# Predicting Home Sale Periods in the City of Sammamish WA.

## Contents

# Introduction

This project attempts to build a classifier that will be able to classify homes in the city of Sammamish Washington State into categories based on how long it will take to sell a home once it goes on the market. The homes being considered were limited to the ones within the $400 k - $1million price range. The data used for the analysis is from the King County assessor's database and MLS (Multiple Listing Service) records retrieved from the Redfin.com website for homes sold in 2014. It is expected that the engine will be able to predict the days (within a pre-determined range) a house will take to sell with a high level of accuracy.

After the data was extracted from the assessor's database, the listing dates were filled in by hand from MLS records because the assessor's records didn't have the listing dates available. The resulting data then went into a cleaning process to ensure that only useful data is used in the analysis. The cleaning steps included:

1. Removal of duplicate information (more than one record for the same house)
2. Removal of irrelevant features.
3. Normalization of attributes (replacing categorical values with numeric values)

Typical use of the results will be an agent/seller trying to figure out how long it will take to sell a house. If the resulting sale period is unsatisfactory, it should be relatively easy to see what needs to be done to improve the time within which the house will be sold.

# Data Processing

The King County assessor web site has a publicly available database that was used to retrieve the data. The format made it easy to transfer the data to an excel spreadsheet for processing. The listing date was added to these to make the needed data complete.

Given the fact that the listing dates had to be added manually, finding a match for each item in the database was a bit of a challenge. While most of the houses had matching information in the MLS data, some houses had to be removed from the list because matching records could not be found. The author opted not to attempt to add sale periods obtained using any statistical/probabilistic methods for fear of the fact that this may skew the analysis one way or the other. This is because the listing period is directly related to the class the house will fall into. Had this missing data been any other attribute, other methods would have been used to fill in the data. Initially, the data retrieved from the assessor was a little over 400 records. After the removal process, the total came to 353.

It was also necessary to adjust attributes that were not going to affect the overall results. The first attribute that was targeted was the street address. While this was useful for finding matching records in the MLS, it wasn't going to be valuable for the classifier. Another attribute that was remove was the "number of living units" which had the same value for all the homes in the data set. The listing dates were changed to listing months to see if there is a relationship between the month of listing and the time to sell. The sale date was also replaced with the sale ranks described in the table below.

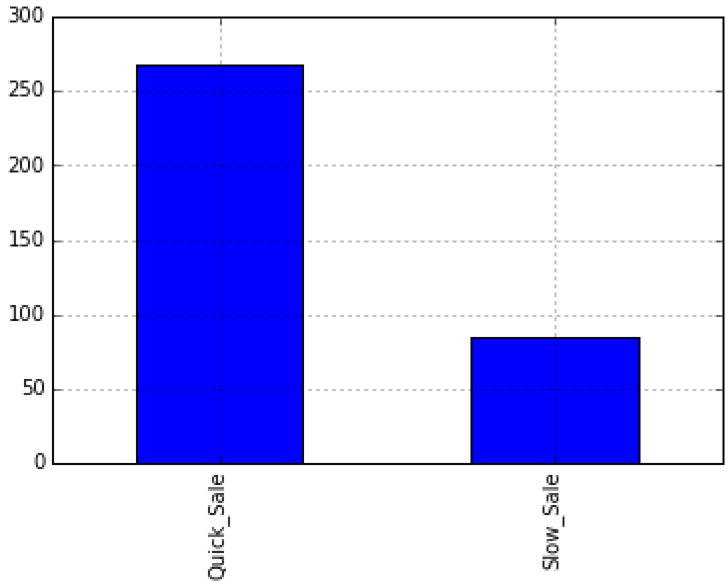
| Days on Market | Sale_Rank |
|---|---|
| 1 to 60 | Quick_Sale |
| 61 and above | Slow_Sale |

The sale rank is what the classifier will attempt to predict.

A little statistical analysis was done on the data in order to place the values into meaningful groups based on the amount of time they stayed on the market. This exercise also subjectively took into account what home owners will consider as desirable as well as the opinion of one realtor. The mean time for a sale was 52.69 days. The lowest time was 6 days and the longest time was 323 days. It worth noting that the city of Sammamish is one of the most desirable cities for home buyers is western Washington.

## EDA

Before attempting to build a classifier, a detailed EDA on the data should reveal some interesting facts that will help in choosing the right attributes to use for the classifier. First some basic statistics are observed and then each attribute is analyzed to see what information can be hypothesized. The chart below shows that majority of the home sales fall into the "quick sale" class.



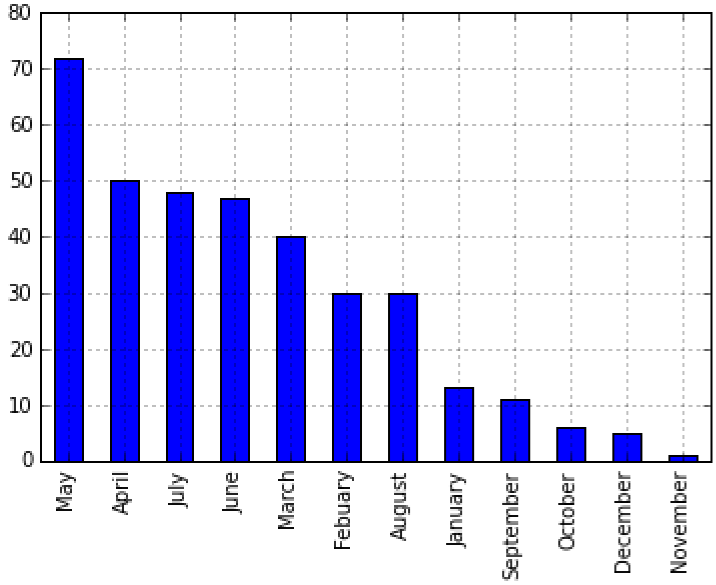
The following charts have some basic statistical analysis on the attributes from the data set. The first tablet has analysis on non-categorical data while the second table has analysis on categorical data.

|  | Sale_Price | Built_Renovated | Living_Area | Total_Basement | Finished_Basement | Bedrooms | Bathrooms | Lot_size | Per_Sqft_Price | Rank |
|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 353.000000 | 353.000000 | 353.000000 | 353.000000 | 353.000000 | 353.000000 | 353.000000 | 353.000000 | 353.000000 | 353.000000 |
| **mean** | 670605.345609 | 1993.524079 | 2679.308782 | 143.031161 | 118.356941 | 3.764873 | 2.650850 | 13568.288952 | 257.580589 | 1.240793 |
| **std** | 144233.823842 | 11.970376 | 738.256579 | 393.344693 | 340.481974 | 0.660330 | 0.528397 | 14468.396331 | 46.305457 | 0.428172 |
| **min** | 406430.000000 | 1932.000000 | 1010.000000 | 0.000000 | 0.000000 | 2.000000 | 1.000000 | 1842.000000 | 143.318965 | 1.000000 |
| **25%** | 548000.000000 | 1988.000000 | 2110.000000 | 0.000000 | 0.000000 | 3.000000 | 2.500000 | 6056.000000 | 234.417344 | 1.000000 |
| **50%** | 660000.000000 | 1993.000000 | 2650.000000 | 0.000000 | 0.000000 | 4.000000 | 2.500000 | 8433.000000 | 254.180602 | 1.000000 |
| **75%** | 775000.000000 | 2003.000000 | 3220.000000 | 0.000000 | 0.000000 | 4.000000 | 2.750000 | 13609.000000 | 271.296296 | 1.000000 |
| **max** | 1000000.000000 | 2014.000000 | 5220.000000 | 2270.000000 | 2120.000000 | 7.000000 | 4.500000 | 108464.000000 | 762.831858 | 2.000000 |

|  | Listing_Month | Grade | Condition | Zoning | Sale_Rank |
|---|---|---|---|---|---|
| **count** | 353 | 353 | 353 | 353 | 353 |
| **unique** | 12 | 6 | 4 | 11 | 2 |
| **top** | May | Good | Average | R4 | Quick_Sale |
| **freq** | 72 | 137 | 287 | 147 | 268 |

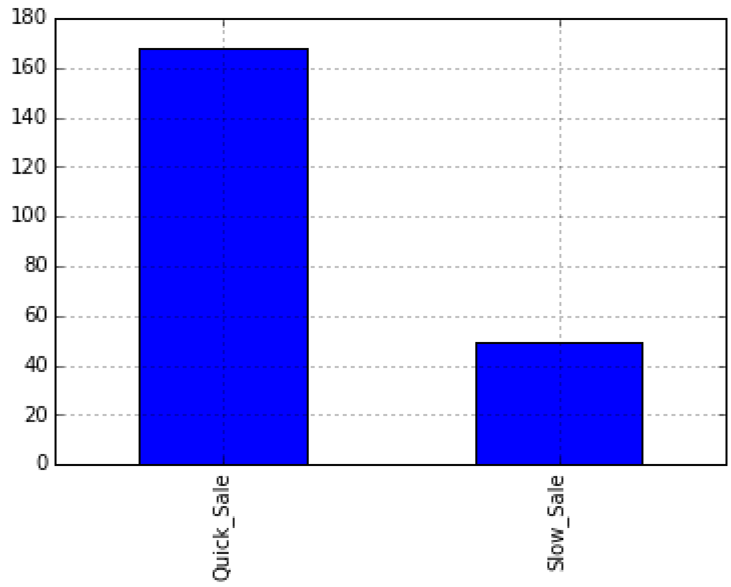
## Listing Month

As stated in the introduction, the “listing month” was added as an attribute in order to see if there’s a pattern for different periods of the year. Realtors in Washington usually say spring/early summer is the best time to list a home for sale. Testing this assertion should be easy. The chart below shows the distribution of sales by month in the data set.

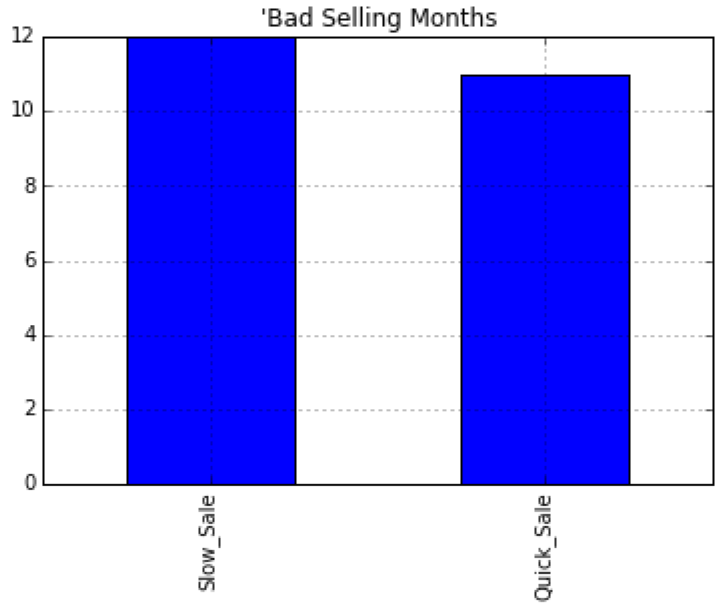


The data suggests that people are listening to their realtors but is there really a benefit to selling within this period? It will be interesting to compare the slow sales in this period to the slow sales for the months Realtors claim are bad listing periods.

First we look at the sales for the "good" listing period



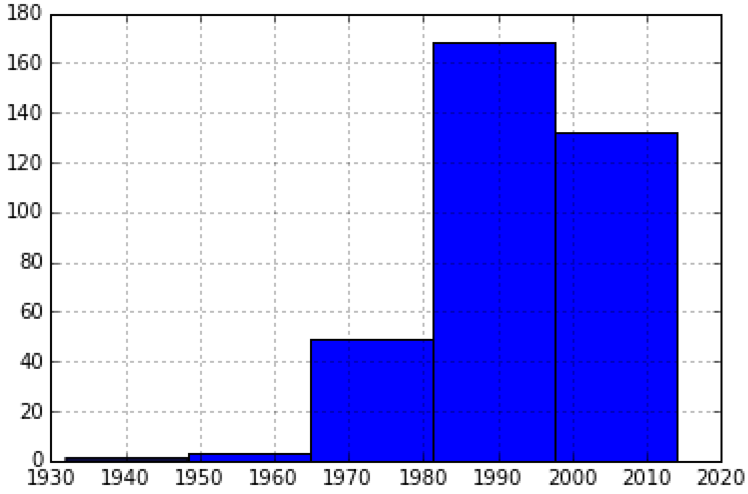
The values seem to reflect the global trend. Now let’s look at the "slow" period sales.



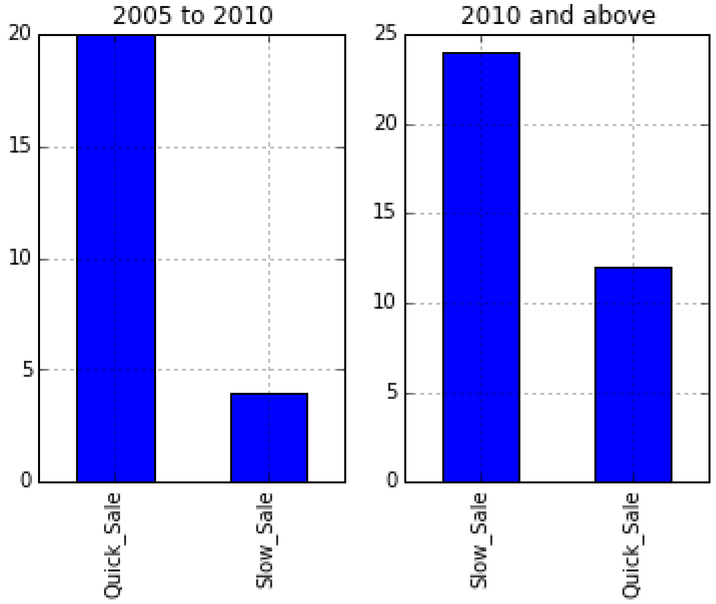
The evidence shows that the "slow" period is not a myth and one should probably avoid trying to sell a home during the winter period. During this time despite fewer listings, homes take longer to sell. It is conceded thought that the amount of data may not necessary be enough to treat this as absolute.

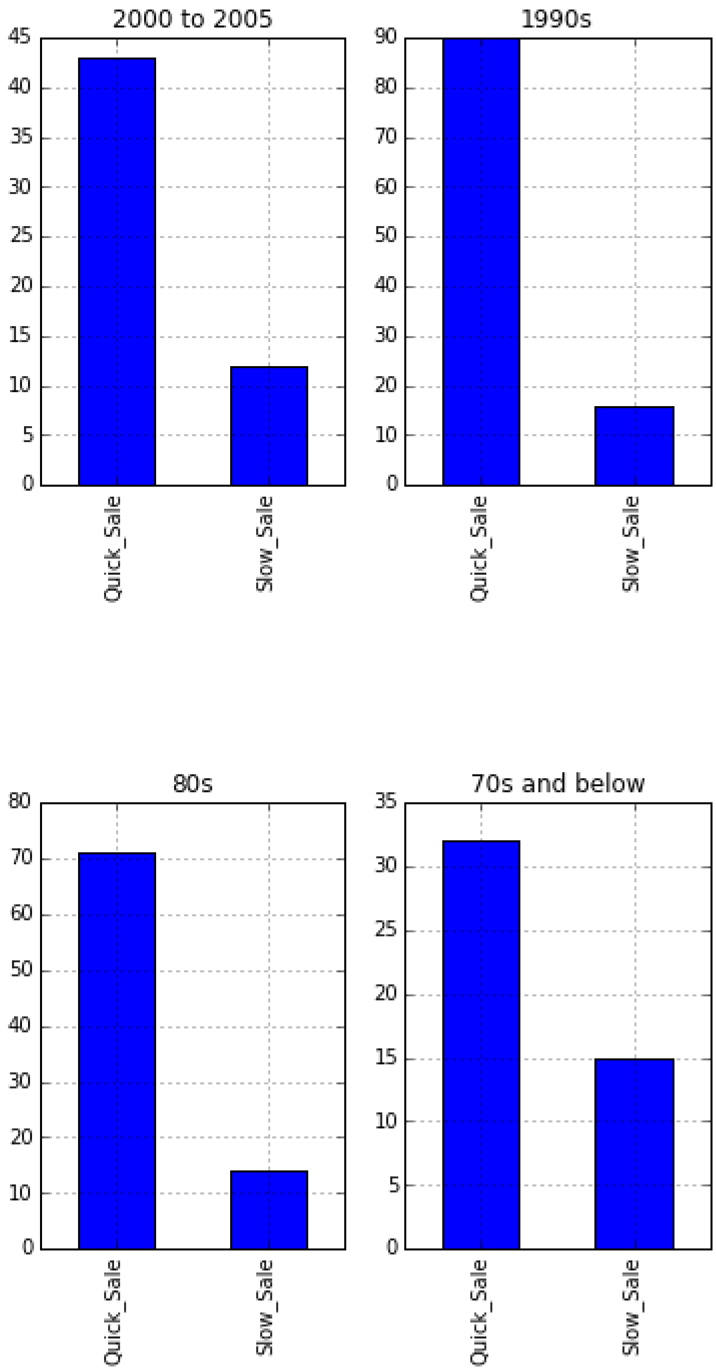
## Age

The age of the house (the age since the last major renovation) is likely to have a significant effect on how fast a house will sell. The chart below show the distribution of homes in the data set by age.

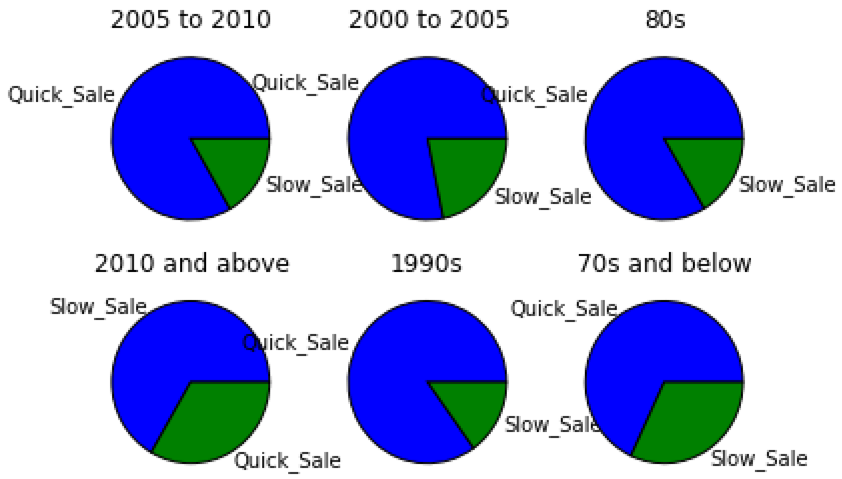


Based on the distribution of the data, homes were broken into different periods and sale time was observed for each of the chosen periods as shown in the following charts:





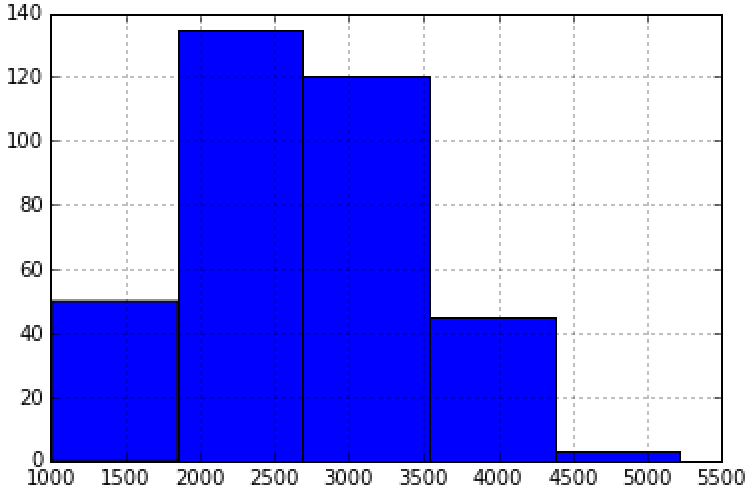
2000 to 2005
0
5
10
15
20
25
30
35
40
45
Quick_Sale
Slow_Sale
1990s
0
10
20
30
40
50
60
70
80
90
Quick_Sale
Slow_Sale
80s
0
10
20
30
40
50
60
70
80
Quick_Sale
Slow_Sale
70s and below
0
5
10
15
20
25
30
35
Quick_Sale
Slow_Sale



Given the charts above, there really is no evidence to suggest that the age of a house will have a definitive impact on how quickly a house will sell. One interesting fact is that the newest homes seems to take longer to sell. This is likely due to practice in new construction where homes are listed long before they are built.

## Living Area

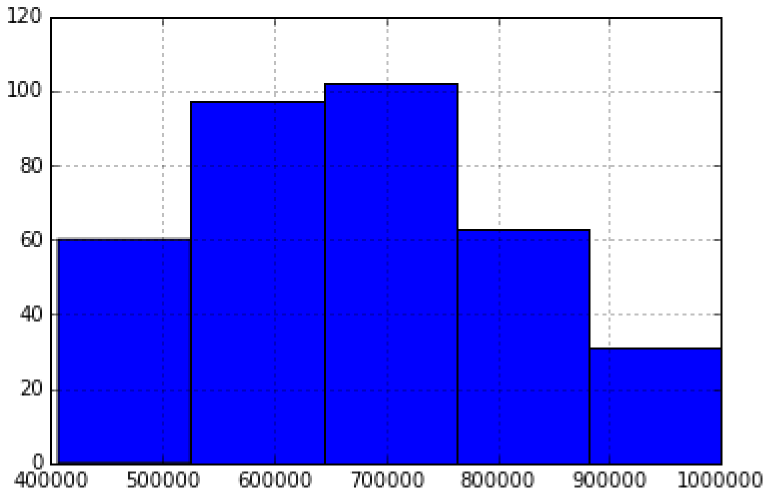
It's difficult to make any assumptions about the living area as this particular attribute is more likely to be affected by the price and other attributes. The distribution is shown below



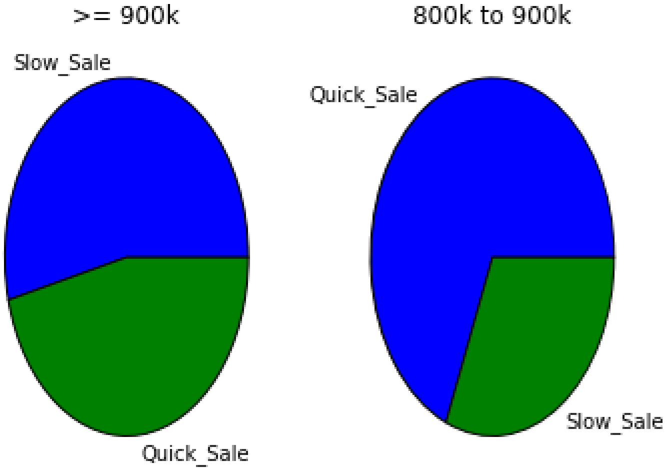
The attribute by itself isn't expected to show anything interesting. Consequently, more effort was investing in analyzing the price per square feet instead as that attribute would likely reveal value for money which is something buyers are likely to pay attention to.

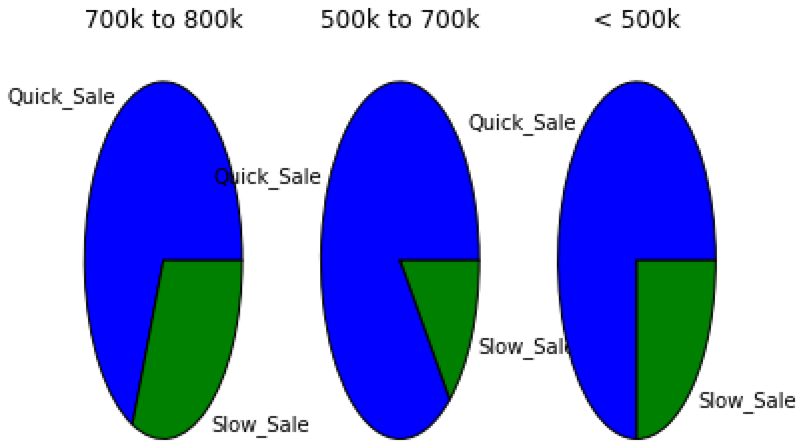
## Price

The price of the house is very likely to be a significant determining factor for the simple fact that fewer people are likely to be in the market for expensive homes. As a result, expensive homes will likely be on the market for longer than relatively affordable homes. The chart below shows the distribution by price.



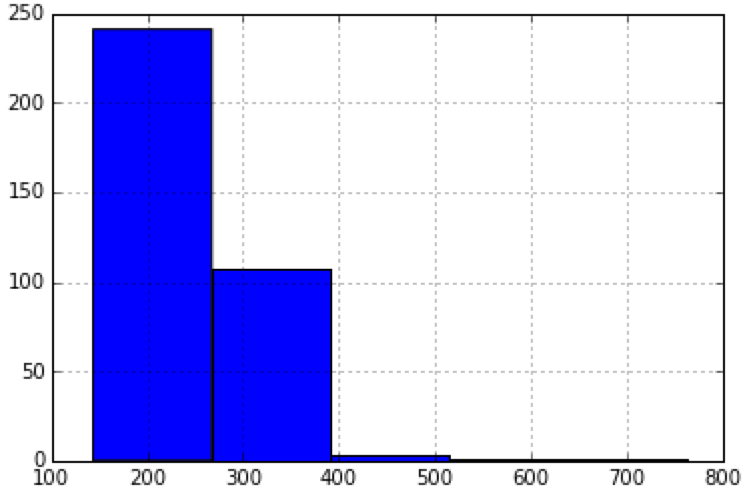
As suspected, the top end of the homes took longer to sell as show in the charts below which categorize home sales base on the price ranges. It appears like the $900k price point, is when the price really starts to have a significant effect on the time it takes to sell the home.

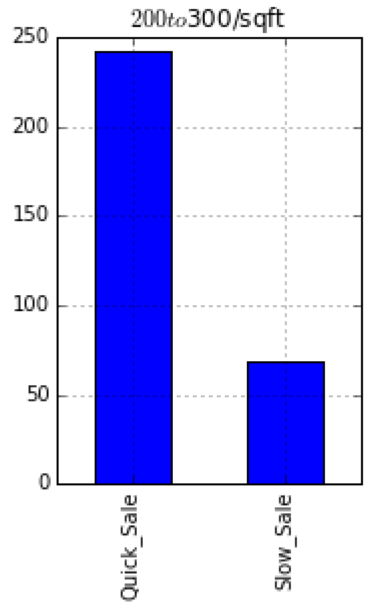


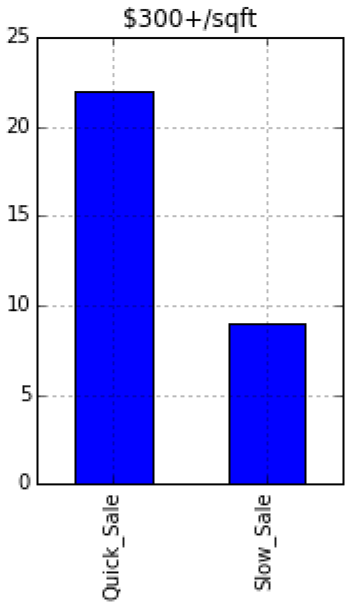


## Price per Square Foot

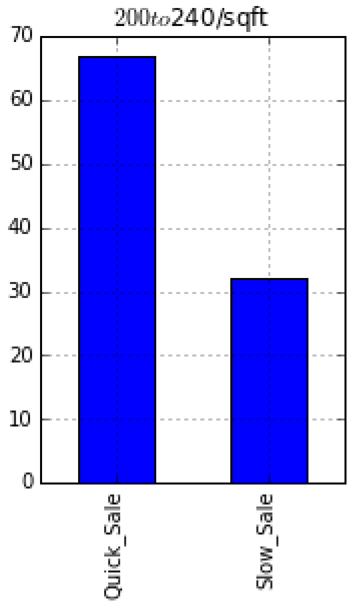
Unfortunately there isn't much variation in our data set. However, the effect of this is examined all the same.

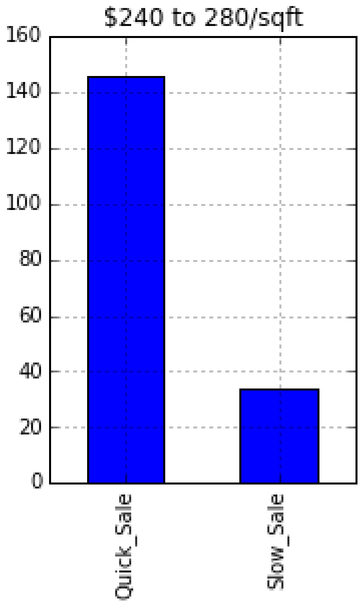


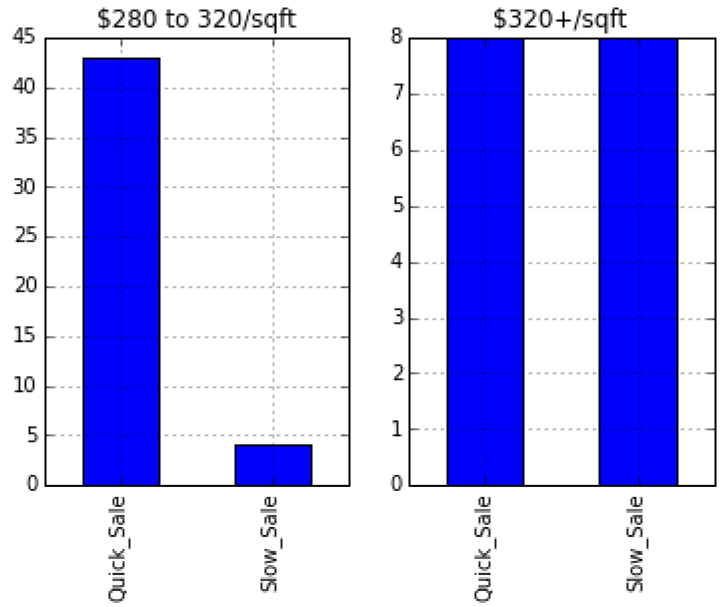




As expected the limited variation didn't yield anything conclusive. This will seem to indicate that the price per squared feet is of little importance. However, given the expectation that this factor is important, further analysis was done to see if a pattern could be found. The charts below use smaller ranges of this attribute.



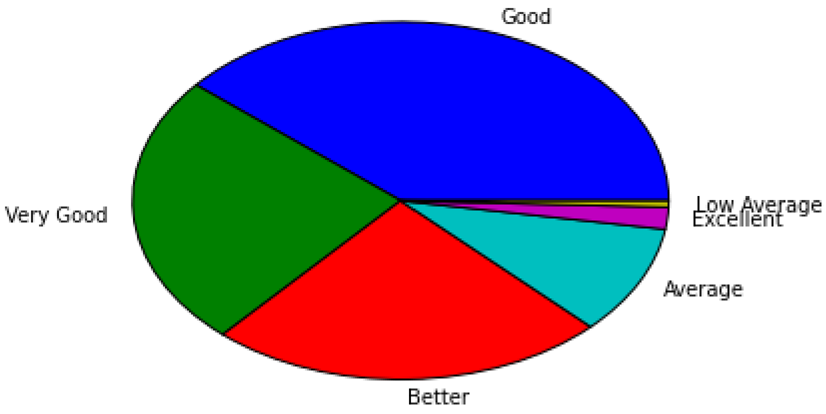




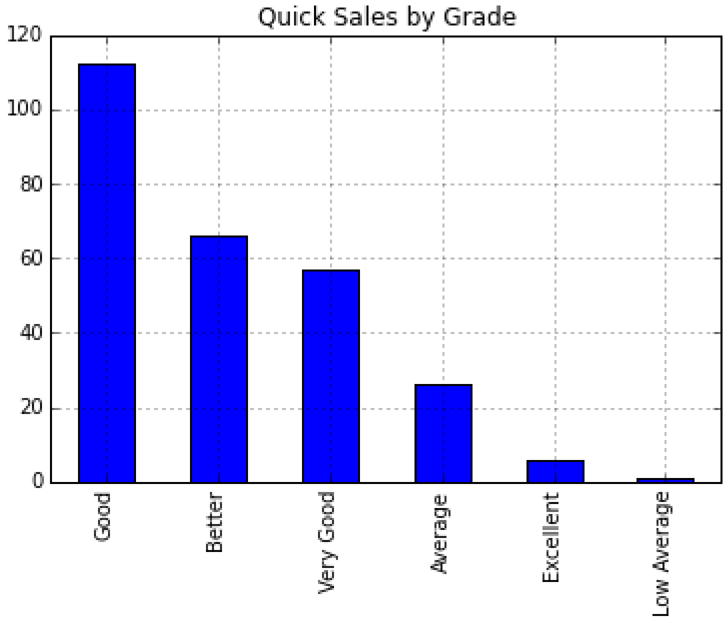
Breaking down the data into smaller ranges reveals an interesting pattern. As price per square feet increases the percentage of slow sales appear to decrease. However, once the $320 mark is reached there is a spike in the percentage of slow sales to the point where the slow sales are equal to the quick sales. This represents a very significant departure from the trend observed in the whole data set. This appears like a clear indication that the price per square feet starts to weigh in significantly once the $320 price point is reached.

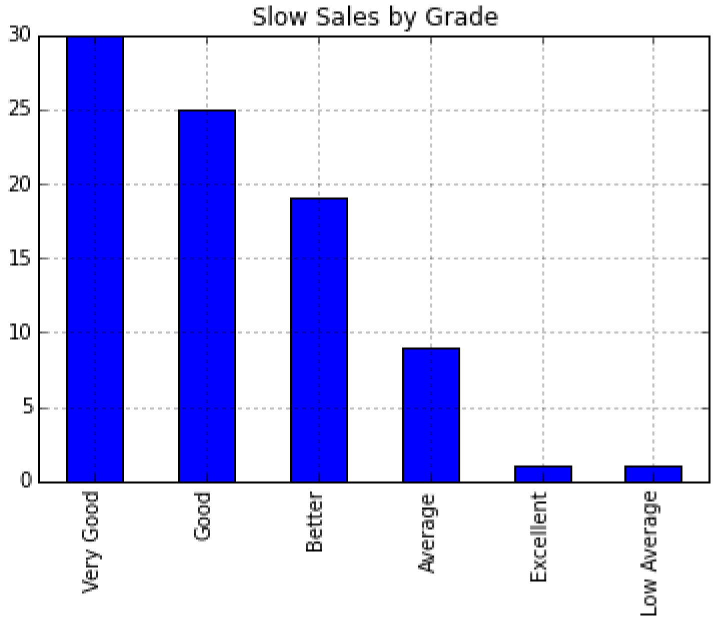
## Grade

The meaning of the "Grade" attribute (and particularly how it was determined) wasn't completely obvious so some analysis of the data was done to see what the dominant categories were. The distribution of the data by grade is shown in the following pie chart.



Viewing this data by sales reveal the following patterns.

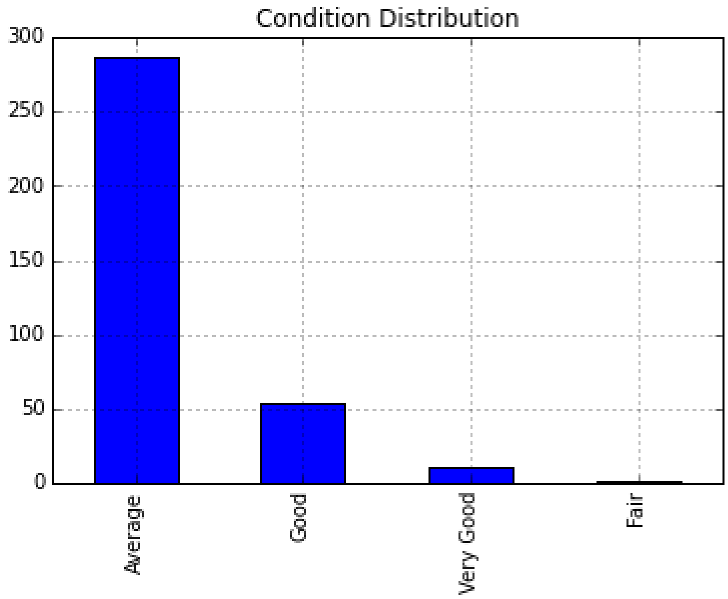




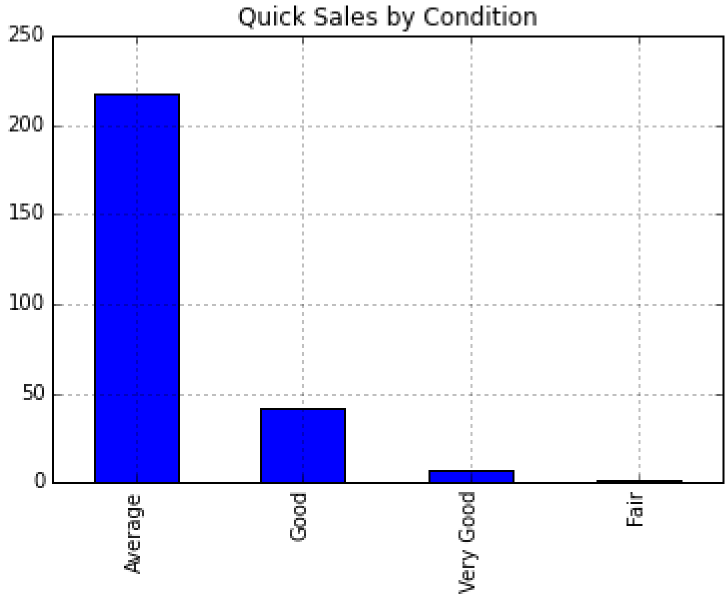
One can't really draw any interesting conclusion about the attribute given the grade that has the largest share of slow sales was the second best grade. The small population size may be having a bad effect on the analysis here. There’s also a good chance that there was a significant amount subjectivity in the assignment of values to this attribute.

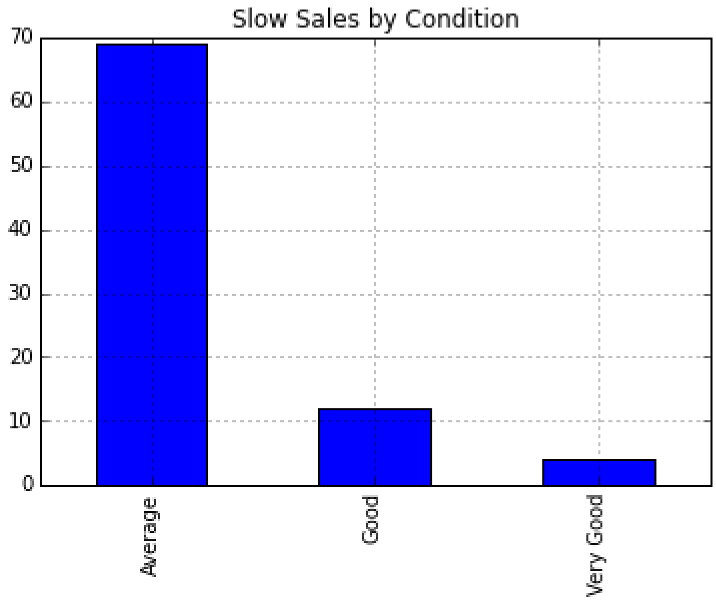
## Condition

One would expect that the condition of the house will have an effect on how quickly a house will sell. However, it is likely that mostly houses put up for sale will be in fairly good condition and the data as shown in the chart below appears to back up this assumption



Given, this skewed data showing that most homes were in average condition, it is not expected that this feature will have much of an effect on sales. This is assumption is examined in the charts below.

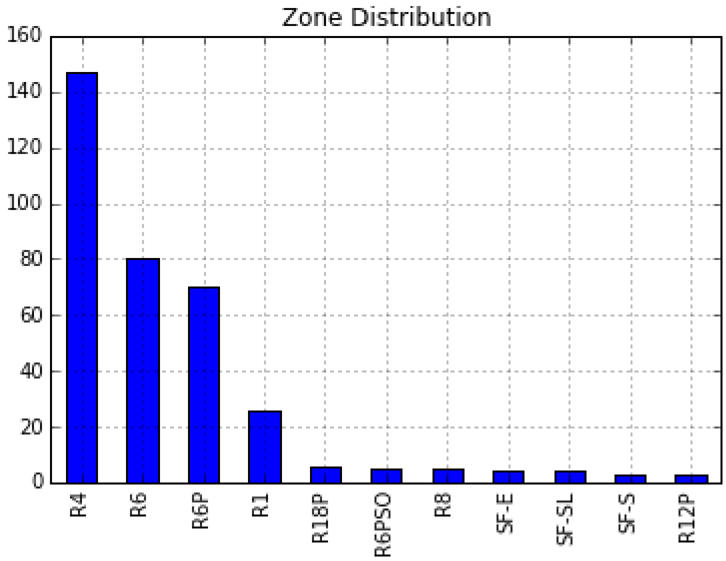




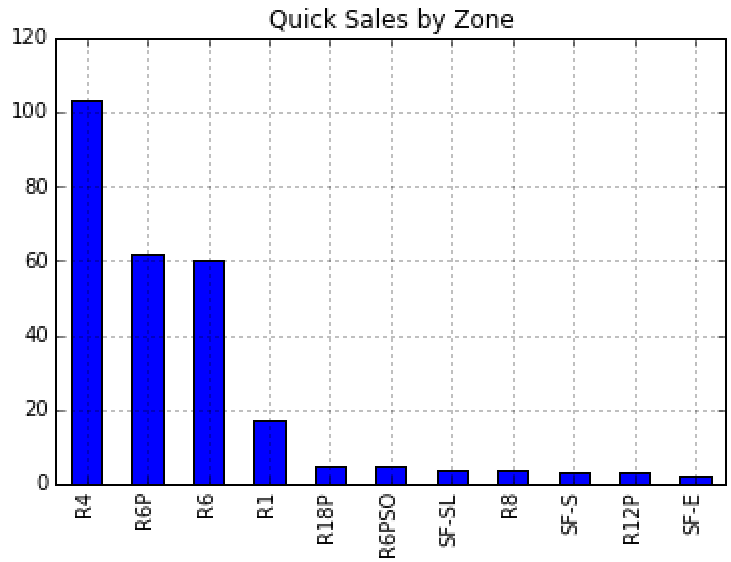
As suspected, there's no evidence of any significant contribution of this attribute to the sale of a house.

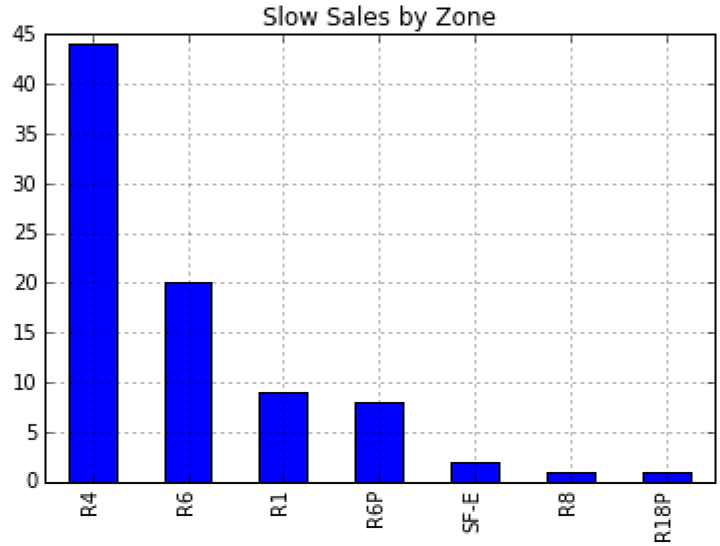
## Zoning

The Zoning attribute is a little mysterious given the odd values. Some statistical analysis below may give some clues about its relevance.



R4 is the zone with the highest number and everything after R1 is practically irrelevant. The charts below show sale by Zones.

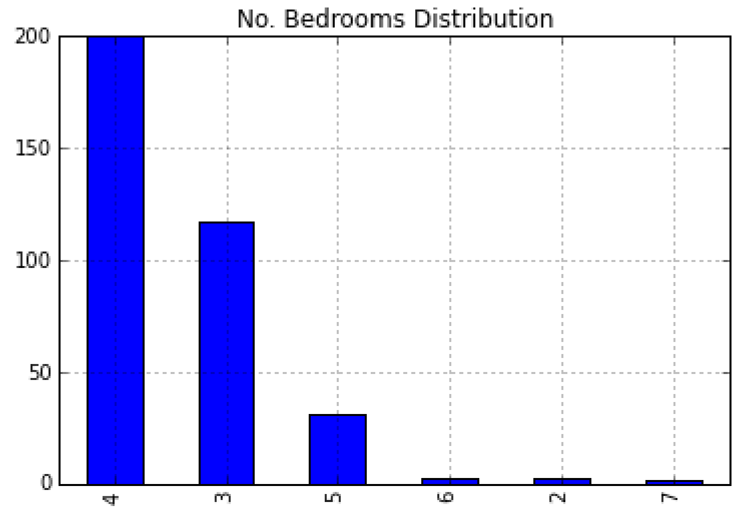




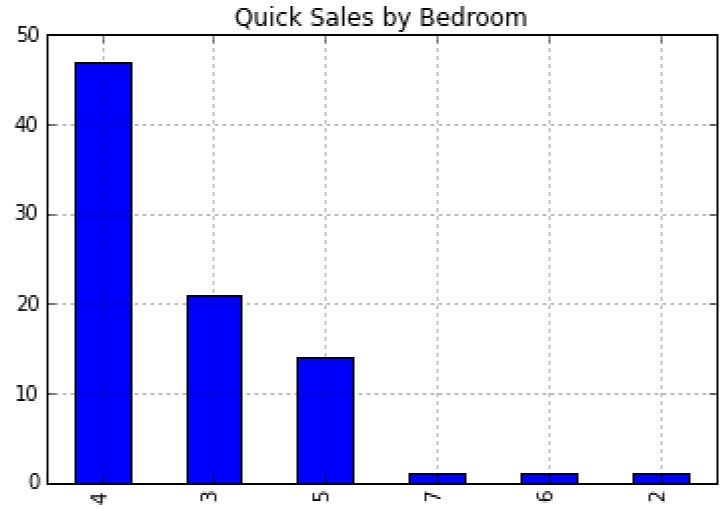
Given these charts, it appears that the zoning has little effect on the sale time.

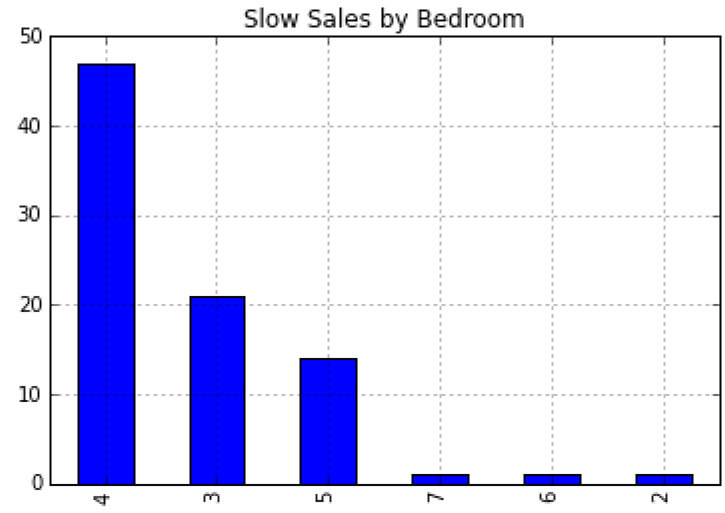
## Bedrooms

The distribution of bedrooms is shown below. While it’s not clear what to expect from number of bedrooms, the combination of bedrooms and bath rooms may reveal interesting patterns.

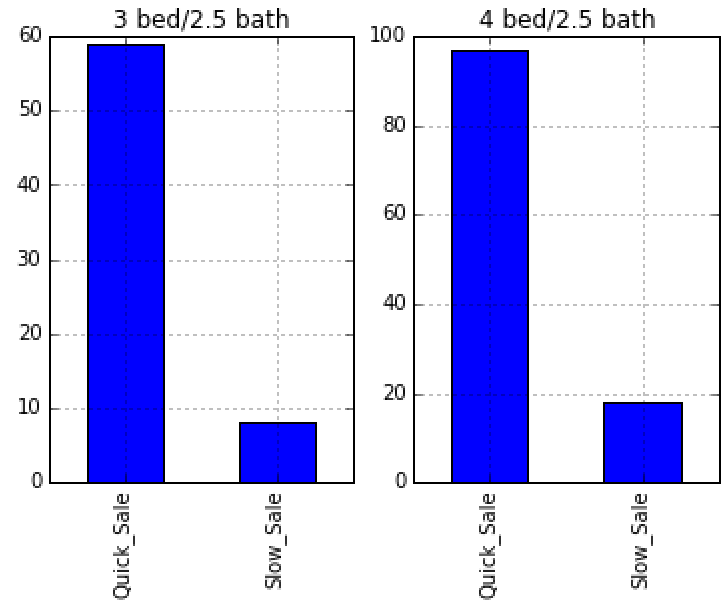


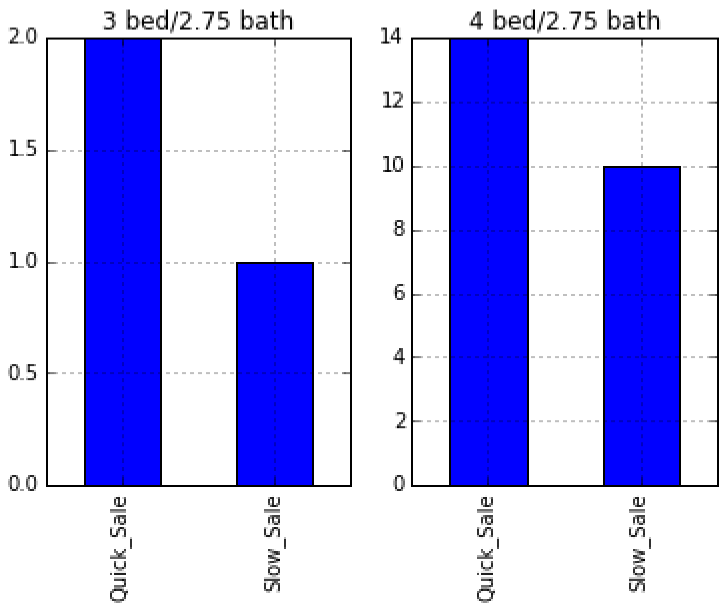
Since 4 bedroom and 3 bedroom homes take up the lion share of homes in the population it will be interesting to see the bathroom combinations that do well. Before that, it’s important to examine the distribution of bedrooms within the 2 classes we’re examining.





There appears to be an even spread in both classes so the number of bed rooms don't seem to play much of a role in how fast a house will sell. Now we consider some 3/4 bedroom bathroom combinations.

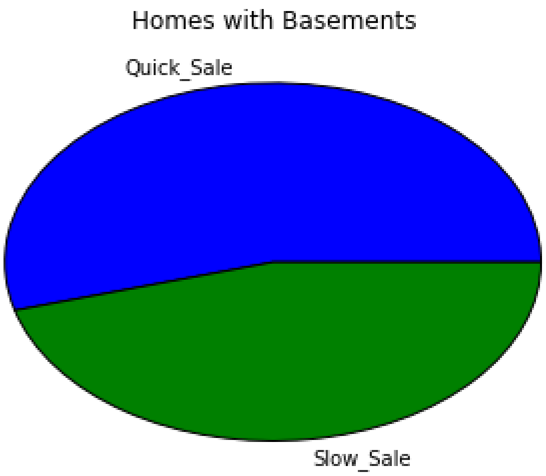




Looking at this data in isolation, it would appears that a 3 bed/2.5 bath house is likely to sell faster than the other combinations as it has the lowest percentage of slow sales.

## Basement

Basements are not very common in the state of Washington so their impact on sale time is of interest. There are a total of 46 homes with basements in the data set making up about 13% of the homes. Looking at the sales data as shown in the pie chart below, there appears to be a near even split between quick and slow sales so it’s safe to conclude that buyers aren't very interested in the presence of a basement. However, given the fact the slow sales are significantly fewer in the data set, having a basement may be a negative thing.



### EDA Summary

In the EDA, all attributes were individually examined for trends that can give clues to how important the attributes are for a quick sale of a home. The price of the home appears to be the most important feature. Price per sq. foot was expected to be a significant fact but it only appears to have an effect at the top end of the range. The relevance of attributes such as condition grade and zoning appears to be relatively insignificant.

## Random Forest Classification

A random forest is a machine learning algorithm that's a combination of individual decision trees that in all has a better result than an individual decision tree. Random forests can be used for classification, regression and other machine learning tasks.

### Classifier

The random forest classifier available in scikit learn (a machine learning library for python) was used to classifier the data. The data was split into a training set and test set. In order to avoid the effects of over-fitting, cross validation was used. Since accuracy doesn't always tell the full story when running a classifier, the area under the curve (AUC) along with precision and recall was also used to assess the performance of the classier.

### Performance

After the first run. The accuracy score was 0.8169 while the AUC score was 0.736. The upper bound cross validation AUC score was 0.7917.

The  classification report is listed below:

```
               precision    recall  f1-score   support

          0       0.50      0.31      0.38        13
          1       0.86      0.93      0.89        58

avg / total       0.79      0.82      0.80        71
```

### Precision:

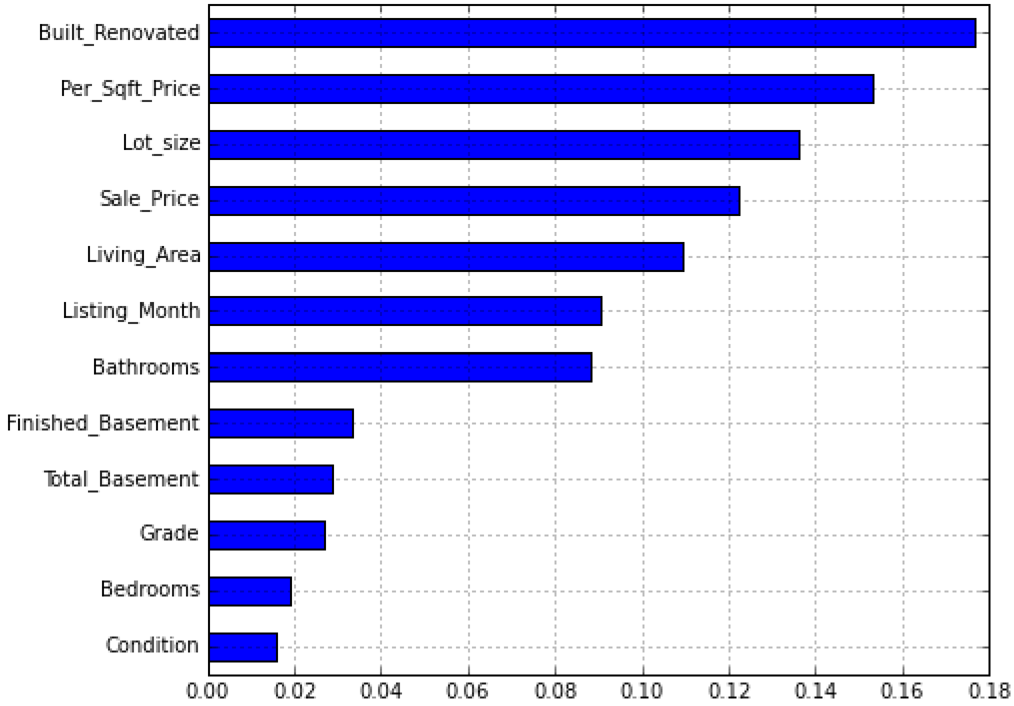
The classifier will accurately identify 86% of quick sales. There's a 14% chance that it will classify slow sales as quick sales.

### Recall

Of all the sales that were classified as quick, 93% of them were actually quick.

## Most Important Attributes

The feature importance as identified by the random forest classifier is as follows:



This insight triggered an attempt to get a better score by dropping the 3 least important features and running the model again.

## Second Model Performance

The second model resulted in a slight better accuracy of 0.831. The AUC score was 0.7241 while the upper bound cross validation AUC score was 0.8003. The classification report is as follows:

```
               precision    recall  f1-score   support


          0       0.56      0.38      0.45        13
          1       0.87      0.93      0.90        58

avg / total       0.81      0.83      0.82        71
```

## Precision:

The classifier will accurately identify 87% of quick sales. There's a 13% chance that it will classify slow sales as quick sales. On the other handle the precision for slow sales is only 50%. It's worth noting that the precision on slow sales was better than the first model.

### Recall

Of all the sales that were classified as quick, 93% of them were actually quick. But the recall was only 38 % for slow sales. Again this is better than the first model.

In all, the second model had modest improvements for all the performance measurements except the A UC score before cross validation.

## Conclusions

The results of the classifier were less accurate than expected but this is likely due to the limited amount of data that was used. There are also some factors that are not accounted for by the attributes available. One of which is the proximity to lake Sammamish which will likely make people have a higher tolerance for homes with high price per square foot.

It appears that preparing to sell a home has to start from the time the home is bought. If a consumer plans to sell their home at some point, it is important for the consumer to consider a number of factors that will affect the sale of the home. From the analysis that has been done in this project, the most important attributes to pay attention to is age and the price per sq feet. Listing the home in the summer also appears to be a good idea.