Homework3

Priya

10/22/2023

Question1

In your own words, describe the data and how you obtained the data.

Answer

I have downloaded " Fruit and Vegetable Prices" dataset from the Data.gov site. It is the United States government's open data website. It provides access to datasets published by agencies across the federal government.

Direct link to download this dataset is given below: https://catalog.data.gov/dataset/fruit-and-vegetable-prices

Fruit and Vegetable Prices dataset is updated on: February 24, 2021

Objective of this dataset is how much do fruits and vegetables cost?

Question2

How did you address cleaning the data? What did you do for missing data (if any)? What rows and columns did you delete? Provide the R code, if any. If you did the cleaning in Excel, describe your step-by-step approach.

Answer:

This is the clean dataset. I think it is mention somewhere on the website that it contains 153 observations but after download I checked it has 62 observations only. It did not have any missing data.

Dataset contain 8 columns:

Fruit: name of the different fruits (character) Form: form of the fruit such as fresh or frozen (categorical variable) RetailPrice : price in integer RetailPriceUnit: unit (character/categorical variable) Yield: integer CupEquivalentSize: integer CupEquivalentUnit : categorical variable CupEquivalentPrice: integer

```
fruit <- read.csv("Fruit Prices 2020.csv")
head(fruit, 5)
```

```
##                       Fruit   Form RetailPrice RetailPriceUnit Yield
## 1                    Apples  Fresh      1.5193       per pound  0.90
```

```
## 2         Apples, applesauce Canned      1.0660       per pound  1.00
## 3     Apples, ready-to-drink  Juice      0.7804        per pint  1.00
## 4 Apples, frozen concentrate  Juice      0.5853        per pint  1.00
## 5                   Apricots  Fresh      2.9665       per pound  0.93
##   CupEquivalentSize CupEquivalentUnit CupEquivalentPrice
## 1            0.2425            pounds             0.4094
## 2            0.5401            pounds             0.5758
## 3            8.0000      fluid ounces             0.3902
## 4            8.0000      fluid ounces             0.2926
## 5            0.3638            pounds             1.1603
```

Question3:

Use summary() to generate summary statistics of your three numeric variables (e.g. var1, var2, var3) by one of your categorical variables (e.g. summary statistics for "region 1" versus "region 2" versus "region 3"). Use boxplot() to visually show the summaries. Provide the R code Followed by graphs Followed by a brief description of what you see (in complete sentences).

Answer

```
#summary(fruit)
# convert to factor
fruit$Form <- as.factor(fruit$Form)
summary(fruit)
```

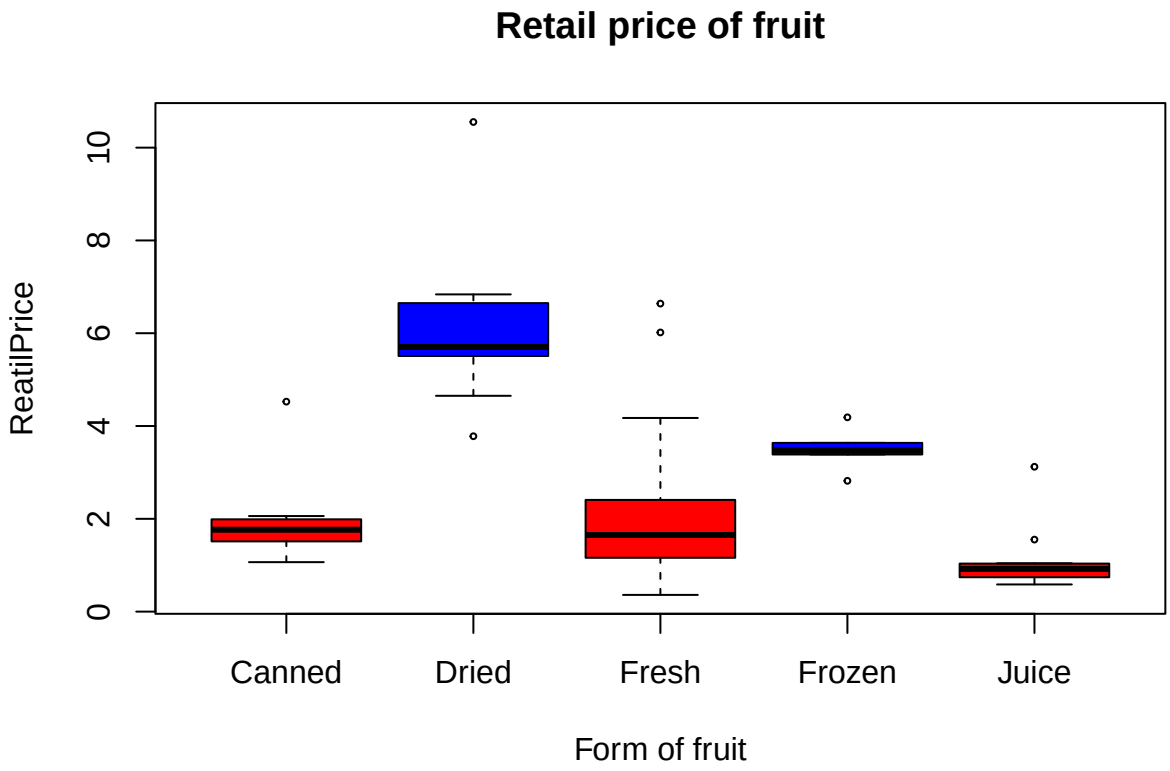
```
##     Fruit               Form     RetailPrice      RetailPriceUnit
##  Length:62          Canned:12   Min.   : 0.3604   Length:62
##  Class :character   Dried : 9   1st Qu.: 1.1559   Class :character
##  Mode  :character   Fresh :24   Median : 1.8684   Mode  :character
##                     Frozen: 6   Mean   : 2.6160
##                     Juice :11   3rd Qu.: 3.5256
##                                 Max.   :10.5527
##      Yield        CupEquivalentSize CupEquivalentUnit  CupEquivalentPrice
##  Min.   :0.4600   Min.   :0.1232    Length:62          Min.   :0.2292
##  1st Qu.:0.7225   1st Qu.:0.3225    Class :character   1st Qu.:0.5793
##  Median :0.9800   Median :0.3638    Mode  :character   Median :0.8952
##  Mean   :0.8761   Mean   :1.7050                       Mean   :0.9197
##  3rd Qu.:1.0000   3rd Qu.:0.5401                       3rd Qu.:1.1505
##  Max.   :1.0000   Max.   :8.0000                       Max.   :3.0700
```

After checking the summary of the variables Retail price and CupEquivalentSize shows big variation in the data as max value is very high as compare to the other stats for these both variable.

boxplot for fruits retail price

```
boxplot(fruit$RetailPrice~fruit$Form,
    main = "Retail price of fruit",
    cex = 0.4,
```

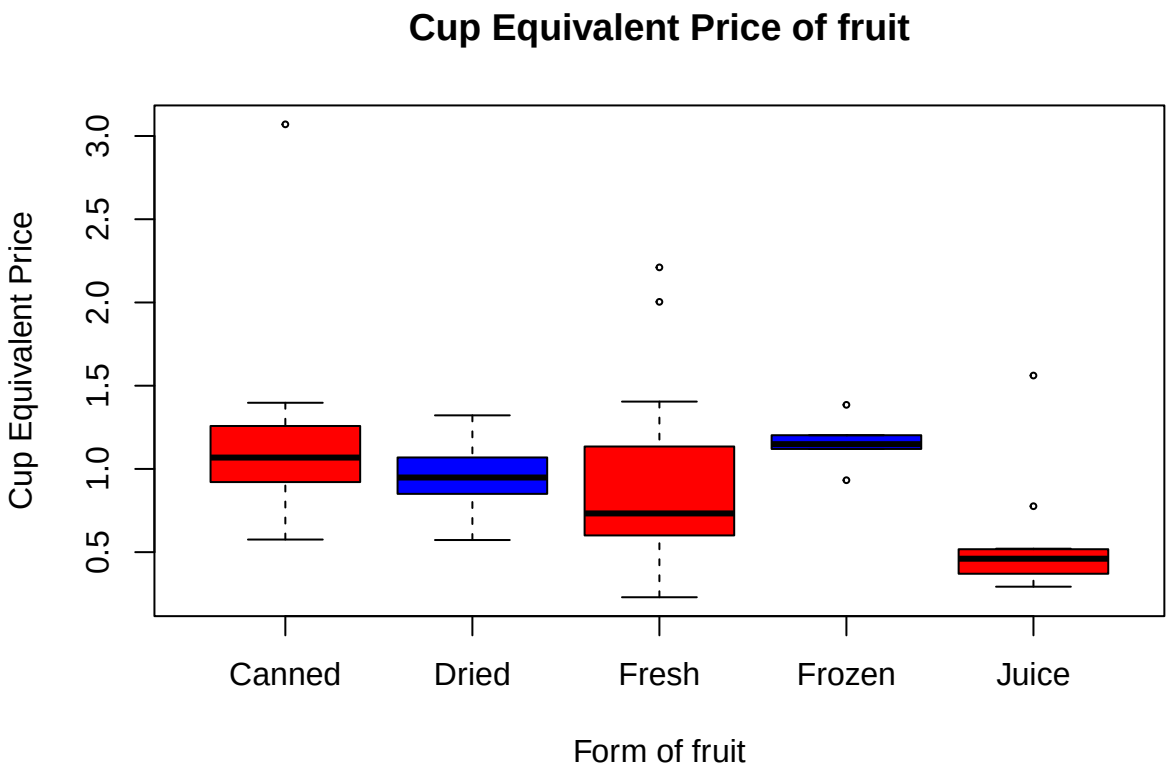
```
    col = c("red","blue"),
    xlab = "Form of fruit", ylab = "ReatilPrice")
```



The above boxplot shows the prices of the Form of the fruit varies. Retail price of the dried fruit is highest. It indicates dried fruits are most expensive in all these form of the fruit.

boxplot for fruits Cup equivalent price

```
boxplot(fruit$CupEquivalentPrice~fruit$Form,
    main = "Cup Equivalent Price of fruit",
    cex = 0.4,
    col = c("red","blue"),
    xlab = "Form of fruit", ylab = "Cup Equivalent Price")
```



When we checked the cup equivalent price for the form of the fruits, it does not show big variation in the price across these form of the fruit. Juice showed the cheapest.

Summary and boxplot for fruits cup equivalent size

```
# summary of Canned fruit form
summary(fruit$CupEquivalentSize[fruit$Form == "Canned"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4409  0.4409  0.4905  0.4905  0.5401  0.5401
```

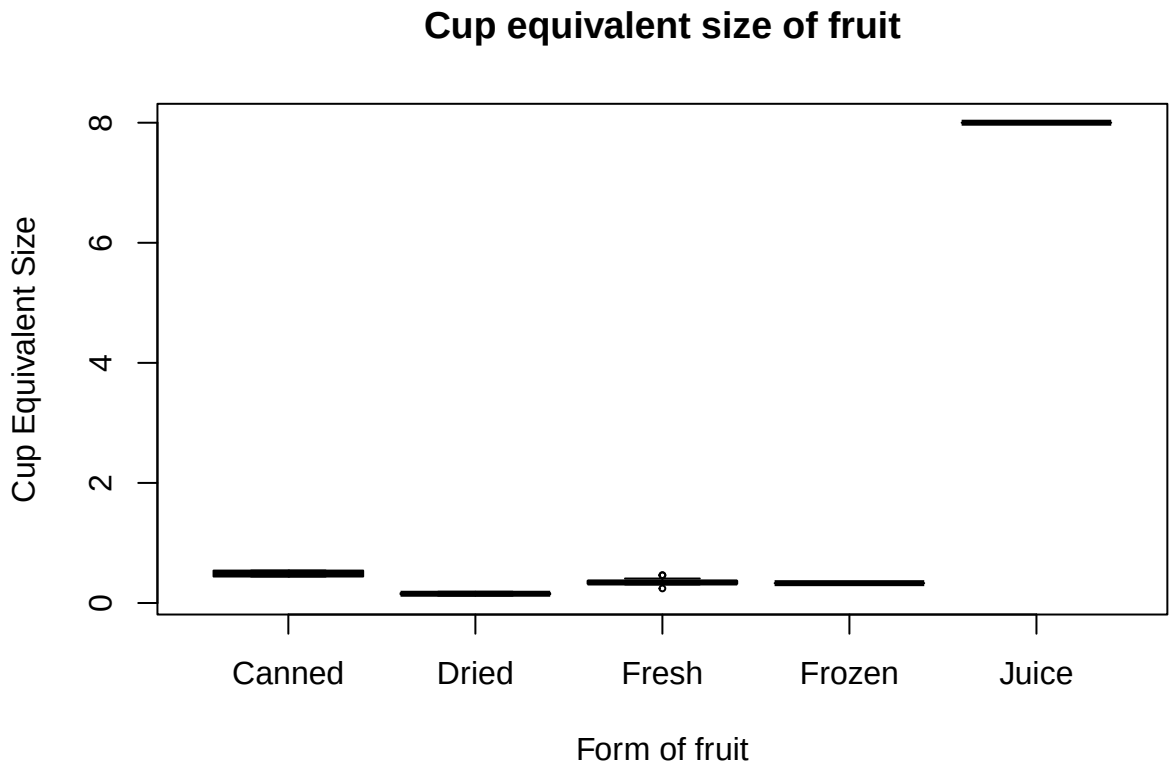
```
# summary of Fresh fruit form
summary(fruit$CupEquivalentSize[fruit$Form == "Fresh"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2425  0.3197  0.3417  0.3523  0.3665  0.4630
```

```
# summary of Juice fruit form
summary(fruit$CupEquivalentSize[fruit$Form == "Juice"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       8       8       8       8       8       8
```

```r
# boxplot
boxplot(fruit$CupEquivalentSize~fruit$Form,
    main = "Cup equivalent size of fruit",
    cex = 0.4,
    col = c("red","blue"),
    xlab = "Form of fruit", ylab = "Cup Equivalent Size")
```



These boxplots represent the cup equivalent size and we see line instead of boxplot here. It is beacuse if we see the stats lower quartile, median, upper quartile all are almost same for these form. I have shown the summary of Canned, fresh and Juice form of the fruit. For example: In the Juice form lower quartile, median and upper quartile, everything is 8, so we will see a line instead of boxplot. It is because this is unit of measurement.
As Juice cup eqivalent size is 8, which is much higher than other. It is because juice is measured in fluid ounces here and other unit are in the pounds. That means it is not appropriate to visualize it on the different unit for cup equivalent size. We probably need to convert fluid ounces in pound (pound = ounces/16) to see everything at the same unit.

Boxplot: After converting the unit in pound for cup equivalent size

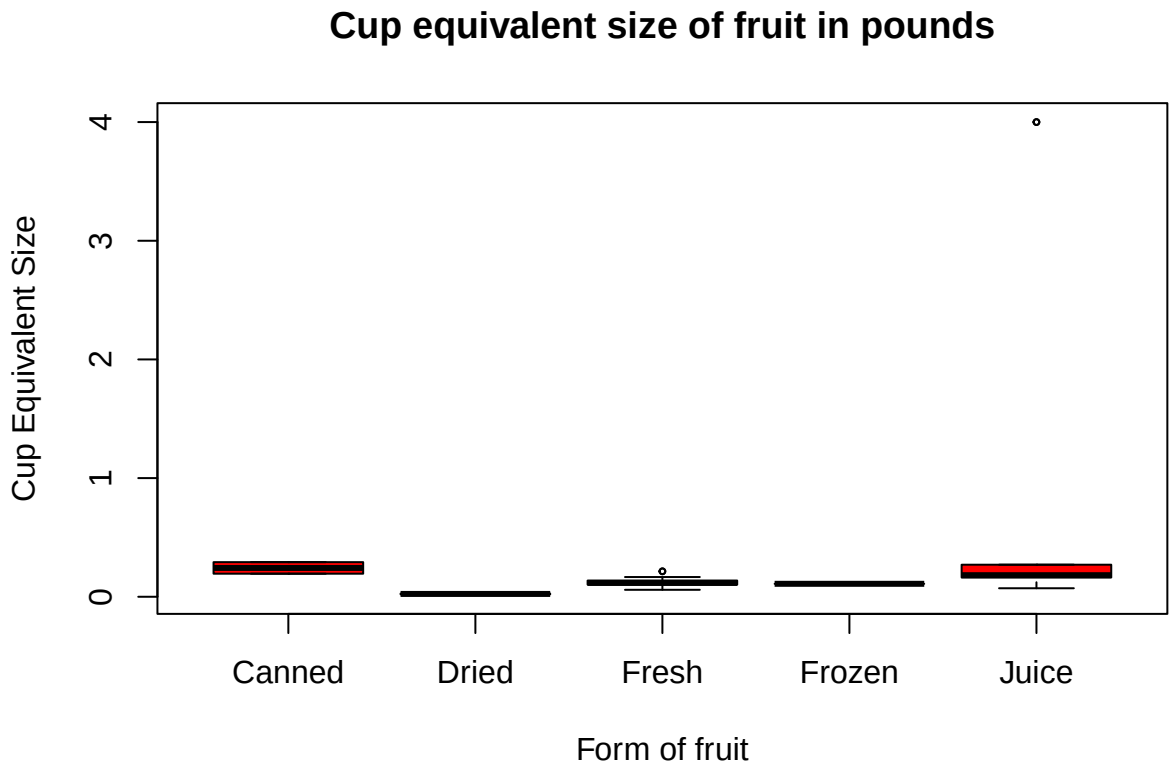
```r
newCupEquivalentSize <- c()
newCupEquivalentSize <- fruit$CupEquivalentSize
newCupEquivalentSize[fruit$Form == "Juice"] <- fruit$CupEquivalentSize/16
```

```
## Warning in newCupEquivalentSize[fruit$Form == "Juice"] <-
```

```
## fruit$CupEquivalentSize/16: number of items to replace is not a multiple of
## replacement length
```

```r
# add into dataframe
fruit <- cbind(fruit, newCupEquivalentSize)

boxplot(fruit$newCupEquivalentSize*fruit$CupEquivalentSize ~ fruit$Form,
    main = "Cup equivalent size of fruit in pounds",
    cex = 0.4,
    col = c("red","blue"),
    xlab = "Form of fruit", ylab = "Cup Equivalent Size")
```



After converting the ounces in pounds we do see all form of fruit have similar cup equivalent size. Juice has one outlier.

Question4:

Generate a three histograms with an overlaid density plot, one for each of your three numeric variables, separated by one of your categorical variables.

Provide the R code Followed by graphs Followed by a brief description of what you see (in complete sentences). You can compare the histograms/density plots of the categories.

Answer

Histogram of retail price of dried fruit with overlaid density plot

```
dens_dried <- density(fruit$RetailPrice[fruit$Form == "Dried"], bw = 0.5)
#plot(dens_dried)

hist(fruit$RetailPrice[fruit$Form == "Dried"],
     freq = F,
     cex.axis = 0.8,
     col=rgb(1,0,0,0.25),
     xlim= c(2,12),
     ylim = c(0,0.4),
     main = "Dried fruit retail price",
     xlab = "Dried fruit retail price",
     breaks = seq(2,12,1))

 lines(dens_dried, lwd = 2, col = "red", lty = 2)
```

Histogram and density plot represent that most of the retail price for dried fruit around 6. There are some dried fruit which are even more expensive and can be found around 10-11 bin.

Histogram of retail price of fresh fruit with overlaid density plot

```
dens_fresh <- density(fruit$RetailPrice[fruit$Form == "Fresh"], bw = 0.5)
#plot(dens_fresh)

hist(fruit$RetailPrice[fruit$Form == "Fresh"],
     freq = F,
     cex.axis = 0.8,
     col=rgb(0,1,0,0.25),
     xlim= c(0,8),
     ylim = c(0,0.5),
     main = "Fresh fruit retail price",
     xlab = "Fresh fruit retail price",
     breaks = seq(0,10,1))

 lines(dens_fresh, lwd = 2, col = "green", lty = 2)
```

Fresh fruit are cheaper than dried fruit and start from 0 to 4. Most of then can be purchased from $1 to $2 per pound. There are some fruit which are expensive those are blackberries.

Question 5

Create a bar plot showing the count of one of your categorical variables.

Provide the R code Followed by graphs Followed by a brief description of any patterns that you might see.

Answer

```
library(stringr)
```

```
##
## Attaching package: 'stringr'
```

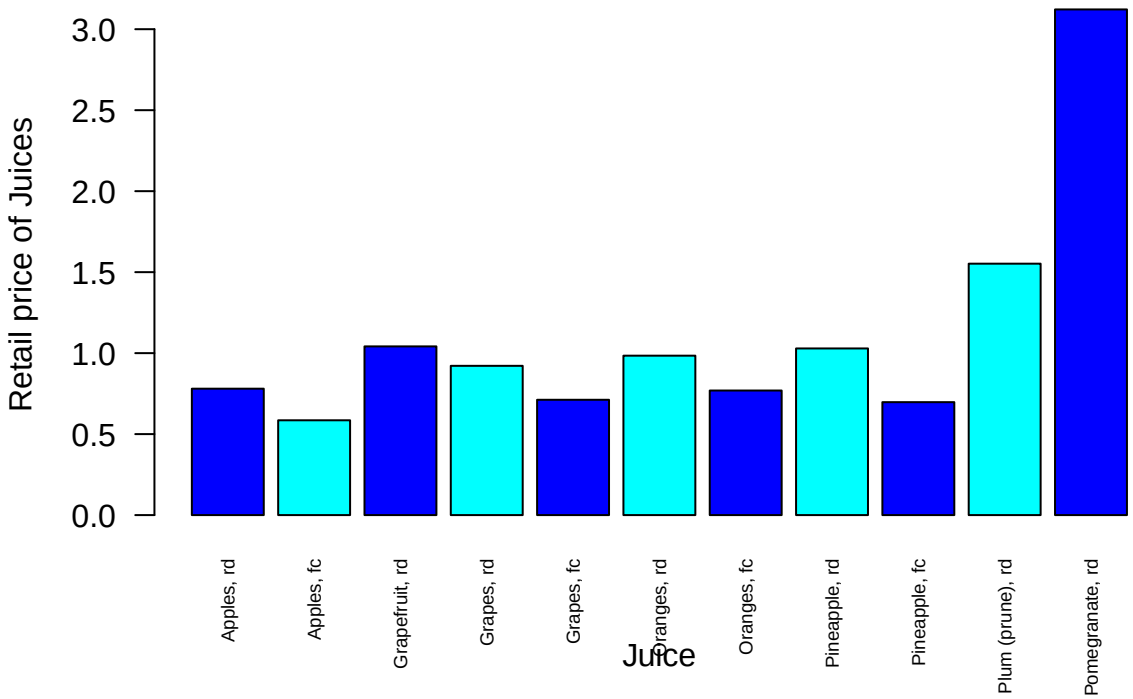
```
## The following object is masked _by_ '.GlobalEnv':
##
##     fruit
```

```
frozen_fruit <- fruit$RetailPrice[fruit$Form == "Juice"]
frozen_fruit_name <- fruit$Fruit[fruit$Form == "Juice"]
frozen_fruit
```

```
##  [1] 0.7804 0.5853 1.0415 0.9215 0.7119 0.9842 0.7690 1.0288 0.6973 1.5522
## [11] 3.1220
```

```
frozen_fruit_name <- str_replace(frozen_fruit_name, "ready-to-drink","rd")
frozen_fruit_name <- str_replace(frozen_fruit_name, "frozen concentrate","fc")

barplot(frozen_fruit,
        ylab = "Retail price of Juices",
        xlab= "Juice",
        names.arg = frozen_fruit_name,
        cex.names=0.60,
        col = c("blue","cyan"),
 las=2)
```



The barplot shows the retail prices of variety of Juice flavors. Here rd means Ready to drink and fc means frozen concentrate. Pomegranate read to drink shows relatively expensive than other juices.

Question 6:

Create a bar plot showing the by-group (category) average of one of your numeric variables:

Provide the R code Followed by graphs Followed by a brief description of any patterns that you might see.

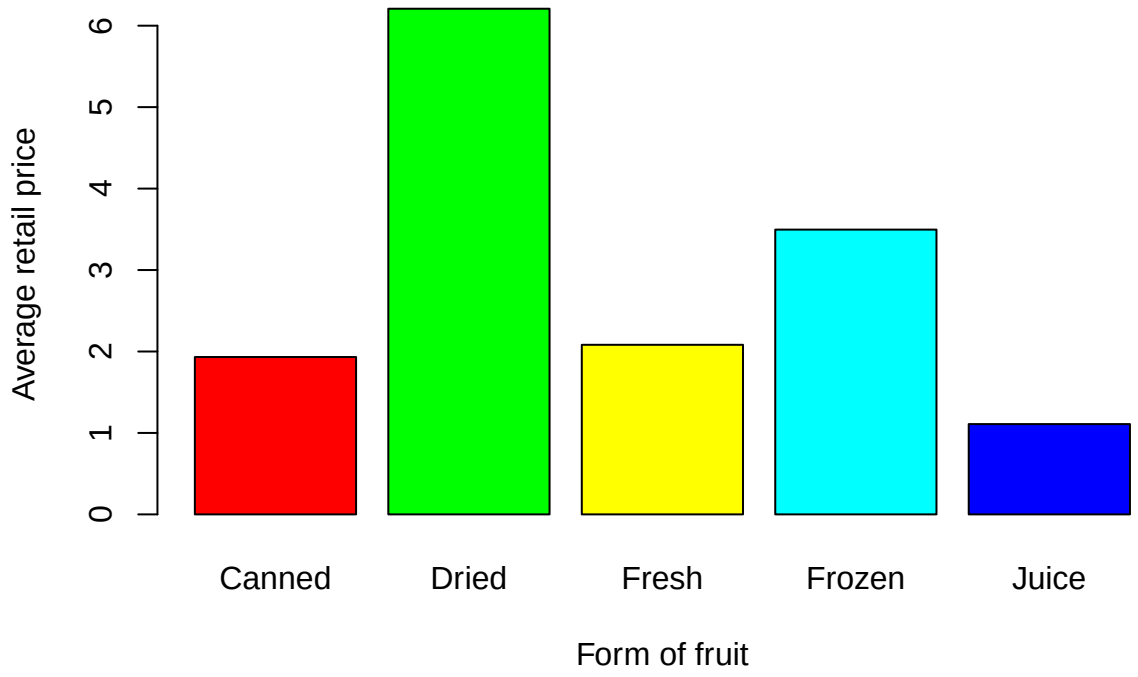
Answer

Barplot for average retail price for all form of the fruit

```
avg_retail_price <- aggregate(RetailPrice~Form, data = fruit, FUN = mean)
avg_retail_price
```

```
##     Form RetailPrice
## 1 Canned    1.931942
## 2  Dried    6.208178
## 3  Fresh    2.081929
## 4 Frozen    3.496300
## 5  Juice    1.108555
```

```
barplot(avg_retail_price$RetailPrice,
        names.arg = avg_retail_price$Form,
        xlab = "Form of fruit",
        ylab = "Average retail price",
        col= c("red","green","yellow","cyan","blue"))
```



Barplot explains that average value of dried fruit are highest and follow by frozen fruits. Juices are the cheapest but I will conclude that fresh fruit are not very expensive and good to health.

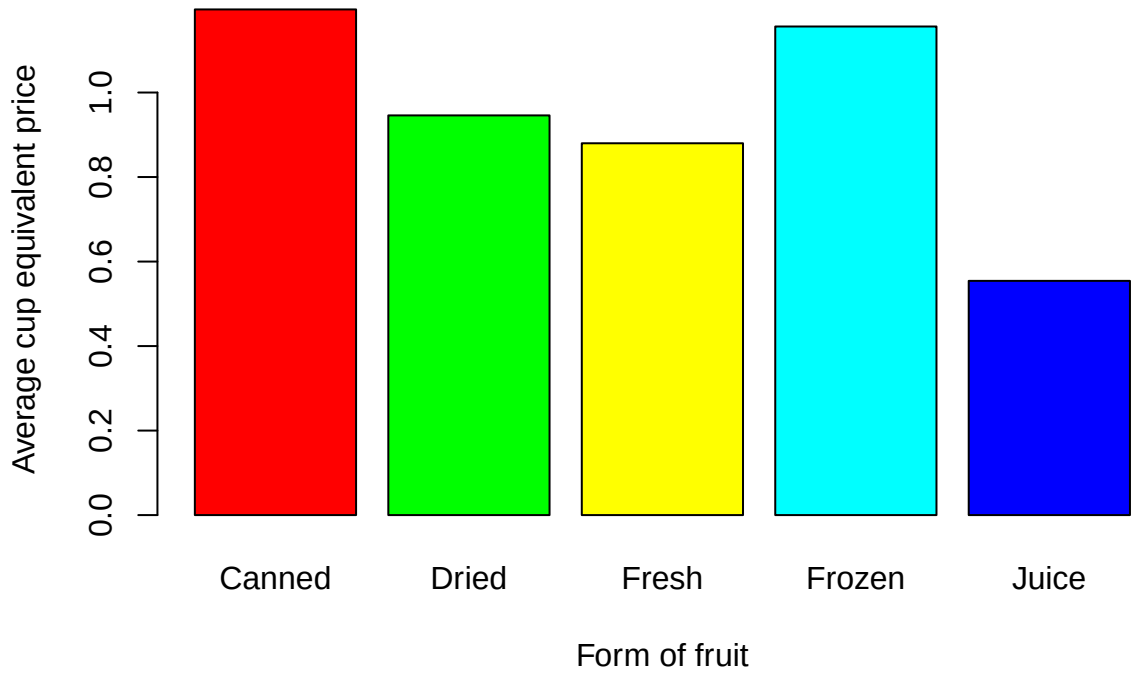
Barplot for average cub equivalent price for all form of the fruit

```
avg_cup_price <- aggregate(CupEquivalentPrice ~ Form, data = fruit, FUN = mean)
avg_cup_price
```

```
##     Form CupEquivalentPrice
## 1 Canned          1.1966417
## 2  Dried          0.9458111
## 3  Fresh          0.8799000
## 4 Frozen          1.1562333
## 5  Juice          0.5542636
```

```
barplot(avg_cup_price$CupEquivalentPrice,
        names.arg = avg_cup_price$Form,
```

```
        xlab = "Form of fruit",
        ylab = "Average cup equivalent price ",
        col= c("red","green","yellow","cyan","blue"))
```



I have also plotted barplot for the average cup equivalent price. This is based on the cup size but this one is not as clear as the retail price.

Question 7:

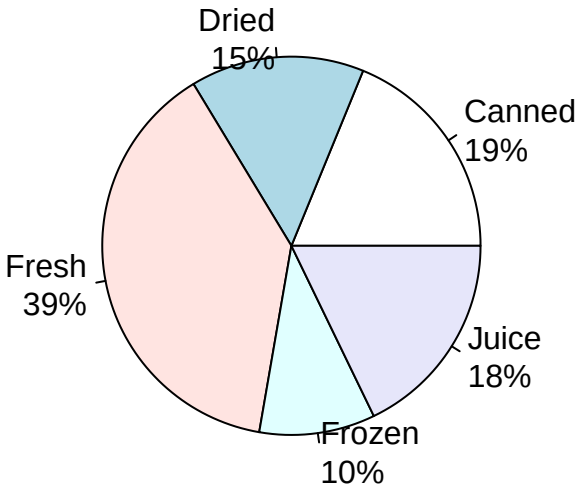
Create an interesting pie chart.

Provide the R code Followed by graphs Followed by a brief description of any patterns that you might see.

Answer

```
fruit_form_table <- table(fruit$Form)
fruit_form_table <- round(prop.table(fruit_form_table), 2)
fruit_form_lab <- paste(names(fruit_form_table), "\n",fruit_form_table *100,  "%", sep= "" )

pie(fruit_form_table,
    label = fruit_form_lab,
    main = "Number of different form of fruit",
    radius = 0.8)
```



This pie chart is showing percentage of the different form of the fruit available in the datasets. Pie charts inform us that variety of fresh fruits are available which is 39%. Frozen fruit has a least variety in the market or datasets.