Executive Summary – NYC Real Estate Analysis

Project Goal:

The objective behind this project is to build a platform which can help users find recommendations of property based on the specifications provided by them. The project also helps in depicting the houses on a geographical map of the USA. It shows the houses labeled as clusters categories which in turn helps us examine the geographical difference between the clusters found.

Data Source and Schema:

This dataset is a record of every building or building unit (apartment, etc.) sold in the New York City property market over a 12-month period. This dataset contains the location, address, type, sale price, and sale date of building units sold. The dataset has 84549 instances, we will be using 80% as training and 20% as testing data. The dataset consists of 22 fields. The dataset depicts records of every building, apartments etc. sold in New York City.

Project Modules:

We broke down our project into below modules which are described in detail in the team project report.

1. Importing Libraries
2. Data Preprocessing
3. Data Visualization
4. Clustering
5. Cluster Analysis
6. Classification -Random Forest
7. Visualizations and Geographical Plots
8. Recommendation of the best match

Methods Used:

1) Unsupervised Learning - Clustering
2) Supervised Learning -Random Forest Classification
3) Visualization and Geographical Plotting – GeoPandas

Conclusion:

After clustering we could conclude that cluster 0 (red) has the highest values from approximately 0.9-1.0 (has 50,000 values), and other clusters are not well defined by sales price, but we hypothesize that there is a geographical significance in separating cluster 1 and cluster 2. We can also observe that cluster 0 has most values and we can consider them as the high-priced houses and cluster 1 has low prices and cluster 2 has medium price range. We conclude that k means gives us clusters with high prices as a separate cluster namely cluster 0 and the other clusters are separated by geographical significance and residential units. Residential unit significance is known from the important feature bar graph of Random Forest classifier.

The Recommend functionality helped us give the closest significant match based on Gross Square Feet and Cluster value. It gave us significant result i.e. SALES PRICE, ADDRESS and NEIGHBORHOOD corresponding to the given input i.e., Gross Square Feet and Cluster value these values were given as recommendation to the users.