Project Proposal: Analyzing Real Estate Market Trends and Price Prediction

Data Description:

The dataset for this project is the "New York Housing Market" dataset sourced from Kaggle. It contains prices of New York houses, providing valuable insights into the real estate market in the region. It includes information such as broker titles, house types, prices, number of bedrooms and bathrooms, property square footage, addresses, state, administrative and local areas, street names, and geographical coordinates.

Motivation:
Scrutinizing this data is critical to understanding the dynamics of one of the most complex and risky real estate markets in the world. Analyzing the New York City real estate market reveals patterns and factors that affect property values, helping stakeholders make more informed decisions.

Problem Statement:

1. What are the most important factors affecting house prices?
2. Can we develop an accurate model that predicts home prices based on property characteristics?
3. Are there unique clusters of properties based on characteristics and pricing that reveal potential market segments?

Techniques:

a) **Clustering Analysis**: Using K-Means clustering and hierarchical clustering, I aim to identify clusters of properties with similar characteristics and price ranges. I will consider features such as location coordinates (latitude and longitude), property type and physical attributes.
b) **Regression analysis:** apply multivariate linear regression to predict property sales prices based on independent variables such as property type, size, number of units, age of building and location details.

Potential Obstacles:

Data quality issues: There are missing values, especially for sales prices and property attributes. Entries with zero or very low sales prices may represent non-independent transactions or errors.

Outliers: Extreme variations in real estate prices, high value transactions may distort the analysis.