Analysing and visualizing the Datasets using R MarkDown

By Akhila Kvns

16 January, 2017

# Introduction

This is an introduction to R Markdown. Here i am going to visualize the following datasets:

* ChickWeight
* Iris

## Description

**1. CHICKWEIGHT**

The ChickWeight data frame has 250 rows and 4 columns fro an experiment on the effect of diet on early growth of chicks.

library(datasets)  
data(ChickWeight)  
summary(ChickWeight)

## weight Time Chick Diet   
## Min. : 35.0 Min. : 0.00 13 : 12 1:220   
## 1st Qu.: 63.0 1st Qu.: 4.00 9 : 12 2:120   
## Median :103.0 Median :10.00 20 : 12 3:120   
## Mean :121.8 Mean :10.72 10 : 12 4:118   
## 3rd Qu.:163.8 3rd Qu.:16.00 17 : 12   
## Max. :373.0 Max. :21.00 19 : 12   
## (Other):506

head(ChickWeight)

## weight Time Chick Diet  
## 1 42 0 1 1  
## 2 51 2 1 1  
## 3 59 4 1 1  
## 4 64 6 1 1  
## 5 76 8 1 1  
## 6 93 10 1 1

**2. IRIS**

Theiris dataset gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

library(datasets)  
data(iris)  
summary(iris)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100   
## 1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300   
## Median :5.800 Median :3.000 Median :4.350 Median :1.300   
## Mean :5.843 Mean :3.057 Mean :3.758 Mean :1.199   
## 3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800   
## Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500   
## Species   
## setosa :50   
## versicolor:50   
## virginica :50   
##   
##   
##

head(iris)

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## 1 5.1 3.5 1.4 0.2 setosa  
## 2 4.9 3.0 1.4 0.2 setosa  
## 3 4.7 3.2 1.3 0.2 setosa  
## 4 4.6 3.1 1.5 0.2 setosa  
## 5 5.0 3.6 1.4 0.2 setosa  
## 6 5.4 3.9 1.7 0.4 setosa

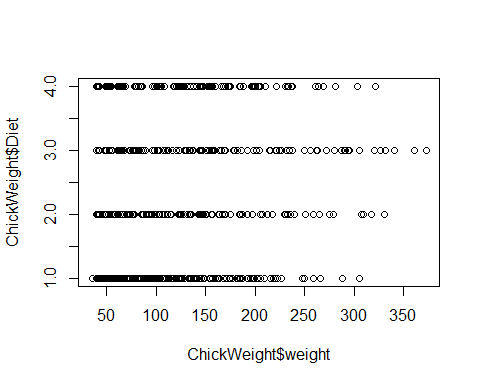
Now we are using the following functions on the datsets.

1. plot()
2. hist()

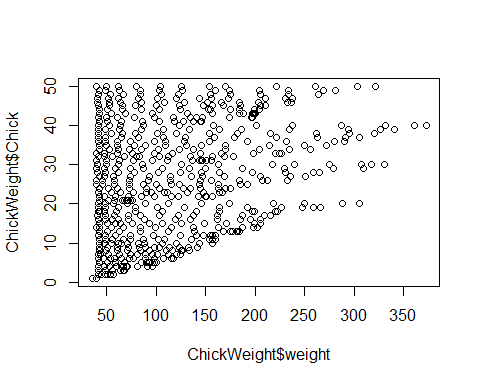
### Plot()

* ChickWeight

plot(ChickWeight$weight , ChickWeight$Diet)

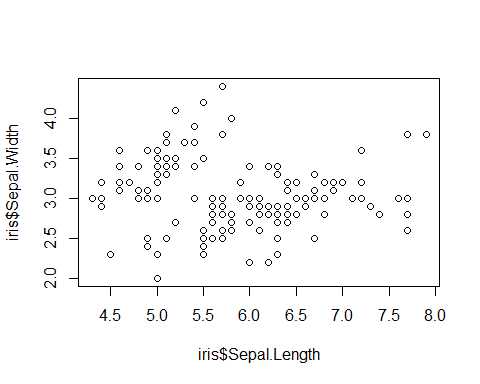


plot(ChickWeight$weight , ChickWeight$Chick)

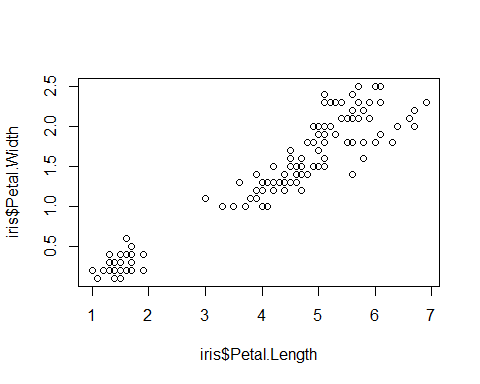


* Iris

plot(iris$Sepal.Length, iris$Sepal.Width)



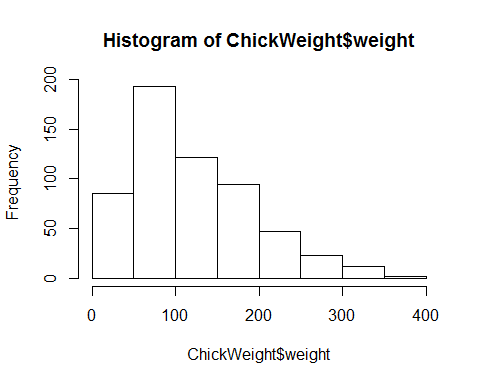
plot(iris$Petal.Length, iris$Petal.Width)



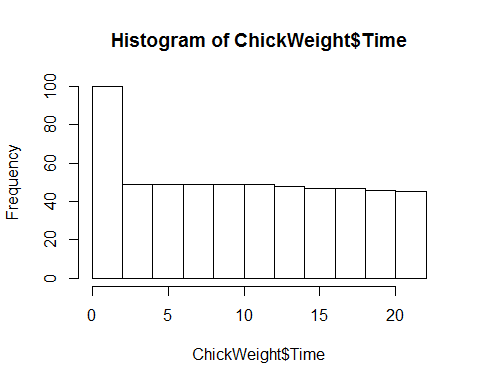
### Hist()

* ChickWeight

hist(ChickWeight$weight)

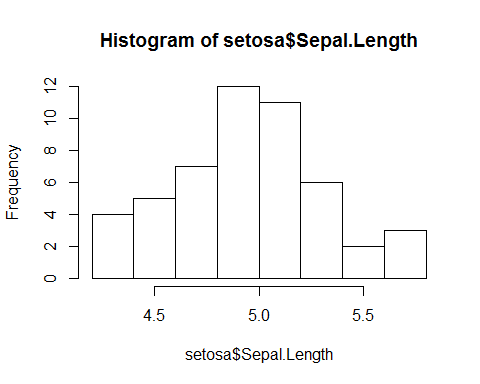


hist(ChickWeight$Time)

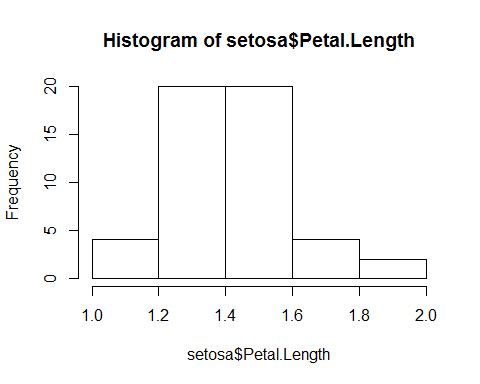


* Iris

iris\_obs <- iris$Species == "setosa"  
setosa <- iris[iris\_obs,]  
hist(setosa$Sepal.Length , breaks=10)



hist(setosa$Petal.Length, breaks=5)

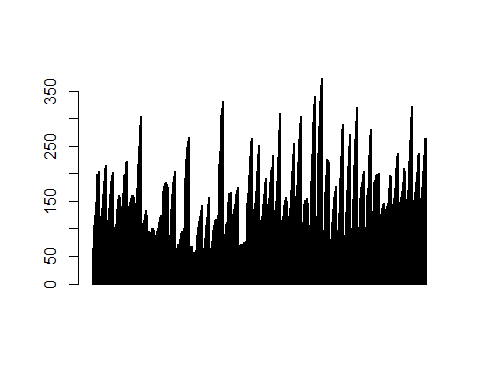


## Bar plot()

Here we are using bar plot to plot our datsets

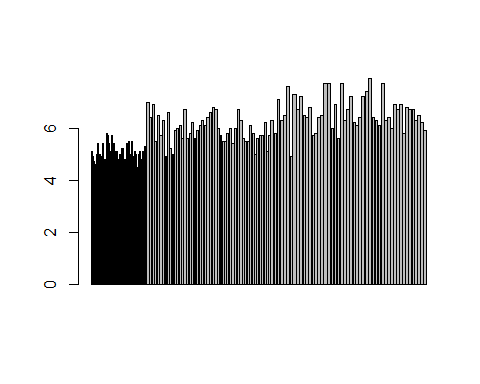
* ChickWeight

barplot(ChickWeight$weight, ChickWeight$Time)



* Iris

barplot(iris$Sepal.Length, iris$Petal.Length)

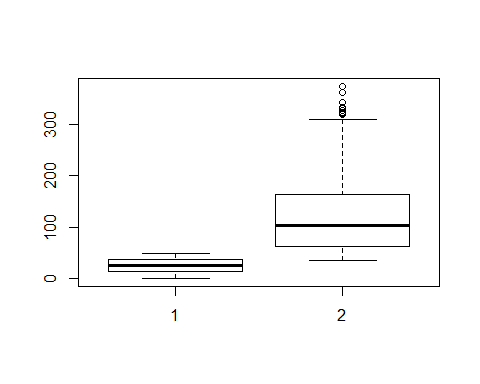


## Box plot()

Here we use box plot to plot both the datasets.

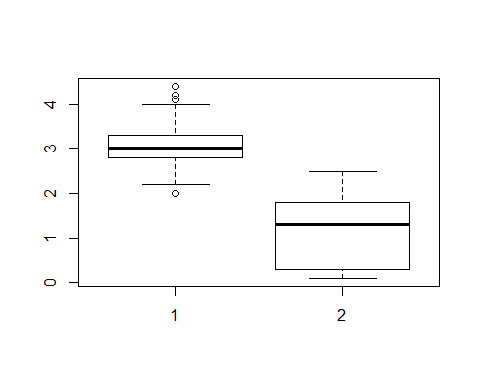
1. ChickWeight

boxplot(ChickWeight$Chick, ChickWeight$weight)



1. Iris

boxplot(iris$Sepal.Width, iris$Petal.Width)

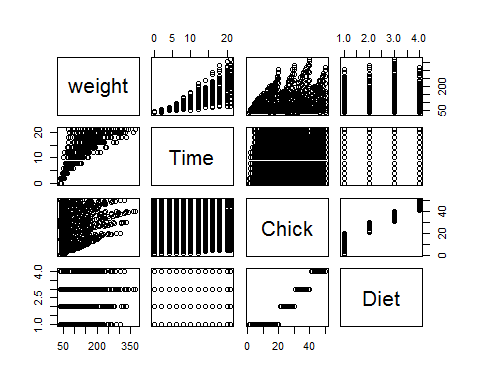


## Pairs()

For finding the relation between two entities of a dataset, we use pairs() function.

* ChickWeight

pairs(ChickWeight)



* Iris

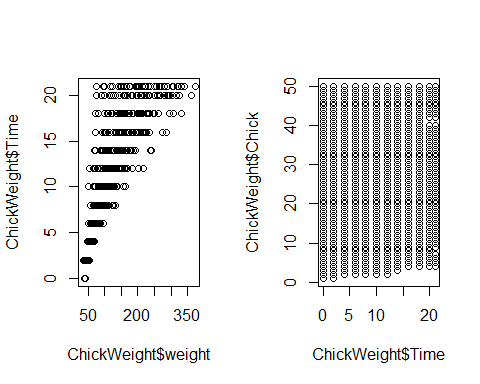
pairs(iris)



Now, we use the par() function which is used for listing all the parameters and plotting multiple plots.

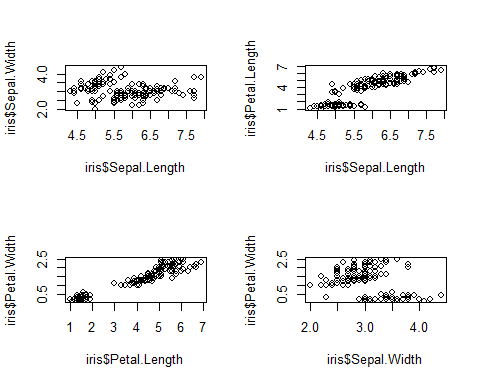
## Par() for ChickWeight

par(mfrow=c(1,2))  
plot(ChickWeight$weight, ChickWeight$Time)  
plot(ChickWeight$Time, ChickWeight$Chick)



## Par() for Iris

par(mfcol=c(2,2))  
plot(iris$Sepal.Length, iris$Sepal.Width)  
plot(iris$Petal.Length, iris$Petal.Width)  
plot(iris$Sepal.Length, iris$Petal.Length)  
plot(iris$Sepal.Width, iris$Petal.Width)

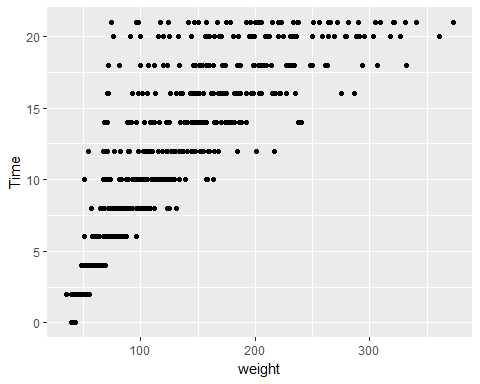


# Using ggplots

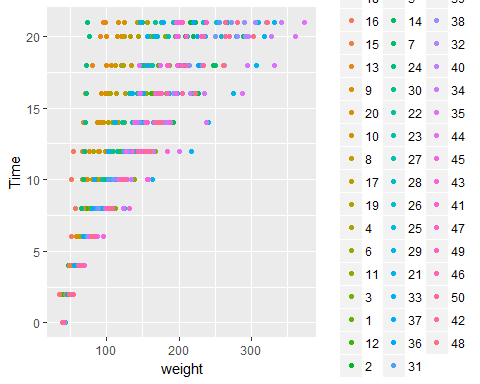
For plotting the ggplot we use qplot() function.

1. ChickWeight

library(ggplot2)  
qplot(weight, Time , data=ChickWeight)

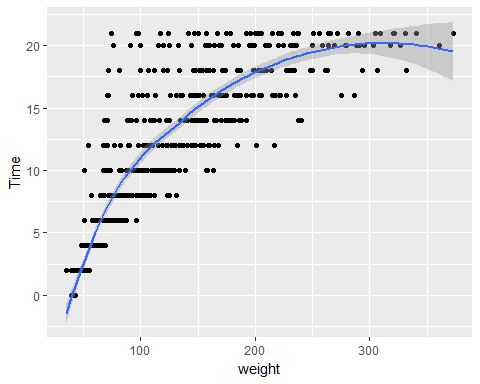


qplot(weight, Time, data=ChickWeight, color=Chick)



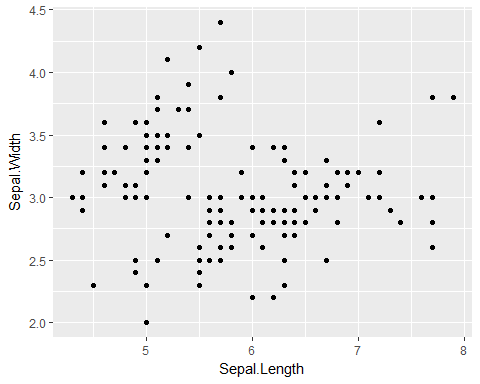
qplot(weight, Time, data=ChickWeight, geom= c("point", "smooth"))

## `geom\_smooth()` using method = 'loess'

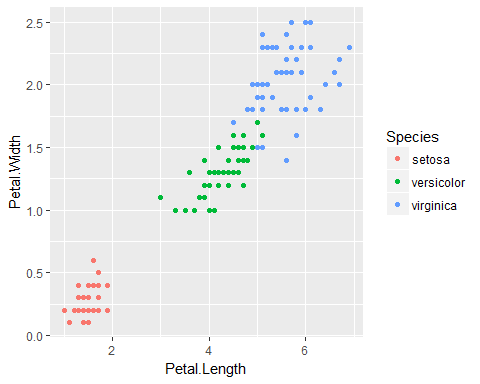


1. Iris

library(ggplot2)  
qplot(Sepal.Length, Sepal.Width, data=iris)

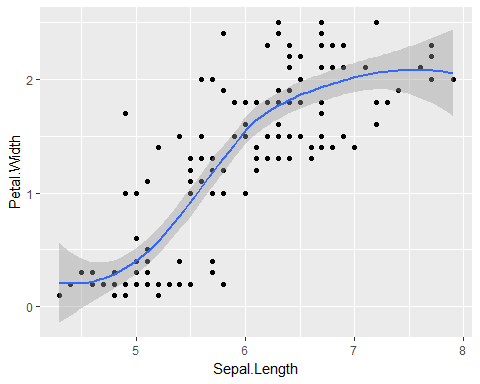


qplot(Petal.Length, Petal.Width, data=iris, color=Species)



qplot(Sepal.Length, Petal.Width, data=iris, geom= c("point", "smooth"))

## `geom\_smooth()` using method = 'loess'

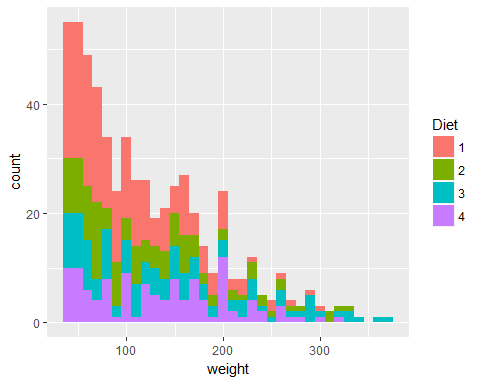


## Using histograms

We have to specify only one variable for plotting a histogram in ggplot.

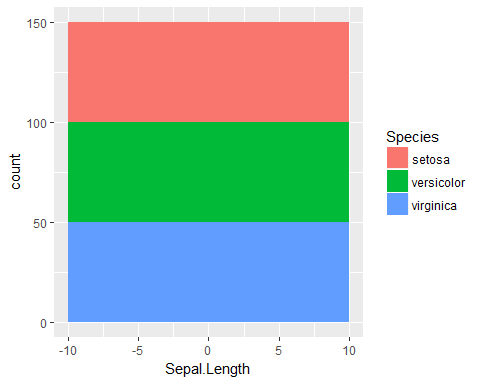
* ChickWeight

qplot(weight, data=ChickWeight, fill=Diet, binwidth=10)



* Iris

qplot(Sepal.Length, data=iris, fill=Species, binwidth=20)



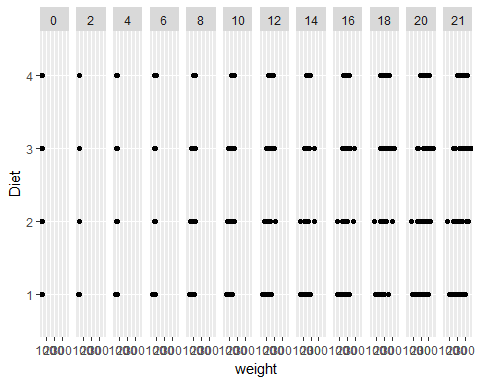
### Facets

We create separate plots using facets.

* ChickWeight

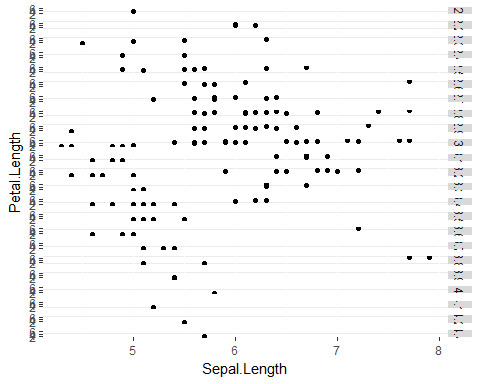
qplot(weight, Diet, data=ChickWeight, facets= .~Time, binwidth=30)

## Warning: Ignoring unknown parameters: binwidth



* Iris

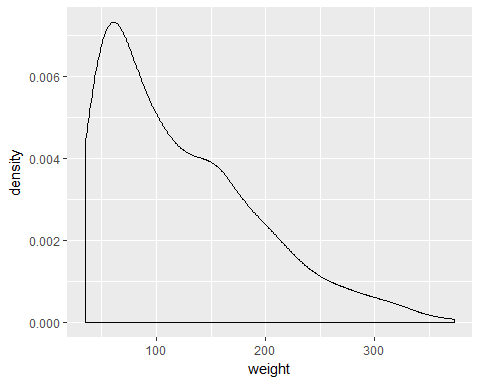
qplot(Sepal.Length, Petal.Length, data=iris, facets=Sepal.Width~.)



## Density smooth

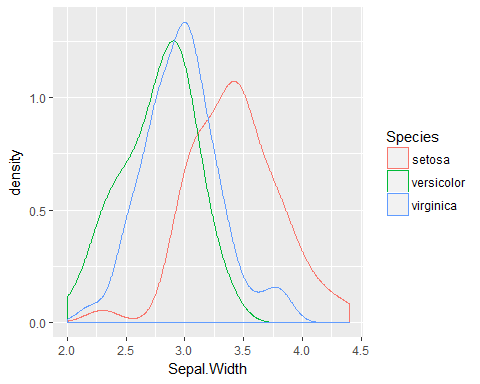
* ChickWeight

qplot(weight, data=ChickWeight, geom="density", color=Time)



* Iris

qplot(Sepal.Width, data=iris, geom="density", color=Species)



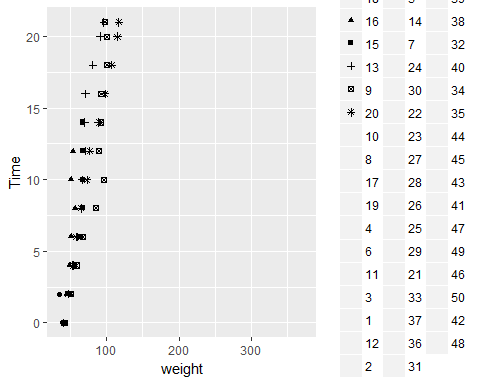
## Scatter plots with shape

* Scatter plot for ChickWeight.

qplot(weight,Time, data=ChickWeight, shape=Chick)

## Warning: The shape palette can deal with a maximum of 6 discrete values  
## because more than 6 becomes difficult to discriminate; you have  
## 50. Consider specifying shapes manually if you must have them.

## Warning: Removed 525 rows containing missing values (geom\_point).



In order to find out the missing values, we use the following:

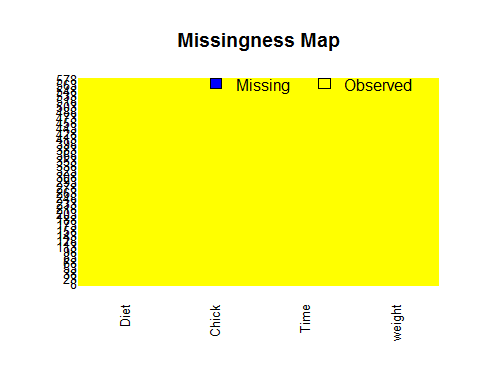
library(Amelia)

## Loading required package: Rcpp

## ##   
## ## Amelia II: Multiple Imputation  
## ## (Version 1.7.4, built: 2015-12-05)  
## ## Copyright (C) 2005-2017 James Honaker, Gary King and Matthew Blackwell  
## ## Refer to http://gking.harvard.edu/amelia/ for more information  
## ##

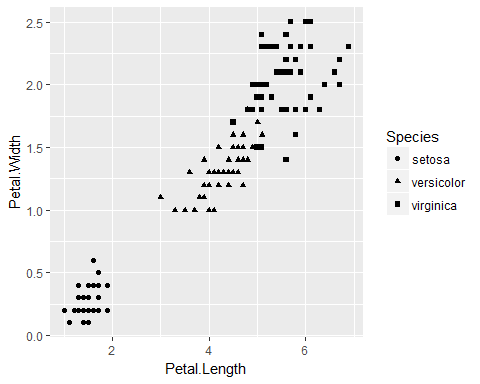
library(mlbench)  
missmap(ChickWeight, legend=TRUE, col=c("blue","yellow"))

## Warning in if (class(obj) == "amelia") {: the condition has length > 1 and  
## only the first element will be used



* Scatter plot for the dataset iris.

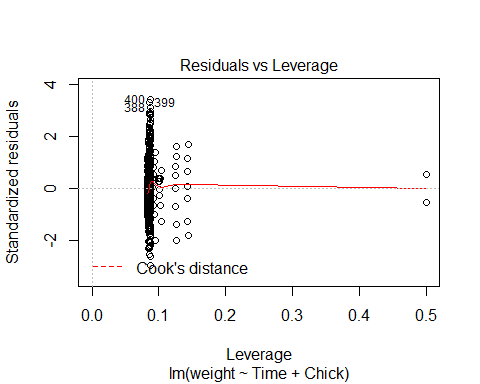
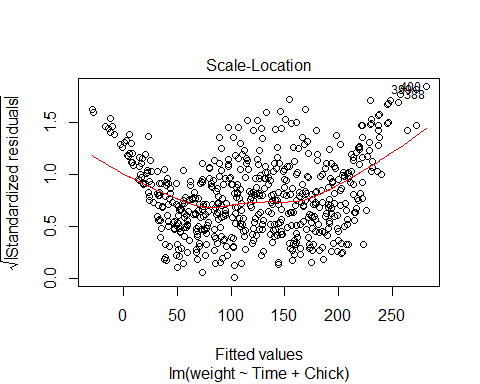
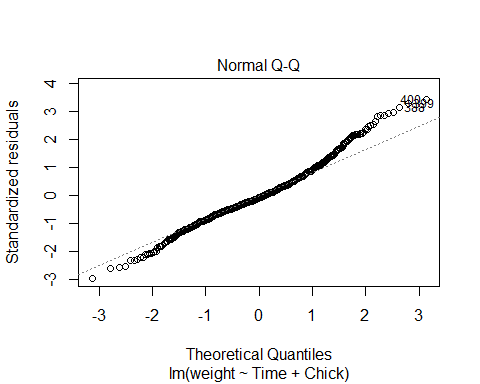
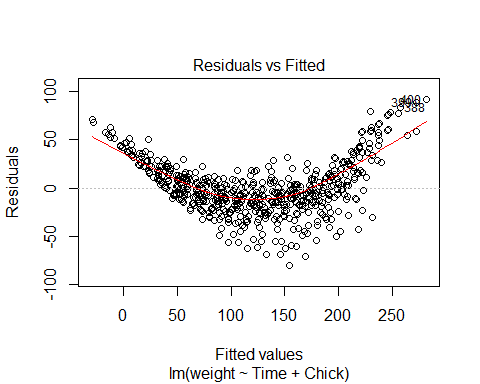
qplot(Petal.Length, Petal.Width, data=iris, shape=Species)



## Fitting linear models

1. ChickWeight

Data1fit <- lm(weight ~ Time + Chick, data=ChickWeight)  
plot(Data1fit)



1. Iris

Data2fit <- lm(Sepal.Length + Petal.Length ~ Sepal.Width + Petal.Width, data=iris)  
plot(Data2fit)

