleaned and structured Excel file generated by Power Query.