NUTRITION PHYSICAL ACTIVITY AND OBESITY BEHAVIORAL RISK FACTOR SURVELLIANCE SYSTEM

Group -03

Akhil Venkatesh AmbooriVarsha Alle

Varshitha Velkanti

UNIVERSITY OF NORTH TEXASADTA 5940

Dr. Denise Philpot

Spring 2024

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER – 1: INTRODUCTION

To prevent the chronic disease's spread at every stage of life is imperative in any country.

To preserve lives, there are a few divisions that focus on the analysis of the prior health difficulties. DNPAO, or the component of Nutrition, Physical Activity, and Obesity, is a component of the US Centers for Disease Control and Prevention (CDC). In response to the expanding public health concerns of obesity, physical inactivity, and poor nutrition—all of which have a significant influence on people's health and well-being, both individually and collectively—the division was established.

The objectives of DNPAO are to reduce obesity, increase physical activity, improve nutrition, and achieve health equity by reducing disparities. In terms of health status or access to care, there are differences between different geographic, racial, ethnic, and socioeconomic groups.

Reason for the selection of dataset

More than 75% of adults do not engage in as much physical activity as is advised to help prevent and reduce disease. Less than 10% of adults and adolescents in the US consume adequate fruits and vegetables. Adult obesity rates in the US are 42%. It is crucial to recognize these issues and inform everyone about them in order to save lives by raising awareness that these issues can be

reduced by up to 90% just by maintaining a healthy diet and engaging in physical exercise. We also want to know about the main issues and current trends in this field.

In order to work on the capstone project, we took into consideration a data collection called Nutrition, Physical Activity, and Obesity, which is used for the DNPAO Data, Trends, and Maps database. It provides information on physical activity, diet, obesity, and fruit and vegetable consumption.

Dataset Description

There are 22 variables and 93250 observations in the dataset. Three types of variables exist: 12 nominal, 7 discrete, and 3 string. Below are the data kinds and descriptions of the variable.

Variable Name	Datatype	Description
1)Year Start	Discrete	Contains the year in which the data collection started
2)Year End	Discrete	Contains the year in which the data collection ended
3)Location Abbreviation	Nominal	Contains the data of the location name
4)Location Description	String	Contains the location where the data is collected
5)Class	Nominal	Contains the category of the class ex: Physical Activity, Fruits and Vegetables, Obesity.
6)Topic	Nominal	Contains the topic of the class (Behavior)
7)Question	String	Contains the question which tells about the category of the data
8)Data Value	Discrete	Contains the average value of the category among the other data
9) Low Confidence limit	Discrete	Contains the lower confidence limit for the data value.
10) High Confidence Limit	Discrete	Contains the Upper confidence limit for the data value.
11)Sample Size	Discrete	Contains the size of the sample among the other data
12)Age(years)	Discrete	Contains the age of the population
13)Education	Nominal	Contains the educational level of the population
14)Income	Nominal	Contains the Income level of the population
15)Race and ethnicity	Nominal	Contains the Race and Ethnicity of the population
16)Geo Location	String	Contains the coordinates of the location where data is collected
17)Class Id	Nominal	Contains the Id for the category of the class

18)Topic Id	Nominal	Contains the Id for the topic of the class
19)Question Id	Nominal	Contains the Id question which tells about the category of the data
20)Location Id	Nominal	Contains the id of the location based on Alphabetic order

21)Stratification Category 1	Nominal	Contains the grouping of data based on specific criteria.
22)Stratification 1	Nominal	Contains the grouping of data based on specific criteria in terms of Income.

EDA AND RESEARCH QUESTIONS AND APPROACH

Discussion, concerns, potential issues (tools, techniques, computing power).

- Selecting the right statistical technique and selecting the right approach for the analysis could be the possible challenges that we might face during working on the project.
- Choosing the right technology to work on the analysis may take time.

- Conducting weekly checks to make sure that the project is going in the right track, in order to eliminate any potential risks.
- We are not worried about high computing power as we have chosen a dataset that requires minimum data storage capacity and processing speed.
- Our primary concern is that no two members of the team use the same data analysis tool. Every member of our team is proficient with a distinct data analysis tool. As a result, we must set aside time to comprehend the different tools that we are all familiar with and choose the best tool for the project.

CHAPTER 2: LITERATURE AND SCHOLARLY REVIEW

Introduction: To prevent chronic disease's spread at every stage of life is imperative in any country. To preserve lives, there are a few divisions that focus on the analysis of the prior health difficulties. Within the US Centers for Disease Control and Prevention (CDC) involves the division known as DNPAO, or Nutrition, Physical Activity, and Obesity. The division was created as a reaction to the growing number of public health issues, including obesity, physical inactivity, and poor nutrition—all of which have a substantial effect on people's health and well-being, both personally and socially.

Through the reduction of disparities, DNPAO seeks to improve nutrition, decrease obesity, boost physical activity, and promote overall health. There are variations in health care, health condition, and treatment accessibility among different geographic, racial, ethnic, and socioeconomic groups.

Cause: All the data that we have taken is from various places from which we understood there can be numerous reasons for the cause of the health issue.

From our findings, the reasons can be as follows:

1) Financial status

Its known that healthy food is comparatively costly when compared to normal food. Poor can't

afford the healthy food and the fee for exercise club which leads to obese issues.

2) Educational background

People from no food educational background eventually end up having no knowledge about healthy food and does not give importance to the diet.

3) Cultural practices

Certain cultures may have diets with unhealthy food activities and traditions that do encourage exercises making it harder to stay healthy. One of such activity is too much fasting.

4) Access to health care.

Few areas do not have access to good health care with no healthy food markets around, pollution, no gym. Some poor places do not even have hospitals and doctors.

5) Environment they live in.

Recently developed areas have lots of fast-food restaurants making people lazy to cook. Most of the food new food culture has unhealthy food. These places sometimes do not even have fresh vegetables and fruits but good restaurants leading to unhealthy food, which can lead to obesity.

Impact:

Due to the aforementioned several reasons there is significant effect on the people's health and well-being. Higher rates of obesity and least physical activity can lead to chronic diseases such as diabetes and heart problems. These physical health issues also stress the peoples mental health leading to other mental health issues which in turn effect their already existing health

issues. When certain groups have health issues, this worsens the existing inequalities. For instance, when lower income of people from minority backgrounds who face more health issues related to obesity have more health problems, disparities arise among rich and poor. Health problems such as obesity, lack of physical activity can also affect the productivity of the

employees and weaken the economy, as people struggle to be active and concentrate on the work with health issues. To address these challenges, it's important to comprehend the causes, understand the possible strategies to overcome the issues by creating healthier environments, to make sure everyone has a chance to be healthy and well.

Existing Solutions: Studies suggest that there are few ways to get people stimulated and improve their activeness. They say that people need support, and they are interested in using technology to help them be more active. If provided easy access and less cost for the public transportation, people start using those which leading to physical activity. Workspaces generally make people sit in a single place for long time. Such rules should be changed encouraging employees to move and adding extra activities in workspace. Encouraging people to use parks and gardens by planning such places in nearer circle to the living places or crowded areas rather than planning those in far larger distances can make people feel better mentally and physically. Making sure everyone can use these spaces and making neighborhoods feel friendly is important. Lastly, using the correct methods to study health and setting clear rules for measuring malnutrition in kids can help us to understand how to make things better. Understanding why people with obesity don't exercise much can also help us make better plans to help them.

Possible Solutions: We can get a lot of analysis considering different variables in groups at a time. In our study, we are looking at different things like age, income, and where people live to see how much they exercise and if they are overweight. By using EDA, we can obtain the visualizations to clearly display the data. We can determine whether the person is overweight and active or not. We can also find out whether the physical activity and weight change with time by examining data. To get precise understanding of the variations in the exercise routines we can divide the subgroups within the major groups.

To determine how often people workout based on the factors such as age and income, we will employ machine learning algorithms. We feel direct conversations with people will be more helpful to understand more about how whether people exercise or not. We will initiate programs to encourage people to work out more, especially those who don't already do many exercises. We will also consider implementing policies and changes for participation of all the members from communities. We are going to create programs in such a way that those should educate individuals understand the importance of physical activity to overcome the obstacles.

Literature review:

It's known that obesity is a big problem in the United States and researchers are working on the data to analyze and understand the reasons for the outcomes. They are doing in-depth research such as counting the number of people trying to lose weight and the methods, they follow to achieve it. A large survey was conducted over the phone in 2000, the data covers lots of people across the various states. This research is published by the author Puerto Rico (Bish et al., 2005). They were surprised by knowing that almost half of the women and one third of the men are actually trying hard to lose weight. Women are starting exercises at a lower weight when compared to men. Furthermore, it is also found that individuals who went to doctor and got advice more likely trying to lose weight. People who had higher education were also more likely trying to lose weight. From the above results its clear that most of the people are trying to lose weight yet there are no positive results globally because they are not following the recommended methods such as consuming healthy food and exercising regularly which are most effective practices.

In 2018, the Physical Activity Guidelines Advisory Committee Systematics Review inspected the

strategies for motivating people to be active and not sitting in one place using the data from research published between 2011 and 2016 (King et al., 2019). Researchers identified that variety of approaches to encourage healthy habits help at various levels of effectiveness.

These include designing communities in such a way that they should be easily accessible to facilities like parks and public transportation, offering each other support, or leveraging technology to help people change their activities and behavior. Example for one such thing is fitness watches are motivating people to exercise.

We can also find ways to reduce sitting, especially for young and IT working individuals We believe that future studies should focus on making sure all these methods are affordable and available to everyone and from all different backgrounds and should be able to evaluate their effectiveness across a wide range of circumstances. All the aforementioned explanations emphasize how significant it is to design spaces that encourage people to be more active and less lazy resulting healthy communities. We acknowledge the importance of regular exercises in maintaining good health preventing illness as stated by the author (Lacombe et al., 2019). Its also important to comprehend how physical workouts interacts with other lifestyles behaviors and personal issues. Studies followed the habits of active and inactive people for last 8 years were examined between 2010 and 2017.

Researchers used the Newcastle-Ottawa Scale to measure the quality of the data they collected for every study. They emphasized on the connection between physical activity and other habits involving eating , drinking, smoking, and being sedentary , and these are responsible for death rates, cancer rates and heart disease.

The outcomes described that individuals who were engaged in physical activity and other healthy habits were less likely to develop heart problems or any disease death and experience overall mortality compared to people who are less active with weak health habits.

These results were consistent across different age groups, genders, and study durations. Although there were fewer studies on cancer outcomes, they also demonstrated similar patterns. A study explains the positive effects of urban green spaces on interpersonal relationships, happiness, and health which is by the author (Jennings & Bamkole, 2019). It underscores the value of social togetherness for human well-being and the contribution of urban green spaces, such as gardens and parks, to the development of a feeling of community. The places that encourage social connection and improves peoples both mental and physical activities take a step to maintain good public health. For these things to happen, they start walking and community meetings activities.

There are studies which highlights the need for investigation to completely understand the effect the green areas in helping individuals' social cohesiveness and health. As part of investigation there is need to focus on green spaces. It also suggests that future study should consider the elements like neighborhood cohesion. Ultimately, the key to improving people well-being in cities is the integration of psychosocial variables into planning and health promotion programs. Statistical analysis is important in research, but many researchers make mistakes because they are not able to completely comprehend it (Serdar et al., 2021). Misconceptions, no enough examples, and the exaggeration of outcomes are few general mistakes made in research.

Always check the number of examples that are included in the research to prevent overuse of them which will lead to the incorrect outcomes. When size and practical significance are considered, these both reveals the strength of the findings. While choosing a sample size for

research, ethical issues are crucial. The significance of the results and the possibility of the errors, software will give the appropriate sample size. Only depending on the statistical significance is insufficient. Adhering to the norms and protocols, various technologies support statistical analysis for research studies. We can completely trust the findings from the analysis if they are thoroughly planned with all the relevant elements. We know that malnutrition is very bad among kids, so to measure the level of malnutrition, new rules have been created by a group of experts from WHO and UNICEF according to (De Onís et al., 2018). The team looked at data from various countries and introduced five levels of malnutrition: 'very low', 'low', 'medium', 'high', and 'very high'. These rules help countries in determining where to concentrate their efforts and how bad their malnutrition issues are. The guidelines were fairly taken by the specialists using the WHO's definition of normal for kids. These guidelines can now be utilized by everyone to increase understanding and communication about malnutrition. To make sure that the correct steps are performed and to prevent ambiguity, it is important that every individual uses the same terminology.

It is suggested that all country's adopt these new rules to assist enhance the health of children. The study (Baillot et al., 2021) inspected the reasons why obese individuals do not exercise despite the fact that even they know which it is beneficial for them if they do exercise, these reasons were examined by the author (Baillot et al., 2021). The study discovered that many individuals who are suffering from obesity are actually interested in work-out, walking, feel more energized and receive social support.

They also face challenges such as not unmotivated, experiencing pain while not having enough time. Surprisingly, many of them prefer walking as their favorite way to be fit. Understanding these reasons can help doctors create better ways, plans to assist people with obesity and keep motivating.

Due to COVID-19 pandemic, many people have to start work from home and its still continuing (Oakman et al., 2020). Researchers wanted to know how this affects workers' mental and physical health. They looked at studies from 2007 to 2020 to find out. They found 23 papers thattalked about things like pain, feeling healthy, stress, and happiness. The results showed that having support from the company, friends, and family, as well as managing work and home life well, can make working from home better for health. They also saw that women may not benefit as much from working from home as men do. The researchers say it's important for companies tohave clear rules and support systems in place to help workers stay healthy while working from home. The report (BRFSS Prevalence & Trends Data: Home | DPH | CDC, n.d.) is a dynamic webpage that has categories and classes that visually shows the behavioral factors that impacts lifestyle and causes disease among people residing in united states. Several factors and geographical locations will affect these trends. Using this visual, we can analyze our findings of the trends and patterns of states by comparing them. Big Health Organization CDC provides the reports about health and behavior. We can get eating habits, physical activity, and obesity information from these papers. Everyone has access to these reports in CDC website, if searched for “BRFSS reports”. In order to locate the most recent news within the desired time range, we may further narrow our search.

Heart issues are a big deal worldwide, especially in Chine where lots of people get sick each year. There are two main reasons why people get these heart problems: one is because of stuff like high blood pressure and diabetes, and the other is because of things like smoking and not moving around enough. To understand how common these behaviors were, a data was taken into account. The considered data was from over 17,000 older Chinese adults. This was studied was by the author (Ding et al., 2020). The research found that few people in some areas of China are more likely smoking, drinking a lot, being overweight or not moving around much. Interestingly, they noticed that people who already had problems like high blood pressure or diabetes tended to not smoke or drink as much, but they were more likely to not move around much and be overweight. Understanding all this can help figure out how to help people live healthier lives and avoid heart problems in China.

Conclusion

For general health and illness prevention, regular physical activity—even walking—is essential. Although obesity is a major public health concern, a sizable section of the population actively tries to lose weight. Having access to natural areas and socializing opportunities enhance wellbeing and promote physical activity People who receive social support from friends, family, and medical professionals are better able to overcome obstacles and stay motivated. Working from home can have a good effect on one's physical and mental well-being, but it's important to have clear boundaries and support networks, especially for women.

Chapter 3: Methodology

Software

Making the right software selection is crucial when working with huge data sets. The ideal program must be able to quickly and accurately produce results while managing large amounts of data. In order to achieve this, we used Python:

Python is a flexible programming language that works incredibly well with handling and modifying data. Because of its user-friendly syntax, wide library support, and community-driven development, Python is our preferred option. To improve our data manipulation skills even more, we use the Numpy and Pandas frameworks in addition to Python. With the help of these frameworks, we can extract insights from large, complex data sets and streamline our workflow with a variety of effective tools like array operations and data analysis algorithms.

We deployed Jupyter Notebook as an IDE(Integrated Development Environment). It is an open-source web tool which gives us the ability to create and execute code and serves as a platform for data visualization, analytical documentation, and team collaboration. It is an efficient tool for scientific computing and data analysis because of its adaptability, interaction, and simplicity.

Data pre-processing

Data collection

The source of our dataset is data.gov, it was last updated on December 8th, 2023. In search for the datasets for the capstone project we have gone through various websites and lastly found nutrition dataset on data.gov. The dataset comes from the Behavioral Risk Factor Surveillance

System (BRFSS), which is spread across multiple states and geographical areas in the United States. With an emphasis on behaviors related to physical activity, food, and obesity, it aims to

collect a significant amount of data concerning health-related behaviors, chronic illnesses, and adult Americans use of preventative services.

With regard to research projects and public health campaigns that aim to address and comprehend issues related to obesity, physical activity, and nutrition in the US, this dataset is particularly significant because it offers statistics about the nation's health state.

This dataset holds particular significance for research initiatives and public health campaigns that seek to understand and solve obesity, physical activity, and nutrition-related concerns in the US, as it provides data on the country's health. There are 22 variables and 93250 observations in the dataset. Three types of variables exist: 12 nominal, 7 discrete, and 3 string.

Data Wrangling

We had to perform data wrangling process for our project improve quality and ensure that it aligns with our research objectives, making it more efficient so that we can obtain focused analysis. Data Value Unit and Data Value Type were eliminated because it contained information about measurement units and natures such as percentages and count. Since our analysis focuses on the data values themselves rather than their units, these variables were considered unnecessary. Our goal is to analyze the primary data values; hence, Data Value Alt was removed. Data Value Footnote Symbol and Data Value Footnote indicates footnotes associated with the data values. Our analysis does not delve into footnotes. So, these two variables are removed as they irrelevant for the analysis.Variable Total was deleted as its unclear what it represented.

Coming to the missing values, we replaced missing values in the columns low confidence limit, high confidence limit and geo-location with the calculated median of the specific variable because our dataset contains outliers, it will be more suitable to use the median in place of missing numbers since it won't be impacted by extreme values.

Chapter 4: Exploratory Data Analysis

In order to obtain insights, spot patterns, and investigate correlations between variables, exploratory data analysis, or EDA, is a crucial stage in the data analysis process.

The initial stage of data analysis involves comprehending the columns and their categories, as well as determining whether any values are missing. A high-level summary of the data is shown in the following graphic.

Rangeindex: 93249 entries,. 0 to 93248Data columns (

total33 columns}:

	Column	Non-NullCount	Dtype
0	YearStart	93249 non-null	int64
1	YearEnd	93249 non-null	int64
2	LoeationAbbr	93249 non-null	object
3	LocationDesc	93249 non-null	object
4	Datasource	93249 non-null	object
5	Class	93249 non-null	object
6	Top.:ic	93249 non-null	object
7	Question	93249 non-null	object
8	Data Value Unit		
9	Data_Value_Type		
10	Data Value		
11	Data Value Alt		
12	Data_Value_Footnote_Symbol		
13	Data Value Footnote		
14	Lo Confidence Limit		
15	High_Confidence_Lirnit		
16	Sample_Size		
17	Total		
18	Age(years)		
19	Ecl ucat ion	0 rnon-null 93249 non-null84014 non-null 840l4 non-null9235 non-null 9235 non-null84014 non-null84014 non-null84014	
20	Gen der	non-null 3330 non-null 19980 non-null 13320 non-null6660 non-null 23310 non-null	
21	Inc om e		

float64 object	o		t object
float64 float64	b		object
object object	j		object
float64 float64	e		
float64 object	c		
22 Race/Ethnicity	26640	**non-null**	object
23 GeoLocation	915:B	**non-null**	object
24 ClassID	93249	**non-null**	object
25 Top.i.cID	93249	**non-null**	object
26 Questior11ID	93249	**non-null**	object
27 DataValueTypeID	93249	**non-null**	object
28 LocationID	93249	**non-null**	int64
29 StratificationCategoryl	93240	non-null	object
30 Stratification1	93240	non-null	object
31 StratificationCategoryid1	93240	non-null	object
32 StratificationID1	93240	**non-null**	object

memory usage: 23.5+ MB
dtypes: float64(6) int64(3} object(:24)

Figure 1: Overview of the data

Data Visualization

Understanding the general trends and distributions of health outcomes across different states

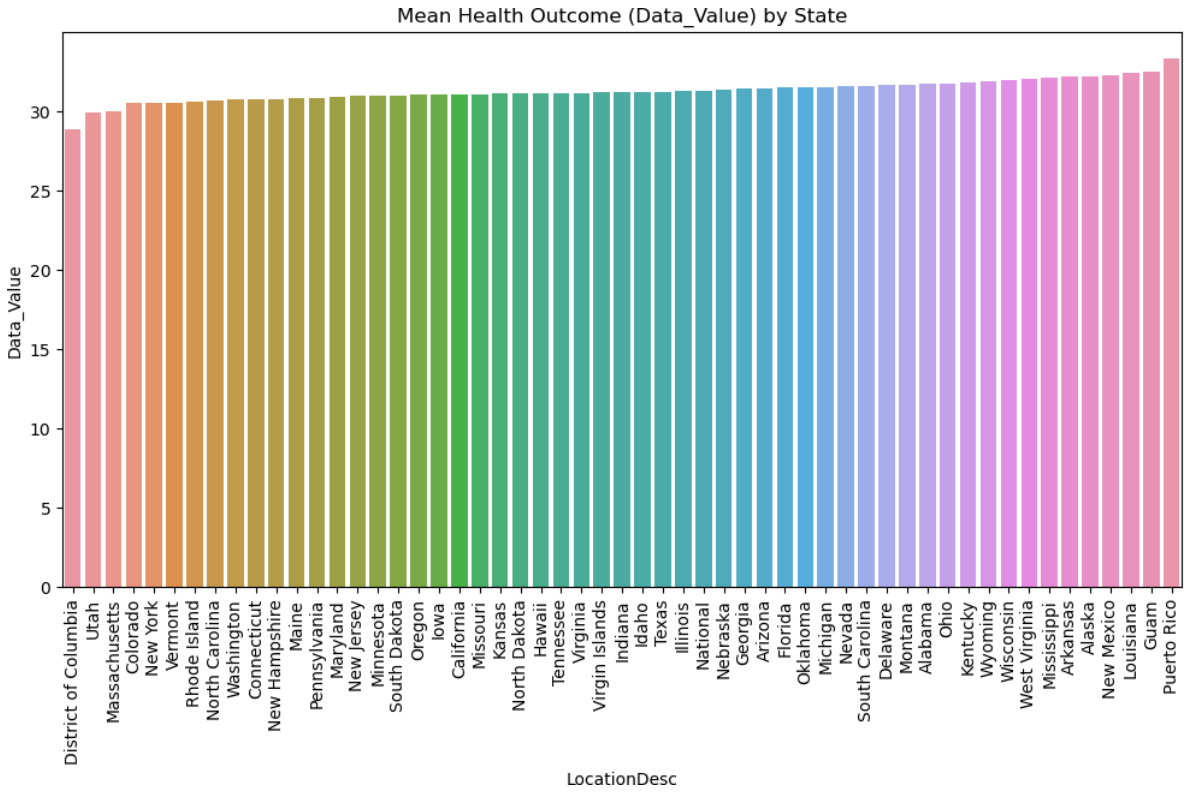


Figure2: Analysis between Location Desc and Data Value on average parameter

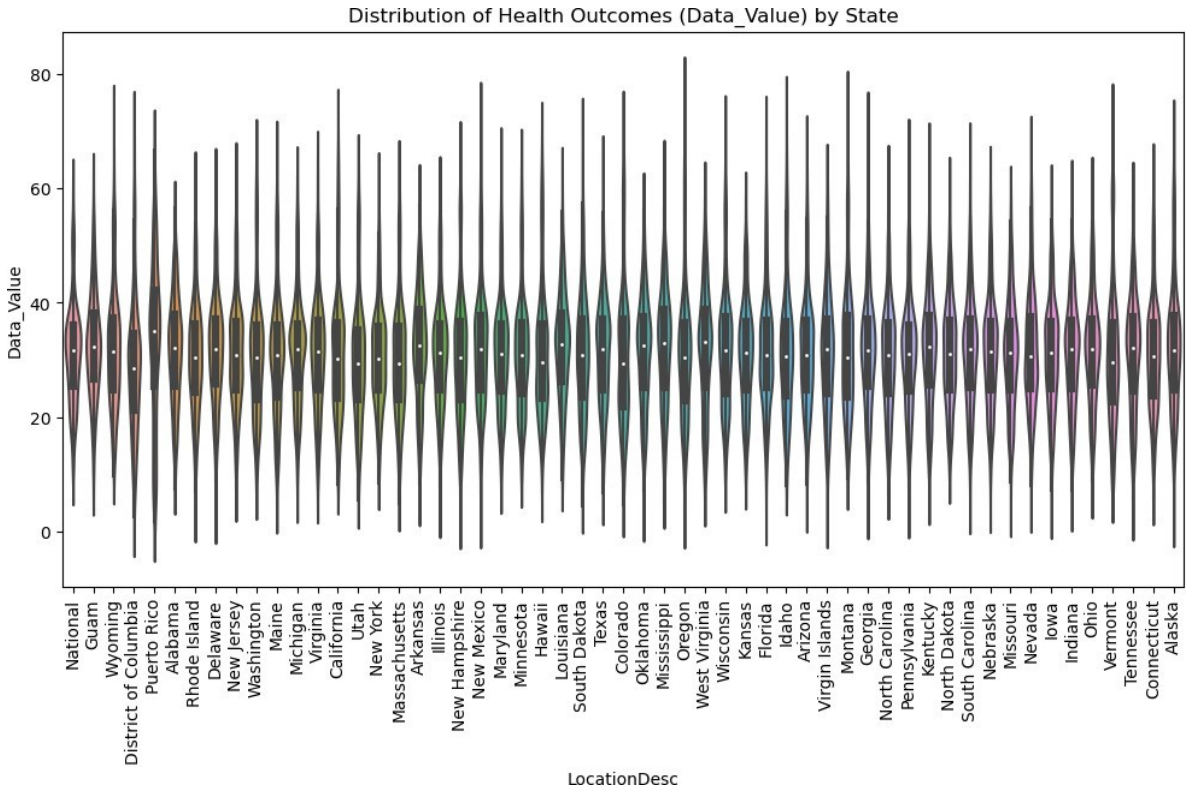


Figure3: Distribution Analysis between location description and data value

Finding the average and median Data_Value for every state can reveal information about the central and average health outcomes in each state.The broad trends, patterns, and distributions of health outcomes can be found by visualising the distribution of Data_Value across the states using histograms or boxplots. This can indicate potential clusters, outliers, and inequalities among states.

The study showed broad distributions, trends, and patterns of health outcomes (Data_Value) in several states (LocationDesc), suggesting possible regional variances, disparities, and state-to-state changes in health outcomes. The central tendency, variability, spread, and dispersion of health outcomes across states were revealed by calculating the mean and median Data_Value, observing

the standard deviation, and visualizing the distribution of Data_Value using histograms or boxplots. These methods also highlighted potential clusters, outliers, disparities, challenges, and

influencing factors affecting health outcomes among states.

Identifying whether there have been improvements or deteriorations in specific behavioral areas and is there any correlation between the sample size and the reported health outcomes

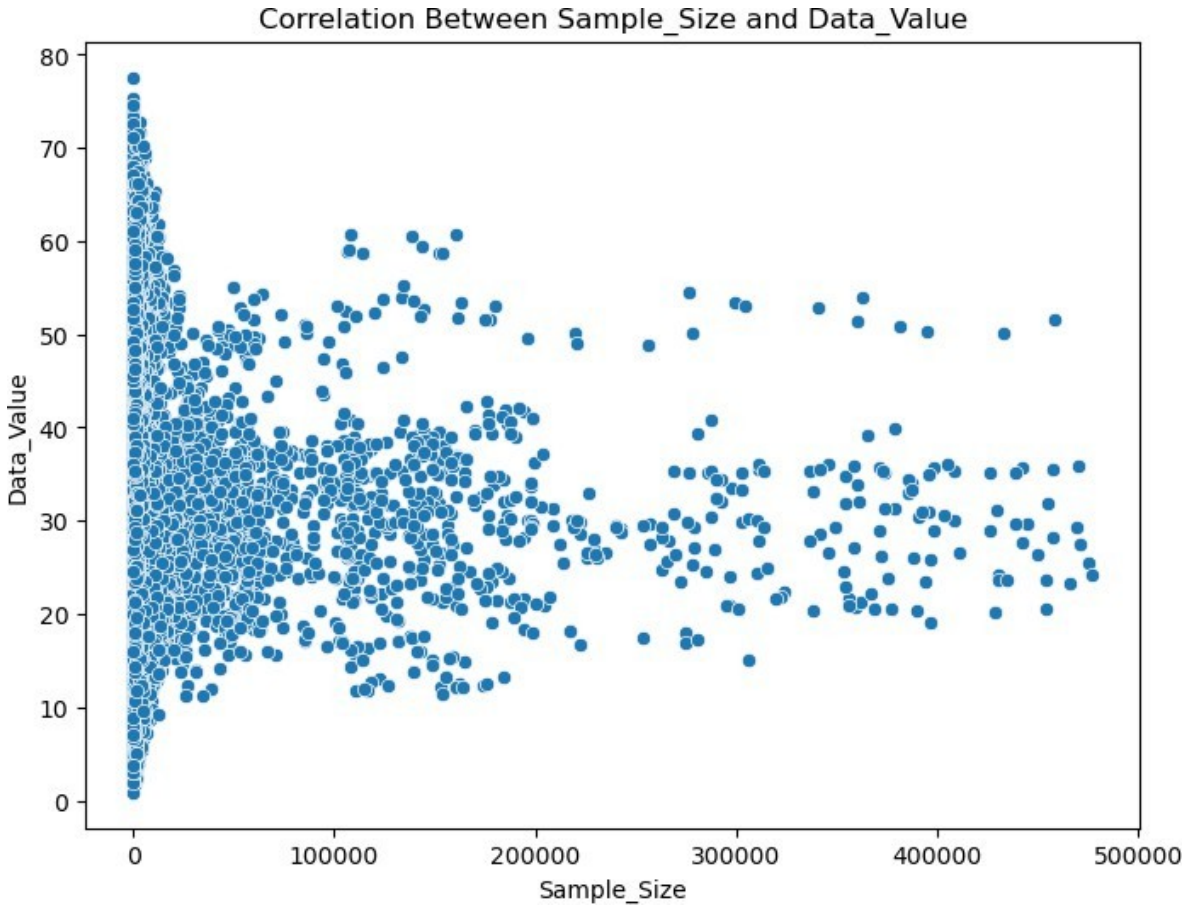


Figure4: Analysis between sample size and data Value

Correlation between Sample_Size and Data_Value: -0.004523193707697787

Sample Size (Sample_Size) and Reported Health Outcomes (Data_Value) Correlation: The purpose of the analysis is to look at the relationship between reported health outcomes (Data_Value) and sample size (Sample_Size). The relationship between Sample_Size and

Data_Value can be seen using a scatter plot, and the strength and direction of the linear relationship between the two variables can be measured using the Pearson correlation coefficient.

Analysis of Data_Value trends over time, broken down by various health themes, showed either advancements or regressions in particular domains. Plotting data value patterns for many health subjects and evaluating trends over time can assist in identifying key areas for focused interventions and strategies to address identified health issues, advance health equity, and enhance health outcomes. It can also help identify obstacles and areas with significant changes.

Using scatter plots to examines if greater sample sizes correlate with more standard or specific health results measures.

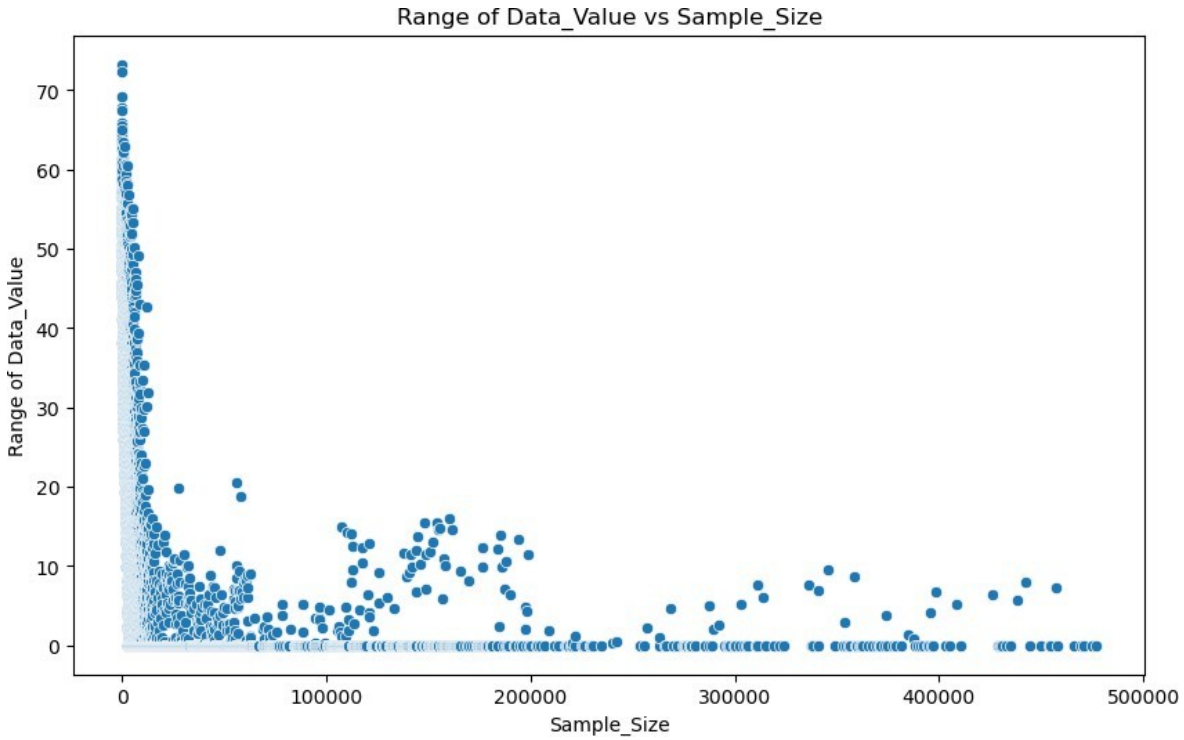


Figure5: Range of data value and sample size

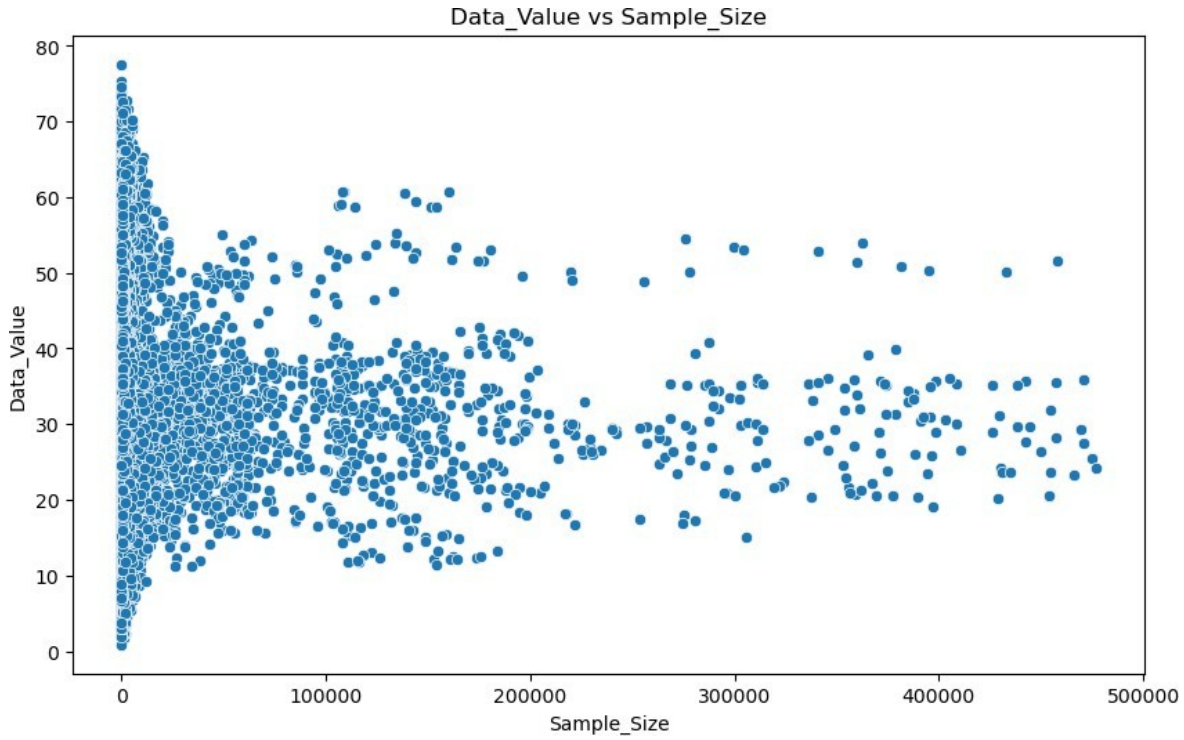


Figure6: Analysis between sample size and data value

Plotting Data Value vs Sample Size By examining and visualizing the link between sample size and health outcome indicators, scatter plots can be used to find possible trends, patterns, correlations, and relationships between the two variables. With a smaller range, lower variability, and more consistency in Data_Value, the scatter plot may reveal trends, patterns, and correlations between higher sample numbers and more consistent or targeted health outcome indicators.

The correlation between sample size (Sample_Size) and the stability, variability, and specificity of health outcome indicators (Data_Value) was revealed by the scatter plot analysis. A smaller range, reduced variability, and consistency in Data_Value may reveal possible trends, patterns, and correlations between higher sample numbers and more consistent or targeted health outcome measurements. These can be found by looking at the scatter plot.

Chapter 5: Research Analysis

Research questions:

1. Are there any temporal changes or shifts in the StratificationCategory1 used for obesity data collection across different time periods (YearStart, YearEnd)?
2. How has the prevalence of obesity changed over the years covered?
3. "Is there a substantial variation in obesity incidence (Data_Value) among various subgroups of the 'Class' attribute?"
4. Are there significant differences in obesity rates between different geographic locations?
5. How does the sample size ('Sample_Size') affect the reliability and generalizability of the reported health behavior estimates?
6. Are there any patterns or trends in the distribution of obesity rates (Data_Value) across different classes (ClassID), topics (TopicID), or locations (LocationID) over time?

Basic Importing:

In [1]:
```
import nltk
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.impute import SimpleImputer, KNNImputer
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression, LogisticRegression
from sklearn.ensemble import RandomForestRegressor, RandomForestClassifier
from sklearn.metrics import mean_squared_error, accuracy_score
```

In [2]:
```
df= pd.read_csv("Nutrition__Physical_Activity__and_Obesity_-_Behavioral_Risk_Factor_Surveillance_System2.csv")
```

In [3]:
```
df.head()
```

Out[3]:

	YearStart	YearEnd	LocationAbbr	LocationDesc	Class	Topic	Question	Data_Value	Low_Confidence_Limit	High_Confidence_Limit	...	Education	Inc
0	2020	2020	US	National	Physical Activity	Physical Activity - Behavior	Percent of adults who engage in no leisure-tim...	30.6	29.4	31.8	...	NaN	
1	2014	2014	GU	Guam	Obesity / Weight Status	Obesity / Weight Status	Percent of adults aged 18 years and older who ...	29.3	25.7	33.3	...	High school graduate	
2	2013	2013	US	National	Obesity / Weight Status	Obesity / Weight Status	Percent of adults aged 18 years and older who ...	28.8	28.1	29.5	...	NaN	50, – 74

We have initially imported the following key libraries: pandas as pd for handling of data, seaborn as sns, and matplotlib.pyplot as plt for displaying visualizations. We also imported several other sklearn libraries for performing various tasks like one hot encoding, linear regression, random forest regression, etc. We then loaded our dataset in the CSV format using the pd.read_csv() command and stored it in the variable and we displayed the first 5 records using the df.head() command.

About Dataset:

```
In [4]: df.columns

Out[4]: Index(['YearStart', 'YearEnd', 'LocationAbbr', 'LocationDesc', 'Class',
               'Topic', 'Question', 'Data_Value', 'Low_Confidence_Limit',
               'High_Confidence_Limit ', 'Sample_Size', 'Age(years)', 'Education',
               'Income', 'Race/Ethnicity', 'GeoLocation', 'ClassID', 'TopicID',
               'QuestionID', 'LocationID', 'StratificationCategory1',
               'Stratification1'],
              dtype='object')

In [5]: df.shape

Out[5]: (93249, 22)

In [6]: # Identify numerical and categorical columns
        numerical_cols = df.select_dtypes(include=['int', 'float']).columns
        categorical_cols = df.select_dtypes(include=['object']).columns

        print("Numerical Columns:")
        print(numerical_cols)

        print("\nCategorical Columns:")
        print(categorical_cols)

        Numerical Columns:
        Index(['YearStart', 'YearEnd', 'Data_Value', 'Low_Confidence_Limit',
               'High_Confidence_Limit ', 'Sample_Size', 'LocationID'],
              dtype='object')

        Categorical Columns:
        Index(['LocationAbbr', 'LocationDesc', 'Class', 'Topic', 'Question',
               'Age(years)', 'Education', 'Income', 'Race/Ethnicity', 'GeoLocation',
               'ClassID', 'TopicID', 'QuestionID', 'StratificationCategory1',
               'Stratification1'],
              dtype='object')
```

- The dataset contains 93249 rows and 22 columns. It has 7 numerical columns and 15 categorical columns.
- Numerical Columns: 'YearStart', 'YearEnd', 'Data_Value', 'Low_Confidence_Limit', 'High_Confidence_Limit ', 'Sample_Size', 'LocationID'
- Categorical Columns: 'LocationAbbr', 'LocationDesc', 'Class', 'Topic', 'Question', 'Age(years)', 'Education', 'Income', 'Race/Ethnicity', 'GeoLocation', 'ClassID', 'TopicID', 'QuestionID', 'StratificationCategory1', 'Stratification1'

Data Cleaning:

```
In [8]: df.isnull().sum()

Out[8]: YearStart                  0
        YearEnd                    0
        LocationAbbr               0
        LocationDesc               0
        Class                      0
        Topic                      0
        Question                   0
        Data_Value                 0
        Low_Confidence_Limit       0
        High_Confidence_Limit      0
        Sample_Size                0
        Age(years)             73269
        Education              79929
        Income                 69939
        Race/Ethnicity         66609
        GeoLocation                0
        ClassID                    0
        TopicID                    0
        QuestionID                 0
        LocationID                 0
        StratificationCategory1    9
        Stratification1            9
        dtype: int64
```

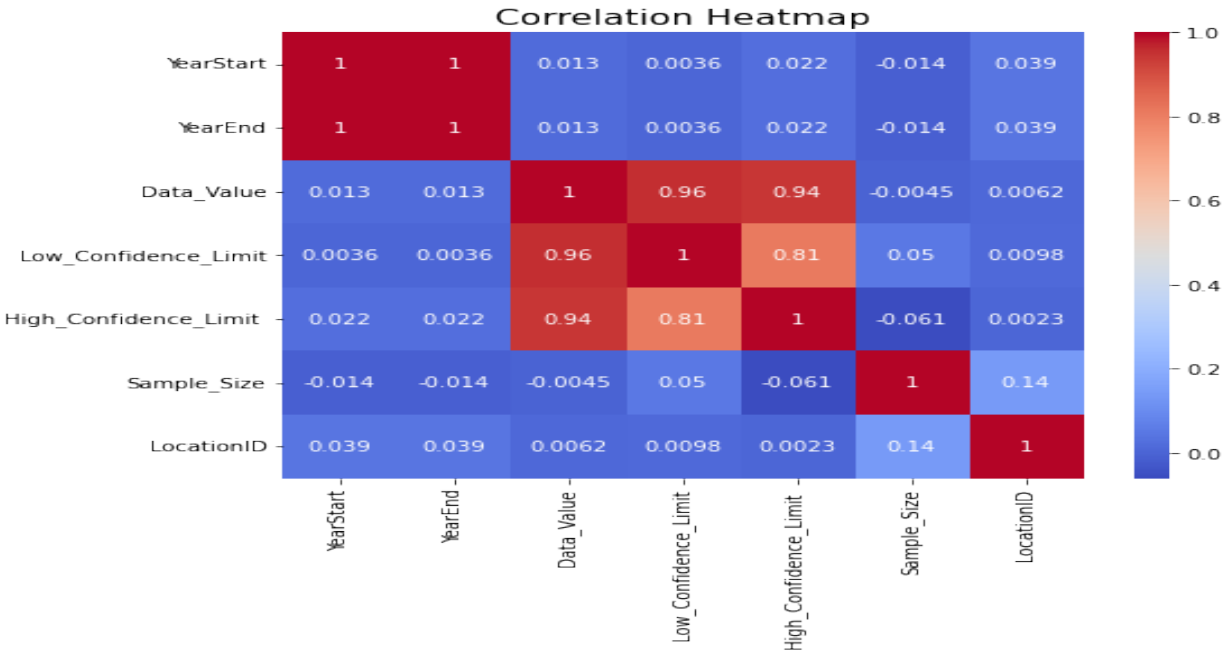
We used the isnull().sum() method to calculate the count of missing values for each attribute. We could see those attributes like Age(years), Education, Income and Race/ Ethnicity have large number of missing values. To treat the problem of missing values, we calculated the null value percentage of each attribute.

```
In [9]: total_attributes = len(df)
        attributes = ['YearStart', 'YearEnd', 'LocationAbbr', 'LocationDesc', 'Class',
               'Topic', 'Question', 'Data_Value', 'Low_Confidence_Limit',
               'High_Confidence_Limit ', 'Sample_Size', 'Age(years)', 'Education',
               'Income', 'Race/Ethnicity', 'GeoLocation', 'ClassID', 'TopicID',
               'QuestionID', 'LocationID', 'StratificationCategory1',
               'Stratification1']
        null_percentages = {}
        for attribute in attributes:
            null_count = df[attribute].isnull().sum()
            null_percentage = (null_count / total_attributes) * 100
            null_percentages[attribute] = null_percentage
            print(f"{attribute}: {null_percentage:.2f}%")
```

```
YearStart: 0.00%
YearEnd: 0.00%
LocationAbbr: 0.00%
LocationDesc: 0.00%
Class: 0.00%
Topic: 0.00%
Question: 0.00%
Data_Value: 0.00%
Low_Confidence_Limit: 0.00%
High_Confidence_Limit : 0.00%
Sample_Size: 0.00%
Age(years): 78.57%
Education: 85.72%
Income: 75.00%
Race/Ethnicity: 71.43%
GeoLocation: 0.00%
ClassID: 0.00%
TopicID: 0.00%
QuestionID: 0.00%
LocationID: 0.00%
StratificationCategory1: 0.01%
Stratification1: 0.01%
```

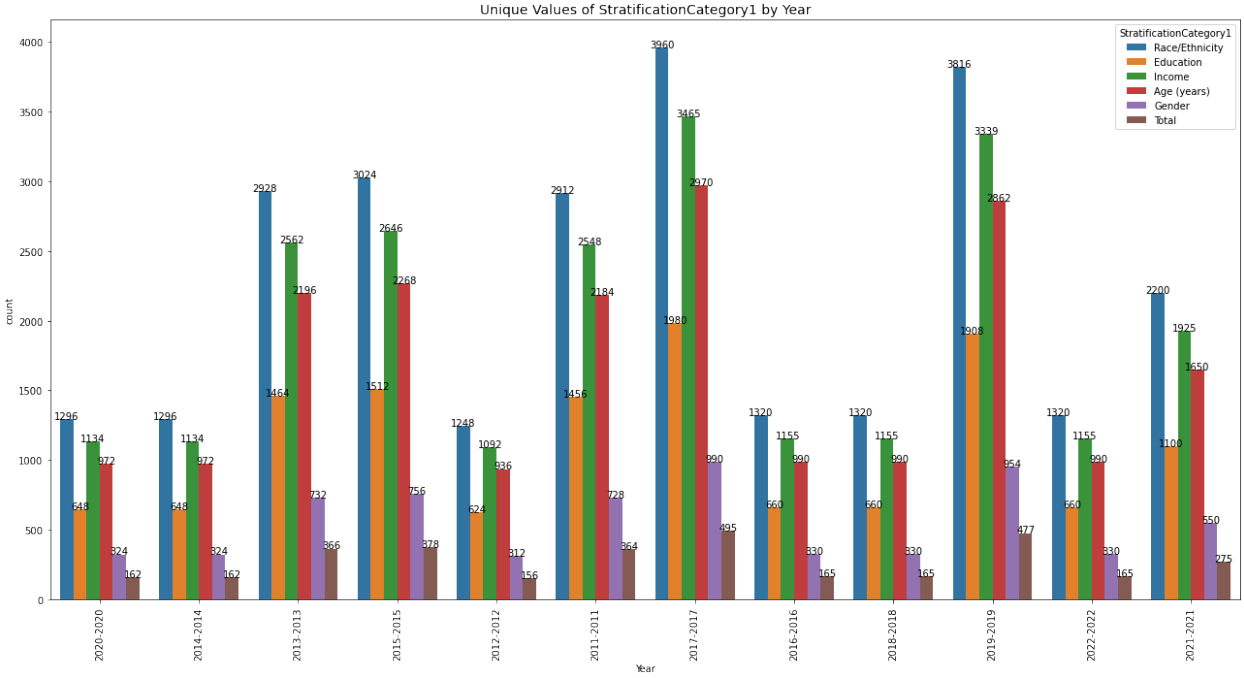
As part of data cleaning, we have checked if there are any null values in the dataset. We have set criteria to remove attributes with null values that had null value percentage above 50%. We found that the following attributes have the respective null values percentages, Age(years): 78.57% Education: 85.72%, Income: 75.00% Race/Ethnicity: 71.43%. We then dropped these attributes as they do not contribute much to our analysis. We also removed records that had missing values. The resultant dataset consists of 93240 rows and 18 columns.

Correlation Matrix:



Research Question 1:

Are there any temporal changes or shifts in the StratificationCategory1 used for obesity data collection across different time periods (YearStart, YearEnd)?



The bar graph indicates that the statistics for a number of demographic categories, such as age, gender, race/ethnicity, education, and total counts, fluctuate yearly. In particular, the highest numbers in each of these

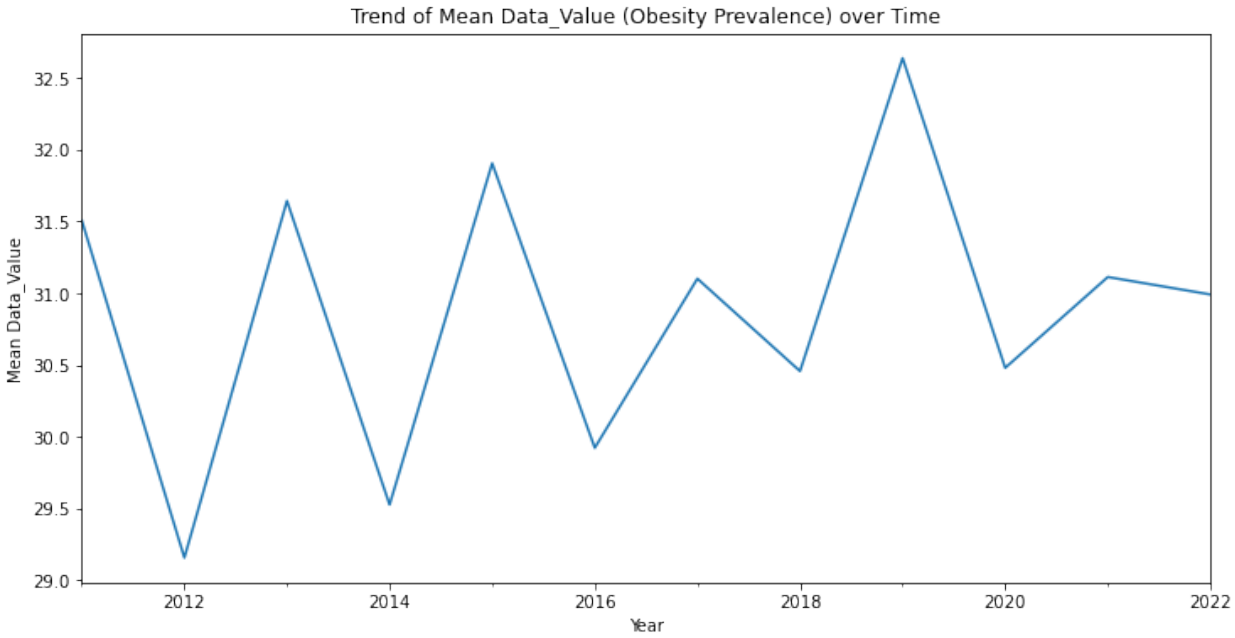
groups are found in the years 2017 and 2019. In contrast, the years with the lowest counts are 2012, 2014, 2018, 2020, and 2022. This unpredictability could be a sign of shifts in survey responses over time or adjustments made to data collection techniques. The causes of these tendencies and their consequences for comprehending demographic trends and developing focused measures to alleviate inequities should be investigated further through analysis.

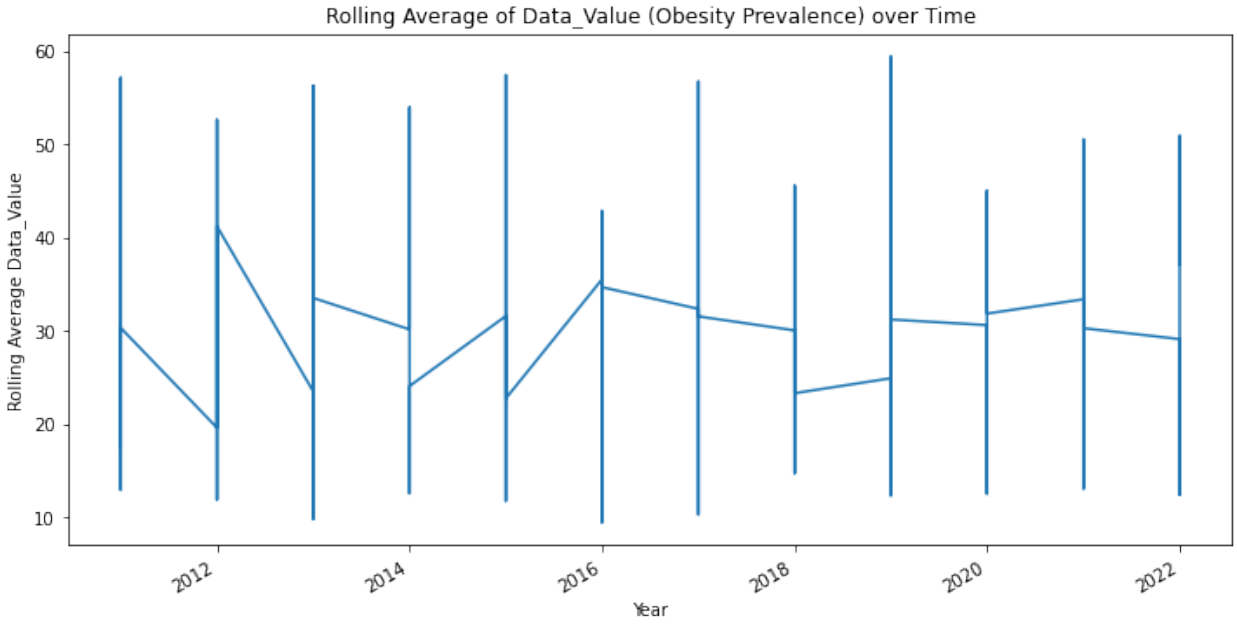
Hypothesis Testing:

```
In [42]: from scipy.stats import chi2_contingency

contingency_table = pd.crosstab(df['Year'], df['StratificationCategory1'])
chi2, p_value, dof, expected = chi2_contingency(contingency_table)

print("Null Hypothesis (H0): There is no association between 'StratificationCategory1' and the time period ('YearStart', 'YearEnd
print("Alternate Hypothesis (H1): There is an association between 'StratificationCategory1' and the time period ('YearStart', 'Ye

print(f"\nChi-square statistic: {chi2:.2f}")
print(f"Degrees of freedom: {dof}")
print(f"p-value: {p_value:.4f}")

if p_value < 0.05:
    print("\nWe reject the null hypothesis. There is evidence of an association between 'StratificationCategory1' and the time pe
else:
    print("\nWe fail to reject the null hypothesis. There is no evidence of an association between 'StratificationCategory1' and

Null Hypothesis (H0): There is no association between 'StratificationCategory1' and the time period ('YearStart', 'YearEnd').
Alternate Hypothesis (H1): There is an association between 'StratificationCategory1' and the time period ('YearStart', 'YearEn
d').

Chi-square statistic: 0.00
Degrees of freedom: 55
p-value: 1.0000

We fail to reject the null hypothesis. There is no evidence of an association between 'StratificationCategory1' and the time pe
riod ('YearStart', 'YearEnd').
```

The incredibly low chi-square statistic of 0.00 and the p-value of 1.0000 suggest that there is insufficient data to disprove the null hypothesis based on the findings. The p-value of 1.0000 indicates that, assuming there is no correlation between "StratificationCategory1" and the time period, the information found precisely matches the anticipated distribution. The stratification category utilized for the gathering of obesity data appears to be independent of the time period, since the chi-square test was unable to find any significant variation from the expected frequencies. Put differently, the findings imply that the stratification category in this dataset has neither altered or changed over time. Significant evidence is provided by the low chi-square statistic and high p-value to support the null hypothesis that there is no correlation between these two variables.

Research Question 2: How has the prevalence of obesity changed over the years covered?

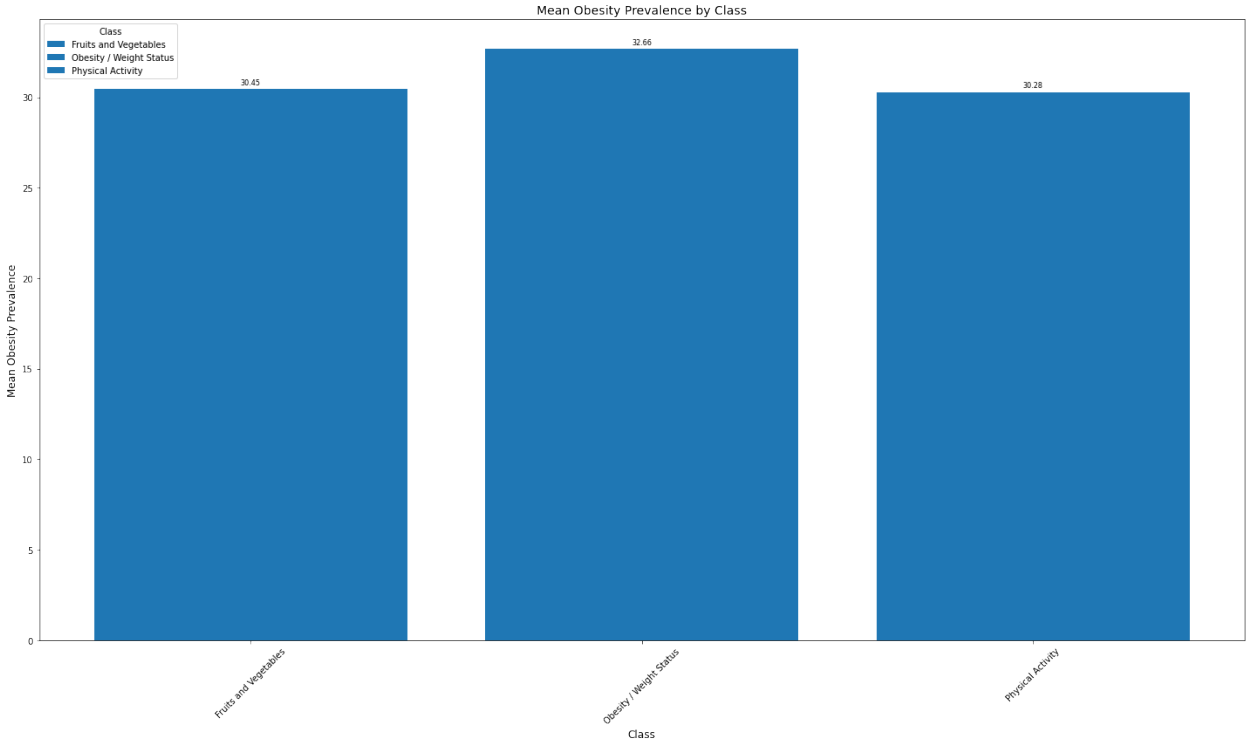




The line chart displays the 'Data_Value' (probably corresponding to a health-related action or result) for the years 2011 through 2022. According to the pattern's analysis, there is an annual trend in which the 'Data_Value'

fluctuates between greater and smaller values on an annual basis, roughly between 29% and 33%. The 'Data_Value' appears to have been quite consistent in the past few years (2020–2022), varying between 30 and 31%. 2019 saw the greatest value at 32.639698, while 2014 saw the lowest value at 29.525000.

Research Question 3 - Is there a substantial variation in obesity incidence (Data_Value) among various subgroups of the 'Class' attribute?




```
In [26]: import statsmodels.api as sm
from statsmodels.formula.api import ols

print("Null Hypothesis (H0): There is no significant difference in obesity prevalence (Data_Value) across different categories of
print("Alternative Hypothesis (Ha): At least one category of the variable of interest has a significantly different mean obesity

variable_of_interest = 'Class'
model = ols('Data_Value ~ C(' + variable_of_interest + ')', data=df).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
print(anova_table)
```

```
Null Hypothesis (H0): There is no significant difference in obesity prevalence (Data_Value) across different categories of the
variable of interest (e.g., Class, Topic, LocationAbbr).
Alternative Hypothesis (Ha): At least one category of the variable of interest has a significantly different mean obesity preva
lence (Data_Value).
                sum_sq       df           F         PR(>F)
C(Class)  1.228747e+05      2.0  646.906129  9.631410e-280
Residual  8.854818e+06  93237.0         NaN           NaN
```

If there is a significant difference in obesity rates (Data_Value) across different subgroups of the 'Class' attribute, it is inquired in the query. The bar graph indicates that there are three categories under the 'Class' attribute: 'Physical Activity, Fruits and Vegetables, and Obesity'.

The class 'Obesity' has a mean obesity rate of 32.66, a significantly higher rate than the classes 'Fruits and Vegetables' (30.45) and 'Physical Activity' (30.28). This implies that there is heterogeneity among the various classes in the rates of obesity.

In contrast, the difference in the 'Physical Activity' and 'Fruits and Vegetables' classes seems to be quite minimal (30.45 vs. 30.28). The 'Obesity' class appears to differ from the other two classifications in a more significant way.

Thus, in light of the bar graph, I would deduce that:

A significant amount of variance exists in the Data_Value for obesity rates between the subgroups of the 'Class' attribute, namely between the 'Obesity' class and the 'Fruits and Vegetables' and 'Physical Activity' classes.

Comparing the 'Obesity' class to the other two, the mean obesity rate is noticeably greater, suggesting that the data in this class are probably unique to the prevalence of obesity or its risk factors.

The 'Physical Activity' and 'Fruits and Vegetables' classes have comparatively similar mean obesity rates, indicating that these classes may be associated with dietary practices and physical activity levels, two factors that significantly impact obesity.

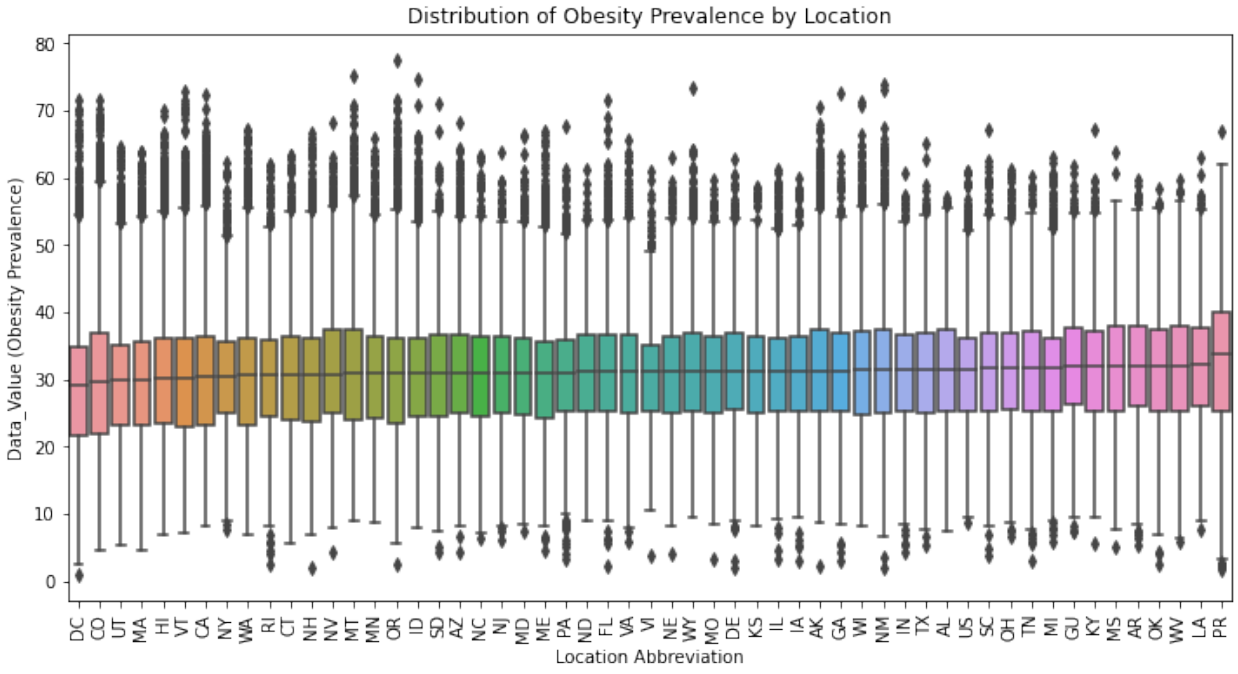
Data Preprocessing:

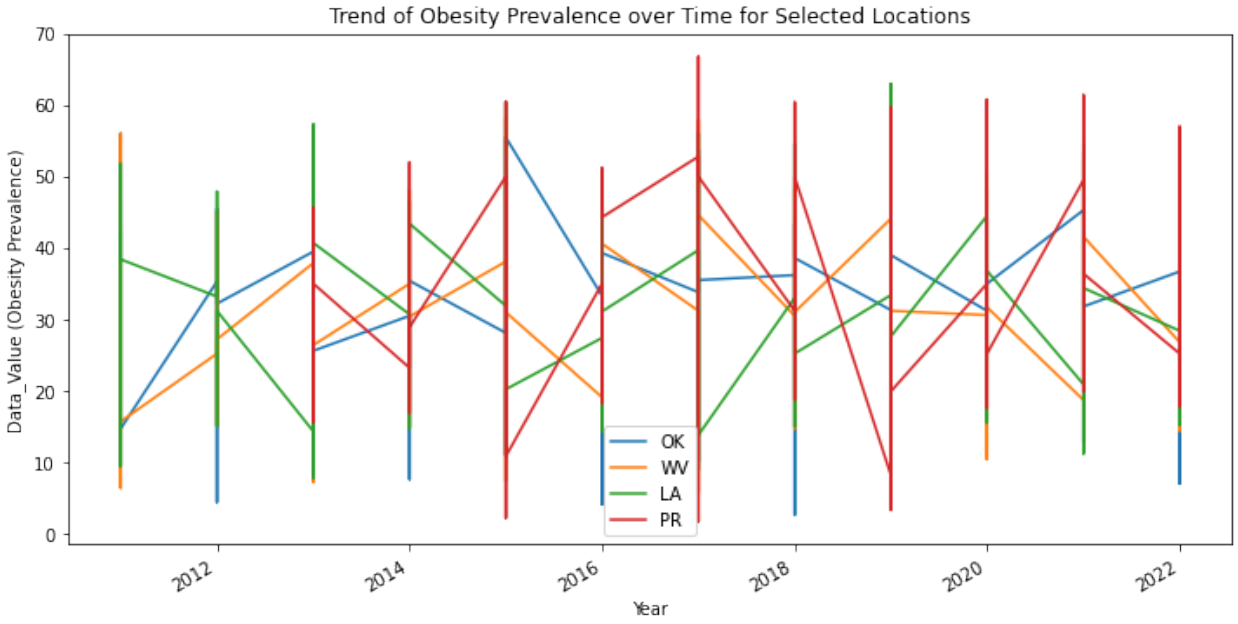
In [17]:
```
# Separate features and target variable
features = df.drop('Data_Value', axis=1)
target = df['Data_Value']


# Encoding categorical variables
categorical_cols = features.select_dtypes(include=['object']).columns
encoder = OneHotEncoder(handle_unknown='ignore')
encoded_features = pd.DataFrame(encoder.fit_transform(features[categorical_cols]).toarray())
encoded_features.columns = encoder.get_feature_names_out(categorical_cols)
features = features.drop(categorical_cols, axis=1)
features.reset_index(drop=True, inplace=True)
encoded_features.reset_index(drop=True, inplace=True)
features = pd.concat([features, encoded_features], axis=1, join='inner')


# Feature scaling
scaler = StandardScaler()
scaled_features = scaler.fit_transform(features)

# Split data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(scaled_features, target, test_size=0.2, random_state=42)
```

- As part of data pre-processing, first we performed the separation of the features and the target variable from the dataset.

- Secondly, we performed one-hot encoding on the categorical attributes to convert them into binary columns.
- We also performed scaling on the features to improve the performance of the models.
- Eventually, we performed the splitting of the cleaned dataset into training and test datasets. This allows us to train a certain portion of the data and use the test data set to determine the model's performance on the hidden patterns of data.

Research Question 4 - Are there significant differences in obesity rates between different geographic locations?

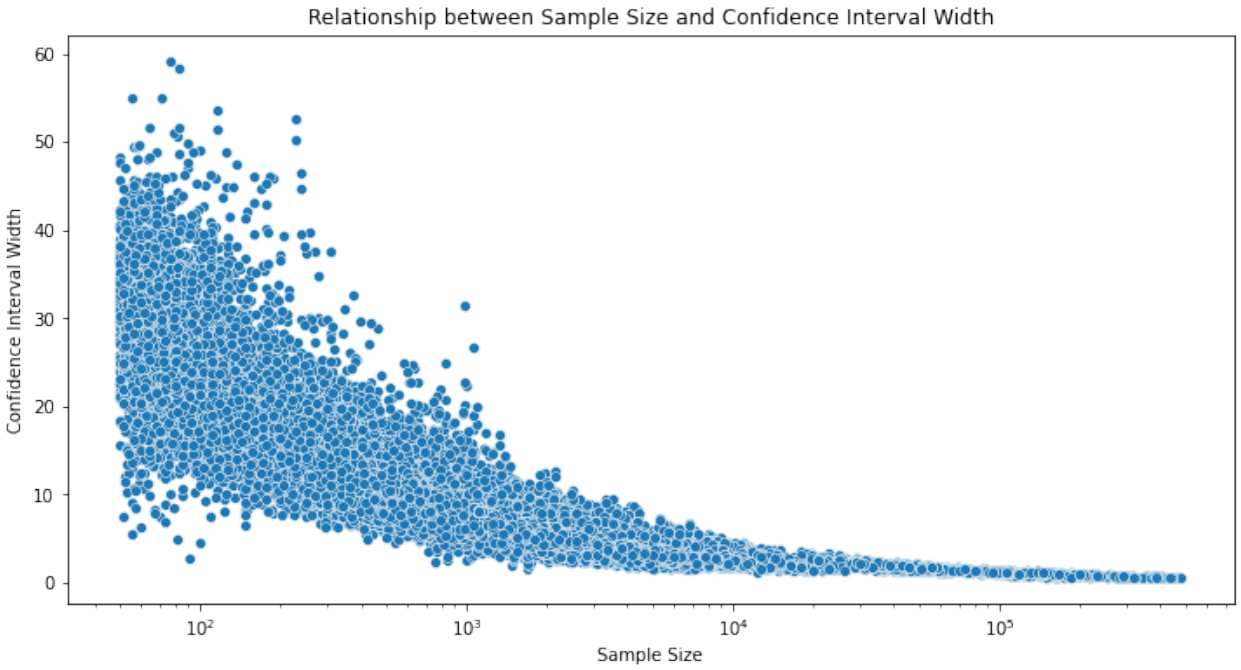




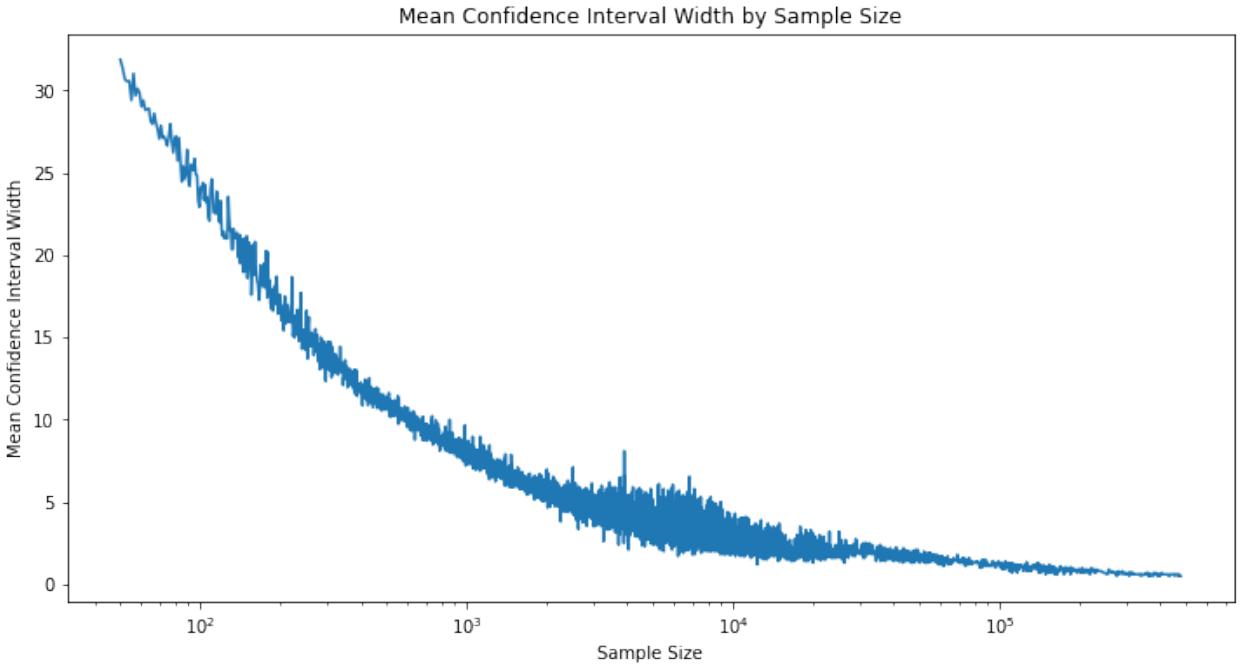
- The box plot illustrates the median rates of obesity prevalent in various states, with values that range from approximately 29% in DC, CO, and UT to approximately 33.8% in PR. The plot also showed a substantial number of outliers on the upper quartile for all locations, with obesity prevalence estimates as high as 80%. The US as a whole has a median rate of obesity of 31.6%, however, states vary greatly in this regard; some have relatively low rates (such as MA, HI, and VT) and others have comparatively high rates (like AR, OK, and MS). Significantly greater obesity rates may be caused by certain demographic groupings, socioeconomic difficulties, or other fundamental reasons, as suggested by the existence of outliers on the upper end of the distribution.

```
In [62]: # Create contingency table
contingency_table = pd.crosstab(df['LocationAbbr'], df['Data_Value'])

# Perform chi-square test
chi2, p_value, dof, expected = chi2_contingency(contingency_table)

print(f"Chi-Square Statistic: {chi2:.2f}")
print(f"Degrees of Freedom (DOF): {dof}")
print(f"P-Value: {p_value:.4f}")

Chi-Square Statistic: 45008.39
Degrees of Freedom (DOF): 37476
P-Value: 0.0000
```

Research Question 5 - How does the sample size ('Sample_Size') affect the reliability and generalizability of the reported health behavior estimates?



- The scatter plot demonstrates the inverse relationship between sample size and confidence interval (CI) width, wherein bigger sample sizes yield more accurate predictions and smaller CI width ranges, signifying decreased ambiguity and increased confidence in the estimations. The CI width for a sample size of 100 spans from 0 to 60, suggesting a comparatively high degree of ambiguity. The range narrows to 0 to 30 when the study size grows to 1000, lowering the level of uncertainty but maintaining a sizable range of potential CI widths. The range progressively decreases to 0 to 10 with a sample size of 10,000, greatly lowering the uncertainty and producing more accurate estimations. Finally, the range of values of CI width is smallest for a sample size that is exceedingly large of 100,000, ranging from 0 to 5, showing minimal ambiguity and continuously tiny CI widths. Larger sample sizes typically result in more accurate estimates, but the CI width can also be affected by other variables like data fluctuation, the fundamental distribution, and the confidence level selected.



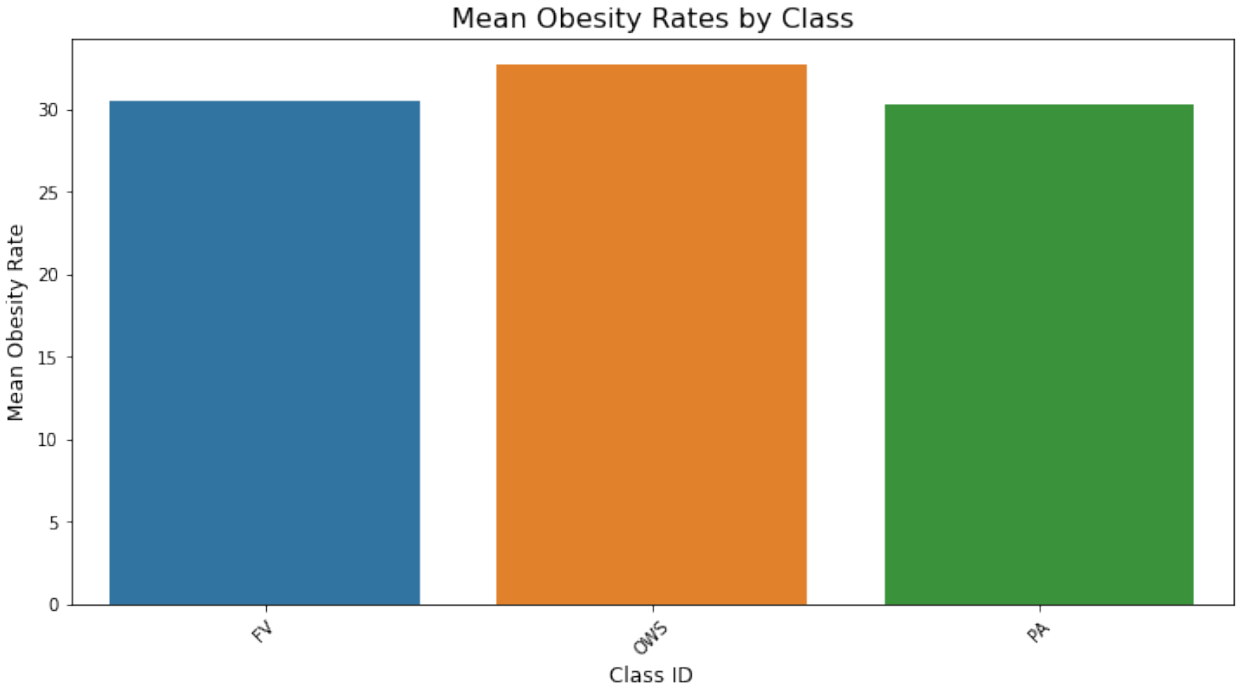
Sample Size

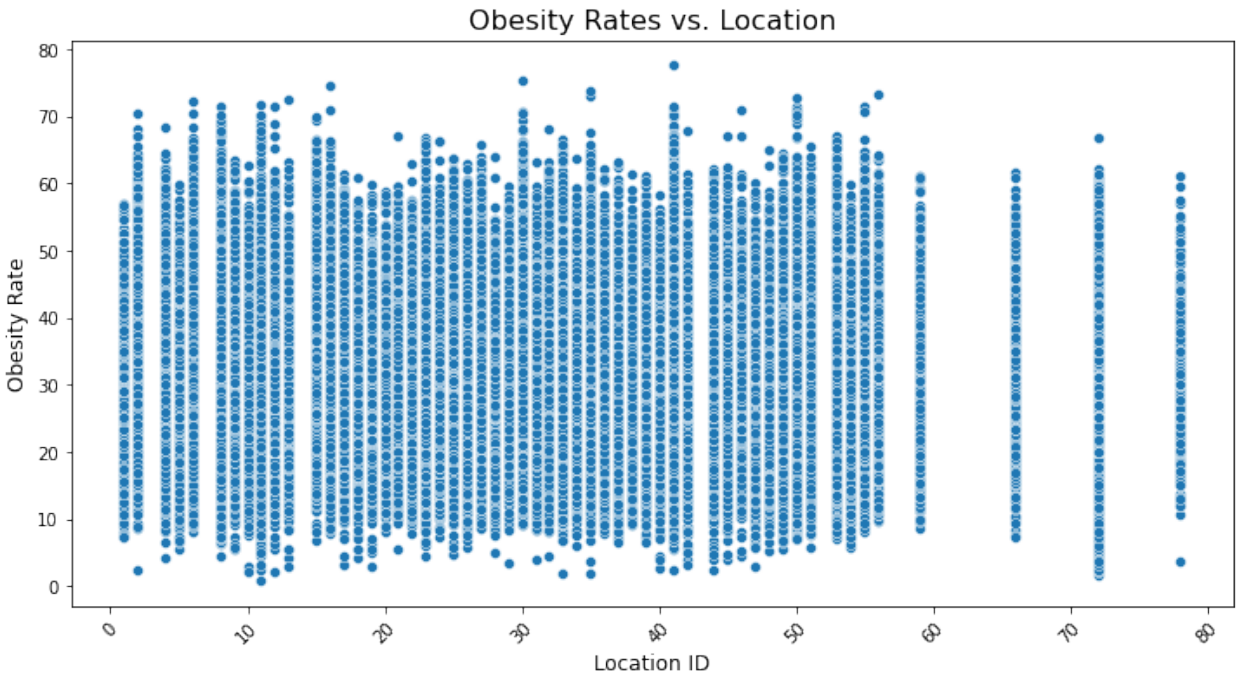
```
In [31]: # Create contingency table
contingency_table = pd.crosstab(df['Sample_Size'], df['CI_Width'])

# Perform chi-square test
chi2, p_value, dof, expected = chi2_contingency(contingency_table)

print(f"Chi-Square Statistic: {chi2:.2f}")
print(f"Degrees of Freedom (DOF): {dof}")
print(f"P-Value: {p_value:.4f}")
```

```
Chi-Square Statistic: 12778510.81
Degrees of Freedom (DOF): 15218958
P-Value: 1.0000
```

The accuracy and adaptability of the observed health-related behavior predictions appear to be unaffected by the sample size, as indicated by the extraordinarily high p-value of 1.0000 and the inability to reject the null hypothesis. The precision or accurate representation of the estimations did not correlate with the sample size, according to the results of the chi-square test. Assuming that there is no correlation between sample size and estimate reliability/generalizability, this suggests that the observed data exactly match the expected distribution.

Research Question 6 - Are there any patterns or trends in the distribution of obesity rates (Data_Value) across different classes (ClassID), topics (TopicID), or locations (LocationID) over time?


Mean Obesity Rates by Class
Mean Obesity Rate
0
5
10
15
20
25
30
FV
OWS
PA
Class ID


Obesity Rates vs. Location
Obesity Rate
0
10
20
30
40
50
60
70
80
Location ID

Topic ID

In [35]:
```
# Discretize the 'Data_Value' column into categories (e.g., normal, overweight, obese)
bins = [0, 25, 30, float('inf')]
labels = ['Normal', 'Overweight', 'Obese']
df['Obesity_Category'] = pd.cut(df['Data_Value'], bins=bins, labels=labels, include_lowest=True)

# Create a contingency table
contingency_table = pd.crosstab([df['Year'], df['ClassID'], df['TopicID'], df['LocationID']], df['Obesity_Category'])

# Perform the chi-square test
chi2_statistic, p_value, dof, expected = chi2_contingency(contingency_table)

# Print the results
print("Chi-Square Test Results:")
print(f"Chi-Square Statistic: {chi2_statistic:.2f}")
print(f"Degrees of Freedom: {dof}")
print(f"P-Value: {p_value:.4f}")
```

```
Chi-Square Test Results:
Chi-Square Statistic: 18956.86
Degrees of Freedom: 2910
P-Value: 0.0000
```

The chi-square statistic of 18956.86, which is extraordinarily big, and the p-value of 0.0000, which is nearly zero, offer compelling evidence against the null hypothesis. The conclusion is that the distribution of obesity rates among various classes, subjects, places, and eras exhibits statistically significant patterns or trends. There is a significant deviation between the observed data and the anticipated distribution assuming no association, suggesting that the variables have strong correlations. When taking into account combinations of class, topic, place, and time in this dataset, the results point to the existence of noteworthy patterns or trends in the prevalence of obesity.

Chapter 6: Evaluation Metrics

Performing evaluation metrics for any data analysis is crucial to quantify the performance, workout benchmarking, help decision-making, identifying strengths and weaknesses and simplifying improvement. They provide norm for testing methods, guiding the selection of the most efficient strategies, and driving perpetual enhancement of solutions over time.

Mean Absolute Error (MAE)

The average absolute difference between the expected and actual values is measured by the Mean Absolute Error, or MAE. Because it is in the same unit as the objective variable and is simple to understand, it is a frequently used metric for regression problems. Because it signifies the model is forecasting values that are closer to the actual values, a lower MAE denotes higher model performance.

Mean Squared Error (MSE)

An indicator of the average squared difference between expected and actual values is the mean squared error, or MSE. Squared differences allow for the identification of models with substantial mistakes, making it more sensitive to flaws than MAE. When the MSE is smaller, the model is performing better since it is predicting values that are more in line with reality.

Root Mean Squared Error (RMSE)

The square root of the mean square error (MSE), or root mean squared error (RMSE), returns the error metric to the same unit as the target variable. It is a widely used measure that strikes a compromise between the interpretability of MAE and the sensitivity to significant errors of MSE.

Better model performance is indicated by a lower RMSE since it suggests the model is producing predictions that are more in line with the actual values.

R-squared (R^2)

The degree to which the model's predictions agree with the variance in the actual values is indicated by the R-squared (R^2) metric. Higher values suggest a better fit. The range is 0 to 1. A model that fully explains all of the variation in the data is said to have an R2 of 1, whereas a one that partially explains all of the variation is said to have an R2 of 0. Better model performance is indicated by a greater R2 since it signifies the model is able to explain a larger percentage of the variation in the target variable.

```
# Code snippet 6
# Displaying statistical descriptive summary of numerical variables
data.describe()
```

	YearStart	YearEnd	Data_Value	Low_Confidence_Limit	High_Confidence_Limit	Sample_Size	LocationID
count	93249.000000	93249.000000	84014.000000	84014.000000	84014.000000	84014.000000	93249.000000
mean	2016.308068	2016.308068	31.226492	26.890256	36.134303	3649.343597	30.953447
std	3.308679	3.308679	10.021059	9.816064	10.978276	18680.688957	17.532688
min	2011.000000	2011.000000	0.900000	0.300000	3.000000	50.000000	1.000000
25%	2013.000000	2013.000000	24.400000	20.100000	28.700000	511.000000	17.000000
50%	2017.000000	2017.000000	31.200000	26.900000	36.000000	1103.000000	30.000000
75%	2019.000000	2019.000000	37.000000	32.900000	42.200000	2405.000000	45.000000
max	2022.000000	2022.000000	77.600000	70.200000	87.700000	476876.000000	78.000000

Figure 13: *Calculation if Mean, Median, Mode*

Insights and Conclusion

Data Time Range: The YearStart and YearEnd columns have identical statistics, which implies that for each record, the start and end years are the same. The data spans from 2011 to 2022.

Data Volume: There are 93,249 observations in the dataset as indicated by the count for YearStart, YearEnd, and LocationID.

Data_Value Analysis: The Data_Value column has a mean (average) of approximately 31.23, with a standard deviation of 10.02, which suggests moderate variability around the mean.

The minimum value is very close to 0 (0.9), and the maximum value is 77.6, indicating a wide range of values. The 25th percentile is 24.4, the median (50th percentile) is 31.2, and the 75th

percentile is 37, indicating that the data is somewhat skewed since the mean is less than the median.

Confidence Intervals: The Low_Confidence_Limit has a mean of approximately 26.89 and the High_Confidence_Limit has a mean of approximately 36.13.

The confidence interval widens significantly from the minimum (0.3 to 3) to the maximum values (70 to 87.7), which may indicate that the variability or uncertainty in the data increases with higher values.

Sample Size Variability: The Sample_Size column has a very high standard deviation (18,680.69) relative to its mean (3,649.34), indicating a large variability in sample sizes across different observations. The minimum sample size is 50, and the maximum is a very large number of 476,876, which suggests that some data points are based on much larger samples than others.

Location Distribution: The LocationID column has a minimum of 1 and a maximum of 78, which could represent different locations or regions coded numerically from 1 to 78. The mean and median values are around 30, suggesting that locations are evenly distributed across the dataset if they are uniformly distributed.

This statistical summary provides a general overview of the dataset, indicating the presence of a wide range of values for Data_Value and Sample_Size and an even distribution of years and locations. However, to gain deeper insights, more context on what these variables represent would be necessary, as well as further exploratory data analysis such as visualizations and hypothesis testing.

Model Evaluation:

In [18]:
```
from sklearn.metrics import mean_squared_error, r2_score

# Create and fit the linear regression model
linear_reg = LinearRegression()
linear_reg.fit(X_train, y_train)

# Make predictions on the test set
y_pred = linear_reg.predict(X_test)

# Calculate evaluation metrics
mse_linear_reg = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print(f'Mean Squared Error: {mse_linear_reg:.2f}')
print(f'R-squared: {r2:.2f}')
```

```
Mean Squared Error: 0.10
R-squared: 1.00
```

In [19]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
# Create and fit the logistic regression model
logistic_reg = LogisticRegression(random_state=42)
logistic_reg.fit(X_train, y_train > y_train.median())

# Make predictions on the test set
y_pred = logistic_reg.predict(X_test)

accuracy_log_reg = accuracy_score(y_test > y_test.median(), y_pred)
print(f'Logistic Regression Accuracy: {accuracy_log_reg:.2f}')
```

```
Logistic Regression Accuracy: 1.00
```

- The Logistic Regression model performed exceptionally well, as evidenced by its accuracy of 1.0 (or 100%), which shows that it properly identified every instance in the test set based on the binarized target variable (i.e., whether the target value is above or below the median). This flawless accuracy score, however, can also be a sign of overfitting, in which the model has become overly adept at handling the noise and abnormalities present in the training set, which could result in subpar performance when dealing with fresh, untested data.

- The outcome that has been provided indicates that the linear regression model has an R-squared (R^2) value of 1.00 and a Mean Squared Error (MSE) of 0.10. The forecasts made by the model are, on average, extremely close to the actual values, with a reasonably little error, as indicated by the low MSE. The ideal fit is indicated by the R2 value of 1.00, which indicates that the model can use the given features to explain 100% of the variance in the target variable. The correlation between the linear regression model's high R^2 and low MSE indicates that it is operating extraordinarily effectively on the provided dataset.

[41]:
```
# Random Forest Regression
rf_reg = RandomForestRegressor(n_estimator
rf_reg.fit(X_train, y_train)
y_pred = rf_reg.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
print(f'Random Forest Regression MSE: {mse
```

```
Random Forest Regression MSE: 0.01
```

[42]:
```
# Random Forest Classification
rf_clf = RandomForestClassifier(n_estimato
rf_clf.fit(X_train, y_train > y_train.medi
y_pred = rf_clf.predict(X_test)
accuracy_rf_clf = accuracy_score(y_test >
print(f'Random Forest Classification Accur
```

```
Random Forest Classification Accuracy: 0.9
```

- The Random Forest Classification model demonstrates an exceptional accuracy of 0.99 (or 99%), implying it accurately categorized 99% of the test set instances depending on the binarized target variable. The Random Forest Regression model accomplishes an exceptionally low Mean Squared Error (MSE) of 0.01 in the provided output, demonstrating that its forecasts are extremely similar to the actual values with little variance. These assessment indicators indicate that both models are operating at a very high level; the Random Forest Classification model is exhibiting outstanding effectiveness in categorizing instances based on the attributes provided, and the Random Forest Regression model is producing extremely precise forecasts for the continuous target variable.

Chapter 7: Conclusion

Discussion

Research topic that's been given in the report addresses common health issues that are that are focused on Laziness, Fatness i.e. Obesity and poor Nutrition in the United States. By using the data from Behavioral Risk Factor Surveillance System (BRFSS), the study is done to reveal the stats of the Mental Health, Chronic Health issues, and their preventive methods utilization among the adult Americans. From the selection of software, collecting the data, technology used for analyzing the data are tough and chosen perfectly for the research purpose. Literature review in the research paper gives the valuable information by highlighting the identity of health issues and surveying by intervention and solving in all sides. In context of study, it helps to put together all the questions in the research and let us the next analysis steps. Methodology chapter in the research paper gives the rough idea of step-to-step process of processing the data and Make sure that the dataset that's been used is worth for Exploring Data analysis and Research analysis.

The Exploratory Data Analysis (EDA) and Research Analysis chapters offers us the understanding the topics such as regional variations, inequalities, and social class effect on the health issues.
Usage of Visualization topics such as histograms, scatter plots, and box plotting helps us to recognize the patterns and relation between each variable available in the dataset. In conclusion the study gives the overview of understanding what are the factors that are affecting the common health issues that are focused on Laziness, Fatness i.e Obesity and poor Nutrition in the United States.

Applications

From the research, we can find the health policies and interventions that mainly aimed at addressing obesity, physical activity, and poor nutrition. By knowing about the regional variations and socioeconomic influences we can filter the people who are at high risk. The insights from the research can lead to healthier lifestyle as it would help in developing community-based programs and educational initiatives.

At the same times the insights from the research would draw a foundation to the future studies (i.e. the dataset and analysis) in health-related fields. Alongside, the researchers can also find the longitudinal studies in order to tract trends over time, exploring more factors that can influence the health outcomes and to know how effectively the intervention strategies are working.

Future Works

Based on the current study, there could be an expansion in the further research, this can happen by integrating the datasets and the further more data from various eras to analyze the patterns over time. In the long haul, panel studies can help outburst the changes in policies and interruptions. To understand this deeper, focus groups and interviews can be a great source of interpretative research, which furnish the fundamental elements of health conduct and result.

More distantly, to magnify the predictive modeling potential of the research we can embody measurements to approximate a model, like ML algorithms. Researcher can spot people at high probability for obesity kind health issues and develop individual -centered blueprint to diminish tasks, this can be done by supporting advanced analytics.

Limitations

Apart from the positives the research and the dataset have many limitations that should be noticed. Firstly, the data is obtained from the survey which might subject to the biases and the inaccuracies.

Furthermore, the dataset may not contain all the related factors that influence the health results, such as genetic or natural personal effects. The scope of the research is limited to the data obtained from only BRFSS source which may not cover all the aspects of physical activity, nutrition and obesity. As such, outcomes may not completely capture the complexity of these problems.

On top of that, the analysis is only performed on the observed data, which includes building causality between variables. While correlations are found, causative relationships cannot be inferred in the research. Finally, the research may face complexities related to generalizability, as the outcomes may not represent all the demographic groups or geographic places within the United States.

References

U.S. Department of Health & Human Services - Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System. (2023, December 8). https://catalog.data.gov/dataset/nutrition-physical-activity-and-obesity-behavioral-risk-factor-surveillance-system

https://www.cdc.gov/nccdphp/dnpao/division-information/aboutus/index.html

Baillot, A., Chenail, S., Polita, N. B., Simoneau, M., Libourel, M., Nazon, É., Riesco, É., Bond, D. S., & Romain, A. J. (2021). Physical activity motives, barriers, and preferences in people with obesity: A systematic review. *PLOS ONE*, *16*(6), e0253114. https://doi.org/10.1371/journal.pone.0253114

Bish, C. L., Blanck, H. M., Serdula, M. K., Marcus, M., Kohl, H. W., & Khan, L. K. (2005). Diet and Physical Activity Behaviors among Americans Trying to Lose Weight: 2000 Behavioral Risk Factor Surveillance System. *Obesity Research, 13*(3), 596–607. https://doi.org/10.1038/oby.2005.64.

BRFSS Prevalence & Trends Data: Home | DPH | CDC. (n.d.). https://www.cdc.gov/brfss/brfssprevalence/index.html

De Onís, M., Borghi, E., Arimond, M., Webb, P., Croft, T., Saha, K. K., De-Regil, L. M., Thuita, F., Heidkamp, R., Krasevec, J., Hayashi, C., & Flores-Ayala, R. (2018). Prevalence thresholds for wasting, overweight and stunting in children under 5 years. *Public Health Nutrition, 22*(1), 175–179. https://doi.org/10.1017/s1368980018002434

Ding, L., Liang, Y., Tan, E. C., Hu, Y., Zhang, C., Liu, Y., Xue, F., & Wang, R. (2020).

Smoking, heavy drinking, physical inactivity, and obesity among middle-aged and older adults in China: cross-sectional findings from the baseline survey of CHARLS 2011–

2012. *BMC Public Health*, *20*(1). https://doi.org/10.1186/s12889-020-08625-5

Jennings, V., & Bamkole, O. (2019). The Relationship between Social Cohesion and Urban Green Space: An Avenue for Health Promotion. *International Journal of Environmental Research and Public Health, 16*(3), 452. https://doi.org/10.3390/ijerph16030452

Napolitano, M. A., Jakicic, J. M., Fulton, J. E., & Tennant, B. (2019). Physical Activity Promotion: Highlights from the2018 Physical Activity Guidelines Advisory Committee Systematic Review. *Medicine and Science in Sports and Exercise, 51*(6), 1340–1353. https://doi.org/10.1249/mss.0000000000001945

Lacombe, J., Armstrong, M., Wright, F. L., & Foster, C. (2019). The impact of physical activity and an additional behavioural risk factor on cardiovascular disease, cancer and all-cause mortality: a systematic review. *BMC Public Health, 19*(1). https://doi.org/10.1186/s12889-019-7030-8

Oakman, J., Kinsman, N., Stuckey, R., Graham, M., & Weale, V. (2020). A rapid review of mental and physical health effects of working at home: how do we optimise health? *BMC Public Health, 20*(1). https://doi.org/10.1186/s12889-020-09875-z

Serdar, C. C., Cihan, M., Yücel, D., & Serdar, M. (2021). Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochemia Medica, 31*(1), 27–53. https://doi.org/10.11613/bm.2021.010502

Appendix -1: Code

```
import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sns
# Load the dataset

obesity_data = pd.read_csv(r"C:\Users\velkanti varshitha\Dropbox\My PC (DESKTOP-
PRRVEFF)\Downloads\Nutrition_Physical_Activity_and_Obesity_-
_Behavioral_Risk_Factor_Surveillance_System2.csv")
# Filter the dataset for the desired columns
data = obesity_data[['Data_Value', 'Education', 'YearStart', 'YearEnd', 'GeoLocation']]



# Remove any rows with missing values
data = data.dropna()



# Analyze trends in Data_Value relative to Education levels across YearStart and YearEnd range
for education in data['Education'].unique():
    education_data = data[data['Education'] == education]
    plt.figure(figsize=(12, 6))
    sns.lineplot(x='YearStart', y='Data_Value', data=education_data, label='YearStart')
```

```
    sns.lineplot(x='YearEnd', y='Data_Value', data=education_data, label='YearEnd')
    plt.title(f'Data_Value Trend for Education Level: {education}')
    plt.xlabel('Year')
    plt.ylabel('Data_Value')
```

```
    plt.legend()
    plt.show()


# Analyze geographic differences in reported health outcomes
plt.figure(figsize=(12, 6))
sns.boxplot(x='GeoLocation', y='Data_Value', data=data)
plt.xticks(rotation=90)
plt.title('Geographic Differences in Data_Value')
plt.xlabel('Geographic Location')
plt.ylabel('Data_Value')
plt.show()



import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sns



# Load the dataset

obesity_data = pd.read_csv(r"C:\Users\velkanti varshitha\Dropbox\My PC (DESKTOP-
PRRVEFF)\Downloads\Nutrition_Physical_Activity_and_Obesity_-
```

```
_Behavioral_Risk_Factor_Surveillance_System2.csv")



# Filter the dataset for the desired columns
```

```
data = obesity_data[['Data_Value', 'Income', ]]



# Remove any rows with missing values
data = data.dropna()



# Create a box plot to visualize the distribution of Data_Value across income levels for different
racial groups
plt.figure(figsize=(12, 8))

sns.boxplot(x='Income',  y='Data_Value',  data=data,)
plt.xticks(rotation=90)
plt.title('Distribution of Data_Value by Income and Race')
plt.xlabel('Income Level')
plt.ylabel('Data_Value')
plt.show()



import pandas as pd

import  matplotlib.pyplot  as  plt
import seaborn as sns
```

```
# Load the dataset


# Replace 'data.csv' with the actual file path and name


obesity_data = pd.read_csv(r"C:\Users\velkanti varshitha\Dropbox\My PC (DESKTOP-
PRRVEFF)\Downloads\Nutrition_Physical_Activity_and_Obesity_-
```

```
_Behavioral_Risk_Factor_Surveillance_System (4).csv")


# Check for missing values in relevant columns

relevant_cols = ['Data_Value', 'LocationAbbr', 'GeoLocation']
print(obesity_data[relevant_cols].isnull().sum())


# Drop rows with missing values in relevant columns
obesity_data = obesity_data.dropna(subset=relevant_cols)


# Convert 'GeoLocation' to latitude and longitude columns

obesity_data[['Latitude', 'Longitude']] = obesity_data['GeoLocation'].str.split(',', expand=True)


# Convert 'Latitude' and 'Longitude' to numeric

obesity_data['Latitude'] = pd.to_numeric(obesity_data['Latitude'], errors='coerce')
obesity_data['Longitude'] = pd.to_numeric(obesity_data['Longitude'], errors='coerce')
```

```
# Plotting Data_Value across various locations
plt.figure(figsize=(15, 8))
sns.boxplot(data=obesity_data, x='LocationAbbr', y='Data_Value', palette='Set2')
plt.title('Comparison of Data_Value across Various Locations')
plt.xlabel('LocationAbbr')
plt.ylabel('Data_Value')
plt.xticks(rotation=90)
```

```
plt.show()
```

EDA

```
import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sns



# Load the dataset

obesity_data = pd.read_csv('Nutrition_and_Obesity.csv')



 Filter the dataset for the desired columns

data = obesity_data[['Data_Value', 'LocationDesc']]



# Remove any rows with missing values
data = data.dropna()



# Group the data by state and calculate the mean health outcome
```

```
state_means = data.groupby('LocationDesc')['Data_Value'].mean().reset_index()



# Sort the states by mean health outcome

state_means = state_means.sort_values(by='Data_Value')



# Plot the mean health outcome for each state

plt.figure(figsize=(12, 6))
```

```
sns.barplot(x='LocationDesc', y='Data_Value', data=state_means)
plt.xticks(rotation=90)
plt.title('Mean Health Outcome (Data_Value) by State')
plt.show()


# Plot the distribution of health outcomes for each state
plt.figure(figsize=(12, 6))
sns.violinplot(x='LocationDesc', y='Data_Value', data=data)
plt.xticks(rotation=90)
plt.title('Distribution of Health Outcomes (Data_Value) by State')
plt.show()


import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sns


# Load the dataset

obesity_data = pd.read_csv(r"C:\Users\velkanti varshitha\Dropbox\My PC (DESKTOP-
PRRVEFF)\Downloads\Nutrition_Physical_Activity_and_Obesity_-
```

```
_Behavioral_Risk_Factor_Surveillance_System2.csv")



# Filter the dataset for the desired columns


data = obesity_data[['Data_Value', 'Topic', 'Sample_Size']]
```

```
# Remove any rows with missing values
data = data.dropna()


# Plot Data_Value trends over time segmented by different health topics
plt.figure(figsize=(12, 6))
for topic in data['Topic'].unique():
    topic_data = data[data['Topic'] == topic]
    sns.lineplot(x='Sample_Size', y='Data_Value', data=topic_data, label=topic)
plt.title('Data_Value Trends Over Time by Health Topic')
plt.xlabel('Year')
plt.ylabel('Data_Value')
plt.legend()
plt.show()


# Check for correlation between Sample_Size and Data_Value
plt.figure(figsize=(8, 6))
sns.scatterplot(x='Sample_Size', y='Data_Value', data=data)
plt.title('Correlation Between Sample_Size and Data_Value')
plt.xlabel('Sample_Size')
plt.ylabel('Data_Value')
```

```
plt.show()
```

```
# Calculate and print the correlation coefficient
correlation = data['Sample_Size'].corr(data['Data_Value'])
print(f'Correlation between Sample_Size and Data_Value: {correlation}')



import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sns


# Load the dataset

obesity_data = pd.read_csv(r"C:\Users\velkanti varshitha\Dropbox\My PC (DESKTOP-
PRRVEFF)\Downloads\Nutrition_Physical_Activity_and_Obesity_-
_Behavioral_Risk_Factor_Surveillance_System (4).csv")



# Filter the dataset for the desired columns

data = obesity_data[['Data_Value', 'Sample_Size']]
```

```
# Remove any rows with missing values
data = data.dropna()



# Calculate the range of Data_Value for each Sample_Size


data_range = data.groupby('Sample_Size')['Data_Value'].agg(['min', 'max'])
data_range['range'] = data_range['max'] - data_range['min']
```

```
# Plot the range of Data_Value against Sample_Size
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Sample_Size', y='range', data=data_range.reset_index())
plt.title('Range of Data_Value vs Sample_Size')
plt.xlabel('Sample_Size')
plt.ylabel('Range of Data_Value')
plt.show()


# Plot Data_Value against Sample_Size
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Sample_Size', y='Data_Value', data=data)
plt.title('Data_Value vs Sample_Size')
plt.xlabel('Sample_Size')
plt.ylabel('Data_Value')
plt.show()
```