Ex2 – Submission practical part

<u>סעיף א':</u> איתור בעיות בקבצי הנתונים.

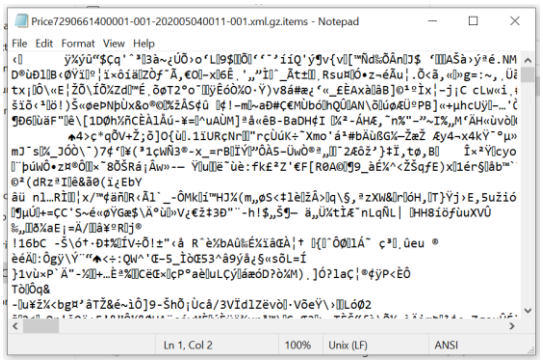
1. **קבצים מושחתים** (Not approachable):

- תיאור הבעיה: קבצים מסוימים מוגדרים מושחתים ועל כן אין אפשרות לעבוד איתם.
- דוגמות: בפרט הקבצים בתיקיית "freshmarket", ובסה"כ כ-20 קבצים שאין אפשרות לפתוח אותם (UNZIP).

2. **קבצים לא קריאים** (Bad Encoding):

- תיאור הבעיה: ישנם קבצים שניתן לפתוח אותם (UNZIP), אבל הקידוד שלהן לא בוצע כהלכה והם אינם ניתנים לעיבוד.
- דוגמות: צפינו ב-1415 מקרים בתיקייה hashook\2020-05-04 , כאשר לשמות הקבצים נוספו .item.

דוגמא לקובץ שכזה –



3. **פער בקבצי נתונים** (Missing Data files):

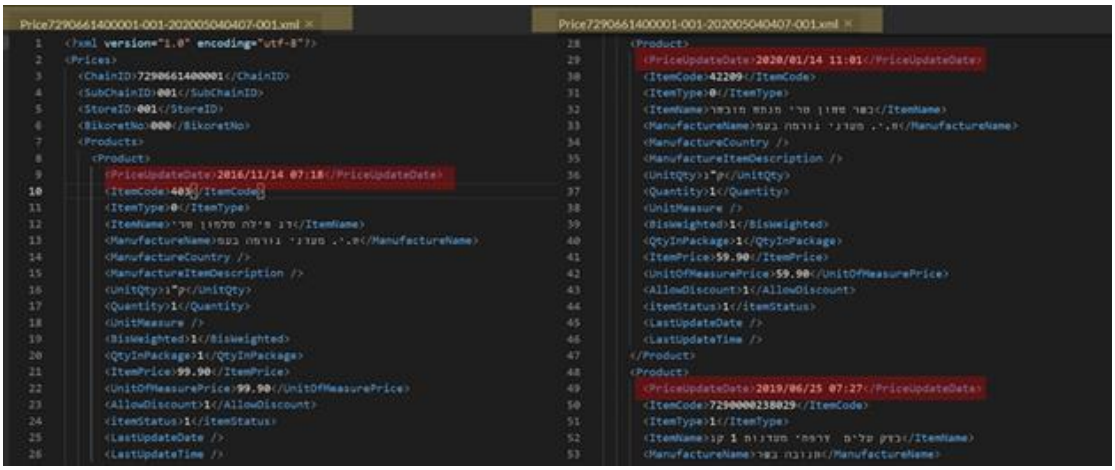
- תיאור הבעיה: בהתאם למבנה העקרוני הנדרש של הנתונים, בחברות רבות אין את כל הקבצים שאמורים להיות, דבר שיפגע ביכולת ההשוואה בין רשתות ובניתוח הנתונים בנוגע לכל רשת, בגלל היעדר בסיס להשוואה.
- דוגמות:

א. קיום קובץ "חנויות" בלבד (ללא פירוט מחירים ומבצעים כנדרש), ברשתות: "קשת", "אושר עד", "רמי לוי", טיב טעם" ועל כן לא ניתן לעשות השוואת מחירים ברשתות אלו

ב. קיום רק פלט קבצי "מבצעים" (ללא פירוט מחירים ורשימת חנויות כנדרש) ברשת "מגה".

4. **בעיית נתוני זמן** (Temporality):

- תיאור הבעיה: טווחי תאריכים בשדות עצמם (מועד עדכון מחיר המוצרים) לא חופפים, ולא מתואמים עם נתוני תיארוך נוספים (בשם הקובץ). הדבר מונע בחינה ברורה בהתאם לתאריך, או לכל הפחות ידרוש הסדרה של השימוש בתאריכים לפי אחד מהשדות המדוברים ובמודל אחיד (מועד הפקת הדו"ח – בשם הקובץ, מועד העדכון של הפרטים). זאת, כדי לאפשר השוואת מחירים של מוצרים בין חנויות שונות (באותה תקופה).
- דוגמות: לשני מוצרים שונים תחת אותו הקובץ יש שנת עדכון מחיר שונה.



5. **ריבוי ערכים חסרים בקובץ** (Missing Data):

- תיאור הבעיה: על פי המבנה הנדרש של הקבצים, בכל רשומה אמורים להיות נתונים מסוימים. היעדרו בולטת של נתונים בקבצים מסוימים פוגעת משמעותית ביכולת הניתוח שלו.
- דוגמות: בקבצים עם הסיומת FULL נמצאו ערכים חסרים. בריצה על 20 קבצים עם הסיומת FULL נתקלנו ב-9 קבצים עם ערכים חסרים, כלומר 45% מהקבצים שנדגמו לא היו בפורמט המצופה. בבדיקה ידנית שעשינו (פתיחת הקבצים באקסל) אכן קיימות עמודות או ערכים ריקים, כפי שהקוד התריע (כלומר, לא מדובר בבעיית פרסור). להלן פלט הריצה:

```
C:\Users\ASUS\PycharmProjects\HW2\converted\PriceFull7290696200003-086-202005040327-001.csv
there are missing values in this file True
```

```
C:\Users\ASUS\PycharmProjects\HW2\converted\PriceFull7290696200003-088-202005040300-001.csv
there are missing values in this file True
```

```
C:\Users\ASUS\PycharmProjects\HW2\converted\PriceFull7290696200003-089-202005040331-001.csv
there are missing values in this file True
```

6. **סוג ערך לא תקין** (Datatype invalid):

- תיאור הבעיה: על פי המבנה הנדרש של הקבצים, הערך של עמודת UnitQty בקבצי המחירים חייבים להיות אותיות בלבד. פער במאפיין זה פוגע הן בתהליך עיבוד המידע וכן מעיד על שימוש לא תקין בעמודה זו (על חשבון נתונים שאמורים להימצא בעמודת Quantity.
- דוגמות: פלט של קבצים שנמצאה בהם טעות:

```
in file: PriceFull7290696200003-095-202005040324-001.csv
 found: non alphabetic chars in UnitQty
['ק"ג' ,'l']
in file: PriceFull7290696200003-095-202005040324-001.csv
 found: non alphabetic chars in UnitQty
['ק"ג' ,'l']
in file: PriceFull7290696200003-095-202005040324-001.csv
 found: non alphabetic chars in UnitQty
['ק"ג' ,'l']
```

7. **יחס בין משקלים למחירים** (Uncorrelated corresponding data fields):

- תיאור הבעיה: ישנן שתי עמודות שמצופה בהן קוהרנטיות גבוהה: ItemPrice, UnitOfMeasurePrice (עד כדי סטייה הנובעת ממגבלת תמחור לפי אגורות, כלומר מיחידת המידה הקטנה ביותר). בשני אופנים:

1. אם יחידות המדידה (UnitMeasure) הן ביטוי זהה עד כדי פרמטר של יחידת המחיר הכללי (UnitQty), כמו במקרה לדוגמה של ק"ג-100 גרם, הקשר בין ItemPrice, UnitOfMeasurePrice אמור לקיים:

   ItemPrice = UnitOfMeasurePrice * UnitMeasure * quantity

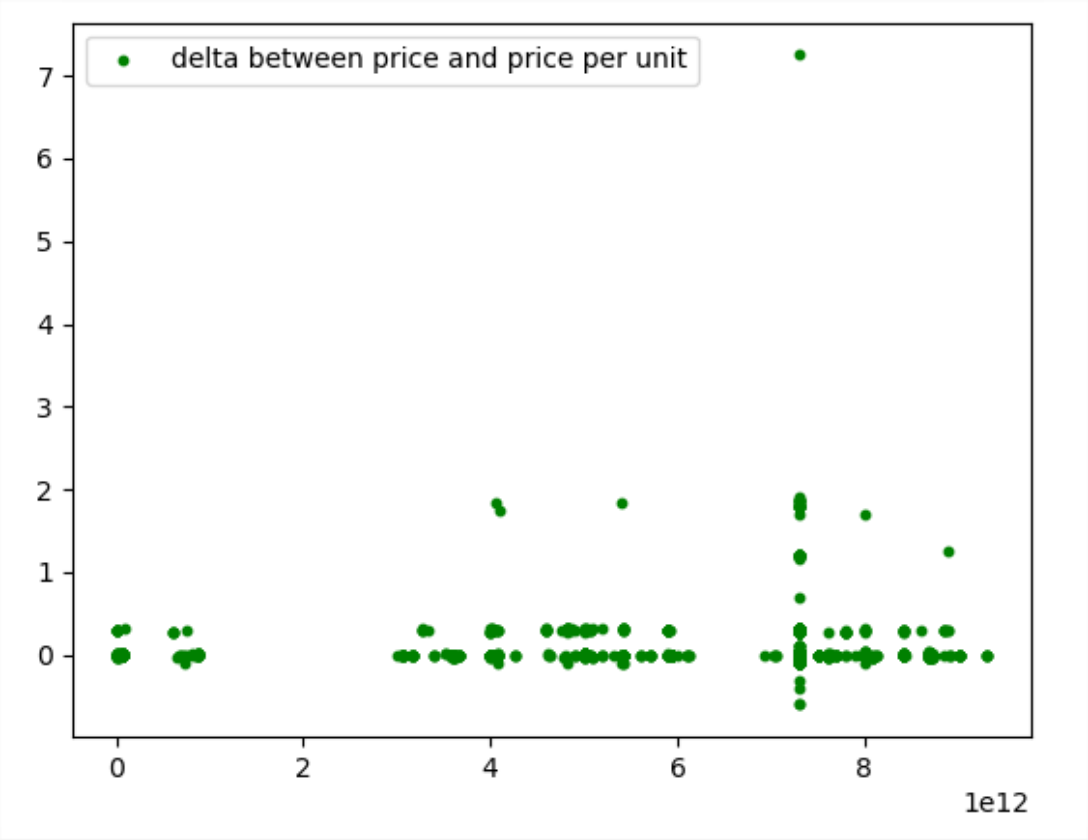
2. אם יחידות המדידה שונות בסקאלה שלהן (למשל במקרה של חישוב לפי יחידות שונות ליטר-ק"ג), הקשר בין ItemPrice, UnitOfMeasurePrice אמור לקיים:

   ItemPrice = UnitOfMeasurePrice * quantity

בהתאם לאמור, חריגות מקשר בין נתונים, שאמור להיות שיקוף טכני (ערך נגזר מערך בין אם ידנית או במסגרת הדו"ח), מעידות ששישנם פערים במידע. יתכן שלאור טעויות הזנה – ובכך פגיעה באיכות הנתונים; או לחלופין, פער מחירים במציאות – ובכך, תדרש הסדרה של ניתוח המחירים לפי אחד הערכים באופן עקבי, כדי לנתחם בצורה אפקטיבית (מעבר לכך שהנתון משקף פגיעה בלקוחות).

- דוגמות: להלן ניתוח בסיסי של נתונים החורגים מקשר זה וגרף פיזור של הפערים בין הערכים הנצפים.

price coha	itOfMea	mPrice	ylnPack	sWeight		itMeasu	uantity	itQty	mCode
	49.9	49.9	1	1		ק"ג	1	ק"ג	9391
envalid	59.9	59.9	1	1		100 גרם	1	ק"ג	23489
	49.9	49.9	1	1		ק"ג	1	ק"ג	47997
envalid	5	50	1	1		100 גרם	10	ק"ג	58078
envalid	2.2	22	1	1		100 גרם	10	ק"ג	58238
envalid	6.9	69	1	1		100 גרם	10	ק"ג	318103
envalid	7.9	79	1	1		100 גרם	10	ק"ג	318165
envalid	7.9	79	1	1		100 גרם	10	ק"ג	318288
envalid	8.5	85	1	1		100 גרם	10	ק"ג	318349
envalid	7.9	79	1	1		100 גרם	10	ק"ג	318363
envalid	6.9	69	1	1		100 גרם	10	ק"ג	318387
envalid	7.5	75	1	1		100 גרם	10	ק"ג	318851
envalid	7.2	72	1	1		100 גרם	10	ק"ג	318868
envalid	4.9	49	1	1		100 גרם	10	ק"ג	318899
envalid	4.9	49	1	1		100 גרם	10	ק"ג	319049
envalid	7.2	72	1	1		100 גרם	10	ק"ג	7.29E+12
envalid	4.9	49	1	1		100 גרם	10	ק"ג	319759
envalid	8	80	1	1		100 גרם	10	ק"ג	7.29E+12
valid		0.43	6.5	1	0	100 מ"ל	15	יח`	7.29002E+12
envalid		0.82	12	1	0	100 מ"ל	13.2	יח`	7.29002E+12
envalid		0.82	12	1	0	100 מ"ל	13.2	יח`	7.29002E+12
envalid		0.82	12	1	0	100 מ"ל	13.2	יח`	7.29002E+12
valid		4.6	6.9	1	0	1 ליטר	1.5	יח`	7.29002E+12
valid		4.6	6.9	1	0	1 ליטר	1.5	יח`	7.29002E+12
valid		4.6	6.9	1	0	1 ליטר	1.5	יח`	7.29002E+12
envalid		0.82	12	1	0	100 מ"ל	13.2	יח`	7.29002E+12
envalid		1.27	13.9	1	0	100 מ"ל	10	יח`	7.29002E+12
envalid		1.27	13.9	1	0	100 מ"ל	10	יח`	7.29002E+12
envalid		1.27	13.9	1	0	100 מ"ל	10	יח`	7.29002E+12
valid		426.67	6.4	1	0	100 מ"ל	0.015	יח`	7.29002E+12
valid		7.98	39.9	1	0	100 מ"ל	5	יח`	7.29002E+12



הסבר על הגרף: כל ערך שגדול מאפס מעיד על פער נצפה בין המחיר לבין הערך המחושב של המחירים כפונקציה של המחיר ליחידה והכמות הנמכרת

סעיף ב': בדיקת מציאות

בדקנו את התפלגות המחירים של כ-30% מקבצי PricesFull (10 קבצים מתיקיית victory ו-10 נוספים מתיקיית hashook). בחרנו לדגום את הקבצים האלו מכיוון שהם מאפשרים דגימה של יחסית הרבה מאוד מחירים. בבדיקה שלנו בחנו שני מאפיינים:

- ראשית, האם המחירים עומדים בבדיקת ולידיות בסיסית: ערכם גדול מאפס.
- שנית, האם המחירים מתפלגים כמצופה: ציפייה לטווחי מחירים של סופר, לפיהם ישנם הרבה מוצרים זולים ומעט מוצרים יקרים, עד כדי אלפים בודדים. ובהתאם, התפלגות המחירים הצפויה הינה לפי Power-law.

כדי לבחון את הנתונים בנוחות, לאחר שאכן התקבלה התפלגות התאמנו את גודל הbins והכנסנו ערכי קצה ל-collection bin כדי שהפריסה שתתקבל תהיה נוחה ככל הניתן.

לסיכום ניתן לראות שאכן הערכים עמודים בשני הקריטריונים והמחירים בכללם נראים כערכים תקינים.