Regression Abalone

Parker Lambert

2024-09-08

Linear Regression Example

Libaries

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(e1071)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

Data Introductions

For this project I will be utilizing a database on Abalone's that was sourced from https://archive.ics.uci.edu/dataset/1/abalone

What is an abalone

An abalone is a type of marine mollusk belonging to the family Haliotidae. It is a single-shelled sea snail, known for its ear-shaped shell, which is lined with a beautiful layer of iridescent nacre, or "mother of pearl." The outer shell is rough and often encrusted with marine organisms, while the inside is smooth and colorful.

Abalones are prized both for their meat, which is considered a delicacy in many cultures, and for their shells, which are often used in jewelry and decorative items. They are typically found in cold coastal waters, clinging to rocks and feeding on algae. In some areas, abalone populations have been severely reduced due to overfishing and environmental changes, leading to various conservation efforts.

Load Data

```
(abalone <- read.csv('abalone/abalone.data')) |>
  head(3)
```

```
##   M X0.455 X0.365 X0.095 X0.514 X0.2245 X0.101 X0.15 X15
## 1 M   0.35  0.265  0.090 0.2255  0.0995 0.0485 0.070   7
## 2 F   0.53  0.420  0.135 0.6770  0.2565 0.1415 0.210   9
## 3 M   0.44  0.365  0.125 0.5160  0.2155 0.1140 0.155  10
```

Now I will add column names

```
colnames(abalone) <- c("Sex", "LongestShell", "Diameter", "Height", "WholeWeight", "ShuckedWeight", "Vi
abalone |>
  head(3)
```

```
##   Sex LongestShell Diameter Height WholeWeight ShuckedWeight VisceraWeight
## 1   M         0.35    0.265  0.090      0.2255        0.0995        0.0485
## 2   F         0.53    0.420  0.135      0.6770        0.2565        0.1415
## 3   M         0.44    0.365  0.125      0.5160        0.2155        0.1140
##   ShellWeight Rings
## 1       0.070     7
## 2       0.210     9
## 3       0.155    10
```

Exploritory Data Analysis

Getting to know the Data

Data Types

```r
abalone |>
  str()
```

```
## 'data.frame':    4176 obs. of  9 variables:
##  $ Sex          : chr  "M" "F" "M" "I" ...
##  $ LongestShell : num  0.35 0.53 0.44 0.33 0.425 0.53 0.545 0.475 0.55 0.525 ...
##  $ Diameter     : num  0.265 0.42 0.365 0.255 0.3 0.415 0.425 0.37 0.44 0.38 ...
##  $ Height       : num  0.09 0.135 0.125 0.08 0.095 0.15 0.125 0.125 0.15 0.14 ...
##  $ WholeWeight  : num  0.226 0.677 0.516 0.205 0.351 ...
##  $ ShuckedWeight: num  0.0995 0.2565 0.2155 0.0895 0.141 ...
##  $ VisceraWeight: num  0.0485 0.1415 0.114 0.0395 0.0775 ...
##  $ ShellWeight  : num  0.07 0.21 0.155 0.055 0.12 0.33 0.26 0.165 0.32 0.21 ...
##  $ Rings        : int  7 9 10 7 8 20 16 9 19 14 ...
```

above we see that we are dealing with 8 numerical variables and one factor in `Sex` however currently `Sex` is type character

Convert to Sex to Factor

```r
abalone$Sex <- as.factor(abalone$Sex)
```

Later on we might want to `one-hot, lable or target encode` Sex however for now factor is enough

Check for missing

```r
colSums(is.na(abalone))
```

```
##           Sex  LongestShell      Diameter        Height   WholeWeight
##             0             0             0             0             0
## ShuckedWeight VisceraWeight   ShellWeight         Rings
##             0             0             0             0
```

We are fortunate to have a data set will NA missing values. With further tests we will check if this is becasue the dataset is complete or imputation has occurred.

Descriptive Statistics

Column Summary Statistics

```r
abalone |>
  summary()
```

```
##  Sex       LongestShell      Diameter          Height         WholeWeight
##  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:1527   Median :0.545   Median :0.4250   Median :0.1400   Median :0.7997
##           Mean   :0.524   Mean   :0.4079   Mean   :0.1395   Mean   :0.8288
##           3rd Qu.:0.615   3rd Qu.:0.4800   3rd Qu.:0.1650   3rd Qu.:1.1533
##           Max.   :0.815   Max.   :0.6500   Max.   :1.1300   Max.   :2.8255
##  ShuckedWeight    VisceraWeight      ShellWeight         Rings
##  Min.   :0.0010   Min.   :0.00050   Min.   :0.0015   Min.   : 1.000
##  1st Qu.:0.1860   1st Qu.:0.09337   1st Qu.:0.1300   1st Qu.: 8.000
##  Median :0.3360   Median :0.17100   Median :0.2340   Median : 9.000
##  Mean   :0.3594   Mean   :0.18061   Mean   :0.2389   Mean   : 9.932
##  3rd Qu.:0.5020   3rd Qu.:0.25300   3rd Qu.:0.3290   3rd Qu.:11.000
##  Max.   :1.4880   Max.   :0.76000   Max.   :1.0050   Max.   :29.000
```

Check for Skewness Kurtosis

```r
numeric_cols <- abalone[, sapply(abalone, is.numeric)]

skewness_values <- apply(numeric_cols, 2, function(x) skewness(x, na.rm = TRUE))
kurtosis_values <- apply(numeric_cols, 2, function(x) kurtosis(x, na.rm = TRUE))

data.frame(Variable = colnames(numeric_cols), Skewness = skewness_values, Kurtosis = kurtosis_values) |>
  print()
```

```
##                    Variable   Skewness    Kurtosis
## LongestShell   LongestShell -0.6397802  0.06171758
## Diameter           Diameter -0.6090196 -0.04847050
## Height               Height  3.1269930 75.91573416
## WholeWeight     WholeWeight  0.5301946 -0.02696777
## ShuckedWeight ShuckedWeight  0.7182081  0.59057946
## VisceraWeight VisceraWeight  0.5910385  0.08056177
## ShellWeight     ShellWeight  0.6201013  0.52758786
## Rings                 Rings  1.1143559  2.32915419
```

Link for Skewness & Kurtosis

To Interpret Skewness:

- -0.5 to 0 and 0 to 0.5: Near Symmetrical
- -1 to -0.5 and 0.5 to 1: moderate negative/left skew and moderate positive/right Skew
- < -1 and > 1: high negative/left skew and high positive/right skew

To Interpret Kurtosis:

- Expected value is 3 for a `Normal Distribution`
- <3 negative/low kurtosis or `Platykurtic` aka slight squish or heavy tails
- >3 positive/high kurtosis or `Leptokurtic` aka slight pull up or light tails

High Kurtosis signals the presence of outliers

Low Kurtosis means fewer extreme outliers

With these interpretations in mind we can see that for our variables we have:

`LongestShell:` Moderate Left Skew Slight Negative Kurtosis Squish

`Diameter:` Moderate Left Skew Slight Negative Kurtosis Squish

`Height:` High Right Skew Extreme Positive Kurtosis Pull

`WholeWeight:` Moderate Right Skew Slight Negative Kurtosis Squish

`ShuckedWeight:` Moderate Right Skew Slight Negative Kurtosis Squish

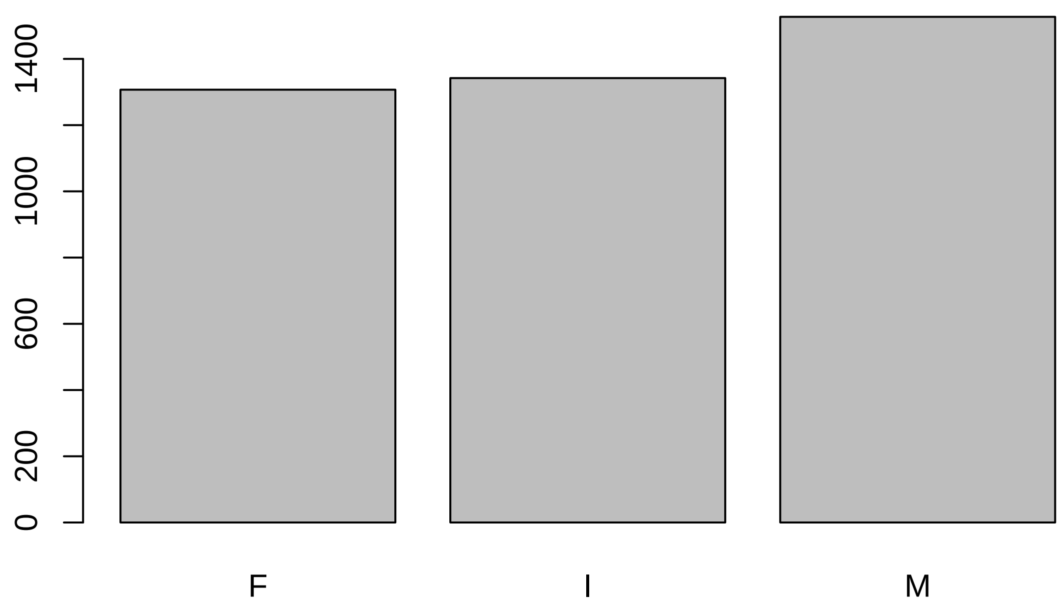
`VisceraWeight:` Moderate Right Skew Slight Negative Kurtosis Squish

`ShellWeight:` Moderate Right Skew Slight Negative Kurtosis Squish

`Rings:` High Right Skew Normal Kurtosis

Visualize the Data

```
barplot(table(abalone$Sex))
```



Sex

`Sex`

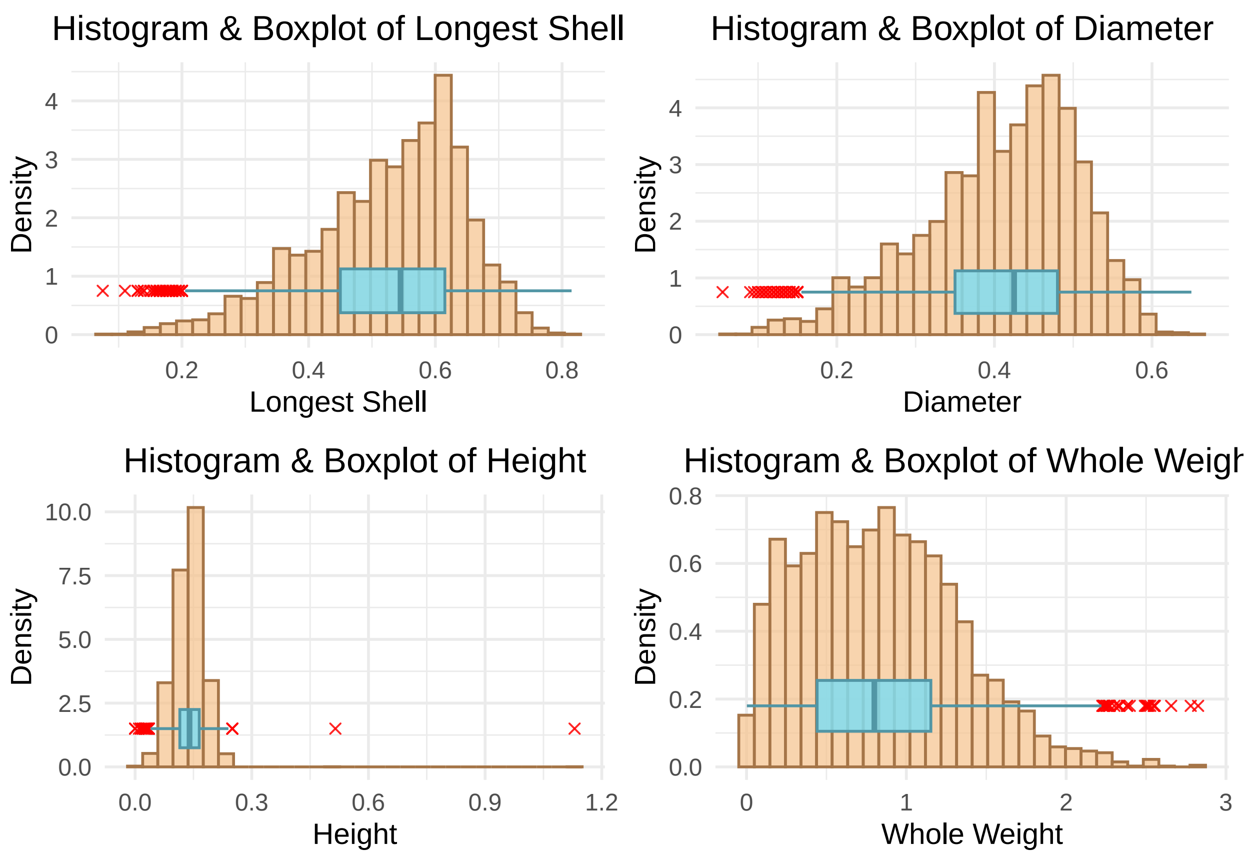
We can see that the groups are roughly evenly split between male female and infant

```
g1 <- ggplot(abalone, aes(x = LongestShell)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = "#f6c28b", color = "#a57548", alpha = 0.7) +
  geom_boxplot(aes(y = 0.75), width = 0.75, fill = "#82ddf0", color = "#5296a5", alpha = 0.85, outlier.s
  theme_minimal() +
  labs(title = "Histogram & Boxplot of Longest Shell", x = "Longest Shell", y = "Density") +
  theme(plot.title = element_text(hjust = 0.5))
g2 <- ggplot(abalone, aes(x = Diameter)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = "#f6c28b", color = "#a57548", alpha = 0.7) +
  geom_boxplot(aes(y = 0.75), width = 0.75, fill = "#82ddf0", color = "#5296a5", alpha = 0.85, outlier.s
  theme_minimal() +
  labs(title = "Histogram & Boxplot of Diameter", x = "Diameter", y = "Density") +
  theme(plot.title = element_text(hjust = 0.5))
g3 <- ggplot(abalone, aes(x = Height)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = "#f6c28b", color = "#a57548", alpha = 0.7) +
  geom_boxplot(aes(y = 1.5), width = 1.5, fill = "#82ddf0", color = "#5296a5", alpha = 0.85, outlier.sha
  theme_minimal() +
  labs(title = "Histogram & Boxplot of Height", x = "Height", y = "Density") +
  theme(plot.title = element_text(hjust = 0.5))
g4 <- ggplot(abalone, aes(x = WholeWeight)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = "#f6c28b", color = "#a57548", alpha = 0.7) +
  geom_boxplot(aes(y = 0.18), width = 0.15, fill = "#82ddf0", color = "#5296a5", alpha = 0.85, outlier.s
  theme_minimal() +
  labs(title = "Histogram & Boxplot of Whole Weight", x = "Whole Weight", y = "Density") +
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(g1, g2, g3, g4, nrow = 2, ncol = 2)
```

LongestShell - Diameter - Height - WholeWeight

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



`LongestShell`

The data appears slightly negatively skewed, confirmed by the left tail in the histogram. The boxplot also shows several outliers on the lower end, suggesting some unusually short shell lengths.

`Diameter`

This variable is slightly negatively skewed, with the bulk of the data concentrated in the middle range. The boxplot shows a few outliers on the lower end, indicating some observations with smaller diameters

`Height`

The histogram shows a strong positive skew with a sharp peak at a low value, and the boxplot reveals significant outliers at higher heights, highlighting the extreme values.

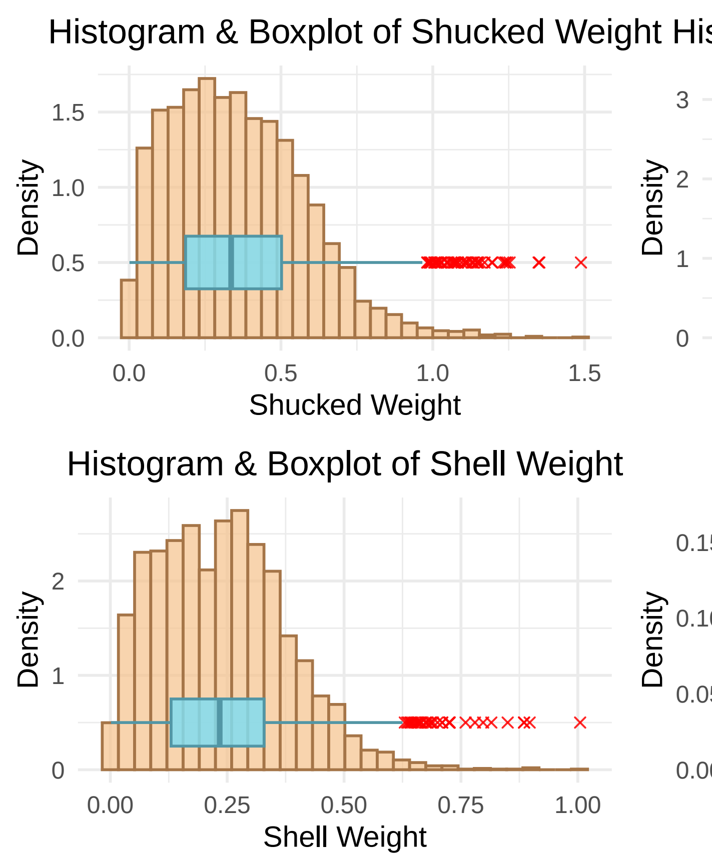
`WholeWeight`

The distribution is moderately positively skewed, with most data concentrated around lower weights. The boxplot shows several high outliers, indicating a few instances of unusually heavy whole weights

```
g5 <- ggplot(abalone, aes(x = ShuckedWeight)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = "#f6c28b", color = "#a57548", alpha = 0.7) +
  geom_boxplot(aes(y = 0.5), width = 0.35, fill = "#82ddf0", color = "#5296a5", alpha = 0.85, outlier.s
  theme_minimal() +
  labs(title = "Histogram & Boxplot of Shucked Weight", x = "Shucked Weight", y = "Density") +
  theme(plot.title = element_text(hjust = 0.5))
g6 <- ggplot(abalone, aes(x = VisceraWeight)) +
```

```
  geom_histogram(aes(y = ..density..), bins = 30, fill = "#f6c28b", color = "#a57548", alpha = 0.7) +
  geom_boxplot(aes(y = 0.5), width = 0.5, fill = "#82ddf0", color = "#5296a5", alpha = 0.85, outlier.sha
  theme_minimal() +
  labs(title = "Histogram & Boxplot of Viscera Weight", x = "Viscera Weight", y = "Density") +
  theme(plot.title = element_text(hjust = 0.5))
g7 <- ggplot(abalone, aes(x = ShellWeight)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = "#f6c28b", color = "#a57548", alpha = 0.7) +
  geom_boxplot(aes(y = 0.5), width = 0.5, fill = "#82ddf0", color = "#5296a5", alpha = 0.85, outlier.sha
  theme_minimal() +
  labs(title = "Histogram & Boxplot of Shell Weight", x = "Shell Weight", y = "Density") +
  theme(plot.title = element_text(hjust = 0.5))
g8 <- ggplot(abalone, aes(x = Rings)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = "#f6c28b", color = "#a57548", alpha = 0.7) +
  geom_boxplot(aes(y = 0.05), width = 0.035, fill = "#82ddf0", color = "#5296a5", alpha = 0.85, outlier
  theme_minimal() +
  labs(title = "Histogram & Boxplot of Rings", x = "Rings", y = "Density") +
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(g5, g6, g7, g8, nrow = 2, ncol = 2)
```



ShuckedWeight - VisceraWeight - ShellWeight - Rings

`ShuckedWeight`

The distribution is positively skewed, with the majority of values concentrated on the lower end. The boxplot indicates the presence of multiple high outliers, showing instances of unusually high shucked weights.

`VisceraWeight`

This variable is also positively skewed with a dense concentration of lower values. The boxplot reveals several high outliers, suggesting some instances of heavier viscera weights than expected.

`ShellWeight`

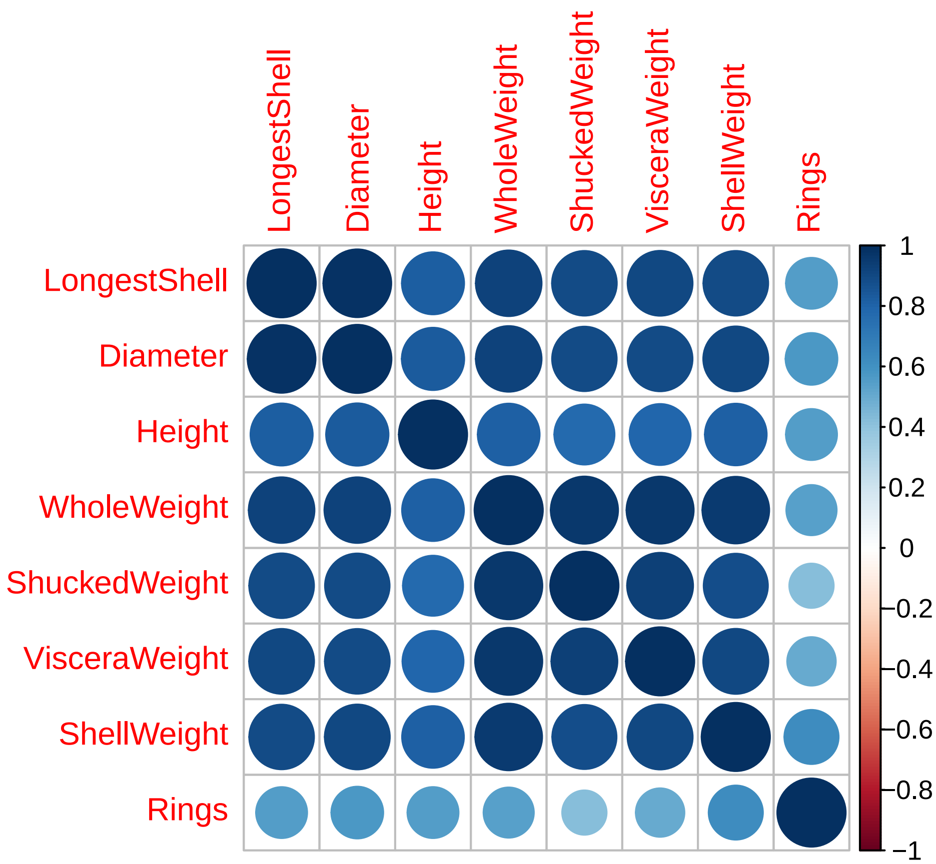
The distribution shows positive skewness with most of the data near the lower values. The boxplot highlights many high outliers, which point to a few unusually heavy shells.

`Rings`

The distribution is moderately positively skewed with a peak near the center of the range. The boxplot also reveals a number of outliers on the higher end, indicating some individuals with a greater number of rings than typical.

Correlation

```
# cor_matrix <- cor(abalone[, sapply(abalone, is.numeric)])
corrplot(cor(abalone[, sapply(abalone, is.numeric)]), method = "circle")
```



As we can see there is a strong correlation between most of the weight related variables. The Rings variable has a relatively low correlation with some attributes, indicating that it behaves somewhat independently of other variables like weight or size.