




# NORTHWESTERN COUNTY HOUSING.

# OVERVIEW.

The King County housing data is a dataset that provides information about real estate properties in King County, Washington, USA. It is commonly used in the field of data analysis, particularly for studying housing market trends, predicting house prices, and exploring factors that influence property values. Some features included in the dataset that describe different aspects of residential houses. The main target variable is price which represents the monetary value which the property was sold. This makes the dataset suitable for regression analysis given the various features to be able to predict the value of a house. The dataset contains a lot of records of houses this allows for flexibility in model building and statistical analysis.



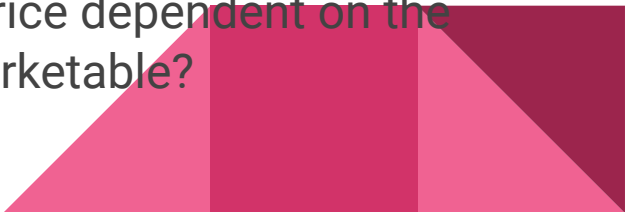
# BUSINESS UNDERSTANDING.

## **DATA ANALYTICS QUESTION:**

The data used in this analysis and the recommendations that come up can be used as a guide for people looking to buy houses in Northwester county. The findings will help guide customers on which are the best features to look at when considering buying a house in the aforementioned county

## **PROBLEM STATEMENT:**

What are the most important features to look for when seeking to buy a new home. Do renovations increase the price of a house? Is price dependent on the location? What aspects generally make a house more marketable?



## METRIC OF SUCCESS:

This project will be considered a success if it can provide actionable insights that would help guide customers on making good selections for houses and improve customer satisfaction with their choices. More than likely, these insights are all influenced ample knowledge following CRISPDM methodology.

The model will be considered a success if the R-squared is between 50-80% as well as looking at other metrics of success like mean absolute error, mean squared error, root mean squared error and the mean absolute percentage error.

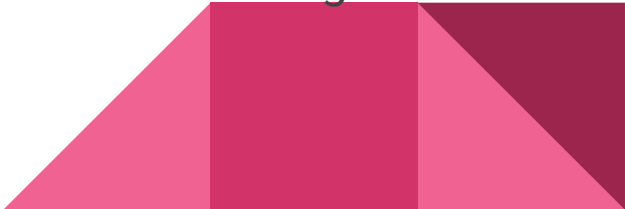


# OBJECTIVES.

## **Main Objective:**

- . To find out which aspects of a house make it more marketable.

## **Specific Objectives:**

- . Do renovations increase the value of a house.
  - . Do any specific features play a role in the price of a house.
  - . To find out what features are most important to consider when choosing a house.
- 

# DATA UNDERSTANDING.

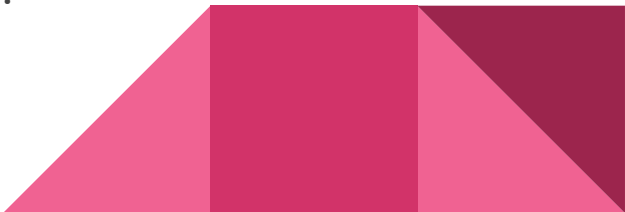
- .The dataset in this analysis is from: . kc\_house\_data.csv.
- .Ample description of the columns and what they describe about the entire dataset is contained in: . column\_names.md
- .The data is about housing in King county.
- . The data has both categorical and numeric data.
- .This columns were cleaned, wrangled and later used to build models to provide insights on what features are the best to use when selecting houses.



# MODELLING.

.Started off by building a simple linear regression model that takes in one independent variable and compares it against the target variable. The method used was OLS from sklearn and then prints out the summary to be able to interpret the results. The model yielded about 43% of variance in sales.

.Next, a multiple regression model that takes in 10 selected features that show the highest correlation with the target variable. The method used was splitting the data into testing and training data to be able to fit the model and print out the results. The model yielded about 52% of variance in sales.




. The final model took in all the independent variables which included dummy variables created from the categorical variables. The total variables used was 22 against the target variable. The method employed was OLS which I later printed out the summary. The model yielded about 66% of variance in sales which improved from the previous model.

. Of all the models the final model had the lowest  $r\_squared$  value indicating that the model had improved performance from the two preceding models.

. The next slides show the print out of the summary of the results of the three models that I used in the analysis.

Model 1, a baseline model that employs simple linear regression model while the other two were multiple linear regression models.





## Baseline model.

### OLS Regression Results

```
=====
Dep. Variable:          price    R-squared:                0.436
Model:                  OLS      Adj. R-squared:           0.436
Method:                 Least Squares    F-statistic:            1.653e+04
Date:                   Fri, 07 Jul 2023    Prob (F-statistic):      0.00
Time:                   16:18:53    Log-Likelihood:         -2.9878e+05
No. Observations:       21420    AIC:                    5.976e+05
Df Residuals:           21418    BIC:                    5.976e+05
Df Model:                1
Covariance Type:        nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
const	2.19e+05	3135.561	69.837	0.000	2.13e+05	2.25e+05
sqft_living	140.2088	1.091	128.558	0.000	138.071	142.347

```
=====
Omnibus:                16944.584    Durbin-Watson:           1.999
Prob(Omnibus):           0.000    Jarque-Bera (JB):        1456266.916
Skew:                    3.211    Prob(JB):                 0.00
Kurtosis:                42.880    Cond. No.                 4.77e+03
=====
```

## Model 2.(Multiple regression model)

This model employs ten selected independence variables against the target variable.

The mean absolute error is: {153352.6073863592}

The mean\_squared error is: {57127602809.03634}

The root mean\_squared error is: {239013.81300886426}

The r\_squared is: {0.5282468900343114}

The mean absolute percentage error is: {32.712066260090374}

# OLS Regression Results

```

=====
Dep. Variable:          price    R-squared:          0.663
Model:                  OLS      Adj. R-squared:       0.663
Method:                 Least Squares    F-statistic:       2009.
Date:                   Fri, 07 Jul 2023    Prob (F-statistic): 0.00
Time:                   16:20:44    Log-Likelihood:    -2.9324e+05
No. Observations:      21420    AIC:               5.865e+05
Df Residuals:          21398    BIC:               5.867e+05
Df Model:               21
Covariance Type:        nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
const	7.634e+06	1.29e+05	59.013	0.000	7.38e+06	7.89e+06
bedrooms	-1.1451	37.739	-0.030	0.976	-75.117	72.827
bathrooms	7.49e+04	3059.084	24.486	0.000	6.89e+04	8.09e+04
sqft_living	49.5518	1.365	36.295	0.000	46.876	52.228
sqft_lot	-1.2047	0.140	-8.606	0.000	-1.479	-0.930
floors	2.198e+04	3478.084	6.319	0.000	1.52e+04	2.88e+04
yr_built	-3557.9648	66.325	-53.645	0.000	-3687.966	-3427.963
waterfront_YES	7.709e+05	1.8e+04	42.939	0.000	7.36e+05	8.06e+05
condition_Fair	-4.01e+04	1.71e+04	-2.350	0.019	-7.35e+04	-6660.505
condition_Good	2.235e+04	3631.546	6.154	0.000	1.52e+04	2.95e+04
condition_Poor	-3.256e+04	4.07e+04	-0.800	0.424	-1.12e+05	4.72e+04
condition_Very Good	5.937e+04	5850.317	10.148	0.000	4.79e+04	7.08e+04
grade_11 Excellent	2.382e+05	1.3e+04	18.345	0.000	2.13e+05	2.64e+05
grade_12 Luxury	7.905e+05	2.44e+04	32.390	0.000	7.43e+05	8.38e+05




This snippet along with the one above shows a print out of the results of the final model.


```
grade_13 Mansion      2.269e+06    5.99e+04    37.842    0.000    2.15e+06    2.39e+06
grade_3 Poor          -6.885e+05    2.14e+05    -3.220    0.001   -1.11e+06   -2.69e+05
grade_4 Low           -6.398e+05    4.22e+04   -15.154    0.000   -7.23e+05   -5.57e+05
grade_5 Fair          -6.681e+05    1.67e+04   -40.125    0.000   -7.01e+05   -6.35e+05
grade_6 Low Average   -6.042e+05    9924.522   -60.879    0.000   -6.24e+05   -5.85e+05
grade_7 Average       -4.984e+05    8185.637   -60.886    0.000   -5.14e+05   -4.82e+05
grade_8 Good          -3.695e+05    7741.903   -47.725    0.000   -3.85e+05   -3.54e+05
grade_9 Better        -1.834e+05    7901.981   -23.215    0.000   -1.99e+05   -1.68e+05
=====
Omnibus:                13721.606    Durbin-Watson:                1.983
Prob(Omnibus):           0.000    Jarque-Bera (JB):             746219.240
Skew:                    2.413    Prob(JB):                     0.00
Kurtosis:                31.510    Cond. No.                     2.24e+06
=====
```

Notes:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 2.24e+06. This might indicate that there are strong multicollinearity or other numerical problems.

# MODELLING RESULTS.

- . The simple model linear regression model explains 43.6% using only one independent variable against the target variable.
  - . The second model which is a multiple regression model explains about 52% of variance in price. This model uses 10 selected independent variables against the target variable.
  - . The final model takes in all the independent variables against the target variable and explains about 66.3% of variance in the price of houses.
- 

- . All the models are statistically significant with a probability f-statistic of below the threshold alpha value of 0.05%.
  - . For the final model however, it takes in all the variables where others are more significant in making the model while others are not for example bedrooms and condition poor have a very high p statistic indicating that they are not really significant to the model.
  - . The mean squared error of the third model is the lowest showing an improvement from the previous model indicating better performance.
  - . The model features were selected on the basis of correlation which might have some bias thus making the model to not perform as effectively.
  - . The final model however did not indicate the model to be as normal as indicated by the qqplot.
- 

# CONCLUSIONS.

- . The model that explains the highest variance in sales was the final model that takes into account multiple independent variables inclusive of those selected. However, some variables like bedrooms and condition poor proved to be insignificant for the model since their p statistic values are above the alpha value of 0.05.
- . From the models above, it is important to be careful during feature selection in order to be able to choose features that make a model perform exceptionally rather than optimally.



- . Square foot living and lot seemed to be significant when predicting the amount a house sells for also, if the condition is good and it has a waterfront the house seems to be selling for more. Thus when choosing a house clients should be keen on these features in order to make the right selection.
- . The final model was the best performing however it lacked some degree of normality which i think could be improved through transformations in order to make the model better performing.

***THANKYOU!***

