In21-S4-CS3121 - Introduction to Data Science

Activity - Data Collecting and Cleaning - In Class Activity

Group K

Data Collecting and Cleaning

To guarantee that the conclusions drawn from data analysis are correct and trustworthy, data quality is crucial. Unfortunately, a lot of datasets have different quality problems that can jeopardise the analysis's objectivity. The most prevalent problems with data quality are highlighted in this paper along with some practical solutions.

Common data quality issues

- Missing Data

When there are null or partial values in the dataset, they are referred to as missing data. Inaccurate inferences can be drawn from analysis results if there are missing data.

- Duplicate Data

Identical records or entries that exist more than once in the dataset are referred to as duplicate data. These data can cause redundancy, skew statistical analysis, and exaggerate counts.

- Inconsistent Data Formatting

When data values vary throughout the dataset, inconsistent data formatting happens. Analysis can be hampered by inconsistent formatting, which also makes it difficult to compare data.

- Incorrect Data Entries

Inaccuracies, typos, or mistakes in the dataset are examples of incorrect data entries. Inaccurate data entry can impair decision-making by producing inaccurate analytical results.

- Invalid Data Entries

Data that falls outside of specified ranges or violates business regulations, such as negative product prices, may indicate errors or fraudulent activities.

Mechanisms to resolve those data quality issues

- Missing Data

Imputation techniques such as mean, median, or mode imputation to fill missing values.
Advanced methods like predictive modelling or interpolation to estimate missing values based on existing data.

- Duplicate Data

De-duplication processes to identify and remove duplicate records.
Implementing unique identifiers or keys to prevent duplicate entries during data collection.

- Inconsistent Data Formatting

Standardising data formats using scripts or tools to ensure consistency.
Regular data validation checks to enforce formatting rules and guidelines.

- Incorrect Data Entries

Manual data verification processes to identify and correct errors.
Implementing data validation rules and constraints to prevent erroneous entries.

- Invalid Data Entries

Implementing data validation checks to enforce preset ranges and business standards during data entry.
Performing regular audits and anomaly detection algorithms to discover and report anomalies.

Conclusion

Data quality issues present challenges to data analysis and decision-making processes. Organisations may, however, improve the integrity of their datasets by recognising the most frequent forms of data quality issues and establishing effective resolution methods. Addressing data quality challenges is critical for obtaining accurate insights and making informed decisions based on data analysis.

Group Members

210518H Ranasinghe K.S
210450P Pathirana L.P.T.R
210483T Prabashwara D.G.H
210588U Senarathna L.P.S.U.K
210460V Perera I.T.M