Problem 1

#import data  
url<-"https://raw.githubusercontent.com/jcbonilla/BusinessAnalytics/master/BAData/Cheesemakers\_v2.csv"  
  
cheeseMakers<-read.csv(url, header=TRUE,stringsAsFactors=FALSE)  
str(cheeseMakers)

## 'data.frame': 94547 obs. of 12 variables:  
## $ Contact.method : chr "Both" "Both" "Both" "Both" ...  
## $ Customer.ID : int 18565 22873 38254 46726 41695 18348 4985 16630 45494 46847 ...  
## $ Date : chr "1/1/12" "1/1/12" "1/1/12" "1/1/12" ...  
## $ Item.ID : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Item.name : chr "Calistoga Cheddar" "Calistoga Cheddar" "Calistoga Cheddar" "Calistoga Cheddar" ...  
## $ Order.ID : int 106 145 179 207 188 102 29 92 202 208 ...  
## $ Row.ID : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ State : chr "CA" "CA" "CA" "CT" ...  
## $ Gross.profit : int 8 10 10 8 21 10 10 10 18 50 ...  
## $ Number.of.Records: int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Sale.amount : int 18 24 24 18 48 24 24 24 43 117 ...  
## $ Sales.target : int 22 22 22 22 67 22 22 22 44 156 ...

head(cheeseMakers)

## Contact.method Customer.ID Date Item.ID Item.name Order.ID  
## 1 Both 18565 1/1/12 1 Calistoga Cheddar 106  
## 2 Both 22873 1/1/12 1 Calistoga Cheddar 145  
## 3 Both 38254 1/1/12 1 Calistoga Cheddar 179  
## 4 Both 46726 1/1/12 1 Calistoga Cheddar 207  
## 5 Both 41695 1/1/12 1 Calistoga Cheddar 188  
## 6 Both 18348 1/1/12 1 Calistoga Cheddar 102  
## Row.ID State Gross.profit Number.of.Records Sale.amount Sales.target  
## 1 1 CA 8 1 18 22  
## 2 1 CA 10 1 24 22  
## 3 1 CA 10 1 24 22  
## 4 1 CT 8 1 18 22  
## 5 1 DC 21 1 48 67  
## 6 1 FL 10 1 24 22

# Compute the summary statistics for gross profit in cheese? What does this mean to you?

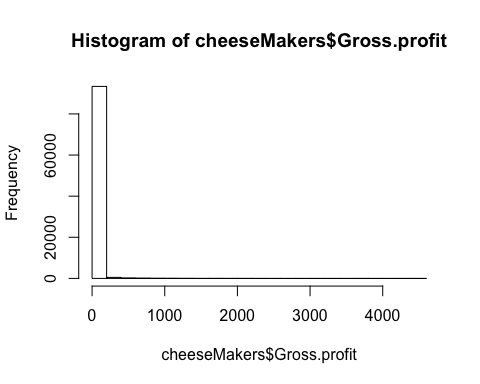
summary(cheeseMakers$Gross.profit)

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

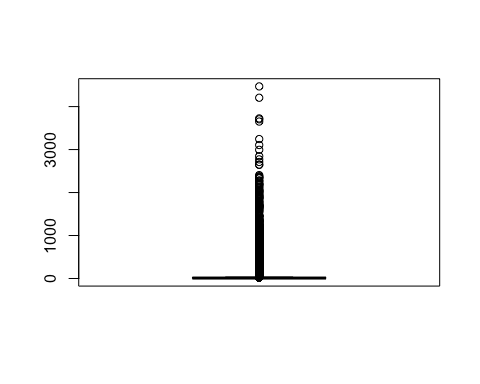
The maximum is pretty large comparing to the mean, meaning that there could be one or more outliers contain(s) in this dataset.

# Plot a histogram and a box plot of gross profits. What do you see? What is normal/abnormal?

hist(cheeseMakers$Gross.profit)

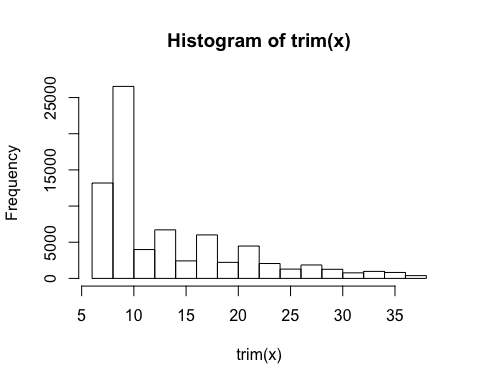


boxplot(cheeseMakers$Gross.profit)

 There are many extreme outliers in the dataset that are abnormal, making noise and affecting the graphs.

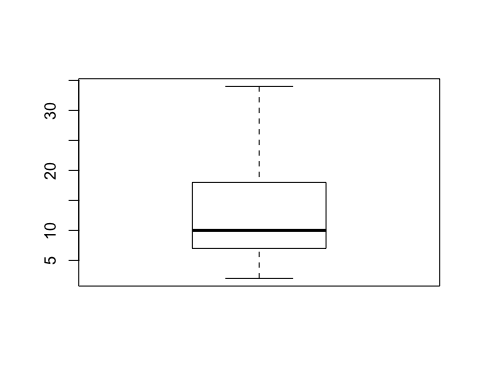
# histogram after dropping the outliers by using trim function

x <- cheeseMakers$Gross.profit  
trim <- function(x){  
 x[(x > mean(x)-1.5\*IQR(x)) & (x < mean(x)+1.5\*IQR(x))]  
}  
hist(trim(x))



# boxplot after dropping the outliers

boxplot(cheeseMakers$Gross.profit, outline = FALSE)



The histogram without outliers is right-skewed, with the majority ranging from 5 to 10. The boxplot without outliers shows that the dataset is ranging from approximately 1 to 34, the majority is ranging from approximately 7 to 17, with the median of 10.

# Using the CustomerID column, identify the number of customers who have done recurring purchases.

n\_occur <- data.frame(table(cheeseMakers$Customer.ID))  
nrow(n\_occur[n\_occur$Freq > 1,])

## [1] 26169

The number of customers who have done recurring purchases is 26169.

# What is the average number of purchases of the recurring clients?

avg\_nop <- mean(n\_occur[n\_occur$Freq > 1,]$Freq)  
avg\_nop

## [1] 2.803049

The average number of purchases of the recurring clients is 2.803049.

# What is the average spent by recurring clients?

recurring <- cheeseMakers[cheeseMakers$Customer.ID %in% cheeseMakers$Customer.ID[duplicated(cheeseMakers$Customer.ID)],]  
recurring\_sale <- aggregate(Sale.amount ~ Customer.ID ,data = recurring, sum)  
mean(recurring\_sale$Sale.amount)

## [1] 191.0412

The average spent by recurring clients is 191.0412.

# What is the variance in gross profits between recurring clients vs clients who buy 1 cheese?

onetime <- cheeseMakers[!cheeseMakers$Customer.ID %in% cheeseMakers$Customer.ID[duplicated(cheeseMakers$Customer.ID)],]  
var(onetime$Gross.profit)

## [1] 1591.282

recurring\_profit <- aggregate(Gross.profit ~ Customer.ID ,data = recurring, sum)  
var(recurring\_profit$Gross.profit)

## [1] 101761.3

The variance of clients who buy 1 cheese is 1591.282, and variance of recurring clients is 101761.3.

# Which are the most profitable clients?

subset(onetime,onetime$Gross.profit == max(onetime$Gross.profit))

## Contact.method Customer.ID Date Item.ID Item.name Order.ID  
## 51849 Neither 47708 4/5/13 1 Calistoga Cheddar 47188  
## Row.ID State Gross.profit Number.of.Records Sale.amount Sales.target  
## 51849 1 CA 1805 1 4223 4053

subset(recurring\_profit,recurring\_profit$Gross.profit == max(recurring\_profit$Gross.profit))

## Customer.ID Gross.profit  
## 26169 47911 18968

Customer who buy 1 cheese with ID 47708 has highest profit 1805; Recurring client with ID 47911 has highest profit 18968 Overall the customer with ID 47911 is the most profitable client.

# How many clients are paying more than 2 standard deviations of the mean price? What does that mean in english?

onetime\_sale<-data.frame('Customer.ID' = onetime$Customer.ID, 'Sale.amount' = onetime$Sale.amount)  
total\_sale<-c(onetime\_sale,recurring\_sale)  
length(which(total\_sale$Sale.amount>(mean(total\_sale$Sale.amount)+2\*sd(total\_sale$Sale.amount))))

## [1] 461

There are 461 clients pay more than 2 standard deviations of the mean price, meaning that very few customers like to spend much higher price buying cheese.

# Compute number of unique clients per state

NumberPerState<-aggregate(Number.of.Records ~ State, data = onetime, sum)  
NumberPerState$min\_max<-(NumberPerState$Number.of.Records-min(NumberPerState$Number.of.Records))/(max(NumberPerState$Number.of.Records)-min(NumberPerState$Number.of.Records))  
NumberPerState

## State Number.of.Records min\_max  
## 1 AL 377 0.142915476  
## 2 AR 150 0.050224581  
## 3 AZ 442 0.169456921  
## 4 CA 2361 0.953042058  
## 5 CO 507 0.195998367  
## 6 CT 336 0.126173949  
## 7 DC 222 0.079624336  
## 8 DE 70 0.017558187  
## 9 FL 1482 0.594120049  
## 10 GA 621 0.242547979  
## 11 IA 238 0.086157615  
## 12 ID 27 0.000000000  
## 13 IL 809 0.319314006  
## 14 IN 335 0.125765619  
## 15 KS 246 0.089424255  
## 16 KY 181 0.062882809  
## 17 LA 333 0.124948959  
## 18 MA 399 0.151898734  
## 19 MD 518 0.200489996  
## 20 ME 111 0.034299714  
## 21 MI 622 0.242956309  
## 22 MN 282 0.104124132  
## 23 MO 373 0.141282156  
## 24 MS 192 0.067374439  
## 25 MT 75 0.019599837  
## 26 NC 558 0.216823193  
## 27 ND 75 0.019599837  
## 28 NE 175 0.060432830  
## 29 NH 98 0.028991425  
## 30 NJ 556 0.216006533  
## 31 NM 162 0.055124541  
## 32 NV 190 0.066557779  
## 33 NY 1008 0.400571662  
## 34 OH 889 0.351980400  
## 35 OK 289 0.106982442  
## 36 OR 77 0.020416497  
## 37 PA 726 0.285422621  
## 38 RI 51 0.009799918  
## 39 SC 214 0.076357697  
## 40 SD 101 0.030216415  
## 41 TN 441 0.169048591  
## 42 TX 2476 1.000000000  
## 43 UT 124 0.039608003  
## 44 VA 961 0.381380155  
## 45 VT 47 0.008166599  
## 46 WA 177 0.061249490  
## 47 WI 311 0.115965700  
## 48 WV 73 0.018783177  
## 49 WY 106 0.032258065

# Is there an association (correlation) between client volume and sales?

SalesPerState<-aggregate(Sale.amount ~ State, data = onetime, sum)  
cor(NumberPerState$Number.of.Records,SalesPerState$Sale.amount)

## [1] 0.8172261

Yes, the correlation is 0.8172261, meaning that there is a high correlation between volume and sales in unique clients per state.

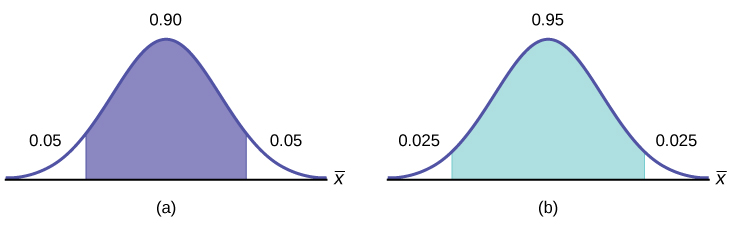
PROBLEM 2

# Compared to the 95% confidence interval, is the 90% confidence interval for toy sales wider or narrower? Please explain your answer very briefly.

The 90% confidence interval for toy sales will be narrower than the 95% confidence interval, because we are increasing the alpha value, which is the area at the tail of the curve, from 5% to 10%.

A screenshot of a cell phone

Description automatically generated



# How many more store toy sales fall within the 95% confidence interval for the population mean toy sales than within the 90% confidence interval?

There are 3 more stores fall within the 95% confidence interval than within the 90% confidence interval.

A screenshot of a cell phone

Description automatically generated

Problem 3

# Auditors of Independent Bank are interested in comparing the reported value of all the 1775 customer saving account balances (the bank has 1775 savings accounts) with their own findings regarding the actual value of such assets.  Rather than reviewing the records of each savings account at the bank, the auditors randomly selected a sample of 100 savings account balances.  The sample mean and sample standard deviations were $505.75 and 360.95, respectively.

* **Construct a 90% confidence interval for the total value of all savings account balances within this bank.**

Sample Mean = *x* = 505.75

Standard Deviation of Population = σ= 360.95

Sample Size = *n* = 100

Significance Level = α = 0.1

Critical Value = zα/2 = z0.05 =1.6604

Critical Value , zα/2 = 1.6604

Margin of Error = zα/2 x =1.6604 x =59.9321

Confidence Interval Limits of 90%= X - E = 505.75 – 59.9321 =445.82

X + E = 505.75 + 59.9321 =565.68

90% Confidence Interval = 505.75 59.9321 = (445.82, 565.68)

Table

Description automatically generated

As there are 1775 savings account a confidence interval of 90% for total value of all savings account balance within this bank is: (1775\*445.82, 1775\*565.68)

The confidence interval of total value is ($792,313.5$1,004,082).

* **Interpret the 90% confidence interval constructed above**

The bank is 90% confident that the total balance of 1775 customers are between $792,313.5 and $1,004,082