Course: Research Method in Software Engineering

WEEK 7 – Data Pre-processing Strategies and Analysis

Lemlem Kassa (Ph.D.)
Addis Ababa Science and Technology University,
Ethiopia

Week-7 Data Pre-processing Strategies and Analysis

Contents

Learning Outcome

- Understand data preprocessing and its importance
- Describe the process of data pre-processing
- Understand methods of data analysis
- Understand qualitative and quantitative data and how to analyze

1. Introduction to Data Pre-processing

Data pre-processing

- The critical first step in analyzing data to transform raw data into an understandable and usable format for analysis.
- It’s a comprehensive process that ensures the data is aware and ready for the subsequent exploration, modeling, and interpretation stages-→influences the accuracy of analysis directly.
- Pre-processed data, devoid of irrelevant noise and inconsistencies, allows models to discern and learn from important features, enhancing prediction accuracy and decision-making ability.

[1]. Astera, What Is Data Preprocessing: Definition, Importance, and Steps, Fasih Khan, May 10, 2024. https://www.astera.com/type/blog/data-preprocessing

1.Introduction to Data Pre-processing ….cont'd

Data Preparation vs Data Preprocessing

- They are used synonymously, but they can have different connotations.
  - **Data preparation :-** can be a broader category, including preprocessing, data collection, and integration.
    - It encompasses the entire process of getting data ready for analysis, from when it's gathered to when it's fed into analytical tools.
  - **Data pre-processing:-** part of the preparation, is specifically focused on transforming and conditioning data before analysis.

[1]. Astera, What Is Data Preprocessing: Definition, Importance, and Steps, Fasih Khan, May 10, 2024. https://www.astera.com/type/blog/data-preprocessing

1.Introduction to Data Pre-processing ….cont'd

Importance of Data Preprocessing

a) Eliminating Errors

- Cleaning is a pivotal data preprocessing technique to eliminate errors, impute missing values, and rectify inconsistencies.

b) Making Data Uniform

- Disparate measures can be adjusted to a uniform scale, enabling equitable comparisons.
- For example normalization techniques such as min-max, to convert all stock prices into a common currency.

[1]. Astera, What Is Data Preprocessing: Definition, Importance, and Steps, Fasih Khan, May 10, 2024. https://www.astera.com/type/blog/data-preprocessing

1. Introduction to Data Pre-processing ….cont'd

Importance of Data Preprocessing …..cont'd

c) Big Data Preprocessing

- As datasets grow in size and complexity, preprocessing becomes even more critical.
- Big data has a large volume, is heterogeneous, and needs to be processed rapidly.
- Preprocessing transforms raw big data into a cleaner, more structured format, removing noise and making it easier to process.
- Similarly, advanced techniques such as parallel processing, distributed computing, and automated preprocessing pipelines are indispensable for processing big data effectively.

[1]. Astera, What Is Data Preprocessing: Definition, Importance, and Steps, Fasih Khan, May 10, 2024. https://www.astera.com/type/blog/data-preprocessing

2. Data Pre-processing Steps

Data Pre-processing Steps

- Data preprocessing involves several key stages that transform raw data into a format ready for analysis.

Step-1. Data Profiling

- Involves examining the data using summary statistics and distributions to understand its structure, content, and quality.
    - This step can reveal patterns, anomalies, and correlations crucial for informed preprocessing.

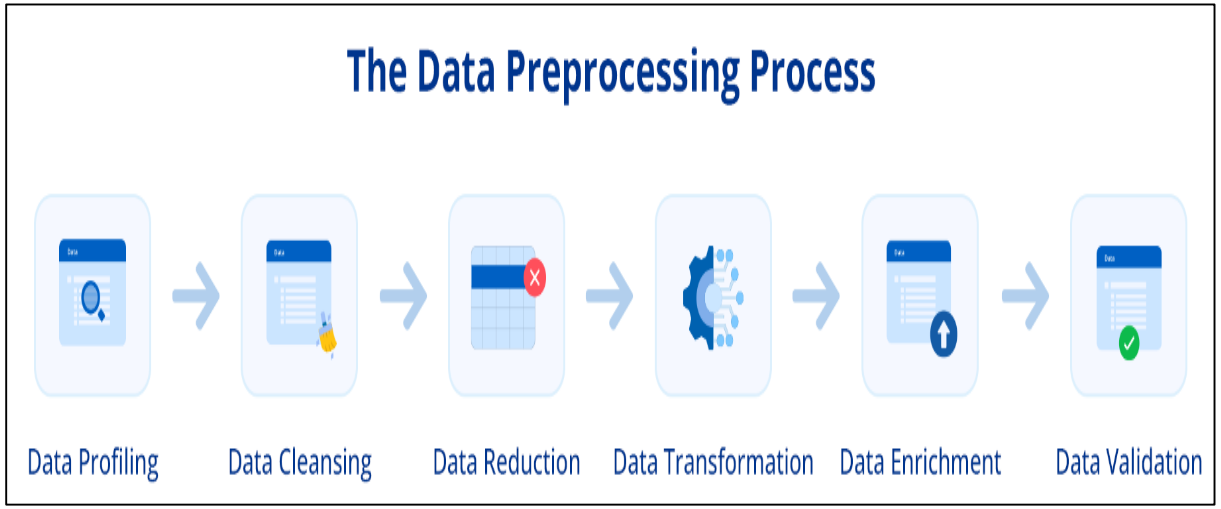

Fig. Data pre-processing Processhttps://images.app.goo.gl/7z5znvwrDr2i65qn6

[1]. Astera, What Is Data Preprocessing: Definition, Importance, and Steps, Fasih Khan, May 10, 2024. https://www.astera.com/type/blog/data-preprocessing

2. Data Pre-processing Steps ….cont'd

Data Pre-processing Steps

Step-2. Data Cleansing

- Detects and corrects corrupt or inaccurate data records such as errors, outliers, duplicates, and missing values.
- Methods like imputation for missing data or trimming for outliers help ensure the accuracy of dataset.
    - **Example:** Correct misspelled product categories or remove duplicate records in sales data.

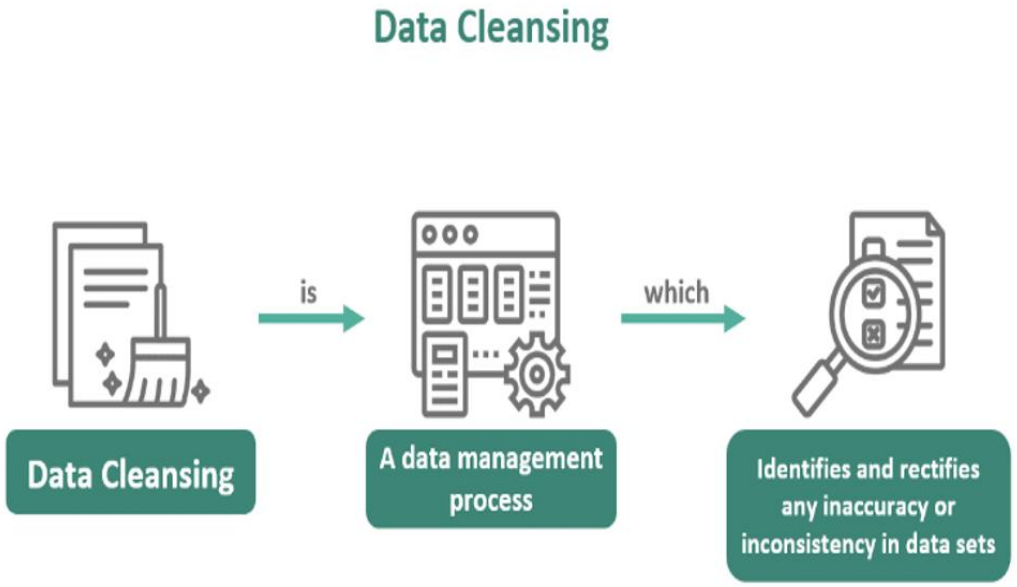

Fig. Data Cleaning.
https://images.app.goo.gl/72BekwNNNL4aJZLXA

[1]. Astera, What Is Data Preprocessing: Definition, Importance, and Steps, Fasih Khan, May 10, 2024.
https://www.astera.com/type/blog/data-preprocessing

2. Data Pre-processing Steps ….cont'd

Data Pre-processing Steps ….cont'd

Step-3. Data Reduction

- This aims to decrease the data volume while producing the same or similar analytical results.
- Techniques like dimensionality reduction, binning, histograms, clustering, and principal component analysis can simplify the data without losing informative patterns and trends.

Example: Data reduction techniques:

- **Compression** is a data reduction technique that reduces the size of files by using algorithms to encode information more efficiently.
- This process works by finding and eliminating statistical redundancies in data, reducing the space required to store it.
- **Feature Selection** involves selecting a subset of relevant features from the dataset to remove irrelevant or redundant features from the dataset.

2. Data Pre-processing Steps ….cont'd

Data Pre-processing Steps ….cont'd

Step-4. Data Transformation

- Helps modify data for specific needs.
- It encompasses a variety of steps such as aggregation, normalization, and sorting, among others, each playing a vital role in understanding data.
- **Example**:- Feature creation devises new variables from the existing dataset, which aids in more effectively discriminating the intrinsic trends within the data.
- A healthcare data analyst leverages mathematical expressions to create new features like Body Mass Index (BMI) through existing features like height and weight.

2. Data Pre-processing Steps ….cont'd

Data Pre-processing Steps ….cont'd

Step-5. Data Enrichment

- Enhancing data with additional sources or derived attributes can provide more depth and context.
- **Example**:- Incorporating demographic information into customer data or adding weather data to sales figures to account for seasonal effects. -→ add weather data to a retailer's sales data to see if weather patterns affect buying trends.

[1]. Astera, What Is Data Preprocessing: Definition, Importance, and Steps, Fasih Khan, May 10, 2024. https://www.astera.com/type/blog/data-preprocessing

2. Data Pre-processing Steps ….cont'd

Data Pre-processing Steps ….cont'd

Step-6. Data Validation

- Before moving on to analysis, it's crucial to ensure the integrity of the data.
- Data validation checks that the data meets specific criteria, such as constraints, relations, and ranges . It helps to confirm that the data is accurate, complete, and reliable.
- **Example:** A finance executive checks whether all entries in a transaction dataset fall within expected date ranges and transaction amounts.

[1]. Astera, What Is Data Preprocessing: Definition, Importance, and Steps, Fasih Khan, May 10, 2024. https://www.astera.com/type/blog/data-preprocessing

3. Methods of Data Analysis

Qualitative Data

- Qualitative data is known as the type of data that represents sentiment in any medium of expression mostly textual.
- It is more categorical than numeric and goes beyond numbers to gain insights from the experiences of people.
- Qualitative data doesn't help in statistical analysis but quantitative data does.

Fig. Qualitative data. https://www.grepsr.com/wp-content/uploads/2024/01/Open-Minded-1.png

[2]. Grepsr, Qualitative Data Analysis: Explore Types, Methods, and Examples, Jeena Timalsina, January 15, 2024. https://www.grepsr.com/blog/qualitative-data-analysis-explore-types-methods-examples/

3. Methods of Data Analysis ..cont'd

Types of Qualitative Data

- Qualitative data is typically generated through:
  - Surveys with open-ended questions
  - Contact center transcripts
  - Reviews, emails or complaints
  - Audio and video recordings
  - Employee notes
  - Social Media Content

[2]. Grepsr, Qualitative Data Analysis: Explore Types, Methods, and Examples, Jeena Timalsina, January 15, 2024. https://www.grepsr.com/blog/qualitative-data-analysis-explore-types-methods-examples/

3. Methods of Data Analysis ..cont'd

Qualitative Data Analysis

- Qualitative Data Analysis is the process of organizing, analyzing, and interpreting the data collected from qualitative research.
- Analyzing qualitative data is quite complicated, so data analysts separate it into 3 categories.
  - **Binary Data:** Numerically represented by the combination of zeros and ones which is used for studying a subject that has either one or the other result. Such as yes/no, positive/negative, up/down, right/wrong based scale, etc.
  - **Nominal Data:** This is used to label mutually exclusive categories that cannot be denoted as a numeric value. Such as gender, ethnicity, college major, mode of transportation, etc

[2]. Grepsr, Qualitative Data Analysis: Explore Types, Methods, and Examples, Jeena Timalsina, January 15, 2024. https://www.grepsr.com/blog/qualitative-data-analysis-explore-types-methods-examples/

3. Methods of Data Analysis ..cont'd

Categories of qualitative Data Analysis ….cont'd

- **Ordinal Data**: Categorized in an order or a ranging scale.
  - The values have clear rank order but lack an even distribution. Such as income: low/medium/high, education level: undergraduate/graduate/postgraduate, etc.
- After data collection, the process of analyzing the review, ratings, and feedback data is known as **sentiment analysis**

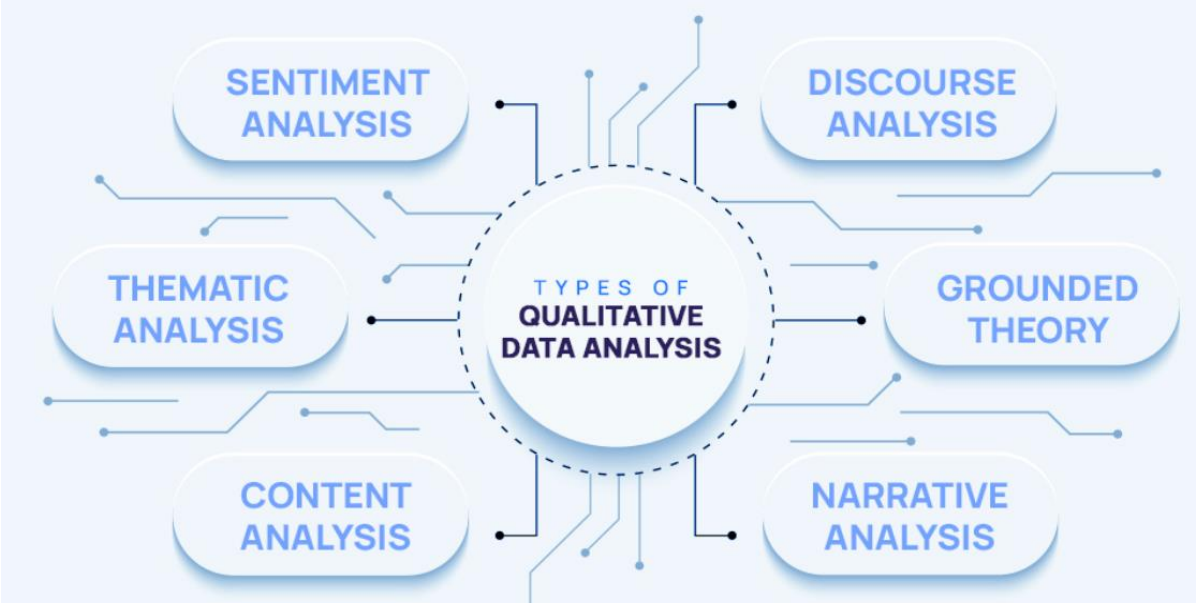

Fig. Types of Qualitative Data Analysis https://www.grepsr.com/wp-content/uploads/2024/01/Types.jpg

[2]. Grepsr, Qualitative Data Analysis: Explore Types, Methods, and Examples, Jeena Timalsina, January 15, 2024. https://www.grepsr.com/blog/qualitative-data-analysis-explore-types-methods-examples/

3. Methods of Data Analysis ..cont'd

Types of qualitative data analysis

- **Thematic analysis**: This is a way to evaluate the **patterns of meaning** in a data set from the answers of focus groups or interview transcripts to provide rich and detailed results from the phenomenon under investigation.
- **Content analysis:** technique used to systematically analyze the **content of communication**, such as texts, images, or videos, to identify patterns, themes, or insights.
- Used to explore the relevance, frequency, and engagement of certain textual, and visual data on social media, search engines, and more.

[2]. Grepsr, Qualitative Data Analysis: Explore Types, Methods, and Examples, Jeena Timalsina, January 15, 2024. https://www.grepsr.com/blog/qualitative-data-analysis-explore-types-methods-examples/

3. Methods of Data Analysis ..cont'd

Qualitative Data Analysis ….cont'd

- **Discourse analysis:** perform discourse analysis to assess a certain speech or conversation that takes place in a specific context, culture, and social setting. This includes the social, cultural, historical, and institutional factors that influence how language is interpreted..
- **Grounded theory**: formulate a theory from the data we have at hand itself.
  - By revising it and having it go through a series of tests, themes, and categories are formed organically during the analysis process.
- **Narrative analysis:** This is essentially analyzing the story telling of people and breaking down its meanings for interpretation and insights. It involves examining the structure, content, and meaning of narratives.

3. Methods of Data Analysis ..cont'd

Quantitative Data

- Data that expresses quantity, amount, or range.
- Quantitative data may be understood as variables in an equation, and these variables can be independent, dependent, or even inessential.
- Some examples of quantitative data include:
  - Counts or units, stored as raw numbers
  - Currency amounts, which are frequently stored as decimal
  - Percentages
  - Ratios
  - Measures of central tendency

3. Methods to Data Analysis ..cont'd

Quantitative Data Analysis

- It is the process of analyzing and interpreting numerical data.
- It helps you make sense of information by identifying patterns, trends, and relationships between variables through mathematical calculations and statistical tests.
- Quantitative data analysis is a more traditional form of analysis. This process crunches numbers to get results.
- The methods used in quantitative data analytics range from basic calculations like mean, median, and mode to more advanced deductions such as correlations and regressions.

[3]. Hotjar, Quantitative Data Analysis: A Complete Guide. August 22, 2023. https://www.hotjar.com/quantitative-data-analysis/

3. Methods to Data Analysis ..cont'd

What quantitative data analysis is not?

- **It only gives the what, not the *why*.** For example, it can tell *how many* website visitors or conversions we have on an average day, but it can't tell *why* users visited our site or made a purchase.
- For the why behind user behavior, we need qualitative data analysis, a process for making sense of qualitative research like open-ended survey responses, interview clips, or behavioral observations.
- By analyzing non-numerical data, we gain useful contextual insights to shape our strategy, product, and messaging.

[3]. Hotjar, Quantitative Data Analysis: A Complete Guide. August 22, 2023. https://www.hotjar.com/quantitative-data-analysis/

3. Methods of Data Analysis ..cont'd

Steps to effective quantitative data analysis

Step-1. Collect data

- This involves conducting quantitate before we can have research and collecting numerical data from various sources, including:
  - Interviews or focus groups , Website analytics, Observations, from tools like heat maps or session recordings, Questionnaires, like surveys.
  - Ensure the questions ask in surveys are close-ended questions—providing respondents with select choices to choose from instead of open-ended questions that allow for free responses

[3]. Hotjar, Quantitative Data Analysis: A Complete Guide. August 22, 2023. https://www.hotjar.com/quantitative-data-analysis/analysis/

3. Methods of Data Analysis ..cont'd

Steps to effective quantitative data analysis …. Cont'd

Step-2. Clean data

- Look through our results to find errors, duplicates, and omissions.
- Keep an eye out for outliers, too. -→ Outliers are data points that differ significantly from the rest of the set—and they can skew results if we don't remove them.
- By taking the time to clean our data set, we ensure the data is accurate, consistent, and relevant before it's time to analyze.

Step-3. Analyze and interpret data

- This step involves crunching the numbers to find patterns and trends via mathematical and statistical methods.

[3]. Hotjar, Quantitative Data Analysis: A Complete Guide. August 22, 2023. https://www.hotjar.com/quantitative-data-analysis/

3. Methods of Data Analysis ..cont'd

Two main branches of quantitative data analysis

1. Descriptive analysis: methods to summarize or describe attributes of a data set. The goal of descriptive statistics is to tell a story that summarizes data in aggregate.

- There are **3 main types of descriptive statistics**:
  - The **distribution** concerns the frequency of each value.
  - The **central tendency** concerns the averages of the values.
  - The **variability** concerns how spread out the values are.

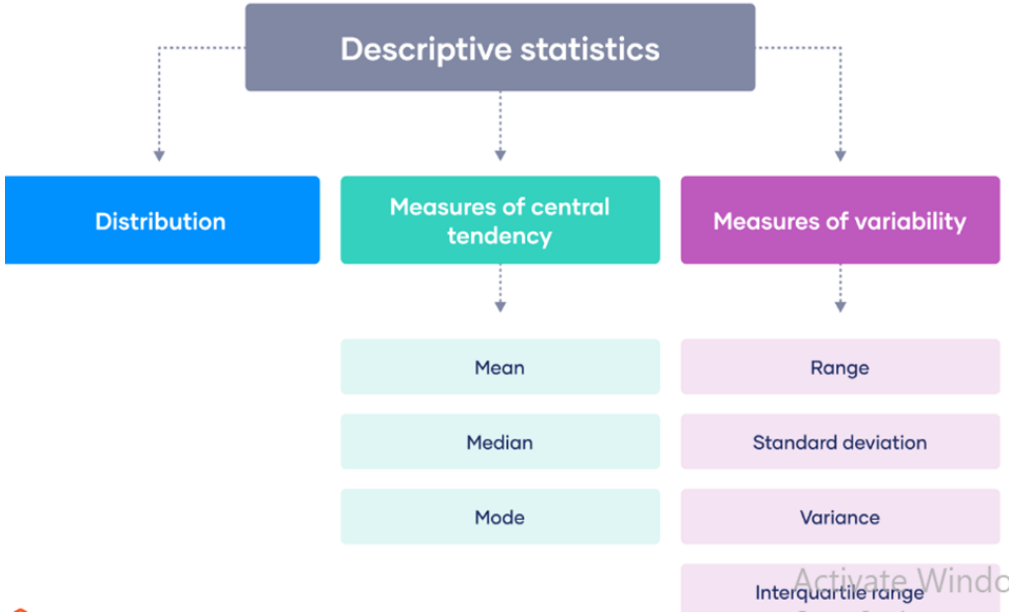

Fig. Distributed Statistics
https://www.scribbr.com/statistics/descriptive-statistics/

[4]. Alation, What's the Difference: Quantitative vs Qualitative Data, Jim Barker. October 12, 2022. https://www.alation.com/blog/quantitative-vs-qualitative-data/

3. Methods to Data Analysis ...Cont'd

2. Inferential analysis: methods that let us draw conclusions from statistics→ include t-tests, cross-tabulation, and factor analysis.

Descriptive Statistics		Inferential Statistics	
Measures of Central Tendency	Measures of Dispersion	Hypothesis Testing	Regression Analysis
Mean Median Mode	Range Standard Deviation Variance Absolute Deviation	Z test F test T test	Linear Regression

Two of the most common types of inferential statistics are:

- **Regression analysis**. This is the act of evaluating across a population how one variable will change with respect to another. Linear regression is most common and is based on changes to an independent variable based on the values of its dependent variable.
- **Hypothesis testing** is one type of inferential statistics that is used to ask a question and test the answers.

[4]. Alation, What's the Difference: Quantitative vs Qualitative Data, Jim Barker. October 12, 2022. https://www.alation.com/blog/quantitative-vs-qualitative-data/

3. Methods to Data Analysis ..cont'd

Steps to effective quantitative data analysis …. Cont'd

Step-4. Visualize and share data

- Once analyzed and interpreted data, create easy-to-read, engaging data visualizations—like charts, graphs, and tables—to present results.
- Data visualizations highlight similarities and differences between data sets and show the relationships between variables.

[3]. Hotjar, Quantitative Data Analysis: A Complete Guide. August 22, 2023. https://www.hotjar.com/quantitative-data-analysis/

3. Methods to Data Analysis ..cont'd

Data analysis software

- Data analysis tools are software platforms that process, analyze, and visualize large data sets to extract meaningful insights and support decision-making.
- These tools range from simple spreadsheet applications, like Microsoft Excel, to more complex data analytics software like SAS, and SPSS, and Python-based libraries like Pandas and NumPy.
- These tools enable users to manipulate data, perform statistical analyses, create predictive models, and present findings in an understandable format through charts, graphs, and dashboards.
- Essential in various fields, including business, finance, healthcare, and research, helping stakeholders identify trends and optimize processes.

[1]. Astera, What Is Data Preprocessing: Definition, Importance, and Steps, Fasih Khan, May 10, 2024. https://www.astera.com/type/blog/data-preprocessing

Summary

- Data preprocessing is critical in analyzing data to transform raw data into an understandable and usable format for analysis.
- Data Preprocessing is important to eliminate errors, maintain data uniformity, and big data preprocessing.
- Data preprocessing involves several key stages that transform raw data into a format ready for analysis such as data profiling, data cleansing , data reduction, data transformation, Data Enrichment and validation.
- Quantitative Data Analysis is the process of analyzing and interpreting numerical data.
- Qualitative data represents sentiment in any medium of expression (mostly textual). And It is more categorical than numeric and goes beyond numbers to gain insights from the experiences of people.

References

1. Astera, What Is Data Preprocessing: Definition, Importance, and Steps, Fasih Khan, May 10, 2024. https://www.astera.com/type/blog/data-preprocessing
2. Grepsr, Qualitative Data Analysis: Explore Types, Methods, and Examples, Jeena Timalsina, January 15, 2024. https://www.grepsr.com/blog/qualitative-data-analysis-explore-types-methods-examples
3. Hotjar, Quantitative Data Analysis: A Complete Guide. August 22, 2023. https://www.hotjar.com/quantitative-data-analysis.
4. Alation, What's the Difference: Quantitative vs Qualitative Data, Jim Barker. October 12, 2022. https://www.alation.com/blog/quantitative-vs-qualitative-data/

Thank You !