Project G6: Kaggle - Coffee Shop Sales Analytics

Team Member: Nikita Jakovlev
The link of the repository:
https://github.com/DaddyIPanda/Sales---Sissejuhatus-andmeteadusess/tree/main

1.Business understanding

This report outlines the business understanding for my project, which focuses on analysing coffee shop sales data. The data for this project is sourced from publicly available datasets on Kaggle, specifically the "Maven Roasters" and "Coffee Shop Sales" files. The main beneficiary of this analysis is a coffee shop manager or owner who needs to make smarter day-to-day decisions.

The main business problem this project addresses is the balance between saving money and keeping customers happy. A coffee shop loses money if it has too much unsold food that gets thrown away or if it schedules too many staff members during quiet periods. On the other hand, customers become unhappy if their favourite items are often out of stock or if they have to wait in long lines because there aren't enough staff during busy times. Currently, many of these decisions are based on guesswork. My project aims to replace that guesswork with insights from the available sales data.

The primary business goals are to increase profitability by reducing waste and optimising labour costs, and to enhance customer satisfaction by ensuring product availability and efficient service. Success would signify a measurable reduction in food waste and enhanced customer service during peak hours for the business. I have delineated two primary objectives for data analysis to ensure effective outcomes. The first is to conduct a comprehensive exploration of the sales data, which involves delving into the numbers to identify clear patterns, determine the most popular products, ascertain which of the three shop locations experiences the highest foot traffic, and pinpoint the times of day or days of the week that attract the most customers. These insights directly help a manager know what to order and when to schedule more staff.

The second goal is to build a forecasting model. This model will predict how much the coffee shop will sell each day. Having a reliable daily sales forecast is incredibly useful for planning. It informs the manager about the quantities of coffee, milk, and pastries to order for the upcoming week, providing a solid foundation for creating the staff schedule to ensure adequate staffing on predicted busy days.

The success of my data work will be measured in two ways. For data exploration, success means I can provide clear, easy-to-understand answers and charts that directly address the manager's questions about sales patterns. For the forecasting model, I will measure its accuracy. A good target is for the model's daily predictions to be, on average, within 15 per cent of the actual sales. This level of accuracy is considered useful for practical business planning.

For this project, I will leverage sales data files from Kaggle and utilise Python programming for my analysis. It is important to highlight that this is an independent endeavour, which allows me to maintain full control over the scope and direction of the work. I will concentrate on the key objectives of identifying sales patterns and developing a robust forecast, without getting sidetracked by more complex analyses that could require excessive time or resources. I am confident that this focused approach will lead to meaningful insights.

There are some potential challenges. The data might be messy and require significant cleaning before any analysis can begin. The forecasting model might not be perfectly accurate. To manage these risks, the initial phase of my work will be dedicated to preparing the data, and I will test a few different simple forecasting methods to find the one that works best with the data I have.

In terms of costs, the only investment in this project is the time spent on the analysis. The potential benefit for a business, however, is significant. The insights gained could lead to direct cost savings from reduced waste and more efficient staff scheduling, and even help increase sales by highlighting opportunities to sell more of the most profitable items. Ultimately, my project aims to provide a data-driven foundation for running a coffee shop more effectively.

2.Data understanding

This report gives an overview of the insights from the datasets used in the coffee shop sales project. I obtained two datasets from Kaggle about coffee shop sales. After reviewing both, I found that the second dataset is better for my analysis due to its detailed information, which will improve my findings.

The data is in a CSV file format, making it easy to work with and analyse using Python. The dataset includes 149,116 individual sales from a coffee shop chain with three locations. Each row represents a single sold item. For example, if a customer buys both a coffee and a muffin, this will appear as two separate entries. The dataset contains the following key information:

- Transaction ID: A unique number for each customer purchase, similar to a receipt number.
- Date and Time: Each transaction has a specific date and time, which helps analyse busy times.
- Store ID: This shows which of the three shops made the sale.
- Product Information: This includes a product ID, general category, specific type, a detailed description, unit price, quantity sold, and total amount spent for each item.

I found some notable trends during my initial analysis. Saturdays and Sundays are the busiest days, with Fridays also having significant sales. The daily sales pattern shows two peak times: one from 7 AM to 9 AM and another from 12 PM to 2 PM during lunch. One shop location consistently reports higher sales than the others, but all locations follow similar weekly trends. Coffee drinks are the most popular items, making up over half of total sales, while bakery products, such as pastries and sandwiches, rank second.

Examining sales data from January to July 2023 reveals a generally stable business environment, with no major increases or decreases in sales. However, some spikes in sales occur during holidays or special events.

The data quality is generally high, but some inconsistencies exist. There are no complete duplicate records, but some product categories have variations in spelling or formatting that require standardisation. All transaction times are within normal business hours and appear credible. The prices for identical products are consistent in all transactions, and the total price calculations are correct. Some transactions show unusually high quantities, like customers buying ten of the same item. These cases might relate to office meetings or be errors; further investigation is needed in later project phases.

Overall, this dataset offers valuable information to analyse sales patterns and predict future sales. As I work on this project independently, I will focus on the most important aspects: transaction dates and times, product details, and sales amounts. The quality issues I identified can be resolved during the data preparation phase. The seven months of data collected should be enough to create a reliable forecasting model and identify important patterns to help shop management make informed decisions.

3.Planning of project

3.1. Plann

- Data Preparation and Cleaning (10 hours): I will load the dataset and clean it. It includes handling any missing values, fixing inconsistencies in product names or categories, and ensuring the date and time formats are correct for analysis.
- Exploratory Data Analysis (EDA) - Part 1 (8 hours): I will perform an initial exploration to answer foundational business questions. It involves creating summaries and visualisations to identify top-selling products, compare performance across the three store locations, and analyse overall sales trends.
- Exploratory Data Analysis (EDA) - Part 2 (7 hours): I will dive deeper into the time-based patterns. The focus will be on creating charts to understand daily sales patterns and weekly seasonality, which are crucial for staffing and inventory decisions.

- Time-Series Forecasting (10 hours): I will develop a model to predict future daily sales. This task involves preparing the data for modelling, testing different forecasting methods, and selecting the best-performing model based on its error rate (MAPE).
- Final Reporting and Presentation (5 hours): I will compile all findings, visualisations, and model results into a final report and presentation. It will summarise the key insights for a business owner and present the forecasting model.

Total Estimated Hours: 40 hours

3.2. Methods and Tools

Programming Language: Python.

Key Libraries: Use Pandas to handle data, Matplotlib and Seaborn for creating visualisations, and Scikit-learn or Statsmodels/Prophet for building forecasting models.

For managing the code, use GitHub to save and organise everything.