מגישות:
ענבל ברזל 316421320
רותם כספין 206395246
נעה ששון 316013994

פרויקט גמר בינה עסקית

I. שאלה עסקית + KPIs

השאלה העסקית:
כיצד נוכחות של רכישות מאומתות משפיעה על הדירוגים שניתנו על ידי לקוחות במערך הנתונים?

מטרת הפרויקט היא לשקף תמונת מצב כללית בנוגע לנוכחות של רכישות מאומתות והשפעתן על דירוג המוצרים שמופיע באתר "אמזון". בעקבות התובנות שנקבל, נוכל להמליץ על אופן הטיפול בתגובות באתר, למשל, האם כדאי לשים דגש על סינון תגובות שאינן מאומתות ולהציג את אימות הרכישה לצד התגובה.

הגדרנו 3 KPIs עבור השאלה העסקית:

1. הגדרה: אחוז הרכישות המאומתות מתוך סך הרכישות.
הסבר SMART:
- ספציפי: ה-KPI מתמקד במדידת שיעור הרכישות המאומתות מסך הרכישות.
- ניתן למדידה: ניתן לחשב את האחוז בצורה מדויקת באמצעות הנתונים מהשאילתות, מה שמאפשר מדידה כמותית והשוואה לאורך זמן.
- בר השגה: הנתונים הדרושים לחישוב ה-KPI הזה זמינים במסד הנתונים, מה שהופך אותו לאפשרי מעקב וניטור.
- רלוונטי: ה-KPI מתייחס ישירות לשאלת המחקר מכיוון שהוא עוזר להבין את השכיחות של אימות רכישה במערך הנתונים.
- מוגבל בזמן: ניתן למדוד את שיעור הרכישות המאומתות עבור תקופת זמן מוגדרת.

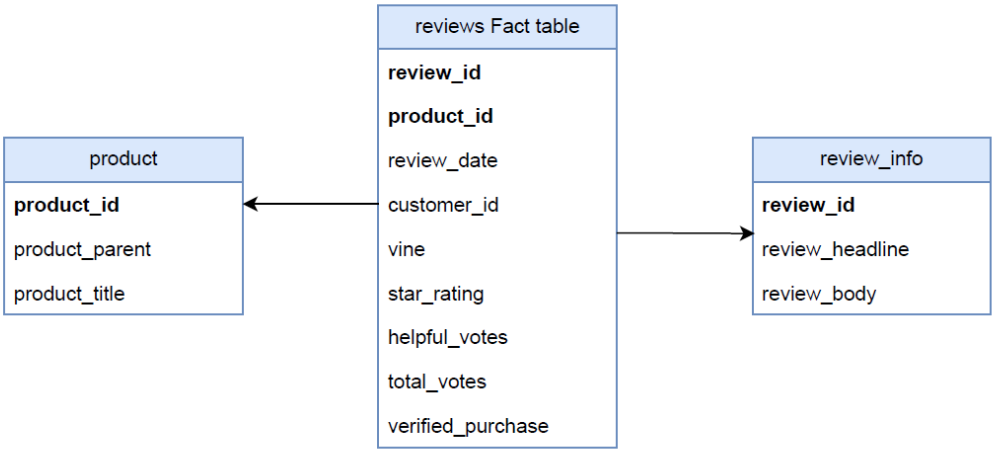
2. הגדרה: דירוג הכוכבים הממוצע עבור רכישות מאומתות ורכישות לא מאומתות.
הסבר SMART:
- ספציפי: ה-KPI מתמקד במדידת דירוג הכוכבים הממוצע עבור כל סטטוס אימות רכישה, ומספק מדדים ספציפיים להשוואה.
- ניתן למדידה: ניתן לחשב את דירוג הכוכבים הממוצע באמצעות הנתונים מהשאילתות, מה שמאפשר מדידה כמותית והשוואה.
- בר השגה: הנתונים הדרושים לחישוב KPI זה זמינים במסד הנתונים, מה שהופך אותו בר השגה למעקב ולניטור.
- רלוונטי: ה-KPI מתייחס ישירות לשאלת המחקר על ידי הערכה כיצד נוכחות אימות רכישה משפיעה על הדירוגים שניתנו על ידי הלקוחות.
- מוגבל בזמן: ניתן למדוד את שיעור הרכישות המאומתות עבור תקופת זמן מוגדרת.

3. הגדרה: אחוז ההערות החיוביות (מעל סף דירוג מוגדר) מבין הרכישות המאומתות.
הסבר SMART:

- ספציפי: ה-KPI מתמקד במדידת שיעור ההערות החיוביות במיוחד בקרב רכישות מאומתות.
- ניתן למדידה: ניתן לחשב את האחוז באמצעות הנתונים מהשאילתות, מה שמאפשר מדידה והשוואה כמותית.
- בר השגה: הנתונים הדרושים לחישוב KPI זה זמינים במסד הנתונים, מה שהופך אותו בר השגה למעקב ולניטור.
- רלוונטי: ה-KPI מתייחס ישירות לשאלת המחקר על ידי בחינת הקשר בין אימות רכישה להערות חיוביות.
- זמן מוגבל: ניתן למדוד ולנטר את ה-KPI לאורך זמן כדי לעקוב אחר כל שינוי או מגמה באחוז ההערות החיוביות בקרב רכישות מאומתות.

II. הגדרת Data-Warehouse

1. בחרנו בסכמת Star עבור הדאטה. הטבלה המרכזית (Fact) היא טבלת Reviews שמכילה את הנתונים שאנו מנתחות. הסכמה קלה להבנה מכיוון שדרך הטבלה המרכזית ניתן להגיע בקלות לטבלאות הממדים וכך לשלוף נתונים במהירות בעזרת שאילתות SQL.
2.



3. בחרנו בסכמת Star משום שהיא היעילה ביותר עבור הדאטה הנתון. להלן שני use-case המדגימים את יעילות הסכמה:

- ניהול ותחזוקה פשוטים של נתונים: הפרדת הנתונים לטבלאות מרובות בהתבסס על הקשרים הלוגיים ביניהן מסייעת לפשט את ניהול הנתונים והתחזוקה. ניתן לעדכן או לשנות כל טבלה באופן עצמאי מבלי להשפיע על הטבלאות האחרות, מה שמקל על הטיפול בעדכוני נתונים, הוספות ומחיקות. מבנה מודולרי זה מאפשר יותר גמישות ומדרגיות ככל שמערך הנתונים גדל או מתפתח עם הזמן. לדוגמה, ניתן להפריד מידע הקשור לביקורות בטבלת fact ו-dimReviewInfo, ומידע הקשור למוצר ל-dimProduct.

- צמצום כמות JOIN ואחזור מהיר יותר: על ידי מבנה הנתונים המבוזר בטבלאות נפרדות (dimReviewInfo, dimProduct ,fact) והקמת קשרי מפתחות זרים מתאימים, מסד הנתונים מאפשר פעולות JOIN יעילות. לדוגמה, בעת שליפת מידע על ביקורת ספציפית, כגון תוכן הביקורת ופרטי המוצר המשויכים, ניתן לבצע JOIN בין טבלאות fact, dimReviewInfo ו-dimProduct על סמך מפתחות review_id ו-product_id. בנוסף, סכמת הכוכב מפחיתה את הצורך באחסון נתונים מיותרים בטבלת fact, המרכזית והשימושית ביותר. בכך, ממזערת את גודל מערך התוצאות המצורפות במהלך שליפת השאילתות ומשפרת את מהירות שליפת הנתונים.

III. תהליך הETL

שלבי הETL בפרויקט שלנו:

- EXTRACT- חילוץ הנתונים שלנו התבצע מקובץ tsv לקובץ csv. לאחר שהבנו שהשימוש בקובץ json יהיה יעיל יותר, המרנו את קובץ הcsv לjson.
- TRANSFORM- ביצענו סינון על העמודות והשארנו את העמודות הרלוונטיות לנו, בדקנו האם קיים מידע לא תקין והסרנו אותו.
- LOAD- ייצאנו את הטבלאות המעובדות שקיבלנו ל3 קבצי csv, בהתאם לסכמת הכוכב (טבלת Fact, טבלת dimReviewInfo וטבלת dimProduct).

מיפוי העמודות ומימוש תהליך הSTTM נמצאים בJupyter Notebook המצורפת.

IV. ניתוח הData Warehouse

שאילתות הSQL שכתבנו התבצעו על 5332 שורות (כתובות גם בJupyter Notebook):

1. מהו מספר התגובות בעלות אימות רכישה עבור כל דירוג (1-5)?

```
1
2 SELECT star_rating, COUNT(*) AS verified_purchase_count
3 FROM fact
4 WHERE verified_purchase = 'Y'
5 GROUP BY star_rating
6 ORDER BY star_rating;
7
```

star_rating	verified_purchase_count
1	365
2	231
3	358
4	605
5	3097

2. האם יש הבדל משמעותי, כלומר פער של יותר מ0.5 בדירוג הכוכבים הממוצע של תגובות בעלות אימות רכישה לבין זה של תגובות ללא אימות רכישה?

```
SELECT DISTINCT verified_purchase, AVG(star_rating)
over (partition BY (verified_purchase)) AS AvgRating
FROM fact;
```

verified_purchase	AvgRating
N	4.343703703703704
Y	4.253865979381444

3. הצג את כמות התגובות בעלות אימות רכישה וללא אימות רכישה. לתגובות בעלות הצבעות מועילות (helpful votes) בעלות יותר מ-2 הצבעות.

```
SELECT DISTINCT verified_purchase, COUNT(*)
over (partition BY verified_purchase) AS Counter
FROM fact
WHERE helpful_votes>2;
```

verified_purchase	Counter
N	61
Y	256

4. הצג את ממוצע הדירוגים של תגובות ללא אימות רכישה ובעלות אימות רכישה, עבור ביקורות ארוכות (review_body), כלומר שאורכן גדול יותר מאורך התגובה הממוצע.

```
SELECT DISTINCT verified_purchase, AVG(star_rating)
over (partition BY (verified_purchase)) AS AvgRating
FROM fact AS f JOIN dimReviewInfo AS dRI
ON f.review_id = dRI.review_id
WHERE LENGTH(dRI.review_body) > (
  SELECT AVG(LENGTH(review_body))
  FROM dimReviewInfo
);
```

verified_purchase	AvgRating
N	4.298611111111111
Y	4.065481758652947

5. מהו דירוג הכוכבים הממוצע עבור ביקורות עם מספר גבוה של הצבעות מועילות (מעל 2) בקרב תגובות ללא אימות רכישה ובעלות אימות רכישה?

```
SELECT DISTINCT verified_purchase, AVG(star_rating)
over (partition BY (verified_purchase)) AS AvgRating
FROM fact
WHERE helpful_votes>2;
```

verified_purchase	AvgRating
N	4.377049180327869
Y	3.8515625

6. הצג את product_id ושם המוצר שיש להם את הדירוג הממוצע הנמוך ביותר, 10 תגובות לפחות ואימות רכישה. נדרש להציג רק את 10 המוצרים הראשונים.

```
SELECT p.product_id, p.product_title, AVG(f.star_rating) AS average_rating
FROM dimProduct AS p
INNER JOIN fact AS f ON p.product_id = f.product_id
WHERE f.verified_purchase = 'Y'
GROUP BY p.product_id, p.product_title
HAVING COUNT(*) >= 10  -- Limit to products with at least 10 comments
ORDER BY average_rating ASC
LIMIT 10;
```

product_id	product_title	average_rating
B00NBAKXIA	Radha Beauty Eye Cream for Dark Circles, Puffiness, Bags & W...	3
B00M1E5F1W	Philips Norelco Rq12+ Replacement Head For Series 8000 (Se...	3.25
B0000YUXI0	Mavala Switzerland Mavala Stop Nail Biting	3.3333333333333335
B000BY2N7S	NOW Foods Biotin 5000mcg	3.5
B00DPE9EQO	OZNaturals- Vitamin C Serum For Your Face (Packaging May V...	3.5
B00IDWP4IA	Vitamin C Serum with Hyaluronic Acid & Vit E - Natural & Organi...	3.5
B003V264WW	Remington AC2015 T|Studio Salon Collection Pearl Ceramic Ha...	3.6666666666666665
B00461F4PA	Baby Foot Exfoliant Foot Peel, Lavender Scented, 2.4 Fl. Oz.	3.6666666666666665
B00KCFAZTE	InstaNatural Eye Gel Cream - Wrinkle, Dark Circle, Fine Line & ...	3.75
B00ARF42H0	Philips Norelco Multigroom 3100 with 5 attachments and skin-fri...	3.888888888888889

7. הצג את כמות התגובות עם אימות רכישה וללא אימות רכישה לתגובות עם המילה "love".

```
SELECT DISTINCT verified_purchase, COUNT(*)
over (partition BY verified_purchase) AS Counter
FROM dimReviewInfo AS ri JOIN fact AS f ON ri.review_id = f.review_id
WHERE LOWER(ri.review_body) LIKE '%love%';
```

verified_purchase	Counter
N	193
Y	925

8. המוצרים ואחוז התגובות החיוביות (מעל 4 כוכבים) מסך כל התגובות, עבור 10 product_id עם הכי הרבה תגובות.

```
SELECT p.product_title,
       COUNT(*) AS total_comments,
       ROUND((COUNT(*) FILTER (WHERE f.star_rating > 4) * 100.0 / COUNT(*)), 2) AS percentage_positive
FROM dimProduct AS p
JOIN fact AS f ON p.product_id = f.product_id
GROUP BY p.product_title
ORDER BY COUNT(*) DESC
LIMIT 10;
```

product_title	total_comments	percentage_positive
Crest 3D White Brilliance Toothpaste, Teeth Whitening and Dee...	2304	43.75
HSI Professional Glider | Ceramic Tourmaline Ionic Flat Iron Hai...	374	83.42
Secret Outlast Clear Gel Antiperspirant and Deodorant Scent, 2....	196	50
Crest Pro-Health Advanced Mouthwash with Extra Deep Clean	144	58.33
Philips Norelco Multigroom 3100 with 5 attachments and skin-fri...	100	40
YEOUTH Best Anti Aging Vitamin C Serum with Hyaluronic Acid...	81	44.44
Waterpik Aquarius Water Flosser	81	55.56
Thayers Alcohol-free Rose Petal Witch Hazel Toner (3 Pack) 12...	81	88.89
Radha Beauty Tea Tree Essential Oil 4 oz - 100% Pure and Nat...	64	62.5
Hydrating Argan Oil Hair Mask and Deep Conditioner By Arvaza...	64	62.5

9. הצג את product_id, כמה פעמים מופיע הביטוי "great" בכותרת התגובה עבור כל product_id,עבור כל תגובה ללא אימות רכישה לפי סדר יורד.

```
SELECT DISTINCT fact.product_id, COUNT(*) over (PARTITION BY fact.product_id) AS Occurrences
FROM fact
JOIN dimReviewInfo ON fact.review_id = dimReviewInfo.review_id
WHERE dimReviewInfo.review_headline LIKE 'great%' AND fact.verified_purchase = 'N'
ORDER BY Occurrences DESC;
```

product_id	Occurrences
B00ZKLLZAI	4
B007K6K0IY	2
B00HJDIAGC	2
604113452X	1
B000BNRKPY	1
B000G1MT2U	1

10. הצג את customer_id, ממוצע הדירוגים שלו ומספר התגובות שלו אשר בעלות אימות רכישה בסדר יורד.

```sql
SELECT f1.customer_id,
       AVG(f1.star_rating) OVER (PARTITION BY f1.customer_id) AS average_rating,
       COUNT(f2.customer_id) AS purchase_count
FROM fact f1
JOIN fact f2 ON f1.customer_id = f2.customer_id
WHERE f1.verified_purchase = 'Y'
GROUP BY f1.customer_id
ORDER BY purchase_count DESC;
```

customer_id	average_rating	purchase_count
25415089	3	121
51832538	5	121
11599687	5	64
18930380	5	64
43301034	5	64
45348717	5	64
1488387	3	49
10532106	5	49
10721966	5	49

V. מסקנות

לאור תוצאות השאילתות שקיבלנו נוכל להסיק מספר מסקנות:

1. אין הבדל מובהק בין ממוצע דירוגי הכוכבים של תגובות עם אימות רכישה וללא אימות רכישה, לכן משאילתה זו בלבד לא ניתן להסיק שתגובות ללא אימות רכישה מדרגות מוצרים בציון גבוה יותר. עם זאת, כאשר בדקנו גם את ממוצע דירוגי הכוכבים לתגובות עם סימוני 'תגובה מועילה' (לפחות 2), מצאנו כי תגובות בעלות אימות רכישה דירגו את המוצרים בציון נמוך יותר מאשר תגובות ללא אימות רכישה. מכאן ניתן להסיק כי תגובות ללא אימות רכישה אשר קיבלו פידבק חיובי כן מדרגות מוצרים בציון גבוה יותר.
2. מצאנו כי תגובות עם אימות רכישה קיבלו יותר סימונים של 'תגובה מועילה' (לפחות 2). ניתן להסיק כי קיימת השפעה של אימות רכישה על אינטראקציות בין הלקוחות.
3. בדקנו האם תגובות ללא אימות רכישה, אשר חשודות כמזויפות, נוטות להיות ארוכות יותר אך לא מצאנו תשובה ודאית. בנוסף, בדקנו האם תגובות רבות מסוג זה מכילות את המילה "great" בכותרת התגובה. אולם גם כאן לא הגענו למסקנה כי תגובות ללא אימות רכישה גוררות נוכחות גבוהה של מילה זו.

4. זיהינו 10 מוצרים אשר יש להם מספר רב של רכישות עם אימות רכישה והדירוג הממוצע שלהם הוא הנמוך ביותר. נסיק כי קיים חוסר שביעות רצון ממוצרים אלו וכדאי לבדוק מה הסיבה. בנוסף, זיהינו 10 מוצרים בעלי מספר התגובות הרב ביותר ומהו אחוז התגובות שקיבלו דירוג 4 ומעלה. ניתן להסיק מהם המוצרים המובילים מבחינת שביעות רצון הלקוחות. הופתענו לגלות שקיימים מספר לקוחות בעלי כמות רכישות מאומתות גבוהה וממוצע דירוגים 5. נסיק כי כדאי לשמר לקוחות אלו בדרכים של מתן הנחות, דוגמיות של מוצרים בחינם וכו'.
5. מצאנו כי רוב התגובות שהכילו את המילה "love" הן בעלות אימות רכישה. ניתן להסיק כי תגובות ללא אימות רכישה החשודות כמזויפות, משתמשות פחות במילה "love" כדי לתאר את המוצר והרכישה.

קישור לGitHub שלנו: https://github.com/InbalTb/BI-project1.git