# 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 # Importing the data.table  
# ---  
#   
library("data.table")  
  
# Reading our dataset  
# ---  
#   
customer\_signature <- fread('http://bit.ly/CustomerSignatureforChurnAnalysis')  
  
# Previewing the first 6 records of the dataset  
# ---  
#   
head(customer\_signature)

## recordID state account\_length area\_code international\_plan voice\_mail\_plan  
## 1: 1 HI 101 510 no no  
## 2: 2 MT 137 510 no no  
## 3: 3 OH 103 408 no yes  
## 4: 4 NM 99 415 no no  
## 5: 5 SC 108 415 no no  
## 6: 6 IA 117 415 no no  
## number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge  
## 1: 0 70.9 123 12.05  
## 2: 0 223.6 86 38.01  
## 3: 29 294.7 95 50.10  
## 4: 0 216.8 123 36.86  
## 5: 0 197.4 78 33.56  
## 6: 0 226.5 85 38.51  
## total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes  
## 1: 211.9 73 18.01 236.0  
## 2: 244.8 139 20.81 94.2  
## 3: 237.3 105 20.17 300.3  
## 4: 126.4 88 10.74 220.6  
## 5: 124.0 101 10.54 204.5  
## 6: 141.6 68 12.04 223.0  
## total\_night\_calls total\_night\_charge total\_intl\_minutes total\_intl\_calls  
## 1: 73 10.62 10.6 3  
## 2: 81 4.24 9.5 7  
## 3: 127 13.51 13.7 6  
## 4: 82 9.93 15.7 2  
## 5: 107 9.20 7.7 4  
## 6: 90 10.04 6.9 5  
## total\_intl\_charge number\_customer\_service\_calls churn customer\_id  
## 1: 2.86 3 no 23383607  
## 2: 2.57 0 no 22550362  
## 3: 3.70 1 no 59063354  
## 4: 4.24 1 no 25464504  
## 5: 2.08 2 no 691824  
## 6: 1.86 1 no 24456543

# Previewing the first 6 rows of this dataset  
# ---  
#

# Finding the mean  
# ---  
#   
customer.accountlength <- mean(customer\_signature$account\_length)  
  
customer.accountlength

## [1] 100.6766

# Finding the median   
# ---  
#  
customer.acclength.median <- median(customer\_signature$account\_length)  
customer.acclength.median

## [1] 100

# Finding the mode  
# ---  
#   
  
# 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  
# ---  
#  
customer.acclength.mode <- getmode(customer\_signature$account\_length)  
  
# Then printing out athletes.dist.mode   
# ---  
# OUR CODE GOES BELOW  
#   
customer.acclength.mode

## [1] 87

## 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  
#  
athletes.dist.max

## [1] 28

#### 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  
#   
library("data.table")  
  
# Reading our dataset  
# ---  
#   
signature <- fread('http://bit.ly/CustomerSignatureforChurnAnalysis')  
  
# Previewing the first 6 records of the dataset  
# ---  
#   
head(signature)

## recordID state account\_length area\_code international\_plan voice\_mail\_plan  
## 1: 1 HI 101 510 no no  
## 2: 2 MT 137 510 no no  
## 3: 3 OH 103 408 no yes  
## 4: 4 NM 99 415 no no  
## 5: 5 SC 108 415 no no  
## 6: 6 IA 117 415 no no  
## number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge  
## 1: 0 70.9 123 12.05  
## 2: 0 223.6 86 38.01  
## 3: 29 294.7 95 50.10  
## 4: 0 216.8 123 36.86  
## 5: 0 197.4 78 33.56  
## 6: 0 226.5 85 38.51  
## total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes  
## 1: 211.9 73 18.01 236.0  
## 2: 244.8 139 20.81 94.2  
## 3: 237.3 105 20.17 300.3  
## 4: 126.4 88 10.74 220.6  
## 5: 124.0 101 10.54 204.5  
## 6: 141.6 68 12.04 223.0  
## total\_night\_calls total\_night\_charge total\_intl\_minutes total\_intl\_calls  
## 1: 73 10.62 10.6 3  
## 2: 81 4.24 9.5 7  
## 3: 127 13.51 13.7 6  
## 4: 82 9.93 15.7 2  
## 5: 107 9.20 7.7 4  
## 6: 90 10.04 6.9 5  
## total\_intl\_charge number\_customer\_service\_calls churn customer\_id  
## 1: 2.86 3 no 23383607  
## 2: 2.57 0 no 22550362  
## 3: 3.70 1 no 59063354  
## 4: 4.24 1 no 25464504  
## 5: 2.08 2 no 691824  
## 6: 1.86 1 no 24456543

# Find the minimum of total day calls  
# ---  
signature.daycalls.min <- min(signature$total\_day\_calls)  
#   
signature.daycalls.min

## [1] 0

#

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

## [1] 165

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

## [1] 0 165

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

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

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

## [1] 397.8691

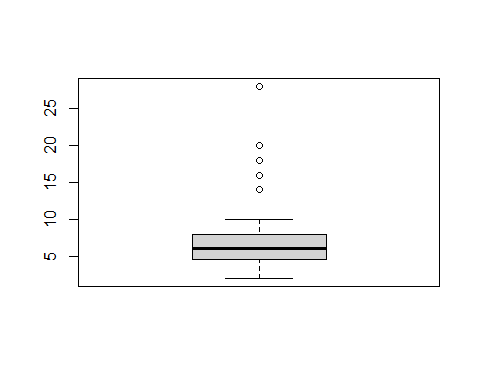
# Find the standard deviation of total day calls  
# ---  
# OUR CODE GOES BELOW  
#   
signature.daycalls.sd <- sd(signature$total\_day\_calls)  
#   
signature.daycalls.sd

## [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.

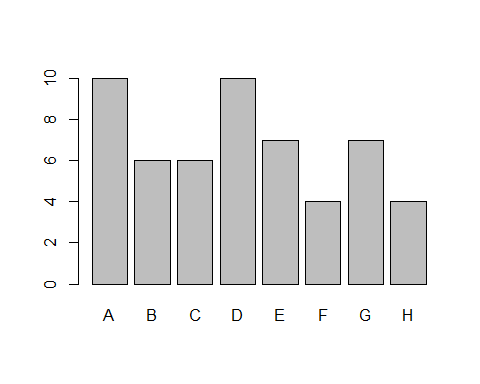
## Example   
# ---  
# Create a frequency distribution of the School variable  
# ---  
# Dataset Info: For this example, we will use an R built-in database named painters.   
# ---  
# OUR CODE GOES BELOW  
#   
  
# Previewing the first six rows of the painters dataset  
# ---  
# OUR CODE GOES BELOW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

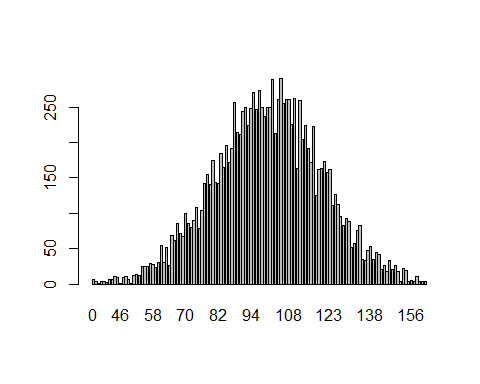
# Then applying the barplot function to produce its bar graph  
# ---  
#   
barplot(school\_frequency)



## 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  
#  
# fetching the total day calls   
total.day.calls <- signature$total\_day\_calls  
total.day.calls\_freq <- table(total.day.calls)  
total.day.calls\_freq

## total.day.calls  
## 0 30 34 35 36 39 40 42 44 45 46 47 48 49 50 51 52 53 54 55   
## 6 4 1 3 4 2 7 6 10 9 1 9 11 7 1 12 14 12 25 25   
## 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75   
## 25 29 27 23 30 54 31 52 26 69 61 86 71 67 100 86 80 90 108 79   
## 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95   
## 104 142 155 141 174 143 142 184 164 195 171 192 256 214 211 243 250 224 248 270   
## 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115   
## 247 274 249 236 250 249 289 213 261 290 255 261 260 225 262 163 259 204 224 192   
## 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135   
## 171 223 125 162 163 173 157 162 111 126 112 96 83 92 88 51 57 75 83 35   
## 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 156 157 158   
## 33 47 53 34 45 42 21 26 18 33 21 26 18 4 22 19 4 5 4 11   
## 160 163 165   
## 4 4 4

# Then applying the barplot function to produce its bar graph  
barplot(total.day.calls\_freq)

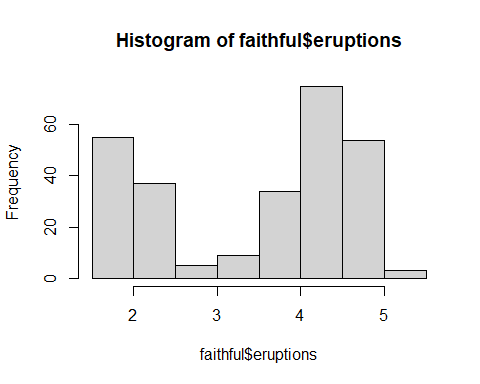
 #### Histogram Code Example 3.3

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  
#   
  
head(faithful)

## eruptions waiting  
## 1 3.600 79  
## 2 1.800 54  
## 3 3.333 74  
## 4 2.283 62  
## 5 4.533 85  
## 6 2.883 55

# 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  
#   
head(signature)

## recordID state account\_length area\_code international\_plan voice\_mail\_plan  
## 1: 1 HI 101 510 no no  
## 2: 2 MT 137 510 no no  
## 3: 3 OH 103 408 no yes  
## 4: 4 NM 99 415 no no  
## 5: 5 SC 108 415 no no  
## 6: 6 IA 117 415 no no  
## number\_vmail\_messages total\_day\_minutes total\_day\_calls total\_day\_charge  
## 1: 0 70.9 123 12.05  
## 2: 0 223.6 86 38.01  
## 3: 29 294.7 95 50.10  
## 4: 0 216.8 123 36.86  
## 5: 0 197.4 78 33.56  
## 6: 0 226.5 85 38.51  
## total\_eve\_minutes total\_eve\_calls total\_eve\_charge total\_night\_minutes  
## 1: 211.9 73 18.01 236.0  
## 2: 244.8 139 20.81 94.2  
## 3: 237.3 105 20.17 300.3  
## 4: 126.4 88 10.74 220.6  
## 5: 124.0 101 10.54 204.5  
## 6: 141.6 68 12.04 223.0  
## total\_night\_calls total\_night\_charge total\_intl\_minutes total\_intl\_calls  
## 1: 73 10.62 10.6 3  
## 2: 81 4.24 9.5 7  
## 3: 127 13.51 13.7 6  
## 4: 82 9.93 15.7 2  
## 5: 107 9.20 7.7 4  
## 6: 90 10.04 6.9 5  
## total\_intl\_charge number\_customer\_service\_calls churn customer\_id  
## 1: 2.86 3 no 23383607  
## 2: 2.57 0 no 22550362  
## 3: 3.70 1 no 59063354  
## 4: 4.24 1 no 25464504  
## 5: 2.08 2 no 691824  
## 6: 1.86 1 no 24456543

# Then applying the hist() function to produce the histogram of the total day minutes variable   
# ---  
#   
hist(signature$total\_day\_minutes)

