“Analysis of Leading Causes of Death in United States”

Ruchit Tripathi

STAT 6020: Introduction to Statistical Computing Using SAS

Clemson University

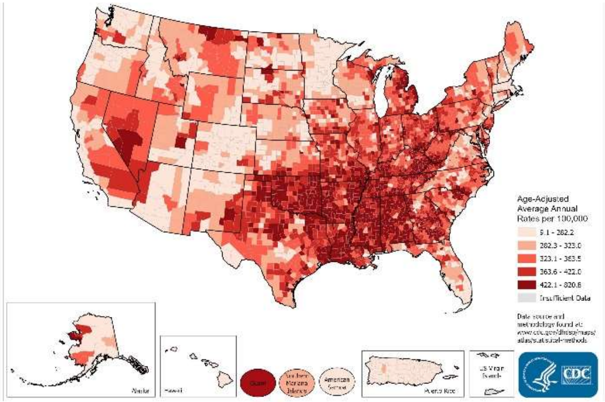
December 15, 2022

INTRODUCTION

In this project, I have tried to find a pattern in the total deaths that have occurred in the last 18 years. Many new diseases have been identified in the last few decades which are impacting the human population across the age spectrum. This project is quite relevant as it tries to find the top diseases that have been proved harmful for humans and took more lives in the past 18 years in the United States. Using the available data, I have tried to answer the following questions that highlights the hidden insights:

- The top diseases responsible for the most death in last 18 years (1999-2017)
- Total deaths by common 10 diseases in the United States
- Average death counts by all causes in the United States along with maximum and minimum deaths by each disease
- Top 3 states and years where most deaths have been identified
- Trends/Variation of each common diseases in past 18 years across United States

Along with the analysis, I have also tried to visualize the data in various graph format to give more insights about the data. As this project requires extensive statistical analysis, SAS stands out to be the best platform to perform the analysis and the same has been used to carry out each operation. At last, the results have been exported to popular output formats such as PDF, RTF, EXCEL etc.



DATA

Data Source

The data has been acquired from the official website of the Centers for Disease Control and Prevention (CDC). It has been published by the National Center for Health Statistics. The dataset is available for public download on the CDC website: link.

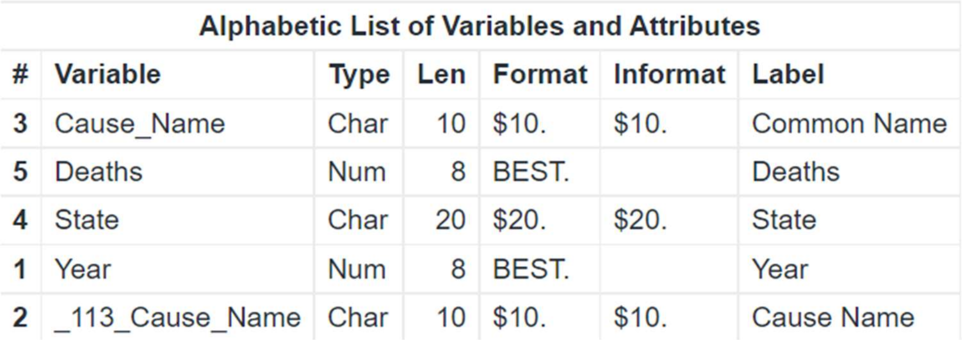
Origin of Data & Purpose

This dataset presents the death rates for the 10 leading causes of death in the United States beginning in 1999. Data are based on information from all resident death certificates filed in the 50 states and the District of Columbia using demographic and medical characteristics. Age-adjusted death rates (per 100,000 population) are based on the 2000 U.S. standard population. Populations used for computing death rates after 2010 are postcensal estimates based on the 2010 census, estimated as of July 1, 2010. Rates for census years are based on populations enumerated in the corresponding censuses. Rates for non-census years before 2010 are revised using updated intercensal population estimates and may differ from rates previously published.

Causes of death classified by the International Classification of Diseases, Tenth Revision (ICD–10) are ranked according to the number of deaths assigned to rankable causes. Cause of death statistics are based on the underlying cause of death

Description of Data

The dataset contains 6 columns: Year, Cause Name, Common Name of the Cause, State, Deaths, and Age-Adjusted Death Rates. The year column has the range from 1999 to 2017. The Cause Name column contains the scientific name of the disease responsible for death. The Common Name column is the name of the cause generally known to the common people. The state columns have all the 50 states where the deaths have been identified for a particular disease. The death column contains the numeric value and represents the death count.
Below is the technical detail of each column obtained by PROC CONTENTS statement:



Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
3	Cause_Name	Char	10	$10.	$10.	Common Name
5	Deaths	Num	8	BEST.		Deaths
4	State	Char	20	$20.	$20.	State
1	Year	Num	8	BEST.		Year
2	_113_Cause_Name	Char	10	$10.	$10.	Cause Name

Complete SAS Programming Procedure

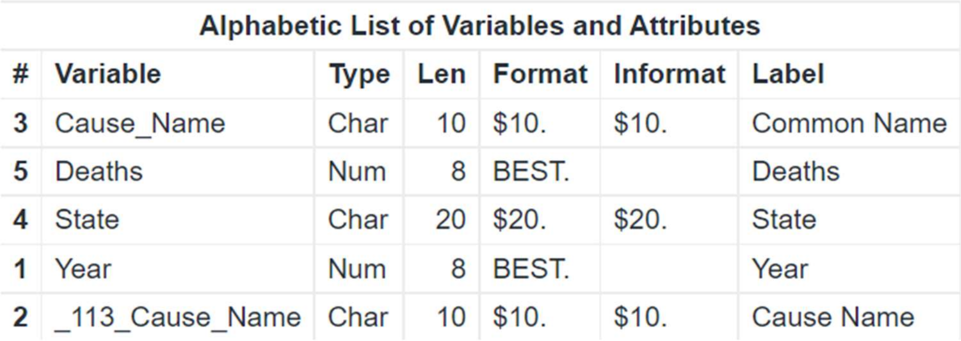
Step 1: Data Access

The raw data is available in the .xlsx format with only one sheet (NCHS). It contains all the death counts for each state (including the nationwide count), each common disease (including summation of all diseases) from 1999 to 2017. Before importing the excel sheet to SAS environment, I separated the ‘All Disease’ count for each state (and the nationwide count) from the main sheet and saved it into a new worksheet. This helps to distinguish the count by each disease and overall disease for a particular state and the year.

After the initial amendments to the raw file, the excel sheet was imported to the SAS system using PROC IMPORT procedure. Since the file was in excel format, OPTIONS VALIDVARNAME=V7 was used to put variable name constraints on the excel columns. Two PROC IMPORT statements have been used to import two different sheets mentioned above (Count by All diseases and Count by individual disease).

Step 2: Exploring Data

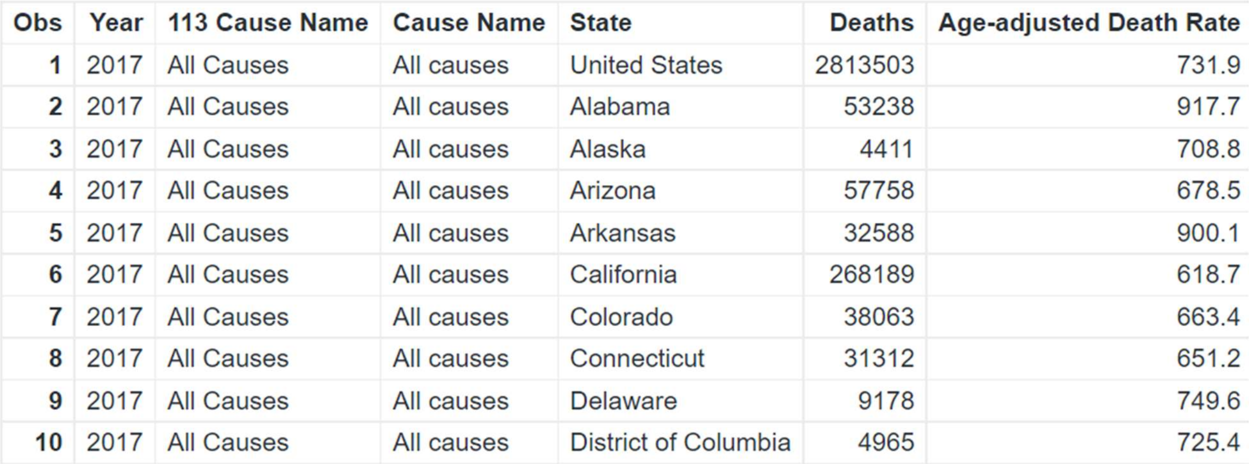
First, the PROC CONTENTS procedure has been used to explore the data. It gave the skeleton of the imported SAS table along with the variable name, their datatype and respective length in the SAS data table. Secondly, the PROC UNIVARIATE procedure was used to check if any column contains missing values. This can also be done by simple lookup in excel before importing it to the SAS system. The PROC CONTENTS showed the following attributes:



Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
3	Cause_Name	Char	10	$10.	$10.	Common Name
5	Deaths	Num	8	BEST.		Deaths
4	State	Char	20	$20.	$20.	State
1	Year	Num	8	BEST.		Year
2	_113_Cause_Name	Char	10	$10.	$10.	Cause Name

To get a glimpse of the imported data tables, the PROC PRINT statement was used to list the first 10 observations for the table. Below is the sample output of the PROC PRINT statement:



Obs	Year	113 Cause Name	Cause Name	State	Deaths	Age-adjusted Death Rate
1	2017	All Causes	All causes	United States	2813503	731.9
2	2017	All Causes	All causes	Alabama	53238	917.7
3	2017	All Causes	All causes	Alaska	4411	708.8
4	2017	All Causes	All causes	Arizona	57758	678.5
5	2017	All Causes	All causes	Arkansas	32588	900.1
6	2017	All Causes	All causes	California	268189	618.7
7	2017	All Causes	All causes	Colorado	38063	663.4
8	2017	All Causes	All causes	Connecticut	31312	651.2
9	2017	All Causes	All causes	Delaware	9178	749.6
10	2017	All Causes	All causes	District of Columbia	4965	725.4

Step 3: Preparing Data

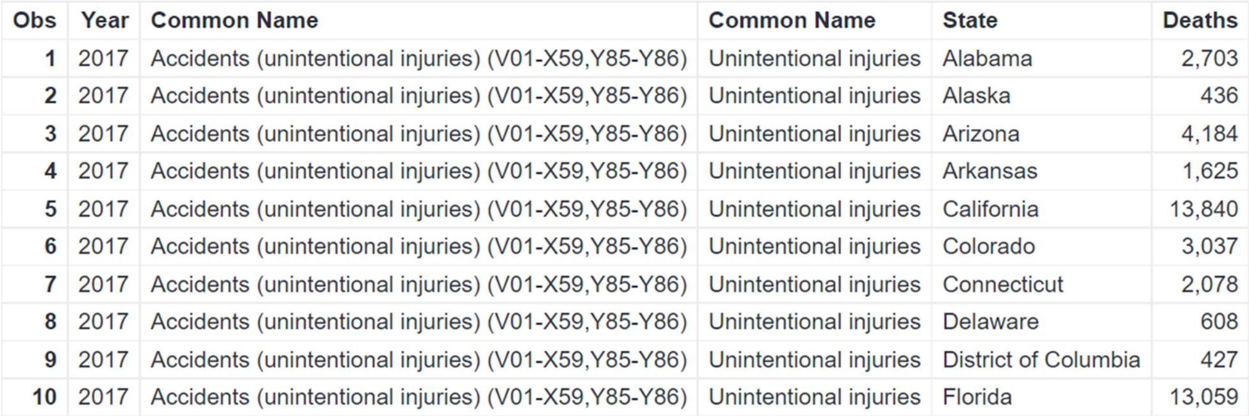
After importing the raw data into SAS system, the tables were rearranged and modified by dropping/keeping the relevant columns. Since the Age-Adjusted Death Rate was not a primary attribute for the analysis, it was dropped from the existing SAS table at the DATA step. Along with dropping the irrelevant columns, the columns with cryptic names were renamed using label statement at the DATA step. For example, '113 Cause Name' column was renamed as "Cause Name" and "Cause Name" was renamed as "Common Name".

The initial data was already sorted by the year, the cause name and then by the name of the states. Hence PROC SORT procedure was not required in our operation.

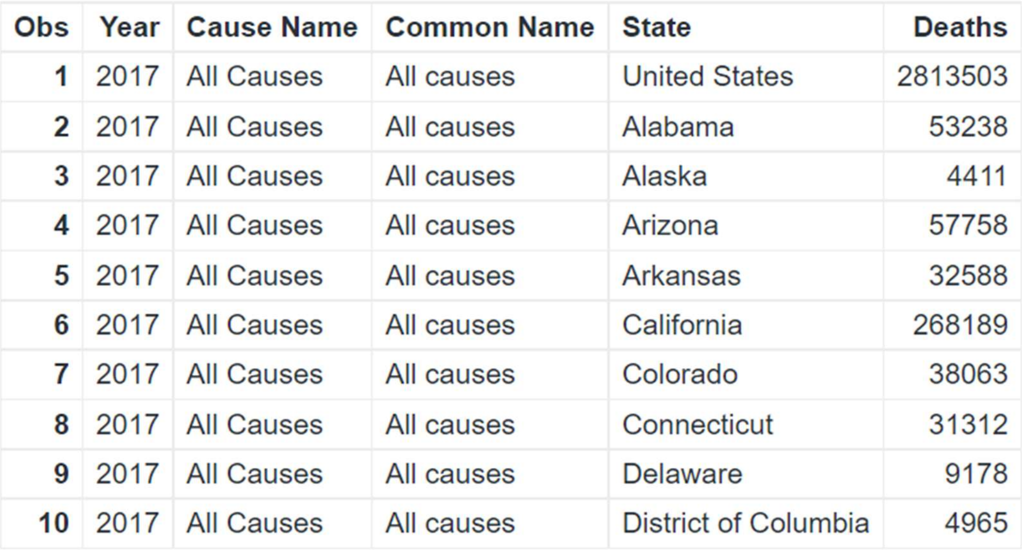
A total of 5 new tables were created from the cleaned SAS data table using DATA STEP to ease the future analysis process. The description of each table has been given below:

Tables	Description
Table_all_cause	Imported table from Excel (Raw) with all columns
Total_all_cases	Restructured SAS table by dropping off unnecessary columns from Table_all_cause
Total_deaths_USA_overall	Death counts by all causes for USA from 1999-2017 (derived from Total_all_cases)
Total_death_state_overall	Death counts by all causes for each 50 states from 1999-2017
Deaths_USA_each_cause	Death counts by each of the 10 common causes for USA from 1999-2017
Deaths_each_state_cause	Death counts by each of the 10 common causes for 50 states from 1999-2017

Below is the snapshot of the cleaned data (keeping only the relevant columns with appropriate labels):



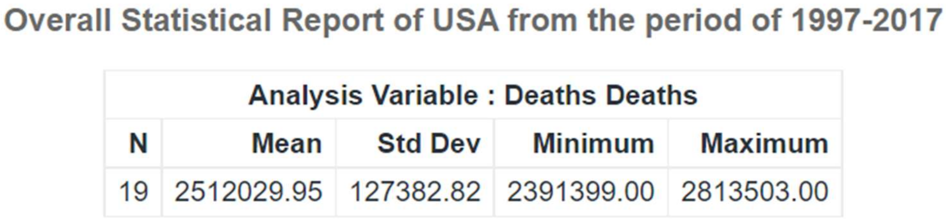
Obs	Year	Common Name	Common Name	State	Deaths
1	2017	Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional injuries	Alabama	2,703
2	2017	Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional injuries	Alaska	436
3	2017	Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional injuries	Arizona	4,184
4	2017	Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional injuries	Arkansas	1,625
5	2017	Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional injuries	California	13,840
6	2017	Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional injuries	Colorado	3,037
7	2017	Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional injuries	Connecticut	2,078
8	2017	Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional injuries	Delaware	608
9	2017	Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional injuries	District of Columbia	427
10	2017	Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional injuries	Florida	13,059



Obs	Year	Cause Name	Common Name	State	Deaths
1	2017	All Causes	All causes	United States	2813503
2	2017	All Causes	All causes	Alabama	53238
3	2017	All Causes	All causes	Alaska	4411
4	2017	All Causes	All causes	Arizona	57758
5	2017	All Causes	All causes	Arkansas	32588
6	2017	All Causes	All causes	California	268189
7	2017	All Causes	All causes	Colorado	38063
8	2017	All Causes	All causes	Connecticut	31312
9	2017	All Causes	All causes	Delaware	9178
10	2017	All Causes	All causes	District of Columbia	4965

Step 4: Analyzing and Reporting

After preparing and creating the appropriate tables, we can jump to the most important part of the SAS programming – Analysis & Reporting. As described in the introduction section, the main purpose of the project is to get the deep insights of death counts caused by the common 10 diseases. The answer to the first question can easily be found by the PROC MEAN procedure on the *Total_deaths_USA_overall* table. As described in *Section 3,* the table contains overall death counts across the United States from 1999 to 2017 caused by all diseases. A snapshot of the result has been pasted below for reference.

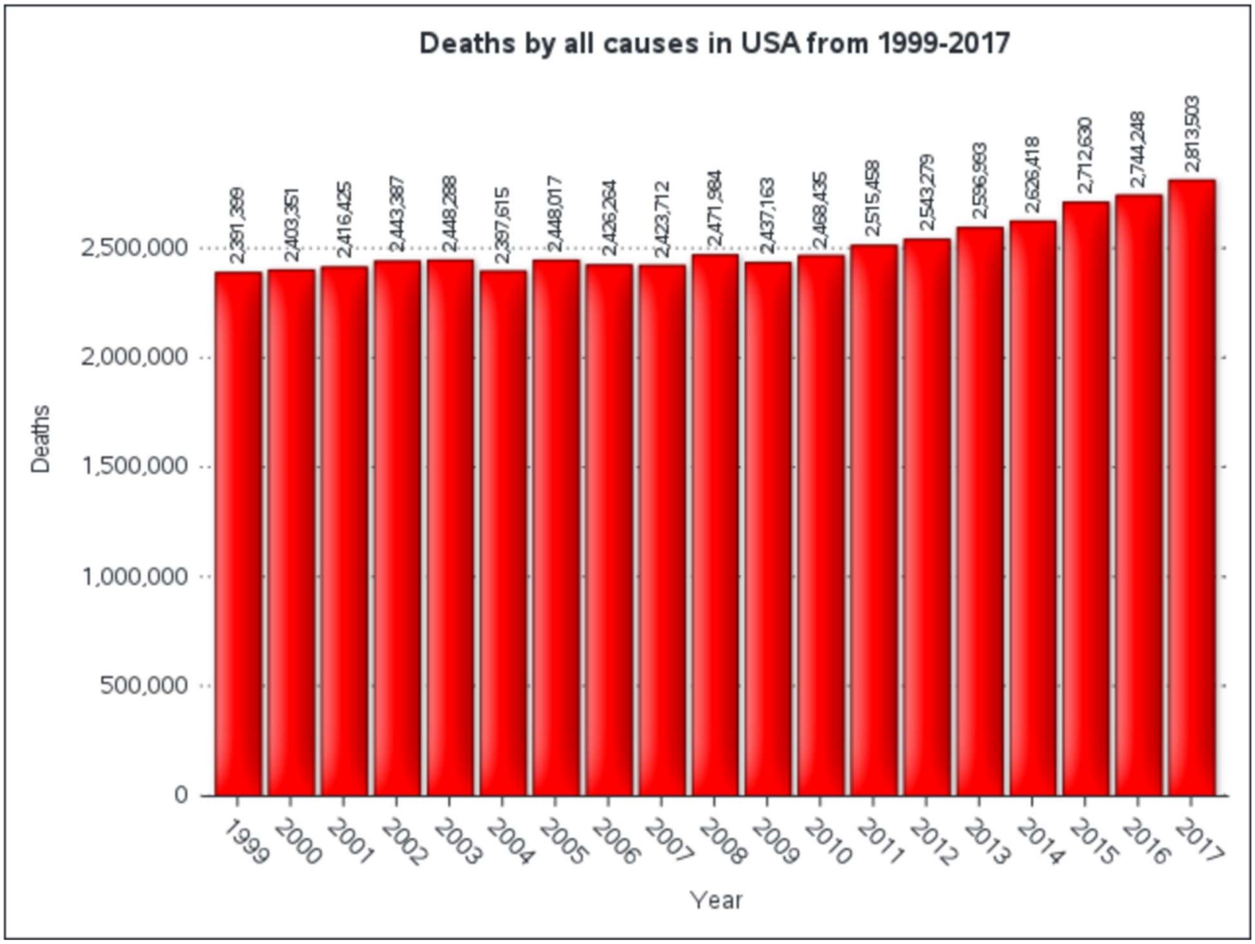


Overall Statistical Report of USA from the period of 1997-2017

Analysis Variable : Deaths Deaths				
N	Mean	Std Dev	Minimum	Maximum
19	2512029.95	127382.82	2391399.00	2813503.00

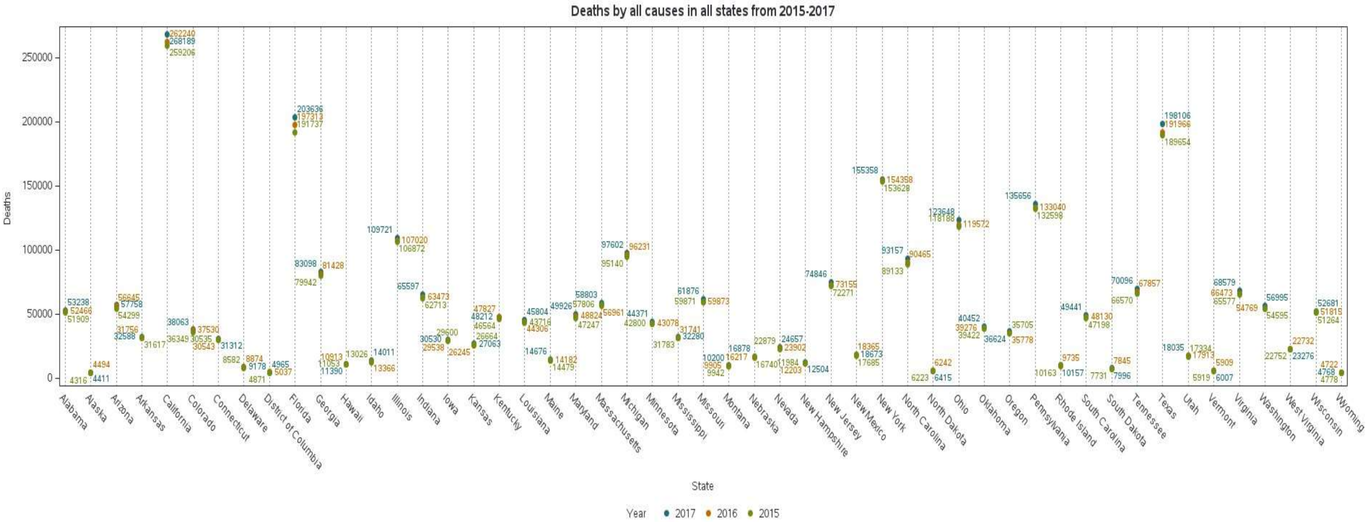
From the snapshot, it can be interpreted that the maximum death count occurred by all causes for a particular year is 2,813,503 and the minimum death count is 2,391,399. The average death count over the period of 18 years is 2,512,029.95.

The graphs can be used to see the variation in the total death counts from 1999 to 2017. It can also help us to identify the top 3 years where the most deaths happened by all causes. To achieve the objective, PROC SGPLOT procedure has been used on data table *Total_deaths_USA_overall.*



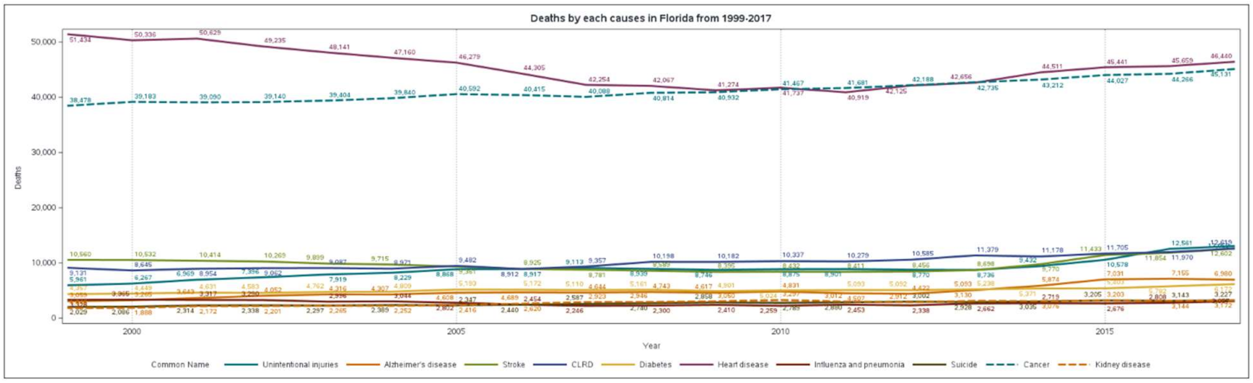
As shown in the graph, it can be easily interpreted that the most death counts (by all causes) were occurred from 2015 to 2017.

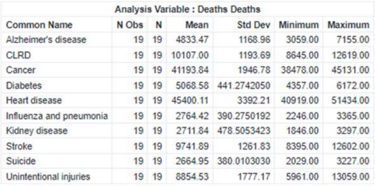
Now, to get the states contributing the most death counts (by all causes) from 2015 to 2017, PROC SGPLOT can again be used on data table *total_deaths_state_overall*.



The top 3 states with highest death counts (by all causes) are California, Florida, and Texas. We can analyze these states to get the overall idea about the top 3 diseases contributing the death counts. Since the SAS code is going to be the same for all three states, I chose to use MACRO variables to ease my analysis. The PROC SGPLOT and PROC MEANS have been again used on data table *deaths_each_state_cause* to identify the diseases/causes helping in highest death counts in Florida, California and Texas.

Florida:



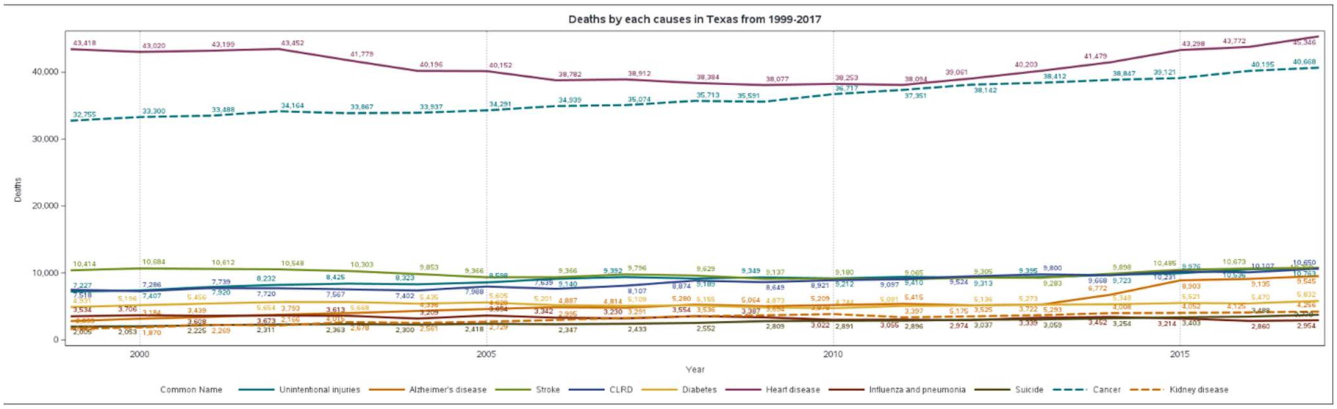


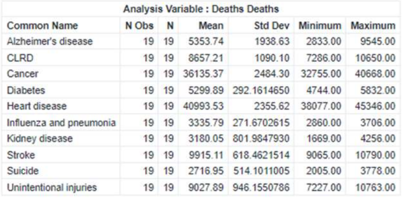
Statistical Report of deaths by each disease in Florida over the period of 1999-2017

Analysis Variable : Deaths Deaths

Common Name	N Obs	N	Mean	Std Dev	Minimum	Maximum
Alzheimer's disease	19	19	4833.47	1168.96	3059.00	7155.00
CLRD	19	19	10107.00	1193.69	8645.00	12619.00
Cancer	19	19	41193.84	1946.78	38478.00	45131.00
Diabetes	19	19	5068.58	441.2742050	4357.00	6172.00
Heart disease	19	19	45400.11	3392.21	40919.00	51434.00
Influenza and pneumonia	19	19	2764.42	390.2750192	2246.00	3365.00
Kidney disease	19	19	2711.84	478.5053423	1846.00	3297.00
Stroke	19	19	9741.89	1261.83	8395.00	12602.00
Suicide	19	19	2664.95	380.0103030	2029.00	3227.00
Unintentional injuries	19	19	8854.53	1777.17	5961.00	13059.00

Texas:



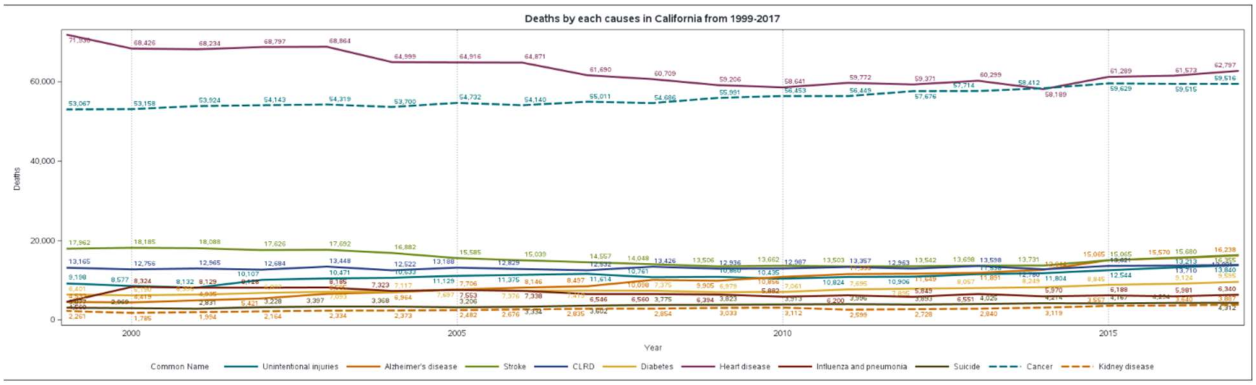


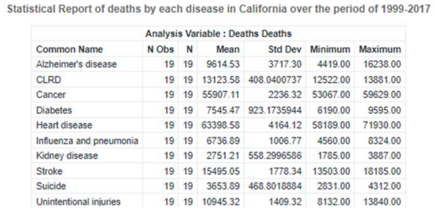
Statistical Report of deaths by each disease in Texas over the period of 1999-2017

Analysis Variable : Deaths Deaths

Common Name	N Obs	N	Mean	Std Dev	Minimum	Maximum
Alzheimer's disease	19	19	5353.74	1938.63	2833.00	9545.00
CLRD	19	19	8657.21	1090.10	7286.00	10650.00
Cancer	19	19	36135.37	2484.30	32755.00	40668.00
Diabetes	19	19	5299.89	292.1614650	4744.00	5832.00
Heart disease	19	19	40993.53	2355.62	38077.00	45346.00
Influenza and pneumonia	19	19	3335.79	271.6702615	2860.00	3706.00
Kidney disease	19	19	3180.05	801.9847930	1669.00	4256.00
Stroke	19	19	9915.11	618.4621514	9065.00	10790.00
Suicide	19	19	2716.95	514.1011005	2005.00	3778.00
Unintentional injuries	19	19	9027.89	946.1550786	7227.00	10763.00

California:

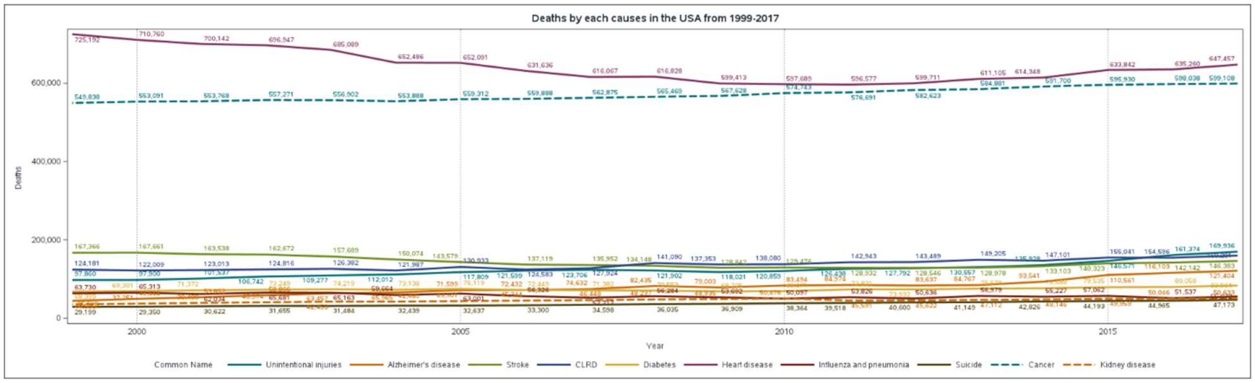


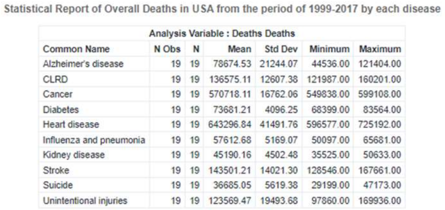


Statistical Report of deaths by each disease in California over the period of 1999-2017

Analysis Variable : Deaths Deaths						
Common Name	N Obs	N	Mean	Std Dev	Minimum	Maximum
Alzheimer's disease	19	19	9614.53	3717.30	4419.00	16238.00
CLRD	19	19	13123.58	408.0400737	12522.00	13881.00
Cancer	19	19	55907.11	2236.32	53067.00	59629.00
Diabetes	19	19	7545.47	923.1735944	6190.00	9595.00
Heart disease	19	19	63398.58	4164.12	58189.00	71930.00
Influenza and pneumonia	19	19	6736.89	1006.77	4560.00	8324.00
Kidney disease	19	19	2751.21	558.2996586	1785.00	3887.00
Stroke	19	19	15495.05	1778.34	13503.00	18185.00
Suicide	19	19	3653.89	468.8018884	2831.00	4312.00
Unintentional injuries	19	19	10945.32	1409.32	8132.00	13840.00

From the graphs above, it can b easily concluded that Heart Diseases, Cancer, and Stroke are the top 3 diseases causing highest death counts in Florida, California, and Texas. To check if the same is applicable on the nationwide data, we can use the PROC MEANS procedure along with PROC SGPLOT to get the death counts by each of the 10 common diseases over the period of 18 years (1999-2017). The data table *deaths_usa_each_cause* can be used to get desired result.





Statistical Report of Overall Deaths in USA from the period of 1999-2017 by each disease

Analysis Variable : Deaths Deaths						
Common Name	N Obs	N	Mean	Std Dev	Minimum	Maximum
Alzheimer's disease	19	19	78674.53	21244.07	44536.00	121404.00
CLRD	19	19	136575.11	12607.38	121987.00	160201.00
Cancer	19	19	570718.11	16762.06	549838.00	599108.00
Diabetes	19	19	73681.21	4096.25	68399.00	83564.00
Heart disease	19	19	643296.84	41491.76	596577.00	725192.00
Influenza and pneumonia	19	19	57612.68	5169.07	50097.00	65681.00
Kidney disease	19	19	45190.16	4502.48	35525.00	50633.00
Stroke	19	19	143501.21	14021.30	128546.00	167661.00
Suicide	19	19	36685.05	5619.38	29199.00	47173.00
Unintentional injuries	19	19	123569.47	19493.68	97860.00	169936.00

It can be seen in the graph and the MEAN procedure table that the top 2 diseases (Heart Attack and Cancer) present in the 3 states are also present in the nationwide count. The stroke disease has been replaced by Unintentional injuries (Accidents).

Step 5: Exporting Data

The data results have been exported to various output platforms such as RTF, PDF, and EXCEL. I have used ODS EXCEL, ODS RTF and ODS EXCEL procedure to export the results. All the result sets code has been written between ODS (EXCEL/RTF/PDF) and ODS CLOSE statements.

SUMMARY

The aim to identify the diseases causing the highest death counts is finally achieved. As seen in the analyzing and reporting section, The Heart Attack and Cancer are the most harmful diseases which occupied top 2 position for 18 years straight and caused the most death counts in each state. It was also observed that the total death count is keep on getting increased year by year. California, Florida and Texas are the three states where the death counts are the highest.

This report can work as a baseline for the future reports and authorities to spread awareness about these deadly diseases and their harmful effects on human population.

During the analysis process, it was evident that visualizing the data using graphs along with statistical procedure is the best way to obtain desired results.

Though the dataset was almost perfect as I did not have spend much time on pre-processing, it would have been better if it had frequent data of past 3 years. Also, additional columns such as active cases and recovered patient counts would have enriched the analysis.

As a result, I would like to conclude that the findings were really useful and almost show a overall trends of fatal diseases and its impact on the people of United States.

BIOGRAPHY

My name is Ruchit Tripathi, and I am a graduate student pursuing M.S. in Computer Science at Clemson University. I have been in the Data and Analytics domain for almost 4 years and wish to keep myself updated with the trending topics in the analytics space. Before coming to Clemson, I was working in the consulting domain as a Data Analytics Consultant. After my graduate studies, I wish to apply my analytics learning in the supply chain and pharmaceutical domain.

REFERENCES

- https://data.cdc.gov/NCHS/NCHS-Leading-Causes-of-Death-United-States/bi63-dtpu
- National Center for Health Statistics. Vital statistical data available. Mortality multiple cause files. Hyattsville, MD: National Center for Health Statistics. Available from: https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm
- Murphy SL, Xu JQ, Kochanek KD, Curtin SC, and Arias E. Deaths: Final data for 2015. National vital statistics reports; vol 66. no. 6. Hyattsville, MD: National Center for Health Statistics. 2017. Available from: https://www.cdc.gov/nchs/data/nvsr/nvsr66/nvsr66_06.pdf.
- https://blogs.sas.com/content/iml/2015/02/25/plotting-multiple-series-transforming-data-from-wide-to-long.html
- https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/grstatproc/n121lznfa1jnlvn1q95t0r5sd2gq.htm#n0439r26gfqvy9n1tzvs9ejj42rh
- https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2794043#:~:text=Because%20the%20data%20are%20publicly,require%20institutional%20review%20board%20review.&text=From%20March%202020%20to%20October%202021%2C%20heart%20disease%20(20.1%25),of%20death%20in%20the%20US.