# Data Science Task

## Objectives:

1. Performing Exploratory data Analysis.
2. Predicting Unit Price Using Prediction Modelling.

## Exploratory Data Analysis

- **Comparing Quantity of Stock purchased by Country.**

**There are no missing values in the dataset.**

| Country | Count of Quantity | Max of Quantity | Min of Quantity |
|---|---|---|---|
| **0** | 619 | 960 | -120 |
| **1** | 217 | 240 | -48 |
| **2** | 7 | 96 | 2 |
| **3** | 1007 | 272 | -12 |
| **4** | 13 | 24 | 2 |
| **5** | 73 | 288 | 2 |
| **6** | 351 | 200 | -2 |
| **7** | 296 | 288 | -33 |
| **8** | 21 | 72 | -24 |
| **9** | 211 | 256 | -25 |
| **10** | 3631 | 1440 | -288 |
| **11** | 28 | 24 | 1 |
| **12** | 332 | 144 | -2 |
| **13** | 4198 | 912 | -120 |
| **14** | 4731 | 600 | -288 |
| **15** | 71 | 48 | -1 |
| **16** | 84 | 240 | 2 |
| **17** | 129 | 100 | -32 |
| **18** | 394 | 100 | -12 |
| **19** | 160 | 1488 | -624 |
| **20** | 25 | 15 | 2 |
| **21** | 15 | 36 | 6 |
| **22** | 60 | 45 | -2 |
| **23** | 1165 | 2400 | -144 |
| **24** | 537 | 240 | -12 |
| **25** | 157 | 48 | -6 |
| **26** | 756 | 120 | -12 |
| **27** | 32 | 12 | 1 |
| **28** | 4 | 12 | -5 |
| **29** | 104 | 100 | -1 |
| **30** | 1256 | 360 | -288 |
| **31** | 213 | 576 | -240 |
| **32** | 887 | 144 | -120 |
| **33** | 135 | 72 | -36 |
| **34** | 33 | 24 | 2 |
| **35** | 177917 | 74215 | -80995 |
| **36** | 131 | 25 | 1 |
| **Grand Total** | **200000** | **74215** | **-80995** |

The total count of the **quantity of stock sell** is **200000**.
The **top 3** countries with maximum quantity purchased is **74215** with country code **35** followed by country code **23** with purchased quantity is **2400** and country code **19** with purchased quantity **1488**. The top 3 countries with **maximum quantity returned** is country with country code **35** with quantity **-80995** followed by country code **19** with quantity **-624** and country code **30,14** and **10** with quantity **-288** each. Country with code **35** has the highest quantity **sold and return**.

- **Comparing Unit Price of Stocks by Country.**

| Country Code | Sum of Unit Price | Average of Unit Price | Max of Unit Price | Min of Unit Price |
|---|---|---|---|---|
| **0** | 1681.19 | 2.715977383 | 14.95 | 0 |
| **1** | 961.8 | 4.432258065 | 40 | 0.36 |
| **2** | 31.45 | 4.492857143 | 9.95 | 1.25 |
| **3** | 3562.52 | 3.53775571 | 29.95 | 0.12 |
| **4** | 55.01 | 4.231538462 | 8.25 | 0.85 |
| **5** | 171.65 | 2.351369863 | 12.75 | 0.1 |
| **6** | 1760.73 | 5.016324786 | 293 | 0.19 |
| **7** | 1991.29 | 6.727331081 | 320.69 | 0.21 |
| **8** | 70.38 | 3.351428571 | 40 | 0.29 |
| **9** | 707.96 | 3.355260664 | 18 | 0.29 |
| **10** | 18703.86 | 5.15115946 | 1687.17 | 0 |
| **11** | 144.55 | 5.1625 | 15 | 0.55 |
| **12** | 1574.38 | 4.742108434 | 40 | 0.12 |
| **13** | 22293.76 | 5.310566937 | 4161.06 | 0 |
| **14** | 17593.57 | 3.718784612 | 599.5 | 0 |
| **15** | 459.13 | 6.466619718 | 50 | 0.19 |
| **16** | 236.01 | 2.809642857 | 12.75 | 0.25 |
| **17** | 526.75 | 4.083333333 | 125 | 0.06 |
| **18** | 1690.41 | 4.290380711 | 40 | 0.12 |
| **19** | 355.92 | 2.2245 | 17.55 | 0.21 |
| **20** | 151.65 | 6.066 | 14.95 | 0.85 |
| **21** | 43.45 | 2.896666667 | 5.95 | 1.25 |
| **22** | 402.23 | 6.703833333 | 65 | 0.19 |
| **23** | 3113.04 | 2.672137339 | 110 | 0.19 |
| **24** | 3591.92 | 6.68886406 | 700 | 0 |
| **25** | 622.14 | 3.962675159 | 40 | 0.19 |
| **26** | 4582.29 | 6.061230159 | 557.72 | 0.12 |
| **27** | 150.54 | 4.704375 | 14.95 | 0 |
| **28** | 9.2 | 2.3 | 2.95 | 1.65 |
| **29** | 12438.86 | 119.6044231 | 2382.92 | 0.19 |
| **30** | 6464.97 | 5.147269108 | 1715.85 | 0 |
| **31** | 822.94 | 3.863568075 | 40 | 0.19 |
| **32** | 2923.31 | 3.29572717 | 40 | 0.12 |
| **33** | 311.27 | 2.305703704 | 16.95 | 0.42 |
| **34** | 115.34 | 3.495151515 | 14.95 | 0.29 |
| **35** | 580127.274 | 3.26066241 | 38970 | 0 |
| **36** | 427.06 | 3.26 | 16.95 | 0.29 |
| **Grand Total** | **690869.804** | **3.45434902** | **38970** | **0** |

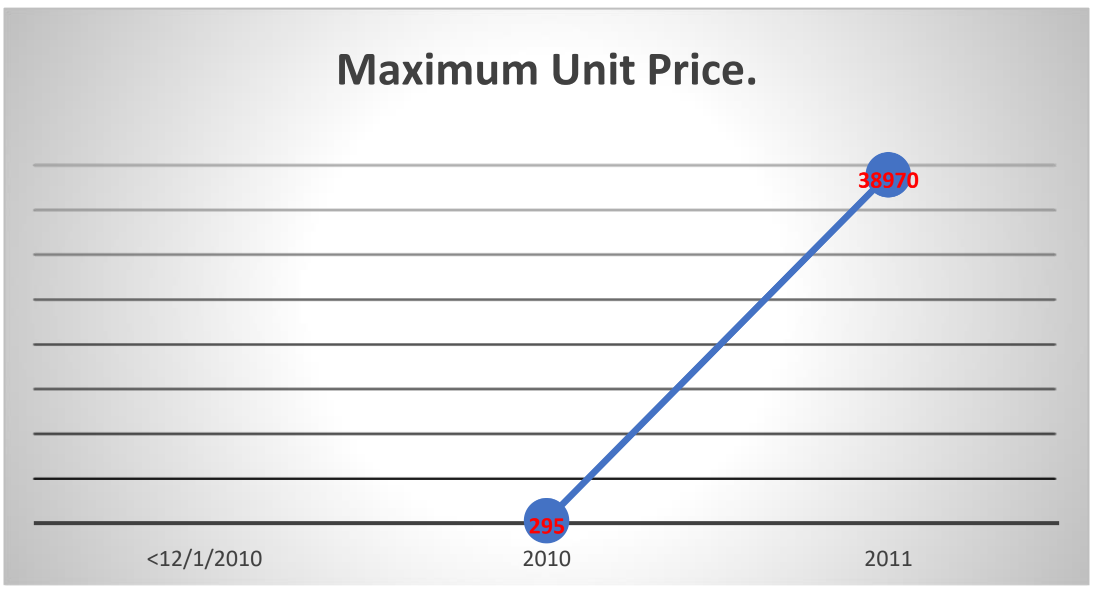
As, we can observe from the above table the maximum unit price of the stock is 38970 in country with country code 35 and Maximum total unit price is also in the country code 35. Also, the maximum average unit price of the stock is 119.60 that is with the country code 29.

- **Comparing Quantity of stocks sold and purchased by Year and Quarter.**

| Row Labels | Sum of Quantity Purchased | Count of Quantity Purchased | Max of Quantity Purchased | Min of Quantity Purchased |
|---|---|---|---|---|
| **<12/1/2010** | | | | |
| **<12/1/2010** | | | | |
| **Qtr1** | | | | |
| **2011** | 501360 | 34395 | 74215 | -600 |
| **Qtr2** | | | | |
| **2011** | 495499 | 39362 | 4300 | -1930 |
| **Qtr3** | | | | |
| **2011** | 627609 | 47239 | 3186 | -756 |
| **Qtr4** | | | | |
| **2010** | 147004 | 13166 | 2400 | -240 |
| **2011** | 627628 | 65838 | 12540 | -80995 |
| **Grand Total** | **2399100** | **200000** | **74215** | **-80995** |

The highest total number of stocks purchased or sold was in the 4th Quarter of 2011 and minimum in the 4th Quarter of 2010. Maximum of stocks sold, returned and the total transaction of the stocks was in the 4th Quarter of 2011 and we can verify with the fact that the stocks market is on growth from comparing from 2010 and hence, the data is true without any manipulations.

1. **Comparing Unit Price of stocks by Year and Quarter.**

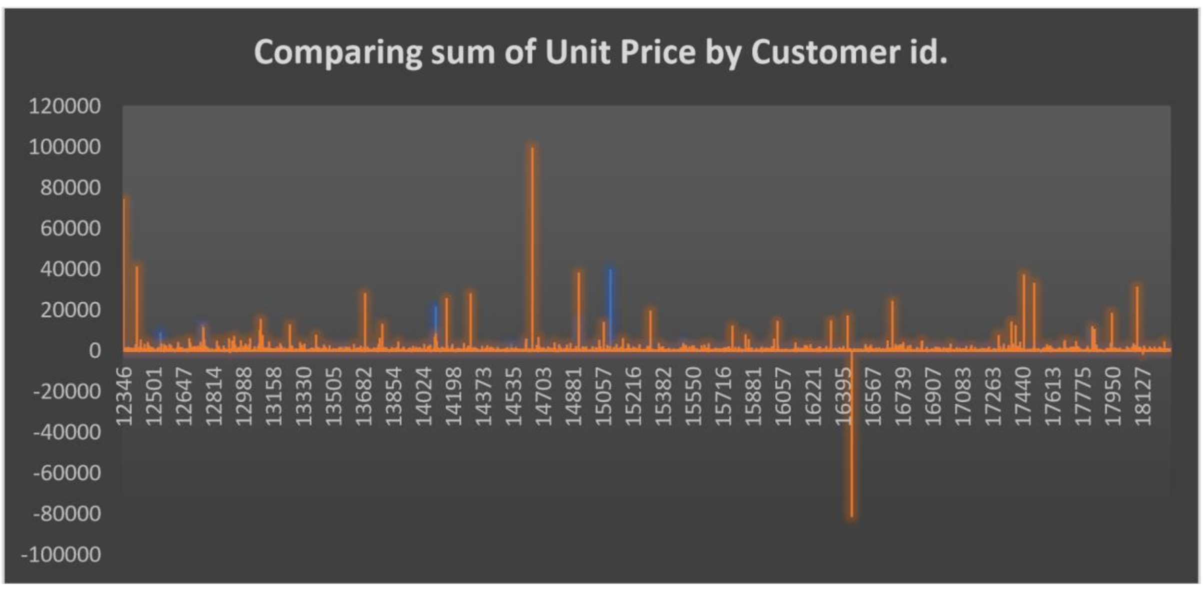


| Row Labels | Sum of Unit Price | Min of Unit Price | Average of Unit Price | Max of Unit Price |
|---|---|---|---|---|
| **<12/1/2010** | | | | |
| **<12/1/2010** | | | | |
| **Qtr1** | | | | |
| **2011** | 115999.38 | 0 | 3.372565198 | 1715.85 |
| **Qtr2** | | | | |
| **2011** | 175317.191 | 0 | 4.453970606 | 38970 |
| **Qtr3** | | | | |
| **2011** | 144912.883 | 0 | 3.067653485 | 3155.95 |
| **Qtr4** | | | | |
| **2010** | 42120.86 | 0 | 3.199214644 | 295 |
| **2011** | 212519.49 | 0 | 3.227915338 | 4161.06 |
| **Grand Total** | **690869.804** | **0** | **3.45434902** | **38970** |

Maximum unit price was in the 4th Quarter of 2011 and minimum was in the 4th Quarter of 2010. The average price was highest in the 2nd Quarter of 2011 and it was minimum in the 4th Quarter of 2010. The Maximum sum of unit price was in the 4th Quarter of 2011.

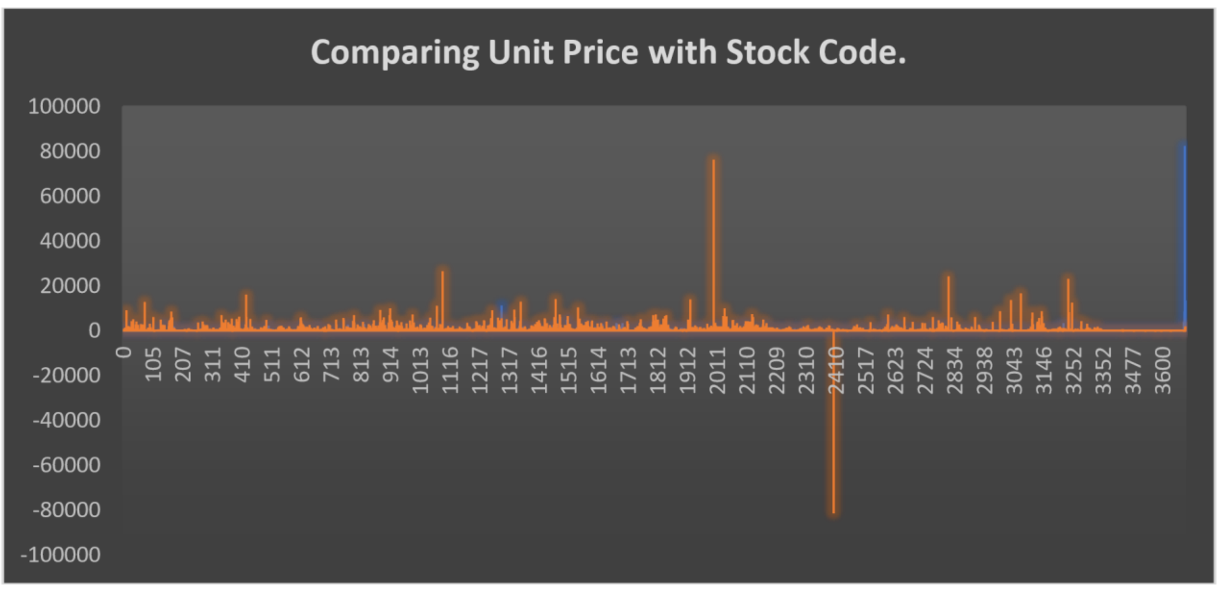
As, from the above tables we have seen that the maximum stocks quantity, maximum unit price maximum transaction was done in the 4th Quarter of 2011 that prove the fact that the stocks market are on demand and are increasing as more and more people are getting involve with the stock market.

- **Comparing Unit Price of stocks by Customer Id.**



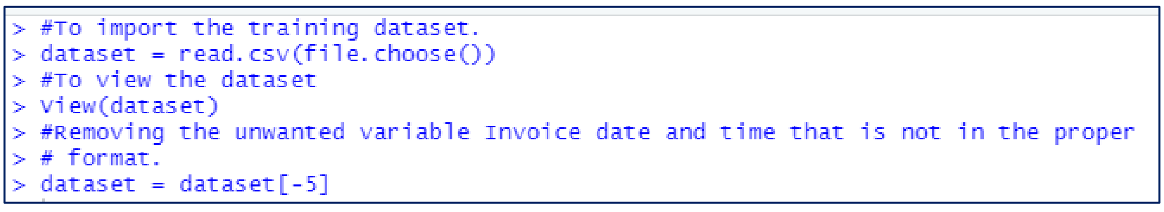
Customer with customer id 15098 has the maximum total unit price for its stocks and with total of 908 Stock Quantity.

- **Comparing Unit Price of stocks by Stock Code.**

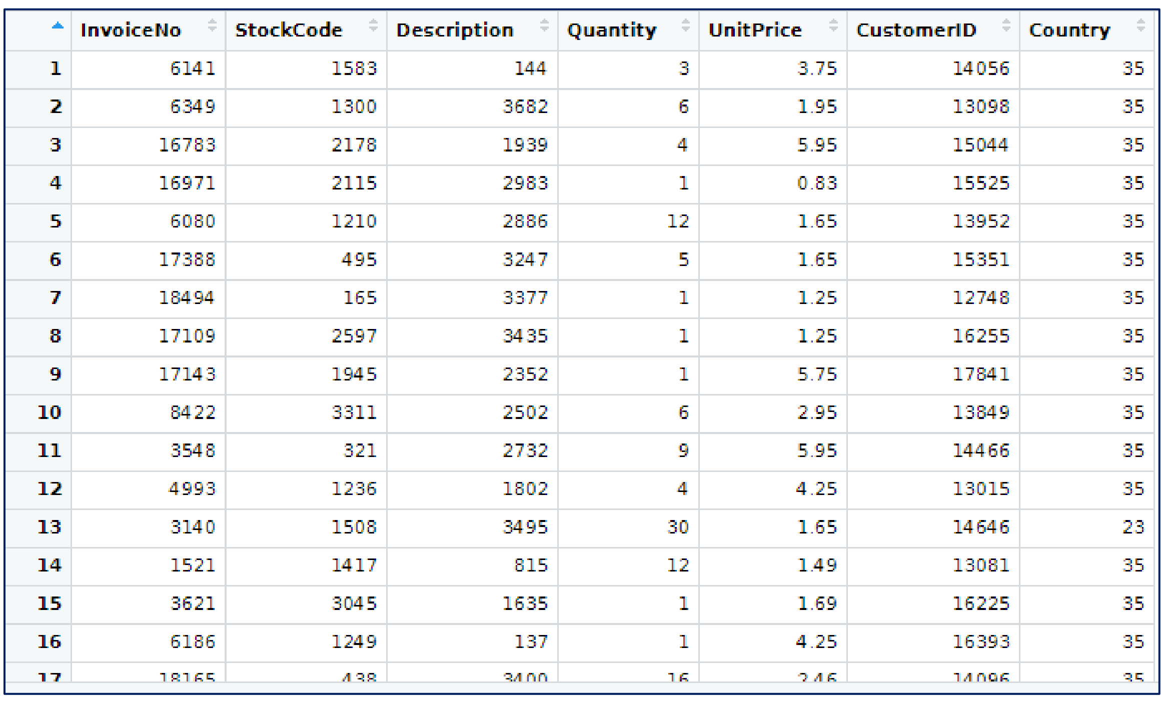


Stock code 3683 has the maximum total unit price for its stocks and has the stock quantity of 1510 stocks.

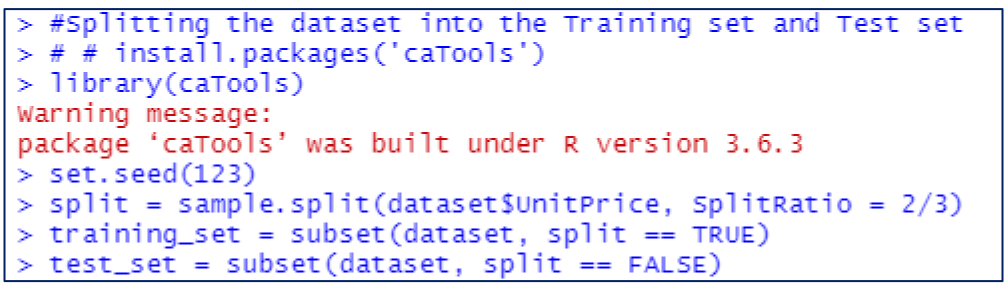
## 2. Predictive Modelling using R-software.



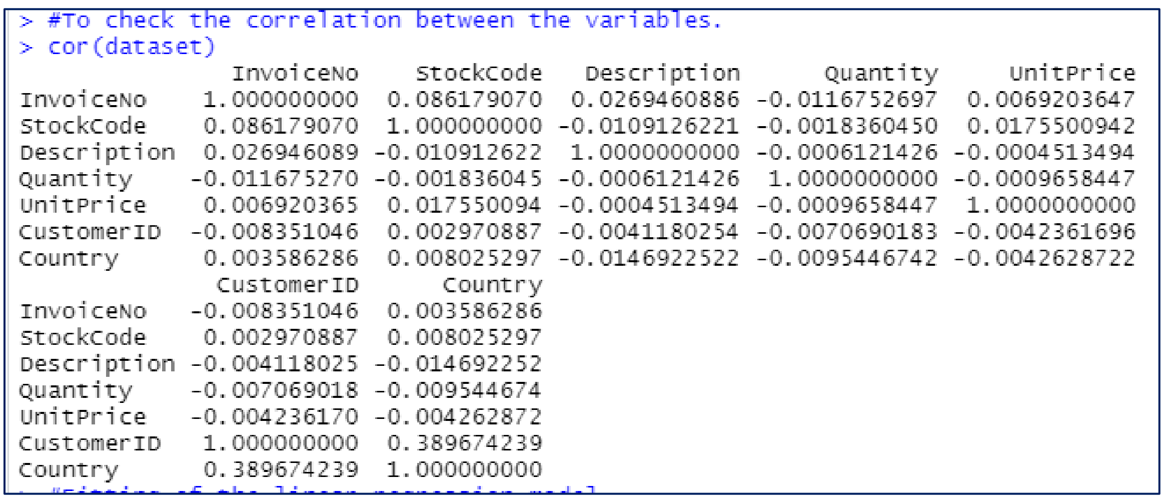
```
> #To import the training dataset.
> dataset = read.csv(file.choose())
> #To view the dataset
> View(dataset)
> #Removing the unwanted variable Invoice date and time that is not in the proper
> # format.
> dataset = dataset[-5]
```



| | InvoiceNo | StockCode | Description | Quantity | UnitPrice | CustomerID | Country |
|---|---|---|---|---|---|---|---|
| 1 | 6141 | 1583 | 144 | 3 | 3.75 | 14056 | 35 |
| 2 | 6349 | 1300 | 3682 | 6 | 1.95 | 13098 | 35 |
| 3 | 16783 | 2178 | 1939 | 4 | 5.95 | 15044 | 35 |
| 4 | 16971 | 2115 | 2983 | 1 | 0.83 | 15525 | 35 |
| 5 | 6080 | 1210 | 2886 | 12 | 1.65 | 13952 | 35 |
| 6 | 17388 | 495 | 3247 | 5 | 1.65 | 15351 | 35 |
| 7 | 18494 | 165 | 3377 | 1 | 1.25 | 12748 | 35 |
| 8 | 17109 | 2597 | 3435 | 1 | 1.25 | 16255 | 35 |
| 9 | 17143 | 1945 | 2352 | 1 | 5.75 | 17841 | 35 |
| 10 | 8422 | 3311 | 2502 | 6 | 2.95 | 13849 | 35 |
| 11 | 3548 | 321 | 2732 | 9 | 5.95 | 14466 | 35 |
| 12 | 4993 | 1236 | 1802 | 4 | 4.25 | 13015 | 35 |
| 13 | 3140 | 1508 | 3495 | 30 | 1.65 | 14646 | 23 |
| 14 | 1521 | 1417 | 815 | 12 | 1.49 | 13081 | 35 |
| 15 | 3621 | 3045 | 1635 | 1 | 1.69 | 16225 | 35 |
| 16 | 6186 | 1249 | 137 | 1 | 4.25 | 16393 | 35 |
| 17 | 18165 | 438 | 3400 | 16 | 2.46 | 14096 | 35 |



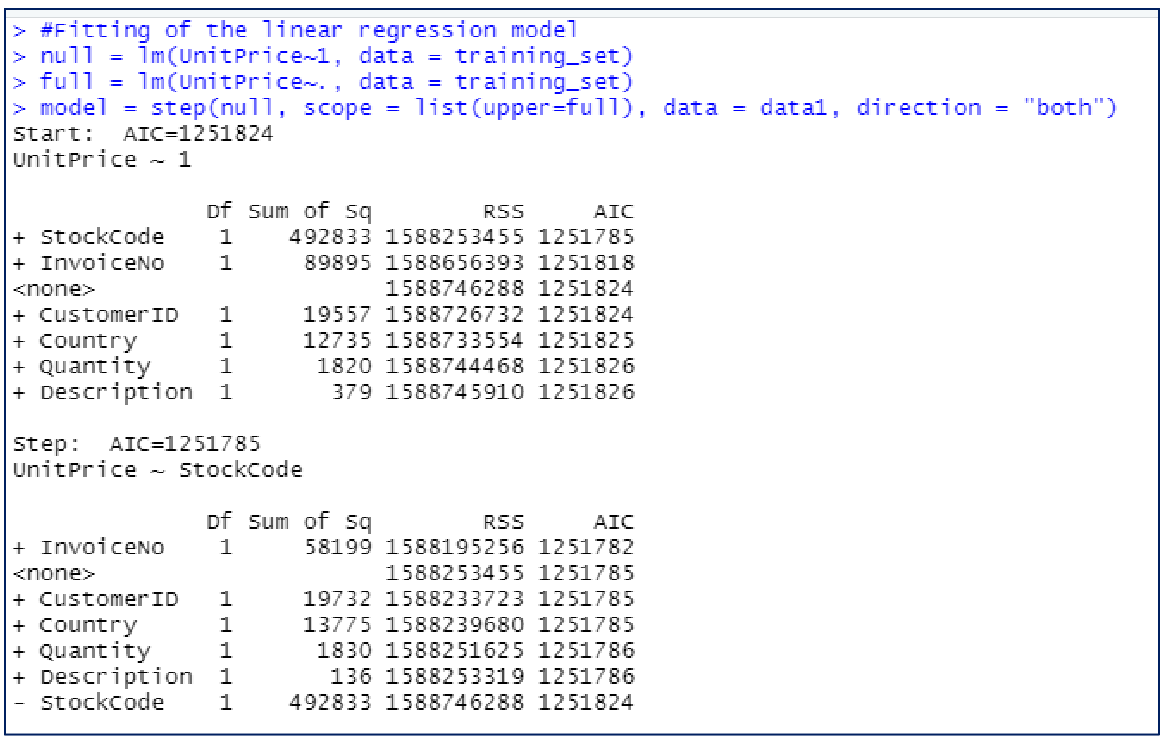
```
> #Splitting the dataset into the Training set and Test set
> # # install.packages('caTools')
> library(caTools)
Warning message:
package ‘caTools’ was built under R version 3.6.3
> set.seed(123)
> split = sample.split(dataset$UnitPrice, SplitRatio = 2/3)
> training_set = subset(dataset, split == TRUE)
> test_set = subset(dataset, split == FALSE)
```



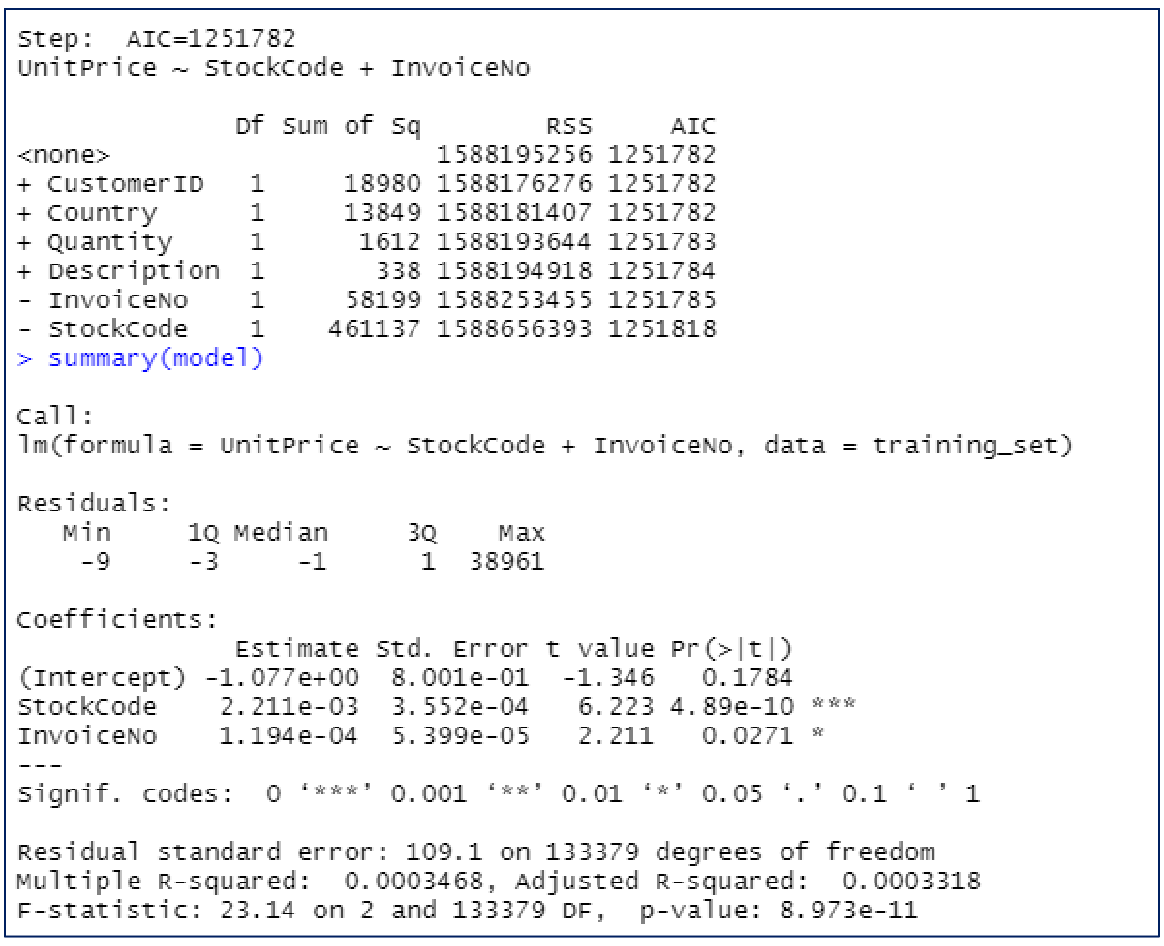
```
> #To check the correlation between the variables.
> cor(dataset)
              InvoiceNo    StockCode   Description      Quantity     UnitPrice
InvoiceNo    1.000000000  0.086179070  0.0269460886 -0.0116752697  0.0069203647
StockCode    0.086179070  1.000000000 -0.0109126221 -0.0018360450  0.0175500942
Description  0.026946089 -0.010912622  1.0000000000 -0.0006121426 -0.0004513494
Quantity    -0.011675270 -0.001836045 -0.0006121426  1.0000000000 -0.0009658447
UnitPrice    0.006920365  0.017550094 -0.0004513494 -0.0009658447  1.0000000000
CustomerID  -0.008351046  0.002970887 -0.0041180254 -0.0070690183 -0.0042361696
Country      0.003586286  0.008025297 -0.0146922522 -0.0095446742 -0.0042628722
              CustomerID      Country
InvoiceNo   -0.008351046  0.003586286
StockCode    0.002970887  0.008025297
Description -0.004118025 -0.014692252
Quantity    -0.007069018 -0.009544674
UnitPrice   -0.004236170 -0.004262872
CustomerID   1.000000000  0.389674239
Country      0.389674239  1.000000000
> #Fitting of the linear regression model
```

From the above sniped we can check the correlation between various variables. We can see that the there is weak positive relation between Invoice no, Stock code, and unit price. There is weak negative relation between Description, Quantity, Customer id, Country with Unit price.

### 3. Fitting Multiple Linear Regression Model using stepwise regression method in both directions.



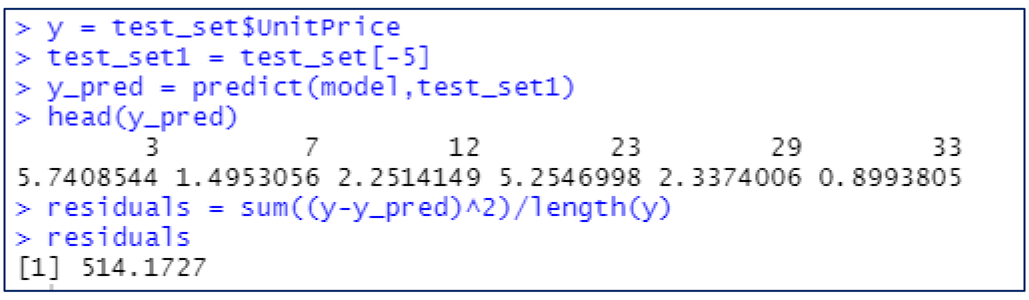
```
> #Fitting of the linear regression model
> null = lm(UnitPrice~1, data = training_set)
> full = lm(UnitPrice~., data = training_set)
> model = step(null, scope = list(upper=full), data = data1, direction = "both")
Start:  AIC=1251824
UnitPrice ~ 1

              Df Sum of Sq        RSS     AIC
+ StockCode    1    492833 1588253455 1251785
+ InvoiceNo    1     89895 1588656393 1251818
<none>                     1588746288 1251824
+ CustomerID   1     19557 1588726732 1251824
+ Country      1     12735 1588733554 1251825
+ Quantity     1      1820 1588744468 1251826
+ Description  1       379 1588745910 1251826

Step:  AIC=1251785
UnitPrice ~ StockCode

              Df Sum of Sq        RSS     AIC
+ InvoiceNo    1     58199 1588195256 1251782
<none>                     1588253455 1251785
+ CustomerID   1     19732 1588233723 1251785
+ Country      1     13775 1588239680 1251785
+ Quantity     1      1830 1588251625 1251786
+ Description  1       136 1588253319 1251786
- StockCode    1    492833 1588746288 1251824
```



```
Step:  AIC=1251782
UnitPrice ~ StockCode + InvoiceNo

              Df Sum of Sq        RSS     AIC
<none>                     1588195256 1251782
+ CustomerID   1     18980 1588176276 1251782
+ Country      1     13849 1588181407 1251782
+ Quantity     1      1612 1588193644 1251783
+ Description  1       338 1588194918 1251784
- InvoiceNo    1     58199 1588253455 1251785
- StockCode    1    461137 1588656393 1251818
> summary(model)

Call:
lm(formula = UnitPrice ~ StockCode + InvoiceNo, data = training_set)

Residuals:
   Min     1Q Median     3Q    Max 
    -9     -3     -1      1  38961 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.077e+00  8.001e-01  -1.346   0.1784    
StockCode    2.211e-03  3.552e-04   6.223 4.89e-10 ***
InvoiceNo    1.194e-04  5.399e-05   2.211   0.0271 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 109.1 on 133379 degrees of freedom
Multiple R-squared:  0.0003468,	Adjusted R-squared:  0.0003318 
F-statistic: 23.14 on 2 and 133379 DF,  p-value: 8.973e-11
```

Hence, the best multiple Linear Regression Model for predicting unit price of stock is **Unit Price ~ Stock code + Invoice No** with adjusted R-square as 0.033 % that is too low and hence multiple linear regression model is not working well with this model.

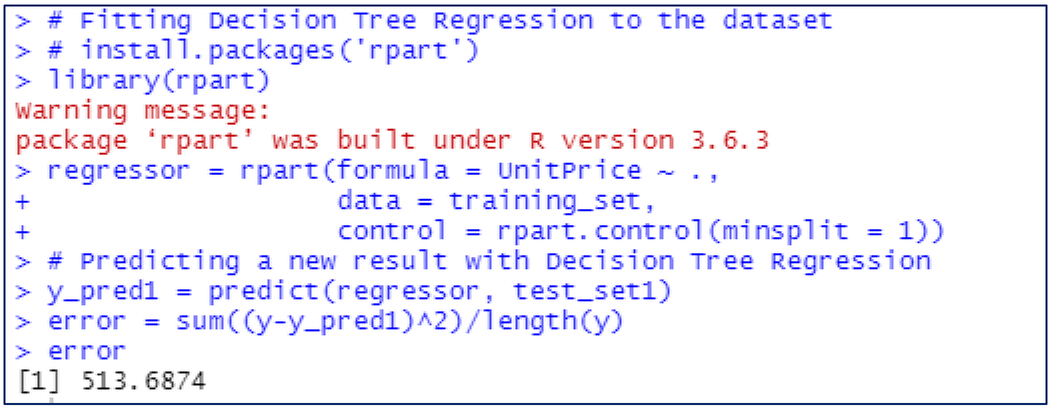
### 4. Prediction of Unit Price with Multiple Linear Regression Model.



```
> y = test_set$UnitPrice
> test_set1 = test_set[-5]
> y_pred = predict(model,test_set1)
> head(y_pred)
        3         7        12        23        29        33 
5.7408544 1.4953056 2.2514149 5.2546998 2.3374006 0.8993805 
> residuals = sum((y-y_pred)^2)/length(y)
> residuals
[1] 514.1727
```

The mean error sum of squares for multiple linear regression model is **514.1727.**

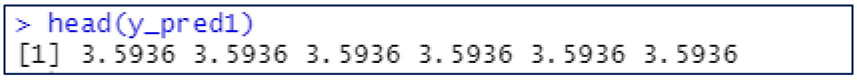
### 5. Fitting Decision Tree Regression Model.



```
> # Fitting Decision Tree Regression to the dataset
> # install.packages('rpart')
> library(rpart)
Warning message:
package ‘rpart’ was built under R version 3.6.3 
> regressor = rpart(formula = UnitPrice ~ .,
+                   data = training_set,
+                   control = rpart.control(minsplit = 1))
> # Predicting a new result with Decision Tree Regression
> y_pred1 = predict(regressor, test_set1)
> error = sum((y-y_pred1)^2)/length(y)
> error
[1] 513.6874
```

With decision tree regression model the mean error sum of squares is **513.6874** that is less than the multiple linear regression model. Hence, we will prefer Decision Tree Regression model over multiple linear regression model.

Checking the predicted values of unit price for test data set.



```
> head(y_pred1)
[1] 3.5936 3.5936 3.5936 3.5936 3.5936 3.5936
```

- ## Observations:

1. Stock code 3683 has the maximum total unit price for its stocks and has the stock quantity of 1510 stocks.
2. Customer with customer id 15098 has the maximum total unit price for its stocks and with total of 908 Stock Quantity.
3. Maximum unit price was in the 4th Quarter of 2011 and minimum was in the 4th Quarter of 2010. The average price was highest in the 2nd Quarter of 2011 and it was minimum in the 4th Quarter of 2010. The Maximum sum of unit price was in the 4th Quarter of 2011.
4. There is weak positive relation between Invoice no, Stock code, and unit price. There is weak negative relation between Description, Quantity, Customer id, Country with Unit price.

- ## Conclusion:

1. Decision Tree regression works well than the multiple linear regression model for predicting Unit Price.