R Data Analysis Markdown

This analysis leverages R Programming to explore a dataset through data manipulation, statistical evaluation, and visualization techniques. It includes data cleaning, variable identification, and the creation of new variables through mathematical transformations. Statistical functions such as mean, median, and mode are calculated, along with data visualization through scatter and bar plots. The study also employs linear regression to examine correlations between variables, offering insights into the dataset's dynamics and patterns.

#Print the structure

```
data <- read.csv("C:/Users/sathi/Downloads/HumberMart.csv")
str(data)
```

```
## 'data.frame':    499 obs. of  12 variables:
##  $ Date         : chr  "01-05-2019" "03-08-2019" "03-03-2019" "1/27/2019" ...
##  $ Invoice.ID   : chr  "750-67-8428" "226-31-3081" "631-41-3108" "123-19-1176" ...
##  $ City         : chr  "Toronto" "Vancouver" "Toronto" "Toronto" ...
##  $ Customer.type: chr  "Member" "Normal" "Normal" "Member" ...
##  $ Gender       : chr  "Female" "Female" "Female" "Male" ...
##  $ Product.line : chr  "Health and beauty" "Electronic accessories" "Home and lifestyle" "Health and
##  $ Unit.price   : num  74.7 15.3 46.3 58.2 86.3 ...
##  $ Quantity     : int  7 5 7 8 7 7 6 10 2 3 ...
##  $ Tax          : num  26.14 3.82 16.22 23.29 30.21 ...
##  $ Payment      : chr  "Ewallet" "Cash" "Credit card" "Ewallet" ...
##  $ cogs         : num  522.8 76.4 324.3 465.8 604.2 ...
##  $ Rating       : num  9.1 9.6 7.4 8.4 5.3 4.1 5.8 8 7.2 5.9 ...
```

#list the variables

```
columns <- colnames(data)
print(columns)
```

```
##  [1] "Date"          "Invoice.ID"    "City"          "Customer.type"
##  [5] "Gender"        "Product.line"  "Unit.price"    "Quantity"
##  [9] "Tax"           "Payment"       "cogs"          "Rating"
```

#Top 15 rows in the dataset

```
top_rows <- head(data, 15)
show(top_rows)
```

```
##          Date  Invoice.ID      City Customer.type Gender           Product.line
## 1  01-05-2019 750-67-8428   Toronto        Member Female      Health and beauty
## 2  03-08-2019 226-31-3081 Vancouver        Normal Female Electronic accessories
## 3  03-03-2019 631-41-3108   Toronto        Normal Female     Home and lifestyle
## 4   1/27/2019 123-19-1176   Toronto        Member   Male      Health and beauty
## 5  02-08-2019 373-73-7910   Toronto        Normal   Male      Sports and travel
## 6   3/25/2019 699-14-3026 Vancouver        Normal Female Electronic accessories
## 7   2/25/2019 355-53-5943   Toronto        Member Female Electronic accessories
## 8   2/24/2019 315-22-5665 Vancouver        Normal Female     Home and lifestyle
## 9  01-10-2019 665-32-9167   Toronto        Member Female      Health and beauty
```

```
## 10  2/20/2019 692-92-5582  Montreal        Member Female     Food and beverages
## 11 02-06-2019 351-62-0822  Montreal        Member Female    Fashion accessories
## 12 03-09-2019 529-56-3974  Montreal        Member Female Electronic accessories
## 13 02-12-2019 365-64-0515   Toronto        Normal Female Electronic accessories
## 14 02-07-2019 252-56-2699   Toronto        Normal Female     Food and beverages
## 15  3/29/2019 829-34-3910   Toronto        Normal Female      Health and beauty
##    Unit.price Quantity     Tax     Payment   cogs Rating
## 1       74.69        7 26.1415     Ewallet 522.83    9.1
## 2       15.28        5  3.8200        Cash  76.40    9.6
## 3       46.33        7 16.2155 Credit card 324.31    7.4
## 4       58.22        8 23.2880     Ewallet 465.76    8.4
## 5       86.31        7 30.2085     Ewallet 604.17    5.3
## 6       85.39        7 29.8865     Ewallet 597.73    4.1
## 7       68.84        6 20.6520     Ewallet 413.04    5.8
## 8       73.56       10 36.7800     Ewallet 735.60    8.0
## 9       36.26        2  3.6260 Credit card  72.52    7.2
## 10      54.84        3  8.2260 Credit card 164.52    5.9
## 11      14.48        4  2.8960     Ewallet  57.92    4.5
## 12      25.51        4  5.1020        Cash 102.04    6.8
## 13      46.95        5 11.7375     Ewallet 234.75    7.1
## 14      43.19       10 21.5950     Ewallet 431.90    8.2
## 15      71.38       10 35.6900        Cash 713.80    5.7
```

#User defined function

```
filter_product_lines <- function(data, target_city, min_rating = 9) {
  filtered_lines <- data[data$City == target_city & data$Rating > min_rating, "Product.line"]
  return(filtered_lines)
}

target_city <- "Toronto"

high_rating_lines_in_city <- filter_product_lines(data, target_city)

print(unique(high_rating_lines_in_city))
```

```
## [1] "Health and beauty"      "Electronic accessories" "Home and lifestyle"
## [4] "Food and beverages"     "Sports and travel"
```

#Filtering data

```
library(dplyr)
filteredData <- filter(data, Quantity > 5)
head(filteredData)
```

```
##         Date  Invoice.ID      City Customer.type Gender           Product.line
## 1 01-05-2019 750-67-8428   Toronto        Member Female      Health and beauty
## 2 03-03-2019 631-41-3108   Toronto        Normal Female     Home and lifestyle
## 3  1/27/2019 123-19-1176   Toronto        Member   Male      Health and beauty
## 4 02-08-2019 373-73-7910   Toronto        Normal   Male      Sports and travel
## 5  3/25/2019 699-14-3026 Vancouver        Normal Female Electronic accessories
## 6  2/25/2019 355-53-5943   Toronto        Member Female Electronic accessories
```

```
##   Unit.price Quantity     Tax     Payment   cogs Rating
## 1      74.69        7 26.1415     Ewallet 522.83    9.1
## 2      46.33        7 16.2155 Credit card 324.31    7.4
## 3      58.22        8 23.2880     Ewallet 465.76    8.4
## 4      86.31        7 30.2085     Ewallet 604.17    5.3
## 5      85.39        7 29.8865     Ewallet 597.73    4.1
## 6      68.84        6 20.6520     Ewallet 413.04    5.8
```

#Reshaping Data

```
data$TotalSales <- data$Unit.price * data$Quantity + data$Tax

aggregated_df <- data %>%
  group_by(Customer.type, Product.line, Payment) %>%
  summarise(TotalSales = sum(TotalSales), .groups = 'drop')
head(aggregated_df)
```

```
## # A tibble: 6 x 4
##   Customer.type Product.line           Payment     TotalSales
##   <chr>         <chr>                  <chr>            <dbl>
## 1 Member        Electronic accessories Cash             6487.
## 2 Member        Electronic accessories Credit card      4140.
## 3 Member        Electronic accessories Ewallet          3986.
## 4 Member        Fashion accessories    Cash             1332.
## 5 Member        Fashion accessories    Credit card      2483.
## 6 Member        Fashion accessories    Ewallet          3135.
```

```
library(tidyr)


wide_df <- aggregated_df %>%
  pivot_wider(names_from = Product.line, values_from = TotalSales)
head(wide_df)
```

```
## # A tibble: 6 x 8
##   Customer.type Payment     `Electronic accessories` `Fashion accessories`
##   <chr>         <chr>                          <dbl>                 <dbl>
## 1 Member        Cash                           6487.                 1332.
## 2 Member        Credit card                    4140.                 2483.
## 3 Member        Ewallet                        3986.                 3135.
## 4 Normal        Cash                           7118.                 3122.
## 5 Normal        Credit card                    4373.                 1410.
## 6 Normal        Ewallet                        3809.                 2262.
## # i 4 more variables: `Food and beverages` <dbl>, `Health and beauty` <dbl>,
## #   `Home and lifestyle` <dbl>, `Sports and travel` <dbl>
```

#Performing Join to append Avg_City_Rating

```
library(dplyr)
library(tidyr)

avg_rating_df <- data %>%
```

```
  group_by(City, Product.line) %>%
  summarise(Avg_Rating_For_City = mean(Rating, na.rm = TRUE), .groups = 'drop')


df_with_avg_rating <- data %>%
  left_join(avg_rating_df, by = c("City", "Product.line"))

head(df_with_avg_rating)
```

```
##         Date  Invoice.ID      City Customer.type Gender           Product.line
## 1 01-05-2019 750-67-8428   Toronto        Member Female      Health and beauty
## 2 03-08-2019 226-31-3081 Vancouver        Normal Female Electronic accessories
## 3 03-03-2019 631-41-3108   Toronto        Normal Female     Home and lifestyle
## 4  1/27/2019 123-19-1176   Toronto        Member   Male      Health and beauty
## 5 02-08-2019 373-73-7910   Toronto        Normal   Male      Sports and travel
## 6  3/25/2019 699-14-3026 Vancouver        Normal Female Electronic accessories
##   Unit.price Quantity     Tax     Payment   cogs Rating TotalSales
## 1      74.69        7 26.1415     Ewallet 522.83    9.1   548.9715
## 2      15.28        5  3.8200        Cash  76.40    9.6    80.2200
## 3      46.33        7 16.2155 Credit card 324.31    7.4   340.5255
## 4      58.22        8 23.2880     Ewallet 465.76    8.4   489.0480
## 5      86.31        7 30.2085     Ewallet 604.17    5.3   634.3785
## 6      85.39        7 29.8865     Ewallet 597.73    4.1   627.6165
##   Avg_Rating_For_City
## 1            6.946667
## 2            7.151429
## 3            6.807500
## 4            6.946667
## 5            7.195000
## 6            7.151429
```

#Omit Missing values

```
data <- na.omit(data)
head(data)
```

```
##         Date  Invoice.ID      City Customer.type Gender           Product.line
## 1 01-05-2019 750-67-8428   Toronto        Member Female      Health and beauty
## 2 03-08-2019 226-31-3081 Vancouver        Normal Female Electronic accessories
## 3 03-03-2019 631-41-3108   Toronto        Normal Female     Home and lifestyle
## 4  1/27/2019 123-19-1176   Toronto        Member   Male      Health and beauty
## 5 02-08-2019 373-73-7910   Toronto        Normal   Male      Sports and travel
## 6  3/25/2019 699-14-3026 Vancouver        Normal Female Electronic accessories
##   Unit.price Quantity     Tax     Payment   cogs Rating TotalSales
## 1      74.69        7 26.1415     Ewallet 522.83    9.1   548.9715
## 2      15.28        5  3.8200        Cash  76.40    9.6    80.2200
## 3      46.33        7 16.2155 Credit card 324.31    7.4   340.5255
## 4      58.22        8 23.2880     Ewallet 465.76    8.4   489.0480
## 5      86.31        7 30.2085     Ewallet 604.17    5.3   634.3785
## 6      85.39        7 29.8865     Ewallet 597.73    4.1   627.6165
```

#Identify and remove duplicated data

```r
data <- data[!duplicated(data), ]
head(data)
```

```
##         Date  Invoice.ID      City Customer.type Gender           Product.line
## 1 01-05-2019 750-67-8428   Toronto        Member Female      Health and beauty
## 2 03-08-2019 226-31-3081 Vancouver        Normal Female Electronic accessories
## 3 03-03-2019 631-41-3108   Toronto        Normal Female     Home and lifestyle
## 4  1/27/2019 123-19-1176   Toronto        Member   Male      Health and beauty
## 5 02-08-2019 373-73-7910   Toronto        Normal   Male      Sports and travel
## 6  3/25/2019 699-14-3026 Vancouver        Normal Female Electronic accessories
##   Unit.price Quantity     Tax     Payment   cogs Rating TotalSales
## 1      74.69        7 26.1415     Ewallet 522.83    9.1   548.9715
## 2      15.28        5  3.8200        Cash  76.40    9.6    80.2200
## 3      46.33        7 16.2155 Credit card 324.31    7.4   340.5255
## 4      58.22        8 23.2880     Ewallet 465.76    8.4   489.0480
## 5      86.31        7 30.2085     Ewallet 604.17    5.3   634.3785
## 6      85.39        7 29.8865     Ewallet 597.73    4.1   627.6165
```

#Sorting Quantity in Descending order

```r
data <- data[order(-data$Rating), ]
head(data)
```

```
##           Date  Invoice.ID      City Customer.type Gender
## 61   2/15/2019 285-68-5083 Vancouver        Member Female
## 63  02-03-2019 347-34-2234  Montreal        Member Female
## 160  3/27/2019 423-57-2993  Montreal        Normal Female
## 388  2/20/2019 725-56-0833   Toronto        Normal Female
## 24   2/17/2019 636-48-8204   Toronto        Normal Female
## 68  01-07-2019 109-28-2512  Montreal        Member Female
##               Product.line Unit.price Quantity     Tax     Payment   cogs
## 61      Home and lifestyle      24.74        3  3.7110 Credit card  74.22
## 63      Home and lifestyle      55.07        9 24.7815     Ewallet 495.63
## 160      Sports and travel      93.39        6 28.0170     Ewallet 560.34
## 388      Health and beauty      32.32       10 16.1600 Credit card 323.20
## 24  Electronic accessories      34.56        5  8.6400     Ewallet 172.80
## 68     Fashion accessories      97.61        6 29.2830     Ewallet 585.66
##     Rating TotalSales
## 61    10.0    77.9310
## 63    10.0   520.4115
## 160   10.0   588.3570
## 388   10.0   339.3600
## 24     9.9   181.4400
## 68     9.9   614.9430
```

#Renaming columns

```r
colnames(data)[colnames(data) == "Customer.type"] <- "Customer_type"
colnames(data)[colnames(data) == "Unit.price"] <- "Unit_price"
colnames(data)[colnames(data) == "Invoice.ID"] <- "Invoice_ID"
columns <- colnames(data)
print(columns)
```

```
##  [1] "Date"          "Invoice_ID"    "City"          "Customer_type"
##  [5] "Gender"        "Product.line"  "Unit_price"    "Quantity"
##  [9] "Tax"           "Payment"       "cogs"          "Rating"
## [13] "TotalSales"
```

#Creating a new variable

```
data$DoubleUnitPrice <- data$Unit_price * 2
head(data)
```

```
##           Date  Invoice_ID      City Customer_type Gender
## 61   2/15/2019 285-68-5083 Vancouver        Member Female
## 63  02-03-2019 347-34-2234  Montreal        Member Female
## 160  3/27/2019 423-57-2993  Montreal        Normal Female
## 388  2/20/2019 725-56-0833   Toronto        Normal Female
## 24   2/17/2019 636-48-8204   Toronto        Normal Female
## 68  01-07-2019 109-28-2512  Montreal        Member Female
##               Product.line Unit_price Quantity     Tax     Payment   cogs
## 61      Home and lifestyle      24.74        3  3.7110 Credit card  74.22
## 63      Home and lifestyle      55.07        9 24.7815     Ewallet 495.63
## 160      Sports and travel      93.39        6 28.0170     Ewallet 560.34
## 388      Health and beauty      32.32       10 16.1600 Credit card 323.20
## 24  Electronic accessories      34.56        5  8.6400     Ewallet 172.80
## 68     Fashion accessories      97.61        6 29.2830     Ewallet 585.66
##     Rating TotalSales DoubleUnitPrice
## 61    10.0    77.9310           49.48
## 63    10.0   520.4115          110.14
## 160   10.0   588.3570          186.78
## 388   10.0   339.3600           64.64
## 24     9.9   181.4400           69.12
## 68     9.9   614.9430          195.22
```

#Traning set using random number generator

```
set.seed(123)
trainingIndex <- sample(1:nrow(data), 0.8 * nrow(data)) # 80% for training
trainingSet <- data[trainingIndex, ]
head(trainingSet)
```

```
##           Date  Invoice_ID      City Customer_type Gender       Product.line
## 209  3/28/2019 573-58-9734  Montreal        Normal Female Food and beverages
## 23   3/15/2019 273-16-6619  Montreal        Normal   Male Home and lifestyle
## 208  3/18/2019 263-87-5680 Vancouver        Member Female Home and lifestyle
## 383  1/14/2019 868-52-7573  Montreal        Normal Female Food and beverages
## 44  03-04-2019 228-96-1411 Vancouver        Member Female Food and beverages
## 213  3/20/2019 142-63-6033  Montreal        Normal   Male Home and lifestyle
##     Unit_price Quantity     Tax     Payment   cogs Rating TotalSales
## 209      30.37        3  4.5555     Ewallet  91.11    5.1    95.6655
## 23       33.20        2  3.3200 Credit card  66.40    4.4    69.7200
## 208      28.53       10 14.2650     Ewallet 285.30    7.8   299.5650
## 383      99.69        5 24.9225        Cash 498.45    9.9   523.3725
## 44       98.70        8 39.4800        Cash 789.60    7.6   829.0800
```

```
## 213      92.36        5 23.0900     Ewallet 461.80    4.9   484.8900
##     DoubleUnitPrice
## 209           60.74
## 23            66.40
## 208           57.06
## 383          199.38
## 44           197.40
## 213          184.72
```

#Summary statistics of the dataset

```
summary(data)
```

```
##      Date              Invoice_ID            City           Customer_type
##  Length:499         Length:499         Length:499         Length:499
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     Gender           Product.line         Unit_price       Quantity
##  Length:499         Length:499         Min.   :10.59   Min.   : 1.000
##  Class :character   Class :character   1st Qu.:30.51   1st Qu.: 3.000
##  Mode  :character   Mode  :character   Median :52.59   Median : 6.000
##                                        Mean   :54.86   Mean   : 5.689
##                                        3rd Qu.:77.83   3rd Qu.: 8.000
##                                        Max.   :99.96   Max.   :10.000
##       Tax            Payment               cogs            Rating
##  Min.   : 0.627   Length:499         Min.   : 12.54   Min.   : 4.000
##  1st Qu.: 6.413   Class :character   1st Qu.:128.27   1st Qu.: 5.600
##  Median :12.835   Mode  :character   Median :256.70   Median : 7.000
##  Mean   :15.714                      Mean   :314.29   Mean   : 7.013
##  3rd Qu.:22.923                      3rd Qu.:458.45   3rd Qu.: 8.450
##  Max.   :49.980                      Max.   :999.60   Max.   :10.000
##    TotalSales      DoubleUnitPrice
##  Min.   :  13.17   Min.   : 21.18
##  1st Qu.: 134.68   1st Qu.: 61.02
##  Median : 269.54   Median :105.18
##  Mean   : 330.00   Mean   :109.71
##  3rd Qu.: 481.38   3rd Qu.:155.65
##  Max.   :1049.58   Max.   :199.92
```

Performing Statistical OPerations: Mean, median, mode, range

```
mean(data$Quantity)
```

```
## [1] 5.689379
```

```r
median(data$Quantity)
```

```
## [1] 6
```

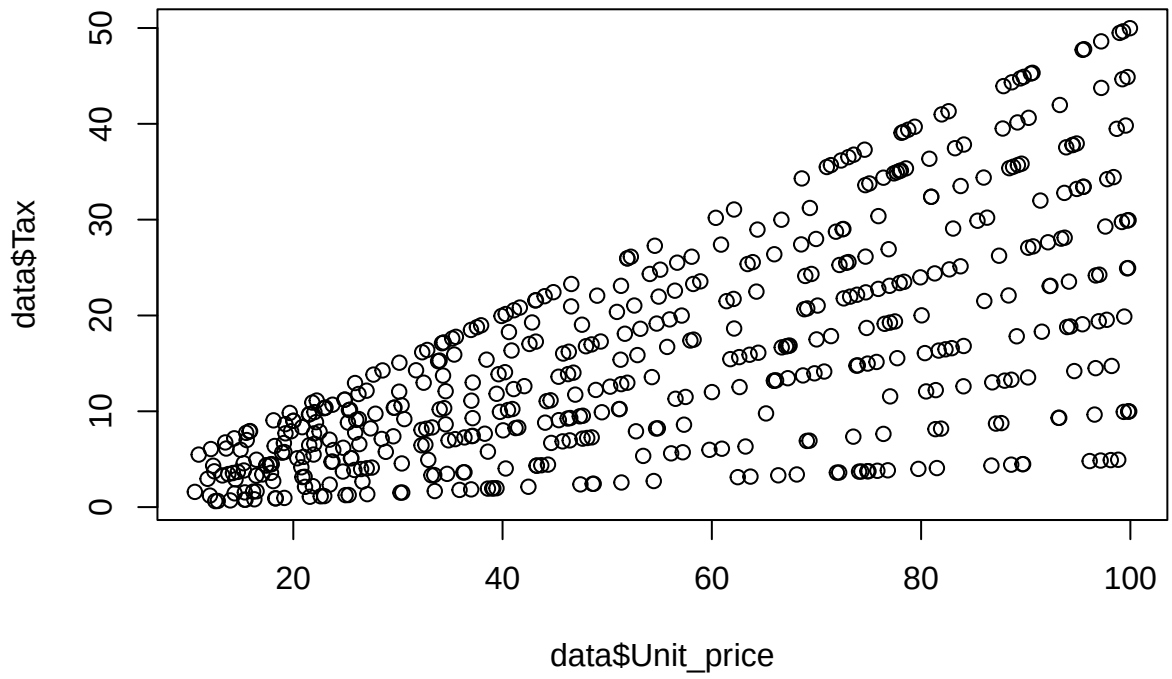
```r
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
Mode(data$Quantity)
```

```
## [1] 10
```

```r
range(data$Quantity)
```

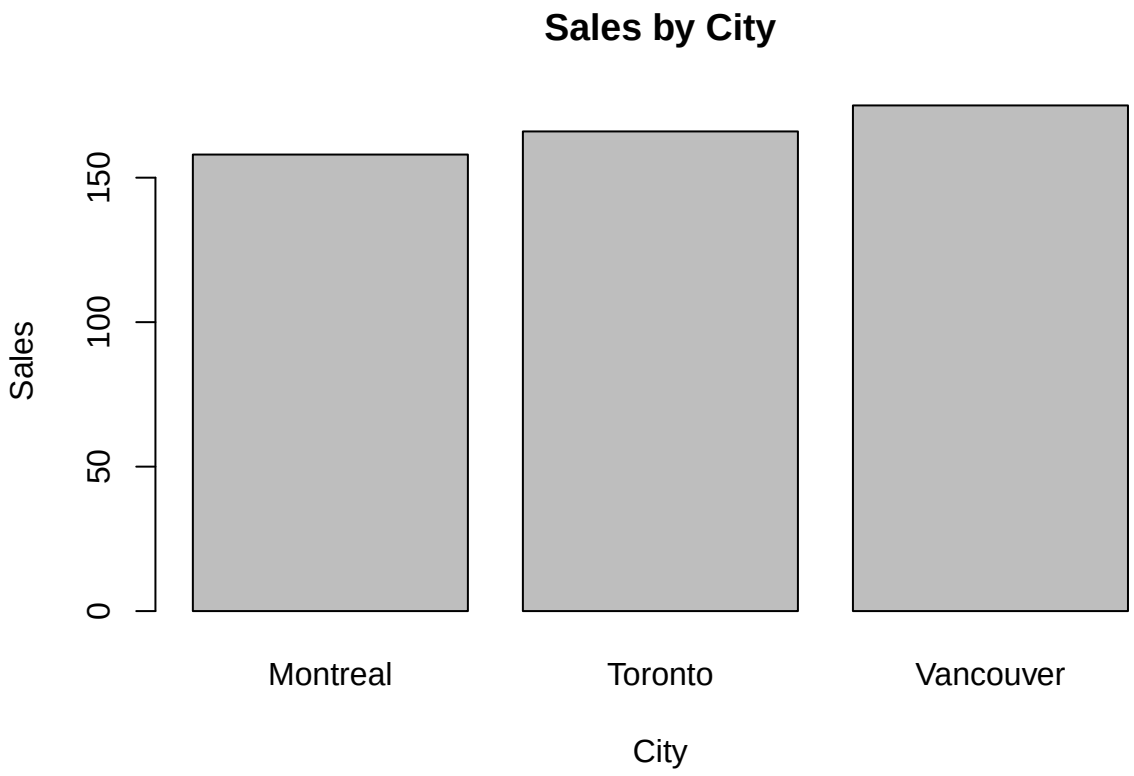
```
## [1]  1 10
```

```r
plot(data$Unit_price,data$Tax)
```



#Barplot Sales By City

```r
barplot(table(data$City), main="Sales by City", xlab="City", ylab="Sales")
```



#Correlation & linear regression model

```
cor(data$Unit_price, data$Rating)
```

```
## [1] -0.0162657
```

```
model <- lm(Rating ~ Unit_price, data=data)
summary(model)
```

```
##
## Call:
## lm(formula = Rating ~ Unit_price, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.04585 -1.44892 -0.02823  1.43856  3.02661
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.070428   0.174987  40.405   <2e-16 ***
## Unit_price  -0.001039   0.002865  -0.363    0.717
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.718 on 497 degrees of freedom
```

```
## Multiple R-squared:  0.0002646,  Adjusted R-squared:  -0.001747
## F-statistic: 0.1315 on 1 and 497 DF,  p-value: 0.717
```