* **Problem Statement:**

Here we are interested in understanding the relationship between area (Square Meter ) of House (independent variable) and their price (dependent variable). The goal is to build a simple linear regression model to predict House price based on the area of house.

* **Data - Set :**

<kaggle datasets download -d muhammadbinimran/housing-price-prediction-data>

* **R-Code for Simple regression Analysis :**

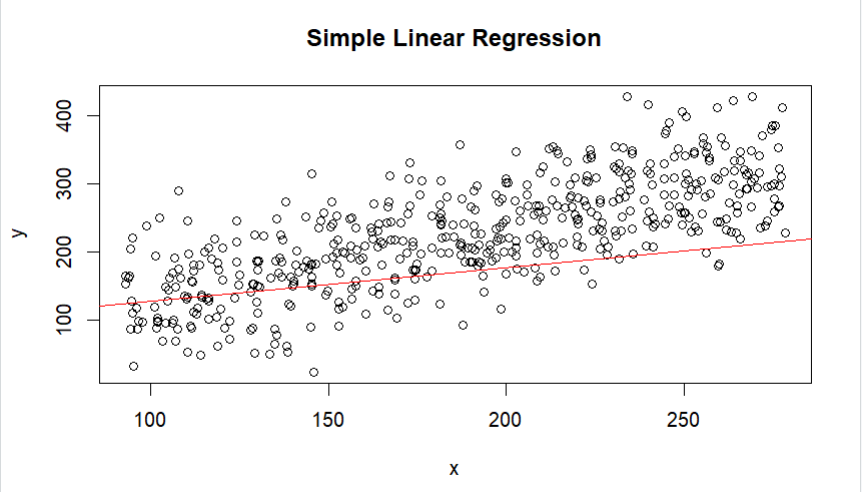
data=read.csv(file.choose(),header = T)

model=lm(SquareMeter ~ PriceK, data = data)

summary(model)

plot(data$SquareMeter, data$PriceK, main = "Simple Linear Regression", xlab = "x", ylab = "y")

abline(model,col="red")



* **Model Description and Interpretation:**

The model equation would be something like:

House price = *β*0 + *β*1 × House area + Ɛ

*β*0 is the Intercept.

*β*1 is the Slope Coefficient.

Ɛ is the error

* **Interpretation:**

The intercept represents the expected House price when House area are zero, and the slope represents the change in House price for a one-unit increase in House Area.

* **Model Adequacy Test:**
* **Checking correlation**

**Strength of Correlation**:

r is to 1 or -1, the stronger the correlation.

If∣r∣<0.3, the correlation is considered weak.

If 0.3≤∣r∣<0.7, the correlation is considered moderate.

If∣r∣≥0.7, the correlation is considered strong.

**R** **Code for correlation**

cor(data$SquareMeter, data$PriceK)

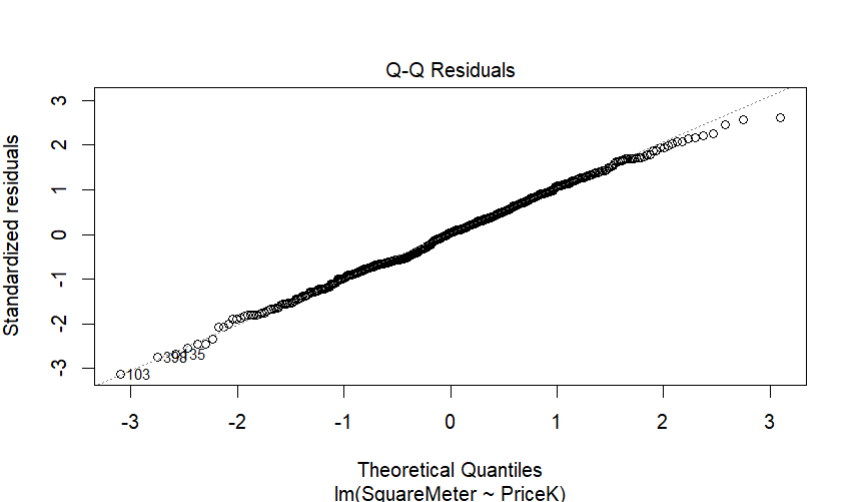
Here , r is **0.7331684.**Hence we can conclude that the house area and house price has strong correlation.

* **Normality(Q-Q Residual plot):**

The points on the QQ Residual plot fall approximately

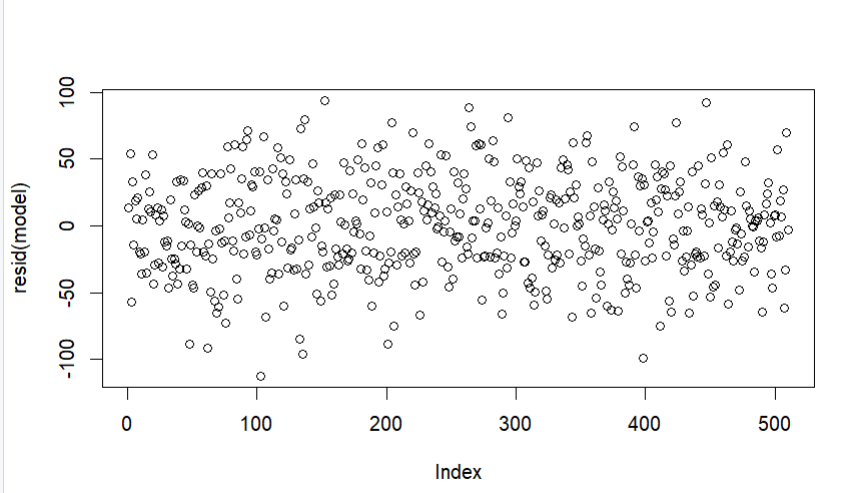
along the straight line, it suggests that the residuals are normally

distributed.



* **Residual Plot**

Check for homocsedasticity and normality of residuals.



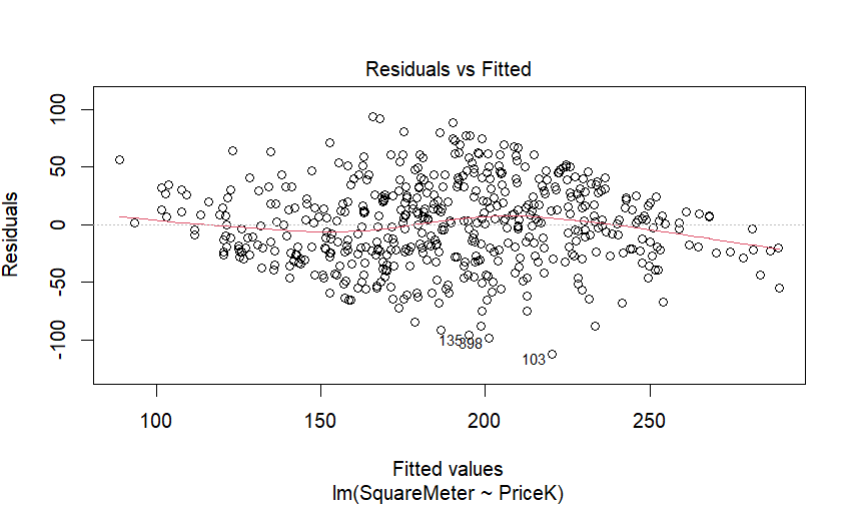
* **Residual vs Fitted Plot**

A residual vs. fitted plot helps assess whether the chosen regression model adequately captures the underlying patterns in the

data. By plotting the residuals (the differences between observed and predicted values) against the fitted values (the predicted values from the regression model), you can visually examine whether the model captures the variation in the data appropriately. Ideally, the residuals should exhibit a random scatter around the horizontal

line at 0. If there are systematic patterns (e.g., curves, clusters, or

heteroscedasticity), it suggests that the model may not be suitable for the data.



The residuals exhibit a random scatter around the horizontal line at 0. There are no systematic patterns. Therefore, it suggests that the model may not be suitable for the data.

* **Summary**

Residual standard error: 36.05 on 508 degrees of freedom

Multiple R-squared: 0.5375,

Adjusted R-squared: 0.5366

1. statistic: 590.5 on 1 and 508 DF,

p-value: < 2.2e-16

The residual standard error of **36.05 i**ndicates the typical distance between observed and predicted values in your regression model. The multiple R-squared value of **0.5375** suggests that approximately **53.75%** of the variability in the dependent variable is explained by the independent variable(s). The adjusted R-squared, which accounts for the number of predictors, is **0.5366** indicates a slight

decrease in explanatory power compared to Multiple R-squared

**Hypothesis testing**

The high F-statistic of **590.5** with a p-value < **2.2e-16** suggests

that the overall model is **statistically significant**. These results collectively

indicate a reasonably **well-fitting model** with a **significant relationship**

between the **House price**( dependent variable ) and **House area** ( independent variable) **,** providing a strong justification for the use of this regression

model in explaining the variation in the House price ( dependent variable)