rahulakrish / phase2_project Private

linear regression model

0 stars 0 forks

Star Unwatch

Code Issues Pull requests Actions Projects Security Insights S

main

rahulakrish updated per feedback from review ... 1 hour ago 52

View code

README.md

Phase_2 Project

Description

To help clients estimate selling prices of their home based on established factors. Also, to recommend any renovations that can be made to improve the resale value of the house

Data Source

King County House Sales dataset, which can be found in kc_house_data.csv

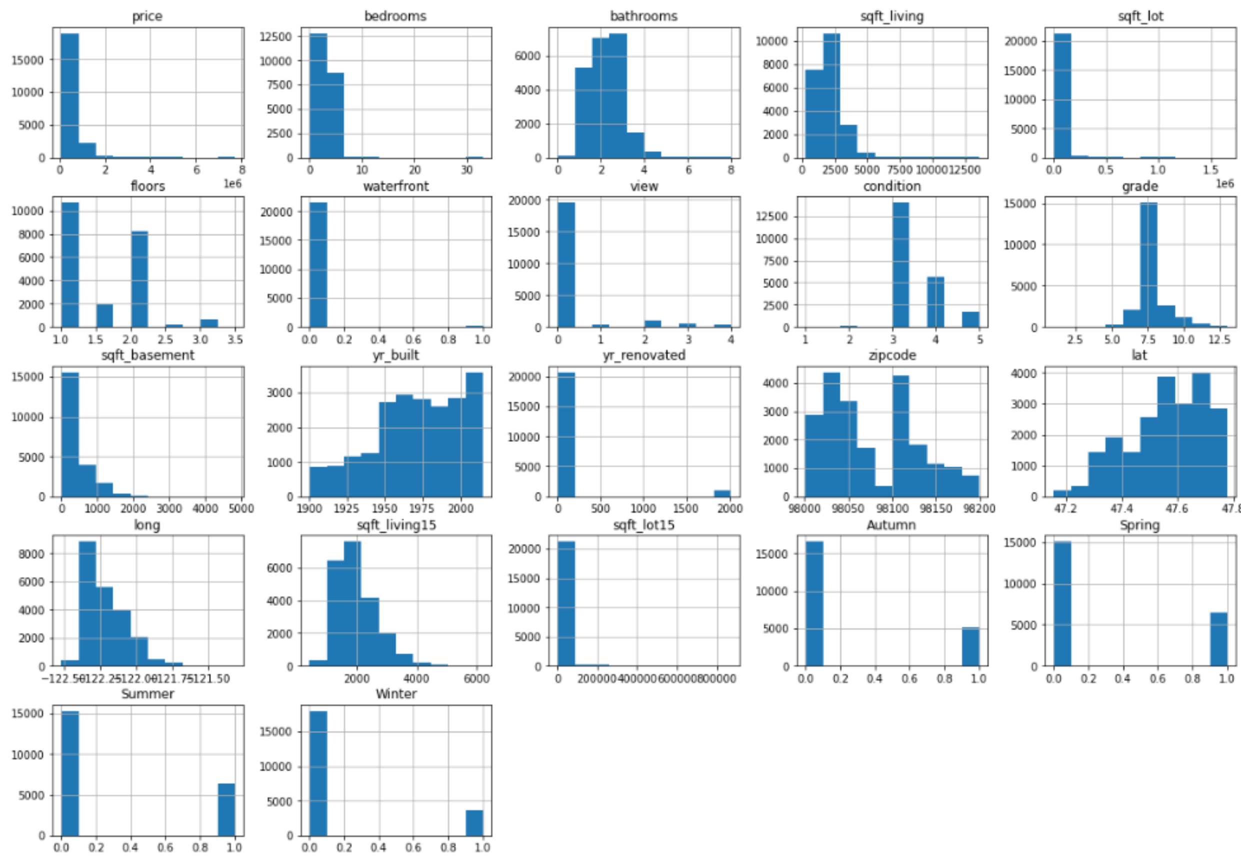
Methodology

Develop linear regression model to help predict house values. Dependent variable 'X' = price. y = Independent(predictor) variables are the rest of the factors in the dataset.

```
Index(['id', 'date', 'price', 'bedrooms', 'bathrooms', 'sqft_living',
       'sqft_lot', 'floors', 'waterfront', 'view', 'condition', 'grade',
       'sqft_above', 'sqft_basement', 'yr_built', 'yr_renovated', 'zipcode',
       'lat', 'long', 'sqft_living15', 'sqft_lot15'],
      dtype='object')
```

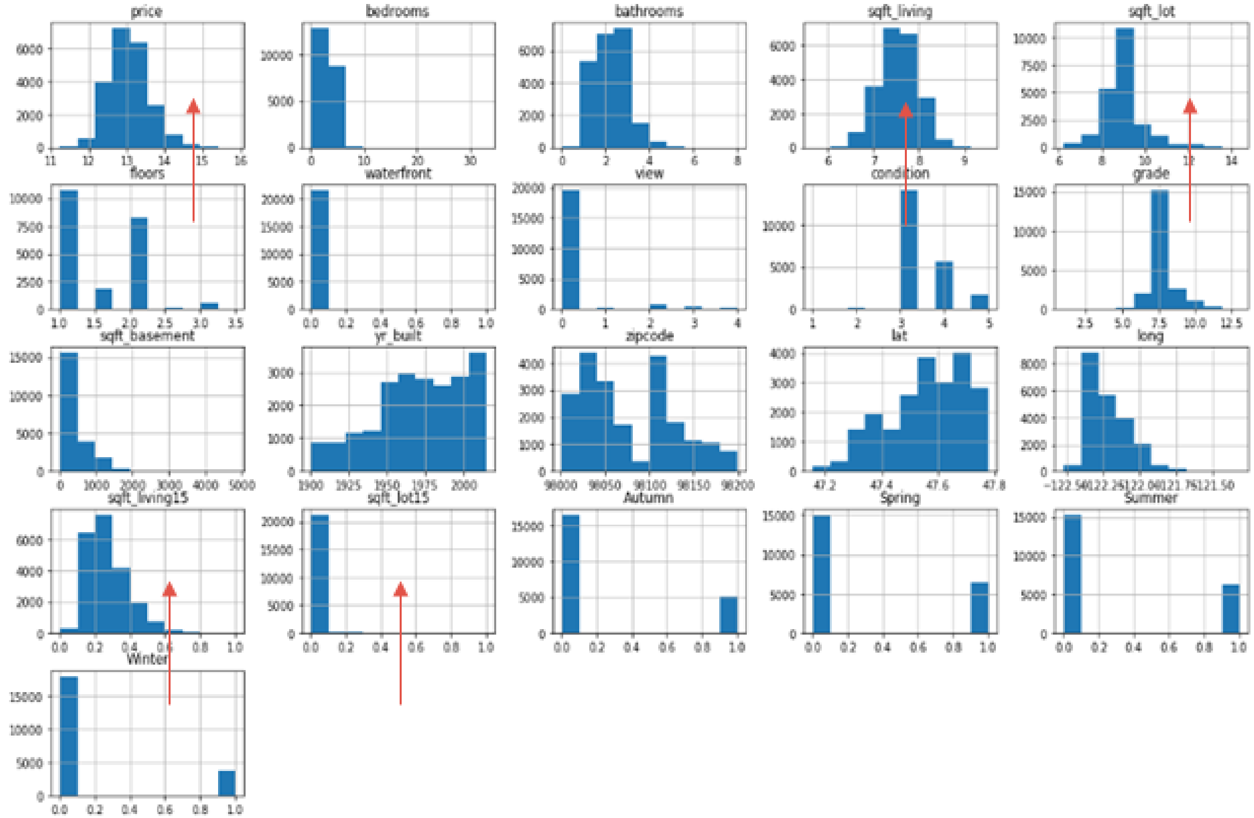
Step -1

Initially, we will look at the distribution of the variables in the dataset to check for normality. Though not mandatory, having normally distibuted varibales help the efficiency of the model.



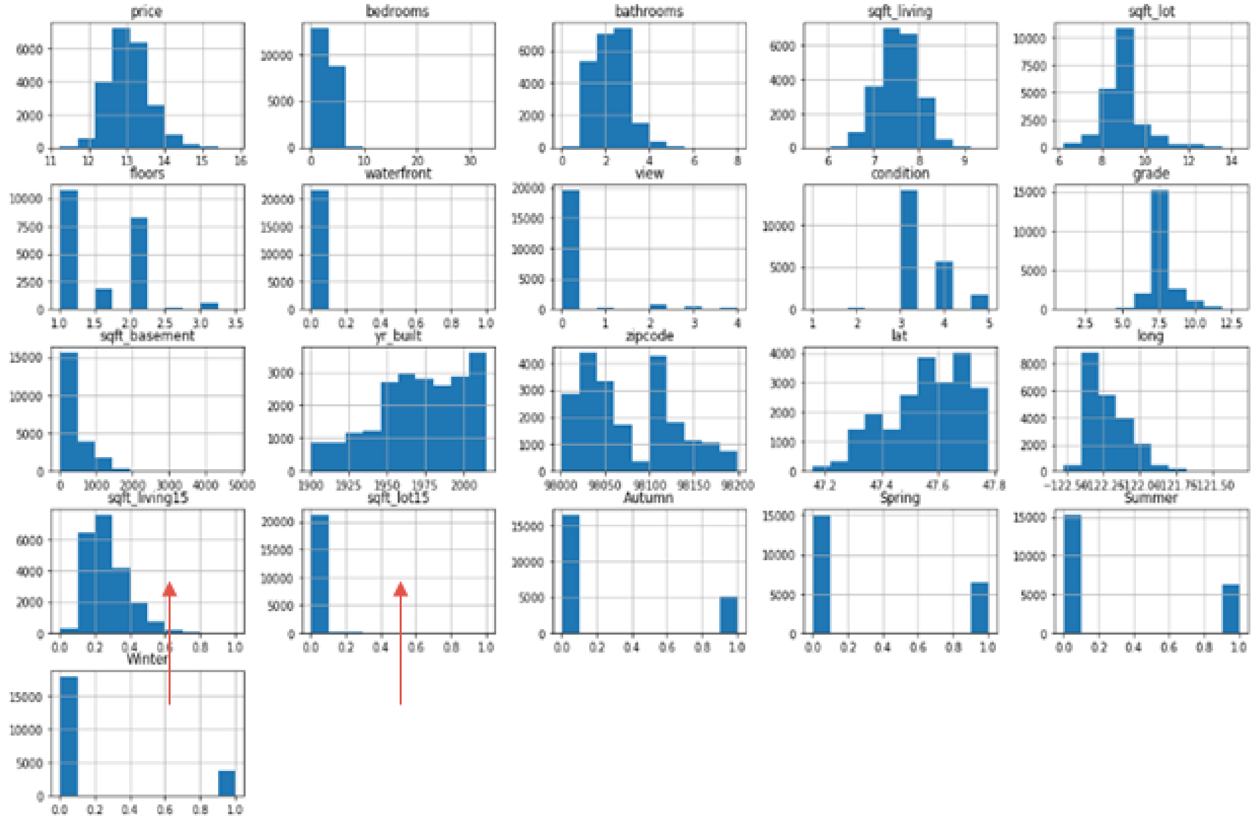
Step-2

We will then apply log transformation of the non-nomral variables to make them normal and help increase the efficiency of the model.



Step-3

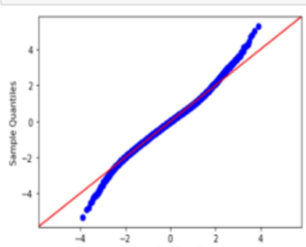
Next, since different variables have different values, we will apply scaling methods to make them more interpretable for the model. For eg. sqft_living is measure in thousands of sqft while price is in hundreds of thousands of dollars. Bringing them to a common scale will help boost model efficency

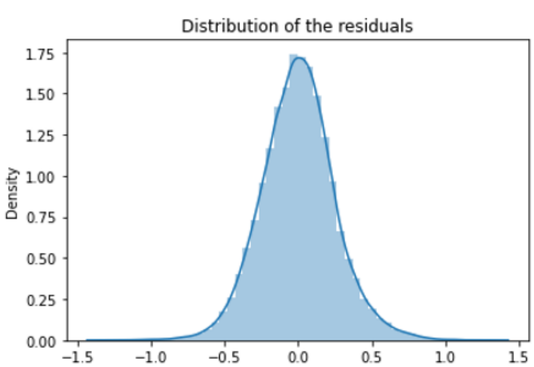


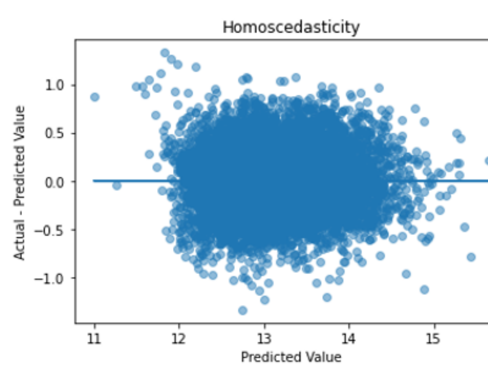
Step-4

Model diagnostics : Using metrics like R-squared, RMSE values and Q-Q plots to interepret the fit of the final model. Also, check to see if the model fit violates linear regression assumptions

```
# Q-Q plots
import scipy.stats as stats
residuals = model_6.resid
fig = sm.graphics.qqplot(residuals, dist=stats.norm, line='45', fit=True)
fig.show()
```







```
R-squared value for the baseline_model =  0.7009279563815052
R-squared value for model_1 =  0.7005211415851306  - removing yr_renovated
R-squared value for model_2 =  0.7715173571062647  - transforming price using log transform
R-squared value for model_3 =  0.7735521196953915  - transforming sqft_living using log transform
R-squared value for model_4 =  0.7743750042560188  - transforming sqft_lot using log transform
R-squared value for model_5 =  0.7743750042560187  - using scaled values
R-squared value for model_6 =  0.7726801965727106  - dropping seasons and the basement
```

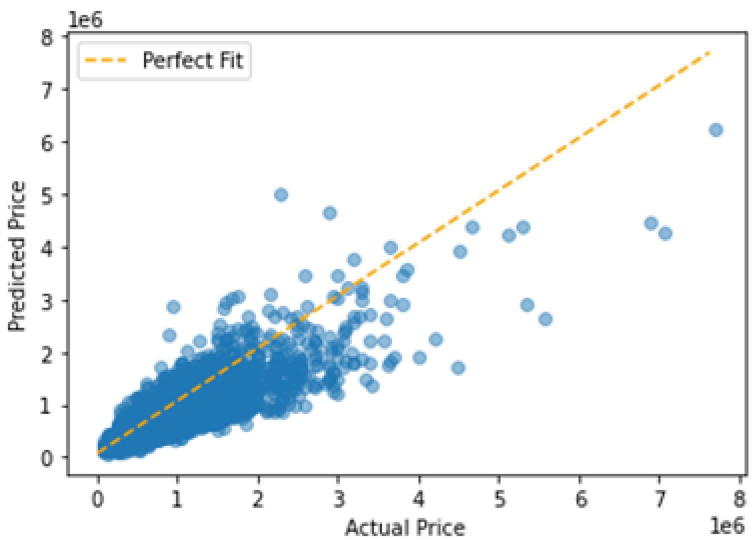
Step-5

Interpret the coeffecients of the predictor variables and pick out 2 to recommend to clients that will have the highest effect on the sale price of a house.

```python
# waterfront -  interpreted differenty since it was not log-transformed
percent_increase = round((np.exp(model_6.params[6])-1)*100,2)
print(f'Expected precentage increase in the house value if the house is on the waterfront is {percent_increase}%')
```

```
Expected precentage increase in the house value if the house is on the waterfront is 50.38%
```

Model Results

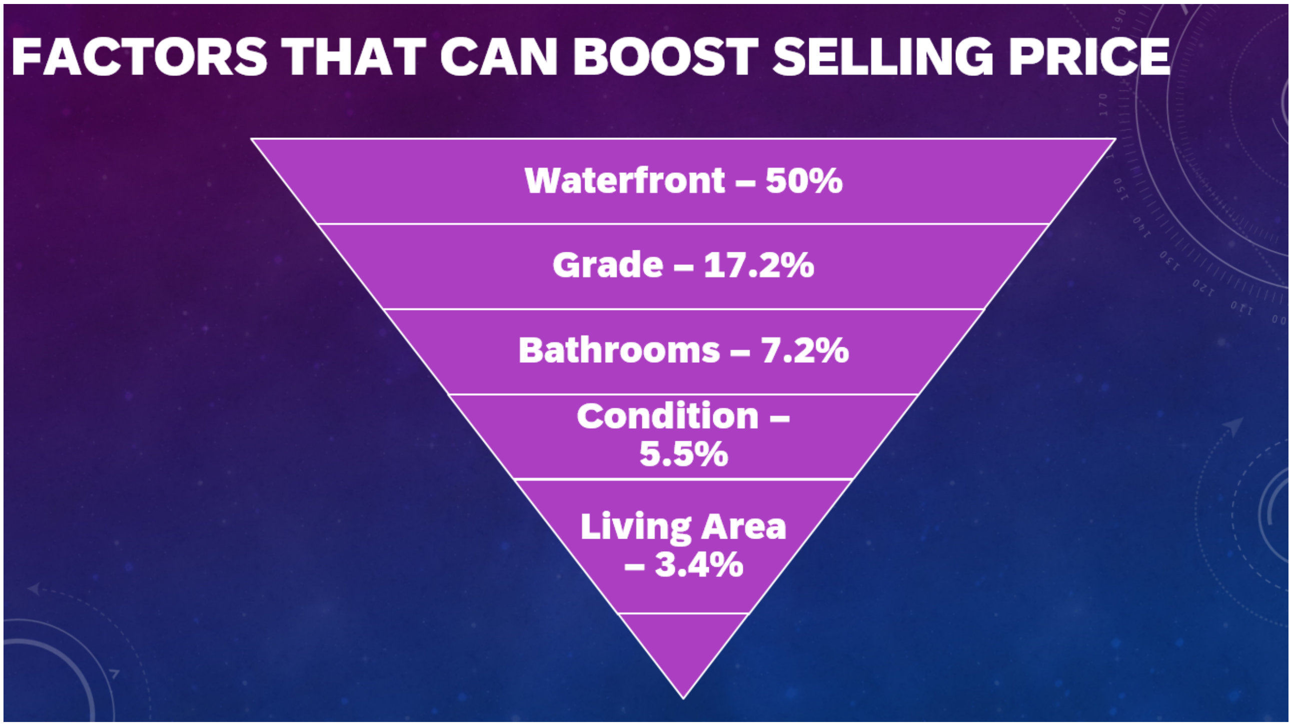



```
mean_squared_error = 33979601170.14924
root_mean_squared_error = 184335.5667529987
```

Conclusions

1. The model will be off by $184,335 when predicting the price of a house

2. Being on the waterfront is the most valuable asset when it comes to selling the house. It increases the value by nearly 50%. A unit increase in the grade, number of bathrooms and condition of the house yield 17.6%, 7.2% and 5.5% increase respectively



Releases

No releases published

Create a new release

Packages

No packages published

Publish your first package

Languages

- **Jupyter Notebook** 100.0%