Introduction: Obtaining and Scrubbing Data

Obtaining data means finding and collecting the data you need for your analysis. Sometimes you might be given data to work with, but other times you might need to find data sources on your own. There are many sources available, some of which are free to use. Scrubbing data refers to the process of cleaning and preparing the data for analysis. This involves removing any errors or inconsistencies in the data, organizing it in a structured format, and making sure it is ready for further analysis. These two steps are important because they lay the foundation for the rest of the data analytics process.

The sources are categorized into three buckets: freely accessible databases, company-specific data, and intentionally collected data. Freely accessible databases are provided by government institutions and organizations and contain a wealth of information on various topics. Company-specific data is specific to a particular company and may include sales data or website analytics. Intentionally collected data is obtained through methods like questionnaires or observations. It's important to explore these different data sources to obtain the necessary data for data analytics projects.

Where to look for data

1. Freely accessible, open-source databases
2. Data specific to your company
3. Data you intentionally collect

COMMON DATA FORMATS

Numeric Data

Data that is expressed in numerical form and that can be measured or counted

Numeric data, also known as quantitative data, refers to data that is expressed in numerical form and can be measured or counted. It represents quantities or values that can be used for mathematical calculations and analysis. Numeric data can include various types of information, such as age, income, sales figures, stock prices, and more. It is typically stored in a tabular or spreadsheet format, often saved as a .CSV file, where values are organized in rows and columns. Numeric data is commonly used in data analytics to draw conclusions, make predictions, and perform statistical analysis.

Text Data

Data that is in written or textual form

Text data, also known as unstructured data, refers to data that is in written or textual form. It consists of letters, words, sentences, and paragraphs. Text data can come from various sources, such as social media posts, emails, tweets, blog posts, customer reviews, and more. Unlike numeric data, which can be easily measured or counted, text data is more complex and requires different techniques for analysis.Text data is often analyzed using natural language processing (NLP) techniques to extract insights and patterns from the data.

Visual Data

Data presented in a visual format, such as images and videos

Visual data, also known as visual content, refers to data that is presented in a visual format, such as images, videos, graphs, charts, maps, and diagrams. It represents information that can be perceived by the human visual system. Visual data is used to convey complex information in a more intuitive and easily understandable way.

Examples of visual data include photographs, drawings, videos, geographical maps, and graphs. Visual data can be used for various purposes, depending on the context and industry.

Sampled data

Data from a subset of a larger population or a larger dataset that is used to represent the entire population or dataset.

Sample data refers to a subset of data that is taken from a larger dataset. It is used to represent the characteristics and properties of the entire dataset. Sample data is often used in data analysis and statistical inference to make conclusions or predictions about the population from which the data is drawn.The process of selecting sample data involves randomly selecting a representative portion of the population. This ensures that the sample accurately reflects the characteristics of the larger dataset. By analyzing the sample data, researchers can make inferences about the population as a whole.Sample data is commonly used in various fields, including market research, social sciences, healthcare, and more. It allows researchers to study and analyze data without having to collect and process the entire dataset, which can be time-consuming and resource-intensive.

First Party Data

Data collected by a business directly from its customers, website visitors, or other internal sources

First-party data refers to data that is gathered directly by a business from its customers, website visitors, or other internal sources. It includes any information that is within the company's control and is collected through its own channels. Here are some key characteristics of first-party data:

1. *Directly collected: First-party data is gathered by the business itself, either through interactions with customers or through its own website or internal systems.*
2. *Examples of first-party data: Some common examples of first-party data include website usage statistics, point-of-sale purchase information, customer feedback collected through surveys, and data collected from customer relationship management systems.*
3. *Control and ownership: Since first-party data is collected by the business, it has full control and ownership over the data. The business knows exactly where the data came from and how it was collected.*
4. *Reliability and accuracy: First-party data is generally considered to be more reliable and accurate because it is collected directly from the source. The business has control over the data collection process and can ensure its quality.*
5. *Insights and decision-making: First-party data is valuable for businesses as it provides insights into customer behavior, preferences, and interactions. It can be used to make informed decisions, improve products or services, and personalize customer experiences.*

Overall, first-party data is an important asset for businesses as it provides valuable insights and helps them better understand their customers and make data-driven decisions.

Examples of First Party Data

- A company's own website usage statistics
- Company point-of-sale purchase information
- Customer feedback survey responses collected by the company

Third Party Data

Data gathered by outside parties that are not affiliated with the business itself

Third-party data refers to data that is gathered by outside parties that are not affiliated with the business itself. It is data about a company that is collected by another party, rather than being collected directly by the company. Here are some key characteristics of third-party data:

1. *Collected by external parties: Third-party data is gathered by market research agencies, media agencies, or other external entities that specialize in collecting data about multiple companies.*
2. *Examples of third-party data: Some common examples of third-party data include government census or demographic data, data on behavioral patterns and customer habits collected by market research firms, and economic indicators released by government agencies.*
3. *Less control and ownership: Since third-party data is collected by external parties, the company has less control and ownership over the data. It relies on the data collected by these external sources.*
4. *Insights and market research: Third-party data is often used for market research purposes, advertising campaigns, and competitive analysis. It provides insights into potential target audiences that businesses might not have access to through their own data.*
5. *Purchased or accessed: Companies can purchase third-party data from market research agencies or other data providers. This data can be used for various purposes, such as predictive analytics, marketing campaigns targeting specific demographics, or gaining insights into industry trends.*

It's important to note that knowing the source of data is crucial for understanding its validity and reliability. While first-party data is collected directly by the business, third-party data comes from external sources. Both types of data have their own advantages and use cases, and a combination of both can provide a more comprehensive understanding of customers and markets.

Examples of Third Party Data

- Government census data
- Behavioral and consumer data collected by market research agencies
- Economic indicators