Big Data Analysis and Project

Assignment 2: Part B (Big Data Analysis)

Possakorn A1873765

July 04, 2023

Part01: Data description

In response to the sophisticated inquiries derived from the preceding report, specifically, "How does the sentiment expressed in customer feedback impact the customer lifetime value in the E-commerce industry, particularly when analysed within each specific customer segment?", there are several essential datasets and key features that directly address this query, as following list below:

- **Dataset01** - customer_df: This fundamental dataset comprises a customer table which collates the unique customer identifiers along with corresponding geolocation details.
- **Dataset02** - order_df: This set represents possible customers and encompasses their order status and associated order tracking dates, embodying the entire customer transaction history.
- **Dataset03** - order_item_df: This dataset contains key significant attributes, including the price of the product and freight value, signifying the overall order value. Furthermore, it facilitates the conjunction with the customer's review table.
- **Dataset04** - customer_review_df: This principal dataset, instrumental in answering the proposed question, comprises review scores and customer feedback.

There exist additional auxiliary tables that amplify the analytical perspective, as per the following list:

- **Dataset05** - geolocation_df: This dataset pertains to geographical coordinates, expressed as latitude and longitude.
- **Dataset06** - payment_df: This dataset consists of the key feature related to pay instalments and prevalent payment methods.
- **Dataset07** - seller_df: This dataset which stores the seller information explicates the connection between the customer's and the seller's location. seller.

In conclusion, data summaries are provided as additional information in an appendix.

Part02: Clustering/Pattern

In the context of the previously refined question detailed in **Part01**, an interesting relationship emerges between customer lifetime value (CLTV) and sentiment, as evident in **Figure01**. In our dataset, no associated cost data is available, hence total sales are utilized as a proxy for customer lifetime value. However, the correlation between the average review score and the customer lifetime value is not robust, and the review scores do not exhibit a notable distinction when churn customers are considered.

As shown in **Figure02**, order performance over time illustrates an increasing sales trend within a specified period. Nevertheless, the rate of repeat orders, quantified as 2.87, is marginally low, indicating potential issues in customer retention and predicting customer attrition. The platform exhibits a high volume of orders from only new users.

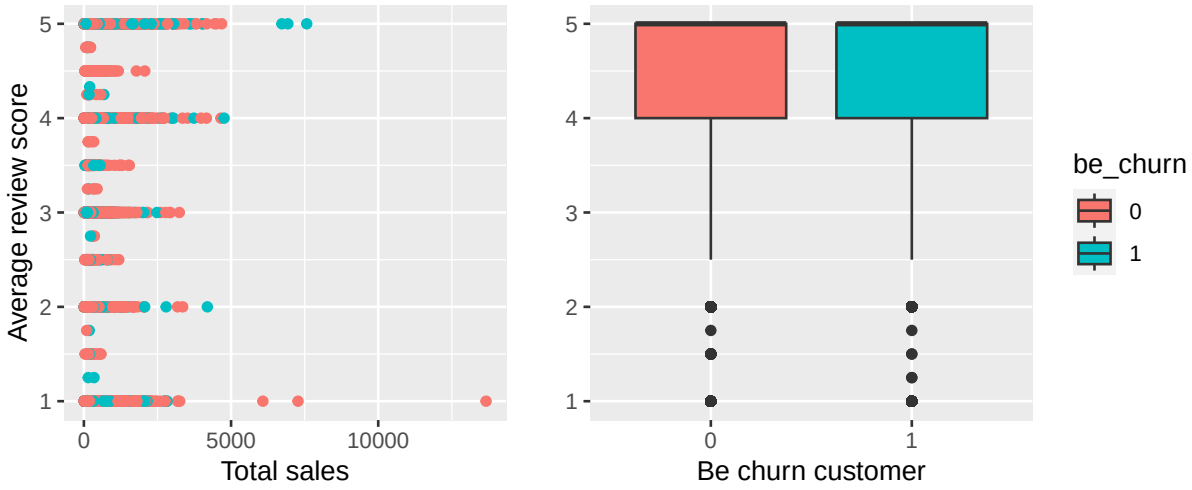


Figure 1: Relationship between customer life time value and average review score

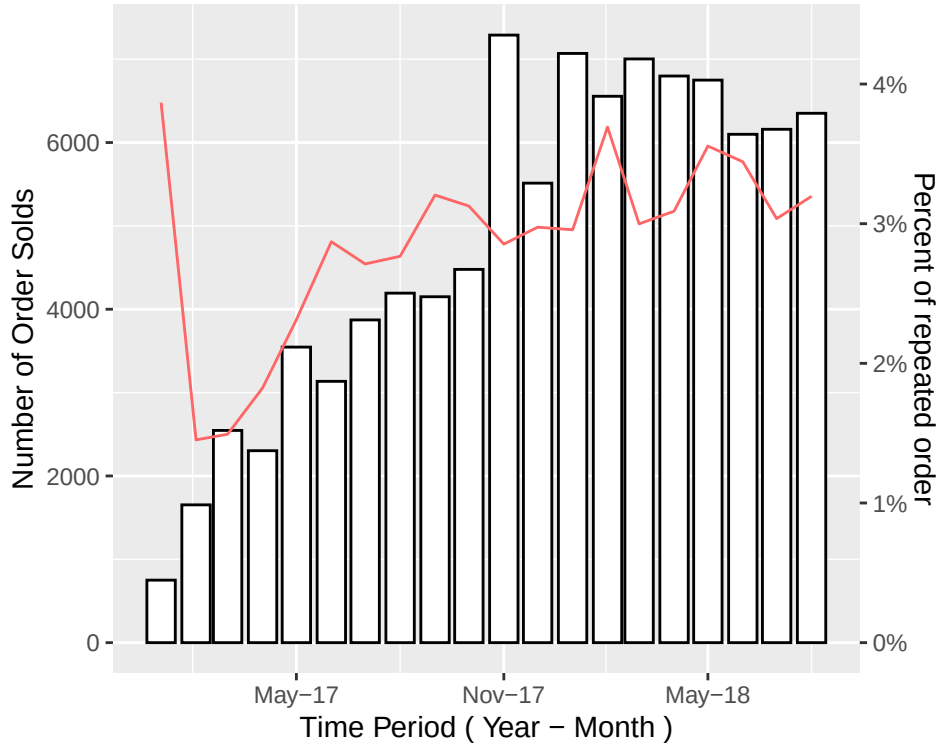


Figure 2: performance of total order (Bar) and percent of repeated order from total over time (Line)

The RFM model, an acronym for Recency, Frequency, and Monetary value, represents specific customer behaviour that correlate slightly with the customer lifetime value. Recency, as shown in **Figure03**, indicates the time elapsed since the last customer activity. The majority of customers exhibit a median of around 220, with the distribution being right-skewed and multimodal.

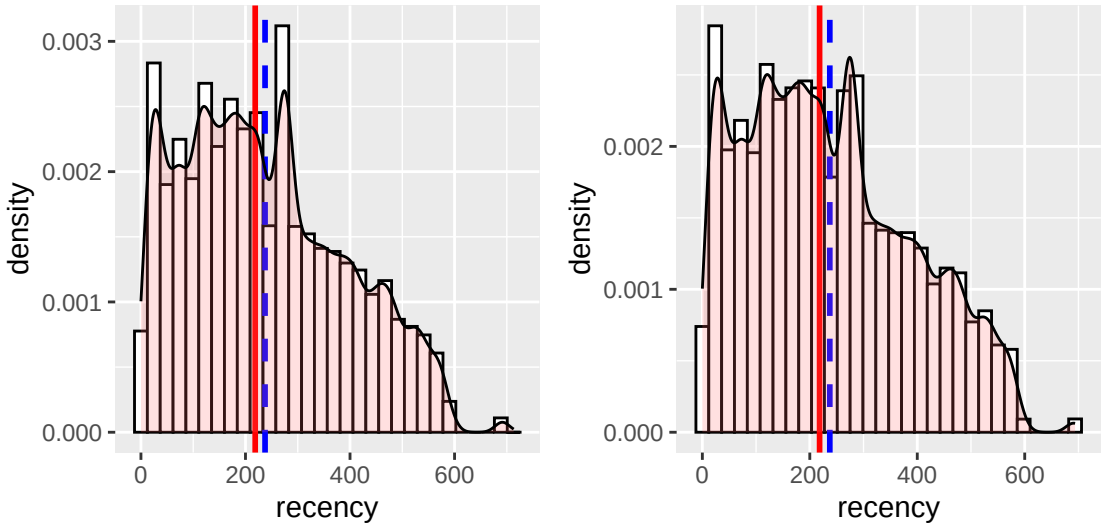


Figure 3: Distribution of recency metrics via histrogram plot - with and without outliers

Frequency, represented in **Figure04**, illustrates customer engagement with our product. The plot reveals the dominance of one-time order customers, corroborating the trend highlighted in earlier plots.

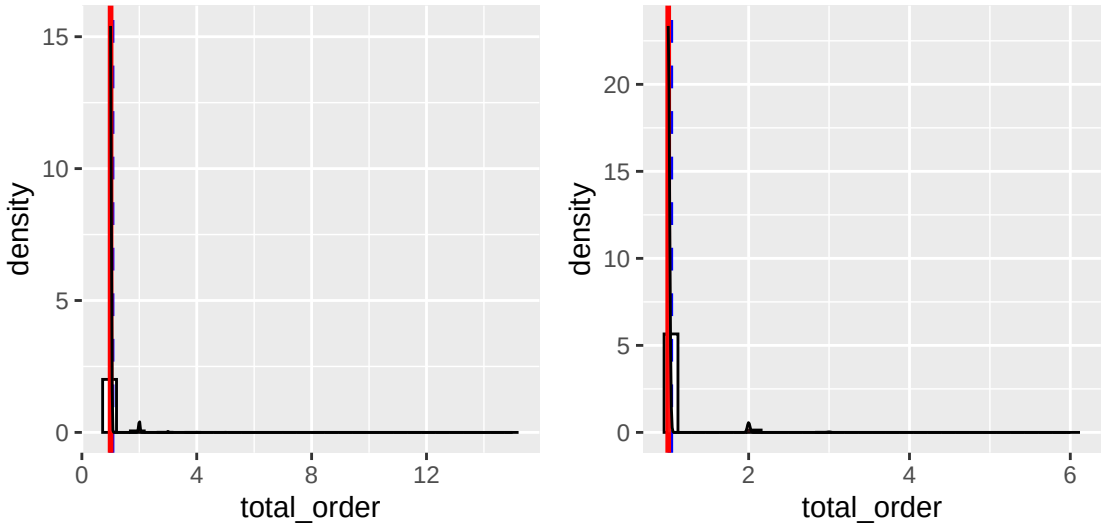


Figure 4: Distribution of frequency metrics via histrogram plot - with and without outliers

Monetary value, shown in **Figure05**, reflects the customer-generated revenue through the platform. However, there are some extreme outliers presented through datasets. After removing the outliers following the interquartile range to detect outliers, the median sales for specific customers, centred around 100, reveal a unimodal right-skewed distribution.

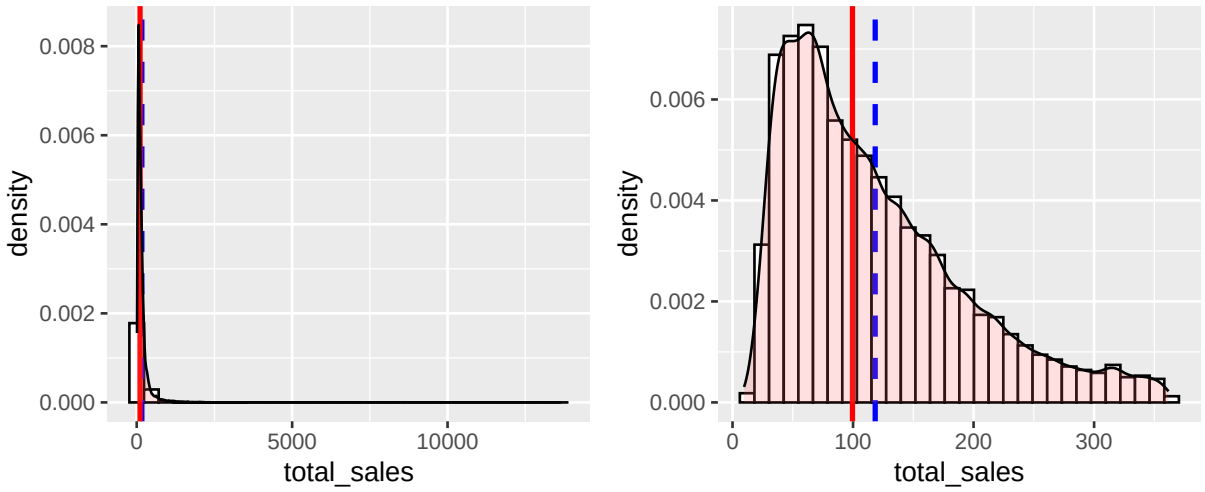
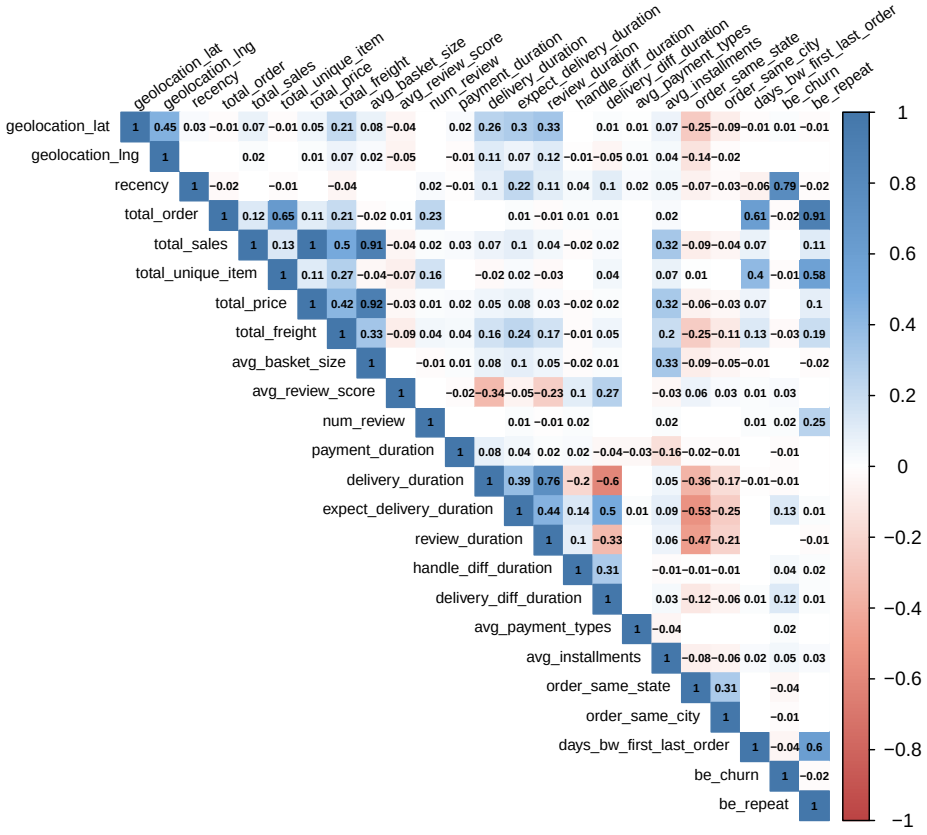


Figure 5: Distribution of monetary metrics via histrogram plot - with and without outliers

Based on the previous plots, there is no significant pattern between total sales and review scores. Nevertheless, potential correlation and issues concerning repeat orders and customer churn warrant further exploration. As such, Matrix chart represents a correlation plot that is instrumental in visualizing and analyzing the interrelationships among variables within the relevant dataset. It measures the statistical association between two variables, thereby assisting in understanding how these potential features change. This correlation matrix is essential for feature selection as variables with high correlation may provide superfluous information.



Correlation plots can determine the significance of variables in predicting the target variable(S and I, 2018). Variables with higher correlation coefficients are more likely to substantially impact the target variable, considered integral in modeling. **Figure06** demonstrates that total sales and average payment instalments exhibit a weak relationship with no discernible pattern with respect to churn customers. Conversely, total prices, total freight, and average basket size are highly correlated variables, offering redundant information following the application of feature engineering methods.

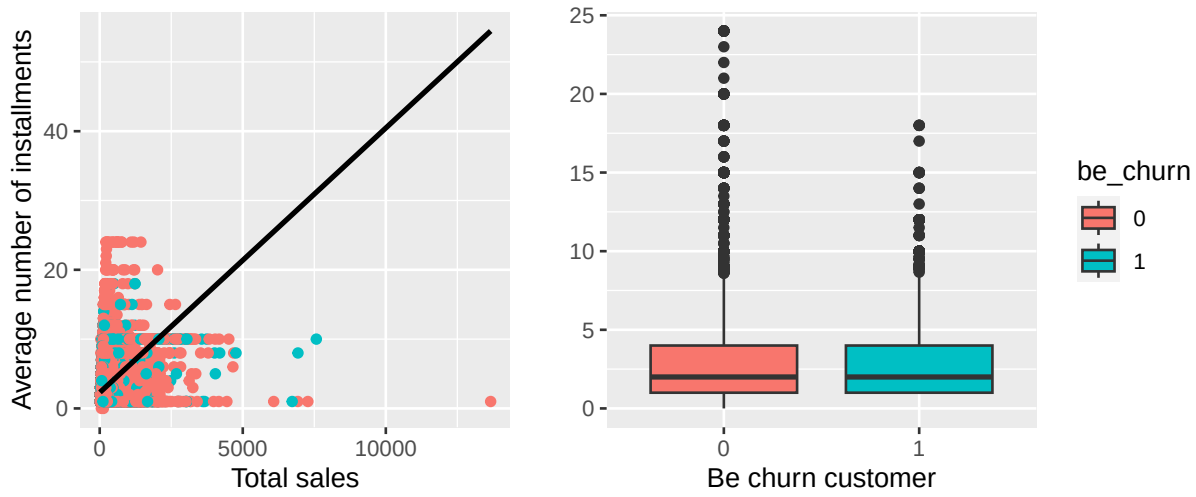


Figure 6: Relationship between total sales and average installments

Following plots of the relationship between total sales and expected delivery duration as shown in **Figure07**, there does not demonstrate a significant correlation between these variables. Although, there is weak relationship between respect to churn customers and expected delivery duration can be discerned.

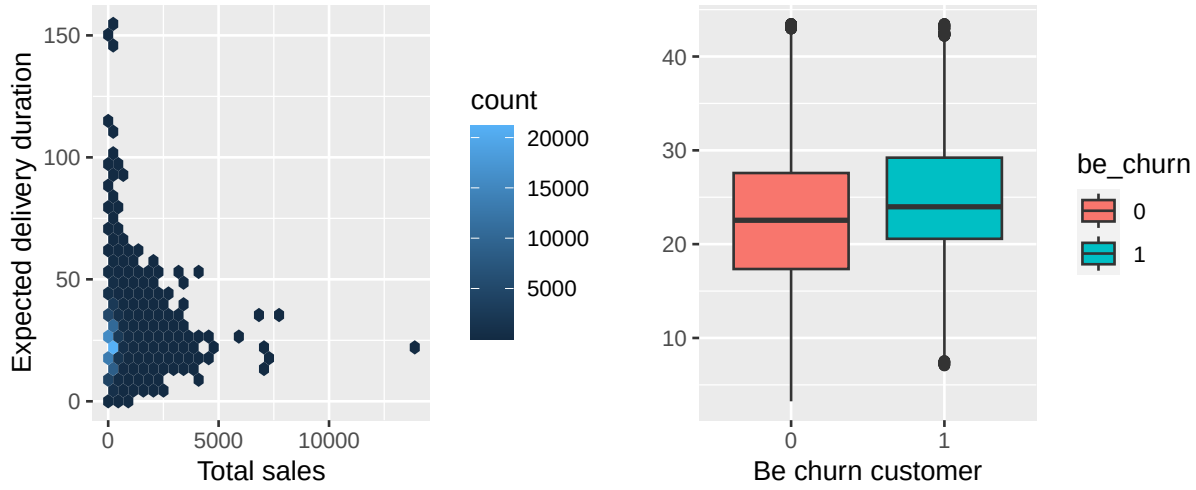


Figure 7: Relationship between total sales and expected delivery duration

Part03: Visualisation

< all visualizations present in **Part02** to identify the pattern and relationship crossing data sets >

Part04: Problem refinement

Previous Problem statement: "How does the sentiment expressed in customer feedback impact the customer lifetime value in the E-commerce industry, particularly when analysed within each specific customer segment?"
As indicated by the patterns and relationships delineated in **Part02**, there appears to be an absence of a significant correlation between customer lifetime value and review scores. Moreover, the data reveal a remarkable concentration of customers who have only placed a single order. However, the platform records an extremely low percentage of repeat orders alongside a substantial number of churn customers. Therefore, our attention towards the effects of platform information on the churn rate.
On the other hand, there are some limitations of perspectives by lacking additional data sources including customer demographics, and web and app session logs. In summary, the new refined question will be "Which key platform features have the potential to most effectively optimize customer churn?".

Appendix

Reference will locate at the end of this report after appendix section.

Appendix A: Data Summary

Table: order_df

Table 1: Data summary

Name	Piped data
Number of rows	99441
Number of columns	5
Column type frequency:	
POSIXct	5
Group variables	None

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
order_purchase_timestamp	0	1.00	2016-09-04 21:15:19	2018-10-17 17:30:18	2018-01-18 23:04:36	98875
order_approved_at	160	1.00	2016-09-15 12:16:38	2018-09-03 17:40:06	2018-01-19 11:36:13	90733
order_delivered_carrier_date	1783	0.98	2016-10-08 10:34:01	2018-09-11 19:48:28	2018-01-24 16:10:58	81018
order_delivered_customer_date	2965	0.97	2016-10-11 13:46:32	2018-10-17 13:22:46	2018-02-02 19:28:10	95664
order_estimated_delivery_date	0	1.00	2016-09-30 00:00:00	2018-11-12 00:00:00	2018-02-15 00:00:00	459

Table: customer_df

Table 3: Data summary

Name	Piped data
Number of rows	99441
Number of columns	5
Column type frequency:	
character	5
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
customer_id	0	1	32	32	0	99441	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
customer_unique_id	0	1	32	32	0	96096	0
customer_zip_code_prefix	0	1	5	5	0	14994	0
customer_city	0	1	3	32	0	4119	0
customer_state	0	1	2	2	0	27	0

Table: order_item_df

Table 5: Data summary

Name	Piped data
Number of rows	112650
Number of columns	3
Column type frequency:	
numeric	3
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
order_item_id	0	1	1.20	0.71	1.00	1.00	1.00	1.00	21.00
price	0	1	120.65	183.63	0.85	39.90	74.99	134.90	6735.00
freight_value	0	1	19.99	15.81	0.00	13.08	16.26	21.15	409.68

Table: customer_review_df

Table 7: Data summary

Name	Piped data
Number of rows	99224
Number of columns	3
Column type frequency:	
character	2
numeric	1
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
review_comment_title	87658	0.12	1	26	0	4178	0
review_comment_message	58256	0.41	1	208	0	35743	18

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
review_score	0	1	4.09	1.35	1	4	5	5	5

Table: customer_review_df

Table 10: Data summary

Name	Piped data
Number of rows	1000163
Number of columns	2
Column type frequency:	
numeric	2
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
geolocation_lat	0	1	-21.18	5.72	-36.61	-23.60	-22.92	-19.98	45.07
geolocation_lng	0	1	-46.39	4.27	-101.47	-48.57	-46.64	-43.77	121.11

Table: payment_df

Table 12: Data summary

Name	Piped data
Number of rows	103886
Number of columns	2
Column type frequency:	
factor	1
numeric	1
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
payment_type	0	1	FALSE	5	cre: 76795, bol: 19784, vou: 5775, deb: 1529

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
payment_installments	0	1	2.85	2.69	0	1	1	4	24

Table: seller_df

Table 15: Data summary

Name	Piped data
Number of rows	3095
Number of columns	2
Column type frequency:	
character	2
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
seller_city	0	1	2	40	0	611	0
seller_state	0	1	2	2	0	23	0

Reference

S, K. and I, C. (2018) *Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states.* Int J Environ Res Public Health. Available at: https://doi.org/10.3390/ijerph15122907.