Introduction:

Data Set

https://data.cdc.gov/NCHS/NCHS-Leading-Causes-of-Death-United-States/bi63-dtpu

The dataset that is being used is taken from United States Center of disease control and prevention (CDC) from 1990 to 2017. This study analyzes the leading causes of deaths in United States of America between 1999 and 2017. Dataset were downloaded as CSV file.

Data Preprocessing and Cleaning:

Pandas read_csv() is used to read the file. After checking df.info and dtypes, df.head() gave a preview to the first 5 rows of dataset. Using shape function on dataset we see approximately 10868 cases of death cases were recorded in different US states. Using python built in function null values were searched. There were no null values in our dataset. Datetime is designated as object which is converted to python date time using pandas datetime (pd.to_datetime) function.

113 Cause Name column is splited into other two column cause 1 and cause 2. Cause 2 column is dropped whereas cause1 column is renamed to disease cause. Column '113 cause name' is dropped.

Data for United States were only used for further analysis. Column named as 'State' is renamed to 'Country' column.

Unique Death Causes in the United States were searched using unique function. There are approximately 11 unique causes excluding row containing 'All' causes.