Assignment 1

2025-03-28

```r
avocado_csv <- read.csv("avocado.csv")
# Structure
str(avocado_csv)
```

```
## '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" ...
```

```r
# List variables
names(avocado_csv)
```

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

```r
# Top 15 rows
head(avocado_csv, 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
```

https://github.com/michelle912/4066_assignment1

```
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
```

```r
# User defined function
get_highest_price_date <- function() {
    print(avocado_csv[which.max(avocado_csv$AveragePrice), ]$Date)
}
get_highest_price_date()
```

```
## [1] "2016-10-30"
```

```r
# Filter rows
subset(avocado_csv, AveragePrice < 0.5)$Date
```

```
## [1] "2015-12-27" "2017-02-05" "2017-03-05" "2017-02-26" "2017-03-05"
```

```r
# New dataframe
new_df <- cbind(avocado_csv$AveragePrice, avocado_csv$Total.Volume)

# Remove missing values
avocado_csv <- na.omit(avocado_csv)

# Remove duplicates
avocado_csv <- unique.data.frame(avocado_csv)

# Reorder in desc order
avocado_desc <- avocado_csv[order(-avocado_csv$AveragePrice),]

# Rename columns
avocado_rename <- avocado_csv %>% rename(plu4225_sold_count = "X4225",
```

https://github.com/michelle912/4066_assignment1

```
plu4770_sold_count = "X4770")


# Add new variables
avocado_csv <- transform(avocado_csv, average_price_cad =
avocado_csv$AveragePrice * 1.43)

# Training set
set.seed(231)
training_index <- sample(1:nrow(avocado_csv), 10, replace = FALSE)
avocado_csv[training_index, ]
```

```
##        X       Date AveragePrice Total.Volume      X4046      X4225
X4770
## 682    5 2015-11-22         1.09    270529.34   95526.76   55945.23
63404.02
## 8233  19 2017-08-20         1.42     60657.76     796.92   36561.00
19.35
## 11141 38 2015-04-05         1.54      4893.26     235.10    2349.78
523.63
## 6786   3 2017-12-10         1.07   2404221.58 1024592.52  396712.41
62000.95
## 12307  9 2016-10-23         1.64      9342.71     147.98    4509.02
811.96
## 43    42 2015-03-08         1.07     40507.36     795.68   30370.64
159.05
## 5640  23 2017-07-23         1.49     84416.61    1905.52   72533.28
940.58
## 1645  32 2015-05-17         1.23    243265.56  168886.63   19797.32
173.63
## 4319   2 2016-12-11         1.32   3381321.05  102683.58 2390081.54
19968.52
## 12454  0 2016-12-25         1.75      4561.12     543.08    2287.13
0.00
##       Total.Bags Small.Bags Large.Bags XLarge.Bags         type year
## 682     55653.33   47388.76    1257.23     7007.34 conventional 2015
## 8233    23280.49   21641.10     392.72     1246.67 conventional 2017
## 11141    1784.75    1330.00     454.75        0.00      organic 2015
## 6786   920915.70  868361.67   16284.72    36269.31 conventional 2017
## 12307    3873.75    3400.36     473.39        0.00      organic 2016
## 43       9181.99    8827.55     354.44        0.00 conventional 2015
## 5640     9037.23    3250.30    4896.93      890.00 conventional 2017
## 1645    54407.98   38066.02   16341.96        0.00 conventional 2015
## 4319   868587.41  733966.40  132518.00     2103.01 conventional 2016
## 12454    1730.91    1588.03     142.88        0.00      organic 2016
##                region average_price_cad
## 682           Detroit            1.5587
## 8233         Syracuse            2.0306
## 11141 RichmondNorfolk            2.2022
## 6786       LosAngeles            1.5301
```

https://github.com/michelle912/4066_assignment1

```
## 12307       Charlotte            2.3452
## 43             Albany            1.5301
## 5640           Albany            2.1307
## 1645          Orlando            1.7589
## 4319        Northeast            1.8876
## 12454        Columbus            2.5025
```

```
# Summary stats
summary(avocado_csv)
```

```
##        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
##      X4046              X4225              X4770          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          average_price_cad
##  Min.   :2015   Length:18249       Min.   :0.6292
##  1st Qu.:2015   Class :character   1st Qu.:1.5730
##  Median :2016   Mode  :character   Median :1.9591
##  Mean   :2016                      Mean   :2.0105
##  3rd Qu.:2017                      3rd Qu.:2.3738
##  Max.   :2018                      Max.   :4.6475
```

```
# Stat function
getMode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
mean(avocado_csv$AveragePrice, na.rm = TRUE)
```

```
## [1] 1.405978
```

```
median(avocado_csv$AveragePrice, na.rm = TRUE)
```

```
## [1] 1.37
```

https://github.com/michelle912/4066_assignment1

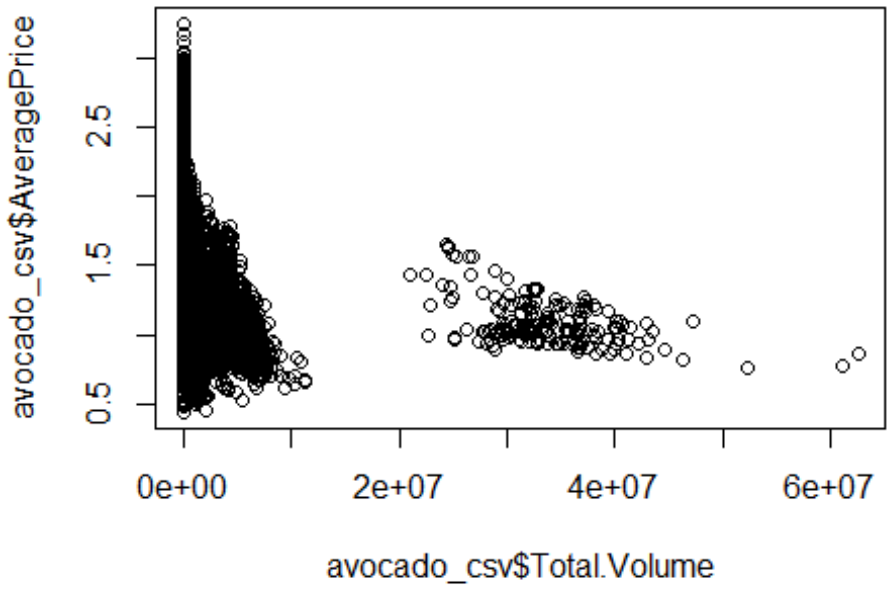
```r
getMode(avocado_csv$region)
```

```
## [1] "Albany"
```

```r
range(avocado_csv$AveragePrice, na.rm = TRUE)
```

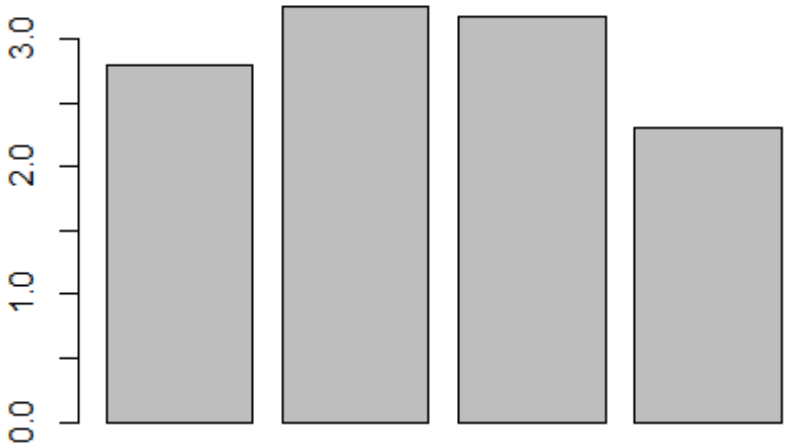
```
## [1] 0.44 3.25
```

```r
# Scatter plot
plot(avocado_csv$Total.Volume, avocado_csv$AveragePrice)
```




```r
# Bar plot
max_by_year <- group_by(avocado_csv, year) %>% summarise(AveragePrice =
max(AveragePrice))
barplot(max_by_year$AveragePrice, max_by_year$year)
```

https://github.com/michelle912/4066_assignment1




```r
# Pearson correlation
cor(avocado_csv$AveragePrice, avocado_csv$Total.Volume, method = "pearson")
```

```
## [1] -0.1927524
```

https://github.com/michelle912/4066_assignment1