# Data analyst – משימה מסכמת

**מגישות:**

רויטל קנירש
ילנה בוברוב

טעינת ה data

טבלאות המידע של chinook נטענו ל DBeaver המקומי לסכמת stg

**פייתון - PANDAS**

הקבצים נטענו ל data frames, הטבלאות נוצרו לפי ההסברים בקובץ של ה script.

**DBT**

לאחר בדיקה של תאריך העדכון בכל הטבלאות, הנחנו שהעדכון של ה stg נעשה לכל ה data base ביחד, ולכן כאשר אחדנו טבלאות, השארנו תאריך עדכון אחד בכל טבלה.

יצרנו קובץ YML לסכימה עם שמות הטבלאות אליהן נפנה, ובכל המודלים השתמשנו ב source.

בכל טבלאות ה dimension ה materialization הוגדר כ table, מאחר ומדובר בטבלאות קטנות יחסית שאינן צפויות להתעדכן בתדירות גבוהה.

Dim_playlist

הבאנו את הרשומות מטבלת playlist ו playlisttrack
שמות טבלאות הוכנסו ל YML
מאחר ומדובר בטבלת dimensions קטנה יחסית, שמחזיקה מידע על כל ה tracks ואינה צפויה להתעדכן בתדירות גבוהה, ה materialization הוגדר כ table.

Dim_customer

הסברים לפעולות כתובים בקובץ.

מאחר והוצאת ה domain אינה פעולה חד פעמית, כתבנו פונקצית macro אשר תשמש אותנו במודלים נוספים.

Dim_employee

הסברים לפעולות כתובים בקובץ.

Dim_track

הסברים לקוד בקובץ.

אין צורך להביא מטבלת track את albumid, genreid, mediatypeid, מאחר ועמודות אלה מופיעות בטבלאות המצורפות.

Fact_invoice

הסברים לקוד בקובץ.

צריך להביא את עמודות הכתובת מטבלת invoice מאחר וה billing address אינה תמיד זהה לכתובת הלקוח (address מטבלת customers).

הטבלה הוגדרה כ incremental כך שמתעדכנות בה שורות אשר תאריך ה last update שלהן מאוחר יותר ממה שיש כרגע בטבלה, ז״א רק שורות חדשות או כאלה שעברו שינוי.

Fact_invoiceline

הטבלה הוגדרה כ incremental כך שמתעדכנות בה שורות אשר תאריך ה last update שלהן מאוחר יותר ממה שיש כרגע בטבלה, ז״א רק שורות חדשות או כאלה שעברו שינוי.

**API-currencies**

בקובץ הפייתון יצרנו בקשה למשיכת נתוני שער הדולר מאתר בנק ישראל. את טווח תאריכי שער הדולר המבוקשים קבענו בין התאריכי ה invoice המינימאלי והמקסימאלי, אותם הכנסנו לבקשת ה API כמשתנים בכדי שהקוד יהיה דינאמי, ז״א כאשר יתווספו או ירדו תאריכים מטבלת ה invoice, טווח התאריכים ישתנה בהתאם (כאשר מריצים שוב את הקוד) . כנראה, שבקשת ה API מוגבלת ל 1221 שורות, ולכן לא כל התאריכים בטווח שקבענו נכללים. בכדי לכלול בכל זאת את כל התאריכים בטווח השלמנו את המידע ע״י merge לטבלת תאריכים רציפים (בין טווח התאריכים שבחרנו). את השער הדולר עבור תאריכים שלא קבלנו באמצעות ה API השלמנו עם ערך שער הדולר בתאריך הקודם שאינו NULL (מתוך הנחה שהשינוי בשער הדולר בטווח של ימים בודדים אינו גדול וגם מתוך הנחה שחלק מהתאריכים החסרים הם ימים שלא יתקיים בהם מסחר ולכן שער הדולר נשאר זהה לקודם). את התאריך הראשון בטווח (שהיה חסר) השלמנו כערך שער הדולר ביום הבא.

**Power BI**

הגדרנו פלטת צבעים על פי צבעי תמונת הלוגו של החברה (Chinook) ועיצבנו את שני ה dashboards לפי צבעים אלה.

כל המדדים המחושבים נמצאים בתיקיית all-measures תחת טבלת fact invoiceline.

את הויזואליזציות הנדרשות חילקנו לשני dashboards :

Main Dashboard

ה dashboard הראשי מתמקד בתובנות כלליות של הזמנות ורווחים על פני ציר הזמן או כתלות בקריטריונים מסוימים :

1. **מטריצה בחלוקה לשנים ורבעונים**. עבור כל שנה ורבעון ניתן לראות את :

- מספר ההזמנות : המדד count_of_orders. בעזרת conditional formatting הוספנו data bar שמאפשר לזהות בקלות שנה/רבעון עם מס הזמנות גבוה או נמוך במיוחד.
- מס הלקוחות הפעילים : המדד count_of_customers, מאפשר לזהות רבעונים/ שנים עם ירידה או עלייה בפעילות לקוחות
- סכום ההכנסות : המדד sum_sales (revenue). את סכום הכנסות חישבנו לפי המחיר ליחידה בטבלת Fact_invoiceline כדי להמנע מכפילויות בחיתוך ל trackid.

- אחוז ההכנסות: המדד sum_sales%, אחוז ההכנסות מכל בכל שנה ביחס לכלל ההכנסות ואחוז ההכנסות בכל רבעון ביחס להכנסות בכל השנה (revenue %).

2. **גרף של סכום ההכנסות לפי חודש ושנה** (סעיף 6 במשימה) – סכום הכנסות לפי start of month (עמודה שהוספה לטבלת ה invoice בשלב המודל.

3. **הפלייליסט עם הכי הרבה שירים** (סעיף 4 במשימה) – שם הפלייליסט ומספר השירים <u>הייחודי</u>, מאחר ומצאנו שיש שני פלייליסטים משוכפלים, אחד מהם הוא הפליליסט music (מופיע פעמיים תחת שני playlisted (1 ו 8)) אם לא נבחר ב distinct count של trackid, כמות השירים תוכפל. הטבלה מפולטרת ל top1 לפי מספר השירים.

4. **הפלייליסט עם הכי מעט שירים** (סעיף 4 במשימה) – שם הפלייליסט ומספר השירים הייחודי. הטבלה מפולטרת ל 1 bottom לפי מספר השירים.

5. **ממוצע השירים בפלייליסט** – המדד: AvgTrackCount, השתמשנו ב summarize, המאגרג את המידע לפי playlisted ומחשב את מספר השירים לכל פלייליסט. בעזרת filter הוצאנו מהחישוב את שני הפלייליסטים הכפולים (3 ו 8), כדי שחישוב הממוצע יהיה מודיק ולא מושפע מכפילויות. על המידע המפולטר והמאוגרג חישבנו ממוצע.

6. **גרף מפה המראה את כמות ההזמנות לכל מדינה** - באופן הזה ניתן לזהות מדינות עם ביקוש גבוה או נמוך ולבמע החלטות עסקיות בהתאם.

7. **תרשים עמודות** המאפשר לבחון את כמות ההזמנות או סכום המכירות לפי חיתוכים שונים (שם האלבום, ג'אנר ומדינה), ע"י בחירת הפרמטרים להצגה והחיתוכים משני boxes הנמצאים מעל הגרף. בצורה זו ניתן לזהות טרנדים שונים (לפי הצורך) ע"י שימוש בגרף אחד (שימוש במדדים sum_sales, count_of_orders).

ב dashboard הזה ישנו סלייסר של זמן (המשותף ל Top dashboard). בנוסף ישנו כפתור המאפס את הסלייסרים ב dashboard ובראש העמוד ישנו כפתור המעביר ל dashboard השני (Top dashboard) .

<u>Top Dashboard</u>

ה dashboard השני מתמקד ב-5 המובילים בהזמנות, מכירות או במספר אלבומים בחיתוך לקריטריונים שונים ומספק מיידע על:

1. **5 הלקוחות עם סכום התשלומים הגבוה ביותר** (סעיף 5 במשימה) – טבלה המפרטת את מספר זיהוי הלקוח, שם ושם משפחה, סכום התשלומים בדולרים (עם background color לבן-ירוק ככל שהסכום גדול) וסכום התשלומים בשקלים (בהתבסס על שער הדולר מטבלת dim_currency). פלטור ה top 5 מתבסס על סכום התשלומים בדולרים. שימוש במדדים: sum_sales, sum_sales_ILS.

2. **5 האומנים עם מספר האלבומים הגדול ביותר** (סעיף 1 במשימה) - גרף עמודות של כמות האלבומים (ספירת albumid) לכל אמן, המפולטר ל 5 האמנים עם מס האלבומים הגבוה ביותר.

3. **5 האומנים עם מספר השירים הגבוה ביותר** (סעיף 2 במשימה) – טבלה המפרטת את שם האמן ומספר השירים (ספירת trackid) לכל אמן. הטבלה מפולטרת ל 5 האמנים עם מס השירים הגבוה ביותר,בפועל מופיעים 6 אמנים מאחר ולשניים (במקום ה 5) יש מספר זהה של שירים.

4. **5 הג'אנרים עם מספר השירים הגבוה ביותר** (סעיף 3 במשימה) - גרף עמודות של כמות השירים (ספירת trackid) לכל ג'אנר המפולטר ל 5 הג'אנרים עם מס השירים הגבוה ביותר.

5. **5 הארצות בהן סכום המכירות הגבוה ביותר** (סעיף 8 במשימה)
6. **5 הארצות בהן סכום המכירות הנמוך ביותר** (סעיף 8 במשימה)

- בשני הגרפים האחרונים הוספנו tooltip של אחוז המכירות של כל ז'אנר מתוך סך המכירות במדינה (סעיף 9 במשימה).

7. **גרף scatter chart של סכום המכירות כתלות באורך השיר** (סעיף 7 במשימה) מראה שהשירים הנמכרים ביותר הם בטווח של 170-350 שניות. טווח נוסף בו רואים עלייה במכירות הוא באזור 2600 שניות.

8. **שם השיר עם מספר ההזמנות הגבוה ביותר** - בראש ה dashboard מוצג השיר המוזמן ביותר עם tooltip שמציג את כמות ההזמנות של השיר.

ב dashboard הזה ישנם סלייסרים של זמן (המשותף ל dashboard הראשי) וג'אנר. בנוסף ישנו כפתור המאפס את הסלייסרים ב dashboard ובראש העמוד ישנו כפתור המחזיר ל dashboard הראשי.