PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

B.Tech, Sem III

Session : Aug-Dec, 2018

UE17CS203 – INTRODUCTION TO DATA SCIENCE

REPORT

EXPLORATORY ANALYSIS ON (DATA SET NAME)

DATA SET LINK :	**https://www.kaggle.com/harlfoxem/housesalesprediction**
TEAM MEMBERS	**NAME : YASH GOEL** **SRN :PES1201701853** **EMAIL ID :goelyash@hotmail.com**
	NAME : YASH GOEL **SRN :PES1201701853** **EMAIL ID :goelyash@hotmail.cm**

Data set : kc_house_data .csv

Analysis of house price in US in a perticular area.

This report provides details of our attempt to analyse house price with different features of house using various methods .We will make use of data of house prices in King County, Washington State, USA for sales in 2014 and 2015.

Data set contains 19 house features plus the price and id columns ,along with 21613 observations.

Our main quetions are :

1) How the price is varying with different features of house ?
2) how the grading provided by King County grading system affects the price of the house ?
3) How is the demand of house overe time ?
4) How the percentage of living space to that of total space varies with time?
5) What number of room houses are sold the most?
6)Comparing the number of rooms in demand over time .
6)What number of floor houses are in demand?
7) Comparing the prices of house with basement and without basement.
8)how the grades based on King County grading system depends on price of house?

DATA SET

Our chosen data set is "**kc_house_data .csv** ". The Data set is collected from kaggle.com. The Data set contains 19 house features plus the price and id columns ,along with 21613 observations.

Data set contain the following columns:

1) id
   Door number for a house.Every house has different id.
2) date
   Date house was sold
3) price
   Price of the house
4) bedrooms
   Number of Bedrooms/House
5) bathrooms

Number of bathrooms/bedrooms(if its a whole number then bathroom contains tub, stower, or bott.If not tten is will eontain any one of bott)

6)sqft_living

square footage of the home

7)sqft_lot

square footage of the lot

8)floors

Total floors (levels) in house. If its not a whole number then the float value of it is considered as balcony.

9)waterfront

House which has a view to a waterfront

10)view

Has been viewed

11)condition

How good the condition is (Overall)

12)grade

overall grade given to the housing unit, based on King County grading system

13)sqft_above

square footage of house apart from basement

14)sqft_basement

square footage of the basement

15)yr_built

Built Year

16)yr_renovated

Year when house was renovated

17)zipcode

zip

18) lat

Latitude coordinate

19) long

Longitude coordinate

20)sqft_living15

Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area

21)sqft_lot15

lotSize area in 2015(implies-- some renovations)

Purpose of the Data set: To compare the price with different features of dataset.

DATA CLEANING :

Number of bedrooms and number of bathrooms:

In the given data set one of the house had 33 bedrooms, it could be a typing error, when compared to 3 bedrooms houses, pricing was not matching, therefore its safe to not include it in the analysis and has been discarded.

There were 13 entries with 0 bedroom and 3 entries with 0 bathroom, these were discarded from analysis. The other features of houses with 0 bedrooms were incomplete, so therefore it was not considered in the analysis.

Date :

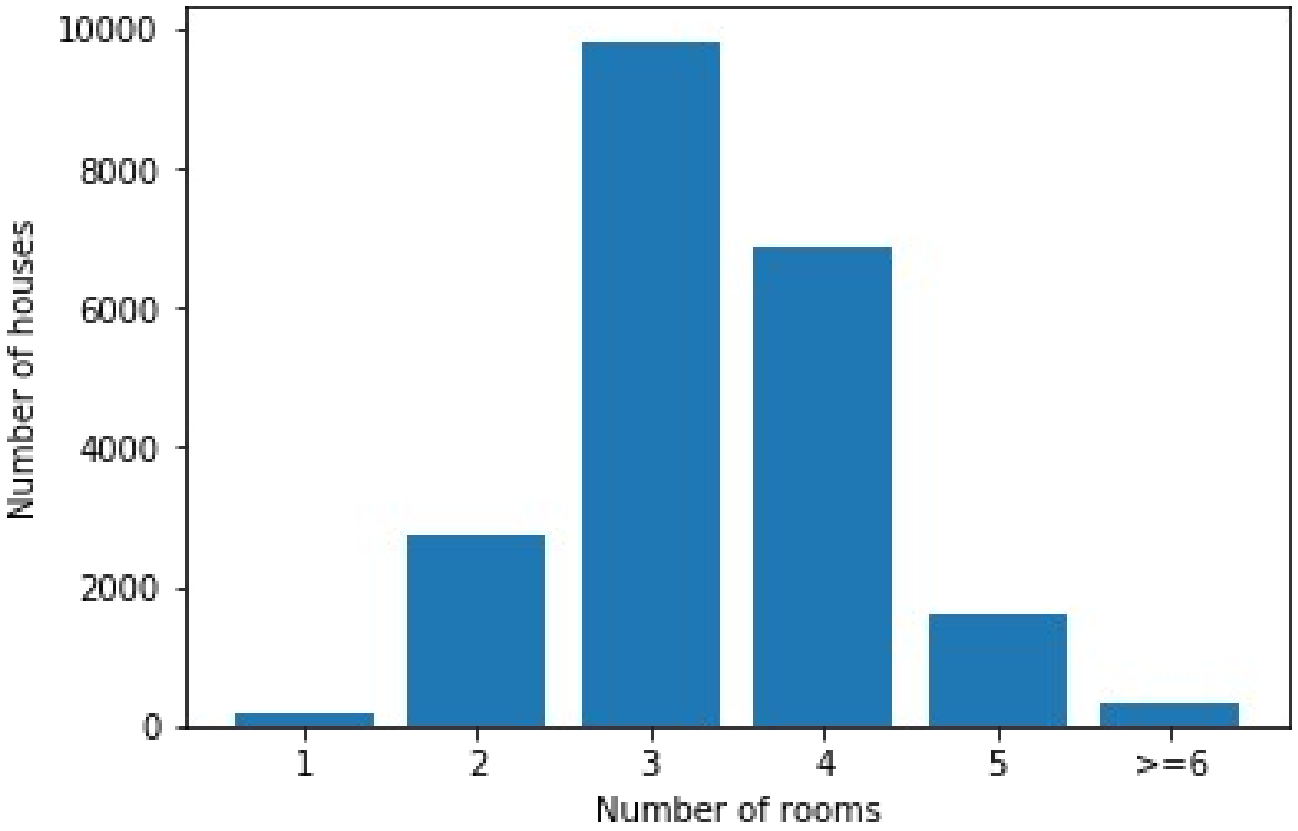
Dates were not in a perticular order ("20150312T000000") we made it 20150312

Id :

we have not used ids of the houses for any of our analysis

Analysis:

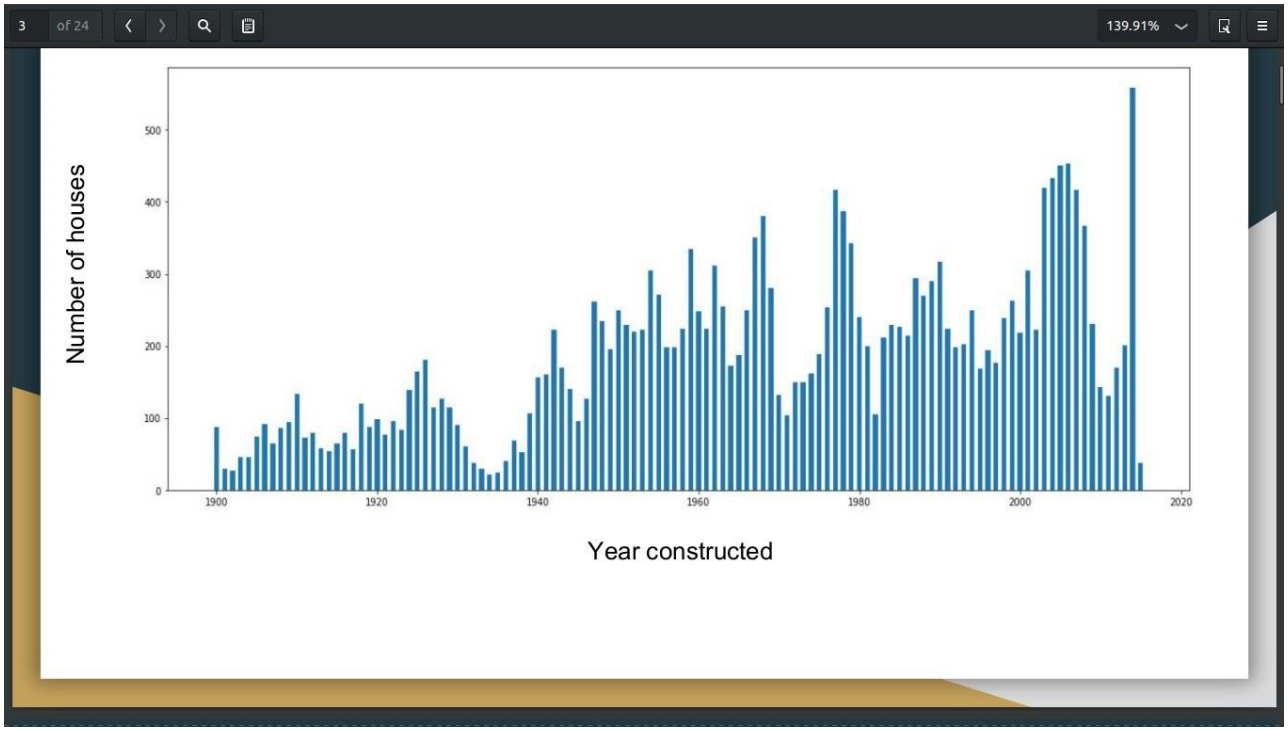
1) A bar graph for number of rooms and number of houses:



* The number of houses of having 3 rooms are large in number as compared to all other different number of room

houses having 1 room and more then or equal 6 rooms are less

2)**graph between year constructed and number of houses.**

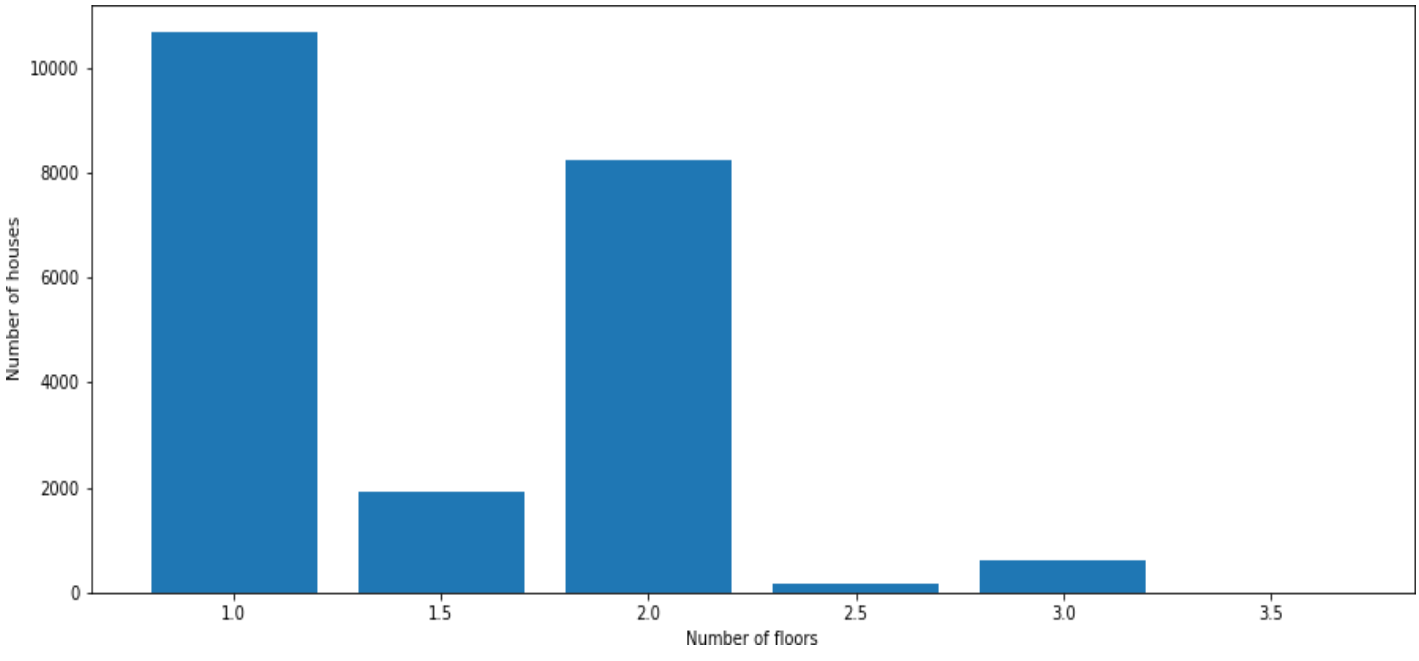


*Number of old houses bought are less, houses constructed in 2000s are more sought.

*By this we can infer that demand for recently built houses is slightly more that old houses.

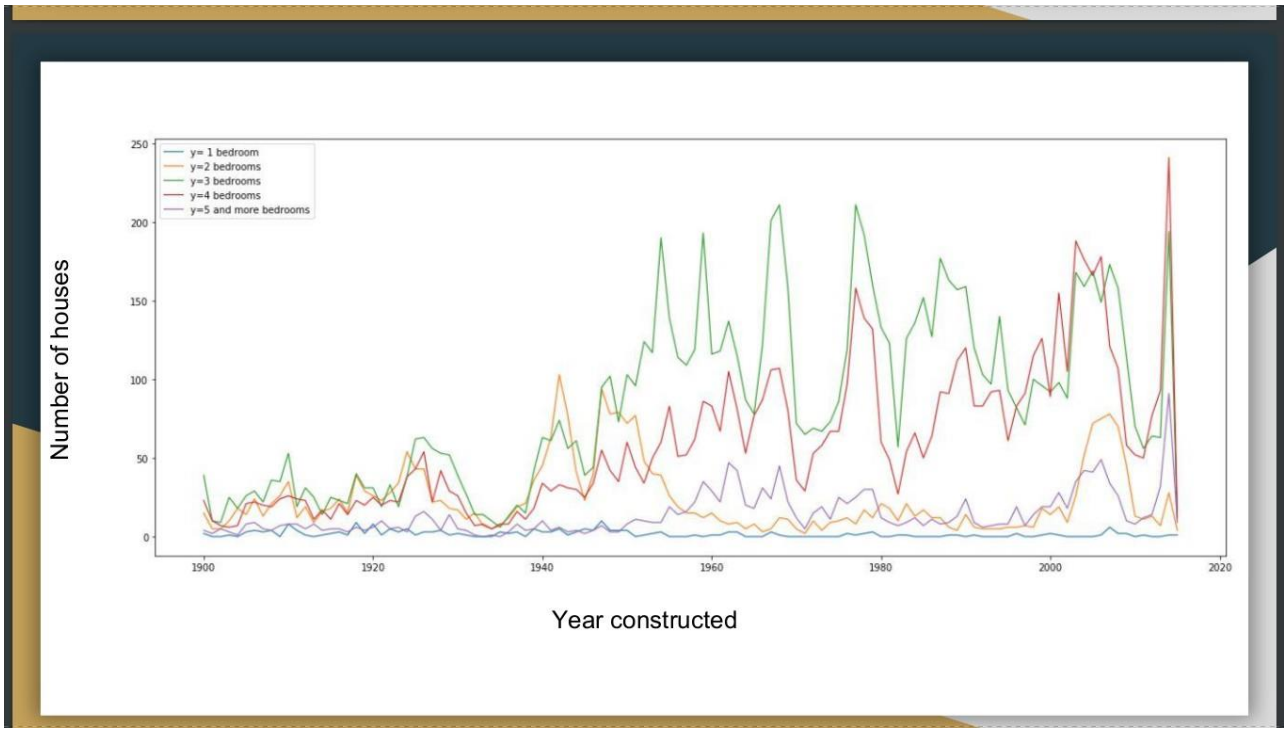
* demand for house is increades by time

3) Graph between number of floors and number of houses.



*Almost 96% of people opt for houses within 2 floors.
* houses having single floors ie only ground floor are more in demand

4)number of houses and year constrected

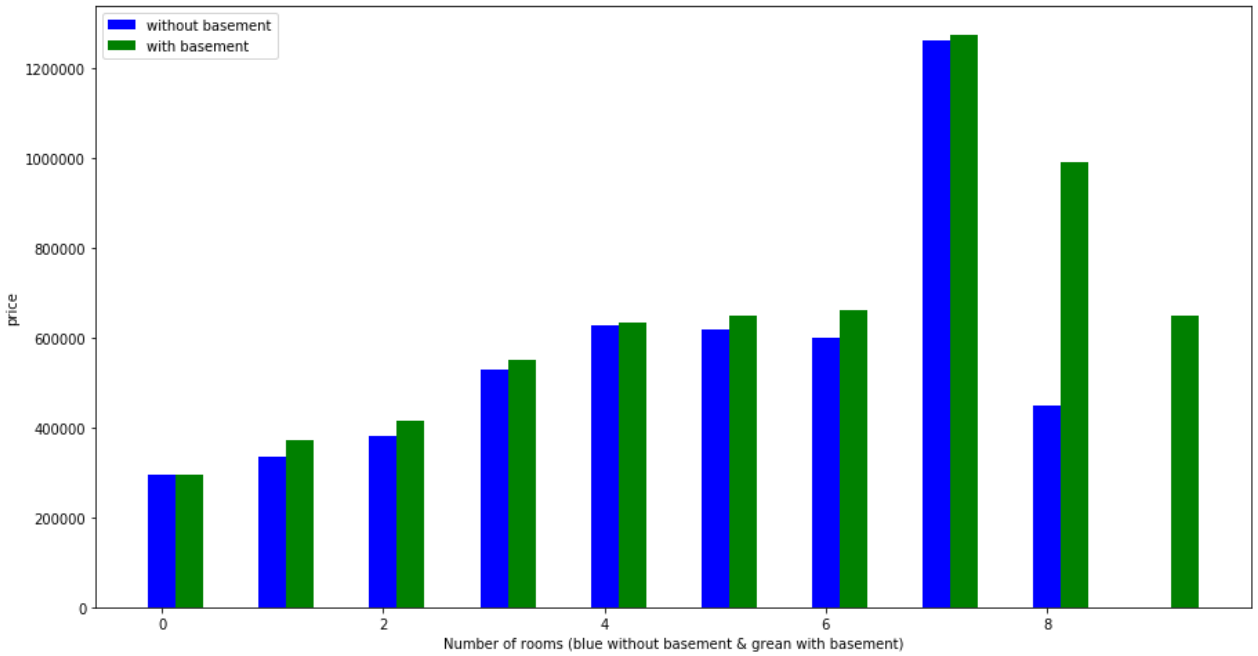


It's clear houses with 3 and 4 bedrooms are the most sought houses.
There is also a good demand for 2 and 5 bedroom houses.
1 bedroom or 6 and more bedroom houses bought are less.

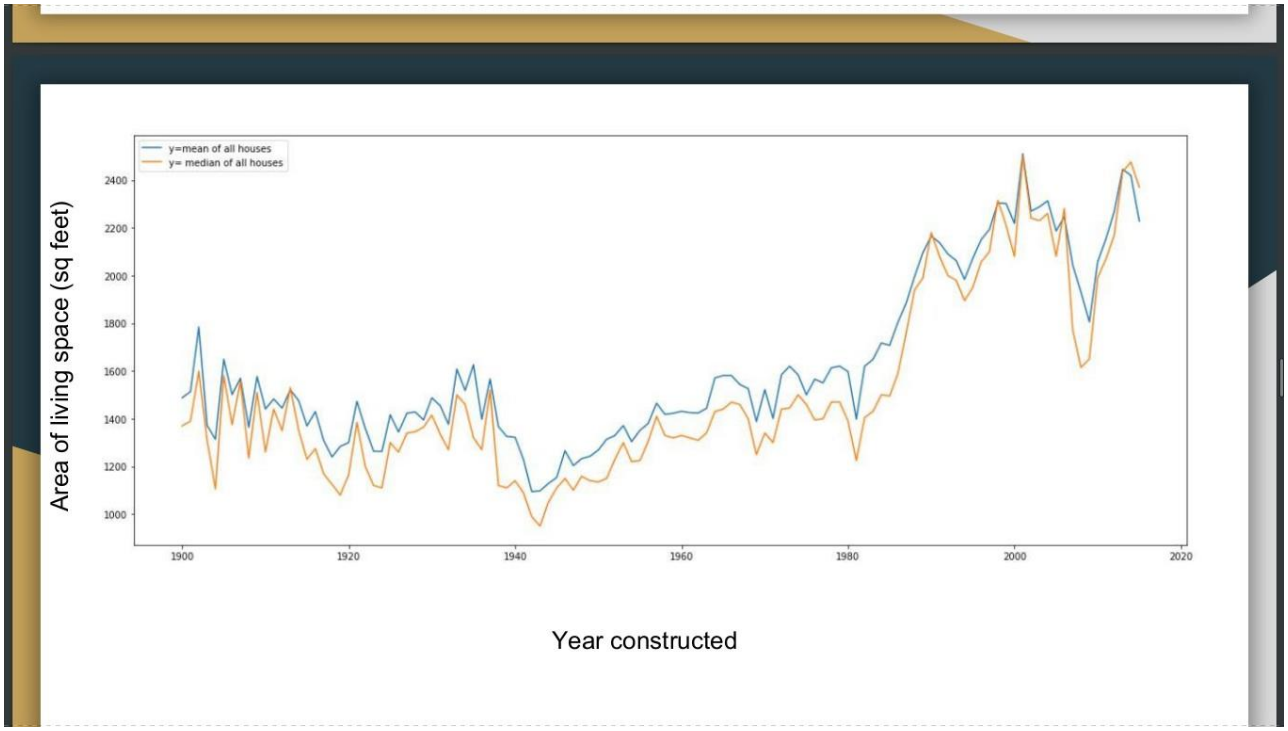
5) comparision between a house having basement and house not havinig basement with price

It can be seen that except for 8 bedroom houses, there is not much difference in the pricing for houses with and without basement.

By this we can see that larger bedroom houses have no basement

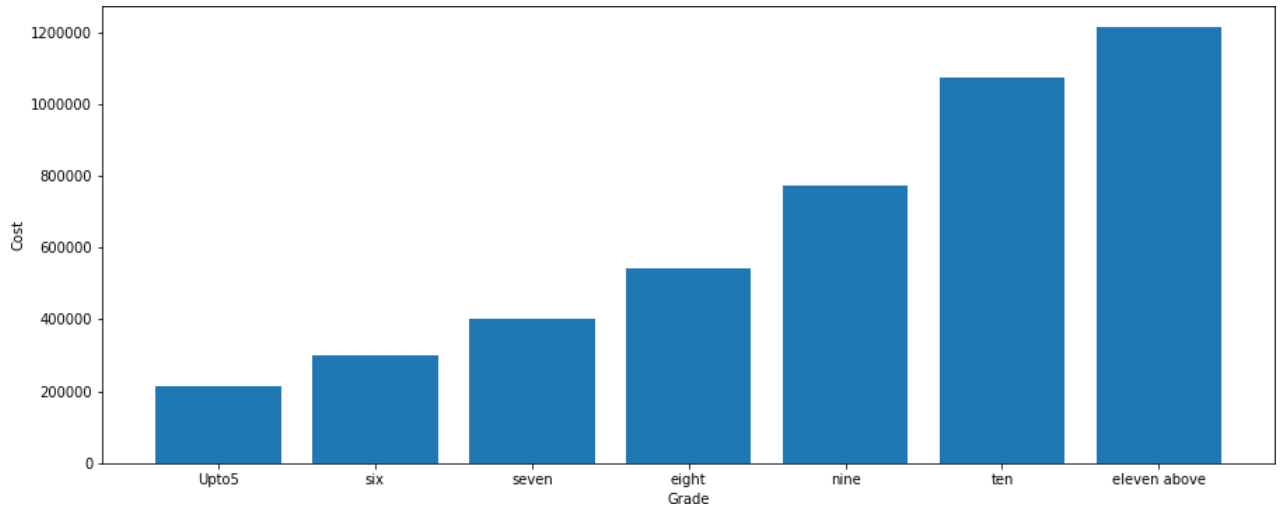


6)graph between year constructed and area living:



The lot size decreases significantly after the year 2000.
Also the is a little increase in the living space.
Considering both of these factors the reason for increase in the percentage of living space to that of lot size is justified.

7) comparision between grade and price of house:

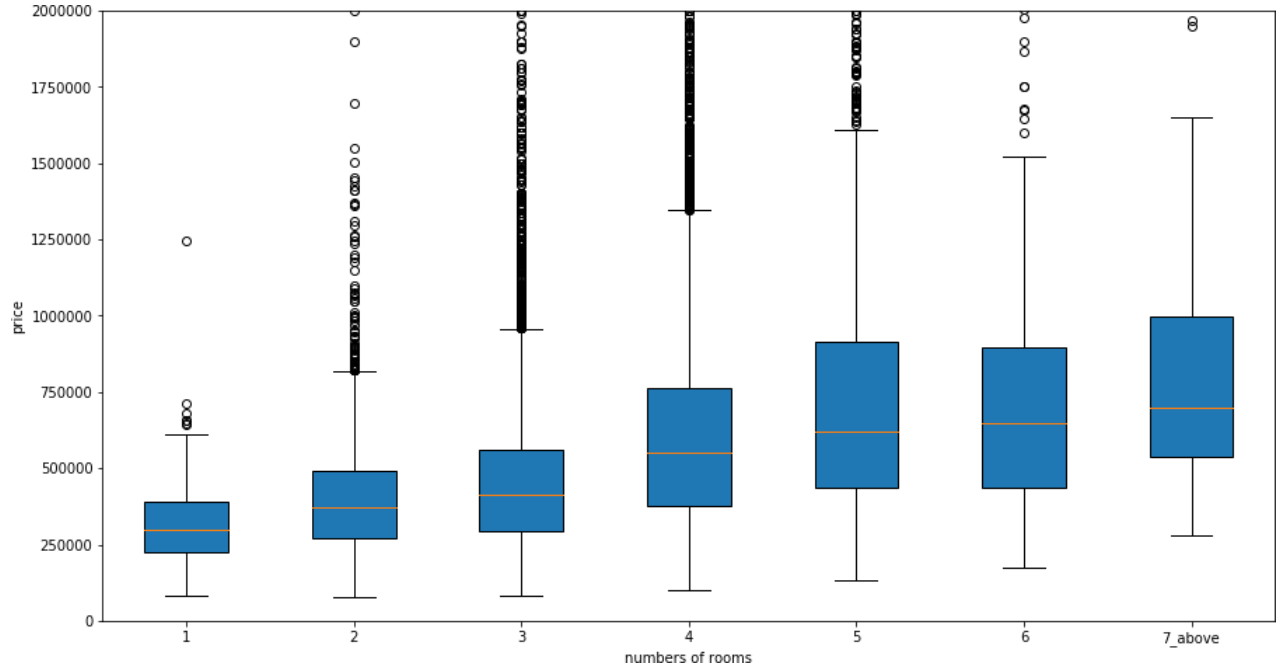


Grade is assigned by the King County authorities which represents the construction quality. Grades run from grade 1 to 13.

Its evident from the graph that higher the grade higher the price

By this we can see that heigher the grade price is more

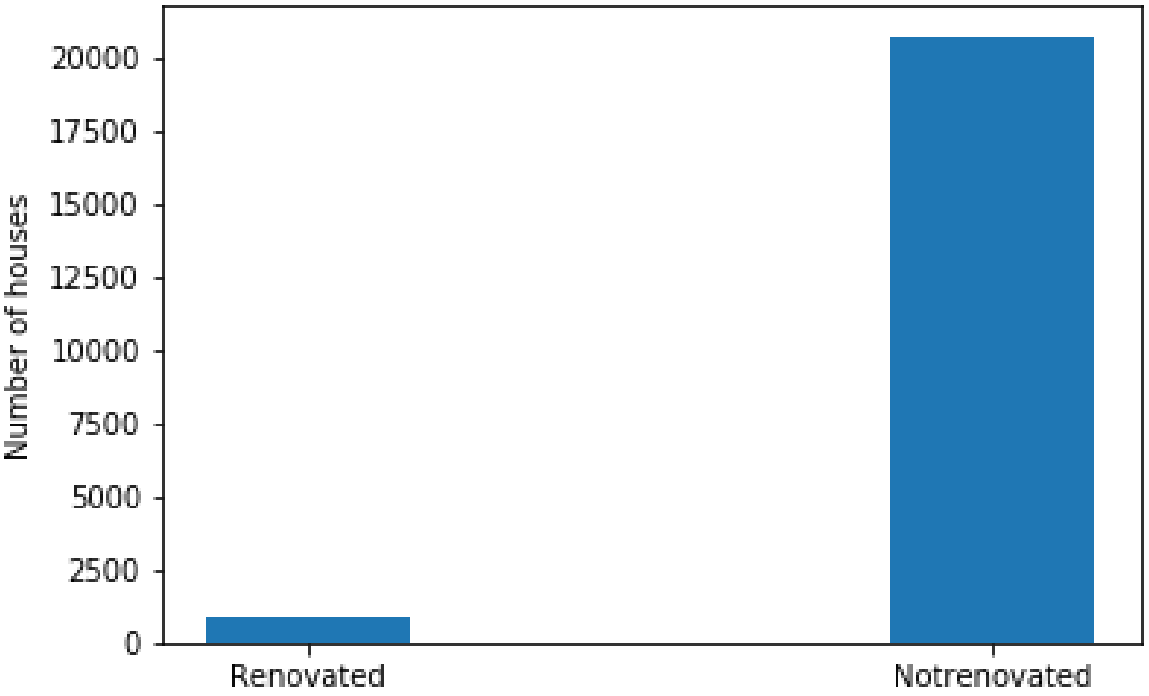
8) <u>Graph between number of rooms and price:</u>



The median values of the houses increase with the increase in the number of rooms.
However there is no clear distinction of prices for different bedroom houses as it can due to the reason of different luxuries.
Eg: A 2 bedroom house might have a modular kitchen, luxury lights, etc and a 3 bedroom

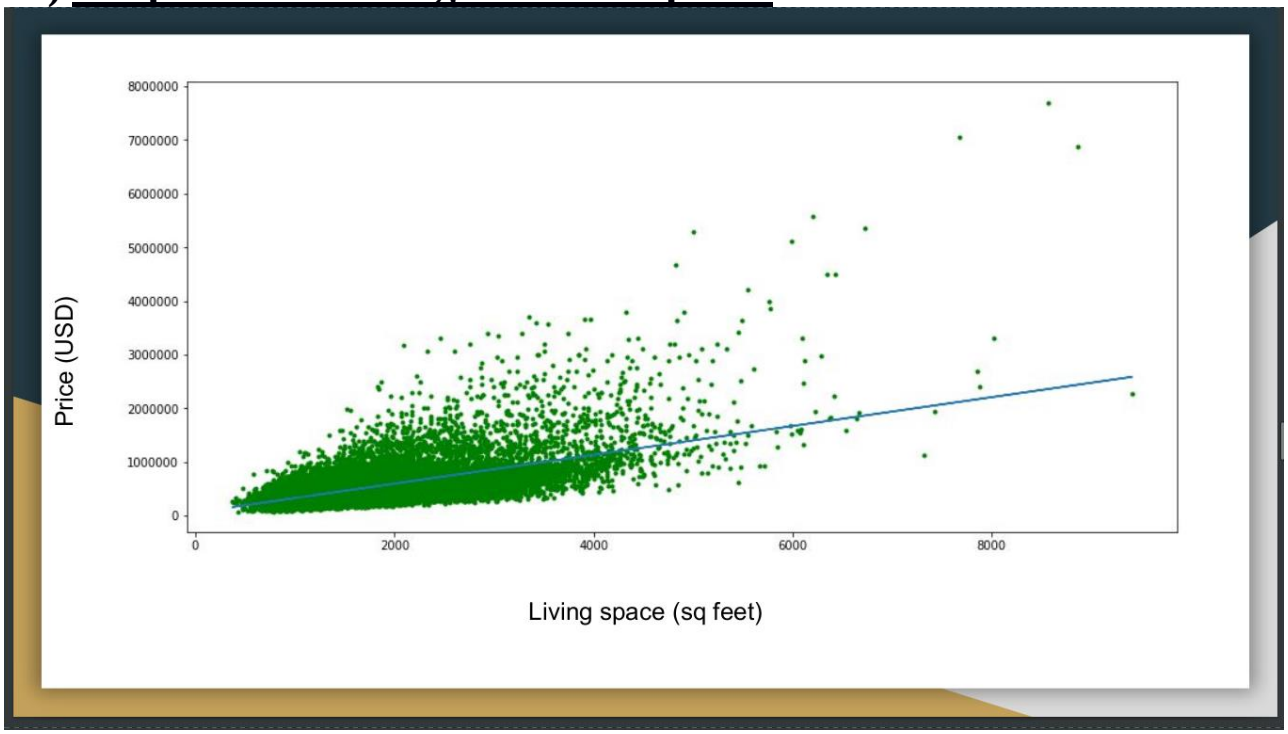
house might not be having these features and may cost lower

9)**Notrenovated and renovated:**



* most of the housed are not renovated

10) comparision of living areas with price:

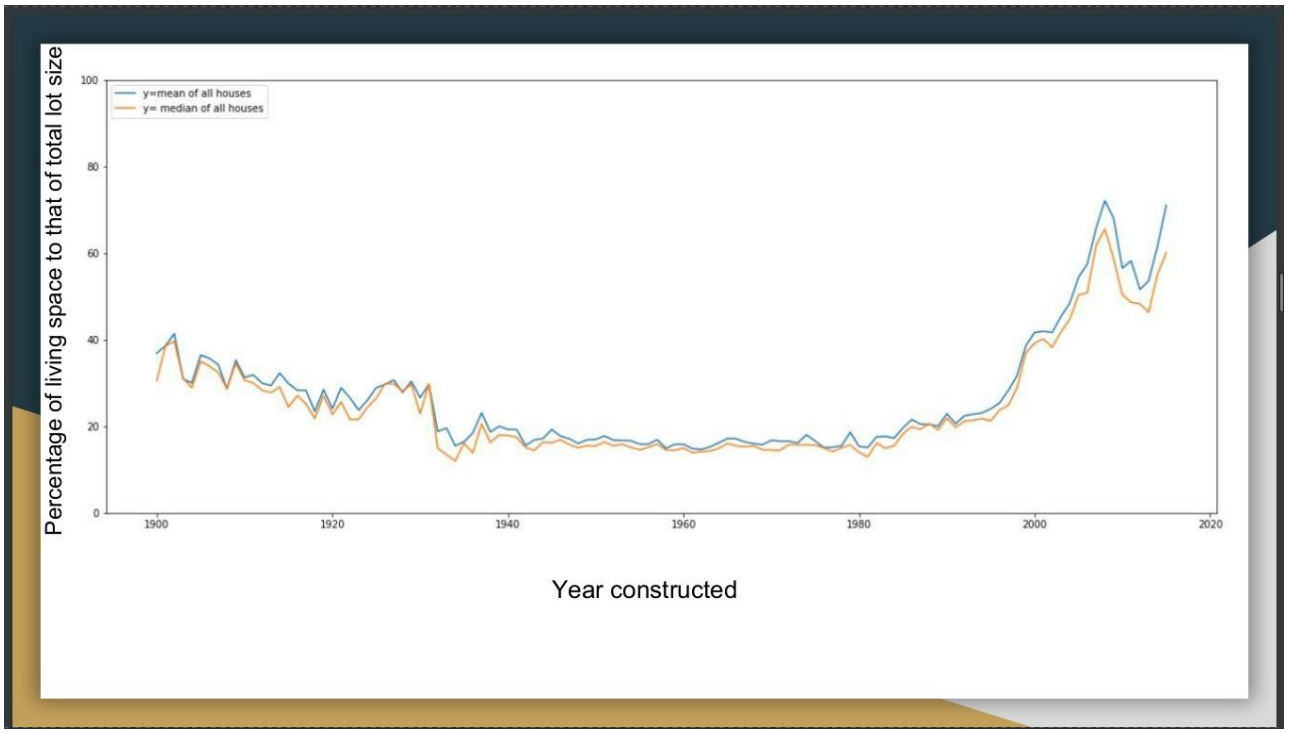


There is a small positive correlation between living space and price.

It's obvious that there is no proper positive correlation, due to fact that

price also depends on other factors such as materials used, furnished or unfurnished, etc.

11) graph between year constrected and percentage of living space to that of total space:



The lot size decreases significantly after the year 2000.
Also the is a little increase in the living space.
Considering both of these factors the reason for increase in the percentage of living space to that of lot size is justified.

Conclusion:

As the grading is increasing the price of house is increasing.
Demand of house is going on increasing over time.
Percentage of living spave over total is going decreasing till 1990 and sudden increase is seen .
Rooms with 3 bed rooms are in more demand os they are more in number.
House with 3,4 are almose increasing in a same rate over time.
House with single floor are in demand
As the grade of house is heigher the price is heigh