Final Report

Kaggle Store: Sales Data Analysis

Introduction

Kaggle is currently selling three products - Kaggle Mugs, Kaggle Hats, and Kaggle Stickers - through two store chains: KaggleMart and KaggleRama in three countries - Finland, Norway, and Sweden. The company is in the process of determining which store chain would be the best official outlet for their products. To aid in this decision-making process, I created a model that can predict future sales and evaluated the performance of both store chains. My objective is to recommend the most suitable store for an official outlet. I also examined and extracted useful insights from the sales data of Kaggle products by analyzing and visualizing sales patterns, trends, and seasonality in each store. This project will also assist in optimizing inventory management, thereby improving the overall operational efficiency of the stores.

Data

I obtained the dataset from Kaggle, which is a time series data that shows the sales of three different types of Kaggle products from January 01, 2015 to December 31, 2018. The raw data consists of 6 columns - row_id, data, country, product, store, and num_sold - with a total of 26298 rows.

Exploratory Data Analysis

The overall sales trend by country, store, and product is summarized in Figure 1. The analysis shows that the Kaggle Hat was the most popular product, while the Kaggle Stickers had the lowest sales. Norway had the highest sales among the countries,

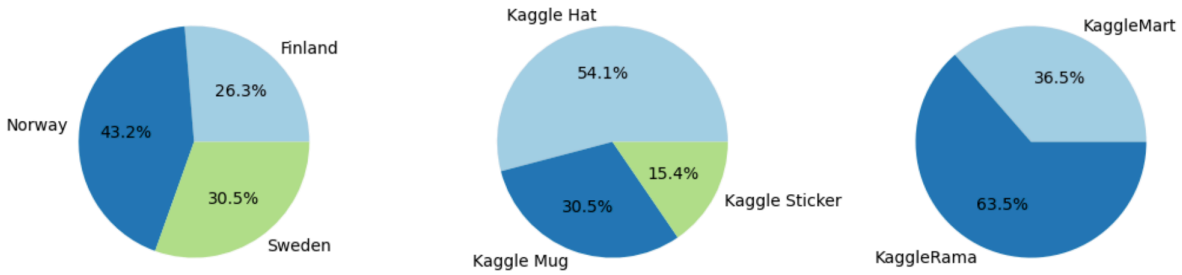

Figure1. Total sales by country, product and store

indicating a strong demand for the products. Sweden and Finland followed in the second and third positions, respectively. Moreover, KaggleRama sold 1.7 times more products than KaggleMart, indicating a significant difference in sales volume between the two stores.

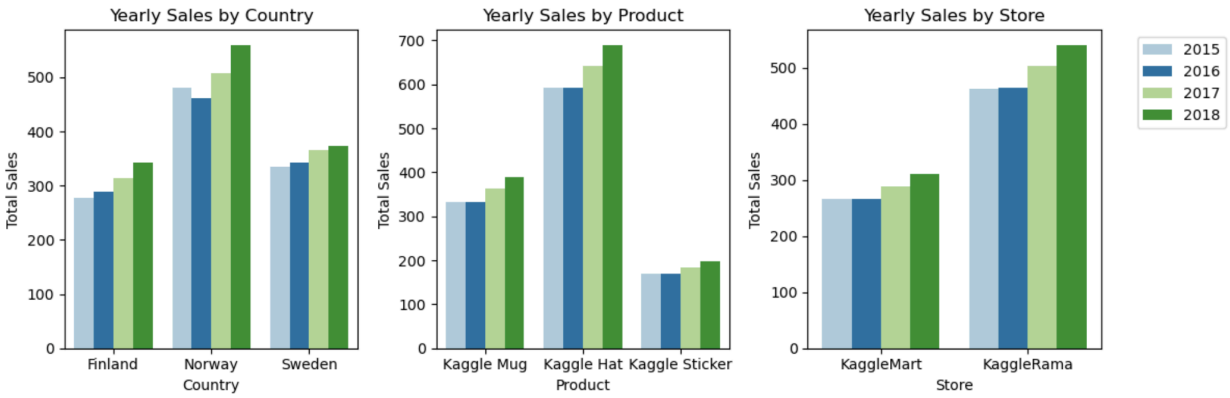

Figure 2. Yearly sales by country, product and store

Figure 2 shows that there has been a steady increase in sales across all three countries, indicating a consistent upward trend of growth. While most countries saw a constant rise in sales between 2015 and 2018, Norway experienced a decline in sales from 2015 to 2016. However, this temporary decrease did not impede the overall upward trend in sales.

Several key insights were discovered by analyzing seasonal patterns and trends across various stores and countries. It was observed that the highest sales occurred during the spring (March to May) and winter (December and January) seasons. The sales plot for each month demonstrated a consistent pattern across all stores and countries, indicating a synchronized demand trend. However, unique seasonal patterns were observed for each product. For instance, Kaggle Hat experienced the most significant seasonal variation, with its highest sales peak in April and December but a dip in September and October.

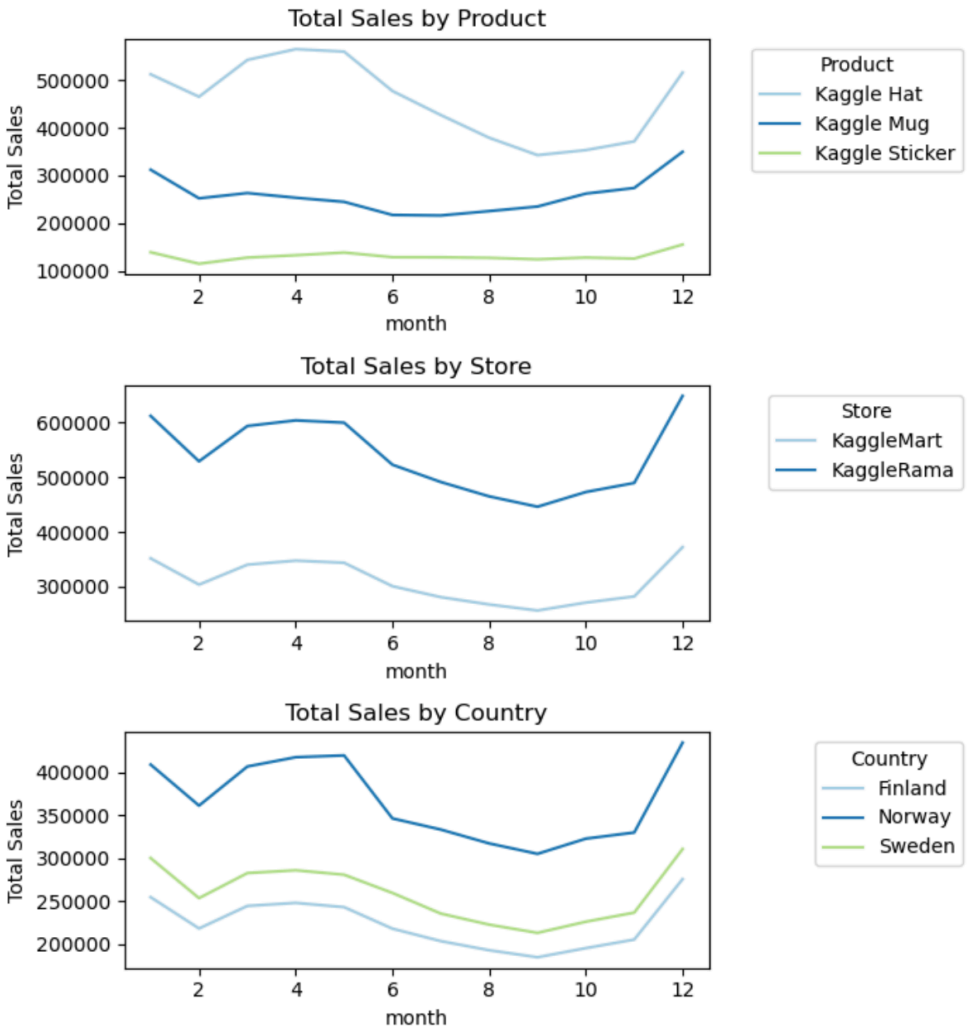

Figure 3. Seasonal Sales Patterns

On the other hand, Kaggle Mug had its peak sales in December but declined in July and August. Surprisingly, stickers had steady sales throughout the year, regardless of the store

or country. Additionally, it was observed that sales remained stable during weekdays but spiked on weekends. Interestingly, all three products experienced a global surge in sales during December, indicating a strong correlation between holiday seasons and increased consumer demand. These insights provide valuable information for stores to strategically plan inventory, promotions, and marketing campaigns to maximize profitability and customer satisfaction.

Time Series Decomposition and Stationarity

Breaking down data into a trend, seasonality, and residual (or error) components is known as decomposing time series. This process helps us gain a better understanding of long-term patterns, identify seasonal effects, analyze residuals for anomalies, and improve forecasting and modeling accuracy. Since we witnessed the presence of two seasonal effects in the previous EDA part, I used the MSTL decomposition technique to inspect the seasonal components. MSTL stands for multiple seasonal trend decomposition using Loess. Figure 4 shows that the weekly and yearly seasonality is well captured using MSTL. Seasonal_7 depicts the weekly seasonality, with lower weekday sales and a peak

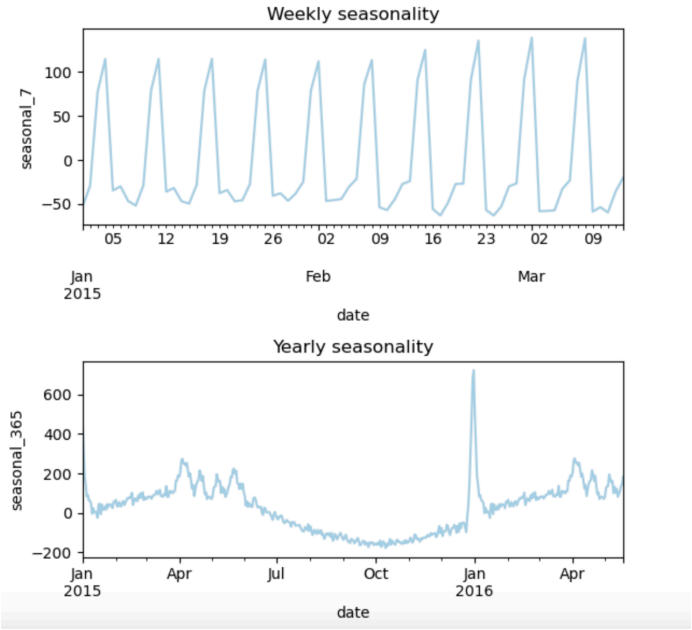

Figure 4. MSTL decompostion

on weekends. Seasonal_365 shows the seasonality that changes over the month, confirming the yearly sales pattern identified in the previous EDA part.

When it comes to forecasting time series, the statistical approach relies on the assumption of stationarity. A time series is considered stationary when its properties remain consistent regardless of when it is observed. Unfortunately, our time series has trends and seasonality, which means it is not stationary. I used the Augmented Dickey-Fuller (ADF) test to confirm this. However, I was able to transform the non-stationary original time series into a stationary one through differencing.

Modeling

When splitting time series data into train and test sets, temporal order must be preserved, unlike random sampling used for other data types. I utilized the most recent 100 days' data as the test set.

I first build a naive seasonal model as a baseline. I then test 3 different models: SARIMA, BATS, and Prophet. To compare the performance of different forecasting models, I relied on MAPE (Mean Absolute Percentage Error). MAPE measures the average percentage difference between the predicted and actual values, offering a relative indicator of forecasting accuracy.

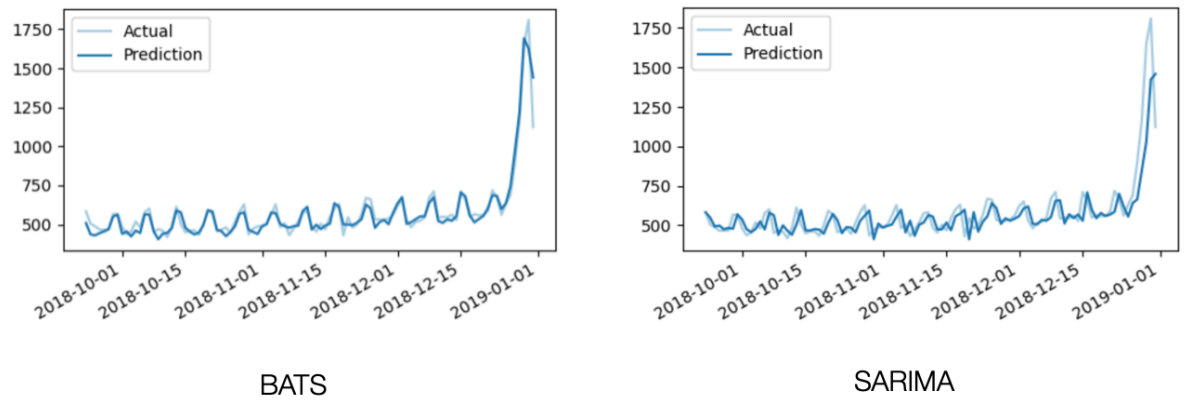

Figure 5. Classical approach

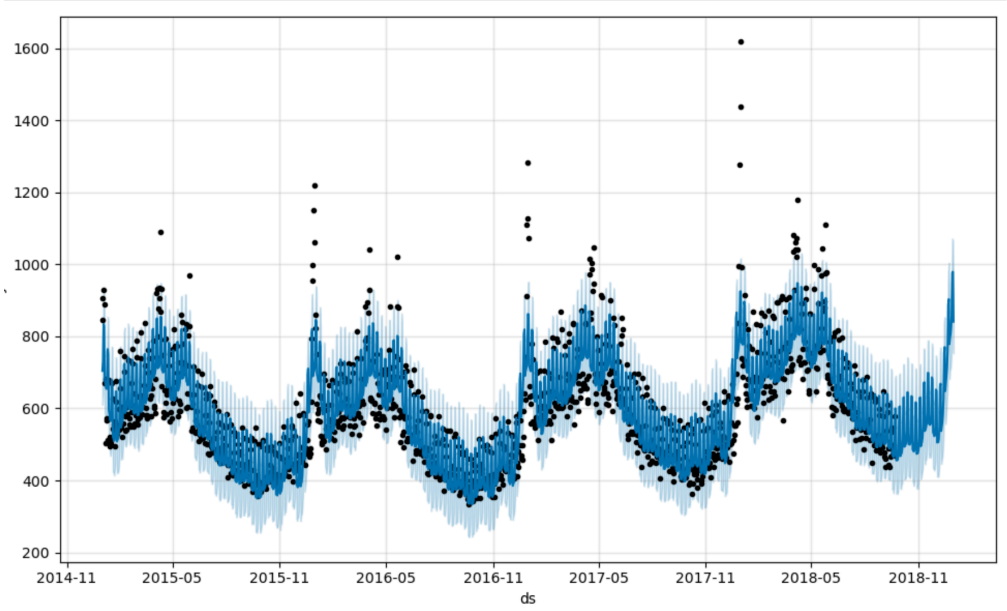

Figure 6. Prophet

The baseline model, seasonal naive, achieved a MAPE of 0.131, serving as a reference for model comparison. The SARIMA model outperformed the baseline model, yielding a lower MAPE of 0.103. SARIMA model, which takes into account the autoregressive, moving average, and seasonal components of the data, provides improved forecasting accuracy. However, the SARIMA model is only designed to handle a single seasonality. On the other hand, the BATS model, which uses exponential smoothing, ARIMA model, residuals, and Box-Cox transformation, can handle data with multiple seasonal patterns and produced more robust predictions with the lowest MAPE of 0.053. As Figure 5 shown, it appears that BATS is more effective at capturing end-of-year seasonality compared to SARIMA. Despite incorporating holiday features, the Prophet model with a MAPE of 0.264 did not outperform the other models in this analysis.

Future work

To enhance forecasting performance, it may be beneficial to fine-tune and optimize models, especially by adjusting the configuration of the Prophet model. Incorporating additional relevant features, such as economic indicators or country-

specific holiday features, can also assist in capturing more underlying relationships to improve forecast accuracy. Advanced time series models such as recurrent neural networks (RNNs) or long short-term memory (LSTM) networks are known for effectively capturing complex temporal patterns. Applying those kinds of time series model can provide improved forecasting capabilities. Outliers or anomalies can have a significant impact on time series analysis. Identifying and handling them appropriately can improve the reliability of the models.