

Georgetown University
Medical Center
Dept. of Biochemistry and Molecular & Cellular Biology
BINF-6970-01: Databased for Bioinformatics
Spring 2024
Project II

In this project, we want the following modified dataset:

Hospital_General_Information.csv (available in the course data directory)

Implement the following cleaning steps:

- Remove duplicates.
- Remove bank rows.
- Fix all text-based columns.
- Check if there are outliers in certain columns and delete the corresponding rows.
- Make sure the values in each column are consistent.
- Export and submit your cleaning project.

Write and submit a report describing all the cleaning steps you have implemented in this project, including the following:

- The columns you have cleaned and the type of cleaning you have done.
- Columns that have outliers
- Total number of rows in your cleaned dataset