



DATABASE OVERVIEW

- 1. טעינת מסד הנתונים דרך CMD
 - 2. עבודה מול Phyton
 - 3. איחוד טבלאות
 - 4. ניקיון ותיקון מסד הנתונים
- Postgres טעינת מסד נתונים לשרת 5.
 - API Service .6
 - 7. עבודה מול DBT
- 8. עיצוב והצגת הנתונים בתוכנת Microsoft PowerBI

PHYTON - PART 1

שלב 3 – איחוד טבלאות

- איחוד טבלאות לטבלה אחת "department_table".
- הטבלה כוללת מזהה מחלקה, שם המחלקה ותקציב לכל מחלקה
 - יצירת עמודה חדשה בשם "total_budget". •

שלב 4 - טעינה

טעינת הטבלה "department_table" מסד הנתונים "Chinook_Database". בתוכנת "dbeaver".

שלב 1 - קריאה

- קריאת הדאטה בעזרת קובץ dump באמצעותCMD.
- פתיחת database חדש בשם "Chinook_Database".

שלב 2 - חילוץ

- יבוא הספריות הרלוונטיות ב-"phyton" בתוכנת "pycharm".
 - "json" קריאת שלושה קבצים מסוג

PHYTON - PART 2 API SERVICE

טעינה

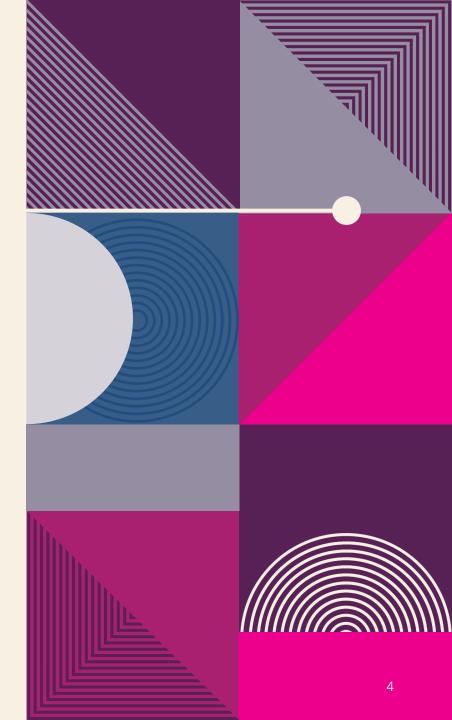
בחירת אתר

לאחר התהליך הנ"ל, מתבצעת המרה ל-"dataframe" והעלאת התוצאה אל שרת "postgres".

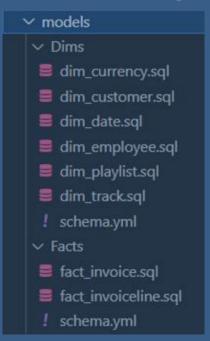
https://openexchangerates.org/

הסבר תהליך

- התבקשנו להוציא המרות בין דולר לשקל לאורך השנים בהם המסד נתונים קיים.
- באמצעות יבוא ספריות "pandas, numpy, requests" הצלחנו לשלוח בקשה לאתר שנבחר, קיבלנו את ההמרה הרלוונטית.
- לאחר טעינה ראשונה של בקשות "API" נסינו לקבל מידע על טווח התאריכים 2018-2020. בנקודה זו זמן הטעינה היה ארוך מדי והחלטנו לעשות שינויים בקוד.
 - הצלחנו להתגבר על זמן הטעינה באמצעות ספריית "concurrent" אשר מאפשרת לחלק את בקשות . "threads" שונים שפועלים במקביל. הורדנו את זמן ההמתנה לדקות בודדות.
 - קיים גם סינון תוצאות שמחזירות "none".



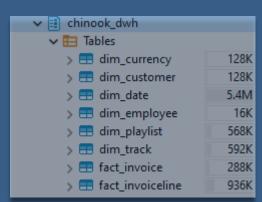
SOLUTION - DBT

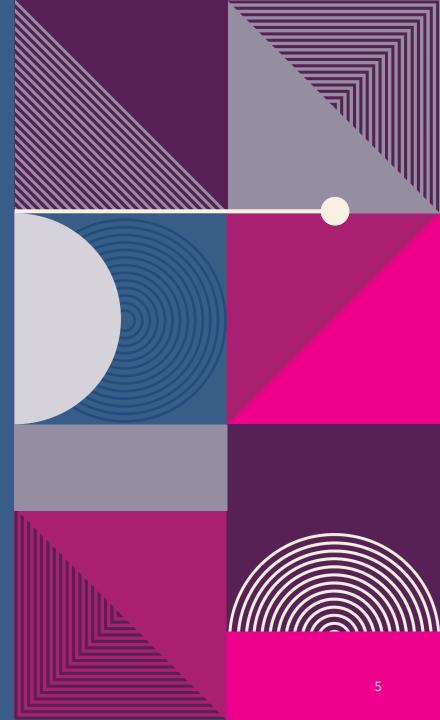


- name: Chinook_Dims
database: Chinook_Database
schema: stg
tables:

- name: Department_Table
- name: album
- name: artist
- name: customer
- name: employee
- name: genre
- name: mediatype
- name: playlist
- name: playlist
- name: track
- name: Exchange Table

- C: פתיחת תייקיה חדשה לפרויקט בכונן 🗸
- ע "git bash"). עריקויה הפרויקט ("git init"). ✓
 - ."schema.yml" + "facts" + "dims" יצירת ✓
- ."dims"- אשר מבילה את כל פירוט ה "schema.yml". ✓
- ."facts"-אשר מבילה את כל פירוט ה "schema.yml". ✓
- ."stg"- אשר מטרתו לכוון אילו טבלאות אנו נמשוך מהסכמת ה-"stg". ✓
 - ב-"source" קיים בנוסף פירוט לכל טבלה, סכמת המודלים, סכמת הדאטה. ✓
 - עריבת קוד ב-"dbeaver" בשפת "lsql" והעתקה שלו אל ה-"dbt". ✓
 - ."postgres" בשרת "Chinook_DWH" בשרת " בשרת " העלאת מסד הנתונים אל סכמת
 - בהצלחה. שי'dbt run" בהצלחה. ✓







POWER BI MODEL





ביססנו את הפרויקט שלנו על מגמות שוק לפי קטגוריות שירים ומדינות.



ABSTRACT

תובנות על התנהגות השוק והתנהגות צרכנים ברכישות שירים בחתך של זמנים שונים



DESIGN

מינימלי, קריא ונוח לשימוש.

MARKET OVERVIEW

ניתן לחלק את כל ה



השנה הרווחית ביותר

• שנת 2019 ₪ 10.3K *>*



סך ההכנסות 2018-2022

₪ 47.60K •



המדינה הרווחית ביותר

מדינת הודו

₪ 4.34K >



2023

ממוצע רכישה לשיר

回 12.49 •



הקטגוריה עם מירב השירים

'קטגוריה 'רוק' •

40.60% >



כמות שירים שנמכרו

13.46K •

DATA TRANSFORM



- הסרת עמודות לא רלוונטיות בטבלת לקוחות לדוגמא: 'fax'.
- . 'company' בגון עמודת 'null' מחיקת עמודות עם הרבה תאי
 - יצירת עמודות תאריך מחושבות: •
 - סופ"ש, יום חול, עונות שנה, רבעונים 🗲
 - יצירת עמודות לוגיקה מחושבות:
 - עמודת 'full name', עמודת 🦫
 - מחיקת עמודת תאריך עברי
 - הסרת כפילויות
 - סינון טבלת פלייליסט מעבר לשקופית



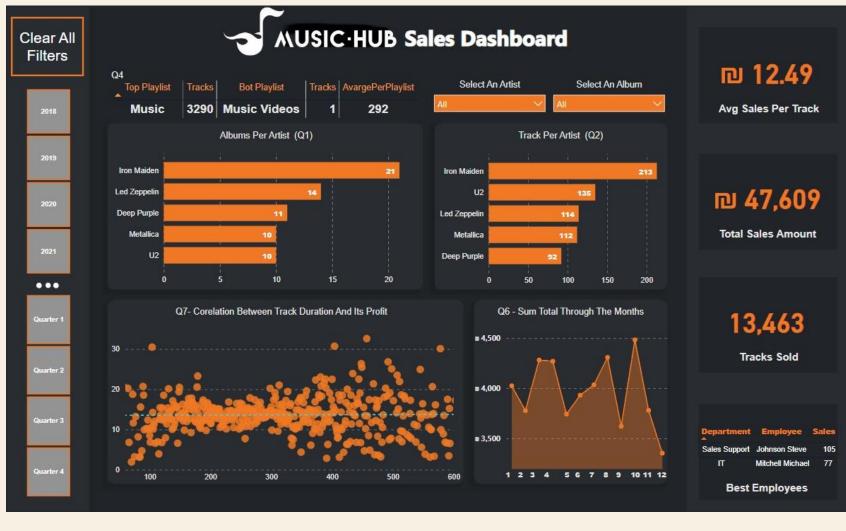
- "Postgres" ייבאנו את הדאטה באמצעות התחברות לשרת
 - יבוא כל טבלאות ה-"dims" וה-
- בניית ERD, ניגשנו לחלון "model view" ובוצעו חיבורים בין טבלאות, ובנוסף חיבור טבלת
 - קביעת מפתחות עקיריים בכל אחת מהטבלאות
- טיפול בסוגי עמודות סיווג הדאטה לסוגיהם השונים: גאוגרפיה, תאריך, מתמטיות ושמיות לעמודות הרלוונטיות.
 - יצירת היררכית מיקום ותאריך:
 - מדינה > מחוז > עיר > כתובת
 - שנה> רבעון > חודש > יום 🗲
- יצירת טבלה "all_measures" בה הכנסנו את כל המאג'רים שהשתמשנו בהם במהלך הפרויקט , בנוסף הם ממיוינים לפי שאלות.
 - יצירת כפתורים ומסננים מתקדמים המאפשרים חלוקה של הדאטה לרזולוציות שונות.
 - שילוב רקע מתוכנה צד שלישית ושימושם לעיצוב הדאשבורדים.
 - שימוש בכלי 'Selection' על מנת לצור את הדשבורד בצורה מדויקת.



DASHBOARD'S OVERVIEW

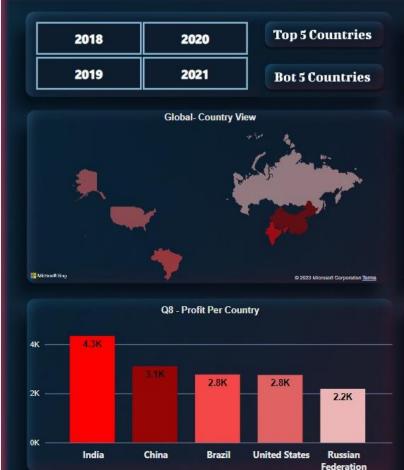
□ SALES□ WORLDWIDE□ GENRE & TRACKS□ QUESTIONS

SALES DASHBORD





WORLDWIDE DASHBORD



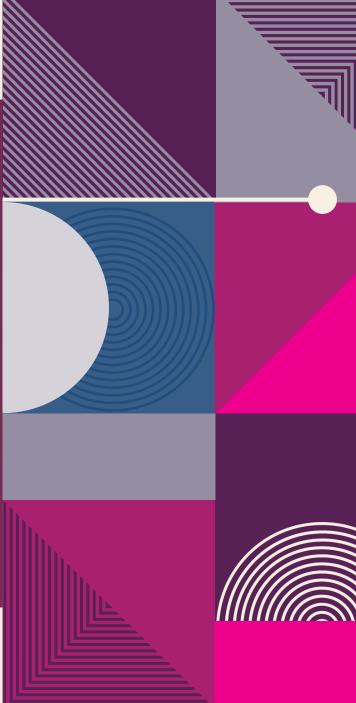


Q5 - Top 10 Customers Buyers & From Which Country

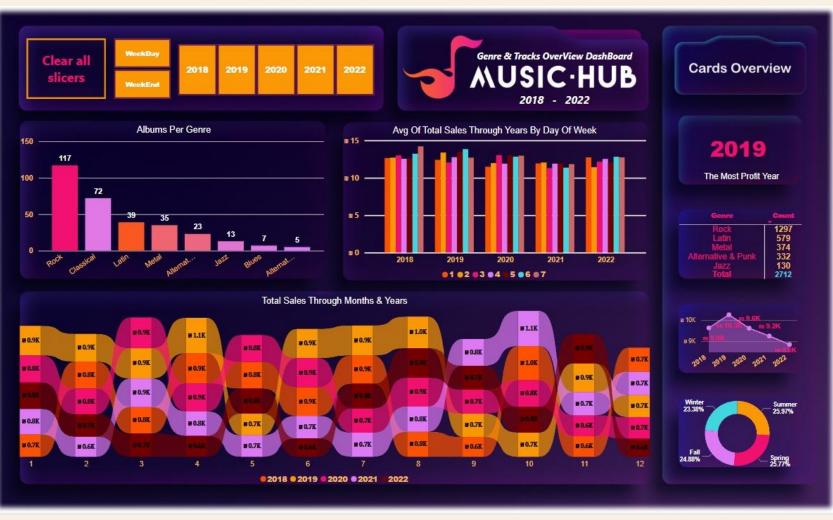
full_name	USD	ILS	Buyer's Country
Puja Srivastava	71	₪ 245.84	India
Denise Kelly	65	₪ 226.71	Brazil
Josephine Gomez	55	₪ 186.05	India
Helena Holý	50	m 171.95	Czech Republic
Richard Cunningham	48	nu 157.99	USA
Kathryn Coleman	48	回 165.05	Ukraine
Annie Russell	48	ங 165.95	Mozambique
Luis Rojas	47	回 165.20	Chile
Hugh O'reilly	46	₪ 157.41	Ireland
Ladislav Kovács	46	m 159.39	Hungary

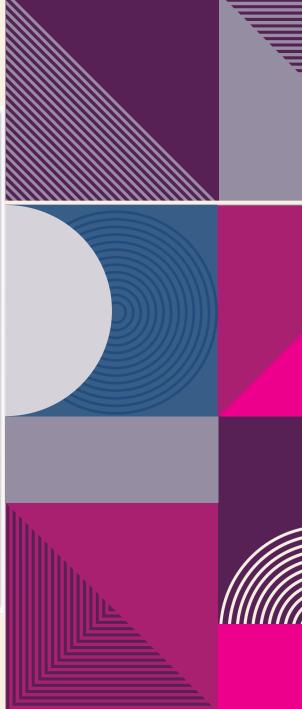
Q9 Bonus - Profit I	Per Genre IN Th	e Selected Country
genre_name	% Of Total	invoiceline_total
Rock	40.58%	18,295.21
Latin	18.46%	8,322.64
Metal	12.01%	5,415.52
Alternative & Punk	11.42%	5,149.59
Jazz	4.37%	1,969.25
Blues	2.85%	1,286.48
R&B/Soul	1.63%	734.89
Reggae	1.56%	703.71
Soundtrack	1.42%	641.07
Heavy Metal	0.87%	390.94
World Total	0.82% 100.00%	370.27 45.078.89

Cards Overview Afghanistan Current Country 12.49 Avg Sales 13.46K **Quantity Sold** 回 47.60K **Total Sales**



GENRE & TRACKS DASHBORD





13



TOP 5 WITH MOST ALBUM & TRACKS PER GENRE AND ARTIST

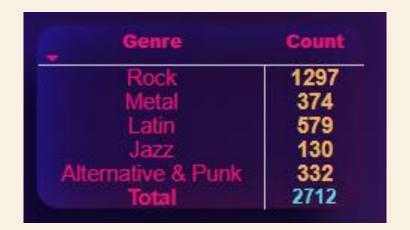


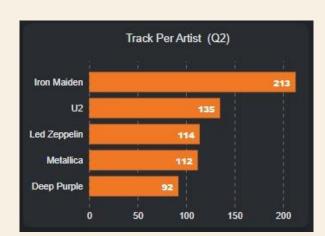
מסקנות עבור אמן ואלבום:

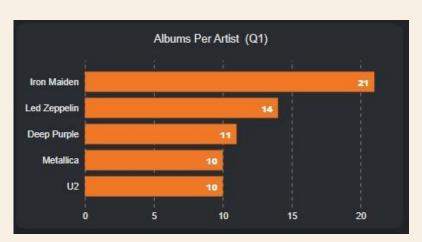
- ניתן להסיק מי האמן שיצר הכי הרבה אלבומים, בשנים -2018 2021 הוא סטיב הריס.
 - ניתן לראות מי חמשת האמנים עם הכי הרבה אלבומים.
 - תדירות הוצאת כמות הגבוהה ביותר של שירים לפי אמן.

מסקנות עבור שירים לפי קטגוריה:

- בטבלה הנ"ל ניתן לראות חמשת הקטגוריות עם הכי הרבה שירים.
 - סך כל השירים שנכללים בחמשת המובילים הוא 2,712 שירים,
 בעוד שמספר השירים הכולל 3,257 כלומר, כ-83% מהשירים
 נמצאים בחמש הקטגוריות רוק, מטאל, לטין, ג'אז, אלטרנטיב
 ופנק.
 - הקטגוריה שיש ממנה הכי הרבה שירים היא "rock" והוא מונה כ-40% מכלל הנתונים.







TOP 5 BY PLAYLIST MOST \ LESS \ AVERAGE



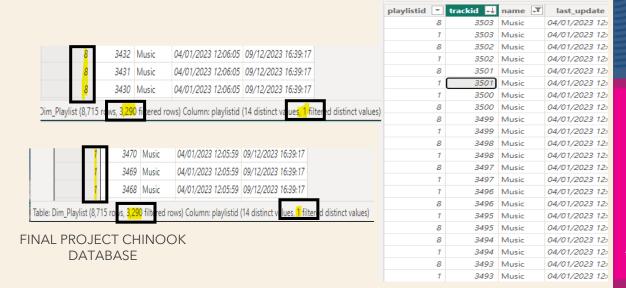
הסבר:

מסקנות:

- הסוג פלייליסט שיש בו הכי הרבה שירים הוא "music" באשר יש בו 3,290 שירים, ניתן לראות שיש פער מאד גדול בין פלייליסט זה לממוצע פר פלייליסט 292. עוד ניתן להסיק כי מתוך 5,425 שירים 60% הם פלייליסט
 "music"
 - ניתן לראות את הפלייליסט עם הכי מעט שירים הוא "music videos".• תחום שהוא פחות נכלל ב-iTunes.
- * Bot Playlist פלייליסט עם כמות השירים הנמוכה ביותר (bot = bottom).

- כאשר ביצענו את הבדיקה של מספר השירים עבור כל פלייליסט ראינו
 כי אנחנו מקבלים מספר גבוה מידי של שירים, מה שהוביל אותנו
 לבדוק את הדאטה ולהבין איזה נתון גורם לניפוח הנתונים.
- בתמונה המצורפת ניתן לראות כי גילנו שקיימים שני ID (1,8) בטבלה שהם מוכפלים, מבדיקה נוספת של מספר השורות המכילות 8 והמכילות 1 ראינו את אותו מספר ערכים מה שהביא אותנו למסקנה שזה הכפלה של הנתונים והחלטנו להוציא החוצה את ID = 8.
- * מסומן בצהוב עם מסגרת שחורה על גבי התמונה להדגשה ולהמחשה.

Top Playlist	Tracks	Bot Playlist	Tracks	AvargePerPlaylist
Music	3290	Music Videos	1	292



SONGS

TOP 5 CUSTOMER BUYER



מסקנות:

- ניתן לראות את סכומי ההזמנות של 10 הלקוחות הכי מכניסים, שמם, המחיר בדולר ושקל ומדינה ממנה מגיעים.
- ניתן לראות את הלקוח שרכש בשקלים הכי הרבה (245 ₪) בהסתכלות על כל התקופה וניתן לראות את שמו של הלקוח.
- 2 מתוך 3 הלקוחות ששילמו הכי הרבה הם מהודו (הודו המדינה שמכניסה הכי הרבה מכירות)
 - ניתן לראות מהטבלה כי סך התשלום של עשרת הלקוחות המכניסים הם פי 2 ומעלה מעל הממוצע הכנסות שהוא כ-72 ₪

הסבר:

- בשאלה התבקשנו להציג את חמשת הלקוחות המכניסים, אך מטעמי
 עניין וניתוח נתונים החלטנו להציג את עשרת הלקוחות המכניסים.
- כמו כן החלטנו להציג את הנתונים על הלקוחות בדשבורד המסונן לפי
 מדינות ושנים. כי ראינו ערך חשוב בלדעת מי הלקוחות המכניסים בכל
 מדינה או לחלופין אילו מדינות עם הלקוחות שהכי פחות מכניסים.

²³ avg_sales_per_customer 72.4478538813

```
select
    (sum("total_ILS")
    /
    count(distinct customerid))
    as avg_sales_per_customer
from "Fact_Invoice" fi
```



full_name	USD	ILS	Buyer's Country
Puja Srivastava	71	₪ 245.84	India
Denise Kelly	65	₪ 226.71	Brazil
Josephine Gomez	55	回 186.05	India
Helena Holý	50	回 171.95	Czech Republic
Richard Cunningham	48	回 157.99	USA
Kathryn Coleman	48	回 165.05	Ukraine
Annie Russell	48	回 165.95	Mozambique
Luis Rojas	47	回 165.20	Chile
Hugh O'reilly	46	₪ 157.41	Ireland
Ladislav Kovács	46	回 159.39	Hungary

SUM OF SALES BY MONTH



מסקנות:

- ניתן להסיק באילו חודשים יש הכי הרבה מכירות בפילוח של שנים.
- ניתן לראות ששנת 2019 נמצאת ברוב החודשים הכי גבוהה במכירות, וכן שברבעון האחרון של השנה נראה שיש ירידה דרסטית בכמות המכירות.
 - שנת 2021 נראית במגמה הפוכה כך שרוב השנה היא פחות במכירות מהשאר
 ובסוף רבעון האחרון של השנה יש עליה גדולה במכירות בעוד שבשנים האחרות
 הרבעון האחרון נחשב כחלש יותר.
- מהגרף והכרטיסייה ניתן לראות ששנת 2019 היא השנה הרווחית ביותר וניתן לראות את הירידה שהיה בשנים שלאחריה.

- התבקשנו להציג סכום מכירות עבור כל חודש, החלטנו להציג את זה בגרף להלן: גרף שמציג את כלל השנים ולפי מספר חודשים כאשר לכל שנה יש צבע משלה וגודל המכירות כגודל עובי. (מצג מספר 1)
- גרף הקווים ממחיש את ביצועי המכירות לאורך חמש השנים מ-2018 עד 2022, כאשר כל נקודה בציר ה- X מייצגת חודש. ציר ה- Y מציג את סכומי המכירות, המציג את התנודות החודשיות בהכנסות. מגמות כלפי מעלה או ירידה בקו מדגישות את הצמיחה או הירידה הכוללת במכירות לאורך השנים. ייצוג חזותי זה מאפשר הבנה ברורה של דפוסי המכירות, ומאפשר תובנות לגבי וריאציות עונתיות, חודשי שיא והמסלול הכללי של הביצועים העסקיים במהלך שצוין מסגרת זמן. (מצג מספר 2)
 - הטבלה נמצאת בדשבורד בו ניתן לסנן לפי השנים וכן לפי המכירות בסופי שבוע לעומת ימי השבוע. (מצג מספר 3)
- הוספנו בנוסף לטבלה כרטיסיה שמראה את השנה הכי רווחית וגם עוד גרף שמראה את הסך מכירות פר שנה (במאקרו, מצג מספר 4)
 - גרף הקווים מציג את ביצועי המכירות לאורך 5 השנים מ-2018 עד 2022, כאשר כל נקודה מייצגת את סך המכירות לכל שנה ובהתאם ליעד בציר ה-Y. (מצג מספר 5)





FINAL PROJECT CHINOOK DATABASE

CORRELATION BETWEEN THE LENGTH OF THE SONG AND ITS SALES



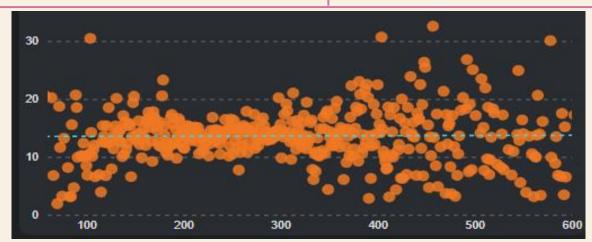
הסבר:

מסקנות:

- החלטנו להציג את הנתונים בגרף 'scatter' על מנת לבדוק האם קיים קשר בין אורך השיר, המופיע בציר ה- X, לבין סכום המכירות שלו שמוצג בציר ה- Y. באמצעות הגרף, אנו יכולים להבין את היחס בין המשתנים על פי שיפוע הקו, וכן לזהות את הפיזור הנראה סביבו שיש בין המשתנים.
 - במקרה שלנו בהסתכלות על כל השנים נראה שאין קשר בין אורך השיר לבין סכום המכירות שלו, ניתן לדעת זאת כיוון שהקו תצפית מקביל לקו הממוצע. נסיק כי אין קשר בין המשתנים.
 - עוד נוכל לראות כי רוב השירים מקובצים בין ב 2 ל5 דקות (100-300 שניות)
 - . גם בסינון על פי שנים עדיין לא נראתה קורלציה.

- כאשר נגשנו לשאלה ראינו שיש ערכים בודדים שהם קיצוניים מאוד. בבדיקה של אותם הערכים גילנו שהם מסוג מדיה שונה משירים כמו למשל וידאו קליפס, קונצרטים שהם באורך של מעל 40 דקות. עקב כך החלטנו לסנן תוצאות שאינם רלוונטיות לשירים, הסינון היה לשירים עד האורך 10 דקות שיר.
- הצמצום של הנתונים מצד אחד גורם לנו לאבד מעט מידע של שירים מעל 10 דק אך זה שולי כיוון שזה מדובר בתצפיות בודדות. הייתרון בכך הוא שניתן לראות את המגמה באמצעות הקו ישר בצורה נוחה יותר וכן לראות את הפיזור סביב הקו.
- בנוסף בדשבורד שלהלן יש סינון לפי שנים וכן לפי רבעונים, ניתן לשנות רזולוציות זמן ולראות האם היו זמנים בהם כן נראתה קורלציה. לדוגמה שנת 2018 רבעון השני.





5 COUNTRIES WITH THE MOST AND LEAST SALES



מסקנות:

- בגרפים שלהלן ניתן לראות את כמות ההכנסות לפי מדינה, יש את חמש המדינות הכי פחות מכניסות (Bot 5 Countries) וחמש הכי מכניסות (Top 5 Countries).
 - ניתן לראות שבהודו כמות ההכנסות הגדולה ביותר כ-4,300 ₪. לעומת מדינת גמביה שמשם ההכנסות הם כ-7 ₪ בלבד.
 - סך כל ההכנסות של חמשת המדינות המכניסות ביותר הוא 15.2 אלף ₪ שזה כשליש מכלל ההכנסות.
- בדיקה בעתיד ההינו רוצים לבדוק את ההכנסה של כל מדינה באופן פרופורציונלי לכמות האוכלוסייה בה, למסקנה זו הגענו כשראינו שהמדינות עם הכי הרבה מכירות הם גם המדינות עם הכי הרבה אזרחים בעולם.

הסבר:

- בדשבורד בו מופיעים הגרפים יש סינון לפי שני כפתורים האחד יחזיר את המדינות עם הכי הרבה הכנסות והשני יחזיר את אלו עם הכי פחות הכנסות.
 - את המידע של הסכומים הבאנו בשקלים.

Bot 5 Countries

- בנוסף בדשבורד עצמו יש ויזואליזציה של מפה שמראה איפה המיקומים של המדינות ומכך ניתן להסיק על יבשות או אזורים ספציפיים בהם. לדוגמה אנו רואים שכל המדורגים הגבוהים הם או מאסיה או מאמריקה בעוד שהמדורגות נמוך הן מאפריקה בעיקר וגם מאירופה
 - סך המבירות מתייחס לכלל המבירות ללא קשר לחלוקת המדינות.

Top 5 Countries







FINAL PROJECT CHINOOK DATABASE

THE PERCENTAGE OF SALES AMOUNT PER GENRE BY COUNTRY

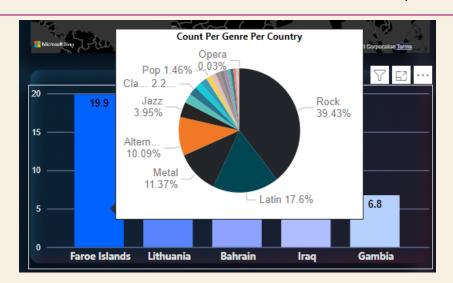


הסבר:

- החלטנו בוויזואליזציה את כמות המכירות לפי טבלה שבה ניתן לראות את הסכום הכנסות לפי קטגוריה בתצוגת אחוזים.
- בנוסף הוספנו 'tooltip' שמציג גרף עוגה שמחולקת לפי קטגוריות שירים. לפיכך, אם נעביר את העכבר על הגרף של המדינות נראה פילוח של קטגוריות שירים לפי המדינה.
- יש עמוד ברטיסיות שמציג את שם המדינה, ממוצע מכירות, כמות מכירות וסך המכירות
 בהתאם לסינון שנבחר.

מסקנות:

- אפשר להסיק מהחלוקה לפי מדינות שקטגוריית 'רוק' היא עם אחוז השמיעה הגבוהה ביותר.
 - ניתן לראות את עשרת הקטגוריות המובילות ביותר
- ניתן לראות שארבעת הקטגוריות נתפסים כחלק נכבד מסך כל המכירות כ-82% מכלל הקטגוריות.
 - המדינה עם הכי הרבה הכנסות היא ביבשת אסיה, הודו
- עם סך המכירות של 4,340 של כמות מכירות כ-1,240 והכנסה ממוצעת מלקוח
 בהודו היא \$11.61
 - ניתן לראות לעומת זאת את המדינה 'גאמביה' •



genre_name	% Of Total	invoiceline_total
Rock	40.58%	18,295.21
Latin	18.46%	8,322.64
Metal	12.01%	5,415.52
Alternative & Punk	11.42%	5,149.59
lazz	4.37%	1,969.25
Blues	2.85%	1,286.48
R&B/Soul	1.63%	734.89
Reggae	1.56%	703.71
Soundtrack	1.42%	641.07
Heavy Metal	0.87%	390.94
World Total	0.82% 100.00%	370.27 45.078.89



FINAL PROJECT CHINOOK DATABASE

OUR COMPETITION

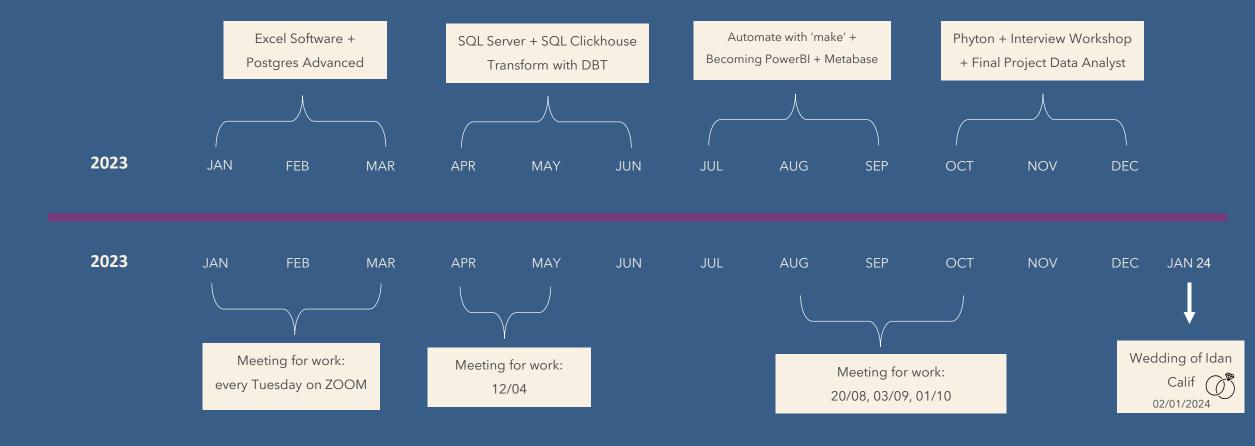


- ניתוח מגמות והסקת מסקנות בדגש על שינויים בקטגוריות
 השירים ומספר שנים
 - עיצוב פשוט וקל לשימוש, בהשוואה למתחרים.
 - שימוש ב-"dbt" עבור טרנספורמציות של מסד הנתונים "chinook_database".
 - איכות הנתונים והניקוי
- אופי הצוות מצייר תמונה ברורה של הבנה ואיכות ניתוח הנתונים.
 - הצגת הנתונים מושכת ויזואלית ואינפורמטיביות שניתן להסיק מסקנות בקלות.



- ✓ הפרויקט שלנו בולט בגישה של התייעלות, מינימליות ופשטות להבנה.
- ✓ אנו מציגים מגוון רחב של גרפים על מנת לקבל תובנות עמוקות ככל הניתן.
- ✓ ההתמקדות העיקרית של הפרויקט שלנו היא בחוויית משתמש, הניכרת בעיצוב האינטואיטיבי ובניווט חלק, המבדיל אותו מהמתחרים.
- ר הצגנו בעקביות מדדי ביצועים מעולים, ביצועים טובים יותר מהמתחרים בתחומים כמו Total Sales" "Through Months & Years

ONE YEAR ACTION PLAN



23

MEET THE TEAM



MATAN HILLEL

308341056



LIRAZ COHEN

316934207



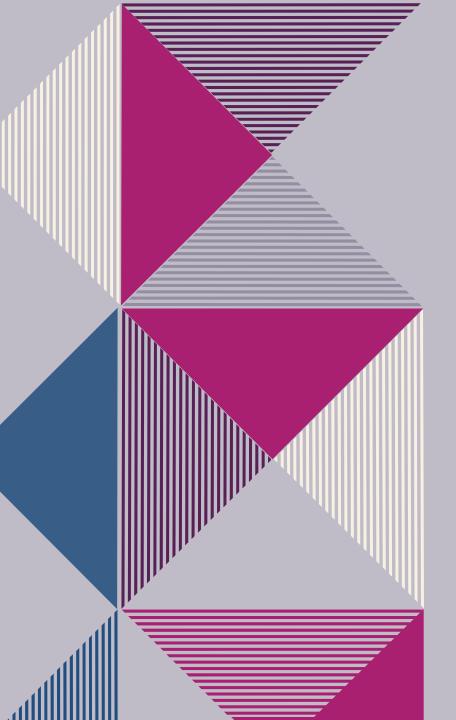
OMER BOARON

313133894



IDAN CALIF

305227274



SUMMARY

2023

בפרויקט הגמר, השקענו המון זמן ומחשבה בכל שאלה, גרף וניתוח של הנתונים. התחייבנו למוטו של "הכל אפשרי אז לא מתפשרים", ובאמת מרוצים מהתוצאות. אף שהשקענו המון בפרויקט, התרכזנו גם בלמידה המתקדמת שהעניק לנו הקורס במכללה.

הרחבנו את הקריאה שלנו ולמדנו פעולות מתקדמות בתוכנות שבהן השתמשנו. יצרנו והבנו לוגיקות מורכבות, השילוב של כל הכלים שלמדנו נתן לנו הבנה עמוקה של התהליכים הנדרשים ליצירת מסד נתונים יעיל ואינפורמטיבי, עם ויזואליזציות מתקדמות.

שדרגנו את ויזואליזציות והוספנו דאשבורדים נוספים, כל אחד עם יעד ספציפי ועיצוב מותאם. רצינו להרחיב את גבולות ההתנסות שלנו בתחום הוויזואלי.

לסיכום, אנו מאד שמחים על ההזדמנות להתנסות בפרויקט בסדר גודל כזה. למדנו שיעורים בצמיחה מקצועית ובעבודת צוות, והצלחנו לשדרג את יכולתנו הויזואלית והאנליטית.

