פרויקט גמר

דוח הבנת הנתונים

ניהול צריכה באמצעות סריקת חשבוניות קבלה

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

תוכן עניינים:

דוח הבנת הנתונים

1.איסוף נתונים:

1.1 מקורות הנתונים:

הפרויקט עוסק בניתוח חשבוניות קבלה המתקבלות לאחר רכישה, מתמקד בזיהוי וחיזוי אי התאמות בכמויות המוצרים שנרכשו ביחס לנתוני היסטוריית הרכישות של הלקוחות.

מקורות הנתונים מבוססים על חשבוניות קבלה ממשתמשים שביצעו קניות בחנויות שונות. אספנו 540 חשבוניות, אשר המידע המופיע בהן, כגון מק"ט, פריטים שנרכשו, מחירים, מס' חשבונית וכמויות מכל מוצר. הועבר לפורמט טבלאי באמצעות Excel. תהליך זה מאפשר לנו להשתמש בנתונים קיימים ואמיתיים לצורך ניתוח וחיזוי אי התאמות בכמויות המופיעות בחשבוניות. בנוסף, השלמנו את מספר החשבוניות ל-1000 באמצעות הגרלת נתונים מתוך החשבוניות הקיימות. בחירת השלמת החשבוניות נבעה מכך, שכמות נתונים גדולה מאפשרת לזהות את מגוון הדפוסים השונים בצורה טובה יותר ומספקת תמונה רחבה ומדויקת יותר על אפשרויות אי ההתאמה. בהמשך לתהליך הוספנו את עמודת מספר לקוח המכילה בתוכה 84 לקוחות, על מנת שנוכל לחזות את אי ההתאמה של הכמות לפי היסטוריית הרכישות של כל לקוח.

1.2 בדיקה ראשונית של הנתונים:

<u>מאפיינים מבטיחים:</u>

המאפיין המרכזי בנתונים שלנו שמספק ערך משמעותי לחיזוי אי התאמות בכמויות הוא הכמות של כל מוצר בחשבונית. מאפיין זה מצביע על מספר הפריטים שנרכשו, וזהו המידע הקריטי לזיהוי אי התאמות.

בנתונים שלנו כל העמודות שנמצאות בנתונים הן רלוונטיות לצורך חיזוי אי ההתאמה בכמויות מלבד עמודת חנות.

מאפיינים רלוונטיים:

מק"ט- מספק זיהוי מדויק של המוצר, **שם המוצר-** מתאר את פרטי המוצר.

כמות- היא קריטית, כיוון שהיא מציינת את מספר הפריטים שנרכשו ומסייעת בזיהוי אי התאמות, **מחיר-** רלוונטי במקרים בהם נרצה לבדוק את ההתאמה בין המחיר לכמות שנרשמה. **מספר חשבונית-** רלוונטית, כיוון שהיא מאפשרת לבדוק את אי ההתאמה של הכמויות לכל חשבונית בנפרד. בנוסף, העמודה שהוספנו של מספר לקוח רלוונטית עבור התהליך של החיזוי בכך שהיא מאפשרת לנו לנתח את היסטוריית הרכישות של כל לקוח. לכן, כל העמודות בנתונים שלנו תורמות באופן משמעותי לתהליך חיזוי אי ההתאמה.

הסקת מסקנות:

כמות הנתונים שנאספה והושלמה, הכוללת 1000 חשבוניות, מספקת בסיס טוב להסקת מסקנות כלליות ולביצוע תחזיות מדויקות. עם כמות זו של נתונים, ניתן לזהות דפוסים וסטיות בכמויות, לבצע ניתוחים סטטיסטיים משמעותיים ולהבטיח שהתחזיות שנפיק יהיו אמינות. כמות הנתונים מאפשרת גם לאמן מודלים סטטיסטיים בצורה שתספק תחזיות מבוססות.

שיטות מידול:

לא קיימים מאפיינים רבים ביחס לשיטת המידול המיועדת. הנתונים שלנו כוללים בעיקר את המאפיינים הרלוונטיים, כגון המוצר, הכמות והמחיר, שהם חיוניים לצורך חיזוי אי התאמות בכמויות. אנחנו מתכננים להשתמש במודל רגרסיה ומודל עצי החלטה, שיכולים להתמודד בצורה טובה עם כמות הנתונים והמאפיינים הקיימים.

מקורות נתונים חיצוניים:

בנתונים שלנו, כל המידע מגיע מקבצי חשבוניות קבלה סרוקים, ולכן אין שילוב עם מקורות נתונים חיצוניים נוספים. עם זאת, כדי להבטיח דיוק ולמנוע בעיות מיזוג, הסרנו כפילויות ואחדנו בין המוצרים הדומים שנמצאים באותה חשבונית. תהליך זה חשוב מאוד, שכן הוא מבטיח שהנתונים לא יכילו חזרות מיותרות.

ערכים חסרים:

לא קיימים ערכים חסרים, כך שאין צורך להתמודד עם בעיה זו. כל השדות הרלוונטיים, כמו המוצר והכמות, מלאים באופן מלא בכל רשומה. זהו יתרון חשוב שמפשט את תהליך הניתוח ומקטין את הסיכון לטעויות הנובעות מערכים חסרים. במקרה כזה, אפשר להתמקד בנתונים עצמם ולבצע ניתוחים או תחזיות ללא הצורך בטיפול מיוחד בערכים חסרים.

2. תיאור הנתונים:

2.1 כמות הנתונים:

הכמות הזמינה לניתוח כוללת 7,424 תצפיות ו-6 מאפיינים.כל תצפית מייצגת נתון בנוגע למאפיינים שונים של המידע, כולל משתנים שקשורים לפרטי המוצר, מחירים, כמויות ועוד.

2.2 סוגי ערכים:

בנתונים שלנו קיימים מאפיינים מסוגים שונים:

מק"ט-מכיל ערכים מספריים שמזהים באופן ייחודי כל מוצר.

מוצר- מכיל ערכים קטגוריאליים (מחרוזות) המתארים את שם המוצר.

חנות- מכילה ערכים קטגוריאליים ומתארת את שם החנות שבה בוצעה הרכישה.

כמות- מכילה ערכים מספריים שלמים ומתארת את הכמות של כל מוצר שנרכשה.

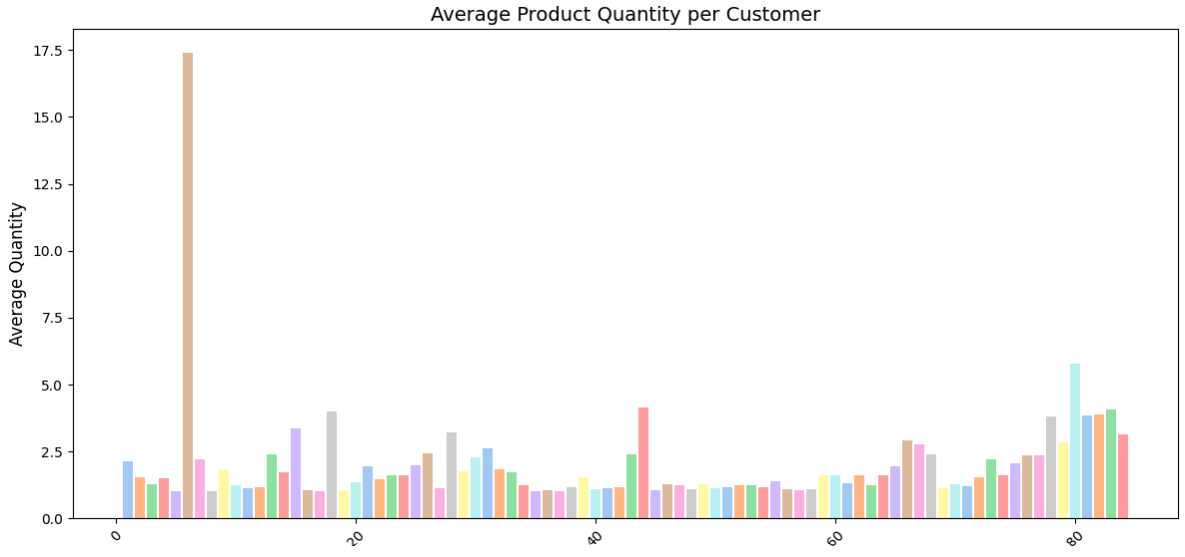
מחיר -מכיל ערכים מספריים ומציג את המחיר עבור כמות היחידות של המוצר.

מספר חשבונית- מכילה ערכים מספריים ומשמשת לזיהוי החשבונית שבה נרשמה העסקה.

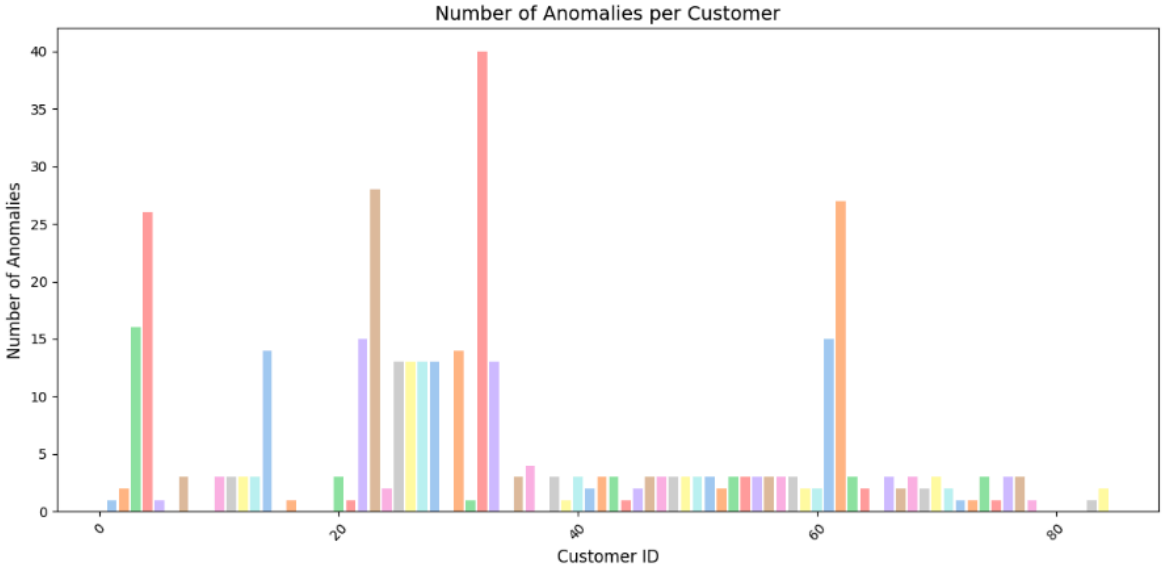
<u>שיטות קידוד:</u>

בשלב חישוב האי-התאמה אין שימוש בשיטות קידוד לערכים קטגוריאליים, שכן חישובי אי-התאמה מתבססים בעיקר על עמודת הכמות - באמצעות חישוב ממוצע הכמות של כל מוצר והשוואתה לרכישה מסוימת. לכן, אין צורך להמיר ערכים כמו שם המוצר או שם החנות לערכים מספריים לצורך ניתוח זה. עם זאת, בעת בניית המודלים החיזויים, כן נשתמש בשיטות מידול לקידוד ערכים קטגוריאליים, ובעיקר בשיטת One-Hot Encoding, על מנת לאפשר למודלים לנצל מידע זה כחלק מתהליך הלמידה והחיזוי, גם אם הוא אינו רלוונטי לחישוב האי-התאמה עצמו.

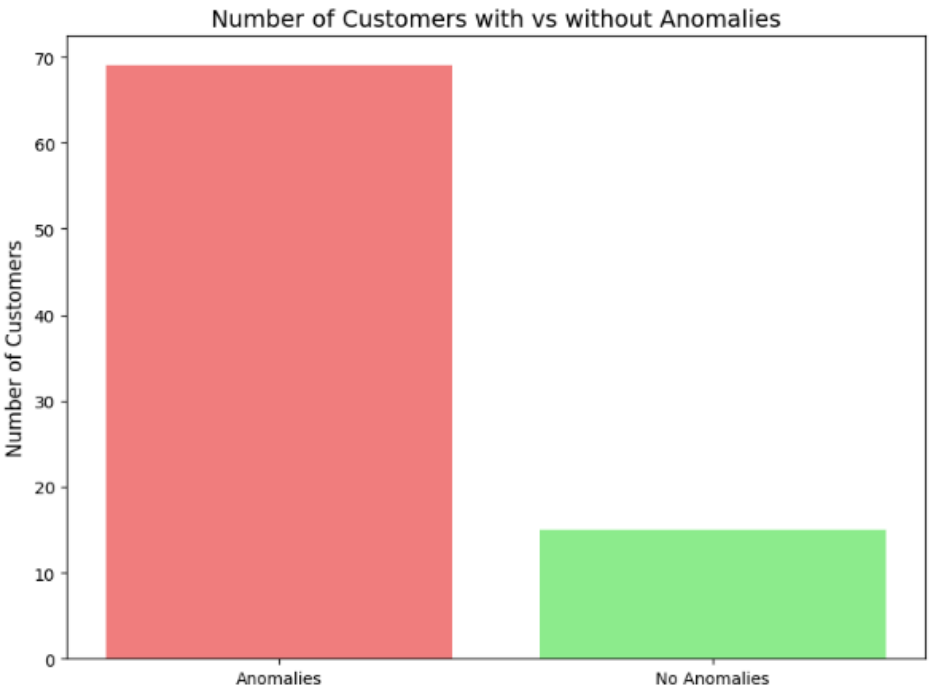
3. חקר נתונים:



הגרף מציג את הכמות הממוצעת של המוצרים לכל לקוח. ניתן לראות שלקוח מספר 6 קיימת כמות הממוצעת הגבוהה ביותר, מה שעשוי להעיד על כך שהוא רוכש מוצרים בכמויות גדולות יותר באופן קבוע בהשוואה ללקוחות אחרים.



הגרף מציג את מספר האי התאמות לכל לקוח. ניתן לראות שלרוב הלקוחות קיימים אי התאמות לפי סטיית התקן, כמות ממוצעת וגבולות תחתונים ועליונים ברכישות שלהם.



הגרף מציג את כמות הלקוחות עם אי-התאמה בכמות המוצרים לעומת הלקוחות שאין להם אי-התאמה. ניתן לראות כי קיימת כמות גדולה יותר של לקוחות עם אי-התאמות בכמות המוצרים.

ההשערות שגיבשנו לגבי הנתונים אושרו במהלך החקר, כאשר גילינו שאכן קיימות אי התאמות בכמויות המוצרים בהיסטוריית הרכישות. המאפיינים המרכזיים שמבטיחים להוות תרומה משמעותית להמשך הניתוח כוללים את הכמות של המוצרים, שם המוצר ומספר הלקוח. נתונים אלו מספקים בסיס חזק להמשך ניתוח ולחיזוי מדויק של אי ההתאמות בכמויות בהתאם להתנהגות הרכישה של כל לקוח. במהלך חקר הנתונים, גילינו מאפיינים חדשים הקשורים לאי-התאמות בכמויות המוצרים. חישבנו את הכמות הממוצעת ואת סטיית התקן של המוצרים עבור כל מספר לקוח, ובהתבסס על כך חישבנו את הגבולות התחתון והעליון באמצעות שתי סטיות תקן מהממוצע. גבולות אלו מאפשרים לזהות חריגות, שבהן הכמויות חרגו מטווח זה. באמצעות חישובים אלו בדקנו אם קיימת אי התאמה ברכישות, כלומר עברנו באמצעות הפייתון על כמות של כל מוצר מלקוח מסוים ובדקנו אם אכן קיים בטווח הזה, ואכן הצלחנו לזהות שקיימות אי התאמות ברכישות. בנוסף, גילינו שקיימים יותר לקוחות עם חריגה בכמות המוצרים. החקר לא רק חיזק את ההשערות הראשוניות שלנו אלא גם הדגיש את החשיבות של ניתוח ממוקד לקוח, שבו כל אי התאמה מזוהה ומנותחת בהקשר של רכישות קודמות. המטרות המקוריות של הפרויקט, המתמקדות בחיזוי אי התאמות בכמויות ושיפור הדיוק בזיהוי תבניות נתונים חריגות, נותרו ללא שינוי. עם זאת, החקר תרם לתובנה שיש לבצע ניתוחים מותאמים אישית לכל לקוח על בסיס היסטוריית הרכישות שלהם, כדי להבטיח שמודלי החיזוי יניבו תוצאות אמינות ומדויקות.

4. איכות הנתונים:

במהלך העבודה עם נתוני החשבוניות, זיהינו מספר בעיות איכות פוטנציאליות שיכולות להשפיע על תוצאות הניתוח. חשוב לציין כי לא היו נתונים חסרים, כלומר כל הערכים שנדרשו היו קיימים. עם זאת, גילינו כי היה צורך להוסיף נתונים חשובים שהיו חסרים, הכמות הממוצעת של המוצרים עבור כל אחד מהלקוחות, סטיית התקן וחישוב הגבולות התחתון והעליון, המידע הזה לא היה קיים בנתונים המקוריים, והיינו צריכים לחשב אותם באמצעות הפייתון, על מנת שנוכל לבצע בדיקה עבור כמויות המוצרים הרשומים בנתונים שהם אכן נמצאים בטווח, דבר שיביא לנו לתובנות מדויקות יותר בנוגע לאי ההתאמות בכמויות. כמו כן, נדרשו לבצע תיקונים בנתונים הקיימים, במיוחד בהקשר של כפילויות בשמות המוצרים. לעיתים קרובות שמות מוצרים דומים הופיעו מספר פעמים באותה החשבונית, דבר שגרם לבעיות בהתאמה ובחישוב. על מנת למנוע בעיות אלו, ביצענו איחוד בין המוצרים הדומים שהיו רשומים באותה חשבונית באופן ידני, בכך שמרנו על אחידות בנתונים. בנוסף, ביצענו בדיקות תקינות באמצעות הפייתון לכל העמודות בנתונים שלנו, תוך וידוא שהערכים בכל עמודה תואמים את הציפיות. לדוגמה, בעמודת 'מוצר' בדקנו שכל הערכים הם טקסטואליים ואינם מכילים ערכים מספריים. בנוסף, וידאנו שאין ערכים חסרים בעמודות השונות. כמו כן, בדקנו שאיחוד הנתונים אינו מכיל כפילויות של מוצרים באותה חשבונית, כדי להבטיח שהנתונים מדויקים ומלאים. בנוסף לכך, לא נתקלנו בבעיות מטא-נתונים פגומים. כל שדה נתון שהכנסנו או עבדנו איתו היה מוגדר בצורה ברורה, מדויקת ועקבית. לדוגמה, שדה "שם המוצר" היה ברור בתוכנו והכיל את שם המוצר בלבד, ללא נתונים נוספים או ערכים שגויים, כגון מספרים או תיאורים נוספים שאינם שייכים לשדה זה. כלומר, לא היו מצבים שבהם השם או ההגדרה של השדה "שם המוצר" לא תאמו את התוכן שהוזן בו, דבר שהבטיח עקביות מלאה בנתונים ומנע בלבול בתהליך הניתוח.