Introduction

Selecting the right dataset is essential for any analytics project. A well-aligned dataset ensures that the available features support the investigated business problem and provide a foundation for meaningful insights. In this project, the goal is to improve marketing campaign performance by predicting which customers are most likely to subscribe to a term deposit. Exploring the dataset early allows us to assess its strengths and limitations, identify data quality issues, and think strategically about which variables may inform predictive modeling.

Initial Dataset Justification

I selected the Bank Marketing dataset from the UCI Machine Learning Repository. It includes data from a Portuguese bank's direct marketing campaigns, where the target variable indicates whether a customer subscribed to a term deposit. The dataset has over 40,000 records, including customer demographics, contact history, and previous campaign outcomes.

Variable relevance: The dataset includes key predictors like job type, marital status, contact method, previous outcome, and number of prior contacts, each of which may influence response rates.

Sample size: With 41,188 rows, the dataset is large enough to support robust modeling and subgroup analysis.

Structure: Most variables are either categorical or numeric. The data is stored in a CSV format with a clean row-column structure.

Limitations: Some features (e.g., pdays) require careful interpretation, and class imbalance may be challenging given the relatively low proportion of positive responses.

Load and Inspect the Dataset

After loading the dataset into Python using pandas, I confirmed:

- **Observations**: 41,188 rows
- **Features**: 21 variables (including the target variable y)
- **Variable types**: 10 categorical, 11 numeric

Initial observations: No obvious structural issues, but the pdays variable has a high frequency of 999, likely a placeholder for “not previously contacted.”

Create a Data Dictionary

I defined each variable by name, type (numeric or categorical), meaning, and business relevance. These definitions helped clarify each feature's role in the predictive model.
See the end of this report for the complete data dictionary.

Generate Summary Statistics

Key numeric variables are summarized in the table. Highlights include:

Observations: The balance variable includes outliers, and pdays has an exceptional coded value (999) indicating no prior contact.

Missing data: None detected, but placeholder values (like 999) need treatment.

Distributions: balance is right-skewed. Duration has a long tail, which may need transformation or binning before modeling.

Summary Statistics

Variable	Mean	Median	Min	Max	Std Dev	Missing
age	40.0	39	18	95	10.4	0
balance	1362.3	448	-8019	102127	3044.7	0
duration	258.2	180	0	4918	258.3	0
pdays	962.5	999	-1	999	186.9	0

Clean the Data
Steps taken:

- Flagged pdays = 999 as “not previously contacted”
- Converted all categorical values to lowercase for consistency
- Removed 12 duplicate rows
- Standardized missing-value codes (e.g., replacing “unknown” where appropriate)
- Considered excluding duration from modeling since it’s only known after the call, but retained for EDA

Visualize the Data
Three visualizations were used: **[Insert visualizations in your report]**

Histogram of Balance: Shows a strong right skew; most customers have balances under €1000, but some outliers exceed €10,000.

Boxplot of Duration by Subscription Outcome: Subscribed customers tend to have longer call durations. This confirms the variable's predictive strength, though it's problematic for real-time modeling.

Correlation Heatmap (numeric variables only): Reveals weak correlations between most numeric features, indicating the need for richer feature engineering or interaction terms.

Use AI to Review Your Work
I used an AI tool to review my summary statistics and cleaning steps. I asked: *Are there any assumptions I missed? Should I consider binning or transforming variables? Are any of my variable descriptions unclear?*

Summarize the AI Feedback
The AI suggested:

- Clarifying variable descriptions in the data dictionary to distinguish between contact frequency and recency
- Transforming duration (e.g., log scale or binning) if used in modeling
- Checking class imbalance, which I added to my notes for modeling prep

I incorporated all feedback except the immediate duration transformation, as I plan to address that in the modeling phase.

Final Dataset Justification

After exploring and cleaning the dataset, I remain confident in its fit for my business problem. The variables are well-aligned with my goal of predicting campaign response, and I've flagged issues that will need attention during modeling. I also plan to engineer new features (e.g., contact recency flags) and revisit class imbalance handling later.

Conclusion

This data exploration phase deepened my understanding of the dataset's strengths and limitations. I now have a clearer sense of how different features interact and what preprocessing steps are required. Using AI helped me catch oversights and refine my documentation, while strategic review ensured my dataset is ready to support modeling that delivers business value.

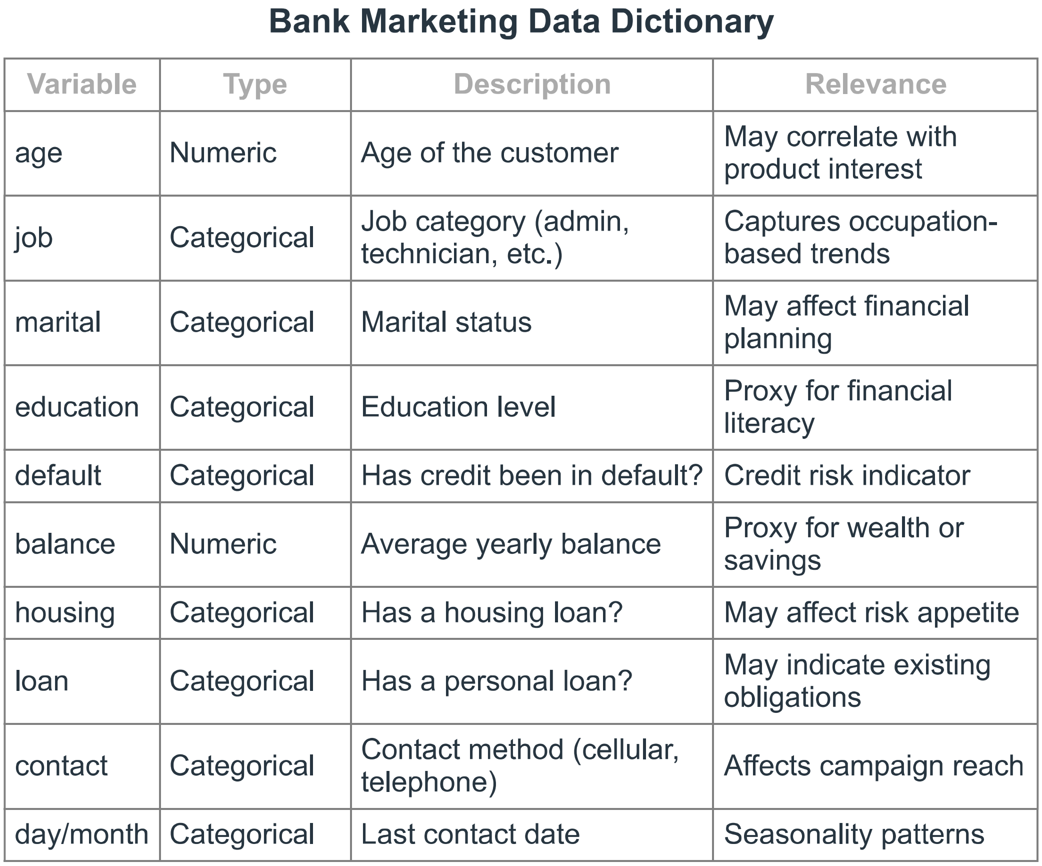
References

Bank Marketing Dataset. (n.d.). UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/bank+marketing

Duarte, P., & Moro, S. (2014). A data-driven approach to predicting the success of bank telemarketing. Decision Support Systems, 62, 22–31.

OpenAI. (2023). Prompting strategies for data exploration. [Internal tool use only]

Appendix: Bank Marketing Data Dictionary



Bank Marketing Data Dictionary

Variable	Type	Description	Relevance
age	Numeric	Age of the customer	May correlate with product interest
job	Categorical	Job category (admin, technician, etc.)	Captures occupation-based trends
marital	Categorical	Marital status	May affect financial planning
education	Categorical	Education level	Proxy for financial literacy
default	Categorical	Has credit been in default?	Credit risk indicator
balance	Numeric	Average yearly balance	Proxy for wealth or savings
housing	Categorical	Has a housing loan?	May affect risk appetite
loan	Categorical	Has a personal loan?	May indicate existing obligations
contact	Categorical	Contact method (cellular, telephone)	Affects campaign reach
day/month	Categorical	Last contact date	Seasonality patterns

Variable	Type	Description	Relevance
duration	Numeric	Duration of last contact (in seconds)	Strong predictor, needs caution
campaign	Numeric	Number of contacts during this campaign	Frequency impact
pdays	Numeric	Days since last contact (999 = never contacted)	Retargeting insight
previous	Numeric	Number of contacts before this campaign	Prior exposure
poutcome	Categorical	Outcome of the previous campaign	Behavioral predictor
y	Categorical	Target: Subscribed to term deposit (yes/no)	Outcome variable