Analysis and Visualisation of Medicare data in the U.S.A

Himanshu Rathee
School of Computing
National College of Ireland
Dublin, Ireland
x20132689@student.ncirl.ie
Group K

Ananya Chandel
School of Computing
National College of Ireland
Dublin, Ireland
x19237529@student.ncirl.ie
Group K

Rohan Koli
School of Computing
National College of Ireland
Dublin, Ireland
x19224842@student.ncirl.ie
Group K

Komal Bhalerao
School of Computing
National College of Ireland
Dublin, Ireland
x20135386@student.ncirl.ie
Group K

***Abstract*—This research paper examines four related data-sets from the health care domain. Extract Transform and Load (ETL) methodology has been implemented over each dataset. Data has been extracted from CMS.gov website application programming interface (API) by using socrata and the unstructured data is stored in MongoDB. Virtual machine (VM) has been used to mimic distant networks. The unstructured data has been transformed and stored in postgreSQL database. The structured data has been fetched using SQL queries and stored as pandas data-frames in python. Numerous visualisations has been performed using plotly and seaborn libraries. In the first data, around 5000 hospitals across various states in USA has been explored to gain insights by visualising mortality rates. In the second data, number of health deficiency surveys by state and city-wise is visualised. Also, overall ratio of the number of surveys conducted around US from year 2015 to 2020 has been visualised. In third data, medicare hospital spending by claim has been examined for ease of access of medical facilities. Payments received by the hospitals has been visualised against various ailments to infer the revenue generated and identifying most common type of ailment in the fourth data.**

***Index Terms*—ETL, VM, MongoDB, PostgreSQL, JSON, API**

I. INTRODUCTION

The current scenario has made everyone realise that medical facilities are of utmost importance. Human beings and their future is closely dependent on medical domain, that improvement in current Heath-care facilities would be a blessing. The future of human beings depend critically on the betterment of the healthcare facilities. Governments all over the world are spending heavily in medical domain to provide ease of access of medical facilities to the people. A large amount of data is generated by hospitals and given the importance of this data it is appropriate to visualise medicare datasets to gain useful insights from it. Throughout the world everyone encounters hospital at some stage of their lives. It is important to know the types of care services provided in different organisation that provide medical care. In this project, various factors are considered with respect to hospital domain. The main objective is to find four parallel data-sets that are closely related to hospital domain and perform exploratory analysis on the data. The insights driven from the process has been been visualized for better understanding and interpretation. After thorough research, we observed that there are various factors included in providing good medicare facilities such as health deficiencies, complication and deaths, medicare spending, Medicare claims and Value of care and payment. The main objective includes exploring, analyzing and visualizing health facilities.

A. OBJECTIVE

The objective of this project is to examine and perform different techniques to extract data from online sources and then visualising it to gain meaningful insights that would help in taking learned decision for the benefit of both patients and hospitals. Four different but related datasets has been examined with the visualisations Also, the project aims to understand and investigate the background of various research to carry out an overall research with respect to hospital domain. Various tie-ups of hospitals with medical claim providers and timely settlement of the claims is also analysed. The project also aims to study the ratio of deaths to surveys conducted and the factors affecting the same. The overall payment and care for different type of diseases including the value of care needed is also analysed.

B. MOTIVATION

Medical expenses are increasing each day .There is a growing demand for health care facilities like nursing homes and many more. Also the complications and deaths caused by different measures should be generalized .Payment measure and range is different for the type of patients encountered. This project aims at knowing the rates of health care by type of patients, claims provided by different hospitals, value of care for different measure and the average spending levels during hospitalising. All the data is gathered to create a better understanding for the types of spending, claims and other factors that are included and affect our visualisations.

C. RELEVANCE OF CHOSEN TOPIC

The data-sets have been extracted from cms.gov API under the hospital domain.The overall research focuses on different aspects of hospitals including complications and deaths, the care required for the treatment, regular surveys taken, total payment and value of care integrated together to understand the background thoroughly.

D. RESEARCH QUESTION

1) Which are the top 15 and bottom 15 hospitals based on mortality rate compared to national rate?
2) Which factors are a high cause for mortality due to complications in hospital procedures?
3) How many cities have centers available for deficiency survey checkup?
4) What is the ratio of surveys conducted so far from 2015 to 2020?
5) What is the overall Medicare hospital spending by claim?
6) What is the average time taken by hospitals to settle medical claims?
7) What are the state-wise revenue generated by hospitals?
8) Which is the most common ailment for which people pay at hospitals?

II. RELATED WORK

In the research paper by [1] , the author Anthony Nokrach compares NOSQL and SQL databases to store data. Based on the performance, easy setup, coding, and popularity of SQL databases are preferred over NoSQL (Mongodb).

In another research paper [2] the authors, Revina Rebbecca and Elizabeth Shanthi have proposed NoSQL soultion to store medical images. They compare the performance of both MongoDb and MySQL based on time for storing the data for i3 and i5 processors.

From an analysis conducted under healthcare included integration of Nosql database for Saudi Arabia.The followed data integration process included extracting, transforming and cleaning data. Several ways have been employed for integration of data. The method incorporated is the ETL process.In conclusion, the process of loading and transforming data is done using GENE2D system. The performance of query and the system interaction interface is validated with growing requirement of efficiency, scalability and security. [3]

Anjali chauhan discusses about different aspects of MongoDb databases,the theory survey includes comparison of RDBMS, No-SQL, MongoDB and Non-relational databases.After conducting analysis it was concluded that MongoDB maybe most popularly used but it is not the robust.MongoDB is an outstanding tool for forming data warehousing. [4]

In [5] the main causes for low MSPB (Medicare Spending per Beneficiary) index are found out by analyzing the data and solutions on how to reduce the MSPB index are suggested. It was found out that a lot of patients were readmitted to the hospital within 7 days of discharge which resulted in high MSPB index. Also, after the suggestions the MSPB index was reduced below 1.0 [6]. includes all Medicare Part A and Part B claims paid during the period from 3 days prior to an inpatient hospital admission through 30 days after discharge. [7]

III. METHODOLOGY

For the implementation of this project ETL methodology has been used where E stands for Extraction, T stands for transform and L stands for Load. Fig. 1

ETL Extract: Extract phase in the ETL involves identifying

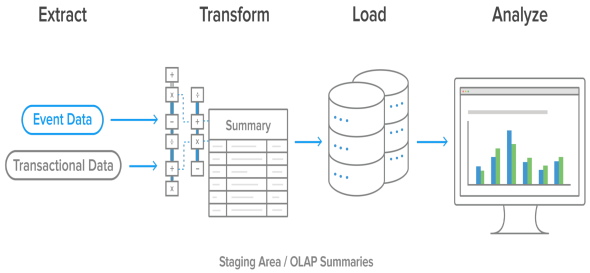

Fig. 1. ETL Flow

the data sources and number of rows, columns that are to be extracted from them. Sources of data can be Cloud based transactional databases or simple databases that can be fetched via APIs. This stage also involves planning for volume that might be required to store the data. The extracted data shall not have any negative impact on the system response time. For our assignment we have retrieved data from data.cms.gov website by using the APIs and the number of rows taken varied from 20,000 to 90,960.

ETL Transform: Transform phase involves the data transformation in which data is cleaned and checked for redundancies. This is an important phase, before performing any analysis on the data it needs to be transformed. It can be done in two ways, The classic approach involves cleaning the data before it is loaded in the data warehouse(database), Alternate approach involves In-house data transformation where data is transformed after loading it in the data warehouse. For our assignment we have used the traditional approach by transforming the data before loading it to the database.

ETL Load: Load marks the last phase of the ETL process flow. After the data is loaded in the database what we intent to do with the data is answered in this phase, we might be performing some analysis on the data or creating any machine learning algorithm etc. The data can be loaded in two ways into a warehouse, Full load in which the entire data is dumped into the data warehouse all at once. Other method involves incremental load in which the data is dumped part by part. For this assignment data was dumped all at once using full load into database.

Analyze: This marks the last stage of an ETL phase in which the data after being extracted and transformed is loaded into the database. The database is queried to fetch the data and later perform visualizations on the data. For this assignment the data is fetched from the postgreSQL database by SQL queries and stored as data-frame to analyze and visualise.

A. Data Description

The data sets that are used for Analysis are taken from data.cms.gov. This repository contains medicare provider data for USA and is accessible publicly.Throughout the world everyone encounters Hospital at some stage of their lives.

Some of the hospitals are certified and some are not. There are several types of health care providers available today.It is important to know the types of quality care services provided in different organisation providing medical care. In this survey,various factors are considered with respect to Hospital domain. The main objective is to find four parallel data-set that are closely related to Hospital domain and perform exploratory analysis of the data. The insights driven from the process is being visualized for better understanding and interpretation. After thorough research the we observed that there are various factors included for instance, Health Deficiencies, Complication and Deaths, Medicare Spending, Medicare claims and Value of care and payment.The main objective includes Exploring, analyzing and visualizing health facilities.

1) Health Deficiencies: Health Deficiency in medical terms can be termed a shortage of useful or functional factor, which is less than usual or important and necessary for functioning. Although there are many categories that fall under Heath Deficiency, namely Environmental, Nutrition, Quality of life and care and many more. Health Deficiency surveys are conducted with respect to the type of survey. This Dataset includes the Provider name, Inspection date, tag number, description of deficiency, scope and severity etc. It is observed that the surveys conducted were likely to decrease in the year 2020 because of COVID 19. These surveys are helpful for people on regular health checkup. The observations are noted and visualised to retrieved brief information [8].
2) Medicare Hospital Spending by Claim: This dataset has 67826 rows with major column being average spending by hospitals on claims of patients for numerous states of U.S.A. Medicare Spending per Beneficiary (MSPB) episodes by Medicare claim type is an important variable that should be visualised for quality and ease of medicare facilities. It also has information regarding the number of days a hospital takes to settle the claim along with the claim type for which any patient applies. Moreover, it has a total of 13 columns with information such as average spending on claims by states, hospital names, percentages of spending by hospitals, etc. [9].
3) Payment and value of care: This data set contains the information about the Hospitals and their locations.It contains payments, lower estimates of expenditures and higher estimates of expenditures. Data also displays value of care and payment measures. The value of care contains different measures according to payment and mortality. [7].
4) Complications and Deaths: The dataset contains information for every hospital in the USA describing the death rates arising by complications due to various factors such as heart failures, pneumonia, respiratory to name some. The dataset contains 90,960 rows and 18 columns. The columns contain information on hospitals such as hospital names, city, county. Next, the dataset contains information on factor measures by lower and higher estimates in rates. Also, there exists a column comparing the measures to the national rate; if they differ, if they are worse or if they are better for each hospital describing for each factors [10]

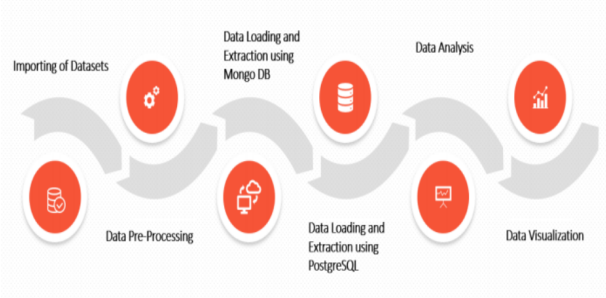

Fig. 2. Data Analysis Flow

B. Data Extraction

The website data.cms.gov contains all the data for medicare providers for USA. This repository was used to extract the data-sets required for our Analysis. Data sets were extracted from the website by using API. On fetching the data from the API, Unstructured JSON data is obtained which is available for storing in appropriate database. To extract the data from API using python, sodapy package from python is used. 3

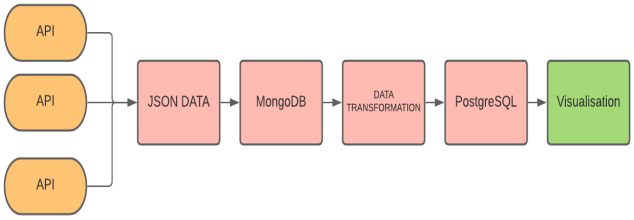

Fig. 3. Process Flow

C. Unstructured Data Storage

MongoDB is a NoSQL database, It can be used to store huge volumes of data. Instead of storing data in the form of tables and rows it used documents and collections which are key-value pairs and basic unit for MongoDB data. After the extraction of the data from the API in the the form of unstructured JSON format We used Mongo database to store the JSON data. Mongo database gives us an advantage of storing the unstructured JSON data directly without making any changes or formatting to the data. Another reason for using MongoDB is that it doesn't need the schema to be defined before hand.

D. Transformation of Data

All four datasets were stored in the MongoDB. To perform operations on the data, we fetch the data by using pymongo

package in python which is used to establish the connection with the database using the IP address and the port number of the Virtual Machine. Once all the four data-sets are fetched from mongoDB we store them as data frames using pandas library in python. Once the data sets are stored as data frames we perform cleaning and transformations. All the data was checked for redundant values, All the unnecessary columns from the datasets were dropped. Data was checked for the Null and NA values, all the rows with null values were dropped as we have sufficient rows in our datasets. Few columns were renamed for the for better interpretation and understanding of the data.

E. Structured Data Storage

PostgreSQL is an open source relational database system. Its capable of scaling and working with most complicated datasets using SQL. Postgre allows you to write codes even from different programming languages with no recompilation of database required. The data sets after being cleaned and transformed, is stored in PostgreSQL database. Pythons package sqlalchemy is used to create database for PostgreSQL. Storing the data in the PostgreSQL makes it easy to retrieve the columns on which we need to perform the visualisations. It also eliminates the purpose of repeatedly cleaning the data for creating the visualisations. The psycopg2 package in python provides us the functionality to connect to the PostgreSQL database.

F. Data Analysis

All the data sets After being cleaned, Transformed and structured are stored in PostgreSQL database. To perform any operations on database, the data is fetched from postgreSQL using pythons psycopg2 package to established the connection. SQL is used to extract the data from the postfreSQL. Once data is fetched from the postgreSQL operations like transformations and visualizations can be performed on the data. All the four datasets were fetched from postgreSQL and stored as dataframes by using pandas library in python in order to perform visualizations on them.

G. Data Visualization

Data is fetched from the PostgreSQL database using SQL query methods, and stored in data frames using pandas library in python to create visualisations. For creating visualisation on data various python libraries were used. Few of them are Plotly, Matplotlib, Seaborn.

IV. Results

A. Health Deficiencies

The Donut chart is created in python using matplotlib library. It depicts the overall providers for health deficiency surveys where there are 100+ providers available state-wise. As observed California is the state with highest providers up to 519. The donut chart gives the count for each state by hovering over. California holds 28% of the overall percent whereas the state with lowest survey provider is Arizona with 7.28%.
The Bar chart depicts the top 10 cities providing health

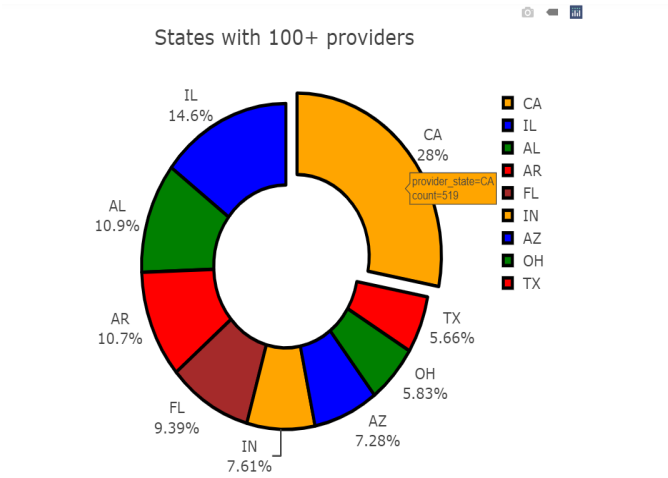

Fig. 4. Donut chart for state-wise providers

surveys.It is important for people to know the hospitals in different cities of providers located.As Los Angeles is crowded and a well known city in America it is also the top city to provide surveys as compared to all other cities. Followed by Los Angeles, Phoenix is the top second in the list.Whereas, the city with lowest providers is Rogers. All the survyes conducted in the cities are monitored and the data is saved to take follow-ups on the survey results.

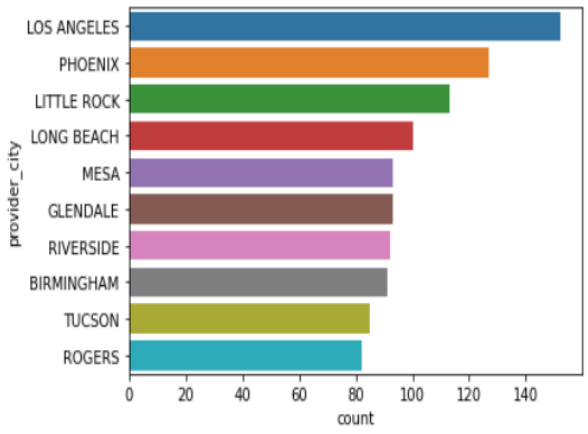

Fig. 5. Health service providers city wide

To extract the year part from the survey date column, the data type for the column is changed from object datatype to datetime to extract using sql query.The pie chart gives a brief description of the surveys conducted from 2015 to 2020.As the total count for the year 2015 is 4 it is not included in the pie chart as the ratio against other years is low. It is observed that the highest surveys conducted were in year 2018 and 2019 followed by year 2017. Fig. 6 represents the percentage of surveys conducted over the years.

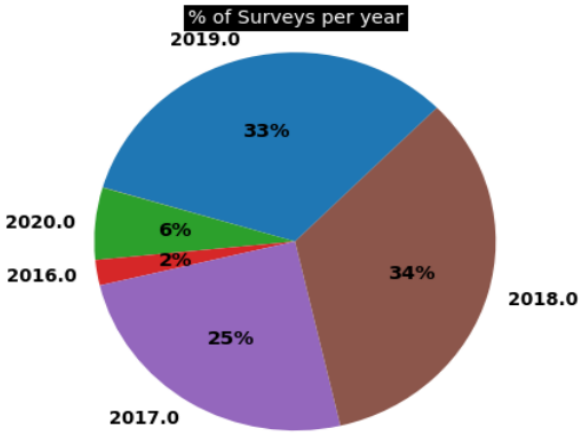

Fig. 6. Survey conducted over years

B. Medicare hospital spending by claim

- It has been noted that this data set consists of 899 unique hospitals for 15 different states of the United States of America (U.S.A).
- Fig. 7 shows the average spending on claims by all the hospitals of a particular state. It has been noted that Washington DC and Delaware have the given the highest amount in claims to patients whereas Texas hospitals have paid the minimum. This information might help the government of those states in deciding the future steps on their medicare policies.

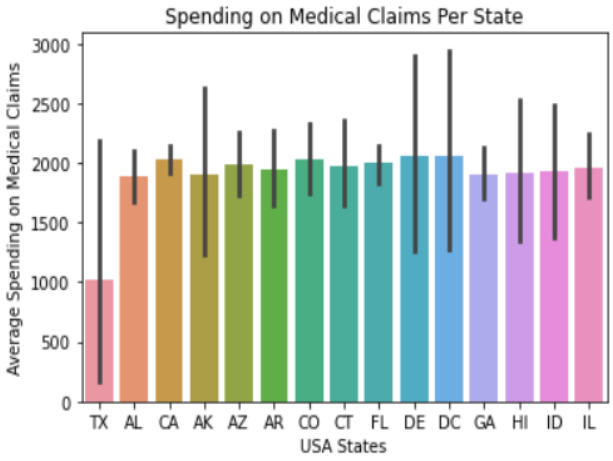

Fig. 7. State wise average hospital spending on claims

- Fig. 8 displays a pie chart with the time period taken by hospitals to settle medical claims of patients. It has been noted that most of the hospitals settled the claims after the completion of the episode i.e. after the patient has completed his medication period. Also, only 1.7% of the total hospitals settle the claim within 1-3 days of admission of a patient in the hospital. This shows that the governments should tighten the insurance companies and the hospitals for timely settlement of the claims in order to help the people.

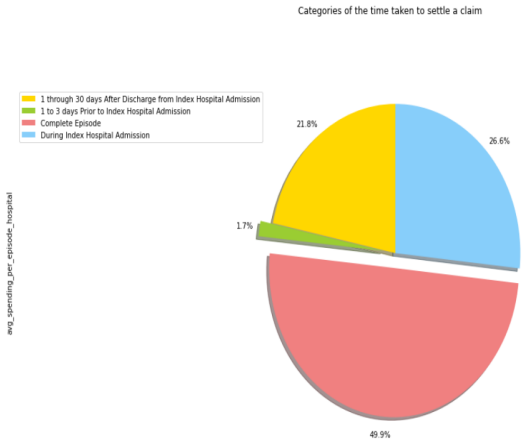

Fig. 8. Settlement of claim on the basis of time period

- It has also been noted that inpatient claim type is the the most settled claim type whereas hospice is the lowest settled claim type by all the hospitals. Therefore, this could help the hospitals in increasing their efficiency by giving more flexibility to the claim types with lower settlement rates.

C. Payment and value of care

This data contained information about Medical Spending's of the beneficiaries and payments received for pneumonia, heart patients and hip or knee replacement patients and value of care for different ailments.

As can be seen from fig.9 we can see the payments received in dollars depending on different ailments. Bars in the plot show payments that are made by the patients that are arranged according to different ailments or measures they paid for. As can be seen from the figure 1 Maximum payments are made by the patients undergoing Heart attacks. By this we can infer that value or the number of patients undergoing heart attack treatment are highest.

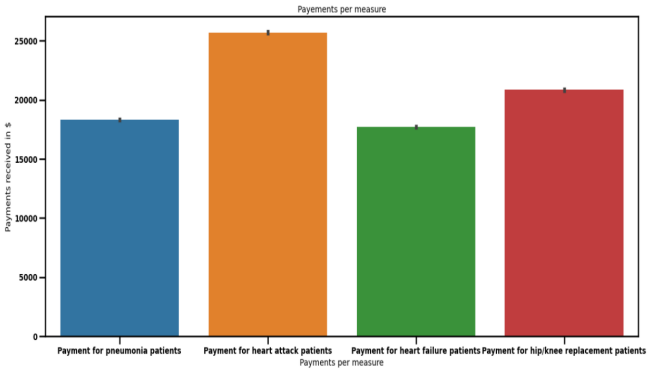

Fig. 9. Payments analysis per measure

As can be seen in the fig.10 The payments received by multiple hospitals are plotted state wise. This plot gives us information about all the payments received by hospitals in a particular state in us. This shows maximum revenue generated by the hospitals lie in NJ(New Jersy or DC(Washington DC). This visualisation can help in Identifying the revenues generated state wise and also gives us an idea about the number of hospitals that might per present per state.

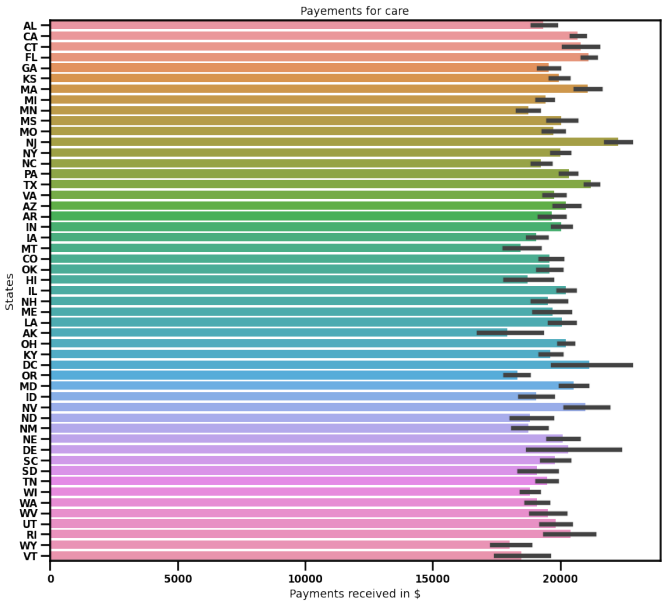

Fig. 10. Payments received by hospitals per state

As can be seen from fig.11 the value of care is assessed by the mortality rates and the payments. The hospital's payments are compared with average national payment and the mortality rate is compared with national rate. This is done to evaluate the value of care from the fig.11 The value of care can be visualised according to its intensity and payment estimates.

value_of_care_category	payment	lower_estimate	higher_estimate
Average Complications and Average Payment	19925346	18600866	21360126
Average Complications and Higher Payment	12847104	11971746	13798047
Average Complications and Lower Payment	17513465	16801451	18260119
Average Mortality and Average Payment	115786897	104416919	128088286
Average Mortality and Higher Payment	23663652	21896542	25507327
Average Mortality and Lower Payment	18606934	16968555	20310873
Better Complications and Average Payment	84244	82873	85673
Better Complications and Higher Payment	131126	127621	134750
Better Complications and Lower Payment	855357	837673	873365
Better Mortality and Average Payment	3748038	3540411	3963073
Better Mortality and Higher Payment	2986943	2829883	3149606
Better Mortality and Lower Payment	1180719	1119004	1243894
Worse Complications and Average Payment	250211	239771	261217
Worse Complications and Higher Payment	619629	591994	648970
Worse Complications and Lower Payment	150001	145290	154942
Worse Mortality and Average Payment	4069451	3794147	4355351
Worse Mortality and Higher Payment	2330109	2197145	2465251
Worse Mortality and Lower Payment	1309727	1223395	1397867

Fig. 11. Value of care for hospitals

D. Complications and Deaths

To visualize the map of USA for the number of hospitals present in each county,a dataframe has been extracted from postgres sql and grouped by unique shortcodes of the states. Next, plotly and choropleth function has been used to plot the pandas dataframe as depicted in fig.12. The function plots the world map according to location shortcodes and values in the dataframe. According to the map, we can see Texas (count 329) and California(315 count) has the highest number of hospitals, suggesting highest reach and accessibility for medical facilities.

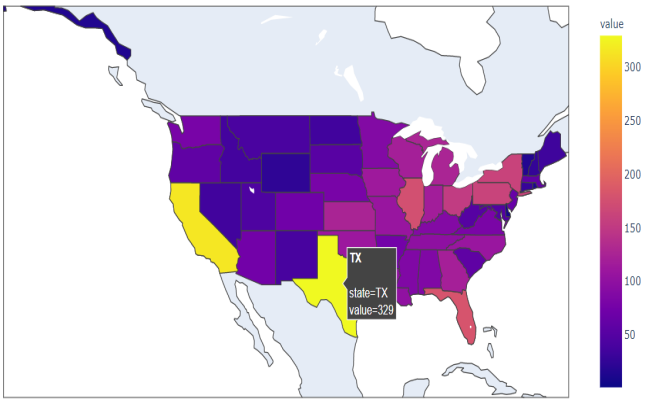

Fig. 12. U.S.A map based on number of hospitals

In Fig. 13, the dataset has been grouped using measure_id. Next, seaborn library has been used in combination with matplotlib to plot the extracted pandas dataframe. From the bar-chart we can see the average mortality rate to be the highest where complications arise due to pneumonia whereas the average mortality rate arising from hip fracture surgery is the lowest.

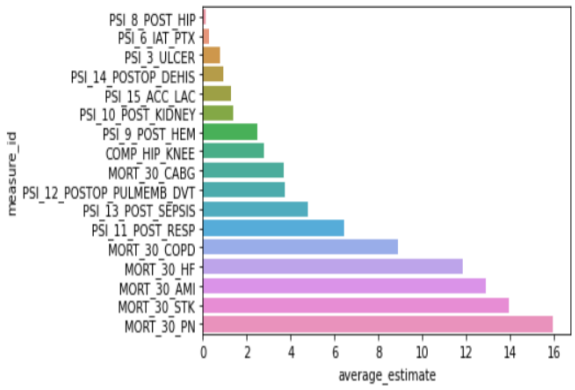

Fig. 13. Hospital mortality rate based on various factors

In fig.14 and fig.15 the names top and bottom 15 hospitals are recognized by comparing each hospital with the national rates for each factors. The sum() function has been used to get

```
In [111]: ####Getting top 15 best US hospitals
          plot2.head(15)
```

Out[111]:

	hospital_name
2475	NEW YORK UNIVERSITY LANGONE MEDICAL CENTER
2383	MORRISTOWN MEDICAL CENTER
771	CLEVELAND CLINIC
2403	MOUNT SINAI HOSPITAL
403	BEAUMONT HOSPITAL ROYAL OAK
404	BEAUMONT HOSPITAL, TROY
3064	RUSH UNIVERSITY MEDICAL CENTER
2041	MASSACHUSETTS GENERAL HOSPITAL
34	ADVENTHEALTH ORLANDO
624	CENTINELA HOSPITAL MEDICAL CENTER
1544	HOUSTON METHODIST HOSPITAL
2969	READING HOSPITAL
2556	NORTHSHORE UNIVERSITY HEALTHSYSTEM - EVANSTON ...
3489	ST LUKE'S HOSPITAL BETHLEHEM
1341	GULF COAST MEDICAL CENTER LEE HEALTH

Fig. 14. Top 15 hospitals based on complications

```
          plot2.tail(15)
```

Out[112]:

	hospital_name
126	ANDERSON REGIONAL MEDICAL CTR
3463	ST JOSEPH REGIONAL MEDICAL CENTER
2940	PROVIDENCE ST PETER HOSPITAL
3818	UAMS MEDICAL CENTER
4124	WESTCHESTER MEDICAL CENTER
2231	MERCY MEDICAL CENTER
3875	UNIVERSITY MEDICAL CENTER
2603	O U MEDICAL CENTER
3359	SPARTANBURG MEDICAL CENTER
3713	THE MEDICAL CENTER (BOWLING GREEN)
2438	MUSC MEDICAL CENTER
1648	JACKSON-MADISON COUNTY GENERAL HOSPITAL
3095	SAINT FRANCIS MEDICAL CENTER
1359	HALIFAX HEALTH MEDICAL CENTER
2235	MERCY MEDICAL CENTER REDDING

Fig. 15. Bottom 15 hospitals based on complications

the counts of each factors. Using the head() and tail() function, top and bottom 15 hospitals are listed from the dataframe.

V. Conclusion and Future Work

Various visualisations are created under this research paper. These visualisations can be used by healthcare service providers and Governments of various countries for evaluating the present and past health care services. From these visualisations, we gained various insights about revenue generated per state, value of care for hospitals, mortality rate, top hospitals in the state and many more. These insights can be used to formulate business rules and enhance performance of the preexisting frameworks. Healthcare services are a necessity and of critical importance the inputs from this research paper can be used to improve the Health care services by adequately allocating health care resources to enhance availability. The frame work and databases were successfully implemented using virtualbox to run mongoDB and postgreSQL. Storage of the unstructured JSON data in mongoDB was very fast (as we stored unstructured data in mongodb) and later storing and structuring data on postgresSQL was successful.

Although the technology and methodologies used are very competent and act as powerful tools for databases storage and analysis, they lag behind due to limited storage and computing power. In future, we would implemented various ideas by modifying the ETL process by inculcating cloud technologies and utilising various statistical and Machine Learning algorithms which would deliver numerous business intelligence. As a part of Business Intelligence, it is possible to integrate data visualization softwares such as Tableau and Power Bi. Further, we would automate and visualize in real-time which was difficult due to lack of time in hand.

References

[1] A. Nokrach, "Comparing the databases mssql and mongodb for the web-based environment ozlab," 2018.

[2] W. L. Schulz, B. G. Nelson, D. K. Felker, T. J. Durant, and R. Torres, "Evaluation of relational and nosql database architectures to manage genomic annotations," *Journal of Biomedical Informatics*, vol. 64, pp. 288 – 295, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046416301526

[3] A. L. a. S. "Halima Samra, OrcID, ""gene2d: A nosql integrated data repository of genetic disorders data"," 2020.

[4] A. Chauhan, "A review on various aspects of mongodb databases," 2019. [Online]. Available: https://www.ijert.org/a-review-on-various-aspects-of-mongodb-databases

[5] N. Bowers, J. Cirelli, A. Andrzejewski, J. Lang, F. Aqlan, and A. Pedersen, "Analysis of medicare spending per beneficiary (mspb)," vol. 2018, no. SEP, 2018, pp. 1171–1179.

[6] J. Bender, T. Nicolescu, S. Hollingsworth, K. Murer, K. Wallace, and W. Ertl, "Improving operating room efficiency via an interprofessional approach," *American Journal of Surgery*, vol. 209, no. 3, pp. 447–450, 2015.

[7] 2020. [Online]. Available: https://data.cms.gov/provider-data/dataset/c7us-v4mf

[8] 2020. [Online]. Available: https://ec.europa.eu/eurostat/view/nrgcbe

[9] 2020. [Online]. Available: https://data.cms.gov/provider-data/dataset/nrth-mfg3

[10] 2020. [Online]. Available: https://data.cms.gov/provider-data/dataset/r5ix-sfxw