Part 1: Proposal

Introduction

1.1 Overview of the Research Question

Driven by the continuous development of sophisticated technology and the impact of the Covid-19 pandemic, traditional businesses are increasingly shifting towards the digital economy. E-commerce, a significant component of this digital economy, has seen its share of global retail sales increase by 8% since the pandemic, according to the International Trade Administration (International Trade Administration, n.d). Within this highly competitive industry, customer satisfaction has become a crucial factor for businesses striving to retain their customer base and develop loyalty. The most critical component influencing customer satisfaction is logistics performance. Therefore, the research question revolves around the aim to explore: How does delivery performance affect customer satisfaction in the e-commerce sector?

1.2 Motivation for the Research

This research question is important because timely and reliable delivery is a fundamental expectation in e-commerce. According to Ahn et al. (2004), customers expect their purchases to arrive within the promised timeframe and the companies that meet or exceed these expectations are often associated with higher levels of customer satisfaction. Alternatively, delays and unreliable deliveries can lead to customer frustration, negative reviews, and loss of business.

Existing literature highlights the importance of delivery performance in shaping customer satisfaction. However, there are gaps in understanding how this relationship varies across different regions and product types. Geographic factors such as distance from the distribution center and the efficiency of regional delivery services can affect delivery speed and reliability. Moreover, customer expectations for delivery speed may differ based on product categories.

This research aims to fill these gaps by studying the relationship between delivery performance and customer satisfaction. By doing so, it aims to provide actionable insights for e-commerce businesses to improve their logistics operations and enhance customer satisfaction.

Literature Review

2.1 E-commerce and Customer Satisfaction

While most businesses in this century have shifted from a product-centric or sales-centric approach to focus on a customer-centric approach and have become an important factor for success. As mentioned by Oliver (1980), customer satisfaction is defined as the measure of how well a product or service provided by a business meets or exceeds customer expectations. From an e-commerce perspective, it is a critical metric as satisfied customers are more likely to return for future purchases, leave positive reviews, and recommend the business to others, thus driving loyalty and growth (Lee, Choi, & Kang, 2009; Tsai & Huang, 2007).

Common measurements of customer satisfaction include customer satisfaction score, customer effort score and more. However, Liu et al. (2021) proposed that analyzing customer reviews and ratings of a purchased good is a more direct and comprehensive way to obtain customer satisfaction information. Customer review often provides qualitative feedback of a specific product that can be leveraged for targeted improvements and customer ratings offer quantitative data that are often summarized in a star or numeric system. By analyzing both, businesses can capture a snapshot of customer sentiment to gauge the overall customer satisfaction level easily.

2.2 Delivery Performance

Delivery performance in e-commerce encompasses several dimensions such as delivery speed, reliability, and accuracy (Vasić et al., 2020). Delivery speed refers to how quickly an order reaches the customer, reliability and accuracy revolves around the consistency of delivery times and the correctness or condition of the delivered items, respectively (Mentzer, Gomes, & Krapfel, 1989; Shang and Liu 2011; Mentzer, Flint & Hult, 2001). Effective logistics performance ensures that customers receive their orders on time and in the expected condition, significantly impacting overall customer satisfaction.

2.3 Impact of Delivery Performance on Customer Satisfaction

Among the factors usually considered as the drivers of customer satisfaction, delivery performance stands out as one of the most critical components (Demoulin & Djelassi, 2013; Lin, Wu, & Chang, 2011). According to Ahn et al. (2004), a timely and reliable delivery is a

fundamental expectation in e-commerce. It is a standard expectation of customers to receive their purchases within the promised timeframe or even sooner. Businesses that are able to provide delivery performance that meets or exceeds their customer expectations can achieve higher satisfaction levels. This is because timely delivery not only fulfills a key promise made to the customer but also enhances the overall shopping experience by reducing uncertainty and inconvenience. In contrast, delivery delays, inaccuracies, and unreliable deliveries can lead to frustration, negative reviews, and ultimately, reducing the perceived value of the customer to the transaction (Hernández, Jiménez & Martín, 2009). The importance of delivery quality in affecting customer satisfaction is often mentioned in existing literature. However, there are gaps in understanding how this relationship varies across different regions and product types.

2.3.1 Geographical factors

Geographic factors, such as the distance from distribution centers and the efficiency of regional delivery services can significantly affect delivery performance. For example, customers located further from distribution centers may experience longer delivery times, which may lead to dissatisfaction. In addition, the efficiency of regional delivery services varies. Past research by Janjevic and Ndiaye (2014) compared delivery performance in urban and rural areas across Europe. It was found that urban customers enjoyed faster and reliable deliveries, leading to higher satisfaction levels while rural customers faced significant delays and inconsistencies which results in lower satisfaction.

2.3.2 Product Categories

Different types of products have different delivery requirements and customer expectations. For instance, customers may expect faster delivery for perishable goods or daily-use items compared to standard products. The characteristics of different product types, such as size, weight, and perishability, can also affect the logistics processes involved, influencing transportation mode, shipping cost and delivery speed.

Data Description

3.1 Dataset Overview

Dataset Name: Brazilian E-Commerce Public Dataset by Olist

Author: Olist (André Sionek)

Data Collection: The dataset was released by the company.

Source: Kaggle

Content: The dataset consists of nine interrelated datasets:

No	Dataset	Unit of Observation	Column	Row
1	Customers	Each observation identifies unique customers (orders dataset) associated with the orders delivery location.	5	99,441
2	Geolocation	Each observation contains the geolocation of the user and seller within the database.	5	100,0163
3	Order Items	Each observation includes data of each purchase.	7	112,650
4	Order Payments	Each observation includes the data of the payment detail of each transaction.	5	103,886
5	Order Reviews	Each observation includes the timestamp, review description & rating of the reviews.	7	99,224
6	Orders	Each observation includes detail regarding the shipment of the transaction.	8	99,441
7	Products	Each observation includes detail regarding the product.	9	32,951
8	Sellers	Each observation includes detail regarding the seller.	4	3,095
9	Product Category Name Translation.	Each observation includes the translation of the product category name to english.	2	71

3.2 General Description of the Variables

There is a total of 9 dataset with a combined total of 52 columns. The core dataset (Order Dataset), centered around customer orders, comprises 8 key columns which are:

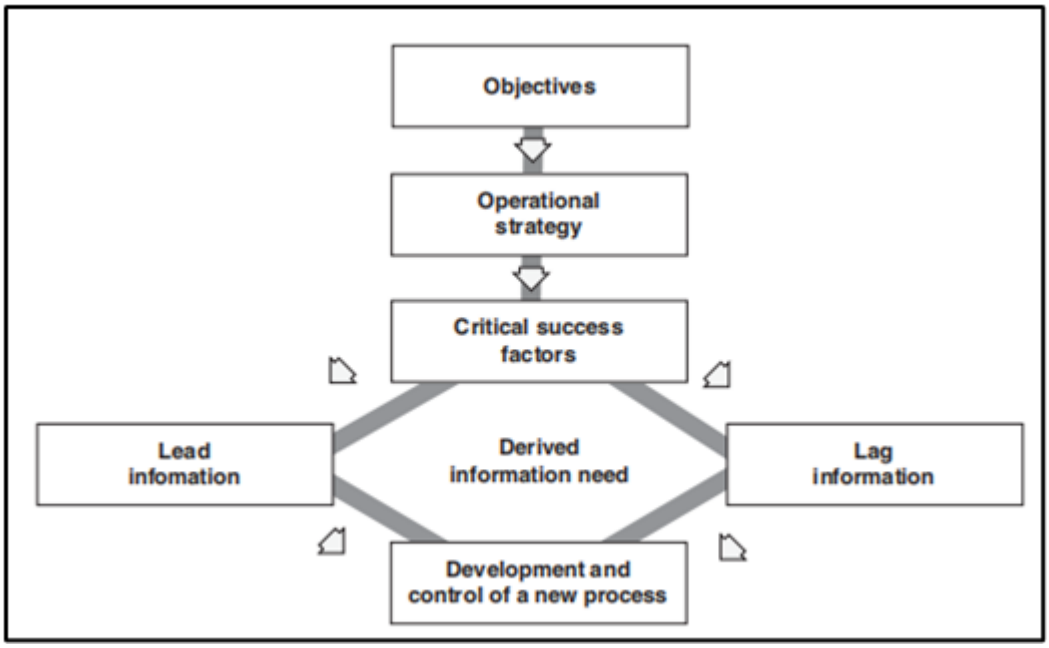
No	Variables	Description

1	order_id	Unique order ID of the purchase.
2	customer_id	Unique customer ID
3	order_status	Order Status of the shipment
4	order_purchase_timestamp	Timestamp of the order purchase
5	order_approved_at	Timestamp of the approved payment
6	order_delivered_carrier_date	The date which the logistic receive the package
7	order_delivered_customer_date	The date which the customer receives the package
8	order_estimated_delivery_date	Estimate delivery duration

These variables capture the essential information about each order, including its unique identifier, the customer associated with it, the status of the order, and various timestamps that track the order's journey from purchase to delivery. This structure allows for data analysis of order processing and delivery performance.

Methodology

4.1 Rockart Model



(Laursen & Thorlund, 2018, pg 60)

The proposal aims to implement a robust Rockart model as an overarching methodology, guided by the company's vision and mission. Objectives are developed to align with the current vision and mission, ensuring strategic alignment. Information flows within the organization are optimized to align key performance indicators (KPIs) with these objectives, creating a systematic approach towards achieving the business goals (Laursen & Thorlund, 2018). Moreover, the success of developing new processes relies on the implementation of critical success factors (CSFs). According to Nah and Delgado (2006), identifying key factors during different phases is essential for success. Therefore, the constant monitoring of the objective of each phase is vital to optimizing the critical success factors ensuring the successful development and deployment of new processes.

Part 2: Report Write Up

Introduction

5.1 Overview of the Research Question

Driven by the continuous development of sophisticated technology and the impact of the Covid-19 pandemic, traditional businesses are increasingly shifting towards the digitalised economy. E-commerce contributes a significant component of this digital economy and has seen its share of global retail sales increase by 8% since the pandemic (International Trade Administration, n.d). Within this highly competitive industry, customer satisfaction has become a crucial factor for businesses striving to maintain market share within the competitive industry. Critical components such as logistics performance do influence customer satisfaction. Therefore, the research question revolves around the aim to explore: How does delivery performance affect customer satisfaction in the e-commerce sector?

5.2 Motivation for the Research

This research question is important because timely and reliable delivery is a fundamental expectation in e-commerce. According to Ahn et al. (2004), customers expect their purchases to arrive within the promised timeframe and the companies that meet or exceed these expectations are often associated with higher levels of customer satisfaction. Alternatively, delays and unreliable deliveries can lead to customer frustration, negative reviews, and loss of business.

Existing literature highlights the importance of delivery performance in shaping customer satisfaction. However, there are gaps in understanding how this relationship varies across different regions and product types. Geographic factors such as distance from the distribution center and the efficiency of regional delivery services can affect delivery speed and reliability. Moreover, customer expectations for delivery speed may differ based on product categories.

This research aims to fill these gaps by studying the relationship between delivery performance and customer satisfaction. By doing so, it aims to provide actionable insights for e-commerce businesses to improve their logistics operations and enhance customer satisfaction.

5.3 Dataset Overview

Dataset Name: Brazilian E-Commerce Public Dataset by Olist

Author: Olist (André Sionek)

Data Collection: The dataset was released by the company.

Source: Kaggle

Content: The dataset consists of nine interrelated datasets:

No	Dataset	Unit of Observation	Column	Row
1	Customers	Each observation identifies unique customers (orders dataset) associated with the orders delivery location.	5	99,441
2	Geolocation	Each observation contains the geolocation of the user and seller within the database.	5	100,0163
3	Order Items	Each observation includes data of each purchase.	7	112,650
4	Order Payments	Each observation includes the data of the payment detail of each transaction.	5	103,886
5	Order Reviews	Each observation includes the timestamp, review description & rating of the reviews.	7	99,224
6	Orders	Each observation includes detail regarding the shipment of the transaction.	8	99,441
7	Products	Each observation includes detail regarding the product.	9	32,951
8	Sellers	Each observation includes detail regarding the seller.	4	3,095
9	Product Category Name Translation.	Each observation includes the translation of the product category name to english.	2	71

5.4 General Description of the Variables

There is a total of 9 datasets with a combined total of 52 columns. The core dataset (Order Dataset), centered around customer orders, comprises 8 key columns which are:

No	Variables	Description
1	order_id	Unique order ID of the purchase.
2	customer_id	Unique customer ID
3	order_status	Order Status of the shipment
4	order_purchase_timestamp	Timestamp of the order purchase
5	order_approved_at	Timestamp of the approved payment
6	order_delivered_carrier_date	The date which the logistic receive the package
7	order_delivered_customer_date	The date which the customer receives the package
8	order_estimated_delivery_date	Estimate delivery duration

These variables capture the essential information about each order, including its unique identifier, the customer associated with it, the status of the order, and various timestamps that track the order's journey from purchase to delivery. This structure allows for data analysis of order processing and delivery performance.

5.5 Exploratory Data Analysis

The descriptive statistics reveal significant variability across key variables. Order Item Price has a mean of $120.65 and ranges from $0.85 to $6,735.00, indicating a diverse price value in regard to the purchase item. Order Payment Value averages $154.10 with a range up to $13,664.08, and Order Payment Installments average 2.85, with up to 24 installments. Order Item Freight Value averages at $19.99, ranging from $0.00 to $409.68. Order Review Scores are exceptionally high, with a mean of 4.09 and most scores being 4 or 5. Product Weight averages 2,276.47 grams, with a wide variation and the product dimensions also display substantial variation. Order Delay averages at 10.03 days, ranging from -452 to 246 days, indicating huge variability in delivery times.

Column	count	mean	std	min	25%	50%	75%	max
Order Item Price	112650	120.65	183.63	0.85	39.90	74.99	134.90	6735.00
Order Item Freight Value	112650	19.99	15.81	0.00	13.08	16.26	21.15	409.68
Order Payment Payment Installments	103886	2.85	2.68	0.00	1.00	1.00	4.00	24.00
Order Payment Payment Value	103886	154.10	217.49	0.00	56.79	100.00	171.84	13664.08
Order Review Review Score	99224	4.09	1.35	1.00	4.00	5.00	5.00	5.00
Product Product Weight (g)	32951	2276.47	4281.91	0.00	300.00	700.00	1900.00	40425.00
Product Product Length (cm)	32951	30.82	16.91	7.00	18.00	25.00	38.00	105.00
Product Product Height (cm)	32951	16.94	13.64	2.00	8.00	13.00	21.00	105.00
Product Product Width (cm)	32951	23.20	12.08	6.00	15.00	20.00	30.00	118.00
Delay Order (D)	112372	10.03	26.85	-452.00	6.00	12.00	16.00	246.00

Analysis, Storytelling and Visualization

6.1 Introduction to Analysis

Brief Overview

The analysis aims to explore the inter-relation between delivery performance and customer satisfaction within the e-commerce sector. Its primary aim is to exhibit light on the dependability between delivery services that influences consumer satisfaction, measured through review ratings. In addition, the study aims to identify the variables influencing delivery delays, recognizing the potential impact on review ratings towards the company. By applying a quantitative approach methodology, the research aims to enlighten the nuanced relationship between delivery performance and consumer satisfaction in the e-commerce landscape within

Brazil. These data driven insights hold significance for retaining market share, providing opportunities to re-evaluate its logistical strategies and improve customer satisfaction.

Problem Statement

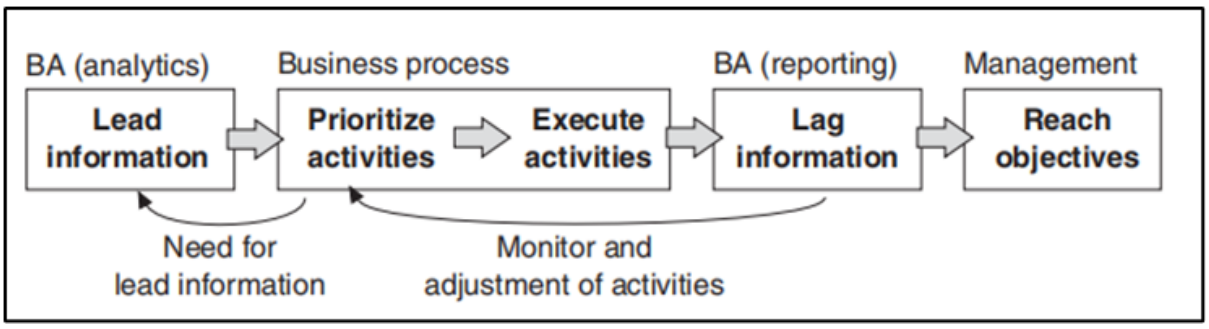
The research question revolves around understanding the influence of delivery performance on customer satisfaction within the e-commerce industry. Though the consumer expected timely and reliable delivery as an expectation within the industry. It is crucial to understand the impact on customer satisfaction as it is crucial for businesses to increase market share within the industry amongst the competitors.

Objectives

1. To explore the relation between delivery performance and customer satisfaction.
2. To identify the factors influencing delivery performance and effects on customer satisfaction.
3. To provide actionable insights towards the e-commerce businesses to optimize the logistics operations aspect and improve customer satisfaction levels.

6.2 Approach Selection

Due to the nature of the Rockart Model, it operates as a closed-loop system with a continuous flow of information throughout the organization's hierarchy, where actions taken become lead information for the organization to analyze. This provides live feedback on the impact of newly implemented processes, using quantitative methods (Palanisamy, 2005; Laursen & Thorlund, 2018).



(Laursen & Thorlund, 2018, pg 60)

By quantifying the organization's vision and mission, departments can create objectives that align with these guiding principles. On the second level, the department heads are required to identify an operational strategy which involves determining the critical success factors essential to this strategy. Provided that the resources are limited, it is crucial to focus the operational strategy strategically instead of attempting to improve all processes and expecting significant change. This requires allocating resources to areas with the highest potential return. Key performance indicators (KPIs) will then be developed to monitor the performance of the newly developed processes. Higher management will monitor and re-evaluate new process performance and the cycle will restart.

6.3 Analysis Process

Python was used to conduct the data cleaning process and tableau was used to visualize the finding of the analysis.

6.3.1 Data Preparation

1. Missing Values
    a. Numeric values were replaced by using the mean function to fill in the missing value.
    b. Within the datasets, there were two types of missing value which require different methods of tackling the data problem.
    c. The image below exhibits that the missing represent a string format thus using the mode function could skew the data distribution thus it was replaced with ‘No Title’ and ‘No Comment’ and fill in the NaN values.
    d. The second missing value represents a numeric format thus the mean function was applied to find the average and fill in the NaN values.

```
# Calculate the sum of missing value in the DataFrame by column
# In this scenario, null value is not remove because review score is more essential towards gauging the customer experiences.
order_review_null = order_review_df.isnull().sum()
print(f'Order Review DataFrame\nNull Value:\n{order_review_null}')
```

```
Order Review DataFrame
Null Value:
review_id                     0
order_id                      0
review_score                  0
review_comment_title      87656
review_comment_message    58247
review_creation_date          0
review_answer_timestamp       0
dtype: int64
```

```
[18] # Therefore the missing value is replace with NA
order_review_df['review_comment_title'] = order_review_df['review_comment_title'].fillna('No Title')
order_review_df['review_comment_message'] = order_review_df['review_comment_message'].fillna('No Comment')
```

```
[19] # Calculate the sum of missing value in the DataFrame by column after replacing values
order_review_null = order_review_df.isnull().sum()
print(f'Order Review DataFrame After Removal\nNull Value:\n{order_review_null}')
```

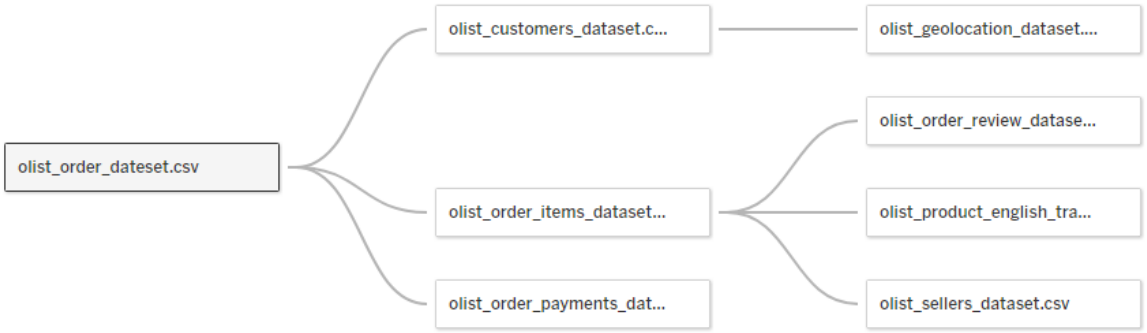
```
Order Review DataFrame After Removal
Null Value:
review_id                 0
order_id                  0
review_score              0
review_comment_title      0
review_comment_message    0
review_creation_date      0
review_answer_timestamp   0
dtype: int64
```

2. Removing Duplicate Values
    a. The duplicated function is used to identify the number of identical rows within the dataset. The observation was removed if found.

```
[39] # Calculate the sum of duplicate rows
order_review_duplicate = order_review_df.duplicated().sum()
print('Order Review DataFrame\nDuplicate:',order_review_duplicate)
```

```
Order Review DataFrame
Duplicate: 0
```

3. Data Integration
    a. Multiple datasets were connected by using inner joints to develop a semantic model for tableau.



b. Multiple datasets were merged to create a master dataset, enabling effective machine learning analysis.

Country

```
[ ] # Perform a many-to-one merge
    order_merged_df = pd.merge(left=order_df, right=order_item_df, on='order_id', how='inner')
    order_merged_df = pd.merge(left=order_merged_df, right=sellers_df, on='seller_id', how='inner')
    order_merged_df = pd.merge(left=order_merged_df, right=customer_df, on='customer_id', how='inner')
    order_merged_df = pd.merge(left=order_merged_df, right=product_merged_df, on='product_id', how='inner')
    order_merged_df = pd.merge(left=order_merged_df, right=order_review_df, on='order_id', how='inner')

    columns_to_drop = ['order_purchase_timestamp',
                       'order_approved_at',
                       'order_delivered_carrier_date',
                       'order_estimated_delivery_date',
                       'shipping_limit_date',
                       'order_delivered_customer_date',
                       'order_id',
                       'order_status',
                       'order_difference',
                       'customer_id',
                       'product_id',
                       'seller_id',
                       'customer_unique_id',
                       'seller_city',
                       'customer_city',
                       'seller_zip_code_prefix',
                       'customer_zip_code_prefix',
                       'seller_state',
                       'customer_state',
                       'product_category_name_english',
                        'order_id',
                        'review_comment_title',
                        'review_comment_message',
                        'review_creation_date',
                        'review_answer_timestamp',
                        'review_id']

    country_df = order_merged_df.drop(columns=columns_to_drop)
```

4. Data Transformation
    a. Python Environment
        i. Convert string into date time format
            1. Order Estimated Delivery Date
            2. Order Delivered Customer Date
        ii. Convert string into categorical format
            1. Seller State
            2. Customer State
            3. Product Category Name English

b. Tableau Environment

i. *Actual Delivery Days* = [Order Delivered Customer Date]-[Order Purchase Timestamp]

ii. *Delay Delivery Day* = [Actual Delivery Days]-[Estimated Delivery Days]

iii. *Estimated Delivery Days* = [Order Estimated Delivery Date]-[Order Purchase Timestamp]

iv. *Number of Late Shipment* = IF DATEDIFF('day', [Order Estimated Delivery Date], [Order Delivered Customer Date]) > 0 THEN 1 ELSE 0 END

v. *Demand of Product* = COUNT([Product Id])

vi. *Order Count* = COUNT([Order Id])

vii. *Percentage of Early Shipment* = SUM(IF ([Delivery Delay (Days)]) < 0 THEN 1 ELSE 0 END) / COUNT([Order Id])

viii. *Percentage of Late Shipment* = SUM(IF DATEDIFF('day', [Order Estimated Delivery Date], [Order Delivered Customer Date]) > 0 THEN 1 ELSE 0 END) / COUNT([Order Id])

6.3.2 Exploratory Data Analysis (EDA)

1. Univariate Analysis

Univariate analysis utilizes statistical methods to describe a single variable, emphasizing its distribution and central tendencies, thus providing a descriptive understanding of the data (Akboğa and Baradan, 2017). Within the report, the only univariate analysis conducted was on the count of review ratings which is known as frequency analysis. Additionally, descriptive analysis was applied to the remaining variables to understand their distributions and central tendencies.

2. Bivariate Analysis

Bivariate analysis aims to explore the relationships between two variables, using a more exploratory approach in nature. By utilizing scatter plots, box plots, and other visualization methods, to identify patterns, trends, and correlations (Akboğa and Baradan, 2017).

6.3.3 Dashboard

Based on Burkhard's (2004) research, visualization provides several advantages, which include:

1. Reducing information overload
2. Misinterpretation
3. Misuse of information

The research establishes that by applying visualization as a means to transfer knowledge retains the quality of information transfer. Moreover, visualization requires less cognitive effort to process compared to text formats, therefore retaining the viewer's attention span more effectively during presentations.

6.4 Exploratory Data Analysis (EDA)

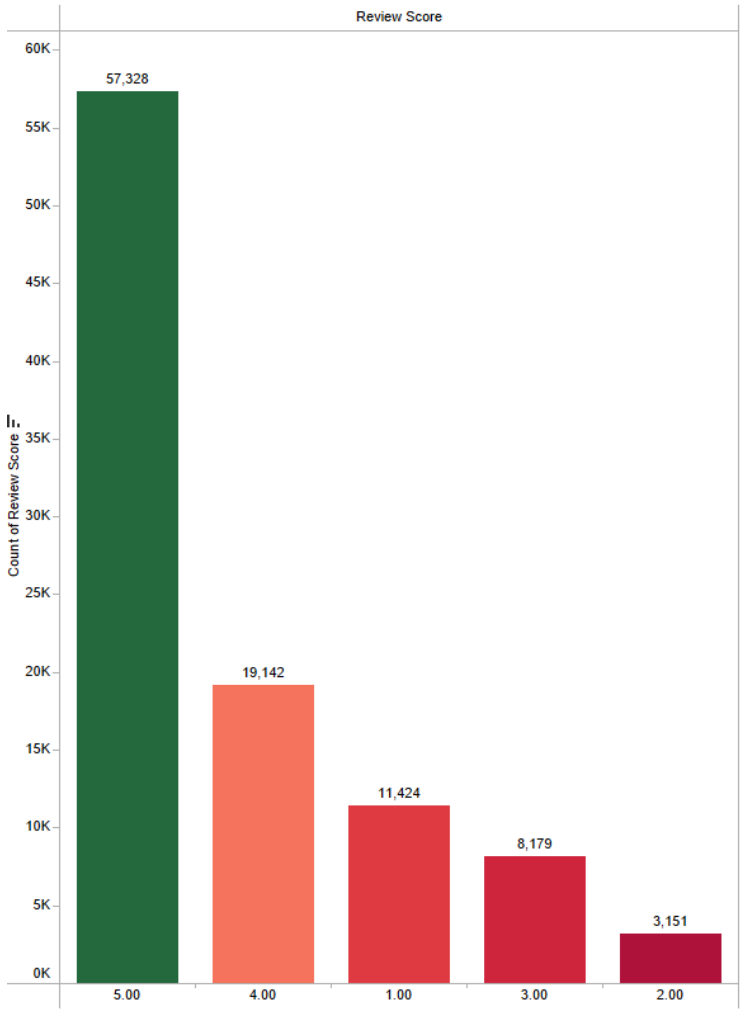
6.4.1 Univariate Analysis

The descriptive statistics reveal significant variability across key variables. Order Item Price has a mean of $120.65 and ranges from $0.85 to $6,735.00, indicating a diverse price value in regard to the purchase item. Order Payment Value averages $154.10 with a range up to $13,664.08, and Order Payment Installments average 2.85, with up to 24 installments. Order Item Freight Value averages at $19.99, ranging from $0.00 to $409.68. Order Review Scores are exceptionally high, with a mean of 4.09 and most scores being 4 or 5. Product Weight averages 2,276.47 grams, with a wide variation and the product dimensions also display substantial variation. Order Delay averages at 10.03 days, ranging from -452 to 246 days, indicating huge variability in delivery times.

Column	count	mean	std	min	25%	50%	75%	max
Order Item Price	112650	120.65	183.63	0.85	39.90	74.99	134.90	6735.00
Order Item Freight Value	112650	19.99	15.81	0.00	13.08	16.26	21.15	409.68
Order Payment Payment Installments	103886	2.85	2.68	0.00	1.00	1.00	4.00	24.00
Order Payment Payment Value	103886	154.10	217.49	0.00	56.79	100.00	171.84	13664.08

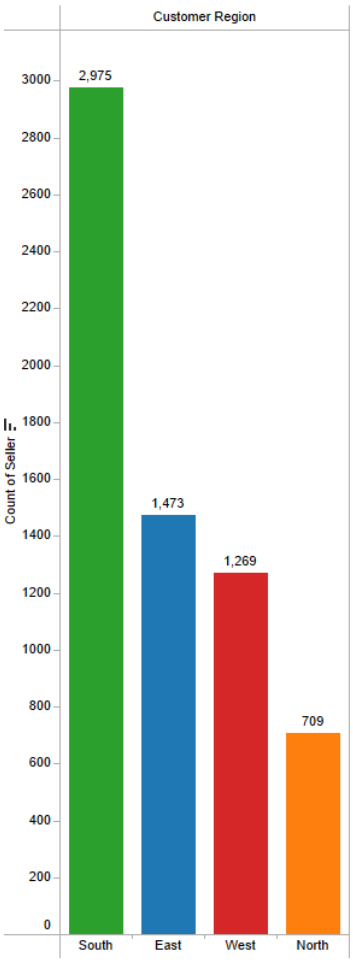
Order Review Review Score	99224	4.09	1.35	1.00	4.00	5.00	5.00	5.00
Product Product Weight (g)	32951	2276.47	4281.91	0.00	300.00	700.00	1900.00	40425.00
Product Product Length (cm)	32951	30.82	16.91	7.00	18.00	25.00	38.00	105.00
Product Product Height (cm)	32951	16.94	13.64	2.00	8.00	13.00	21.00	105.00
Product Product Width (cm)	32951	23.20	12.08	6.00	15.00	20.00	30.00	118.00
Delay Order (D)	112372	10.03	26.85	-452.00	6.00	12.00	16.00	246.00

Based on the bar chart, the distribution of review scores reveals the distribution in the frequency of each rating. The review score of 1 has a count of 11,424, indicating the number of negative reviews. On the other hand, the score of 2 has the lowest count, with only 3,151 counts. At the other end of the range, the review score of 5.00 stands out as the most frequent, with a count of 57,328, highlighting a significant proportion of positive reviews. Subsequently, the score of 4 has a substantial count of 19,142, while the score of 3 has 8,179 count. This distribution suggests that while there is a considerable amount of high satisfaction (scores of 4 and 5), there is also a fair amount of extremely low satisfaction.

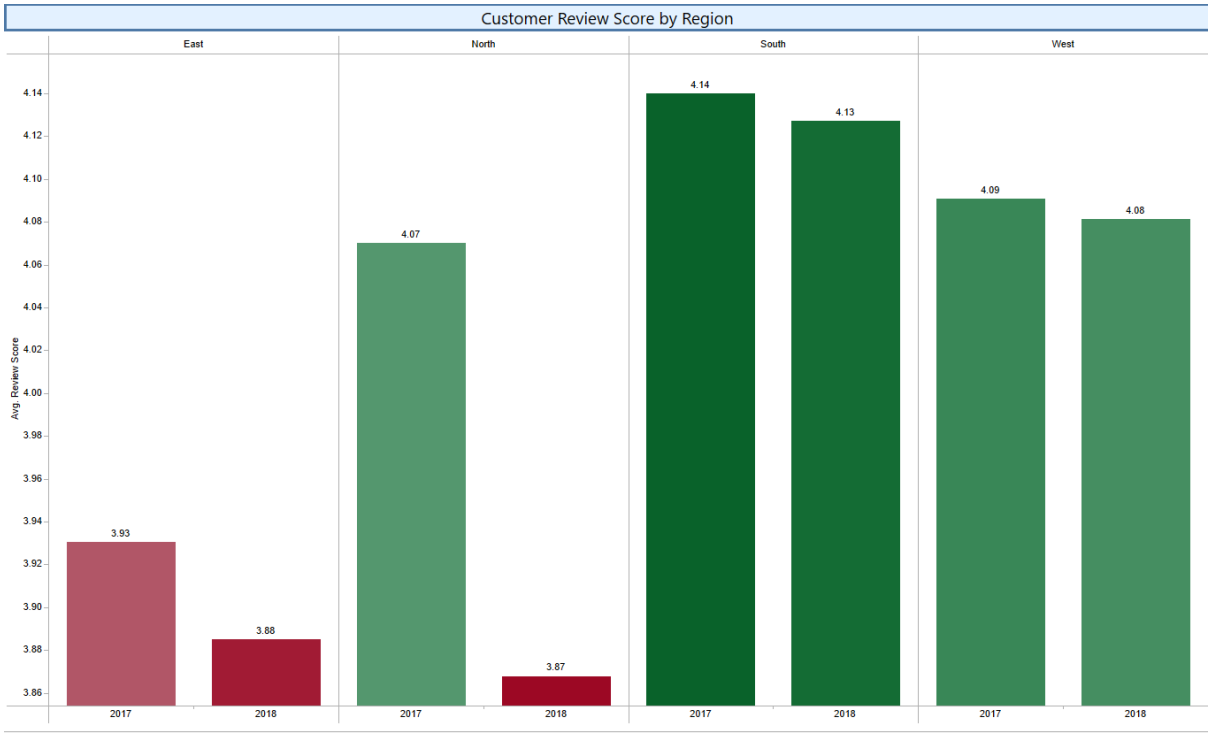


6.4.2 Bivariate Analysis

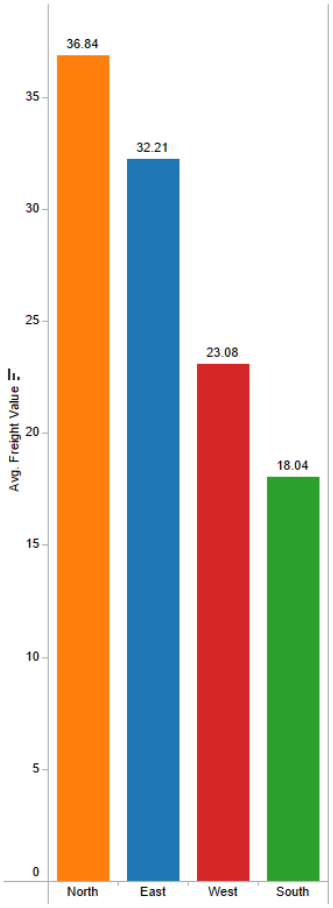
Based on the image below, the distribution of sellers across different regions reveals significant discrepancies. The South region emerges as the area with the highest number of sellers, totaling 2,975. On the other hand, the North region exhibits the lowest count, with only 709 sellers. The East and West regions only account for 1,473 and 1,269 sellers respectively. This distribution shows the varying levels of seller presence across regions, with the South dominating while the North lags significantly.



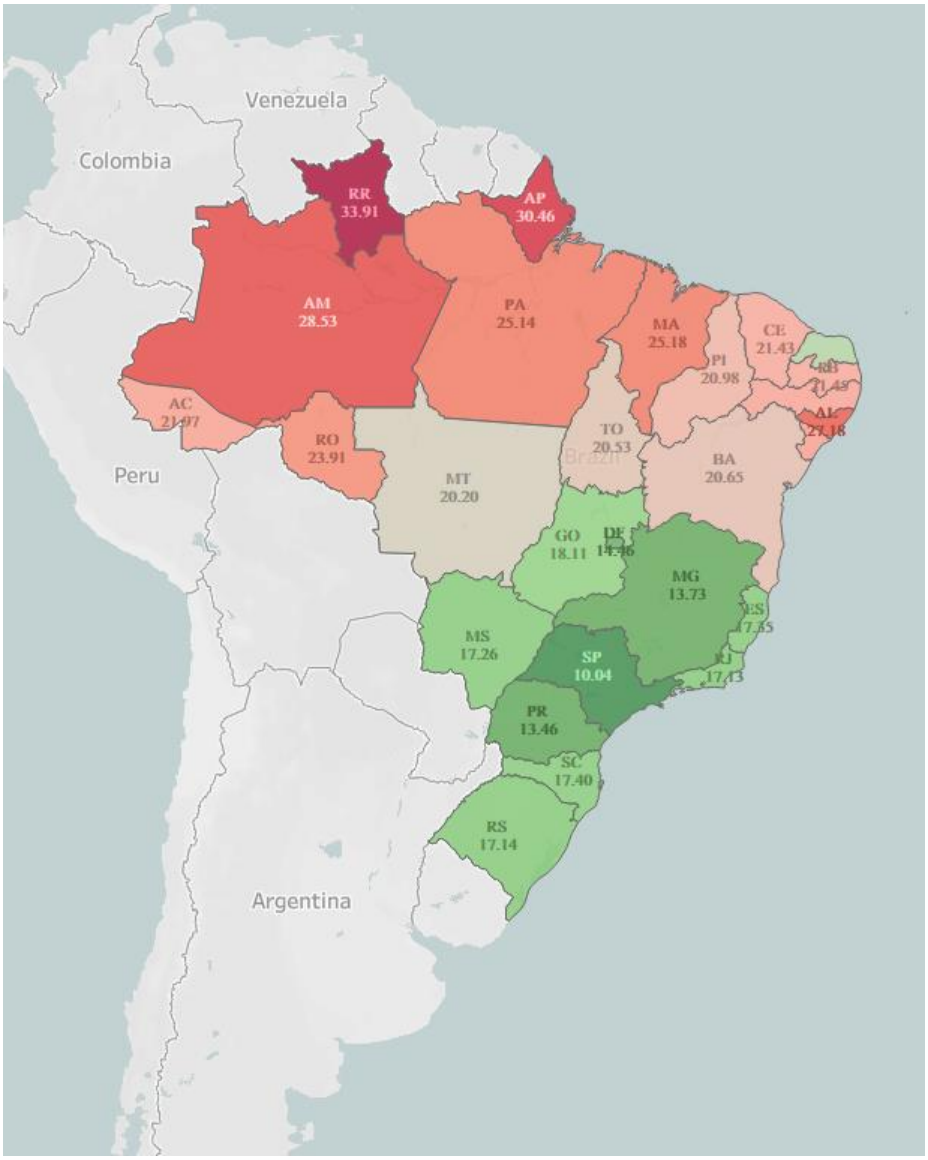
The image below indicates a downward trend from 2017 to 2018. The northern region experienced the most significant change, with a rating decrease of 4.91%. This was followed by the eastern region, which saw a decrease of 1.27%. The remaining regions maintained their ratings relatively well compared to the north and east, although there was still a slight decline of 0.24%. Overall, this suggests a broad decline in ratings across all regions during this period.



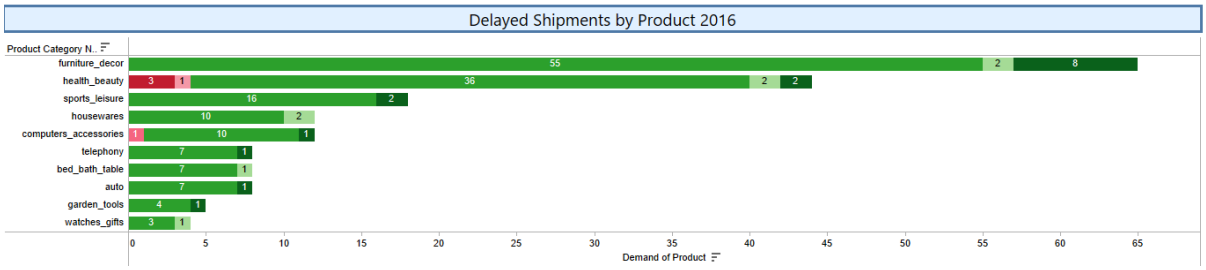
Based on the image below, the North region stands out with the highest average freight value, reaching $36.84, while the South region displays the lowest at $18.04. Following closely, the East region records an average freight value of $32.21, and the West region follows with $23.08. This discrepancy exhibits varying cost structures across regions, with the North exhibiting the highest freight costs on average and the South, the lowest.



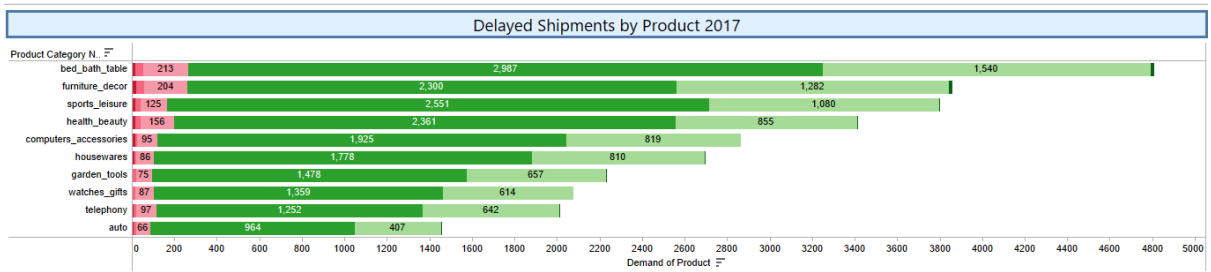
Based on the image below, it is shown that the northern region typically exhibits higher delivery delays compared to the southern region. Especially, the state of RR in the north has the highest average delivery delay, taking an estimate of 33.91 days. In contrast, the southern region enjoys significantly lower delays, with the lowest being an average of 10.04 days. Interestingly, the state of RN, located in the eastern region, exhibits an average delivery delay of 19 days, which is notably lower than its surrounding states, where delays surpass 20 days on average. This makes RN an outlier in the east, a region otherwise characterized by higher average delivery delays.



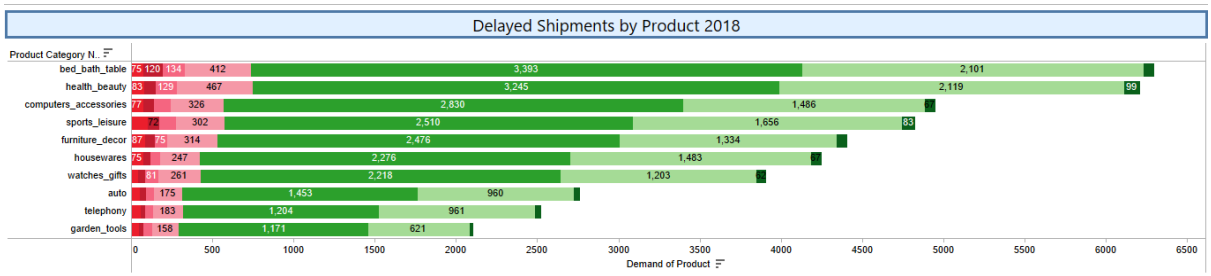
Based on the horizontal bar chart below, it is shown that despite the data covering only several months of 2016, there was minimal performance delay in logistics. This visualization was selected because it highlights the product with the highest total count, providing a clear representation of logistical efficiency during this period.



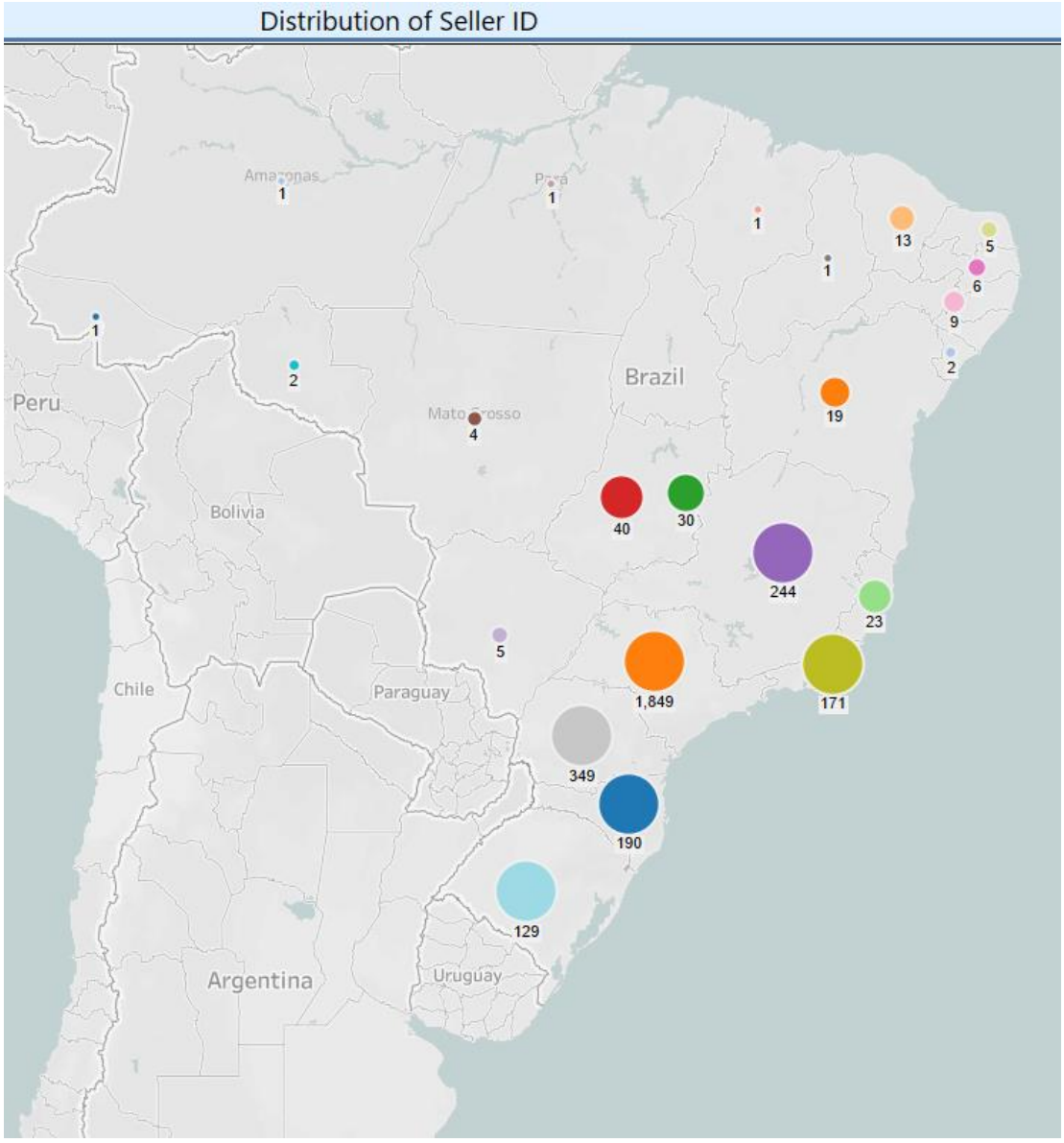
The horizontal bar chart for 2017 displays an increase in the count of product delays across various categories compared to 2016. This trend suggests a decaying in logistical performance when compared to 2016.



The horizontal bar chart for 2018 shows an increase in the count of product delays across various categories compared to 2017. This rise suggests that logistical performance is drastically deteriorating over time.

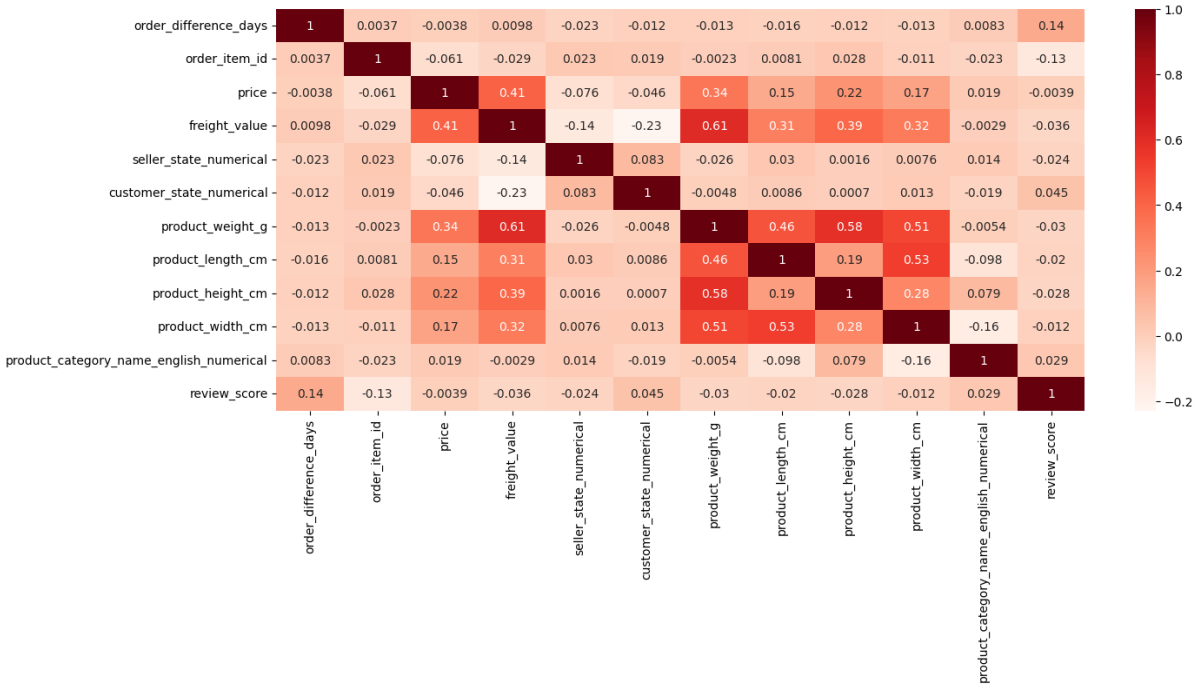


Based on the map below, it shows the distribution of sellers across various states. SP stands out with the highest number of sellers per state (1833). In contrast, the northern region has a median of only one seller per state. The eastern region shows a slightly higher number of sellers per state, ranging from 1 to 19. Lastly, the southern states exhibit the highest average number of sellers compared to the other regions.



Country Level

The heatmap below acts as a continuation of the pair plot analysis. Revolving on the key variables which are review score and order difference days. It is indicated that there is no strong correlation between these variables. The color intensity on the heatmap confirms that the relationship between review scores and order difference days is weak, suggesting that these variables do not significantly influence each other.

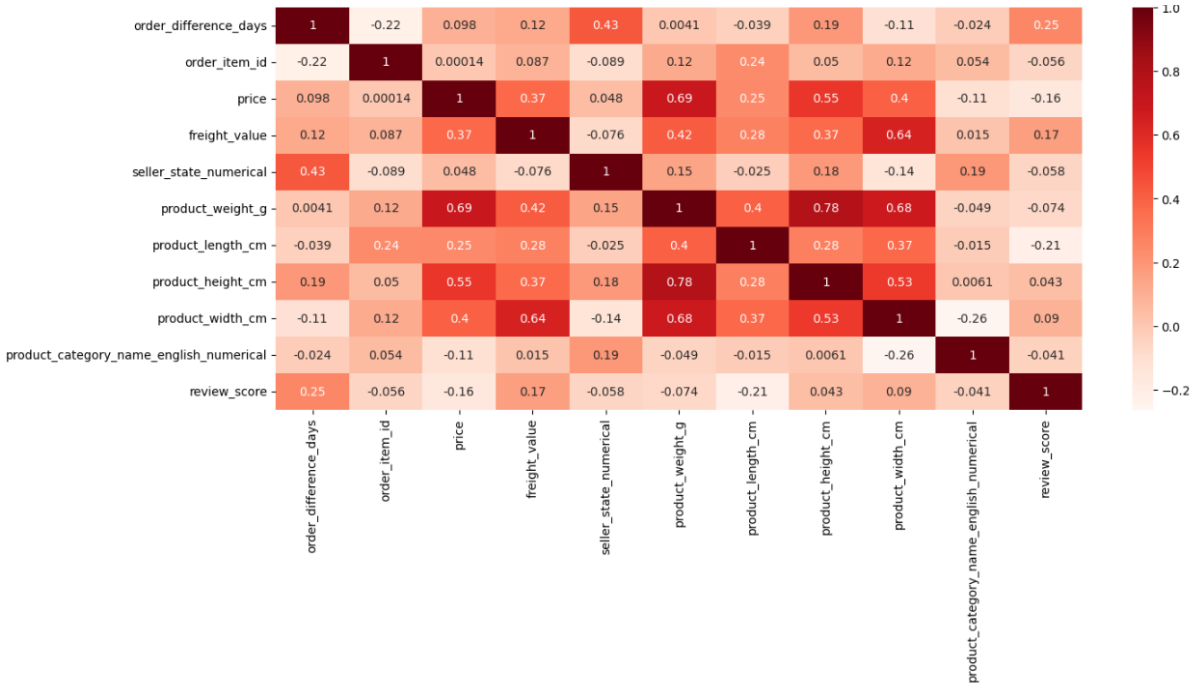


To address this problem, the data was filtered down at the state level within Brazil to explore the correlation to achieve more granularity. As there are currently 26 states in Brazil, only two states were selected for in depth exploration. While each state may differ in its specific variables, this approach serves as an example to show how state-level analysis can reveal different factors influencing the target variable (review ratings). This method helps in understanding the localized impact of various factors on customer satisfaction.

Least populated State of Brazil - Roraima (RR)

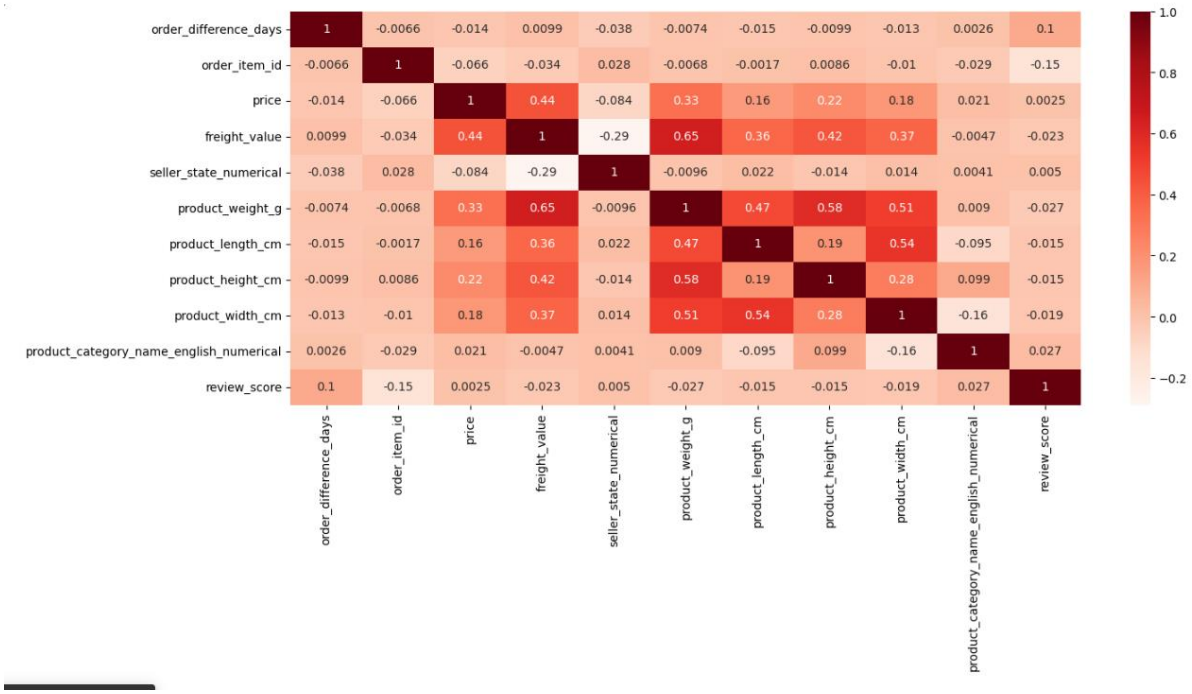
Based on the heatmap, it shows that the variable order difference days have the highest influence on the review score, with a correlation of 0.25. This is followed by freight value, which has a correlation of 0.17 and product width, with a correlation of 0.09. These findings align with the research conducted by Janjevic and Ndiaye (2014) in which indicates that rural areas typically experience significant delivery delays which heavily influence review rating. The data confirms that logistical factors such as delivery times and freight costs are significant determinants of customer satisfaction in these regions.

In addition, from a rural location standpoint, the variables such as the seller's state have the highest correlation with order difference days, with a correlation of 0.43, followed by the height of the product at 0.19 and freight value at 0.12. This re-emphasizes the impact of logistical challenges on a geographic aspect on delivery performance and customer satisfaction in rural areas.



Most populated State of Brazil - São Paulo (SP)

On the other hand, the most populated state does reflect the same findings as Roraima. Within this state, the heatmap indicates that order difference days had the highest correlation with the review score, but it was only 0.1. This is followed by the seller's state with a correlation of 0.05, and product category with a correlation of 0.02. Furthermore, there were no significant factors that could heavily influence order difference days or the review rating, suggesting that the state benefits from more efficient logistics operations. This contrast highlights how logistical efficiency can vary greatly between different regions, affecting customer satisfaction in different ways.

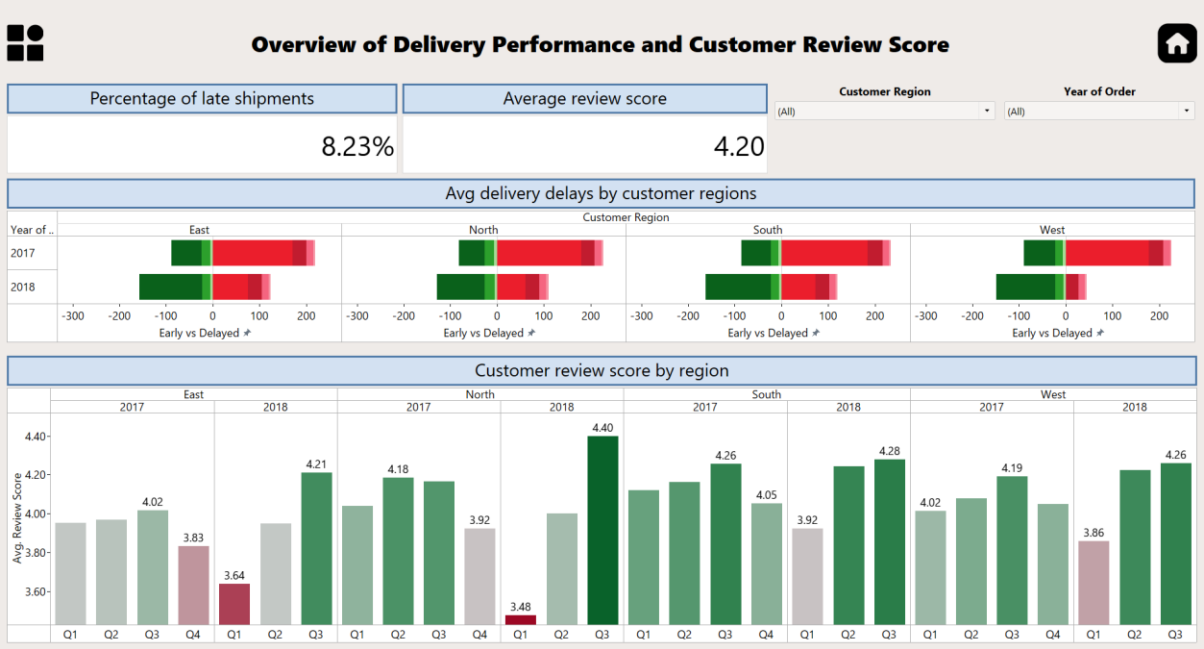


6.4.3 Storytelling and Visualization

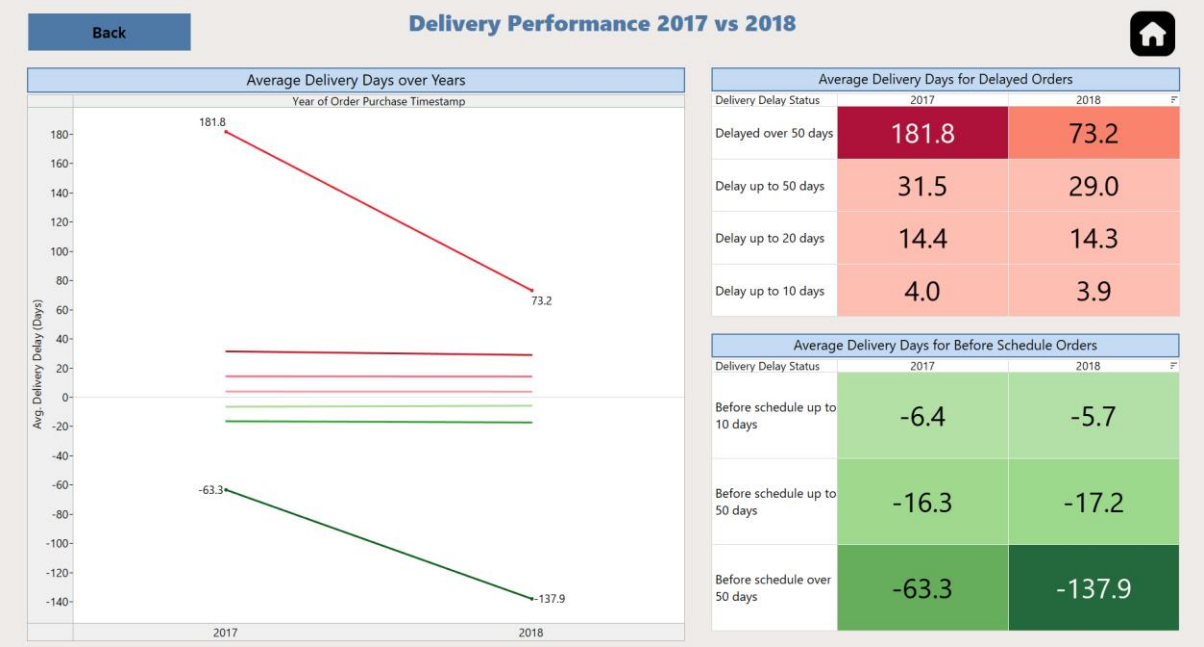
Dashboard 1 opens with an 8.23% overall late shipment rate and an average customer review score of 4.20, setting the benchmark for delivery performance.

In 2018, a significant increase in early shipments can be observed compared to 2017 across all regions. This improvement is visually represented by the expanded green bars in the dashboard, indicating a higher volume of shipments arriving earlier than scheduled. The increase in early shipments suggests that logistical operations have been optimized over the year, possibly due to enhanced supply chain strategies or better management of delivery expectations.

Despite improvements, all regions experienced a notable decline in customer satisfaction in the first quarter of 2018, suggesting a temporary systemic issue affecting delivery performance that requires further investigation. However, the Northern region, with the smallest number of sellers, managed to recover impressively in customer ratings by the third quarter of 2018, possibly due to less demanding customer expectations due to fewer options.



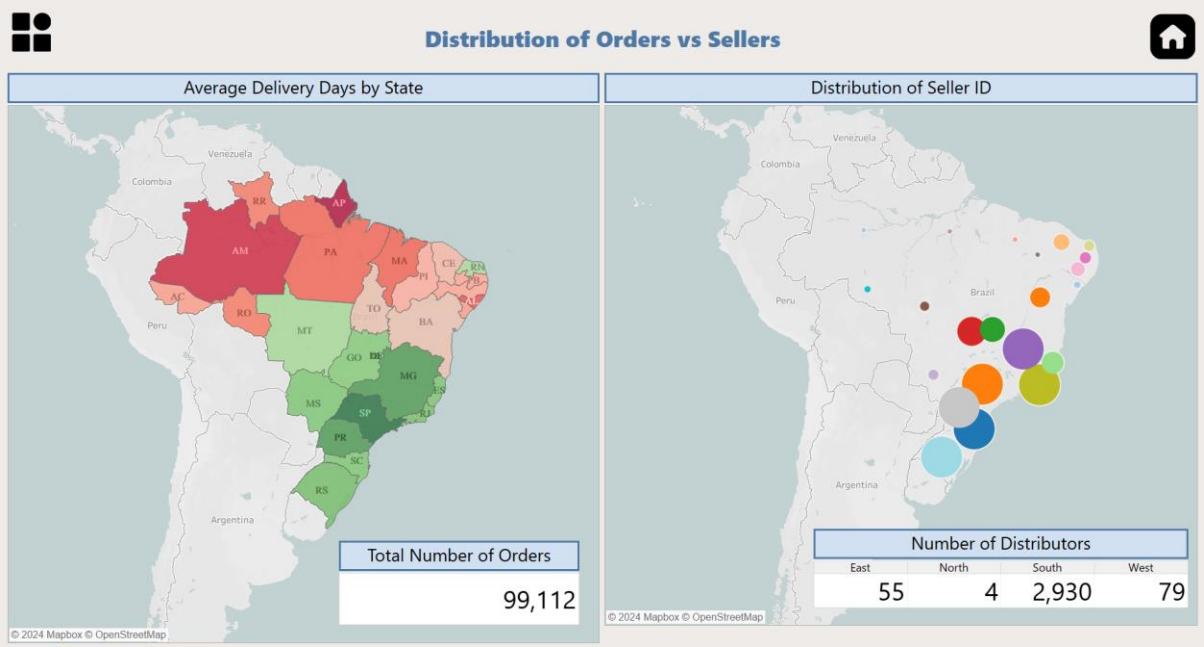
Dashboard 1: Overview of Delivery Performance and Customer Review Score



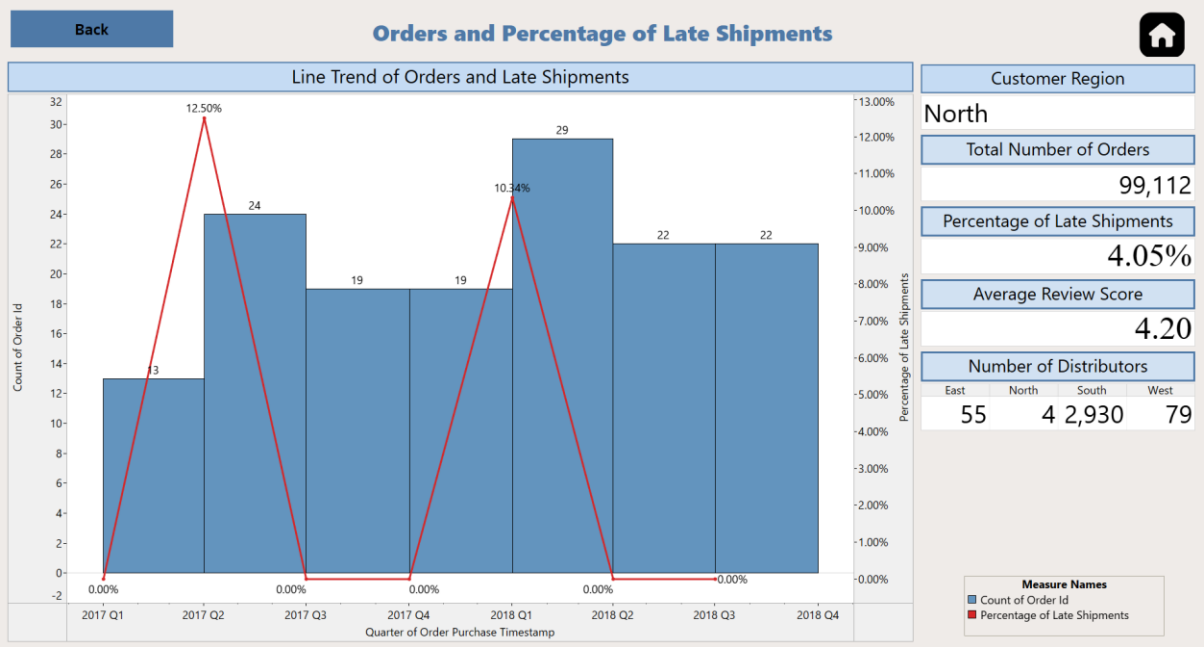
Dashboard 1a: Comparison of Delivery Performance in 2017 and 2018

This dashboard explores deeper into the average delivery days by state, revealing longer delivery times in northern states like Roraima (RR) also records the lowest customer review scores, directly correlating poor delivery performance with low customer satisfaction. This contrast with Amapá (AP), which maintains relatively higher satisfaction despite similar delivery challenges suggested regional logistical deficiencies in certain states.

Besides, the current distribution strategy also reveals inefficiencies. A high concentration of distributors in the South that results in shorter delivery times and higher customer satisfaction scores can be observed in these states. However, the analysis also indicates an inorganization routing system, where sellers located in the South are tasked with covering deliveries for all regions, including the distant North, and conversely, sellers in the North are delivering to the South. This arrangement likely contributes to the extended delivery times observed in Roraima and other regions, underlining the need for a more strategically organized distribution route to enhance efficiency across all areas.



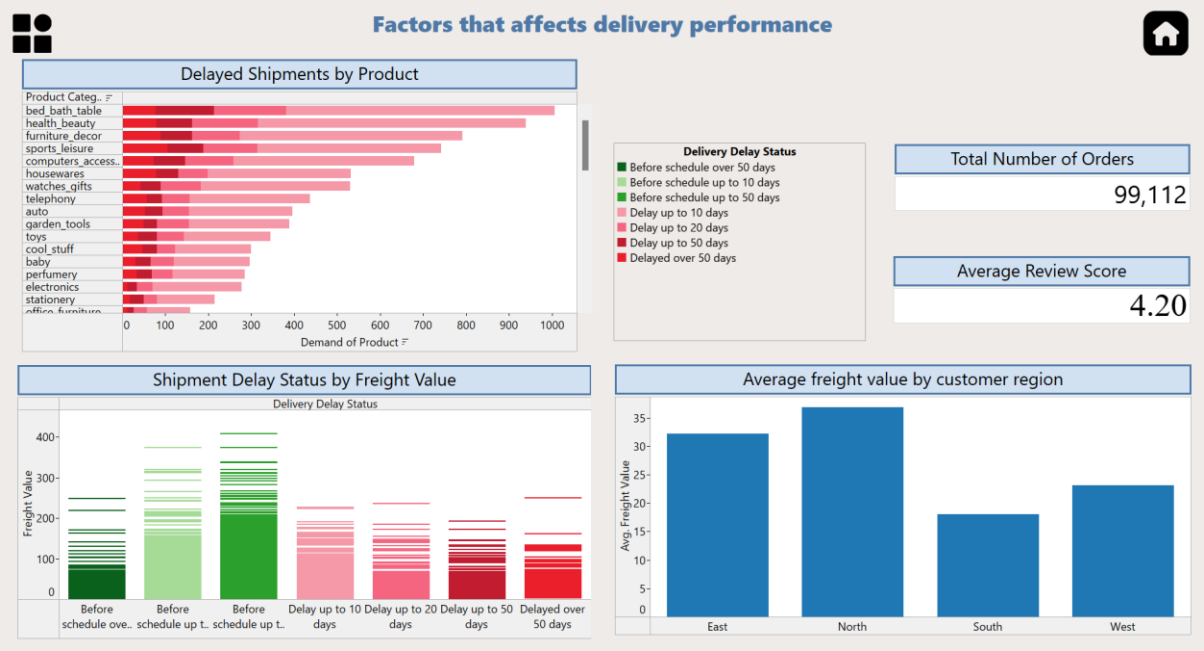
Dashboard 2: Geographic analysis and order distribution



Dashboard 2a: Comparison of total number of order and percentage of late shipments

This analysis explores how different product categories affect delivery performance. Product categories such as health, beauty, bed, bath & table show the most significant delays. This could influence customer perceptions and satisfaction differently due to their essential nature. Most customers may expect timely deliveries for these products as they are essential for daily-use and maintaining quality of life. Therefore, it is encouraged that the e-commerce company should prioritize inventory management and shipment of these items, while enhancing communication about order status and regularly analyze delivery processes to identify and address the root causes of delays.

In addition, the last dashboard also shows that higher-value shipments generally arrive on schedule more frequently, highlighting a possible prioritization in the logistics process. This suggests an opportunity to re-evaluate freight logistics to ensure equitable delivery performance across all shipment values for better customer experience.

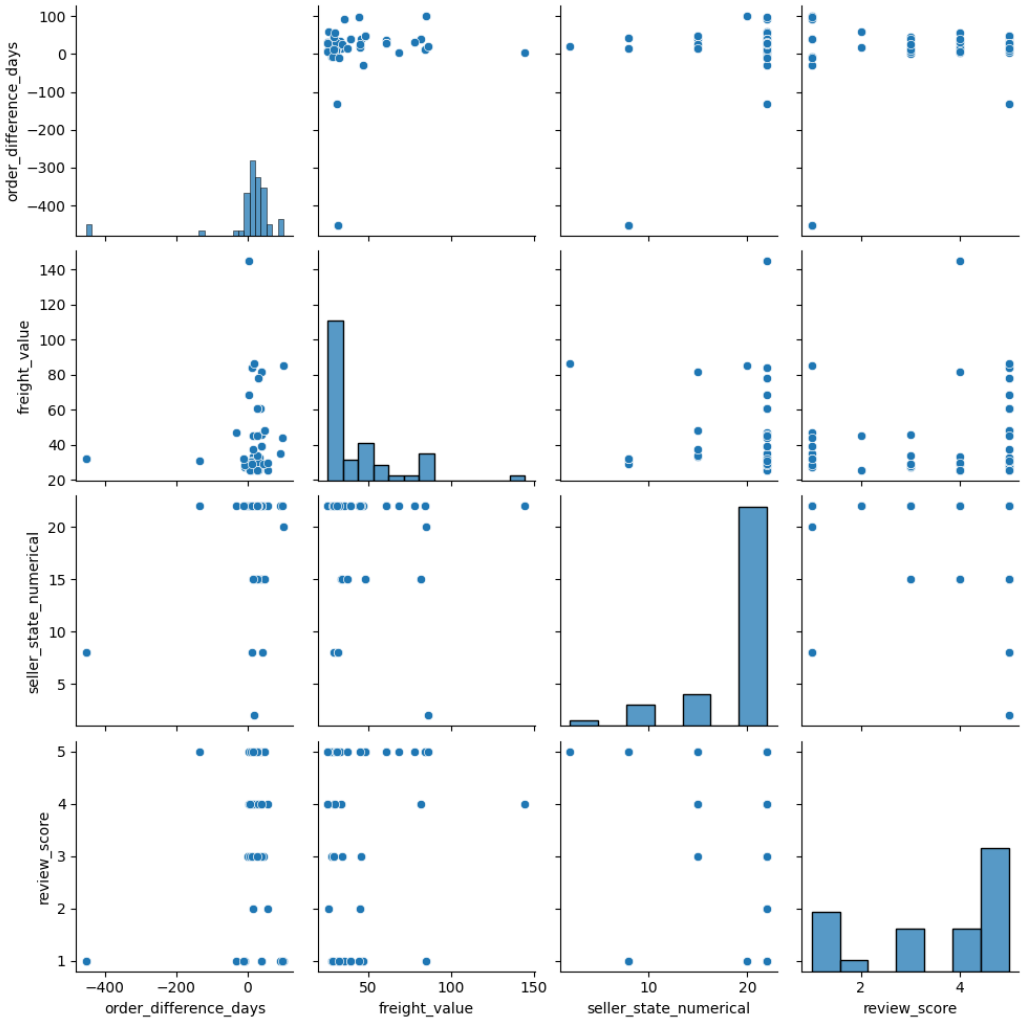


Dashboard 3: Product Categories and Freight Value Effects

Discussion

7.1 Roraima

Based on the Roraima State pairplot generated using the Seaborn library in Python, it can be observed that higher freight values in rural areas was associated with lower delivery difference days. Specifically, freight values above $50 tend to correspond with lower order differences. Suggesting that higher freight costs may be a positive correlation with better delivery performance in rural areas. Furthermore, it can be speculated that poor logistics infrastructure in rural areas was heavily influenced by the location of the seller's state therefore affecting order difference days. Some states show higher order difference days, indicating variability in logistics efficiency. According to Dashboard 2, the southern regions have a higher number of sellers compared to the northern regions, which suggests that logistics efficiency may degrade as the packages travel towards the northern region. Therefore, the customers within the state of Roraima are especially sensitive towards order difference days. Hence it is crucial to improve logistics operations to maintain customer satisfaction.


order_difference_days
freight_value
seller_state_numerical
review_score
100
0
−100
−200
−300
−400
140
120
100
80
60
40
20
20
15
10
5
5
4
3
2
1
−400
−200
0
50
100
150
10
20
2
4

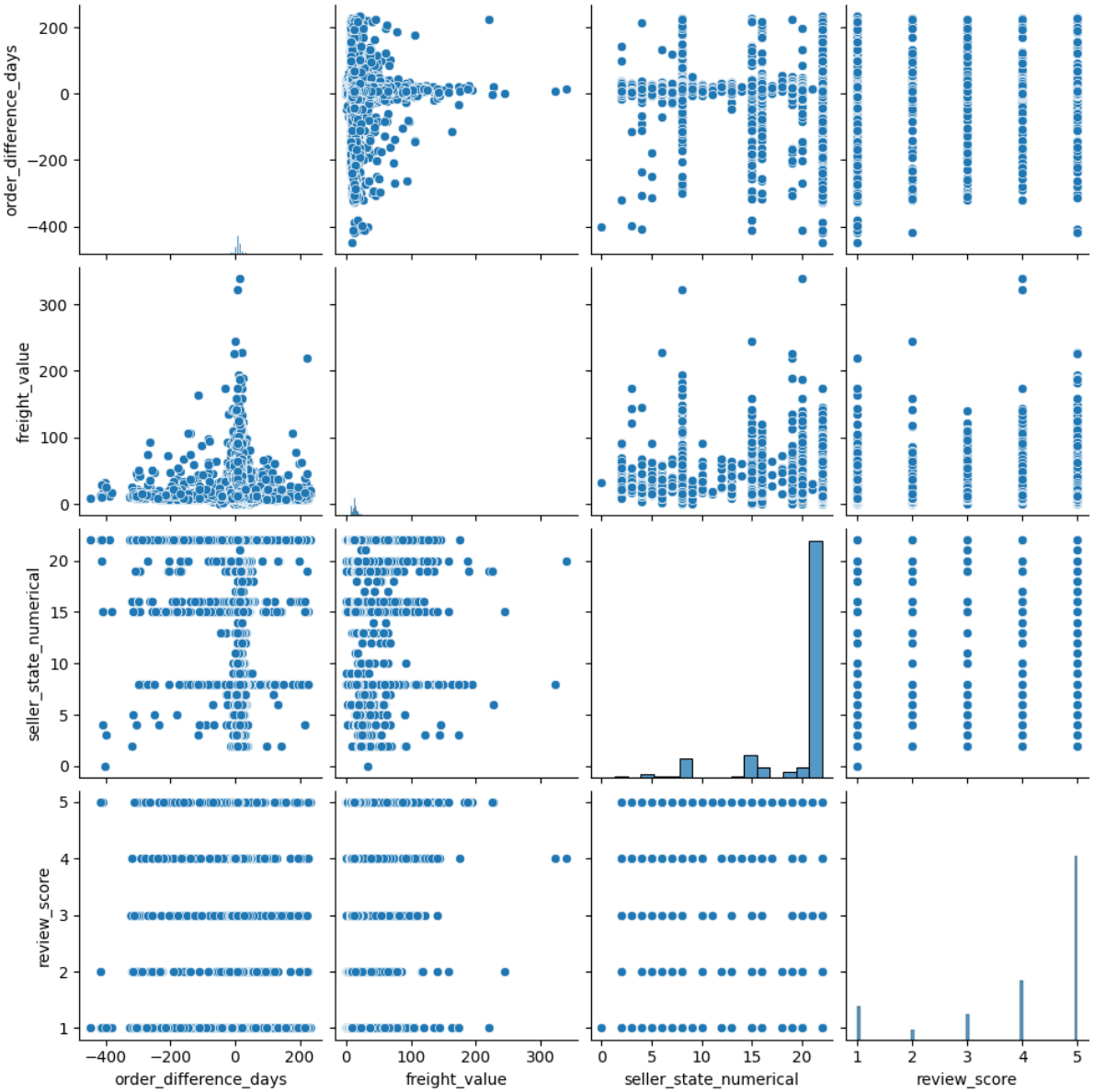
7.1.1 Roraima State Rockart Strategy

Level 1 - Objective	Maintain and Enhance Customer Satisfaction
Level 2 - Identifying an operation strategy	Strategically re-evaluating the operation of the logistic to identify inefficiencies and areas for improvement which includes: 1. Analyzing current delivery processes. 2. Assessing the changes within the supply chain network. 3. Benchmarking against industry best practices.
Level 3 - Identifying Critical Success Factors	1. Implementation of Six Sigma methodologies to systematically improve logistics processes and reduce inconsistency. 2. Incorporate advanced logistics software for real-time tracking and improved route optimization.
Level 4 - Identifying Lead & Lag information (KPI)	1. Reduce average delivery delays by 10%. 2. Decrease logistics-related complaints by 20%.

7.2 São Paulo

On the other hand, São Paulo does not exhibit the same characteristics as the state of Roraima. Although the heatmap for São Paulo shows a lower correlation for review scores when compared to Roraima (difference of 0.15), there is still a small pattern within the scatter plot of order difference days & freight values. Despite not being obvious, the logistics within the state show inconsistency in order difference days, particularly in the lower range of freight values. Delivery delays range from 200 days late to 400 days early, suggesting the lack of consistency within the operations.

Although São Paulo demonstrates a more efficient average delivery delay of 10.04 days compared to Roraima's 33.91 days, this finding could align with Dashboard 2, which shows a higher count of sellers in the southern regions. Suggesting that there could be more logistical development within the south, enabling shorter delivery times. However, the findings also suggest that the logistics operation in São Paulo is not consistent and could lack quality checks to optimize estimated delivery dates. While the factors influencing customer satisfaction may differ from those in Roraima, São Paulo requires a different methodology to identify which attributes truly influence review ratings. Therefore, the exploration of customer satisfaction based on the dimensions of e-SERVQUAL would be more effective than focusing only on logistic improvements (Lin et al., 2016). This approach would provide insights into various aspects of service quality, including efficiency, reliability, fulfillment, and responsiveness, which are crucial for enhancing customer satisfaction in São Paulo's e-commerce sector.


order_difference_days
freight_value
seller_state_numerical
review_score
200
0
−200
−400
300
200
100
0
20
15
10
5
0
5
4
3
2
1
−400
−200
0
200
0
100
200
300
0
5
10
15
20
1
2
3
4
5

7.2.1 São Paulo State Rockart Strategy

Level 1 - Objective	Maintain and Enhance Customer Satisfaction
Level 2 - Identifying an operation strategy	Strategically re-evaluating the dimension of e-SERVQUAL to identify areas for improvement which includes: 1. Website design 2. Effectiveness and efficiency of online browsers 3. Security issues
Level 3 - Identifying Critical Success Factors	1. Customer Feedback Analysis 2. Employee Training
Level 4 - Identifying Lead & Lag information (KPI)	1. Customer Satisfaction Scores increase by 5% 2. Web traffic increase by 5% 3. Unique user ID increase by 10%

Limitations

8.1 Data Problem

8.1.1 Missing values

The datasets used in this analysis contained missing values, which may have led to inaccuracies in the results and insights derived.

8.1 2 Inconsistent Data

There were inconsistencies within and across the datasets, making data cleaning and preprocessing more complex and potentially impacting the reliability of the analysis.

8.1.3 Limited Data Period

The analysis was based on only 2.5 years of data, which is insufficient to observe long-term trends and seasonal patterns. A more extended dataset would provide a better understanding of the trends.

8.1.4 Lack of Transportation and Logistics Data

The datasets did not include information on the mode of transportation, or the logistics service companies used by Olist. This data is crucial for a more detailed assessment of delivery performance and identifying potential areas for improvement.

8.2 Tableau

8.2.1 Data Preprocessing

Tableau is primarily a visualization tool and offers limited capabilities for data preprocessing. Additional tools are necessary to handle data cleaning, transformation, and integration before visualization.

8.2.2 Limited 'Filter' function

The filter function in Tableau has limitations. It cannot be used in 'Story' mode, affecting presentations to stakeholders. Additionally, filters in Tableau are interconnected across multiple dashboards, which means that selecting a filter in one dashboard without deselecting it can cause subsequent dashboards to display incorrect numbers. Furthermore, creating numerous filters can make the Tableau file heavy, which may slow down the performance.

8.3 Organizational Insights

The project did not include information of the organization itself, which limits the ability to determine the root causes of the issues identified in the data. A more comprehensive approach including organizational data would provide deeper insights.

8.4 Data Modeling

The current data model requires further development to be robust enough for deployment. Improvements in model accuracy, validation, and testing are necessary to ensure it can be reliably used.

Conclusion

This study has highlighted the critical impact of delivery performance on customer satisfaction within the e-commerce sector. By analyzing the Brazilian E-Commerce Public Dataset provided by Olist, it was found that there were significant variations in delivery performance and customer satisfaction across different regions and product categories.

The analysis revealed that longer delivery times and higher delivery delays are associated with lower customer satisfaction scores, particularly in rural regions like Roraima. This underscores the importance of efficient logistics operations and the need for strategic distribution routes. In contrast, more populous regions like São Paulo demonstrated more efficient logistics operations, although inconsistencies in delivery performance still exist. This indicates that while São Paulo benefits from a more developed logistics network, there is still room for improvement to ensure consistent delivery performance.

The study also emphasized the need for e-commerce businesses to prioritize timely and reliable deliveries, particularly for essential products like health and beauty items, where customer expectations for prompt delivery are higher.

Future work could involve a deeper exploration of the impact of different logistics strategies and the role of various stakeholders in the supply chain to further enhance delivery performance and customer satisfaction. Additionally, incorporating more comprehensive data on transportation modes and logistics service providers would provide a more detailed understanding of the factors influencing delivery performance.

In conclusion, improving delivery performance is crucial for enhancing customer satisfaction in the e-commerce sector. By addressing logistical inefficiencies and optimizing distribution routes, e-commerce businesses can better meet customer expectations, leading to increased loyalty and sustained business growth.

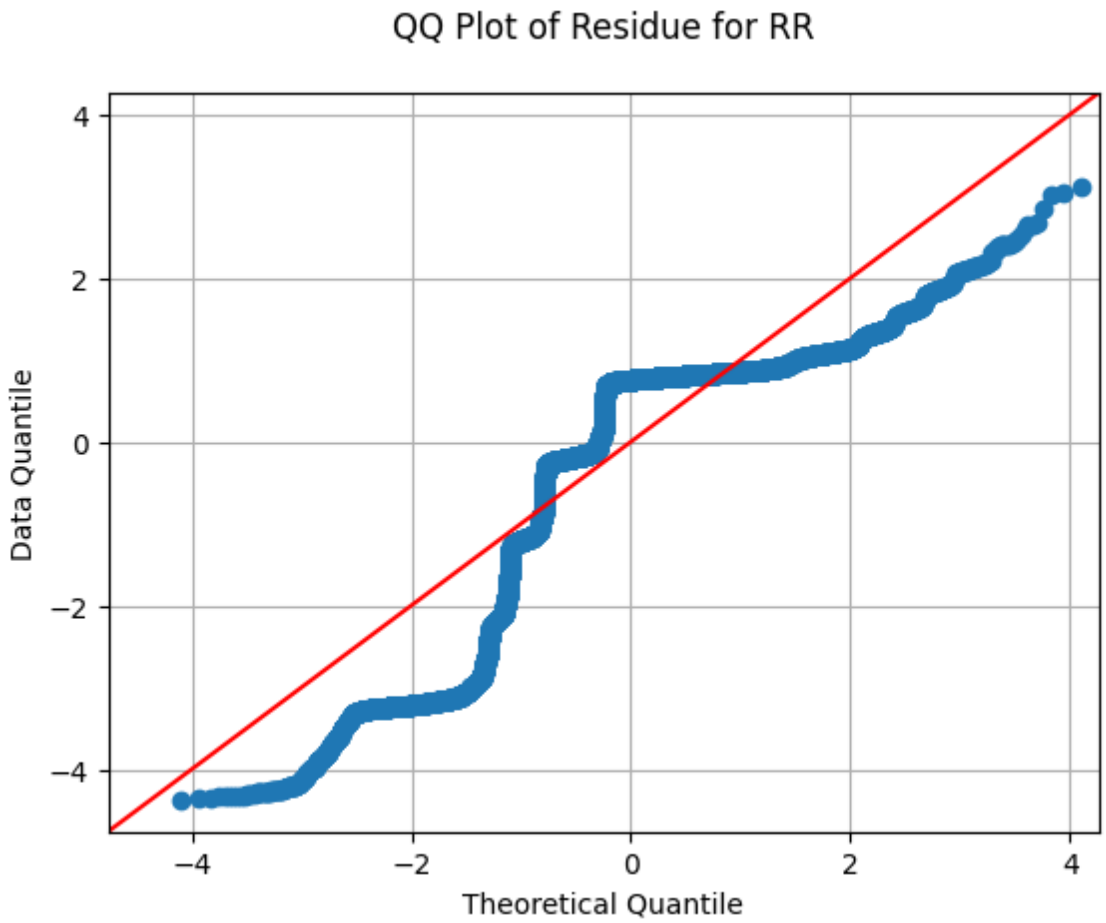
Additional Work

Based on the provided dataset, the analysis focuses on two states: São Paulo and Roraima. The objective of the experiment is to understand how the variables within each state explain the variability of the dependent variable (review score).

Based on the metric below, XGBRegressor would be most suitable among the selection as it explains 22.3% of the dependent variable.

São Paulo

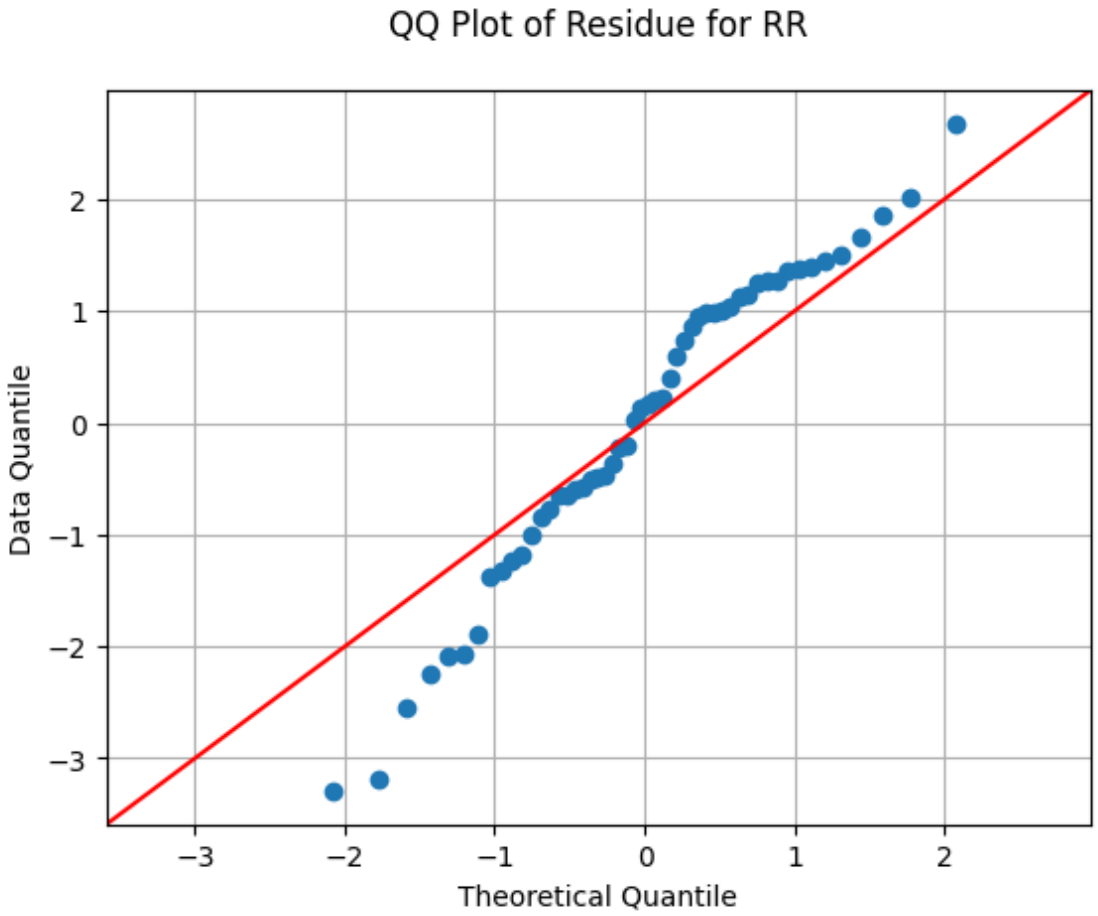
Model	**Mean Squared Error**	**R-squared:**	**Time Taken (s)**
XGBRegressor	1.337	0.223	0.394
RandomForestRegressor	1.443	0.161	30.174
DecisionTreeRegressor	2.674	-0.555	0.423



Based on the metric below, XGBRegressor would be most suitable among the selection as it explains 26.4% of the dependent variable.

Roraima

Model	Mean Squared Error	R-squared:	Time Taken (s)
XGBRegressor	1.802	0.264	0.034
RandomForestRegressor	1.808	0.261	0.133
DecisionTreeRegressor	1.909	0.220	0.005



References

1. Ahn, T., Ryu, S., & Han, I. (2004). The impact of the online and offline features on the user acceptance of Internet shopping malls. Electronic Commerce Research and Applications, 3(4), 405–420.
2. Akboğa, Ö. and Baradan, S. (2017). Safety in ready mixed concrete industry: descriptive analysis of injuries and development of preventive measures. Industrial Health, 55(1), 54-66. https://doi.org/10.2486/indhealth.2016-0083
3. Burkhard, R. A. (2004). Learning from architects: the difference between knowledge visualization and information visualization. IEEE Xplore. https://doi.org/10.1109/IV.2004.1320194
4. Demoulin, N. and Djelassi, S. (2013). customer responses to waits for online banking service delivery. International Journal of Retail & Distribution Management, 41(6), 442-460. https://doi.org/10.1108/09590551311330825
5. Hernández, B., Jiménez, J., & Martín, M. J. (2009). Key website factors in e-business strategy. International Journal of Information Management, 29(5), 362-371. https://doi.org/10.1016/j.ijinfomgt.2008.12.006.
6. International Trade Administration. (n.d.). Impact of COVID pandemic on eCommerce. Retrieved May 17, 2024, from https://www.trade.gov/impact-covid-pandemic-ecommerce.
7. Janjevic, M. and Ndiaye, A. B. (2014). Development and application of a transferability framework for micro-consolidation schemes in urban freight transport. Procedia - Social and Behavioral Sciences, 125, 284-296. https://doi.org/10.1016/j.sbspro.2014.01.1474
8. Laursen, G. H. N., & Thorlund, J. (2018). Business Analytics for Managers : Taking Business Intelligence beyond Reporting (Second Edition). Hoboken, New Jersey: Wiley.
9. Lee, H., Choi, S. Y., & Kang, Y. S. (2009). Formation of e-satisfaction and repurchase intention: moderating roles of computer self-efficacy and computer anxiety. Expert Systems With Applications, 36(4), 7848-7859. https://doi.org/10.1016/j.eswa.2008.11.005
10. Lin, C. C., Wu, H. Y., & Chang, Y. F. (2011). The critical factors impact on online customer satisfaction. Procedia Computer Science, 3(1), 276–281.

11. Lin, Y., Luo, J., Cai, S., Ma, S., & Rong, K. (2016). Exploring the service quality in the e-commerce context: a triadic view. Industrial Management & Data Systems, 116(3), 388–415. https://doi.org/10.1108/imds-04-2015-0116
12. Liu, Y., Wan, Y., Shen, X., Ye, Z., Wen, J (2021). Product Customer Satisfaction Measurement Based on Multiple Online Consumer Review Features. Information 2021, 12, 234. https://doi.org/10.3390/info12060234
13. Mentzer, J. T., D. J. Flint and G. T. M. Hult (2001). Logistics Service Quality as a Segment-Customized Process. Journal of Marketing 65 (4): 82–104. https://doi:10.1509/jmkg.65.4.82.18390
14. Mentzer, J. T., R. Gomes and R. E. Krapfel Jr. (1989). Physical Distribution Service: a Fundamental Marketing Concept? Journal of the Academy of Marketing Science 17 (1): 53–62. https://doi:10.1177/009207038901700107.
15. Nah, F. F.-H., & Delgado, S. (2006). Critical Success Factors for Enterprise Resource Planning Implementation and Upgrade. Journal of Computer Information Systems, 46(5), 99–113. https://doi.org/10.1080/08874417.2006.11645928
16. Oliver, R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. Journal of Marketing Research, 17(4), 460–469.
17. Palanisamy, R. (2005). Strategic information systems planning model for building flexibility and success. Industrial Management &Amp; Data Systems, 105(1), 63-81. https://doi.org/10.1108/02635570510575199.
18. Tsai, H. L. and Huang, H. (2007). Determinants of e-repurchase intentions: an integrative model of quadruple retention drivers. Information &Amp; Management, 44(3), 231-239. https://doi.org/10.1016/j.im.2006.11.006
19. Vasić, N., Kilibarda, M., Andrejić, M., & Jović, S. (2020). Satisfaction is a function of users of logistics services in e-commerce. Technology Analysis &Amp; Strategic Management, 33(7), 813-828. https://doi.org/10.1080/09537325.2020.1849610