Classification ML Model

Predictive Analysis for Customer Retention in Brazilian E-Commerce

André Teixeira

Index

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Conclusions
- Business Recommendations

Business Understanding


novo pedido
amazon
americanas
Magalu
olist

Problem Definition

Predictability of customer loyalty to the e-commerce

Project Objective

Making predictions about customer repurchase trends using ML

Marketplace Benchmarking

- [1] Customers who make repeat purchases buy 67% more than first-time buyers.

- [1] For marketplace sellers, the repeat purchase ratio is 20-40%

- [2] The retention rate for Walmart and Target over 16 months is 14%, for Temu it's 28%, and for Amazon it's 56%.

Sources:
[1] Salesduo, [2] Earnest Analytics

Marketplace Benchmarking

- [3] Approximately 90% of transactions involve disintermediation after the first purchase.
- [4] Less priority on delivery speed and more on delivery reliability.
- [4] Importance of free shipping and discounts, and retention programs.

Sources:
[3] Zhu et al. 2018, [4] Mckinsey

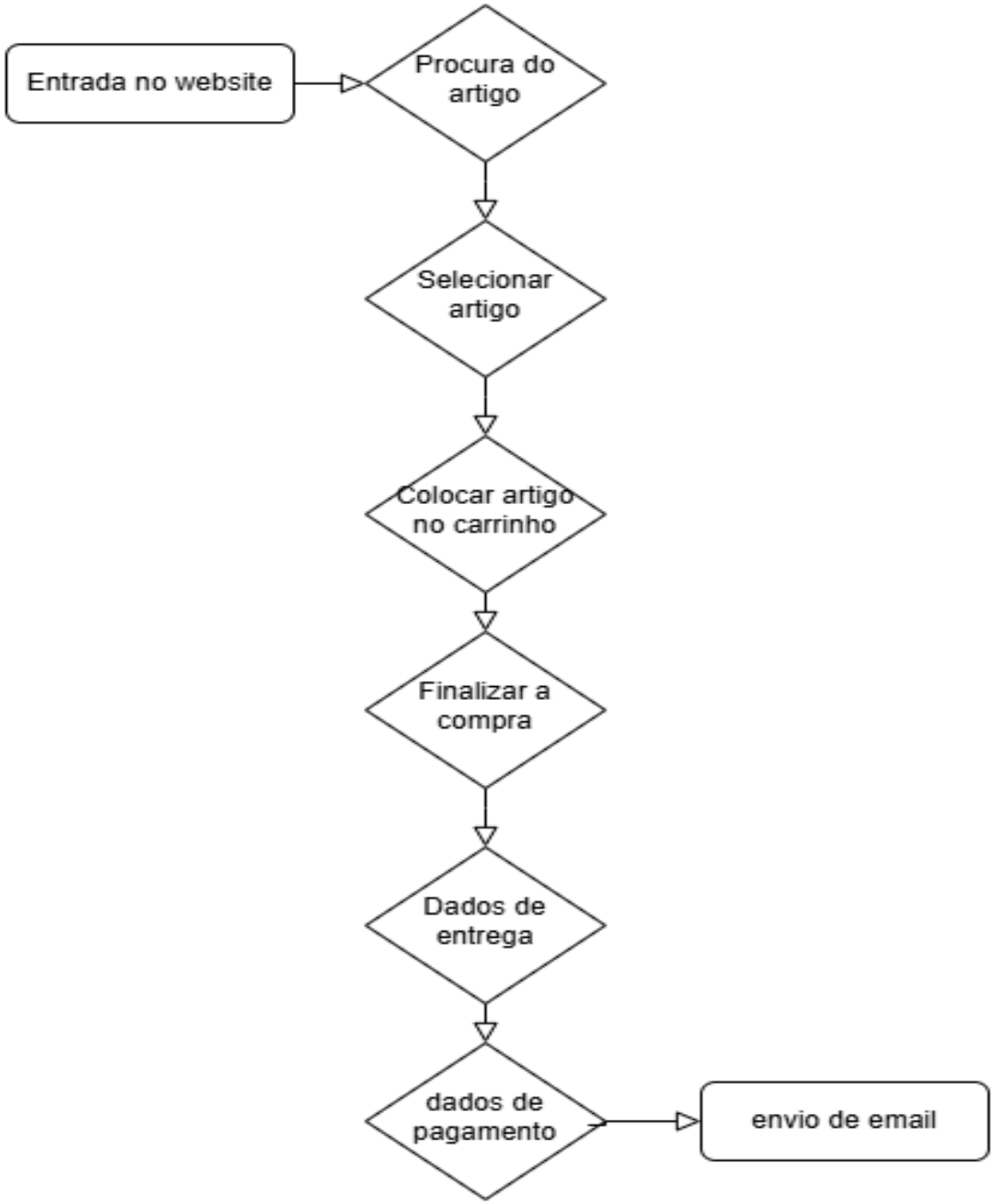
[5] Mercado Livre: Brazilian Marketplace Example

Good coupon policy.

Good customer communication policy.

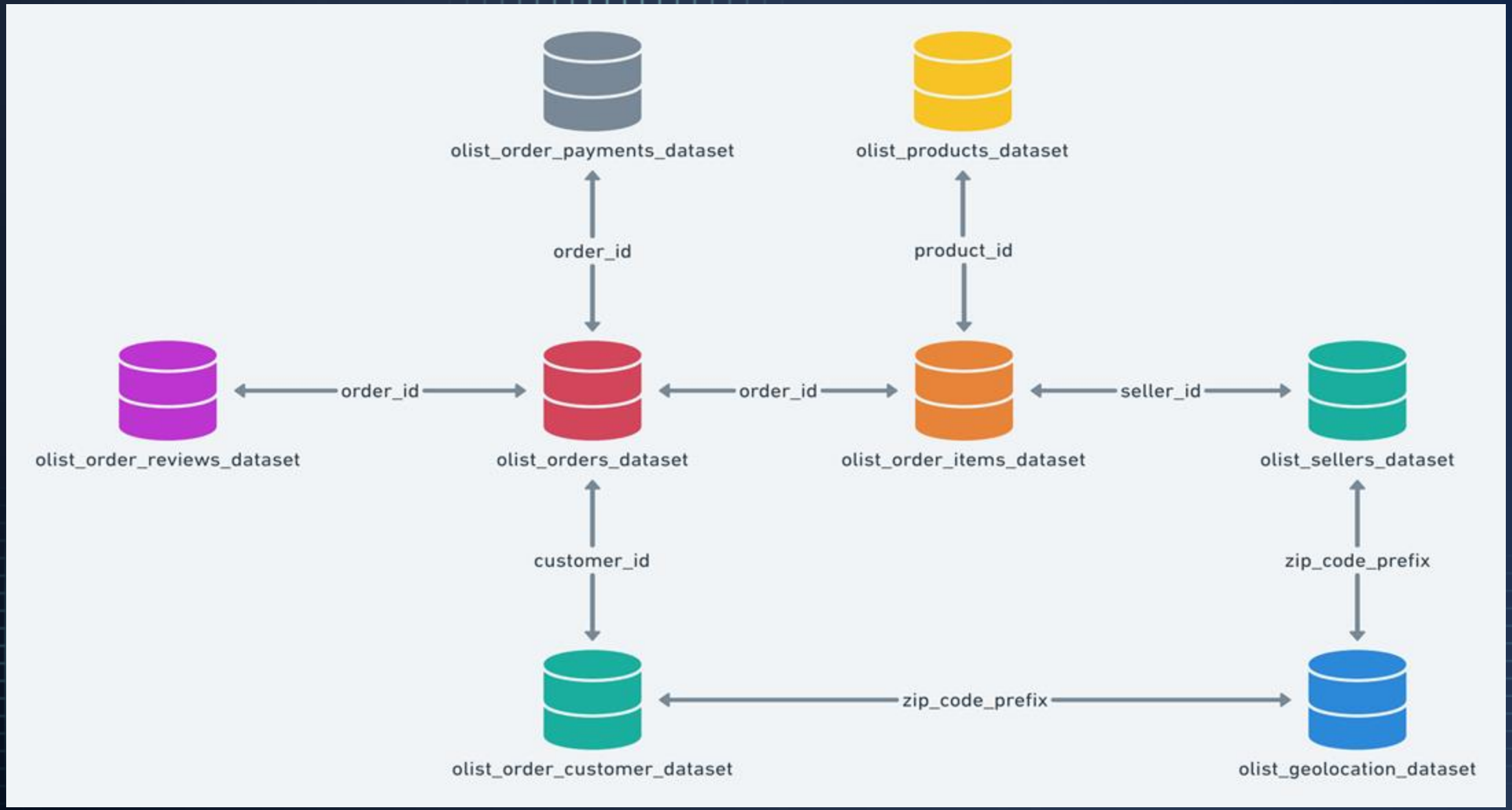
Good free shipping policy.

Sources:
[5] Mercado Livre


Entrada no website
Procura do artigo
Selecionar artigo
Colocar artigo no carrinho
Finalizar a compra
Dados de entrega
dados de pagamento
envio de email

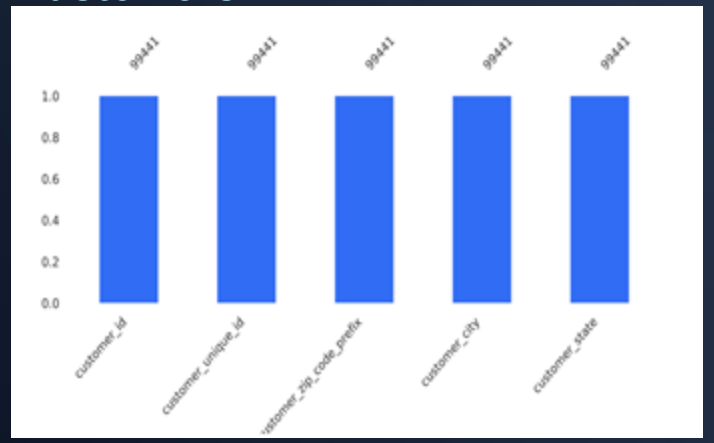
Data Understanding

Olist Database Schema



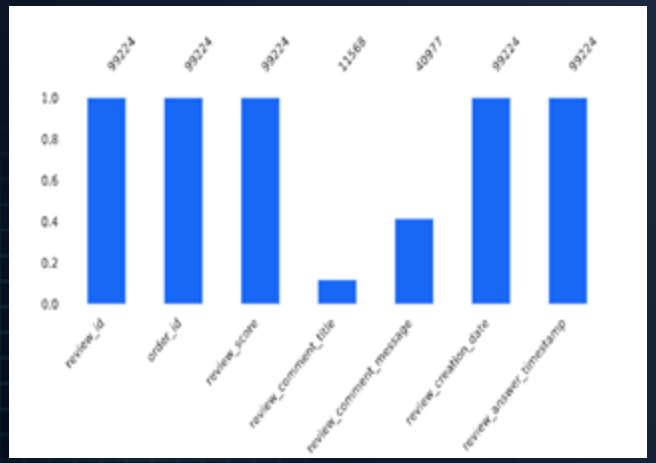
Tables Selection

Customers



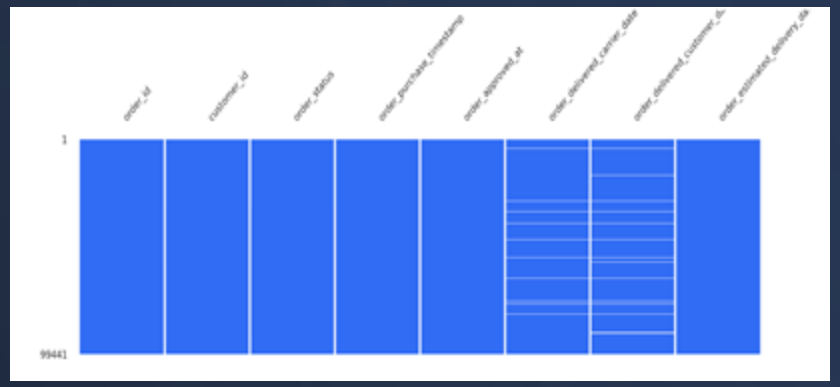
No Missing Values

Reviews



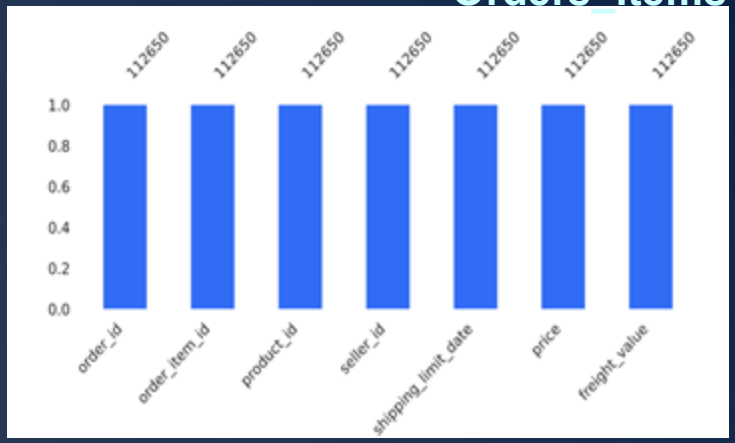
review_comment_title 88.3% missing values
review_comment_message 58.7% missing values

Orders



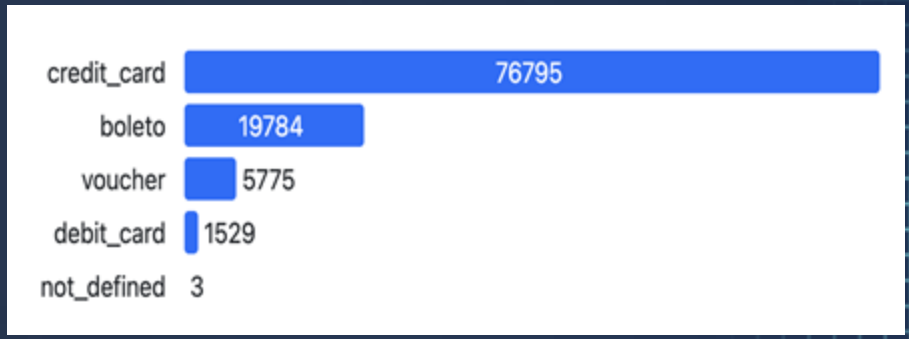
Order_delivered_carrier_date 1.8% missing values
Order_delivered_customer_date 3.0% missing values

Orders_Items



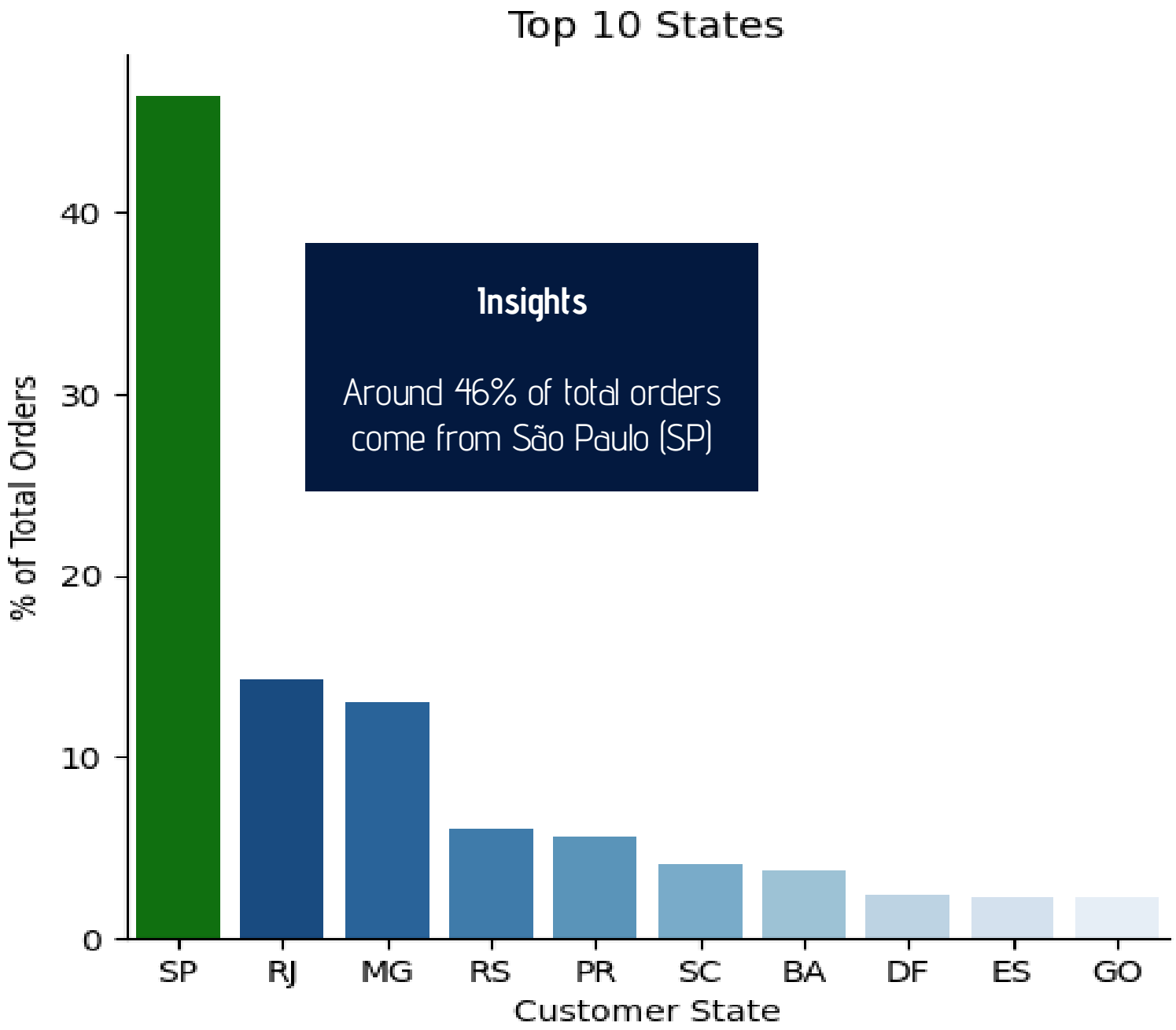
No Missing Values

Payments

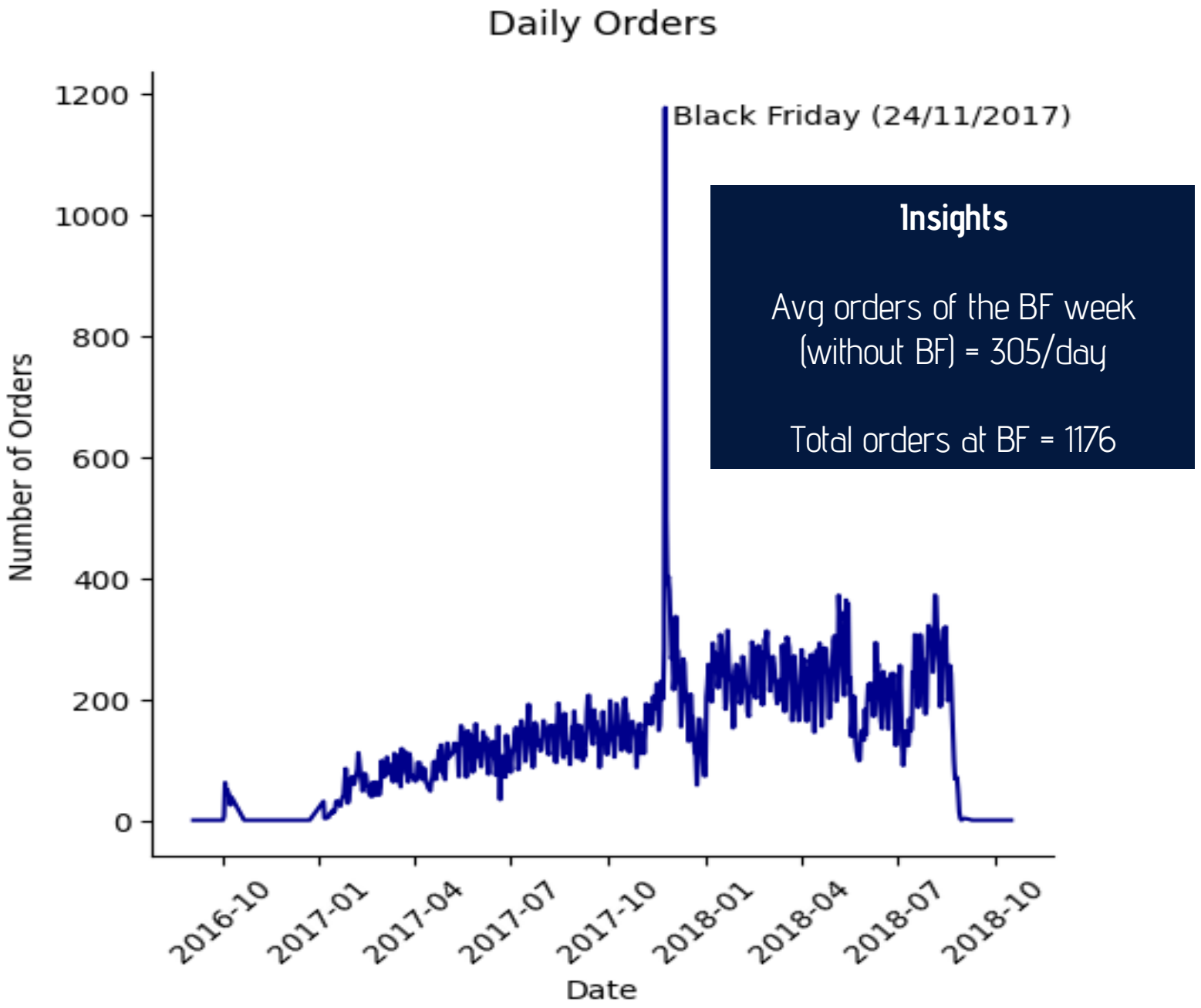


No Missing Values

Initial EDA: Important Insights



Initial EDA: Important Insights

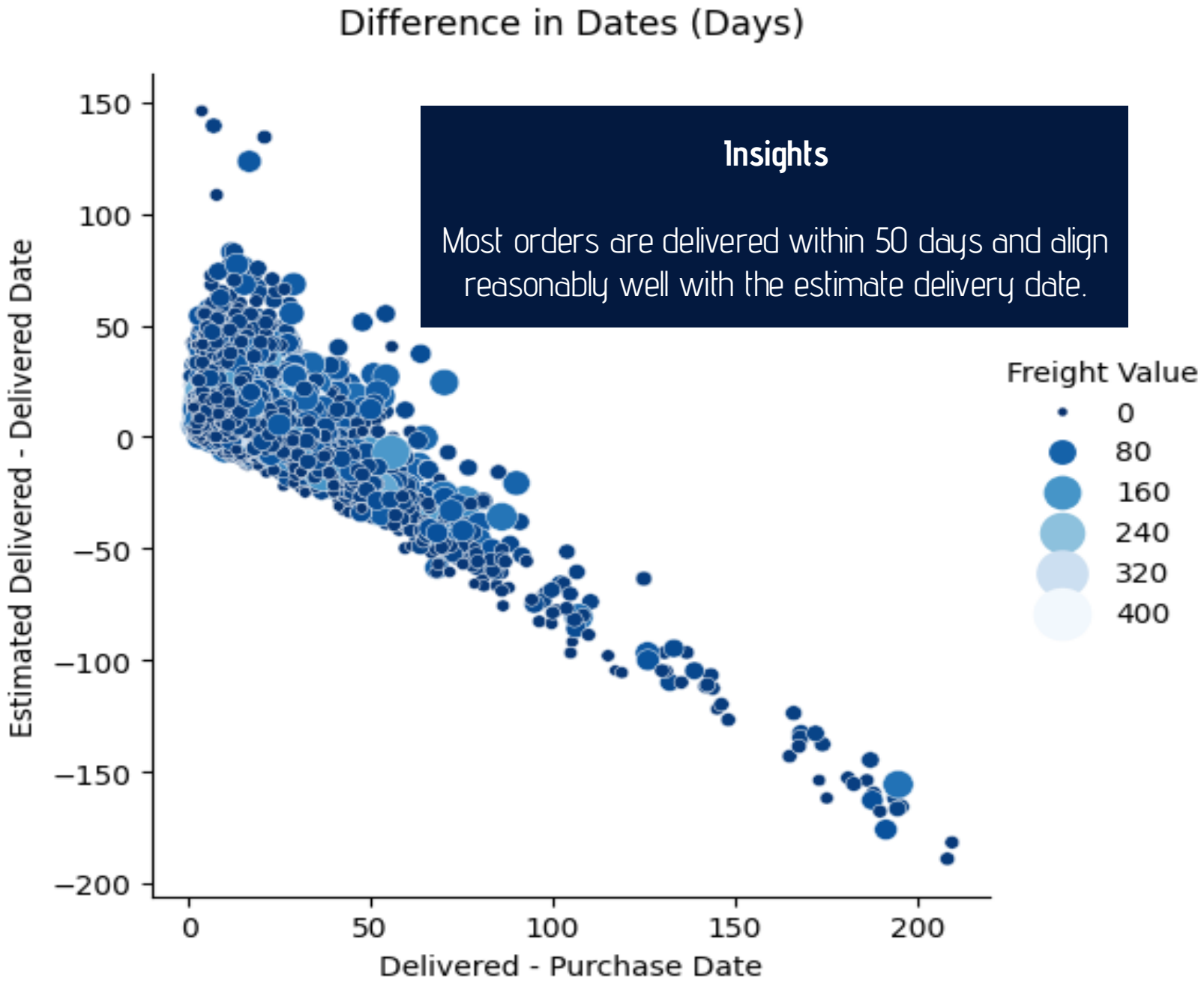


Insights

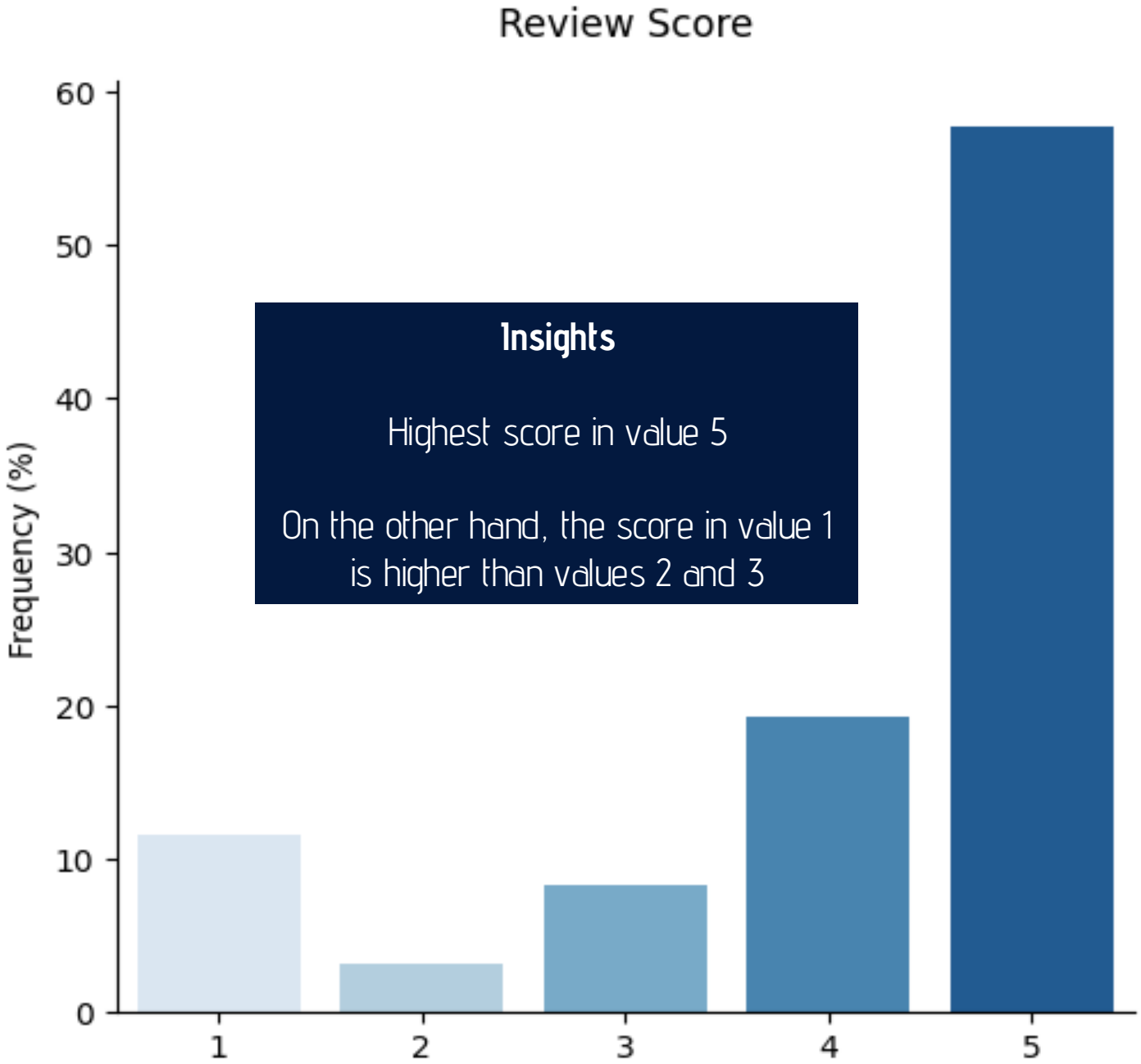
Avg orders of the BF week (without BF) = 305/day

Total orders at BF = 1176

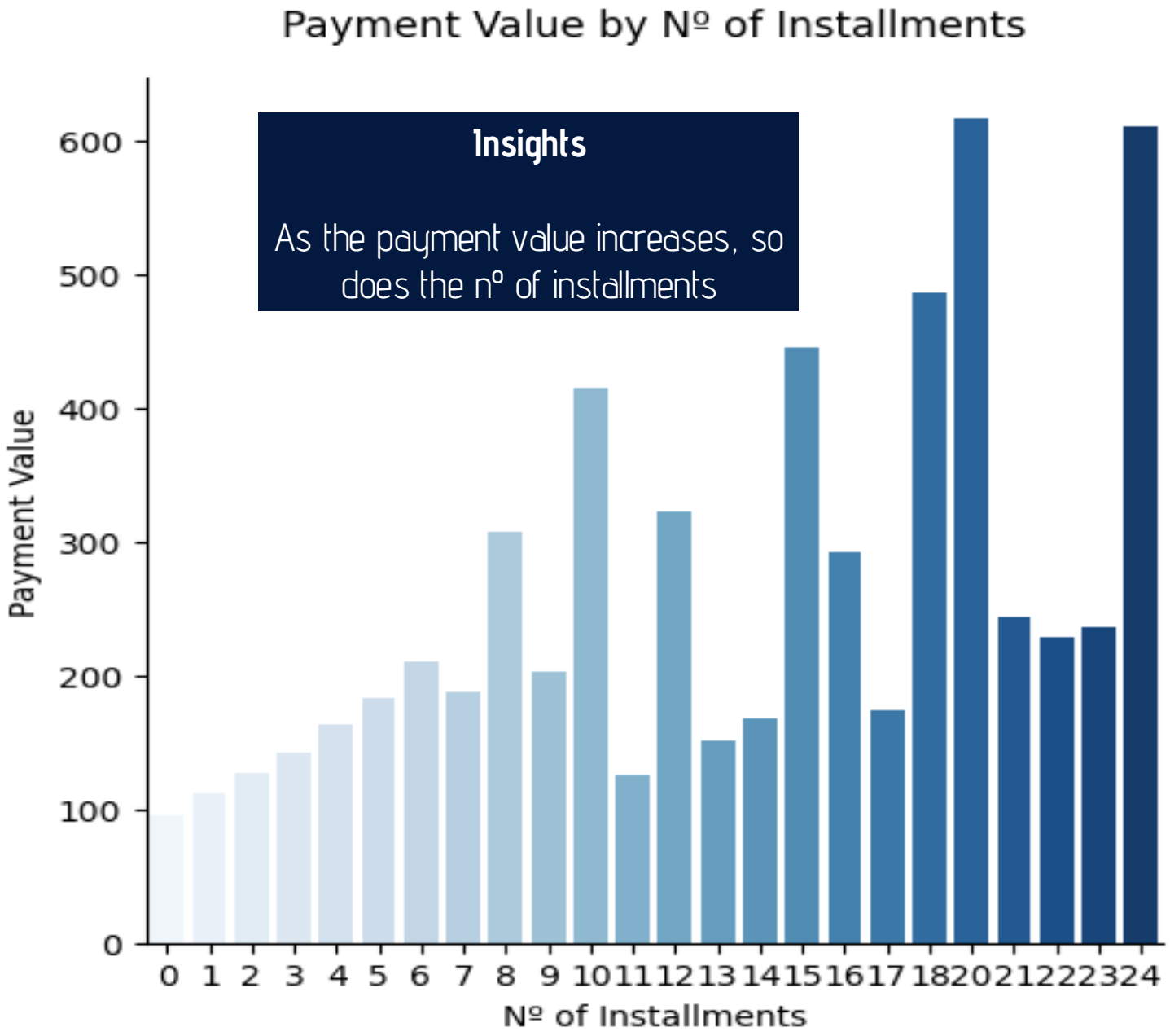
Initial EDA: Important Insights



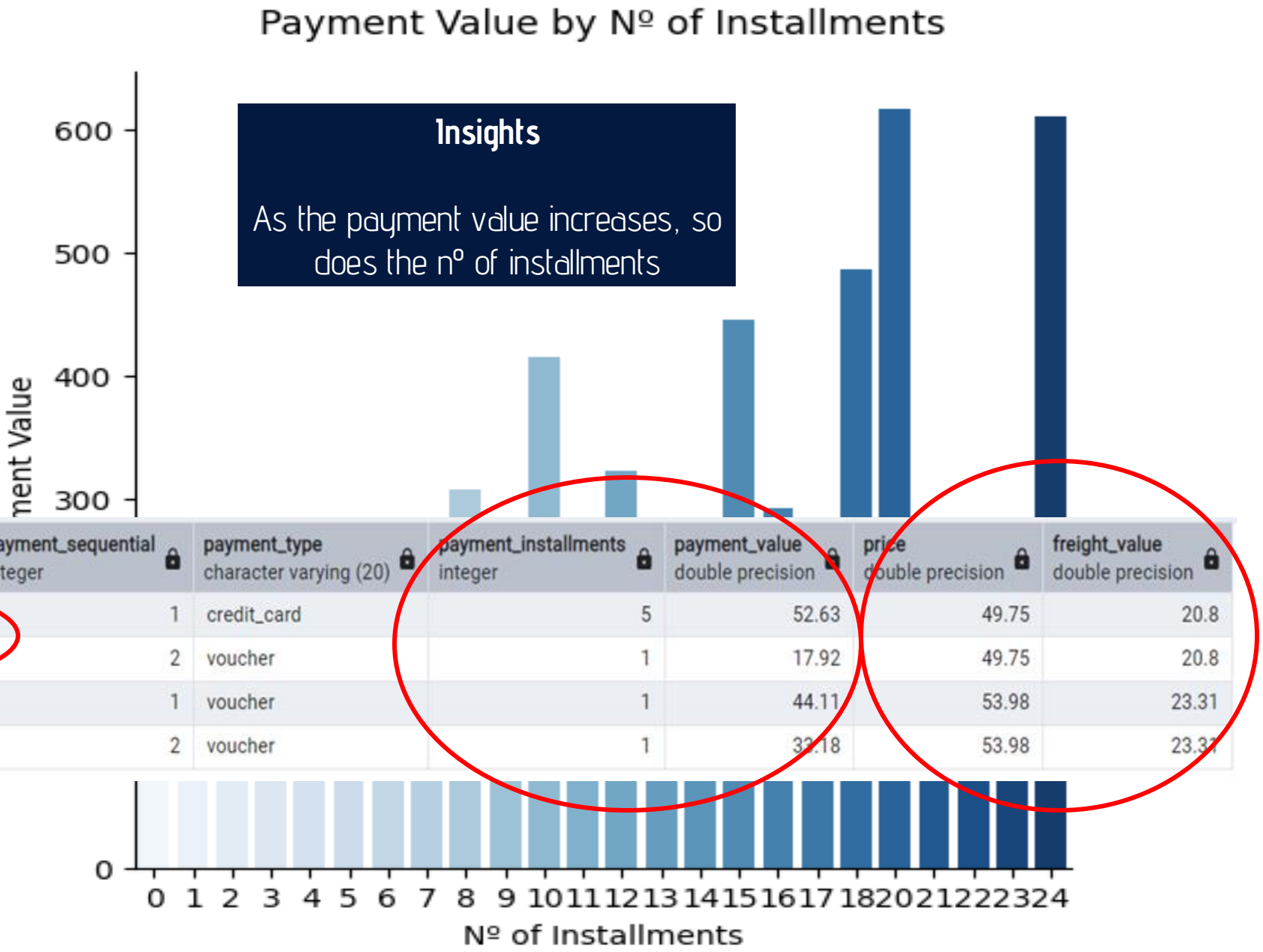
Initial EDA: Important Insights



Initial EDA: Important Insights



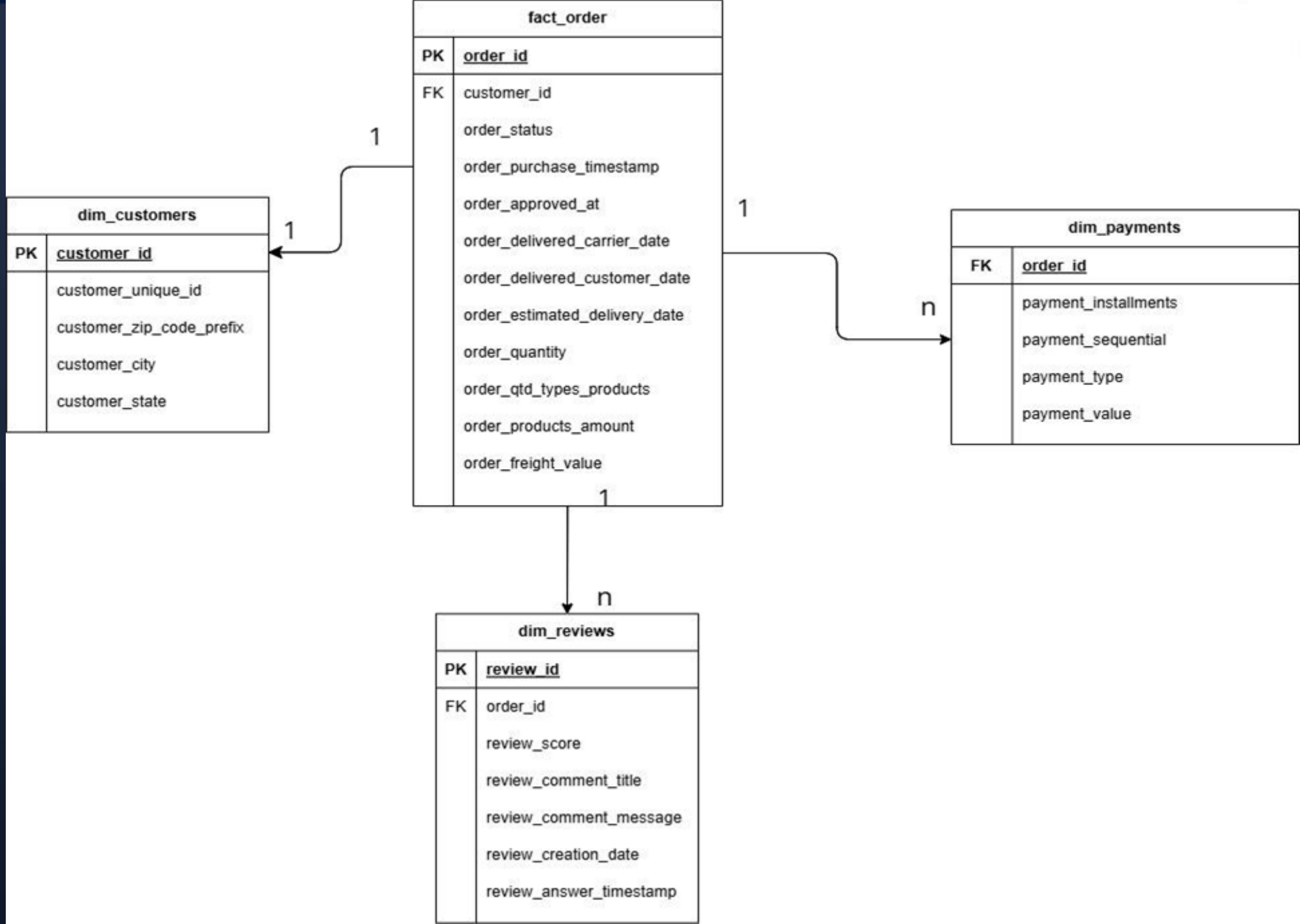
Initial EDA: Important Insights



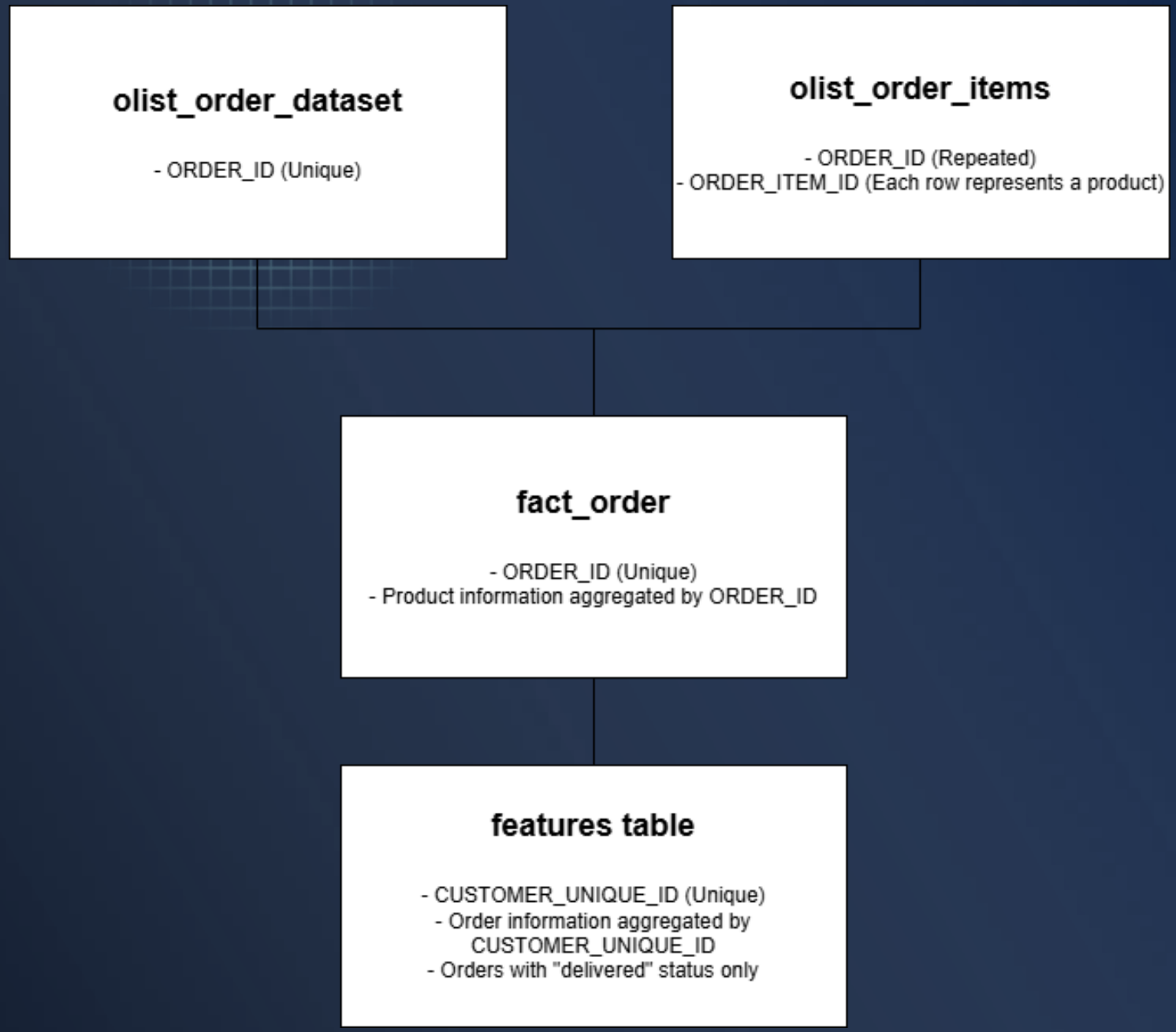
	order_id character varying (50)	payment_sequential integer	payment_type character varying (20)	payment_installments integer	payment_value double precision	price double precision	freight_value double precision
1	0016dfedd97fc2950e388d2971d718c7	1	credit_card	5	52.63	49.75	20.8
2	0016dfedd97fc2950e388d2971d718c7	2	voucher	1	17.92	49.75	20.8
3	002f19a65a2ddd70a090297872e6d64e	1	voucher	1	44.11	53.98	23.31
4	002f19a65a2ddd70a090297872e6d64e	2	voucher	1	32.18	53.98	23.31

Data Preparation

Star Schema Building



Data Hierarchization



TARGET

1- A temporal separation was made in the data, in order to create a 'past' moment and a 'future' moment.

2- We chose the date 05/31/2017 to be the dividing point

3 - We created an SQL query that builds the predict_next_order_delivered table.

4 - The granularity of this table is each customer in the Olist database who made a purchase up to 05/31/2017.

5 - The TARGET variable ('retention') receives the value 0 if the customer who purchased up to 05/31/2017 did not buy again after that date, and the value 1 if they did buy again..

6 - The remaining columns are the features that describe the behavior of these customers.

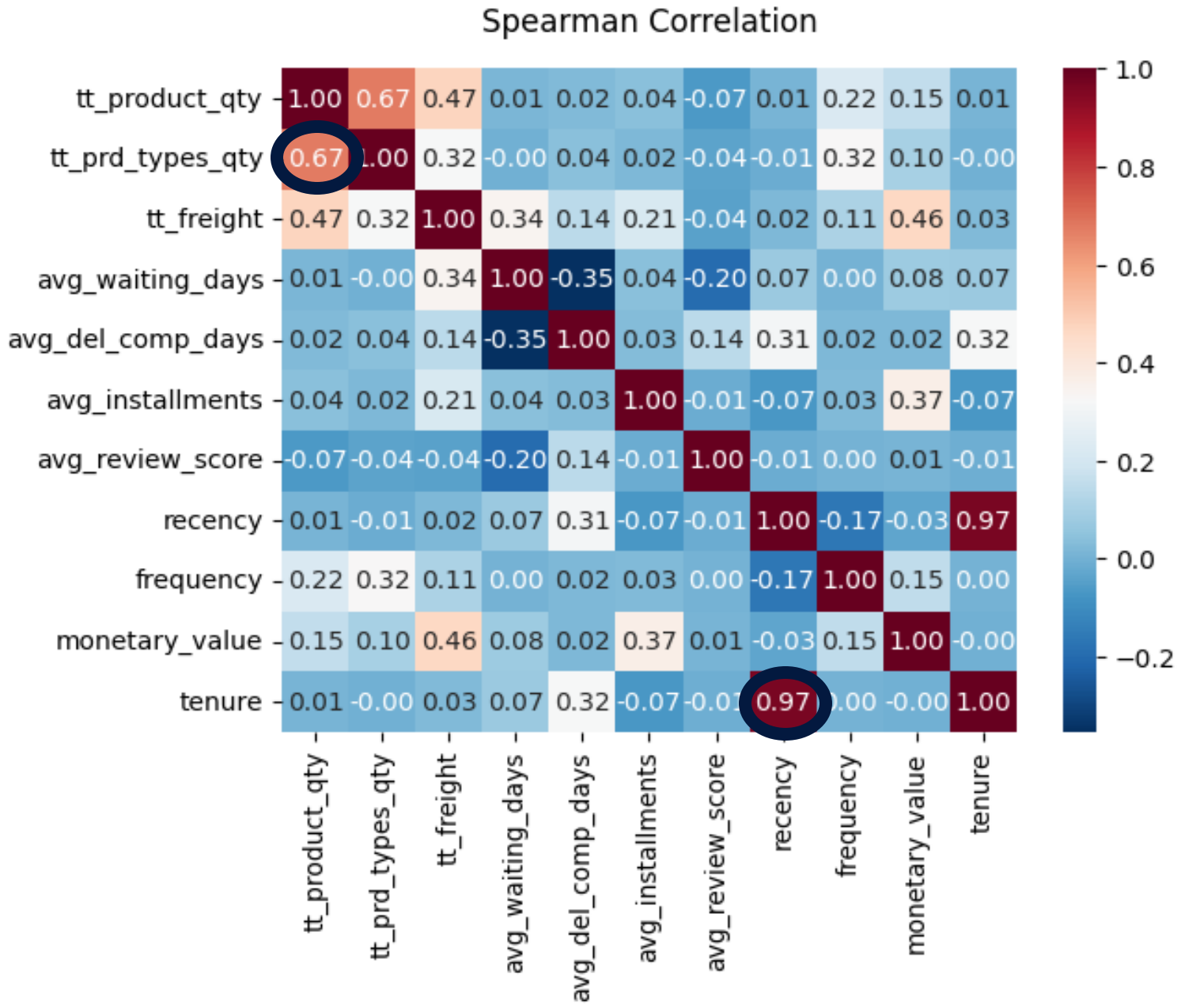
Note: The first order was placed on 09/04/2016 and the last one on 10/17/2018.

Features Table Preview

Customer	Retention	Products	Types of Products	Freight	Avg Waiting Days	Avg Delivery Compliance Days	Avg Installments	Avg_ Review	Day/Night	Week/Weekends	Payment Type	State
client_123	1	5	3	35.00	7	2	2	4.5	day	weekdays	credit_card	sp_rj_mg
client_456	0	2	2	18.50	10	1	1	3.0	night	weekend	boleto	other_state
client_789	1	1	1	12.00	5	3	3	5.0	day	weekdays	credit_card	sp_rj_mg
client_abc	0	3	2	25.00	8	0	1	4.0	day	weekdays	debit_card	other_state

No retention: 97%
Retention: 3%

Correlation for Feature Selection



Modelling

Clustering

Customer Segmentation

K-Means Methodology

Feature Selection

- Various feature combinations across multiple iterations
- Features related to customer characteristics, such as location and spending habits

Data Preprocessing

- One-hot encoding
- Scaling numerical features to the 0-1 range

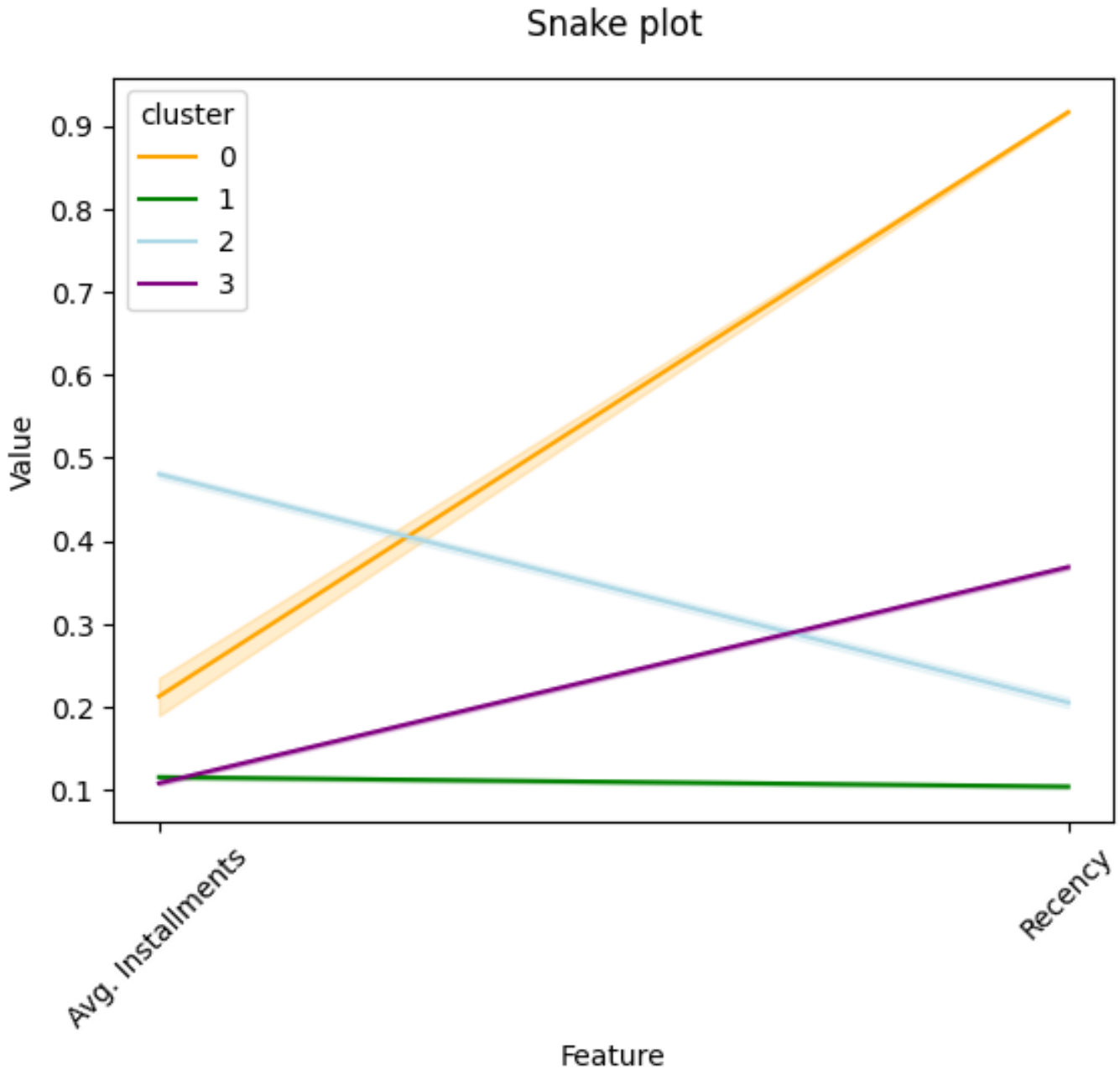
Cluster Analysis

- Elbow method and Silhouette score

Challenges

- High cardinality in categorical variables (K-Modes for future reference)
- Monetary value distribution: Highly skewed

Cluster Analysis



Cluster Analysis

Long-Lapsed Spenders

Cluster 0

Characteristics:

- Smallest segment
- Low spending
- Long inactivity

Marketing Strategy:

- Evaluate if the effort to re-engage them is cost-effective

New Value Spenders

Cluster 1

Characteristics:

- Largest segment
- Recent purchases

Marketing Strategy:

- Onboarding and Welcome campaigns

Premium Installment Spenders

Cluster 2

Characteristics:

- High Spending
- Comfortable with spreading out payments

Marketing Strategy:

- Offer flexible payment options
- Premium products

Reactivation-Ready Spenders

Cluster 3

Characteristics:

- Recent buyers but not brand new

Marketing Strategy:

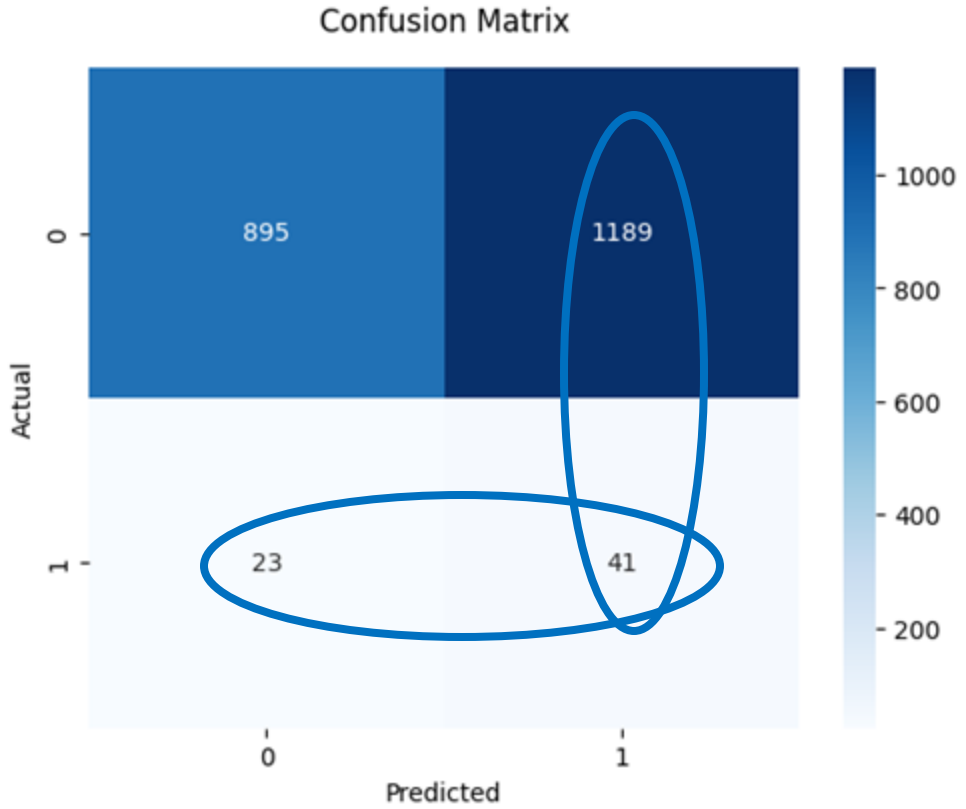
- Re-Engagement campaigns

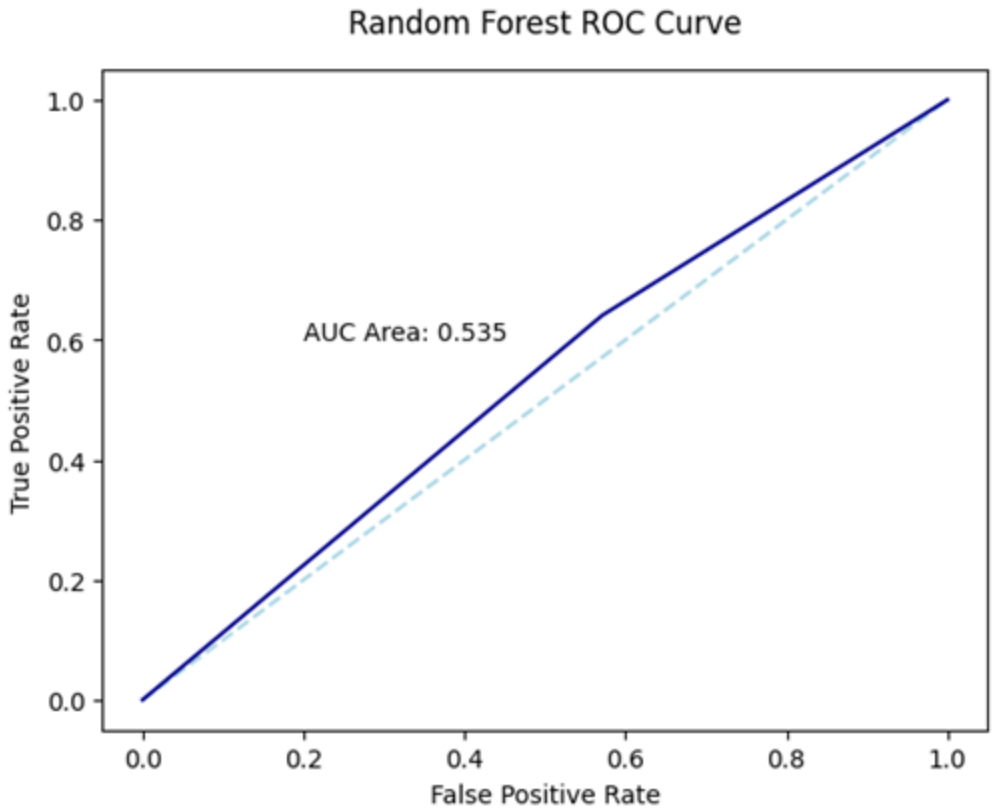
Classification

Random Forest Methodology

Feature Selection	• Various feature combinations across multiple iterations • Features that could impact customer's likelihood to make a repeat purchase
Data Preprocessing	• One-hot encoding • Numerical imputation for average score feature • Pipeline to avoid data leakage
Model Training	• Train/ Test split • RandomSearchCV (K-fold = 10) trained on the training dataset with Recall score and then evaluated on the independent test dataset
Challenges	• Introducing bias through mean and median imputation (Other imputation for future reference) • Imbalanced class

RF Retention Analysis





Train Recall	Test Recall
0.68	0.64

Test Set	Recall	Precision
0	0.43	0.97
1	0.64	0.03

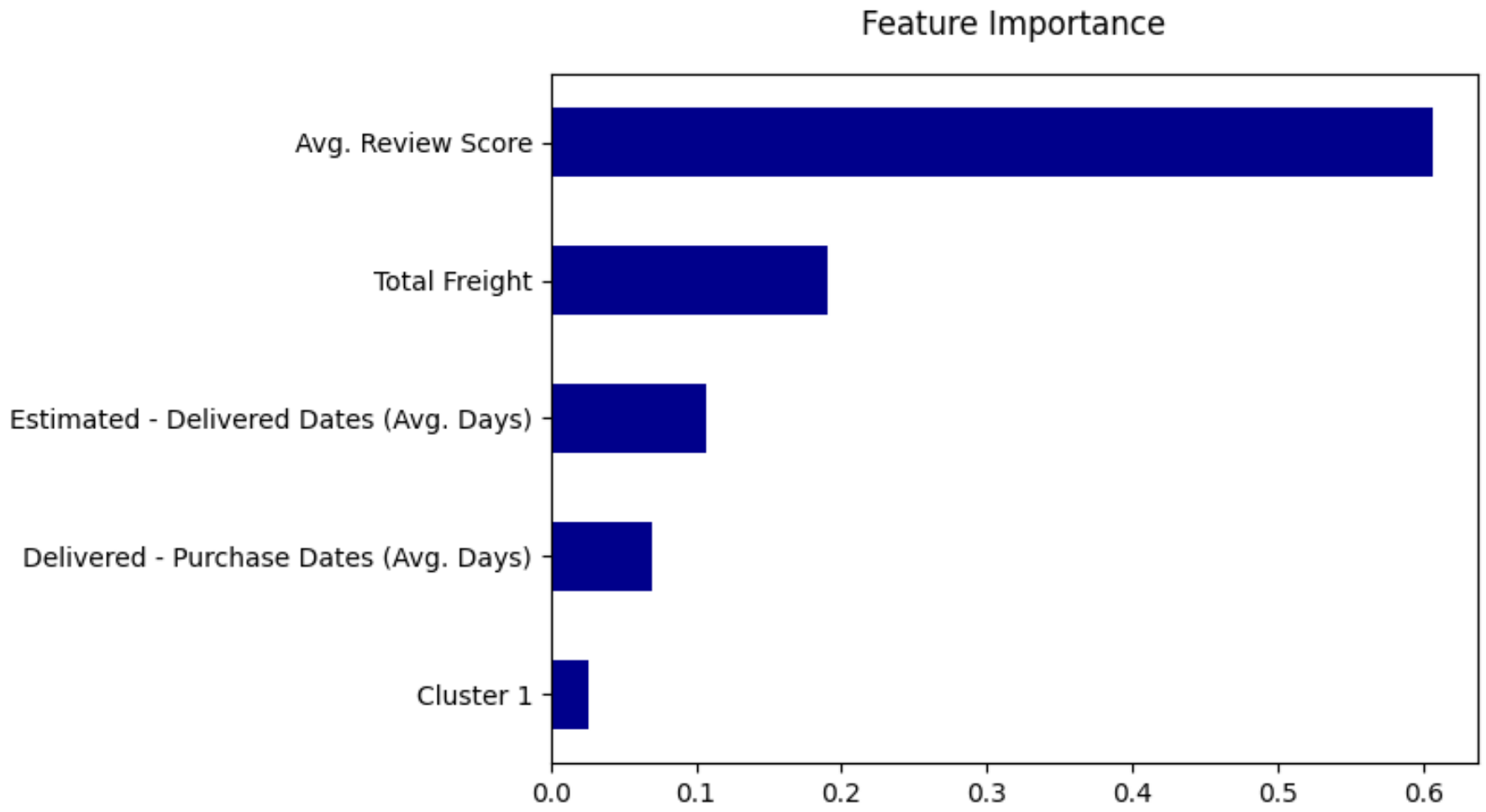
Retention Analysis

Relevant Factors

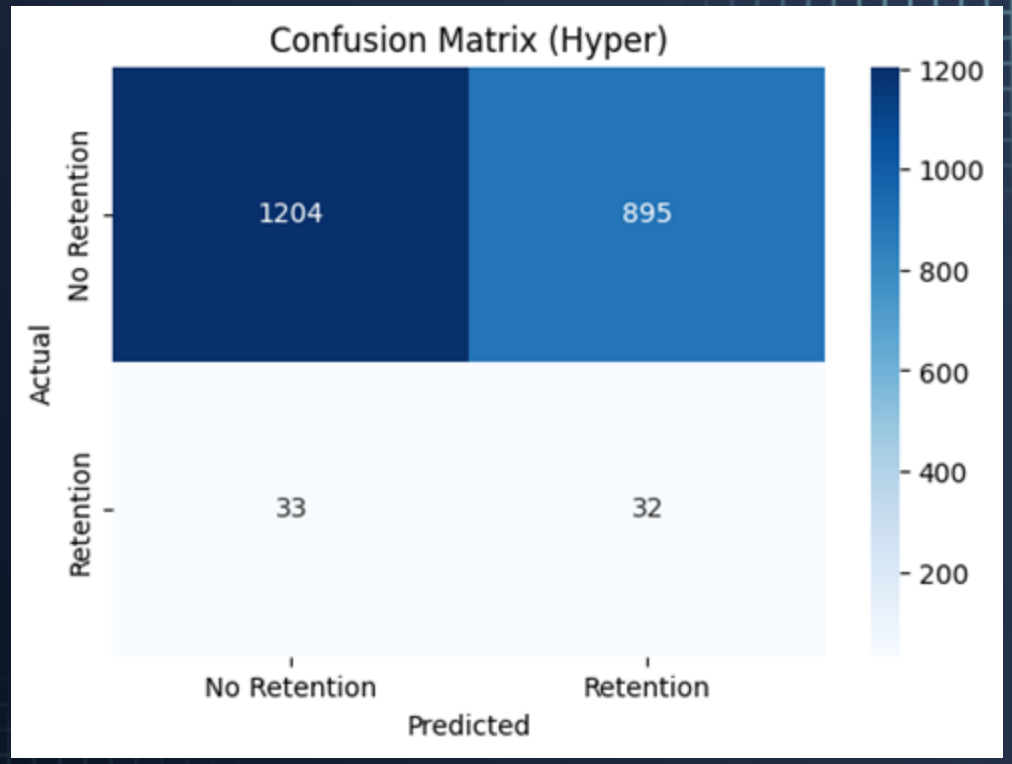
Product/ Service quality

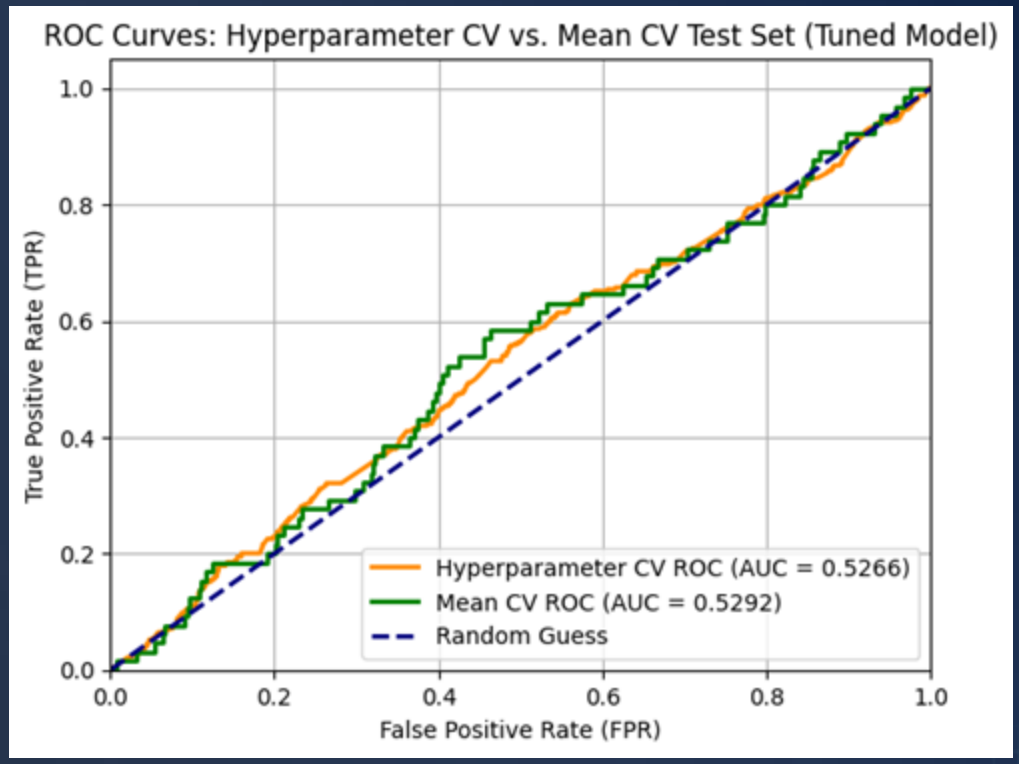
Logistic issues

The model's performance overview is poor due to imbalanced data



LR - Retention Analysis





Test Set	Recall	Precision
0	0.57	0.97
1	0.49	0.03

Train Recall	Test Recall
0.57	0.49

Conclusions

Conclusion: Key Takeaways

Imbalanced Data Impact

The dataset's imbalance caused the model to favor the majority class, affecting its performance.

Missing Data Challenges

Absence of data limited the model's ability to learn comprehensive patterns, contributing to performance constraints.

Best Performance Model

Random Forest performed better than Logistic Regression for this case scenario.

+ Retention Data

To predict the next order more accurately, more retention data was needed.

Conclusion: The diamond customers

True Positive = 41 customers

The model detected correctly 41 customers that purchased again.

These customers are very important because they represent direct revenue to the business and must be nurtured by the marketing department.

Conclusion: Possible hidden opportunities

False Positive = 1189 customers

The model detected that 1189 customers purchased again, but actually they didn’t.

Even though this detection is incorrect, it's important to conduct a cost analysis in order to potentially run campaigns to attract these customers.

Conclusion: Black Friday Insights

Orders at BF = 1176

Number of customers that purchased before the time split (31/05/2017) and purchased again at BF

=

2 Customers

Based on the numbers above, it is possible to see that Black Friday is an excellent tool for a massive injection of revenue at one time.

However, it did not prove to be a good tool for increasing customer retention.

Business Recommendations

The client needs to increase the customer retention rate in order to have a Classification Model with better performance metrics.

Some recommendations based on client data

Customer Segments

- Use the information from the clusters to apply marketing campaigns adapted to each type of customer.

Logistic improvements

- Work with sellers to improve logistic KPIs (On-time delivery rate)
- Create “free freight” campaigns

Loyalty programs

- Create loyalty programs to prevent customers from bypassing the marketplace and buying directly from the Seller

Bibliographic References

[1] https://salesduo.com/blog/customer-retention-on-amazon/

[2] https://www.earnestanalytics.com/insights/temus-retention-grows-over-time-leads-walmart-trails-amazon

[3] https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2021/03/Gu-Grace_Technology-and-Disintermediation-in-Online-Marketplaces.pdf

[4] https://www.mckinsey.com/industries/logistics/our-insights/what-do-us-consumers-want-from-e-commerce-deliveries

[5] https://www.mercadolivre.com.br/l/promocoes

Thank you!

Do you have any questions?