King County Price Prediction

|  |  |  |
| --- | --- | --- |
| Created by | Rajasayanam | |
| Git hub | https://github.com/sayanam/KingCountyPricePrediction.git | |
| Version | Date | Comments |
| 1 | 24-05-2020 | Initial Draft |
|  |  |  |

Contents

[Problem Statement 3](#_Toc41222422)

[Objective 3](#_Toc41222423)

[EDA and anomaly detection 3](#_Toc41222424)

[Univariate analysis 3](#_Toc41222425)

[id 3](#_Toc41222426)

[Price 3](#_Toc41222427)

[bedrooms 4](#_Toc41222428)

[bathrooms 4](#_Toc41222429)

[sqft\_living 5](#_Toc41222430)

[sqft\_lot 5](#_Toc41222431)

[floors 6](#_Toc41222432)

[waterfront 7](#_Toc41222433)

[View 7](#_Toc41222434)

[condition 7](#_Toc41222435)

[grade 8](#_Toc41222436)

[sqft\_above 8](#_Toc41222437)

[sqft\_basement 9](#_Toc41222438)

[yr\_build 10](#_Toc41222439)

[yr\_renovated 10](#_Toc41222440)

[zipcode 10](#_Toc41222441)

[Columns Ignored 10](#_Toc41222442)

[Bivariate analysis 10](#_Toc41222443)

[Analysis on continuous variables 10](#_Toc41222444)

[Analysis on Categorical but numerical variable 11](#_Toc41222445)

[New Features 12](#_Toc41222446)

[Analysis on the newly created Features 12](#_Toc41222447)

[Model Building 12](#_Toc41222448)

[Model Summary 12](#_Toc41222449)

[Conclusion 14](#_Toc41222450)

# Problem Statement

The client being a property investor is looking to acquire properties in King County area.

The client wanted to understand the house prices for those houses for which price is not listed with the help of the houses for which selling prices are listed.

# Objective

To build a model with the best R2 and least RMSE value for price prediction of housing using Linear Regression.

# EDA and anomaly detection

* Data provided was already subdivided into train, validation and test data.

|  |  |  |
| --- | --- | --- |
| **File name** | **Type of Data** | **No. of Rows** |
| wk3\_kc\_house\_train\_data.csv | Training | 9761 |
| wk3\_kc\_house\_valid\_data.csv | Validation | 9635 |
| wk3\_kc\_house\_test\_data.csv | Test Data | 9761 |

* Exploratory analysis has been performed on training data set. The inference and finding of the analysis are assumed to be same even in case of validation data set.
* There are no missing values in the given dataset
* Attributes: id, price, bedrooms, bathrooms, sqft\_living, sqft\_loft, floors, waterfront, view, condition, grade, sqft\_above, sqft\_basement, yr\_built, yr\_renovated, zipcode, lat, long, sqft\_living15, sqft\_lot15

## Univariate analysis

### id

**Attribute Description**: id is supposed to be a number that should be able to address each row elements uniquely.  
**Scale of Measure**: Nominal Scale  
**Observation**:

There are 40 records having duplicate id values (i.e.) a total of 80. Each data is having two duplicates as per the training data. Only difference among the data are with date and price attributes.

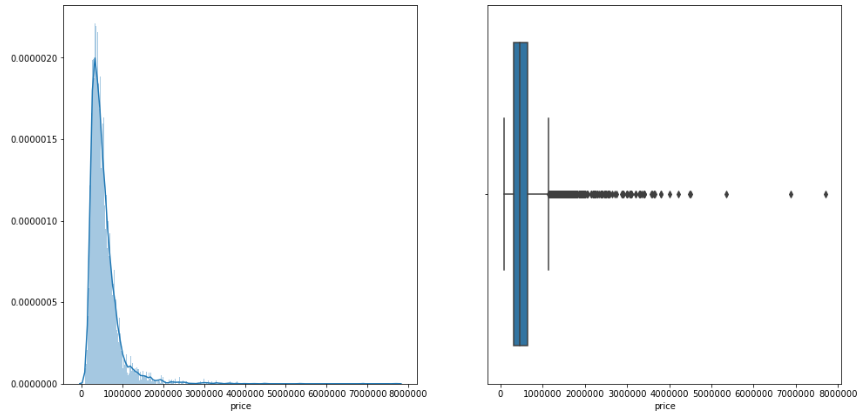
So, the record with the latest date could only be considered for training the data.

### Price

**Attribute Description:**It is the price of the house. It is our target variable.  
**Scale of Measure**: Ratio Scale  
**Observation**:

1) Since our target variable is ratio scale, regression techniques can be preferred for training the data

2) Data is right skewed and has 5% of outliers.

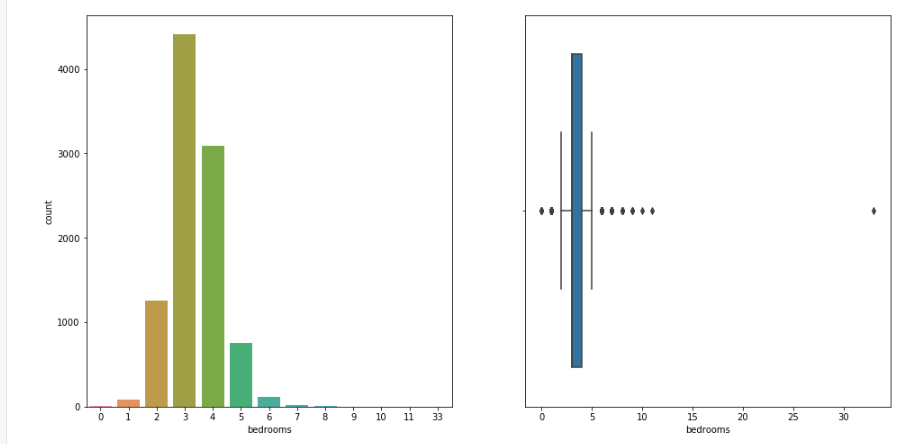


### bedrooms

**Attribute Description:**  Represents the number of bedrooms  
**Scale of Measure:** Ordinal Scale  
**Observation:**

1) As per outlier analysis 6,1,7,8,9,0,11,10,33 are considered to be outliers. But considering them as outliers would not be ideal.

2) As per the percentage of no. of number of bedrooms available, percentages less 0.05 could be considered as outliers. (i.e.) 9, 10, 11, 33

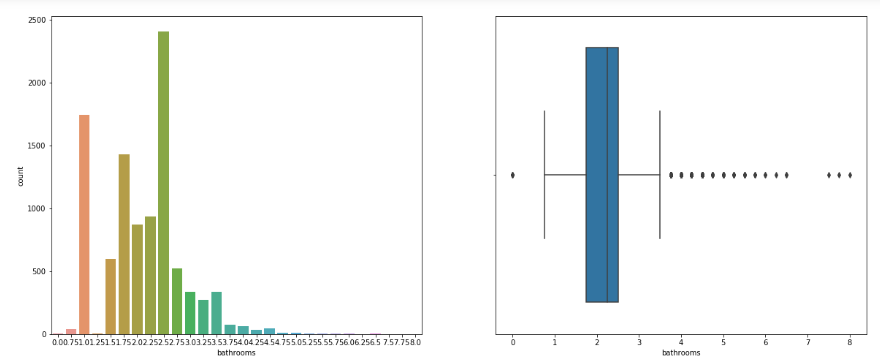


### bathrooms

**Attribute Description:** bathrooms - Represents the number of bathrooms  
**Scale of Measure:** Ordinal Scale  
**Observation:**

1) As per outlier analysis 0, 3.75, 4, 4.25, 4.5, 4.75, 5, 5.25, 5.5 5.75, 6, 6.25, 6.5, 7.5, 7.75, 8 are considered to be outliers. But considering them as outliers would not be ideal.

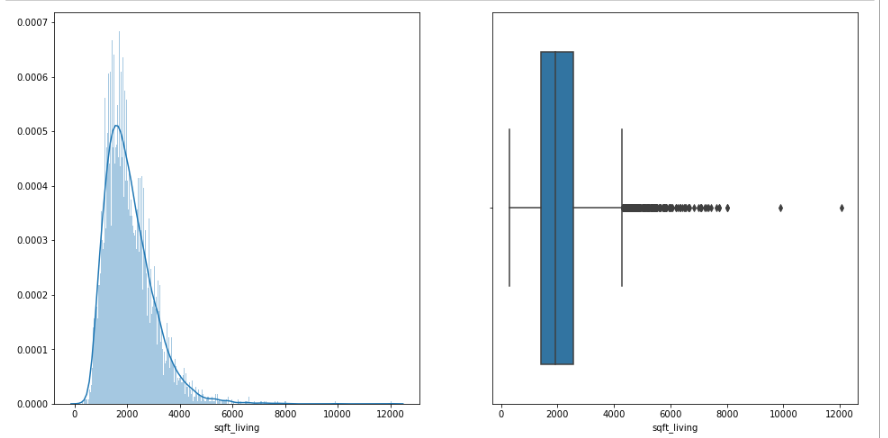
2) As per the percentage of no. of number of bedrooms available, 3% could be considered as outliers.



### sqft\_living

**Attribute Description:** Represents the area occupied by the building  
**Scale of Measure:** Ratio Scale  
**Observations:**

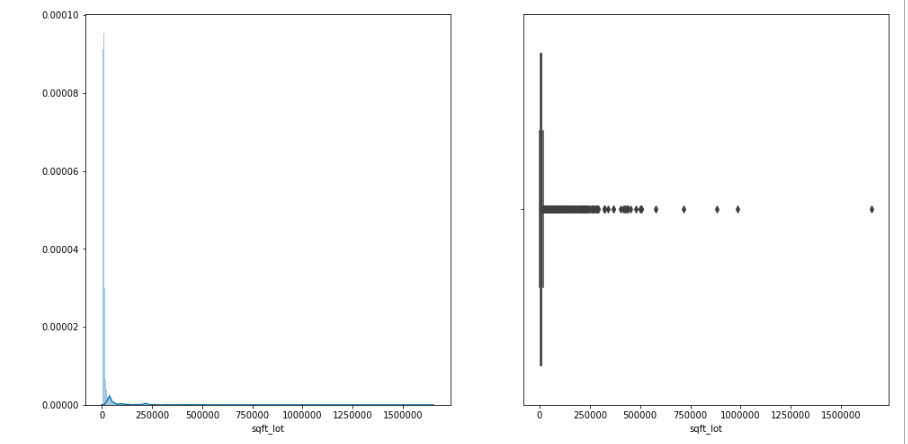
There are about 2% of the data that are considered to be outliers. But on the exterior look, this attribute need not undergo an outlier treatment.



### sqft\_lot

**Attribute Description:** sqft\_lot - Represents the area in which the building is built  
**Scale of Measure:** Ratio Scale  
**Observations:**

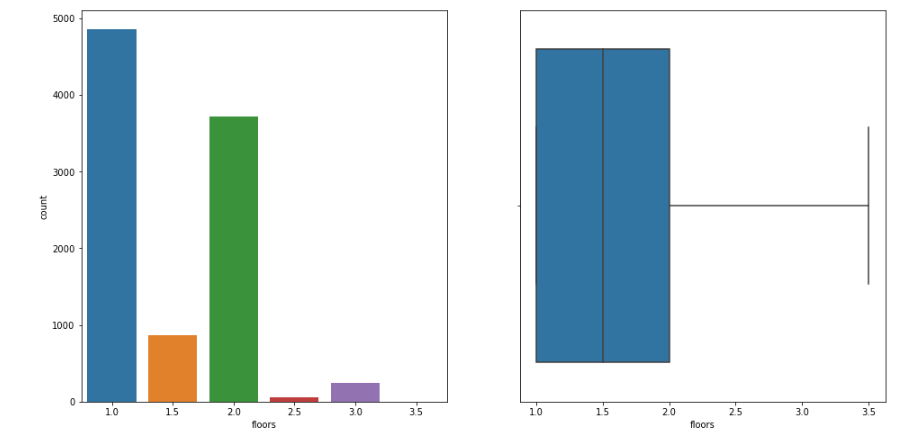
There are about 11% of the data that are considered to be outliers. But on the exterior look, this attribute need not undergo an outlier treatment.



### floors

**Attribute Description:** Floors - Represents the number of floors in a building  
**Scale of Measure:** Ordinal Scale  
**Observations:**

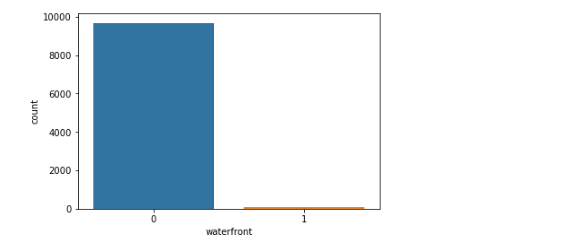
It is a categorical variable. Might not require one hot encoding, as it is already a numerical. But the interpretation of such variable in linear regression could be inconclusive.



### waterfront

**Attribute Description:** waterfront - Says if the property has a waterfront  
**Scale of Measure:** Nominal Scale  
**Observations:**

It is a binary variable where 1 denotes the presence of waterfront and 0 the absence of it. As per the training data there are only 0.84 of the property have waterfront.

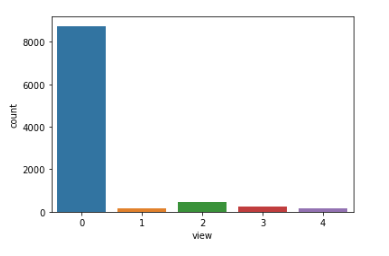


### View

**Attribute Description:** view - Represents the numerical representation of how good view of the house. Number varies from 1 to 4(Assumption: 1 is poor view and 4 being the view is good)  
**Scale of Measure:** Ordinal  
**Observations:**

1) Most of the house listed do not have a good view

2) It’s is already numerical, would not need one hot encoding for linear regression.

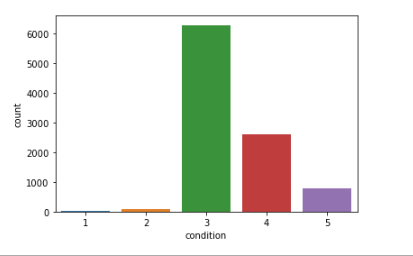


### condition

**Attribute Description:** Condition - Represents the condition of house. Number varies from 1 to 5(Assumption: 1 is the condition is very bad and 5 being the condition is very good)  
**Scale of Measure:** Ordinal  
**Observations:**

1) Most of the house listed are list in condition rating no 3 which could be understood as moderate condition.

2) It’s is already numerical, would not need one hot encoding for linear regression.

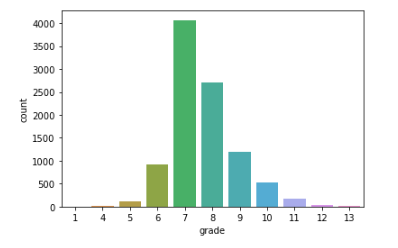


### grade

**Attribute Description:** grade - Numerical grade varying from 1 to 13(Assumption: Increase in graded number means increase in the quality)  
**Scale of Measure:** Ordinal  
**Observations:**

1) Most of the grade appear between 6 to 10.

2) It’s is already numerical, would not need one hot encoding for linear regression.

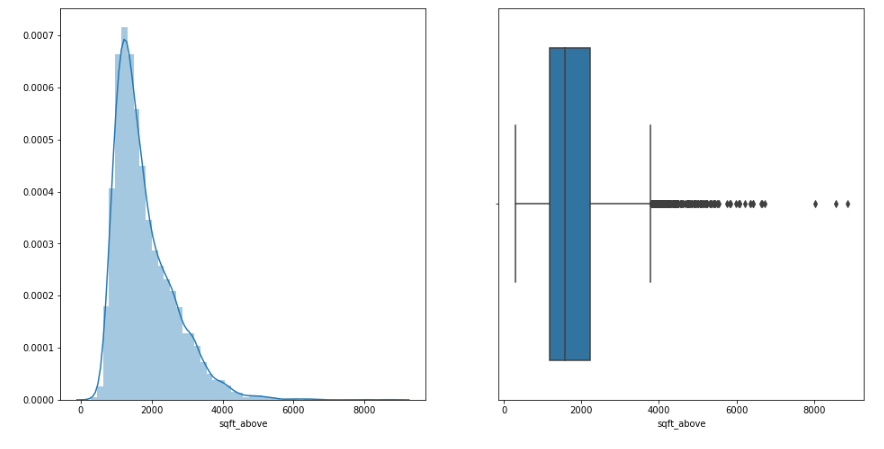


### sqft\_above

**Attribute Description:** sqft\_above - Represents the area of the building that are not inclusive of basement  
**Scale of Measure:** Ratio Scale  
**Observations:**

1) There are about 2.7% of the data that are considered to be outliers. But on the exterior look, this attribute need not undergo an outlier treatment.

2) This variable can be avoided as sqft\_living is the sum of sqft\_above + sqft\_basement



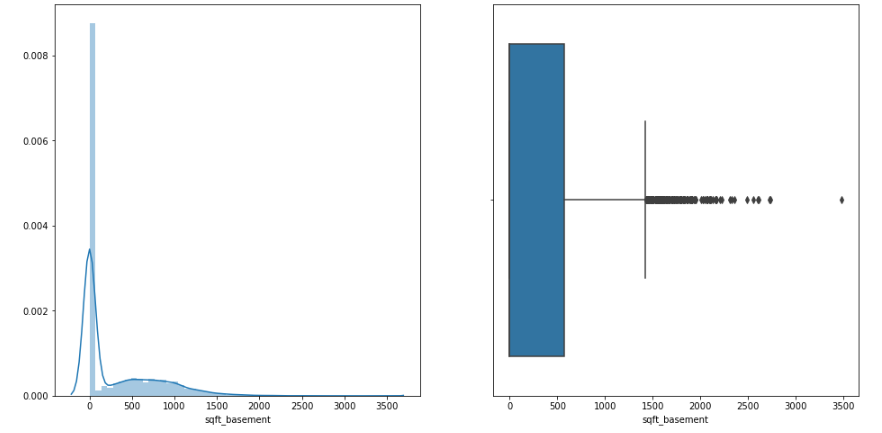
### sqft\_basement

**Attribute Description:** sqft\_basement - Represents the area of the building that are not inclusive of basement  
**Scale of Measure:** Ratio Scale  
**Observations:**

1) There are about 2.7% of the data that are considered to be outliers. But on the exterior look, this attribute need not undergo an outlier treatment.

2) This variable can be avoided as sqft\_living is the sum of sqft\_above + sqft\_basement

3) Another binary variable can be constructed based on if the basement is present or not. 1 for basement present and in case of 0 basement is not there.



### yr\_build

**Attribute Description:** yr\_built - Year in which the building was built  
**Scale of Measure:** Interval Scale  
**Observations:**

1) The number of building built increases with time.

2) It’s the year in which the building was built. Not exactly sure of how this could be used in regression. But this could provide us with another continuous column of age of the building which could definitely be used in regression.

### yr\_renovated

**Attribute Description:** yr\_renovated - Year in which the buildings are renovated  
**Scale of Measure:** Interval Scale  
**Observations:**

1) The attribute has renovated year values to be a year or 0. 0 represents there were no renovations on the building.

2) Only 4% of the building are renovated.

3) Not entirely sure how this attribute could contribute for my regression model. But a binary attribute can be created with yes/no for if\_renovated.

### zipcode

**Attribute Description:** zip code  
**Scale of Measure:** Nominal Scale  
**Observations:**

One hot coding is required for this data for it to be applied in linear regression

### Columns Ignored

lat, long - Yet to figure how this would help in model building.

sqft\_living15 and sqft\_lot15 - unaware of the meaning of the attribute.

## Bivariate analysis

### Analysis on continuous variables

**Observation:**

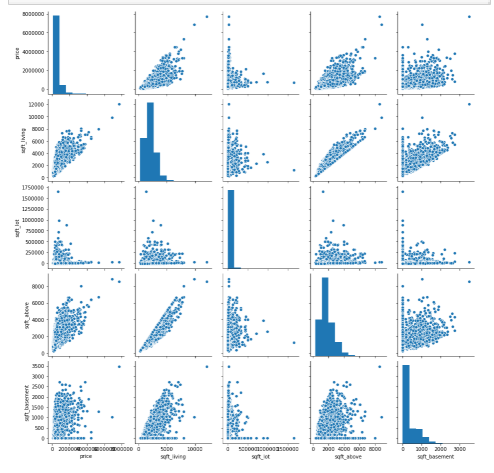
1) Correlation with Target Variable (Price)

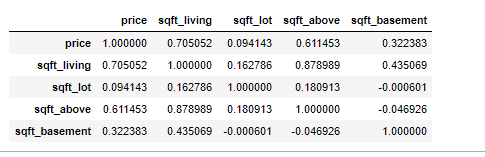
* Attributes with strong co-relation (mod value > 0.5): - sqft\_living, sqft\_above
* Attributes with medium co-relation (mod value < 0.5 && mod value > 0.1): - sqft\_basement
* Attributes with low co-relation (value close to zero): - sqft\_lot

2) Attributes with very high correlation among themselves:

* sqft\_living with sqft\_above
* sqft\_living with sqft\_basement

Note: The two attributes that have strong correlation with Price, is highly corelated among themselves. All three should not be used in building linear model as they will lead to multicollinearity.





### Analysis on Categorical but numerical variable

**Observations:**

1) Correlation with Target Variable (Price)

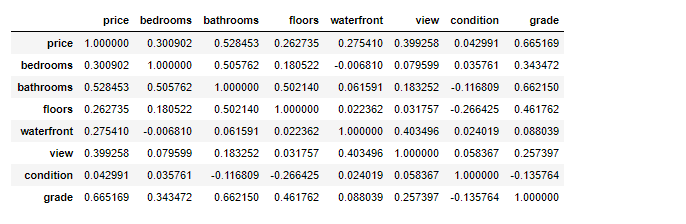
* Attributes with strong co-relation (mod value > 0.5): - bathrooms, grade
* Attributes with medium co-relation (mod value < 0.5 && mod value > 0.1): - bedrooms, floors, waterfront, view
* Attributes with low co-relation (value close to zero): - condition

2) Attributes with correlation among themselves:

* bathrooms and grades are highly colinear
* grades are colinear with bedrooms, bathrooms, floors
* bathrooms are colinear with bedrooms, floors
* View and waterfront are moderately colinear

Note:

* floors and bedrooms attributes cannot be used as they are colinear with both bathrooms and grade which have high collinearity with price
* View and waterfront cannot be used together



### New Features

#### building\_age

**Description**: Derived from yr\_built, provides us a continuous attribute which represents the age of the building at the time when entry was made

#### if\_yr\_renovated

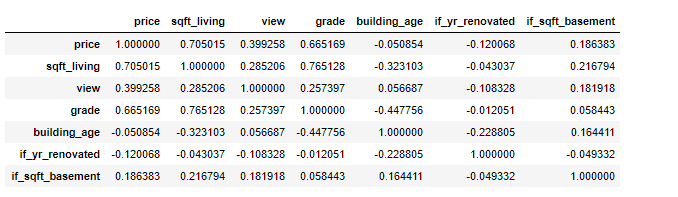
**Description**: Derived yr\_renovated. 0 - building is not renovated, 1 - building is renovated

#### if\_sqft\_basement

**Description:** Derived sqft\_basement. 0 - no basement, 1 - presence of basement

### Analysis on the newly created Features

**Observation:** None of the newly created attribute seem to have any linear correlation with our Target Variable.



# Model Building

After exploratory data analysis on the training data, began building models step by step with inclusion and exclusion of variables. Initially the selection of attributes was made solely based on the correlation value of the attributes with the target variable (Price) and the attributes with no correlation among each other.

There were series of model build based on various reason, which are listed on the model summary subsection.

## Model Summary

The following table captures the types of model build, reason behind the selection of the features and their results.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **SNO** | **Description** | **Predictors** | **Training** | | **Validation** | | **Comments** |
| **R2 (%)** | **RMSE** | **R2(%)** | **RMSE** |
| 1 | Contains only continuous variables, which has high correlation with price and less collinearity with other attributes | ­ sqft\_living | 49 | 269129 |  |  | 1) Initial model with continuous variable. 2) sqft\_living has high correlation with price, significant variable in price prediction. |
| 2 | Contains only categorical numerical variables which have high collinearity with price and no collinearity with other attributes | ­ grade  ­ view | 49 | 268834 |  |  | 1) Initial model with all the categorical numerical variables. 2) grade has high correlation with price, could be significant variable in price prediction. |
| 3 | Model to check accuracy with sqft\_living and view | ­ sqft\_living  ­ view | 53 | 257436 |  |  | 1) grade and sqft\_living have high correlation value 2) model 3 proves better than model 2, but yet we will retain model 2 and 3 for further improvements. |
| 4 | Including one hot encoded zip code with model 2 attributes | ­ grade ­ view ­ zipcode | 69 | 210551 |  |  | Inclusive of zip code significantly reduces the prediction variance |
| 5 | Including one hot encoded zip code with model 3 attributes. Note: encoded in this context means one hot encoded | ­ sqft\_living ­ view  ­ encoded zipcode | 75 | 186159 | 75 | 175045 | Model 5 delivers better result compared with the model having grade. Will be enhancing on Model 5 going ahead. So, using the validation data to determine the prediction. |
| 6 | Including waterfront and grade | ­ sqft\_living ­ view  ­ encoded zipcode ­ grade ­ waterfront | 79 | 172957 | 79 | 162966 | Since our focus is on prediction and not on interpretation of models, inclusion of correlated variables improved the prediction result. |
| 7 | Applying log transform to sqft\_living and price | ­ sqft\_living\_log ­ view ­ encoded zipcode ­ grade  ­waterfront | 85 | 143681 | 86 | 129510 | Applying log transformations on sqft\_living and price improved the training and validation result. |

Model 7 with attributes sqft\_living\_log, view, one hot encoded zip code, grade and waterfront provided the best result of 85% and 86% of R2 value in training and validation data respectively and 1,45,681 and 1,29,510 of RMSE values for training and validation data respectively.

So, on testing the model with test data following are the results obtained,

|  |  |
| --- | --- |
| **R2 (%)** | **RMSE** |
| 86 | 132303 |

# Conclusion

Further analysis needs to be conducted on the model to understand the influence of outliers on the attributes and the tests for checking if the assumptions of Linear Regression are maintained.

At the point the model has a higher chance of getting the actual price between the range of predicted value 132303 at 86% of variation in data being accounted for.