**Battle of King and Snohomish**

**(Comparing housing prices of King and Snohomish:**

**The two neighborhoods in the US and predicting future prices)**

Indra Adhikary

September , 2019

1. **Introduction**
   1. **Background**

The two neighborhoods King and Snohomish in the US are pretty similar to each other but on doing a deeper analysis we can figure out patterns and differences the prices of products there. Snohomish covers a comparatively less land area, yet it is considered corporate area and people are generally busier there. Considering the population growth, there is a rise of number of houses with a rise in their prices too. Thus, a prediction of new houses in the two neighborhoods can be really useful and of a great help to citizens there and the new people planning to stay there.

* 1. **Problem**

The project aims to figure out a better neighborhood with the help of housing data in terms of prices of the houses based upon the features that the houses have. It also aims to predict the price any new house based upon which neighborhood it is being built in and other of its specifications.

* 1. **Interest**

Obvious enough that the citizens of the two neighborhoods and also the new citizens would be the ones having the prime interest in this data analysis. Also the companies over here might also be interested to settle its employees here.

1. **Data acquisition and cleaning**
   1. **Data sources**

The housing data of this region of US can be found in :

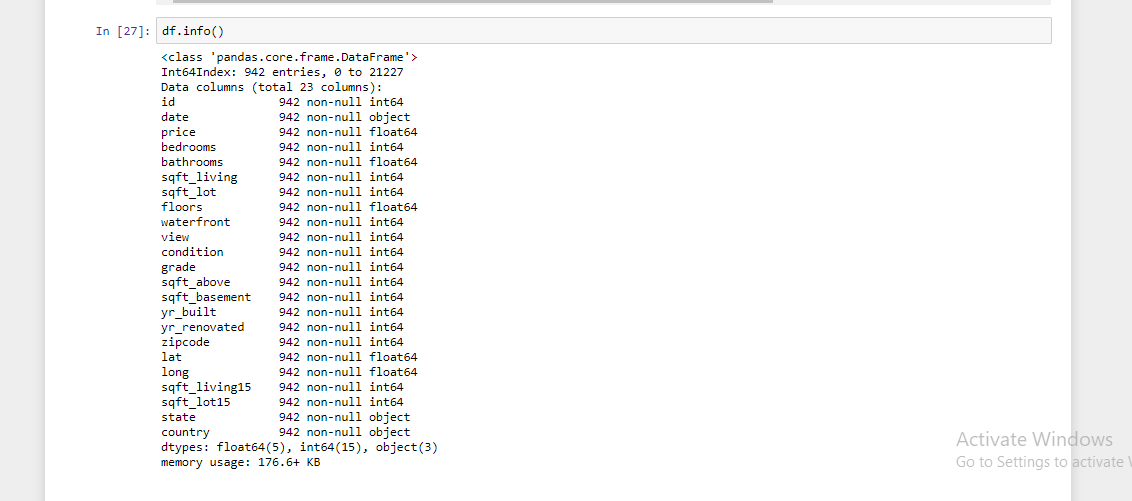
[www.kaggle.in](http://www.kaggle.in)

However, this data lacked the data of State, city and zip code.

For which I used a dataset available at:

<https://simplemaps.com/data/us-zips>

Thus, creating my required dataset was quite a task.



* 1. **Data cleaning**

The dataset created till now after joining the two datasets that was downloaded and merged has a lot of data that are not really required in our analysis and prediction.

Thus, here we can get rid of a few columns first and keep only the following columns:

1. Id
2. Price
3. Bedrooms
4. Bathrooms
5. Sqft\_lot
6. Floors
7. Waterfront
8. Condition
9. Yr\_built
10. Yr\_renovated
11. Zipcode
12. Lat
13. Long
14. State
15. Country

Also, we find that the bathroom data are of datatype “str” thus, we convert it to “np.int”. Similar is the case with floors , so we take care of it, too.

* 1. **Feature conversions**

There were various features with str values which needed to be converted to readable int values.

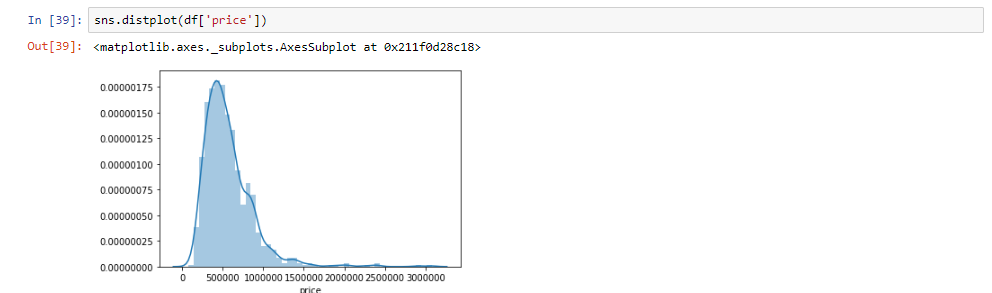
Waterfront column was converted to values 0 and 1 with 0 for not waterfront and 1 for a waterfront house.

Low and high in the new column range which we create later are converted to 0 and 1 respectively.

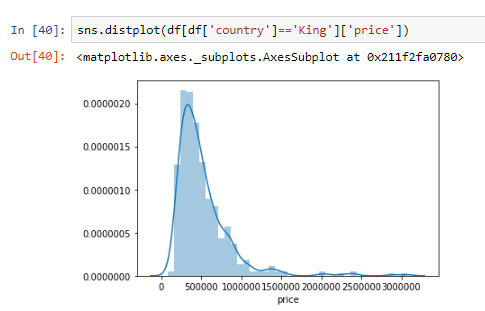
Later, for data analysis or model training, King and Snohomish of the state column was converted to 0 and 1 respectively.

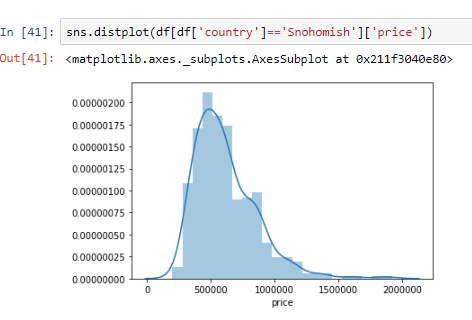
1. **Exploratory Data Analysis**
   1. **Price Analysis**

We can plot a distribution graph of the price column.

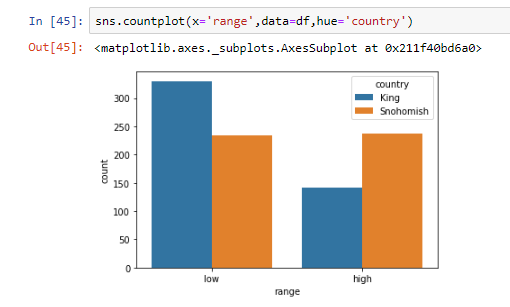


We can also plot the prices separately for the two cities.





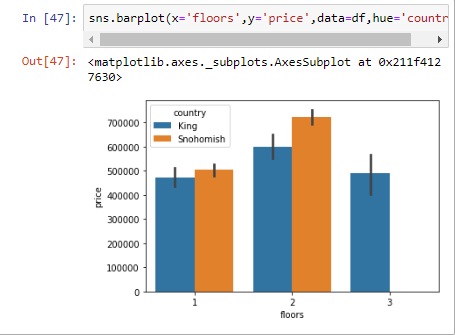
We can classify and create a new column to signify whether the mentioned price is high or low based upon greater or lesser than the mean price which is calculated to be **568247.1560509554**



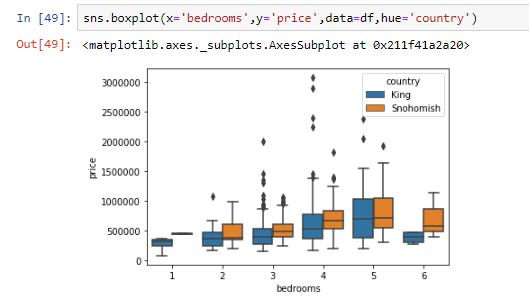
**Relationships with Floors, Bathrooms and Bedrooms**

Another conclusion from our analysis can be that there are no 3 story houses yet constructed in Snohomish.

For obvious reasons with increase in the number of floors in the houses, prices increase. Yet this increase is a bit higher in Snohomish than in King.

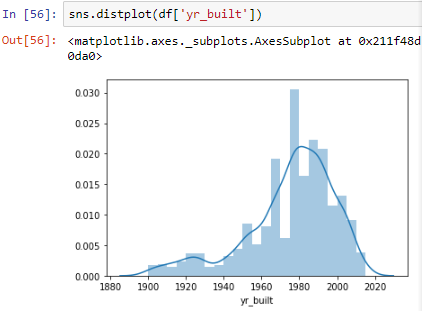


Similar is the case with bedrooms.



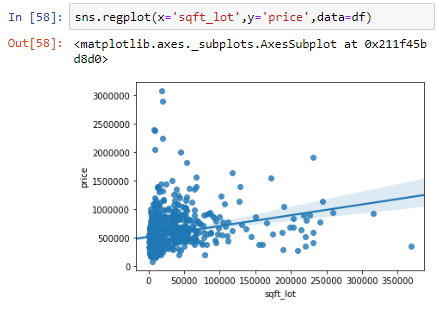
**Relationship with Yr\_built**

There is a great diversity in the different years when the houses were built. But, mostly The data is clustered to years between the years 1980 - 2000.



And it follows a regression line along with the price.

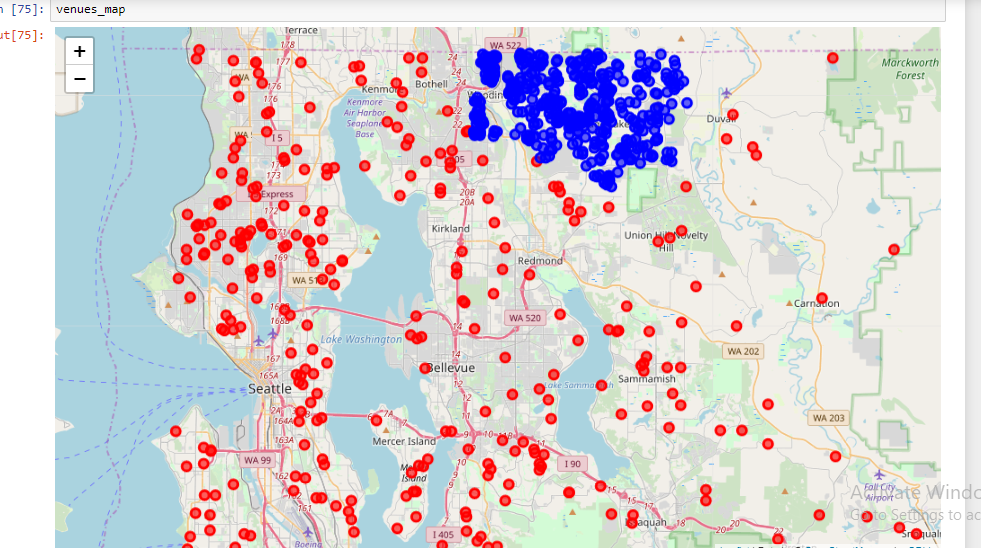
We find it is positively correlated.



**Relation with location**

We just display this for a view of the scatter of the various houses across the cities.

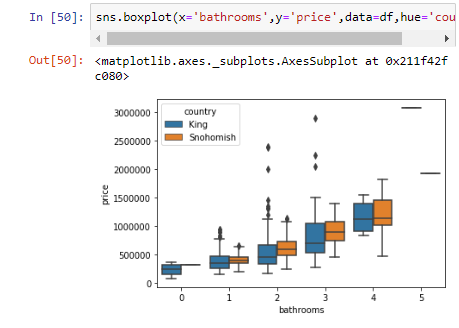
Houses in King are marked with red and the ones in Snohomish are marked with blue.



**Other Relations**

With increase in number of bathrooms, it is very obvious of the increase in the prices of the houses.

Also, a waterfront view contributes to increases in prices.



1. **Conclusions from the data analysis**

### *According to my analysis, prices of the houses in Snohomish is comparatively more than the prices of the houses in King. The prices increase with increase in sqft rea as well as no. of bedrooms, but we saw this difference to be greater in case of houses in Snohomish than in King. Most of the houses are built in year around 1980. Also, if one is looking for a house of 3 stories, It is only available in King and not in Snohomish*

1. **Predictive model**

# We will now be Building a Machine Learning Model for predicting future prices of any new houses constructed in this region based on the values.

For our predictions, we use logistic regression model

We use train\_test\_split and use 70% of the test data to train our LogisticRegression model and the rest 30% to test it.

We match our predictions and match with the original data to figure out our accuracy

