King County House Price Prediction

Author: Aswathi Sasikumar

Summary

- This projects helps stakeholder,Real-estate agency to provide advice to homeowners about how home renovations might increase the estimated value of their homes, and by what amount.

- The regression model developed in the project predicts the price of the house and how certain parameters affect the price of the house.

Outline

- Business Problem
- Data
- Methods
- Results
- Conclusions

Business Problem

- The aim of this analysis is to build a multiple regression model that can predict house prices.

- The results will help home owners interested in selling their homes about the most important factors to consider for improving sale prices.The parameters they must improve like the renovation they need to carry out to increase the estimated price of house.

Data

- The King County Housing Data Set contains information about the size, location, condition, and other features of houses in King County.

- The aim of this project is to develop a multiple regression model than can predict a house's price as accurately as possible.

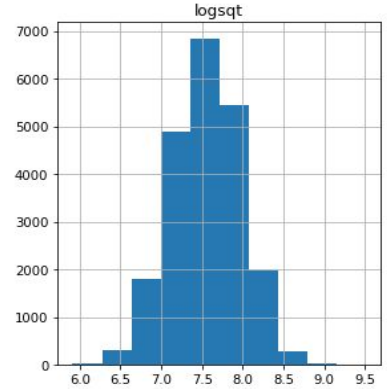
Methods

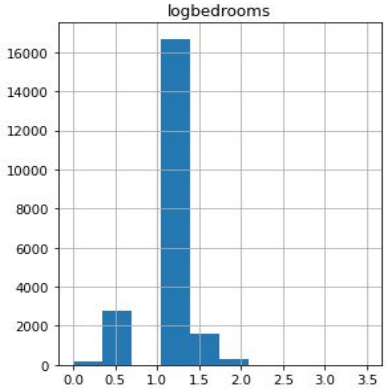
- After exploring and preprocessing the data, simple and multiple linear regression models were built in OLS statsmodels, with price as the dependent variable.

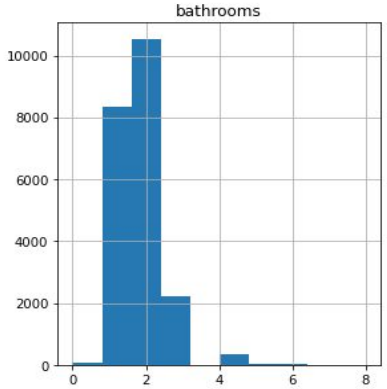
Results

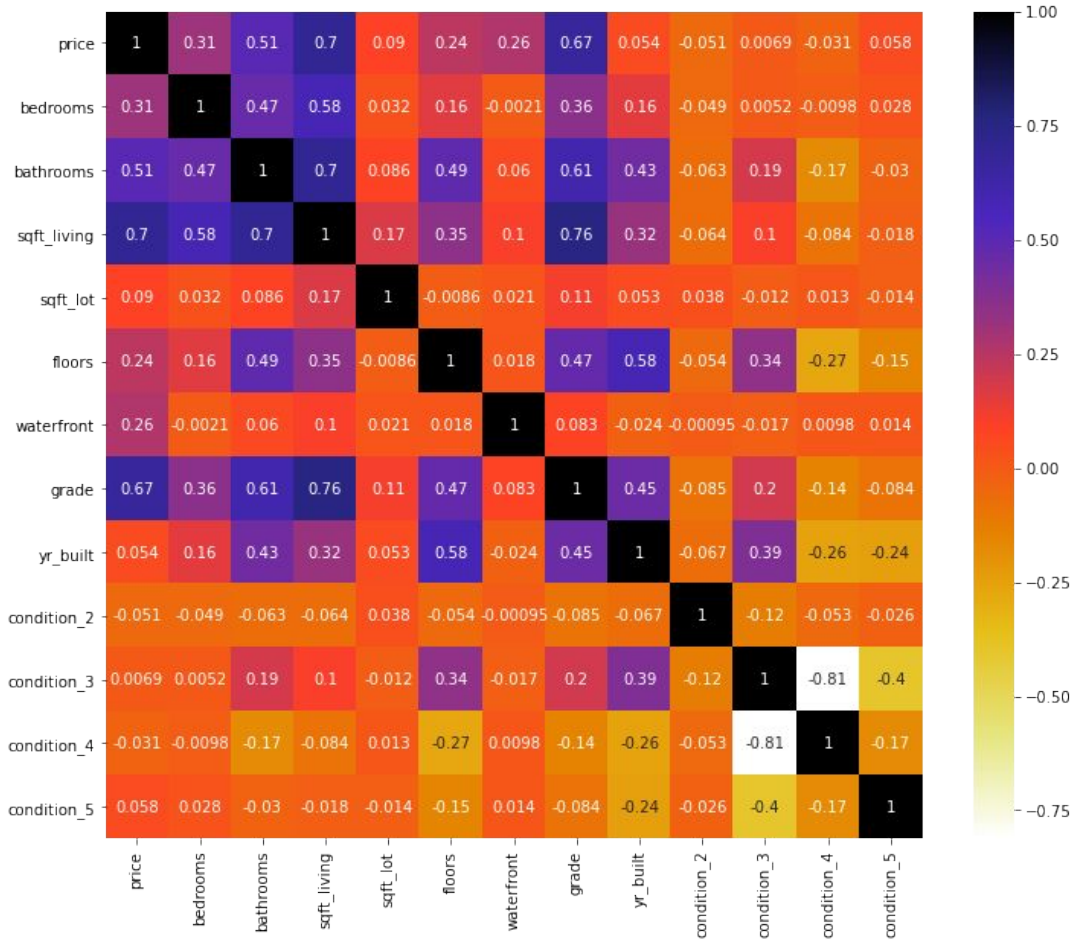
Together, square footage, bedrooms and bathrooms are the best predictors of a house's price in King County. These features were included in the final multiple regression model. The model satisfied all multiple regression assumptions, and p-values for each predictor variable were 0. The r-squared value of the model was .477


price
7000
6000
5000
4000
3000
2000
1000
0
12
13
14
15
16


logsqt
7000
6000
5000
4000
3000
2000
1000
0
6.0
6.5
7.0
7.5
8.0
8.5
9.0
9.5


logbedrooms
16000
14000
12000
10000
8000
6000
4000
2000
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5


bathrooms
10000
8000
6000
4000
2000
0
0
2
4
6
8


price 1 0.31 0.51 0.7 0.09 0.24 0.26 0.67 0.054 -0.051 0.0069 -0.031 0.058
bedrooms 0.31 1 0.47 0.58 0.032 0.16 -0.0021 0.36 0.16 -0.049 0.0052 -0.0098 0.028
bathrooms 0.51 0.47 1 0.7 0.086 0.49 0.06 0.61 0.43 -0.063 0.19 -0.17 -0.03
sqft_living 0.7 0.58 0.7 1 0.17 0.35 0.1 0.76 0.32 -0.064 0.1 -0.084 -0.018
sqft_lot 0.09 0.032 0.086 0.17 1 -0.0086 0.021 0.11 0.053 0.038 -0.012 0.013 -0.014
floors 0.24 0.16 0.49 0.35 -0.0086 1 0.018 0.47 0.58 -0.054 0.34 -0.27 -0.15
waterfront 0.26 -0.0021 0.06 0.1 0.021 0.018 1 0.083 -0.024 -0.00095 -0.017 0.0098 0.014
grade 0.67 0.36 0.61 0.76 0.11 0.47 0.083 1 0.45 -0.085 0.2 -0.14 -0.084
yr_built 0.054 0.16 0.43 0.32 0.053 0.58 -0.024 0.45 1 -0.067 0.39 -0.26 -0.24
condition_2 -0.051 -0.049 -0.063 -0.064 0.038 -0.054 -0.00095 -0.085 -0.067 1 -0.12 -0.053 -0.026
condition_3 0.0069 0.0052 0.19 0.1 -0.012 0.34 -0.017 0.2 0.39 -0.12 1 -0.81 -0.4
condition_4 -0.031 -0.0098 -0.17 -0.084 0.013 -0.27 0.0098 -0.14 -0.26 -0.053 -0.81 1 -0.17
condition_5 0.058 0.028 -0.03 -0.018 -0.014 -0.15 0.014 -0.084 -0.24 -0.026 -0.4 -0.17 1
price bedrooms bathrooms sqft_living sqft_lot floors waterfront grade yr_built condition_2 condition_3 condition_4 condition_5
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

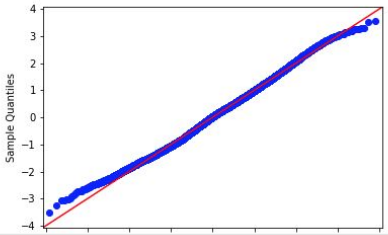
Dep. Variable:	price	**R-squared:**	0.477
Model:	OLS	**Adj. R-squared:**	0.477
Method:	Least Squares	**F-statistic:**	6568.
Date:	Sun, 05 Jun 2022	**Prob (F-statistic):**	0.00
Time:	19:47:10	**Log-Likelihood:**	-9789.8
No. Observations:	21597	**AIC:**	1.959e+04
Df Residuals:	21593	**BIC:**	1.962e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.6247	0.062	106.813	0.000	6.503	6.746
logsqt	0.8813	0.010	91.231	0.000	0.862	0.900
logbedrooms	-0.3121	0.012	-25.943	0.000	-0.336	-0.288
bathrooms	0.0780	0.005	16.237	0.000	0.069	0.087

Omnibus:	104.310	**Durbin-Watson:**	1.980
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	96.727
Skew:	0.131	**Prob(JB):**	9.91e-22
Kurtosis:	2.802	**Cond. No.**	193.

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

In [41]:
```
import scipy.stats as stats
residuals = model3.resid
sm.graphics.qqplot(residuals, dist=stats.norm, line='45', fit=True, )
plt.show;
```



Conclusions

- Together, square footage, bedrooms and bathrooms are the best predictors of a house's price in King County. Homeowners who are interested in selling their homes at a higher price should focus on expanding square footage. When expanding square footage, homeowners should consider building additional bathrooms,bedrooms as this analysis suggests that number of bathrooms is positively related to price.

Limitations

The model does have some limitations: given that some of the variables needed to be log-transformed to satisfy regression assumptions, any new data used with the model would have to undergo similar preprocessing.

Thank You!

Email: kukkuaswathi@gmail.com
GitHub: https://github.com/AswathiSasikumar
LinkedIn:
https://www.linkedin.com/in/aswathi-sasikumar-21989333/