KING COUNTY HOUSING

GAGANDEEP (0766564)

JARMAN SINGH (0767192)

SIMRANJIT KAUR (0773402)

URVI NILESHKUMAR PATEL (0770850)

YASHPREET SINGH (0767186)

DAB 402, M018, 2020-2021

A capstone project submitted

in partial fulfilment of the requirements for the degree of

DATA ANALYTICS FOR BUSINESS

in

Data Analysis

SAINT CLAIR COLLEGE FOR APPLIED ARTS AND TECHNOLOGY

Mississauga, ONTARIO, CANADA

SUBMITTED TO

Professor Savita Sherawat

Abstract

To foresee King County's home costs, I picked the lodging cost dataset that was sourced from Kaggle. This dataset contains house deal costs for King County, which incorporates Seattle. It incorporates homes sold between May 2014 and May 2015

Introduction:

In the course of one's life, the most costly and biggest buy that the individual makes is generally a home. People should know the sensible worth of their resources. Forecast on house cost will help the two property holders and homebuyers to settle on choices regardless of whether to sell or purchase a house at a specific cost. Notwithstanding, it is regularly hard to decide the cost of a house, as there are many elements included, like the age of the house, climate, area and so on In this work, we will apply a few relapse and prescient techniques to concentrate on house deal cost in King County, Washington, USA and investigate the best model for expectation.

A few scientists and groups have investigated the KC house dataset. For instance, highlight positioning with Random Forest, RFE, and direct models was examined, and straight models were assessed in certain works. Numerous relapses, rope relapse and k-Nearest Neighbors Regression were additionally examined.

While past works have shown convincing outcomes, the R-squared qualities (regularly <0.9) may be additionally improved. In this review, we utilize a few relapses and machine learning procedures, just as a model stacking (consolidating) way to deal with survey their forecast execution and to acquire a model that is best for expectation inside our structure.

The general thought of relapse is to look at two things:

1. Which locations within the King County area have the highest average house prices?

2. Which house attributes increase sale price?

3. Does time of the year have an impact on house sales

Software: R Studio, Excel, Word, PowerPoint.

Depiction

In this dataset the business cost of houses in King County, Seattle is available. It incorporates homes sold between May 2014 and May 2015. Prior to doing anything we should initially think about the dataset that it contains, what are its provisions and what is the design of information.
By noticing the information, we can realize that the cost is reliant upon different elements like bedrooms (which is most reliant element), washrooms, sqft_living(second most significant component), sqft_lot, floors and so forth The cost is likewise reliant upon the area of the house where it is available.
Different provisions like waterfront, see are less reliant upon the cost.

There are 21 Attributes in this dataset.

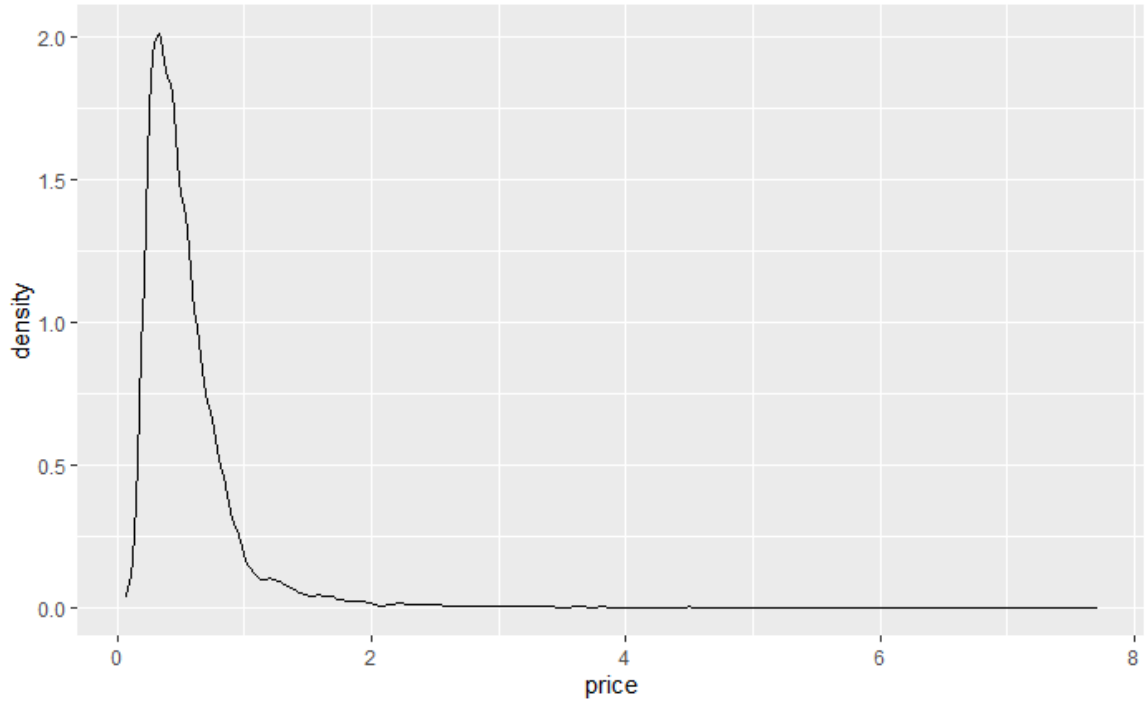
Detailed data dictionary:

Variable	Description
id	Identification
date	Date sold
price	Sale price
bedrooms	Number of bedrooms
bathrooms	Number of bathrooms
sqft_liv	Size of living area in square feet
sqft_lot	Size of the lot in square feet
floors	Number of floors
waterfront	‘1’ if the property has a waterfront, ‘0’ if not.
view	An index from 0 to 4 of how good the view of the property was
condition	Condition of the house, ranked from 1 to 5
grade	Classification by construction quality which refers to the types of materials used and the quality of workmanship. Buildings of better quality (higher grade) cost more to build per unit of measure and command higher value
sqft_above	Square feet above ground
sqft_basmt	Square feet below ground
yr_built	Year built
yr_renov	Year renovated. ‘0’ if never renovated
zipcode	5 digit zip code

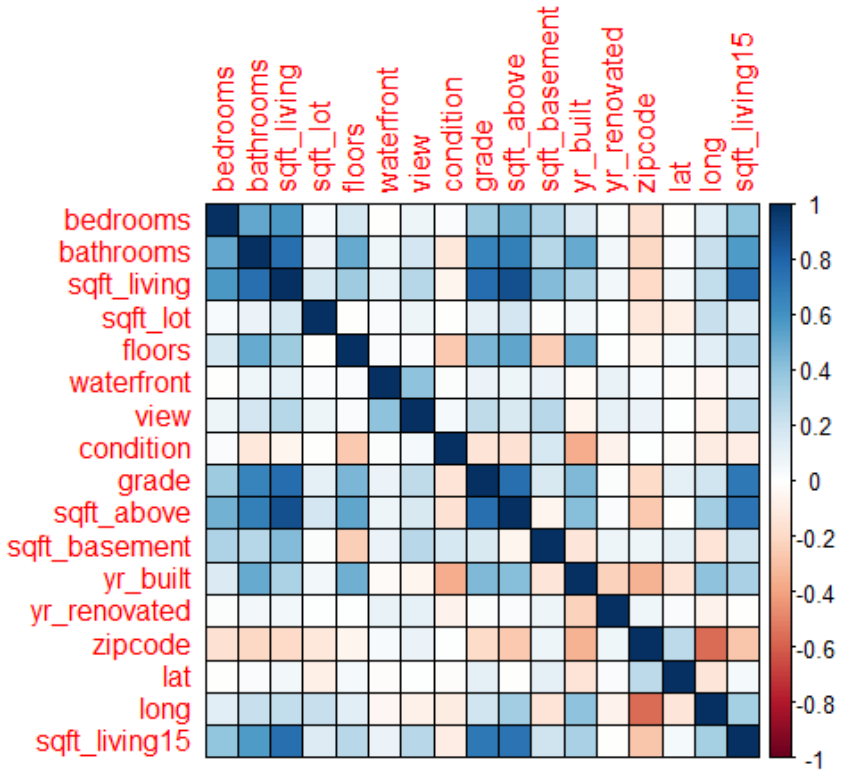
lat	Latitude
long	Longitude
squft_liv15	Average size of interior housing living space for the closest 15 houses, in square feet
squft_lot15	Average size of land lots for the closest 15 houses, in square feet
Shape_leng	Polygon length in meters
Shape_Area	Polygon area in meters

Methods:

density of the price to get a first impression on its distribution

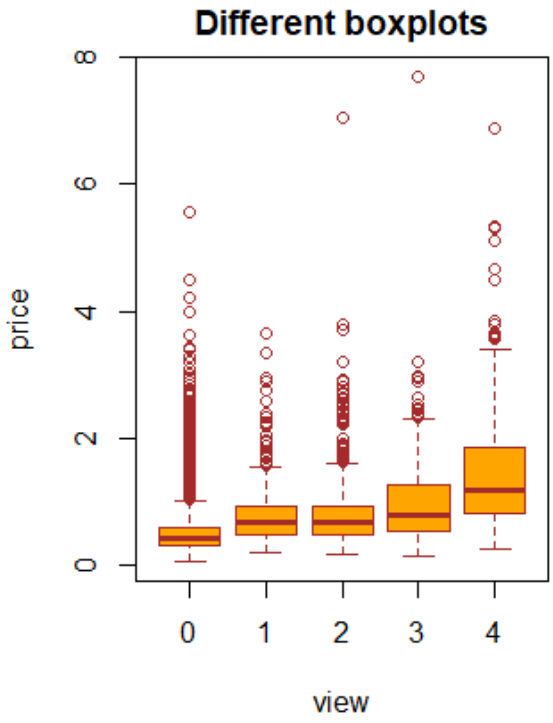


Determining the association between variables

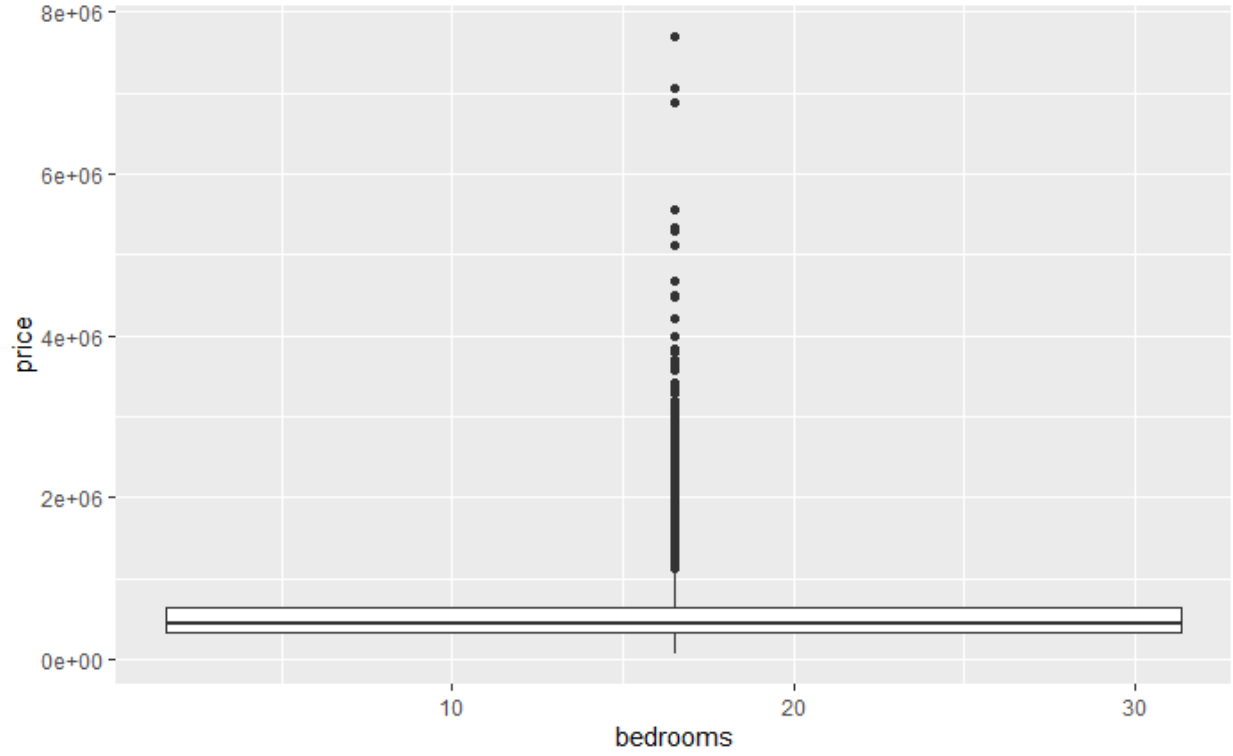


from these scatter plots, we conclude that the relationship between price and bedroom, bathroom, Sqft_living,sqft_above, sqft_basement, lat, sqft_living 15 is linear.

For the two categorical variables(view and grade) we draw boxplots to understand the relationship.



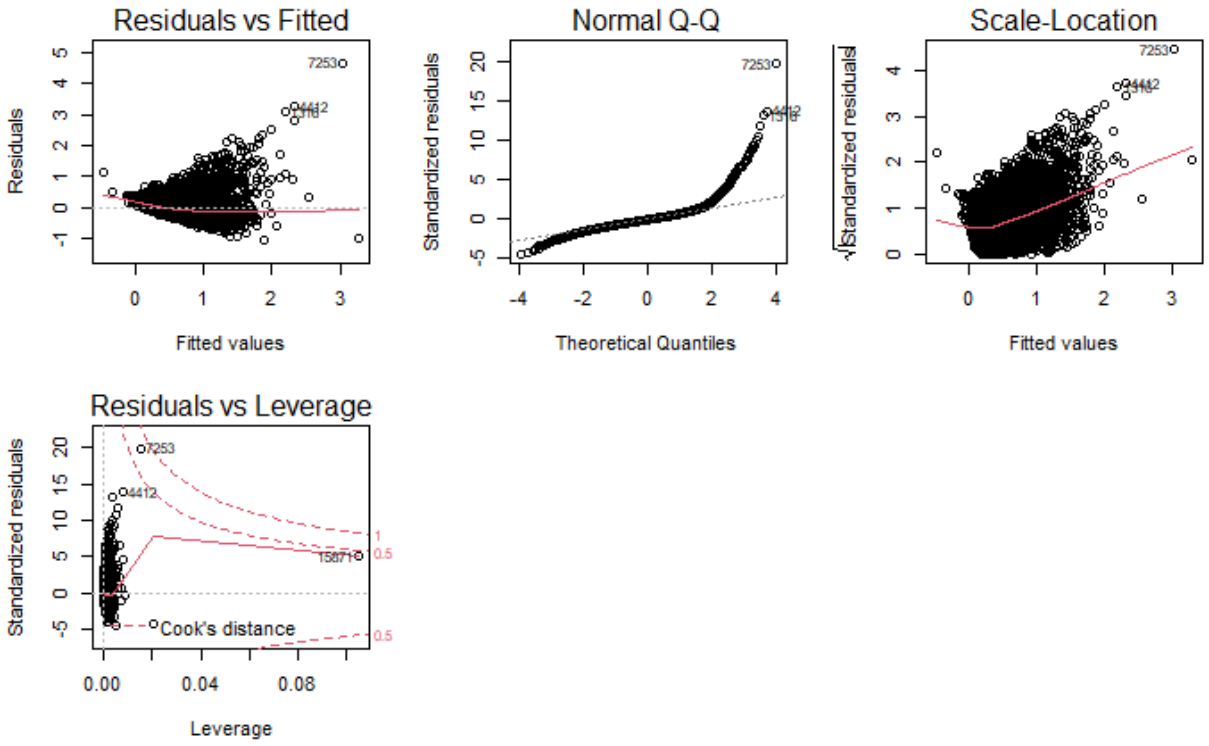
#now we check for outliers in the dependent variable(price) using a boxplot.

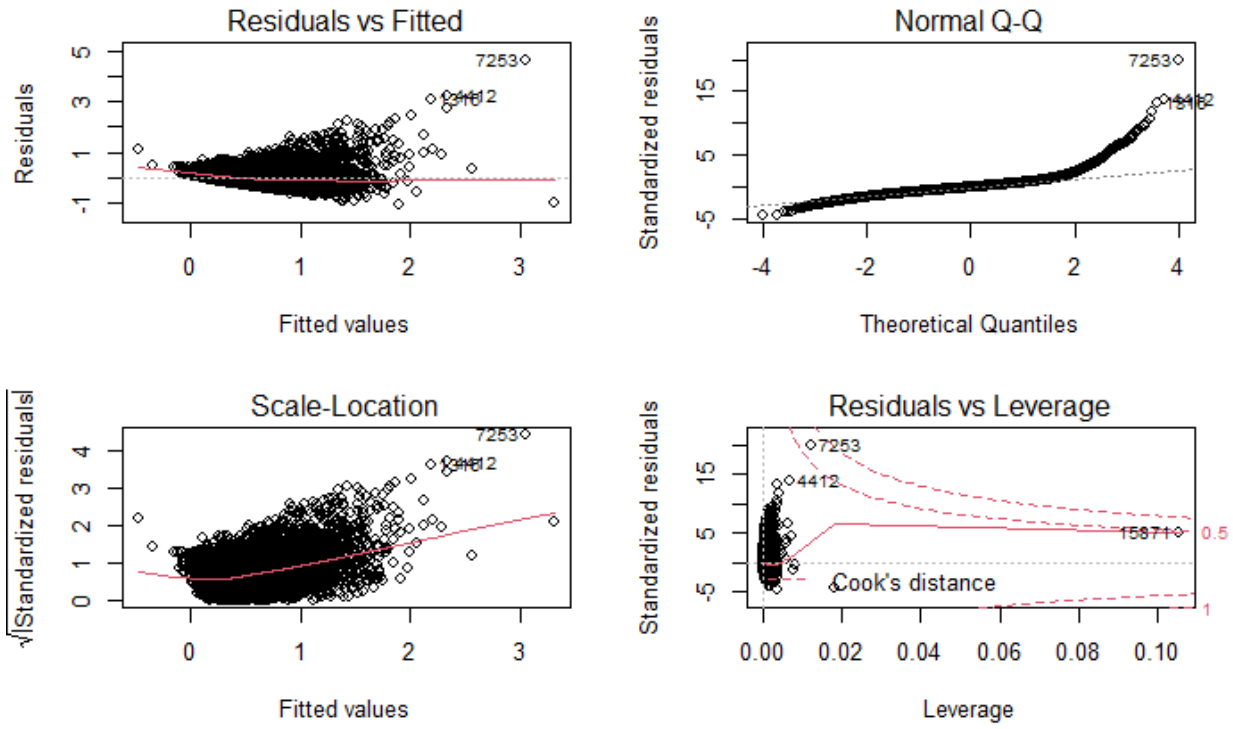


#we obtain 872 observations as outliers.

#Now we plot the data with and without outliers.

Notice the change in slope of the best fit line after removing the outliers. It is evident that if we remove the outliers to train the model, our predictions would be exagerated (high error) for larger values of price because of the larger slope.





Reference :

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

https://udspace.udel.edu/bitstream/handle/19716/21667/RR17-10.pdf?sequence=1&isAllowed=y

https://geodacenter.github.io/data-and-lab/KingCounty-HouseSales2015/