Data Cleansing and Transformation

Katrina Kirby

LISUM43

5/2/2025

Submit a pdf document and ipynb notebook which should contain following details:

Team member's details : Data Collector, Katrina Kirby, katkirby9217@gmail.com, United States, Virginia Commonwealth University, Specialization (Data Analyst)

Problem description: XYZ company is collecting the data customer using google forms/survey monkey and they have floated in number of forms on the web. Company wants to create a pipeline which will collect all the data of these google forms/survey monkey and visualize the data in the dashboard. Company wants clean data and if there is any data issue present in the data then it should be treated by this pipeline (duplicate data or junk data). dedup check should be performed on the email id of the customer

GitHub Repo link:

Data cleansing and transformation done on the data.

Try at least 2 techniques to clean the data (for NA values : mean/median/mode/Model based approach to handle NA value/WOE and like this try different techniques to identify and handle outliers as well)

Up to Week 9, I have:

- Merged and labeled two datasets
- Cleaned up column names and email addresses
- Deduplicated based on Email
- Added a timestamp column
- Saved the cleaned data to a master file
- Handled NA values by dropping rows with NA and Imputation using mean and mode
- Performed outlier detection by using boxplots to visualize outliers in numerical fields
- Retained genuine outlier values instead of removing
- Cleaned “Email Address” with simple string manipulation to ensure consistency