נתוני עזריאלי משאיות ובניו – דו"ח ראשוני

גיא לבבי

כחלק מיצירת סיכום ביצועים של נהגי החברה שלכם, בדו"ח זה מפורטות בעיות שנמצאו בנתונים הגולמיים שהתקבלו. ישנן בעיות מהותיות שדורשות שיח עמכם, אך גם בסוף מצורפות בעיות פחות מהותיות שנמצאו וטופלו ללא צורך במשוב מכם. פתרון הבעיות מהסוג הראשון קריטי ליצירת סיכום ביצועי עובדים מהימן.

# בעיות בנתונים שדורשות בירור

להלן בעיות בנתונים אשר למען פתרונן נדרשת התייעצות עמכם:

* ברשימות הנהגים, הן הוותיקים והחדשים, לעיתים המגדר חסר. האם לאגד את כל מקרים אלו תחת קטגוריה חדשה של מגדר לא ידוע (למשל, קטגוריית "unknown"), או לחלופין למשל להציב בכל מקרים אלו מגדר זכר, כיוון שהרוב המוחלט של הנהגים הם ממין זכר? אופציה נוספת תהיה להתעלם מרשומות אלו ולא להחשיבן בסיכום הסופי.
* ברשימות הנהגים, ישנם תאריכי לידה חסרים. כמו כן, לאחר מבט על הרשומות, לעיתים יש נהגים עם תאריך לידה "1.1.1900", אשר נראה כממלא מקום לתאריך לידה לא ידוע. כיוון שבסיכום נדרש גיל של כל נהג, כיצד לנהוג במקרים אלו?
* בטבלת התעריפים ישנן שתי שורות בעייתיות:
  + חברת חשמל ("electricity") – חסר תעריף מוגבר ("extra milage"). כמו כן הבונוס לסופ"שים בשורה זו מופרז (עומד על 70,000( בהשוואה לשאר התעריפים.
  + ביטוח לאומי – נראה שכל התעריפים חסרים.
* נראה שבתיעוד הנסיעות מופיעות שלוש חברות נוספות שלא מוזכרות בטבלת התעריפים:
  + hot
  + yes
  + aminach

ולהפך, יש שלוש חברות בטבלת התעריפים שלא מופיעות בתיעוד הנסיעות:

* + telecommunication\_ltd
  + dbs
  + mizranei\_kfar\_saba

נראה שמדובר באותן חברות תחת שמות שונים, משום ששמה המלא של Hot הוא Hot Telecommunication Systems Ltd ובעבר Yes נקראה D.B.S. Satellite Services. מאלימנציה והגיון בריא (עמינח גם היא חברת מזרנים) ניתן לזווג את שתי החברות הנותרות. האם התאמה זו מקובלת עליכם?

* בתיעודי הנסיעות יש מקרים בהם נחצים 3000, 6000, ואף 9000 קילומטרים בטווח של יום-יומיים, מה שנראה בלתי הגיוני. האם מדובר בטעות ביחידות מידה, או שגיאה מסוג אחר? אם אכן מדובר בשגיאה, כיצד תרצו שנטפל במקרים חריגים אלו?
* סופשבוע מוגדר כימי שישי ושבת?
* לילה מוגדר כ-10 בלילה עד 6 בבוקר?
* היות שידועים רק סך הקילומטרים לכל נסיעה, האם חישוב התעריפים המוגברים יעשה לפי החלק היחסי מזמן הטיול הכולל? למשל נהג נוסע מ-10 בלילה בשבת עד 8 בבוקר בראשון. 8 שעות סופש, 6 מתוכן לילה, שעתיים אחרונות בתעריף ללא בונוס. אז 80 אחוז מבונוס הלילה המלא ו-60 אחוז מבונוס הלילה המלא.
* האם תעריף מוגבר הוא בנוסף לתעריף הבסיסי או שהוא מחליף אותו? נשאל כי לעיתים תעריף זה לבדו נמוך מהתעריף הבסיסי.
* כאחוז מרשומות הנסיעות חסרות תאריך התחלה או סיום. כאחוז מתוך אלו חסרים את שניהם. כיצד לשערך את זמני הנסיעה אז? (למען חישוב בונוסי הסופ"ש והלילה)

# בעיות בנתונים שאינן דורשות בירור

כאן מצוינות בעיות שנמצאו בנתונים אך הפתרון להן היה יחסית פשוט ואין צורך לבצע בירור מולכם. בכל זאת חשוב לנו להעלות אותן בפניכם, למשל למען תהליכי ביקורת פנימית.

* בחלק מהטבלאות שסופקו על ידיכם (למשל רשימות הנהגים, הן הוותיקים והחדשים, ותיעודי הנסיעות) ישנה עמודה מיותרת ששמה "Unnamed: 0". הסרנו אותה כחלק מניקוי הנתונים. כמו כן שורה ראשונה ריקה מיותרת בטבלת התעריפים.
* ברשימות הנהגים, תאריכי הלידה בפורמט לא עקבי. רוב הרשומות בפורמט dd/mm/yyyy, אך לעיתים רשום dd.mm.yyyy, או month name day, year (למשל September 5, 1977). טופל על ידי מעבר לפורמט אחיד. כנ"ל לגבי פורמט לא עקבי של מגדר.
* בעוד שברשימות הנהגים החדשים הותק נראה שהותק נמדד בימים, ברשימות הנהגים הוותיקים הותק נמדד בשנים. כיוון שהסיכום נדרש להכיל את הותק בשנים, נמיר את הותק של החדשים לשנים.