GlobalMart Business Overview

Executive Summary (TL;DR)

GlobalMart generated approximately **6.69 million orders** and generated **~$4.42 million in revenue** from the start of the year up to **May 9, 2018**. The average transaction value was **~$50.83** (median **~$52.64**). About **20% of orders** included a discount, with average monthly discount spending at **~$46,000** (roughly **0.0045% of total revenue**).

Data Preparation

- **Re-labeled product taxonomy**: Transformed the original product list into a structured **29-category** hierarchy (e.g., Alcohol, Apparel, Baker, etc.).
- **Standardized country codes**: Converted all country identifiers to **ISO-3166** standard and cleaned latitude/longitude data for all **96 cities**.
- **Temporal data cleaning**: Identified and removed **1% of records without timestamps** (67,000 orders = **~$3.42 million in revenue**) from the analysis.

Business Performance Insight

- **Top-performing category**: Alcohol (accounted for ~10.8% of orders and ~9.9% of revenue).
- **Lowest-performing categories**: Energy Drink, Apparel, and Medicine (each contributing **less than 5%** in both orders and revenue).
- **Price tier distribution**: Balanced across low, medium, and high price ranges — each tier made up **approximately 34%** of total products.

Forecasting Result (Facebook Prophet, Daily Granularity)

- **Orders**: MAE = 153.51, MAPE = 0.30% -> **forecast ~18.92M** orders for the full year.
- **Revenue**: MAE = 95,599.57, MAPE = 0.30% -> **forecast $ ~965M** revenue for the full year.

Document Log

Actions	Date
Initial Document	2025-09-03

Background

The executive leadership of GlobalMart is preparing for a high-level strategic meeting to define the company's next growth direction. To enable data-driven decision-making, the data team will deliver comprehensive insights and analytical support that illuminate current business performance, customer behavior, and product trends across the grocery portfolio. This foundation will empower executives to evaluate opportunities, mitigate risks, and set actionable priorities for the upcoming fiscal year.

Problem Statement

Despite having a rich set of transactional and master data (e.g., sales, customers, products, locations), GlobalMart currently lacks a unified view that transforms these data sources into actionable intelligence. Key challenges include:

- **Fragmented data assets**: Seven raw data tables are siloed across systems, with inconsistent naming conventions and incomplete attribute definitions—making integration and analysis difficult.
- **Inadequate category alignment**: The original product-category taxonomy does not accurately reflect the true product assortment mix, limiting the relevance and accuracy of category-level insights.
- **Limited forecasting capability**: No robust time-series forecasting model is in place to predict core business metrics—such as order volume and revenue—over the full fiscal year, hindering proactive planning and strategic resource allocation.

These gaps prevent leadership from confidently identifying performance drivers, understanding customer behavior, and making forward-looking decisions with certainty

Objective and Goals

The primary objective is to deliver an end-to-end deep-dive analysis of GlobalMart's grocery sales that:

- **Reviews historical business performance** across key dimensions—product, geography, and customer segment.
- **Generates accurate forecasts** for the 2018 fiscal year using a proven time-series model (Facebook Prophet).
- **Creates a reusable analytical framework**, including cleaned and enriched data tables, that can be extended for future reporting cycles.

This comprehensive analysis will provide the executive team with the insight and confidence needed to set strategic priorities and drive sustainable growth.

Analysis Methodology

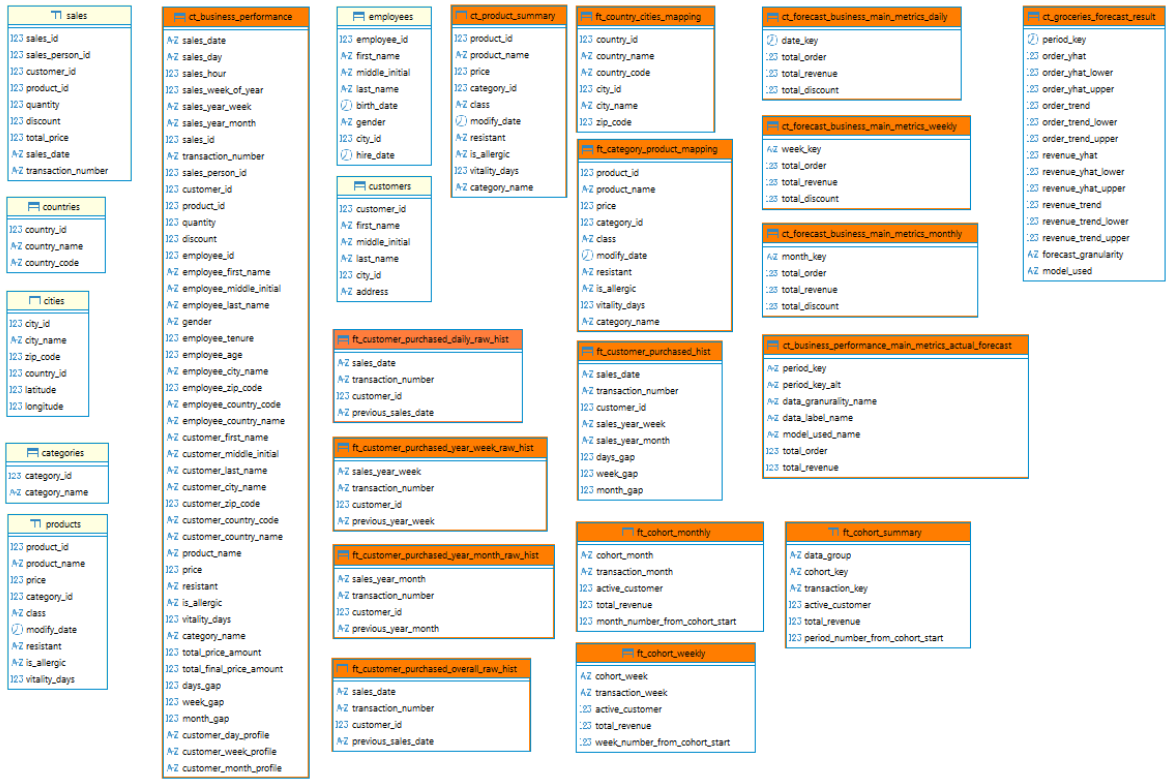
Data Preparation

Data Overview

Table Name	Description	Total Data
categories	Reference list of product categories used throughout the dataset.	29
cities	Reference list of cities, including latitude and longitude coordinates.	96
countries	Reference list of countries.	205
customers	Master table containing basic customer profiles.	98,759
employees	Master table containing basic employee profiles.	23
products	Catalog of available products and their attributes.	452
sales	Transactional table with all order details and purchase information.	6,756,125

Table 01 - Groceries Origin Data Sources

Based on the seven data points, generate the additional seventeen data points that will be used to perform analysis, derive insights, and forecast the primary business metrics. The schema details of the data points, including an ER diagram, can be presented as shown below.



Picture 01 - ER Diagram

Data Cleaning and Re-Labeling

There are four datasets being adjusted to ensure accurate and meaningful analysis, with details as follows:

Table: category

The objective is to redefine the list of categories to ensure better alignment with product names.

Old Category List	New Category	
1. Beverages 2. Cereals 3. Confections 4. Dairy 5. Grain 6. Meat 7. Poultry 8. Produce	1. Alchohol 2. Apparel 3. Bakery 4. Baking 5. Beverages 6. Cake & Tarts 7. Cereals 8. Cleaning	16. Frozen Food 17. Fruits 18. Herbs & Spices 19. Household Items 20. Instant Food 21. Kitchen Supply 22. Meat & Poultry 23. Medical

9. Seafood 10. Shell fish 11. Snails	9. Condiments & Oils 10. Cookies & Biscuits 11. Cooking 12. Dairy 13. Dessert 14. Energy Drink 15. Food Ingredients	24. Processed Food 25. Produce 26. Seafood 27. Snack 28. Staple Foods 29. Vegetables

Table 02 - Comparison of Old Category Vs New Category

Table: product

There are a few adjustments that have been made for:

Column Name	Notes
category_id	The product's category ID, updated according to the latest category reference.
resistant	Indicates the product's resistance level to spoilage, categorized as: Weak - Medium - High. The value depends on the product's category.
is_allergic	Specifies whether the product contains ingredients that may trigger allergic reactions: ● **TRUE** → Contains potential allergens ● **FALSE** → No known allergens
vitality_days	Represents the expected shelf life or "freshness duration" of the product, based on its type: - **Fresh Products** (e.g., fresh meat, dairy): **2 to 7 days** - **Agricultural Products** (e.g., fresh fruits, vegetables): **7 to 21 days** - **Dry or Preserved Products** (e.g., spices, dried goods): **180 to 365 days** - **Non-Food Items** (e.g., household goods): **9,999 days** (non-perishable) - **Ready-to-Eat Products** (e.g., pre-packaged meals): **1 to 3 days** - **Other Products**: **14 to 90 days**

Table 03 - Logic Applied and notes of adjustment in the table product

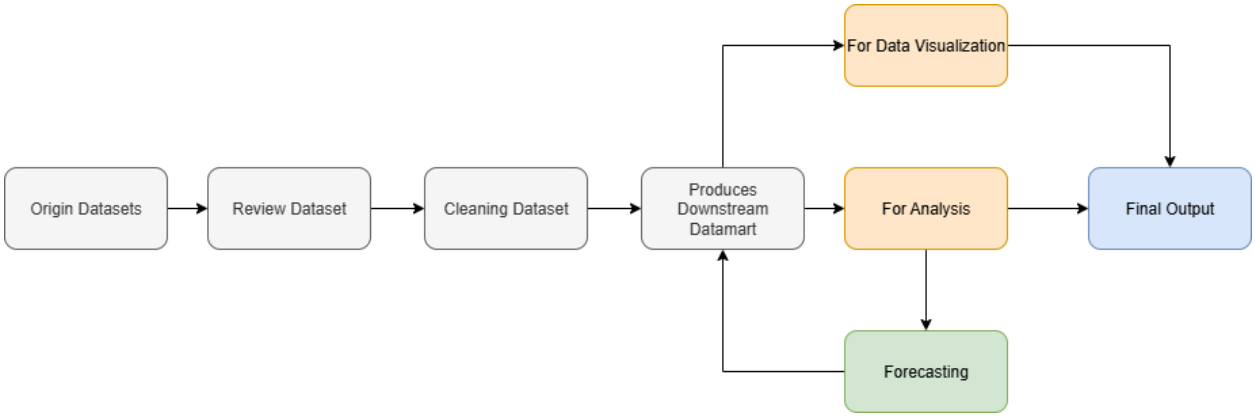
Table: country

country_code – Updated to follow ISO 3166, the internationally recognized standard for country codes. This ensures that the values are consistent with the global naming convention for nations.

Table: cities

longitude and latitude – Now formatted to be fully compatible with data visualization tools. The coordinates are stored in standard decimal degrees (e.g., -6.2088, 106.8456), ensuring they can be accurately plotted on maps and visualized without issues.

Overall flow



Picture 02A -Overall Flow Analysis

Figure 02A illustrates the end-to-end workflow of the project.

1. **Ingestion & Initial Review** – Receive the raw source dataset and examine it thoroughly. This step involves:
    a. Mapping the relationships between tables,
    b. Understanding the meaning and permissible values of each column, and
    c. Conducting a sanity-check to flag obvious errors or inconsistencies.
2. **Data Cleansing** – Apply transformations to bring the data into a consistent, standardized format, which includes:
    a. Normalizing data types,
    b. Standardizing naming conventions
    c. Redefining category values where necessary.
3. **Datamart Construction** – Create a downstream datamart that will serve as the primary source for analytical workloads, especially forecasting. The datamart is also designed to receive periodic refreshes (re-pushed data).
4. **Deep Dive Analysis -** Perform analysis to extract insight and information from the data, including the forecasting data.
5. **Visualization Development** – Build interactive visualizations on top of the datamart to enable stakeholders to explore insights easily.

Dataset Produced

Several datasets have been generated and are now available for analysis and data visualization.

Dataset Produced	Dataset Sources
ft_customer_purchased_hist	- sales
ct_business_performance	- categories - cities - countries - employees - customers - products - sales - ft_customer_purchased_hist
ct_product_summary	- products - categories
ft_country_cities_mapping	- countries - cities
ft_category_product_mapping	- products - categories
ft_customer_purchased_daily_raw_hist	- sales
ft_customer_purchased_year_week_raw_hist	- sales
ft_customer_purchased_year_month_raw_hist	- sales
ft_customer_purchased_overall_raw_hist	- sales
ft_cohort_weekly	- ct_business_performance
ft_cohort_monthly	- ct_business_performance
ft_cohort_summary	- ft_cohort_weekly - ft_cohort_monthly
ct_forecast_business_main_metrics_daily	- ct_business_performance
ct_forecast_business_main_metrics_weekly	- ct_business_performance
ct_forecast_business_main_metrics_monthly	- ct_business_performance
ct_groceries_forecast_result	The forecasting results are pushed back to the database using Python.
ct_business_performance_main_metrics_actual_forecast	- ct_business_performance

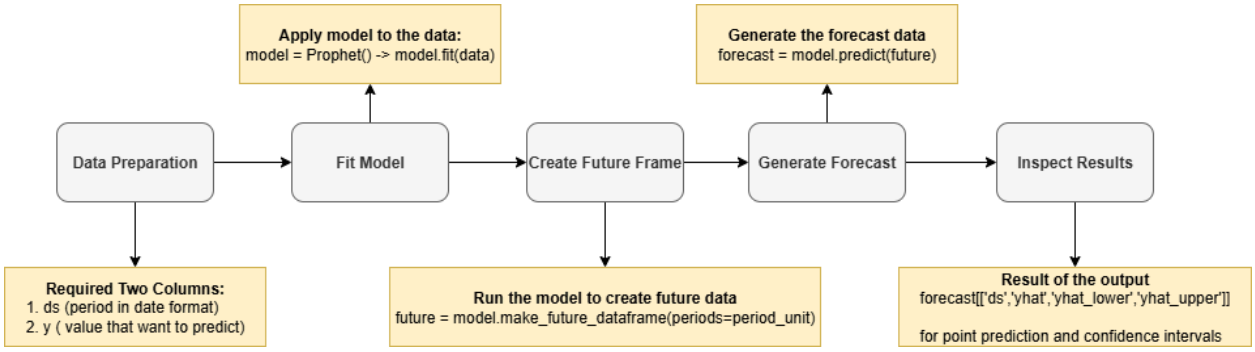
	- ct_groceries_forecast_result

Table 04 - List of Tables Downstream and Upstream

Forecasting

Forecasting involves using historical data to predict future values of a time-dependent metric (e.g., orders or revenue). A forecasting model captures trends, seasonality, and irregular patterns, then projects these patterns forward in time. This capability enables businesses to anticipate performance gaps over specific periods and plan accordingly.

In this project, **Facebook Prophet** is employed as the forecasting model. As an open-source library, it supports easy and robust forecasting across multiple time granularities—including daily, weekly, and monthly intervals—and effectively handles strong seasonal effects and holiday impacts.



Picture 02B -Overall Process of Prediction

Analysis Result

Business Overview

The available data span the period from **2018-01-01 to 2018-05-09**. During this interval, **1 % of the records lack a transaction timestamp**. Consequently, these unidentified entries will be omitted from subsequent analysis and forecasting. The missing **1 % corresponds to 67,526 orders**, which translates to **approximately $3,424,911 in revenue**.

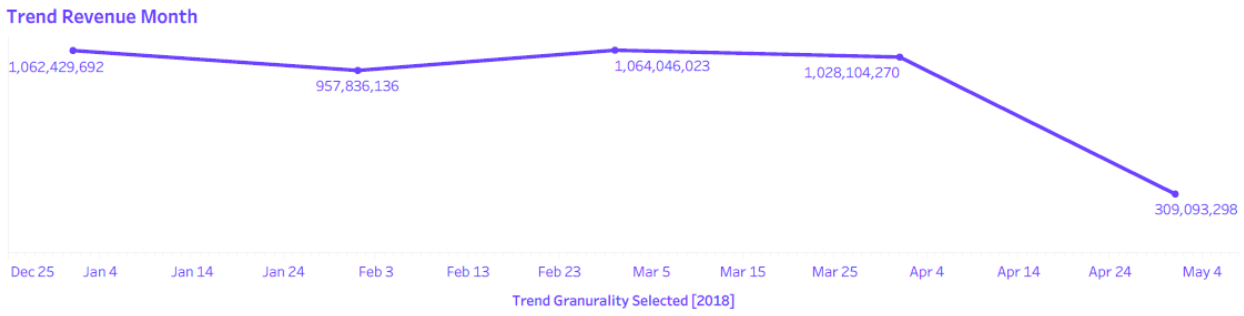
Overall Business Performance

In 2018, operations were confined to the United States, spanning approximately **96 cities** nationwide. The product portfolio comprises **27 categories** and **452 distinct items**. The dataset contains **98,759 unique customers** supported by **29 front-line employees**. On **average**, each customer places about **~4 purchases per week** (~12 purchases per month).

In 2018, up to the most recent data point, 129 days have elapsed – roughly ~35 % of the year – leaving about 236 days (~ 65 %) remaining until year-end. To date, the business has generated **~ 6.69 million unique orders**, corresponding to **~ $4.421 million in revenue**.

- **Average transaction value**: $ 50.83
- **Median transaction value**: $ 52.64
- **Gap between average and median**: ~ 3 % (about $ 1.81)

The modest 3 % difference indicates a relatively tight central tendency, even though the distribution is slightly left-skewed (more low-value orders than high-value ones). Consequently, extreme outliers are scarce, and most transactions cluster around the typical basket size.



Picture 03 - Trend Revenue Month-on-Month

Based on **Picture 03**, the business showed a fluctuating pattern throughout the observed period, with a pronounced dip in May.

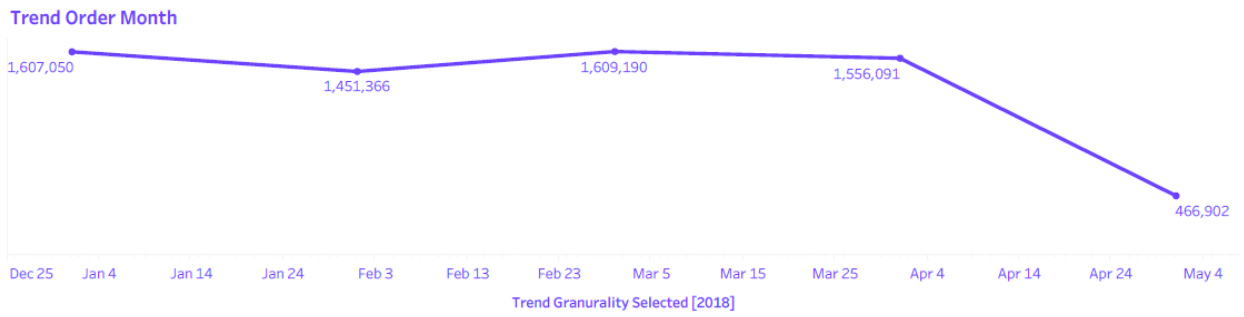
- **February**: orders declined by approximately ~9 % compared with the previous month.
- **March**: orders rebounded, increasing by roughly ~11 % month-on-month.
- **May**: the apparent drop is largely attributable to the data cut-off at 2018-05-09; the month is not fully captured, so the figures underestimate the true activity for May.

Overall, the trend indicates short-term volatility, with February representing the steepest decline and March the strongest recovery, while the incomplete May data should be interpreted with caution.

When the metric is shifted to **revenue per day**, the picture becomes clearer and more equitable. The daily-growth analysis shows that the business is **highly stable**: the day-to-day changes in revenue are all **under 1 %**, alternating between slight decreases and slight increases. Details can be found in **Table 05** below:

Period	Total Days	Total Revenue in Month	Total Revenue per day	Growth Per Month	Growth per Day
Jan	31	1,062,429,692	34,271,926		
Feb	28	957,836,136	34,208,433	-9.84%	-0.19%
Mar	31	1,064,046,023	34,324,065	11.09%	0.34%
Apr	30	1,028,104,270	34,270,142	-3.38%	-0.16%
May	9	309,093,298	34,343,700	-69.94%	0.21%

Table 05 - Comparison MoM and Per Day Revenue

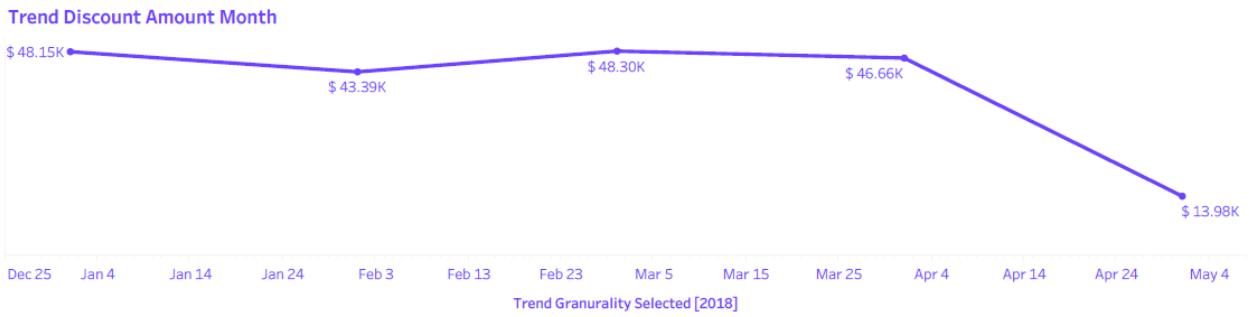


Picture 04 - Trend Order Month-on-Month

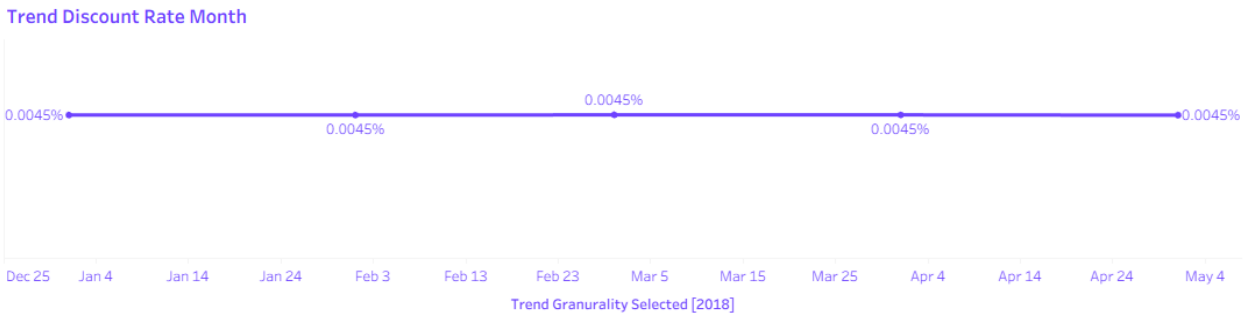
Based on the month-on-month analysis, the **order volume** showed a ~9 % decline in February, followed by a ~10 % increase in March. However, when the metric is shifted to **orders per day**, the apparent volatility disappears – daily changes stay within ~1 %, indicating a very stable ordering pattern throughout the period. Details can be found in Table 06 below:

Period	Total Days	Total Order in Month	Total Order per day	Growth Per Month	Growth per Day
Jan	31	1,607,050	51,840		
Feb	28	1,451,366	51,835	-9.69%	-0.01%
Mar	31	1,609,190	51,909	10.87%	0.14%
Apr	30	1,556,091	51,870	-3.30%	-0.08%
May	9	466,902	51,878	-70.00%	0.02%

Table 06 - Comparison MoM and Per Day Order



Picture 05 - Trend Discount Amount Month-on-Month

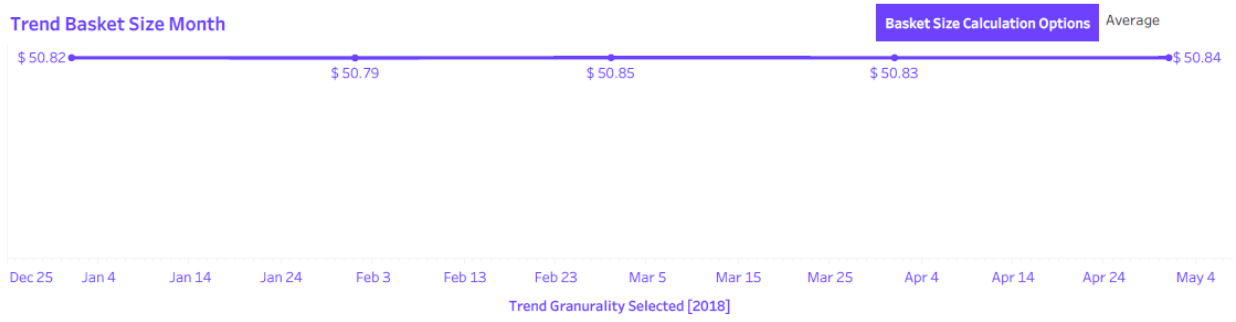


Picture 06 - Trend Discount Rate Month-on-Month

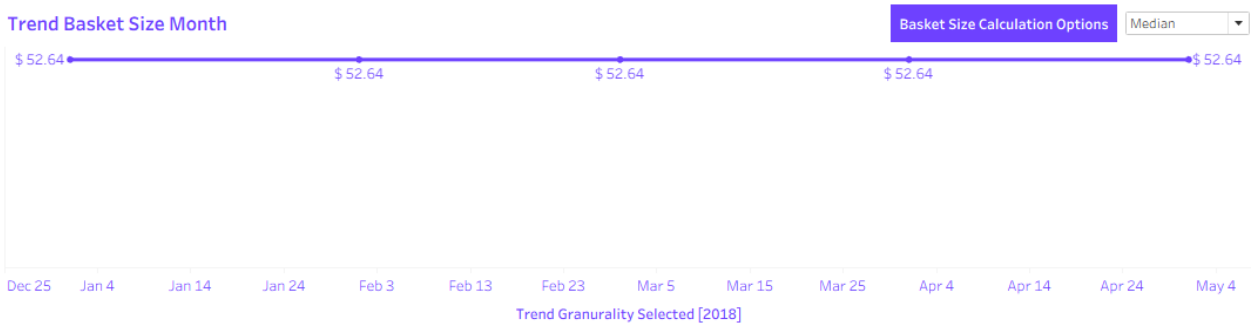
The data presented in **Picture 05** and **Picture 06** indicate that the company consistently allocates funds to discounts. On **average**, **monthly discount** outlays amount to **~ $ 46 K**, which translates to roughly **~ $1.5 K per day**. The proportion of discount expense relative to total

revenue remains stagnant at **0.0045 %**, demonstrating that the discount burden is minimal compared with overall sales.

Across the entire observation period, **~ 20 % of all orders include a discount**, and this share stays stable month-to-month, showing no notable fluctuation.



Picture 07 - Trend Basket Size Month-on-Month (Average)

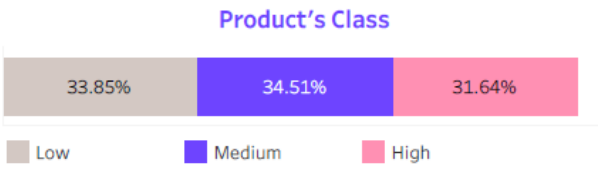


Picture 08 - Trend Basket Size Month-on-Month (Median)

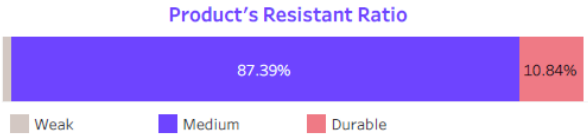
The term *basket size* denotes the total amount spent by a customer in each transaction. Across the observed period, the **month-on-month average basket size** was **~ $ 50.83**, while the **median basket size** stood at **~$ 52.64**. When the orders are segmented by discount status, the difference in basket size is marginal—only about **0.7 %**. Specifically, orders that included a discount had an average basket size of **~ $ 50.79**, whereas orders without any discount recorded a slightly higher average of **~ $ 50.83**. This negligible variance indicates that the presence of a discount does not materially affect the amount customers spend per transaction.

Product and Category Insight

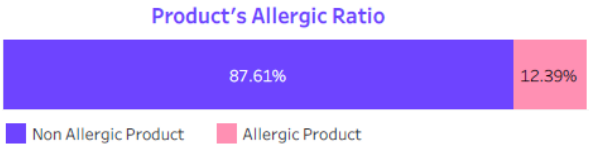
The company offers a catalog of **452 unique products** distributed across **29 categories**. The **average price** per product stands at **~ $50.8**, while the **median** price is slightly higher at **~ $52.5**, indicating a fairly balanced price distribution with only modest skewness toward higher-priced items



Picture 09 - Product's Class



Picture 10 - Product's Resistant Ratio



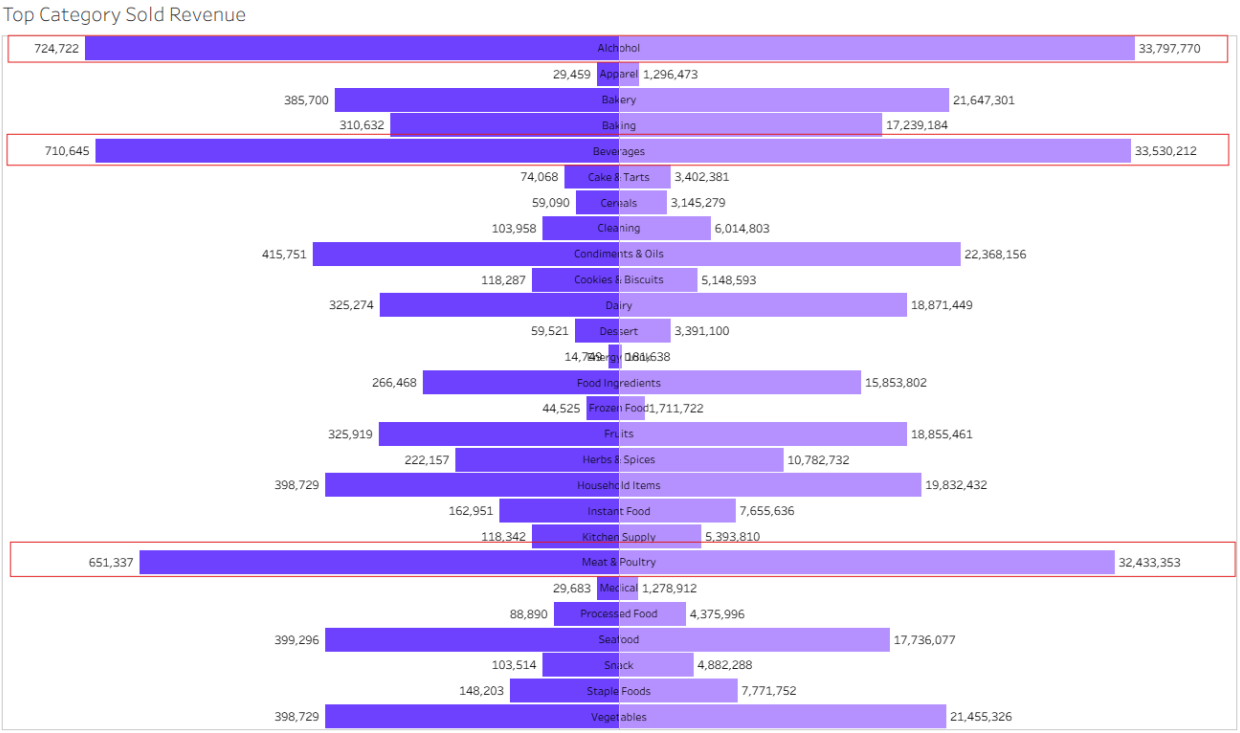
Picture 11 - Product's Resistant Ratio

The product class denotes a grouping of items according to their perceived value or price segment. Across the entire catalog, the distribution of these classes is notably balanced, indicating that the company offers a full spectrum of price tiers—low, medium, and high—to meet the purchasing preferences of customers in every segment. This equilibrium suggests a deliberate strategy to cater equally to price-sensitive shoppers as well as those seeking premium offerings, thereby supporting a diversified market reach. The details of the class are:

- **Low (33.85%)** for everyday staples or budget-friendly items.
- **Medium (35.51%)** for mid-range products in terms of price and perceived value.
- **High (31.64%)** for premium, specialty, or high values.

The product's resistance refers to the durability or shelf life of a product.

- **Weak (~1.77%)**: Perishable items with very short shelf-life.
- **Medium (~87.39%)**: Products with moderate shelf-life.
- **Durable (~10.84%)**: Refers to non-perishable products with a long shelf-life.



Picture 12 - Product Most Ordered and Most Generates Revenue

Based on the butterfly chart presented in Picture 12, it's evident that certain categories have consistently led in terms of both order volume and revenue over the observed period. The top-performing categories are:

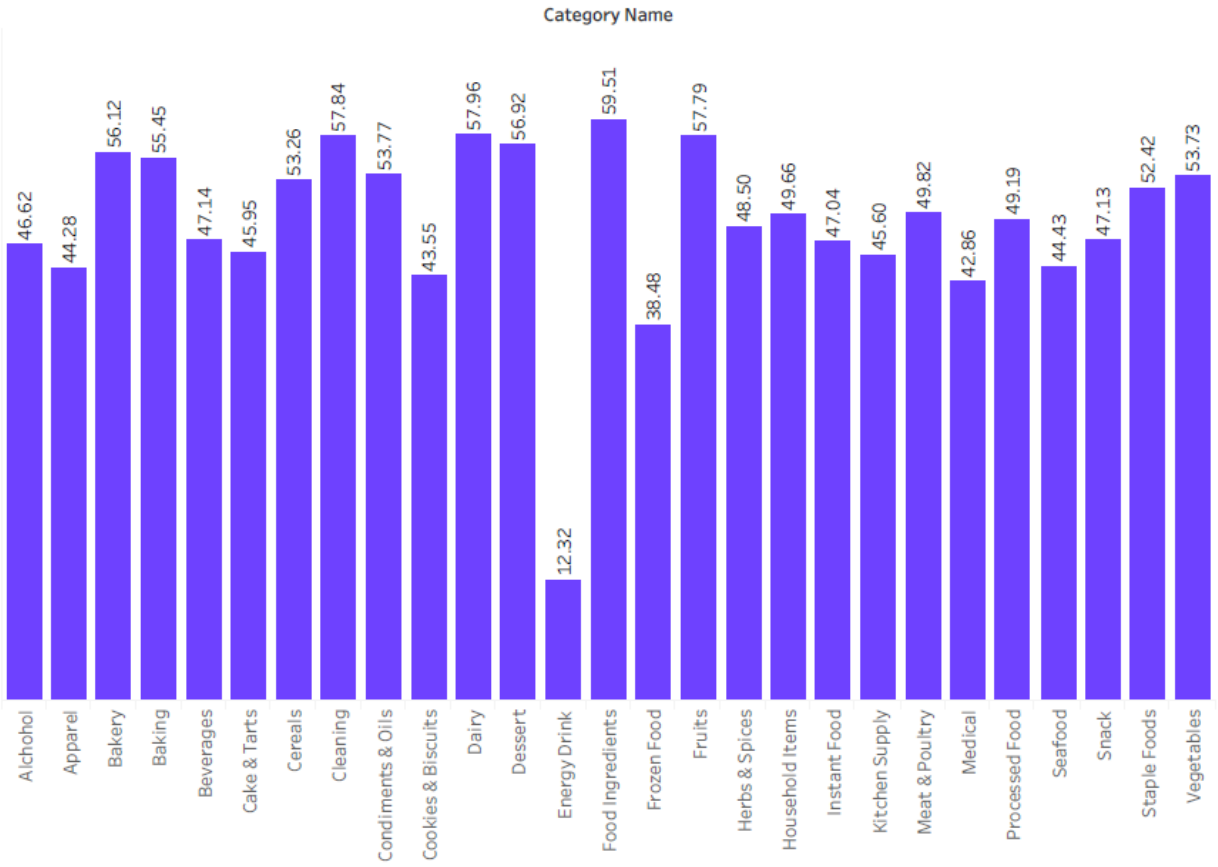
- Alcohol: This category demonstrates strong leadership with approximately 10.83% of total orders and 9.94% of total revenue.
- Beverages: Following closely, Beverages account for around 10.62% of orders and 9.86% of revenue, indicating its significant contribution.
- Meat & Poultry: This category secures a notable position with approximately 9.73% of orders and 9.54% of revenue.

Conversely, the analysis also highlights categories that have contributed the least to the overall order and revenue figures:

- Energy Drink: This category shows the lowest contribution, with roughly 0.22% of orders and 0.05% of revenue.

- Apparel: With about 0.44% of orders and 0.38% of revenue, Apparel is among the least contributing categories.
- Medicine: Similar to Apparel, Medicine accounts for approximately 0.44% of orders and 0.38% of revenue.

This distribution underscores the varied performance of different product categories, indicating areas of strength and those that may require strategic attention.



Picture 13 - Category's Price Unit

Each category may contain many products with a wide range of prices. To simplify the understanding of pricing at the category level, the **Unit Price** metric is used. This metric provides an approximate average price among all products within a category, offering a clear overview without the complexity of individual product prices.

The Unit Price is calculated using the formula:

$$\textbf{\textit{Unit Price}} = \frac{\Sigma\, Total\ Price\ in\ Category}{\Sigma\, Total\ Unique\ Product}$$

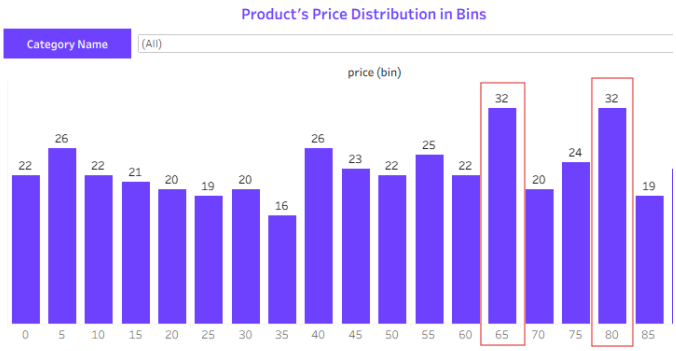
Formula 01 - Unit Price Formula

By utilizing the Unit Price metric, it becomes easier to identify pricing patterns across categories. For example, based on the given data:

- **The top categories** with the **highest** unit prices are **Food Ingredients** ($59.51), **Dairy** ($57.96), and **Cleaning** ($57.84).
- **The categories** with the **lowest** unit prices include **Energy Drink** ($12.32), **Frozen Food** ($38.48), and **Medical** ($42.48).

This metric provides valuable insight into how product prices vary across categories, enabling better pricing strategy and category management.

Picture 14 illustrates the distribution of product prices across categories using **price bins with an interval of $5**. **On average**, each price **bin contains** approximately **22 products**, with a median value close to this number as well. Additionally, about **40% of all bins contain more than 22 products**, indicating a moderate concentration of products within certain price ranges.



Picture 14 - Product's Price Distribution in Bins for All Categories

Notably, there are **extreme outliers** in the price distribution **at the $65 and $80 bins**, each **containing 32 product IDs**. These outliers could represent premium or specialty items within the catalogue.

Given the variability in price ranges across categories, this analysis opens opportunities for deeper insights into pricing and marketing strategies. For example, further study using market basket size analysis could help identify potential upsell opportunities by targeting specific product segments effectively.

Move to other topics. Product vitality refers to the number of days a product remains safe and in good condition for consumption or use, assuming proper handling, such as avoiding exposure to heat or direct sunlight.

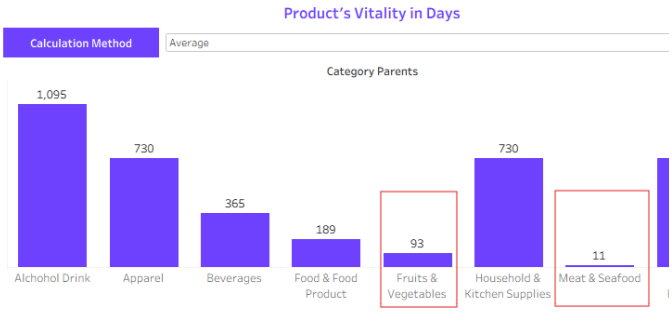
Most product categories report both the average and median vitality days, with these numbers generally aligning closely—except in the cases of **Fruits & Vegetables** and **Meat & Seafood** categories, which show noticeable differences between the two metrics.

A closer examination reveals the following details within these parent categories:

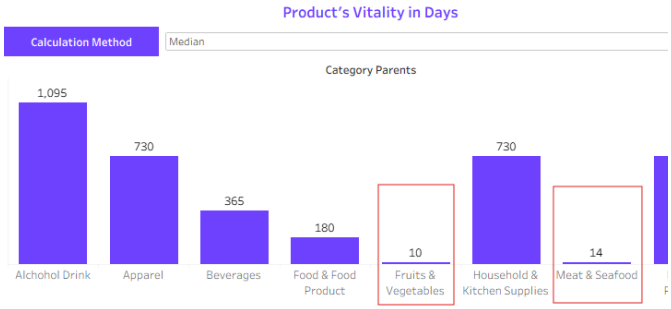
- Fruits & Vegetables:
  - Fruits: 10 days
  - Vegetables: 10 days
  - Herbs & Spices: 365 days
- Meat & Seafood:
  - Meat & Poultry: 14 days
  - Seafood: 7 days

The discrepancy observed in the **Fruits & Vegetables** category is mainly caused by the long vitality of **Herbs & Spices**. Meanwhile, in the **Meat & Seafood** category, the difference arises primarily due to the vitality variation between **Meat & Poultry** and **Seafood**.

This insight is useful for managing inventory and shelf-life expectations within these product groups.



Picture 15 - Product Vitality in Days (Average)



Picture 16 - Product Vitality in Days (Median)

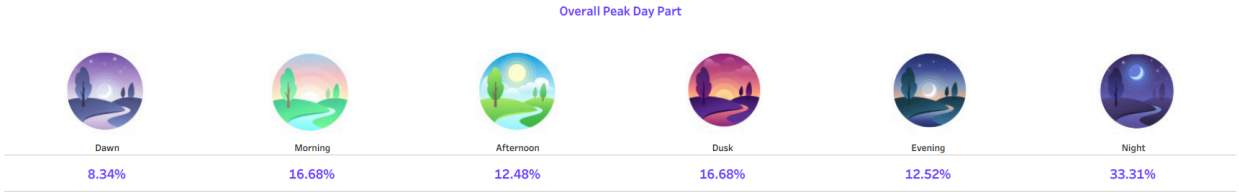
Customer's Behaviour Insight

This segment focuses on understanding the interaction dynamics between customers and the business. By examining the store's peak operational times, a typical day is divided into six distinct time-of-day categories, which will be detailed as follows:

No	Day Part Name	Time Range
1	Dawn	06:00 - 07:59
2	Morning	08:00 - 11:59
3	Afternoon	12:00 - 14:59
4	Dusk	15:00 - 18:59
5	Evening	19:00 - 21:59
6	Night	22:00 - 05:59

Table 07 - Day Part Definition Time

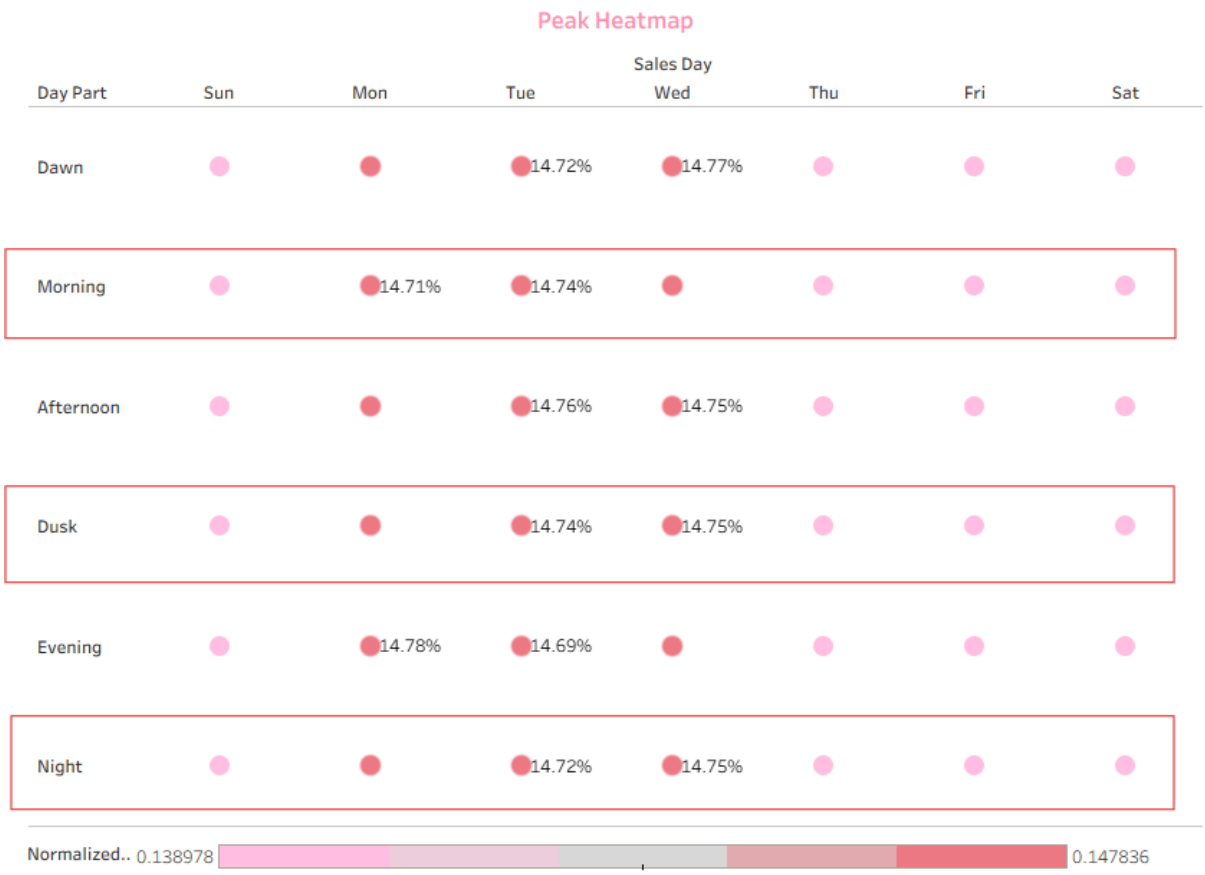
Picture 17 illustrates key transactional patterns, revealing that the **majority of transactions occur during the Night period**, contributing **approximately 33% of the total**. This is followed by **Dusk and Morning**, each accounting for around **16% of transactions**. This distribution highlights specific periods of heightened customer engagement and transactional volume within the store's operational hours.



Picture 17 - Overall Peak Day Part

A more in-depth analysis, combining dayparts with specific days and using **normalized transactions as the primary metric**, reveals that **most peak** activity occurs **midweek**, specifically from **Monday to Wednesday**. Within these peak days, the most **significant transactional volumes** are concentrated during the **Night, Dusk, and Morning** dayparts.

This indicates a clear pattern of customer engagement and provides valuable insights for optimizing staffing and operational strategies during these high-traffic periods.



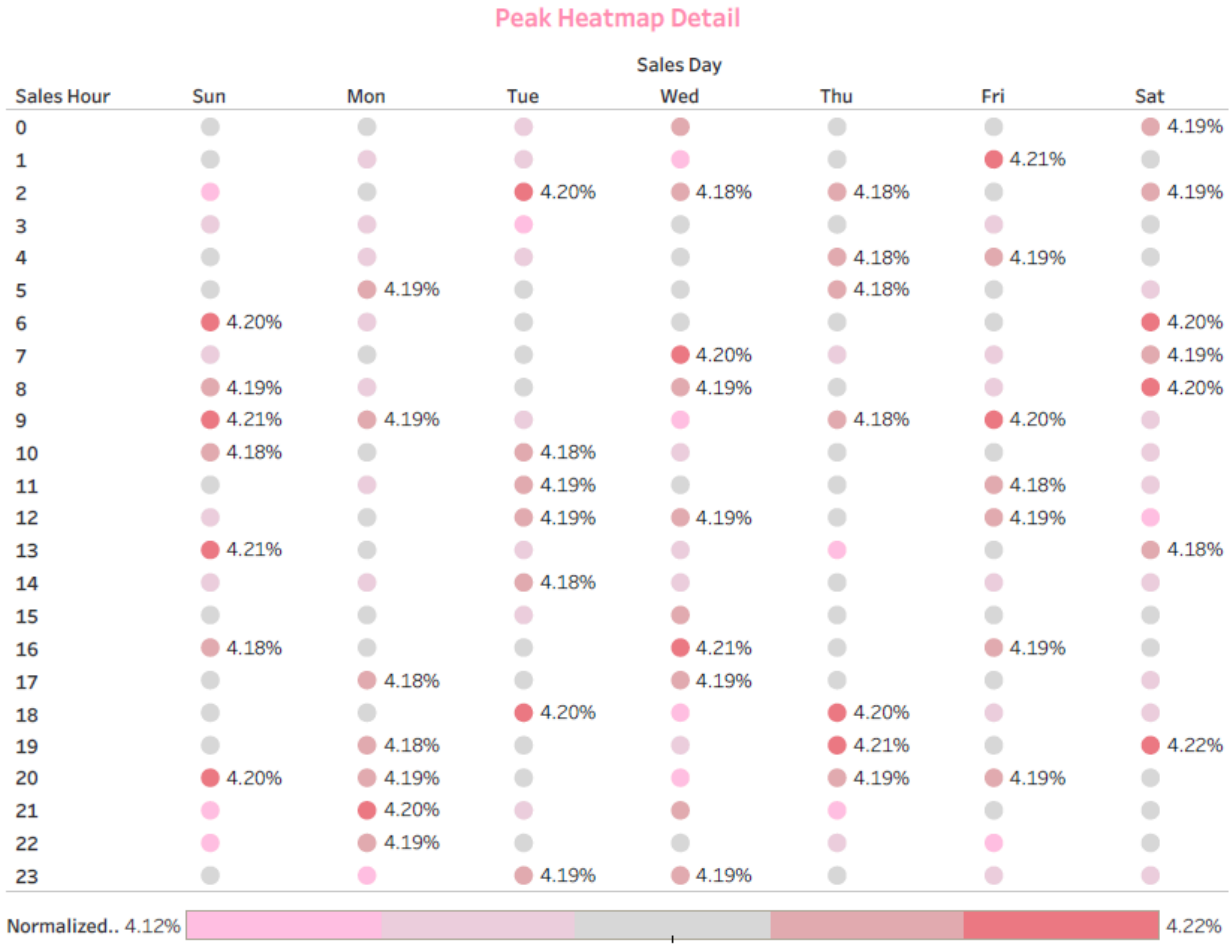
Picture 18 - Peak Heatmap (Day x Daypart) Order

Further insights into customer behavior are gained by examining the combination of day and hour, as depicted in Picture 18. While the **general peak periods across dayparts may exhibit similarities**, a more granular look at **specific hours reveals distinct patterns throughout the week**.

Here's a breakdown of the peak hours for each day:

- **Sunday**: The peak hours are observed at 6 AM and 8 PM.
- **Monday**: A singular peak occurs at 9 PM.
- **Tuesday**: Peak activity is seen at 2 AM and 6 PM.
- **Wednesday**: The peak hour is at 7 AM.
- **Thursday**: Peak hours are concentrated at 6 PM and 7 PM.
- **Friday**: A peak is noted at 1 AM.
- **Saturday**: The peak hour is at 6 AM.

This detailed hourly analysis provides valuable information for optimizing operational schedules, staffing, and marketing efforts to align with precise customer activity patterns.



Picture 19 - Peak Heatmap Detail (Day x Hour) Order

Based on the detailed **peak-hour heatmap analysis**, the **company should enhance operational preparedness to ensure a smooth shopping experience for customers**. This includes **optimizing staffing levels and resource allocation during identified peak hours** to better handle customer flow, minimizing wait times, and improving overall service quality.

Being proactive in operational adjustments aligned with customer activity patterns will contribute significantly to customer satisfaction and business efficiency.

Customer's Cohort

Cohort Data Group: Weekly | Cohort Retention Calculation: Percentage

Period Number From Cohort Start

Cohort Key	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1/1/2018	100.0%	98.7%	98.8%	98.8%	98.3%	48.1%	48.1%	48.2%	48.1%	48.1%	48.1%	48.1%	48.1%	48.1%	48.1%	48.1%	48.1%	48.1%	39.2%
1/8/2018	100.0%	97.6%	97.7%	97.6%	97.2%	97.2%	97.8%	97.5%	97.6%	97.5%	97.5%	98.0%	97.2%	97.5%	97.7%	97.7%	98.0%	79.0%	
1/15/2018	100.0%	98.5%	98.5%	100.0%	100.0%	97.1%	98.5%	94.1%	98.5%	97.1%	98.5%	98.5%	97.1%	95.6%	97.1%	98.5%	75.0%		
1/22/2018	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%			

Picture 20 - Customer's Cohort (Weekly)

Based on the actual data, four distinct weeks have been identified as the starting points for individual cohorts. For the cohort commencing on January 1, 2018, retention remained quite stable, hovering around 98% from the first to the fifth week. However, a notable decline in retention was observed from the sixth week onwards, with the primary reason for this drop yet to be identified.

An examination of cohorts originating in other weeks reveals generally stable and positive retention rates across various starting points and weekly periods. Furthermore, among all cohort starting points, the week of January 22, 2018, stands out as exceptionally strong, exhibiting a perfect 100% retention rate.

This analysis provides crucial insights into customer behavior patterns, highlighting areas for further investigation to understand retention dynamics better.

CLTV Period Data Group: Weekly

Period Number From Cohort Start

Cohort Key	1	2	3	4	5	6	7	8	9	10	11	12
1/1/2018	18,440,598	36,388,007	54,400,580	72,404,234	90,385,176	108,333,937	126,286,415	144,274,407	162,240,860	180,223,576	198,229,702	216,269,235
1/8/2018	453,494	898,996	1,350,565	1,804,663	2,244,974	2,692,201	3,138,381	3,582,611	4,033,073	4,490,000	4,944,011	5,395,768
1/15/2018	11,909	26,015	38,143	50,817	62,245	74,178	87,284	98,684	111,856	124,582	137,470	150,002
1/22/2018	338	386	615	769	937	1,053	1,203	1,364	1,677	1,775	1,844	2,326

Picture 21 - CLTV Period Weekly

Customer Lifetime Value (CLTV) serves as a **critical metric for businesses to quantify the long-term profitability of customers**. This metric offers a detailed perspective on how customer value evolves and accrues over time. As depicted in Picture 21, the **CLTV in this analysis is presented based on a weekly time period**. The corresponding mathematical formula will be:

$$CLTV_{Cohort,N} = \Sigma_{i=1}^{N} Revenue_{Cohort,i}$$

Formula 02 - CLTV Formula

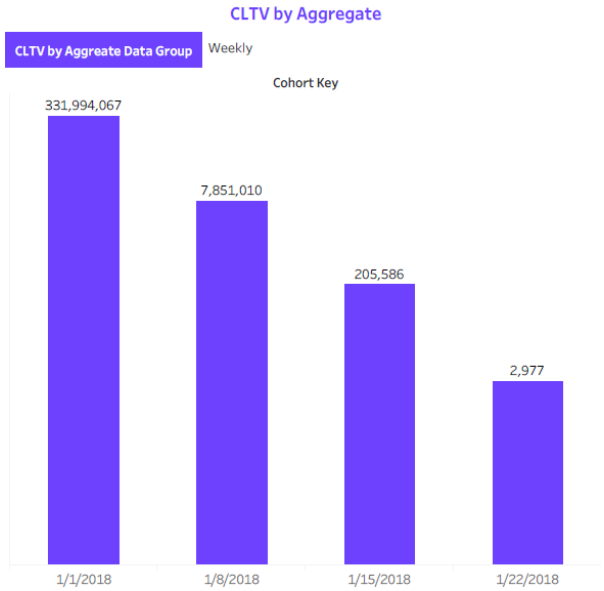
Where:

- $CLTV_{Cohort,N}$ is the cumulative CLTV for the specific Cohort up to Week N
- $Revenue_{Cohort,i}$ is the total revenue generated by that Cohort in Week i

Example interpretation for the cohort starting on January 1, 2018:

- Week 1 CLTV: $18,440,598
- Week 2 CLTV: $36,388,007 (This value represents the cumulative sum of Week 1 revenue plus Week 2 revenue for the cohort starting on January 1, 2018)
- The progression continues similarly up to the longest observed week period, where the cohort's CLTV reaches $216,269,235.

This example clarifies how CLTV accumulates over time, providing insight into the total value generated by a customer cohort over multiple weeks.



Picture 22 - CLTV by Aggregate for Period Weekly

Picture 22 provides a more granular view of Customer Lifetime Value (CLTV), with data already **aggregated for weekly periods**. Among all periods presented, it becomes clearer that the **period commencing on January 1, 2018, generated the highest CLTV when compared to other periods**.

This visualization allows for a more detailed understanding of the performance of different time periods in terms of customer value generation.

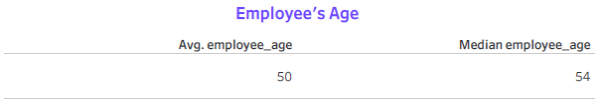
An analysis of the **cohort and Customer Lifetime Value (CLTV)** charts **reveals interesting contrasts**. Despite the **cohort** starting on **January 1, 2018**, exhibiting the **lowest retention** rate among the various cohort weeks, the customers within this specific **cohort generated the highest Customer Lifetime Value**.

Conversely, the cohort commencing on **January 22, 2018**, appears to have the **most favorable retention rate**. However, the customers associated with this cohort key regrettably **generated the lowest CLTV**.

This observation highlights that high retention does not always translate to the highest lifetime value, and vice-versa, suggesting a need for a nuanced understanding of customer segments and their contributions.

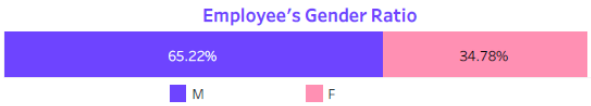
Employee's Insight

Based on the actual data, it is observed that **employees** have an **average age of 50 years**, with a **median age of 54 years**.



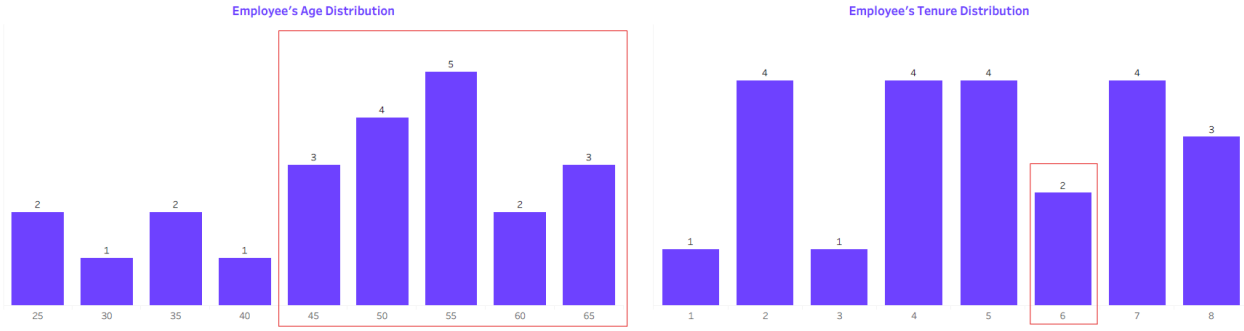
Picture 23 - Employee Ages

Regarding the gender ratio among employees, the **majority are male**, accounting for approximately **66%**, while **female** employees comprise about **34%**. Additionally, the **average tenure** for employees is approximately **4.87 years**, with a **median** tenure of **5 years**.



Picture 24 - Employee Gender Ratio

Picture 25 - Employee Working Tenure



Picture 26 - Employee Age and Tenure Distribution

Based on Picture 26, the distribution of employee ages exhibits a rightward skew, indicating that a significant portion of the workforce is in the **mid-to-late** career stage (**ages 45-54 years**) and approaching typical **retirement** ages (**over 55 years**), as categorized by the **Bureau of Labor Statistics of the USA**.

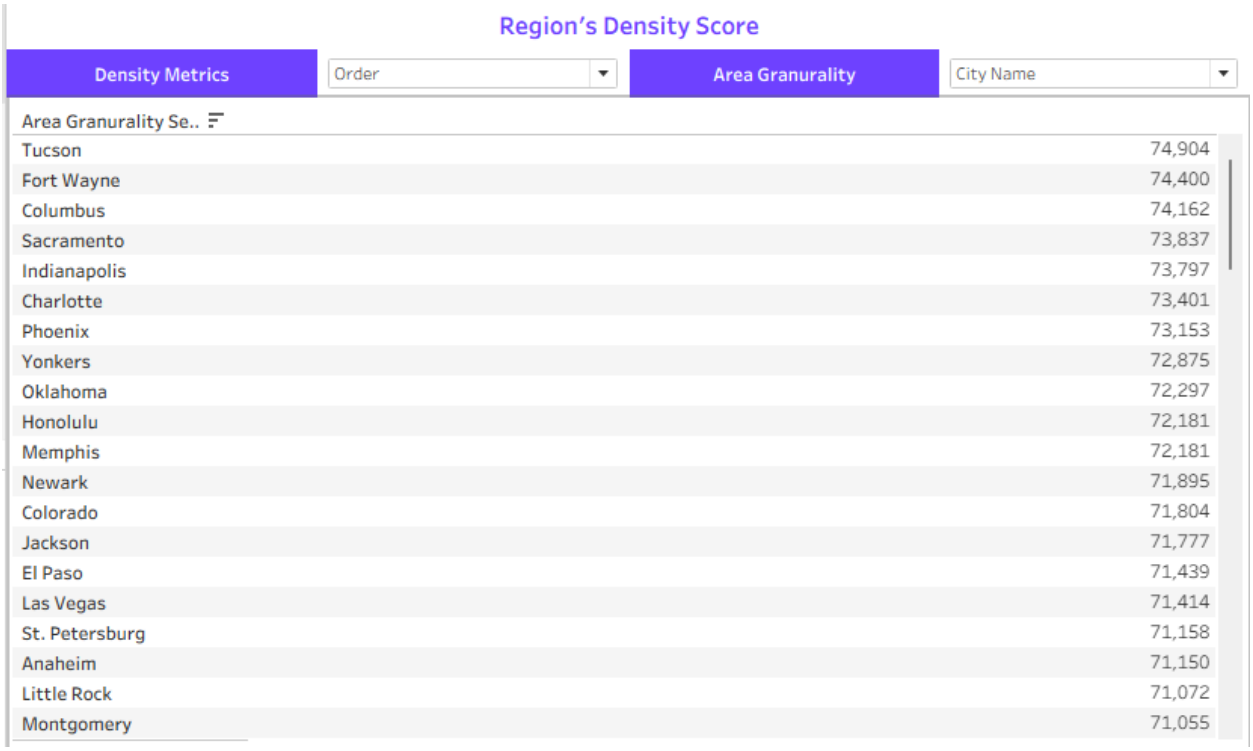
Regarding employee tenure, the distribution is relatively balanced, with approximately **56% of employees** having more than **5 years of tenure** and about **44% having less than 5 years**. However, it can be concluded that a substantial **majority of employees, approximately 70%, are elderly**.

Given these facts from the provided data, it is imperative for the company to address this topic proactively. Initiatives supporting hiring for replacement and regeneration are crucial to ensure a smooth business transition.

Region's Business Performance

This section delivers detailed information about several metrics at the city level for specific regions. The metrics and the corresponding rankings by city are as follows:

1. **Order** (Tucson -> Fort Wayne -> Columbus -> Sacramento -> Indianapolis)
2. **Revenue** (Tucson -> Columbus -> Indianapolis -> Sacramento -> Charlotte)
3. **Order Per Customer** (San Antonio -> Mobile -> Newark -> Fort Wayne -> St. Paul)



Picture 27 - Region's Density Score (Order)

Region's Density Score

Density Metrics | Revenue | Area Granurality | City Name

Area Granurality Se..	
Tucson	3,795,660
Fort Wayne	3,777,699
Columbus	3,755,175
Indianapolis	3,749,651
Sacramento	3,735,164
Charlotte	3,717,506
Phoenix	3,717,317
Yonkers	3,706,802
Oklahoma	3,681,719
Jackson	3,670,528
Colorado	3,669,080
Newark	3,667,452
Honolulu	3,664,738
Memphis	3,663,849
El Paso	3,631,867
St. Petersburg	3,626,732
Little Rock	3,624,349
Las Vegas	3,621,725
Mesa	3,616,701
Anaheim	3,615,944

Picture 28 - Region's Density Score (Revenue)

Pictures 27 and 28 illustrate the Region's Density Score based on Order and Revenue, **revealing a consistent pattern among the top-listed cities**. Beyond these metrics, an examination of different indicators provides a clearer understanding of profitability.

Picture 29, which presents **Revenue by Customer**, offers metrics that **represent the spending power of each customer within each geographical area**. Notably, the list of cities differs somewhat from those presented in Pictures 27 and 28, which focused on overall order and revenue metrics.

Possessing this type of information enables the company to strategically align marketing and sales efforts, making them more precise to effectively boost company growth.

Region's Density Score

Density Metrics | Revenue by Customer | Area Granurality | City Name

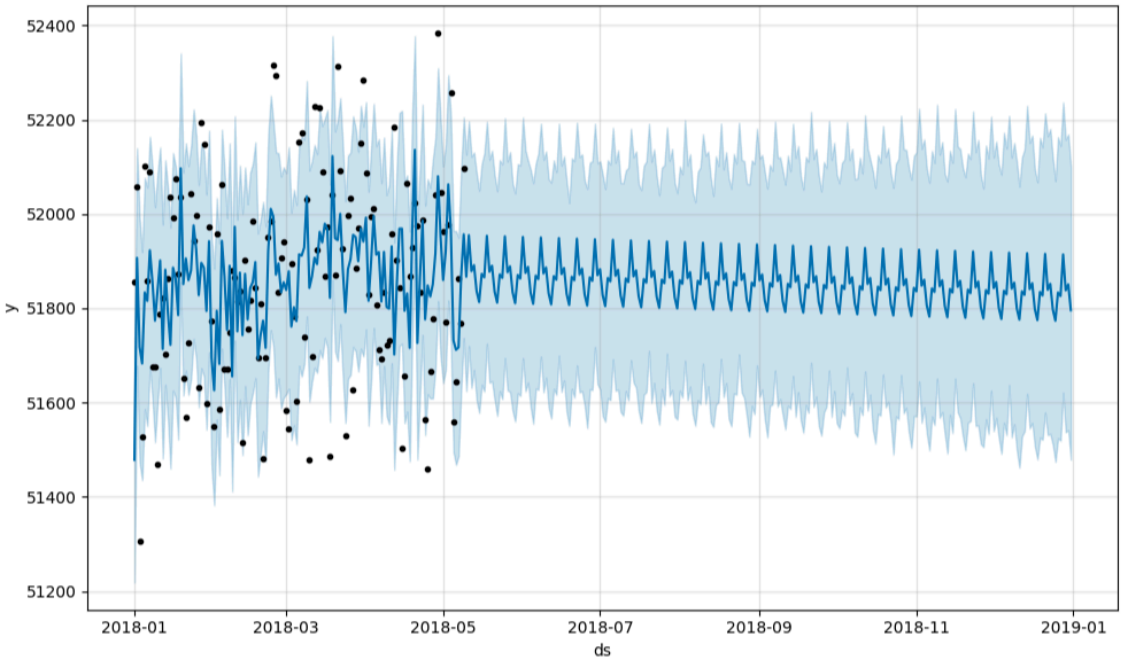
Area Granurality Se..	
San Antonio	3,477
Mobile	3,475
Newark	3,473
Fort Wayne	3,472
St. Paul	3,471
San Diego	3,468
Pittsburgh	3,466
Tampa	3,464
Jacksonville	3,463
Philadelphia	3,462
Wichita	3,462
Tacoma	3,462
Lubbock	3,461
Mesa	3,461
Spokane	3,458
Glendale	3,458
St. Louis	3,457
Baltimore	3,457
Albuquerque	3,455
Colorado	3,455

Picture 29 - Region's Density Score (Revenue by Customer)

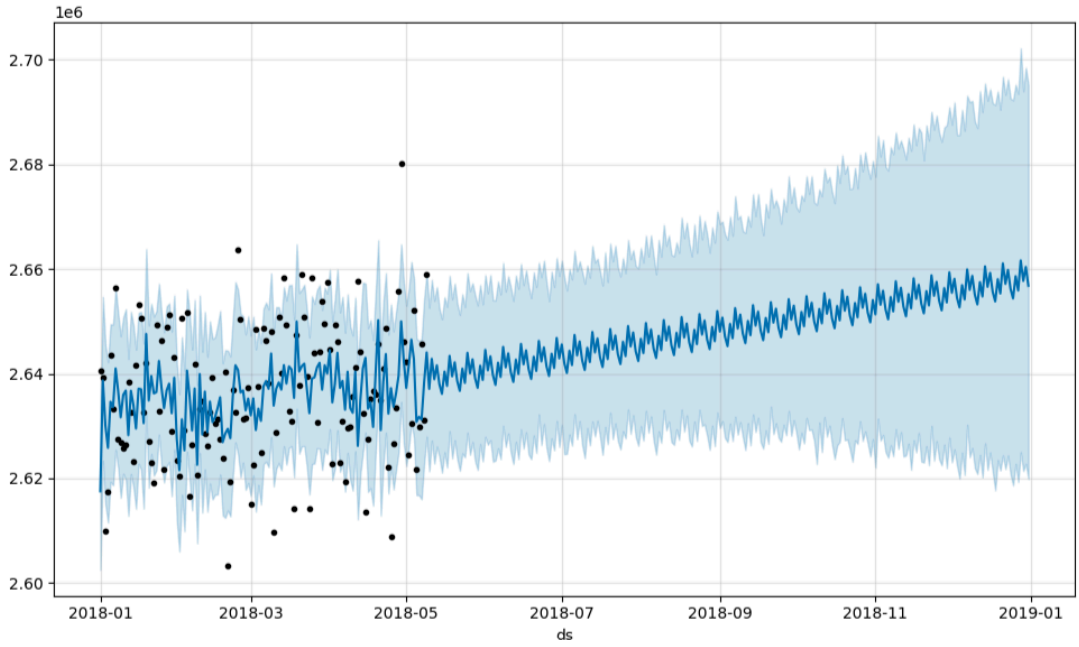
Business Forecasting

It is important to note that forecasting ideally requires data spanning one to two years. However, in this particular case, only less than 5 months of data are available. Furthermore, **forecasting is not about "certainty," but rather about "probability with a confidence interval"**.

Given the limited data, the most ideal granularity for this scenario is daily, even though weekly and monthly data are also available.



Picture 30 -Forecast Order Metrics to the End of the Year
MAE : 153.51 ; MAPE : 0.30%



Picture 31 -Forecast Revenue Metrics to the End of the Year
MAE : 95,599.57 ; MAPE : 0.30%

Pictures 30 and 31 provide visualizations of forecasting results utilizing Facebook Prophet. The details on how to interpret these visualizations are as follows:

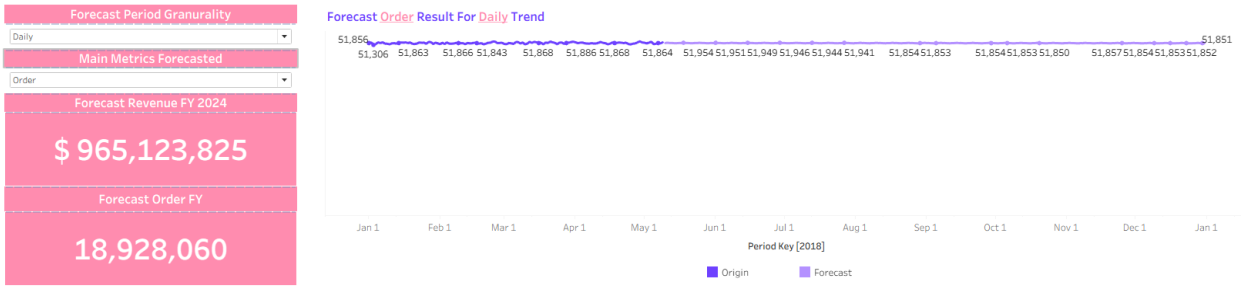
1. **X-axis**: This axis represents the time series, which in this specific case is daily.
2. **Y-axis**: This axis represents the metrics that are being forecast, namely order volume and revenue.
3. **Black-colored dots**: Each black dot symbolizes an actual data point. These actual data points serve as observations for the model's "learning" process.
4. **Dark-blue colored line**: This line represents the prediction or forecast value generated by the model.
5. **Soft-blue colored line**: This line indicates the confidence interval, which illustrates the likely range of values that could occur in the future.

Additionally, two key error metrics are provided:

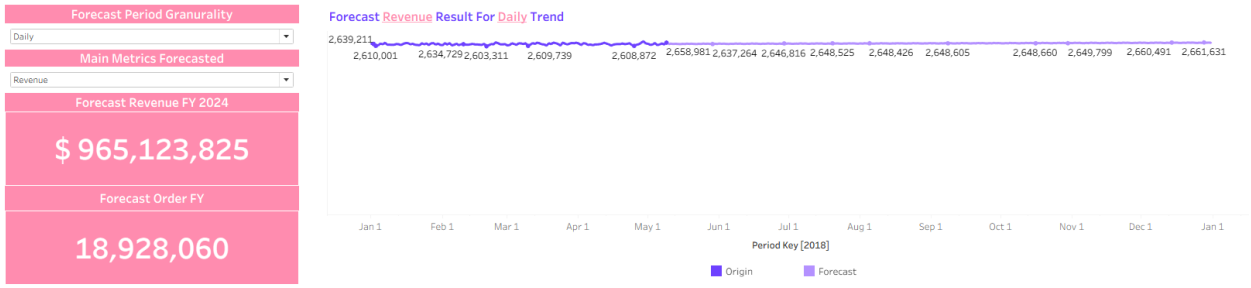
1. **MAE (Mean Absolute Error)**: This metric represents the average absolute difference between the actual values and the predicted values.
2. **MAPE (Mean Absolute Percentage Error)**: This metric represents the average percentage difference between the actual values and the predicted values.

Both forecasted metrics demonstrate a Mean Absolute Percentage Error (MAPE) of approximately 0.30%. This value is categorized as a very good result, indicating high accuracy in the predictions. **Ideally, a MAPE result should be less than 20%.** For reference, accuracy levels can generally be interpreted as follows:

- < 10%: **Highly Accurate**
- 10% - 20%: **Good Accuracy**
- 20% - 50%: **Reasonable Accuracy**
- > 50%: **Inaccurate**



Picture 32 - Forecast Order Result for Daily Trend



Picture 33 - Forecast Revenue Result for Daily Trend

Based on the **forecasting results** for both orders and revenue from the beginning to the end of 2018, the company is projected to generate approximately **~18.92 million orders**. This volume of orders translates into a **revenue of about $~965.12 million**.

This forecast provides a valuable outlook for annual performance, supporting strategic planning and resource allocation for the year.

Recommendations

1. **Standardise Data Architecture**
    - Consolidate the origin datasets into a single source of truth, a well-documented datamart.
    - Adopt consistent naming conventions (snake_case) and enforce ISO-3166 country codes, latitude/longitude precision, and unified date formats.
2. **Refine Product Taxonomy**
    - Deploy the newly defined 29-category structure across all upstream and downstream systems (e.g., POS, ERP, BI).
3. **Enhance Forecasting Capability**
    - Regularly revisit the forecasting model to evaluate and enhance its accuracy.
    - Integrate external drivers (e.g., promotion, holiday, weather) to improve predictive performance.
4. **Optimise Inventory & Assortment**
    - Use the forecast output to generate safety-stock recommendations by SKU-category and region.

- Identify low-velocity items for rationalisation or promotion bundling to free shelf space for high-margin categories.

5. **Leverage Insight - Driven Marketing**
    - Initiate the development of customer segments based on criteria such as purchase frequency, basket size, and category affinity..
    - Design targeted campaigns (e.g., email, loyalty app) that promote cross-category upsell opportunities.
    - Utilise data-driven marketing to optimise budget allocation and increase ROI.

Conclusion

The deep dive into GlobalMart's grocery sales uncovers three core insights that drive subsequent recommendations:

1. **Data Inconsistencies Area Limiting Insight**
    - The original taxonomy (e.g., "Snack" vs "Chips") and non-standard country/geo code fragmented reporting, leading to misaligned forecasts and inventory mismatches.
    - Cleaning effort – standardising ISO-3166 codes, correcting latitude/ longitude, and redefining categories into a coherent 29-level hierarchy – creates a reliable foundation for analytics.
2. **Forecast Accuracy Hinges on Enriched Context**
    - With millions of sales data spanning 452 products across 96 cities, the current models lack external drivers (promotion, holidays, weather); these things may contribute to systematic forecast bias observed in the analysis.
    - Incorporating these variables will tighten the error margin, enabling more precise safety-stock calculation and reducing stock-outs on high-margin items.
3. **Customer Segmentation Reveals Untapped Revenue Levers**
    - The 98K unique customers might expose distinct purchasing behaviours; high-frequency shoppers gravitate towards fresh and frozen categories, while low-frequency segments show potential for cross-category upsell through personalized campaigns.

These findings directly justify the recommendation above:

- **Unified datamart + standardize taxonomy** resolves the root data-quality issues, ensuring all downstream models consume consistent inputs.
- **Enhance forecasting with external drivers**, translate cleaned data into actionable inventory and assortment decisions.
- **Insight-driven marketing** leverages the rich customer dimension to boost ROI and improve loyalty.

In short, the analysis shows that **clean, well-structured data is the catalyst for smarter forecasting, efficient inventory, and profitable marketing**. By executing the recommendation, GlobalMart will transform its grocery business into **a data-first, agile operation** capable of sustaining growth in an increasingly competitive market.