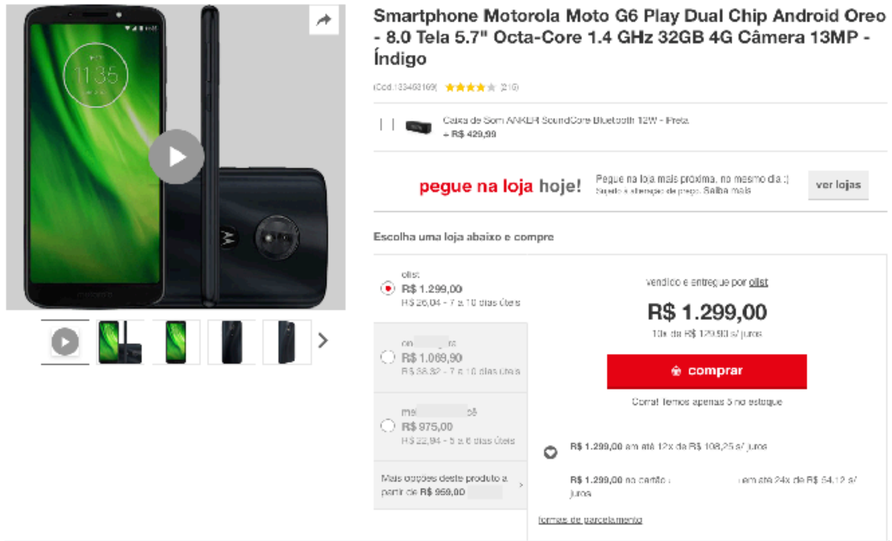
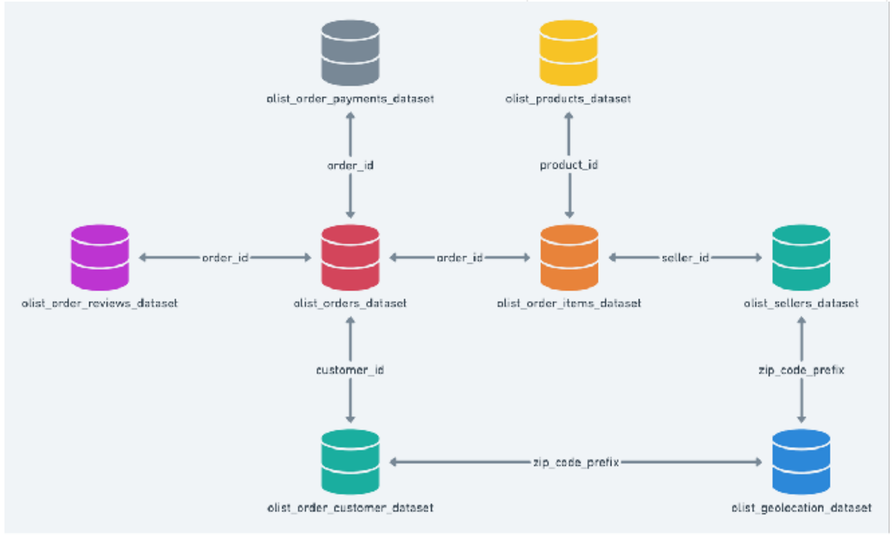
# CAPSTONE PRESENTATION

*E-commerce Analytics*

# BACKGROUND AND OVERVIEW OF THE DATA

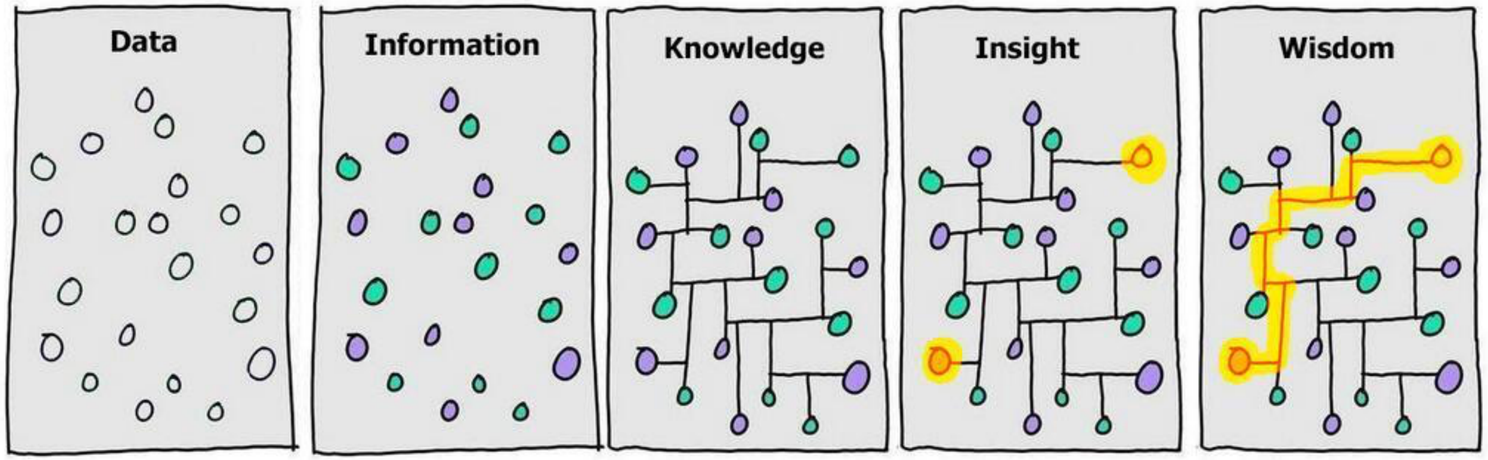
- **What is Olist?**
  - The largest online department store in Brazil. It connects small businesses from all over Brazil to sell products directly to customers.
- The relational database (from Kaggle) contains **110k orders from 2016 to 2018.**
  - Split into 8 datasets that allow viewing an order from multiple dimensions: order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers.
  - Real commercial dataset that has been anonymised and released for public use



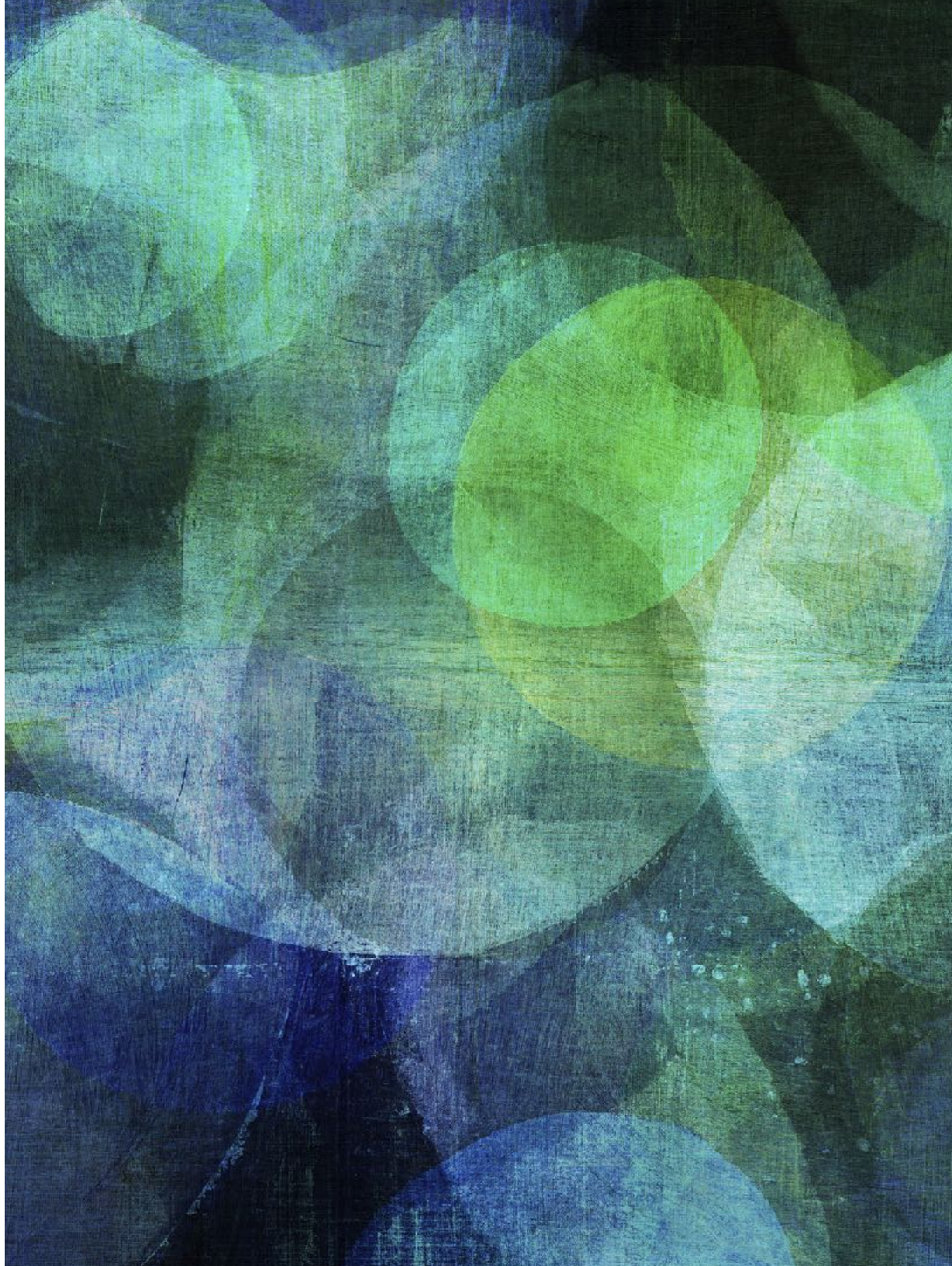
# 3 BUSINESS OBJECTIVES GUIDING THE DATA SCIENCE PROCESS

1. **Platform insights** - What are the geographical and category-level revenue patterns across Olist?
2. **Customer understanding** - How can customers be segmented based on their purchase habits? How much will a customer bring in future revenue?
3. **Customer satisfaction** - What are the drivers of positive customer reviews?



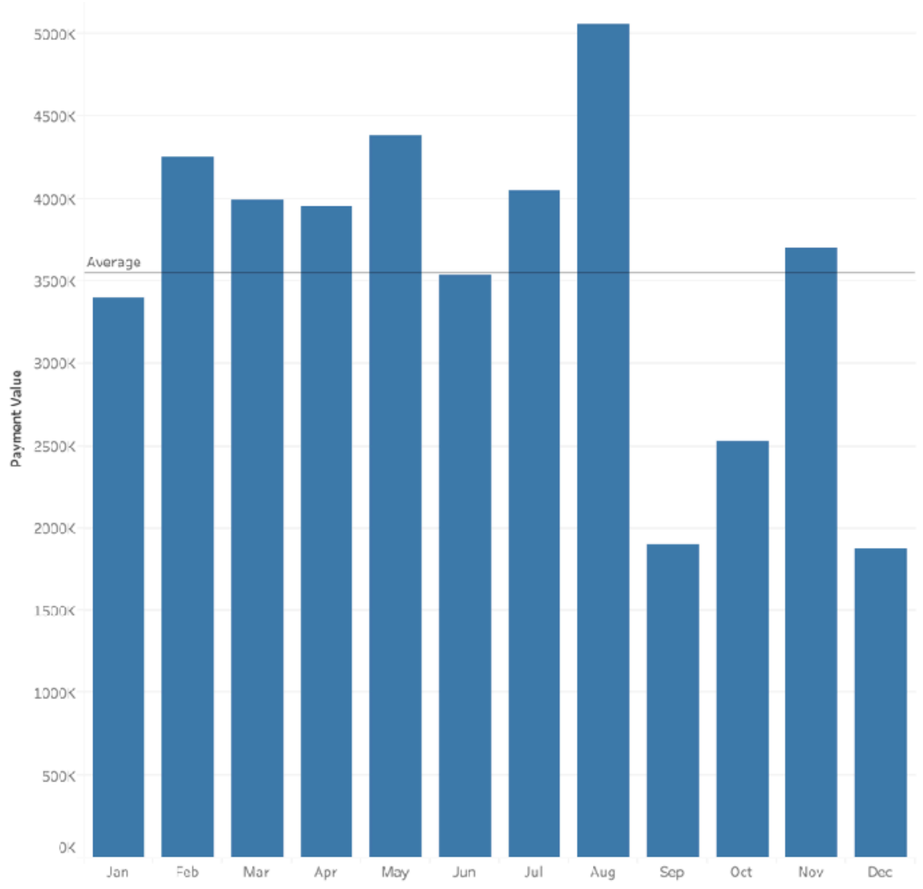
# PLATFORM INSIGHTS

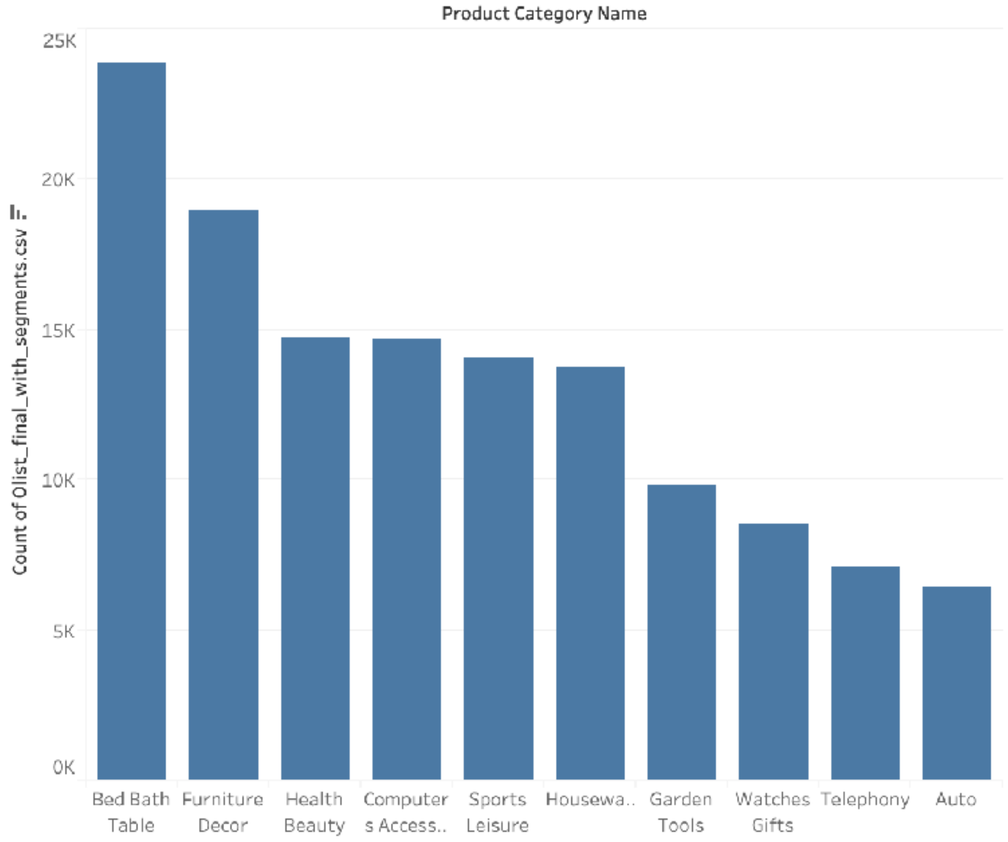
*What are the patterns in the data?*



# MACRO-LEVEL PATTERNS IN REVENUE AND CATEGORY

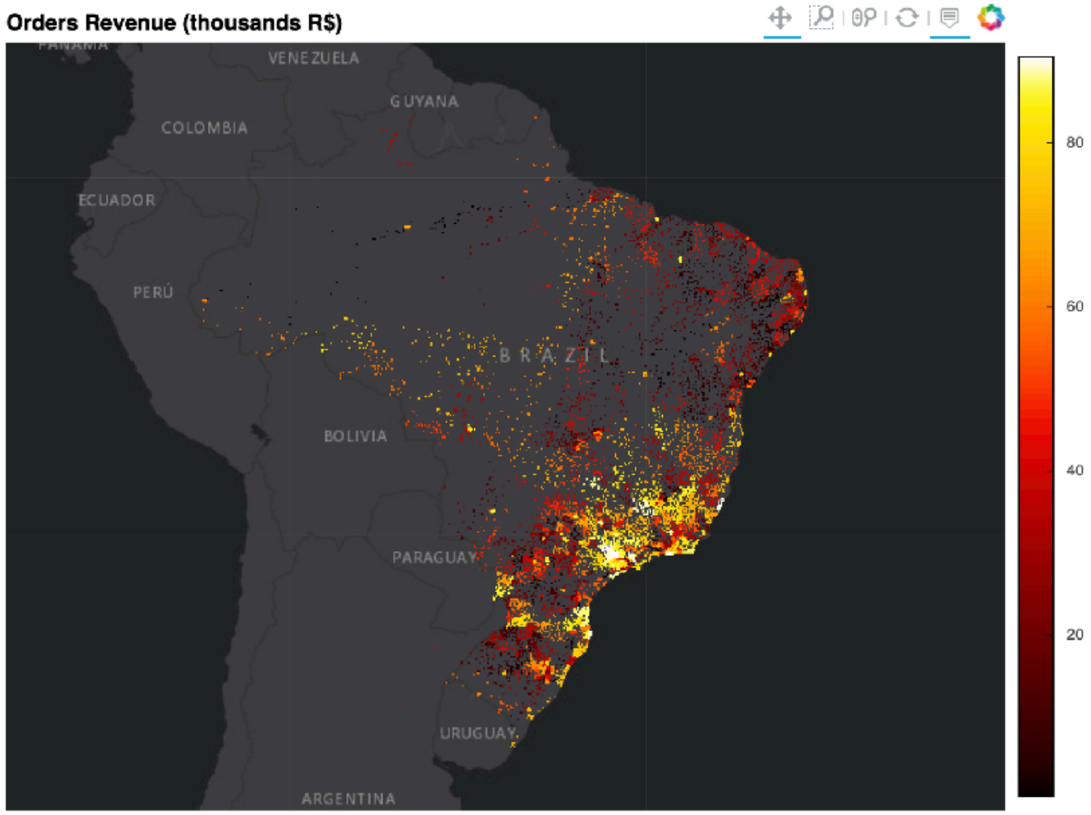
- Order volume peaks in August, sees a sharp decline and picks up just before Black Friday.
- Bed/bath/table category is the biggest by volume

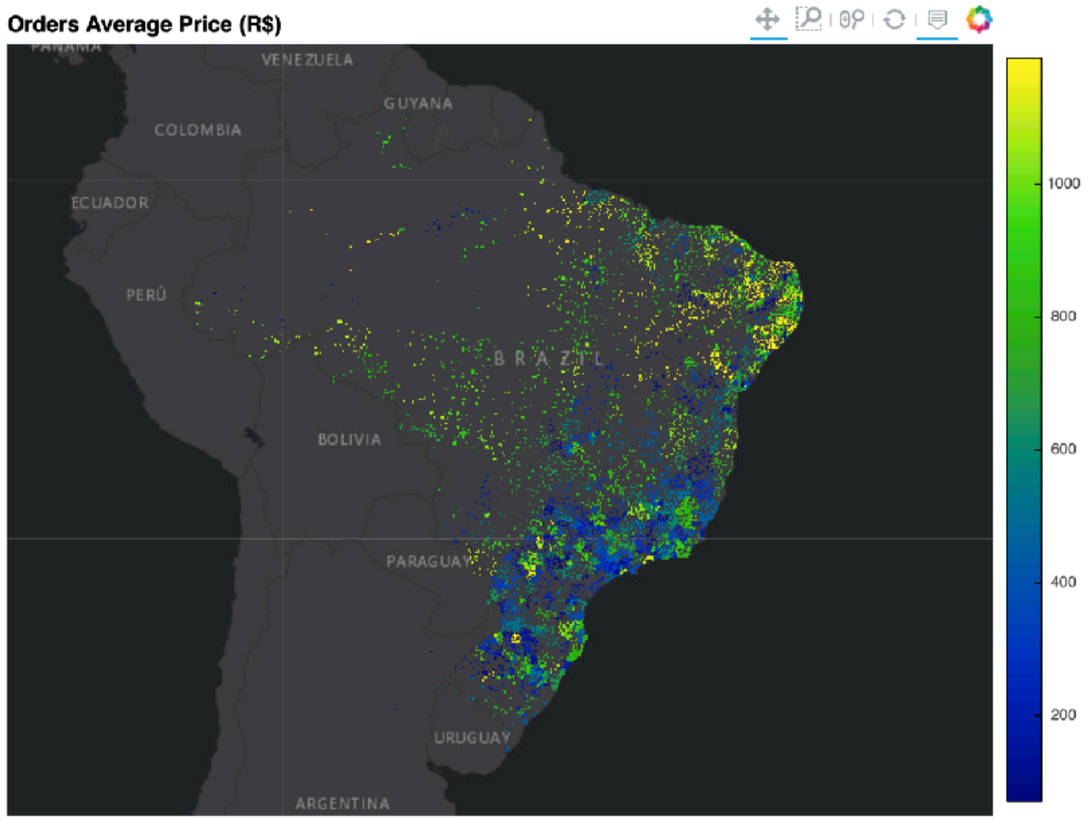


# GEOSPATIAL ANALYSIS SHOWS THAT REVENUE MOSTLY COMES FROM SAO PAOLO

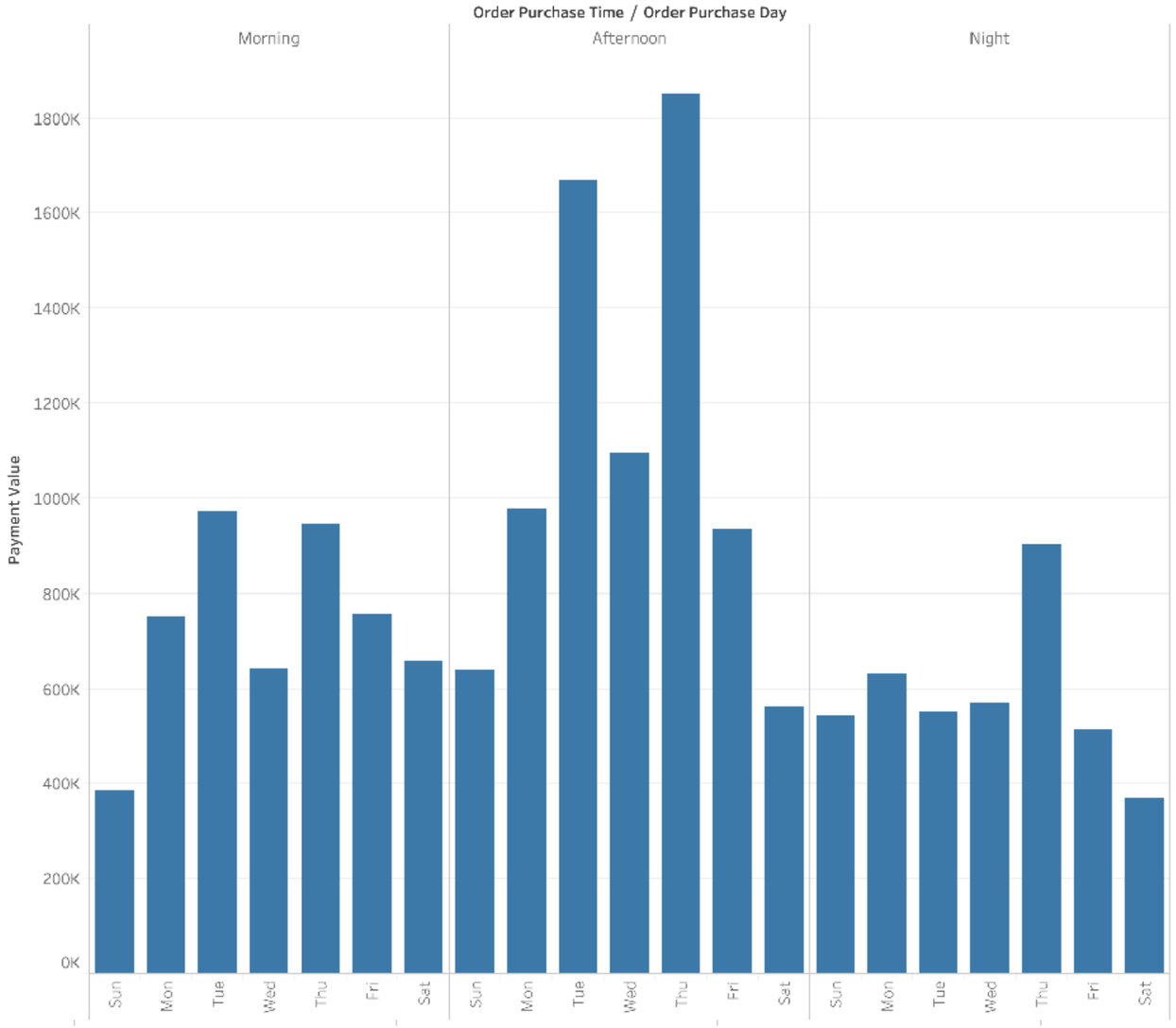
- While Sao Paolo has the most amount of orders and lowest average payment value, states further away from distribution centres pay more freight, driving up average price per order



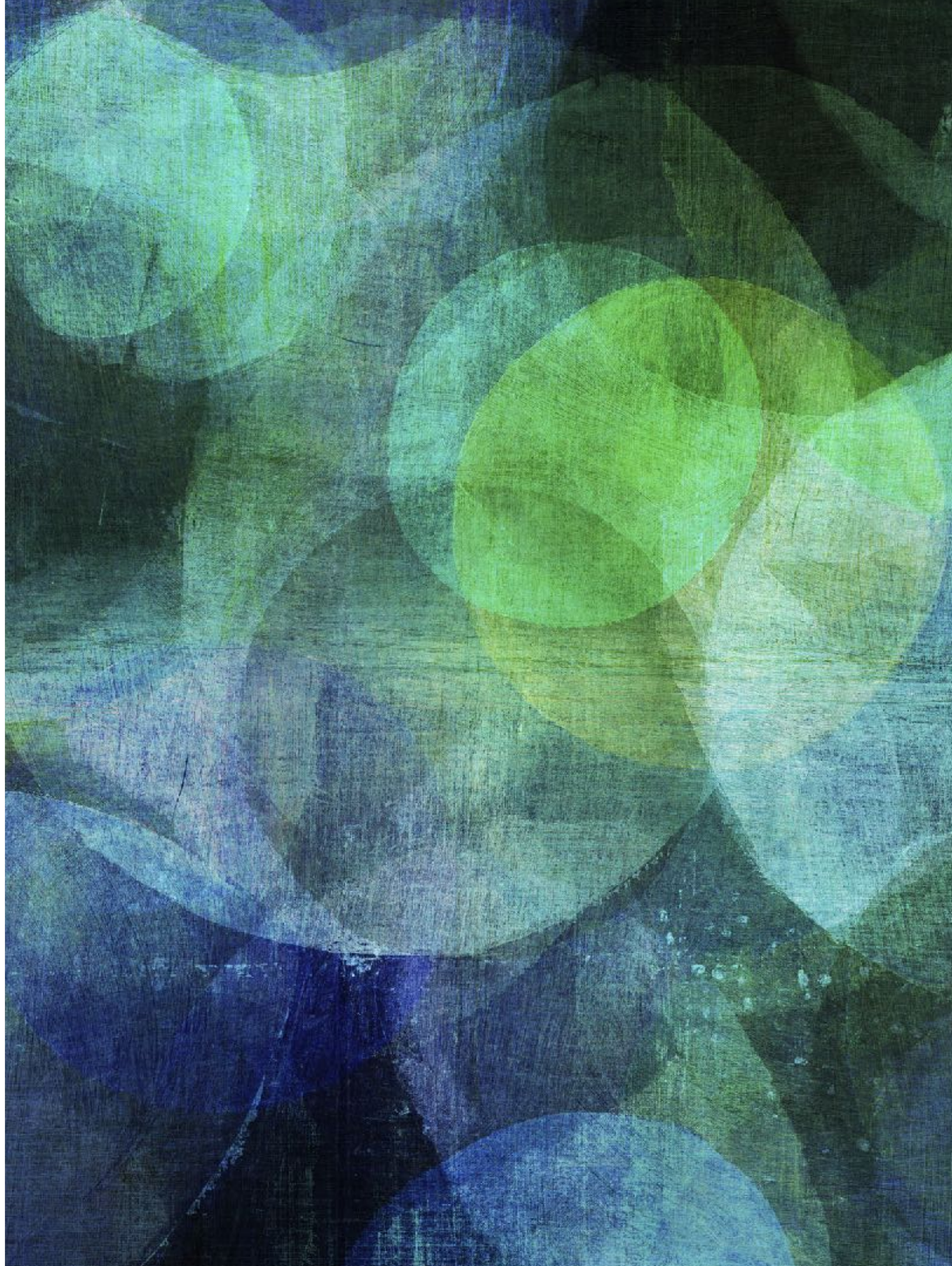
# CLEAR BUYING PATTERNS EMERGE WHEN LOOKING AT DAY OF WEEK AND TIME OF DAY

- Tuesdays and Thursdays are the most popular days for online shopping, and afternoon (12pm-4pm) is peak shopping time in Brazil.
- At a category-level: sales of automotive category is very popular on Wednesday mornings and computer accessories on Thursday afternoon.



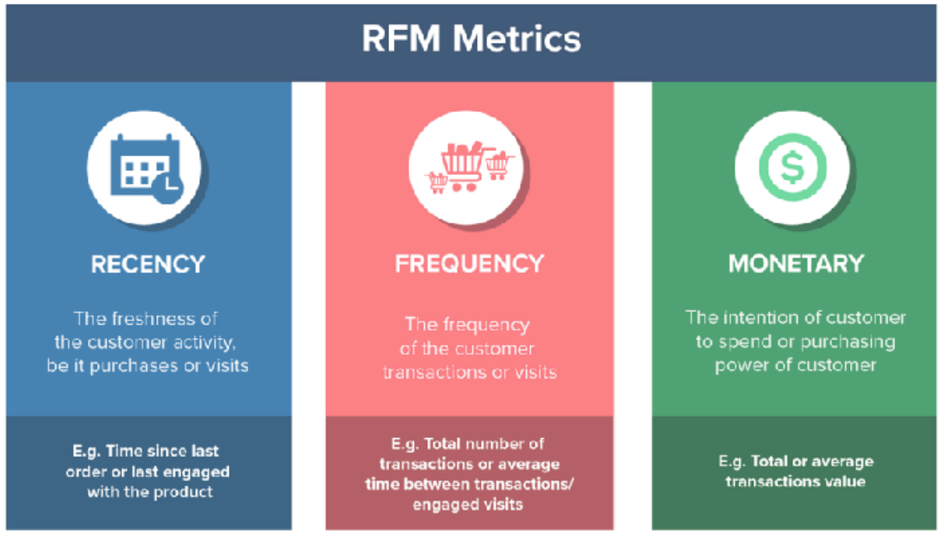
# CUSTOMER UNDERSTANDING

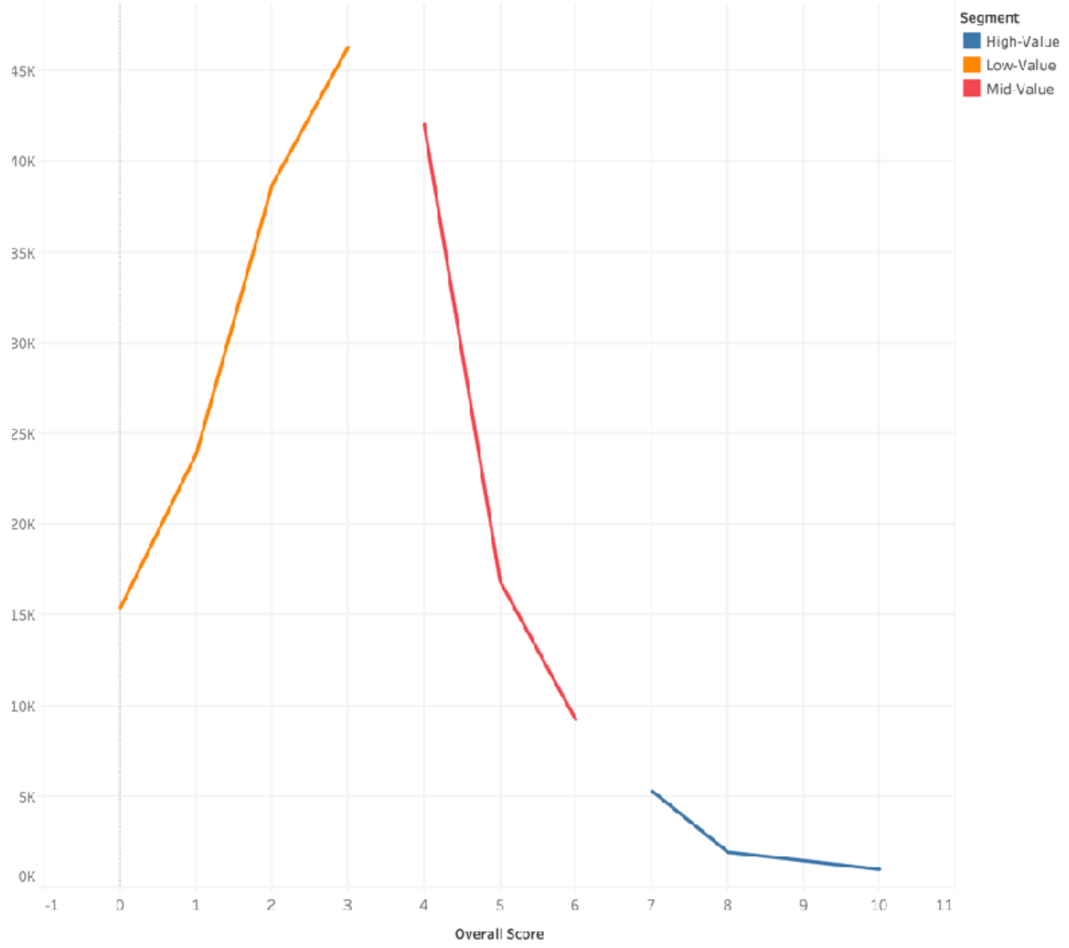
*How does shopper behaviour differ?*



# RECENCY-FREQUENCY-MONETARY ANALYSIS FOR CUSTOMER SEGMENTATION

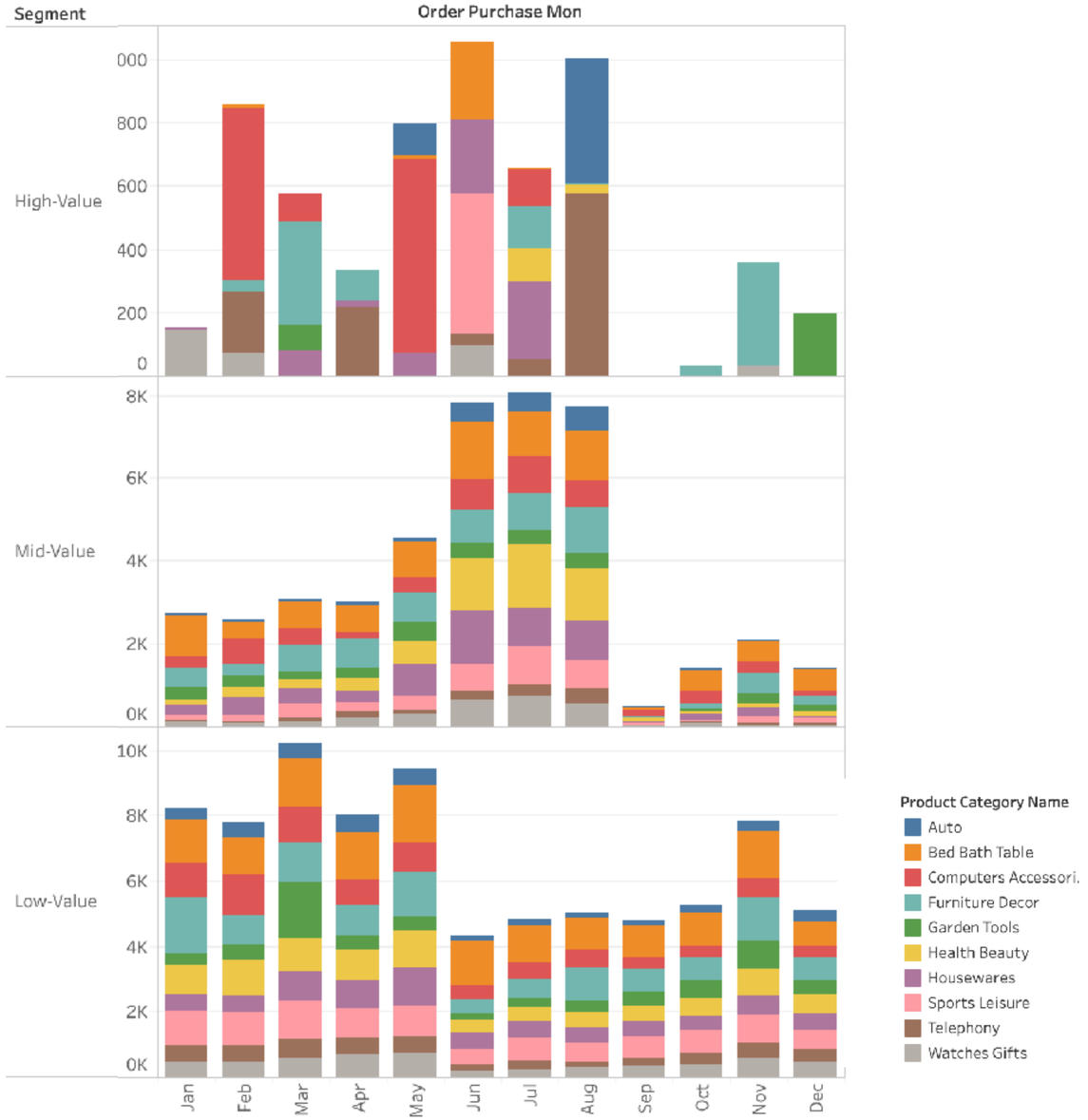
- Instead of analysing the entire customer base, it is better to segment them into homogeneous groups, understand the traits of each group, and engage them with relevant campaigns rather than segmenting on just customer age or geography.



# SEGMENTATION REVEALS CRITICAL BEHAVIOURAL DIFFERENCES

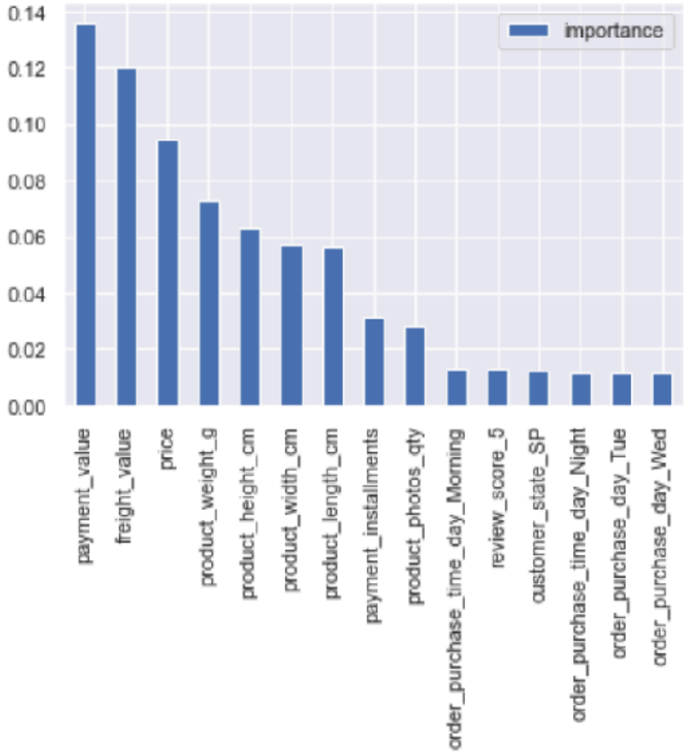
| | Category shopped most often: | Majority located in: | Purchase most frequently in: |
|---|---|---|---|
| **High value** | Automotive | Maranhão | Q2 |
| **Mid Value** | Health and Beauty | São Paulo | Q2 |
| **Low Value** | Bed Bath Table | Rio de Janeiro | Q1 |



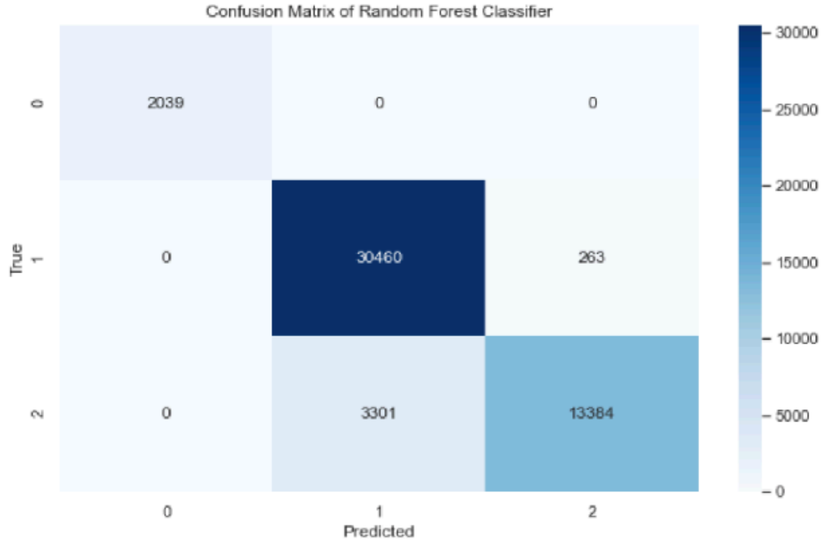
# PAYMENT VALUE IS THE KEY PREDICTOR OF CUSTOMER SEGMENTS

- Monetary factors are the key determinant for segments - intuitive but not insightful.
- A non-monetary based model (for predicting customer value on an individual level) would help the business to -

1. Distinguish active customers from inactive customers.
2. Generate transaction forecasts for individual customers.
3. Predict the purchase volume of the entire customer base.
4. Predict customer churn risk.



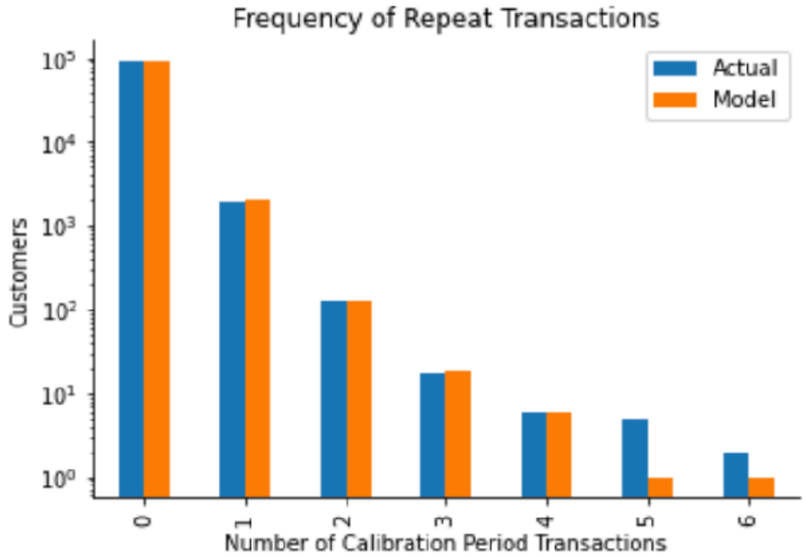

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| High-Value | 1.00 | 1.00 | 1.00 | 2039 |
| Low-Value | 0.90 | 0.99 | 0.94 | 30723 |
| Mid-Value | 0.98 | 0.80 | 0.88 | 16685 |
| accuracy | | | 0.93 | 49447 |
| macro avg | 0.96 | 0.93 | 0.94 | 49447 |
| weighted avg | 0.93 | 0.93 | 0.93 | 49447 |

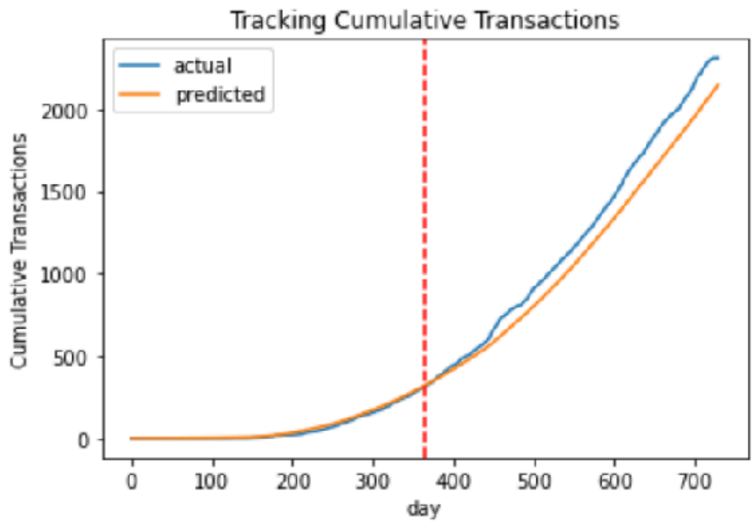



***Model**: Random Forest | **F1 Score**: 93%*

# PREDICTING CUSTOMER LIFE VALUE USING ONLY RECENCY AND FREQUENCY

- Lifetimes links the RFM paradigm with customer lifetime value (CLTV). The stochastic model presented here, featuring BG/NBD framework to capture the flow of transactions over time. BG/NBD portrays the story being about how/when customers become inactive.
- This model requires only two pieces of information about each customer's purchase history: **Recency and Frequency**

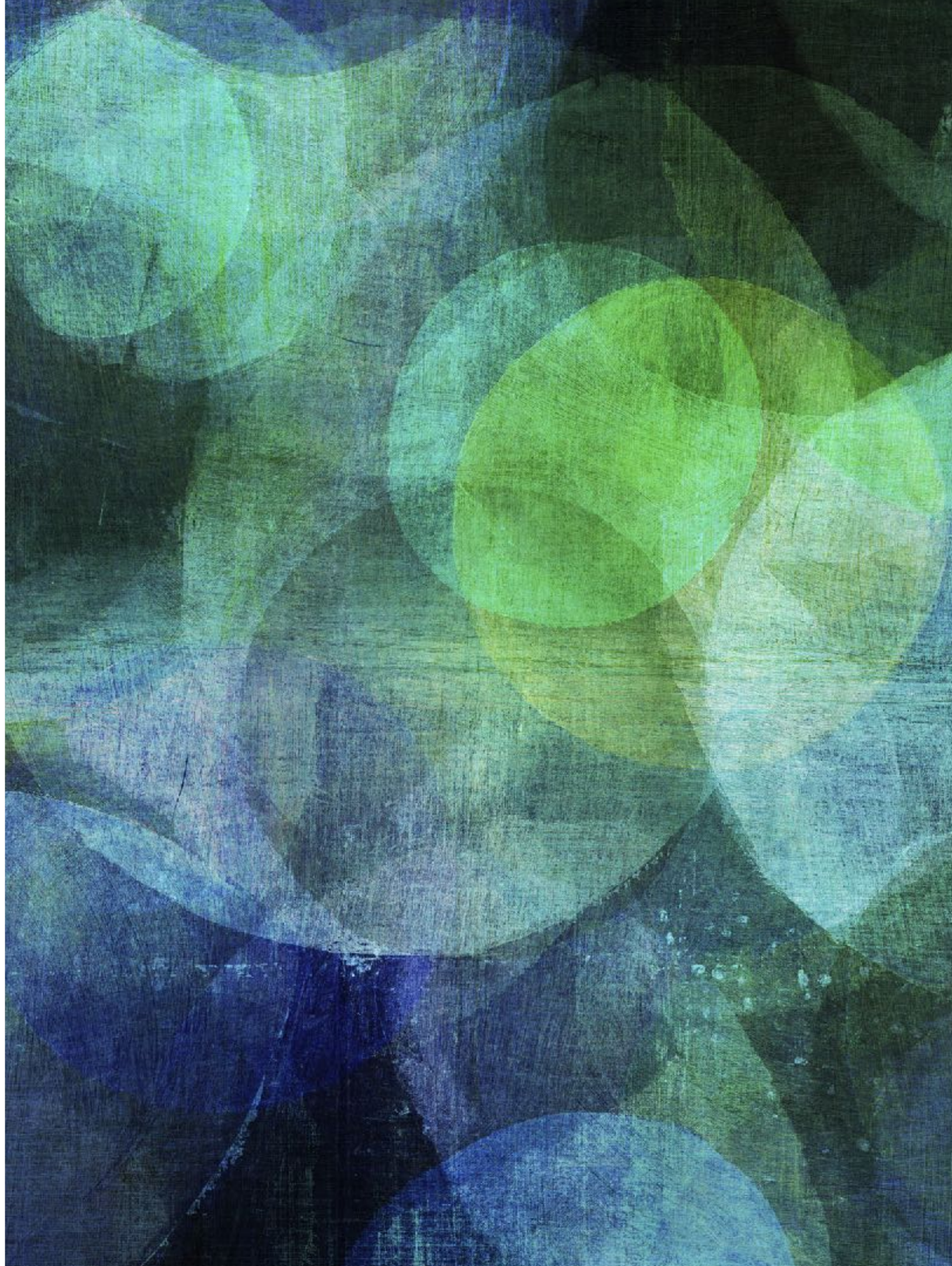



*Model: BG/NBD | Prediction error: 6%*

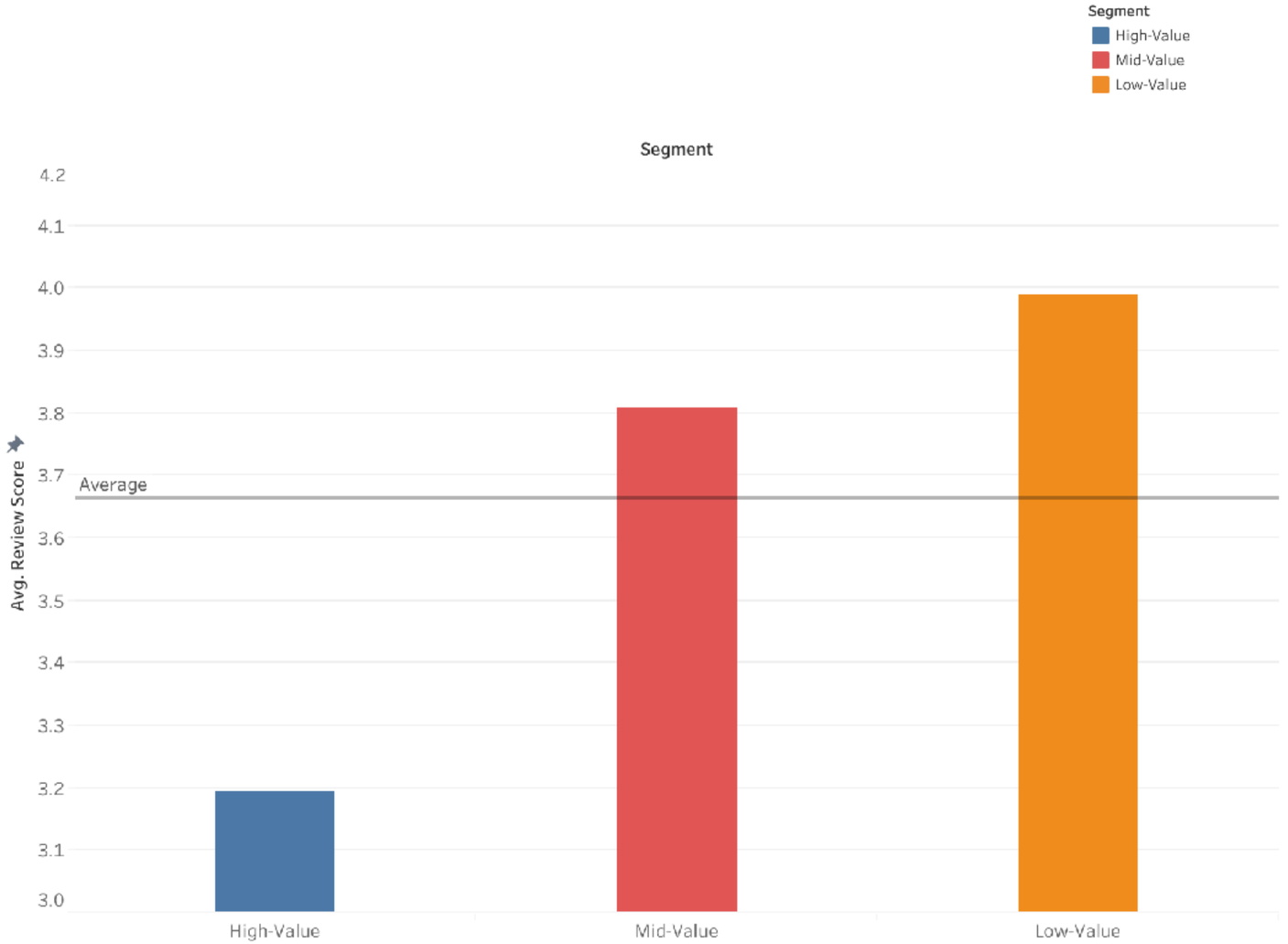
# CUSTOMER SATISFACTION

*What drives positive reviews?*



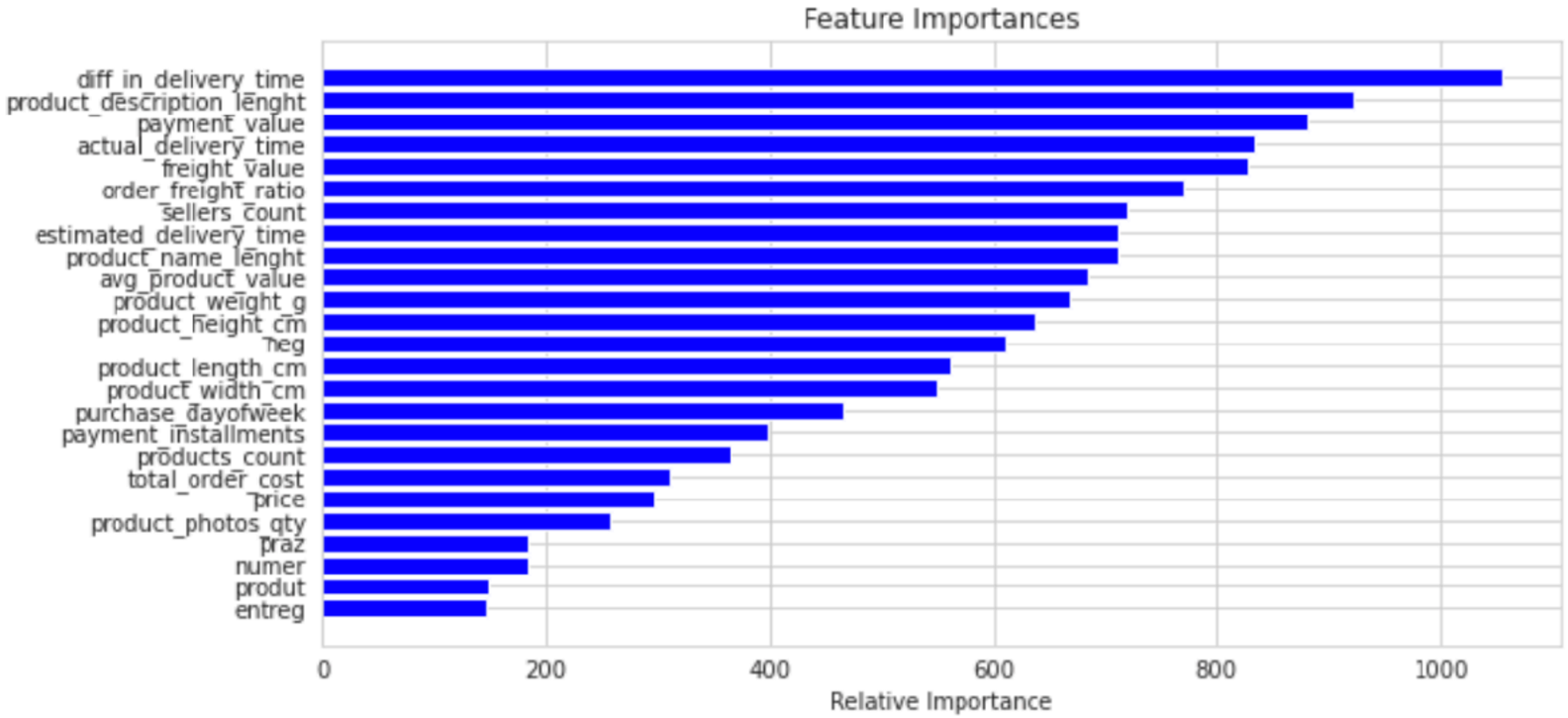
# A NEED TO UNDERSTAND DRIVERS OF REVIEW SCORES

- There is an obvious difference in how segments leave review scores.
- For any eCommerce ecosystem, reviews are critical towards nurturing the seller-buyer relationship.
- Therefore predicting customer review scores and understanding factors is the logical next step.



# 2 KEY LEARNINGS ON HOW TO BOOST REVIEWS FOR OUR HIGH-VALUE CUSTOMERS

- Reviews were binarised into 0 for negative and 1 for positive.
- Key predictor of satisfaction is the difference between the actual and estimated date. Customers will be more satisfied if the order arrives sooner than expected, or unhappy if received after expected.
- Effort in product description length is a strong predictor of positive reviews. This clearly shows that setting expectations can help to drive positive reviews.




***Model**: XGBoost | **F1 Score**: 94%*

# LIMITATIONS AND NEXT STEPS

- Limitations of this analysis -
    1. Data is limited to a sample. It would be more insightful (and computationally draining) if all of the data was available.
- Next steps for the business -
    1. Recommender system to build up customer basket size, based on RFM segmentation [WIP]
    2. A/B testing for the 3 segments to understand differences in campaign response

# THANK YOU