In these studies, we used multiple regression models because we want to predict independent variable Market price with reference to the dependent variable. These models contain many dependent variables and independent variable.

Before building multiple linear regressions, Let us first look at the data; we have total 15 different year wise sample data for different variable. In this data we have different variable named Sydney price Index, Annual % change, Total number of square meters, age of the house (years) and Market price. According to the data Market price variable depending on remaining variables.

Firstly, we did an exploratory data analysis for finding missing value or any outlier. This is first step of data analysis which is called is data pre-processing. We find missing value, there is no missing value. Next we find outliers, outlier is also called as extreme value. There can be many reasons for the presence of outliers in the data. Sometimes the outliers may be genuine, while in other cases, they could exist because of data entry errors. It is important to understand the reasons for the outliers before cleaning them.

We will start the process of finding outliers by running the summary statistics on the variables.

**Identifying Outliers with Interquartile Range (IQR):**

The interquartile range (IQR) is a measure of statistical dispersion and is calculated as the difference between the 75th and 25th percentiles. It is represented by the formula *IQR = Q3 − Q1*. The lines of code below calculate and print the interquartile range for each of the variables in the dataset.

Then we find lower and upper region by formula which is given in below,

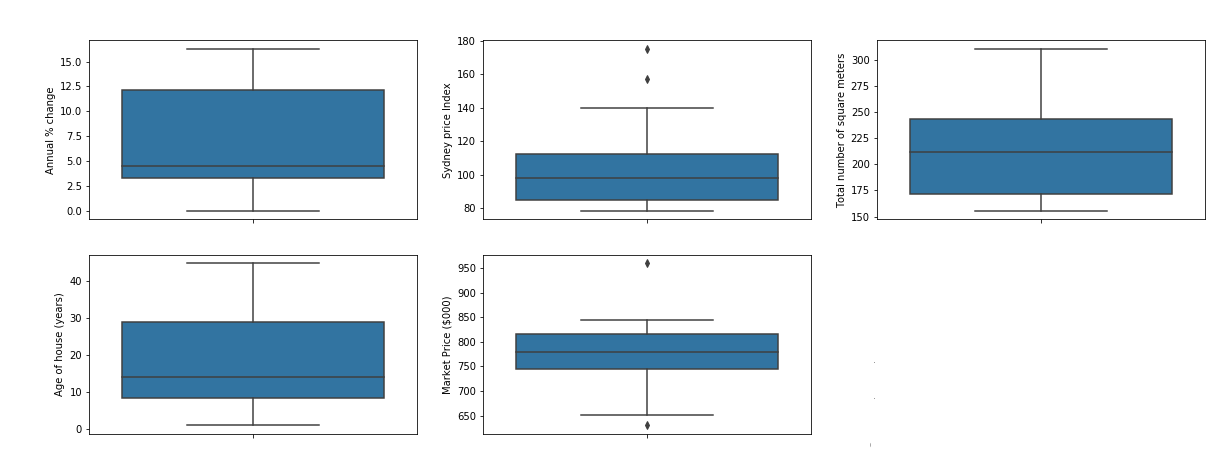
Upper range=maximum (value) + IQR\*1.5

Lower range=minimum (value) - IQR\*1.5

If Upper range< given value and Lower range> given value then this value is called as outliers.

**Identifying Outliers by visualization:**

We also find outliers in data by visualization method which is called as boxplot, so first we plot boxplot then we see what happends.you have seen in boxplot diagram which is in below,

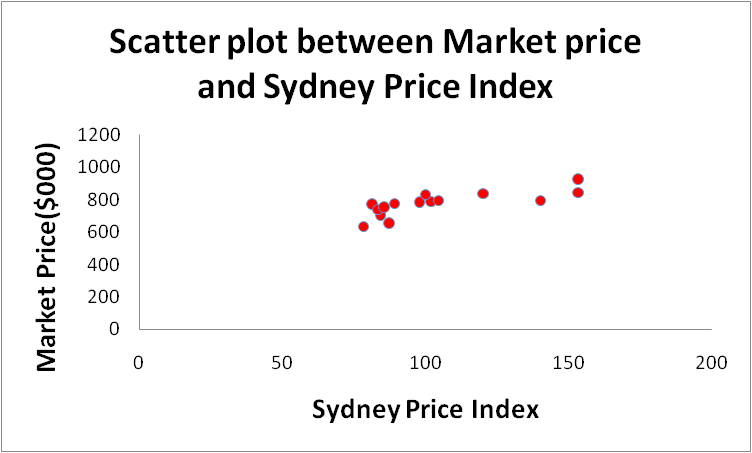


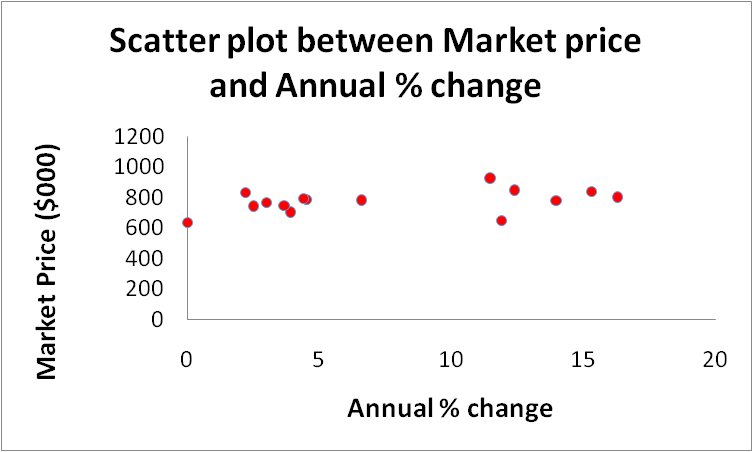
In this figure you have seen the Sydney price Index and Market price has outliers. So first we need to remove these outliers by data pre-processing.

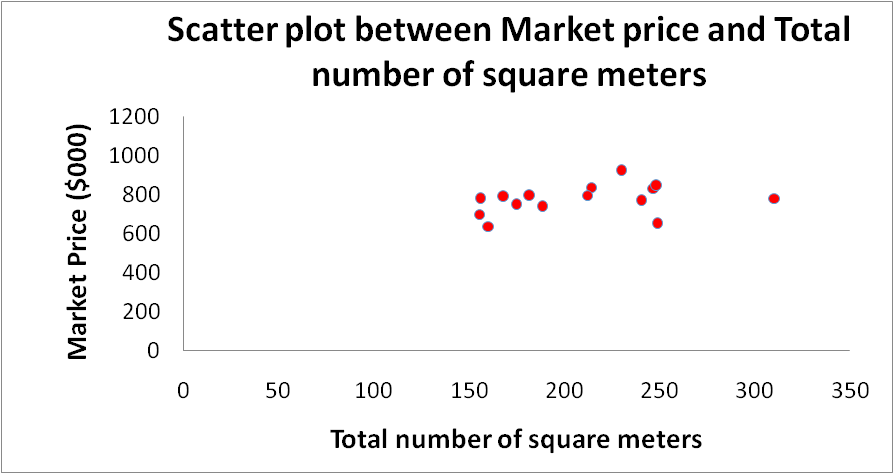
After adjusting outlier, we just plot a scatter plot to see the relationship between independent variable and dependent variable.

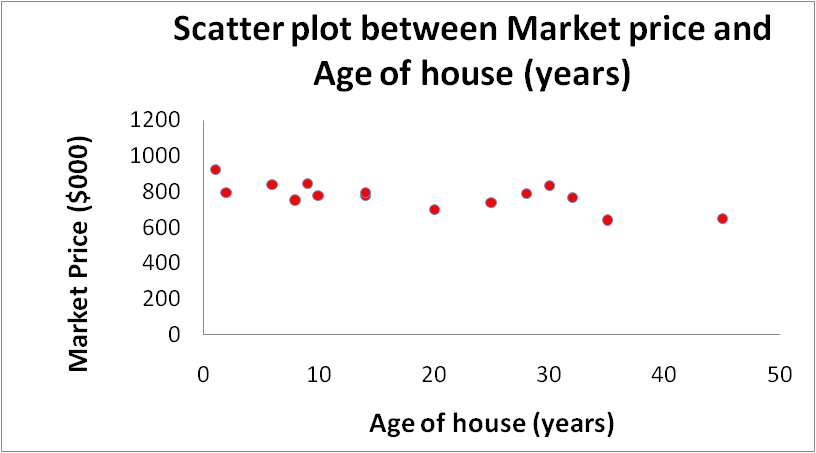
We plot a scatter plot to look which variable relationship with market price.

There are different scatter plot in below.









In the scatter plot we have seen Market price and the Sydney price Index have some relationship, Market price and Age of house also has some relationship, so we first find some statistics and correlation matrix.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Statistics** | **Sydney price index** | **Annual % change** | **Total number of square meters** | **Age of house (years)** | **Market Price ($000)** |
| **Minimum** | 78.20 | 0.00 | 155.30 | 1.00 | 636.75 |
| **Maximum** | 153.58 | 16.30 | 310.50 | 45.00 | 922.75 |
| **Average** | 104.12 | 7.48 | 209.11 | 18.60 | 774.97 |
| **Standard deviation** | 25.83 | 5.45 | 45.05 | 13.29 | 74.52 |
| **Q1** | 84.95 | 3.35 | 171.50 | 8.50 | 744.00 |
| **Q2/Median** | 97.80 | 4.50 | 212.10 | 14.00 | 780.00 |
| **Q3** | 112.40 | 12.15 | 243.70 | 29.00 | 815.50 |
| **IQR** | 27.45 | 8.80 | 72.20 | 20.50 | 71.50 |
| **Lower range** | 43.78 | -9.85 | 63.20 | -22.25 | 636.75 |
| **Upper range** | 153.58 | 25.35 | 352.00 | 59.75 | 922.75 |
| **Correlation** | 0.77 | 0.41 | 0.32 | -0.68 | 1.00 |
|  |  |  |  |  |  |

In the above table, we found the correlation between Market price and Sydney Price Index is a 0.77 means when the Sydney price index is increased Market price is also increased and correlation between Market price and Age of house is -0.68 means when age of the house increases Market price is decreased.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Sydney price Index** | **Annual % change** | **Total number of square meters** | **Age of house (years)** | **Market Price ($000)** |
| **Sydney price Index** | 1 |  |  |  |  |
| **Annual % change** | 0.684 | 1 |  |  |  |
| **Total number of square meters** | 0.323 | 0.547 | 1 |  |  |
| **Age of house (years)** | -0.581 | -0.403 | -0.0150 | 1 |  |
| **Market Price ($000)** | 0.771 | 0.406 | 0.317 | -0.682 | 1 |

In correlation matrix we have highlighted which variable is correlated with respected to other variables.

Let us fit regression model for see relationship and also predict the output. First, we learn about linear regression and how to implement in Python. Linear Regression is nothing but statistical model we can learn the linear relationship between two means one is independent variable and one is dependent variable or if we have more independent variable and one dependent variable you must used multiple regression. Regression is similar as correlation but difference is that if you want to predict dependent variable then regression is useful.

Linear formula for multiple linear regression is Yi = b0+ b1X1+b2X2+b3X3 +---+bnXn

In above equation, *Y* is the dependent variable, we are trying to predict or estimate. X is the independent variable, this variables are using to make predictions; b1,b2,----,bn is the slope of the regression line it represent the effect *X*  on *Y*. *b0*is a constant, also known as the Y-intercept.

In a regression model, we are trying to minimize mean squared error or sum of squared error.

X is dependent variable data and Y is independent variable data. Using the statsmodel python Api we fit ordinary least squares and find a summary of this model which is given in below table.

In a regression model, we are trying to minimize mean squared error or sum of squared error.

We can fit multiple regression model in python using two library which is namely statsmodel and scikit-learn.

In our multiple regression model we used statsmodel api, it is just python module using this model we can estimate of many different statistical models, as well as conducting statistical test and data exploration.

For python first you have to installed statsmodels using pip command then import library using command

Import statsmodel.api as sm

import pandas as pd

first, we load data in pandas dataframe using below command

data1=pd.read\_csv("data.csv", delimiter='\t')

data1.head(5)

Next we split this dataset into X and y where X is our independent variable and y is dependent variable using below command

X = data1[['Sydney price Index', 'Annual % change', 'Total number of square meters', 'Age of house (years)']]

y = data1['Market Price ($000)']

X.head(5)

y.head(5)

Let us fit model in python, commands are given in below

X = sm.add\_constant(X) # adding a constant

model = sm.OLS(y, X).fit()

predictions = model.predict(X)

print\_model = model.summary()

print(print\_model)

In the above we used OLS model which is stands for ordinary least squares and in this methods we’re trying to fit a regression line that would minimize the square of distance from the regression line this model will give in output Date and Time ,number of observations, Degree of freedom of residuals and models.

|  |  |  |  |
| --- | --- | --- | --- |
| **OLS Regression Results** | | | |
| **Dep. Variable:** | Market Price ($000) | **R-squared:** | 0.77 |
| **Model:** | OLS | **Adj. R-squared:** | 0.679 |
| **Method:** | Least Squares | **F-statistic:** | 8.388 |
| **No. Observations:** | 15 | **Prob (F-statistic):** | 0.0031 |
| **Df Residuals:** | 10 | **Log-Likelihood:** | -74.397 |
| **Df Model:** | 4 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025 0.975]** |
| **const** | 543.4979 | 85.606 | 6.349 | 0 | 352.755 734.240 |
| **Sydney price Index** | 2.0329 | 0.674 | 3.014 | 0.013 | 0.530 3.536 |
| **Annual % change** | -5.8192 | 3.25 | -1.79 | 0.104 | -13.061 1.423 |
| **Total number of square meters** | 0.5221 | 0.311 | 1.678 | 0.124 | -0.171 1.216 |
| **Age of house (years)** | -2.4655 | 1.082 | -2.28 | 0.046 | -4.875 -0.056 |

See in the above table, dependent variable is Market price, R-squared value is 0.77 is means 77% of the data fit the regression model.

In the above table you see P-value of the Sydney price index is less than 0.05 means we can say Sydney price index 95% Statistically significant with respect to Market price and similar P-value of age of the house is also less than 0.05 means this variable is also statistically significant with respect to Market price. Remaining two variables are not statistically significant.

If we want to predict Market price after fitting this model for the Sydney price Index=80, Annual % change =10, total no of squared meter=300, age of house=2yrs.This models give Market Price is $799.64

Note: In all this Analysis I used Excel, Python 2.7 kernel use in Jupyter Notebook.