בינה עסקית – מטלה 1

מגישים : סתיו אטיאס 318641024

עידו שטרן 315471920

ניב בן אברהם 313581373

1. **איסוף נתונים**

בחרנו במאגר נתונים מאתר KAGGEL, העוסק באיסוף נתונים של מכירות חנות.
הנתונים מחולקים לפי איזה סוג מוצר קנה כל לקוח.
https://www.kaggle.com/imakash3011/customer-personality-analysis
הנתונים מחולקים לשתי טבלאות:

א. טבלת לקוח – מכילה את פרטי הלקוח.

ב. טבלת מכירות – מכילה את כלל הקניות שלקוח עשה בחנות.

2. **שאלות המחקר**

- **<u>Supervised:</u>** האם לקוח יקנה בחנות (בהתבסס על נתוניו האישיים) כתוצאה מפרסום של קמפיין.

<u>KPIs</u>:

1. אחוז השתתפות אקטיבי בלפחות 10% בקמפיין לבעלי הכנסה מעל $30,000 בשנה.
2. אחוז השתתפות אקטיבי בלפחות 15% בקמפיין ללקוחות בעלי ותק בחנות של לפחות שנה.

<u>SMART</u>: המדדים ספציפיים בהיותם מתייחס לשאלת המחקר. רלוונטי מכיוון שמודד את כמות האנשים שביצעו קנייה בתגובה לקמפיין בהתאם למאפייני הלקוח, בנוסף נוכל למדוד אותו לאורך הזמן הרצוי.

- **<u>Unsupervised:</u>** חלוקת הלקוחות לקבוצות בעלות מאפיינים זהים, אשר תתרום להבנת הקבוצה בעלת הפוטנציאל הגבוה לרכישה.

<u>KPIs</u>:

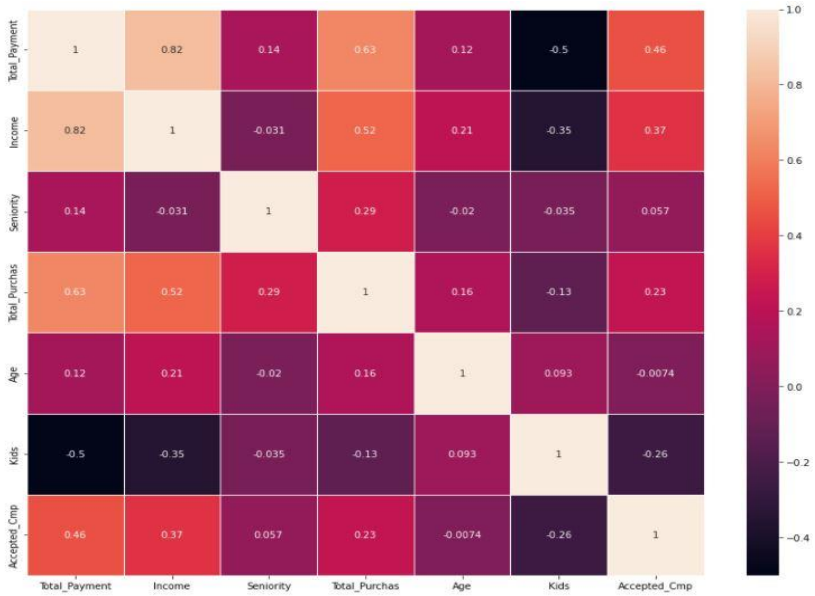
קבוצה בעלת ממוצע קניה של לפחות $600 בעת קנייה בחנות.

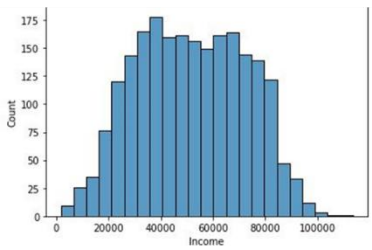
<u>SMART</u>: המדד ספציפי כי רק סוג אוכלוסייה מסוים כלול בתוכו, רלוונטי לשאלת המחקר מכיוון שלעמודה יש קשר עם הצלחת החנות, בנוסף נוכל למדוד אותו לאורך הזמן הרצוי.

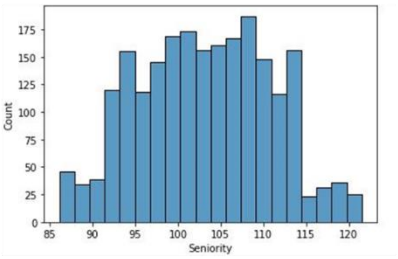
3. **הבנת הנתונים**

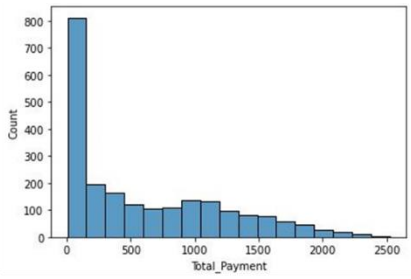
א. Supervised learning

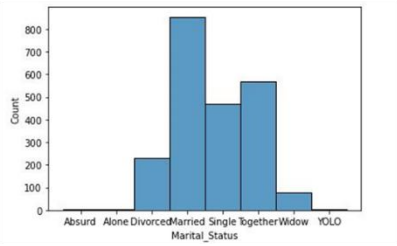
חקירה ראשונית של בסיס הנתונים :

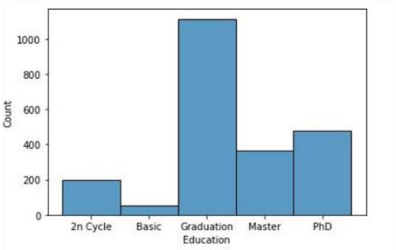












	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds
count	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000
mean	2.325000	4.084821	2.662054	5.790179	5.316518	303.935714	26.302232	166.950000	37.525446	27.062946	44.021875
std	1.932238	2.778714	2.923101	3.250958	2.426645	336.597393	39.773434	225.715373	54.628979	41.280498	52.167439
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	2.000000	0.000000	3.000000	3.000000	23.750000	1.000000	16.000000	3.000000	1.000000	9.000000
50%	2.000000	4.000000	2.000000	5.000000	6.000000	173.500000	8.000000	67.000000	12.000000	8.000000	24.000000
75%	3.000000	6.000000	4.000000	8.000000	7.000000	504.250000	33.000000	232.000000	50.000000	33.000000	56.000000
max	15.000000	27.000000	28.000000	13.000000	20.000000	1493.000000	199.000000	1725.000000	259.000000	263.000000	362.000000

ב. Unsupervised learning

חקירה ראשונית של בסיס הנתונים:

	ID	Year_Birth	Income	Kidhome	Teenhome	Recency
count	2240.000000	2240.000000	2216.000000	2240.000000	2240.000000	2240.000000
mean	5592.159821	1968.805804	52247.251354	0.444196	0.506250	49.109375
std	3246.662198	11.984069	25173.076661	0.538398	0.544538	28.962453
min	0.000000	1893.000000	1730.000000	0.000000	0.000000	0.000000
25%	2828.250000	1959.000000	35303.000000	0.000000	0.000000	24.000000
50%	5458.500000	1970.000000	51381.500000	0.000000	0.000000	49.000000
75%	8427.750000	1977.000000	68522.000000	1.000000	1.000000	74.000000
max	11191.000000	1996.000000	666666.000000	2.000000	2.000000	99.000000

4. <u>אנטרופיה:</u>

בהמשך לשאלת ה- Supervised, חישבנו את ערך האנטרופיה עבור כל העמודות המתארות את תכונותיו של הלקוח.

Income entropy: 10.881

Dt_Customer: 9.151

Year_Birth entropy: 5.515

Education entropy: 1.849

Marital_Status entropy: 2.048

5. מחישוב האנטרופיה מעלה ניתן לראות ששני הערכים הנמוכים ביותר הם:

- Education
- Marital_Status

עבור תכונות אלה נחשב את מדדי ה- gini-index:

Education gini: 0.664

Dt_Customer gini: 0.726

info_gain_Education: 0.0064

info_gain_Education: 0.0072

6. מניתוח ראשוני של הנתונים ניתן לראות כי קיים קשר בין מצב המשפחתי של הלקוח לבין ההוצאות הקנייה שלו בחנות בנוסף לעמודת האם יש ללקוח ילדים יש קורלציה שלילית לכמות הכסף שלקוח יוציא בחנות. לקוחות בעלי הכנסה גבוהה משתתפים ביותר קמפיינים וסכום הוצאות הקנייה שלהם הממוצע שלהם גבוה מממוצע ההוצאות הכללי של שאר אוכלוסיית הלקוחות.