**Executive Summary**

This report analyses bakery sales data alongside weather data from January 2021 to September 2022 to help the bakery reduce waste and optimise product availability. Key findings indicate that sales peak during summer months and weekends, with temperature being a significant factor influencing sales. Recommendations include product bundling and themed special menus.

**Data Overview**

The data presented several challenges, necessitating specific assumptions for pre-processing to ensure accuracy.

**Assumptions**

**Bakery df**

1. **Unit Price Correction**:
   * Zero unit prices were assumed incorrect and replaced using the mode for each 'Article'.
   * Articles without a mode price were dropped from the dataset.
2. **Negative Quantities**:
   * Negative quantities were treated as returns and removed from the dataset.
3. **Large Quantities**:
   * Quantities greater than 100 were considered erroneous and removed.
4. **Duplicate Rows**:
   * All 1,210 duplicate rows were dropped to maintain data integrity.

**Weather df**

1. **Missing Data**:
   * Columns with predominantly missing values, such as 'snow' and 'tsun', were dropped.
   * Missing data in other columns were handled using linear interpolation.

**Analysis**

Focusing on 'Traditional Baguette', the highest-selling and most profitable product, the analysis revealed:

1. **Seasonal Trends**:
   * Sales peak in summer and decline in winter, indicating higher demand during warmer months.
   * Weekly patterns show higher sales on weekends, especially Sundays.
   * Monthly variations indicate the highest sales in August and the lowest in January and February.
2. **Weather Impact**:
   * Warmer temperatures positively correlate with higher sales.
   * Precipitation and wind have minimal impact.
3. **General Trends**:
   * Significant variability with recurring peaks suggests that sales are influenced by holidays, weekends, and seasonal preferences.

A graph showing the weather

Description automatically generated

*Fig. 1: Monthly sales trends with summer months highlighted to emphasise peak sales periods.*

A graph showing the temperature of a temperature

Description automatically generated with medium confidence

*Fig. 2: Relationship between average temperature and quantity sold, with a trend line indicating the overall trend.*

A graph of different colored lines

Description automatically generated

*Fig. 3: Compares the quantity sold on weekends versus weekdays, with trend lines showing overall patterns for each.*

**Model**

To address missing dates (approximately 37), a complete dataset was created with forward-filled missing quantities. Seasonality was confirmed through decomposition analysis, and stationarity was achieved via differencing, using ADF test. Feature engineering involved removing redundant columns such as 'Revenue', 'WindGust', 'Pressure', and 'Product'. Numeric data was scaled, and categorical variables were encoded.

* **Linear Regression** and **Random Forest** models were employed to forecast 'Quantity Sold'.
* Random Forest outperformed Linear Regression.

**Limitations**

1. **Data Quality**:
   * Challenges such as negative values and incorrectly inputted 'Article' names affect the reliability of the data.
   * Missing data on specific dates and holidays limits the completeness.
2. **Assumptions Impact**:
   * Assumptions made during data cleaning could introduce bias if they do not hold.
3. **External Factors**:
   * The analysis did not account for other external factors such as promotions, or local events.

**Conclusion**

The insights gained from the analysis can help in making data-driven decisions.

**Recommendations**

* Implement bundled product promotions once a month, pairing best-selling items with least-selling ones.
* Offer a special menu with a holiday theme.

**Future Work**

* Incorporate additional sources, such as local events.