Types of Data

Contents

1 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]. This is listed below.

- **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.

In this project, we will deal with structured datasets as they are readily available online and are easier to analyse. Due to time constraints, other types of datasets were not looked at in detail, though the interested raeder may consult the references [1], [2] and [3] for details. Having chosen to look at structured datasets, we next choose the application domain. Data is used in many application domains; we provide a brief and certainly not exhaustive list of some domains, along with a simple example for each domain below.

1. **Healthcare**: Using patient records and doctor notes to optimize treatment plans.
2. **Retail and E-commerce**: Analyzing purchase data and customer reviews for personalized marketing.

3. **Finance**: Combining transaction data and customer feedback for fraud detection.
4. **Social Media Analytics**: Monitoring brand sentiment by analyzing engagement metrics and user-generated content.
5. **Manufacturing**: Utilizing sensor data and maintenance logs to improve operational efficiency.
6. **Telecommunications**: Merging call data records and customer service interactions for churn prediction.
7. **Education**: Combining student performance data and open-ended survey responses to improve learning outcomes.

The list above is adapted from [4]. We choose an application domain and justify our choice in the next section.

2 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 to time series related to stock prices, interest rates and economic indicators. 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 methods of exploratory data analysis to uncover valuable insights and trends. We will also later be able to apply various machine learning techniques and tools on the data.

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. We next explore what data we can analyse in this application domain.

3 Types of Financial Data

Financial data is very rich and varied. It can be used to analyze trends, behaviors, and patterns within financial markets, institutions, and transactions. 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.
- **Customer Data**: Data on customer demographics, purchasing behaviors, and financial habits, often used in personalized marketing.
- **Financial Statements**: Data extracted from balance sheets, income statements, and cash flow statements of companies, providing insight into financial health and performance.

The list above is adapted from [5] and [6]. We now look at sources of financial data to help us make our choice of datasets for analysis.

4 Sources of Financial Data

Several repositories and databases offer access to a wealth of financial datasets for analysis. We give a list below:

- **Kaggle**: A popular platform for data science competitions that hosts a wide range of financial datasets, including those related to credit card fraud detection.
- **UCI Machine Learning Repository**: A repository of datasets which can be used for classification and regression.
- **Monash Time Series Forecasting Repository**: A repository for time series.
- **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. Data can be accessed freely through its API (application interface) – see for example the package `quantmod` in R.
- **Quandl**: A data platform that provides access to financial datasets.

Access to varied and rich financial datasets ensures the ability to perform comprehensive analysis and create meaningful visualizations, which are crucial for developing our data science skills and to uncover insights into the data.

We chose to use Kaggle and the Yahoo Finance API as they provided easy ways to access data. From Kaggle, we chose to analyse the Credit Card Fraud dataset, and from the Yahoo Finance API, we analysed Apple stock data.

5 References

Online resources were helpful in our research. We provide a list of such resources below.

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.
4. **Application Domains**
5. **Types of Financial Data: Egnyte**
6. **Types of Financial Data: Indeed**

Note: These links were accessed from 10:00 to 12:00 on 27/09/2024.

5.1 List of Data Sources

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

1. **Kaggle Datasets**
2. **UCI Machine Learning Repository**
3. **Monash Time Series Forecasting Repository**
4. **Google Dataset Search**
5. **Yahoo Finance**
6. **Quandl**
7. **`quantmod` `R` package**: For fetching stock data from Yahoo Finance.