

Learning from the Dead – Perspectives from the CDC Data

By

Aditya Damle

Sarika Dhoot

Tony(Yutong) Chen

Aditya Safi

Aasish Sipani

Bo Tang

**Dataset And Cleaning**

The dataset is composed of a main DeathRecords table, EntityAxisConditions and RecordAxisConditions tables showing the death related health issues, as well as other tables containing the data definitions. Before analyzing the dataset, the first step is to clean the dataset for easier analysis. The target is mainly the DeathRecords table since it is our main table for analysis and the methods we used include: 1. Removing redundant columns that have less significance 2. Unifying metrics measurement 3. Removing some of the rows that contains unknown data while limiting the impact on the integrity of the full dataset, etc.

The technology we used include SQL, Pig, Hive, Spark, Tableau, VM - Hadoop, as well as the AWS platform

**Analysis**

The analyses are guided by the following questions:

I. **Are there systematic patterns to the health issues that associated with deaths across 50 states? Are there systematic patterns across race, gender, and education in terms of the reasons of death?**

**Method:Hive**

1. **What are the top 3 causes of deaths and the relative number of deaths?**

Result: The top is “Atherosclerotic heart disease” with count of 161951, followed by “Malignant Neoplasm: Bronchus or lung, unspecified”, with count of 154858, followed by “Unspecified Dementia”, with count of 122019.

1. **What are the relationship between age, conditions and death?**

Result: people in the age zone of 85 years and over with conditions of Unspecified dementia, Alzheimer disease- unspecified and Atherosclerotic heart disease have the highest number of deaths among all the combinations.

1. **What are the leading causes of death of black people? Comparing to white people? (See Appendix A-1)**

Interpretation: causes seem to match between Black, White and overall. However, black people also suffer from acute myocardial infarction, which is heart attack. Black people has more deaths caused by heart disease, which may relate to high fat diet patterns.

1. **What are the leading causes of death of people’s age between 15 to 24?**

Result: The top 3 leading causes of death are “assault by other and unspecified firearm discharge”, followed by “person injured in unspecified motor-vehicle accident”, followed by “intentional self-harm by hanging.” So gun control could be justified because of this finding.

1. **What are the leading causes of death between the genders and how do they vary depending on race?**

Result: The top 3 leading causes of death among female includes dementia but this is not the case for males. This is because white males and black males constitute only about 30% of death caused by dementia while females account for 70% in both the races.

**Method: Pig On AWS**

1. **What are the percentages of suicide deaths by marital status? (See appendix A-2)**

Result: The single has the highest suicide percentage.

1. **What are the leading causes of deaths of infants with age < 1?**

Result: The top is “Extreme immaturity”, followed by “Sudden infant death syndrome”, followed by “Other ill-defined and unspecified causes of mortality.”

1. **What are the numbers of deaths in “Sports and athletics area” by four seasons? (See Appendix A-3)**

Result: Spring and fall has the higher numbers of deaths in “Sports and athletics areas” than summer and winter, which makes sense because spring and fall have good weathers.

1. **Special Finding:** What are the estimated life expectancies of male and female (with additional comparison while being married) (See Appendix A-4 for more information)

Result: Sometimes the result can be deceptive. In general, women tend to live a longer life than men. However, when the marital status is married, we might come up with the wrong conclusion based on the data aggregation (as shown on the graph on the right in Appendix A-4) – men actually have a higher life expectancy than women. This is because most women who died after the men are actually classified as “widowed” instead of “married” so that they are not counted in the estimated life expectancy calculation for the graph on the right.

II. **Where do gun-related death occur? When do they occur? What are correlating factors when guns or shootings are implicated?**

* Across all the deaths about 1.263% of the deaths are due to gun related issues.
* An interesting find in the case of resident status; the percentage of gun related deaths is much higher for foreign residents compared to Interstate and Intrastate residents.
* The percentage of autopsy is 67.5% in the gun related deaths compared to only 7.7% across all the deaths. This is a correlating factor when gun related deaths are implied.
* 62.6% of gun related deaths occur at home while 11.22% are occurring on streets and highways. This indicates that safety within homes is a concern. The percentage of deaths on streets and highways is also quite high which points to the fact that these places are a danger zone and security should be increased in these places.
* For season wise deaths we get following values: Winter: 23.505%, Spring: 24.961%, Summer: 26.09%, Fall: 25.43688%. This analysis is not very interesting as there more or less equitable distribution. For month wise we get highest number of gun related deaths in the month of August.

III. **Develop a prediction model for whether a person is cremated or buried?(Method: PySpark on AWS)**

1. Naïve Bayes Model. (See Appendix B-1)

Result: Accuracy of 0.565

1. Logistic Regression Model. (See Appendix B-2)

Result: Accuracy of 0.653

Additional Notes: Our first approach to tackle this process was to clean the dataset in our hands. We transformed the column “MethodofDisposition” by plugging all cremated entries as 1 and all the other values as 0. Since we were having many categorical variables, we factored them into the appropriate levels.  The data is split into 80% training and 20% validation.

**Policy Suggestions**

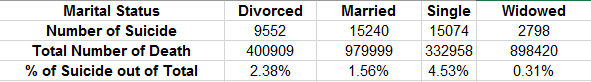
* As we could see, more single people tend to commit suicide compared to widowed, married or divorced people, so the government might want to create support groups for the single people who are stressed.
* Most of the deaths at “Sports and athletics area” are in the seasons of spring and fall, so the government can try to renovate the sporting facilities at the start of those seasons as well as sending more supervising personnel to those areas.
* The top cause of death for infants who are under one year of age is “Extreme immaturity.” So the government might want to create programs focusing on the pre-natal care to reduce the amount of infant death.
* The number of patients getting affected by Alzheimer's occur in the Age group of 85 years of above. This is useful to recognize health policy decisions of what efforts may be targeted to what age group for special diseases which are a cause of concern for the government.

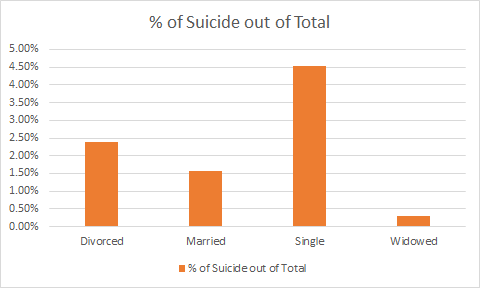
**Appendix**

**A-1**

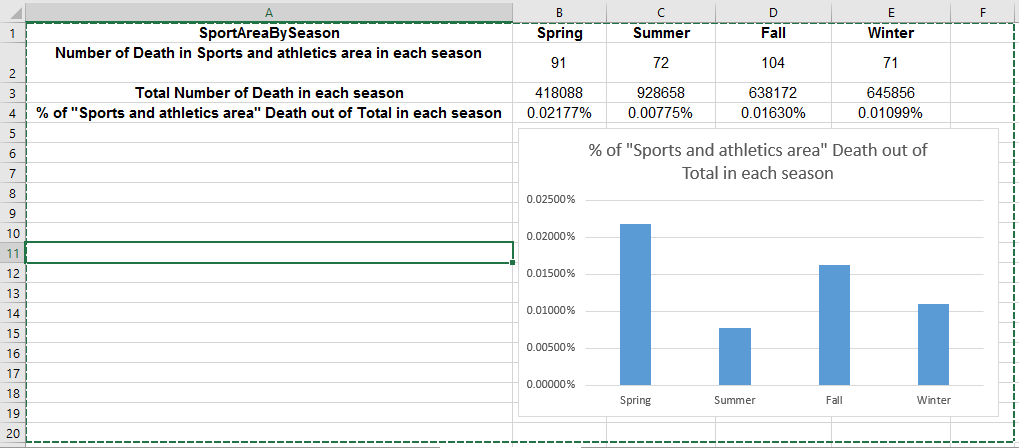
|  |  |  |
| --- | --- | --- |
| Leading Cause of Death | Black | White |
| #1 | Malignant Neoplasm: Bronchus or lung | Atherosclerotic heart disease |
| #2 | Atherosclerotic heart disease | Malignant Neoplasm: Bronchus or lung |
| #3 | Acute myocardial infarction | Unspecified dementia |

**A-2**

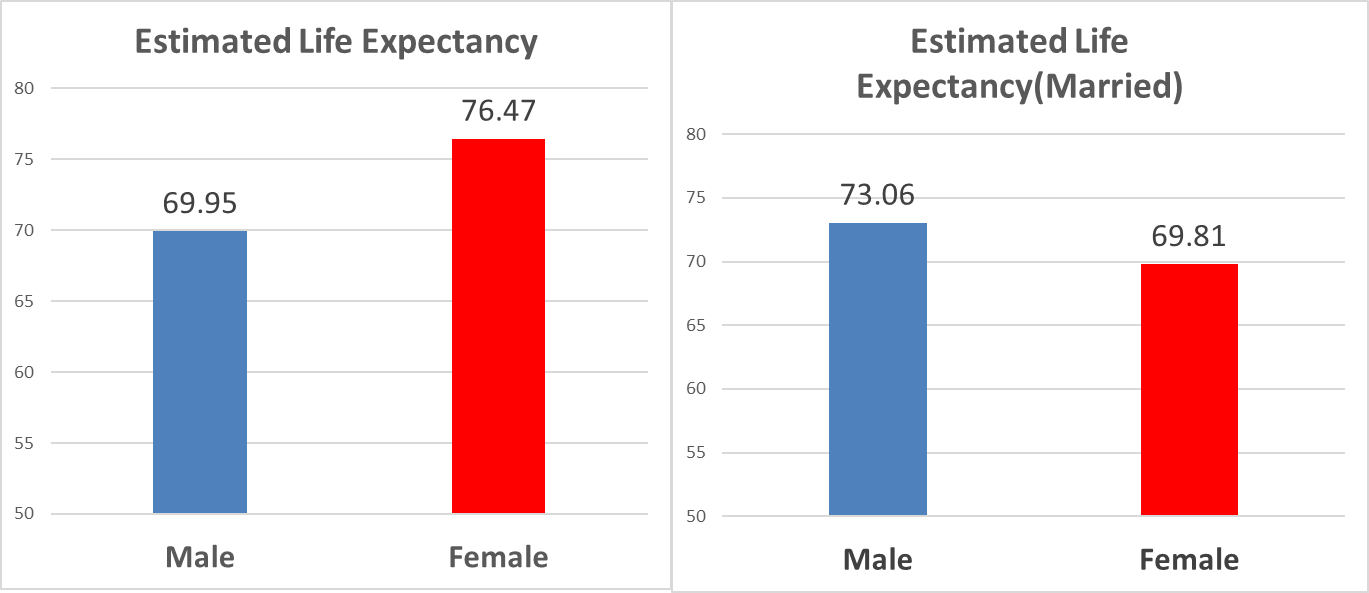




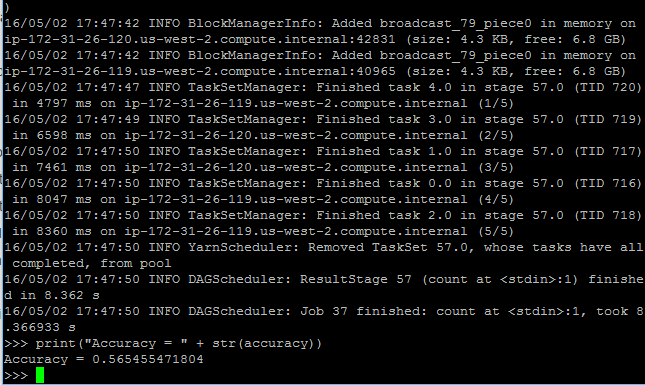
**A-3**



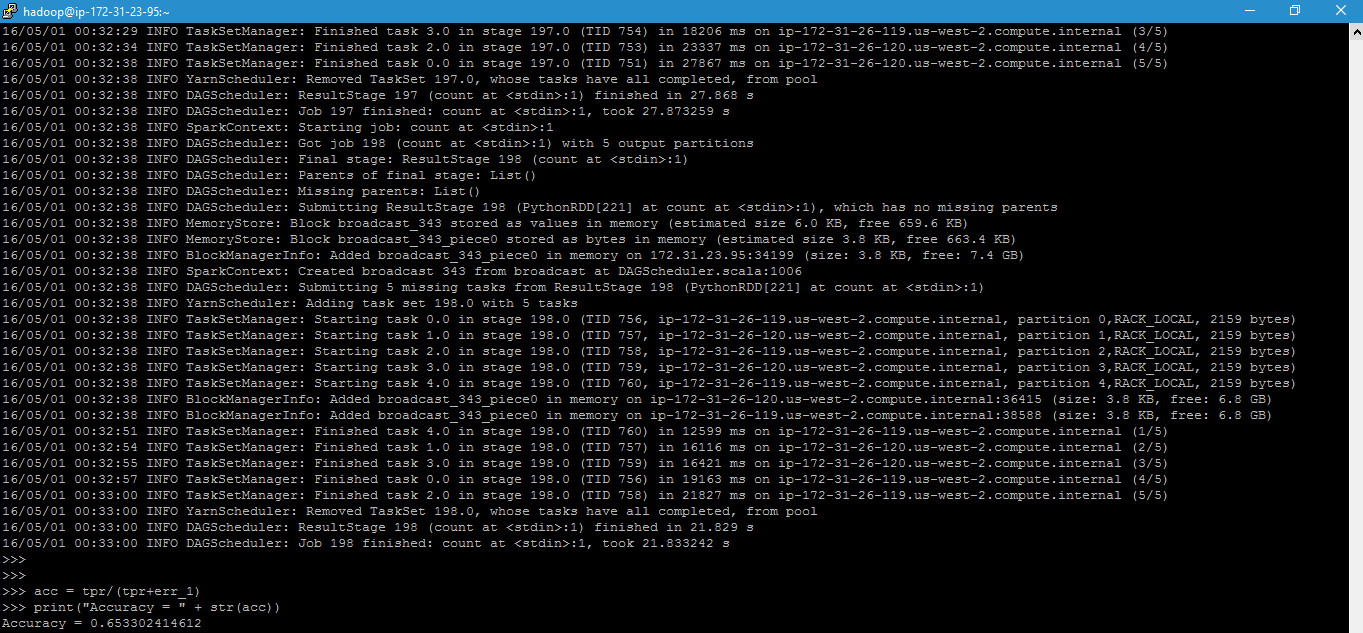
**A-4**



**B-1**

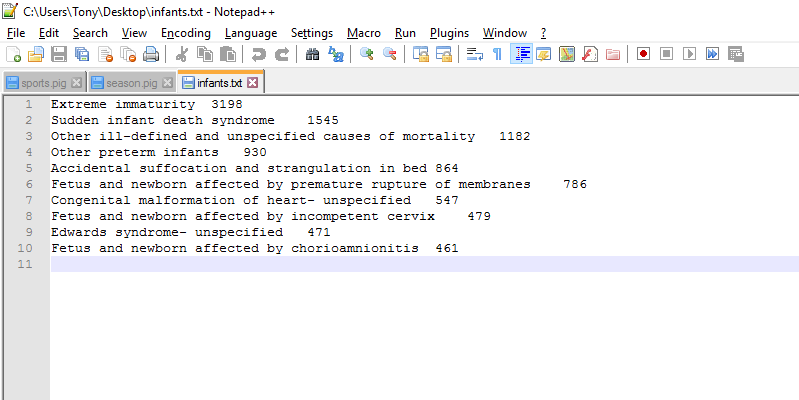
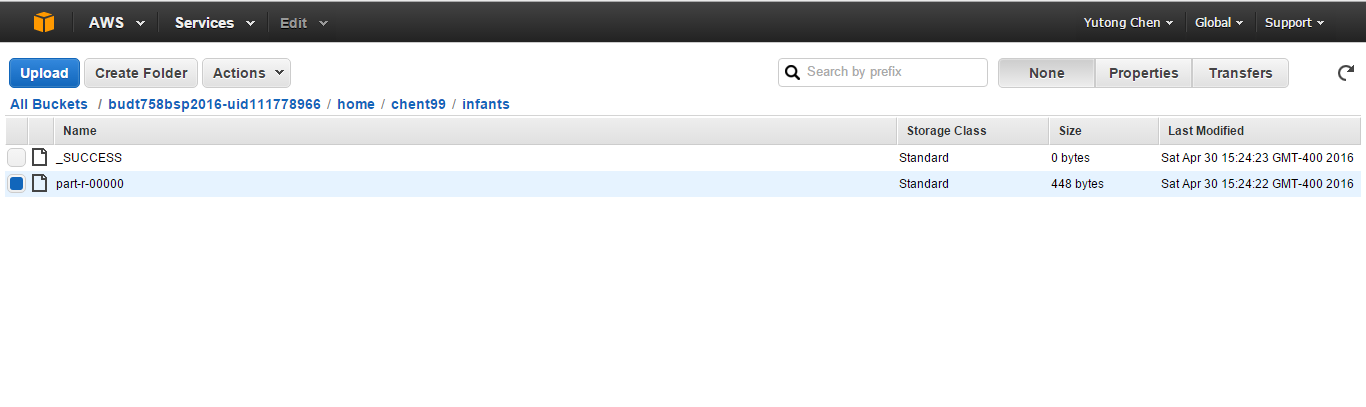


**B-2**



Sample Processing Screenshots

Running a Pig job:



Running a Hive Query:

