KING COUNTY HOUSING

Multiple Linear Regression Analysis

Yesim Cebeci
June 23rd, 2022

TABLE OF CONTENTS

BUSINESS UNDERSTANDING

We have been approached by an investor wants to invest real estate business about how to accurately appraise homes in Kings County so that they can have idea about home prices it comes to buying and selling homes. We've been given a data set that contains various information about the different homes within King County.

In this study, we hope to highlight the features available to us in the data that were the most indicative of a property's sale and buy prices.

DATA UNDERSTANDING

The data provide to us consist of information pertaining to over **20,000** house sales carried out between **2014 and 2015**, located in the **data/kc_house_data.csv** file in this repository. Data dictionary summarizing the information contained in each of the **20 relevant features**.

DATA UNDERSTANDING

Metrics for Evaluation

There are 2 key metrics for evaluation to be used to assess if our model is considered successful.

Coefficients :

The coefficients of the features describe the mathematical relationship between each independent variable and the dependent variable, which in this case is the price of the house. The coefficient value demonstrates how much the mean of the target variable changes given a one-unit change in the featurevariable when the other features are unchanged.

Adjusted R2:

The Adjusted R2 is a key metric for evaluation of a multivariate linear regression model, as it accounts for the number of predictors in a model when calculating the model's goodness-of-fit.

DATA UNDERSTANDING

Nearly all practical datasets will contain **null** values. However, only three columns had missing values to be converted.

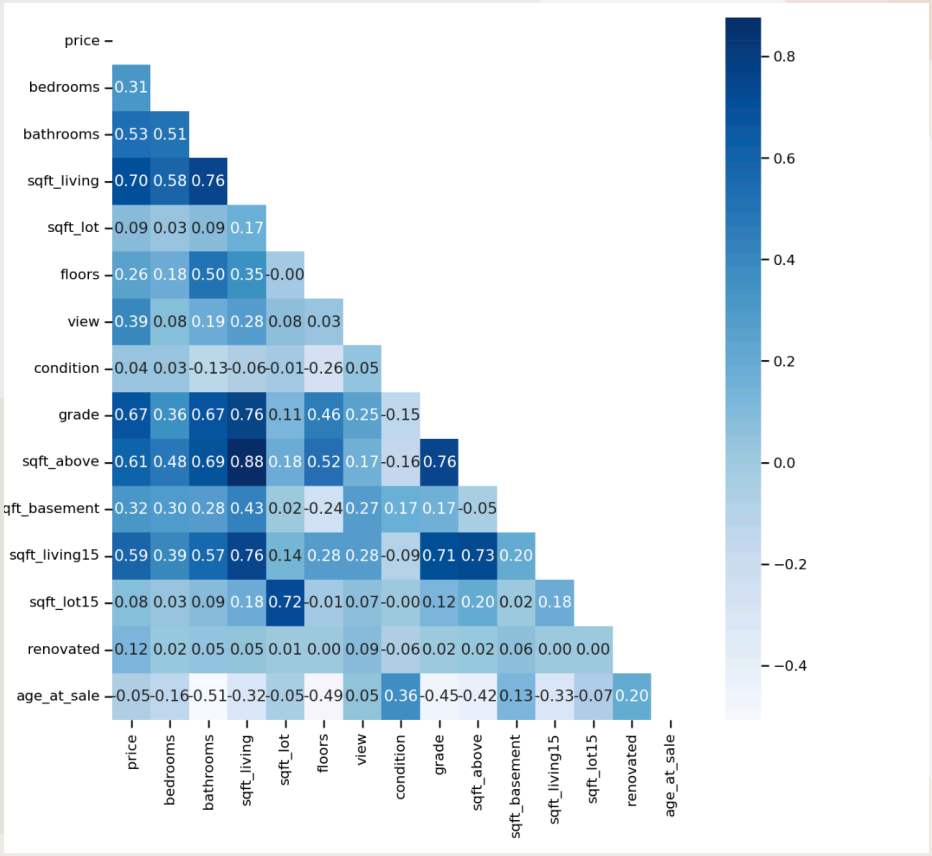
view - Quality of view from house
waterfront - whether the house was located next to a body of water
yr_renovated - the year a house was renovated (if it ever had been)

In each of these cases, we found it appropriate to fill these columns with their **modes,** which represented the overwhelming majority of values pertaining to each feature (most houses hadn't been viewed, most were not waterfront properties, etc.)

DATA PREPARATION

- Removing unnecessary features
- Checking for the completeness of data(missing values)
- Convert to types to proper types

MODELING



Correlation with Price

From heatmap and matrix plot, it seems to be there is a high correlation between

- **`sqft_living`**,
- **`sqft_above`**,
- **`grade`**

`with` **`price`**

MODELING

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:	Tue, 07 Jun 2022	**Prob (F-statistic):**	0.00
Time:	16:15:09	**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

First (Simple) Model

We conducted our first model with highest correlated feature `'sqft_living'` with our target and we saw that **49%** of the variance in the target variable can be explained by the features.

MODELING

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.639
Model:	OLS	**Adj. R-squared:**	0.639
Method:	Least Squares	**F-statistic:**	2726.
Date:	Tue, 07 Jun 2022	**Prob (F-statistic):**	0.00
Time:	16:17:09	**Log-Likelihood:**	-2.9640e+05
No. Observations:	21597	**AIC:**	5.928e+05
Df Residuals:	21582	**BIC:**	5.929e+05
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-9.727e+05	1.68e+04	-57.775	0.000	-1.01e+06	-9.4e+05
bedrooms	-4.215e+04	2079.000	-20.273	0.000	-4.62e+04	-3.81e+04
bathrooms	4.579e+04	3578.962	12.794	0.000	3.88e+04	5.28e+04
sqft_living	108.6067	19.829	5.477	0.000	69.741	147.473
sqft_lot	-0.0316	0.052	-0.602	0.547	-0.134	0.071
floors	2.729e+04	3873.336	7.045	0.000	1.97e+04	3.49e+04
view	6.902e+04	2151.125	32.085	0.000	6.48e+04	7.32e+04
condition	2.096e+04	2546.010	8.232	0.000	1.6e+04	2.6e+04
grade	1.195e+05	2307.257	51.776	0.000	1.15e+05	1.24e+05
sqft_above	57.6153	19.798	2.910	0.004	18.810	96.420
sqft_basement	60.3652	19.650	3.072	0.002	21.850	98.881
sqft_living15	20.6684	3.681	5.615	0.000	13.454	27.883
sqft_lot15	-0.5303	0.080	-6.615	0.000	-0.687	-0.373
renovated	3.872e+04	8684.702	4.459	0.000	2.17e+04	5.57e+04
age_at_sale	3570.8733	71.860	49.692	0.000	3430.022	3711.725

Omnibus:	17273.621	**Durbin-Watson:**	1.982
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1351836.123
Skew:	3.303	**Prob(JB):**	0.00
Kurtosis:	41.192	**Cond. No.**	5.71e+05

Second(Multiple) Model

If we consider all features for the model . R-squared seems to be higher than simple model. So we can say that we captured better model and still not enough for the best fit model

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.628
Model:	OLS	**Adj. R-squared:**	0.628
Method:	Least Squares	**F-statistic:**	3208.
Date:	Tue, 14 Jun 2022	**Prob (F-statistic):**	0.00
Time:	16:40:07	**Log-Likelihood:**	-5144.0
No. Observations:	20928	**AIC:**	1.031e+04
Df Residuals:	20916	**BIC:**	1.041e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.9178	0.031	347.632	0.000	10.856	10.979
bedrooms	-0.0342	0.003	-10.603	0.000	-0.041	-0.028
bathrooms	0.0887	0.005	17.280	0.000	0.079	0.099
sqft_living	0.1450	0.005	27.194	0.000	0.135	0.155
floors	0.0836	0.005	16.460	0.000	0.074	0.094
view	0.0636	0.003	20.399	0.000	0.058	0.070
condition	0.0469	0.004	12.949	0.000	0.040	0.054
grade	0.2027	0.003	61.280	0.000	0.196	0.209
sqft_living15	0.0681	0.004	18.313	0.000	0.061	0.075
sqft_lot15	-0.0733	0.007	-11.213	0.000	-0.086	-0.060
renovated	0.0373	0.012	3.001	0.003	0.013	0.062
age_at_sale	0.0057	0.000	55.606	0.000	0.005	0.006

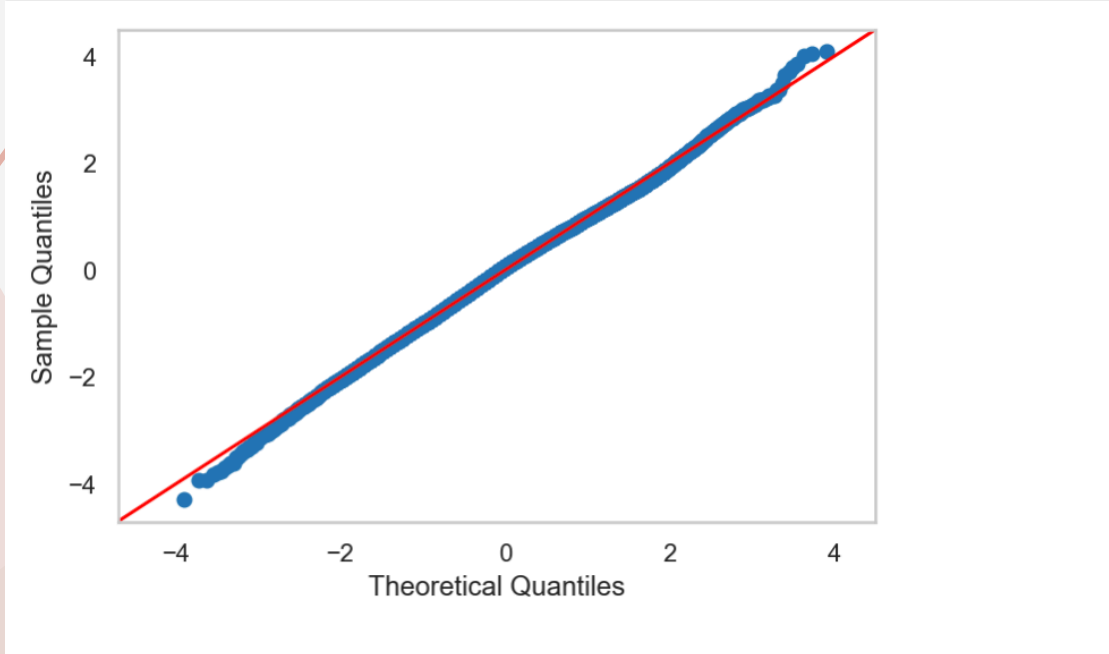
Omnibus:	60.476	**Durbin-Watson:**	1.968
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	62.266
Skew:	-0.117	**Prob(JB):**	3.01e-14
Kurtosis:	3.129	**Cond. No.**	789.

Final Model

MODELING

After dropped the not statistically significant features, considered multicollinearty issues and removed the outliers our final model became **63%** of the variance in the target variable can be explained by the features.

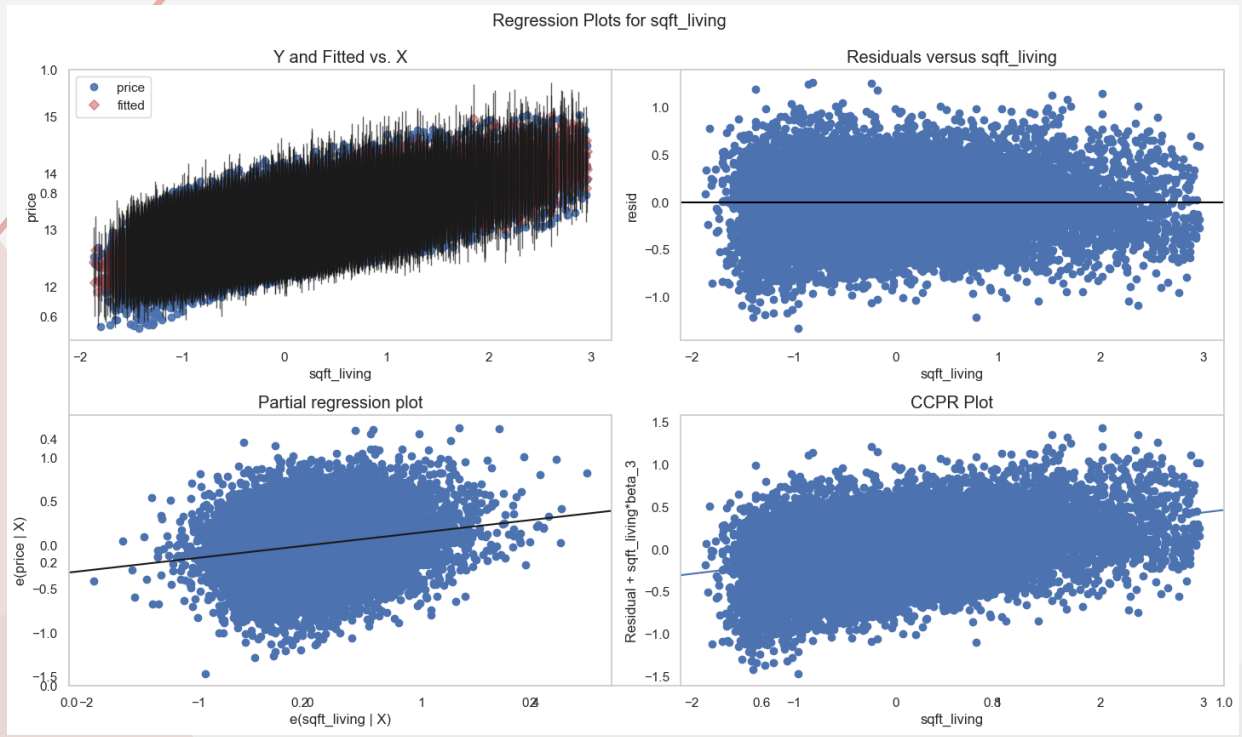
CHECKING FOR ASSUMPTIONS



Final Model

- Final model met linearty assumptions

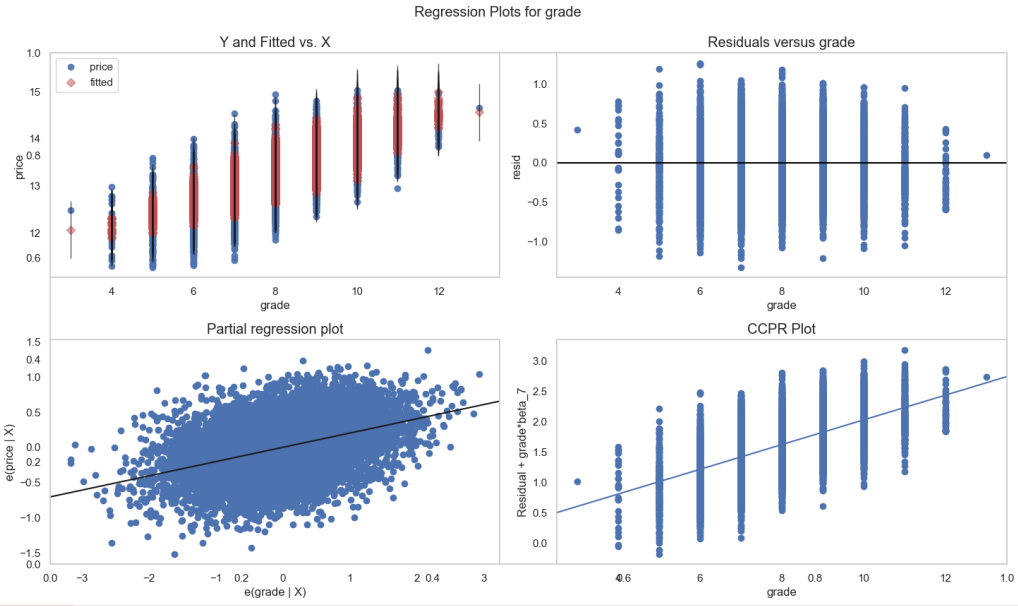
CHECKING FOR ASSUMPTIONS



sqft_living

- sqft_living is **corraleted** with Target
- Residuals are somewhat **homoskedastic** (meaning the variance doesnt decrease or increase as the independent variable gets bigger)

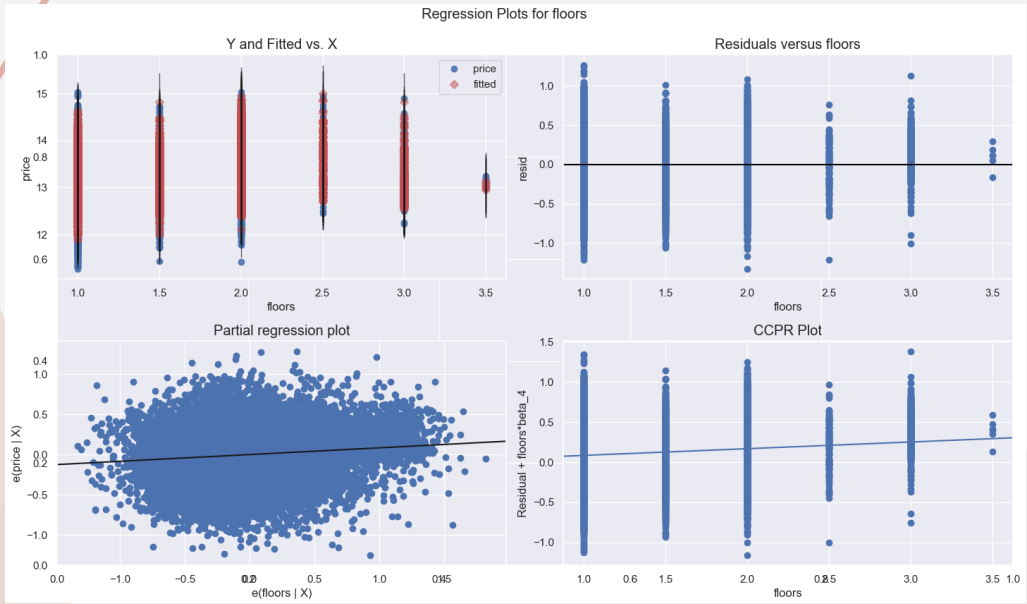
CHECKING FOR ASSUMPTIONS



Grade

- Grade is **corraleted** with Target
- Residuals are somewhat **homoskedastic**

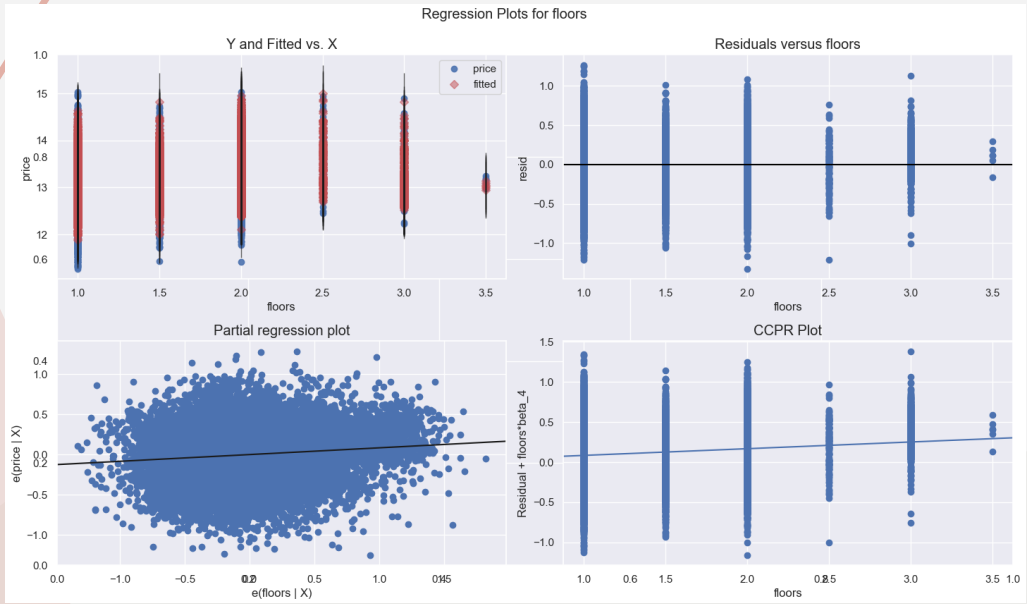
CHECKING FOR ASSUMPTIONS



Floors

- Floors is not **corraleted** with Target
- Residuals are not **homoskedastic**

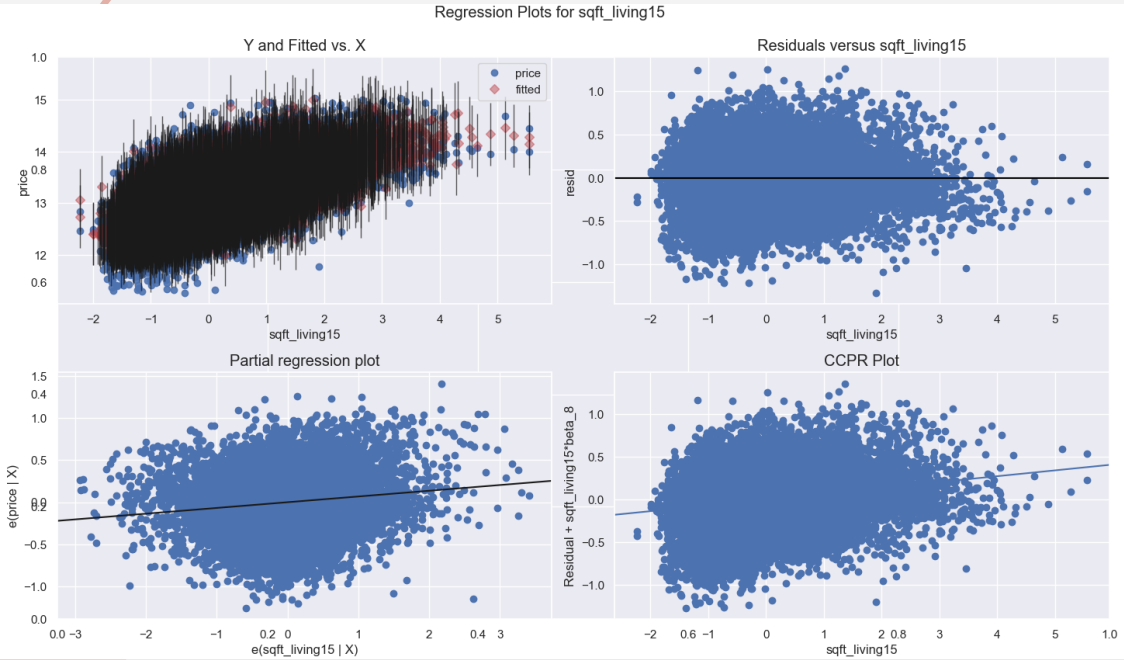
CHECKING FOR ASSUMPTIONS



Floors

- Floors is not **corraleted** with Target
- Residuals are not **homoskedastic**

CHECKING FOR ASSUMPTIONS

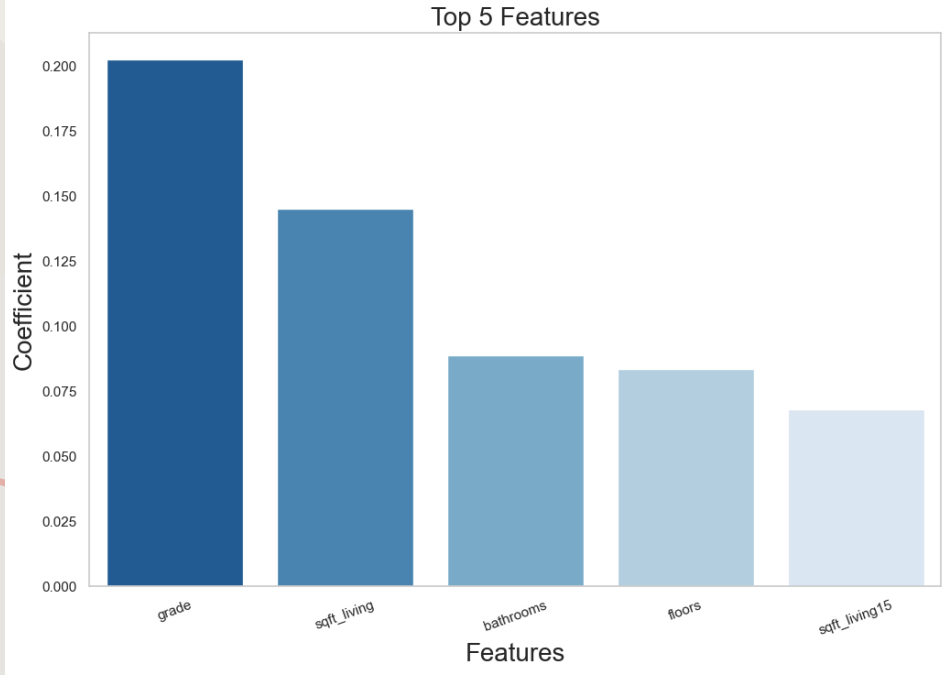


sqft_living15

- sqft_living15 is **corraleted** with Target
- Residuals are not **homoskedastic**

CONCLUSION

Interpreting Regression Coefficients



Coefficients

When we increase the features with one unit the price will increase in the following way:

- grade : **+20.27%**
- sqft_living : **+14.50%**
- bathrooms : **+8.87%**
- floors : **+8.87%**

And also when we increase the features with one unit the price will decrease in the following way:

- bedrooms : **-3.42%**
- sqft_lot15: **-7.33%**

CONCLUSION

Recommendations

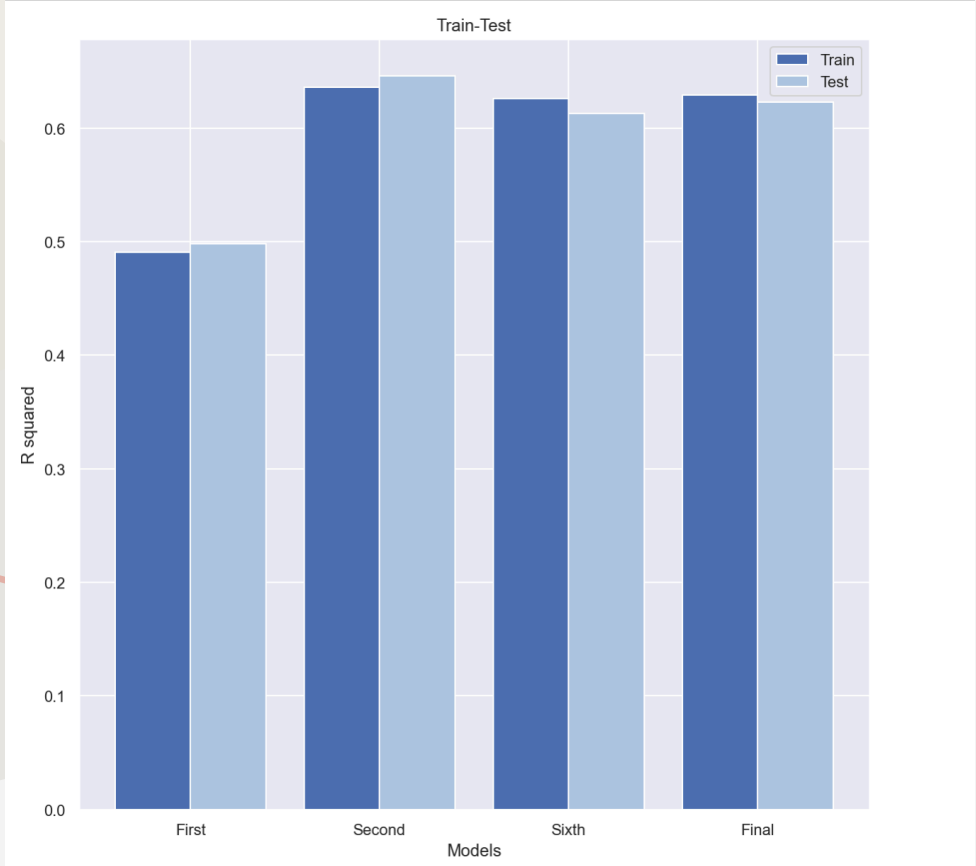
- Grade is referring to the classification based on a structures construction quality. This mainly has to do with the types of materials used and the quality of the work done. Trying to get **at least grade 8** which is an average in construction and design according to the King County Department of Assessment. It can be achieved by using better materials in both the exterior and interior finishes. As grade increases, the house price tends to be grow.

- Most preferable house floor(levels) can be reached **up to 2.5** in order to stay increased in price. Houses with floors(levels) between 3-3.5 are not desirable since prices getting sharply decreasing.

- Increasing the square footage of the living area along with the square footage of interior housing living space for the nearest 15 neighbors will also tend positively effects the price increase.

- Renovating house impacts positively its value.

CONCLUSION



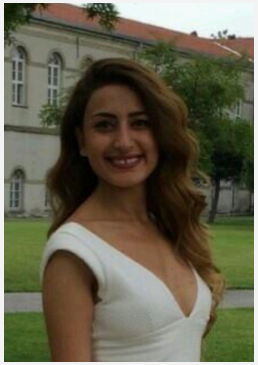
R square

Next Step

- Based on the adjusted R-squared we got more than 35% of the variance in housing prices cannot be explained by the selected principal components.In Future analysis I would like to add more features such house locations, demographics , security of a neighborhood etc to our regression model.

- Also I would like to apply machine learning tools on future home sales to find a better fit model.

THANK YOU



www.linkedin.com/in/yesim-cebeci

https://github.com/yesimcebeci