# BA Homework 1

*Devarshi Pancholi*

*9/9/2019*

***PROBLEM 1***

**You are given the closing stock prices of 4 companies for one year. With the help of the what you are being taught in the first class (basic arithmetic in R) answer the following questions.**

*Q1. Compute the average price of each company's share for the given year.*
R CODE:

```
library(readr)
Stocks<-read_csv(file= "/Users/devarshipancholi/Desktop/Stock prices HMK1.csv")
```

```
## Parsed with column specification:
## cols(
##   Date = col_character(),
##   AMZN = col_double(),
##   KMX = col_double(),
##   GOOG = col_double(),
##   GE = col_double()
## )
```

```
mean(Stocks$AMZN)
```

```
## [1] 977.5859
```

```
mean(Stocks$KMX)
```

```
## [1] 65.83
```

```
mean(Stocks$GOOG)
```

```
## [1] 928.3296
```

```
mean(Stocks$GE)
```

```
## [1] 25.67647
```

*Q2. What are the data types of all the variables in the dataset?*
R CODE:

```
class(Stocks$Date)
```

```
## [1] "character"
```

```
class(Stocks$AMZN)
```

```
## [1] "numeric"
```

```
class(Stocks$KMX)
```

```
## [1] "numeric"
```

```
class(Stocks$GOOG)
```

```
## [1] "numeric"
```

```
class(Stocks$GE)
```

```
## [1] "numeric"
```

*Q3. Calculate the returns for each company's share for the given year on daily basis.*
R CODE:

```
Stocks$RAMZN <- 0
Stocks$RKMX <- 0
Stocks$RGOOG <- 0
Stocks$RGE <- 0
for(a in 2:nrow(Stocks)){
  Stocks$RAMZN[a]<-(Stocks$AMZN[a]-Stocks$AMZN[a-1])/Stocks$AMZN[a-1]
  Stocks$RKMX[a]<-(Stocks$KMX[a]-Stocks$KMX[a-1])/Stocks$KMX[a-1]
  Stocks$RGOOG[a]<-(Stocks$GOOG[a]-Stocks$GOOG[a-1])/Stocks$GOOG[a-1]
  Stocks$RGE[a]<-(Stocks$GE[a]-Stocks$GE[a-1])/Stocks$GE[a-1]
}
head(Stocks,10)
```

```
## # A tibble: 10 x 9
##    Date        AMZN   KMX  GOOG    GE    RAMZN     RKMX    RGOOG      RGE
##    <chr>      <dbl> <dbl> <dbl> <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
##  1 29/12/2017 1169.  64.1 1046.  17.4  0        0        0        0
##  2 28/12/2017 1186.  64.8 1048.  17.4  0.0142   0.0106   0.00166 -0.00516
##  3 27/12/2017 1182.  64.9 1049.  17.4 -0.00324  0.00154  0.00117  0.00115
##  4 26/12/2017 1177.  65.6 1057.  17.4 -0.00465  0.00986  0.00702  0.00288
##  5 22/12/2017 1168.  65.2 1060.  17.5 -0.00714 -0.00488  0.00320  0.00402
##  6 21/12/2017 1175.  66   1064.  17.5  0.00548  0.0118   0.00331 -0.00171
##  7 20/12/2017 1178.  68.5 1065.  17.4  0.00243  0.0374   0.00124 -0.00114
##  8 19/12/2017 1187.  67.8 1071.  17.6  0.00829 -0.00906  0.00538  0.00802
##  9 18/12/2017 1191.  68.5 1077.  17.8  0.00270  0.00987  0.00603  0.00966
## 10 15/12/2017 1179.  67.7 1064.  17.8 -0.00961 -0.0115  -0.0120   0.00338
```

*Q4. Calculate the cumulative returns for each company's share for the given year.*
R CODE:

```
as.numeric(as.character(Stocks$RAMZN[1])) - as.numeric(as.character(Stocks$RAMZN[237]))
```

```
## [1] 0.01683164
```

```
as.numeric(as.character(Stocks$RKMX[1])) - as.numeric(as.character(Stocks$RKMX[237]))
```

```
## [1] -0.0001466276
```

```
as.numeric(as.character(Stocks$RGOOG[1])) - as.numeric(as.character(Stocks$RGOOG[237]))
```

```
## [1] 0.01412041
```

```
as.numeric(as.character(Stocks$RGE[1])) - as.numeric(as.character(Stocks$RGE[237]))
```

```
## [1] 0.01218308
```

*Q5. Find out the top 5 top returns for the given year.*
R CODE:

```
Stocks$AMZN[order(Stocks$AMZN, decreasing = TRUE)[1:5]]
```

```
## [1] 1195.83 1193.60 1190.58 1187.38 1186.10
```

```
Stocks$KMX[order(Stocks$KMX, decreasing = TRUE)[1:5]]
```

```
## [1] 76.81 76.59 76.45 76.37 76.33
```

```
Stocks$GOOG[order(Stocks$GOOG, decreasing = TRUE)[1:5]]
```

```
## [1] 1077.14 1070.68 1064.95 1064.19 1063.63
```

```
Stocks$GE[order(Stocks$GE, decreasing = TRUE)[1:5]]
```

```
## [1] 30.52 30.45 30.37 30.37 30.35
```

*Q6. Find out the top 5 worst returns for the given year.*
R CODE:

```
Stocks$AMZN[order(Stocks$AMZN, decreasing = FALSE)[1:5]]
```

```
## [1] 807.64 810.20 812.50 817.88 819.71
```

```
Stocks$KMX[order(Stocks$KMX, decreasing = FALSE)[1:5]]
```

```
## [1] 55.37 55.61 55.82 56.13 56.24
```

```r
Stocks$GOOG[order(Stocks$GOOG, decreasing = FALSE)[1:5]]
```

```
## [1] 795.695 796.790 798.530 801.340 801.490
```

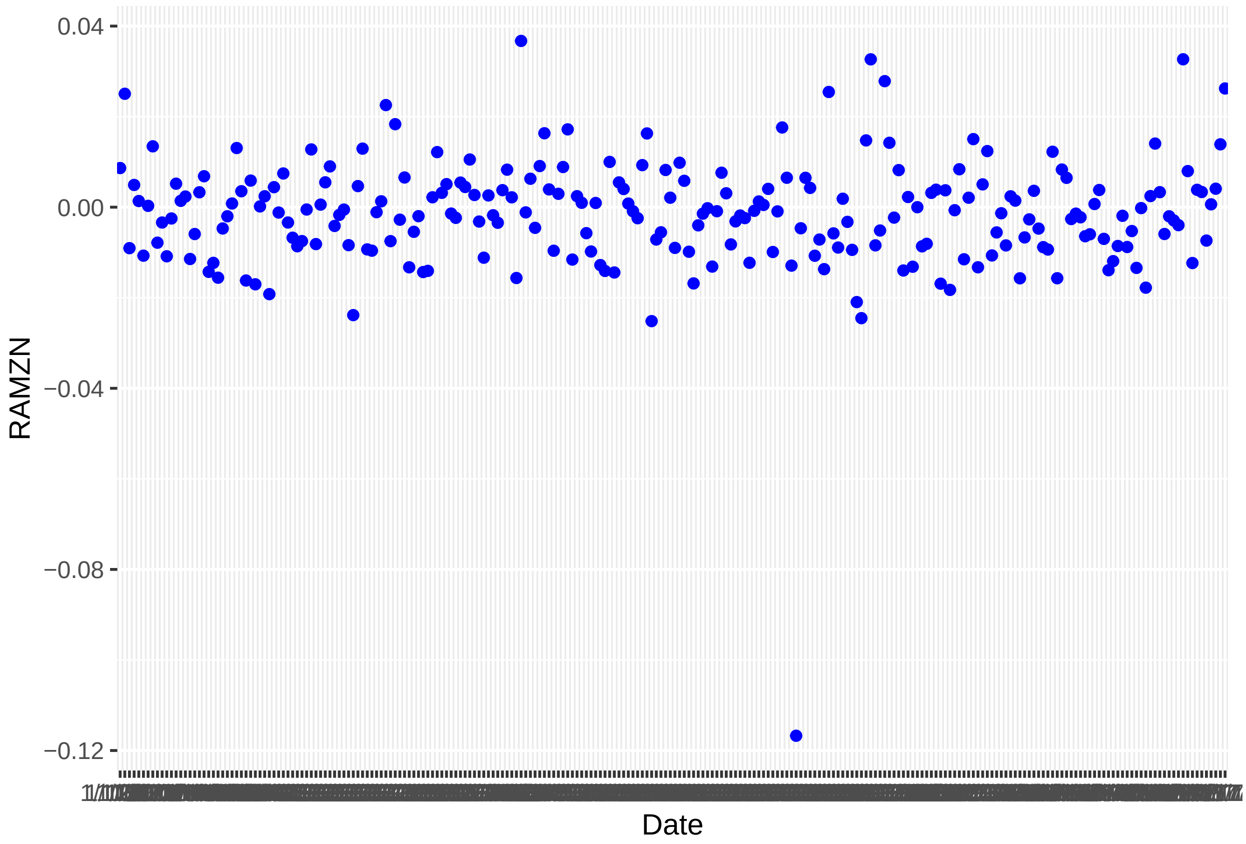
```r
Stocks$GE[order(Stocks$GE, decreasing = FALSE)[1:5]]
```

```
## [1] 17.36 17.38 17.43 17.45 17.45
```

*Q7. Using the function plot(), try to visualize the returns of the stock over one year*
R CODE:

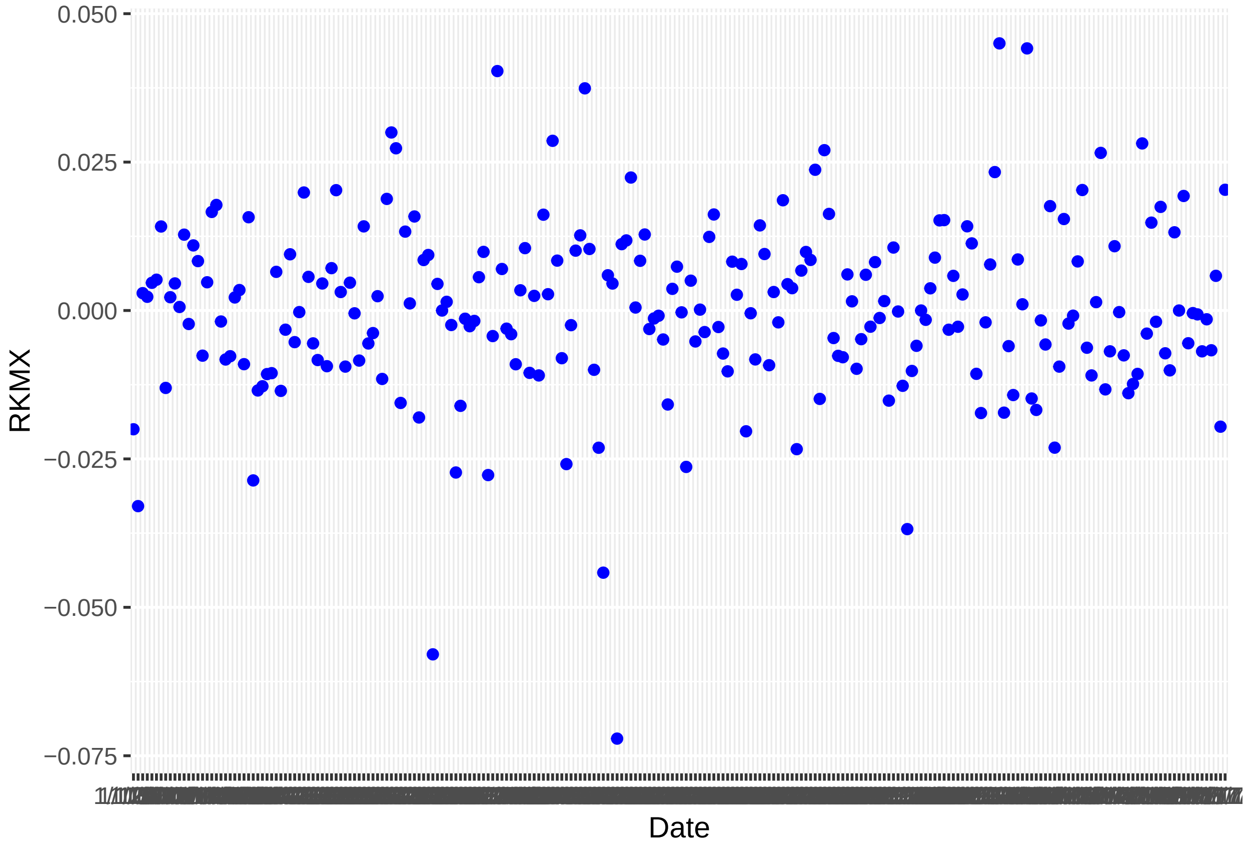
```r
library(ggplot2)
ggplot(data= Stocks, aes( x= Date, y= RAMZN)) + geom_point( color= 'blue') + geom_smooth( method= 'auto
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



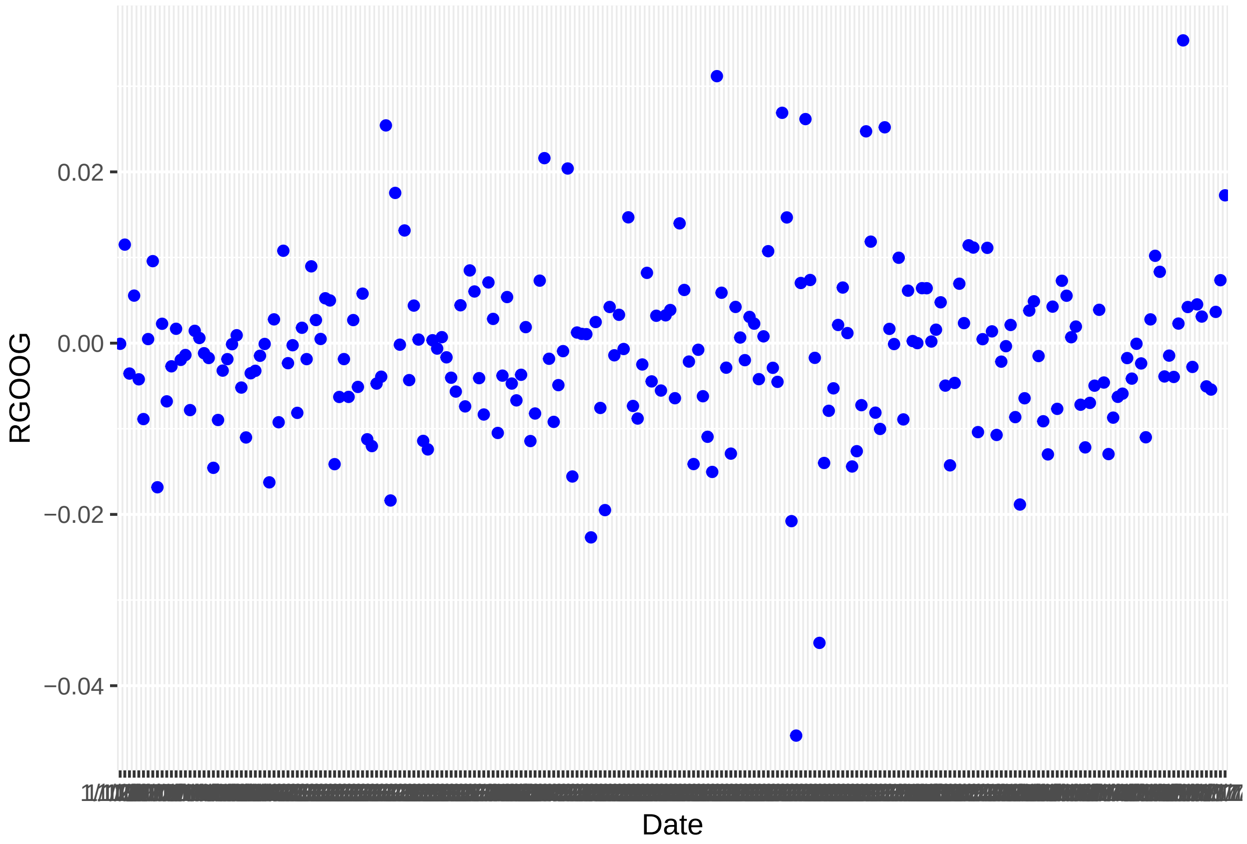
```r
ggplot(data= Stocks, aes( x= Date, y= RKMX)) + geom_point( color= 'blue') + geom_smooth( method= 'auto'
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



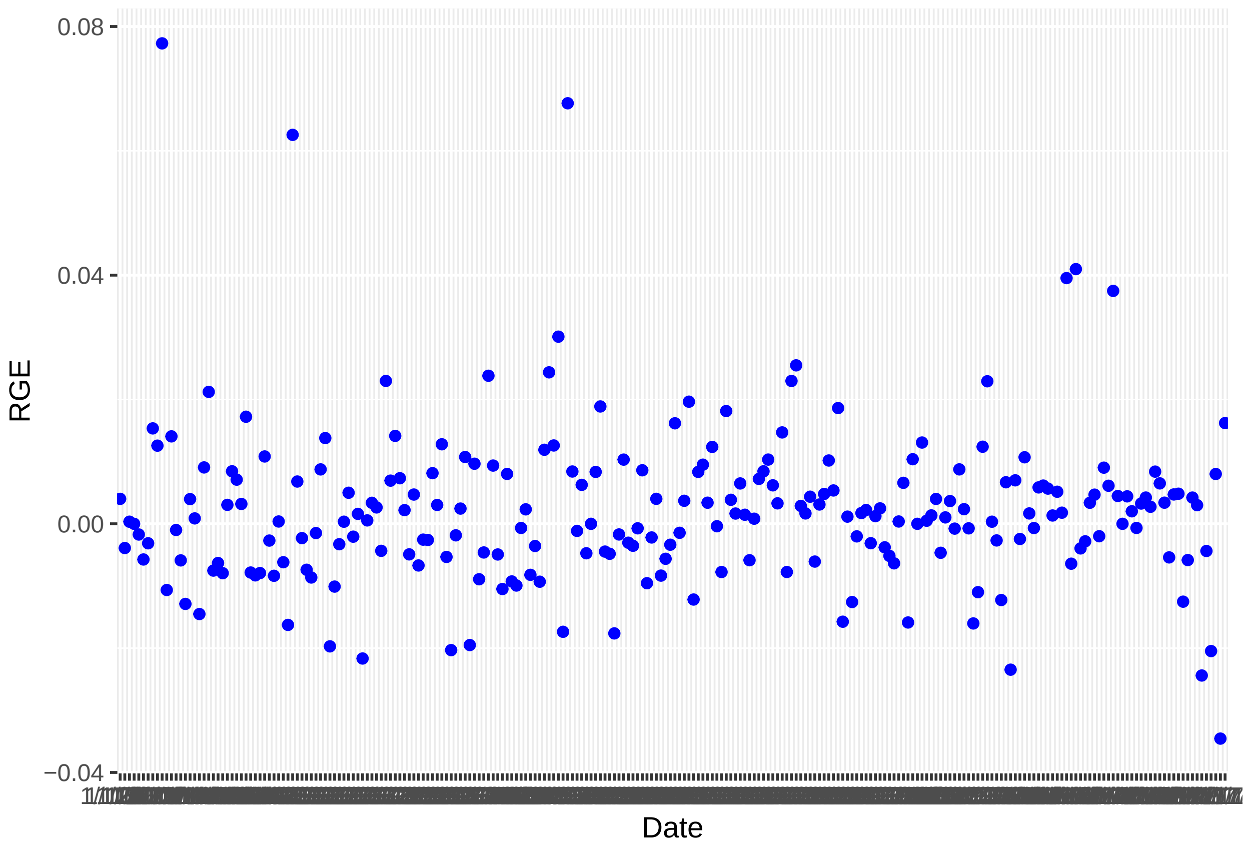
```r
ggplot(data= Stocks, aes( x= Date, y= RGOOG)) + geom_point( color= 'blue') + geom_smooth( method= 'auto
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



```
ggplot(data= Stocks, aes( x= Date, y= RGE)) + geom_point( color= 'blue') + geom_smooth( method= 'auto',
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



***PROBLEM 2*** **Using the Cheesemakers dataset, answer the following questions:**

*Q1. Compute the summary statistics for gross profit in cheese? What does this mean to you?*
R CODE:

```
library(readr)
Cheesemaker<-read_csv(file= "/Users/devarshipancholi/Downloads/Cheesemakers_v2.csv")
```

```
## Parsed with column specification:
## cols(
##   `Contact method` = col_character(),
##   `Customer ID` = col_double(),
##   Date = col_character(),
##   `Item ID` = col_double(),
##   `Item name` = col_character(),
##   `Order ID` = col_double(),
##   `Row ID` = col_double(),
##   State = col_character(),
##   `Gross profit` = col_double(),
##   `Number of Records` = col_double(),
##   `Sale amount` = col_double(),
##   `Sales target` = col_double()
## )
```

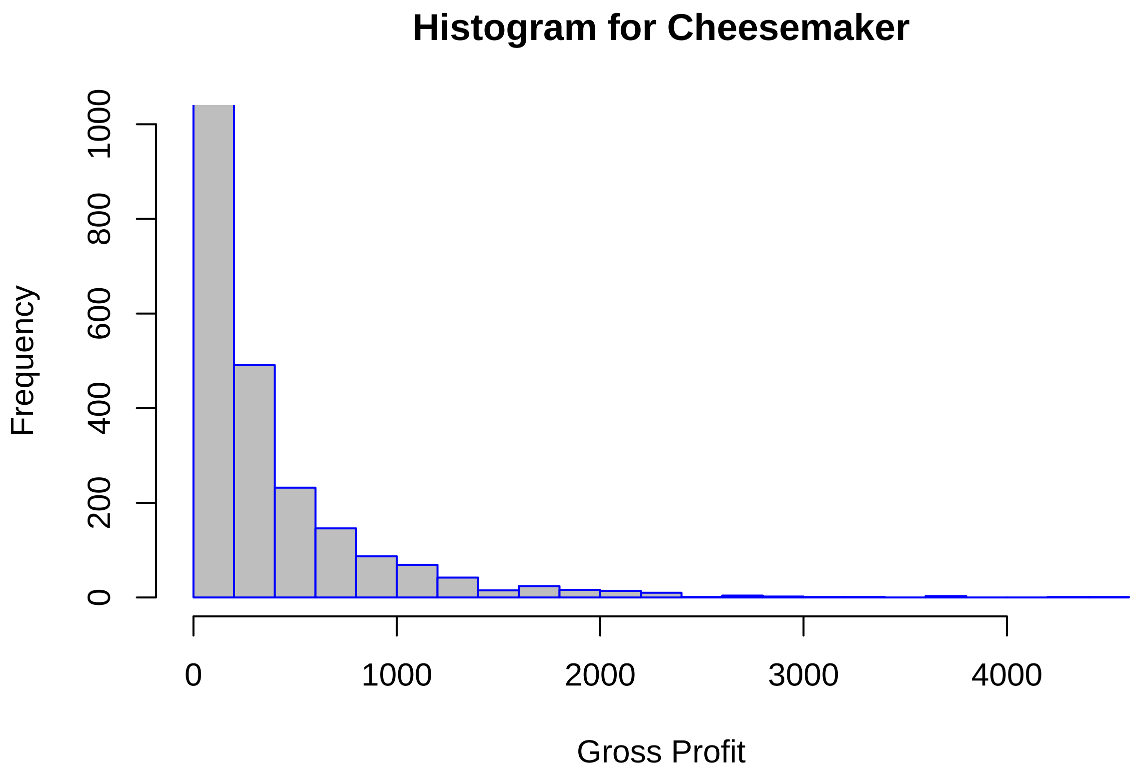
```
summary(Cheesemaker$`Gross profit`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    2.00    7.00   10.00   22.49   18.00 4470.00 
```

DESCRIPTION: As we can see the average profit from all customers is 22.49 with maximum individual of 4470 and minimum of 2.

*Q2. Plot a histogram and a box plot of gross profits. Explain them in English? What do you see? What is normal/abnormal?*
R CODE for histogram:

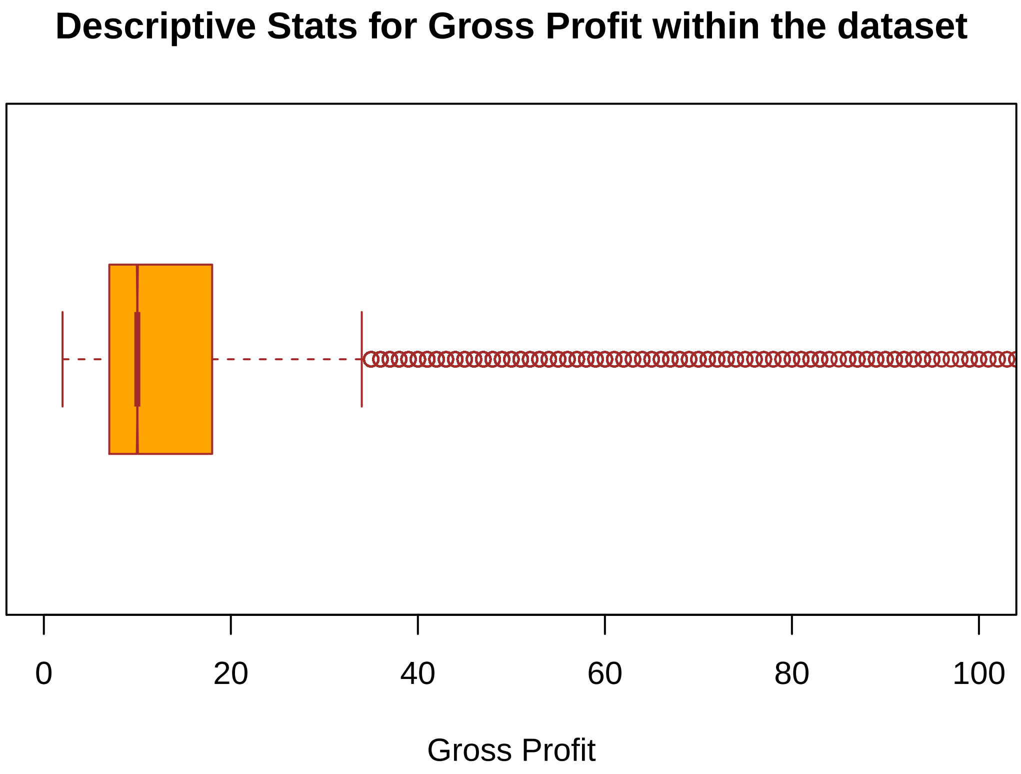
```
hist(Cheesemaker$'Gross profit',
     main="Histogram for Cheesemaker",
     xlab="Gross Profit",
     border="blue",
     col="grey",
     ylim = c(0,1000))
```



DESCRIPTION: In the histogram, I have limited the y-value to 1000 so as to get more clear idea about the frequencies of different groups. it can be clearly seen that the gross profits of 0-100 is the most common/repeated in the dataset. Then comes the gross profit group of 100-200 which is drastically low occuring around just 500 times as compared to approximately 100,000 times for 0-100 group in gross profit.

R CODE for boxplot:

```
boxplot(Cheesemaker$'Gross profit',
        main = "Descriptive Stats for Gross Profit within the dataset",
        xlab= "Gross Profit",
        col = "orange",
        border = "brown",
        ylim= c(0,100),
        horizontal = TRUE,
        notch = TRUE)
```



DESCRIPTION: As there are a lot of outliers, I have limited the Y-axis to 100 as most observation falls in that range which is evident from the histogram we plotted above. Here we can see the minimum is at 2, 1st quartile falls at 7, median falls at 10 and 3rd quartile is at 18. This can be verified from the summary statistics in the above question. However the maximum which is at 4470 is intentionally removed from the plot so as to make the other things clear.
*Q3. Using the CustomerID column, identify the number of customer who have done recurring purchases.*
R CODE:

DESCRIPTION: There are 47,363 repeat entries in Customer ID which means 47,363 customers are repeat customers. I have not included the output here as it prints 47,636 rows and i can't wrap count function inside the head function.
*a. What is the average number of purchases of the recurring clients?*
R CODE:

```
mean(Reccuring$Freq)
```

```
## [1] 1.996221
```

*b. What is the average spent by recurring clients?*
R CODE:

```
mean(subset(Cheesemaker$'Sale amount', Reccuring$Freq > 1 ))
```

```
## [1] 58.35204
```

*c. What is the variance in gross profits between recurring clients vs clients who buy 1 cheese?*
R CODE:

```
var(subset(Cheesemaker$'Gross profit', Reccuring$Freq > 1 ))
```

```
## [1] 6801.187
```

```
var(subset(Cheesemaker$'Gross profit',!Reccuring$Freq > 1 ))
```

```
## [1] 10864.53
```

*Q4. Which are the most profitable clients?*
R CODE:

```
head(sort(Cheesemaker$'Gross profit', decreasing=TRUE), 5)
```

```
## [1] 4470 4206 3725 3704 3652
```

*5. How many clients are paying more than 2 standard deviations of the mean price? What does that mean in english?*
R CODE:

```
sd(Cheesemaker$'Sale amount')*2
```

```
## [1] 500.1844
```

```
length(subset(Cheesemaker$'Customer ID', Cheesemaker$'Sale amount' > 500.1844))
```

```
## [1] 1221
```

DESCRIPTION: As we can see here, 1221 cuctomers are paying above 2 standard deviation of the mean price. This mneans these customers are highly profitable customers.

*6. Compute number of unique clients per state* R Code:

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="AL")))
```

```
## [1] 814
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="AK")))
```

```
## [1] 0
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="AZ")))
```

```
## [1] 1008
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="AR")))
```

```
## [1] 309
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="CA")))
```

```
## [1] 5267
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="CO")))
```

```
## [1] 1110
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="CT")))
```

```
## [1] 714
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="DE")))
```

```
## [1] 160
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="FL")))
```

```
## [1] 3463
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="GA")))
```

```
## [1] 1294
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="HI")))
```

```
## [1] 0
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="ID")))
```

```
## [1] 61
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="IL")))
```

```
## [1] 1879
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="IN")))
```

```
## [1] 798
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="IA")))
```

```
## [1] 545
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="KS")))
```

```
## [1] 617
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="KY")))
```

```
## [1] 432
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="LA")))
```

```
## [1] 666
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="ME")))
```

```
## [1] 218
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="MD")))
```

```
## [1] 1063
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="MA")))
```

```
## [1] 851
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="MI")))
```

```
## [1] 1352
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="MN")))
```

```
## [1] 657
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="MS")))
```

```
## [1] 426
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="MO")))
```

```
## [1] 869
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="MT")))
```

```
## [1] 192
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="NE")))
```

```
## [1] 385
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="NV")))
```

```
## [1] 437
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="NH")))
```

```
## [1] 211
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="NJ")))
```

```
## [1] 1236
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="NM")))
```

```
## [1] 336
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="NY")))
```

```
## [1] 2126
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="NC")))
```

```
## [1] 1246
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="ND")))
```

```
## [1] 201
```

```r
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="OH")))
```

```
## [1] 2013
```

```r
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="OK")))
```

```
## [1] 648
```

```r
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="OR")))
```

```
## [1] 162
```

```r
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="PA")))
```

```
## [1] 1645
```

```r
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="RI")))
```

```
## [1] 110
```

```r
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="SC")))
```

```
## [1] 535
```

```r
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="SD")))
```

```
## [1] 211
```

```r
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="TN")))
```

```
## [1] 976
```

```r
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="TX")))
```

```
## [1] 5545
```

```r
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="UT")))
```

```
## [1] 278
```

```r
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="VT")))
```

```
## [1] 81
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="VA")))
```

```
## [1] 2240
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="WA")))
```

```
## [1] 385
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="WV")))
```

```
## [1] 179
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="WI")))
```

```
## [1] 749
```

```
length(unique(subset(Cheesemaker$'Customer ID',Cheesemaker$'State'=="WY")))
```

```
## [1] 239
```

*a. Normalize the data using min-max scaling*
R CODE:

```
library(normalr)
head(normalize(Cheesemaker$'Gross profit'),10)
head(normalize(Cheesemaker$'Sale amount'),10)
head(normalize(Cheesemaker$'Sale amount'),10)
head(normalize(Cheesemaker$'Sales target'),10)
```

*b. Is there an association (correlation) between client volume and sales?*
R CODE:

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
all<- Cheesemaker %>% select_if(is.numeric)
cor(all, use="all.obs", method="pearson")
```

```
## Warning in cor(all, use = "all.obs", method = "pearson"): the standard
## deviation is zero
```

```
##                   Customer ID      Item ID     Order ID     Row ID
## Customer ID       1.000000000 0.0034851056  0.351940463 0.20229213
## Item ID           0.003485106 1.0000000000  0.038387945 0.10352925
## Order ID          0.351940463 0.0383879451  1.000000000 0.25780168
## Row ID            0.202292131 0.1035292455  0.257801681 1.00000000
## Gross profit      0.163697643 0.0002959835 -0.006255983 0.01997163
## Number of Records          NA           NA           NA         NA
## Sale amount       0.169449487 0.0025858346 -0.002788553 0.02594528
## Sales target      0.172576725 0.0123084008 -0.041873478 0.01013048
##                    Gross profit Number of Records  Sale amount
## Customer ID        0.1636976430                NA  0.169449487
## Item ID            0.0002959835                NA  0.002585835
## Order ID          -0.0062559829                NA -0.002788553
## Row ID             0.0199716300                NA  0.025945280
## Gross profit       1.0000000000                NA  0.981689257
## Number of Records            NA                 1           NA
## Sale amount        0.9816892567                NA  1.000000000
## Sales target       0.9033325040                NA  0.904902929
##                   Sales target
## Customer ID         0.17257673
## Item ID             0.01230840
## Order ID           -0.04187348
## Row ID              0.01013048
## Gross profit        0.90333250
## Number of Records           NA
## Sale amount         0.90490293
## Sales target        1.00000000
```

```r
cor(Cheesemaker$'Sales target',Cheesemaker$'Sale amount')
```

```
## [1] 0.9049029
```

DESCRIPTION: Looking at the output there seems to be minimal co-relation, but the states which has high customers has high sales indicating association between the two variables.