assignment 1

Load data

```
avocado = read.csv(file = "https://github.com/KazuMaeshima/Group-9-/raw/main/avocado.csv",
    header = TRUE)
```

##Provide a introduction of your analysis in the .RMD file so it can be produced in the output # this codes will introduce us how to use R Studio as part of our day to day data analysis. the avocado data set gives you the prices and the types of avocados produces in the USfor the year 2015

head

Print the structure of your dataset.

```
str(avocado)
```

```
## 'data.frame':    18249 obs. of  14 variables:
##  $ X           : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Date        : chr  "2015-12-27" "2015-12-20" "2015-12-13" "2015-12-06" ...
##  $ AveragePrice: num  1.33 1.35 0.93 1.08 1.28 1.26 0.99 0.98 1.02 1.07 ...
##  $ Total.Volume: num  64237 54877 118220 78992 51040 ...
##  $ X4046       : num  1037 674 795 1132 941 ...
##  $ X4225       : num  54455 44639 109150 71976 43838 ...
##  $ X4770       : num  48.2 58.3 130.5 72.6 75.8 ...
##  $ Total.Bags  : num  8697 9506 8145 5811 6184 ...
##  $ Small.Bags  : num  8604 9408 8042 5677 5986 ...
##  $ Large.Bags  : num  93.2 97.5 103.1 133.8 197.7 ...
##  $ XLarge.Bags : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ type        : chr  "conventional" "conventional" "conventional" "conventional" ...
##  $ year        : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
##  $ region      : chr  "Albany" "Albany" "Albany" "Albany" ...
```

##List the variables in your dataset

```
names(avocado)
```

```
##  [1] "X"            "Date"         "AveragePrice" "Total.Volume" "X4046"
##  [6] "X4225"        "X4770"        "Total.Bags"   "Small.Bags"   "Large.Bags"
## [11] "XLarge.Bags"  "type"         "year"         "region"
```

##Print the top 15 rows of your dataset.

```
head(avocado, 15)
```

```
##     X       Date AveragePrice Total.Volume   X4046     X4225  X4770 Total.Bags
## 1   0 2015-12-27         1.33     64236.62 1036.74  54454.85  48.16    8696.87
## 2   1 2015-12-20         1.35     54876.98  674.28  44638.81  58.33    9505.56
## 3   2 2015-12-13         0.93    118220.22  794.70 109149.67 130.50    8145.35
## 4   3 2015-12-06         1.08     78992.15 1132.00  71976.41  72.58    5811.16
## 5   4 2015-11-29         1.28     51039.60  941.48  43838.39  75.78    6183.95
## 6   5 2015-11-22         1.26     55979.78 1184.27  48067.99  43.61    6683.91
## 7   6 2015-11-15         0.99     83453.76 1368.92  73672.72  93.26    8318.86
## 8   7 2015-11-08         0.98    109428.33  703.75 101815.36  80.00    6829.22
## 9   8 2015-11-01         1.02     99811.42 1022.15  87315.57  85.34   11388.36
## 10  9 2015-10-25         1.07     74338.76  842.40  64757.44 113.00    8625.92
## 11 10 2015-10-18         1.12     84843.44  924.86  75595.85 117.07    8205.66
## 12 11 2015-10-11         1.28     64489.17 1582.03  52677.92 105.32   10123.90
## 13 12 2015-10-04         1.31     61007.10 2268.32  49880.67 101.36    8756.75
## 14 13 2015-09-27         0.99    106803.39 1204.88  99409.21 154.84    6034.46
## 15 14 2015-09-20         1.33     69759.01 1028.03  59313.12 150.50    9267.36
##    Small.Bags Large.Bags XLarge.Bags         type year region
## 1     8603.62      93.25           0 conventional 2015 Albany
## 2     9408.07      97.49           0 conventional 2015 Albany
## 3     8042.21     103.14           0 conventional 2015 Albany
## 4     5677.40     133.76           0 conventional 2015 Albany
## 5     5986.26     197.69           0 conventional 2015 Albany
## 6     6556.47     127.44           0 conventional 2015 Albany
## 7     8196.81     122.05           0 conventional 2015 Albany
## 8     6266.85     562.37           0 conventional 2015 Albany
## 9    11104.53     283.83           0 conventional 2015 Albany
## 10    8061.47     564.45           0 conventional 2015 Albany
## 11    7877.86     327.80           0 conventional 2015 Albany
## 12    9866.27     257.63           0 conventional 2015 Albany
## 13    8379.98     376.77           0 conventional 2015 Albany
## 14    5888.87     145.59           0 conventional 2015 Albany
## 15    8489.10     778.26           0 conventional 2015 Albany
```

##Write a user defined function using any of the variables from the data set

```
m <- c(45, 34, 34, 34, 67)
getmode <- function(m) {
    uniqv <- unique(m)
    uniqv[which.max(tabulate(match(m, uniqv)))]
}
getmode(m)
```

```
## [1] 34
```

##Use data manipulation techniques and filter rows based on any logical criteria that exist in your dataset

```
filter(avocado, AveragePrice < 0.5)
```

```
##    X       Date AveragePrice Total.Volume      X4046     X4225    X4770
```

```
## 1  0 2015-12-27         0.49   1137707.43  738314.80 286858.37 11642.46
## 2 47 2017-02-05         0.46   2200550.27 1200632.86 531226.65 18324.93
## 3 43 2017-03-05         0.44     64057.04     223.84   4748.88     0.00
## 4 44 2017-02-26         0.49     44024.03     252.79   4472.68     0.00
## 5 43 2017-03-05         0.48     50890.73     717.57   4138.84     0.00
##   Total.Bags Small.Bags Large.Bags XLarge.Bags         type year
## 1  100891.80   70749.02   30142.78        0.00 conventional 2015
## 2  450365.83  113752.17  330583.10     6030.56 conventional 2017
## 3   59084.32     638.68   58445.64        0.00      organic 2017
## 4   39298.56     600.00   38698.56        0.00      organic 2017
## 5   46034.32    1385.06   44649.26        0.00      organic 2017
##             region
## 1    PhoenixTucson
## 2    PhoenixTucson
## 3 CincinnatiDayton
## 4 CincinnatiDayton
## 5          Detroit
```

##Identify the dependent & independent variables and use reshaping techniques and create a new data frame by joining those variablesfrom your dataset.
Create a new dataset with the selected columns

```
bags <- as.data.frame(avocado %>%
    select(Total.Bags, Small.Bags, Large.Bags, XLarge.Bags))
str(bags)
```

```
## 'data.frame':    18249 obs. of  4 variables:
##  $ Total.Bags : num  8697 9506 8145 5811 6184 ...
##  $ Small.Bags : num  8604 9408 8042 5677 5986 ...
##  $ Large.Bags : num  93.2 97.5 103.1 133.8 197.7 ...
##  $ XLarge.Bags: num  0 0 0 0 0 0 0 0 0 0 ...
```

##Remove missing values in your dataset.

```
x = na.omit(avocado)
head(x, 10)
```

```
##    X       Date AveragePrice Total.Volume   X4046     X4225  X4770 Total.Bags
## 1  0 2015-12-27         1.33     64236.62 1036.74  54454.85  48.16    8696.87
## 2  1 2015-12-20         1.35     54876.98  674.28  44638.81  58.33    9505.56
## 3  2 2015-12-13         0.93    118220.22  794.70 109149.67 130.50    8145.35
## 4  3 2015-12-06         1.08     78992.15 1132.00  71976.41  72.58    5811.16
## 5  4 2015-11-29         1.28     51039.60  941.48  43838.39  75.78    6183.95
## 6  5 2015-11-22         1.26     55979.78 1184.27  48067.99  43.61    6683.91
## 7  6 2015-11-15         0.99     83453.76 1368.92  73672.72  93.26    8318.86
## 8  7 2015-11-08         0.98    109428.33  703.75 101815.36  80.00    6829.22
## 9  8 2015-11-01         1.02     99811.42 1022.15  87315.57  85.34   11388.36
## 10 9 2015-10-25         1.07     74338.76  842.40  64757.44 113.00    8625.92
##    Small.Bags Large.Bags XLarge.Bags         type year region
## 1     8603.62      93.25           0 conventional 2015 Albany
## 2     9408.07      97.49           0 conventional 2015 Albany
## 3     8042.21     103.14           0 conventional 2015 Albany
```

```
## 4     5677.40     133.76           0 conventional 2015 Albany
## 5     5986.26     197.69           0 conventional 2015 Albany
## 6     6556.47     127.44           0 conventional 2015 Albany
## 7     8196.81     122.05           0 conventional 2015 Albany
## 8     6266.85     562.37           0 conventional 2015 Albany
## 9    11104.53     283.83           0 conventional 2015 Albany
## 10    8061.47     564.45           0 conventional 2015 Albany
```

##Identify and remove duplicated data in your dataset.

```
y = avocado[!duplicated(avocado), ]
head(y, 10)
```

```
##    X       Date AveragePrice Total.Volume   X4046     X4225  X4770 Total.Bags
## 1  0 2015-12-27         1.33     64236.62 1036.74  54454.85  48.16    8696.87
## 2  1 2015-12-20         1.35     54876.98  674.28  44638.81  58.33    9505.56
## 3  2 2015-12-13         0.93    118220.22  794.70 109149.67 130.50    8145.35
## 4  3 2015-12-06         1.08     78992.15 1132.00  71976.41  72.58    5811.16
## 5  4 2015-11-29         1.28     51039.60  941.48  43838.39  75.78    6183.95
## 6  5 2015-11-22         1.26     55979.78 1184.27  48067.99  43.61    6683.91
## 7  6 2015-11-15         0.99     83453.76 1368.92  73672.72  93.26    8318.86
## 8  7 2015-11-08         0.98    109428.33  703.75 101815.36  80.00    6829.22
## 9  8 2015-11-01         1.02     99811.42 1022.15  87315.57  85.34   11388.36
## 10 9 2015-10-25         1.07     74338.76  842.40  64757.44 113.00    8625.92
##    Small.Bags Large.Bags XLarge.Bags         type year region
## 1     8603.62      93.25           0 conventional 2015 Albany
## 2     9408.07      97.49           0 conventional 2015 Albany
## 3     8042.21     103.14           0 conventional 2015 Albany
## 4     5677.40     133.76           0 conventional 2015 Albany
## 5     5986.26     197.69           0 conventional 2015 Albany
## 6     6556.47     127.44           0 conventional 2015 Albany
## 7     8196.81     122.05           0 conventional 2015 Albany
## 8     6266.85     562.37           0 conventional 2015 Albany
## 9    11104.53     283.83           0 conventional 2015 Albany
## 10    8061.47     564.45           0 conventional 2015 Albany
```

##Reorder multiple rows in descending order

```
z = avocado %>%
    arrange(desc(AveragePrice))
head(z, 10)
```

```
##     X       Date AveragePrice Total.Volume    X4046    X4225  X4770 Total.Bags
## 1   8 2016-10-30         3.25     16700.94  2325.93 11142.85   0.00    3232.16
## 2  37 2017-04-16         3.17      3018.56  1255.55    82.31   0.00    1680.70
## 3   7 2016-11-06         3.12     19043.80  5898.49 10039.34   0.00    3105.97
## 4  42 2017-03-12         3.05      2068.26  1043.83    77.36   0.00     947.07
## 5  18 2017-08-27         3.04     12656.32   419.06  4851.90 145.09    7240.27
## 6  12 2016-10-02         3.03      3714.71   296.71  2699.80   0.00     718.20
## 7  13 2017-10-01         3.00     10741.93   140.46  4331.20 147.11    6123.16
## 8  18 2017-08-27         3.00     19329.49 10517.41  7907.99   0.00     904.09
## 9   6 2016-11-13         2.99     18930.40  6204.65  9341.41   0.00    3384.34
```

```
## 10 13 2017-10-01         2.99      2819.87   174.13   703.73  12.01    1930.00
##    Small.Bags Large.Bags XLarge.Bags    type year            region
## 1     3232.16       0.00           0 organic 2016      SanFrancisco
## 2     1542.22     138.48           0 organic 2017             Tampa
## 3     3079.30      26.67           0 organic 2016      SanFrancisco
## 4      926.67      20.40           0 organic 2017 MiamiFtLauderdale
## 5     6960.97     279.30           0 organic 2017 RaleighGreensboro
## 6      718.20       0.00           0 organic 2016          LasVegas
## 7     5873.99     249.17           0 organic 2017 RaleighGreensboro
## 8      900.76       3.33           0 organic 2017      SanFrancisco
## 9     3337.67      46.67           0 organic 2016      SanFrancisco
## 10    1736.97     193.03           0 organic 2017      Jacksonville
```

##Rename some of the column names in your dataset.

```
colnames(avocado)
```

```
##  [1] "X"            "Date"         "AveragePrice" "Total.Volume" "X4046"
##  [6] "X4225"        "X4770"        "Total.Bags"   "Small.Bags"   "Large.Bags"
## [11] "XLarge.Bags"  "type"         "year"         "region"
```

```
# Change column names to new names
names(avocado)[names(avocado) == "X4046"] <- "variety1"
names(avocado)[names(avocado) == "X4770"] <- "variety2"
names(avocado)[names(avocado) == "X4225"] <- "variety3"

# Show changed column names
colnames(avocado)
```

```
##  [1] "X"            "Date"         "AveragePrice" "Total.Volume" "variety1"
##  [6] "variety3"     "variety2"     "Total.Bags"   "Small.Bags"   "Large.Bags"
## [11] "XLarge.Bags"  "type"         "year"         "region"
```

##Add new variables in your data frame by using a mathematical function (for e.g. –multiply an existing column by 2 and add it as a new variable to your data frame) ##Create new variable by mutliplying an existing column by 2

```
avocado$Doubleyear = avocado$year * 2
```

##Create a training set using random number generator engine. # Initiate random number generator engine

```
set.seed(1234)
```

Select 80% rows from the main dataset as the training set

```
training = avocado %>%
    sample_frac(0.8, replace = FALSE)
head(training, 10)
```

```
##     X       Date AveragePrice Total.Volume   variety1   variety3 variety2
## 1  33 2017-05-14         1.33    117857.60   51068.92   20279.94  1751.27
## 2  14 2017-09-24         1.33   4405343.75 2350943.03  750401.85  6455.71
## 3   8 2017-11-05         1.45   3492828.10  228843.35 1955086.76  5588.32
## 4  31 2017-05-28         1.38   3916908.69 2252109.64  450910.10  5062.46
## 5  17 2015-08-30         1.53      9298.19    2349.53    3534.12     0.00
## 6  50 2015-01-11         0.76   1128693.04  680572.11  348535.22 11900.83
## 7  22 2017-07-30         1.85     14874.38     286.35    6566.23     0.00
## 8  42 2015-03-08         1.52     12545.98    8912.48    1115.21     0.00
## 9  49 2015-01-18         1.39    229216.05    3347.41  160683.80   286.52
## 10 26 2016-06-26         1.38      2515.91      27.16    2001.10     0.00
##    Total.Bags Small.Bags Large.Bags XLarge.Bags         type year
## 1    44757.47   41341.77    3086.53      329.17 conventional 2017
## 2  1297543.16  965684.93  331517.39      340.84 conventional 2017
## 3  1303309.67 1078991.73  224223.95       93.99 conventional 2017
## 4  1208826.49  789659.56  371172.96    47993.97 conventional 2017
## 5     3414.54    1577.97    1836.57        0.00      organic 2015
## 6    87684.88   67857.83   19801.95       25.10 conventional 2015
## 7     7882.30    2278.78    5603.52        0.00      organic 2017
## 8     2518.29    2513.33       4.96        0.00      organic 2015
## 9    64898.32   62671.58    2226.74        0.00 conventional 2015
## 10     487.65     440.00      47.65        0.00      organic 2016
##                 region Doubleyear
## 1           Pittsburgh       4034
## 2         SouthCentral       4034
## 3            Northeast       4034
## 4            Southeast       4034
## 5              Atlanta       4030
## 6        DallasFtWorth       4030
## 7     CincinnatiDayton       4034
## 8        PhoenixTucson       4030
## 9  HartfordSpringfield       4030
## 10         GrandRapids       4032
```

#Print the summary statistics of your dataset.

```
summary(avocado)
```

```
##        X             Date            AveragePrice    Total.Volume
##  Min.   : 0.00   Length:18249       Min.   :0.440   Min.   :      85
##  1st Qu.:10.00   Class :character   1st Qu.:1.100   1st Qu.:   10839
##  Median :24.00   Mode  :character   Median :1.370   Median :  107377
##  Mean   :24.23                      Mean   :1.406   Mean   :  850644
##  3rd Qu.:38.00                      3rd Qu.:1.660   3rd Qu.:  432962
##  Max.   :52.00                      Max.   :3.250   Max.   :62505647
##     variety1           variety3           variety2         Total.Bags
##  Min.   :       0   Min.   :       0   Min.   :      0   Min.   :       0
##  1st Qu.:     854   1st Qu.:    3009   1st Qu.:      0   1st Qu.:    5089
##  Median :    8645   Median :   29061   Median :    185   Median :   39744
##  Mean   :  293008   Mean   :  295155   Mean   :  22840   Mean   :  239639
##  3rd Qu.:  111020   3rd Qu.:  150207   3rd Qu.:   6243   3rd Qu.:  110783
##  Max.   :22743616   Max.   :20470573   Max.   :2546439   Max.   :19373134
##    Small.Bags         Large.Bags       XLarge.Bags           type
```

```
##  Min.   :       0   Min.   :      0   Min.   :     0.0   Length:18249
##  1st Qu.:    2849   1st Qu.:    127   1st Qu.:     0.0   Class :character
##  Median :   26363   Median :   2648   Median :     0.0   Mode  :character
##  Mean   :  182195   Mean   :  54338   Mean   :  3106.4
##  3rd Qu.:   83338   3rd Qu.:  22029   3rd Qu.:   132.5
##  Max.   :13384587   Max.   :5719097   Max.   :551693.7
##       year         region            Doubleyear
##  Min.   :2015   Length:18249       Min.   :4030
##  1st Qu.:2015   Class :character   1st Qu.:4030
##  Median :2016   Mode  :character   Median :4032
##  Mean   :2016                      Mean   :4032
##  3rd Qu.:2017                      3rd Qu.:4034
##  Max.   :2018                      Max.   :4036
```

##Use any of the numerical variables from the dataset and perform the following statistical functions. Mean

```
mean(avocado$Large.Bags)
```

```
## [1] 54338.09
```

##Median

```
median(avocado$Total.Bags)
```

```
## [1] 39743.83
```

##Mode

```
v <- c(avocado$AveragePrice)
# Calculate the mode using the user defined function
result <- getmode(v)
print(result)
```

```
## [1] 1.15
```

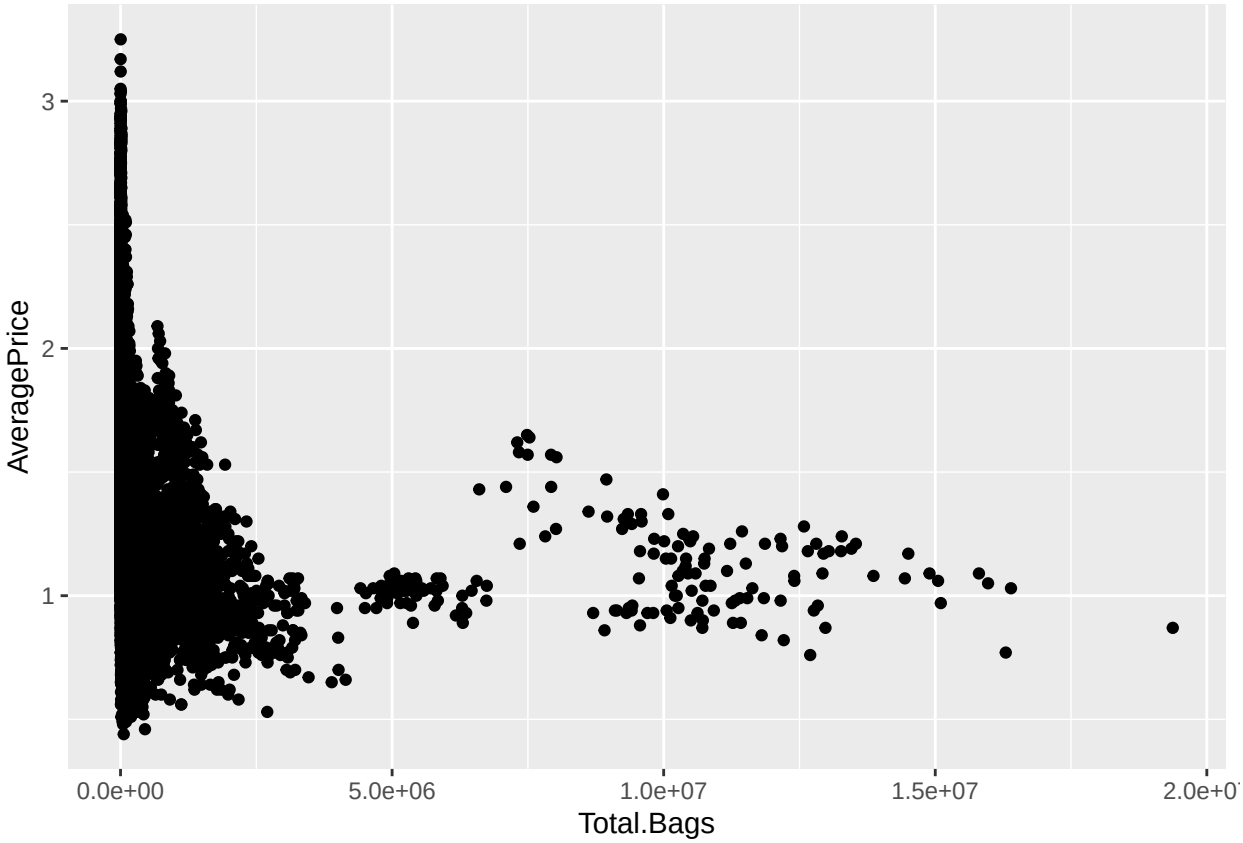
##Range

```
range(avocado$Total.Bags)
```

```
## [1]        0 19373134
```

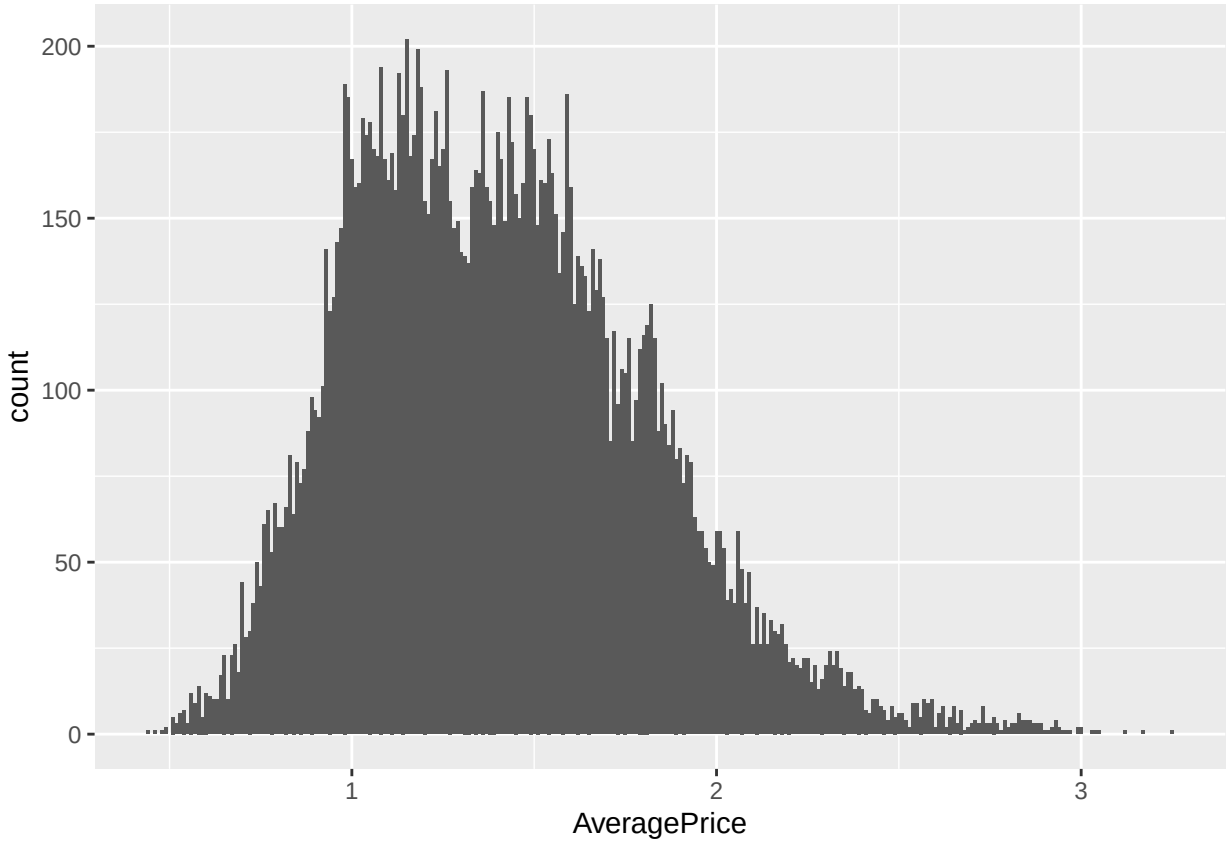
##Plot a scatter plot for any 2 variables in your dataset.

```
ggplot(data = avocado, aes(x = Total.Bags, y = AveragePrice)) + geom_point()
```



##Plot a bar plot for any 1 variables in your dataset

```
ggplot(data = avocado, aes(x = AveragePrice)) + geom_bar()
```



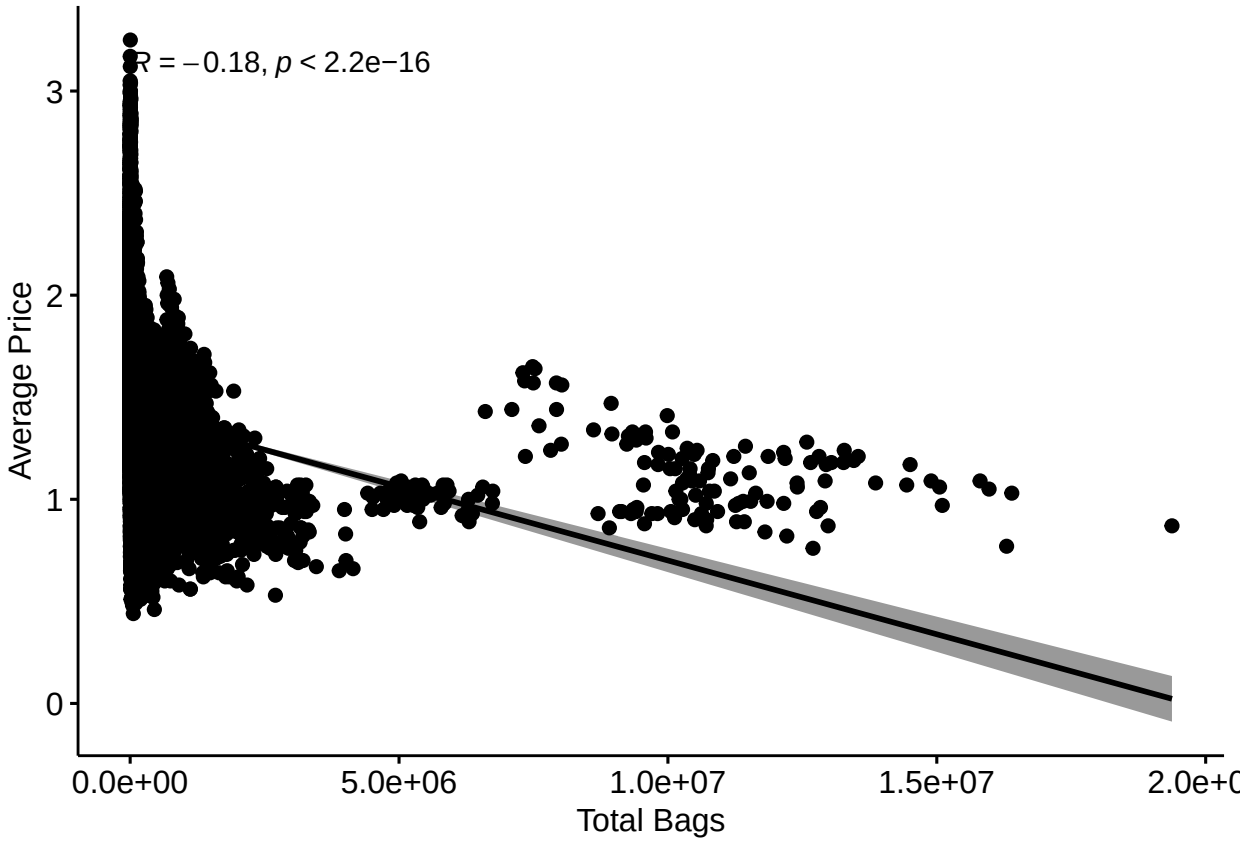
##Find the correlation between any 2 variables by applying least square linear regression model.

```
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.0.5
```

```
ggscatter(avocado, x = "Total.Bags", y = "AveragePrice", add = "reg.line", conf.int = TRUE,
    cor.coef = TRUE, cor.method = "pearson", xlab = "Total Bags", ylab = "Average Price")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



##Provide a conclusion of your analysis if any in the .RMD file. # there are more avocados in the range of $1-2,