Regression Analysis in Excel Project

Task 1

- What is the correlation coefficient between the **Discount** and **Total Amount Spent**?

correlation coefficient	0.406

- How would you interpret this value?

  A correlation coefficient of 0.406 suggests that there is a moderate positive correlation between the two variables.

- Based on the scatter plot, how would you describe the relationship between **Discount** and **Total Amount Spent**?

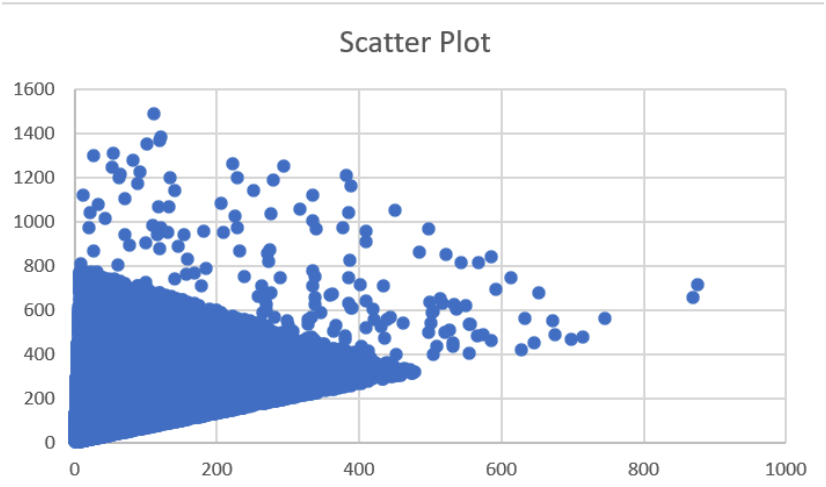

Figure 1. Scatter plot for Discount and Total Amount Spent (Y - total time spent, X - Discount)

The scatter plot shows a general linear relationship between the variables, with total time spent increasing as the discount increases. However, there are still many data points that deviate from this trend.

- Does the scatter plot suggest a positive, negative, or no correlation, and is the relationship linear or non-linear?

The scatter plot suggests a positive linear relationship, where total time spent tends to increase as the discount increases.

Task 2

- What are the independent and dependent variables in this analysis?

The dependent variable is the variable we are trying to predict or explain, in this case its: total time spent

The independent variable is the variable used to predict or explain the dependent variable: discount

- What is the equation of the regression line?

y = slope * x + intercept.

- What are the values of the slope and intercept, and what do they represent in this context?

$Y = 127.29 + 0.751X$

If no discount is offered, the model predicts that the total time spent will be 127.29 units. for each 1% increase in discount, the total time spent increases by approximately 0.751 units.

- How well does the regression model fit the data?

The R-squared value of 0.165 indicates that only 16.5% of the variation in total time spent can be explained by the discount offered. This suggests a weak relationship between the two variables, meaning that discount alone is not a strong predictor of time spent. Most of the variation in user behavior is likely influenced by other factors not captured in this model. Therefore, while a positive trend exists, the low R-squared value shows that the linear regression does not fit the data well.

Task 3

- Based on the skewness values, do any variables appear to have distributions that deviate from normality?

Here are the skewness values extracted from the data:

skew	Without transformation	With transformation (log)
Product price	0.636	-0.933
Discount	1.878	-0.709
Total Amount Spent	1.582	-0.522

Table. 1. Skewness values

The original (untransformed) data showed positive skewness for all variables, meaning each distribution had a longer right tail. Among them, Product Price was closest to a normal distribution (its skew was nearer to zero), while Discount and Total Amount Spent were more heavily skewed.

After applying a log transformation, all three variables became negatively skewed, meaning the tails shifted to the left. However, the transformation successfully reduced the skewness for Discount and Total Amount Spent, bringing them closer to a symmetric (normal-like) distribution. In contrast, Product Price became more negatively skewed, suggesting that for this variable, the log transformation may have overcorrected the original skew.

- Would a log transformation of any variable improve the linearity of the relationship?

As seen above, applying a log transformation to the Discount and Total Amount Spent variables improves their distributions, bringing them closer to normality and potentially enhancing the linearity of their relationship in the regression model.

- How does the R-squared value change (if at all) after transformation? Why?

R Square	Without transformation	With transformation (log)
	0.471	0.618

Table. 2. R squared value before and after log transform

As seen in the table above, the R-squared value is higher after the transformation. This may be because applying a log transformation to variables like Discount and Total Amount Spent improved their distributions, bringing them closer to normality and enhancing the linear relationship with the dependent variable. However, since Product Price did not require transformation, its increased skewness after transformation suggests that it may not have contributed to the improvement.

- Are the coefficients interpretable in the context of the problem, especially after any transformations?

coefficients	Without transformation	With transformation (log)
	-1.487	0.168

Table. 3. R squared value before and after log transform

The change in the coefficient's sign—from negative before the transformation to positive after—suggests that the relationship between the variable and the target may appear different depending on how the data is scaled. This could happen if the original relationship was non-linear or influenced by skewed data. After transformation (e.g., using logs), the relationship may become more linear and reveal the true direction of the effect, leading to a positive coefficient.

Compared to the initial regression model that excluded Product Price, the results of this model are significantly improved.

Task 4

- What is the R-squared of the model, and what does this tell you about the model?

R Square	0.652

Slightly higher than the previous regression where we only added where we only added product cost as another variable.

- Which variables significantly predict the Total Amount Spent, and how do you know?

Product Cost has a coefficient of 0.516 and a p-value of 0.000, indicating it is statistically significant.

Since it has both the highest coefficient and the lowest p-value, we can conclude that Product Cost is the strongest predictor of Total Amount Spent in the model.

- What is the interpretation of the coefficients of the significant variables?

For every 1-unit increase in Product Cost, the Total Amount Spent increases by approximately 0.516 units, assuming all other variables remain constant.