Title:

Forecasting Seasonal Trends for Dairy Queen Using Google Trends Data

Author:
Kenji Wagner

Dataset Name:
Dairy Queen Google Trends Time Series (Monthly Data, 2009–2019)
https://trends.google.com/trends/explore?date=all&geo=US&q=%2Fm%2F021fdl&hl=en

Data Source:
Google Trends – Downloaded from Google Trends and augmented in R

Objective:
To analyze and forecast seasonal search interest in Dairy Queen using time series modeling. The goal is to understand how consumer search interest fluctuates throughout the year and assess the effectiveness of different forecasting models for seasonal business analysis.

Final Models Used:

Manual SARIMA: (1,1,1)(0,1,1)[12]
RMSE: 8.19 | MAPE: 11.65%
Auto ARIMA: (2,0,0)(0,1,1)[12]
RMSE: 5.80 | MAPE: 6.99%
Holt-Winters Additive:
RMSE: 11.95 | MAPE: 15.28%

Conclusions:

The data exhibited strong seasonality and an increasing trend, typical of a business like Dairy Queen that peaks in popularity during the warmer months.
The SARIMA models outperformed the Holt-Winters model, which struggled with capturing more nuanced time dependencies, using AICc, exhibiting the strength of the metric.
The auto.arima-selected SARIMA model was the most accurate, outperforming the manually tuned SARIMA model.
The final model provided a reliable forecasting tool that businesses like Dairy Queen can use to anticipate customer interest and adjust operations accordingly (e.g., marketing, staffing, and inventory decisions).
This project illustrates the value of Google Trends as a data source for analyzing consumer behavior, particularly for seasonal businesses.

Background:

Google Trends is a free service from Google that allows you to visualize the popularity of google search queries over a select period of time. It is useful for understanding and identifying trends, which can be useful for businesses. It is specifically useful for understanding interest in businesses over time, which can be useful for companies that supposedly have seasonal business models. Seeing as how the summer is coming up, I thought about how Dairy Queen is a business that does follow seasonal trends, given that some Dairy Queens aren't open all year (some are closed in the winter). I want to see if this necessarily affects their search trends, which in turn would significantly affect their business.

For ethical reasons, Google Trends doesn't show exact numbers as it pertains to search queries. Instead, it chooses an unbiased random sample of search queries and normalizes those numbers to show us an unbiased representation of search interest for specific terms. Even though this doesn't show us exact numbers and the scale isn't necessarily directly proportional to the exact numbers, the methodology makes sense and should provide a fairly accurate representation of the search trends we are looking for.

Preprocessing & Initial Thoughts on Data:

Google Trends normalizes their data so that trend searches are normalized, then indexed on a scale of 1-100, with each point being divided by the highest point (being 100). So I didn't have to scale the data to make it more interpretable, but I did have to query the years I wanted to use because I downloaded the csv file from Google Trends from the start of when they took down data (2004). Instead of just querying the data in Google Trends, I chose to query it in R for future exploration purposes (meaning that if I wanted to include values after the range I chose a query that had a higher interest than those in the range of query in my forecast test set, I could).

I chose to use approximately 10 years of monthly data (2009–2019) to ensure a robust and representative time series without including outdated or potentially irrelevant trends. I ended the dataset in 2019 to avoid potential issues caused by the COVID-19 pandemic in 2020, which could introduce abnormal behavior. The data was split into an 80/20 ratio, with 2009–2017 as the training set and 2018–2019 for testing the forecast accuracy.

The model showed a positive trend with clear seasonality, with search results peaking during the summer time as shown in Figure 1. The data did appear to have constant variance across the model, so it made sense to plot out the additive decomposition of the model to make sure my conclusions were correct, which they were as shown in Figure 2. This was evidenced by the increasing trend component and strong seasonal pattern evidenced by the predictable fluctuations in the seasonal component.

I used differencing to neutralize the trend component; however, I didn't think to use seasonal differencing given the strong seasonal pattern evidenced by the predictable fluctuations in the seasonal component at first. However, seasonal differencing can sometimes simplify the ARMA modeling of the residuals by removing the strong seasonal autocorrelation,

making it easier to identify the remaining short-term dependencies so I chose to apply seasonal differencing also even though the seasonal component was consistent.

Model selection rationale:

Seeing that there was a trend component as well as a seasonal component, I chose to leverage three models to predict this seasonal time series. I chose to use two seasonal arima models and one Holt Winters model. I started with a manually specified SARIMA model to understand the structure of the data and control the modeling choices more directly, using ACF and PACF of the differenced data to find potential models. I also used the auto.arima() function to automatically select a SARIMA model based on model selection criteria, being AICc (corrected Akaike Information Criterion), given that auto.arima was optimized to find the best model, not restricting the model from choosing a model without a differencing term so it could just find the best model. Additive Holt-Winters is a good model for seasonal time series with trend, particularly when the seasonality is stable over time (as it appears to be here).

I used AICc as the criterion for selecting and comparing SARIMA models because it is a more reliable version of the Akaike Information Criterion (AIC), particularly when working with small to moderately sized time series datasets—which applies here, since I'm using about 8 years of monthly Google Trends data (i.e., around 120 data points) for the model.

Model results:

SARIMA model

For the manually selected SARIMA model, I relied on the ACF and PACF to give me candidate models to choose from (Figure 3). From these plots, I found the ACF had a potential seasonal trend, although it could be argued it is just decaying, and a spike at lag one, so I concluded that there may be an MA(1) component for seasonal and non-seasonal. The PACF plot showed me a spike at lag one in PACF, suggesting a potential AR(1) component, but the seasonal component had no significant values.

From these results I chose to evaluate combination of models that had MA(1) components for seasonal and non-seasonal and AR(1) for the nonseasonal component, also with differencing in the non-seasonal element for all versions of the model and some also containing seasonal differencing, given how useful it was for the steps in the model identification. The candidate models being, SARIMA(1, 1, 1)(0, 0, 0)[12], SARIMA(1, 1, 1)(0, 1, 0)[12], SARIMA(1, 1, 1)(0, 0, 1)[12], SARIMA(1, 1, 1)(0, 1, 1)[12], SARIMA(1, 1, 0)(0, 0, 1)[12], and SARIMA(0, 1, 1)(0, 0, 1)[12]. I used AICc to evaluate which model worked the best and it was the SARIMA(1, 1, 1)(0, 1, 1)[12] model with an AICc of 513.886486796247. The forecast looked somewhat accurate (as shown in Figure 5), forecasting slightly lower values, but still scoring an impressive RMSE of 8.19091790763365 and MAPE of 11.6482847669358 (Figure 8).

Auto ARIMA model:

The auto ARIMA was much simpler, I just evaluated it based on AICc and had a max p value of 5 and max q value of 5, not restricting it to require differencing terms. I ended up finding the best model was a SARIMA(2,0,0)(0,1,1)[12], which was much different than my manually selected model. The forecasting for future values (as shown in Figure 6), RMSE, and MAPE were better than the model I had selected for the manual selection, with the forecasting still having values a little lower than the actual results. The RMSE and MAPE equaling 5.7978080885455 and 6.99309126002932, respectively (Figure 8).

Holt Winters result:

I chose to use an additive Holt Winters model because it is a good model for seasonal time series with trend, particularly when the seasonality is stable over time, as previously mentioned. This model forecasted well as shown in Figure 7, but was far inferior to the SARIMA models and forecasted values had much lower search interest, as evidenced by its respective RMSE and MAPE scores, being 11.9451977395732 and 15.2832730944308 (Figure 8).

Conclusion:

This project was effectively able to demonstrate the value of time series forecasting for understanding seasonal business trends for Dairy Queen using normalized Google Trends data. The data revealed a clear seasonal pattern, with higher search interest occurring in the summer months, which makes sense given the business' seasonal operations.

By applying multiple forecasting methods—including a manually tuned SARIMA model, an automatically selected SARIMA model using auto.arima(), and an additive Holt-Winters model—I was able to compare their respective forecasting accuracies using both RMSE and MAPE metrics. Among the models:

The manually selected SARIMA(1,1,1)(0,1,1)[12] performed decently, capturing the seasonality and trend well with an RMSE of 8.19 and MAPE of 11.65%.

The auto.arima-selected SARIMA(2,0,0)(0,1,1)[12] significantly outperformed the manual SARIMA model, with a lower RMSE of 5.80 and MAPE of 6.99%.

The Holt-Winters additive model, while useful for capturing basic seasonal trends, underperformed with the highest RMSE (11.95) and MAPE (15.28%), suggesting it was less effective for this time series data.

Ultimately, the auto.arima-selected SARIMA model performed the best of the three which makes sense, given the optimization of the auto.arima package. However, the high performance found on all three of the models helped make this project successful and given how close some of these models were it emphasizes the trying multiple approaches. This model in a more broad sense could be used to anticipate customer interest and adjust operations accordingly (e.g., marketing, staffing, and inventory decisions), revealing the true power of time series.

Appendix:

Figure 1:

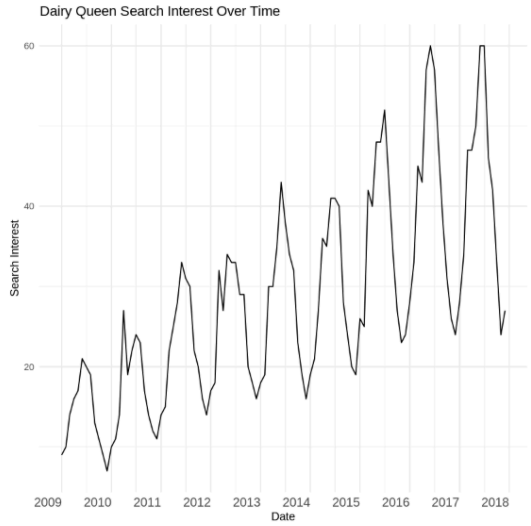

Figure 2:

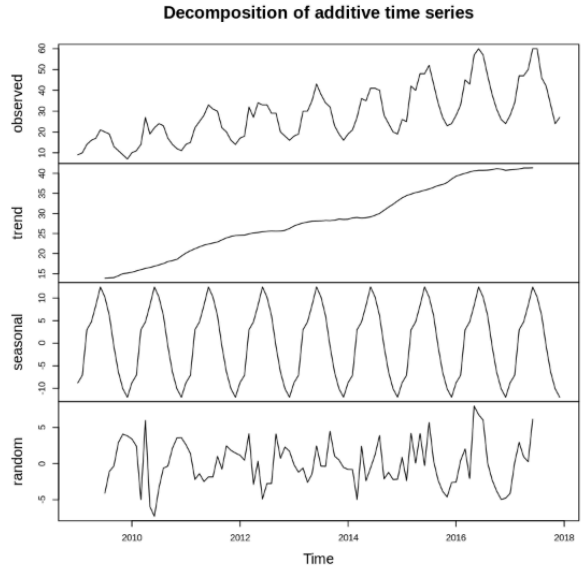

Figure 3:

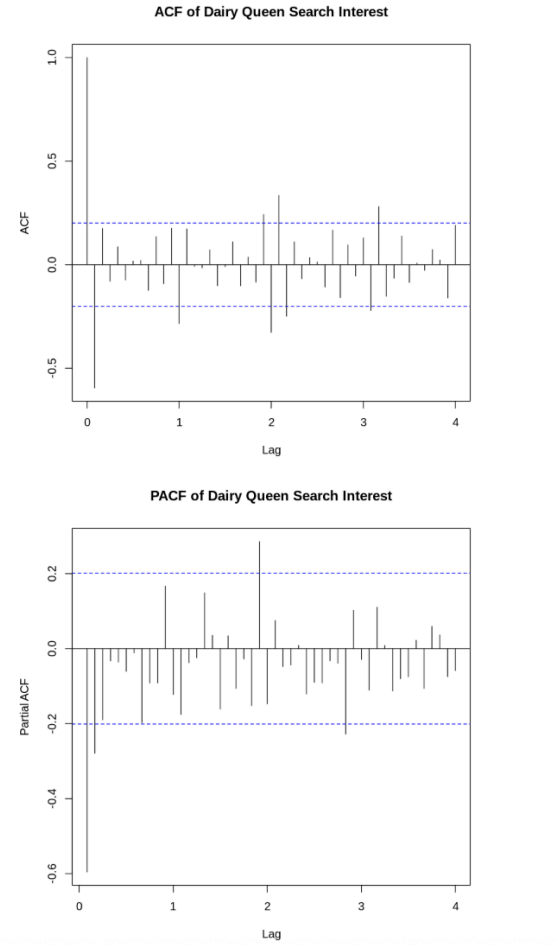

Figure 4:

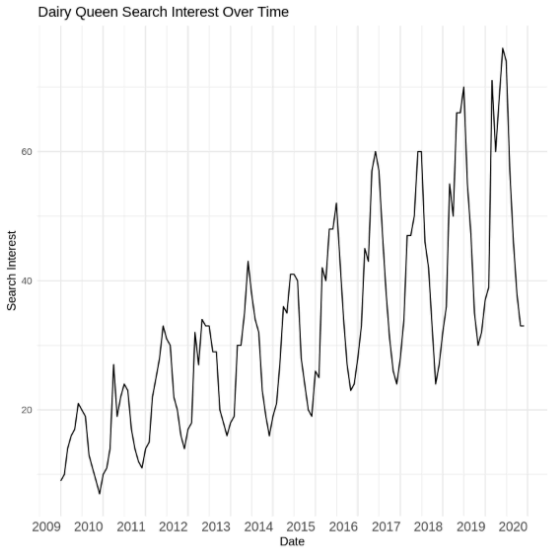

Figure 5:

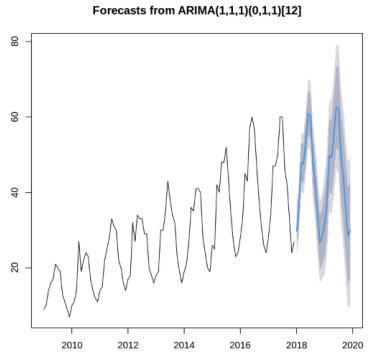

Figure 6:

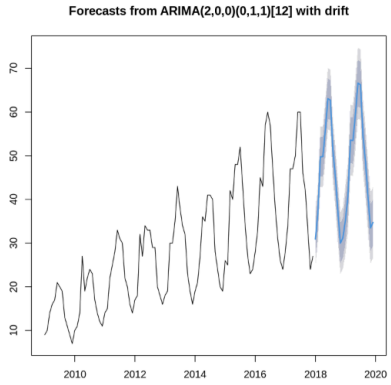

Figure 7:

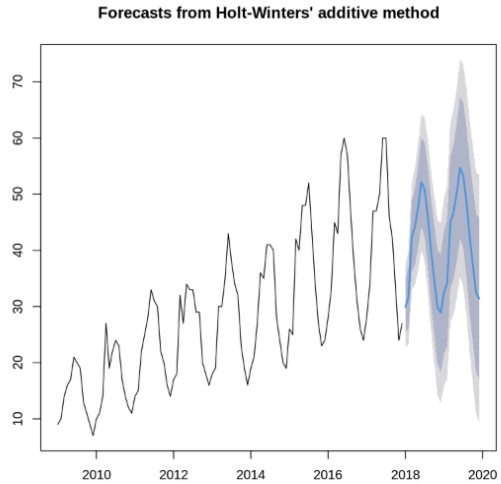

Figure 8:
Manual SARIMA: (1,1,1)(0,1,1)[12]
RMSE: 8.19 | MAPE: 11.65%
Auto ARIMA: (2,0,0)(0,1,1)[12]
RMSE: 5.80 | MAPE: 6.99%
Holt-Winters Additive:
RMSE: 11.95 | MAPE: 15.28%