Types of Data

Contents

Preamble

In the world of data science and statistics, data is a critical asset that supports informed decision-making, pattern recognition, and forecasting. Across various industries, data is collected and analyzed to draw insights that help solve complex problems. Data can be broadly categorized into several types, each with distinct characteristics and applications. However, the categorization of data is not uniform across literature and online resources (see references [1], [2], and [3]). Here, we present a categorization based on [3].

- **Structured datasets**: Organized data that is easy to query and analyze. Examples include patient records in a hospital or a client list for a business.
- **Unstructured datasets**: Unorganized data that may not be designed for data science purposes. This can include diverse data types, such as social media posts, emails, or multimedia content.
- **Hybrid datasets**: Datasets that combine both structured and unstructured data. For example, healthcare records can include structured data in the form of patient data, appointment history, and lab results in a structured format. Unstructured data can include physician notes, discharge summaries, and patient feedback that may be written in narrative form.

Choice of Application Domain

We decided to focus on financial data. This choice stems from the critical role of financial data analysis in modern economies and the vast opportunities it offers for applying data science techniques. Financial data is not only abundant but also highly varied, ranging from transaction data and stock market information to time series related to interest rates and economic indicators.

By choosing financial data, we aim to explore real-world challenges such as credit card fraud detection and stock price forecasting, both of which are areas with significant social and economic impact. Financial markets generate large amounts of data, making them ideal for exploring predictive modelling. Furthermore,

financial datasets are readily accessible from numerous reputable sources, enabling us to experiment with various machine learning techniques and tools to uncover valuable insights and trends.

By focusing on financial data, we seek to enhance our understanding of data science while working in an area of personal interest to the group members.

Introduction to Financial Data

Financial data used to analyze trends, behaviors, and patterns within financial markets, institutions, and transactions. In data science and statistics, the primary types of financial data include:

- **Transaction Data**: Detailed records of individual financial transactions, such as credit card purchases, which are often analyzed to detect fraud.
- **Time Series Data**: Data collected at regular intervals over time, such as stock prices, exchange rates, or economic indicators. Time series analysis is key to forecasting trends in financial markets.
- **Customer Data**: Information on customer demographics, purchasing behaviors, and financial habits, often used in personalized marketing or credit scoring.
- **Financial Statements**: Data extracted from balance sheets, income statements, and cash flow statements of companies, providing insight into financial health and performance.
- **Market Data**: Real-time or historical data on stock prices, bond yields, commodities, and other financial instruments, often used for trading strategies and risk management.

Sources of Financial Data

Several repositories and databases offer access to a wealth of financial datasets for analysis, including:

- **Kaggle**: A popular platform for data science competitions that hosts a wide range of financial datasets, including those related to stock market predictions, credit card fraud detection, and economic indicators.
- **UCI Machine Learning Repository**: A repository of well-curated datasets, which includes financial data used in classification and prediction tasks, such as loan approval and credit risk assessments.
- **Monash Time Series Forecasting Repository**: Specializing in time series data, this resource provides datasets for financial time series analysis, making it useful for predicting market trends and prices.
- **Google Dataset Search**: A powerful tool for finding datasets across various domains, including finance. This search engine aggregates datasets from multiple sources, including academic institutions, government agencies, and commercial platforms.
- **Yahoo Finance**: Offers real-time and historical data on stocks, bonds, and commodities, as well as financial news and economic indicators, commonly used in market analysis.
- **Quandl**: A data platform that provides access to financial, economic, and alternative datasets. Quandl is often used for market research, investment modeling, and forecasting.

In this project, financial data serves as the foundation for understanding and addressing real-world problems in the financial industry. For example, credit card fraud detection leverages transaction data to identify anomalous behavior, while financial time series analysis focuses on predicting stock prices or interest rate movements. By utilizing diverse datasets from reliable sources, the project will enable a deeper exploration of financial phenomena, offering insights into predictive modeling, fraud detection, and market behavior.

Access to varied and rich financial datasets ensures the ability to perform comprehensive analysis and create meaningful visualizations, which are crucial for developing data-driven insights in the finance sector.

References

1. Blog on Dataset Types: Sprinkledata: This allowed us to get an overview of the types of datasets.
2. Office for National Statistics: This source defines datasets in a different way to what we have in this text.
3. Blog on Dataset Types: Brightdata: This helped us get an overview of different types of datasets, and explore the consistency or variation in the categorisation across resources.

The following references are the data sources used throughout the project.

4. Kaggle Datasets
5. UCI Machine Learning Repository
6. Monash Time Series Forecasting Repository
7. Google Dataset Search
8. Yahoo Finance
9. Quandl