פרוייקט בקורס בינה עסקית
קורס מס' 40205
מטלה מספר 1

מוגש לידי מר אור פרץ

<u>מגישות:</u>
מיתר ירון, ת.ז.- 204263818
הדר פרץ, ת.ז.- 315970020
עדי קדם, ת.ז.- 312234818

חלק א'- איסוף נתונים

סט הנתונים שבחרנו הוא- Top 100 Most Streamed Songs on Spotify | Kaggle

חלק ב'- שאלות מחקר

Supervised:
האם שירים בסגנונות "פופ" בעלי BPM מעל 110, שאורכם פחות מ-180 שניות יותר פופולריים מאשר שירים בסגנונות "ראפ" בעלי BPM מעל 110, שאורכם יותר מ-180 שניות?

מדד KPI:
נמדוד את ממוצע עמודת הפופולריות ושיר מעל הממוצע יתויג כפופולרי.

SMART:
ספציפי- קביעת מדדים ספציפיים עם התייחסות לערכי העמודות.
מדיד- יש לנו מידע על הפופולריות, אורך השיר, BPM והסגנון.
בר השגה- באמצעות שימוש במדדים של חוסר וודאות נוכל לראות האם המאפיינים משפיעים אחד על השני.
רלוונטי- אם אכן קיים קשר בין התכונות שנבדוק לבין מידת פופולריות של השיר, עובדה זו יכולה לשמש עובדים בתעשיית המוזיקה.
תחום בזמן- לא רלוונטי.

Unsupervised:
האם לשירים פופולריים קיימים מאפיינים משותפים?

מדד KPI:
מציאת מאפיינים עם קשר חזק לשירים פופולריים.

SMART:
ספציפי- קביעת מדד ספציפי לקורלציה.
מדיד- קורלציה ניתנת למדידה באמצעות מתאם פירסון.
בר השגה- באמצעות שימוש במטריצת קורלציה ניתן לראות באופן ויזואלי את מידת הקשר בין שני פרמטרים נומריים.
רלוונטי- אם אכן קיימים מאפיינים משותפים עבור שירים פופולריים, עובדה זו יכולה לשמש עובדים בתעשיית המוזיקה.
תחום בזמן- לא רלוונטי.

חלק ג'- שאלות מחקר

מדדי פיזור:

ממוצע-

Popularity- 79.67
BPM- 116.97
Length- 214.53

סטיית תקן-

Popularity- 5.875
BPM- 27.332
Length- 35.754

רבעונים-

Popularity:
0.25- 79
0.50- 81
0.75- 83

BPM:
0.25- 95
0.50- 115
0.75- 135.25

Length:
0.25- 190.5
0.50- 210
0.75- 234.25

מספר השירים בסגנון "פופ" בסט הנתונים הוא 55, מספר השירים בסגנון "ראפ" 17 ובסגנונות אחרים 28 שירים.

תלויות וקשרים:

חישוב קורלציה מול עמודת הפופולריות-

Length- -0.009
BPM- -0.006
Genre- Pop: -0.25
Genre- Rap: 0.22

מדדים נוספים:

אנטרופיה-

Popularity- 3.843
BPM- 5.559
Length- 5.931

Gini-index-

Popularity- 0.9094
BPM- 0.9744
Length- 0.982

חישוב Information-gain מול עמודת הפופולריות-

Length- 2.183
BPM- 1.995

מסקנות:
מניתוח הנתונים הראשוני עולה כי התכונות הנומריות שבחרנו להשתמש בהן עבור השאלה המונחית שלנו- קצב השיר לדקה, אורך השיר והפופולריות שלו, אדישות זו לזו. בנוסף, קיבלנו קורלציה חיובית נמוכה בין סגנון "ראפ" לפופולריות השיר, וקורלציה שלילית נמוכה בין סגנון "פופ" לפופולריות השיר.

1. הגדרת ה- Data Warehouse

1. בחרנו בסכמת כוכב. על מנת לענות על השאלה העסקית שלנו יש צורך בטבלה מרכזית אחת המכילה בתוכה את המימדים והמדדים.

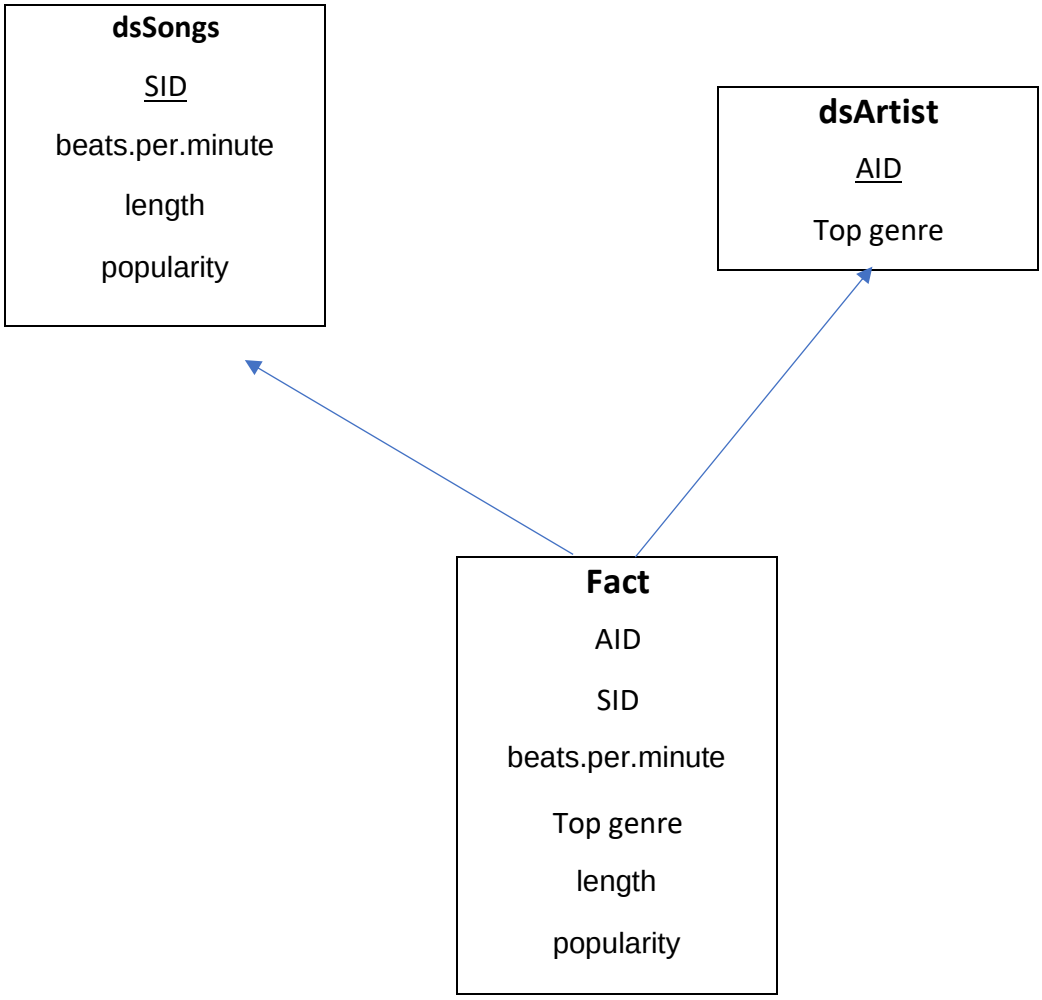
2. בסיס הנתונים שלנו הוצג כטבלה אחת אותה פיצלנו לשתי טבלאות – Artist ו –Songs.

להלן הטבלאות:

Artist
AID
Artist
Top genre
Year

Songs
SID
AID
title
beats.per.minute
energy
danceability
loudness.dB
liveness
valance
length
acousticness
speechiness
popularity

ה- Datawarehouse יהיה בנוי כסכמת כוכב, כלומר טבלה מרכזית אחת (Fact table) המורכבת משני מימדים:



3. מימוש ה- Datawarehouse :

```
CREATE TABLE dsArtist (
        AID INT NOT NULL,
        topGenere VARCHAR(50),
        PRIMARY KEY(AID)
)
```

```
CREATE TABLE dsSongs (
        SID INT NOT NULL,
        bpm int,
        length INT,
        popularity INT,
        PRIMARY KEY(SID)
)

CREATE TABLE songsFact (
        AID INT NOT NULL,
        SID INT NOT NULL,
        topGenere VARCHAR(50),
        bpm int,
        length INT,
        popularity INT,
  FOREIGN KEY(AID) REFERENCES dsArtist(AID),
  FOREIGN KEY(SID) REFERENCES dsSongs(SID)
)
```

4. חברת התקליטים רוצה להפיק שיר חדש ומעוניינת לדעת כיצד המאפיינים של bpm, אורך השיר, והסגנון שלו משפיעים על הפופולריות שלו לאחר צאת השיר. הסכמה שבחרנו יכולה לעזור לחברה בשליפה מהירה של המאפיינים הרצויים ללא ריבוי של joins.

5. מצורף קובץ Excel עם הDW שיצרנו.

2. הגדרה ומימוש ETL

1. הגדרת תהליך ה-ETL עבור אוסף הנתונים:

Extraction – בשלב זה נחלץ את הנתונים מטבלאות המקור (artist ו- songs).

Transformation – בשלב זה העברנו את הנתונים שחילצנו למודל טבלאי אחד. כעת נבצע סינון של העמודות הרלוונטיות בהתאם לשאלה העסקית.

Load – בשלב זה נטעין את הנתונים שחילצנו לDW.

2. הגדרת תהליך ה- ETL Pipeline:

שלב 1 – Reference Data: בשלב זה נגדיר את סט הנתונים המורכב משתי טבלאות, artist ו- songs, אותן נגדיר באמצעות השדות הרלוונטים.

שלב 2 – Extract from Data Reference: בשלב זה חילוץ של הנתונים ע"י קובץ CSV.

שלב 3 – Data Validation: בשלב זה נוודא כי הנתונים הקיימים מתאימים למטרת הפרויקט, כלור נוודא שקיימים המאפיינים של סגנון השיר, bpm, אורך השיר ומידת הפופולריות שלו.

שלב 4 – Transformation Data: בשלב זה נבצע אימות של שילוב המידע מתוך הטבלאות בהן השתמשנו. שלב זה כולל ניקוי של הדאטה, במקרה שלנו לא הופיעו בנתונים ערכים חסרים או חריגים ולכן לא היה צורך בכך.

מחקנו עמודות לא רלוונטיות מתוך סט הנתונים.

מטבלת artist: artist ו -year.

מטבלת songs: title, energy, danceability, loudness.dB, liveness, valance, aousticness ו- speechiness.

שלב 5 – Stage: בשלב זה הוא שלב ביניים שבו כל הנתונים נמצאים באזור ה- staging בתוכנות כמו Tableau או Airflow, דרכן הנתונים עוברים.

שלב 6 – Publish to Data Warehouse – בשלב זה העברנו את הנתונים הרלוונטיים לאחר עיבודם למחסן נתונים שמורכב מטבלה אחת בה יהיו המפתחות של השירים והאמנים.

3. מימוש תהליך ה- ETL Pipeline:

Version	Source Table	Source Column	Source Type	Transformation	Target Table	Target Column	Target Type	Target Length	Default Value	Note
1	Artist	AID	int	NOT NULL	dsArtist	AID	int			
1	Artist	AID	int	NOT NULL	songsFact	AID	int			
1	Artist	topGenere	VARCHAR		dsArtist	topGenere	VARCHAR	50	NoGenere	
1	Artist	topGenere	VARCHAR		songsFact	topGenere	VARCHAR	50	NoGenere	
1	Songs	SID	int	NOT NULL	dsSongs	SID	int			
1	Songs	SID	int	NOT NULL	songsFact	SID	int			
1	Songs	beats_per_minute	int		dsSongs	bpm	int		-1	
1	Songs	beats_per_minute	int		songsFact	bpm	int		-1	
1	Songs	length	int		dsSongs	length	int		-1	
1	Songs	length	int		songsFact	length	int		-1	
1	Songs	popularity	int		dsSongs	popularity	int		-1	
1	Songs	popularity	int		songsFact	popularity	int		-1	