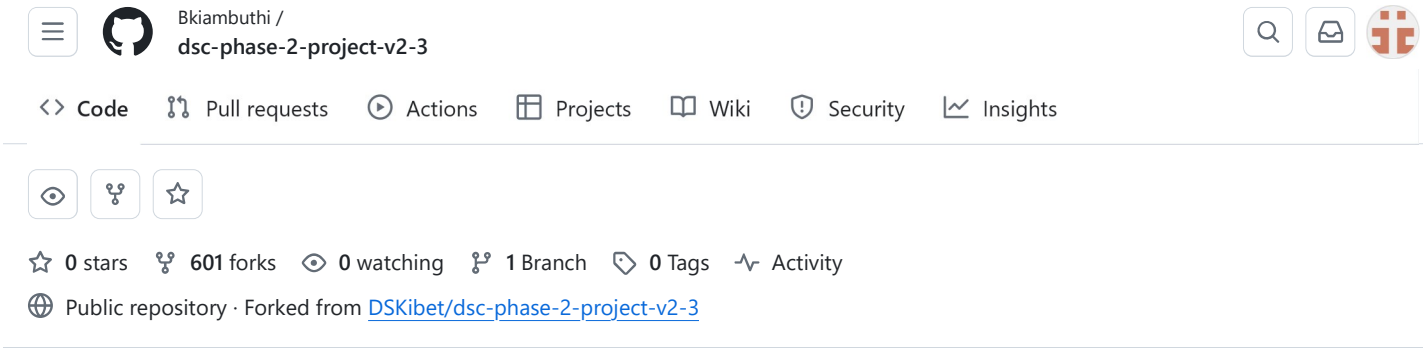

Bkiambuthi /
dsc-phase-2-project-v2-3
Code
Pull requests
Actions
Projects
Wiki
Security
Insights
0 stars
601 forks
0 watching
1 Branch
0 Tags
Activity
Public repository · Forked from DSKibet/dsc-phase-2-project-v2-3

main 1 Branch 0 Tags Go to file t Go to file Add file Code

This branch is 21 commits ahead of DSKibet/dsc-phase-2-project-v2-3:main . Contribute Sync fork

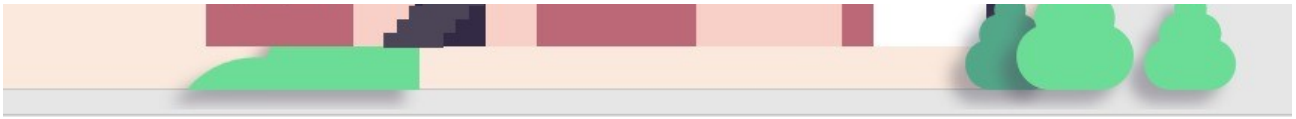
Japhet Kibet Cheboiywo and Japhet Kibet Cheboiywo	created a pdf for notebook & minor edi...	5 minutes ago
.ipynb_checkpoints	created a pdf for notebook & minor ed...	5 minutes ago
data	revisions for v2.3	3 years ago
Group 5 Phase 2 project.pdf	created a pdf for notebook & minor ed...	5 minutes ago
Group 5 phase 2 project.ipynb	created a pdf for notebook & minor ed...	5 minutes ago
README.md	Update README.md	yesterday
Understanding King County's H...	Editing the presentation	39 minutes ago
Understanding King County's H...	Editing the presentation	39 minutes ago
kc_house_data.csv	created a pdf for notebook & minor ed...	5 minutes ago

README

UNDERSTANDING KING COUNTY'S HOUSING MARKET

Introduction

The project aims to find important insights as we explore the huge world of real estate that will aid in the comprehension of the dynamics of the housing market in King County. We will walk you through each step of the process of developing and assessing our linear regression model and go over the main factors that affect home values, talk about the issues with linear regression, and use analysis and visualizations to show off our model's ability to predict outcomes.By the end of the presentation, we hope to provide insights that can be used in similar datasets , in addition to demystifying the complexity of real estate prediction.



Business Problem

The management of a startup real estate company seeks to understand the key factors influencing housing prices in King County. Their capacity to develop focused actions and activities to support housing affordability and fair access to housing is limited by a lack of data-driven insights.

The Data

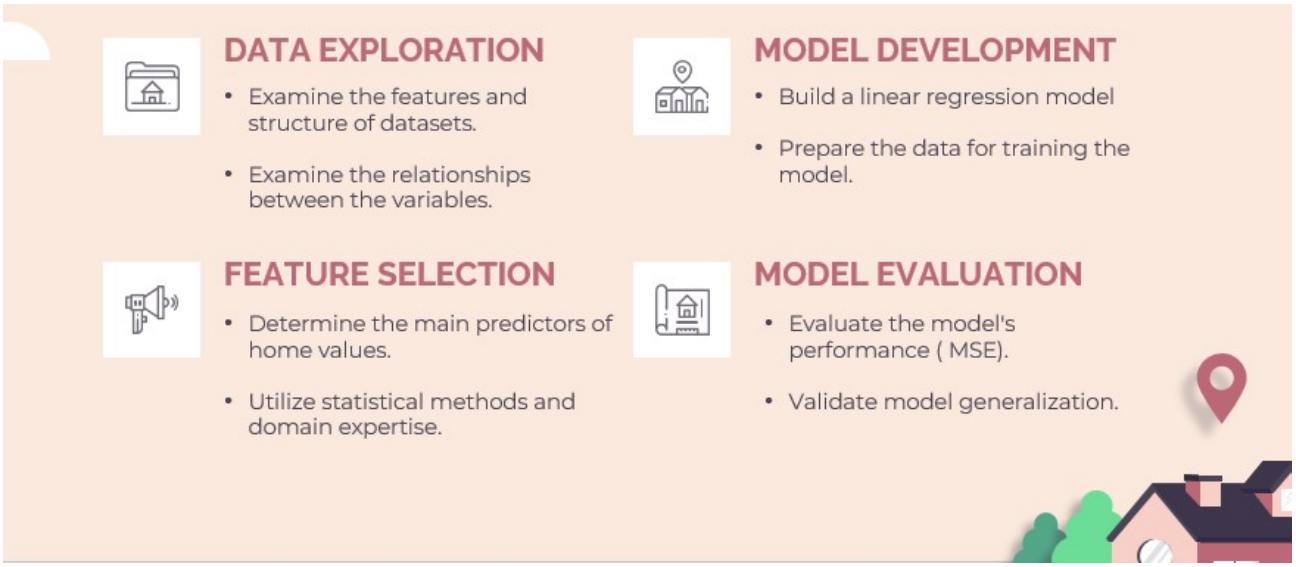
This project uses the King County House Sales dataset, which can be found in `kc_house_data.csv` in the data folder in this project's GitHub repository. The description of the column names can be found in `column_names.md` in the same folder.

Project Goals

The project aims to use CRISP-DM data science methodology to fit and select the best regression model that will predict housing prices. The resultant model will be used by the real estate company to price housing units so as to attract customers and optimize their sales.

Methods

Linear regression analysis will be used to find the relationship between sales prices and various house features other factors.



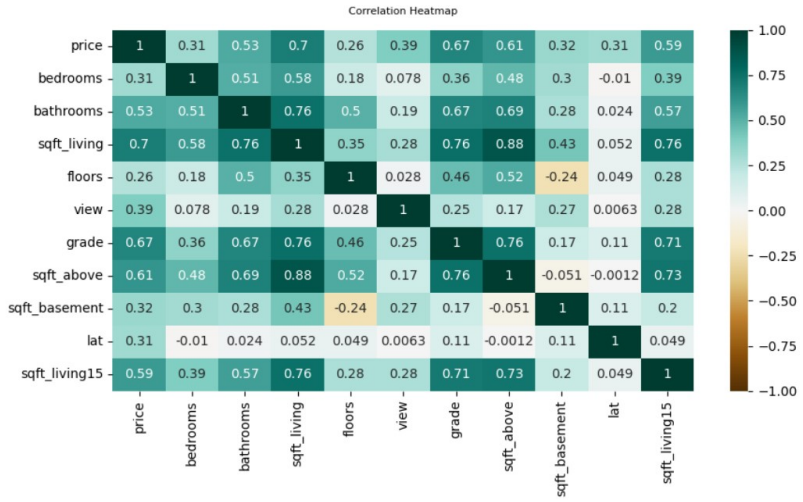
Results

When using linear regression there are four assumptions that need to be followed:

```
.Linearity - relationship between predictor and target should be linear.

.Independence - Observations are independent from each other, low or no multicollinearity.

.Normality - Errors should be normally distributed.

.Homoscedasticity - variance on the errors is the same
```

.Homoscedasticity - variance of the errors is the same.

1.Checking the relationships amongst all the variables in our data set



2. Fitting our baseline model using sqft_living as the independent variable and price as the dependent variable. Sqft_living has the highest correlation with price.

OLS Regression Results

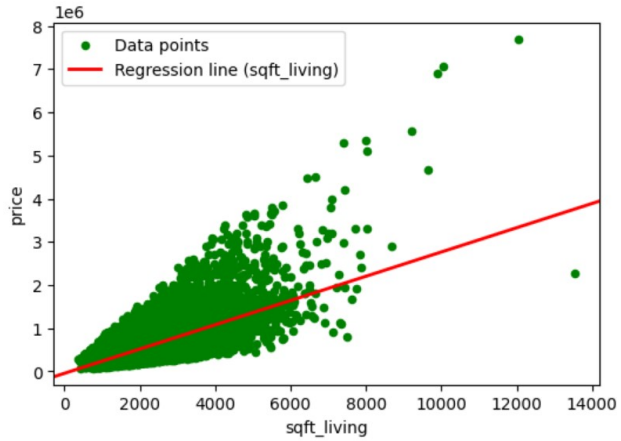
Dep. Variable:	price	**R-squared:**	0.493
Model:	OLS	**Adj. R-squared:**	0.493
Method:	Least Squares	**F-statistic:**	2.097e+04
Date:	Fri, 05 Apr 2024	**Prob (F-statistic):**	0.00
Time:	11:04:59	**Log-Likelihood:**	-3.0006e+05
No. Observations:	21597	**AIC:**	6.001e+05
Df Residuals:	21595	**BIC:**	6.001e+05
Df Model:	1		
Covariance Type:	nonrobust		

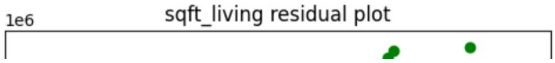
	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.399e+04	4410.023	-9.975	0.000	-5.26e+04	-3.53e+04
sqft_living	280.8630	1.939	144.819	0.000	277.062	284.664

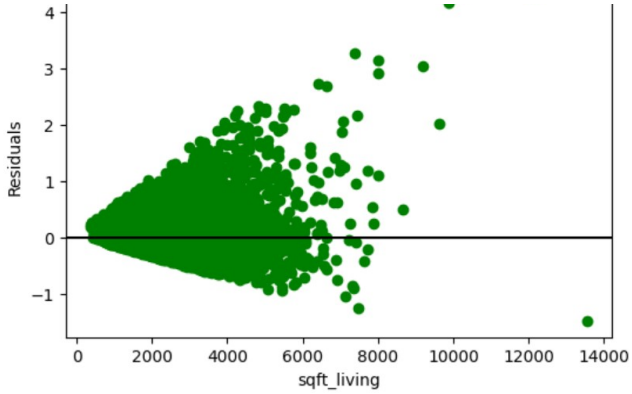
Omnibus:	14801.942	**Durbin-Watson:**	1.982
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	542662.604
Skew:	2.820	**Prob(JB):**	0.00
Kurtosis:	26.901	**Cond. No.**	5.63e+03

The model is explaining 49.3% of the change in price.

Equation for our model Price = 280.86X - 43988.89







3. Multiple linear regression

91]:

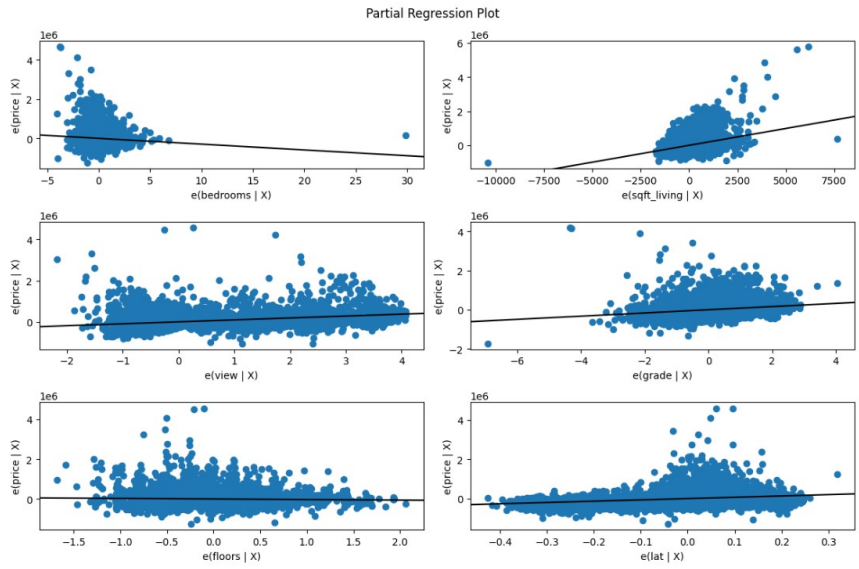
OLS Regression Results

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:	11:07:45	**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

The model is explaining 64% of the change in price.



4. Polynomial regression Using 3-degree polynomial regression, the model yielded the highest score. The results were as below:

   Mean Squared Error (MSE) 192833.78

   R-squared (training) 0.738

R-squared (testing) 0.731

Conclusion

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

For More Information

Please review our full analysis in our Jupyter Notebook or our presentation.

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%