PES UNIVERSITY, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UE18CS203

B.Tech, Sem III
Session : Aug-Dec, 2019

UE18CS203 – INTRODUCTION TO DATA SCIENCE

REPORT
ON
EXPLORATORY ANALYSIS ON
House Sales In King County, USA

SECTION : B

#	SRN	Name	Contact No.	Email ID	Sign
1.	PES1201801308	K A ADEAB	9980324448	adeabayyup@gmail.com	
2.	PES1201801298	Nischal Jain	7869966223	nischaljain143@gmail.com	

ABOUT THE DATA SET

The given dataset consists the details(year sold, square feet, etc) of house sales in King County (a county in Washington), USA. The data set condenses a lot of information using decimal values. Ex: 5.5 bathrooms, means the house consists of 5 bathrooms with toilet and shower and 1 bathroom with only toilet.

ABSTRACT

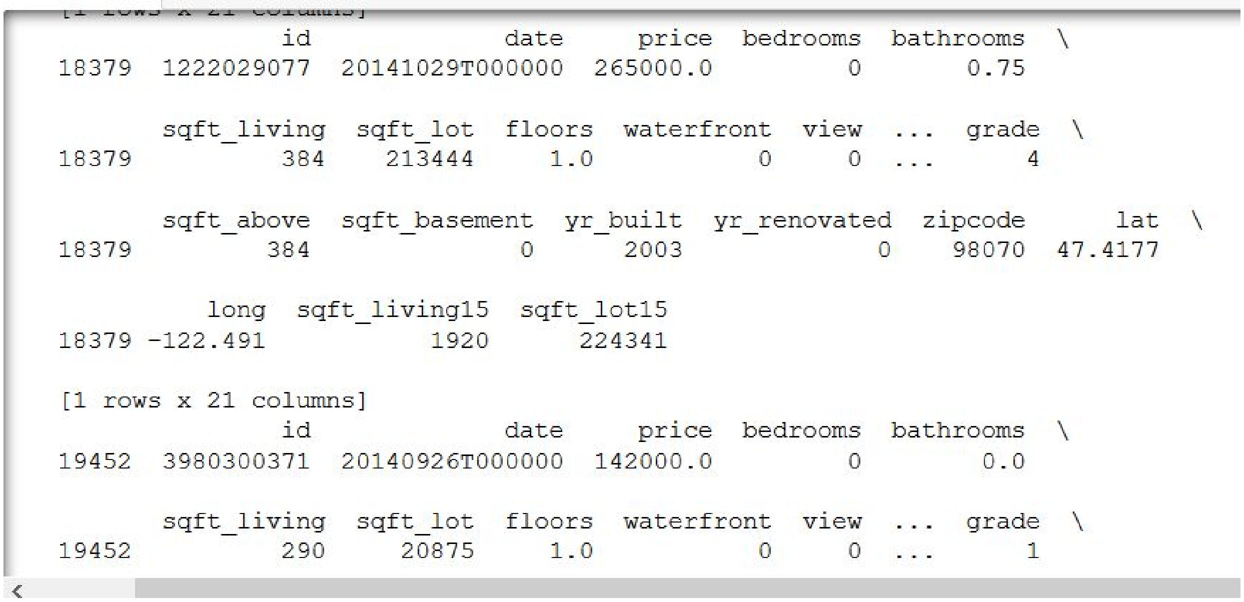
We performed hypothesis tests and used various visualizations to answer specific questions like how does renovation affect the price of a house or how location or a view affect the price.

Explanation of row headers, which we involved in our visualizations/tests
Bathrooms -no of bathrooms(a 0.5 bathroom means it has only toilet, a 0.25 bathroom means it has only sink, a 0.75 bathroom means it has both toilet and an extra sink)
Grade - An index from 1-13, where 1-3 falls short on building construction and design seven has an average level of construction and design 11-13 have high quality of construction and design.

EXPLORATORY ANALYSIS

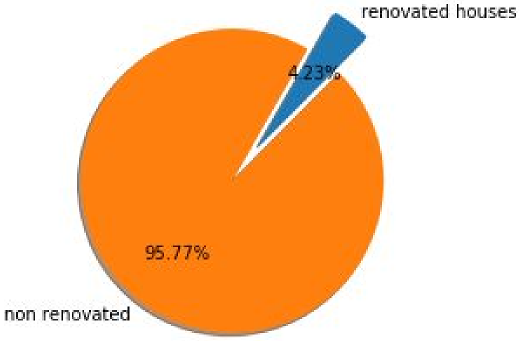
Data cleanup

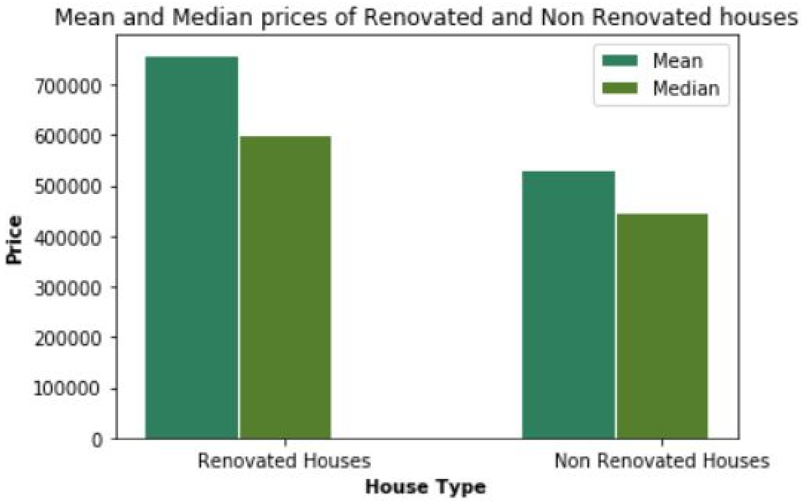
Some rows in the given dataset did not make sense. Ex: A house costing $1.2 million dollars had no bathrooms or bedrooms, there were houses with only 1 bathroom but no bedroom. Some of these houses had grades less than 4, meaning they were under construction. Houses had average grades yet they had no bathrooms/bedrooms. All the details of houses like those were dropped.
We further cleaned the data by rounding the number of bathrooms. We considered a 0.5 or 0.75 bathroom as 1 bathroom to help us better visualize the boxplots.



```
[1 rows x 21 columns]
               id             date     price  bedrooms  bathrooms  \
18379  1222029077  20141029T000000  265000.0         0       0.75

       sqft_living  sqft_lot  floors  waterfront  view  ...  grade  \
18379          384    213444     1.0           0     0  ...      4

       sqft_above  sqft_basement  yr_built  yr_renovated  zipcode      lat  \
18379         384              0      2003             0    98070  47.4177

          long  sqft_living15  sqft_lot15
18379 -122.491           1920      224341

[1 rows x 21 columns]
               id             date     price  bedrooms  bathrooms  \
19452  3980300371  20140926T000000  142000.0         0        0.0

       sqft_living  sqft_lot  floors  waterfront  view  ...  grade  \
19452          290     20875     1.0           0     0  ...      1
```

Comparison of mean and median prices of renovated houses and houses which are not renovated





Both the mean and median prices of renovated houses are significantly higher compared to non renovated houses. From the pie chart we infer that only a small proportion of houses are renovated. We treat the houses in King County as a sample which is coming from a larger distribution of houses(let's say price of every house in washington). We state our null hypothesis as "There is no difference between the median prices of renovated and non renovated houses and alternate hypothesis as "There is a difference in the median prices of renovated and non renovated houses". We used Kruskal Wallis test to test our hypothesis.

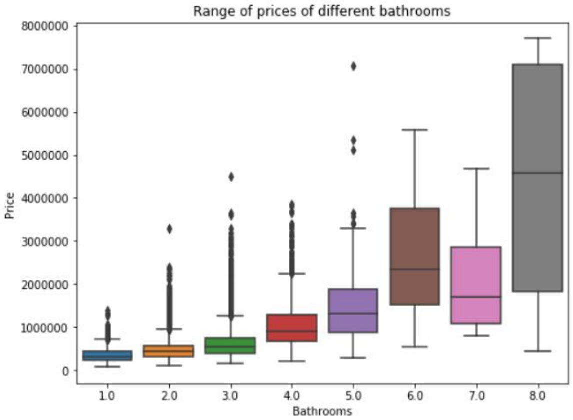
Reasons to use Kruskal Wallis test:

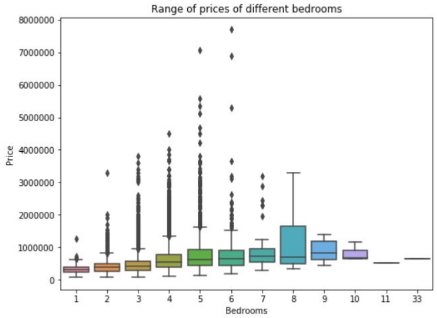
1. It's a distribution free test
2. Data is categorical and both categories come from the same distribution
3. The median better represents the data as many outliers are present

Df=1 and alpha=5%, the critical chi-square value was found to be 3.81
The test statistic value was found to be 220.6 which is more than the critical value. Hence, ***We reject the null hypothesis.***

Do renovated houses have better grades?
We state the null hypothesis as "There is no difference between the mean grades of renovated and non renovated ones" and alternate hypothesis as "There is a difference between the mean grades of renovated and non renovated ones". We used the Mann Whitney U test(data was not normal and independent) to test the hypothesis. The p value(0.028) of test statistic was found to be less than the significance level(5%) hence ***we reject the null hypothesis.***

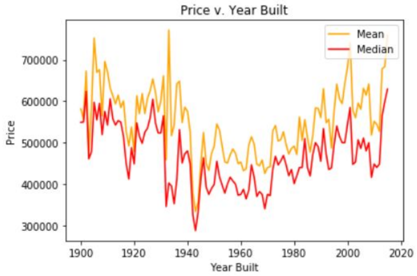
Range of prices of different prices of bathrooms and bedrooms





Number of bathrooms/bedrooms between 2 and 5 had the most number of outliers.

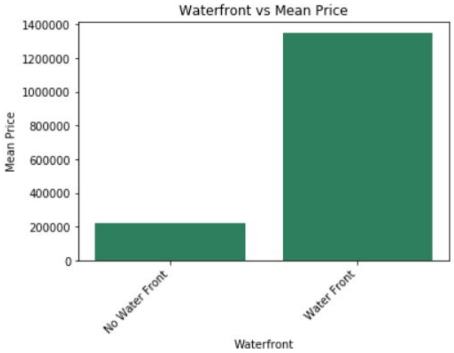
The trends in mean/median prices over the years



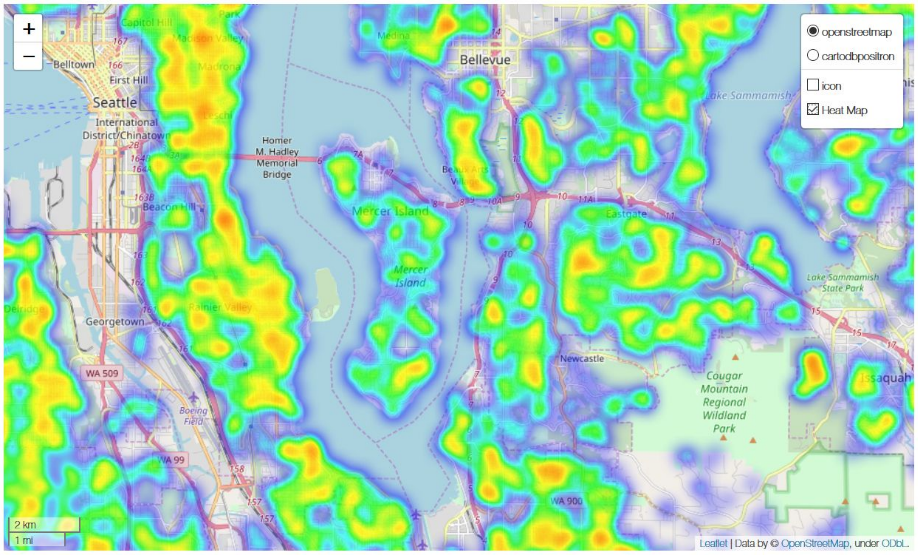
The median graph follows a similar pattern with mean graph. We can observe that houses that were built from 1980's onward, show an overall positive growth in price.

Does the price of a house of a house depend on its location?

Factors like waterfront or view depend on the location of the house. We first compare the average prices of houses with and without a waterfront. The average prices of houses with waterfront were significantly higher.



We next plot a heatmap of price and location



Houses near the water bodies are either yellow or red, implying that they are more expensive. Interior part of the land appears to be green, implying that they are relatively less expensive.

CONCLUSION

We were able to find out that renovated houses had higher median price and their grades were not the same as non renovated houses.House prices showed an overall growth over time. We were able to track how price changes with location with the help of heatmap.