Data Intake Report

Name: Katrina Kirby
Report date: 4/19/2025
Internship Batch: LISUM43
Version:1.0
Data intake by: Katrina Kirby
Data intake reviewer:<intern who reviewed the report>
Data storage location: https://github.com/kirbykt/Week-7-Data-Analyst-Project

Tabular data details:

Total number of observations	20
Total number of files	3
Total number of features	13
Base format of the file	csv
Size of the data	5KB

Total number of observations	10
File name	Product_Survey_Responses
Total number of features	6
Base format of the file	csv
Size of the data	1KB

Total number of observations	10
File name	Service_Survey_Responses
Total number of features	7
Base format of the file	csv
Size of the data	2KB

Total number of observations	20
File name	Master_Combined_Data
Total number of features	13
Base format of the file	csv
Size of the data	3KB

Total number of observations	14
File name	Cleaned_Master_Dataset
Total number of features	13
Base format of the file	csv
Size of the data	2KB

Proposed Approach:

- **Mention approach of dedup validation (identification):**

I created two Google Forms, one for product feedback and another for customer service feedback. Both forms collected qualitative and quantitative responses. I used Python to automate a data intake pipeline that imported both csv files, standardized column names and email formats, merged the data, created a column indicating the data source, performed cleaning and deduplication, and appended batch timestamps for traceability. The primary key I used was “Email Address” and standardized them by lowercasing and trimming whitespace. I removed redundant entries and assumed that the unique email addresses represent a unique customer.

- **Mention your assumptions (if you assume any other thing for data quality analysis):**
  Even though I manually inputted forms and responses, as a consultant for the company I would assume each user has a single unique email address but this could also fail in real world scenarios because of shared emails or typos. NaN values were expected since product and service forms contain different fields, I did not treat the missing values as errors. I assumed simulated data is truthful and no error with submission however in a much more real scenario, safeguards to prevent typos and spelling errors on certain questions will be needed. Since the file size was small, I knew the script would run fast and doesn’t require optimization.