Bootcamp

Data Science

Project title

Does music have an impact on the levels of anxiety, depression, insomnia or compulsive disorders in people?

Analist

Damian Lazos

Technical manual

Technologies summary **3**
Process notes **4**

Technologies summary

Technology	Function
Pandas	Dataset manipulation
Altair	Visualization
Streamlit	App framework
Scikit-Learn	Statistics and Machine learning
Jupyter	Notebook
Jupytext	Python script
Github	Repository service

Process notes

Notebook cell	Process	Description
[6]	Load dataset	It was used the Pandas library to manage datasets. The data was imported using the read_csv() method.
[9]	EDA - View missing data	It was used the “missingno” external library to have a look of the null data using the matrix() method.
[10]	EDA - View missing data	It was used the “missingno” external library to quick view of the correlation between columns with missing values using the heatmap() method
[11]	EDA - View missing data	It was used the “missingno” external library to count true or filled values per column using the bar() method
[12]	EDA - Outliers recognition	It was created the function: createBoxplotChart() This function creates a single boxplot chart to identify outliers using quartiles. Its code is in cell [5].
[13]	EDA - Missing data	It was created the function: calculateMeanOrModePerColumnWithMissingValues() This function helps to get the mean or mode of a given column. That would help to fill up each cell of a given column with its regarding mean o mode based on Favorite Genres Groups in order to fill NA or empty values of every single column with missing values
[14]	EDA - Outliers treatment	Using .loc[ ] and max() methods was founded the outlier was found in the

		column “BPM”.
[15]	EDA - Outliers treatment	Using the at[ ] method the outlier was replaced by the mean of bpm for the Favorite genre 'Video game music'
[17]	EDA - Missing values replacement	It was created the function: fillUpEmptyFieldsByColumn() This method fills up the empty cells of the dataset by column with its respective mean or mode depending on the Favorite Music Genre the respondent selected. It was coded on cell [5].
[21]	EDA - Columns types transformation	It was casted the type of each column in the dataset using the astype() method of pandas.
[22]	EDA - Dataset depuration	Unuseful columns were dropped using the drop() method of pandas.
[23]	EDA - Variables analysis	It was created the function: graphPeopleByPathologyAndMusicGenre() This function creates a single graphic bar that shows the total number of people listening to each music genre with certain frequency VS the level of highness they reported for each pathology.
[24]	EDA - Variables analysis	Using the method cut() of pandas, a series of ranks to group the data by the BPM of music listened.
[25]	EDA - Data process	A new column into the dataset is created to specify whether BPM score is higher than 90 or not using the LabelEncoder() method of scikit learn.
[26]	EDA - Variables analysis	It was created the function: calculateMeanPerListenersOverAndUnderNinetyBPM()

		This function calculates the mean over the total people listening to music over and under 90 BPM levels by a given pathology name and returns a bar chart showing the results to compare both means.
[27]	EDA - Data process	It was transformed the data into numerical values using the OneHotEncoder() object from the scikit learn library.
[28]	EDA - Data process	It was normalized all the data contained in the dataframe using the StandardScaler() method of scikit-learn.
[29]	Modeling - Logistical Regression model	It was trained a Logistical Regression model using the LogisticRegresion() object of scikit learn.
[30]	Modeling - Performance report	It was created a performance report using the classification_report() of scikit learn.