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

library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.0.5
```

```
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
```

```


# Question: Find the mean of the distance covered by the athletes and assigning the mean to the variabl
athletes.dist.mean <- mean(hills$dist)

athletes.dist.mean
```

Mean Code Example 1.1

```
## [1] 7.528571
```

```


# Question: Find the median which is the middle most value of the distance covered dist

athletes.dist.median <- median(hills$dist)

athletes.dist.median
```

Median Code Example 1.2

```
## [1] 6
```

```
# Question: Find the mode which is the value that has highest number of occurrences in a set of data.

# 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)))]
}

athletes.dist.mode <- getmode(hills$dist)

athletes.dist.mode
```

Mode Code Example 1.3

```
## [1] 6
```

```
# Question: Find the mean, median, mode of the total evening calls given the following dataset
library('data.table')
```

```
## Warning: package 'data.table' was built under R version 4.0.4
```

```
url = "http://bit.ly/CustomerSignatureforChurnAnalysis"

# Loading the data
calls <- fread(url)

# Previewing the first 6 rows of this dataset
head(calls)
```

```
##    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
```

```
# Finding the mean
calls$total_eve_calls.mean
```

```
## NULL
```

```
# Finding the median
calls$total_eve_calls.median
```

```
## NULL
```

```
# Finding the mode
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

calls.mode <- getmode(calls$total_eve_calls)

calls.mode
```

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

2. Measures of Dispersion

```

# Question: Find the minimum element of the distance using the min() function

athletes.dist.min <- min(hills$dist)

# And then printing athletes.dist.min to show the minimum element

athletes.dist.min
```

Mininum Code Example 1.4

```
## [1] 2
```

```
# Question: Find the maximum element of the distance using the function max()

athletes.dist.max <- max(hills$dist)

athletes.dist.max
```

Maximum Code Example 1.5

```
## [1] 28
```

```
# Find the maximum element of the distance using the function range() as shown below

athletes.dist.range <- range(hills$dist)

athletes.dist.range
```

Range Code Example 1.6

```
## [1]  2 28
```

```
# Question: Get the first and the third quartile together with the range and the median using the quant

athletes.dist.quantile <- quantile(hills$dist)

athletes.dist.quantile
```

Quantile Code Example 1.7

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

```
# Question: Find the variance of the distance using the var() function as shown below

athletes.dist.variance <- var(hills$dist)

athletes.dist.variance
```

Variance Code Example 1.8

```
## [1] 30.51387
```

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

```r
# Question: Find the standard deviation of vector t using the sd() function

athletes.dist.sd <- sd(hills$dist)

athletes.dist.sd
```

Standard Deviation Code Example 1.9

```
## [1] 5.523936
```

```r
# Question: Find the minimum, maximum, range, quantile, variance and standard deviation for total day c

url = "http://bit.ly/CustomerSignatureforChurnAnalysis"

# Since the data had been loaded earlier, we will continue manipulating the same data

# Find the minimum of total day calls
min(calls$total_day_calls)
```

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

```r
# Find the maximum i.e. max() total day calls
max(calls$total_day_calls)
```

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

```r
# Find the range i.e. range() of total day calls
range(calls$total_day_calls)
```

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

```r
# Find the quantile of total day calls
quantile(calls$total_day_calls)
```

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

```r
# Find the variance of total day calls
var(calls$total_day_calls)
```

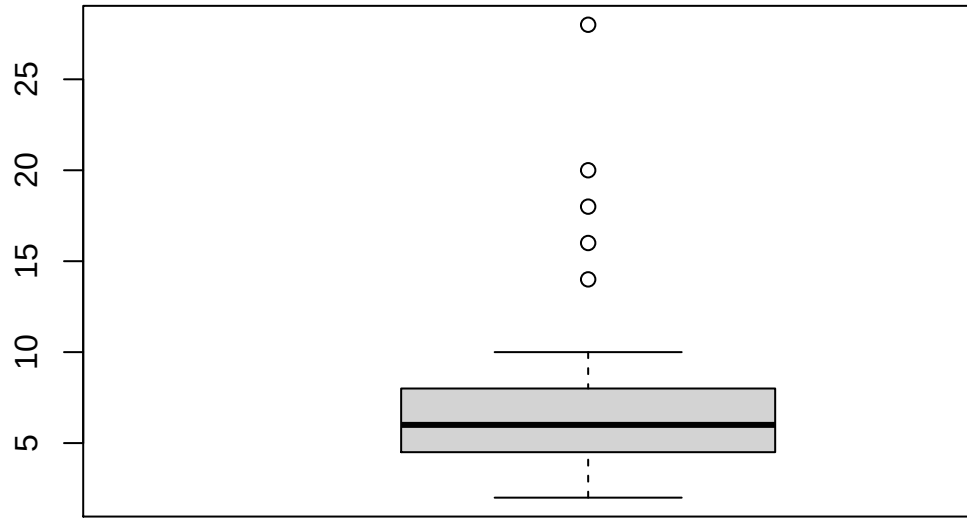
```
## [1] 397.8691
```

```r
# Find the standard deviation of total day calls
sd(calls$total_day_calls)
```

```
## [1] 19.94666
```

3. Univariate Graphical

```r
# Question: Lets create a boxplot graph for the distance using the boxplot() function

boxplot(hills$dist)
```



Box Plots Code Example 3.1

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.

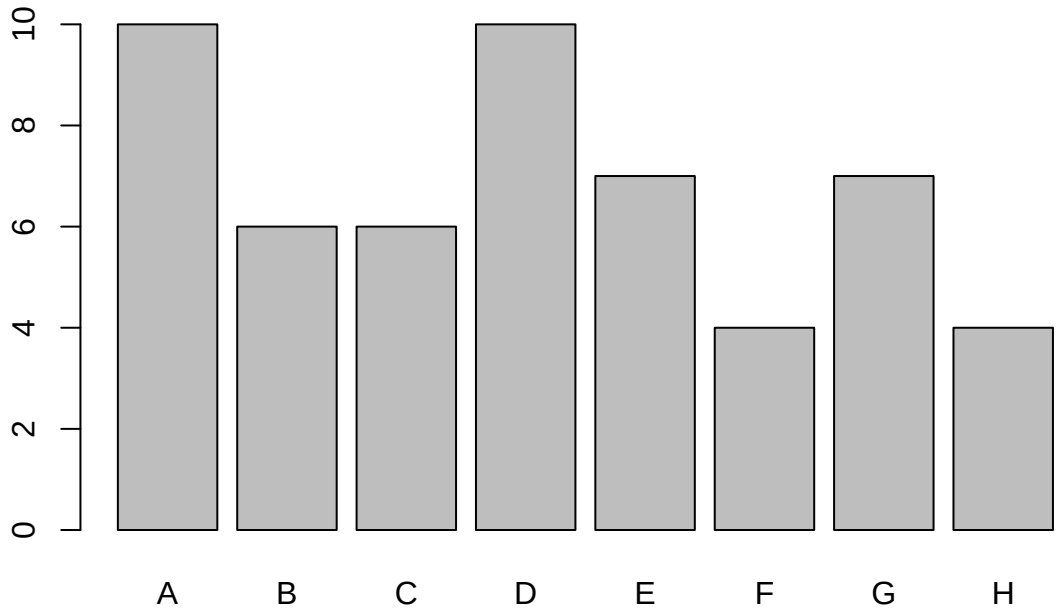
Bar Graph Code Example 3.2 A bar graph of a qualitative data sample consists of vertical parallel bars that shows the frequency distribution graphically.

```r
## Example

# Create a frequency distribution of the School variable

# Dataset Info: For this example, we will use an R built-in database named painters.

# Previewing the first six rows of the painters dataset

head(painters)
```

```
##               Composition Drawing Colour Expression School
## Da Udine               10       8     16          3      A
## Da Vinci               15      16      4         14      A
```

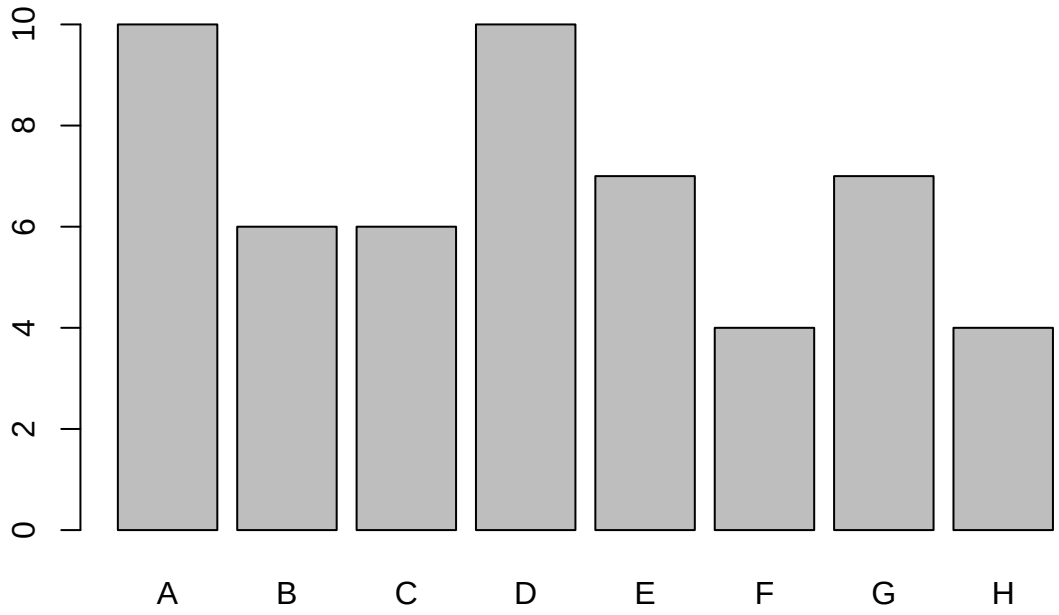
```
## Del Piombo              8      13     16          7      A
## Del Sarto              12      16      9          8      A
## Fr. Penni               0      15      8          0      A
## Guilio Romano          15      16      4         14      A
```

```r
# Creating a vector
school <- painters$School

# Creating a frequency table
freq_table <- table(school)

# Barplot
barplot(freq_table)
```



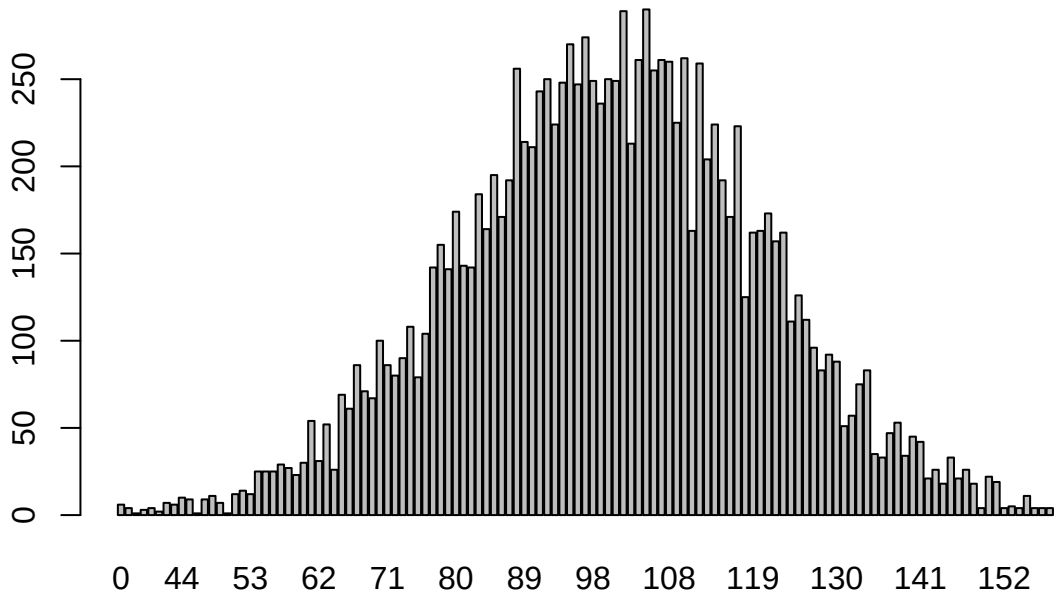

```r
# Fetching the school column

school <- painters$School

# Applying the table() function will compute the frequency distribution of the School variable

school_frequency <- table(school)
```

```r
# 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
# Since this data had been loaded earlier, we will make reference to it without reloading it

day_call_freq <- table(calls$total_day_calls)

barplot(day_call_freq)
```

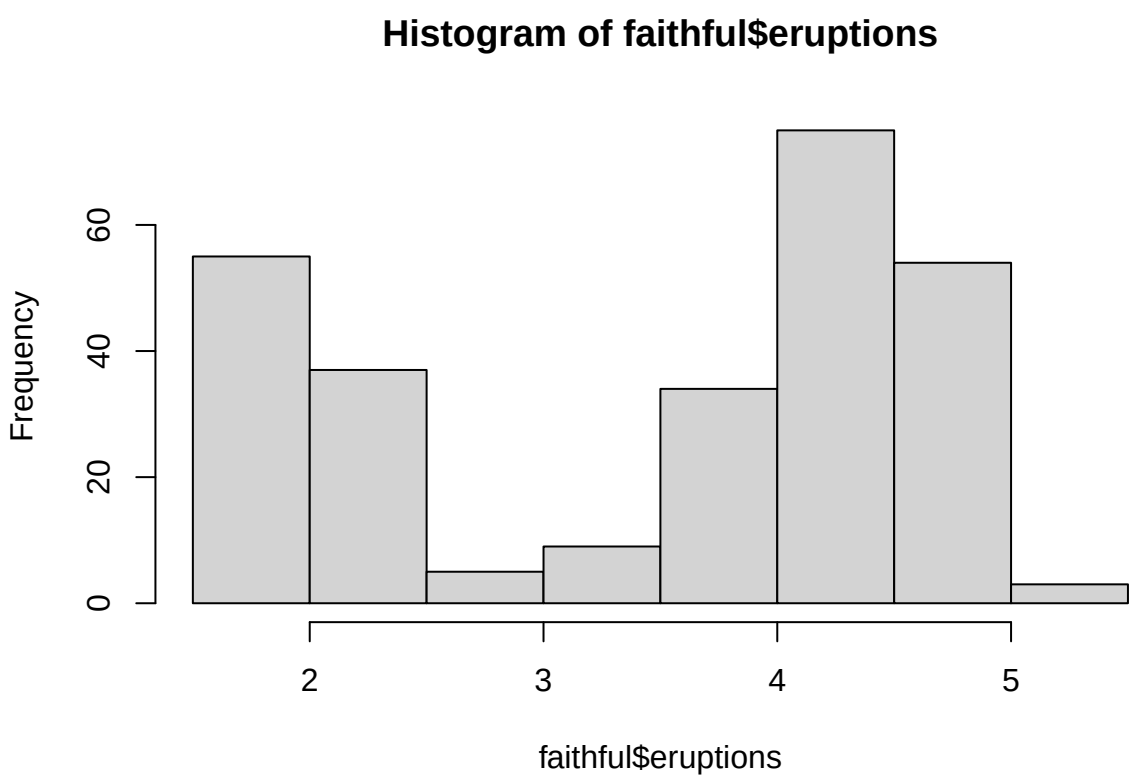


Histogram Code Example 3.3 A histogram shows the frequency distribution of a *quantitative variable.* The area of each bar is equal to the frequency of items found in each class.

```
## Example

# Hint: we will use an R built-in data frame called faithful

# Preview the first six rows of the faithful dataset

head(faithful)
```

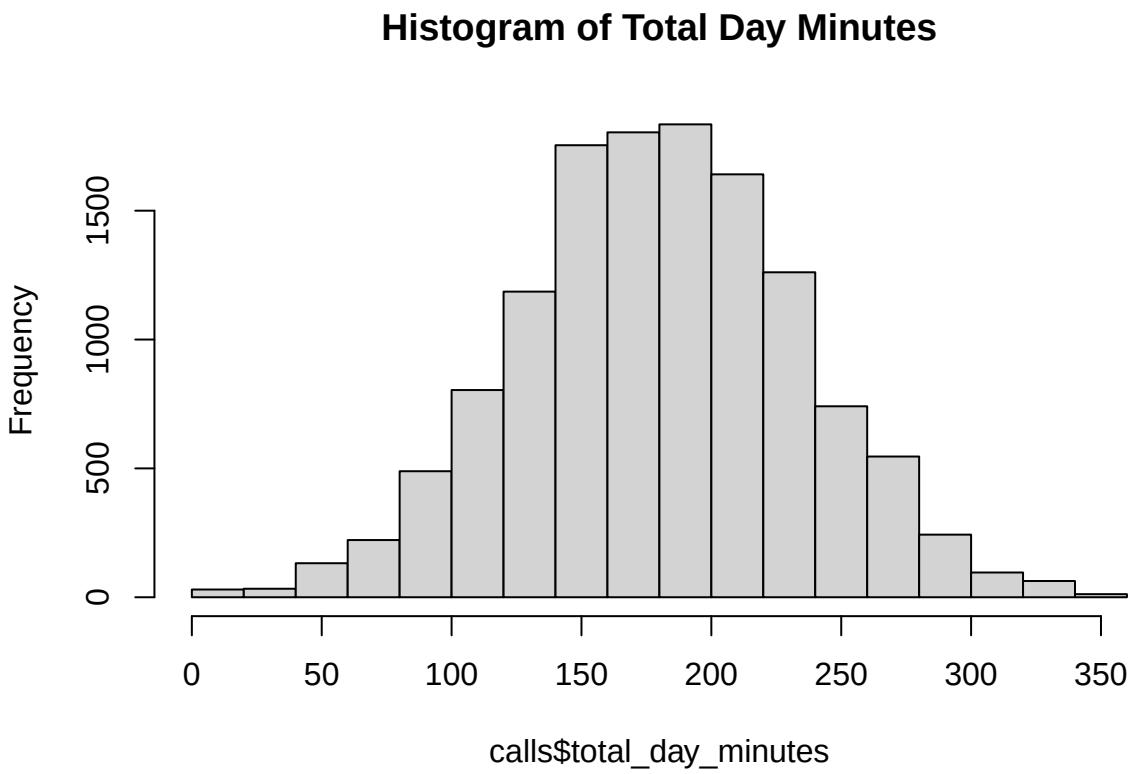
```
##   eruptions waiting
## 1     3.600      79
## 2     1.800      54
## 3     3.333      74
## 4     2.283      62
## 5     4.533      85
## 6     2.883      55
```

```
# 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

# url = "http://bit.ly/CustomerSignatureforChurnAnalysis"

hist(calls$total_day_minutes, main = "Histogram of Total Day Minutes")
```


Histogram of Total Day Minutes
Frequency
0
500
1000
1500
0
50
100
150
200
250
300
350
calls$total_day_minutes