Question 2. Download this data set and then upload the data into RStudio. Each row represents a customer's interactions with the organization's web store.

```
cust_data<- read.csv("customertxndata.csv")
# Reading Dataset
```

Question 3. Calculate the following summative statistics: total transaction amount (revenue), mean number of visits, median revenue, standard deviation of revenue, most common gender.

```
colnames(cust_data) <- c("number of visits", "Transactions", "O.S", "Gender", "Revenue")
attach(cust_data)

cust_data %>%
summarise(total_revenue=sum(Revenue, na.rm = T),mean_visits=mean(`number of visits`,
na.rm = T), med_revenue=median(Revenue, na.rm = T), stdev_revenue=sd(Revenue, na.rm = T))
```

```
##   total_revenue mean_visits med_revenue stdev_revenue
## 1      10372524    12.48673    344.6516      425.9871
```

```
# Finding most common gender
table(Gender)
```

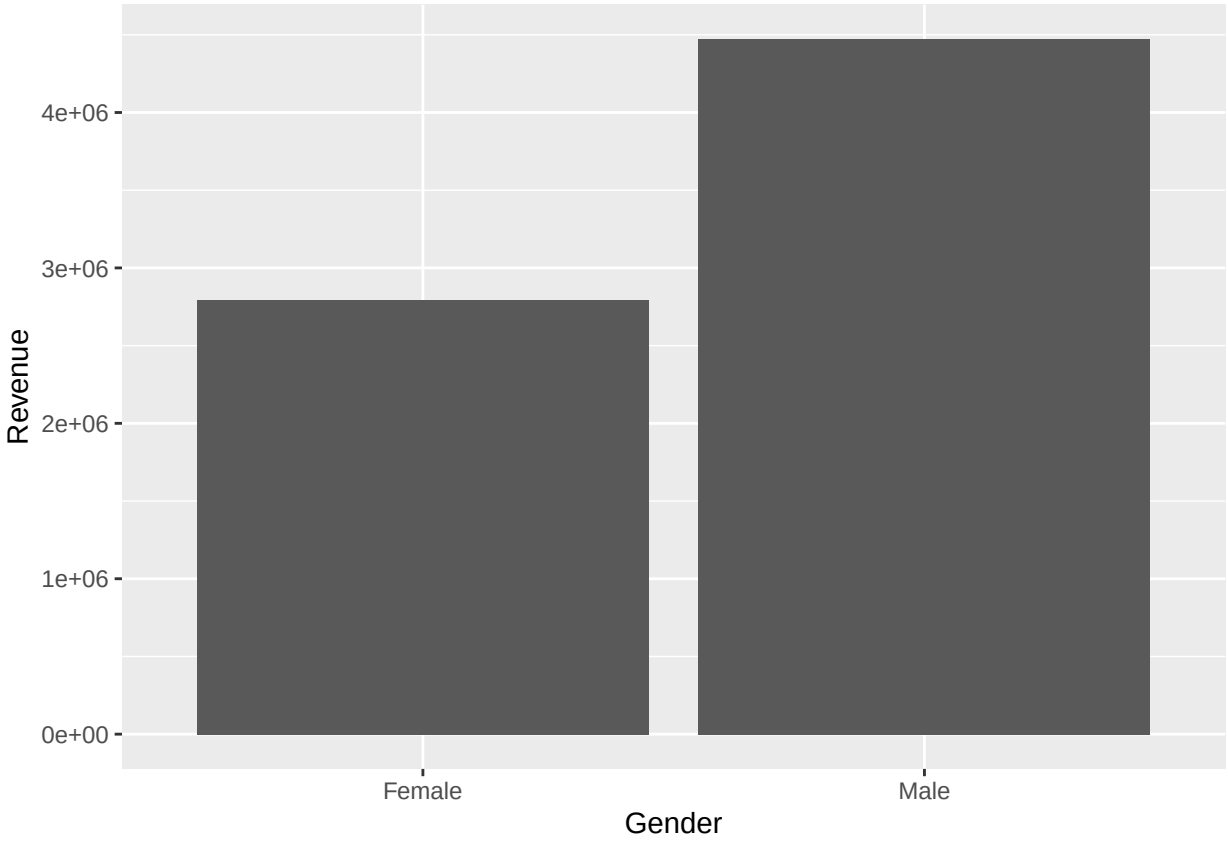
```
## Gender
## Female   Male
##   2670  14729
```

```
print("Most common gender is male")
```

```
## [1] "Most common gender is male"
```

Question 4. Create a bar/column chart of gender (x-axis) versus revenue (y-axis).

```
ggplot(data= na.omit(cust_data), aes(x=Gender, y=Revenue)) +
  geom_bar(stat="identity")
```



Question 5. What is the Pearson Moment of Correlation between number of visits and revenue? Comment on the correlation.

```
cor(`number of visits`, Revenue, method = "pearson")
```

```
## [1] 0.7388375
```

As the pearson correlation is 0.738, this signifies that the variables are correlated and have a positive linear relationship

Question 6. Which columns have missing data? How did you recognize them? How would you impute missing values?

```
table(is.na(Gender))
```

```
##
## FALSE  TRUE
## 17399  5400
```

```
table(is.na(`number of visits`))
```

```
##
## FALSE
## 22799
```

```
table(is.na(Transactions))
```

```
##
## FALSE  TRUE
## 20999  1800
```

```
table(is.na(O.S))
```

```
##
## FALSE
## 22799
```

```
table(is.na(Revenue))
```

```
##
## FALSE
## 22799
```

Computed the frequency of NAs for each columns using table function.Imputation could be done using KNN algo, regression models and statistical measures(mean, median & mode)

Question 7. Impute missing transaction and gender values. Use the mean for transaction (rounded to the nearest whole number) and the mode for gender.

```
impute_trans <- round(impute(Transactions, mean))
impute_gen <- impute(Gender, mode)
head (impute_trans, 20)
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  1  1  2  0  1  2  1  1  2  1  1  0  1  2  2  2 1*  2  1  2
```

```
head (impute_gen, 20)
```

```
##      1      2      3      4      5      6      7      8      9     10     11
##  Male* Female Female   Male   Male   Male   Male   Male   Male   Male  Male*
##     12     13     14     15     16     17     18     19     20
##   Male  Male*   Male Female  Male*   Male   Male   Male Female
```

Question 8. Split the data set into two equally sized data sets where one can be used for training a model and the other for validation. Take every odd numbered case and add them to the training data set and every even numbered case and add them to the validation data set.

```
n <- nrow(cust_data)/2
even <- seq(2, n-1, 2)
odd <- seq(1,n, 2)


valid_data <- cust_data[even,]
train_data <- cust_data[odd,]
```

Question 9. Calculate the mean revenue for the training and the validation data sets and compare them. Comment on the difference.

```r
mean(valid_data$Revenue)
```

```
## [1] 452.5755
```

```r
mean(train_data$Revenue)
```

```
## [1] 458.2594
```

The mean revenue for training dataset is 458.25 and for validation dataset is 452.57, thus they seem slightly significant or close.

Question 10. Use the sample() function to split the data set, so that 60% is used for training and 20% is used for testing, and another 20% is used for validation.

```r
set.seed(77654)

sample <- sample.int(n = nrow(cust_data), size = floor(.60*nrow(cust_data)), replace = F)
train <- cust_data[sample, ]

newdata <- cust_data[-sample,]

sample2 <- sample.int( n= nrow(newdata), size = floor(.50*nrow(newdata)), replace = F)

test  <- newdata[sample2, ]
valid <- newdata[-sample2, ]

mean(train$Revenue)
```

```
## [1] 455.575
```

```r
mean(test$Revenue)
```

```
## [1] 459.7076
```

```r
mean(valid$Revenue)
```

```
## [1] 448.3437
```