Michael Elrod

Jake Brulato

Beau Brown

Gage Peffer

**DSBA 6211 Project Proposal: New York Housing Market**

Project Objectives

* Predict market price of New York houses from property type, property measurements, and location data.
* Exploratory analysis of data and past trends and what were the most significant factors in those specific types of properties.
* Data Visualization on the overall groupings of houses and what they were like depending on that specific region our data is from.
* Analysis on the broker, predict what the pricing structure is based on from their previous listings.
* After Python visualizations, if there is time we will add some Tableau visualizations to showcase some geospatial data base on coordinates.

Data Set

* Dataset: [New York Housing Market](https://www.kaggle.com/datasets/nelgiriyewithana/new-york-housing-market)
* This dataset has prices of houses in New York as well as attributes about the layout of the house and location of the house. The main target variable will be the price of the house.
* There are 17 total columns including the “Price” column with 4802 observations that have multiple levels for most of the columns (ex. Beds, Bathrooms, Type).
* A unique thing is the geospatial data, allowing us to get a view of the surrounding houses.

Proposed Methods

* Linear regression- use the continuous numerical variables available to predict housing price. This will mostly use data like the property dimension and number of bedrooms and bathrooms.
* Random Forest - building a Random Forest Algorithm that will build a series of decision trees to predict price, and select the optimal tree from the produced trees. Specifying a root node that would have the best information gain, then determining the best subset of data with our best tree. If the model becomes too complex, we will implement early stopping or post-pruning methods to create the best model.
* SVM - Using a support vector machine we could either predict price or look at specific houses based on the Types column. By using feature selection, we could specify what would be most likely to influence the pricing and look at the model output. If we think we can make the model better, we will tune the parameters until we find what the best fit would be.
* K-means (Distance, Regression) - By using a K-means clustering algorithm, we can find distinct housing groupings based on the size of the house or the pricing that is similar to others. Using this data, we can create sharp regression models for the model significance to find the best support vector regression model.
* K-Nearest Neighbors - Using feature selection for housing types and normalizing the data, we can specify the neighbors and how they will be classified based on pricing.
* **Price analysis:** Analyze the distribution of house prices to understand market trends and identify potential investment opportunities.
* **Property size analysis:** Explore the relationship between property square footage and prices to assess the value of different-sized houses.
* **Location-based analysis:** Investigate geographical patterns to identify areas with higher or lower property prices.
* **Bedroom and bathroom trends:** Analyze the impact of the number of bedrooms and bathrooms on house prices.
* **Broker performance analysis:** Evaluate the influence of different brokers on the pricing of houses.

This dataset 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.

* **BROKERTITLE**: *Title of the broker*
* **TYPE**: *Type of the house*
* **PRICE**: *Price of the house*
* **BEDS**: *Number of bedrooms*
* **BATH**: *Number of bathrooms*
* **PROPERTYSQFT**: *Square footage of the property*
* **ADDRESS**: *Full address of the house*
* **STATE**: *State of the house*
* **MAIN\_ADDRESS**: *Main address information*
* **ADMINISTRATIVE\_AREA\_LEVEL\_2**: *Administrative area level 2 information*
* **LOCALITY**: *Locality information*
* **SUBLOCALITY**: *Sublocality information*
* **STREET\_NAME**: *Street name*
* **LONG\_NAME**: *Long name*
* **FORMATTED\_ADDRESS**: *Formatted address*
* **LATITUDE**: *Latitude coordinate of the house*
* **LONGITUDE**: *Longitude coordinate of the house*