**Project Report (DSTEAM05)**

**Big Data Analytics for Healthcare**

**Team Members:**

S Venkatesh 21R11A0550

K Mysanthosh 21R11A0579

K Hemanth Reddy 21R11A0582

P Saikumar 21R11A0538

**Introduction**

In today's healthcare landscape, the shift towards digitalization of patient medical records marks a significant turning point, ushering in an era driven by data insights and analysis. Electronic Health Records (EHR) stand as a pivotal foundation, offering numerous advantages such as seamless data sharing, improved access to patient health histories, and robust analytical capabilities for extracting valuable insights from extensive healthcare data repositories. The emergence of big data analytics has further transformed the healthcare sector, presenting unprecedented opportunities to uncover concealed patterns and understandings from vast and diverse data sets.

With a focus on harnessing the potential of Hadoop Distributed File System (HDFS) and MapReduce paradigms, our project aims to unearth valuable insights from a dataset spanning U.S. mortality trends dating back to 1900. By leveraging distributed computing and advanced analytical techniques, we seek to illuminate the factors influencing mortality rates across various demographics and geographic regions. The dataset utilized in this endeavor originates from death certificates filed across the United States, offering a comprehensive overview of age-adjusted death rates for major causes of mortality over the past century.

Through a series of exploratory analyses encompassing population standard deviation, variance, and demographic-based aggregation, our project aims to reveal fundamental trends and patterns underlying mortality dynamics in the U.S. By connecting to a pre-configured Hadoop environment and utilizing tools such as Hive and Pig for data manipulation and analysis, our team strives to showcase the effectiveness of big data analytics in extracting insights from intricate healthcare datasets.

With a meticulous methodology involving data loading, transfer, and exploration, our project seeks to empower stakeholders with actionable insights for informed decision-making and the formulation of healthcare policies.

**Methodology**

1. Begin by connecting to a pre-configured image of the Hadoop ecosystem, such as Cloudera or Hortonworks HDP, either on a virtual machine or any cloud platform.

2. Once connected, proceed to load the data into the Hadoop Distributed File System (HDFS).

3. Transfer the data from HDFS to a Hive table for further processing.

4. Initiate the initial data exploration phase by executing specific analysis queries:

- Determine the Population Standard Deviation of ‘Deaths’ for the years 2013 and 2014.

- Calculate the Population Variance of ‘Deaths’ for the year 2010.

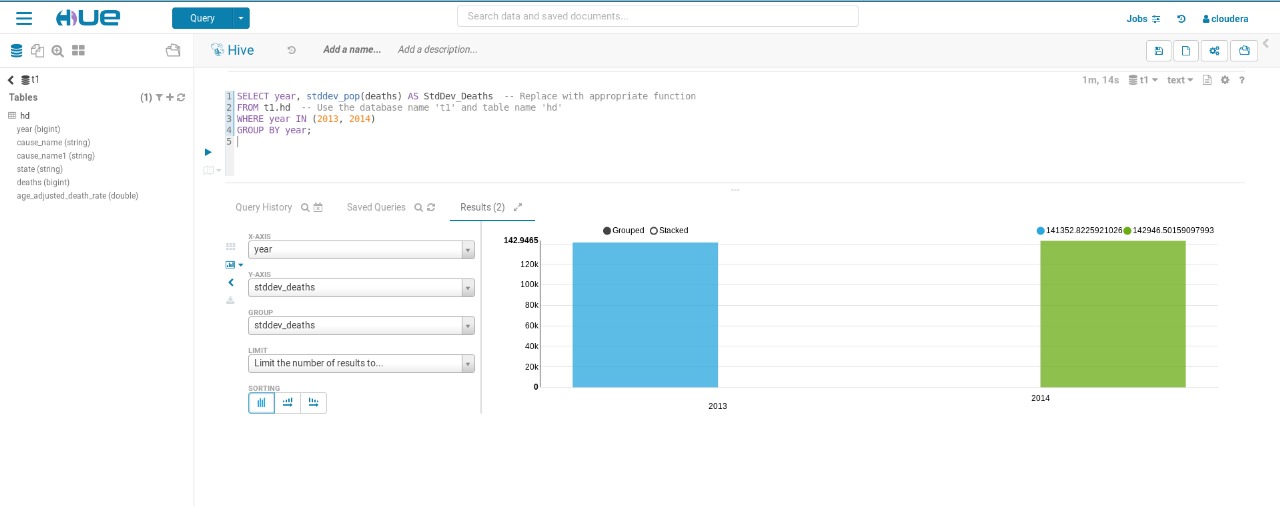
- Compute the Sample Standard Deviation of ‘Deaths’ for the year 2003.

- Compute the Sample Variance of ‘Deaths’ for the year 2009.

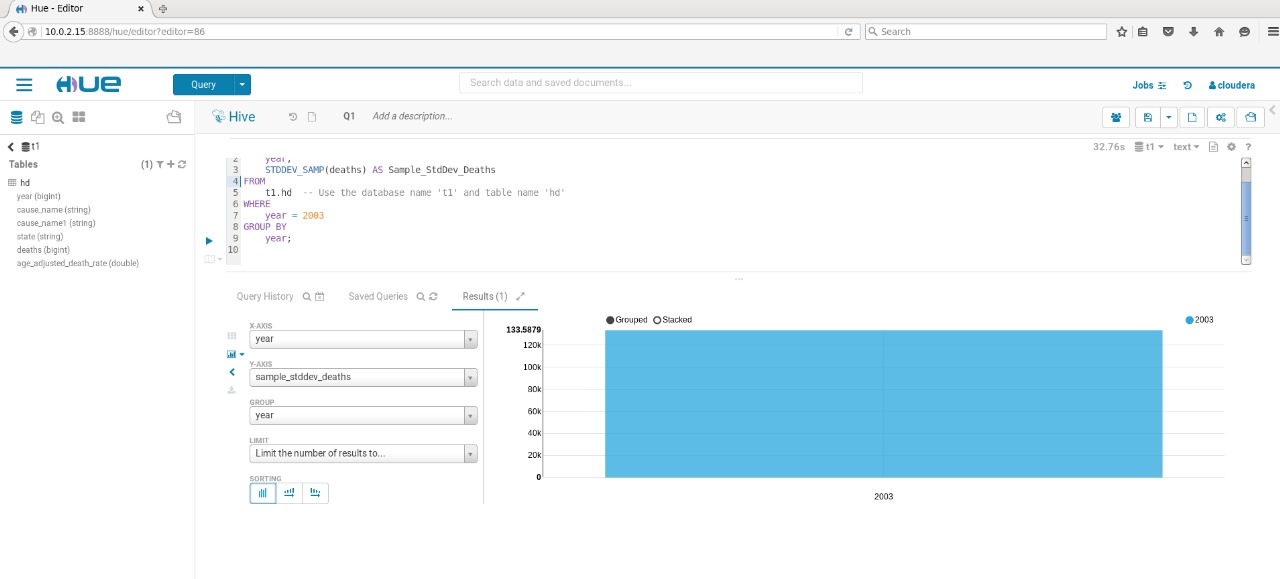
- Calculate the Variance of deaths for the year 2012.

5. Additionally, gather insights on the number of deaths, total deaths, average deaths, minimum and maximum number of deaths, grouped by ‘State’, and order the results by ‘State’.

6. To optimize processing for the above query, create a partitioned table in Hive.

**Query 1.**  
**Observation:**

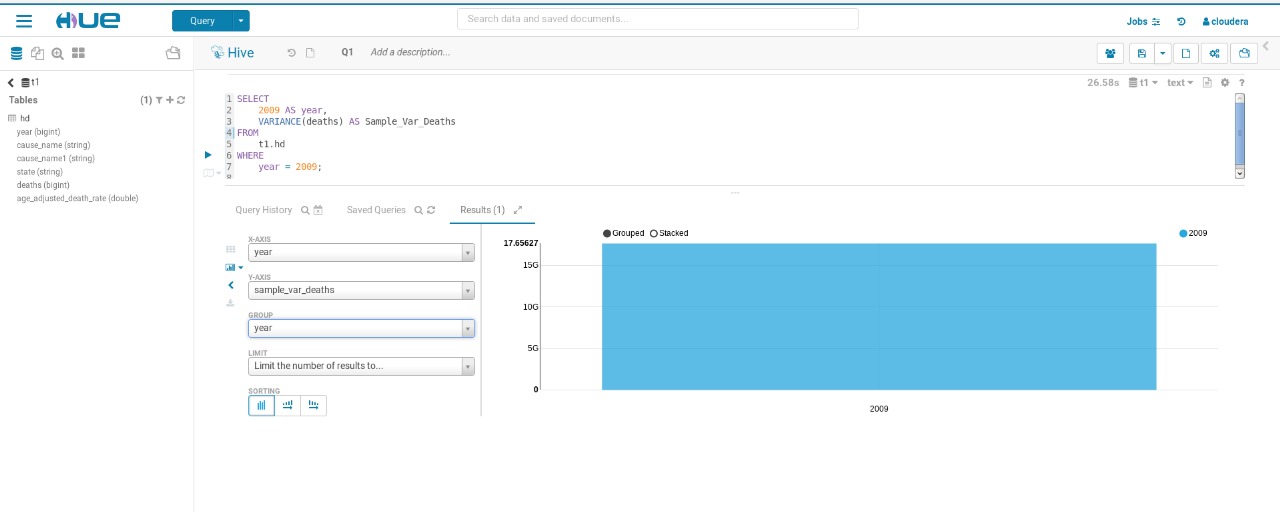
Based on the graph, the Population Standard Deviation of ‘Deaths’ for 2013 and 2014 shows a slightly higher value for 2014 compared to 2013. Specifically, the Population Standard Deviation for 2013 is approximately 141352.8225921026, whereas for 2014, it stands at approximately 142946.50159097993.

**Query 2.**  


**Observation:**

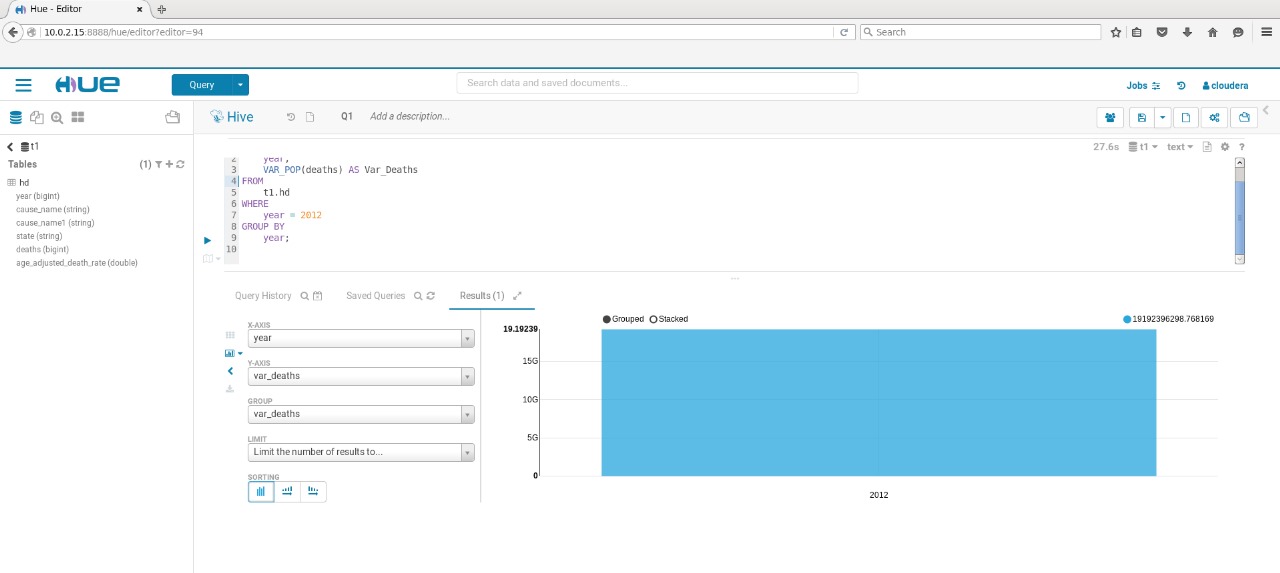
According to the graph, it appears to represent the analysis of the sample standard deviation of deaths for the year 2003. Specifically, it indicates that there were 1,335,879 deaths recorded for the year 2003. However, there seems to be a labeling discrepancy on the graph: the x-axis label incorrectly states "year" instead of specifying "2003," while the y-axis label appears truncated and might refer to the standard deviation of deaths.

**Query 3.**



**Observation:**

The graph illustrates the sample variance of deaths for the year 2009. Specifically, the y-axis labels denote the years, while the x-axis represents the variance values. In the context of the year 2009, the variance value of deaths is observed to be 17.65627

**Query4.**

**Observation:**

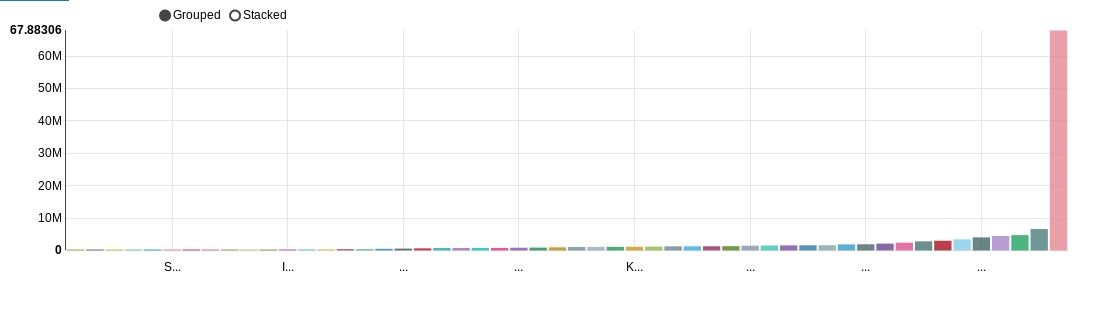
The graph illustrates a single data point representing the year 2012, with a variance value labeled as "Var Deaths" and measured at 19.19239. Notably, when compared to the variance observed in 2009, there is an increase in variance in the year 2012, indicating potential fluctuations or shifts in mortality trends over time.

**Query 5.**

Number of the countries



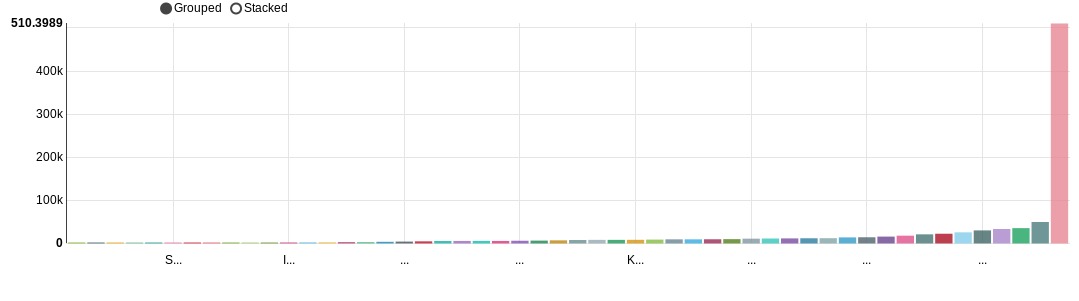
Number of deaths in each country

Observation:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| state | total\_deaths | sum\_deaths | avg\_deaths | min\_deaths | max\_deaths |
| Alabama | 133 | 1285165 | 9662.895 | 772 | 53238 | |
| Alaska | 133 | 94562 | 710.9925 | 24 | 4494 | |
| Arizona | 133 | 1279352 | 9619.188 | 683 | 57758 | |
| Arkansas | 133 | 793194 | 5963.865 | 430 | 32588 | |
| California | 133 | 6635282 | 49889.34 | 4419 | 268189 | |
| Colorado | 133 | 851664 | 6403.489 | 525 | 38063 | |
| Connecticut | 133 | 791835 | 5953.647 | 449 | 31312 | |
| Delaware | 133 | 206923 | 1555.812 | 107 | 9178 | |
| District of Columbia | 133 | 135110 | 1015.865 | 53 | 6076 | |
| Florida | 133 | 4735234 | 35603.26 | 2246 | 203636 | |
| Georgia | 133 | 1882517 | 14154.26 | 1080 | 83098 | |
| Hawaii | 133 | 260125 | 1955.827 | 109 | 11390 | |
| I13 | 988 | NULL | NULL | NULL | NULL | |
| Idaho | 133 | 306459 | 2304.203 | 189 | 14011 | |
| Illinois | 133 | 2819804 | 21201.53 | 1908 | 109721 | |
| Indiana | 133 | 1569270 | 11799.02 | 962 | 65597 | |
| Iowa | 133 | 777774 | 5847.925 | 504 | 30530 | |
| Kansas | 133 | 677697 | 5095.466 | 511 | 27063 | |
| Kentucky | 133 | 1147055 | 8624.474 | 728 | 48212 | |
| Louisiana | 133 | 1129495 | 8492.444 | 683 | 45804 | |
| Maine | 133 | 357903 | 2691 | 205 | 14676 | |
| Maryland | 133 | 1197950 | 9007.143 | 681 | 49926 | |
| Massachusetts | 133 | 1487938 | 11187.5 | 1007 | 58803 | |
| Michigan | 133 | 2409020 | 18112.93 | 1431 | 97602 | |
| Minnesota | 133 | 1072574 | 8064.466 | 529 | 44371 | |
| Mississippi | 133 | 784322 | 5897.158 | 356 | 32280 | |
| Missouri | 133 | 1522983 | 11451 | 914 | 61876 | |
| Montana | 133 | 240519 | 1808.414 | 147 | 10200 | |
| N17-N19 | 988 | NULL | NULL | NULL | NULL | |
| Nebraska | 133 | 421469 | 3168.94 | 266 | 16878 | |
| Nevada | 133 | 519505 | 3906.053 | 174 | 24657 | |
| New Hampshire | 133 | 288617 | 2170.053 | 187 | 12504 | |
| New Jersey | 133 | 1923714 | 14464.02 | 1041 | 74846 | |
| New Mexico | 133 | 420067 | 3158.398 | 248 | 18673 | |
| New York | 133 | 4047365 | 30431.32 | 1357 | 159927 | |
| North Carolina | 133 | 2131964 | 16029.8 | 1456 | 93157 | |
| North Dakota | 133 | 164776 | 1238.917 | 120 | 6415 | |
| Ohio | 133 | 3012874 | 22653.19 | 1743 | 123648 | |
| Oklahoma | 133 | 987388 | 7423.97 | 546 | 40452 | |
| Oregon | 133 | 886996 | 6669.143 | 382 | 36624 | |
| Pennsylvania | 133 | 3448894 | 25931.53 | 2192 | 135656 | |
| Rhode Island | 133 | 263695 | 1982.669 | 161 | 10246 | |
| South Carolina | 133 | 1112967 | 8368.173 | 690 | 49441 | |
| South Dakota | 133 | 198525 | 1492.669 | 134 | 7996 | |
| Tennessee | 133 | 1626698 | 12230.81 | 944 | 70096 | |
| Texas | 133 | 4472578 | 33628.41 | 2833 | 198106 | |
| Unintentional injuries | 988 | NULL | NULL | NULL | NULL | |
| United States | 133 | 67883063 | 510399 | 44536 | 2813503 | |
| Utah | 133 | 383478 | 2883.293 | 245 | 18035 | |
| Vermont | 133 | 146703 | 1103.03 | 52 | 6007 | |
| Virginia | 133 | 1621752 | 12193.62 | 917 | 68579 | |
| Washington | 133 | 1351811 | 10163.99 | 578 | 56995 | |
| West Virginia | 133 | 582287 | 4378.098 | 314 | 23276 | |
| Wisconsin | 133 | 1300064 | 9774.917 | 888 | 52681 | |
| Wyoming | 133 | 117150 | 880.8271 | 89 | 4778 | |
| Y87.0)" | 988 | NULL | NULL | NULL | NULL | |
|  |  |  |  |  |  | |
|  |  |  |  |  |  | |

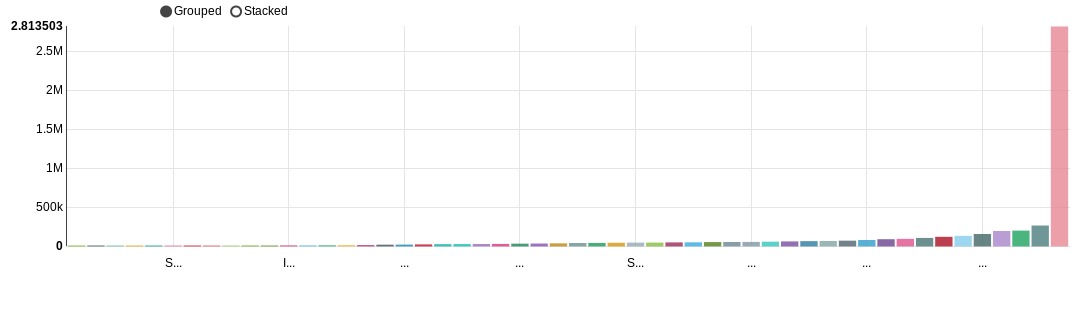
From the above graph and table, the highest number of the death are for the United States of America and the least number of the deaths are for the Alaska

AVERAGE DEATHS



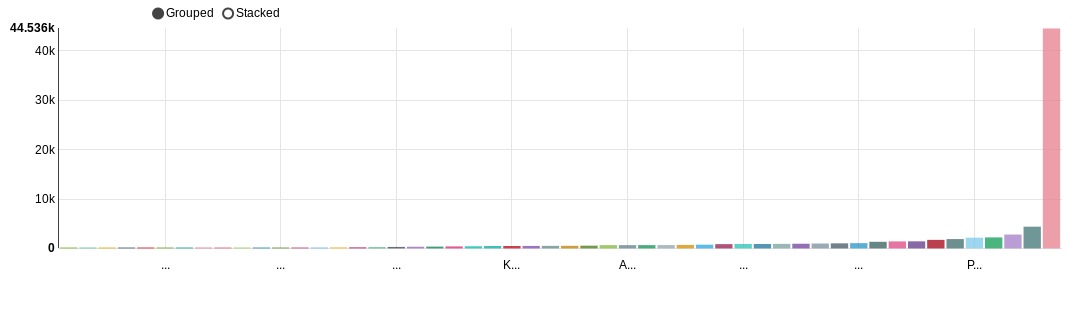
From the above graph the United States has the highest average deaths as 510399

MAXIMUM DEATHS



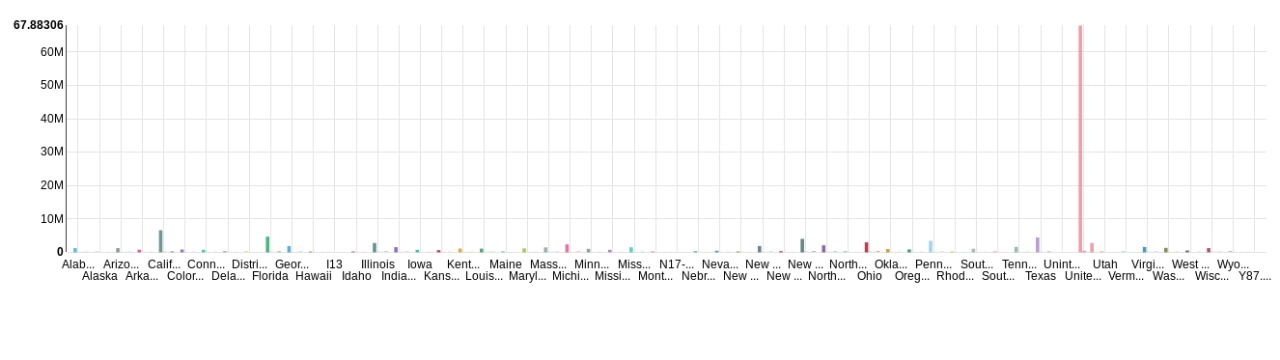
From the above graph we can say that united states of America has highest maximum deaths are 2813503

MINIMUM DEATHS



From the above graph Alaska has minimum number of the deaths of 24

CONSERDING ALL THE PARAMETERS



From the above graph united states has the maximum number of deaths , highest number of deaths whereas Alaska has minimum number of deaths