כריית מידע וייצוג מידע (83676)
דו"ח מסכם - פרויקט חלק 1

עידו שר שלום 21240146, תומר גריבה 325105625

הקדמה:

קמפיינים שיווקיים הם אחת הדרכים היעילות ביותר בגישה לאנשים למכירת מוצר או שירות.
עם זאת, נדרשות השקעות רבות כדי להוציאם לפועל. יתר על כן, המספר הרב של קמפיינים שיווקיים אלו לאורך זמן הפחית את השפעתו על הציבור הכללי. כל אלה כמו גם לחצים כלכליים ותחרותיים הובילו את מנהלי השיווק להשקיע בקמפיינים מכוונים תוך בחירה מדויקת וקפדנית של האנשים.
מטרת הפרויקט היא להגביר את היעילות של קמפיין השיווק הישיר על ידי חיזוי לקוח אשר יענה בחיוב להצעה למוצר או שירות.
חיזוי הלקוח מתבצע על ידי מודל למידת מכונה אשר אומן מראש על נתוני לקוחות עבר.
בחלק זה של הפרויקט נבצע את שלב הכנת הנתונים (preprocessing) תוך חקירה והבנה של הדאטא הקשרים והתלויות בו.
כמו כן, נעזר בכלים סטטיסטיים וויזואליים כדי לנתח את מסד הנתונים ונממש אותם.
נדגיש כי שלב זה הוא חלק עיקרי וחשוב ממטרת הפרויקט שכן, עיבוד הנתונים הוא הכרחי לדוגמה, סינון נתונים אשר לא רלוונטיים לנו ל- שיערוך ערך המטרה.

לכן, בחלק זה של הפרויקט נבצע:

- נלמד וננתח את מסד הנתונים, תכונותיו ופילוג הערכים.
- סטטיסטיקה של הדאטא והתאמתו להתפלגויות מוכרות (נורמלי, אחיד, גיאומטרי, אקספוננציאלי...).
- ננתח קורלציות בין מאפיינים, ונתונים סטטיסטיים כגון, skewness ו- median.
- נציג את ויזואליזציית הדאטא תוך שימוש בכלים שונים (boxplot, scatter plot וכו'...)
  אשר יעזרו לנו להבנה ולמידה על קשרים בין משתנים שונים.
- ננקה את המידע על ידי מילוי ערכים חסרים של מסד הנתונים, בנוסף, נבחן אי התאמה בדאטא.
- נזהה מאפיינים מיותרים אותם נסיר מהדאטאסט גם, נוסיף מאפיינים לשיפור ולשערוך המשתנים.
- ניישם שיטת PCA להורדת מימדים של מערך הנתונים.
- נעשה טרנספורמציה לנתונים נפעיל שיטות לנרמול הדאטא, דיסקרטיזציה של הדאטא.

<u>הכרת המידע ומאפייניו:</u>

המידע שקיבלנו מכיל נתונים הקשורים לקמפיינים שיווקיים שהיו. בין ה- features יש פרטים אישיים של הלקוחות, מידע על הרכישות שלהם ועוד...

למידע שקיבלנו יש 29 מאפיינים והם:

```
#   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   ID                     1680 non-null   int64
 1   Year_Birth             1651 non-null   float64
 2   Education              1672 non-null   object
 3   Status                 1646 non-null   object
 4   Income                 1609 non-null   float64
 5   Num_of_kids            1672 non-null   float64
 6   Num_of_Teen            1660 non-null   float64
 7   Registration_date      1680 non-null   object
 8   Recency                1680 non-null   int64
 9   Mnt_Fruits             1673 non-null   float64
 10  Mnt_Meat               1673 non-null   float64
 11  Mnt_sweet              1659 non-null   float64
 12  Mnt_Wines              1673 non-null   float64
 13  Mnt_Gold_Products      1673 non-null   float64
 14  Mnt_Fish               1673 non-null   float64
 15  Num_Web_Purchases      1651 non-null   float64
 16  Num_Store_Purchases    1673 non-null   float64
 17  Num_Deals_Purchases    1673 non-null   float64
 18  Num_Catalog_Purchases  1673 non-null   float64
 19  Num_Web_Visits         1673 non-null   float64
 20  Response_Campaign_1    1662 non-null   float64
 21  Response_Campaign_2    1673 non-null   float64
 22  Response_Campaign_3    1673 non-null   float64
 23  Response_Campaign_4    1673 non-null   float64
 24  Response_Campaign_5    1673 non-null   float64
 25  Complain               1673 non-null   float64
 26  Cost_Contact           1673 non-null   float64
 27  Revenue                1673 non-null   float64
 28  Response               1680 non-null   int64
dtypes: float64(23), int64(3), object(3)
```

כאשר ליד כל feature מוזכר כמה ממנו הם לא null (כלומר לכמה דגימות של feature זה לא חסרים ערכים), כמו כן גם סוג הערך השמור ב- feature.

ניתן לראות כי הערך ממנו חסרות הכי הרבה דגימות הוא Income, וזה הגיוני מכיוון שלרוב, אנשים לא חושפים את משכורתם.

כעת, נסביר על ה- attributes של מסד הנתונים:

1. ID - מספר הזהות הלקוח
2. Year_Birth - שנת לידת הלקוח
3. Education - רמת ההשכלת הלקוח
4. Status - מצב משפחתי של הלקוח
5. Income - הכנסה שנתית של הלקוח
6. Num_of_kids - מספר ילדים קטנים של הלקוח
7. Num_of_Teen - מספר נערים מתבגרים של הלקוח
8. Registration_date - תאריך רישום הלקוח בחברה
9. Recency - מספר הימים שעברו מאז הרכישה האחרונה של הלקוח
10. Mnt_Fruits - כמות הכסף אשר הוציא הלקוח על פירות בשנתיים האחרונות
11. Mnt_Meat - כמות הכסף אשר הוציא הלקוח על בשר בשנתיים האחרונות
12. Mnt_sweet - כמות הכסף אשר הוציא הלקוח על מוצרי ממתקים בשנתיים האחרונות
13. Mnt_Wines - כמות הכסף אשר הוציא הלקוח על יין בשנתיים האחרונות
14. Mnt_Gold_Products - כמות הכסף אשר הוציא הלקוח על מוצרי זהב בשנתיים האחרונות
15. Mnt_Fish - כמות הכסף אשר הוציא הלקוח על דגים בשנתיים האחרונות
16. Num_Web_Purchases - כמות רכישות הלקוח באתר האינטרנט של החברה בחודש האחרון
17. Num_Store_Purchases - כמות רכישות הלקוח בחנויות החברה (פיזית) בחודש האחרון
18. Num_Deals_Purchases - כמות עסקאות בהנחה שביצע הלקוח בחודש האחרון
19. Num_Catalog_Purchases - כמות הרכישות שביצע הלקוח מהקטלוג בחודש האחרון
20. Num_Web_Visits - כמות הביקורים של הלקוח באתר החברה בחודש האחרון
21. Response_Campaign_1 - בינארי, 1 לקוח קיבל את הצעת החברה בקמפיין הראשון, אחרת 0
22. Response_Campaign_2 - בינארי, 1 לקוח קיבל את הצעת החברה בקמפיין השני, אחרת 0
23. Response_Campaign_3 - בינארי, 1 לקוח קיבל את הצעת החברה בקמפיין השלישי, אחרת 0
24. Response_Campaign_4 - בינארי, 1 לקוח קיבל את הצעת החברה בקמפיין הרביעי, אחרת 0
25. Response_Campaign_5 - בינארי, 1 לקוח קיבל את הצעת החברה בקמפיין החמישי, אחרת 0
26. Complain - בינארי, 1 אם הלקוח התלונן בשנתיים האחרונות, 0 אם לא
27. Cost_Contact - עלות יצירת הקשר עם הלקוח
28. Revenue - הכנסה שהתקבלה לאחר שהלקוח קיבל את ההצעה בקמפיין
29. Response - בינארי, 1 אם הלקוח קיבל את ההצעה בקמפיין האחרון, 0 אם לא

נתונים אלו מייצגים צרכן פוטנציאלי, לקוח כללי.
ערך המטרה אותו נרצה לחזות הוא Response. כלומר, בהינתן פרטים אודות צרכן מסוים נרצה לדעת האם הוא יקבל את ההצעה לקמפיין או לא.
כך נוכל לאתר קבוצות אנשים להם סיכוי גבוה ברכישת מוצרים בחברה, תוך שימוש בנתונים ידועים עליהם.

למערך הנתונים יש את ה- data types הבאים כאשר ה- data type הוא תלוי attribute.

data type	כמות features
float64	23
int64	3
object	3

כלומר, ישנם 26 מאפיינים אשר סוגי הערכים הרשומים בהם הם מספריים (numeric).
3 מאפיינים אשר סוגי הערכים הרשומים בהם הם מסוג object, אובייקט כללי.
עבור תכונות מסוג object נבדוק חוסר תיאום לערך בפועל, במידה וקיים נשנה אותו כדי שנוכל לנתח את הדאטא בצורה טובה יותר.

כעת, בעזרת describe נציג כמה ערכים אודות ה- features מקבלים ערכים נומריים:

	ID	Year_Birth	Education	Status	Income	Num_of_kids	Num_of_Teen	Registration_date	Recency	Mnt_Fruits	...	Num_Web_
count	1680.000000	1651.000000	1672	1646	1609.000000	1672.000000	1660.000000	1680	1680.000000	1673.000000	...	1673.00
unique	NaN	NaN	5	6	NaN	NaN	NaN	634	NaN	NaN	...	
top	NaN	NaN	Graduation	Married	NaN	NaN	NaN	14/02/2013	NaN	NaN	...	
freq	NaN	NaN	830	653	NaN	NaN	NaN	10	NaN	NaN	...	
mean	5584.735714	1969.047244	NaN	NaN	51983.554382	0.454545	0.503614	NaN	48.890476	303.676031	...	10.60
std	3233.716033	11.937421	NaN	NaN	26567.679664	0.538492	0.544011	NaN	29.091872	340.672889	...	5.03
min	0.000000	1893.000000	NaN	NaN	1730.000000	0.000000	0.000000	NaN	0.000000	0.000000	...	0.00
25%	2862.500000	1959.500000	NaN	NaN	34596.000000	0.000000	0.000000	NaN	24.000000	23.000000	...	6.00
50%	5511.000000	1970.000000	NaN	NaN	50611.000000	0.000000	0.000000	NaN	50.000000	167.000000	...	12.00
75%	8395.500000	1978.000000	NaN	NaN	67716.000000	1.000000	1.000000	NaN	74.000000	508.000000	...	14.00
max	11191.000000	1996.000000	NaN	NaN	666666.000000	2.000000	2.000000	NaN	99.000000	1493.000000	...	40.00

לא ניתן היה להציג את כל הטבלה כיוון שיש יותר מדי עמודות.
בטבלה ניתן לראות כמה דגימות שהן לא null יש לכל מאפיין (בעזרת שדה ה- count), ממוצע הדגימות, סטיית התקן, ערך המינימלי, ערך מקסימלי והאחוזון ה- 25, 50, 75.
לדוגמה, ניתן לראות כי שנת הלידה המינימלית של לקוח היא 1893.

כעת, נסביר גם על ה features שאינם נומריים.
כפי שנאמר ישנם 3 כאלו והם: Education, Status, Registration_date.
נבדוק חוסר תיאום לערך בפועל.
עבור Registration_date, תאריך ההרשמה של הלקוח לחברה, נוכל להמיר את התאריך מ object ל- datetime64. כך, נוכל להציג ולנתח אותו בעזרת הכלים של ספריית pandas.
עבור תאריך ההרשמה, השתמשנו ב describe וקיבלנו:

```
count                  1680
mean     2013-07-13 20:00:00
min      2012-01-08 00:00:00
25%      2013-01-25 18:00:00
50%      2013-07-12 00:00:00
75%      2013-12-31 06:00:00
max      2014-12-06 00:00:00
```

כלומר, התאריך הראשון בו לקוח הצטרף לחברה הוא 8.1.2012

עבור Status, סטטוס מערכת היחסים של הלקוח, הערכים אשר הוא יכול לקבל הינם:

- Married - נשוי
- Divorced - גרוש
- Single - רווק
- Together - בזוגיות (חברה/חבר)
- Widow - אלמנה
- Alone - לבד

בנוסף, עבור תכונה זו ישנם גם כמה ערכים חסרים (34)

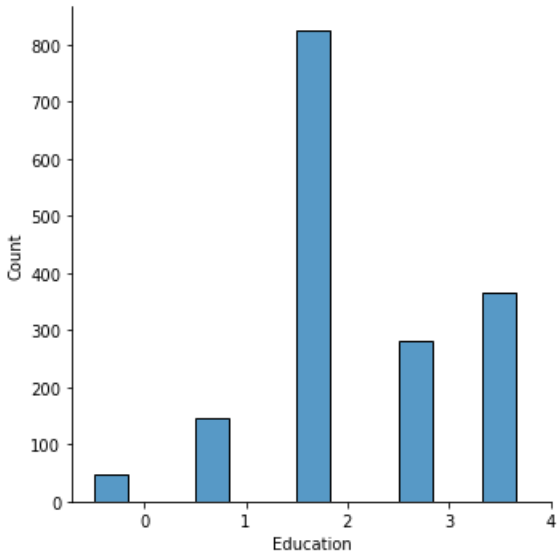
את מאפיין זה נעדיף להשאיר כמו שהוא, נומינלי. זאת מכיוון שאין סדר כלשהו או אפס מוחלט עבורו.

עבור attribute של Education, הערכים אשר הוא יכול לקבל הינם:

- Graduation - סיום לימודי תואר ראשון
- 2n Cycle - תואר דו-שנתי
- Phd - דוקטורט
- Master - תואר שני
- Basic - חינוך בסיסי

בנוסף, עבור תכונה זו ישנם גם כמה ערכים חסרים (8)

ניתן להתייחס למאפיין זה כאורדינלי (Ordinal) שכן, ניתן לדרג לקוח על פי ההשכלה שלו בסדר עולה.
Phd > Master > Graduation > 2n Cycle > Basic
לכן, מיפינו את ה attribute הזה למספרים בצורה הבאה:
"Basic"-> 0, “2n Cycle” -> 1, “Graduation” -> 2, “Master” -> 3, “PhD” -> 4



למרות זאת, גם דירוג זה הוא לאו דווקא אינפורמטיבי שכן ההשכלה של אדם לאו דווקא מעידה על הצורך שלו בשירות או בקניית מוצר.

Data statistics

בחלק זה, נפצל את הדאטא של מסד הנתונים לנומרי ונומינלי.

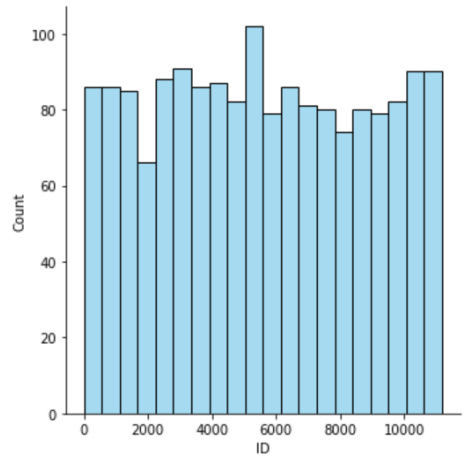
תחילה, נראה את הממוצע, סטיית התקן, החציון, האחוזונים ה- 25, 50, 75, את המקסימום והמינימום של המידע הנומרי, בעזרת שימוש בפונקציית describe:

	ID	Year_Birth	Income	Num_of_kids	Num_of_Teen	Registration_date	Recency
count	1680.000000	1680.000000	1680.000000	1680.000000	1680.000000	1680	1680.000000
mean	5584.735714	1935.057738	49786.630357	0.452381	0.497619	2013-07-13 20:00:00	48.890476
min	0.000000	0.000000	0.000000	0.000000	0.000000	2012-01-08 00:00:00	0.000000
25%	2862.500000	1959.000000	32619.750000	0.000000	0.000000	2013-01-25 18:00:00	24.000000
50%	5511.000000	1970.000000	49095.000000	0.000000	0.000000	2013-07-12 00:00:00	50.000000
75%	8395.500000	1977.000000	66999.000000	1.000000	1.000000	2013-12-31 06:00:00	74.000000
max	11191.000000	1996.000000	666666.000000	2.000000	2.000000	2014-12-06 00:00:00	99.000000
std	3233.716033	256.809015	28025.634948	0.538118	0.543515	NaN	29.091872

ניתן לראות את ה attribute של תאריך ההרשמה, מכיוון שהגדרנו את התאריך בעזרת אובייקט datetime64 של pandas.

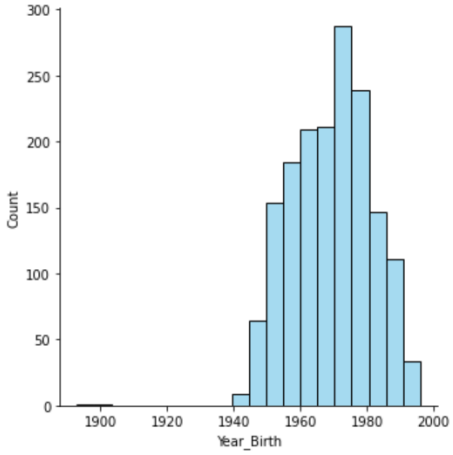
כמו כן, ביצענו ויזואליזציה של המידע בעזרת היסטוגרמות כדי להבין מהי ההתפלגות המתאימה.
עבור כל attribute (נומרי), נציג את כמות הרשומות במסד הנתונים כתלות ב- attribute.
דוגמאות:

- המאפיין ID:



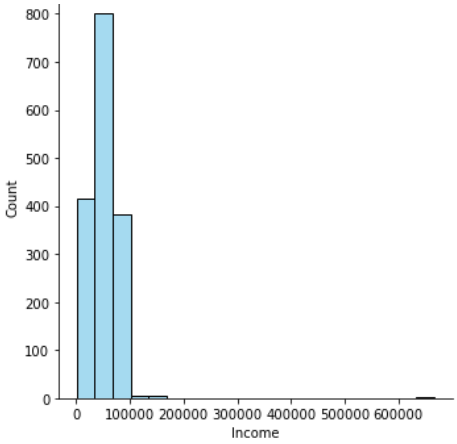
ניתן לראות כי ה- ID מתפלג בקירוב בהתפלגות אחידה.
נזכור כי ID הוא מזהה חד ערכי כלומר, לכל לקוח יש מספר זהות ייחודי משלו ובנוסף ה- ID נבחרים ללא תלות ב attributes האחרים. טווח ערכי ה- ID הוא (11191, 0) כאשר ישנם מספרי ID חסרים בתחום (לדוגמה, 5377 לא מופיע). חוסרים אלו, יוצרים אחידות שאינה נראית במדויק.
קיבלנו בקירוב כמות זהה של ID בכל תא בהיסטוגרמה כלומר, התפלגות אחידה.

- המאפיין Year_Birth



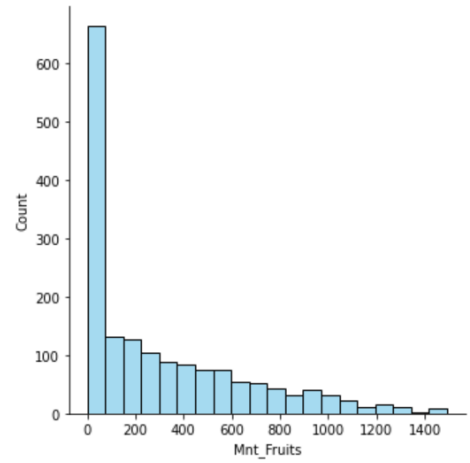
ניתן לראות כי שנת הלידה של הלקוחות מתפלגת בערך בהתפלגות נורמלית, עם $\mu = 1935,\ \sigma = 256.8$.
הדבר מתאים למה שלמדנו ממשפט הגבול המרכזי. מכיוון שאין תלות בין האנשים במדגם, ובפרט בין שנות הלידה שלהם, (בקירוב הם גם שווי התפלגות, עם טווח גילאים זהה (ללא ה outliers)), כך, לפי משפט הגבול המרכזי, ההתפלגות הכוללת היא בקירוב נורמלית.

- המאפיין Income



כמו שנת הלידה, כך גם ההכנסה מתפלגת לפי התפלגות נורמלית.

- המאפיין Mnt_Fruits

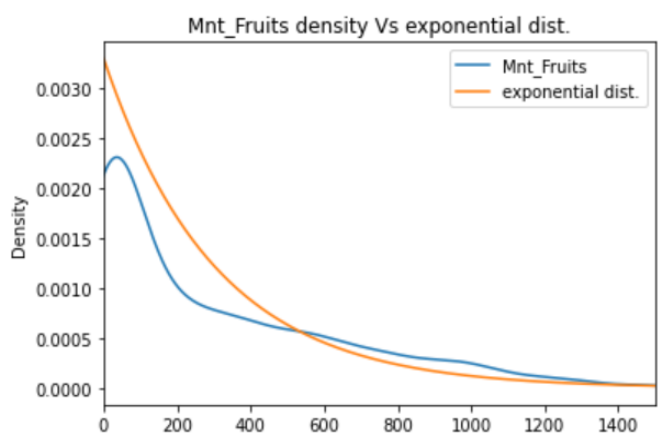


כמות הכסף שהוציאו הלקוחות על פירות, מתפלגות בהתפלגות דומה להתפלגות אקספוננציאלית ($f_X(x) = \lambda e^{-\lambda x}$).

כאשר בהתפלגות אקספוננציאלית, האומדן עבור λ הוא $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \mu^{-1}$.

עבור ה- Mnt_Fruits התוחלת היא $\mu = 302.4$ ולכן $\hat{\lambda} = \mu^{-1} = 0.0033$.

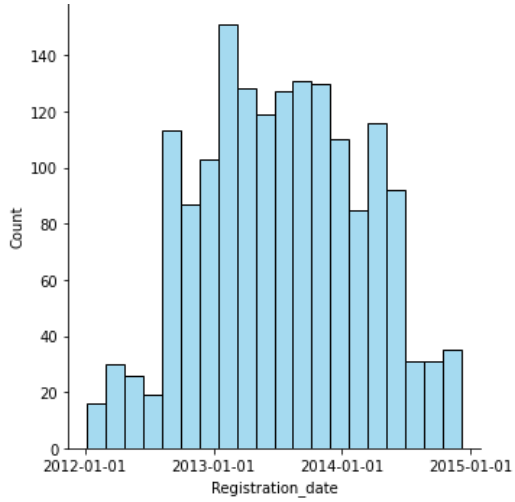
מצורף גרף המראה את ההתאמה של Mnt_Fruits להתפלגות מעריכית:



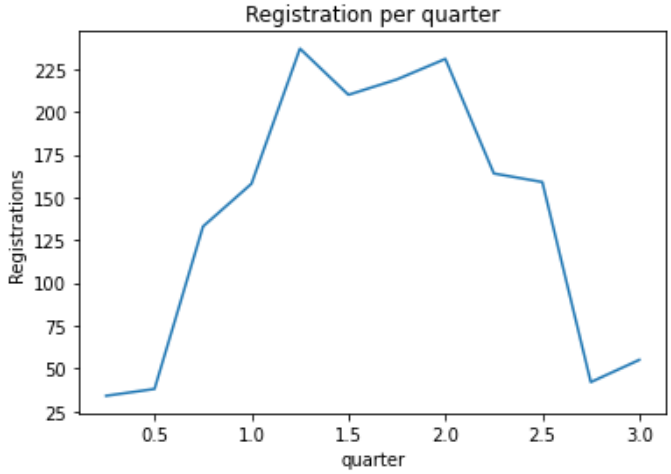
ואכן, התפלגות Mnt_Fruits היא בקירוב התפלגות אקספוננציאלית.

- המאפיין Registration_date

מאחר ואנו מתייחסים למאפיין זה כנומרי ניתן להציגו בהיסטוגרמה.
נציין כי בהצגת היסטוגרמה זו נתקלנו בבעיה בהצגת ערכי ציר x (חפיפה של תאריכים בציר) ולכן, ה- plot של attribute זה הוא בנפרד מהשאר.
התוצאה שקיבלנו:



כפי שניתן לשים לב, נראה כי תאריכי רישום הלקוחות מתפלגים בקירוב בצורה גאוסיאנית על פני טווח השנים, הדבר אינפורמטיבי שכן, נוכל ללמוד מכך שהקמפיינים אשר בוצעו בשנים 2013,2014 היו אפקטיביים ומשכו כמות רבה של לקוחות. כמו כן, נציין כי אין הגיון שתאריכי ההרשמה יתפלגו בצורה נורמלית אך זה המצב.
כדי לנתח מאפיין זה בצורה טובה יותר נרצה להבין באיזה רבעונים בטווח השנים הרישום לקמפיינים היה השכיח ביותר, כך נוכל ללמוד חודשים אפקטיביים בשנה להפעלת שירותי החברה ולאתר בחודשים אלו כמות לקוחות גדולה יותר.



מתוארים כמות הרישומים כתלות ברבעונים בטווח השנים 2012,2013,2014 (טווח השנים הפעיל),
ניתן לראות כי הרבעון האחרון בכל שנה הוא הרבעון החזק ביותר ולכן, נסיק כי ברבעון זה מומלץ להגביר את פעילות הקמפיין.

נסכם מהגרפים את התפלגויות ה- attributes:

From the graphs we can infer the distributions:

- Normal

Year_Birth, Income, Registration_date

- Exponential

Mnt_Fruits, Mnt_Meat, Mnt_sweet, Mnt_Wines, Mnt_Gold_Products, Mnt_Fish

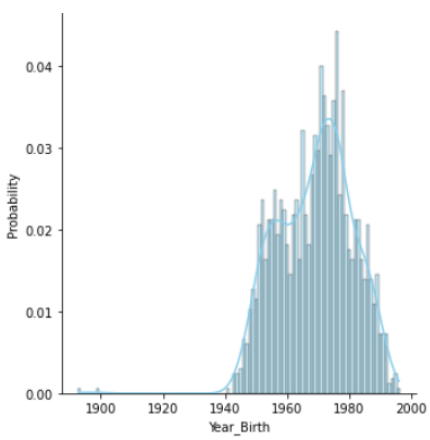
- Uniform

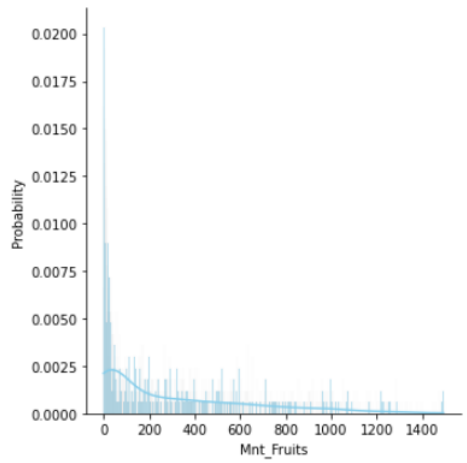
ID, Recency

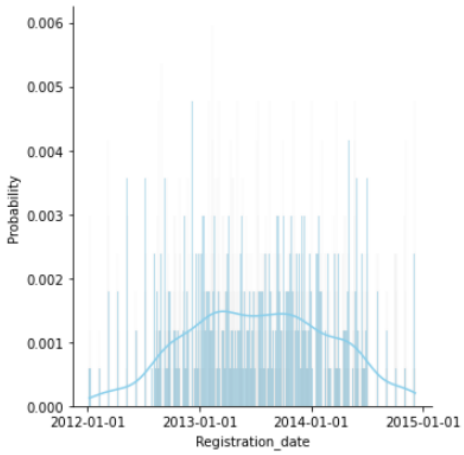
- Other

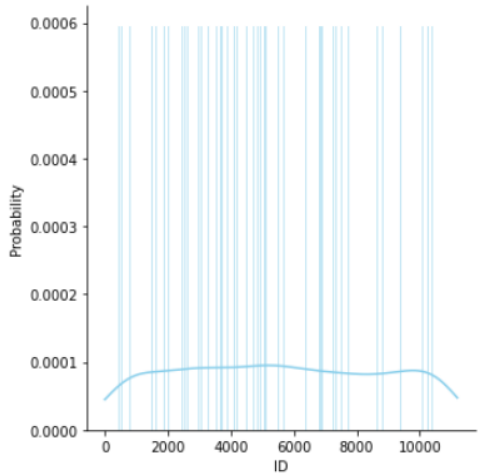
Num_Web_Purchases, Num_Store_Purchases, Num_Deals_Purchases, Num_Catalog_Purchases, Num_Web_Visits, Cost_Contact, Revenue
Response_Campaign_1-5, Complain, Num_of_kids, Num_of_Teen

באופן דומה, ניתן לראות את צפיפות התפלגויות אלו בעזרת עקומה חלקה על פני הגרף (ייצוג בדיד של הנתונים).

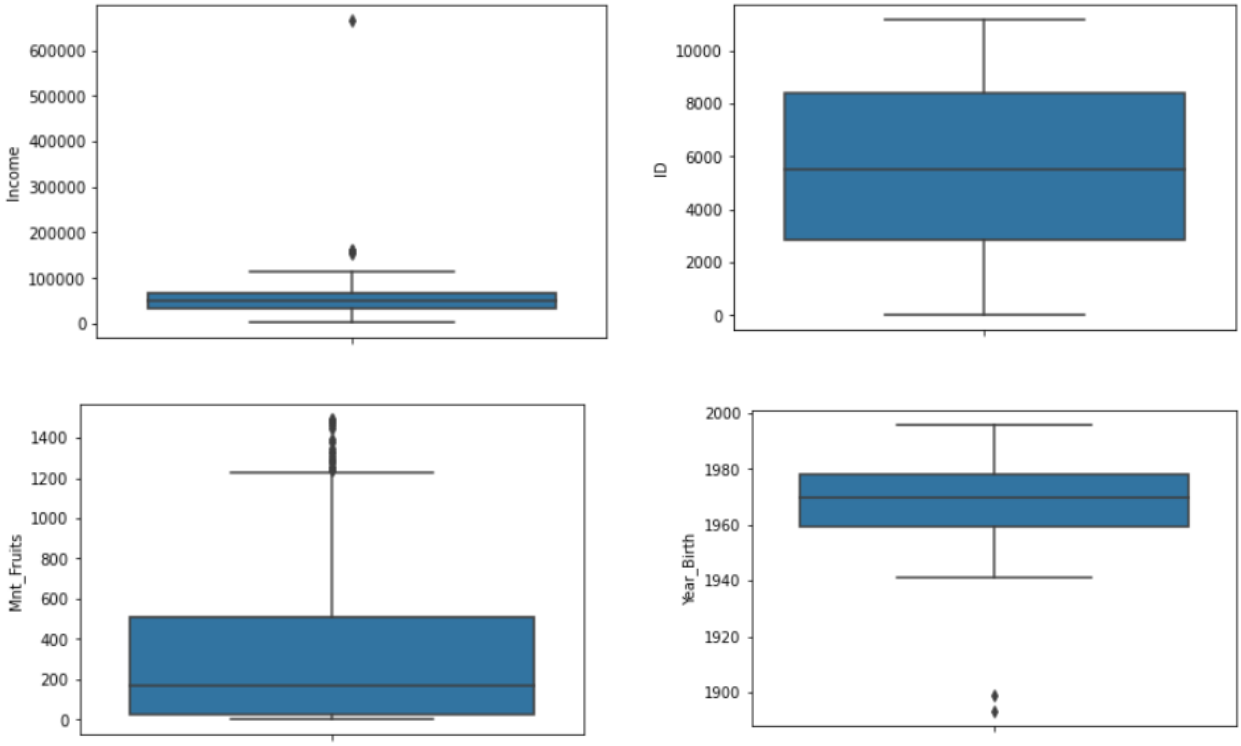








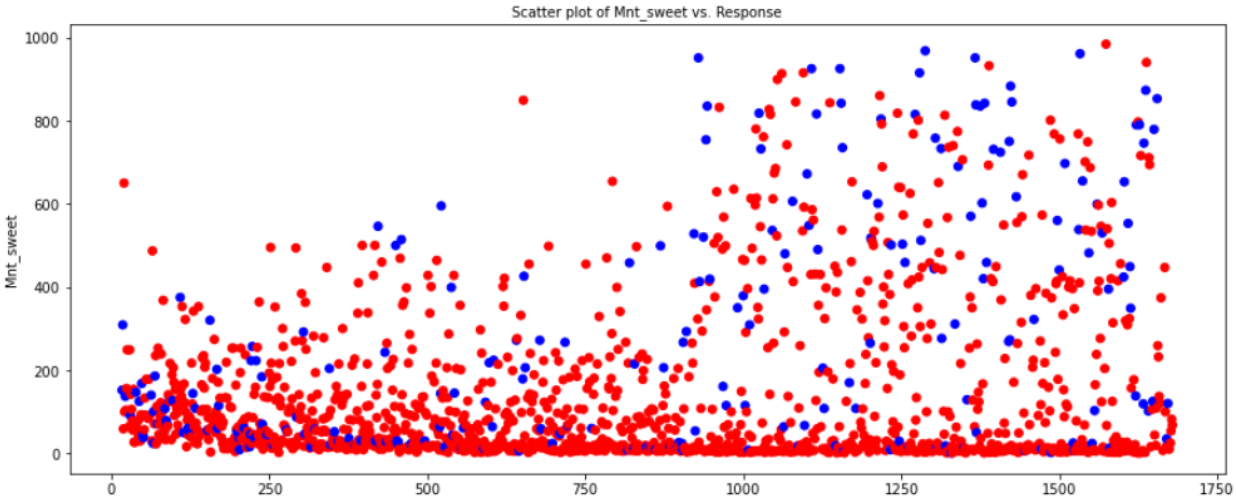
ייצוג נוסף הוא בעזרת Box plot:



כפי שלמדנו, על ידי ייצוג זה ניתן לראות את אחוזוני ה- 25,50,75 כמו כן, טווח ערכי ה- attribute המוצג.
חשוב לשים לב לערכים אשר גדולים פי 1.5 מערכו של ה- IQR, הפרש האחזונים 75 ו- 25.
למשל, ב- Income יש ערך המתקבל בקצה הסכמה שערכו מעל ל- 1.5 x IQR, דבר המעיד על outlier וכמו כן, מסתדר עם התובנות מגרפים קודמים.
בדומה, גם ל- Year_Birth ישנם שני ערכי קיצון מוטים מה- IQR.
עבור Mnt_Fruits, ישנה קבוצת ערכים מוטת דבר היכול להעיד על Collective outliers.

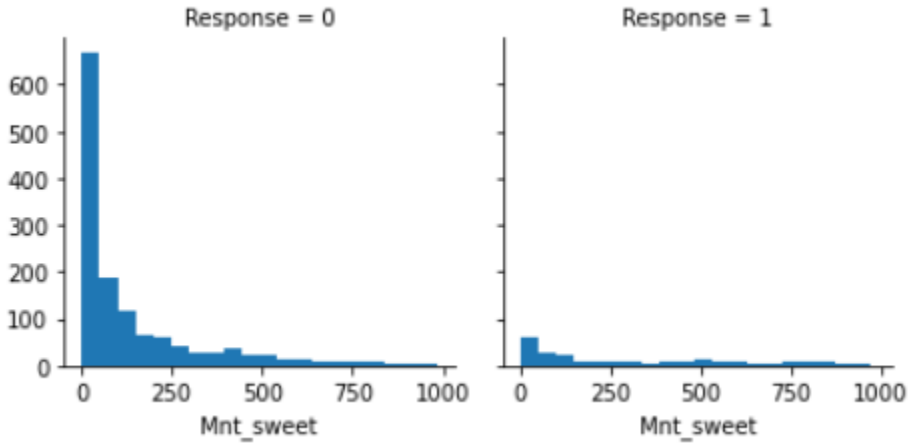
ייצוג נוסף הוא בעזרת scatter plot:

מוצגת ויזואליזציית Mnt_sweet ו- Response.



מבדיקה, קורלציית שני מאפיינים אלו היא גבוהה יחסית קשה לראות זאת מגרף ה- scatter plot.
ניתן לראות כי רוב האנשים אשר סירבו להצעת הקמפיין ערך ה- Mnt_sweet שלהם הוא יחסית נמוך, ניתן לראות בנוסף כי רוב המדגם סירב להצעת הקמפיין.

כמו כן, אפשר להציג זאת כך:



יתרון לשיטה זו הוא הפרדת ערך ה- Response והוא עובד טוב עבור מאפיינים בעלי מספר מופעים סופי (קטגוריאליים, בינאריים וכו'...)

הטיה (skewness)

לאחר תיאור הגרפים בהיסטורמות, נצפה כי ה- attributes אשר יתפלגו בצורה אחידה/נורמלית/אקספוננציאלית יהיו בקירוב בעלי הטיה זהה ובפרט אלו המתפלגים באופן אחיד/גאוסיאני יהיו בעלי הטיה נמוכה שכן עבור התפלגויות אלו ה- skewness הוא 0.

```
skewness of normal distributed attributes
Year_Birth  skewness: -0.3521297334201921
Income  skewness: 7.916831486203255

skewness of exponenital distributed attributes
Mnt_Fruits  skewness: 1.1829109635956956
Mnt_Meat  skewness: 2.1495184429670813
Mnt_sweet  skewness: 2.1092143754162325
Mnt_Wines  skewness: 1.988342728206097
Mnt_Gold_Products  skewness: 2.218833414968302
Mnt_Fish  skewness: 1.9402486173238092

skewness of uniform distributed attributes
ID  skewness: 0.02751801537005124
Recency  skewness: -0.012307768420328446

skewness of other attributes
Num_of_kids  skewness: 0.5841313250351708
Num_of_Teen  skewness: 0.4129909449000085
Num_Web_Purchases  skewness: 2.5125546694750898
Num_Store_Purchases  skewness: 1.5552728492914274
Num_Deals_Purchases  skewness: 2.0657617307381693
Num_Catalog_Purchases  skewness: 0.6997937661170338
Num_Web_Visits  skewness: 0.25966407514401624
Response_Campaign_1  skewness: 3.191823311903448
Response_Campaign_2  skewness: 3.3224579772122715
Response_Campaign_3  skewness: 3.305146751217529
Response_Campaign_4  skewness: 3.5271264290087347
Response_Campaign_5  skewness: 8.359327109203887
Complain  skewness: 10.427719201596041
Cost_Contact  skewness: 0
Revenue  skewness: 0
```

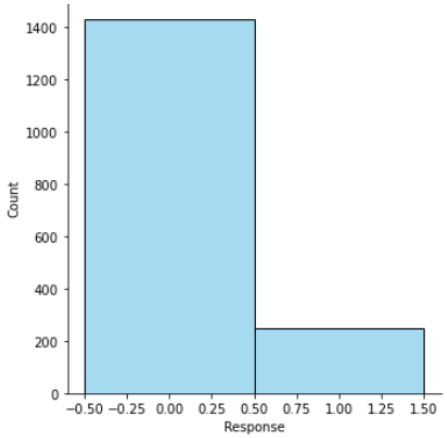
כפי שניתן לראות, אכן הטיית ה- attributes המתפלגים אחיד היא בקירוב 0.
כמו כן, נשים לב כי הטיית מאפיין Income גבוהה יחסית לאופן התפלגותו (גאוסיאני) דבר אשר יכול להעיד על outliers וערכים שגויים. הדבר הגיוני שכן, לא כל אדם מעוניין לחשוף את משכורתו האמיתית ואף אולי יעדיף למלא מידע שגוי.
נשים לב לעוד דבר מעניין, בעוד ה- skewness עבור תכונות Response_Campaign_1-4 הוא בקירוב זהה סביב הערך 3.2, ההטיה עבור Response_Campaign_5 היא 8.35. כלומר, בעוד הקמפיינים הקודמים היו מאוזנים יחסית והניבו תוצאות ללא הטיה רבה בין אחד לשני, הקמפיין החמישי מוטה.
בנוסף, תוצאות ההיסטוגרמות מעידות כי אכן כמות האנשים אשר נענו בחיוב לקמפיין החמישי היא קטנה ביחס לשאר הקמפיינים. דבר אשר הגיוני להטייה היחסית בין קמפיין זה לשאר, נסיק כי הטיה זו היא אינה טובה ומעידה על קמפיין שיווקי לא מוצלח.
נוסף על כך, נראה כי הטיית **כל** ה- attributes המתפלגים אקספוננציאלית היא בקירוב 2.

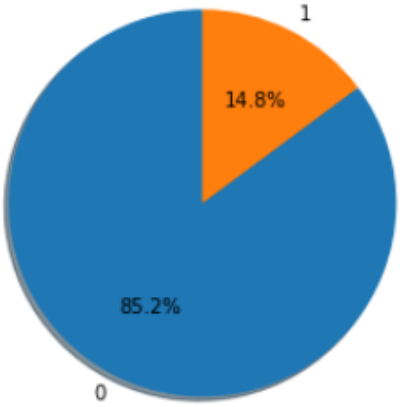
שכיח (mode)

ניתן לראות את ערך השכיח עבור כל attribute בדאטאסט לאחר מילוי הערכים החסרים.

```
Year_Birth                          1976.0
Education                                2
Status                             Married
Income                              7500.0
Num_of_kids                            0.0
Num_of_Teen                            0.0
Registration_date        2013-02-14 00:00:00
Recency                                 56
Mnt_Fruits                             2.0
Mnt_Meat                               0.0
Mnt_sweet                              7.0
Mnt_Wines                              0.0
Mnt_Gold_Products                      0.0
Mnt_Fish                               4.0
Num_Web_Purchases                      1.0
Num_Store_Purchases                    2.0
Num_Deals_Purchases                    0.0
Num_Catalog_Purchases                  3.0
Num_Web_Visits                        14.0
Response_Campaign_1                    0.0
Response_Campaign_2                    0.0
Response_Campaign_3                    0.0
Response_Campaign_4                    0.0
Response_Campaign_5                    0.0
Complain                               0.0
Response                                 0
Status_cat                               5
Mnt_all                               22.0
Name: 0, dtype: object
```

ערך המטרה (Target)





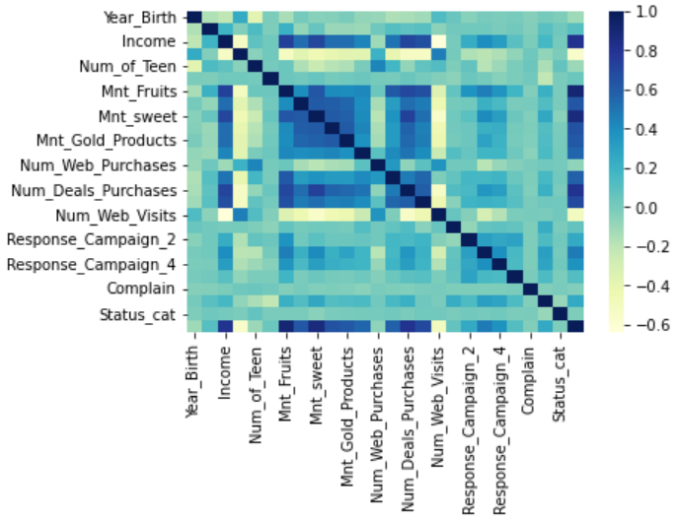
כלומר, הרוב המוחלט של האנשים סירבו להצעת הקמפיין.

Attributes correlations

תחילה, כדי לחשב קורלציות בין מאפיינים היינו צריכים למלא את הערכים החסרים של המאפיינים בערך 0.
וזאת מכיוון שאין הרבה ערכים חסרים ובלאו הכי מילוי שלהם לא ישנה משמעותית את הקורלציות.
את הקורלציות חישבנו לפי מקדם הקורלציה של פירסון, הערכים המתקבלים הם בין 1- ל- 1.
כאשר:

- 1 אומר קורלציה חזקה חיובית, כלומר שני ה attributes עולים ויורדים ביחד.
- 1- אומר קורלציה חזקה שלילית, כלומר attribute אחד עולה כאשר השני יורד ולהפך.

ויזואליזציה של תוצאות הקורלציה:



כמו כן, גם חלק מ- טבלת הקורלציות:

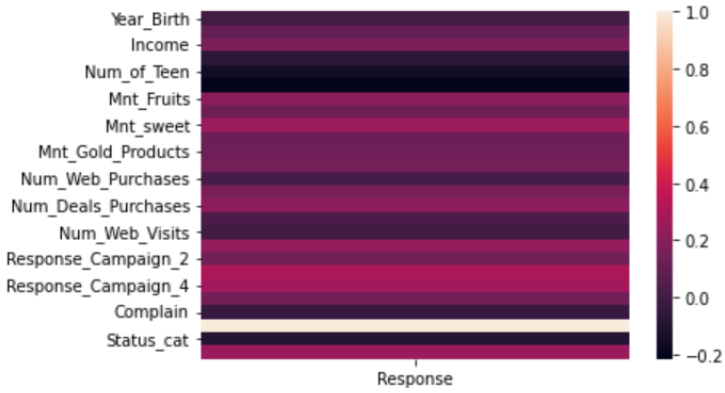
0.078182	0.388410	0.490626	0.371903	0.217967	-0.026639	0.057462	0.057462	0.242603
0.013086	0.013661	0.205846	0.178858	-0.013865	0.001016	0.041989	0.041989	0.119676
0.019370	0.080191	0.373732	0.294261	0.024647	-0.012561	0.046809	0.046809	0.237879
0.003501	0.014160	0.199063	0.240632	-0.000920	-0.015511	0.043818	0.043818	0.109113
0.005263	0.026438	0.272295	0.220876	0.001744	-0.018498	0.041333	0.041333	0.120718
0.123391	0.038210	0.186099	0.176047	0.035119	-0.048622	0.053346	0.053346	0.144707
-0.009770	0.024087	-0.165846	-0.098628	-0.044873	0.000656	0.075751	0.075751	0.011341
0.049795	0.145304	0.147900	0.161158	0.026958	-0.020499	0.092352	0.092352	0.145196
0.110203	0.140266	0.329833	0.297913	0.102269	-0.005425	0.056528	0.056528	0.206802
-0.045532	0.184484	0.198202	0.189784	0.092708	0.006472	0.114098	0.114098	0.039618
0.046725	-0.025100	-0.260752	-0.165507	-0.015011	-0.017961	0.134853	0.134853	-0.010593
1.000000	-0.079313	0.077107	0.089223	0.063196	-0.003206	0.018498	0.018498	0.241937
-0.079313	1.000000	0.325035	0.264743	0.305451	-0.026325	0.017940	0.017940	0.157510
0.077107	0.325035	1.000000	0.393989	0.204943	-0.001967	0.018021	0.018021	0.330907
0.089223	0.264743	0.393989	1.000000	0.176928	-0.025001	0.017038	0.017038	0.277794
0.063196	0.305451	0.204943	0.176928	1.000000	-0.011183	0.007621	0.007621	0.152688
-0.003206	-0.026325	-0.001967	-0.025001	-0.011183	1.000000	0.006140	0.006140	-0.021784
0.018498	0.017940	0.018021	0.017038	0.007621	0.006140	1.000000	1.000000	0.026982
0.018498	0.017940	0.018021	0.017038	0.007621	0.006140	1.000000	1.000000	0.026982
0.241937	0.157510	0.330907	0.277794	0.152688	-0.021784	0.026982	0.026982	1.000000

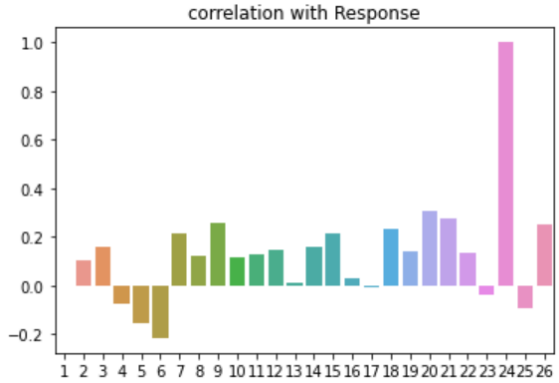
כאשר, אנו נתרכז בחלק המסומן באדום בטבלה.

מתוך גרף ויזואליזציית הקורלציה והטבלה ניתן לראות:

ישנה קורלציה גבוהה בעלת ערך 1 בין שני attributes מסוימים, ולכן ניתן יהיה להוריד את אחד מהם.
כמו כן, מתוך ויזואליזציית הקורלציה אפשר לראות כי קיימים **רק** שני משתנים אשר ביניהם קורלציה 1, למעט האלכסון המייצג קורלציה עצמית, בעלת ערך 1.
מאפיינים אלו הינם: Revenue ו- Cost_Contact (בהמשך, נסביר כי מחקנו אותם כי הם קבועים).

כעת, מכיוון שערך המטרה הוא Response, נרצה לראות את הקורלציות שלו עם משתנים אחרים:





בנוסף, כדי לראות בצורה נוחה יותר עם מי יש ל- Response קורלציה חזקה (גבוהה חיובית, או נמוכה שלילית) נציג את 5 מקדמי הקורלציה הגבוהים והנמוכים ביותר:

```
Response                1.000000
Response_Campaign_3     0.330907
Response_Campaign_4     0.277794
Mnt_Fruits              0.242603
Response_Campaign_1     0.241937
Name: Response, dtype: float64
Recency        -0.211211
Num_of_Teen    -0.141540
Num_of_kids    -0.073626
ID             -0.033697
Complain       -0.021784
Name: Response, dtype: float64
```

ניתן לראות כי לתגובה בקמפיינים הקודמים (הקמפיין השלישי, רביעי וראשון) יש קורלציה חיובית, יחסית גבוהה עם Response.
יתר על כן, כמות הכסף אשר הוצא על פירות גם בעל קורלציה חיובית גבוהה יחסית עם Response.
לעומת זאת, ל Num_of_Teen, Recency יש קורלציה שלילית יחסית גדולה עם Response.

ערכים חסרים (Missing Value)

ישנן רשומות בעלות ערכי attribute חסרים, כדי לנתח את הדאטא בצורה טובה נצטרך להשלים את אותם ערכים.
נעשה זאת בצורה פרטנית עבור כל attribute שחסר ברשומות.

על ידי בדיקה של כמות השדות החסרים ברשומה, ניתן לראות כי ישנן 7 רשומות ב- dataset אשר חסר להן 19 ערכים(!), כלומר, יותר ממחצית מהערכים הם NaN רשומות אלו אינן אינפורמטיביות עבורינו ואף, יכולות לגרום לשגיאות ולירידת ביצועים של המודל.

```
df[df.isnull().sum(axis=1)==19]
```

	ID	Year_Birth	Education	Status	Income	Num_of_kids	Num_of_Teen	Registration_date	Recency	Mnt_Fruits	...	Num_Web_Visits	Response_Can
1659	1419	1950.0	Graduation	Together	34026.0	1.0	1.0	05/08/2013	11	NaN	...	NaN	
1662	9284	1958.0	Graduation	Together	53977.0	0.0	1.0	08/06/2013	21	NaN	...	NaN	
1663	3673	1971.0	Graduation	Single	55239.0	0.0	1.0	14/07/2013	59	NaN	...	NaN	
1665	10983	1952.0	Graduation	Together	75278.0	0.0	0.0	29/01/2013	17	NaN	...	NaN	
1666	2611	1959.0	Master	Together	82576.0	0.0	0.0	01/08/2012	66	NaN	...	NaN	
1673	979	1975.0	Graduation	Single	33249.0	1.0	0.0	20/02/2013	11	NaN	...	NaN	
1675	8278	1990.0	PhD	Married	74214.0	0.0	0.0	26/08/2012	3	NaN	...	NaN	

7 rows × 29 columns

כמו כן, ישנן 8 רשומות כאשר כל אחת מהן בעלת 6 ערכים חסרים, (כמות ערכים לא מעטה) בנוסף, לרשומות אלו אין משהו משותף בתור קבוצה לכן, נסיק כי גם רשומות אלו לא נחוצות לנו.

סך הכל, החלטנו להסיר מה- dataset רשומות בעלות 6 חוסרים או יותר.
(מבדיקה, ישנן רשומות רק עם 1/2/6/19 ערכים חסרים).

```
redundant_rows = df[df.isnull().sum(axis=1)>=6].index
print(redundant_rows,'\n')

df = df.drop(redundant_rows)  # update the dataframe
df.reset_index(drop=True, inplace=True)

df.info()
```

לאחר הורדת שורות אלו, נשארו רק רשומות עם אחד או שניים ערכים חסרים.

נחשב את כמות הערכים החסרים ב- dataset.

```
155 rows has  1  missing values
7 rows has  2  missing values
0 rows has  3  missing values
0 rows has  4  missing values
0 rows has  5  missing values
0 rows has  6  missing values
0 rows has more than  7  missing values
169
```

ישנם:

$$\#Missing\ values\ =\ 155\ *\ 1\ +\ 7\ *\ 2\ =\ 169$$

ערכים חסרים ב- database.

כעת, נרצה להשלים את חוסרים אלו על ידי טיפול פרטני עבור כל attribute, נבדוק עבור כל מאפיין את כמות הערכים אשר חסרים לו.

ישנם 9 מאפיינים בעלי ערכי NaN (צמצום משמעותי מאוד ביחס ללפני הסרת הרשומות):

```
Attribute  Year_Birth has  21  missing values
Attribute  Status has  26  missing values
Attribute  Income has  63  missing values
Attribute  Num_of_Teen has  12  missing values
Attribute  Mnt_sweet has  14  missing values
Attribute  Num_Web_Purchases has  22  missing values
Attribute  Response_Campaign_1 has  11  missing values
169
```

כמו כן, נבצע "בדיקת שפיות" (sanity check) שאכן כמות ערכי שדות ה- NaN הינו 169.
נוכל לחשב את כמות הערכים החסרים ב- database, נקבל:

$$\#Missing\ values\ =\ 21\ +\ 26\ +\ 63\ +\ 12\ +\ 14\ +\ 22\ +\ 11\ =\ 169$$

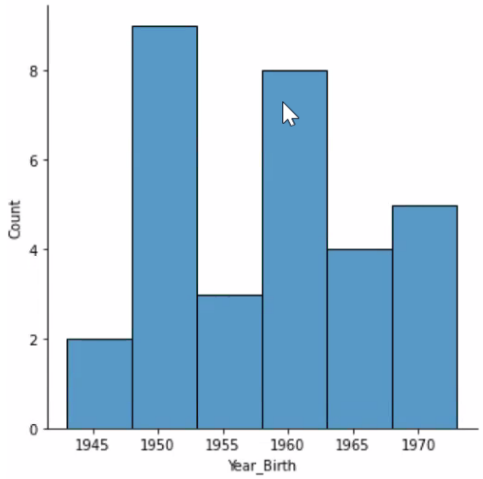
ואכן, התוצאה תואמת את המצופה.

- המאפיין Year_Birth

הפרדנו את הרשומות ב- dataframe בהן שדה זה הוא NaN ניתן לשים לב כי המשותף לרשומות אלו הוא שדה ה- Status אשר ברובן הערך הוא Widow (אלמן).

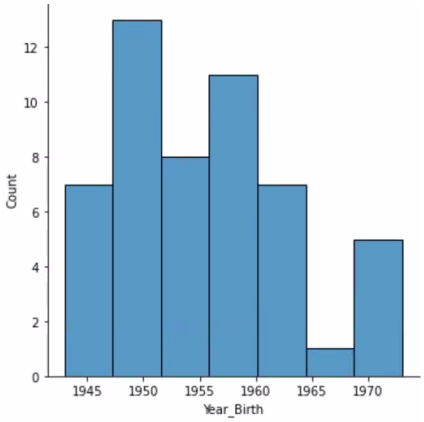
A	B	C	D	E
	ID	Year_Birth	Education	Status
204	9204		2	Widow
211	702		1	Widow
335	9336		4	Widow
346	6379		3	Widow
465	6878		2	Widow
484	8842		2	Widow
502	5985			
509	9699			
630	8594		4	Widow
634	2587			
777	10664		3	Widow
838	10591		2	Widow
840	4945		2	Widow
933	3921		1	Widow
936	1544			
945	8629		2	Widow
1144	8650		2	Widow
1151	2431			
1153	10955		2	Widow
1235	5084		2	Widow
1236	4587		3	Widow
1358	7851		2	Widow
1439	9213		2	Widow
1464	10451			
1502	6437			
1522	6609		2	Widow
1564	10102		2	Widow
1568	9058		2	Widow
1639	7627			

נוכל לשערך את ערך ה- Year_Birth עבור Widow.
נציג בהיסטוגרמה את התפלגות שנות הלידה של אלמנים:



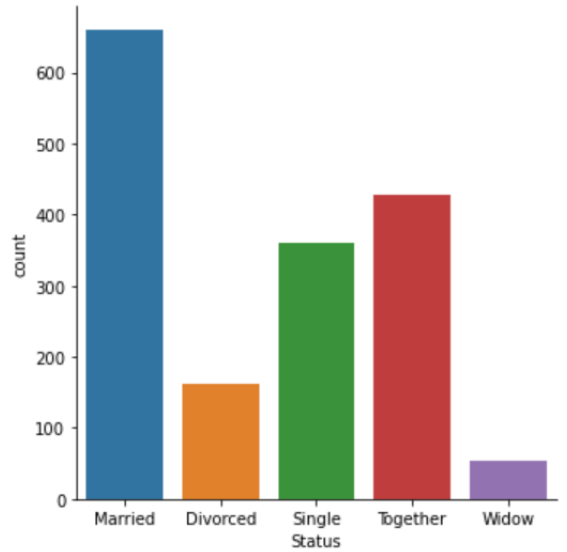
כדי לא לפגום בשונות המדגם לא נרצה לבחור ערך קבוע עבור כל הרשומות (למשל, החציון או השכיח)
כמו כן, ישנן 29 רשומות כאלו, לא מעט.
ולכן, נפלג בצורה אחידה את גילאי ה- Year_Birth בטווח השנים (1945,1970), טווח השנים מ- ההיסטוגרמה.

הפילוג לאחר השינוי כלומר, פילוג Year_Birth של רשומות בעלי סטטוס Widow
(לאחר השלמת שדה Year_Birth):



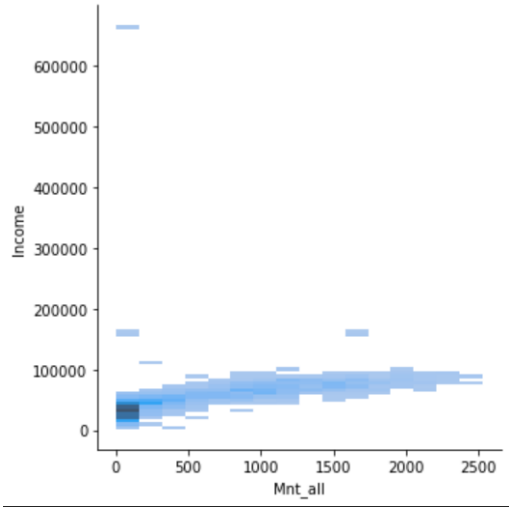
- המאפיין Status

תחילה, החלטנו להוריד את הערך "Alone" מכיוון שרק שני אנשים סימנו את ערך זה, ולכן הוא outlier.
מסטטיסטיקת ה- attribute ניתן לראות כי אין התפלגות גורפת.



למרות זאת, החלטנו לשמור על הפילוג המקורי ולמלא את הרשומות על ידי הגרלה מ- ההתפלגות המקורית.
וזאת מכיוון שאמנם ההתפלגות היא לא מוכרת אך, אין ערכים גבוהים בהרבה מערכים אחרים כלומר הפילוג הוא לא קיצוני ובחרנו להסתמך עליו.

- המאפיין Income



ראשית כל, מצאנו קורלציה של Income עם attributes אחרים, עשינו זאת כדי לנסות לשערך אותו לפיהם.
מצאנו כי ה- attributes עם Mnt (כמה כסף הוציאו על דברים בשנתיים האחרונות) הם בעלי קורלציה גבוהה עם Income דבר הגיוני שכן, ככל שה- Mnt גבוה כך הוציא האדם יותר כסף אשר הוא יכול להרשות לעצמו לפי משכורתו. קיבלנו קורלציה מקסימלית של Income עם attribute Mnt_all (מאפיין אשר הוספנו).

כדי לטפל ברשומות בעלות ערך Income חסר, מצאנו את כל הרשומות להן ערך Mnt_all הרחוק ב- לכל היותר 5 מהערך של Mnt_all עבור הרשומה עם ערך ה- Income החסר.
את ערך ה Income השלמנו לפי ממוצע ערכי הדגימות הללו.
אם לא היו דגימות במרחק של לכל היותר 5, חיפשנו דגימות במרחק של לכל היותר 15, ואם גם לא היו דגימות כאלו, השלמנו את ערך ה Income לפי ממוצע ערכי ה Income של כל ה df (שהם לא NaN).

- Mnt_sweet

ראשית, מצאנו את הקורלציה של משתנה זה עם משתנים אחרים. הקורלציה החזקה ביותר הייתה עם Mnt_all, אבל, ערכי ה- NaN ב- Mnt_sweet גוררים ערכי NaN גם ב- Mnt_all ולכן, לא יכולנו להשתמש ב Mnt_all למילוי החוסרים.
במקום זאת, השתמשנו ב Num_Deals_Purchases, לו גם הייתה קורלציה חזקה עם Mnt_sweet.
כעת, השלמנו את הערכים החסרים בצורה דומה ל- Income, למעט המרחקים.

- Num_of_Teen

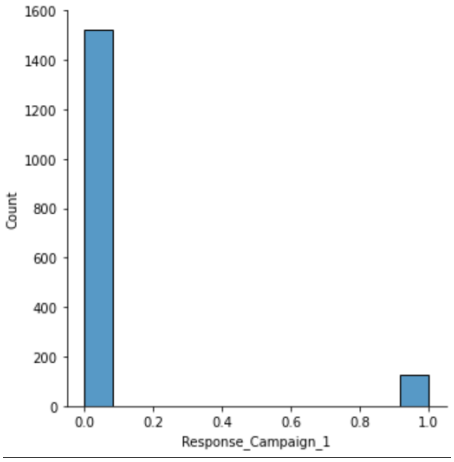
תחילה, מצאנו את הקורלציה של משתנה זה עם משתנים אחרים. החזקה ביותר הייתה עם Num_Web_Purchases. ולכן, השתמשנו בו למילוי החוסרים.
גם כאן מילוי הערכים החסרים התבצע בצורה דומה לזו של Income, למעט המרחקים.

- Num_Web_Purchases

באותו אופן כמו מקודם למעט כך ש- הקורלציה החזקה ביותר מתקבלת עם המאפיין Num_of_Teen.

- Response_Campaign_1

עבור attribute זה, ראינו כי רוב הערכים שלו הם 0:



לכן, מילאנו ערכים חסרים של מאפיין זה ב- 0 (שהוא גם השכיח).

Outliers

בחלק זה, מטרתנו היא מציאת outliers בדאטא באמצעות שיטות סטטיסטיות.
נזכור כי סוגי ה- outliers הינם: Global/Contextual/Collective outliers לאחר שנזהה את סוג ה- outlier נבצע טיפול פרטני בהתאם.

זיהוי ה- outliers התבצע בכמה דרכים שונות:

- ניתוח ערך סטיית התקן, std ומרחקו מהממוצע.
בהרצאה ראינו כלל אצבע של 3 סטיות תקן מהממוצע לזיהוי outliers בפועל לוא דווקא השתמשנו בערך זה וזאת מכיוון שהיו מקרים בהם ירדו יותר/פחות מידי דגימות או ערכי קיצון גבוליים לא נכנסו עבור ערך זה.
- בדיקת חוסר תיאום של הדאטא.
לדוגמה, ציפייה של ערכי מאפיין לקבלת ערכים של מספרים שלמים ואי שליליים (כמו מספר ילדים, ביקורים באתר וכו'...) ובפועל קבלת ערך מספר רציונלי שלילי.
בנוסף, גם בערכי מאפיינים בינאריים, נצפה לקבל **רק** 0 או 1.

כמו כן, מצאנו גם חוסר עקביות (inconsistent) על חלק מהמידע (חלקים המתנגשים עם חלקים אחרים), נרחיב על כך בהמשך.

כעת, ננתח outliers שונים עבור כל מאפיין ונתעד את המאפיינים המרכזיים.

- המאפיין Year_Birth, על ידי חיפוש דגימות שונות מהממוצע ב- std*3 או יותר. מצאנו 2 outliers והם:

	ID	Year_Birth	Education	Status	Income	Num_of_kids	Num_of_Teen	Registration_date	Recency	Mnt_Fruits	...	Response_Campaign_3
1184	1150	1899.0	4	Together	83532.0	0.0	0.0	2013-09-26	36	755.0	...	1.0
1217	11004	1893.0	1	Single	60182.0	0.0	1.0	2014-05-17	23	8.0	...	0.0

כלומר, מדובר בשני אנשים, האחד נולד בשנת 1899, והשני בשנת 1893
מכיוון שמדובר רק ב- שתי דגימות, והנחנו שלא סביר כי אדם בגיל 100+ שייך לדאטא, החלטנו למחוק את דגימות אלו.

- המאפיין Education, חיפוש ה- outliers התבצע על ידי בדיקת ערכי המאפיין, הקשר ביניהם ו- מציאת הגיון ביניהם.
התחלנו מ-שינוי הערכים של Education לערכים מספריים.

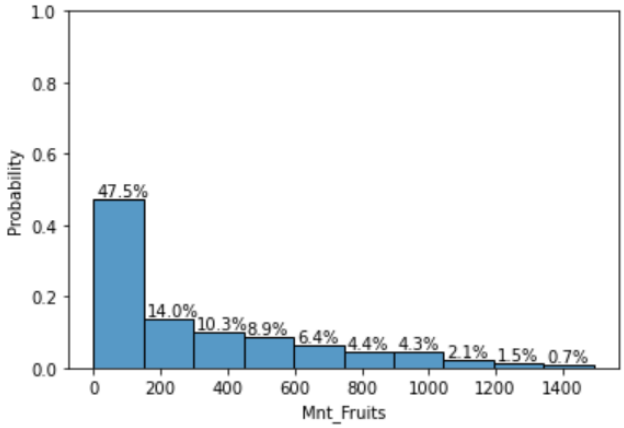
```
df["Education"].unique()
# no outliers

array([2, 1, 4, 3, 0], dtype=int64)
```

על ידי בדיקת כמות המופעים של כל ערך יכולנו להסיק אם מדובר ב- outlier בפרט, כמות מופעים קטנה של ערך מעידה על כך.

השתמשנו בשיטה זו עבור כל ה- attributes שמקבלים ערכים בדידים בטווח ידוע (קטגורי).

בדומה ל- Education, המאפיין Status המקבל ערכים קטגוריאליים בעל ערך Alone המופיע **רק** בשתי רשומות ב- database.
ערך סטטוס זה הוא לא הגיוני שכן הוא לא מאפיין סטטוס של בן אדם, במילים אחרות רשומות אלו הן outlier והחלטנו להסיר אותן.

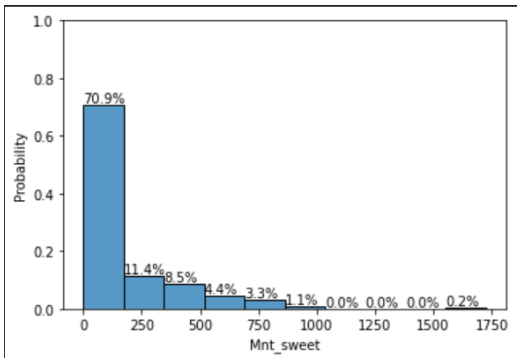
- המאפיין Mnt_Fruits, חיפוש ה- outliers התבצע על ידי הסתכלות על גרף ההיסטוגרמה:



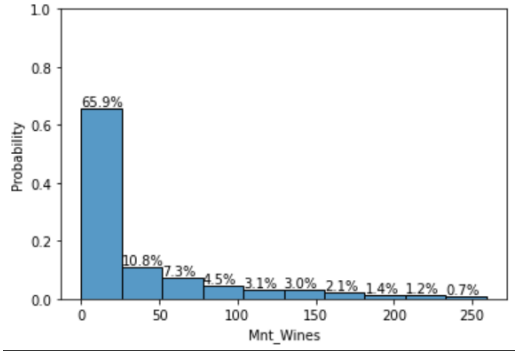
בדומה להרצאה, bins אשר ערכם הוא מתחת ל- 1% מעידים על outliers.
ערכו של bin זה מתחיל ב 1343.7. לכן, נסיר כל דגימה בעלת ערך גדול מערך זה (ערך לא שלם ולכן, ניתן להוריד כל מה שגדול מזה ולא גדול-שווה). ישנן 11 רשומות כאלו, ואותן אנו מחשיבים כ outliers.

באותו אופן, ביצענו אותה השיטה עבור המאפיינים הבאים:

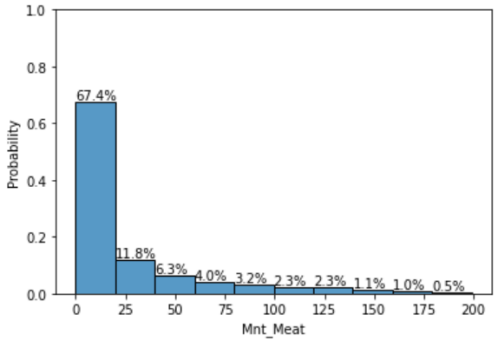
1. Mnt_sweet



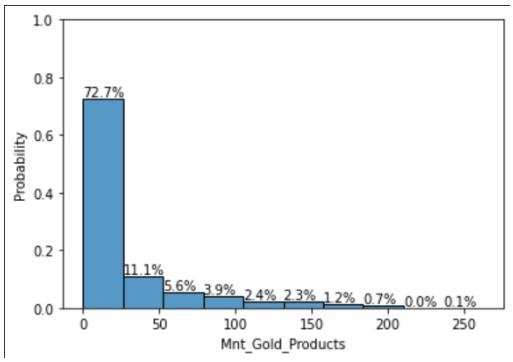
2. Mnt_Wines



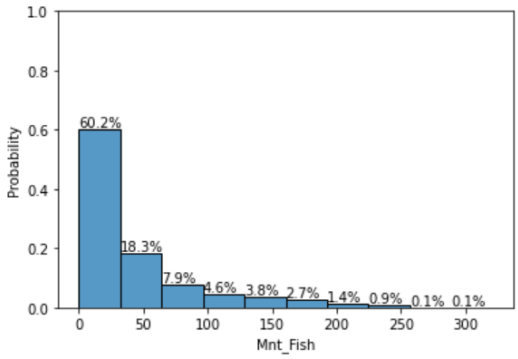
3. Mnt_Meat



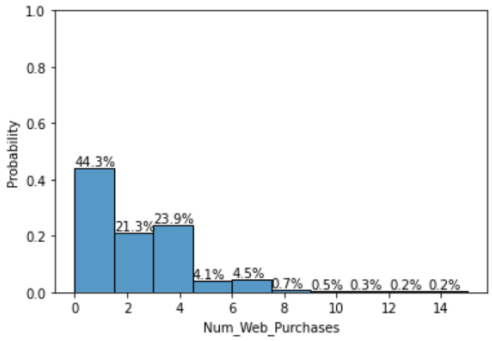
4. Mnt_Gold_Products



5. Mnt_Fish

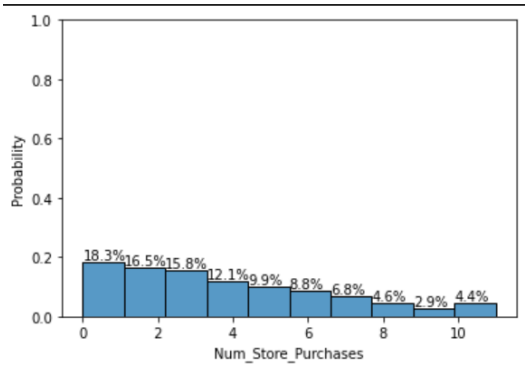


6. Num_Web_Purchases

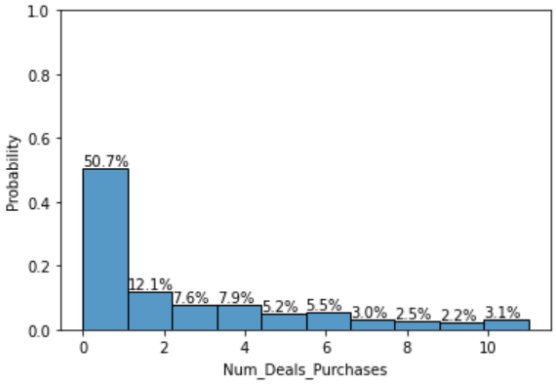


ישנם מאפיינים עבורם תוצאת ההיסטוגרמה לא העידה על outliers.
לדוגמה,

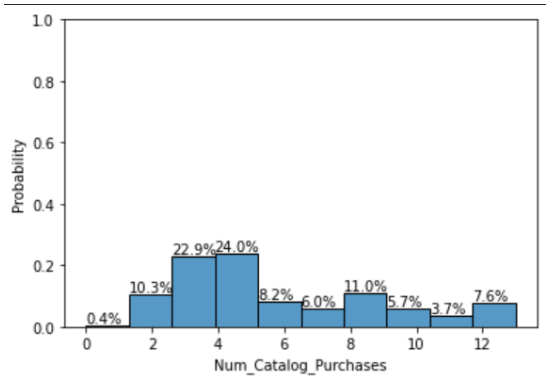
1. Num_Store_Purchases



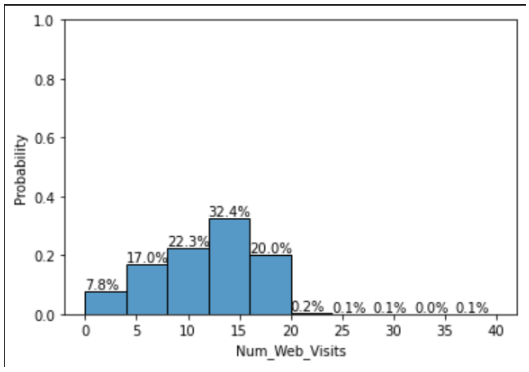
2. Num_Deals_Purchases



3. Num_Catalog_Purchases



4. Num_Web_Visits



קיימים attributes נוספים עבורם טיפלנו ובדקנו קיום של outliers.

כעת, נותר לטפל בחוסר עקביות בדאטא (inconsistent data).

נשים לב לשני מאפיינים מעניינים: Num_Web_Visits ו- Num_Web_Purchases.
קיימת תלות ביניהם, הרי לא ניתן לקנות באתר האינטרנט ללא כניסה לתוך האתר ולכן, רשומות בהן ערך כמות הקניות החודשיות באתר (Num_Web_Purchases) גדול מ- כמות כניסות חודשיות לאתר (Num_Web_Visits) הן רשומות לא עקביות המהוות טעות.
מאחר ויש רק 29 דגימות כאלו, וקיימת בהן טעות, זה לא משהו לא סביר סטטיסטית, אלא טעות.
החלטנו למחוק את כל הרשומות הללו.

<u>מחיקה והוספת מאפיינים</u>

לאחר התבוננות ב- database במטרה למצוא מאפיינים בעלי ערכים קבועים ותלויות בין מאפיינים.
שמנו לב כי המאפיינים: Cost_Contact, Revenue קבועים לכל הדגימות, ולכן לא תורמים כלל.
דבר היוצר יתירות בדאטא שאינה הכרחית לכן, החלטנו להסיר מאפיינים אלו.

כמו כן, החלטנו להוריד את attribute ה- ID מכיוון שזהו מספר סידורי שניתן לכל אדם, ללא כל תלות בשאר ה- attributes.

כדי לנתח את הדאטא ופילוגו בצורה טובה יותר, החלטנו להוסיף את ה- attributes הבאים:

Status_cat - העברת ערכי הסטטוס מערך קטגוריאלי לערך מספרי.

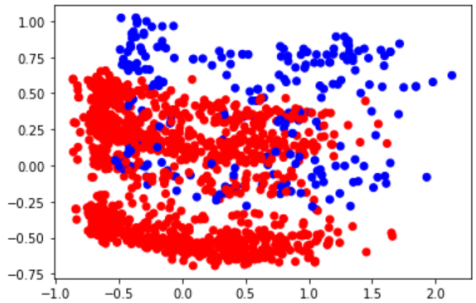
Mnt_all - שילוב מאפייני ה- Mnt על ידי סכומם וכתוצאה מכך, קבלת קורלציה גבוהה וחזקה יותר עם שאר המאפיינים (למשל Income), דבר אשר הועיל בתהליך מילוי הערכים החסרים בדאטא.

Data reduction

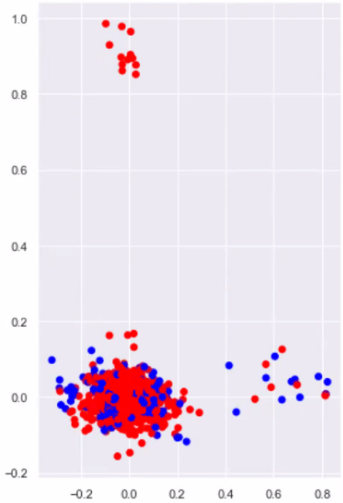
לשם ביצוע data reduction, השתמשנו ב PCA.

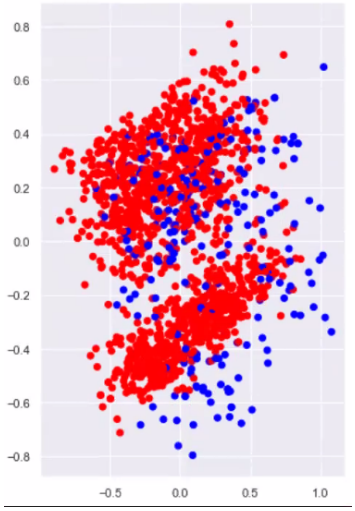
מטרת ה- PCA הינה ביצוע מזעור ממדים תוך שימור על שונות מקסימלית.
כלומר, ביצוע מזעור ממדים תוך שמירה על חיזוי נכון של ערך המטרה, Response.
השימור על שונות מקסימלית, גורם לכך שיש יותר מידע בנתונים אותם אנו שומרים.

אם ניקח את שני הוקטורים העצמיים אשר ממקסמים את השונות.



אם ניקח וקטורים עצמיים אחרים נקבל שונות גדולה מאוד ודאטא קשה להפרדה:





Normalization

על ידי נרמול הדאטא נגביל את ערכי המקסימום והמינימום וכך נמנע "התפוצצות" או "העילמות" לערך גבוה או נמוך. כמו כן, הנרמול גורם לכך שהערכים יהיו בסקאלה זהה.
אנחנו בחרנו להשתמש בנרמול Min-max, כך הערכים של כל attribute יהיו בין 0 ל-1.

	Year_Birth	Education	Income	Num_of_kids	Num_of_Teen	Registration_date	Recency	Mnt_Fruits	Mnt_Meat	Mnt_sweet
count	1576.000000	1576.000000	1576.000000	1576.000000	1576.000000	1576.000000	1576.000000	1576.000000	1576.000000	1576.000000
mean	0.545213	0.616751	0.297956	0.231282	0.252221	0.522095	0.498936	0.220071	0.134375	0.158745
std	0.222813	0.252471	0.130171	0.270740	0.270792	0.216982	0.294165	0.247250	0.205426	0.215401
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.372549	0.500000	0.194094	0.000000	0.000000	0.366290	0.252525	0.017267	0.005618	0.014242
50%	0.568627	0.500000	0.289484	0.000000	0.000000	0.518832	0.505051	0.119745	0.039326	0.060529
75%	0.705882	0.750000	0.396386	0.500000	0.500000	0.680791	0.757576	0.371809	0.157303	0.217701
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Discretization

דיסקרטיזציה היא חלוקה של המידע לערכים בדידים.
נבצע דיסקרטיזציה למאפיינים אשר ניתן לדרג אותם ולהפריד אותם לקבוצות ערכים מסוימות.
כגון: Income, Year_Birth, ה- attributes של ה Mnt.

לדוגמה, עבור Income, חילקנו ל-5 קבוצות,

3.0	667
2.0	410
1.0	250
0.0	169
4.0	80

כאשר,

$$Income \leq 25,000 \rightarrow 0$$
$$25,000 < Income \leq 35,281 \rightarrow 1$$
$$35,281 < Income \leq 51,170 \rightarrow 2$$
$$51,170 < Income \leq 82949 \rightarrow 3$$
$$Income > 82949 \rightarrow 4$$

וגבולות כל קבוצה, נבחרו בעזרת היסטוגרמה.

שמנו לב כי ב- ביצוע דיסקרטיזציה ונורמליזציה ואחר כך הפעלת אלגוריתם PCA, קיבלנו תוצאות שהן פחות טובות (בהשוואה לאי ביצוע דיסקרטיזציה).
התוצאה הסופית אינה טובה יותר ומניבה דאטא אשר לא ניתן להפרדה בקלות ואף שהשונות בחלק מהרכיבים גדולה יותר. ולכן, החלטנו שלא להשתמש בתוצאות של דיסקרטיזציה.

תוצאת ה- PCA עם דיסקרטיזציה: