

The background features a stylized illustration of several houses in various colors (red, orange, green, teal, grey) with white windows and chimneys. A large, semi-transparent teal circle is positioned on the right side, partially overlapping the houses. The text is overlaid on this background.

Multiple Linear Regression: Predicting King County, WA, Housing Prices

GROUP 2 PHASE 2 PROJECT

Tobias, Phelix, Wambui, Muthoni, John, Peggy & Catherine

OVERVIEW

- 
- 
- ❑ The project aims to help the real estate agencies and homeowners in making informed decisions about home selling and buying by utilizing the King County House Sales dataset.
 - ❑ By analyzing and modeling the dataset, we can determine the influence of various factors on house prices, ultimately providing valuable insights to real estate agencies and homeowners regarding the potential change in the estimated value of their homes through different choices.

OBJECTIVES

1. To understand which factors determines the price of a home.
2. To understand how square feet living affect the value of a home.
3. To explore how condition affect the price of a home.
4. To explore features which decrease and increase value of the house.



RESEARCH QUESTIONS

1. Which factors determines the price of a home?
2. How square feet living affect the value of a home?
3. How condition affect the price of a home?
4. Which features will decrease and increase value of the house?



DATA UNDERSTANDING



DATA SOURCE

- ❑ This project uses the King County House Sales dataset. The dataset contains house sale prices for King County, which includes Seattle. It also includes homes sold between May 2014 and May 2015.

DESCRIPTION

- ❑ The dataset includes various features such as the number of bedrooms, bathrooms, square footage of living space, lot size, number of floors, waterfront status, view quality, condition, grade, presence of a basement, year built and renovated, ZIP code, latitude, longitude, and information about the nearest 15 neighbors.
- ❑ The dataset has both numerical and categorical columns. date, waterfront, view, condition, grade and sqft_basement are categorical columns while the rest are numerical columns.



DATA PREPARATION AND CLEANING

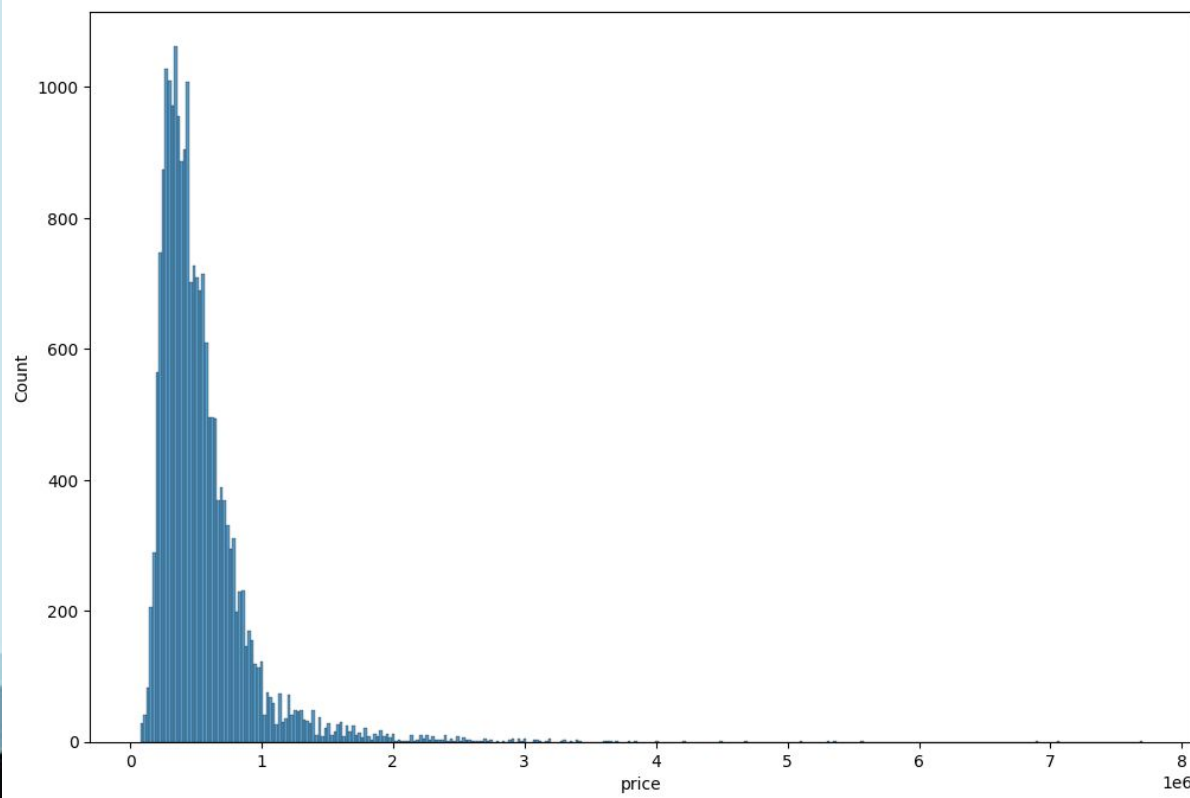
The data was cleaned through the following steps.

1. Checking for missing values where we found that the view, waterfront and year renovated column had missing values. We decided to drop the view column and replace Nans with mode in the waterfront column.
2. Checking for duplicates and outliers
3. Feature engineering using the date column to create a new column named age which shows the age of the house.
4. Dealing with categorical features through one hot encoding



EDA

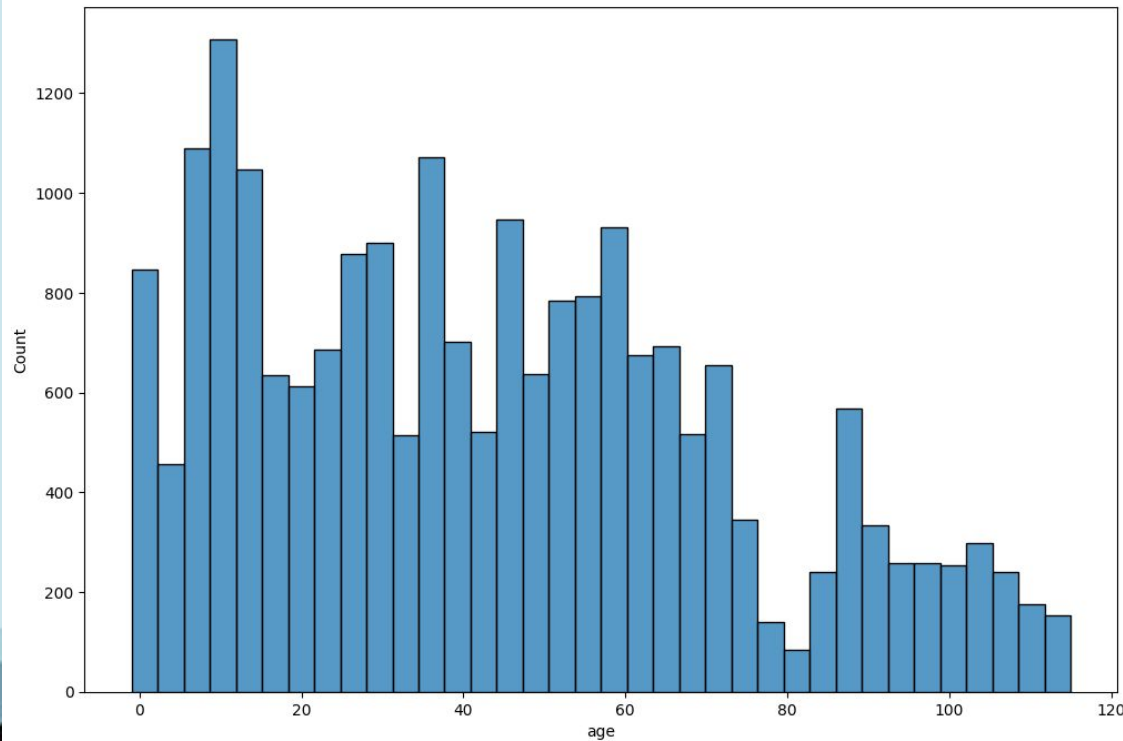
Distribution of Price



- ❑ The price of the majority of the houses ranges between 200,000 to 1,200,000.
- ❑ Some house have very prices of more than 4,000,000.
- ❑ There are some outliers in the price data

EDA

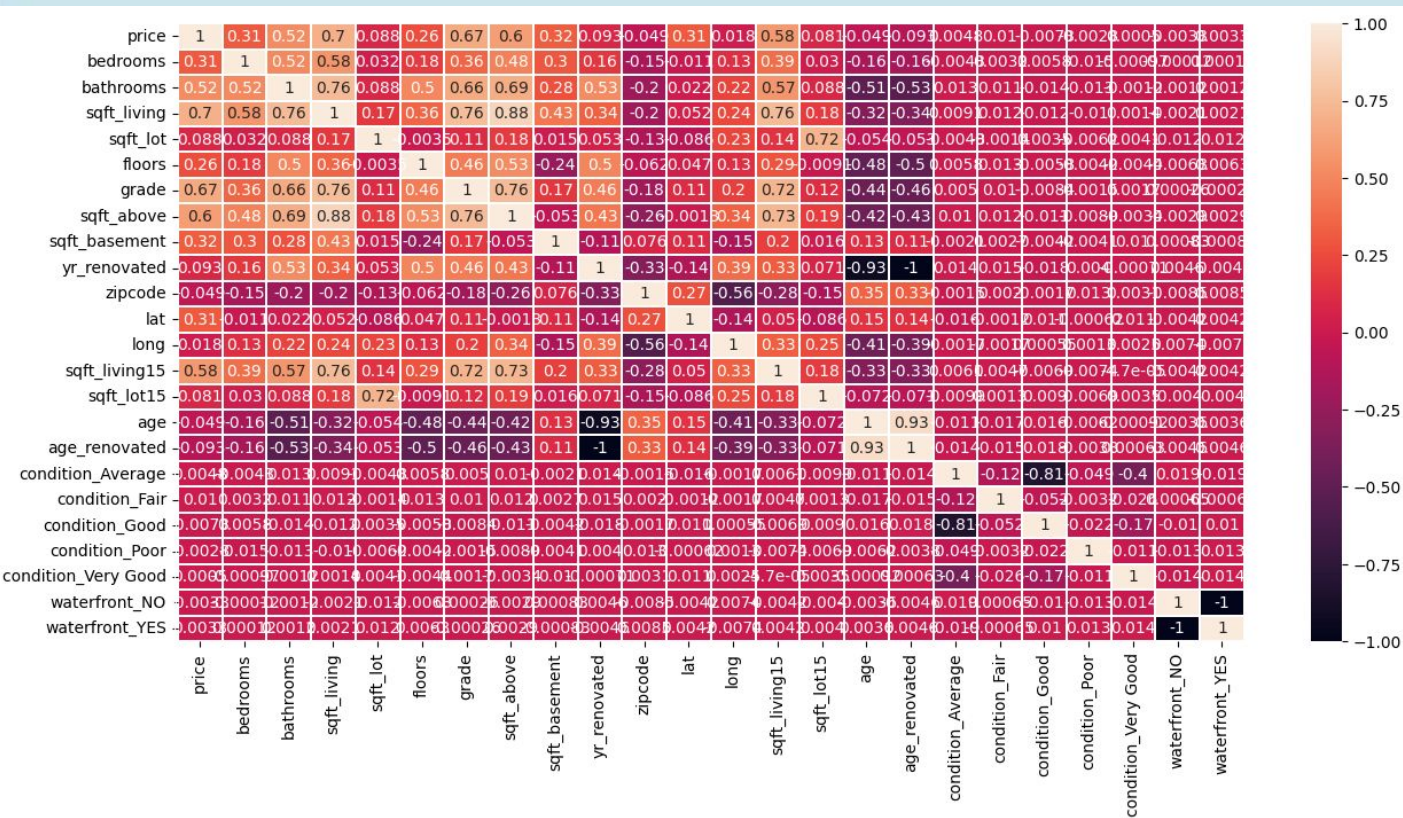
Distribution of Age



- ❑ The number of houses sold generally decreases with the increase in the number of years from the year built to the year sold.

EDA

Correlation Matrix



According to this heatmap, there are some variables which are highly correlated which will be considered in linear regression



MODELLING

The following is a list of all models built and general description of changes between each model.



Model 1: All Features

Our first model has an adjusted r-squared of .660 All features with p_values that are significant, let's check our residuals.

```
Dep. Variable: price R-squared: 0.660
Model: OLS Adj. R-squared: 0.660
Method: Least Squares F-statistic: 1964.
Date: Fri, 02 Jun 2023 Prob (F-statistic): 0.00
Time: 22:59:57 Log-Likelihood: -2.9085e+05
No. Observations: 21244 AIC: 5.818e+05
Df Residuals: 21222 BIC: 5.819e+05
Df Model: 21
Covariance Type: nonrobust
```

	coef	std err	t	P> t	[0.025	0.975]
const	-3.371e+07	4.13e+06	-8.159	0.000	-4.18e+07	-2.56e+07
bedrooms	-4.55e+04	2024.056	-22.479	0.000	-4.95e+04	-4.15e+04
bathrooms	4.745e+04	3497.152	13.569	0.000	4.06e+04	5.43e+04
sqft_living	109.1784	22.653	4.820	0.000	64.777	153.580
sqft_lot	0.1422	0.051	2.792	0.005	0.042	0.242
floors	9096.6499	3867.668	2.352	0.019	1515.728	1.67e+04
grade	1.033e+05	2309.367	44.711	0.000	9.87e+04	1.08e+05
sqft_above	76.7000	22.666	3.384	0.001	32.272	121.128
sqft_basement	74.6117	22.657	3.293	0.001	30.202	119.021
yr_renovated	2.549e+04	3143.506	8.108	0.000	1.93e+04	3.17e+04
zipcode	-522.2306	34.950	-14.942	0.000	-590.736	-453.725
lat	5.477e+05	1.14e+04	48.031	0.000	5.25e+05	5.7e+05
long	-2.482e+05	1.41e+04	-17.635	0.000	-2.76e+05	-2.21e+05
sqft_living15	38.0384	3.657	10.401	0.000	30.870	45.207
sqft_lot15	-0.3115	0.078	-3.992	0.000	-0.464	-0.159
age	3854.6952	135.390	28.471	0.000	3589.321	4120.069
age_renovated	2.487e+04	3144.132	7.909	0.000	1.87e+04	3.1e+04
condition_Average	-6.755e+06	8.26e+05	-8.176	0.000	-8.37e+06	-5.14e+06
condition_Fair	-6.735e+06	8.26e+05	-8.150	0.000	-8.35e+06	-5.11e+06
condition_Good	-6.757e+06	8.26e+05	-8.179	0.000	-8.38e+06	-5.14e+06
condition_Poor	-6.703e+06	8.27e+05	-8.102	0.000	-8.32e+06	-5.08e+06
condition_Very Good	-6.758e+06	8.26e+05	-8.180	0.000	-8.38e+06	-5.14e+06
waterfront_NO	-1.685e+07	2.07e+06	-8.159	0.000	-2.09e+07	-1.28e+07
waterfront_YES	-1.685e+07	2.07e+06	-8.159	0.000	-2.09e+07	-1.28e+07
Omnibus:	19043.086	Durbin-Watson:	1.309			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1873733.100			
Skew:	3.914	Prob(JB):	0.00			
Kurtosis:	48.338	Cond. No.	6.88e+20			

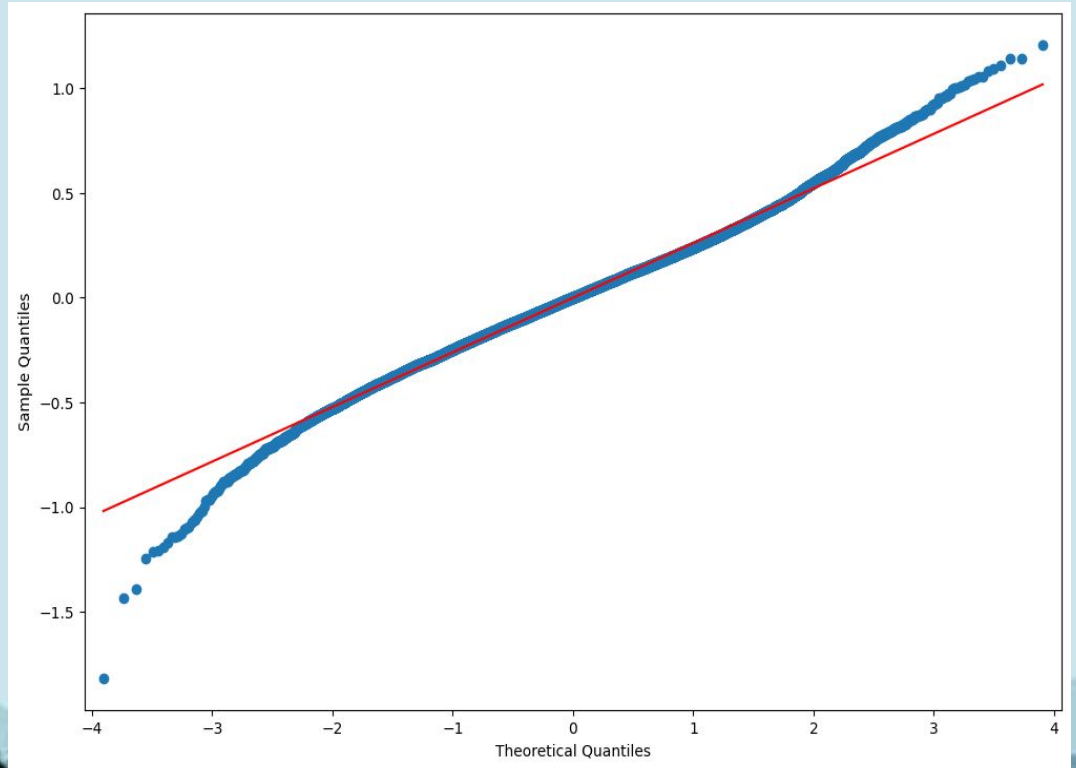
Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The smallest eigenvalue is 4.55e-28. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

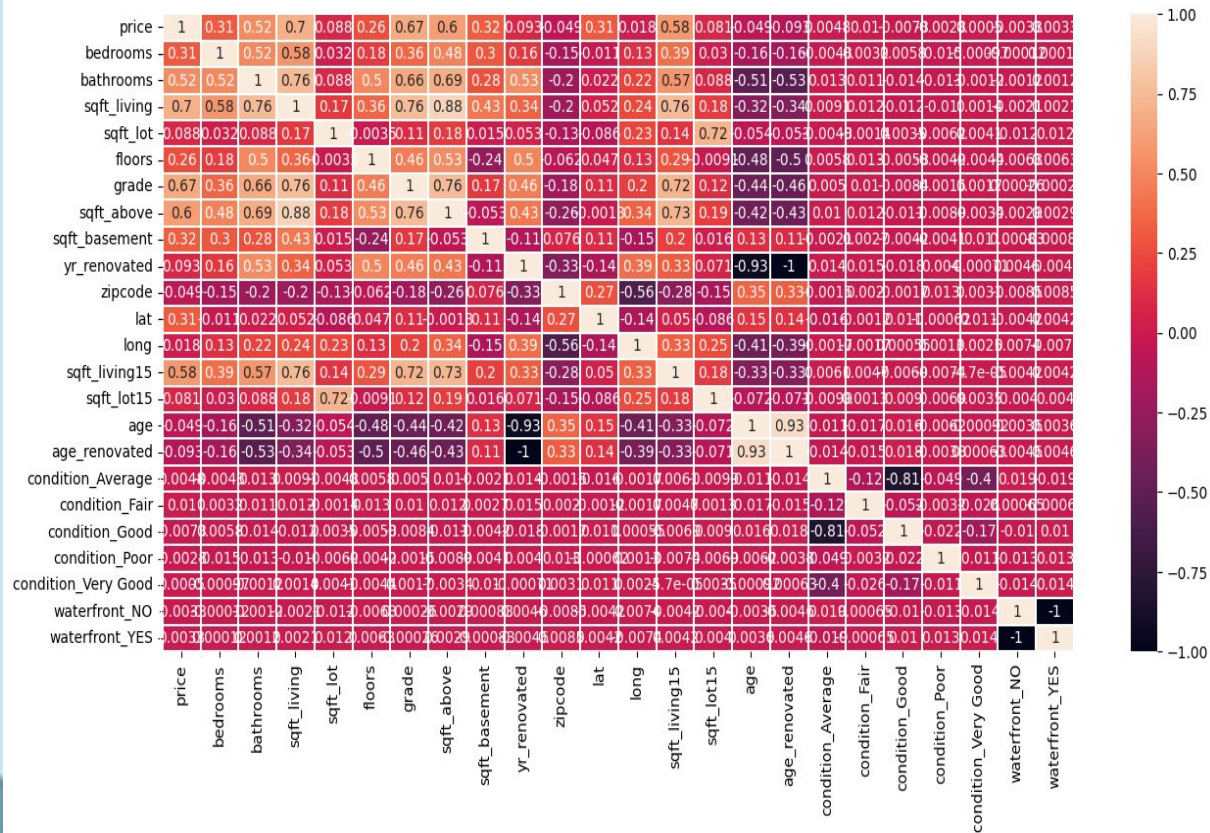
Model 2: with log transformed y variables

After log transformation of the dependent variables residuals are much closer to a normal distribution



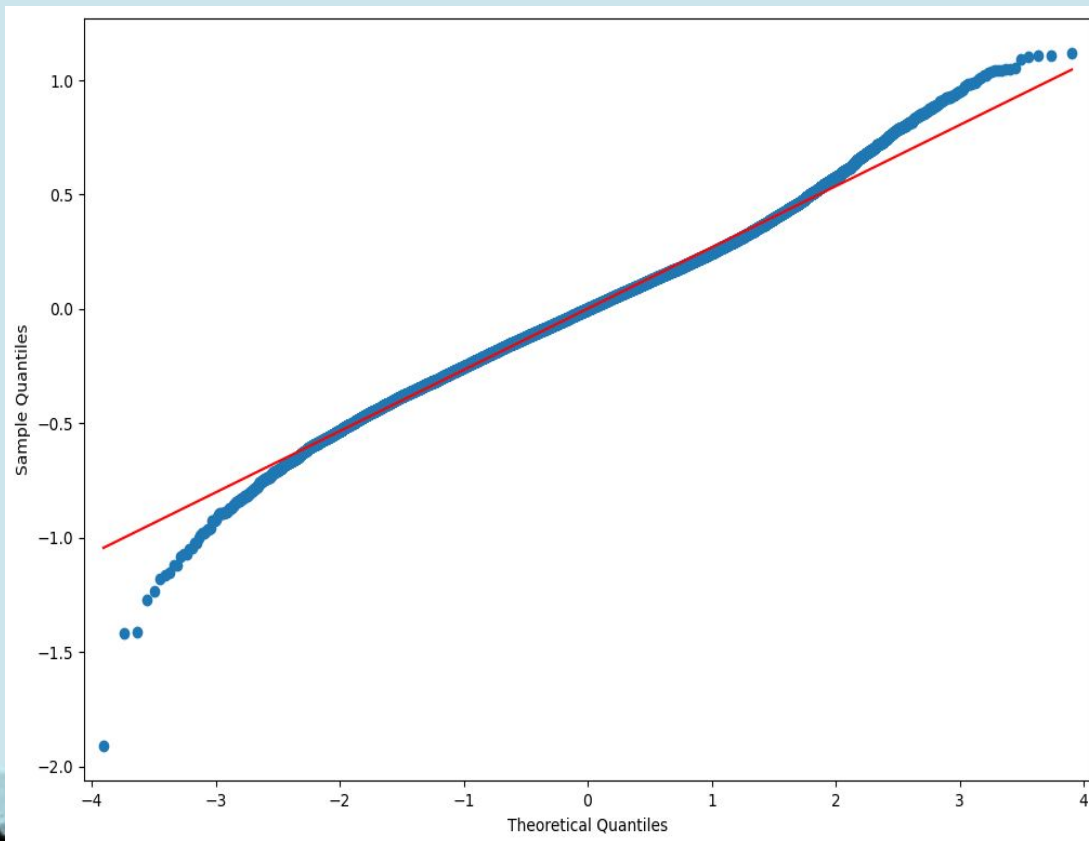
Model 3: Dealing with Multicollinearity

Conclusion: There are several features that seem to have multicollinearity. Rather than just dropping some of these features, let's first look at the variance inflation factor to understand the severity of the multicollinearity.



Model 4: Dropping Insignificant Features

Conclusion: Our adjusted R squared still stays the same at .740 and all features are significant. Next, we will further refine our data by removing additional, potential outliers



Model 5: Standardizing Features

Interpretation: Since the p-value (0.0024) is less than the significance level (e.g., 0.05), we can reject the null hypothesis. This suggests that there is significant evidence of heteroscedasticity in the data. It implies that the variance of the errors is not constant across the range of the predictors.

In summary, F statistic and p-value, there is evidence of heteroscedasticity in the data, indicating that the assumption of constant error variance may not hold.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.740
Model:	OLS	Adj. R-squared:	0.740
Method:	Least Squares	F-statistic:	4654.
Date:	Fri, 02 Jun 2023	Prob (F-statistic):	0.00
Time:	23:00:29	Log-Likelihood:	-2145.4
No. Observations:	21244	AIC:	4319.
Df Residuals:	21230	BIC:	4430.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	-19.1124	2.512	-7.609	0.000	-24.036	-14.189
bedrooms	-0.0198	0.003	-7.819	0.000	-0.025	-0.015
bathrooms	0.0725	0.004	16.586	0.000	0.064	0.081
sqft_living	0.0002	5.77e-06	35.087	0.000	0.000	0.000
sqft_lot	4.287e-07	4.63e-08	9.256	0.000	3.38e-07	5.19e-07
floors	0.0657	0.005	13.593	0.000	0.056	0.075
grade	0.1609	0.003	55.785	0.000	0.155	0.167
sqft_above	-5.477e-05	5.75e-06	-9.523	0.000	-6.6e-05	-4.35e-05
zipcode	-0.0006	4.37e-05	-12.865	0.000	-0.001	-0.000
lat	1.3575	0.014	95.255	0.000	1.330	1.385
long	-0.2469	0.017	-14.131	0.000	-0.281	-0.213
sqft_living15	0.0001	4.57e-06	24.031	0.000	0.000	0.000
age_renovated	0.0037	8.93e-05	41.553	0.000	0.004	0.004
waterfront_NO	-9.5608	1.256	-7.614	0.000	-12.022	-7.099
waterfront_YES	-9.5516	1.256	-7.605	0.000	-12.013	-7.090

Omnibus:	426.976	Durbin-Watson:	1.143
Prob(Omnibus):	0.000	Jarque-Bera (JB):	873.336
Skew:	0.089	Prob(JB):	2.28e-190
Kurtosis:	3.977	Cond. No.	9.00e+20

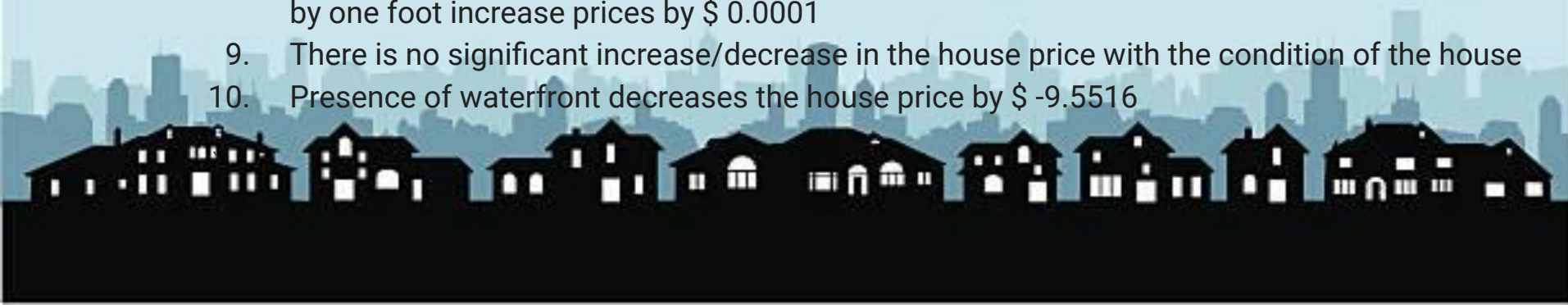
Notes:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The smallest eigenvalue is 2.6e-28. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.


REGRESSION RESULTS

Analyzing the models the following conclusion can be made.

1. An increase with one bedroom decreases the house sale by \$ 0.002.
2. An increase with one bathroom increases the house price by \$ 0.0725
3. An increase in Square footage of the home by one square foot increases the price of the house by \$ 0.0002
4. An increase in Square footage of the by one square feet decreases the house price by \$ 4.287e-07
5. An increase in floors by one increases price by \$0.0657
6. An increase in grade rating by one increases the price by \$ 0.1609
7. An increase in one square foot from basement decrease price by \$ -5.477e-05
8. An increase in Square footage of interior housing living space for the nearest 15 neighbors by one foot increase prices by \$ 0.0001
9. There is no significant increase/decrease in the house price with the condition of the house
10. Presence of waterfront decreases the house price by \$ -9.5516



RECOMMENDATIONS

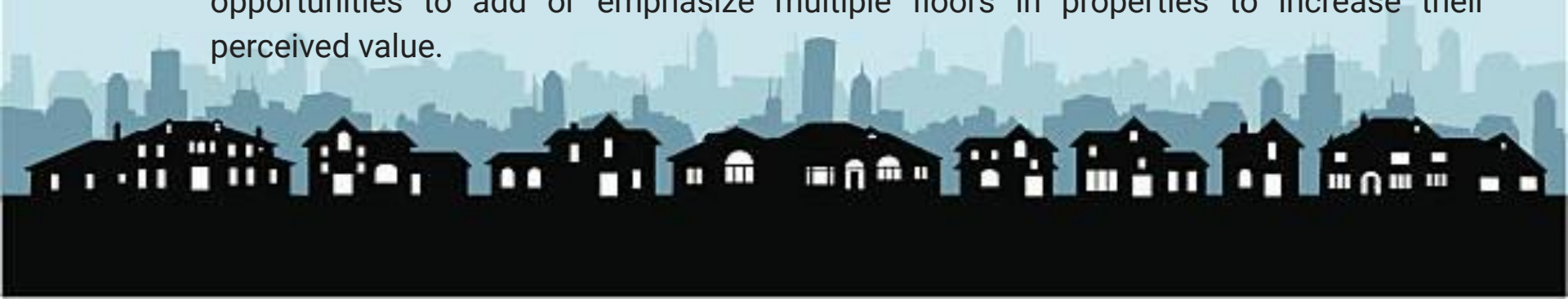


1.Bedrooms: Consider the impact of the number of bedrooms on the house prices, and buyers preferences.


3 Bathrooms: If feasible, consider adding or upgrading bathrooms in properties to attract buyers who value this feature.

4.Square Footage: Pay attention to the square footage of the living space and lot size. Consider the preferences of the target market and the local real estate trends to find the right balance.

5.Floors: Houses with multiple floors tend to have higher prices. If feasible, explore opportunities to add or emphasize multiple floors in properties to increase their perceived value.



RECOMMENDATIONS

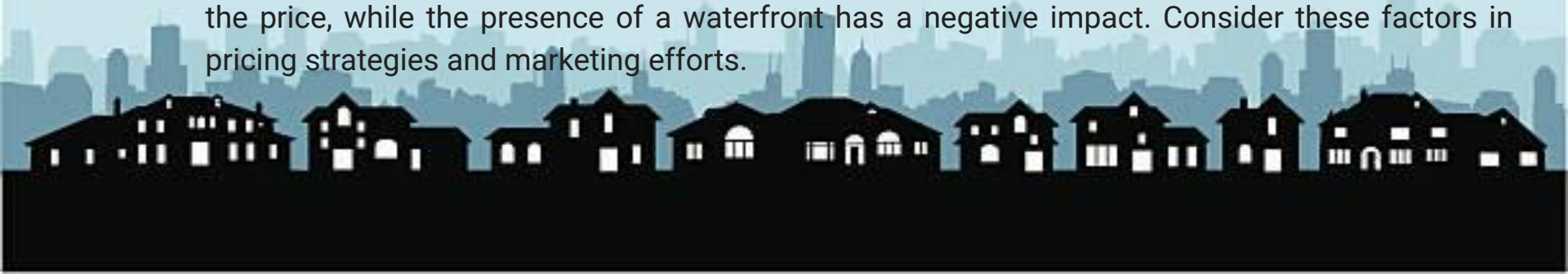


6. Grade Rating: Consider improving the grade rating through renovations, upgrades, or focusing on properties with higher grade ratings.

7. Basement: The square footage of the basement has a small negative impact on the house price. Evaluate the demand for basements in the target market and consider their condition and potential for improvement.

8. Neighborhood: Consider the impact of the square footage of interior living space for the nearest 15 neighbors on the house price. Properties with larger nearby living spaces may attract higher prices. Evaluate the desirability and market demand for properties in different neighborhoods.

9. Condition and Waterfront: The condition of the house does not have a significant impact on the price, while the presence of a waterfront has a negative impact. Consider these factors in pricing strategies and marketing efforts.



-
- A horizontal banner image featuring a dark silhouette of a row of houses in the foreground. Behind the houses is a light blue silhouette of a city skyline with various skyscrapers and buildings. The entire image is set against a white background.

THANK YOU



Questions

Contact information:

Catherine Gakii: <https://ke.linkedin.com/in/gakii-catherine-234a89b7>

Phelix Okumu: <https://www.linkedin.com/in/phelix-okumu-631baa78>

Muthoni Kahura: <https://www.linkedin.com/in/muthoni-kahura/>

Wambui Thuku: <https://www.linkedin.com/in/wambui-thuku-24228675>

John Gikonyo: <https://www.linkedin.com/in/john-gikonyo-16b719ba/>

Tobias Ng'ong'a: <https://www.linkedin.com/in/tobias-ng-ong-a-6ba471a9/>

Peggy Obam: <http://www.linkedin.com/in/peggy-obam-b96b50187>

