Dataset Exploration (Part 4)

bdat1005 | Math for data analytics

Submitted by | Navid Saleh Sadik (200544271)

Submitted to | Dr. Ehsan Pourjavad

Submitted on | 04/16/2023

Contents

[Introduction 2](#_Toc132519761)

[Dataset and Workflow Description 2](#_Toc132519762)

[Data Dictionary 4](#_Toc132519763)

[Research Questions 6](#_Toc132519764)

[Univariate Analysis 7](#_Toc132519765)

[Hypotheses test 11](#_Toc132519766)

[Interpolation and Extrapolation 14](#_Toc132519767)

[Reference 15](#_Toc132519768)

# Introduction

The document contains a Dataset Exploration report based on a chosen dataset. This 4 part exploration has primarily revolved around choosing an appropriate dataset, developing FINER research questions, cleaning and coding relevant data, univariate and multivariate statistical analyses on attributes/variables in focus, developing hypotheses and finally perform regression analysis and forecasting. The later sections cover detailed explanation on the exploration.

# Dataset and Workflow Description

The dataset that has been chosen for exploration is ‘Spotify Tracks Dataset’ by Maharshi Pandya, collected from Kaggle, and covers data on 11,400 individual soundtracks. So there are 11,400 records each portraying 20 different features. The dataset was collected and further cleaned by the author using Python and Spotify’s web API.

As the dataset has been recorded to be published and modified 3 months ago, an assumption has been made that the data depicts the scenario of the year 2022. It contains a fair mix of qualitative and quantitative data and appears to be full of insights to extract information from. Although the data is rich, it could have had one or two extra columns containing variables like ‘likes count’, ‘times streamed’. The research questions formed for the project in one of the following sections do not need any other data that has not been provided in the excel spreadsheet. It is safe to say that it has the potential to generate a ‘Song Recommendation System’ based on the dataset.

Upon choosing the dataset, 6 research questions have been formed, which are presented in a later section. The dataset has been further cleaned and modified according to the research questions which led to smoother and more efficient analyses. This part of the exploration consisted of:

* Cleaning the dataset:
  + Got rid of rows consisting of empty fields.
  + Got rid of irrelevant variables for the analysis, such as: Speachability, Instrumentality and liveness.
  + Got rid of duplicate data.
  + Got rid of rows consisting of Outliers.
* Coding variables and creating dummy variables for simplifying the analysis process for the future:
  + Tempo >> Tempo\_coded, Duration >> Duration\_coded, Explicit >> Expicit\_coded.
* Commencing univariate analysis of the variables:
  + Genre, Popularity, Albums, Valence, Explicit\_coded, Tempo\_coded, Duration\_coded and Key.

In the further exploration stage, 5 Hypotheses had been formed and tested according to the needs of research questions. This stage of the exploration consisted of the below hypotheses:

* ANOVA\_Duration Hypothesis: There are significant differences in song popularity amongst different Duration Groups. The variables Popularity and Duration\_Coded (coded in the previous stage) were filtered and sorted and presented in a new sheet called ‘Anova data’. Then using Data Analysis tool from excel, ANOVA analysis has been done. Please refer to the Screenshots.
* ANOVA\_Tempo Hypothesis: There are significant differences in song popularity amongst different Tempo Groups. The variables Popularity and Tempo\_Coded (coded in the previous stage) were filtered and sorted and presented in the stated sheet ‘Anova data’. Then using Data Analysis tool from excel, ANOVA analysis has been done. Please refer to the Screenshots.
* T-test\_Valnece Hypothesis: The popularity of Negative music (Sad/Dark songs) is higher than Positive music (Happy songs). The variable ‘Valence’ has been sorted to show below 0.5 and above 0.5, respectively representing Music with Negative human emotion and Positive human emotion. Then using Data Analysis tool from excel, Descriptive statistics have been presented and T-test has been conducted. Please refer to the screenshots.
* T-test\_Genre Hypothesis: The POP Music genre is more popular than ROCK Music genre. The variable ‘Genre’ has been sorted to show Popularity of POP and ROCK types. Then using Data Analysis tool from excel, Descriptive statistics have been presented and T-test has been conducted. Please refer to the screenshots.
* Odds Ratio Hypothesis: Odds of a song being Popular were higher in those with Explicit lyrics, compared with those without Explicit lyrics. By filtering and sorting Explicit\_coded and Popularity variables, 4 columns were created
  + Ex1Np (Songs with explicit lyrics, Not popular)
  + Ex0Np (Songs without explicit lyrics, Not popular)
  + Ex1P (Songs with explicit lyrics, Popular)
  + Ex0P (Songs without explicit lyrics, Popular)

Then using Data Analysis tool from excel, Descriptive statistics have been presented and T-test has been conducted. Please refer to the screenshots.

For the next phase, Interpolation (regression analysis) and extrapolation (forecast) has been performed which are showcased in later section of the document.

# Data Dictionary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Column** | **Data type** | **Variable type** | **Ranges** | **Description** |
| track\_id | string | nominal |  | Unique ID of the soundtrack |
| artists | string | nominal |  | The name of the musician or band or orchestra |
| album\_name | string | nominal |  | Name of the album the soundtrack was released in |
| track\_name | string | nominal |  | Name of the soundtrack |
| popularity | number | continuous | 0-100 | Spotify’s algorithm-based popularity of the song (0-100, 10 being the most popular) |
| duration\_ms | number | continuous | 0.00-5.24 | The length of the soundtrack in milliseconds |
| duration\_coded | number | continuous | 1-6 | Duration < 199999 ms = ‘1’,  199999<= Duration <399999ms = ‘2’,  399999<= Duration <599999ms = ‘3’,  599999<= Duration <799999ms = ‘4’,  799999<= Duration <999999ms = ‘5’,  Duration>= 999999ms = ‘6’ |
| explicit | boolean | dichotomous | TRUE/FALSE | If the soundtrack has explicit lyrics or not |
| Explicit\_coded | number | dichotomous | 1/0 | If the soundtrack has explicit lyrics, value is 1, otherwise 0. |
|  |  |  |  |  |
| danceability | number | continuous | 0.00-1.00 | How suitable the soundtrack is for dancing |
| energy | number | continuous | 0.00-1.00 | Perceptual measurement of activity and intensity (Usually loud and fast tracks are considered energetic) |
| key | number | discrete |  | Which key the song is in (0-11 represents respectively C,C#,D,D#,E,E#,F,F#,G,G#,A,A#,B ); -1 is used when no key had been detected |
| loudness | number | continuous | (-49.531)- 4.532 | Loudness of the soundtrack in decibals |
| mode | number | continuous | 0-1  0= Minor  1= Major | Modality of the track (Major or Minor) |
| valence | number | continuous | 0.00-1.00 | How positive or negative the soundtrack is. Closer to 1 represents Happy, Euphoric music, whereas closer to 0 represents Sad, Angry or Depressed kind of music |
| tempo | number | continuous | 30.200- 243.372 | The pace of the song, in BPM (Beats Per Minute) |
| tempo\_coded | string | categorical |  | Tempo <=74 bpm = slow,  74<=Tempo <=110 bpm= moderate,  110<=Tempo<=146 bpm = fast,  Tempo >=146 bpm = super-fast |
| time signature | number | discrete | 1-5 | How many beats in each bar |
| track genre | string | nominal |  | In which genre the soundtrack belongs, the type of that song |

# Research Questions

1. Which is a more popular music genre between Pop and Rock music?
2. What are the top 10 albums of 2022?
3. What aspects should a rising music producer consider to successfully launch a hit soundtrack and or an album?
4. Are people more drawn toward positive and happy music or are they more attracted to dark and sad songs?
5. Does explicitness (songs with explicit lyrics) have any impact on likability of a soundtrack?
6. Does the duration of a song (how long or short the track is) have any impact on its popularity?

# Univariate Analysis

Graphical user interface, application, table, Excel

Description automatically generatedGraphical user interface, application, table, Excel

Description automatically generatedGraphical user interface, chart, table, Excel

Description automatically generatedGraphical user interface, application, table, Excel

Description automatically generatedGraphical user interface, application, table, Excel

Description automatically generatedGraphical user interface, chart, application

Description automatically generatedGraphical user interface, application, table, Excel

Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

# Hypotheses test

**ANOVA\_Duration**

Graphical user interface, application, table, Excel

Description automatically generated

**ANOVA\_Tempo**

Graphical user interface, application, table, Excel

Description automatically generated

**Odds Ratio**

Graphical user interface, application, table, Excel

Description automatically generated

**T-test\_Genre**

Graphical user interface, application, table, Excel

Description automatically generated

**T-test\_Valence**

Graphical user interface, application, table, Excel

Description automatically generated

# Interpolation and Extrapolation

Graphical user interface, application, table, Excel

Description automatically generatedChart, scatter chart

Description automatically generated

# Reference

* Maharshi Pandya. (2022). <i>🎹 Spotify Tracks Dataset</i> [Data set]. Kaggle. <https://doi.org/10.34740/KAGGLE/DSV/4372070>