Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

THE POWER OF MUSIC

Proyecto final de primer año

Grado en Ciencia de Datos

Autores: Diego Cotino Bolufer
Miguel Camuñas Castelló
Marcos Carrasco Panadero
Pablo Gil Martínez
Sergio Samaniego Hernández

Tutor: Jose Manuel Gil

1ºB1

18/06/2024

Annexes

Annex A

In the preprocessing stage, the data is prepared for the analysis and for its modeling, for this it will be carried out an analysis of its content.

Identification and treatment of outliers

In this stage the outliers have been analysed and treated with different tools. Fisrt of all, the values were ordered and the values that where out of range

Outliers observed in the sample will then be identified and addressed. The otuliers are only present on numerical variables so they will be worked with our 7 numerical variables with the objective of finding outliers.

The first variable that will be analyzed will be **Hours per day** because in this variable can be a lot of outliers and values out of range, out of the 24 hours in one day. The second one will be **BPM**, the third one will be **Age** and the following ones will be **Anxiety**, **Depression**, **Insomnia** and **ODC** in this order.

Hours per day:

First of all, it will be seen what values don’t make sense because they are out of range. The individuals with this values will be deleted. The individuals will be ordered as a descending form and the values bigger or equal than 24 will be deleted.

	Timestamp	Age	Primary streaming service	Hours per day
				24 hours
	Texto	Numérico	Texto	Numérico
1	8/29/2022 2:40:16	23	Other streaming service	45
2	8/27/2022 21:40:40	61	YouTube Music	25
3	8/28/2022 11:55:54	37	YouTube Music	25
4	8/28/2022 14:24:10	16	Spotify	25
5	8/28/2022 19:18:53	42	YouTube Music	25
6	8/28/2022 19:56:46	25	Apple Music	25
7	8/29/2022 2:49:37	16	Spotify	25
8	8/29/2022 9:07:42	13	Spotify	25
9	9/1/2022 19:44:33	71	I do not use a streaming service.	25
10	9/13/2022 0:48:14	42	Other streaming service	25
11	8/27/2022 23:40:55	17	Spotify	24
12	8/29/2022 12:32:30	16	Spotify	24
13	9/28/2022 17:25:48	89	Spotify	24
14	8/29/2022 2:46:27	27	Spotify	20
15	10/23/2022 20:50:27	18	Apple Music	18
16	8/29/2022 9:42:23	18	Spotify	16
17	8/27/2022 19:57:31	63	Pandora	15
18	8/28/2022 18:03:50	25	Spotify	15

Figure 1

After that the outliers identification will start.

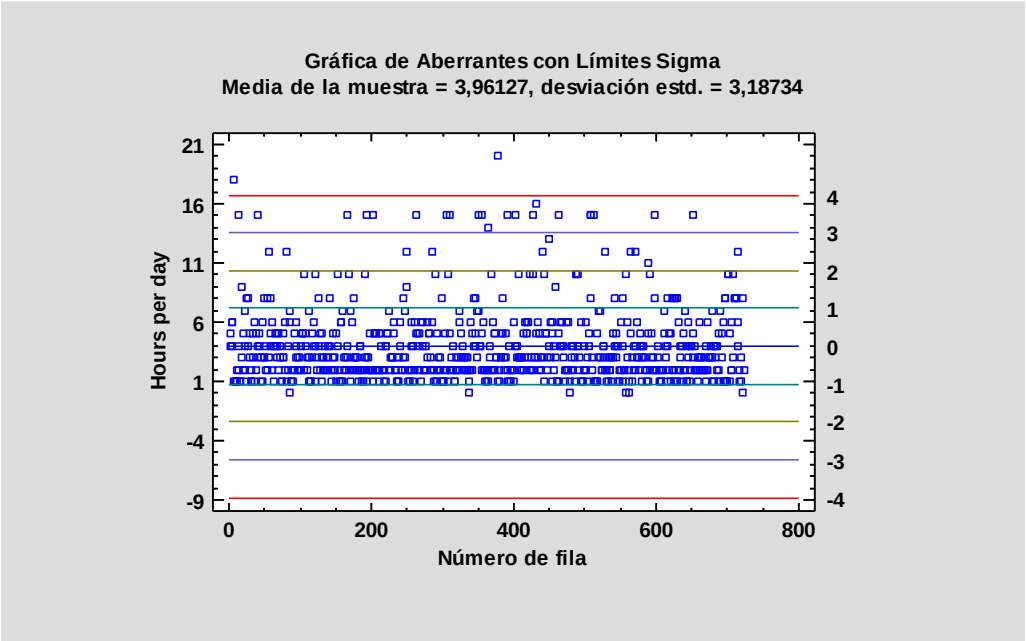

Figure 2

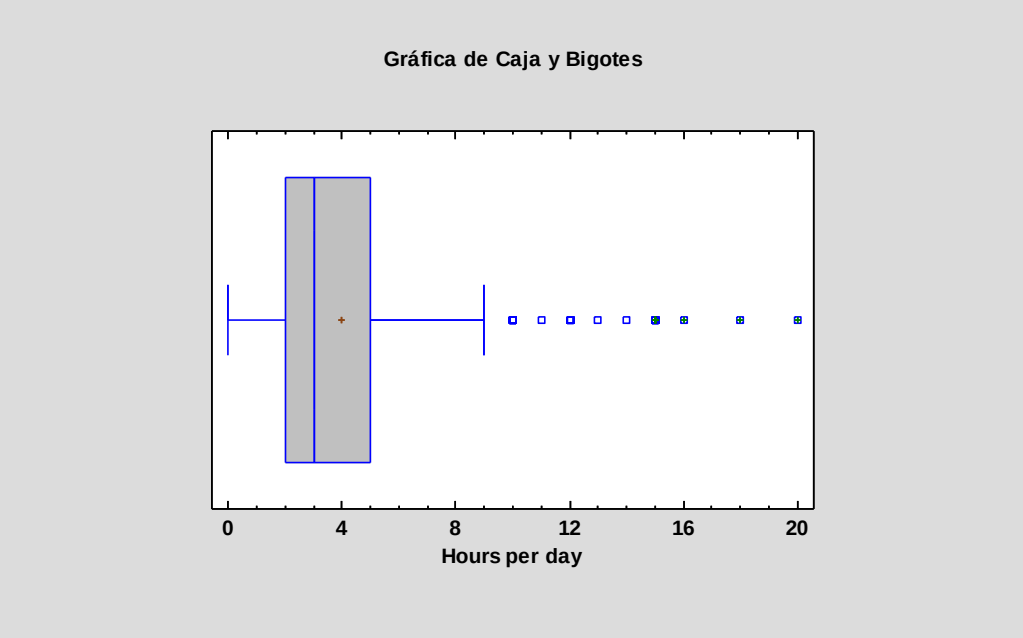

Figure 3

After seeing this graphs it can be said that there are a positive asimetry and there are extreme values that are not representative. So the values biggers or equal to 14 will be changed for the median because they are not representative. After this process it will be checked again.

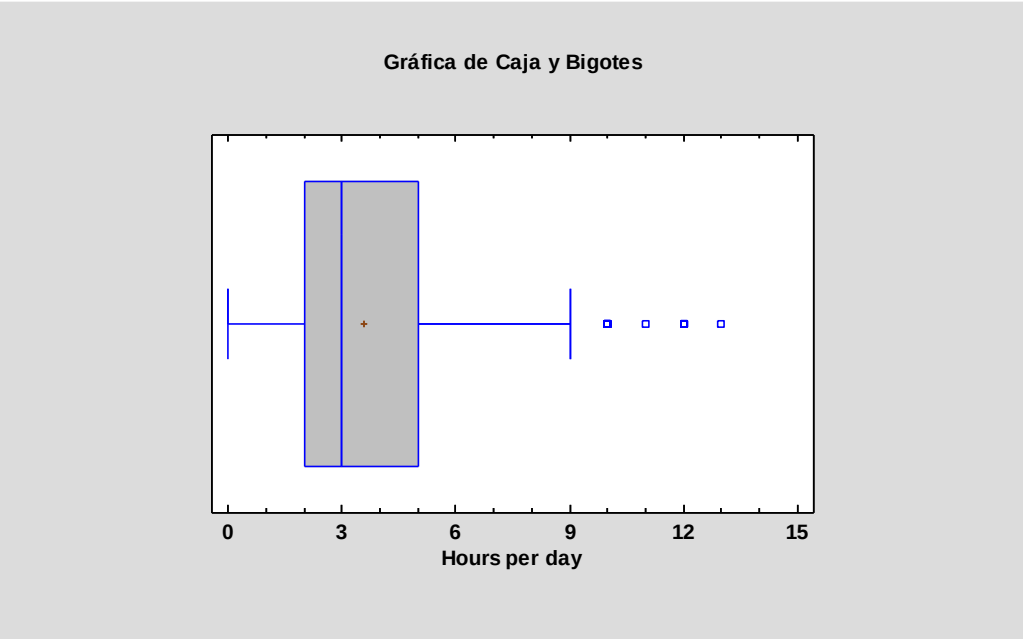

Figure 4

How it can be seen there are extrem values yet so it will be changed the individual with hours biggers than 10 and it will be checked again.

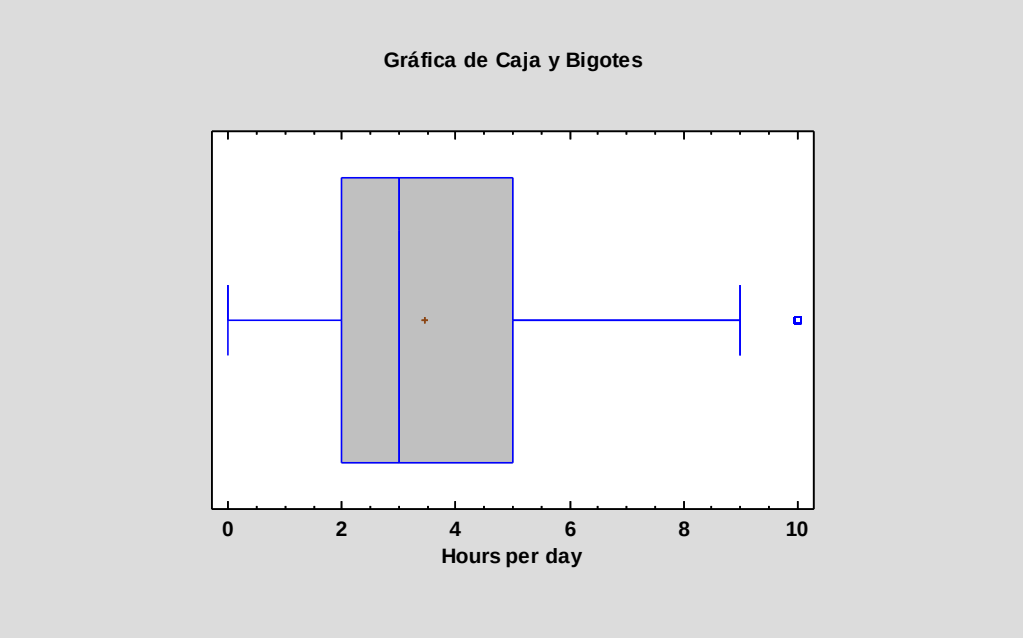

Figure 5

After this process the variable hours per day is treated.

BPM:

There are one value that don't make sense because it is imposible and this value affects all the sample parameters. So it will be deleted.

Foreign languages	BPM	Frequency [Cl
YES or NO		never, ra sometimes, frequen
Texto	Numérico	Texto
lo	999999999	Never
es	624	Sometimes
es	220	Rarely
es	220	Rarely
es	218	Sometimes
es	210	Never

Figure 6

After that it will be done the identification of outliers and extreme values.

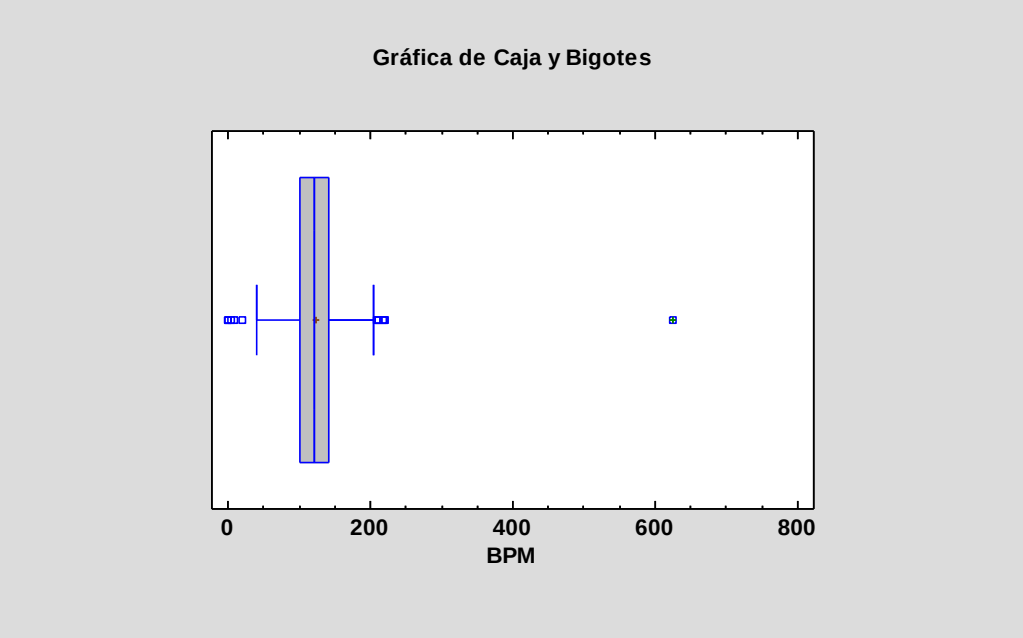

Figure 7

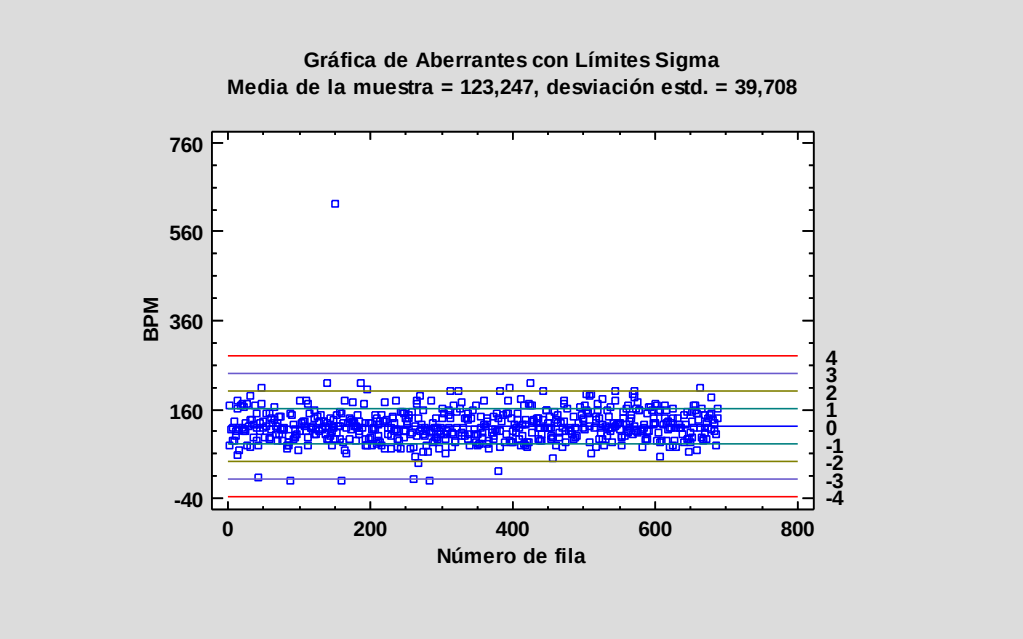

Figure 8

After seeing this graphs it can be seen that there is an extreme value that is not representative from above and there are also some values that are not representative under. So to avoid eliminating all the individuals the values will be changed for the median. After that, the variable is already treated.

Age:

There are not values that don’t make sense so it will be done the identification of outliers and extreme values.

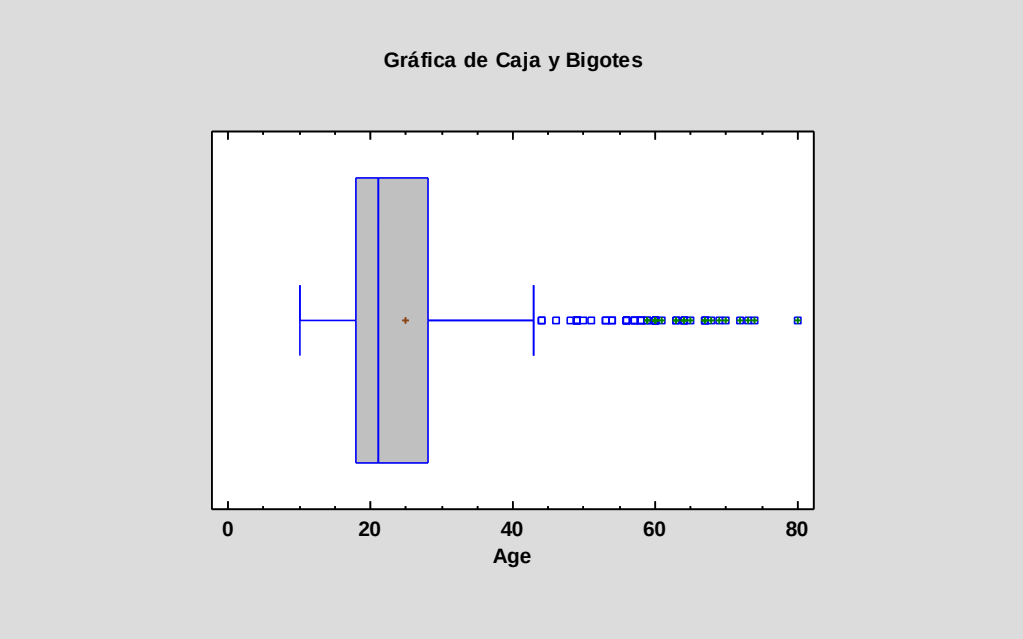

Figure 9

It can be seen that there is a positive asimetry, so how the values make sense and this variable will be discretized now it won’t receive changes by our team.

Anxiety:

After seeing the variable and order it can be seen that there is a value that is out of range. This value is 75 but the range is (0 – 10) so it will be deleted.

ly, sometimes, very requently	0 - 10	
Texto	Numérico	
	75	5
	10	10
	10	3
	10	9
	10	9
	10	8
	10	10
ıtly	10	7

Figure 10

With this process the variable is ready to be analyzed by the team.

Depression:

After seeing the variable and order it can be seen that there are two values that are out of range. These values are 35 but the range is (0 – 10) so both will be deleted.

ısic]	Anxiety	Depression	Insomnia	OCD	
very	0 - 10	0 - 10	0 - 10	0 - 10	Wo
	Numérico	Numérico	Numérico	Numérico	
	10	35	5	10	Ir
	7	35	35	55	Ir
	10	10	6	6	Ir
	10	10	10	8	Ir
	10	10	8	10	Ir

Figure 11

ısic]	Anxiety	Depression	Insomnia	OCD	
very	0 - 10	0 - 10	0 - 10	0 - 10	Wo
	Numérico	Numérico	Numérico	Numérico	
	10	35	5	10	Ir
	7	35	35	55	Ir
	10	10	6	6	Ir
	10	10	10	8	Ir
	10	10	8	10	Ir

Figure 12

With this process the variable is ready to be analyzed by the team.

Insomnia:

In insomnia there are not extreme values or outliers so it won't be changed.

ODC:

After seeing the variable and order it can be seen that there is a value that is out of range. This value is 85 but the range is (0 – 10) so this individual will be deleted.

Numérico	Numérico	Texto
	85	Improve
	10	Improve

Figure 13

With this process all the variables are ready to be analyzed by the team.

Identification and treatment of missing values

After the identification and treatment of outliers the sample has **687** values. So with the tool data visor it can be seen if in the variables there are missing values. There are some missing values in the variables. In the following board will be each variable with its respective missing values.

Age	**Primary streaming service**	**Hours per day** *(24 hours)*	**While working** *(YES or NO)*
1	1	0	2
Instrumentalist *(YES or NO)*	**Composer** *(YES or NO)*	**Fav genre**	**Exploratory** *(YES or NO)*
3	1	0	0
Foreign language *(YES or NO)*	**BPM**	**Frequency [Classical]** *(never, rarely, sometimes, very frequently).*	**Frequency [Country]** *(never, rarely, sometimes, very frequently).*
3	101	0	0
Frequency [Edm] *(never, rarely, sometimes, very frequently).*	**Frequency [Folk]** *(never, rarely, sometimes, very frequently).*	**Frequency [Gospel]** *(never, rarely, sometimes, very frequently).*	**Frequency [Hip hop]** *(never, rarely, sometimes, very frequently).*
0	0	0	0
Frequency [Jazz] *(never, rarely, sometimes, very frequently).*	**Frequency [Kpop]** *(never, rarely, sometimes, very frequently).*	**Frequency [Latin]** *(never, rarely, sometimes, very frequently).*	**Frequency [Lofi]** *(never, rarely, sometimes, very frequently).*
0	0	0	0
Frequency [Metal] *(never, rarely, sometimes, very frequently).*	**Frequency [Pop]** *(never, rarely, sometimes, very frequently).*	**Frequency [R&B]** *(never, rarely, sometimes, very frequently).*	**Frequency [Rap]** *(never, rarely, sometimes, very frequently).*
0	0	0	0
Frequency [Rock]	**Frequency [Video games music]**	**Anxiety** *(0 – 10)*	**Depression** *(0 – 10)*

(never, rarely, sometimes, very frequently).	*(never, rarely, sometimes, very frequently).*		
0	0	0	0
Insomnia *(0 – 10)*	**Ocd** *(0 – 10)*	**Music effects** *(worsen, no effect, improve)*	**Permision and timestamp**
0	0	6	0

Figure 28

It will be started working with the most significant missing values case, the **BPM** missing values. It should be the case with more missing values because a lot of individuals don't know what **BPM** means. So, in this case, because there are a lot of individuals with this problem, all the missing values will be deleted with the finality to no affect the variable **BPM** in the analysis.

After that process the other missing values were practicaly disapeared but there are some yet. So this missing values will be deleted too.

The missing values and the outliers are correct, so now is time to recode the variables with the goal to obtain a perfect sample to work.

Recodification

Now the variables that need a transformation will be recoded to be analyzed correctly. The variables, **Age** and **BPM** will be discretized because they have a lot of different values but the original variable won't be deleted because the variables can be needed. In **BPM** the values will be discretized on intervals of 40 values and in **Age** the values will be discretized on intervals of 15 values. The new variables will be called **Age_recod** and **BPM_Recod**.

BPM intervals: (0 – 40], (40 – 80], (80 – 120], (120 – 160], (160 – 200], (200 – 240], (240 – 300]

Age intervals: (0 – 16], (16 – 30], (30 – 45], (45 – 60], (60 – 75], (75 – 90)

With this last process the preprocess has finished. The final sample will be a sample with 33 useful variables and with 586 individuals.

The mission of this project is to find whether or not the type of music that an user listens to influences their mental state. And, if so, to give a guide on how to develop a Spotify playlist with the objective of improving your mental health.