Chocolate Sales Analysis

Charlotte Soeurng

The idea of this dataset is to analyze chocolate sales across different categories, seasons, and customer demographics to uncover trends and predict future sales based on historical data.

The analytics I am planning to use are:

- Descriptive: I will analyze past chocolate sales and identify their contributing sales factors, as well as studying how the sales are currently to identify if anything has changed. This will uncover problems in the chocolate industry and provide opportunities for change.
- Predictive: I will use current and historical data to predict what will happen involving chocolate sales in the future, and what is causing it to happen. This will involve data mining and forecasting and will provide accurate predictions of future occurrences in chocolate sales.
- Prescriptive: By predicting future sales and problems, I will make recommendations on strategies to minimize those problems and improve sales.

The dataset I am using will be "Chocolate Sales Data" by Atharva Soundankar on Kaggle. It contains product details, sales, and customer information. The columns are:

- Date: Date of sales
- Product Name: Brand or type of chocolate
- Category: Kind of chocolate (dark, milk, white…)
- Units sold: Number of units sold
- Boxes shipped: Total boxes shipped
- Region: Location of sales

The analysis will contain a start schema of a fact table and dimension tables described below.

Fact Table: Date, Product, Amount

Dimension Tables

- Product table
  - Product Name
  - Category
  - Salesperson
- Date table
  - Date ID
  - Date
  - Month
  - Year
  - Quarter
- Salesperson table
  - Products
  - Amount
  - Boxes shipped
  - Country
  - Date

Measures

- Total Sales Revenue
- Average Revenue per Sale
- Total Units Sold
- Sales per Product Category
- Regional Sales Distribution

Snow Flaking will be added by normalizing

- Date table: Create a month dimension and a year dimension
- Product table: Create a product name dimension and a category dimension

Data Consolidation Step

- Select needed data from the dataset
- Connect table relationships

Data Cleaning Step

- Checking date formats
- Finding and removing duplicated entries
- Impute values

Data Transformation Step

- Converting categorical variables into numerical representations for easier analysis
- Normalize data and making a star schema
- Creating attributes

Data Reduction Step

- Removing unneeded columns
- Checking outliers that highly affect results
- Reducing volumes and balancing data

High-Level System Architecture

Identify Data Sources, Technical Staff, Business Users, & Managers/Executives

- Data Sources: Kaggle Chocolate Sales dataset

- Technical Staff: Data engineers, BI analysts, data scientists
- Business Users: Marketing teams, sales teams
- Managers/Executives: Business managers, financial analysts

Describe Data Sources, Data Warehouse, & User Interface

- Data Sources: Raw chocolate sales data stored in Power BI or SQL database.
- Data Warehouse: Central repository where structured data is stored after ETL processing.
- User Interface: Power BI dashboard for visual analytics and reporting.

Label Data Warehouse Environment, Business Analytics Environment, and Performance & Strategy Environment

- Data Warehouse Environment: Stores historical sales data, customer data, and product details.
- Business Analytics Environment: Uses Power BI for visualization, trend analysis, and performance monitoring.
- Performance & Strategy Environment: Business managers use insights from analytics to optimize sales strategies and decision-making.

Design Considerations

1. Data Visualization Type Selection

In my project, I carefully selected chart types that best suited the data and the purpose of each analysis. For example, I used a pie chart to show sales distribution by country because it clearly illustrates how total sales are divided among different regions, making it easy to see which markets contribute the most. To analyze sales trends over time, such as by product category or by month, I chose line charts since they effectively reveal patterns and changes over periods. For comparing sales totals across months or individual products, I used bar charts to provide clear and straightforward comparisons between different items or time frames. I also incorporated a scatter plot with a regression line to explore the relationship between the number of boxes shipped and total sales by category, which helped me identify correlations and understand how shipment volume affects revenue. Lastly, to forecast future sales, I implemented a forecast line chart with prediction intervals which not only shows the expected trend but also the uncertainty around those predictions, supporting the predictive analytics aspect of my project.

2. User Needs and Business Goals

My analyses are designed to deliver actionable insights tailored to the needs of marketing teams, sales teams, and business managers. For instance, by visualizing sales by country and product, I helped prioritize which markets and product lines should receive more focus, supporting strategic decision-making. The monthly and daily sales trends provide crucial information for understanding seasonality, which assists in inventory management and promotion planning. The forecasting analysis enables business planners to anticipate future demand and allocate resources effectively. Furthermore, the scatter plot examining shipment volume versus sales supports logistics and pricing strategies, helping identify areas where operational improvements can be made.

3. Data Granularity and Aggregation

To address different business questions, I worked with the data at various levels of granularity. I used daily data to capture fine-grained sales trends and detect short-term patterns. For product-level insights, I aggregated data by category, allowing me to identify which types of chocolate are driving revenue. Finally, I summarized sales by country to provide a high-level overview of geographic performance, supporting market expansion decisions. This approach ensures that the analyses are flexible and relevant for multiple stakeholders.

4. Assumptions Transparency

Throughout my project, I made sure to clearly state the assumptions behind each analysis to maintain transparency and credibility. For example, I assumed that the country information in the dataset accurately represents where sales occurred and that sales figures are comparable across regions without currency conversion issues. I also assumed that product categories were correctly classified and mutually exclusive. By documenting these assumptions, I helped ensure that users of the analysis understand the context and limitations, which is important for making informed decisions.

5. Star Schema and Snowflaking

For the data model, I designed a star schema consisting of a fact table and several dimension tables, such as Product, Date, and Sales dimensions. This structure allows efficient querying and supports flexible analysis. To further optimize the model, I applied snowflaking by normalizing certain dimensions — for example, breaking the Date dimension into separate Year and Month tables and splitting Product into Product Name and Category tables. This normalization improves data integrity and reduces redundancy, which enhances overall system performance and maintainability.

Analysis and Design Considerations

1. Sales by Country (Pie Chart)

Idea Behind This Analysis:

This pie chart visualizes the distribution of total sales amount across different countries where the products are sold. It highlights market contributions and relative sales share by geographic region.

Why It’s Needed:

Understanding which countries contribute most to sales helps direct marketing efforts, resource allocation, and strategic market expansions.

Assumptions:

- Country data is accurate and corresponds to actual sales transactions.
- Sales amounts are comparable across countries without currency conversion issues.

How to Use It:

Business managers can use this chart to prioritize markets for promotions or inventory adjustments. It also supports strategic planning for international sales.

Implications:

Countries with larger sales shares, such as Australia and the USA, might warrant more focused campaigns, while smaller markets may require growth strategies.

2. **Sales by Category (Line Chart)**

Idea Behind This Analysis:

This line chart shows the sum of sales amounts grouped by product categories, such as Chocolate types (e.g., Dark, White, Fruit).

Why It’s Needed:

Identifying which categories drive the highest revenue allows product managers to tailor product offerings and promotions accordingly.

Assumptions:

- Categories are properly classified and mutually exclusive.
- Sales data is aggregated accurately.

How to Use It:

Category managers can use this to identify star products and underperforming categories, driving inventory and marketing decisions.

Implications:

Categories with declining sales may require product reformulation or promotional support, whereas high-performing categories suggest stable demand.

3. Sales by Month (Bar Chart)

Idea Behind This Analysis:

This horizontal bar chart illustrates monthly sales trends, displaying total revenue by month.

Why It's Needed:

Month-to-month sales trends help identify seasonal patterns and cyclical effects that impact revenue.

Assumptions:

- Sales data is consistently captured over months without missing periods.
- Months are accurately labeled and aligned to the fiscal calendar.

How to Use It:

This analysis guides inventory stocking, budgeting, and promotional calendar planning by revealing high and low sales months.

Implications:

Months with dips may require special promotions, while peak months might stress supply chains, requiring pre-planning.

4. Sales by Product (Bar Chart)

Idea Behind This Analysis:

This chart breaks down sales of individual products, showing which items generate the most revenue.

Why It's Needed:

Pinpointing best-selling products assists in optimizing product portfolios and discontinuing underperforming items.

Assumptions:

- Product names are uniquely identified.
- Sales data is current and reflective of market behavior.

How to Use It:

Product managers can identify star products to prioritize marketing and inventory and identify products that need review.

Implications:

Resources should be allocated towards best sellers; underperforming products may be candidates for discounts or removal.

5. Boxes Shipped vs Sales by Category (Scatter Plot with Regression Line)

Idea Behind This Analysis:

This scatter plot compares the number of boxes shipped against the total sales amount, segmented by product category, with a regression line to illustrate correlation.

Why It's Needed:

Understanding the relationship between units shipped and revenue is critical for supply chain and financial forecasting.

Assumptions:

- Shipping data and sales amounts are accurate and properly linked by category.
- Linear relationships are a reasonable assumption.

How to Use It:

Logistics and finance teams can use this to forecast revenue based on shipment volumes and to detect anomalies.

Implications:

A strong positive correlation suggests that shipping volume is a reliable sales predictor. Deviations could indicate pricing or product issues.

6. Forecast of Total Sales by Month Number (Line Chart with Prediction Interval)

Idea Behind This Analysis:

This analysis presents a time series forecast of monthly total sales, including confidence intervals to indicate uncertainty.

Why It’s Needed:

Forecasting future sales enables strategic planning for production, budgeting, and market initiatives.

Assumptions:

- Historical sales data is representative of future trends.
- Forecast model assumptions such as seasonality and trend are valid.

How to Use It:

Executives and planners use this forecast to make informed decisions on investments, staffing, and supply chain.

Implications:

Forecasts with wide confidence intervals indicate higher uncertainty, suggesting a need for flexible planning.

7. Forecast of Total Sales Amount and Boxes Shipped by Day (Column Chart)

Idea Behind This Analysis:

This analysis provides a daily time series forecast of the total sales amount and the total number of boxes shipped, including confidence intervals to illustrate the range of uncertainty in the predictions.

Why It’s Needed:

Daily forecasting of sales revenue and shipped quantities helps optimize day-to-day operational decisions, such as inventory management, workforce scheduling, and logistics planning.

Assumptions:

Historical daily sales and shipment data accurately reflect future patterns.

The forecasting model properly accounts for daily seasonality, trends, and any anomalies in the data.

How to Use It:

Operations managers, supply chain coordinators, and sales teams use this forecast to align daily resources and inventory levels with expected demand, minimizing stockouts or overstock situations.

Implications:

Wide prediction intervals suggest volatility or uncertainty in daily sales or shipments, indicating the need for contingency plans such as safety stock or flexible staffing to handle fluctuations.

8. Sales by Day and Country (Bar Chart)

Idea Behind This Analysis:

This analysis visualizes the total sales amount aggregated by each day and segmented by country, using a bar chart to compare sales performance across different regions over time.

Why It’s Needed:

Understanding daily sales distribution across countries allows businesses to identify regional demand patterns, evaluate market performance, and tailor sales strategies or promotions by location.

Assumptions:

The sales data by day and country accurately reflect true transaction volumes without significant missing or erroneous entries.

Variations in daily sales between countries are meaningful and representative of market differences.

How to Use It:

Sales managers and regional directors use this chart to monitor and compare daily revenue generation in each country, enabling targeted marketing efforts, resource allocation, and localized decision-making.

Implications:

Significant differences in daily sales by country may highlight market opportunities or challenges. Sudden drops or spikes could indicate issues such as supply disruptions or successful campaigns that require further investigation.

Design Considerations for Data Analysis

1. Goals

Provide clear, actionable insights for business decision-makers to optimize sales strategies, marketing efforts, and inventory management.

Enable identification of key markets, product categories, and time periods that drive sales performance.

Facilitate forecasting to support proactive planning and resource allocation.

Ensure accessibility and usability of visualizations for diverse stakeholders.

2. Challenges

Ensuring data accuracy and consistency, especially across multiple countries and categories.

Handling seasonal and daily fluctuations in sales data while maintaining meaningful trends.

Managing different granularities of data (daily, monthly, by product, by category, by country).

Avoiding misinterpretation of visualizations due to assumptions like currency comparability and classification accuracy.

Presenting complex forecasts with confidence intervals in an understandable manner.

3. Factors Influencing Analyses

The business context of international chocolate sales, with multiple countries and product lines.

Availability and granularity of historical sales and shipment data.

The need for predictive analytics alongside descriptive and diagnostic analyses.

Stakeholders' focus on strategic planning, operational efficiency, and market growth.

4. Implementation of Features for Data Analysis

Data Modeling in Power BI: Use of star schema with fact tables for sales transactions and dimension tables for products, categories, countries, and time.

DAX Measures: Creation of custom DAX calculations for aggregated sales amounts, boxes shipped and forecast metrics.

5. Visualizations

Pie Charts for sales distribution by country.

Line and Bar Charts show trends over time by category, month, product, and day.

Scatter Plot with regression line to analyze correlation between shipment volumes and sales revenue.

Forecast visuals using Power BI's built-in forecasting feature with prediction intervals to display uncertainty.

Filtering and Slicing: Use of slicers to enable drill-down by country, category, and date ranges for dynamic exploration.

Time Intelligence: Leveraging Power BI's time intelligence functions to accurately aggregate and forecast sales across daily and monthly periods.

Data Refresh and Quality Checks: Scheduled data refreshes to maintain up-to-date analysis, with validation rules to handle missing or inconsistent data.

Tooltips and Annotations: Enhancing charts with explanatory tooltips and annotations to guide users in interpretation.

Key Findings

- There is no inherent difference between the sales of each country
- Dark chocolates are better selling than whites, fruits, and cocoa drinks
- January is the best-selling month, with a decreasing slope throughout the year
- There should not be any big change in the sales of chocolate in the last quarter of the year, considering the trend goes down in the later quarters and then goes back up in new years
- People buy more chocolate at the start of the month
- Preferred chocolate types don't differ by country

Conclusion

The comprehensive analysis of chocolate sales across multiple countries and product categories reveals several important insights that can guide strategic business decisions. The distribution of sales by country indicates that markets such as Australia and the USA dominate revenue generation, underscoring the need for targeted marketing and resource allocation in these regions. Meanwhile, the consistent preference for dark chocolate across countries highlights a strong product category that warrants continued focus.

Seasonal trends show that sales peak in January and gradually decline throughout the year, with notable dips in the last quarter before rising again at the start of the next year. This cyclical pattern suggests opportunities for promotional campaigns during slower months to maintain steady demand. Daily sales data further emphasize increased consumer purchases at the beginning of each month, which could inform inventory and staffing decisions to better align with customer behavior.

The correlation between boxes shipped and sales amount reinforces the reliability of shipment volume as a sales predictor, enabling more accurate forecasting and supply chain optimization. Forecast models with prediction intervals provide a valuable tool for planning, although wider intervals highlight the need for flexible strategies to manage uncertainty.

Overall, this analysis equips business leaders with actionable insights to optimize product portfolios, enhance marketing efforts, improve operational efficiency, and plan proactively for future demand. Continued attention to data quality, seasonal trends, and market-specific behaviors will be essential to sustain growth and competitiveness in the dynamic international chocolate market.