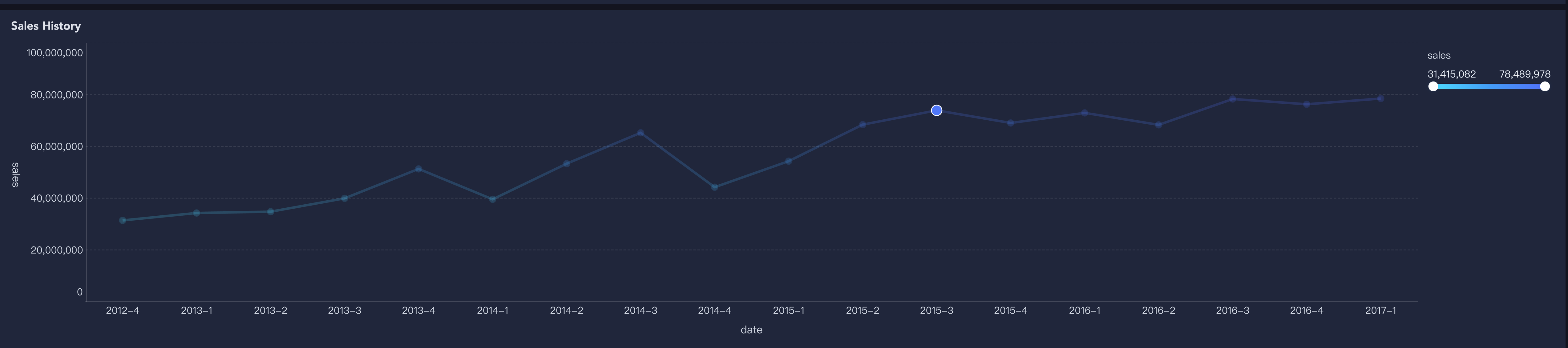
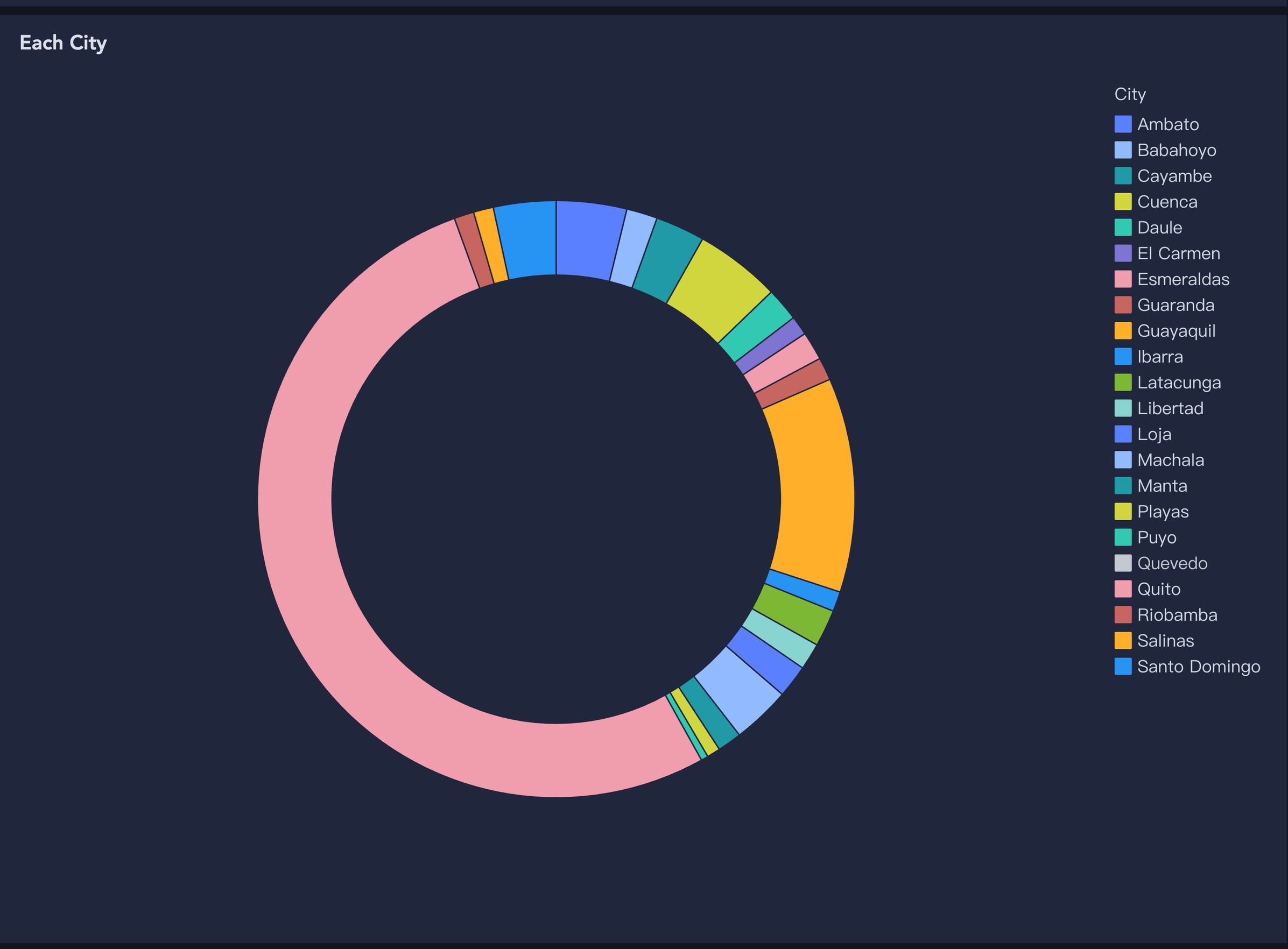
Supermarket Business Data Analysis and Forecasting

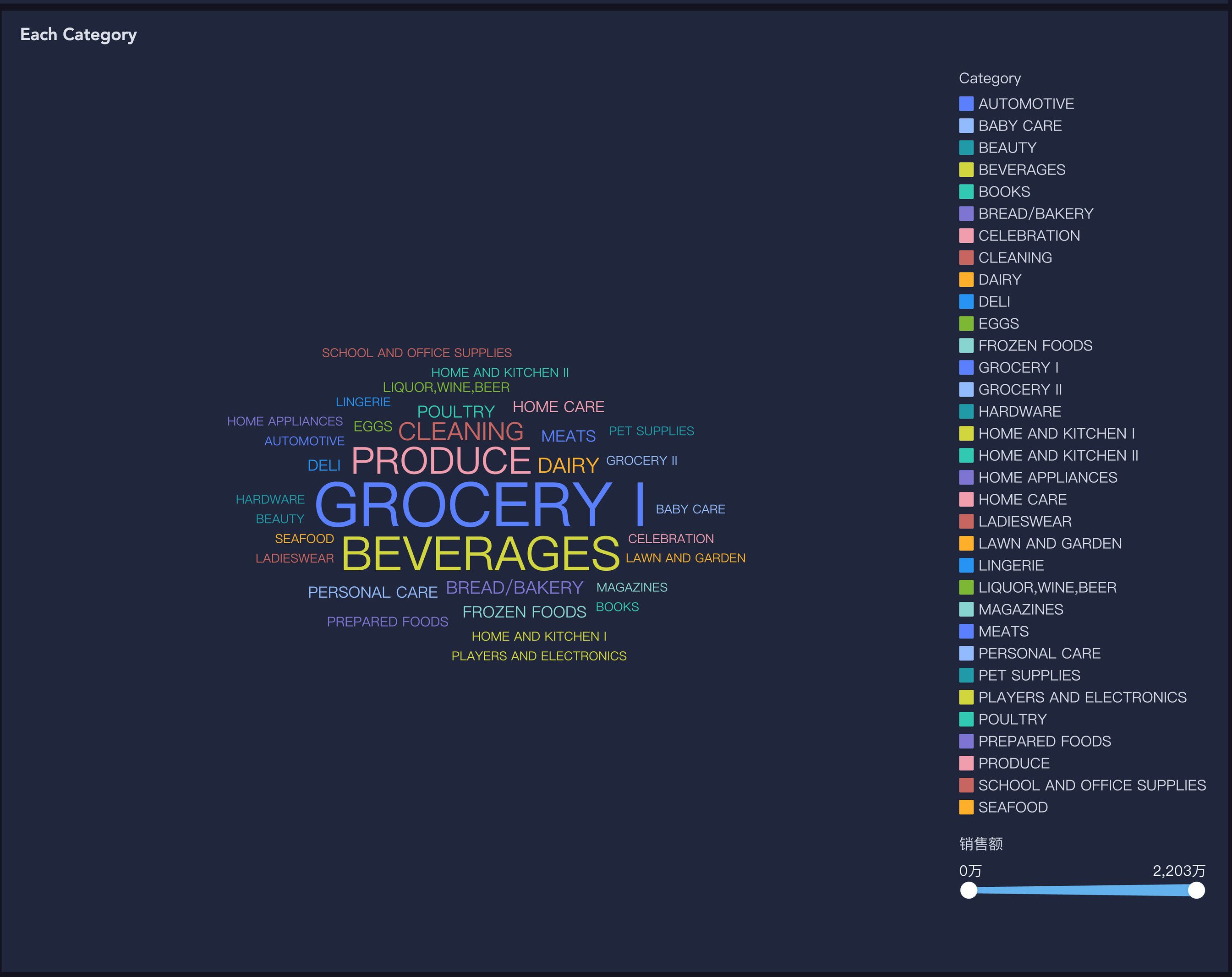
--Albert Wang

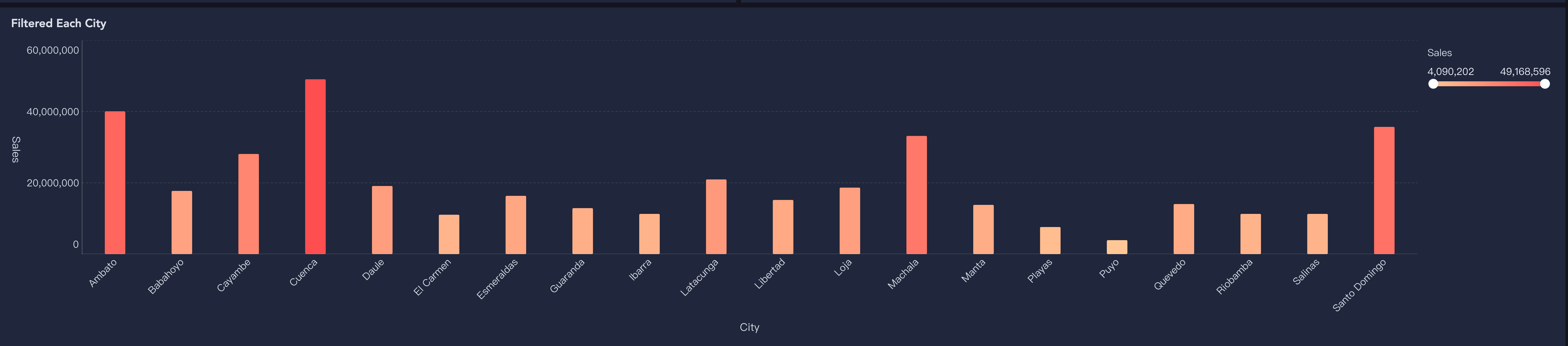
Data Feature: In the time dimension, the total sales of the 54 supermarket chains in Ecuador are steadily increasing. In terms of quarterly granularity, the fourth quarter of each year is the quarter with the highest sales for that year, mainly because Christmas falls in the fourth quarter, similar to the Chinese New Year period in China, when consumers spend heavily.



Data Feature: Guayaquil, Quito are an order of magnitude larger than the rest of the cities, the above chart shows us that Guayaquil, Quito are the key cities to focus on, based on this, if you want to look at the rest of the cities in detail, you can eliminate these two cities and then generate a bar chart to analyze. Distinguished by category, the larger share is for groceries, beverages, production appliances, cleaning products, bread, and dairy products.







Sales Rules and Strategies in Different Dimensions

1. Aggregate by 7 days of the week and analyze the pattern of sales volume and sales per day. There is not much difference in the volume of transactions between the days of the week, but there is a large difference in sales. The average sales per order are higher on Sundays, followed by Saturdays, and the average sales per order are lower on weekdays. This indicates that customers consume goods with the lowest unit price on Thursdays and the highest unit price on Sundays.

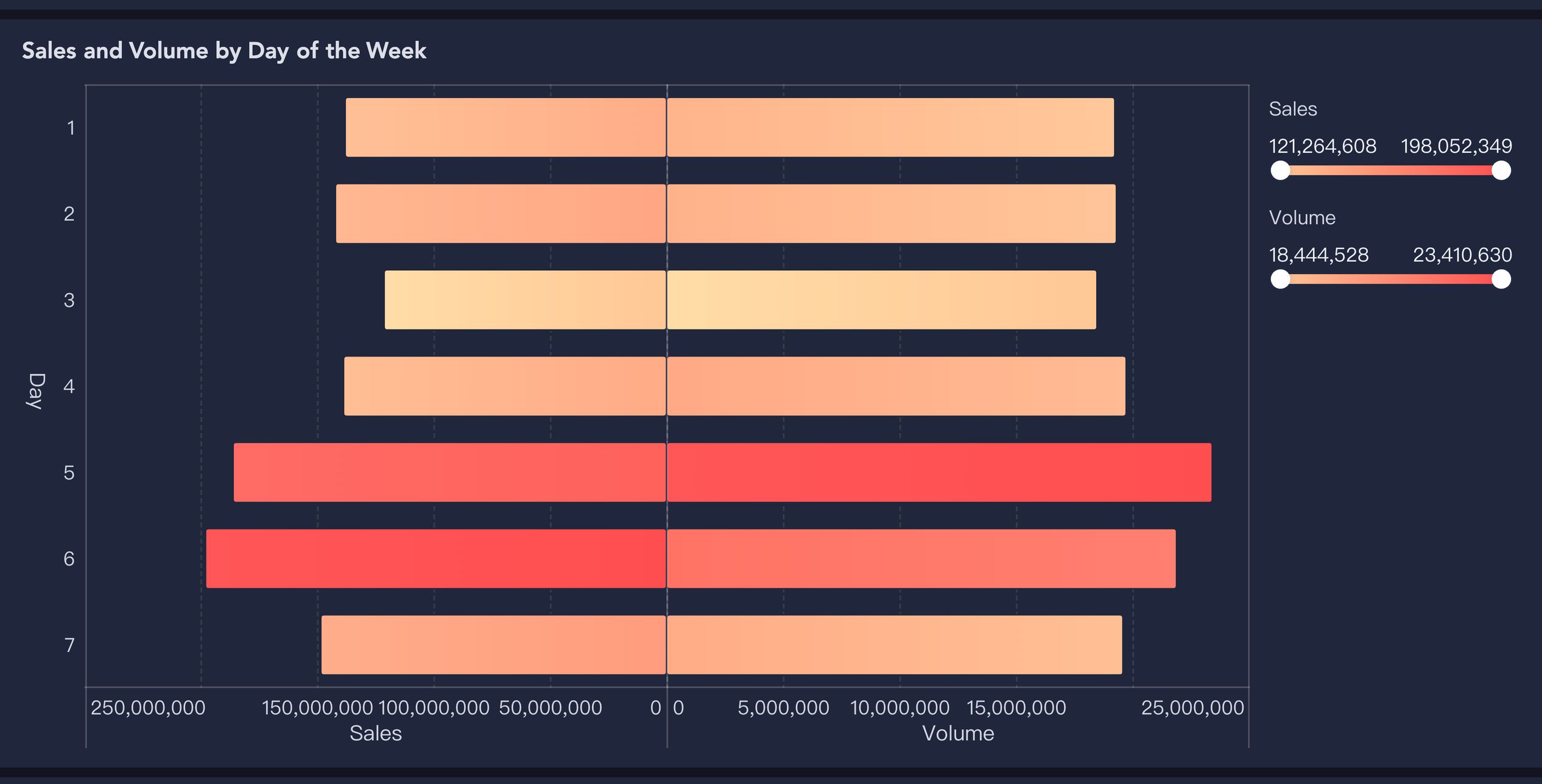
Sales strategy: Conduct promotional activities for goods with high unit prices on weekends to increase sales during the time when customers are more willing to buy.

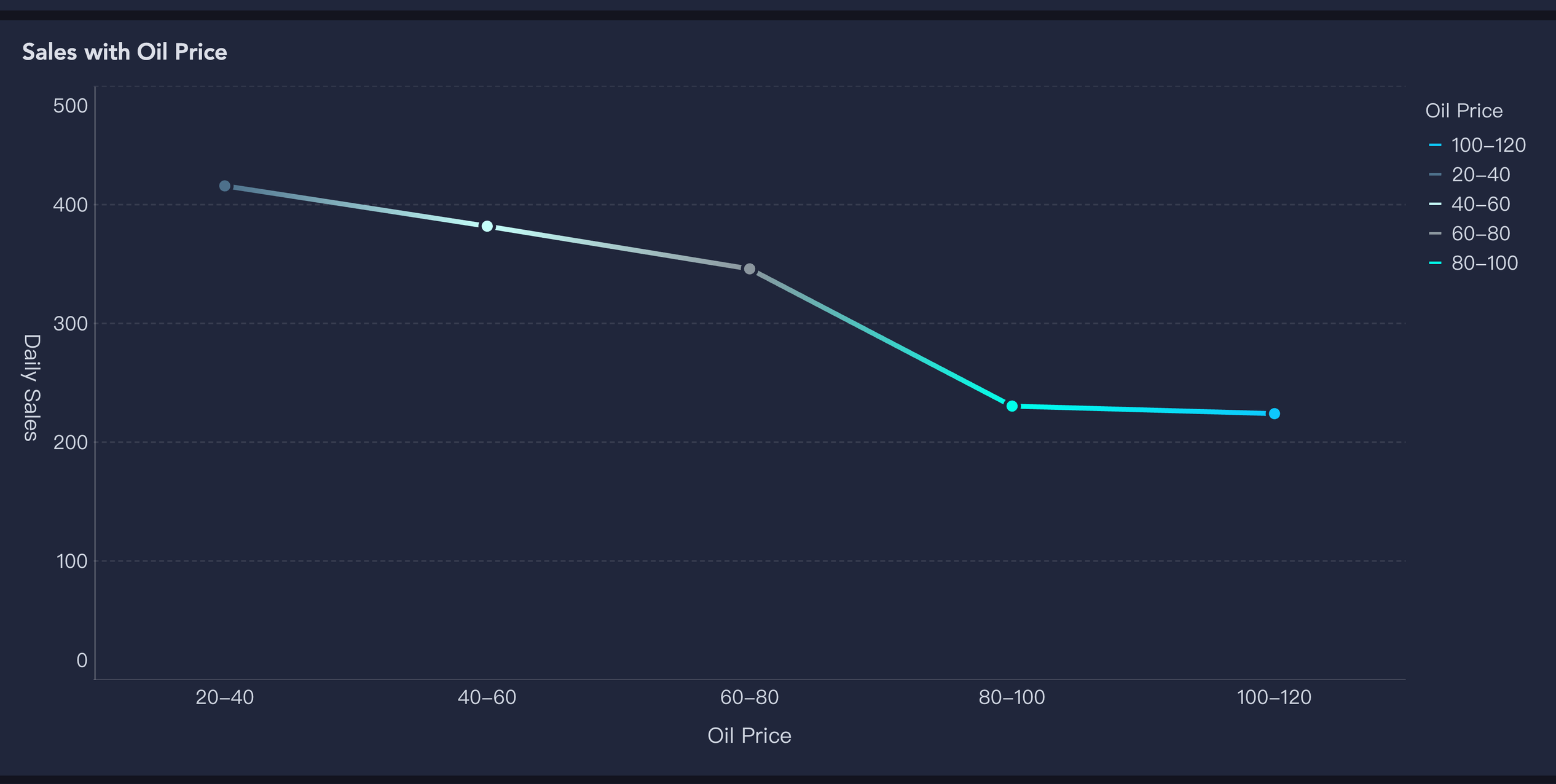
2. Among the categories with high sales, bread and dairy products are goods with a short shelf life and need to be supplied according to the sales time pattern. Bread accounted for 19.13% of the week's sales on Sunday, followed by 16.38% on Saturday, and the lowest was Thursday with 11.57%.

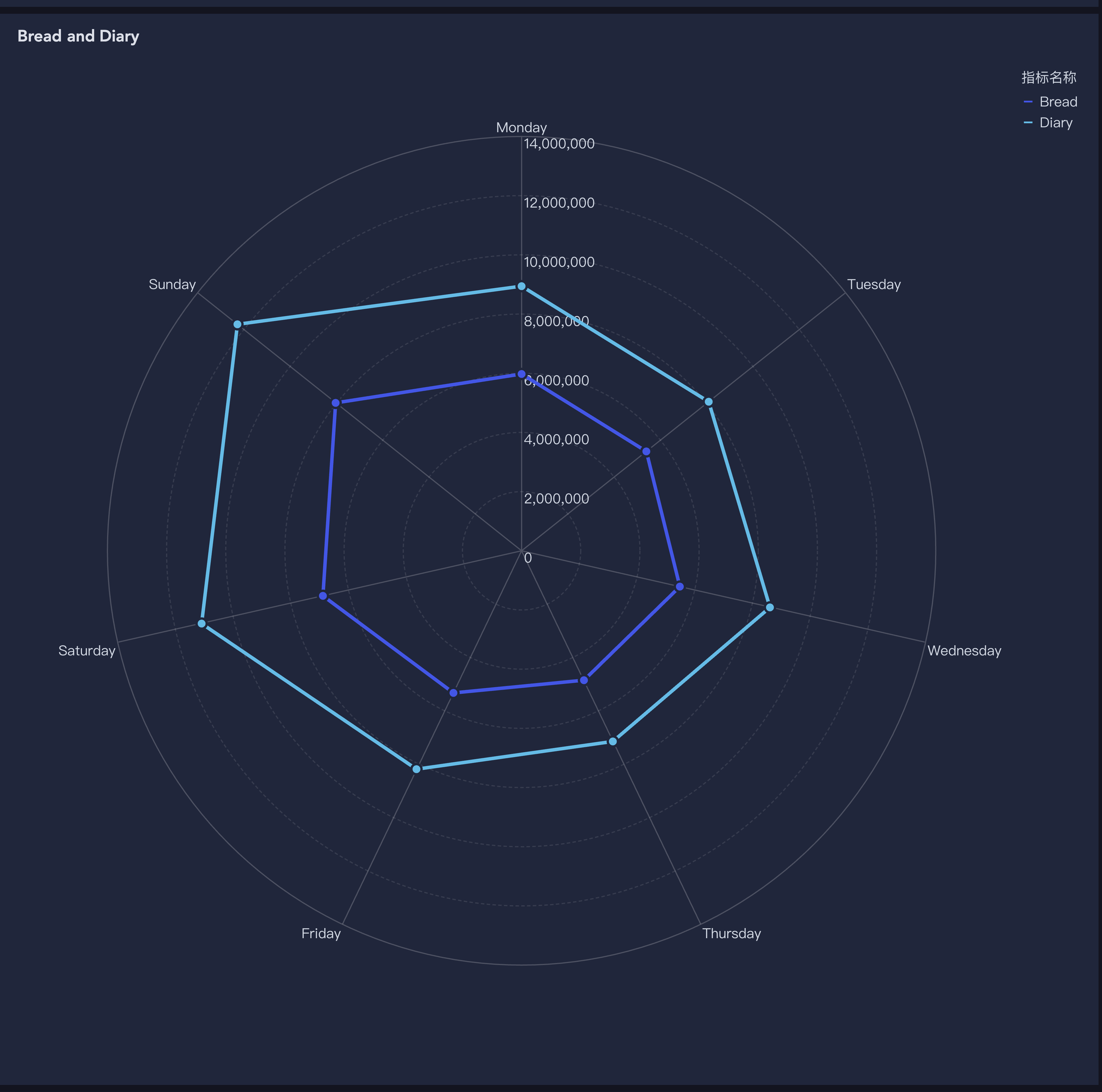
Sales strategy: offer bread on Saturday morning at 1.5 times the average weekday supply. Supply and consumer demand should be synchronized with cyclical changes. In order to make it easy for consumers to buy fresh bread at any time, the total amount of production for the week should be decided based on the total weekly demand for bread, and then supplied in accordance with the proportional relationship of weekdays: weekends = 1:1.5. Avoid too much supply on weekdays and consumers buy expired bread on weekends; or too little supply on weekdays, resulting in consumers buying expired bread produced on weekends.

3. Between 2013 and 2017, the price of oil experienced significant fluctuations, and Ecuador's economy is heavily dependent on oil, there is a chain of causality between: the price of oil ➡️ the economy ➡️ the purchasing power of consumers ➡️ the sales of supermarkets. It is possible to analyze what are the trends in the average daily sales of supermarkets on dates with different oil price ranges.

Sales strategy: increase the supply of goods during periods of low oil prices; reduce the supply during periods of high oil prices to avoid large stockpiles of goods.







Forecasting Sales of Various Categories of Goods by LightGBM Machine Learning Algorithm

Model Principle: Ecuador is highly dependent on oil, and during periods of low oil prices, the domestic economy is in better shape and average daily sales are higher; and vice versa. Combined with the holiday data and promotional data, the coefficient of influence of oil prices on sales can be calculated historically. The current price of oil can then be used to predict the sales of each item in the following period.

Model Accuracy

The LightGBM algorithm is utilized, and the data set is historical data such as oil prices, holidays, promotions, etc. combined with historical transaction data, and the scoring criteria areRMSLE:

The evaluation metric for this competition is **Root Mean Squared Logarithmic Error**.

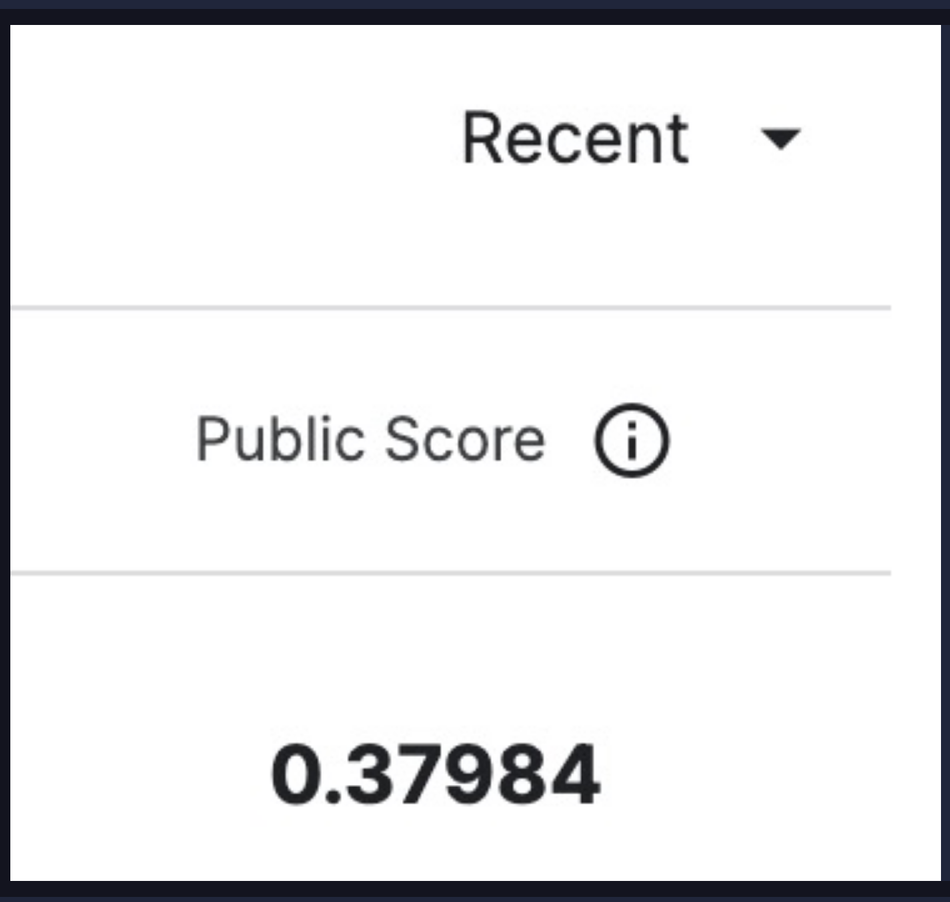
The **RMSLE** is calculated as:

[\sqrt{ \frac{1} {n} \sum_{i=1}^n \left(\log (1 + \hat{y}_i) - \log (1 + y_i)\right)^2}]

where:

- *n* is the total number of instances,
- *ŷ i*^ is the predicted value of the target for instance (i),
- *yi* is the actual value of the target for instance (i), and,
- log is the natural logarithm.

According to the RMSLE scoring criteria, the prediction result score of the LightGBM model is 0.37984. Typically, a score lower than 0.5 is a better prediction, indicating that the prediction has a smaller error from the actual.



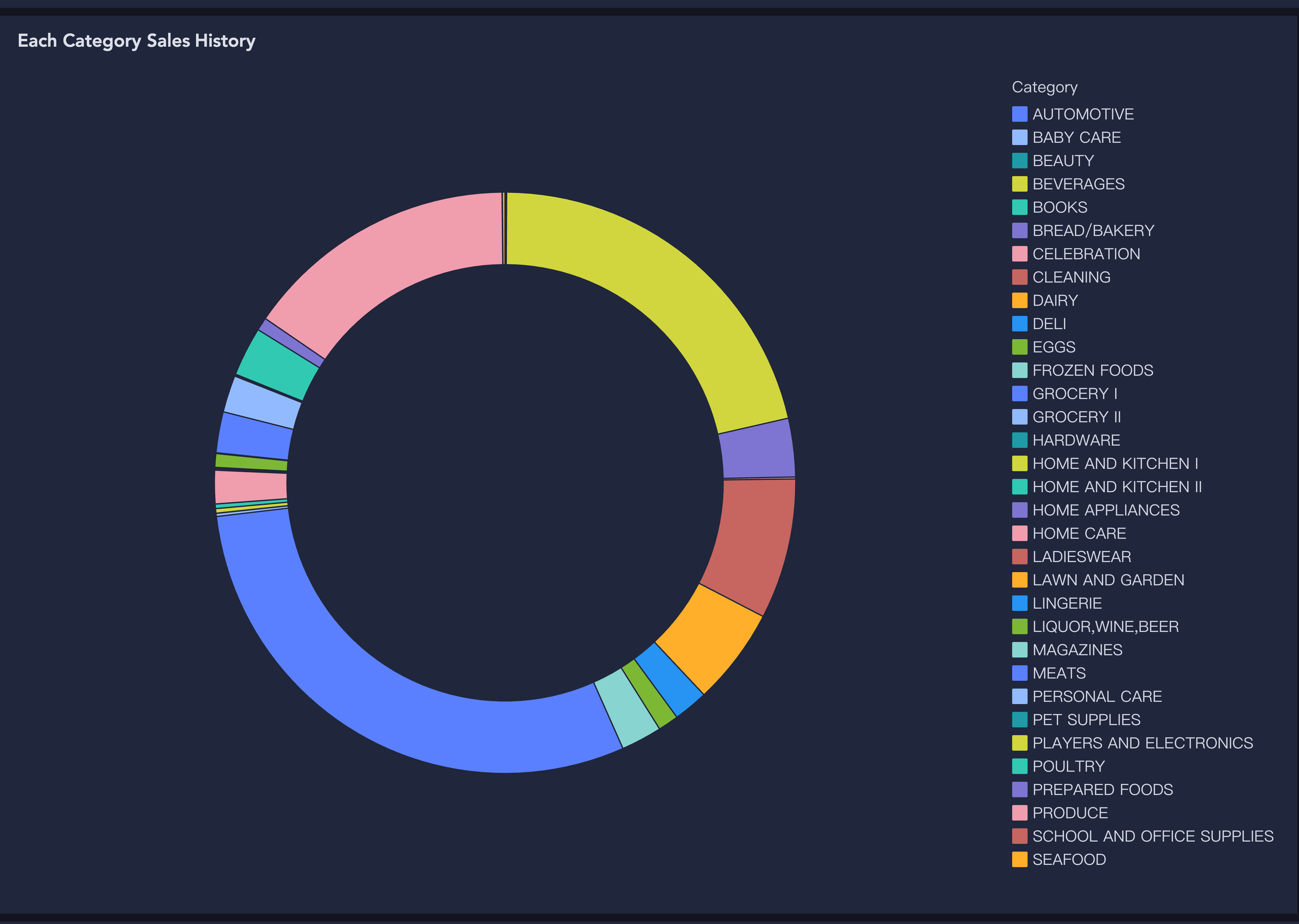
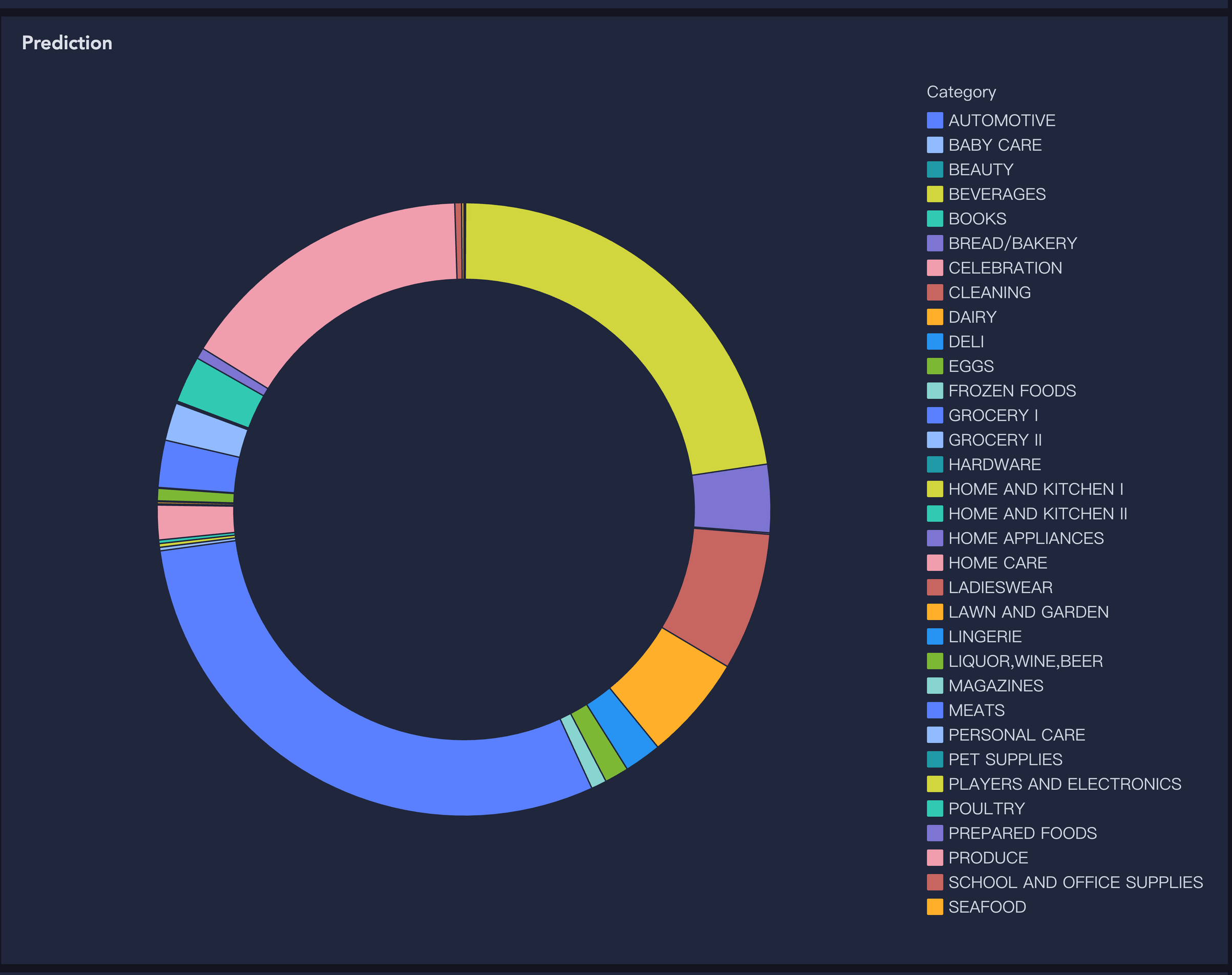
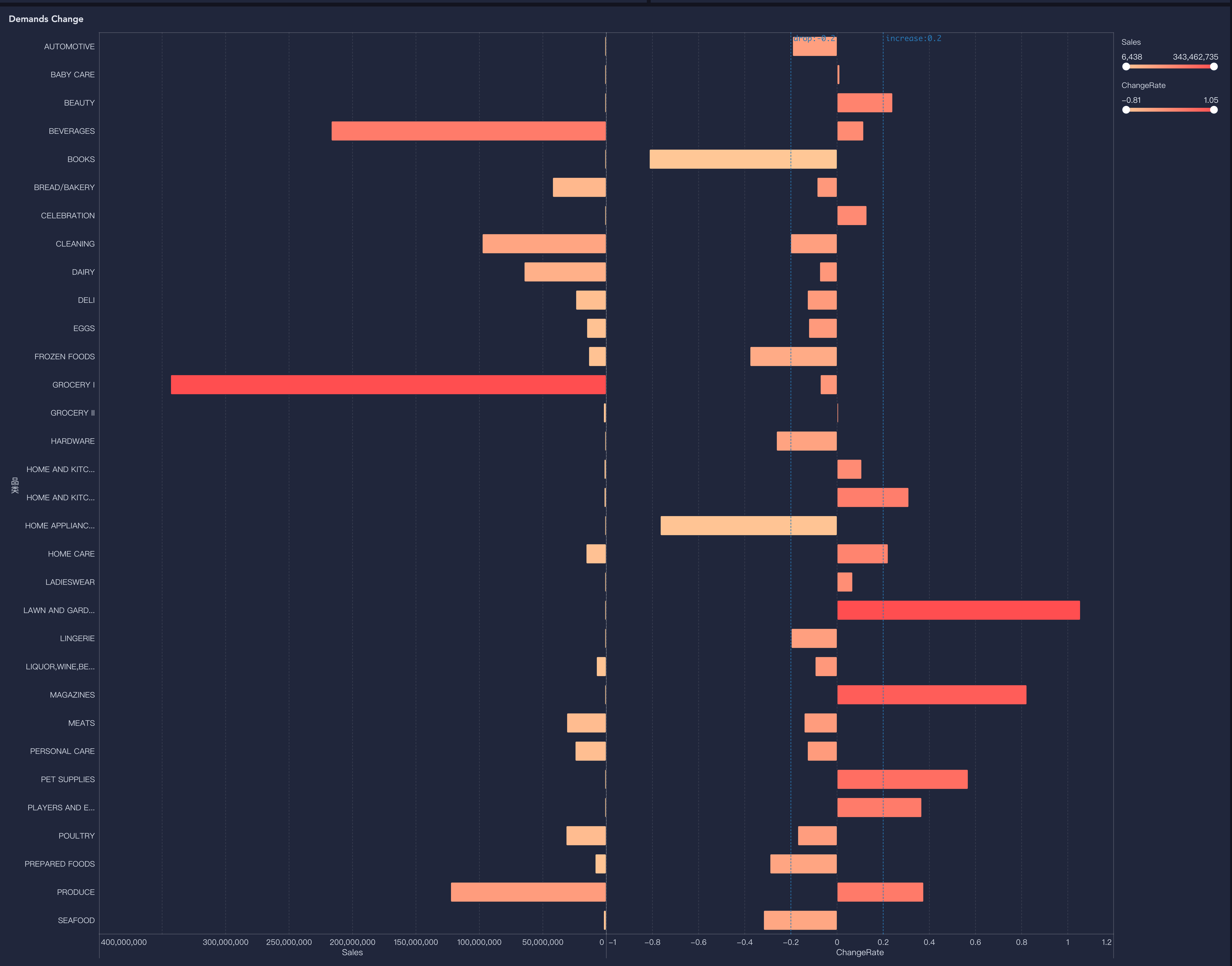
Recent ▾

Public Score ⓘ

0.37984

Forecast Results and Business Recommendations

1. Among the higher-volume product categories, production and beverages are projected to see larger sales increases. In the next phase, supermarkets can increase stock levels of beverages and produce without worrying about overstocking and expiration dates, while better meeting rising consumer demand for beverages. 2.

2. Frozen food and cleaning products, which have a history of high sales volume and a significant decrease in demand, should not be stocked more in the next period to avoid waste or buildup of large quantities of expired food products, which could lead to consumer dissatisfaction.