# Introduction

## Business Overview:

Mzalendo ltd is a housing stakeholder that gives advice to homeowners so that they can buy or sell homes.

The company wants to help homeowners to be able to predict the current and future prices of their houses depending on different features.

### Challenges:

1. Data quality issues: The dataset may contain errors, inconsistencies, or missing values, which can impact the accuracy of the regression analysis.

2. Complexity of the housing market: The housing market can be complex and dynamic, and factors such as economic conditions, interest rates, and demographic changes can all impact home values. This can make it challenging to accurately predict the impact of renovations on home values.

3. Limited resources: Building an accurate regression model can be time-consuming and resource-intensive, and may require specialized expertise in data analysis and statistics.

4. Communication with stakeholders: It is important to effectively communicate the findings and recommendations to the stakeholders, who may have varying levels of

technical expertise or understanding of the analysis. Clear and concise communication is essential to ensure that the stakeholders can make informed decisions based on the results.

## Proposed Solution:

Our proposed solution to address the business problem and meet the project objectives is to use multiple linear regression modeling to analyze the King County House Sales dataset, which contains information on various factors that can influence the value of a home.

## Conclusion:

We aim to use multiple linear regression modeling,So that we can analyze the King County House Sales dataset to identify the most important factors that impact the value of a home, build an accurate regression model to predict home values, and provide guidance to homeowners on how to increase the estimated value of their homes through renovations.

# Business Problem

The business problem for this project is to provide guidance to homeowners to determine house prices based on the house descriptions.

# Objectives

1. To identify the most important factors that affect the value of a home in a northwestern county, and to determine the extent to which these factors impact home values.
2. To build a multiple linear regression model that accurately predicts the value of a home based on these factors.
3. To evaluate the performance of the regression model using appropriate metrics, and to compare the results of different models to determine the best approach.
4. To communicate the findings and recommendations to the stakeholders in a clear and concise manner, and to ensure that they understand the implications of the results for their business.

# Analytical Questions

1. Which features have the strongest correlation with home sale price?
2. Are there any variables that need to be transformed or standardized before being used in the model?
3. How can we evaluate the performance of our model and ensure that it is not overfitting to the data?
4. How can we communicate the results of our analysis to stakeholders in a clear and actionable manner?

# Data Understanding

## Dealing with missing values

1. We found the water-front column with 11% of missing values  and the view columns with 63 missing values which we decided to fill with the mode.
2. Water-front and view are categorical variables so we checked the value counts and filled with the most occurring feature.
3. We dropped the yr_renoveted column with 18% of missing values since it was not useful to our model

## Data Analysis

We conducted univariate and bivariate analysis and visualized some variables.
We also conducted statistical analysis on the price column, conducted analysis on outliers, missing values and null values.

# Modelling

## Metrics of success

The metric of success we used are the R-squared adjusted and the MSE because they provide a more interpretable measure of the goodness of fit of the model. Further,

R-squared also takes into account the number of variables included in the model, which can be helpful in comparing models with different numbers of variables.

## Linear regression model

Overall this model is statistically significant and explains about 47% of the variance in price.

The model is off by about $160,438.91.

The intercept is at about $34k. This means that a zero square foot living house would sell for $34k.

The coefficient for square foot living is about $239. This means for each additional squarefoot, the house costs about $239 more.

The model explains 47% of the variance in  price but we can improve the model to 50% by doing a multi regression model

# Multiple regression model

## One hot encoded with Condition column

Overall this model is statistically significant and explains about 53% of the variance in price.

The model is off by about $149,576.71.

The intercept is at about $5,114,000.

The coefficient for square foot living is about $256, bathrooms $73110, one rating of average condition is $19120, good condition $16690, very good condition $51530 .

The model explains 53% of the variance in  price but we can improve the model to 60% by doing log transformation on the multi regression model

## Log transformed model

The model is statistically significant

The model explains about 50% of the variance in price. However the model has a lower adjusted r squared value compared to the previous model

To improve the model, we one hot code the grade of the house column to get to the model to 60%

One hot encoded with Grade column

Overall this model is statistically significant and explains about 63% of the variance in price.

The model is off by about $130,794.15.

The intercept is at about $6,956,000.

The coefficient for square foot living is about $118, bathrooms $58970, grade of low grade is $39730, fair $18260, low average $54850, average $147700, good $263000, better $430000 , very good  $606900, excellent $789100, luxury $999400, mansion $1519000.

The model explains 63% of the variance in  price which has attained our target of a model explaining 60% of variance.

The price equation for this models is

price= 118sqftliv + 58970bathroom +39730lowgrade +18260fair + 54850lowaverage +147700average +263000good +430000better +606900verygood +789100excellent +999400luxury +1519000mansion -24170bedrooms -2528yrbuilt -0.1571sqftlot

## Train and Test model

According to the train tests performed the One hot encoded with Grade column is the best model because it has a lower MSE value which indicates better performance, and the difference between the train and test MSE values is also small which indicates that the model is not overfitting to the training data.

# Results

1. The base linear model had a adjusted R squared of 47%
2. The second model was a multiple regression model which was one hot coded with condition column and had an adjusted rsquared of 53%
3. The third mode was a log transformed model of the second model which had an adjusted rsquared of 50%. This model was lower compared to the second model and we opted an approach of one hot coding another categorical variable column called grade.
4. The fourth model was a multiple one hot coded model of grade column and had an adjusted r squared of 63% which achieved our target of above 60%.

5. All our models were statistically significant as they had a p value less than the significant value
6. The Rsquared of the model improved from the first to second but dropped on the third log transformed model but improved to the forth to 63%
7. We choose R-squared over RMSE because it provides a more interpretable measure of the goodness of fit of the model. Further, R-squared also takes into account the number of variables included in the model, which can be helpful in comparing models with different numbers of variables.

# Recommendations

1. Based on our findings, we recommend that prospective homebuyers focus on properties with larger living spaces and more bedrooms and bathrooms, as these tend to have higher values. Additionally, we recommend that real estate agents use our model to help set appropriate listing prices for homes, taking into account the features of the property.

# Next Steps

1. Limited variables: The dataset used in this analysis only included a limited number of variables. There could be other variables that could also impact housing prices that were not included in the dataset.

2. Data quality: The dataset used in this analysis is assumed to be accurate, complete, and representative of the population. However, there may be errors or biases in the data that were not detected.

3. Model assumptions: The regression models used in this analysis make certain assumptions, such as linearity, independence, and normality. Violations of these assumptions could impact the accuracy and reliability of the results.

4. Limited geographic scope: The dataset used in this analysis only covers housing prices in King County, Washington. The results may not be generalizable to other locations.

5. Timeframe: The dataset used in this analysis only covers housing prices from May 2014 to May 2015. Housing prices may have changed since that time, and the results may not be applicable to the current housing market

6. Log transformation can be performed on variables that have huge skewness and kurtosis and see if it improves the model.