KCHPM(King County House Pricing Model)

# Introduction

I have chosen this project to use my understanding to apply regression analysis in real world scenario.

It is modeling and analyzing several variables, when the focus is on the relationship between dependent variables and one independent variable.

I have selected a dataset that contains house sale details for King County, which includes Seattle. The price of the house vary according to number of bedrooms, view, location, lot-size etc. We need to have a model to predict the price of the house taking account of attributes which influence it to great extent. It will help both seller/buyer to predict the price of house.

# Literature Review

I reviewed my course notes which I have been taught in 'ckme 642' . The professor taught us how to find the House Prices with several attributes. There were only 2-3 attributes. I also referred some of the sites to get detail information about regression analysis .

e.g

http://sites.stat.psu.edu/~sesa/stat460/Project/greg.pdf

https://ww2.amstat.org/publications/jse/v16n2/datasets.pardoe.pdf

I picked the dataset from kaggle for this project.

# Dataset

I picked the dataset from kaggle for this project.

URL:- https://www.kaggle.com/harlfoxem/housesalesprediction

|  |  |
| --- | --- |
| **Data-Original-Attributes** | **data-type** |
| id | number |
| date | date |
| price | number |
| bedrooms | integer |
| bathrooms | number |
| sqft\_living | integer |
| sqft\_lot | integer |
| floors | number |
| waterfront | integer |
| view | integer |
| condition | integer |
| grade | integer |
| sqft\_above | integer |
| sqft\_basement | integer |
| yr\_built | integer |
| yr\_renovated | integer |
| zipcode | integer |
| lat | number |
| long | number |
| sqft\_living15 | integer |
| sqft\_lot15 | integer |

Using stepAIC algorithm in R helped me to get attributes mentioned below.

|  |  |
| --- | --- |
| **Selected-Attributes** | **data-type** |
| price | number |
| bedrooms | integer |
| bathrooms | number |
| sqft\_living | integer |
| sqft\_lot | integer |
| waterfront | integer |
| view | integer |
| condition | integer |
| grade | integer |
| sqft\_above | integer |
| sqft\_basement | integer |
| yr\_built | integer |
| yr\_renovated | integer |
| zipcode | integer |
| lat | number |
| long | number |
| sqft\_living15 | integer |
| sqft\_lot15 | integer |

# Approach

compare the RMSE values for training and testing and get the best model.

Compare the models

store RMSE values for training and testing for each model

find RMSE for our models for training and testing dataset

Make separate models for each different approach

x=60%, y=[100-x]%

split the data between training and testing where training has x% and testing has y%

make separate linear regression model from training dataset

Do predictions for test datasets

Step-13

Step-12

Step-11

Step-10

Step-09

Step-08

Step-07

Step-06

Step-05

Step-04

Step-03

Step-02

Step-01

Filter the attributes using stepAIC, multi-correlation techniques, histograms,boxplots,boruta

clean data

check statistics of data

open data in R

import data from kaggle

Except below steps, all steps are self explanatory, as I have provided the detailed flowchart of this project.

## Step 03: <Check Statistics of data>

I verified the type of data as statistics is concerned like qualitative /quantitative using R function

## Step 04: <Clean Data>

Check for NA values

Check for missing value

Check for noisy value

## Step 05: <Filter attributes>

Use co-relation matrix to determine variance between quantitative data

Use ANOVA to determine variance between qualitative data

Use histogram and boxplot to check data distribution

Use stepAIC algorithm and boruta algorithm to get attributes which can influence dependent variable