**פרויקט בינה עסקית**

**שלב 3**

**https://github.com/EladHaziza94/BI-Project**

|  |  |
| --- | --- |
| **מגישים – קבוצה F** | |
| **ת.ז** | **שם** |
| **307854422** | **תומר בן- יהודה** |
| **312326788** | **אלעד חזיזה** |
| **313522534** | **עידן כהן** |

**חלק 1**

ניתן לעיין גם תחת STTM בגיט.

1. בקישור [זה](https://docs.google.com/spreadsheets/d/1Qt04K8XrtRYiA6kd3CntQ9rzEvoZBsSEmlLc6L5bDd0/edit?usp=sharing), בטאב הראשון, מצורף הSTTM של הפרויקט שלנו.
2. בקישור [זה](https://docs.google.com/spreadsheets/d/1Qt04K8XrtRYiA6kd3CntQ9rzEvoZBsSEmlLc6L5bDd0/edit?usp=sharing), בטאב השני, מצורף הסכמה המבוקשת בשאלה.

**חלק 2**

1. בחרנו לעבוד בשיטת ה- Descriptive לאורך מרבית ביצוע ה-KDD ובמהלך הפרויקט כולו.

ניסינו לנתח ולהפיק תובנות עסקיות על מאגר המידע (כמו למשל לאילו קבוצות ניתן לחלק את הלקוחות של החנות?) ומנגד ניסינו גם במקרה מסוים לבצע חיזוי המתבסס על מאגר המידע ולבצע פרדיקציה (מה הוא פוטנציאל ההכנסות של חנות ברשת?).

1. נעזר בטכניקות של כריית מידע בכדי לנתח את הנתונים ולקבל תשובות לשאלות העסקיות אותן שאלנו, להלן:  
   * **לאילו קבוצות ניתן לחלק את הלקוחות של החנות כדי לדעת איזה קהלים שונים רלוונטיים לקמפיינים פרסומיים לפי סוגים מוצרים שונים?**  
     לשאלה זו, נשתמש בשיטת Clustering אשר תסייע לנו לחלק את הלקוחות של החנות, וזאת בכדי לפרסם להם תכנים רלוונטים שיגרמו להם לרכוש יותר מהחנות.

למשל, בעזרת מפת חום ניתן יהיה לאתר קורלציה (בין ערכים מספריים כמו ממוצע רכישה לסוג מוצר, או מיקום הלקוח לסוג המוצר, ולגזור מכך את הערכים שנותנים את הקולרציה הטובה ביותר. לאחר מכן נכניס את הערכים לאלגוריתם של KNN כדי לקבוע את קבוצות הלקוחות ולאחר מכן ננתחם שוב כדי להבין מה מייחד אותם ובכך נתאים להם פרסומות.  
לדוגמא: מהאלגוריתם KNN התקבלו קבוצות של נתונים. מניתוח עלה כי לקבוצה מסוימת בריו דה ז'נרו ממוצע הרכישות הינו 85 ריאל ורוב המוצרים של לקוחות אלו הן מוצרי חשמל. לכן, נבקש לבצע קמפיין פרסומי ללקוחות אלו עם מוצרי חשמל כמו גאדג'טים שאינם יקרים יתר על המידה.

* + **מה הוא פוטנציאל ההכנסות בחנויות ברשת של סוג מוצר מסוים?**

נוכל להיעזר בשיטת רגרסיה לינארית כאשר עלינו לקחת בחשבון מספר פרמטרים במודל במטרה לזהות מה כמות ההכנסות הצפויה של חנות ברשת במידה וחנות נוספת תיפתח ותמכור סוג מוצר X.  
הפרמטרים שניתן לקחת בחשבון הינם סוג מוצרים, מיקום החנות, כמות לקוחות פוטנציאלים ועוד ובעזרת פרמטרים אלו ננסה לחזות את כמות ההכנסות.  
לדוגמא: נפתחת חנות חדשה באיזור מרכזי ואטרקטיבי למשל במחוז באהיה. נוכל לחלץ את ממוצע הרכישות באזור זה ואת כמות הלקוחות הפוטנציאילים הקיימים באזור. בנוסף נבין איזה סוגי מוצרים היא תמכור ומה המחיר הממוצע אותו היא תדרוש עבור כל מוצר. את כל פרמטרים נבדוק אל מול משתנה המטרה שהוא כמות ההכנסות של מוצר ספציפי ולאחר מכן המשתנים שמציגים קורלציה נכניס אותם למודל ובכך ננסה לחזות את ההכנסות של חנות זו בסוגי מוצרים מסוימים.

למשל, נניח כי התגלה כי ממוצע הרכישות באזור, מחיר ממוצע, וכמות מוצרים משלימים התגלו כקורלטיבים לפוטנציאל ההכנסות אז נכניס למשוואה עם הרגרסיה הלינארית עם המשקלים שחושבו ונקבל:   
 ממוצע רכישות = 100 (משקל במשוואה 77.69)

מחיר ממוצע לרכישה = 50 (משקל במשוואה 11)  
 כמות מוצרים משלימים הנמכרים בחנות = 7 (משקל במשוואה 230)  
   
7937450 = 100\*77.69 + 50\*11 + 1000\*230 → פוטנציאל ההכנסות של מוצרים מסוג X בחנות החדשה.

1. מדד הדמיון אשר מתאים ביותר עבור ה-DW שבנינו הינו sorensen dice.

מדד זה הוא הרלוונטי ביותר משום שהוא תואם עבור ערכים נומריים וקטגוריאליים יחדיו, כפי שאנו מציגים ב-DW ובתהליך ה-KDD ובנוסף הוא ממדל את הדמיון וחוסר הדמיון בין קבוצות שונות ויכול לסייע במענה על השאלה לאילו קבוצות ניתן לחלק את הלקוחות של החנות.

1. ההשערות הן להלן:
   * עבור השאלה: **לאילו קבוצות ניתן לחלק את הלקוחות של החנות כדי לדעת איזה קהלים שונים רלוונטיים לקמפיינים פרסומיים לפי סוגים מוצרים שונים?**

H0: Customers’ Consumption habits **aren’t** divided by geographical residential place

H1: Customers’ Consumption habits **mainly** divided by geographical residential place

נקבל את H1 במידה ובמפת החום/ בדיקת הקורולציה נגלה כי ישנה קורלציה של 0.7 ומעלה בין מיקום גיאוגרפי וסוג מוצר. נקבל את השערת האפס במידה והקורלציה תהיינה נמוכה מ0.3, אחרת לא נוכל לקבוע.

* + עבור השאלה: **מה הוא פוטנציאל ההכנסות בחנויות ברשת של סוג מוצר מסוים?**

H0: The potential revenue of an electronic product in Bahaia county is **higher** than 500K a month per shop.

H1: The potential revenue of an electronic product in Bahaia county is **less (or equal)** than 500K a month per shop.

נקבל את HO במידה ובריגרסיה הלינארית אשר נבצע נגיע לניבוי רווח פוטנציאלי הגבוהה מ-500,000 לחודש. כל מספר הנמוך מצפי הכנסות זה יוביל לדחות את השערת ה0 ולקבל את H1.

**חלק 3**

שאלת המחקר: **מה הוא פוטנציאל ההכנסות בחנויות ברשת של סוג מוצר מסוים?**

1. SELECT seller\_id, product\_id, SUM(amount\_of\_sales) AS "**sum of orders per store and product**"

FROM FactSales AS A,

FactCustomerPurchase AS B

WHERE A.customer\_id=B.customer\_id

Group by seller\_id,product\_id

1. SELECT product\_id,count(distinct seller\_id) AS "**amount of stores per product**"

FROM FactSales AS A,

FactCustomerPurchase AS B

WHERE A.customer\_id=B.customer\_id

Group by product\_id

ORDER BY count(distinct seller\_id) DESC

1. SELECT seller\_id, A.order\_id , product\_id, AVG(total\_spend) OVER (PARTITION BY product\_id) AS "**AVG cost per product**"

FROM FactSales AS A,

FactCustomerPurchase AS B

WHERE A.customer\_id=B.customer\_id

שאלת מחקר: **לאילו קבוצות ניתן לחלק את הלקוחות של החנות כדי לדעת איזה קהלים שונים רלוונטיים לקמפיינים פרסומיים לפי סוגים מוצרים שונים?**

1. SELECT customer\_id,order\_item\_id , SUM(amount\_of\_itmes\_purchase) OVER (PARTITION BY customer\_id) AS "**amount\_of\_itmes\_purchase per customer**"

FROM FactCustomerPurchase

1. SELECT seller\_id,A.order\_id , product\_id, SUM(amount\_of\_sales) OVER (PARTITION BY A.customer\_id) AS "**amount of orders per customer**"

FROM FactSales AS A,

FactCustomerPurchase AS B

WHERE A.customer\_id=B.customer\_id

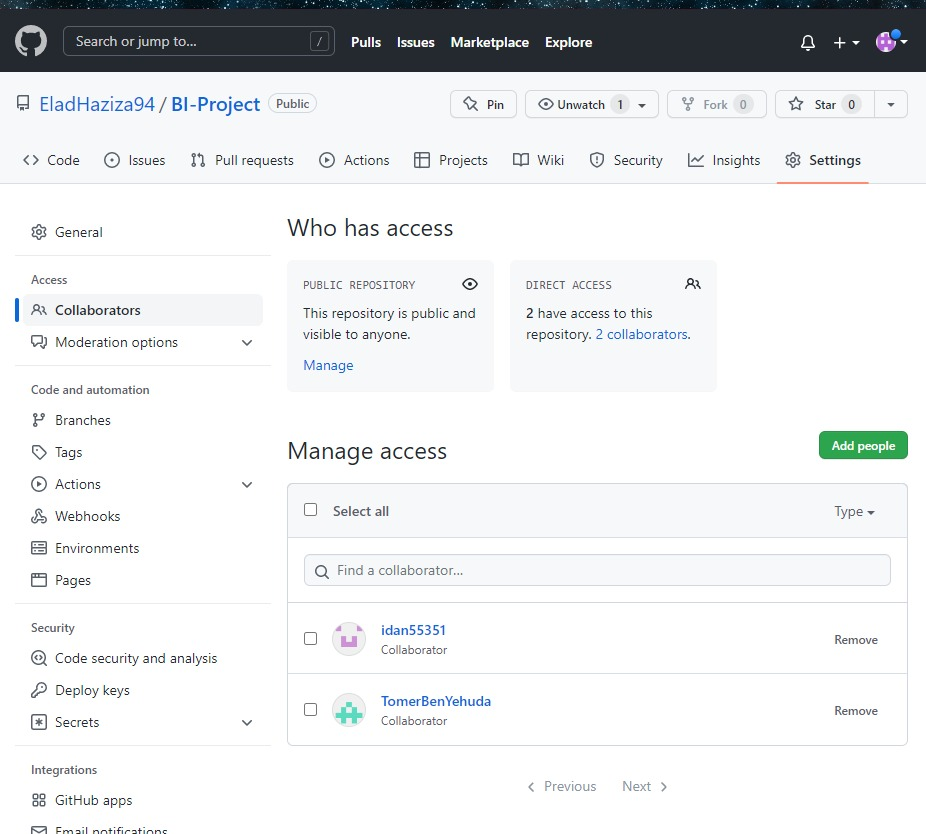
Group by product\_id

1. SELECT customer\_id , AVG(total\_spend) OVER (PARTITION BY customer\_id) AS "**AVG total\_spend per customer**"

FROM FactCustomerPurchase

**חלק 4**

בקישור [זה](https://github.com/EladHaziza94/BI-Project), ניתן לראות את הפרויקט שלנו בGITHUB.

****