כריית נתונים ב- R // תרגיל בית מספר 3

חברת *הייטקס* היא חברה המשווקת ציוד סטריאופוני, מחשבים אישיים ומוצרים אלקטרוניים אחרים. *הייטקס* מפרסמת את מוצריה על ידי דיוור קטלוגים ללקוחותיה, וכל ההזמנות שלה נלקחות דרך הטלפון. על מנת ללמוד את יעילות השיווק באמצעות קטלוגים, החברה אספה נתונים על 1000 לקוחות בסוף השנה הנוכחית. (ראה את הקובץ Catalogs.csv באתר הקורס ; כל שורה במסד מתאימה ללקוח מסוים).

להלן תיאור משתני המסד :

שם משתנה	תיאור
Age	משתנה גיל הינו אורדינלי בעל 3 ערכים : 1 – צעיר 2 – אמצע 3 – מבוגר
Gender	מין הלקוח
Married	סטטוס משפחתי (רווק/ נשוי)
Location	האם הלקוח מתגורר בסמיכות לסניף של החברה (קרוב/ רחוק)
Salary	משכורת חודשית ממוצעת
Children	מספר ילדים
Catalogs	מספר קטלוגים שנשלחו בשנה האחרונה
AmountSpent	הוצאות בחנות

<u>שאלות</u>

1. קראו את המסד (השתמשו בפרמטר stringsAsFactors = TRUE). פצלו את המסד ל- training (60%) ו- validation (40%).
2. (3 נק') לכל אחד מהמשתנים האורדינלים במסד (age, children, catalogs) :
   א. הציגו את הקשר בין המשתנה האורדינלי למשתנה התלוי (AmountSpent) על ידי גרף עמודות (אין צורך להציג את הגרף בקובץ הפלט). זכרו לכלול רק רשומות שנמצאות ב- training set.
   ב. קבעו האם יש להתייחס למשתנה כמשתנה כמותי או נומינלי (רמז : האם הקשר לינאריז?). דווח את מסקנתך לפלט.
3. (2 נק') בחנו את המשתנים הכמותיים Salary ו- AmountSpent. האם משתנים אלו מתפלגים נורמלית, או לפי התפלגות זנב ימין? דווח את מסקנתך לפלט.
4. (3 נק') על סמך מסקנותיכם מסעיף 2 ו-3, בנו מודל רגרסיה מרובה לחיזוי ההוצאות בחנות (AmountSpent), הכולל את כל המשתנים במודל. בדקו את ביצועי המודל על ה-

training set וה- validation set. דווחו לפלט את טבלאות הרגרסיה (פונקציית summary()) ואת מדדי ה- RMSE וה- MAPE על שני המדגמים. באיזה מקרה קיבלתם תוצאות טובות יותר? הסבירו את המשמעות של המדדים RMSE ו MAPE. יש להשתמש בפונקציות הנלמדו בהרצאה.

5. (2 נק') חשבו באמצעות ChatGPT RMSE ו MAPE (החישוב צריך להופיע בקוד). האם הוא ביצע שימוש בספריות הנלמדו בהרצאה? האם קיבלתם תוצאה זהה לסעיף הקודם? דווחו לפלט מה ביקשתם והאם התוצאה היתה זהה.

<u>אופן הגשה:</u>

- ✓ הגשה דרך אתר למידה
- ✓ הגשה בזוגות או ביחידים (רק אחד מבני הזוג צריך להגיש. על הקובץ יופיעו השמות של שני המגישים)
- ✓ יש להגיש קובץ R **מתועד**, וקובץ **פלט** המכיל את התשובות לשורות המסומנות בצהוב