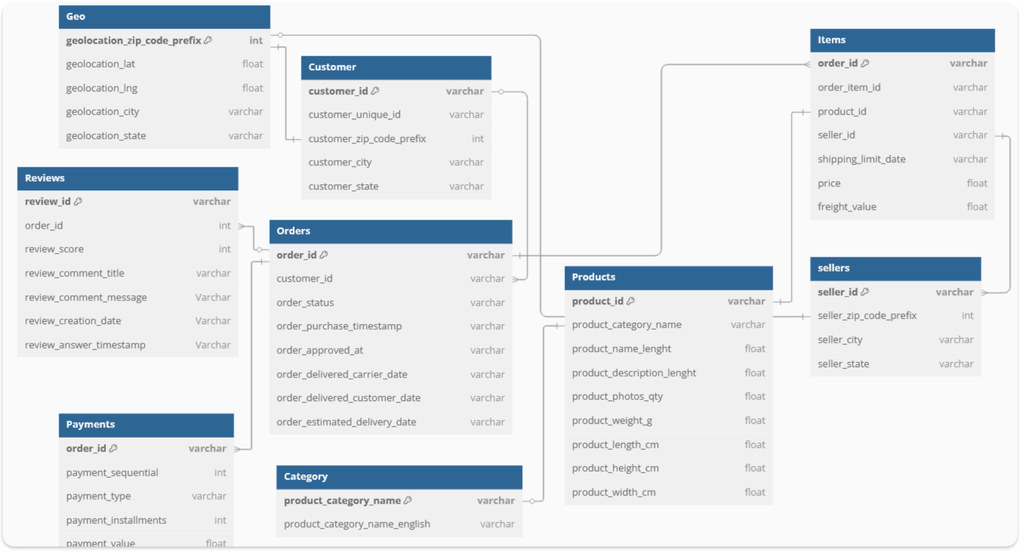

Business Questions:

- Orders:

A. For order status, 'delivered', 'shipped', 'others', what is the average duration for status change from 'shipped' to 'delivered'? This metric measures the effectiveness of logistic partner's services.
B. Can breakdown 'others'? Task this to operation team to answer.
C. From 'order_purchase_timestamp' to 'order_approved_at', what is the average duration? Max and Min too. And what is the order approval process? Why some orders took a few mins to be approved, while some were took days to be approved? Again, operation team to provide details.
D. From 'order_approved_at' to 'order_delivered_carrier_date', what is the average duration? Max and Min too. Tracing order_id -> seller_id, aggregate the top sellers that are providing superior services to minimize this duration. Experience sharing by top sellers to other sellers on this, to further improvement on this metric.
E. From 'order_delivered_carrier_date' to 'order_delivered_customer_date', what is the average duration? Max and Min too. Aggregate to find out the most effective carrier. And suggest sellers to consider their engagement with better carrier supported by data.
F. From 'order_delivered_customer_date' to 'order_estimated_delivered_date', what is the average duration? Max and Min too. Compare this with item E average, max and min result, what is the delta? This metric shows the expectation given to customer and their real experience in terms of delivery speed. Is it over-promised, under-promised, or the prediction is just fine? What is the basis of the delivery duration prediction? From carrier provider? And why their estimation has such delta? Details could be provided for improvement.

- Payments:

A. 'payment_sequential' field reveals that majority of the customers pay in a single payment, can we introduce more voucher or debate scheme to reward returning customers? Measure the effectiveness once implemented on this column vs 'payment_value'.
B. 'payment_type' field reveals that majority of the customers pay by credit card payment mode, can we narrow down to which bank's credit cards are most popular? And think of a business profit-sharing scheme with the selected bank to offer customers with better rebate and faster shipping method.
C. What is the correlation between 'payment_installment' and 'payment_value'? What is the threshold amount that makes customers consider installment rather than one-time payment? Again can work out the profit-sharing scheme with the selected bank to offer customers with better choice in both one-time payment and installment incentives.

- Customers:

A. What is the geographical distribution of customers? The lower density areas reveal more business opportunities through advertisement on mass media.
B. Link 'customer_id' to 'payment_value', find out the bigger spending group of customers, analyze their buying patterns, and introduce product recommendation to them via apps or social media channels.
C. Plot out the geographical distribution by customer spending tiers. Marketing strategies shall be aligned with this statistics closely.

- Reviews:

A. From 'review_score', groupby 'customer_unique_id', to figure out the happy customer groups and unhappy groups. It could also reveals the challenging customers so as to set a ceiling for their review entries.
B. The average 'review_score' is the telltale sign of the service being rendered, as the corporate goal to be set for improvement.
C. 'review_comment_title' needs language translation, we could leverage on web scraping skill to automate it? Sort 'review_comment_title', the result shall show more insights on major areas of services for improvement. This can be in conjunction with item A, to bypass challenging customers' reviews for more informative data.
D. From 'review_creation_date' to 'review_answered_timestamp', what is the average duration? This metric shows the resolution effectiveness. And it could be in conjunction with 'seller_id' and 'product_id' to find out the complaints are for products or services?

- Order_items:

A. From 'product_id' column, easily to find out what are the most popular products that being sold. Does sellers hold enough stock for popular products? Can the e-commerce platform stock-up popular products to push for more sales? Can it combine with less popular products to create bundle-deal? Can increase price for better profit margin?
B. From 'shipping_limit_date' and 'seller_id', can easily find out the efficient seller groups, provide incentives to level up the overall efficiency.
C. From 'price', what are the most comfortable price range of customers for them to choose online purchase? It can link back to customer tiers and geographical distribution to find out more insights.
D. From 'freight_value', find out the correlation with 'product_weight', 'product_length', 'product_height' and 'product_width'. This yields the best approach for packaging strategy to be shared with sellers, so as to increase profitability.

- Products:

A. From 'product_name', 'product_description' and 'product_photo', aggregate to find out the most effective 'product_name' length, the way to present it online. Same rom 'product_name', 'product_description' and 'product_photo', aggregate to find out the most effective 'product_name' length, the way to present it online. Same applies to 'product_description', how to make it clear and concise, the best practice could share among sellers. 'product_photo' will be the more the better or less is more? Photo quality, informative indication?

- Sellers:

A. What is the geographical distribution of existing sellers? Will the current seller-customer pairs are using the best algorithm to match? Re-pair based on geographical location can be done? This increases customer satisfaction by shorten the delivery time. As well as for 'out-of-stock' scenario.

- Geolocations:

A. This table serves as the basis for all geographical distribution analysis, which shall be used for joins and plots.