**Linear Regression**

You work for the Marketing division of a company, and you want to understand what causes increase in Sales of the company.

So, we hypothesize that Sales increases whenever marketing Cost Increases.

**Correlation:**

Correlation tells us about the relationship between the two variables whereas

Corr (Sales, Marketing cost) =+0.7

* The Equation means that 70% of the times when Marketing Cost/Budget increases, Sales increases.
* But we cannot understand the magnitude of Increase. For this purpose, we need Regression

**Regression Equation:**

Regression tells us by how much amount Sales increases when Marketing Cost Increases.

Sales=f (Marketing Cost)

* Assume that the Regression output gives us 0.7
* Then we can interpret this as, when Marketing Cost increases by 1 unit, Sales increases by 0.7

**When can we use Linear Regression?**

Typically, a linear relationship is denoted by a straight line. The Equation is as follows:

***Y = A + BX, where***

**A** = Intercept, **B** = Slope

In this example,

Sales=A+B\*Marketing Cost

What is intercept? It’s the value of Y when X = 0.

* + Here, Intercept =A=Base Sales i.e. when BX=0 which means Base Sales when Marketing Cost =0
  + What if A = 0? Then the line passes through the origin, and Y is directly proportional to X

What is Slope? It is the rate of change of Y when X changes, or the magnitude of impact of changes in X on Y

* + Here Slope =B i.e. when A=0.This gives us the value of change in Sales when Marketing Cost changes by 1 unit.
  + What if B = 0? Then Y is a constant so there is no relationship between Y and X, because, however much X changes, Y does not change

**How to Estimate the Beta Coefficients?**

The most commonly used method is the Ordinary Least Squares (OLS)



Say we have a plot between #Drivers and Fatalities. The questions to be analysed is:

* How many straight lines can be fit through these points?
* How does OLS method help us decide this?

**OLS Method** chooses that line which has the least Error.

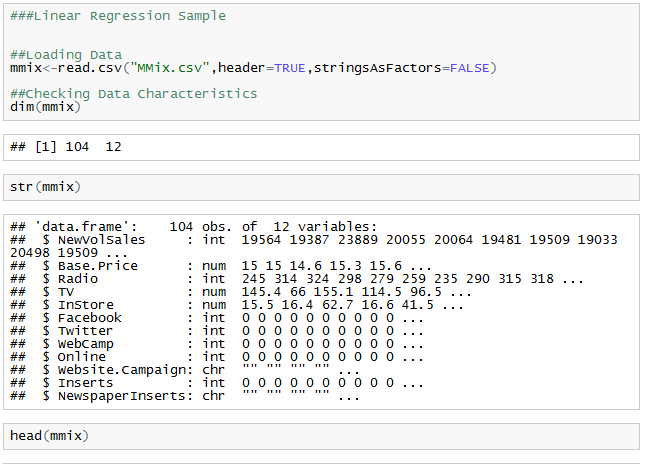
Error is defined as:

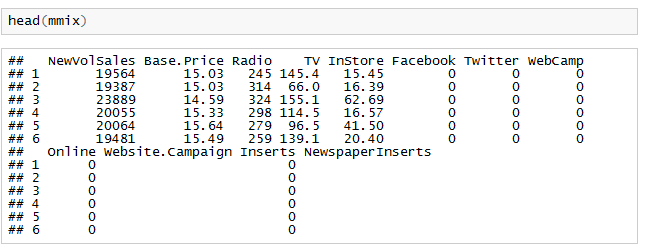
*sum (the perpendicular difference between the points on each line and actual Y) ^2*

**Linear Regression in R**

**Objective:** To try and build a regression model to understand what drives the Sales in a company based on different kinds of marketing costs.

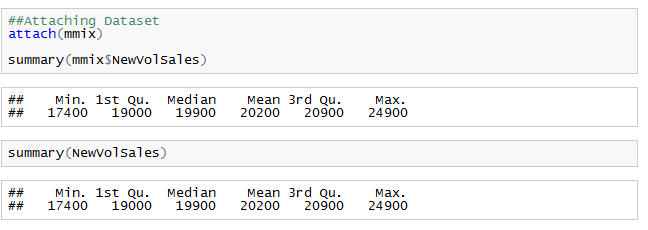
**Data Understanding:** We first read in the “MMix.csv” data, take a glimpse of the dataset and understand the structure of the dataset



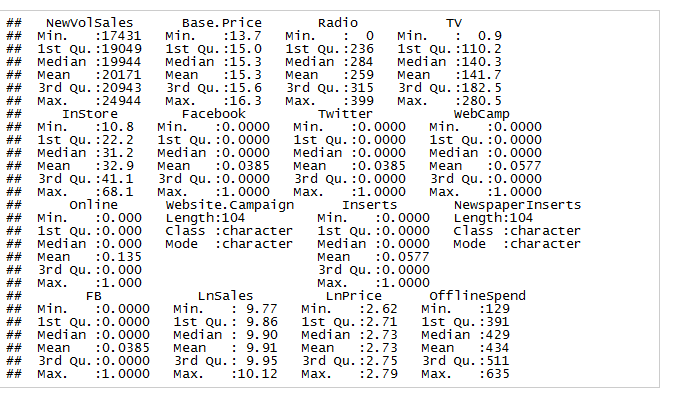


We take a look at the variables present in the Dataset. Our variable of interest is “Newville’s” which will be our DV (Dependent Variable)

The summary command gives us the distribution of the variable and the number of NA’s If present. You can also quickly try and understand if there is any outlier by looking at the max value.

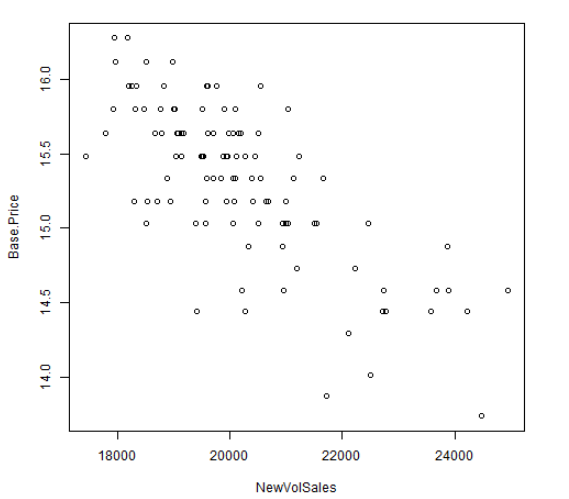


*Summary(mmix)*

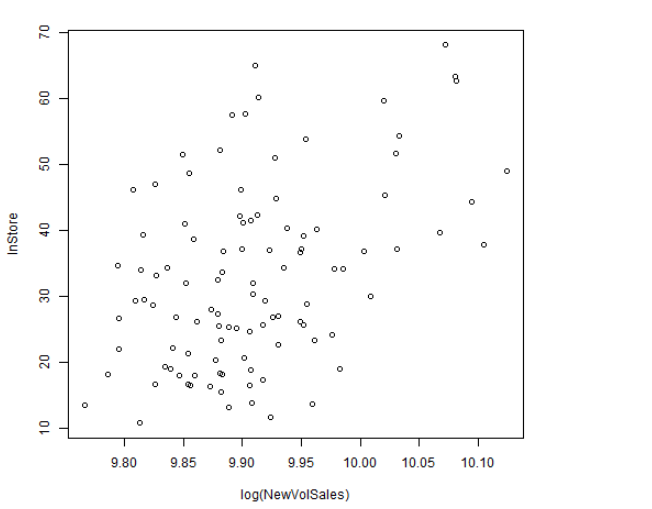


**Data Visualisations**: Try and see if there is any linear trend between NewVolSales and any other variable. In this case Base. Price



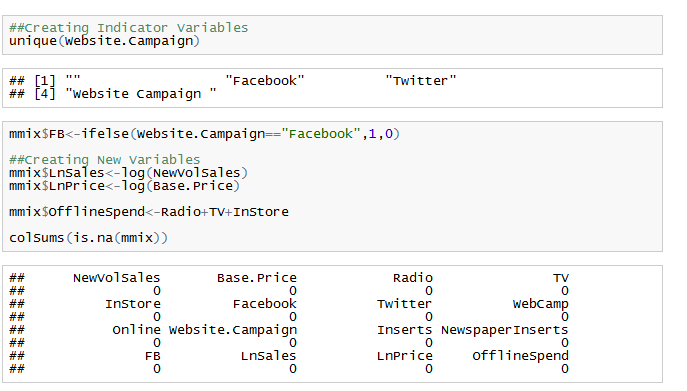


We see that an inverse Linear Relationship exists between NewVolSales and Base. Price So this variable “Base. Price” can be used as an IDV (Independent Variable)



Next, we are going to plot log of “NewVolSales” Vs In store Spend. Please refer to the Variable Transformation section to understand better.

**Data Preparation Steps:**



**Creating Indicator Variables**:

Now before we build a model, we need to convert all Categorical Variables to dummy variables/indicators. This is because the model will throw an error when we use Character variables as IDV in the model.

Using “*if else*” statements indicator variables can be created.

For Example, in the Column “website. Campaign” we are creating 1 if “Facebook” else 0.

**Creating Derived Variables:**

Now we want to understand the impact of Total Offline Spend on the Sales(DV). So we create Offline Spend=Radio Spend+TV Spend+Instore Spend

**Log Transformations:**

Similarly, you can also try few variables transformations as shown above. This is using log ().Log transformations help to capture the range or the variance in the data.

**Missing Values:**

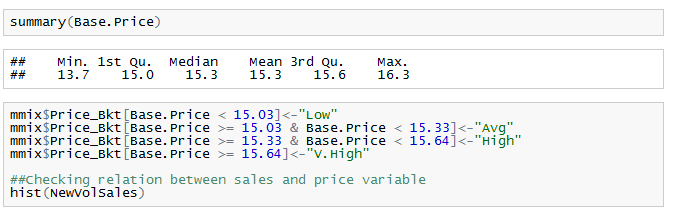
is.na () is used for checking missing values. The dataset provided above is very clean. In case you have missing values, you can replace them with mean/median.

*mmix$variable<-if else(is.na(mmix$variable), mean (mmix$variable, na.rm=T), mmix$variable)*

This is a simple case where we replace the missing values with the column average. But on a higher level of analysis, you can split the dataset into different groups and replace the missing values with group means to get higher accuracy.

**Outliers:** We can find out outliers through summary Function or even Boxplots

**Creating Buckets:**

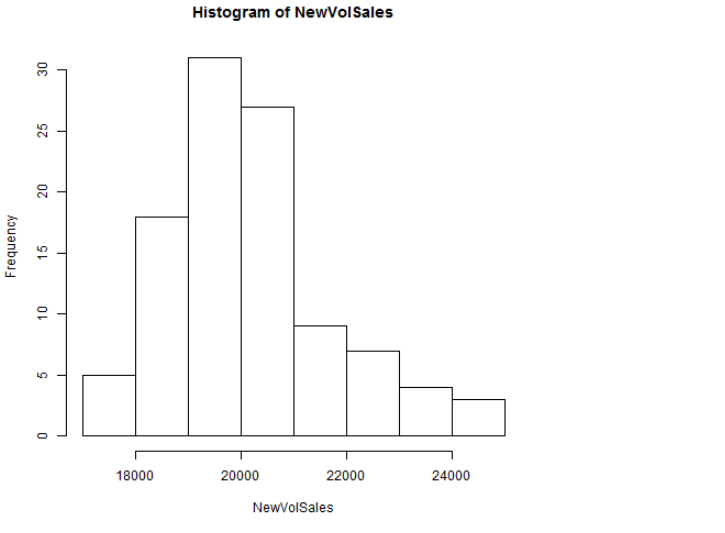


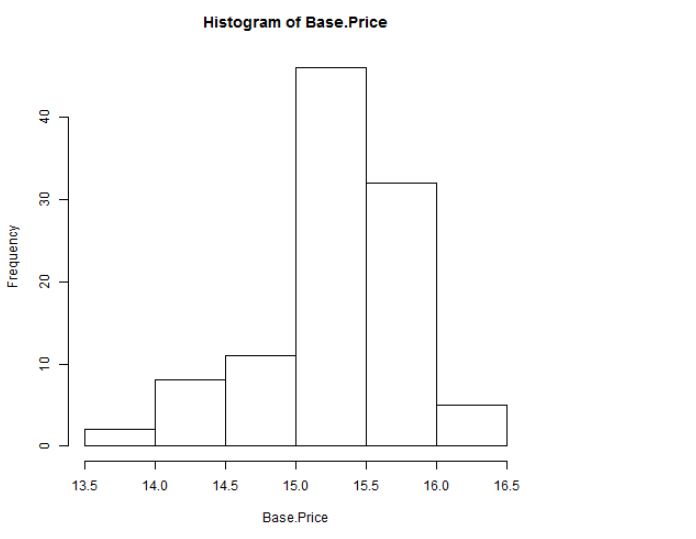
By looking at the Base Price, we can create buckets (Price\_Bkt) such as V.High, High, Avg, Low groups. These can be further then converted to dummy variables (1/0) and used in the model. This is one way of trying to feed in the IDV (capture the trend) as a classification variable and not a numeric variable

**More Visualisations to understand the Trends Better:**

**Histograms:**

The purpose of histograms is to understand the Frequency distribution of variables

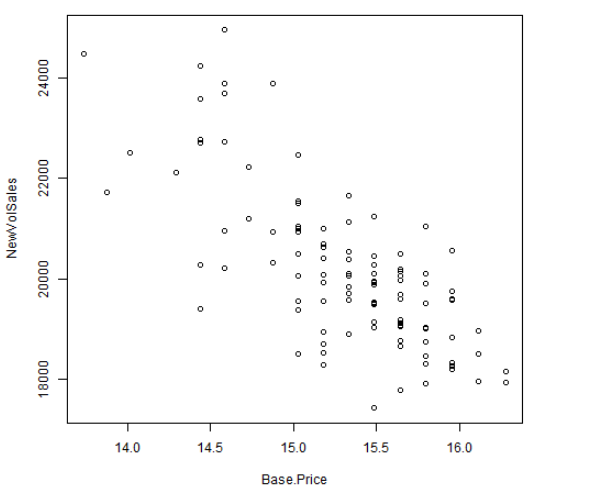


From this histogram we can interpret that 30 data points (Majority of the population) have an average NewVolSales in the bucket 19K – 20K 0

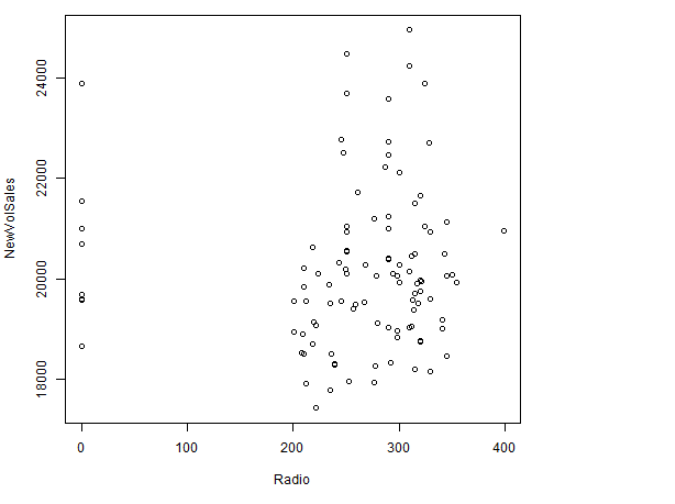
Now looking at the Freq Distribution of Base Price, we see that majority of the data points have a slightly higher Base Price

**Corr and Plot Function:**

The Correlation here is -0.7 which shows that Base Price and NewVolSales have a very high inverse correlation. The plot explains itself.



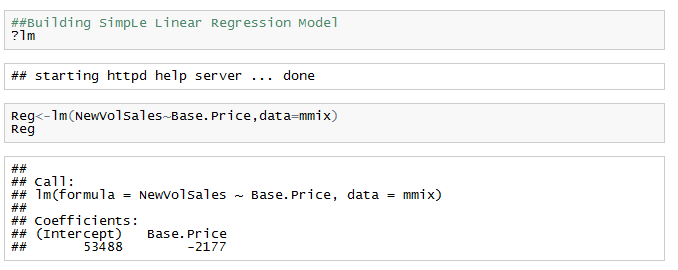
Looking at NewVolSales Vs Radio Spend, we see that majority of the data, Radio Spend >200$(higher side of Radio Spend). But there is no Linear Trend



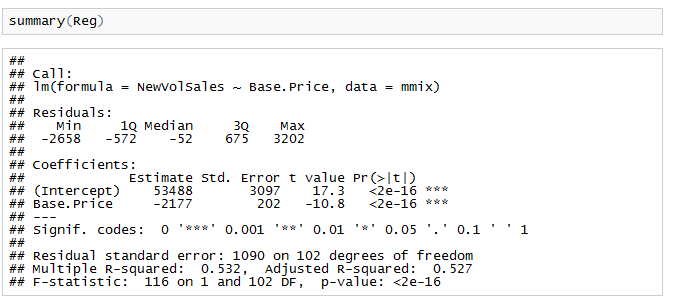
**Simple Linear Regression Model**:

Iteration 1:

Building the Model:



The Function used is lm (). It takes in parameters like the Equation. In this case DV=NewVolSales and IDV=Base. Price and the model are stored in the object “Reg”. Print the object Reg.



Here also we test a hypothesis on the regression model.

**Null Hypothesis**: Base Price has no impact on NewVolSales

**Alternate Hypothesis**: Base Price has an impact on NewVolSales

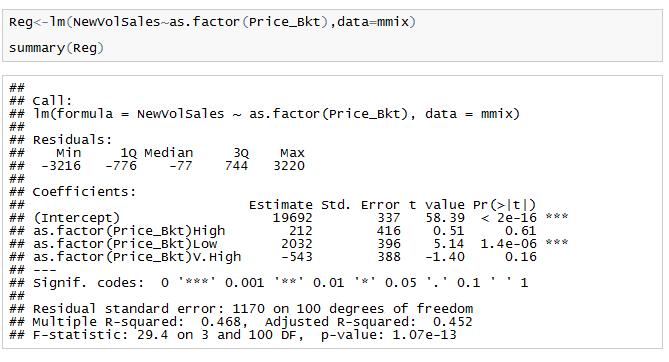
The summary of the object “Reg” gives us all the details about the linear model:

1. The formula: This is the regression Equation used.
   1. **NewVolSales = 53488 - 2177\*Base. Price**
2. Residuals: The distribution of the error (Actual –Predicted)
3. Coefficients:
4. Estimate: Gives the Coefficients (Beta Value) of the IDv’s.
   1. In the results table, we can say that when Base. Price increase by 1 unit, NewVolSales decreases by 2177 units.
5. PR (>t) :
   1. This is the P Value of each variable. Tells us whether the variable is significant or not based on the T test.
   2. In the results table, the p-value on the Base. Price variable is < 0.001. Hence the variable is significant. This means we should accept the alternate hypothesis that as Base Price decreases, New Volume Sales will increase
6. Multiple R Square:
   1. R square is the model performance metric. The value ranges between 0 and 1. The higher the value, the better the model performance is. It explains how much % of my DV is explained by the IDV’s.
   2. However, this value increases of the number of IDV’s increase.
   3. Hence Adjusted R square is used as it captures the true R square value without the impact of #IDv’s

Now we can do a trial an error method of adding variables which are significant till we arrive with the best R square value of the model.

**Iteration2:**

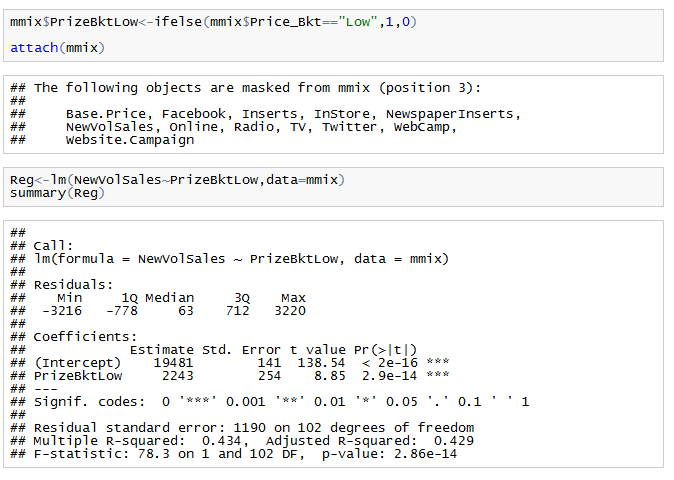
* IDV= Price\_Bkt
* Whenever we have a variable with many factors, instead of creating dummy variables, you can directly use as. factor (IDV)
* R-square: 0.452



Here we see that only Price\_bkt=Low is significant. So, we can create an indicator variable as shown below.

**Iteration2:**

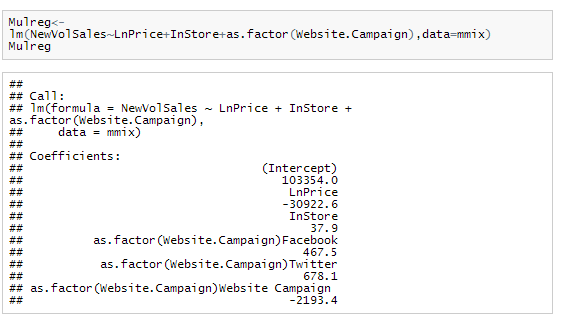
* Adding Price\_BktLow
* R-square: 0.429



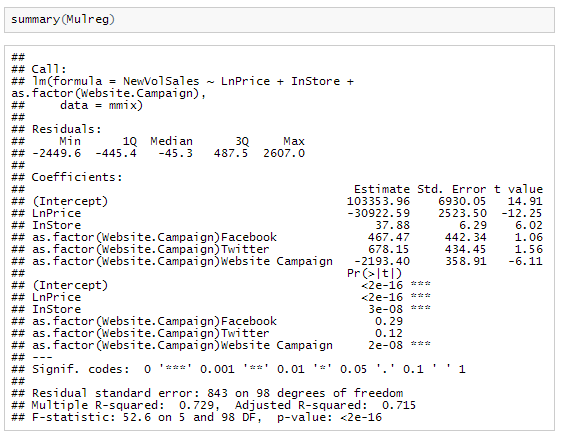
**Multivariate Regression Model**:

Adding more than 1 IDV

Adding IDV’s: LnPrice, Instore, as. factor (Website. Campaign)

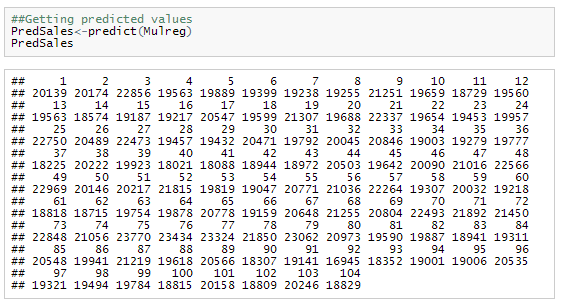


Looking at the summary, we see that variables instore, LnPrice and Website Campaign is significant

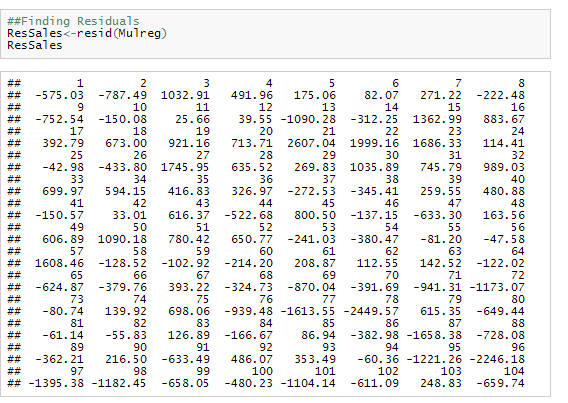


Prediction:

Use Predict Function to map the equation back on the dataset and get predicted values



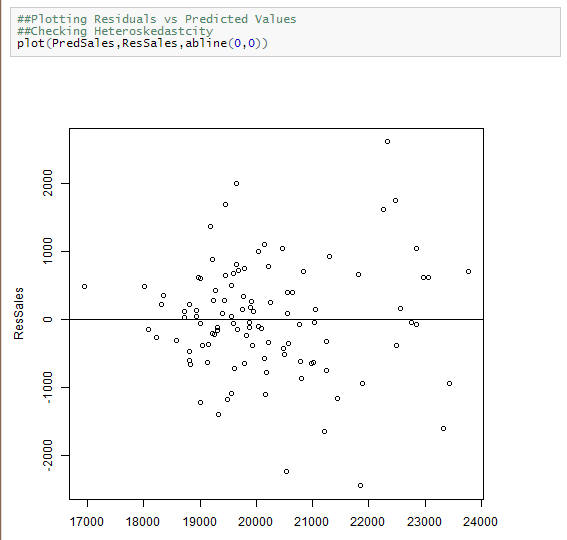
Residuals: This is the error value (Actual – Predicted)



**Checking for Heteroskedasticity**

The plot Between Residuals and Predicted values should be a random scatterplot. This is because if they follow a pattern then the model is not a good model and we say that it suffers from **Heteroskedasticity.**

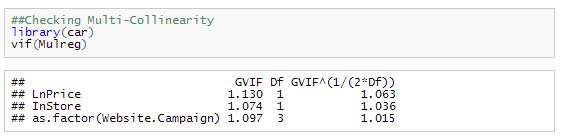
The following plot does not have Heteroskedasticity



**Multicollinearity:**

When there is a high correlation between the IDv’s then the problem of multicollinearity comes up.

To check that you can use the vif function. The vif values should be <10



**Predicted Vs Actuals Plot**