Final Project HackerU

Mohammad Shkir

Anwar Abu Alheja

06 March 2024

Data Load

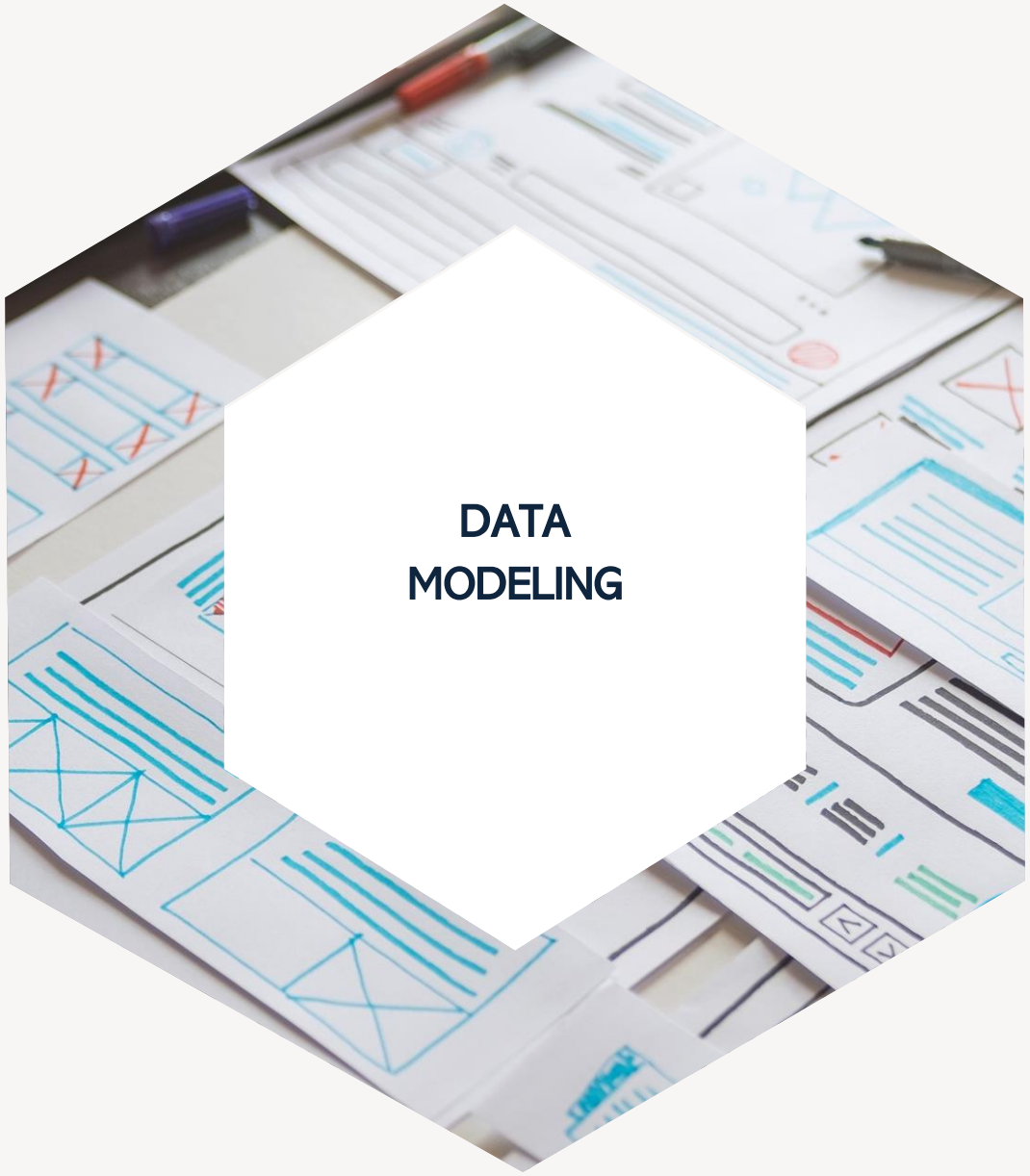
1. **טעינת הנתונים דרך ממשק DBeaver**

- **יצירת Database חדש בשם Chinook.**
- **Tools > execute script > בחירת קובץ chinook dump.**
- **Start.**

2. מסד הנתונים מוכן לשימוש.

Python {Pandas}

1. אנו מייבאים את הספריות הדרושות (Pandas, Sqlalchemy).
2. יוצרים חיבור חדש עם מסד הנתונים .
3. קוראים את שלושת הקבצים השונים.
   - הקובץ הראשון (raw-department-budget) קוראים כ- json ומוסיפים lines=True.
   - הקובץ השני (raw-department-budget2) קוראים כ- json.
   - הקובץ השלישי (raw-department) קוראים כ- csv ומוסיפים sep = '-'.
4. משרשרים את שתי הטבלאות של תת מחלקות .
5. מחשבים תקציב דרך Group by.
6. מכניסים תקציבי המחלקות לטבלה חדשה שמכילה גם department_id ו department_name.
7. משכפלים את ה- Data frame לטבלת SQL חדשה במסד הנתונים.



DBT

Dim_playlist

שדות להביא		האם יש משהו שדורש התייחסות עקב כך שמדובר בחיבור של שתי טבלאות שונות עם שני תאריכי עדכון שונים ?
• כל השדות מטבלת playlisttrack. • כל השדות מטבלת playlist.		כן, כאשר מדובר בחיבור של שתי טבלאות עם תאריכי עדכון שונים, חשוב להתייחס למדיניות עדכון של כל טבלה בנפרד, יתכן גם שיהיה צורך ביישום פתרונות טכניים כמו Materialization על מנת לשמור על יעילות ומהירות בגישה לנתונים.

Dim_customer

שדות להביא	תיקון את השמות	עמודת domain
• כל השדות מטבלת customer.	• בשם פרטי ושם משפחה האות הראשונה גדולה • השתמשנו בפונקציית INITCAP.	• עמודה שמכילה הדומיין מכתובת האימייל • השתמשנו בפונקציית substring substring(email, position ('@' in email) +1,length (email)) as domain

Dim_employee

עמודת is_manager	עמודת domain	עמודת exp_years	שדות להביא
• עמודה שמציבעה האם העובד הוא מנהל (1) או לא (0). • CASE WHEN e.employeeid IN (SELECT DISTINCT reportsto FROM stg.employee e WHERE reportsto IS NOT NULL) THEN 1 ELSE 0 END AS is_manager	• עמודה שמכילה הדומיין מכתובת האימייל • השתמשנו בפונקציית substring substring(email, position ('@' in email) +1,length (email)) as domain	• עמודה שמחשבת מספר השנים בהם העובד כבר מעוסק. • השתמשנו בפונקצית DATE_PART ו AGE. • DATE_PART('year', AGE(current_date , e.hiredate)) AS exp_years,	• כל השדות מטבלת customer. • שם המחלקה ותקציבה מטבלת department_budget.

Dim_track

שדות להביא	עמודת track_length_ss	עמודת track_length_mm
• כל השדות מטבלת track. • כל השדות מהטבלת genere, mediatype, artist, album. • לא צריך להביא את המפתחות הראשיות בטבלאות אחרי ה- join, כי הן שוות למפתחות הזרות בטבלת track.	• אורך השיר בשניות. • מחלקים עמודת milliseconds בטבלת track ב- 1000. • milliseconds / 1000 AS track_length_ss	• אורך השיר בדקות ושניות בפורמת mm:ss. • לחישוב הדקות מחלקים milliseconds ב- 1000ו ב- 60, לחישוב השניות מחלקים milliseconds ב- 1000ושאר החילוק ב- 60 ומשתמשים בפונקציית LPAD כדי לקבל שני אותיות. LPAD(FLOOR(t.milliseconds / 1000 / 60)::TEXT, 2, '0') || ':' || LPAD(FLOOR(t.milliseconds / 1000 % 60)::TEXT, 2, '0') AS track_length_mm,

Fact_invoice

Materialization		שדות להביא
• כדי להפוך את הטבלה ל incremental אנו מוסיפים {{ config(materialized='incremental') }} בראש הקובץ או בקובץ dbt_project.yml תחת models.		• כל השדות מטבלת invoice. • לא צריך להביא את הכתובת מטבלת invoice, כי יש לנו את הערכים בטבלת dim_customer ושני הטבלאות מקושרות.

Fact_invoiceline

Materialization		שדות להביא
• כדי להפוך את הטבלה ל incremental אנו מוסיפים {{ config(materialized='incremental') }} בראש הקובץ או בקובץ dbt_project.yml תחת models.		• כל השדות מטבלת invoiceline.

API Currencies

Import Libraries

1. אנו מייבאים את הספריות הדרושות (Pandas, requests, json, numpy, sqlalchemy) .
2. יוצרים חיבור חדש עם מסד הנתונים באמצעות sqlalchemy.

get_exchange_rates

1. בחרנו exchangerate API המאפשר 1,500 בקשות בחודש וגם גישה לנתונים היסטוריים
2. יוצרים פונקציית פיתון חדשה בשם get_echange_rates שלוקחת את התאריך כקלט ומחזירה שיעור השקל בתאריך שהועבר כפלט.

ILS historical rates

1. משכפלים את הטבלאות מ- SQL– ל Data frames .
2. הוצאת כל התאריכים מהטבלאות.
3. שמירת התאריכים הייחודיים ל Data frame חדש.
4. הפעל את הפונקציה על ה- Data frame החדש על מנת לקבל תעריפי שקלים בכל התאריכים

API Currencies

continue..

Joins

1. כדי לחבר הטבלאות עלינו להמיר את כל התאריכים לפורמט yyyy-mm-dd .
2. אחרי שיצרנו טבלה עם כל התאריכים והתעריפים השקלים, עכשיו אנחנו מצטרפים שערי השקל עם שלוש הטבלאות בסכימה המכילות סכומים בדולר.
3. יוצרים עמודה חדשה בשם ils_total ששווה ל- Amount * ils_rate.
4. הגדרת הנקודה העשרונית בערכים עשרוניים לשניים.

PostgreSQL Tables

1. הסרת עמודת התאריך (עמודה משוכפלת)
2. המרת ה- Data frames לטבלאות SQL בסכמה stg.

Visualization

Playlist Dashboard

Playlist Dashboard

Music Videos

Playlist Name With The Least Tracks

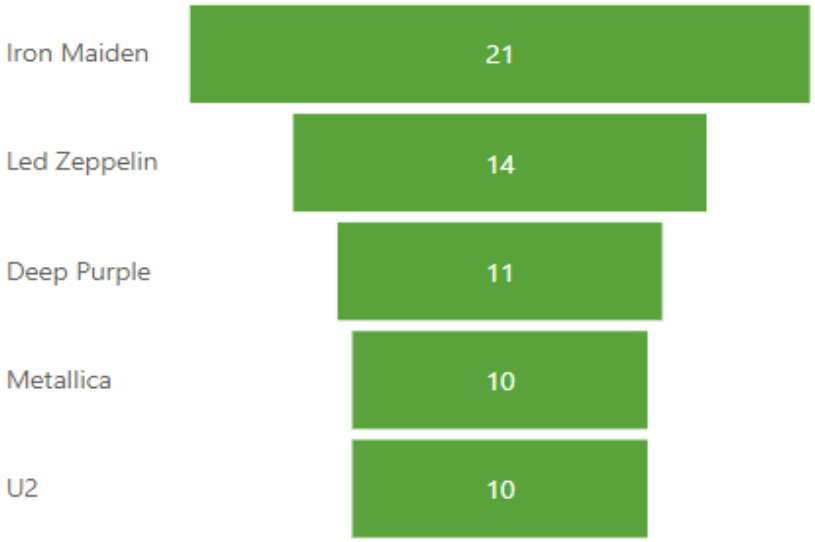
Music

Playlist Name With The Most Tracks

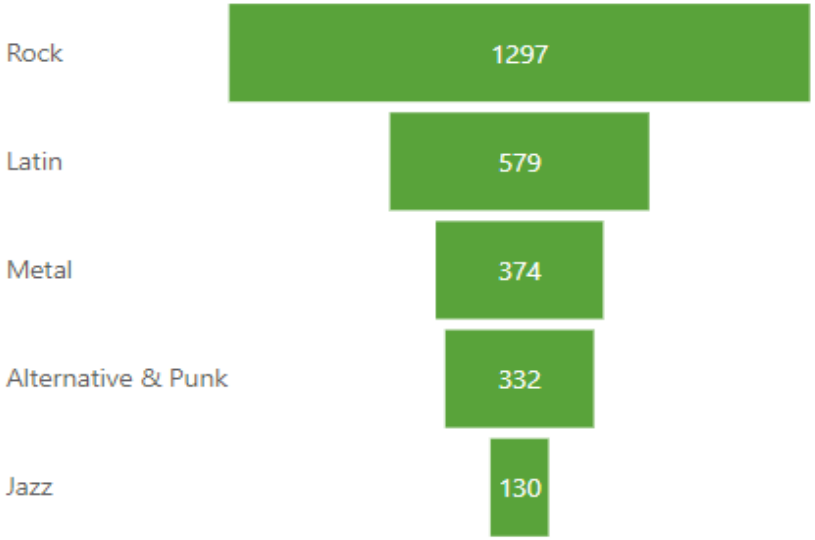
622.50

AverageTracks In An Album

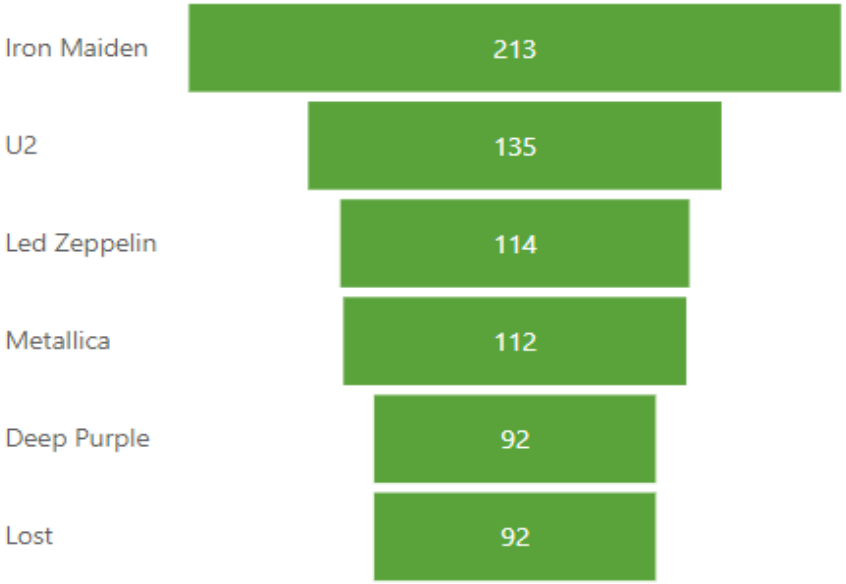
Top 5 artists by albums number



Top 5 genres by tracks number



Top 6 artists by tracks number



genre : All

Alternative	Blues	Classical
Alternative & Punk	Bossa Nova	Comedy

Clear all slicers

media type: All

AAC audio file	Protected AAC audio file	Purchased AAC audio file
MPEG audio file	Protected MPEG-4 video file	

Sales Dashboard

Top 5 Customers By Rental Amount

	total_usd	total_ils
Puja Srivastava	71.31	246.30
Denise Kelly	65.36	226.85
Josephine Gomez	55.44	185.94
Helena Holý	49.62	172.36
Richard Cunningham	47.62	158.20

Top 5 Countries By Revenue

	total_usd	total_ils
India	1,265.60	4,343.95
China	906.18	3,111.69
Brazil	803.15	2,773.87
United States	803.13	2,755.40
Russian Federation	634.83	2,188.62

Bottom 5 Countries By Revenue

	total_usd	total_ils
Gambia	1.98	6.77
Bahrain	3.96	12.75
Iraq	3.96	12.75
Lithuania	3.97	13.58
Faroe Islands	5.94	20.08

Country	Genre	Sales Percentage
Brazil	Rock	40.43%
	Latin	17.38%
	Metal	9.37%
	Alternative & Punk	8.63%
	Blues	3.70%
	TV Shows	3.47%
	Jazz	3.08%
	Reggae	1.97%
	Soundtrack	1.85%
	Drama	1.49%
	R&B/Soul	1.36%
	World	1.23%
	Heavy Metal	1.11%
	Science Fiction	0.99%
	Classical	0.74%
	Pop	0.62%

2018

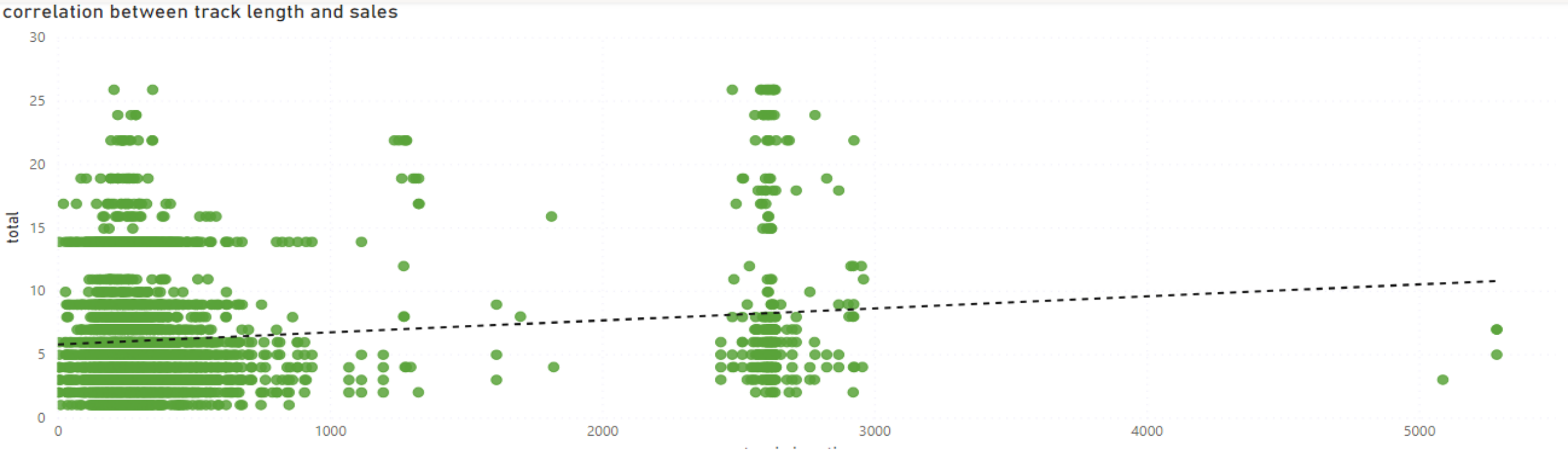
2019

2020

2021

2022

Appendix



there is a positive correlation between ***track length*** and the ***sales*** it made, but it is very very small