

פרויקט רגרסיה לינארית



קבוצה 30

315625301_310050893_313191884_208849265

תוכן עניינים

חלק א' – בחירת בסיס נתונים וניתוח סטטיסטי של הנתונים	3
1. בחירת מאגר נתונים	4
2. יצירת טבלת משתנים	4
3. תיאור המשתנים	5
4. תיאור קשרים בין משתנים	6
5. ניתוח תיאורי של המשתנים (נספח 2)	7
6. ניתוח חריגים (נספח 3)	9
7. פונקציית צפיפות והתפלגות מצטברת (נספח 4)	11
8. ייצוג קשרים בעזרת תרשימים (נספח 5)	13
9. טבלאות שכיחויות (נספח 6)	15
חלק ב' – ניתוח פורמאלי של מאגר הנתונים	19
1. תקציר מנהלים	19
2. עיבוד מקדים	19
2.1. הסרה של משתנים (נספח 7)	19
2.2. התאמה של משתנים (נספח 8)	20
2.3. משתני דמה (נספח 9)	22
2.4. הוספת משתני אינטראקציה (נספח 10)	23
3. התאמת המודל ובדיקת הנחות המודל	26
3.1. בחירת משתני המודל (נספח 11)	26
3.2. בדיקת הנחות המודל (נספח 12)	27
3.3. דוגמא לשימוש המודל הנבחר	27
4. שיפור המודל (נספח 13)	28
נספחים	31
1. חלק א' סעיף 4	31
2. חלק א' סעיף 5	32
3. חלק א' סעיף 6	33
4. חלק א' סעיף 7	34
5. חלק א' סעיף 8	34
6. חלק א' סעיף 9	34
7. חלק ב' סעיף 2.1	35

36	חלק ב' סעיף 2.2	8.
36	חלק ב' סעיף 2.3	9.
37	חלק ב' סעיף 2.4	10.
37	חלק ב' סעיף 3.1	11.
41	חלק ב' סעיף 3.2	12.
42	חלק ב' סעיף 4	13.

רשימת איורים

13	איור 1- פיזור בין קטגוריית מוצר לבין מחיר ליחידת מוצר
13	איור 1- פיזור בין קטגוריית מוצר לבין מחיר ליחידת מוצר
13	איור 2- פיזור בין שווי הסחורה לבין מחיר ליחידת מוצר
13	איור 2- פיזור בין שווי הסחורה לבין מחיר ליחידת מוצר
14	איור 3- פיזור בין רווח הקנייה ברוטו לבין שווי הסחורה
14	איור 3- פיזור בין רווח הקנייה ברוטו לבין שווי הסחורה
14	איור 4- פיזור בין דירוג לקוחות לבין המחיר הסופי של הרכישה
14	איור 4- פיזור בין דירוג לקוחות לבין המחיר הסופי של הרכישה
14	איור 5- פיזור בין מספר המוצרים שנרכשו לבין המחיר הסופי של הרכישה
14	איור 5- פיזור בין מספר המוצרים שנרכשו לבין המחיר הסופי של הרכישה
15	איור 6- פיזור בין חודש הקנייה לבין המחיר הסופי של הרכישה
15	איור 6- פיזור בין חודש הקנייה לבין המחיר הסופי של הרכישה
20	איור 7- תרשים פיזור בין דירוג המוצרים לסכום הרכישה הסופי
20	איור 7- תרשים פיזור בין דירוג המוצרים לסכום הרכישה הסופי
24	איור 8- הקשר בין כמות המוצרים שנרכשו לבין מגדר הלקוח
24	איור 8- הקשר בין כמות המוצרים שנרכשו לבין מגדר הלקוח
24	איור 9- הקשר בין מחיר ליחידת מוצר לבין מגדר הלקוח
24	איור 9- הקשר בין מחיר ליחידת מוצר לבין מגדר הלקוח
25	איור 10- הקשר בין מספר המוצרים שנרכשו לבין סוג הלקוח
25	איור 10- הקשר בין מספר המוצרים שנרכשו לבין סוג הלקוח
25	איור 11- הקשר בין מחיר ליחידת מוצר לבין סוג הלקוח
25	איור 11- הקשר בין מחיר ליחידת מוצר לבין סוג הלקוח
28	איור 12- טרנספורמצית Box-Cox

רשימת טבלאות

4	טבלה 1- יצירת משתנים
9	טבלה 2- ניתוח חריגים

חלק א' – בחירת בסיס נתונים וניתוח סטטיסטי של הנתונים

1. בחירת מאגר נתונים

בפרויקט זה נחקר את הגורמים המשפיעים על המחיר הסופי של סל קניות בכלבו. המשתנה המוסבר הוא המחיר הסופי של סל קניות בכלבו, המשתנה המסביר הם כמות הפריטים בסל, מחיר ליחידה, קטגוריית הפריט, תאריך קניה, אמצעי התשלום ועוד. האתר ממנו נלקחו הנתונים מוצגים בקישור הבא:

<https://www.kaggle.com/aungpyaeap/supermarket-sales>

שינויים שביצענו בבסיס הנתונים:

- א. הפכנו את משתנה תאריך הקנייה לחודש הקנייה לפי מספר החודש (1-ינואר, 2-פברואר, 3-מרץ).
- ב. הפכנו את משתנה שעת הקנייה לזמן קנייה ביום (עד 12:00 בצהריים- בוקר, 12:00-17:00 צהריים, 18:00-00:00 ערב).

2. יצירת טבלת משתנים

טבלה 1- יצירת משתנים

סוג המשתנה - מוסבר/מסביר	סימון	יחידת מידה	סוג המשתנה - רצף / קטגוריאל	הסבר קצר על המשתנה
מוסבר	Y	\$	רצף	מחיר סופי של סל הקניות
מסביר	X ₁	-	קטגוריאל	הסניף בו התבצעה הרכישה (C,B,A)
מסביר	X ₂	-	קטגוריאל	עיר בה ממוקם הכלבו (Mandalay, Yangon, Napyitaw)
מסביר	X ₃	-	קטגוריאל	סוג הלקוח (רגיל, חבר מועדון)
מסביר	X ₄	-	קטגוריאל	מגדר הלקוח (זכר, נקבה)
מסביר	X ₅	-	קטגוריאל	קטגוריית המוצר (אלקטרוניקה, אופנה, מזון, בריאות ויופי, בית ולייף סטייל, ספורט וטיולים)
מסביר	X ₆	\$	רצף	מחיר ליחידת מוצר
מסביר	X ₇	-	כמותי	מספר המוצרים שנרכשו ע"י הלקוח
מסביר	X ₈	\$	רצף	5% מיסים מסך הקנייה
מסביר	X ₉	-	קטגוריאל	החודש בה התבצעה הקנייה (1-ינואר, 2-פברואר, 3-מרץ)
מסביר	X ₁₀	-	קטגוריאל	השעה בה התבצעה הקנייה (בוקר, צהריים, ערב)
מסביר	X ₁₁	-	קטגוריאל	סוג התשלום (Ewallet, cash, credit card)
מסביר	X ₁₂	\$	רצף	שווי הסחורה
מסביר	X ₁₃	%	רצף	כמה הכלבו מרוויח מכל קנייה באחוזים
מסביר	X ₁₄	\$	רצף	רווח הקניה ברוטו בדולרים
מסביר	X ₁₅	-	רצף	דירוג הלקוחות עפ"י חוויית הקנייה הכוללת שלהם (בסולם של 1 עד 10)

3. תיאור המשתנים

1. X_1 - הסניף בו התבצעה הרכישה

הסניף בו התבצעה עשוי להשפיע על סכום הקנייה הכולל לפי כמות ושוני האוכלוסייה המגיעה לכלבו.

2. X_2 - עיר בה ממוקם הכלבו

העיר בה ממוקם הכלבו עשוי להשפיע על סכום הקנייה הכולל של הלקוח לפי המיקום הגיאוגרפי של העיר, האם היא נגישה ומהי מידת הפופולריות שלה אשר תמשוך קהל יעד. בנוסף מיקום העיר משפיע על מנעד המחירים של המוצרים הנמכרים דבר המשפיע על כמות הרכישות באותה העיר.

3. X_3 - סוג הלקוח

סוג הלקוח משפיע על ההטבות שכל לקוח מקבל ובכך משפיע באופן ישיר על סכום הקנייה הכולל. בנוסף משתנה מסביר זה יכול להשפיע על מאפייני הקנייה (לדוגמא כמות מוצרים נרכשת).

4. X_4 - מגדר הלקוח

מגדר הלקוח ונטיות הרכישה שלו יכולים להשפיע על כמות הפריטים הנקנים ובכך להשפיע על סכום הקנייה הכולל.

5. X_5 - קטגוריית המוצר

קטגוריות מסוימות הן יותר פופולריות מאחרות ונוגעות בקהל יעד רחב יותר. לכן יותר לקוחות יקנו מוצרים מאותם קטגוריות דבר אשר משפיע על סכום הקנייה הכולל.

6. X_6 - מחיר ליחידת מוצר

ככל שמוצר יותר זול ככל הנראה ירצו לקנות ממנו בכמות גדולה יותר ולכן מחיר המוצר הוא גורם השפעה ישיר על סכום הקנייה הכולל.

7. X_7 - מספר המוצרים שנרכשו ע"י הלקוח

מספר מוצרים גדול יותר יגדיל את סכום הקנייה הכולל של הלקוח ולהפך. לכן זהו גורם ישיר המשפיע על סכום הרכישה הכולל.

8. X_8 - 5% מיסים מסך הקנייה

ככל שמחיר הקנייה ברוטו גבוה יותר כך גודל המיסים עבור הקנייה גדול יותר וכתוצאה מכך מחיר הקנייה הכולל (נטו) גבוה יותר.

9. X_9 - החודש בה התבצעה הקנייה

החודש בו התבצעה הקנייה עשוי להשפיע על מחיר קנייה כולל כיוון וסביר להניח כי העונות והחגים בחודשים אלו משפיעים על כל אחד מהם בצורך הצריכה.

10. X_{10} - השעה בה התבצעה הקנייה

ניתן לומר שהשעה בה התבצעה הקנייה משפיעה על סכום הקנייה הכולל בכך שבכל זמן נתון יש הבדל בכמות האנשים שנמצאים בכלבו הנחשפים למוצרים בחנות.

11. X_{11} - סוג התשלום

ניתן לומר כי ישנם סוגי תשלומים יותר פופולריים הנוגעים בקהל יעד רחב יותר שזהו דבר המעודד רכישה ומשפיע על סכום הקנייה הכולל

12. X_{12} - שווי הסחורה

שווי סחורה גבוה יותר נותן לכלבו לגיטימציה למכור את המוצרים בחנות במחירים

גבוהים יותר לצורך יצירת רווח גדול יותר וכתוצאה מכך להגדיל את סכום הקנייה הכולל של הלקוח.

13. X_{13} - כמה הכלבו מרוויח מכל קנייה באחוזים

אחוז הרווח של הכלבו משפיע על מחירי המוצרים ובכך משפיע ישירות על סכום הקנייה הכולל שהלקוחות ישלמו.

14. X_{14} - רווח הקנייה ברוטו בדולרים

רווח הקנייה ברוטו פר רכישה משפיע על מחירי המוצרים ובכך משפיע ישירות על סכום הקנייה הכולל שהלקוחות ישלמו.

15. X_{15} - זירוג הלקוחות

ניתן לומר שככל שהלקוחות מדרגים את חווית הקניה שלהם גבוה יותר אז יותר לקוחות ירצו להגיע בשנית לסניף ולבצע רכישות נוספות. חווית קניה טובה מעודדת את הלקוחות לקנות יותר מוצרים ולהגדיל את סכום הרכישה הכולל.

4. תיאור קשרים בין משתנים

כתוצאה מהגרף (נספח 1) ומהסתכלות על הנתונים, אלה הקשרים בין המשתנים:

קשרים מדגמיים:

- X_1 - הסניף בו התבצעה הרכישה, X_2 - עיר בה ממוקם הכלבו ($X_1 \leftarrow X_2$)
הסניף בו מתבצעת הרכישה מוכל בעיר בה ממוקם הכלבו ולכן אנחנו מצפים לראות מקדם מתאם חיובי בין שני משתנים אלו. ככל שיש יותר רכישות בסניף מסוים כך יהיו יותר רכישות בעיר בה הסניף ממוקם. משתנים אלו מתבססים על לקוחות מתוך מדגם (לקוחות מסניף מסוים) ועל כן הקשר המדגמי.
- X_6 - מחיר ליחידת מוצר, X_7 - מספר המוצרים שנרכשו ע"י הלקוח ($X_6 \leftarrow X_7$)
מחיר המוצר מאוד משמעותי עבור מספר המוצרים שיירכשו ע"י הלקוחות כיוון שככל שהמחיר יהיה יותר זול אז יש פוטנציאל גבוה יותר שהלקוח ירכוש מספר גדול יותר של מוצרים מאותו הסוג. מצד שני, ככל שמספר המוצרים ברכישה גדול יותר נוכל להסיק שהמחיר ליחידה נמוך יותר. כלומר, נצפה לראות קשר הפוך בין מחיר ליחידת מוצר לבין מספר המוצרים שיירכשו ע"י הלקוח אבל לא ברור מי הגורם של מי ולכן אנו סבורים כי מדובר בקשר מדגמי.
- X_7 - מספר המוצרים שנרכשו ע"י הלקוח, X_9 - החודש בה התבצעה הקנייה, X_{10} - השעה בה התבצעה הקנייה ($X_{11}, X_{10} \leftarrow X_7$)
החודש והשעה בהם התבצעה הרכישה עשויים להשפיע על כמות המוצרים שיירכשו באותה הקנייה כיוון שיש השפעה רבה לחודש והשעה ביום על הרכישה. לדוגמא: במידה ובחודש מסוים יש יותר חגים אז דבר זה משפיע על כמות המוצרים שיירכשו. בנוסף, במידה ומדובר בשעות הערב אז יותר אנשים מגיעים לרכוש מוצרים אחרי שעות העבודה ולכן מספר המוצרים שיירכשו יהיה ככל הנראה גבוה יותר. מצד שני, מספר המוצרים הנרכשים יכול להעיד על השעה והחודש שנקנו (כמויות גדולות ממוצרים מסוימים יכולים להעיד על חגים מסוימים או מבצעים בחודשים מסוימים). לכן, נצפה לראות קשר חיובי בין

השעות והחודשים הפופולריים יותר לבין מספר המוצרים שיירכשו ע"י הלקוח
אך אין לדעת מי גרם למי ולכן אנו סבורים שמדובר בקשר מדגמי.

קשרים סיבתיים:

- X_5 - קטגוריית המוצר, X_6 - מחיר ליחידת מוצר ($X_5 \rightarrow X_6$)
למוצרים הנמצאים בקטגוריה מסוימת יכול להיות מנעד מחירים שונה מאשר
בקטגוריה אחרת. לדוגמא: מוצרים הנמצאים בקטגוריית אלקטרוניקה יותר
יקרים מאשר מוצרים בקטגוריית מזון. נצפה לראות קשר חיובי בין קטגוריית
יקרות יותר לבין מחירי המוצרים הנמצאים בה.
- X_3 - סוג הלקוח, X_{15} - דירוג הלקוחות ($X_{15} \rightarrow X_3$)
סוג הלקוח מושפע מדירוג הלקוחות שבה לידי ביטוי בחוויית הקנייה של הלקוח.
ככל שחוויית הקנייה של הלקוח טובה יותר, הוא ירצה לשוב לקנות בכלבו ולכן
ככל הנראה ירצה לעשות חבר מועדון ולשנות את סטטוס סוג הלקוח שלו. לכן,
נצפה לראות קשר חיובי בין דירוג גבוה לבין סטטוס חבר מועדון של הלקוח.
- X_6 - מחיר ליחידת מוצר, X_{12} - שווי הסחורה, X_{13} - כמה הכלבו מרוויח מכל
קנייה באחוזים ($X_{12}, X_{13} \rightarrow X_6$)
ככל ששווי הסחורה גדול יותר, כך יתמחרו את המוצרים במחיר גבוה יותר זאת
מפני שאחוז הרווח של הכלבו פר מוצר הוא קבוע. מכאן, נצפה לראות קשר חיובי
בין עלות הסחורה שנקנתה מהספק והרווח באחוזים של הכלבו לבין מחיר
ליחידת מוצר.
- X_7 - מספר המוצרים שנרכשו ע"י הלקוח, X_{14} - רווח הקנייה ברוטו בדולרים
ככל שמספר המוצרים שנרכשו בקנייה מסוימת גדול יותר, המחיר הכולל של
הרכישה יגדל גם כן וכתוצאה מכך רווח הקנייה ברוטו של הכלבו יגדל. לכן,
נצפה לראות קשר חיובי בין מספר המוצרים שנרכשו ע"י הלקוח לבין רווח
הקנייה ברוטו בדולרים של הכלבו.

5. ניתוח תיאורי של המשתנים (נספח 2)

Total price including tax - Y

נתון סטטיסטי	תוצאה	הסבר
ממוצע	322.966749	ניתן לראות מהנתונים שהממוצע גבוהה מחציון וסטיית התקן גדולה יחסית, הסבר אפשרי לניתוח הני"ל שישנם מספר רכישות שסכומם היה גדול בהרבה מהממוצע ולכן משך את הממוצע מעל לחציון, בעקבות כך נצפה להתפלגות א- סימטרית עם זנב ימני. דבר המתיישב עם סטיית התקן הגדולה שמשמעותה פיזור גדול בנתונים.
א-סימטריה (skewness)	0.892569805	
סטיית תקן	245.8853351	
רבעון ראשון	124.422375	
חציון	253.848	
רבעון שלישי	471.35025	

Unit price - X_6

הסבר	תוצאה	נתון סטטיסטי
מהנתונים על משתנה זה ניתן לראות כי החציון והממוצע קרובים מאוד וערך הא-סימטריה קטן מאוד. בצירוף הנתונים האלה נצפה להתפלגות סימטרית. סטיית התקן הינה גדולה ומשמעותה פיזור גדול של הנתונים, על הגרף היא תיראה כרחבה ונמוכה. ניתן לראות שהמחיר הממוצע למוצר בכלבו הינו 55.6 דולר.	55.67213	ממוצע
	0.007077448	א-סימטריה (skewness)
	26.49462835	סטיית תקן
	32.875	רבעון ראשון
	55.23	חציון
	77.935	רבעון שלישי

Number of products purchased by customer - X_7

הסבר	תוצאה	נתון סטטיסטי
מהנתונים על משתנה זה ניתן לראות כי החציון והממוצע קרובים וערך הא-סימטריה קטן יחסית. מכך ניתן לצפות להתפלגות סימטרית. סטיית התקן הינה קטנה ומשמעותה גרף כגבוהה וצרה. כל לקוח רוכש מספר רב של פריטים מאותו מוצר ואינו מסתפק בפריט אחד. ייתכן שמדובר בטקטיקה השיווקית של הכל-בו.	5.51	ממוצע
	0.012941048	א-סימטריה (skewness)
	2.923431	סטיית תקן
	3	רבעון ראשון
	5	חציון
	8	רבעון שלישי

הסבר	תוצאה	נתון סטטיסטי
כאן ניתן לראות שקיים מרחק בין הממוצע לחציון, סטיית התקן גדולה ומדד הסימטריות חיובי, הדבר מתאר התפלגות א-סימטרית בעלת זנב ימני ארוך כאשר מרבית הנתונים הם ברביע השלישי. נתונים אלה נובעים מחישוב המס על העלות הכוללת של הקנייה	15.379369	ממוצע
	0.892569805	א-סימטריה (skewness)
	11.70882548	סטיית תקן
	5.924875	רבעון ראשון
	12.088	חציון
	22.44525	רבעון שלישי

tax fee for customer buying - X_8

Cost of goods sold - X_{12}

הסבר	תוצאה	נתון סטטיסטי
קיימת סטיית תקן עצומה דבר המעיד על פיזור רב בין הנתונים. בנוסף, ניכר שהחציון נמוך מהממוצע, בדר המעיד על תצפיות אשר מושכות מעלה את הממוצע. נתון זה מתיישב עם הנתון של א-סימטריה שמעיד על זנב ימני. היות שהמדד נקבע עפ"י עלות הסחורה שנמכרה ניתן להסיק ישנם מספר מוצרים שמחירים גבוה בהרבה מן הממוצע וזה מה שיוצר את הזנב הימני.	307.58738	ממוצע
	0.892569805	א-סימטריה (skewness)
	234.1765096	סטיית תקן
	118.4975	רבעון ראשון
	241.76	חציון
	448.905	רבעון שלישי

Gross income - X_{14}

הסבר	תוצאה	נתון סטטיסטי
<p>היחס בין הנתונים של משתנה זה דומים ליחס הנתונים שראינו ב X_{13}. ניתן להבין זאת מכיוון שקיימת תלות בין המשתנים, ההכנסה של הסופר תלויה בעלות הסחורה שנמכרה. כלומר שגם כאן קיים פיזור רב בנתונים ומדד הא-סימטריה גדול שמעיד על זנב ימני.</p>	15.379369	ממוצע
	0.892569805	א-סימטריה (skewness)
	11.70882548	סטיית תקן
	5.924875	רבעון ראשון
	12.088	חציון
	22.44525	רבעון שלישי

Rate of costumers - X_{15}

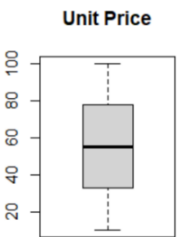
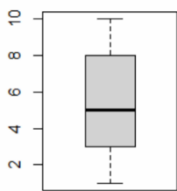
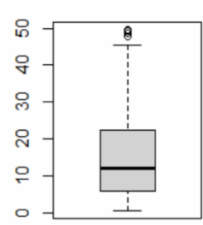
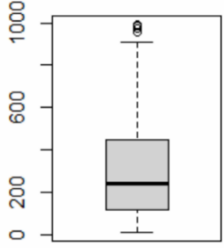
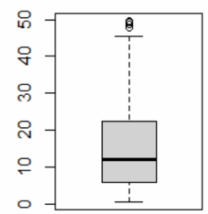
הסבר	תוצאה	נתון סטטיסטי
<p>הממוצע והחציון במשתנה זה קרובים מאוד ומדד הא-סימטריה קטן. לכן, נצפה להתפלגות נורמלית של הנתונים. דבר המתיישב עם העובדה שהנתונים מייצגים דירוג לקוחות. עפ"י משפט הגבול המרכזי, כאשר יש די תצפיות בלתי תלויות הן מתקרבות להתפלגות נורמלית לאחר תקנון מסוים.</p>	6.9727	ממוצע
	0.009009649	א-סימטריה (skewness)
	1.718580294	סטיית תקן
	5.5	רבעון ראשון
	7	חציון
	8.5	רבעון שלישי

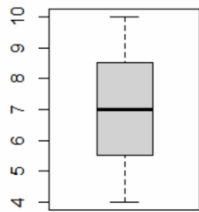
6. ניתוח חריגים (נספח 3)

טיפול בתצפיות חריגות נועד כדי לנפות או להתעלם מתצפיות אשר לא מתיישבות עם המודל ונמצאות מאוד רחוק מהחציון. כדי למצוא את התצפיות החריגות נשתמש בפונקציית ה boxplot בתוכנת ה- R. את הערכים החריגים נצפה לראות מסומנים בעיגול מחוץ לגבולות הקופסא. בנוסף ערכים חריגים מאוד מסומנים בכוכבית. יצוין כי ביצוע תרשימי הקופסא חל על הנתונים הכמותיים בלבד וכי כלל הנתונים מתייחסים לתוצאות הללו בלבד. את התוצאות ניתן לראות בטבלה הממוקמת בעמוד הבא והקוד שהוזן ליצירת התרשימים נמצא בנספחים.

טבלה 2- ניתוח חריגים

משתנה	תרשים	הסבר
Y Total price including tax	<p>Total Price</p> 	<p>בתרשים ניתן לראות שהחציון נמוך יחסית בעוד שישנם חריגים הגבוהים מהגבול העליון, דבר זה מסביר את הזנב הימני הגדול של ההתפלגות. הנתונים מייצגים סכום קנייה סופי ומהגרף ניתן לראות שישנם מספר קניות יקרות באופן חריג. אנו נחליט שלא לנפות חריגות אלה כדי לשמור מידע גם על קניות גדולות ולשקף את המציאות. ייתכן שהסופר ירצה להבין מה המאפיינים של קניות גדולות ואיך ניתן להגדיל רווחים.</p>

משתנה	תרשים	הסבר
X_6 Unit Price		נראה כי לא קיימים נתונים חריגים במדד זה. הגרף הוא סימטרי ועל כן ניתן להסיק שהמחירים נקבעו באופן מחושב כך שמחירי המוצרים לא יהיו יקרים מאוד או זולים מאוד באופן חריג.
X_7 Number of products purchased by customer		נראה כי לא קיימים נתונים חריגים במדד זה. נראה כי החציון קרוב יותר לרבעון השני, דבר אשר מצביע על טווח ערכים מצומצם.
X_8 tax fee for customer buying		נראה כי הגרף נראה זהה לגרף של Total price (מלבד המספרים), שכן הדבר מתיישב עם כך שה Tax מחושבים מתוך ה Total. כלומר בתרשים ניתן לראות שהחציון נמוך יחסית בעוד שישנם חריגים הגבוהים מהגבול העליון, דבר זה מסביר את הזנב הימני הגדול של ההתפלגות. אנו נחליט שלא לנפות חריגות אלה כדי לשמור מידע גם על קניות גדולות ולשקף את המציאות.
X_{12} Cost of goods sold		נראה כי הגרף נראה זהה לגרף של Total price (מלבד המספרים), שכן הדבר מתיישב עם כך שה Cogs מחושבים מתוך ה Total. כלומר בתרשים ניתן לראות שהחציון נמוך יחסית בעוד שישנם חריגים הגבוהים מהגבול העליון, דבר זה מסביר את הזנב הימני הגדול של ההתפלגות. אנו נחליט שלא לנפות חריגות אלה כדי לשמור מידע גם על קניות גדולות ולשקף את המציאות.
X_{14} Gross income		נראה כי הגרף נראה זהה לגרף של Total price (מלבד המספרים), שכן הדבר מתיישב עם כך שה Gross income מחושבים מתוך ה Total. כלומר בתרשים ניתן לראות שהחציון נמוך יחסית בעוד שישנם חריגים הגבוהים מהגבול העליון, דבר זה מסביר את הזנב הימני הגדול של ההתפלגות. אנו נחליט שלא לנפות חריגות אלה כדי לשמור מידע גם על קניות גדולות ולשקף את המציאות.

משתנה	תרשים	הסבר
X_{15} Rate of costumers	<p>Rate Of Costumers</p> 	<p>נראה כי לא קיימים נתונים חריגים במדד זה. הגרף הוא סימטרי ומייצג את דירוג לקוחות ושביעות רצונם. כלומר ישנם לקוחות שמרוצים יותר ונתנו דירוג גבוה, לעומת לקוחות שמרוצים פחות ונתנו דירוג נמוך.</p>

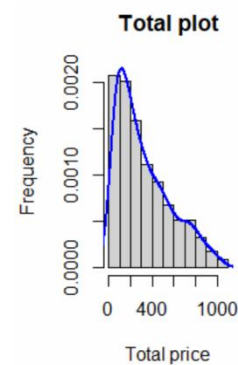
7. פונקציית צפיפות והתפלגות מצטברת (נספח 4)

חלק זה מתייחס לפונקציית הצפיפות וההתפלגות המצטברת של מספר משתנים.

Total price including tax - Y

סוג המשתנה: מוסבר

פ' צפיפות:



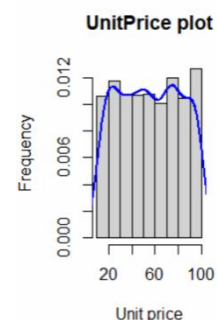
הסבר:

ניתן לראות עבור משתנה שזה שיש זנב ימני ארוך. כלומר ההתפלגות הינה בעלת אסימטריה חיובית. ריכוז הנתונים הגבוה ביותר הוא סביב \$200-\$1 שלפי פ' ההתפלגות המצטברת מהווים 40% מכלל הצרכנים.

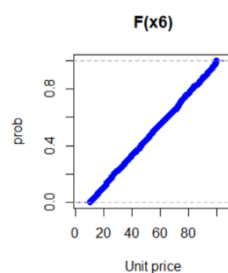
Unit price - X6

סוג המשתנה: מסביר

פ' צפיפות:



פ' ההתפלגות מצטברת:



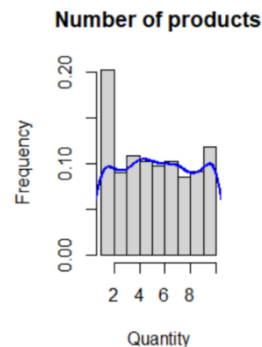
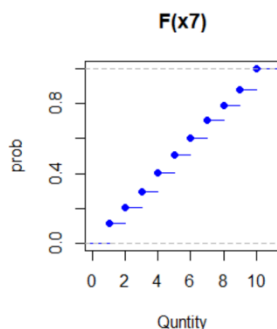
הסבר : נראה כי פ' הצפיפות דומה לפונקציית צפיפות של התפלגות אחידה, כלומר הסיכוי של מחיר של מוצר מסוים להיות בין 1 ל100 הוא אחד, הסיכוי נע בין 0.006 ל0.00125.

Number of products purchased by customer -X7

סוג המשתנה : מסביר

פ' צפיפות :

פ' התפלגות מצטברת :



הסבר :

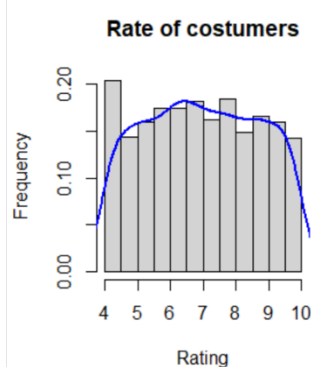
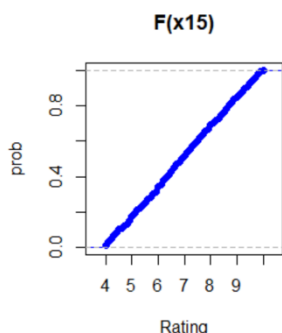
זהו משתנה בדיד ועל כן פ' ההתפלגות המצטברת היא פ' מדרגות. ניתן לראות בפ' הצפיפות שמספר המוצרים הנקנה מאותו סוג לרוב יהיה 1 ושאר כמויות המוצרים הנקנים מאותו סוג הם כמעט זהים. ניתן לראות זאת בפ' ההתפלגות המצטברת בה יש פיזור נתונים אחד ובפ' הצפיפות בה מלבד העמודה של קניית פריט 1, שאר העמודות דומות להתפלגות אחידה.

Rate of costumers- X₁₅

סוג המשתנה : מסביר

פ' צפיפות :

פ' התפלגות מצטברת :



הסבר : נראה כי פ' הצפיפות דומה לפונקציית צפיפות של התפלגות אחידה, כלומר הסיכוי של דירוג להיות בין 4 ל10 הוא יחסית שווה כאשר ניתן לראות שהדירוג הנפוץ ביותר הוא בין 4 ל4 וחצי והדירוגים הכי פחות נפוצים הם בין 9 וחצי ל10 ובין 4 וחצי ל5.

8. ייצוג קשרים בעזרת תרשימים (נספח 5)

החלטנו לבצע תרשימי פיזור משתי זוויות :

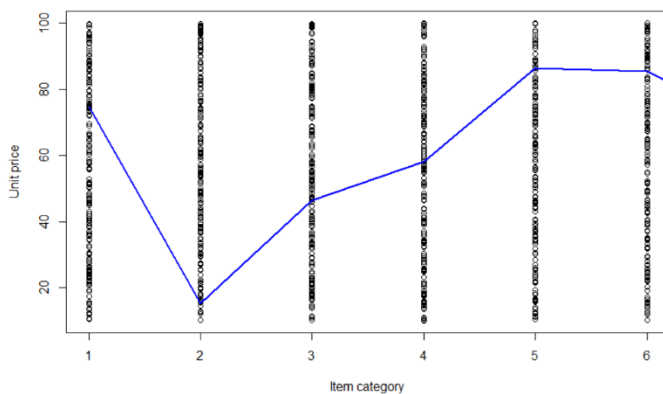
- א. תרשימי פיזור בין המשתנים המסבירים- זאת על מנת להבין האם קיימת תלות בין המשתנים המסבירים ואם כן, כיצד באה לידי ביטוי וכיצד ניתן להסביר אותה.
- ב. תרשימי פיזור בין המשתנה המוסבר לבין המשתנים המסבירים הנבחרים- זאת על מנת לקבל אינטואיציה האם בבסיס הנתונים קיימים משתנים מסבירים אשר אכן מסבירים את המשתנה המוסבר.

ראשית נציג גרפים המתארים קשרים בין המשתנים המסבירים:

פיזור בין קטגוריית המוצר לבין מחיר ליחידת מוצר:

משתנה קטגוריאל "קטגוריות" לפי מספור :

1= אביזרים אלקטרוניים, 2= אביזרי אופנה, 3= אוכל ומשקאות, 4= בריאות ויופי, 5= בית ולייף סטייל, 6= ספורט וטיולים



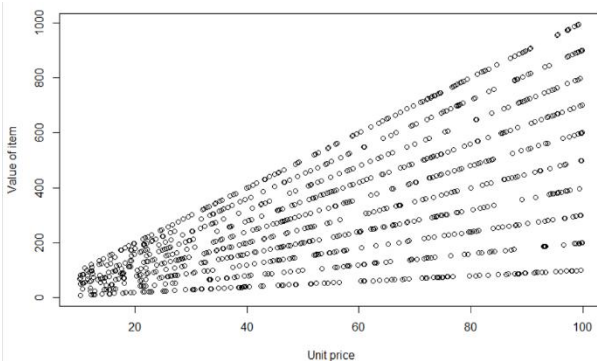
איור 1- פיזור בין קטגוריית מוצר לבין מחיר ליחידת מוצר

הסבר: ניתן לראות מגרף פיזור הנ"ל את קטגוריית המוצרים כפונקציה של המחיר ליחידה. בגרף ניתן לראות את קו המגמה של הקטגוריה כפונקציה של המחיר עליו ניתן להסיק מספק דברים. ראשית, כפי שצפינו עבור קטגוריות יקרות יותר ניתן לראות שהמחיר ליחידת מוצר הוא גבוה יותר, לדוגמא: התפלגות המחירים בקטגוריית ספורט וטיולים היא גבוהה יותר ביחס לשאר הקטגוריות המוצגות בגרף.

פיזור בין שווי הסחורה לבין מחיר

ליחידת מוצר:

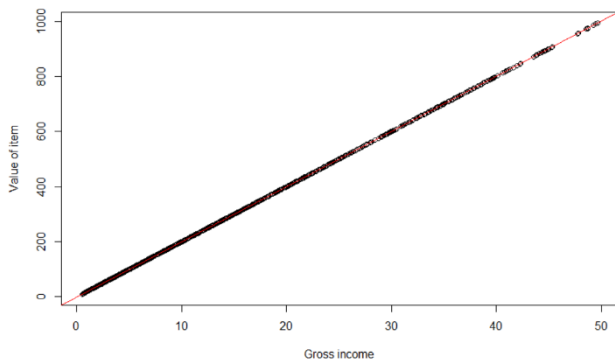
הסבר: ניתן לראות בגרף הפיזור הבא את המחיר ליחידת מוצר כפונקציה של שווי הסחורה. בגרף נראית מגמת עלייה ופיזור ברורה. ככל ששווי הסחורה נמוך יותר, כך פיזור הנקודות צפוף יותר והמחיר ליחידה נמוך יותר. ניתן להסיק מכך שככל הנראה קיימים מוצרים רבים בעלי מחירים נמוכים בבסיס הנתונים של הכלבו לעומת מוצרים בעלי מחירים גבוהים. כפי שצפינו, באמת קיים קשר חיובי בין שווי הסחורה למחיר ליחידת מוצר.



איור 3- פיזור בין שווי הסחורה לבין מחיר ליחידת מוצר

איור 4- פיזור בין שווי הסחורה לבין מחיר ליחידת מוצר

פיזור בין רווח הקנייה ברוטו בדולרים לבין שווי הסחורה:



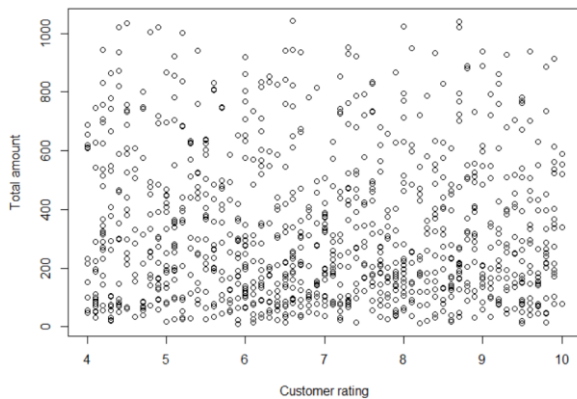
איור 5- פיזור בין רווח הקנייה ברוטו לבין שווי הסחורה

הסבר: ניתן לראות בגרף הפיזור הבא את רווח הקנייה כפונקציה של שווי הסחורה. בגרף זה נראית מגמה לינארית עולה באופן מובהק כלומר, יש קשר חיובי בין המשתנים בדיוק כמו שהיינו מצפים. הקשר בין המשתנים נובע כתוצאה מכך ששווי הסחורה ורווח הקנייה משלימים למחיר הסופי של הרכישה ומהווים ביחד 100% ממנו וזאת כיוון שאחוז הרווח הוא קבוע עבור כל סל קניות.

שנית נציג גרפים המתארים קשרים בין

המשתנה המוסבר למשתנים המסבירים:

פיזור בין דירוג הלקוחות לבין המחיר הסופי של הרכישה:



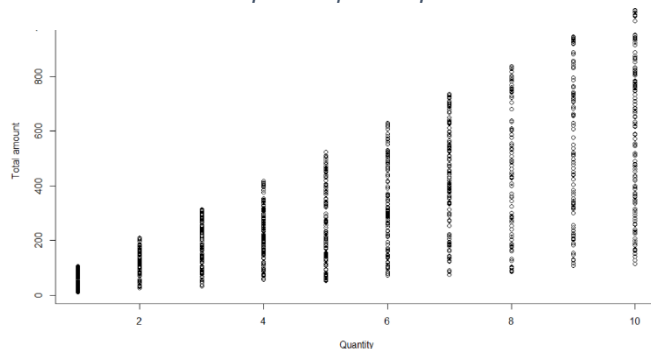
איור 7- פיזור בין דירוג לקוחות לבין המחיר הסופי של הרכישה

הסבר: ניתן לראות בגרף הפיזור הבא את דירוג הלקוחות כפונקציה של המחיר הסופי. הפיזור בגרף זה הוא יחסית גבוה ומכאן ניתן לומר כי לא קיים קשר מובהק או מגמה כלשהי בין שני המשתנים. מהגרף ראינו שקיימים יותר דירוגים כאשר המחיר הסופי של סלי הקניות היו נמוכים מאשר בסלים גבוהים.

פיזור בין מספר המוצרים שנרכשו ע"י הלקוח

לבין המחיר הסופי של הרכישה:

איור 8- פיזור בין דירוג לקוחות לבין המחיר הסופי של הרכישה



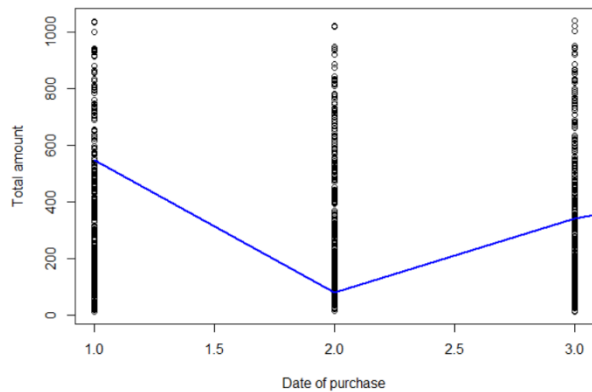
איור 9- פיזור בין מספר המוצרים שנרכשו לבין המחיר הסופי של הרכישה

הסבר: ניתן לראות בגרף הפיזור הבא את מספר המוצרים שנרכשו כפונקציה של המחיר הסופי. הפיזור בכמויות הנמוכות הוא צפוף יותר מאשר בכמויות הגבוהות זאת בגלל שככל שיש פחות מוצרים אז המחיר של הסל יהיה יותר נמוך. זאת בניגוד לציפייה שלנו שיהיה פיזור גם בכמויות הנמוכות (אם למשל לקוח יקנה מוצר שמחירו גבוה מאוד).

כמו כן, ניתן לראות כי ככל שכמות המוצרים בסל עולה כך גם המחיר הסופי עולה ולכן יש קשר חיובי בין שני המשתנים.

איור 10- פיזור בין מספר המוצרים שנרכשו לבין המחיר הסופי של הרכישה

פיזור בין חודש הקנייה לבין המחיר הסופי של הרכישה:



הסבר: ניתן לראות בגרף הפיזור הבא את מספר החודש כפונקציה של המחיר הסופי של סל הקניות. הפיזור במחירים הנמוכים צפוף יותר מהפיזור במחירים הגבוהים. בהסתכלות על קו המגמה נראה שיש קשר בין המחיר הסופי של הסל לבין החודש בו התבצעה הקנייה. כפי שצפינו, ניח כי הבדל זה מעיד על מועדים מיוחדים ומבצעים המתקיימים בחודש זה.

איור 11- פיזור בין חודש הקנייה לבין המחיר הסופי של הרכישה

9. טבלאות שכיחויות (נספח 6)

בסעיף זה מוצגת התייחסות לשתי טבלאות שכיחות מסוגים שונים:

טבלאות שכיחות חד ממדיות, טבלאות אשר מראות לנו ידע כמותי והסתברותי על משתנה יחיד.

טבלאות שכיחות דו ממדיות, אשר מציגות נתונים אודות שני משתנים והקשר ביניהם.

טבלאות חד ממדיות:

טבלת השכיחות החד ממדית מראה לנו את מספר הפעמים להיות בטווח ערכים מסוים וכן את ההסתברות להיות בטווח הנתון.

טבלת השכיחות של מספר המוצרים שנרכשו

טווח ערכים	תדירות	הסתברות
(0,2]	203	0.203
(2,4]	199	0.199
(4,6]	200	0.2
(6,8]	187	0.187
(8,10]	211	0.211

בטבלת שכיחות זו ביצענו איחוד רשומות בקפיצות של 2. טווח הערכים שבו מספר המוצרים הוא הגדול ביותר הוא (8,10] ובכך ניתן להסיק כי סל קניות בעל מספר מוצרים אלה הוא הכי שכיח בקרב המבקרים בכלבו. בנוסף, ניתן לראות שפיזור נתוני הממצאים הוא יחסית שווה בין כל הטווחים.

טבלת השכיחות של דירוג חוויית הקנייה

טווח ערכים	תדירות	הסתברות
(0,1]	0	0

0	0	(1,2]
0	0	(2,3]
0.11	11	(3,4]
0.163	163	(4,5]
0.167	167	(5,6]
0.178	178	(6,7]
0.173	173	(7,8]
0.157	157	(8,9]
0.151	151	(9,10]

שכיחות זו ביצענו בטבלת

איחוד רשומות בקפיצות של 1. ניתן לראות בטבלה שהדירוג הכי נפוץ בקרב מבקרי הכלבו הנו בטווח ערכים (6,7] ומכך ניתן להסיק כי רוב מבקרי הכלבו מרוצים ברמה בינונית מחוויית הקנייה. כמו כן, ניתן לראות מטבלת השכיחות כי הפונקציה דומה באופן יחסי להתפלגות נורמלית א-סימטרית עם זנב שמאלי ופיזור צפוף סביב טווח נתונים גבוה יותר. טבלת שכיחות זו מתיישבת עם המציאות כך שהדירוג בשטח תואם לממצאים שהצגנו והדירוג הוא סביב הממוצע.

טבלאות זו ממדיות:

טבלת השכיחות של שווי הסחורה למול מחיר ליחידת מוצר

תדירות:

מחיר ליחידת מוצר											שווי הסחורה
(90,100]	(80,90]	(70,80]	(60,70]	(50,60]	(40,50]	(30,40]	(20,30]	(10,20]	(0,10]		
13	12	15	11	10	19	32	41	71	0	(0,100]	
14	10	9	14	22	26	29	51	35	0	(100,200]	
12	11	16	20	20	26	24	26	0	0	(200,300]	
13	6	15	15	24	16	23	0	0	0	(300,400]	
10	16	13	16	18	20	0	0	0	0	(400,500]	
13	12	18	12	14	0	0	0	0	0	(500,600]	
14	9	12	13	0	0	0	0	0	0	(600,700]	
14	13	22	0	0	0	0	0	0	0	(700,800]	
13	16	0	0	0	0	0	0	0	0	(800,900]	
11	0	0	0	0	0	0	0	0	0	(900,1000]	

הסתברות:

מחיר ליחידת מוצר											שווי הסחורה
(90,100]	(80,90]	(70,80]	(60,70]	(50,60]	(40,50]	(30,40]	(20,30]	(10,20]	(0,10]		
0.013	0.012	0.015	0.011	0.01	0.019	0.032	0.041	0.071	0	(0,100]	
0.014	0.01	0.009	0.014	0.022	0.026	0.029	0.051	0.035	0	(100,200]	

0.012	0.011	0.016	0.02	0.02	0.026	0.024	0.026	0	0	(200,300]
0.013	0.006	0.015	0.015	0.024	0.016	0.023	0	0	0	(300,400]
0.01	0.016	0.013	0.016	0.018	0.02	0	0	0	0	(400,500]
0.013	0.012	0.018	0.012	0.014	0	0	0	0	0	(500,600]
0.014	0.009	0.012	0.013	0	0	0	0	0	0	(600,700]
0.014	0.013	0.022	0	0	0	0	0	0	0	(700,800]
0.013	0.016	0	0	0	0	0	0	0	0	(800,900]
0.011	0	0	0	0	0	0	0	0	0	(900,1000]

בטבלת שכיחות דו ממדית זו, ניתן לראות את שווי הסחורה למול מחיר ליחידת מוצר.
 בטבלת שכיחות זו ביצענו איחוד רשומות בקפיצות של 100 לשווי הסחורה ובקפיצות של 10 למחיר ליחידת מוצר. ניתן לראות כי פיזור הנתונים של שווי הסחורה גדול יותר ככל שהמחיר ליחידת מוצר גדול יותר. ניתן לראות שכאשר שווי הסחורה הוא בטווח גבוה, לא קיימים מוצרים בטווח זה עם מחיר ליחידת מוצר בטווח מחירים נמוך, דבר המתיישב עם ההיגיון שבמידה ושווי הסחורה גבוה אז לא יתמחרו מוצר במחיר זול יותר כי הכלבו רוצה להרוויח ועל כן רוב עמודות האפס במשולש התחתון.

טבלת השכיחות של דירוג הלקוחות למול מספר המוצרים שנרכשו

תדירות :

דירוג הלקוחות											מספר המוצרים
(9,10]	(8,9]	(7,8]	(6,7]	(5,6]	(4,5]	(3,4]	(2,3]	(1,2]	(0,1]		
21	14	15	25	19	16	2	0	0	0	(0,1]	
11	14	15	20	14	17	0	0	0	0	(1,2]	
15	9	24	15	14	11	2	0	0	0	(2,3]	
14	20	18	25	12	18	2	0	0	0	(3,4]	
20	17	14	17	14	20	0	0	0	0	(4,5]	
16	12	20	15	21	13	1	0	0	0	(5,6]	
14	18	18	16	17	18	1	0	0	0	(6,7]	
13	19	11	16	11	14	1	0	0	0	(7,8]	
13	12	23	11	16	15	2	0	0	0	(8,9]	
14	22	15	18	29	21	0	0	0	0	(9,10]	

הסתברות :

דירוג הלקוחות											מספר המוצרים
[9,10]	[8,9]	[7,8]	[6,7]	[5,6]	[4,5]	[3,4]	[2,3]	[1,2]	[0,1]		
0.021	0.014	0.015	0.025	0.019	0.016	0.002	0	0	0	[0,1]	
0.011	0.014	0.015	0.02	0.014	0.017	0	0	0	0	[1,2]	

0.015	0.009	0.024	0.015	0.014	0.011	0.002	0	0	0	[2,3)	
0.014	0.02	0.018	0.025	0.012	0.018	0.002	0	0	0	[3,4)	
0.02	0.017	0.014	0.017	0.014	0.02	0	0	0	0	[4,5)	
0.016	0.012	0.02	0.015	0.021	0.013	0.001	0	0	0	[5,6)	
0.014	0.018	0.018	0.016	0.017	0.018	0.001	0	0	0	[6,7)	
0.013	0.019	0.011	0.016	0.011	0.014	0.001	0	0	0	[7,8)	
0.013	0.012	0.023	0.011	0.016	0.015	0.002	0	0	0	[8,9)	
0.014	0.022	0.015	0.018	0.029	0.021	0	0	0	0	[9,10)	

בטבלת שכיחות דו ממדית זו, ניתן לראות את דירוג הלקוחות למול מספר המוצרים שנרכשו. בטבלת שכיחות זו ביצענו איחוד רשומות בקפיצות של 1 לדירוג הלקוחות ובקפיצות של 1 למחיר למספר המוצרים שנרכשו. ניתן לראות כי פיזור הנתונים של מספר המוצרים גדול יותר ככל שדירוג הלקוחות עולה. אין קשר מובהק בין שני המשתנים כיוון שאין פיזור ממוקד סביב טווח מסוים, ניתן לראות שאין מגמה מסוימת או דפוס. בנוסף, מהטבלה החד ממדית שיצרנו של דירוגי הלקוחות ראינו כי לא קיימים דירוגים בטווחים הנמוכים ולכן דבר מתיישב עם העובדה כי בטווחים הנמוכים עבור הדירוג אין מוצרים שנרכשו ודורגה חוויית הקנייה (שלושת העמודות הראשונות).

חלק ב' – ניתוח פורמאלי של מאגר הנתונים

1. תקציר מנהלים:

מטרת הפרויקט היא בניית מודל רגרסיה לינארית ושיפורו באמצעות כלים אשר נלמדו בכיתה בעזרת תוכנת ה-R. בפרויקט זה בחרנו לחקור את הגורמים המשפיעים על סכום הרכישה הסופי של סל קניות בכלבו. המשתנה המוסבר הוא הסכום הסופי של הרכישה, המשתנים המסבירים הם כמות מוצרים ברכישה, מחיר ליחידת מוצר, שווי הסחורה של המוצרים ועוד. בחלק לזה בדקנו את כל המשתנים שפורטו בחלק א ובחנו את השפעתם של המשתנים המסבירים על המשנה המוסבר בעזרת מקדם מתאם פירסון ותרשימי פיזור וע"י ניפיו משתנים בעלי מתאם נמוך. כמו כן, במשתנים הקטגוריאליים ביצענו איחוד קטגוריות. ככלל, משתני הרגרסיה הלינארית הינם בעלי אופי רציף, ולכן התאמנו למשתנים הקטגוריאליים משתנה דמה ובחנו משתני אינטראקציה. בנוסף, לאחר הגדרת המודל בדקנו באמצעות אלגוריתמים היוריסטיים מהו מבנה המודל המיטבי. בחלק זה, כלל האלגוריתמים הצביעו על המודל להלן:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_6 x_6 + \hat{\beta}_7 x_7$$

$$R^2_{adj} = 0.8899$$

לאחר הצעת המודל הסופי, בדקנו שאכן ההנחות עליהן מושתתת הרגרסיה הלינארית מתקיימות, הנחת הלינאריות, שוויון שונויות וההתפלגות הנורמלית של השגיאות וגילינו כי השתיים הראשונות לא התקיימו ואילו השלישית כן התקיימה. לאחר מבחנים בבדיקות שצוינו לעיל, בדקנו את יכולותיו של מודל הרגרסיה הלינארית אותו הצענו ע"י לקיחת רשומה מבסיס הנתונים והשוואה של הסכום הסופי האמיתי לסכום הסופי המתקבל מהמודל. לבסוף, על מנת לשפר את המודל, בוצעה בחינת המודל עם טרנספורמציה Box-Cox כאשר $\lambda = 0.37$.

המודל שהתקבל בעקבות הטרנספורמציה זהה למודל המקורי, כלומר מכיל את כמות המוצרים ברכישה ואת המחיר עבור כל יחידת מוצר. כפי שהזכרנו, לא בחרנו במודל זה ונשארו עם המודל המקורי כיוון שכל הנחות המודל לא מתקיימות במודל זה.

2. עיבוד מקדים:

2.1. הסרה של משתנים (נספח 7)

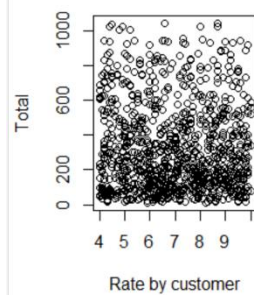
מקדם מתאם –

תחילה בדקנו את מקדם המתאם של פירסון עבור כל משתנה מסביר רציף עם המשתנה המוסבר וכתוצאה מכך הגענו אל מדד טיב ההתאמה של המשתנים. מדד ההתאמה מאפשר לבחון עד כמה המודל מדויק, ועד כמה המשתנה המסביר אכן מסביר את המשתנה המוסבר בהתחשב בכך שערך 1 מייצג קשר חזק וערך שקרוב ל-0 מייצג קשר חלש. המדד אינו רלוונטי עבור המשתנים קטגוריאליים.

variable	Pearson
Unit Price	0.633
Quantity	0.705
Tax	1
Cost Of Goods	1
Gross Income	1

Rating	-0.036
--------	--------

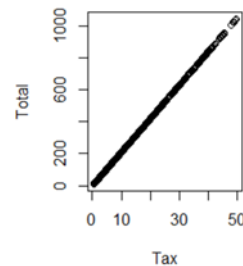
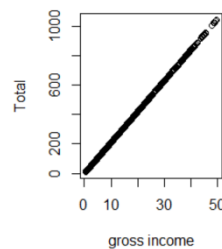
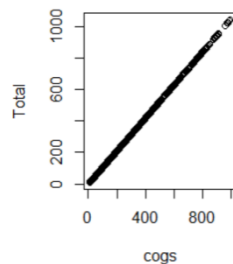
נראה כי מקדם המתאם של פרמטר דירוג הוא היחידי שקרוב לאפס בערך מוחלט ולכן בחרנו לבדוק אותו ולהציג עבורו את תרשים פיזור שלו:



ניתן להתרשם כי לא קיים קשר ליניארי בין דירוג המוצרים על ידי הלקוחות לסכום התשלום הסופי של כל עסקה. בנוסף מקדם המתאם הוא -0.036 - לכן נרצה להסיר משתנה זה.

איור 13- תרשים פיזור בין דירוג המוצרים לסכום הרכישה הסופי

בנוסף נראה כי עבור המשתנים Gross, Cost Of Goods Tax, Income יש התאמה מושלמת בינם לבין המשתנה המוסבר ובנוסף ישנה קורלציה חמורה בין משתנים אלה שכן הם מסבירים את אותו הדבר. הדבר מתיישב עם העובדה שמשתנים אלה "נגזרים" מהסכום הכולל ועל כן נרצה להסירם.



2.2. התאמה של משתנים (נספח 8)

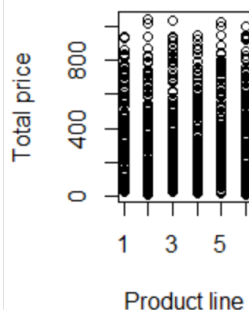
בחלק זה נבדוק אילו קטגוריות ניתן לאחד בעזרת תרשימי הפיזור:

:Product line

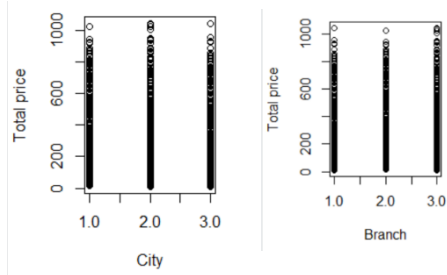
הבחנו כי במשתנה הקטגוריאל Product line ישנו דמיון בין הפיזור של מספר קטגוריות ועל כן החלטנו לאחד קטגוריות אלה. הקטגוריות שהחלטנו לאחד הם:

1=fashion accessories ו-home and lifestyle

2=Food and beverages ו-Sports and travel



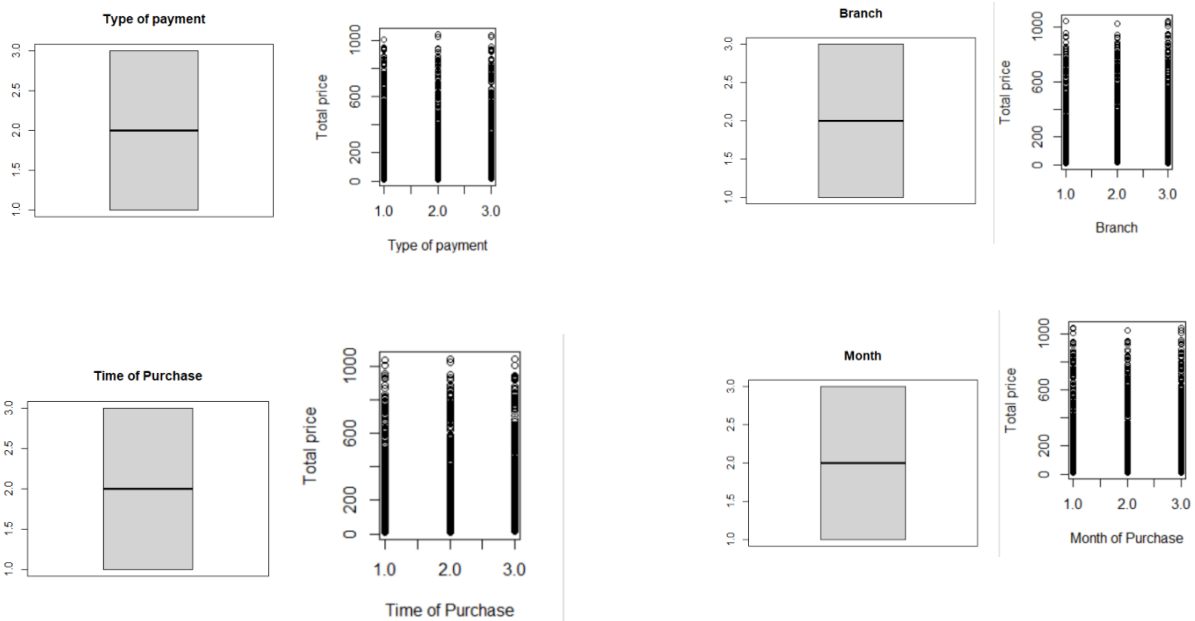
:City



הבחנו כי למשתנה city יש פיזור משתנים זהה לפיזור המשתנים של brunch, דבר זה מתיישב עם כך שבכל עיר יש סוג סניף אחד (בעיר Yangon יש רק את סניף A בעיר Naypyitaw יש רק את סניף C ובעיר Mandalay יש רק את סניף B). כלומר יש הלימה מלאה בין העיר לסניף ועל כן החלטנו להסיר את המשתנה city.

: Time of purchase ,Month ,Type of payment ,Branch

נראה כי עבור משתנים אלה הקטגוריות משפיעות באותה המידה, כלומר המשתנים לא מפלגים את המשתנה המוסבר בצורה שונה, ולכן נרצה להסיר משתנים אלה. ניתן לראות זאת גם בטבלת boxplot.



2.3. משתני דמה (נספח 9)

לאחר התאמת המשתנים הקטגוריאליים נרצה לשלבם ברגרסיה כמשתנים מסבירים. לצורך כך נגדיר משתני דמה. משתני הדמה יצינו את התרומה השולית של המשתנים (אשר לא נמצאים בקבוצת הבסיס) לחותך.

קבוצת הבסיס שלנו :

(A, Normal, Female, Health and beauty, Morning, Cash, January)

- עבור המשתנה הקטגוריאלי Branch נגדיר משתנה דמה כמספר הקטגוריות פחות 1. המשתנה שלנו בעל 3 קטגוריות ולכן 2 משתני דמה :

$$X_{11} = \begin{cases} 1, & \text{If branch is B} \\ 0, & \text{else} \end{cases}$$

$$X_{12} = \begin{cases} 1, & \text{If branch is C} \\ 0, & \text{else} \end{cases}$$

- עבור המשתנה הקטגוריאלי Customer Type נגדיר משתנה דמה כמספר הקטגוריות פחות 1. המשתנה שלנו בעל 2 קטגוריות ולכן נגדיר משתנה דמה אחד :

$$X_{31} = \begin{cases} 1, & \text{If Customer type is member} \\ 0, & \text{else} \end{cases}$$

- עבור המשתנה הקטגוריאלי Gender נגדיר משתנה דמה כמספר הקטגוריות פחות 1. המשתנה שלנו בעל 2 קטגוריות ולכן נגדיר משתנה דמה אחד :

$$X_{41} = \begin{cases} 1, & \text{If Gender is male} \\ 0, & \text{else} \end{cases}$$

- עבור המשתנה הקטגוריאלי Product line נגדיר משתנה דמה כמספר הקטגוריות פחות 1. המשתנה שלנו בעל 4 קטגוריות ולכן נגדיר 3 משתני דמה :

$$X_{51} = \begin{cases} 1, & \text{If Category is Electronic accessories} \\ 0, & \text{else} \end{cases}$$

$$X_{52} = \begin{cases} 1, & \text{If Category is Fashion or Home and life Style} \\ 0, & \text{else} \end{cases}$$

$$X_{53} = \begin{cases} 1, & \text{If Category is Food and beverages or Sport} \\ 0, & \text{else} \end{cases}$$

- עבור המשתנה הקטגוריאלי Date נגדיר משתנה דמה כמספר הקטגוריות פחות 1. המשתנה שלנו בעל 3 קטגוריות ולכן נגדיר 2 משתני דמה :

$$X_{91} = \begin{cases} 1, & \text{If Date is February} \\ 0, & \text{else} \end{cases}$$

$$X_{92} = \begin{cases} 1, & \text{If Date is March} \\ 0, & \text{else} \end{cases}$$

- עבור המשתנה הקטגוריאלי Time נגדיר משתנה דמה כמספר הקטגוריות פחות 1. המשתנה שלנו בעל 3 קטגוריות ולכן נגדיר 2 משתני דמה :

$$X_{101} = \begin{cases} 1, & \text{If Time is Evening} \\ 0, & \text{else} \end{cases}$$

$$X_{102} = \begin{cases} 1, & \text{If Time is Noon} \\ 0, & \text{else} \end{cases}$$

- עבור המשתנה הקטגוריאלי Payment נגדיר משתנה דמה כמספר הקטגוריות פחות 1. המשתנה שלנו בעל 3 קטגוריות ולכן נגדיר 2 משתני דמה :

$$X_{111} = \begin{cases} 1, & \text{If Payment is Ewallet} \\ 0, & \text{else} \end{cases}$$

$$X_{112} = \begin{cases} 1, & \text{If Payment is Credit card} \\ 0, & \text{else} \end{cases}$$

2.4. הוספת משתני אינטראקציה (נספח 10)

הוחלט לבחון את הצורך במשתני האינטראקציה הבאים :

1. זכר/נקבה :

א. כמות המוצרים שנרכשו למול מין הלקוח שרכש אותם.

ב. מחיר המוצרים שנרכשו למול מין הלקוח שרכש אותם.

2. חבר מועדון/רגיל :

א. כמות המוצרים שנרכשו למול סוג הלקוח שרכש אותם.

ב. מחיר המוצרים שנרכשו למול מין הלקוח שרכש אותם.

בחירת אינטראקציות אלו נובעת מהמחשבה שתמהילים אלו עשויים להיות מכפילי כוח

במודל הרגרסיה. באם נבין כי קיים שוני בין הפילוגים השונים של המשתנים

הקטגוריאליים בשילוב עם המשתנים הרציפים שנבחרו, תהיה הצדקה להוספת

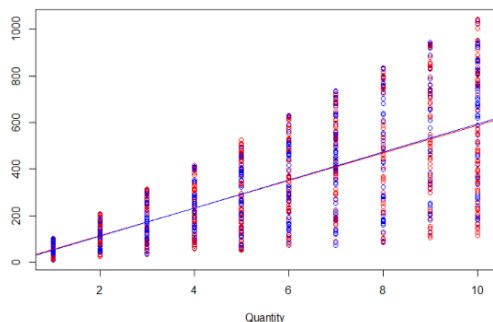
המשתנים הללו למודל. בפרק זה נבחן את הקומבינציות השונות שהוזכרו על מנת להבין

האם מיטיבות עם מודל הרגרסיה בשתי דרכים : תרשימי פיזור ומובהקות משתני

האינטראקציה למול המשתנה המוסבר Y (נספח 8).

בתרשימים שלהלן מוצג הקשר בין משתנה הדמה Gender למשתנים המסבירים
(Male-אדום, Female-כחול):

X_7 - מספר המוצרים שנרכשו ע"י הלקוח ו- X_4 - מגדר הלקוח



איור 15- הקשר בין כמות המוצרים שנרכשו לבין מגדר הלקוח

```
model <- lm(formula = dataset$Total ~ dataset$Quantity*GenderFactor, data = dataset)
summary(model)
```

Call:
 lm(formula = dataset\$Total ~ dataset\$Quantity * GenderFactor, data = dataset)

Residuals:

Min	1Q	Median	3Q	Max
-471.69	-98.44	-0.26	102.26	455.88

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2.1577	17.3364	-0.124	0.901
dataset\$Quantity	58.8930	2.7039	21.780	<2e-16 ***
GenderFactorMale	-3.5792	23.6393	-0.151	0.880
dataset\$Quantity:GenderFactorMale	0.9126	3.7885	0.241	0.810

 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 174.5 on 996 degrees of freedom
 Multiple R-squared: 0.4978, Adjusted R-squared: 0.4963
 F-statistic: 329.1 on 3 and 996 DF, p-value: < 2.2e-16

ניכר כי כמות המוצרים שנרכשו בקנייה משפיע באופן

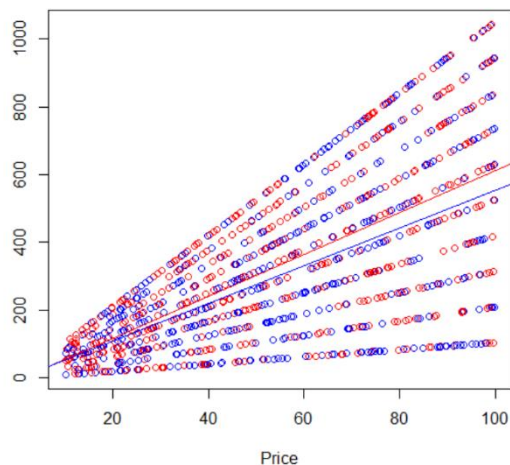
חיובי על הסכום הסופי ברכישה. בנוסף, ניכר כי

מגמת הקווים זהה. עוד ניתן לראות כי ערך P-value

של מקדם האינטראקציה גדול מ0.05 לכן לא מצריך הגדרת משתנה אינטראקציה.

איור 16- הקשר בין כמות המוצרים שנרכשו לבין מגדר הלקוח

X_6 - מחיר ליחידת מוצר ו- X_4 - מגדר הלקוח



איור 17- הקשר בין מחיר ליחידת מוצר לבין מגדר הלקוח

```
model <- lm(formula = dataset$Total ~ dataset$Unit.price*GenderFactor, data = dataset)
summary(model)
```

Call:
 lm(formula = dataset\$Total ~ dataset\$Unit.price * GenderFactor, data = dataset)

Residuals:

Min	1Q	Median	3Q	Max
-505.6	-114.1	-2.9	114.7	490.2

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-7.2050	19.2107	-0.375	0.708
dataset\$Unit.price	6.1939	0.3120	19.855	<2e-16 ***
GenderFactorMale	6.4523	27.9776	0.231	0.818
dataset\$Unit.price:GenderFactorMale	-0.6388	0.4538	-1.408	0.160

 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 189.7 on 996 degrees of freedom
 Multiple R-squared: 0.4066, Adjusted R-squared: 0.4048
 F-statistic: 227.5 on 3 and 996 DF, p-value: < 2.2e-16

ניכר כי המחיר ליחידת מוצר משפיע באופן

חיובי על הסכום הסופי ברכישה. בנוסף ניכר כי

מגמת הקווים פחות או יותר זהה. עוד ניתן לראות כי ערך P-value של מקדם

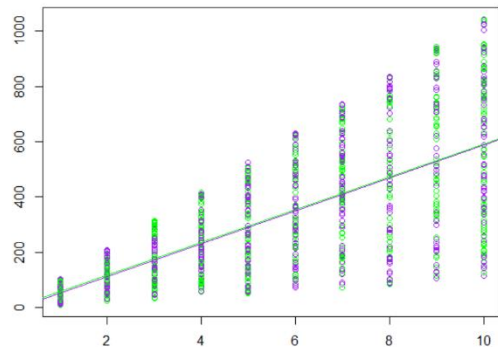
האינטראקציה גדול מ0.05 לכן לא מצריך הגדרת משתנה אינטראקציה.

איור 18- הקשר בין מחיר ליחידת מוצר לבין מגדר הלקוח

בתרשימים שלהלן מוצג הקשר בין משתנה הדמה Customer Type למשתנים המסבירים

(Normal-סגול, Member-ירוק):

X_7 - מספר המוצרים שנרכשו ע"י הלקוח ו- X_3 - סוג הלקוח



```
model <- lm(formula = dataset$Total ~ dataset$Quantity*CustomerTypeFactor, data = dataset)
summary(model)
```

Call:
lm(formula = dataset\$Total ~ dataset\$Quantity * CustomerTypeFactor, data = dataset)

Residuals:

Min	1Q	Median	3Q	Max
-472.71	-98.64	-0.98	99.00	451.72

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-6.3160	16.7293	-0.378	0.706
dataset\$Quantity	59.4110	2.7089	21.932	<2e-16 ***
CustomerTypeFactorMember	4.7429	23.5635	0.201	0.841
dataset\$Quantity:CustomerTypeFactorMember	-0.1609	3.7792	-0.043	0.966

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 174.5 on 996 degrees of freedom
Multiple R-squared: 0.4978, Adjusted R-squared: 0.4963
F-statistic: 329.1 on 3 and 996 DF, p-value: < 2.2e-16

איור 19- הקשר בין מספר המוצרים שנרכשו לבין סוג הלקוח

ניכר כי כמות המוצרים ברכישה משפיע באופן

חיובי על הסכום הסופי של הרכישה. כמו כן,

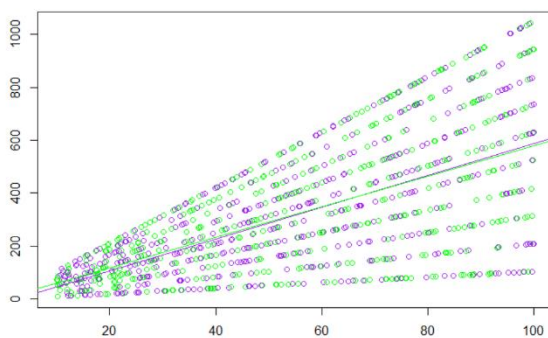
ניתן לראות באמצעות התרשים כי לא קיים

הבדל הן בשיפוע והן בחיתוך עם הצירים עבור מדד זה. בנוסף, ניתן לראות כי ערך P-

value גדול מרמת המובהקות 0.05 ולכן נגיד כי אין צורך להגדיר משתנה אינטראקציה

זה.

X_6 - מחיר ליחידת מוצר ו- X_3 - סוג הלקוח



```
model <- lm(formula = dataset$Total ~ dataset$Unit.price*CustomerTypeFactor, data = dataset)
summary(model)
```

Call:
lm(formula = dataset\$Total ~ dataset\$Unit.price * CustomerTypeFactor, data = dataset)

Residuals:

Min	1Q	Median	3Q	Max
-480.46	-117.50	1.19	117.98	465.96

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-12.2952	19.8383	-0.620	0.536
dataset\$Unit.price	5.9929	0.3249	18.445	<2e-16 ***
CustomerTypeFactorMember	15.4376	28.0418	0.551	0.582
dataset\$Unit.price:CustomerTypeFactorMember	-0.2169	0.4550	-0.477	0.634

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 190.4 on 996 degrees of freedom
Multiple R-squared: 0.4021, Adjusted R-squared: 0.4003
F-statistic: 223.3 on 3 and 996 DF, p-value: < 2.2e-16

איור 21- הקשר בין מחיר ליחידת מוצר לבין סוג הלקוח

ניכר כי המחיר ליחידת מוצר שנרכש משפיע

באופן חיובי על הסכום הסופי הכולל

ברכישה. כמו כן, ניתן לראות באמצעות התרשים כי קיימים הבדלים הן בשיפוע והן

בחיתוך עם הצירים עבור מדד זה. עם זאת, נראה כי מדובר בהבדלים מאוד קטנים. יתרה

מכך, ניכר כי משתנה האינטראקציה לא משפיע בצורה מובהקת על הסכום הסופי הכולל ברכישה ולכן נבחר שלא להגדיר משתנה אינטראקציה זה.

3. התאמת המודל ובדיקת הנחות המודל:

3.1. בחירת משתני המודל (נספח 11)

על מנת לבחור את המודל האופטימלי, השוונו את המודל המלא שקיבלנו למודלים שהתקבלו באמצעות אלגוריתמים היוריסטיים שלמדנו. לצורך בחינת החלופות

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$$

שקיבלנו, ביצענו השוואה באמצעות קריטריון

1. **Backward Elimination**: ההתחלה ממודל מלא המכיל את כל המשתנים. בכל שלב מוציאים את המשתנה הכי פחות מובהק, כלומר בעל הערך הסטטיסטי הקטן ביותר. תוצאות הפלט:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_6 x_6 + \hat{\beta}_7 x_7$$
$$R^2_{adj} = 0.8899$$

2. **Forward Selection**: ההתחלה ממודל ריק ללא משתנים. בכל שלב מכניסים את המשתנה המובהק ביותר, כלומר את הערך הסטטיסטי הגדול ביותר. תוצאות הפלט:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_6 x_6 + \hat{\beta}_7 x_7$$
$$R^2_{adj} = 0.8899$$

3. **Stepwise Regression**: שילוב שתי השיטות לעיל. בכל שלב בודקים האם להכניס או להוציא משתנים אלו שנוספו למודל. תוצאות הפלט:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_6 x_6 + \hat{\beta}_7 x_7$$
$$R^2_{adj} = 0.8899$$

ניתן לראות כי שלושת השיטות הובילו למודל זהה בעל מדד R^2_{adj} זהה ועל כן זהו המודל אותו נבחר.

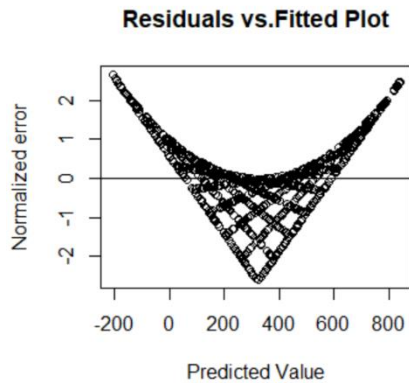
המודל הנבחר:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_6 x_6 + \hat{\beta}_7 x_7$$

3.2. בדיקת הנחות המודל (נספח 12)

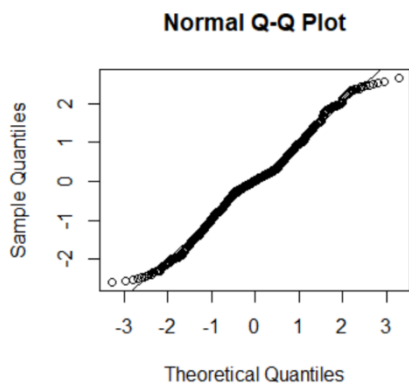
בדיקת הנחת הלינאריות

על מנת לבחון הנחה זו בנינו תרשים פיזור והסתכלנו על התצפיות סביב האפס. ניתן לראות שפיזור השגיאות הוא לא אחיד סביב האפס ודבר זה מעיד על כך שהנחת הלינאריות של המודל לא מתקיימת.



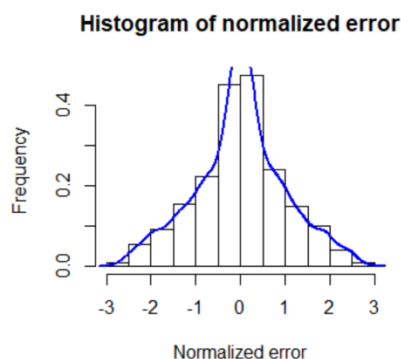
בדיקת הנחת שוויון שוניות:

קיים שוני בין השוניות ככל שמתקדמים על ציר \hat{y} , כלומר עבור כל ערך של המודל שלנו נקבל פיזור שונה של שגיאות. לכן נובע כי הנחת שוויון השוניות אינה נכונה כאן.



בדיקת הנחת הנורמאליות של השגיאות:

על מנת לבדוק האם השגיאות במודל מתפלגות נורמלי, נתבונן ב-QQPLOT שמציג השוואה בין הנתונים בפועל לבין הערכים שהיינו מצפים לראות במידה והם היו מגיעים מהתפלגות נורמלית. מהתבוננות ב-QQPLOT של השגיאות המתוקנות ניתן לראות שהתקבלו נקודות שנמצאות על הקו, מלבד סטייה קלה בקצוות. על פי ההיסטוגרמה ניתן לראות שהשגיאות מתפלגות בצורה הנראית כנורמלית.



בנוסף נבצע מבחן סטטיסטי על מנת לאשש את ההשערה כי השגיאות מפולגות נורמלי. המבחן שנבצע הוא KS, מבחן זה משווה את ערכי המדגם לערכים אשר היו מתקבלים לו הנתונים היו מגיעים מהתפלגות נורמלית. השערת האפס היא שהשגיאות מתפלגות נורמלי לעומת ההשערה האלטרנטיבית שתאמר אחרת.

לאחר ביצוע המבחן, ניתן לראות כי ה-P-Value הוא 0.056, כלומר גדול מ-0.05 ולכן לא נדחה את השערת האפס, ונסיק כי הנחת הנורמאליות מתקיימת.

One-sample Kolmogorov-Smirnov test

```
data: dataset$stan_residuals
D = 0.070991, p-value = 8.387e-05
alternative hypothesis: two-sided
```

3.3. דוגמא לשימוש המודל הנבחר

המודל הנבחר אמור להעריך את המחיר הסופי של סל קניות ברכישה בהינתן הערכים הבאים: מחיר ליחידת מוצר וכמות המוצרים שנרכשו. לאחרונה, הוחלט לבחון את נתוני סל

הקניות של רכישה מסוימת שהתבצעה ע"י לקוח בכלבו. להלן פרטי הרכישה שהתבצעו בחודש פברואר בשעות הצהריים:

מחיר ליחידת מוצר (\$)	כמות המוצרים שנרכשו
54.84	3

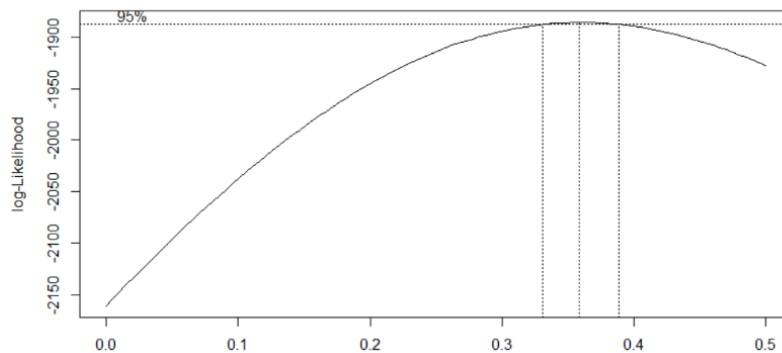
נשתמש במודל שמצאנו על מנת לבחון את המשתנה המוסבר:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_6 x_6 + \hat{\beta}_7 x_7 = -324.52221 + 5.81364 * 54.84 + 58.77155 * 3 = 170.6124576$$

קיבלנו כי המחיר הסופי לפי המודל הוא \$170.612 בעוד שהתוצאה המקורית הנה \$172.746. הטעות בין הערך החזוי לערך האמיתי אינה גדולה. מכאן שהמודל נתן הערכה קרובה לתוחלת המחיר הסופי של רכישה מסוימת על סמך המשתנים המסבירים. ככלל ניתן להשתמש במודל רגרסיה זה כאשר נדרש להבין את המחיר הסופי של סל קניות מסוים. מודל זה יכול לעזור לחברות סופרמרקטים וכלבו בחיזוי מחירים סופיים של רכישות שיכולות להתבצע בהם והן בהבנת הגורמים המשפיעים לכך.

4. שיפור המודל (נספח 13)

לאחר ביצוע התהליכים הקודמים בבניית המודל הסופי, והעובדה שהמודל לא עומד בהנחת שוויון השונויות והנחת הלינאריות, החלטנו לבצע מבחן Box-Cox על מנת לראות איזו טרנספורמציה על המשתנה המוסבר היא המתאימה ביותר. להלן פלט הבדיקה:



איור 23- טרנספורמציה Box-Cox

ניתן לראות כי הערכים 1 ו-0 אינם נמצאים ברווח הסמך בגרף הנייל ולכן, נבחר להשתמש בטרנספורמציה כאשר:

$$\lambda = 0.37$$

משפחת טרנספורמציות החזקה של Box-Cox הנה מהצורה הבאה:

$$y(\lambda) = \begin{cases} (y^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \ln(y), & \lambda = 0 \end{cases}$$

נציב את $\lambda = 0.37$ בנוסחה ונבצע טרנספורמציה על המשתנה המוסבר בשביל לנסות להשפיע על המודל ולשפרו.

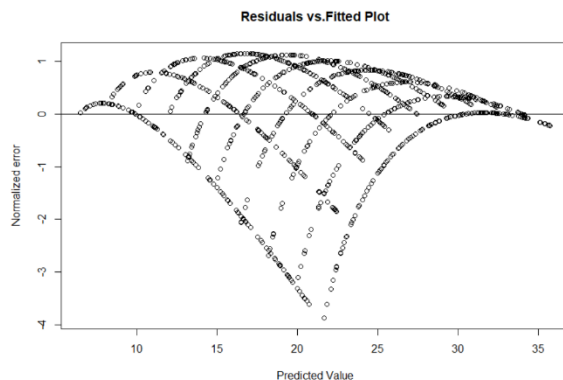
הטרנספורמציה	R^2_{adj}
$y(\lambda)$	0.9458

נשתמש בשיטת רגרסיה לאחר על מנת לבדוק את טיב המודל באמצעות הטרנספורמציה. לאחר ביצוע רגרסיה לאחר קיבלנו מודל זהה, אך מצד שני קיבלנו מדד R^2_{adj} גבוה יותר ועל כן מדובר במודל מדויק יותר.

נבדוק האם יש שינוי בהנחות המודל:

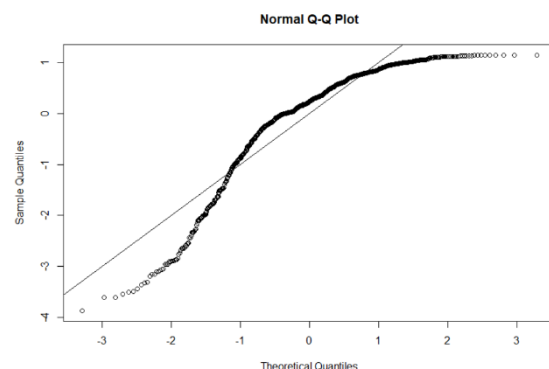
הנחת שוויון השונות:

ניתן לראות ביחס למודל הקודם ולגרף שהציג כי הגרף השתפר בפריסתו אך עדיין לא נוכל להגיד כי יש שוויון שונות סביב הציר.



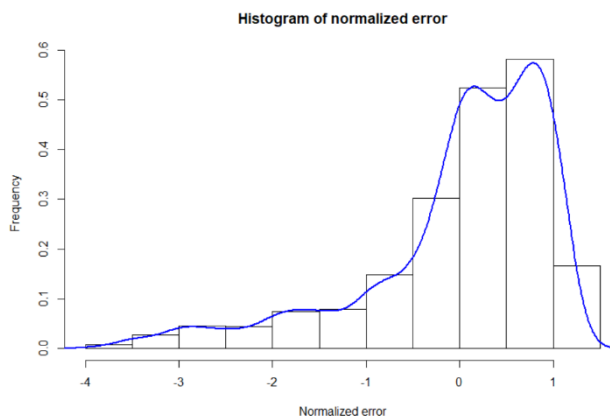
הנחת הלינאריות:

ניתן לראות ביחס למודל הקודם כי דווקא הנחה זו לא השתפרה ואפילו ניראה כי הגרף פחות לינארי. מכאן ניתן להגיד כי לא נוכל להסיק שההנחה מתקיימת.



הנחת הנורמליות:

ניתן לראות מהגרף הבא כי בהשוואה למודל הקודם, הגרף לא דומה לגרף התפלגות נורמלית. כמו כן, לאחר ביצוע מבחן קולמוגורוב-סמירנוף ניתן לראות כי התוצאה מובהקת ולכן נדחה את השערת האפס כי המודל נורמלי.



```
One-sample Kolmogorov-Smirnov test
```

```
data: dataset$stan_residuals  
D = 0.1516, p-value < 2.2e-16  
alternative hypothesis: two-sided
```

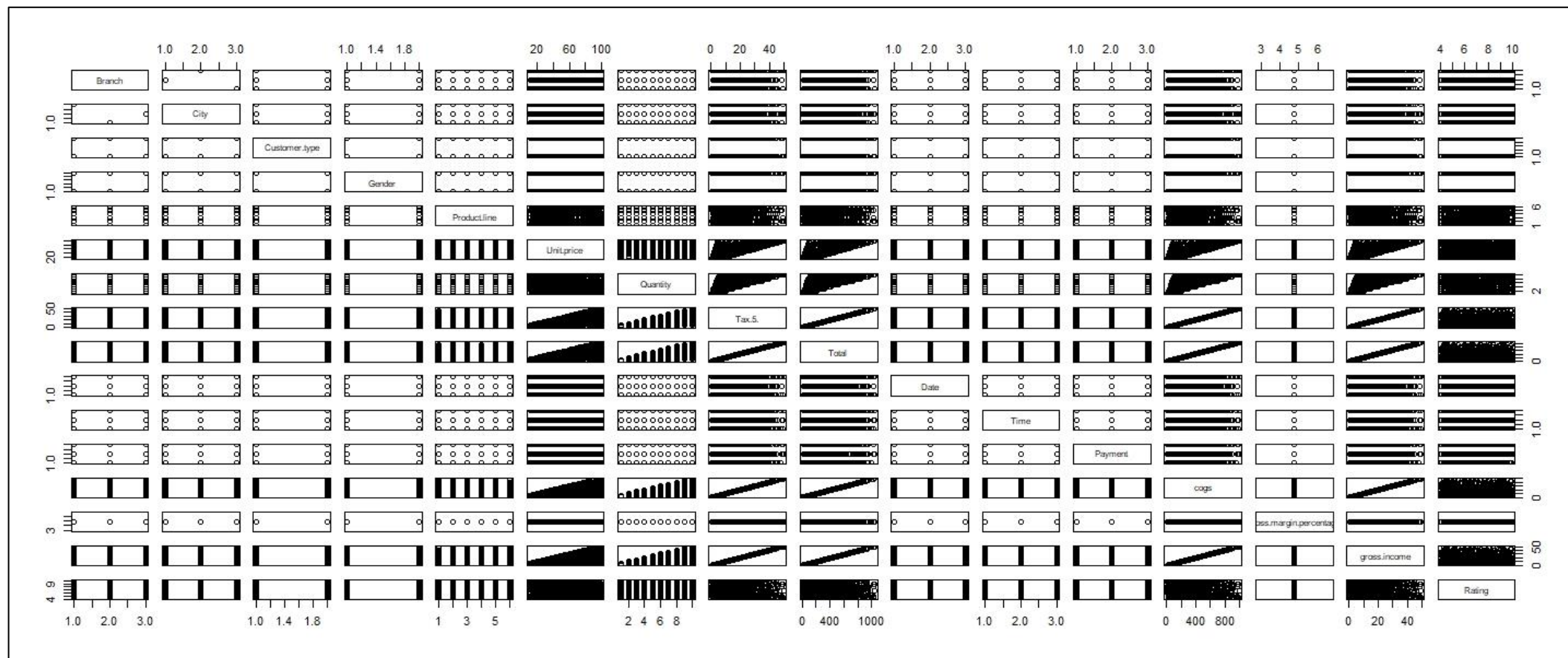
ניתן להסיק מתוצאות טרנספורמציות החזקה של Box-Cox כי אומנם מדד R^2_{adj} משתפר אך מצד שני, אף אחת מהנחות המודל לא מתקיימות בהשוואה למודל הקיים שבו הנחת הנורמליות התקיימה, ולכן נעדיף להישאר עם המודל המקורי:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_6 x_6 + \hat{\beta}_7 x_7$$

נספחים

1. חלק א' סעיף 4

```
dataset <- read.csv(file.choose(), header = T)
plot(dataset)
```



2. חלק א' סעיף 5

```
> summary(data)
Invoice.ID      Branch      City      Customer.type
Length:1000    Length:1000    Length:1000    Length:1000
Class :character Class :character Class :character Class :character
Mode :character Mode :character Mode :character Mode :character

Gender      Product.line      Unit.price      Quantity
Length:1000 Length:1000      Min. :10.08      Min. : 1.00
Class :character Class :character 1st Qu.:32.88      1st Qu.: 3.00
Mode :character Mode :character Median :55.23      Median : 5.00
Mean :55.67      Mean : 5.51
3rd Qu.:77.94      3rd Qu.: 8.00
Max. :99.96      Max. :10.00

Tax.5.      Total      Date      Time
Min. : 0.5085 Min. : 10.68 Length:1000 Length:1000
1st Qu.: 5.9249 1st Qu.: 124.42 Class :character Class :character
Median :12.0880 Median : 253.85 Mode :character Mode :character
Mean :15.3794 Mean : 322.97
3rd Qu.:22.4453 3rd Qu.: 471.35
Max. :49.6500 Max. :1042.65

Payment      cogs      gross.margin.percentage gross.income
Length:1000 Min. : 10.17 Min. :4.762 Min. : 0.5085
Class :character 1st Qu.:118.50 1st Qu.:4.762 1st Qu.: 5.9249
Mode :character Median :241.76 Median :4.762 Median :12.0880
Mean :307.59 Mean :4.762 Mean :15.3794
3rd Qu.:448.90 3rd Qu.:4.762 3rd Qu.:22.4453
Max. :993.00 Max. :4.762 Max. :49.6500
```

```
Rating
Min. : 4.000
1st Qu.: 5.500
Median : 7.000
Mean : 6.973
3rd Qu.: 8.500
Max. :10.000
```



```

> skewness(data$Total)
[1] 0.8898939
> sd(data$Total)
[1] 245.8853
> View(data)
> View(data)
> ##Y
> skewness(data$Total)
[1] 0.8898939
> sd(data$Total)
[1] 245.8853
> ##x6
> skewness(data$Unit.price)
[1] 0.00705623
> sd(data$Unit.price)
[1] 26.49463
> ##x7
> skewness(data$Quantity)
[1] 0.01290225
> sd(data$Quantity)
[1] 2.923431
> ##x8
> skewness(data$Tax.5.)
[1] 0.8898939
> sd(data$Tax.5.)
[1] 11.70883
> ##x12
> skewness(data$cogs)
[1] 0.8898939
> sd(data$cogs)
[1] 234.1765
> ##x14
> skewness(data$gross.income)
[1] 0.8898939
> sd(data$gross.income)
[1] 11.70883
> ##x15
> skewness(data$Rating)
[1] 0.008982638
> sd(data$Rating)
[1] 1.71858

```

3. חלק א' סעיף 6

```

boxTotalPrice<-boxplot(data$Total, main='Total Price')
boxUnitPrice<-boxplot(data$Unit.price, main='Unit Price')
boxNumberOfProduct<-boxplot(data$Quantity, main='Number Of Product')
boxTax<-boxplot(data$Tax.5., main='Tax')
boxCogs<-boxplot(data$cogs, main='Cogs')
boxGrossIncome<-boxplot(data$gross.income, main='Gross income')
boxRateOfCostumers <-boxplot(data$Rating, main='Rate Of Costumers')

```

4. חלק א' סעיף 7

```
##y
hist(data$Total, prob=TRUE,xlab= "Total price", ylab="Frequency", main='Total plot')
lines(density(data$Total), col= 'blue', lwd=2)
plot.ecdf(data$Total, prob=TRUE, xlab= "Total price", ylab="prob", main="F(y)", col='blue' )

##x6
hist(data$Unit.price, prob=TRUE,xlab= "Unit price", ylab="Frequency",main='UnitPrice plot')
lines(density(data$Unit.price), col= 'blue', lwd=2)
plot.ecdf(data$Unit.price, prob=TRUE, xlab= "Unit price", ylab="prob", main="F(x6)", col='blue' )

##x7
hist(data$Quantity, prob=TRUE,xlab= "Quantity", ylab="Frequency",main='Number of products')
lines(density(data$Quantity), col= 'blue', lwd=2)
plot.ecdf(data$Quantity, prob=TRUE, xlab= "Quantity", ylab="prob", main="F(x7)", col='blue' )

##x15
hist(data$Rating, prob=TRUE,xlab= "Rating", ylab="Frequency",main='Rate of costumers')
lines(density(data$Rating), col= 'blue', lwd=2)
plot.ecdf(data$Rating, prob=TRUE, xlab= "Rating", ylab="prob", main="F(x15)", col='blue')
```

5. חלק א' סעיף 8

```
## Category items VS price of items

VectorCategory <- as.factor(dataset$Product.line)
levels(VectorCategory) <- 1:length(levels(VectorCategory))
VectorCategory <- as.numeric(VectorCategory)%>%print()

plot(x = VectorCategory, y = dataset$Unit.price, xlab = "Item category", ylab = "Unit price")
lines(dataset$Unit.price, col = "blue", lwd = 2)
```

```
## Price of items VS value of goods

plot(x = dataset$Unit.price, y = dataset$cogs, xlab = "Unit price", ylab = "Value of item")
```

```
## Value of goods VS Gross income

plot(x = dataset$gross.income, y = dataset$cogs, xlab = "Gross income", ylab = "Value of item")
abline(lm(dataset$cogs ~ dataset$gross.income), col = "red", lwd=1)
```

```
## Total VS customer rate

plot(x = dataset$Rating, y = dataset$Total, xlab = "Customer rating", ylab = "Total amount")
```

```
## Total VS Quantity

plot(x = dataset$Quantity, y = dataset$Total, xlab = "Quantity", ylab = "Total amount")
```

```
## Total VS Month

plot(x = dataset$Date, y = dataset$Total, xlab = "Date of purchase", ylab = "Total amount")
lines(dataset$Total, col = "blue", lwd = 2)
abline(lm(dataset$Total ~ dataset$Date), col = "red", lwd=1)
```

6. חלק א' סעיף 9

טבלת השכיחות של מספר המוצרים שנרכשו:

```
> table(cut(dataset$Quantity, breaks = seq(0,10,2)))

(0,2] (2,4] (4,6] (6,8] (8,10]
  203   199   200   187   211
```

טבלת השכיחות של דירוג הלקוחות:

```
> table(cut(dataset$Rating, breaks = seq(0,10,1)))
```

(0,1]	(1,2]	(2,3]	(3,4]	(4,5]	(5,6]	(6,7]	(7,8]	(8,9]	(9,10]
0	0	0	11	163	167	178	173	157	151

טבלת השכיחות של שווי הסחורה למול מחיר ליחידת מוצר:

```
> cbind(Freq=table(cut(dataset$cogs, breaks = seq(0,1000, 100)), cut(dataset$Unit.price, breaks = seq(0,100,10))),
+       relative= prop.table(table(cut(dataset$cogs, breaks = seq(0,1000, 100)), cut(dataset$Unit.price, breaks = seq(0,100,10)))))
```

	(0,10]	(10,20]	(20,30]	(30,40]	(40,50]	(50,60]	(60,70]	(70,80]	(80,90]	(90,100]	(0,10]	(10,20]	(20,30]	(30,40]	(40,50]	(50,60]
(0,100]	0	71	41	32	19	10	11	15	12	13	0	0.071	0.041	0.032	0.019	0.010
(100,200]	0	35	51	29	26	22	14	9	10	14	0	0.035	0.051	0.029	0.026	0.022
(200,300]	0	0	26	24	26	20	20	16	11	12	0	0.000	0.026	0.024	0.026	0.020
(300,400]	0	0	0	23	16	24	15	15	6	13	0	0.000	0.000	0.023	0.016	0.024
(400,500]	0	0	0	0	20	18	16	13	16	10	0	0.000	0.000	0.000	0.020	0.018
(500,600]	0	0	0	0	0	14	12	18	12	13	0	0.000	0.000	0.000	0.000	0.014
(600,700]	0	0	0	0	0	0	13	12	9	14	0	0.000	0.000	0.000	0.000	0.000
(700,800]	0	0	0	0	0	0	0	22	13	14	0	0.000	0.000	0.000	0.000	0.000
(800,900]	0	0	0	0	0	0	0	0	16	13	0	0.000	0.000	0.000	0.000	0.000
(900,1e+03]	0	0	0	0	0	0	0	0	0	11	0	0.000	0.000	0.000	0.000	0.000

	(60,70]	(70,80]	(80,90]	(90,100]
(0,100]	0.011	0.015	0.012	0.013
(100,200]	0.014	0.009	0.010	0.014
(200,300]	0.020	0.016	0.011	0.012
(300,400]	0.015	0.015	0.006	0.013
(400,500]	0.016	0.013	0.016	0.010
(500,600]	0.012	0.018	0.012	0.013
(600,700]	0.013	0.012	0.009	0.014
(700,800]	0.000	0.022	0.013	0.014
(800,900]	0.000	0.000	0.016	0.013
(900,1e+03]	0.000	0.000	0.000	0.011

טבלת השכיחות של דירוג הלקוחות למול מספר המוצרים שנרכשו:

```
> cbind(Freq=table(cut(dataset$Quantity, breaks = seq(0,10, 1)), cut(dataset$Rating, breaks = seq(0,10,1))),
+       relative= prop.table(table(cut(dataset$Quantity, breaks = seq(0,10, 1)), cut(dataset$Rating, breaks = seq(0,10,1)))))
```

	(0,1]	(1,2]	(2,3]	(3,4]	(4,5]	(5,6]	(6,7]	(7,8]	(8,9]	(9,10]	(0,1]	(1,2]	(2,3]	(3,4]	(4,5]	(5,6]	(6,7]	(7,8]	(8,9]	(9,10]
(0,1]	0	0	0	2	16	19	25	15	14	21	0	0	0	0.002	0.016	0.019	0.025	0.015	0.014	0.021
(1,2]	0	0	0	0	17	14	20	15	14	11	0	0	0	0.000	0.017	0.014	0.020	0.015	0.014	0.011
(2,3]	0	0	0	2	11	14	15	24	9	15	0	0	0	0.002	0.011	0.014	0.015	0.024	0.009	0.015
(3,4]	0	0	0	2	18	12	25	18	20	14	0	0	0	0.002	0.018	0.012	0.025	0.018	0.020	0.014
(4,5]	0	0	0	0	20	14	17	14	17	20	0	0	0	0.000	0.020	0.014	0.017	0.014	0.017	0.020
(5,6]	0	0	0	1	13	21	15	20	12	16	0	0	0	0.001	0.013	0.021	0.015	0.020	0.012	0.016
(6,7]	0	0	0	1	18	17	16	18	18	14	0	0	0	0.001	0.018	0.017	0.016	0.018	0.018	0.014
(7,8]	0	0	0	1	14	11	16	11	19	13	0	0	0	0.001	0.014	0.011	0.016	0.011	0.019	0.013
(8,9]	0	0	0	2	15	16	11	23	12	13	0	0	0	0.002	0.015	0.016	0.011	0.023	0.012	0.013
(9,10]	0	0	0	0	21	29	18	15	22	14	0	0	0	0.000	0.021	0.029	0.018	0.015	0.022	0.014

7. חלק ב' סעיף 2.1

מתאם פירסון:

```
dataset<- read.csv(file.choose(),header=T)
DBContinuous<-subset(dataset,select=c( Unit.price, Quantity, Tax.5., Total, cogs,gross.income,Rating))
res<- cor(DBContinuous)
```



```
plot(x=dataset$Rating,y=dataset$Total,xlab= "Rate by customer", ylab="Total")
plot(x=dataset$Tax.5.,y=dataset$Total,xlab= "Tax", ylab="Total")
plot(x=dataset$gross.income,y=dataset$Total,xlab= "gross income", ylab="Total")
plot(x=dataset$cogs,y=dataset$Total,xlab= "cogs", ylab="Total")
```

	Unit.price	Quantity	Tax.5.	Total	cogs	gross.income	Rating
Unit.price	1.000000000	0.01077756	0.6339621	0.6339621	0.6339621	0.6339621	-0.008777507
Quantity	0.010777564	1.00000000	0.7055102	0.7055102	0.7055102	0.7055102	-0.015814905
Tax.5.	0.633962089	0.70551019	1.0000000	1.0000000	1.0000000	1.0000000	-0.036441705
Total	0.633962089	0.70551019	1.0000000	1.0000000	1.0000000	1.0000000	-0.036441705
cogs	0.633962089	0.70551019	1.0000000	1.0000000	1.0000000	1.0000000	-0.036441705
gross.income	0.633962089	0.70551019	1.0000000	1.0000000	1.0000000	1.0000000	-0.036441705
Rating	-0.008777507	-0.01581490	-0.0364417	-0.0364417	-0.0364417	-0.0364417	1.000000000

8. חלק ב' סעיף 2.2

קוד תרשימי פיזור של משתנים קטגוריאליים:

```
dataset_Numric <- data.matrix(dataset)

boxBranch<-boxplot(x=dataset_Numric[,1], y=dataset_Numric[,9], main='Branch')
boxCity<-boxplot(x=dataset_Numric[,2], y=dataset_Numric[,9], main='City')
boxCustomerType<-boxplot(x=dataset_Numric[,3], y=dataset_Numric[,9], main='Customer Type')
boxGender<-boxplot(x=dataset_Numric[,4], y=dataset_Numric[,9], main='Gender')
boxProductLine<-boxplot(x=dataset_Numric[,5], y=dataset_Numric[,9], main='Product Line')
boxMonth<-boxplot(x=dataset_Numric[,10], y=dataset_Numric[,9], main='Month')
boxTimeOfPurchase <-boxplot(x=dataset_Numric[,11], y=dataset_Numric[,9], main='Time of Purchase')
boxTypeOfPayment <-boxplot(x=dataset_Numric[,12], y=dataset_Numric[,9], main='Type of payment')

plot(x=dataset_Numric[,1], y=dataset_Numric[,9], xlab= "Branch", ylab="Total price")
plot(x=dataset_Numric[,2], y=dataset_Numric[,9], xlab= "City", ylab="Total price")
plot(x=dataset_Numric[,3], y=dataset_Numric[,9], xlab= "Customer Type", ylab="Total price")
plot(x=dataset_Numric[,4], y=dataset_Numric[,9], xlab= "Gender", ylab="Total price")
plot(x=dataset_Numric[,5], y=dataset_Numric[,9], xlab= "Product line", ylab="Total price")
plot(x=dataset_Numric[,10], y=dataset_Numric[,9], xlab= "Month of Purchase", ylab="Total price")
plot(x=dataset_Numric[,11], y=dataset_Numric[,9], xlab= "Time of Purchase", ylab="Total price")
plot(x=dataset_Numric[,12], y=dataset_Numric[,9], xlab= "Type of payment", ylab="Total price")
```

9. חלק ב' סעיף 2.3

```
BranchFactor <- factor(dataset$Branch)
levels(BranchFactor)
BranchFactor <- relevel(BranchFactor, ref = c('A'))

CustomerTypeFactor <- factor(dataset$Customer.type)
levels(CustomerTypeFactor)
CustomerTypeFactor <- relevel(CustomerTypeFactor, ref = c('Normal'))

GenderFactor <- factor(dataset$Gender)
levels(GenderFactor)
GenderFactor <- relevel(GenderFactor, ref = c('Female'))

ProductTypeFactor <- factor(dataset$Product.line)
levels(ProductTypeFactor)
ProductTypeFactor <- relevel(ProductTypeFactor, ref = c('Health and beauty'))

TimeFactor <- factor(dataset$Time)
levels(TimeFactor)
TimeFactor <- relevel(TimeFactor, ref = c('Morning'))

PaymentFactor <- factor(dataset$Payment)
levels(PaymentFactor)
PaymentFactor <- relevel(PaymentFactor, ref = c('Cash'))

DateFactor <- factor(dataset$Date)
levels(DateFactor)
DateFactor <- relevel(DateFactor, ref = c('1'))
```

10. חלק ב' סעיף 2.4

קוד R עבור משתנה דמה מין הלקוח:

```
GenderFactor <- factor(dataset$Gender)
levels(GenderFactor)
GenderFactor <- relevel(GenderFactor, ref = c('Female'))
coloring <- ifelse(dataset$Gender == "Female", c("red"), c("blue"))
plot(x=dataset$Quantity, y=dataset$Total, xlab = "Quantity", ylab = "Total price", col=coloring)
abline(lm(dataset$Total ~ dataset$Quantity, data = dataset, subset = coloring=="red"), col="red")
abline(lm(dataset$Total ~ dataset$Quantity, data = dataset, subset = coloring=="blue"), col="blue")

model <- lm(formula = dataset$Total ~ dataset$Quantity*GenderFactor, data = dataset)
summary(model)

coloring <- ifelse(dataset$Gender == "Female", c("red"), c("blue"))
plot(x=dataset$Unit.price, y=dataset$Total, xlab = "Price", ylab = "Total price", col=coloring)
abline(lm(dataset$Total ~ dataset$Unit.price, data = dataset, subset = coloring=="red"), col="red")
abline(lm(dataset$Total ~ dataset$Unit.price, data = dataset, subset = coloring=="blue"), col="blue")

model <- lm(formula = dataset$Total ~ dataset$Unit.price*GenderFactor, data = dataset)
summary(model)
```

קוד R עבור משתנה דמה סוג הלקוח:

```
CustomerTypeFactor <- factor(dataset$Customer.type)
levels(CustomerTypeFactor)
CustomerTypeFactor <- relevel(CustomerTypeFactor, ref = c('Normal'))
coloring <- ifelse(dataset$Customer.type == "Normal", c("purple"), c("green"))
plot(x=dataset$Quantity, y=dataset$Total, xlab = "Quantity", ylab = "Total price", col=coloring)
abline(lm(dataset$Total ~ dataset$Quantity, data = dataset, subset = coloring=="purple"), col="purple")
abline(lm(dataset$Total ~ dataset$Quantity, data = dataset, subset = coloring=="green"), col="green")

model <- lm(formula = dataset$Total ~ dataset$Quantity*CustomerTypeFactor, data = dataset)
summary(model)

coloring <- ifelse(dataset$Customer.type == "Normal", c("purple"), c("green"))
plot(x=dataset$Unit.price, y=dataset$Total, xlab = "Price", ylab = "Total price", col=coloring)
abline(lm(dataset$Total ~ dataset$Unit.price, data = dataset, subset = coloring=="purple"), col="purple")
abline(lm(dataset$Total ~ dataset$Unit.price, data = dataset, subset = coloring=="green"), col="green")

model <- lm(formula = dataset$Total ~ dataset$Unit.price*CustomerTypeFactor, data = dataset)
summary(model)
```

11. חלק ב' סעיף 3.1

המודל המלא:

```
Call:
lm(formula = dataset$Total ~ BranchFactor + CustomerTypeFactor +
    GenderFactor + ProductTypeFactor + TimeFactor + PaymentFactor +
    dataset$Unit.price + dataset$Quantity, data = dataset)

Residuals:
    Min       1Q   Median       3Q      Max
-224.975  -47.886   2.657   46.860  216.549

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -325.8538    12.6845  -25.689  <2e-16 ***
BranchFactorB         1.7511     6.3349   0.276   0.782
BranchFactorC         7.9187     6.3897   1.239   0.216
CustomerTypeFactorMember -3.4846     5.2061  -0.669   0.503
GenderFactorMale     -2.9316     5.2313  -0.560   0.575
ProductTypeFactorElectronic accessories -1.4921     9.1645  -0.163   0.871
ProductTypeFactorFashion accessories    0.1460     9.0808   0.016   0.987
ProductTypeFactorFood and beverages    0.4745     9.1154   0.052   0.958
ProductTypeFactorHome and lifestyle    6.8565     9.3027   0.737   0.461
ProductTypeFactorSports and travel    0.6665     9.2323   0.072   0.942
TimeFactorEvening    -5.2701     6.4725  -0.814   0.416
TimeFactorNoon      -2.5552     6.3203  -0.404   0.686
PaymentFactorCredit card 10.4476     6.4533   1.619   0.106
PaymentFactorEwallet   0.2338     6.2696   0.037   0.970
dataset$Unit.price    5.8179     0.0981  59.306  <2e-16 ***
dataset$Quantity     58.6467     0.8934  65.645  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 81.83 on 984 degrees of freedom
Multiple R-squared:  0.8909,    Adjusted R-squared:  0.8892
F-statistic: 535.7 on 15 and 984 DF, p-value: < 2.2e-16
```

רגרסיה לאחור:

Start: AIC=8825.19
dataset\$Total ~ BranchFactor + CustomerTypeFactor + GenderFactor +
ProductTypeFactor + TimeFactor + PaymentFactor + dataset\$Unit.price +
dataset\$Quantity

	Df	Sum of Sq	RSS	AIC
- ProductTypeFactor	5	6834	6596023	8816.2
- TimeFactor	2	4440	6593629	8821.9
- BranchFactor	2	11271	6600460	8822.9
- GenderFactor	1	2103	6591292	8823.5
- CustomerTypeFactor	1	3000	6592189	8823.6
- PaymentFactor	2	22514	6611704	8824.6
<none>			6589189	8825.2
- dataset\$Unit.price	1	23552201	30141391	10343.7
- dataset\$Quantity	1	28855847	35445036	10505.7

Step: AIC=8816.22
dataset\$Total ~ BranchFactor + CustomerTypeFactor + GenderFactor +
TimeFactor + PaymentFactor + dataset\$Unit.price + dataset\$Quantity

	Df	Sum of Sq	RSS	AIC
- TimeFactor	2	4281	6600304	8812.9
- BranchFactor	2	10540	6606564	8813.8
- GenderFactor	1	2144	6598168	8814.5
- CustomerTypeFactor	1	2747	6598770	8814.6
- PaymentFactor	2	22198	6618221	8815.6
<none>			6596023	8816.2
- dataset\$Unit.price	1	23618781	30214805	10336.1
- dataset\$Quantity	1	29054712	35650735	10501.5

Step: AIC=8812.87
dataset\$Total ~ BranchFactor + CustomerTypeFactor + GenderFactor +
PaymentFactor + dataset\$Unit.price + dataset\$Quantity

	Df	Sum of Sq	RSS	AIC
- BranchFactor	2	10095	6610399	8810.4
- GenderFactor	1	2321	6602625	8811.2
- CustomerTypeFactor	1	2583	6602887	8811.3
- PaymentFactor	2	22409	6622713	8812.3
<none>			6600304	8812.9
- dataset\$Unit.price	1	23671807	30272112	10334.0
- dataset\$Quantity	1	29254801	35855105	10503.2

Step: AIC=8810.4
dataset\$Total ~ CustomerTypeFactor + GenderFactor + PaymentFactor +
dataset\$Unit.price + dataset\$Quantity

	Df	Sum of Sq	RSS	AIC
- CustomerTypeFactor	1	2405	6612804	8808.8
- GenderFactor	1	2909	6613308	8808.8
- PaymentFactor	2	21859	6632258	8809.7
<none>			6610399	8810.4
- dataset\$Unit.price	1	23715142	30325542	10331.7
- dataset\$Quantity	1	29273067	35883466	10500.0

Step: AIC=8808.76
dataset\$Total ~ GenderFactor + PaymentFactor + dataset\$Unit.price +
dataset\$Quantity

	Df	Sum of Sq	RSS	AIC
- GenderFactor	1	2723	6615527	8807.2
- PaymentFactor	2	20970	6633774	8807.9
<none>			6612804	8808.8
- dataset\$Unit.price	1	23716530	30329334	10329.9
- dataset\$Quantity	1	29271193	35883998	10498.0

Step: AIC=8807.17
dataset\$Total ~ PaymentFactor + dataset\$Unit.price + dataset\$Quantity

	Df	Sum of Sq	RSS	AIC
- PaymentFactor	2	21429	6636956	8806.4
<none>			6615527	8807.2
- dataset\$Unit.price	1	23714479	30330005	10327.9
- dataset\$Quantity	1	29475289	36090816	10501.8

Step: AIC=8806.41
dataset\$Total ~ dataset\$Unit.price + dataset\$Quantity

	Df	Sum of Sq	RSS	AIC
<none>			6636956	8806.4
- dataset\$Unit.price	1	23698836	30335792	10324.1
- dataset\$Quantity	1	29487290	36124246	10498.7

`summary(aic.table)`

סיכום-

```
Call:
lm(formula = dataset$Total ~ dataset$Unit.price + dataset$Quantity,
    data = dataset)

Residuals:
    Min       1Q   Median       3Q      Max
-211.831  -47.400   0.434   45.925  217.304

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -324.52221    7.69334  -42.18  <2e-16 ***
dataset$Unit.price    5.81364    0.09744   59.67  <2e-16 ***
dataset$Quantity    58.77155    0.88305   66.56  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 81.59 on 997 degrees of freedom
Multiple R-squared:  0.8901,    Adjusted R-squared:  0.8899
F-statistic: 4038 on 2 and 997 DF,  p-value: < 2.2e-16
```

גרסיה לפנים:

```
Start: AIC=11010.73
dataset$Total ~ 1

              Df Sum of Sq  RSS   AIC
+ dataset$Quantity  1  30063346 30335792 10324
+ dataset$Unit.price 1  24274893 36124246 10499
+ TimeFactor        2    311342 60087796 11010
+ GenderFactor      1    147700 60251438 11010
<none>              60399138 11011
+ CustomerTypeFactor 1    23370 60375769 11012
+ BranchFactor       2    106988 60292151 11013
+ PaymentFactor      2     9825 60389313 11015
+ ProductTypeFactor  5    102506 60296633 11019
```

```
Step: AIC=10324.08
dataset$Total ~ dataset$Quantity

              Df Sum of Sq  RSS   AIC
+ dataset$Unit.price 1  23698836 6636956 8806.4
<none>              30335792 10324.1
+ CustomerTypeFactor 1     3717 30332075 10326.0
+ GenderFactor       1     525 30335267 10326.1
+ BranchFactor       2    54768 30281025 10326.3
+ TimeFactor         2    51529 30284263 10326.4
+ PaymentFactor      2     5787 30330005 10327.9
+ ProductTypeFactor  5    68051 30267741 10331.8
```

```
Step: AIC=8806.41
dataset$Total ~ dataset$Quantity + dataset$Unit.price

              Df Sum of Sq  RSS   AIC
<none>              6636956 8806.4
+ PaymentFactor     2   21429.3 6615527 8807.2
+ GenderFactor      1    3182.0 6633774 8807.9
+ CustomerTypeFactor 1   1346.8 6635609 8808.2
+ BranchFactor      2   10030.3 6626926 8808.9
+ TimeFactor        2    3973.6 6632983 8809.8
+ ProductTypeFactor  5    5506.8 6631449 8815.6
```

סיכום-

```
Call:
lm(formula = dataset$Total ~ dataset$Quantity + dataset$Unit.price,
    data = dataset)

Residuals:
    Min       1Q   Median       3Q      Max
-211.831  -47.400   0.434   45.925  217.304

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -324.52221    7.69334  -42.18  <2e-16 ***
dataset$Quantity    58.77155    0.88305   66.56  <2e-16 ***
dataset$Unit.price    5.81364    0.09744   59.67  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 81.59 on 997 degrees of freedom
Multiple R-squared:  0.8901,    Adjusted R-squared:  0.8899
F-statistic: 4038 on 2 and 997 DF,  p-value: < 2.2e-16
```

רגרסיה בצעדים:

```
Start: AIC=11010.73
dataset$Total ~ 1

      Df Sum of Sq  RSS   AIC
+ dataset$Quantity 1 30063346 30335792 10324
+ dataset$Unit.price 1 24274893 36124246 10499
+ TimeFactor        2   311342 60087796 11010
+ GenderFactor      1   147700 60251438 11010
<none>                                60399138 11011
+ CustomerTypeFactor 1    23370 60375769 11012
+ BranchFactor       2   106988 60292151 11013
+ PaymentFactor      2     9825 60389313 11015
+ ProductTypeFactor  5   102506 60296633 11019
```

```
Step: AIC=10324.08
dataset$Total ~ dataset$Quantity

      Df Sum of Sq  RSS   AIC
+ dataset$Unit.price 1 23698836 6636956 8806.4
<none>                                30335792 10324.1
+ CustomerTypeFactor 1    3717 30332075 10326.0
+ GenderFactor       1     525 30335267 10326.1
+ BranchFactor       2   54768 30281025 10326.3
+ TimeFactor         2   51529 30284263 10326.4
+ PaymentFactor      2    5787 30330005 10327.9
+ ProductTypeFactor  5    68051 30267741 10331.8
- dataset$Quantity   1 30063346 60399138 11010.7
```

```
Step: AIC=8806.41
dataset$Total ~ dataset$Quantity + dataset$Unit.price

      Df Sum of Sq  RSS   AIC
<none>                                6636956 8806.4
+ PaymentFactor      2   21429 6615527 8807.2
+ GenderFactor       1    3182 6633774 8807.9
+ CustomerTypeFactor 1    1347 6635609 8808.2
+ BranchFactor       2   10030 6626926 8808.9
+ TimeFactor         2    3974 6632983 8809.8
+ ProductTypeFactor  5    5507 6631449 8815.6
- dataset$Unit.price 1 23698836 30335792 10324.1
- dataset$Quantity   1 29487290 36124246 10498.7
```

סיכום-

```
Call:
lm(formula = dataset$Total ~ dataset$Quantity + dataset$Unit.price,
    data = dataset)

Residuals:
    Min       1Q   Median       3Q      Max
-211.831  -47.400    0.434   45.925  217.304

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -324.52221    7.69334  -42.18  <2e-16 ***
dataset$Quantity   58.77155    0.88305   66.56  <2e-16 ***
dataset$Unit.price   5.81364    0.09744   59.67  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 81.59 on 997 degrees of freedom
Multiple R-squared:  0.8901,    Adjusted R-squared:  0.8899
F-statistic: 4038 on 2 and 997 DF,  p-value: < 2.2e-16
```


12. חלק ב' סעיף 3.2

```
##----- Q 3.2. Part B----- ##
mod <-lm(dataset$Total ~ dataset$Unit.price +dataset$Quantity, data = dataset)
Predicted <- predict(mod)
unstandardizedResiduals <- resid(mod)
Residuals <- (unstandardizedResiduals - mean(unstandardizedResiduals)) / sd(unstandardizedResiduals)
plot(Predicted, Residuals, main = "Residuals vs.Fitted Plot", xlab = "Predicted Value", ylab = "Normalized error")
abline(0,0)

#---QQplot and histogram: Normal assumption
mod <-lm(dataset$Total ~ dataset$Unit.price +dataset$Quantity, data = dataset)
summary(mod)
dataset$fitted<-fitted(mod)
dataset$residuals<-residuals(mod)
s.e_res <- sqrt(var(dataset$residuals))
dataset$stan_residuals<-(residuals(mod)/s.e_res)
qqnorm(dataset$stan_residuals)
abline(a=0, b=1)
hist(dataset$stan_residuals,prob=TRUE, main="Histogram of normalized error", xlab ="Normalized error",ylab = 'Frequency',col="white")
lines(density(dataset$stan_residuals),col="blue",lwd=2)
#-----KS : Normal assumption
ks.test(x= dataset$stan_residuals,y="pnorm",alternative = "two.sided", exact = NULL)
```

13. חלק ב' סעיף 4

קוד למציאת גרף Box-Cox:

```
## Box-Cox find lamda ##
model.boxcox <- boxcox(dataset$Total ~ dataset$Unit.price + dataset$Quantity, data=dataset, lambda = seq(0,0.5,0.01))
```

המודל המלא:

```
> # Full model
> mod <- lm(y ~ BranchFactor + CustomerTypeFactor + GenderFactor + ProductTypeFactor + TimeFactor + PaymentFactor + (dataset$Unit.price) + (dataset$Quantity) , data=dataset)
> summary(mod)

Call:
lm(formula = y ~ BranchFactor + CustomerTypeFactor + GenderFactor + ProductTypeFactor + TimeFactor + PaymentFactor + (dataset$Unit.price) + (dataset$Quantity), data = dataset)

Residuals:
    Min       1Q   Median       3Q      Max
-6.2067 -0.4537  0.3504  1.0917  2.0231

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.249010    0.241508   13.453 <2e-16 ***
BranchFactorB   -0.001858    0.120614   -0.015  0.988
BranchFactorC   -0.117845    0.121658   -0.969  0.333
CustomerTypeFactorMember -0.119426    0.099122  -1.205  0.229
GenderFactorMale -0.059409    0.099601   -0.596  0.551
ProductTypeFactorElectronic accessories -0.108252    0.174488   -0.620  0.535
ProductTypeFactorFashion accessories -0.145946    0.172894   -0.844  0.399
ProductTypeFactorFood and beverages    0.067931    0.173553    0.391  0.696
ProductTypeFactorHome and lifestyle    0.107809    0.177119    0.609  0.543
ProductTypeFactorSports and travel     -0.073309    0.175780   -0.417  0.677
TimeFactorEvening -0.018925    0.123234   -0.154  0.878
TimeFactorNoon    0.043723    0.120335    0.363  0.716
PaymentFactorCredit card 0.128390    0.122868    1.045  0.296
PaymentFactorWallet 0.064365    0.119371    0.539  0.590
dataset$Unit.price 0.159265    0.001868   85.270 <2e-16 ***
dataset$Quantity   1.672454    0.017010  98.322 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.558 on 984 degrees of freedom
Multiple R-squared:  0.9463,    Adjusted R-squared:  0.9455
F-statistic: 1157 on 15 and 984 DF,  p-value: < 2.2e-16
```

גרסיה לאחר:

```
> step.backw <- step(mod,direction= "backward",trace = TRUE)
Start: AIC=902.72
y ~ BranchFactor + CustomerTypeFactor + GenderFactor + ProductTypeFactor + TimeFactor + PaymentFactor + (dataset$Unit.price) + (dataset$Quantity)

              Df Sum of Sq    RSS    AIC
- ProductTypeFactor    5      8.5 2397.1  896.3
- TimeFactor            2      0.7 2389.3  899.0
- PaymentFactor         2      2.7 2391.3  899.8
- BranchFactor          2      3.0 2391.6  900.0
- GenderFactor          1      0.9 2389.5  901.1
- CustomerTypeFactor    1      3.5 2392.1  902.2
<none>                  2388.6  902.7
- dataset$Unit.price    1 17650.0 20038.6 3027.7
- dataset$Quantity      1 23466.9 25855.5 3282.5

Step: AIC=896.28
y ~ BranchFactor + CustomerTypeFactor + GenderFactor + TimeFactor + PaymentFactor + dataset$Unit.price + dataset$Quantity

              Df Sum of Sq    RSS    AIC
- TimeFactor            2      0.7 2397.8  892.6
- PaymentFactor         2      2.9 2400.1  893.5
- BranchFactor          2      3.1 2400.2  893.6
- GenderFactor          1      0.7 2397.9  894.6
- CustomerTypeFactor    1      3.1 2400.3  895.6
<none>                  2397.1  896.3
- dataset$Unit.price    1 17691.2 20088.3 3020.1
- dataset$Quantity      1 23650.5 26047.7 3279.9
```

```

Step: AIC=892.57
y ~ BranchFactor + CustomerTypeFactor + GenderFactor + PaymentFactor +
  dataset$Unit.price + dataset$Quantity

      Df Sum of Sq    RSS    AIC
- PaymentFactor      2      2.8  2400.6  889.7
- BranchFactor       2      3.0  2400.8  889.8
- GenderFactor       1      0.8  2398.6  890.9
- CustomerTypeFactor 1      3.2  2401.0  891.9
<none>                 2397.8  892.6
- dataset$Unit.price 1  17719.8 20117.6 3017.6
- dataset$Quantity   1  23772.5 26170.3 3280.6

Step: AIC=889.72
y ~ BranchFactor + CustomerTypeFactor + GenderFactor + dataset$Unit.price +
  dataset$Quantity

      Df Sum of Sq    RSS    AIC
- BranchFactor      2      3.2  2403.8  887.1
- GenderFactor       1      0.8  2401.4  888.0
- CustomerTypeFactor 1      2.9  2403.5  888.9
<none>                 2400.6  889.7
- dataset$Unit.price 1  17727.4 20128.0 3014.1
- dataset$Quantity   1  23774.4 26175.0 3276.8

Step: AIC=887.07
y ~ CustomerTypeFactor + GenderFactor + dataset$Unit.price +
  dataset$Quantity

      Df Sum of Sq    RSS    AIC
- GenderFactor       1      0.6  2404.5  885.3
- CustomerTypeFactor 1      3.0  2406.8  886.3
<none>                 2403.8  887.1
- dataset$Unit.price 1  17729.3 20133.2 3010.4
- dataset$Quantity   1  23771.2 26175.0 3272.8

```

```

Step: AIC=885.32
y ~ CustomerTypeFactor + dataset$Unit.price + dataset$Quantity

      Df Sum of Sq    RSS    AIC
- CustomerTypeFactor 1      2.9  2407.4  884.5
<none>                 2404.5  885.3
- dataset$Unit.price 1  17730.9 20135.4 3008.5
- dataset$Quantity   1  23919.5 26324.0 3276.5

Step: AIC=884.53
y ~ dataset$Unit.price + dataset$Quantity

      Df Sum of Sq    RSS    AIC
<none>                 2407.4  884.5
- dataset$Unit.price 1  17729 20136.3 3006.5
- dataset$Quantity   1  23917 26324.7 3274.5

```

סיכום רגרסיה לאחר:

```

Call:
lm(formula = y ~ dataset$Unit.price + dataset$Quantity, data = dataset)

Residuals:
    Min       1Q   Median       3Q      Max
-6.0071 -0.4614  0.3582  1.1475  1.7798

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.171001   0.146521   21.64  <2e-16 ***
dataset$Unit.price 0.159011   0.001856   85.69  <2e-16 ***
dataset$Quantity  1.673811   0.016818   99.53  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.554 on 997 degrees of freedom
Multiple R-squared:  0.9459,    Adjusted R-squared:  0.9458
F-statistic: 8717 on 2 and 997 DF, p-value: < 2.2e-16

```

המודל הסופי אחרי רגרסיה לאחור:

```
Call:
lm(formula = y ~ (dataset$Unit.price) + (dataset$Quantity), data = dataset)

Residuals:
    Min       1Q   Median       3Q      Max
-6.0071 -0.4614  0.3582  1.1475  1.7798

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.171001   0.146521   21.64  <2e-16 ***
dataset$Unit.price 0.159011   0.001856   85.69  <2e-16 ***
dataset$Quantity  1.673811   0.016818   99.53  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.554 on 997 degrees of freedom
Multiple R-squared:  0.9459,    Adjusted R-squared:  0.9458
F-statistic: 8717 on 2 and 997 DF,  p-value: < 2.2e-16
```

בדיקת הנחת שוויון השונויות:

```
## Equality of variance
mod <- lm(y ~ (dataset$Unit.price) + (dataset$Quantity), data = dataset)
Predicted <- predict(mod)
unstandardizedResiduals <- resid(mod)
Residuals <- (unstandardizedResiduals - mean(unstandardizedResiduals)) / sd(unstandardizedResiduals)
plot(Predicted, Residuals, main = "Residuals vs.Fitted Plot", xlab = "Predicted Value", ylab = "Normalized error")
abline(0,0)
```

בדיקת הנחת הלינאריות:

```
> # QQ plot: linear assumption
> mod <- lm(y ~ dataset$Unit.price + dataset$Quantity, data = dataset)
> summary(mod)

Call:
lm(formula = y ~ dataset$Unit.price + dataset$Quantity, data = dataset)

Residuals:
    Min       1Q   Median       3Q      Max
-6.0071 -0.4614  0.3582  1.1475  1.7798

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.171001   0.146521   21.64  <2e-16 ***
dataset$Unit.price 0.159011   0.001856   85.69  <2e-16 ***
dataset$Quantity  1.673811   0.016818   99.53  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.554 on 997 degrees of freedom
Multiple R-squared:  0.9459,    Adjusted R-squared:  0.9458
F-statistic: 8717 on 2 and 997 DF,  p-value: < 2.2e-16

> dataset$fitted<-fitted(mod)
> dataset$residuals<-residuals(mod)
> s.e_res <- sqrt(var(dataset$residuals))
> dataset$stan_residuals<-(residuals(mod)/s.e_res)
> qqnorm(dataset$stan_residuals)
> abline(a=0, b=1)
```

בדיקת הנחת הנורמליות:

```
> ## histogram: Normal assumption
> hist(dataset$stan_residuals,prob=TRUE, main="Histogram of normalized error", xlab = "Normalized error", ylab = 'Frequency',col="white")
> lines(density(dataset$stan_residuals),col="blue",lwd=2)
> ## KS : Normal assumption
> ks.test(x= dataset$stan_residuals,y="pnorm",alternative = "two.sided")
```