**Central Michigan University, School of Business**

**Mount Pleasant, MI 48859**

**BIS581 Business Data Analytics – Final Project**

**Summer 2024**

**Cleaning, Visualizing, Modelling, and analyzing Using R**

**Professor: Vishal Shah**

**Student: Sai Kalyan Gurumanchi**

**Email: gurum1s@cmich.edu**

**Dataset:**

Data set was taken from Kaggle.

Dataset Link - <https://www.kaggle.com/datasets/pipchu/sales-2020-2021>

**The dataset consists of 286,392 entries and 36 columns. Here’s a brief overview of the key columns:**

* order\_id: Unique identifier for each order.
* order\_date: Date when the order was placed.
* status: Status of the order (e.g., received, complete).
* item\_id: Unique identifier for each item.
* sku: Stock-keeping unit, which is a unique identifier for each product.
* qty\_ordered: Quantity of items ordered.
* price: Price per item.
* value: Total value of the items ordered.
* discount\_amount: Discount applied to the order.
* total: Final total after applying the discount.
* category: Product category.
* payment\_method: Method of payment used.
* year: The year the order was placed.
* month: The month the order was placed.

Given that the dataset spans multiple years, filtered the data to focus only on sales data from 2021.

**Data Cleaning Issues**

**What issues did you see with your dataset in terms of data cleaning?**

* The dataset contained unnecessary columns that were not relevant for analysis, duplicate records, and inconsistent values in certain fields (e.g., payment types and order status). Additionally, the customer\_since field was not in the proper date format, and the SSN data needed to be masked for privacy.

**Before and after explanation in terms of data cleaning:**

* **Before:** The dataset had irrelevant columns, inconsistent values in categorical fields, duplicates, and improperly formatted date columns. Sensitive data like SSN was not masked, and the customer since field was not in a usable format.
* **After:** Unnecessary columns were removed, categorical values were categorized, duplicates were eliminated, and the customer\_since field was converted to the correct format. SSN data was masked to display only the last four digits for privacy.

**New Package Explanation**

**If you are using new packages, explain what they do. What is their general purpose?**

* **ggplot2** is a package used for data visualization. It allows the creation of complex plots from data in a data frame and is known for its flexibility and ease of use.
* **dplyr:** A package used for data manipulation. It provides a set of functions to perform common tasks such as filtering rows, selecting columns, and summarizing data.
* **options (scipen = 999):** This is not a package but an R setting to prevent scientific notation in numeric outputs, making numbers easier to read.

**Visualization Insights**

* **Age Distribution:** Customers are evenly distributed across various age groups, with a concentration in the 45–55 age range, suggesting targeted marketing could be effective.
* **Gender Distribution:** Male customers slightly outnumber female customers, but efforts should focus on appealing to both genders equally.
* **Monthly Sales:** April, March, and June had the highest sales, indicating these months have successful strategies that could be replicated in other months.
* **Category-wise Sales:** Over 50% of sales come from mobiles and tablets, entertainment, computing, and appliances, indicating these categories are key drivers of revenue.
* **Sales by Region:** The South region leads in sales, while the Northeast and West regions lag, suggesting a need for region-specific marketing strategies.
* **Sales by State:** TX, CA, and NY have the highest sales, with Houston, Detroit, and San Antonio being the top cities in Texas.
* **Customer Tenure:** Old customers (before 2000) generate more sales, indicating a need to focus on retaining and rewarding long-term customers.

**Predictions**

* This project uses logistic regression to predict whether an order will result in a **refund or cancellation** based on factors **like price, discount amount, category, payment type, customer age, and region.**

### **Interpretation:**

* **Low Prices:**

1. At lower prices (near 0), there is a wide spread of predicted probabilities, with many transactions having low to moderate chances of refunds or cancellations.
2. However, as the price increases from 0, the predicted probability also rises sharply.

* **High Prices:**

1. As the price continues to increase (around 25,000 and beyond), the predicted probability of a refund or cancellation approaches 1 (or 100%).
2. This suggests that higher prices are strongly associated with a higher likelihood of refund or cancellation.

* **Data Distribution:**

Most of the data points are concentrated on the left side of the graph (lower price range), with fewer points as the price increases, indicating that most transactions involve lower-priced items.

### **Potential Insights:**

* **Refund/Cancellation Sensitivity:**

There is a strong correlation between price and the likelihood of a refund or cancellation, with higher prices leading to a greater chance of these outcomes.

* **Model Fit:**

The logistic regression model seems to capture this trend well, as indicated by the smooth curve fitting the data.