# REPORT

The model which I liked the most is Linear Regression. Because of it gave better prediction in comparison to K-Nearest Neighbor algorithm and Random Forest.

## Basic description of how Linear regression works:

1.It's designed for more statistically-oriented approaches to data analysis, with an emphasis on econometric analyses.

2. It integrates well with the pandas and numpy libraries.

3. It also has built in support for many of the statistical tests to check the quality of the fit and a dedicated set of plotting functions to visualize and diagnose the fit.

4. The best part is that, we use multiple packages to design a linear regression model.

## Evaluation of how it works:

1. In more technical way, we are able to predict whether the quantities on x-axis and y-axis are positively or negatively correlated.

2. If the slope of the plot is positive. Then, we can say that the variables are positively correlated. And vice-versa.

3. Statistically speaking, more the value of R-square, and it's respective Adjusted R-square; better is the correlation between these variables.

4. Moreover, I tried using two different libraries for plotting the linear regression model. First library is statsmodel.formula.api

and the second library is sklearn.linear_model. In both the models, the accuracy was more or less the same.

### Mathematical principles behind Linear Regression model:

1. The model forms a best fit line, considering all the data points.

2. The procedure for plotting the best fit line is as follows:

- Initially few random points are considered.

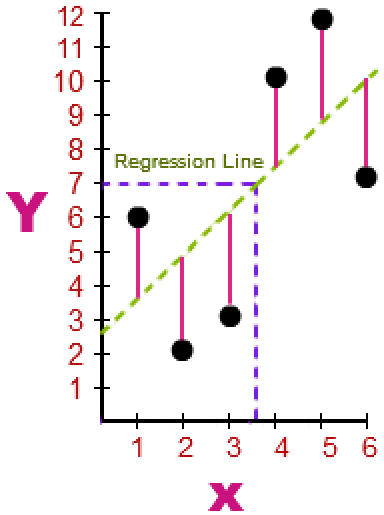
- The perpendicular distance from the data point to the line is calculated.

- Similar process is repeated for all the points.

- And we know, that we can draw infinite number of lines in a plane.

- Hence, the line which is closed to all the points, is known as best fit line.

- Visually, the line which has minimum perpendicular distance from all the points, is the best fit line.



3. Formula wise:  SSE=∑ni=1(xi−xi^)2

- We calculate the sum of squares of the perpendicular distance from the data point to the line.

- After comparing between all the lines, the line with least sum of square of error is considered as the best fit line.

## Interesting experiences or surprises:

1. When I performed two Linear regression models. One between, 'basementsqft' => First; while another between 'logerror' and 'numberofstories + basementsqft' => Second.

2. Considering, the Adjusted R-square value of 'First' plot as X and mean-square value as Y. And Adjusted R-square value of 'First' plot as A and mean-square value as B.

3. I see that, the X>A;  and even, Y<B.

4. Considering the above situation. I think that the model is over fitting 'Second' time.

5. Also, practically speaking I don't see any true relation in the 'Second'.