יישומי R בכריית נתונים, תרגיל 3 – קובץ פלט

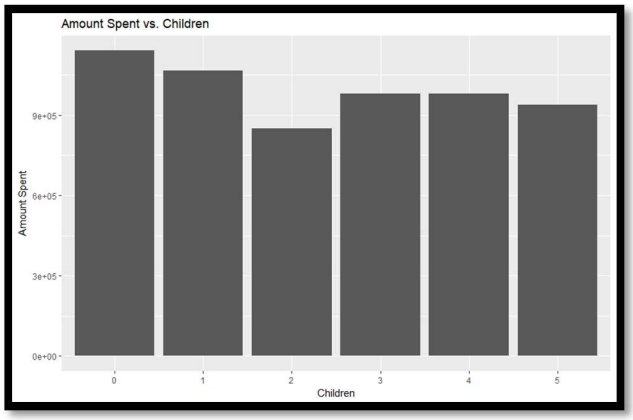
2. הקשר בין המשתנים האורדינאלים למשתנה התלוי-Amount spent:

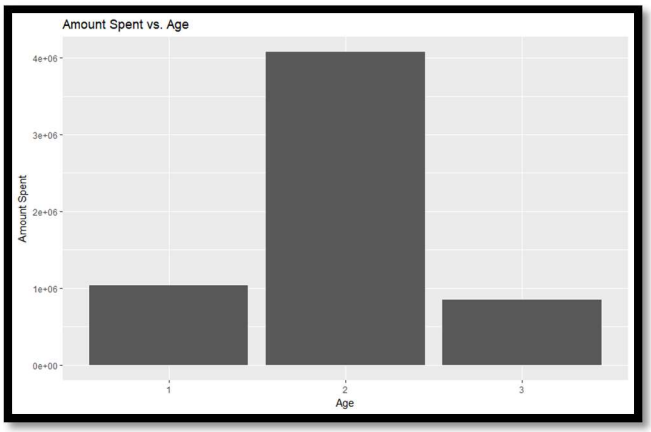
א. i. **ביחס למשתנה גיל-** צעירים מוציאים יותר ממבוגרים, ואילו אלה שבאמצע מוציאים הרבה יותר משני הקודמים יחד.

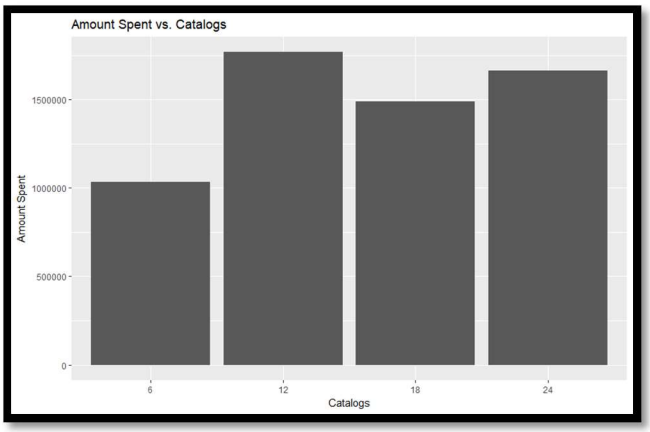
ii. **ביחס למשתנה ילדים-** לקוחות בעלי 2 ילדים מוציאים פחות מהשאר, לקוחות חסרי ילדים מוציאים יותר מכולם, לקוחות בעלי 3 ו-4 ילדים מוציאים סכום דומה הגבוה יותר מאלו בעלי 2 ילדים ונמוך מאלו חסרי ילדים או בעלי ילד 1.

iii. **ביחס למשתנה קטלוגים-** לקוחות להם נשלחו 6 קטלוגים מוציאים פחות מכולם, לקוחות להם נשלחו 12 קטלוגים מוציאים יותר מכולם, לקוחות להם נשלחו 18/24 קטלוגים מוציאים ערכים בין שני הקיצונים מעלה.

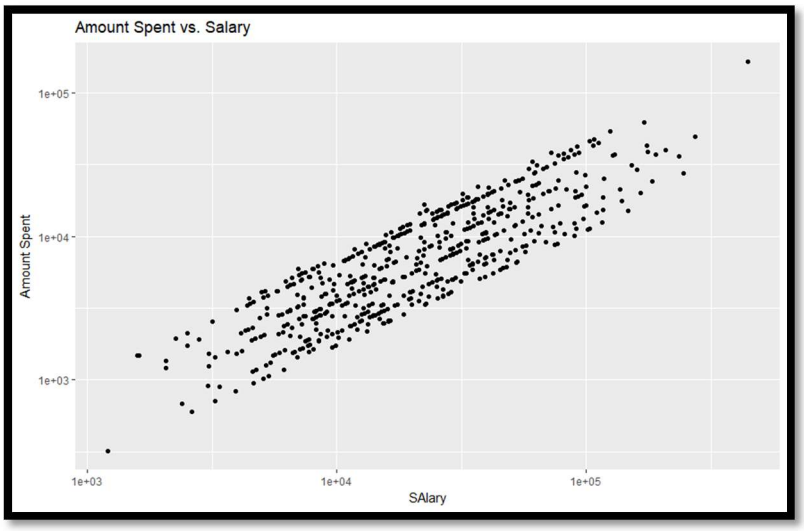
ב. מאחר והקשר לא לינארי, יש להתייחס למשתנה כאל משנה נומינאלי (הוגדר בתור as.factor).







3. שני המשתנים הכמותיים – AmountSpent ו-Salary מתפלגים לפי התפלגות זנב ימין, רק אחרי שמפעילים לוג (במקרה זה- לוג ע"ב 10) <u>על שניהם</u>, רואים קשר לינארי חיובי בין שניהם. כלומר פונקציית לוג תרמה ל"פתיחת" התפלגות זנב ימין ולהצגת הקשר.



4.

```
> summary(reg)

Call:
lm(formula = log(AmountSpent) ~ log(Salary) + as.factor(Age) +
    as.factor(Catalogs) + as.factor(Children) + Gender + Married +
    Location, data = train.df)

Residuals:
      Min        1Q    Median        3Q       Max
-0.233291 -0.018947 -0.008226  0.002297  0.213046

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             0.220248   0.027646   7.967 8.54e-15 ***
log(Salary)             0.809157   0.002361 342.646  < 2e-16 ***
as.factor(Age)2         0.710329   0.005740 123.747  < 2e-16 ***
as.factor(Age)3         0.011480   0.006874   1.670   0.0955 .
as.factor(Catalogs)12   0.192359   0.006454  29.805  < 2e-16 ***
as.factor(Catalogs)18   0.199448   0.006868  29.038  < 2e-16 ***
as.factor(Catalogs)24   0.199915   0.006863  29.131  < 2e-16 ***
as.factor(Children)1    0.011115   0.007763   1.432   0.1527
as.factor(Children)2   -0.009095   0.007827  -1.162   0.2457
as.factor(Children)3   -0.003668   0.007925  -0.463   0.6437
as.factor(Children)4    0.104578   0.008064  12.969  < 2e-16 ***
as.factor(Children)5    0.106482   0.007979  13.346  < 2e-16 ***
GenderMale              0.184545   0.012056  15.307  < 2e-16 ***
MarriedSingle          -0.389350   0.004885 -79.703  < 2e-16 ***
LocationFar             0.006617   0.005067   1.306   0.1921
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05625 on 585 degrees of freedom
Multiple R-squared:  0.9962,    Adjusted R-squared:  0.9961
F-statistic: 1.097e+04 on 14 and 585 DF,  p-value: < 2.2e-16

> pred <- predict(reg, newdata = train.df)
> accuracy(exp(pred), train.df$AmountSpent)
               ME     RMSE      MAE        MPE     MAPE
Test set 42.82825 1295.141 362.3903 -0.1509736 2.71515
>
> pred<- predict(reg, newdata = valid.df)
> accuracy(exp(pred), valid.df$AmountSpent)
                ME   RMSE      MAE        MPE     MAPE
Test set -77.76938 1082.9 333.1728 -0.4992875 3.065753
```