Project Proposal: Geographical and Housing Market Insights in New York

1. Project Overview

The objective of this project is to derive insights into real estate pricing based on the house type, property square footage, neighbourhood characteristics and geographical location that influence the housing market in reference to New York, USA.

Key goals include:

- Acquisition of data using APIs and scraping techniques.
- Storing the data systematically in NoSQL database (MongoDB)
- Integration of several datasets for comprehensive analysis.
- Ensuring data quality and profiling to prepare for advanced analytics and visualizations.

2. Data Acquisition

Data acquisition is based on APIs and web scraping.

- **Kaggle API**: Dataset based on the house price, house type, broker titles, number of bedrooms and bathrooms, property square footage, address, state, administrative and local areas, street names and geographical coordinates for New York City will be assessed through the Kaggle API.
  Dataset :- https://www.kaggle.com/datasets/nelgiriyewithana/new-york-housing-market

- **OpenStreetMap (OSM) API**: To extract the specific amenities like schools, hospitals and public transport stations and stop locations in New York, Overpass API, an OpenStreetMap (OSM) query tool along with requests library is used to retrieve data.

- **Web Scraping**: To get insights further about the neighbourhood of the listed house locations, data is extracted from this website: https://www.walkscore.com/score/. Extracted data contains information on walk score, transit score and bike score which corresponds to neighbourhood of the houses that listed in the dataset obtain from Kaggle platform. Walk score is a metric that measures the walkability of a specific location which evaluates how convenient to access everyday amenities. Transit score evaluates the availability and usefulness of public transportation. Bike score measures how bike-friendly the area is.

  Data is scrapped from similar web pages corresponds to neighbourhood such as:
  https://www.walkscore.com/score/astoria-new-york-11102
  https://www.walkscore.com/score/astoria-new-york-11105
  https://www.walkscore.com/NY/New_York/Bedford-Stuyvesant etc.

3. Data Storage

To handle the diversity and scale of data, a NoSQL database (MongoDB) is selected for storage. Following are some queries that we intend to analyse to gain a comprehensive understanding of the housing market in New York.

- Unveiling the geographical patterns in New York's housing market by identifying high and low-value areas in real estate.
- Impact of property features such as space, number of rooms vs price.
- Variations in New York housing by neighbourhood and its features.

4. Data Integration/ Data Enrichment

- **Datasets to Integrate**:
  - Data scraped from various web pages regarding walk score, transit score and bike score integrated into a one dataset.
  - Extracted data from web scraping technique and the dataset obtained from the Kaggle platform is merged based on the neighbourhood location name.
  - Geographic data fetched from OpenStreetMap (OSM), including coordinates of the amenities like schools, hospitals and public transport stations and stop locations are merged into one dataset.
- **Data Enrichment:**
  - Enrich housing data with additional attributes, such as walk score, transit score and bike score.
  - Integrated demographic location of amenities, such as schools, hospitals, public stations and public stop positions for feature creation of proximity from the listed housing location to the nearest schools, hospitals, public stations and public stop positions.

5. Data Profiling/ Data Quality

Ensuring the quality and reliability of data is critical for meaningful analysis. Steps that are going to be include:

- **Data Profiling**
  - Analysing the distributions, missing values and outliers present in the data,
  - Identifying duplicate entries.
  - Cross-referencing geographical data to ensure accuracy of the locations and boundaries.
- **Data Quality**
  - Handling missing values.
  - Standardizing location formats for uniformity in mapping locations.
  - Removing outliers based on statistical thresholds.

6. Deliverables

- **Unified Dataset**: A fully integrated MongoDB database containing housing, neighbourhood and geographical data for New York City.
- **Data science pipeline**: Documentation of the pipeline (data acquisition, storage, integration, profiling and analysis) including method used, challenges faced, and solutions implemented.
- **Validation Insights**: Metrics and visualizations to demonstrate the reliability and consistency of the data.