```
# Univariate Graphical Exploratory Data Analysis

## 1. Measures of Central Tendency

## 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
```

```
## 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
```

```
## 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
# ---
url_1 <- 'http://bit.ly/CustomerSignatureforChurnAnalysis'
# ---
# OUR CODE GOES BELOW

# Previewing the first 6 rows of this dataset
# ---
#
library(data.table)
churn = fread(url_1)
head(churn)
```

```
##    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
```

```
eve.calls.mean <- mean(churn$total_eve_calls)
eve.calls.median <- median(churn$total_eve_calls)
eve.calls.mode <- getmode(churn$total_eve_calls)

eve.calls.mean
```

```
## [1] 100.1371
```

```
eve.calls.median
```

```
## [1] 100
```

```
eve.calls.mode
```

```
## [1] 105
```

```r
## 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
```

```r
## 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
```

```r
#### 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
```

```
#### 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
#
day.calls.min <- min(churn$total_day_calls)
day.calls.max <- max(churn$total_day_calls)
day.calls.range <- range(churn$total_day_calls)
day.calls.quantile <- quantile(churn$total_day_calls)
day.calls.variance <- var(churn$total_day_calls)
day.calls.std <- sd(churn$total_day_calls)

day.calls.min
```

```
## [1] 0
```

```
day.calls.max
```

```
## [1] 165
```

```
day.calls.range
```

```
## [1]   0 165
```

```
day.calls.quantile
```

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

```
day.calls.variance
```

```
## [1] 397.8691
```

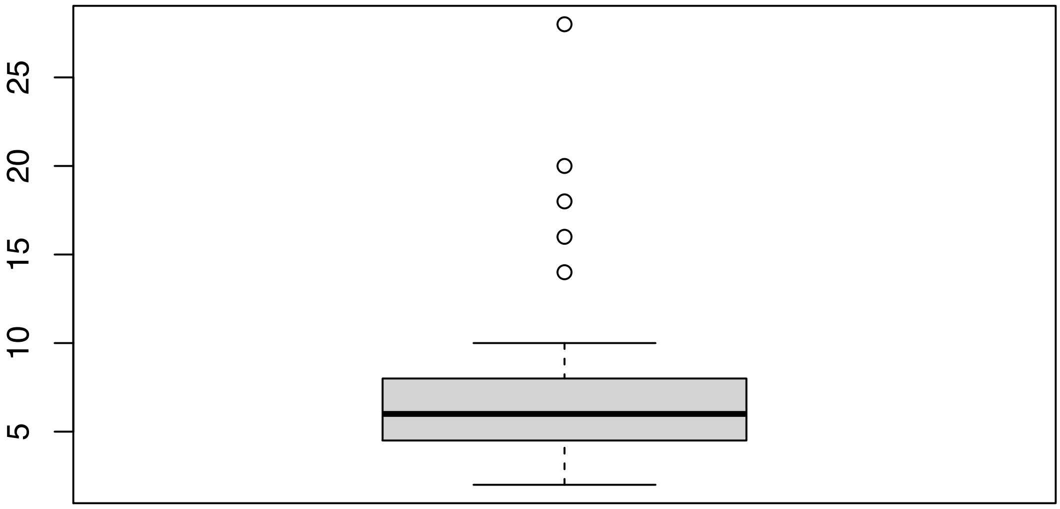
```
day.calls.std
```

```
## [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)
```



```
#### Bar Graph Code Example 3.2
## 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
```

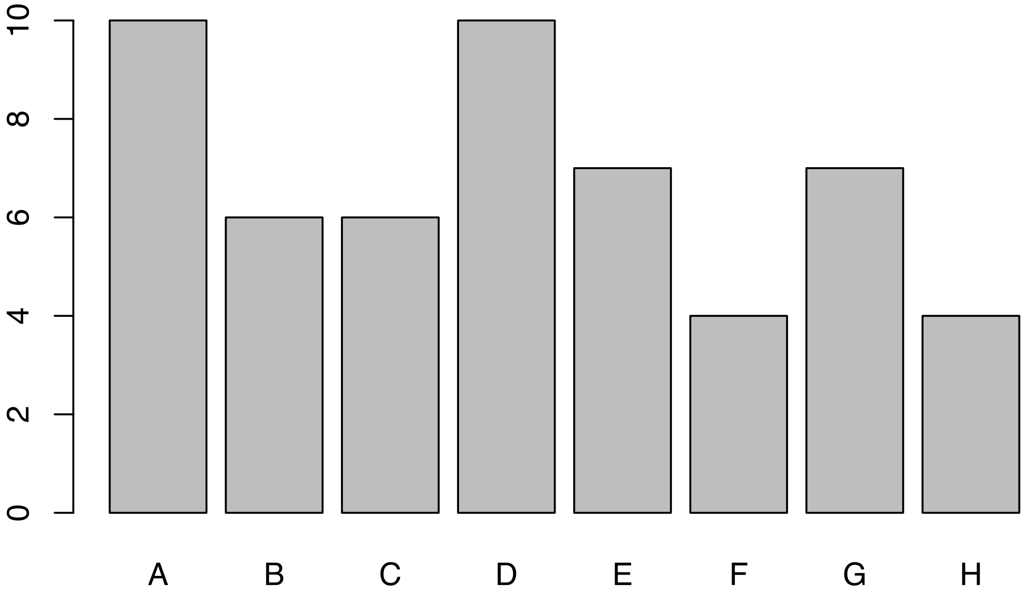
```
# Fetching the school column
# ---
#
school <- painters$School

# Applying the table() function will compute the frequency distribution of the School variable
# ---
#
school_frequency <- table(school)

# Printing school_frequency below
# ---
#
school_frequency
```

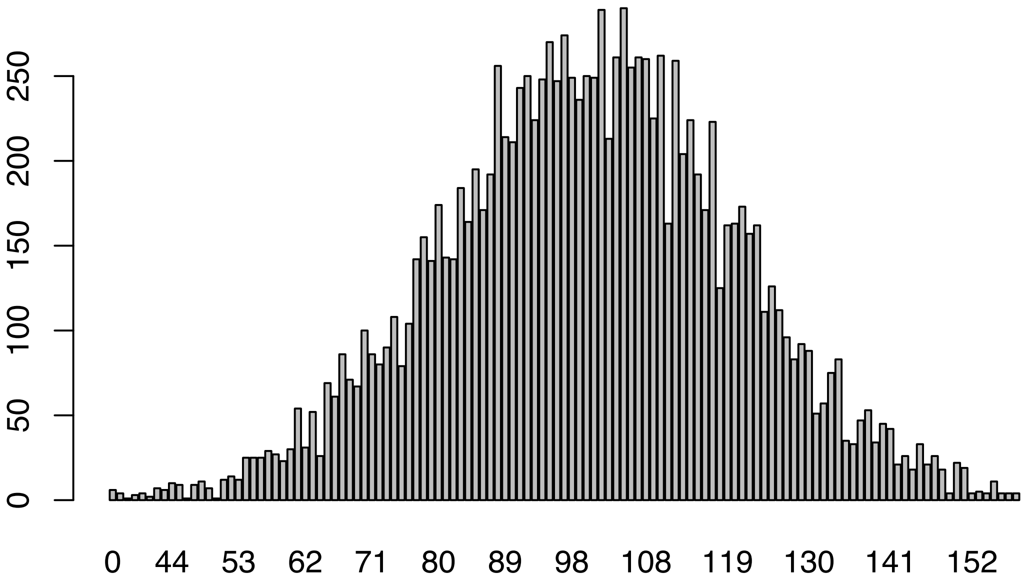
```
## school
##  A  B  C  D  E  F  G  H
## 10  6  6 10  7  4  7  4
```

```
# 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
#
day_calls <- churn$total_day_calls
```

```r
day.calls_frequency <- table(day_calls)
barplot(day.calls_frequency)
```



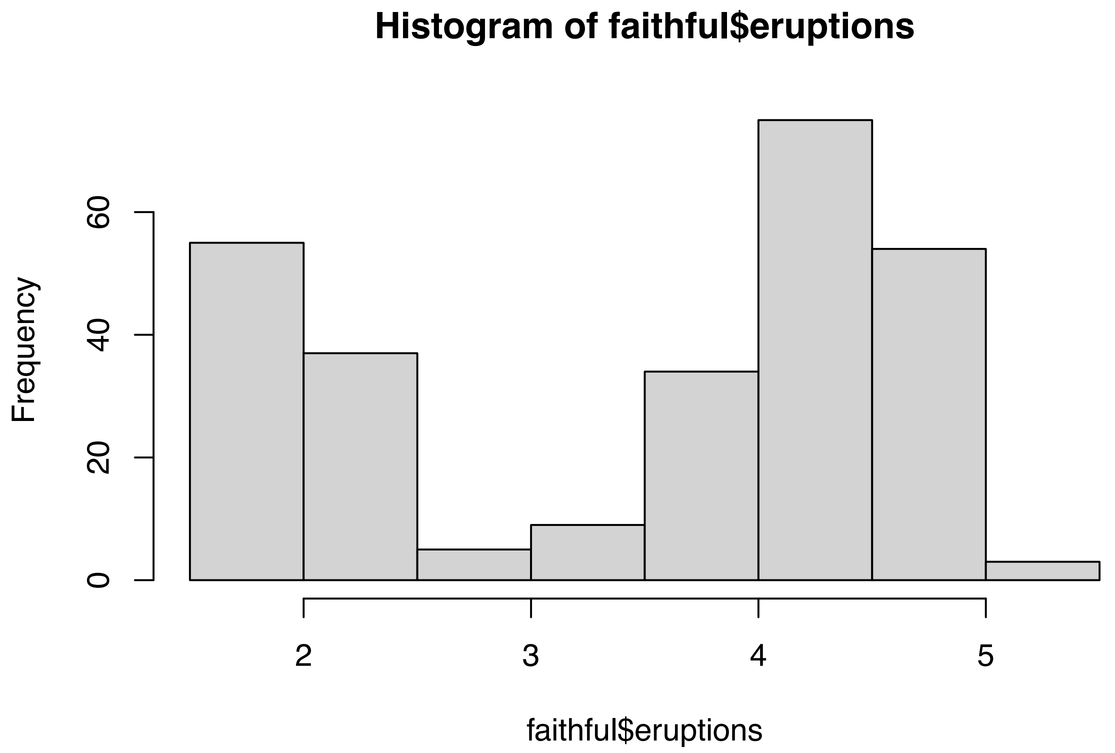
```r
#### Histogram Code Example 3.3

## 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)
```

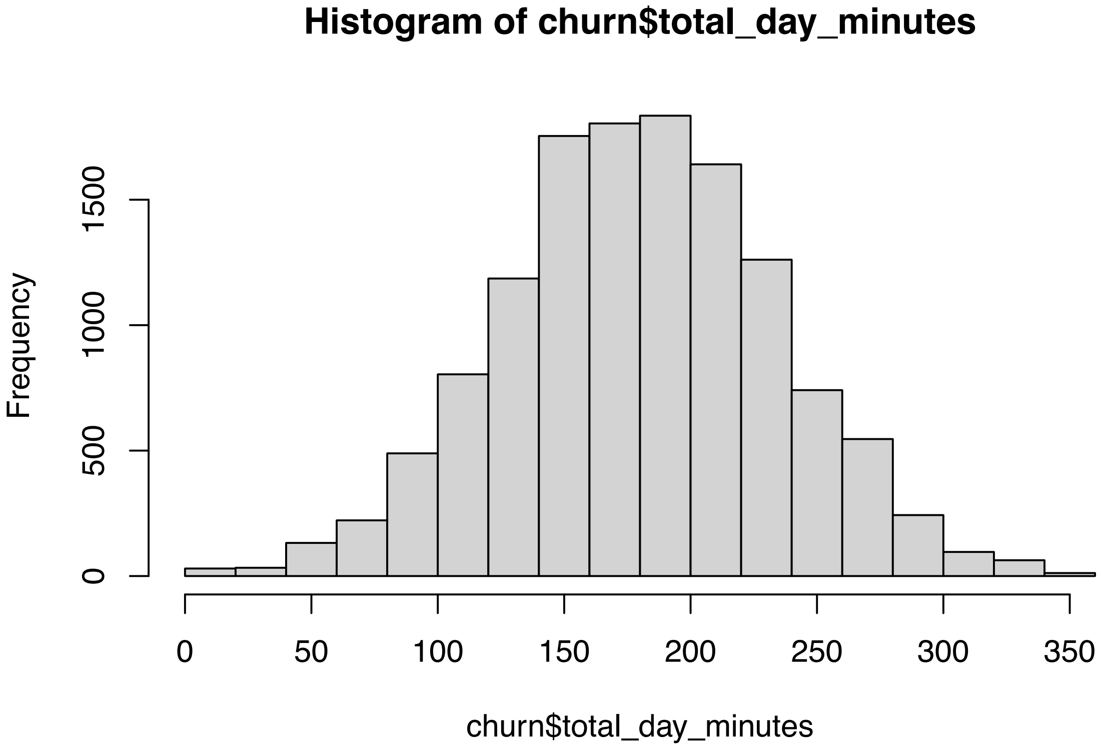
```
##   eruptions waiting
## 1     3.600      79
## 2     1.800      54
## 3     3.333      74
## 4     2.283      62
## 5     4.533      85
## 6     2.883      55
```

```r
# 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

hist(churn$total_day_minutes)
```


Histogram of churn$total_day_minutes
Frequency
0
500
1000
1500
0
50
100
150
200
250
300
350
churn$total_day_minutes