Introduction

This project marks the culmination of the Capstone course in the "Python for Everybody" specialization offered by the University of Michigan on Coursera, under the guidance of Professor Severance. The challenge was to select a dataset of interest, perform a series of data processing steps – including extraction, reading, parsing, cleaning, and visualization – and derive meaningful insights.

I chose the "Nutrition, Physical Activity, and Obesity – Behavioral Risk Factor Surveillance System" dataset from data.gov, spanning from 2011 to 2021. This dataset offers extensive information on adult diet, physical activity, and weight status, essential for the DNPAO's Data, Trends, and Maps database. My interest in this dataset stems from my background as a dietitian and my current pursuit of a career in data analytics, with a focus on nutrition, medicine, healthcare, wellness, and fitness. This project seemed like an excellent opportunity to align my professional expertise with newfound analytical skills.

A particular aspect that piqued my interest was the relationship between income levels and physical activity. I hypothesized a positive correlation: higher income levels would correspond to increased physical activity, possibly due to greater access to exercise facilities and resources.

Analysis

Although we had learned SQL in the course, I found Pandas to be a more efficient tool for data manipulation, requiring fewer lines of code to achieve similar outcomes. The original dataset contained over 88,000 records. Post-cleaning, this number was reduced to 9,700 relevant entries. The visualization of these cleaned records revealed some expected and some surprising patterns.

Consistent with my hypothesis, individuals with an income of $75,000 or higher were more physically active compared to lower income groups. However, the data also revealed a startling trend: across all income levels, the percentage of individuals engaged in adequate physical activity was lower than anticipated. In none of the categories did the percentage of sufficiently active individuals exceed 50%, suggesting that irrespective of income, a significant portion of the population falls short of the recommended 150 minutes of moderate-intensity activity and two days of muscle-strengthening activity weekly.

Conclusion

While income level does impact physical activity, it's clear that it's not the sole determinant. Time constraints, job demands, and lifestyle choice could also all play a role in individuals being unable to achieve the recommended activity level. Even among higher earners, only a marginal majority meet the physical activity guidelines. This leads me to theorize that factors such as time availability, job nature, environmental aspects, and educational background might significantly influence physical activity levels.

For future analysis, a comparative study between education, income, and physical activity levels could be enlightening. It would be interesting to explore whether there's a stronger correlation between higher education and physical activity.

Reflecting on my journey, I am immensely proud of what I've accomplished in just three months of learning Python. This project is not only a testament to my growth as a Python programmer but also an exciting step towards integrating these skills into my professional domain.

Potential Improvements

Cosmetic Enhancements:

Adding percentages within or atop each bar in the graphs for clearer data representation.

Ensuring uniformity in the x-axis labels, particularly the inconsistent appearance of the dollar sign.

Rearranging the "Less than $15,000" category for improved visual flow in the graphs.

Exploring more visually appealing color schemes and designs to enhance the overall aesthetic of the graphs.