Predicting Hit Music

The Chiefs

Main Points

Business Understanding

Describing the business context, and the expected end-result of the project.

Data Limitation /Preparation

limitation, erroneous, and missing data. Attributes and The process of preparing the data

Data Understanding

Data sources, sample size, data types, stats and visualization

Proposed Models

Proposed models and the reasons of choosing the them and the expected output

Business Understanding

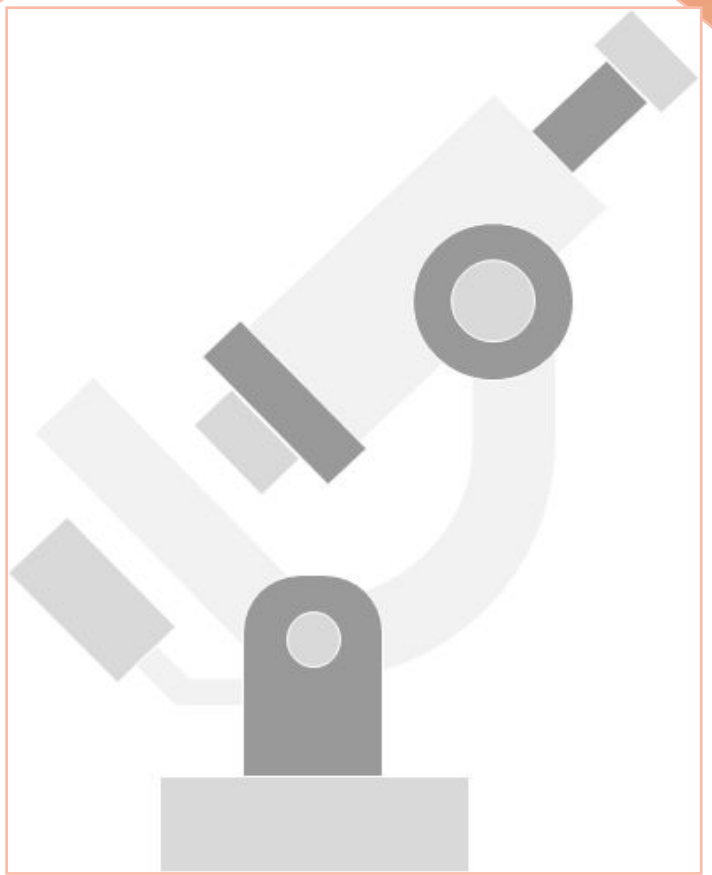
Project Objectives

Predicting a Hit Song

Finding Data Sources

Cleaning/Enriching the Data

Choosing the Model



Data Understanding

Prediction Details

What Is a hit or non-hit song .

How Popular is it at the streaming platforms?

Understand The Data. Most Important variables

The main source of data will be the Spotify API. The API provides 14 audio features


```
"audio_features": [
    {
        "danceability": 0.808,
        "energy": 0.626,
        "key": 7,
        "loudness": -12.733,
        "mode": 1,
        "speechiness": 0.168,
        "acousticness": 0.00187,
        "instrumentalness": 0.159,
        "liveness": 0.376,
        "valence": 0.369,
        "tempo": 123.99,
        "type": "audio_features",
        "id": "4JpKVNYnVcJ8tuMKjAj50A",
        "uri": "spotify:track:4JpKVNYnVcJ8tuMKjAj50A",
        "track_href": "https://api.spotify.com/v1/tracks/4JpKVNYnVcJ8tuMKjAj50A",
        "analysis_url": "http://echonest-analysis.s3.amazonaws.com/TR/WhpYUARk1kNJ_qP0AdKGcDDFKOQTT
```

To obtain the most streamed songs on Spotify from 2013-2021

Covers charts from 2013/09/29 to 2021/04/01.
Totals do not include time spent outside the weekly chart.

Pos	Artist and Title	Wks	T10	Pk (x?)	PkStreams	Total
1	Ed Sheeran - Shape of You	218	34	1 (x14)	64,217,796	2,726,385,847
2	Tones and I - Dance Monkey	89	41	1 (x17)	52,055,226	2,117,315,285
3	Post Malone - rockstar	172	27	1 (x17)	46,995,997	2,072,812,831
4	The Weeknd - Blinding Lights	70	64	1 (x13)	52,375,259	2,072,001,165
5	Post Malone - Sunflower - Spider-Man: Into the ...	128	31	1 (x2)	34,579,416	1,829,554,088
6	Lewis Capaldi - Someone You Loved	117	15	4	24,962,682	1,793,994,813
7	Shawn Mendes - Señorita	93	23	1 (x14)	67,237,638	1,759,436,341
8	Billie Eilish - bad guy	105	24	1 (x6)	50,342,324	1,717,931,688
9	The Chainsmokers - Closer	183	28	1 (x11)	46,300,740	1,699,978,914
10	James Arthur - Say You Won't Let Go	234	15	7	19,297,939	1,648,127,324

Hit or Non-Hit

HIT

A hit song, will be the upper bound outlier on our hit/non-hit variables

Your music sucks?

NON-HIT

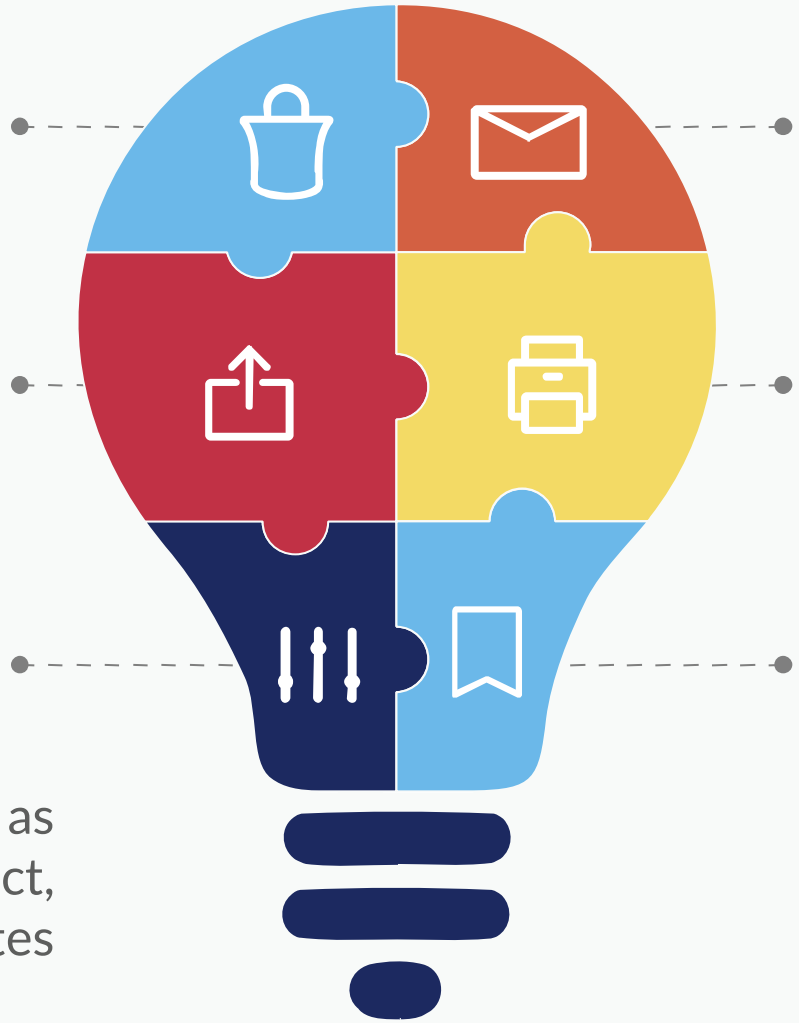
A non-hit song are the non outliers

Hit or Non-Hit

Top X Songs
A hit song, will be the top X songs streamed on Spotify

Top 200 Weekly Chart
A hit song has stayed a certain number of weeks in the TOP 200 weekly chart

Stream Count
We are taking Stream Count as the numerical variable to predict, based on our song attributes



Non-Hit Song
A non-hit song are the bottom N - top X songs streamed on Spotify

Top 10 Weekly Chart
A hit song has stayed a certain number of weeks in the TOP 10 weekly chart

Peak
The highest position a song has reached in the charts, and how many weeks has been there

Attributes

	Variable	Variable Description	Variable Type	Analysis
Audio Features	**Acoustics**	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.	Quantitative: numerical	Supervised: linear regresión/ Unsupervised PCA
	Danceability	Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.	Quantitative: numerical	Supervised: linear regresión/ Unsupervised PCA
	Durations	The duration of the track in milliseconds.	Quantitative: numerical	Supervised: linear regresión/ Unsupervised PCA
	Energy	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.	Quantitative: numerical	Supervised: linear regresión/ Unsupervised PCA
	Instrumentalness	Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentals value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.	Quantitative: numerical	Supervised: linear regresión/ Unsupervised PCA
	Liveness	Detects the presence of an audience in the recording. 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.	Quantitative: numerical	Supervised: linear regresión/ Unsupervised PCA
	loudness	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.	Quantitative: numerical	Supervised: linear regresión/ Unsupervised PCA

Attributes

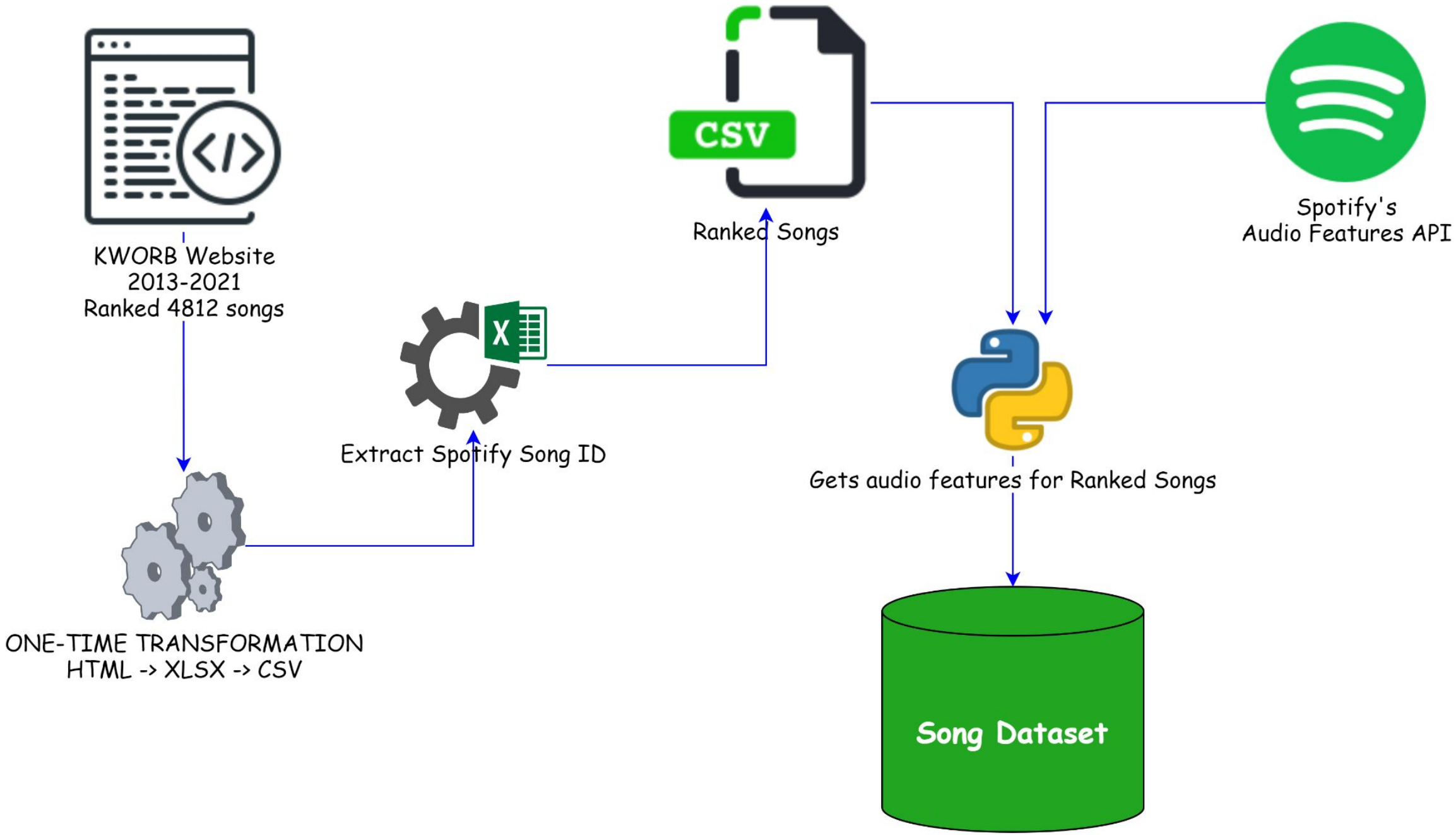
	Variable	Variable Description	Variable Type	Analysis
Audio Features	**Mode**	Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.	Categorical: binary (Major/minor)	Supervised: linear regression (dummy) classification/ Unsupervised PCA
	Speechiness	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.	Quantitative: numerical	Supervised: linear regresión/ Unsupervised PCA
	Tempo (BPM)	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.	Quantitative: numerical	Supervised: linear regresión/ Unsupervised PCA
	Time_signature	An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).	Quantitative: numerical	Supervised: linear regresión/ Unsupervised PCA
	Valence	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).	Quantitative: numerical	Supervised: linear regression/ Unsupervised PCA
	Key (C, D, E…)	The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D ♭ , 2 = D, and so on.	Categorical: nominal	supervised classification unsupervised PCA

Attributes

	Variable	Variable Description	Variable Type	Analysis
Artist Information	**Artist**	Name of the artist	Categorical: nominal	Unsupervised PCA
	Artist's genre	whether the artist is female or male	Categorical: binary (F/M)	Supervised: linear regresión(dummy)/ classification Unsupervised PCA
	Artist's age	date of birth of the artist	Quantitative: Date	
	Followers on IG	number of followers the artist has on instagram	Quantitative: numerical	Supervised: linear regresión/ Unsupervised PCA

	Variable	Variable Description	Variable Type	Analysis
Song information	**Song's name**	Name of the song	Categorical: nominal	
	Released	date the song was released	Quantitative: Date	
	Season	Season in which the song was released	Categorical: nominal	supervised. linear regression(dummy), classification unsupervised. PCA
	album	name of the album	Categorical: nominal	
	Song's genre	genre of the song	Categorical: nominal	supervised: classification
	popularity	The popularity of the track. The value will be between 0 and 100, with 100 being the most popular.	Quantitative: numerical	Supervised: linear regresión/ Unsupervised PCA
	streams	daily number of plays on spotify.	Quantitative: numerical	Supervised: linear regression
	T10	how many weeks the song has been in the top 10	quantitative: numerical	supervised: linear regression

Data Preparations


KWORB Website
2013-2021
Ranked 4812 songs
ONE-TIME TRANSFORMATION
HTML -> XLSX -> CSV
Extract Spotify Song ID
CSV
Ranked Songs
Spotify's
Audio Features API
Gets audio features for Ranked Songs
Song Dataset

KWORB

Scraping HTML into Tabular data

To obtain the most streamed songs on Spotify from 2013-2021. **4812** Songs

Pos	Artist and Title	SONG NAME	ARTIST NAME	Wks	T10	Pk	(x?)	PkStreams	Total	ARTIST URL	TRACK
1	Ed Sheeran - Shape of You	Shape of You	Ed Sheeran	218	34	1	(x14)	64,217,796	2,726,385,847	6eUKZXaKkcviH0Ku9	7qiZfU4dY1lWllzX
2	Tones and I - Dance Monkey	Dance Monkey	Tones and I	89	41	1	(x17)	52,055,226	2,117,315,285	2NjfBq1NflQcKSeiDo	1rgnBhdG2JDFTbY
3	Post Malone - rockstar	rockstar	Post Malone	172	27	1	(x17)	46,995,997	2,072,812,831	246dkjvS1zLTtiykXe5	0e7ipj03S05BNilyu
4	The Weeknd - Blinding Lights	Blinding Lights	The Weeknd	70	64	1	(x13)	52,375,259	2,072,001,165	1Xyo4u8uXC1ZmMp	0VjIjW4GlUZAMYd
5	Post Malone - Sunflower - Spider-Man: Into the Spider-Verse	Sunflower - Spider-Man: Into the Spider-Verse	Post Malone	128	31	1	(x2)	34,579,416	1,829,554,088	246dkjvS1zLTtiykXe5	3KkXRkHbMCARz0
6	Lewis Capaldi - Someone You Loved	Someone You Loved	Lewis Capaldi	117	15	4		24,962,682	1,793,994,813	4GNC7GD6oZMSxPG	7qEHsqek33rTcFN
7	Shawn Mendes - Señorita	SeÃ±orita	Shawn Mendes	93	23	1	(x14)	67,237,638	1,759,436,341	7n2wHs1TKAczGzO7	6v3KW9xbzN5yKL

Song Dataset

SpotifyChartsTableCreation (Python)

chillin & clusterin

```
5  df_top
```

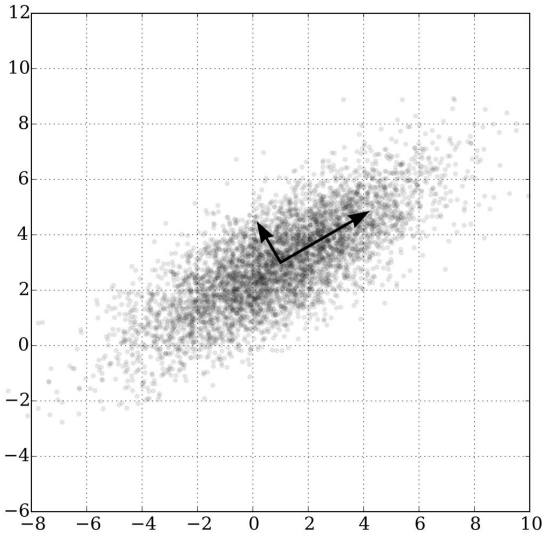
```
0    4778
1      34
Name: Top, dtype: int64
Out[30]:
```

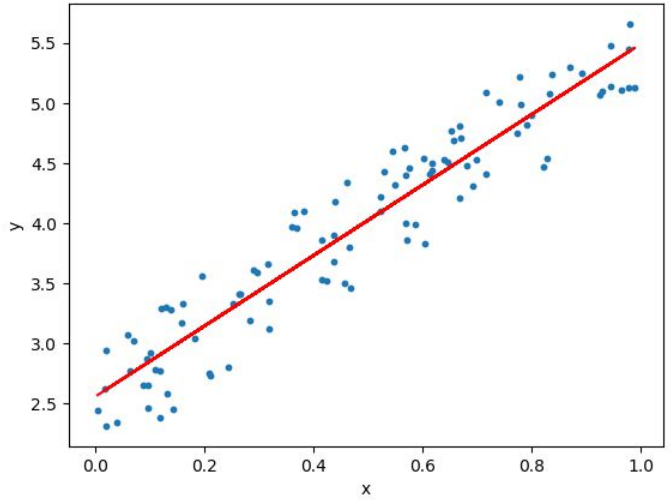
song name	album	artist	release_date	length	popularity	danceability	acousticness	energy	instrumentalness	liveness	loudness	speechiness	tempo	time_signature
Shape of You	÷ (Deluxe)	Ed Sheeran	2017-03-03	233712	87	0.825	0.58100	0.652	0.000000	0.0931	-3.183	0.0802	95.977	4
Dance Monkey	Dance Monkey	Tones And I	2019-05-10	209754	69	0.825	0.68800	0.593	0.000161	0.1700	-6.401	0.0988	98.078	4
rockstar (feat. 21 Savage)	beerbongs & bentleys	Post Malone	2018-04-27	218146	86	0.585	0.12400	0.520	0.000070	0.1310	-6.136	0.0712	159.801	4
Blinding Lights	After Hours	The Weeknd	2020-03-20	200040	97	0.514	0.00146	0.730	0.000095	0.0897	-5.934	0.0598	171.005	4
Sunflower - Spider-Man: Into the Spider-Verse	Spider-Man: Into the Spider-Verse (Soundtrack ...	Various Artists	2018-12-14	158040	85	0.760	0.55600	0.479	0.000000	0.0703	-5.574	0.0466	89.911	4

	X_0	X_1	X_2
Cool	1	0	0
Cooler	0	1	0
Coolest	0	0	1

Choosing Variables

- Build a linear regression model
- Reduce the cardinality by using PCA
- For the non-numerical variable we will use dummy variables

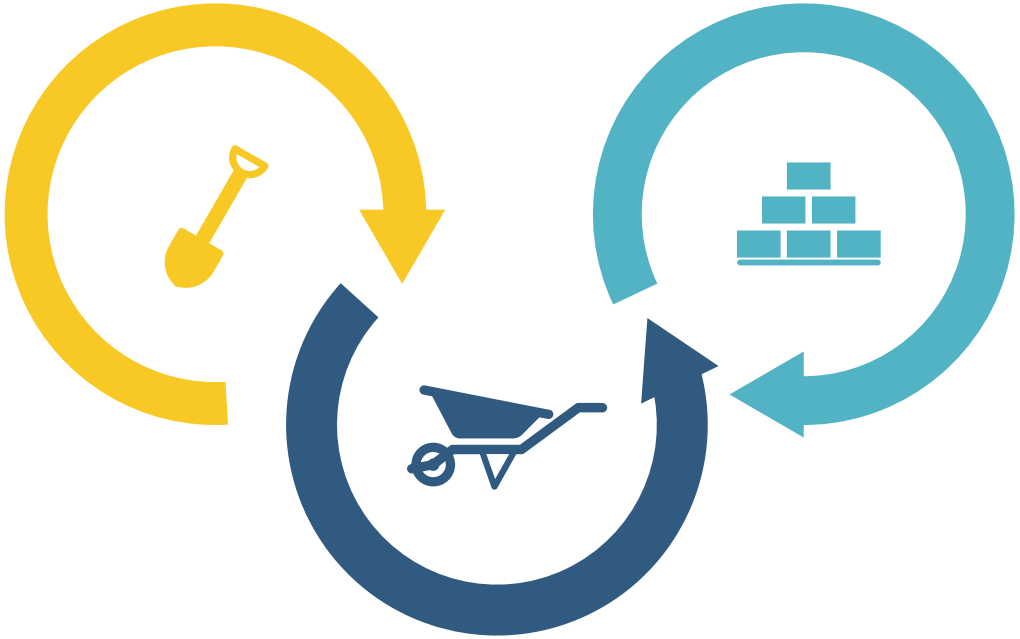




Data Exploration

Testing Spotify API

Testing Spotify API and the feasibility of creating a prototype model



Cleaning and scaling

Cleaning and scaling the data : Standard Scaler

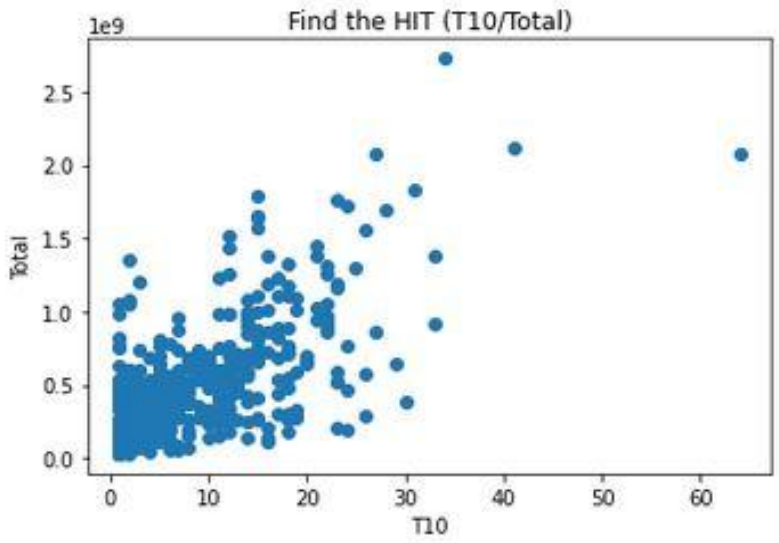
First confusion matrix

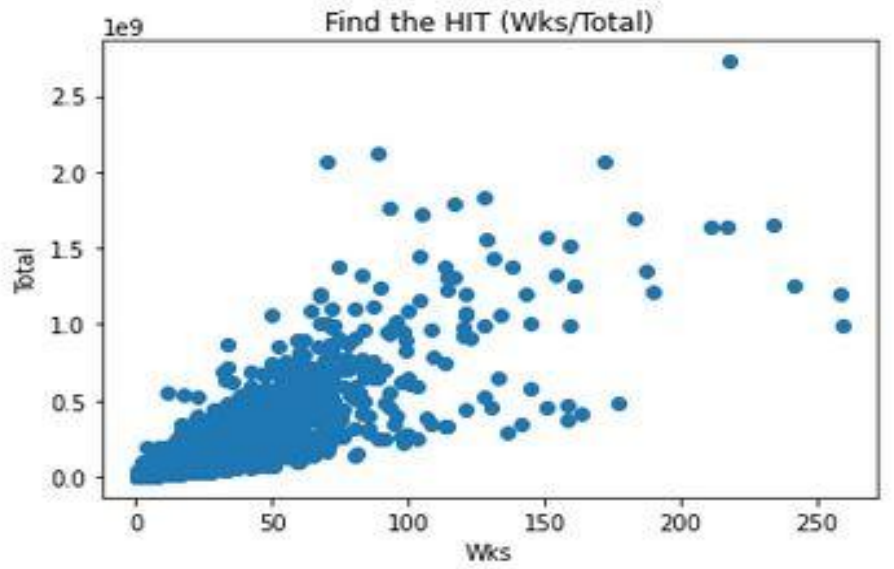
```
Cross validation scores : 0.6000046542934067
[[ 64 101]
 [  7 282]]
Precision: 0.736
Recall: 0.976
F1 Score: 0.839
```

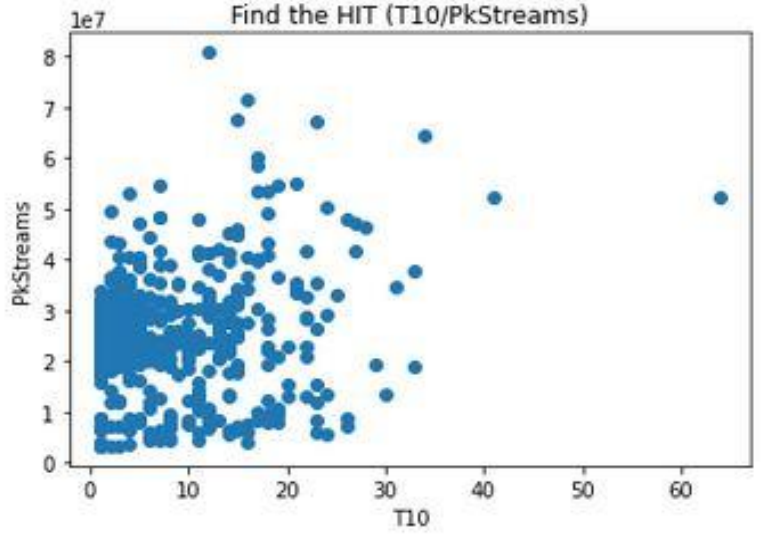
Training the model

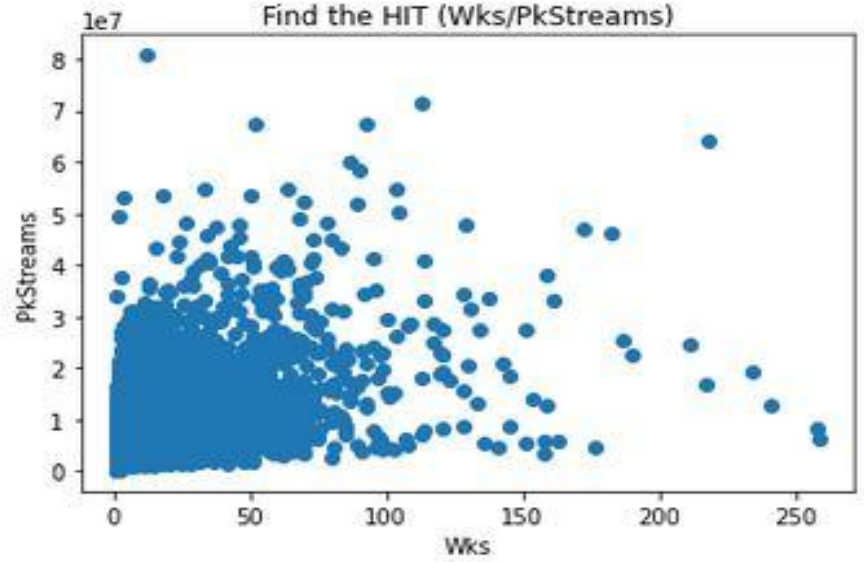
LinearSVC - > Classification model with less than 10 000 data

Correlation?



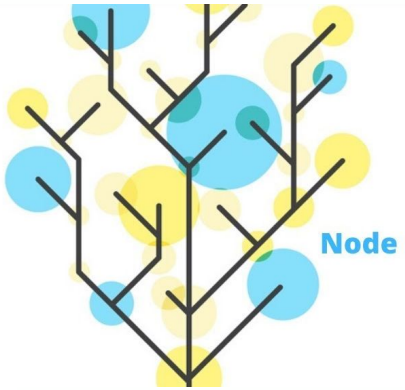






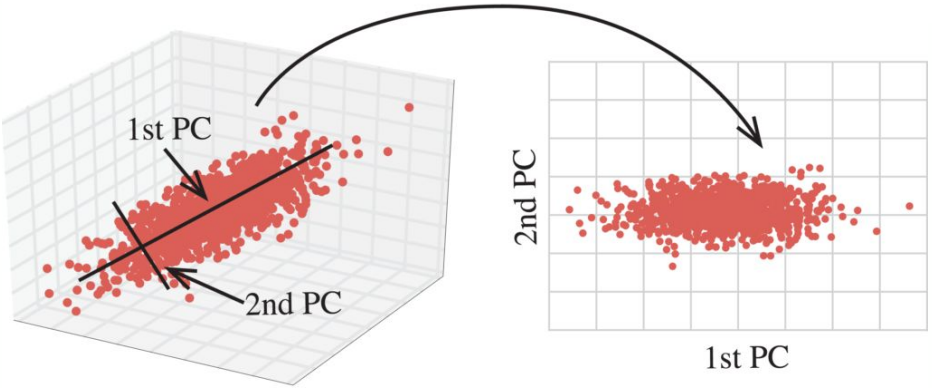
Proposed Models

Proposed Algorithms



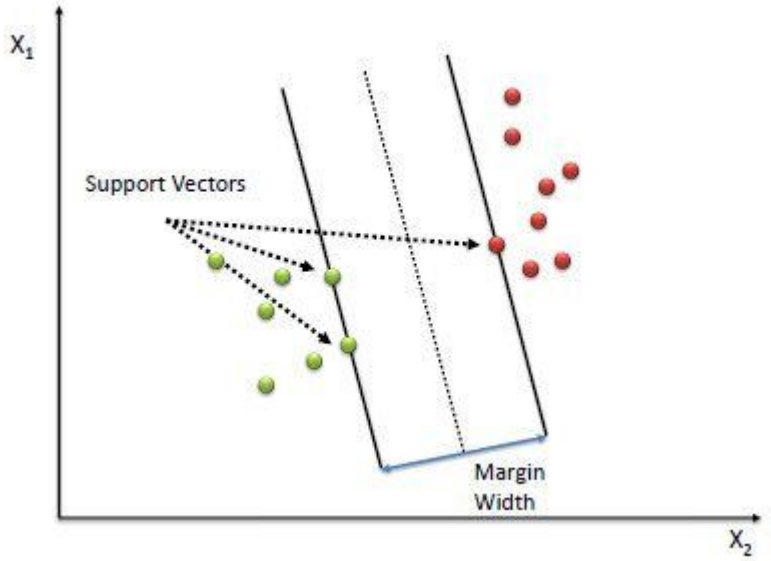
1) **Hit or Non-Hit?**

- **Classification Model**
  - **Dimensionality Reduction**
    Ex: Principal Component Analysis (PCA)



  - **Models**
    Ex: SVM - good for outlier detection and well suited for complex, small and medium sized datasets

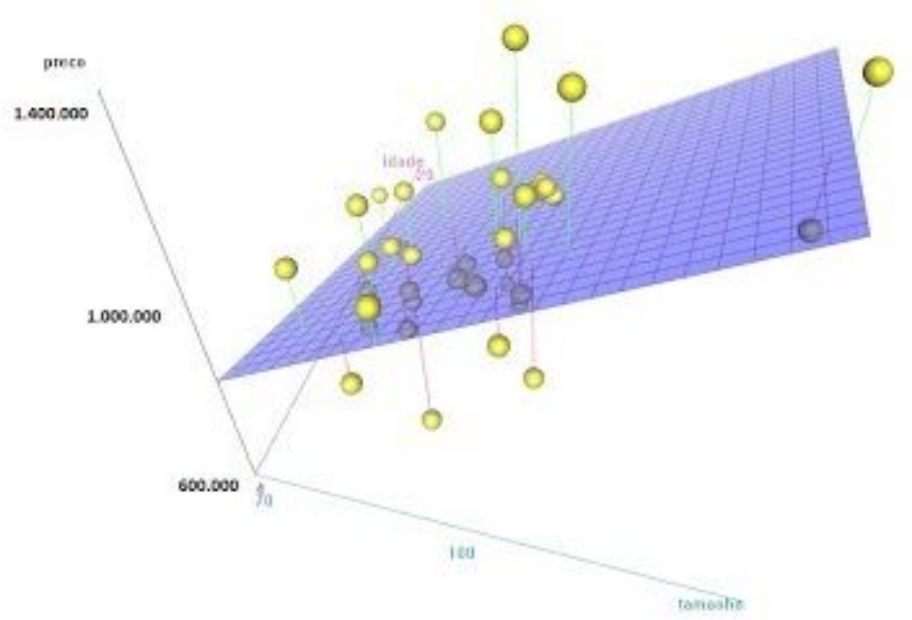
    Ex: Naive Bayes - could be applied for numerous data points and many variables to train the dataset. Faster comparing to other classification algorithm.



Algorithms we can use

2) Number of streams prediction

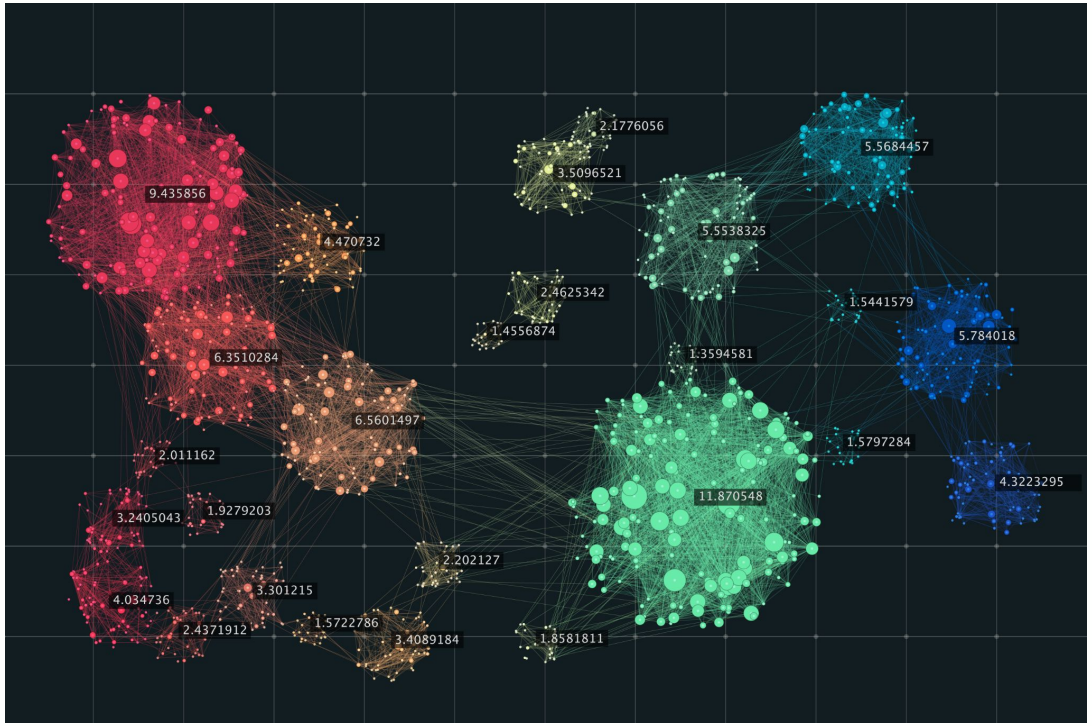
- **Regression Model**
  - **Linear Regression with Multiple Variables**



Algorithms we can use

3) Exploratory Analysis

- **Clustering Models**
  - **Models**
    Ex: K-Means: One of the simplest model and it can be applied in small datasets.




THANK
YOU
GRACIAS
ARIGATO
SHUKURIA
JUSPAXAR
BOLZÏN
MERCI
BÏYAN
SHUKRIA
TINGKI
DANKSCHEEN
YAQHANYELAY
TASHAKKUR ATU
SUKSAMA
EKHMET
GRAZIE
MEHRBANI
KOMAPSUMNIDA
MAAKE
GOZAIMASHITA
EFCHARISTO
SPASIBO
DENKAUJA
NENACHALHYA
SHACHALHUYA
CHALTU
NUHUH
SPASSIBO
TAVTAPUCH
MEDAHAGSE
BAIKA
AGUYJE
FAKAAUE
GAEJTHO
SAIKO
MERASTANHY
LAH
PALDIES
MERSI
ATTO
ANIHA
DHANYABAAD
WABEEJA
MAITEKA
HUI
YUSPAGARÀTAM
GUI
HATUR
EKOJU
SIKOMO
MAKETAJ
MIHMONCHAR