International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 21s, 2025
https://theaspd.com/index.php

Preventing The Preventable: Analysis On The Behavioral And Environmental Factors Behind U.S. Chronic Diseases

Karnam Akhil[1], Sumanth Veluvolu[1,] Venkata Ramana Kaneti[2], Rushika Arvapalli[3], Sainath Kammiti[4]
[1,2]Assistant Professor, VNR Vignana Jyothi Institute of Engineering and Technology
[3,4]Student, VNR Vignana Jyothi Institute of Engineering and Technology
akhilresearch18@gmail.com[1], sveluvo@gmu.edu[1,2*,] kvramana_2001@yahoo.com[2]
rushika0807@gmail.com[3], sainathkammiti@gmail.com[4]

Abstract. *Chronic diseases are a leading cause of mortality in the United States, substantially impacting public health and healthcare costs. This paper examines trends in chronic diseases using the U.S. Chronic Disease Indicators dataset. Cleaning and analysis were performed using data science techniques. The results point out strong region-specific disparities, highly correlated variables of physical inactivity versus chronic diseases, and negatively correlated vaccination rates versus preventable diseases. EDA was performed to interpret the behavior of data. Resulted findings could underscore the necessity for specific interventions that may reduce chronic disease prevalence or improve behavior. Analysis also shows cases the environmental conditions for the symptoms in US.*
Keywords: *Chronic Diseases, Physical Inactivity, Public Health, Vaccination Rates.*

1 INTRODUCTION

Chronic diseases include diabetes, cardiovascular diseases, and obesity as the leading causes of high mortality and expenditure on health in the United States. There is a great need to recognize the differentials in the prevalence of chronic diseases to correctly formulate applicable public health policies. This research paper finds the correlations found in the U.S. Chronic Disease Indicators dataset across states, temporally, and correlates of behavioral and prevention factors with chronic diseases as shown in Table 1.[1]

Table 1: Attributes of data

Columns/ Variables	Description	NOIR Data Types
YearStart / YearEnd	The start and end year of the data collection.	Interval
LocationAbbr	Represents the abbreviation for the state.	Nominal
LocationDesc	Represents the name of the state.	Nominal
DataSource	Indicates the origin of the data.	Nominal
Topic	The health-related category.	Nominal
Question	Highlights the specific question for the topic.	Nominal
Response	The type of measurement.	Nominal
DataValueUnit	Unit of measurement for the value.	Nominal
DataValueType	Explains how the measured value is categorized.	Nominal
DataValue	Actual value measured.	Ratio
DataValueAlt	Alternative representation of the measured value.	Ratio
LowConfidenceLimit	Lower limit of the confidence interval.	Interval
HighConfidenceLimit	Upper limit of the confidence interval.	Interval
Stratification1	Groups data based on key demographic variables.	Nominal
StratificationCategory 1	Identifies the specific demographic group.	Nominal
Geolocation	Geographic coordinates for mapping.	Nominal

2 LITERATURE REVIEW

Hacker's research brings into focus the dramatic transition in global health from a predominance of infectious diseases to one now dominated by chronic conditions in morbidity and mortality rates. Chronic diseases, such as diabetes, heart disease, and cancer, are on the rise; lifestyle factors contribute to these increases. In that perspective, the COVID-19 pandemic has made things even worse, resulting in a decline in healthcare access and more health disparities among under-resourced communities. What this paper highlights is that not only are these chronic diseases avoidable, but also highly manageable with

proper public health measures and interventions at the community level. This brings into perspective how very urgent it is that greater focus be directed toward the prevention aspect of managing the diseases rather than focusing singly on treatment. The paper also discusses ways technology can help in handling these chronic diseases, with prospects for wearable devices and mobile health applications that will enable individuals to take part in monitoring their personal metrics of health.[2]

Benavidez et al.underlines the complicated interaction between the prevalence of chronic diseases and SDOH. In doing so, it determines significant gaps in the current literature; among them is the way analyses are being conducted at a county-level geography instead of at smaller units like the ZCTA. Such masking may blur variations in chronic disease rates and SDOH factors within counties. The paper identifies several dimensions of SDOH, including socioeconomic status, housing, transportation, and physical disability, because these factors interactively affect health outcomes. Previous studies have identified these dimensions as contributing to premature death, such as in Chicago, thus setting the stage for a nuanced understanding of how SDOH impacts chronic disease prevalence.[3][9]

Molnár, D et al. proposed the literature on the level of physical activity of children and adolescents is complex, and the findings are often contradictory. Children do not take part in sufficient levels of moderate to vigorous physical activities to promote health. For instance, Simon-Morton et al. concluded that children from North America often do not attain recommended levels of activity; similarly, the reviews by Cale and Almond of several studies of both young and secondary school-aged children reached the same conclusions about the low activity levels among each age group. The diversity in methodologies of the different studies, in terms of self-reporting, observational measures, and physiological assessment, makes it more difficult to make any specific conclusion about children's activity levels. Riddoch and Boreham further iterated that such discrepancies in the findings are more due to differences in methodologies rather than due to real variations in activity levels among various populations of children. These findings are therefore wide-ranging in their implications, especially regarding childhood obesity. This paper discusses how obesity among children has turned into a long-run negative health outcome, casting significant effects on the risk of developing chronic diseases in adulthood. Therefore, promoting physical activity among children is crucial, even though there is an inadequate agreed-upon precise quantity and intensity of exercise necessary for health maintenance. These data suggest that although many children might achieve threshold levels of activity, overall patterns are low to moderate rather than vigorous activities. This raises concerns regarding the health consequences of such activity patterns and the need for clarity on types and amounts of physical activity that can effectively prevent obesity and promote health in children.[4]

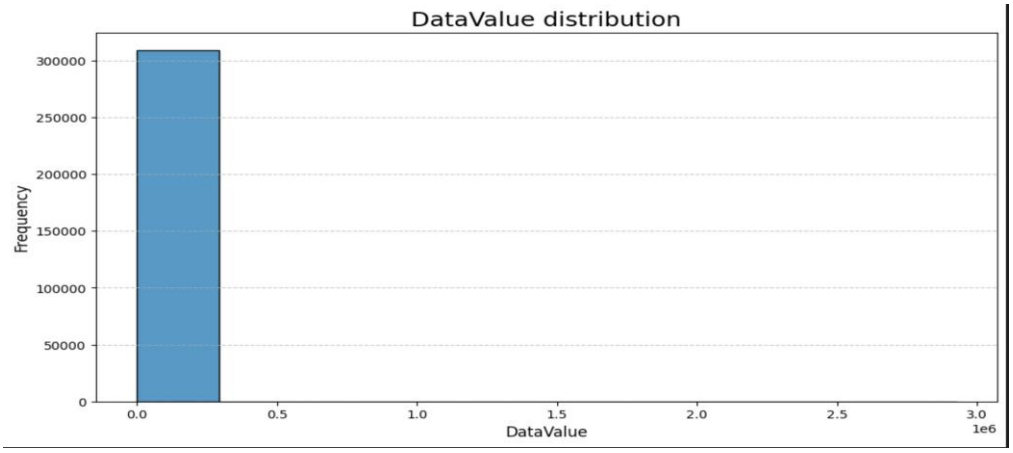
Santoro et al.focusses up on the results of similar research showing that vaccination coverage and the occurrence of outbreaks, particularly in settings with low vaccination rates, are related. According to the results obtained, although seropositivity rates regarding rubella and varicella are commendably high, the rates regarding hepatitis B are concerningly low, thus showing the need for better vaccination strategies among students of healthcare.[5][6][7][8][10]

3 Research Questions

1. How are chronic disease rates varying across US States and regions?
2. What is the relation between physical inactivity and the prevalence of chronic diseases?
3. How have chronic disease prevalence rates changed over time?

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 21s, 2025
https://theaspd.com/index.php

4 Analysis on the Data [1]



4.1 Univariate Analysis

Fig 1: Data distribution on univariate analysis

The above figure as shown in Fig 1 is a histogram of the distribution of the variable DataValue. The x-axis is DataValue, and the y- axis is the count of frequency. This graph indicates that the data is not very even since all data falls within a small range with few or no values distributed across the higher range of the DataValue axis.

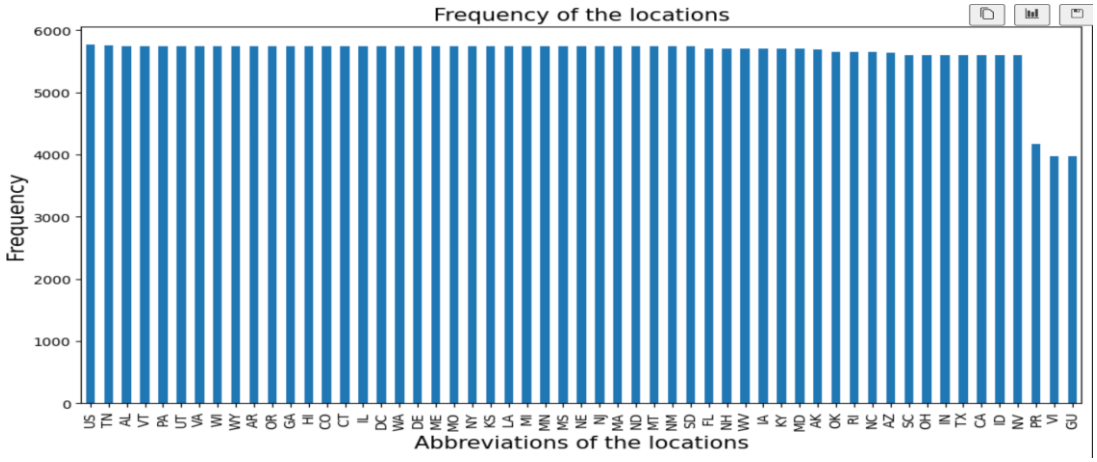
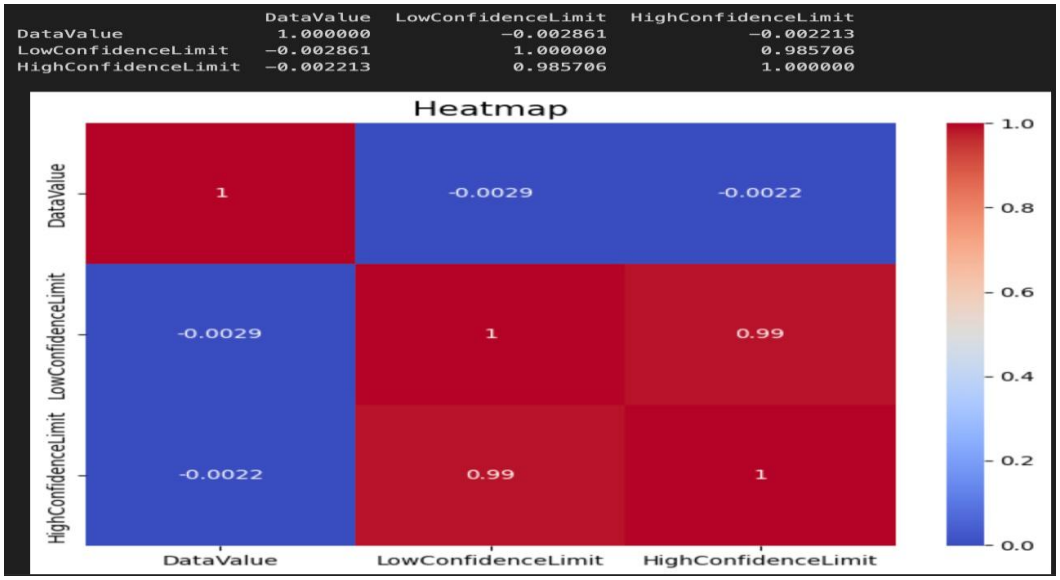


Fig 2: Frequency of the locations



As shown in Fig 2 a bar chart representing the frequency of occurrences for various location abbreviations within the dataset is visualized. On the x-axis, one can find the abbreviations of the locations; however, the y-axis illustrates the frequency associated with each abbreviation.

Fig 3: Heatmap visualization on the data

4.2 Multivariate Analysis

The univariate data analysis indicated that the DataValue variable is quite highly skewed with its mean being roughly around zero telling of the existence of outliers or data that is poorly distributed. The frequency distribution of location abbreviations (LocationAbbr) was uniform which indicated a fair representation of all the locations. A correlation heat map showed that LowConfidenceLimit and HighConfidenceLimit had very strong relationship (0.98576) as the coefficients between all the DataValue and each of the two confidence limits were all weak and slightly negative indicating very low linear relationships.

Fig 3 represents the visualization of a Heatmap on multi variate analysis.

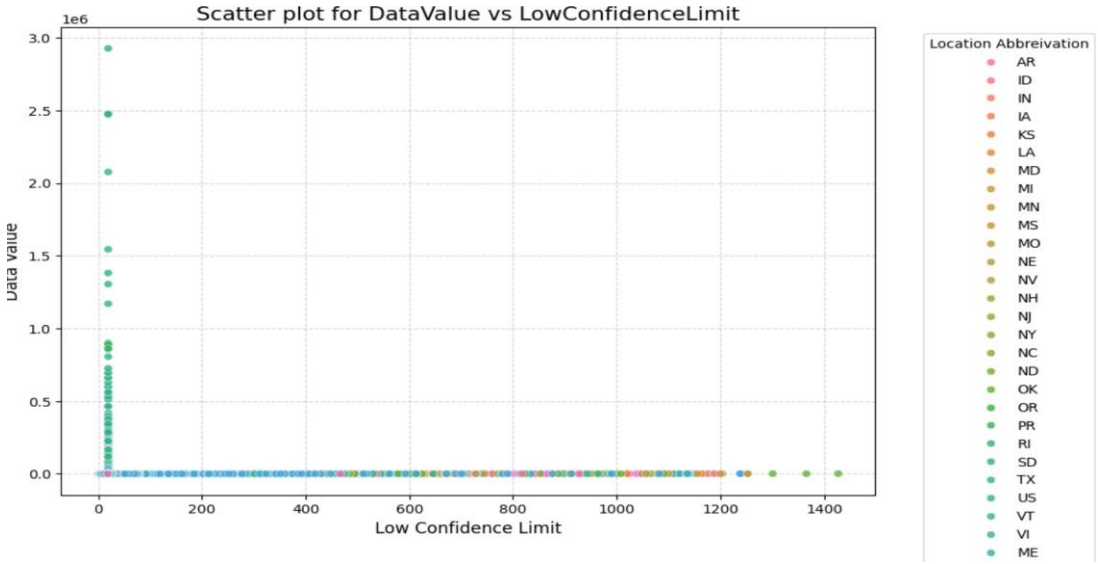
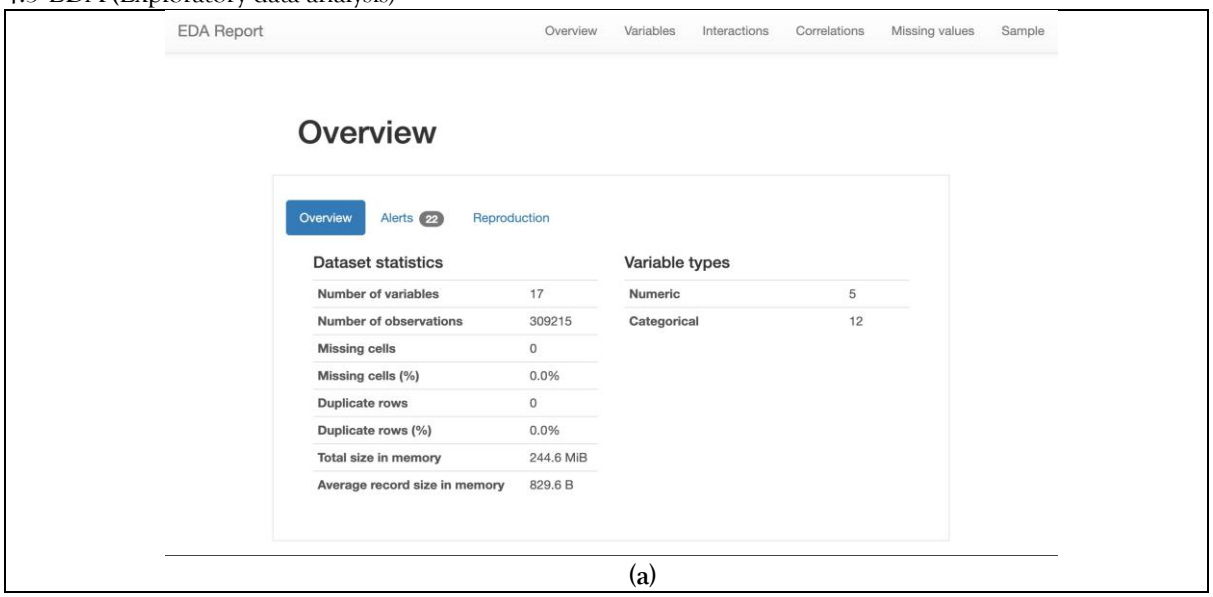


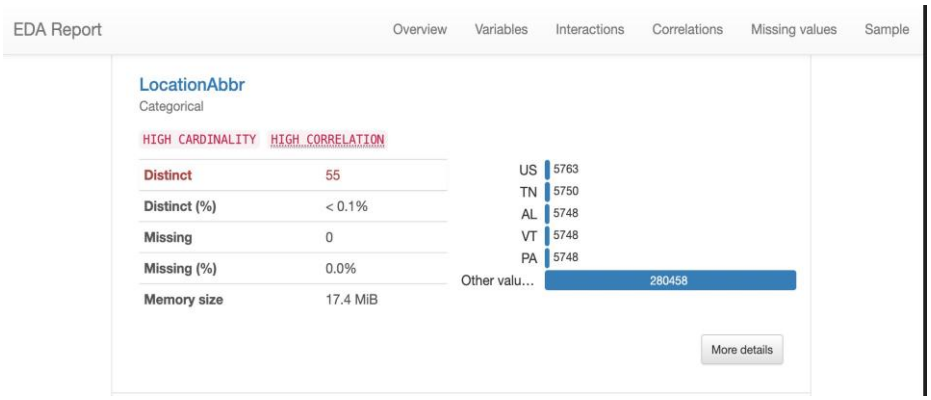
Fig 4: Scatterplot visualization for Data value and confidence value

As shown in fig 4 the scatter plot illustrates the dependence of DataValue and LowConfidenceLimit with the respective dots representing the data points by the location's abbreviation. It can be seen from the graph that more than half of the points are elbowed towards LowConfidenceLimit lower values, while DataValue has a large range and even a few extreme outliers in the upper range. Most of both y and x metrics show cluster near their minimum which means that for most of the data points the LowConfidenceLimit measurement used was not varying much. The color-coded legend explains the geographical spread of these data and hence regions can be compared.

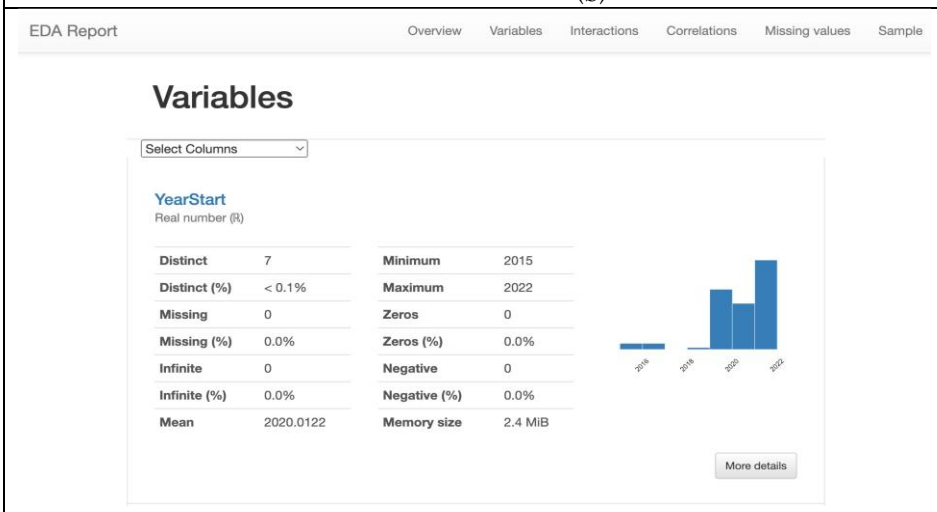
4.3 EDA (Exploratory data analysis)



(a)


EDA Report
Overview
Variables
Interactions
Correlations
Missing values
Sample
LocationAbbr
Categorical
HIGH CARDINALITY
HIGH CORRELATION
Distinct 55
Distinct (%) < 0.1%
Missing 0
Missing (%) 0.0%
Memory size 17.4 MiB
US 5763
TN 5750
AL 5748
VT 5748
PA 5748
Other valu... 280458
More details

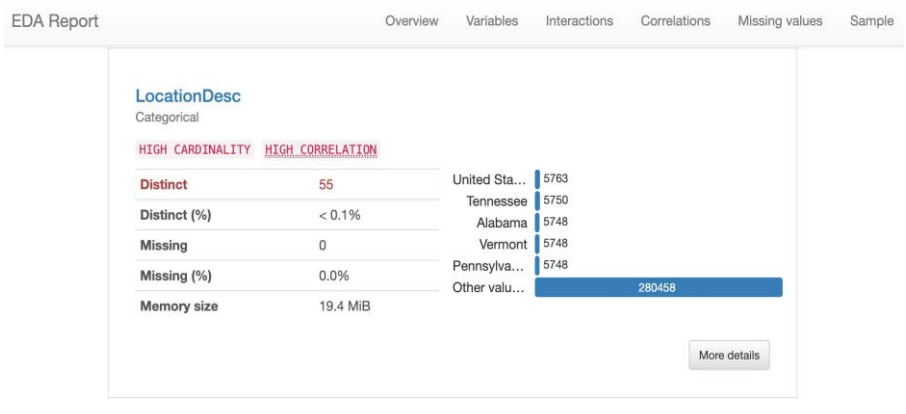
(b)


EDA Report
Overview
Variables
Interactions
Correlations
Missing values
Sample
Variables
Select Columns
YearStart
Real number (ℝ)
Distinct 7
Distinct (%) < 0.1%
Missing 0
Missing (%) 0.0%
Infinite 0
Infinite (%) 0.0%
Mean 2020.0122
Minimum 2015
Maximum 2022
Zeros 0
Zeros (%) 0.0%
Negative 0
Negative (%) 0.0%
Memory size 2.4 MiB
More details

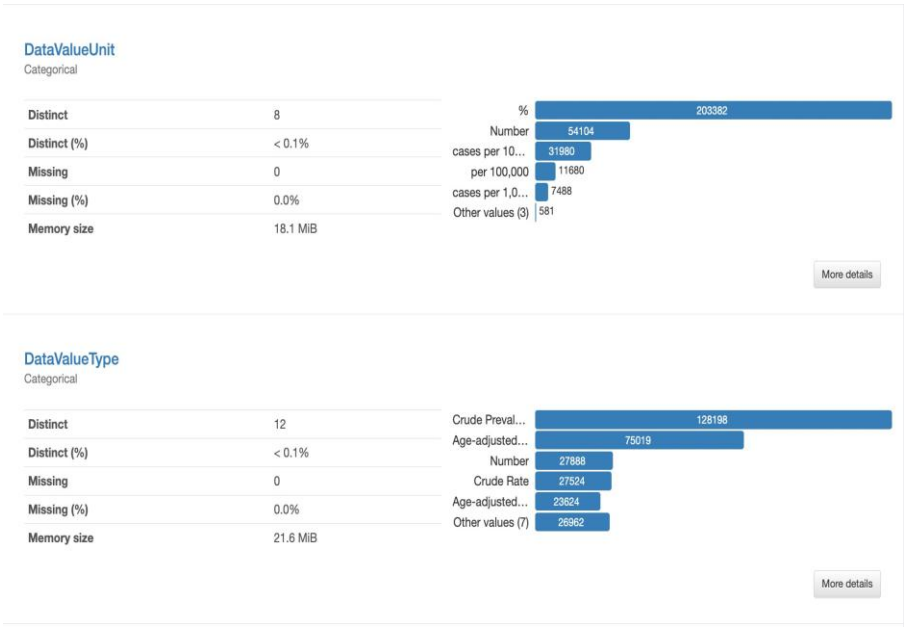
(c)


EDA Report
Overview
Variables
Interactions
Correlations
Missing values
Sample
DataSource
Categorical
HIGH CORRELATION
IMBALANCE
Distinct 14
Distinct (%) < 0.1%
Missing 0
Missing (%) 0.0%
Memory size 18.6 MiB
BRFSS 195594
NVSS 50232
YRBSS 17710
US Cance... 17520
CMS Part ... 11232
Other valu... 16927
More details

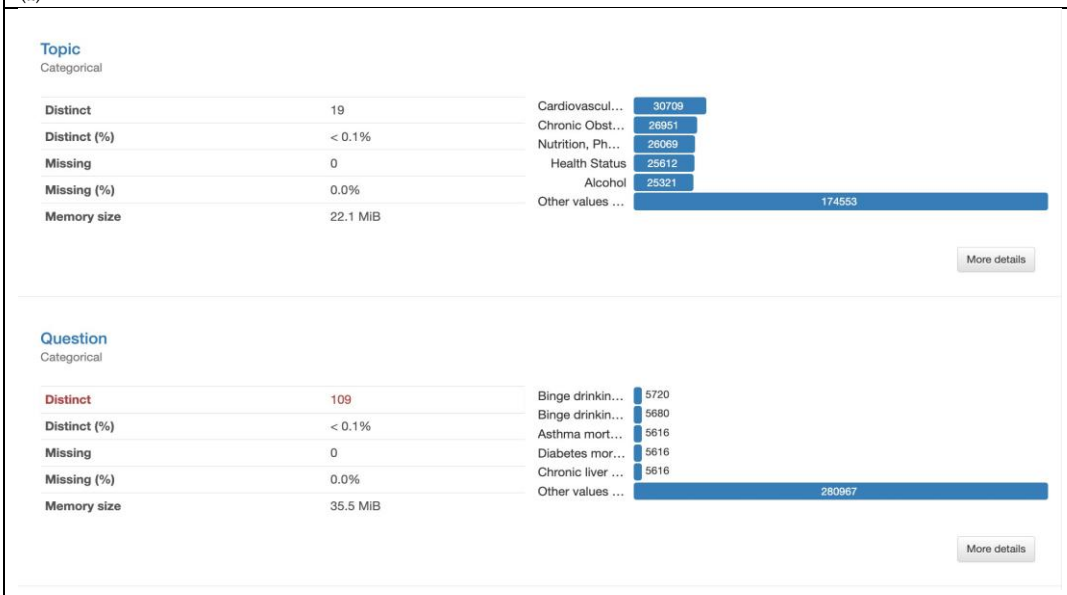
(d)


EDA Report
Overview
Variables
Interactions
Correlations
Missing values
Sample
LocationDesc
Categorical
HIGH CARDINALITY
HIGH CORRELATION
Distinct 55
Distinct (%) < 0.1%
Missing 0
Missing (%) 0.0%
Memory size 19.4 MiB
United Sta... 5763
Tennessee 5750
Alabama 5748
Vermont 5748
Pennsylva... 5748
Other valu... 280458
More details

(e)


DataValueUnit
Categorical
Distinct 8
Distinct (%) < 0.1%
Missing 0
Missing (%) 0.0%
Memory size 18.1 MiB
% 203382
Number 54104
cases per 10... 31980
per 100,000 11680
cases per 1,0... 7488
Other values (3) 581
More details
DataValueType
Categorical
Distinct 12
Distinct (%) < 0.1%
Missing 0
Missing (%) 0.0%
Memory size 21.6 MiB
Crude Preval... 128198
Age-adjusted... 75019
Number 27888
Crude Rate 27524
Age-adjusted... 23624
Other values (7) 26962
More details

(f)

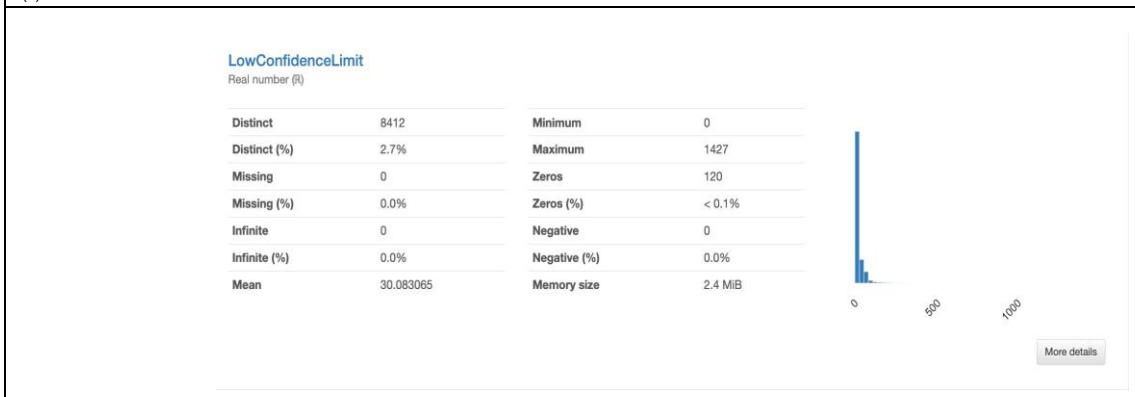

Topic
Categorical
Distinct 19
Distinct (%) < 0.1%
Missing 0
Missing (%) 0.0%
Memory size 22.1 MiB
Cardiovascul... 30709
Chronic Obst... 26951
Nutrition, Ph... 26069
Health Status 25612
Alcohol 25321
Other values ... 174553
More details
Question
Categorical
Distinct 109
Distinct (%) < 0.1%
Missing 0
Missing (%) 0.0%
Memory size 35.5 MiB
Binge drinkin... 5720
Binge drinkin... 5680
Asthma mort... 5616
Diabetes mor... 5616
Chronic liver ... 5616
Other values ... 280967
More details

(g)

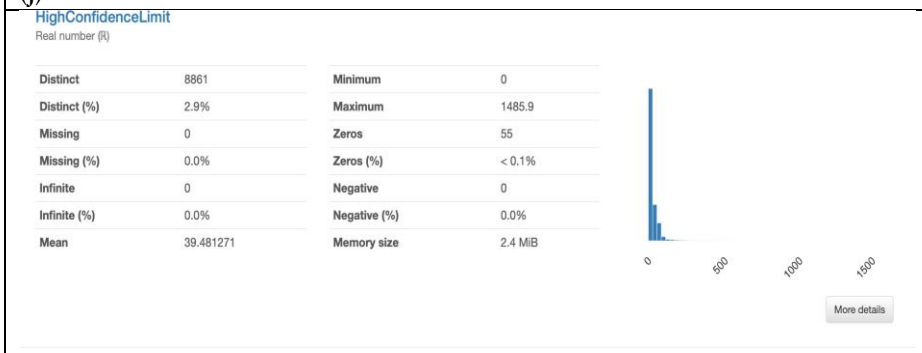

DataValue
Real number (ℝ)
HIGH CORRELATION SKEWED
Distinct 14852
Distinct (%) 4.8%
Missing 0
Missing (%) 0.0%
Infinite 0
Infinite (%) 0.0%
Mean 478.49292
Minimum 0
Maximum 2925456
Zeros 269
Zeros (%) 0.1%
Negative 0
Negative (%) 0.0%
Memory size 2.4 MiB
0
1
2
3
1e6
More details

(h)

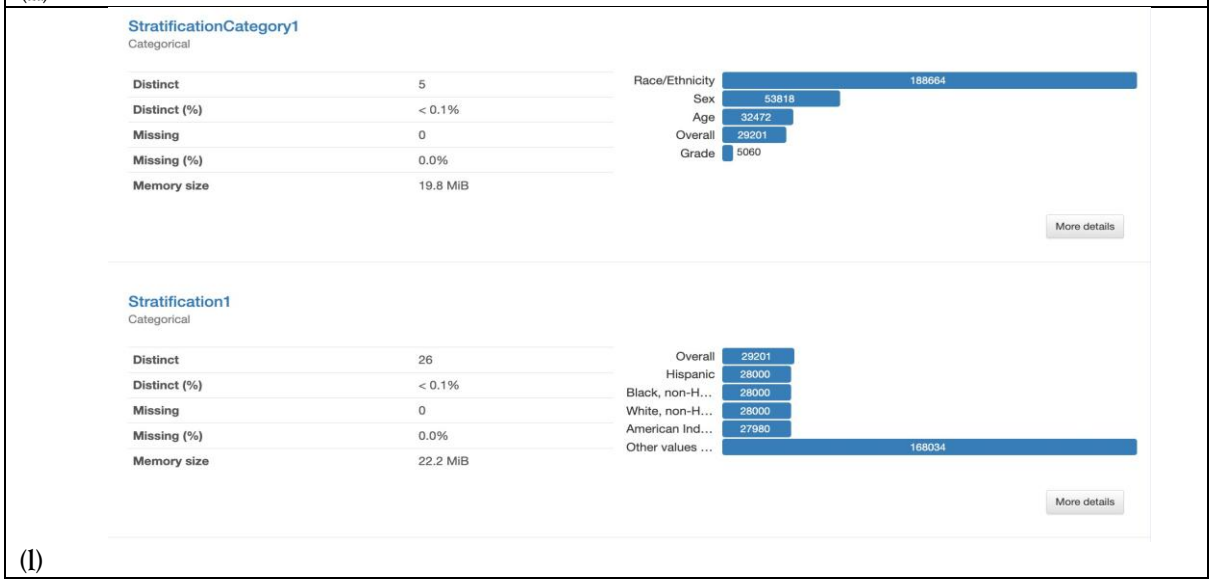
(i)



(j)



(k)



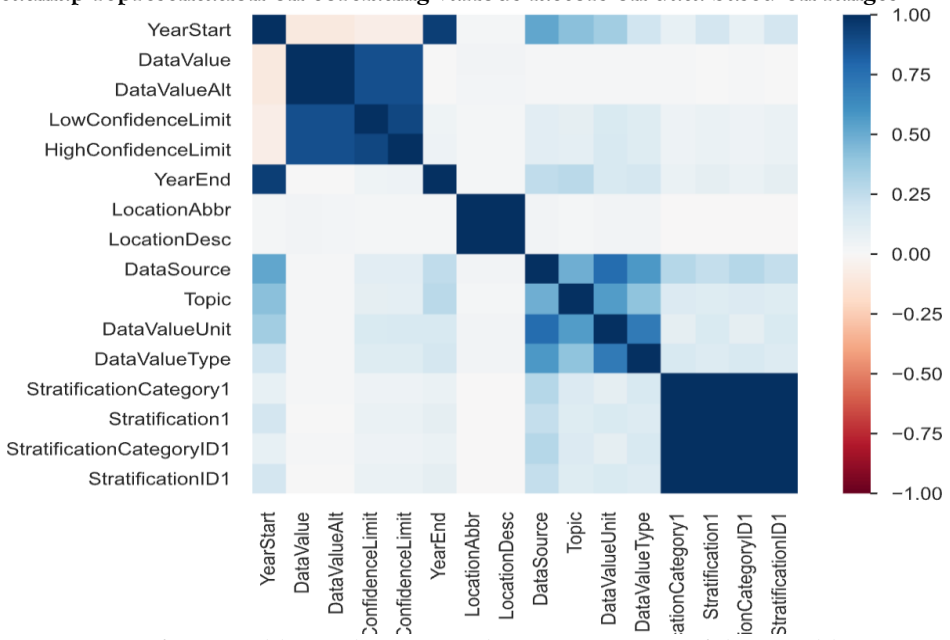
(l)

Fig 5: EDA analysis on data

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 21s, 2025
https://theaspd.com/index.php

As shown in Fig 5 the descriptive statistics EDA for the data in the dataset is shown. It indicates some important aspects and features of the variables. The above figures (a)(b)(c)(d)(e)(f)(g)(h)(i)(j)(k)(l) represents the EDA

Fig 6: Heatmap representation on corelating various factors on data based on ranges



The dataset consists of 17 variables and 309,215 observations. Out of the variables, 5 are numeric while 12 are categorical. The YearStart variable is taking values from 2015 all the way to 2022, a total of 7 unique values with no outliers such as those that are blank, zero or less than zero. Fig 6 represents Heatmap analysis of data.[1]

5 Research Questions Answers

1. How are chronic disease rates varying across U.S. States and regions?

To understand how the chronic disease rates are varying across U.S. States and regions, a bar graph helps in understanding about the variation of chronic disease rates.

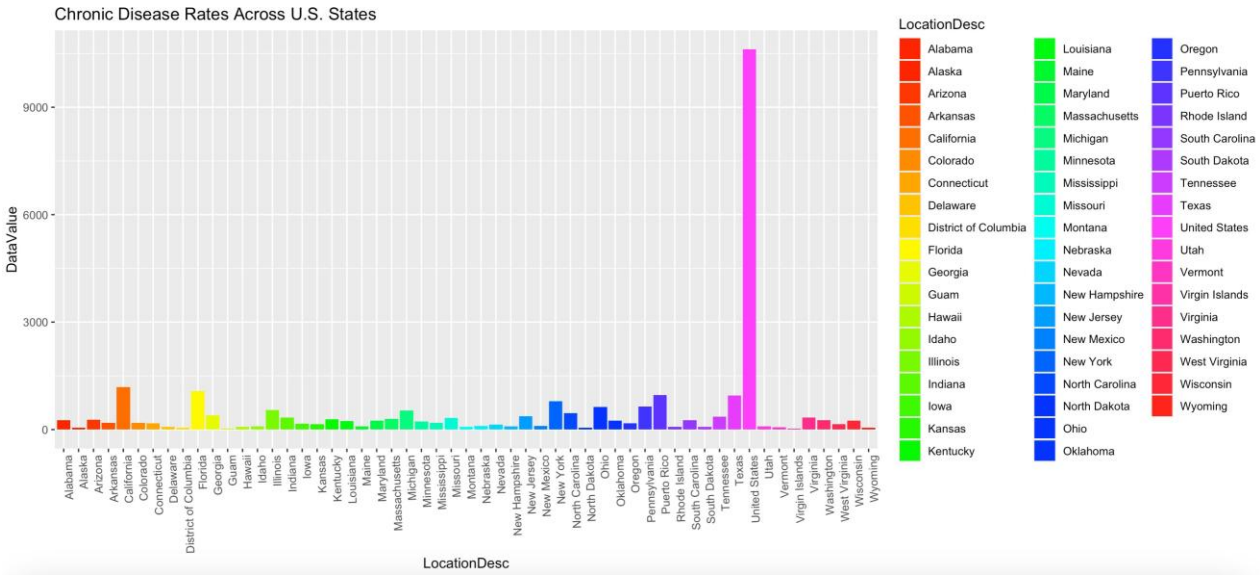
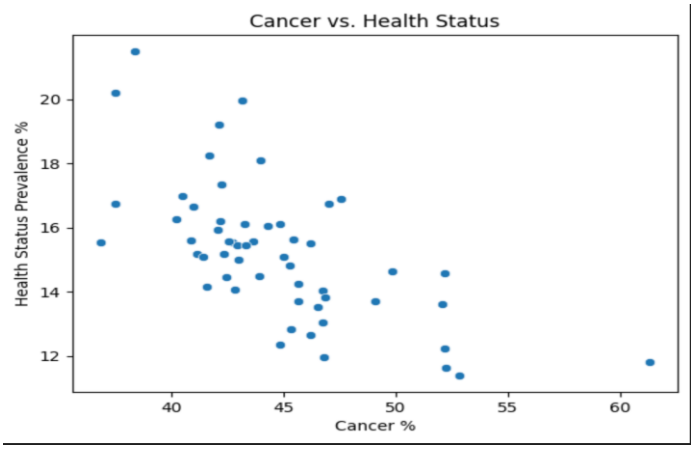


Fig 7: Bar chart representation on chronic diseases varying in US states

The above bar chart as ahown in Fig 7 the different chronic disease rates in the various states, here represented by the variable DataValue. Each bar represents a state, and different colors have been assigned to each state to facilitate comparison. United States has a much larger chronic disease rate compared with the other states, shown here by the tall pink bar. Most states have relatively low or moderate chronic disease rates, with small variations among them. The graph highlights geographical disparities in health outcomes, emphasizing the need for targeted interventions in states with higher rates.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 21s, 2025
https://theaspd.com/index.php

2. What is the relation between physical inactivity and the prevalence of chronic diseases?
A scatterplot helps understand the relation between physical inactivity and the prevalence of chronic diseases as shown in fig 8 like cancer. The above scatter plot illustrates the relationship between cancer prevalence percentages and health status prevalence percentages. The data points suggest a generally negative correlation, indicating that higher cancer prevalence is associated with lower health status



percentages. The general health status of a population usually tends to worsen with an increase in the burden of cancer as shown in Fig.8.

Fig 8: scatter plot visualization to indicate physical inactivity and prevalence of symptoms

```
Correlation Summary:
Diabetes: Correlation = -0.62, P-value = 3.50e-07
Asthma: Correlation = -0.69, P-value = 4.18e-09
Mental Health: Correlation = -0.51, P-value = 6.09e-05
Arthritis: Correlation = -0.20, P-value = 1.36e-01
Health Status: Correlation = -0.64, P-value = 1.24e-07
```

Fig 9: Correlation summary
As shown in fig 9 the correlation summary of different diseases is represented.

3. How have chronic disease prevalence rates changed over time?
A line plot helps understand the chronic disease prevalence rates over time.

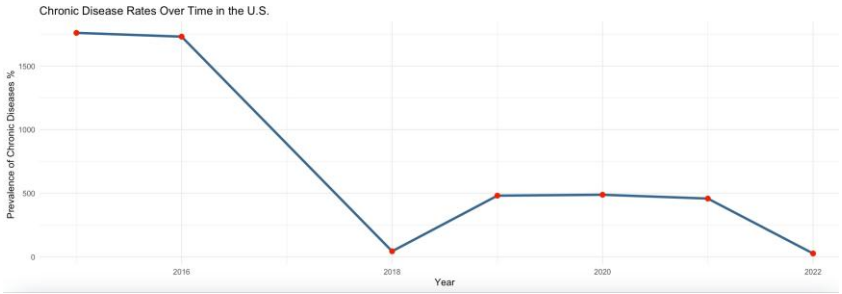


Fig 10: Time constraint prevalence se

The above line as shown in Fig 10 plot depicts the time trend of chronic diseases in the U.S., usually represented by the years against prevalence percentages. Rates of chronic disease during the time frame represented went from sharp declines between 2016 and 2018 through a robust period of stability between 2018 and 2020; the rates again decrease after 2020. The lowest point of chronic disease rates came in 2022. This set of data possibly shows effective interventions or measures that should have taken place between the periods being reviewed, particularly in or after 2016 and 2020. The interval between 2018 and 2020 would seem to imply that there was a leveling-off in progress.

6 Limitations and Future Scope

The dataset has some limitations. First, because it covers only numerical data, the data lacks important contextual information. Contextual variables that explain chronic diseases, such as access to health care, socioeconomic conditions, and cultural influences, are also not included in the data set. The data may prove very useful in understanding the rates of chronic diseases and the risk factors, it still lacks data such as individual-level behaviors, environmental conditions, and regional health care system performance that might further detail the trends in public health. Indicators at vaccination rates and prevalence, for example, are quite different from access to healthcare rates or treatment adherence rates in representing other important indicators; this includes longitudinal health outcomes that have the potential to capture far more realistic dynamics of chronic diseases within the U.S. population than is possible with these data, therefore allowing much greater utility in targeted public health interventions.

Adding other contextual variables help in providing more insights. Including a health care accessibility dimension along with socioeconomic factors, ecological and lifestyle habit factors increase the value towards a much greater understanding of underlying diseases determinants. Inclusion of demographic variables as age, race, and income while also behavioral patterning relationships can add up knowledge upon relationships between them and respective health outcomes. Moreover, the capturing of temporal dynamics-policy changes, economic developments, and environmental shifts-provides a necessary context in which to interpret the observed trends and design more effective public health strategies. Extending the data in these directions would also allow researchers and policymakers to better devise more specific, evidence-based interventions needed to meet chronic disease challenges in the United States.

7 CONCLUSION

The bar graph on chronic diseases for most of the states in that graph appears to show that most of them are moderate and few are higher in the value. This, therefore, presents regional differences that may be attributed to the difference in health outcomes based on varying access to healthcare, socio-economic conditions and lifestyle behaviors. These are the tallest bars which represent the highest rate for states and may indicate an area where more targeted interventions and public health initiatives might be needed. The scatter plot shows that health status prevalence and cancer percentage have a negative correlation; this means that as health status prevalence goes down—representing more physical inactivity—the percentage of cancer goes up. What this trend suggests is that communities with higher levels of physical inactivity are likely to increase their chronic diseases. While the relationship is clearly seen, the scatter of points suggests variability, hence suggesting that other factors may also be involved in this relationship. This again calls for the promotion of physical activity as a preventive measure to reduce the prevalence of chronic diseases. The line graph explains that in the U.S. the current rates of chronic diseases seem to have undergone a particular change. This is shown by a reduction in the number of times the disease occurred for 2016-2018, which I think is due to the possible success of health care interventions or public health programs that it was the time for thereof. From 2018 to 2020, the rates leveled off over this period, with little to absolutely no change. After 2020, prevalence rates began to decline again and hit their lowest in 2022. The general trend would indicate that over time, there is a decline in the prevalence of these chronic diseases; however, this period of stability from 2018 to 2020 indicates that continuous efforts have to be implemented to ensure that progress is sustained.

REFERENCES

1. Data.gov. (2024, March 9). U.S. Department of Health & Human Services - U.S. Chronic Disease Indicators. https://catalog.data.gov/dataset/u-s-chronic-disease-indicators
2. Hacker, K. (2024). The burden of chronic disease. Mayo Clinic Proceedings: Innovations, Quality & Outcomes, 8(1), 112-119.
3. Benavidez, G. A., Zahnd, W. E., Hung, P., & Eberth, J. M. Sociodemographic and Geographic Variations by Zip Code Tabulation Area.
4. Molnár, D., & Livingstone, B. (2000). Physical activity in relation to overweight and obesity in children and adolescents. European journal of pediatrics, 159, S45-S55.
5. Santoro, P. E., Paladini, A., Borrelli, I., Amantea, C., Rossi, M. F., Fortunato, C., ... & Moscato, U. (2024). Vaccine-preventable diseases: Immune response in a large population of healthcare students. Vaccine, 42(4), 930-936.
6. El-Mawla, Nesma Abd, et al. "Integrating IoT, Green AI, and Big Data Analytics in Climate Change Mitigation and Adaptation: Sustainable Smart Healthcare Systems as a Case Study." Handbook of Climate Change Mitigation and Adaptation.

Springer, New York, NY, 2025. 1-37.
7. Anderson, Gerard, and Jane Horvath. "The growing burden of chronic disease in America." Public health reports 119.3 (2004): 263-270.
8. Wilper, Andrew P., et al. "A national study of chronic disease prevalence and access to care in uninsured US adults." Annals of internal medicine 149.3 (2008): 170-176.
9. Bauer, Ursula E., et al. "Prevention of chronic disease in the 21st century: elimination of the leading preventable causes of premature death and disability in the USA." The Lancet 384.9937 (2014): 45-52.
10. Kahn, Jeremy M., et al. "The epidemiology of chronic critical illness in the United States." Critical care medicine 43.2 (2015): 282-287.