**MACHINE LEARNING PROJECT**

**Course code:INT-254**

**Submitted by:**

**Akash Deep-12006825**

**Roll No:11**

**Dinkar-12007839**

**Roll No:13**

**Topic-Zillow Home Value Prediction**

**Code link:** [Zillow Home Value Prediction.ipynb - Colaboratory (google.com)](https://colab.research.google.com/drive/14BEOUuaL7-w7xho4kfDdovwZK86a0FQM#scrollTo=saE9O2devczH)

**Dataset:** <https://www.kaggle.com/code/dasarimohana/zillow-house-price-prediction-by-dasari-mohana/notebook>

INTRODUCTION

Buying a house that suits their choices is every person's desire, and it is thus known as their dream house. One considers several aspects while purchasing a home, starting from the budget, the location, the number of rooms available, and many more. But how to find a house that satisfies one's requirements? This is not a quick and easy task.

But no need to worry; homebuyers can nowadays find their dream home with a click of a button. Zillow is a popular estimator for house evaluation available online. It is considered one of the top real estate marketplaces for buying a house in the United States. Zillow's Zestimate allows the homebuyers to search for a home that satisfies their location, area, budget, etc.

The Zillow Zestimate provides the homebuyers with information on the actual worth of the house based on public data. The accuracy of the Zestimate information depends on the location and availability of the data of a specific area. Hence the more data available, the more is the accuracy of the Zestimate.

**Aim:**

To predict the sale prices of the houses and improve the log error i.e., the error due to the difference between the actual and the predicted home values.

**My Approach**

1. Importing the required libraries and reading the dataset.

a. Merging of the two datasets

b. Understanding the dataset

1. Exploratory Data Analysis (EDA) –

a. Data Visualization

1. Feature Engineering

a. Duplicate value removal

b. Missing value imputation

c. Rescaling of incorrectly scaled data

d. Standardization

e. Encoding of categorical variables

f. Generation of new feature wherever required.

g. Dropping of redundant feature columns

h. Checking for multi-collinearity and removal of highly correlated features

i. Check for the outliners and removal of outliers.

4. Model Building

a. Performing train test split

b. Feature Scaling

c. Dropping features if necessary

d. Linear Regression Model

e. Elastic Net

f. Ridge Regression

g. Lasso Regressor

h. XG Boost Regressor

i. Ada boost Regressor

j. Gradient Boosting Regressor

k. Decision Tree Regressor

l. Forest Regressor

5. Model Validation

a. Mean Absolute Error

b. Mean Squared Error

c. Root Mean Squared Error

1. Hypermeter Tuning (Grid Search CV)

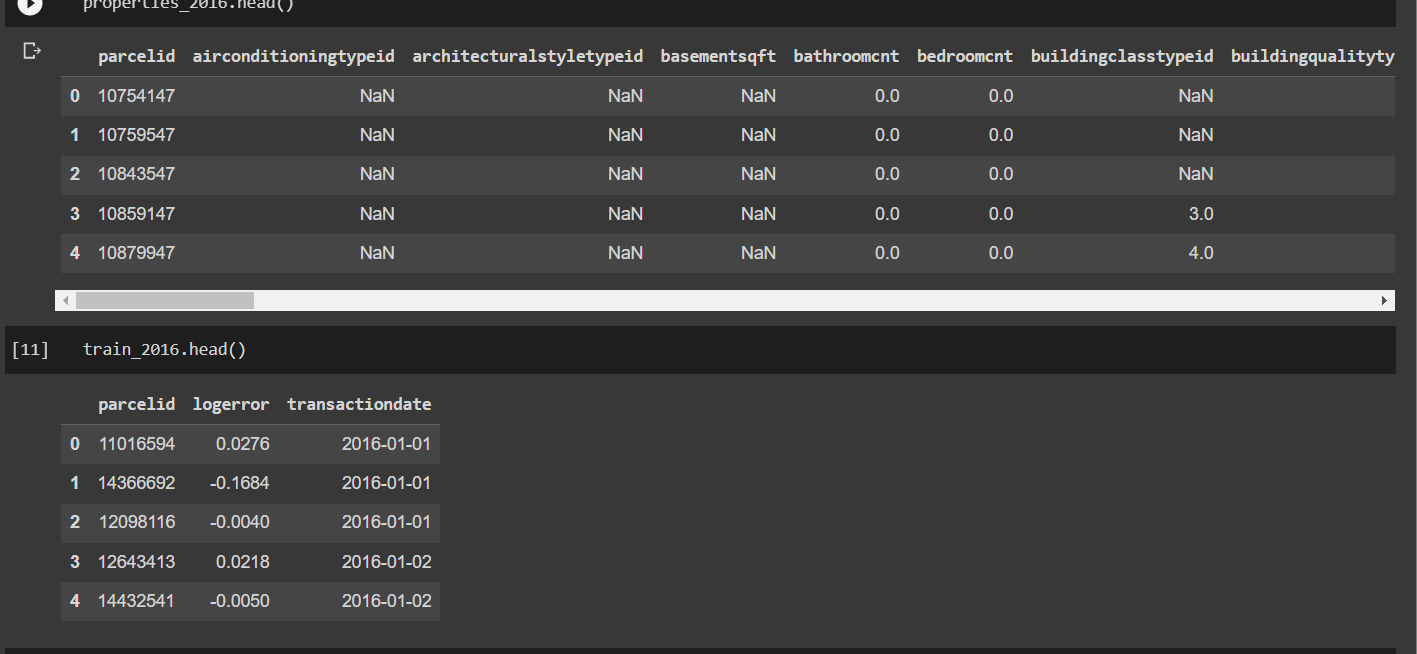
a. For Random Forest Regressor

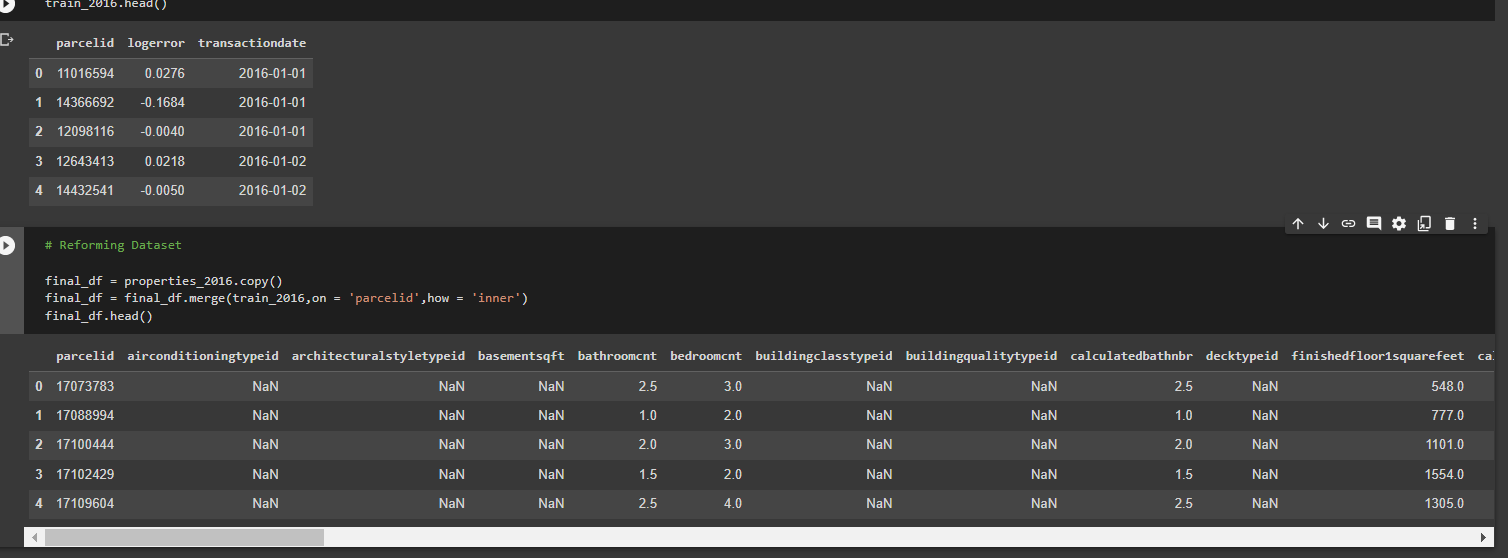
1. Checking for Feature Importance
2. Creating the final model and making predictions
3. Conclusion

### **Importing Libraries & Loading Dataset:**



OUTPUT:



****

**Feature Description:**

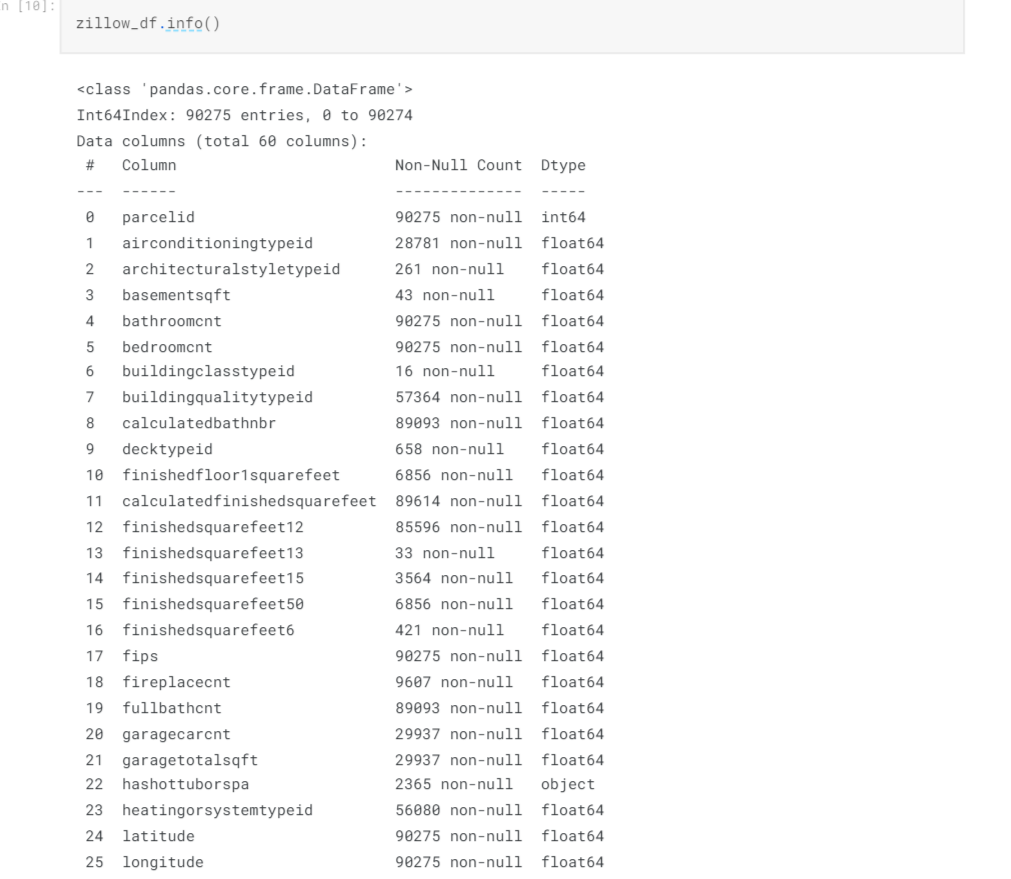
****

# **Exploratory Data Analysis:**

The above reformed dataset contains 90275 rows and 60 columns.

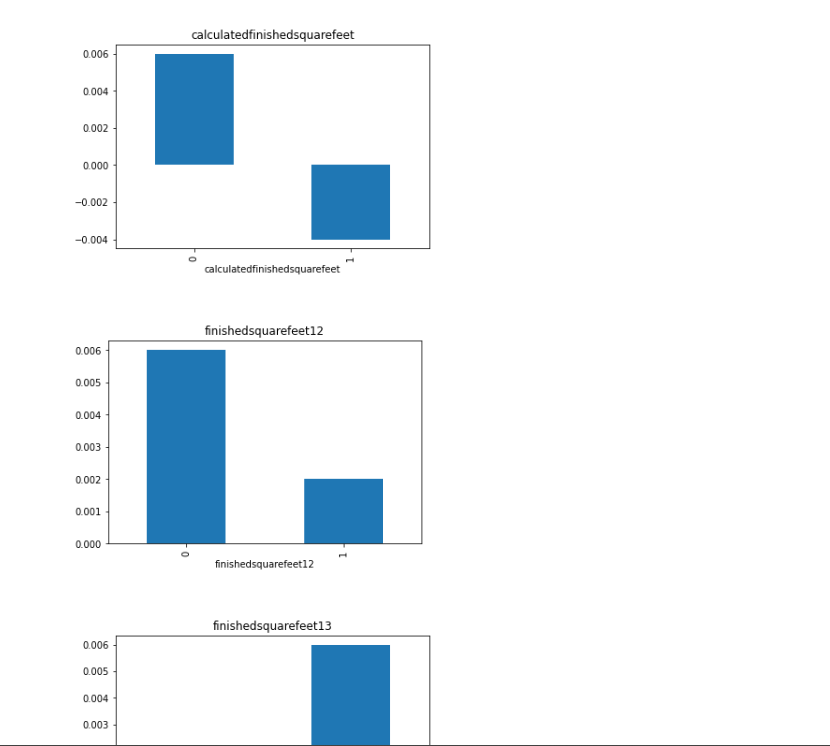
Let's analyse the dataset to identify the following:

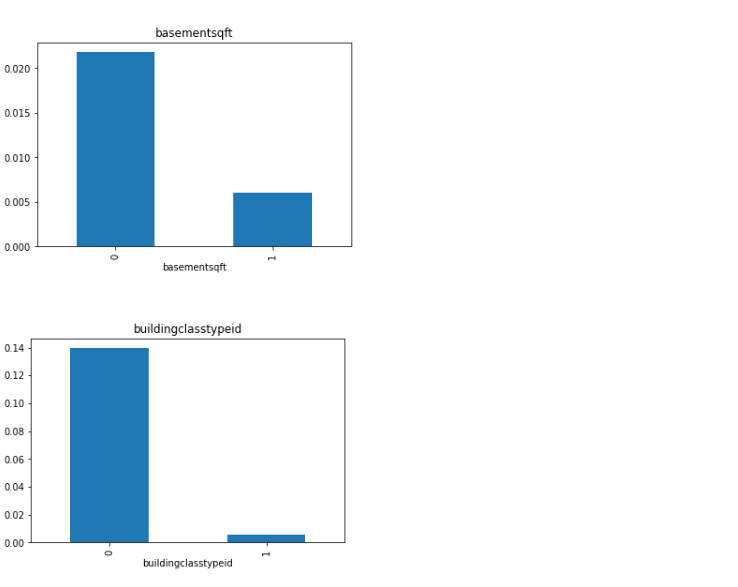
1. Missing values
2. Distribution of the numerical variables
3. Outliers
4. Distribution of the categorical variables
5. Potential relationship between the variables and the target

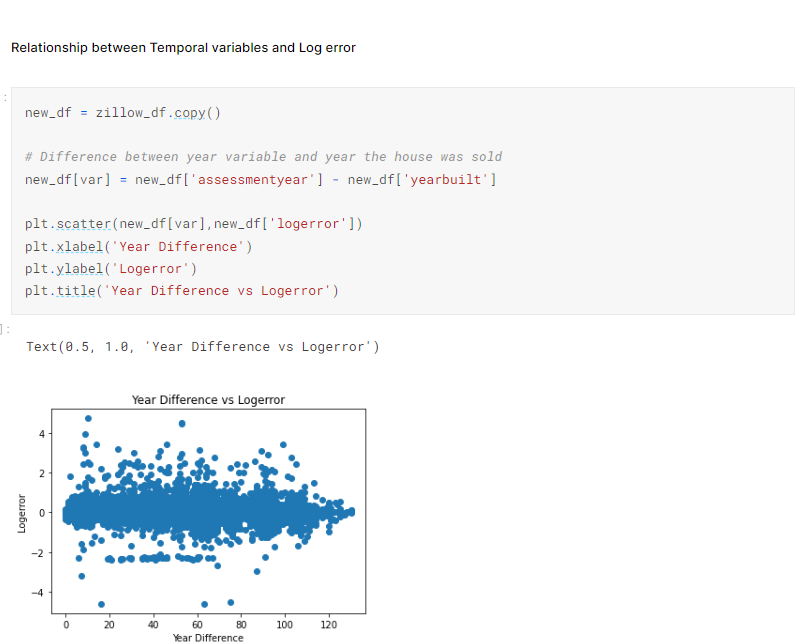


### **Relationship between missing values and log error (Target Variable)**

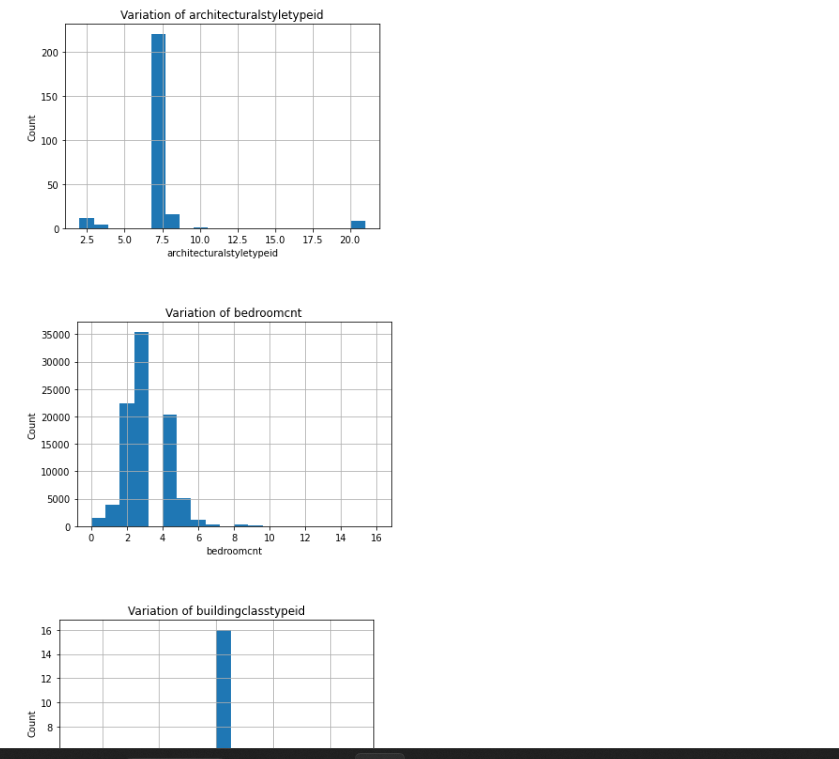
****

****

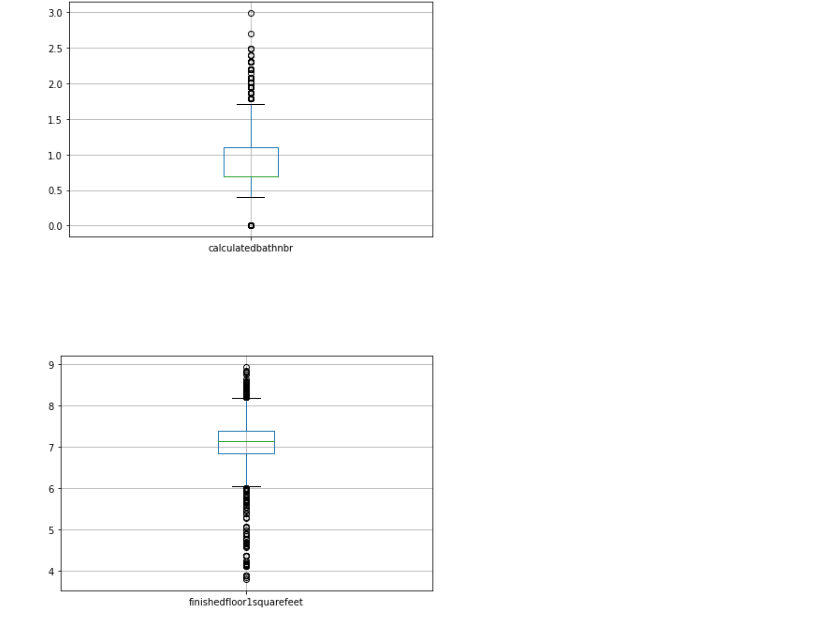
****

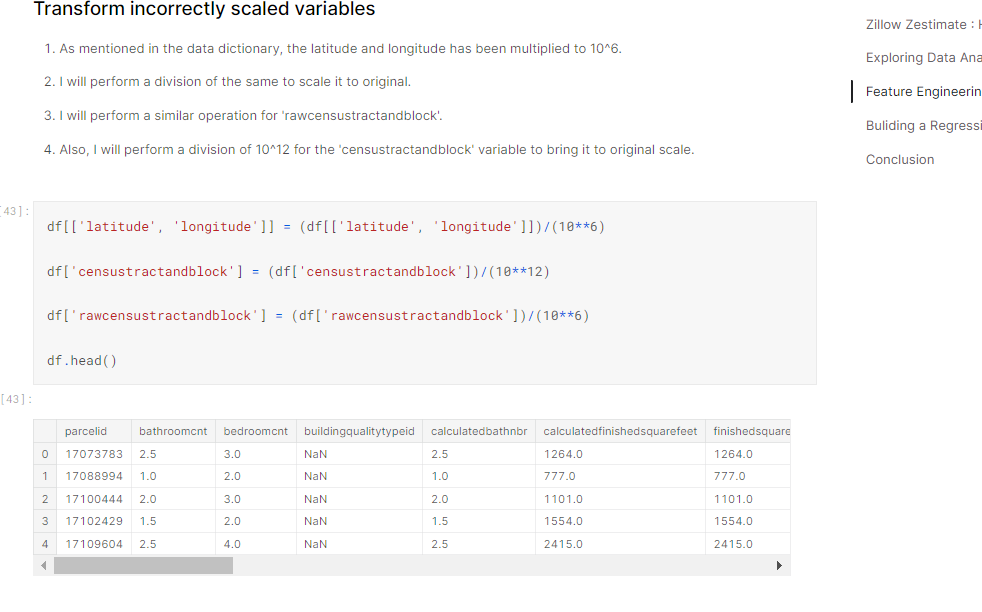
****

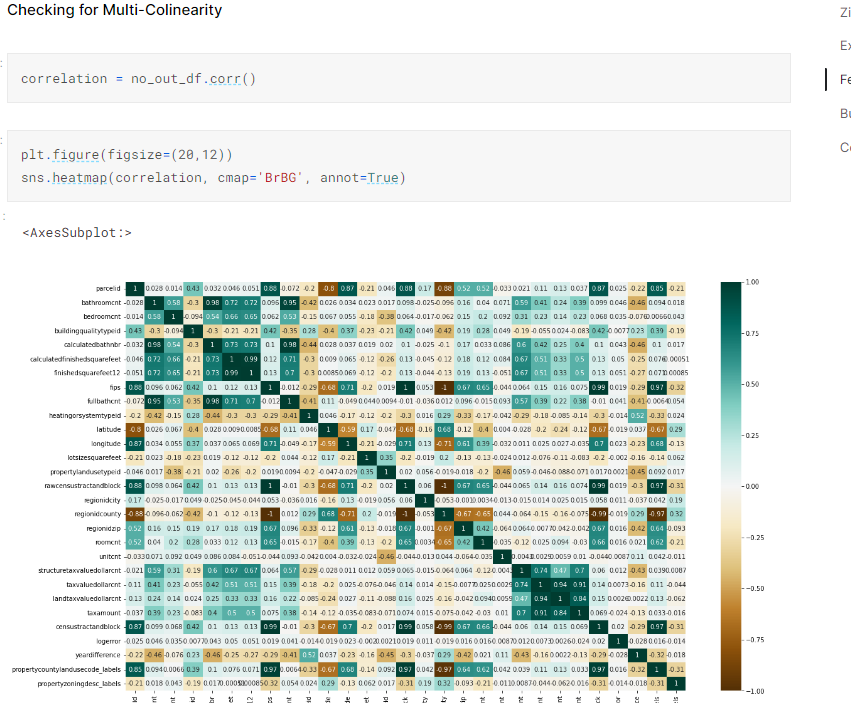
**Distribution of Discrete variables**

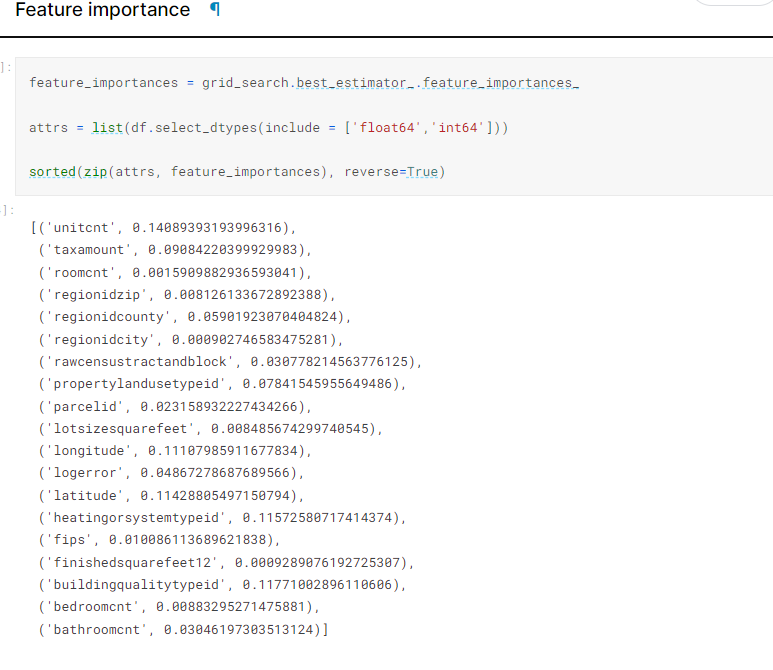
****

### **Analysing Outliers**







****

****

# 

# **Conclusion**

1. I have performed all the feature engineering steps necessary to ensure the dataset is ready to be fed into Machine Learning algorithms.
2. After Pre-processing and Feature Engineering the raw dataset we spitted the dataset into train and test sets.
3. Performed Feature scaling on data for better performance.
4. Trained multiple models using different ML regression algorithms on dataset.
5. Applied Performance metrics such as MAE, MSE, RMSE to find out best prediction model.
6. With the help of Grid Search CV, we found out best estimator with least Root mean squared error.
7. Saved best predictor in. pickle format for future predictions.
8. Done prediction on test data and saved predictions into .csv