

Data Glacier – Week 8 Deliverables

Title: Data Cleansing and Transformation


Team Member's Details

- Group Name: Visionary Analysts
- Name: Rishabh Raman
- Email: internships.apply1@gmail.com
- Country: United States
- College/Company: Georgia Tech
- Specialization: Data Science

Problem Description

The goal is to clean and transform a dataset to prepare it for analysis. This includes handling missing values, removing outliers, and cleaning textual data using regex and NLP techniques. Multiple approaches were explored and reviewed.

GitHub Repository Link

 <https://github.com/rishabhraman/data-glacier-week8-cleansing>

Data Cleansing and Transformation Techniques Used

Missing Value Handling

- Mean Imputation: Used for `age`
- KNN Imputation: Used for `bmi` and `salary`

Outlier Handling

- IQR Method: Capped outliers in `income`
- Z-Score: Removed extreme `expenses` values

NLP Text Cleaning

- Removed punctuation, digits, HTML using regex
- Performed lemmatization and stopword removal
- Applied TF-IDF and CountVectorizer

Code Review

- Inline # REVIEW COMMENT blocks included in notebook
- Compared strengths and limitations of each approach