ניהול צריכה באמצעות סריקת חשבוניות קבלה

עלית בן חמו- 206851081
נופר גרשוני- 322773680
למנחה- גב' קרן סגל

רקע

המטרה שלנו היא זיהוי אי- התאמה באמצעות סריקת חשבוניות קבלה, שיאפשר למשתמשים לקבל תובנות עסקיות ממוקדות באמצעות סריקת קבלות מחנויות שונות, תוך ניתוח נתונים על רכישות והוצאות. נתמקד במעקב אחר הוצאות המשתמשים, זיהוי אי-התאמות בכמויות ברשימות החשבוניות, ונציג את המידע בצורה ברורה ומסודרת. המערכת תספק למשתמשים כלי ניתוח שיאפשרו להם לעקוב אחר השינויים בהוצאות לאורך זמן, לזהות חריגות ולבצע השוואות בין רכישות שונות. בנוסף, נאפשר למשתמשים לזהות דפוסי רכישה ולספק התרעות על אי-התאמות, כמו במקרה שבו קיימת חוסר התאמה בין כמות הרכישות בפועל לבין הרשום בחשבונית.

תכנון הפרויקט

שלב 1- רעיון

הפרויקט מתמקד בזיהוי חריגות ברכישות של לקוחות על בסיס חשבוניות קבלה והיסטוריית הרכישות שלהם. בשלב זה נקבעה מטרת הניתוח – להבין את דפוסי הקנייה של כל לקוח ולזהות רכישות שהכמות שלהן חורגת מהכמויות הצפויות על פי דפוסי הרכישה הקודמים.

שלב 2- תכנון

הנתונים נאספו מהחשבוניות, ונוצרו מהם מאפיינים סטטיסטיים כמו ממוצע, סטיית תקן, טווח ומספר רכישות לפי לקוח ומוצר. העיבוד כלל נרמול נתונים מספריים באמצעות Standard Scaler והמרת ערכים קטגוריאליים ב־One Hot Encoding.

שלב 3- הביצוע

בשלב זה הופעלו מודלי חיזוי על מערך הנתונים שעבר עיבוד מקדים, במטרה לאתר חריגות בכמויות הרכישה. החיזוי בוצע הן ברמת הלקוח והן ברמת המוצר, והמודלים אפשרו זיהוי של דפוסי רכישה חריגים, לקוחות עם התנהגות יוצאת דופן ומוצרים עם שיעור חריגות. תוצאות המודלים הוצגו בטבלאות לפי מדדי דיוק שונים. בנוסף, הוצגו גרפים אינטראקטיביים אשר יקלו על חוויית המשתמשים.

הערכה ראשונית וטכניקות

סביבת פיתוח עיקרית

בבחירה ראשונית בחרנו בכלי יעיל למדעי הנתונים , בחרנו ב Python , כפלטפורמה העיקרית שלנו. פייתון בולט בזכות הרבגוניות והספריות הנרחבות ,מה שהופך אותו לבחירה אידיאלית עבור הפרויקט שלנו.

ספריות נלוות

שפת התכנות של פייתון , לרבות ספריות עוצמתיות כמו , pandas scikit-learn Numpy , מספקת לנו ערכת כלים מקיפה לכל שלב במדעי הנתונים שלנו.

איסוף נתונים

באמצעות סריקת חשבוניות קבלה וניתוח קפדני, הניתוח המעמיק שביצענו מאפשר הבנה רחבה ומעמיקה של דפוסי הקנייה של המשתמשים, ותורם לקבלת החלטות מושכלות, תכנון אסטרטגי ואופטימיזציה של הנתונים.

תכונות מובילות

התכונות המובילות שעליהן מתבסס הפרויקט הן כמות המוצרים והיסטוריית הרכישות של הלקוחות. תכונות אלו מאפשרות לבצע ניתוח מעמיק של דפוסי הקנייה, במטרה לזהות אי-התאמות בכמויות.

איסוף נתונים

בפרויקט נאספו כ-540 חשבוניות רכישה מ-84 לקוחות שביצעו קניות בחנויות שונות, תוך התמקדות בכמויות המוצרים והיסטוריית הרכישות של כל לקוח. הנתונים הועברו לפורמט טבלאי ועברו תהליך של עיבוד והשלמת נתונים, ובמסגרתו הורחבו הנתונים המקוריים לכדי כ-1,000 חשבוניות באמצעות הגרלה אקראית של החשבוניות הקיימות, על מנת ליצור בסיס מידע רחב ומגוון יותר. לאחר מכן בוצע ניתוח סטטיסטי שמטרתו לנתח דפוסי קנייה לפי חשבוניות הקבלה.

שילוב זה מאפשר להבין לעומק את התנהגות הקנייה של המשתמשים ולבנות מודלים לחיזוי ואיתור מוקדם של אי-התאמות.

הכנת מודלים ומידול

- בחירת משתנים רלוונטיים – בפרויקט נעשה שימוש במשתנים כמו כמות בכל מוצר בחשבונית, שם המוצר, היסטוריית רכישות קודמות של כל לקוח, תדירות הקנייה שלו וסך כל הרכישות, כדי לזהות דפוסי רכישה ולהשוות אותם לכמויות שנרשמו בחשבוניות הנוכחיות.
- איכות הנתונים – הנתונים נאספו באמצעות סריקת חשבוניות קבלה שקובצו לקבצי pdf ונותחו כדי לוודא שהם מלאים, מדויקים וללא חוסרים, כך שניתן יהיה לזהות חריגות בכמויות באופן אמין.
- הבנת הבעיה – מטרת הפרויקט היא לאתר אי-התאמות בכמויות המוצרים בחשבוניות, ביחס לדפוסי הרכישה האישיים של כל לקוח, וזיהוי חריגות סטטיסטיות על בסיס ממוצע וסטיית תקן.
- הערכת המודל – הערכת ביצועים של המודלים מתבצע באמצעות מדדים סטטיסטיים (Precision, Recall, F1, Accuracy) כדי לבחון את יכולת החיזוי של החריגות ולהשוות את ביצועי המודלים השונים על קבוצת הנתונים.

בחירת טכניקות המודל

בבפרויקט שלנו נעשה שימוש בטכניקות מידול מתקדמות לצורך ניתוח מדויק של דפוסי רכישה וזיהוי אי-התאמות בכמויות. תהליך המידול התבסס על מספר שלבים עיקריים:

1. עיבוד נתונים:

בשלב זה בוצעה בדיקה והסרה של כפילויות, על מנת לוודא שכל עסקה נלקחת בחשבון פעם אחת בלבד. בנוסף, מוצרים דומים אוחדו כדי לייצר תמונה מדויקת של דפוסי הצריכה. לאחר חלוקת הנתונים לסט אימון וסט בדיקה, בוצע איזון קטגוריות באמצעות טכניקת SMOTE, שנועדה להתמודד עם חוסר איזון בין נתונים רגילים לנתוני חריגות ולשפר את ביצועי המודלים.

2. איכות הנתונים:

כדי להבטיח אמינות, הנתונים הועברו לפורמט טבלאי מסודר ועברו בדיקות תקינות. תהליך זה אפשר המשך עבודה עקבית עם הנתונים ושיפור דיוק החיזוי.

3. יצירת מאפיינים:

נבנו מאפיינים חדשים כמו ממוצע, סטיית תקן, טווח, טווח תחתון וטווח עליון עבור כל לקוח. הממוצע מצביע על הכמות הצפויה לרכישה, וסטיית התקן מצביעה על יציבות הרכישות. שילוב מאפיינים אלה אפשר זיהוי חריגות בצורה אמינה יותר.

4. התאמת פורמט נתונים:

בוצע קידוד One-Hot-Encoding לנתונים קטגוריאליים ונרמול לנתונים מספריים באמצעות StandardScaler, על מנת להתאים את הנתונים לדרישות המודלים השונים ולשפר את תפקודם.

סוגי המודלים

מכונת וקטורים תומכת (SVM): מודל סיווג בינארי שנועד לחזות האם קיימת אי-התאמה בכמות הרכישה. קודם סומנו חריגות בעזרת חישובים סטטיסטיים – כגון ממוצע וסטיית תקן – ליצירת משתנה מטרה בינארי. המודל משתמש בגרעין RBF להפרדה לא ליניארית בין קבוצות, ומתבסס על מאפיינים מספריים כמו כמות, מחיר וסטטיסטיקות נוספות, כדי לבצע תחזיות מדויקות של חריגות.

- **עצי החלטה (Decision Trees):** מודל פשוט וברור לסיווג מבוסס סדרת תנאים לוגיים, המסייע בזיהוי גורמים מרכזיים המשפיעים על אי-ההתאמות.
- **Random Forest:** שיטה של עצים רבים, המשפרת את הדיוק ומפחיתה סיכון של התאמת יתר, ומתמודדת היטב עם נתונים מורכבים ומגוונים.
- **אשכולות K-Means :** טכניקת אשכולות ללא פיקוח שמטרתה לקבץ תצפיות דומות יחדיו. בפרויקט השתמשנו בה כדי לזהות קבוצות רכישות רגילות וחריגות, מה שנתן תובנה נוספת על דפוסי רכישה לא שגרתיים.

לסיכום, המודל שנבחר לפרויקט הוא SVM, בזכות יכולתו להתמודד עם סיווג בינארי במרחבים לא ליניאריים והעניק תחזיות מדויקות ואמינות לגבי אי-התאמות בכמויות הרכישה.

מדדים מרכזיים:

Recall:
מדד זה מראה איזה חלק מהחריגות האמיתיות המודל הצליח לזהות.

בפרויקט שלנו זהו המדד החשוב ביותר, כי המטרה היא לא לפספס רכישות שבהן הכמות חריגה מהמצופה.

Precision:
מודד מתוך כל המקרים שהמודל סיווג כחריגים – כמה מהם אכן היו חריגים.

F1 Score:
מדד שמחבר בין Precision ל־Recall.
מספק תמונה כוללת של ביצועי המודל: עד כמה המודל מצליח לאתר חריגות אמיתיות ובמקביל להימנע מסימון יתר של רכישות רגילות כחריגות.

Accuracy:
אחוז התחזיות הנכונות מתוך כלל המקרים.

Confusion Matrix (טבלת בלבול):
מאפשרת להבין אילו טעויות המודל עושה – כמה מקרים אמיתיים הוא פספס, וכמה חזה בטעות כחיוביים.

הערכת תוצאות

במהלך תהליך ההערכה הופעלו שלושה מודלים חיזויים – Decision Tree, Random Forest ו־SVM – במטרה לאתר אי־התאמות בכמות המוצרים. הנתונים חולקו לסט אימון וסט בדיקה ביחס של 70% לאימון ו־30% לבדיקה, והשוואת התוצאות ביניהם אפשרה להסיק מסקנות לגבי רמת היעילות של כל אחד מהמודלים.
מסקנה סופית, ניתן לראות שבמודל SVM ה- recall הכי גבוה שהוא הנתון הקריטי לזיהוי אי התאמה.

מודל	Precision	Recall	F1 Score	Accuracy	Silhouette
מכונת וקטורים תומכת(SVM)	0.947	0.989	0.968	0.996	-
עצי החלטה	0.889	0.978	0.932	0.991	-
Radom Forest	0.607	0.956	0.742	0.960	-
k-means	-	-	-	-	0.74

פירוט על המודל הנבחר:

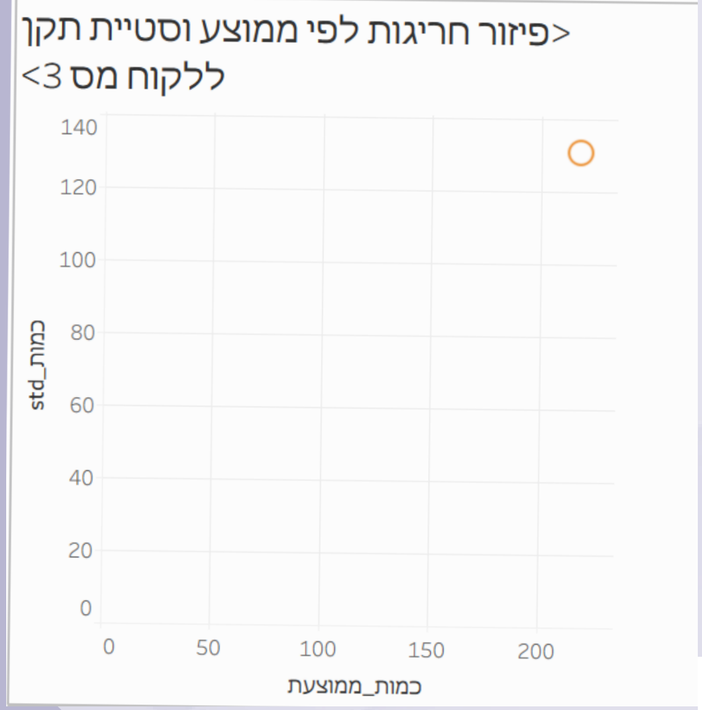
בפרויקט שלנו, מודל SVM (Support Vector Machine) הציג את הביצועים הטובים ביותר בהערכת חריגות בכמויות.
המדד המרכזי לפרויקט, Recall, הגיע ל־0.989, מה שמעיד על כך שהמודל הצליח לזהות כמעט את כל הרכישות שבהן הכמות חרגה מהממוצע האישי של הלקוחות. בנוסף, F1 Score של 0.968 משקף איזון טוב בין היכולת לזהות חריגות אמיתיות (Recall) לבין היכולת למנוע סימון יתר של רכישות רגילות כחריגות (Precision). Accuracy של 0.996 מציין שהמודל חזה נכון את רוב התחזיות הכלליות, אך בפרויקט שלנו הוא פחות מהווה את הקריטריון המרכזי. סך הכל, הנתונים מראים ש-SVM הוא כלי חזק ומדויק לאיתור מוקדם של חריגות בכמויות, המספק תובנות אמינות על דפוסי רכישה חריגים.

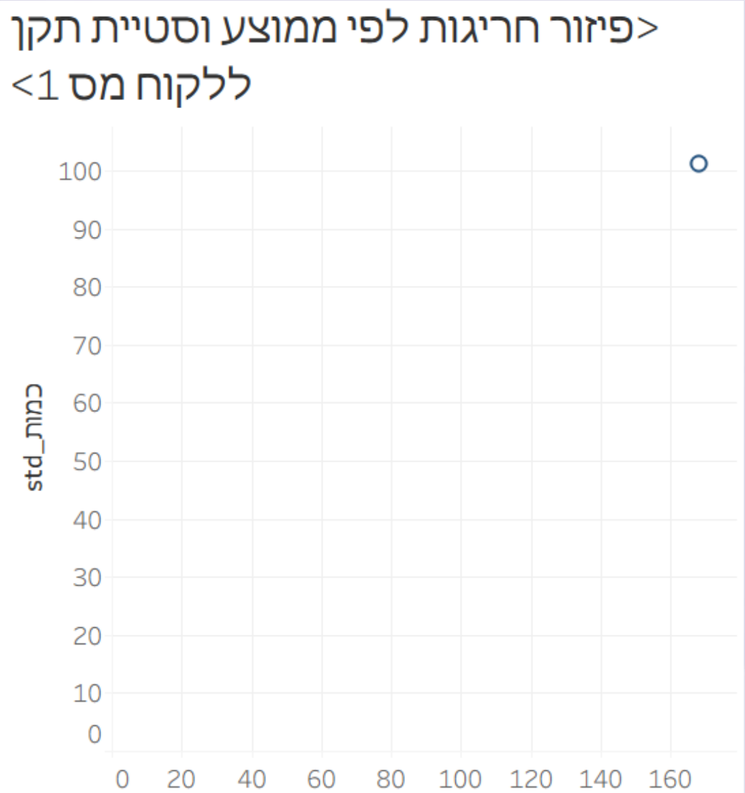
הטמעת המודל

בפרויקט התמקדנו בזיהוי אי־התאמות בכמויות הרכישה של המשתמשים, באמצעות ניתוח היסטוריית דפוסי הקנייה האישיים שלהם, המבוסס על חישוב ממוצע וסטיית תקן עבור כל לקוח. המטרה הייתה לאתר מצבים שבהם הכמות שנרכשה שונה באופן מובהק מהכמות הצפויה לפי ההיסטוריה האישית. לצורך הצגת הממצאים בצורה ברורה ונגישה, יצרנו לוח מחוונים אינטראקטיבי ב-Tableau, הכולל גרפים ויזואליים המדגישים את נקודות החריגה ומסייעים להבין את סיבותיהן האפשריות. הצגת הנתונים בצורה גרפית אפשרה לזהות בקלות מצבים של חריגה, להבחין בין רכישות רגילות לחריגות, ולהפיק תובנות משמעותיות על דפוסי הקנייה של המשתמשים.

מסכי לוח מחוונים

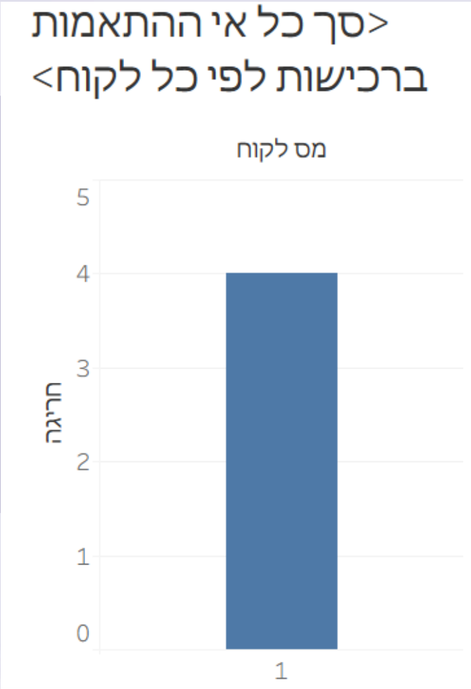
תרשים הפיזור מציג נתונים עבור כל לקוח בנפרד. בציר האופקי מוצגת הכמות הממוצעת שרכש, ובציר האנכי סטיית התקן של הכמויות – המייצגת את פיזור הרכישות שלו. גודל הנקודה משקף את מספר החריגות שנמצאו אצל הלקוח, ומאפשר לזהות את דפוסי הרכישה והחריגות באופן פרטני. לדוגמה, בגרף אחד נראית נקודה גדולה עם כמות ממוצעת של כ-200 וסטיית תקן של כ-130, מה שמעיד על לקוח שקונה הרבה ובאופן משתנה ויש לו הרבה חריגות. בגרף אחר נראית נקודה קטנה עם כמות ממוצעת של כ-170 וסטיית תקן של כ-100, מה שמעיד על לקוח שקונה הרבה אך באופן משתנה מועט יחסית, עם מספר קטן של חריגות.

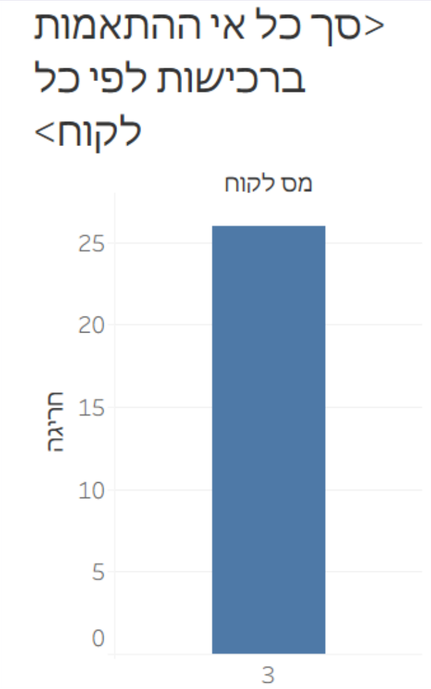




מסכי לוח מחוונים

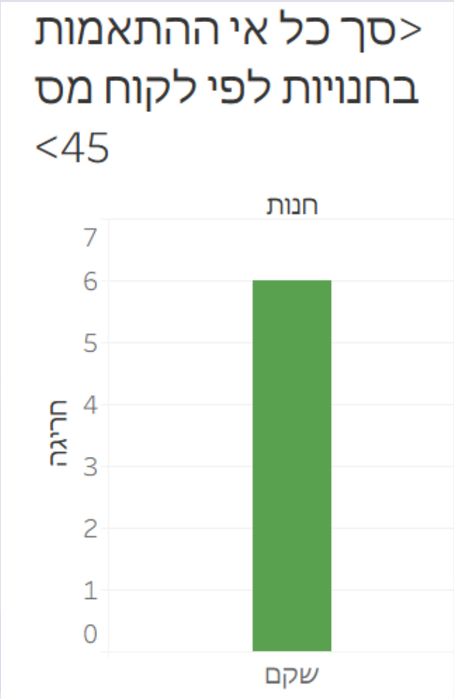
התרשים מציג את כמות החריגות של כל לקוח בנפרד, כפי שחושבו על בסיס ממוצע וסטיית התקן של הכמויות שרכש. כל עמודה מייצגת רכישה שבה הכמות חרגה מהטווח הצפוי (ממוצע ± שתי סטיות תקן). התרשים מאפשר לראות כמה חריגות נמצאו, ולהעריך את חוסר היציבות בכמויות שרכש הלקוח. לדוגמה, בגרף של לקוח מס׳ 1 מספר החריגות בכל הרכישות שלו עומד על 4, ואילו בגרף של לקוח מס׳ 3 מספר החריגות גבוה יותר ועומד על 25.

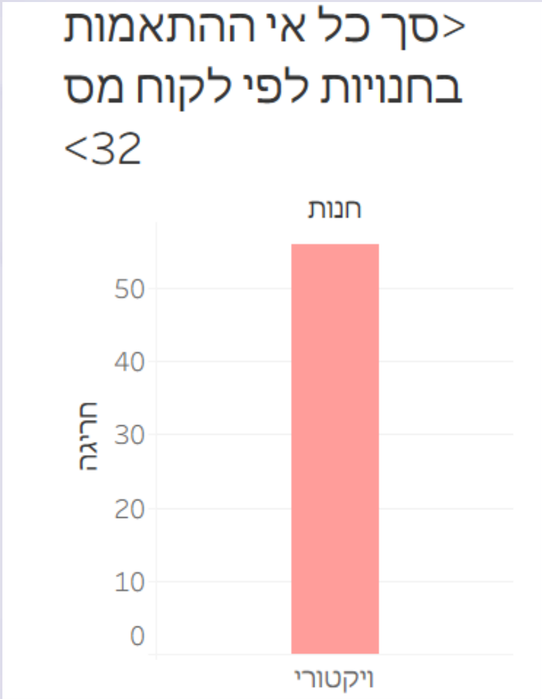




מסכי לוח מחוונים

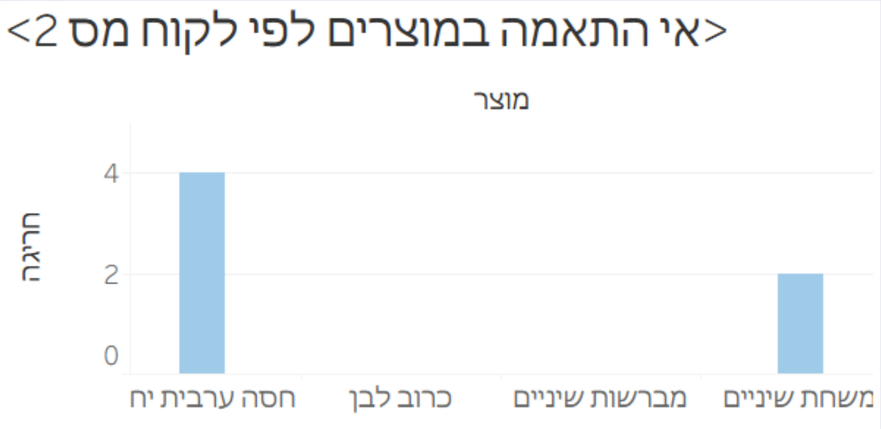
תרשים העמודות המוערם מציג את כמות החריגות שחושבו לפי ניתוח סטטיסטי של ממוצע וסטיית תקן עבור כל לקוח בנפרד, בחנויות שבהן רכש. כל עמודה מייצגת חנות אחת, וגובהה מראה את מספר החריגות באותה חנות. במקרה זה, הלקוח ביצע רכישות רק בחנות אחת, ולכן מוצגת עמודה יחידה בגרף. לדוגמה, בגרף של לקוח מס׳ 32 קיימות כ־55 חריגות בחנות ויקטורי, ואילו בגרף של לקוח מס׳ 45 קיימות כ־6 חריגות בחנות שקם.





מסכי לוח מחוונים

תרשים העמודות המוערם מציג את כמות החריגות של כל לקוח בנפרד, שחושבו לפי ניתוח סטטיסטי של ממוצע וסטיית תקן עבור כל המוצרים שרכש. כל עמודה מייצגת מוצר מסוים, וגובהה מציין את מספר החריגות של הלקוח במוצר זה. התרשים מאפשר לראות הן את סך החריגות של הלקוח בכל המוצרים, והן את הפירוט – באילו מוצרים התרחשו החריגות וכמה חריגות היו בכל אחד מהם.



סיכום

הפרויקט נועד לספק כלי יעיל ואמין לאיתור וניתוח חריגות בכמויות הרכישה המופיעות בחשבוניות קבלה, באמצעות השוואה בין הכמות שנרשמה לבין הכמות הצפויה על פי דפוסי הרכישה של הלקוח. החשיבות העסקית נובעת מהיכולת לאתר חריגות חוזרות בכמויות הרכישה, לזהות לקוחות או מוצרים בעלי דפוסי סטייה עקבי, ולספק תובנות שיכולות לסייע בצמצום טעויות ושיפור חווית הצריכה.

במהלך עיבוד הנתונים, חישוב סטטיסטי של ממוצע וסטיית תקן עבור כל לקוח ומוצר שימש ליצירת עמודת המטרה של החריגות, שהצביעה האם רכישה מסוימת חרגה מהכמות הצפויה (ממוצע ± שתי סטיות תקן). עמודת המטרה הזו שימשה להכנת הנתונים למודלים חישוביים, במטרה לזהות חריגות ברכישות. הנתונים וממצאי החישוב הוצגו בצורה ויזואלית ברורה באמצעות גרפים ומדדים מרכזיים, מה שאיפשר לזהות נקודות בעייתיות ולהשוות בין גורמים שונים, ולספק ראייה מקיפה ומדויקת של דפוסי הצריכה.

THANK YOU