Regression Analysis in Excel Project

Exploring the Influence of Various Determinants on Customer Spending

Case Description

In this Regression Analysis in Excel project, you'll be working with data from a company in the e-commerce sector, The Trendy Shopper—a fast-growing e-commerce business that offers a diverse selection of contemporary products across various categories, including fashion, electronics, and home decor. The company has a broad customer base and a wide range of products. However, the enterprise is unsure about the impact of product prices and discounts on customers' spending, which hampers its ability to create an effective pricing strategy to maximize sales and profits. It currently uses a one-size-fits-all marketing approach. But given the diversity in the customer base and the wide range of products, there may be more effective approaches. That's why The Trendy Shopper aims to leverage its transactional data to understand how factors (product prices, quantity purchases, discounts) affect overall expenditure by utilizing predictive modeling (simple and multiple linear regression). Therefore, your task will be to help the company optimize its pricing and discount strategies, potentially increasing sales and customer satisfaction.

Project files

Excel's Regression Analysis Dataset.xlsx file consists of four sheets: Task 1, Task 2, Task 3, Task 4, and Task 5. Each sheet contains specific information regarding the project's tasks and corresponding data. Use the data in these sheets to answer the questions that follow.

In this regression analysis in Excel, you'll delve into a dataset from The Trendy Shopper. This dataset is derived from the company's transactional records. It contains the following fields.

1. **Age:** customer's age
2. **Gender:** customer's gender
3. **User Region:** customer's location for The Trendy Shopper's distribution
4. **Product Category:** a diverse selection at The Trendy Shopper
5. **Product Price:** the monetary cost assigned to a specific product

6. **Discount:** the reduction in the original product price (given in absolute monetary value and not presented as a percentage)
7. **Total Amount Spent:** the shopper's total expenditure derived from multiplying the product price by quantity

Tasks:

1) Correlation Analysis:

As part of a broader assignment to perform predictive modeling for The Trendy Shopper, the first task is to explore and understand the relationship between Discount and Total Amount Spent. You must ascertain whether there is a correlation between these two variables and, if so, how strong it is. Visualize this relationship using a scatter plot to better understand the nature of the relationship.

- What is the correlation coefficient between the **Discount** and **Total Amount Spent**?
- How would you interpret this value?
- Based on the scatter plot, how would you describe the relationship between **Discount** and **Total Amount Spent**?
- Does the scatter plot suggest a positive, negative, or no correlation, and is the relationship linear or non-linear?

2) Simple Linear Regression:

The next task aims to create a linear regression model to better understand the relationship between the **Discount** and **Total Amount Spent**.
This analysis is necessary because it allows us to quantify the impact of discounts on the total amount spent. If there’s a significant relationship between these two variables, we can use this information to make informed decisions about discount strategies to maximize sales.

- What are the independent and dependent variables in this analysis?
- What is the equation of the regression line?
- What are the values of the slope and intercept, and what do they represent in this context?

- How well does the regression model fit the data?

3) Multiple Linear Regression:

Your next task is to refine the linear regression model by incorporating "Product Price" as an explanatory variable. Please open the Regression Analysis Dataset.xlsx file and navigate to the Task 3 sheet.

Using the dataset provided, perform the following steps:

1. **Data Cleaning**:
    - Check for missing values in **Product Price**, **Discount**, and **Total Amount Spent**. Handle these missing values appropriately, as they can distort the regression analysis results and lead to inaccurate conclusions.
2. **Skewness Analysis**:
    - Compute the skewness values for **Product Price**, **Discount**, and T**otal Amount Spent**.
    - Create histograms for **Product Price**, **Discount**, and **Total Amount Spent** to visually inspect their distributions.
    - Consider a log transformation if the data deviates from normality.
3. **Regression Analysis (Before Transformation)**:
    - Perform a multiple linear regression analysis using the original variables.
    - Examine the residuals plot against each predictor variable (especially **Product Price**) to check for heteroscedasticity. To do that check the "Residuals" box in the "Output Options" section. If there's a systematic pattern or a funnel shape, it indicates heteroscedasticity.
4. **Regression Analysis (After Transformation)**:
    - Perform another multiple linear regression analysis using the transformed variables.
    - Examine the residuals plot to ensure the transformed model addresses any heteroscedasticity observed previously.

6. **Interpretation:**
    - Interpret the results of your analysis, focusing on:
        - The R-squared value of your model and its implications regarding the fit.
        - The coefficients of **Product Price** and **Discount**.
        - The significance of **Product Price** and **Discount** in predicting **Total Amount Spent**.

Questions for interpretation:

- Based on the skewness values, do any variables appear to have distributions that deviate from normality?
- Would a log transformation of any variable improve the linearity of the relationship?
- How does the R-squared value change (if at all) after transformation? Why?
- Are the coefficients interpretable in the context of the problem, especially after any transformations?

4) Advanced Multiple Linear Regression:

Now that you better understand how each variable relates to the **Total Amount Spent**, your next task is to use this knowledge to build a more advanced multiple linear regression model. Please open the Linear Regression Dataset.xlsx file and navigate to the Task 4 sheet.

In this task, you will utilize all available variables to predict the log-transformed **Total Amount Spent**. The variables you can use in this analysis include:

- Age
- Gender
- User Region
- Product Category
- Log-transformed Product Price
- Log-transformed Discount

Before you start, check for missing values and clean the data if necessary.

Once you have your clean dataset, perform the following steps:

-Convert categorical variables into dummy variables. Remember, when you create dummy variables, you should make one less than the number of unique categories to avoid multicollinearity.

-Use the VIF analysis to check for multicollinearity among the independent variables. The VIF for each independent variable is as follows:

- log_Product Price: 1.488
- log_Discount: 1.487
- Age: 1.000
- Gender_Male: 1.001
- User Region_North: 1.496
- User Region_South: 1.497
- User Region_West: 1.501
- Product Category_Electronics: 1.599
- Product Category_Clothing: 1.607
- Product Category_Home & Kitchen: 1.595
- Product Category_Sports & Outdoors: 1.608

-Evaluate the model by interpreting the coefficients' R-squared, F-statistic, and p-values.

Questions for interpretation:

- What is the R-squared of the model, and what does this tell you about the model?
- Which variables significantly predict the Total Amount Spent, and how do you know?
- What is the interpretation of the coefficients of the significant variables?

5) Predictive Modeling and Next Steps:

Now that you've built a multiple linear regression model using the training data, your next task is to use this model to make predictions on a testing dataset. The testing dataset will have the same structure as the training dataset but will represent new observations not included in the model-building process. To execute the following task, utilize the data presented in the Task 5 tab of the Regression Analysis Dataset.xlsx spreadsheet. The spreadsheet consists of a training dataset previously employed for regression estimation in Task 3, alongside a testing dataset available for evaluating your model's predictive performance. For model construction, both dependent and independent variables underwent log transformation.

Your task involves the following steps:

- **Predict Total Amount Spent**. Use the regression model to predict the log of **Total Amount Spent** for each observation in the log-transformed testing dataset.

- **Assess Model Performance**. Compare the predicted log of **Total Amount Spent** with the actual log of **Total Amount Spent** in the log-transformed testing dataset.

Questions for interpretation:

1. What are some critical insights gained from utilizing a scatter plot to analyze the variance between predicted and actual values?
2. Based on your results, does your model fit the data well? Why or why not? What improvements, if any, would you suggest for this model?