**Sertac Balbaros**
alagozse@gmail.com

## Data Science Project

# Coffee Sales Data Analysis Conceptual Design Report

**5 October 2025**

## Abstract

Coffee shops often don’t know what customers want or how to plan staff and products. Sales change a lot during the day, the week or the season. In this project, I studied more than 3,500 coffee sales. The data have product, amount, payment and time. I want to find sales patterns and predict future demand.

I will use EDA and simple plots to see trends. I will then use **statistical tests** like chi-squared and correlation to uncover relationships between variables. Time series analysis can show daily and weekly sales.

I expect to see busy hours, popular coffees and the effect of the payment method. This can help with stock, staffing and marketing. Even if the data is from one shop, the results can also help other small shops.

## Table of Contents

**Abstract** 1
**Table of Contents** 2
**1 Project Objectives** 3
**2 Methods** 3
**3 Data** 4
**4 Metadata** 6
**5 Data Quality** 6
**6 Data Flow** 6
**7 Data Model** 7
**8 Documentation** 7
**9 Risks** 7
**10 Conclusions** 8
**Acknowledgements** 8
**Statement** 8

## 1 Project Objectives

Small coffee shops often cannot plan well. Sales change with time of day, weekday and season. If shop owners do not understand customer demand, they may buy too little or too many products, or wrong staff planning. Coffee sales are more sensitive to habits and busy hours than other shops. Because of this, data can help operations run smoothly.
In this project, we study more than 3,500 sales. The data includes product, amount, payment and time. Our goal is to see sales patterns, like the most popular product and busy hours. We also want to check how payment methods relate to customer behavior. If they work, results can help with stock, staffing and business planning.

## 2 Methods

The data for this project comes from Kaggle[1]. I use Google Collab because it is free and easy to use online. I write the code in Python. For software libraries, I plan to use pandas and Numpy for data cleaning and preparation, and matplotlib and seaborn for visualization.

For statistics I will start with descriptive analysis, for example mean, median and frequency. I will also use correlation to check links between variables, like sales and payment methods.

I will also use time series methods to check daily and weekly sales patterns. For daily analysis I will group the data by date and plot total sales. For weekly analysis I will resample the data by week and study the trend. I can also use moving averages to see the trend more clearly. This will help to understand busy days and weeks.

I also plan to use hypothesis testing with Python's Statsmodels package. For example, I can run statistical tests to see if average spending is different between cash and card payments, or if sales are different between weekdays. ANOVA can be used to compare sales between product groups or time periods. This helps to know if differences are real or just random. I do not know all the methods, but during the CAS I expect to learn more and maybe use additional tools.

## 3 Data

The dataset for this project comes from Kaggle. This is open data and free to use. It contains more than 3,500 transactions from a coffee shop. Each row is one sale. The main columns are product name, amount, payment method, time of day, weekday, month and datetime.
I will show some simple plots.

[1]Coffee Sales : https://www.kaggle.com/datasets/navjotkaushal/coffee-sales-dataset

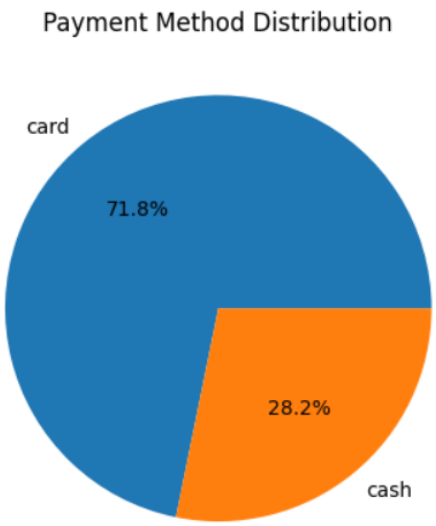



*Figure 1: Pie Chart of Payment Method Distribution*

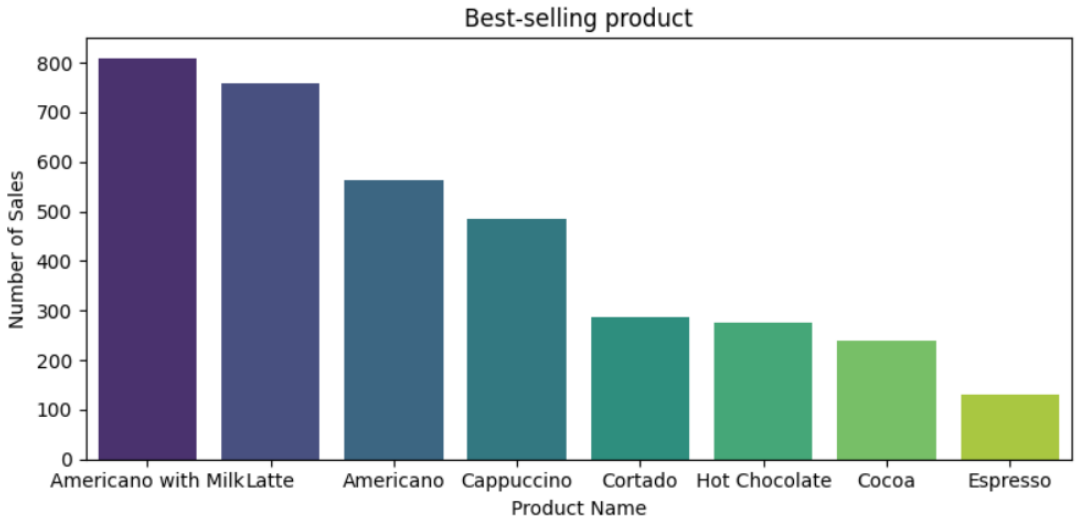



*Figure 2: Histogram for the Best-Selling Products Distribution*

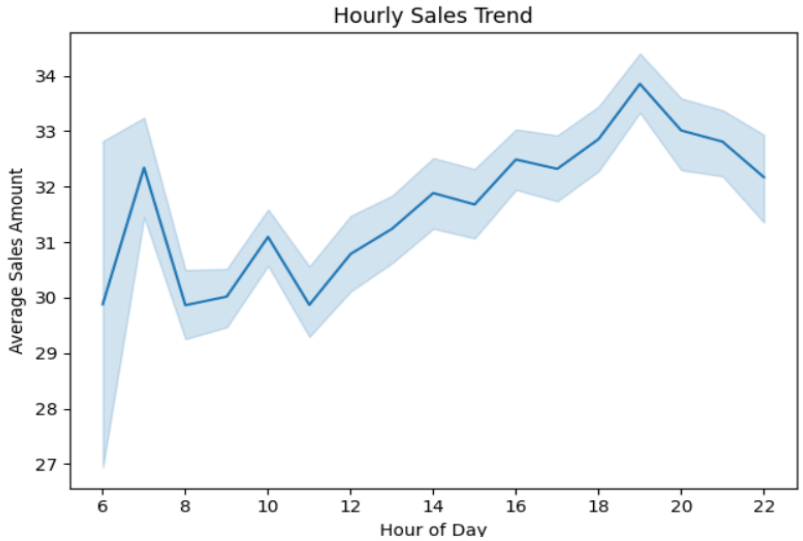



***Figure 3: Line Char for Hourly Sales Trend***

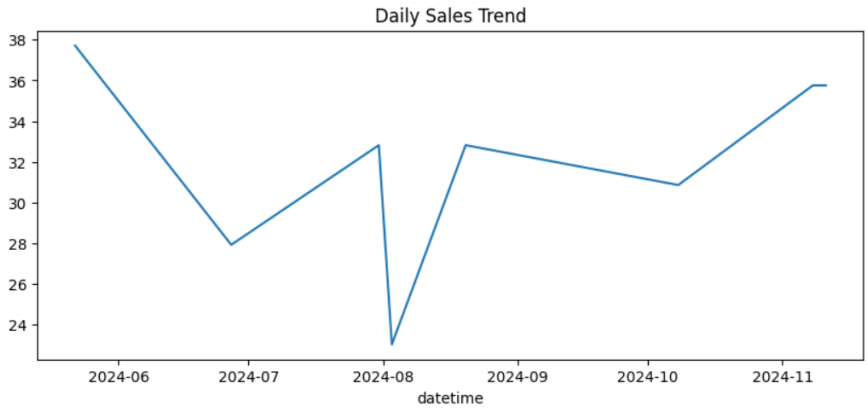



***Figure 4: Line Chart for Daily Sales Trend over Months***

On a simple note, there are no big security issues, because the data has no personal information. It is only transaction data from one coffee shop. The data is safe to use for analysis.

## 4 Metadata

To repeat the analysis, we need metadata like the column names, data types and a short description of the Coffee Sales dataset.

The dataset has details of Coffee Shop sales, with columns such as Date Time, Product, Quantity, Payment Method, Price and Total Sale. Each row shows one sale, what was bought, how it was paid, and the total amount.

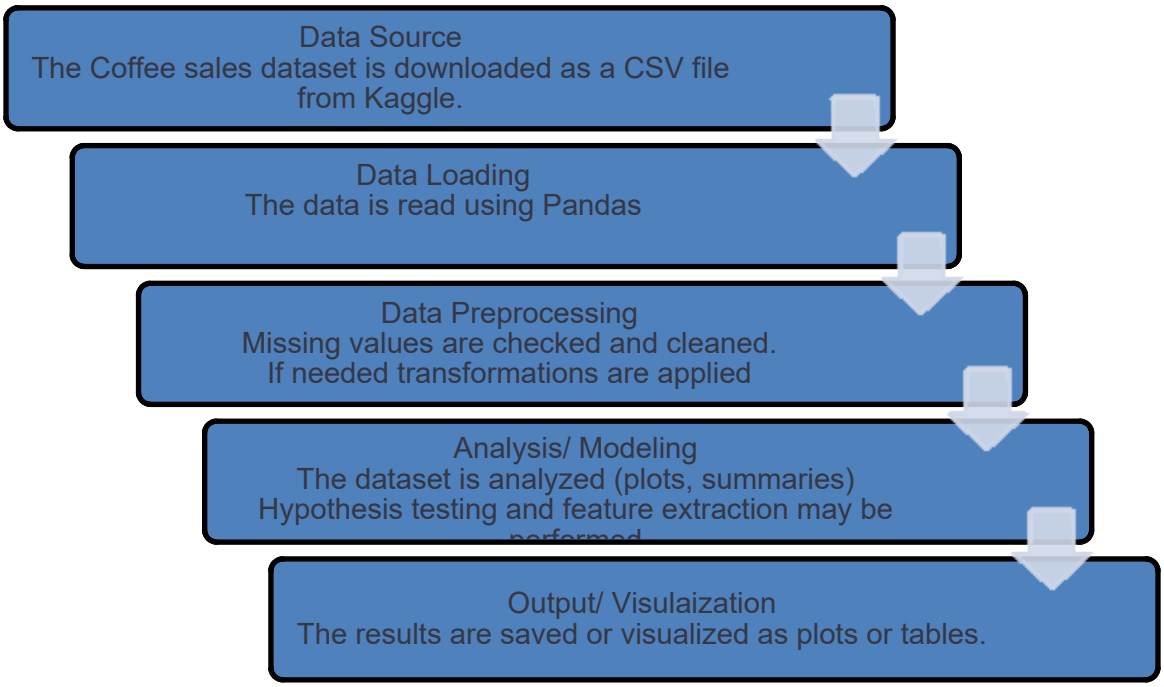
The data can be used to study sales trends, customer choices, and payment methods and it covers the years 2024 and 2025.

The metadata for the Coffee Sales dataset, such as column names, data types and description, is stored together with the dataset in the Kaggle repository. It is publicly available, and anyone can access it directly from the Kaggle dataset page.

## 5 Data Quality

I checked the dataset and found no duplicated nor missing values. The data size is reasonable for this analysis, with 11 columns and more than 3,500 rows. The data types are consistent, and values appear to be reliable. Since the dataset is clean, we do not expect a significant negative impact on our results. To further improve quality, we could validate the data against the source, monitor for outliers, an ensure regular updates,

## 6 Data Flow




## 7 Data Model

For this work, we define the specific columns/features to be used for analysis and modeling:

- Sales Price → coffee_price (float)
- Merchandise Type → beverage_type (categorical)

Timestamp (derived features): From the original timestamp, we generate several time-related variables:

- Time of Day → time_of_day (categorical)
- Day of Week → week_day (categorical)
- Month of Year → month (categorical)
- Weekday → weekday (categorical)
- Date → date (datetime)
- Time → time (datetime)

## 8 Documentation

This PDF file has been prepared to document the analysis, using the insights obtained from the Google Collab Notebook.

## 9 Risks

Possible issues include missing or incorrectly entered data. The dataset represents only one shop and does not have enough data for reliable time series forecasting. A check was made for missing or duplicate records, but no such problems were found. Therefore, the data cleaning process was simple, and the dataset is ready for analysis.

## 10 Conclusions

In this work, I used the coffee sales dataset and analyzed the data. I created plots to show the main patterns. In addition, I performed a hypothesis test. Using the Chi-Square test, I found that the distribution of products is related to the day of week. This means that product sales differ depending on the weekday.

## Acknowledgements

## Statement

The following part is mandatory and must be signed by the author or authors.

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.“

Date: 05.10.2025 Signature(s):