Week10 Assignment

Mohit Mehndiratta

10/14/2021

# DATA PROCESSING AND VISUALISATION

## Question 1

### 1. Import the abalone dataset and explore.

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library("ggplot2")

## Warning: package 'ggplot2' was built under R version 4.1.1

library(psych)

## Warning: package 'psych' was built under R version 4.1.1

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

abalone <- read.csv("../datasets/abalone.csv", stringsAsFactors = TRUE)  
head(abalone)

## Sex Length Diameter Height Whole.weight Shucked.weight Viscera.weight  
## 1 M 0.455 0.365 0.095 0.5140 0.2245 0.1010  
## 2 M 0.350 0.265 0.090 0.2255 0.0995 0.0485  
## 3 F 0.530 0.420 0.135 0.6770 0.2565 0.1415  
## 4 M 0.440 0.365 0.125 0.5160 0.2155 0.1140  
## 5 I 0.330 0.255 0.080 0.2050 0.0895 0.0395  
## 6 I 0.425 0.300 0.095 0.3515 0.1410 0.0775  
## Shell.weight Rings  
## 1 0.150 15  
## 2 0.070 7  
## 3 0.210 9  
## 4 0.155 10  
## 5 0.055 7  
## 6 0.120 8

dim(abalone)

## [1] 4177 9

summary(abalone)

## Sex Length Diameter Height Whole.weight   
## F:1307 Min. :0.075 Min. :0.0550 Min. :0.0000 Min. :0.0020   
## I:1342 1st Qu.:0.450 1st Qu.:0.3500 1st Qu.:0.1150 1st Qu.:0.4415   
## M:1528 Median :0.545 Median :0.4250 Median :0.1400 Median :0.7995   
## Mean :0.524 Mean :0.4079 Mean :0.1395 Mean :0.8287   
## 3rd Qu.:0.615 3rd Qu.:0.4800 3rd Qu.:0.1650 3rd Qu.:1.1530   
## Max. :0.815 Max. :0.6500 Max. :1.1300 Max. :2.8255   
## Shucked.weight Viscera.weight Shell.weight Rings   
## Min. :0.0010 Min. :0.0005 Min. :0.0015 Min. : 1.000   
## 1st Qu.:0.1860 1st Qu.:0.0935 1st Qu.:0.1300 1st Qu.: 8.000   
## Median :0.3360 Median :0.1710 Median :0.2340 Median : 9.000   
## Mean :0.3594 Mean :0.1806 Mean :0.2388 Mean : 9.934   
## 3rd Qu.:0.5020 3rd Qu.:0.2530 3rd Qu.:0.3290 3rd Qu.:11.000   
## Max. :1.4880 Max. :0.7600 Max. :1.0050 Max. :29.000

str(abalone)

## 'data.frame': 4177 obs. of 9 variables:  
## $ Sex : Factor w/ 3 levels "F","I","M": 3 3 1 3 2 2 1 1 3 1 ...  
## $ Length : num 0.455 0.35 0.53 0.44 0.33 0.425 0.53 0.545 0.475 0.55 ...  
## $ Diameter : num 0.365 0.265 0.42 0.365 0.255 0.3 0.415 0.425 0.37 0.44 ...  
## $ Height : num 0.095 0.09 0.135 0.125 0.08 0.095 0.15 0.125 0.125 0.15 ...  
## $ Whole.weight : num 0.514 0.226 0.677 0.516 0.205 ...  
## $ Shucked.weight: num 0.2245 0.0995 0.2565 0.2155 0.0895 ...  
## $ Viscera.weight: num 0.101 0.0485 0.1415 0.114 0.0395 ...  
## $ Shell.weight : num 0.15 0.07 0.21 0.155 0.055 0.12 0.33 0.26 0.165 0.32 ...  
## $ Rings : int 15 7 9 10 7 8 20 16 9 19 ...

The abalone data set has 4177 observations of 9 variables. Each variable is numeric in nature except the variable “Sex” which is a factor variable.

### 2. Subset the dataset by selecting the variables Length, Diameter and Height.

new\_data <- abalone %>% select(Length, Diameter, Height)  
head(new\_data)

## Length Diameter Height  
## 1 0.455 0.365 0.095  
## 2 0.350 0.265 0.090  
## 3 0.530 0.420 0.135  
## 4 0.440 0.365 0.125  
## 5 0.330 0.255 0.080  
## 6 0.425 0.300 0.095

### 3. Subset the dataset by selecting observations of female abalone.

female <- abalone %>% filter(Sex == "F")  
head(female)

## Sex Length Diameter Height Whole.weight Shucked.weight Viscera.weight  
## 1 F 0.530 0.420 0.135 0.6770 0.2565 0.1415  
## 2 F 0.530 0.415 0.150 0.7775 0.2370 0.1415  
## 3 F 0.545 0.425 0.125 0.7680 0.2940 0.1495  
## 4 F 0.550 0.440 0.150 0.8945 0.3145 0.1510  
## 5 F 0.525 0.380 0.140 0.6065 0.1940 0.1475  
## 6 F 0.535 0.405 0.145 0.6845 0.2725 0.1710  
## Shell.weight Rings  
## 1 0.210 9  
## 2 0.330 20  
## 3 0.260 16  
## 4 0.320 19  
## 5 0.210 14  
## 6 0.205 10

### 4. Subset the dataset by selecting the weight variables of Male abalone.

male\_weight <- abalone %>% filter(Sex == "M") %>% select(Sex, Whole.weight,  
 Shucked.weight,  
 Viscera.weight,  
 Shell.weight)  
head(male\_weight)

## Sex Whole.weight Shucked.weight Viscera.weight Shell.weight  
## 1 M 0.5140 0.2245 0.1010 0.150  
## 2 M 0.2255 0.0995 0.0485 0.070  
## 3 M 0.5160 0.2155 0.1140 0.155  
## 4 M 0.5095 0.2165 0.1125 0.165  
## 5 M 0.4060 0.1675 0.0810 0.135  
## 6 M 0.5415 0.2175 0.0950 0.190

### 5. Summarize the mean-length by gender

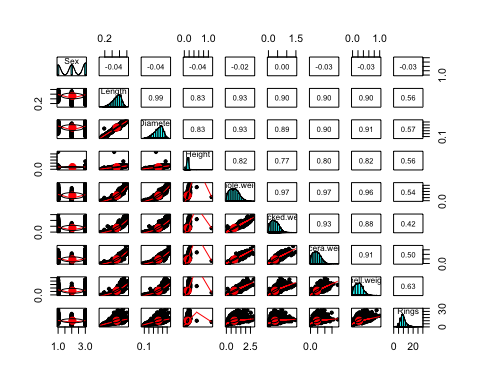
mean\_length <- abalone %>% group\_by(Sex) %>% summarise(mean = mean(Length))  
mean\_length

## # A tibble: 3 × 2  
## Sex mean  
## <fct> <dbl>  
## 1 F 0.579  
## 2 I 0.428  
## 3 M 0.561

## Question 2

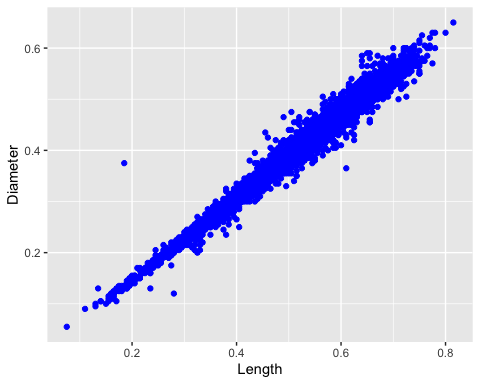
### 1. Create a scatter plot of matrices for the dataset.

pairs.panels(abalone)



### 2. Create a scatter plot of Diameter Vs Length

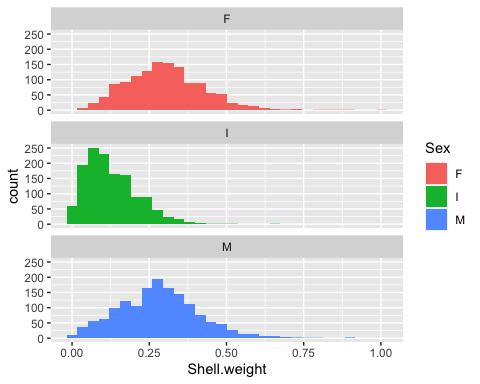
ggplot(data = abalone, mapping = aes(Length, Diameter)) + geom\_point(col = "blue")



### 3. Create a set of histograms of Shell-weight by Sex category and fill with different colurs. (use facet\_wrap with 1 column)

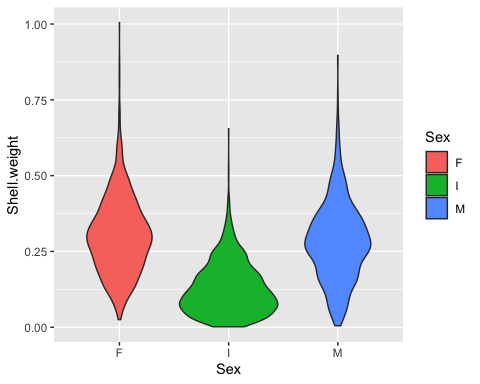
ggplot(data=abalone, mapping=aes( x=Shell.weight, fill=Sex)) +  
 geom\_histogram()+  
 facet\_wrap(~Sex, ncol=1)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



### 4. Create violin plot for Shell-weight by Sex category.

ggplot(data = abalone, mapping = aes(x = Sex, y = Shell.weight, fill = Sex)) +  
 geom\_violin()



### 5. Create box plot for Shell-weight by Sex category.Enhance the plot by adding title, labels etc. . .

ggplot(data = abalone, mapping = aes(x = Sex, y = Shell.weight, fill = Sex)) +  
 geom\_boxplot() +   
 labs(title = "Boxplots for Sex Categories w.r.t Shell-Weight", x = "Sex"  
 , y = "Shell Weight")

