**כריית וייצוג מידע – פרויקט**

**חלק ראשון**

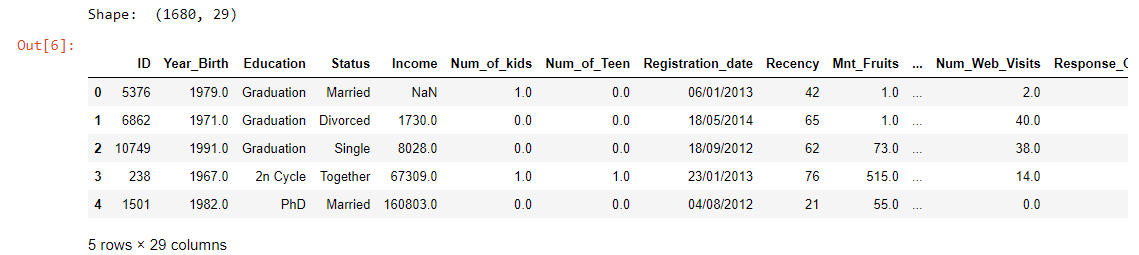
**הקדמה**

**שלב 1 - Data information**

ראשית, התקנו וייבאנו את כל הספריות הנדרשות לפרויקט, ולאחר מכן טענו את קובץ הcsv בעזרת ספריית panda (pd מעתה) לתוך משתנה df ( dataFrame).

הרצנו פקודת print("Shape: ", df.shape) כדי להבין את סדר הגודל של מסד המידע שלנו. וקיבלנו שהוא בגודל של 1680 רשומות מידע, עם 29 פיצ'רים לכל אחד.

ביצענו פקודת head כדי להציג את 5 השורות הראשונות ולהבין קצת איך נראה המידע.



כעת ביצענו אתה הפקודה הבאה כדי להבין בראייה מאקרו על המידע כולו:



נשתמש בה פעמים רבות בהמשך בזמן הכנת התרגיל כדי לוודא אילו חוסרים עדיין קיימים:

תמונה שמכילה שולחן

התיאור נוצר באופן אוטומטי

וכבר ניתן לראות שיש מספר רב של פיצרים חסרים. שבהם נטפל אחכ.

* המרנו את שנת הלידה לגיל (Age ) שיהיה מידע יותר נח מול העיניים לניתוח



* **הצגת המידע הסטטיסטי**

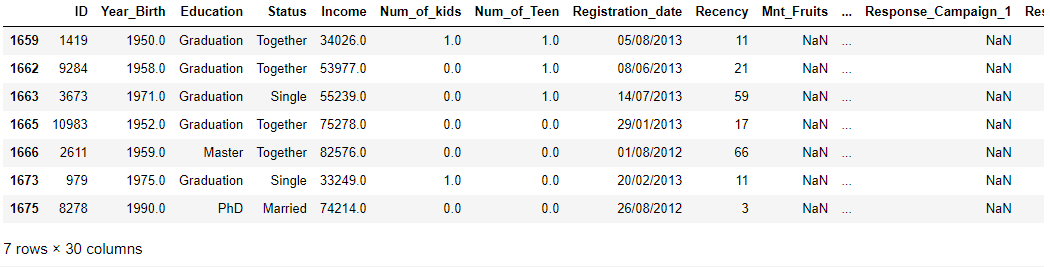
תמונה שמכילה שולחן

התיאור נוצר באופן אוטומטי

שממנו ניתן ללמוד על חלוקת הערכים לפי חציונים, רבעונים ועוד, וכן על הערך שחוזר הכי הרבה ועוד שנראה בהמשך.

* **נקיון מוקדם**

בשלב נקיון והכנת המידע שיגיע בהמשך גילינו שישנם 7 רשומות מאוד בעיתיות ומלבלבלות, אלו 7 אנשים שאין להם שום מענה ויחס להיענות שלהם לקמפיינים בכלל, ולכן בעצם אין להם ערך, יתרה מכך ,ראינו שאם נסיר אותם- שאר הטיפול במידע נהיה הרבה יותר סביר.



תמונה שמכילה שולחן

התיאור נוצר באופן אוטומטי

לכן ביצענו מחיקה של שורות אלו בפקודה הבאה:



ולאחר מכן הצגנו שוב את המידע, ורואים שאכן כעת כמות הפיצרים החסרים יותר סבירה.

תמונה שמכילה שולחן

התיאור נוצר באופן אוטומטי

* לאחר מכן פיצלנו את כל המידע שלנו לערך מטרה בנפרד.  
  ואת כל שאר המידע בנפרד (בשלב זה כבר הורדנו גם את ID שאין לו משמעות בכלל)

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

* כעת, חילצנו את כל העמודות בהם הערכים הם מספריים ולא מילוליים

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

* ואז חילצנו את כל המידע לשני חלקים, מידע נומרי ונומינלי

תמונה שמכילה טקסט

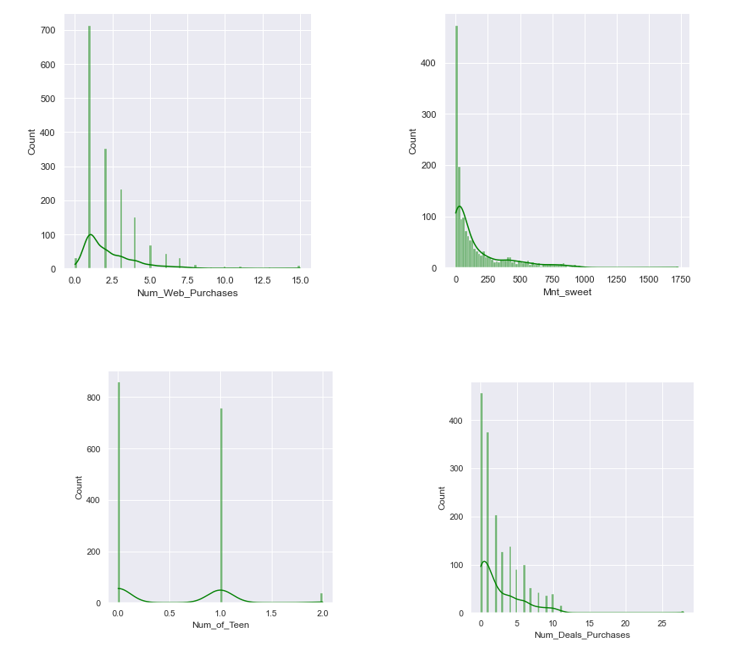
התיאור נוצר באופן אוטומטי

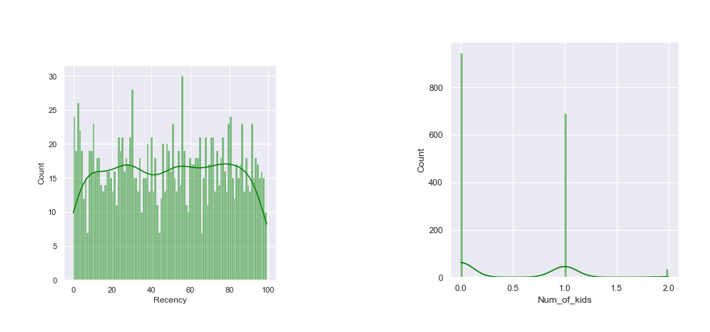
כעת – בשלב זה הצגנו לעצמינו (ונראה בהמשך) את הפלטים בצורת של היסטוגרמה והתפלגויות, והבחנו בכך שה"היענות לקמפיין 1-5" נמצא ביחד כחלק מה"ערכים המספריים" למרות שערכיהם בינארים (1-0) ולכן החלטנו להוסיף מסד נתונים אחר בשם nominal\_data2 שהוא יכיל את אותו מידע נומילי לעיל, בנוסף לכל התגובות ל"מענה לקמפיינים" כי הם מתפרשים עבורינו כקטוגריאלים.

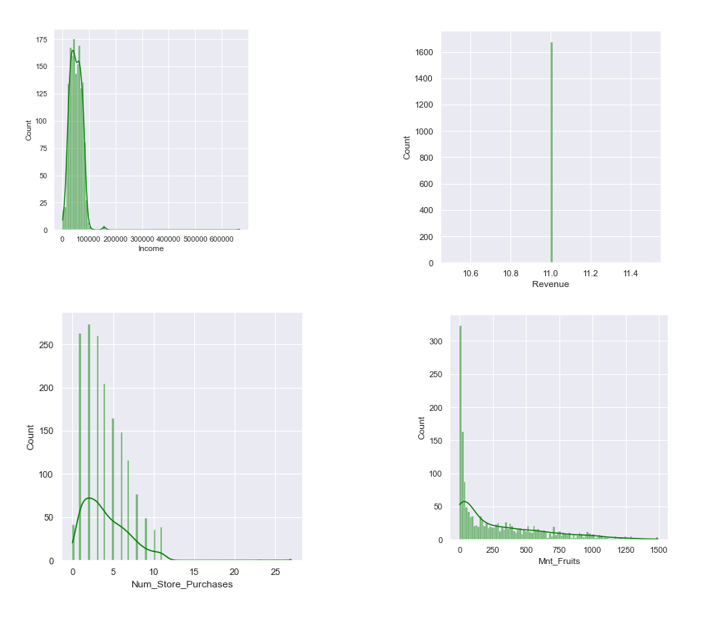
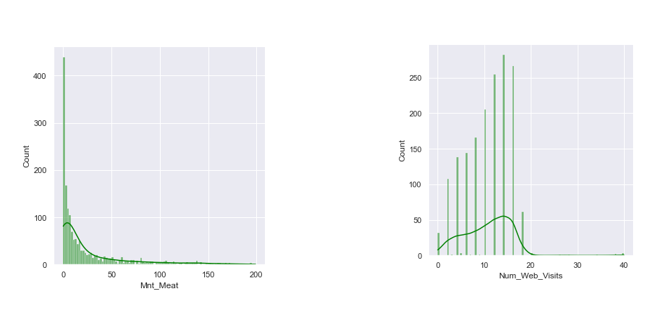
תמונה שמכילה טקסט

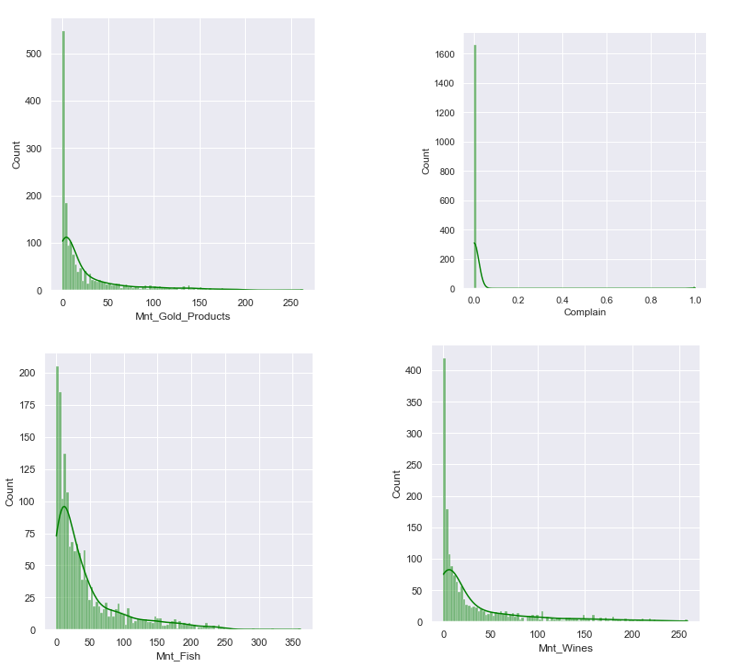
התיאור נוצר באופן אוטומטי

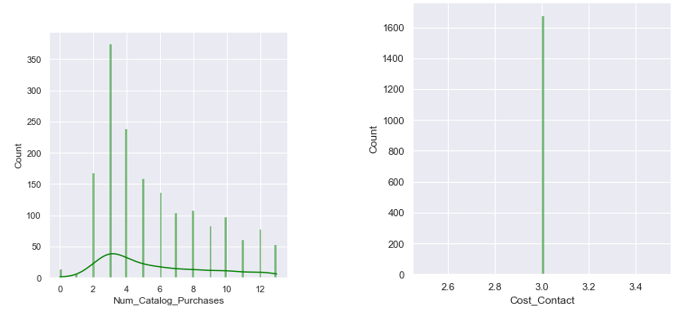
**הצגת הגרפים**

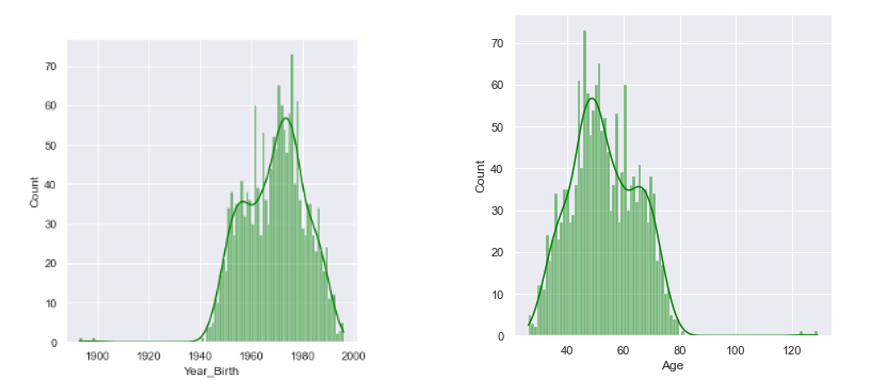












* שמנו לב שיש שני תמונות נוספות עם ערך קבוע ולא רלוונטי Revenue וה Cost\_contatc לכן העפנו אותם מהמידע בהמשך.
* חישוב Skewness :

תמונה שמכילה טקסט

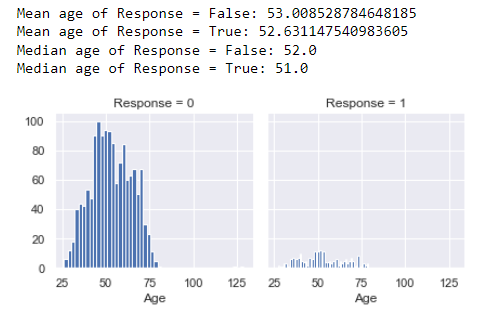
התיאור נוצר באופן אוטומטי

ניתן לראות שבערכים בהם היינו מצפים לקבל שההתפלגות תהיה נורמלית יחסית בלי הרבה ערכי קיצון הSKEW אכן נמוך כמו למשל בAGE, וכן במספר הילדים והמתבגרים בבית.

לעומת הזנב של ההכנסות, שבו ייתכן ויהיו ערכי קיצון (כמו שנגלה בהמשך שיש ערך אחד קיצוני ביותר שנשמיט אותו בהמשך) וכמובן שהזנב לערכי המענה לקמפיינים הוא לא רלוונטי כי הערכים שם בינארים ואין התפלגות נורמלית כלל. (לכן הסטנו אותם לערכי הנומינלים)

עצרנו בשלב זה לנסות לנתח ולגלות קולרציה אחת בעצמינו בין ערך המטרה לבין הגילאים:  
בהמשך בעזרת מפת החום, נמצא עוד קשרים רבים.  
להלן התוצאה:  
**תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי**



**המידע הנומינלי**

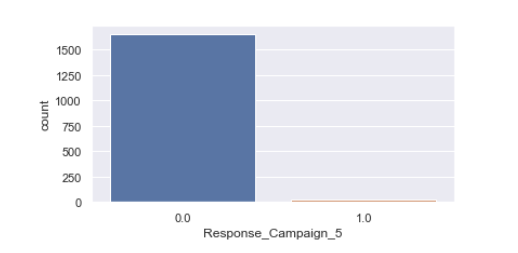
כעת, התחלנו לנתח ולהציג ויזואליזציה של המידע הנומינלי, וכן מידע סטטיסטי לגביו (לשים לב, בכוונה בהצגת המידע הסטטיסטי הצגנו את nominal\_data1 כי אם היינו מציגים את 2 אז היה הולך להציג רק את הערכים המספריים של הקמפיינים, לכן השארנו בשקט את המידע המילולי פה להצגה)

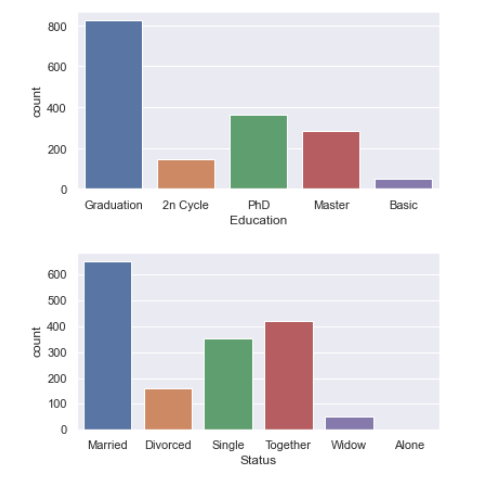
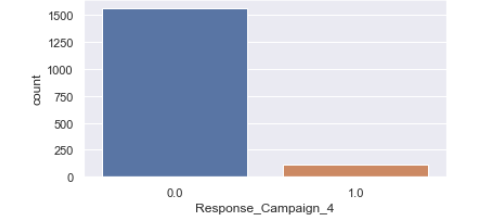
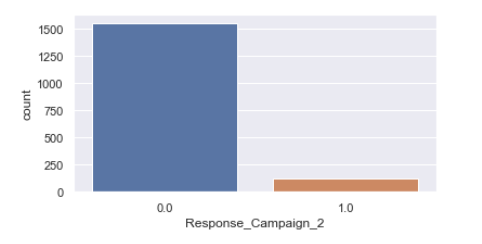
תמונה שמכילה טקסט

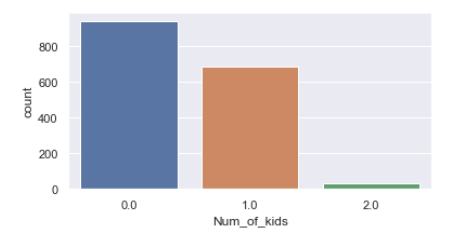
התיאור נוצר באופן אוטומטי

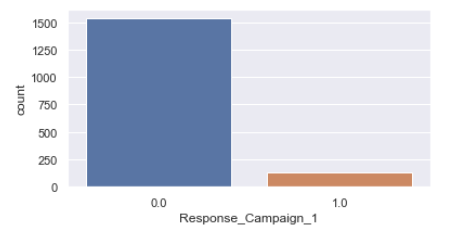
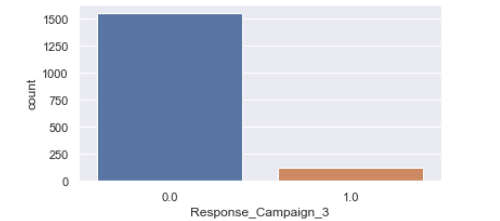
תמונה שמכילה טקסט

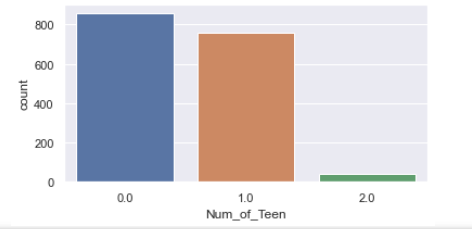
התיאור נוצר באופן אוטומטי

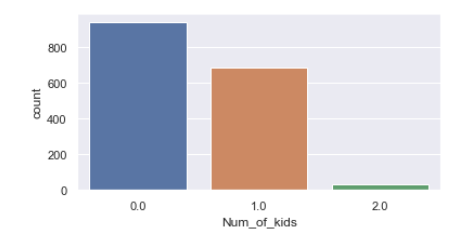
כעת הצגנו את המידע הויזאולי:





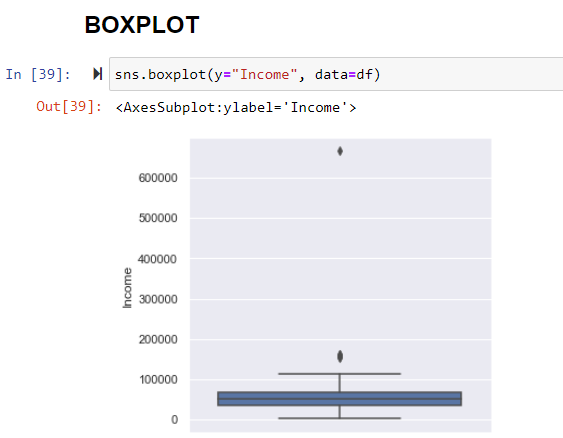


****

****

נציג חלק מהמידע בתצורת BOXPLOT

לדוגמא, המידע על השכר

****

זיהינו פה חריגה, לכן הלכנו לברר מי מסיט את הגרף והתמונה וגילינו שיש אדם אחד שמרוויח פי כמה יותר מכולם באופן חריג

תמונה שמכילה שולחן

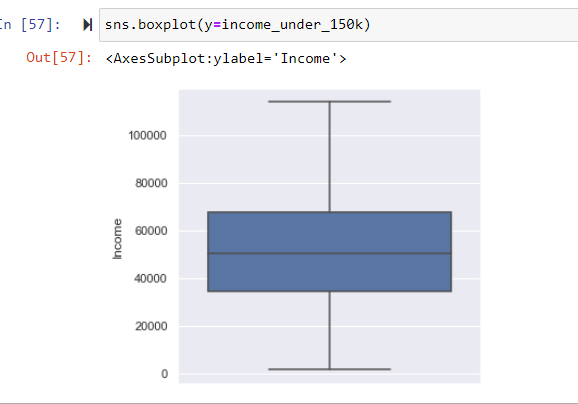
התיאור נוצר באופן אוטומטי

לכן החלטנו להעיף אותו מהמאגר המידע שכן הוא ערך חריג מאוד.  
ולאחר מכן יצרנו ערך חדש עם האנשים שמרוויחים פחות מ150 אלף בשנה.  
(ישנם אומנם כמה ספורים שמרוויחים יותר אבל לא בצורה קיצונית, השארנו אותם

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

וניתן לראות שכאשר הוצאנו את החריגים אזיי ההתלפגות נראית הגיונית ונורמלית מאוד.



ניסינו לבצע גם הצגה של SCATTER PLOT עם ערכים שונים, אך לא קיבלנו משו משמועתי ללמוד ממנו כמו שראינו בתרגול.  
נצרף דוגמא אחת שמצאנו



**קורלציות בין פיצרים והשלמת מידע**

כשלב בפני עצמו וכן כהכנה לנקיון המידע, רצינו לבצע פה את שיטת הקורלציה של פירסון ולהציג במפת חום את היחס בין כל המשתנים. גם מעניין לעצמו וגם יעזור בהמשך להשלים מידע עם הגיון.

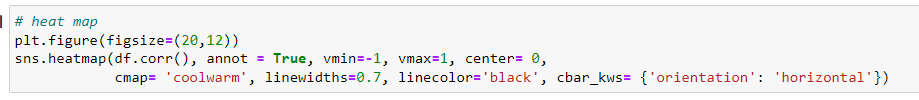
לצורך כך , קבענו את הערכים המילולים כמו סטטוס וכדומה לערכים מספריים.

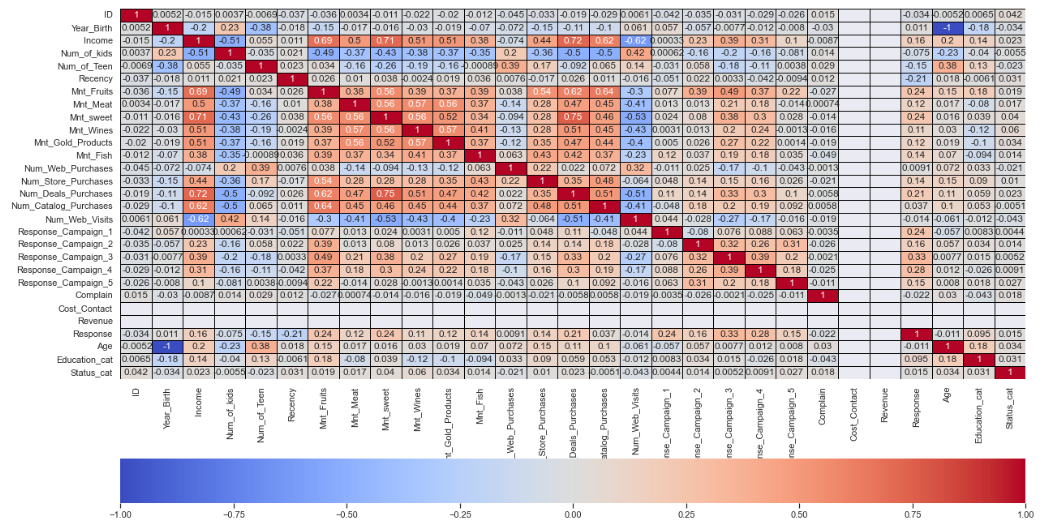
והפעלנו את פונקציה שמוצאת קורלציה

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

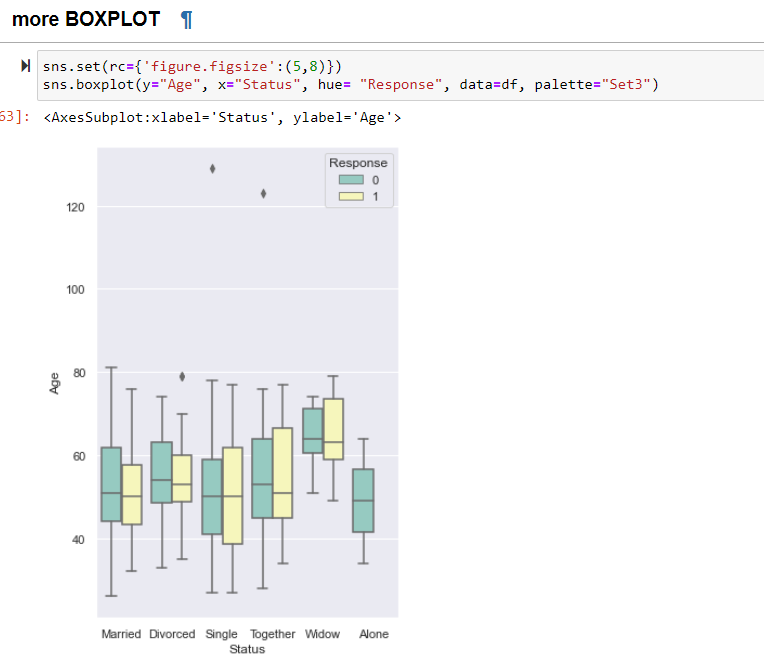
מפת החום:





לא לפי הסדר, אבל כך מופיע במחברת, הצגנו עוד כמה boxplot ששופכים מידע מעניין

סטטוס לעומת ערך מטרה



חינוך והשכלה לעומת ערך מטרה



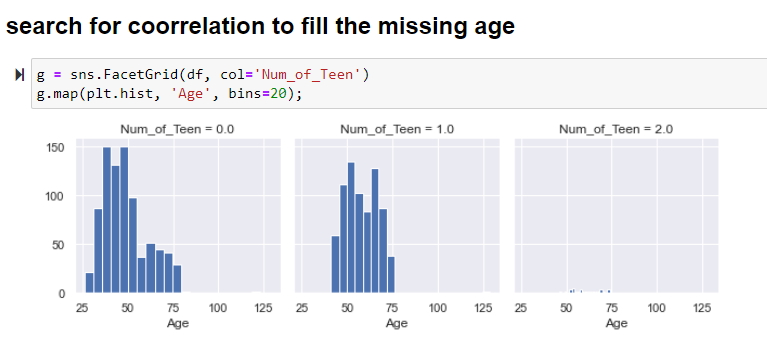
**Data Cleaning**

ניגש לשלב נקיון המידע מדברים לא רלנווטים וכן השלמת המידע של דברים חסרים. נעזרנו במפת חום כדי לחפש קולרציה והגיון מושכל להשלמת המידע

**טיפול בחוסר בAGE:**

ראינו שיש קשר במפת חום בין הגיל לבין כמות הילדים המתבגרים בבית. החלטנו לבדוק קורלציה ביניהם וכתלות במספר המתבגרים.

ואילו התוצאות:

****

**תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי**

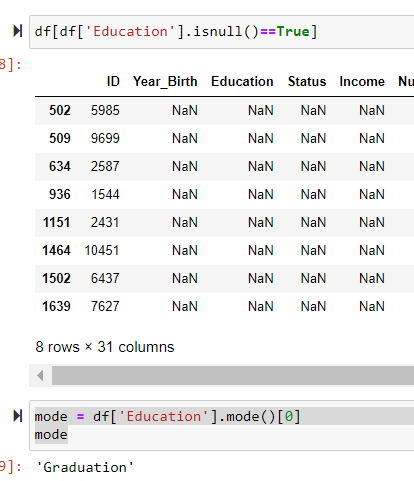
יצא בדומה לכיתה, לכן השלמנו את הגילאים החסרים לפי הCLASS שיצא שהם נמצאים בו לפי כמות המתבגרים:

ואת השאר מילאנו לפי הממוצע שמצאנו

**תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי**

**טיפול בחוסרים של השכלה.**

****

למרות קשר מעניין שראינו בהמשך בין השכלה לבין צריכת פירות, מכיוון שכמות החוסרים לא גדולה החלטנו למלא את הערכים החסרים פשוט בערך השכיח שזה "בוגר תואר ראשון"

**טיפול בחוסרים של סטטוס זוגי**

גם חלק זה ניסינו לנתח לפי קורלציות לפיצרים אחרים, אך לא ראינו משהו חזק במיוחד, בין השאר, כמות החוסרים פה קטנה ולכן מילאנו שוב לפי הערך השכיח

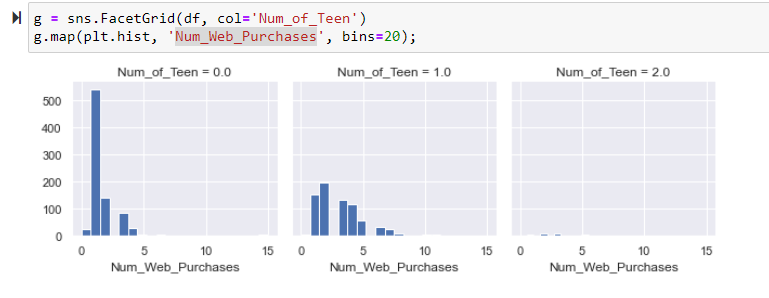
תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

**טיפול בחוסרים של רכישות באינטרנט:**

פה שמנו לב ליחס בין כמות הרכישות לבין מספר המתבגרים בבית ,שיש לזה הסבר דווקא מאוד הגיוני😊

ולכן ביצענו בדיקות והשלמות בדומה לחלק לעיל.



תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

את ההשלמות ביצענו בצורה חכמה, לפי הממוצע בכל קבוצה בפני עצמה.

**טיפול בחוסרים של כמות ילדים ומתבגרים בבית**

גם פה ראינו שכמות החוסרים לא גדולה, ולכן השלמנו אותם ע"י הערך השכיח שקיים. (כמובן שיכול לצאת מצב להשלים כמות ילדים לבן אדם לא נשוי וכדומה אבל זה ערכים זניחים וגם ייתכן שלאדם לא נשוי יהיו ילדים)

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

**טיפול בחוסרים של רמת הכנסה**

זיהינו יחס מעניין בין רמת ההכנסה לבין כמות הרכישות שבוצעו עם הנחה. (מעניין אם אולי לאנשים עם משכורת גבוהה יותר יש יותר מודעות כלכלית לקניות מושכלות ובהנחה וכו')

תמונה שמכילה שולחן

התיאור נוצר באופן אוטומטי

גם פה זיהינו חלוקה למספר מחלקות מסוימות ואת השלמת המידע החסר ביצענו בהתאם.

**פיצרים אחרונים חסרים – כמות מתוקים והיענות לקמפיין 1**

בשני האחרונים הללו, זיהינו שוב כמות קטנה מאוד של חוסרים ולכן ביצענו השלמה.  
במתוקים שזו כמות נומרית ביצענו לפי הממוצע,  
ובהיענות לקמפיין שזה ערכים בינארים השלמנו לפי השכיח.

סיימנו את שלב הכנת המידע ונקיון המידע והסרנו עמודות לא רלוונטיות, והשלמנו את החוסרים בעמודות שבהם היה חסר, ככל הניתן לפי הגיון ושכל.

**And Data reduction Data transformation**

ביצענו שיטות לנורמליזציה שלמדנו בכיתה, הן כדי להתלמד בהן והן כדי להכין את המידע לקראת תהליך הPCA שדורש ערכים מנורמלים.

השתמשנו בשיטת הMIN\_MAX לנרמל מספר עמודות בבת אחת שעליהם נרצה לערוך pca בהמשך

**תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי**

עוד הכנה שלא טופלה לפני הpca היא המרת התאריך שהיה נתון בפורמט מסוים לערכים מחולקים.  
(נכון, שבסוף מידע זה לא השתתף בPCA אבל תיקנו אותו בכל זאת)

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

**נציג את המידע שיצא מהPCA**

**(הערה:** לפני הגשת התרגיל הרצנו הרצה חוזרת של הכלומשהו לא עבד כשורה, איבדנו את תמונת הPCA, דווקא יצא מידע מעניין וחלוקה יפה

נצרף את הקוד בתקווה שבהרצה נוספת הדברים יסתדרו)

**כעת ביצענו Discretization**

ביצענו שני שיטות שלמדנו בכיתה

לחלק את קבוצת הגילאים לפי קבוצות בגודל שווה

או לפי קבוצות שהגדרנו מראש את טווחי הגילאים בכל קבוצה (גם אם גודלם לא שווה