Spotify Project Report:

The project requirements were obtained from a Capstone project for the Afterwork Data Science Nairobi meetup group. Though it was done individually and not for any review, under the meetup group.

**Background Information**

Spotify Technology is a Swedish music streaming and media services provider that provides an audio streaming platform, the "Spotify" platform, that offers DRM-restricted music and podcasts from record labels and media companies.

The Spotify platform provides access to over 50 million tracks. Users can browse by parameters such as artist, album, or genre, and can create, edit, and share playlists.

The service is available in most of Europe and the Americas, Australia, New Zealand, and parts of Africa and Asia, and on most modern devices, including Windows, macOS, and Linux computers, and iOS, and Android smartphones and tablets.

## Business understanding:

**Problem Statements**

1. Conduct an EDA on the dataset and come up with some data visualisations.
2. Identify popular songs by building a machine learning model that predicts track popularity. Then present the results to the senior management of Spotify. → to increase their revenue
3. Segment tracks on the platform by building a model that segments the tracks. Then present the results to the senior management of Spotify. → To identify a new genre of music.

### **Success criteria**

The data mining success criteria for the project is:

* To have a data visualization for the data to be used in the project.

While the Business success criteria for this project would be:

1. To identify the most popular song using a machine learning model.
2. To have a track segmentation and to be able to identify a new genre of music.

### **Project plan**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Plan** | **Duration** | **Resources** | **Input** | **Output** | **Dependencies** |
| **Problem definition** | 5 minutes | Project requirements | N/A | Definition of the project | Business understanding |
| **Success criteria** | 5 minutes | Project requirements | N/A | Success criteria of the project | Problem definition |
| **Project Plan** | 15 minutes | Crisp\_DM methodology | N/A | Intended plan to achieve the goals of the project | N/A |
| **Data sourcing** | 1 hour | Project requirements and internet | N/A | Raw data acquired for the project | Problem definition |
| **Data preparation and Quality** | 1 hour | Python libraries | Raw Data | Data description report and quality data | Data sourcing |
| **Data cleaning** | 1 hour | Python libraries | Prepared data | Data cleaning report and final data to for analysis | Data preparation |
| **Data Analysis** | 1 hour | Python libraries | Clean data set | Analysis report | Data cleaning |
| **Conclusion** | 15 minutes | Analyzed data | Analysis report | Conclusion to the project | Data analysis |
| **Recommendation and next step** | 15 minutes | Conclusion | Conclusions | Recommendation | Conclusion |

## Data understanding:

### **Data sourcing**

The data for the project was sources from : <https://bit.ly/SpotifySongsDs> . While the glossary data set was from, [[Link](https://drive.google.com/file/d/11Bf2kBcesrZvOkGlrK-47ubqhmaHctYk/view?usp=sharing)].

**Acknowledgements:** The data for this project comes from Spotify via the [spotifyr](https://www.rcharlie.com/spotifyr/) package by Charlie Thompson, Josiah Parry, Donal Phipps, and Tom Wolff.

### 

### **Data Preparation and Quality**

The spotify data set had 32833 rows by 23 columns.

The glossary data set contained the name of the columns, their data types and what they implied in our case study.

Notable column descriptions as shown in the glossary data set were:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| **danceability** | How suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity.  range(0.0-1.0) |
| **Energy** | Perceptual measure of intensity and activity.Typically, energetic tracks feel fast, loud, and noisy. Contributed by include dynamic range, perceived loudness, timbre, onset rate, and general entropy.  For example, death metal has high energy, while a Bach prelude scores low on the scale.  range(0.0-1.0) |
| **Key** | Estimated overall key of the track  Integers map to pitches using standard Pitch Class notation . E.g. 0 = C, 1 = Câ™¯/Dâ™­, 2 = D, and so on.  If no key was detected, the value is -1. |
| **Loudness** | Overall loudness of a track(averaged across the entire track) in decibels(dB).  It is the quality of a sound that is the primary psychological correlate of physical strength (amplitude).  Values typical range between -60 and 0 db. |
| **Mode** | Modality of a track, the type of scale from which its melodic content is derived.  major/minor i.e (1/0) |
| **Speechiness** | Detects presence of spoken words n a track  Closer to 1.0 → The more exclusively speech-like the recording (e.g. talk show, audio book, poetry); Above 0.66 describe tracks that are probably made entirely of spoken words.  0.33-0.66 → tracks may contain both music and speech, either in sections or layered, including such cases as rap music.  Below 0.33 → most likely represent music and other non-speech-like tracks. |
| **Acousticness** | range(0.0-1.0). A confidence measure of whether a track is acoustic |
| **Instrumentalness** | Predicts whether a track contains no vocals or Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live. |
| **Liveness** | Detects the presence of an audience in the recording.  Higher liveness → increased probability that the track was performed live.  A value above 0.8 provides strong likelihood that the track is live. |
| **Valance** | High → sound more positive (e.g. happy, cheerful, euphoric)  Low → sound more negative (e.g. sad, depressed, angry).  Range (0.0-1.0) |
| **Tempo** | Overall estimated tempo of a track in beats per minute(BPM).  It is the speed or pace of a given piece and derives directly from the average beat duration. |

### **Data Cleaning**

In terms of quality, I deleted columns that I won’t be needing in the analysis. These were; track\_id , track\_album\_id, 'track\_album\_name, 'playlist\_name, playlist\_id, playlist\_genre, playlist\_subgenre.

The datatype of the track\_album\_release\_date column was changed from string to date-time with a format of year-month-day. Then the year and month variables were extracted into individual columns. This was done because we needed to work with these variables separately in our analysis.

The durations of the track in the duration\_ms column were converted to minutes from milliseconds, for easy manipulation of the data.

We then checked for the number of duplicate observations in our dataset and found that there were 4,484 which summed up to 13.66% of the whole data set. These observations were dropped leaving only 86.34%.

The data set was checked for missing observation per column. They were found in only two, that is track\_name and track\_artist which had 4 missing observations each. These observations were kept as they would greatly influence our analysis.

The final cleaned data set was exported to a new csv file that could be used for the analysis

## Data Analysis

For the analysis the exported data set was loaded into a data frame.

### **Outliers:**

A subset of the continuous variables(numeric variables that have an infinite number of values between any two values) of the track characteristics was created from the data frame. These continuous variables were; danceability, energy, speechiness, acousticness, instrumentalness, liveness and valence. Something to note is that all these continuous variables range between 0.0 and 1.0, which were their respective minimum and maximum respectively.

This subset was checked for outliers using a boxplot, and the findings were as follows;

* Danceability had outliers on it’s lower(left) end hence showing it was skewed left
* Energy too had a number of outliers on it’s left end, giving the same observations
* Speechiness had numerous outliers on it’s top(right) end hence was skewed right
* Acousticness too had many outliers on its right end
* Instrumentalness had a whole lot of observations on its right end. This observation goes to show that most of the songs are comprised of a lot of instrumentals.
* Valence had no outliers, showing that none of the songs portrayed feelings that were on extreme ends.

These so called outliers were not removed from the data set because the variables affect the songs differently, for instance a song could have low energy and be speechy.

The outliers on loudness were on both ends but more of skewed to the left…..

Outliers were also checked in the track\_duration variable. The observation was that there were outliers on both ends of the boxplot but more on the right/upper end.

There were no outliers in the track popularity indicating that there were no instances of a track being extremely popular or extremely unpopular or rather not listened to.

### **Distribution:**

A histogram was also plotted for the whole dataset to show the distribution. The summary is as follows:

* The histogram graph on acousticness is a decreasing slope graph with the lower values being highly populated indicating that many songs were less acoustic.
* The danceability graph was skewed to the left as indicated in the boxplot.
* The duration graph was slightly skewed to the right, as observed in the boxplot, with most tracks being between the 3min to roughly 4.2min mark.
* The energy graph was clearly skewed to the left with most observations being roughly between the 0.7 to 0.9 mark indicating that there were more songs with relative intensity.
* The instrumentalness graph was extremely skewed to the right with most of the observations being between the 0.0 to 0.1 mark.This indicated that there were more tracks that were not performed live.
* The liveness graph was skewed to the right, with most observations being around th 0.1 to 0.2 mark. This shows that most tracks were not performed live.
* Speechiness was skewed to the right, with most observations 0.0 to 0.1 indicating that most songs were less speechy, the limit of speechiness was recorded at around 0.55
* The tempo is more skewed to the right with most values being at the central part.
* The popularity graph is skewed to the right, with most of the tracks in the 0 to 10 mark and the least in the 85 to 100 mark.
* Valence is symmetrically skewed.
* While the year graph is an increasing graph with a steep increase from around the year 2010.

### **What affects track popularity?**

A correlation matrix of the data set was done basically to see how the variables affect track popularity. The summary is as follows:

None of the variables had correlation coefficients above or even close to +/-0.5 to the track\_popularity. Therefore none are worth mentioning

This shows that none of the variables are significantly proportional to track\_popularity, that is; track popularity cannot be defined solely on a particular variable but a number or rather a combination of them.

From there on a decision was made to focus the analysis on those with a coefficient of 0.1 or close to that, these were:

* acousticness with a corelation co-efficient of 0.091759
* months with 0.080462
* energy with -0.103579
* instrumentalness with -0.124229
* duration\_min with -0.139600

The most were done virtually with scatter plot:

From the scatter plot it is clear that tracks with really low instrumentalness dominate the top most popular positions with only a few being popular with high instrumentalness.

The three months with most track releases over the years were; January, November and October in that order. The same order was observed when checking for the months with most popular track releases over the years.

Virtually, from the scatter plot above, the month of track release doesn't seem to affect its popularity. Though it would be nice to note that the month of October(10) has a continuous number of popular track releases(having a popularity of above 90) while March(3) has only one and April(4) none.

From the scatter plot above it is safe to say that the duration of the track affects it's popularity to some extent:

This is because towards high popularity index the scatter plots tend to come together around the 3 minutes mark.Though it is not a guarantee that a track at around 3 minutes will have high popularity, it is a good starting point.

It is also nice to note that the tracks highly close to the zero mark are more likely to be less popular.

The scatter plot observations were backed up after finding out that most popular tracks had a duration of 2.7 minutes followed by 4.0, 3.5, 3.3 and 3.1.

## Conclusions

* Track popularity does not rely on one variable but rather it is affected by a number of factors.
* The factors that affect track popularity are not straight forward but do contain a number of outliers.

## Recommendations and Next step

The model to predict popular tracks should incorporate the variables that showed meaningful relationship with track popularity

The next step would be to build a model that predicts track popularity.

Later another one to segment tracks to identify whether there is a new genre.