**HOUSING PRICE PREDICTION USING MULTIPLE LINEAR REGRESSION**

# Business Understanding

Buying a new house is, without a doubt, a daunting process, both for the agency and the individual. The journey is often filled with lots of fraud and the research needed to strike a good deal is exhausting.

Our client is a real estate agency who plans on expanding their real estate business within the northwestern county.

They want us to help them investigate the relationship between the price of their properties and several predictor variables including square footage, number of bathrooms, and number of bedrooms so as to identify the key factors that influence the sale price of a house.

# Research Question

What factors influence housing prices in King County, and how well can a multiple linear regression model predict housing prices based on those factors?

# Objectives

* Perform Exploratory Data Analysis of the dataset to understand each variable, and their relationship among each other, and to the target the housing sale price.
* To identify the variables affecting house prices.
* To create a linear model that quantitatively relates house prices with variables
* Evaluating our model on how well these variables can predict house prices.
* Come up with recommendations upon interpreting our regression model.

# Data Understanding

## **Data Source**

The data used in this project was obtained from Canvas Project Description. It was shared on a shared repository where it can be pulled from the repository in the data folder consisting of a csv file named kc\_house\_data.csv.

## **Data Description**

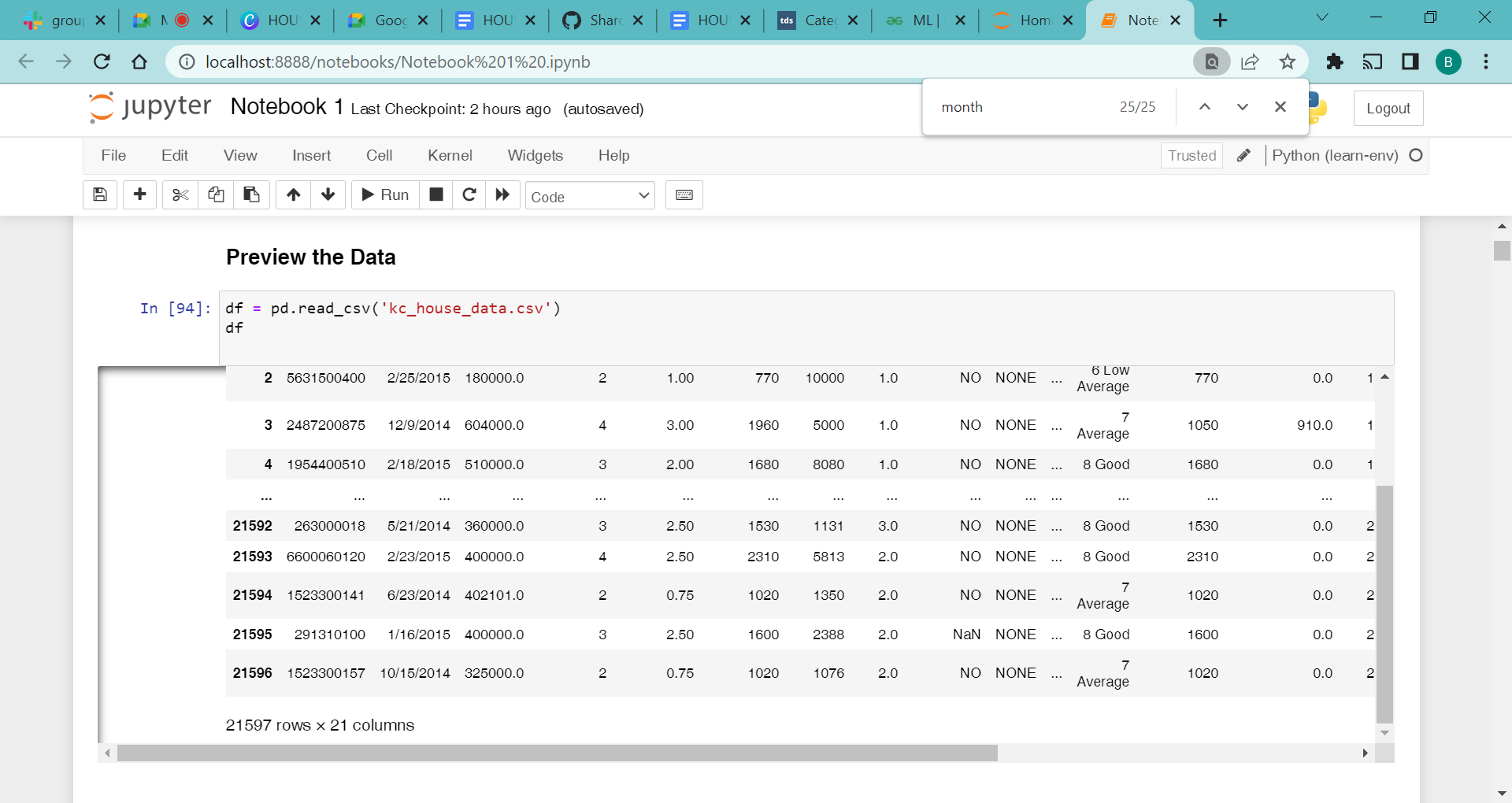
The data for this study includes information on (no. of homes) properties sold in King County. Our data has 21597 rows and 20 columns.

Below are the Column Names and Description for King County Data Set:

1. **`id`** - Unique identifier for a house
2. **`date`** - Date house was sold
3. **`price`** - Sale price (prediction target)
4. **`bedrooms`** - Number of bedrooms
5. **`bathrooms`** - Number of bathrooms
6. **`sqft\_living`** - Square footage of living space in the home
7. **`sqft\_lot`** - Square footage of the lot
8. **`floors`** - Number of floors (levels) in house
9. **`waterfront`** - Whether the house is on a waterfront
10. **`view`** - Quality of view from house
11. **`condition`** - How good the overall condition of the house is. Related to the maintenance of the house.
12. **`grade`** - Overall grade of the house. Related to the construction and design of the house.
13. **`sqft\_above`** - Square footage of house apart from basement
14. **`sqft\_basement`** - Square footage of the basement
15. **`yr\_built`** - Year when the house was built
16. **`yr\_renovated`** - Year when the house was renovated
17. **`zipcode`** - ZIP Code used by the the United States Postal Service
18. **`lat`** - Latitude coordinate
19. **`long`** - Longitude coordinate
20. **`sqft\_living15`** - The square footage of interior housing living space for the nearest 15 neighbors
21. **`sqft\_lot15`** - The square footage of the land lots of the nearest 15 neighbors

# Data Preparation - Loading the Data

The data was stored in a CSV file and loaded into Python using the pandas library as shown below:



# Cleaning the data

We began by cleaning the data to ensure that our dataset is consistent and accurate.

Clean data ensures that our results are not only accurate but also reliable. Our data cleaning process involved the following steps:

### **Handling missing values**

We identified any missing values in our data and handled them by either

replacing or dropping based on its importance.

### **Removing outliers**

By checking for outliers in our data, we identified strange values and treated them by dropping them.

### **Converting data types**

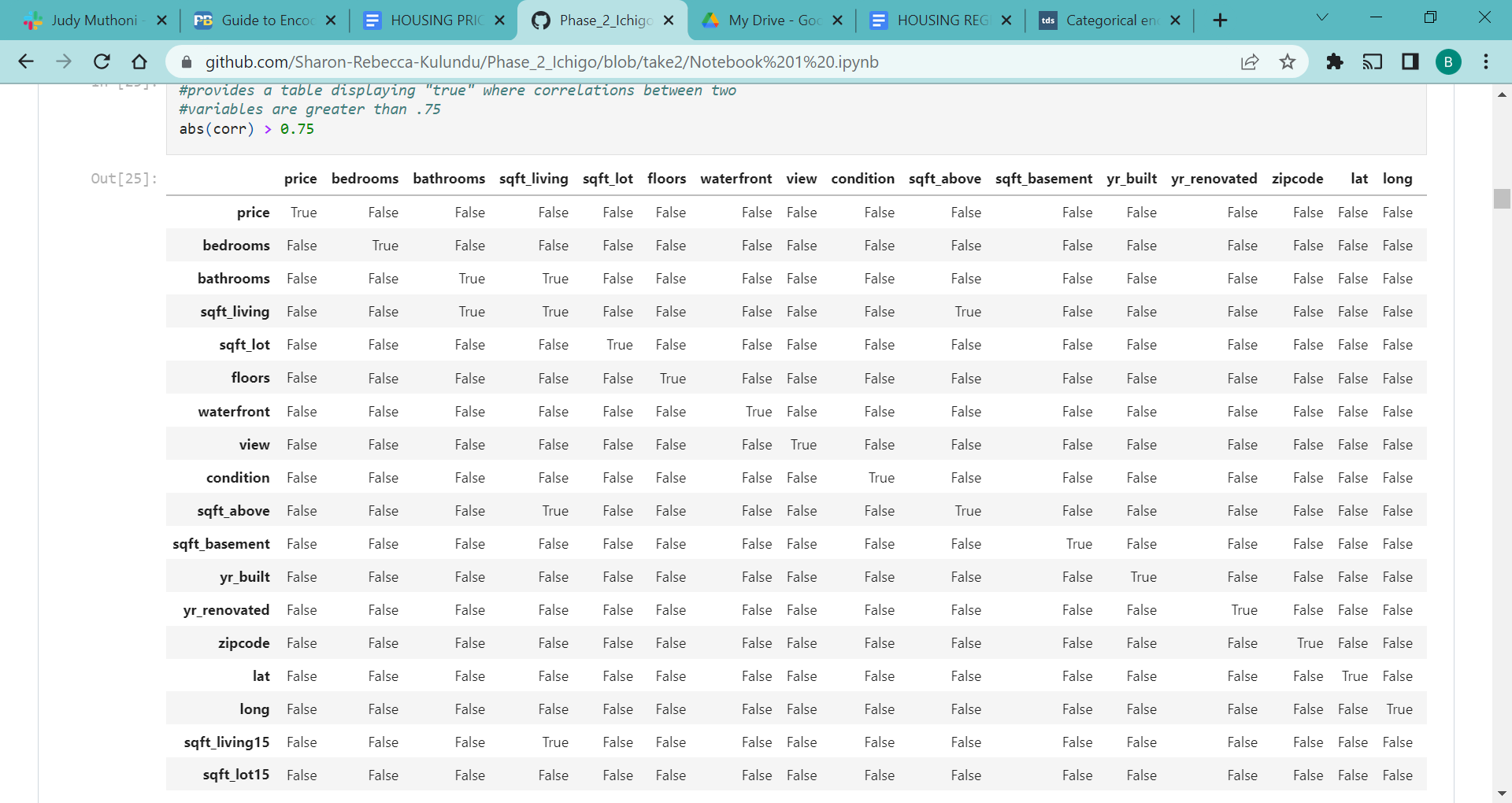
We checked the data types of all our variables and converted them as needed based on what works best for the modeling process.

### **Removing Duplicates**

We checked the data for duplicates and kept the last and dropped the rest mainly on the ID column

### **Checking for Multicollinearity**

The data had a lot of data and therefore there will be a lot of multicollinearity between the columns. A true false correlation of over .75 was used and columns dropped.



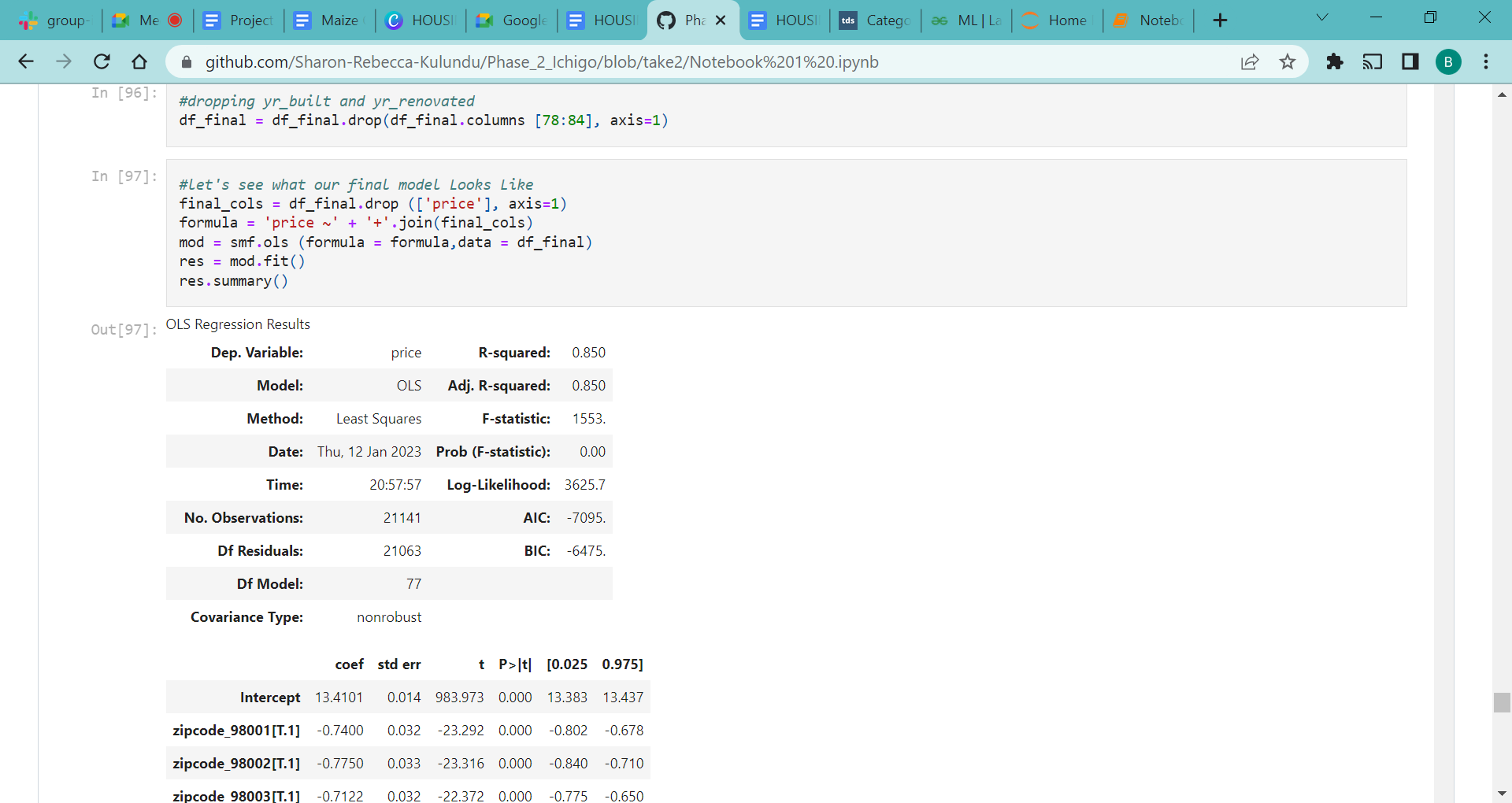
# Modeling

First, we split the data into categorical and numerical data since they needed to be analyzed differently. The categorical data was then binned , encoded and normalized.

A multilinear regression model was used to predict the sale of property based on various variables.

The model was trained on the training set and evaluated on the set using Mean Squared Error and R-squared as the evaluation metrics for success.

The model was evaluated using mean squared error and R^2 score. It gave a mean squared error of 0.04 on the test set, and was able to achieve an R-squared score of 0.85, which shows that 85% of the variance in the data is explained by the predictors, thus a good fit. This indicates that the model is able to accurately predict the sale price of a property based on the input variables.



# Conclusion

Our final model only includes the predictor variables that had r2 values of .1 or higher

The final predictors were sqft\_living15 with an r2 value of 0.385, sqft\_above with 0.362, bathrooms with 0.304, lat with 0.202, bedrooms with 0.123,

sqft\_basement with .101, and zipcode with .531. Floors and view both had r2 values of .097 which we rounded up to .1.

All of our predictors are statistically significant with p-values smaller than .05

Our r-squared value overall was .85

# Recommendations:

Square footage is the best predictor of a house's price in King County and homeowners who are interested in selling their homes at a higher price should focus on expanding it.

When expanding square footage, homeowners should consider building additional bathrooms and bedrooms, as this analysis suggests that number of bathrooms and bedrooms is highly positively related to price.

Use the current model to predict price ranges by zip code so that home buyers can see what their macro options are for housing by location and price. The model can also provide general price modifications to a house such as knowing how much an additional bedroom or bathroom would cost, and how the price changes depending on the addition or subtraction of square footage.

# Limitations and Next Steps To Take

The accuracy of our linear regression models might not be sufficient for launching the business, but they are a step in the right direction. An expert could do a better job making predictions, but of course not at the scale and speed of our models.

To make further improvements to our model, we need additional data. To start, better location information, specifically neighborhood level data, historical price information for each house and real time localized information about the housing market.