DATA MINING: HOMEWORK 2
NAME: MEYA STEPHEN KENIGBOLO

1a. What are the column names of the dataset?

The column names in the data set are as follows

1. Gender
2. Length
3. Diameter
4. Height
5. Weight
6. Rings

R code => colnames(abalone)length

1b. How many observations (i.e. rows) are in this data frame?

There are 1000 obs (rows) in abalone

R code => nrow(abalone)

1c. Print the first 3 lines from the dataset. What are the values of feature rings of the printed observations?

	Gender	Length	Diameter	Height	Weight	Rings
1	F	0.505	0.385	0.135	0.6185	12
2	F	0.650	0.475	0.165	1.3875	9
3	I	0.520	0.380	0.135	0.5395	8

R code used => head(abalone, 3)

The Values of rings are as follows

	Rings
1	12
2	9
3	8

R code used => abalone[0:3,"Rings", drop=FALSE]

1d. Extract the last 2 rows of the data frame. What is the weight of these abalones?

The last two rows of abalone are

	Gender	Length	Diameter	Height	Weight	Rings
999	I	0.525	0.400	0.130	0.6455	8
1000	M	0.515	0.395	0.135	1.0070	8

R code used => tail(abalone, 3)

The weight of the abalones are

	Weight
999	0.6455
1000	1.0070

R code use => abalone[999:1000,"Weight", drop=FALSE]

1e. What is the value of diameter in the row 755?

The value of Diameter in row 755

	Diameter
755	0.385

R code => abalone[755:755,"Diameter", drop=FALSE]

1f. How many missing values are in the height column?

	Height
21	NA
60	NA
126	NA
168	NA

R code => na_height <- abalone[is.na(abalone$Height),]

R code => na_height[,"Height", drop=FALSE]

1g. What is the mean of the height column? Exclude missing values from this calculation.

The mean of the Height column is

=0.1398092

R code => mean_height <- abalone[,"Height", drop=FALSE]

R code => mean(mean_height$Height, na.rm=TRUE)

1h. Extract the subset of rows of the data frame where gender is M and weight values are below 0.75.

The subset has 199 observable

R code => new_abalone <- subset(abalone, Gender == "M" & Weight < 0.75)

What is the mean of diameter in this subset?

The mean of the Diameter in the subset

=0.3426471

R code => mean_diameter <- new_abalone[,"Diameter", drop=FALSE]

R code => mean(mean_diameter$Diameter, na.rm=TRUE)

1i. What is the most frequent rings value?

Most frequent Value = 9 (Appears 176 times)

R code => sort(table(abalone$Rings), decreasing=TRUE)[1]

1j. What is the minimum of length when rings is equal to 18?

Minimum value - 0.465

R code => minLength <- subset(abalone, Rings == 18)

R code => min(minLength$Length, na.rm = TRUE)

2a. What is the data about?

2b. What are different features and their types (discrete, continuous etc)?

Length => Continuous

Weight => Continuous

Height => Continuous

Diameter => Continuous

Rings => Discrete

Gender => Discrete

2c. How many rows are in the dataset?

There are 1000 rows in the dataset (1001 if we count the row of column names)

R code => nrow(abalone)

2d. For each feature show it's: mean, median, max, min, standard deviation, distribution (with a plot)

=> Comment on the outcome (is the distribution skewed - more small/large values, etc).

Length

Mean = 0.52276

R code => mean(abalone$Length, na.rm = TRUE)

Median = 0.545

R code => median(abalone$Length, na.rm = TRUE)

Max = 0.815

R code => max(abalone$Length, na.rm = TRUE)

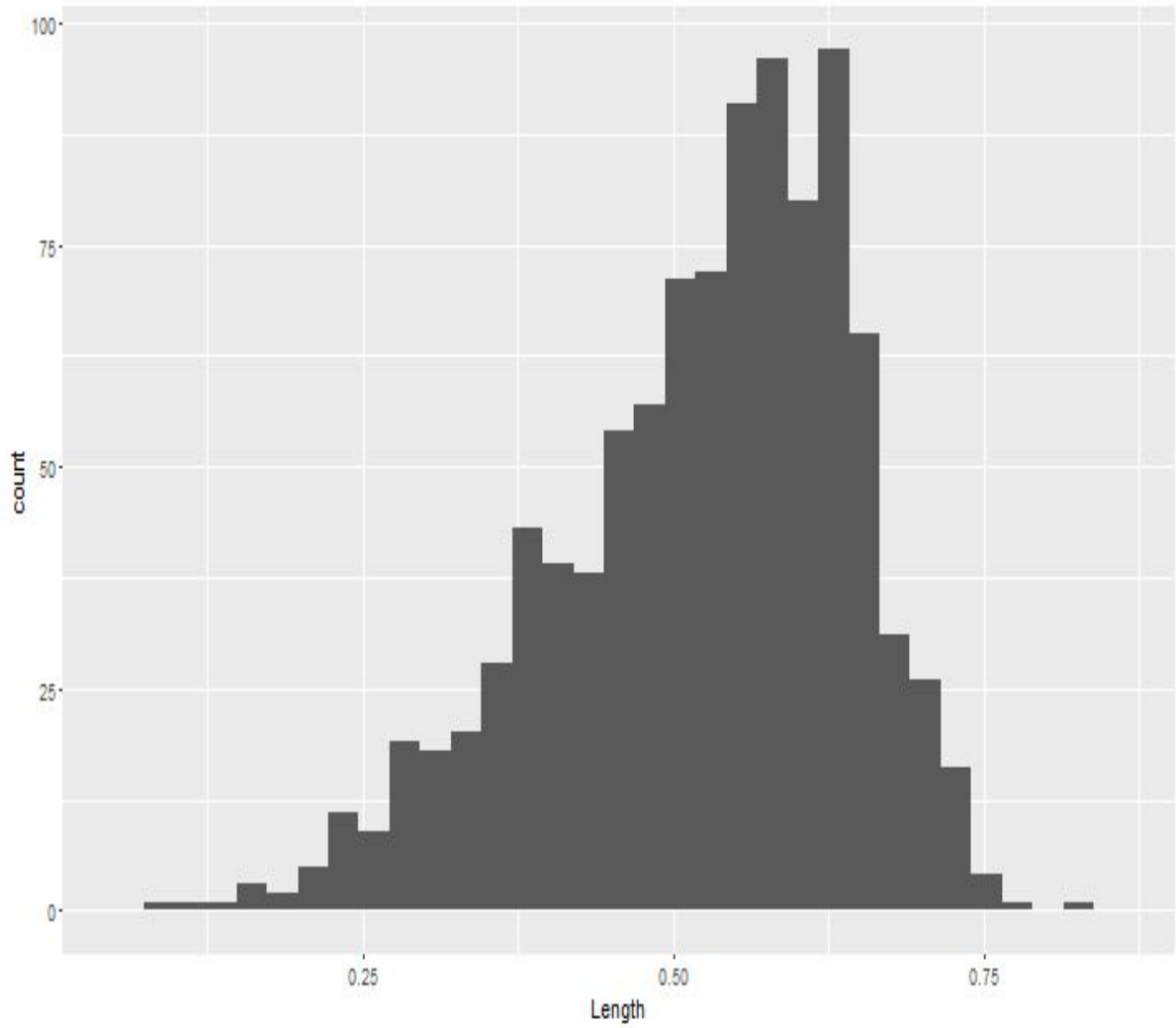
Min = 0.075

R code => min(abalone$Length, na.rm = TRUE)

Standard Deviation = 0.1200641

R code => sd(abalone$Length, na.rm = TRUE)

Distribution (with plot)



R code => qplot(data = abalone, x = Length)

Weight

Mean = 0.8255405

R code => mean(abalone$Weight, na.rm = TRUE)

Median = 0.801

R code => median(abalone$Weight, na.rm = TRUE)

Max = 2.555

R code => max(abalone$Weight, na.rm = TRUE)

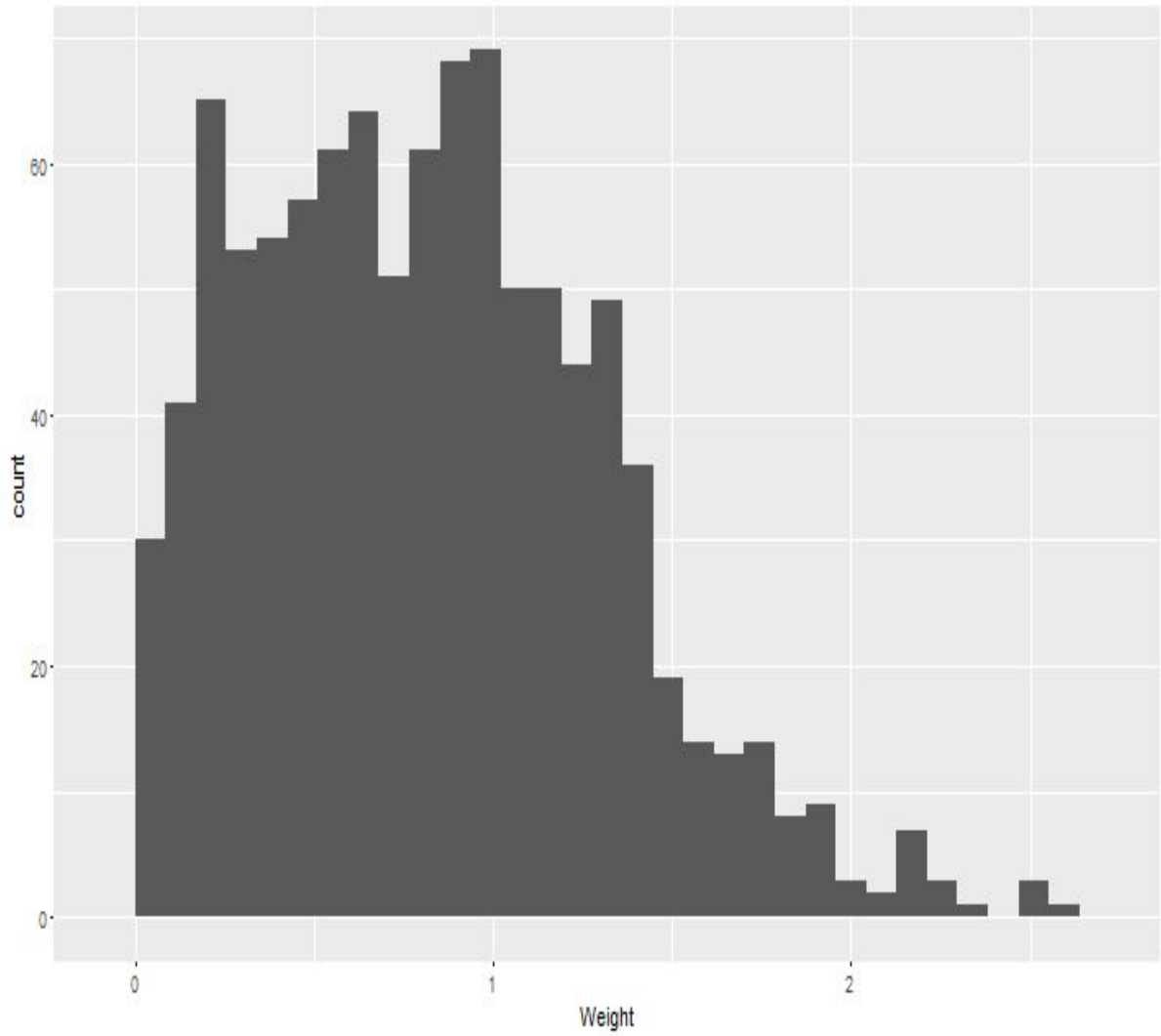
Min = 0.002

R code => min(abalone$Weight, na.rm = TRUE)

Standard Deviation = 0.4903796

R code => sd(abalone$Weight, na.rm = TRUE)

Distribution (with plot)



R code => qplot(data = abalone, x = Weight)

Height

Mean = 0.1398092

R code => mean(abalone$Height, na.rm = TRUE)

Median = 0.1425

R code => median(abalone$Height, na.rm = TRUE)

Max = 1.13

R code => max(abalone$Height, na.rm = TRUE)

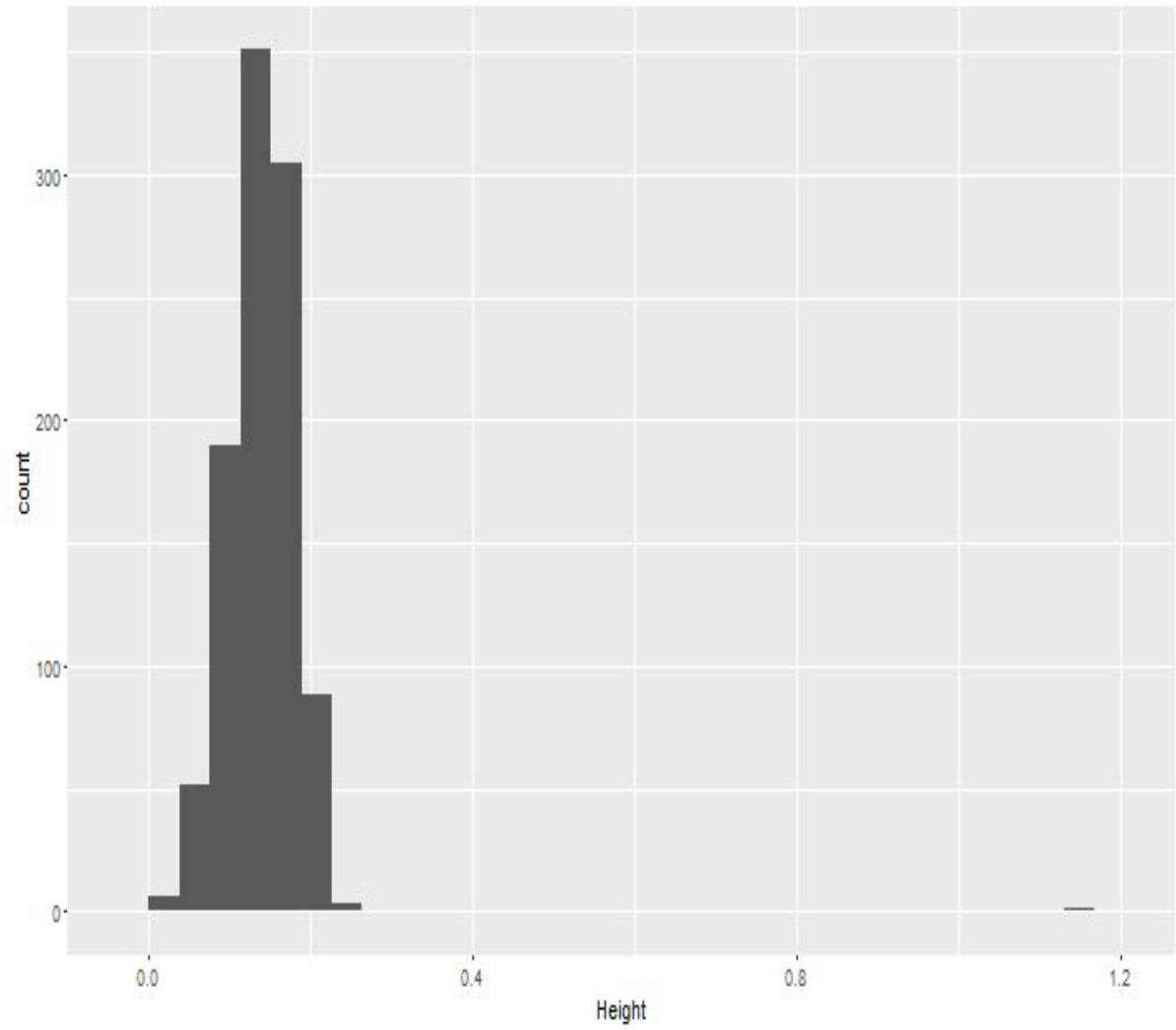
Min = 0.010

R code => min(abalone$Height, na.rm = TRUE)

Standard Deviation = 0.04942548

R code => sd(abalone$Height, na.rm = TRUE)

Distribution (with plot)



R code => qplot(data = abalone, x = Height)

Diameter

Mean = 0.405955

R code => mean(abalone$Diameter, na.rm = TRUE)

Median = 0.42

R code => median(abalone$Diameter, na.rm = TRUE)

Max = 0.65

R code => max(abalone$Diameter, na.rm = TRUE)

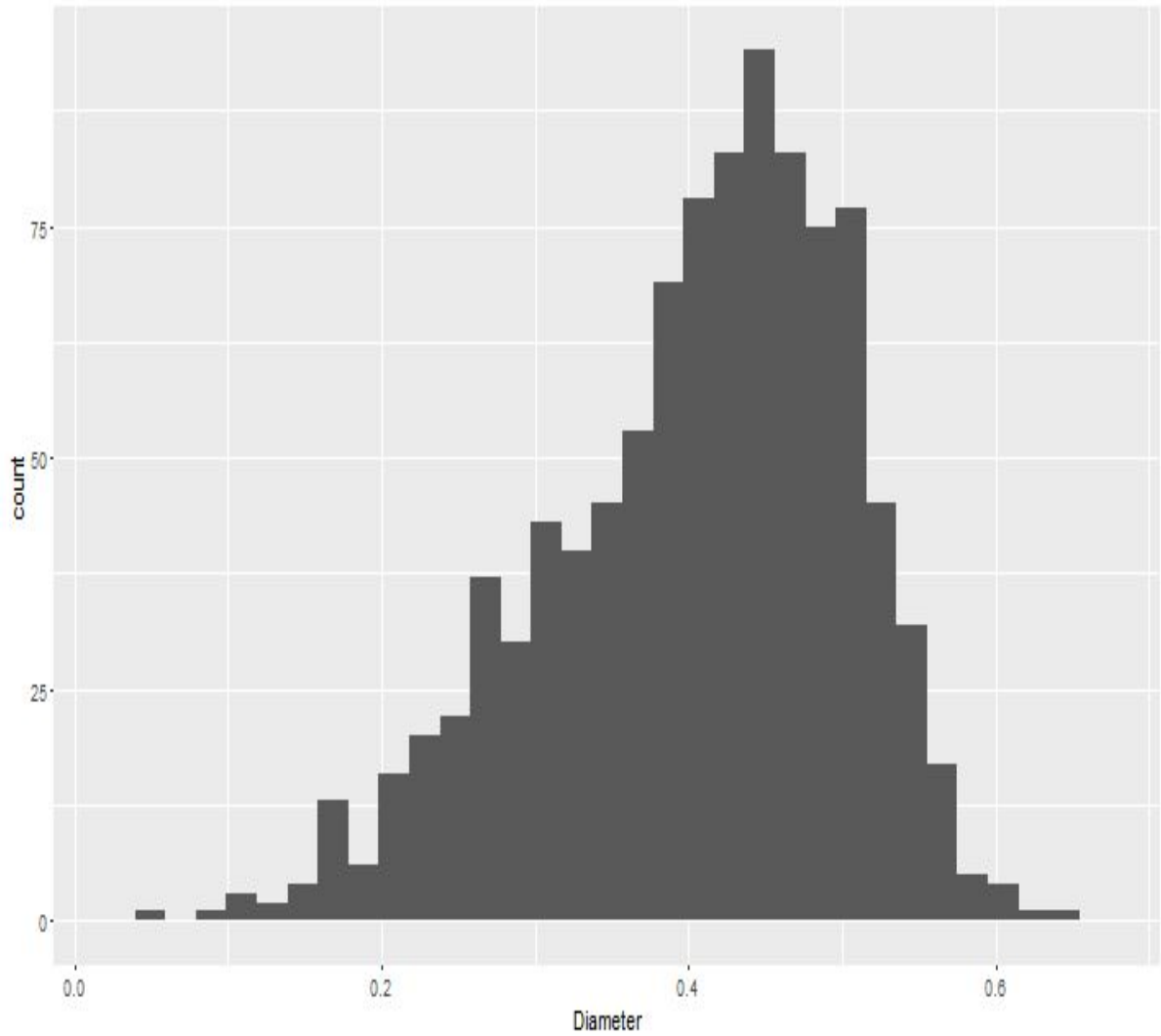
Min = 0.055

R code => min(abalone$Diameter, na.rm = TRUE)

Standard Deviation = 0.09883183

R code => sd(abalone$Diameter, na.rm = TRUE)

Distribution (with plot)



R code => qplot(data = abalone, x = Diameter)

Rings

Mean = 11.318

R code => mean(abalone$Rings, na.rm = TRUE)

Median = 9

R code => median(abalone$Rings, na.rm = TRUE)

Max = 1500

R code => max(abalone$Rings, na.rm = TRUE)

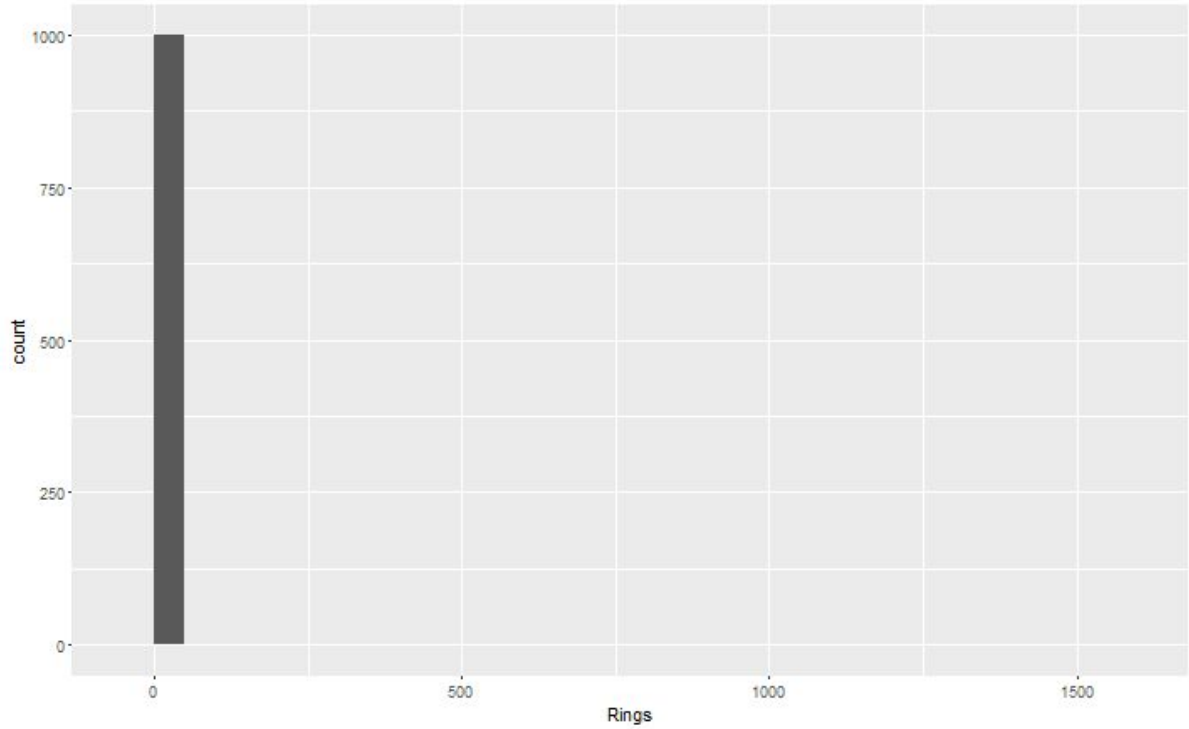
Min = 1

R code => min(abalone$Rings, na.rm = TRUE)

Standard Deviation = 47.2277

R code => sd(abalone$Rings, na.rm = TRUE)

Distribution (with plot)

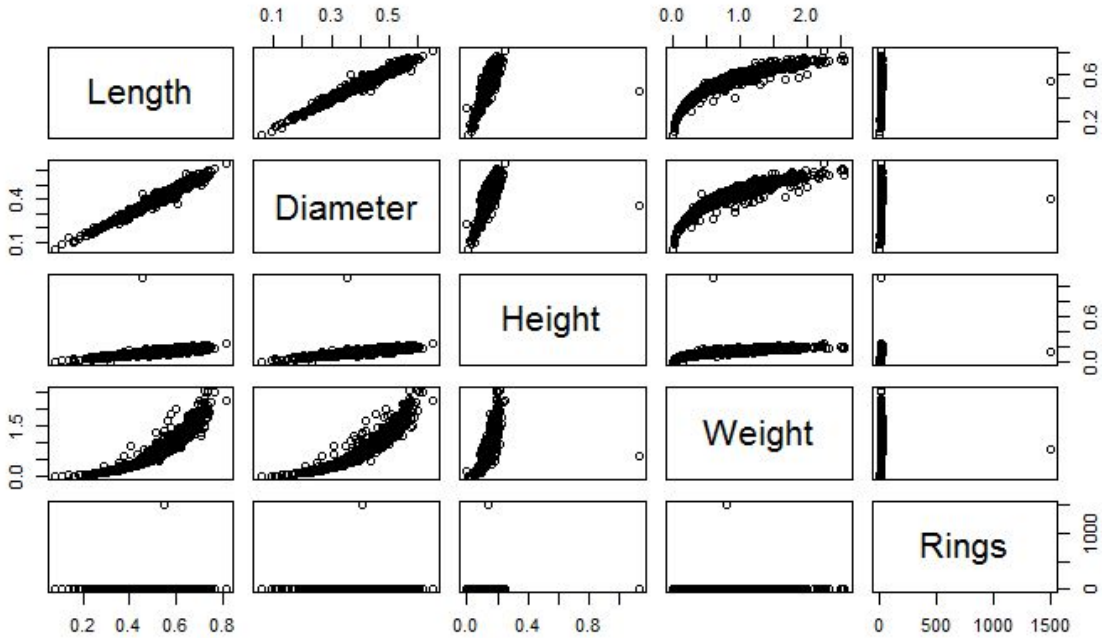


R code => qplot(data = abalone, x = Rings)

Comment on the outcome (is the distribution skewed - more small/large values, etc).

3. Create scatterplots between all variables except gender (in R and Python you can do a scatter matrix between all variables at once, you don't have to do it manually for each pair). Calculate also the correlations between all variables. Plot separately the scatterplot of rings and the variable most correlated with it. Also plot separately two most correlated variables. Explain what you observe.

Scatterplot between all variables



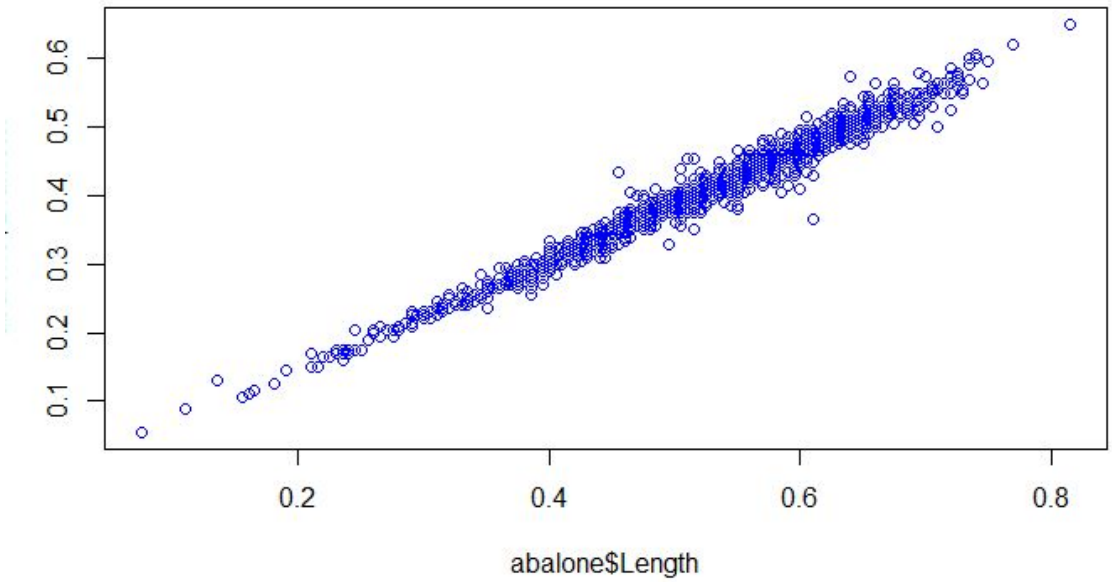
R code => plot(na.omit(abalone[,-1]))

Correlation between all variables

	Length	Diameter	Height	Weight	Rings
Length	1.00000000	0.98747396	0.68725941	0.92071393	0.04407577
Diameter	0.98747396	1.00000000	0.69249343	0.92348680	0.03787742
Height	0.68725941	0.69249343	1.00000000	0.67393746	0.03112661
Weight	0.92071393	0.92348680	0.67393746	1.00000000	0.03385043
Rings	0.04407577	0.03787742	0.03112661	0.03385043	1.00000000

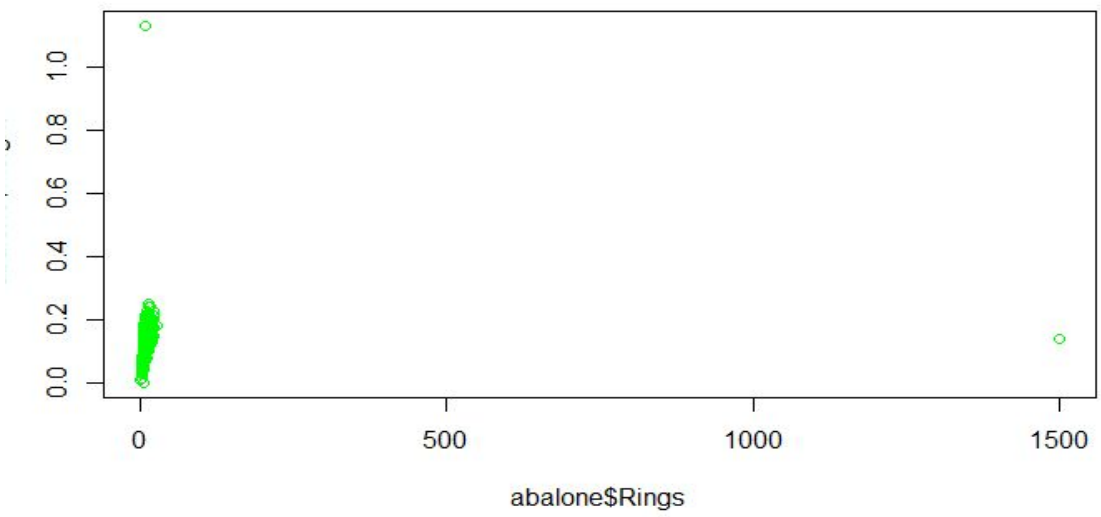
R code => cor(na.omit(abalone[,-1]))

Two most correlated variables



R code => plot(abalone$Length, abalone$Diameter, col = 'blue')

Scatterplot of Rings and Height



R code => plot(abalone$Rings, abalone$Height, col = 'green')

My Observation

4. Identify potential outliers in the dataset by applying inter-quartile ranges (IQR) (range between the 25% and 75% quartiles) - are there values that are over k*IQR larger the upper quartile or k*IQR below lower quartile. Use k=1.5. Are there any multidimesional outliers - observations that are outliers in many variables? *Note: you can also make boxplots to gain some intuition about the outliers.*

Summary of abalone dataframe

Gender	Length	Diameter	Height	Weight	Rings
F:301	Min. :0.0750	Min. :0.055	Min. :0.0000	Min. :0.0020	Min. : 1.00
I:317	1st Qu.:0.4500	1st Qu.:0.345	1st Qu.:0.1150	1st Qu.:0.4389	1st Qu.: 8.00
M:382	Median :0.5450	Median :0.420	Median :0.1425	Median :0.8010	Median : 9.00
	Mean :0.5228	Mean :0.406	Mean :0.1398	Mean :0.8255	Mean : 11.32
	3rd Qu.:0.6150	3rd Qu.:0.480	3rd Qu.:0.1650	3rd Qu.:1.1462	3rd Qu.: 11.00
	Max. :0.8150	Max. :0.650	Max. :1.1300	Max. :2.5550	Max. :1500.00
			NA's :4		

R code => summary(abalone)

IQR of variables in Abalone dataframe

Length IQR =	0.165	R code => IQR(abalone$Length)
Diameter IQR =	0.135	R code => IQR(abalone$Diameter)
Height IQR =	0.05	R code => IQR(abalone$Height)
Weight IQR =	0.707375	R code => IQR(abalone$Weight)
Rings IQR =	3	R code => IQR(abalone$Rings)

*IQR * k (where k = 1.5)*

Length =	0.247
Diameter =	0.2025
Height =	0.075
Weight =	1.0610625
Rings =	4.5

*Values that are over k*IQR larger the upper quartile or k*IQR below lower quartile*

Length (k*IQR below lower quartile)

0.135 0.245 0.235 0.235 0.230 0.225 0.180 0.230 0.210 0.165 0.210 0.235 0.245 0.160 0.210 0.240 0.215 0.235 0.240 0.110 0.190 0.155 0.220 0.075

R code => lenghtValue <- subset(abalone, Length < 0.247)

R code => print(lenghtValue$Length)

Diameter (k*IQR below lower quartile)

0.130 0.170 0.160 0.170 0.165 0.195 0.175 0.125 0.175 0.170 0.190 0.115 0.150 0.175 0.190 0.175
0.110 0.150 0.170 0.150 0.170 0.190 0.200 0.175 0.090 0.145 0.105 0.195 0.165 0.055 0.195

R code => DiameterValue <- subset(abalone, Diameter < 0.205)

R code => print(DiameterValue$Diameter)

Height (k*IQR below lower quartile)

0.040 0.060 0.065 0.070 0.060 0.060 0.050 0.055 0.060 0.050 0.070 0.055 0.060 0.035 0.065 0.045
0.050 0.015 0.045 0.040 0.070 0.055 0.025 0.070 0.000 0.055 0.050 0.055 0.070 0.055 0.070 0.055
0.060 0.030 0.065 0.060 0.040 0.070 0.050 0.070 0.050 0.055 0.010 0.065

R code => HeightValue <- subset(abalone, Height < 0.075)

R code => print(HeightValue$Height)

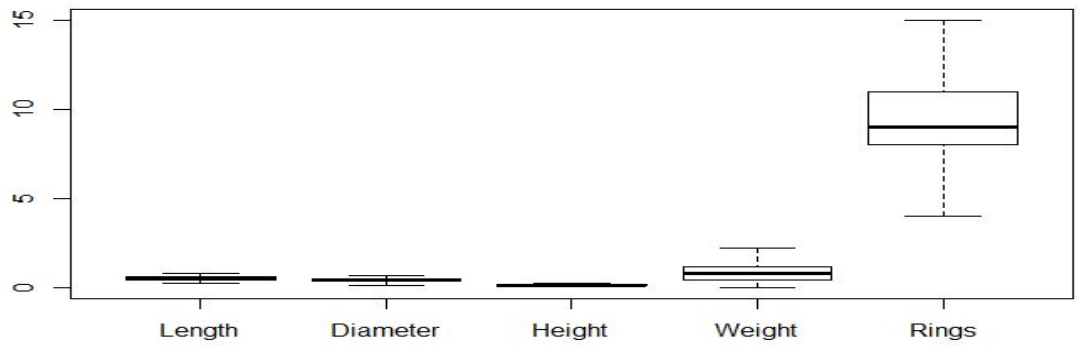
Rings (k*IQR below lower quartile)

4 4 4 4 4 4 4 4 4 4 4 3 3 4 4 4 1

R code => RingsValue <- subset(abalone, Rings < 4.5)

R code => print(RingsValue$Rings)

Using Box plot to analyse outliers



Multidimensional Outliers

0.165

0.160

0.110

0.190

5. Explain if you should remove outliers and if so, then which ones (you do not need to decide for each observation separately, try to explain in general). Remove the outliers you decided to remove and recalculate statistics from ex. 2d. What do you observe?

Considering the multidimensional outliers are few I believe they are negligible hence I will remove them from the data frame and recalculate

```
new_Abalone <- subset(newabalone, Diameter = 0.165 | 0.160 | 0.110 | 0.190)
new_Abalone <- subset(new_Abalone, Diameter = 0.165 | 0.160 | 0.110 | 0.190)
```

Length

Mean = 0.52276

R code => mean(new_Abalone$Length, na.rm = TRUE)

Median = 0.545

R code => median(new_Abalone$Length, na.rm = TRUE)

Max = 0.815

R code => max(new_Abalone$Length, na.rm = TRUE)

Min = 0.075

R code => min(new_Abalone$Length, na.rm = TRUE)

Standard Deviation = 0.1200641

R code => sd(new_Abalone$Length, na.rm = TRUE)