Data Intake Report

Name: CancerMine
Report date: 26/08/22
Internship Batch: LISUM 11
Version:1.0
Data intake by:
Data intake reviewer: Disha Lamba, Somasundaram Palaniappan, Kiarash Rastegar, Mercy Oyekanmi
Data storage location: https://zenodo.org/record/6811941#.YwdZGC8w1aJ

Tabular data details:

Total number of observations	5
Total number of files	
Total number of features	24
Base format of the file	.csv
Size of the data	186.6 MB

Proposed Approach:

- Dedupe Validation: differences between strings will be calculated and assigned a numeric score. This will indicate if there is any similarity between pairs of data. Predicate blocks can be used to group similar values in the data.
- The assumption being made is that the information given regarding different cancers is accurate and up to date.