Summary

Dashboard vs EDA: need to decide if I want to create this like a regular ongoing e-commerce dashboard where I display all potentially important metrics (now or in the future) or an EDA where I just present insights.

- Potential employers would like the dashboard to show I know what visuals and insights are important in the industry.
- I don't want to waste too much time putting together visuals that I know won't show much since the dataset is not ongoing.
- Can punt and decide later. At least have a main dashboard page with most important metrics.

First Page (Dashboard)

Revenue & Revenue Drivers

- **Absolute Revenue**
    - Chart showing revenue over time up to present with ability to adjust time axis
    - Optional additional lines in the revenue chart:
        - Average Order Value (AOV)
        - Avg units per order
        - Avg price per unit ordered
        - # of unique products purchased (per category - stacked bar)
        - # of unique customers purchasing
- **Revenue Growth YoY**
    - Optional additional contribution to growth stacked bars/area:
        - By product category
        - New vs returning customers
        - Growth in avg units ordered vs price per unit

Revenue Forecast?

- Could use premade model like Meta's and would have adjustable time period but default to next 30 days.

Marketplace Growth

- Chart(s) with: total transactions, total active sellers, total inactive sellers, new sellers, total active customers, total inactive customers, new customers, new products,mn.kjh.

Best Sellers (Products & Sellers)

- Viz with adjustable time axis, preset for last week.

- (Table) Top 25 best selling products with absolute revenue and YoY growth
- (Table) Top 10 best selling product categories with absolute revenue and YoY growth.

Problems / Alerts

- Table for Late Deliveries:
    - By degree of lateness
    - Shows lateness in days, expected date, seller, customer, revenue/value.
    - Will probably have to toggle by how late, default should be latest orders.
- Seller Alerts
    - Table of worst sellers defaulting to last 30 days view
    - Score based on: canceled orders; late deliveries; undelivered orders; negative reviews.

Seasonality / Time View

Overall Revenue Seasonality (ML?)

- Use R packages to identify seasonal patterns, like holiday spikes, etc. Need list of Brazilian holidays.
- ML?

Product Category Seasonality

- Find which product categories have highest sales in which quarter, month, week, etc. See if you can create clusters of categories with significantly seasonal sales and then group them into peak season/month categories.

Purchase Timing

- Transactions by day of week and time of day
- Is this breakdown relatively constant over the 2 years of data or shifting?

Market Basket Analysis

Frequently Bought Together (FBT)

- Products commonly bought in the same order
- Products commonly bought at various time delays (24 hrs, 48hrs, 7D, 30D).
- Generally popular add-on products for any order?
- Do these change or become more frequent by season (quarter or month)
- Does customer location impact FBT pairs/groups?

Logistics/Geospatial

Buyer (Seller) Location

- Map views with dots sized for amount of revenue from that location (state or zip)
    - Same adjusted for area population (revenue per capita)
    - Same adjusted for retail sales or income (revenue per income)
- Table for buyer location with categorizations: urban/rural, coastal/inland, by state, by region with adjustments per capita and per avg income (share of income or something like that).
- Can do the same or part for sellers as well since marketplaces need to market to them as well. Also important for considering any logistical help like warehousing and/or delivery.

Ship Cost/Time Driver ML

- Goal: what factors have the strongest influence on shipping cost or shipping time?
- Variables: buyer-seller distance, buyer-seller urban to urban, urban to rural, etc., package dim weights, product price, time of year, seller volume (# of units), seller experience (time since first sale),
- May want to separately see if different sellers shipping the same product to the same buyer zip have different shipping costs and if this is related to their volume or experience.

Impact of Freight Costs on Sales

- Correlation of freight/shipping share of item price against number of orders or total revenue.
    - Can control for confounding variables or use multivariate regression. Variables like price, delivery time, seasonality, etc.
    - Can look for shipping cost sensitivity by customer segment (high vs low spenders, frequent vs infrequent buyers).

Logistics Failure

- Cancellations and delays by seller and buyer location. Are cancellations and delays more common when shipping from certain regions or to certain regions?

Customer Segmentation

Customer Clustering ML

- Based on factors like: customer total spend, # of orders, AOV, geolocation (urban/rural, coastal/inland, region, state), payment type, avg number of payment installments, order

time (DoW, ToD), new or returning customer, number of reviews left, avg reviews left per item ordered, product categories purchased.
- K-means seems to be used most but various other algos available.

Customer Lifetime Value (CLV)

- There are multiple examples of calculating CLV in published Kaggle notebooks with this dataset. However, repeat customers are a very small percentage of total in the dataset so CLV may not be a very reliable calc in the way it is usually done.
- CLV can be calculated using ML or with non ML methods
    - Traditional CLV = avg purchase value x average purchase frequency X avg customer lifespan.
        - We won't have lifespan with only 2 yrs of data.
    - Cohort CLV - calculate CLV by customer segments.
- Something like 93% of customers are one-time buyers which makes CLV calculations less reliable unless we truly believe these customers will never buy again.
    - Probably more useful to focus on the 7% that are repeat buyers and see what characteristics are common among them.

Characteristics of High-Value Customers

- See what factors tend to correlate to a customer being a repeat purchaser
- See what factors tend to correlate with customers having high AOVs.
- RFM (recency, frequency, monetary) Matrix
    - Create a score for customers based on these factors with the higher scoring being the most promising target marketing audience.

Pricing Analysis

Price Distribution

- Histogram of pricing across dataset
- ECDF plot of prices

Price Changes

- Avg price for all products over time
- Avg freight price for all products over time
- Avg price for all products by category over time
- Products with significant price fluctuations - alerts?

Price Elasticity of Demand

- Need to ensure a decent sample size of products bought enough at different prices to determine this.
- Can classify products into groups by elasticity or create an interactive table where you can input product id and get back elasticity.

Degree of Price Competition

- I looked into how often we see seller competition by looking at the number of sellers selling the same product in any given month and there is almost no competition apparent in the dataset. I.E. there was almost always one seller per product per month.

Sellers

- Seller growth: unique sellers with sales per month
    - Avg # of products sold per seller or histogram
    - Revenue per seller or histogram
- Top sellers T30D (user can move time period)
- Any sellers with large change in revenue T30D (user can move time period)
- Seller's products avg review scores
- Seller's delivery record - get products to freight carrier on time?

Market Concentration (Sellers)

- Degree to which sales are dominated by small group of sellers (HHI?)

Products

Product Category Stats

- Revenue per category over time
- Unique products per category over time
- Avg product price per category
    - Any shifts in price distribution or avg?
- Purchase frequency per category over time

Market Concentration (Products)

- Degree to which sales dominated by few products or visa versa (HHI)

Top Selling Products

- Table of top sellers with products_id, category, sales T30D, sales growth MoM.

- Any top sellers showing major changes in revenue, AOV, etc.

Listing Feature Analysis

- Do features of listings like # of pics, length of title, length of description correlate to product revenue?

Product Size & Weight

- Per category avg dimweight or major changes in dimweight
- Can I divide dim-weights based on Brazil's shipping size tiers?
- What dim-weights sell best?
- *Need to answer some basic shipping questions first:*
    - *Does an item always have the same ship cost no matter what other items are ordered with it or time of year?*
    - *If not, lots of calcs will change to see if customers prefer to bundle for shipping savings, etc.*

Payments

- Payment method breakdown - change over time.
- Payment installments - change over time
- Methods and installments by price tier
- Methods and installments by season

Reviews

- What causes bad reviews? (evidence that it is delivery failures or delays)
- Avg review score by category
- Alerts for products with multiple negative reviews or avg score falling significantly.