

# מבוא למדעי הנותנים – מטלה 1

עומר כהן 208715813

אילון פישמן 318404282

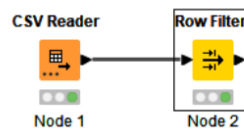
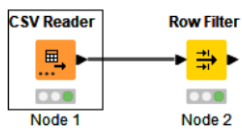
## שאלה 1

סעיף א':


לפני הסינון: 896 ערכים.

לאחר הסינון: 475 ערכים.

סה"כ סוגנו 421 ערכים.



Console

 Node Monitor

×

Node:

CSV Reader (4:1)

State:

EXECUTED

Port Output

Port 0

Load data

Rows: 896, Columns: 6

| ID   | Store ID | Store_Area | Items_Available | Daily_Customer_Count | Store_Sales | City     |
|------|----------|------------|-----------------|----------------------|-------------|----------|
| Row0 | 1        | 1659       | 1961            | 530                  | 66490       | Tel Aviv |
| Row1 | 2        | 1461       | 1752            | 210                  | 39820       | Tel Aviv |

Console Node Monitor

Node: Row Filter (4:2)

State: EXECUTED

Port Output Port 0 Load data Rows: 475, Columns: 6

| ID   | Store ID | Store_Area | Items_Available | Daily_Customer_Count | Store_Sales | City     |
|------|----------|------------|-----------------|----------------------|-------------|----------|
| Row1 | 2        | 1461       | 1752            | 210                  | 39820       | Tel Aviv |
| Row2 | 3        | 1340       | 1609            | 720                  | 54010       | Tel Aviv |

סעיף ב':

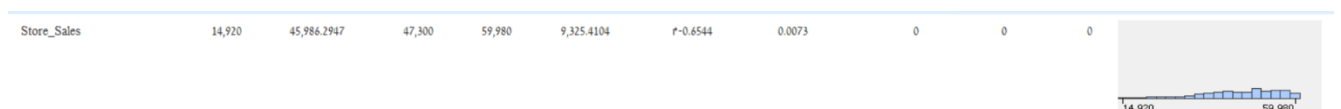
ממוצע: 45986.29

סטיית תקן: 9325.41

סטיית תקן מנורמלת: 0.2027

$$cv = \frac{sd}{\bar{x}}$$

חציון: 47300



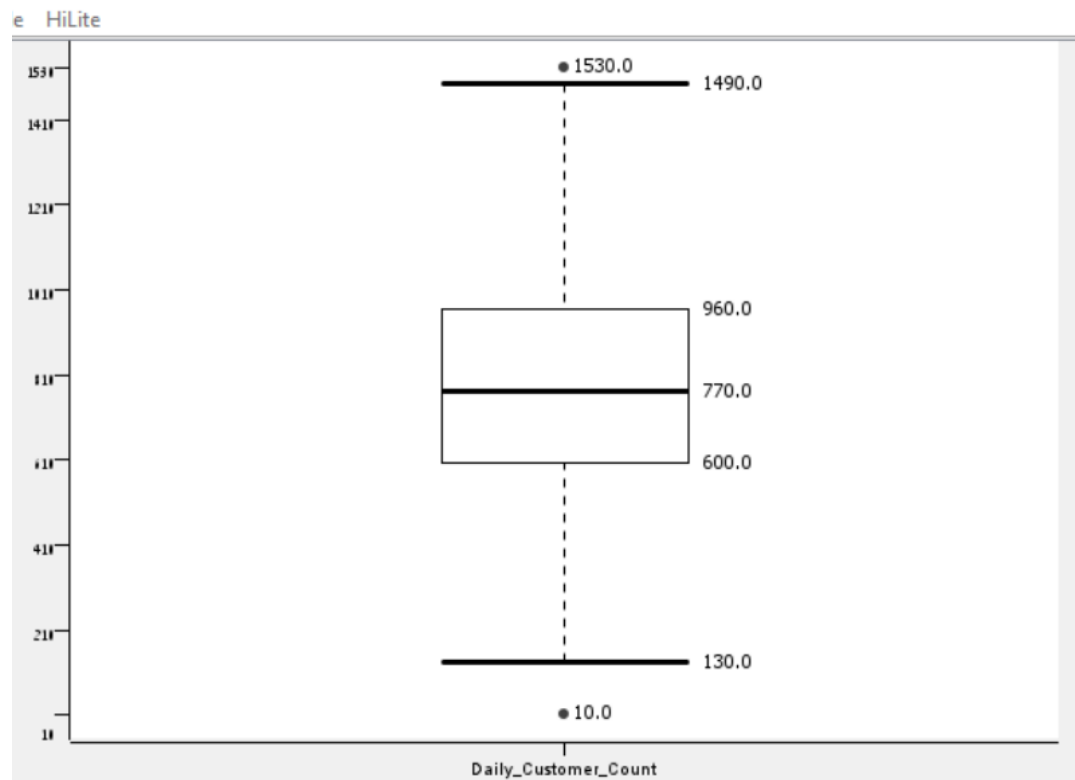
## סעיף ג':

$$Q_1 = 600, Q_3 = 1490$$

$$IQR = Q_3 - Q_1 = 1490 - 600 = 890$$

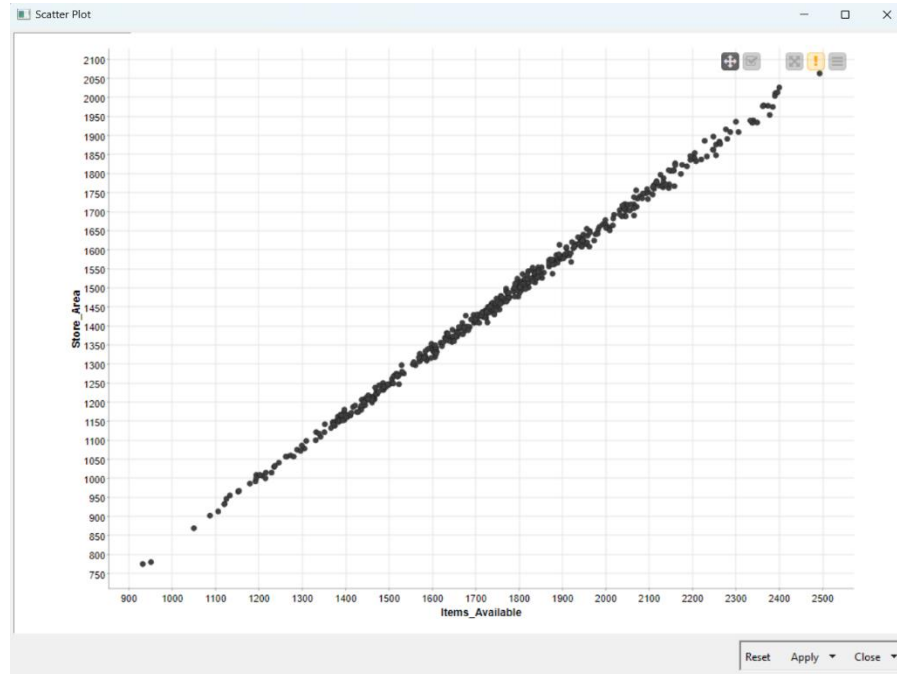
$$Minimum = Q_1 - 1.5 \cdot IQR = 600 - 1.5 \cdot 890 = -735$$

$$Maximum = Q_3 + 1.5 \cdot IQR = 1490 + 1.5 \cdot 890 = 2825$$



## סעיף ד' a:

הייתי מצפה לקבל בגרף קשר בין גודל החנות לפריטים הזמינים. כפי שניתן להניח התנהגות טיפוסית, אמור להיות קשר פרופורציונלי, ככל שהחנות תהיה גדולה יותר יהיו יותר פריטים בה.



## סעיף ד' b:

הגרף מצביע על קשר ליניארי חיובי עולה. כפי שצויין בסעיף a ככל שהחנות גדולה יותר כמות הפריטים הזמינה גדלה.

סעיף ה':








































מקדם המתאם בין שני משתנים אלו הוא 0.99892

| Row ID          | D Store ID    | D Store_Area         | D Items_Available     | D Daily_Customer_...  | D Store_Sales       | D City |
|-----------------|---------------|----------------------|-----------------------|-----------------------|---------------------|--------|
| Store ID        | 1.0           | -0.05079497265031... | -0.043664416344423... | -8.391279631761148... | 0.08337212058242... | ?      |
| Store_Area      | -0.0507949... | 1.0                  | 0.9989249423064273    | -0.02699260477597359  | 0.06851568443320... | ?      |
| Items_Available | -0.0436644... | 0.9989249423064273   | 1.0                   | -0.028879161736706... | 0.07384147722798... | ?      |
| Daily_Custom... | -8.3912796... | -0.02699260477597... | -0.028879161736706... | 1.0                   | 0.03978270112158... | ?      |
| Store_Sales     | 0.08337212... | 0.06851568443320674  | 0.07384147722798493   | 0.03978270112158677   | 1.0                 | ?      |
| City            | ?             | ?                    | ?                     | ?                     | ?                   | 1.0    |

Correlation Matrix - 4:6 - Linear Correlation

File View

**Some numeric column(s) contained missing**

|  |   |   |   |   |   |   |
|--|---|---|---|---|---|---|
|  corr = -1  | Store ID  |   |   |   |   |   |
|  corr = +1  | Store...  |   |   |   |   |   |
|  corr = n/a | Items...  |   |   |   |   |   |
|  | Daily_...   |   |   |   |   |   |
|  | Store...  |   |   |   |   |   |
|  | City  |   |   |   |   |   |
| Store ID   |  |  |  |  |  |  |
| Store_Area   |  |  |  |  |  |  |
| Items_Available  |  |  |  |  |  |  |
| Daily_Customer_...   |  |  |  |  |  |  |
| Store_Sales  |  |  |  |  |  |  |
| City   |  |  |  |  |  |  |

## שאלה 2

כפי שניתן לראות תוחלת הלוקחות המגיעים ביום היא בקירוב 787.7 עבור חיפה ו798.544 עבור תל אביב. סטיית התקן בקירוב היא 255.72 עבור חיפה ו280.41 עבור תל אביב.

קיבלנו במבחן לווין לשוויון שוניות כי  $P_{value} = 0.0466 = 4.66\% < 5\%$ . מכאן אנו מבינים כי ההנחה של שוניות שוות לא מתקיימת.

$H_0$  היא שלא קיים הבדל מובהק סטטיסטית בין מספר הלוקחות בין חיפה לתל אביב.

מסתכלים על השורה התחתונה ורואים כי ערך ה  $p\text{-value}(2\text{ tailed})$  שווה ל0.6167 שהוא גבוה מ5%. מכאן ניתן להסיק כי לא ניתן לדחות את השערת האפס לכן אנו מסיקים כי אין הבדל מובהק סטטיסטית בספירת הלוקחות היומית.

### Independent groups t-test

#### Group Statistics

|                      | Group    | N   | Missing Count | Missing Count (Group Column) | Mean     | Standard Deviation | Standard Error Mean |
|----------------------|----------|-----|---------------|------------------------------|----------|--------------------|---------------------|
| Daily_Customer_Count | Haifa    | 287 | 0             | 0                            | 787.7003 | 255.7225           | 15.0948             |
| Daily_Customer_Count | Tel Aviv | 323 | 0             | 0                            | 798.5449 | 280.4157           | 15.6027             |

286 rows have been ignored. Their value in the grouping column is neither "Haifa" nor "Tel Aviv".

#### Levene Test

The Levene Test is used to test for the equality of variances.

|                      | F      | df 1 | df 2 | p-Value |
|----------------------|--------|------|------|---------|
| Daily_Customer_Count | 3.9775 | 1    | 608  | 0.0466  |

#### Independent Groups Statistics

Confidence Interval (CI) Probability: 95.0%

Differences are reported of the groups: Haifa - Tel Aviv

|                      | Variance Assumption         | t         | df       | p-value (2-tailed) | Mean Difference | Standard Error Difference | CI (Lower Bound) | CI (Upper Bound) |
|----------------------|-----------------------------|-----------|----------|--------------------|-----------------|---------------------------|------------------|------------------|
| Daily_Customer_Count | Equal variances assumed     | t=-0.4968 | 608      | 0.6195             | t=-10.8445      | 21.8277                   | t=-53.7114       | 32.0223          |
| Daily_Customer_Count | Equal variances not assumed | t=-0.4995 | 607.5838 | 0.6176             | t=-10.8445      | 21.7094                   | t=-53.4792       | 31.7901          |

### שאלה 3

#### סעיף א':

סה"כ נמצאו 14 ערכים חריגים.

המשתנה בו היו הכי הרבה ערכים חריגים הם *store\_area*, *items\_available* עם 5 ערכים כל אחד.

| Row ID | \$ Outlier ... | I Membe... | I Outlier ... | D Lower ... | D Upper ... |
|--------|----------------|------------|---------------|-------------|-------------|
| Row0   | Store_Area     | 877        | 5             | 810.625     | 2,161.625   |
| Row1   | Items_Avail... | 896        | 5             | 962         | 2,594       |
| Row2   | Daily_Custo... | 896        | 3             | 45          | 1,525       |
| Row3   | Store_Sales    | 896        | 1             | 8,205       | 110,005     |

#### סעיף ב':

1. על מנת להשלים ערכים חסרים חובר הצומת *Numeric outliers* אל צומת חדשה-

*Missing values*. לכל משתנה הותאם ערך ההשלמה לפי סוגו.

משתנה איכותי חריג הושלם ע"י הערך השכיח.

משתנה כמותי חריג הושלם ע"י ממוצע או חציון.

*Store\_Area*: נשתמש בחציון מכיוון שהוא פחות רגיש לערכים קיצוניים ועשוי לספק ערך מייצג יותר של חנות טיפוסית.

*Items\_Available*: נשתמש בחציון, בדומה ל*store area* תמנע השפעה עם ספירת מלאי גבוהה או נמוכה במיוחד.

*Daily\_Customer\_Count*: נשתמש בחציון, כך ימנע מצב שבהם יש ימים עם תנועת לקוחות נמוכה או גבוהה במיוחד.

*Store\_Sales*: נשתמש בממוצע שכן הממוצע מתאים במקרה זה.

*City*: משתנה איכותי קטגורי, נשתמש בערך השכיח ביותר.

*Store\_id*: מזהה ייחודי, ערכים חסרים במשתנה זה יהוו בעיה, חוסר בערך זה יהיה הגיוני להשמיט את השורה כולה.

בחרנו לבצע בצורה זו את השלמת הערכים, בגלל שמשתנה איכותי נקבע ע"י תכונה שלא ניתן למדוד באופן כמותי – כך שחשובים סטטיסטיים, לדוגמת ממוצע וחציון, אינם בעלי משמעות.

במשתנים כמותיים יש משמעות לחישובים הסטטיסטיים, לכן נבחר להשתמש בממוצע או חציון. סיננו את המספרים הקיצוניים.

## שאלה 4

