Team members: No.4
Name (GT ID):
Zilin Ma(zma338)
Wenhui Ma(wma98)
Xin Ye (xye85)
Biyao Zhou (bzhou301)

MGT 6203
Nov.4th 2023

Project Progress Report

Project Title: US Home Price Prediction

Github Repository : https://github.gatech.edu/MGT-6203-Fall-2023-Canvas/Team-4

Problem to tackle: Use past home sold history in the US to come up with a home price prediction model. Features related to the home are used for prediction. Once the model is built, we can try to use the model to evaluate the reasonability of each home's listing price.

Progress of Work:

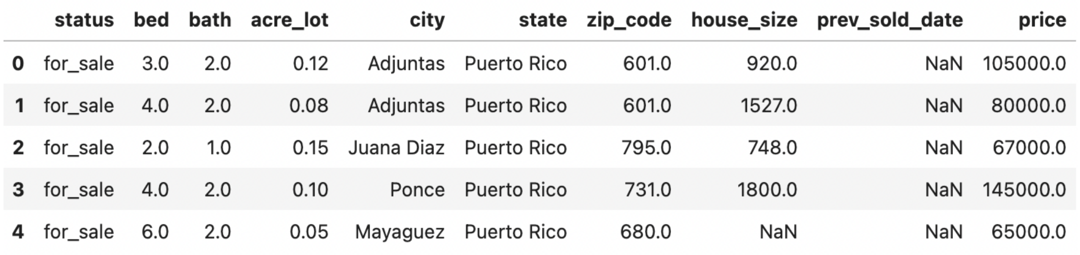
In our project, we've accomplished crucial data cleaning, addressed missing values, delved into data relationships, performed feature engineering, and laid the groundwork for modeling. Here are the specific steps we've taken thus far.

1.Data Cleaning

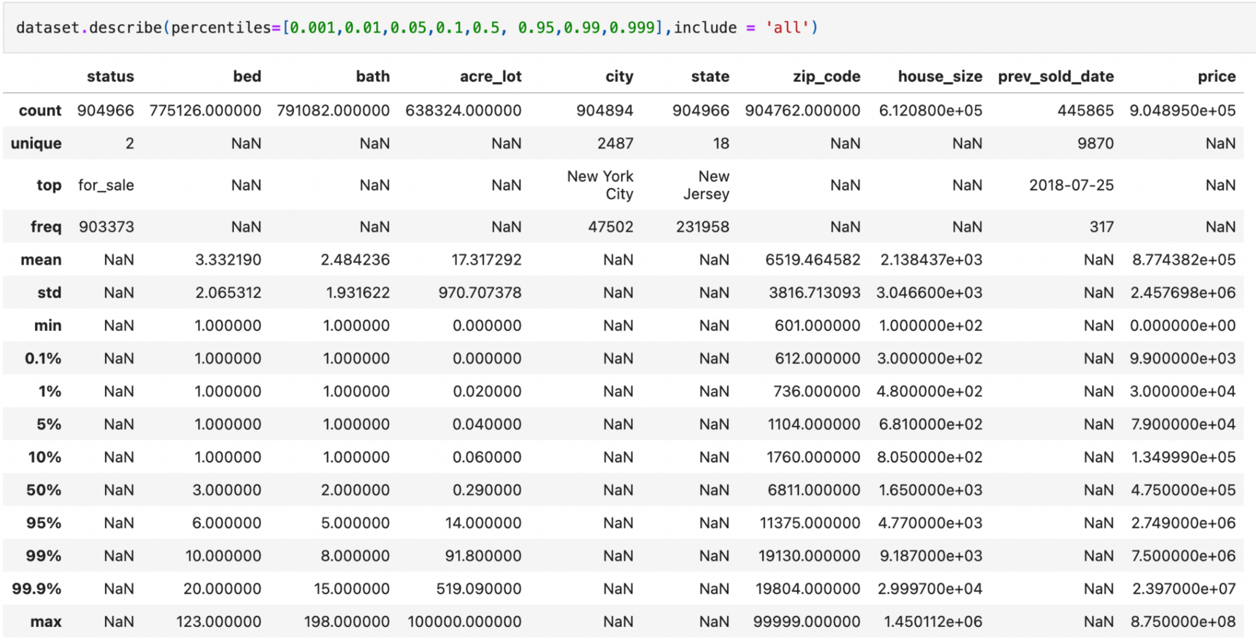
Ensuring the quality and suitability of a dataset for modeling is a pivotal initial step in data science. This involves meticulous exploration and preprocessing of the data. To this end, we undertook the following steps as part of our data cleaning process:

Step1: Initial Setup and Preliminary Exploration of the U.S. Real Estate Dataset

Upon importing the requisite libraries, including 'pandas', 'numpy', 'matplotlib', 'seaborn', and 'sklearn', we configured 'pandas' to display a maximum of 50 rows and columns. This adjustment facilitated an effective visual inspection of the data. Subsequently we utilized the 'head()' function to display the initial few records of the dataset. Additionally, the 'describe()' function was employed to extract comprehensive statistical data, thereby providing an overview of the data distribution and potential anomalies.



	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



```
dataset.describe(percentiles=[0.001,0.01,0.05,0.1,0.5, 0.95,0.99,0.999],include = 'all')
```

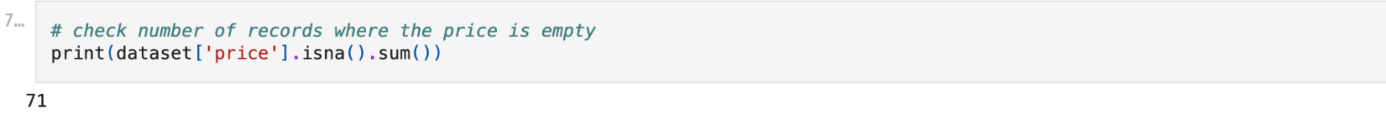
	status	bed	bath	acre_lot	city	state	zip_code	house_size	prev_sold_date	price
count	904966	775126.000000	791082.000000	638324.000000	904894	904966	904762.000000	6.120800e+05	445865	9.048950e+05
unique	2	NaN	NaN	NaN	2487	18	NaN	NaN	9870	NaN
top	for_sale	NaN	NaN	NaN	New York City	New Jersey	NaN	NaN	2018-07-25	NaN
freq	903373	NaN	NaN	NaN	47502	231958	NaN	NaN	317	NaN
mean	NaN	3.332190	2.484236	17.317292	NaN	NaN	6519.464582	2.138437e+03	NaN	8.774382e+05
std	NaN	2.065312	1.931622	970.707378	NaN	NaN	3816.713093	3.046600e+03	NaN	2.457698e+06
min	NaN	1.000000	1.000000	0.000000	NaN	NaN	601.000000	1.000000e+02	NaN	0.000000e+00
0.1%	NaN	1.000000	1.000000	0.000000	NaN	NaN	612.000000	3.000000e+02	NaN	9.900000e+03
1%	NaN	1.000000	1.000000	0.020000	NaN	NaN	736.000000	4.800000e+02	NaN	3.000000e+04
5%	NaN	1.000000	1.000000	0.040000	NaN	NaN	1104.000000	6.810000e+02	NaN	7.900000e+04
10%	NaN	1.000000	1.000000	0.060000	NaN	NaN	1760.000000	8.050000e+02	NaN	1.349990e+05
50%	NaN	3.000000	2.000000	0.290000	NaN	NaN	6811.000000	1.650000e+03	NaN	4.750000e+05
95%	NaN	6.000000	5.000000	14.000000	NaN	NaN	11375.000000	4.770000e+03	NaN	2.749000e+06
99%	NaN	10.000000	8.000000	91.800000	NaN	NaN	19130.000000	9.187000e+03	NaN	7.500000e+06
99.9%	NaN	20.000000	15.000000	519.090000	NaN	NaN	19804.000000	2.999700e+04	NaN	2.397000e+07
max	NaN	123.000000	198.000000	100000.000000	NaN	NaN	99999.000000	1.450112e+06	NaN	8.750000e+08

Step2: Handling Missing Values

In this phase, we aimed to enhance the integrity and reliability of the dataset through a series of data refinement and imputation steps.

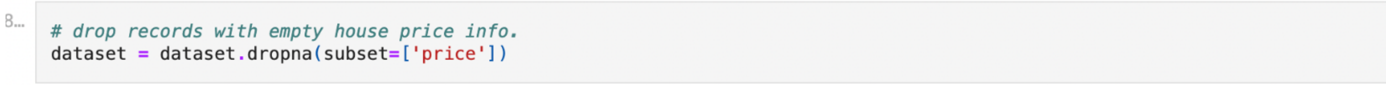
- Identification and Removal of Records with Empty Prices.

  Initially, we conducted an examination of the 'price' variable to ascertain the presence of empty records. A total of 71 instances were identified and subsequently removed from the dataset, ensuring accuracy in our predictive model.



```
7…
# check number of records where the price is empty
print(dataset['price'].isna().sum())

71
```

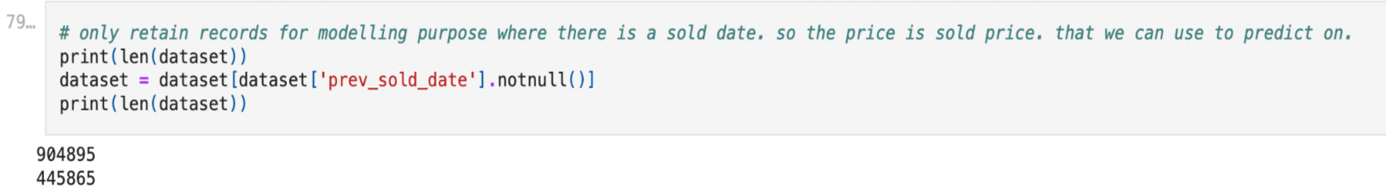


```
8…
# drop records with empty house price info.
dataset = dataset.dropna(subset=['price'])
```

  (please check out our github for all the codes we have input so far)

- Filtering for Relevant Records

Following that, we filtered the dataset to retain only records with a ‘prev_sold_date’, signifying a previous sale of the property. This step was crucial in ensuring that the prices analyzed represented actual transaction values.



```
# only retain records for modelling purpose where there is a sold date. so the price is sold price. that we can use to predict on.
print(len(dataset))
dataset = dataset[dataset['prev_sold_date'].notnull()]
print(len(dataset))
```

```
904895
445865
```

- Comprehensive Missing Value Analysis and Imputation

  To address missing data, we utilized a combination of indicator variable creation and strategic imputation techniques. Specifically, binary indicators such as ‘bed_missing’, ‘bath_missing’, ‘acre_lot_missing’ and ‘house_size_missing’ were established to denote missing data instances. These indicators can be used at the later modeling stage as an interaction term to account for the missing info. At the same time, considering that the missing data percentage for ‘bed’ and ‘bath’ was less than 5%, we evaluated various imputation strategies, including applying the mean, median, and even considering the possible exclusion of such records. For ‘acre_lot’and ‘house_size’, missing values were imputed using their respective means, while for 'bed' and ‘bath’, missing values were replaced with the corresponding medians.

  Additionally, we explored a category based imputation strategy wherein missing values would be filled with group medians with respective price bins. However, we abandoned this approach because it had the potential to oversimplify the data, introduce bias, and fail to accurately represent the diversity among properties in the same price categories.

Step3: Data Capping and Truncation

To mitigate the impact of extreme values and outliers, we introduced caps and floors for certain attributes. Specifically: The ‘house_size_fiiled’ was capped and floored at the 99th and 1st percentiles, respectively, resulting in a range of 500 to 7000. The ‘bed_filled’.‘bath_filled’ and ‘acre_lot_filled’ were capped at their respective 99th percentiles.

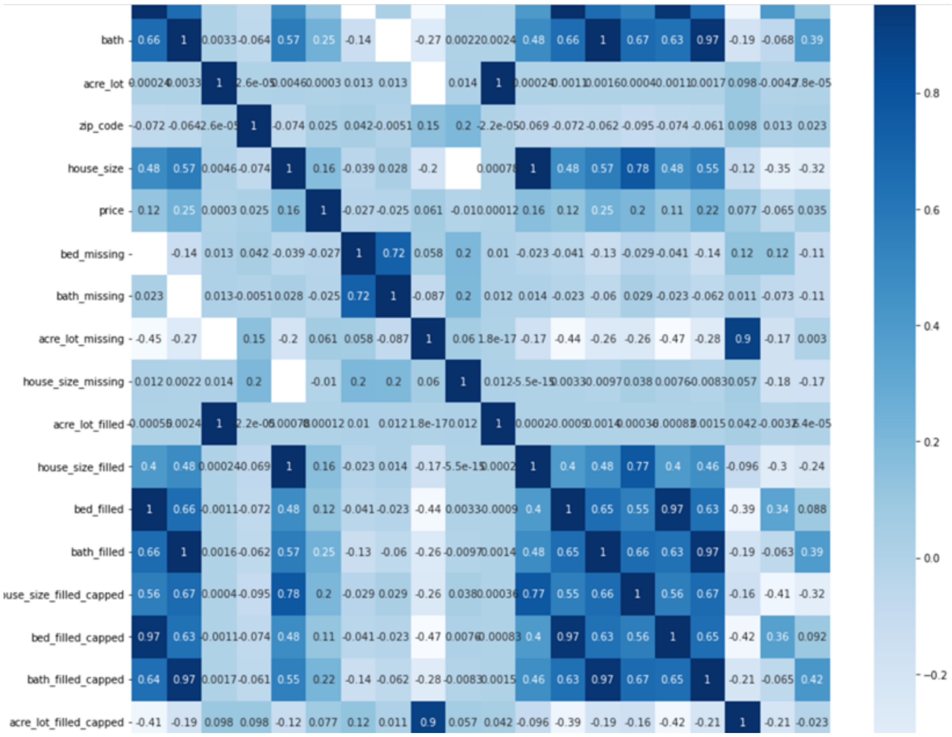
Based on the summary stats, we also noticed the response variable ‘price’ seems to have a wide distribution. With min of 0 and max of 8.75*e+08. This wide range will cause issues when trying to build a solid predictive model, given the limited amount of input features available. We are considering either to floor/cap the price at 1%&99%, or drop the outliers beyond those two thresholds. And later at the modeling stage we will try out logging the response variable price so as to achieve a better model performance.

We posit that the dataset has undergone substantial optimization through the execution of meticulous and systematic procedures. Consequently, it is now suitably primed for subsequent modeling and analysis, thereby ensuring that the data utilized in predictive modeling is both representative and robust, bolstering the reliability of the outcomes.

2.Data EDA

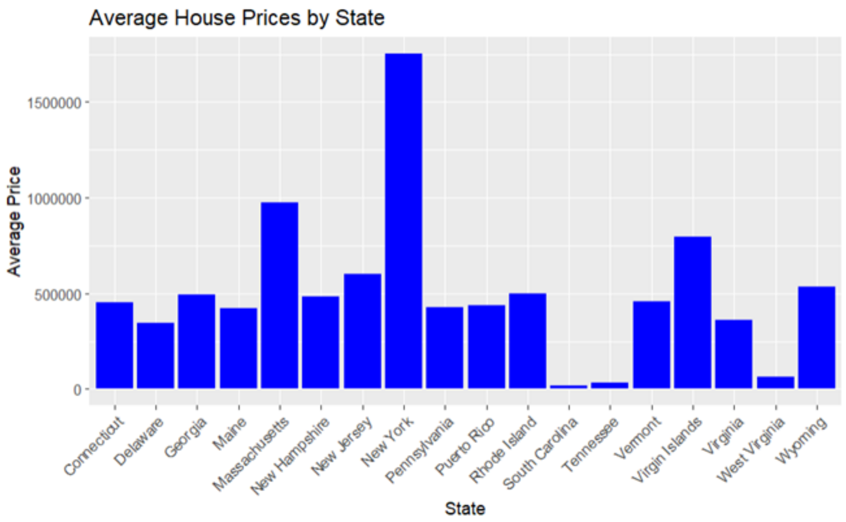
Based on the cleaned up dataset, we have performed initial EDA on the independent variables. We try to get the average price for each level of each independent variable so we can get a general idea what would be the linear relationship between the independent variables & dependent variable price.

To explore the relationship between the predicting variable and the response (price), we utilized the heatmap to visualize how the predicting variable correlated to the response (price), as well as the correlation between each pair of predicting variables. As our goal was to create a price predicting model with the existing predicting factors in the dataset, we first ran an initial multilinear regression with all the original form of variables to test whether our existing data would be able to fulfill the underlying assumption for multilinear regression and whether any transformation of variables would be needed.



The assumptions for multilinear regression are linearity, constant variance, uncorrelated errors and normality. We plot each predicting variable to the response variable to check for the linear relationship. The variance inflation factor (VIF) from the model also indicated multicollinearity did not exist, which confirmed what we can see from the heatmap. Then we also plot the residual from the initial model to the fitted value to check whether it meets the constant variance criteria. We also used qq-plot and histogram to check for normality.

However, from most of the plots that we got, we noticed an issue regarding price differences among states that are quite significant. As we can see from the bar chart below, the average house price of a few states (New York and Massachusetts) in our dataset are at very high end, around $1000k range, compared to some of them are on the lower side of $250k (South Carolina, Tennessee and West Virginia). This posed an uncertain factor to our process of model building. We are discussing the necessity of building models to fit in different price ranges so it can better reflect the income level or purchasing power across different states. And another optional solution is to use State as an categorical input variable to account for the inherent difference between the price of different states.



3.Feature Engineering

The dataset under consideration encompasses a limited selection of features. Consequently, to enhance the richness of the dataset and facilitate a more comprehensive analysis, we envisage incorporating several additional variables derived from the existing features. These encompass ratios such as the number of bedrooms and bathrooms per unit of acreage (bed/bath per acre_lot), the ratio of house size to acreage (house_size/acre_lot), and a categorical variable delineating the relationship between the number of bedrooms and bathrooms. Preliminary feature engineering efforts have been undertaken, and we are currently engaged in the meticulous process of appending the remaining variables to the dataset. We also investigated the other features, like sold year, and tried to find the relationship between sold year and house price.

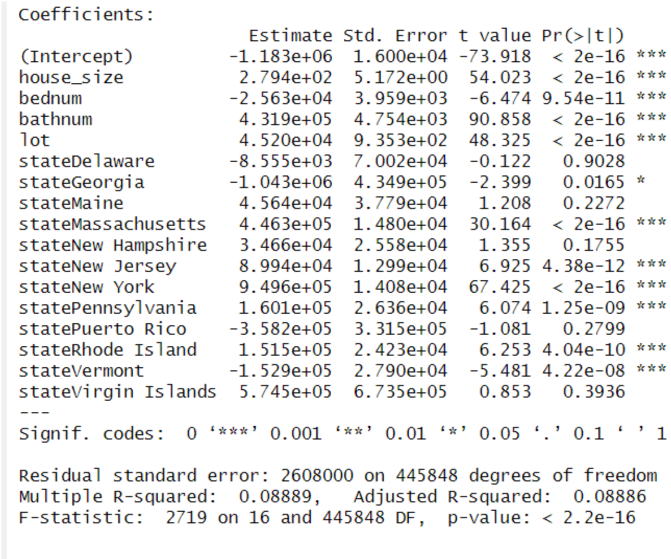
4.Modeling

We have performed random sampling on the entire modeling data to split into 60/20/20 as for training/validation/holdout. This will help ensure we are building a model that is neither under-predictive nor over-predictive.

We will try out a few different approaches for this prediction project. We started with a simple linear regression model, using the price as response and bed/bath/acre_lot/house_size as input features. The first attempt is flagging a few things to look out for at a later modeling stage, such as the scale of the response variable price is too large and the range of values is also too wide, compared with the input features.

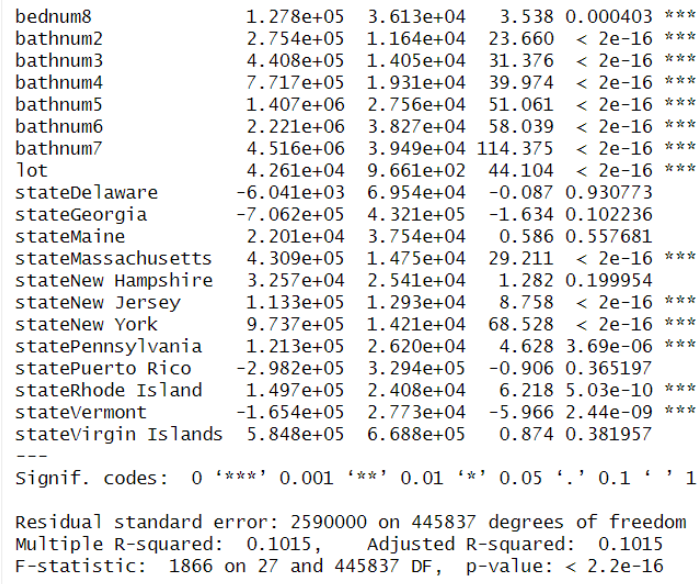
We also tried the approach to turn some of the quantitative variables (bath/ bedroom) to qualitative variables to see how this change would affect the model.

We will consider using BoxCox transformation to decide whether we can use log(price) as the response variable in future model builds or any other kinds of transform would be appropriate.



```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -1.183e+06  1.600e+04 -73.918  < 2e-16 ***
house_size           2.794e+02  5.172e+00  54.023  < 2e-16 ***
bednum              -2.563e+04  3.959e+03  -6.474 9.54e-11 ***
bathnum              4.319e+05  4.754e+03  90.858  < 2e-16 ***
lot                  4.520e+04  9.353e+02  48.325  < 2e-16 ***
stateDelaware       -8.555e+03  7.002e+04  -0.122   0.9028    
stateGeorgia        -1.043e+06  4.349e+05  -2.399   0.0165 *  
stateMaine           4.564e+04  3.779e+04   1.208   0.2272    
stateMassachusetts   4.463e+05  1.480e+04  30.164  < 2e-16 ***
stateNew Hampshire   3.466e+04  2.558e+04   1.355   0.1755    
stateNew Jersey      8.994e+04  1.299e+04   6.925 4.38e-12 ***
stateNew York        9.496e+05  1.408e+04  67.425  < 2e-16 ***
statePennsylvania    1.601e+05  2.636e+04   6.074 1.25e-09 ***
statePuerto Rico    -3.582e+05  3.315e+05  -1.081   0.2799    
stateRhode Island    1.515e+05  2.423e+04   6.253 4.04e-10 ***
stateVermont        -1.529e+05  2.790e+04  -5.481 4.22e-08 ***
stateVirgin Islands  5.745e+05  6.735e+05   0.853   0.3936    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2608000 on 445848 degrees of freedom
Multiple R-squared:  0.08889,	Adjusted R-squared:  0.08886 
F-statistic:  2719 on 16 and 445848 DF,  p-value: < 2.2e-16
```



```
bednum8              1.278e+05  3.613e+04   3.538 0.000403 ***
bathnum2             2.754e+05  1.164e+04  23.660  < 2e-16 ***
bathnum3             4.408e+05  1.405e+04  31.376  < 2e-16 ***
bathnum4             7.717e+05  1.931e+04  39.974  < 2e-16 ***
bathnum5             1.407e+06  2.756e+04  51.061  < 2e-16 ***
bathnum6             2.221e+06  3.827e+04  58.039  < 2e-16 ***
bathnum7             4.516e+06  3.949e+04 114.375  < 2e-16 ***
lot                  4.261e+04  9.661e+02  44.104  < 2e-16 ***
stateDelaware       -6.041e+03  6.954e+04  -0.087 0.930773    
stateGeorgia        -7.062e+05  4.321e+05  -1.634 0.102236    
stateMaine           2.201e+04  3.754e+04   0.586 0.557681    
stateMassachusetts   4.309e+05  1.475e+04  29.211  < 2e-16 ***
stateNew Hampshire   3.257e+04  2.541e+04   1.282 0.199954    
stateNew Jersey      1.133e+05  1.293e+04   8.758  < 2e-16 ***
stateNew York        9.737e+05  1.421e+04  68.528  < 2e-16 ***
statePennsylvania    1.213e+05  2.620e+04   4.628 3.69e-06 ***
statePuerto Rico    -2.982e+05  3.294e+05  -0.906 0.365197    
stateRhode Island    1.497e+05  2.408e+04   6.218 5.03e-10 ***
stateVermont        -1.654e+05  2.773e+04  -5.966 2.44e-09 ***
stateVirgin Islands  5.848e+05  6.688e+05   0.874 0.381957    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2590000 on 445837 degrees of freedom
Multiple R-squared:  0.1015,	Adjusted R-squared:  0.1015 
F-statistic:  1866 on 27 and 445837 DF,  p-value: < 2.2e-16
```

5.Model evaluation & comparison

Since we are planning to try out a few different models, we want to use the same evaluation metrics to compare different model techniques. MSE/RMSE is our current pick for model evaluation metric.

6.Review Of Literatures:

1. An, X., & Lo, H. P. (2015). Determinants of House Prices: A Quantile Regression Approach. The Journal of Real Estate Finance and Economics, 50(4), 601-622.

https://www.researchgate.net/publication/23534659_Determinants_of_House_Prices_A_Quantile_Regression_Approach

The authors have examined how various variables affect house prices across different percentiles of the price distribution and tried to understand whether the factors have different impacts on house prices at different price levels. They used the quantile regression and analyzed the result by identifying which factors are significant in influencing house prices and how their effects vary at different price percentiles. It can provide some insight and direction when we do similar analysis with our data.

2. Determinants of the Price of Housing in the Province of Alicante (Spain): Analysis Using Quantile Regression"by Raúl-Tomás Mora-García, María-Francisca Céspedes-López, V. Raúl Pérez-Sánchez, Pablo Martí, and Juan-Carlos Pérez-Sánchez\
https://www.mdpi.com/2071-1050/11/2/437

The authors here also applied the quantile regression technique in their analysis of the house pricing in the Province of Alicante in Spain . The quantile regression approach enables a more nuanced examination of how the relationships between variables change at different percentiles of the housing price distribution. It is another example which provided us with more information and understanding of how to use quantile regression. We can consider this methodology in our project.

Summary:

In this project, our data cleaning process involved meticulous exploration, handling of missing values, and data capping to ensure the dataset's quality and suitability for modeling. Subsequently, we conducted initial exploratory data analysis (EDA) to understand the relationships between independent variables and the response variable, price. We tested our data against the assumptions for multilinear regression, including linearity, constant variance, uncorrelated errors, and normality.

As we delved deeper into our dataset, we discovered significant variations in house prices among different states, which posed challenges for our modeling process. This led us to consider the necessity of building models tailored to different price ranges to better reflect regional income levels and purchasing power.

To enhance the richness of our dataset, we embarked on feature engineering, introducing additional variables derived from existing features. We also explored the relationship between the year of sale and house price.

Our project's next phases will involve modeling and model evaluation, where we will use various approaches, including linear regression and possibly log-transforming the response variable to address its wide range. Model evaluation will be based on metrics like Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). We aim to select the best-performing model to make reliable predictions in the field of real estate.