Deliverable 2: Data Wireframe

This project aims to use artificial intelligence (AI) and machine learning (ML) techniques to analyze customer behavior and product performance. The analysis will focus on creating customer segmentation based on purchase behavior, predicting customer churn, predicting customer lifetime value, and sentiment analysis. These insights will help optimize real-time pricing and tailor marketing strategies for high-valued customers.

On the product side, AI and ML will be used to analyze product performance and optimize the product portfolio. The analysis will look at how different products perform and suggest adjustments to the portfolio based on the findings.

This project aims to provide valuable insights into customer behavior and product performance, ultimately helping the business to make data-driven decisions and optimize their strategies.

Here are the project data wireframe with the explanations of each column:

order_id	product_id	seller_id	price	freight
00010242fe8c5a6d1ba2dd792cb16214	4244733e06e7ecb4970a6e2683c13e61	48436dade18ac8b2bce089ec2a041202	58.9	13.29
130898c0987d1801452a8ed92a670612	4244733e06e7ecb4970a6e2683c13e61	48436dade18ac8b2bce089ec2a041202	55.9	17.96
532ed5e14e24ae1f0d735b91524b98b9	4244733e06e7ecb4970a6e2683c13e61	48436dade18ac8b2bce089ec2a041202	64.9	18.33
6f8c31653edb8c83e1a739408b5ff750	4244733e06e7ecb4970a6e2683c13e61	48436dade18ac8b2bce089ec2a041202	58.9	16.17

order_id	payment_type	payment_installments	Amount	customer_id
00010242fe8c5a6d1ba2dd792cb16214	credit_card	2	72.19	3ce436f183e68e07877b285a838db11a
130898c0987d1801452a8ed92a670612	boleto	1	73.86	e6eecc5a77de221464d1c4eaff0a9b64
532ed5e14e24ae1f0d735b91524b98b9	credit_card	2	83.23	4ef55bf80f711b372afebcb7c715344a
6f8c31653edb8c83e1a739408b5ff750	credit_card	3	75.07	30407a72ad8b3f4df4d15369126b20c9

order_id	Date	review_id	review_score	product_category_name	order_status
00010242fe8c5a6d1ba2dd792cb16214	2017-09-13 8:59	97ca439bc427b48bc1cd7177abe71365	5	cool_stuff	delivered
130898c0987d1801452a8ed92a670612	2017-06-28 11:52	b11cba360bbe71410c291b764753d37f	5	cool_stuff	delivered
532ed5e14e24ae1f0d735b91524b98b9	2018-05-18 10:25	af01c4017c5ab46df6cc810e069e654a	4	cool_stuff	delivered
6f8c31653edb8c83e1a739408b5ff750	2017-08-01 18:38	8304ff37d8b16b57086fa283fe0c44f8	5	cool_stuff	delivered

product_id	product_category	product_weight_g	product_length_cm	product_height_cm	product_width_cm
1e9e8ef04dbcff4541ed26657ea517e5	perfumaria	225	16	10	14
3aa071139cb16b67ca9e5dea641aaa2f	artes	1000	30	18	20
96bd76ec8810374ed1b65e291975717f	esporte_lazer	154	18	9	15
cef67bcfe19066a932b7673e239eb23d	bebes	371	26	4	26
9dc1a7de274444849c219cff195d0b71	utilidades_domesticas	625	20	17	13
41d3672d4792049fa1779bb35283ed13	instrumentos_musicais	200	38	5	11

Zip_code	geolocation_lat	geolocation_lng
1037	-23.54562128	-46.63929205
1046	-23.54608113	-46.6448203
1046	-23.54612897	-46.64295148
1041	-23.54439216	-46.63949931
1035	-23.54157796	-46.64160722

Columns Descriptions:

1. **Order_id:** Unique identifier for each order.
2. **product_id**: Unique identifier for each product.
3. **Seller_id:** Unique identifier for each seller.
4. **Price:** Price of the product.
5. **Freight_value:** The cost of freight for shipping the product.
6. **Payment_type:** Type of payment used by the customer (e.g., credit card, debit card, etc.).
7. **Payment_installments:** The number of installments for the payment.
8. **Payment_value:** The total value of the payment.
9. **Customer_id:** Unique identifier for each customer.
10. **Order_status:** The status of the order (e.g., delivered, cancelled, etc.).
11. **Order_purchase_timestamp:** The timestamp of when the purchase was made.
12. **Review_id:** Unique identifier for each review.
13. **Review_score:** Score given by the customer in the review.
14. **Postal_code:** Postal code of the customer.
15. **City:** City of the customer.
16. **State:** State of the customer.
17. **Product_category_name:** Name of the product category.
18. **Product_weight_g:** Weight of the product in grams.
19. **Product_length_cm:** Length of the product in centimeters.
20. **Product_height_cm:** Height of the product in centimeters.
21. **Product_width_cm**: Width of the product in centimeters.
22. **Seller_zip_code_prefix: Zip** code prefix of the seller.
23. **Seller_city:** City of the seller.
24. **Seller_state**: State of the seller.
25. **geolocation_zip_code_prefix:** postal codes
26. **geolocation_lat:** latitude based on the zip code
27. **geolocation_lng:** longitude based on zip code.

What clean-up steps do you anticipate?

Handling Missing Data: From the preliminary inspection, it appears that some columns like 'product_category_name', 'product_weight_g', 'product_length_cm', 'product_height_cm', and 'product_width_cm' have missing values. These will need to be handled either through imputation, deletion or other suitable methods depending on the nature and proportion of the missing data.
Dealing with Outliers: It will be necessary to check for and handle outliers in numerical fields like 'price', 'freight_value', 'payment_value', etc., as these can influence the results of your analysis and model performance.
Data Formatting: The 'order_purchase_timestamp' column, currently an object type, will need to be converted to a datetime format for more accurate time series analysis.
Checking for Duplicates: The data will need to be scrutinized for any duplicate entries which could skew the analysis.

How do you want your perfect data set to look?

The perfect data set for this project would be clean, well-structured, and devoid of any missing values, outliers, or duplicates. It would be formatted correctly with appropriate data types for each column. The data should be rich and diverse enough to allow for robust customer segmentation, churn prediction, lifetime value prediction, sentiment analysis, and product performance analysis.

END