Final Project

Answers file

"chinook"

DWH – SQL :

הסיבה ליצירת סכמת dwh2 (במקום dwh) :

במהלך העבודה יצרתי תחילה את הסכמה בשם dwh אך זיהיתי בה טעויות ושגיאות שהחלו להצטבר עוד בשלבים הראשונים.

במקום לנסות לתקן נקודתית ולהסתכן בשגיאות ו/ או כפילויות, החלטתי להקים סכמה חדשה בשם dwh2.

בסכמה הזו בניתי את כל הטבלאות מחדש, בצורה נקייה ומסודרת, תוך הקפדה על שימוש אך ורק בטבלאות ובקבצים שנדרשו במטלה וכך הבטחתי שהעבודה שתוגש תתבסס על נתונים עקביים, עדכניים ונכונים.

1. dim_playlist:

בבניית טבלת dim_playlist חיברתי את הטבלאות playlist ו- playlisttrack בעזרת INNER JOIN.

הסיבה היא שרציתי לכלול רק פלייליסטים שיש להם בפועל שירים מקושרים.

אילו הייתי משתמשת ב - LEFT JOIN, הייתי מקבלת גם פלייליסטים "ריקים" ללא תוכן, דבר שהיה מוסיף נתונים שהם לא רלוונטיים.

2. dim_track :

בעת בניית dim_track, בחרתי להשמיט עמודות מיותרות כגון, עמודות מזהי מערכת פנימיים או שדות שאין להם ערך אנליטי.

המטרה הייתה לשמור על טבלה נקייה, ורלוונטית לניתוחים, מבלי להעמיס מידע שאינו תורם להבנה העסקית.

3. Fact_invoice :

בבניית טבלת fact_invoice , החלטתי שלא להביא את שדות הכתובת מתוך טבלת ה־ invoice המקורית.

ההיגיון הוא שטבלאות fact אמורות להכיל אך ורק מדדים כמותיים (כמו סכום החשבונית, מפתחי חיבורים לטבלאות מימד, תאריכים רלוונטיים).

כל מידע תיאורי שאינו נמדד באופן ישיר, שייך לטבלאות dim למשל, הכתובת של הלקוח, היא נתון ש"מתאר" את הלקוח ולכן מקומה של עמודה זו הוא בטבלת dim_customer, הכנסת שדות כדוגמת "כתובת" ל־ fact הייתה "מכבידה" על הטבלה, מייצרת כפילויות, ומפרה את עקרונות מודל הכוכב Star Schema.

חשוב להדגיש כי אני מודעת לכך שייתכנו מצבים בהם עמודות הכתובת יציגו את ה"כתובת לחיוב" ולא את כתובת המגורים של הלקוח ובכל זאת, את עמודות הכתובת (כולל כתובת לחיוב) בחרתי לרכז ב – dim_customer ובכך, שמרתי על הפרדה ברורה בין נתונים תיאוריים (ב- dim) לבין נתונים עסקיים/כמותיים (ב- fact).

2 אנליזות נוספות:

1. שאלת SQL :

מציאת הלקוחות שנמצאים בסיכון נטישה על-ידי בחינת הנתונים הבאים (RFM) ודירוגם לפי דרגות סיכון:

R (recency) - בדיקה כמה זמן עבר מאז הרכישה האחרונה של כל לקוח (ככל שטווח הזמן גדול יותר, כך סיכויי הנטישה גדולים יותר).

F (frequency) – בדיקת כמות החשבוניות של כל לקוח.

M (monetary) – כמה כסף הוצא כל לקוח.

שם הקובץ: *churn_risk_RFM.sql*

הבחירה לשאול את השאלה "מי הם הלקוחות שבסיכון נטישה" נבעה מתוך הבנה עסקית מהותית, לפיה, כל ארגון שמבוסס על לקוחות חוזרים, חייב לדעת לזהות מבעוד מועד מי הלקוחות שעלולים להפסיק לרכוש ממנו וזאת, כדי שיוכל לבצע פעולות שימור ממוקדות. השאלה שבחרתי לנתח, בוחנת את איכות הקשר עם הלקוח עצמו, בדקתי מתי הוא קנה לאחרונה, כמה פעמים הוא ביצע רכישה ומה ההיקף הכספי של הרכישות שביצע ובעצם, רציתי לשלב בין נתונים טכניים ופשוטים, לבין הפקת תובנות עסקיות עמוקות. תוך כדי תהליך החשיבה ויישום השאילתה, למדתי שמודל RFM הינו שיטה מוכרת ויעילה למדידה ראשונית של ערך ושל נאמנות לקוחות. לכן, בניתי שאילתה שמביאה תחילה את כלל העסקאות של כל לקוח, מחשבת את סכום ההוצאות, סופרת את מספר הרכישות שביצע ושולפת את תאריך הרכישה האחרון שלו.

לאחר מכן, דירגתי את הלקוחות לפי שלושת הפרמטרים של המודל: R= סיכון גבוה לנטישה ו- F /M = סיכויים נמוכים יותר והחזרתי טבלה שמציגה סיווג של כל לקוח לפי רמת "סיכון-נטישה" שונה ("נאמן", "בינוני" ו"סיכון גבוה").

הפלט שמתקבל, מציג לכל לקוח את השם המלא שלו, כמה פעמים רכש, כמה כסף הוציא בסה"כ, מתי ביצע רכישה אחרונה, כמה ימים עברו מאז, הדירוגים היחסיים שלו בכל אחד מהמדדים ולבסוף, את הדירוג הכולל שלו לפי רמות הסיכון שהגדרתי.

תצוגת דוגמאות של לקוחות בדירוגים שונים:

customer_id integer	customer_name text	num_invoices bigint	total_spent numeric	last_purchase date	days_since_last integer	r_tile integer	f_tile integer	m_tile integer	churn_flag text
595	Clinton Buford	6	29.71	2021-10-10	1417	2	3	2	High Risk
57	Luis Rojas	7	46.62	2021-10-14	1413	2	2	1	High Risk
[illegible]	[illegible]	7	31.69	2021-10-25	1402	2	2	2	High Risk
[illegible]	[illegible]	7	37.62	2021-11-01	1395	2	3	1	High Risk
[illegible]	[illegible]	10	32.69	2021-11-09	1387	2	1	1	High Risk
[illegible]	[illegible]	9	35.68	2021-11-10	1386	2	1	1	High Risk
[illegible]	[illegible]	7	37.62	2021-11-14	1382	2	2	1	High Risk
[illegible]	[illegible]	7	21.79	2021-11-18	1378	2	2	3	High Risk

customer_id integer	customer_name text	num_invoices bigint	total_spent numeric	last_purchase date	days_since_last integer	r_tile integer	f_tile integer	m_tile integer	churn_flag text
497	Alexander Fen...	7	25.76	2022-05-13	1202	3	2	2	Medium ...
409	Jack Foust	4	12.87	2022-05-16	1199	3	4	4	Medium ...
610	Hugh Waldrop	1	4.95	2022-05-17	1198	3	5	5	Medium ...
152	Norma Gonzal...	12	45.55	2022-05-18	1197	3	1	1	Medium ...
28	Julia Barnett	7	43.62	2022-05-19	1196	3	2	1	Medium ...
160	Crystal Ford	12	33.67	2022-05-19	1196	3	1	1	Medium ...
490	Edwin Burk	4	17.82	2022-05-20	1195	3	4	3	Medium ...
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Medium ...
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Medium ...
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Medium ...

customer_id integer	customer_name text	num_invoices bigint	total_spent numeric	last_purchase date	days_since_last integer	r_tile integer	f_tile integer	m_tile integer	churn_flag text
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
284	Maureen Little	3	8.92	2021-04-26	1584	1	5	5	Loyal
341	Felicia Sutton	2	4.95	2021-05-09	1571	1	5	5	Loyal
607	Nelson Christ...	1	3.96	2021-05-10	1570	1	5	5	Loyal
406	Roger Quintan...	2	5.95	2021-05-25	1555	1	5	5	Loyal
400	Harold Martino	4	14.86	2021-05-26	1554	1	4	4	Loyal
557	Marc Outlaw	3	10.91	2021-05-29	1551	1	4	4	Loyal
462	Stanley Scrog...	4	10.89	2021-05-29	1551	1	4	4	Loyal
585	Cory Meehan	3	10.91	2021-05-31	1549	1	4	4	Loyal
640	Andy Vanhorn	1	1.98	2021-06-10	1539	1	5	5	Loyal

2. שאלת python:

מציאת הקשר, אם קיים קשר, בין חודשי השנה לבין מכירות שירים לפי ז'אנרים.

שם הקובץ: *seasonality_by_genre.py*

הבחירה לבחון אם קיים קשר עונתי-תקופתי לפי מכירות השירים לפי ז'אנרים, עמד לנגד עייני הצורך להבין את סך התפלגות המכירות לפי ז'אנרים אך יתרה מזו, עניינה אותי "התנהגות" שנמשכת לאורך זמן.

בעולם האמיתי, חברה כמו "chinook" נדרשת לזהות עונות "חמות" ועונות "רגועות" יותר עבור סוגי תוכן שונים, על מנת לייעל החלטות שיווקיות, תמחור וניהול מלאי .

השלב הראשון היה לזהות את חמשת הז'אנרים המרכזיים מבחינת היקף המכירות הכולל שלהם, מתוך כוונה להתרכז באזורים שבהם יש השפעה עסקית גדולה יותר ולא "לבזבז" זמנים על ניתוח ז'אנרים ובעצם, למנוע "רעש מיותר" מהקטגוריות הקטנות.

לאחר מכן, איחדתי את המכירות לפי חודש לאורך כל השנים (ולא בפירוק לפי שנה) משום שכך, אפשר לחשוף דפוסי עונתיות עקביים שאינם תלויים במקריות של שנה מסוימת (כך אפשר להבדיל בין "אירוע חד־פעמי" לבין מגמה יציבה).

ניתוח מסוג זה, מאפשר לחברה (דוגמת "chinook") לשאול שאלות כגון:

- באילו חודשים כדאי להשיק קמפיינים פרסומיים מותאמים ז'אנר?
- האם קיימת עונתיות מובנית (למשל, ז'אנרים שנמכרים יותר בקיץ או בחגים)?
- כיצד ניתן לתזמן מבצעים חכמים או שיתופי פעולה עסקיים כדי למקסם הכנסות?

וכיוצא בזה.

תצוגת גרף :