Impact of Mortality on GDP in the USA

Sankara Subramanian Venkatraman
School of Computing
National College of Ireland
Dublin, Ireland
x18179541@student.ncirl.ie

Iswaria Nagarajan
School of Computing
National College of Ireland
Dublin, Ireland
x18183379@student.ncirl.ie

Tejas Sanjay Shinde
School of Computing
National College of Ireland
Dublin, Ireland
x18180159@student.ncirl.ie

Shreya Merkaje Ravi
School of Computing
National College of Ireland
Dublin, Ireland
x18190910@student.ncirl.ie

***Abstract*—This study examines the impact of GDP on mortality rate in the United States of America by considering various causes such as murder, violent crimes, HIV/AIDS and natural causes such as cancer, heart disease and stroke. A collective analysis is done by combining all these causes and life expectancy is calculated for all the states of America. A combination of unstructured and structured data sources is used to conduct this analysis using a variety of technologies including relational and NoSQL databases, cloud computing platform and data processing engines. Individual analysis of each data source is performed with the help of data mining methodologies such as data integration, data pre-processing, data cleaning and visualized for better interpretation. Multiple regression is performed on the dataset to determine the spurious correlation by combining the mortality due to different causes and the GDP for each state. From the analysis, assumptions of the regression are tested, and finally model is developed to predict the GDP based on significant factors.**

***Index Terms*—OpenStack, GDP, MongoDB, PostgreSQL, Spark, Scala, Python, Multiple Regression**

I. Introduction

A study on the influence of economy and mortality on one another has been illustrated many times in the past. However, the hypothesis has neither been proved or rejected. In the recent past, gathered evidence has suggested that an increase in income and personal well-being are inversely proportional to each other. From this, we can infer that growth in the economy causes more death[1]. GDP and life expectancy have a strong relationship between them which infers that more the money better is life expectancy. Christopher Ruhm illustrated that when the US economy was rising, people faced more medical issues and died faster but when the economy was down, people lived longer[2].

In our project, we try to understand deaths due to drug usage per city and different age groups of the people who tend to use drugs. HIV data is taken to understand the observed deaths per state. The relationship between observed and expected deaths due to natural causes is also drawn, along with identifying the safest and unsafe states in the US-based on the murder rate. By taking all the above-mentioned causes for death we try to see its impact on economic growth.

We have seen many papers/studies on how growth in the economy influences life expectancy. It is often assumed that the country with a higher economy has a better life expectancy. However, we never investigate how the mortality rate affects the economy. Hence, we will try to analyze different causes of death such as drugs, HIV, natural causes (cancer, heart disease and unexpected injury) and murder based on different states in the US and analyze the impact it has on GDP.

II. Related Work

A. Research on Mortality

A considerable number of researches is done in this area of analysing the mortality rate. The researches in the area of event recognition where different approaches have been used for various objectives are discussed below.

[1] has analysed the factors causing death due to cardiac arrest and has also predicted the death due to cardiac arrests at an early stage i.e. one hour before its occurrence using heart rate variability analysis. ECG signals from only one database have been used for healthy patients and Sudden Cardiac Death patients in order to predict the death using K-NN classifier which strongly predicted the death of patients at an early stage.

Statistical analysis of factors causing death cases at hospitals in China was done by [2]. A detailed study is done for three years (2007-2010) by considering factors such as death rate, gender, age, occupation and causes of death. These factors were analysed to reduce the death rate, improve the health conditions and to strengthen disease prevention. This study analysed that the death rate dropped every year and the men were higher than the women.

A state-of-art approach [3] makes use of a set of procedures by analysing the National Violent death reporting database in order to predict the types of violent death that have occurred. A series of data mining methodologies were utilized to conduct

[1]https://www.weforum.org/agenda/2016/10/the-relationship-between-gdp-and-life-expectancy-isnt-as-simple-as-you-might-think

[2]https://academic.oup.com/qje/article-abstract/115/2/617/1840483

this analysis and they have introduced a new set of association rules that could be used by the government to predict deaths due to violent crimes. Stroke is considered to be the fifth leading cause of death in the United States of America.

This study [4] has proposed a novel approach called as real-time stroke early detection system in order to prevent deaths and to achieve a good clinical outcome. They have performed this analysis with the help of sleep data of healthy patients and stroke patients using KNN and SVM classifiers.

Cancer is one of the major causes of natural deaths. [5] has proposed an approach that causes death due to lung cancer. They have considered Malignant Associated Changes (MAC) which is an indicator of lung cancer. SVM classifiers are used to classify the MAC and non-MAC cells in microscopic sputum cytology images in order to prevent deaths due to lung cancer by detecting them at the earlier stage.

In a similar way, [6] has studied how the treatment interruptions during chemotherapy may be used to control the drug-resistance in Cancer and HIV/AIDS. They have made use of mathematical models and optimal control algorithms to frame an optimized treatment visit schedules in order to prevent disease progression and death.

In Japan, it is found that the effect of HIV is greater on the working population than among the senior population. This approach [7] has analysed the effect of HIV on the overall population of Japan considering the fertility rate of the population of Japan and has finally predicted that there is an overall decrease in population of Japan due to the deaths of HIV/AIDS.

B. Technologies

OpenStack is an open-source cloud-based platform used for deploying Information as a Service (IaaS) applications. This technology is used in this analysis to remotely access all the data sources and databases. The problems encountered on the effective use of the physical and virtual resources based on OpenStack cloud platform are discussed in [8]. Also, the structural and functional analysis of this cloud platform is focused upon.

The core problems of computer labs in universities and schools due to lack of open-source cloud-based platforms along with the tools of OpenStack is discussed in [9]. This study has analysed that the application of OpenStack tool can efficiently deploy the cloud of a university computer lab with good functionality and performance.

Big data has an enormous attraction in recent years. Analysis of Big Data is a very common requirement nowadays as data is increasing in size and number day by day. The tools used to analyze them should be proficient enough to avoid facing difficulties using analysis. It is challenging to analyze the bulk amount of emerging big data in a timely manner. [10] has made use of Apache Spark which is one of the emerging tools of industries to analyze the tweets of Twitter.

[11] focuses on the working sets of data across multiple parallel platforms using a variety of machine learning algorithms and data analysis tools using Apache Spark. They have made use of Spark's Resilient Distributed Datasets (RDD) to achieve this goal. An RDD is a collection of objects in a read-only format that can be sub-divided across a set of machines and can be re-constructed if a partition is lost.

[12] has focused upon the features, abstractions and components of Spark for implementing the machine learning algorithms on big data pipelines, graph analysis and stream processing. Also, some of the R & D directions for big data analytics using this data processing engine is discussed in this approach.

PostgreSQL is a relational database management system for structured data analysis. The unstructured data sources used in this analysis are converted to structured data through a series of data pre-processing techniques and are stored in PostgreSQL. [13] has dedicated a benchmark to PostgreSQL for data warehousing as a low-cost platform. These benchmarks include executing complex and nested queries, aggregations, multi-joins and others. With these benchmark analyses, they were able to figure out the outstanding issues and problems encountered in the execution.

Also, they have demonstrated that this RDBMS can be used as an alternative to data warehousing with some improvements in structures and optimizes. [14] has focused upon the design and implementation decisions of using PostgreSQL backend functions for the analysis of structured datasets.

MongoDB is a NoSQL database management system used to store and query unstructured data. [15] has presented a comparative study on relational and non-relational databases by considering MySQL and MongoDB respectively. This analysis has finally derived its result that MongoDB is more efficient when compared to MySQL because of its scalable performance metrics.

The comparative study of MySQL and MongoDB by evaluating their respective performances in the area of Super Market Management system is discussed in [16]. They have concluded that as the size of the data increases with an increase in the number of records, MongoDB shows a significant reduction in time taken for execution when compared to MySQL.

Hence it can be inferred that NoSQL databases can handle high transaction loads when compared to relational databases. Another paper [17], has discussed the variation in performance when trying to reduce the query execution time in MongoDB due to embedding and normalization.

III. Methodology

A. Data Sources

Our main objective is to analyze the impact Mortality has on the economy and hence we had to find as much data possible related to mortality from relevant sources. Also, since the economy is important for our study, we had to gather data on GDP for each state in the US. We have used 5 datasets as mentioned below:

1) Drug Death: Deaths due to drugs is sourced from[3]. Drug usage has been a concern all over the world. Even after many substances being considered illegal, we often read about cases of drug overdose and fatality. For our project, the dataset has been procured from the US government website with the help of an API key and contains data on death count due to drug usage per state and year. The data is semi-structured and hence we loaded it into MongoDB. Metadata of the dataset is provided in the ER diagram below.

2) HIV Death: Deaths due to HIV is sourced from[4]. Medical field has made an enormous advancement in the treatment of HIV virus but we still have a huge number of deaths. This dataset contains data on death count due to HIV for various years per state. This dataset has been procured from the US government as well. The data is semi-structured and metadata for this dataset is provided in the ER diagram below.

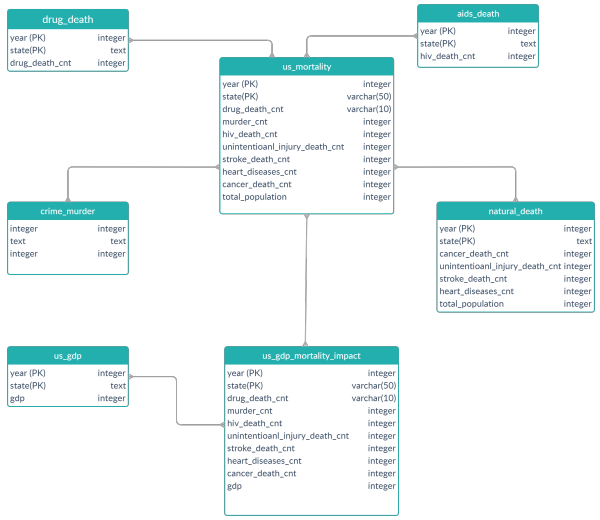

Fig. 1. ER Diagram

3) Natural Death: Deaths due to natural Causes is sourced from[5]. In this dataset, we have data related to deaths due to cancer, stroke, heart disease and unintentional injury. This data is semi-structured as well and attributes of the dataset can be found in the ER diagram.

4) Murder: Deaths due to violent crime: Murder is sourced from[6]. The dataset has the count of murders per state for various years. Data is structured and hence after pre-processing the data, is loaded into PostgreSQL.

5) GDP: GDP data is sourced from[7]. Gross Domestic Product measures the economic value of a country. Our dataset which has been procured from kaggle has data on GDP for individual states over various years.

Datasets related to drugs, HIV and natural death were procured from the US government website with the help of API in Python. Since the data is in JSON format which is semi-structured we have loaded them into MongoDB.

For our data cleaning, we have loaded our data from MongoDB to Pandas framework. Post the data cleaning we loaded the data frames into PostgreSQL. Murder and GDP datasets are structured and hence after the ETL process, we have loaded them on to PostgreSQL.

The four datasets related to deaths are joined using common attributes (Year and State) in spark. The combined table (US_Mortality) consists of all the causes of death, year, state its respective population. This table is in PostgreSQL and is joined with the GDP table to obtain our final table **us_gdp_mortality_impact**.

B. Architecture

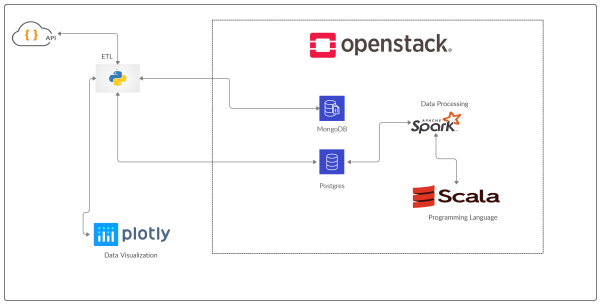

Fig. 2. Data Lake Architecture

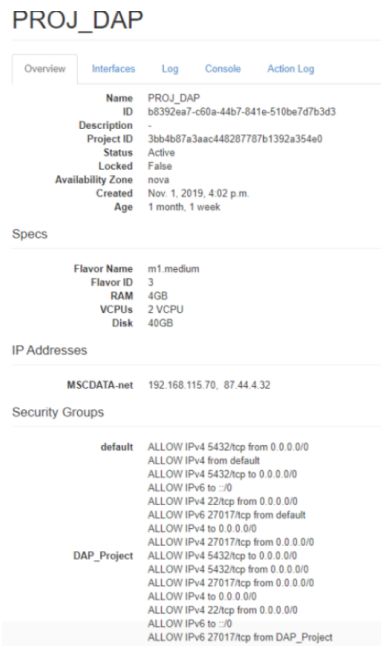

Fig. 3. Machine Specification

OpenStack, a cloud instance is used as Infrastructure as a Service (IaaS) provisioned for installation of MongoDB, Postgres and Spark. The cloud instance is provided by National College of Ireland. Security is ensured by authenticating the instance through a valid .pem or .ppk file.

[3]https://https://data.ct.gov/api/views/rybz-nyjw/rows.json
[4] https://catalog.data.gov/dataset/dohmh-hiv-aids-annual-report
[5] https://data.cdc.gov/api/views/vdpk-qzpr/rows.json
[6] https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/tables
[7]https://www.kaggle.com/solorzano/gdp-per-capita-in-us-states

Machine Specification:

- Machine_Type - Ubuntu-Bionic 18.04.3 LTS
- RAM - 4 GB
- Virtual CPUs - 2
- Hard disk - 40 GB
- Availability Zone - nova
- Java version - 8

The above figure represents an overview of the machine specification. New security (DAP_Project) is created for opening the ports 27017 and 5432. 27017 is for MongoDB and 5432 is for Postgres. Two super users mongodb and postgres were created to access NoSQL and SQL databases respectively.

MongoDB:

- MongoDB version 3.6.3 is installed and **myTester** user is created and Read-Write permission is provided to access it from python with password authentication.
- We have created a database **test** and store all our collections.

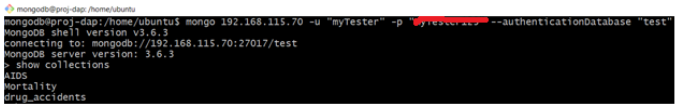

Fig. 4. MongoDB

PostgreSQL:

- PostgreSQL version 12.0 is installed and **group_a** user is created with Read-Write access, and it is accessed from python using password authentication.

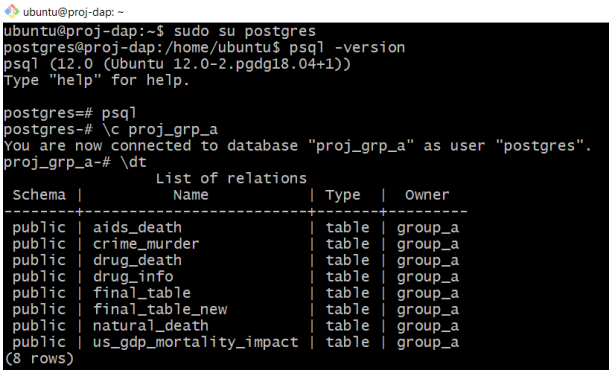

Fig. 5. PostgreSQL

- **Proj_grp_a** database is created, and the tables are stored in the database.

Apache Spark:

- Apache Spark version 2.4.4 is also installed in the OpenStack instance. Three key reason to choose Spark are **simplicity, speed and support**[8].
- It can be easily integrated with any services and scale the data. Data can be processed in-memory as well as in disk.
- As we are processing the data in-memory it performs faster than Hadoop MapReduce and other services. It supports language such as Java, Python, Scala and R.

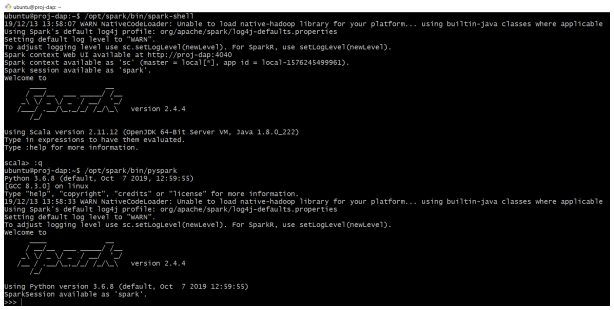

Fig. 6. SPARK

- Scala version is 2.11.12 and Python version is 3.6.8 is installed in the instance.
- For our analysis, Scala is used for processing the data from postgres using JDBC connection and combine the tables and load back into a new table.
- Spark-SQL function such as filter, join, groupBy, count and sum are used for joining and processing the dataset. The advantage of using Spark is schema-less.
- It processes the data in-memory and loads data into postgres without specifying its type and there is no need of creating a table, unlike python.

Data Processing:

- Python is used for Extraction, Transformation and Loading (ETL) the dataset into respective databases. Also pre-processing, cleaning of the dataset is carried out in python's library such as Pandas, NumPy, etc.
- Regex, map and lambda functions are also used to transform the datasets. For extracting the API data from websites, we used JSON imports of python.
- MongoClient from pyMongo for MongoDB, psycopg2 for postgres is used to connect the databases from python.
- A user-defined function (UDF) is created for exceptional handling in the code. The tables are loaded to postgres from CSV file using sqlio of Pandas library.
- The visualizations are performed using Plotly library of python.
- Various Plotly objects such as express, graph_objects, chlorepth maps, seaborn chart, heat map, histogram, bubble chart, pie-charts, donut charts, sun-burst chart and widgets drop down boxes are used for interactive visualization.

Automation:

- The automation process is carried out in the python code, whenever the data in the API changes, the code can pull the data and process it and if there is an error it can be caught using try catch exception.
- Also, the codes are stored in a common GIT branch repository and retrieved by each member of the group irrespective of the dependencies. The codes are peer

[8]https://mapr.com/blog/spark-101-what-it-what-it-does-and-why-it-matters/

reviewed by every member of the group and comments are updated. After the changes are implemented the final codes are pushed to the master branch using GIT Bash.
- To automate the process we are using **subprocess** module which stores all the .py file in a sequence as a program list.

```
(base) D:\NCI\SEM-1\DatabaseAndAnalyticsProgramming-1\Project\code>python sub.py
Finished:Read_Json_To_Mongodb.py
Table created successfully
Finished:Table_Creation_Postgres.py
```

Fig. 7. Sub Process Automation

- Each .py file will be executed in the same order when the previous program is executed successfully.

Firewall Settings:

```
ubuntu@proj-dap: ~
ubuntu@proj-dap:~$ sudo ufw status
Status: active

To                         Action      From
--                         ------      ----
OpenSSH                    ALLOW       Anywhere
27017                      ALLOW       Anywhere
5432/tcp                   ALLOW       Anywhere
OpenSSH (v6)               ALLOW       Anywhere (v6)
27017 (v6)                 ALLOW       Anywhere (v6)
5432/tcp (v6)              ALLOW       Anywhere (v6)
```

Fig. 8. Firewall Setting

- In the firewall settings, we have opened the ports 22, 5432 and 27017 for OpenSSH, PostgreSQL and MongoDB respectively to connect from other machines.

IV. Results

A. Murders in the USA for year 2014-16

- High/Low murder rates by states.
- Bar chart is used to visualize the murder rate per state, which is calculated per 100,000 population.
- Plotly graph object library's 'Bar' object was used in these charts.

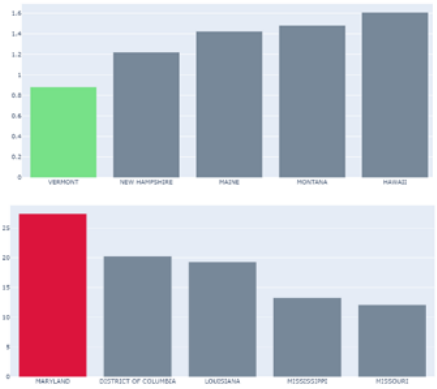

Fig. 9. Safe & Unsafe City

- The first chart shows Vermont to be the most safe state among the rest with murder rate of over 0.8 for the year 2016.
- Similarly, Maryland is identified to have had a high murder rate in that period followed by the district of Columbia and Louisiana.

B. HIV Mortality in the USA for year 2011-15

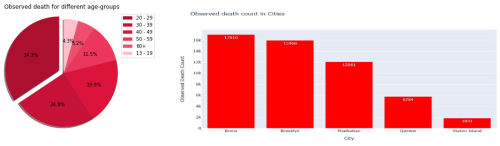

Fig. 10. Observed deaths due to HIV

- Pie chart to shows proportion of HIV deaths by age group.
- It shows that the max proportion(34.3%) of people died in the age group of 20-29 followed by 30-39 and 40-49. Also, comparatively least percentage(4.3%) of people died in the age group of 13-19.
- The graph above explains the observed death count for different cities in United states due to HIV.
- From this graph, we can infer that, the city Bronx of USA has the highest number of deaths due to HIV, while Staten Islands has the least number of deaths.

C. Natural Deaths in the USA for year 2005-15

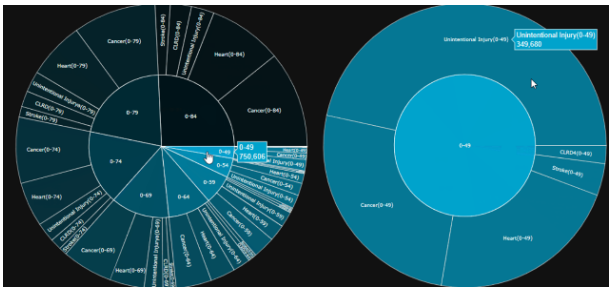

Fig. 11. Sun Burst Chart Natural Deaths

- Natural Deaths by Year, Age Range and type of deaths.
- Here we have used a sunburst chart which is incorporated in a custom function that takes 'Year' as input and returns a sunburst chart of natural deaths by age range and type of death for that year.
- Below is the screenshot of the interactive sunburst chart for the year 2014 which shows that a significantly low number of people of age range less than 49 have died (75 thousand) as compared to the other age ranges.
- Major cause of the death for the people below 84 were cancer and heart disease.

D. Drug Mortality in the USA for year 2012-18

- The interactive visualization will help us to select an option based on state, city, year and different age groups

as input and determine the death count and average drug score as output.

- The drug score is calculated on the sum of drugs consumed by individuals and the average function is applied based on year, city, state and age groups.

Fig. 12. Death Count and Average Drug Count

- Similarly, the overall death count and individual state-wise death count for 2012-2018 is calculated on year, state, city and age groups.
- A SQL query will help us to calculate and store in pandas dataframe and the visualization takes part in Plotly graphs with figure widgets.
- In this case, for the year 2013, all states, all city and all age groups are selected and observing the death count is 428 and the average drug score is 1.61.

E. GDP Vs Mortality Multiple Regression

The below fitted vs residual graph shows random noise in the variance as we move along the fitted values. Also there seems to be a linear relation between the fitted vales and residuals. The Durbin-Watson statistic result is 1.76 which is close to 2, this shows that the model has no auto correlation. Thus, satisfies the assumption of homoscedasticity and linearity[9].

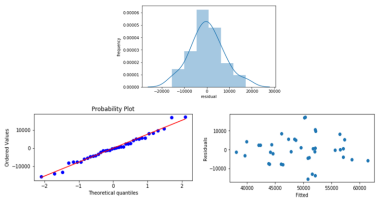

Fig. 13. Regression Assumptions

- After combing all 4 tables, the final table was joined with GDP table based on States and Year. The resulting table was used to identify the impact of Mortality on GDP using OLS regression in Python.
- From the summary of OLS we observed that only cancer death count and heart disease death count were

[9]https://github.com/bhattbhavesh91/linear-regression-assumptions/blob/master/lr-assumptions-notebook.ipynb

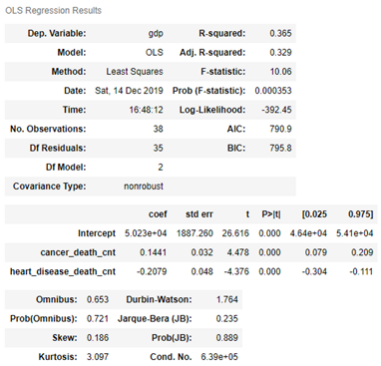

Fig. 14. Summary of OLS

significant($p<0.05$). Thus, rest of the factors were removed from the model. Below is the output of the improved OLS model summary

- Here we can observe that the model's accuracy is 32% which shows the model is not a reliable model.

V. Conclusion and Future Work

In this paper, different technologies, languages and studies on mortality and the GDP of the US are carried out. Also, in terms of analysis of the impact of GDP based on mortality justifies the model meets assumptions of multiple regression with the above figure we come up with an equation as stated below, which shows that number of deaths due to cancer and heart disease have a significant impact on GDP in the USA.

Out of 7 independent variables chosen only 2 are statistically significant. The below figure represents the correlation matrix between the dependent variable GDP and the 7 independent variables.

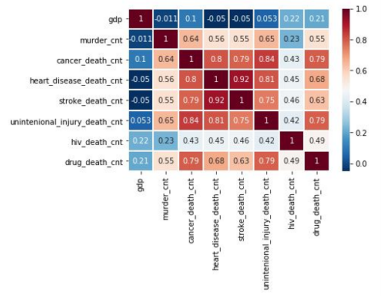

Fig. 15. Pearson Correlation

It can be inferred that when one person dies due to cancer the GDP is likely to grow by 0.1441, whereas when one person dies due to heart disease the GDP is likely to fall by 0.2079.

The assumption of spurious correlation is satisfied. Also the final equation for the GDP on the mortality rate is predicted by

$$Y = 50230 + 0.1441(\text{cancer\_death}) - 0.2079(\text{heart\_disease\_death})$$

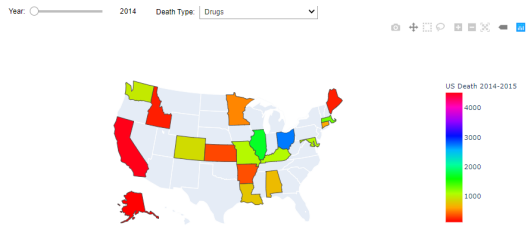

Fig. 16. US Geographical Heat Map

The USA geographical interactive heat map explains the different types of death in various states for the years 2014 and 2015. From this map, the US government can take action in a particular state for drugs, murder or any natural deaths. In the future scope of this project, we will try to improve the model performance by adding factors related to heart and cancer death such as smokers, non-smokers, people who eat high fat and cholesterol foods.

In addition to this, we have planned to implement the automation process using Apache Airflow which is a data pipeline tool written in python which helps us to view the process in DAG.

REFERENCES

[1] R. Devi, H. K. Tyagi and D. Kumar, "Early stage prediction of sudden cardiac death," 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2017.

[2] H. Shengli and L. Yi, "Statistical Analysis of 2115 Hospitalization Death Cases," 2015 7th International Conference on Information Technology in Medicine and Education (ITME), 2015.

[3] S.-H. Kim, C. Dunham, S. Muljono , A. Lee and T. Wang, "Discovery of Association Rules in National Violent Death Data Using Optimization of Number of Attributes," 2009 WRI World Congress on Computer Science and Information Engineering, 2009.

[4] S. Jeon, T. Park, Y. Soo Lee and S. Hyuk Son, "RISK-Sleep: Real-Time Stroke Early Detection System During Sleep Using Wristbands," 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2018.

[5] R. K. Sudheesh , J. Rajan, V. S. Veena and K Sujathan, "Study of malignancy associated changes in sputum images as an indicator of lung cancer," 2016 IEEE Students' Technology Symposium (TechSym), 2016.

[6] M. M. Hadjiandreou and G. D. Mitsis, "Taking a break from chemotherapy to fight drug-resistance: The cases of Cancer and HIV/AIDS," 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2013.

[7] K. Koide, H. Matsuura, N. Noda and T. Nemoto, "Effect of HIV on Japanese Population," 2008 3rd International Conference on Innovative Computing Information and Control, 2008.

[8] D. Grzonka, "The Analysis of OpenStackCloud Computing Platform:Features and Performance," Journal of Telecommunications and Information Technology 2015(3):52-57, 2015.

[9] L. Wang and D. Zhang, "Research on OpenStack of open source cloud computing in colleges and universities' computer room," IOP Conference Series Earth and Environmental Science, 2017.

[10] A. Ghaffar Shoro and T. Rahim Soomro, "Big Data Analysis: Apache Spark Perspective," Global Journal of Computer Science and Technology, 2015.

[11] M. Zaharia, M. Chowdhury, M. J. Franklin and S. Shenker, "Spark: Cluster Computing with Working Sets," International Journal of Data Science and Analytics, 2016.

[12] S. Salloum, R. Dautov and X. Chen, "Big data analytics on Apache Spark," 016 IEEE 24th Annual International Symposium on Big Data Analytics, 2016.

[13] M. Sfair Sunye , "BENCHMARKING POSTGRESQL FOR DATA WAREHOUSING," Proceedings of the IADIS International Conference on Applied Computing, 2005.

[14] M. Stonebraker, L. A. Rowe and M. Hirohama, "THE IMPLEMENTATION OF POSTGRES," Defense Advanced Research Projects Agency through NASA Grant, 2012.

[15] C. GYŐRÖDI, R. GYŐRÖDI, G. PECHERLE and A. OLAH, "A Comparative Study: MongoDB vs. MySQL," The 13th International Conference on Engineering of Modern Electric Systems, 2015.

[16] "MongoDB Vs MySql : A comparitive study of performance in super market management system," International journal of computational science and information technology, 2016.

[17] A. Kanade, A. Gopal and S. Kanade, "A study of normalization and embedding in MongoDB," IEEE International Advance Computing Conference (IACC), 2014.

VI. APPENDIX - INDIVIDUAL CONTRIBUTION

1) X18179541:

- Installation of MongoDB in OpenStack instance and creating a new user and assigning the permission to access from local machine.
- This helps the group to collectively work on the same database and create their own collections.
- Also installation of Spark is carried out in the same instance to process the data to create master table for the analysis.
- Consolidated the individual portion of the report from each member of the team and compiled as single final report and architecture diagram.

2) X18180159:

- Worked on the data visualization and the analysis of multiple regression on GDP Vs Mortality Rate.
- In the visualization, extensively explored the graphical objects and interactive widgets to integrate with Chloropleth map which is out of the box approach.
- Automation process is carried out using subprocess function and implemented successfully by sequential execution of process.
- Results and the conclusion part of the report is prepared along with the improvements in the visualization.

3) X18183379:

- Installation of PostgreSQL in the OpenStack and created a new super user and assigned required permissions to the user and check the connection properties.
- Handled the database activities throughout the project.
- Entire literature review is taken care in terms of the dataset and the technology part.
- Collectively worked in the team to get an overall idea of the project and prepared the presentation.

4) X18190910:

- The entire data modelling of the project is taken care and the entity relationship diagram is prepared to maintain the flow.
- The data source files and codes are maintained in the GitHub repository and periodical updates on the peer review is carried out to maintain the integrity of the project.
- Along with Iswaria, worked on managing the administration of MongoDB and PostgreSQL databases.
- Introduction and Methodology of the report is prepared. Also, the correlation among the factors is analysed.