1

Liaising with DCU Registry to get datasets from the student registration and results systems.

Answer: Gathering

Anonymising student comments that include identifying details.

Answer: Processing

Conducting student surveys to answer the key questions about their experience.

Answer: Gathering

Calculating the average satisfaction levels based on the sentiment ratings.

Answer: Analysing

Creating a document to share with senior university management summarising the findings.

Answer: Presenting

Removing incorrect entries from the student datasets.

Answer: Processing

Documenting the data formats used in the study and saving all of the created datasets.

Answer: Preserving

Converting student words into sentiment ratings and correlating with the field of study.

Answer: Analysing

Processing:

Definition: Involves preparing the raw data for analysis. This stage focuses on cleaning, transforming, and structuring the data so it's ready for meaningful analysis.

Examples:

Data Cleaning: Removing incorrect or incomplete entries.

Data Transformation: Anonymizing personal details or converting data into a usable format.

Data Aggregation: Summing, averaging, or combining data points (before deeper insights).

Think of it as: Making the data usable.

Analysing:

Definition: Involves examining the processed data to uncover insights, patterns, and conclusions. This stage applies statistical methods, algorithms, or visualization techniques to extract meaning from the data.

Examples:

Calculating Metrics: Determining satisfaction levels or sentiment scores.

Identifying Trends: Correlating survey results with fields of study.

Drawing Conclusions: Interpreting patterns to make recommendations.

Think of it as: Extracting insights from the prepared data.

Okay, let's explore the characteristics and differences of Gathering, Processing, Analysing, Presenting, and Preserving in the context of data. The sources provide a good framework for understanding these concepts, particularly within a Data Analytics Pipeline.

### Data Analytics Pipeline

The sources outline a generic Data Analytics Pipeline represented as a flowchart with five key stages:

* **Gathering**
* **Processing**
* **Analysing**
* **Presenting**
* **Preserving**

### Gathering

**Gathering** focuses on the initial acquisition of data from various sources. This can involve:

* **Capturing** data directly from sensors or user inputs.
* **Importing** data from existing files or databases.
* **Surveying** to collect data through questionnaires or structured feedback.

The sources emphasize the diversity of data sources, including websites, user surveys, sensors, legacy databases, and more.

### Processing

Once data is gathered, the **Processing** stage aims to prepare the data for analysis. This involves tasks like:

* **Cleaning:** Addressing data quality issues, handling missing values, and correcting errors.
* **Aligning:** Ensuring consistency in data formats and structures across different sources.
* **Integrating:** Combining data from multiple sources into a unified dataset.

The sources highlight the importance of data cleaning, noting that data scientists often spend a significant amount of time on this task.

### Analysing

**Analysing** the data involves extracting meaningful insights and patterns. This stage encompasses techniques such as:

* **Statistics:** Applying statistical methods to summarize and describe the data.
* **Machine Learning:** Utilizing algorithms to identify patterns, predict outcomes, or classify data.
* **Exploring:** Interactively examining the data to uncover trends and relationships.

The sources mention that the choice of analysis techniques depends on the type of data being analyzed.

### Presenting

**Presenting** involves communicating the findings derived from the analysis in a clear and understandable way. This often involves:

* **Visualisations:** Creating charts, graphs, and other visual representations of the data.
* **Communication:** Effectively conveying insights through reports, presentations, or other forms of communication.
* **Actionable:** Presenting information in a manner that facilitates decision-making and action.

The sources stress the importance of data visualization as both an analysis and communication tool.

### Preserving

**Preserving** data ensures its long-term accessibility, integrity, and usability. This stage involves:

* **Storing:** Selecting appropriate data storage solutions to ensure data integrity and longevity.
* **Management:** Implementing data management practices to maintain data organization and accessibility.
* **Re-use:** Facilitating the future re-use of the data for further analysis or other purposes.

The sources discuss various data storage approaches, including traditional relational databases, NoSQL databases, and data warehouses.

### Distinguishing Characteristics

The five stages are distinct but interconnected:

* **Gathering** sets the foundation by acquiring the raw data.
* **Processing** transforms the data into a usable format.
* **Analysing** extracts meaning from the processed data.
* **Presenting** communicates the insights to the intended audience.
* **Preserving** ensures the data's availability and integrity for future use.

These stages form a cyclical process, where insights gained from presenting data can inform future data gathering and analysis efforts.

### Key Takeaways

* Each stage in the Data Analytics Pipeline serves a distinct purpose and requires specific skills and tools.
* The stages are interconnected, with each stage influencing subsequent steps in the process.
* Understanding the characteristics and differences between these stages is crucial for effectively managing and analyzing data.

It's important to note that these are broad descriptions. The specific tasks and methods employed within each stage can vary significantly depending on the context, the data itself, and the goals of the analysis.

Let me know if you have any further questions or would like to explore a specific aspect in more detail!

2

Big data is typically defined by the 3 Vs:

Volume: Refers to the massive amount of data generated or collected.

Velocity: Refers to the speed at which data is generated and processed.

Variety: Refers to the different types and sources of data

Records from Spotify:

This involves data on listening behavior for approximately 232 million users, generating a huge volume of data. It also updates in real-time (high velocity) and includes diverse data types like user preferences, song metadata, and listening times (variety). Therefore, it fits the classical big data definition.

3

**1. Qualitative Data (Categorical Data):定性的**

Definition: Descriptive data that represents categories or characteristics. It cannot be measured numerically but can be classified into groups.

Examples:

Types of pets (dog, cat, bird)

Colors (red, blue, green)

Gender (male, female)

Key Traits:

Descriptive or categorical

Often uses labels or names

No inherent numerical value or order (except for ordinal qualitative data（比如满意度Qualitative, Ordinal

，没有数学含义）)

**2. Quantitative Data (Numerical Data):定量的**

Definition: Data that can be measured or counted. It involves numerical values that express quantities.

Examples:

Number of pets (2 dogs, 3 cats)

Height of a person (170 cm)

Weight of an object (500 grams)

Key Traits:

Numerical data

Can be analyzed mathematically

Subdivided into discrete and continuous data

**3. Discrete Data:**

Definition: A type of quantitative data that can only take specific, distinct values (often whole numbers). There are no in-between values.

Examples:

Number of pets (1, 2, 3—not 1.5 pets)

Number of students in a class

Number of cars in a parking lot

Key Traits:

Countable, finite values

Gaps between values (e.g., no 2.3 students)

Often results from counting

**4. Continuous Data:**

Definition: A type of quantitative data that can take any value within a given range. It can be infinitely subdivided into smaller increments.

Examples:

Weight of pets (e.g., 4.5 kg, 4.56 kg)

Height of a person (e.g., 170.3 cm)

Time (e.g., 2.5 hours)

Key Traits:

Measurable, can take any value within a range

Includes fractional or decimal values

Often results from measuring

****Is there a natural order?****

* + No → ****Nominal****
  + Yes → Check next

****Are the intervals equal?****

* + No → ****Ordinal****
  + Yes → Check next

****Is there a true zero?****

* + No → ****Interval****
  + Yes → ****Ratio****

1. Nominal Data:

Definition: Data that represents categories or labels with no inherent order or ranking.

Key Feature: No order or numerical significance. Categories are simply different.

Examples:

Types of pets (dog, cat, bird)

Colors (red, blue, green)

Gender (male, female)

🔍 Tip: You cannot sort nominal data in a meaningful way.

2. Ordinal Data:

Definition: **Data that represents categories with a specific order or ranking**, but the intervals between values are not consistent.

Key Feature: Order matters, but the difference between ranks is not necessarily uniform.

Examples:

Happiness rating (1 = very unhappy, 5 = very happy)

Education level (high school, bachelor’s, master’s, PhD)

Customer satisfaction (poor, fair, good, excellent)

🔍 Tip: You can sort ordinal data, but the difference between "good" and "excellent" might not be the same as between "poor" and "fair."

3. Interval Data:

Definition: Numerical data with equal intervals between values, but no true zero (zero does not mean "none").

Key Feature: You can add and subtract values, but ratios (multiplication/division) are meaningless.

Examples:

Temperature in Celsius or Fahrenheit (0°C does not mean "no temperature")

Dates on a calendar (the year 0 is arbitrary)

IQ scores

🔍 Tip: You cannot say "30°C is twice as hot as 15°C" because there's no absolute zero point.

4. Ratio Data:

Definition: Numerical data with equal intervals and a true zero point (zero means "none").

Key Feature: You can add, subtract, multiply, and divide values. Ratios are meaningful.

Examples:

Weight (0 grams means no weight)

Height (0 cm means no height)

Number of pets (0 pets means no pets)

Income (0 means no money)

🔍 Tip: You can say "20 kg is twice as heavy as 10 kg."

point

line

area

position

colour (hue)

colour (saturation)

slope

size (length)

size (area)

quantity

Choropleth Map

Area: Represents geographical regions.

Position: Corresponds to the geographical location.

Colour (hue/saturation): Indicates categories or intensity.

Line Chart

Point: Marks individual data values.

Line: Connects data points to show trends.

Position: Determines placement along axes.

Slope: Shows the rate of change.

Scatterplot

Point: Represents individual data observations.

Position: Plots values on x and y axes.

Colour (hue): Differentiates categories.

Size (area): Represents a third variable (optional).

Slope Chart

Point: Marks data values at two time points.

Line: Connects points to show changes.

Slope: Indicates magnitude and direction of change.

Box Plot

Point: Shows outliers.

Line: Represents whiskers and median.

Area: Represents the interquartile range (IQR).

Bar Chart

Position: Determines the placement of each bar.

Colour (hue): Differentiates categories.

Size (length): Represents the value of each bar.

1. Point

Function: Represents individual data values or observations at specific points along the x and y axes.

Usage: Often shown as dots or markers on the line chart to highlight key values or intersections.

2. Line

Function: Connects data points to show trends or patterns over time or categories.

Usage: Helps in understanding the direction and magnitude of changes in the data.

3. Area

Function: Refers to the filled space beneath the line in an area chart, emphasizing the magnitude of the values.

Usage: Useful for showing cumulative values or emphasizing volume.

4. Position

Function: Determines the location of data points along the x (horizontal) and y (vertical) axes.

Usage: Crucial for accurately representing data values based on their coordinates.

5. Colour (hue)

Function: Differentiates categories or data series by assigning distinct colors.

Usage: Helps distinguish between multiple lines or groups in a multi-series line chart.

6. Colour (saturation)

Function: Adjusts the intensity or brightness of a color to convey additional information, such as confidence or importance.

Usage: Used for indicating data density or highlighting particular areas of interest.

7. Slope

Function: Represents the rate of change between data points.

Usage: A steeper slope indicates a rapid change, while a flatter slope indicates a slower change. Important for trend analysis.

8. Size (length)

Function: Not typically used directly in line charts but can be applied to bar lengths or line thickness to represent values.

Usage: Sometimes applied to emphasize the importance of specific lines or segments.

9. Size (area)

Function: Refers to the visual area occupied by an element (like points or markers).

Usage: Used in bubble charts or scatter plots but not commonly in line charts.

10. Quantity

Function: Represents the numeric values or amounts being plotted.

Usage: Displayed on the y-axis, showing the magnitude or value of data points.

Choropleth Map

Definition: A map where geographical regions are shaded or colored based on a data variable (e.g., population density, income).

Use Case: Visualizes data distribution across geographic areas.

Line Chart

Definition: A chart that connects data points with lines to show changes over time or trends.

Use Case: Ideal for showing time-series data, such as sales growth over years.

Scatterplot

Definition: A chart that plots data points on two axes to show the relationship between two continuous variables.

Use Case: Used to identify correlations or trends between variables, like temperature vs. water consumption.

Slope Chart

Definition: A chart that shows changes between two points in time by connecting data points with lines.

Use Case: Effective for comparing rankings or values before and after a change (e.g., pre- and post-performance).

Box Plot

Definition: A statistical chart that displays the distribution of data based on a five-number summary: minimum, lower quartile (Q1), median (Q2), upper quartile (Q3), and maximum.

Use Case: Used to show the spread and skewness of data and to identify outliers.

Bar Chart

Definition: A chart that represents data with rectangular bars, where the length of each bar is proportional to the value it represents.

Use Case: Ideal for comparing discrete categories, such as travel methods or survey responses.

Conventional measures of data quality

Accuracy : The data was recorded correctly.

Completeness : All relevant data was recorded.

Uniqueness : Entities are recorded once.

Timeliness : The data is recent or kept up to date.

Date published vs Data captured ...

Consistency : The data agrees with itself (internal).

Problems with conventional measures

Unmeasurable: Accuracy and completeness are extremely difficult, perhaps impossible to measure.

Context independent: No accounting for what is important. E.g., if you are computing aggregates, you can tolerate a lot of inaccuracy.

Incomplete: What about interpretability, accessibility, metadata, analysis, etc.

Vague: The conventional definitions provide no guidance towards practical improvements of the data.

Data delivery issues such as transmission problems that may result in loss of network connectivity, buffer overflows or corruption. → Gathering,

No documentation provided on format for missing values (“”, Nan, -999) → Preserving,

Time synchronization errors resulting in missing values. → Processing,

Unnecessary precision of generated numerical data. → Analysing

Inconsistent units used for recording heights (cm, inches, feet) Processing

Okay, let's explore what constitutes a good measure of data quality using the information provided in the sources.

### Defining Data Quality

The sources primarily focus on identifying and addressing poor data quality but provide insights into what constitutes good data quality. A key takeaway is that **data quality is contextual and dependent on the intended use of the data**. What might be considered good quality for one application may be inadequate for another.

* For example, if you are computing aggregates, a certain level of inaccuracy might be tolerable. However, for sensitive applications like scientific research or medical diagnosis, a much higher level of accuracy is essential.

### Conventional Measures of Data Quality

The sources outline some conventional measures of data quality:

* **Accuracy:** The data was recorded correctly.
* **Completeness:** All relevant data was recorded.
* **Uniqueness:** Entities are recorded only once.
* **Timeliness:** The data is recent or kept up to date.
* **Consistency:** The data agrees with itself (internally).
* **Credibility:** The data comes from a recognised (or official) source.

### Problems with Conventional Measures

The sources also highlight some limitations of these conventional measures:

* **Unmeasurable:** Accuracy and completeness can be extremely difficult, if not impossible, to measure definitively.
* **Context Independent:** They don't account for the specific needs and priorities of different applications.
* **Incomplete:** They don't address other important aspects of data quality, like interpretability, accessibility, metadata, and the suitability for analysis.
* **Vague:** They provide limited guidance on practical steps to improve data quality.

### Alternative Approaches to Data Quality Measurement

Given these limitations, the sources suggest alternative approaches to measuring data quality:

* **Inventory:** A comprehensive inventory of data assets, though expensive, can provide valuable insights into data quality.
* **Proxy Measures:** Tracking customer complaints related to data errors can be a practical indicator of data quality issues.
* **Formal Measures of Accessibility:** Assessing how easily and effectively data can be accessed and used by different stakeholders.
* **Test Cases:** Using test cases with known results to identify glitches or errors in analysis can highlight data quality problems that impact usability.
* **End-to-End Process Completion:** Successfully completing an end-to-end data process (e.g., data ingestion, processing, indexing, querying, and summarisation) can demonstrate the overall fitness of the data for its intended purpose.

### Data Quality Constraints and Metrics

The sources introduce the concepts of data quality constraints and metrics as measurable ways to assess and improve data quality:

* **Data Quality Constraints:** Rules or conditions that data should adhere to, often based on the schema or workflow.
  + Static constraints are fixed rules (e.g., data type restrictions, mandatory fields).
  + Dynamic constraints depend on specific situations or business rules (e.g., order processing workflows).
* **Data Quality Metrics:** Quantifiable measures that indicate specific aspects of data quality. These can be used to track progress, identify problem areas, and incentivize improvements.

The key point is that **adherence to constraints is measurable**, providing a concrete way to assess data quality. Similarly, well-chosen metrics can highlight areas for improvement and drive efforts to enhance data quality.

### Practical Suggestions for Initial Data Quality Checks

The sources offer practical tips for initial data quality checks:

* Checking for missing values, records, or variables.
* Identifying erroneous values (typos or values out of place).
* Addressing inconsistencies in capitalization, units of measurement, etc.
* Eliminating duplicate records.
* Updating out-of-date data.
* Removing leading or trailing spaces and ensuring consistent formatting.
* Performing "sanity checks" to look for extreme values or outliers and verify record counts.

These checks are essential first steps in ensuring data quality before undertaking more complex analysis or processing.

### Key Takeaways

* **There is no single universally agreed-upon measure of data quality.** It depends on the specific context, the intended use of the data, and the resources available.
* **Conventional measures have limitations and should be supplemented with more practical and measurable approaches.**
* **Data quality constraints and metrics provide a framework for defining, measuring, and improving data quality.**
* **Practical checks and a focus on data cleaning are essential for ensuring data is fit for its intended purpose.**

By applying these principles, you can establish effective strategies for assessing and enhancing data quality to support reliable analysis, decision-making, and ultimately, the achievement of your data-driven goals. Let me know if you have any other questions!

Here are the key differences between a **Bar Chart** and a **Histogram**:

### 1. ****Data Type****

* **Bar Chart**: Used to display **categorical data**. Each bar represents a different category, and the length or height of the bar indicates the size or value of that category. It's commonly used to compare different categories.
  + Example: Displaying sales figures for different cities, or counts of different products.
* **Histogram**: Used to display **continuous data** or the **distribution of data**. It groups data into intervals (bins), and each bar represents the frequency or count of data points that fall within a specific range.
  + Example: Displaying the distribution of students' exam scores, or the distribution of income levels.

### 2. ****How Data is Displayed****

**Bar Chart**: Bars are separate from each other with gaps between them, each bar representing a different category. The X-axis typically shows discrete category labels.

* + **X-axis**: Categorical data.
  + **Y-axis**: The value or frequency of the category.

**Histogram**: Bars are adjacent to each other, with no gaps, showing the frequency of data points in a range. The X-axis represents intervals or ranges of continuous data.

* + **X-axis**: Continuous data intervals (bins).
  + **Y-axis**: The frequency or count of data points within each interval.

### 3. ****X-axis and Y-axis Differences****

* **Bar Chart**:
  + **X-axis**: Represents distinct categories.
  + **Y-axis**: Represents values or frequencies for each category.
* **Histogram**:
  + **X-axis**: Represents continuous data intervals (bins), not distinct categories.
  + **Y-axis**: Represents the frequency or count of data points in each interval.

### 4. ****Use Cases****

* **Bar Chart**: Used to compare different categories or groups. It's ideal for comparing the size or count across distinct categories.
  + Example: Comparing sales figures for different products, or comparing the budgets for different departments.
* **Histogram**: Used to show the distribution of data. It helps to understand the spread, frequency, and pattern of data within intervals.
  + Example: Displaying the distribution of exam scores, income distribution, or the frequency of a certain range of values.

### 5. ****Example****

**Bar Chart**: If you have the following sales data:

* + Product A: 200 units
  + Product B: 300 units
  + Product C: 150 units You can use a bar chart to show the sales of each product. The X-axis would represent the product categories, and the Y-axis would represent the sales figures.

**Histogram**: If you have exam scores for 100 students, ranging from 0 to 100, you can use a histogram to show the distribution of scores. You might divide the scores into intervals like 0-10, 11-20, 21-30, etc., and the height of each bar would show how many students fall into each score range.

### Summary

| **Feature** | **Bar Chart** | **Histogram** |
| --- | --- | --- |
| Data Type | Categorical data | Continuous data or distribution data |
| X-axis | Discrete categories | Continuous intervals (bins) |
| Y-axis | Value or frequency of each category | Frequency or count of data points in each bin |
| Purpose | Compare different categories or groups | Show data distribution and frequency |

In short, a **Bar Chart** is used for comparing the values of different categories, while a **Histogram** is used for displaying the distribution of continuous data within certain ranges (bins).

Let's categorize the given data attributes using the definitions from the sources:

* **Qualitative data** describes qualities or characteristics. It is **categorical** and generally **non-numerical**. Examples include gender, country of origin, favorite color. **Limited mathematical functions** can be performed on qualitative data.
* **Quantitative data** represents quantities or measurements. It is **numerical**, and **"all the maths"** can be performed on it. Shoe size, temperature, and bank balance are examples of quantitative data.
* **Discrete data** is quantitative data that can only take **certain values** with "gaps" between them. This data type often results from **counting**, like the number of times someone attended an event or the number of crimes reported.
* **Continuous data**, also quantitative, can theoretically take **any value** within a given range. The **precision of the measurement** depends on the measuring device's capability. Continuous data usually results from **measurements**, like cholesterol levels or the time required to complete a task.

The **four levels of data measurement**, NOIR (Nominal, Ordinal, Interval, Ratio), are:

* **Nominal:** These are labels or categories with **no inherent order**. Examples are gender, department, and language. Nominal data can only be compared for equality or set membership.
* **Ordinal:** Ordinal data has labels **with order**. Examples are temperature (very hot, hot, warm, mild), medals (gold, silver, bronze), and Likert scales. While they can be arranged in order, they cannot be added or subtracted.
* **Interval:** Interval data uses numbers with **proportionate spaces between values**. This means we can talk about the "difference" between values. Income, shoe size, and temperature in Celsius or Fahrenheit are examples of interval data. Interval data **lacks a true zero point**.
* **Ratio:** Ratio data also uses numbers but **includes a true zero point**. This means we can multiply and divide these values. Age, amount of rainfall, book sales, and temperature in Kelvin are examples of ratio data.

### Categorizing the Pet Attributes

Now let's categorize the provided pet attributes based on the above definitions:

#### (i) Type of Pet

* **Qualitative:** It describes a characteristic (the species of the pet).
* **Nominal:** There's no inherent order among "cat," "dog," "bird," and "fish."

#### (ii) Number of Pets Currently Owned

* **Quantitative:** It represents a countable quantity.
* **Discrete:** You can only own whole numbers of pets.
* **Ratio:** There's a true zero point (meaning "no pets").

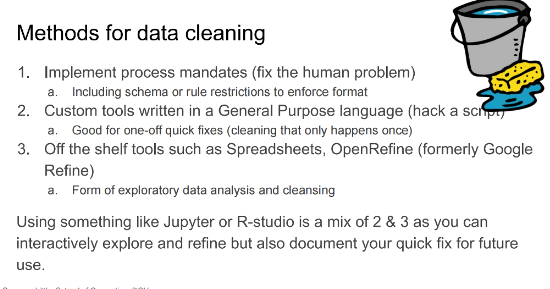
#### (iii) Weight of the Pets

* **Quantitative:** It's a measurable value.
* **Continuous:** A pet's weight can theoretically take any value within a range.
* **Ratio:** There's a true zero point (meaning "no weight").

#### (iv) Happiness of Pet Owners (Self-Rated from 1 to 5)

* **Qualitative:** It represents a subjective perception of happiness.
* **Ordinal:** The scale from 1 to 5 implies an order (1 being the least happy and 5 the most happy), but the differences between each level are not necessarily equal.

This categorization provides a clear understanding of the nature of each attribute and how it can be analyzed.



Let's discuss data cleaning methods and the specific anomalies each method can address.

### Data Cleaning Methods and the Specific Anomalies They Address

The sources provide a range of data cleaning methods, and while they don't always directly connect them to specific anomalies, we can infer from the context and apply common data cleaning practices.

* **Implement Process Mandates (Fix the Human Problem)**:
  + This involves establishing and enforcing rules and restrictions in the data gathering and entry process. This proactively addresses anomalies caused by human error, typos, inconsistencies, and data format issues.
  + **Examples:**
    - Enforcing schema or rule restrictions for data format (e.g., ensuring all dates follow a specific format like DD/MM/YYYY). This addresses inconsistencies in date formats, like those mentioned in the practical suggestions.
    - Building in integrity checks and entry constraints (process architecture). This could prevent anomalies like a gender value appearing in an age column.
    - Process management - rewarding accurate human data entry, promoting data sharing, and establishing data stewards and redundancy.
* **Custom Tools (General Purpose Language - "Hack a Script")**:
  + These tools provide flexibility to address specific, often one-off, anomalies that might not be handled by standard tools.
  + **Examples:**
    - Cleaning that only needs to happen once.
    - Handling a situation where a house price census is missing data for a specific avenue (likely due to an oversight during data gathering). This could involve writing a script to retrieve the missing data from another source or to estimate the values based on similar properties.
    - Correcting temperature values entered in Fahrenheit rather than Celsius. This might involve writing a script to convert the values to the desired unit.
* **Off-the-Shelf Tools (Spreadsheets, OpenRefine)**:
  + These tools offer a visual and often interactive way to explore, identify, and clean data anomalies.
  + **Examples:**
    - Spreadsheets (Google, Excel) are useful for tasks like sorting, filtering, and applying formulas to address basic inconsistencies. For example, you could use a formula to calculate the correct "Total Expenses" for January, correcting the calculation error in the example you provided.
    - Purpose-built tools like OpenRefine (formerly Google Refine), RapidMiner, and Tableau Prep offer more advanced features for tasks like data profiling, transformation, and validation. OpenRefine, in particular, is highlighted as being good for exploratory data analysis and cleansing. These tools are often suitable for cleaning data where entries are overwritten during merging, such as when combining company records.
* **LLMs (Large Language Models)**:
  + LLMs like ChatGPT are emerging tools for data cleaning. They can be beneficial for quick cleaning tasks and can be used without coding experience. However, their reliability and consistency in data cleaning still need further evaluation.
  + The sources don't specify the types of anomalies LLMs are good at addressing.
* **Combination of Approaches ("Mix of 2 & 3")**:
  + Using tools like Jupyter or R-studio allows you to combine the interactive exploration and refinement capabilities of off-the-shelf tools with the scripting and documentation possibilities of custom tools.

The sources emphasize that the choice of data cleaning methods will depend on the nature of the data, the specific errors or anomalies, the available resources, and the goals of the analysis. A combination of different methods is often required to comprehensively address data quality issues.

It's important to note that the provided examples are based on general knowledge and inferences from the sources. The sources do not explicitly link each cleaning method to specific anomalies in all cases. You may want to verify this information based on your specific data cleaning needs.