First Delivery Statistics

Eduardo Alarcón & Alfonso Pineda

2022-11-20

1.Introduction:

The first delivery of the final project of Statistics of the degree of Computer Science. In this document, we (Alfonso Pineda and Eduardo Alarcón) will be showing the histogram of the main variable we have chosen for our project, is the **energy** of a song, which is a value assigned by the spoify algorith to try and categorize songs and if they make people more or less energetic. As well, we will have a Box Plot and the statistical Measures on the same block as the histogram, the first one.

On the second block, we will show the **rithm** of a song which we have tested to be the variable that has the most relation with the main variable. We are also going to show a Histogram, a Box Plot and the Statistical Measurements.

Lastly, we will show the Scatter Plot and the Linear Model between the energy and the loudness

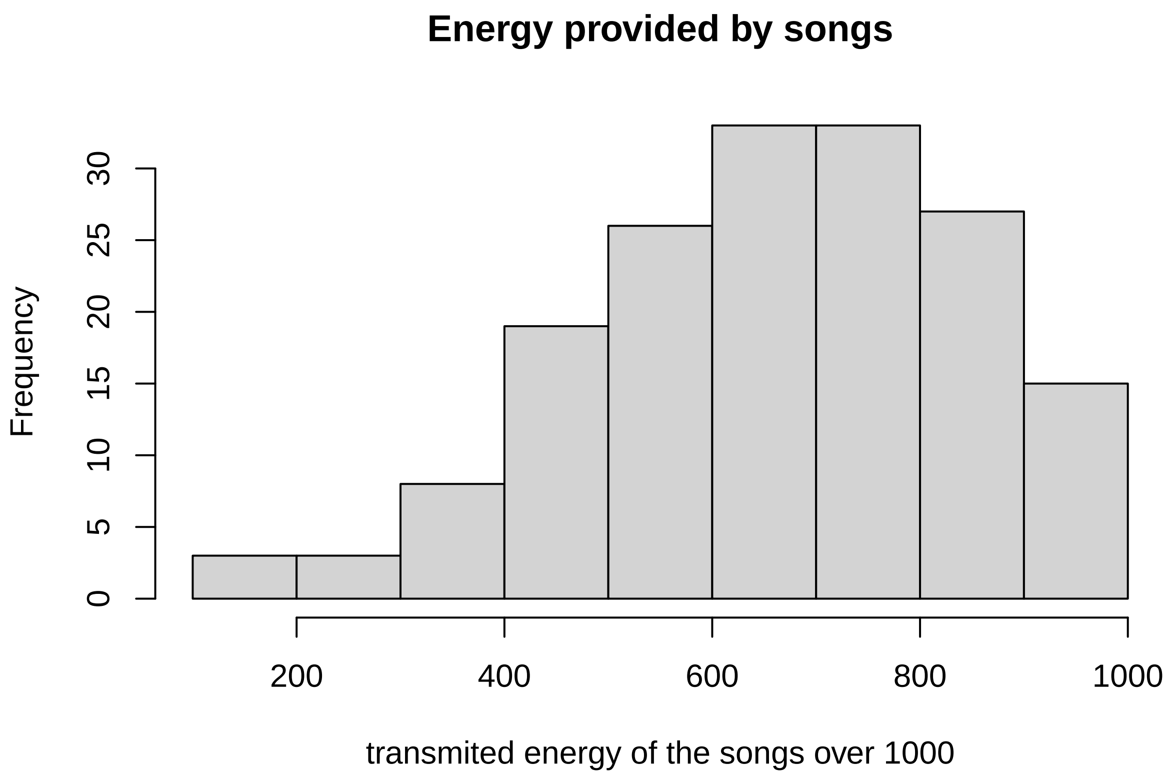
1st Block:

Including the data from the excel: The first thing we need to do is import the data we are going to work with.

```
library(readxl)
SpotifySongs <- read_excel("songstats.xlsx")
View(SpotifySongs)
```

Histogram Then, we need to create the histogram, using R

```
SpotifySongs <- read_excel("songstats.xlsx")
energy <- SpotifySongs$energy
hist(energy, xlab = "transmited energy of the songs over 1000",
     main = "Energy provided by songs")
```



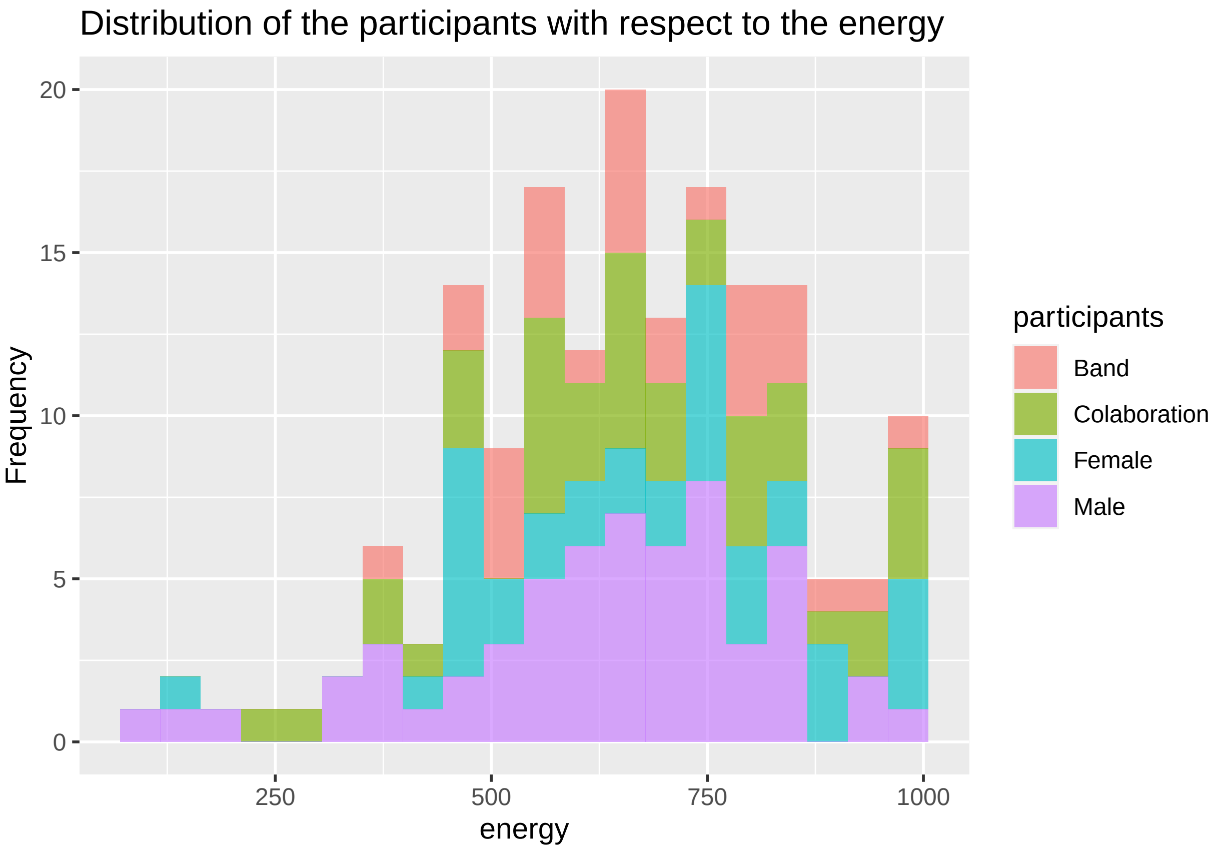
We have also created the histogram differentiating if the artist is a Male, Female, Band or Collaboration between different artists (we used different colors to view them):

```
# {r, fig.height = 4, fig.width = 6}

suppressWarnings(library(ggplot2))
SpotifySongs$participants <- "Male"
SpotifySongs$participants[SpotifySongs$GenderGroup == "F"] <- "Female"
SpotifySongs$participants[SpotifySongs$GenderGroup == "Band"] <- "Band"
SpotifySongs$participants[SpotifySongs$GenderGroup == "Colab"] <- "Colaboration"

qplot(energy, data=SpotifySongs, geom=c("histogram"), fill=participants,
      alpha=I(.65), main="Distribution of the participants with respect to the energy",
      xlab="energy", ylab="Frequency", bins=20)
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
```

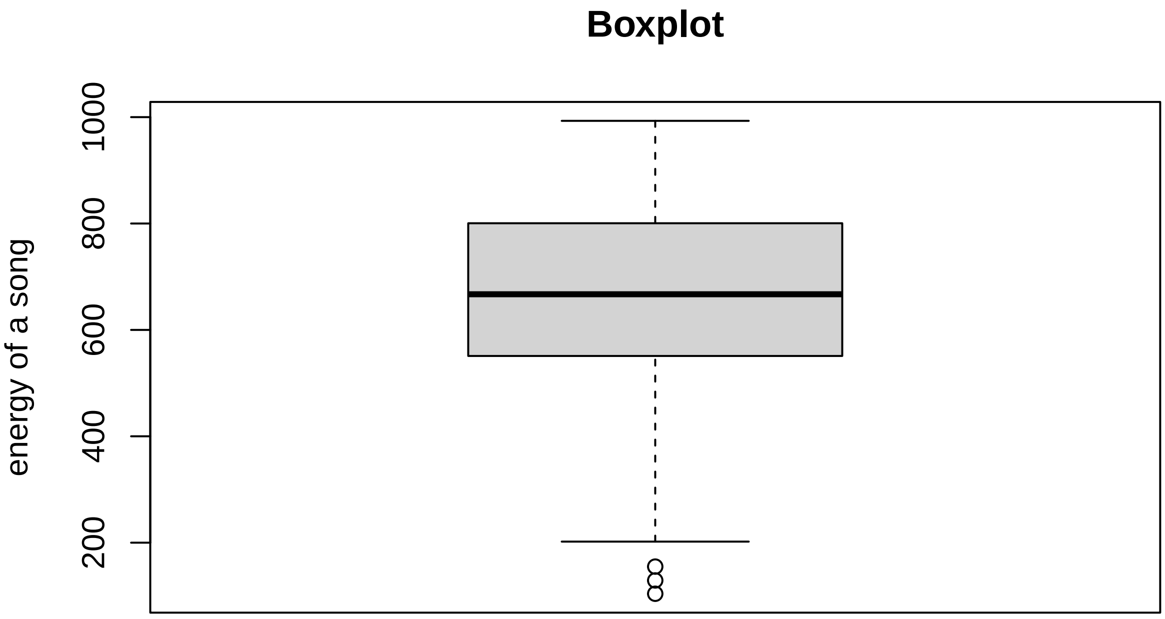


```
#Use bins=num to set the number of intervals
```

Boxplot

Then, we also need to create a Box Plot:

```
#{r, fig.height = 4, fig.width = 5}
boxplot(energy, ylab = "energy of a song",
        main = "Boxplot")
```



As we can see from the Box Plot there are some extreme outliers.

We can see from the Box Plot that the Histogram is not symmetric at all.

Statistical Measures

Now, it's time for us to calculate the statistical measures of the main variable, **energy** These measures are: the *mean*, the *median*, the *mode*, the *percentiles*, the *range*, the *variance*, and the *standarddeviation*.

First, we need to store the variable as Data in R, then, we ask R to describe the variable, which outputs the measures we need, as well as the number of elements there are, in this case N: 167

```
energy<-SpotifySongs$energy
suppressWarnings(library(summarytools))
# Describe the variable energy
descr(energy)
```

```
## Descriptive Statistics
## energy
## N: 167
##
##                       energy
## ----------------- --------
##              Mean   660.92
##           Std.Dev   185.68
##               Min   104.00
##                Q1   551.00
##            Median   667.00
##                Q3   804.00
##               Max   993.00
##               MAD   188.29
##               IQR   249.50
##                CV     0.28
##          Skewness    -0.44
##       SE.Skewness     0.19
##          Kurtosis     0.07
##           N.Valid   167.00
##         Pct.Valid   100.00
```

2nd Block:

On this second part we will test which of the variables we have on our study has the best correlation with the main variable. To asses this, we will use the next block of R:

```r
SpotifySongs <- read_excel("songstats.xlsx")
View(SpotifySongs)
dance <- SpotifySongs$danceability
energy <- SpotifySongs$energy
rithm <- SpotifySongs$rithm
loud <- SpotifySongs$loudness
speech <- SpotifySongs$speechiness
accous <- SpotifySongs$acousticness
live <- SpotifySongs$liveness
valence <- SpotifySongs$valence
tempo <- SpotifySongs$tempo
duration <- SpotifySongs$duration_s
# Choose best second variable
cat("Correlation between loud and Danceability\n")
```

```
## Correlation between loud and Danceability
```

```r
cor(loud, SpotifySongs$danceability)
```

```
## [1] -0.4811005
```

```r
cat("Correlation between loud and energy\n")
```

```
## Correlation between loud and energy
```

```r
cor(loud, SpotifySongs$energy)
```

```
## [1] -0.5156052
```

```r
cat("Correlation between loud and rithm\n")
```

```
## Correlation between loud and rithm
```

```r
cor(loud, SpotifySongs$rithm)
```

```
## [1] -0.4599229
```

```r
cat("Correlation between loud and Loudness\n")
```

```
## Correlation between loud and Loudness
```

```r
cor(loud, SpotifySongs$loudness)
```

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

```r
cat("Correlation between loud and speechiness\n")
```

```
## Correlation between loud and speechiness
```

```r
cor(loud, SpotifySongs$speechiness)
```

```
## [1] -0.2661057
```

```
cat("Correlation between loud and acousticness\n")
```

```
## Correlation between loud and acousticness
```

```
cor(loud, SpotifySongs$acousticness)
```

```
## [1] 0.66467
```

```
cat("Correlation between loud and liveness\n")
```

```
## Correlation between loud and liveness
```

```
cor(loud, SpotifySongs$liveness)
```

```
## [1] -0.07253627
```

```
cat("Correlation between loud and valence\n")
```

```
## Correlation between loud and valence
```

```
cor(loud, SpotifySongs$valence)
```

```
## [1] -0.3773111
```

```
cat("Correlation between loud and tempo\n")
```

```
## Correlation between loud and tempo
```

```
cor(loud, tempo)
```

```
## [1] -0.335712
```

```
cat("Correlation between loud and duration_s\n")
```

```
## Correlation between loud and duration_s
```

```
cor(loud, SpotifySongs$duration_s)
```

```
## [1] 0.2083979
```

```
cor(loud, energy)
```

```
## [1] -0.5156052
```

```
cor(loud, log10(energy))
```

```
## [1] -0.5322306
```

```
cor(energy, rithm)
```

```
## [1] 0.9083597
```

```
#The best correlation found is: Energy & Loudness, with a correlation of 0.8125021
```

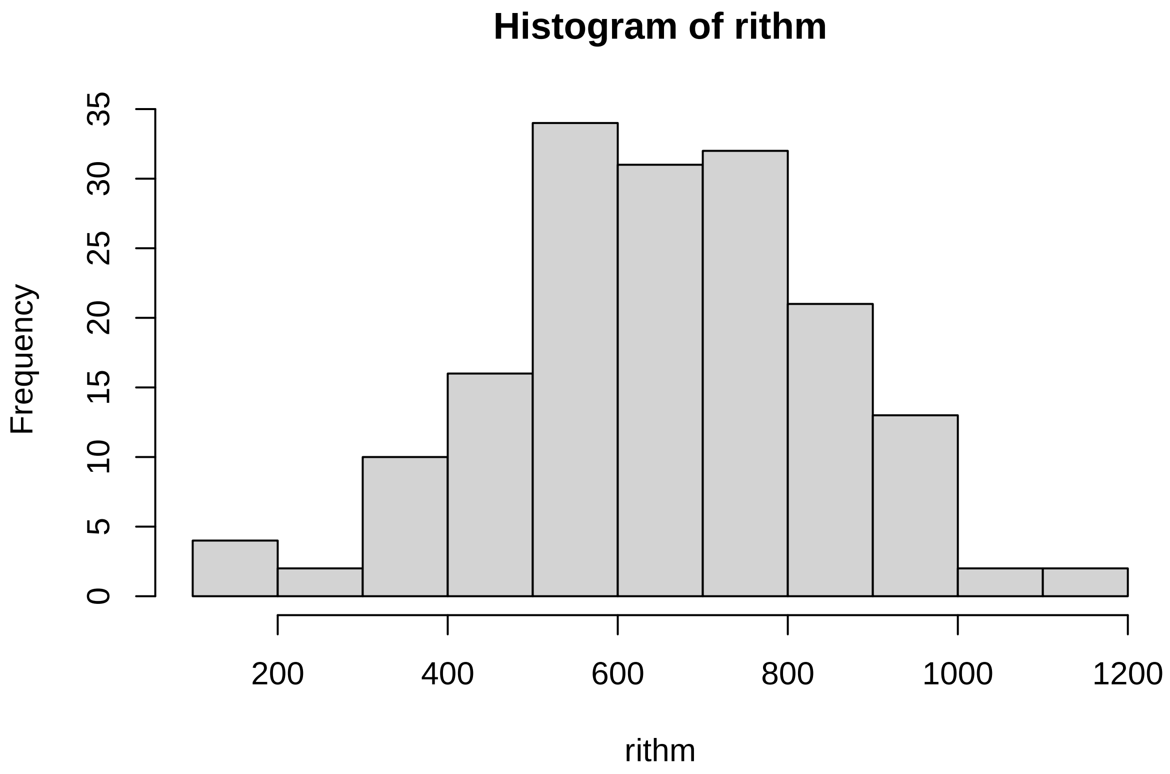
With the previous results, we choose the variable *rithm* These are the statistical variables of the rithm, as well as the Histogram and Box Plot:

```
SpotifySongs <- read_excel("songstats.xlsx")
rithm<-SpotifySongs$rithm
descr(rithm)
```

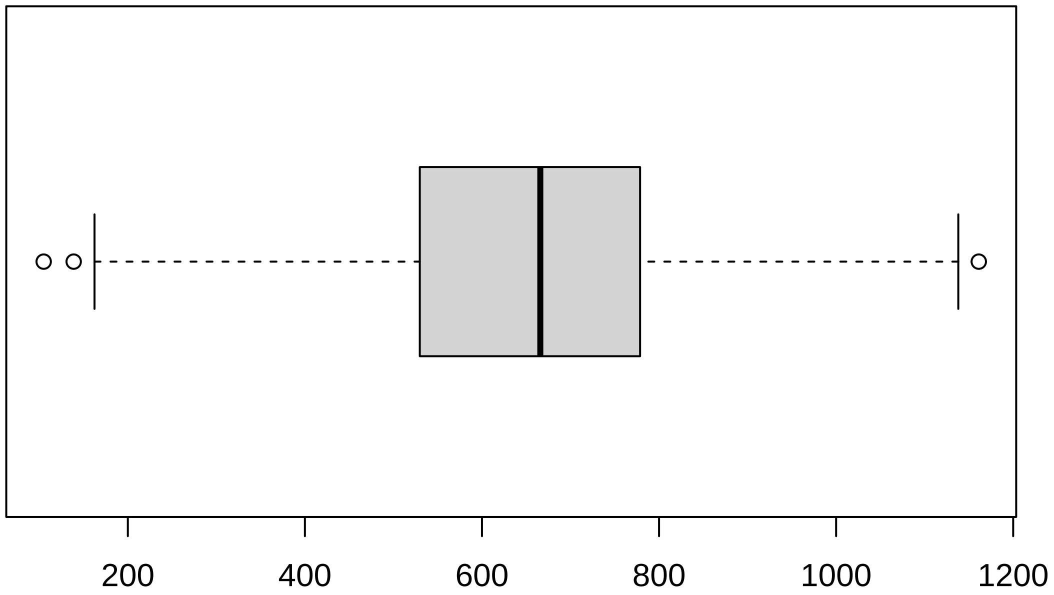
```
## Descriptive Statistics
## rithm
## N: 167
```

```
##
##                           rithm
## ----------------- ---------
##              Mean    654.43
##           Std.Dev    196.49
##               Min    104.93
##                Q1    529.07
##            Median    665.87
##                Q3    778.78
##               Max   1161.16
##               MAD    192.25
##               IQR    248.84
##                CV      0.30
##          Skewness     -0.19
##       SE.Skewness      0.19
##          Kurtosis      0.02
##           N.Valid    167.00
##         Pct.Valid    100.00
```

```r
# Histogram/Box-Plot of the secondary variable
hist(rithm)
```



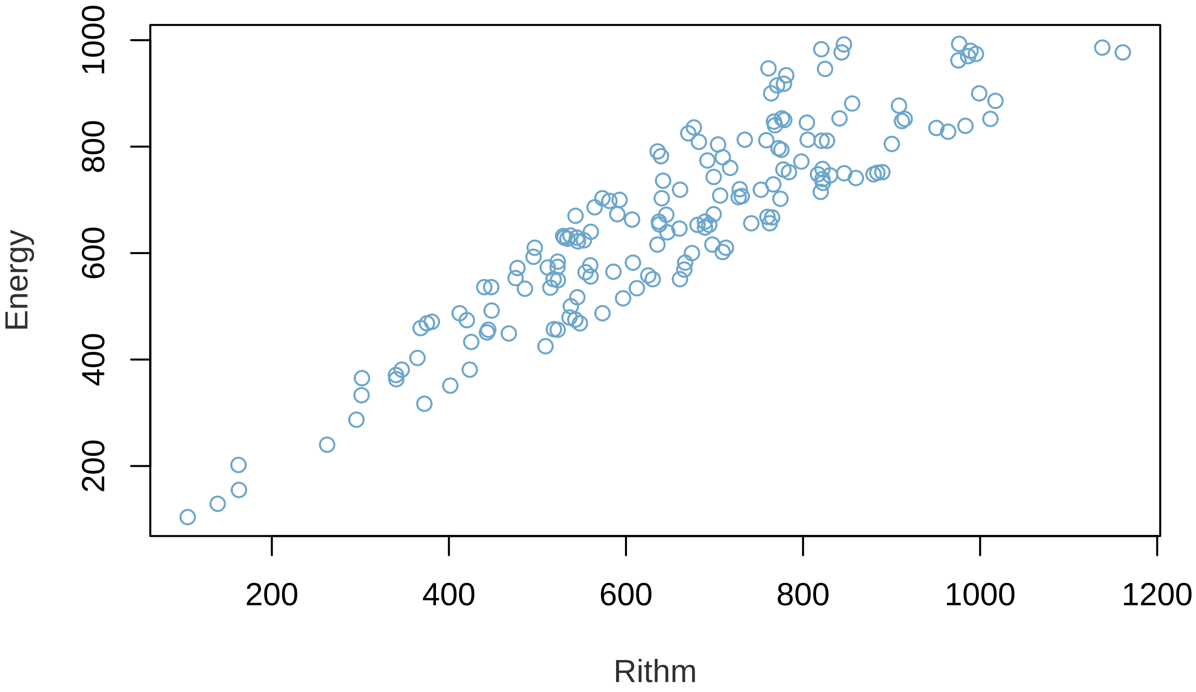
```r
boxplot(rithm, horizontal = TRUE)
```



3rd Block:

On the last block, we will see the Scatter Plot and Linear Model between the main variable, the tempo and the speechiness

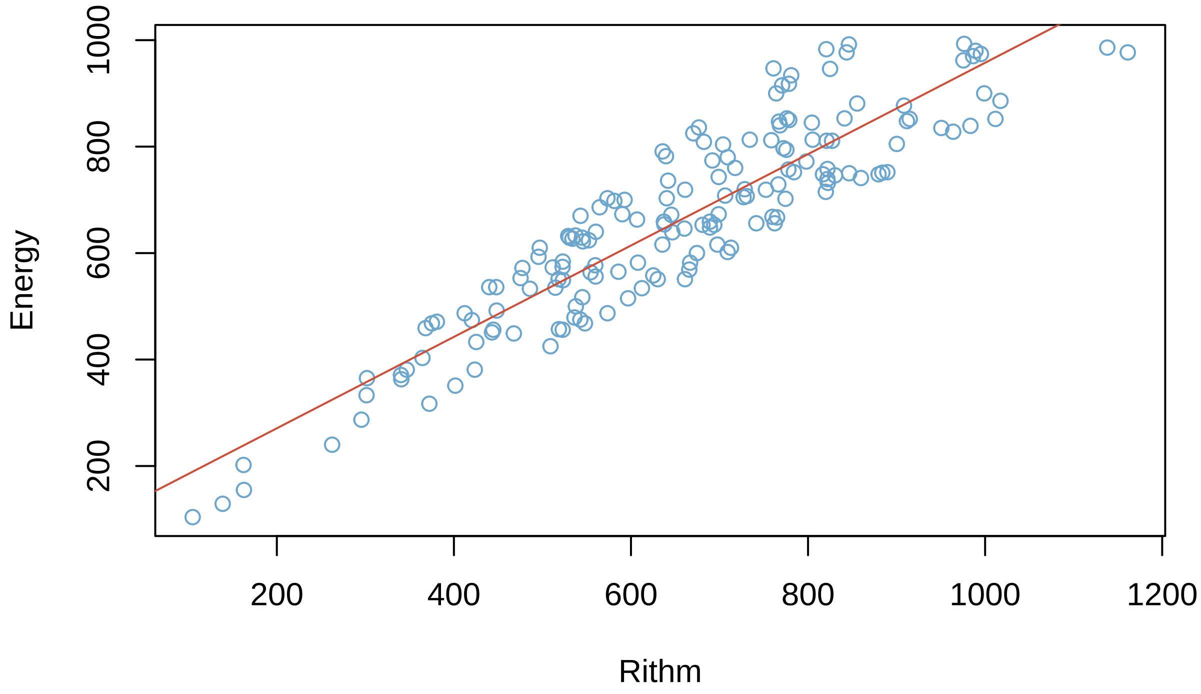
```
# Scatter plot without linear model of tempo and speechiness
plot(
    rithm,
    energy,
    xlab = "Rithm",
    ylab = "Energy",
    col.lab = "gray19",
    col="skyblue3"
)
```



Scatter Plot with the Linear Model:

The scatter plot created with the log_{10} of the energy is:

```
#energy <- log(energy)
energy <- SpotifySongs$energy
rithm <- SpotifySongs$rithm
plot(
    rithm,
    energy,
    xlab = "Rithm",
    ylab = "Energy",
    col="skyblue3",
)
RegressionModel <- lm(energy~ rithm, data=SpotifySongs)
abline(lm(energy ~ rithm), col="tomato3")
```



```
RegressionModel
```

```
##
## Call:
## lm(formula = energy ~ rithm, data = SpotifySongs)
##
## Coefficients:
## (Intercept)        rithm
##     99.1878       0.8584
```

```
print(RegressionModel)
```

```
##
## Call:
## lm(formula = energy ~ rithm, data = SpotifySongs)
##
## Coefficients:
## (Intercept)        rithm
##     99.1878       0.8584
```

```
summary(RegressionModel)
```

```
##
## Call:
## lm(formula = energy ~ rithm, data = SpotifySongs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -118.881  -74.468   -9.043   54.895  194.641
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 99.18779   21.01553    4.72 5.01e-06 ***
## rithm        0.85836    0.03076   27.90   < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.88 on 165 degrees of freedom
## Multiple R-squared:  0.8251, Adjusted R-squared:  0.8241
## F-statistic: 778.5 on 1 and 165 DF,  p-value: < 2.2e-16
```