UNIVERSITETI I EVROPËS JUGLINDORE
УНИВЕРЗИТЕТ НА ЈУГОИСТОЧНА ЕВРОПА
SOUTH EAST EUROPEAN UNIVERSITY

Capstone Project
Faculty of Contemporary Sciences and Technologies (CST)
Study Program: Business Informatics

Title:

What makes the top songs Popular?

Mentor:	Student:
Assit.PhD. Marika Apostolova	Arbesa Kajtazi
________________	_____________

June, 2018

This page is intentionally left blank

ABSTRACT

The aim of this research is to put forward an overview of applying data mining and data visualization within the scope of audio data from a dataset of Spotify. The research starts by presenting background information of these two fields and their influence on music industry.

This includes explanation of the most essential concepts and their role. The thesis is concentrated on analysis of audio features of the tracks in Spotify’s Top Songs of 2017 playlist and try highlight the common patterns behind the audio features of these songs. The thesis is concentrated on Spotify datasets as practical scenario. For this reason, more detailed information is given about songs features, what are they, what do these top songs have in common and why do people like them.

The result of the study showcase how singers and song makers can leverage the power of data visualization and data mining to help trying to predict one audio feature based on the others, look for patterns in the audio features of the songs(Why do people stream these songs the most.) and see which features correlate the most.

ACKNOWLEDGEMENT

Initially, I would like to thank my mentor, Assit.PhD. Marika Apostolova, for her valuable suggestions throughout the thesis. I am very grateful for the advices received and the lead offered. I found quite beneficial the insightful comments that pointed to the needed improvements.

My special thanks goes to prof. Visar Shehu who introduced and led me to the world of data mining applications. I genuinely appreciate his efforts for advising and helping us and being available for all the questions I had on understanding better the data mining world better.

This thesis was inspired from the idea of the project I had for the Data mining class during the VI semester.

Table of Contents

List of Figures

1 INTRODUCTION

In the recent years, we started witnessing Big Data. This trend has influenced all industries. More and more data is generated and stored in different formats. There are many challenges we face with these exaggerated volumes of data, starting from the step of capturing them till properly displaying valuable information.

*"**Data is the new oil.** It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals etc. to create a valuable entity that drives profitable activity; so data must be broken down, analyzed for it to have value"*[1]. Big data comes with big promises also, however consisting of only 1s and 0s, the data is imperceptible to the public. For organizations to excel in their work, sometimes conducting only an analysis of data is not enough, effective dissemination of the information is required.

The merit of data mining and visualization here is that they, together, utilize and make sense of the data. Since both fields are emerging, research on how to best combine these two areas is conducted every day. The field of music industry can benefit a lot from this synthesis. As new technology rapidly is being developed, the way of storytelling is changed by integrating data visualization tools.

The purpose of this thesis is to see what these songs have in common and why people like them by using big data analytics. Hopefully, with the new knowledge, it can contribute to the prediction of audio features based on the others. To achieve this purpose, I will be using a Spotify dataset where the audio features where extracted using the Spotify Web API and the Spotify Python library. The credit goes to Spotify for calculating the audio feature values.

This is done by exploring the research questions:

1. Can you predict what is the rank position or the number of streams a song will have in the future?
2. How long does songs "resist" on the top 3, 5, 10, 20 ranking?

[1] Clive Humby, UK mathematician and Data Science Innovator at Tesco

3. What are the signs of a song that gets into the top rank to stay?
4. Do continents share same top ranking artists or songs?
5. Are people listening to the very same top ranking songs on countries far away from each other?
6. How long time does a top ranking song takes to get into the ranking of neighbor countries?

In doing so, this study will be of interest for:

a) The field of music industry where the organizations are trying to adapt to changes in the industry.

b) Scholars wanting to expand their knowledge in the application forms of data mining and data visualization.

2 BACKGROUND

2.1 Data Mining

"**Data mining** is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems"[2]. This new powerful technology enhances the value of existing data by automating the process of extracting knowledge from it. Massive databases are analyzed in minutes which enables users to conduct more studies in shorter amount of time and at the same time helps understand complex data.

Data mining often is used as synonym for knowledge discovery from data (KDD), but in fact, it only depicts a step of that process.

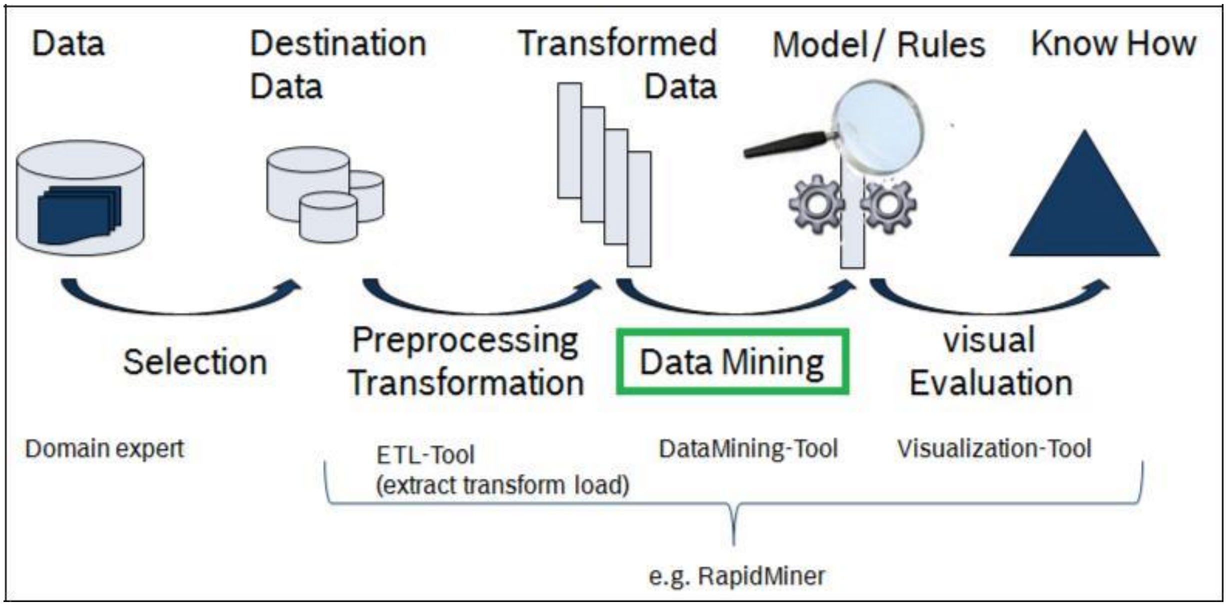


Figure 1 Data mining in Knowledge discovery process [2]

Hidden patterns are found following various techniques and implementing the corresponding algorithms. Most popular techniques that provide diverse insights are: classification analysis, association rule learning, anomaly detection, clustering analysis and regression analysis.

Data mining offers a solution to the problem by intelligently analyzing the present data in the database. This process often includes discovering patterns among data, finding specific connection between them that will be used for prediction of further actions.

The importance of data mining as resource is that it elucidates the hidden knowledge in large datasets. Applying data mining has proved to lead to new insights, better decision making and sometimes bring competitive advantage to companies. Data mining is interdisciplinary computer science field with broad implementation in other fields as well. This allows it to be used not only for business purposes, but also in journalism. Later, is given more detailed information about how journalism benefits from data mining.

2.2 Data visualization

Visualization is often associated with coordinated system, points, lines, tables, charts or graphs. This application provides unique perspective on the information. Additionally, it can be viewed as a form of humanizing statistics. The capability of combining data visualizations with stories gave birth to new form of storytelling known as "***narrative visualization***".

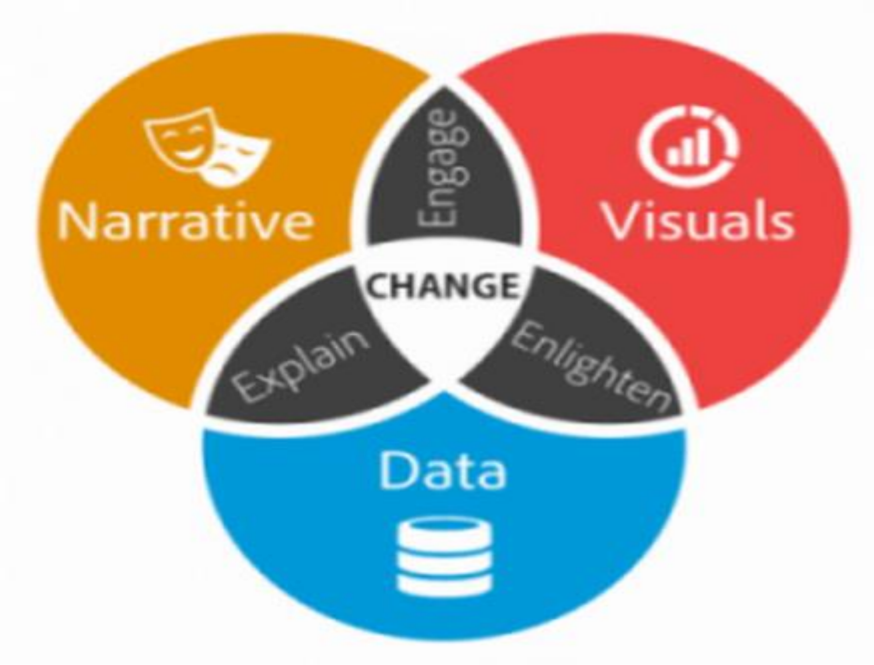


Figure 2 Narrative visualization

Data visualization plays perhaps one of the most important roles in successfully communicating information. From human-computer interaction perspective, it has an ability to capture the attention of the human eye and brain. Ware explains the crucial role that it has by proclaiming that "*Visual displays provide the highest bandwidth channel from the computer to the human. We acquire more information through vision than through all of the other senses combined. The 20 billion or so neurons of the brain devoted to analyzing visual information provide a pattern-finding mechanism that is a fundamental component in much of our cognitive activity."* [4].

The two main reasons to use visualization as a form of presentation in stories are:

a) exploratory: to understand what is present in the data

b) Communicative: to display and discuss visualizations with the public by telling stories based on the data and the facts. [5]

1.3 Music Psychology and the Exposure Effect

Surprisingly, the question of "Why we like or not a particular song?" has received little attention from music psychology. Although music preference is recognized as a central aspect of modern identities, the field is "still in its infancy". The issue of liking per se is indeed difficult to study directly, and music psychologists have traditionally focused on less elusive, more directly measurable phenomena such as memorization, recognition or learning.

In our context, a central issue in trying to explain music hits is exposure, that is, the simple fact of listening to a musical piece. What is the effect of exposure on preference or liking? Studies on exposure show that there is indeed an impact of repeated exposure on liking, but also that this impact is far from simple. Parameters such as the context, type of music or listening conditions (focused or incidental), seem to influence the nature of this impact, and many contradictory results have been published.

The popular idea that repeated exposure tends to increase liking was put forward early and was confirmed experimentally in a wide variety of contexts and musical genres. The so-called mere exposure effect, akin to the familiarity principle, or perceptual fluency, is considered by many psychologists to be a robust principle, pervading many facets of music listening.

However, as noted by Schellenberg [29], this increase in liking may be restricted to musically impoverished or highly controlled stimuli. Indeed, other studies have shown a more subtle effect of repeated exposure. The study by Siu-Lan et al. [31] showed different effects of exposure on intact and patchwork compositions. An inverted U-curve phenomena was observed in particular by Szpunar et al. [33] and Schellenberg [30], it explained in large part by the “two factor model” of Berlyne [3]. In this model, two forces compete to build up liking: (1) the arousal potential of the stimulus (the music), which decreases with repeated listening, thereby increasing liking (with the habituation to this arousal potential), and (2) familiarity, which tends to create boredom. These two forces combined produce typical inverted U-shapes that have been observed in many studies of preference. This model is itself related to the famous “Wundt curve” [36]. The Wundt curve describes the typical experience of arousal as being optimal when achieving a compromise between repetition/boredom and surprise.

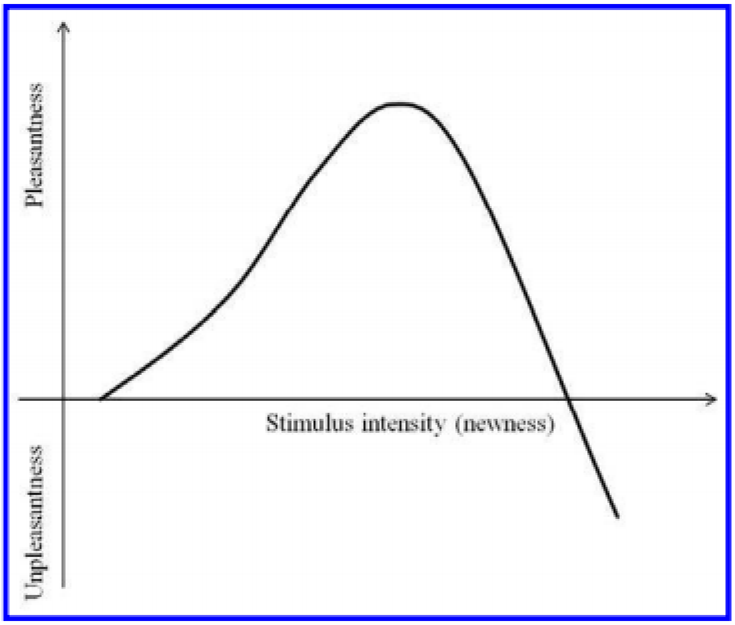


Figure 3 The Wundt curve describes the optimal "hedonic value" as the combination of two conflicting forces

Yet, other studies [35] show in contrast a polarization effect, whereby repeated exposure does not influence initial likings but makes them stronger, both positively and negatively. Finally, Loui et al. [19] studied exposure effects by considering exotic musical temperaments, to study the relation between learning and preference. They showed that passive exposure to melodies built in an entirely new musical system led to learning and generalization, as well as increased preference for repeated melodies. This work emphasizes the importance of learning in music preference. These psychological experiments show that a relation between exposure and liking exists, but that this relation is complex and still not well understood, in particular for rich, emotionally meaningful pieces. It is therefore impossible to simply consider, from a psychological point of view, that repeated exposure necessarily increases liking: it all depends on a variety of factors.

1.4 Data-driven Music Industry

Big data is a term that reflects the amount of information people generate – and a lot. Some estimate that today, humans generate more information in one minute than in every moment from the earliest historical record through 2000. Unsurprisingly, harnessing this data has shaped music industry in radical new ways.

Being a data driven company means using the data in almost any part of the organization. Spotify is one of the data-driven companies where users create 600 Gigabyte of data per day and 150 Gigabyte of data per day via different services. Music companies, artists and producers are shifting their gaze towards digital music distribution while maintaining other revenue streams. However, it's worth mentioning that these traditional channels are also being disrupted by technological innovations. With data becoming easier to acquire and analyze, it has now become possible to crunch large numbers and predict audience preference and buying tendencies.

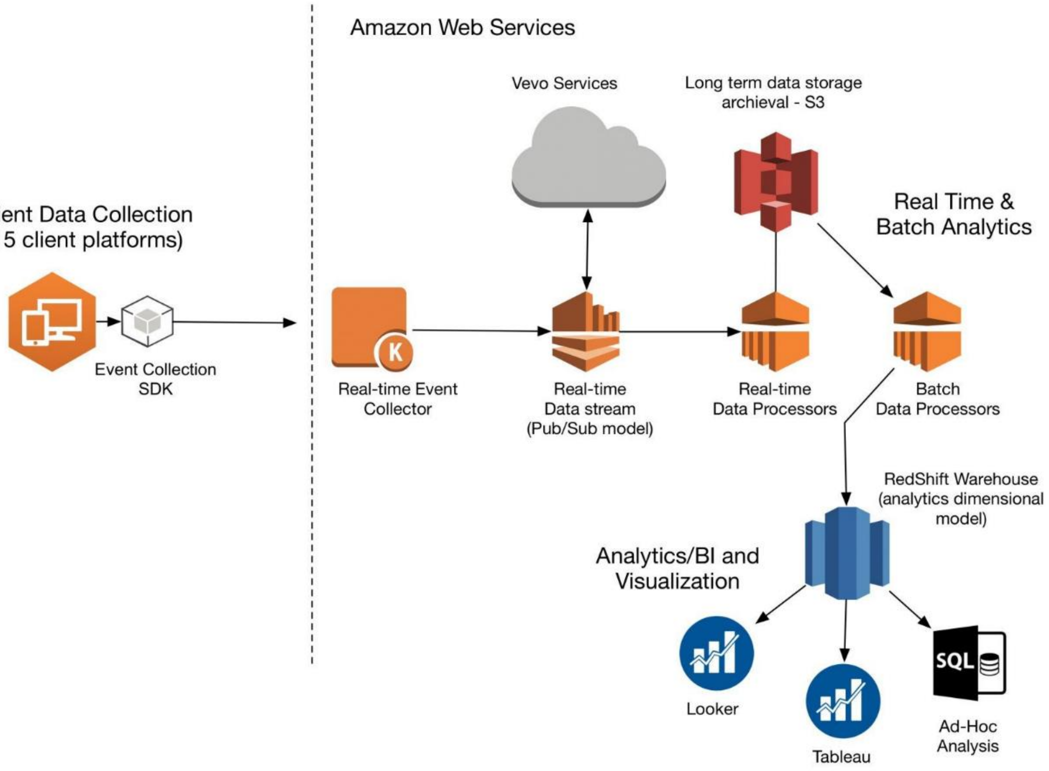

Amazon Web Services
Vevo Services
Long term data storage archieval - S3
Client Data Collection (15 client platforms)
Real Time & Batch Analytics
Event Collection SDK
K
Real-time Event Collector
Real-time Data stream (Pub/Sub model)
Real-time Data Processors
Batch Data Processors
RedShift Warehouse (analytics dimensional model)
Analytics/BI and Visualization
Looker
Tableau
SQL
Ad-Hoc Analysis

Figure 4 Data driven music industry

Whereas in the past, the industry relied primarily on sales and how often a song was played on the radio, they can now see what specific songs people are listening to, where are they hearing it and how they are consuming it.

On a daily basis, people generate 2.5 exabytes of data, which is equivalent to 250,000 times all of the books in the Library of Congress. Obviously, not all of this data is useful to the music industry. But analytical software can utilize some of it to help the music industry understand the market.

Big Data is being integrated into nearly every field. It should be no surprise that the multi-billion dollar music industry wants in. There are two major ways big data is already influencing the music industry; music creation and music selection.

Numerous studies show that people like music that sounds familiar. This means that there is a certain circular quality to pop music. The more you are forced to listen to that inescapable new song, the more you like it. You liking means more music of the same type will be produced. Data, however, is making the new era of music one of the most populistic. For example; in our case Spotify knows what listeners like and want. The ability to listen to music is only the most basic feature of Spotify. They constant compile data and create algorithms to suggest new music to listeners.

The logical extension of data-driven music is data-created music. Don't just wait around for the next big star to pop up – engineer them from scratch. For, many, this is a horrifying, terrible idea. What if the art of music creation could entirely remove from the process?

3 PROBLEM STATEMENT

The music industry is in the middle of something new and bold, thanks to digital technology. Traditional distribution venues have already been replaced with digital ones and this has surprisingly boosted the popularity of artists. Instead of slowing things down, downloads and media streaming have helped artist mount more successful tours and sold out concerts than ever before. This is a positive example of what digital (data-driven) innovations and bold ideas can do.

After years of steady decline, the global music industry revenue grew from $14.5 billion to $15 billion between 2016 and 2017. The sales of physical media continued to slow down, accounting for only about a third of the 2017 revenues. In the meantime, digital sales continue to grow, with a 17, 7% increase from 2016.

Subscription services, which include Spotify, had 68 million users by 2017 and total revenues of $2 billion. The majority of users, however, prefer to listen for free, via ad-supported sites which include YouTube. Total industry earning was only $634 million in 2017. Another trend to note is the growth of streaming services due to the growing buying power of millennials – an increasingly powerful market – which is currently estimated to be $1 trillion.

3.1 Music Industry

From a global music currency to virtual multi-camera recording setups, there are a lot of ways that technology is helping the music industry grow.

Fifteen years ago, Steve Jobs introduced the iPod. Since then, most music fans have understood this has radically changed how they listen to music. Data from what we download and listen to can now be mined to create and promote future songs. Less understood are the ways that raw information – accumulated via downloads, apps and online searches – is influencing not only what songs are marketed and sold, but which songs become hits.

Decisions about how to market and sell music, to some extent, still hinge upon subjective assumptions about what sounds good to an executive, or which artists might be easier to market. Increasingly, however, businesses are turning to big data and the analytics that can help turn this information into actions.

But, the big question I arise in this research is: Can someone predict whether your recently produced song will become a hit?

Any pop song composer would probably laugh at this question and respond: How could someone predict the success of what took so much craft, pain, and immeasurable creativity to produce? I myself do not even have a clue! This question raises a recurring fantasy in our culture: wouldn't it be thrilling to understand the "laws of attraction" that explain how this sort of preference system of music in human beings works, to the point of being able to predict the success of a song or any other cultural artifact before it is even released? This fantasy is elaborated in detail in Malcom Gladwell's story "The Formula" [4]. In this fiction, a—needless to say, fake—system is able to predict the success of movies by analyzing their script automatically. The system is even smart enough to propose modifications of the script to increase the success of the movie, with a quantified estimation of the impact in revenues. In the introduction, Gladwell begins by describing the reasons why we like a movie as resulting from a combination of small details.

He writes:

> Each one of those ... narrative details has complicated emotional associations, and it is the subtle combination of all these associations that makes us laugh or choke up when we remember a certain movie... Of course, the optimal combination of all those elements is a mystery. [4]

This process is also true for music: what makes us like a song or not probably has to do with a complex combination of micro-emotions, themselves related to our personal history, to the specifics of the song and too many other elusive elements that escape our direct understanding. In spite of the many claims that writing hit songs is just a matter of technique (see, for example, Blume), it is likely that, as the highly successful Hollywood screenwriter William Goldman said: "Nobody knows anything" [5].

However daring, I will challenge this assumption by precisely undertaking the task of making these kinds of predictions. Several companies now claim to be able to automatically analyze songs in order to predict their success (HSS, PlatiniumBlue) and to sell their results to record labels.

3.2 Related Work

While I researched among different webpages, I did not find such a predictive analysis with data visualization, some of them were only research paper without including data mining and data visualization. There is one project on Kaggle in python [] that is really close to mine but that used different kind of techniques for an attempt to build a classifier that can predict whether or not he likes a song.

4 METHOD

4.1 System architecture

For this research, a traditional knowledge discovery process was used: first the data was gathered, pre-processed & cleaned, analyzed and at the end the results were visualized. The general system architecture followed is presented in Figure 4.

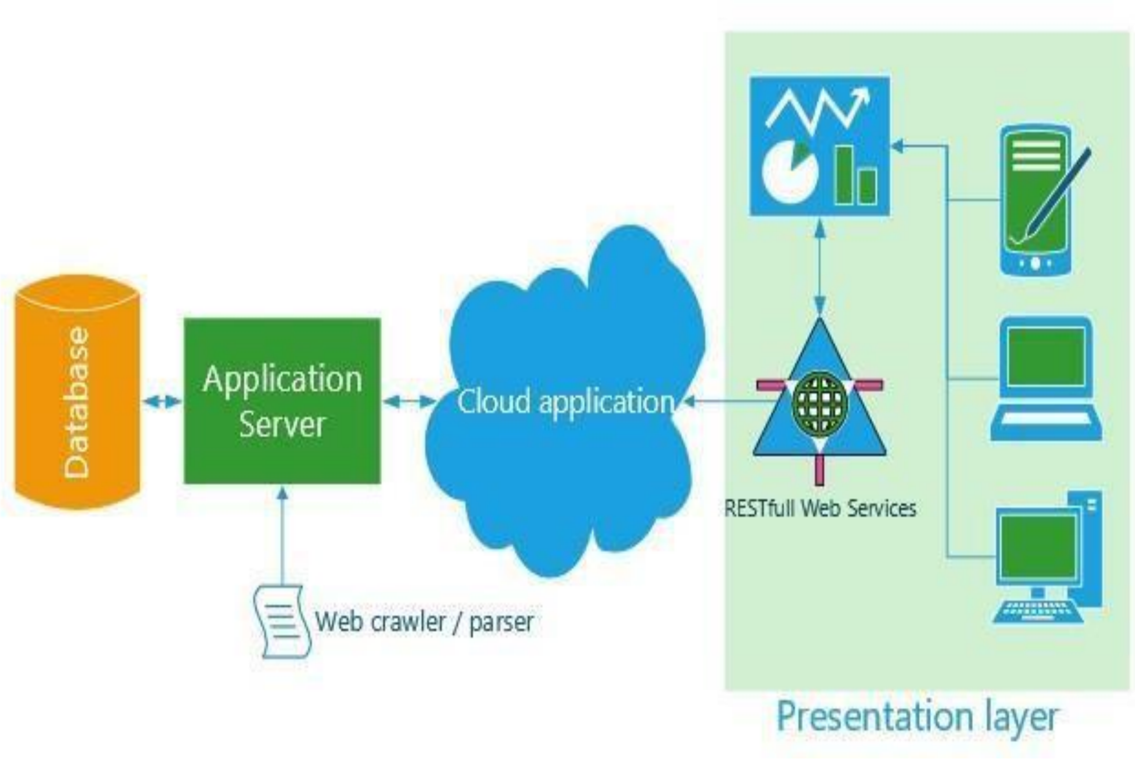


Figure 5 General system architecture

4.2 Data gathering and pre-processing

Even though the data is open and free, there are costs associated with the process of getting the dataset as a resource. Often the data should be converted into specific format, annotated or by necessity a certain metadata to be created. All of this requires human resources with specialized skills to do the work and technology resources (both hardware and software).

Spotify has developed a workflow manager, Luigi, which it is open source. Luigi is a Python framework for data flow definition and execution. Luigi is used to crunch a lot of data. Most of the data is user-centric data, such as billions of log messages that allows Spotify to provide music recommendations or select for example the next song heard on the radio. The data however is also used in decision-making, providing forecasting information and business analytics. I found the dataset from Kaggle.

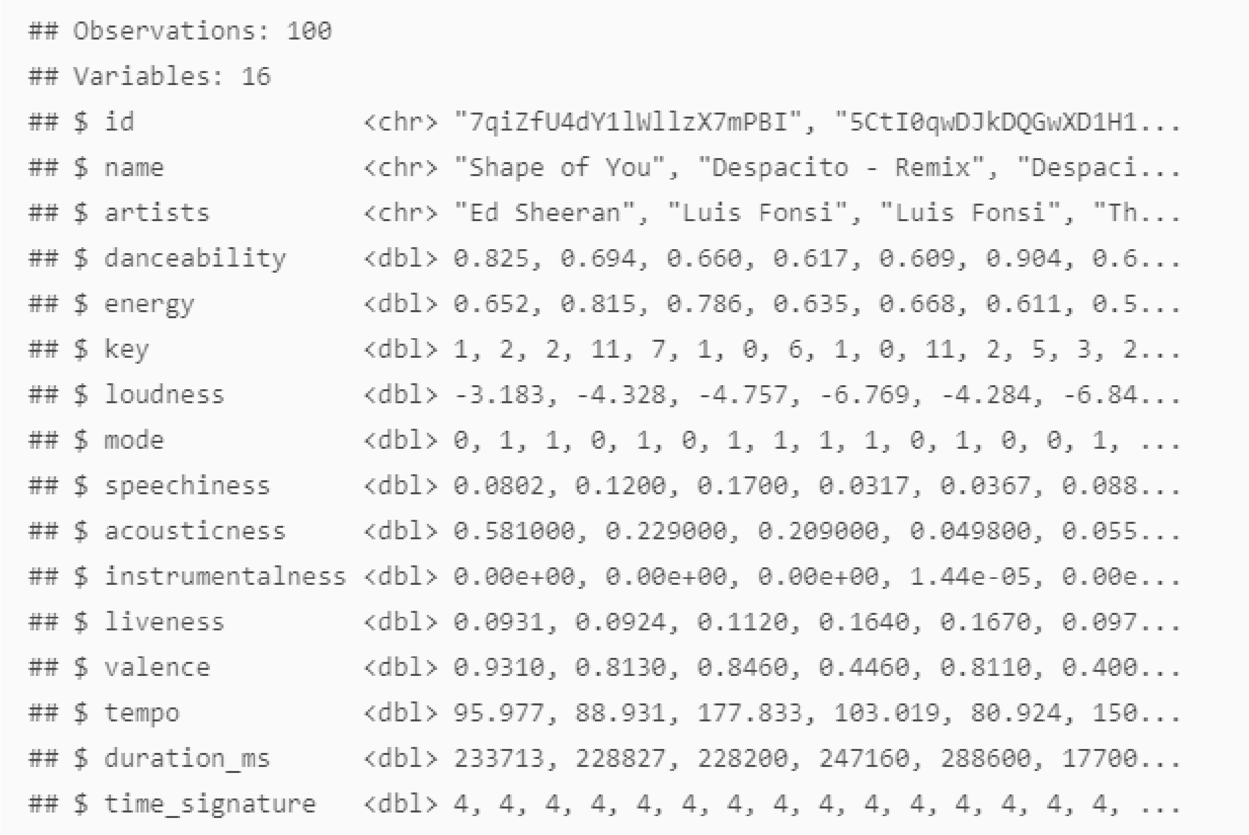
4.3 Data analysis

The dataset contains the daily ranking of the 200 most listened songs in 53 songs from 2017 and 2018 by most Spotify users. It contains more than 2 million rows, which comprises 6629 artists, 18598 songs for a total count of one hundred five billion streams count. The data spans from 1st January 2017 to 9th January 2018 and will be kept up-to-date on following versions. It has been collected from Spotify's regional chart data.

For manipulating with data and visualizing the results, RStudio as integrated development environment (IDE) for R was used. R is designed to read entire datasets into memory and work with them there. The dataset imported in RStudio was of *.csv* format. For some of the cases, it was necessary to read the data into a data frames and use that data frame to conduct further manipulations.

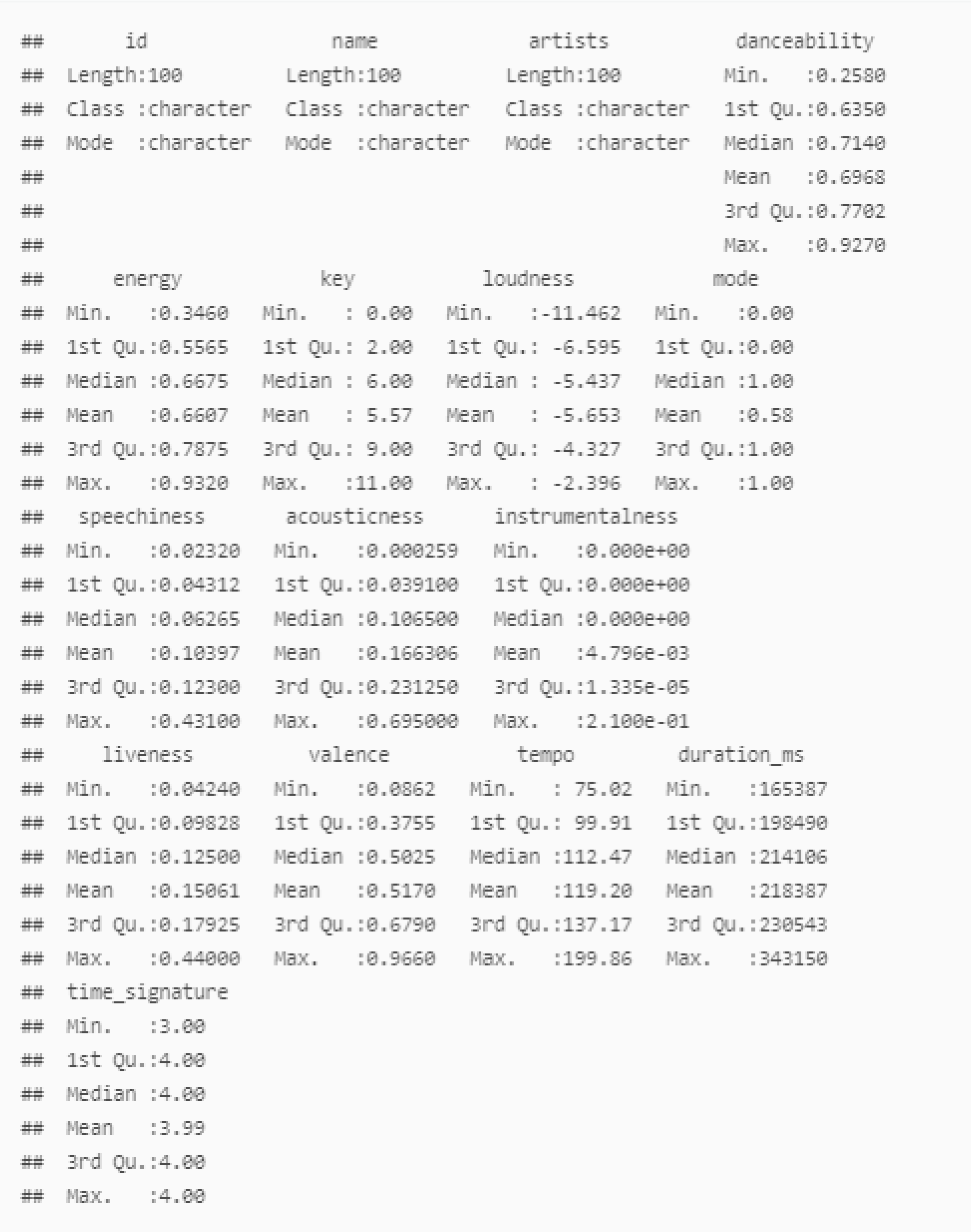
Data visualization was possible with the use of ggplot2 and plotrix package. This plotting system of R allows customization of the diagram and offers different built-in functions. Combination of data, aesthetic mappings and geom (geometric object) are specified for every diagram. The details of how data values are translated into visual representation were controlled by changing the scales, limiting the values of the axes, modifying labels and legend and adding theme to the plot.

First of all I will begin with the implementation of data and get an idea of what I am working with:



```
## Observations: 100
## Variables: 16
## $ id               <chr> "7qiZfU4dY1lWllzX7mPBI", "5CtI0qwDJkDQGwXD1H1...
## $ name             <chr> "Shape of You", "Despacito - Remix", "Despaci...
## $ artists          <chr> "Ed Sheeran", "Luis Fonsi", "Luis Fonsi", "Th...
## $ danceability     <dbl> 0.825, 0.694, 0.660, 0.617, 0.609, 0.904, 0.6...
## $ energy           <dbl> 0.652, 0.815, 0.786, 0.635, 0.668, 0.611, 0.5...
## $ key              <dbl> 1, 2, 2, 11, 7, 1, 0, 6, 1, 0, 11, 2, 5, 3, 2...
## $ loudness         <dbl> -3.183, -4.328, -4.757, -6.769, -4.284, -6.84...
## $ mode             <dbl> 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, ...
## $ speechiness      <dbl> 0.0802, 0.1200, 0.1700, 0.0317, 0.0367, 0.088...
## $ acousticness     <dbl> 0.581000, 0.229000, 0.209000, 0.049800, 0.055...
## $ instrumentalness <dbl> 0.00e+00, 0.00e+00, 0.00e+00, 1.44e-05, 0.00e...
## $ liveness         <dbl> 0.0931, 0.0924, 0.1120, 0.1640, 0.1670, 0.097...
## $ valence          <dbl> 0.9310, 0.8130, 0.8460, 0.4460, 0.8110, 0.400...
## $ tempo            <dbl> 95.977, 88.931, 177.833, 103.019, 80.924, 150...
## $ duration_ms      <dbl> 233713, 228827, 228200, 247160, 288600, 17700...
## $ time_signature   <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, ...
```

Figure 6 Implementation of data



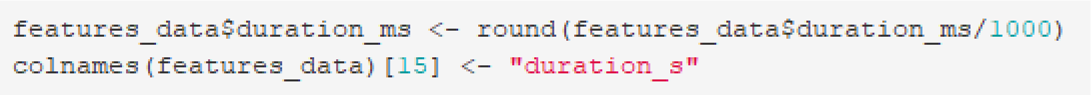
```
##       id                name              artists           danceability   
##  Length:100         Length:100         Length:100         Min.   :0.2580  
##  Class :character   Class :character   Class :character   1st Qu.:0.6350  
##  Mode  :character   Mode  :character   Mode  :character   Median :0.7140  
##                                                           Mean   :0.6968  
##                                                           3rd Qu.:0.7702  
##                                                           Max.   :0.9270  
##      energy            key           loudness           mode     
##  Min.   :0.3460   Min.   : 0.00   Min.   :-11.462   Min.   :0.00  
##  1st Qu.:0.5565   1st Qu.: 2.00   1st Qu.: -6.595   1st Qu.:0.00  
##  Median :0.6675   Median : 6.00   Median : -5.437   Median :1.00  
##  Mean   :0.6607   Mean   : 5.57   Mean   : -5.653   Mean   :0.58  
##  3rd Qu.:0.7875   3rd Qu.: 9.00   3rd Qu.: -4.327   3rd Qu.:1.00  
##  Max.   :0.9320   Max.   :11.00   Max.   : -2.396   Max.   :1.00  
##   speechiness       acousticness      instrumentalness  
##  Min.   :0.02320   Min.   :0.000259   Min.   :0.000e+00  
##  1st Qu.:0.04312   1st Qu.:0.039100   1st Qu.:0.000e+00  
##  Median :0.06265   Median :0.106500   Median :0.000e+00  
##  Mean   :0.10397   Mean   :0.166306   Mean   :4.796e-03  
##  3rd Qu.:0.12300   3rd Qu.:0.231250   3rd Qu.:1.335e-05  
##  Max.   :0.43100   Max.   :0.695000   Max.   :2.100e-01  
##     liveness          valence           tempo         duration_ms    
##  Min.   :0.04240   Min.   :0.0862   Min.   : 75.02   Min.   :165387  
##  1st Qu.:0.09828   1st Qu.:0.3755   1st Qu.: 99.91   1st Qu.:198490  
##  Median :0.12500   Median :0.5025   Median :112.47   Median :214106  
##  Mean   :0.15061   Mean   :0.5170   Mean   :119.20   Mean   :218387  
##  3rd Qu.:0.17925   3rd Qu.:0.6790   3rd Qu.:137.17   3rd Qu.:230543  
##  Max.   :0.44000   Max.   :0.9660   Max.   :199.86   Max.   :343150  
##  time_signature
##  Min.   :3.00  
##  1st Qu.:4.00  
##  Median :4.00  
##  Mean   :3.99  
##  3rd Qu.:4.00  
##  Max.   :4.00  
```

Figure 7 Implementation of data

4.4 Results

A. Case

First of all we will convert the duration in seconds for the sake of simplicity of analysis.



```
features_data$duration_ms <- round(features_data$duration_ms/1000)
colnames(features_data)[15] <- "duration_s"
```

Here we create a variable where we get the column "*duration_ms*" and convert it into seconds with a name of "*duration_s*".

B. Case

Here, on the next step the bar graph shows "top artists based on their appearance and songs."

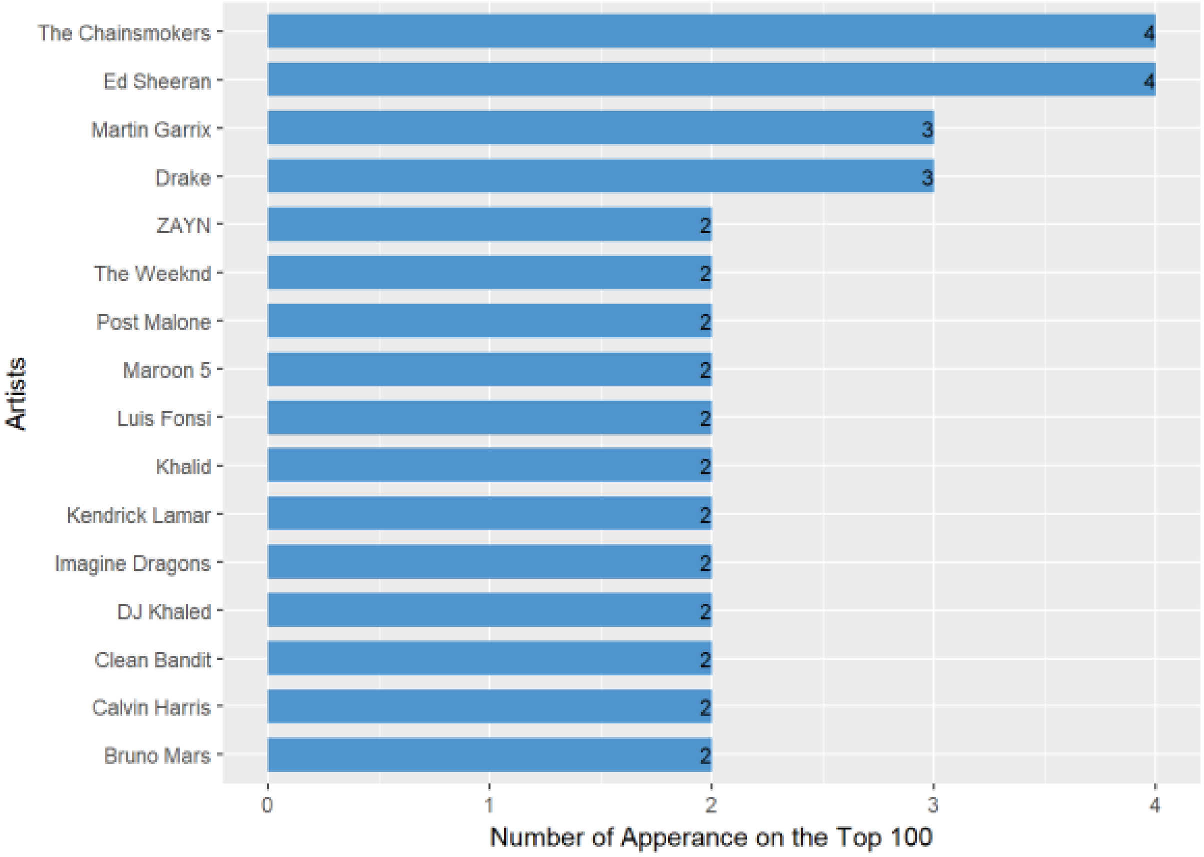
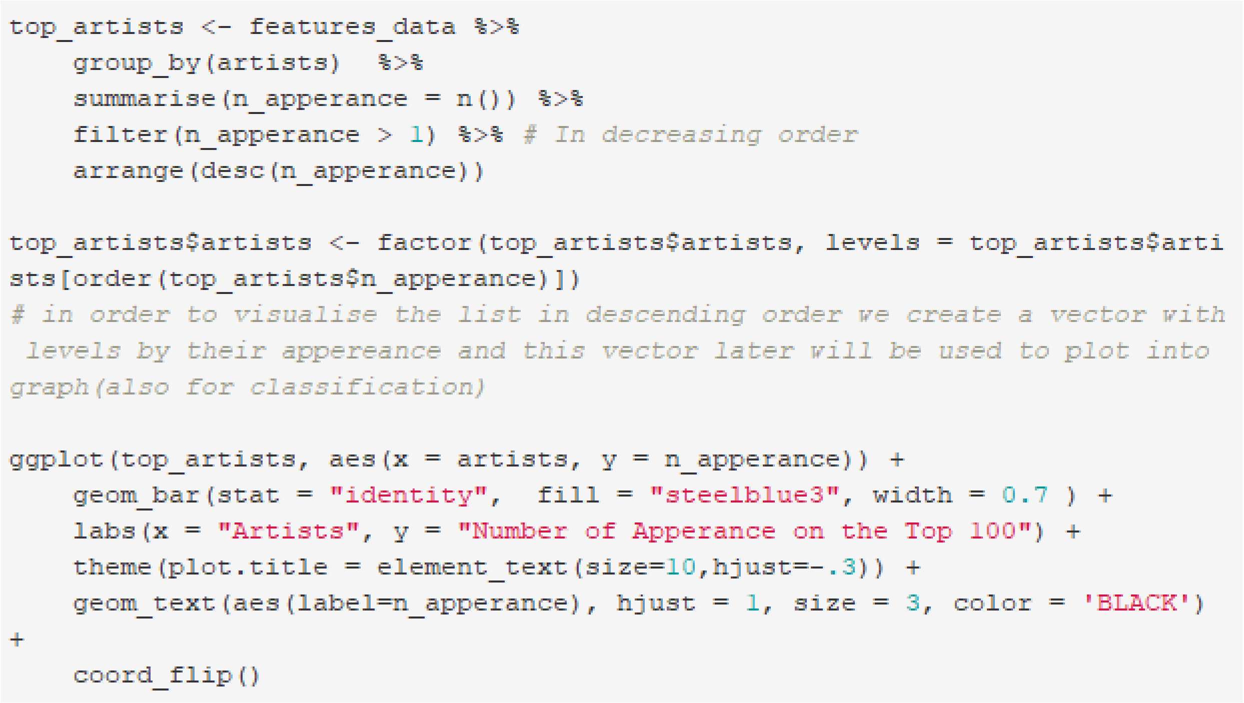


Figure 8 Top artist based on their appearance

R code for figure 8:



```
top_artists <- features_data %>%
    group_by(artists)  %>%
    summarise(n_apperance = n()) %>%
    filter(n_apperance > 1) %>% # In decreasing order
    arrange(desc(n_apperance))

top_artists$artists <- factor(top_artists$artists, levels = top_artists$arti
sts[order(top_artists$n_apperance)])
# in order to visualise the list in descending order we create a vector with
 levels by their appereance and this vector later will be used to plot into
graph(also for classification)

ggplot(top_artists, aes(x = artists, y = n_apperance)) +
    geom_bar(stat = "identity",  fill = "steelblue3", width = 0.7 ) +
    labs(x = "Artists", y = "Number of Apperance on the Top 100") +
    theme(plot.title = element_text(size=10,hjust=-.3)) +
    geom_text(aes(label=n_apperance), hjust = 1, size = 3, color = 'BLACK')
+
    coord_flip()
```

We here first, create a variable top artists and take the dataset features_data and group all the artists where and summarize all the appearance in the data set of the artist and filter them in descending order.

For creating the graph we used ggplot where we use the list top artists we created and with two axes, x which equals to artists and y which equals the appearance of an artist. Then we use the R studio features to design the graph with the colors and other features we like.

C. Case

The graph below shows Top Artist in Central Europe by playing time

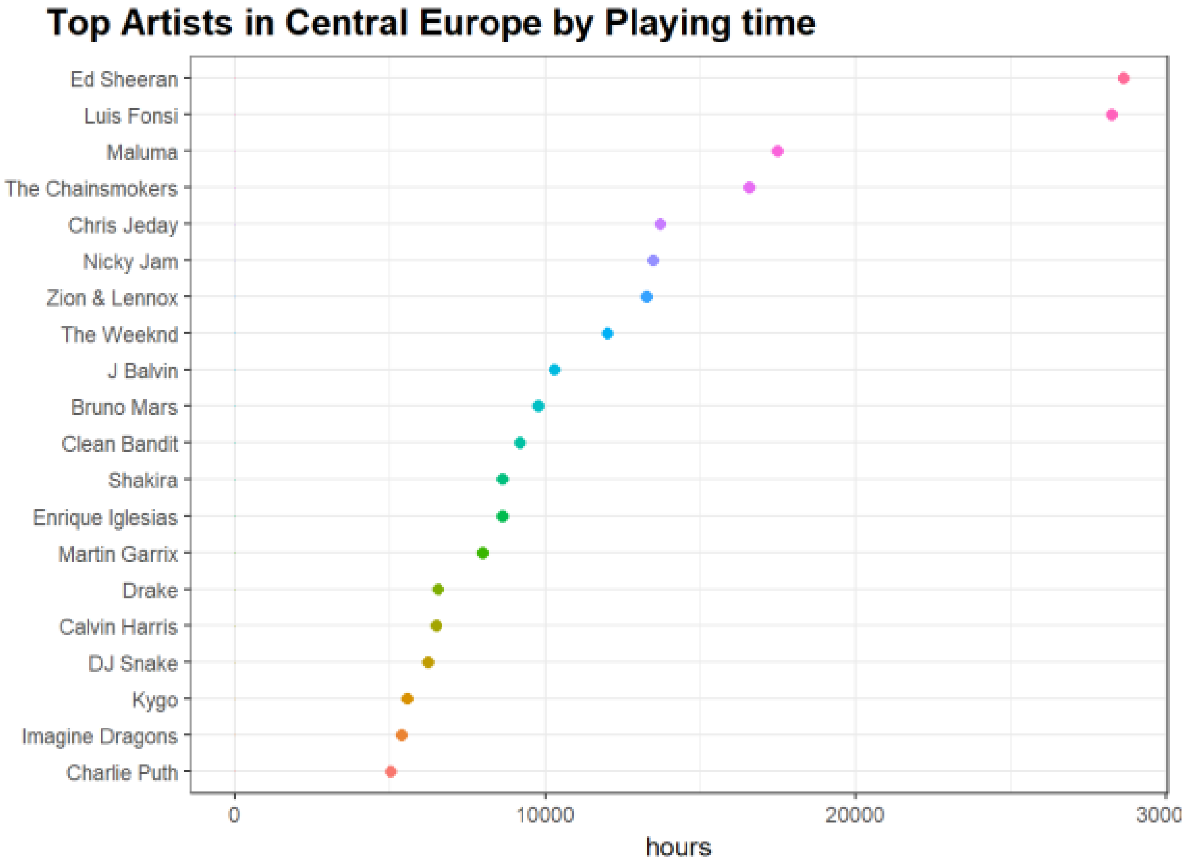
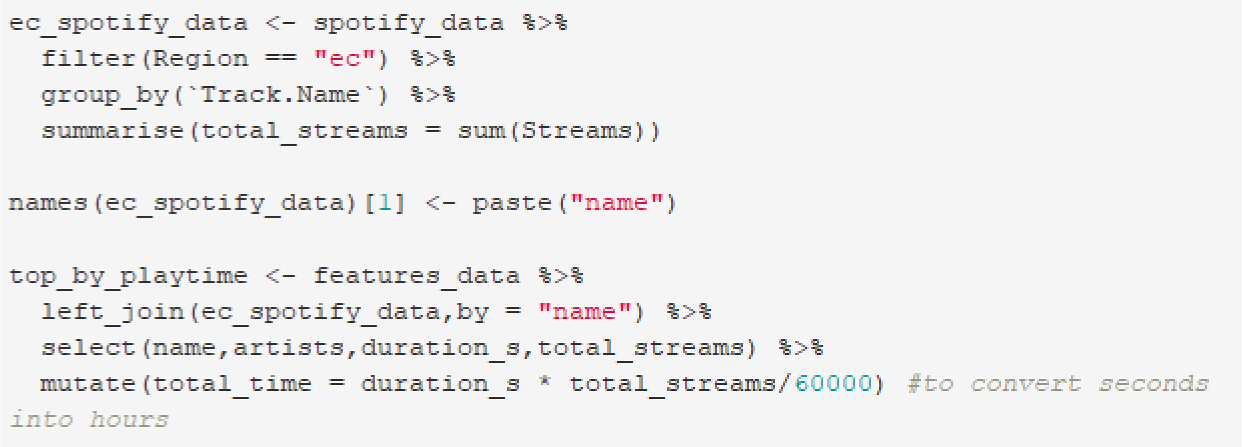


Figure 9. Top Artists in Central Europe by paying time

As we can see on the above graph the artist Ed Sheeran was played the most by the users, reaching nearly 30000 hours played, followed by artist Luis Fonsi and so on going down.

From the graph we can see that artists like Lois Fonsi and Maluma and Chris Jeday now started to be on top of the list.

R code for figure 9:

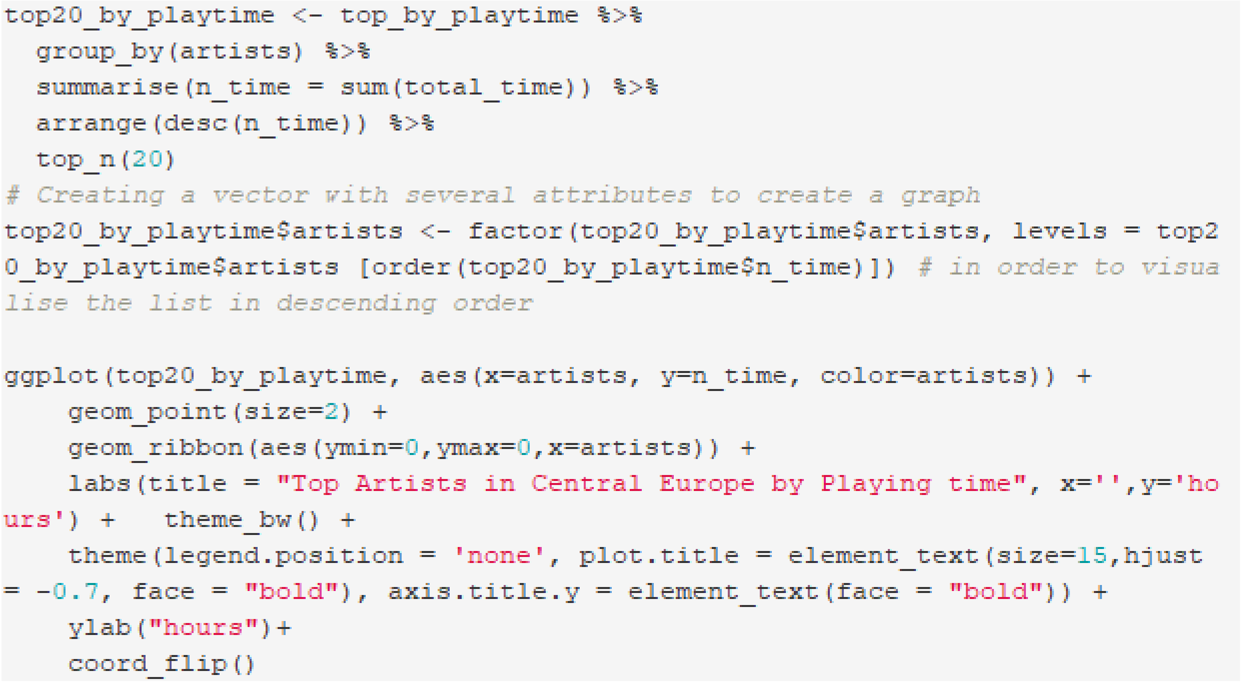


```
ec_spotify_data <- spotify_data %>%
  filter(Region == "ec") %>%
  group_by(`Track.Name`) %>%
  summarise(total_streams = sum(Streams))

names(ec_spotify_data)[1] <- paste("name")

top_by_playtime <- features_data %>%
  left_join(ec_spotify_data,by = "name") %>%
  select(name,artists,duration_s,total_streams) %>%
  mutate(total_time = duration_s * total_streams/60000) #to convert seconds
into hours
```

Top Artist by playing time in this case we will use the “Streams” variable of Spotify_data and “duration” of features_ data, data sets.



```
top20_by_playtime <- top_by_playtime %>%
  group_by(artists) %>%
  summarise(n_time = sum(total_time)) %>%
  arrange(desc(n_time)) %>%
  top_n(20)
# Creating a vector with several attributes to create a graph
top20_by_playtime$artists <- factor(top20_by_playtime$artists, levels = top2
0_by_playtime$artists [order(top20_by_playtime$n_time)]) # in order to visua
lise the list in descending order

ggplot(top20_by_playtime, aes(x=artists, y=n_time, color=artists)) +
    geom_point(size=2) +
    geom_ribbon(aes(ymin=0,ymax=0,x=artists)) +
    labs(title = "Top Artists in Central Europe by Playing time", x='',y='ho
urs') +    theme_bw() +
    theme(legend.position = 'none', plot.title = element_text(size=15,hjust
= -0.7, face = "bold"), axis.title.y = element_text(face = "bold")) +
    ylab("hours")+
    coord_flip()
```

Here at the first code we create a new list with the name top20_by_playtime, by using the list top_by_playtime we group by the artist’s name, create a variable n_time where we save the sum from the total_time and arrange by descending

D. Case

In the below graph I visualized finding the relationship between the variables of the features_data dataset.

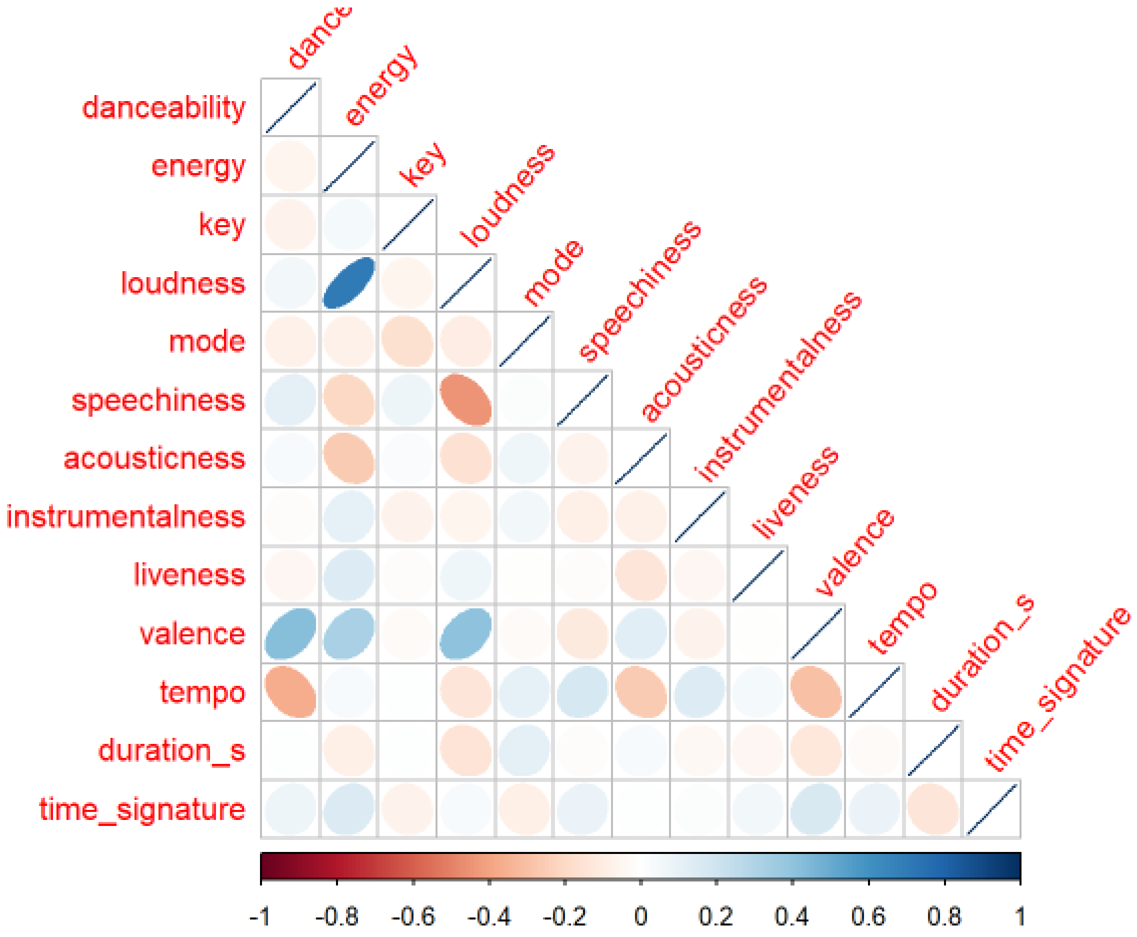


Figure 10. The relationship between the variables

After finding the relationship between the variables we can conclude that “energy” and “loudness” are highly positively correlated. Also “valence” is more correlated with “danceability, energy and loudness”. The variable “loudness” and “Speechness”, “Tempo” and “danceability”, “tempo” and “valence” are negatively correlated.

R code for figure 10:

```
library(corrplot)
features_data_num <- features_data[,-(1:3)]
mtCor <- cor(features_data_num)
corrplot(mtCor, method = "ellipse", type = "lower", tl.srt = 50)
```

Here we will use the library corrplot for visualizing the relationship between the variables. Will create a new variable fetures_data_num and will avoid the three firs columns because the firs columns are not numeric and include the name of the artist, song etc. And then we will use the method “ellipse” of type “lower”.

E. Case

Here is the distribution of the above 3 variables and how are they distributed among top 100 songs.

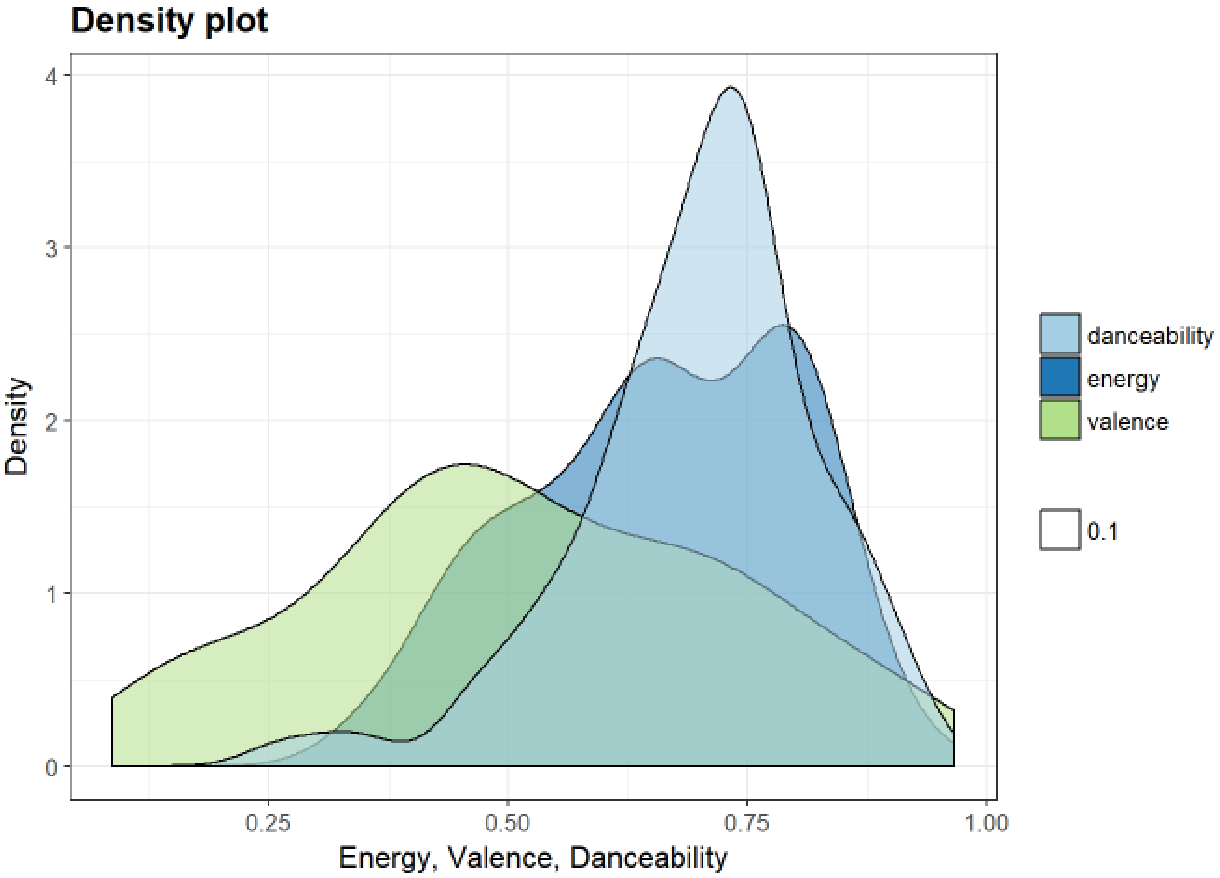
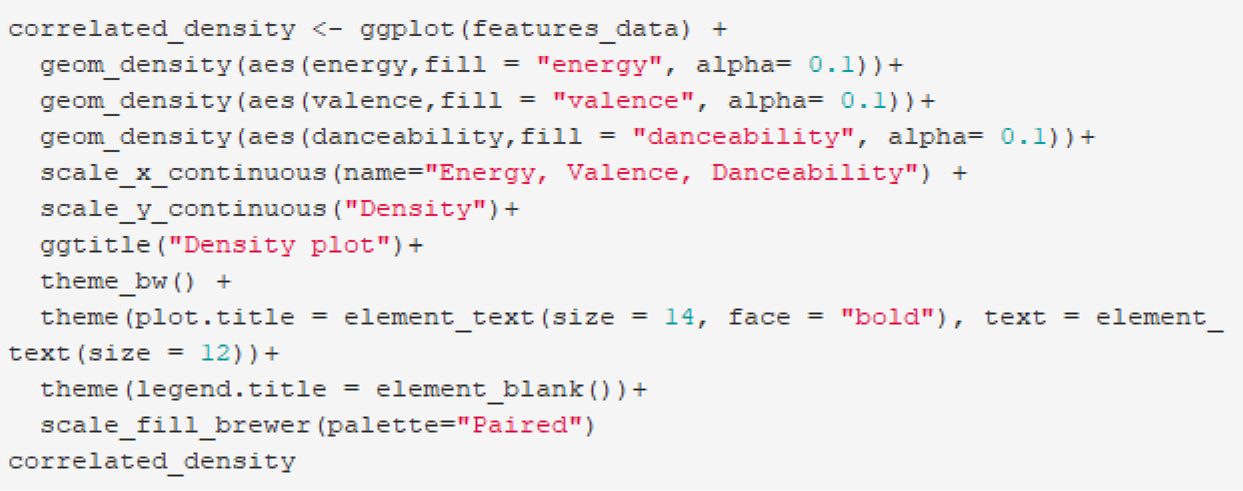


Figure 11. The distribution of Energy, Valence and Danceability

R code for figure 11:



```
correlated_density <- ggplot(features_data) +
  geom_density(aes(energy,fill = "energy", alpha= 0.1))+
  geom_density(aes(valence,fill = "valence", alpha= 0.1))+
  geom_density(aes(danceability,fill = "danceability", alpha= 0.1))+
  scale_x_continuous(name="Energy, Valence, Danceability") +
  scale_y_continuous("Density")+
  ggtitle("Density plot")+
  theme_bw() +
  theme(plot.title = element_text(size = 14, face = "bold"), text = element_
text(size = 12))+
  theme(legend.title = element_blank())+
  scale_fill_brewer(palette="Paired")
correlated_density
```

First we create a variable correlated_density where this variable contains the ggplot of features_data datasets then we find the density of the variables energy, valence and danceability and we use the R studio features to design the graph depending on our preferences.

F. Case

Among all the 13 Keys that were appeared on this top 100 songs in Spotify by users we counted the amount of time each key song appeared the most.

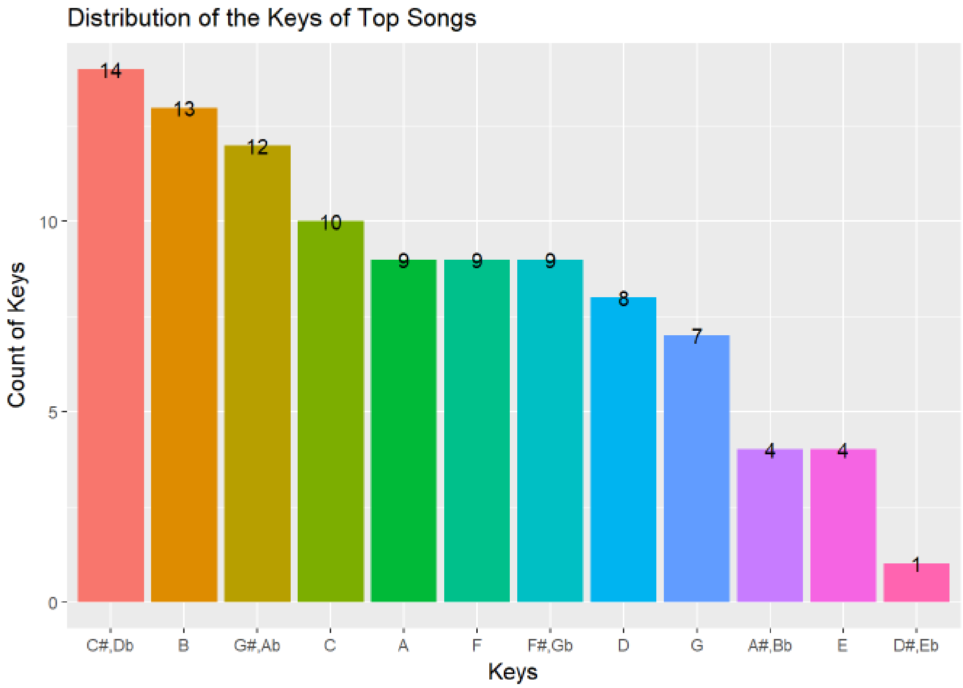
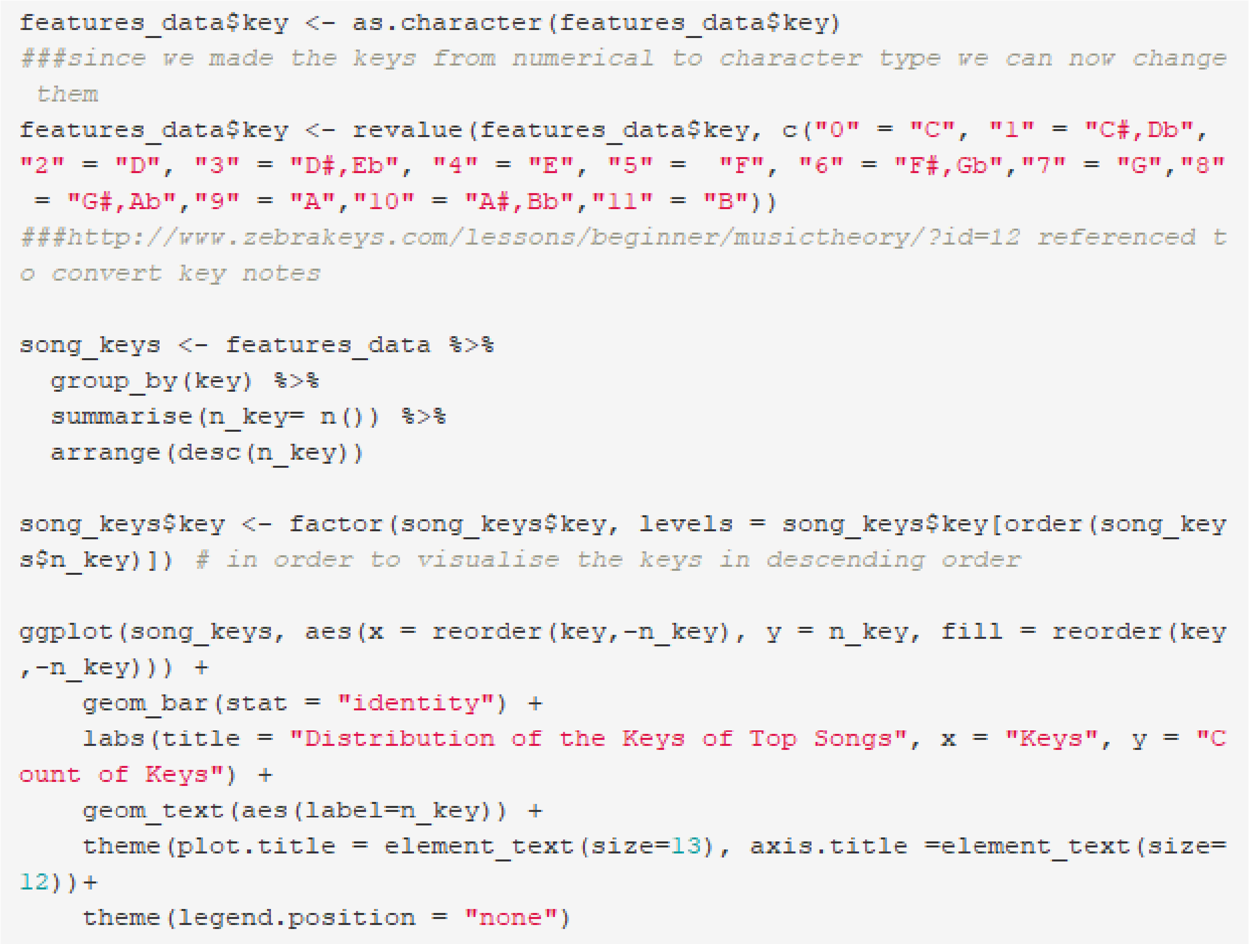


Figure 12. Distribution of the keys of top songs

R code for figure 12:



```
features_data$key <- as.character(features_data$key)
###since we made the keys from numerical to character type we can now change
 them
features_data$key <- revalue(features_data$key, c("0" = "C", "1" = "C#,Db",
"2" = "D", "3" = "D#,Eb", "4" = "E", "5" =  "F", "6" = "F#,Gb","7" = "G","8"
 = "G#,Ab","9" = "A","10" = "A#,Bb","11" = "B"))
###http://www.zebrakeys.com/lessons/beginner/musictheory/?id=12 referenced t
o convert key notes

song_keys <- features_data %>%
  group_by(key) %>%
  summarise(n_key= n()) %>%
  arrange(desc(n_key))

song_keys$key <- factor(song_keys$key, levels = song_keys$key[order(song_key
s$n_key)]) # in order to visualise the keys in descending order

ggplot(song_keys, aes(x = reorder(key,-n_key), y = n_key, fill = reorder(key
,-n_key))) +
    geom_bar(stat = "identity") +
    labs(title = "Distribution of the Keys of Top Songs", x = "Keys", y = "C
ount of Keys") +
    geom_text(aes(label=n_key)) +
    theme(plot.title = element_text(size=13), axis.title =element_text(size=
12))+
    theme(legend.position = "none")
```

For the sake of analysis, I converted the keys into their original symbols. First, here we create a variable fetures_data$key where we convert the keys into their original symbols. After that we create another variable where we summarize all the keys and arrange by descending order. We then use the ggplot to show the distribution of the keys and here from the graph we can see that the note C# was the one that appeared the most on the top 100 list of Spotify songs played by users.

G. Case

Here at the graph below we will use Decision tree analysis to see which of the songs have the highest chances to be on the top 100 list payed by users.

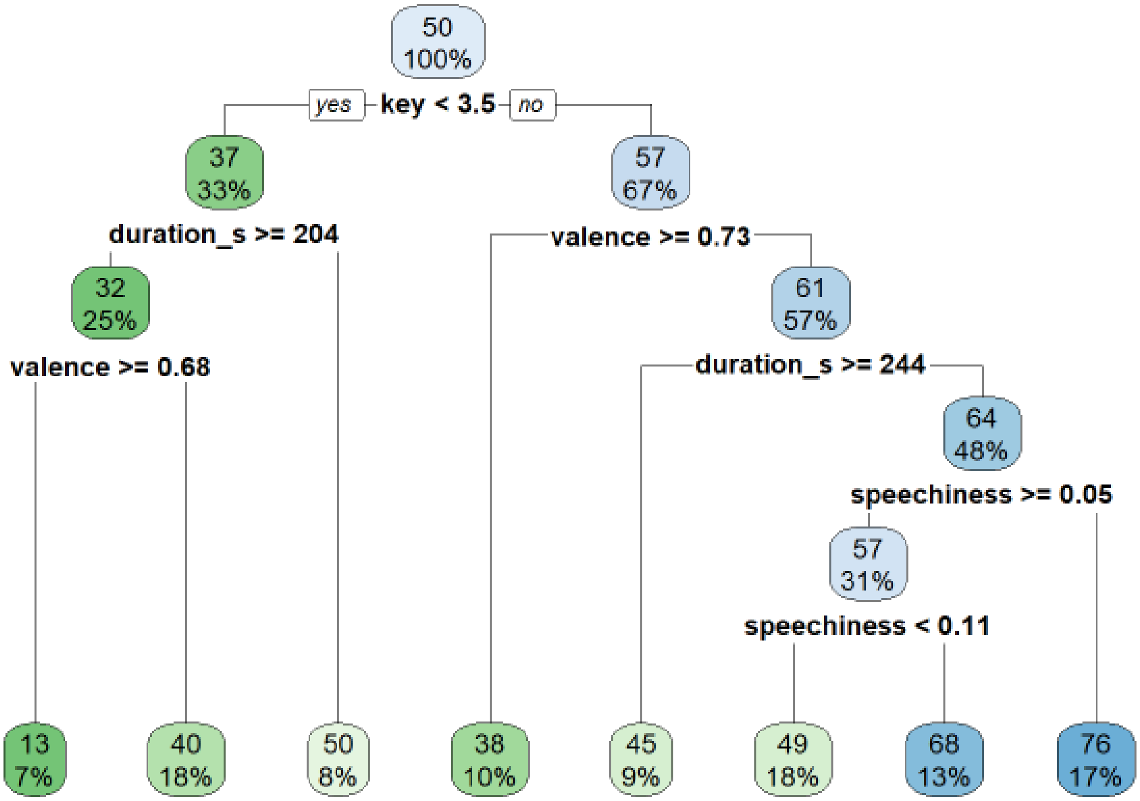
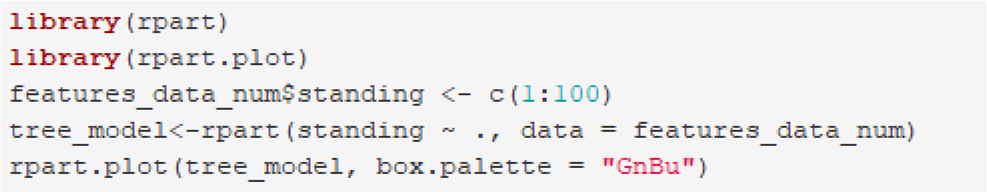


Figure 13 Decision Tree

R Code for figure 13:



```
library(rpart)
library(rpart.plot)
features_data_num$standing <- c(1:100)
tree_model<-rpart(standing ~ ., data = features_data_num)
rpart.plot(tree_model, box.palette = "GnBu")
```

First, for conducting a decision tree analysis we need two libraries rpart and rpart.plot. Then, create a list features_data_num$standing with all top 100 songs and then, create a variable tree model where we generate the decision tree model.

5 CONCLUSION

- From the tree we can see that the songs on which the key values are less than 3.5 (‘C’, ’C#’, ’D’, ’D#’, ’Eb’),
- Duration is more than 204 seconds and
- Valence is more than 0.68 have highest chance to be around the Top 100 list.

REFERENCES

[1] [Online]. Available: https://en.wikipedia.org/wiki/Data_mining. [Accessed June 2017].

[2] A. Rieger, "Large scale data analysis and predictive modeling in data mining," 07 11 2011. [Online]. Available: http://blog.bosch-si.com/categories/technology/2011/11/large-scale-data-analysis-and-predictive-modeling-in-data-mining/.

[3] B. Dykes, "Forbes," 31 March 2016. [Online]. Available: https://www.forbes.com/sites/brentdykes/2016/03/31/data-storytelling-the-essential-data-science-skill-everyone-needs/#2b29d1e252ad. [Accessed June 2017].

[4] M. Gladwell. The formula. The New Yorker, October 16, 2006

[5] W. Goldman. Adventures in the Screen Trade. Warner Books, New York, 1983.

[6] Augur, H. (2016, May 30). Will Big Data Write The Next Hit Song? Retrieved from http://dataconomy.com/2016/01/will-big-data-write-the-next-hit-song/

[7] Spotifycharts.com. (2018). Spotify Charts. [online] Available at: https://spotifycharts.com/regional [Accessed 24 Jun. 2018].

[8] GeorgeMcIntire. “Spotify Song Attributes | Kaggle.” Countries of the World | Kaggle, 4 Aug. 2017, www.kaggle.com/geomack/spotifyclassification/data