# R Assignment

# AIT Educamp 2022

# Submitted By:


# Dhruv Dixit

# United College of Engineering and Management, Prayagraj

# Introduction

The data provided is a massive collection of supermarket data having different details of various parameters regarding what factors supports in sales of products like gender, Unit Price, Quantity etc.

The task is to perform descriptive statistics, correlation analysis and linear along with multiple regression analysis.

# Code & Inferences

**STATISTICAL ANALYSIS WITH R - AIT Assignment**

**Date of Submission: 14th Sep 2022**

**Package Installation :**

```
install.packages(c('psych', 'MASS', 'GGally', 'VGAM', 'ggplot2', 'truncreg', 'boot',
                   'foreign', 'Hmisc', 'aod', 'margins', 'reshape2'))
```

**Calling the libraries downloaded :**

```
lapply(c('psych', 'MASS', 'GGally', 'VGAM', 'ggplot2', 'truncreg', 'boot',
         'foreign', 'Hmisc', 'aod', 'margins', 'reshape2'), library, character.only = TRUE)
```

## Task 1 - DESCRIPTIVE STATISTICS

**Setting up your working directory :**

```
setwd("C:/Users/dhruv/Downloads/R Assignment")
```

**Uploading the first dataset to RStudio :**

```
dataset <- read.csv(file.choose(), header = TRUE)
attach(dataset)
```

**Summary of the descriptive statistics:**

```
summary(dataset)
```

**Output :**

```
     X.            Customer.type        Gender          Product.line        Unit.price       Quantity
Length:1000        Length:1000        Length:1000        Length:1000        Min.   :10.08   Min.   : 1.00
Class :character   Class :character   Class :character   Class :character   1st Qu.:32.88   1st Qu.: 3.00
Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :55.23   Median : 5.00
                                                                            Mean   :55.67   Mean   : 5.51
                                                                            3rd Qu.:77.94   3rd Qu.: 8.00
                                                                            Max.   :99.96   Max.   :10.00
    Tax.5.            Total           Payment                cogs        gross.margin.percentage  gross.income
Min.   : 0.5085   Min.   :  10.68   Length:1000        Min.   : 10.17   Min.   :4.762            Min.   : 0.5085
1st Qu.: 5.9249   1st Qu.: 124.42   Class :character   1st Qu.:118.50   1st Qu.:4.762            1st Qu.: 5.9249
Median :12.0880   Median : 253.85   Mode  :character   Median :241.76   Median :4.762            Median :12.0880
Mean   :15.3794   Mean   : 322.97                      Mean   :307.59   Mean   :4.762            Mean   :15.3794
3rd Qu.:22.4453   3rd Qu.: 471.35                      3rd Qu.:448.90   3rd Qu.:4.762            3rd Qu.:22.4453
Max.   :49.6500   Max.   :1042.65                      Max.   :993.00   Max.   :4.762            Max.   :49.6500
    Rating
Min.   : 4.000
1st Qu.: 5.500
Median : 7.000
Mean   : 6.973
3rd Qu.: 8.500
Max.   :10.000
```

**Seeing the column names of dataset :**

names(dataset)

**Output :**

```
 [1] "X."                "Customer.type"     "Gender"                "Product.line"
 [5] "Unit.price"        "Quantity"          "Tax.5."                "Total"
 [9] "Payment"           "cogs"              "gross.margin.percentage" "gross.income"
[13] "Rating"
```

**Summary of the descriptive statistics :**

describe(Unit.price)

describe(Quantity)

describe(Tax.5.)

describe(Total)

describe(Rating)

describe(cogs)

**Output :**

```
> describe(Unit.price)
Unit.price
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    1000        0      943        1    55.67     30.6    15.28    19.31    32.88    55.23    77.94    93.12    97.22

lowest : 10.08 10.13 10.16 10.17 10.18, highest: 99.82 99.83 99.89 99.92 99.96
> describe(Quantity)
Quantity
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    1000        0       10     0.99     5.51     3.36        1        1        3        5        8       10       10

lowest :  1  2  3  4  5, highest:  6  7  8  9 10

Value          1     2     3     4     5     6     7     8     9    10
Frequency    112    91    90   109   102    98   102    85    92   119
Proportion 0.112 0.091 0.090 0.109 0.102 0.098 0.102 0.085 0.092 0.119
> describe(Tax.5.)
Tax.5.
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    1000        0      990        1    15.38    12.89    1.956    3.243    5.925   12.088   22.445   34.234   39.166

lowest :  0.5085  0.6045  0.6270  0.6390  0.6990, highest: 48.6900 48.7500 49.2600 49.4900 49.6500
> describe(Total)
Total
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    1000        0      990        1      323    270.7    41.07    68.10   124.42   253.85   471.35   718.91   822.50

lowest :   10.6785   12.6945   13.1670   13.4190   14.6790, highest: 1022.4900 1023.7500 1034.4600 1039.2900 1042.6500
> describe(Rating)
Rating
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    1000        0       61        1    6.973    1.985    4.295    4.500    5.500    7.000    8.500    9.400    9.700

lowest :  4.0  4.1  4.2  4.3  4.4, highest:  9.6  9.7  9.8  9.9 10.0
> describe(cogs)
cogs
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    1000        0      990        1    307.6    257.8    39.11    64.86   118.50   241.76   448.91   684.68   783.33

lowest :  10.17  12.09  12.54  12.78  13.98, highest: 973.80 975.00 985.20 989.80 993.00
```

**Checking Missing Values :**

sum(is.na(dataset))

**Output :**

```
> sum(is.na(dataset))
[1] 0
```

**Calculating Average Unit price for each category :**

library(dplyr)

prodVar <- group_by(dataset,Product.line)

summarise(prodVar, avgUnit = mean(Unit.price))

**Output :**

```
# A tibble: 6 x 2
  Product.line           avgUnit
  <chr>                    <dbl>
1 Electronic accessories    53.6
2 Fashion accessories       57.2
3 Food and beverages        56.0
4 Health and beauty         54.9
5 Home and lifestyle        55.3
6 Sports and travel         57.0
```

## Task 2 - DATA VISUALIZATION/ CORRELATION ANALYSIS

➔ **Pearson's Product Moment Correlation for Unit.price and Quantity :**

```
cor.test(Unit.price,Quantity,alternative = c("two.sided"), method = c("pearson"),
        exact = NULL, conf.level= 0.95, continuity = FALSE)
```

**Output :**

```
        Pearson's product-moment correlation

data:  Unit.price and Quantity
t = 0.3405, df = 998, p-value = 0.7336
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.05124976  0.07272206
sample estimates:
       cor
0.01077756
```

**Result:** *It is observed clearly that p-value is not less than 0.05 and Correlation Percentage is 1%. Hence, Unit.price and Quantity are not correlated and null hypothesis is not rejected.*

➔ **Pearson's Product Moment Correlation for Unit.price and Tax.5.**

```
cor.test(Unit.price,Tax.5.,alternative = c("two.sided"), method = c("pearson"),
        exact = NULL, conf.level= 0.95, continuity = FALSE)
```

**Output :**

```
        Pearson's product-moment correlation

data:  Unit.price and Tax.5.
t = 25.897, df = 998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5953677 0.6696376
sample estimates:
      cor
0.6339621
```

**Result :** *It is observed clearly that p-value is less than 0.05 and Correlation Percentage is 63.39%. Hence, Unit.price and Tax.5. are highly correlated and null hypothesis is rejected.*

➔ **Pearson's Product Moment Correlation for Unit.price and Total**

```
cor.test(Unit.price,Total,alternative = c("two.sided"), method = c("pearson"),
        exact = NULL, conf.level= 0.95, continuity = FALSE)
```

**Output :**

```
        Pearson's product-moment correlation

data:  Unit.price and Total
t = 25.897, df = 998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5953677 0.6696376
sample estimates:
      cor
0.6339621
```

**Result :** *It is observed clearly that p-value is less than 0.05 and Correlation Percentage is 63.39%. Hence, Unit.price and Total are highly correlated and null hypothesis is rejected.*

➔ **Pearson's Product Moment Correlation for Unit.price and cogs**

```
cor.test(Unit.price,cogs,alternative = c("two.sided"), method = c("pearson"),
        exact = NULL, conf.level= 0.95, continuity = FALSE)
```

**Output :**

```
        Pearson's product-moment correlation

data:  Unit.price and cogs
t = 25.897, df = 998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5953677 0.6696376
sample estimates:
      cor
0.6339621
```

**Result :** *It is observed clearly that p-value is less than 0.05 and Correlation Percentage is 63.39%. Hence, Unit.price and cogs are highly correlated and null hypothesis is rejected.*

➔ **Pearson's Product Moment Correlation for Unit.price and Rating**

```
cor.test(Unit.price,Rating,alternative = c("two.sided"), method = c("pearson"),
    exact = NULL, conf.level= 0.95, continuity = FALSE)
```

**Output :**

```
        Pearson's product-moment correlation

data:  Unit.price and Rating
t = -0.2773, df = 998, p-value = 0.7816
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.07073210  0.05324455
sample estimates:
        cor
-0.008777507
```

**Result :** *It is observed clearly that p-value is not less than 0.05 and Correlation Percentage is 0%. Hence, Unit.price and Rating are not correlated and null hypothesis is not rejected.*

➔ **Pearson's Product Moment Correlation for Quantity and Total**

```
cor.test(Quantity,Total,alternative = c("two.sided"), method = c("pearson"),
        exact = NULL, conf.level= 0.95, continuity = FALSE)
```

**Output :**

```
        Pearson's product-moment correlation

data:  Quantity and Total
t = 31.449, df = 998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6729497 0.7353418
sample estimates:
      cor
0.7055102
```

**Result :** *It is observed clearly that p-value is less than 0.05 and Correlation Percentage is 70.55%. Hence, Quantity and Total are highly correlated and null hypothesis is rejected.*

➔ **Pearson's Product Moment Correlation for Quantity and Rating**

```
cor.test(Quantity,Rating,alternative = c("two.sided"), method = c("pearson"),
        exact = NULL, conf.level= 0.95, continuity = FALSE)
```

**Output :**

```
        Pearson's product-moment correlation

data:  Quantity and Rating
t = -0.49967, df = 998, p-value = 0.6174
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.07773178  0.04622350
sample estimates:
       cor
-0.0158149
```

**Result :** *It is observed clearly that p-value is not less than 0.05 and Correlation Percentage is 0%. Hence, Quantity and Rating are not highly correlated and null hypothesis is not rejected.*

➔ **Pearson's Product Moment Correlation for Quantity and Gross Income**

```
cor.test(Quantity,gross.income,alternative = c("two.sided"), method = c("pearson"),
        exact = NULL, conf.level= 0.95, continuity = FALSE)
```

**Output :**

```
        Pearson's product-moment correlation

data:  Quantity and gross.income
t = 31.449, df = 998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6729497 0.7353418
sample estimates:
      cor
0.7055102
```

**Result :** *It is observed clearly that p-value is less than 0.05 and Correlation Percentage is 70.55%. Hence, Quantity and Gross Income are highly correlated and null hypothesis is rejected.*

➔ **Pearson's Product Moment Correlation for Total and Gross Income**

```
cor.test(Total,gross.income,alternative = c("two.sided"), method = c("pearson"),
        exact = NULL, conf.level= 0.95, continuity = FALSE)
```

**Output :**

```
        Pearson's product-moment correlation

data:  Total and gross.income
t = Inf, df = 998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 1 1
sample estimates:
cor
  1
```

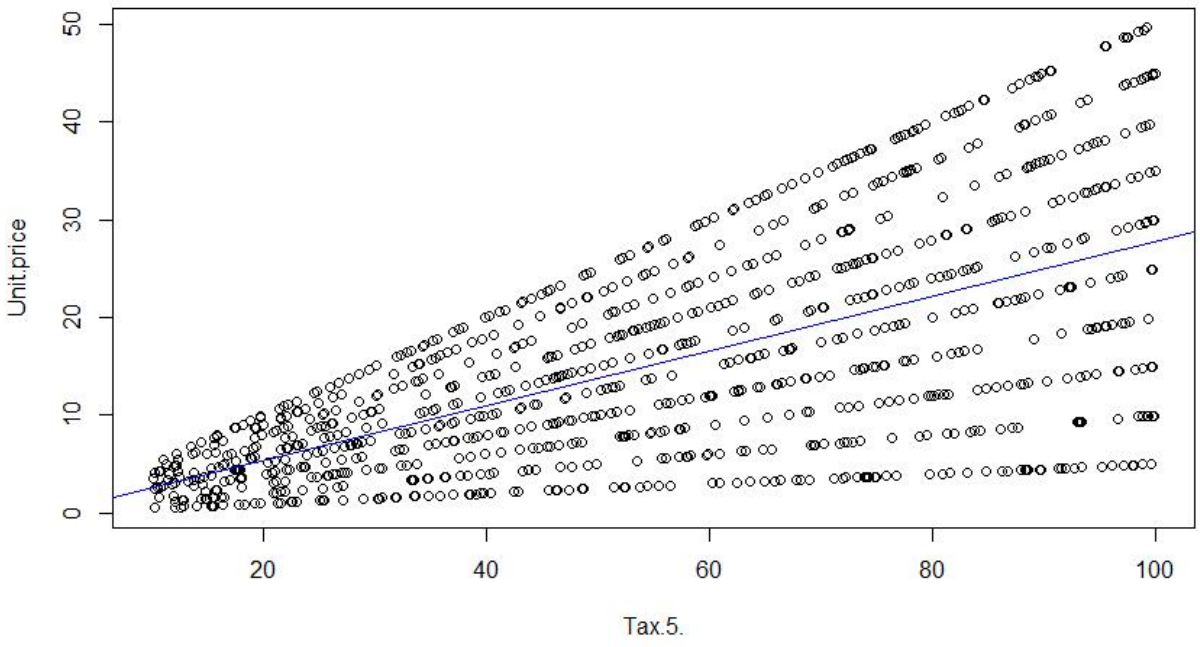
**Result :** *It is observed clearly that p-value is less than 0.05 and Correlation Percentage is 100%. Hence, Quantity and Gross Income are highly correlated and null hypothesis is rejected.*

**Scatterplot Data Visualization**

**A.) Unit.price and Tax.5.**

```
plot(Unit.price,Tax.5., xlab = 'Tax.5.', ylab = 'Unit.price')
#Scatterplot with linear regression line plotting
abline(lm(Tax.5. ~ Unit.price), col = "blue")
```

**Output:**

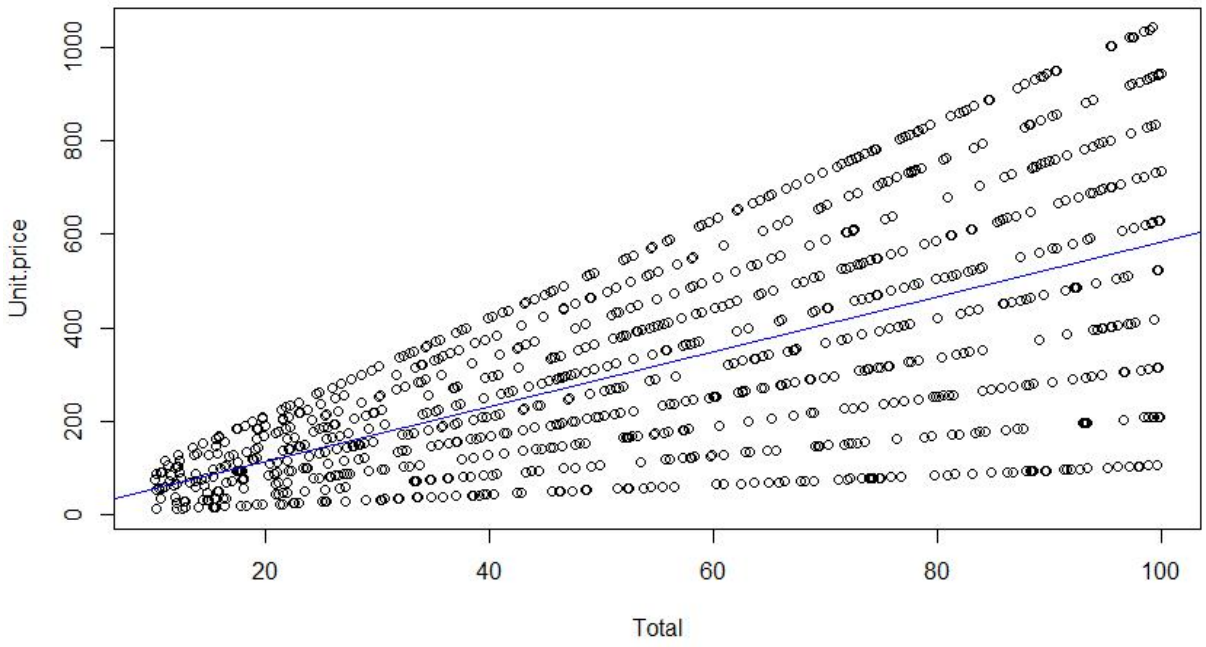



**Result:** *We can conclude that Unit.price and Tax.5. have positive correlation*

**B.) Unit.price and Total**

```
plot(Unit.price,Total, xlab = 'Total', ylab = 'Unit.price')
#Scatterplot with linear regression line plotting
abline(lm(Total ~ Unit.price), col = "blue")
```

**Output:**

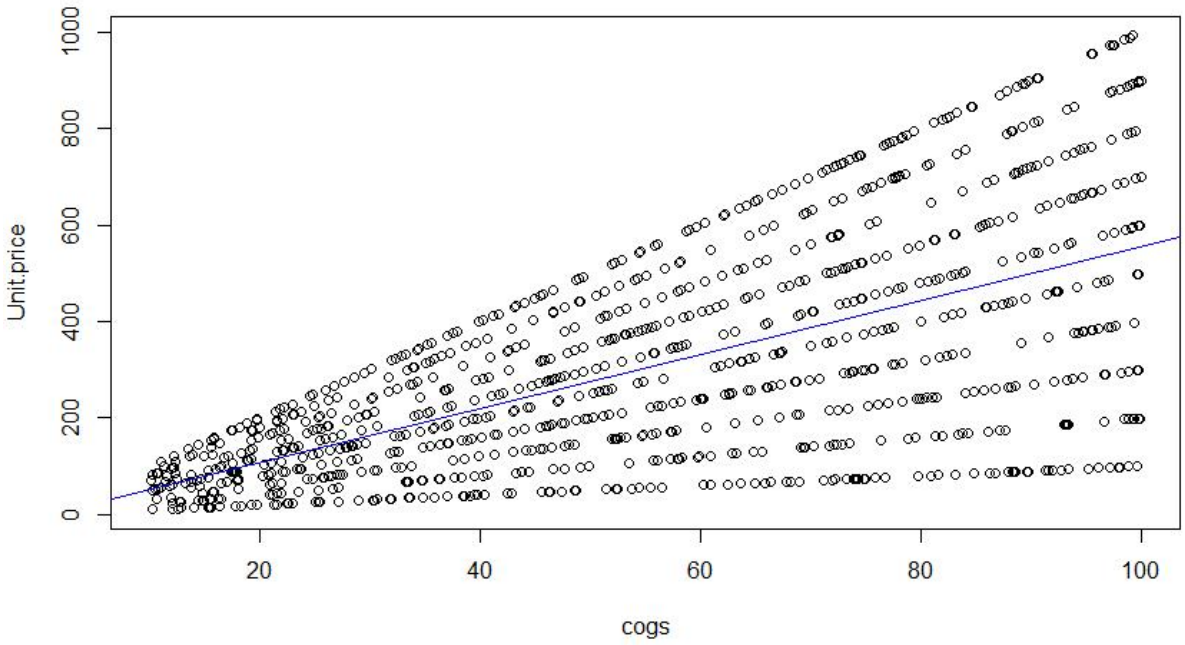



**Result:** *We can conclude that Unit.price and Total have positive correlation*

**C.) Unit.price and cogs**

```
plot(Unit.price,cogs, xlab = 'cogs', ylab = 'Unit.price')
#Scatterplot with linear regression line plotting
abline(lm(cogs ~ Unit.price), col = "blue")
```

**Output:**

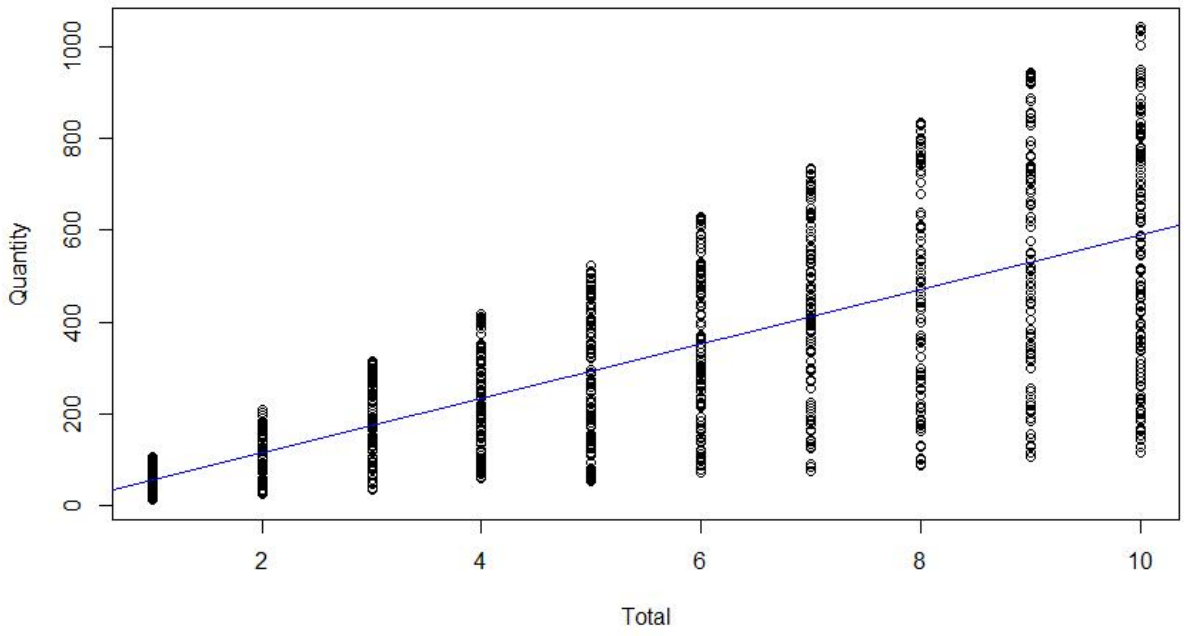



**Result:** *We can conclude that Unit.price and cogs have positive correlation*

**D.) Quantity and Total**

```
plot(Quantity,Total, xlab = 'Total', ylab = 'Quantity')
#Scatterplot with linear regression line plotting
abline(lm(Total ~ Quantity), col = "blue")
```

**Output:**

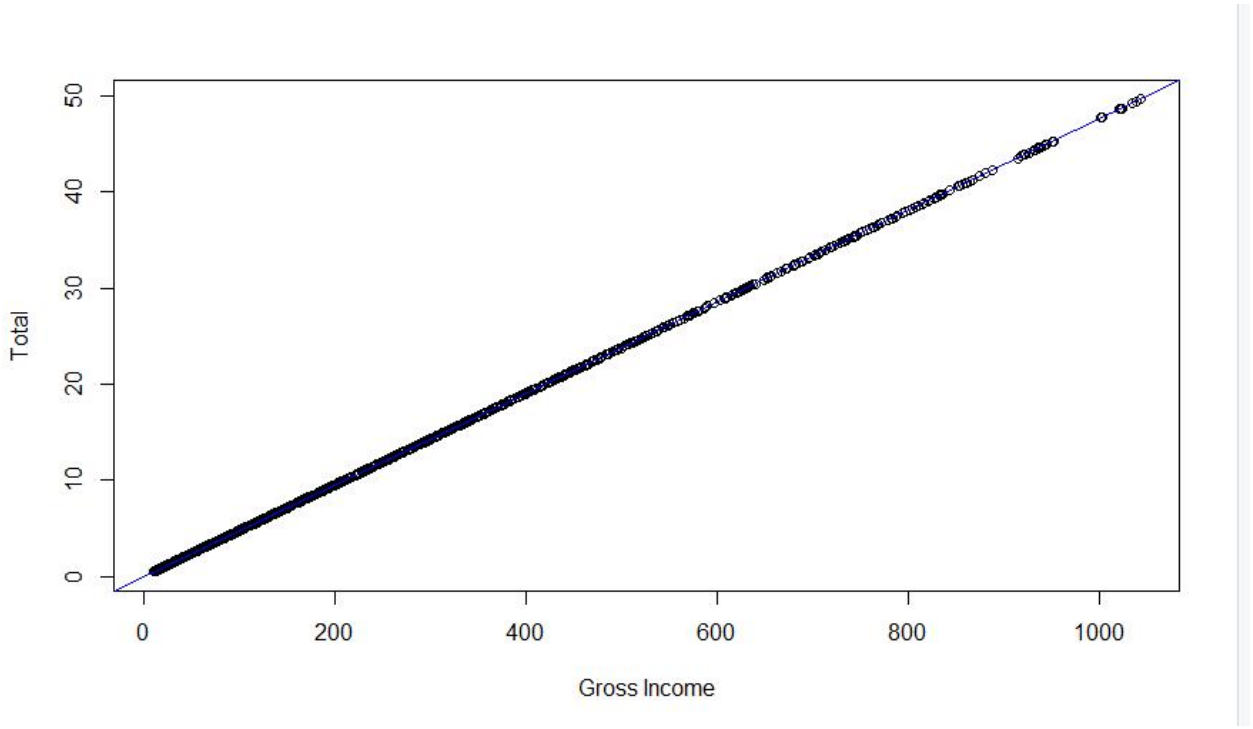



**Result:** *We can conclude that Quantity and Total have positive correlation*

**E.) Total and Gross Income**

```
plot(Total,gross.income, xlab = 'Gross Income', ylab = 'Total')
#Scatterplot with linear regression line plotting
abline(lm(gross.income ~ Total), col = "blue"
```

**Output :**




**Result :** *We can conclude that Total and Gross Income have Strong positive correlation*

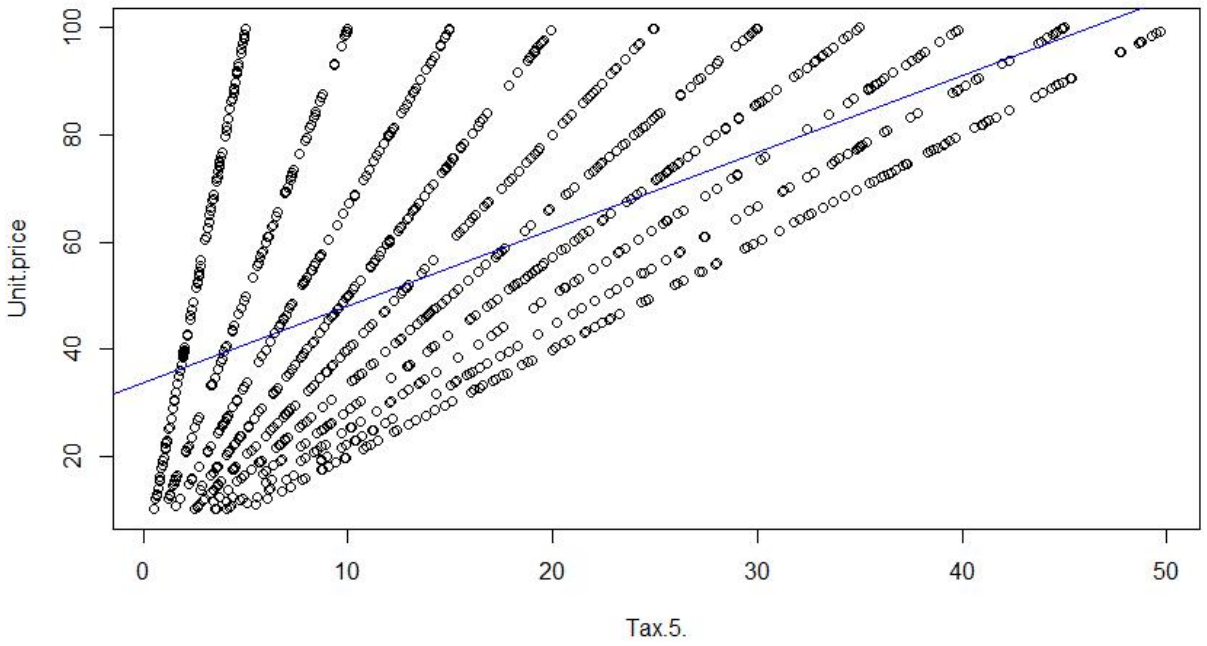
## Task 3 - REGRESSION ANALYSIS (LINEAR REGRESSION, MULTIPLE LINEAR REGRESSION, NON-LINEAR REGRESSION)

### Simple linear regression

**1. Unit.price and Tax.5**.

```
simpreg <- lm(Unit.price ~ Tax.5.)
plot(Unit.price ~ Tax.5.)
abline(simpreg, col = "blue")
summary(simpreg)
```

**Output:**



```
Call:
lm(formula = Unit.price ~ Tax.5.)

Residuals:
    Min      1Q  Median      3Q     Max
-30.511 -16.204  -4.338  12.760  58.930

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.61006    1.07053    31.4   <2e-16 ***
Tax.5.       1.43452    0.05539    25.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.5 on 998 degrees of freedom
Multiple R-squared:  0.4019,    Adjusted R-squared:  0.4013
F-statistic: 670.6 on 1 and 998 DF,  p-value: < 2.2e-16
```

**2. Unit.price and Total**

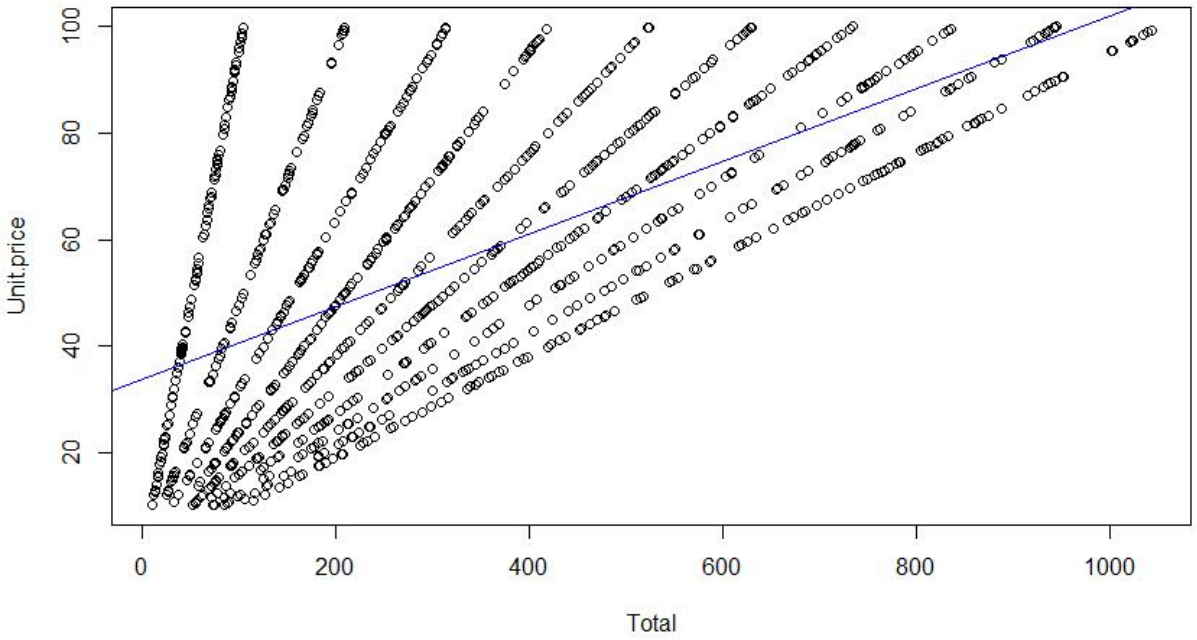
simpreg <- lm(Unit.price ~ Total)

plot(Unit.price ~ Total)

abline(simpreg, col = "blue")

summary(simpreg)

**Output:**



```
Call:
lm(formula = Unit.price ~ Total)

Residuals:
    Min      1Q  Median      3Q     Max
-30.511 -16.204  -4.338  12.760  58.930

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.610058   1.070532    31.4   <2e-16 ***
Total        0.068311   0.002638    25.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.5 on 998 degrees of freedom
Multiple R-squared:  0.4019,    Adjusted R-squared:  0.4013
F-statistic: 670.6 on 1 and 998 DF,  p-value: < 2.2e-16
```

**3. Unit.price and cogs**

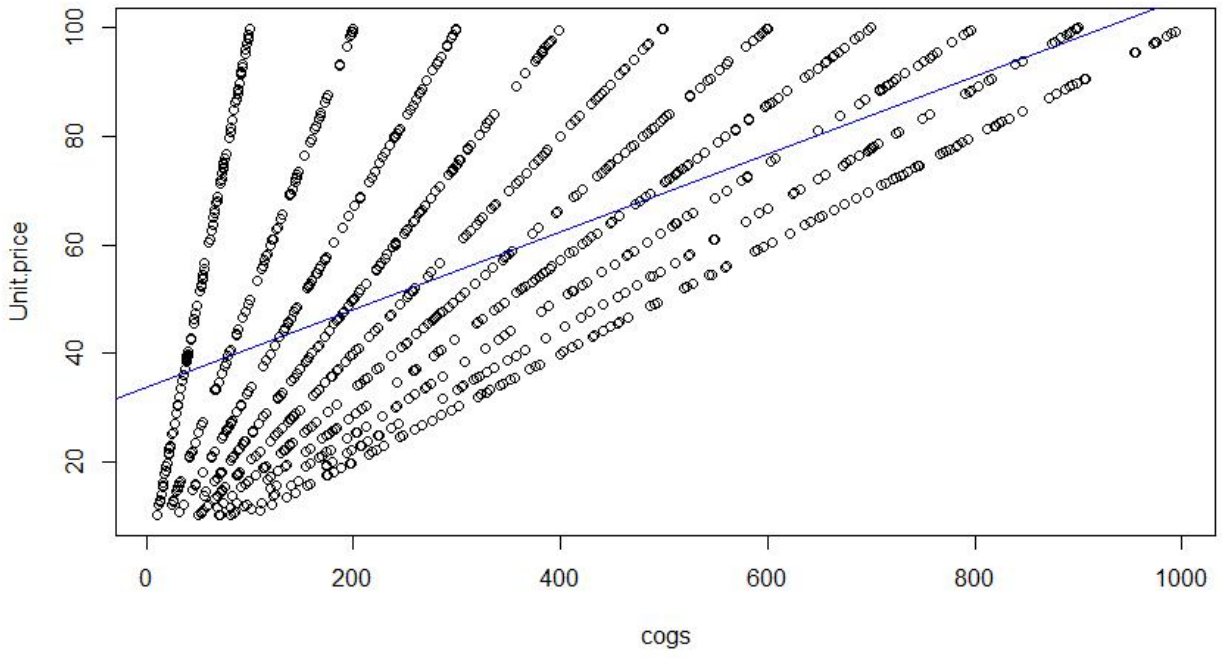
simpreg <- lm(Unit.price ~ cogs)

plot(Unit.price ~ cogs)

abline(simpreg, col = "blue")

summary(simpreg)

**Output:**



```
Call:
lm(formula = Unit.price ~ cogs)

Residuals:
    Min      1Q  Median      3Q     Max
-30.511 -16.204  -4.338  12.760  58.930

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.61006    1.07053    31.4   <2e-16 ***
cogs         0.07173    0.00277    25.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.5 on 998 degrees of freedom
Multiple R-squared:  0.4019,    Adjusted R-squared:  0.4013
F-statistic: 670.6 on 1 and 998 DF,  p-value: < 2.2e-16
```

**4. Quantity and Total**

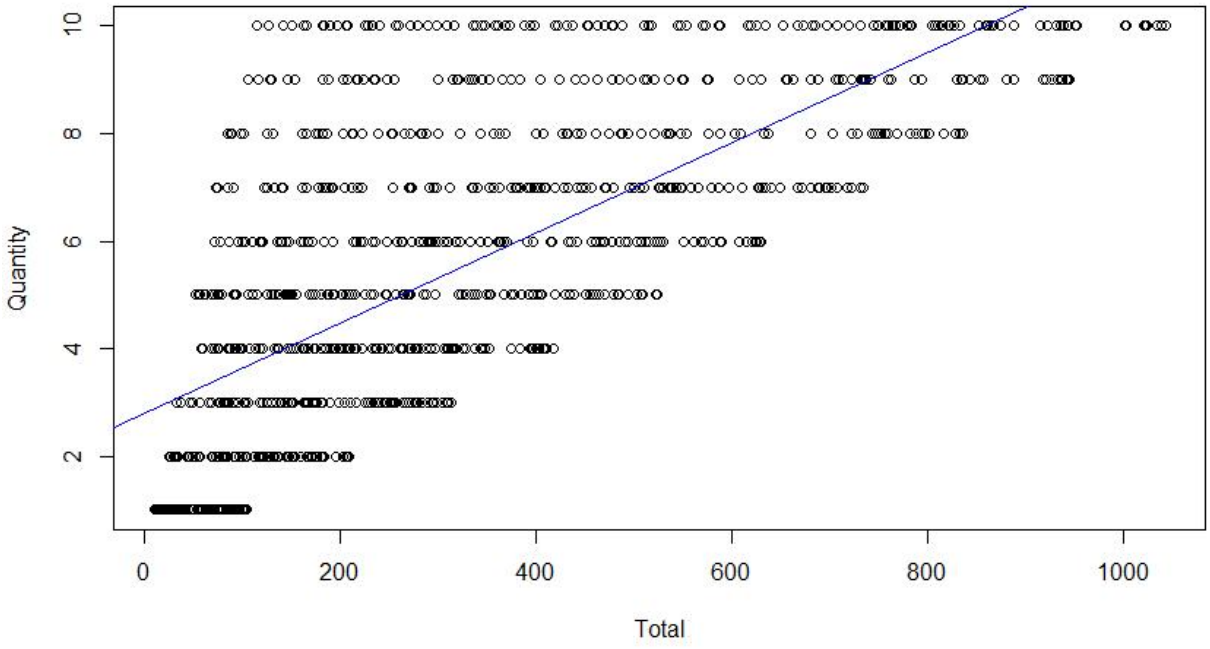
simpreg <- lm(Quantity ~ Total)

plot(Quantity ~ Total)

abline(simpreg, col = "blue")

summary(simpreg)

**Output:**



```
Call:
lm(formula = Quantity ~ Total)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6789 -1.6822 -0.5127  1.2203  6.2338

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.8009236  0.1082462   25.88   <2e-16 ***
Total       0.0083881  0.0002667   31.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.073 on 998 degrees of freedom
Multiple R-squared:  0.4977,    Adjusted R-squared:  0.4972
F-statistic:   989 on 1 and 998 DF,  p-value: < 2.2e-16
```

**5. Total and Gross Income**

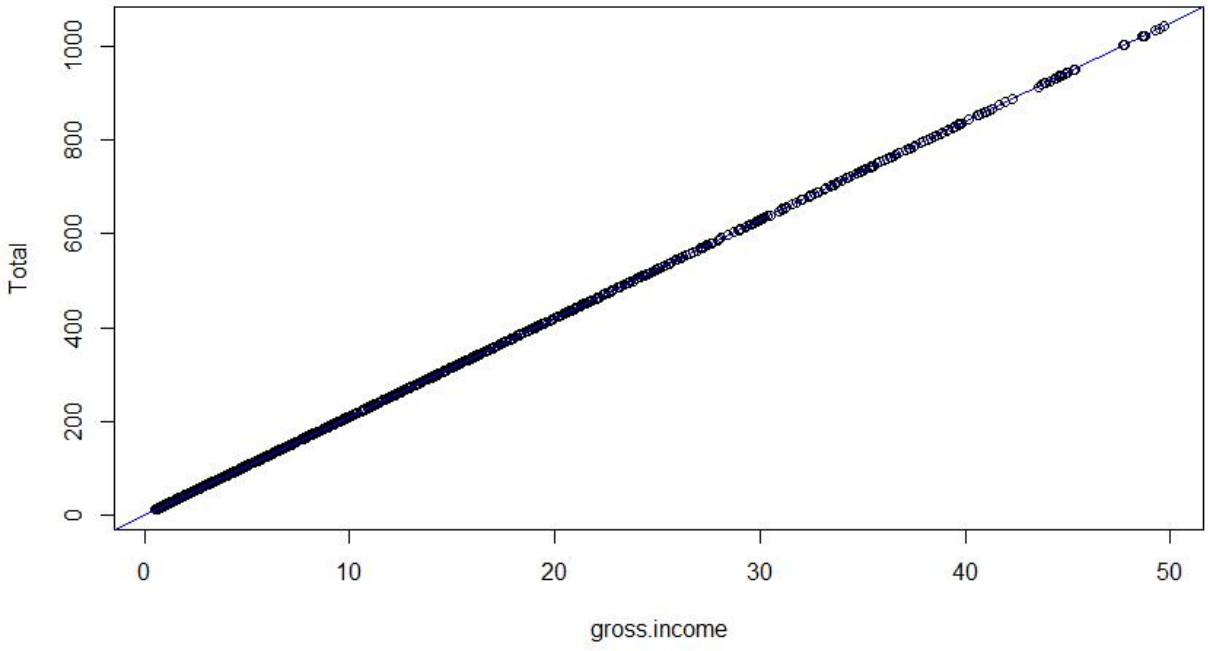
simpreg <- lm(Total ~ gross.income)

plot(Total ~ gross.income)

abline(simpreg, col = "blue")

summary(simpreg)

**Output:**



```
Call:
lm(formula = Total ~ gross.income)

Residuals:
       Min         1Q     Median         3Q        Max
-2.995e-13 -6.010e-14 -3.700e-14 -1.800e-14  2.871e-11

Coefficients:
              Estimate Std. Error   t value Pr(>|t|)
(Intercept)  1.438e-12  5.144e-14 2.795e+01   <2e-16 ***
gross.income 2.100e+01  2.662e-15 7.889e+15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.851e-13 on 998 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 6.224e+31 on 1 and 998 DF,  p-value: < 2.2e-16
```

## Multiple Linear Regression

### 1. Unit.price and Tax.5. with Quantity

reg <- lm(Unit.price ~ Tax.5. + Quantity)

summary(reg)

### Output :

```
Call:
lm(formula = Unit.price ~ Tax.5. + Quantity)

Residuals:
    Min      1Q  Median      3Q     Max
-39.059  -6.469   0.101   6.211  37.830

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  55.6705     0.8373   66.49   <2e-16 ***
Tax.5.        2.8219     0.0473   59.67   <2e-16 ***
Quantity     -7.8761     0.1894  -41.58   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.4 on 997 degrees of freedom
Multiple R-squared:  0.7812,    Adjusted R-squared:  0.7808
F-statistic:  1780 on 2 and 997 DF,  p-value: < 2.2e-16
```

**2. Unit.price and Tax.5. with Total and cogs**

```
reg <- lm(Unit.price ~ Tax.5. + Total + cogs)
summary(reg)
```

**Output:**

```
Call:
lm(formula = Unit.price ~ Tax.5. + Total + cogs)

Residuals:
    Min      1Q  Median      3Q     Max
-30.511 -16.204  -4.338  12.760  58.930

Coefficients: (2 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.61006    1.07053    31.4   <2e-16 ***
Tax.5.       1.43452    0.05539    25.9   <2e-16 ***
Total             NA         NA      NA       NA
cogs              NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.5 on 998 degrees of freedom
Multiple R-squared:  0.4019,    Adjusted R-squared:  0.4013
F-statistic: 670.6 on 1 and 998 DF,  p-value: < 2.2e-16
```

**3. Quantity and cogs with Total and rating**

```
reg <- lm(Quantity ~ cogs + Total + Rating)
summary(reg)
```

**Output:**

```
Call:
lm(formula = Quantity ~ cogs + Total + Rating)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6967 -1.6930 -0.5175  1.2383  6.2511

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.6820146  0.2904540   9.234   <2e-16 ***
cogs        0.0088120  0.0002804  31.431   <2e-16 ***
Total              NA         NA      NA       NA
Rating      0.0168547  0.0382018   0.441    0.659
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.074 on 997 degrees of freedom
Multiple R-squared:  0.4978,    Adjusted R-squared:  0.4968
F-statistic: 494.2 on 2 and 997 DF,  p-value: < 2.2e-16
```

# Conclusion

From the above observations, it can be concluded that the variables Unit.price and Quantity are highly dependent variables while Total, gross.income, Tax.5. etc are independent variables.

We have clearly analyzed the variables using Bi-variate Analysis and Correlation Analysis along-with Linear Regression and Multiple Linear Regression after specifying the descriptive statistics of the dataset.

###Created by :- Dhruv Dixit - United College of Engineering and Management ###