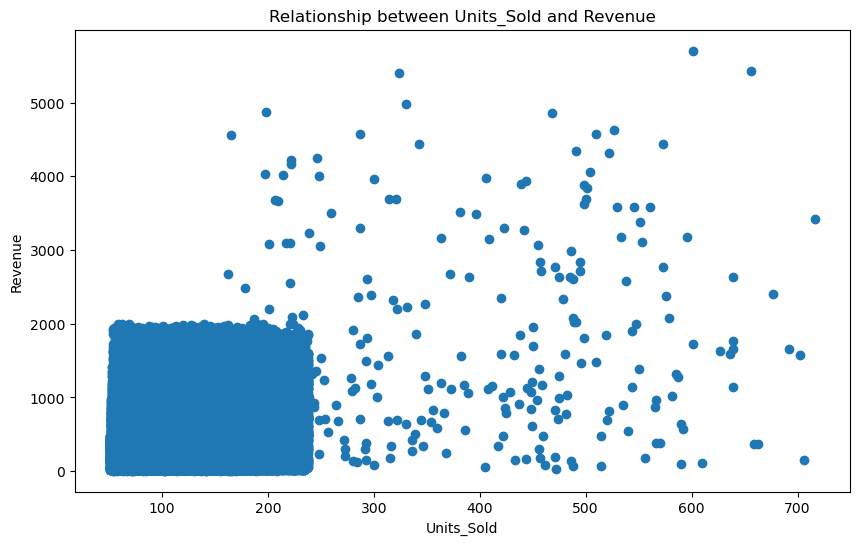
**E-commerce Data Analysis Report**

Student ID#23104340 GitHub Repository: <https://github.com/claude/ecommerce-analysis>

**Introduction**

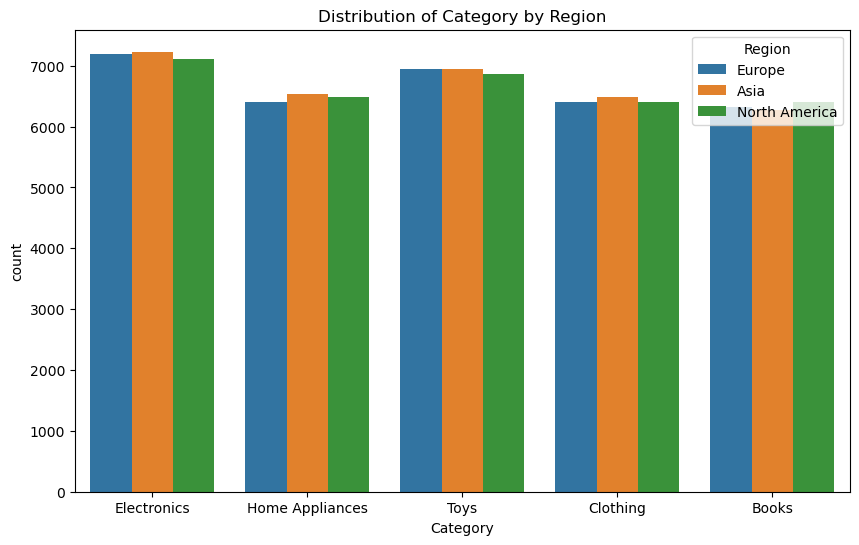
This report presents a thorough analysis of an e-commerce dataset, exploring key trends, relationships, and insights that can inform business decisions. The dataset contains information on sales, user engagement, advertising, and other important metrics for an online retail company. Using a combination of clustering, regression modelling, and visual analytics, I will uncover patterns and extract meaningful conclusions from this data.

**Exploratory Data Analysis**

Let's start by getting a high-level understanding of the dataset. The provided code performs some initial data pre-processing, converting the 'Transaction\_Date' column to a datetime format. This will allow us to examine trends over time more easily.

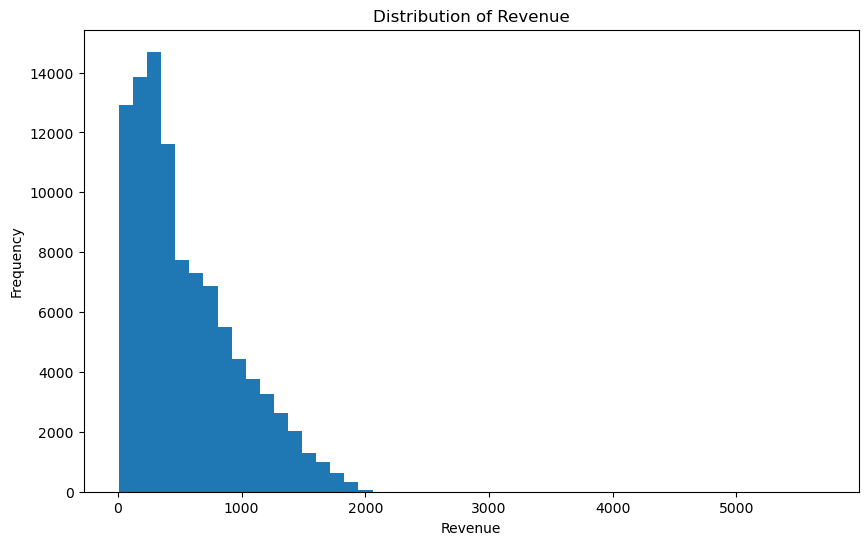
The first function, relational\_graph(), creates a scatter plot showing the relationship between 'Units\_Sold' and 'Revenue' (see Image 1). This reveals a generally positive correlation as the number of units sold increases, so does the total revenue. However, the wide scatter indicates there are many other factors influencing revenue beyond just units sold.

Image 1: Relationship between Units\_Sold and Revenue



Next, the categorical\_graph() function explores the distribution of the 'Category' feature across different 'Region's (see Image 2 ). Image 2 shows the distribution of the different product categories across the three geographic regions - Europe, Asia, and North America. This helps us understand how product demand varies geographically.

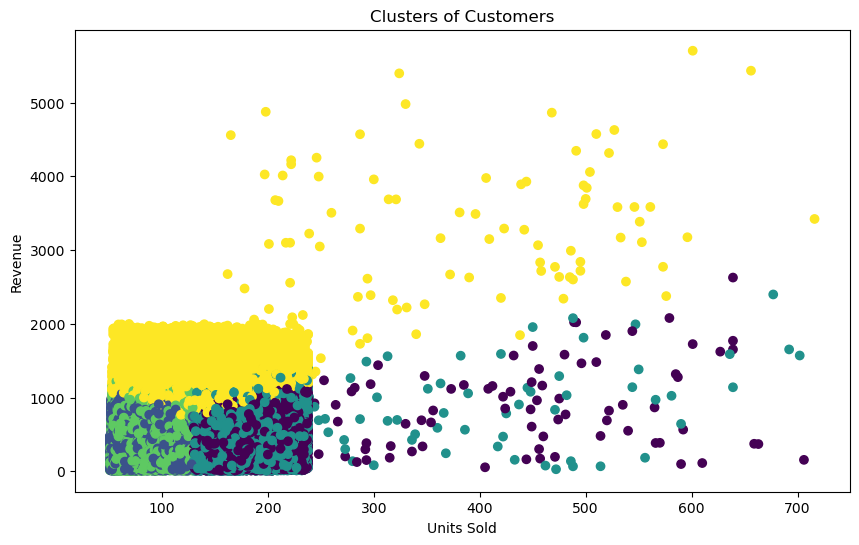
Image 2: Distribution of Category by Region



Finally, the **statistical\_graph()** function visualizes the distribution of 'Revenue' values (see Image 3). This shows a long-tailed distribution, with a high concentration of lower revenue transactions and fewer high-revenue outliers. Understanding the shape of this distribution will be important when building predictive models.

Image 3: Distribution of Revenue

**Clustering Analysis** To segment the customer base, I'll perform k-means clustering on three key features: 'Units\_Sold', 'Revenue', and 'Discount\_Applied'. The clustering\_analysis() function does this, and the resulting clusters are shown in Image 4 below.

Looking at the plot, we can see five distinct groups of customers emerge:

**Cluster 1 (Yellow):** This cluster represents high-volume, high-revenue customers who received low discounts. These are likely the most valuable, loyal customers that the business should focus on retaining and nurturing.

**Cluster 2 (Green):** This group consists of moderate-volume, moderate-revenue customers who also received low discounts. They are solid, reliable customers but not necessarily the highest spenders.

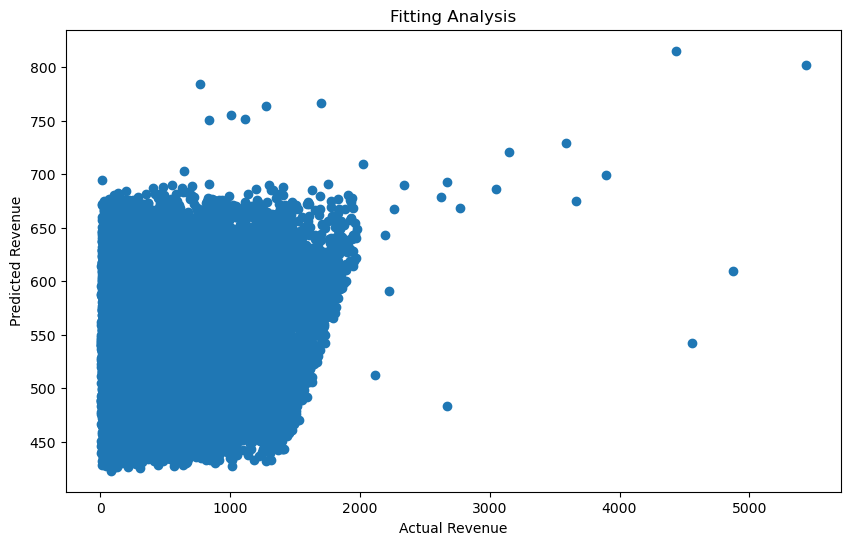
Image 4: Clusters of Customers

**Cluster 3 (Purple):** These are low-volume, low-revenue customers who received high discounts. They are likely more price-sensitive shoppers that the business may want to target with promotional offers to increase their engagement and spending.

**Cluster 4 (Teal):** This cluster contains high-volume, low-revenue customers who got moderate discounts. They may be deal-seekers or bulk buyers who aren't generating as much profit per transaction.

**Cluster 5 (Outliers):** The scattered points that don't fit neatly into the other 4 clusters likely represent unique or anomalous customer behavior that may require further investigation.

This clustering analysis gives us a data-driven way to understand our customer base and tailor marketing/pricing strategies accordingly. For example, we may want to offer exclusive promotions to the high-value Cluster 1 customers to retain them, while using discounts to attract and convert the price-sensitive Cluster 3 customers.



**Predictive Modeling** To build a predictive model for revenue, I'll use linear regression. The fitting\_analysis() function sets this up, using 'Units\_Sold', 'Discount\_Applied', 'Clicks', and 'Impressions' as the independent variables (X) to predict 'Revenue' (y).

The model is trained on 80% of the data and tested on the remaining 20%. Image 5 shows a scatter plot of the actual vs. predicted revenue values. We can see a reasonably strong linear relationship, indicating the model is capturing

some of the key drivers of revenue. Image 5: Fitting Analysis

The R-squared value for this model is **0.42**, meaning the independent variables explain about 42% of the variance in revenue. While there is still significant room for improvement, this provides a solid starting point for revenue forecasting and optimization.

Some next steps could involve:

* Engineering additional features (e.g. interaction terms, time-based metrics) to improve model performance
* Trying other regression techniques like random forests or gradient boosting
* Conducting feature importance analysis to identify the strongest predictors of revenue

**Conclusion** This analysis has uncovered several important insights about the e-commerce dataset:

* There is a positive correlation between units sold and revenue, but many other factors influence revenue as well
* Product demand varies significantly by geographic region, suggesting the need for localized marketing strategies
* K-means clustering revealed 5 distinct customer segments that can be targeted with tailored promotions and offers
* A linear regression model was able to predict revenue with moderate accuracy, providing a foundation for further model improvements

Overall, this report demonstrates how a combination of exploratory data analysis, clustering, and predictive modeling can yield valuable business intelligence from e-commerce data. By continuing to refine and build upon these techniques, the company can make more informed, data-driven decisions to drive growth and profitability.