# GSB 521 Chapter 2 Assignment

Student Name:

Date:

Selected Dataset: countries\_fuzzy\_search.csv

*Please submit a csv file with the manipulated dataset along with the assignment document.*

## Section 1: Removing duplicate rows

The following dataset has some duplicate rows. Remove these duplicate rows and give the number of rows in the original dataset.

Download the dataset with duplicated rows:

📂 Duplicated Dataset: http://127.0.0.1:5000/static/csv/duplicated\_countries\_fuzzy\_search.csv

## Section 2: Replacing missing values

Please find the dataset with missing values below. Fill in the missing values of the numeric columns using the mean imputation method.

Download the dataset with missing values:

📂 Missing Values Dataset: http://127.0.0.1:5000/static/csv/missing\_countries\_fuzzy\_search.csv

## Section 3: Clean Incorrect Formats

Below are the replacement rules that are provided. Apply these cleaning rules to the dataset to standardize the values in the selected column. Once this task is done answer the following questions

|  |  |
| --- | --- |
| Incorrect Text | Corrected Text |
| Asian | Asia |
| Europen | Europe |

## Section 4: Creating a new categorical column

*Create a new categorical variable that contains a categorized version of the numeric variable* **'mean\_years\_of\_schooling'***. Please specify the thresholds and what each range should be categorized as. The first few rows are shown here*

### Values from column: 'mean\_years\_of\_schooling'

|  |
| --- |
| mean\_years\_of\_schooling |
| 2.98506999 |
| 5.417390823 |
| 11.28645502 |
| 12.69402981 |
| 11.14726927 |
| 11.33030033 |
| 12.72681999 |
| 12.25669003 |
| 10.5429697 |
| 3.129266883 |

## Section 5: Create a Boolean Variable that Represents a Categorical Variable

*Please review the column***'continent'***below. Please find a few rows here*

### Values from column: 'continent'

|  |
| --- |
| continent |
| North America |
| Asian |
| Africa |
| Europe |
| Asia |
| South America |
| Oceania |
| Europen |

## Create a Boolean Variable

Create a Boolean variable where:  
- Rule 1: True if continent == 'Africa'  
- Rule 2: False if continent != 'Africa'

Based on the new column answer the following questions

How many of 'True' are in the new column?

How many of 'False' are in the new column?