Nguyen_Tyler_Final_Exam

Stats 20 Spring 2022

6/8/2022

```
load("final_data.RData")
load("Store.RData")
data <- merge(yr2016, yr2018, all = TRUE)
data2 <- merge(data, yr2020, all = TRUE)
finalIowa <- merge(data2, store, by.x = "Store Number")
```

Firstly, I have to load the data into environment and merge data sets together

1

A

```
nrow(finalIowa)
```

```
## [1] 11882
```

```
ncol(finalIowa)
```

```
## [1] 16
```

The data set has 11882 rows and 16 columns. The name of the columns are:

```
names(finalIowa)
```

```
##  [1] "Store Number"          "Invoice/Item Number"   "Date"
##  [4] "Category"              "Vendor Number"         "Item Number"
##  [7] "State Bottle Retail"   "Bottles Sold"          "Sale (Dollars)"
## [10] "Volume Sold (Liters)"  "Volume Sold (Gallons)" "Store Name"
## [13] "Address"               "City"                  "Zip Code"
## [16] "County"
```

Now, I look at what type of variables I am working with.

```
sapply(finalIowa, class) |> table()
```

```
##
## character      Date
##        15         1
```

I know by the names of the variables that I need to convert some of these character variables to numeric. I look at the first few rows of data to know which ones to convert

```
head(finalIowa)
```

```
##   Store Number Invoice/Item Number       Date Category Vendor Number
## 1         2106     INV-11726900055 2018-04-26  1081600           421
## 2         2106     INV-12519400024 2018-06-07  1031200           434
## 3         2106        S33152600063 2016-06-30  1081200           260
## 4         2106        S32413900065 2016-05-19  1081200           260
## 5         2106     INV-01363700059 2016-11-03  1081300           065
## 6         2106     INV-02026600060 2016-12-08  1081300           434
##   Item Number State Bottle Retail Bottles Sold Sale (Dollars)
## 1       64866              $13.50           24        $324.00
## 2       35179              $10.50           12        $126.00
## 3       68037              $24.75           12        $297.00
## 4       74090              $25.50           12        $306.00
## 5       85526               $7.88           24        $189.12
## 6       79336               $7.50           12         $85.56
##   Volume Sold (Liters) Volume Sold (Gallons)                 Store Name
## 1                   18                 4.76 Hillstreet News and Tobacco
## 2                    9                 2.38 Hillstreet News and Tobacco
## 3                   12                 3.17 Hillstreet News and Tobacco
## 4                    9                 2.38 Hillstreet News and Tobacco
## 5                   18                 4.76 Hillstreet News and Tobacco
## 6                    9                 2.38 Hillstreet News and Tobacco
##        Address        City Zip Code     County
## 1 2217 College Cedar Falls    50613 BLACK HAWK
## 2 2217 College Cedar Falls    50613 BLACK HAWK
## 3 2217 College Cedar Falls    50613 BLACK HAWK
## 4 2217 College Cedar Falls    50613 BLACK HAWK
## 5 2217 College Cedar Falls    50613 BLACK HAWK
## 6 2217 College Cedar Falls    50613 BLACK HAWK
```

From this, I know to convert the columns: “State Bottle Retail”, “Bottles Sold”, “Sale (Dollars)”, “Volume Sold(Liters)”, and “Volume Sold (Gallons)”.

```
finalIowa$`State Bottle Retail` <- gsub("\\$|,", "", finalIowa$`State Bottle Retail`)
finalIowa$`Sale (Dollars)` <- gsub("\\$|,", "", finalIowa$`Sale (Dollars)`)
finalIowa[7:11] <- sapply(finalIowa[7:11], as.numeric)
```

Now we can see what we are working with:

```
numeric <- unlist(lapply(finalIowa, is.numeric), use.names = FALSE)
char <- unlist(lapply(finalIowa, is.character), use.names = FALSE)
summary(finalIowa[numeric])
```

```
##  State Bottle Retail  Bottles Sold     Sale (Dollars)     Volume Sold (Liters)
##  Min.   :  1.34      Min.   :   1.00   Min.   :    0.00   Min.   :   0.050
##  1st Qu.:  8.51      1st Qu.:   2.00   1st Qu.:   32.26   1st Qu.:   1.500
##  Median : 12.39      Median :   6.00   Median :   73.53   Median :   4.800
```

```
##  Mean   : 15.69      Mean   :  10.81   Mean   :  144.09   Mean   :   9.393
##  3rd Qu.: 19.49      3rd Qu.:  12.00   3rd Qu.:  148.56   3rd Qu.:  10.500
##  Max.   :375.00      Max.   :1728.00   Max.   :39191.04   Max.   :3024.000
##  Volume Sold (Gallons)
##  Min.   :  0.010
##  1st Qu.:  0.400
##  Median :  1.270
##  Mean   :  2.479
##  3rd Qu.:  2.770
##  Max.   :798.860
```

```
sapply(names(finalIowa)[char], function(x) table(finalIowa[x], useNA = "always", dnn = x) |> s
ort() |> tail())
```

```
##      Store Number Invoice/Item Number Category Vendor Number Item Number
## 2512           69                   1      496           593          65
## 2670           69                   1      578           699          69
## 4829           78                   1      617           916          78
## 2572           87                   1      730           940         101
## 2603           88                   1     1159           987         106
## 2633          101                   1     1386          2063         116
##      Store Name Address City Zip Code County
## 2512         69      69  349      206    393
## 2670         69      69  352      225    594
## 4829         78      78  355      230    663
## 2572         87      87  542      267    762
## 2603         88      88  786      317   1017
## 2633        101     136  930      317   2132
```

Time to assess missing values

```
sapply(names(finalIowa), function(ix) round(prop.table(table(is.na(finalIowa[ix]), dnn = ix)),
 digit = 5))
```

```
## $`Store Number`
## Store Number
## FALSE
##     1
##
## $`Invoice/Item Number`
## Invoice/Item Number
## FALSE
##     1
##
## $Date
## Date
## FALSE
##     1
##
## $Category
## Category
```

```
##   FALSE    TRUE
## 0.99815 0.00185
##
## $`Vendor Number`
## Vendor Number
## FALSE
##     1
##
## $`Item Number`
## Item Number
## FALSE
##     1
##
## $`State Bottle Retail`
## State Bottle Retail
## FALSE
##     1
##
## $`Bottles Sold`
## Bottles Sold
## FALSE
##     1
##
## $`Sale (Dollars)`
## Sale (Dollars)
## FALSE
##     1
##
## $`Volume Sold (Liters)`
## Volume Sold (Liters)
## FALSE
##     1
##
## $`Volume Sold (Gallons)`
## Volume Sold (Gallons)
## FALSE
##     1
##
## $`Store Name`
## Store Name
## FALSE
##     1
##
## $Address
## Address
##   FALSE    TRUE
## 0.99705 0.00295
##
## $City
## City
##   FALSE    TRUE
## 0.99705 0.00295
##
## $`Zip Code`
```

```
## Zip Code
##   FALSE    TRUE
## 0.99705 0.00295
##
## $County
## County
##   FALSE    TRUE
## 0.99344 0.00656
```

We see that there are barely any missing values so we are good to start analyzing the data and make the desired table

```
finalIowa$Year <- format(finalIowa$Date, format = "%Y")
finalIowa$Month <- format(finalIowa$Date, format = "%m")
tapply(finalIowa$`Sale (Dollars)`, finalIowa$Year, FUN = sum)
```

```
##     2016     2018     2020
## 565769.8 669415.4 476914.5
```

I created a new variable “Year” to easily analyze the three years we want.

B

1

```
head(sort(table(finalIowa$`Zip Code`[which(finalIowa$Year == 2016)]), decreasing = TRUE), n =
10)
```

```
##
## 52402 50010 52240 50613 50401 51501 52001 52807 50317 52404
##   125   100    81    80    65    63    63    59    58    56
```

Year 2016

```
head(sort(table(finalIowa$`Zip Code`[which(finalIowa$Year == 2018)]), decreasing = TRUE), n =
10)
```

```
##
## 50010 52240 52402 50265 52001 50613 51501 52404 52804 52241
##   134   129   124   104   104    99    97    83    82    76
```

Year 2018

```
head(sort(table(finalIowa$`Zip Code`[which(finalIowa$Year == 2020)]), decreasing = TRUE), n =
10)
```

```
##
## 50010 52402 52001 50322 52240 50702 52241 52404 50265 50613
##    83    68    58    57    57    53    53    53    51    51
```

Year 2020

2

```
zipcodeSale2016 <- tapply(finalIowa$`Sale (Dollars)`[which(finalIowa$Year == 2016)], finalIowa
$`Zip Code`[which(finalIowa$Year == 2016)] , FUN = sum)
head(sort(zipcodeSale2016, decreasing = TRUE), n=10)
```

```
##    50266    51501    50320    52402    50010    52240    52807    52001
## 55928.19 33665.11 26091.85 25883.44 21514.89 16288.45 13995.96  9770.42
##    52241    50314
##  9043.10  8971.18
```

2016

```
zipcodeSale2018 <- tapply(finalIowa$`Sale (Dollars)`[which(finalIowa$Year == 2018)], finalIowa
$`Zip Code`[which(finalIowa$Year == 2018)] , FUN = sum)
head(sort(zipcodeSale2018, decreasing = TRUE), n=10)
```

```
##    50320    52240    50613    50314    51501    52807    52001    52402
## 34197.25 32995.67 27933.58 22340.96 18446.86 17066.42 16472.71 15790.78
##    52722    50010
## 15096.32 14039.98
```

2018

```
zipcodeSale2020 <- tapply(finalIowa$`Sale (Dollars)`[which(finalIowa$Year == 2020)], finalIowa
$`Zip Code`[which(finalIowa$Year == 2020)] , FUN = sum)
head(sort(zipcodeSale2020, decreasing = TRUE), n=10)
```

```
##    50322    52402    50314    52722    52807    52001    50320    50010
## 18921.37 13603.39 12350.17 12294.81 12243.93 12199.97 11121.07 10800.45
##    52314    52002
##  9771.23  9433.56
```

2020

3

```
medianSale2016 <- tapply(finalIowa$`Sale (Dollars)`[which(finalIowa$Year == 2016)], finalIowa$
`Zip Code`[which(finalIowa$Year == 2016)] , FUN = median)
head(sort(medianSale2016, decreasing = TRUE), n=10)
```

```
##  50109  50628  50530  52057  50435  50514  52036  52571  52333  50421
## 484.92 340.08 338.76 287.94 283.56 282.24 279.12 279.00 276.12 271.32
```

2016

```
medianSale2018 <- tapply(finalIowa$`Sale (Dollars)`[which(finalIowa$Year == 2018)], finalIowa$
`Zip Code`[which(finalIowa$Year == 2018)] , FUN = median)
head(sort(medianSale2018, decreasing = TRUE), n=10)
```

```
##   50169   50435   50061   51109   50261   50207   52224   50048   50071   52227
## 413.760 340.080 288.000 284.535 284.280 282.240 280.320 269.880 230.160 206.880
```

2018

```
medianSale2020 <- tapply(finalIowa$`Sale (Dollars)`[which(finalIowa$Year == 2020)], finalIowa$
`Zip Code`[which(finalIowa$Year == 2020)] , FUN = median)
head(sort(medianSale2020, decreasing = TRUE), n=10)
```

```
##  50124  52031  50212  50543  51453  52158  52324  50129  51109  50854
## 592.80 452.37 351.36 323.28 322.80 319.44 294.21 280.68 276.00 265.50
```

2020

C

```
tapply(finalIowa$`State Bottle Retail`, finalIowa$Year, FUN = median)
```

```
##  2016  2018  2020
## 12.38 12.38 13.11
```

```
tapply(finalIowa$`State Bottle Retail`, finalIowa$Year, FUN = mean)
```

```
##     2016     2018     2020
## 15.04960 15.61447 16.56893
```

It appears that over the years, the price of alcohol has increased. We can see this from the code above when I calculated the mean and median prices by year. The reason for this is probably due to inflation, but also COVID. This is because we see a larger jump from 2018 to 2020 than from 2016 to 2018, meaning that COVID affected the economy. This may because it was harder to export alcohol as a result of COVID 19 restrictions on trade.

D

I believe that the most influential vendors means the most volume sold. As I result, I will analyze the most amount of alcohol sold in liters by retailer.

```
head(sort(tapply(finalIowa$`Volume Sold (Gallons)`, finalIowa$`Store Name`, FUN = sum), decrea
sing = TRUE), n= 10)
```

```
##        Hy-Vee #3 / BDI / Des Moines         Costco Wholesale #788 / WDM
##                              957.83                              900.70
##     Hy-Vee Food Store / Cedar Falls                     Wilkie Liquors
##                              626.42                              521.96
```

```
## Hy-Vee Wine and Spirits / Iowa City                    Central City 2
##                              513.30                             488.12
##      Sam's Club 8162 / Cedar Rapids        Sam's Club 8238 / Davenport
##                              478.63                             385.35
##        I-80 Liquor / Council Bluffs                   Hy-Vee / Waukee
##                              309.64                             296.95
```

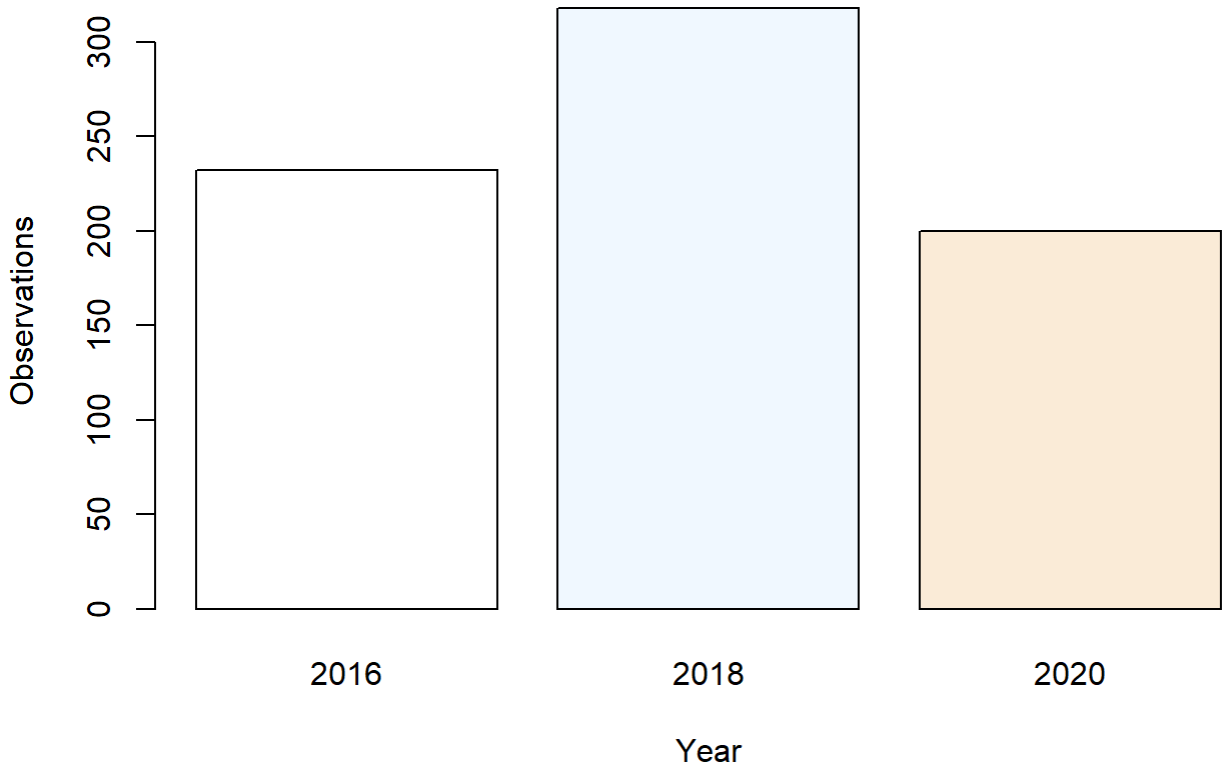
Hy-Vee appears to be the retailer who is most influential as they have the most alcohol sold by volume.

E

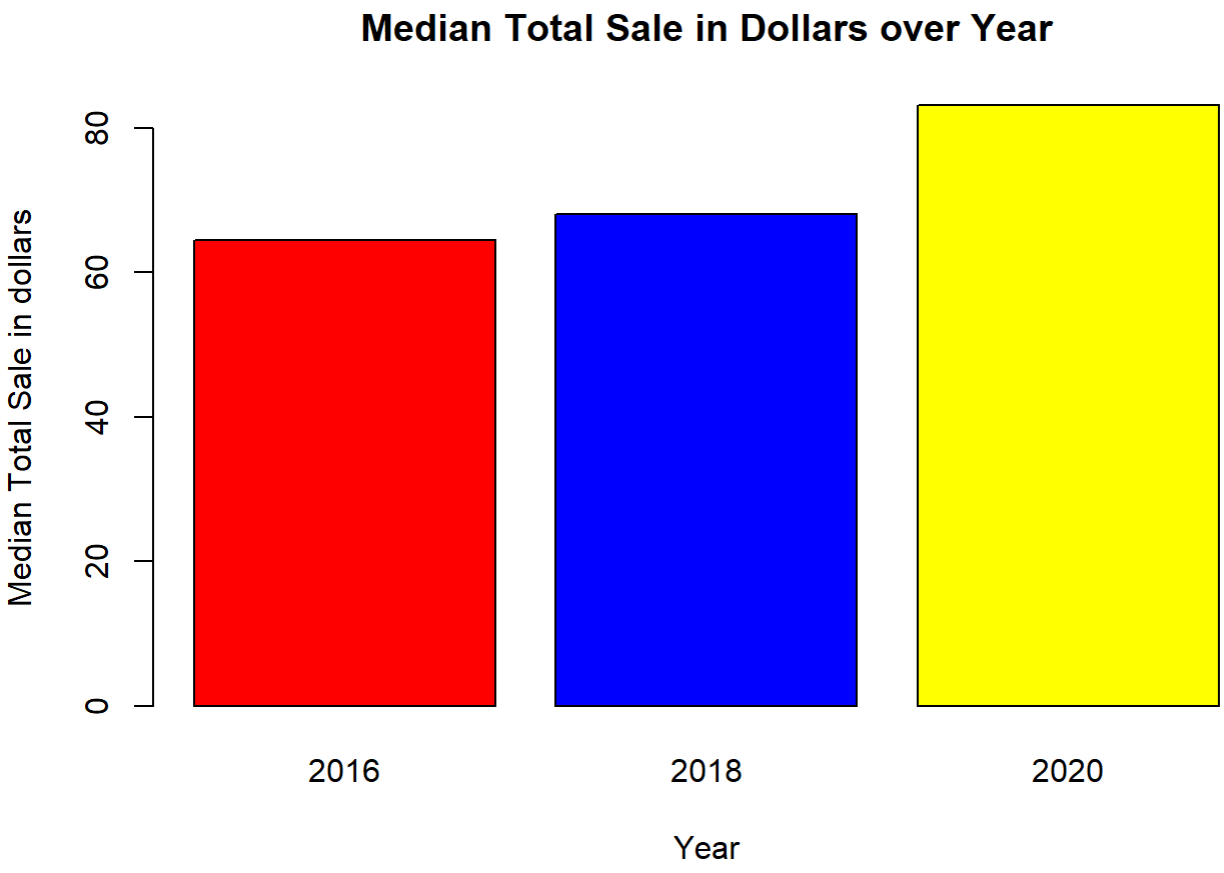
I have used the apply functions in part B, C, and D of problem 1. Also in problem 2 and problem 3.

2

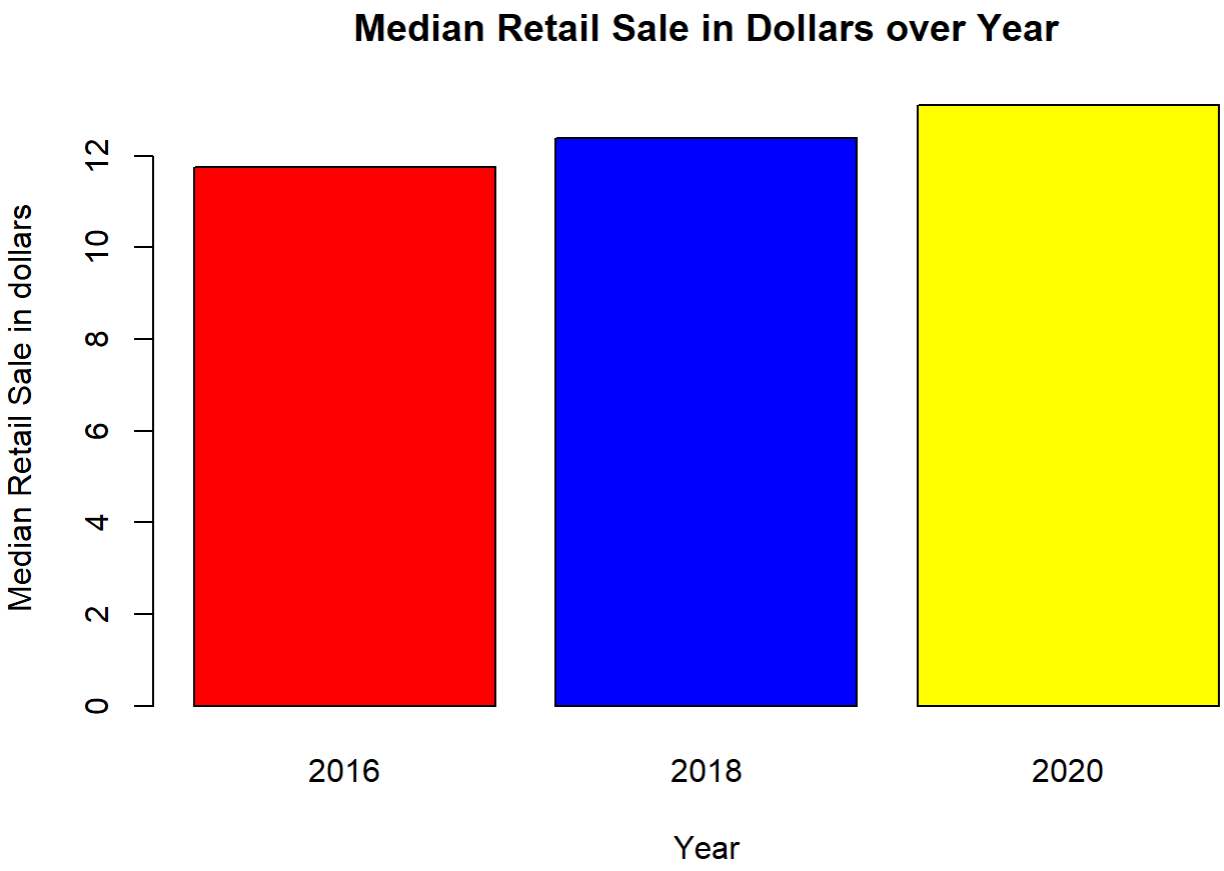
```
set.seed(105764460)

index <- sample(nrow(finalIowa), 750)

sampleData <- finalIowa[index,]

sampleVolumebyYear <- tapply(sampleData$`Volume Sold (Liters)`, sampleData$Year, median)
sampleBottlesbyYear <- tapply(sampleData$`Bottles Sold`, sampleData$Year, median)
samplePricebyYear <- tapply(sampleData$`Sale (Dollars)`, sampleData$Year, median)
sampleRetailbyYear <- tapply(sampleData$`State Bottle Retail`, sampleData$Year, median)

barplot(table(sampleData$Year), col = colors(), xlab = "Year", ylab = "Observations")
```




```
barplot(samplePricebyYear, xlab = "Year", ylab = "Median Total Sale in dollars", col = c("red"
, "blue", "yellow"), main = "Median Total Sale in Dollars over Year")
```




```
barplot(sampleRetailbyYear, xlab = "Year", ylab = "Median Retail Sale in dollars", col = c("red", "blue", "yellow"), main = "Median Retail Sale in Dollars over Year")
```



3

I would outline the effect of COVID-19 on alcohol sales by computing the median price of sales by Year. The reason I would do this is to see the effect which COVID-19 had on the alcohol prices. And as we could see from question 2 and its bar plots, it is safe to say that COVID 19 did affect the price of alcohol as we can see a sharp increase in the median total sales and the median retail price of alcohol. For instance, the median total price increased from 68.04 dollars to 83.19 dollars, a drastic increase when compared to the increase of 64.44 dollars to 68.04 dollars from 2016-2018.

```
samplePricebyYear <- tapply(sampleData$`Sale (Dollars)`, sampleData$Year, median)
samplePricebyYear
```

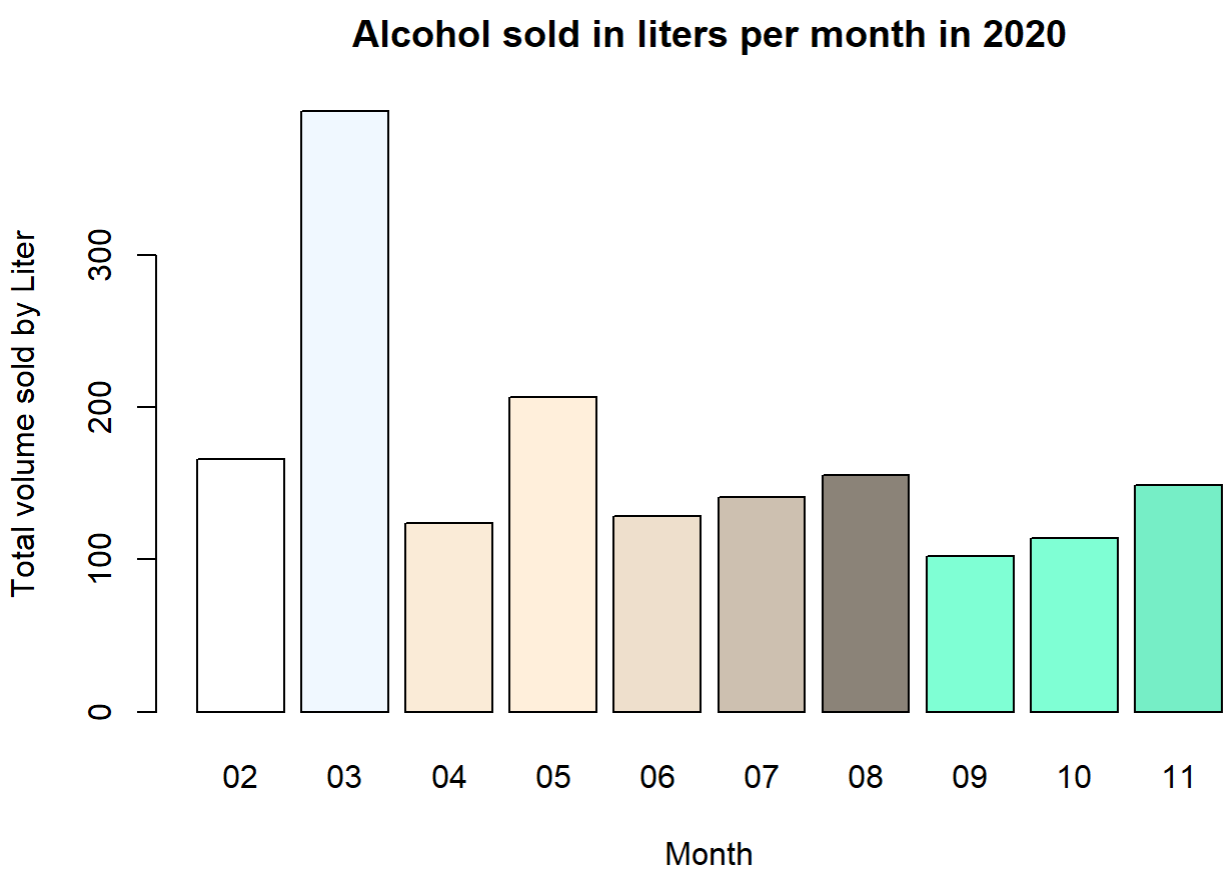
```
##  2016  2018  2020
## 64.44 68.04 83.19
```

I used median to nullify the effect of outliers, and the reason COVID 19 affected the price is because it may have been harder to export alcohol from other countries as a result of COVID 19 restrictions.

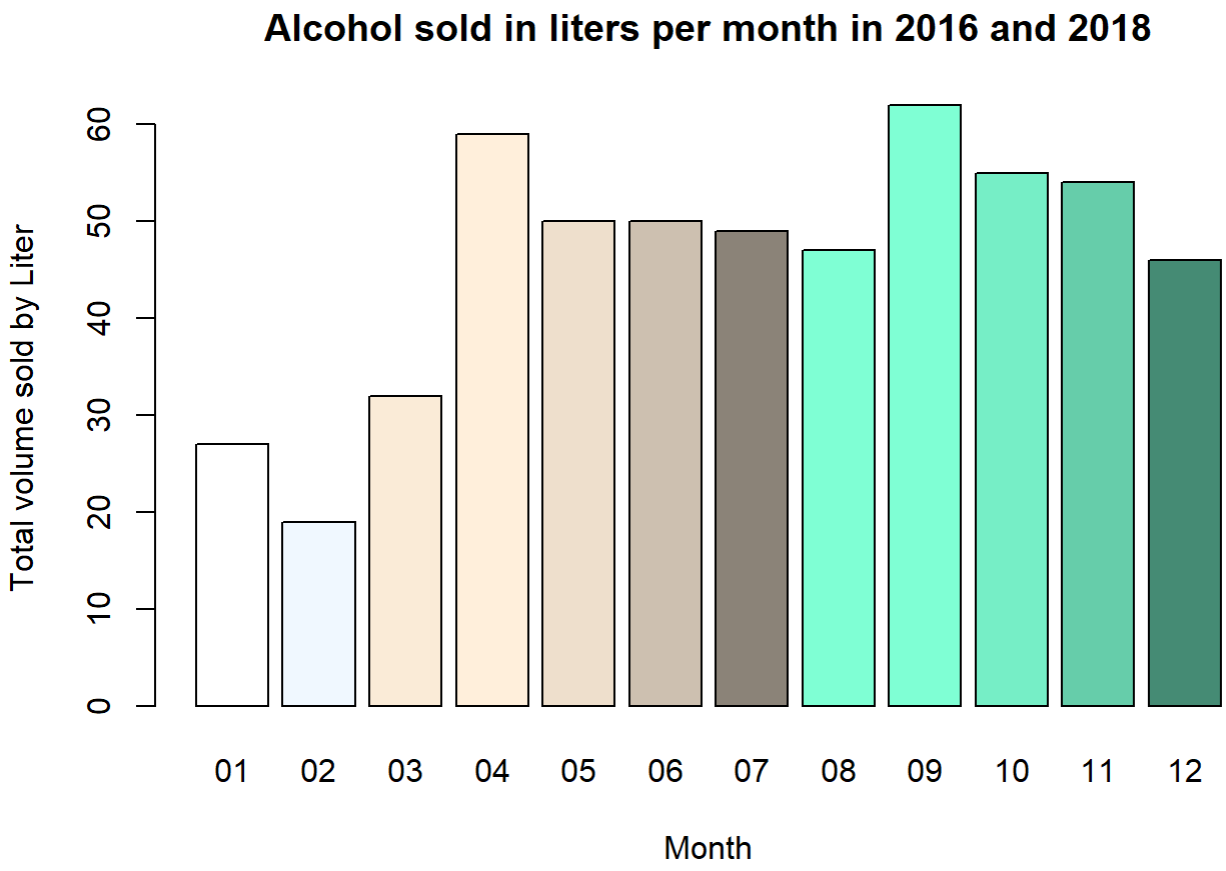
In addition, I will attempt to find the effect that COVID 19 had on the amount of alcohol sold. To illustrate the point that the amount of volume sold is effected by COVID 19, I will construct two barplots by month. The reason why I do this is to find the effect which the national emergency, implemented in March, had on the amount of alcohol sold. One barplot is of 2020 and the other is of 2016&2018 combined.

```
sampledata2020 <- sampleData[which(sampleData$Year ==2020),]
volumebyMonth2020 <- tapply(sampledata2020$`Volume Sold (Liters)`, sampledata2020$Month, sum)
```

```
barplot(volumebyMonth2020, col = colors(), xlab = "Month", ylab = "Total volume sold by Liter"
, main = "Alcohol sold in liters per month in 2020")
```




```
sampledata20162018 <- sampleData[which(sampleData$Year ==2016 | sampleData$Year ==2018),]
volumebyMonth20162018 <- tapply(sampledata20162018$`Volume Sold (Liters)`, sampledata20162018$
Month, sum)
barplot(table(sampledata20162018$Month), col = colors(), xlab = "Month", ylab = "Total volume
sold by Liter", main = "Alcohol sold in liters per month in 2016 and 2018")
```



As we can see from the data, the national emergency did in fact have an effect on the amount of alcohol sold. This is because in March 2020, the month of the declared national emergency, we see a huge spike in the amount of alcohol sold. To be sure that this is not just a coincidence, I plotted amount of alcohol sold by month in the years 206 and 2018 combined, and we can see that in March, there is no spike in the amount of alcohol sold, showing that COVID 19 did have an affect the amount of alcohol sold.