MGT 6203 Group 4 Project Proposal

TEAM INFORMATION (1 point)

Team # 4:

Team Members:

1. Zilin Ma ; GT ID : 903856403
2. Wenhui Ma; GT ID :903836445
3. Xin Ye; GT ID:903846327
4. Biyao Zhou; GT ID :903856517

OBJECTIVE/PROBLEM (5 points)

Project Title: US Home Price Prediction

Background Information on chosen project topic:

The U.S. real estate market is a multifaceted and pivotal component of the national economy, deeply interconnected with economic well-being and progress. Housing is at the heart of this market, serving as a crucial gauge of financial stability for individuals and communities. Whether someone is a potential homebuyer, a discerning renter, a prudent investor, a real estate developer, a policymaker shaping housing policies, or simply a keen observer of economic trends, understanding the complex nuances of housing dynamics is essential for making wise decisions. It involves grasping evolving housing trends, deciphering price fluctuations, and considering location-specific attributes. Real estate represents a substantial financial investment for both individuals and institutions, demanding meticulous analysis to identify opportunities and assess risks. Policymakers can draw valuable insights from housing market analysis to formulate effective housing policies, addressing issues like affordability. Housing prices and market activity act as strong economic indicators, aiding economists and decision-makers in forecasting and policymaking. Thus, analyzing real estate data fosters market transparency, promoting fairness and efficiency, benefiting industry professionals and the general public alike.

Problem Statement (clear and concise statement explaining purpose of your analysis and investigation):

The purpose of this analysis is to explore the trends and characteristics of the U.S. real estate market, aiming to provide in-depth insights for potential homebuyers, investors, developers,and policymakers. We will focus on variables such as changes in the number of bedrooms, bathrooms, and house size, as well as the impact of geographic location on housing prices. Through this study, our goal is to uncover the key factors of the real estate market, gaining a better understanding of market dynamics.

State your Primary Research Question (RQ):

What are the key variables that contribute to the final sale price of a US home? How can we interpret those relationships? How to build models with the given home features? What are the different variables affecting the home sale price across different states/cities? How can we use these models for prediction in the real world?

Add some possible Supporting Research Questions (2-4 RQs that support problem statement):

1. Try to build a ML model can help predict the final sale price of a given property.
2. Use the ML model to assist in evaluating the feasibility of the listing price.

Business Justification: (Why is this problem interesting to solve from a business viewpoint? Try to quantify the financial, marketing or operational aspects and implications of this problem, as if you were running a company, non-profit organization, city or government that is encountering this problem.)

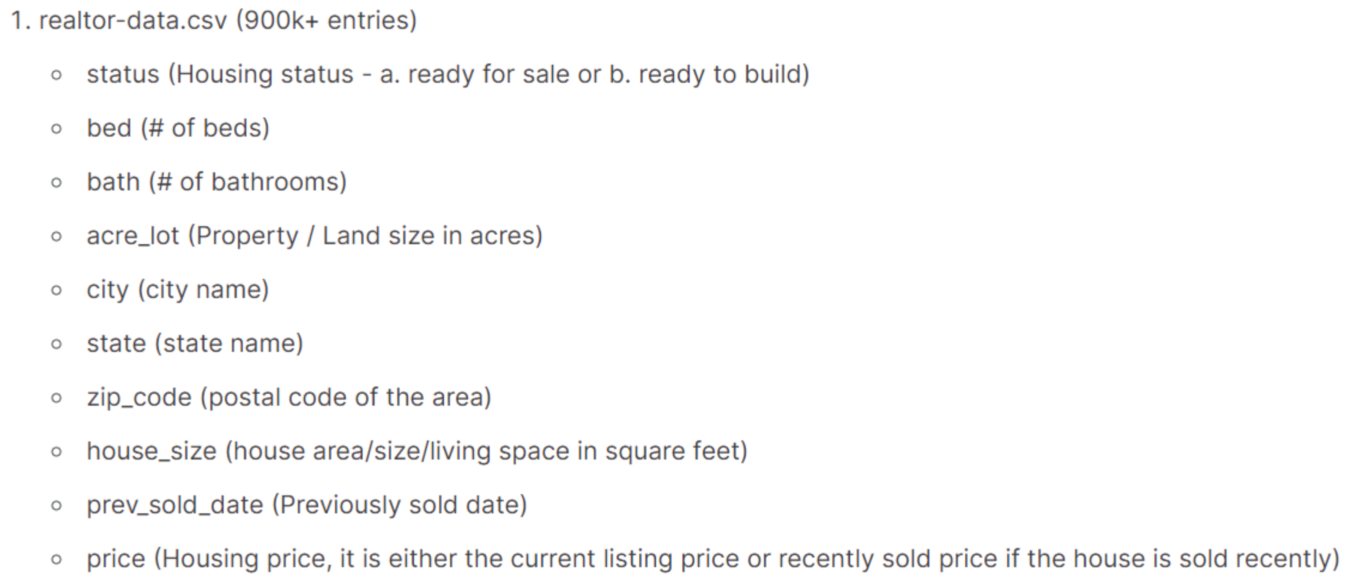
The increasing purchase activities in the house market and rising home price is a signal of the economy going upward. It suggests people have the income as well as the positive outlook for the economy. The housing market condition can pose pressure on interest rate change. It may indirectly influence how the central bank makes decisions and conducts monetary policy. The home price in different regions of a country also demonstrate the income level as well as the living affordability of the region. It can significantly affect private sector investment and how the local government makes decisions on infrastructure spending and fiscal policy(taxation). If we are able to observe the housing market trend of different regions, we can predict and compare the housing market of each region. We may be able to notice whether the past trend is a sustainable and stable growth to the economy. If we find out there is a potential over-valuation in the housing market, it may lead to an economic bubble. Monetary policy may be required as the public intervention. More conservative investment strategies will be more appropriate for private investors.

DATASET/PLAN FOR DATA (4 points)

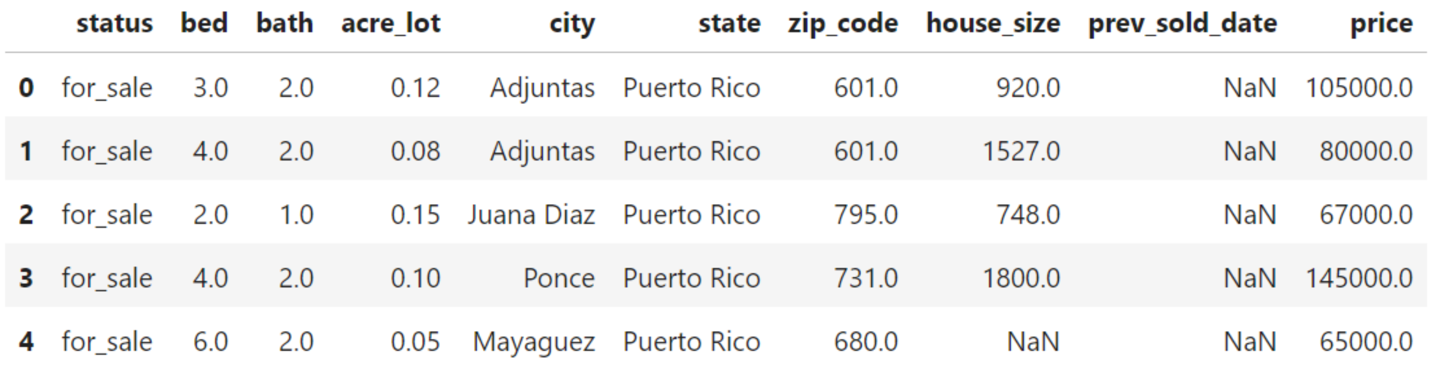
Data Sources (links, attachments, etc.):

https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset/data updated every 2-4 weeks

Data Description (describe each of your data sources, include screenshots of a few rows of data):



1. realtor-data.csv (900k+ entries)
    - status (Housing status - a. ready for sale or b. ready to build)
    - bed (# of beds)
    - bath (# of bathrooms)
    - acre_lot (Property / Land size in acres)
    - city (city name)
    - state (state name)
    - zip_code (postal code of the area)
    - house_size (house area/size/living space in square feet)
    - prev_sold_date (Previously sold date)
    - price (Housing price, it is either the current listing price or recently sold price if the house is sold recently)



	status	bed	bath	acre_lot	city	state	zip_code	house_size	prev_sold_date	price
0	for_sale	3.0	2.0	0.12	Adjuntas	Puerto Rico	601.0	920.0	NaN	105000.0
1	for_sale	4.0	2.0	0.08	Adjuntas	Puerto Rico	601.0	1527.0	NaN	80000.0
2	for_sale	2.0	1.0	0.15	Juana Diaz	Puerto Rico	795.0	748.0	NaN	67000.0
3	for_sale	4.0	2.0	0.10	Ponce	Puerto Rico	731.0	1800.0	NaN	145000.0
4	for_sale	6.0	2.0	0.05	Mayaguez	Puerto Rico	680.0	NaN	NaN	65000.0

Key Variables: (which ones will be considered independent and dependent? Are you going to create new variables? What variables do you hypothesize beforehand to be most important?)

Price is our dependent variable for this model. And all other variables, except prev_sold_date, are our independent variables.

Given we don't have too many variables readily available, we will try to create new variables through feature engineering.

A few variables that are intuitively to be very significant for our model are:

1. City, State. The geo location of the property should be very closely correlated with its sale price. Based on different location, the price would vary much based on economy, supply/demand, etc.
2. House size. The size of the house, assuming it's related to above ground living space. The bigger the living space, the higher the sale price should be.

APPROACH/METHODOLOGY (8 points)

Planned Approach (In paragraph(s), describe the approach you will take and what are the models you will try to use? Mention any data transformations that would need to happen. How do you plan to compare your models? How do you plan to train and optimize your model hyper-parameters?))

Data Preprocessing:

In data preprocessing, we address missing values through imputation or removal. For categorical variables, we transform them into numerical representations using either binary column creation (one-hot encoding) or assigning unique numerical labels (label encoding). Numerical features are then scaled for uniformity using methods like Min-Max scaling or Z-score normalization, optimizing the dataset for U.S. real estate market analysis and modeling.

Exploratory Data Analysis (EDA):

In the exploratory data analysis (EDA) phase, we deeply examine the dataset. This involves visualizing housing price trends over time, across regions, and relevant variables. We also study attribute distributions and correlations, aiming to discover regional patterns and insights into U.S. real estate dynamics.

Feature Engineering:

In the feature engineering stage, we will enhance the dataset by introducing new features like price per square foot, property age, or proximity to amenities, where applicable, to provide additional insights for analysis. Furthermore, we will judiciously curate the feature set by identifying and retaining those attributes that exhibit meaningful correlations with our target variable and align with domain expertise. This process of feature engineering aims to refine the dataset for modeling, enhancing its predictive capabilities and relevance to the U.S. real estate market analysis.

Model Selection：

In the model selection phase, we will explore a range of regression models, including linear regression, random forest regression, gradient boosting regression, or neural networks models, to predict housing prices.

Linear Regression is a valuable and interpretable predictive model when there's an approximate linear relationship in the data. However, it struggles with complex nonlinear patterns. To ensure its validity, we analyze assumptions using techniques like scatterplots, residuals, and tests. If assumptions are violated, adjustments like transformations are needed to enhance model performance. This diagnostic process ensures the reliability of regression analysis for making predictions or conclusions.

Random Forest Regression is a method designed for regression tasks. It comes with numerous advantages, such as resistance to overfitting, efficient handling of large datasets, and the ability to capture complex non-linear relationships with minimal hyperparameter tuning. However, due to its use of multiple decision trees in making predictions, it can be challenging to understand and explain the specific reasons behind individual predictions.

Gradient Boosting Regression is a powerful ensemble learning technique used for regression tasks. It is good at capturing complex relationships in data and has high predictive accuracy, making it a popular choice today when dealing with large data sets and complex patterns. However, it requires careful hyperparameter tuning and can be sensitive to overfitting. While it provides excellent predictive performance, it may sacrifice a degree of model interpretability compared to simpler models such as linear regression or decision trees

*Neural Networks are powerful, versatile, and complex machine learning models ideal for handling diverse tasks like regression and classification, particularly with large datasets. They consist of interconnected layers of neurons that learn through optimization. While they provide strong predictive capabilities, they can be hard to interpret and demand significant computational resources and data for training.

Model Training and Optimization:

In the model training and optimization phase, the dataset is first divided into three subsets: 60% for training, 20% for validation, and 20% for testing using random splitting. Machine learning models are then trained on the training data and assessed for their performance using the validation set. To improve model accuracy and prevent overfitting, hyperparameters are fine-tuned systematically, often utilizing methods like grid search or random search. Moreover, we use regularization techniques to promote generalization, ensuring that the models excel in practical scenarios while remaining reliable and effective.

Cross Validation:

To choose the best model for our real estate analysis, we will use techniques such as k-fold cross validation. This includes dividing the training data into subsets, training the model on some subsets, and evaluating others. We will consider the characteristics of the data set and balance the complexity and interpretability of the model to provide personalized insights, rather than focusing solely on accurate predictions.

Evaluate and Compare Models:

In the evaluation and model comparison phase, we use selected criteria to see how each model performs on the validation set. We use some common regression measures, such as mean absolute error (MAE), mean square error (MSE), or R-MSE, to measure how good the model is. We compare these metrics, hoping to find out which model is best suited to meet the goals and needs of the project. This step guides which model we ultimately choose to ensure that it is best suited to handle the special situations in the analysis of US real estate datasets.

Final Evaluation:

In the final evaluation stage, we choose a model through cross-validation and parameter tuning. Then, we test this selected model with an independent dataset to see how well it adapts to unseen data. This crucial step ensures the model remains effective beyond training and validation. By assessing its performance on the test dataset, we aim to provide a reliable estimate of its real-world effectiveness and reliability for various stakeholders in the U.S. real estate market, including homebuyers, investors, developers, and policymakers.

Anticipated Conclusions/Hypothesis (what results do you expect, how will you approach lead you to determining the final conclusion of your analysis) Note: At the end of the project, you do not have to be correct or have acceptable accuracy, the purpose is to walk us through an analysis that gives the reader insight into the conclusion regarding your objective/problem statement

Hypothesis:

1. Price Trends and Geographic Variation: We expect to identify trends in housing prices over time and geographic regions. Our approach will involve visualizing price trends and exploring whether certain areas experience higher or lower price growth rates.
2. Impact of Housing Characteristics: We hypothesize that factors such as the number of bedrooms, bathrooms, and house size will significantly influence housing prices. Our analysis will include correlation assessments to quantify these relationships.

Ultimately, our approach will lead us to determine the final conclusion of our analysis by considering factors such as model performance, price and regional trends, and housing characteristics. While our hypotheses guide our analysis, the goal is to provide stakeholders with a comprehensive understanding of the key factors and trends shaping the U.S. real estate market, allowing them to make informed decisions. The accuracy of our conclusions will depend on the quality and representativeness of the dataset and the effectiveness of the chosen models and methodologies.

What business decisions will be impacted by the results of your analysis? What could be some benefits?

Governments can employ this tool as a valuable lagging indicator to carefully assess the impact of their existing monetary and fiscal policies. By analyzing historical data and trends, policymakers can gain insights into how their decisions have influenced economic conditions, helping them make more informed choices for the future. For private companies, this tool can be a powerful resource when making strategic investment decisions, especially when contemplating entry into regional markets. Individual home buyers and investors can also benefit significantly from our models. When considering the purchase of a home or making an investment, they can utilize our insights to gain a deeper understanding of the key factors that impact home prices. Armed with this knowledge, they can make more informed budgetary decisions, ultimately increasing their chances of making sound and profitable investments.

PROJECT TIMELINE/PLANNING (2 points)

Project Timeline/Mention key dates you hope to achieve certain milestones by:

Data cleaning: By 10/14/2023

EDA: 10/28/2023

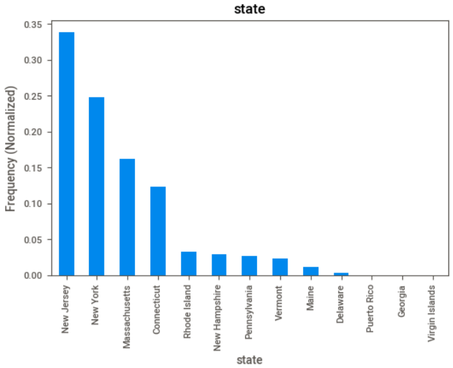
Feature Engineering: 10/28/2023

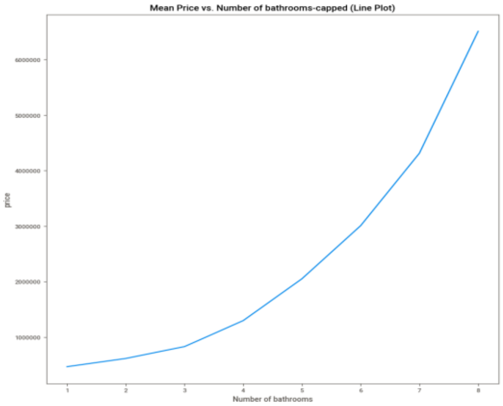
Modeling: 11/11/2023

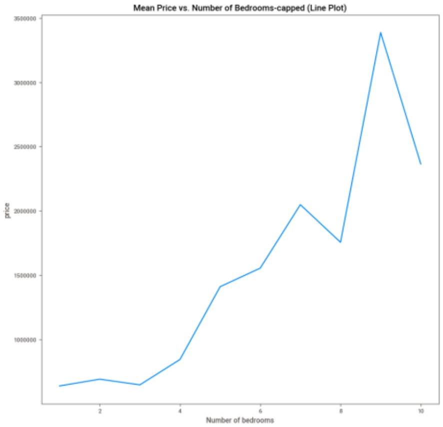
Analysis: 11/ 18/2023

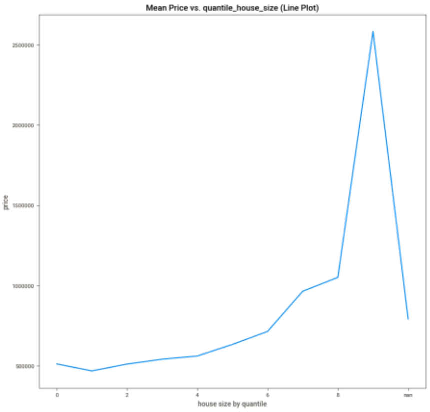
Final report: 11/25/2023

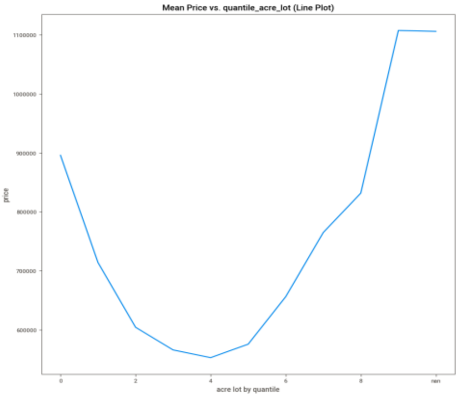
Appendix (any preliminary figures or charts that you would like to include):


state
Frequency (Normalized)
New Jersey
New York
Massachusetts
Connecticut
Rhode Island
New Hampshire
Pennsylvania
Vermont
Maine
Delaware
Puerto Rico
Georgia
Virgin Islands
state


Mean Price vs. Number of bathrooms-capped (Line Plot)
price
Number of bathrooms


Mean Price vs. Number of Bedrooms-capped (Line Plot)
price
Number of bedrooms


Mean Price vs. quantile_house_size (Line Plot)
price
nan
house size by quantile


Mean Price vs. quantile_acre_lot (Line Plot)
price
nan
acre lot by quantile