Contents

[Overview 2](#_Toc197710731)

[Objectives 2](#_Toc197710732)

[Scope 2](#_Toc197710733)

[Dataset Overview 2](#_Toc197710734)

[Methodology 3](#_Toc197710735)

[Data Cleaning and Preprocessing 4](#_Toc197710736)

[Analysis and Results 6](#_Toc197710737)

[Inventory management analysis 6](#_Toc197710738)

[Customers Analysis 9](#_Toc197710739)

[Marketing Performance Analysis 9](#_Toc197710740)

[Delivery Performance Analysis 11](#_Toc197710741)

[Orders and Sales Analysis 14](#_Toc197710742)

[Predictive Analysis 19](#_Toc197710743)

Project Documentation (Blinkit sales analysis)

## Overview

This project aims to conduct an indepth analysis of Blinkit’s operational data to uncover valuable insights across multiple business areas. The analysis will cover customer behavior, delivery performance, marketing effectiveness, inventory management, and order trends, providing a holistic view of the company’s performance and opportunities for growth.

## Objectives

Descriptive Analysis: Summarize historical trends in customer activity, delivery times, product sales, and marketing performance. Prescriptive Analysis: Recommend strategies to optimize inventory, improve customer satisfaction, and enhance marketing ROI. Predictive Analysis: Potentially, forecast future order volumes and customer demand.

## Scope

* Customer Analysis: Understand customer demographics, purchasing patterns, and feedback.
* Delivery Performance: Evaluate delivery times and success rates
* Marketing Insights: Assess the impact of marketing campaigns on sales and customer retention.
* Inventory Management: Track stock levels and inventory efficiency
* Order and Sales Trends: Analyze order frequency, revenue patterns, and product performance.

## Dataset Overview

This dataset contains sales transaction data from Blinkit, an online grocery delivery platform. It provides valuable insights into customer purchasing behavior, product demand, revenue trends, and sales performance over time.

* Contains sales data from Blinkit, including product details, order quantities, revenue, and timestamps.
* Useful for demand forecasting, price optimization, trend analysis, and business insights.
* Helps in understanding customer behavior and seasonal variations in online grocery shopping.

A screenshot of a computer screen

AI-generated content may be incorrect.

## Methodology

* Data Tools: Python (pandas, matplotlib, seaborn), Excel, Power BI, Tableau.
* Techniques: Data cleaning, exploratory data analysis (EDA), feature engineering, predictive modeling, clustering, and recommendation system.
* Visualizations: Graphs, heatmaps, trend lines, and interactive dashboards.

## Data Cleaning and Preprocessing

**1. blinkit\_customer\_feedback**

* Removed missing values (NaN) and duplicate rows.
* Converted `rating` column to integer data type.
* Cleaned text in `feedback\_text` by:
  + Converting to lowercase.
  + Removing punctuation and special characters using regex.
  + Stripping extra whitespace.

**2. blinkit\_orders**

* Removed rows with missing values in critical columns: `order\_id`, `product\_name`, and `order\_date`.
* Removed duplicate entries.
* Converted `order\_date` to datetime format and extracted the hour field for timebased analysis.

**3. blinkit\_products**

* Removed duplicate entries.
* Stripped whitespace from all stringbased fields (e.g., `product\_name`, `category`, `brand`).
* Converted the `price` column to numeric values and dropped rows with invalid/missing price data.

**4. blinkit\_delivery\_performance**

* Handling Missing Values
  + Numerical columns: `delivery\_time\_minutes` and `distance\_km`: Filled missing values with the median to avoid skewing distributions.
  + Categorical columns:
    - `delivery\_status`: Filled missing entries with "Unknown" to indicate missing status.
    - `reasons\_if\_delayed`: Filled missing entries with "No Reason" for clarity.
* Removed duplicate rows to ensure uniqueness of records.
* Handling Outliers: applied the Interquartile Range (IQR) method to `delivery\_time\_minutes`:
* Data Type Conversion
* String conversion: `order\_id` and `delivery\_partner\_id` were converted to strings to treat them as identifiers.
* Datetime conversion: `promised\_time` and `actual\_time` converted to datetime format.
* Created new columns:
  + `promised\_date` and `promised\_time` (split from `promised\_time`).
  + `actual\_date` and `actual\_time` (split from `actual\_time`).

**5. blinkit\_order\_items, blinkit\_marketing\_performance**

* Validated data integrity (no missing values/duplicates).
* Visualized outliers for further analysis (no removal performed).
* Preserved original data types.
* Exported cleaned dataset.

**6. blinkit\_inventory**

The inventory data was merged from 2 different CSV files and underwent basic cleaning to ensure uniqueness of entries, consistent schema, data types and date format.

**7. blinkit\_customer**

* A check was performed to detect duplicate customer IDs.
* Missing values were assessed across all columns.
* The phone numbers and pincodes were converted to string format to preserve formatting and leading zeros.

## Analysis and Results

### Inventory management analysis

The main objective of this analysis was to identify products with the highest levels of damage, observe damage trends over time, and determine the frequency of damage for each product.

The following analyses were conducted to answer key questions about the damage trends in the inventory:

* **Total Damaged Stock per Product**:
  + We aggregated the total damage for each product based on the product\_id.
  + The products with the highest damaged stock were identified.
* **Monthly Damage Trends**:
  + By extracting the month from the date column, we grouped the data by month and calculated the total damaged stock per month.
  + This helped us to observe any seasonal patterns in damaged stock over time.
* **Frequency of Damage by Product**:
  + We analyzed how often products received damages, counting how many times each product experienced damage.

#### Findings

* **Top 10 Products with Highest Damaged Stock**:
  + These are the products that experienced the highest overall damage in stock.
* **Top 10 Months with Highest Damaged Stock**:
  + These months had the highest total damaged stock, indicating when the most significant inventory issues occurred.
* **Top 10 Products with Most Frequent Damage**:
  + These are the products that experienced damage the most frequently, which could help pinpoint products prone to frequent damage.

#### Visualizations

**Bar Chart of Damaged Stock per Product**:

A bar chart was plotted showing the total damaged stock for each product. This helps visualize which products are the most problematic.

A graph of damage products

AI-generated content may be incorrect.

**Line Chart of Monthly Damaged Stock Trends**:

A line chart was used to show how damaged stock trends changed over time, helping to identify any patterns in damages across months.

A graph with orange lines and numbers

AI-generated content may be incorrect.

**Heatmap of Correlation Between Stock Received and Damaged Stock:**  
A heatmap was created to visualize the correlation between the amount of stock received and the amount of damaged stock. This helps determine whether an increase in stock quantity is associated with more damages. The correlation coefficient was also calculated to support this analysis.

A red and blue squares

AI-generated content may be incorrect.

#### Recommendations

* **Focus on High Damage Products**: Consider reducing stock of products that consistently show high damage rates or implement better quality control processes.
* **Improve Inventory Handling**: Analyze the root causes of damage, especially during highdamage months, and implement changes to reduce it.
* **Track Damage Frequency**: Regularly monitor products that frequently experience damage to identify and address any recurring issues in the inventory chain.

### Customers Analysis

The purpose of this analysis was to explore customer purchasing patterns and segment performance using transactional and demographic data. Key questions addressed include identifying highvalue segments, top regions, and customer loyalty indicators.

The dataset contains detailed information on 2,500 customers, including customer IDs, segment classification, contact details, geographic area, registration date, total orders, and average order value. Initial exploration ensured data types were appropriate for analysis, especially converting dates for timebased insights.

### Marketing Performance Analysis

#### Key Business Questions & Insights

* **Which customer segments contribute most to high average order values in the topperforming areas?**
  + By focusing on the top 10 areas with the highest average order values, we found significant variation across customer segments. Some segments consistently outperformed others, indicating concentrated highvalue customers in specific regions.
* **How are customer segments distributed within the top 20 regions by total customer count?**
  + A stacked bar chart revealed that certain customer segments dominate specific geographic areas. Understanding this distribution helps tailor marketing strategies and regional service planning.
* **Which five regions generate the highest average order value?**
  + The analysis identified five key areas with significantly higher average order values, suggesting these are premium markets with customers who tend to spend more per transaction.
* **Who are the top five customers based on the total number of purchases?**
  + Top individual customers were highlighted based on the total number of orders placed. These loyal customers can be targeted for premium membership programs, loyalty rewards, or personalized offers.
* **Which customer segment exhibits the highest repeat purchase behavior?**
  + By analyzing total orders and computing the average number of orders per customer in each segment, we determined which segments show strong loyalty and repeat purchasing. These insights support customer retention strategies and segmentation refinement.

#### Key Business Questions & Findings

1. **What are the best marketing channels in terms of ROAS and conversions?**
   * **Topperforming channels:** *Email* and *App* had the highest **average ROAS** and conversion rates.
   * **Least effective:** *SMS* campaigns showed higher spend but lower conversions and ROAS.
2. **Which target group generates the highest return or conversion rate?**
   * *Returning Customers* showed the **highest average ROAS**, followed by *New Users*.
   * *Loyal Customers* had high conversions but were costlier to acquire.
3. **Are there time trends in campaign performance?**
   * Campaign performance varied across months.
   * A **peak in ROAS** was observed during festive periods, indicating seasonality impact.
4. **Top 5 campaigns by revenue generated:**
   * The campaigns with IDs: C504, C311, C129, C289, and C490 contributed the highest revenue.
5. **How are campaigns distributed across channels and target groups?**
   * Most campaigns targeted *New Users* and used the *App* channel.
   * A heatmap confirmed dense activity in specific channelaudience combinations.
6. **What’s the average Customer Acquisition Cost (CAC)?**
   * The **average CAC** was calculated as:

CAC=Total SpendTotal Conversions\text{CAC} = \frac{\text{Total Spend}}{\text{Total Conversions}}CAC=Total ConversionsTotal Spend​

* + - The average CAC across campaigns was approximately **₹126.78**.

#### Visualizations

* **Bar Charts** comparing ROAS by channel and audience.
* **Time Series Plot** of ROAS over time.
* **Heatmap** showing campaign density by channel and audience.
* **Scatter Plots** for impressions vs conversions and clicks vs revenue.

### Delivery Performance Analysis

#### Objective

This analysis aims to enhance delivery efficiency by understanding the key factors affecting delivery time and identifying recurring causes of delays. The insights gathered can be used to develop strategies for improving overall performance.

#### Data Overview

The dataset contains information on order status, delivery time, distance traveled, and reasons for delays. Other potential influencing factors include:

* **Traffic congestion**
* **Order timing (peak vs. offpeak hours)**
* **Availability of delivery personnel**

#### Performance Analysis

* **Total deliveries:** total\_deliveries
* **Ontime deliveries:** on\_time\_deliveries
* **Success rate:** success\_rate%
* **Average delivery time:** average\_delivery\_time in minutes

A pie chart with numbers and a percentage

AI-generated content may be incorrect.

#### Delay Pattern Analysis

To assess delay patterns, we examined whether:

* **Delays occur more frequently during peak hours?**
* **Certain weekdays show higher delay rates?**

**A chart showing the main reasons for delay

AI-generated content may be incorrect.**

Visualizations

Various charts were used to illustrate key insights, including:

* **Pie charts** showing the distribution of delivery statuses
* **Bar charts** highlighting the most common reasons for delays
* **LOWESS regression** indicating trends between distance and delivery time\

A graph with blue and green lines

AI-generated content may be incorrect.

#### Relationship Between Distance and Delivery Time

* **Pearson correlation coefficient:**
* The correlation between distance and delivery time is **very weak**, suggesting that distance alone does not significantly impact delivery duration.

A diagram of a number of blue dots

AI-generated content may be incorrect.

#### Recommendations for Improvement

Based on our findings, the following actions may help reduce delays and enhance success rates:

* Optimize scheduling during peak hours.
* Avoid highly congested routes.
* Train drivers to improve response times.
* Utilize predictive models based on historical data.

### Orders and Sales Analysis

#### Project Overview

This report presents a comprehensive data analysis for Blinkit, an online delivery platform. The aim is to explore sales and order trends, understand product and customer behavior, and provide strategic recommendations to enhance performance and customer satisfaction.

#### Data Overview

* Rows: 5000
* Key Columns: order\_id, product\_id, quantity, unit\_price, order\_date, category, brand, revenue, margin\_percentage
* Revenue Column was calculated as: revenue = quantity unit\_price

#### Analysis Questions

1. Which products and categories are the best sellers?
2. How do payment methods impact total sales?
3. What is the overall sales performance over time?
4. Which days of the week have the highest sales?

**Which products and categories are the best sellers?**

A bar graph with different colored bars

AI-generated content may be incorrect.

**How do payment methods impact total sales?**

A graph of sales by payment method

AI-generated content may be incorrect.

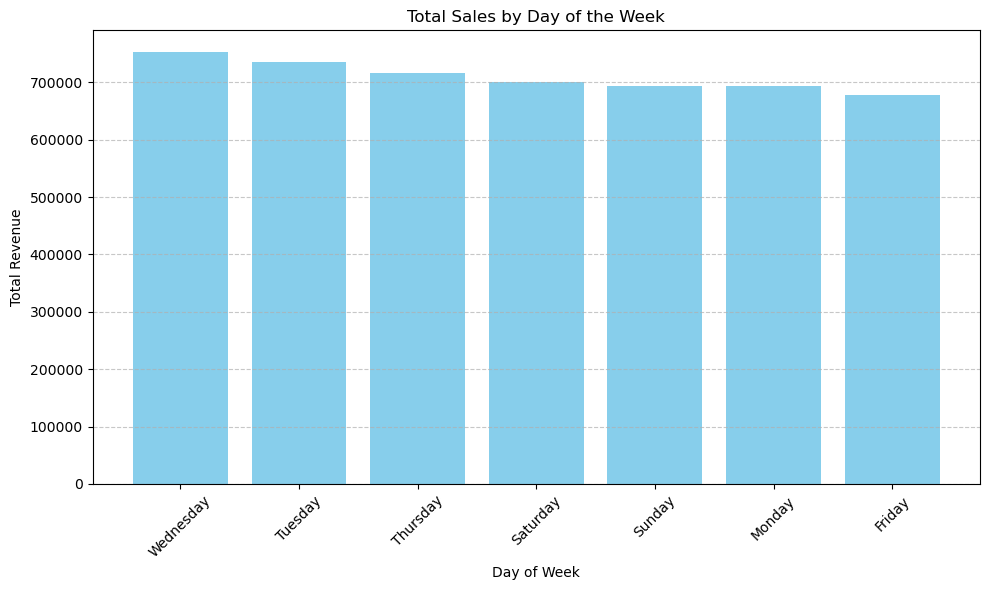
**What is the overall sales performance over time? :**

A graph with a line going up

AI-generated content may be incorrect.

1. **Sharp Increase at the Start**: There is a noticeable jump in total revenue from March 2023 to April and May 2023. This could be due to a successful marketing campaign or the launch of a new product.
2. **Peak in August 2023**: Sales reached their highest point in August 2023, followed by a gradual decline.
3. **Significant Drop in November 2024**: There's a very sharp decline in revenue in November 2024. This could be due to a system issue, temporary shutdown, or possibly incomplete data for that month.
4. **Relative Stability After August 2023**: Postpeak, revenue shows relative stability with some fluctuations through to October 2024.

**Which days of the week have the highest sales?:**



1. **Wednesday leads** in total revenue among all days, suggesting it's the strongest sales day.
2. **Tuesday and Thursday** follow closely behind, also showing strong performance.
3. **Friday and Sunday** appear to have the lowest total revenue, although the difference across all days is relatively minor — sales are fairly balanced throughout the week.
4. **Saturday and Monday** fall in the middle range

### Predictive Analysis

In this analysis, our main goal is to predict the total revenue (order\_total) generated from customer orders on the Blinkit platform.

We combined multiple datasets (orders, inventory, customer feedback, etc.), performed feature engineering, and trained a machine learning model (XGBoost) to forecast revenue trends.

By creating key features like average revenue per order, price effect, and growth rate, we built a model capable of predicting revenue for the next 30 days, helping the business make better planning and strategy decisions.

#### EDA

A blue line graph with white text

AI-generated content may be incorrect.

Temporal Trend of Total Revenue

* There are regular fluctuations throughout the timeline, which is typical for retail or ecommerce sales.
* Certain spikes in revenue stand out and are likely linked to seasonal events, promotions, or highdemand periods.
* Overall, despite some noise, the revenue trend appears relatively stable over time without significant longterm increases or declines.

A graph showing a number of colored squares

AI-generated content may be incorrect.

**Observations**

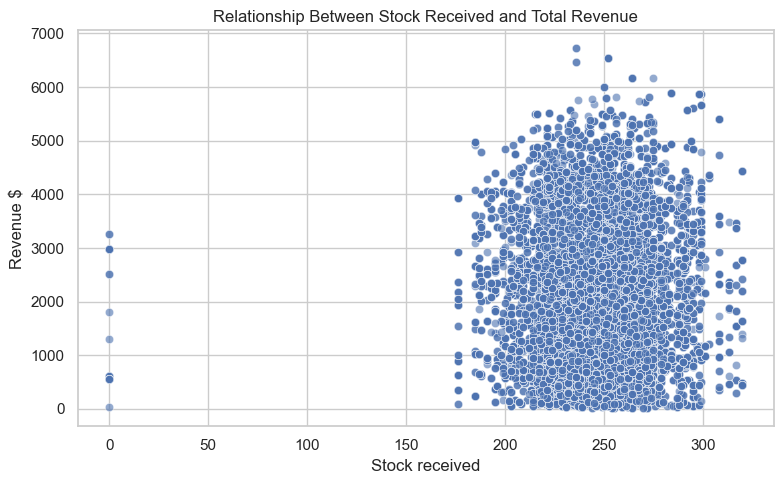
* Consistent distributions across all days:
  + The median remains almost identical throughout the week.
  + This indicates that sales volume is relatively stable on most days.
* Slight increase on certain days:
  + The boxplots for days 3, 4, and 6 (Thursday, Friday, and Sunday) show a slightly higher median and spread.
  + This suggests there may be more activity towards the end of the week.
* Presence of outliers:
  + Some days show highvalue outliers.
  + These outliers indicate occasional spikes in revenue, possibly due to large orders, promotions, or special days.

A graph showing different colored squares

AI-generated content may be incorrect.

**Observations:**

* Similar medians across months:
  + The median values (middle line in the boxes) are fairly consistent, indicating that monthly income remains balanced throughout the year.
* Slightly higher spreads in months 5, 7, and 12:
  + These months (May, July, and December) show a wider range of values and more highvalue outliers, suggesting potential seasonal boosts in sales.
* Outliers in mid and end months:
  + The spikes in May and July may indicate midyear promotions or special events.
  + The spikes in December are likely a result of the holiday season, which is common for ecommerce sales performance.



**Observations:**

* Stock levels mostly range between 200 and 300 units:
  + Most data points are concentrated within this range, indicating it is the typical restocking level for the store.
* No clear linear correlation:
  + Revenue varies significantly even when stock received remains constant.
  + This suggests that factors beyond stock levels — such as demand, pricing strategies, and promotions — strongly influence revenue.
* Stock = 0 with revenue > 0:
  + A few instances show sales occurring despite no new stock received.
  + This likely points to sales from existing inventory or timing differences in stock recording.

#### Feature Engineering

To improve model accuracy by adding more useful information, we will implement several feature engineering techniques on the data:

* **Average Revenue per Order:** Measures the average revenue generated for each order placed.
* **Revenue per Unit:** Shows how much revenue is earned for each individual unit sold.
* **Revenue Rolling Mean (7 days):** Smooths shortterm fluctuations to reveal longerterm revenue trends.
* **Revenue Rolling Std (7 days):** Captures the variability in revenue over a 7day period, highlighting periods of instability.
* **Revenue Growth Rate (% change from 7 days ago):** Compares the current day's revenue to the revenue from the same day in the previous week, showing growth or decline.
* **StocktoRevenue Ratio:** Indicates how much stock is needed to generate a certain amount of revenue, useful for evaluating operational efficiency.
* **Price Influence (Expected Revenue vs. Actual Revenue):**
* `**expected\_revenue`:** Calculated as unit price multiplied by quantity sold.
* `**price\_effect\_gap`:** The difference between the actual and expected revenue, which helps identify pricing strategies, promotions, or potential anomalies.

Finally, we will remove rows containing NaNs caused by the rolling calculations.

A screen shot of a graph

AI-generated content may be incorrect.

This heatmap reveals that several of the newly engineered features have a clear correlation with the target variable, suggesting they could play a valuable role in enhancing the model’s predictive accuracy.

## 

New Engineered features VS Target Variable

* avg\_revenue\_per\_order vs Total Sum
  + Displays a strong positive correlation — higher order averages tend to align with increased revenue.
* revenue\_per\_unit vs Total Sum
  + Shows a clear upward trend, suggesting it’s a reliable indicator of revenue performance.
* rolling\_revenue\_avg\_7 vs Total Sum
  + Presents a moderate positive correlation, reflecting gradual trends in revenue over time.
* revenue\_growth\_rate vs Total Sum
  + Mostly fluctuates around zero, offering little consistent pattern — less useful for prediction.
* stock\_to\_revenue\_ratio vs Total Sum
  + Shows an inverse relationship — higher ratios may signal overstocking or operational inefficiencies.
* expected\_revenue vs Total Sum
  + Lacks a visible pattern — being a theoretical measure, it doesn't align closely with actual results.
* price\_effect\_gap vs Total Sum
  + Reveals a strong positive correlation — variations in pricing effectiveness have a noticeable impact on revenue.

#### Results

A screenshot of a computer program

AI-generated content may be incorrect.

A graph showing the difference between the average and the average

AI-generated content may be incorrect.

This graph shows that the predicted values closely follow the actual values, reflecting the model’s strong accuracy and reliability.

A graph with blue lines and numbers

AI-generated content may be incorrect.

Prediction Summary:

The line chart illustrates the model's forecast for total revenue over the next 30 days. While there are some daytoday fluctuations and occasional peaks suggesting potential sales surges, the overall trend remains steady, indicating a consistent and stable revenue pattern for the coming month.