**Housing Price Prediction**

Jyoti Sahu

Bellevue University

DSC680- Applied Data Science

July 17, 2022

**Topic**

This project's objective is to forecast the selling price of a house, given the presence of specific qualities or attributes in that house.  
  
A model like this has many of useful applications in real life. For instance, a builder of new homes can decide on the ideal number of bedrooms, bathrooms, or a popular neighborhood to maximize selling price. This model could be used by a homeowner to choose the best asking price for their house. Alternately, a lender can use the model to estimate the worth of a house and then utilize that information to approve loans or set interest rates.

**Business Problem**

In the US, the number of homes has recently increased dramatically. The Zillow Home Value Index shows that prices increased by a startling 18% between 2020 and 2021. Now, purchasing a home in this market might bring in money for real estate investors. Real estate investors usually buy houses, fix them up, and resell them for a profit. Using a machine learning algorithm on previous data, the company will obtain a full understanding of the market. Several different types of questions, such as

**Research Questions:**

* What is the potential cost of a new home?
* Where in a certain County is it most advantageous to buy a new piece of property?
* Is a home's renovation necessary?
* What elements determine if a real estate investor wants to improve a house they recently purchased?

**Background**

Since loans are used to finance the purchase of homes, increased financial stability and an accelerated loan growth rate tend to boost housing demand, possibly by driving up property values. Theoretically, this arrangement is known as the house price channel. The process of altering total production and price level through a change in monetary policy that impacts real estate prices, such as those of housing and land, and consequently the household's investments and outlays, is known as the "house prices channel." High interest rates cause people to save more money and reduce the market for housing that is geared toward investors. However, housing prices rise when interest rates decline. The effect is a rise in overall demand.

**Datasets**

**Data Source URLs:**

* <https://www.kaggle.com/code/konerunikhil/king-county-sales-regression/data>

**Column definition:**

We have got 20 columns with 21590 entries and the columns are

* Id – ID of row
* Date – House sold date
* Price – House Price
* Bedrooms – Number of bedrooms
* Bathrooms - Number of bathrooms
* sqft\_living – Living Square foot
* sqft\_lot – Lot Size
* floors – Number of floors
* waterfront – water bodies view
* view – has view?
* condition – House condition
* grade – County Grade
* sqft\_above -
* sqft\_basement
* yr\_built – Built Year
* yr\_renovated – if renovated then renovated year
* zipcode
* latitude
* longitude
* sqft\_living15 – Size of renovation in 2015
* sqftlot15 – total lot size in 2015

**Methods**

I am going to follow below approach to build my ML model.

* Missing Value check
* Understand the data
* Model fitting check
* Analysis of all the features and check correlation to house price
* Visualization for for easy understanding
* Feature engineering
* Training model
* Hyperparameter -tuning
* Model accuracy check
* Conclusion

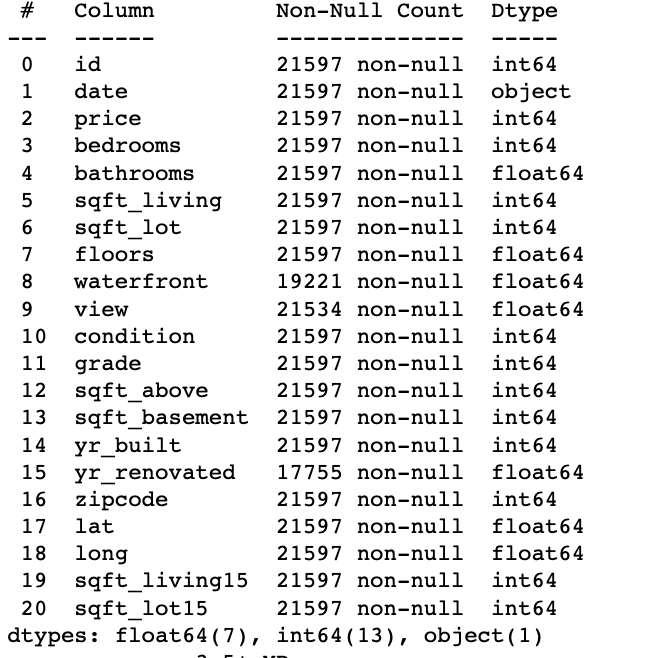
**Analysis**

21 columns and 21K observations make up the data. As we can see from the above table, "id" appears to be a distinctive identifier for each sale record, and "sqft living" is the total of "sqft above ground" and "sqft below ground."

Let's use the info() method to quickly get a description of the data.

Out of 21 variables, one is categorical (an object), while the other 20 are numerical (int/float), according to the data.

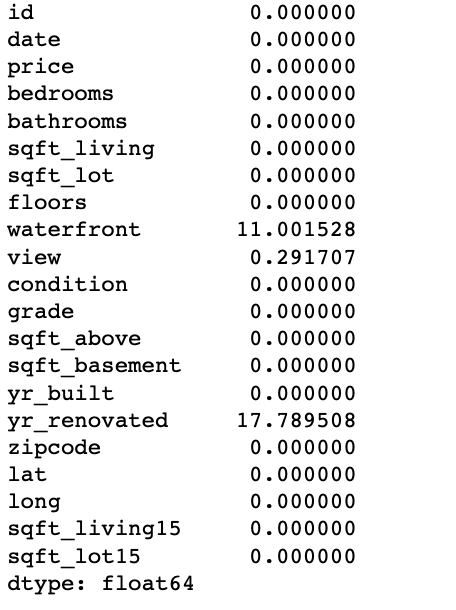
However, categorical variables like waterfront, view, condition, grade, year built, year renovated, and zipcode are listed in the data description provided with the dataset.



Checking to see whether any columns have any missing values is important since missing values might cause issues for linear models. If so, the issues can be fixed during data cleansing.

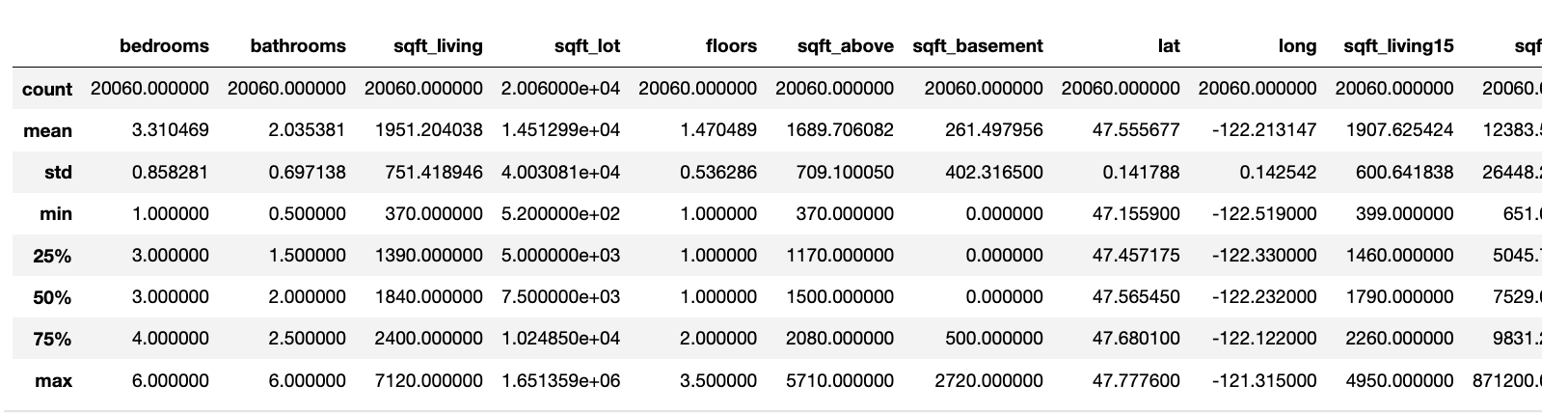
Let's look at the percentage of these columns that have missing values.

We can see waterfront and ye\_renovated has missing values



Let's see the distribution of the price. The describe method shows , 75% percentile of the data are below $645K priced houses. The real estate investors focus primarily on investing in homes that are appropriate for people of intermediate income. As a result, the investors are guaranteed a consistent stream of income because these homes are more likely to be sold quickly. Therefore, I will continue to include in the data any homes that are effectively under $1 million in value or have the typical 2 to 6 bedrooms.

The following table provides a brief overview of the numerical variables and lists their average value along with their minimum and maximum values. Let's investigate further by displaying the numerical factors and combining them with prices to determine what type of relationship each one offers to the price of the house.



Observation made from EDA Please see the appendix for the visualization

1. Price and number of bedrooms - A home's price rises as the number of bedrooms grows. With a little digging, we can find that there is only a $5K difference, but this is being stated without accounting for the impact of other amenities on price.

2. The relationship between price and bathrooms appears to be linear, thus when there are more bathrooms in a home, the price of the home also rises.

3. House price and interior square footage (sqft) - The price of the house rises as the interior square footage does. Their structure is linear.

4. Price and lot area(sqft) of the house\*\* - Price and sqft\_lot do not have any relationship. This feature doesn’t seem very predictive in determining house price.

5. Price and the Number of Floors in the House\*\* - The cost of a home rises a little bit with the number of floors.

6. The cost and square footage (sqft) of the home. With each additional square foot of living space, the price of the home rises. Their structure is linear.

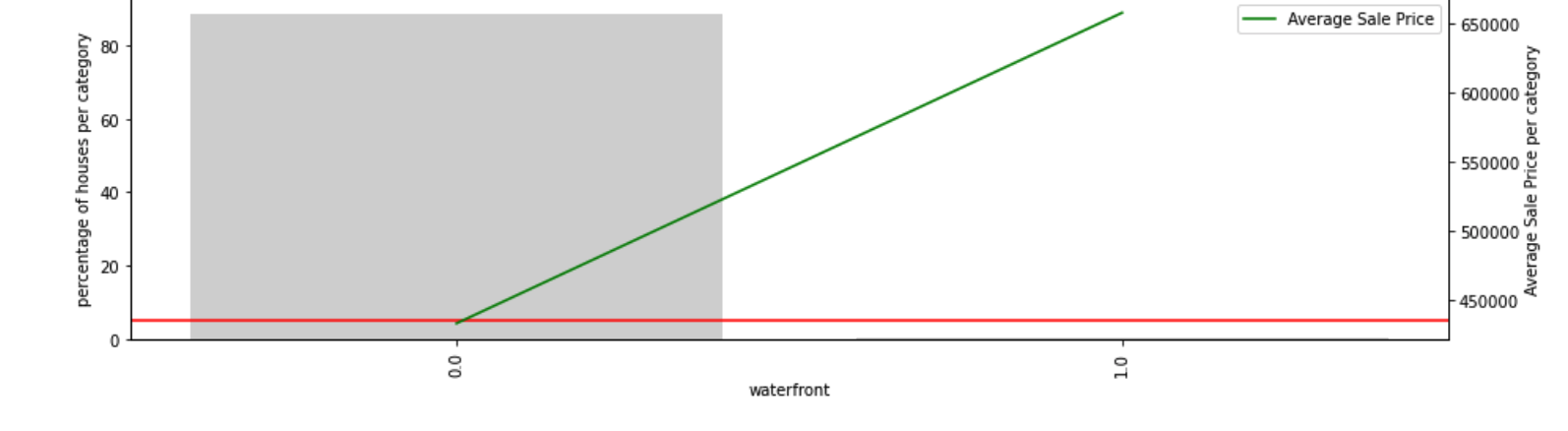
7. There is no correlation between the price and the basement's square footage (sqft). This function doesn't seem to be particularly accurate in predicting home prices.

8. The price of the house rises with the interior living space in square feet (sqft) for the 15 closest neighboring properties. Their structure is linear.

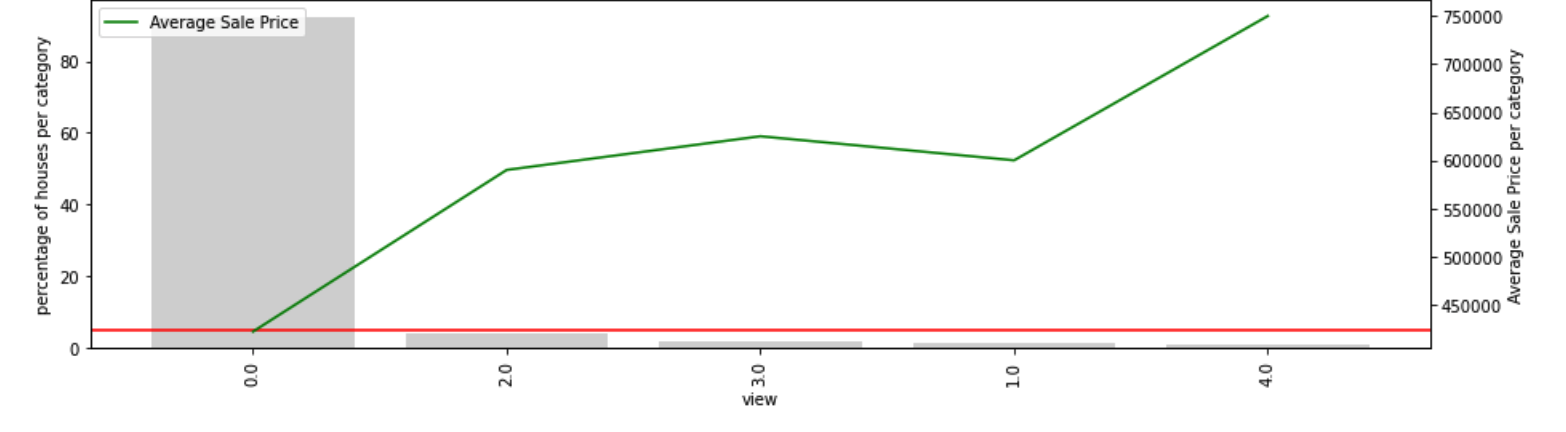
9. The price of the home and the average lot size of the 15 closest neighboring homes (in square feet) show no relationship. This function doesn't seem to be particularly accurate in predicting home prices.

Now moving to exploration of categorial values Date, waterfront, view, condition, grade, year built, year renovated, and zipcode are categorical variables, as we have seen in the dataset description.

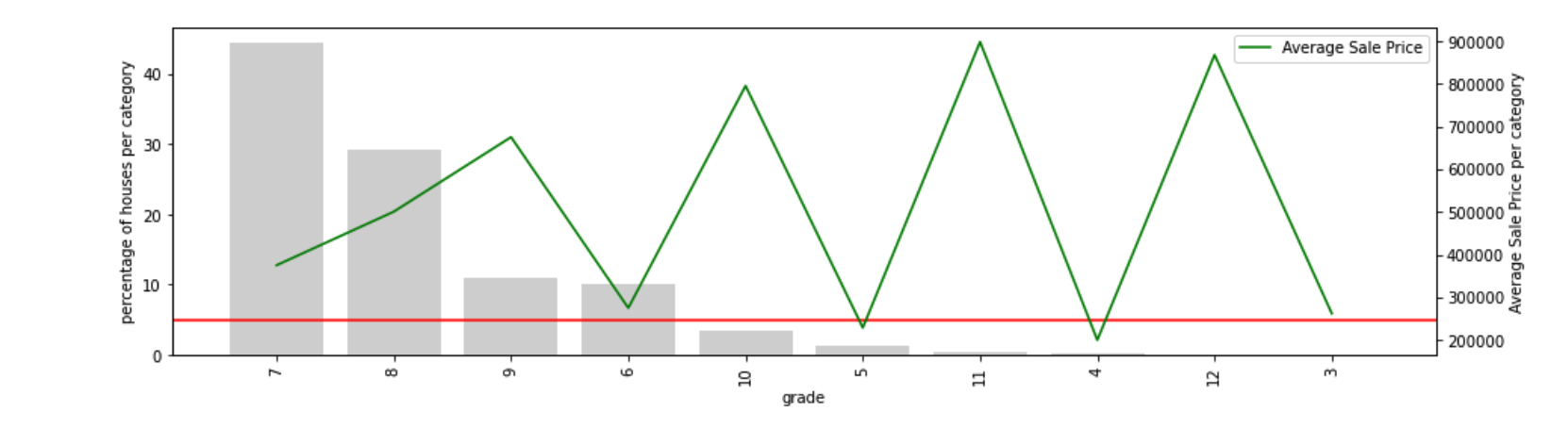
Better is the view of the property, higher is the price . However, better view (1-4) comprises less than 5% houses



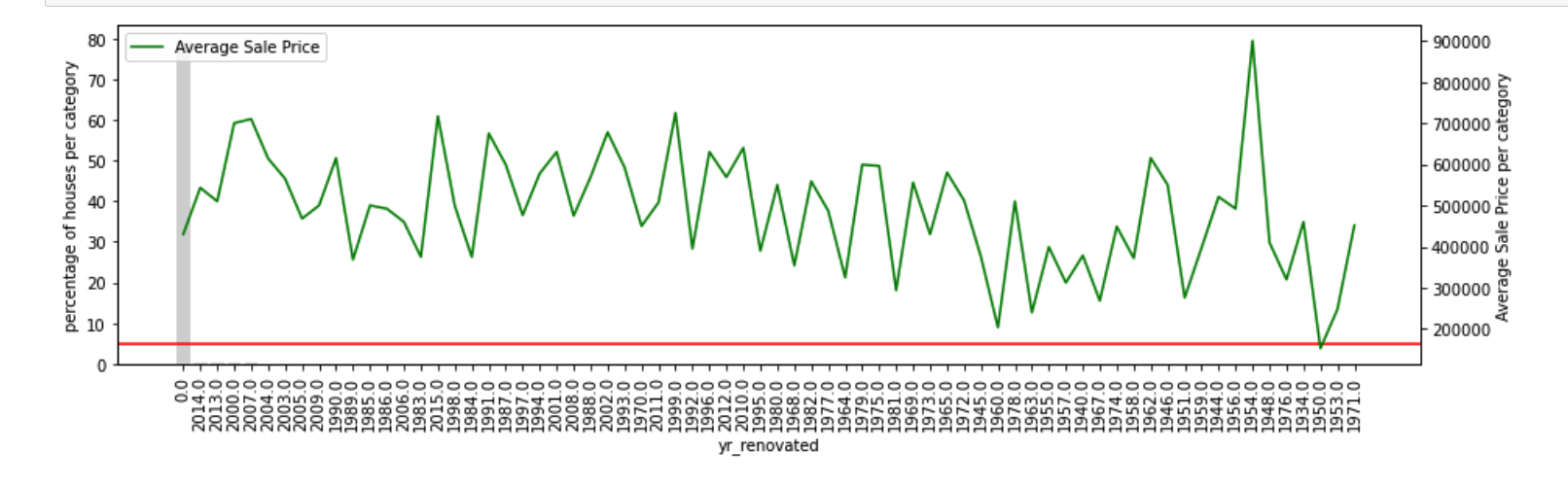
Higher the number in condition, higher is the average sale price.



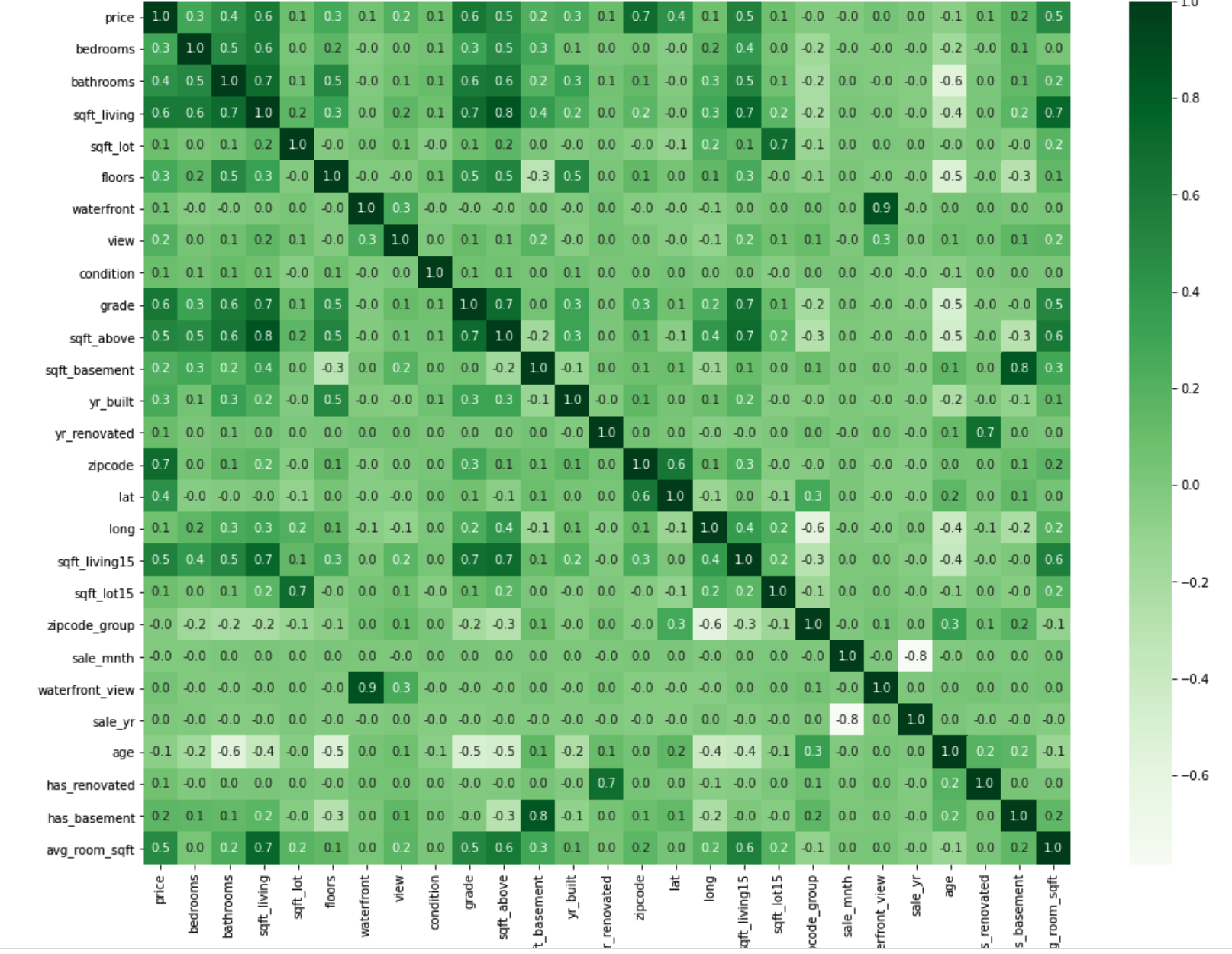
Higher the grade of the house, higher is the average sale price.

****

From the below diagram we see, maximum number of houses were not renovated.

****

After the feature engineering we have dropped the few variables which are not needed anymore and other variables are already been derived. to eventually determine which factors are connected with the price variable, a correlation chart is created. The graph demonstrates the strong correlation between price and attributes like square footage, bathrooms, and grade.

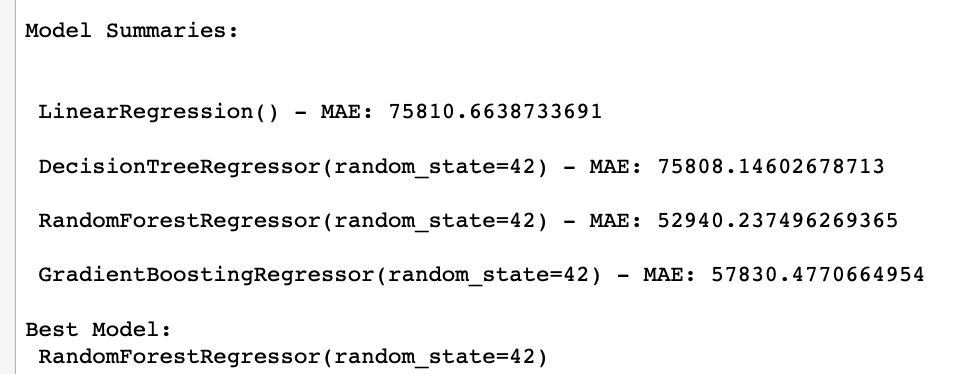
****

The simplest way to anticipate a house price is to build a baseline model using the median house price across all zipcode groups. The price of a property is the median price of all the houses in the zipcode group to which it belongs when there is no ML model in place.

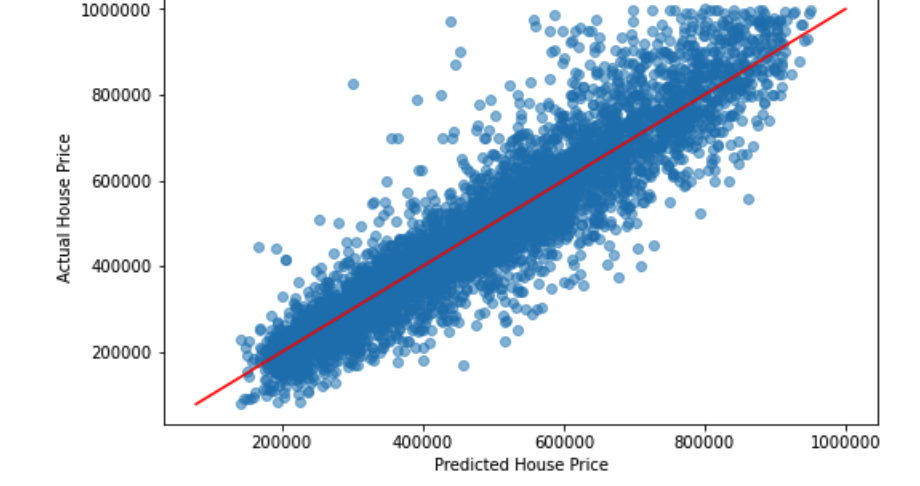
Separating the primary data into train and validation data allows us to finalize the model using the training data and then test its effectiveness using validation data. This is crucial so that we can understand how our model predicts the price of a new home with these amenities when it hits the market. I have splited the train and test as 70:30 ratio.

We can see from the exercise below that Random Forest has the lowest MAE and performs the best for our data. This indicates that on average, our projection is only $52K off.

I did some hypertuning of my model. The model’s estimate is only $51K off from the actual worth of the house after the hyperparameters have been adjusted and tested on hypothetical data. Additionally, the model accounts for 86% of the price variation in the dataset.

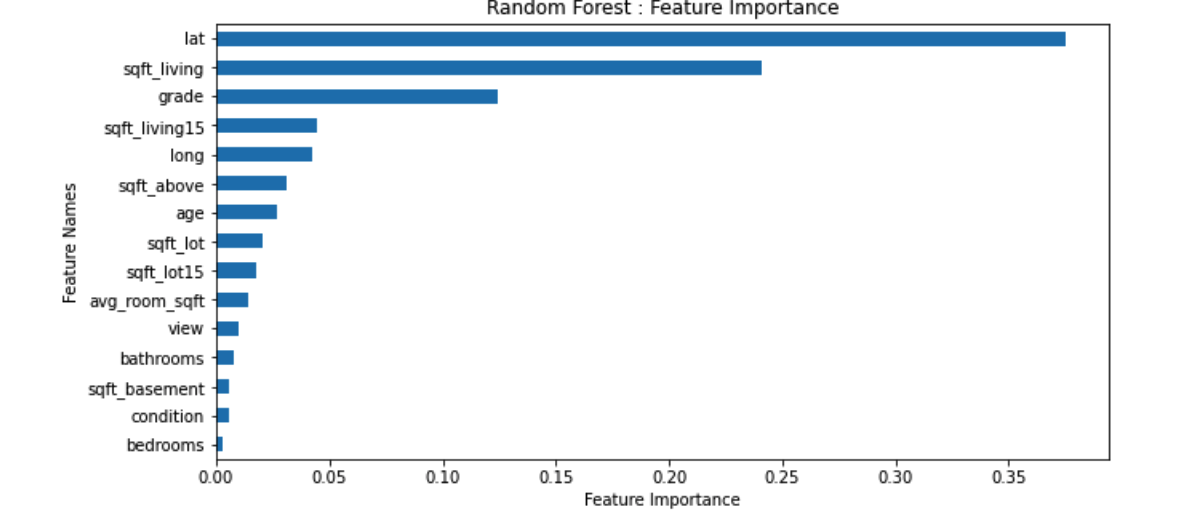


We can also detect a good connection between the predicted and actual values by plotting the residuals, or the predicted and actual values. The points must be relatively near to the fitted line for a satisfactory fit. Few outliers indicate that some forecasts are wrong, but the positive correlation shows that there is generally good fit.



After the model has been fitted, we now search for the features that have contributed most to this model. When investing in buying or selling a home in King County, investors can take this into consideration and pay more attention to these aspects.

We can see those factors like the house's latitude and longitude, or its location, size of its interior living space, grade, and condition, as well as the typical size of a room, are crucial to consider when estimating the price of homes in King County.



**Conclusion/Recommendation**

The best indicators of a home's price in King County are its location, its square footage of living space, its grade, and the size of its neighbors' houses taken together. The model can considerably assist real estate investors in taking all of these things into consideration when purchasing a home. The approach does, however, have a constraint in that it only applies to residences that cost less than $1,000,000 and typically have 2–6 bedrooms. Further research on this data revealed that the machine learning model creates significant residual errors since there is a lack of data on expensive residences.

**Assumptions and Limitations**

Duplicates and a small enough dataset. After I perform the EDA and massaging the data, my model should be useful. Due to the fact that we will be dramatically reducing the number of training points, our accuracy may be excessively high.

**Ethical Considerations**

Many groups, including people of color, those with disabilities, and other protected classes, are restricted in their neighborhood and housing choices by public policies and private discrimination that concentrate poverty, block investments from going into struggling areas, or make housing unaffordable in areas with good schools and access to good jobs. We may nevertheless build a deep picture of the housing circumstances, quality, affordability, and choice in a neighborhood even while there aren't many direct fair housing complaints or lawsuit data for tiny areas. This section examines housing in terms of property attributes, tenure, lending, affordability, subsidized housing, blight and abandonment, and describes data sources for each component. Real estate professionals collect, maintain, and disseminate customer data utilizing certain technologies that link the ideal features for people looking for a home or those wanting to create one using physical features, location, and other relevant criteria. Customers' personal information, including images of their own homes or homes with qualities they value, would also be used by AI techniques to create homes that have all of their preferences taken into account. It is usually a touchy subject when personal information is discussed as a significant factor

**Challenges/Issues**

* Duplicate data.
* Unavailability of real time dataset
* Data may highly skewed as only very minimal housing data.
* Python package issues
* Data accuracy and verification is not defined.
* Kaggle dataset may be too small to predict.

**Future Uses/Additional Applications**

On these data, four supervised machine learning models are being evaluated. In order to further reduce our mistake rate, we can try to utilize deep learning models like neural network models as our next step. Additionally, the algorithm can only anticipate homes under $1 million in value due to the lack of data on high-priced residences. We can get around this restriction by collecting more data, which can then be used to estimate pricey, high-priced homes.

**References**

* Akinsomi, Omolokolade, Aye Godness, C. Babalos Vassilios, Fotini Economou, and Rangan Gupta. 2016. Real estate returns predictability

<https://link.springer.com/article/10.1007/s00181-015-1037-5>

* <https://www.hud.gov/program_offices/fair_housing_equal_opp/online-complaint>

**Appendix**

