

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | |
|  | | Analysis of Spotify song features Dataset | | |  | |
|  |  | | | | |  |
|  | | |  |  | | |
|  | | | SWARUPA VIJAYKUMAR SRN: PES1UG19CS533  **SUPREETH G KURPAD**  SRN: PES1UG19CS520  **SUVIGYA JAIN**  SRN: PES1UG19CS5  **SHRUVI D**  SRN: PES1UG19CS5  **SECTION**  H  **DATE**  15-11-2020  **COURSE**  STATISTICS FOR DATA SCIENCE  (UE19CS203) |  | | |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | |  |  | | |  | |
|  | ABSTRACT | | | | | | |  | |
|  |  | | |  |  | | |  | |
|  |  | |  | | |  | | |  | |
|  |  |  | In the recent era, the use of data science and analytics has increased tremendously in all fields. It is especially used in the entertainment industry to understand the reasons for popularity of certain songs/movies and to make customized recommendations to users by analyzing the song/movie features.  Hence, we have decided to perform exploratory analysis on the various features of songs from the Spotify song features dataset and make interesting conclusions. The dataset consists of approximately 27k songs with 15 features like tempo, danceability, valence and genres. The process includes visualizing the dataset to remove outliers, cleaning the dataset by filling missing values, standardization and normalization of the data, testing research hypotheses and finally performing correlation analysis on the columns to make conclusions. | | |  |  | |  | |
|  | | | |  | | |  | |
|  |  |  | | |  | |
|  |  | | |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | |  |  | | |
|  | | INTRODUCTION | | | | |  | | |
|  | |  |  | | |  |  | | |
|  | A song is characterized by various features like its acousticness, loudness, energy etc. It is the combination of such features that determine the genre of songs. Such features are also important in classification of similar songs and are used heavily in song recommendation systems like Spotify. They also help in studying the trends in songs and help us understand how music has changed over the years.  For the purpose of our project we aim to study and analyze song features along with their distributions. We also tested three research hypotheses based on popular opinions and data available on the internet. Further, a correlation analysis was performed on the features to find out linear correlation between variables. | | | | | | |  | |
|  |  | | |  |  | | | |  | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | DATASET | | |  | | |
|  | The dataset was originally taken from Kaggle. It was titled “ Spotify Dataset 1921-2020, 160k+ Tracks “. The dataset contains 16 columns with name of artists, song features and a list of genres. The song features are numeric data with include- acousticness, danciblity, duration, energy, liveness, instrumentalness, loudness, tempo, valence, popularity, key, mode, count and speechiness. Among them, popularity, key, mode and count are discrete variables while the remaining were continuous data. The dataset contains 27,621 songs which were analyzed. | | | | |  | |
|  |  | |  |  | | |  | |
|  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | DATASET | | | | | | | | |  | | | | |
|  | | |  |  | | | | |  | | |  | | | | |
|  | The dataset was originally taken from Kaggle. It was titled “ Spotify Dataset 1921-2020, 160k+ Tracks “. The dataset contains 16 columns with name of artists, song features and a list of genres. The song features are numeric data with include- acousticness, danciblity, duration, energy, liveness, instrumentalness, loudness, tempo, valence, popularity, key, mode, count and speechiness. Among them, popularity, key, mode and count are discrete variables while the remaining were continuous data. The dataset contains 27,621 songs which were analyzed. | | | | | | | | | | | | |  | | |
|  |  | | | |  | |  | | | | | | | |  | | |
|  |  |  | | | |  | | | |
| |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  | | DATASET | | | | |  | | | |  | |  |  | | |  |  | | | |  | The dataset was originally taken from Kaggle. It was titled “ Spotify Dataset 1921-2020, 160k+ Tracks “. The dataset contains 16 columns with name of artists, song features and a list of genres. The song features are numeric data with include- acousticness, danciblity, duration, energy, liveness, instrumentalness, loudness, tempo, valence, popularity, key, mode, count and speechiness. Among them, popularity, key, mode and count are discrete variables while the remaining were continuous data. The dataset contains 27,621 songs which were analyzed. | | | | | | |  | | |  |  | | |  |  | | | |  | | |  | | | | | | | | | | | |  | | |
|  | | |  |  | | | |  | | |  | | | | |
|  |  | | | | | | | | | | | |  | | |