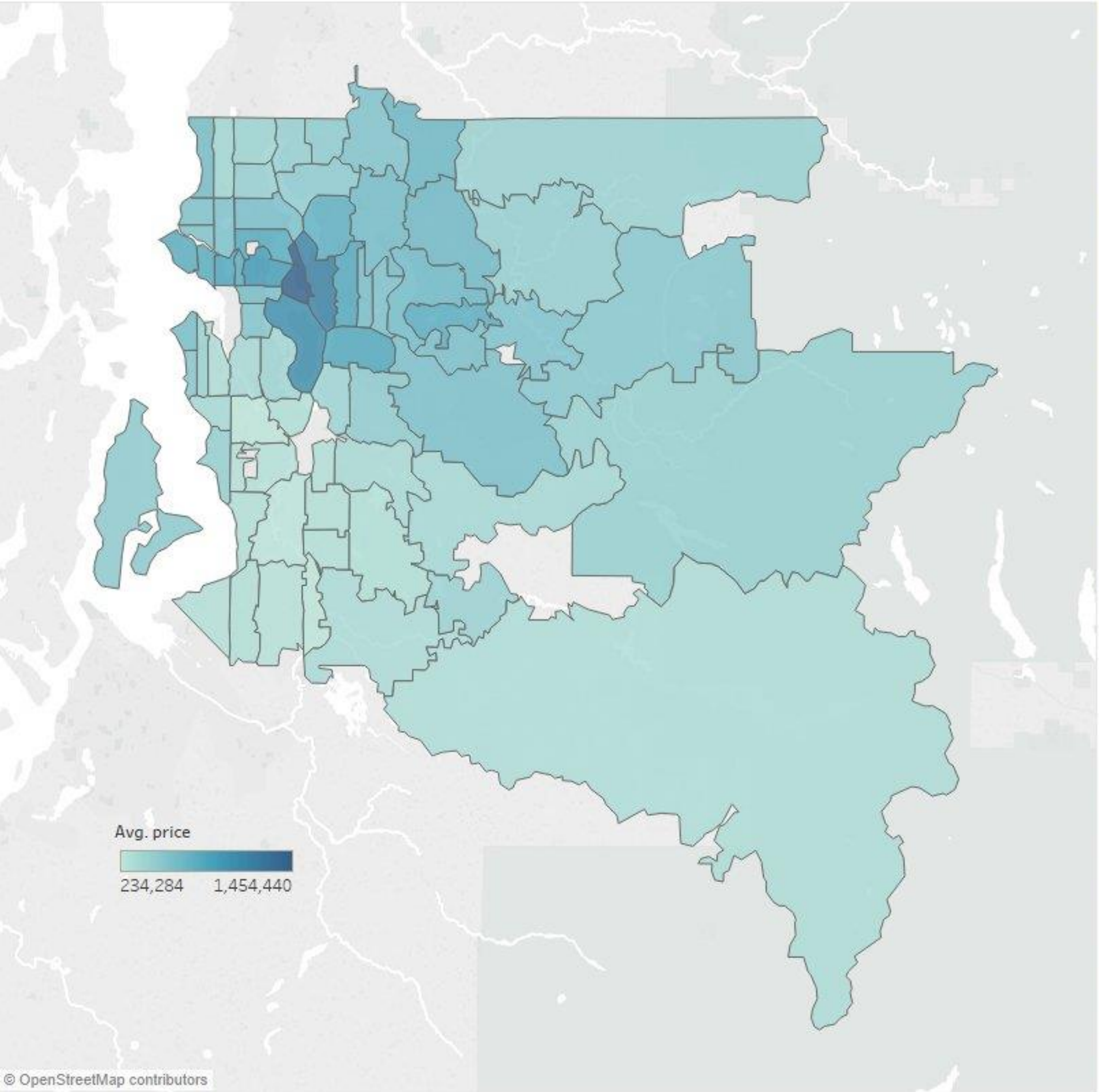
Analysis of house sale prices, for King Country during May 2014 and May 2015



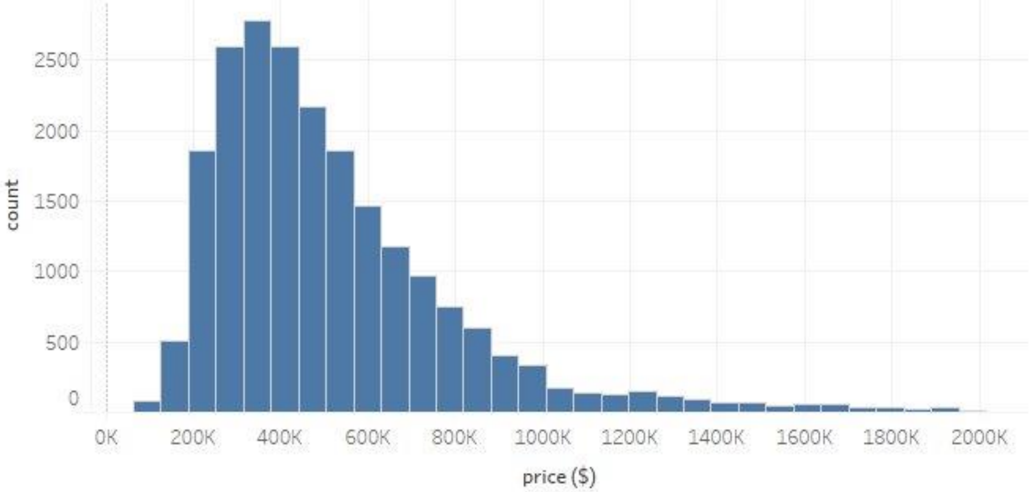
Real Estate is one of the biggest market in the world providing terrific opportunities for investors. Well know investors Donald Trump and Robert Kyosake created their capital investing in Real Estate. To be succesful in this business, it is imoprtant to understand market, price formation, price evolution and to analize the available property data.

This dataset was obtained from Kaggle. The identified main feature is the price and the other features are living area, lot area, basement area, grade, zipcode and condition. The trends and relationships of the price with other variable in the dataset will be identified.

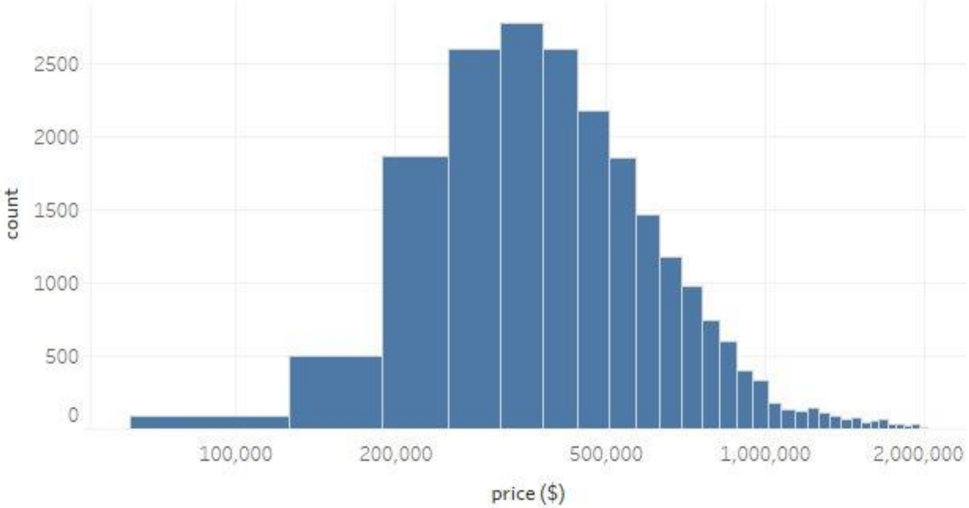
Prepared by Boris Kushnarev

Price is the main characteristic of the preperty which can immediatly attract or repel a buyer. Property price forms based on a number of features such as the lot area, the living area, its condition and location, for instance. Moreover, price can be influenced by different life situations such as a divorce or urgently moving to another area. What is the price distribution in this dataset and how it correlates with other variables?

Property price distribution



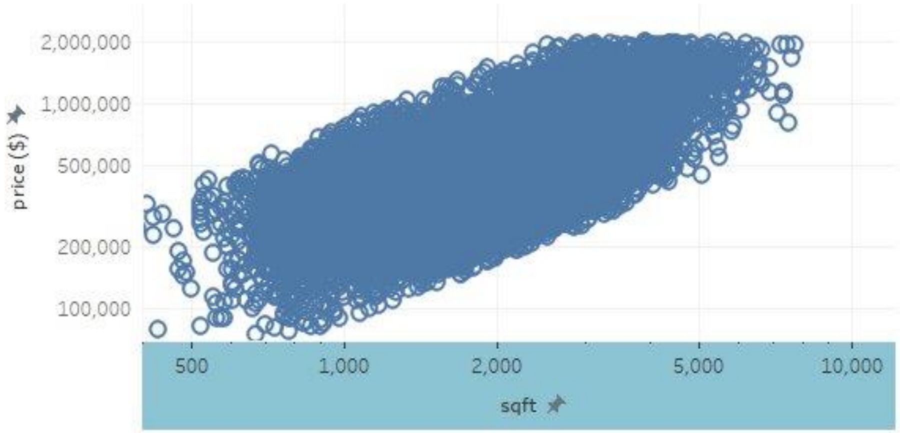
Log-normal price distribution of the property



The created histograms and the scatter plot are interactive and choosing all zipcodes or only some of them we can investigate the entire dataset or just part of it.

The property price distribution is right skewed. An attempt to transform the skewed histogram to the log-normal was succesful with identified peak at about $400000. Correlation between price and one of the most important feature living area is postive moderate, with correlation coefficient 0.68. The logarithmic transformation of both axes shows a linear relationship between these parameters.

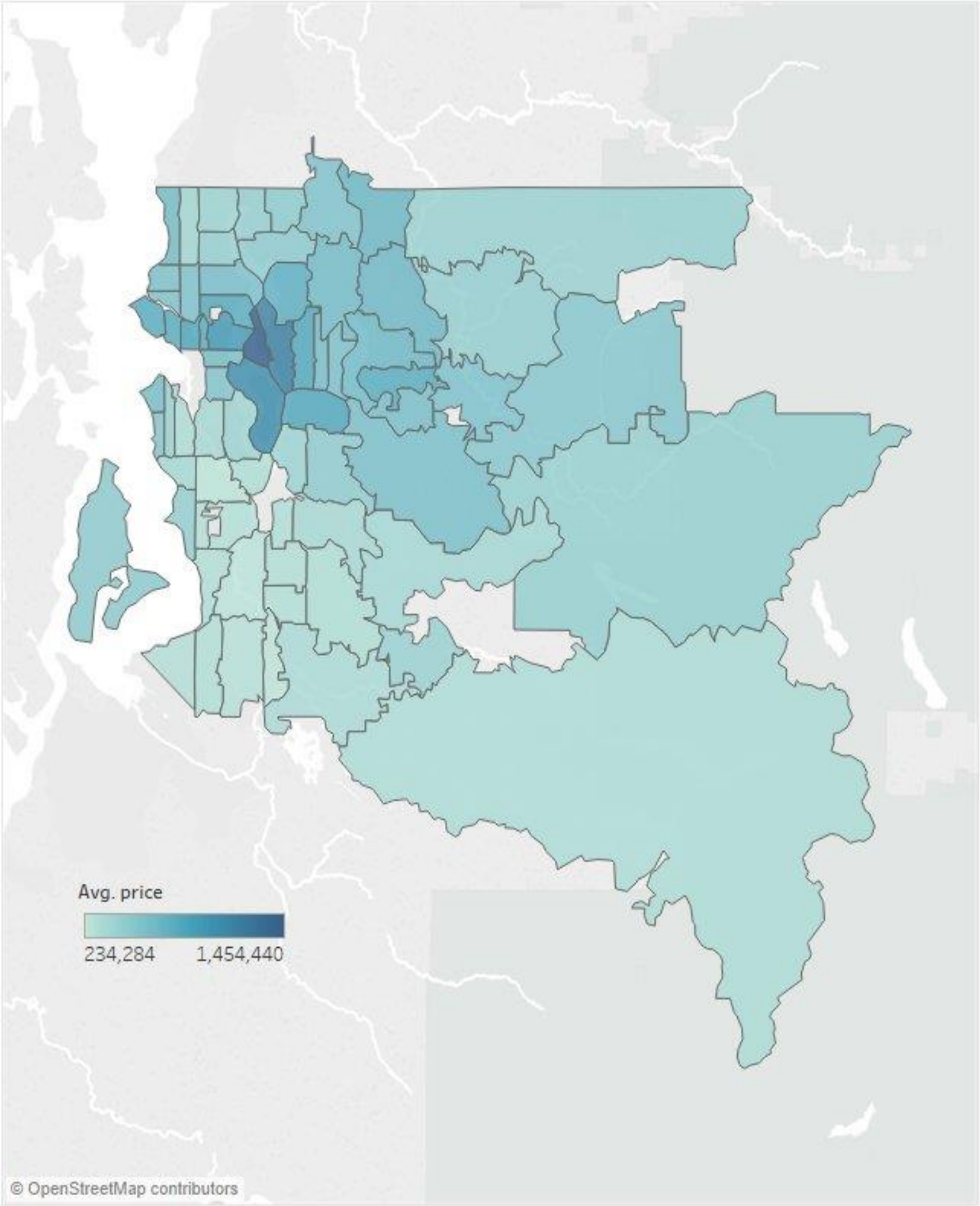
Logarithmic transformation of both price and living area, their correlation.



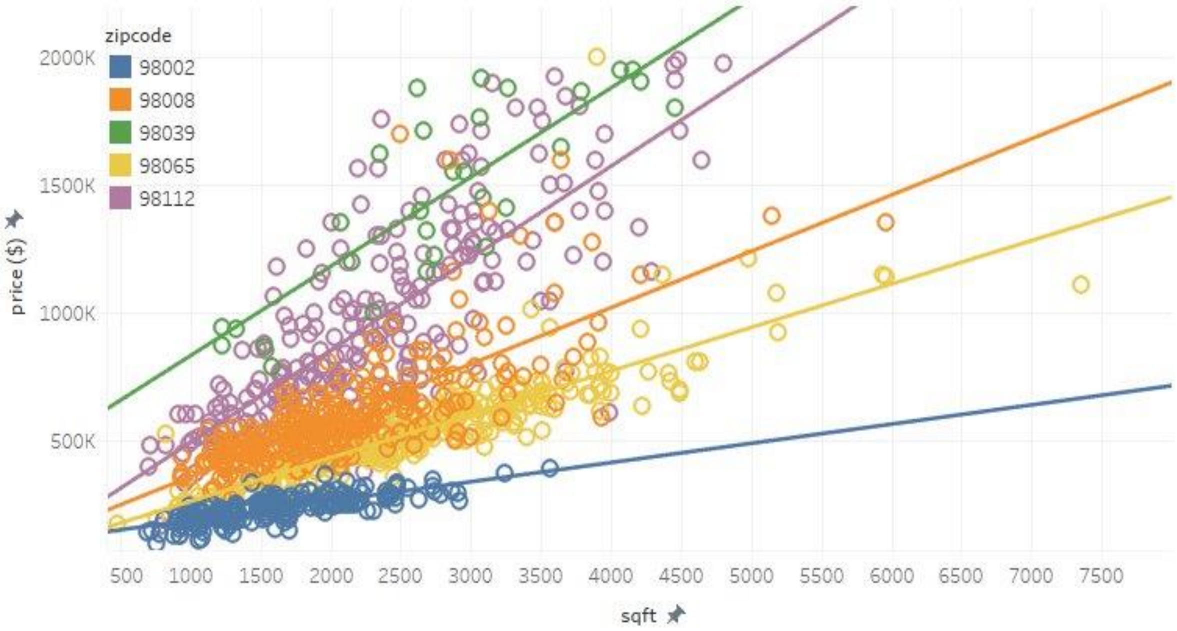
zipcode

- ☑ 98001
- ☑ 98002
- ☑ 98003
- ☑ 98004
- ☑ 98005
- ☑ 98006
- ☑ 98007
- ☑ 98008
- ☑ 98010
- ☑ 98011
- ☑ 98014
- ☑ 98019
- ☑ 98022
- ☑ 98023
- ☑ 98024
- ☑ 98027
- ☑ 98028
- ☑ 98029
- ☑ 98030
- ☑ 98031
- ☑ 98032
- ☑ 98033
- ☑ 98034
- ☑ 98038
- ☑ 98039
- ☑ 98040
- ☑ 98042
- ☑ 98045
- ☑ 98052
- ☑ 98053
- ☑ 98055
- ☑ 98056
- ☑ 98058
- ☑ 98059
- ☑ 98065
- ☑ 98070
- ☑ 98072
- ☑ 98074
- ☑ 98075
- ☑ 98077
- ☑ 98092
- ☑ 98102
- ☑ 98103
- ☑ 98105

Analysis of house sale prices, for King Country during May 2014 and May 2015



Scatter plot of price with respect to living area and zipcodes



On the map we can see how the average prices depends on location, i.e. on a zipcode. We can see that the location has a significat affect on the price.

On the scatter plot we can see a similar situation where we see that prices are clusterred for certain zipcodes.

zipcode

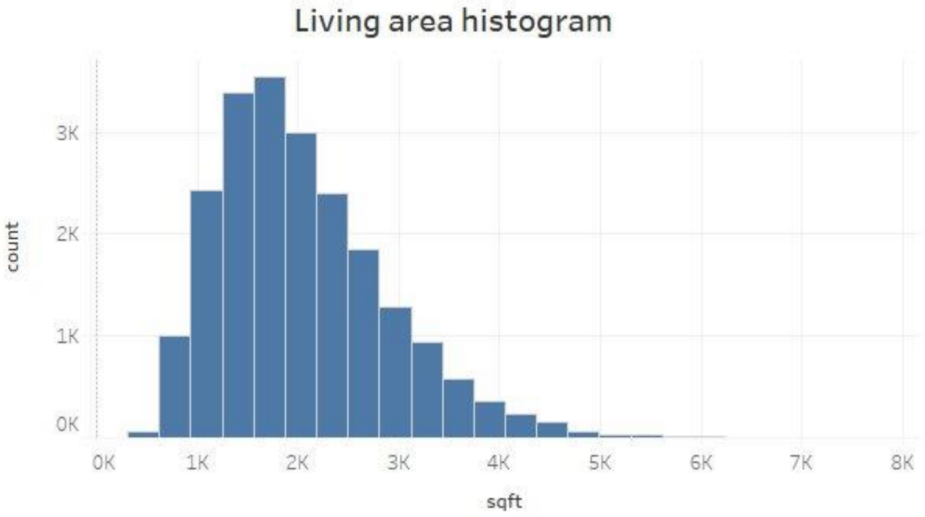
- ☑ (All)
- ☑ 98001
- ☑ 98002
- ☑ 98003
- ☑ 98004
- ☑ 98005
- ☑ 98006
- ☑ 98007
- ☑ 98008
- ☑ 98010
- ☑ 98011
- ☑ 98014
- ☑ 98019
- ☑ 98022
- ☑ 98023
- ☑ 98024
- ☑ 98027
- ☑ 98028
- ☑ 98029

grade

- ☑ (All)
- ☑ 3
- ☑ 4
- ☑ 5
- ☑ 6
- ☑ 7
- ☑ 8
- ☑ 9
- ☑ 10
- ☑ 11
- ☑ 12
- ☑ 13

condition

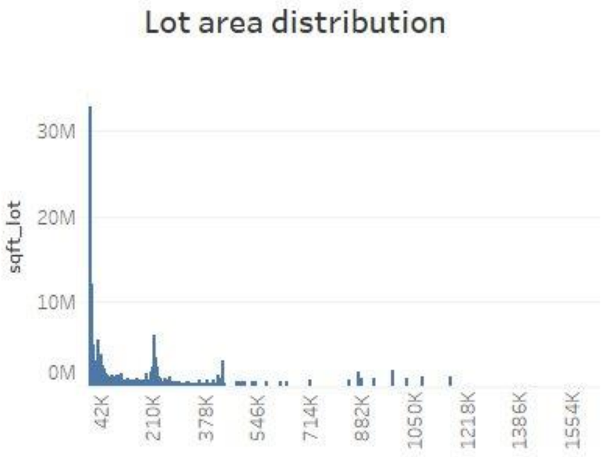
- ☑ (All)
- ☑ 1
- ☑ 2
- ☑ 3
- ☑ 4
- ☑ 5

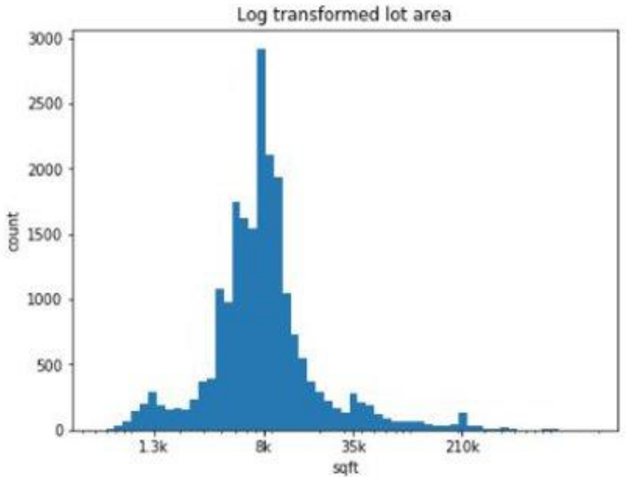

Living area histogram

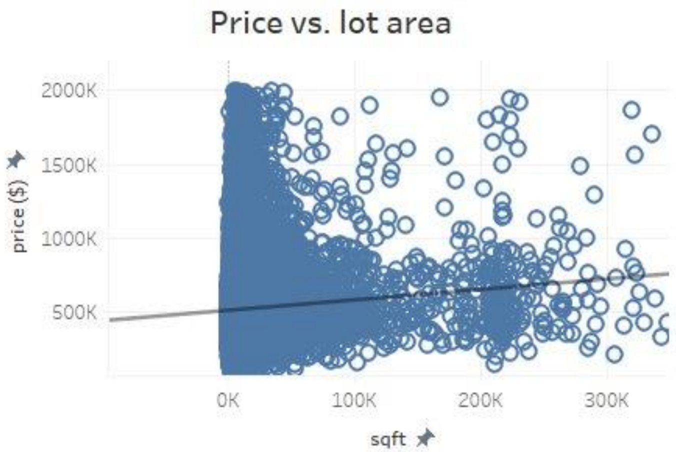
Taking a look at other feature distributions is also important. We can see that the most common living area is about 1800 sqft and the histogram is right skewed with a long tail.

The values of the lot area are mostly concentrated below 40000 sqft with a long tail. Furthermore, some lot areas can reach even huge numbers, more than 210000 sqft. The logarithmic scale was used again where a multimodal distribution is observed. The main peak is about 8000 sqft with abour 3000 occurances. Other peaks at 1300 sqft, 35000 sqft and 210000 sqft have significantly lower occurance.

In additon, price vs lot area scatter plot shows low positive correlation, with correlation coefficient 0.1. This is probably is not a good predictor to consider.

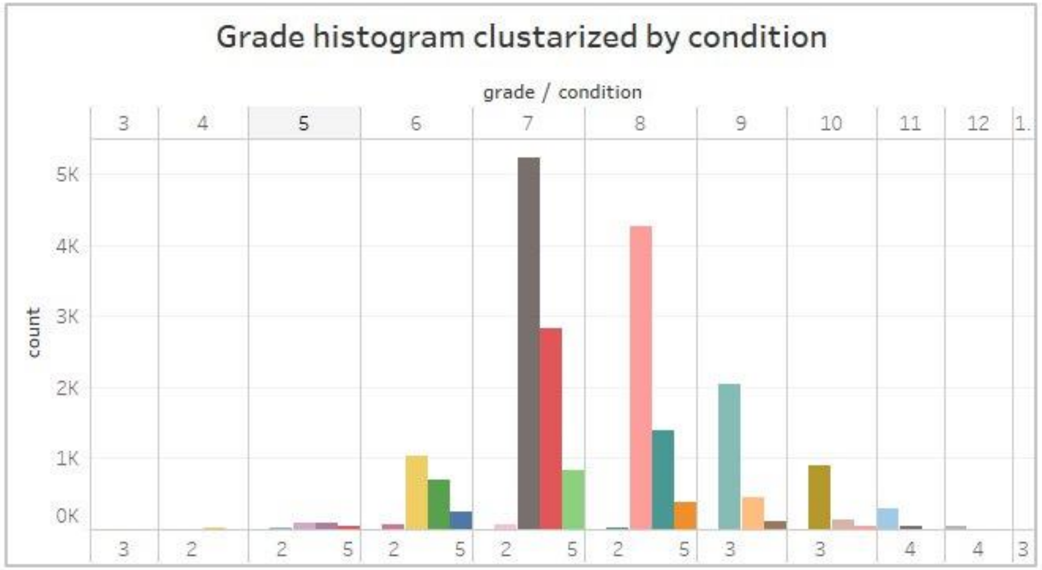

Lot area distribution


Log transformed lot area

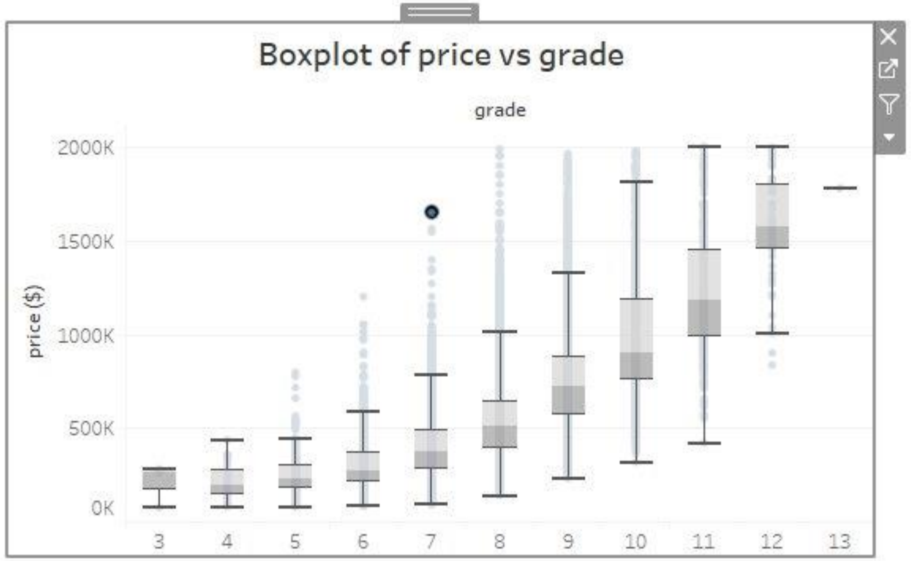

Price vs. lot area

zipcode

- ☑ (All)
- ☑ 98001
- ☑ 98002
- ☑ 98003
- ☑ 98004
- ☑ 98005
- ☑ 98006
- ☑ 98007
- ☑ 98008
- ☑ 98010
- ☑ 98011
- ☑ 98014
- ☑ 98019
- ☑ 98022
- ☑ 98023
- ☑ 98024
- ☑ 98027
- ☑ 98028
- ☑ 98029
- ☑ 98030
- ☑ 98031
- ☑ 98032
- ☑ 98033
- ☑ 98034
- ☑ 98038
- ☑ 98039
- ☑ 98040
- ☑ 98042
- ☑ 98045
- ☑ 98052
- ☑ 98053
- ☑ 98055
- ☑ 98056
- ☑ 98058
- ☑ 98059
- ☑ 98065
- ☑ 98070
- ☑ 98072
- ☑ 98074
- ☑ 98075
- ☑ 98077
- ☑ 98092
- ☑ 98102
- ☑ 98103
- ☑ 98105
- ☑ 98106



Grade histogram clustarized by condition



Boxplot of price vs grade

zipcode

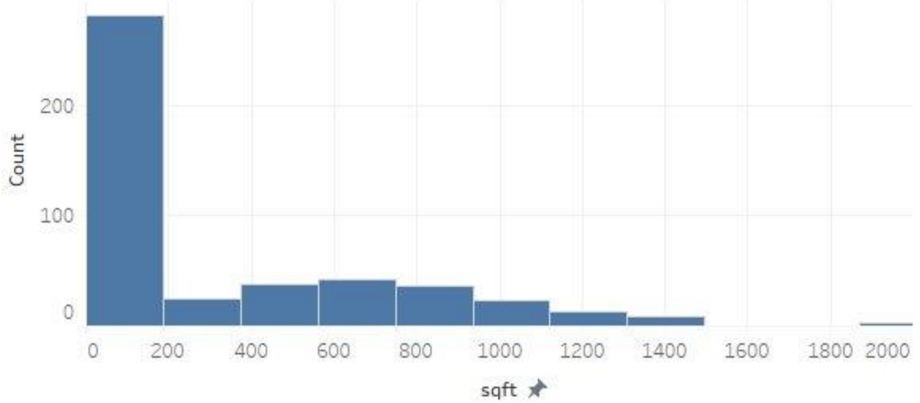
☑ (All)
☑ 98001
☑ 98002
☑ 98003
☑ 98004
☑ 98005
☑ 98006
☑ 98007
☑ 98008
☑ 98010
☑ 98011
☑ 98014
☑ 98019
☑ 98022
☑ 98023
☑ 98024
☑ 98027
☑ 98028
☑ 98029
☑ 98030
☑ 98031
☑ 98032
☑ 98033
☑ 98034
☑ 98038
☑ 98039
☑ 98040
☑ 98042
☑ 98045
☑ 98052
☑ 98053
☑ 98055
☑ 98056
☑ 98058
☑ 98059
☑ 98065
☑ 98070
☑ 98072
☑ 98074
☑ 98075
☑ 98077
☑ 98092
☑ 98102
☑ 98103
☑ 98105
☑ 98106

Two categorical variables grade and condition plays an important role in the price formation. From the grade histogram which was clustarized by condition we can infer that grade 7 is the most common and then grade 8.

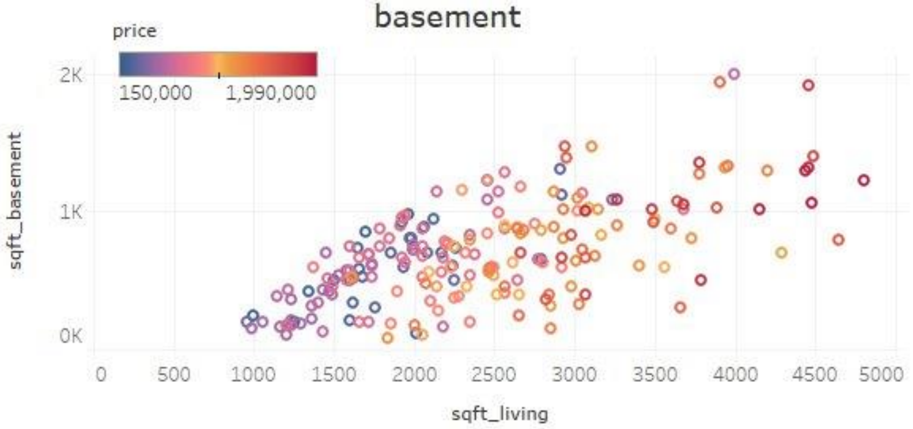
The most common condition is for grade 6, 7, 8, 9, 10 and 11.

The boxplot of the price vs grade shows a parabolic dependance. The price gradually increasing from values aroun $250000 upto velues above $1500..

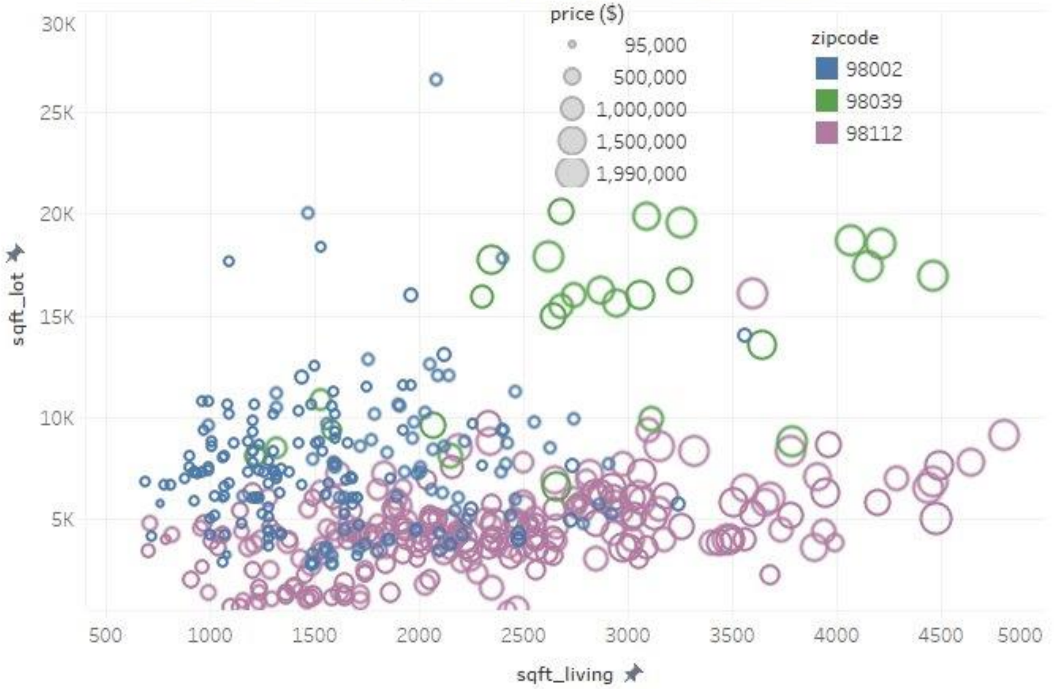
Basement area histogram



Scatter plot of price with respect to living area and basement



Price with respect to living area and lot area



From the basement area histogram we can see that mostly buildings do not have basements. However, when property has a basement the area can vary from small abour 200 sqft to sizes upto 1600 sqft.

To investigate farther this question, the scatter plot of price with respect to basement area and living area was obtained. To reduce the overlapping only a few number of zipcodes were chosen.

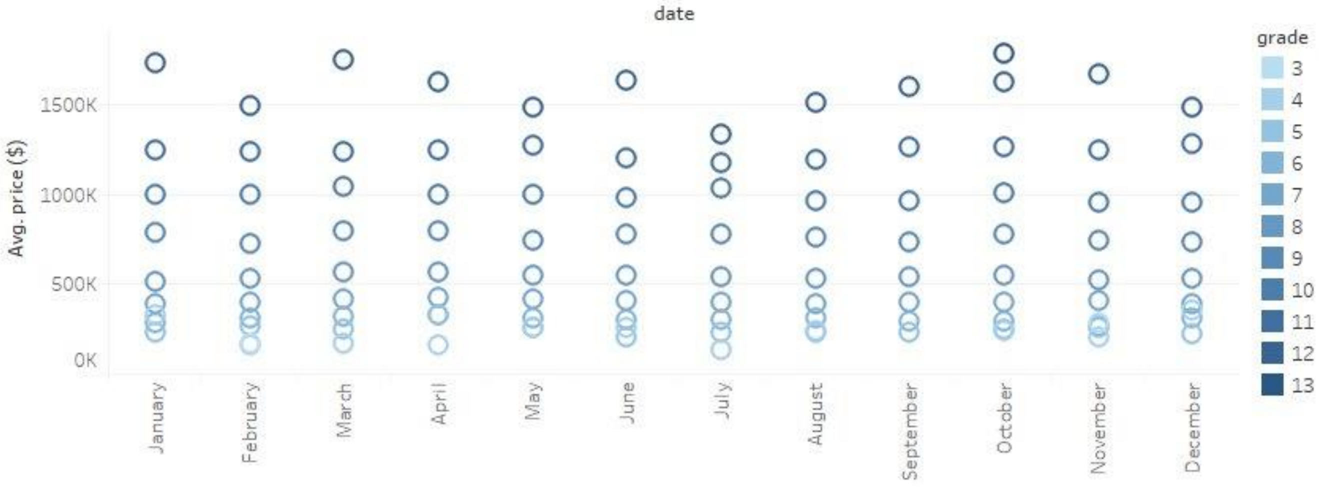
The scatter plot of the price vs living and lot areas for a few randomly chosen zipcodes shows that the higher living area the higher price. However, we cannot say the same for the lot area. We can see the price increase only for the zipcode 98039 with respect to lot area but at the same time the living area was also increased.

zipcode

□ (All)
□ 98001
☑ 98002
□ 98003
□ 98004
□ 98005
□ 98006
□ 98007
□ 98008
□ 98010
□ 98011
□ 98014
□ 98019
□ 98022
□ 98023
□ 98024
□ 98027
□ 98028
□ 98029
□ 98030
□ 98031
□ 98032
□ 98033
□ 98034
□ 98038
☑ 98039
□ 98040
□ 98042
□ 98045
□ 98052
□ 98053
□ 98055
□ 98056
□ 98058
□ 98059
□ 98065
□ 98070
□ 98072
□ 98074
□ 98075
□ 98077
□ 98092
□ 98102
□ 98103
□ 98105
□ 98106

The evolution of the price with respect to time is another important aspect of analysis. How does the price evolved during the year when properties were sold? How does the price depends on the year when the property was built? Does the year of renovation effects the price? Can we observe any dependance on grade and condition?

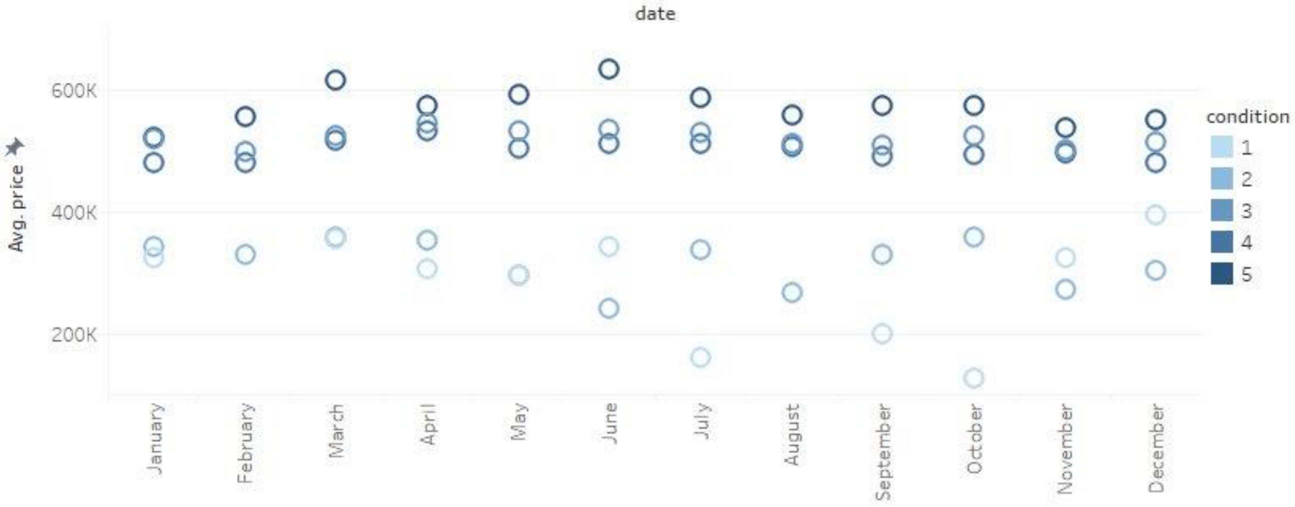
Average price vs month sold and grade

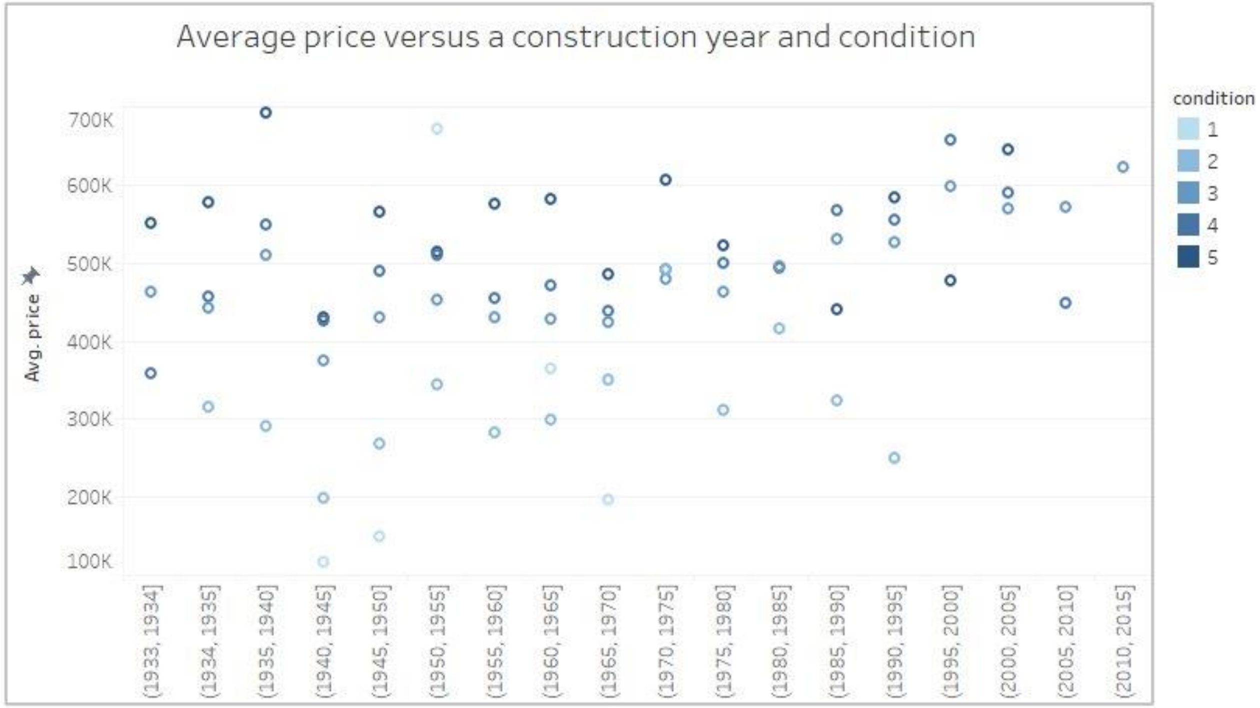


Considering these two plots we can say that price does not show any trends with respect to time. However, we can see a strong dependence on the grade such that the higher the grade the higher price.

The same applies to the condition. The higher condition the price. In addition, we can see a jump between condition 2 and 3.

Average price vs month sold and condition

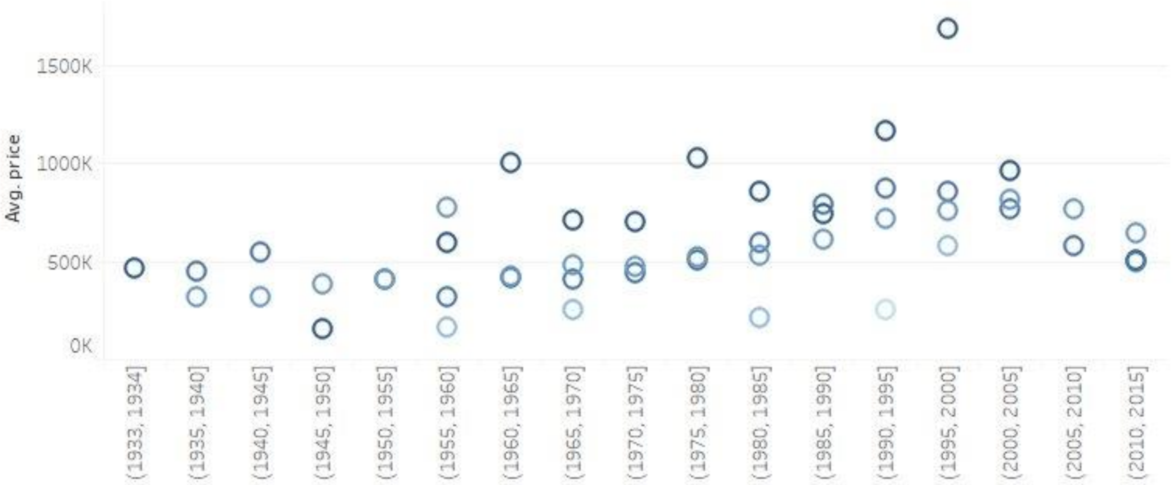




Average price versus a construction year and condition

The first plot shows an uptrend. We can see that the younger the property the higher the price. Morevere, as previously, the higher condiion the higher price.

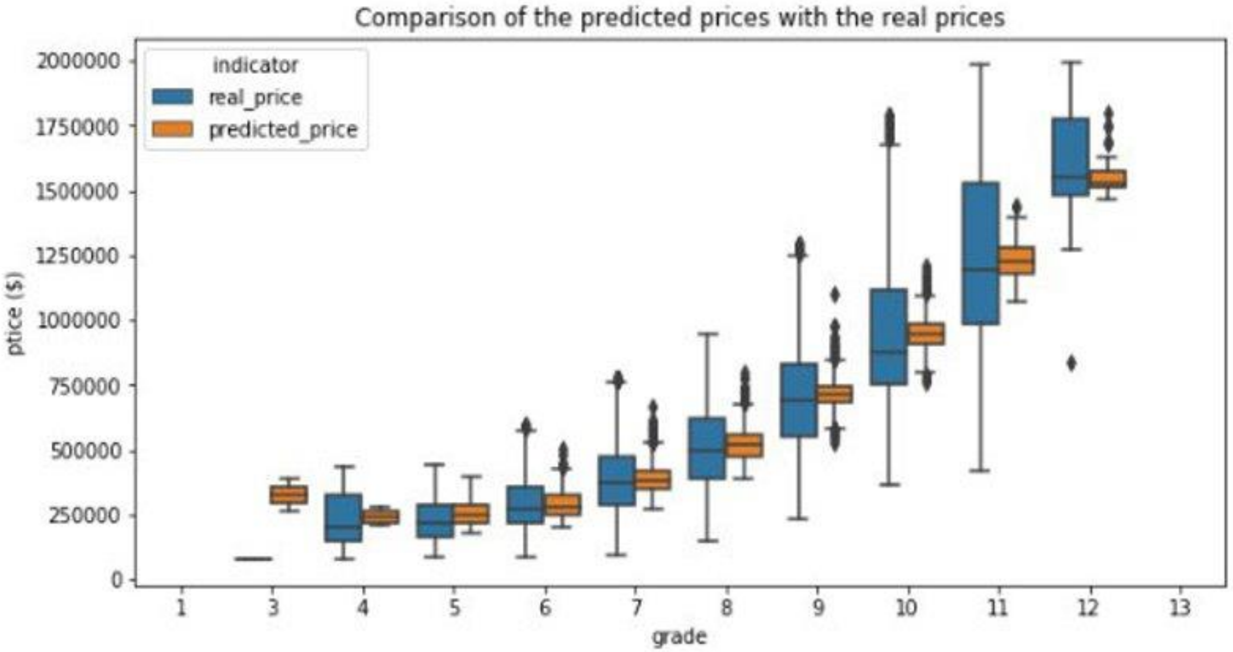
Similarly, we can conclude from the second plot that the earlier was renovation the price is lower. At the same time the higher condition the higher price.



Average price versus a renovation year and condittion

Nowadays, there are number of techniques allow to predict the price such as, for example, neuron networks or regression. Can we predict our main fearture ***price*** based on some other features and the analysis which was presented?

The answer is yes! The regression model was used. The dataset was randamly splitted to train and test sets. In addition, we found a parabolic relationship of the price with respect to the grade. The quadratic regression was used for this analysis with the predictor variables living area, grade a..



We can see that predicted variable fits our data. We observe a parabolic dependance which mimics the real price behavior.