# Univariate Graphical Exploratory Data Analysis

## 1. Measures of Central Tendency

Before embarking on developing statistical models and generating predictions, it is essential to understand our data. This is typically done using conventional numerical and graphical methods.

## Example   
# ---  
# We will be using the hills dataset in this section,   
# this dataset contains information on hill climbs made by various athletes  
# ---  
# OUR CODE GOES BELOW  
#   
  
# Printing the first six rows of the dataset   
# ---  
#   
library(MASS)  
head(hills)

## dist climb time  
## Greenmantle 2.5 650 16.083  
## Carnethy 6.0 2500 48.350  
## Craig Dunain 6.0 900 33.650  
## Ben Rha 7.5 800 45.600  
## Ben Lomond 8.0 3070 62.267  
## Goatfell 8.0 2866 73.217

#### Mean Code Example 1.1

## Example   
# ---  
# Question: Find the mean of the distance covered by the athletes   
# and assigning the mean to the variable athletes.dist.mean  
# ---  
# OUR CODE GOES BELOW  
#   
  
athletes.dist.mean <- mean(hills$dist)  
  
# Printing out  
# ---  
#  
athletes.dist.mean

## [1] 7.528571

#### Median Code Example 1.2

## Example   
# ---  
# Question: Find the median which is the middle most value of the distance covered dist  
# ---  
# OUR CODE GOES BELOW  
#   
athletes.dist.median <- median(hills$dist)  
  
# Printing out athletes.dist.median  
# ---  
#   
athletes.dist.median

## [1] 6

#### Mode Code Example 1.3

## Example   
# ---  
# Question: Find the mode which is the value that has highest number of occurrences in a set of data.   
# ---  
# OUR CODE GOES BELOW  
#   
  
# Unfotunately, R does not have a standard in-built function to calculate mode so we have to build one  
# We create the mode function that will perform our mode operation for us  
# ---  
#   
getmode <- function(v) {  
 uniqv <- unique(v)  
 uniqv[which.max(tabulate(match(v, uniqv)))]  
}  
  
# Calculating the mode using out getmode() function  
# ---  
#  
athletes.dist.mode <- getmode(hills$dist)  
  
# Then printing out athletes.dist.mode   
# ---  
# OUR CODE GOES BELOW  
#   
athletes.dist.mode

## [1] 6

## Challenge   
# ---  
# Question: Find the mean, median, mode of the total evening calls given the following dataset   
# ---  
# Dataset url = http://bit.ly/CustomerSignatureforChurnAnalysis  
# ---  
# OUR CODE GOES BELOW   
  
# Previewing the first 6 rows of this dataset  
# ---  
#   
library("data.table")  
df <- fread('http://bit.ly/CustomerSignatureforChurnAnalysis')  
View(df)  
# Finding the mean  
# ---  
#   
df.mean <- mean(df$total\_day\_calls)  
df.mean

## [1] 100.2666

# Finding the median   
# ---  
#  
df.median <- median(df$total\_day\_calls)  
  
# Finding the mode  
# ---  
#   
getmode <- function(v) {  
 uniqv <- unique(v)  
 uniqv[which.max(tabulate(match(v, uniqv)))]  
}  
df.mode <- getmode(df$total\_day\_calls)

df.mean

## [1] 100.2666

df.median

## [1] 101

df.mode

## [1] 105

## 2. Measures of Dispersion

#### Mininum Code Example 1.4

## Example   
# ---  
# Question: Find the minimum element of the distance using the min() function  
# ---  
# OUR CODE GOES BELOW  
#   
athletes.dist.min <- min(hills$dist)  
  
# And then printing athletes.dist.min to show the minimum element  
#   
athletes.dist.min

## [1] 2

#### Maximum Code Example 1.5

## Example  
# ---  
# Question: Find the maximum element of the distance using the function max()   
# ---  
# OUR CODE GOES BELOW   
#   
athletes.dist.max <- max(hills$dist)  
  
# Then printing out the variable athletes.dist.max to show that maximum element  
# ---  
# OUR CODE GOES BELOW  
#

#### Range Code Example 1.6

## Example   
# ---  
# Find the maximum element of the distance using the function range() as shown below  
# ---  
#   
athletes.dist.range <- range(hills$dist)  
  
# Printing out the variable athletes.dist.range to show the range   
# ---  
#  
athletes.dist.range

## [1] 2 28

#### Quantile Code Example 1.7

## Example   
# ---  
# Question: Get the first and the third quartile together with the range   
# and the median using the quantile() function  
# ---  
# OUR CODE GOES BELOW  
#   
athletes.dist.quantile <- quantile(hills$dist)  
  
# Printing out the variable athletes.dist.quantile to show the range   
# ---  
# OUR CODE GOES BELOW  
#   
athletes.dist.quantile

## 0% 25% 50% 75% 100%   
## 2.0 4.5 6.0 8.0 28.0

#### Variance Code Example 1.8

## Example   
# ---  
# Question: Find the variance of the distance using the var() function as shown below  
# ---  
# OUR CODE GOES BELOW   
#   
  
athletes.dist.variance <- var(hills$dist)  
  
# Printing out the the variable athletes.dist.variance to show the variance   
#   
athletes.dist.variance

## [1] 30.51387

The variance is a numerical measure of how the data values is dispersed around the mean.

#### Standard Deviation Code Example 1.9

## Example   
# ---  
# Question: Find the standard deviation of vector t using the sd() function   
# ---  
# OUR CODE GOES BELOW   
#   
athletes.dist.sd <- sd(hills$dist)  
  
# Printing out the variable athletes.dist.sd to show the variance   
# ---  
#  
athletes.dist.sd

## [1] 5.523936

# Challenge   
# ---  
# Question: Find the minimum, maximum, range, quantile, variance   
# and standard deviation for total day calls using the given dataset  
# ---  
# Dataset url = http://bit.ly/CustomerSignatureforChurnAnalysis  
# ---  
# OUR CODE GOES BELOW  
#   
  
  
# Find the minimum of total day calls  
# ---  
# OUR CODE GOES BELOW  
#   
df.min <- min(df$total\_day\_calls)  
df.min

## [1] 0

# Find the maximum i.e. max() total day calls  
# ---  
# OUR CODE GOES BELOW  
#   
df.max <- max(df$total\_day\_calls)  
df.max

## [1] 165

# Find the range i.e. range() of total day calls  
# ---  
# OUR CODE GOES BELOW  
#   
  
df.range <- range(df$total\_day\_calls)  
df.range

## [1] 0 165

# Find the quantile of total day calls  
# ---  
# OUR CODE GOES BELOW  
#   
df.quantile <- quantile(df$total\_day\_calls)  
df.quantile

## 0% 25% 50% 75% 100%   
## 0 87 101 114 165

# Find the variance of total day calls  
# ---  
# OUR CODE GOES BELOW  
#   
df.variance <- var(df$total\_day\_calls)  
df.variance

## [1] 397.8691

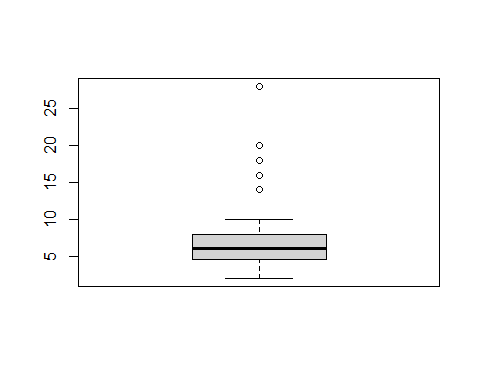
# Find the standard deviation of total day calls  
# ---  
# OUR CODE GOES BELOW  
#   
df.std <- sd(df$total\_day\_calls)  
df.std

## [1] 19.94666

## 3. Univariate Graphical

#### Box Plots Code Example 3.1

## Example   
# ---  
# Question: Lets create a boxplot graph for the distance using the boxplot() function  
# ---  
# OUR CODE GOES BELOW  
#   
  
boxplot(hills$dist)



The box plot of an observation variable is a graphical representation based on its quartiles, as well as its smallest and largest values. It attempts to provide a visual shape of the data distribution.

#### Bar Graph Code Example 3.2

A bar graph of a qualitative data sample consists of vertical parallel bars that shows the frequency distribution graphically.

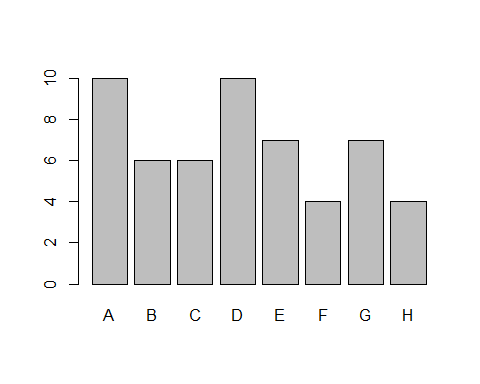
## Example   
# ---  
# Create a frequency distribution of the School variable  
# ---  
# Dataset Info: For this example, we will use an R built-in database named painters.   
# ---  
# OUR CODE GOES BELOW  
#   
  
# Previewing the first six rows of the painters dataset  
# ---  
# OUR CODE GOES BELOW  
#   
head(painters)

## Composition Drawing Colour Expression School  
## Da Udine 10 8 16 3 A  
## Da Vinci 15 16 4 14 A  
## Del Piombo 8 13 16 7 A  
## Del Sarto 12 16 9 8 A  
## Fr. Penni 0 15 8 0 A  
## Guilio Romano 15 16 4 14 A

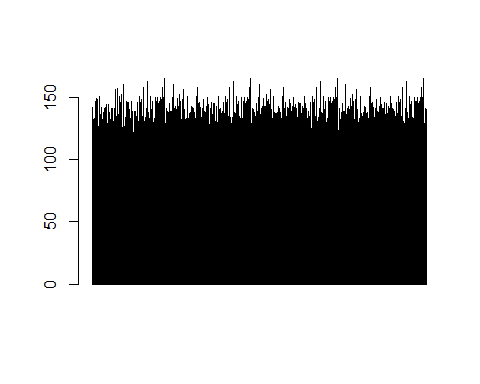
# Fetching the school column  
# ---  
#   
school <- painters$School  
  
# Applying the table() function will compute the frequency distribution of the School variable  
# ---  
#   
school\_frequency <- table(school)  
  
# Printing school\_frequency below  
# ---  
#  
school\_frequency

## school  
## A B C D E F G H   
## 10 6 6 10 7 4 7 4

# Then applying the barplot function to produce its bar graph  
# ---  
#   
barplot(school\_frequency)



## Challenge  
# ---  
# Question: Create a bar graph of the total day calls in the customer signature dataset  
# ---  
# Dataset url = http://bit.ly/CustomerSignatureforChurnAnalysis  
# ---  
# OUR CODE GOES BELOW  
#  
  
barplot(df$total\_day\_calls)

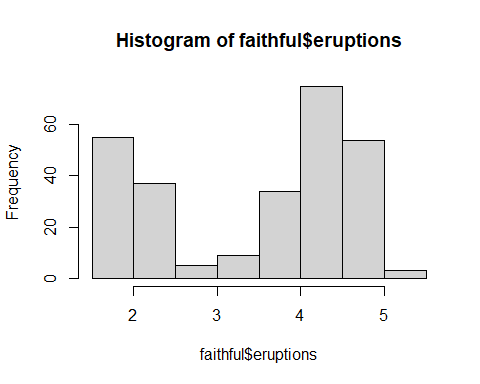


#### Histogram Code Example 3.3

A histogram shows the frequency distribution of a quantitative variable. The area of each bar is equal to the frequency of items found in each class.

## Example  
# ---  
# Create a histogram using the faithful dataset   
# ---   
# Hint: we will use an R built-in data frame called faithful   
# ---  
# OUR CODE GOES BELOW  
#   
  
# Preview the first six rows of the faithful dataset  
# ---  
# OUR CODE GOES BELOW  
#

# Then applying the hist() function to produce the histogram of the eruptions variable   
# ---  
#   
hist(faithful$eruptions)



## Challenge   
# ---  
# Question: Create a histogram of the total day minutes in the customer signature dataset   
# ---  
# Dataset url = http://bit.ly/CustomerSignatureforChurnAnalysis  
# ---  
# OUR CODE GOES BELOW  
#   
df.hist <- hist(df$total\_day\_minutes)

