**Project Title: Predicting Wake County, NC House Prices with Machine Learning**

**Project Overview:**

The real estate market in Wake County, NC is highly competitive, with properties receiving an average of 5 offers and selling in just 37 days. To gain a competitive edge in this market, accurately assessing house prices is crucial. By leveraging historical sale data, property features and economic indicators, we aim to provide accurate and actionable pricing and remodeling recommendations to buyers and sellers.

**Objectives:**

Build a Predictive Model: Develop a machine learning model that can accurately predict house prices in Wake County, North Carolina.

**Data Sources:**

To achieve the objectives of this project, we will utilize the following data sources:

1. Historical Sale Data: Information on house sales in Wake County, NC, including sale prices and sale dates.
2. Property Features: Data on property attributes such as square footage, bathrooms, location, and utilities.
3. Economic Indicators: Information on the economy, such as unemployment rate, population.
4. Market Indicators: Information on real estate market, such as mean days to pending, median days to pending and median sale to list ratio.

**Methodology:**

Data Collection: Collect and consolidate data from public records, real estate databases, US census bureau and other relevant sources.

Data Preprocessing: Clean and preprocess the data to handle missing values, outliers, and inconsistencies.

Feature Selection: Identify and select the most influential features using statistical analysis and domain knowledge.

Model Training: Experiment with various machine learning algorithms, such as linear regression, decision trees, random forests.

Hyperparameter Tuning: Fine-tune the model hyperparameters to optimize performance.

Model Evaluation: Assess MAE, MSE, RMSE and R2 score. Utilize techniques like cross-validation to avoid overfitting.

**Conclusion:**

1. Based on the raw data from Wake County, NC real estate website which has 441329 rows and 87 columns, the dataset has datetime variables, continuous variables, categorical variables, location variables and text (natural language) variables. The target is a continuous variable called total\_sale\_price, a good dependent variable for regression analysis.
2. There are 52 variables with missing value and 27 variables with missing value percentage greater than 70. It will be a good dataset for data wrangling learning.
3. Predicting house prices in Wake County, NC using machine learning is a valuable endeavor in the competitive real estate market. This project aims to develop a machine learning model that can predict house prices in Wake County, North Carolina accurately and provide actionable remodel suggestions. By successfully completing this project, we can contribute to the optimization of property transactions and enhance the real estate experience for buyers and sellers in Wake County, NC.