Capstone Data 205 Final Report

Project Plan

My Capstone project was to analyze the public health trends in the United States, primarily focusing on the obesity epidemic that is plaguing our nation. My project was to analyze historical trends over the past two decades and provide statistical analysis that would give public healthcare policy makers and medical professional tools to combat obesity crisis. The National Health and Nutrition Examination Survey or known as NHANES dataset, was the bedrock of my capstone project. In addition, I also used the BRFSS as a supplement to my analysis beyond 2018. It wanted to continue the analysis of the obesity epidemic from 2018 till 2022. Both the NHANES and the BRFSS surveys are observation studies conducted by the Center for Disease Control and Prevention (CDC). The NHANES is conducted bi-annually and the BRFSS is conducted yearly. NHANES dataset contains demographic, physical examination, nutritional, laboratory, and interview questionnaires. BRFSS is a telephone survey that contains health- related behaviors and risk factors amongst adults such as tobacco use, alcohol consumption diet and nutrition. Managing large datasets from NHANES and BRFSS datasets, I used Python for initial exploratory data analysis (EDA) and created an interactive dashboard using Power BI. Python was used for data cleaning, manipulation, exploration, and visualization. With libraries like Pandas, NumPy, Matplotlib, Seaborn, and Scripy.Stat allowed me to handle the dataset complexity. Power BI is a great tool that makes complex coding of python and turns it into a user-friendly interface. Although it can do all the data exploration analysis like python but without coding knowledge. Power BI was used as an interactive dashboard as my capstone product. The interactive dashboard allows users to explore data in real time. The project goals were showing a holistic view of the health trends in the United States by extracting insights from the NHANES dataset specifically emphasizing the obesity epidemic. It was to see the obesity trends such as BMI and its relationship with other health metrics such as blood pressure or cholesterol and illustrate the roles that social economic levels play with BMI. Since the NHANES dataset is limited to 2018 due to the pandemic I chose the BRFSS as a companion dataset. The BRFSS data set I included where another piece of data focused on sleeping habits, mental health, obesity levels, and access to healthcare per state. BRFSS is not rigorous as the NHANES because the BRFSS is based on self-reporting information. To demonstrate the obesity epidemic, I would have to conduct several statistical analyses and visual representation of the data to highlight the obesity epidemic. The impact of this project was to provide valuable insight to researchers, policymakers, and health professionals so they could address the obesity epidemic. The dashboard was to provide interactive was to highlight the states with the most obesity levels and its impact of mental health and accessibility to health care to its their integrate role with each other.

EDA

My Data Exploration Analysis for the NHANES and BRFSS datasets were done in python. First gathering all the datasets from 1999-2018 from the NHANES which was in SAS files. The data had many columns that were the names were changed, missing, or merged into another column, trying to dig through each dataset took tremendous amount of time and cumbersome. Once you find the pattern on how to find each dataset and their columns of interest the data wrangling process becomes easy. The data sets for the NHANES were complicated and demanded labor intensive data wrangling. First had to load the data from the official website from NHANES and create a function that would allow the dataset to be from the data dictionary, everything needed to be decoded and the information was scattered everywhere. I was able to sequence the datasets from demographic, dietary, examination, laboratory, questionnaire datasets together for each year and participants. I was able to combine each dataset and concatenate them into a large data frame. After forming the large data frame, I created a column of interest leading with the SEQN which is a unique identifier that corresponds to a participant in that survey year. The SEQN identifier connected all the columns that were an interest to me for example SDDSRVYR (year of the survey), RIAGENDR (Age during time of examination), etc. The data was encoded numerically. I had to decode the numerical data which each code was defined in the data dictionary. I took a numerical code then I decoded it into a readable string. Then I used a .map function in pandas which grabbed the numerical codes and turned them into new categorical assigned values. The dataset also included missing value for numerical codes. I used np.nan for the numerical codes so they could know in the dataset to be strings. I did combine certain datasets together for example DMDBORN refers to a demographic question about participants place of birth to determine if they were natural born citizens or they were born in a different county. DMDBORN was missing for certain years because the column names changed to DMDBORN2 or DMDBORN4. So, I combined them all into DMDBORN.

Descriptive Statistics

One of the initial statistical analyses I conducted was to explore the relationship between BMI (body mass index) and ethnic race. I created a line graph that would illustrate the average BMI amongst different racial groups from 1999 to 2018. To further understand the relationship, I created a bar graph that would give me a comparison between BMI, gender, and different ethnic groups. As to deepen my understanding I created another bar graph comparing the relationship between males and females. Through the data you can see the impact of obesity as the avg BMI was in the range of overweight. The data that these graphs were displaying clearly showed how Americans were becoming overweight. Obesity not only affects us physically, but it was also so as an impact on our cardiovascular system. I wanted to see the relationship between obesity and our blood pressure. NHANES data provided physical examination data with systolic blood pressure and diastolic pressure. Through the data I was able to create a line chart by different races throughout the years and see the avg blood pressure was rising as BMI increased. Another visualization I wanted to create was that the percentage of the population in the data set was overweight. Since NHANES has over 5000 people that they survey it’s a good data set because the sample size is large enough. In this line graph you’re able to see the trends of obesity that is rising. In 2018 41% of the population had a BMI > 30 and 9% of the population had a BMI > 40. I also created a heatmap showing the intersection between BMI, race, and income levels. I wanted to show that obesity is related to a social economic factor. Certain ethnic groups are more prone to be obese as the income rise for example African Americans. As the income rose you can also see Mexican Americans BMI also rose as income levels increased. I did various laboratory tests as well such as cholesterol, HDL, and LDL. Then I did a chi square test that categorized different BMI categories with labels of underweight, normal, overweight, and obese. That test came out to be statistically significant as well. One of the statistical analyses that confirmed my suspicion was the ANOVA test between different racial groups which showed itself to be statistically significant. Then I did a post-hoc Tukey HSD test to compare the BMI between different races to each other.

[Data Product](https://github.com/asingh2024/DATA_205/blob/main/NHANES/projectfiles/Obesity%20Dashboard%20BRFSS%202018-2022.pbix)

My final data product took the form of an interactive Power BI dashboard. There was a gap in the NHANES dataset from 2018-2022. To address this gap so I used the BRFSS data set to fill in the missing years. The dashboard offered a multi-dimensional view of key health indicators. This included the BMI distribution, obesity rates, mental health status, uninsured population, and average sleep patterns all segmented by state and year. The dashboard featured nine interactive tiles that interlinked with each other to provide a holistic picture of the health status of the United States. One of the primary tiles presented a map of the United States that color coded BMI classifications and another tile showed the percentage of each racial group obesity levels in that state. I had a mental health tile that provided insight into the mental health status of each state. Another tile that also what percent of the population was uninsured. This interactive dashboard enhances our understanding of health indicators that are plaguing each state. It helps the public policy maker in their perspective states. Obesity Dashboard in Power BI offers a visual representation of critical health data in real-time, eliminating the need for complex Python coding. Its interactivity allows for the exploration of various datapoints, including obesity rates, uninsured percentage, and average sleep duration. Its significance lies in its ability to illuminate the interconnectedness of obesity with other health factors. Highlighting the obesity epidemic’s this dashboard provides importance of addressing this multifaceted issue. For example, individuals without healthcare access may neglect seeking medical treatment, resulting in unhealthy coping mechanisms that may result in further downhill pathologies of obesity. I believe this product achieved its intended goal. This dashboard not only presents statistics but also adds a layer of visualization that enhances our understanding of complex health relationships. The code behind my capstone project is reproduceable, meaning that if someone wished to replicate the dashboard, they could. I recognize the importance of reproducibility and to make sure that the data is valid. I believe there is a potential to further develop this tool and possibly offer it to healthcare organizations. The dashboard serves as an excellent resource for the public delivering straightforward factual insights. The dashboard provides immediate information on various health indicators by state, for example percentage of overweight individuals, mental health status, sleep patterns, and access to healthcare. This makes it a powerful tool at your fingertips.

[Data story](https://github.com/asingh2024/DATA_205/blob/main/capstone%20project%202.pptx)

Technology has been transformative affecting every aspect of our lives. Technologies have made our lives easier and more enjoyable but ever since the latter half of the 20th century and epidemic has been slowly rearing its ugly head. Obesity has been on an upward trend because our lifestyle has been increasingly sedentary. The ramification of obesity extends beyond our physical appearance. Obesity is a multi-faceted issue that is influenced by a variety of factors such as environmental, behavioral, genetic, social, and economic. Obesity has a profound impact on individuals and communities and there should be proactive measures to combat this crisis that is plaguing our nation. Obesity is full of weight of emotions. Obesity has many health implications, for example increase in type 2 diabetes, heart disease, psychological problems, stroke which all lead to increase morbidity and mortality. Living large dude to our lifestyle and dietary choices. Sedentary lives and overconsumption of high caloric and processed foods results in weight gain. Sedentary lifestyles have become common place due to digital screens dominating our time, from office desk to digital entertainment. Digital products such as video games, streaming platforms, encouraging indoors and prolonged periods of inactivity. Obesity is also a drain on economy and society. Health care costs associated with obesity related illnesses such as chronic conditions like diabetes, heart disease, stroke and certain cancers are exacerbated by obesity. Another problem with obesity is its stigma and psychological repercussions that lead to social isolation which further reinforcing self-esteem and body image problems. What is BMI and how to measure obesity? BMI is a tool to measure obesity. It is calculated by dividing weight in kilograms by square of height in meters. The formula is BMI = weight (kg) / height(m) ^ 2. Obesity can be categorized in 6 different categories. BMI below 18.5 falls in to the “Underweight” category. “Normal Weight” BMI is between 18.5 to 24.9. “Overweight” individuals have BMI between 25 to 29.9. “Obese” is referred to as to any BMI over 30.

The NHANES dataset from 1999 to 2018 provides concrete evidence that obesity is not a matter of perception but there is substantial factual data. NHANES is critical for public health planning, policy development, research, and gives health practitioners a gameplan to combat obesity. The heavy truth about obesity is that 41% of the adult population is obese. Out of that nearly 9% of individuals with BMI > 40%. Another problem that is facing obesity in the children of our nation. Close to 40% of children ages 2-19 were classified as obese according to 2018 NHANES dataset. Obesity exhibits variations among different ethnic groups and gender. According to the NHANES dataset females were more suspectable to obesity than males. African Americans experienced higher obesity rates compared to other ethnic groups, Mexican, White, other Hispanic groups weren’t far beyond also faced significant obesity rates. Mixed racial groups were slightly overweight that could be attributed to genetic interplay resulting in healthier offspring driven by expression of dominant genes from different gene pools. The intersection of obesity, race and income demonstrated a noteworthy impact. Among African Americans and interesting trend emerges in the data. As their income increased, so did their obesity rates. The threshold for obesity for African Americans with BMI over 30 appeared to be income levels ranging form 45k to 50k and notably those that were earning over 75k and 100k and beyond were classified as obese. In contrast, multiracial groups displayed a relatively lower BMI levels hovering around 25 or slightly below 25. One of the consequences of obesity is the impact it has on the cardiovascular system. You can clearly see the impact it has on systolic and diastolic blood pressures. Normal blood pressure is less than 120/80. People who are obese, the avg blood pressure trend was increasing as the years went on causing concern causing the heart to pumper harder dude to increase blood pressure. Cholesterol also plays a huge role in obesity causing another effect on our cardiovascular system by affecting the arteries flow of blood. The NHANES dataset showed an increase in cholesterol levels. Doing hypothesis testing such as ANOVA to find difference between the MEAN BMI with different ethnic groups was important. The p-value was extremely low. To devel deeper and find which ethnic groups were being impacted, I did a post-hoc test. Tukey HSD test to confirm my suspicions. African Americans and Mexican Americans had a greater mean difference then another ethnic groups. The Chi-square test was done to show the significant association between BMI categories. Conducting hypothesis testing such as ANOVA to examine difference in mean BMI across various ethnic group was pivotal step in my analysis. The p-value was extremely low, indicating that there were significant differences among the ethnic groups. To delve deeper into these disparities and pinpoint which ethnic groups were affected I did a post hoc test. The Tukey HSD test validated my initial suspicions. During the test results revealed that the African American and Mexican Americans exhibited more mean difference compared to other groups. In addition, Chi-square test was performed to see the relationship between BMI categories.

Overall Experience

My overall experience with the capstone project was meaningful and allowed me to grow. When I took the math 117, I knew about the data science certificate, but I wasn’t sure about it and I am glad I finished it. Starting the data science certificate helped me develop critical thinking skills and gave me superpowers that I didn’t have before. Now I can understand the impact of data science. I was pushed to the limits where I have a firm understanding of data science. My experience with the NHANES dataset was challenging, particularly when it came to data wrangling, which consumed many weeks of effort. However, this challenge pushed me to the limits and in the end was rewarding. It teaches me valuable skills of handling complex and messy datasets. Initially I was frustrated but towards the end I became more confident, and I can tackle any dataset in the future. The learning curve for Python and Power BI was rough. Learning these tools broadens my skillset. Python demanded most of my time and energy but now it’s in my toolkit for data analysis. Reflecting on my journey, I should have recognized my strong suit was R utilizing it would have saved me considerable amount of time and frustration. Unfortunately, that wasn’t the case. I am glad I stuck it out with Python. Overall, my capstone project has been a great experience and provided a strong foundation for future projects and my career aspirations. I can’t wait till I dive into computer science and health care research; it feels like I have acquired a set of superpowers through my data science certification journey. With my keyboard as my tools and the concepts I have gained I am ready to build the future.

Recommendations

The obesity dashboard has the potential to serve as a real-time resource accessible to a wide range of users, including journalists, researchers, healthcare professionals, policymakers, providing valuable insights. Government organizations, nonprofits, and for profit can use health dashboards to understand and address obesity related issues. To enhance this the CDC could unify all the datasets into a large data frame or dataset. This would ensure that the column names are consistent across survey years. This would minimize combing through the data dictionary, spending hours upon hours trying to search for the columns. An integrated dataset that connects several databases such as the BRFSS and NHANES and other health related statistics could give an holistic approach and better decision making capabilities for battling obesity.

Acknowledgement

I want to express my appreciation to the faculty at Montgomery College for their guidance and support throughout my data science journey. Firstly, I am thankful to Prof. Alaree of Math 117, whose encouragement was pivotal in my decision to pursue a data science certificate. Prof. R. Saidi’s passion for Data 110 and Math 217 brought life to visual storytelling and statistics. I am grateful to Prof. Mohammad Adbirisak for his patience and expertise in teaching Python and machine learning. Although this semester presented many challenges, Prof. Lapalucci’s not only taught me Data 101 but also was mentor during my Capstone Project in Data 205. This certificate program has been transformative, filling the gaps left by various online courses like Dataquest, Datacamp, Udemy, Coursera. The structed learning environment, the motivation to perform and accountability of earning grades pushed me to succeed. I strongly recommend the Data Science Certificate program to students to deepen their understanding in the increasingly data-driven world. I want to thank my Mentors Dr. Ami Ostchega (NIH) and Dr. Wilfred Kokou Letsou (Sr. Data Scientist) who guided me and supported me. Dr. Ami Ostchega to help me with NHANES dataset and Dr. Wilfred Kokou Letsou whose knowledge data science is greatly appreciated. Thank you!