Capstone Project; Data Wrangling

1. What kind of cleaning steps did you perform?

Our dataset (here) has 12 different .csv files whose descriptions are listed in an additional .txt file. Among .csv files 11 of them are distinct. Three of them incorporate definitions of the acronyms and parameters, where they can be found, explanation of the values and missing values and health indicators in 2010 (DATA_ELEMENT_DESCRIPTION, DEFINED_DATA_VALUE, HEALTHY_PEOPLE_2010). The 8 remaining files contain the data. The structure is based on demographics where each row represents one county. The total number of rows is 3141, that is equal to the total number of counties in the USA. The columns are thorough categories of information, e.g., type of diseases, the leading cause of death for different ethnicities within different age groups, preventative care, environmental and social conditions etc.
In order to have a meaningful dataset and define the correlation between these parameters they should be normalized and concatenated.

```
LEADING_CAUSES_OF_DEATH.head()
```

	State_FIPS_Code	County_FIPS_Code	CHSI_County_Name	CHSI_State_Name	CHSI_State_Abbr	Strata_ID_Number	A_Wh_Comp	CI_Max_F_Hi_Cancer	LCD_Time_Span
0	1	1	Autauga	Alabama	AL	29	-1111	-1111	1999-2003
1	1	3	Baldwin	Alabama	AL	16	57	-1111	2001-2003
2	1	5	Barbour	Alabama	AL	51	-1111	-1111	1999-2003
3	1	7	Bibb	Alabama	AL	42	-1111	-1111	1994-2003
4	1	9	Blount	Alabama	AL	28	34	-1111	1999-2003

5 rows × 235 columns

• Normalization:

Data for each county is collected in different time spans –table above shows the leading cause of death data for Autauga is collected in 1999-2003 while it is 2001-2003 for Baldwin. Using a keyword search function first, I identified the *time-dependent* columns, then I divided their values by the number of years in which data was taken.

• Neutral parameters:

There are some columns whose number of unique values are equal to 1, which means they are neutral and they make no impact on the dataset. I dropped them.

• Redundant columns:

Columns whose descriptions include: *Favorable indicator*, *Confidence interval lower limit*, *Confidence interval upper limit* and *percentile* are not necessarily needed for analysis; they can be calculated whenever they are required. I dropped them.

• Not-available/reported data cleaning:

It is shown in DEFINED_DATA_VALUE file that the not-available/reported values are coded as: nan_values = [-1111, -1111.1, -1, -9999, -2222, -2222.2, -2]; I replaced them with 0. Knowing that all the values in this dataset are linear functions of time or they are indicators of varieties of magnitudes, it can be concluded that all the values

must be positive. Thus removing negative values does not hurt the data analysis. The normalization step has to be done before this step, otherwise nan_values will be divided by time spans and creates more numbers to be cleaned.

2. How did you deal with missing values, if any?

In this dataset missing or NaN values are encoded by the nan_value list which is explained above.

3. Were there outliers, and how did you handle them?

With the preliminary analysis the data seems to be proper and does not show any outliers.

One more cleaning step: A couple of column names were not capitalized and they would not be shown in proper alphabetic order; they are taken care of!
Finally after all of the constituent .csv files are wrangled I merged them on the list of county_info = ['State_FIPS_Code', 'County_FIPS_Code', 'CHSI_County_Name', 'CHSI_State_Name', 'CHSI_State_Abbr', 'Strata_ID_Number'], which is the common list of columns in all the .csv files. The final dataframe has 3141 rows corresponding each county and 201 columns of information. Without cleaning the merged dataframe would have 500> columns!