

AIT-580-Project-Assignment-4

Data Analytics Research Project

Pramath Rajprasad Rao

prajpras@gmu.edu

Section-010

G01483865

**“Nutrition Physical Activity and Obesity – Behavioral Risk Factor
Surveillance System”**

ISSAC GANG

GEORGE MASON UNIVERSITY

ABSTRACT

This dataset includes information about adults' food, physical activity levels, and weight status that was gathered via the Behavioral Risk Factor Surveillance System (BRFSS). The dataset's main goal is to provide information for the Data, Trends, and Maps database of the Division of Nutrition, Physical Activity, and Obesity (DNPAO). It provides information on obesity prevalence, behavioral patterns, and related dietary and physical activity practices at the federal, state, and local levels. The dataset has been organized in a way that makes it easier for researchers, policymakers, and public health officials to monitor trends, and create, and carry out focused initiatives meant to enhance public health outcomes on obesity, physical activity, and diet.

Introduction

In the US, obesity is a serious public health concern since it is linked to several chronic illnesses like diabetes, heart disease, and some types of cancer. Effective public health efforts require an understanding of the behavioral factors that contribute to obesity and related disorders. To fill this gap, the Centers for Disease Control and Prevention (CDC) manage the Behavioral Risk Factor Surveillance System (BRFSS), which gathers information on adult dietary practices, levels of physical activity, and weight status in all 50 states as well as the District of Columbia and three U.S. territories.

This dataset is an essential part of the Data, Trends, and Maps database maintained by the Division of Nutrition, Physical Activity, and Obesity (DNPAO). It offers thorough, useful data that illustrates the continuous obstacles and advancements in the fields of physical activity and nutrition at the federal and state levels. For public health professionals, researchers, and policymakers, this dataset provides insights into the prevalence of obesity, nutritional practices, and physical activity involvement. These stakeholders use the data to track patterns, assess the effectiveness of health programs, and customize actions meant to enhance the health of communities all around the nation.

Every year, the dataset is updated to make sure the information is still accurate and representative of recent developments in the field. It allows for a more sophisticated investigation of how various demographic groups are impacted by and react to public health policies linked to nutrition and physical activity since it incorporates characteristics like age, gender, race, and socioeconomic status.

We aim to extract useful information from the dataset that will help to explain the obesity pandemic as it currently exists and shape future public health policy by using robust statistical approaches and models. It is expected that the data and findings from this study will significantly impact the current discussions about health, lifestyle, and socioeconomic status and how they all relate to obesity.

OBJECTIVE

The main objective of this research is to use the Behavioral Risk Factor Surveillance System (BRFSS) dataset to examine and comprehend the major behavioral and demographic variables that affect obesity, physical inactivity, and the consequences associated with these conditions in adult Americans. The project specifically seeks to answer the following research questions:

State and Territorial Analysis of Physical Inactivity: Determine the average percentage of adults who do not engage in any physical activity during their leisure time in each state and territory to determine which areas have the highest rates of physical inactivity and could be the focus of more public health initiatives.

Education and Obesity Correlation: Examine the relationship that exists between the percentage of adults who are obese and their level of education. This investigation will assist in determining whether lower obesity rates are correlated with greater education levels, which can guide community-based and educational health promotion initiatives.

Gender Differences in Obesity Prevalence: Examine the disparities in obesity prevalence across both genders to customize public health initiatives that target the unique requirements and risk factors of each gender.

Literature Review

State at least 3 well-formulated research questions that could be answered by exploring the dataset.

1 What is the average percentage of adults engaging in no leisure-time physical activity across different states and territories?

We must investigate public health planning such as how many people are engaging in outdoor activities so that it can prevent initial prevention of diseases. Regular activities can help people be active and attentive. Frequent physical activity helps people with many health advantages, including a lower chance of developing chronic conditions like diabetes, obesity, heart disease, and some types of cancer. This can help the government and other organizations so that they can initiate funds to improve the health of public people.

2 Is there a correlation between education level and the prevalence of obesity?

Identifying differences in health outcomes depending on their education level can be made easier by understanding the connection between obesity prevalence and education level. Policies and initiatives targeted at resolving health disparities and advancing health equity can be informed by this knowledge. Early-life prevention initiatives can be informed by the identification of schooling as a potential risk factor for obesity. Health education is something that educational institutions can include in their curriculum to encourage students to adopt healthy habits and avoid obesity.

3 How does the prevalence of obesity differ between genders?

The prevalence of obesity varies by gender, and this information can be used to customize interventions to meet the unique requirements and difficulties that men and women encounter. For instance, we can create specialized programs to encourage healthy behaviors and prevent obesity in a particular demographic if one gender has a greater prevalence of obesity. We can learn about the obesity of genders by patterns and trends where we can bring in many programs and activities to improve the health of public people.

Data Cleaning

	YearStart	YearEnd	LocationAbbr	LocationDesc	\
0	2020	2020	US	National	
1	2014	2014	GU	Guam	
2	2013	2013	US	National	
3	2013	2013	US	National	
4	2015	2015	US	National	
				Datasource	Class \
0				Behavioral Risk Factor Surveillance System	Physical Activity
1				Behavioral Risk Factor Surveillance System	Obesity / Weight Status
2				Behavioral Risk Factor Surveillance System	Obesity / Weight Status
3				Behavioral Risk Factor Surveillance System	Obesity / Weight Status
4				Behavioral Risk Factor Surveillance System	Physical Activity

First, import the data set into Python read the CSV file print the data set, and look for null values and unnecessary columns.

	YearStart	LocationAbbr	LocationDesc	Class	\
0	2020	US	National	Physical Activity	
1	2014	GU	Guam	Obesity / Weight Status	
2	2013	US	National	Obesity / Weight Status	
3	2013	US	National	Obesity / Weight Status	
4	2015	US	National	Physical Activity	
					Topic \
0					Physical Activity - Behavior
1					Obesity / Weight Status
2					Obesity / Weight Status
3					Obesity / Weight Status
4					Physical Activity - Behavior
				Question	Data_Value \
0				Percent of adults who engage in no leisure-tim...	30.6
1				Percent of adults aged 18 years and older who ...	29.3
2				Percent of adults aged 18 years and older who ...	28.8
3				Percent of adults aged 18 years and older who ...	32.7
4				Percent of adults who achieve at least 300 min...	26.6
	Low_Confidence_Limit	High_Confidence_Limit	Sample_Size	...	\
0	29.4	31.8	31255.0	...	
1	25.7	33.3	842.0	...	
2	28.1	29.5	62562.0	...	
3	31.9	33.5	60069.0	...	

Removing all the unnecessary columns Datasource, Data_Value_Unit, Data_Value_Type, Data_Value_Alt, Data_Value_Footnote_Symbol, Data_Value_Footnote , Total, YearEnd from the data set. From this visualization can be made easily.

	Year	LocationAbbr	LocationDesc	Class	\
0	2020	US	National	Physical Activity	
1	2014	GU	Guam	Obesity / Weight Status	
2	2013	US	National	Obesity / Weight Status	
3	2013	US	National	Obesity / Weight Status	
4	2015	US	National	Physical Activity	

Renaming the YearStart as Year because there were two columns in the data set named YearStart and YearEnd so dropped YearEnd and renamed YearStart as Year.

```

Year                0
LocationAbbr        0
LocationDesc         0
Class               0
Topic               0
Question            0
Data_Value          9235
Low_Confidence_Limit 9235
High_Confidence_Limit 9235
Sample_Size         9235
Age(years)          73269
Education            79929
Gender              86589
Income              69939
Race/Ethnicity      66609
GeoLocation         1736
ClassID             0
TopicID             0
QuestionID          0
DataValueTypeID     0
LocationID          0
StratificationCategory1 9
Stratification1     9
StratificationCategoryID1 9
StratificationID1   9
dtype: int64

```

First, check for empty values in the dataset.

	GeoLocation	ClassID	TopicID	QuestionID	DataValueTypeID	\
0	NULL	PA	PA1	Q047	VALUE	
1	(13.444304, 144.793731)	OWS	OWS1	Q036	VALUE	
2	NULL	OWS	OWS1	Q036	VALUE	
3	NULL	OWS	OWS1	Q037	VALUE	
4	NULL	PA	PA1	Q045	VALUE	

Filling the empty values as NULL so the other columns are not affected by removing the entire row.

RESULTS AND ANALYSIS

Research Question 1: What is the average percentage of adults engaging in no leisure-time physical activity across different states and territories?



First, import the CSV file into SQL and show all the tables in the database.

Result Grid

Filter Rows:





Export:

Wrap Cell Content:

LocationDesc	Average
Alabama	30.36866666666665
Alaska	22.37875816993464
Arizona	24.236249999999988
Arkansas	31.662251655629152
California	21.267912772585664
Colorado	19.191900311526467
Connecticut	25.04487179487181
Delaware	27.9190635451505
District of Columbia	21.203973509933775
Florida	27.143150684931495
Georgia	27.143809523809523
Guam	28.784888888888865
Hawaii	21.922115384615388
Idaho	22.782105263157888
Illinois	25.770805369127498
Indiana	28.841401273885346
Iowa	26.3429054054054
Kansas	26.197468354430377

Result 7

The average percentage of adults in each U.S. state and territory who do not participate in leisure-time physical exercise is determined by the SQL query. This information is retrieved from a table called `nutrition_physical_activity_and_obesity`, with a focus on entries that have the question "Percent of adults who engage in no leisure-time physical activity" matching the column. The location description is used to group the results, and the names of the states and territories are arranged alphabetically. This inquiry offers a clear picture of the degrees of physical inactivity in different areas.

Result Grid |  Filter Rows: | Export:  | Wrap Cell Content:  | Fetch rows: 

	LocationDesc	Average
▶	Puerto Rico	43.86931216931214
	Mississippi	33.24029304029306
	Arkansas	31.662251655629152
	Louisiana	30.7138613861386
	Kentucky	30.692333333333323
	Oklahoma	30.457928802589002
	Alabama	30.36866666666665
	Tennessee	30.266779661016958
	West Virginia	30.01872791519432
	Indiana	28.841401273885346

This query shows the top 10 states with average percentages of adults who do not engage in leisure-time physical activity. Puerto Rico has the highest rate (43.86%), meaning that most adults there do not engage in leisure activities. With more than 30% each, Mississippi, Arkansas, and Louisiana come next. The database focuses on obesity, physical activity, and nutrition. It is most likely being used to analyze how physical inactivity is distributed geographically to potentially address public health issues.

Research Question 2: Is there a correlation between education level and the prevalence of obesity?

Stratification1	College graduate	High school graduate	Less than high school	Some college or technical school
Stratification1				
College graduate	1.000000	0.971158	0.883151	0.994806
High school graduate	0.971158	1.000000	0.879299	0.966473
Less than high school	0.883151	0.879299	1.000000	0.899499
Some college or technical school	0.994806	0.966473	0.899499	1.000000

The correlation matrix contrasts various degrees of educational achievement. The correlation coefficient between two sets of educational categories is displayed in each cell of the matrix. These coefficients have a range of -1 to 1, where a perfect negative correlation is represented by a value of -1, a perfect positive correlation by a value of 1, and little to no correlation is suggested by values close to 0.

College graduate and high school graduate: The correlation coefficient, 0.971158, indicates a strong positive link and may suggest that the same factors that affect college graduation also affect high school graduation.

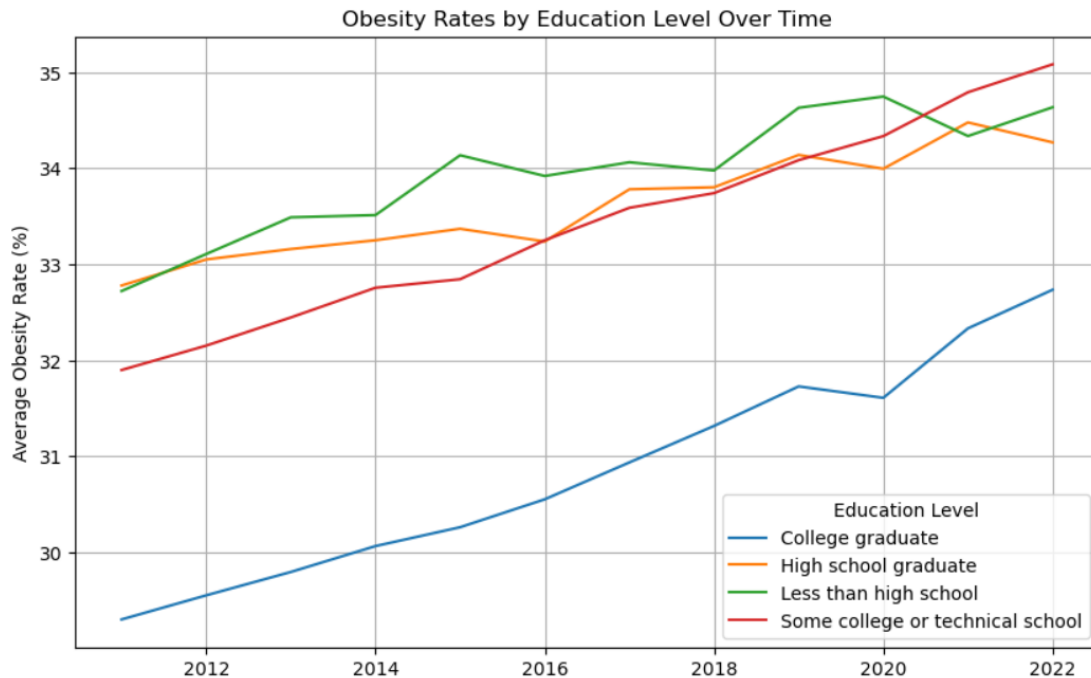
College graduate and less than high school: Although it is marginally smaller than the correlation for high school graduates, the correlation is still quite strong at 0.883151.

College graduate and some college or technical school: The extraordinarily high correlation of 0.994806 may be due to the close relationships between these categories in terms of social characteristics or educational experiences.

High school graduate and less than high school: A strong correlation of 0.879299 may be due to underlying socioeconomic or demographic factors that have a similar impact on both groups.

High school graduate and some college or technical school: A strong correlation of 0.966473 indicates a substantial overlap in the traits or results of these groups.

Less than high school and some college or technical school: The correlation coefficient is 0.899499, suggesting that there is a strong positive association. This is noteworthy, given that the individuals' educational backgrounds differ greatly.



The trends in obesity rates by educational attainment between 2011 and 2022 are shown in the graph. It demonstrates that the highest obesity rates are constantly found in people with "less than high school" education, with rates peaking slightly over 34% by 2022. This suggests a clear association between obesity rates and lower educational attainment. On the other hand, obesity rates among "College graduates" are the lowest; by 2022, they had noticeably climbed but remained below 32%. Over the decade, obesity rates for "High school graduates" and "Some college or technical school" attendance were both intermediate and variable, suggesting a complex link between obesity prevalence and educational attainment. Overall, the graph points to a possible correlation between reduced obesity rates and higher educational attainment.

Research Question 3: How does the prevalence of obesity differ between genders?



The average obesity rates for men and women are displayed in the graph named "Average Obesity Rates by Gender". The red bar indicates the average obesity rate for females, which is 29.46%. The rate is higher for men (32.84%), as indicated by the blue bar.

In this graphic representation, the average obesity rate among men is much greater than that of women in the dataset under analysis. To address the gender disparity in obesity and develop focused interventions, health policymakers, educators, and public health efforts may find this information to be extremely important. The labels directly on the bars emphasize the importance of gender as a determinant in obesity prevalence and offer a clear and quick grasp of the quantifiable differences in obesity rates.

AWS Cloud

npo-dataset

S3 Nutrition_Physical_Activity_and_Obesity_-_Behavioral_Risk_Factor_Surveillance_S...

Run data profile Create project with this dataset Actions JOB DETAILS

Dataset preview Data profile overview Column statistics Data quality rules Data lineage

Dataset preview

Grid Schema Text Tree

#	YearStart	#	YearEnd	ABC LocationAbbr	ABC LocationDesc	ABC Datasource
2020		2020		US	National	Behavioral Risk Factor Surve
2014		2014		GU	Guam	Behavioral Risk Factor Surve
2013		2013		US	National	Behavioral Risk Factor Surve
2013		2013		US	National	Behavioral Risk Factor Surve
2015		2015		US	National	Behavioral Risk Factor Surve
2015		2015		GU	Guam	Behavioral Risk Factor Surve
2012		2012		WY	Wyoming	Behavioral Risk Factor Surve
2012		2012		DC	District of Columbia	Behavioral Risk Factor Surve
2015		2015		PR	Puerto Rico	Behavioral Risk Factor Surve
2011		2011		AL	Alabama	Behavioral Risk Factor Surve
2015		2015		GU	Guam	Behavioral Risk Factor Surve
2015		2015		RI	Rhode Island	Behavioral Risk Factor Surve
2011		2011		US	National	Behavioral Risk Factor Surve
2012		2012		WY	Wyoming	Behavioral Risk Factor Surve

The dataset preview is accessed through a web interface that looks to be a component of an AWS (Amazon Web Services) environment. It is most likely using AWS Glue DataBrew for data preparation and analysis and Amazon S3 for data storage. According to the DataSource column, the "npo-dataset" originates from the "Behavioral Risk Factor Surveillance System".

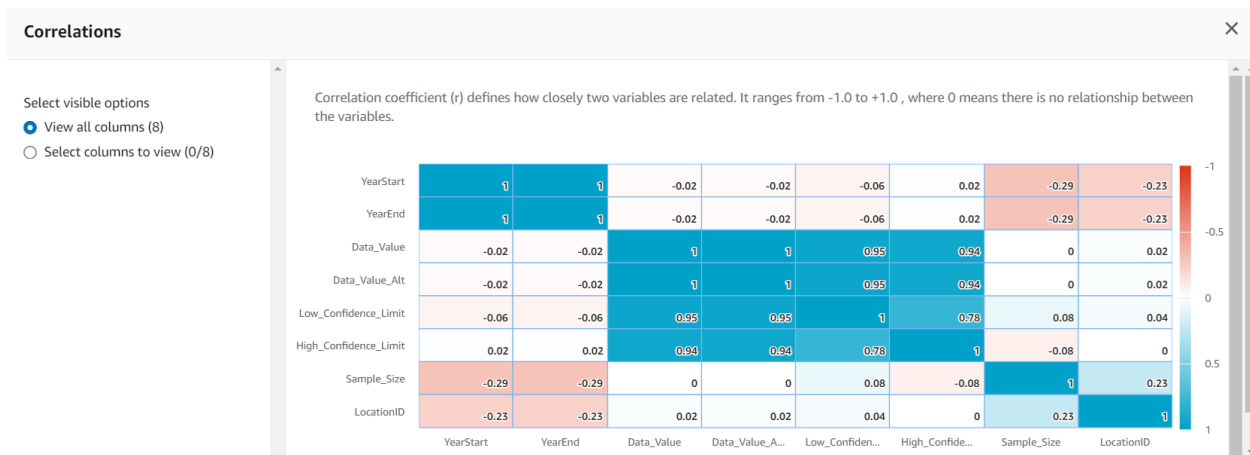
Amazon S3 : The data is stored on S3 (Simple Storage Service). AWS offers a safe and scalable cloud storage solution.

AWS Glue DataBrew: It is clear from the interface options such as "Run data profile" and "Create project with this dataset" that AWS Glue DataBrew is a data preparation tool that enables users to clean and normalize data without writing code.

YearStart and YearEnd: The years that the data was gathered are shown in these columns. The data appear to be annual observations based on the matching start and end years.

LocationAbbr (Location Abbreviation) and LocationDesc (Location Description): These columns include national data (US) as well as the full and abbreviated names of locales, including U.S. states and territories like Guam (GU) and Wyoming (WY). This demonstrates that the dataset contains both compiled national data and in-depth regional data.

Data source: The Behavioral Risk Factor Surveillance System is the source of all records, and it is a reputable source of data for public health research.

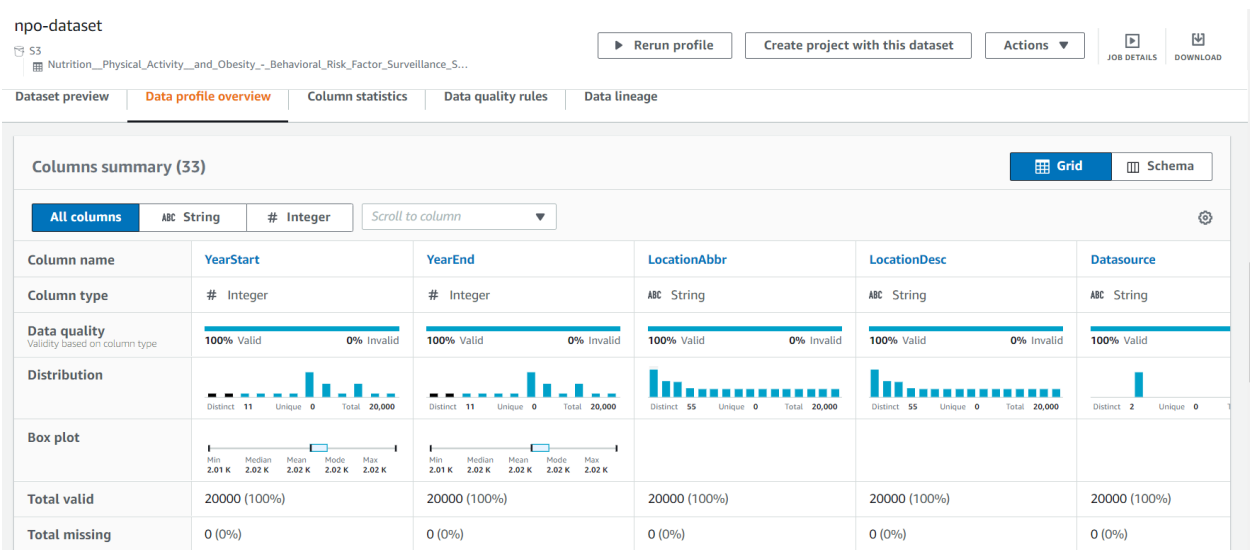


This dashboard, which includes a correlation matrix and a data preview, comes from a data analysis tool that was probably designed with a dataset about obesity, physical activity, and nutrition in mind. Using a color-coded heatmap, the correlation matrix shows the links between different metrics, including YearStart, YearEnd, Data_Value, and others. The variables' strong and weak correlations can be found using this matrix. The data preview section provides a summary of the structure and contents of the dataset by showcasing national and regional data derived from the Behavioral Risk Factor Surveillance System in columns such as YearStart, YearEnd, LocationAbbr, and LocationDesc.

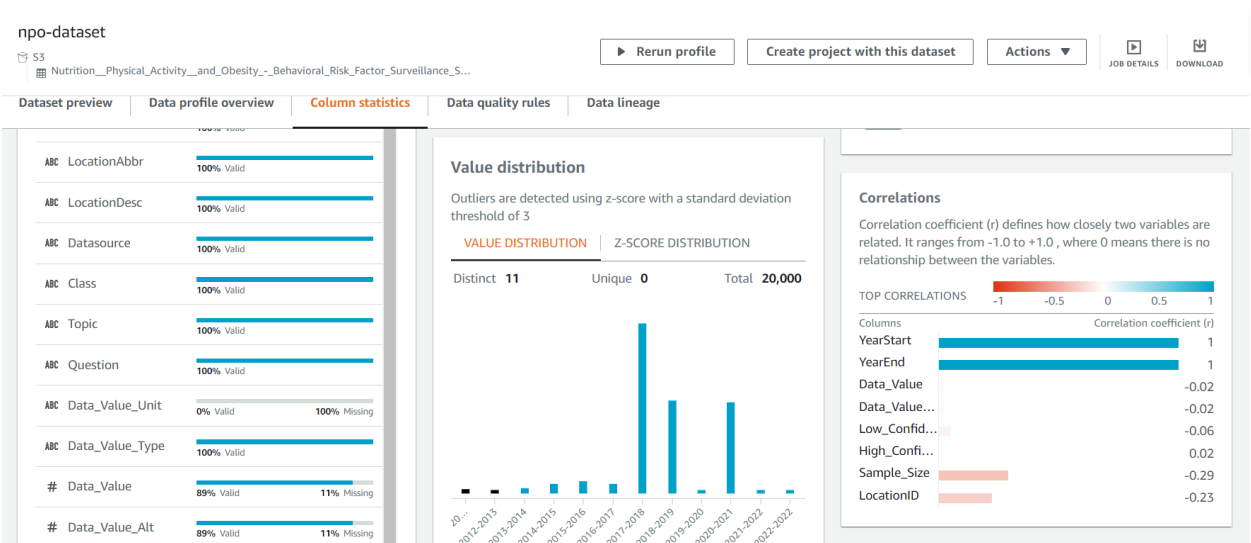


The statistics about obesity can be understood by combining tabular and visual aids. The data profile summary uses a box plot to display normalized distributions for variables like YearStart, YearEnd, and Data_Value, while the box chart illustrates the average obesity rate for females, which is almost 30%. The average obesity rate for each of the 50 states and territories in the United States is shown in a result grid; Puerto Rico has the highest rate, at 43.87%, while Indiana

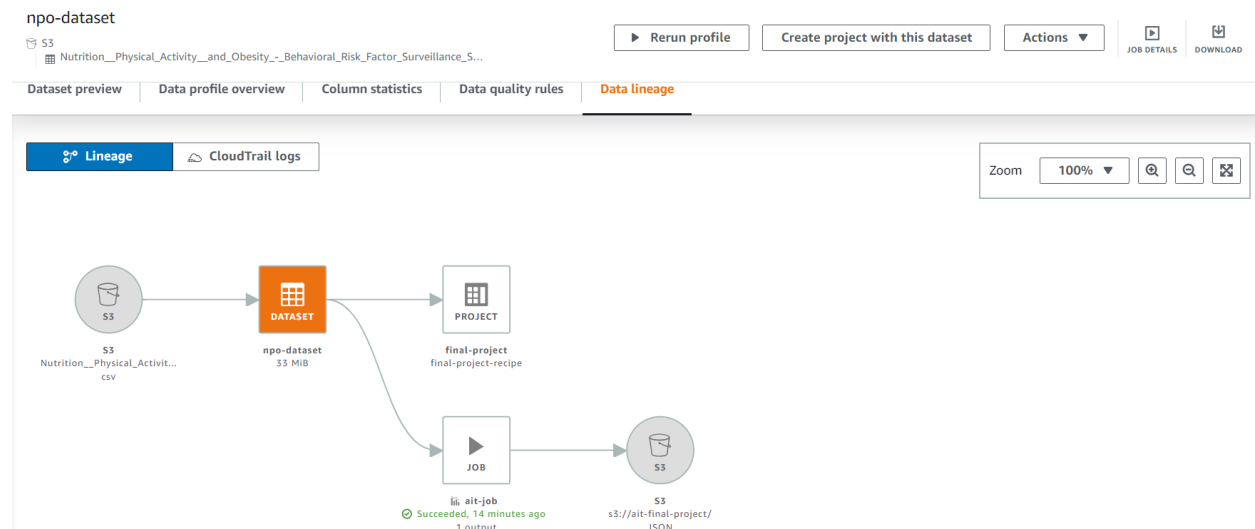
has the lowest, at 28.84%. When combined, these resources make it easier to analyze obesity trends in terms of gender, geography, and other factors.



This analysis explores obesity rates using several visual aids. A line graph that shows different patterns for groups such as college graduates and those with less education than a high school diploma shows how obesity rates have changed over time dependent on educational degrees. The resulting grid highlights regional differences in obesity prevalence by listing all U.S. states and territories along with their respective average obesity rates. The metadata insights provided by the data profile section further validate the accuracy of dataset columns like LocationAbbr, LocationDesc, and Datasource. When combined, these components provide an in-depth understanding of disparities, trends, and data quality related to obesity from a variety of angles.



The completeness and validity of different columns are shown by the column statistics; some, like LocationAbbr and LocationDesc, are fully complete, while others, like Data_Value_Unit, are completely absent. The value distribution histogram highlights the temporal coverage of the dataset by showing how unique data points are distributed over various years. The correlation heatmap displays negative correlations with Sample_Size in addition to substantial positive correlations between YearStart and YearEnd. This thorough study aids in comprehending the distribution, quality, and correlations between the dataset's important variables.



This shows the data lineage from the S3 bucket showing the dataset splitting into the final project lineage and job create ait-job then back to the S3 bucket.

Discussions and Limitations

The Behavioral Risk Factor Surveillance System (BRFSS) provides a dataset that is extremely useful for analyzing national trends in adult obesity, physical activity, and food. It makes data on key health indicators at the state and federal levels analyzed for scholars and policymakers. Due to recall bias and social desirability, the dataset's reliance on self-reported data adds biases including underreporting or overreporting. When comparing results, care must be taken due to sampling variability and possible differences in data-gathering techniques between states. Furthermore, because the dataset covers several years, it's critical to take definitional and survey methodology changes into account when assessing trends over time.

The dataset has significant constraints that affect the scope and precision of the research. First off, non-binary people might not be included in the dataset due to its low gender representation, which limits the scope of thorough gender-related analyses. Analyzing the relationship between education and obesity may contain errors due to the dependence on self-reported educational levels. The primary source of information from observational data is correlation rather than causality, and the dataset may not take confounding factors like socioeconomic position or

access to healthcare into consideration. Furthermore, results may be skewed by missing data points and geographical bias in data collecting; therefore, it is important to handle these gaps carefully to guarantee that the results are genuine and representative.

FUTURE SCOPE

The dataset from the Behavioral Risk Factor Surveillance System (BRFSS), focusing on nutrition, physical activity, and weight status, provides a useful resource for public health research and policymaking. It provides a multitude of perspectives for comprehending the complex interaction between lifestyle decisions and health consequences because of its extensive national and state-specific data. This information can be used by researchers to identify long-term patterns and differences in obesity and physical activity across different demographic groups, which will help them design more focused interventions. Its longitudinal design also makes it possible to assess the success of public health campaigns over time, which aids in the improvement of obesity-fighting tactics. This dataset can be updated frequently and mined for insights into changing public health issues as new factors and patterns appear.

The prevalence of adult inactivity at the regional level. Researchers can discover places with greater rates of inactivity by using the dataset to identify geographical patterns in physical activity levels. This information can help direct local health campaigns to encourage active living and assist in the efficient use of resources to discourage sedentary behavior.

Examining the correlation between education level and obesity prevalence can provide insight into the differences in health outcomes between socioeconomic groups. Education frequently has an impact on finances, lifestyle decisions, and access to knowledge that affects exercise and nutrition routines. We can determine whether and how education-based interventions might reduce obesity by evaluating the dataset and determining how education affects obesity rates.

Understanding the variations in obesity incidence between genders can shed light on biological, behavioral, and cultural elements that impact well-being. The dataset can identify notable trends that may indicate difficulties each gender has managing their weight by comparing the obesity rates of men and women. Given that men and women have different requirements and risk factors for obesity, these findings can help shape interventions that are specifically designed to address the issue.

Conclusion

This dataset provides thorough insights into the complex links between gender, education, and physical activity in obesity prevalence in the United States. The necessity for region-specific treatments is highlighted by the differences in leisure-time physical activity between states and territories. The relationship between obesity and education highlights the role that education plays in health outcomes, and the gender variations in obesity prevalence need the development of customized treatments for each group. These insights can be used by educators, policymakers, and medical experts to create focused initiatives that lower obesity rates and enhance public health.

The dataset analysis indicates that a considerable number of adults in several states and territories do not participate in leisure-time physical activity, which calls for region-specific health interventions. Furthermore, there is a strong association between the prevalence of obesity and educational attainment, with higher rates of obesity being associated with lower levels of education. These data suggest that educational interventions may play a critical role in the prevention of obesity. Lastly, the data reveals gender-specific disparities in obesity rates, with variances in prevalence between males and females across demographic categories and geographies. These disparities call for the development of sophisticated, gender-targeted health initiatives to successfully address them.

REFERENCES

[1]

“Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System - Data.gov,” *Data.gov*, 2019. <https://catalog.data.gov/dataset/nutrition-physical-activity-and-obesity-behavioral-risk-factor-surveillance-system>

[2]

Centers for Disease Control and Prevention, “Adult Physical Inactivity Prevalence Maps by Race/Ethnicity,” *Centers for Disease Control and Prevention*, Jan. 16, 2020. <https://www.cdc.gov/physicalactivity/data/inactivity-prevalence-maps/index.html>

[3]

Centers for Disease Control and Prevention, “Adult Obesity Prevalence Maps,” *Centers for Disease Control and Prevention*, Sep. 21, 2023. <https://www.cdc.gov/obesity/data/prevalence-maps.html>

[4]

D. W. Brock, O. Thomas, C. D. Cowan, D. B. Allison, G. A. Gaesser, and G. R. Hunter, “Association Between Insufficiently Physically Active and the Prevalence of Obesity in the United States,” *Journal of Physical Activity and Health*, vol. 6, no. 1, pp. 1–5, Jan. 2009, doi: <https://doi.org/10.1123/jpah.6.1.1>.

[5]

R. Kanter and B. Caballero, “Global Gender Disparities in Obesity: A Review,” *Advances in Nutrition*, vol. 3, no. 4, pp. 491–498, Jul. 2012, doi: <https://doi.org/10.3945/an.112.002063>.