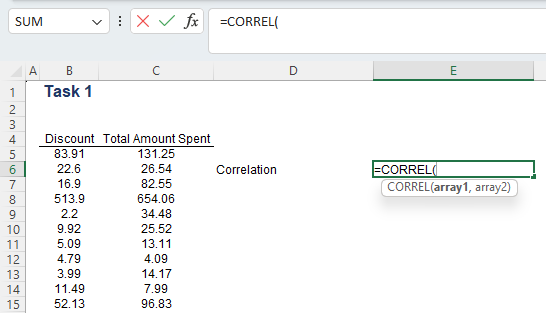
**Part 1:** Correlation Analysis

**Task 1**

*You can find the solution to this problem in the Correlation Analysis\_Solution.xlsx file:*

*Sheet: Task 1*

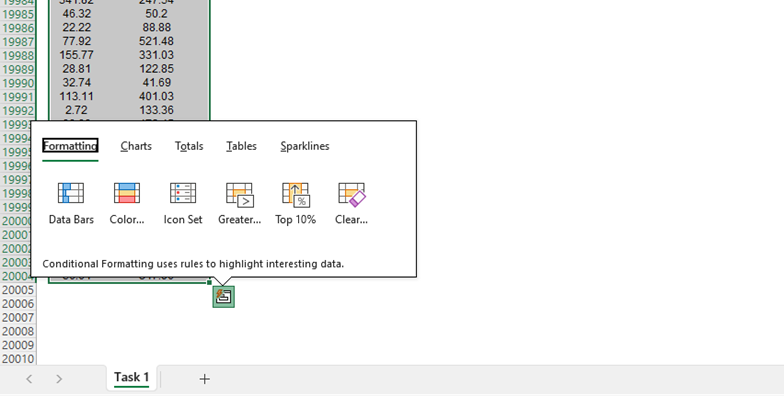
1. The first step is calculating the correlation between the **Discount** and **Total Amount Spent.**Use Excel’s **CORREL()** function to do this.



2. The next step is to create a scatter plot to better understand the nature of the relationship between the two variables. Note the steps below to create a scatter plot in Excel using the Quick Analysis feature.

Select both the **Discount** and **Total Amount Spent** columns.

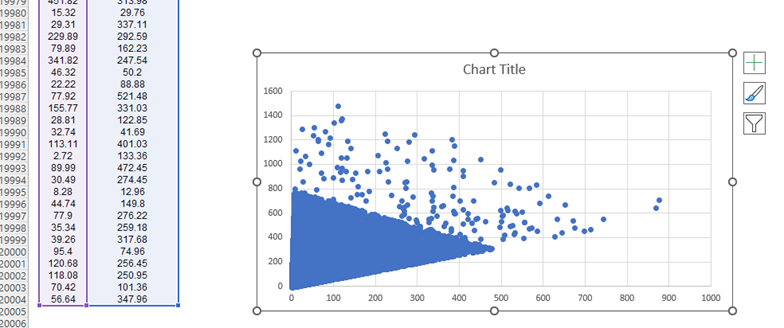
Once you've selected your data, the Quick Analysis button will automatically appear at the bottom right of your selected data. Click on this button.



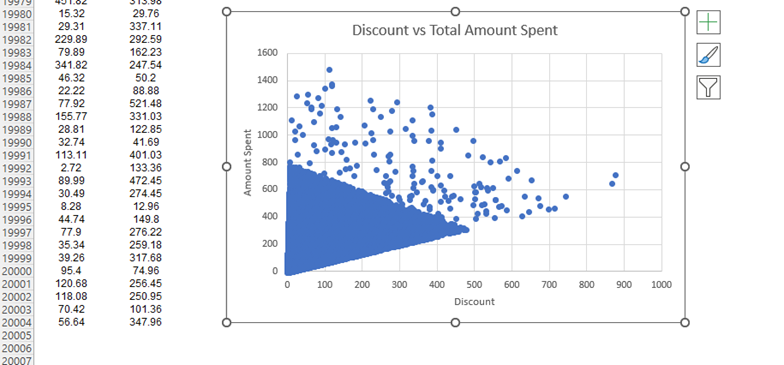
A new menu will appear. Select Charts from this menu.



In the Charts tab, click on Scatter. Excel will automatically generate a scatter plot from your selected data.



Click on the scatter plot to select it. Once the scatter plot is selected, you can use the tools in the Excel Ribbon Chart Tools section to customize the appearance of your plot.



A positive correlation of approximately 0.406 between the **Discount** and the **Total Amount Spent**exists. This correlation is moderate, suggesting a relationship between these two variables but is not particularly strong.

The scatter plot visualizes the relationship between the **Discount** and the **Total Amount Spent** for each transaction in the dataset. Each point on the plot represents a single transaction, with its position along the x-axis indicating the discount and its status along the y-axis indicating the total amount spent.

Looking at the scatter plot, you can see that as the **Discount** increases, the **Total Amount Spent** also tends to increase, which aligns with the positive correlation we calculated. This trend suggests that customers spend more when the discount is higher—perhaps due to customers perceiving higher value in discounted products and willing to spend more.

But the spread of points on the plot shows variability in the data. While the general trend is upward, many transactions do not follow this trend. For example, there are transactions with high discounts but relatively low total amounts spent and transactions with low discounts but high total amounts expended.

These outliers could be due to many factors. For instance, some customers might be more price-sensitive than others and more influenced by discounts. Or certain products might have been discounted heavily but are not popular or do not meet customers' needs, leading to lower total spending. Alternatively, some products might have low discounts but are very popular or in high demand, leading to higher total expenditures. Outliers are not always wrong; they can be essential to the data and provide valuable information.

In conclusion, while there’s a positive relationship between the **Discount** and the **Total Amount Spent**, there’s significant variability in the data, and many transactions do not follow the overall trend. This suggests that while discount is a factor that can influence total spending, there are also other aspects at play.

Correlation Analysis\_Solution.xlsx

**Part 2:** Simple Linear Regression

**Task 2**

*You can find the solution to this problem in the Simple Linear Regression\_Solution.xlsx file:*

*Sheet: Task 2*

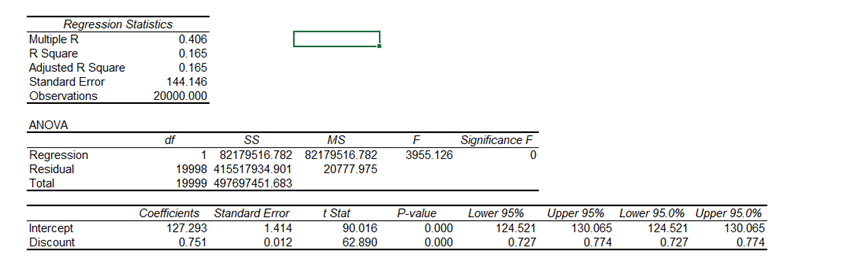
In this case, you're trying to predict the **Total Amount Spent** based on the **Discount** given:

The **dependent variable** is the **Total Amount Spent**. This is the variable that you’re interested in predicting.

The **independent variable** is the **Discount**. This is the variable used to make the prediction.

In general, the dependent variable is what you're interested in understanding or predicting, and the independent variable(s) are the factors you hypothesize will impact the dependent variable.

The next step is to perform linear regression using the Data Analysis ToolPak. Note summary of the results below.



Interpret the results of the linear regression analysis:

**Slope (0.751)**. The slope of the regression line is approximately 0.751. This means that for each unit increase in **Discount**, the **Total Amount Spent** increases by about 0.751 units, assuming all other variables are held constant. This indicates that the **Discount** positively impacts the **Total Amount Spent**.

**Intercept (127.293)**. The intercept of the regression line is approximately 127.293. This is the estimated value of the **Total Amount Spent** when the **Discount** is zero—assuming all other variables are held constant. It acts as a baseline value for the **Total Amount Spent**.

**Prob (F-statistic) (also P-value)**. This is the p-value associated with the F-statistic. It tells us the probability of obtaining an F-statistic (3955.126) as extreme as ours if the null hypothesis is true—i.e., if the true slope is zero. A small p-value indicates strong evidence to reject the null hypothesis. Our p-value is 0.00, less than the commonly used significance level of 0.05, so we would reject the null hypothesis and conclude that our model provides a significant fit to the data. In other words, the **Discount** significantly explains the **Total Amount Spent**.

**Correlation Coefficient (R, 0.406)**. The correlation coefficient (R-value) is approximately 0.406. This indicates a moderate positive linear relationship between the **Discount** and the **Total Amount Spent**. A positive R-value signifies that the **Discount** increases as the **Total Amount Spent** increases.

**Coefficient of Determination (R-squared, 0.165)**. The coefficient of determination (R-squared) is approximately 0.165. This means that the **Discount** can explain about 16.5% of the variability in the **Total Amount Spent**. In other words, the **Discount** explains about 16.5% of the variation in the **Total Amount Spent**. The remaining 83.5% could be explained by other factors not included in the model.

So, while there’s a positive relationship between the **Discount** and the **Total Amount Spent**—and the **Discount** explains some of the variability in the **Total Amount Spent**—other factors are likely influencing the **Total Amount Spent** not captured in this simple linear regression model.

Simple Linear Regression\_Solution.xlsx

**Part 3:** Multiple Linear Regression

**Task 3**

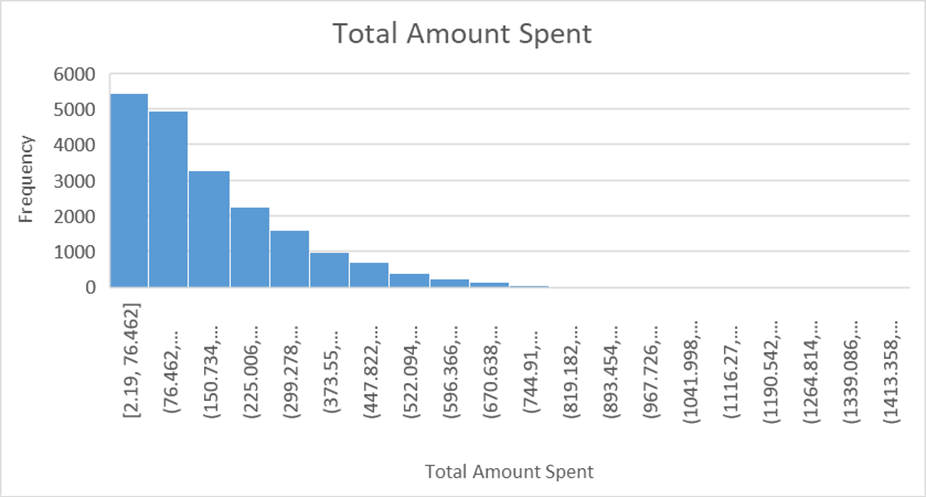
*You can find the solution to this problem in the Multiple Linear Regression\_Solution.xlsx file:*

*Sheet: Task 3*

Here are the histograms of the three variables:







**Distributions**:

**Discount**: This distribution appears right-skewed, with most of the values clustered on the left and a long tail on the right.

**Product Price**: This distribution also appears to be slightly right-skewed, though not as pronounced as the **Discount**.

**Total Amount Spent**: The distribution is noticeably right-skewed, similar to **Discount**.

**Skewness Values**:

**Discount**: 1.87

**Product Price**: 0.63

**Total Amount Spent**: 1.58

Recall the general guidelines for skewness:

If skewness is between -0.5 and 0.5, the distribution is approximately symmetric.

If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed.

If skewness is less than -1 or greater than 1, the distribution is highly skewed.

Based on the skewness values and visual inspection:

Both **Discount** and **Total Amount Spent** are highly skewed to the right, and log transformation could help in making their distributions more symmetric.

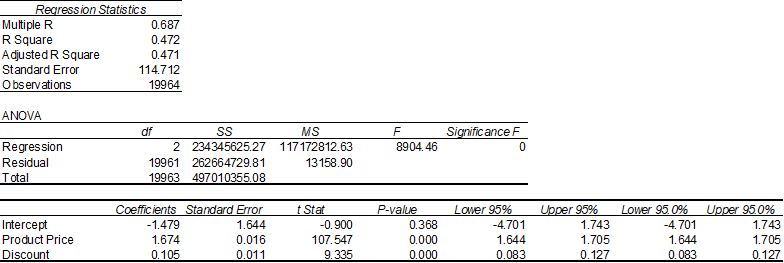
**Product Price** is moderately skewed to the right, and a log transformation might also be beneficial for this variable.

**Regression Analysis**

Given this analysis, it seems that log-transforming the variables could be beneficial in achieving more symmetrical distributions, especially for **Discount** and **Total Amount Spent**.

Use the Data Analysis ToolPak to perform the multiple linear regression in Excel. Note the results below:

No log transformation



The regression results without transformations are as follows:

**R-squared**: The model explains approximately 47.2% of the variance in the **Total Amount Spent**.

**Coefficients**:

**Discount**: For every unit increase in Discount, the **Total Amount Spent** increases by approximately 0.1051 units, holding all else constant.

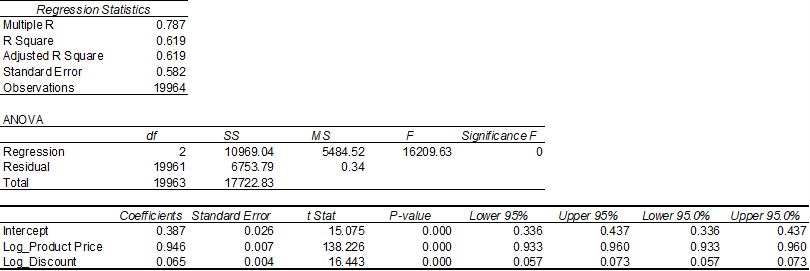
**Product Price**: For every unit increase in **Product Price**, the **Total Amount Spent** increases by approximately 1.6740 units, holding all else constant.

Both **Discount** and **Product Price** are statistically significant in predicting **Total Amount Spent**.



Note that when we plot the residuals from the regression model against the original Product Price values, there is a funnel shape. This funnel shape suggests that the residuals have increasing variance, which is indicative of heteroscedasticity. Therefore, Product Price should be log-transformed.

With log transformation



The regression results after the log-transformations are as follows:

**R-squared**: The model with the transformed variables explains approximately 61.9% of the variance in the **Log\_Total Amount Spent**, which is an improvement from the original model's 47.2%.

**Coefficients** (with transformed variables):

**Log\_Discount**: A 1% change in Discount results in an approximately 0.0711% change in the **Log\_Total Amount Spent**, holding all else constant.

**Log\_Product Price**: A 1% change in **Log\_Product Price** results in an approximately 0.9470% change in the **Log\_Total Amount Spent**, holding all else constant.

Both **Log\_Discount** and **Log\_Product Price** remain statistically significant in predicting the **Log\_Total Amount Spent**.

**Interpretation**:

Based on the R-squared values, the model with the log-transformed variables offers a better fit compared to the model with the original variables.

After the transformation, the interpretation of the coefficients changes. For the **Log\_Discount** coefficient, it's now about percentage changes due to the log transformation.

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

Yes, the log transformation of **Discount** and **Total Amount Spent** improved the linearity of their relationships with other variables, as evidenced by the higher R-squared value in the transformed model.

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

The R-squared value increased from 47.2% in the original model to 61.9% in the model with transformed variables, indicating a better fit.

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

With all variables log-transformed, the coefficients can be interpreted as elasticities:

**Log\_Discount**: A 1% change in Discount results in an approximately 0.065% change in the **Log\_Total Amount Spent**, holding all else constant.

**Log\_Product\_Price**: A 1% change in **Product Price** results in an approximately 0.9460% change in the **Log\_Total Amount Spent**, holding all else constant.

The interpretations are in terms of percentage changes due to the log transformations, and it's crucial to keep this in mind when discussing the results.

Multiple Linear Regression\_Solution.xlsx

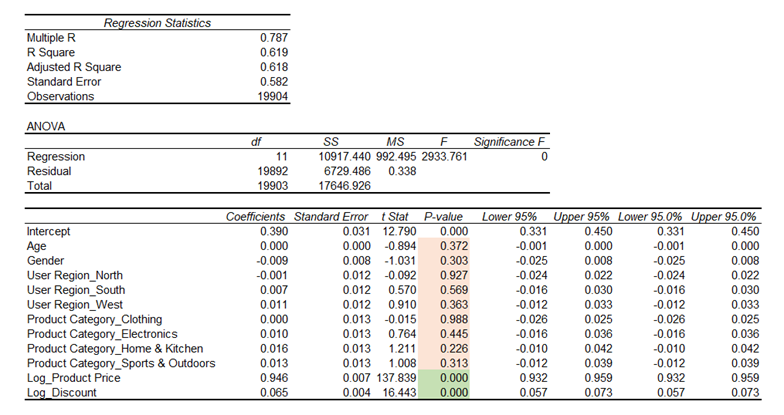
**Part 4:** Advanced Multiple Linear Regression

**Task 4**

*You can find the solution to this problem in the Advanced Multiple Linear Regression\_Solution.xlsx file:*

*Sheet: Task 4*

Note the results of the regression below:



**R-squared**. The R-squared value is 0.619, which means our model explains about 61.9% of the variation in the **Log\_Total Amount Spent**. In other words, about 61.9% of changes in the **Log\_Total Amount Spent** can be explained by differences in our independent variables.

**F-statistic and Prob (F-statistic)**. The F-statistic is 2933, and the p-value is 0.00, less than 0.05. This indicates that our independent variables are statistically significant and contribute to predicting the **Log\_Total Amount Spent**.

**Coefficients and significance of coefficients (P>|t|)**.

The coefficient for the constant term is 0.390.

The coefficient of **Log\_Product Price** is 0.946. This means that, on average, a 1% increase in the product price is associated with an approximately 0.946% increase in the **Log\_Total Amount Spent**, holding all other variables constant.

The coefficient of **Log\_Discount** is 0.065. This suggests that, on average, a 1% increase in the discount is associated with an approximately 0.065% increase in the **Log\_Total Amount Spent**, holding all other variables constant.

Other variables **are not** **significant predictors** because their p-values are more than 0.05, including **Age, Gender, User Region\_North, User Region\_South, User Region\_West, Product Category\_Clothing, Product Category\_Electronics, Product Category\_Home & Kitchen, and Product Category\_Sports & Outdoors**.

**Regression Equation**. Based on these results, the multiple linear regression equation is as follows:

**log(Total Amount Spent)=0.3905+0.9468 × log(Product Price) + 0.0712 × log(Discount)+Other variables**

The Trendy Shopper should then consider the following strategies:

**Pricing Strategy**. The company could continue to offer products at a higher price if the quality and brand justify the cost. This is supported by the positive coefficient for the Product Price, suggesting that customers are willing to pay more for the company's products.

**Discount Strategy**. The positive coefficient for Discount indicates customers will likely spend more when discounts are offered. The company could consider offering strategic discounts to drive sales and increase customers’ spending.

**Customer Segmentation**. The lack of significant results for Age, Gender, User Region, and Product Category suggests that these factors may not significantly influence the total amount spent. But the company should not disregard these factors entirely since they could be significant in other aspects of customer behavior, and further analysis could provide insights for better customer segmentation and targeted marketing strategies.

It's essential to note that these strategies are based on the regression analysis results. Other factors, such as **market trends**, **competitive landscape**, **company objectives**, and **resources**, should also be considered when making strategic decisions. Additionally, the relationships identified in the regression analysis are associative, not causal. It’s also vital to validate the model with new data and assess its predictive performance.

Advanced Multiple Linear Regression\_Solution.xlsx

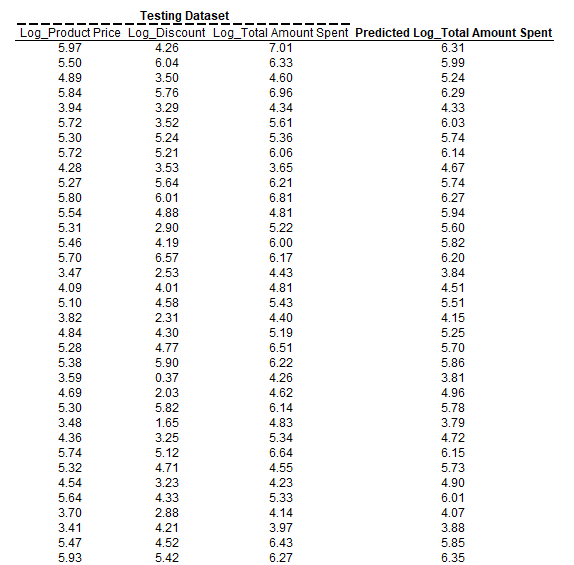
**Part 5:** Predictive Modeling and Next Steps

**Task 5**

*You can find the solution to this problem in the Predictive Modeling\_Solution.xlsx file:*

*Sheet: Task 5*

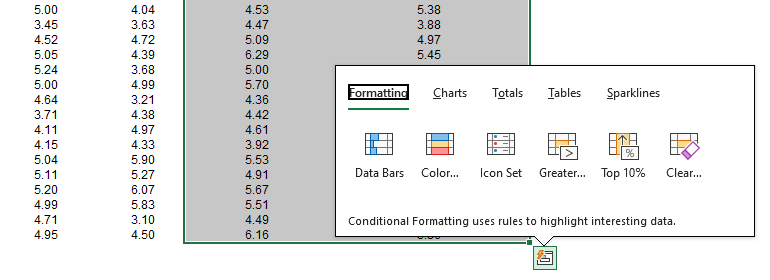
Note the results of the calculation below.



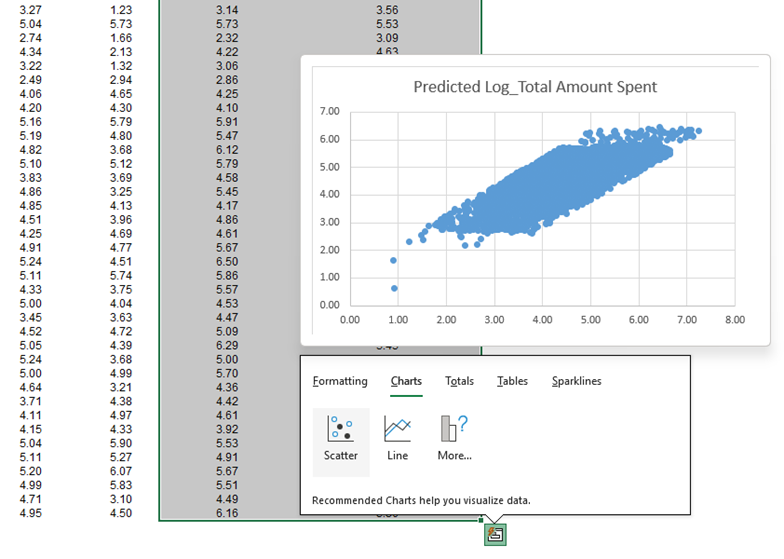
Once you've obtained the predicted values for the Total Amount Spent, you can use Excel's Quick Analysis tool to create a scatter plot of your actual and predicted values:

1. **Select your data**. Click and drag to select both columns (**Predicted Log\_Total Amount Spent** and **Log\_Total Amount Spent**).

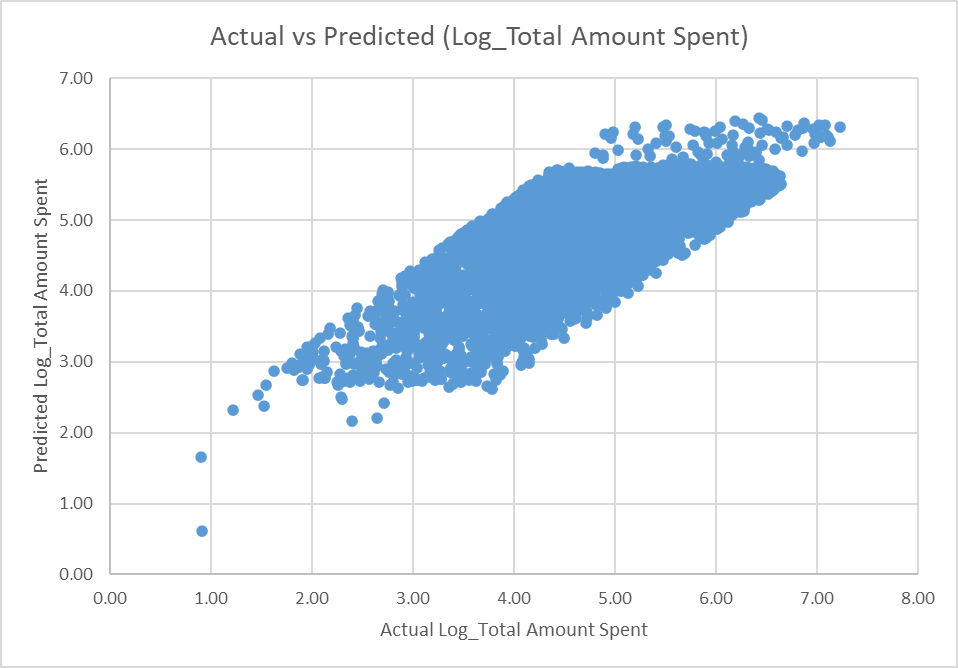
2.**Open Quick Analysis**. Once your data is selected, a small square icon (the Quick Analysis tool) will appear at the bottom right of the chosen area. Click on this icon.



3. **Insert scatter plot**. In the Quick Analysis dialog box, navigate to the Charts tab. Click on Scatter. A scatter plot will automatically be created in your workbook.



4. **Adjust the chart title and axis labels**. Click on the chart title and axis labels to adjust the text. For example, change the chart title to Actual vs. Predicted **Log\_Total Amount Spent**, the x-axis label to **Log\_Total Amount Spent**, and the y-axis label to **Log\_Predicted Total Amount Spent**.



The scatter plot compares the actual and predicted values of the log-transformed **Total Amount Spent**. The x-axis displays the actual log values, while the y-axis shows the model's predicted log values. Each point on the plot corresponds to a transaction from the test dataset.

A perfectly accurate model would result in all points lying on a 45-degree line, indicating that the actual value matches the predicted value for each point. Deviations from this line suggest discrepancies between the model's predictions and the actual data.

The plot shows that the model's predictions are not perfectly aligned with the actual values. The scatter plot reveals some potential outliers, which could be unduly influencing the model's predictions. These could be transactions with exceptionally high or low values of Total Amount Spent that do not follow the same pattern as the rest of the data.

While the model does capture some of the relationships between the **Discount**, **Product Price**, and **Total Amount Spent**, the scatter plot and MSE (observed in output in Task 4) suggest that the model's predictive performance could be improved. This could be achieved by including **additional predictors** in the model, **handling outliers**, or using a **more complex model** that can capture non-linear relationships or interactions between variables.

Predictive Modeling\_Solution.xlsx