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

Pablo Gil Martínez

Sergio Samaniego Hernández

Marcos Carrasco Panadero

Miguel Camuñas Castelló

Tutor: Jose Manuel Gil

1ºB1

18/06/2024

Index

Introduction

On the street, at the public transport, at the university… People listens to music and use it as a method to escape their problems. According to Cristina Caron (2023) “music therapy has grown over the last decade.”[1] But does music really impact on mental health? Experts say yes, but how should people consum music to maximize their effects in mental wellness?

Motivation

This project is driven by the team's commitment to address the crucial issue of mental health. This topic is considered a public issue by important universities, like the university of Tulane (2021)[2], so contributting to this cause is an honour and a powerfull motivation. Obtaining significant results that can contributte to the topic is a pleasure and the team will be delighted to share the results with the society.

The idea about doing it about music came because it was an unexplored topic that was not talked about, outside the musical therapy. That is why, after looking for other variables, such as geographic conditions, it was finally decided by the team to work with a potential relationship with music.

Objectives

These are the goals that are aimed to fulfill:

- To find in the data analysis a series of potential relations between a certain music consumption habits and mental health.
- To develop a website that helps people to understand how music affects mental health.
- To transmit in our report how music consumption behavior could impact on the mental health of the user.
- To create a video about the project made, with its analysis, results and conclusions extracted.

Methodology

In order to make the project succesfully and obtain relevant results it is used the following methodology. First of all, the project has to be planned, with the topic, the objectives and the products the project wants to have. With the objectives and the topic defined the dataset has to be found, this dataset has to be relevant and it has to have the information needed to make all the project. For this project it was found a dataset that relates mental health with music (Music & Mental Health Survey Results), it was obtained from kaggle, a web platform where the data sciencie community share their projects and different tools like datasets.

After the dataset is chosen, the analysis of the variables to obtain results about our topics, starts. All the analysis is done with an statistical software called Statgraphics. This software was chosen because it is the best software to develop the project. The software has all the necessary tools and functions to make the projec,t and it has a visual and atractive form to show the information and the data.

Knowing the variables the preprocess stage can be done. In this stage the individuals and the data with outliers and missing values is transformed or deleted with the goal to have an easier source to analyse. The outliers are found with the *outlier identificator* tool that the software has, and are treated being deleated or being changed by the media depending on the case. The missing values are found by the *data visor* and are treated in the same way as the outliers. In the preprocessing stage some variables are also recodificated because it is easier to make correlations with intervals instead of different values.

When the data is ready to be analysed, the unidimensional analysis begun. In this stage each variable is going to be analysed to find patterns in its behavior. This analysis is done in stargraphics too using the unidimensional analysis tools. Depending on if the variable is categorical or numerical the analysis is done different, using different visuals like *Box and Whiskers* or *the probability graph* and using different parameters.

With the unidimensional analysis finished the multidimensional analysis is carried out. The multidimensional analysis consists in compare different variables in order to find relations. For this part it is used the multidimensional tools that the software has. The *Two factors* funcionality and the *Subgroup analysis*. The variables with relations and correlations are taken for making the conclusions because are the principal variables that are interesting.

After that, the next thing that has to be done is to formulate conclusions that emerge from the work. This conclusions are obtained from the results of the analysis, with the knowledge gained from the state of the art and other parts of the project such as the problem analysis. All the conclusions have to make sense and have to be relevant to the principal objectives of the project.

In order to transmit the results there are some products. To create these products it is key to start from the objectives and link it to what has been learned. The products are thinked as a form to transmit, so they should be understandable and visually attractive. They are a website, a report and a video that provide more depth to the project, and they are the principal form to communicate the project’s results.

Structure

The project is structured in several important parts that give it a cohesive and orederly form and a cohesion that makes the project more understandable to everyone.

The first part of the project is the introduction, where the project's topic and its problem is raised. This first section includes the motivation, the objectives, the methodology and the structure of the project. After the introduction, in the State of Art the problem on which the project focuses is analyzed in-depth, and it is proposed a possible solution to the problem after. After that, it is planteated the project scope where it is shown what the project will try to be. This part includes the products, the success criteria that is ,when the project is considered as a success, and the alignment with the sustainable development goals. Following the project scope is the working plan that is used during the project.

Regarding the data analysis, the data source and from where it comes is explained. With the data source undestood the analysis starts with the unidimensional analysis where each variable is analysed individually. After the unidimensional analysis the multidimensional analysis is explained, in this part the relations between the variables is shown in oreder to find evidences. After the analysis it is explained what tools were used during the process.

After that the results of the analysis and the conclusions are shown. Finally, it is displayed the bibliography and the annexes with additional information and more explanations about what has been done in the project.

State of art

Mental health has been a recurrent topic of debate in recent years for several reasons. Firstly, it has gone from being a taboo subject in certain areas and social groups to being something on everyone's lips, in addition in recent times there has been an increase in problems in this area mainly among young people, which is a compelling reason to carry out a project focused on this issue. This project is going to analyse whether there is a relationship between the music listened to and the state of mental health, making use of a large database with different variables related to the amount of music consumed, as well as the specific styles and the state of mental health, focusing on different related problems.

According to Amanda MacMillan (2023) "more people in the US are living with mental and emotional distress"[3]. This may extend globally, as Andrew Moose and Ruma Bhargava (2024) point out "Since 2006, levels of reported youth happiness have declined in North America, South America, Europe, South Asia, the Middle East and North Africa."[4] Areas which concentrate the vast majority of the world's population, so it can assumed that this is a global problem.

Due to the increase in mental health problems, investigations have been carried out to examine the potential factors contributing to this rise. Among the many factors under consideration, MacMillan A. (2023)[5] identifies four key ones: Social media usage, Covid-19, Isolation and Loneliness, and Lack of access to care. Being the first one something that only increases due to modern era and causes problems mainly because over-comparison with what is seen on social media that often is not true. The other main factor to note is the third: Isolation and loneliness, which are enhanced factors due to social trends like decreased community involment and fewer people starting a family.

In recent times, other factors that may affect mental health to a lesser extent have been studied, among which is the influence of music, whether the person suffering from mental health problems sings or plays it or simply listens to it. Even therapies based on this, known as music therapy, have been developed.

As Priyanjana Pramanik wrote in her recent article in medical.net (2023) music therapies have been used for physical recovery and in motor rehabilitation. However, the relationship between music and mental health continues to be poorly understood even though listening to music has been associated with mental health improvements especially in cognitive function.[6] It is evident in the scientific community that there is a consensus on the need for further development of music therapy.

These music therapies are based on the neurological effect that music has on people, as Alison Pearce Stevens (2024) explains "Highly emotional music causes networks in the brain to release dopamine. This brain chemical, a type of neurotransmitter, plays a role in feelings of pleasure."[7] This, in addition to Fatima Reynold's explanation of what music allows individuals to express "Through music, individuals can express their unique experiences, struggles, and triumphs, forging connections with others who share similar backgrounds. Research has shown that exposure to diverse musical genres and artists can broaden perspectives, challenge stereotypes, and foster empathy among listeners especially when dancing together"(2023).[8] So from a neurological such as from a social point of view experts outline the effect of music on mental health.

Other benefits of music for mental health have been outlined by Lynne Gilmour (2024)[9] in a recent article. The project developed at the University of Stirling outlines the health benefits of live music. They propose to make it easier for young people with mental health problems to access live music events, as their condition is greatly improved when they are able to participate in such events.

In addition, medical studies have shown improvements in patients with various diseases such as schizophrenia, Parkinson or Alzheimer, as highlighted by Pearce Stevens (2024).[10] As far as can be seen, music is beneficial not only to face mental health problems but also diseases that affect memory or motor skills as mentioned above.

Problem analysis

The state of the art has shown that experts agree that there is an increase in mental health problems, such as depression, in society (Goodwin et al. 2022)[11]. And that is why it is considered a problem to be addressed.

In addition, the state of the art has shown several relationships between music and mental health, but all of them are focused from the point of view of treatments and not on how music consumption affects mental health.

It is well known that people use music to improve their mood. Is this measure sufficient? In short, the problem is that society is not aware of the extent to which music affects their mental health and society don't know how to use the music properly to obtain all the benefits that brings to.

Proposed solution

A bad mental health, like any other problem, can be remedied in several ways. The State of Art has shown that music therapy contributtes positively to mood and other diseases like Parkinson or Alzheimer. People know the power of music but again, it is not known in which grade consumption habits affect wellness. If the *How* is not explained, how would people act in consequence to improve their lives? Said this, the solution wouldn't be to find evidence whether music affects mental health. Instead, it would be to find correlations between the characteristics of music listeners and their habits with factors like anxiety, insomnia or depression. To do so, a database which cointains info about the listeners stats according to music and mental health is being analysed.

An important part of this solution will be the way it is presented. If people has not an accesible way to get the conclusions the efforts are done for nothing. For this reason, a web page will be made to show all parts of the project in an eye-catching way.

Project scope

Deliverables (products)

The project is formed by the following three products:

-The most representative is the official web page of the project as it cointains the rest of the products. It will be done with HTML and CSS. It is meant to be the an aesthetic way to deliver the correlations that the team has found. It is a public domain web because its goal is to reach as many people as posible. In order to let the user know how to flow throught the website, it contains an userguide to indicate what to see in order. The website also tells about the team, its motivacion and where they come from.

-The website offers a downloadable report. It is made by the team and it is the heart that exlpains the project's fundamentals. All the processes, methodology, analysis and conclusions about the project are explained from the beggining of the project, together with a presentation. It is thought for people who are used to statistic terms and knwoledge.

-One of the website's pages shows the promotional video. It is a short 16:9 video edited with free Software. This video's goal is to help the website reach as many people as posible as the video is what is being sent when the project launches. It explains the general aspects and serves likea a tutorial to teach users how to use the web properly.

Success criteria

It can be said that success has been achieved if the proposed objectives are achieved or fulfilled. To comfirm that they have been achieved, they have been more measurable and quantifiable, and are the following:

- To find in the data analysis at least three potential correlations that may occur between variables.
- To develop a website that helps people to understand how music affects mental health that provides value and information to the user, having at least 250 visits by the due date.

Alignment with the Sustainable Development goals (SDG)

ALARGAR

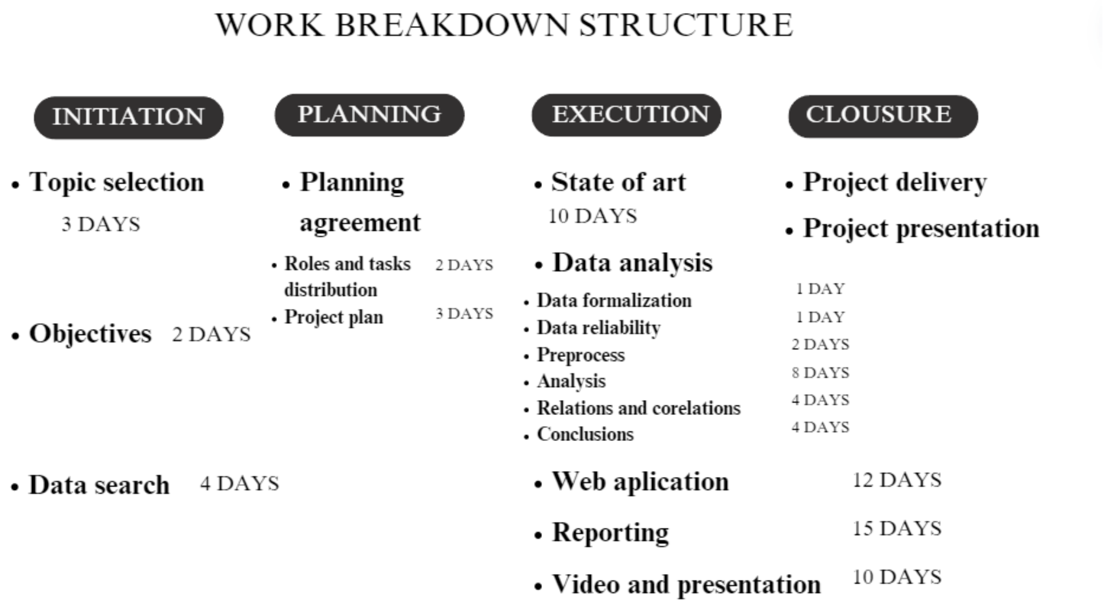
The project is aligned with the Sustainable Development Goal number 3 for the UPV: "Health and Wellbeing". This is because the team has been investigating the correlation between music taste and mental health, and the latter is a very spoken about topic in recent times, especially regarding youth's mental health.

The goal of the team, at the end of the day, is to contribute to improving all people's mental health, in one way or another, fixating on underrated factors such as the music listened to.

Working plan

When the project started, a plan was defined with the objective to serve as a guide to be followed upon, that uses a breakdown structure and a Gantt chart. This planning has changed over time redefining the times and the Gantt chart with it. The following graphs are the last Gantt chart and the last breakdown structure.

This figure is actualized with the information about the project in the final stage. The work breakdown structure has practicaly the same parts that it had when the project started, but some changes were done. Some sections from web aplications were removed because this parts were considered unnecesary and far from the principal purpose of the project, and now is only one principal section. The time of the state of art and the time of the web aplication were reduced and the time of all the parts of initation and the time of reporting were expanded.



WORK BREAKDOWN STRUCTURE

INITIATION

- Topic selection 3 DAYS
- Objectives 2 DAYS
- Data search 4 DAYS

PLANNING

- Planning agreement
  - Roles and tasks distribution 2 DAYS
  - Project plan 3 DAYS

EXECUTION

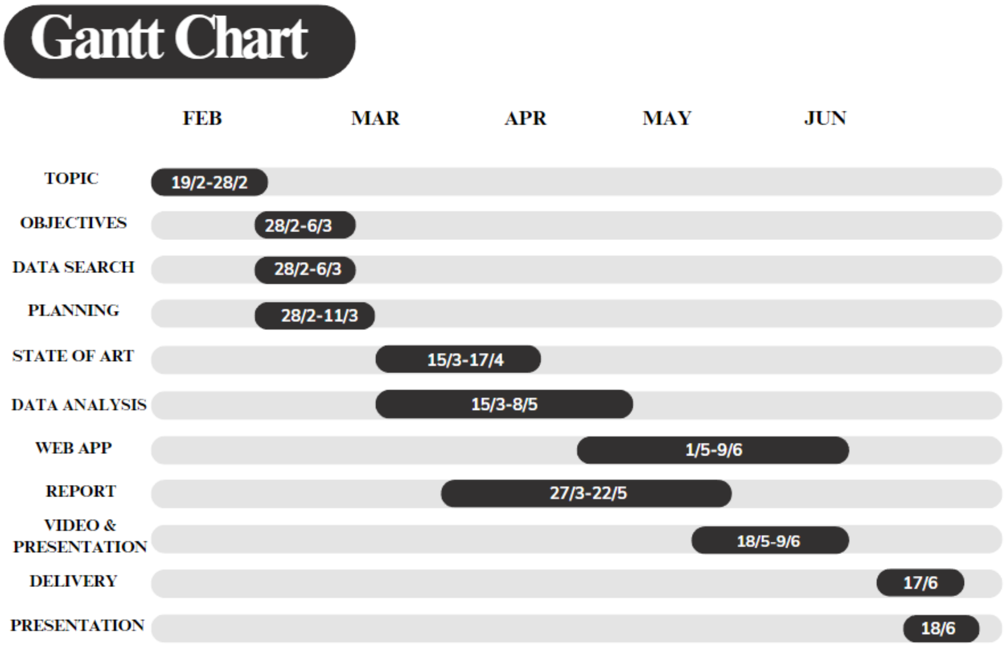
- State of art 10 DAYS
- Data analysis
  - Data formalization 1 DAY
  - Data reliability 1 DAY
  - Preprocess 2 DAYS
  - Analysis 8 DAYS
  - Relations and corelations 4 DAYS
  - Conclusions 4 DAYS
- Web aplication 12 DAYS
- Reporting 15 DAYS
- Video and presentation 10 DAYS

CLOUSURE

- Project delivery
- Project presentation

The Gantt Chart has also been updated during the project, as it has been said before this is the last Gantt Chart that has been made which represents the project timings better than the ones made before.

As can be seen in figure 2, the project has 3 distinct phases after the choice of topic. On 28 February the first phase started, in which the objectives, data research and planning were carried out. Then stated the second phase which is composed of the state of art, the data analysis and the report, this last started later due to it depends on the previous ones. The last phase could be call the products one because it consisted in the development of the web app, the video and the presentation.



Gantt Chart

TOPIC	19/2-28/2
OBJECTIVES	28/2-6/3
DATA SEARCH	28/2-6/3
PLANNING	28/2-11/3
STATE OF ART	15/3-17/4
DATA ANALYSIS	15/3-8/5
WEB APP	1/5-9/6
REPORT	27/3-22/5
VIDEO & PRESENTATION	18/5-9/6
DELIVERY	17/6
PRESENTATION	18/6

Preparation and data understanding

Introduction

The main part of the project is to analyze the data that the group has collected about the main topic. This analysis has the objective of finding relations and correlations in the variables to verify previously found information and to confirm the hypotesis the team had. The analysis started with the identification and reliability of the data source and its characteristics. It continued with the treatment of missing values and outliers and the analysis ended with the unidimensional and the multidimensional analysis. After this analysis the team drew conclusions.

Data source

DECIR NUMERO VARIABLES y mencionarlas cuando se habla de ellas, meter link y añadir la tabla de las variables de Pablo

The database was obtained from a website called "Kaggle", and it's name is "Music & Mental Health Survey Results". From all the other databases related to the topic, this one has by far the most upvotes. It is also very well rated, with lots of users opining that the database is clean and well-documented. Another good sign is that it has been downloaded more than 20 thousand times. This is the web page's link https://www.kaggle.com/datasets/catherinerasgaitis/mxmh-survey-results .

It is needed to keep in mind that some of the variables, precisely the ones about mental health disorders, are self-reported in a rating of 1 to 10, so there is a reason to watch out if aiming to a very precise analysis. There are also other details, such as the frequency of listening to a certain type of music, as it is a discrete variable with options such as "Sometimes", or "Never". There cannot be an exact measurement because It is not known if an answer such as "Very frequently" means 1 in every 2 songs, or 1 in every 5.

To summarize this, and taking into account everything discussed before, it has been considered to qualify this database as helpful, useful and trustable, despite the small inaccuracies that may come with this subjective way of responding. As said before, it is very likely that this database has also been helpful to other data analysts, and so it is believed that it will be useful to the team, too.

The content of the data file contains the information about 736 individuals and collects of each one, information about 33 different variables. The data file's content is structured in columns for the variables and in rows for the individuals.

The data file contains information about the using of music, his different genres and the feellings or sensations that the different individuals have. In this information there is the individual age, the streaming service, the visualization hours, if they listen to music while working, among other interesting variables for the analysis.

The following variables are the variable the data file has, their description, a comment and the variable type (if the variable is numerical or categorical). The variables in colour purple are the variables that were used by the team for the analysis:

Variables	Description	Comments	Variable type
Age	This variable offers the age o the individual.		Numerical
Primary streaming service	This variable offers the name of the principal streaming service the individual use to listen music.		Categorical
Hours per day	This variable offers the amount of hours the individual listens to music in one day.	*(24 hours)*	Numerical
While working	This variable define if the individual listens music while working.	*(YES or NO)*	Categorical
Instrumentalist	This variable define if the individual plays an instrument or not.	*(YES or NO)*	Categorical
Composer	This variable define if the individual compose music or not.	*(YES or NO)*	Categorical
Fav genre	This variable offers the favourite genre of the individual.		Categorical
Exploratory	This variable define if the individual actively explores new genres or artists.	*(YES or NO)*	Categorical
Foreign language	This variable define if the individual listens to music in other languages that he can't speak fluid or understand.	*(YES or NO)*	Categorical
BPM	This variable offers the beats per minute of the favourite genre.		Numerical
Frequency [Classical]	This variable define how frequently the individual listens to this genre.	*(never, rarely, sometimes, very frequently).*	Categorical
Frequency [Country]	This variable define how frequently the individual listens to this genre.	*(never, rarely, sometimes, very frequently).*	Categorical
Frequency [Edm]	This variable define how frequently the individual listens to this genre.	*(never, rarely, sometimes, very frequently).*	Categorical
Frequency [Folk]	This variable define how frequently the individual listens to this genre.	*(never, rarely, sometimes, very frequently).*	Categorical
Frequency [Gospel]	This variable define how frequently the individual listens to this genre.	*(never, rarely, sometimes, very frequently).*	Categorical
Frequency [Hip hop]	This variable define how frequently the individual listens to this genre.	*(never, rarely, sometimes, very frequently).*	Categorical
Frequency [Jazz]	This variable define how frequently the individual listens to this genre.	*(never, rarely, sometimes, very frequently).*	Categorical
Frequency [Kpop]	This variable define how frequently the individual listens to this genre.	*(never, rarely, sometimes, very frequently).*	Categorical
Frequency [Latin]	This variable define how frequently the individual listens to this genre.	*(never, rarely, sometimes, very frequently).*	Categorical

Frequency [Lofi]	This variable define how frequently the individual listens to this genre.	*(never, rarely, sometimes, very frequently).*	Categorical
Frequency [Metal]	This variable define how frequently the individual listens to this genre.	*(never, rarely, sometimes, very frequently).*	Categorical
Frequency [Pop]	This variable define how frequently the individual listens to this genre.	*(never, rarely, sometimes, very frequently).*	Categorical
Frequency [R&B]	This variable define how frequently the individual listens to this genre.	*(never, rarely, sometimes, very frequently).*	Categorical
Frequency [Rap]	This variable define how frequently the individual listens to this genre.	*(never, rarely, sometimes, very frequently).*	Categorical
Frequency [Rock]	This variable define how frequently the individual listens to this genre.	*(never, rarely, sometimes, very frequently).*	Categorical
Frequency [Video games music]	This variable define how frequently the individual listens to this genre.	*(never, rarely, sometimes, very frequently).*	Categorical
Anxiety	This variable offers the level of anxiety of the individual.	*(0 – 10)*	Numerical
Depression	This variable offers the level of depresion the individual thinks he has.	*(0 – 10)*	Numerical
Insomnia	This variable offers the level of insomnia of the individual.	*(0 – 10)*	Numerical
Ocd	This variable offers the level of Obsessive compulsive disorder the individual has.	*(0 – 10)*	Numerical
Music effects	This variable offers the level of music affects the individual.	*(worsen, no effect, improve)*	Categorical
Permision			Categorical
Timestamp			Categorical

The description specifies what the variable means.The comments were added with the goal to specify what values can the variables have. The variable type says if the variable is categorical or numerical.

Exploratory analysis and data quality

With the data source analysed the team knows how to work with the data obtained. In the preprocess the outliers and the missing values were found and treated with the objective of having a data file analysable to search the necessary information to the project. In this stage, some variables were recoded with the finality of providing different types of variables and different forms of the same variable.

After the identification of outliers, some of the individuals with outliers were deleted because the data didn't make sense, for example individuals that say that listen to music more than 24 hours per day. Other outliers that made sense but were no representative and were so far from the media were changed for this parameter.

There were some missing values. The individuals with this kind of values were deleted because individuals with no information were no representative to the project. After these processes the data file was left with 586 individuals compared to 736 at the beginning.

The variables **Age** and **BPM** (beats per minut of favourite genre) were discretized because in the analysis could be needed to be used these variables like qualitative ones to find relations. These new variables were called **Age_recod** (age recodificated) and **BPM_recod** (beats per minute of favourite genre recodificated).

With all the preprocesses done the data file has 586 individuals (rows) and 35 variables (columns) and the data analysis can be started. Of all these variables only 10 are going to be analysed in order to find the relations interesting to the project.

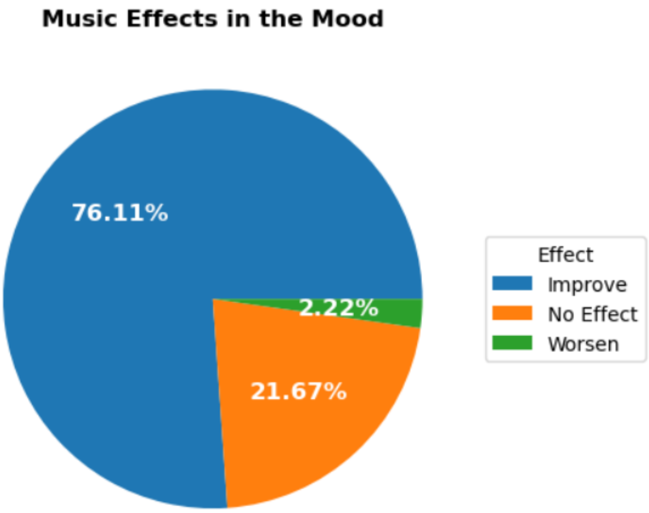
For more information about the preprocess, see the annex A.

Unidimensional Analysis

The variables that have been used in this project are related to two different themes, the first one is the music, and the second one is the mental health. Regarding the music, the most relevant variables are the hours per day hearing music, the quantity of people who are more likely to explore new type of music and the effect that music has in the mood of the individuals.

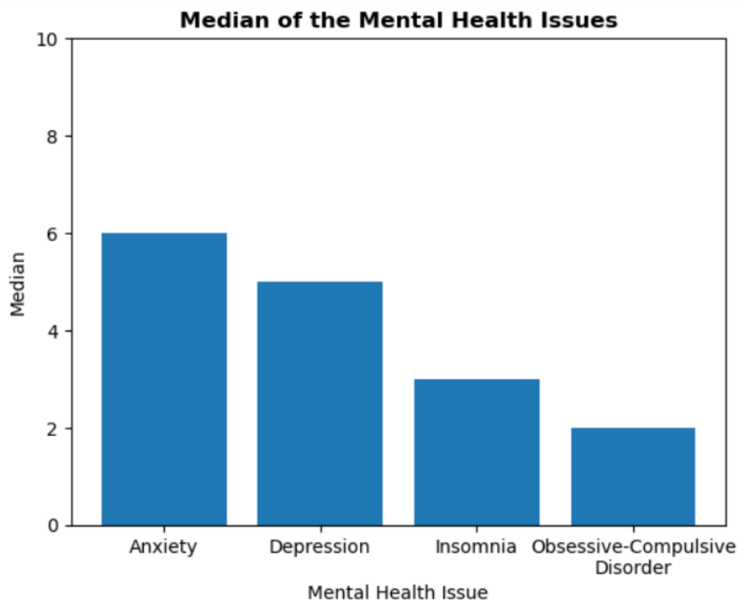
The individuals of the sample listen to music between zero and ten hours each day. The variable has a great positive asymmetry, which means that the data is concentrated in low values, being the median three. Most people of the sample listen among one and five hours of music per day (*reference the hours per day bar chart*). Besides, a 73,21% of the individuals are willing to search and explore new types and genres of music (*reference the exploratory pie chart*).

A 76,11% of people in the sample think that music has an improvement effect on their mood, 2,22% that it has a negative effect, and 21,67% that music does not have an effect in them. That is perfectly appreciable in the next pie chart:

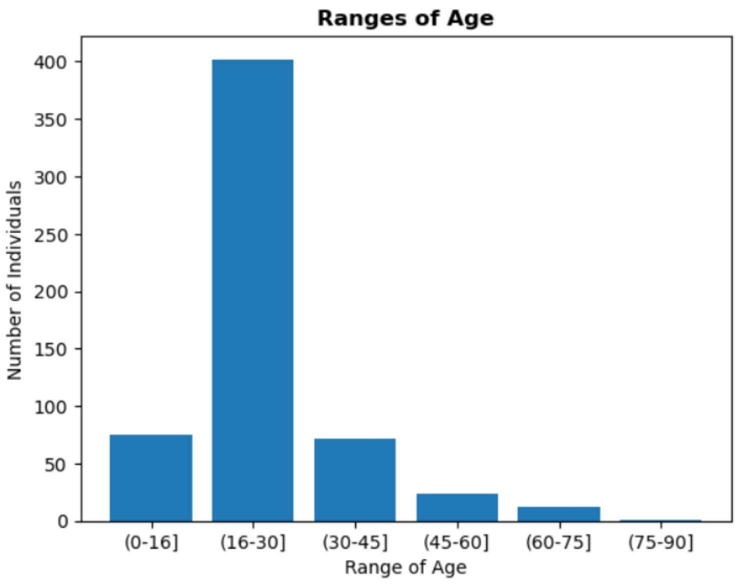


Furthermore, the variables that talk about mental health are the anxiety, depression, insomnia and obsessive-compulsive disorder that individuals think that suffer, being all of them a self-diagnosis.

Every value of these four variables ranges from zero to ten. Anxiety variable has a negative asymmetry, that means its data tends to have a high value (*reference anxiety bar chart*), also its median is six. The values of depression are more equilibrated in the sample (*reference depression bar chart*), with a median of five. Besides, insomnia variable has a median of three and obsessive-compulsive disorder has a median of two. Both have a positive asymmetry, due to the fact that most of their data has low values. In order to compare visually, this bar chart shows the medians of the four mental health issues:



Moreover, the individuals of the sample have been divided in ranges of age* (*footnote: "(" this specific age is not included in the interval, "]" this specific age is included in the interval.*): (0-16], (16-30], (30-45], (45-60], (60-75] and (75-90]. The age variable has a median of 21 and its quartiles are 18 and 27, being its interquartile range 9. The 68,6% of the individuals are included in the interval that comprise the ages between 17 and 30, the (16-30] one. That is seen clearly in this bar chart:



For more information and the complete unidimensional analysis of the variables check the Annex B.

Multidimensional Analysis

A big part of the project lays on the Multidimensional Analysis. It gives answer to questions like, does listening to music more hours influence in insomnia levels?, how much does believing that music is positive has to do in your current mental state?

Standardization

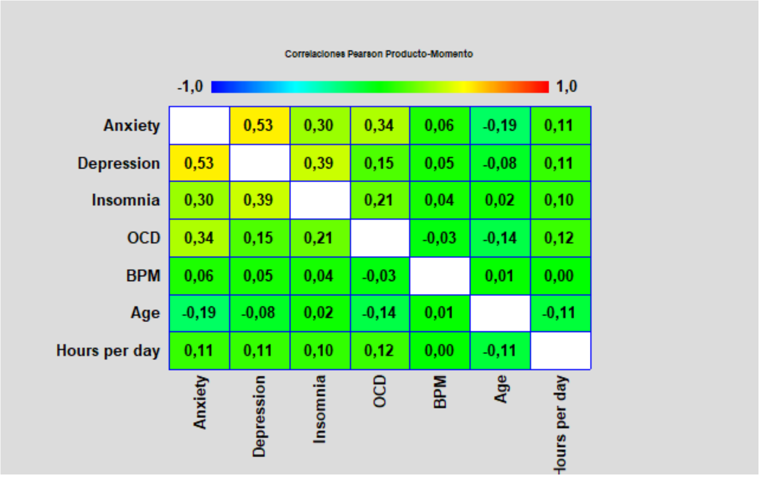
The Standardization is the procedure by which the comparision among values in different units is possible. In this descriptive analysis, the quantitive variables have no unit as they are collected in a scale from 0 to 10 (**Anxiety**, **Depression**, **Insomnia**, **OCD**). In the other hand, the variables **BPM**, **Hours per day** and **Age** have every value in beat per minute, hours and years respectively. Everything said, a standardization is not necessary.

Identification, study and description of relationships

The goal of the multidimensional analysis is to relate two or more variables in order to look for an explanation of a fenomenon. To put it in other words, is the same as investigating the cause and showing how intense the consequences are.

Multidimensional Analysis of Quantitative Variables

Age (without recodification), **BPM, Hours per day, Anxiety, Depression, Insomnia** and **OCD** are quantitative variables**.** These variables will be related in order to find some relations. The relations are obtained from the covariance matrix, in this graphic each variable has the number of correlation with the other variables. The boxes with a value close to 0 has no correlation between them (boxes in green). It can be seen that there is a relation between **Depression** and **Anxiety** (box yelllow)**,** This relation show that people who has more **Anxiety** has higher levels of **Depression**.

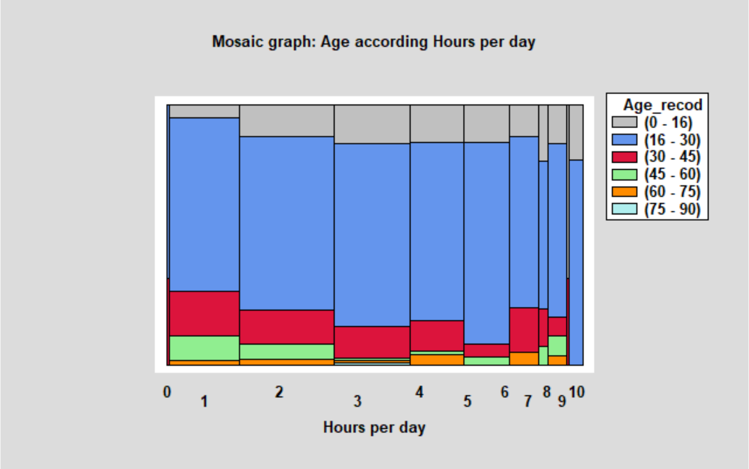


Correlaciones Pearson Producto-Momento

	Anxiety	Depression	Insomnia	OCD	BPM	Age	Hours per day
Anxiety		0,53	0,30	0,34	0,06	-0,19	0,11
Depression	0,53		0,39	0,15	0,05	-0,08	0,11
Insomnia	0,30	0,39		0,21	0,04	0,02	0,10
OCD	0,34	0,15	0,21		-0,03	-0,14	0,12
BPM	0,06	0,05	0,04	-0,03		0,01	0,00
Age	-0,19	-0,08	0,02	-0,14	0,01		-0,11
Hours per day	0,11	0,11	0,10	0,12	0,00	-0,11	

Multidimensional Analysis of Qualitative Variables

The distribution of **Age_recod** was seen in the unidimensional analysis, now this variable is compared with other variables in order to find correlations. The finality of finding correlations is to know if the different ranges of age should be treated separately. In order to see if these correlations exist it is used the “Mosaic grapf” relating **Age_recod** with the rest of variables. It can be seen that the variable **Age_recod** has no correlation with any variable because the colored areas has the same size for each correlation how in the following example:

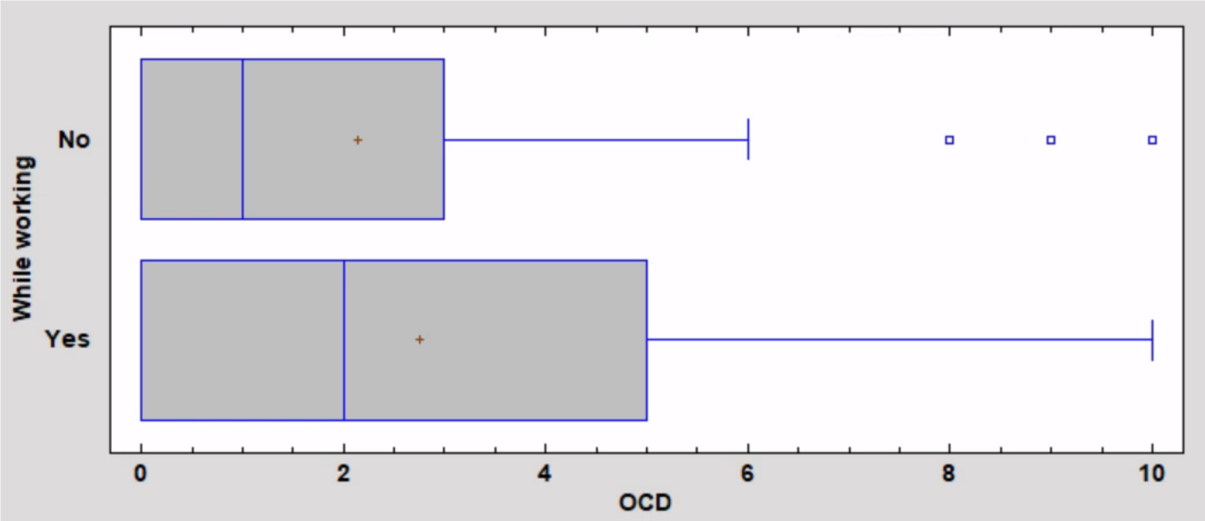


How the variable **Age_recod** has not a relation with other variables we can use all the sample instead of different age ranges in the rest of the analysis.

After that it will be seen if the hours of listening to music affects to mental status. The variables **Hours per day, Anxiety, Depression** and **Insomnia** are quantitative variables but how the variables has few values they will be treated as ordinal qualitative ones. How the variables has no corelations it is shown that listen more hours of music per day does not contribute to lower or higher levels of **Anxiety, Depression** or **Insomnia** than listening to less hours.

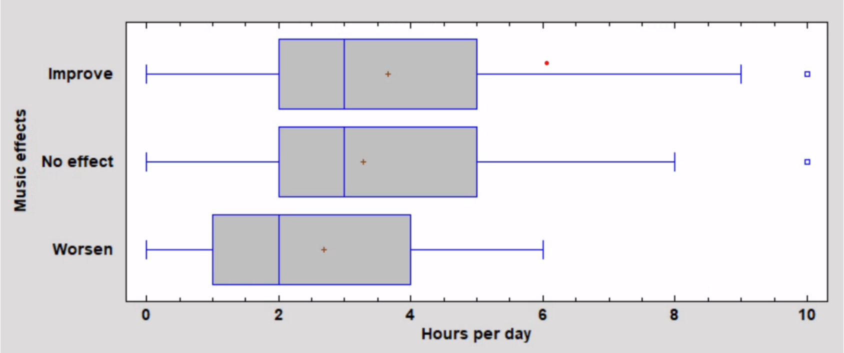
Now the variables **OCD** (Obsesive compulsive disorder) and **While working** are analysed in order to find correlations. **OCD** was previously unused. It also goes from 0 to 10. It will be treated as a quantitative variable by using the **Subgroup Analysis.** This tool makes an univariant analysis for every value of the qualitative variable **(While working)** along with every value of the quantitative one (**OCD**, **Anxiety**, **Depression**, **Insomnia**…).

The results are shown on the **Box&Whiskers** graph. How it can be seen the the distribution on the **Yes** is more to the right than on the **No**. So, it can be said that people with highs levels of **OCD** use to use more music while working.



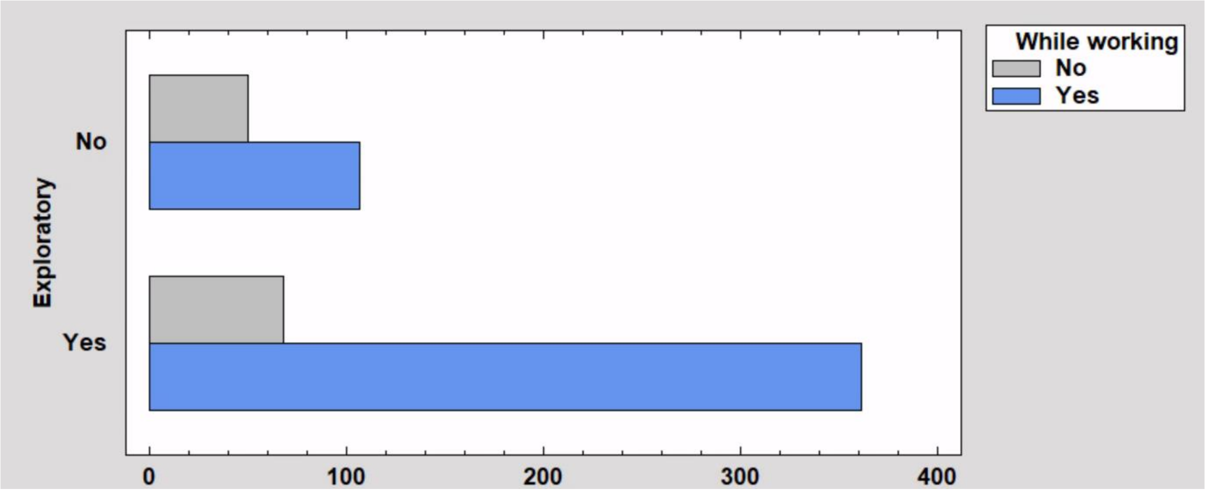
After that, it will be proved if there is a relation between the **Mussic effects** and the **Hours per day** the music is listened. **Hours per day** has been used with **Anxiety, Depression** and **Insomnia** but treating them as ordinal qualitative variables. Now **Hours per day** will be treated as a discrete quantitative variable.

How it can be seen in the **Box&Whiskers** the distribution on people who think that music worsens their lives has lower values of **Hours per day** than people who think the opposite or nothing.



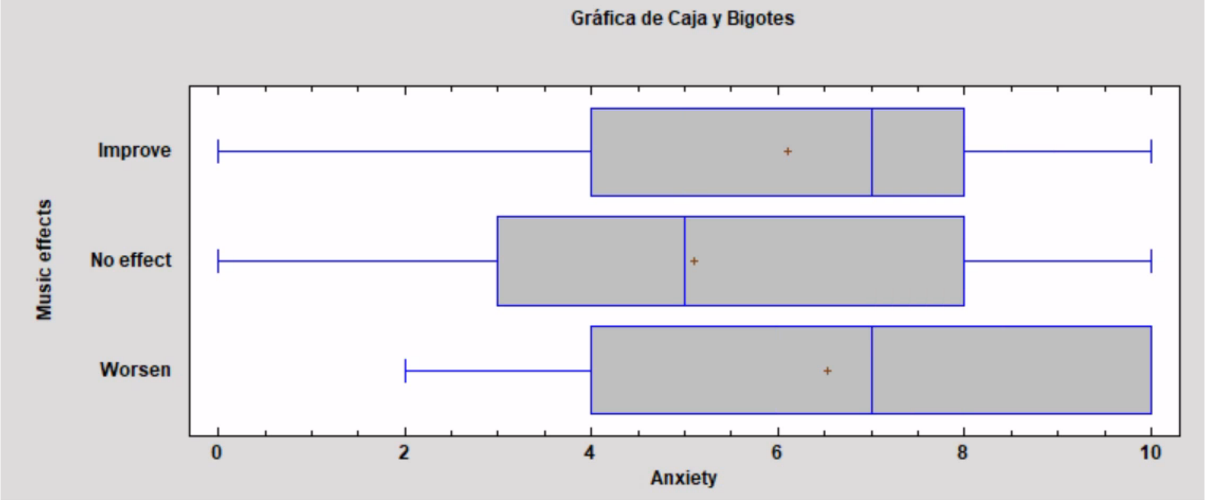
Now, do people who listen to music **while working** tend to be more **exploratory**? Or are they more likely to be less **exploratory while working**?

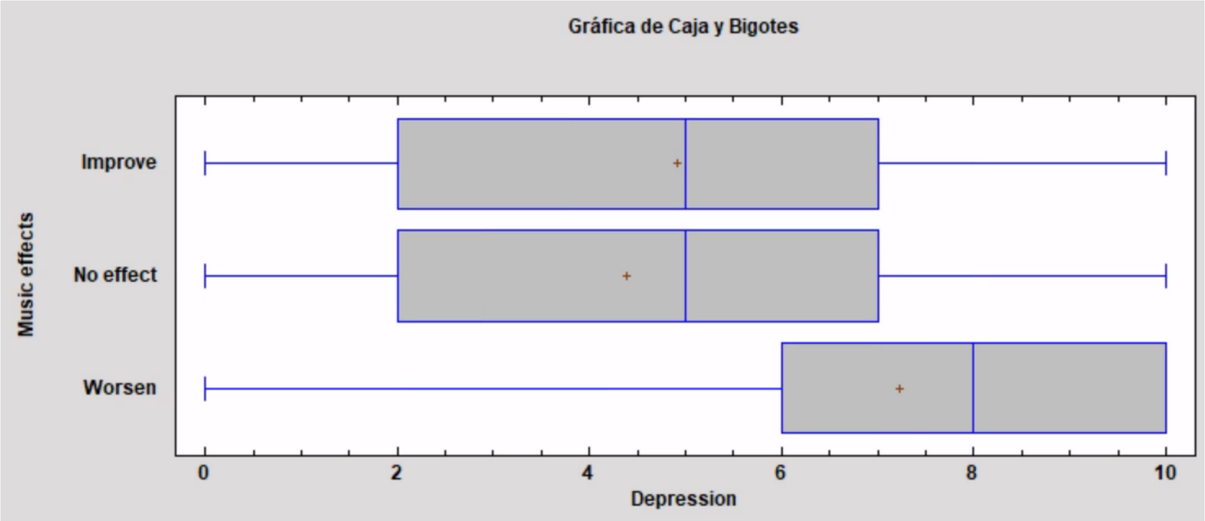
Although being more people who listen to music while working in both scenarios of the **exploratory** variable, there is a significant difference in the frequencies. There are many more who are exploratory and listen to music while working 85% than people who are not exploratory and listen to music while working 68% .

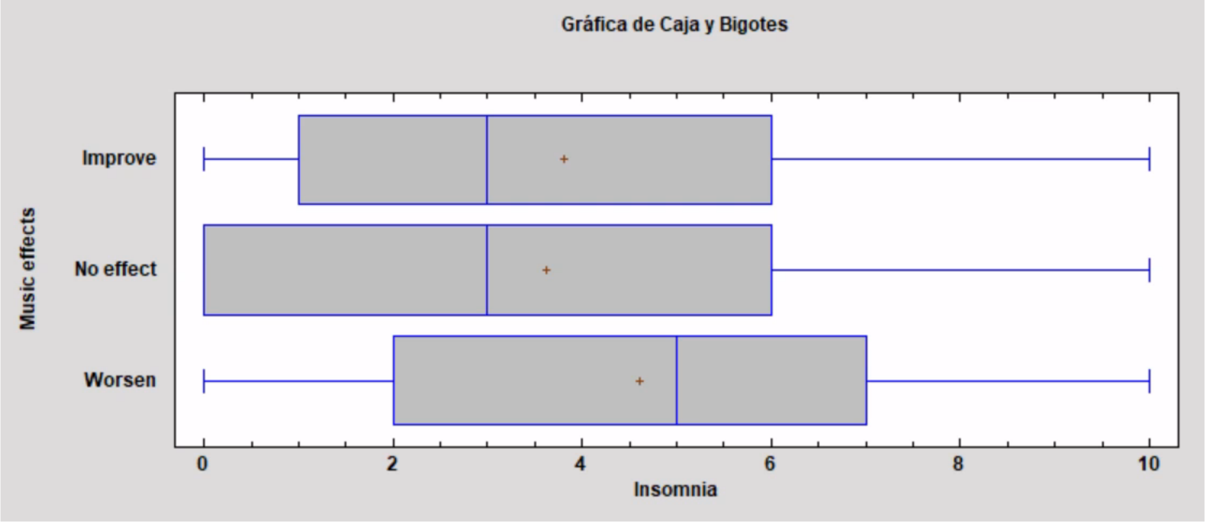


After all, it will be checked how the mental status of the individuals es related with if the music affects them. For this it will be used **Music effects**, a qualitative varible with 3 possible values (Improve, No effect, Worsen) and the thre principal mental health problems in this project **(Anxiety, Depression and Insomnia)**. Will the fact of considering music to be positive be reflected in the values of **Anxiety, Depression** and **Insomnia**?

There are clearly higher levels of **depression** and **anxiety** in people who believe that music worsens their lives than people who believe the opposite. Distribution of *Improve* and *No effect* is quite the same in every scenario.







For more details abour the multidimensional analysis check the annex C.

Tools utilized

In order to carry out the analysis different tools have been used. The main tools were Statgraphics and excel. The team used Statgraphics instead of other tools like Python (with jupyter notebook) because the team is acquainted with this program and the team can use this properly and saved time of learning a new tool. Excel was used to prepare the data and to make the data file usable to Statgraphics.

In the preprocess stage were used the following tools: The tool used for the identification of outliers was the outliers identificator from analysis of one variable and the tool used for treat the data was the data book. The missing values were found by the data viewer and they were treated by the data book. The new variables were recoded, by the recodification tool that statgraphics provides, after copying the original variables.

First of all, in the unidimensional analysis of the variables, their type and subtype have been defined. In addition to that, the frequency table has been used to determine the absolute frequency and the relative frequency of every value of the variables. Furthermore, in the qualitative variables the bar chart and pie chart have been used to give a more visual representation of the result. In regard to the quantitative variables, the Box and Whisker chart has been used to find atypical data. In order to determine the form of the variable, the standardized asymmetry coefficient of Fisher meets the objective. About the position of the values, the parameters used were the median and the quartiles. Besides, the interquartile range have been used to see the dispersion of the values. Moreover, for the purpose of giving a visual representation, the bar chart and the scatter plot, being the quantitative variables discrete, and also the histogram and the normal probability graph, being continuous, have been used.

In the multivariant analysis was used StatGraphics and its functionalities: for categorical variables the functionality of *Two Factors*, for numerical variables the functionality of *Multivariable* and when both kinds were worked at the same time, the *Subgroup Analysis*.

Results

The results are in form of the relationship between the variables previously obtained. In summary, they relate the grade of association between variables which talk about habits consumption, listener's characteristics and mental paramets. There are some correlations stronger than others. To see discarded relations due to their weakness but which are interesting to get rid of incorrect assumptions, visit Annex C. These are the relations said in natural language:

People who have **higer levels of OCD** are more likely to listen to **music while working** than people who have lower levels. Music, especially the one with repetitive and predictable rhythms, can provide a structure that resonates with the organizational and routine needs of a person with OCD. This can help create a more comfortable and manageable work environment to people with OCD.

People who **think that music worsens** their **lives**, **listen to music less hours** than people who think the opposite or nothing. This is obvious but can be due to multiple factor, like for example having asociated bad experiences with music.

People who are not e**xploratory** with music listen less to **foreign language** music than those who are e**xploratory**. Again, this is another obvious statement. If you like to explore new music genres, it is more likely to listen to foreign language music.

If you listen to **music while working**, it is more likely for you to be an **exploratory** person than if you do not listen to music while working.

Higher levels of **anxiety** and **depression** abound more in people who think that music worsens their lives. This can be possible since people with more anxiety or depression tend to feel a negative perception of many activities.

Conclusions

Extraída la información de los datos, se ponen sobre la mesa las siguientes conclusiones sobre la base de datos: Se ha visto como las variables de Ansiedad, Depresión, Insomnia y OCD se distribuyen de la misma manera para todos los rangos de edad que se han establecido, aun cuando el rango más popular es el de (16-30). Con esto se puede afirmar que todas las edades tienden a tener valores similares autopercibidos de Ansiedad, Depresión, Insomnia y OCD. #Aquí hay que hacer StateOfArt para interpretarlo con estudios científicos# #Se deberán mostrar los 4 gráficos de mosaicos y explicar cómo entenderlas en la web#

#Informarse en State of Art cómo la salud mental afecta la percepción de actividades# La salud mental puede afectar la percepción de actividades cotidianas, como escuchar música. La variable Music effects habla sobre las percepciones. En la muestra, hay un 76,11% de personas que creen que la música es positiva para su salud mental. El 21,67% no cree nada y el 2,22% restante cree que empeora. #Mostrar diagrama de secotres de Marcos#. Si se comparan las percepciones con los niveles de Ansiedad y Depresion se puede observar como claramente la gente que piensa que la música es mala para ella tiene más ansiedad y depresión #Inserto los 2 Box & Whiskers que los relacionan#. El nivel de Ansiedad y Depresión promedio, redondeados, es de 6. Como estas variables siguen una distribución normal, se considerará que un valor >= 6 será un nivel alto de Ansiedad y uno >=5 será un nivel alto de Depresión. ¿Cuáles son las probabilidades, de tener esos niveles altos? Mediante la llamada tipificación, se consiguen. P(Anxiety >= 6) = 0.644 & P(Depression >= 5) = 0.5195. Si se filtra la muestra para la gente que cree que la música mejora su salud y se hace un diagrama de sectores según niveles de depresión y ansiedad, y luego se repite lo mismo filtrando para la gente que cree que empeora su salud mental, se obtiene: . #Mostrar tabla de 4 diagramas de sectores#. Es notable cómo los diagramas de sectores de las personas Worsen tienen más trozo de nivel mayor que 7 que las personas Improve.

Como suena lógico, la gente que crea que la música es mala para su salud escuchará menos horas de música, como se muestra en esta gráfica #Insertar Box&Whiskers de Music effects y hours per day#. La media de horas escuchadas para la gente que considera que la música les empeora es de 2,692 h contra las 3,661h de los que piensan lo contrario. Se aprovecha para decir que las horas de música escuchadas y la edad son indiferentes entre sí, pues hay un 0,11 de relación sobre 1. Esto significa que la edad no afecta a cuántas horas de música escuches. Continuando con variables cuantitativas, las horas de consumo y la edad tampoco tienen una correlación con la ansiedad, depresión, insomnio o OCD, tal y como se puede observar en la tabla de correlaciones. #Mostras tabla de correlaciones de Pearson#

Prestando atención a la tabla de correlaciones de Pearson se pueden apreciar como hay una leve relación entre la gente con ansiedad y trastorno obsesivo compulsivo (OCD), del 0,34 sobre 1. El 68,94% de la muestra tiene valores menores o iguales para OCD significando que no es común que haya valores altos de OCD. #Muestro diagramas de sectores de niveles de OCD filtrados con gente While working Yes & No. La frecuencia relativa o la probabilidad de escuchar música mientras trabajas es de 0,7986. La probabilidad de NO escucharla es de 0,2014. Si consideramos los valores de OCD > 3 como aquellos que son menos comunes, la probabilidad acumulada sería del 31,06%. Sabiendo esto es posible sacar la Probabilidad(While Working=Yes & OCD>3) = 0,7986 * 0,3106 = 0,2480 = 24,8%. Mientras tanto la Probabilidad(While Working = No & OCD > 3) = 0,2014 * 0,3106 = 0,0625 = 6,25% En otras palabras, es cuatro veces más probable que tengas valores anormlaes de OCD si escuchas música mientras trabajas que si no. #Hacer búsqueda científica sobre este tema#. Las personas con niveles más altos de OCD son más propensas a escuchar música mientras trabajan. La música, especialmente aquella con ritmos repetitivos y predecibles, puede proporcionar una estructura que resuena con las necesidades de organización y rutina de alguien con OCD.

Destacar también la relación que hay entre la gente que escucha música mientras trabaja y a la que le gusta explorar (exploratorios) nuevos géneros. #Inserto tabla de contingencia reducida a cuatro campos de probabilidad#. Hablando de géneros, se ha visto cómo cualquier clase de género musical que tengas por favorito no afecta a los niveles de insomnia, depresión, ansiedad u OCD. Si se observan los siguientes box&whiskers, es apreciable como las cajas están dsitribuidos muy a la par. #Meto imagen de análisis subgrupos de géneros favoritos e insomnia, asniedad…etc#

Bibliography

References

1. Caron C. (2023) *How Music Can Be Mental Health Care*. The New York Times.
2. Tulane University (2021) *Understanding Mental Health as a Public Health Issue*. Tulane.edu
3. MacMillan A. (2023). *4 Possible Reasons Why Mental Health Is Getting Worse*. health.com
4. Moose A. and Bhargava R. *A generation adrift: Why young people are less happy and what we can do about it*. weforum.org
5. MacMillan A. (2023). *4 Possible Reasons Why Mental Health Is Getting Worse*. health.com
6. Pramanik P. (2023) *Music could hold the key to developing effective mental health interventions*. news-medical.net
7. Pearce Stevens A. (2024). *Music has the power to move us physically and emotionally. Here's why*. ScienceNewsExplores
8. Reynolds F. (2023). *The transformative power of music in mental well-being*. American Psychiatric Association
9. Gilmour L. (2024). *New project supports children and young people to access the mental health benefits of live music*. University of Stirling
10. Pearce Stevens A. (2024). *Music has the power to move us physically and emotionally. Here's why*. ScienceNewsExplores
11. Goodwin, Dierker, Wu, Galea, Hoven and Weinberger (2022) *Trends in U.S. Depression Prevalence From 2015 to 2020: The Widening Treatment Gap*. American Journal of Preventive Medicine.

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

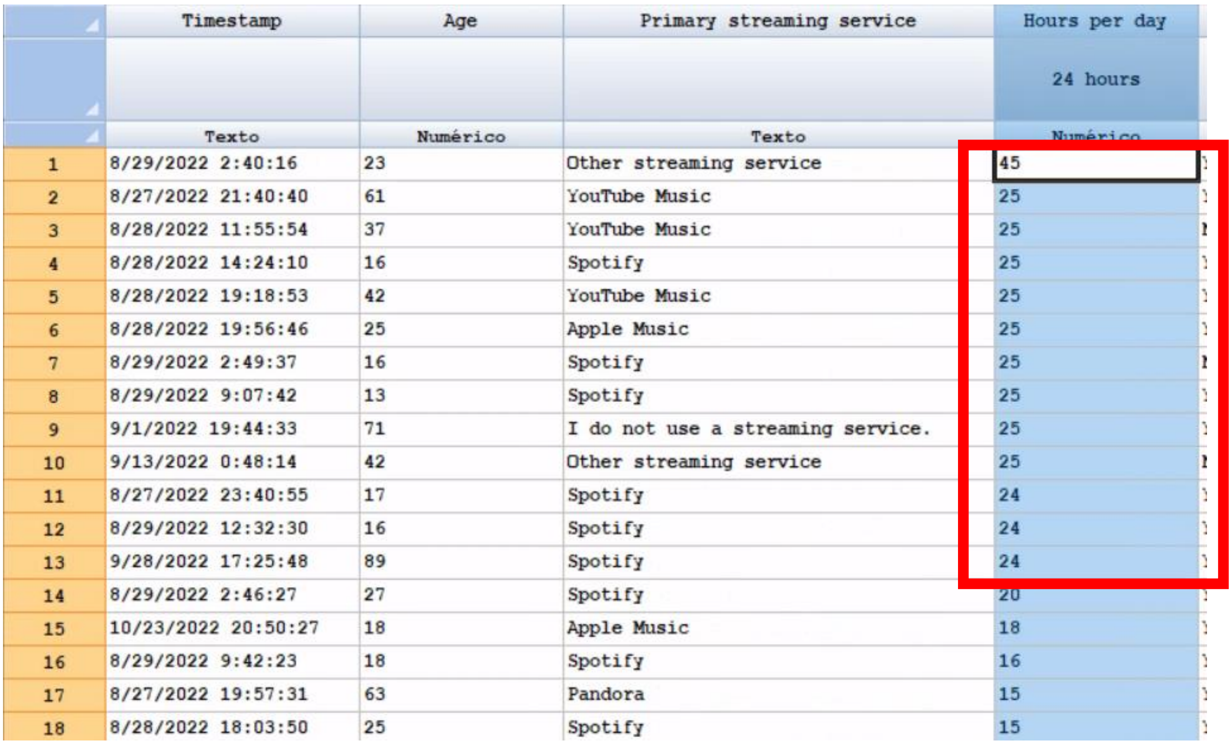
In this stage the outliers have been analysed and treated with different tools. Fisrt of all, the values were ordered and the values that where out of range

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.

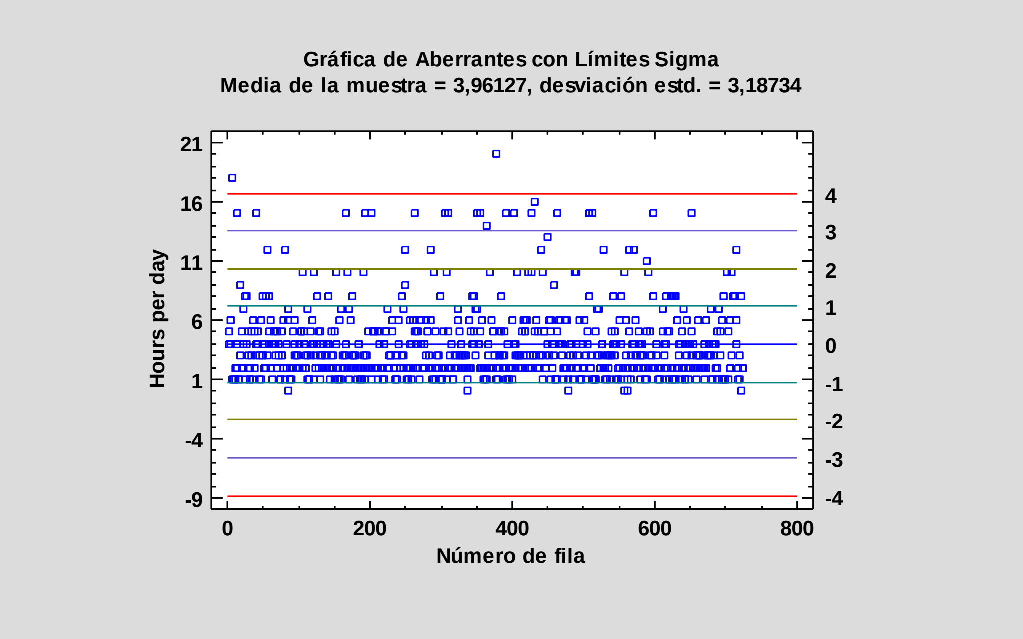
Hours per day:

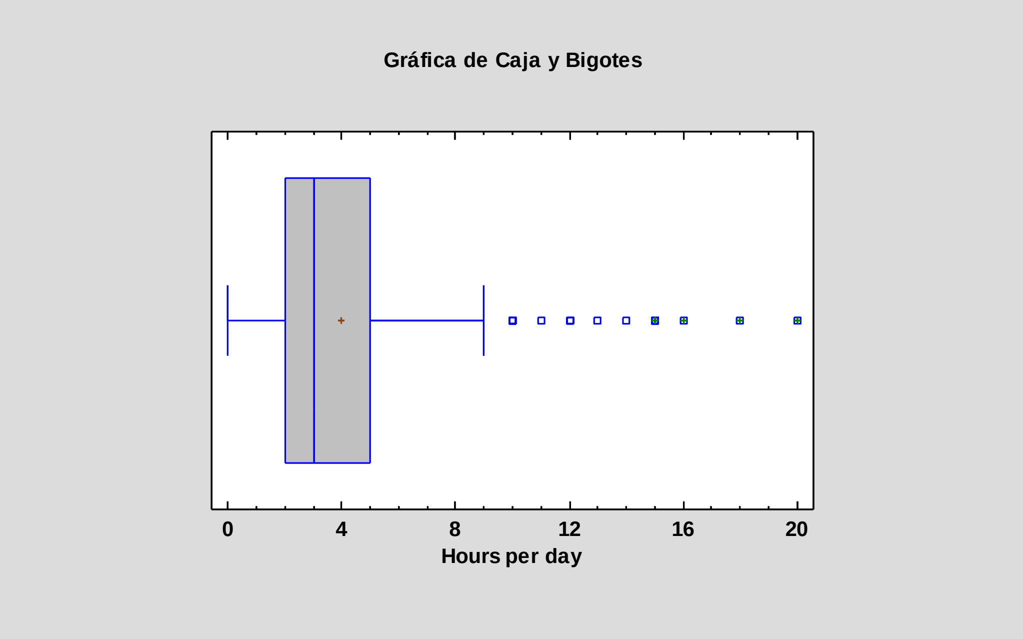
First of all, it will be seen what values don't make sense because they are out of range. The individuals with this values will be deleted. The individuals will be ordered as a descending form and the values bigger or equal than 24 will be deleted.



	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

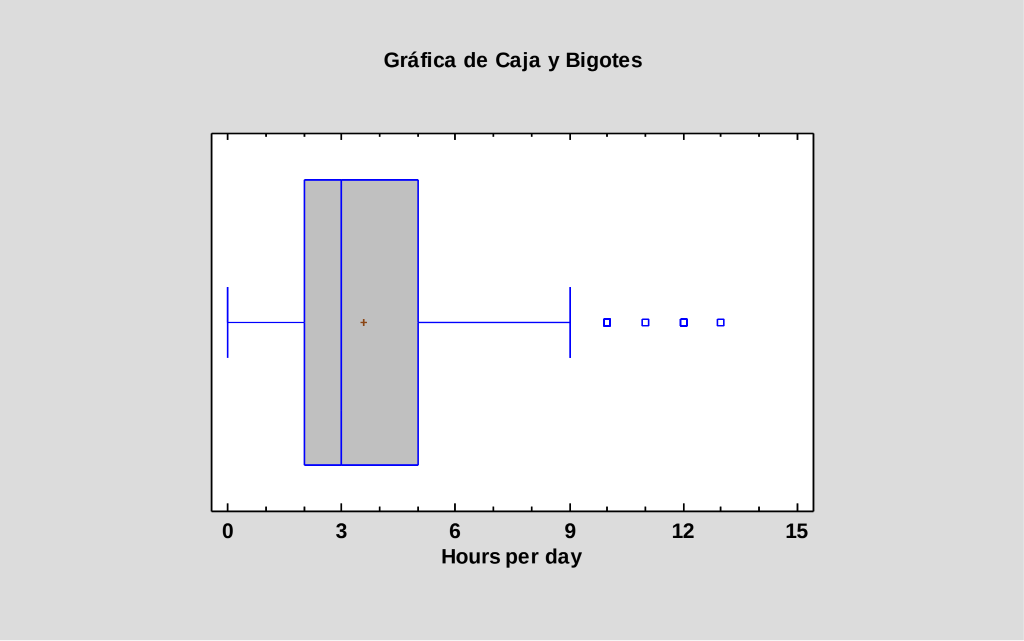
After that the outliers identification will start.



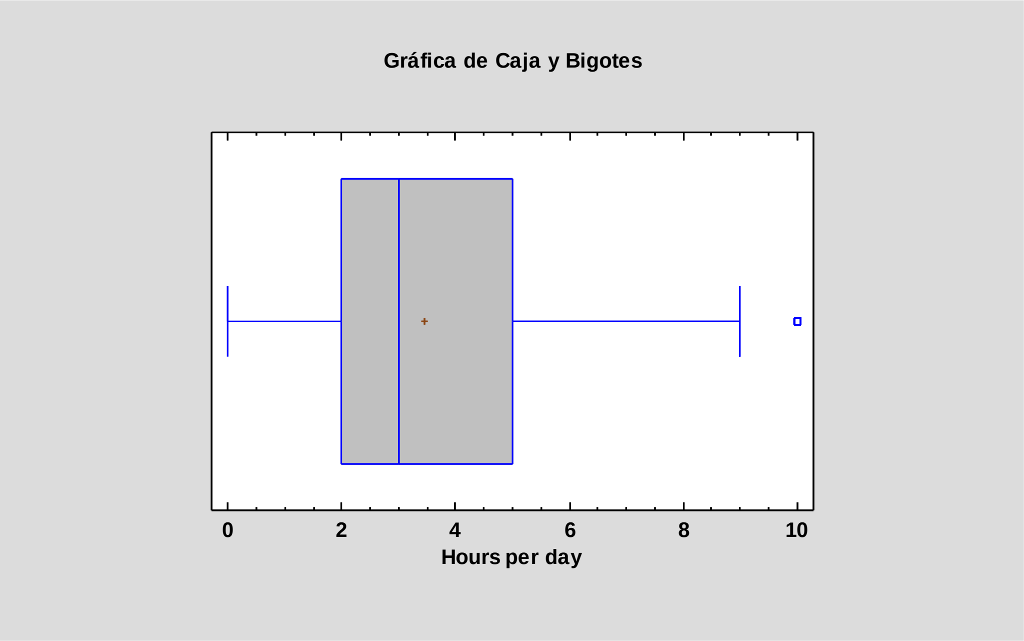


After seeing this graphs it can be said that there are a positive asimetry and there are extreme values that are not representative. So the values biggers or equal to 14 will be changed for the median because they are not representative. After this process it will be checked again.

BPM:



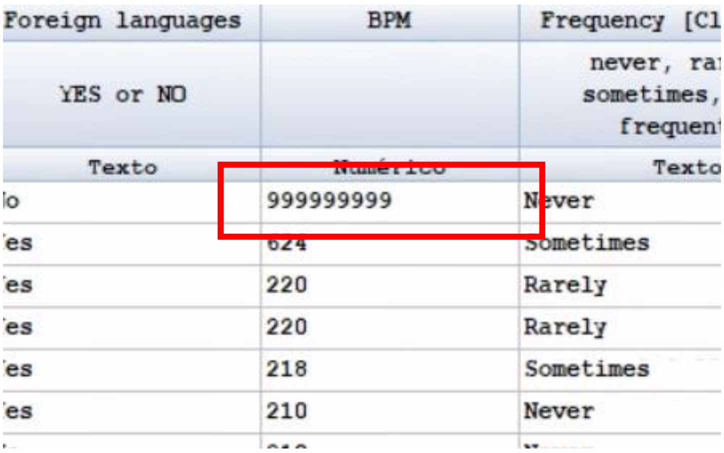
How it can be seen there are extrem values yet so it will be changed the individual with hours biggers than 10 and it will be checked again.



After this process the variable hours per day is treated.

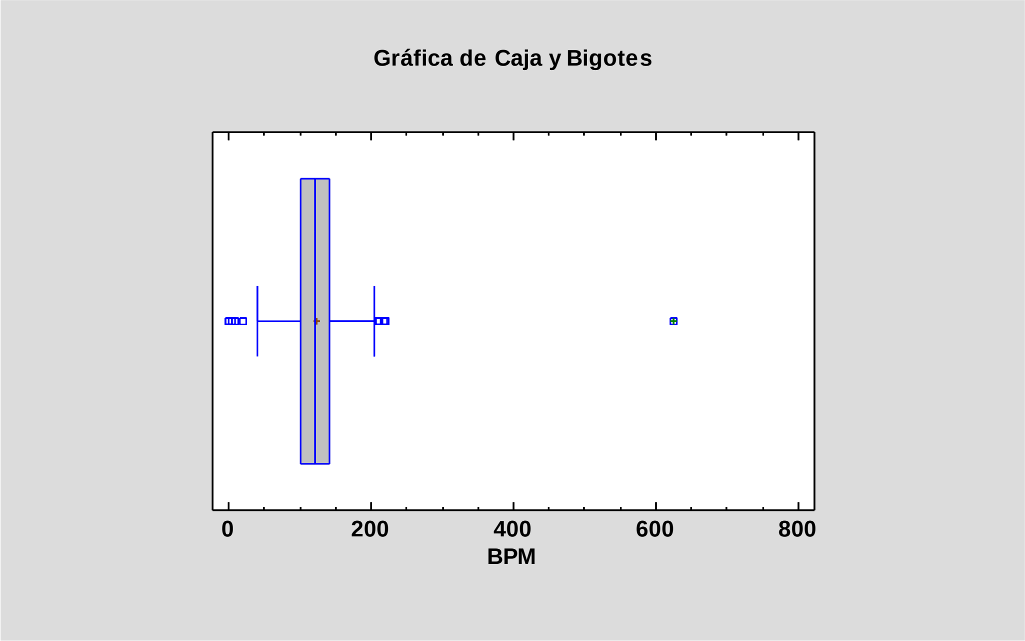
BPM:

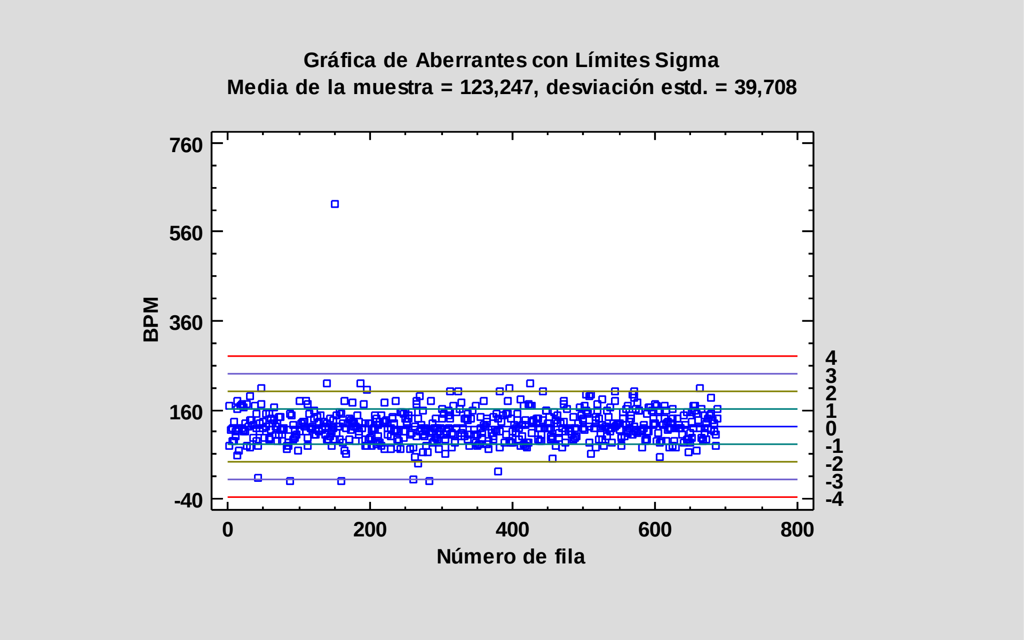
There are one value that don’t make sense because it is imposible and this value affects all the sample parameters. So it will be deleted.



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

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

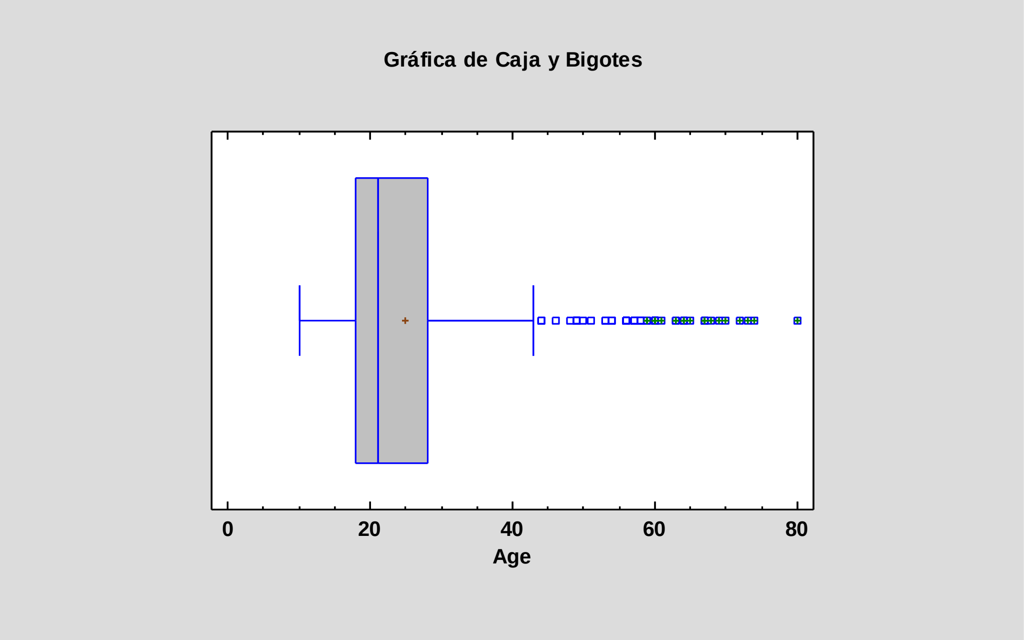




After seeing this graphs it can be seen that there is an extreme value that is not representative from above and there are also some values that are not representative under. So to avoid eliminating all the individuals the values will be changed for the median. After that, the variable is already treated.

Age:

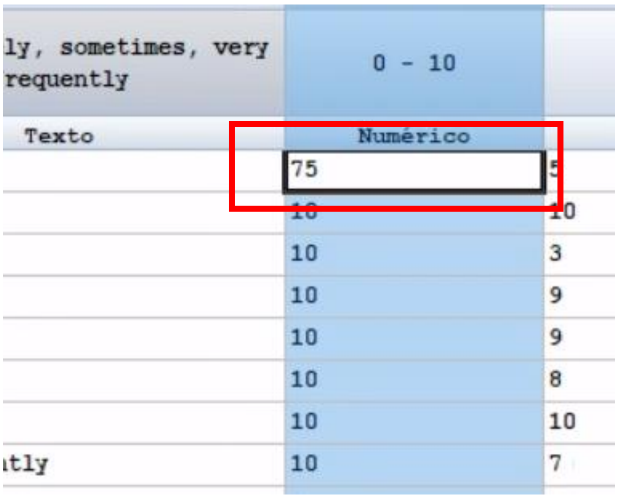
There are not values that don’t make sense so it will be done the identification of outliers and extreme values.



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.

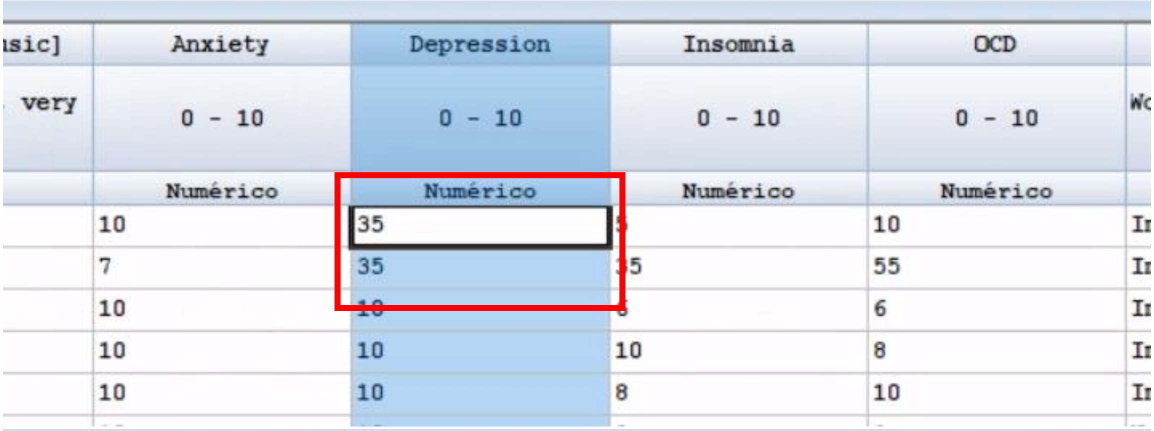


ly, sometimes, very requently	0 - 10	
Texto	Numérico	
	75	5
	10	10
	10	3
	10	9
	10	9
	10	8
	10	10
ıtly	10	7

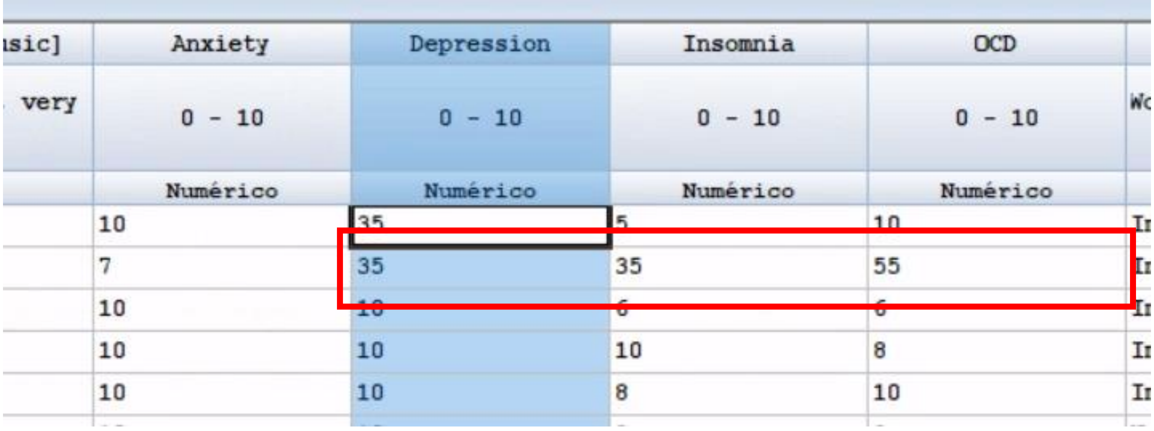
With this process the variable is ready to be analyzed by the team.

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	Wc
	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



ısic]	Anxiety	Depression	Insomnia	OCD	
very	0 - 10	0 - 10	0 - 10	0 - 10	Wc
	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

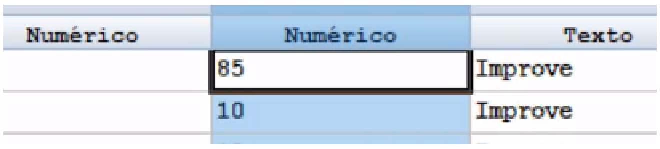
With this process the variable is ready to be analyzed by the team.

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

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] *(never, rarely, sometimes, very frequently).*	**Frequency [Video games music]** *(never, rarely, sometimes, very frequently).*	**Anxiety** *(0 – 10)*	**Depression** *(0 – 10)*
0	0	0	0
Insomnia *(0 – 10)*	**Ocd** *(0 – 10)*	**Music effects** *(worsen, no effect, improve)*	**Permision and timestamp**
0	0	6	0

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.

Annex B (Unidimensional analysis)

Tools Used for the Qualitative Variables Analysis

In addition to defining the type of qualitative variable, for its analysis the frequency table has been used to determine the number of individuals that have each value of the variable, its absolute frequency; and how much that quantity represents in the total of the sample, its relative frequency.

Moreover, for a more visual representation of the result, the bar chart has been used to show the absolute frequency and the pie chart to show the relative frequency.

Tools Used for the Quantitative Variables Analysis

In addition to defining the type of quantitative variable, for its analysis the frequency table has been used to determine the number of individuals that have each value of the variable, its absolute frequency; and how much that quantity represents in the total of the sample, its relative frequency. In order to do this table, the quantitative variables have been discretized, if it had been necessary.

First of all, the Box and Whisker chart has been used to find out if there is atypical data in the sample. Furthermore, with the objective of defining the form of the values of the variable, the standardized asymmetry coefficient of Fisher has been used to determine the asymmetry of the variable. If the quantitative variable was continuous, in case of being symmetric and having a normal distribution too, the standardized kurtosis coefficient has been used to determine its aiming or fussiness.

Regarding the position of the values of the variable, the parameters used have been the median, the quartiles and, if the sample was symmetric, the winsorized mean too. In regard to the dispersion of the values, the interquartile range has been used and, being the sample symmetric, also the winsorized standard deviation. The parameters mentioned before are robust, extreme values in the sample do not affect them.

Besides, the bar chart and the scatter plot, being the quantitative variables discrete, and also the histogram and the normal probability graph, being continuous, have been used to represent visually the form, position and dispersion of the values of the sample.

Anxiety

This variable indicates the self-perception of the level of anxiety. It is a discrete quantitative variable, whose values range from 0 to 10.

The frequency table of the variable is the next:

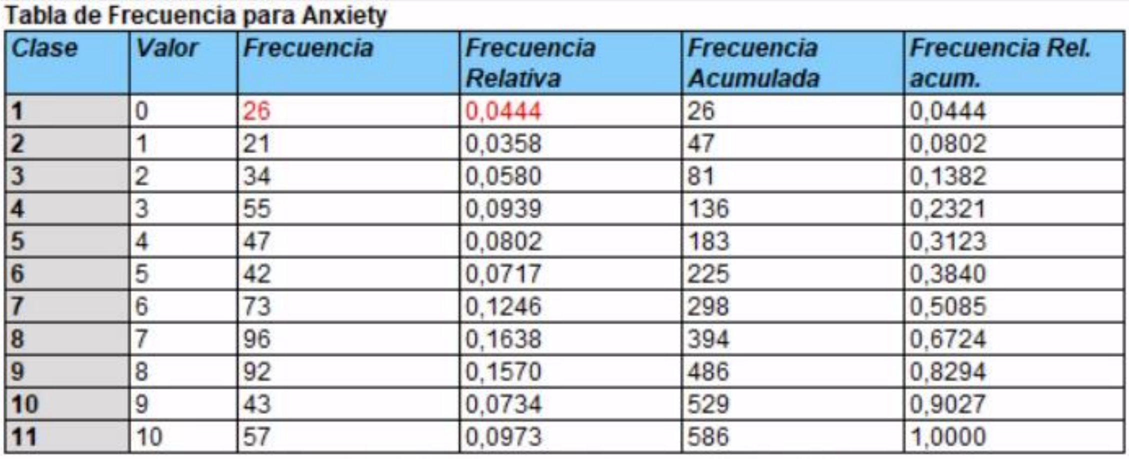
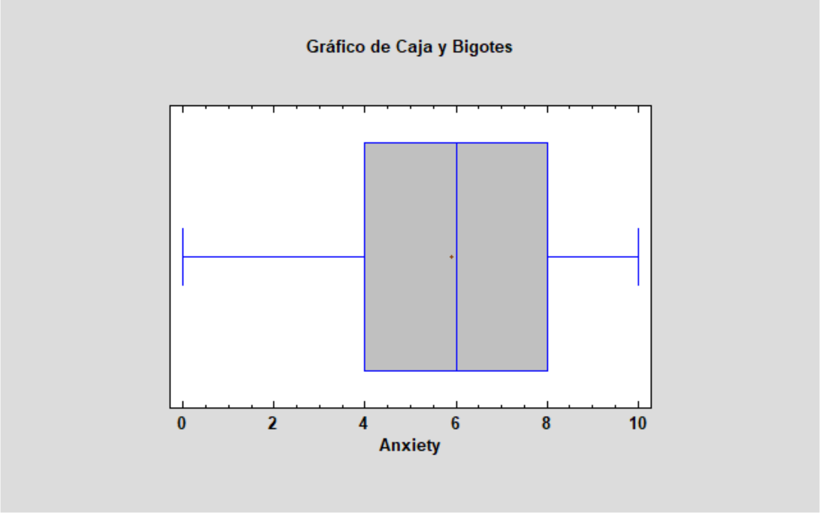
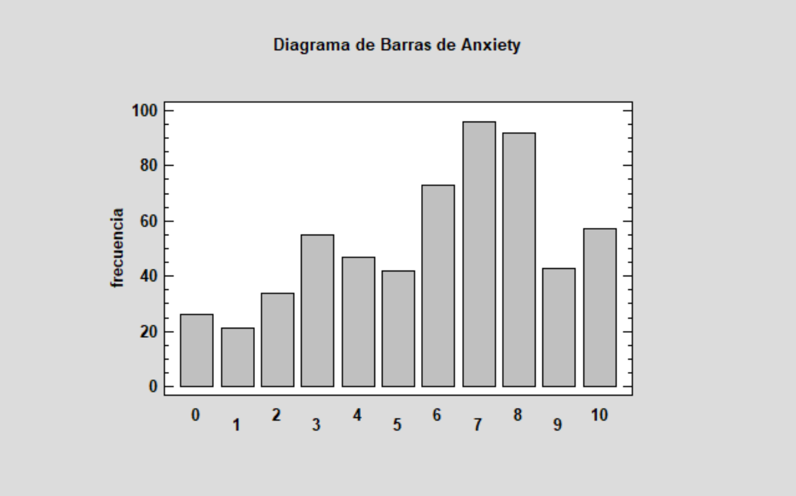


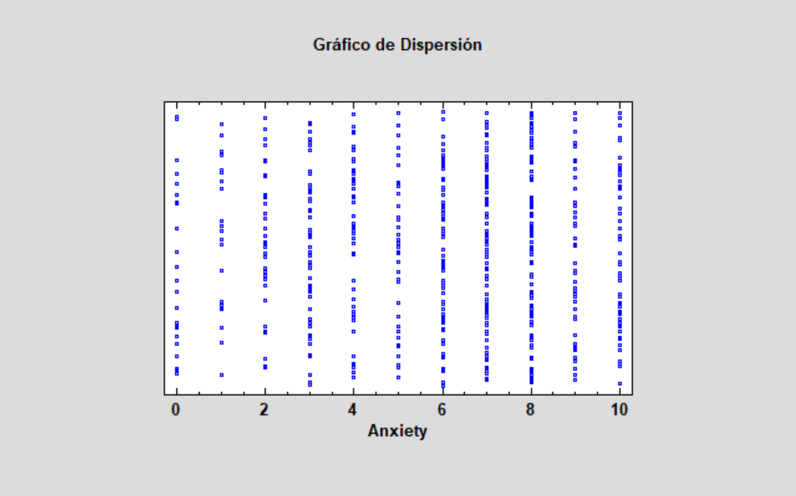
Tabla de Frecuencia para Anxiety

Clase	Valor	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Rel. acum.
1	0	26	0,0444	26	0,0444
2	1	21	0,0358	47	0,0802
3	2	34	0,0580	81	0,1382
4	3	55	0,0939	136	0,2321
5	4	47	0,0802	183	0,3123
6	5	42	0,0717	225	0,3840
7	6	73	0,1246	298	0,5085
8	7	96	0,1638	394	0,6724
9	8	92	0,1570	486	0,8294
10	9	43	0,0734	529	0,9027
11	10	57	0,0973	586	1,0000



As it can be seen in the Box and Whisker chart there are not atypical data in the sample. Furthermore, to describe its form, the standardized asymmetry coefficient of Fisher of –4,21 indicates a negative asymmetry in the variable. It can be seen better in the bar chart and the scatter plot.





Regarding the position of the values, the median is 6, the lower quartile is 4 and the upper quartile is 8. Besides, about the dispersion of the sample, the interquartile range is 4.

Depression

This variable indicates the self-perception of the level of depression. It is a discrete quantitative variable, whose values range from 0 to 10.

The frequency table of the variable is the next:

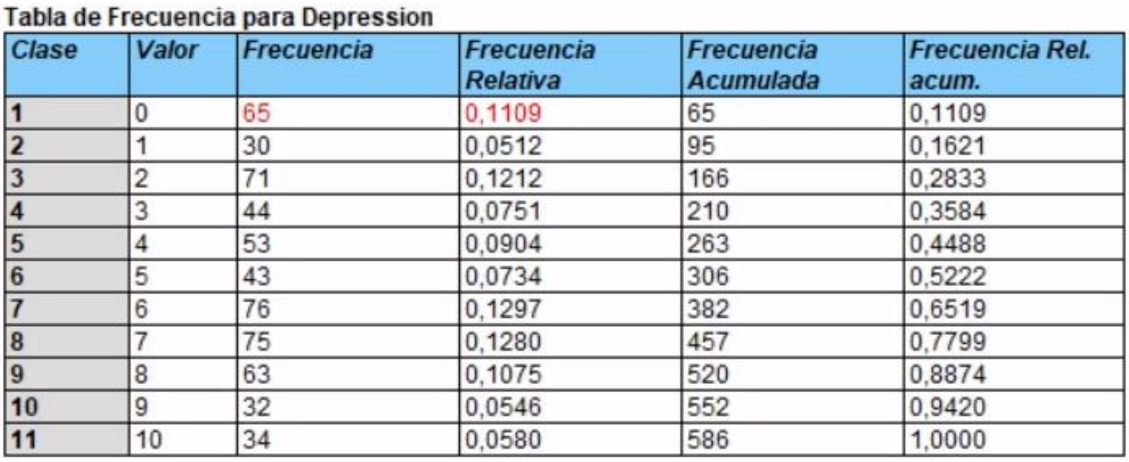
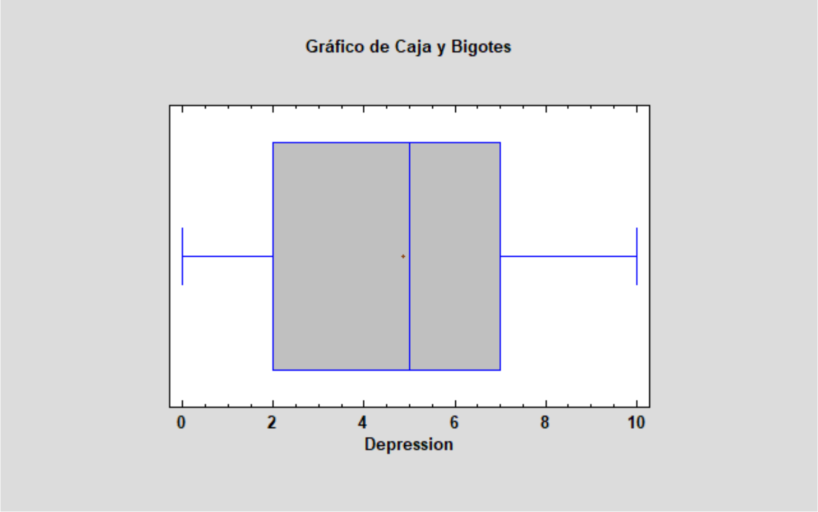
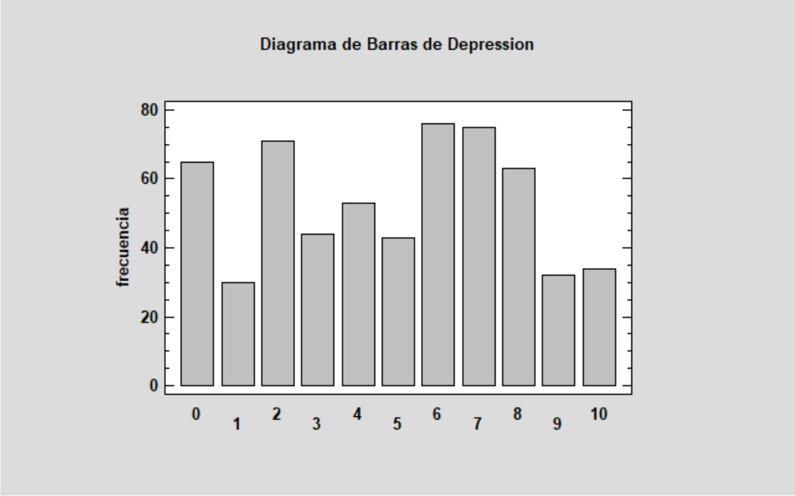


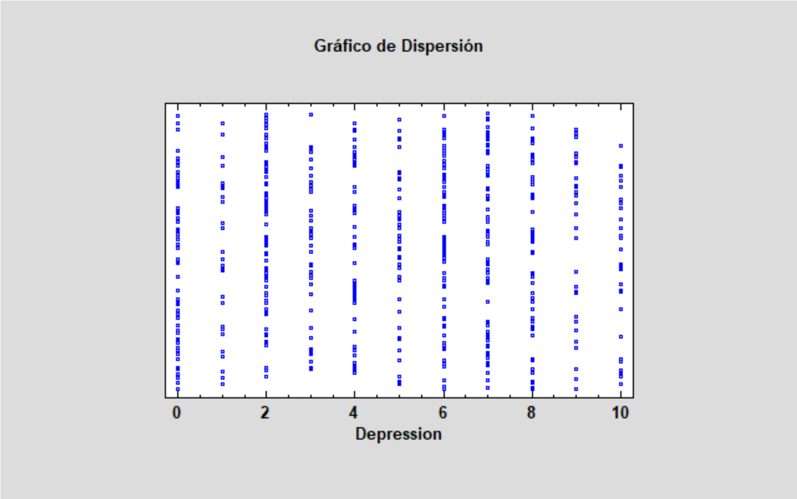
Tabla de Frecuencia para Depression

Clase	Valor	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Rel. acum.
1	0	65	0,1109	65	0,1109
2	1	30	0,0512	95	0,1621
3	2	71	0,1212	166	0,2833
4	3	44	0,0751	210	0,3584
5	4	53	0,0904	263	0,4488
6	5	43	0,0734	306	0,5222
7	6	76	0,1297	382	0,6519
8	7	75	0,1280	457	0,7799
9	8	63	0,1075	520	0,8874
10	9	32	0,0546	552	0,9420
11	10	34	0,0580	586	1,0000



As it can be seen in the Box and Whisker chart there are not atypical data in the sample. Furthermore, to describe its form, the standardized asymmetry coefficient of Fisher of –0,86 indicates a symmetry in the variable. It can be seen better in the bar chart and the scatter plot.





Regarding the position of the values, the median is 5, the lower quartile is 2 and the upper quartile is 7. Besides, about the dispersion of the sample, the interquartile range is 5.

Insomnia

This variable indicates the self-perception of the level of insomnia. It is a discrete quantitative variable, whose values range from 0 to 10.

The frequency table of the variable is the next:

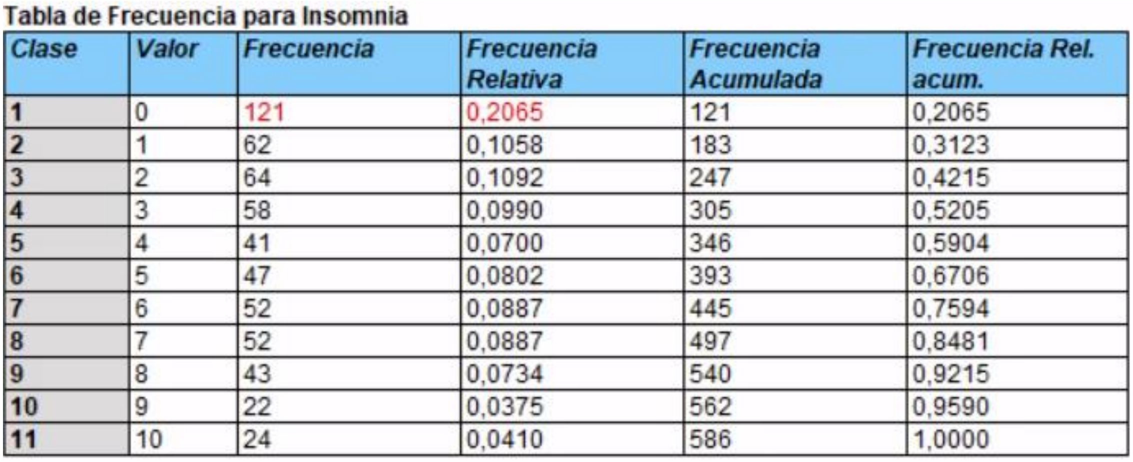
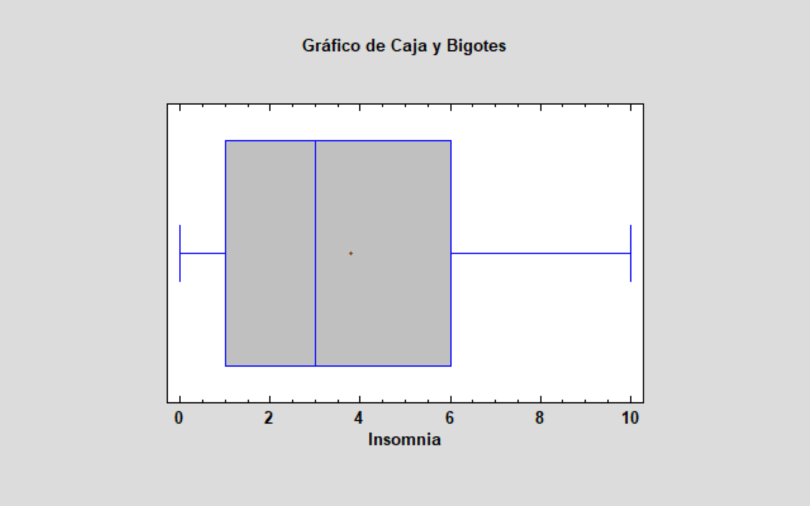
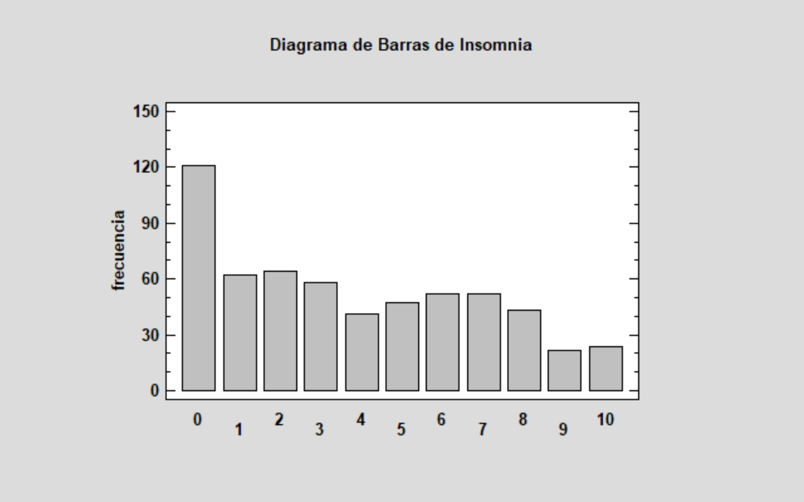


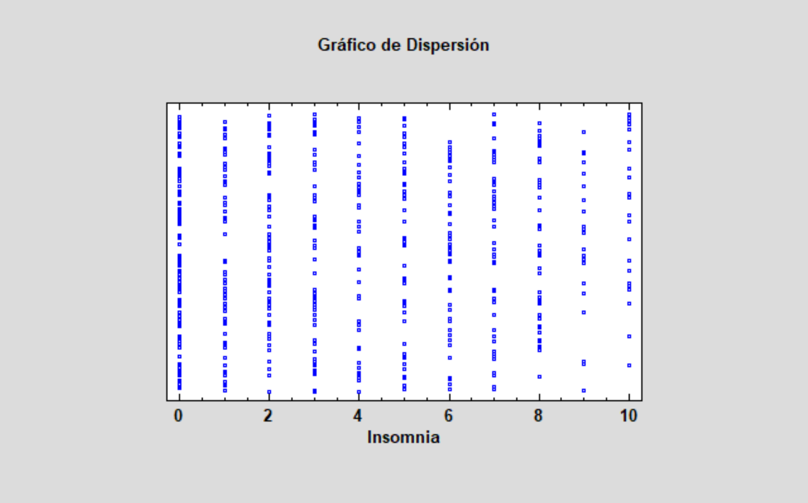
Tabla de Frecuencia para Insomnia

Clase	*Valor*	*Frecuencia*	*Frecuencia Relativa*	*Frecuencia Acumulada*	*Frecuencia Rel. acum.*
1	0	121	0,2065	121	0,2065
2	1	62	0,1058	183	0,3123
3	2	64	0,1092	247	0,4215
4	3	58	0,0990	305	0,5205
5	4	41	0,0700	346	0,5904
6	5	47	0,0802	393	0,6706
7	6	52	0,0887	445	0,7594
8	7	52	0,0887	497	0,8481
9	8	43	0,0734	540	0,9215
10	9	22	0,0375	562	0,9590
11	10	24	0,0410	586	1,0000



As it can be seen in the Box and Whisker chart there are not atypical data in the sample. Furthermore, to describe its form, the standardized asymmetry coefficient of Fisher of 3,48 indicates a positive asymmetry in the variable. It can be seen better in the bar chart and the scatter plot.





Regarding the position of the values, the median is 3, the lower quartile is 1 and the upper quartile is 6. Besides, about the dispersion of the sample, the interquartile range is 5.

OCD

This variable indicates the self-perception of the level of obsessive-compulsive disorder. It is a discrete quantitative variable, whose values range from 0 to 10.

The frequency table of the variable is the next:

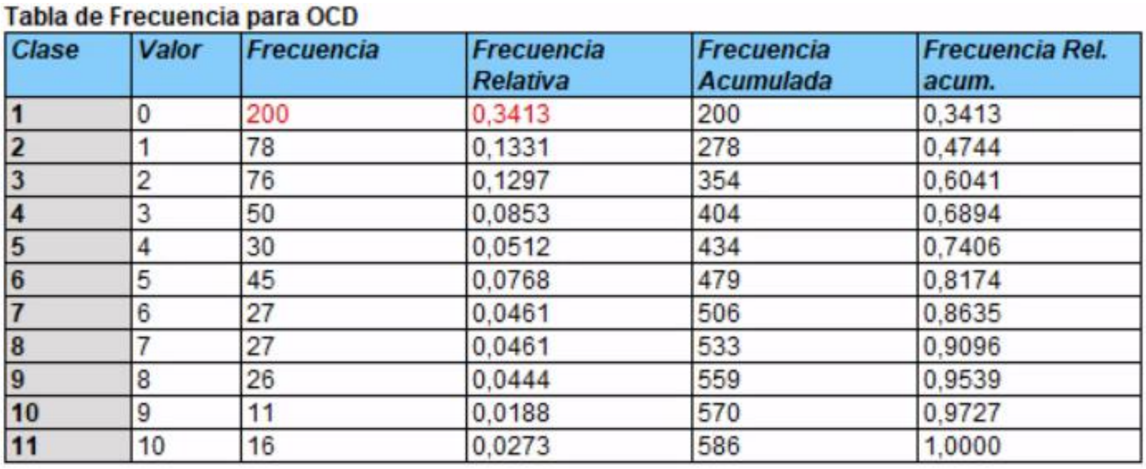
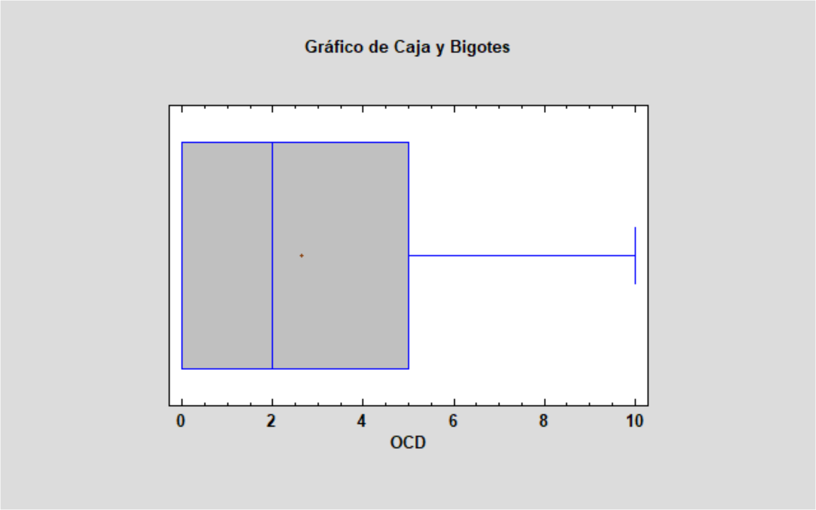
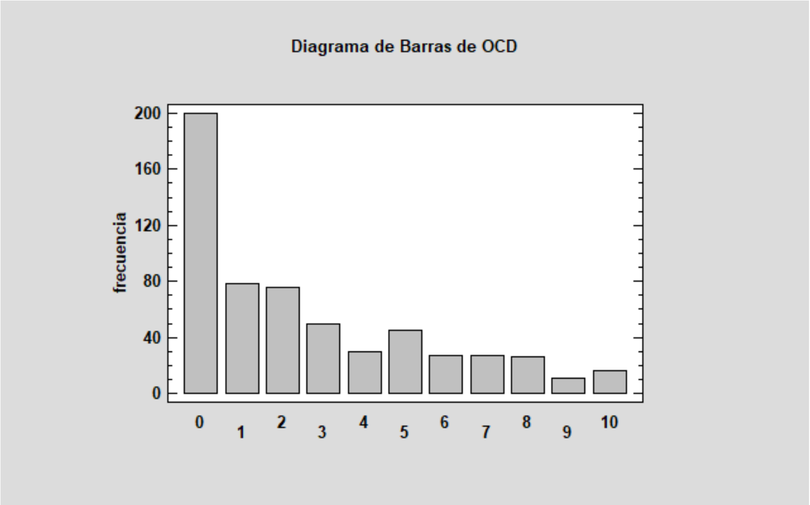


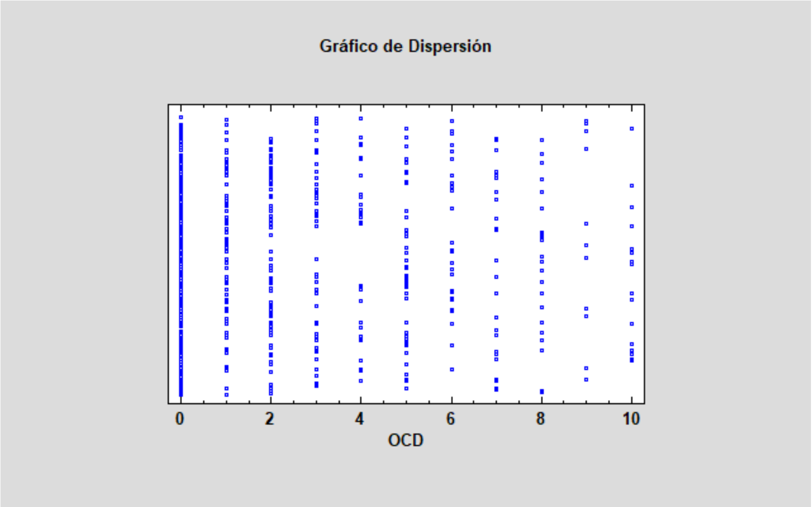
Tabla de Frecuencia para OCD

Clase	Valor	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Rel. acum.
1	0	200	0,3413	200	0,3413
2	1	78	0,1331	278	0,4744
3	2	76	0,1297	354	0,6041
4	3	50	0,0853	404	0,6894
5	4	30	0,0512	434	0,7406
6	5	45	0,0768	479	0,8174
7	6	27	0,0461	506	0,8635
8	7	27	0,0461	533	0,9096
9	8	26	0,0444	559	0,9539
10	9	11	0,0188	570	0,9727
11	10	16	0,0273	586	1,0000



As it can be seen in the Box and Whisker chart there are not atypical data in the sample. Furthermore, to describe its form, the standardized asymmetry coefficient of Fisher of 9,45 indicates a great positive asymmetry in the variable. It can be seen better in the bar chart and the scatter plot.





Regarding the position of the values, the median is 2, the lower quartile is 0 and the upper quartile is 5. Besides, about the dispersion of the sample, the interquartile range is 5.

Hours per day

This variable indicates the number of hours that each individual listens to music. It is a discrete quantitative variable, and its values go from 0 to 10.

The frequency table of the variable is the next:

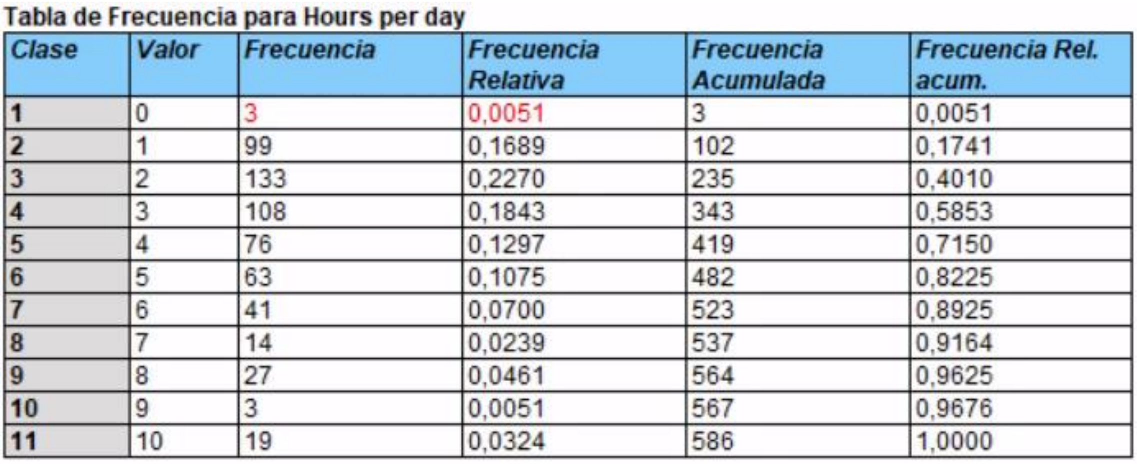
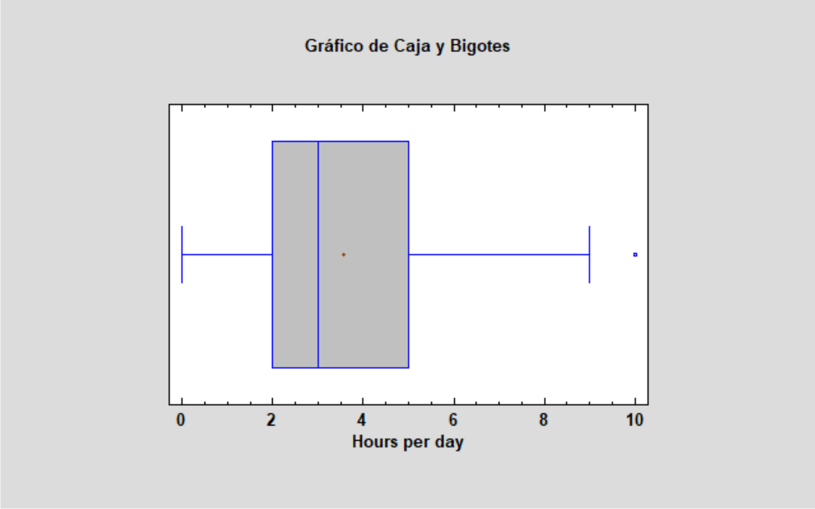
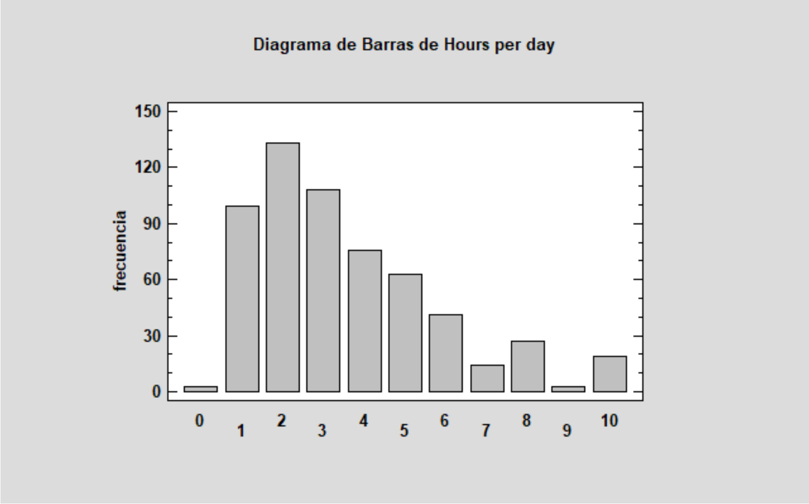


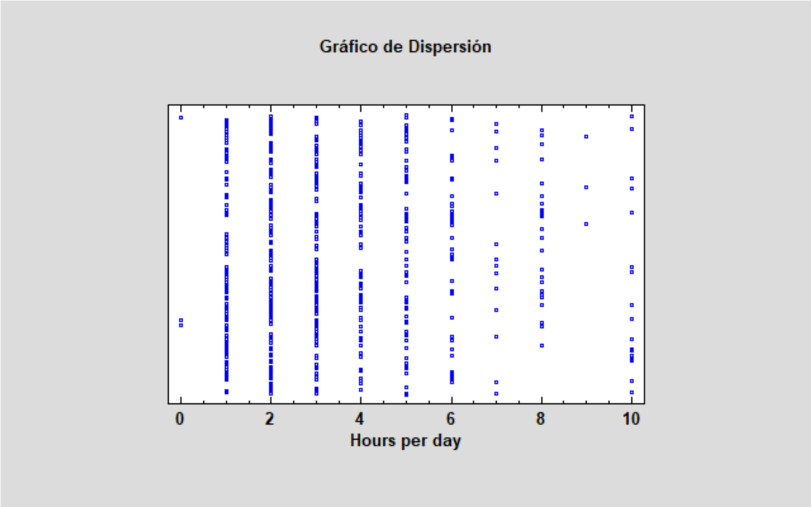
Tabla de Frecuencia para Hours per day

Clase	Valor	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Rel. acum.
1	0	3	0,0051	3	0,0051
2	1	99	0,1689	102	0,1741
3	2	133	0,2270	235	0,4010
4	3	108	0,1843	343	0,5853
5	4	76	0,1297	419	0,7150
6	5	63	0,1075	482	0,8225
7	6	41	0,0700	523	0,8925
8	7	14	0,0239	537	0,9164
9	8	27	0,0461	564	0,9625
10	9	3	0,0051	567	0,9676
11	10	19	0,0324	586	1,0000



As it can be seen in the Box and Whisker chart there are 19 atypical data in the sample, which value is 10 hours. Furthermore, to describe its form, the standardized asymmetry coefficient of Fisher of 10,14 indicates a great positive asymmetry in the variable. It can be seen better in the bar chart and the scatter plot.





Regarding the position of the values, the median is 3, the lower quartile is 2 and the upper quartile is 5. Besides, about the dispersion of the sample, the interquartile range is 3.

While working

Indicates if each individual listens to music while he or she is working. It is a binary or dichotomous qualitative variable, since it only has two values: “Yes” or “No”.

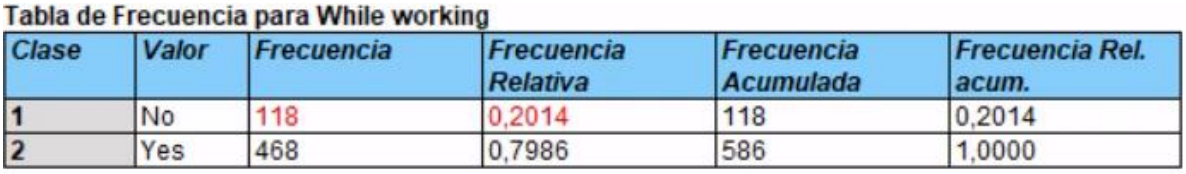
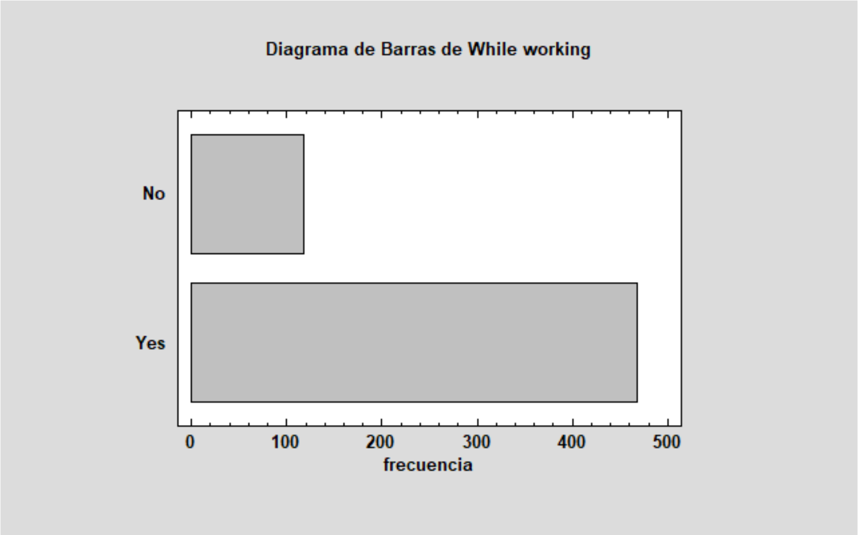
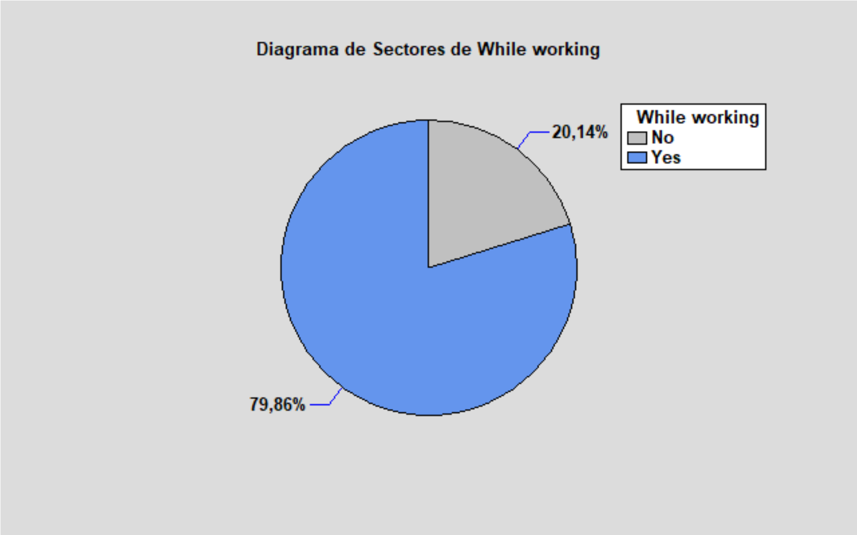


Tabla de Frecuencia para While working

Clase	Valor	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Rel. acum.
1	No	118	0,2014	118	0,2014
2	Yes	468	0,7986	586	1,0000

In the frequency table it can be observed that 468 individuals of 586 listen to music while they are working, a much higher value than the 118 that do not. The individuals that listen to music working are a 79,86% of the sample, while the ones who do not do it are a 20,14%. It is a thing easily appreciable in the bar chart and the pie chart.





Exploratory

In this variable is indicated if each individual is willing to search and explore new types and genres of music. It is a binary or dichotomous qualitative variable, since it only has two values: “Yes” or “No”.

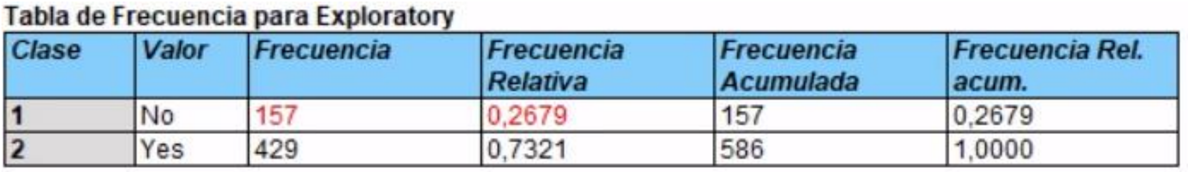
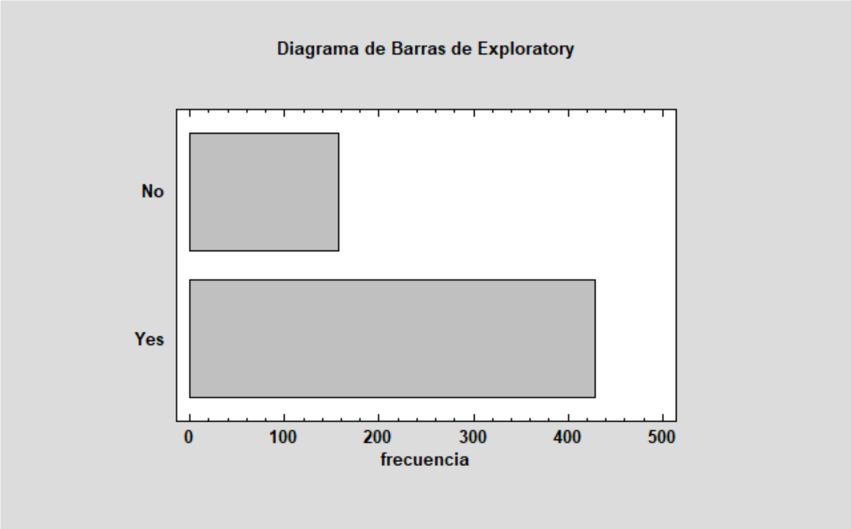
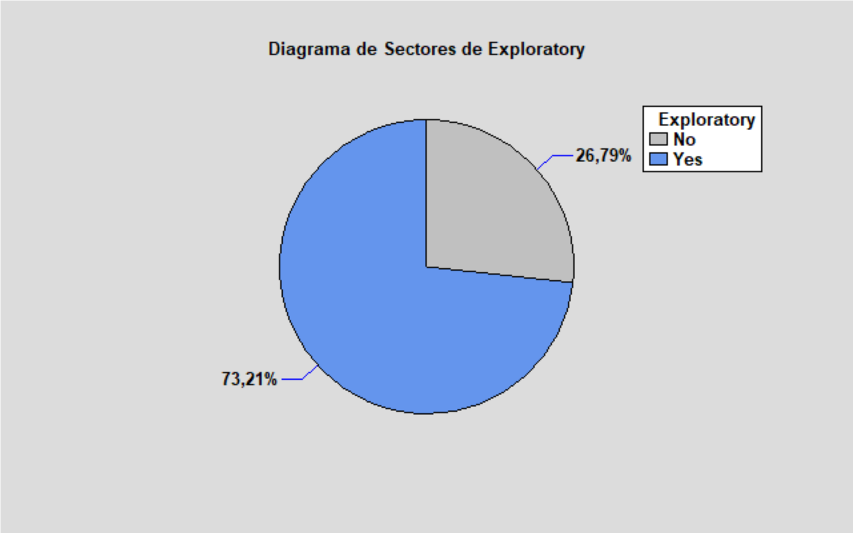


Tabla de Frecuencia para Exploratory

Clase	Valor	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Rel. acum.
1	No	157	0,2679	157	0,2679
2	Yes	429	0,7321	586	1,0000

In the frequency table it can be seen that the 73,21% of the individuals of the sample are willing to explore new music, instead, 26,79% are not willing to do it. The difference between both is observed in a better way in the bar chart and the pie chart.





Foreign languages

Here it is shown if the individuals listen to music in a different language of their mother tongue. It is a binary or dichotomous qualitative variable, since it only has two values: “Yes” or “No”.

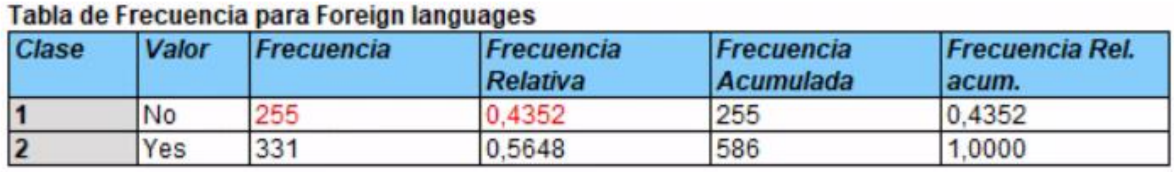
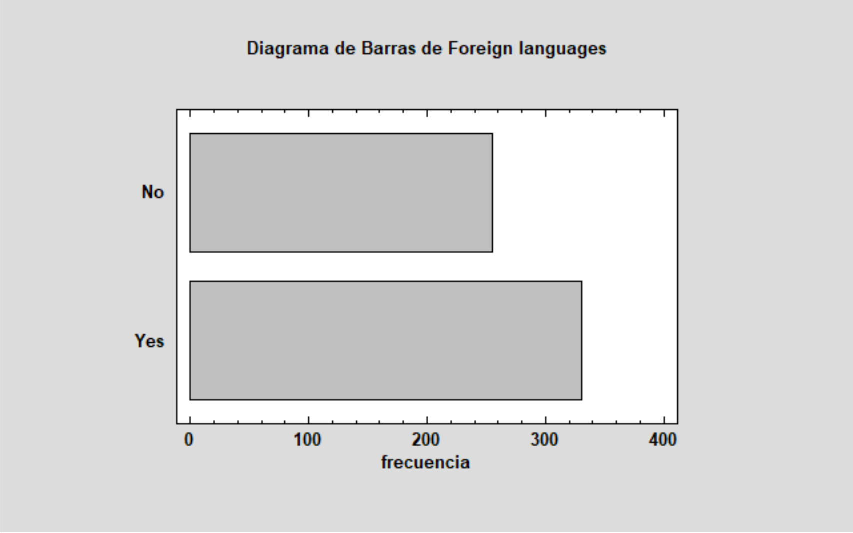
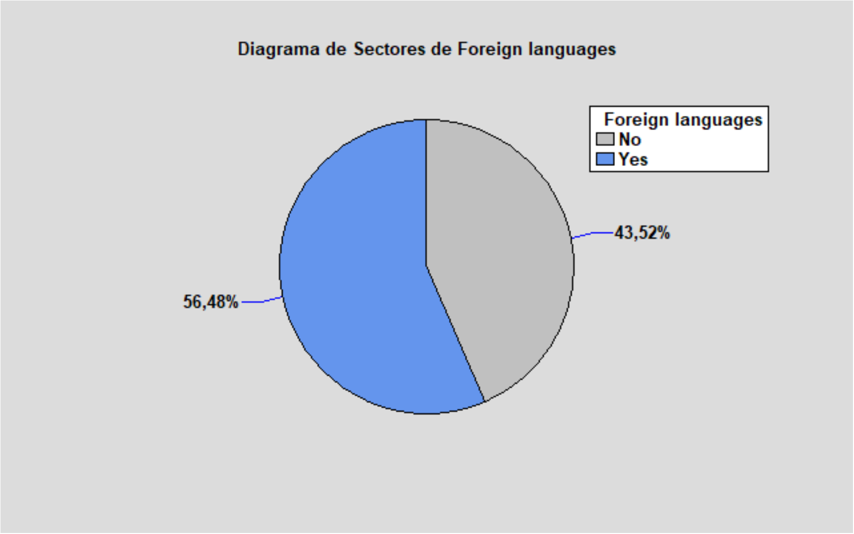


Tabla de Frecuencia para Foreign languages

Clase	Valor	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Rel. acum.
1	No	255	0,4352	255	0,4352
2	Yes	331	0,5648	586	1,0000

In the frequency table as in the bar chart and the pie chart, it can be appreciated that both values are quite even. A 56,48% of the individuals listen to music in a foreign language, while a 43,52% do not.





Composer

In this variable it can be seen if each individual of the sample composes or has composed any type of music. It is a binary or dichotomous qualitative variable, because it only has two values: “Yes” or “No”.

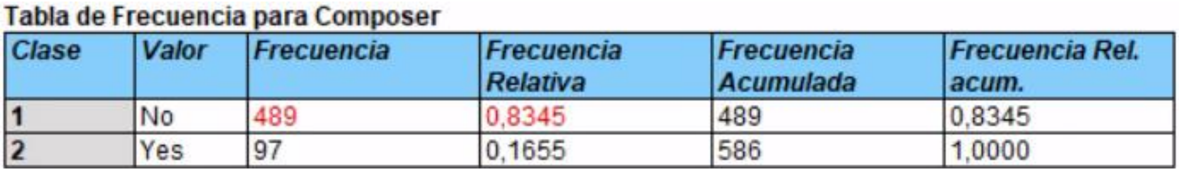
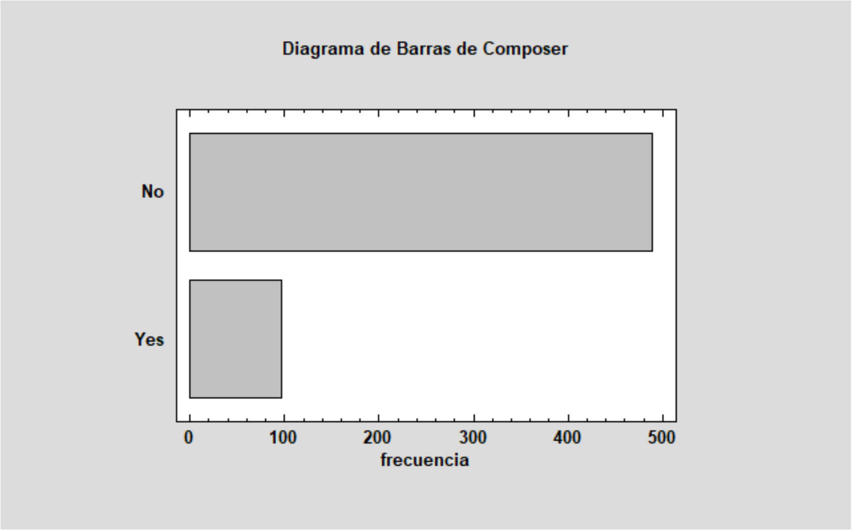
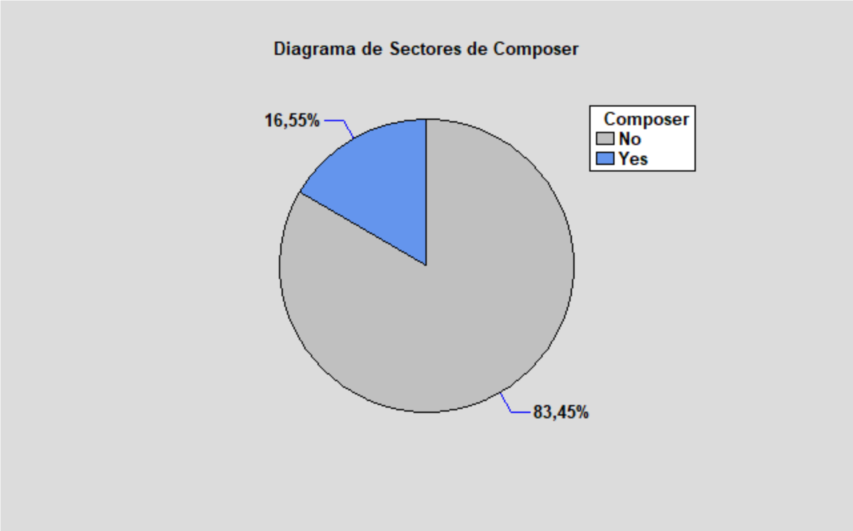


Tabla de Frecuencia para Composer

Clase	Valor	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Rel. acum.
1	No	489	0,8345	489	0,8345
2	Yes	97	0,1655	586	1,0000

In the frequency table the existence of a great difference between both values is appreciated. Between all the individuals, 489, a 83,45%, have never composed anything; while 97, a 16,55%, have composed something. This can be observed better in the bar chart and pie chart.





Music effects

This variable shows if the individuals feel that music improves or not their mood. It is an ordinal qualitative variable, due to the fact that the values follow an order. The values of the variable, ordered from best to worst, are: “Improve”, “No effect” and “Worsen”.

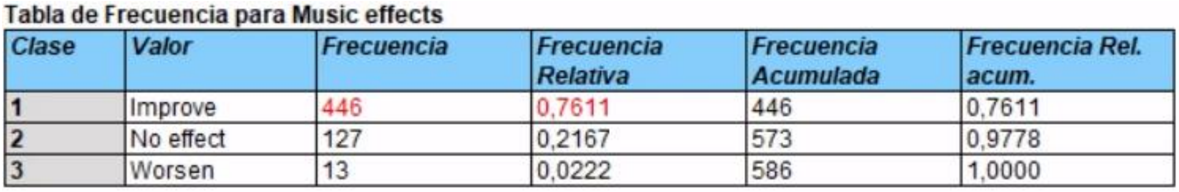
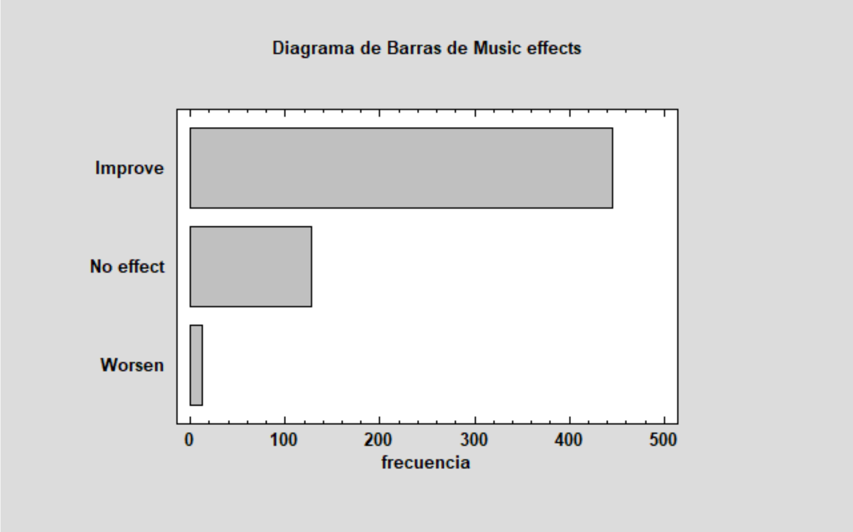
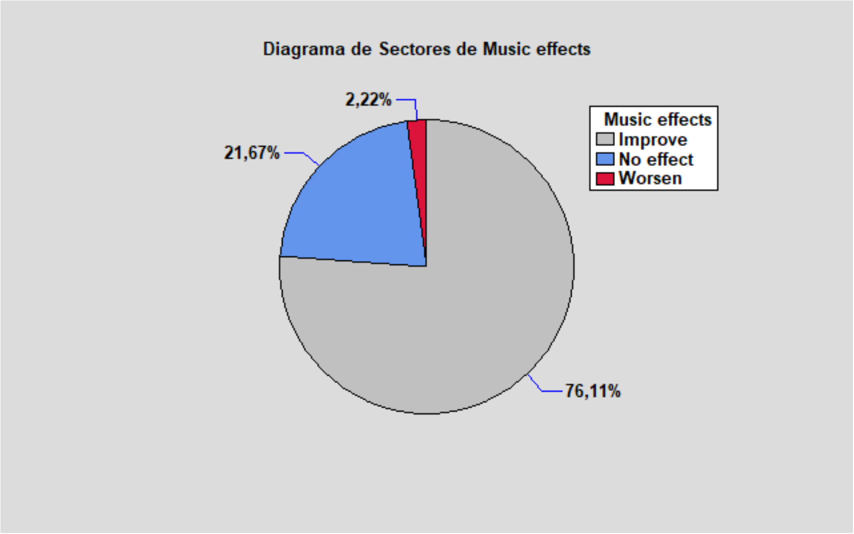


Tabla de Frecuencia para Music effects

Clase	Valor	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Rel. acum.
1	Improve	446	0,7611	446	0,7611
2	No effect	127	0,2167	573	0,9778
3	Worsen	13	0,0222	586	1,0000

In the frequency table it is seen that a 76,11% of the individuals of the sample think that music improves their mood, while a 2,22% think that it worsens their mood. Moreover, music does not affect to a 21,67% of the individuals in their mood. Between them, highlights that music affect to the mood with 446 of the people; this is better appreciable in the bar chart and pie chart.





Fav genre

This variable shows the favourite genre of music of each individual of the sample. It is a nominal qualitative variable, and its values are: “Classical”, “Country”, “EDM”, “Folk”, “Gospel”, “Hip hop”, “Jazz”, “K pop”, “Latin”, “Lofi”, “Metal”, “Pop”, “R&B”, “Rap”, “Rock” and “Video game music”.

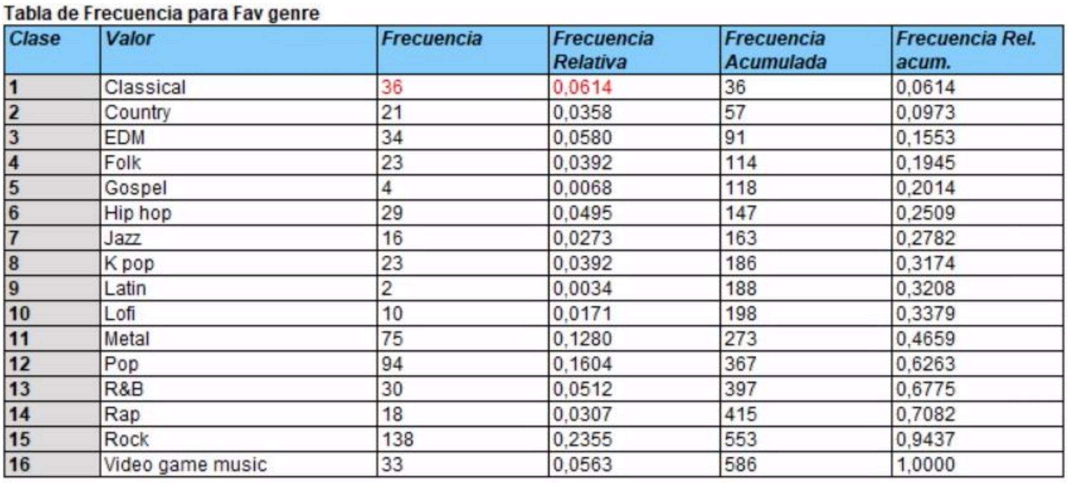
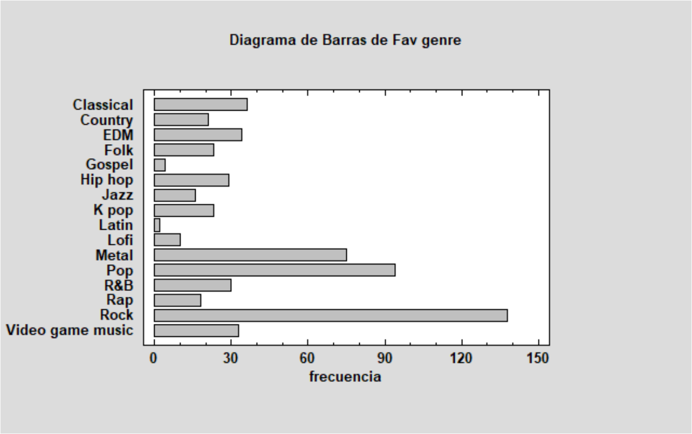
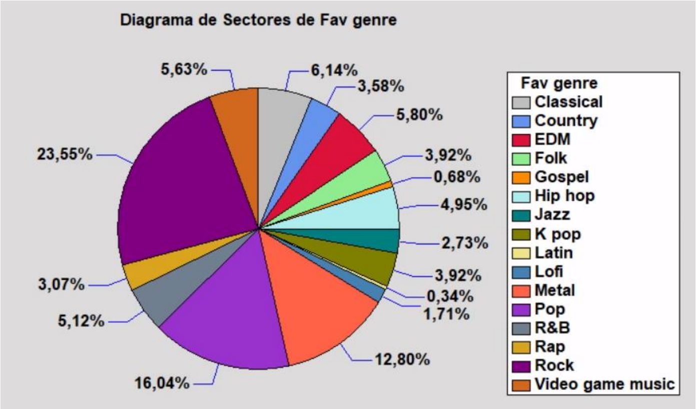


Tabla de Frecuencia para Fav genre

Clase	Valor	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Rel. acum.
1	Classical	36	0,0614	36	0,0614
2	Country	21	0,0358	57	0,0973
3	EDM	34	0,0580	91	0,1553
4	Folk	23	0,0392	114	0,1945
5	Gospel	4	0,0068	118	0,2014
6	Hip hop	29	0,0495	147	0,2509
7	Jazz	16	0,0273	163	0,2782
8	K pop	23	0,0392	186	0,3174
9	Latin	2	0,0034	188	0,3208
10	Lofi	10	0,0171	198	0,3379
11	Metal	75	0,1280	273	0,4659
12	Pop	94	0,1604	367	0,6263
13	R&B	30	0,0512	397	0,6775
14	Rap	18	0,0307	415	0,7082
15	Rock	138	0,2355	553	0,9437
16	Video game music	33	0,0563	586	1,0000

As it is seen in the frequency table, the three more chosen genres of music as favourite for the individuals are: “Rock”, with a 23,55% of people of the sample; “Pop”, with a 16,04%; and “Metal”, with a 12,8%. On the other hand, the three ones less chosen are: “Lofi”, with a 1,71% of the individuals; “Gospel”, with a 0,68%; and “Latin”, with a 0,34%. That can be seen more visually in the bar chart and the pie chart.





BPM

This variable shows the beats per minute in the favourite genre of every individual in the sample. It is a continuous quantitative variable, and its values range from 40 to 220 bpm.

This variable has been discretized in intervals of 40 in order to do the frequency table and the bar chart. Here are both:

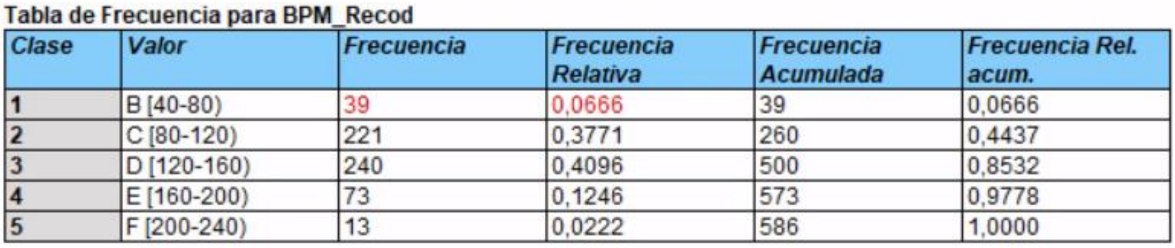
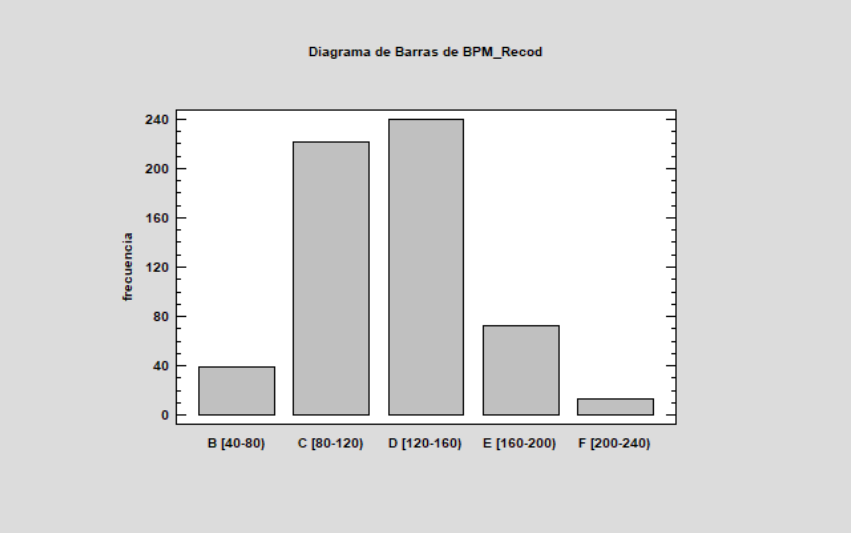
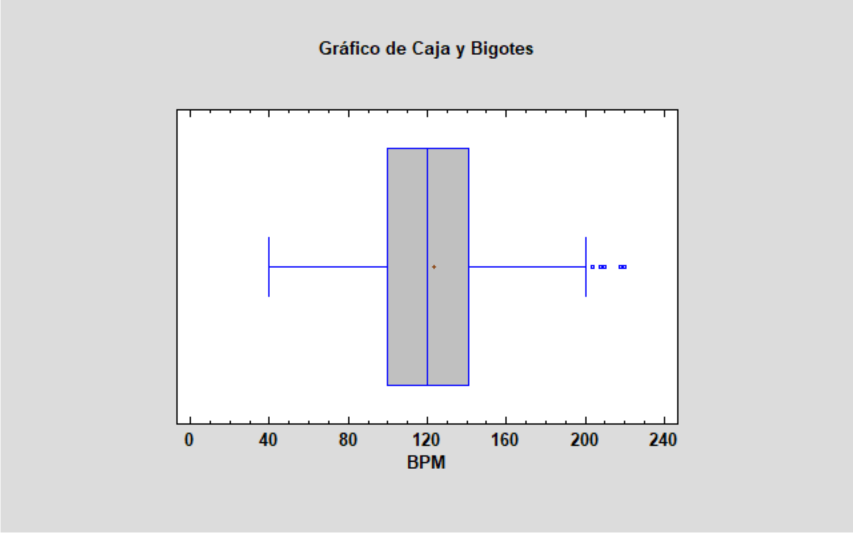


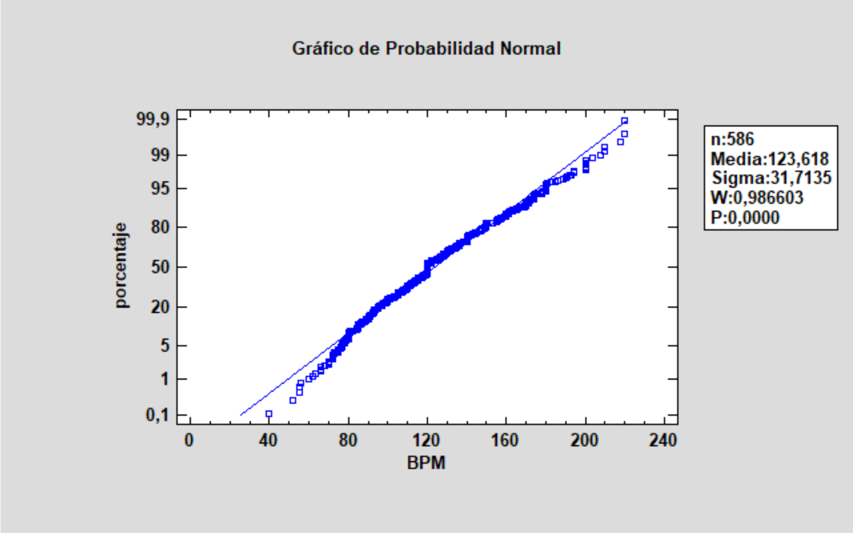
Tabla de Frecuencia para BPM_Recod

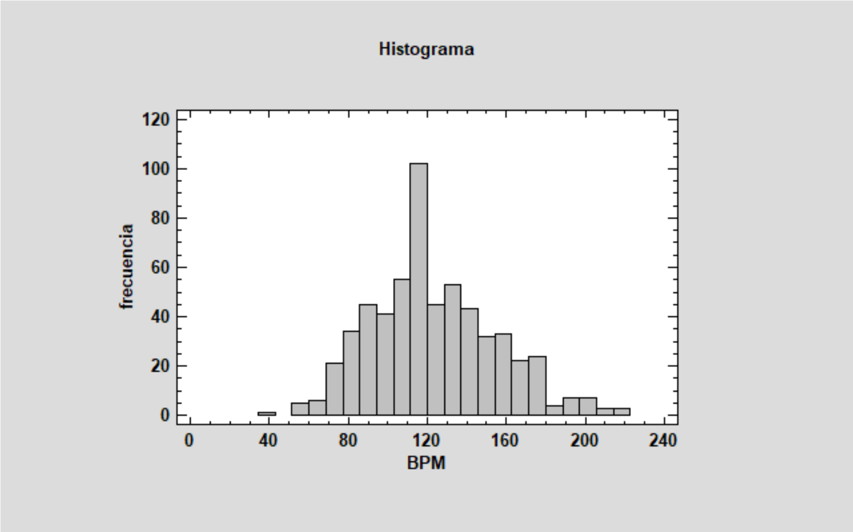
Clase	Valor	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Rel. acum.
1	B [40-80)	39	0,0666	39	0,0666
2	C [80-120)	221	0,3771	260	0,4437
3	D [120-160)	240	0,4096	500	0,8532
4	E [160-200)	73	0,1246	573	0,9778
5	F [200-240)	13	0,0222	586	1,0000

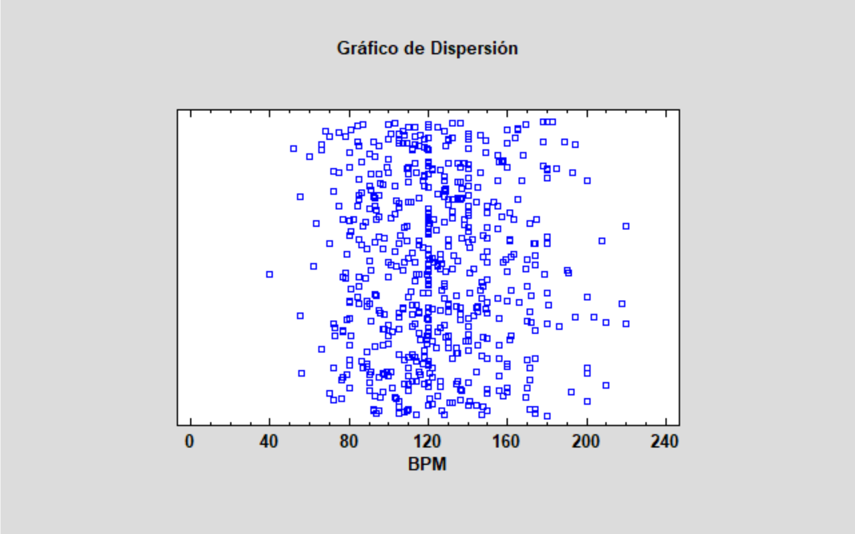




As it can be seen in the Box and Whisker chart there are 13 atypical data in the sample, which values go from 204 to 220 bpm. Furthermore, to describe its form, the standardized asymmetry coefficient of Fisher of 3,96 indicates a positive asymmetry in the variable. Moreover, it can be seen in a more visual way in the normal probability graph, the histogram and the scatter plot.







Regarding the position of the values, the median is 120, the lower quartile is 100 and the upper quartile is 141. Besides, about the dispersion of the sample, the interquartile range is 41.

Age

This variable shows the age of each individual in the sample. It is a continuous quantitative variable, and its values range from 10 to 80 years.

This variable has been discretized in intervals in order to do the frequency table and the bar chart. Here are both:

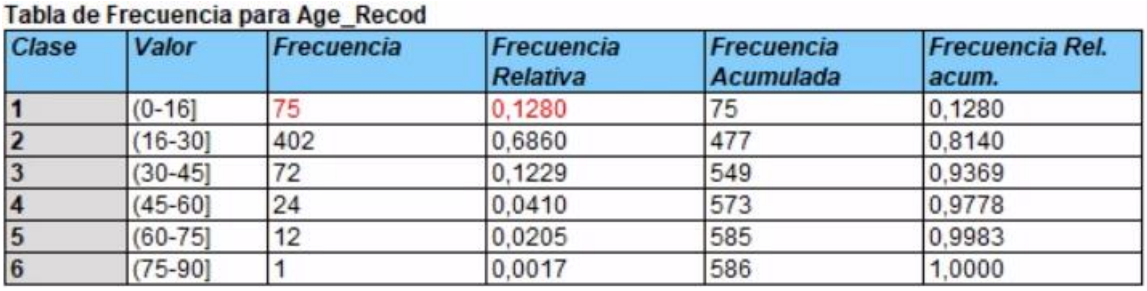
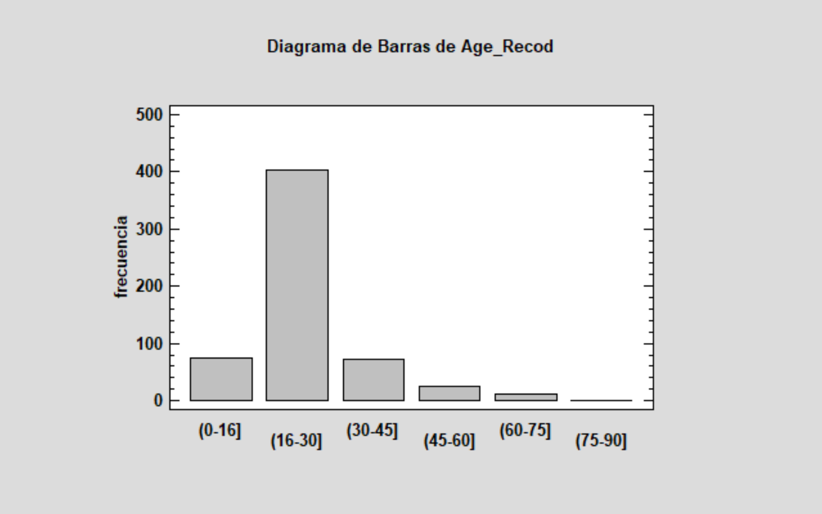
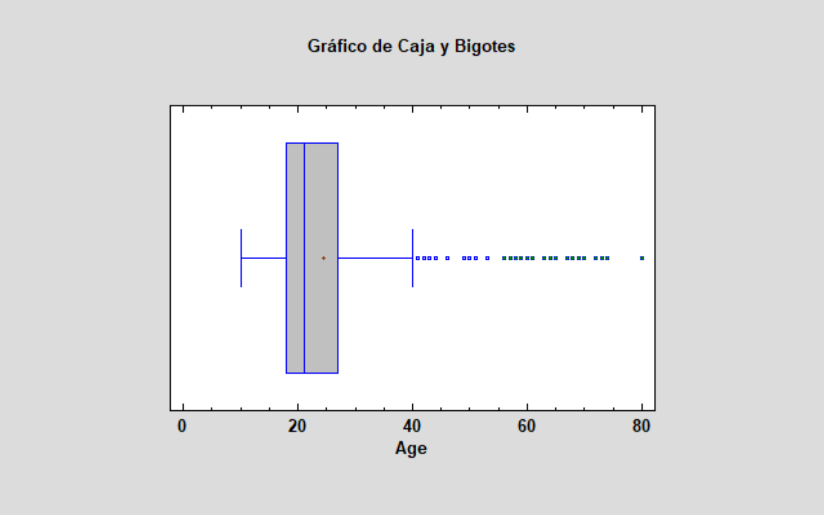


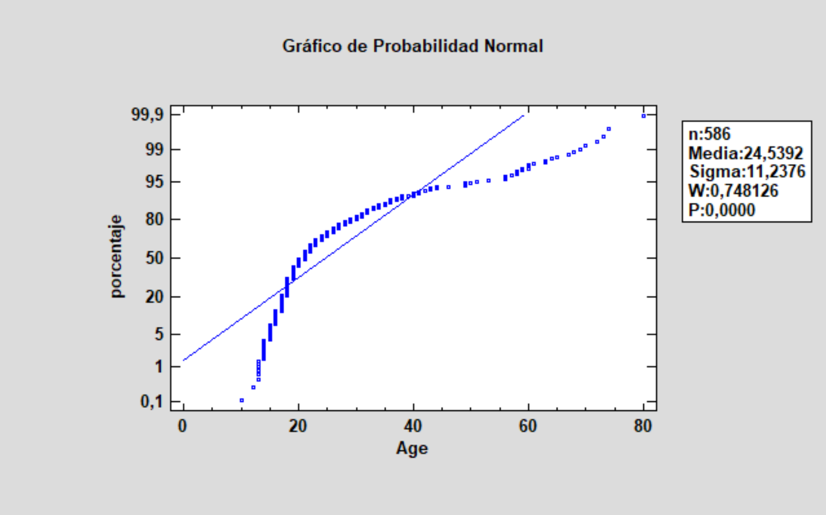
Tabla de Frecuencia para Age_Recod

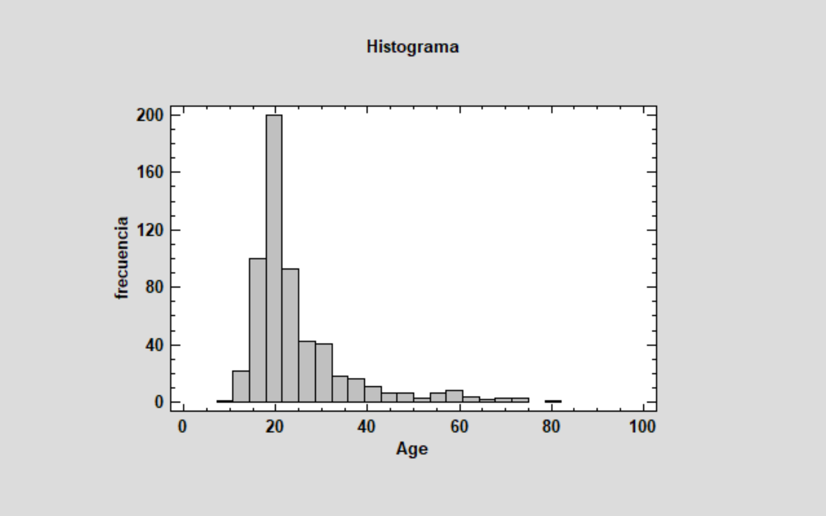
Clase	Valor	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Rel. acum.
1	(0-16]	75	0,1280	75	0,1280
2	(16-30]	402	0,6860	477	0,8140
3	(30-45]	72	0,1229	549	0,9369
4	(45-60]	24	0,0410	573	0,9778
5	(60-75]	12	0,0205	585	0,9983
6	(75-90]	1	0,0017	586	1,0000

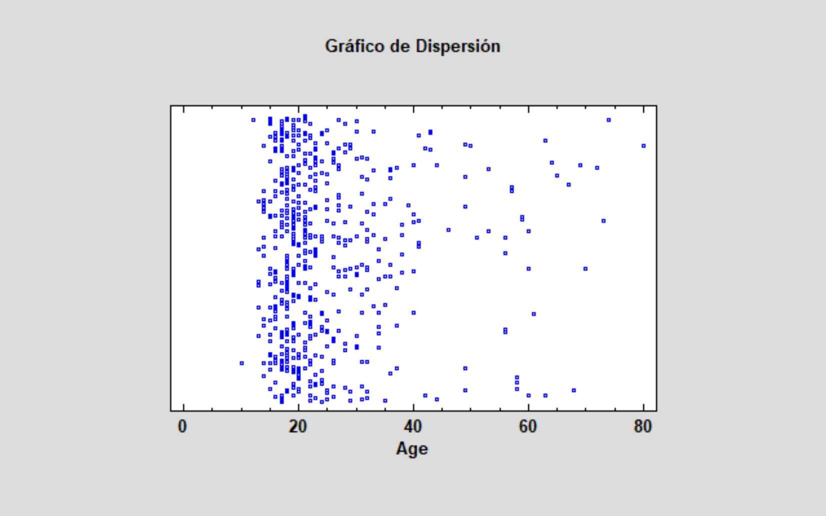




As it can be seen in the Box and Whisker chart there are 48 atypical data in the sample, which values go from 41 to 80 years. Furthermore, to describe its form, the standardized asymmetry coefficient of Fisher of 21,99 indicates a great positive asymmetry in the variable. Moreover, it can be seen in a more visual way in the normal probability graph, the histogram and the scatter plot.







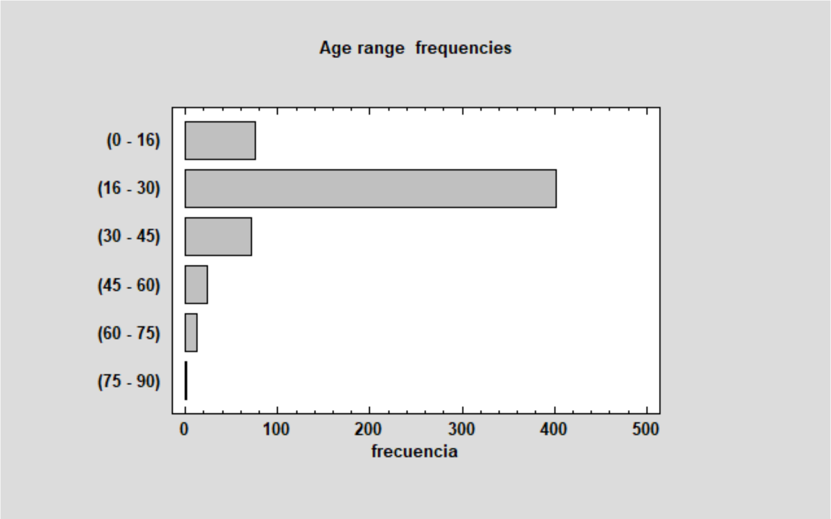
Regarding the position of the values, the median is 21, the lower quartile is 18 and the upper quartile is 27. Besides, about the dispersion of the sample, the interquartile range is 9.

Annex C (Multidimensional analysis)

As not every categorical variable has been treated, it would be adequate to bring them despite they don't show correlationship. This annexe helps to discard relationships that might seem logical but statiscally are not possible.

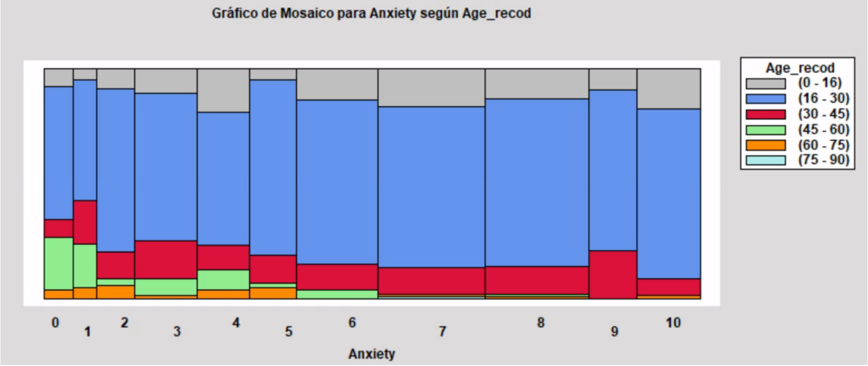
Multidimensional Analysis of Qualitative Variables

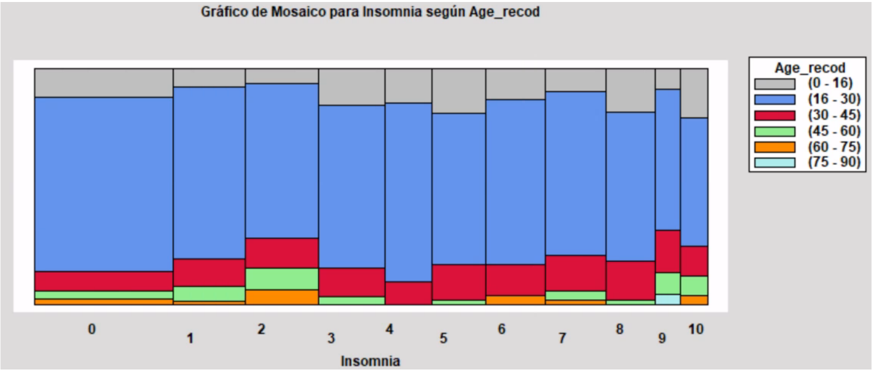
The distribution of the age ranges can be kwown with the variable **Age_recod**. Imagine that **Insomnia** is compared with **Hours per day**. It'll be necessary to check first if every range of age has the same distribution of **hours** listened. In that case, the whole sample will be used but if it is not, the conclusions will be brought to (16-30) range, as it is the largest. The image down below shows the frequencies of every age range. (16-30) range is the most popular.

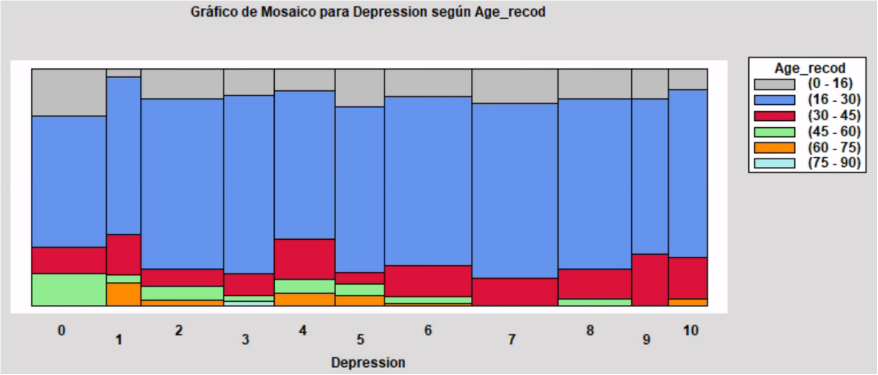


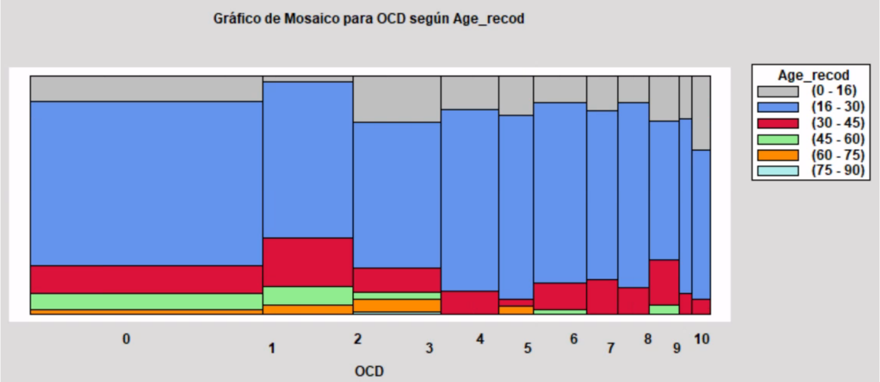
Age range – Anxiety/Depression/Insomnia/OCD

Do anxiety, depression, insomnia and OCD have the same distribution through every age range? The answer is Yes. Squares with the same color, which represent an age range, have similar sizes in every level of anxiety, depression, insomnia and OCD. This means these 4 variables share same distribution for every age range. This analysis is necessary in order to know whether it is better to make a relationship with one range or the every age range.



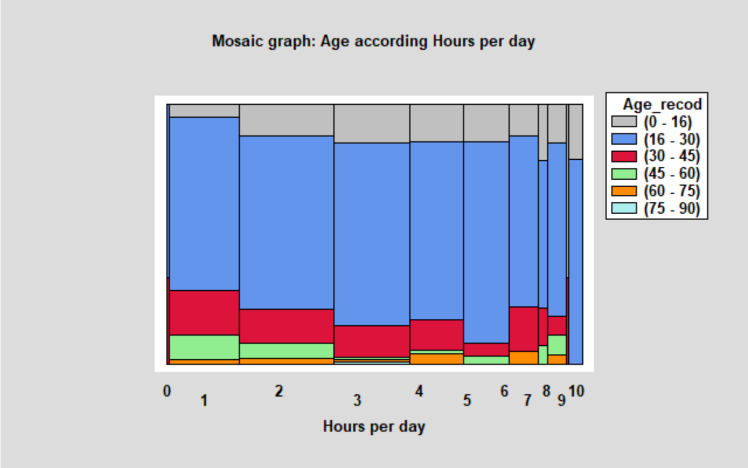






Hours per day – anxiety/depression/insomnia

Hours per day, Anxiety, Depression and **Insomnia** are ordinal qualititative variables (because of their few values) with values between 0 and 10. Let’s compare its distribution with **Age_recod** as said.



The proportions of each age range in every amount of **Hours per day** are quite the same. This means the hours of music you consume are not related to age.

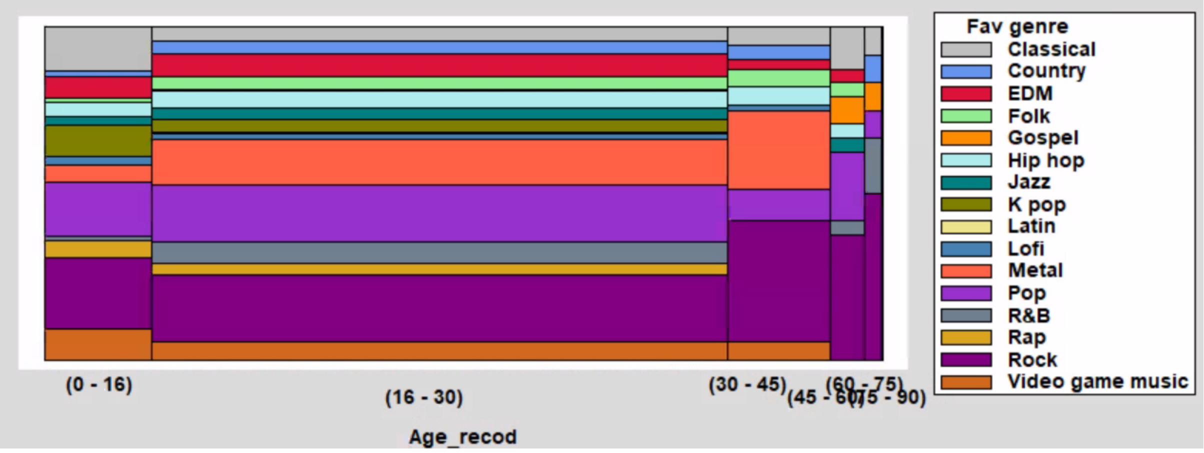
So that the variables **Hours per day** and can be considered related, Kendall parameter should approach to –1 or 1. In this case they’re clearly closer to 0 so there’s no relationship.

ANXIETY	DEPRESSION	INSOMNIA
Coef. De Contingencia 0,3131 Cramer's V **0,1475** Kendall's Tau b 0,0016	Coef. De Contingencia 0,3953 Cramer's V **0,1361** Kendall's Tau b 0,0914	Coef. De Contingencia 0,4037 Cramer's V **0,1395** Kendall's Tau b 0.0912

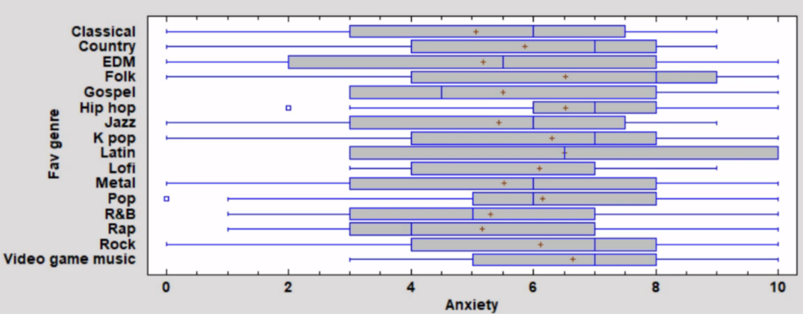
To listen more hours of music per day does not contribute to lower or higher levels of **Anxiety, Depression** or **Insomnia** than listening to less hours.

Fav-genre – anxiety/depression/insomnia

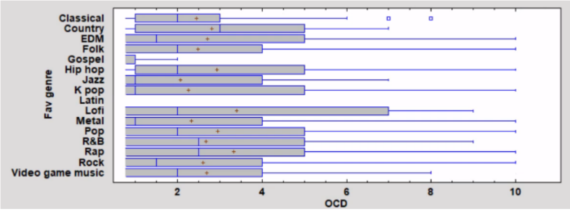
Fav genre is a qualitative variable with 16 possible values. Does it have the same distribution for every poblational group according to age?

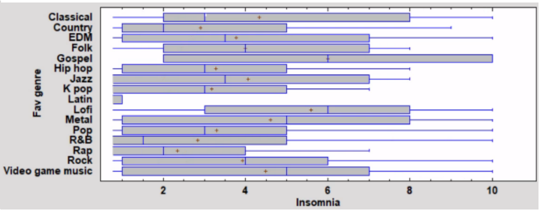


Every range has similar proportions of favourite genres so the whole sample will be used. Now, ¿Has any **genre** in particular higher values of **Insomnia, Anxiety, Depression** or **OCD**? This is relevant at the moment of determining if some genre leads generally to sadness or not.

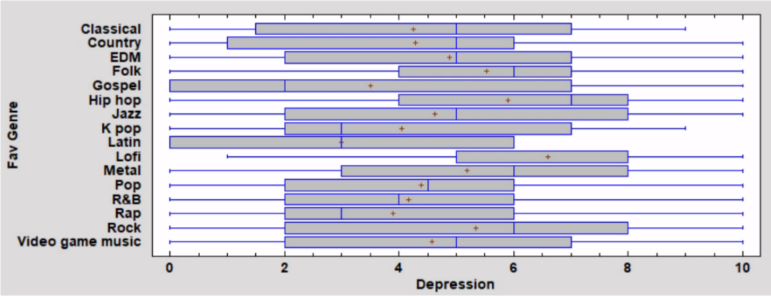


At first sight, every genre has similar distribution in terms of **Anxiety** at mid-high levels.





It is shown that listening to music in any genre gives low levels of **OCD** and **Insomnia**



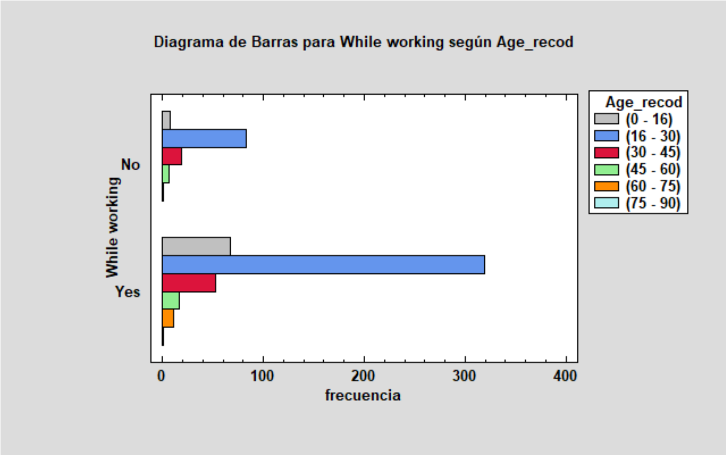
Excepting Latin and Gospel, due to having barely 4 values, every range get together in mid **Depression** levels.

No specific genre has been deviated from the rest drastically, meaning that there’s no need in picking a genre to focus on it.

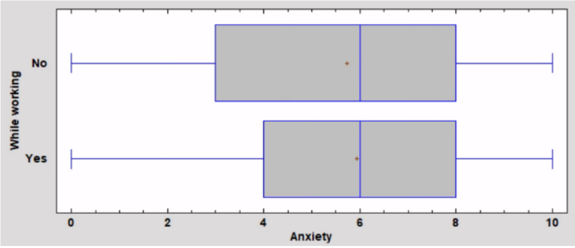
Cramer’s V is used when there are nominal qualitative variables. Besides, Kendall is used when they are ordinal. In this case there are nominal **(Fav genre)** and ordinal **(Insomnia, Depression, Anxiety)** variables. Contingency is removed because it only can be used in case there’are the same ammount of values in both variables. This requirement is not met. Cramer and Kendall rule, they’ll be our reference. None of them are closer to 1 than to 0. There’s no relation.

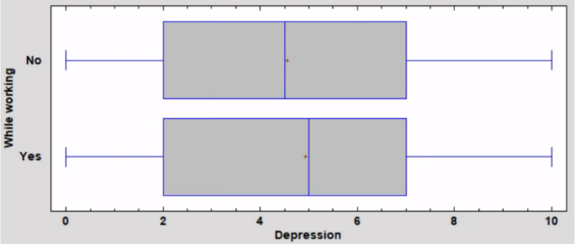
While working – anxiety/depression/insomnia

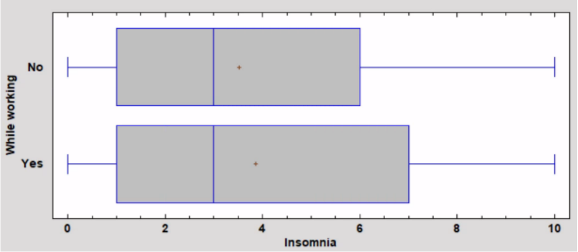
While working is a nominal qualitative variable with two possible outcomes, Yes or No. Will there be clearly higher levels of **Anxiety**, **Depression** or **Insomnia** for people who **listen to music while working**? As always, let’s see how **While working** is distributed through the ages.

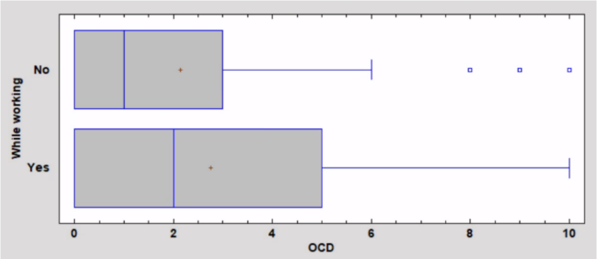


Every range follows a proportion between people who listen to music while working or not, so the whole sample will be used.







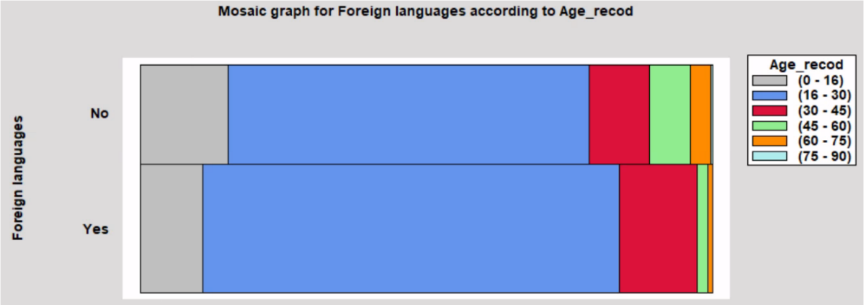


People who listen to music **while working** have higher OCD and **Insomnia** values than people who don't. It happens the opposite in Anxiety and Depression stays the same.

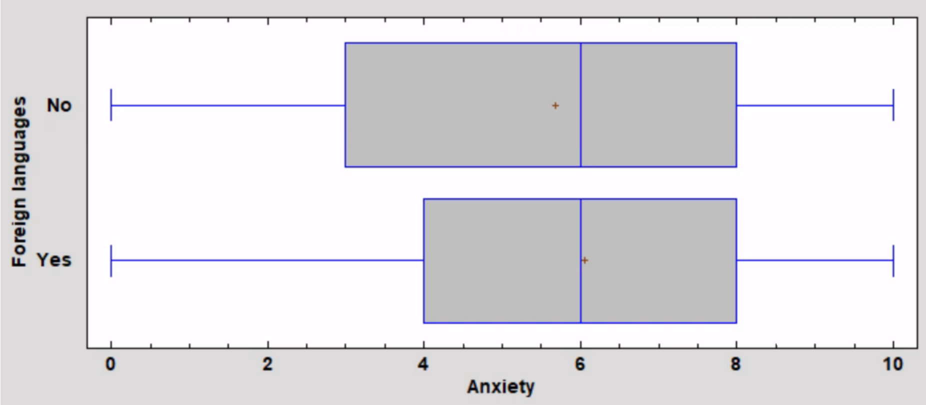
Listening to music while working doesn't affect at all the mental health

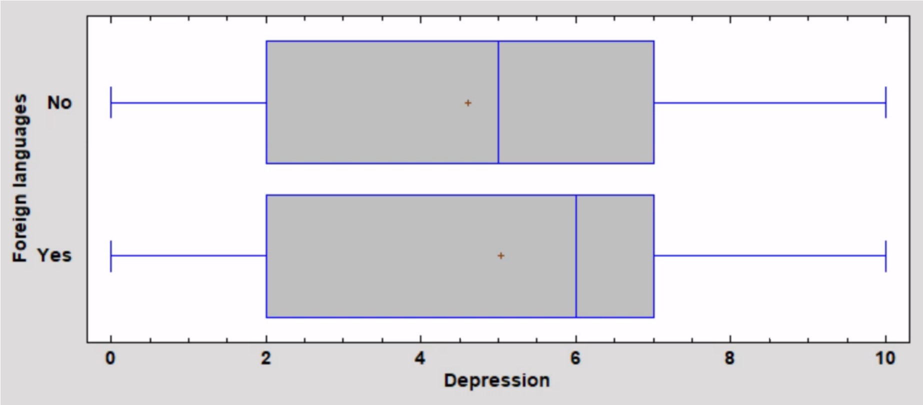
Foreign language – anxiety/depression/insomnia

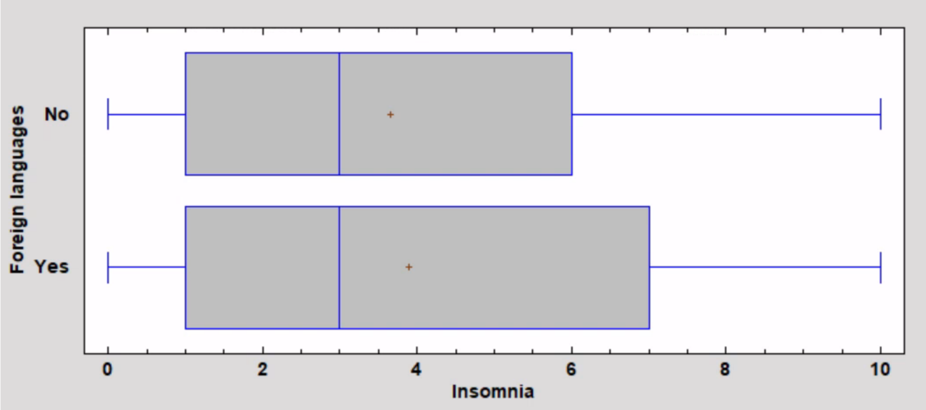
Will listening to **Foreign Language** somehow affect the levels of Anxiety, Depression or Insomnia?. First of all, let's see the how it is distributed among the ages.



The first age intervals, have similar amounts of individuals, besides, ranges are distributed pretty similar in both cases. The whole sample will be used.







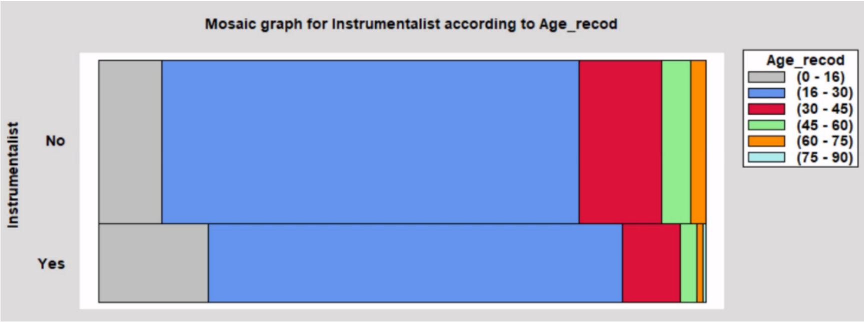
Both **Insomnia** and **Anxiety** have their respective couple of Box&Whisker graphs slightly different distributions. In the case of Anxiety, individuals who don't listen to **foreign music** have their values of Anxiety more scattered below than the ones who do. On the other side, individuals who listen to **foreign language** music have their values of Insomnia more scattered above than the ones who don't. This soft difference is not enough as the Kendall coeficient or Cramer's V are closer to 0.

Anxiety	Insomnia
Cramer's V 0,1758 Kendall 0,0637	Cramer's V 0,1484 Kendall 0,0303

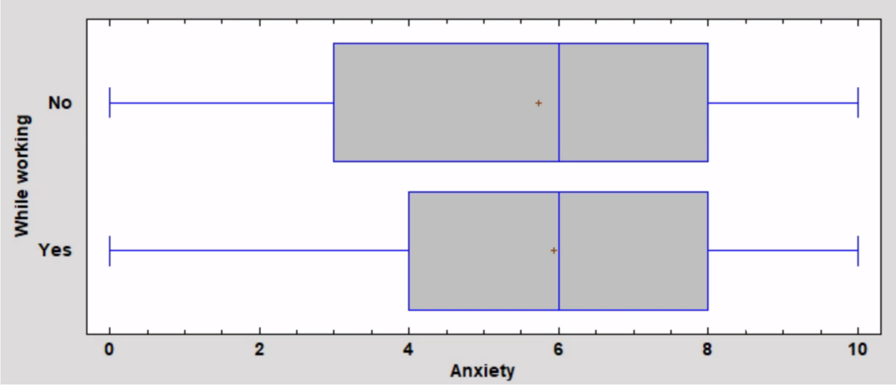
Being most likely to listen to foreign music doesn not affect **Anixety**, **Depression** or **Insomnia** at all.

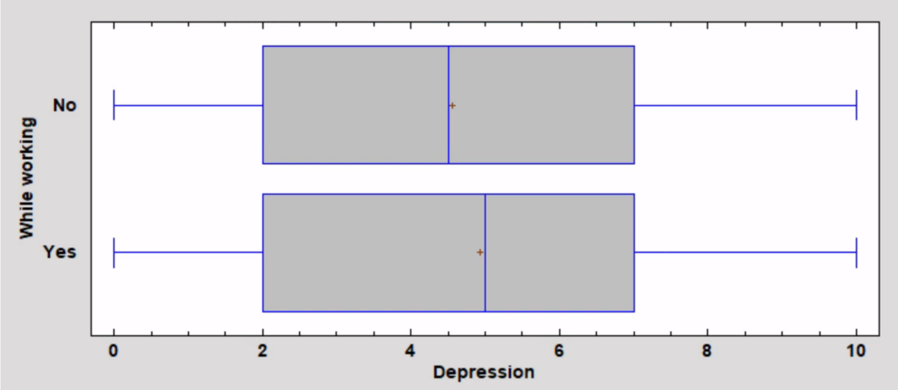
Instrumentalist – anxiety/depression/insomnia

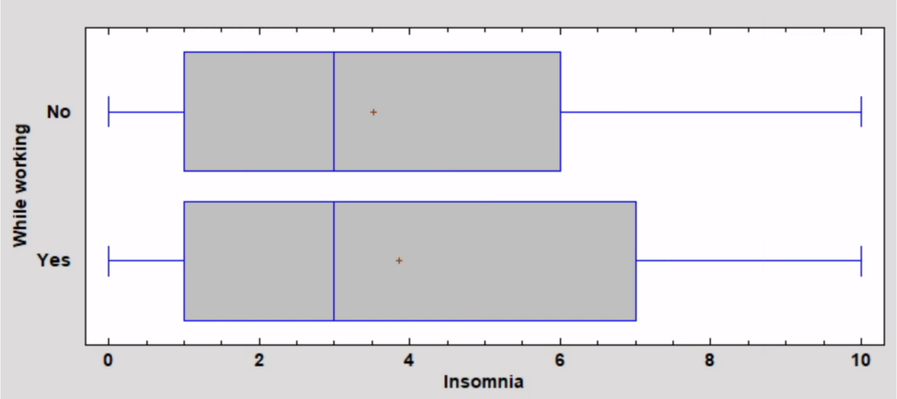
Is the simple fact of being an **instrumentalist** capable of changing the levels of anxiety, depression or insomnia while listening to music?



As expected, there are more non-instrumentalist (67,75%) than instrumentalists (32,25%). However the both kinf of indivuals share a similar distribution of age ranges, so the whole sample will be used.







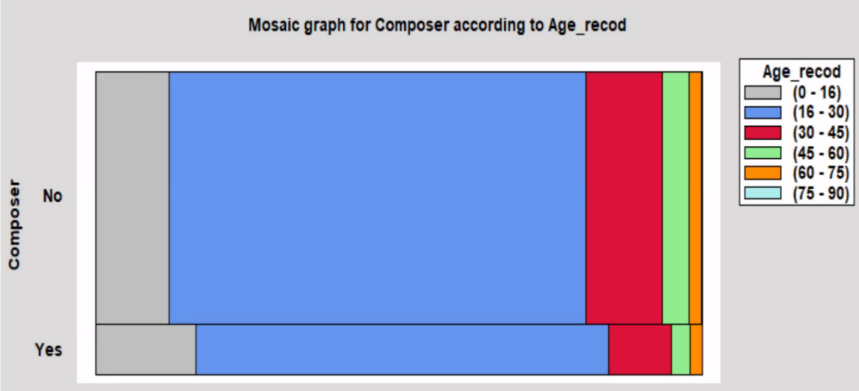
The only two factors that alter minimunly the **Instrumentalist'**s distribution are **Anxiety** and **Depression** although their couples have almost the same average. Again is not enough to determine a relationship. Descriptive parameters are brought to confirm the judgement.

Anxiety	Insomnia
Cramer's V 0,1102 Kendall 0,0378	Cramer's V 0,1170 Kendall 0,0506

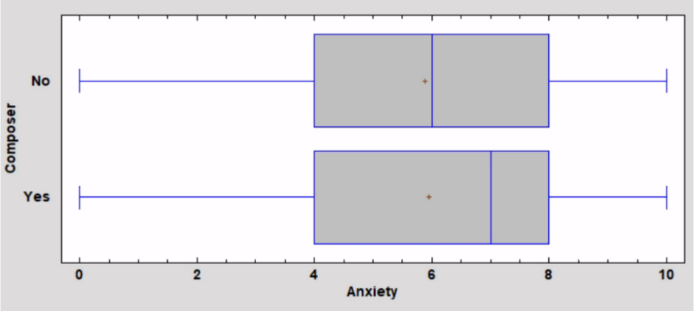
The exposure to music to which an instrumentalist is oftenly subjected doesn't affect the trio.

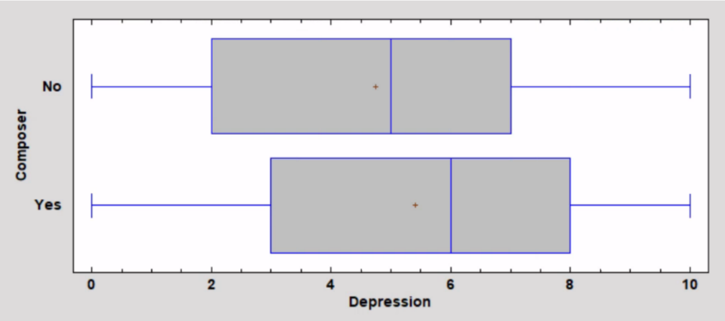
Composer – anxiety/depression/insomnia

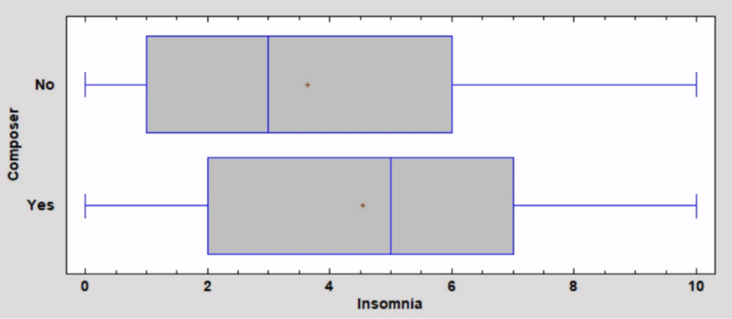
If being an instrumentalist does not affect, will being a composer affect the levels of **anxiety**, **depression** or **insomnia** at the time of listening to music? Firstly, let's compare the variable with the **age_recod**.



Just as thought, there are more composers (16,55%) than people who aren't (83,45%). Both Yes and No share approximately the same amount of values for every age range so the whole sample will be used.







People who don't compose have 50 % of their **Depression** and **Insomnia** values in lower levels than the ones who compose. Besides, there's a difference in the average too as they are lower in the non-composer individuals. The difference is not enough as the descriptive parameters don't get close to 1 in either Cramer or Kendall.

Depression	Insomnia
Cramer's V 0,1171 Kendall 0,0726	Cramer's V Kendall

People who are more familiar to music like instrumentalists or composers have the same distribution in the variable trio than people who don't.