Analysis on Common Diseases in Adults and their Causing Factors

Abdul Azadh Abdul Saleem
x18203621
M.Sc in Data Analytics
National College of Ireland

Sudesh Kumar Narayanasamy Murali
x18199666
M.Sc in Data Analytics
National College of Ireland

Sai Chethan Singu
x18181937
M.Sc in Data Analytics
National College of Ireland

***Abstract*—The major diseases affecting human health is caused by improper nutrition, Abnormal physical activities and Bad Food habits that has a drastic effect on human health. Modernizing lifestyle has reduced the awareness on physical health, which has ultimately led to increased victims to various diseases, which can be controlled by following proper diet and physical exercises. United States of America being a developed country, has been chosen for this analysis since it has been reported with the tremendous increase in these common diseases. This report studies interpreting and visualizing various data about the country, which could help for this analysis. This project involves the extracting unstructured data from authorized government website, then pre-processing it into structured form and then interpreting it. Initially this analysis has been conducted on the entire country and further it is reduced to the parts with major impact, and its Nutrition, Physical activity and common Tobacco usage of the population has been studied.**

Keywords—Disease, Human health, Nutrition, Physical activity, Obesity, Tobacco usage, Survey, Visualization, Data Analysis.

I. INTRODUCTION

In developed nation like United States, people follow an exorbitant/sophisticated lifestyle, which involves food habits with heavy amount of Junk foods and does not possess proper nutrition. Also a lot of things are done in handy, which lacks frequent physical movement and exercises, which is quite dangerous and results in increase in overall health issues and associated diseases, and sometimes reduces the overall health quality & life expectancy of the people in the country. And majority of the people are used to this lifestyle from the very young age, a lot of diseases like Arthritis, Cancer, diabetes, and associated Heart disease seems to occur at the later point in life. So, these activities can be avoided by changing the lifecycle pattern along with proper food habits which can lead to a healthy lifestyle. This project is conducted to study about the common disease in United states and the common causing factors of these disease in a state, which has the highest cases of these disease. This project follows the process of extracting an unstructured data from official United states government repository and further loading the data to a non-relational database. Further this data is cleaned and converted into structured data and fed into a relational database, this data is then retrieved from the database, interpreted and visualized to obtain insights which can help for this analysis.

This research has been conducted with an intent to find the solution for the following questions.

1. What are the states which has the high Crude prevalence value of common disease in United States.
2. Which state in United States has the high Tobacco usage and analyze the category of people who are more prone to this habit.
3. Which category of people in West Virginia, does not have awareness about their daily nutrition and physical activity.

II. LITERATURE REVIEW

This research [1] is about the major cardiovascular disease in developed countries, which is caused by their improper diet Plan. This article discusses about diseases like High Cholesterol, Obesity, Hypertension and Atherosclerosis which is caused by food with high fat content and are hard to digest. This research has been conducted on people across various socioeconomic status and ethnicity. The researchers have concluded that it is a challenge to promote physically active lifestyle and healthy nutrition to the people across different ages. Until 20th century's last decade, the nutrition research on women was not conducted, instead the result obtained from the research on men is directly presented as a common result of the research, eliminating the gender difference. The author [4] here discusses that the metabolism occurring on both the gender are different as they have various impact of sex steroid hormones, gene repertories and environmental factor. This article reviews various underlying impact of food on disease from the perspective of sex-gender. In addition, it also researches about effect of Nutrition on various disease, this research [5] is about the cause of blood pressure due to over intake of nutrients. This study indicates the dietary approach to reduce hypertension, by following DASH diet which includes high fruits, vegetables, and low-fat dairy products. These authors [11] in this article express their study about the importance of exotic fruits and their influence on physiological human health.

This study [2] was about the Environmental tobacco smoke and its effects on the passive smokers. The researcher has conducted the case study over the spouses of smokers and found that they are 30% more prone to ischemic heart disease than the spouse of non-smokers. Smoking Cigarette has the major impact on causing Coronary heart disease, peripheral vascular disease, and stroke. The author [3] explains that the risk of coronary heart disease, caused by smoking cigarette increase with the duration and the amount of smoke. Cessation of smoking cigarette has reduced the risk of disease.

The relationship between arthritis and the physical activity in both genders has been explored in this journal [6], based on the National health survey conducted in 2002. It discusses

about the reduction in pain and improvement in human functions, for the population with regular physical activity. Result of this journal explains that 37% of adults with arthritis are physical inactive and the rest have moderate physical activity. Efforts must be made to promote physical activity for the patients with arthritis, as it is a recommended therapy for adults. Arthritis associated pain management measure and counselling for physical activity must be conducted by the healthcare providers.

This study [7] was about the prevalence of Chronic Obstructive Pulmonary Disease, association between the disease frequency and smoking rate in population. This involves in developing and validating a model based on the COPD prevalence and the smoking rate. The model is based gender and age specific rate of weakening or damage of lungs by the status of smoking. This model is initially tested on US smoking data collected from a national survey and later applied on various European countries to estimate the prevalence of this lung disease. In addition, the previous model to estimate the prevalence of COPD, this article [8] involves selection of small geographical area, to conduct a study of this disease using the Behavioral Risk Factor Surveillance System. Estimation on a small geographical area is a statistical technique which helps to provide a definitive estimate for the survey data.

This paper [9] describes about the Big Data in terms of Data visualization and presentation. This article also speaks about the problems faced in visualizing the Bigdata, like Visual noise, Perception of Large image, Loss of information, Requirement of High performance, Image change at higher rate, slicing of data and has prescribed approaches to rectify those issues. Another article [10] speaks about the choosing the correct visualization technique in order to infer the information conveyed from the analyzed data.

III. METHODOLOGY

A. *Dataset Description:*

The dataset used in this analysis are extracted from official government repository of Centers for Disease Control and Prevention (chronicdata.cdc.gov), which is accessible by the public. The dataset has the data collected and segregated from various Health surveys conducted on different health areas.

1. Dataset for common disease in United States.
   This dataset [12] has the data collected from a census survey conducted over 500 cities in America. This dataset has the crude prevalence value of various common disease, adult overs 18 years of age across the country. Crude prevalence value for various disease is obtained by division of total number of cases and the entire population, at a specific area in given time.
2. Dataset for Tobacco usage in United States.
   This dataset [13] is a collection of data from Behavioral Risk Factor Surveillance System survey data, with State Tobacco Activities Tracking and Evaluation (STATE) system. The state-based surveillance system BRFSS collects data about the risk factors of chronic diseases. This dataset is about the Tobacco Usage among the adults, in all the states of US. Data was provided with the information like States, Smoking Status of adult, Data Value of Tobacco usage, Gender, Race, Age group, Education of the adult who participated in the survey.
3. Dataset for Nutrition, Physical Activity and Obesity in West Virginia state.
   This dataset [14] also contains state-based information collected by BRFSS, to monitor the Health of adults over 18 years of age. This data has the information collected under three classes, Obesity/Weight status, Physical activity, and Food behavior. This data also has details like Gender, Age, Race, Education, and their Geo-location. The main purpose of this data collection is to divide the population based on these categories and to monitor and spread awareness about their nutrition, physical exercise, and weight status.

B. *Data Analysis:*

All the three datasets have been obtained in unstructured format, which is processed through multiple steps in order to visualize a proper structured data and get the insights from the analysis. The steps followed in this analysis for all the three datasets is explained below.

i. Data Extraction.

The three datasets used for this analysis purpose is extracted from government repository **chronicdata.cdc.gov**. The data is initially extracted in JSON format using Socrata Open Data API, also known as SODA API. Installation of sodapy package in python programming, helped in extracting the unstructured data from government repository using API Keys. The extracted data is dumped as .json file.

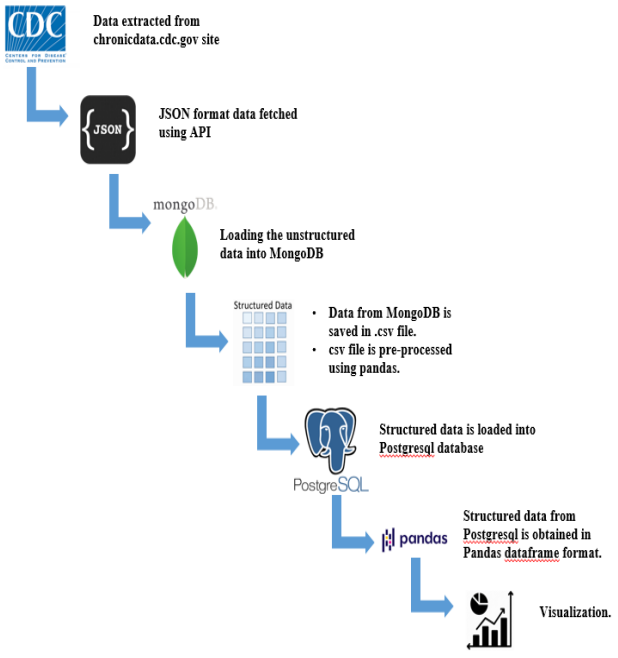

Fig.1 Data Analysis Flow Procedure.

ii. Storage of Unstructured data.

This unstructured data saved into the .json file is then parsed into dictionary data and loaded into the Non-Relational database, Mongo db. In Mongo dB, a new database has been created, along with a new table for loading the data, these steps were followed for every individual dataset. This data from this database is further obtained and written on array for every row, which helped to write the output data in .csv format on a separate csv file.

iii. Data Transformation.

Data transformation process here is done by converting the csv data into dataframe using pandas package. The dataframe is further processed by dropping the missing values, renaming the column name, and dropping the columns which doesn't contribute for this analysis. Now, the cleaned and structured dataset is stored as a new .csv file and saved separately.

iv. Storage of Structured data.

The structured data in csv format is loaded into the Relational database, PostgreSQL database here. Connection is established with the local host IP address and default port for PostgreSQL, with the help of psycopg2 package. Following the creation of a new database in and table for each dataset. The data is then loaded using the dbConnection.cursor(). The data from this table is extracted in the form of Pandas dataframe using pandas.io.sql package in python language.

v. Data Visualization.

The data is extracted from PostgreSQL database in Pandas dataframe format, which is then used for visualizing the data. The visualization has been done with plotly.express and go libraries. Initial analysis starts visualizing the common disease data of the entire country and further it is reduced to the state with most impact along the causing factors, which are being analyzed.

IV. Results.

i. Common diseases in United States.

The analysis has been carried out by considering ten common diseases in United States, and the state with most cases for each disease is plotted. The diseases analyzed here are Arthritis, High Blood Pressure, Cancer, Diabetes, Obesity, Chronic Obstructive Pulmonary Disease (COPD), High Cholesterol, Coronary Heart Disease, Mental Health and Physical Health. This value of each considered in the analysis is Crude Prevalence, which is the presence of the total cases with the particular disease, at a particular place at an instance of time, divided by the population at the same particular time and place.

On analyzing, it has been observed that West Virginia has the highest cases in most of the disease considered here. West Virginia state has highest number of cases in disease like Arthritis, High Blood Pressure, Cancer, COPD, High cholesterol and Coronary Heart Disease, while it ranks fourth in number of cases with Obesity, following Ohio state, and sixth in Mental Health following state Michigan. This analysis has been plotted and the diseases in which West Virginia has highest case among other states were plotted in below figures. The graphs here are represented in three different types, in order to get a clear picture of the fact analyzed here.

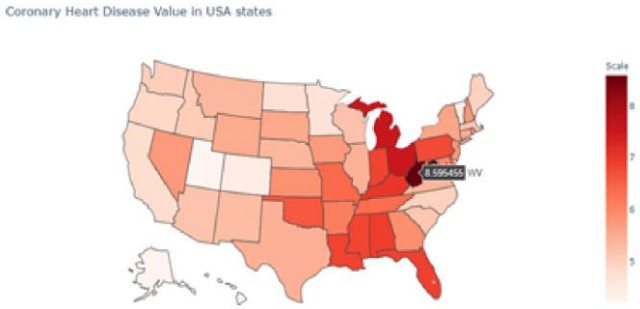

Fig.2 Average Crude Prevalence value of CHD in states of US.

Fig.2 is a choropleth map of the states in United States, with the value of their average Crude prevalence value. This provides an information that West Virginia has the highest value for Coronary Herat disease across all states in US. The cause of this disease is studied to be increase in cholesterol and fatty acids in the body, also by the usage of tobacco.

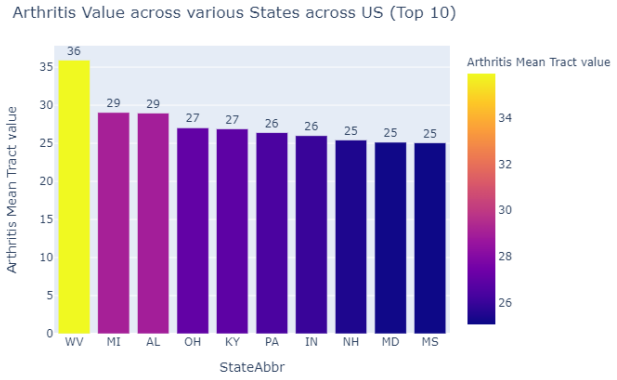

Fig.3 Average Crude Prevalence value of Arthritis in states of US.

Arthritis is a common disease that affects the human being on the process of aging. It is also studied that this disease is also caused by poor nutritious food, absence of proper physical activity in daily routine of people, mostly in the people of mid-age who are working professionals. Fig.3 represents the values of arthritis across top states in US, where it is found that West Virginia has the highest value.

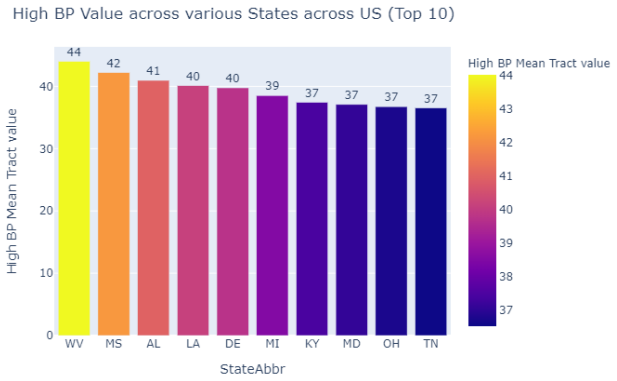

Fig.4 Average Crude Prevalence value of High Blood Pressure in states of US.

High Blood Pressure is a very common disease in this modern world, the victims of this disease are mostly the working professionals who don't have any time leisure, they do not follow any routine body exercise which could help the mind and body to resurrect. Here the graph in Fig.4 shows that West Virginia is the leading state, in this disease as well.

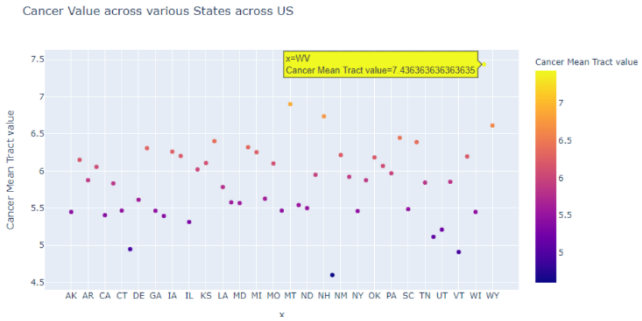

Fig.5 Average Crude Prevalence value of Cancer in states of US.

Cancer is studied to be caused by the high tobacco usage among the people, it is also proven that the people who are associated with active smokers for a considerable duration of time, also has the high chances of being affected by this disease. The Scatter plot in Fig.5 depicts that the state of West Virginia has the highest average number of cases among other states.

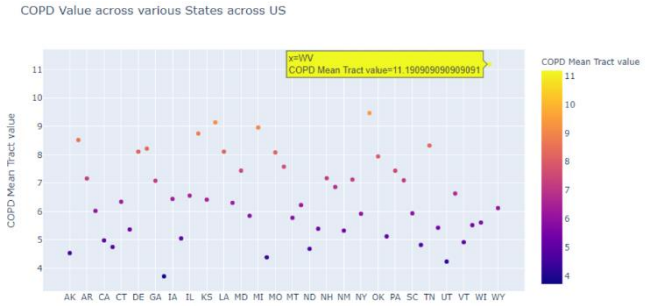

Fig.6 Average Crude Prevalence value of COPD in states of US.

Chronic Obstructive Pulmonary Disease is disease that affects Lungs, causing it hard for the lungs to circulate the air in human body. This disease is caused by tobacco usage, chemical particles in air, and pollution. West Virginia has the highest number of COPD cases visiting the hospital.

From the knowledge gained from researched papers, it has been inferred that, the common factor causing these diseases were usage of Tobacco, Improper food practice, Nutrition and Physical Activities. Further analysis is drilled down to the state of West Virginia which has high impact of these major diseases.

ii. Tobacco usage in states of US.

Tobacco usage in any form has been the cause of diseases like Coronary Heart Disease [2][3], Cancer, Chronic Obstructive Pulmonary Disease [7]. Usage of Tobacco has been analyzed using Choropleth map of Tobacco usage in states of US as shown in figure below.

In the Above Chart, average value of Tobacco usage is shown by the darkness of the pixels. The state with the highest Tobacco usage average value is shown by the darkest shade than the other states. The color grades of the average tobacco usage values are shown in the color scale. The state of West Virginia is inferred as the state with highest average of tobacco usage with the average value of 28.9, which has been shown by hovering the cursor over the state and Fig.7 and the state of California to be the least with 22.55 average value

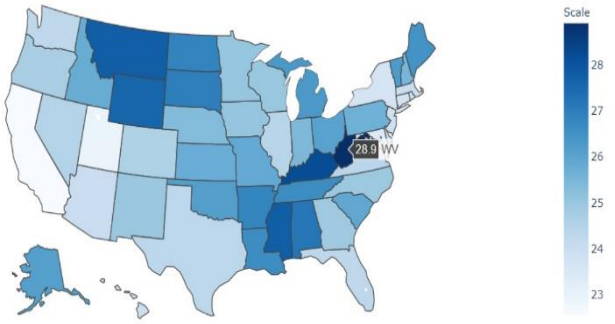

Fig.7 Usage of Tobacco in different states of US.

As observed from the reference [8], conducting the study in the small area could provide more accurate insights about the analysis. This study is further drilled down on analyzing this causing factor on the state of West Virginia. Gender-based analysis has been conducted to analyze proportion of both the gender who has the practice of using tobacco as shown in the figure below.

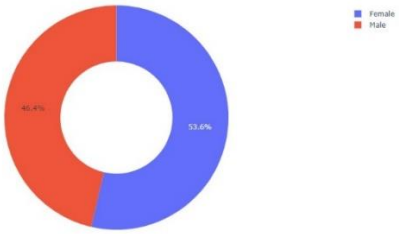

Fig.8 Pie chart showing the Gender proportion in usage of tobacco in West Virginia.

The graphs provide the insight that the Male are more prone to Tobacco usage than the Female gender. On the whole population taken in this analysis, percentage of Male using tobacco is 54%, while Female is 46%, which is not significantly different. Both genders are almost equal when it comes to the usage of Tobacco and its related drugs.

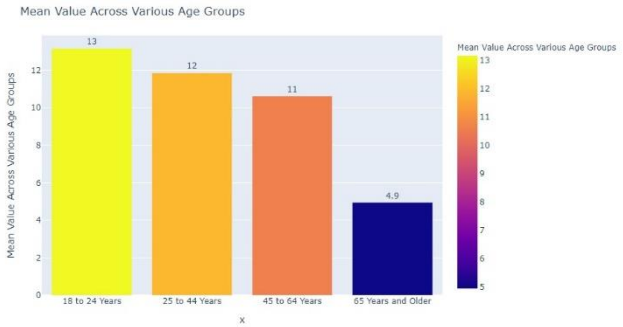

Fig.9 Age proportion of adult's usage of Tobacco in West Virginia.

This analysis has been conducted on adults more than the age 18, about their usage of tobacco. The age of adults was categorized into four, which could help to get the information about which age category has the most impact of Tobacco. This age analysis results that population in West Virginia, has shown that population between the age of 18 to 24 has the high proportion of usage with 32%, followed by the age category between 25 to 44 with 29% of tobacco usage. People over the age of 65 were less addicted to tobacco with the average usage value of 4.9 which contributes 12% of tobacco usage on the entire population.

iii. Nutrition and Physical Activities of adults in West Virginia.

This part of analysis is on Nutrition, Food practice and Physical activities of population in West Virginia state, since these factors are studied to be the cause of some common diseases analyzed in this project. Improper nutrition plays a vital role in causing diseases like High Blood pressure [5], Obesity, Diabetes, High Cholesterol and sometimes they may also result in causing Coronary heart disease [1]. Physical activities are highly recommended cure for the people with Arthritis. This disease causes stiffness in joints of bones, and reduced strength of bones.

Physical activities are a very common cure advised by many physicians, to cure many common occurring diseases in society. People with less physical activities are likely to be suffering the diseases like over-weight, high cholesterol, excess body fat which may even lead to obesity. Professionals who work and has less physical activity been found to be suffering from High Blood pressure. This makes the reason to analyze the physical activity of people in West Virginia as this state has the highest cases caused by the absence of physical activity. The below Fig.10 shows that most of the people in West Virginia has low physical activity, so we could get an insight that this is also the cause of the diseases.

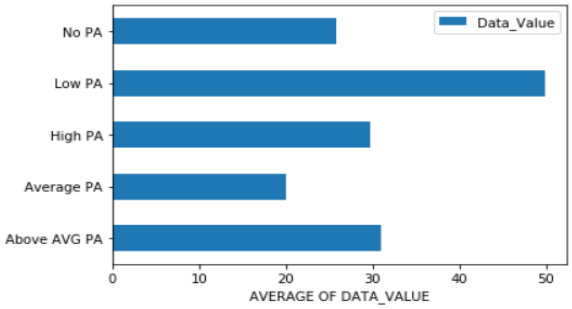

Fig.10 Physical activities of people in West Virginia.

Nutrition has also been an important factor related to these diseases as inferred. So, Adults are advised to follow proper diet for successful aging. Adding fruits and vegetable with the diet has proven to support healthy lifestyle for people [11]. Based on this Gender based analysis has been done on the population in West Virginia, following improper nutrition or diet as shown in the Fig.11. Improper food practice here refers to inadequate nutrients present in their meals.

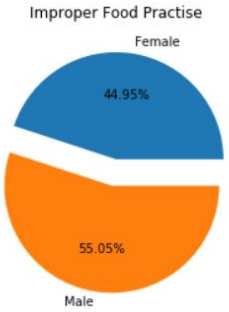

Fig.11 Gender-based analysis of improper nutrition in West Virginia.

It has been observed that on analyzing over gender proportion following improper diet, it has been studied that majority of gender contributing for this factor is Male, with 55%, slightly higher than Female who are 45%. Likewise the Gender based analysis has also been done over on the presence of Obesity among the population in West Virginia. As shown in Fig.12, seems similar to Fig.11, which infers that the people with no proper diet meal is more likely to be affected by Obesity or Over-weight.

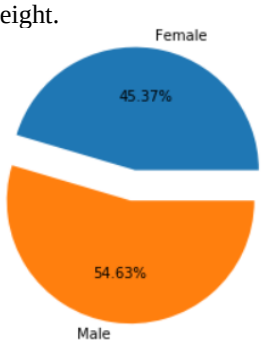

Fig.12 Gender-based analysis of Obesity in West Virginia.

And this Nutrition factor is also analyzed on the age proportion for the people in West Virginia, which is shown in the Fig.13.

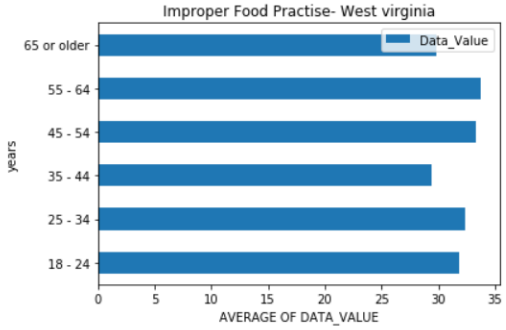

Fig. 13 Age-based analysis of Improper food practice in West Virginia.

This age-based analysis on bad nutrition has been done to identify the category of people who are less aware of proper nutrition in order to live a health life. This age-related information obtained from this study can help the spread of awareness for the age group of people, who are less aware of health nutrition, and did not follow proper diet on their daily life. Since, people of these age groups are more likely to be victims of many common disease which occurs on the process of aging. As analyzed above the age-based population affected with Obesity is also done as shown in the Fig.14. This graph

infers that the people between the age group of 25 to 65 are more prone to be affected with Obesity or over-weight.

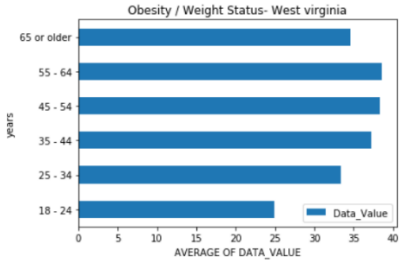

Fig.14 Age-based analysis on Obesity in West Virginia.

V. Conclusion

From the study conducted on three different datasets in this project, it is inferred that among the all the states of United States, West Virginia has the highest cases of most of the diseases. Further the study has been conducted on the Tobacco usage in the states of US, as it has been a main causing factor for most of the disease considered in the project. This analysis resulted that West Virginia state has the highest average usage of Tobacco among other states which relates it to the state, as it tops in the disease caused by this habit. Further, analysis has been done on age and gender proportion of the people in West Virginia, about their tobacco usage, which results that Male gender being the most victims of this habit and people of Mid-age (i.e. 25-44) mostly has this habit. Then, another study has also been done on Nutrition and Physical activities, which resulted that the category of people who do not follow proper diet in their daily meal, are more likely to be suffering from Over-weight or Obesity. The future scope of this project would be conducting this study on the entire states of United States, and about the steps taken to spread awareness to the people of the country about the importance of nutrition, physical activity and the impacts of tobacco usage to the human health.

VI. Reference.

[1] Godfrey S. Getz, Catherine A. Reardon, "*Nutrition and Cardiovascular Disease*", Arteriosclerosis, Thrombosis, and Vascular Biology. Vol 27 (2499-2506), Oct 2007.

[2] Malcolm R Law, Nicholas J Wald, "*Environmental tobacco smoke and ischemic heart disease*", Progress in Cardiovascular diseases. Vol.46 (31-38), 2003.

[3] David M Burns, "*Epidemiology of smoking-induced cardiovascular disease*", Progress in Cardiovascular Diseases. Vol.46 (11-29), 2003.

[4] Maria Marino, Roberta Masella, Pamela Bulzomi, ilaria Campesi, Walter Malorni, Flavia Franconi, "*Nutrition and human health from a sex-gender prespective*", Molecular Aspects of Medicine. Vol.32 (1-70), Feb 2011.

[5] Vincenzo Savica, Guido Bellinghieri, Joel D. Kopple, "*The Effect of Nutrition on Blood Pressure",* Annual Review of Nutrition. Vol.30(365-401), Aug 2010.

[6] Margaret Shih MD PhD, Jennifer M. Hootman PhD, Judy Kruger PhD, Charles G. Helmick MD, "*Physical Activity in Men and Women with Arthritis: National Health Interview Survey, 2002*", American Journal of Preventive Medicine. Vol.30 (385-393), May2006.

[7] Paul Stang PhD, Eva Lydick PhD, Cheryl Silberman PhD, Angela Kempel, Elizabeth T.Keating, "*The Prevalence of COPD: Using Smoking Rates to Estimate Disease Frequency in the General Population*", CHEST. Vol.117(354S-359S), May2000.

[8] Xingyou Zhang, James B.Holt, Hua Lu, Anne G. Wheaton, Earl S. Ford, Kurt J. Greenlund, Janet B, Croft, "*Multilevel Regression and Poststratification for Small-Area Estimation of Population Health Outcomes: A Case Study of Chronic Obstructive Pulmonary Disease Prevalence Using the Behavioral Risk Factor Surveillance System*", American Journal of Epidemology. Vol.179(1025-1033), Apr2014.

[9] Evgeniy Yur'evich Gorodov, Vasil'evich Gubarev, "*Analytical Review of Data Visualization Methods in Application to Big Data*", J.Electrical and Computer, 2013.

[10] Lidong Wang, Guanghui Wang, Cheryl Ann Alexander, "*Big Data and Visualization:Methods, Challenges and Technology Progress*," Digital Technologies, 2015, Vol. 1, No. 1, 33-38.

[11] Valey M. Dembitsky, Sumitra Poovarodom, Hanna Leontowicz, Maria, Leontowicz, Suchada Vearasilp, Simon Traktenberg, Shela Gorinstein, "*The multiple nutrition properties of some exotic fruits: Biological activity and active metabolites*", Food Research International. Vol.44(1671-1701), Aug2011.

[12] "500 Cities: Census Tract-level Data", chronicdata.cdc.gov, 06 10 2019. [Online]. Available: https://chronicdata.cdc.gov/500-Cities/500-Cities-Census-Tract-level-Data-GIS-Friendly-Fo/k86t-wghb

[13] "Behavioral Risk Factor Data: Tobacco Use", chronicdata.cdc.gov, 2019. [Online] Available: https://chronicdata.cdc.gov/Survey-Data/Behavioral-Risk-Factor-Data-Tobacco-Use-2011-to-pr/wsas-xwh5/data.

[14] "Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System", chronicdata.cdc.gov, 10 10 2019. [Online] Available: https://chronicdata.cdc.gov/Nutrition-Physical-Activity-and-Obesity/Nutrition-Physical-Activity-and-Obesity-Behavioral/hn4x-zwk7.