VINUNIVERSITY

FINAL REPORT

PRESENTED BY GROUP 6

Brazilian E-Commerce Public Dataset by Olist

OVERVIEW ABOUT THE DATASET

A 360-degree View of End-to-End Ecommerce Operations

Source

Kaggle – Brazilian E-Commerce Public Dataset by Olist

Size & Structure

100k+ rows across multiple relational tables (orders, products, customers, sellers, payments, reviews, geolocation)

Structured in a relational schema with foreign keys connecting each table

Key Variables

1 Orders Dataset

order_id, customer_id,
order_status,
order_purchase_timestamp,
order_delivered_customer_date

2 Order Items Dataset

product_id,
product_category_name, price,
freight_value

3 Order Payments Dataset

payment_type
payment_installments
payment_value

4 Order Reviews Dataset

review_score,
review_comment_title,
review_comment_message,
review_creation_date

5 Sellers Dataset

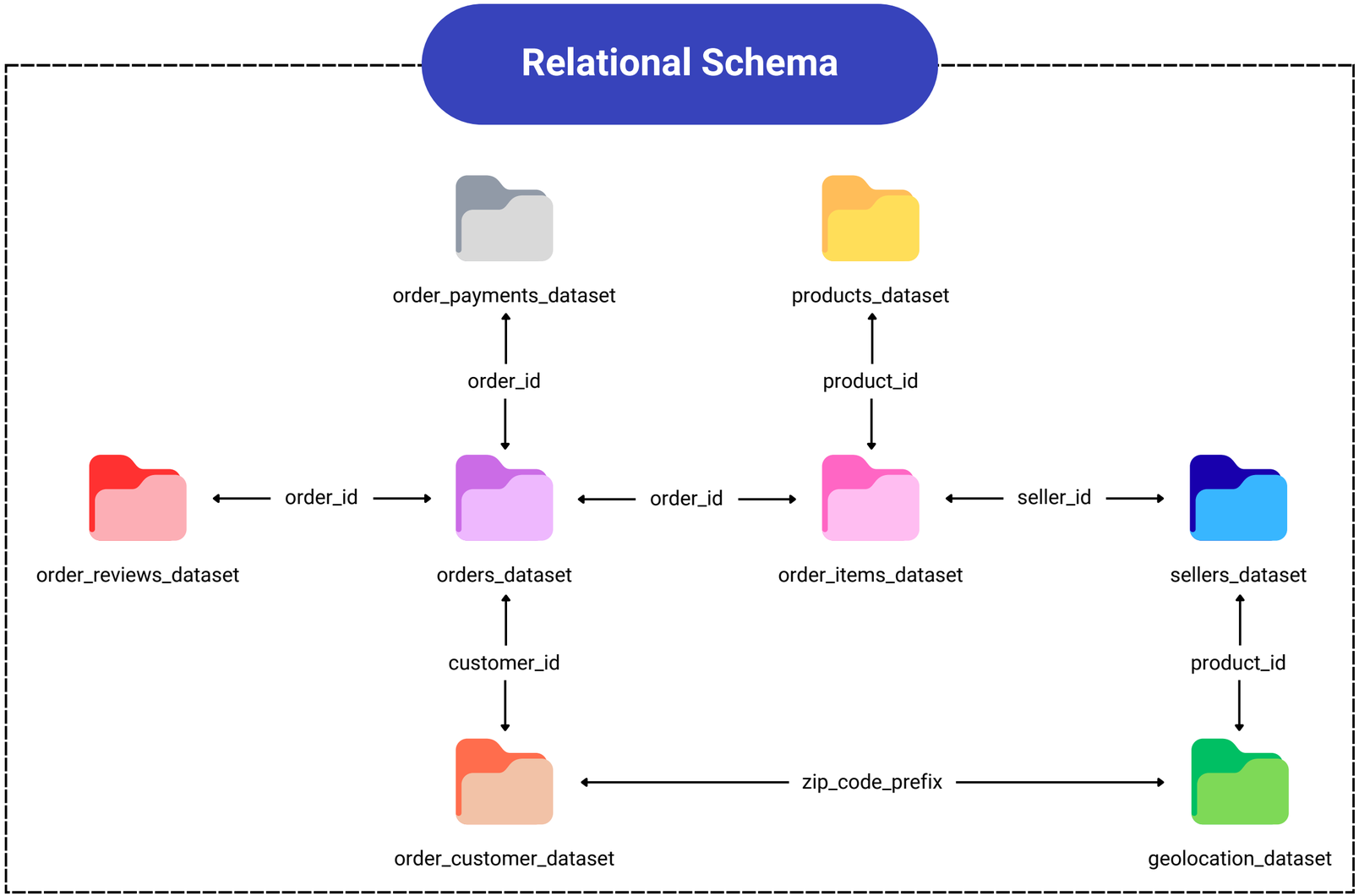
seller_id,
seller_city
seller_state

6 Geolocation Dataset

geolocation_zip_code_prefix,
geolocation_lat,
geolocation_state

OVERVIEW ABOUT THE DATASET

Structure of Relational Schema among the Datasets



The datasets are structured and connected through key fields like order_id, product_id, and zip_code_prefix. It highlights the relational database design with **orders_dataset as the central hub** linking to other tables.

Each dataset serves a unique purpose: tracking orders, payments, reviews, customer and seller details, and geolocation. They **enable a full analysis** of customer journeys, seller performance, and delivery logistics.

EVALUATION OF THE DATASET

Overall Description of Both Dataset Potential and Limitations

Potential for Business Decision Making

Holistic Coverage

Provides a 360° view of e-commerce operations: customer behavior, product categories, delivery logistics, etc.

Use Cases

Useful for developing BI dashboards to inform inventory planning, marketing strategies, etc.

Analysis Friendly

Enables analysis of customer satisfaction, logistics efficiency, pricing, and sales trends

Limitations & Challenges of Dataset

Outdated information

The dataset only covers transactions from 2016 to 2018. Trends and behaviors may have shifted significantly since then, especially due to post-COVID changes in shopping habits.

Brazil-Specific Context

The dataset is based entirely on the Brazilian market, so findings may not be directly applicable to other regions with different consumer behavior, logistics infrastructure, or payment ecosystems

No Real-Time or Live Data

The dataset is static and historical, which means it cannot be used for real-time decision-making, demand forecasting, etc.

INDUSTRY IDENTIFICATION - ECOMMERCE

Latest Trends & Challenges in Ecommerce Sector (BI-focused)

Customer Personalization		Group users based on demographics, purchasing behavior, or geography to tailor marketing
Delivery Optimization		Analyze order-to-delivery times and logistics costs to improve supply chain performance
Product Recommendation		Leverage purchase history and reviews to enhance cross-selling and up-selling
Customer Experience Design		Analyze reviews and return patterns to reduce churn and improve satisfaction
Fraud Detection & Security		Identify anomalies in payment types or order behaviors to prevent fraud

DATA EXPLORATION

Description of Dataset and Potential Insights

Orders Dataset

Description

This dataset contains information about customer orders, including the order status, timestamps, etc.. It forms the core timeline of the customer purchase journey.

Potential Insights

- Average delivery time vs. estimated delivery
- Frequency of different order statuses (e.g., canceled, delivered)
- Seasonal patterns in order volumes

Order Items Dataset

Description

This dataset provides item-level details for each order, including product IDs, seller IDs, pricing, and freight costs. It enables granular analysis of what customers are buying.

Potential Insights

- Best-selling products and categories
- Average price per item and shipping costs
- Seller-wise distribution and performance

Customers Dataset

Description

This dataset includes customer IDs and their geolocation information such as city and state. It helps track demand patterns across different regions.

Potential Insights

- Top customer locations by order volume
- Regional trends in payment methods or delivery delays

DATA EXPLORATION (cont.)

Description of Dataset and Potential Insights

Order Payments Dataset

Description

This table shows how each order was paid for, including payment type, value, and installment details. It supports financial and payment behavior analysis.

Potential Insights

- Most common payment methods (e.g., credit card vs. boleto)
- Average order value by payment type
- Use of installments and correlation with order size

Order Reviews Dataset

Description

This dataset stores customer feedback in the form of star ratings and optional comments, linked to specific orders. It is key to understanding customer satisfaction.

Potential Insights

- Average review scores per product category or seller
- Common issues based on text reviews
- Impact of delivery delays on ratings

Products Dataset

Description

Contains product metadata such as category, dimensions, and weight. It enables product-level trend analysis and logistics evaluation.

Potential Insights

- Popular categories and their performance
- Shipping costs based on product weight/size
- Pricing patterns by category

DATA EXPLORATION (cont.)

Description of Dataset and Potential Insights

Sellers Dataset

Description

This dataset provides location information of sellers, enabling supply-side analysis. It helps assess seller distribution and potential delays.

Potential Insights

- Seller concentration by region
- Seller impact on delivery speed and review scores
- High-performing vs. low-performing sellers

Order Reviews Dataset

Description

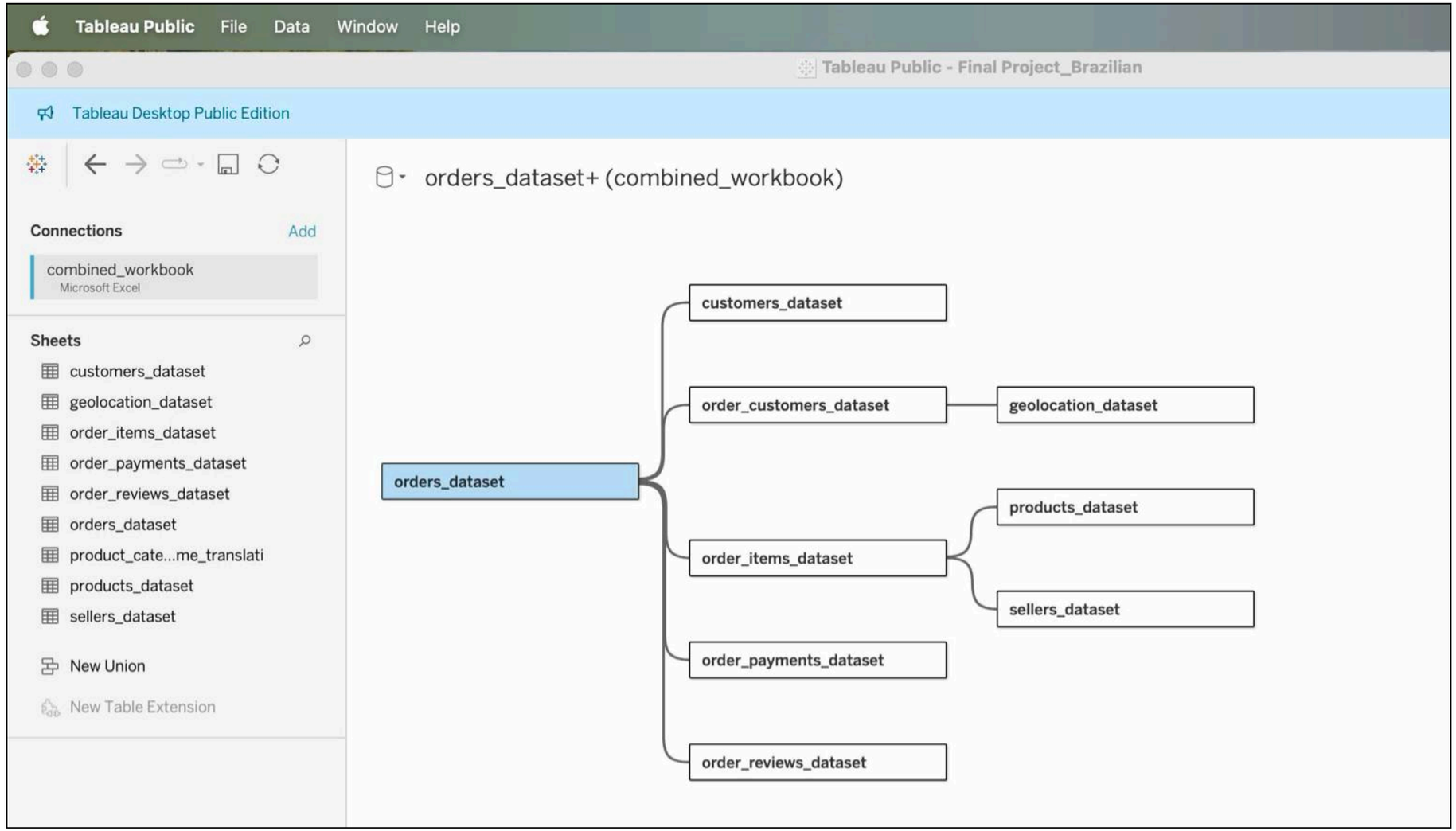
Provides lat/long coordinates and zip code prefixes for both customers and sellers. It is used for geographic visualizations and distance calculations.

Potential Insights

- Regional sales heatmaps
- Delivery bottlenecks in remote areas
- Correlation between geography and delivery delays

DATA EXPLORATION (cont.)

Datasets Matching to Identify Main Relationships and Create Theme

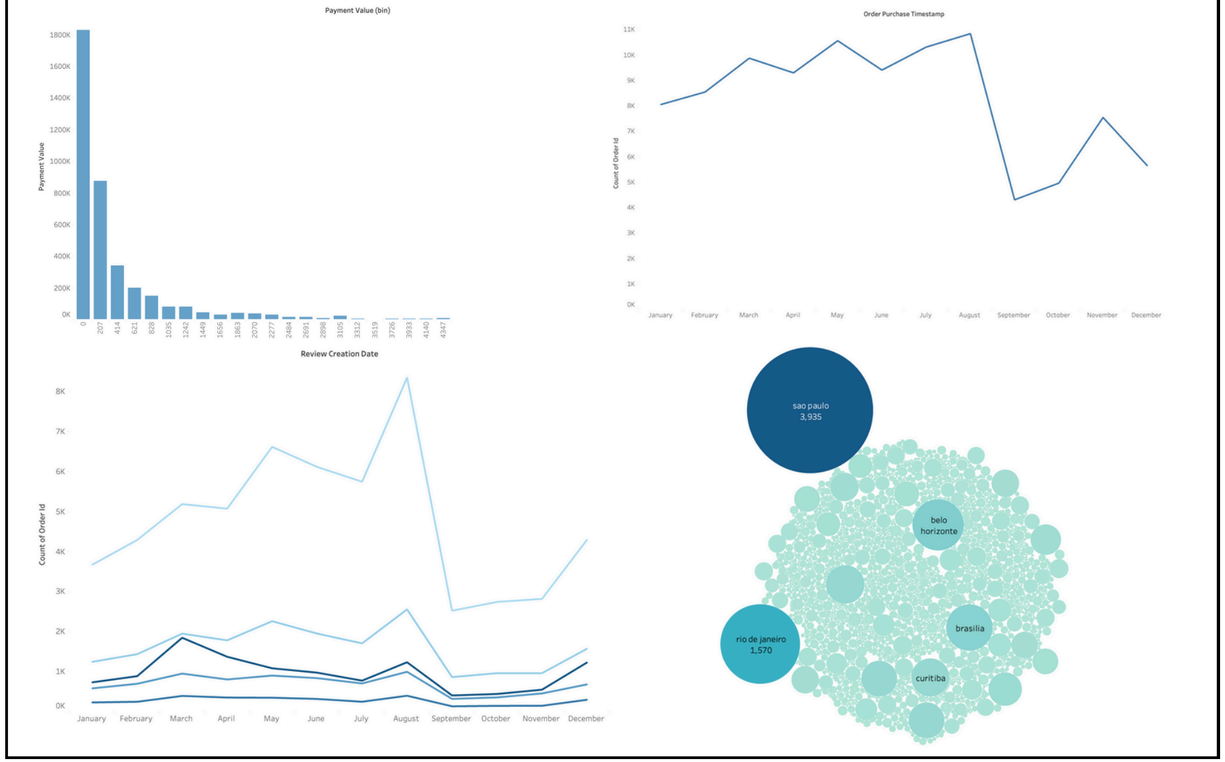


DATA EXPLORATION (cont.)

Data Preprocessing and EDA

TRANSLATE INTO ENGLISH → **EXPLORATORY DATA ANALYSIS** → **FILTER OUT ERROR DATA**

There are records in the category column that require us to replace with **English**-translated versions of the **Portuguese** strings



Data from September 2018 onward appears incomplete or incorrectly recorded, as values abruptly fall to zero. Therefore, these months have been **excluded**.

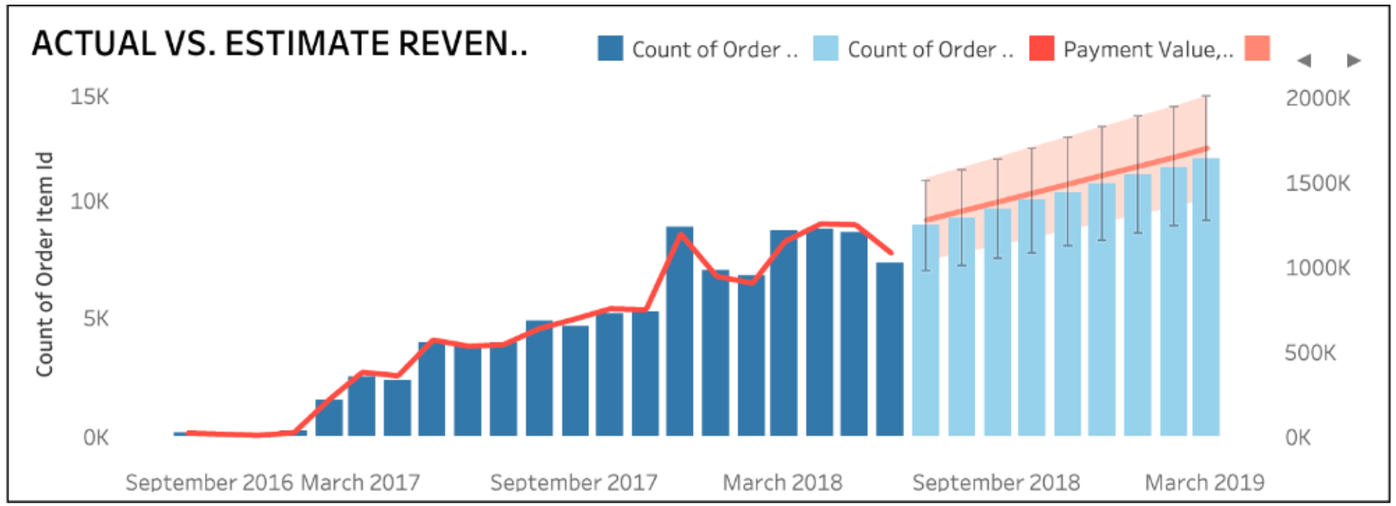
DASHBOARD OBJECTIVES & RESEARCH QUESTIONS

Identify Three Dashboards

	Target Audience	Objectives	KPIs & Metrics
Sales Overview Dashboard	Sales Directors, Regional Managers, Category/Product Managers	Analyze sales performance over time, understand top-selling products and regions, and guide sales strategy	• Sales Performance by Weekdays/ Over Years • Revenue by State • Best Selling Categories
Customer Experience Dashboard	Customer Experience Managers, Product Quality Assurance Analysts	Monitor customer satisfaction trends and identify areas for service and product improvement	• Average Review Score • Highest Rating Products • Review Response Time
Operational Dashboard	Business Operations Managers, Supply Chain Managers, Logistics Teams	The objective of this dashboard is to monitor and improve operational efficiency over KPIs (e.g. order tracking, delivery status)	• Revenue from Top 10 Category • Average Freight Value of Clustered Order Size • Late Orders by Cities

ADVANCED ANALYTICS

Trend Forecast



Options Used to Create Forecasts

Time series: Month of Review Answer Timestamp
Measures: Count of Order Item Id, Sum of Payment Value
Forecast forward: 9 months (July 2018 – March 2019)
Forecast based on: October 2016 – June 2018
Ignore last: 1 month (July 2018)
Seasonal pattern: None (Not enough data to search for a seasonal pattern recurring every 12 Months)

Count of Order Item Id

Initial July 2018	Change From Initial July 2018 – March 2019	Seasonal Effect High	Seasonal Effect Low	Contribution Trend	Contribution Season	Quality
8,865 ± 1,896	2,860	None		100,0%	0,0%	Poor

Sum of Payment Value

Initial July 2018	Change From Initial July 2018 – March 2019	Seasonal Effect High	Seasonal Effect Low	Contribution Trend	Contribution Season	Quality
1,279,545 ± 242,502	423,283	None		100,0%	0,0%	Poor

Trend-Only Forecast

With just 21 months of history, Tableau's automatic model found **no reliable 12-month seasonality.** Both order count and revenue are driven entirely by an additive trend component (ETS = (N,T,N), simple additive error model)

Wide Confidence Intervals

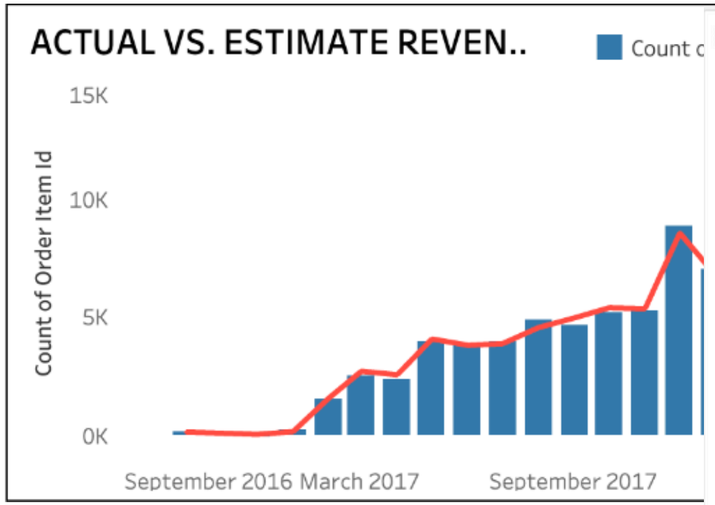
The ±1,896 orders and ±R$242 K revenue bands reflect **high uncertainty**. This "Poor" rating stems from limited data and hold-out validation error.

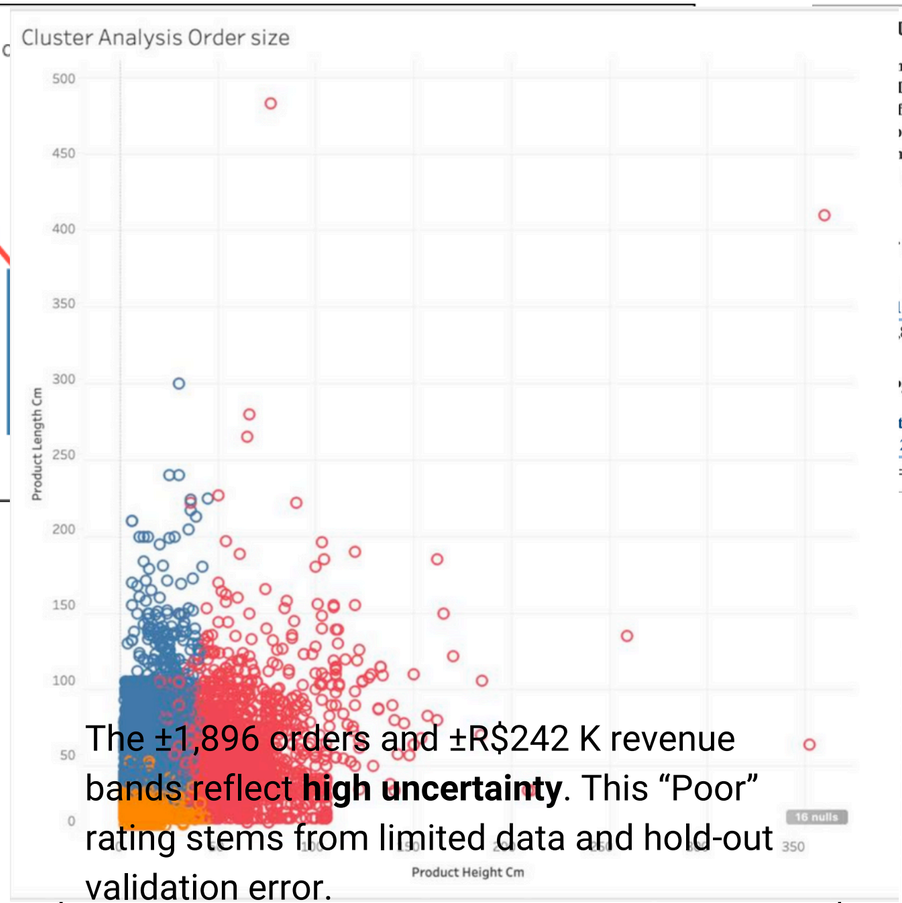
Projected Growth

A clear **upward trajectory** emerges from 2017 through late 2018, and is expected to continue into early-2019.

ADVANCED ANALYTICS

Trend Forecast





Used to Create Forecasts

ıe series: Month of Review Answer Timestamp
leasures: Count of Order Item Id, Sum of Payment Value
forward: 9 months (July 2018 – March 2019)
ıased on: October 2016 – June 2018
ıore last: 1 month (July 2018)
pattern: None (Not enough data to search for a seasonal pattern recurring every 12 Months)

'Order Item Id

18	Change From Initial July 2018 – March 2019	Seasonal Effect High	Seasonal Effect Low	Contribution Trend	Contribution Season	Quality
,896	2,860	None		100,0%	0,0%	Poor

'ayment Value

tial 2018	Change From Initial July 2018 – March 2019	Seasonal Effect High	Seasonal Effect Low	Contribution Trend	Contribution Season	Quality
± 242,502	423,283	None		100,0%	0,0%	Poor

Trend-Only Forecast

With just 21 months of history, Tableau's automatic model found **no reliable 12-month seasonality.** Both order count and revenue are driven entirely by an additive trend component (ETS = (N,T,N), simple additive error model)

The ±1,896 orders and ±R$242 K revenue bands reflect **high uncertainty**. This "Poor" rating stems from limited data and hold-out validation error.

Projected Growth

A clear **upward trajectory** emerges from 2017 through late 2018, and is expected to continue into early-2019.

THANK YOU FOR READING

Group 06
Pham Phuong Hoa
Dang Tran Dang Khanh
Do Minh Quan
Nguyen Trang Nhung
Hoang Nguyen Gia Huy
Tran Tuan Viet

The team would like to express our appreciation to the Teaching Team, for supporting us in this assignment.

Prof. Abhishek Nayak
Lecturer of BANA4010

Mr. Nguyen My Linh
Teaching Assistant of BANA4010