

PROJECT PHASE 2

AN ANALYSIS AND MODELING OF KING'S COUNTY HOME
SALES DATASET

BY G-ONE LIMITED



G-ONE LIMITED MEMBERS

- Nazrah Nyangwara
- Loise Hellen
- Bahati Ndwiga
- Felix Awino
- Robin Mutai
- Stephen Ndirangu
- Daniel Ndirangu

BUSINESS UNDERSTANDING

- G-One Limited is a real estate agency that helps homeowners buy and/or sell homes. Our client, a family of three has approached us to help them settle on a home that will have the highest resell value.
- Our intention is to help the family get insight into the features that will most contribute to the highest or best sales of the housing units.
- To achieve this, we will analyze the **King's County home sales dataset**.

DATA UNDERSTANDING

- The dataset was obtained from Kings County housing dataset contained in a CSV file named `kc_house_data.csv`.
- The file contains information on over 21,000 housing units. The data is organized into a table with several columns containing different information about the houses.
- We noted that the data was collected between the time period of 2014 and 2015
- Some of the challenges encountered during data preparation included the presence of missing values, outliers and placeholders.
- We went through data preparation and modeling to come up with our final conclusions for this project

METHODS USED

- Data cleaning
 - We filled in missing values, and created some additional columns as necessary
 - Dropping some columns that we deemed not too relevant to the modeling.
- Data Modeling:
 - We created a base model and subsequently added additional variables to come up with our final model
 - Our conclusions and recommendations were based on the final model

CORRELATION HEAT MAP: PRICE VS FEATURES



BASELINE MODEL

OLS Regression Results

```
=====
Dep. Variable:          price_log      R-squared:                0.483
Model:                  OLS           Adj. R-squared:          0.483
Method:                 Least Squares  F-statistic:              2.020e+04
Date:                   Thu, 01 Jun 2023  Prob (F-statistic):      0.00
Time:                   20:33:11        Log-Likelihood:           -9662.2
No. Observations:       21597          AIC:                     1.933e+04
Df Residuals:           21595          BIC:                     1.934e+04
Df Model:                1
Covariance Type:        nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
const	12.2188	0.006	1915.383	0.000	12.206	12.231
sqft_living	0.0004	2.81e-06	142.118	0.000	0.000	0.000

```
=====
Omnibus:                3.541          Durbin-Watson:            1.978
Prob(Omnibus):           0.170          Jarque-Bera (JB):         3.562
Skew:                    0.028          Prob(JB):                 0.169
Kurtosis:                2.973          Cond. No.                  5.63e+03
=====
```

FINAL MODEL R-SQUARED

OLS Regression Results

=====			
Dep. Variable:	price_log	R-squared:	0.651
Model:	OLS	Adj. R-squared:	0.651
Method:	Least Squares	F-statistic:	1550.
Date:	Thu, 01 Jun 2023	Prob (F-statistic):	0.00
Time:	20:35:51	Log-Likelihood:	-5411.9
No. Observations:	21597	AIC:	1.088e+04
Df Residuals:	21570	BIC:	1.109e+04
Df Model:	26		
Covariance Type:	nonrobust		

FINAL MODEL COEFFICIENTS

	coef	std err	t	P> t	[0.025	0.975]
-----	-----	-----	-----	-----	-----	-----
const	24.4024	0.201	121.605	0.000	24.009	24.796
sqft_living	0.0002	4.92e-06	37.118	0.000	0.000	0.000
bedrooms	-0.0298	0.003	-9.950	0.000	-0.036	-0.024
bathrooms	0.0791	0.005	15.800	0.000	0.069	0.089
sqft_lot	-3.096e-08	5.25e-08	-0.589	0.556	-1.34e-07	7.2e-08
floors	0.0774	0.005	15.457	0.000	0.068	0.087
yr_built	-0.0058	0.000	-56.178	0.000	-0.006	-0.006
grade_11 Excellent	0.1194	0.018	6.473	0.000	0.083	0.156
grade_12 Luxury	0.2127	0.035	6.031	0.000	0.144	0.282
grade_13 Mansion	0.2291	0.088	2.593	0.010	0.056	0.402
grade_3 Poor	-1.0540	0.312	-3.383	0.001	-1.665	-0.443
grade_4 Low	-1.2108	0.062	-19.593	0.000	-1.332	-1.090
grade_5 Fair	-1.1267	0.025	-45.792	0.000	-1.175	-1.078
grade_6 Low Average	-0.9091	0.015	-59.940	0.000	-0.939	-0.879
grade_7 Average	-0.6303	0.012	-50.571	0.000	-0.655	-0.606
grade_8 Good	-0.3939	0.011	-34.531	0.000	-0.416	-0.372
grade_9 Better	-0.1604	0.011	-14.088	0.000	-0.183	-0.138
condition_Fair	-0.1676	0.024	-6.899	0.000	-0.215	-0.120
condition_Good	0.0190	0.005	3.576	0.000	0.009	0.029
condition_Poor	-0.1476	0.058	-2.530	0.011	-0.262	-0.033
condition_Very Good	0.0863	0.009	10.088	0.000	0.070	0.103
view_EXCELLENT	0.1655	0.024	7.018	0.000	0.119	0.212
view_FAIR	0.0833	0.020	4.191	0.000	0.044	0.122
view_GOOD	0.0352	0.017	2.053	0.040	0.002	0.069
view_NONE	-0.0974	0.011	-9.244	0.000	-0.118	-0.077
waterfront_YES	0.3151	0.032	9.987	0.000	0.253	0.377
Renovated_yes	0.0081	0.012	0.656	0.512	-0.016	0.032
=====	=====	=====	=====	=====	=====	=====

FINDINGS

- The model is statistically significant overall, with an F-statistic p-value well below 0.05
- The model explains about 65% of the variance in price
- The fact that we went from 1 predictors to 26 predictors and increased R-Squared by 17% from 48% to 65% is an indicator that this a fairly good model
- A number of the model coefficients are statistically significant. These are : "sqft_living, bedrooms, bathrooms, floors, yr_built, grade_11 Excellent, grade_12 Luxury, grade_13 Mansion, grade_3 Poor, grade_4 Low, grade_5 Fair, grade_6 Low Average, grade_7 Average, grade_8 Good, grade_9 Better, condition_Fair, condition_Good, condition_Poor, condition_Very Good, view_EXCELLENT, view_FAIR, view_GOOD, view_NONE, waterfront_YES" have p-values below 0.05 and are therefore statistically significant
- sqft_lot and Renovated_yes have p-values above 0.05 and are therefore not statistically significant at an alpha of 0.05

INTERPRETATION OF THE COEFFICIENTS

- The following features will improve the pricing of the houses:
- A unit increase in square foot living will increase the price of a house by 0.02%
- A unit increase in the number of bathrooms will increase the price of a house by 7.91%
- A unit increase in the number of floors will increase the price of a house by 7.74%
- The higher the grading of a house, the higher it's price. For instance, a house graded as excellent will attract a price increase of 11.94%, while a house graded as luxury will attract a price increase of 21.27%, and mansion a price increase of 22.91%
- The better the condition of a house, the higher it's price. A house in "good" condition will attract a price increase of 1.9% while a house in "very good" condition will attract a price increase of 8.63%
- Houses without views attract lower prices compared to houses with views. The model demonstrates that a house with a good view attracts a price increase of 3.52%, fair view 8.33%, and excellent view 16.55% increase in price
- Houses with a waterfront attract a price increase of 31.51%

CONCLUSIONS AND RECOMMENDATIONS

- In conclusion, the model has provided insights into the various features that affect the price of a house in King's County. G-One Limited therefore has the following recommendations for the family to guide their choice of a house in the King's County neighborhood:
- They should consider the number of bathrooms
- They should consider the number of floors
- They should focus on houses graded as excellent, luxury, or mansion
- They should focus on houses whose condition are either good or very good
- Houses with a good view will attract a higher price compared to ones without
- Houses with a waterfront have the highest price value

THANK YOU