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 מבסד 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	מיל הלקוח
Married	סטטוס משפתחי (רווק/ נשוי)
Location	האם הלקוח מתגורר בסמיכות לסניף של החברה (קרוב/ רקוח)
Salary	משכורת חודשים ממוצעת
Children	מספר ילדים
Catalogs	מספר קטלוגים שנשלחו בשנה האחרונה
AmountSpent	הוצאות בחנות

<u>שאלות</u>

1. פצלו את המסד ל- training (60%) ו- validation (40%).
2. (3 נק') לכל אחד מהמשתנים האורדינלים במסד (age, children, catalogs):
   א. הצג את הקשר בין המשתנה האורדינלי למשתנה התלוי (AmountSpent) על ידי גרף עמודות (אין צורך להציג את הגרף בקובץ הפלט). זכור לכלול רק רשומות שנמצאות ב- training set.
   ב. קבע האם יש להתייחס למשתנה כמשתנה כמותי או נומינלי (רמז: האם הקשר לינאריז?). דווח את מסקנתך לפלט.
3. (2 נק') בחן את המשתנים הכמותיים Salary ו- AmountSpent. האם משתנים אלו מתפלגים נורמלית, או לפי התפלגות זנב ימין? דווח את מסקנתך לפלט.
4. (5 נק') על סמך מסקנותיכם מסעיף 2 ו-3, בנו מודל רגרסיה מרובה, ללא אינטרקציות, הכולל את כל המשתנים במודל. בדקו את ביצועי המודל על ה- training set וה-

validation set. דווחו לפלט את טבלת הרגרסיה (פונקציית summary()) ואת מדדי ה- ; וה- MAPE על שני המדגמים.

<u>אופן הגשה</u>:

- ✓ הגשה דרך אתר למידה
- ✓ יש להגיש קובץ R מתועד, וקובץ פלט.