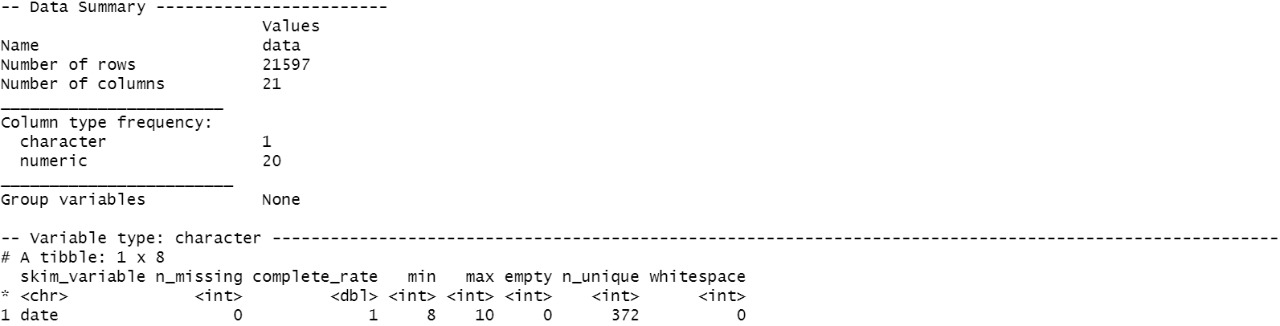
**PROJECT: HOUSE PRICE PREDICTION USING LINEAR REGRESSION**

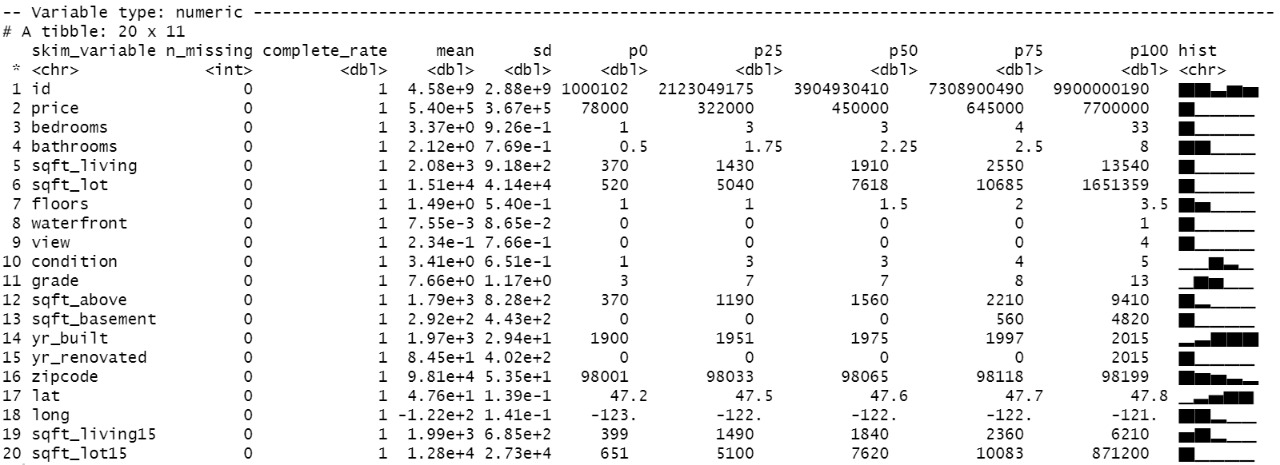
**Objective:** The aim of this report is to predict the house sales in King County, Washington State, USA using Multiple Linear Regression (MLR).

**Dataset description:** The dataset consists of house prices from King County an area in the US State of Washington, this data also covers Seattle. The dataset was obtained from Kaggle. This data was published/released under CC0: Public Domain. The dataset consisted of historic data of houses sold between May 2014 to May 2015.

The dataset consisted of 21 variables and 21613 observations.

**Dataset summary:**





**Variable description:**

id - Unique ID for each home sold

date - Date of the home sale

price - Price of each home sold

bedrooms - Number of bedrooms

bathrooms - Number of bathrooms, where .5 accounts for a room with a toilet but no shower

sqft\_living - Square footage of the apartments interior living space

sqft\_lot - Square footage of the land space

floors - Number of floors

waterfront - A dummy variable for whether the apartment was overlooking the waterfront or not

view - An index from 0 to 4 of how good the view of the property was

condition - An index from 1 to 5 on the condition of the apartment,

grade - An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.

sqft\_above - The square footage of the interior housing space that is above ground level

sqft\_basement - The square footage of the interior housing space that is below ground level

yr\_built - The year the house was initially built

yr\_renovated - The year of the house’s last renovation

zipcode - What zipcode area the house is in

lat - Lattitude

long - Longitude

sqft\_living15 – Updated square footage area of living space based on year 2015

sqft\_lot15 - Updated square footage of land space based on year 2015

**STEPS:**

1. Data cleaning:

* We checked for null values in the given dataset, found none.
* We checked for duplicated house ids and removed them.

1. Feature Engineering:

* Calculated age of house sold by extracting year from the date by subtracting it by year of being built.
* Updated the age of houses based on year renovated.
* Categorized Basement variable into 1s and 0s based on their availability and non-availability.
* Clubbed the 70 unique zip codes into four main categories namely – ‘Average, Good, very good, Posh’ grouping them based on the mean prices for each zip code.

1. EDA:

* Identified extreme data points for each 12 variables using exploratory data analysis.
* Checked the dataset for multicollinearity.

1. Model building:

* Splitting of dataset into train and test data.
* Scaled variables using scaling techniques.
* Build our model cautiously keeping an eye on p-value and used cross-validation technique to select the important/driver variables in our model. But the relationship between predicted and actual values of price weren’t showing linear relationship hence we decided to do log transformation.

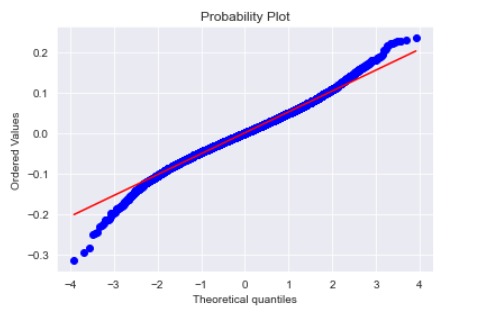
Before log transformation :

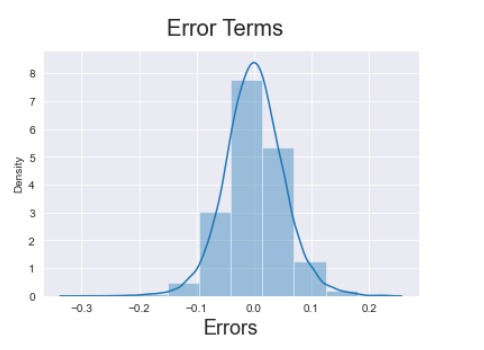


After log transformation:



* We checked the assumption of normality of residuals using Q-Q Plot and histogram as shown below:





* Checked the model accuracy using R2 – value.

