Analysis various Health Care Factors

1st Archana Uday Mahajan
MSc. Data Analytics
National College of Ireland
Dublin, Ireland
x20198825@student.ncirl.ie

2nd Pritish Mehta
MSc. data Analytics)
National College of Ireland
Dublin, Ireland
x20184409@student.ncirl.ie

3rd Rutuja Dinesh Mehta
MSc. Data Analytics
National College of Ireland
Dublin, Ireland
x20129751@student.ncirl.ie

4th Marmik Ajmera
MSc. Data Analytics
National College of Ireland
Dublin, Ireland
x20237723@student.ncirl.ie

I. ABSTRACT

The outbreak of various flu's, virus and diseases have time and again affected the population, the healthcare workers, the healthcare system, the government as well as the entire human race at some point in time. One of the recent example of it is the COVID-19 pandemic which we have experienced closely and has also experienced the damaged it caused to the global economy as well as the personally everywhere around the world. Since the healthcare system started using modern technologies, the use of data comes along with it parallelly. Therefore, it has become and important research area due to the massive availability of the data in this sector. The primary aim of this project is to search four analogous data sets from a selected domain. The four datasets chosen are related to Healthcare system in the United States of America. Exploratory data analysis has been carried on the four datasets. Acumen from this analysis is visualized using various plots. The four data sets are subjected to a series of data analysis using the Python programming language.The analysis is further extended establishing a connection to databases such as MongoDB and PostGreSQL using Python programming language. Visualization is created to understand the trend using Matplotlib, Seaborn and Plotly libraries. ***Index Terms*—MongoDB, Postgre, Visualization, Healthcare, Medical Expenses, Infections, Covid-19, Python, Git**

II. INTRODUCTION

The objective of this project is to analyze the various healthcare factors that are considered in the healthcare industry, which is one of the most important and crucial industry not only specific to country but to the entire human race. Massive volumes of data is generated due to the adaptation of digital technologies that collect patients records, managing the supplies of the hospital equipments, hospital performance data, hospital staffs and nurses data, and the list goes on. Applying analytics on this data can result in the outcomes that can benefit everyone around us. For years, gathering massive amount of data for medical use has been proven difficult, time consuming and at times expensive as well. But with today's modern age developments, it becomes possible to not only save this data but also create comprehensive healthcare reports and turn them into critical insights. With the growth of wearable devices, every person has his own digital health data on their mobile phone. United States of America, even though being one of the most technological and medical advanced country, faces difficulty in solving their healthcare system crisis to the people's advantage. Even in recent times, having the best talent along with ample financial resources, U.S.A remains one of the most impacted country with the COVID-19 virus. Healthcare being one of the most important industry that is inevitable to fade as long as humanity exists, by default has to have ample amount of data. Along with the medical advancement and research carried out at the required pace, healthcare being one of the data rich industry, it becomes imperative to explore the data and find out insights that aid in the betterment of people, government, hospital organizations. Exploring the healthcare data enables the government and the hospital organization to better study the patterns and take appropriate decisions. Exploring this data also helps in planning the allocation of the financial resources and advice the government and the concerned organization when and how much to allocate for a specific illness. For example; Having appropriate data can help the government in infusing the financial capital to COVID-19 situations with respect to it's severity. With the help of the analysis on the healthcare datasets, every aspect and every data of the healthcare system can be worked upon to better understand the trend along with it's accurate time to help the government and hospital organization to plan their spending and help the people. With the help of the data, the inessential or the secondary expenses can be halted or saved, and instead can be used on the process that require more financial aid. Analyzing and researching about the datasets of the healthcare industry helps us to find out insights about the industry and various factors in the healthcare industry. In this project we have taken into consideration four different datasets having various factors trying to highlight the inter reliability as well. The population of the world needs to know the data and try to explore the data and use it to our advantage. We were motivated to carry out the analysis on the chosen topic due to the fact that the healthcare industry as a whole needs to be understood and needs to be explored.

III. Related Work

The evaluation of decision aids such as healthcare interventions must be done before the patient pathway is introduced for providing them with the maximum benefit, prevent harm and achieving their goals. The evaluations that were published by Fiona Collard and Jonathan M Garibaldi differ in outcome measures, making comparisons difficult and limiting progress in this field. [1]

Adopting a Business Process Management practices is becoming crucial for the healthcare process improvement. As a result, the methods and tools that can be used to address the behavioral and performance aspects need to improve quality of healthcare , reduction in costly reworks, and increase the efficiency of BPM approach. This paper by G. Antonacci, A. Calabrese, A. D'Ambrogio, A. Giglio, B. Intrigila and N. Levialdi Ghiron focuses on the process life-cycle's specification and analysis phases, and it gives an introduction to a model-driven technique for the simulation of healthcare process . [2]

Among the many potential causes of rising healthcare costs, the major contributor is claim fraud, and the impact can be mitigated by detecting fraud .Richard A. Bauder and Taghi M. Khoshgoftaar proposed a outlier detection model which is based on Bayesian inference. Unlike most common outlier detection methods, the model gives the probability distributions instead of the point values. Credible intervals were generated for increasing confidence. [3]

Computational intelligence has the concerned in creating and applying various computational models and their simulations, which performs high in computing for solving complex physical problems that were risen in the analysis of engineering, design and natural phenomena. This work by P.A. Harsha Vardhini, S.Shiva Prasad and Seena Naik Korra presents an application that distributes medications to those who have tested positive. This is followed by a review of the patient's medical information, which includes blood pressure, diabetes, cancer, alcoholic habits, and so on. This paper also discusses clustering methods, which are a part of unsupervised learning. Various clustering algorithms were used to evaluate the accuracy score for allotting medicines. [4]

Chiara Antonini, Sara Calandrini, Fabrizio Stracci, Claudio Dario and Fortunato Bianconi proposed the use of a computational framework for modelling the epidemic and analyzing the effects of lockdown. They calibrated a model of COVID-19 clinical progress against daily epidemiological data using a Bayesian method called CRC. The calibrated model was then subjected to a robust analysis for quantifying the influence of model parameters and generating possible scenarios for containment measures. CRC gives us the conformation on hypothesis of underestimating new positive cases and emphasizing the importance of detecting pre-symptomatic transmission in order to reduce the contagion. [5]

A mix of inherited and environmental variables that cause the severe diseases posing a threat to health. Incorporating the links between cases with other data reveals that shared genetic variations and the exposure to environment could be linked to diseases like these. The difficulty of assessing these diseases transfers to a multivariate graph visualization problem. Their design studies contribution to a new visual representation for multivariate graphs, which has an application to genealogies and clinical data. To scale to several families, they present data-driven aggregation approaches. [6]

In this survey, authors collect related data to highlight the importance of data mining in the sector of healthcare. Using data mining models to analyze health datasets obtained by electronic health record systems it helps in insuring claims and health surveys for extreme complications and can be fraught with obstacles. [7]

Recent research in biomedical and genomics exhibit a link between gene mutations, inheritance and the possibility of getting specific malignancies. In this paper, the authors have researched about the distributions of different cancer in breast and ovaries throughout the United States of America across all states and counties. A Web-based Analysis and Visualization Environment, Weave, is described in brief which is used to look into the different cancers and also provide interactive visualizations of family hereditary patterns and genetic distributions. The system was also helpful for other deadly cancers and similar health indicators. The authors have also explored the interaction of Weave with cancer data and it can also be used to relate to other types of epidemiological data, such obesity. [8]

Manual monitoring of hospital-acquired infections is time consuming and is mostly restricted to ICU. The effectiveness can be enhanced using computer-assisted techniques for recognising hospital-acquired illnesses. The authors have also suggested the development of a state-of-the-art knowledge-based e-Health surveillance system for predicting hospital-acquired illnesses. The system may accumulate patient-related data from hospital databases and use knowledge discovery criteria and hospital acquired infections decision standard algorithms to predict patient infection automatically. Both central line-associated bloodstream infections and patient treatment expenses can be significantly reduced with the proposed method. The system's experimental results showed an improvement of 87%. [9]

IV. Methodology

A. Data Acquisition

1) Dataset 1 HRIP: The dataset is obtained from the following source
Source:https://health.data.ny.gov/Health/

Medicaid-Electronic-Health-Records-Incentive-Progr/6ky4-2v6j

HRIP stands for Health Record Incentive Program. This dataset is for Medicaid Electronic Health Records Incentive Program Provider Payments which contains the prices of medical care for the year of 2011. The dataset contains the type of the provider (Physician, Dentist, Nurse etc) and the other details including the price. The data has been downloaded from the the state of New York State Health data.

2) Dataset 2 HEDIS: The dataset is obtained from the following source

Source:https://opendata.utah.gov/Health/Utah-Healthcare-Effectiveness-Data-and-Information/gawi-uz7h

The next dataset is HEDIS which stands for Healthcare effectiveness Data and Information Set which shows the patient care that is under gone in the previous year. This data is from 2010 till 2017. The data set contains multiple variables like the computed value, national average, the plans undertaken by the end user etc. Further, the dataset has been obtained from the Open Data Catalog from the Sate of Utah.

3) Dataset 3 HAIH: The dataset is obtained from the following source

Source:https://opendata.utah.gov/Health/Healthcare-Associated-Infections-By-Hospital-In-Ut/pu6y-f26m

The third dataset was also obtained from the Open Data Catalog from the Sate of Utah. The dataset HAIH which stands for Healthcare associated Infections by Hospitals shows the number of infections caused when the patient is admitted to the hospital. The data is from the year 2013 and shows various different infections that are caused with the total number of people affected by that infection.

4) Dataset 4 CNH: The dataset is obtained from the following source

Source:https://health.data.ny.gov/Health/New-York-State-Statewide-COVID-19-Nursing-Home-and/u2vg-th2g

Finally the last dataset is based on the current situation. The dataset contains the number of reported and lab confirmed casing in the nursing home and adult care facilities. This data like the first one is being obtained from the New York State Health Data. The dataset contains the information of all the cases from April 2020 all the way till March 2021. The dataset contains various variables such as End of the week cases, cases within the facility, cases outside the facility and also the presumed cases.

The reason for choosing these datasets is to visual various aspects of the healthcare industry.

B. Data Prepossessing

All the 4 datasets used in this projects are downloaded in CSV format from their respective websites. Using ”json” and

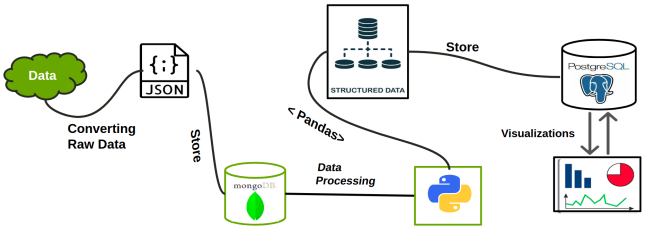

Fig. 1. Flowchart of the entire process

”csv” libraries we coverted the data into a JSON format so that it can be stored onto MongoDB.Which is a NOSQL database which eliminates the issues of storing big data. The traditional database uses tables and rows while in MongoDB Documents and collections are used as key-value pairs. The data which is being stored onto the MongoDB database is unstructured data hence we will have to perform some data cleaning and transformation steps to convert it to structured format. Further, by using ”pymongo” library we established a connection with the MongoDB. Once the connection is established we can now store the JSON converted files onto MongoDB. Once they are stored onto the databases in MongoDB, they can then be converted to dataframes using library ”pandas”. The screenshot below shows the database being uploaded in MongoDB.

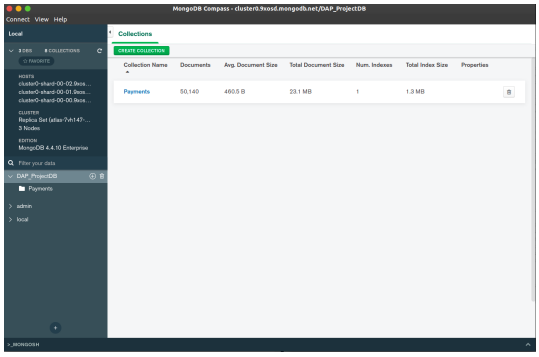

Fig. 2. Flowchart of the entire process

C. Data Cleaning

Data Cleaning is one of the most important steps. If the data that is being obtained is not not cleaned then the visualizations that will be performed on the datasets will not give accurate results. We applied various data cleaning methods on our datasets to make better visualizations. Firstly we removed all the null values present in the datasets. Additionally we checked if there are any empty strings in the dataset and removed them as well. Finally, we replaced few of the ”NA” values with random values so that it does not affect our visualizations.

D. Data Transformation

Data Transformation is a stage where we perform various steps to change the format of the data. In this project we have used the ”ETL” (Extract Transform Load”) process where we are performing the transformation of the data before loading the data for visualizations. For this project first we dropped all the columns that are not required for visualizations. Further we then converted the numeric columns to either ”FLOAT” or ”INT” depending on the data type required. Finally we stored the final DataFrame into a csv file which will be then uploaded to PostGreSQL.

E. Data Loading

In this project we have used 2 databases i.e. MongoDB and PostGreSQL. The data is of 2 types 1. Unstructured Data: The data the is converted in JSON file is the unstructured format and as mentuoned above it uses key pair values to store the data. 2. Structured Data: The data that is transformed and being stored in the PostGreSQL is the structured data. PostgreSQL is an SQL database which uses the ”RDBMS” (Relational Database Management System) to store the data into tables and rows. First a connection to the Postgres server has to be initialised for us to save data in PostgreSQL. The data is uploaded using Python code with the help of the psycopg2 library. The psycopg2 library tries to join the database after importing. If the connection fails, then the error is printed out. After a successful binding, a new database called postgres is created in Postgresql. Once the data is stored onto the PostgreSQL database it can now be explored using multiple libraries to perform various analysis on the data. The libraries such as Pandas and NumPy was use for data transformation and data cleaning while Plotly and Seaborn is used for the visualization of the data. Below screenshots depicts the dataset being uploaded to the PostgreSQL Database ”payments”

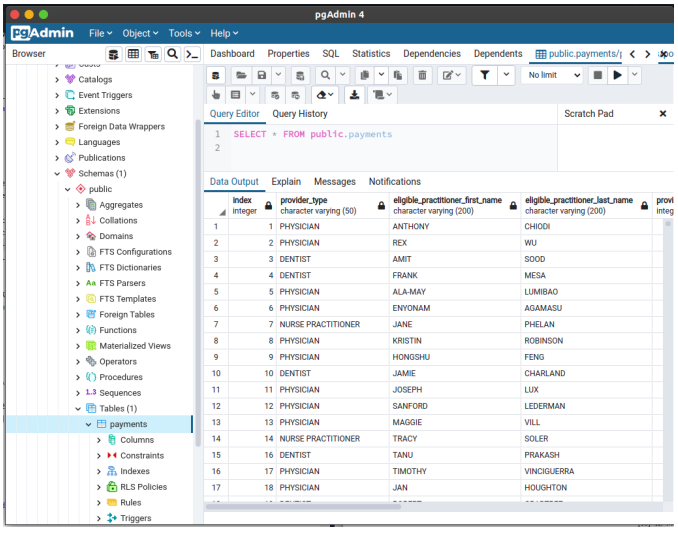

Fig. 3. Pie Chart for different Providers

V. Results and Evaluation

To perform visualizations on the datasets we imported few libraries like ”MathPlotLib”, ”Seaborn” and ”Plotly”. Seaborn is a data visualization library which is based on MathplotLib. Seaborn generates various statistical visualizations which helps the user in making better decisions on the datasets. MathplotLib is another library which help generate various visualizations both static and animated. Finally, Plotly is a library which is used to make interactive plots.

A. Pritish

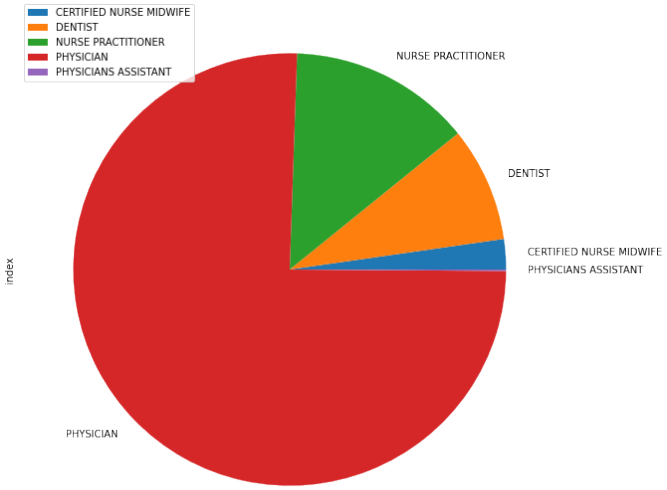

Fig. 4. Preference between different health providers

The figure 4 shows the preference of people between the different health providers. Majority of the people prefer to visit a physician whereas very few prefer to visit a certified nurse midwife.

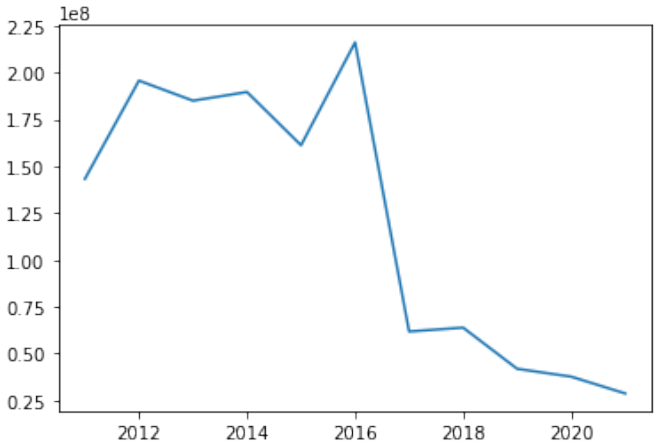

Fig. 5. Total expense on healthcare (in 100 millions) per year

Figure 5 depicts the total expenses done by people across different healthcare providers for every year. There was a sudden spike in 2016 where the expenses reached a record high of more than 200 Million in the last 10 years. After the spike there was a major downfall in 2017 to roughly 75M and had been decreasing steadily since then.

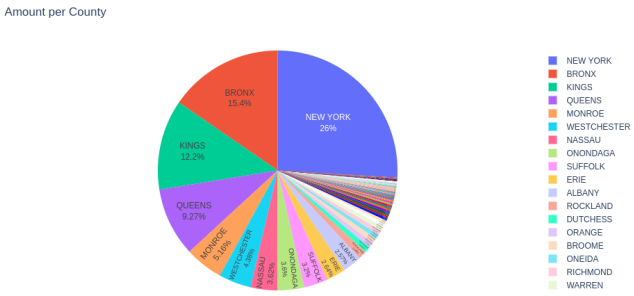

Fig. 6. Total expense across different counties

The figure 6 showcases the total expenses done in different counties from 2011 to 2021. Around 26% of the total expenses are done solely in New York. Following that 15.4% expenditure was done in Bronx and 12.2% of the total expense was in Kings. Rest all the counties have less than 10% of the total expense across different health providers from 2011 to 2021.

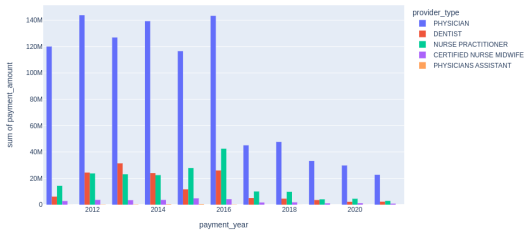

Fig. 7. Expense distribution across different heath providers

Figure 7 depicts the sum of payment amounts across different health providers from 2011 to 2021. It can be clearly seen that the amount of payments made to physician was roughly more than 120M from 2011 till 2016. The payment amount suddenly dropped by half to 40M in 2017 and didn't increase much after that. In 2016, the sum of payments done to Nurse Practitioners reached a record high of approx 40M.

B. Archana

Figure 8 depicts the scores of the infections of every country compared to the bar set by that of the United Nations i.e. the US.In this we can see that 45% of the country have infection scores worse than the US. Only a meager 10% have score better than the US. 30% have no better than the US and 15% are not even available. From this we can infer that the infection

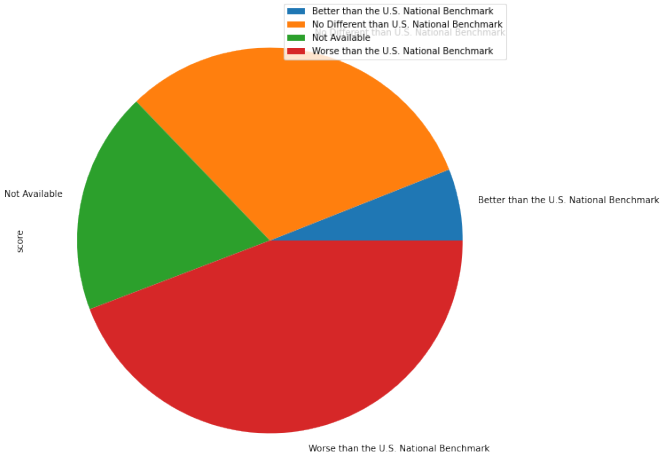

Fig. 8. Pie Chart for comparing scores to US region Providers

rate in almost all countries are higher or equal to that of the US which is a major reason for concern.

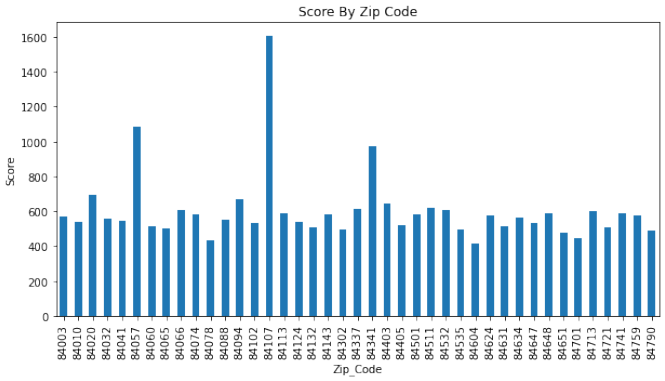

Fig. 9. Bar Chart of scores to different zipcode

Figure 9 depicts the scores of the infections of different countries. Here we can see that the zip_code 84107 has the highest score of 1600 for infection and 84078 has the lowest score of 580. 84057 and 84341 also have very high scores of 1100. Rest all zipcodes have a moderate score. This means that the zipcodes with highest score have to take better measures at preventing infections.

Figure 10 depicts the scores of the infections of different measures. Here we can see that all the infections have some high and low scores in all the zipcodes. The highest number is 30 where as the lowest score is 0. There needs to be takes preventive measures to reduce the infections in all the countries. Some are highly contagious and can be of huge risk

C. Rutuja

The boxplot below in figure 11 shows the computed value for the measures taken by Utah throughout the year 2010 to 2017. It says that the mean 24.98 with the minimum and maximum computed value as 20 and 30 respectively.

Fig. 10. Scatter Plot for scores of different Infections

Fig. 11. Boxplot for the Computed value

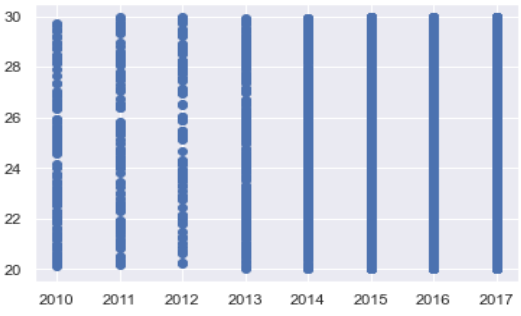

Fig. 12. Scatter Plot for the Utah average

Figure 12 shows the line scatter plot for the Utah average from the year 2010 to 2017. It can be clearly seen that the Utah average stands for approximately for 24 for the years 2010, 2014 and 2015. Rest of the years have 25 as the overall average. The Utah average has been consistent 2013 onwards.

The health plan types offered by Utah are Chip, HMP, PPO, Commercial HMO PPO, marketplace and medicaid and are highlighted in figure 13. The medicaid and PPO had almost the same percentage contribution of 26.98% and 26.65% respectively with marketplace having the least 3.72%.

D. Marmik

The figure 14 indicates the percentage of weekend report count for all the counties in New York. The above pie-chart highlights that county Suffolk has the highest percentage of Week end report count which is 10.7% followed by ERIE 8.51% and Queens 7.8%. This means that the highest number of testing reports were carried out in county Suffolk. The

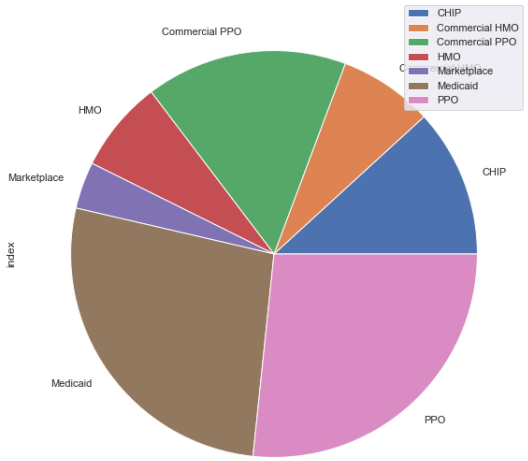

Fig. 13. Piechart for the Health Plan Type

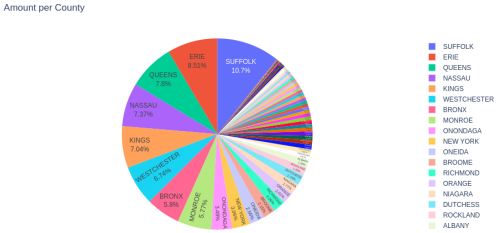

Fig. 14. Piechart for Weekend Report Count

county Albany, Rockland and Dutches had the least weekend_report_count which is less than 1% each. More than 50% of the weekend_report_count was registered from just eight counties while the rest 50% was divided unequally among all the remaining counties in New York city.

The above figure 15 tries to highlight the correlation of facility in the county with respect to various columns such as weekend report count, confirmed in facility, confirmed out facility, presumed in facility. Highest number of positive correlation can be seen between the count Queens and confirmed out of facility. We can notice that county Suffolk has a high positive correlation with confirmed in facility while Schoharie has the least positive correlation among all.

VI. Conclusion

This paper highlights various analytical methods to figure out the impact of health care factors in the healthcare industry. With the analysis from the HRIP data, it can be inferred that how the various incentive programs had the impact on total expenses done by people across different healthcare providers. The HAIH data tells us about the comparative study on the number of infections occurred when the patients were admitted to the hospital from different countries. HEDIS is a dataset

Fig. 15. Correlation Heatmap

that analysis the preventive measures and the care taken and the recovery time for the same for the state of Utah. It can be inferred from the CNH data about the reported and the lab confirmed cases in the nursing and adult care facilities. A analysis in depth could be done by taking into consideration about how the government can help in increasing the effective measures effectively, improving the incentives program and how the daily recorded cases could be reduced. These are the factors that could be analysed and then had a comparison with the time series analysis. Use of some machine learning can also be done for finding the relationship amongst the dependent and independent variables.

REFERENCES

[1] F. Collard and J. M. Garibaldi, "Measuring healthcare decision aid effectiveness," 2012 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS), 2012, pp. 1-6, doi: 10.1109/CBMS.2012.6266353.

[2] G. Antonacci, A. Calabrese, A. D'Ambrogio, A. Giglio, B. Intrigila and N. L. Ghiron, "A BPMN-Based Automated Approach for the Analysis of Healthcare Processes," 2016 IEEE 25th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2016, pp. 124-129, doi: 10.1109/WETICE.2016.35.

[3] R. A. Bauder and T. M. Khoshgoftaar, "A Probabilistic Programming Approach for Outlier Detection in Healthcare Claims," 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 2016, pp. 347-354, doi: 10.1109/ICMLA.2016.0063.

[4] P. A. Harsha Vardhini, S. S. Prasad and S. N. Korra, "Medicine Allotment for COVID-19 Patients by Statistical Data Analysis," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), 2021, pp. 665-669, doi: 10.1109/ESCI50559.2021.9396830.

[5] C. Antonini, S. Calandrini, F. Stracci, C. Dario and F. Bianconi, "Dynamical modeling, calibration and robustness analysis of COVID-19 using Italian data," 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), 2020, pp. 452-457, doi: 10.1109/BIBE50027.2020.00079.

[6] C. Nobre, N. Gehlenborg, H. Coon and A. Lex, "Lineage: Visualizing Multivariate Clinical Data in Genealogy Graphs," in IEEE Transactions on Visualization and Computer Graphics, vol. 25, no. 3, pp. 1543-1558, 1 March 2019, doi: 10.1109/TVCG.2018.2811488.

[7] M. H. Tekieh and B. Raahemi, "Importance of data mining in healthcare: A survey," 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2015, pp. 1057-1062, doi: 10.1145/2808797.2809367.

[8] S. Purushe, G. Grinstein, M. B. Smrtic and H. Lyons, "Interactive Animated Visualizations of Breast, Ovarian Cancer and Other Health Indicator Data Using Weave, an Interactive Web – based Analysis and Visualization Environment," 2011 15th International Conference on Information Visualisation, 2011, pp. 247-252, doi: 10.1109/IV.2011.108.

[9] A. Y. Noaman, N. Al-Abdullah, A. Jamjoom, A. H. M. Ragab, F. Nadeem and A. G. Ali, "Knowledge Based e-Health Surveillance System for Predicting Hospital Acquired Infections," 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), 2018, pp. 345-351, doi: 10.1109/COMPSAC.2018.10255.