Real Estate Prediction Model

Submitted by:-
Simranh Kaur Bhasin

THE DATASET

The dataset on which our prediction model has been applied contains information of or relating to real estate. It includes the date of purchase, house age, location, distance to nearest MRT station, and house price of unit area that constitute 8 columns and 413 rows. It was found out that the dataset is clean and tidy as it did not contain any missing values or inconsistent data. A sample of the data can be seen below:-

	A	B	C	D	E	F	G	H
1	No	transaction dat	house age	distance to the nearest MRT s	number of convenience sto	latitude	longitude	house price of unit area
2	1	2012.917	32	84.87882	10	24.98298	121.54024	37.9
3	2	2012.917	19.5	306.5947	9	24.98034	121.53951	42.2
4	3	2013.583	13.3	561.9845	5	24.98746	121.54391	47.3
5	4	2013.5	13.3	561.9845	5	24.98746	121.54391	54.8
6	5	2012.833	5	390.5684	5	24.97937	121.54245	43.1
7	6	2012.667	7.1	2175.03	3	24.96305	121.51254	32.1
8	7	2012.667	34.5	623.4731	7	24.97933	121.53642	40.3
9	8	2013.417	20.3	287.6025	6	24.98042	121.54228	46.7
10	9	2013.5	31.7	5512.038	1	24.95095	121.48458	18.8
11	10	2013.417	17.9	1783.18	3	24.96731	121.51486	22.1

SOURCES OF THE DATASET

- **The dataset was found on https://www.kaggle.com/.**
- **The complete dataset can be found on the Github link:-**

  https://github.com/simranhbhasin/data-science-project

RESEARCH PROBLEM

- The dataset that was chosen sparked some inquisitiveness in me. I was curious to know how does this data account for the prices of properties.

- I was of the belief that the physical features of a house has the maximum weightage in determining the price of the house. However the data that was selected presented a contradiction. So on some research I understood that a property's physical structure tends to depreciate over time, while the land it sits on typically appreciates in value.

- Moreover, Not all spots within a given area are considered equal. A home by a calm street is usually in higher demand than a home situated near a busy roadway.

- It was found out that these prices also depend on the centrality of a property within a city. If it is surrounded by a well functioning neighborhood that has a good network of roads and transportation facilities it is rated higher. The geography also plays a huge role in this aspect.

- Thus based on these findings I built my research problem. It states that:

> This prediction model and the research around it is done in an attempt to determine exactly which factors play the most important role in deciding the price of a house. It will also be ascertained whether we can accurately estimate and predict this value if we perform our entire research on these few determinants or other factors need to be considered as well.

PREDICTION MODEL
(LINEAR REGRESSION)

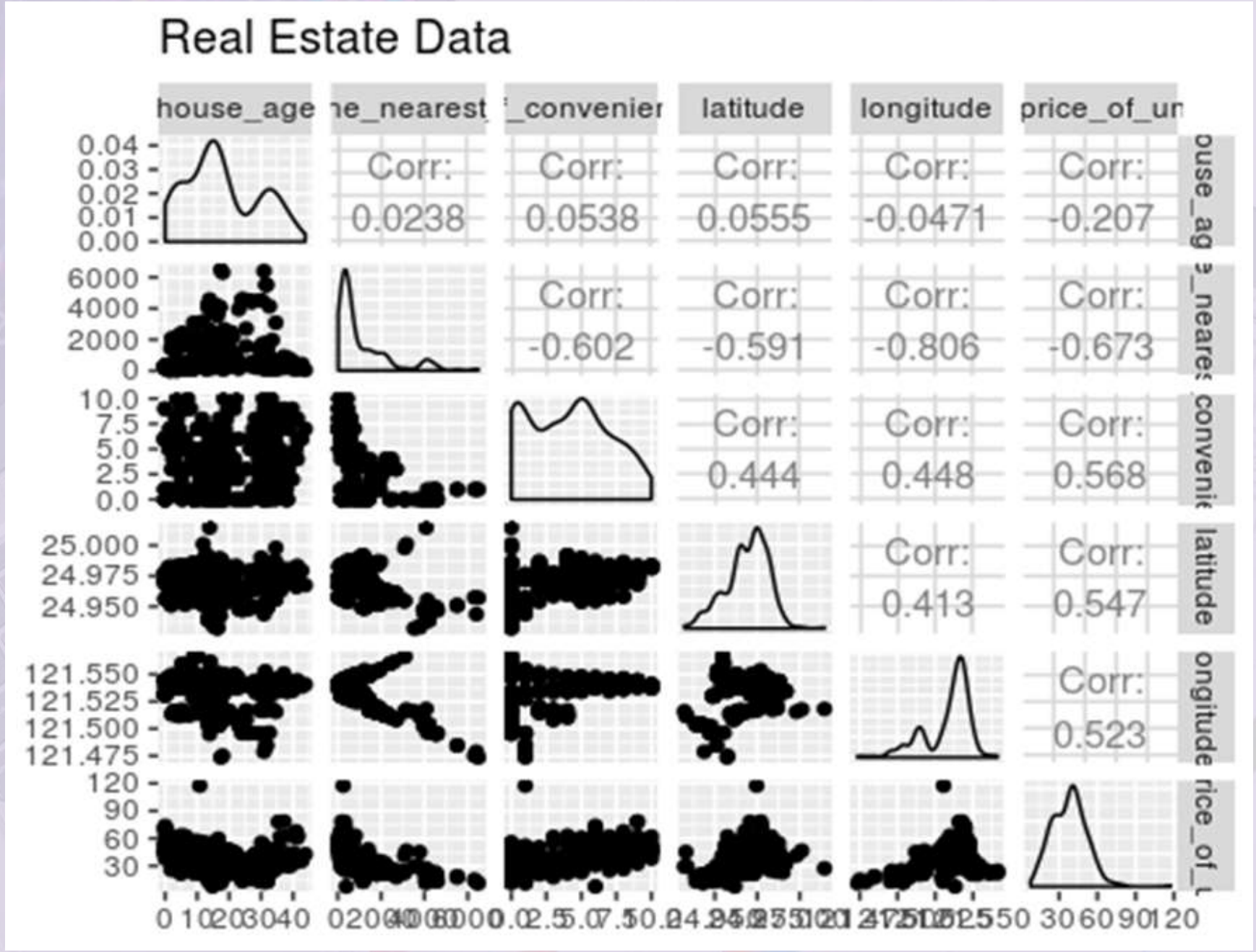
- **Predictive models are extremely useful for forecasting future outcomes and estimating metrics. The model used here is known as linear regression. Linear regression is used to predict the value of an outcome variable *Y* based on one or more input predictor variables *X*. The aim is to establish a linear relationship (a mathematical formula) between the predictor variable(s) and the response variable, so that, we can use this formula to estimate the value of the response *Y*, when only the predictors' (*Xs*) values are known.**

- **Linear regression has been chosen because the selected dataset doesn't contain any categorical data.**

HOW TO IMPLEMENT THE PREDICTION MODEL?

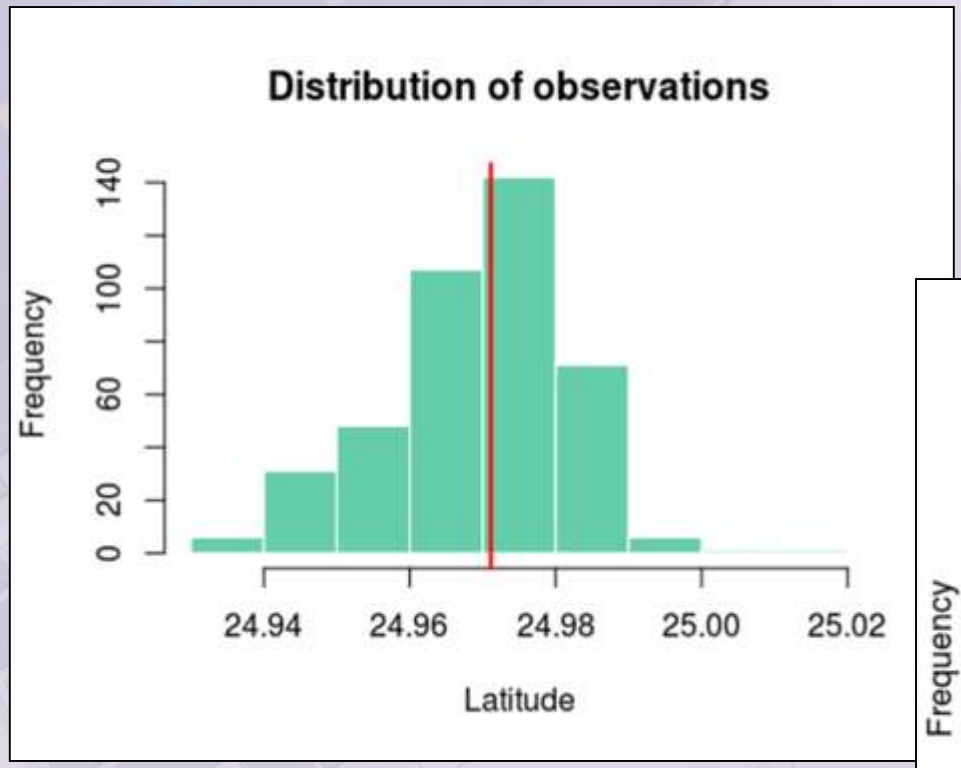
- As stated by the research problem it is known that the price of the house needs to be estimated and therefore it will serve as the dependent variable, *'Y'.*

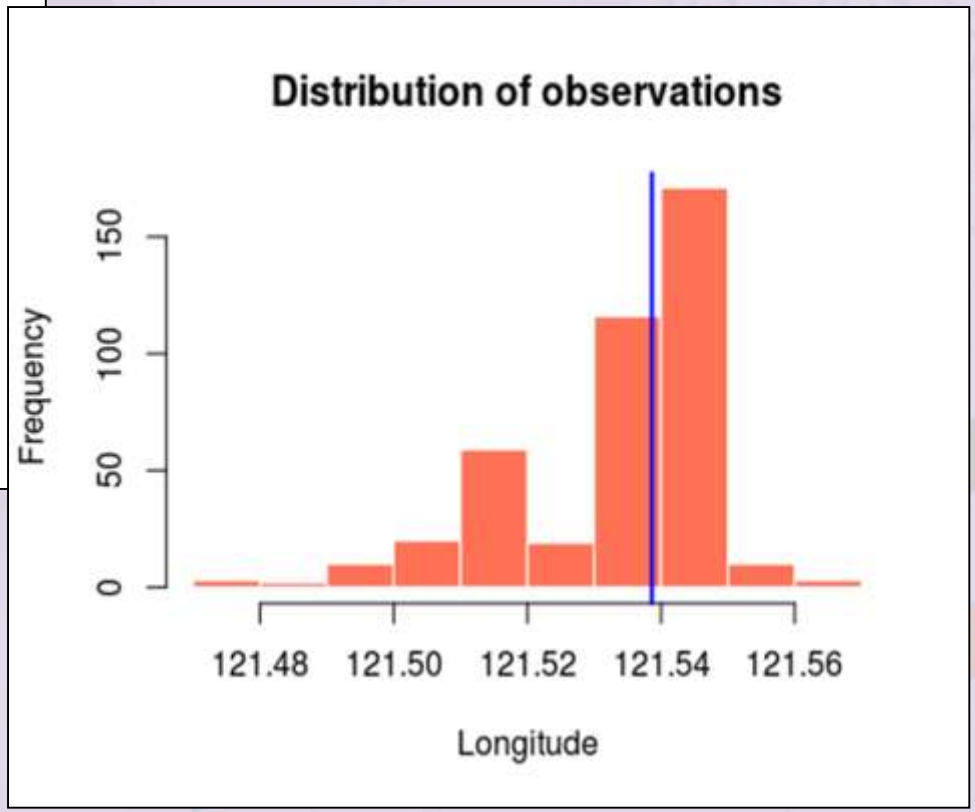
- Thus, now we are required to decide which attributes will serve as the predictor variables. In order to do that we will be making a plot using ggpairs() present in the Ggally package. The plot will help us visualize the dataset, depict the relationship between the different attributes present in it and execute exploratory data analysis.

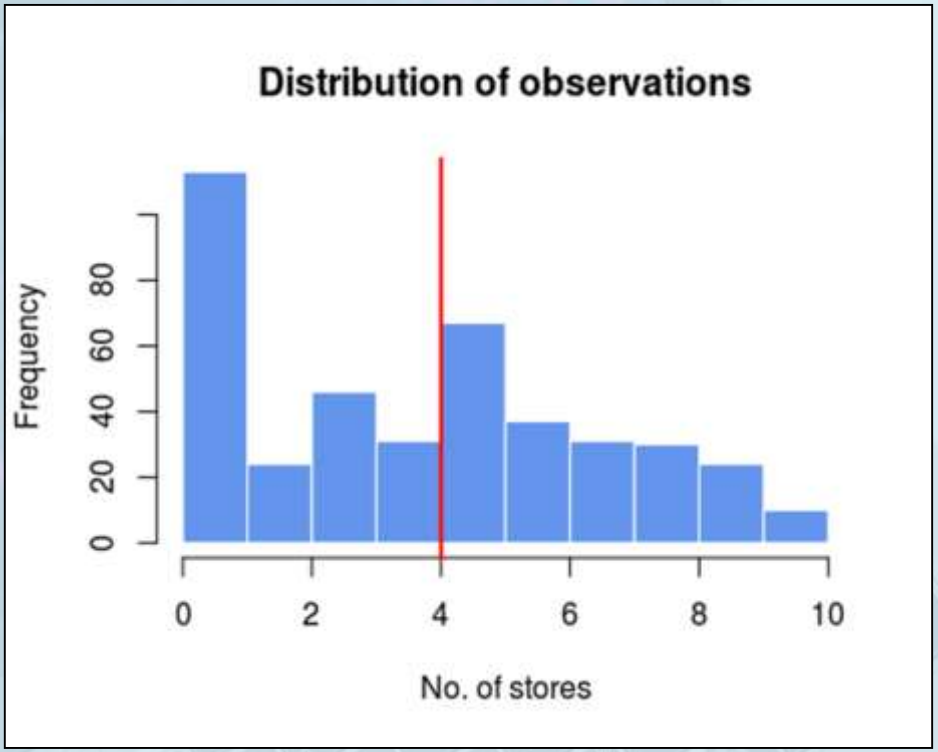

Real Estate Data
house_age
latitude
longitude
price_of_un
Corr:
0.0238
Corr:
0.0538
Corr:
0.0555
Corr:
-0.0471
Corr:
-0.207
Corr:
-0.602
Corr:
-0.591
Corr:
-0.806
Corr:
-0.673
Corr:
0.444
Corr:
0.448
Corr:
0.568
Corr:
0.413
Corr:
0.547
Corr:
0.523

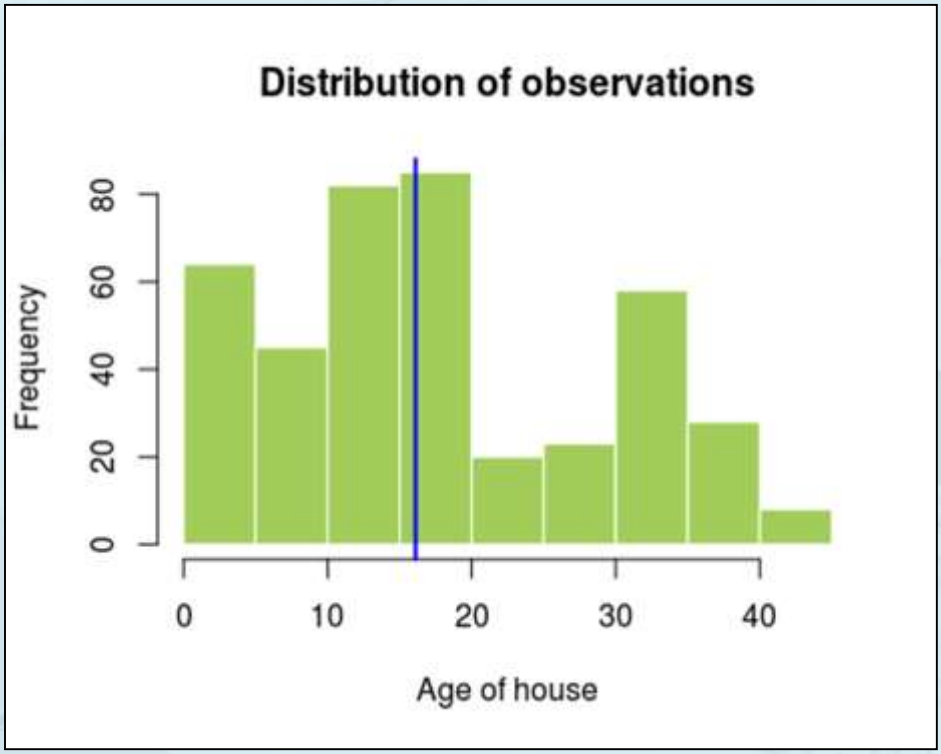
- **The plot matrix gives us scatter plots for each variable combination, as well as density plots for each variable and the strength of correlations between variables. The correlation coefficients provide information about how close the variables are to having a relationship; the closer the correlation coefficient is to 1, the stronger the relationship is. The scatter plots let us visualize the relationships between pairs of variables.**
- **Thus the top 4 variables that seem to have a strong relationship with our dependent variable i.e. house per unit area are:-**
  - **Number of convenience stores**
  - **Latitude**
  - **Longitude**
  - **House Age**
- **These variables will be the *'Xs'* according to our definition of linear regression and will be used to predict the price per unit area of a house.**

- To further carry out data analysis and judge the data distribution of these four attributes that will work as the predictor variables we will plot histograms. To visualize, the line for the median of these observations is also plotted.







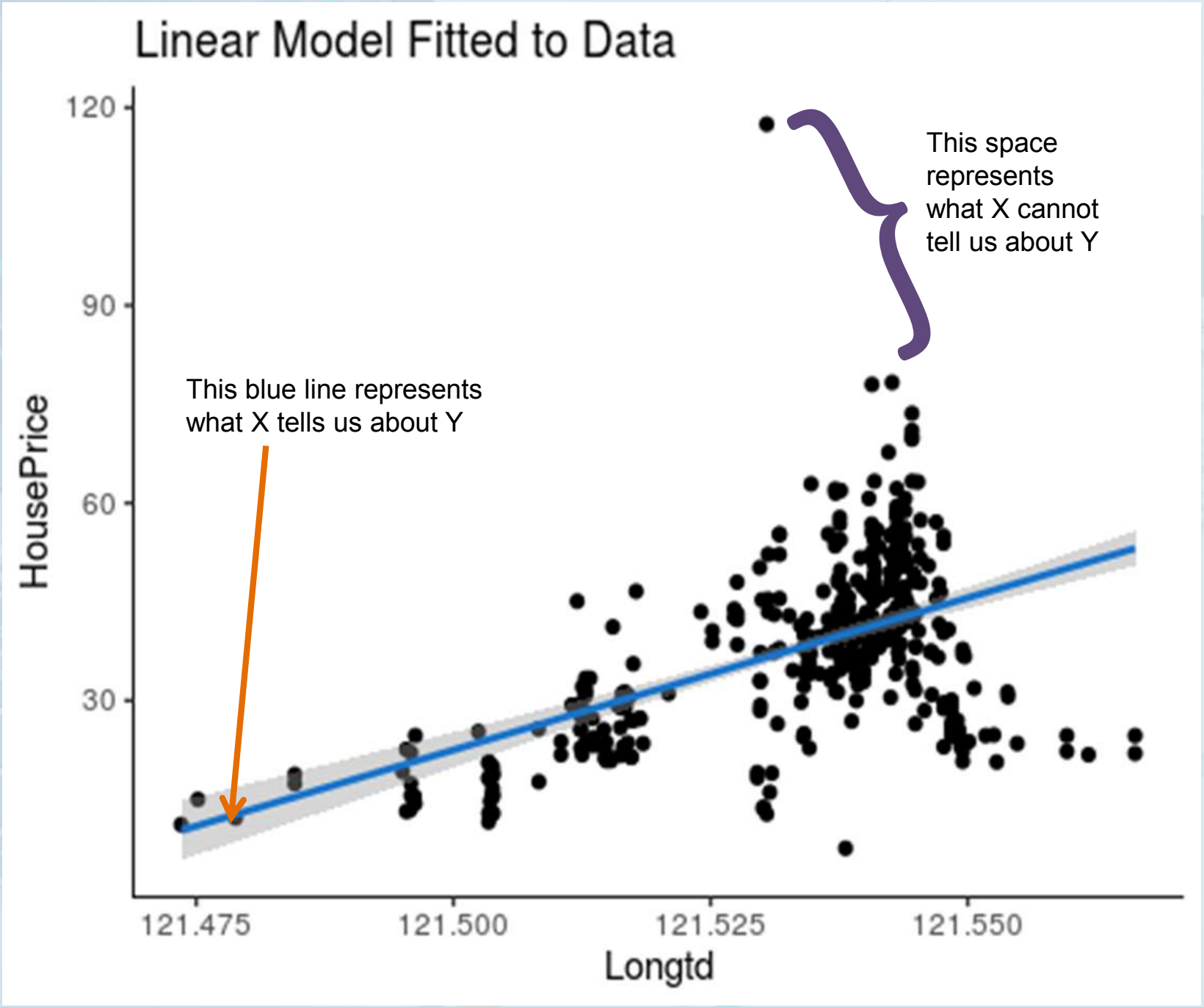


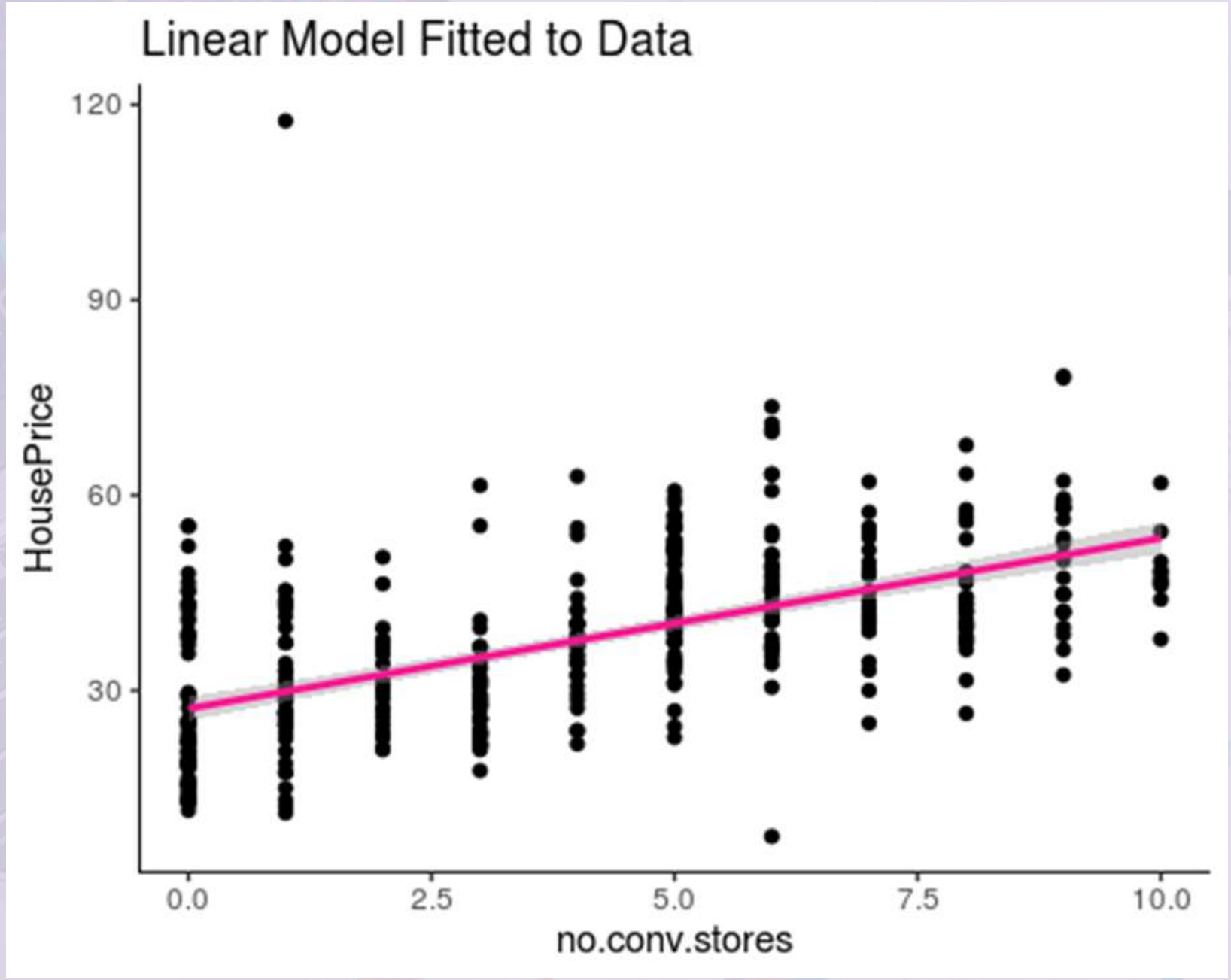
- **The height of the histogram indicates the number of observations in that interval. These distribution plots give us a point to point *univariate analysis* of the raw data.**

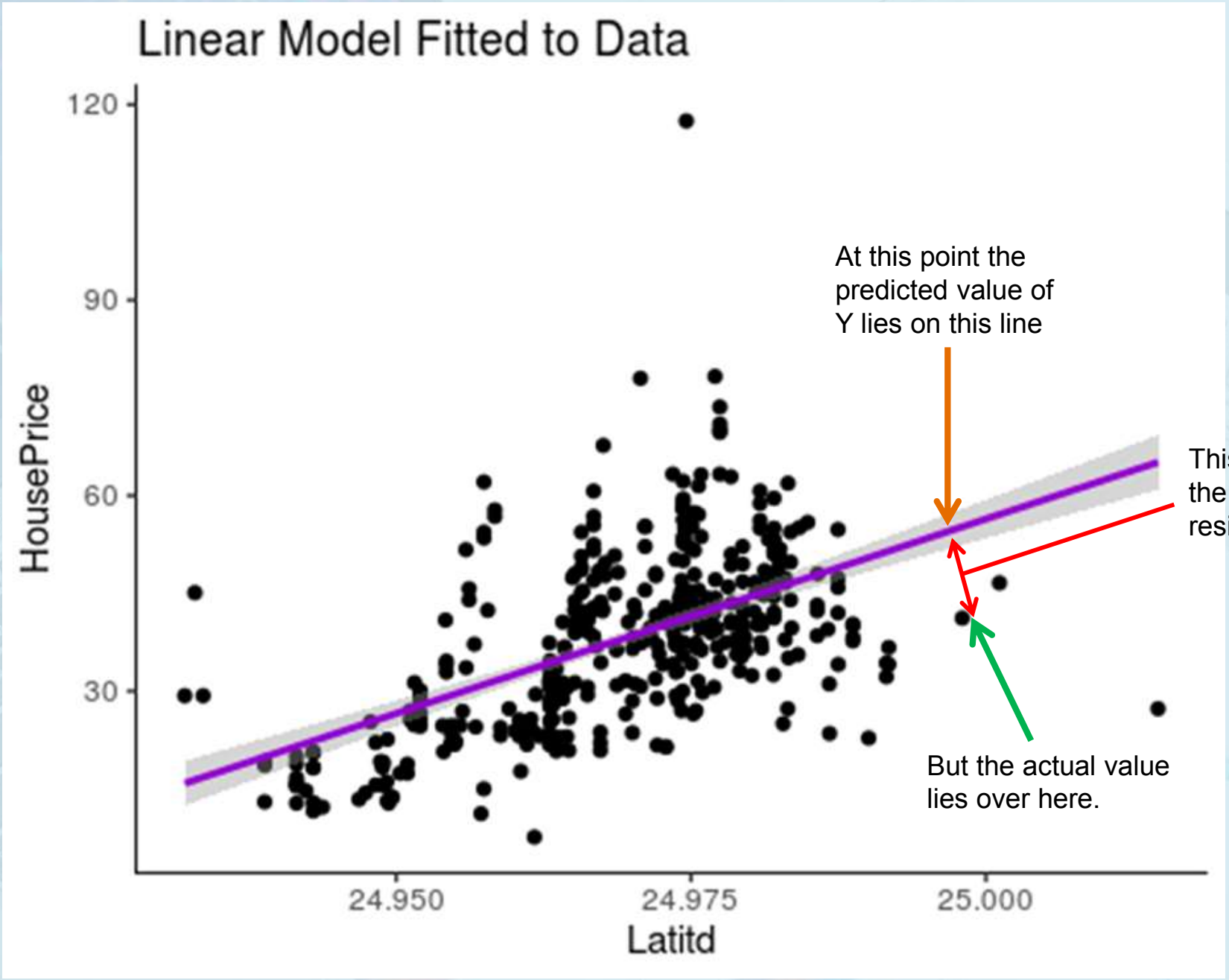
VISUALIZING THE LINEAR REGRESSION MODEL

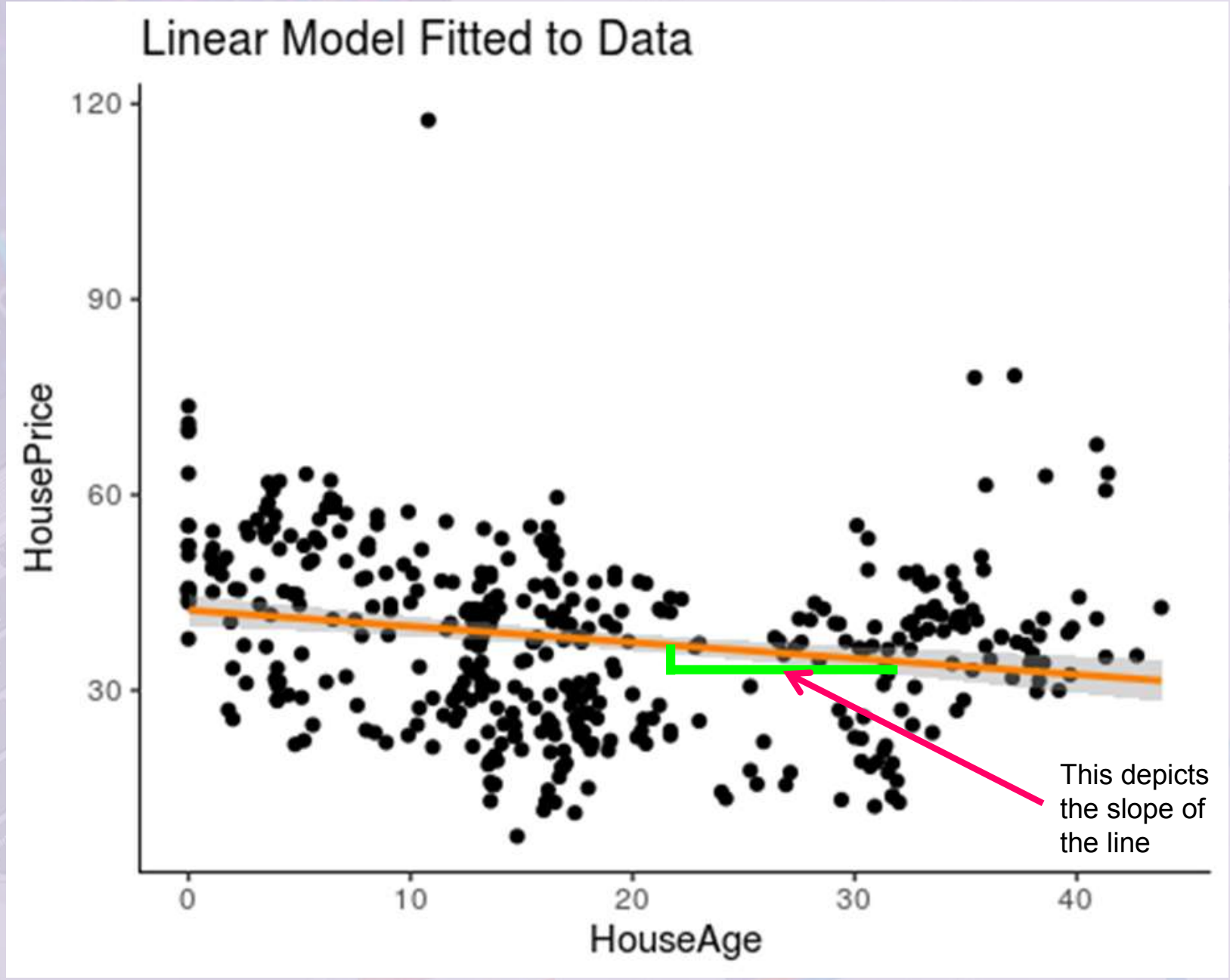
- The lm() function fits a line to our data that is as close as possible to all of our observations. More specifically, it fits the line in such a way that the sum of the squared difference between the points and the line is minimized; this method is known as “minimizing least squares.” Even when a linear regression model fits data very well, the fit isn’t perfect. The distances between our observations and their model-predicted value are called residuals.

- When the regression line was fit to the scatterplot of our data using ggplot(), we obtained the graphs:


Linear Model Fitted to Data
120
90
60
30
HousePrice
121.475
121.500
121.525
121.550
Longtd
This space represents what X cannot tell us about Y
This blue line represents what X tells us about Y


Linear Model Fitted to Data
120
90
60
30
HousePrice
0.0
2.5
5.0
7.5
10.0
no.conv.stores


Linear Model Fitted to Data
HousePrice
120
90
60
30
24.950
24.975
25.000
Latitd
At this point the predicted value of Y lies on this line
This is the residual
But the actual value lies over here.


Linear Model Fitted to Data
120
90
60
30
HousePrice
0
10
20
30
40
HouseAge
This depicts the slope of the line

- Mathematically we can interpret linear regression for our dataset as:

House Price/unit area ≈ Intercept + Slope(Latitude Longitude No. of convenience stores House Age) + Error

- But it can be noticed from the graphs that the regression line is unable to account for many observations and gives high residuals. We will need to fit the regression model more accurately so as to attain the targeted values of the response variable.

- To do this we will be applying polynomial regression. Polynomial Regression is a form of linear regression in which the relationship between the independent variable *X* and dependent variable *Y* is modeled as an nth degree polynomial. Polynomial regression fits a nonlinear relationship between the value of *X* and the corresponding conditional mean of *Y.*

```
trainIndex=createDataPartition(rs_data$No, p=0.7,list=FALSE)

training=rs_data[trainIndex,]
testing=rs_data[-trainIndex,]
```

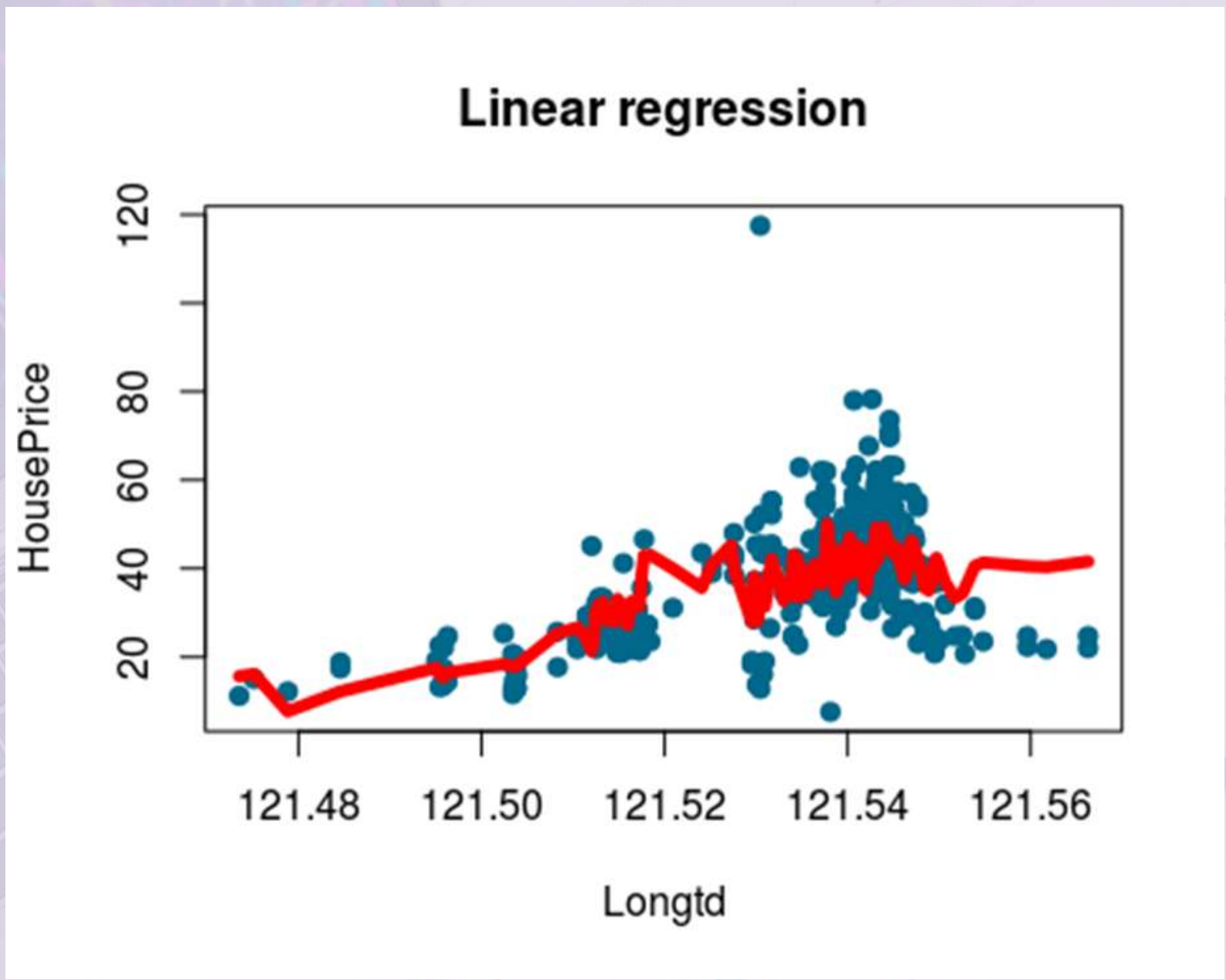
- We will also need to split our data into 2 sets namely testing and training. The training dataset will contain 70% of the original data and it will be modeled. The testing dataset will contain 30% of the data and it will be used to cross validate the observations generated by the predictions made by the implementation of the model.

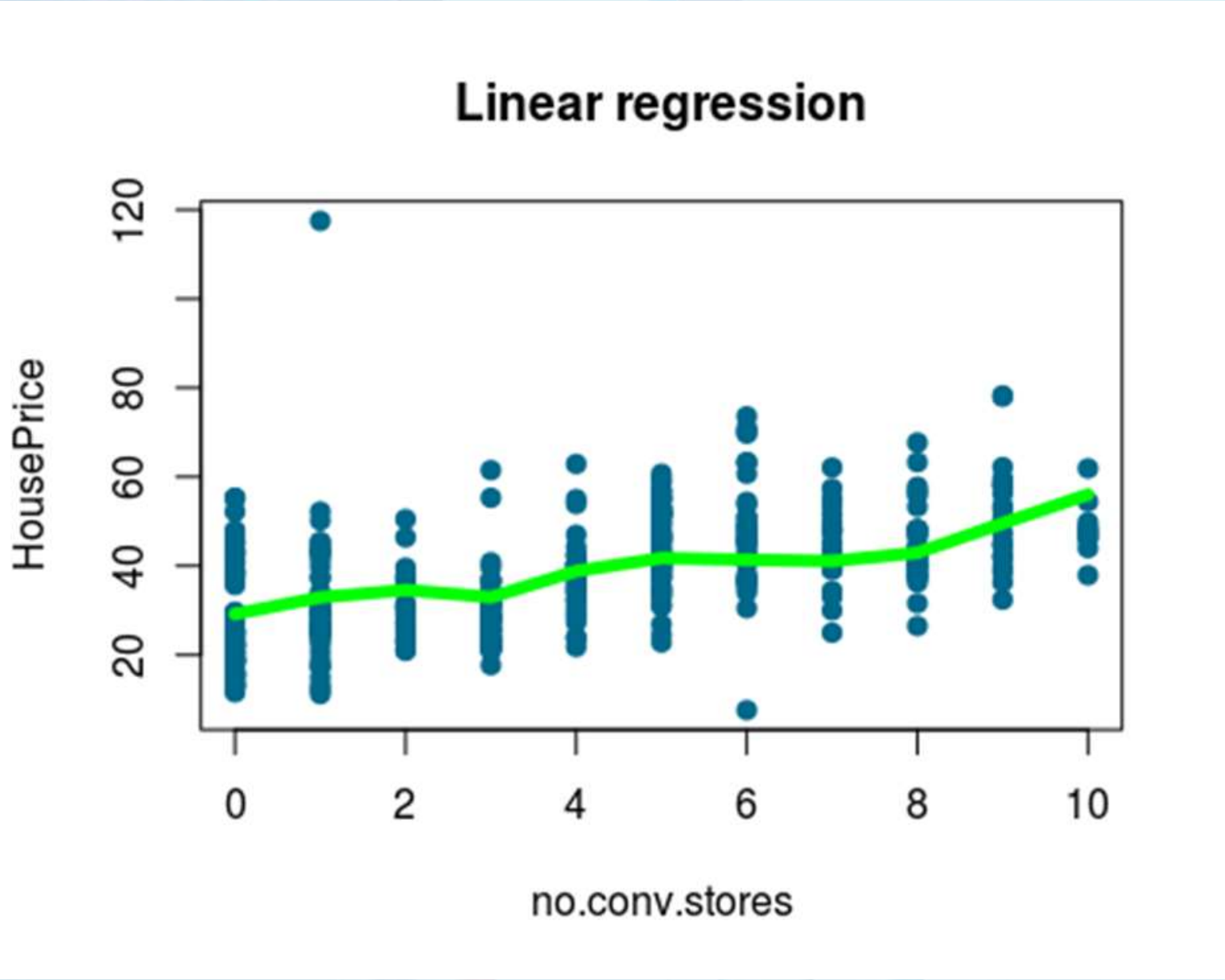
- We will hypothesize the relationship between the predictor and response variables to be curvilinear and thus we will include some polynomial terms to our previously made linear hypothesis. The tests that will be conducted thereafter will prove this hypothesis.

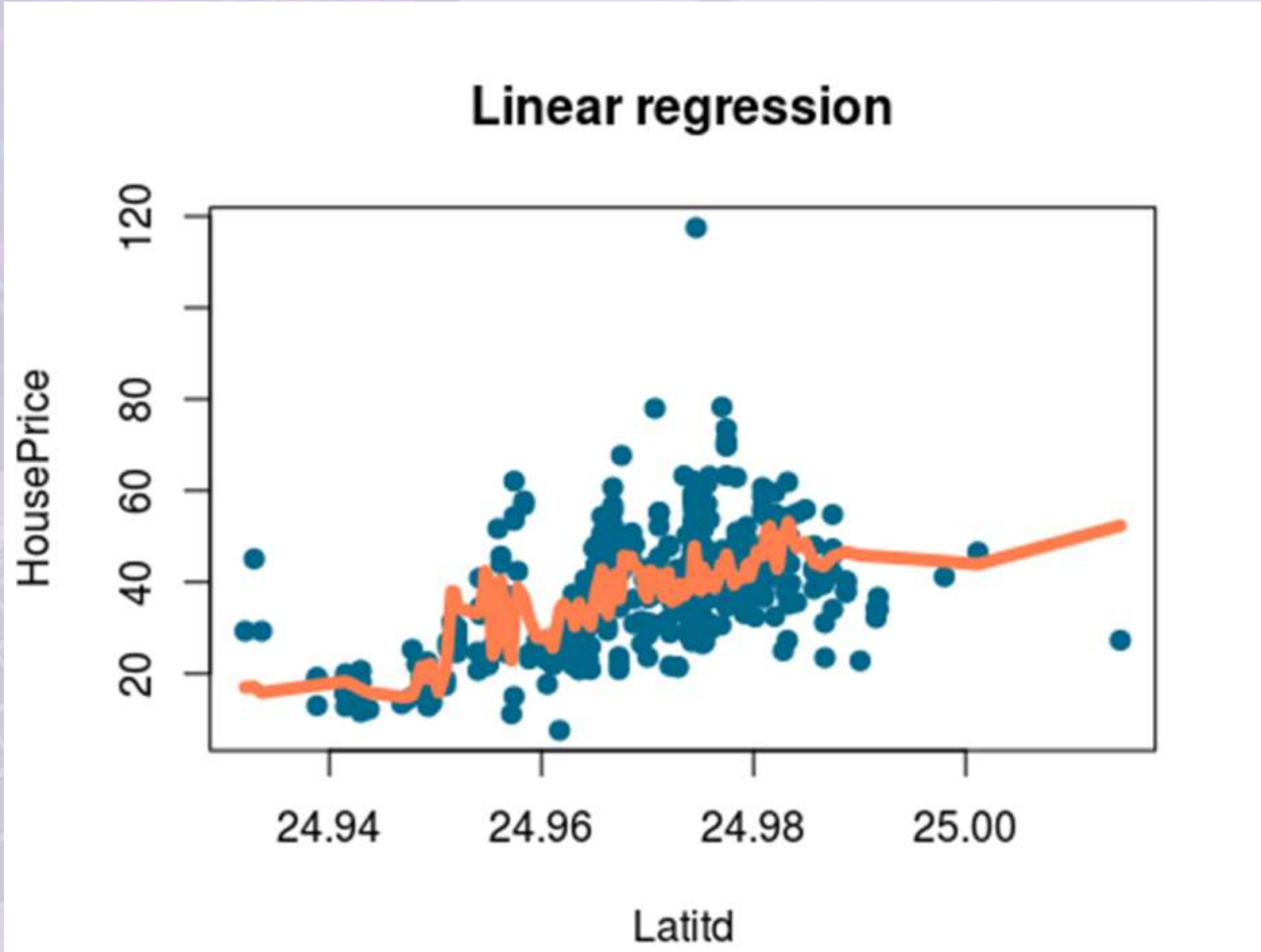
- So now the regression model can be implemented as:

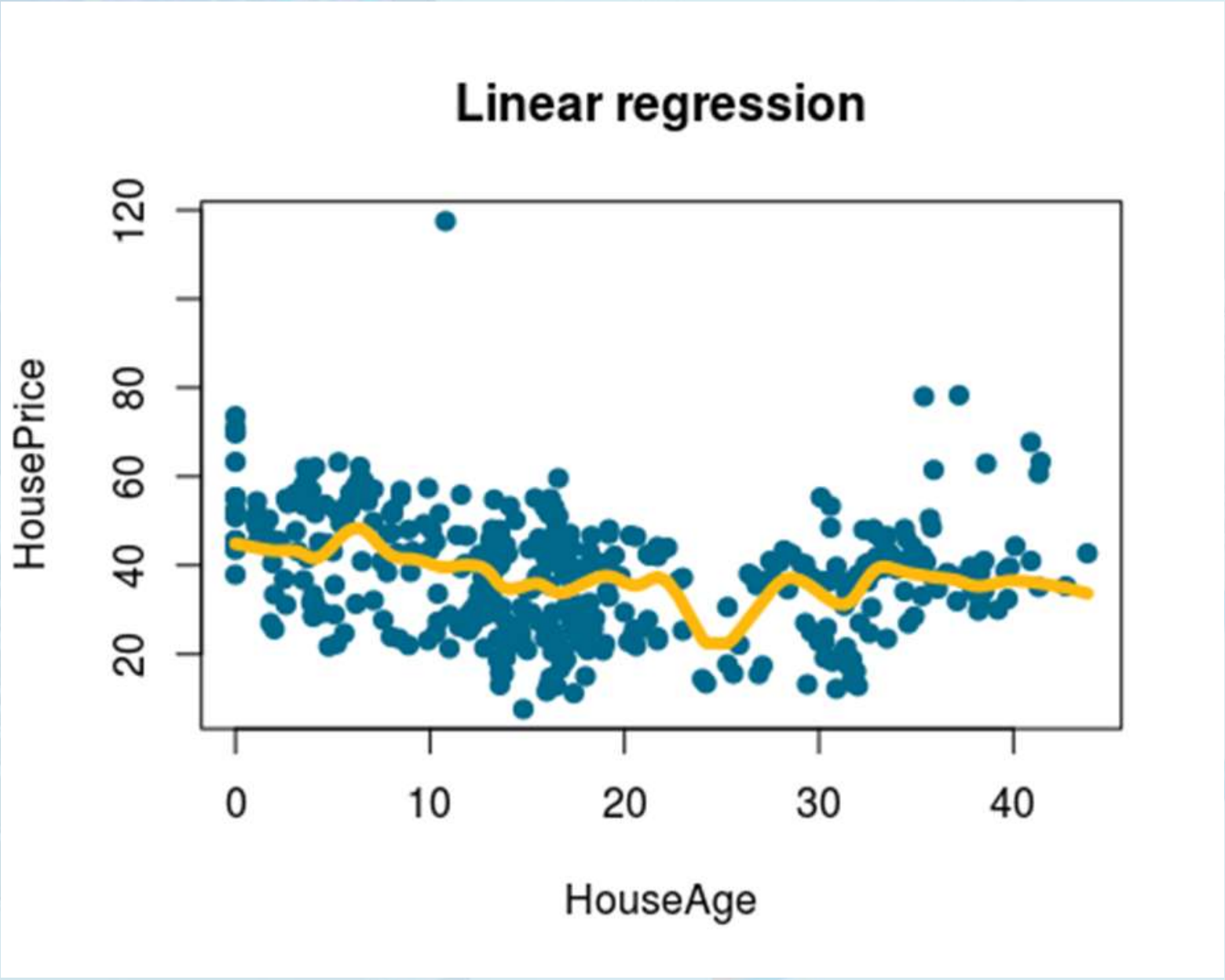
```
linearMod=lm(HousePrice ~ I(Longtd^2) +
I(Latitd^3) + I(no.conv.stores^5) + HouseAge,
data=training)
```

- The degrees are decided based on the correlations and to avoid under-fitting or over-fitting as far as possible. Now, the graph plots will depict how accurately the regression model fits our data


Linear regression
HousePrice
120
80
60
40
20
121.48
121.50
121.52
121.54
121.56
Longtd


Linear regression
HousePrice
20
40
60
80
120
0
2
4
6
8
10
no.conv.stores


Linear regression
HousePrice
120
80
60
40
20
24.94
24.96
24.98
25.00
Latitd


Linear regression
HousePrice
120
80
60
40
20
0
10
20
30
40
HouseAge

ANALYSIS OF THE REGRESSION MODEL

- The visual interpretation of our regression model does depict a better fit now. We can further affirm this conclusion by elucidating the results of the summary of the regression model and by applying the p-test and t-test to it.

```
summary(linearMod)
##
## Call:
## lm(formula = HousePrice ~ I(Longtd^2) + I(Latitd^3) +
I(no.conv.stores^5) +
##     HouseAge, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.007  -5.805  -0.681   4.961  77.584
```

```
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1.967e+04  2.017e+03  -9.752  < 2e-16 ***
## I(Longtd^2)            1.103e+00  1.437e-01   7.675 1.23e-13 ***
## I(Latitd^3)            2.198e-01  2.320e-02   9.473  < 2e-16 ***
## I(no.conv.stores^5)    1.336e-04  2.442e-05   5.473 7.73e-08 ***
## HouseAge              -2.769e-01  4.252e-02  -6.511 2.20e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

- This portion of the summary tells us if the hypothesis is supported or not.
  - The intercept is the expected house price if the value of the predictor variables was zero.
  - The coefficient standard errors tell us the average variation of the estimated coefficients from the actual average of our response variable.

❖ Significance:

- Test of significance shows that there is a strong evidence of a polynomial relationship between the predictor and response variables. This is visually interpreted by the significance stars (***) at the end of the row. This level is a threshold that is used to indicate real findings, and not the ones by chance alone. For each estimated regression coefficient, the variable's *p-Value Pr(>|t|)* provides an estimate of the probability that the true coefficient is zero given the value of the estimate. More the number of stars near the p-Value are, more *significant* is the variable. With the presence of the p-value, there is a test of hypothesis associated with it. In Linear Regression, the Null Hypothesis is that the coefficient associated with the variables is equal to zero. Instead, the *alternative hypothesis* is that the coefficient is not equal to zero and there exists a relationship between the independent variable and the dependent variable.

- So, if p-values are less than the significance level (ideally p-value < 0.05), null hypothesis can be safely rejected. In the current case, p-values are well below the 0.05 threshold, so the model is statistically significant.

- **T-statistic**

It is the measure of likelihood that the actual value of the parameter is not zero. A larger t-value indicates that it is less likely that the coefficient is not equal to zero purely by chance. The t-statistic value should be greater than 1.96 and we see that the variables with high correlation values do show a high t-static value.

❖ **We have therefore proved the statistical significance of the model. Now we will determine how well our model fits our data.**

```
## Residual standard error: 9.746 on 408 degrees of
freedom
## Multiple R-squared:  0.4886, Adjusted R-squared:  0.4836
## F-statistic: 97.46 on 4 and 408 DF,  p-value: < 2.2e-16
```

❖ R-squared

For simple linear regression, R-squared is the square of the correlation between two variables. Its value can vary between 0 and 1: a value close to 0 means that the regression model does not explain the variance in the response variable, while a number close to 1 indicates that the observed variance in the response variable is well explained. In the current case, R-squared suggests the linear model fit explains approximately 50% of the variance observed in the data.

High value of R-squared does not necessarily indicate if a regression model provides an adequate fit to data. A good model could show a low R-squared value, while, on the other hand, a biased model could have a high R-squared value.

❖ F-statistics

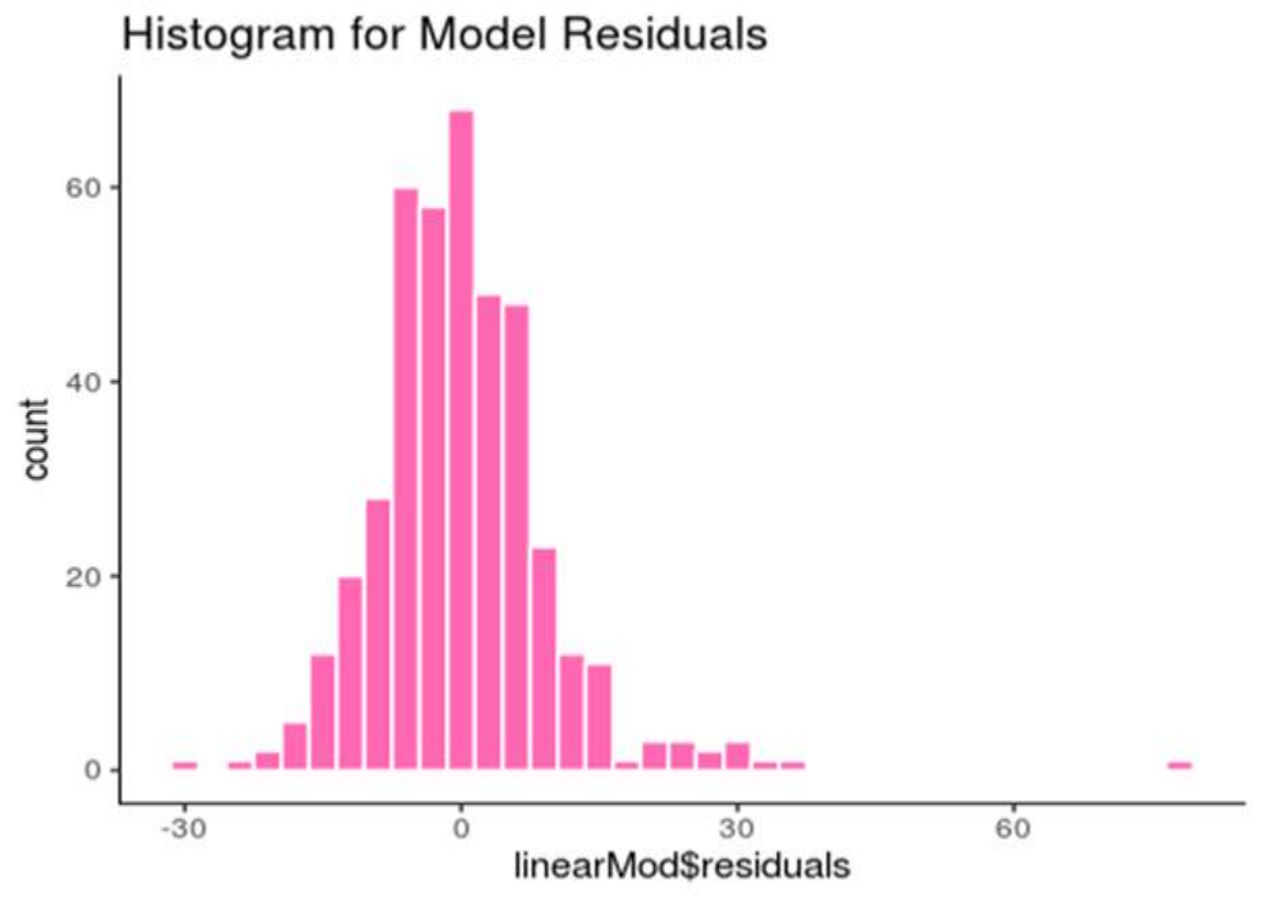
F statistic defines the collective effect of all predictor variables on the response variable. In this model, F=97.46 which is far greater than 1.

Ideally, greater the value of F statistic (higher than 1) , better the fit is (so, this statistic can be intended as a measure of goodness of fit).

❖ Residuals

Residuals show if the predicted response values are close or not to the response values that the model predicts. The residuals should have a symmetrical distribution around zero. Generally, it is required for the residuals to be normally distributed around zero.

To visualize this we will make a histogram using *ggplot2.*



The graph depicts that the residuals are in fact normally distributed around zero and thus gives a clear picture about the accuracy of the model.

HOW ACCURATE ARE THE PREDICTED VALUES

- To do a direct comparison of the forecasted and original values and to estimate the accuracy we will predict our model using the testing dataset and then we will put all the data into a dataframe. Further we will be finding the minimum-maximum accuracy and the mean absolute percentage error.

```
pred=predict(linearMod,testing)

actual_pred=data.frame(cbind(actuals=testing$house_price_
of_unit_area,predicteds=pred))

min_max_accuracy=mean(apply(actual_pred,1,min)/apply(actu
al_pred,1,max))
min_max_accuracy

## [1] 0.7211614
MAPE=mean(abs((actual_pred$predicteds -
actual_pred$actuals)) / actual_pred$actuals)
MAPE     #mean absolute percentage error
## [1] 0.3950446
```

❖ **Min-Max Accuracy**

It is a metric that indicates how close the actual and predicted values are to one another. It is calculated by considering the average between the minimum and the maximum prediction. It can be seen that our model presents an accuracy in the range of 70-73% which is fairly good.

❖ **Mean Absolute Percentage Error**

This metric gives the percentage of error compared to the actual value. It gives a standardized error measure. When applied to the predictions made by our model, this metric gives a result in the range 0.39-0.44.

CONCLUSION AND FUTURE SCOPE

- **The in depth analysis of our prediction model suggests that the geography, location and age of the house are essential factors that dictate the cost of a house.**

- **However as indicated by the accuracy and error values, it may be helpful to incorporate other factors such as physical features etc. to determine the exact pricing.**

- **This regression model that has been further used for prediction can be considered to be a simple yet efficient and effective method for evaluation of costs of not just houses but any land or property in general that needs to be utilized for personal or commercial purposes. Various websites such as www.magicbricks.com/ and https://housing.com/ can amalgamate such models into their procedures of evaluation of selling or re-selling prices of properties.**

FIN.