Research to find critical factors affecting human health

Rishabh Singh Chauhan
MSc. Data Analytics
National College of Ireland
Dublin, Ireland
x21107939@student.ncirl.ie

Himanshu Duragkar
MSc. Data Analytics
National College of Ireland
Dublin, Ireland
x20210639@student.ncirl.ie

Akash Manjunatha
MSc. Data Analytics
National College of Ireland
Dublin, Ireland
x21141797@student.ncirl.ie

Kartik Sharma
MSc. Data Analytics
National College of Ireland
Dublin, Ireland
x21125813@student.ncirl.ie

Abstract— The importance of health in one's life cannot be overstated. Humans have become more aware of the value of health and lifestyle in the last two years. Currently, every group (government or non-governmental) is collaborating to lessen the impact of Covid-19. Every data analyst, statistician, and stakeholder on the planet is working nonstop to discover ways to improve human health and save humanity from terrible infections. There is a lot of data to analyze and extract insights in the health domain. Each of the four datasets chosen is significant to the global healthcare system. We hope to learn more about how numerous risk factors, such as smoking, air quality, and blood pressure, affect human health through this initiative. The four datasets were subjected to exploratory data analysis. The four data sets are subjected to a series of data analyses using the Python programming language. The inquiry is furthered by connecting to databases such as MongoDB and PostgreSQL using the Python programming language. Visualization is created using the Matplotlib, Seaborn, and Plotly tools to better understand the trend.

Keywords—Air quality index, covid data, Lung Cancer, Death rate by risk factor.

I. INTRODUCTION

The objective of our project is to identify various risk factors that affect human health. The adaptation of digital technology produces a huge amount of data to collect such as patient records, death rates, disease trends or seasons etc. Using analytics to analyse health data can lead to outcomes that benefit everyone in our community. In the past years acquiring large amounts of data for medical purposes has proven difficult, time-consuming, and occasionally costly. However, with today's technological advancements, it is possible to save data and also analyse them to build comprehensive healthcare reports and turn them into important insights. Healthcare is one of the data-rich industries, with medical advancement and research occurring at a rapid rate. Through healthcare data hospitals and governments can understand patterns and make appropriate decisions in their sector. Through exploration of health data, we can plan finance resource allocation, as well as advise the government and concerned organizations on when and how much to allocate for a given condition or disease. The government and hospitals can pre-planned themselves for a particular disease and can make the arrangements according to them. Let's take the example of Covid-19. Through insights and exploration from data of Covid-19 government and hospitals can now pre-planned on various components such as Medicines, availability of beds in the hospital, ICU wards, finance allocation, essential goods, etc. We considered four different datasets with various characteristics in this study, intending to highlight inter-reliability as well. We are conducting a study in this area since the healthcare industry needs more data investigation to prepare for future requirements.

II. RELATED WORK

In this paper research has been carried out to find the result of the COVID-19 lockdown, changes in air quality and atmospheric composition have occurred in the United Kingdom As a result, social participation, movement, and non-essential firms and services are restricted. The findings revealed that NO2 concentrations declined by a mean of 14 to 38 percent during a five-year period, owing to traffic reductions of over 70 percent. Yet, there was a continuous increase in average ambient O3 with only sporadic indications of particle matter decline. [1]

This study is based on a fuzzy air quality index for AQI evaluation; It says, utilizing simply the AIQ index to assess AQI cannot provide a correct result; additional parameters involved (That is CO, NO, PM's and O3,) in the cause of pollution must be included in order to estimate the impact on human health.[2]

This study examines the association between air pollution and health and deprivation in order to improve the UK's air quality index management policy. For analysis, they separated the population into four groups depending on income levels, and air pollution was divided into four categories. Low-income neighbourhoods, where the most vulnerable people resided and where the greatest health needs occurred, had the highest levels of air pollution, which had the strongest linkages to respiratory diseases, lung cancer, and mortality.[3]

Manual surveillance of hospital-acquired infections takes time and is usually limited to the intensive care unit (ICU).Computer-assisted strategies for identifying hospital-acquired infections can improve efficacy.The authors also recommend that a cutting-edge knowledge-based e-Health surveillance system be developed for anticipating hospital-acquired infections.[9]

Many causes threaten the lives of every citizen in developing countries, including population increase, a lack of adequate medical resources, and the fear of new diseases on the horizon. As a result, the development of medical resources in these areas is always increasing. Environmental variables, such as people's lifestyle, are another component that causes such medical problems. With these two conditions, the odds of becoming unwell are extremely high. This information can assist clinicians in improving the accuracy of disease diagnosis.[10]

The United States has highest death toll in the world due to Covid-19. After many researhes ,it was found that the rate of Covid-19 infection depends on many factor like race/ethinicity, age and prevailing health conditions. The Statistical Analysis and Machine Learning was used to get more insights on the Covid-19 infection rates in various Individuals.[11]

III. METHODOLOGY

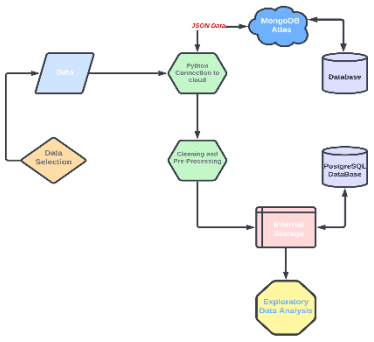

Fig. 1. Project Flow

A. *Technologies used:*

The project has utilized MongoDB Atlas as the cloud storage solution, local deployment of PostgreSQL Database. The whole project is done using Python programming language using Jupyter Notebook on the Microsoft Windows Operating System.

B. *WorkFlow*

1) Data Selection: In this study, we used 4 distinct datasets, the specification of the same provided below.

a) Risk Factors Data: The data is sourced from W.H.O. data repository.

Data Description:

Variables	Description	Variables	Description
Entity	Country name	Diet low in vegetables	Deaths due to diet low in vegetables
Year	Years sample collected	Unsafe sex	Deaths due to unsafe sex
Unsafe water source	Deaths due to unsafe water source	Low physical activity	Deaths due to low physical activity
Unsafe sanitation	Deaths due to unsafe sanitation	High fasting plasma glucose	Deaths due to high fasting plasma glucose
No access to handwashing facility	Deaths due to no access to handwashing facility	High total cholesterol	Deaths due to high total cholestrol
Household air pollution from solid fuels	Deaths due to household air pollution	High body-mass index	Deaths due to high body-mass index
Non-exclusive breastfeeding	Deaths due to discontinued breastfeeding	High systolic blood pressure	Deaths due to high systolic blood
Discontinued breastfeeding	Deaths due to child wasting	Smoking	Deaths due to smoking
Child wasting	Deaths due to child wasting	Iron deficiency	Deaths due to iron deficiency
Child stunting	Annual Deaths due to child stunting	Vitamin A deficiency	Deaths due to vitamin A deficiency
Low birth weight for gestation	Deaths due to low birth weight due to gestation	Low bone mineral density	Deaths due to low bone mineral
Secondhand smoke	Deaths due to secondhand smoke	Air pollution	Deaths due to air pollution
Alcohol use	Deaths due to alcohol use	Outdoor air pollution	Deaths due to outdor air pollution
Drug use	Deaths due to drug use	Diet high in sodium	Deaths due to diet high in sodium
Diet low in fruits	Deaths due to diet low in fruits	Diet low in whole grains	Deaths due to diet low in whole grains
		Diet low in nuts and seeds	Deaths due to diet low in nuts and seeds

Fig. 2. Risk Factors dataset description.

b) Lung Cancer: Lung Cancer classification is taken from kaggle a Google LLC subsidary where data is present publicly all over the globe.

Data Description:

Variables	Description	Variables	Description
Name	Name of the person	Smokes (packs/year)	No. of cigarette packs he/she smokes per year
Member_ID	ID of the person	AreaQ	-
Diagnosis	Stage of cancer	Alkhol	Alchol consumption
Age	Age of the person	family history	outcome if Family had cancer
Smokes	No. of cigarette he/she smokes	Result	Outcome Positive -1 Negative -0
Smokes (years)	No. of cigarette he/she smokes per year		

Fig. 3. Lung Cancer classification dataset description.

c) Pollutants and Air Quality data: Air quality index data: which is taken from the GOVT website link (Liverpool city council website)

Data Description:

Variables	Units	Description	Variables	Units	Description
Organisation	NA	Name of the organization	CO (ppb)	Parts per billion (ppb)	Carbon monoxide
Device name	NA	Name of the devices(Two Stations)	O3 (ppb)	Parts per billion (ppb)	Ozone
datetime	NA	Date and time the data is	AQI PM10	NA	Subindex PM10
Device ID	NA	Device's ID	AQI PM25	NA	Subindex PM2.5
PM1	µg/m	Particulate matter-1	AQI NO2	NA	Subindex No2
PM25	µg/m	Particulate matter-2.5(PM2.5 in µg/m³)	AQI O3	NA	Subindex O3
PM10	µg/m	Particulate matter-3(PM10 in µg/m³)	AQI CO	NA	Subindex Co
Temperature	°C	Temperature	AQI Max	NA	AQI(Max value among the sub indexes)
NO2 (ppb)	Parts per billion (ppb)	Nitrogen dioxide	Geolocation	NA	longitude and lattitude values collected

Fig. 4. Pollutants and Air Quality dataset description.

d) Conditions Contributing to Covid-19 deaths: The data has been sourced from *Centers for Disease Control and Prevention* website is a health agency of US whose only goal is to improve overall public health.

Data Description:

Variables	Description
Data As Of	Data Collected and stored till date
Start Date	Date of starting the data collection
End Date	Date of ending the data collection
Group	The Data is grouped as Total , By Year , By Month
Year	The Year for which the Data is collected
Month	The Month for which the Data is collected
State	The United States
Condition Group	The Conditions which can cause to the Covid-19 Deaths
Condition	Subset to the Condition Group
Age Group	Human age groups from 0-85+ years
Covid-19 Deaths	Total Deaths due to Covid-19 with other heatlth conditions included

Fig. 5. Covid-19 Dataset

2) Data Storage and Retreval: After sourcing the data using python programmig, it was stored on the MongoDB Atlas. All the data was ensured to be in JSON format before uploading it to the MongoDB collections.

The Data was again fetched from the MongoDB database and was cleaned and pre-precessed before storing it to locally deployed PostgreSQL Database.

The Data was then retrieved from the PostgreSQL Database into the *pandas* DataFrame , and the Exploratory Data Analysis was performed on it to generate insights.

3) Data Cleaning and Pre-Processing: During this step all the datasets were thoroughly investigated by each teammate , ensuring all the null values, missing values, duplicates and un-necessary data/columns were handled properly.

This cleaned, pre-processed data was then stored in PostgreSQL Database.

4) Exploratory Data Analysis :

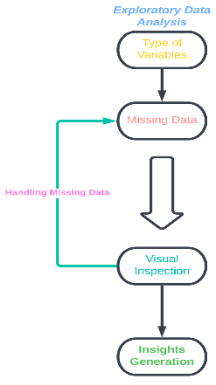

Fig. 6. Exploratory Data Analysis Process

The data was fetched from PostgreSQL database into a data frame and the analysis process began.

To get highly accurate insights, we again performed small cleaning process on the data. Once we were sure about the quality of the data, each team member started visualizing and extracting insights from their respective datasets.

III. RESULTS

Through graphs, charts, and plots we can see a clear and accurate picture of the data. In addition, we can spot a specific pattern, structure, or trend in the data. To visualize the datasets, libraries such as matplotlib, seaborn, and plotly were utilized. Basic graphs, histograms, pie charts, and scatterplots are all created with Matplotlib. We can compile all of the data into a single plot using seaborn and through plotly, we can create interactive plots.

A. Individual Dataset Analysis(Rishabh Singh Chauhan)

The World Health Organization provided this dataset, which shows the number of deaths for each risk factor from 1990 to 2017. This dataset contains 30 different risk factors. High systolic blood pressure, smoking, high fasting plasma glucose, and air pollution are the four biggest risk factors for death in the world, according to this Bar plot.

As may be seen, Because high systolic blood pressure, smoking, high fasting plasma glucose, and air pollution are all on the rise, we can conclude that they are all significantly connected. Their upward tendency is clear, reaching a pinnacle in 2017.

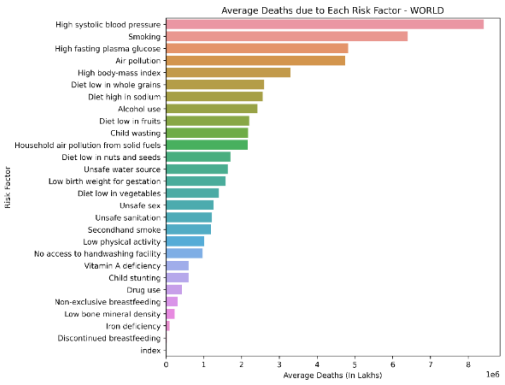

Fig. 7. Risk Factor VS Average Deaths

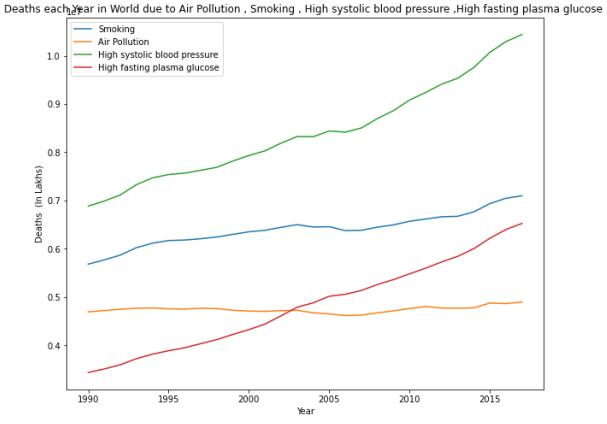

Fig. 8. Death VS Year

We may deduce from the heat map that air pollution is strongly linked to high systolic blood pressure, implying that air pollution is a major contributor to high blood pressure in humans.

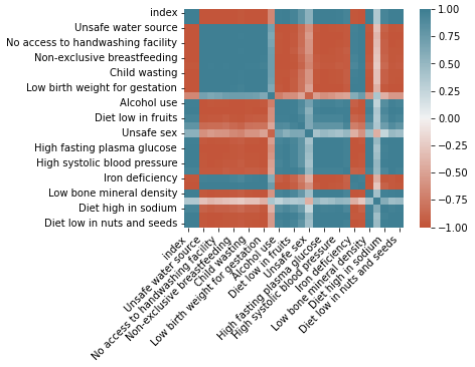

Fig. 9. Relation between factors

We may deduce from the heat map that air pollution is strongly linked to high systolic blood pressure, implying that air pollution is a major contributor to high blood pressure in humans.

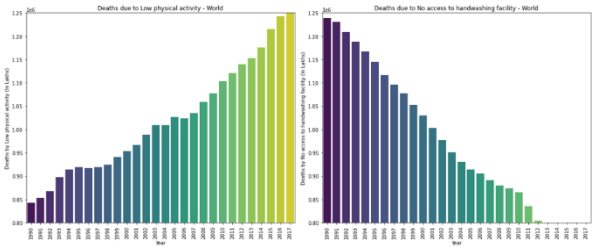

Fig. 10 (Left)Death VS Physical activity over years, (Right) Death VS Hand Sanitization over years

We can infer those human beings are moving away from physical exercise, which leads to higher death rates because the bar graph is rigorously growing. They have a stronger attachment to electronic devices than to physical activity. As we can see, there are fewer deaths today than there were in 1990, and this trend is continuing due to a lack of exercise.

We may say that human beings are heading towards a healthier and more sterilized environment because the graph is strictly declining. Death rates owing to a lack of access to handwashing peaked in 1990, however as time goes on, the rate is decreasing. As a result, we might conclude that humans have adopted the habit of handwashing because death rates are lower.

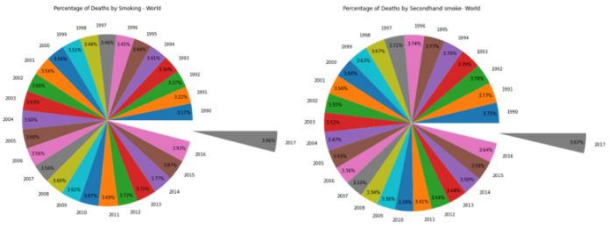

Fig. 11.(Left) Deaths due to smoking, (Right) Deaths due Passive Smoking

Figure 1 shows the number of smoking-related deaths from 1990 to 2017, whereas Figure 2 shows the number of deaths from second-hand smoke or passive smoking from 1990 to 2017. We can deduct from these two pie charts that the number of fatalities caused by smoking is the same as the number of deaths caused by second-hand smoke or passive smoking. This indicates that smoking is harmful not just to those who smoke, but also to those who are near them.

B. Individual Dataset Analysis(Himanshu Duragkar)

Fig. 12: We have divided the age category into three parts. Adult, Young and old depending the Age rank. Age 0 to 18 is

young, 18 to 50 is adult, 50 to 100 is old. The above figure shows the percentage of the age group who are smoking. Adult category has occupied most of the by smoking 66.10% of the whole. Then Old with 32.20% and young category take over 1.69%.

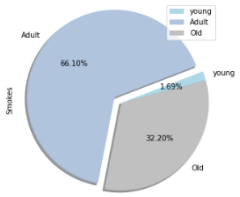

Fig. 12. AgeGroup VS Smoking

Fig. 13: As we can say from the be image, there is a actual growth cancer affection as the consumption of alcohol intake increases. On an average around 18 to 20 units of alcohol intake consumed by any category are mostly to get affected. The more the intake the more likely a individual can catch cancer.

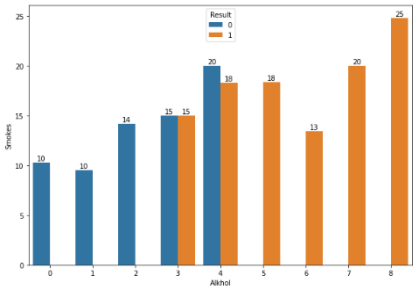

Fig. 13. Alcohol drinking and Smoking Vs Risk to cancer

Fig. 14: The visualization shows which categories alcohol takers belong to, as well as whether or not they have cancer. As we can see age rank of old category are mostly the people who consume alcohol and they are mostly affected by lung cancer. Age category of young and adult consumes alcohol but they are less lightly affected by the end result of cancer.

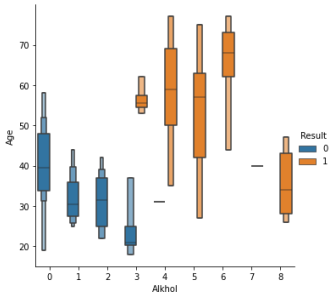

Fig. 14. Alcohol consumption Vs Age Group

Fig. 15: As we can see people who smokes around 120+ packs of smokes per year have the highest number of chances to get affected by cancer. Till the consumption of 100 packs there are less chances of any age category get affected by cancer. It directly shows the relation between smoking person and cancer result.

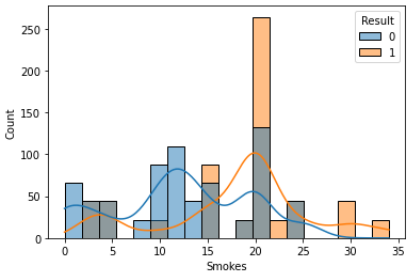

Fig. 15. No. of packs of cigarette vs Risk to cancer

C. *Individual Dataset Analysis(Akash Manjunatha)*

The time series line plot and pie chart depict the information regarding the distribution of Air quality affecting variables from 2021 to 2022(April). for line plot we took the mean value of variables per 24 hours and mean of the same for a month and the distribution of the same can be seen in the Fig.16 for an entire year

Fig16: In the line plot It can be seen from the graph, the gases NO2, CO, and O3 have a higher prevalence in the city's atmosphere. may be Due to malfunctioning heaters, fossil fuel burning, and traditional wood burning at homes [4] in the spring month from April to June Co marks its highest presence in the atmosphere reach near to the value 40 PPB, but in No2 gas, it is the month November. Ozone, on the other hand, had its peak presence in the month of February and has been steadily decreasing since then, reaching its lowest presence in the month of July before rising again. Particulate Matter (ex: dust, dirt, soot, or smoke) which mainly contributes asthma, throat and lung infection, which had been completely flat, suddenly rockets and reaches a peak in the mid-year which almost reach 12 PPB in the air and return to a position similar to that of the first quarter

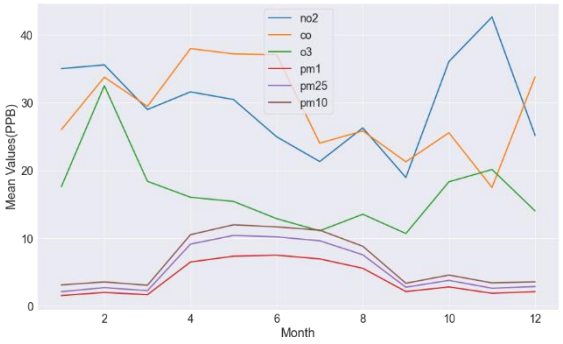

Fig. 16.Mean Values(PPB) Vs Month.

Fig 17 Pie chart: According to the pie chart, the city of Liverpool has managed to keep particulate matters well within the limit, contributing less than 20% of the total. The next constituent is ozone, which solely equates the percentage of particulate matters. O3 is primarily caused by industries, power plants, and trucks, and it primarily affects the lungs, children's health. [3] The two gases No2, Co (nitrogen dioxide and carbon monoxide) that make up, contribute nearly equally to more than 60% of the total, among other factors Co and No gas are emitted mostly as a result of the burning of fossil fuels for power generation, natural gas in homes or for heating

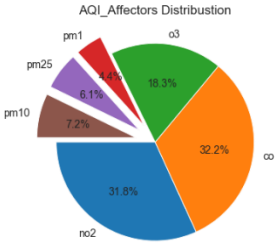

Fig.17.Pie chart Distribution of AQI affectors

Fig18: Overall air quality in the city of Liverpool has been maintained well within the range (considering average value grouped by month) during the period 2021 to 2022(April). Air quality in the first five months increased progressively until may month, as seen in the bar graph. May and June (seasonal change from spring to summer) have the greatest air quality index slightly above 100, while the rest of the month through the end of the year has air quality indexes of less than 40, with September having the lowest and the best air quality index of the year at 20.

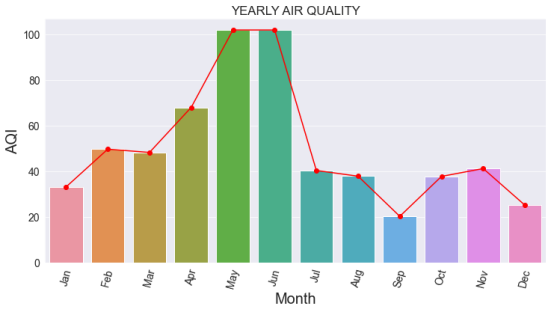

Fig. 18.AQI VS Month

Fig19: How is Air quality is graded? A government website called AQInow [5] which gives the grade basics categorization AQI is divided in to five category with value less than AQI value 50 as Good, between 50 to 100 moderate, between 100 to 150 unhealthy for sensitive people, between 150 to 200 unhealthy, between 200 to 300 very unhealthy and 300 and more Hazardous, pie chart depicts only close to one percentage of unfit air quality index in the city, Moderate and good AQI which constitutes almost three-quarter which consider to be the better AQI to breathe whereas almost quarter proportion that gas emitted over year (100-150) found to be unsensitive grade for the sensitive groups, The values appears to be realistic. The IQ Air report [4] of the city Liverpool Every year, 1040 people die as a result of being exposed to excessive levels of NO2 and PM2.5 in the air.

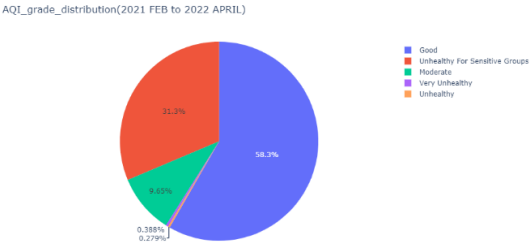

Fig. 19.AQI Grade Distribution

D. *Individual Dataset Analysis(Kartik Sharma)*

Fig. 20: 23% of covid-19 Deaths are majorly due to disease or condition involving the lungs (influenza, pneumonia and Respiratory Failure)

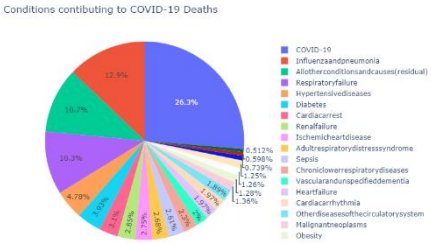

Fig. 20. Condition Contributing to COVID-19 Deaths

Fig. 21: The Graph suggests that a person suffering from Covid 19, and having other ailments (Influenza, pneumonia or Respiratory failure) is more susceptible to the risk of death.

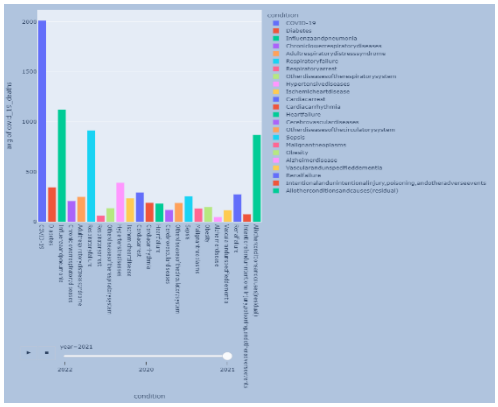

Fig. 21. Age Group Vs Health Condition

Fig. 22: Though the year 2020, the world saw the terror of COVID-19, the world suffered most in the year 2021, by the end of 2021, the vaccination increased, thus 2022 is seeing a decline in the deaths due to Covid-19.

Fig. 22. Covid Deaths over the years

Fig. 23: The Data Suggests, that individuals with an age of more than 54 years are more at risk of Covid19 mortality.

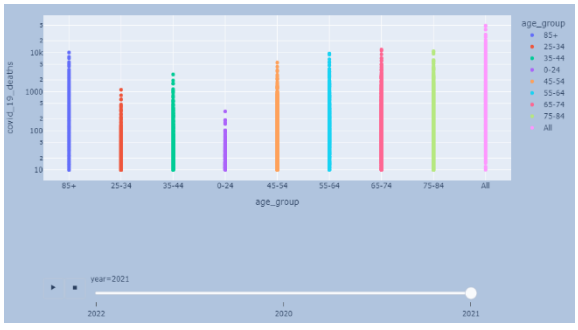

Fig. 23. AgeGroups Vs Covid Deaths over 3 years

Fig. 24: The US states - California, Florida and Texas have the major portion of the adult population above 54 years of age, which confirms the high mortality in these states also these 3 states have the highest population in the United States. According to data in 2019, 26% of the total population of California was above 54 years of age.[6]

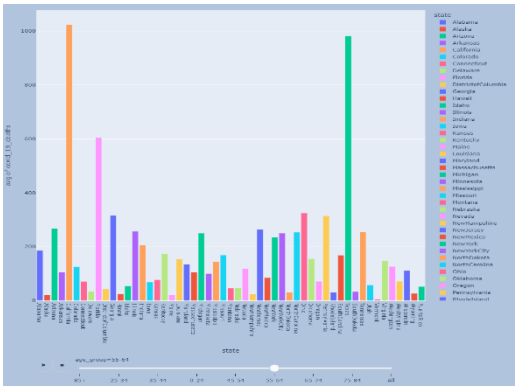

Fig. 24. Sates vs Covid Deaths for different Age Groups

Fig. 25: The Death Rates in the US have drastically fallen as compared to the years 2020 & 2021, but still, California and Texas followed by Florida are having high death rates, The reason behind this could be the high population as California, Texas and Florida are the top 3 most populated cities in the US.[7]

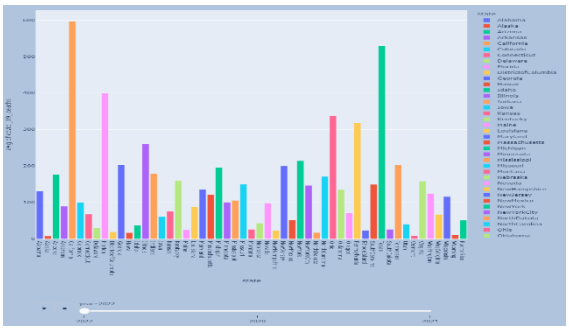

Fig. 25. Sates vs Covid Deaths for different years

The Lasso Regression with GridSearchCV is used find out which features in the dataset can predict the Covid-19 Deaths.The important features are Year , Month , health factors (Alzheimer, Covid-19, Diabetes, Injury, Malignantneoplasms, Obesity, Renal Failure , Respiratory , Sepsis , Vascular, Other Health Conditions), age group (0-24, 25-34, 35-44, 45-54, 65-74, 75-84, 85+).

IV. CONCLUSION AND FUTURE WORK

We've concluded that the quality of human life is heavily influenced by lifestyle, based on visualizations and insights extracted from each dataset. Air quality, physical fitness, lung health issues, smoking and drinking habits all play a significant influence in this way of life. Moreover, to this every individual is also responsible for its own health. This report can help government agencies and various NGOs to take further improvements to on how it can help people with number of services so that every citizen can have a good and healthy life by providing them information regarding environment and health issues and how it can affect a living being.

REFERENCES

[1] K. P. Wyche κ.ά., ‘Changes in ambient air quality and atmospheric composition and reactivity in the South East of the UK as a result of the COVID-19 lockdown’, Science of the Total Environment, τ. 755, σ. 142526, 2021.

[2] M. H. Sowlat, H. Gharibi, M. Yunesian, M. T. Mahmoudi, και S. Lotfi, ‘A novel, fuzzy-based air quality index (FAQI) for air quality assessment’, Atmospheric Environment, τ. 45, τχ. 12, σσ. 2050–2059, 2011.

[3] H. Brunt, J. Barnes, S. J. Jones, J. W. S. Longhurst, G. Scally, και E. Hayes, ‘Air pollution, deprivation and health: understanding relationships to add value to local air quality management policy and practice in Wales, UK’, Journal of Public Health, τ. 39, τχ. 3, σσ. 485–497, 2017.

[4] Iqair.com. 2022. Liverpool Air Quality Index (AQI) and United Kingdom Air Pollution | IQAir. [online] Available at: <https://www.iqair.com/uk/england/liverpool> [Accessed 25 April 2022].

[5] "AQI Basics | AirNow.gov", Airnow.gov, 2022. [Online]. Available: https://www.airnow.gov/aqi/aqi-basics/. [Accessed: 24- Apr- 2022].

[6] "Population Distribution by Age", KFF, 2022. [Online]. Available: https://www.kff.org/other/state-indicator/distribution-by-age/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D. [Accessed: 26- Apr- 2022].

[7] "US States - Ranked by Population 2022", Worldpopulationreview.com, 2022. [Online]. Available: https://worldpopulationreview.com/states. [Accessed: 26- Apr- 2022].

[8] 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.

[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.

[10] JIA WU, PEIYUAN GUAN AND YANLIN TAN "Diagnosis and Data Probability Decision Based on Non-Small Cell Lung Cancer in Medical System" 2019

[11] Chengzhen L Dai, Sergey A Kornilov, Ryan T Roper, Hannah Cohen-Cline, Kathleen Jade, Brett Smith, James R Heath, George Diaz, Jason D Goldman, Andrew T Magis, Jennifer J Hadlock, Characteristics and Factors Associated With Coronavirus Disease 2019 Infection, Hospitalization, and Mortality Across Race and Ethnicity, Clinical Infectious Diseases, Volume 73, Issue 12, 15 December 2021, Pages 2193–2204, https://doi.org/10.1093/cid/ciab154