# **Final Project Report**

# 

A Project Report submitted as a course requirement.

for

BUS895 – Research Projects in Business

Submitted by:   
Kiran Ranganalli



# 

## **Index**

1. Executive Summary ...................................................... 3
2. Introduction ................................................................... 4
3. Background .................................................................. 4
   * 3.1 Client Industry.............................................. 4
   * 3.2 Market Dynamics......................................... 4
   * 3.3 Supporting Research................................... 5
4. Problem Statement ....................................................... 6
   * 4.1 Business Question....................................... 6
   * 4.2 Importance..........……………………………..6
5. Methodology ................................................................. 7
   * 5.1 Data Preparation.......................................... 7
6. Analysis and Results .................................................... 10
   * 6.1 Exploratory Data Analysis .......................... 10
     + 6.1.1 Univariate Analysis……………..…10
     + 6.1.2 Bivariate Analysis……………...…13
   * 6.2 Analytical Approach……………………...….15
     + 6.2.1 Linear Regression (OLS)..............15
     + 6.2.2 Random Forest……………………22
   * 6.3 Model Summary……………………………..22
     + 6.3.1 OLS Model……………….…….…22
     + 6.3.2 Random Forest Model…………..23
   * 6.4 Key Findings ............................................ 24
     + 6.4.1 Business Insights…………...…...25
     + 6.4.2 Recommendations to Clients..…26
7. Discussion and Conclusion ....................................... 27
   * 7.1 Discussion…………………………………..27
   * 7.2 Conclusion……………….……………….…27
8. References ................................................................ 28

# **Analyzing Customer Behaviors and Characteristics to Identify High-Value Users**

## **1.Executive Summary**

Delivered Korea, a leading international e-commerce logistics provider, conducted a comprehensive analysis to uncover the behaviors and characteristics that define high-value customers. This report offers actionable insights derived from customer data, focusing on spending patterns, marketing engagement, and regional trends.[1] The findings reveal that email marketing plays a pivotal role in driving higher spending, as customers who opt into marketing emails consistently demonstrate increased purchase behavior.[2] Additionally, the analysis highlights a strong correlation between premium shipping preferences, such as FedEx IP (2~4 Days), and high-value spending, indicating a preference for speed and reliability among top-tier customers. Specific vendors, including BTS, IU, and BlackPink, also emerge as key drivers of customer loyalty and high-value purchases, presenting significant opportunities for exclusive collaborations and tailored promotions. Geographic analysis further identifies regions like Arizona and Northern Ireland as hotspots for high-value customers, providing a basis for targeted regional marketing strategies.[1]

The report also addresses key limitations, such as the moderate R-squared values observed in predictive models, suggesting the need for further exploration of customer demographics, seasonal trends, and additional predictors to enhance model accuracy. Moreover, it identifies underperforming vendors and regions, offering potential areas for improvement to broaden the customer base.[2] By leveraging these insights, Delivered Korea is well-positioned to enhance customer segmentation, optimize marketing strategies, and promote behaviors that increase customer lifetime value. These findings lay a strong foundation for sustainable revenue growth by aligning the company’s offerings with the preferences and characteristics of its most valuable customers.[1][2]

## **2.Introduction**

Delivered Korea operates in the cross-border e-commerce and logistics sector, facilitating international shipments of Korean goods. The company integrates with platforms like Shopify to ensure seamless global customer access to Korean products. With an extensive network of logistics partners and a commitment to delivering high-quality service, Delivered Korea has established itself as a reliable provider in the competitive e-commerce logistics industry.[1]

In the rapidly growing e-commerce market, understanding customer behavior is essential to maintaining a competitive edge. High-value customers often contribute disproportionately to overall revenue, making it critical to identify their key characteristics and preferences.[1] This report focuses on analyzing customer data to uncover insights that can guide marketing strategies, operational improvements, and customer retention initiatives. By leveraging data-driven methods, Delivered Korea aims to strengthen its position in the market and provide exceptional value to its customers.[1][2]

## **3.Background**

### **3.1 Client Industry**

Delivered Korea is a South Korean company specializing in cross-border e-commerce logistics, facilitating the global distribution of Korean products. The company aims to bridge Korean sellers with international buyers, ensuring seamless access to authentic Korean goods.[1]

The company offers a comprehensive suite of services, including parcel forwarding, e-commerce fulfillment, and dropshipping, tailored to meet the demands of the global market.Strategically headquartered in Busan, Delivered Korea operates a warehouse that serves as a central hub for domestic orders, which are then shipped directly to customers worldwide.The company has attracted investment from notable partners, including Beenos Partners, reflecting confidence in its business model and growth potential.[2][3] Through its official online store, DKshop, Delivered Korea caters to global K-pop fans and enthusiasts of Korean lifestyle products, further solidifying its position in the international e-commerce landscape.[1][3]

### **3.2 Market Dynamics**

Delivered Korea is a prominent player in the international logistics and e-commerce sector, specializing in cross-border shipping solutions that connect Korean products with consumers worldwide.[4] Leveraging a robust infrastructure and strategic partnerships, the company ensures efficient and reliable delivery services across multiple regions. By integrating advanced logistics technologies and maintaining a customer-centric approach, Delivered Korea effectively meets the growing global demand for Korean goods, positioning itself as a key facilitator in the international e-commerce landscape.[4]

The global e-commerce logistics market is experiencing significant growth, with projections indicating an increase from USD 441.55 billion in 2024 to approximately USD 1,903.08 billion by 2032, reflecting a compound annual growth rate (CAGR) of 20.04% over the forecast period.

This expansion is driven by rising consumer demand for international goods and advancements in logistics technologies. Delivered Korea is well-positioned to capitalize on these opportunities by leveraging its expertise and infrastructure to provide efficient cross-border shipping solutions.[4][5]

In the evolving e-commerce landscape, customers increasingly expect faster, reliable, and cost-effective delivery services. Research indicates that online shoppers prioritize faster delivery (48%), affordable shipping (43%), and accurate delivery time information (39%).

Delivered Korea's focus on premium shipping options and seamless integrations ensures it remains competitive in meeting these heightened customer expectations.[6]

The e-commerce logistics sector is highly competitive, with both global logistics giants and regional players striving for market share. To differentiate itself, Delivered Korea emphasizes customer-centric services, such as targeted marketing and optimized delivery networks, aligning with industry trends that prioritize speed, efficiency, and customization to meet modern shopper needs.[4][6]

By focusing on these areas, Delivered Korea enhances its competitive edge in the dynamic e-commerce logistics market.

### **3.3 Supporting Research**

Supporting research underscores the pivotal role that high-value customers play in driving revenue within the e-commerce industry, often aligning with the Pareto Principle, where 80% of revenue is generated by 20% of customers.[6] Retaining and understanding these high-value users is crucial for sustained profitability, as they exhibit greater engagement, higher purchase frequency, and stronger brand loyalty.[7] Studies emphasize the importance of leveraging data to identify and cater to this customer segment effectively.[8]

In recent years, the e-commerce and logistics sector has seen widespread adoption of advanced analytics, enabling businesses to perform customer segmentation and predictive modeling with greater precision.[9] This trend enhances decision-making and fosters growth, as businesses can tailor their strategies to align with customer preferences and behaviors. Personalized marketing strategies have become increasingly prominent, with research showing that a majority of consumers are more likely to make purchases when brands offer personalized experiences.[10] This approach not only improves customer satisfaction but also drives loyalty and repeat purchases.[11]

Additionally, there has been a growing emphasis on environmental sustainability in logistics, driven by consumer and regulatory demands. Companies are innovating to offer eco-friendly solutions such as carbon-neutral shipping and sustainable packaging, which align with global efforts to reduce emissions.[12] Furthermore, advancements in technology, particularly in artificial intelligence (AI) and machine learning, are transforming supply chain operations and improving delivery efficiency. These technologies enable businesses to optimize their logistics networks, reducing costs while enhancing customer satisfaction.[13]

Delivered Korea’s commitment to utilizing advanced analytics and staying abreast of these trends positions the company as a leader in the logistics industry.[6][12] By analyzing customer behaviors and spending patterns, Delivered Korea is able to provide actionable recommendations that enhance operational efficiency and maximize revenue streams, ensuring its continued success in a competitive market.[6][13]

**4.Problem Statement**

### **4.1 Business Question**

Delivered Korea seeks to understand customer behaviors and characteristics to effectively identify high-value users while distinguishing them from low-value users. This analysis aims to uncover actionable insights that can guide strategic decision-making, improve customer segmentation, and enhance operational efficiency. By examining spending patterns, marketing engagement, and shipping preferences, the company hopes to develop targeted strategies that maximize the value of its high-spending customers and explore opportunities to uplift low-spending ones.

### **4.2 Importance**

In the competitive e-commerce logistics market, high-value customers play a critical role in driving profitability and growth. Identifying and understanding the behaviors and characteristics of these users allows Delivered Korea to develop strategies to maximize their engagement and satisfaction. Additionally, analyzing the traits of lower-value customers enables the company to design targeted interventions to improve their value contribution.

This analysis empowers Delivered Korea to:

* **Refine Marketing Strategies**: Gain insights into the behaviors that drive high-value customer actions, enabling more precise and effective marketing efforts.
* **Enhance Customer Segmentation**: Improve segmentation to deliver tailored experiences and increase customer loyalty.
* **Optimize Operations**: Align resources to better serve high-value users, improving shipping, inventory, and customer service strategies.
* **Achieve Sustainable Growth**: Drive long-term profitability by focusing on actionable characteristics and behaviors that differentiate customer segments.

## **5.Methodology**

### **5.1 Data Preparation**

Delivered Korea's dataset is sourced from their Shopify platform, encompassing three primary tables: **Customers**, **Products**, and **Orders**. Each table contains distinct information critical for analyzing customer behaviors and spending patterns.

The **Customers table** comprises over 30,000 observations collected between January 2020 and August 2024. It includes seven key variables, such as Customer ID, Email, Country Code, Total Spent, Total Orders, and Marketing Acceptance indicators. This table provides foundational customer profile information essential for segmentation and targeted analysis.

The **Products table** features more than 4,000 observations from the same time period (January 2020 to August 2024). It contains ten variables, including details like Product Title, Vendor, Type, Product Price, and Compare-at Price. This dataset enables the analysis of product-level trends, vendor performance, and pricing impacts on customer spending behavior.

The **Orders table** captures over 4,000 orders made between January 2023 and June 2024. It includes 21 variables, such as Product Title, Vendor, Type, Product Price, and Compare-at Price. This table provides granular insights into order-level dynamics, such as frequency, order values, and shipping preferences.

This comprehensive dataset spans customer profiles, product details, and order records, creating a robust foundation for analyzing high-value customer characteristics and spending patterns.



**5.1.1 Data Collection**

Data was collected from Delivered Korea’s Shopify database, focusing on two primary datasets: **Customers**, **Orders and Products**. These datasets include a rich variety of features to enable comprehensive analysis:

**Customers Dataset**:

* + Profile information such as Total Spent,Shipping cost and Total orders.
  + Engagement metrics, including acceptance of marketing emails and SMS campaigns.
  + Geographic data such as country and province of residence.
  + There were around 30,000+ observations with seven variables

**Products Dataset**:

* + Includes product-level details, such as product titles, types, and vendor information.
  + Provides pricing information, including current and compare-at prices, to analyze discount effects.
  + Contains product tags and categories for grouping and segmentation.
  + There were around 4,000+ observations with Ten variables

**Orders Dataset**:

* + Detailed order-level information, including order value, line-item details (quantity, price), and shipping costs.
  + Time-based data, such as order placement and fulfillment timestamps, to calculate delivery speeds.
  + Geographic data, such as shipping destination (province and country).
  + There were around 4,000+ observations with Twenty one variables

**5.1.2 Data Cleaning**

**Duplicate Removal:**

* + Identified and removed 135 duplicate entries based on unique customer IDs and order numbers to maintain data integrity.

**Handling Missing Values:**

* + For categorical variables (e.g., Shipping Method, Accepts Marketing), imputed the most frequent value to preserve consistency.
  + For numerical variables (e.g., Total Spent, Line Item Quantity), filled missing values using the median to reduce the impact of extreme values.

**Outlier Detection and Handling:**

* + Used Interquartile Range (IQR) analysis to detect outliers in spending (Total Spent) and order frequency.
  + Capped outliers to the upper limit of Q3+1.5×IQR Q3 + 1.5 \times IQR Q3+1.5×IQR to minimize skewness without removing valuable data.

**Data Standardization:**

* + Standardized numerical columns, like Average Order Value for consistency in scale.
  + Normalized date formats (e.g., Paid At, Fulfilled At) to ensure compatibility for time-based analyses.

**Final Validation:**

* + Ensured no null values remained in critical columns (e.g., Total Spent, Shipping Cost).
  + Verified data consistency across merged datasets (e.g., matching email addresses in Customers and Orders datasets).

**5.1.3 Data Transformation**

**Encoding**:

* + Applied one-hot encoding to categorical variables such as Shipping Methods and Country Codes to enable their use in machine learning models.
  + Used label encoding for ordinal variables like Vendor Types for compact numeric representation.

**Scaling**:

* + Standardized numerical features like Total Spent, Order Quantities, and Shipping Costs to a standard scale (mean = 0, standard deviation = 1) for improved interpretability and model compatibility.

**Feature Engineering:**

* + Created binary indicators (0/1) for email and SMS marketing engagement to quantify customer responses.
  + Grouped shipping methods into broader categories, such as Express and Standard, for high-level analysis.

**Validation**:

* + Confirmed that transformed variables aligned with the original data structure and ensured no information loss during encoding or scaling

**5.2 Integration**:

* + Merged Two datasets (customer profiles and orders with reference to marketing metrics) into a unified dataset by matching unique identifiers such as customer emails.
  + Ensured no duplicate or inconsistent records post-merge.

## **6.Analysis and Results**

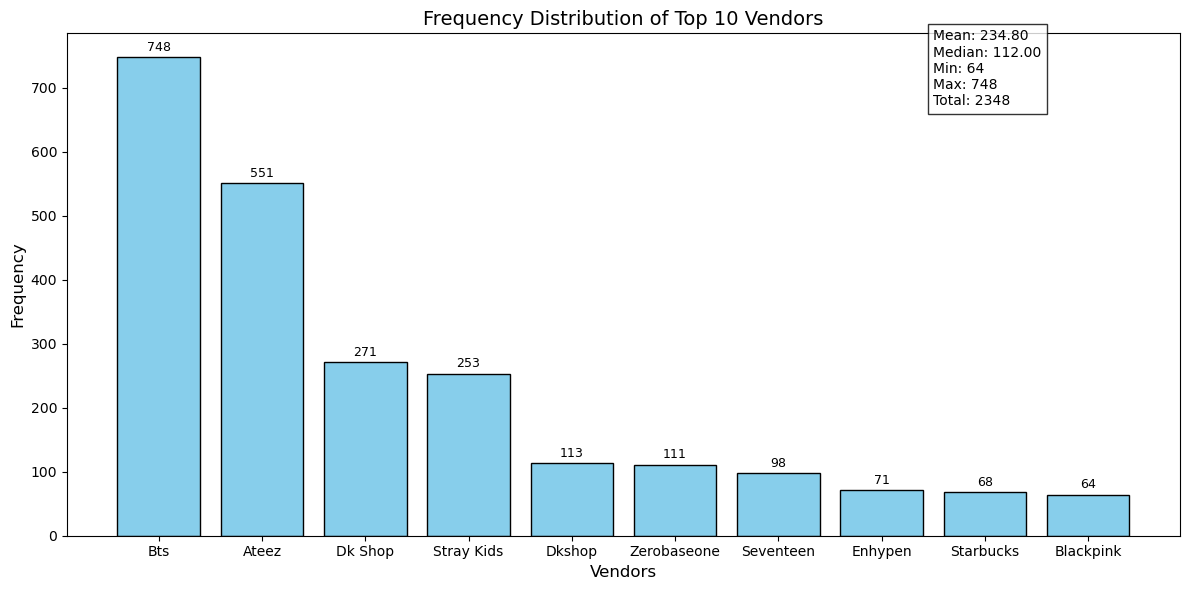
### **6.1 Exploratory Data Analysis**

The exploratory data analysis (EDA) was conducted in two stages to uncover insights into customer behaviors and characteristics:

#### **6.1.1 Univariate Analysis**

**Distribution of Vendor Frequency:**

The bar chart illustrates the frequency distribution of the top 10 vendors in the dataset. BTS stands out as the most frequent vendor, with 748 transactions, followed by Ateez with 551 transactions. These vendors show significant customer engagement, indicating their popularity and strong contribution to overall sales. Other notable vendors include DK Shop (271 transactions) and Stray Kids (253 transactions), which also play a substantial role. Lesser frequent vendors like Zerobaseone, Seventeen, and Enhypen show moderate levels of customer interaction, with frequencies ranging between 98 and 111. The vendors with the least frequency in this top 10 list, including Starbucks and Blackpink, each account for 64 to 68 transactions. The descriptive statistics reveal an average frequency of 234.8 transactions among these top vendors, with a median of 112 transactions, a minimum of 64, and a maximum of 748, collectively contributing to a total frequency of 2,348 transactions.

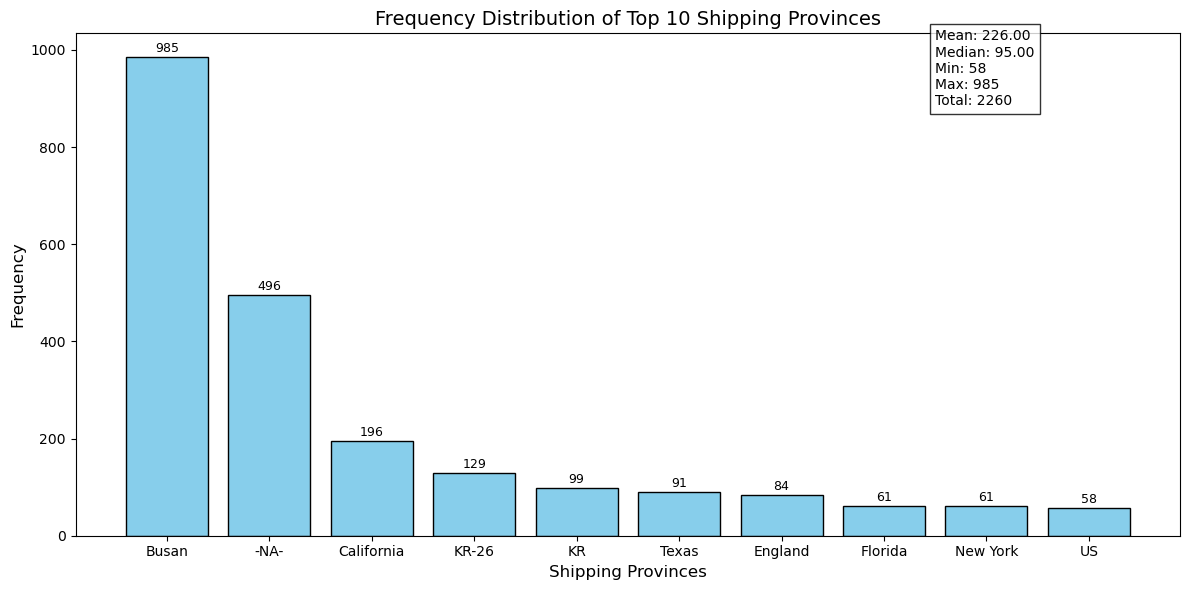


**Distribution of Shipping Provinces:**

The bar chart illustrates the frequency distribution of the top 10 shipping provinces used by customers, highlighting Busan as the most frequently recorded province with 985 occurrences. The next significant contributor is '-NA-', representing regions without specific province information, with 496 entries. California ranks third, with 196 occurrences, indicating its importance as a key shipping destination.

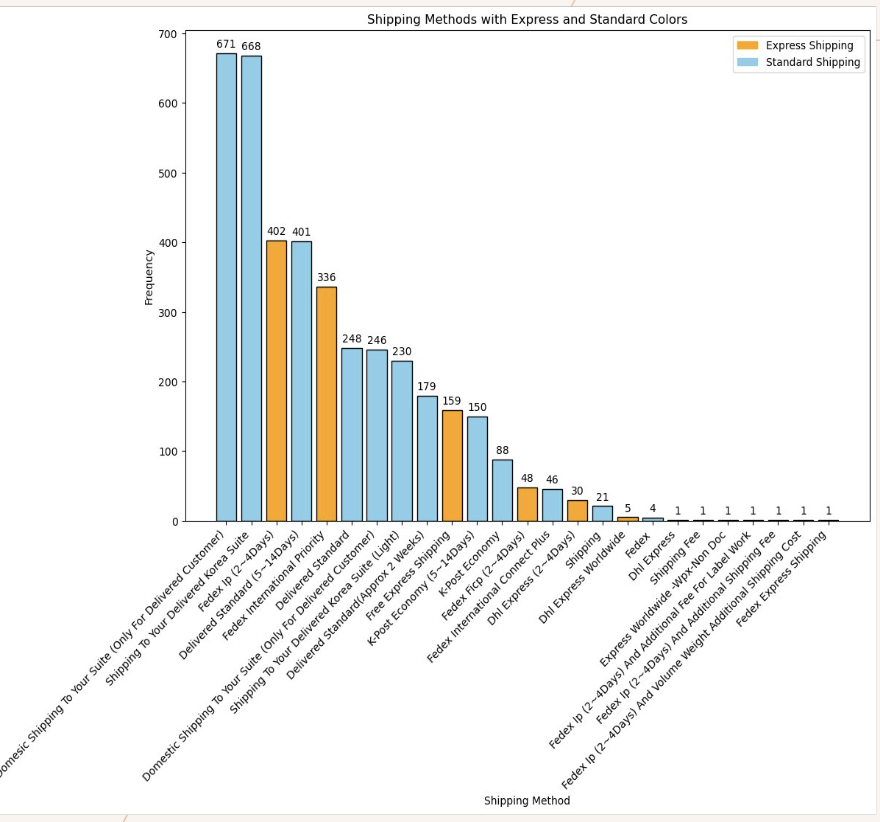
KR-26 follows with 129 entries, showcasing its relevance as a shipping destination in Korea. Other notable provinces include KR (99), Texas (91), England (84), Florida (61), New York (61), and the US (58). The data emphasizes the geographic diversity of Delivered Korea's customer base, reflecting the significance of both domestic and international shipping regions.

The summary statistics further contextualize the data, with a mean frequency of 226.00, a median of 95.00, a minimum frequency of 58, and a maximum of 985. The total number of shipping entries across the top 10 provinces sums to 2,260. These insights underscore the importance of focusing marketing and operational strategies on these high-frequency regions to optimize customer satisfaction and enhance revenue potential.



**Shipping Methods for Express and Standard**:

The chart highlights customer preferences for shipping methods, with standard options like "Domestic Shipping To Your Suite (671)" and "Shipping To Your Delivered Korea Suite (668)" being the most popular. Express methods such as "FedEx IP (2~4 Days)" (402) and "FedEx International Priority" (336) also show significant usage, indicating demand for faster delivery. Standard methods dominate overall, but express options are preferred by high-value customers, showcasing the need for diverse shipping solutions to cater to varying preferences.



#### **6.1.2 Bivariate Analysis**

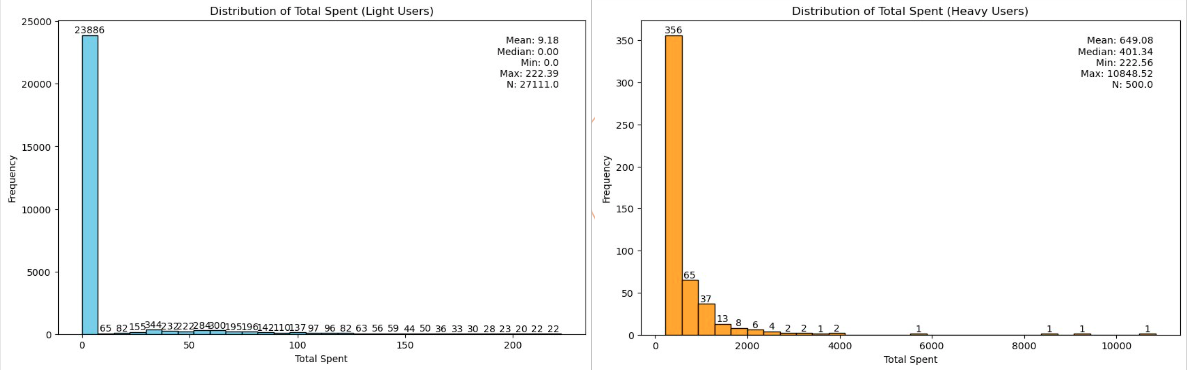
**Heavy vs Light Users**:

Heavy and light users were classified based on their total spending behavior. Heavy users are defined as those whose total spending exceeds $222, while light users are those with spending below this threshold. This cutoff was selected after analyzing the spending distribution, where $222 corresponds to the upper quartile, representing a clear distinction between the majority of users and the top spenders. Heavy users account for the top 25% of total spenders, indicating their significance in revenue contribution.

In the dataset, **98.18% of users (27,111 out of 27,611)** were classified as light users, contributing to **11.8% of total spending**, while **1.82% of users (500 out of 27,611)** were classified as heavy users, accounting for **88.2% of total spending**. This analysis highlights that while light users dominate in terms of user count, heavy users contribute disproportionately to the company's revenue, making them the focal point for targeted strategies.

The decision to focus on heavy users stems from their disproportionate impact on profitability. Heavy users tend to exhibit behaviors such as higher engagement with email marketing campaigns, strong brand loyalty to specific vendors, and a preference for premium shipping options. These characteristics make them highly valuable for strategic marketing initiatives and personalized retention efforts. By focusing on heavy users, Delivered Korea can maximize the return on investment for its marketing and operational resources.

Although the primary focus is on heavy users, light users offer potential for growth if effectively targeted. Future analyses could explore strategies for engaging light users, identifying factors that might encourage higher spending, and converting them into heavy users. Understanding their behaviors and preferences could help expand their lifetime value, providing a balanced approach to customer segmentation and retention strategies.



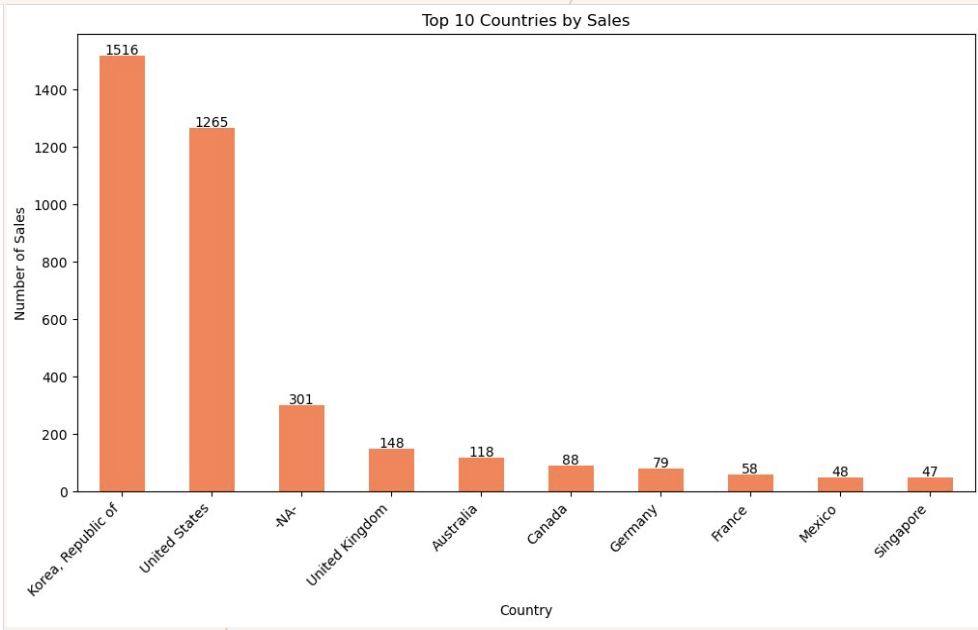
**Number of Orders by Marketing Acceptance**:

1. Customers who accepted marketing communications placed significantly more orders than those who opted out, underscoring the impact of marketing engagement on customer behavior.



**Top 10 Countries by Sales**:

* This analysis highlights the countries contributing the most to Delivered Korea’s total sales. The data reveals strategic opportunities for expansion and targeted marketing in regions with high sales volumes.



**6.2 Analytical Approach Model Selection**:

**6.2.1 Linear Regression (OLS):**

Used for interpretability and identifying statistically significant predictors of customer spending.

**6.2.2 Random Forest:**

Applied for feature importance analysis to understand variable contribution.

**Model Training:**

Split data into training (80%) and testing (20%) sets. Used cross-validation to ensure robust model performance.

**Evaluation Metrics:**

Evaluated models using R-squared, mean squared error (MSE), and feature importance rankings. Assessed p-values in OLS regression to determine statistical significance of predictors.

**Segmentation Analysis:**

Classified customers into high-value and low-value groups based on spending thresholds. Performed comparative analysis of customer characteristics across these groups.

### **Key Findings**

**Regression Analysis (OLS)**:

#### **Objective**

The primary objective of the OLS model is to:

* Understand which factors significantly influence customer spending (Total Spent).
* Quantify the relationship between independent variables (predictors) and the dependent variable (Total Spent).
* Provide actionable business insights to improve marketing, shipping, and vendor strategies.

#### **OLS Model Setup**

Y-Variable

The model uses **Total Spent** as the dependent(Y) variable and the following predictors:

X-Variables

1. **Email\_Marketing\_Binary (0/1):**
   1. **0**: The customer has not opted to receive marketing emails.
   2. **1**: The customer has opted to receive marketing emails.
2. **Shipping Method** (Categorical: One-hot encoded)
3. **Shipping Province Name** (Categorical: One-hot encoded)
4. **Vendor** (Categorical: One-hot encoded)
5. **Discount Amount** (Numeric)

These predictors were chosen based on their relevance to customer spending behavior.

To improve the robustness of the model, we applied a log transformation to the dependent variable, Total Spent. This decision was motivated by the highly skewed distribution of the data and the presence of outliers. The raw spending data exhibited a right skew, with a majority of customers having low spending amounts and a small proportion of heavy spenders creating extreme values. These outliers could disproportionately influence the results, leading to biased coefficients and an inaccurate representation of the predictors' effects.

By applying a log transformation (Log\_Total Spent = log(1 + Total Spent)), we stabilized the variance, reduced the influence of outliers, and achieved a more normal-like distribution for the dependent variable. This transformation ensures that the model captures the relationships between predictors and spending behavior more accurately and enables meaningful interpretation of the coefficients in percentage terms.

**Categorical Variables Converted to Dummy Variables**

When categorical variables are one-hot encoded, each unique category is transformed into a binary (0/1) variable. In this process, one category is selected as the **reference category**, and the coefficients of the other categories are interpreted relative to this reference. The reference categories for each variable in the model are:

1. **Shipping Method (One-hot encoded):**
   * The reference category is the **most frequent or default shipping method**, such as *Standard Shipping*. Other shipping methods (e.g., *FedEx IP (2~4 Days)*, *Domestic Shipping*) are compared to this.
2. **Shipping Province Name (One-hot encoded):**
   * The reference category is a province not explicitly encoded in the dummy variables, often a baseline or less significant location.
3. **Vendor (One-hot encoded):**
   * The reference category is the vendor with the highest occurrence or default grouping, such as *Other*. Vendors like *BTS*, *Stray Kids*, or *IU* are compared to this reference.

### **Importance of Reference Categories**

The reference category plays a crucial role in interpreting the results:

Coefficients of dummy variables represent the effect of that category relative to the reference.

### **Reason for Dummy Variable Transformation**

Dummy variables allow categorical variables to be used in regression models by encoding them numerically. This ensures the model interprets the relationship between categories correctly without assuming an inherent order, as with numerical data. Reference categories are essential to avoid multicollinearity, where variables are too highly correlated, making it difficult for the model to estimate coefficients accurately.

#### **6.3 Model Summary**

**6.3.1 Here is the summary of the OLS model:**

A screenshot of a computer screen

Description automatically generated

#### **Key Metrics**

* **R-squared (0.313)**:
  + 31.3% of the variability in Total Spent is explained by the predictors in the model.
  + This indicates a moderate fit, with room for improvement by incorporating additional predictors or interactions.
* **F-statistic (12.08, p-value < 0.001)**:
  + The model is statistically significant overall, meaning at least one predictor significantly impacts Total Spent.

#### **Coefficients and Interpretations**

* **Email\_Marketing\_Binary**:
  + **Coefficient = 0.5622**
  + Customers engaged in email marketing spend approximately **56.22% more**, on average, compared to those who are not. This highlights the importance of targeted email marketing campaigns in driving higher spending.
* **Vendor\_Bts**:
  + **Coefficient = 0.2878**
  + Customers purchasing from **BTS** spend **28.78% more** than those purchasing from the baseline vendor, **Queen Of Tears**. This indicates a strong customer preference for BTS products.
* **Vendor\_Iu**:
  + **Coefficient = 2.382**
  + Customers purchasing from **IU** spend a significant **138.2% more** than those purchasing from the baseline vendor. This suggests that IU is highly associated with high-value purchases.
* **Shipping Method (Fedex Ip)**:
  + **Coefficient = 0.2614**
  + Customers opting for **FedEx IP (2~4 days)** spend **26.14% more**, on average, compared to customers using the baseline shipping method. This indicates that customers preferring faster shipping options are likely to be high spenders.
* **Shipping Province (Arizona)**:
  + **Coefficient = 3.1688**
  + Customers from **Arizona** spend an impressive **216.88% more** than those from the baseline shipping region, **r8vLZvOfws1l0xPuEIvH3ldcH**. This demonstrates Arizona's prominence as a key region for high-value customers.

**Additional Insights**

#### **Baseline Vendor and Shipping Region:**

* **Baseline Vendor**: Queen Of Tears
* **Baseline Shipping Region**: r8vLZvOfws1l0xPuEIvH3ldcH

#### **Bottom 5 Vendors by Average Total Spent:**

1. R9Gluuc193Oieuaguevuzycf7: $0.00
2. Ryxwww7Ov5P9Utomdupp6Krqi: $0.00
3. Nmixx: $4.26
4. '106: $15.00
5. Fromis\_9: $25.90

#### **Bottom 5 Shipping Regions by Average Total Spent:**

1. FL: $0.00
2. rBX15Sgxp7ax0Q48knpQkJcfO: $1.22
3. rdLXY5NQaGcOq51tNZnNNYD72: $6.74
4. Medford Lakes: $8.90
5. Paran: $12.92

#### **Business Insights**

1. **Email Marketing's Role**:
   * Customers engaged in email marketing consistently demonstrate higher spending. This highlights the importance of **personalized email campaigns** to drive customer engagement and spending.
2. **Shipping Method Influence**:
   * Premium shipping options like **FedEx IP** significantly impact spending. Customers using these methods tend to be high-value users, indicating a strong preference for speed and reliability in shipping.
3. **Vendor Impact**:
   * Vendors such as **BTS**, **BlackPink**, and **IU** are highly associated with higher spending. These brands hold strong customer loyalty, presenting opportunities for **exclusive collaborations** and **tailored promotions** to attract and retain high-value customers.
4. **Regional Spending Trends**:
   * Regions like **Arizona**, **California**, and **Northern Ireland** are consistently linked to higher spending, making them ideal targets for **region-specific marketing strategies** to maximize revenue.

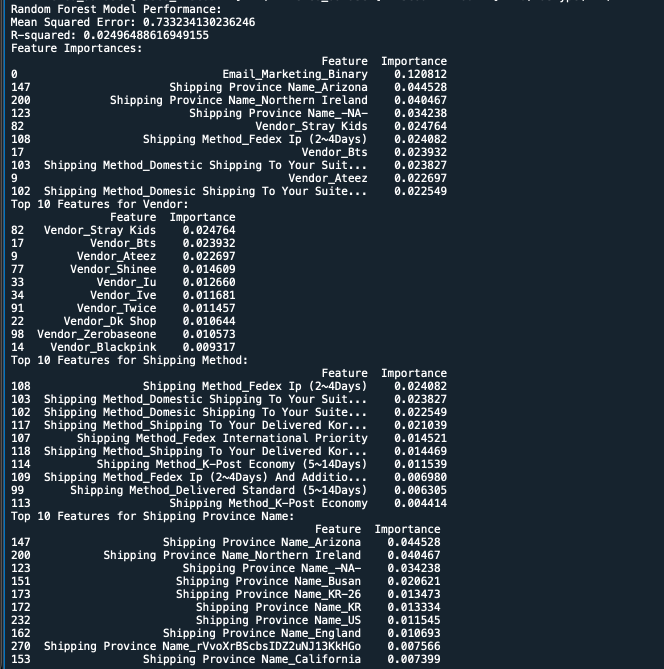
### **Actionable Opportunities for Optimization**

1. **Cross-Selling with High-Impact Vendors**: Collaborate with high-performing vendors like BTS and IU to bundle products or create promotional campaigns that resonate with high-value customers. Offering exclusive items can further boost spending.
2. **Promotional Focus on Premium Shipping**: Highlighting the benefits of premium shipping options such as faster delivery times or enhanced reliability in targeted email campaigns may encourage more customers to select these methods.
3. **Region-Specific Offers**: Tailor promotional offers based on geographic insights. For instance, exclusive promotions in Arizona and Northern Ireland could further leverage the higher spending observed in these areas.
4. **Incentivizing High-Value Behaviors**: Design loyalty programs or incentives for customers who repeatedly engage in high-value behaviors, such as opting into email marketing or choosing premium shipping.

#### **Limitations and Future Directions**

1. **Moderate R-squared**:
   * While the model explains **31.3%** of variability, exploring additional predictors such as customer demographics or seasonal trends could enhance the model's predictive power.
2. **Interpreting Vendor Contributions**:
   * Some vendors with higher coefficients might have fewer observations, which may skew results. Investigate vendor performance in more depth to validate findings.
3. **Geographic Expansion**:
   * Analyze underperforming regions to identify opportunities for growth or investigate barriers that may hinder spending.

**6.3.2 Here is the summary of Random Forest Model:**



#### **Objective**

The primary objectives of the Random Forest model are to:

* Identify the most influential factors driving customer spending (Total Spent).
* Rank features based on their importance to spending behavior.
* Provide actionable insights to improve marketing, shipping, and vendor strategies.

#### **Random Forest Model Setup**

The model uses the following setup:

* **Dependent Variable (Target)**:**X - Variables**
  + Total Spent (log-transformed to stabilize variance and reduce the impact of outliers).
* **Independent Variables (Predictors)**: **Y-Variables**
  + **Email\_Marketing\_Binary**: Whether the customer opted for email marketing (Binary: 0/1).
  + **Shipping Method**: Categorical variable (One-hot encoded to represent each shipping type).
  + **Shipping Province Name**: Categorical variable (One-hot encoded to capture geographic influence).
  + **Vendor**: Categorical variable (One-hot encoded to represent each vendor).
  + **Discount Amount**: Binary feature (Used\_Discount: 1 if a discount was applied, 0 otherwise).
  + **Interaction Term**: Email marketing and discount interaction (Email\_Marketing\_Discount).

These predictors were chosen for their potential to capture key customer behaviors and spending drivers.

#### **Evaluation Metrics**

* **Mean Squared Error (MSE)**: 0.733.
  + Measures the average squared difference between predicted and actual log-transformed Total Spent. While lower is better, this value reflects the inherent complexity of customer behavior.
* **R-squared**: 0.025.
  + The model explains **2.5% of the variance** in Total Spent. Though low, the feature importance rankings still offer actionable insights.

#### **Key Insights from Feature Importance**

#### **Overall Top Features:**

#### **Email\_Marketing\_Binary:**

#### The most influential feature, emphasizing that email marketing opt-ins are strong indicators of higher spending.

#### **Shipping Province Name (Arizona and Northern Ireland):**

#### Key geographic drivers of spending, with Arizona having the highest impact among provinces.

#### **Vendor\_Stray Kids and Fedex Ip (2~4 Days):**

#### Preferences for specific vendors and premium shipping options correlate with high spending.

#### 

#### **Model Performance:**

1. **Mean Squared Error (MSE)**:
   * **0.733**: This value measures the average squared difference between the predicted and actual log-transformed Total Spent. While a lower value is preferred, it’s reasonable given the complexity of customer behavior.
2. **R-squared**:
   * **0.025**: The model explains only **2.5% of the variance** in the target variable (Total Spent). This is quite low and indicates that the predictors may not fully capture the patterns driving customer spending.
   * **Possible reasons for low R-squared**:
     + The relationship between predictors and the target might be highly non-linear or influenced by other variables not included in the model.
     + The log transformation of the target variable may have reduced variability, making it harder for the model to fit.

#### **Top Features Overall:**

1. **Email\_Marketing\_Binary**:
   * **Importance: 0.1208**: This feature has the highest importance, highlighting that customers opting into email marketing are key drivers of spending behavior.
2. **Shipping Province Name\_Arizona**:
   * **Importance: 0.0445**: Customers in Arizona are strongly associated with higher spending.
3. **Shipping Province Name\_Northern Ireland**:
   * **Importance: 0.0405**: Another highly influential region for high-value customers.
4. **Vendor\_Stray Kids** and **Shipping Method\_Fedex Ip (2~4 Days)**:
   * Both are relatively influential, indicating preferences for specific vendors and premium shipping methods contribute to spending.

#### **Top Features by Category:**

1. **Vendor**:
   * **Vendor\_Stray Kids**: Importance **0.0247**, showing it’s the most impactful vendor for spending.
   * Other significant vendors: **Bts,Blackpink Ateez, Shinee**, and **Iu**.
2. **Shipping Method**:
   * **Shipping Method\_Fedex Ip (2~4 Days)**: Importance **0.0241**, suggesting that customers using faster shipping methods tend to spend more.
   * Other significant methods: Domestic Shipping To Your Suite and Domestic Shipping To Your Suite.
3. **Shipping Province Name**:
   * **Shipping Province Name\_Arizona,California** and **Northern Ireland** are the most impactful provinces.
   * Other important locations: Busan, KR, and US.

#### 

#### **Interpretation:**

* **Email Marketing** is a clear driver of high spending. Customers who opt-in for marketing communications are more likely to spend significantly.
* **Geographical Segments**:
  + Certain regions (e.g., **Arizona, Northern Ireland, Busan,California**) are consistently associated with higher spending.
* **Vendor Preferences**:
  + Vendors such as **Stray Kids, Bts**, and **Ateez** attract high-value customers, indicating brand affinity among these shoppers.
* **Shipping Preferences**:
  + Customers opting for faster or more convenient shipping methods like **Fedex Ip (2~4 Days)** are likely to be high spenders.

### **6.4 Key Findings**

#### **6.4.1 Business Insights**

* **Email Marketing**:
  + Both OLS and Random Forest models emphasize that customers engaged in email marketing spend significantly more. Email marketing is a clear driver of high spending and an effective tool for targeting high-value users.
* **Vendor Preferences**:
  + Vendors such as **BTS, IU**, and **Stray Kids** consistently appear as strong predictors of high spending. These vendors attract high-value customers and showcase brand loyalty among their followers.

#### **Shipping Behavior**

* **Premium Shipping Methods**:
  + Customers opting for faster and premium shipping options, such as **FedEx IP (2~4 Days)**, exhibit higher spending habits. This aligns with the characteristics of high-value users who prioritize convenience.
* **Interaction Between Marketing and Discounts**:
  + The interaction between email marketing and discounts suggests customers who are targeted via email campaigns and provided discounts tend to spend more, highlighting the synergy of these two strategies.

#### **Geographical Trends**

* **High-Spending Regions**:
  + Arizona and Northern Ireland stand out as key regions associated with higher spending. These areas offer opportunities for targeted, localized marketing campaigns.
* **Underperforming Regions**:
  + While some regions drive high spending, others exhibit lower contributions. Identifying barriers in underperforming areas could unlock untapped potential.

### **7.Discussion and Conclusion**

### **7.1 Discussion**

This analysis revealed critical factors influencing customer behavior, focusing on the characteristics that define high-value users. High-value customers exhibited significant engagement with email marketing campaigns, preference for premium shipping methods like FedEx, and a strong association with specific vendors and regions such as BTS and California. Conversely, low-value customers showed limited engagement and opted for standard shipping options. These insights underline the importance of tailoring strategies to cater to high-value customers while developing interventions to enhance the value of low-value groups.

The study also demonstrated the effectiveness of data-driven segmentation and regression techniques in predicting customer spending behavior. While the OLS model provided interpretability, Random Forest highlighted the relative importance of various features.

## **7.2 Conclusion**

### **Recommendations for the Client**

#### **1. Optimize Email Marketing Campaigns**

* Focus on **personalized and targeted email campaigns** to engage customers effectively.
* Integrate discounts or exclusive offers into these campaigns to maximize their impact on spending.

#### **2. Strengthen Vendor Partnerships**

* Build strategic collaborations with top-performing vendors like **BTS**, **IU**, and **Stray Kids**.
* Offer **exclusive promotions and product bundles** for these vendors to leverage customer loyalty and drive revenue.

#### **3. Promote Premium Shipping Options**

* Highlight **premium shipping methods** like FedEx IP in marketing campaigns to attract high-value users who prioritize faster delivery.
* Offer incentives, such as free upgrades to premium shipping for high-spending customers.

#### **4. Localized Marketing for High-Spending Regions**

* Invest in **region-specific campaigns** in areas like **Arizona** and **Northern Ireland** to further engage high-value users.
* Consider promotions tailored to these regions, such as shipping discounts or exclusive vendor partnerships.

#### **5. Explore Underperforming Regions**

* Analyze **low-spending regions** to identify barriers or potential improvements. Address factors like limited product availability or shipping constraints to boost spending.

#### **6. Enhance Data Collection for Better Predictive Models**

* Include **additional predictors**, such as customer demographics, product categories, and seasonal trends, to improve the models’ ability to explain spending behaviors.

Key Achievements:

* Identified actionable predictors for customer value segmentation.
* Highlighted the importance of marketing, shipping, and regional factors.
* Developed targeted recommendations to enhance revenue and operational efficiency.

Future Recommendations:

* Strengthen personalized marketing campaigns to maximize engagement.
* Optimize vendor partnerships and premium shipping options.
* Explore additional predictors like product categories and seasonal trends for further analysis.

## **8.References**

* [1] Delivered Korea. "Delivered Korea – E-Commerce and Logistics Overview." Delivered Korea, https://renewal.delivered.co.kr. Accessed 13 Dec. 2024.
* [2] Invest Issue. "Understanding E-Commerce Customer Behavior Analysis." Invest Issue,<https://investissue.com/ecommerce-customer-behavior-analysis/>. Accessed 13 Dec. 2024.
* [3] Delivered Korea Company Profile. "Delivered Korea Company Profile 2024: Valuation, Funding & Investors." PitchBook,<https://pitchbook.com/profiles/company/535661-74>. Accessed 13 Dec. 2024.
* [4] Fortune Business Insights. "E-Commerce Logistics Market Size, Share & Industry Analysis." Fortune Business Insights, https://www.fortunebusinessinsights.com/e-commerce-logistics-market-107945. Accessed 13 Dec. 2024.
* [5] ECommerceDB. "E-Commerce Logistics Report 2023: Global Market Size and Consumer Preferences." ECommerceDB,<https://ecommercedb.com/insights/ecommerce-logistics-2023-report-global-market-size-analysis-consumer-preferences/>. Accessed 13 Dec. 2024.
* [6]"The 80/20 Rule and Customer Lifetime Value." *Think with Google*,<https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/80-20-rule-and-customer-lifetime-value/>. Accessed 13 Dec. 2024.
* [7]"Why Focus on the Top 20% of Your Customers? How to Leverage the Pareto Principle for Maximum Profitability." *Reviews.io*,<https://blog.reviews.io/post/why-focus-on-the-top-20-of-your-customers-how-to-leverage-the-pareto-principle-for-maximum-profitability>. Accessed 13 Dec. 2024.
* [8]"80/20 Rule & RFM Segmentation for Ecommerce." *Unific*,<https://www.unific.com/80-20-rule-ecommerce>. Accessed 13 Dec. 2024.
* [9]"A Comprehensive Guide to the RFM Model." *Omniconvert*,<https://www.omniconvert.com/blog/rfm-model/>. Accessed 13 Dec. 2024.
* [10]"New Epsilon Research Indicates 80% of Consumers Are More Likely to Make a Purchase When Brands Offer Personalized Experiences." *Epsilon*,<https://www.epsilon.com/us/about-us/pressroom/new-epsilon-research-indicates-80-of-consumers-are-more-likely-to-make-a-purchase-when-brands-offer-personalized-experiences>. Accessed 13 Dec. 2024.
* [11]Chhabria, Riddhi. "A Study on the Impact of 'Personalized Marketing' on Customer Buying Behavior." *International Journal of Computer Applications*, vol. 185, no. 11, 2023, pp. 1-5.<https://www.ijcaonline.org/archives/volume185/number11/chhabria-2023-ijca-922707.pdf>. Accessed 13 Dec. 2024.
* [12]"E-Commerce Logistics Sustainability Trends." *Fortune Business Insights*, https://www.fortunebusinessinsights.com/e-commerce-logistics-market-107945. Accessed 13 Dec. 2024.
* [13]"AI-Enhanced Big Data Analytics for Personalized E-Commerce Recommendations." *International Journal of Advanced Engineering and Innovative Technology*, vol. 3, no. 4, 2024, pp. 45-52.<https://ijaeti.com/index.php/Journal/article/download/697/725/1288>. Accessed 13 Dec. 2024.
* Delivered Korea Online website <https://www.delivered.co.kr/en>
* Delivered Korea Shopify Datasets.
* Regression and Feature Importance Models.