Step 3 _analysis

2022-08-11

Importing all the required libraries.

Data read before cleaning the datasets.

First dataset

```
caffeine_df <- read.csv("data/caffeine.csv")
head(caffeine_df)
```

```
##                         drink Volume..ml. Calories Caffeine..mg.   type
## 1                Costa Coffee    256.9937        0           277 Coffee
## 2 Coffee Friend Brewed Coffee    250.1918        0           145 Coffee
## 3          Hell Energy Coffee    250.1918      150           100 Coffee
## 4          Killer Coffee (AU)    250.1918        0           430 Coffee
## 5                Nescafe Gold    250.1918        0            66 Coffee
## 6            Espresso Monster    248.4174      170           160 Coffee
```

```
#summary(caffeine_df)
```

```
names(caffeine_df) <- c('DRINK', 'VOLUME', 'CALORIES', 'CAFFEINE', 'TYPE')
head(caffeine_df)
```

```
##                         DRINK   VOLUME CALORIES CAFFEINE   TYPE
## 1                Costa Coffee 256.9937        0      277 Coffee
## 2 Coffee Friend Brewed Coffee 250.1918        0      145 Coffee
## 3          Hell Energy Coffee 250.1918      150      100 Coffee
## 4          Killer Coffee (AU) 250.1918        0      430 Coffee
## 5                Nescafe Gold 250.1918        0       66 Coffee
## 6            Espresso Monster 248.4174      170      160 Coffee
```

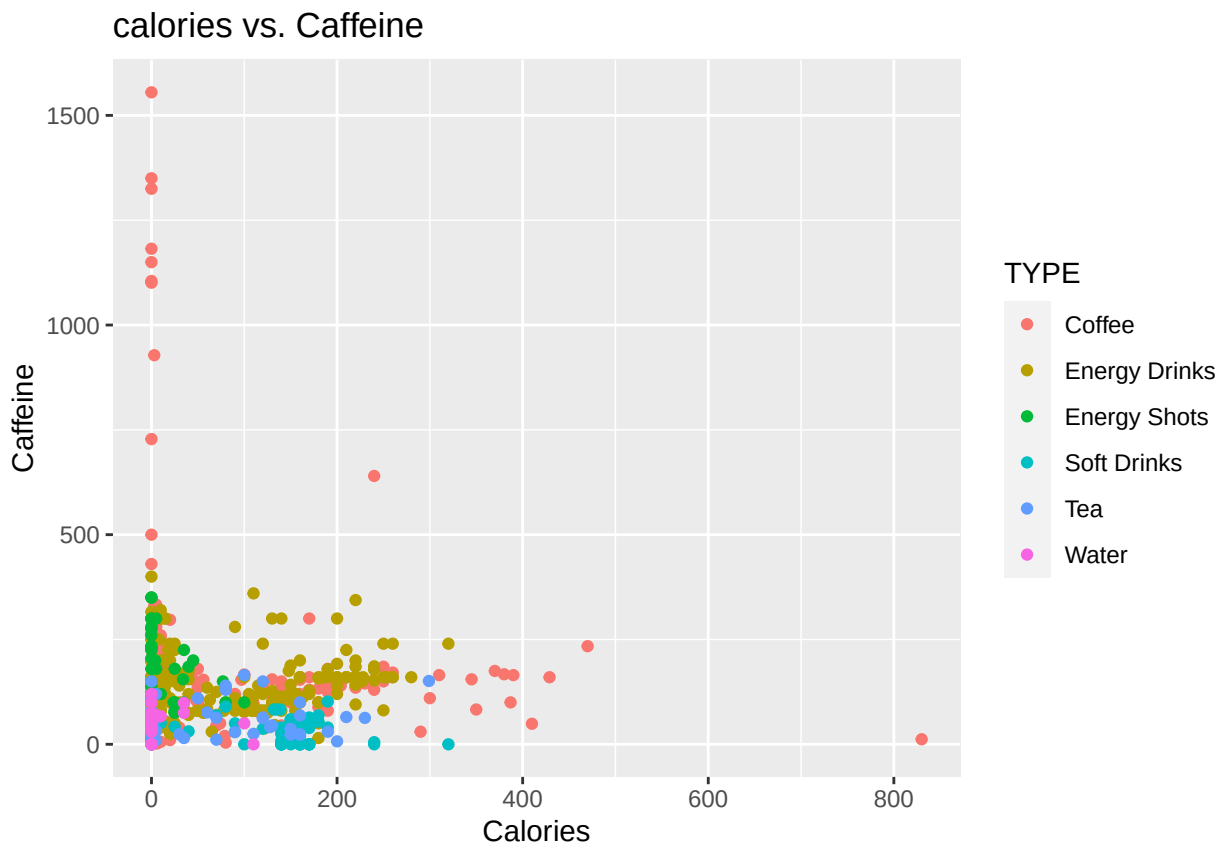
```
caffeine_df <- na.omit(caffeine_df)
head(caffeine_df)
```

```
##                         DRINK   VOLUME CALORIES CAFFEINE   TYPE
## 1                Costa Coffee 256.9937        0      277 Coffee
## 2 Coffee Friend Brewed Coffee 250.1918        0      145 Coffee
## 3          Hell Energy Coffee 250.1918      150      100 Coffee
## 4          Killer Coffee (AU) 250.1918        0      430 Coffee
## 5                Nescafe Gold 250.1918        0       66 Coffee
## 6            Espresso Monster 248.4174      170      160 Coffee
```

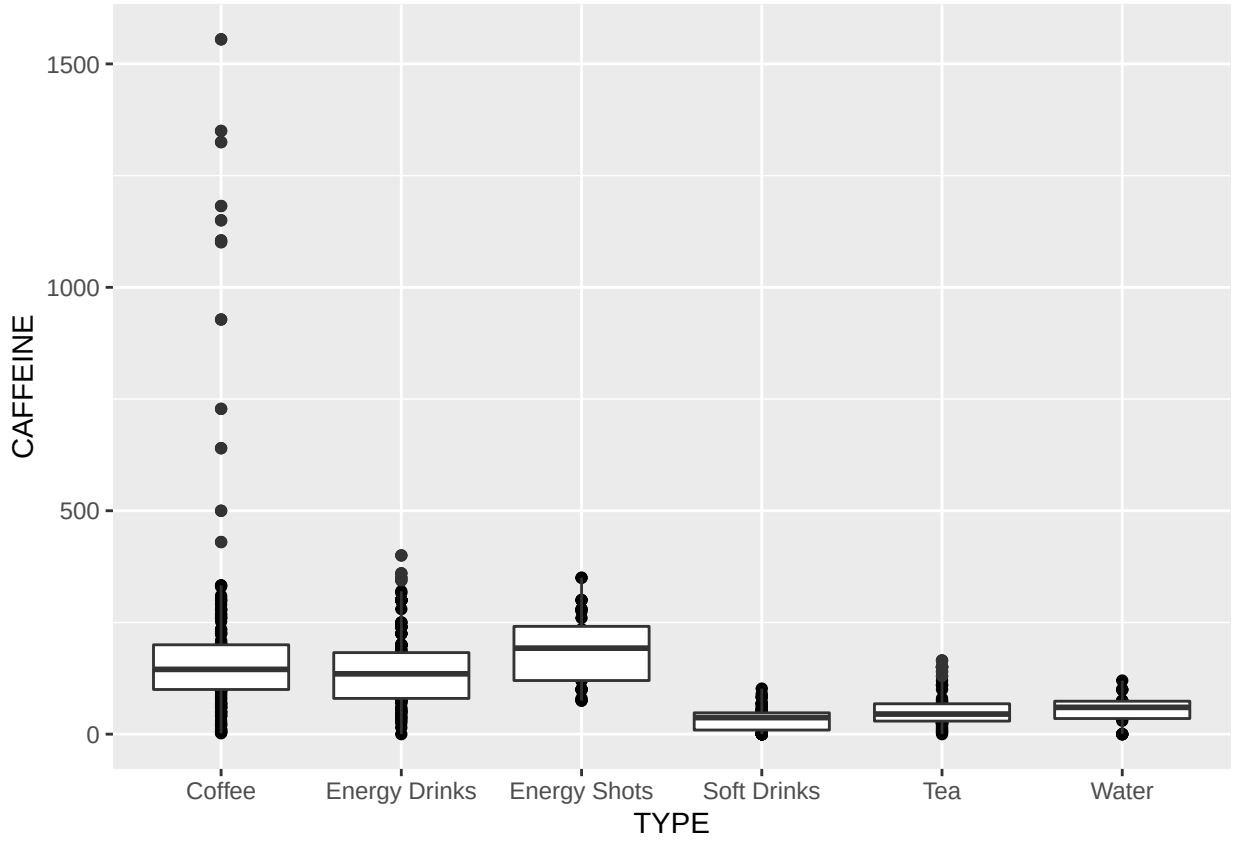
```
summary(caffeine_df$TYPE)
```

```
##    Length     Class      Mode
##       610 character character
```

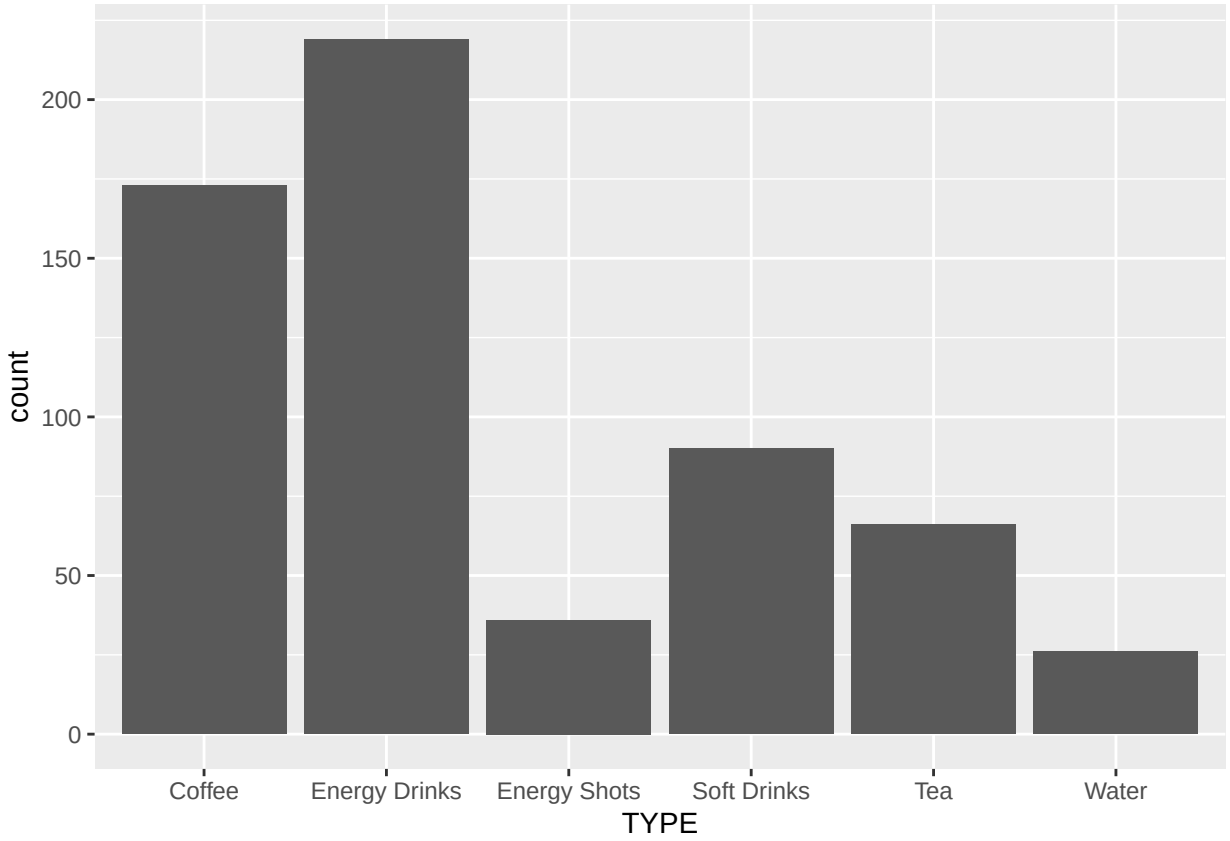
```
ggplot(caffeine_df, aes(x=CALORIES, y=CAFFEINE, col=TYPE)) + geom_point() +
  ggtitle("calories vs. Caffeine") + xlab("Calories") + ylab("Caffeine")
```




```
ggplot(caffeine_df, aes(x=TYPE, y=CAFFEINE)) + geom_point()+ geom_boxplot()
```




```
ggplot(caffeine_df, aes(TYPE)) + geom_bar()
```




```
#caffeine_df$TYPE <- as.factor(caffeine_df$TYPE)

#type_lm <-  lm( TYPE ~ CAFFEINE + CALORIES, data=caffeine_df)
#type_lm
```

```
caffeine_df$TYPE <- as.factor(caffeine_df$TYPE)
caffeine_glm <- glm(TYPE ~ CAFFEINE + CALORIES , data = caffeine_df, family = binomial())
summary(caffeine_glm)
```

```
##
## Call:
## glm(formula = TYPE ~ CAFFEINE + CALORIES, family = binomial(),
##     data = caffeine_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9173  -1.2508   0.6738   0.7896   1.3434
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.6785515  0.1850911   9.069  < 2e-16 ***
## CAFFEINE    -0.0051516  0.0009790  -5.262 1.42e-07 ***
## CALORIES    -0.0005942  0.0009707  -0.612     0.54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 727.52  on 609  degrees of freedom
## Residual deviance: 681.66  on 607  degrees of freedom
## AIC: 687.66
##
## Number of Fisher Scoring iterations: 5
```

Second dataset

```
coffeesurvey2_df <- read.csv("data/Coffee_Survey2.csv")
head(coffeesurvey2_df)
```

```
##   Do.you.drink.coffee.daily.
## 1                          1
## 2                          1
## 3                          1
## 4                          0
## 5                          0
## 6                          1
##   How.many.coffee.you.drink.daily..Starbucks.Grande.cup..
## 1                                                       2
## 2                                                       2
## 3                                                       3
## 4                                                       1
## 5                                                       0
## 6                                                       3
##                              Why.do.you.drink.coffee
## 1                                        Study Stress
## 2                           Refreshing every morning
## 3                                     Living habits
## 4                                     Living habits
## 5                                      Study Stress
## 6 Study Stress;Living habits;Refreshing every morning
##   Do.you.think.coffee.works.for.you.
## 1                                  0
## 2                                  1
## 3                                  2
## 4                                  0
## 5                                  0
## 6                                  1
```

```
names(coffeesurvey2_df) <- c('drinkcoffee', 'totalcups', 'ycoffee', 'coffeeworks')
#head(coffeesurvey2_df)
coffeesurvey2_df$totalcups <- as.integer(coffeesurvey2_df$totalcups)
```

```
## Warning: NAs introduced by coercion
```

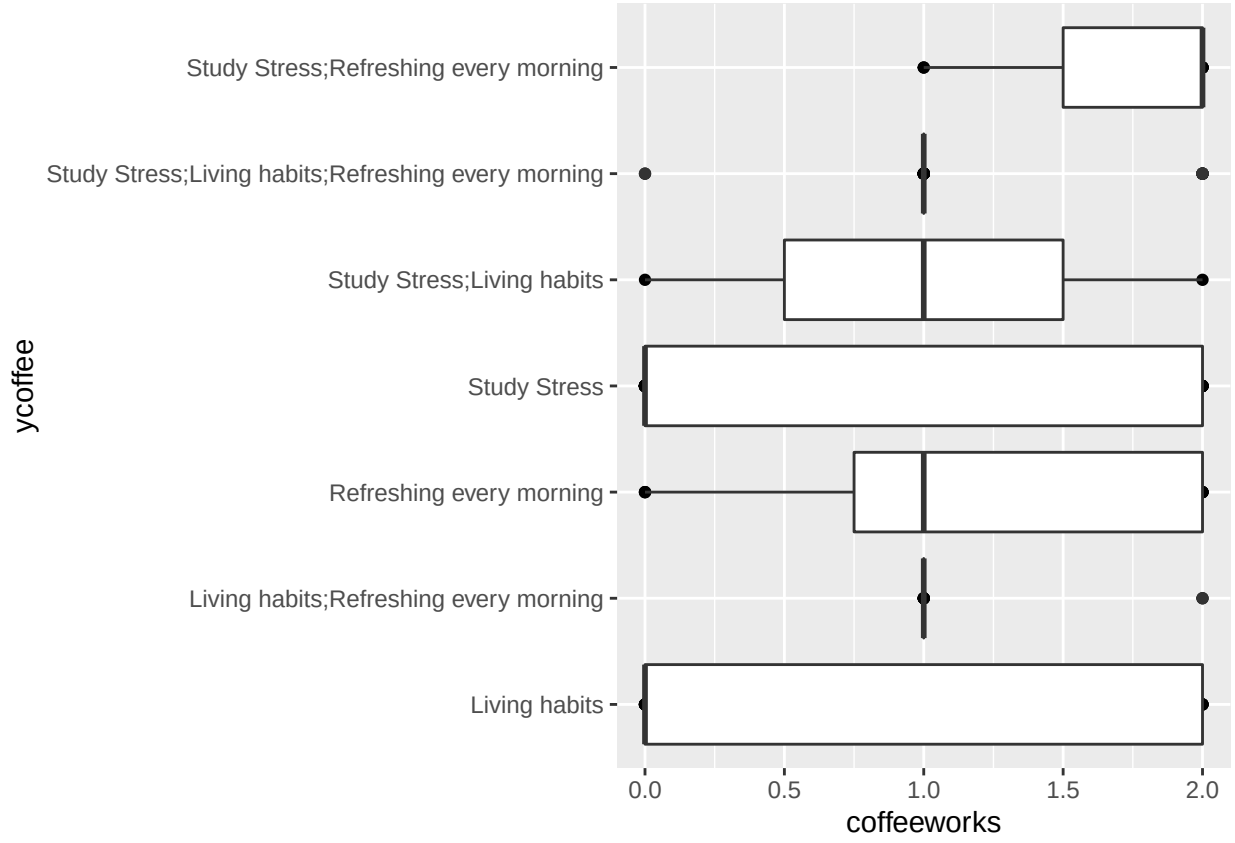
```
coffeesurvey2_df <- na.omit(coffeesurvey2_df)

#summary(coffeesurvey2_df)
```

```
head(coffeesurvey2_df)
```

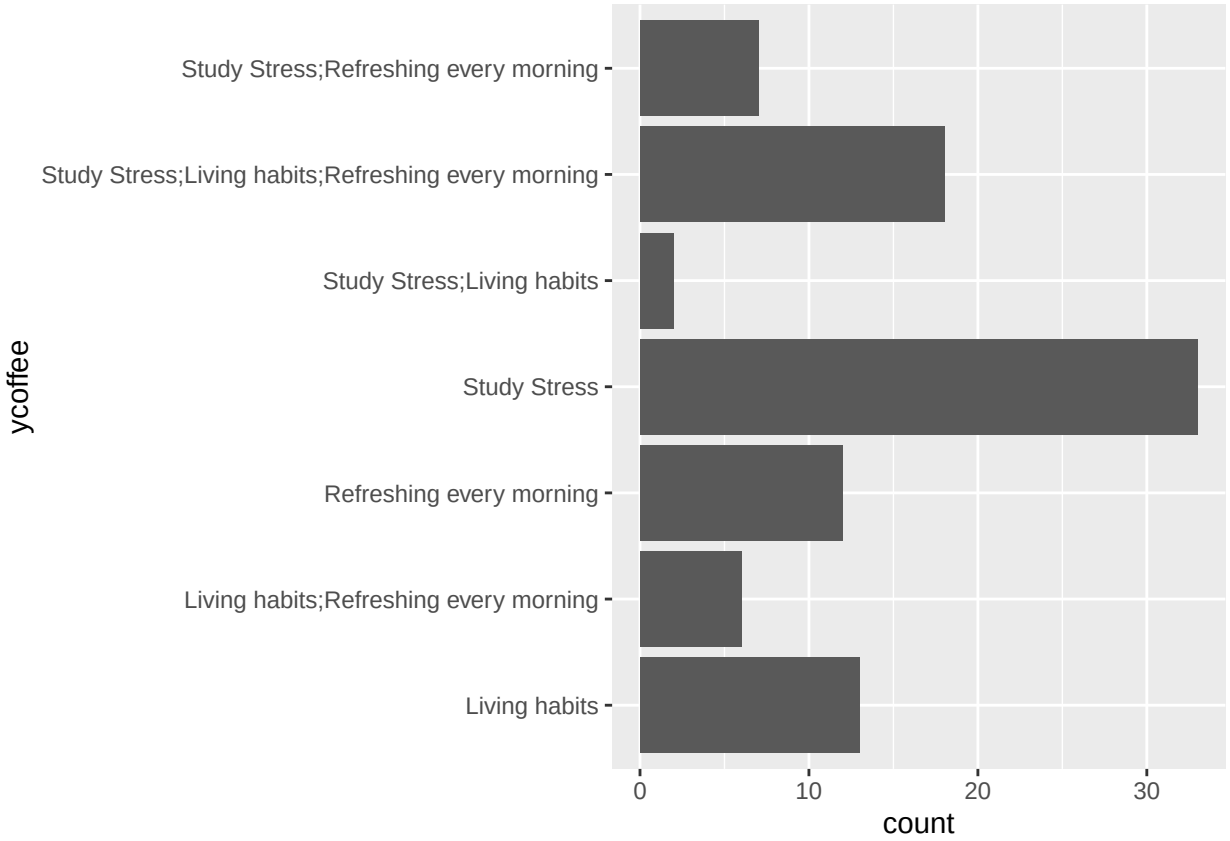
```
##   drinkcoffee totalcups                                    ycoffee
## 1           1         2                                Study Stress
## 2           1         2                     Refreshing every morning
```

```
## 3           1         3                                        Living habits
## 4           0         1                                        Living habits
## 5           0         0                                         Study Stress
## 6           1         3 Study Stress;Living habits;Refreshing every morning
##   coffeeworks
## 1           0
## 2           1
## 3           2
## 4           0
## 5           0
## 6           1
```

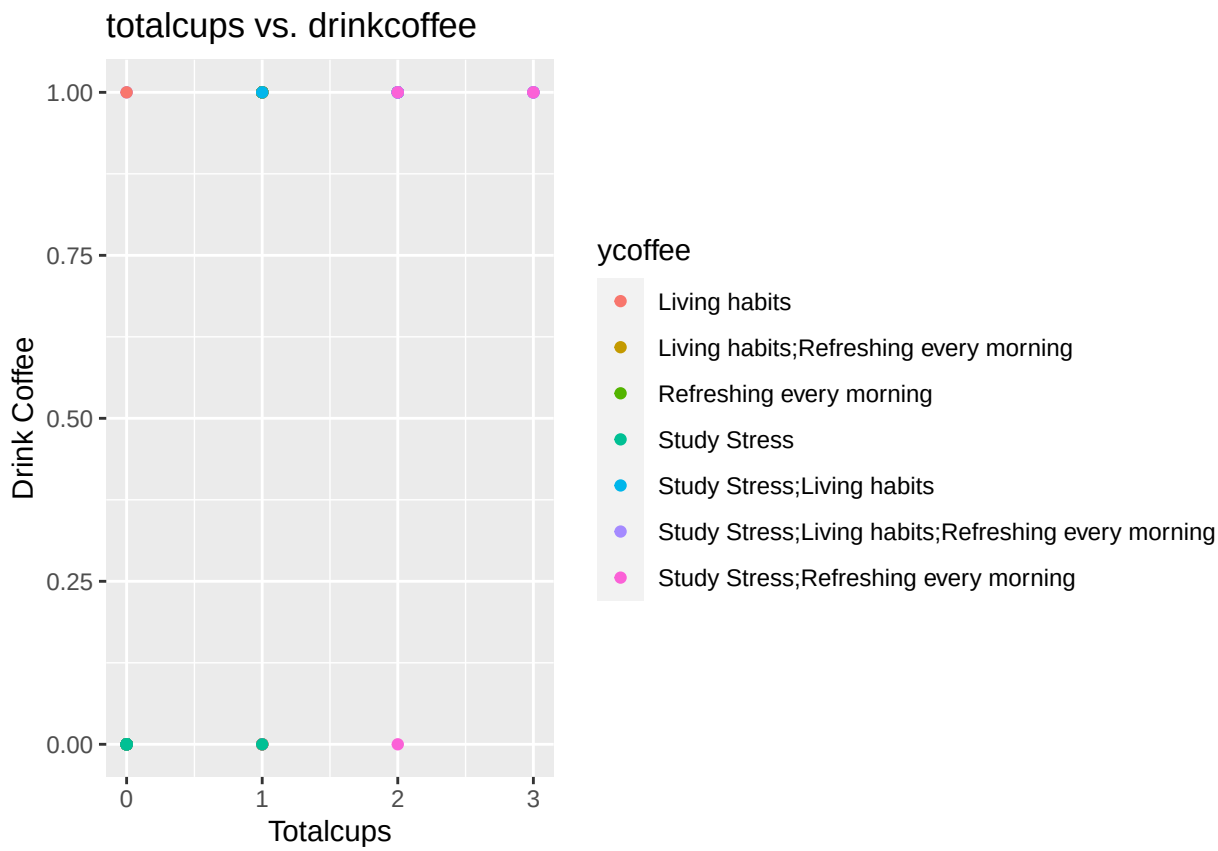
```
ggplot(coffeesurvey2_df, aes(x=coffeeworks, y=ycoffee)) + geom_point()+ geom_boxplot()
```




```
ggplot(coffeesurvey2_df, aes(y=ycoffee)) + geom_bar()
```




```
ggplot(coffeesurvey2_df, aes(x=totalcups, y=drinkcoffee, col=ycoffee)) + geom_point() +
  ggtitle("totalcups vs. drinkcoffee") + xlab("Totalcups") + ylab("Drink Coffee")
```




```
#coffee_lm <-  lm( ycoffee ~ coffeeworks + totalcups + drinkcoffee, data=coffeesurvey2_df)
#coffee_lm
```

```
coffeesurvey2_df$ycoffee <- as.factor(coffeesurvey2_df$ycoffee)
coffee_glm <- glm(ycoffee ~  drinkcoffee+ totalcups + coffeeworks , data = coffeesurvey2_df, family = bi
summary(coffee_glm)
```

```
##
## Call:
## glm(formula = ycoffee ~ drinkcoffee + totalcups + coffeeworks,
##     family = binomial(), data = coffeesurvey2_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2242   0.4264   0.5240   0.5841   0.7540
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.4802     0.4993   2.965  0.00303 **
## drinkcoffee  -0.3678     0.8763  -0.420  0.67472
## totalcups     0.1337     0.4830   0.277  0.78193
## coffeeworks   0.4359     0.4209   1.036  0.30037
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 74.641  on 90  degrees of freedom
## Residual deviance: 73.282  on 87  degrees of freedom
## AIC: 81.282
##
## Number of Fisher Scoring iterations: 4
```

```
summary(coffeesurvey2_df$ycoffee)
```

```
##                                       Living habits
##                                                  13
##              Living habits;Refreshing every morning
##                                                   6
##                            Refreshing every morning
##                                                  12
##                                        Study Stress
##                                                  33
##                           Study Stress;Living habits
##                                                   2
## Study Stress;Living habits;Refreshing every morning
##                                                  18
##               Study Stress;Refreshing every morning
##                                                   7
```

Third data set

```
coffeechainFinal_df <- read.csv("data/CoffeeChainFinal.csv")
head(coffeechainFinal_df)
```

```
##   Area.Code    Ddate  Market  Market.Size           Product Product.Type
## 1       970 1/1/2012 Central Major Market Decaf Irish Cream       Coffee
## 2       719 2/1/2012 Central Major Market Decaf Irish Cream       Coffee
## 3       720 3/1/2012 Central Major Market Decaf Irish Cream       Coffee
## 4       303 4/1/2012 Central Major Market Decaf Irish Cream       Coffee
## 5       720 5/1/2012 Central Major Market Decaf Irish Cream       Coffee
## 6       719 6/1/2012 Central Major Market Decaf Irish Cream       Coffee
##      State  Type Caffeine..mg. Budget.Cogs Budget.Margin Budget.Profit
## 1 Colorado Decaf             2         100           140           110
## 2 Colorado Decaf             2         100           140           110
## 3 Colorado Decaf             2         100           140           110
## 4 Colorado Decaf             2         100           150           120
## 5 Colorado Decaf             2         110           150           120
## 6 Colorado Decaf             2         130           180           140
##   Budget.Sales Coffee.Sales Cogs Inventory Margin Marketing Number.of.Records
## 1          240          234   95       821    139        26                1
## 2          240          232   95       809    137        26                1
## 3          240          234   95       799    139        26                1
## 4          250          245  100       822    145        28                1
## 5          260          256  104       871    152        29                1
## 6          310          301  123       947    178        34                1
##   Number.Of.Records Profit Total.Expenses
## 1                 1    101             38
## 2                 1     99             38
## 3                 1    101             38
## 4                 1    105             40
## 5                 1    112             40
```

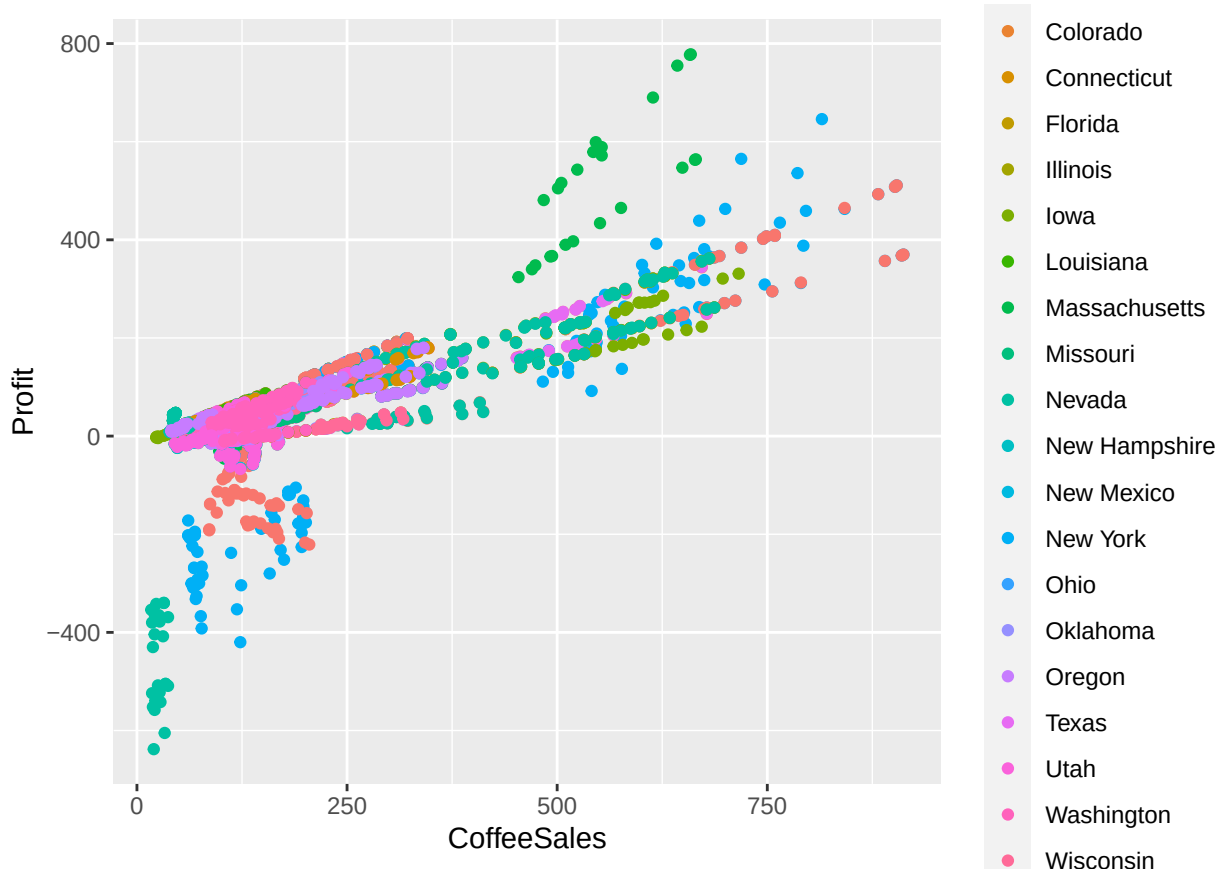
```
## 6                 1    132             46
```

```
colnames(coffeechainFinal_df)[14] <-"CoffeeSales"
colnames(coffeechainFinal_df)[9]<- "Caffeine"
head(coffeechainFinal_df)
```

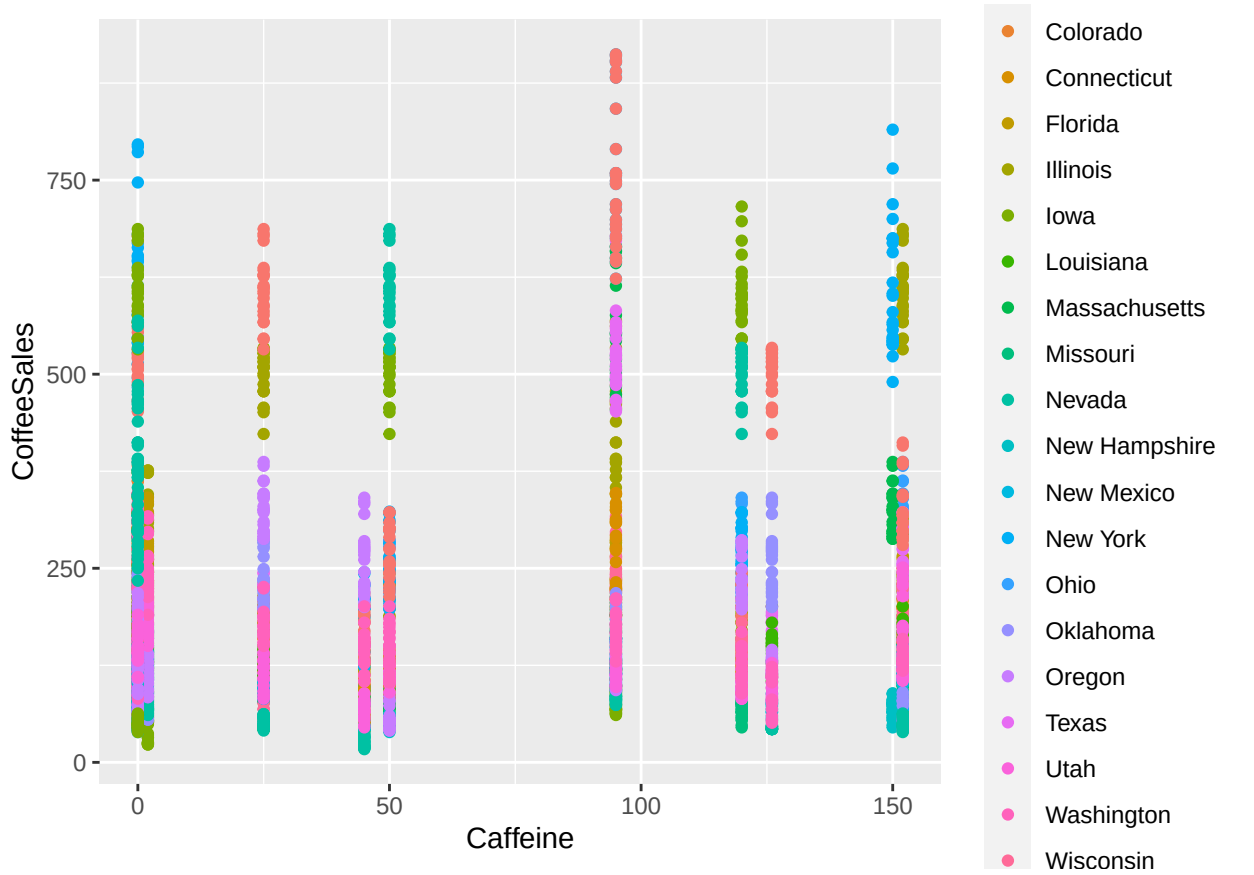
```
##   Area.Code    Ddate  Market  Market.Size           Product Product.Type
## 1       970 1/1/2012 Central Major Market Decaf Irish Cream       Coffee
## 2       719 2/1/2012 Central Major Market Decaf Irish Cream       Coffee
## 3       720 3/1/2012 Central Major Market Decaf Irish Cream       Coffee
## 4       303 4/1/2012 Central Major Market Decaf Irish Cream       Coffee
## 5       720 5/1/2012 Central Major Market Decaf Irish Cream       Coffee
## 6       719 6/1/2012 Central Major Market Decaf Irish Cream       Coffee
##      State  Type Caffeine Budget.Cogs Budget.Margin Budget.Profit Budget.Sales
## 1 Colorado Decaf        2         100           140           110          240
## 2 Colorado Decaf        2         100           140           110          240
## 3 Colorado Decaf        2         100           140           110          240
## 4 Colorado Decaf        2         100           150           120          250
## 5 Colorado Decaf        2         110           150           120          260
## 6 Colorado Decaf        2         130           180           140          310
##   CoffeeSales Cogs Inventory Margin Marketing Number.of.Records
## 1         234   95       821    139        26                1
## 2         232   95       809    137        26                1
## 3         234   95       799    139        26                1
## 4         245  100       822    145        28                1
## 5         256  104       871    152        29                1
## 6         301  123       947    178        34                1
##   Number.Of.Records Profit Total.Expenses
## 1                 1    101             38
## 2                 1     99             38
## 3                 1    101             38
## 4                 1    105             40
## 5                 1    112             40
## 6                 1    132             46
```

```
coffeechainFinal_df <- na.omit(coffeechainFinal_df)
```

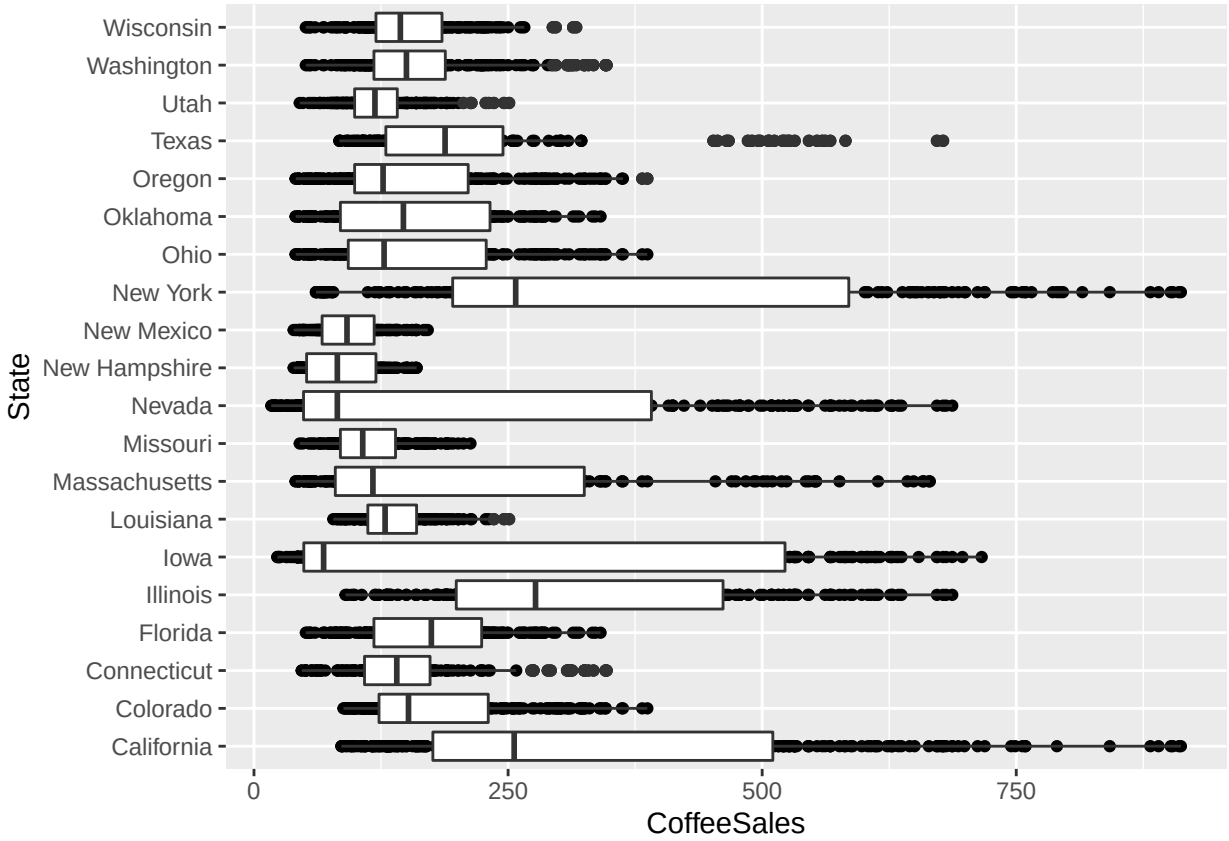
```
ggplot(data = coffeechainFinal_df, aes(x= CoffeeSales, y=Profit, color=State)) +  geom_point()
```



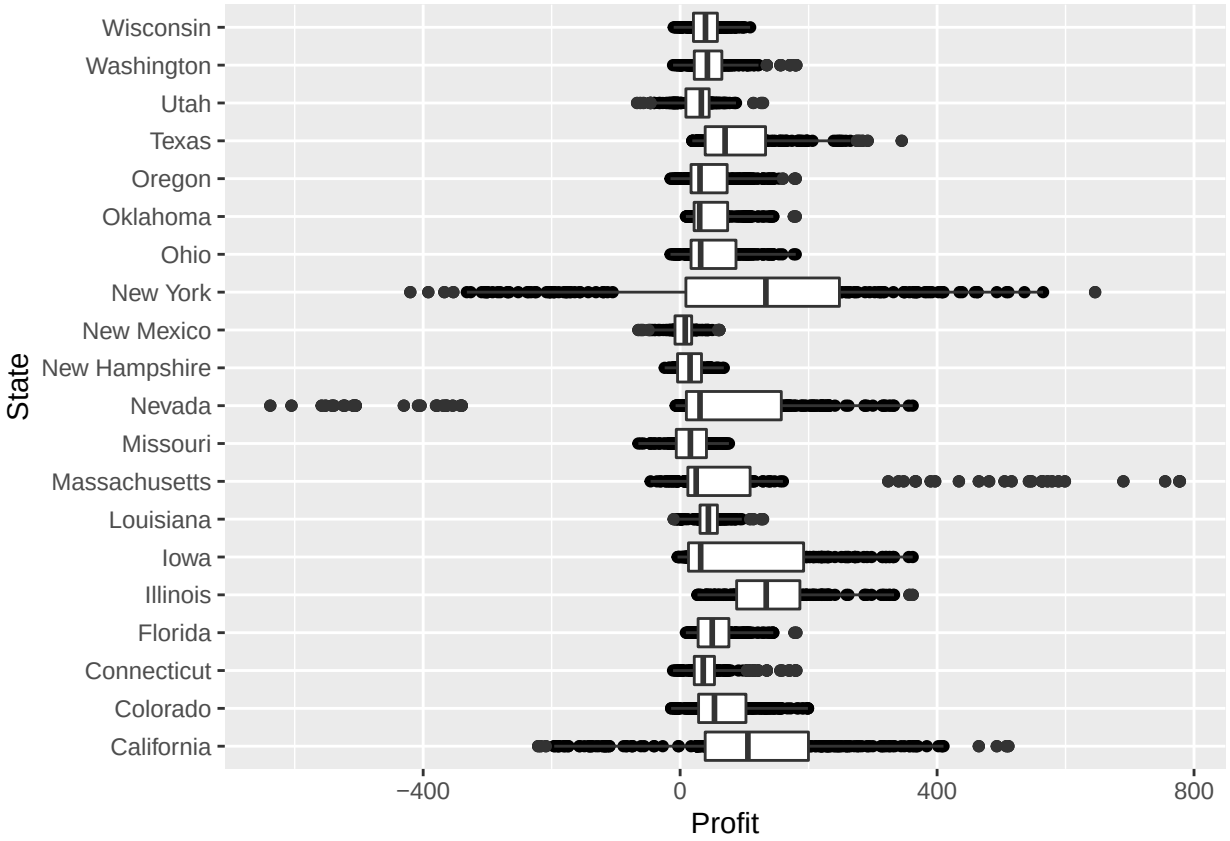

```
ggplot(data = coffeechainFinal_df, aes(x= Caffeine, y=CoffeeSales, color=State)) +  geom_point()
```



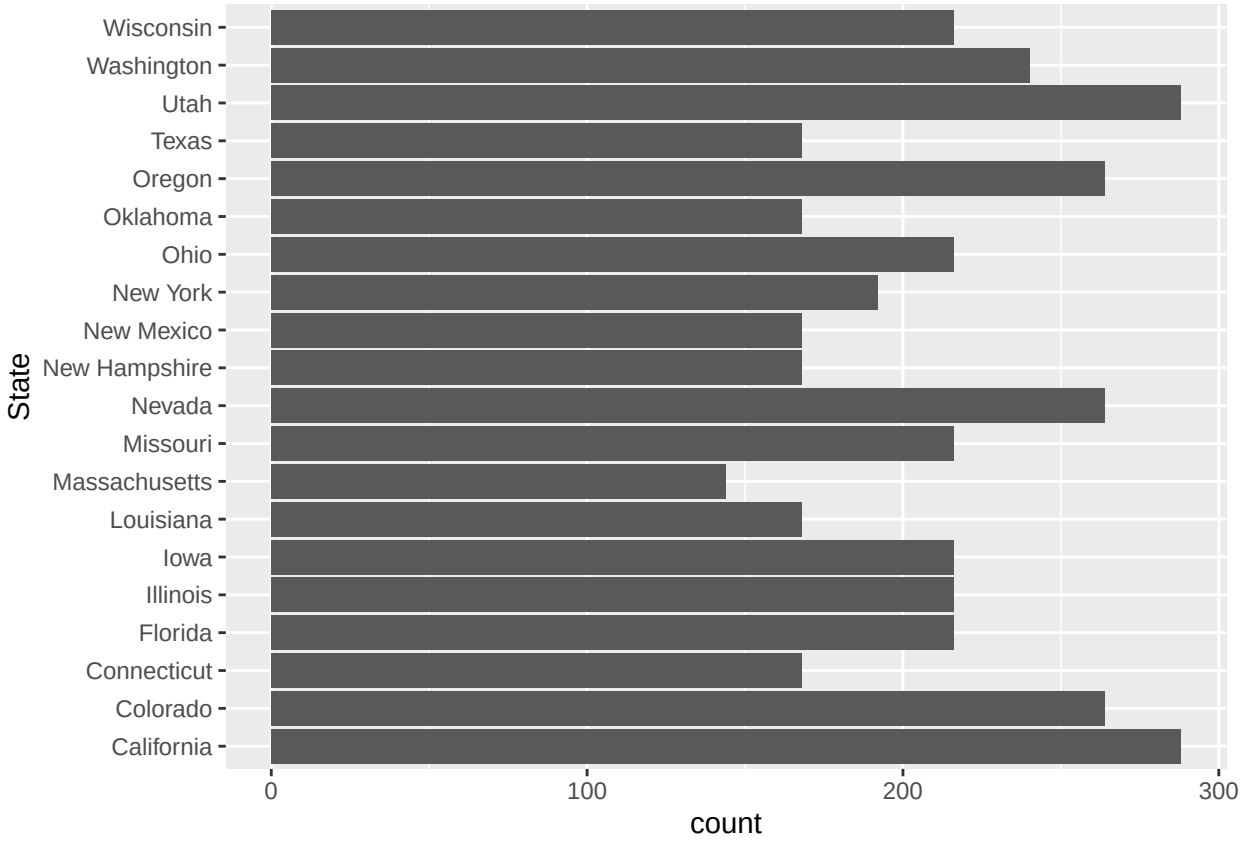

```
ggplot(coffeechainFinal_df, aes(x=CoffeeSales, y=State)) + geom_point()+ geom_boxplot()
```




```
ggplot(coffeechainFinal_df, aes(x=Profit, y=State)) + geom_point()+ geom_boxplot()
```




```
ggplot(coffeechainFinal_df, aes(y=State)) + geom_bar()
```




```
coffeechainFinal_df <- read.csv("data/CoffeeChainFinal.csv")
colnames(coffeechainFinal_df)[14] <-"CoffeeSales"
```

```
summary(coffeechainFinal_df$State)
```

```
##    Length     Class      Mode
##      4248 character character
```

```
coffeesurvey2_df <- na.omit(coffeesurvey2_df)
```

```
coffeechainFinal_df$State<- as.factor(coffeechainFinal_df$State)
coffeech <- glm(State ~ CoffeeSales + Profit, data = coffeechainFinal_df, family = binomial())
summary(coffeech)
```

```
##
## Call:
## glm(formula = State ~ CoffeeSales + Profit, family = binomial(),
##     data = coffeechainFinal_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6751   0.2436   0.2815   0.3320   1.5244
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.0595602  0.1260534  32.205  < 2e-16 ***
## CoffeeSales -0.0076922  0.0005234 -14.697  < 2e-16 ***
## Profit       0.0058622  0.0007311   8.018 1.08e-15 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2106.2  on 4247  degrees of freedom
## Residual deviance: 1852.5  on 4245  degrees of freedom
## AIC: 1858.5
##
## Number of Fisher Scoring iterations: 6
```