PHASE 2 PROJECT:
Modelling House Sales Prices in King County

GROUP 5 MEMBERS

Lisa Mwikali
Japhet Cheboiywo
Purity Gitonga
Cynthia Dalmas
Brian Ochieng
Bethuel Maruru

01

INTRODUCTION

The management of Skyline Ltd, a start-up real estate company is interested in determining the factors that affect the prices of real-estate units in King County. This would help the company optimize its sales revenue, revamp its marketing strategy and position itself as the leading real estate firm in the region.

As their consultant, we sought to explore the huge world of real estate to understand the dynamics of the housing market in King County, and thereafter generate a pricing model that will be used by the company.

02

Project Goals

DATA EXPLORATION

- Examine the features and structure of datasets.
- Examine the relationships between the variables.

MODEL DEVELOPMENT

- Build a linear regression model
- Prepare the data for training the model.

FEATURE SELECTION

- Determine the main predictors of home values.
- Utilize statistical methods and domain expertise.

MODEL EVALUATION

- Evaluate the model's performance (MSE).
- Validate model generalization.

PROJECT GOALS

INSIGHTS GENERATION

- Regression coefficient interpretation.
- Get insights that stakeholders can use.

03 DATA DESCRIPTION

The King County dataset obtained from Kaggle.com was selected for use in the research. The King County dataset contains the below variables.

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
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 house was built
16.`yr_renovated` - Year when house was renovated
17.`zipcode` - ZIP Code used by the United States Postal Service
18. `lat` - Latitude coordinate
19.`long` - Longitude coordinate
21. ‘sqft_living15` - The square footage of interior housing living space for the nearest 15 neighbors
22.`sqft_lot15` - The square footage of the land lots of the nearest 15 neighbors

04 Methodology

Data Loading and Preview:

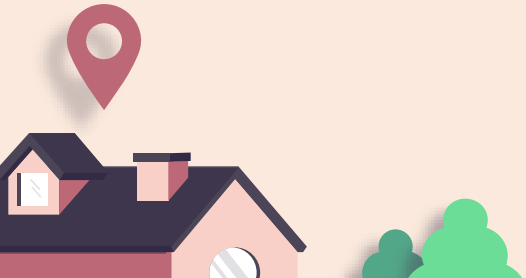
We loaded the dataset and imported the required libraries to see a preview of its structure.

Managing The Missing Values

We determined the columns with missing values. waterfront, view and yr_renovated columns had missing values. We were able to establish that there was a significant percentage of missing data on yr_renovated and waterfront , so we removed those and for view column, we imputed the values with the mode.

Feature engineering

To be able to carry out our analysis we extracted relevant data from already-existing columns, such as the numeric portion of the "grade" column.



EXPLORATORY DATA ANALYSIS

To obtain insight into the data, we investigated correlations between variables and used heatmaps, pair plots, and scatter plots to depict associations.

Correlation Heat Map showing relationships between key features and house prices

Correlation Heat Map Findings

- Based on the correlation results , "sqft_living," "bathrooms," and "grade" show a strong positive correlation with the "price" of the house. This implies that when living area, number of bathrooms, or grade of the house increases, the price will increase as well.

- "zipcode" has a slight negative correlation nearing -1 with price if certain areas tend to have lower prices.

- "yr_built" and "price" have a low correlation coefficient close to 0, indicating that the year a house was built might not have a strong linear impact on its price.

- There is multicollinearity between "sqft_above" and "sqft_living" because they are highly correlated.

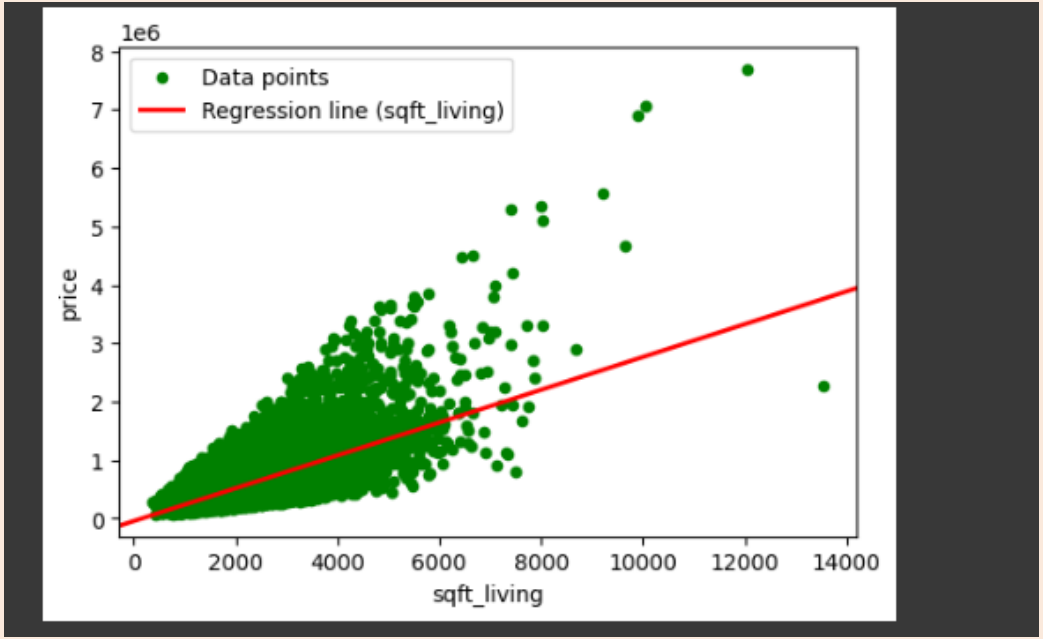
- Based on the heatmap, we decided to select features that have a strong correlation with the "price“ for model building. It helped us in identifying important predictors and potentially dropping less relevant or highly correlated predictors to improve model efficiency

BUILDING REGRESSION MODELS

- **BASELINE MODEL**

Using the "sqft_living" variable, we first created a baseline linear regression model. We then used metrics like R-squared and mean squared error (MSE) to assess the model's performance.

- **Visualizing the scatter plot and the line of best fit for the baseline model**



BUILDING REGRESSION MODELS

Residual Plot for the baseline model

Interpretation of the findings for the baseline model

- The intercept value (-43,988.89) represents the estimated price of a house when the independent variable (sqft_living) is zero. In our case this interpretation is not practical since a house cannot have a living area of zero, so, it signifies the baseline value of the model.

- The coefficient (280.86) indicates that for every one unit increase in (sqft_living), the price of the house is estimated to increase by 280.86. This coefficient represents the change in the (price) per unit over change (sqft_living).

- The R-squared value of 0.493 suggests that approximately 49.3% of the variance in house prices can be explained by the linear relationship with square footage of living space.

- The p-value for the F-statistic is 0.00, meaning that the model is statistically significant.

MULTIPLE LINEAR REGRESSION MODELS

We used multiple features to create a multiple linear regression model, which was then optimized by removing variables which were not statistically significant.

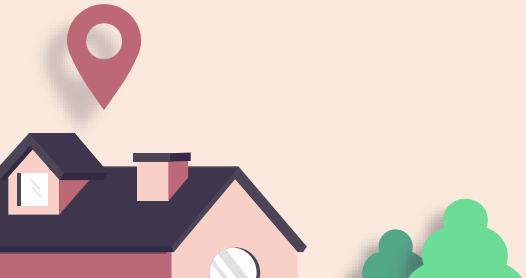
Regression Results using all the features

OLS Regression Results

Dep. Variable:	price	R-squared:	0.640
Model:	OLS	Adj. R-squared:	0.640
Method:	Least Squares	F-statistic:	4803.
Date:	Fri, 05 Apr 2024	Prob (F-statistic):	0.00
Time:	13:40:38	Log-Likelihood:	-2.9635e+05
No. Observations:	21597	AIC:	5.927e+05
Df Residuals:	21588	BIC:	5.928e+05
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.218e+07	5.25e+05	-61.282	0.000	-3.32e+07	-3.12e+07
bedrooms	-2.901e+04	2059.575	-14.087	0.000	-3.31e+04	-2.5e+04
sqft_living	199.5426	3.489	57.199	0.000	192.705	206.380
view	9.39e+04	2103.379	44.642	0.000	8.98e+04	9.8e+04
grade	8.178e+04	2187.692	37.384	0.000	7.75e+04	8.61e+04
bathrooms	-4158.0755	3345.798	-1.243	0.214	-1.07e+04	2399.937
floors	-2.764e+04	3713.135	-7.443	0.000	-3.49e+04	-2.04e+04
sqft_basement	-1.7950	4.491	-0.400	0.689	-10.598	7.008
lat	6.669e+05	1.11e+04	60.214	0.000	6.45e+05	6.89e+05

Omnibus:	18735.051	Durbin-Watson:	1.995
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1702719.450
Skew:	3.729	Prob(JB):	0.00
Kurtosis:	45.855	Cond. No.	8.06e+05



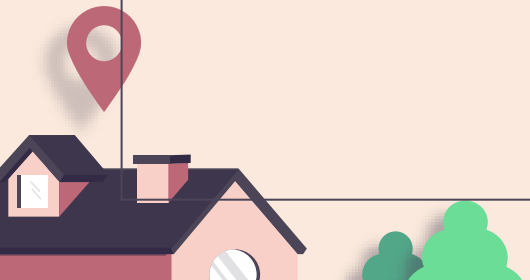
Interpretation of the multiple linear regression model

The intercept value (-3,218,420) represents the estimated price of a house when all the independent variables in the model are zero. In our case having all variables at zero might not make sense so we took the intercept as the baseline value of the model.

Based on the coefficients;

- A one-unit increase in the number of bedrooms is associated with a decrease in price by 2,901.335, when all the other variables are kept constant.
- For every additional square foot of living space, the price is estimated to increase by 199.5426, when all the other variables are kept constant.
- A better view quality is associated with an increase in price by 93,899.58, when all the other variables are kept constant.

- Higher grades correspond to higher prices, with an increase of 81,784.08 for each grade point, when all the other variables are kept constant.
- Each additional floor is associated with a decrease in price by 27,637.31, assuming other variables remain constant.
- Changes in latitude have a substantial impact on prices, with an increase of 666,856.7 per unit change in latitude, when all the other variables are kept constant.
- The R-squared value of 0.64 indicates that approximately 64% of the variance in house prices is explained by the linear relationship with the multiple independent variables.



MULTIPLE LINEAR REGRESSION MODELS

Regression Results after dropping variables that have a P(t) greater than a standard alpha 0.05

OLS Regression Results

Dep. Variable:	price	R-squared:	0.640
Model:	OLS	Adj. R-squared:	0.640
Method:	Least Squares	F-statistic:	6403.
Date:	Fri, 05 Apr 2024	Prob (F-statistic):	0.00
Time:	13:40:38	Log-Likelihood:	-2.9635e+05
No. Observations:	21597	AIC:	5.927e+05
Df Residuals:	21590	BIC:	5.928e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.218e+07	5.18e+05	-62.105	0.000	-3.32e+07	-3.12e+07
bedrooms	-2.955e+04	2021.669	-14.616	0.000	-3.35e+04	-2.56e+04
sqft_living	197.2828	2.979	66.234	0.000	191.445	203.121
view	9.376e+04	2068.421	45.329	0.000	8.97e+04	9.78e+04
grade	8.157e+04	2123.408	38.413	0.000	7.74e+04	8.57e+04
floors	-2.851e+04	3143.876	-9.069	0.000	-3.47e+04	-2.24e+04
lat	6.668e+05	1.09e+04	61.076	0.000	6.45e+05	6.88e+05

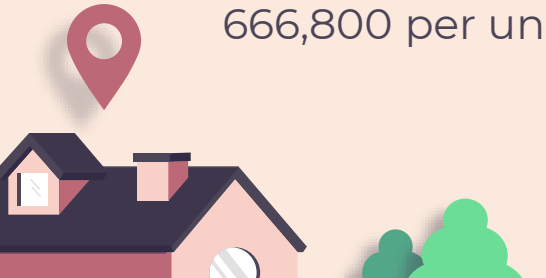
Omnibus:	18733.272	Durbin-Watson:	1.995
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1700656.549
Skew:	3.729	Prob(JB):	0.00
Kurtosis:	45.828	Cond. No.	7.86e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.86e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

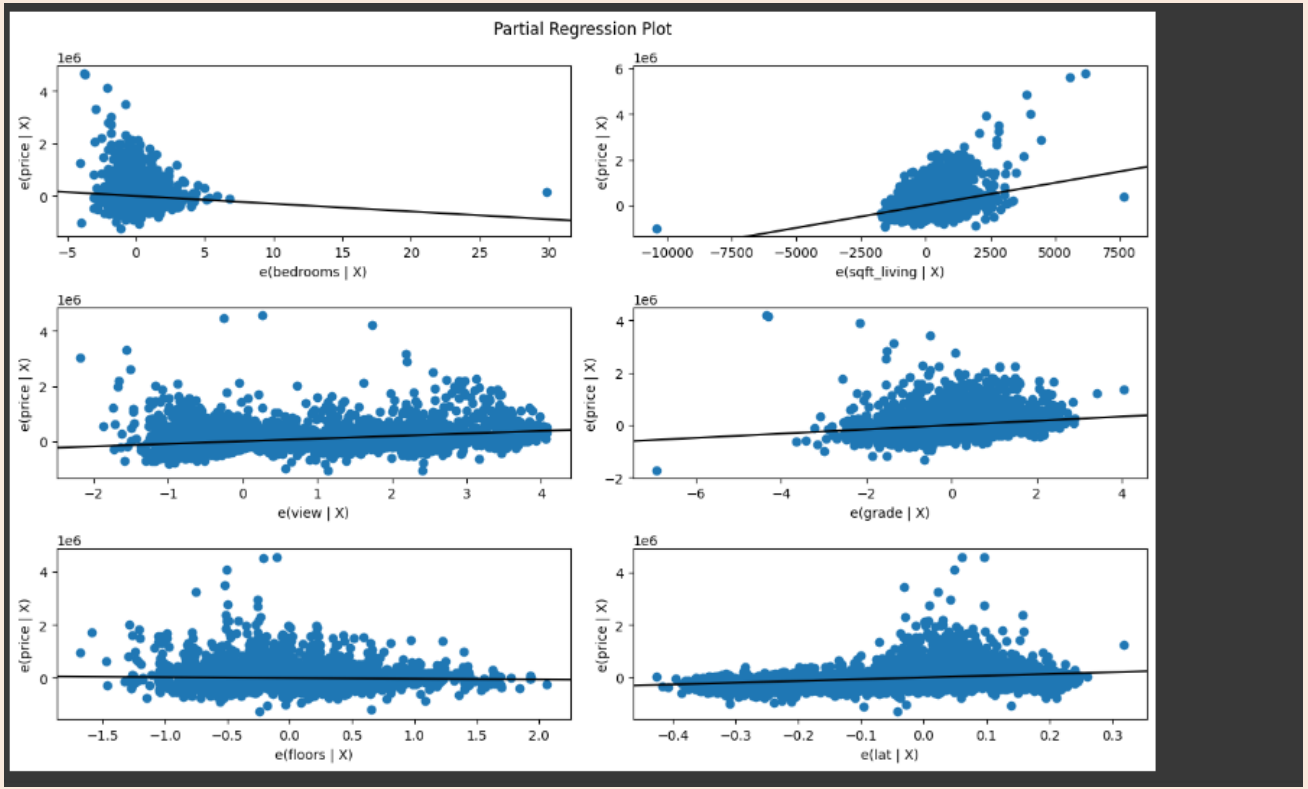
Interpretation of the findings for the multiple linear regression model after dropping statistically insignificant variables

The R-squared value of 0.640 indicates that approximately 64% of the variance in house prices can be explained by the independent variables included in the model.
Based on the coefficients;

- For every square foot increase in living space, the price increases by 197.28
- A better view quality corresponds to an increase in price by 93,760
- Higher grades result in higher prices, with an increase of 81,570 for each unit increase in grade
- Additional floors are associated with a decrease in price by 28,510
- Changes in latitude have a significant impact on prices, with an increase of 666,800 per unit change in latitude



Plotting Partial Regression Plot for the Multiple Linear Regression



POLYNOMIAL REGRESSION

Using R-squared and MSE, we compared the performance of several degrees (2–5) of polynomial regression models in order to identify non-linear correlations between features and the target variable.

2 degrees

- The MSE of approximately 200710.51 indicates the average squared difference between the predicted price and the actual price in the test data set.
- The R-squared value of 0.715 suggests that about 71.5% of the variance in (price) is explained by the independent variables in the training data set.
- The R-squared value of 0.708 indicates that the model explains about 70.8% of the variance in the test data set. This value is close to the training R-squared, suggesting that the model can be used to generalize well to unseen data.

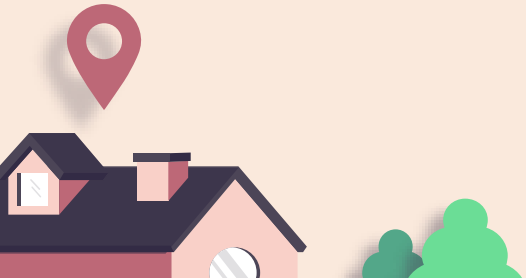
3 degrees

- The MSE of approximately 192833.78 is lower than the MSE for the degree 2 model, indicating potentially better predictive performance.
- The R-squared value of 0.738 on the training data suggests that this degree 3 polynomial model explains about 73.8% of the variance.
- The R-squared value of 0.731 on the test data indicates good generalization performance, as it is close to the training R-squared.

POLYNOMIAL REGRESSION

4 degrees

- The MSE of approximately 192963.83 is similar to that of the degree 3 model.
- The R-squared value of 0.749 on the training data indicates that this degree 4 polynomial model explains about 74.9% of the variance.
- The R-squared value of 0.73 on the test data suggests that the model's generalization performance is comparable to the degree 3 model.



5 degrees

- The MSE of approximately 206093.91 is higher than that of the degree 3 and degree 4 models, indicating potential overfitting or decreased generalization performance.
- The R-squared value of 0.752 on the training data suggests that this degree 5 polynomial model explains about 75.2% of the variance.
- The R-squared value of 0.692 on the test data indicates a drop in generalization performance compared to the lower-degree models, which could be a sign of overfitting.

05 Conclusion

- In conclusion, based on our analysis, the Polynomial Regression Model with a degree of 3 (PR_Model_3) provided the best performance with an R-squared (testing) score of 0.731. Therefore, we concluded that polynomial regression is the best solution for predicting house prices in the King County dataset.
- Square footage of living space in the home is the highest determinant in house pricing. The other factors included Construction and design of the house Number of floors Number of bedrooms Quality of view from house Location

06

Recommendations

The analysis recommends that Skyline Ltd prices higher those house units with a large square footage of living space, less number of floors, higher number of bedrooms. In addition, the better the Location and quality of view, the higher the house unit price.