פייתון מתקדם לכלכלנים

תרגיל מסכם

שם המגיש: דור גאון

שם המרצה: ד״ר אביחי שניר

חלק I

משימה ראשונה

להלן הטבלה המעודכנת עם השדות הבינאריים:

```
         condition       id  baseline_car1  update_car1  Sign_Top  drive_diff
0              NaN      NaN            NaN          NaN         0         NaN
1         Sign Top      1.0          896.0      39198.0         1     38302.0
2              NaN      NaN            NaN          NaN         0         NaN
3      Sign Bottom      2.0        21396.0      63511.0         0     42115.0
4              NaN      NaN            NaN          NaN         0         NaN
...            ...      ...            ...          ...       ...         ...
26971  Sign Bottom  13486.0        30700.0      32916.0         0      2216.0
26972          NaN      NaN            NaN          NaN         0         NaN
26973     Sign Top  13487.0        29884.0      35459.0         1      5575.0
26974          NaN      NaN            NaN          NaN         0         NaN
26975     Sign Top  13488.0        22930.0      37888.0         1     14958.0

[26976 rows x 6 columns]
```

משימה שנייה

הנתונים מקבוצת הניסוי כפי שהודפסו בקוד:

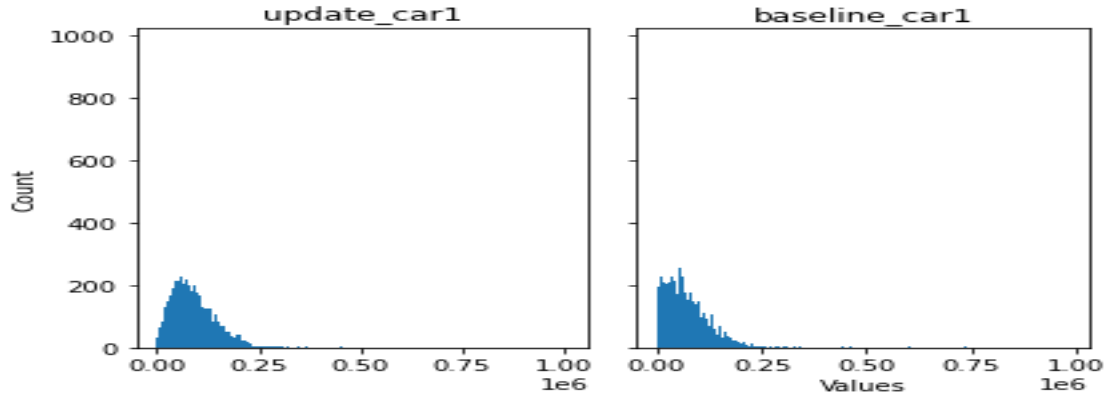
```
'%' in experiment Group ("Sign Top"):   25.3%

Mean of drive distance difference:      24929.0

std of drive distance difference:       14422.46
```

משימה שלישית

בהינתן הגדרת הנתונים המצויים בעמודות, נוכל להסיק כי יתכן ותהיה סטייה קלה מהתפלגות נורמלית סטנדרטית, נסביר:

- מרכז ההתפלגות - ניזכר כי מדובר במרחקים שעברו מכוניות טרם תחילת הניסוי ולכן אנו יכולים לצפות כי מכוניות רבות ידווחו מרחק התחלתי אפס, זאת בשונה מעמודת הנתונים update_car1 המציגה נתונים לאחר תחילת הניסוי, שם אנו צופים לקבל ערכים גדולים יותר. הבדל זה יזיז את התפלגות ימינה של baseline_car1 ביחס ל update_car1
- נקודות קצה (outliers) – איננו יודעים מה אינטרס הדיווח של כל אדם המשתתף בניסוי. במידה ויש לאדם מסוים רווח מדיווח גבוה/נמוך של מרחק, יתכן ונבחין בנקודות קצה, שאינן עולות בקנה אחד עם התפלגות נורמלית.

היסטוגרמות:



נאמת מספר עובדות מתוך היסטוגרמות:

- אכן ההיסטוגרמות משקפות התפלגות נורמלית במראם הכללי.
- נבחין כי baseline_car1 כתומה מצידה השמאלי באופן גס
- כפי שצפינו, ישנה הצטברות של ערכים בצידה השמאלי של ההתפלגות של baseline_car1 – רכבים שדיווחו שלא עברו שום מרחק כלל.
- ההתפלגויות שניהם הם בעלות זנב ימני – ישנה הצטברות של ערכים בין 0 ל – 200,000 מייל אך ישנם דגימות לאורך הציר עד מרחקים של מעל 1,000,000 מייל!

משימה רביעית

נציג את הממוצעים של ההפרשים בשתי הקבוצות:

```
Total average of drive_diff:    24929.0
sign_Bottom average is:     23622.55
sign_Top average is:        26204.83
```

נבחין כי יש פער בין הממוצעים, נחשב אותו:

$$\left(\frac{26204.83}{23622.55} - 1\right) \cdot 100 = 10.93\%$$

ממוצע המרחקים שדווח ע״י קבוצת הניסוי (שחתמו בתחילת הטופס) הוא גדול ב 10.93% מזה שדווח ע״י קבוצת הביקורת. הממצאים עולים בקנה אחד עם השערות החוקרים.

משימה חמישית

נבדוק את מובהקות ההבדלים, בעזרת הרגרסיה שהרצנו :

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             drive_diff   R-squared:                       0.008
Model:                            OLS   Adj. R-squared:                  0.008
Method:                 Least Squares   F-statistic:                     108.9
Date:                Sun, 04 Sep 2022   Prob (F-statistic):           2.08e-25
Time:                        18:48:29   Log-Likelihood:            -1.4825e+05
No. Observations:               13488   AIC:                         2.965e+05
Df Residuals:                   13486   BIC:                         2.965e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   2.362e+04    175.971    134.241      0.000    2.33e+04     2.4e+04
Sign_Top    2582.2822    247.397     10.438      0.000    2097.349    3067.216
==============================================================================
Omnibus:                     9955.573   Durbin-Watson:                   1.983
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              783.625
Skew:                           0.009   Prob(JB):                    6.89e-171
Kurtosis:                       1.819   Cond. No.                         2.63
==============================================================================
```

תוצאות הרגרסיה :

- מקדם ההסברה של הרגרסיה שלנו נמוך מאוד – 0.008
- ביצענו רגרסיה כאשר המשתנה הבלתי תלוי שלנו הוא משתנה המעיד על סוג הקבוצה – ביקורת/ניסוי - משתנה בינארי. בעת ביצוע רגרסיה שכזו, המשתנה הדמי מודד את ההפרש הצפוי בין שתי הקבוצות.
  ערך המקדם – 2582.2822
  נאמת כי זה ההפרש שקיבלנו בסעיף הקודם :
  $$26{,}204 - 23{,}622 = 2{,}582$$
- נבחין כי ה Pvalue של מקדם המשתנה הדמי קטן מאוד ובפרט קטן מהסטטיסטי (t = 10.438) ולכן נסיק כי התוצאות מובהקות.

התוצאות תומכות בהשערה של החוקרים שהניחו כי מי שיחתום על הדף בראשו ידווח תוצאות מהמנות לעומת אלו שיחתמו בתחתית הדף. קיים הפרש בין קבוצות אלו וע״פ הרגרסיה הוא אכן מובהק.

משימה שישית

להלן חישוב מרחקי הנסיעה הממוצעים לפי baseline_car1:

```
Total average of baseline_car1 is:  67356.43

sign_Bottom average is:     74945.71

sign_Top average is:        59945.09
```

לא היינו מצפים לראות שינוי בין הקבוצות בנתונים שנדגמו טרם הניסוי. טרם הניסוי, אין שוני בין קבוצת הביקורת לקבוצת הניסוי ולכן, במידה ודגימת הנתונים נעשתה באופן אקראי ואיכותי, לא נצפה לראות שוני בין הממוצעים. **בפועל**, יש שוני מסוים בין הממוצעים. הצטרך להשתמש בכלים סטטיסטיים על מנת לקבוע האם השוני מובהק או לא.

משימה שביעית

להלן חישוב מרחקי הנסיעה הממוצעים לפי baseline_car1:

```
Total average of update_car1 is:    92285.43

sign_Bottom average is:     98568.26

sign_Top average is:        86149.92
```

במקרה זה, אין התשובה לשאלה "שחור" או "לבן", נבין תחילה את משמעות העמודה update_car1 – ניזכר כי העמודה מציינת את סה"כ המרחק שעבר כל רכב מראשית דרכו על הכביש, כאשר נמדד לאחר סיום הניסוי. המדד אותו אנו מעוניינים למדוד הוא גודל המרחק שדווח מרגע תחילת הניסוי! נסיק מכאן שהמדד update_car מכיל "רעש" שהוא המרחק שעבר כל רכב טרם תחילת הניסוי. בחינה איכותית יותר של הממצאים תהיה בדיקת ההפרש כפי שעשינו בסעיף 4.

נסכם, כי בגלל ה"רעש" שקיים במשתנה הנבדק, נתקשה להסיק מסקנות מהמנות ללא שימוש בכלים סטטיסטיים לבדיקת מובהקות השוני בין הקבוצות, אך כן נוכל לשער כי על סמך ממצאים אלו, לא תהיה מובהקות עבור שוני הממוצעים במקרה זה.

משימה שמינית

נשים לב למובהקות המקדם הב״ת: ערך הסטטיסטי שלילי ולכן אין מובהקות של מקדם המשתנה הדמי.

נסיק מנתונים אלו כי השוני בין ערכי ה – baseline אינו מובהק! כלומר לא ניתן להבחין בשוני משמעותי בין הערכים. המסקנה עולה בקנה אחד עם המסקנות מסעיף 6. – המשתנה מתאר מדד שנקבע טרם בידול בין קבוצת הניסוי והביקורת ולכן היינו מצפים לראות חוסר מובהקות כדי לקבוע שהניסוי אכן בוצע באופן איכותי.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          baseline_car1   R-squared:                       0.017
Model:                            OLS   Adj. R-squared:                  0.017
Method:                 Least Squares   F-statistic:                     239.6
Date:                Sun, 04 Sep 2022   Prob (F-statistic):           1.36e-53
Time:                        18:48:29   Log-Likelihood:            -1.6667e+05
No. Observations:               13488   AIC:                         3.333e+05
Df Residuals:                   13486   BIC:                         3.334e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   7.495e+04    689.249    108.735      0.000    7.36e+04    7.63e+04
Sign_Top     -1.5e+04    969.015    -15.480      0.000   -1.69e+04   -1.31e+04
==============================================================================
Omnibus:                     7476.950   Durbin-Watson:                   1.997
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           172171.456
Skew:                           2.193   Prob(JB):                         0.00
Kurtosis:                      19.944   Cond. No.                         2.63
==============================================================================
```

משימה תשיעית

נשים לב למובהקות המקדם הב״ת: ערך הסטטיסטי שלילי ולכן אין מובהקות של מקדם המשתנה הדמי.

נסיק מנתונים אלו כי השוני בין ערכי ה – updated אינו מובהק, כלומר לא ניתן להבחין בשוני משמעותי בין הערכים. המסקנה עולה בקנה אחד עם המסקנות מסעיף 7.

כפי, שהסברנו בסעיף 7, ממצאים אלו עולים בקנה אחד עם השערותינו. נוסיף כי קבלת החלטה על הצלחת הניסוי ע״פ משתנה זה היא לא מוצלחת ובמידה והניסוי הסתמך על תוצאות אלו, יתכן והגיע למסקנות שגויות.

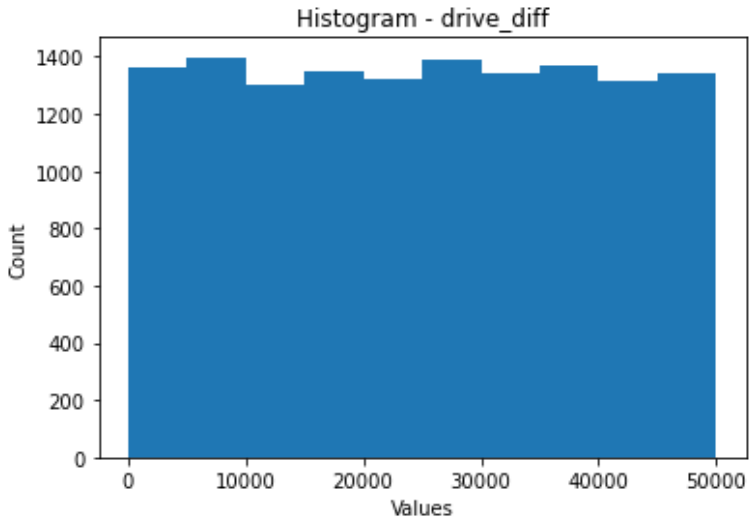
```
                            OLS Regression Results
==============================================================================
Dep. Variable:            update_car1   R-squared:                       0.011
Model:                            OLS   Adj. R-squared:                  0.011
Method:                 Least Squares   F-statistic:                     153.2
Date:                Sun, 04 Sep 2022   Prob (F-statistic):           5.26e-35
Time:                        18:48:29   Log-Likelihood:            -1.6713e+05
No. Observations:               13488   AIC:                         3.343e+05
Df Residuals:                   13486   BIC:                         3.343e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   9.857e+04    713.553    138.137      0.000    9.72e+04       1e+05
Sign_Top   -1.242e+04   1003.184    -12.379      0.000   -1.44e+04   -1.05e+04
==============================================================================
Omnibus:                     6772.629   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           127920.667
Skew:                           1.971   Prob(JB):                         0.00
Kurtosis:                      17.563   Cond. No.                         2.63
==============================================================================
```

משימה עשר

בציור ההיסטוגרמה של drive_diff, נצפה לכאורה להתפלגות נורמלית שכן פעולות כמו חיבור וחיסור בין התפלגויות נורמליות יניבו התפלגות נורמלית, אך ניזכר מה קורה לתוצאה של חיסור התפלגויות נורמליות:

ממוצע ההתפלגויות מחוסר האחד בשני, מנגד, סטיות התקן סכמות. לכן, נוכל לצפות שההתפלגות של ההפרש לא בהכרח תיראה נורמלית.

בפועל, קיבלנו התפלגות שאינה נראה נורמלית כלל, הדגימות מפוזרות לכל אורך טווח הערכים שלנו, אינם ערכים קיצוניים ברורים או ריכוז של ערכים סביב נקודה מסוימת. ננסה לשנות את ערך ה bins כדי ללמוד על הנתונים בצורה טובה יותר:



בהקטנת ערך ה bins, הקטנו את הרזולוציה של ההיסטוגרמה וקיבלו התפלגות שנראית כמעט אחידה, כלומר, כל משתתף שנבחר, יש הסתברות כמעט זהה שהוא נמצא בין אחד מהטווחים המצוינים על ציר X.

ההיסטוגרמה מייצגת את תכונות משתנה ההפרשים וכפי שהיינו מצפים, מציגה התפלגות דיי אקראית של ערכים המתכנסת לפיזור אחיד שלהם לאורך הספקטרום. אין מן הממצאים פרט שגורם לחשד בנתונים.

משימה אחת-עשרה

על סמך כלל הניתוחים והתוצאות שקיבלנו נסיק כי:

1. על סמך הנתונים בקובץ Shu et al, ניתן להסיק כי ניתן להגיע לממצאים מובהקים אודות השוני בין קבוצת הביקורת לקבוצת הניסוי.
2. מובהקות הממצאים, בפרט הפרש הממוצעים בין קבוצת הביקורת לקבוצת הניסוי מאפשר לנו להסיק מסקנות העולות בקנה אחד עם השערות הניסוי – חתימה בראש הדף היא אמצעי יעיל לעודד אנשים לדווח אמת, זאת בגלל שגילינו שקבוצת הניסוי דיווחה כי עברה מרחקים ארוכים יותר ביחס לקבוצת הביקורת (נזכיר כי לכל אחד יש אינטרס לדווח מרחק נמוך)

חלק II

משימה שנייה

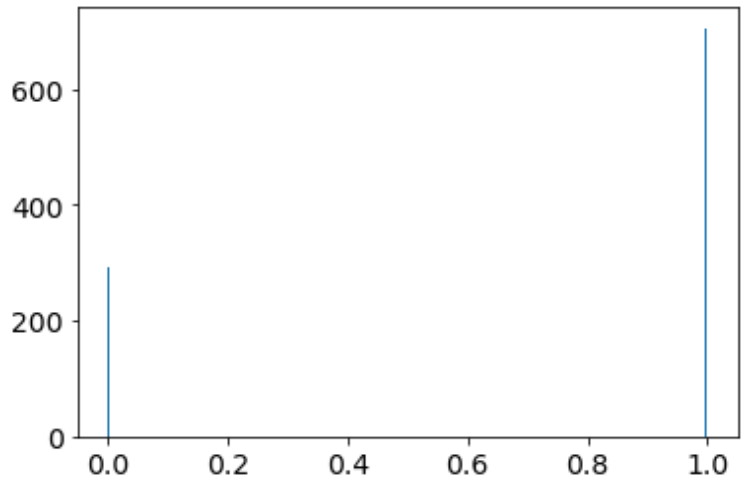
יצרנו משתני דמי עבור המשתנים בעמודות:

- Internet
- owner
- ownernfl
- devtown

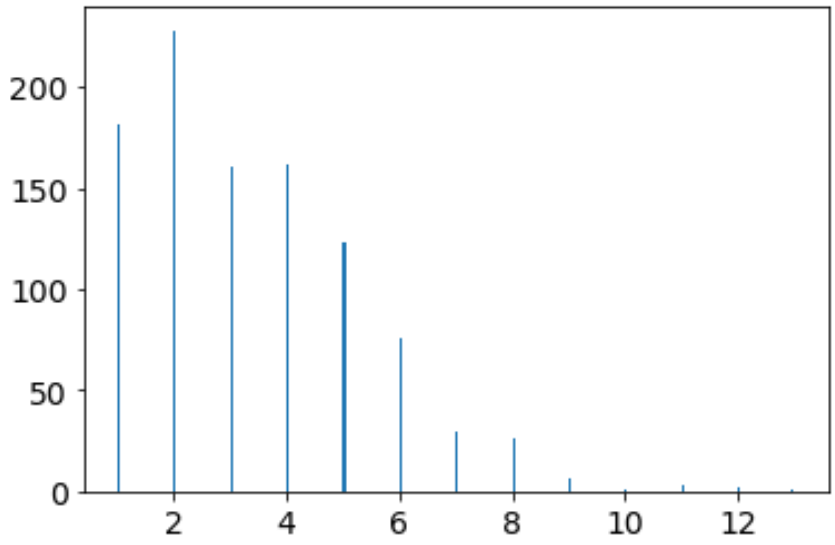
משימה שלישית

להלן ההיסטוגרמות:

Owner – האם ברשות משק הבית יש דירה

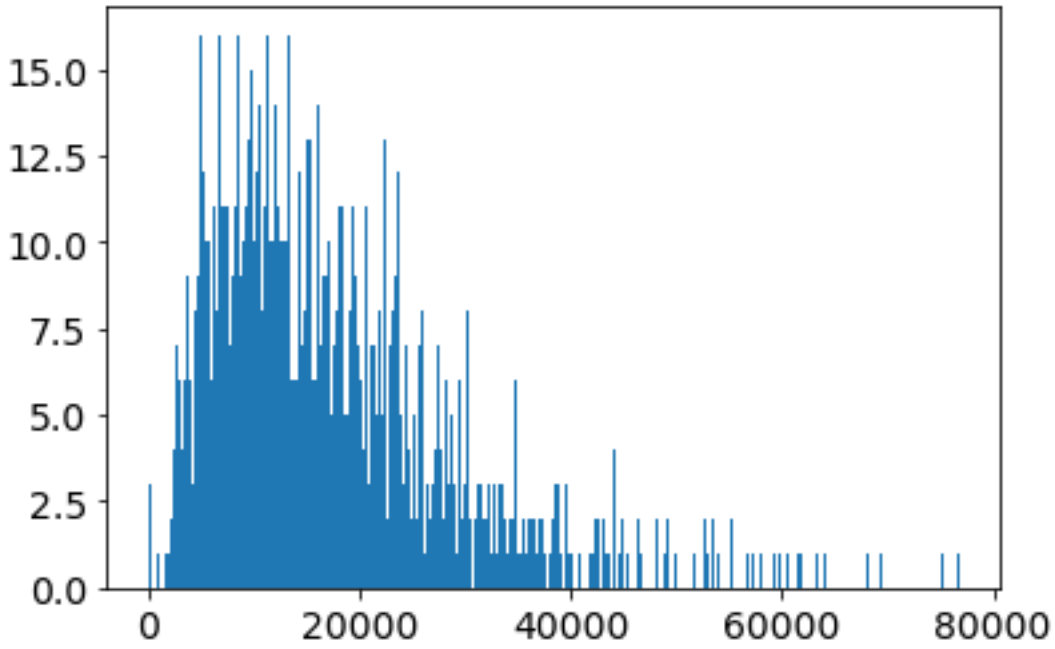


Hrprsns – מספר הנפשות במשק הבית



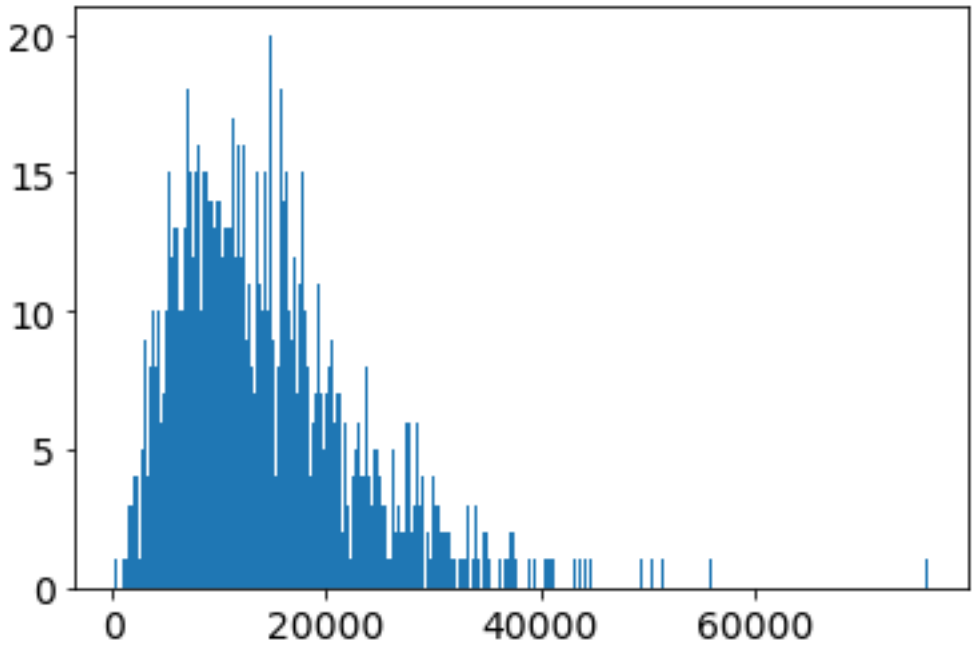
גם כאן נבחין בערכים חריגים בקצה הימני של ציר ה X, הם מעטים בכמותם ביחס לשאר.

Incoment – ההכנסה נטו של משק הבית



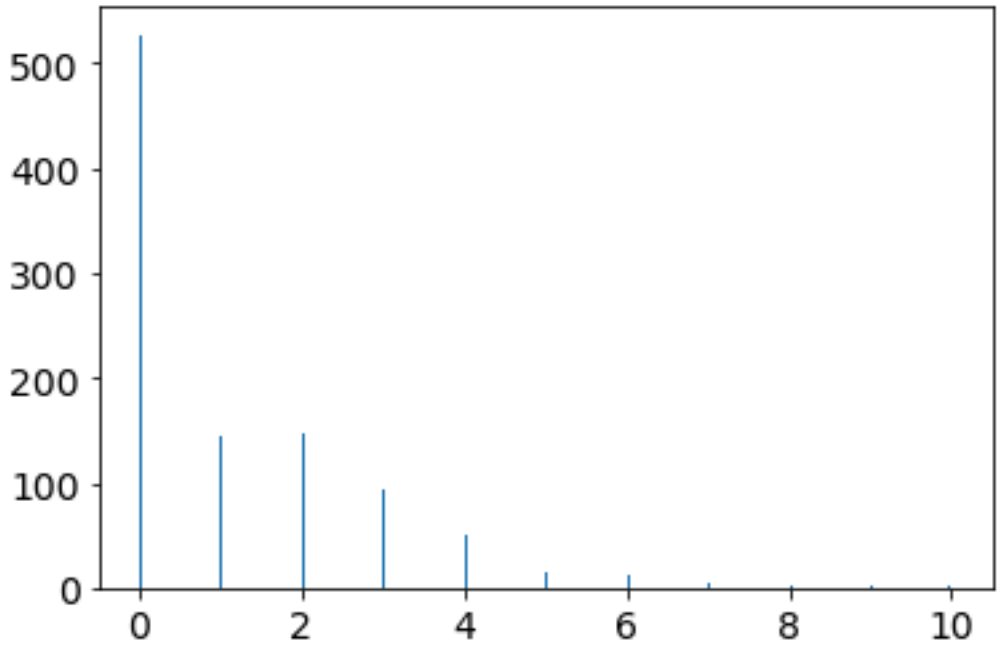
קיימים ערכים חריגים בטווח ה 80 אלף

C3 – סה"כ ההוצאות של משק הבית



C3 המשתנה התלוי שלנו. ולכן נבחין בערכים ממונפים בצד הקיצוני של ציר ה X. קיימים מספר ערכים קיצוניים שכאלו שיהיו ערכים ממונפים באמידת הרגרסיה

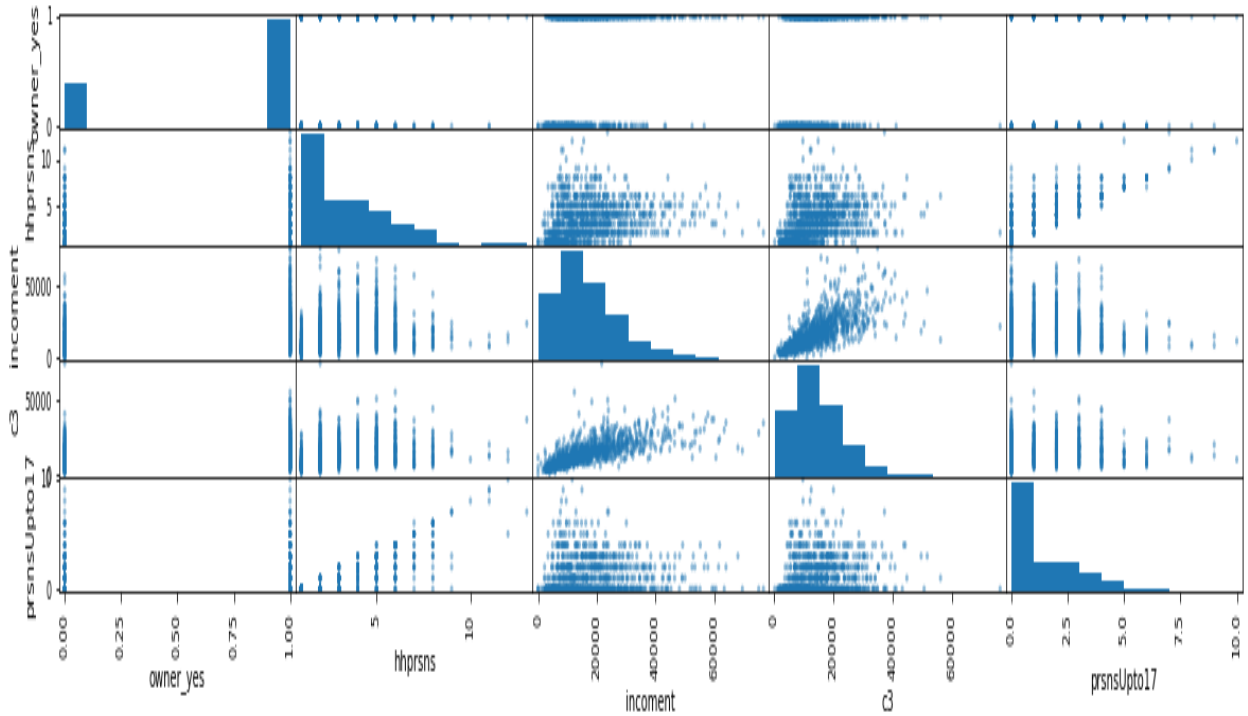
prsnsUpto17 – כמות הנפשות על גיל 17 (כולל)



הירידה אחידה לאורך הציר. למרות זאת, יתכן וניתקל בערכים חריגים כאשר נגדיל את הרזולוציה של ההיסטוגרמה. ניתן לראות מספר כאלו עבור הערך 9.

משימה רביעית

מטריצת הפיזור של המשתנים :



מטריצת מקדמי פירסון:

```
The correlation matrix (Pearson):
             owner_yes   hhprsns  incoment        c3  prsnsUpto17
owner_yes     1.000000  0.146167  0.268814  0.221336     0.057301
hhprsns       0.146167  1.000000  0.214317  0.306931     0.838022
incoment      0.268814  0.214317  1.000000  0.686303     0.071694
c3            0.221336  0.306931  0.686303  1.000000     0.173856
prsnsUpto17   0.057301  0.838022  0.071694  0.173856     1.000000
```

א. נשים לב כי יש מתאם גבוה בין סה"כ ההוצאות בבית לבין ההכנסה נטו בבית. מקדם המתאם בין משתנים אלו הוא הגדול ביותר – 0.686303

ב. בין שאר המשתנים במטריצה, נבחין כי קיים מתאם גדול גם בין כמות הנפשות עד גיל 17 (כולל) לבין סה"כ הנפשות בבית, ערך המתאם עומד על – 0.838022

ג. כן, כאשר יש מתאם בין המשתנה התלוי לב"ת, נוכל לאמת כי קיים קשר כלשהוא ביניהם, בין אם גבוה או נמוך ועל כן הם ניתנים להסברה האחד באמצעות השני. מודל הרגרסיה לצורך העניין יתקשה להסביר את המשתנה התלוי באמצעות משתנים ב"ת שהמתאם שלהם עם התלוי הוא נמוך מאוד.

ד. לא, נרצה מתאם נמוך בין המשתנים הב"ת תלויים שלנו. מתאם גבוה בין משתנים אלו מראה כי הם "מכילים מידע דומה" ועל כן הסברה של משתנה תלוי באמצעות שניהם היא לא אפקטיבית ויכולה לגרום למוטיקולינאריות.

משימה חמישית

להלן תוצאות הרגרסיה:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     c3   R-squared:                       0.499
Model:                            OLS   Adj. R-squared:                  0.497
Method:                 Least Squares   F-statistic:                     247.6
Date:                Sun, 04 Sep 2022   Prob (F-statistic):          1.35e-147
Time:                        18:59:05   Log-Likelihood:                -10133.
No. Observations:                1000   AIC:                         2.028e+04
Df Residuals:                     995   BIC:                         2.030e+04
Df Model:                           4
Covariance Type:            nonrobust
==================================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept   3439.7029    521.747      6.593      0.000    2415.852    4463.554
owner_yes    417.8448    442.972      0.943      0.346    -451.421    1287.111
hhprsns      873.0562    185.920      4.696      0.000     508.215    1237.897
incoment       0.4623      0.017     26.561      0.000       0.428       0.496
prsnsUpto17 -234.1238    228.412     -1.025      0.306    -682.349     214.101
==============================================================================
Omnibus:                      552.382   Durbin-Watson:                   2.058
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             9904.076
Skew:                           2.130   Prob(JB):                         0.00
Kurtosis:                      17.818   Cond. No.                     6.40e+04
==============================================================================
```

משימה שישית

המקדמים המובהקים ברגרסיה הם המקדמים בעלי Pvalue קטן מ 0.05. המשתנים המקיימים זאת:

- hhprsns
- incoment

משימה שביעית

הנש״צ המתקבל לפי הרגרסיה בסעיף 5 הוא 0.38 וזאת ע״פ מקדם המשתנה Incoment. מכיוון שמשתנה זה מעיד על ההוצאה של הפרט. מקדם המשתנה זה ברגרסיה משמעותו השינוי בהוצאה כפונקציה של ההכנסה (בשקלים) – ההגדרה של נש״צ.

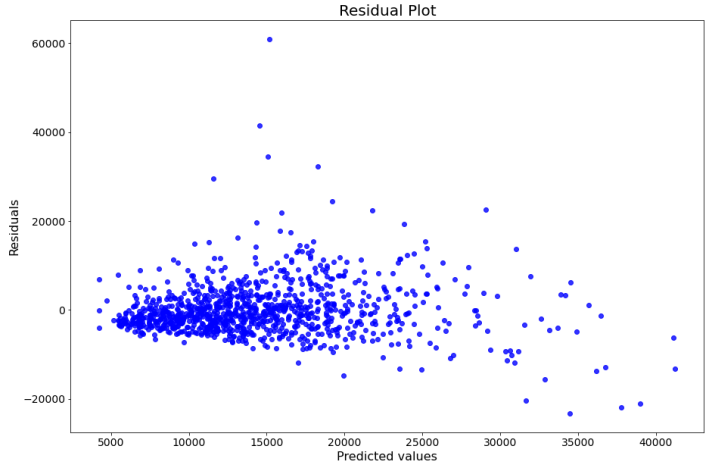
מכאן נסיק כי אף אחת מההשערות אינה נכונה, בפרט השערה ג׳, שכן מקדם מספר הילדים עד גיל 17 כולל הוא 74.3184 ולא 250.

משימה שמינית

נבחין בגרף שציירנו של הטעות כפונקציה של הערכים המנובים. נבחין בצורה המזכירה גרף לינארי בכיוון התפשטות הנקודות. זאת על אף שהנקודות אינם מסודרות באופן סימטרי סביב קו ישר.

קיימים ערכים קיצוניים שפוגעים בנראות הגרף ועלולים לשפר את הביצועים אם ניפטר מהם.

נסיק כי התוצאות שקיבלנו תומכות בהשערה שרגרסיה היא כלי מתאים במקרה זה.



משימה תשיעית

להלן חישוב ה VIF:

נבחין כי ערך ה VIF של כמות הנפשות במשפחה גדול מ 5 ועל כן נותן לסיבה לחשוד במולטיקולינאריות.

	independent_values	VIF
0	owner_yes	3.220166
1	hhprsns	9.721572
2	incoment	3.455773
3	prsnsUpto17	4.653336

משימה עשירית

נציג את התהליך :

- שלפנו נתונים אודות הנקודות הקיצוניות באמצעות הפונקציה get_influence().
- בחרנו את החתך להיות 3 כפי שהגדרנו בכיתה כחתך שנהוג לבחור בדרך כלל.
- הצגנו את הנקודות בעלי הערכים החורגים ואז דגמנו את התצפיות המיוצגות על ידם :

```
Name: 786, dtype: int64
owner_yes          1
hhprsns            4
incoment       23660
prsnsUpto17        0
Name: 799, dtype: int64
owner_yes          0
hhprsns            2
incoment       13840
prsnsUpto17        0
Name: 866, dtype: int64
owner_yes          1
hhprsns            3
incoment       60588
prsnsUpto17        0
Name: 949, dtype: int64
owner_yes          1
hhprsns            4
incoment       15663
prsnsUpto17        1
Name: 953, dtype: int64
owner_yes          1
hhprsns            6
incoment       12272
prsnsUpto17        1
Name: 983, dtype: int64
```

```
owner_yes          1
hhprsns            5
incoment       46636
prsnsUpto17        3
Name: 577, dtype: int64
owner_yes          0
hhprsns            2
incoment       21611
prsnsUpto17        0
Name: 607, dtype: int64
owner_yes          1
hhprsns            4
incoment       18230
prsnsUpto17        3
Name: 629, dtype: int64
owner_yes          1
hhprsns            5
incoment       68084
prsnsUpto17        3
Name: 661, dtype: int64
owner_yes          1
hhprsns            6
incoment       29458
prsnsUpto17        4
```

```
owner_yes          1
hhprsns            2
incoment       22335
prsnsUpto17        0
Name: 53, dtype: int64
owner_yes          1
hhprsns            6
incoment       23885
prsnsUpto17        4
Name: 79, dtype: int64
owner_yes          1
hhprsns            4
incoment       52534
prsnsUpto17        0
Name: 288, dtype: int64
owner_yes          1
hhprsns            2
incoment       39376
prsnsUpto17        0
Name: 299, dtype: int64
owner_yes          1
hhprsns            2
incoment       69615
prsnsUpto17        0
Name: 563, dtype: int64
```

- בסה"כ מצאנו 15 ערכים כאלה.

משימה אחת עשרה

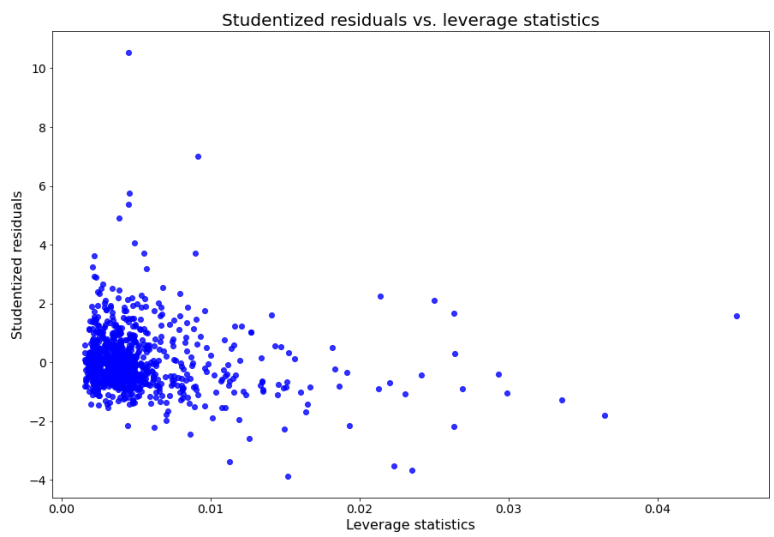
באופן דומה פעלנו גם עבור הערכים הממונפים. את החתך בחרנו להיות: $2 \cdot \frac{p+1}{n}$

הערכים שקיבלנו כממונפים:

```
The cutoff is: 0.01
High leverage points:
 0      0.011263
33     0.011932
42     0.016526
77     0.010963
99     0.014502
         ...
945    0.016682
949    0.015184
956    0.011495
984    0.011111
994    0.019165
Name: hat_diag, Length: 74, dtype: float64
```

- בסה"כ 74 ערכים.

משימה שתיים עשרה



ע"י חיתוך בין התצפיות הממונפות לערכים הממונפים קיבלו 4 תצפיות שנתגלו כחריגות בשתי הבדיקות. ניתן גם לראות אותן בגרף.

התצפיות שהתקבלו כחריגות הן:

[288, 563, 661, 949]

משימה שלוש עשרה

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     c3   R-squared:                       0.526
Model:                            OLS   Adj. R-squared:                  0.524
Method:                 Least Squares   F-statistic:                     274.8
Date:                Sun, 04 Sep 2022   Prob (F-statistic):          6.47e-159
Time:                        18:59:07   Log-Likelihood:                -10066.
No. Observations:                 996   AIC:                         2.014e+04
Df Residuals:                     991   BIC:                         2.017e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    3168.4070    509.723      6.216      0.000    2168.147    4168.667
owner_yes     353.9751    431.702      0.820      0.412    -493.180    1201.130
hhprsns       819.2471    181.636      4.510      0.000     462.812    1175.682
incoment        0.4943      0.017     28.247      0.000       0.460       0.529
prsnsUpto17  -208.8512    223.133     -0.936      0.350    -646.718     229.015
==============================================================================
Omnibus:                      600.415   Durbin-Watson:                   2.085
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            11394.823
Skew:                           2.388   Prob(JB):                         0.00
Kurtosis:                      18.867   Cond. No.                     6.29e+04
==============================================================================
```

להלן תוצאות הרגרסיה ללא התצפיות הקיצוניות.

א. נבחן את ההשפעה על המקדמים:

- Owner - קטן
- Hhprsns – קטן
- Incoment – גדל
- prsnsUpto17 – הפך לשלילי

ב. R^2– גדל

משימה ארבע עשרה

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     c3   R-squared:                       0.517
Model:                            OLS   Adj. R-squared:                  0.515
Method:                 Least Squares   F-statistic:                     246.8
Date:                Sun, 04 Sep 2022   Prob (F-statistic):          5.10e-144
Time:                        18:59:07   Log-Likelihood:                -9338.2
No. Observations:                 926   AIC:                         1.869e+04
Df Residuals:                     921   BIC:                         1.871e+04
Df Model:                           4
Covariance Type:            nonrobust
==================================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept    2590.6485    546.458      4.741      0.000    1518.202    3663.095
owner_yes     132.2658    439.019      0.301      0.763    -729.329     993.860
hhprsns       750.7766    225.040      3.336      0.001     309.125    1192.428
incoment        0.5534      0.021     25.940      0.000       0.512       0.595
prsnsUpto17   -17.5706    278.113     -0.063      0.950    -563.378     528.237
==============================================================================
Omnibus:                      605.542   Durbin-Watson:                   2.047
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            13576.386
Skew:                           2.608   Prob(JB):                         0.00
Kurtosis:                      21.018   Cond. No.                     6.10e+04
==============================================================================
```

הנקודות שהסרנו הם הנקודות שצוינו בסעיף אחת עשרה. להלן פלט הרגרסיה:

א. נבחן את ההשפעה על המקדמים:

- Owner- קטן
- Hhprsns – קטן
- Incoment – קטן
- prsnsUpto17 – הפך לשלילי

ב. R^2– קטן

משימה חמש-עשרה

נבחין בתוצאות הרגרסיה עם המשתנים החדשים:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     c3   R-squared:                       0.540
Model:                            OLS   Adj. R-squared:                  0.536
Method:                 Least Squares   F-statistic:                     129.1
Date:                Sun, 04 Sep 2022   Prob (F-statistic):          3.63e-160
Time:                        18:59:07   Log-Likelihood:                -10090.
No. Observations:                1000   AIC:                         2.020e+04
Df Residuals:                     990   BIC:                         2.025e+04
Df Model:                           9
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept          1890.0943    891.705      2.120      0.034     140.246    3639.943
owner_yes         -1495.7423    971.393     -1.540      0.124   -3401.969     410.484
hhprsns             390.7537    375.172      1.042      0.298    -345.470    1126.977
incoment              0.8127      0.060     13.628      0.000       0.696       0.930
prsnsUpto17          -9.2816    223.808     -0.041      0.967    -448.474     429.911
hhprsns_sq          -13.1253     33.478     -0.392      0.695     -78.822      52.571
incoment_sq       -8.007e-06   8.89e-07     -9.009      0.000   -9.75e-06   -6.26e-06
ownerXhhprsns       191.7577    229.057      0.837      0.403    -257.736     641.252
incomentXowner        0.0560      0.045      1.238      0.216      -0.033       0.145
hhprsnsXincoment      0.0103      0.010      1.040      0.299      -0.009       0.030
==============================================================================
```

```
==============================================================================
Omnibus:                      594.081   Durbin-Watson:                   2.054
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            11087.473
Skew:                           2.345   Prob(JB):                         0.00
Kurtosis:                      18.624   Cond. No.                     5.06e+09
==============================================================================
```

- המקדמים המובהקים:
  - Incoment
  - incoment_sq

נשים לי כי מתוך אלו שהוספנו, מובהקים רק המקדמים בריבוע ולא משתני האינטראקציה

- R^2 עלה בכ – 0.023
- מדד ה AIC השתפר
- מדד ה BIC השתפר

משימה שש-עשרה

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     c3   R-squared:                       0.538
Model:                            OLS   Adj. R-squared:                  0.535
Method:                 Least Squares   F-statistic:                     231.2
Date:                Sun, 04 Sep 2022   Prob (F-statistic):          1.05e-163
Time:                        18:59:07   Log-Likelihood:                -10093.
No. Observations:                1000   AIC:                         2.020e+04
Df Residuals:                     994   BIC:                         2.023e+04
Df Model:                           5
Covariance Type:            nonrobust
==================================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept     600.8547    654.279      0.918      0.359    -683.072    1884.781
hhprsns       679.8699    327.951      2.073      0.038      36.314    1323.426
incoment        0.8638      0.048     18.162      0.000       0.770       0.957
prsnsUpto17   -46.3990    222.479     -0.209      0.835    -482.982     390.184
hhprsns_sq     -8.5989     33.275     -0.258      0.796     -73.897      56.699
incoment_sq  -7.44e-06    8.2e-07     -9.069      0.000   -9.05e-06   -5.83e-06
==============================================================================
Omnibus:                      589.589   Durbin-Watson:                   2.051
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            10723.326
Skew:                           2.330   Prob(JB):                         0.00
Kurtosis:                      18.351   Cond. No.                     3.01e+09
==============================================================================
```

המקדמים מובהקים:

- Hhprsns
- Incoment
- Incoment_sq

R^2 ירד ב 0.02

חלק III

משימה ראשונה

להלן חלוקת הנתונים שלנו לאימון ואימות:

```
x_train, x_test, y_train, y_test = train_test_split(x, y,
                                                    test_size=0.3,
                                                    random_state=42)
```

הממוצעים:

```
The average of Y train vector (c3): 20172.46783625731
The average of Y test vector (c3): 20476.472972972973
```

הערה: לא ביצענו הוספה או השמטה של משתנים נכון לנקודה זו. בהמשך לטובת הרצת האלגוריתמים, הורדנו מכמות המשתנים באופן שרירות על מנת לסיים את ריצה הקוד בזמן סביר.

משימה שנייה

<u>הערה:</u> אלגוריתם הבחירה לקדימה מומש בצמוד למצגות בכיתה והותאם על מנת למצוא את המודל הטוב ביותר גם על סמך "BIC" - זאת בניגוד לאלגוריתם בכיתה שמצא את המודל הטוב ביותר על סמך "AIC" בלבד.

להלן התוצאות המדווחות:

```
Results:
---------------
The best model according to the AIC:
 [list(['incoment', 'aptval', 'hhprsns', 'salaryhh', 'ownernfl_yes', 'religion_other', 'ch1014',
'devtown_south', 'incgross', 'ch1517'])]
The AIC of the best model:  [3485.145]

The best model according to the BIC:
 [list(['incoment', 'aptval', 'hhprsns', 'salaryhh', 'ownernfl_yes'])]
The BIC of the best model:  [3507.224]

The MSE of the best model (AIC):  [5.07411553e+08]
The MSE of the best model (BIC):  [9.14252378e+08]
```

משמה שלישית

התוצאות:

```
Results:
--------------------
The best model according to the AIC:
 [['ch01', 'ch1014', 'prsns18p', 'aptval', 'incoment', 'incgross', 'salaryhh', 'ownernfl_yes',
'devtown_south', 'religion_other']]
The AIC of the best model:  [3483.2]

The best model according to the BIC:
 [['ch1014', 'aptval', 'incoment', 'salaryhh', 'ownernfl_yes']]
The BIC of the best model:  [3506.745]

The MSE of the best model (AIC):  [5.14482389e+08]
The MSE of the best model (BIC):  [9.18037951e+08]
```

משימה רביעית

האלפא האופטימלית שנמצאה ע"י האלגוריתם היא: 5

המשתנים בעלי מקדם אפס:

- Yearsur
- Hhprsns
- Men18p
- Owner_yes

ה MSE שהתקבל: 1.3968451720401488e+19

להלן התוצאות המדווחות במלואם כפי שמוצגות בקוד:

```
---------------------
The 'best' lambda 5.0
The R^2 of the train data: -210901841370.03305
The R^2 of the test data: -140851036157.75806
---------------------
```

	Variables	Coefficients
0	Intercept	22236.402233
1	Unnamed: 0	-287.038029
2	yearsur	0.000000
3	hhprsns	0.000000
4	hhwernrs	-249.163831
5	ch01	981.930839
6	ch24	-217.961573
7	ch59	-192.196915
8	ch1014	1738.460577
9	ch1517	-642.561601
10	prsns18p	577.207928
11	men18p	0.000000
12	wom18p	627.510213
13	carval	744.101628
14	aptval	1894.947403
15	rooms	-2911.604222
16	roomflv	2900.751857
17	incoment	6943.490110
18	incgross	-4858.647384
19	inccap	-196.766895
20	salaryhh	1631.801813
21	salarysp	191.722378
22	salaryot	-687.954634
23	inctax	373.746653
24	internet_yes	-1885.975679
25	owner_yes	0.000000
26	ownernfl_yes	2812.170793
27	devtown_north	543.393083
28	devtown_south	-3305.023367
29	religion_jewish	-766.390813
30	religion_muslim	711.635897
31	religion_other	-9518.828780

משימה חמישית

נשווה את מוסברות המודלים למציאת המודל הטוב ביותר. להלן התוצאות:

```
The MSE of the LASSO model (test):  1.3968451720401488e+19
The Lasso model's score (rSquared):  0.49083509397818 45
Front (AIC) rSquared: 0.9174027333141157
Back (AIC) rSquared: 0.9185987486945754
```

מהממצאים שקיבלנו לאחר בדיקת כלל האלגוריתם קיבלנו מוסברות גבוה ביותר במקרה של אלגוריתם ה – backwards selection עם ערך של 0.92 ביחס לשאר. כאשר forward Selection נמצא מעט מאחוריו.