One of the dictionary definitions of wrangling is herding animals like a shepherd. If the shepherd does not do their jobs, the sheep could end up lost or hurt.

The same concept applies with data wrangling. If data is not organized properly before use, the results of any analysis would be skewed and incorrect.

The definition of data wrangling is: It’s a set of processes for turning raw and messy data into a clean format to answer interesting questions.

There are three phases for data wrangling:

1. Gathering data: in this stage, data is also extracted in the correct format for efficient manipulation
2. Assessing data: checking for any issues in the data. Is there a column for every necessary variable? Is there any missing data? etc.
3. Cleaning data: fixing issues in found during the assessment part.
4. Storing data: Storing happens right after gathering and right after cleaning.

After these phases, the data is used for answering questions and producing visualizations.

Data wrangling vs data inspection (which is when the data that is given is assessed and cleaned):

* Data wrangling focuses on getting data from the source while in data inspection, the data is given.
* Data wrangling places more emphasis on decision making and handling specific data (sometimes complex data).
* Data wrangling focuses more on reproducible workflows, so others can easily replicate the clean data in their system. As part of these workflows, testing will be done to check the work.

Data wrangling goes beyond basic data inspection and manipulation. Using data wrangling techniques ensures better insights into data insights.

What’s the difference between exploratory data analysis and data wrangling?

EDA is used to create visualizations when exploring data. Some visualizations come in handy when data wrangling like for when looking for missing values and outliers to clean. Most data visualizations happen during EDA.

What’s the difference between data wrangling and extract, transform and load (ETL)?

ETL is a systematic process of gathering and combing data from different sources into a central repository called a data warehouse. ETL is used primarily by IT teams on very structured data to get it integrated into a company’s data warehouse for their industry needs. Data wrangling is used by data analysts and data scientists on messy and complex data for many different purposes such as exploration, reporting etc.