This report is to extract and prepare data describing certain characteristics and prices of a number of properties in three different neighbourhoods. It includes two sections:

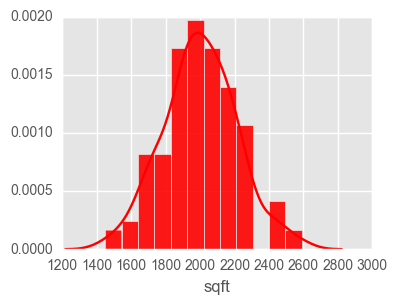
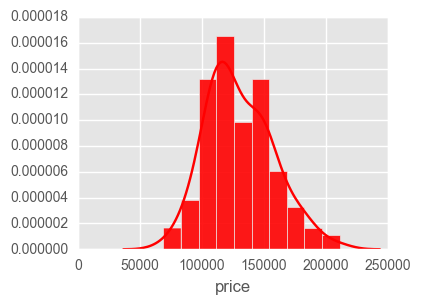
1. Exploratory Data Analysis
2. A simple linear model (1 factor) to predict price given the square footage of a property.

**Exploratory Data Analysis**

There are 128 rows of data with 7 variables. There are no missing values.

nbhd and brick are categorical variables with unique values as 'nbhd02' 'nbhd01' 'nbhd03' and ‘Yes’ and ‘No’ respectively.

offers, bedrooms and bathrooms are discrete variables with 6, 4 and 3 unique values.

Sqft and price are continuous variables. Sqft has a minimum value of 1450 and maximum of 2590 with the mean of 2000.93. **There is no property between 2300 and 2400 sqft.** The minimum price of the property in the dataset is 69100 and maximum is 211200. The mean value of price is 130427.34.

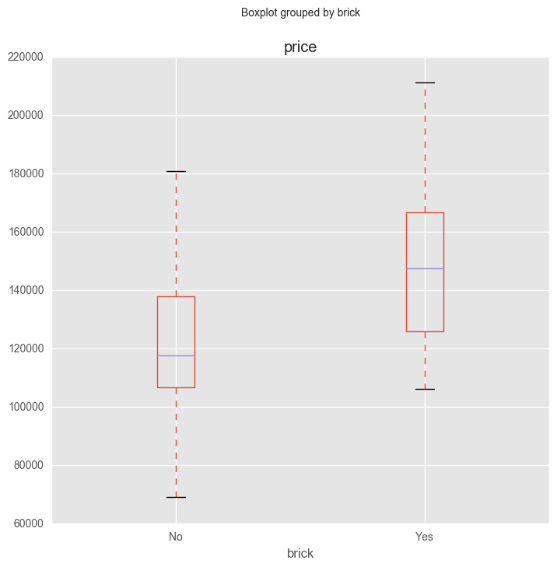
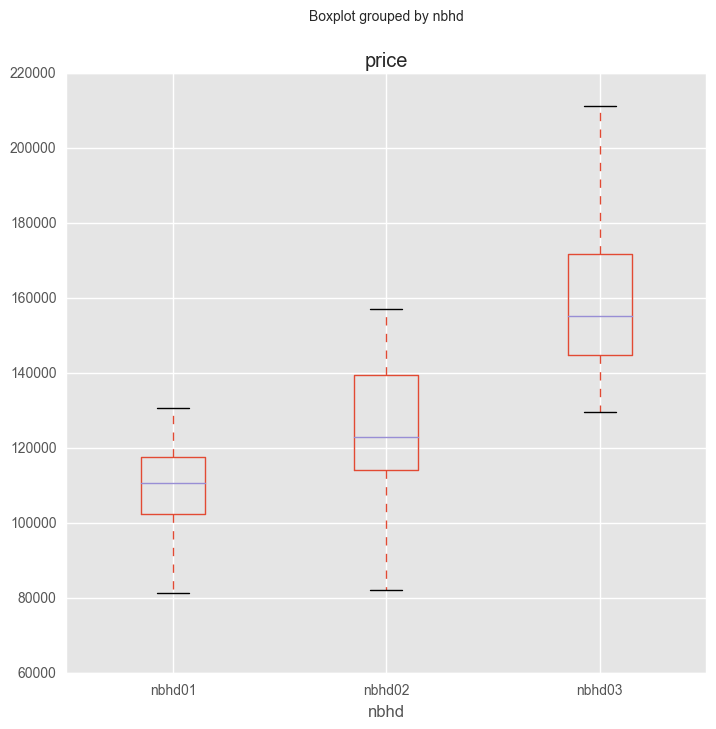
The number of properties distributed according to the neighbourhood is as follows:

nbhd01: 44

nbhd02: 45

nbhd03: 39

The number of properties built with brick is 42 while 86 are not built with brick.

The median price of the home built without brick is 117650.00 while with brick is 147500. While median prices in three different neighbourhoods are 110750 in nbhd01, 123100 in nbhd02 and 155400 in nbhd03.

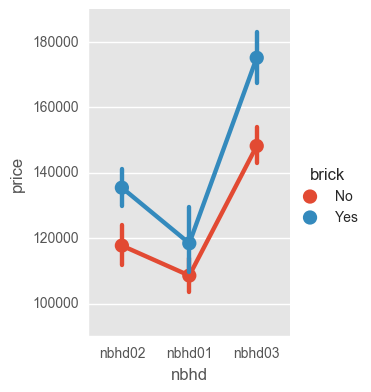
The mean price/ square feet in different neighbourhoods are as follows:

Nbhd01: 57.46

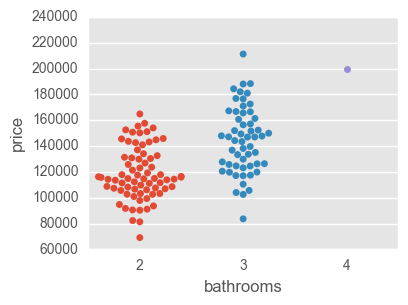
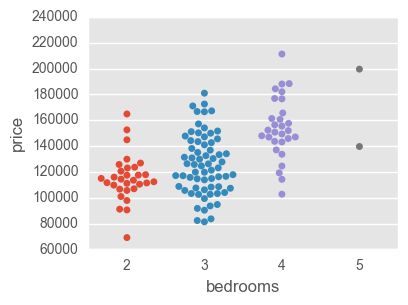
Nbhd02:62.18

Nbhd03: 76.55.

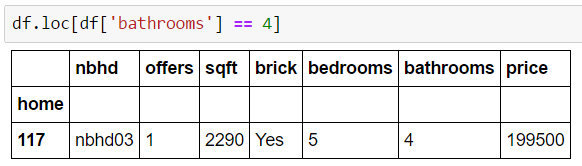
**Therefore, the price of the home built with bricks is more than built without bricks. Also, the homes in nbhd03 are the most expensive and larger in size.**

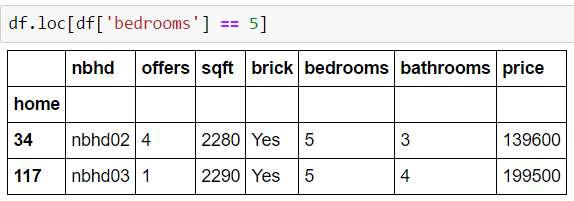


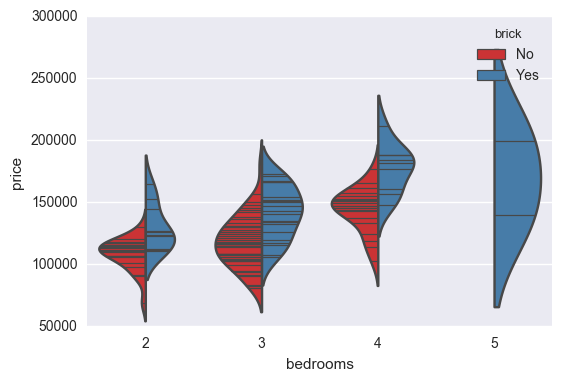
Since the lines representing the bricks in the plot are not parallel, this implies an interaction effect between brick and neighbourhood. **So, how price changes with neighbourhood depends on the brick, and vice versa.**



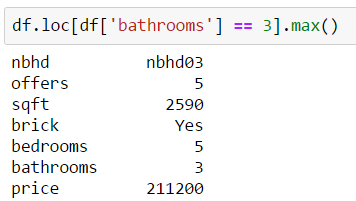
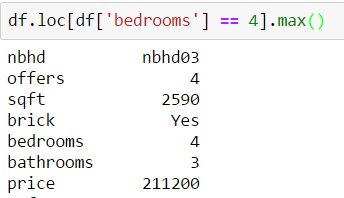
The above plots show that most of the properties have 3 bedrooms and 2 bathrooms. The price increases with the number of bedrooms and bathrooms. There are only 2 properties with 5 bedrooms and 1 property with 4 bathrooms. However, the price for such homes is less than 4-bedroom property and 3-bedroom property respectively. **Therefore, other factors must play a part in deciding the price.**

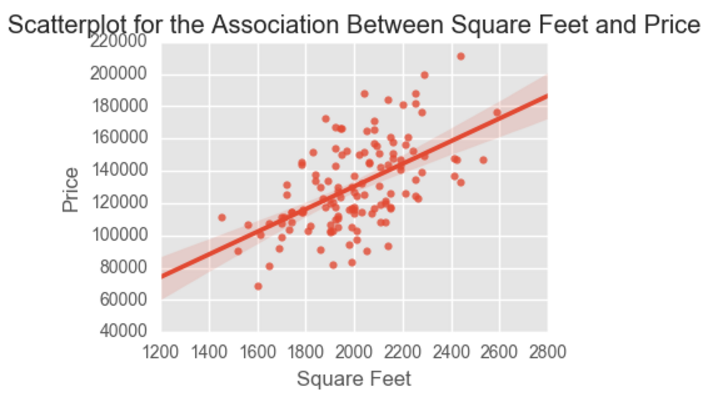




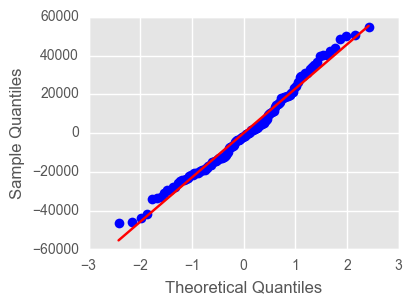


The above plot shows that as the number of bedrooms’ increases, more homes are built with brick and show the increase in the price as well. However, upon further exploration, the two separate properties with more valuation and 3 bathrooms and 4 bedrooms respectively show that **the size and neighbourhood are more important factors than bathrooms and bedrooms** in deciding the price of a property. **Offers show a negative effect on price** because the property with 4 bedrooms has one less offer than the property with 5 bedrooms but still has the same selling price.

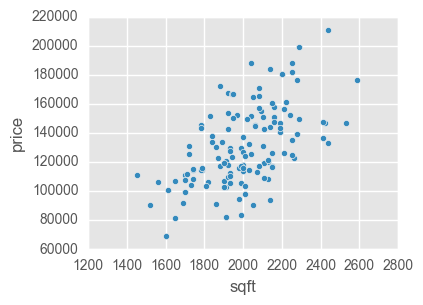
 



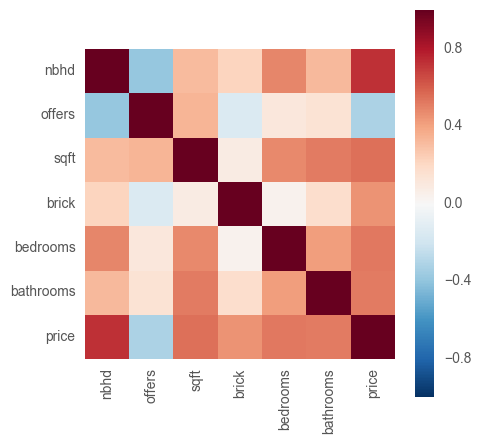
The above scatterplot shows how well price and sqft are related. It is a positive linear relationship.



As the above Q-Q plot shows that the two sets of variables, price and sqft, are normally distributed, one can take a Pearson Correlation Coefficient. With 0.55, it is a strong degree of relationship and does not show a cause-effect relationship.



However, the two properties with the most sqft (2590 and 2530) have less price than 2440 sqft. 2440 sqft property have the same neighbourhood, bedrooms and bathrooms but fewer offers made and is constructed with brick. **Hence, offers has a negative effect while brick have a positive effect on price.** The correlation plot below confirms the relationship between price and other factors.



The features are arranged by their correlation with price in decreasing order:

nbhd: 0.7323335104041765

sqft: 0.5529822425554838

bedrooms: 0.5259260576398529

bathrooms: 0.5232577581859738

brick: 0.4528167861905481

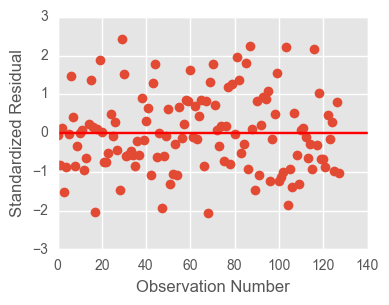
offers: -0.3136358828617982

The variables show medium to strong strength of association with price. Therefore, it is possible to make a probabilistic model with

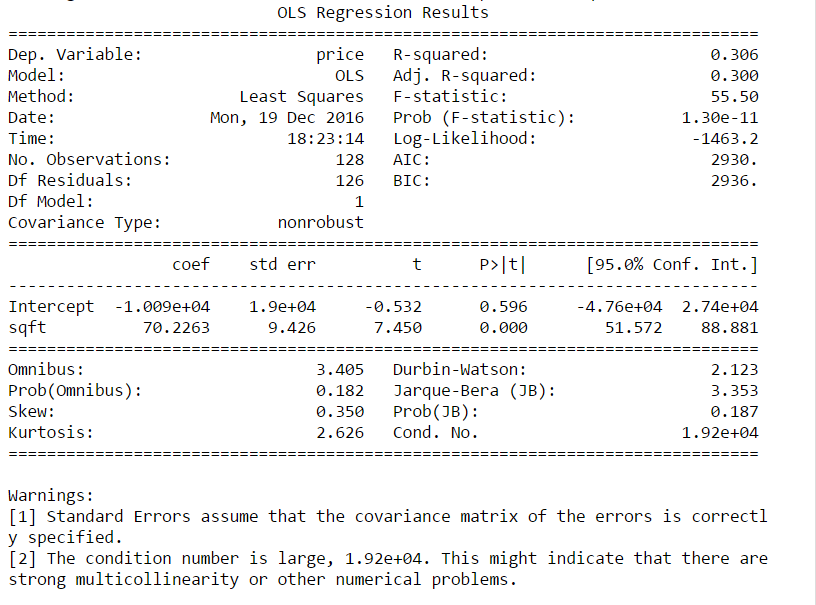
**Response Variable = Deterministic Variable + Random error.**

**Simple Linear Model to predict price given the square footage**

A residual plot determines if a linear regression model is appropriate. Following is a **Residual plot** for price vs sqft



The residual plot above falls in a symmetrical pattern and have a constant spread throughout the range. Hence, the linear regression model is appropriate for the data.

****

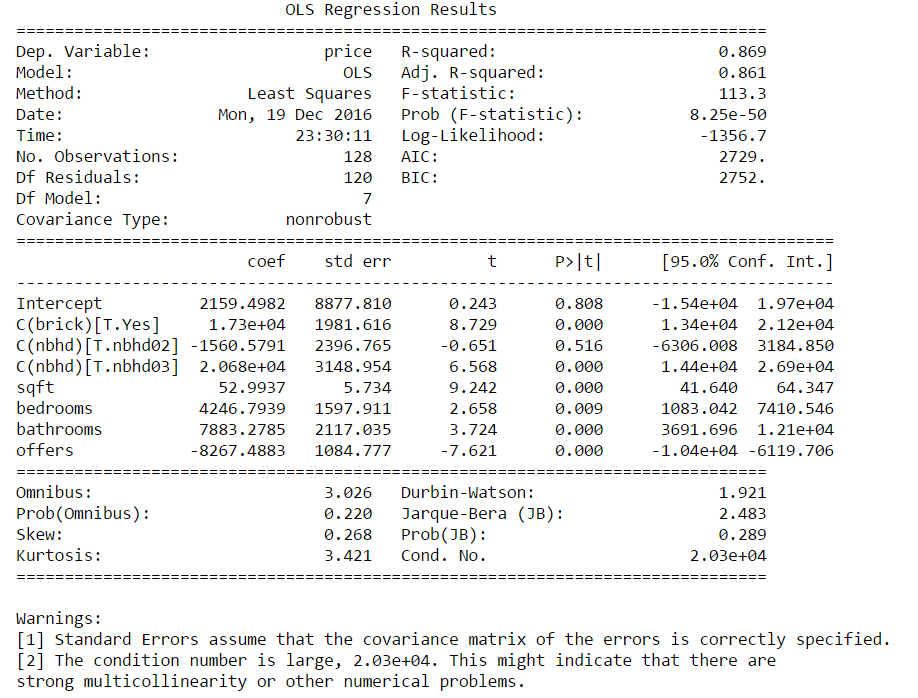
Simple Linear model is:

**Price = -10091.1299 + 70.22 \* sqft**

* Intercept is -10091.12. It is the point at which the line intercepts or cuts through the y-axis when sqft = 0.
* Slope is 70.22. It is a positive slope, hence, the price (y) increases 70.22 for every 1-unit increase in sqft (x).
* Standard error of 9.42 is the quantitate measure of uncertainty – the possible difference between the true price and the square footage of the property.
* F-statistics is 55.50 and Prob(F-statistics) is very small at 1.30e -11. Hence we can reject the null hypothesis.

The coefficient of determination (R-squared value) determines the proportion of variance in price that can be accounted for by knowing sqft and vice versa. It is also a statistical measure of how close the data are to the fitted regression line. The regression model explains only 30.6 % of the variation in price due to square footage of the property. Hence, little variance is accounted for by the regression model and the data points fall further away from the fitted regression line. **Consequently, it is not a very good model fit to the data.**

**A Multivariate model using all the variables (R-squared value of 0.869)**

****