Death Cause by Country
Exploratory Analysis

Name, jschelonka@bellarmine.edu

I. INTRODUCTION

Comparison of the leading causes of death in low and middle income countries. Dataset was found on Kaggle called *Death Cause by Country* at https://www.kaggle.com/majyhain/death-cause-by-country.

II. DATA SET DESCRIPTION

This data set contains 191 unique values with 32 columns with various data types. A complete list of the various data types is shown in Table 1.

this data set contains 398 samples with 7 columns with various data types. A complete listing is shown in **Table 1**. For data types you want to indicate two things (nominal, ordinal, interval, or ratio) and the Pandas data type. For example, age might be ratio/int32. For missing data, indicate what percentage of data from that column are missing. Ensure you check to for NaN, NA, or any other indicators that actually mean missing data.

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
Country Name	Nominal, Object	0%
Covid-19 Deaths	Ratio, float64	0%
Cardiovascular Diseases	Ratio, Int64	0%
Respiratory Diseases	Ratio, int64	0%
Kidney Diseases	Ratio, int64	0%
Neonatal Disorders	Ratio, Int64	0%
Meningitis	Ratio, Int64	0%
Malaria	Ratio, Int64	0%
Interpersonal Violence	Ratio, Int64	0%
HIV / AIDS	Ratio, Int64	0%
Tuberculosis	Ratio, Int64	0%
Maternal Disorders	Ratio, Int64	0%
Lower Respiratory Infections	Ratio, Int64	0%
Alcohol Use Disorders	Ratio, Int64	0%
Diarrheal Diseases	Ratio, Int64	0%
Poisoning	Ratio, Int64	0%
Nutritional Deficiencies	Ratio, Int64	0%
Alzheimer's Disease	Ratio, Int64	0%
Parkinson's Disease	Ratio, Int64	0%
Acute Hepatitis	Ratio, Int64	0%
Digestive Diseases	Ratio, Int64	0%
Cirrhosis and other chronic liver diseases	Ratio, Int64	0%
Protein-energy Malnutrition	Ratio, Int64	0%
Neoplasms	Ratio, Int64	0%
Fire, heat	Ratio, Int64	0%
Drowning	Ratio, Int64	0%
Drug Use Disorders	Ratio, Int64	0%
Road Injuries	Ratio, Int64	0%
Environmental heat and cold exposure	Ratio, Int64	0%
Self-harm	Ratio, Int64	0%

Conflict and Terrorism	Ratio, Int64	0%
Diabetes	Ratio, Int64	0%

III. Data Set Summary Statistics

Table 2 will give a brief summary of the count, mean, std, min, 25th, 50th, and 75th percentiles, and the max of each variable.

Table 2: Summary Statistics for Death Cause by Country

Variable Name	*Count*	*Mean*	*Standard Deviation*	*Min*	*25th*	*50th*	*75th*	*Max*
Country Name	*191*	-	-	-	-	-	-	-
Covid-19 Deaths	*188*	*10461.40*	*35107.42*	*1*	*84.75*	*612.50*	*3986.75*	*353948*
Cardiovascular Diseases	*191*	*9808.93*	*39567.82*	*8*	*464.85*	*1815.70*	*5737.25*	*458427*
Respiratory Diseases	*191*	*2104.24*	*11623.53*	*1*	*52.95*	*252.9*	*711.8*	*116838*
Kidney Diseases	*191*	*7490*	*24085*	*1*	*342*	*1813*	*4773.5*	*222922*
Neonatal Disorders	*191*	*9759.51*	*39729.67*	*0*	*104.5*	*549*	*5296*	*438004*
Meningitis	*191*	*1217.59*	*4486.65*	*0*	*13.5*	*83*	*474*	*44914*
Malaria	*191*	*3275*	*15318.49*	*0*	*0*	*0*	*92*	*191106*
Interpersonal Violence	*191*	*2157.36*	*6791.96*	*0*	*55.5*	*316*	*945*	*65907*
HIV / AIDS	*191*	*4378.55*	*14466.79*	*0*	*22*	*171*	*1813.5*	*143851*
Tuberculosis	*191*	*6071.65*	*31918.46*	*0*	*30.5*	*329*	*2284.5*	*422634*
Neoplasms	*191*	*5372.8*	*21947.5*	*3*	*218.5*	*1008*	*2958.1*	*271655*
Fire, heat	*191*	*580.55*	*2129.85*	*0*	*23*	*117*	*505.5*	*25876*
Drowning	*191*	*1237.52*	*5733.97*	*0*	*37*	*167*	*689.5*	*56524*
Drug Use Disorders	*191*	*684.08*	*4873.22*	*0*	*11*	*53*	*201.5*	*65717*
Road Injuries	*191*	*6246.37*	*24346.92*	*0*	*187.5*	*1049*	*4352.5*	*250025*
Environmental heat and cold exposure	*191*	*248.07*	*1138.29*	*0*	*4*	*24*	*114.5*	*11348*
Self-harm	*191*	*3987.27*	*17190.78*	*0*	*174*	*671*	*2197*	*195336*
Conflict and Terrorism	*191*	*327.72*	*2078.77*	*0*	*0*	*0*	*5*	*24295*
Diabetes	*191*	*8120.7*	*26275.14*	*3*	*412.5*	*1899*	*4736.5*	*273089*

There should be a table for **EACH** categorical variable.

Table 3: Proportions for Death Cause by Country (n=yyy)

Category	*Frequency*	*Proportion (%)*
Country Name	*191*	*100*

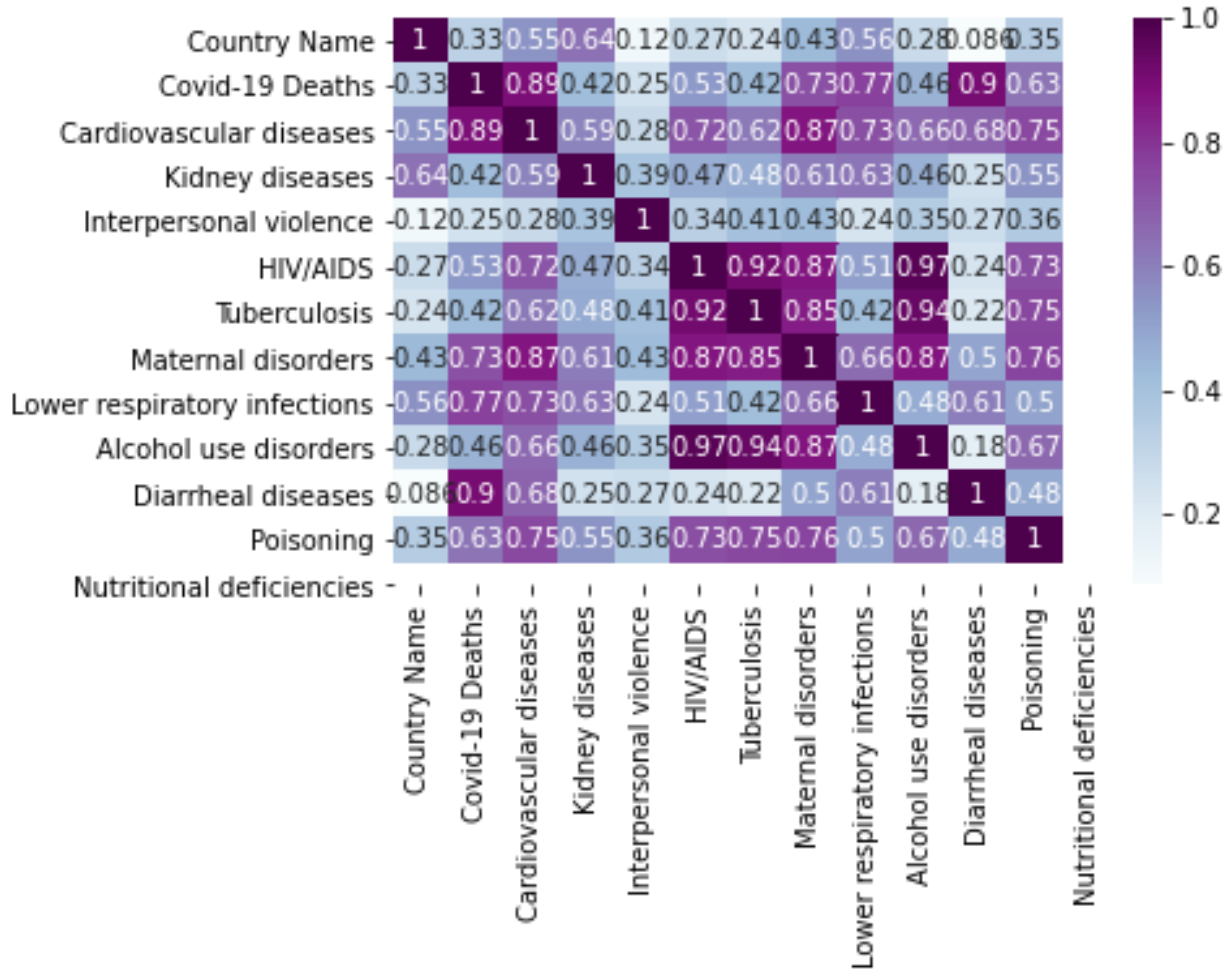
After you summarize the categorical variables, generate a correlation matrix for all continuous variables (not categorical – this doesn't make sense)

Table 4: Correlation Table/Tables

Covid-19 Deaths	*188*	*.035*
Cardiovascular diseases	*191*	*.033*
Respiratory diseases	*191*	*.033*
Kidney diseases	*191*	*.033*
Neonatal diseases	*191*	*.033*
Meningitis	*191*	*.033*
Malaria	*191*	*.033*
Interpersonal Violence	*191*	*.033*
HIV / AIDS	*191*	*.033*

Tuberculosis	*191*	*.033*
Maternal disorders	*191*	*.033*
Lower Respiratory Infections	*191*	*.033*
Alcohol Use Disorders	*191*	*.033*
Diarrheal Disorders	*191*	*.033*
Poisoning	*191*	*.033*
Nutritional deficiencies	*191*	*.033*
Alzheimer's disease	*191*	*.033*
Parkinson's disease	*191*	*.033*
Acute hepatitis	*191*	*.033*
Digestive diseases	*191*	*.033*
Cirrhosis and other chronic liver diseases	*191*	*.033*
Protein-energy malnutrition	*191*	*.033*
Neoplasms	*191*	*.033*
Fire, heat	*191*	*.033*
Drowning	*191*	*.033*
Drug Use Disorders	*191*	*.033*
Road Injuries	*191*	*.033*
Environmental heat and cold exposure	*191*	*.033*
Self-harm	*191*	*.033*
Conflict and terrorism	*191*	*.033*
Diabetes	*191*	*.033*

After the table with the raw data, include a heatmap of the correlation matrix as a figure.



IV. DATA SET GRAPHICAL EXPLORATION

Narrative introduction to the section. In each section below, indicate any interesting distributions, anomalies, imbalance, etc. that you notice.

A. *Distributions – no noticeable distributions or anomalies worth noting. All values follow a consistent ideology that with a higher population the greater the number of deaths per category would be.*
B. *ScatterPlots / Pairwise Plots (continuous variables)*
C. *Barcharts (categorical variables)*
D. *Other Plots - don't skimp – there are likely other plots that would be useful that I haven't already specified. Include those in this section.*

All figures should be cited formatted like this and mentioned in the text.

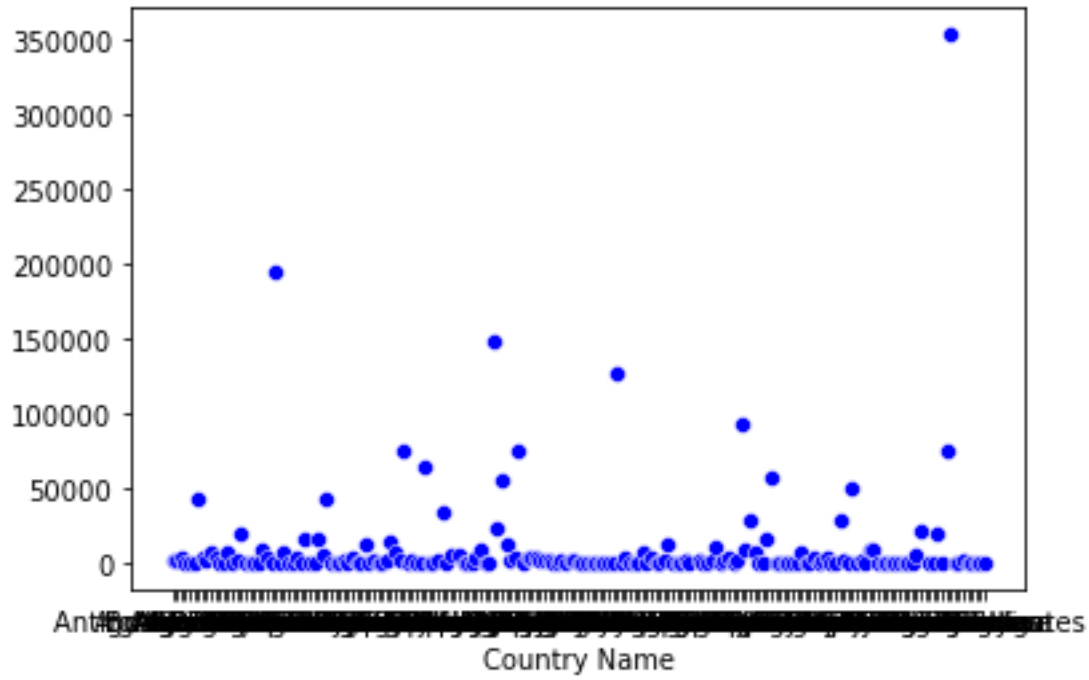


Figure 1: Comparison of Country Name / Covid-19 Deaths from Death Cause by Country

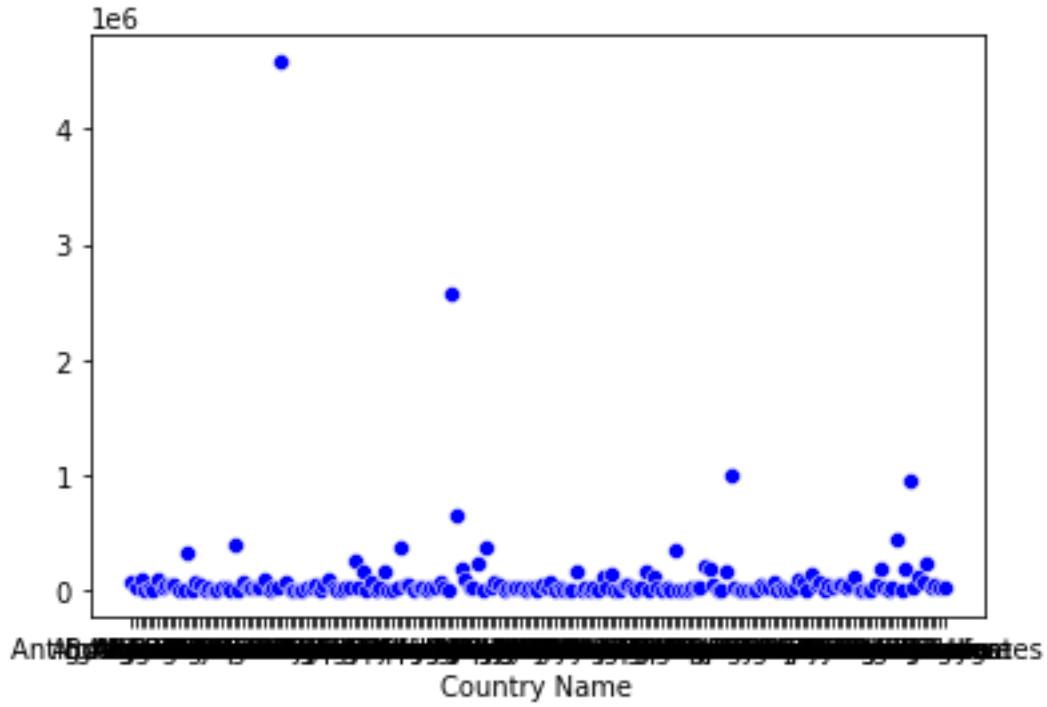


Figure 2: Comparison of Country Name / Cardiovascular diseases from Death Cause by Country

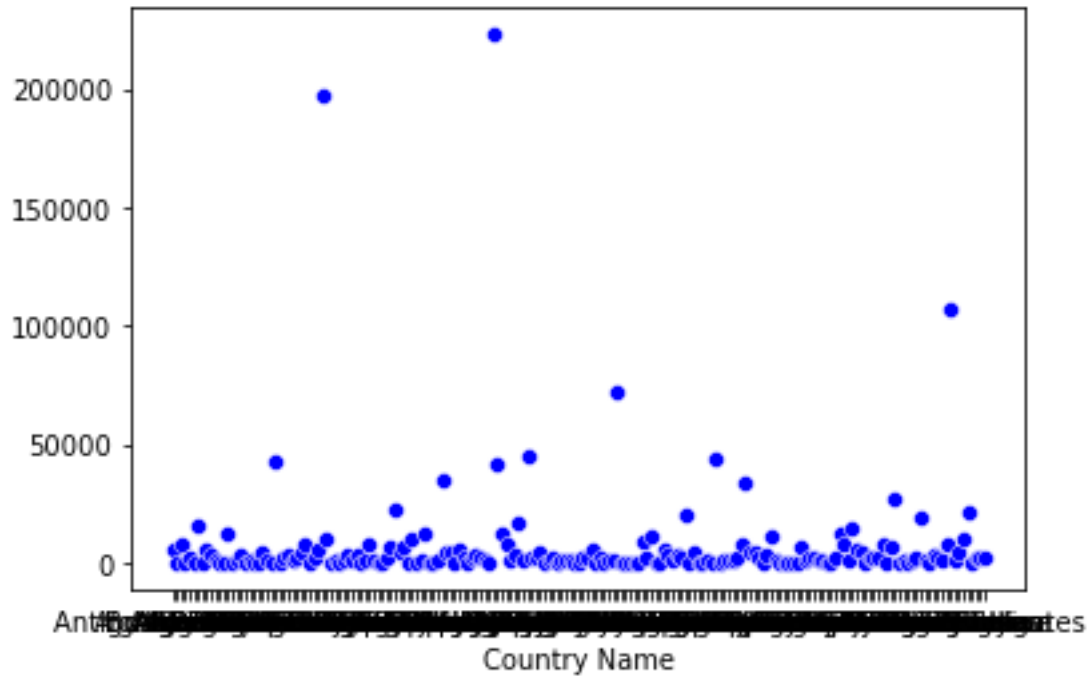


Figure 3: Comparison of Country Name / Kidney Diseases from Death Cause by Country

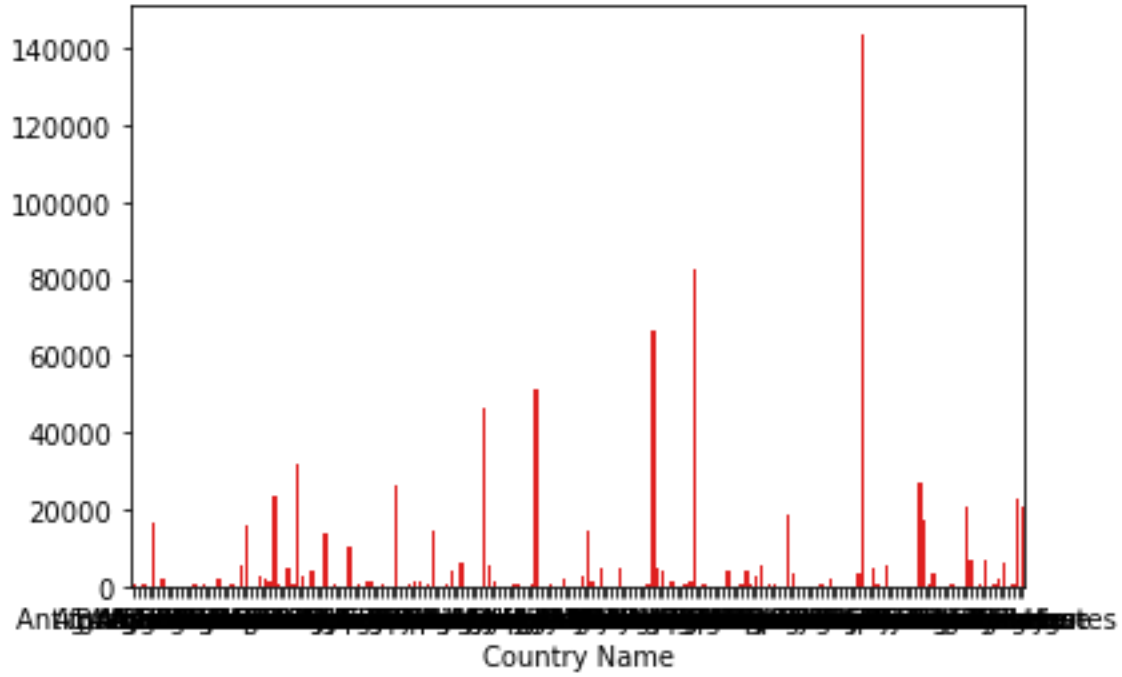


Figure 4: Comparison of Country Name / HIV/AIDS from Death Cause by Country

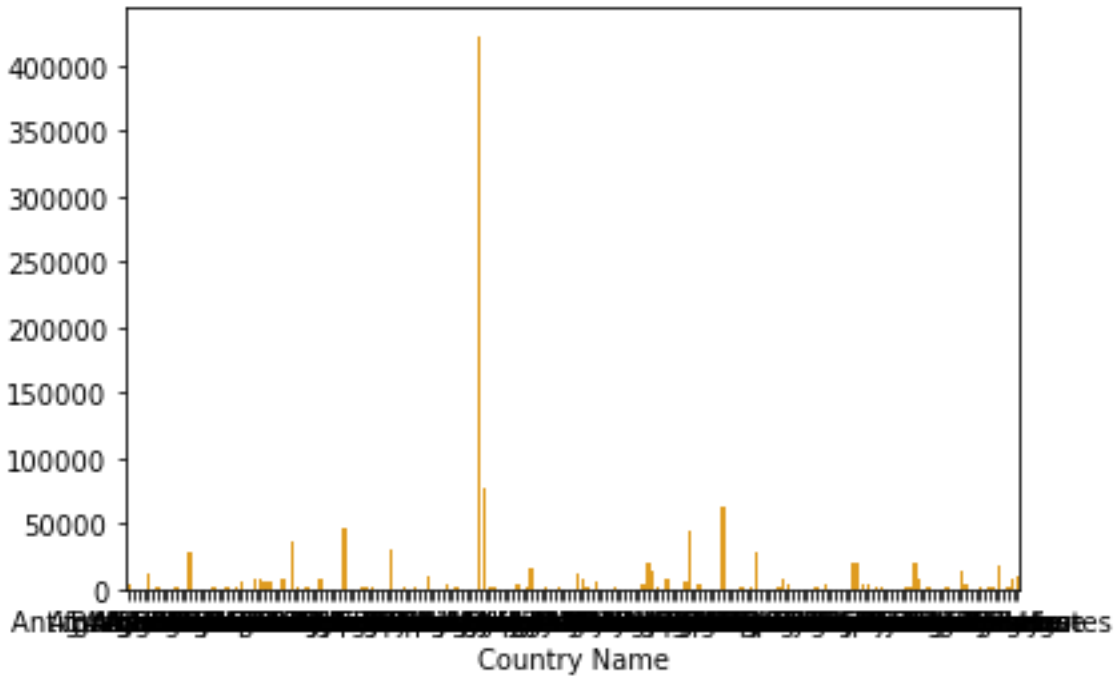


Figure 5: Comparison of Country Name / Tuberculosis from Death Cause by Country

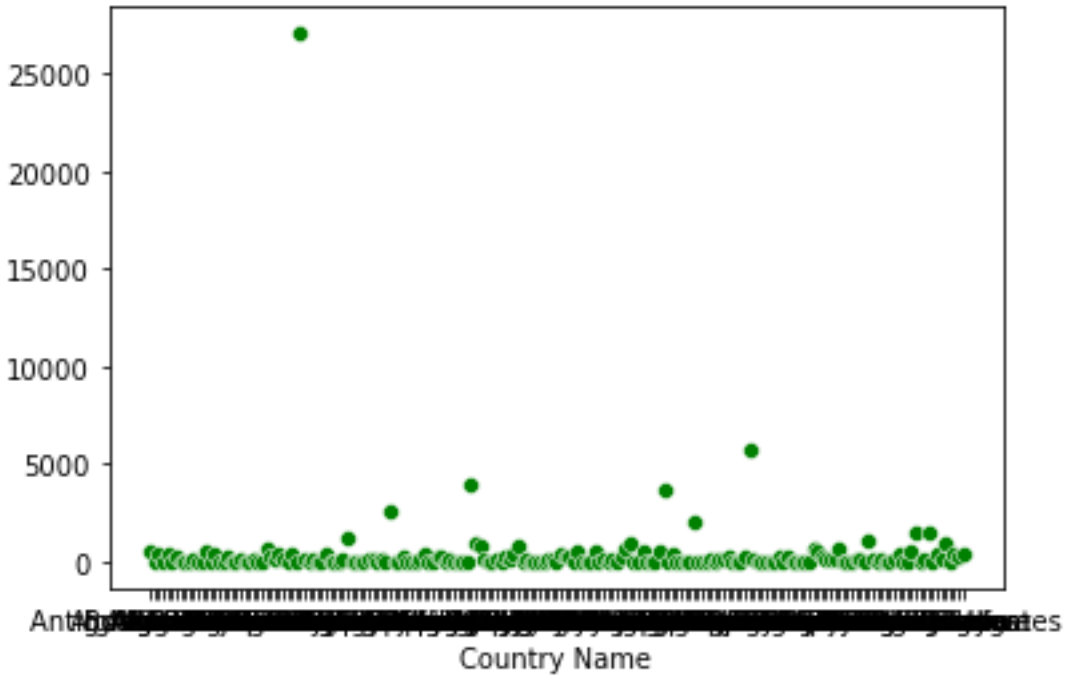


Figure 6: Comparison of Country Name / Poisoning from Death Cause by Country

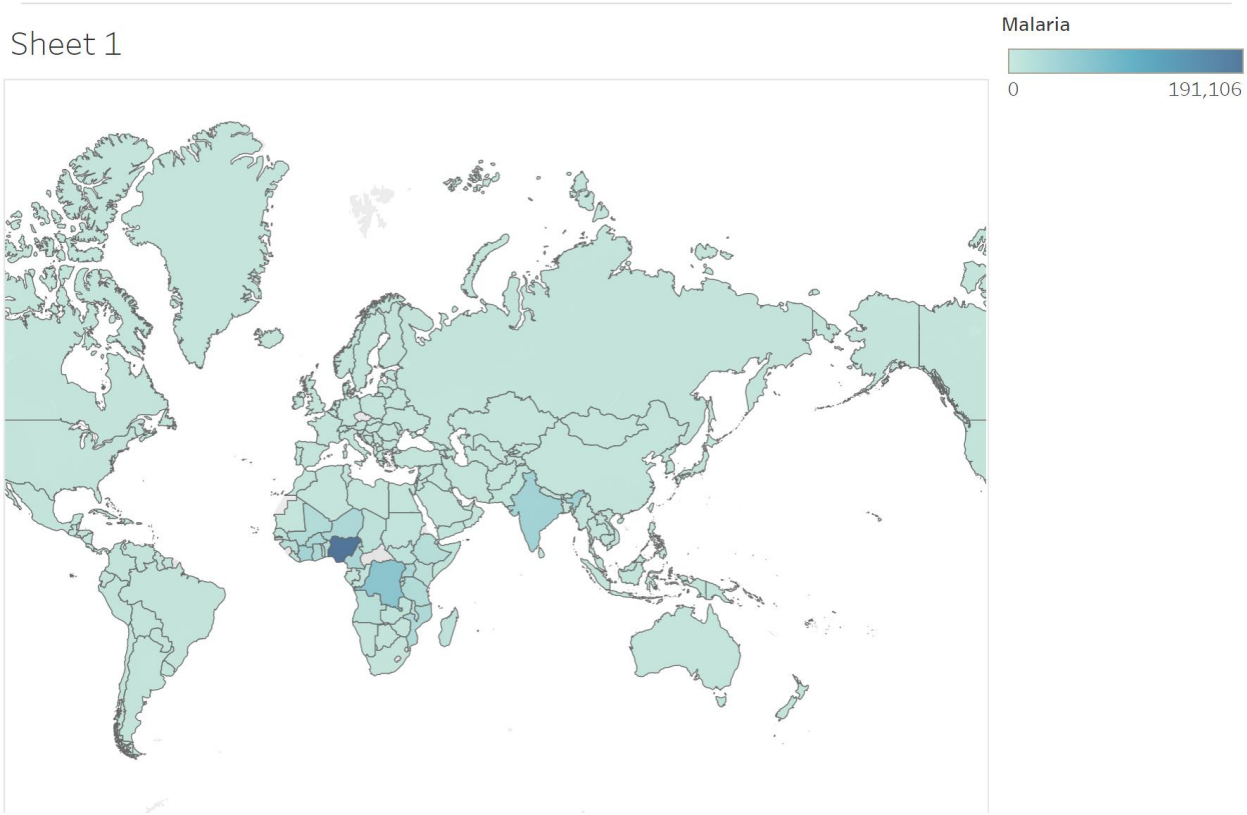


Figure 7: Map comparison of Country Name / Malaria from Death Cause by Country

V. SUMMARY OF FINDINGS

Finish up with a paragraph or two of summarizing your findings about this data set.

During examination of the Death Cause by Country dataset including data from 191 countries and 30 different forms of death present in the countries. I was able to develop an understanding of why certain patterns existed. In countries with greater populations there tends to be a pattern of an increase in each form of death within its respective population. Which makes sense because with more people we expect to see only a greater number of deaths in each category over time, for example if we can say 1 out of every 10 people will die to some sort of cancer but then we take a population of 100 people, we should be able to safely make a reasonable estimate that 10 out of those 100 will die to some sort of cancer. In this data set all categories of cause of death followed that general pattern.