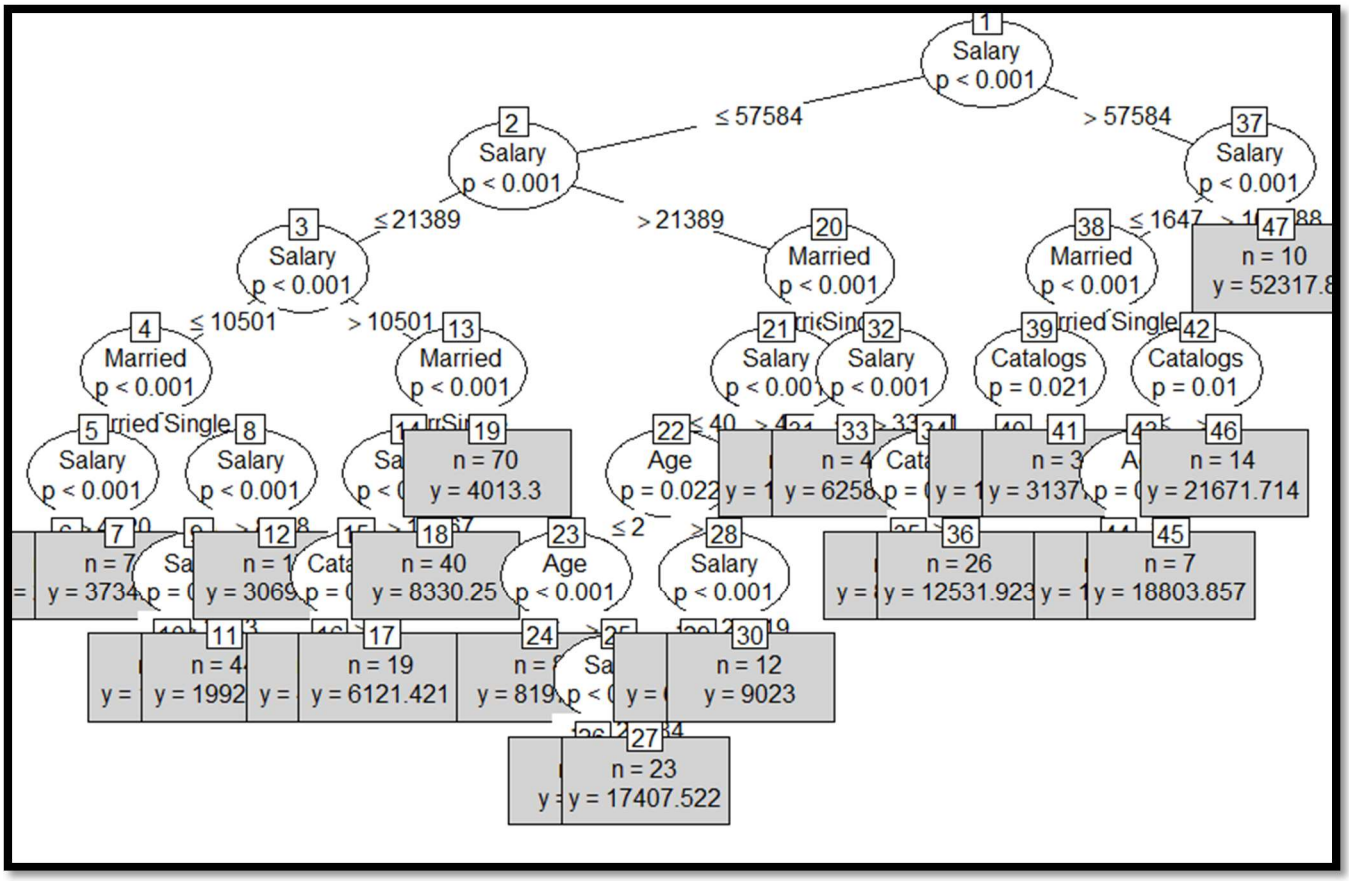
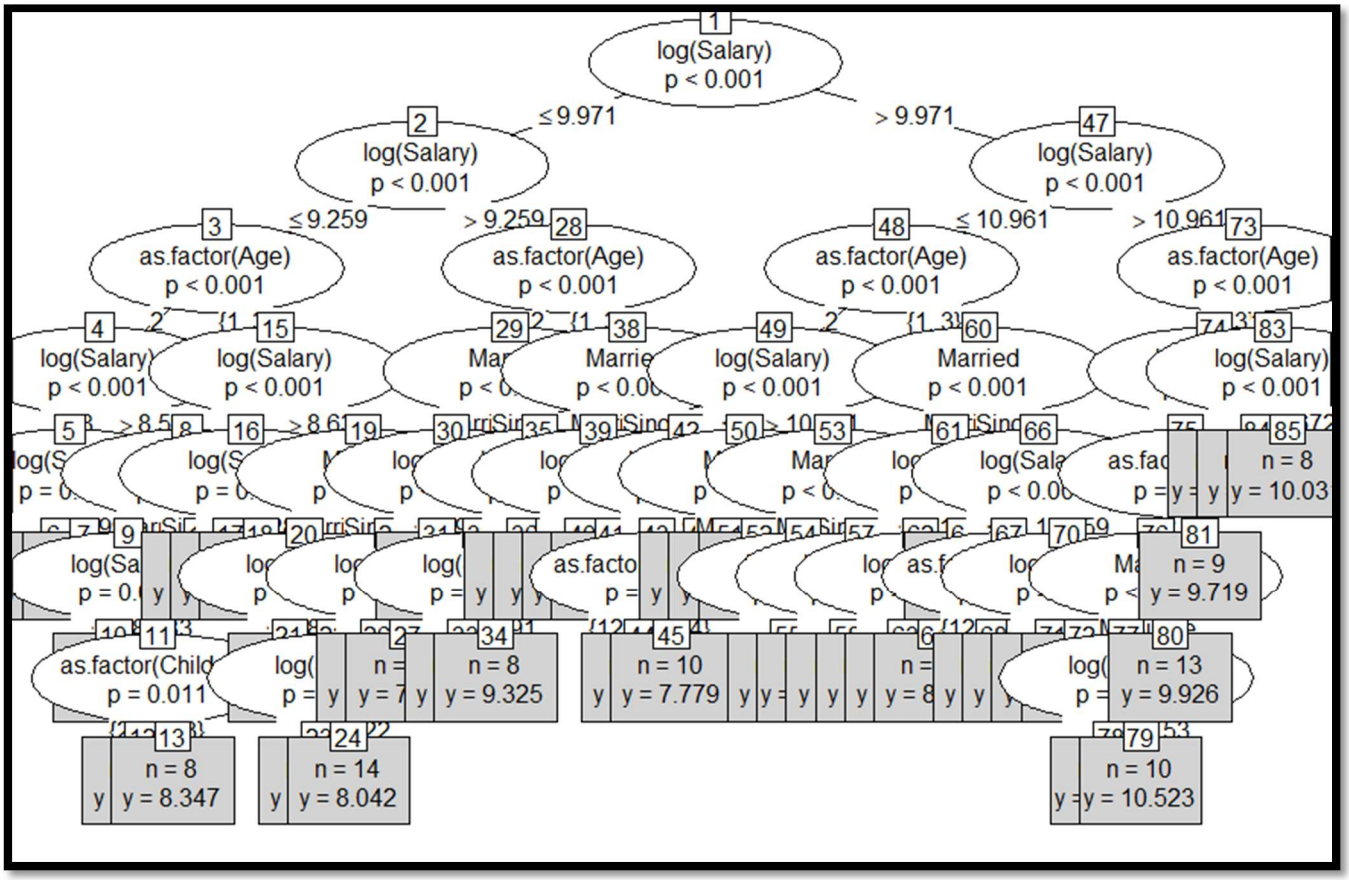
יישומי R בכריית נתונים, תרגיל 5 – קובץ פלט

1. **עץ ההחלטה מבוסס רגרסיה, מאחר והמשתנה התלוי (Amount spent) הינו רציף ולא קטגוריאלי. לדעתנו נכון יותר להריץ עץ רגרסיה כלל שמספר המשתנים הבלתי תלויים רב יותר, מאחר והנ"ל יוצר תופעת OVERFIT ברגרסיה הלינארית, ואילו בעץ הרגרסיה ניתן להציג בצורה ויזואלית ברורה אילו משתנים ב"ת תלויים הם המשפיעים ביותר על המשתנה התלוי. ולאחר מכן, להשתמש במשתנים אלו על מנת להריץ רגרסיה לניארית בצורה מושכלת, רק על המשתנים הקריטיים ביותר.**

2. **התאמת עץ רגרסיה ל-training data תוך שימוש בפונקציית ctree, מניבה את העץ הנ"ל:**



3. התאמת עץ חדש על ידי ביצוע טרנפורמציית לוג-לוג ושינוי המשתנים הקטגוריאליים ל-as.factor, מניבה עץ החלטה חדש tr_new.



4. מדדי RMSE:

א. עץ מסעיף 2:

```
> RMSE.rtree
[1] 6592.309
```

ב. עץ מסעיף 3:

```
> RMSE.rtree2
[1] 5964.473
```

ג. עץ מתרגיל קודם:

```
> pred<- predict(reg, newdata = valid.df)
> accuracy(exp(pred), valid.df$AmountSpent)
                ME    RMSE      MAE        MPE
Test set -77.76938 1082.9 333.1728 -0.4992875
```

כלומר, ההנחה הראשונית שלנו לא הייתה נכונה, ובמקרה זה עדיף להשתמש ברגרסיה לינארית (אולי כי אין יותר מידי משתנים בלתי תלויים ולא נוצרת בעיית overfit ו-multicollinearity) ובכך הרגרסיה הלינארית חוזה את הנתון התלוי ברמת שגיאה נמוכה יותר מעצי ההחלטה במקרה זה.