Optimizing Coffee Shop Operations: A Data-Driven Approach to Understanding Sales Trends

Tarane Javaherpour and Edgar Rosales

Shiley-Marcos School of Engineering, University of San Diego

Exploratory Data Analysis (EDA) and Data Quality Assessment

1. Introduction

This section provides an in-depth analysis of the coffee shop transactional dataset, covering the exploratory data analysis (EDA) steps and the assessment of data quality. The dataset spans from January to June 2023, with variables related to transaction date, transaction quantity, and unit price. The goal is to understand patterns in the data, identify any data quality issues, and assess its suitability for forecasting future coffee shop sales.

2. Data Preparation and Cleaning

a. Date Conversion: To facilitate analysis, the transaction_date variable was converted from a character string to the Date format using lubridate::mdy(). This step ensured that the dates were correctly recognized and enabled time-based aggregation (weekly and daily sales).

b. Aggregation: The sales data were aggregated at two time levels:

- **Weekly Sales:** Sales were aggregated into weekly totals by using the floor_date() function to group by week.
- **Daily Sales:** Sales were also aggregated into daily totals by directly grouping by transaction_date.

c. Handling Missing Values: There were no missing or invalid dates in the dataset. Each transaction had a valid transaction_date, ensuring the integrity of time-series analysis.

d. Outliers: Outliers were identified in both the weekly and daily datasets, particularly in extreme sales values. For example, certain days displayed spikes in sales, likely due to specific promotions or customer behavior. While these outliers were not removed during the initial exploration, their impact was mitigated through smoothing techniques (e.g., moving averages).

3. Exploratory Data Analysis (EDA)

a. Sales Trends: We analyzed both weekly and daily sales data, observing an overall upward trend in sales from January to June 2023. A key finding was that sales increased steadily, particularly in the second quarter of the year, suggesting positive business growth or increased customer demand.

b. Weekly Sales Trend: The weekly sales plot shows a consistent increase in sales from January to June, with notable spikes in April and May. The 3-week moving average smooths out short-term fluctuations and highlights the underlying upward trend, indicating that the coffee

shop is experiencing growing demand. The plot and moving average together offer insights into business performance over time.

c. Daily Sales Trend: The daily sales plot revealed more volatility compared to the weekly data, with sharp fluctuations occurring from day to day. Despite this variability, a general upward trend emerges, particularly towards the end of the period. The 7-day moving average smooths out the daily fluctuations, providing a clearer picture of sales growth.

d. Seasonality and Trends: Due to the short six-month dataset, we were unable to detect clear seasonal patterns. However, the increasing trend across both daily and weekly aggregations suggests that the business is gaining traction. To better understand seasonality, data spanning over a year would be required.

e. Smoothing: To address the variability in daily sales, we applied a 7-day moving average to the daily data, and a 3-week moving average to the weekly data. These transformations helped smooth out extreme fluctuations and provided a clearer view of the underlying sales trends.

4. Data Quality Assessment

a. Completeness: The dataset is complete with no missing values in key variables (e.g., transaction_date, transaction_qty, unit_price). There were no missing records or discrepancies found during the data preparation process.

b. Consistency: The data is consistent across the variables. The unit_price variable, for example, aligns with expected price ranges based on product categories. No anomalies were found in the transactional data that would suggest inconsistency in data recording.

c. Transformation Needs: While the data is generally well-structured, smoothing techniques were applied to both daily and weekly sales to reduce noise and improve trend visibility. No further transformations were required at this stage, although future analyses might benefit from applying logarithmic transformations to stabilize variance if the data exhibits skewness.

d. Outliers and Extreme Values: Outliers in the sales data were observed, especially in daily sales. These outliers likely correspond to specific high-demand days or events. While they were not removed, smoothing techniques were applied to minimize their impact on the analysis. Further analysis could consider the removal or adjustment of extreme values based on business knowledge or statistical methods.

5. Model Selection: In order to forecast the coffee shop's weekly sales, we explored three different ARIMA (AutoRegressive Integrated Moving Average) models. These models were chosen based on the assumption that sales trends could be influenced by previous sales data and potential forecasting errors. We tested the following models:

- **ARIMA(0,1,0)**: The simplest model, using no autoregressive or moving average terms but including a first-order differencing (I(1)) to make the data stationary.
- **ARIMA(1,1,1)**: A more complex model that includes one autoregressive (AR(1)) term and one moving average (MA(1)) term, along with the first-order differencing (I(1)).
- **ARIMA(1,1,2)**: A further refinement that includes one autoregressive (AR(1)) term and two moving average (MA(2)) terms, along with first-order differencing (I(1)).

6. Rational from Model Selection

- **ARIMA(0,1,0)** was chosen as a baseline model to serve as the simplest possible model.

- **ARIMA(1,1,1)** was tested next to capture potential relationships between previous sales (AR term) and forecast errors (MA term).

- **ARIMA(1,1,2)** was explored to capture additional complexities in the error structure by introducing a second moving average term (MA(2)).

7. Model comparison

After fitting the models, we compared their performance using several key metrics:

- **AIC (Akaike Information Criterion)**: A lower AIC indicates a better fit.
- **RMSE (Root Mean Squared Error)**: Lower RMSE suggests better model accuracy.
- **MAE (Mean Absolute Error)**: Smaller MAE values indicate more accurate forecasts.
- **MAPE (Mean Absolute Percentage Error)**: This metric provides a measure of the model's forecast accuracy in percentage terms.

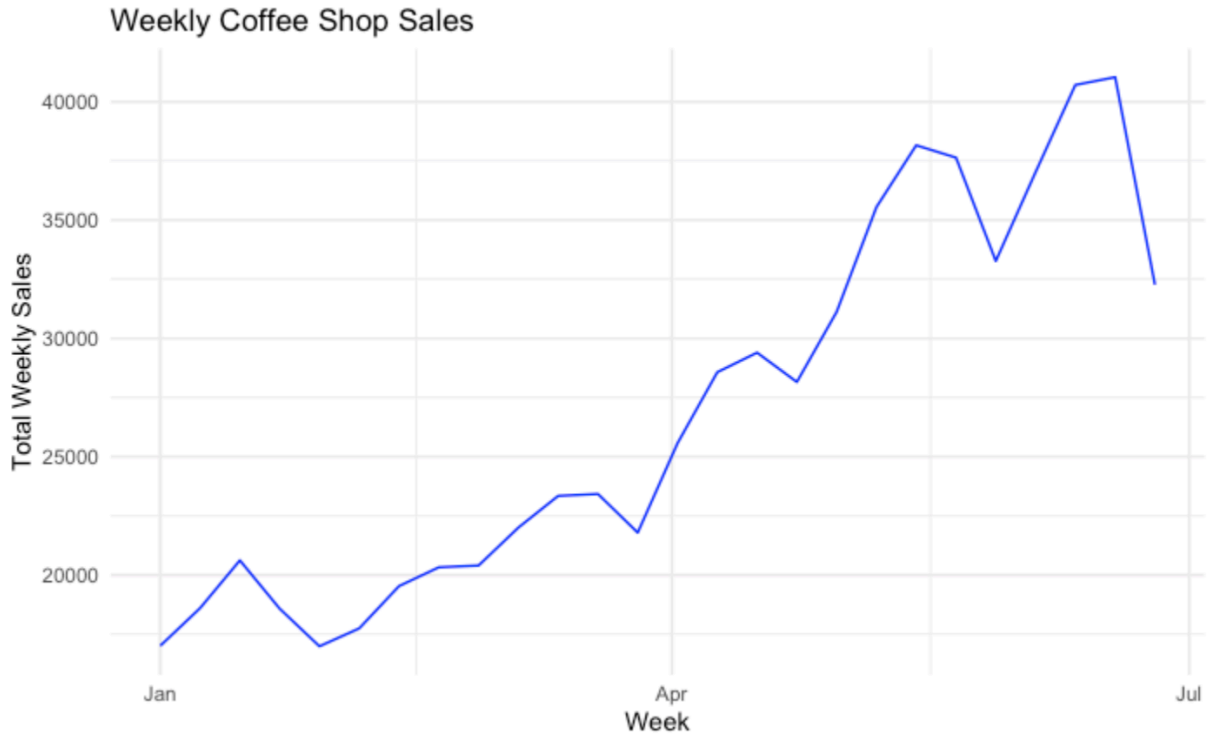
Model	AIC	RMSE	MAE	MAPE
ARIMA(0,1,0)	471.48	2838.57	2138.70	7.71%
ARIMA(1,1,1)	473	2691.23	1870.01	6.73%
ARIMA(1,1,2)	474.14	2645.35	1917.03	6.86%

- **ARIMA(1,1,1)** provides the **best model** based on **MAPE** (6.73%) and shows improvements in **RMSE** and **MAE** compared to **ARIMA(0,1,0)**.

- **ARIMA(1,1,2)**, though slightly higher in **MAPE**, provides the lowest **RMSE** and **MAE**, making it a strong contender for capturing the underlying trends.

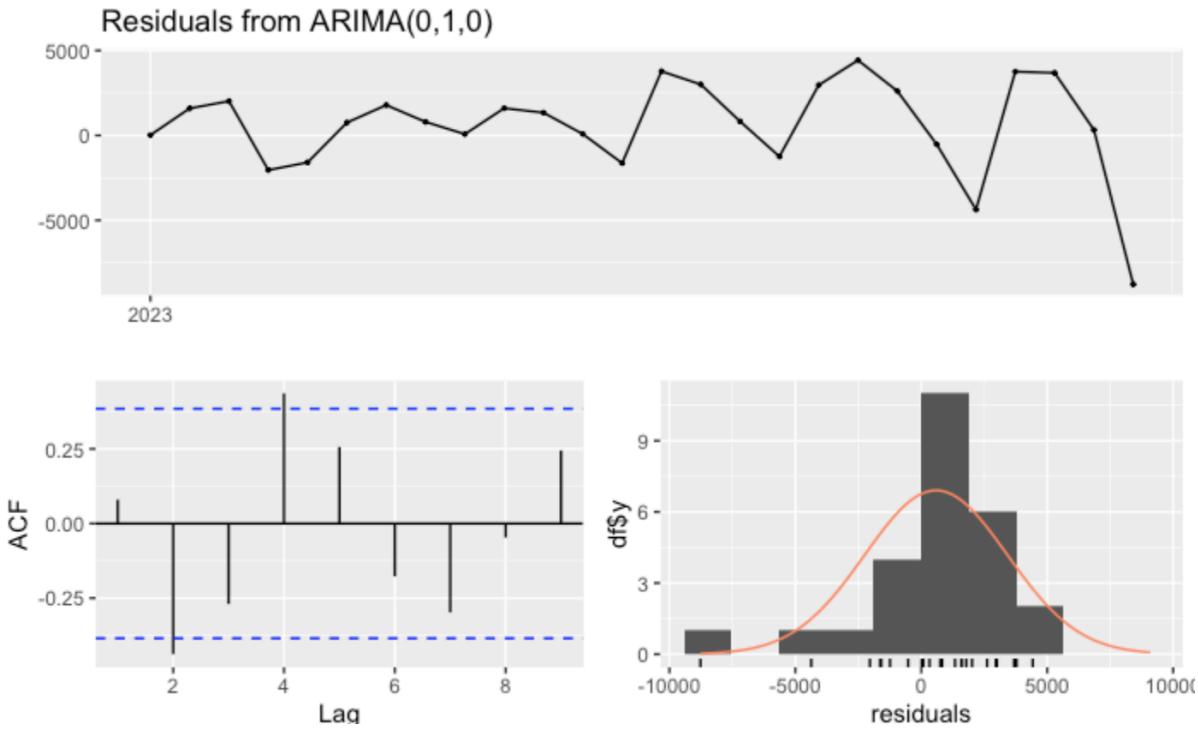
8. Conclusion: After comparing the three models, ARIMA(1,1,1) was selected for forecasting due to its balance of accuracy in both relative (MAPE) and absolute (RMSE, MAE) error metrics. This model offers a reliable prediction of future weekly sales, and its simplicity allows for ease of interpretation while still accounting for the essential temporal dependencies in the data.

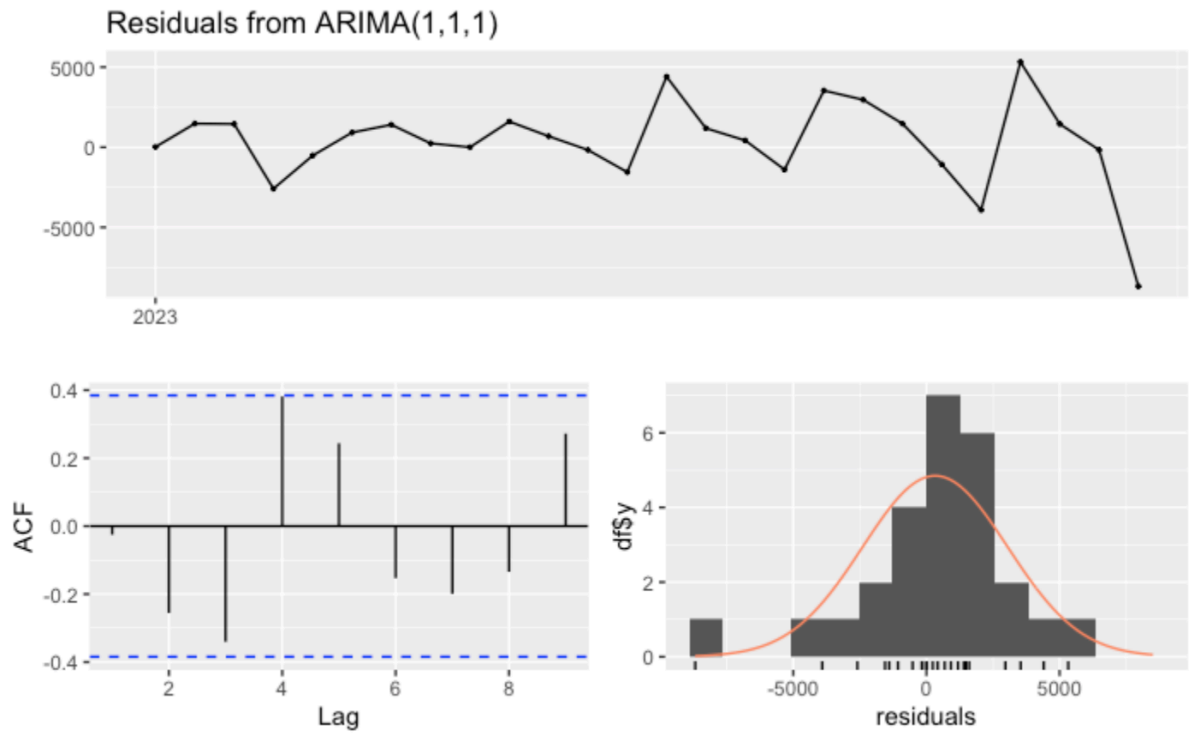

Daily Coffee Shop Sales
Total Daily Sales
6000
5000
4000
3000
2000
Jan
Apr
Jul

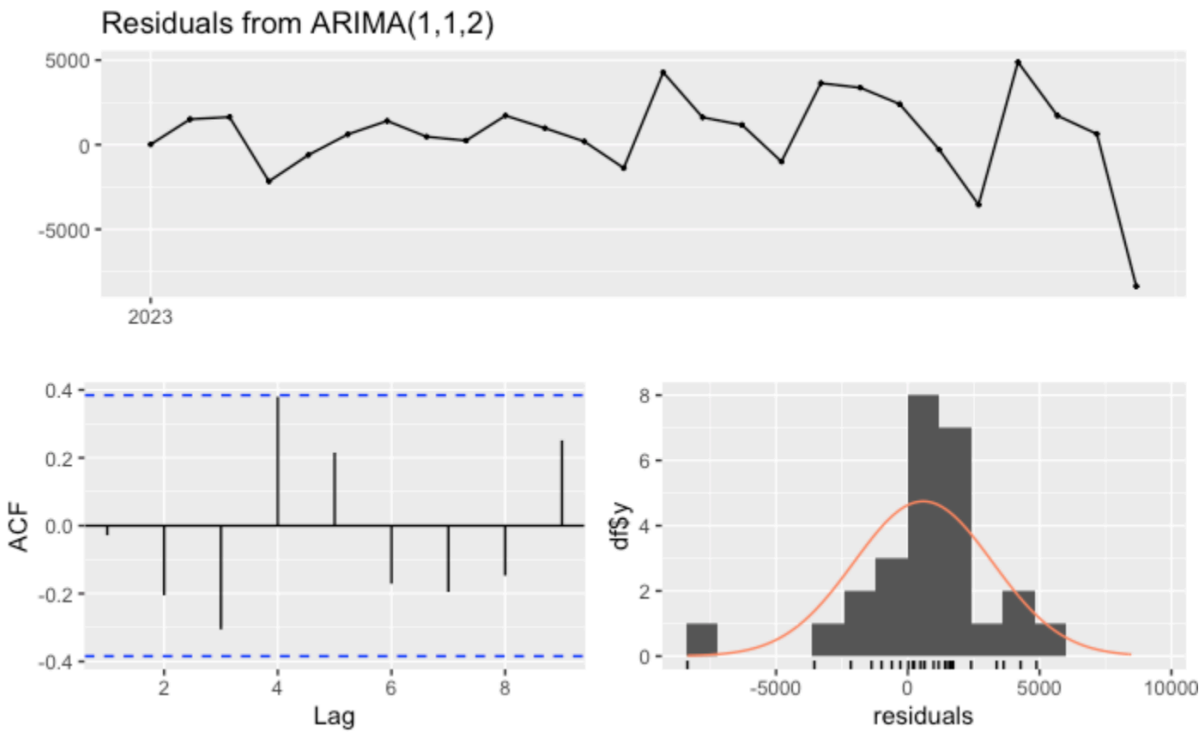

Weekly Coffee Shop Sales
Total Weekly Sales
40000
35000
30000
25000
20000
Jan
Apr
Jul
Week


Daily Coffee Shop Sales with 7-Day Moving Average
Total Daily Sales
6,000
5,000
4,000
3,000
2,000
Jan
Apr
Jul
Date


Weekly Coffee Shop Sales with 3-Week Moving Average
Total Weekly Sales
40,000
35,000
30,000
25,000
20,000
Jan
Apr
Jul
Week


Residuals from ARIMA(0,1,0)
5000
0
-5000
2023
ACF
0.25
0.00
-0.25
2
4
6
8
Lag
df$y
9
6
3
0
-10000
-5000
0
5000
residuals


Residuals from ARIMA(1,1,1)
5000
0
-5000
2023
ACF
0.4
0.2
0.0
-0.2
-0.4
2
4
6
8
Lag
df$y
6
4
2
0
-5000
0
5000
residuals


Residuals from ARIMA(1,1,2)
5000
0
-5000
2023
0.4
0.2
0.0
-0.2
-0.4
ACF
2
4
6
8
Lag
8
6
4
2
0
df$y
-5000
0
5000
10000
residuals