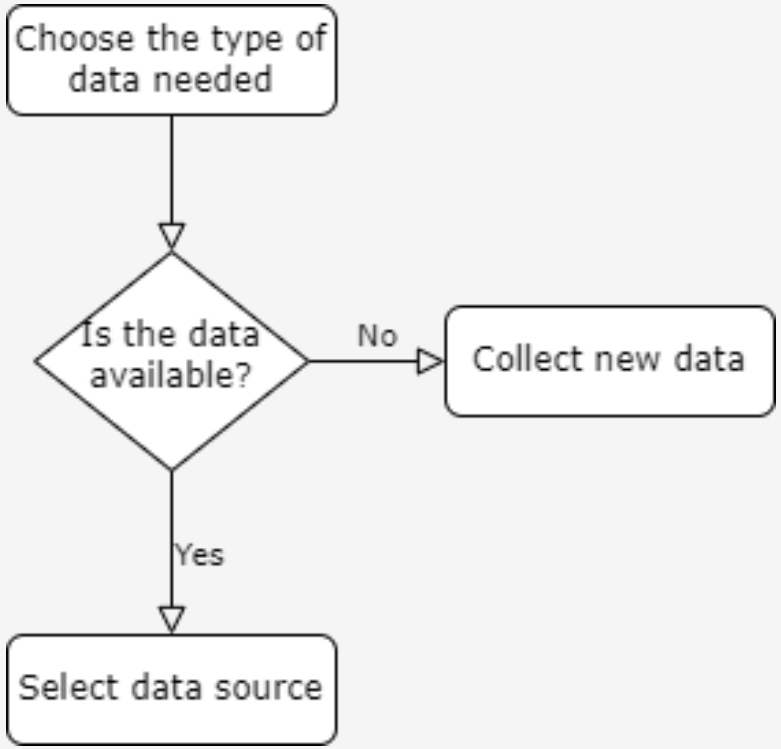
Data Gathering

The data gathering process
Data reliability
Structured and unstructured data
Data extraction from different sources

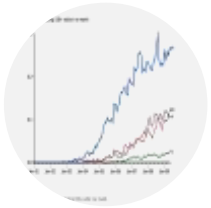
Data Collection

What to do after you have understood the business use case



Data Gathering

How?



Given

Searched

Collated

Itee

Sample Size

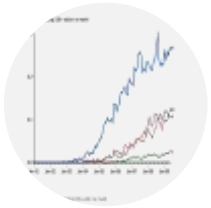
- Don’t use a sample size less than 30. It has been statistically proven that 30 is the smallest sample size where an average result of a sample starts to represent the average result of a population.
- The confidence level most commonly used is 95%, but 90% can work in some cases.
- For a higher confidence level, use a larger sample size
- To decrease the margin of error, use a larger sample size
- For greater statistical significance, use a larger sample size

Sample Size

- Confidence level: The probability that your sample size accurately reflects the greater population.
- Margin of error: The maximum amount that the sample results are expected to differ from those of the actual population.
- Population: This is the total number you hope to pull your sample from.
- Sample: A part of a population that is representative of the population.
- Sample size calculator by SurveyMonkey

Data Gathering

Where?



Different sources

Different formats

How to handle different sources and formats

Itee

Data Formats

Data formats are identified in various forms

Primary and secondary data

Internal and external data

Qualitative and quantitative data

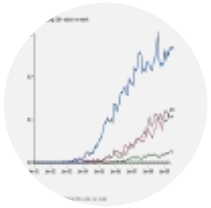
Continuous and discrete data

Nominal and ordinal data

Structured and unstructured data

Data Gathering

Structured data



Flat files

Databases

Semi-structured e.g JSON

Flat Files

A plain text file with one data record per line and fields separated by delimiters

CSV

TSV

Human readable

Simple to understand

Ubiquitous software

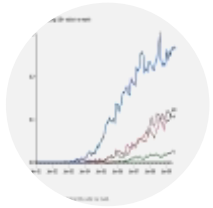
Lack of standards

Sharing data

Not great for large datasets??

*Flat **Files***

Read with Excel, SQL, Python



Data tab; Get data

Import data

Pandas library

pandas.read_csv(file name, separator)

Databases

- Databases store and organize data, making it much easier for data analysts to manage and access information.
- They help us get insights faster, make data-driven decisions, and solve problems.
- A relational database is a database that contains a series of related tables that can be connected via primary and foreign keys.
- A primary key is a unique column that is unique to a particular table
- A foreign key is a field within a table that is a primary key in another table

BigQuery

- BigQuery is a data warehouse on Google Cloud that data analysts can use to query, filter large datasets, aggregate results, and perform complex operations.
- Follow these step-by-step instructions or watch the video, Setting up BigQuery, including sandbox and billing options.
- For more detailed information about using the sandbox, start with the documentation, Using the BigQuery sandbox.
- In your browser, go to console.cloud.google.com/bigquery.
- Watch the How to use BigQuery video for an introduction to each part of the BigQuery SQL workspace.
- To explore public datasets, refer to these step-by-step instructions.
- To upload flat files, refer to these step-by-step instructions.

Additional Resources

Getting started with other databases (if not using BigQuery)

- Getting started with MySQL: This is a guide to setting up and using MySQL.
- Getting started with Microsoft SQL Server: This is a tutorial to get started using SQL Server.
- Getting started with PostgreSQL: This is a tutorial to get started using PostgreSQL.
- Getting started with SQLite: This is a quick start guide for using SQLite.