First Delivery Statistics

Eduardo Alarcón & Alfonso Pineda

2022-10-28

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, namely percentWeeksOnChart, which represents the % of weeks with respect to a year the song has been on the top chart. 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 "loudness" which we think is the one 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

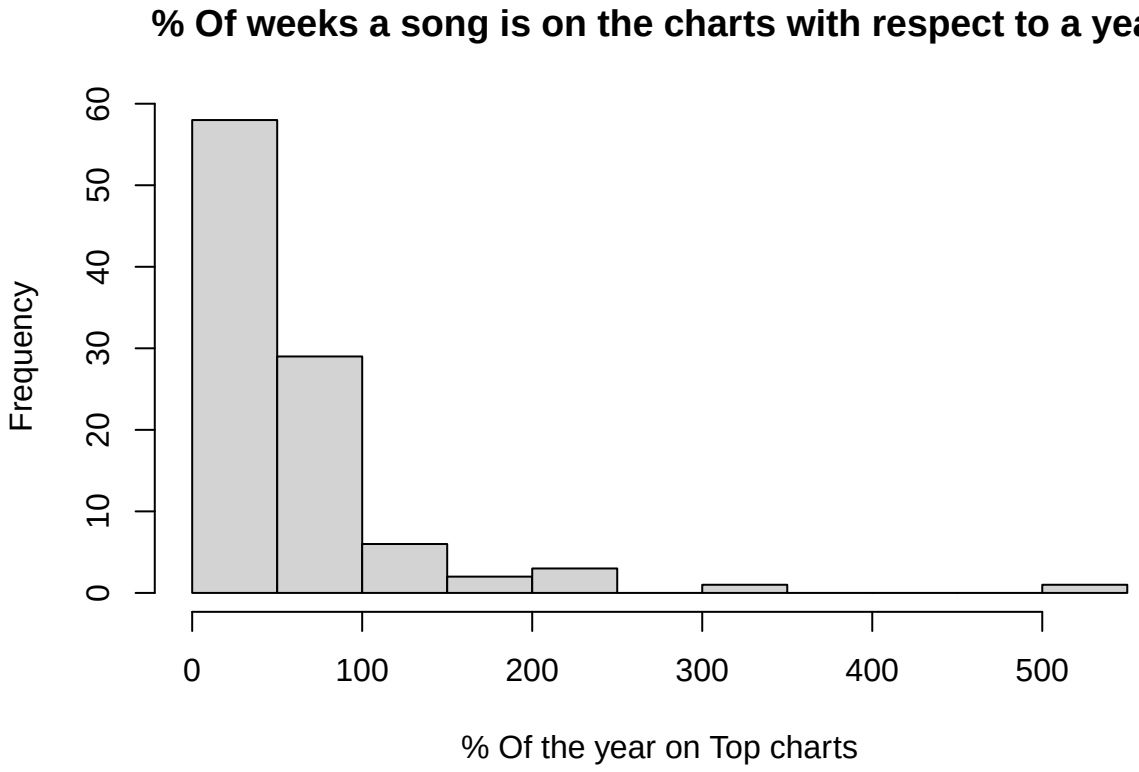
1st Block:

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("spotistats.xlsx")
View(SpotifySongs)
```

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

```
hist(SpotifySongs$percentWeeksOnChart, xlab = "% Of the year on Top charts",
     main = "% Of weeks a song is on the charts with respect to a year")
```

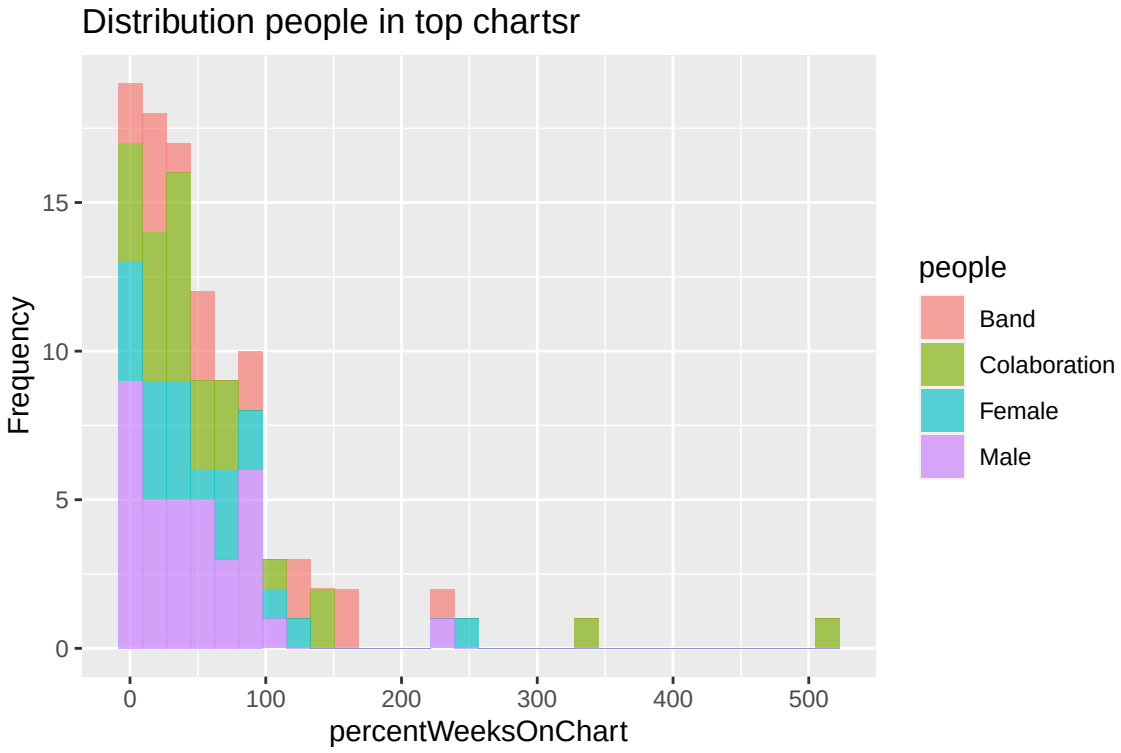


We have also created the histogram differentiating if the artist is a Male, Female, Band or Collaboration between different artists (we used different colors to view them):

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

qplot(percentWeeksOnChart, data=SpotifySongs, geom=c("histogram"), fill=people,
      alpha=I(.65), main="Distribution people in top chartsr",
      xlab="percentWeeksOnChart", ylab="Frequency")
```

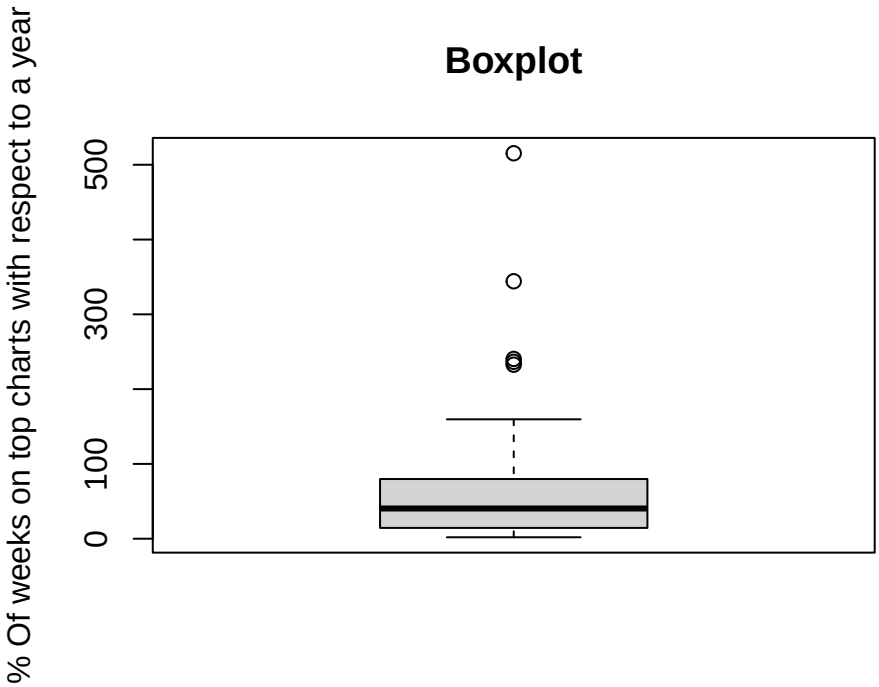
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Boxplot

Then, we also need to create a Box Plot:

```
boxplot(SpotifySongs$percentWeeksOnChart, ylab = "% Of weeks on top charts with respect to a year",
        main = "Boxplot")
```



As we can see from the Box Plot there are some extreme outliers and indicate that some of the songs stay on

the top charts for significant more time than the rest of songs.

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, percentWeeksOnChart. These measures are: the mean, the median, the mode, the percentiles, the range, the variance, and the standard deviation.

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: 100

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

```
## Descriptive Statistics
## weeksonchart
## N: 100
##
##                     weeksonchart
## ----------------- --------------
##              Mean          60.12
##           Std.Dev          74.15
##               Min           1.92
##                Q1          14.42
##            Median          40.38
##                Q3          79.81
##               Max         515.38
##               MAD          45.60
##               IQR          64.43
##                CV           1.23
##          Skewness           3.27
##       SE.Skewness           0.24
##          Kurtosis          14.54
##           N.Valid         100.00
##         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:

```
# Choose best second variable
# Correlation between percentWeeksOnChart and Danceability\n"
cor(SpotifySongs$percentWeeksOnChart, SpotifySongs$Danceability)
```

```
## [1] -0.005667035
```

```
# Correlation between percentWeeksOnChart and Duration\n"
cor(SpotifySongs$percentWeeksOnChart, SpotifySongs$Duration)
```

```
## [1] 0.002479081
```

```
# Correlation between percentWeeksOnChart and Loudness\n"
cor(SpotifySongs$percentWeeksOnChart, SpotifySongs$Loudness)
```

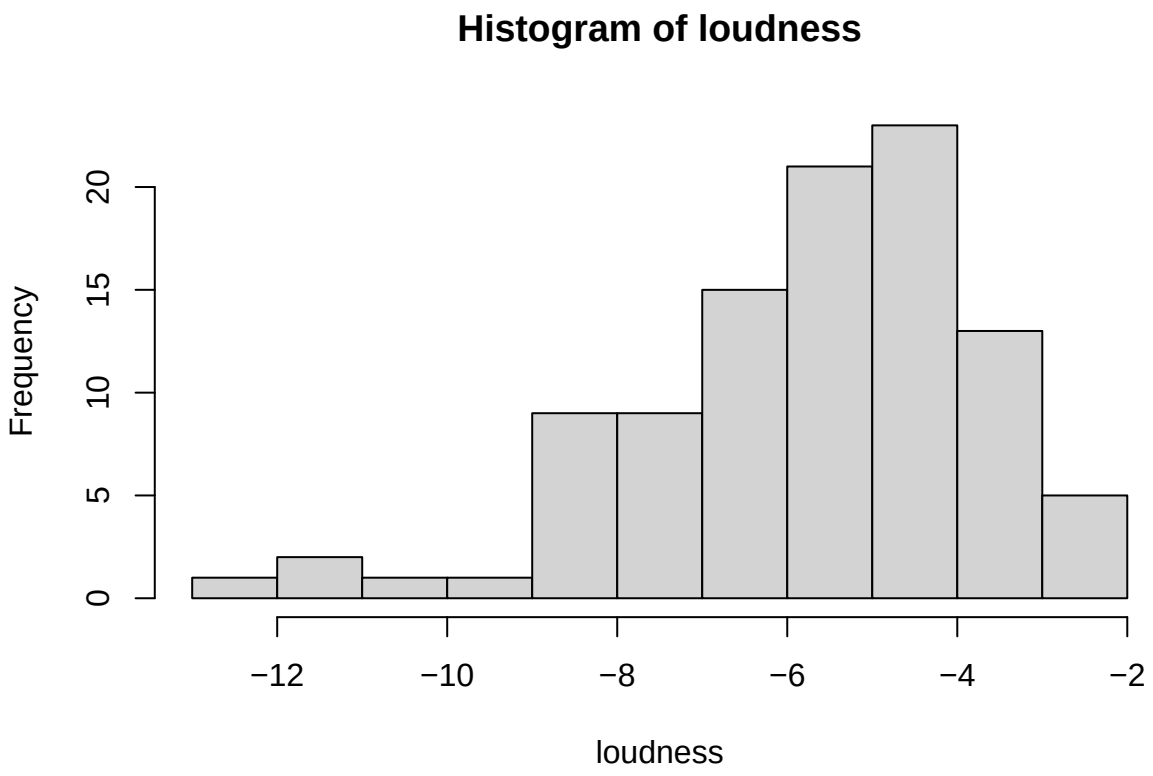
```
## [1] 0.1765837
```

With the previous results, we choose the variable Loudness. Here are the statistical variables of the Loudness, as well as the Histogram and Box Plot:

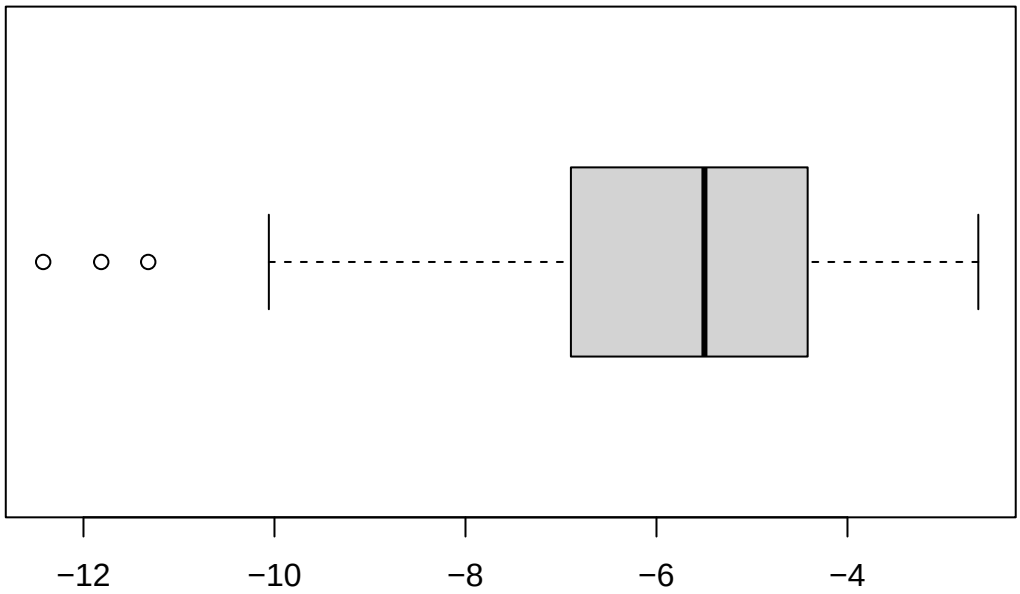
```
loudness<-SpotifySongs$Loudness
descr(loudness)
```

```
## Descriptive Statistics
## loudness
## N: 100
##
##                     loudness
## ----------------- ----------
##              Mean      -5.79
##           Std.Dev       1.96
##               Min     -12.42
##                Q1      -6.90
##            Median      -5.50
##                Q3      -4.42
##               Max      -2.63
##               MAD       1.75
##               IQR       2.47
##                CV      -0.34
##          Skewness      -0.94
##       SE.Skewness       0.24
##          Kurtosis       1.04
##           N.Valid     100.00
##         Pct.Valid     100.00
```

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




```
boxplot(loudness, horizontal = TRUE)
```

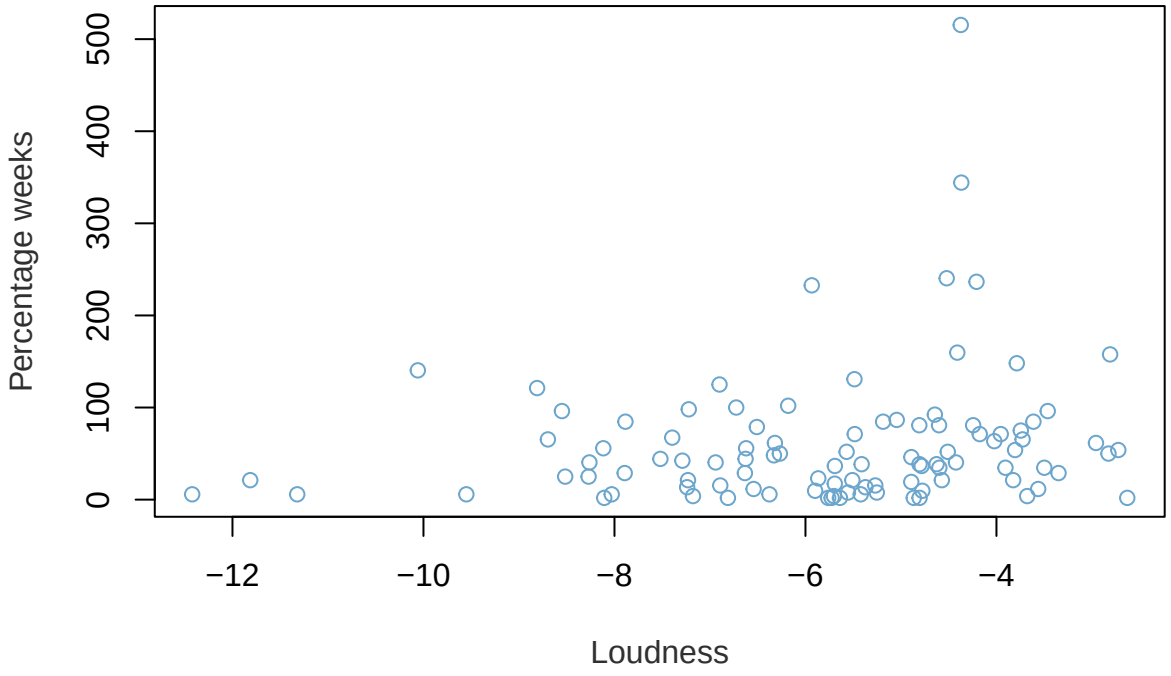


As we can see from the box plot and the histogram, this distribution is more symmetrical than our main variable.

3rd Block:

On the last block, we will see the Scatter Plot and Linear Model between the main variable, the percentWeeksOnChart and the loudness.

```
# Scatter plot without linear model of percentWeeksOnChart and Loudness
plot(
    SpotifySongs$Loudness, # We could also use loudness
    SpotifySongs$percentWeeksOnChart,
    xlab = "Loudness",
    ylab = "Percentage weeks",
    col.lab = "gray19",
    col="skyblue3"
)
```



Scatter Plot with the Linear Model:

The scatter plot created with the Loudness is:

```
RegressionModel <- lm(percentWeeksOnChart ~ Loudness, data=SpotifySongs)
print(RegressionModel)
```

```
##
## Call:
## lm(formula = percentWeeksOnChart ~ Loudness, data = SpotifySongs)
##
## Coefficients:
## (Intercept)     Loudness
##       98.81         6.68
```

```
plot(
    SpotifySongs$Loudness,
    SpotifySongs$percentWeeksOnChart,
    xlab = "Loudness",
    ylab = "Percentage weeks",
    col="skyblue3"
)
abline(RegressionModel, col="tomato3")
```