Correlation Between Strength Training and Income in the United States

Matthew Irvin

Western Governors University

Table of Contents

Project Overview

Research Question

Is there a relationship between strength-training habits and income levels in the United States

Scope of the Project

This project focused on discovering whether consistent participation (muscle-strengthening exercises two or more times a week) was related to higher earnings. The dataset used in my analysis is national data from the CDC's Behavioral Risk Factor Surveillance System (BRFSS). Each data point summarized thousands of adults, who were grouped by income, giving a reliable snapshot of how workout habits vary across income levels.

Overview of the Solution

The primary environment used for data cleaning, filtering, statistical testing, and visualization was Python, accessed through Jupyter Notebook. The workflow utilized well-known libraries, such as Pandas and NumPy, for the cleaning and processing of numeric data. Seaborn and Matplotlib were used for creating visuals, and Scipy and Statsmodels for correlation and regression tests.

Methods:

Pearson correlation and simple linear regression were used to demonstrate the relationship between income and participation in strength training. I ran additional correlation and regression tests (Spearman and robust HC3) to ensure my results were consistent and not influenced by outliers.

Data Scope:

Income ranges were consolidated into numeric midpoints to allow quantitative modeling of the relationship. For example, participants making less than $15,000 were assigned a midpoint value, while those making $75,000 or more were placed in another midpoint.

The analysis shows a strong positive link between income and weight training frequency, while physical inactivity moves in the opposite direction. The results suggest that disciplined exercise habits are related to greater earning potential.

B. Project Execution

Project plan

The plan outlined in Task 2 stayed consistent during the execution. I collected the CDC dataset from Kaggle, isolated the strength-training question, ran correlation and regression tests, and then created visualizations to show the results. I followed the workflow from task 2 closely, and it translated smoothly into practice; however, there was a difference in my projected milestone completion days, as some tasks, such as writing the final summary report and organizing visuals, took a lot longer than I anticipated.

Project Planning Methodology

The project followed a simplified CRISP-DM framework:

Each stage of the framework was represented in some form:

Business Understanding- Defining the question: “Is there a relationship between strength-training habits and income levels in the United States?”

Data Understanding- I collected and explored the CDC dataset from Kaggle, addressing any anomalies or quality concerns.

Data Preparation- I loaded the CDC dataset into Jupyter Notebook and cleaned, transformed, and organized the data using Python libraries such as Pandas and NumPy. This ensured the dataset is accurate and ready for statistical analysis.

Modeling- I used the Pearson and Spearman correlation tests, as well as a simple regression model, to measure the strength and direction of the relationship between strength-training frequency and income.

Evaluation- I reviewed the results of the analysis to determine if the statistical relationships are significant and useful. The final results were summarized through visuals such as scatterplots and correlation graphs to provide additional validation that the results align with the project's objectives.

Deployment- The results are summarized with visuals such as scatterplots and correlation graphs. The analytical work was completed in Jupyter Notebook and then formally summarized in a Panopto presentation to demonstrate the workflow and results.

Project timeline and milestones

Milestone	Duration (days)	Start Date	End Date
Deliverable 1.1.1 – Identify/download CDC BRFSS dataset	2	Oct 21 2025	Oct 22 2025
Deliverable 1.1.2 – Explore/verify dataset structure	2	Oct 23 2025	Oct 24 2025
Deliverable 1.2.1 – Clean and prepare data in Python	2	Oct 25 2025	Oct 26 2025
Deliverable 1.2.2 – Perform data profiling for quality checks	1	Oct 27 2025	Oct 27 2025
Deliverable 1.3.1 – Conduct Pearson/Spearman correlation tests	2	Oct 28 2025	Oct 29 2025
Deliverable 1.3.2 – Regression analysis/summarize results	2	Oct 30 2025	Nov 1 2025
Deliverable 1.4.1 – Create visualizations	2	Nov 2 2025	Nov 3 2025
Deliverable 1.5.1 – Write summary report interpreting visuals and statistics	6	Nov 3 2025	Nov 8 2025
Deliverable 1.5.2 – Record/upload Panopto presentation	2	Nov 8 2025	Nov 9 2025

Note: Some milestones, such as data cleaning, went faster than I originally anticipated, but some tasks, such as writing the summary report, took a lot longer to complete. The report required more time to organize findings, interpret results, and make sure everything was presented clearly and concisely.

C. Data Collection process

Data Source

- The data was collected by downloading a CSV file from Kaggle, with the origin of the data coming from the CDC's annual BRFSS survey. The collection process matched the plan from Task 2 with no unexpected obstacles. The dataset itself included annual self-reported health behaviors, which included the frequency of strength-training and income categories.

Obstacles

- My main struggle with this dataset is trying to navigate its large size and complex structure since it was much larger than I have worked with before. Some column names were unclear and seemed to serve the same purpose; for example, there was a column for income and a column named "stratification" that seemed to contain the same exact values, so I had to verify I was working with the correct one. To solve this issue, I had to dig in and refer to the CDC codebook for each variable's meaning to verify I was using the correct data. I also noticed some missing responses in both income and physical-activity fields, which is pretty normal for a survey of this magnitude, so I dropped incomplete records and recoded the income ranges into numeric midpoint values so the correlation tests would work accurately. Once I fixed these issues, the data became much easier to manage and analyze.

Governance & Security

- No governance or security issues arose during my analysis, and everything stayed compliant with the CDC's public-use license. No personally identifiable information (PII) was used in the data, as the CDC removes all names, contact information, and location identifiers before it is released.

C.1 Advantages and Limitations of the Dataset

Advantages:

- The dataset was big and of high quality, and it is a nationally representative sample collected by the CDC, a very trustworthy source.
- Fully anonymized, free, and safe for educational purposes
- The survey is current and updated annually, and uses reliable collection methods, which aid in keeping the results accurate and comparable over time.

Limitations:

- There is a chance of bias or inaccurate survey inputs since the responses are self-reported.
- Observational data, meaning it can correlate but not prove causation.
- Limited precision due to income being grouped into ranges instead of exact values, for example, one of the income fields reads “less than $15,000”, which is a range.

D. Data Extraction and Preparation

The CDC dataset CSV file was loaded into Jupyter Notebook using Pandas, where I filtered for relevant rows, including income and strength-training questions in this scenario, and dropped missing and unrelated entries. The income values were consolidated into numeric midpoints to allow quantitative modeling. The data was then cleaned for empty values and duplicates and then sorted by income brackets to prepare for analysis.

Since the dataset was already well-organized and did not require complicated extraction from multiple sources, these steps were appropriate. My main goal was to make sure the income and exercise data were usable in quantitative models, and using Pandas, I was able to efficiently handle that while keeping the process clean and repeatable.

E. Data Analysis Process

E.1 Data analysis Methods

Main analytical methods used:

1. Pearson correlation: Measured linear relationships between income and frequency of training.
2. Spearman correlation: Verified rank-based relationships and checked for non-linearity.
3. Simple linear regression: Quantified how much training participation changes across income levels.

These methods were used because they provide straightforward numerical measures of association that are suitable for two continuous variables.

E.2 Advantages and Limitations of Tools and Techniques

Advantages: Utilizing Python's open-source libraries (Pandas, SciPy, Statsmodels, etc) made it easy to explore patterns and verify statistical relationships. Seaborn and Matplotlib provided a simple way to turn raw data into clean, professional visuals. Jupyter Notebook was especially beneficial because I could test and refine my code one step at a time.

Limitations: Linear regression methods assume the data points are independent, which may be hard to guarantee in survey-based data. The income variable had only a few income categories, resulting in smaller samples within each group. However, the group sizes were still sufficient for meaningful correlation insights.

E.3 Application of Analytical Methods

I began the analysis with simple descriptive statistics and checked for missing data to understand the overall shape of the dataset and spot any obvious errors. Next, I tested the relationship between income midpoints and the percent of adults reporting strength-training participation using Pearson and Spearman correlations.

The correlation results were as follows:

Pearson = 0.991, $p < 0.001$

Spearman $\rho = 1.000$, $p = 0$

Both test results showed an extremely positive correlation, meaning that the data suggest that adults who strength train more often also fall into higher income ranges.

An ordinary least squares (OLS) regression model was then used to measure how much income could explain changes in strength-training frequency.

$R^2 = 0.994$

$\beta_1 = 0.0001$

$p < 0.001$

These results show a near-perfect fit, meaning income accounted for around 99% of the differences in strength-training rates across income groups. The positive slope ($\beta_1 = 0.0001$) reveals that even when income went up slightly, strength training participation also rose.

I also checked the scatterplot and residual pattern to verify that the relationship was strongly linear and that no single point appeared to skew the results.

F. Data Analysis Results

F1. Statistical Significance

The results of the analysis were a success and support the **alternative hypothesis**: Individuals who regularly participate in strength-training exercises are more likely to have higher income levels than those who do not engage in strength-training exercises. The theory proposes that regular exercise builds discipline, focus, and goal-oriented behavior; traits that may contribute to stronger professional performance and long-term financial success.

The Pearson and Spearman correlations were very significant at $p < 0.001$, and the regression slope was positive, which means that as income goes up, strength-training participation also goes up. A follow-up secondary regression was used to show the opposite trend, and income had a negative correlation, $r = -0.991$, $p < 0.001$, with inactivity. In simple terms, this just means when

income goes up, inactivity goes down, meaning higher-income individuals tend to stay more active. The results in Figure 1 verify that as income rises, strength training participation also rises, while inactivity decreases.

Figure 1: Strength Training vs Income Level

This is a scatterplot with a regression line, which shows a strong positive relationship between average income per income bracket and the percent of adults who engage in strength training two or more times per week.

Figure 2: Dual-Line Chart: Physical Activity Patterns Across Income Levels

This dual-line chart illustrates the relationship between income and exercise frequency.
Red line = People who reported strength training two or more times per week
Blue line = People who did not report exercising at all

F2. Practical Significance

The practical significance of the results suggests that strength training tends to be associated with higher income groups; however, this does not imply causation. It does shed light on the fact that people who maintain structured health routines may also exhibit traits that contribute to financial success. Positive benefits from strength training, such as motivation, confidence, and discipline, may carry over into the workplace and provide a valuable insight for organizations to use in boosting productivity by implementing wellness programs that encourage regular exercise.

F3. Overall Success

The project was overall a huge success, meeting every objective set and producing interpretable results supported by statistics and visuals. There were no major workflow interruptions, and the tools used were efficient and effective throughout the process. The results were clear and directly answered the research question, showing a distinct, strong, and consistent relationship between income and strength-training frequency. The project improved my confidence in building a full data analysis project from start to finish and sharpened my Python skills for real-world analysis. Overall, I am satisfied with the results and the outcome aligned perfectly with the project and rubric expectations.

G. Conclusion

G.1 Summary of Conclusions

The research findings showed a powerful positive correlation between income and strength-training activity in U.S. adults, with most of the differences in workout frequency could be explained by income level. The results of this analysis reinforce the idea that people who apply consistency with goal-driven behavior in strength training often have the same positive characteristics that drive career and financial growth.

G.2 Effective Storytelling

The visuals were key in making the results understandable and easy for anyone to comprehend. Charts are effective because they make the results undeniable and paint a clear picture of what is happening with the data. A simple rising red line for strength training and a falling blue line for inactivity tell a whole story in seconds. Tools such as Matplotlib and Seaborn make it easy to build powerful charts that clearly communicate these patterns.

G.3 Recommended Courses of Action

1. Implement wellness into professional development

Organizations could implement wellness directly into their professional development programs by giving employees short, structured exercise breaks a few times per week. A small on-site gym or workout area could boost focus, relieve stress, and boost productivity throughout the workweek.

2. Use this model with new variables

Future research can be conducted to identify further factors that influence income, and the necessary data is already included in the CDC dataset used for this project, such as age, gender, location, and education level. By adding variables, we may be able to pinpoint the traits or circumstances that most strongly contribute to creating personal and financial success.

H. Panopto Presentation

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=955a288f-bc02-40cd-a672-b39000f7ae6f

I. Appendices

Appendix A (code evidence)

Figure A.1 : Library Imports

[2]:
```
# Matthew Irvin
# WGU Capstone Task 3
# Correlation Between Strength Training and Income in the United States

# Import libraries and tools for data handling, stats, and visuals
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import pearsonr
import statsmodels.api as sm
```

Python libraries are imported (Pandas, NumPy, Seaborn, Matplotlib, SciPy, and Statsmodels) for handling data, visuals, and statistical analysis.

Figure A.2 : Load/Inspect CDC Dataset

```
[3]: # Read CDC dataset into a DataFrame
     df = pd.read_csv("Untitled Folder/Nutrition__Physical_Activity__and_Obesity_-_Behavioral_Risk_Factor_Surveillance_System 2.csv")

[5]: df.head()
```

[5]:

	YearStart	YearEnd	LocationAbbr	LocationDesc	Datasource	Class	Topic	Question	Data_Value_Unit	Data_Value_Type	...	GeoLocation	Class
0	2011	2011	AL	Alabama	Behavioral Risk Factor Surveillance System	Obesity / Weight Status	Obesity / Weight Status	Percent of adults aged 18 years and older who ...	NaN	Value	...	(32.84057112200048, -86.63186076199969)	O
1	2011	2011	AL	Alabama	Behavioral Risk Factor Surveillance System	Obesity / Weight Status	Obesity / Weight Status	Percent of adults aged 18 years and older who ...	NaN	Value	...	(32.84057112200048, -86.63186076199969)	O
2	2011	2011	AL	Alabama	Behavioral Risk Factor Surveillance System	Obesity / Weight Status	Obesity / Weight Status	Percent of adults aged 18 years and older who ...	NaN	Value	...	(32.84057112200048, -86.63186076199969)	O
3	2011	2011	AL	Alabama	Behavioral Risk Factor Surveillance System	Obesity / Weight Status	Obesity / Weight Status	Percent of adults aged 18 years and older who ...	NaN	Value	...	(32.84057112200048, -86.63186076199969)	O
4	2011	2011	AL	Alabama	Behavioral Risk Factor Surveillance System	Obesity / Weight Status	Obesity / Weight Status	Percent of adults aged 18 years and older who ...	NaN	Value	...	(32.84057112200048, -86.63186076199969)	O

5 rows × 33 columns

Loads the Behavioral Risk Factor Surveillance System (BRFSS) dataset from the CDC into Pandas DataFrame. The .head() cell verifies the dataset has imported correctly and shows the first few rows.

Figure A.3 : Filter, Clean, Statistical Testing

Main Analysis: Income vs Strength Training Participation

This section looks at how strength-training habits change across income groups.
After filtering for the CDC's "muscle-strengthening" question, we averaged the results within each income bracket to create a clean summary dataset.

Then, a Pearson correlation and simple linear regression were used to measure how closely income and training frequency move together.
Higher income groups clearly show higher participation, and the regression line quantifies that upward trend.

[8]:
```
# Filters rows for the strength-training questions only
mask = df["Question"].str.contains("muscle-strengthening", case=False, na=False)
df_strength = df[mask]
# Keep only the records that compare results by income
df_strength = df_strength[df_strength["StratificationCategory1"].str.contains("Income", case=False, na=False)]
# Keep useful columns & drop missing values
df_strength = df_strength[["Stratification1", "Data_Value"]].dropna()
# Average data within each income group. Aggregated across all states and years
df_income = df_strength.groupby("Stratification1")["Data_Value"].mean().reset_index()
```

[9]:
```
# Convert income brackets to numeric midpoints
def income_midpoint(value):
    value = str(value)
    if "Less" in value:
        return 7500
    elif "15,000" in value and "24,999" in value:
        return 20000
    elif "25,000" in value and "34,999" in value:
        return 30000
    elif "35,000" in value and "49,999" in value:
        return 42500
    elif "50,000" in value and "74,999" in value:
        return 62500
    elif "75,000" in value:
        return 90000
    elif "100,000" in value:
        return 110000
    else:
        return np.nan

# Add numeric income values/sort
df_income["Income_Midpoint"] = df_income["Stratification1"].apply(income_midpoint)
df_income = df_income.dropna(subset=["Income_Midpoint"])
```

[10]:
```
df_income = df_income.sort_values("Income_Midpoint")
```

[11]:
```
# Linear relationship strength (Pearson correlation)
r, p = pearsonr(df_income["Income_Midpoint"], df_income["Data_Value"])
print(f"Pearson correlation coefficient (r): {r:.4f}")
print(f"P-value: {p:.4f}")
```

```
Pearson correlation coefficient (r): 0.9970
P-value: 0.0000
```

[12]:
```
# Simple OLS regression model
X = sm.add_constant(df_income["Income_Midpoint"])
y = df_income["Data_Value"]
model = sm.OLS(y, X).fit()
print(model.summary())

from scipy.stats import spearmanr
rho, p_s = spearmanr(df_income["Income_Midpoint"], df_income["Data_Value"])
print(f"Spearman correlation (rho): {rho:.3f}, p-value: {p_s:.4g}")

# Refit OLS with robust standard errors (HC3)
ols_robust = sm.OLS(y, X).fit().get_robustcov_results(cov_type="HC3")
print("\nOLS with robust (HC3) SEs:\n")
print(ols_robust.summary())
```

This is the main analysis and shows how the dataset was filtered for strength-training & income fields by cleaning and transforming using numeric midpoints. The screenshot also includes the correlation, Pearson and Spearman, as well as regression (OLS and robust HC3) testing used to gauge the relationship between income and strength-training participation.

Figure A.4: Scatterplot

```
[13]: sns.set(style="darkgrid")
plt.figure(figsize=(9,5))
sns.regplot(x="Income_Midpoint", y="Data_Value", data=df_income, ci=None, color="deepskyblue", scatter_kws={"s":70, "color":"#D62828"})
plt.title("Strength Training vs. Income Level (CDC Data)")
plt.xlabel("Average Income per Bracket")
plt.ylabel("Adults(%) Strength Training")
plt.tight_layout()
plt.show()
```

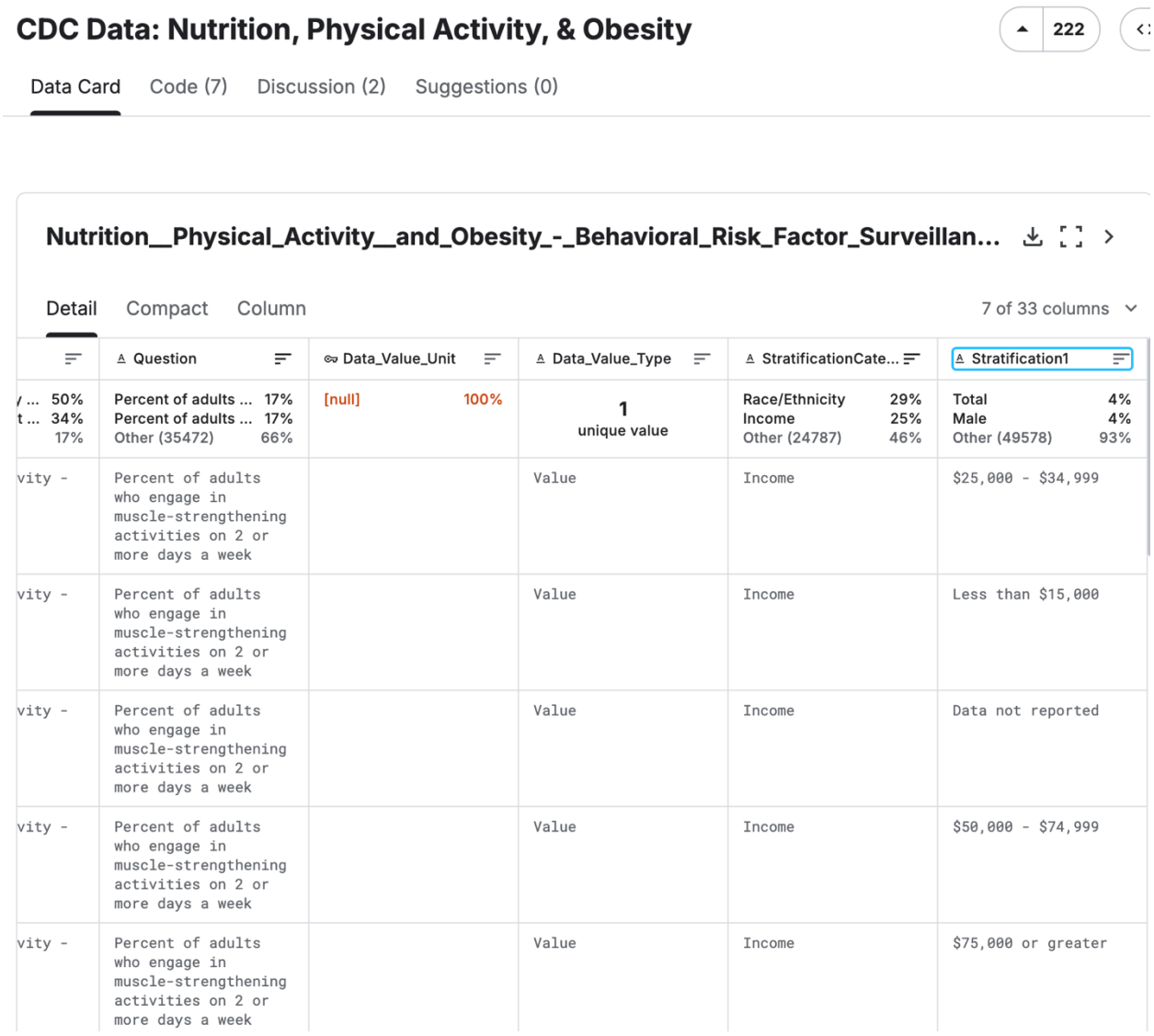
The chart visualizes the relationship between average income and the percentage of adults who strength train. A fitted regression line confirms a strong positive relationship between income and strength training.

Appendix B (Dataset Source)

The Behavioral Risk Factor Surveillance System (BRFSS), Nutrition, Physical Activity, and Obesity dataset.

Retrieved from Kaggle

URL: https://www.kaggle.com/datasets/spittman1248/cdc-data-nutrition-physical-activity-obesity



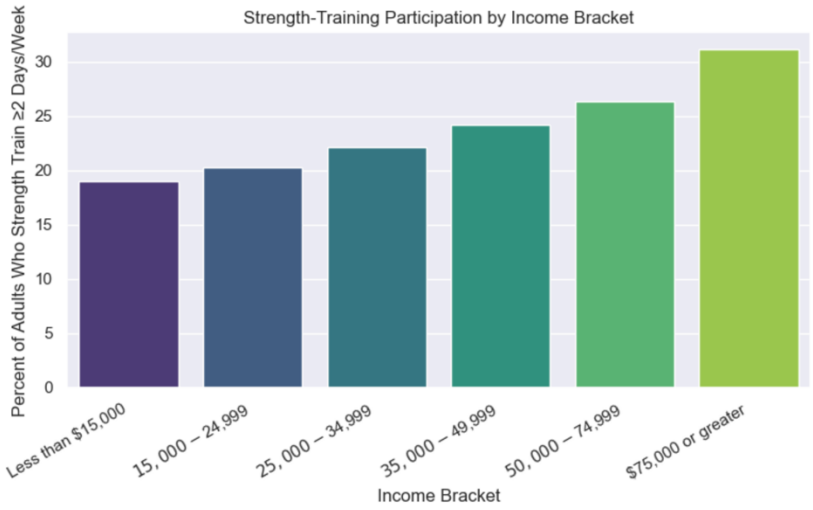
The question column was used for the strength training variable, and the StratificationCategory1 column was used for the income variable.

Appendix C (Visuals)

Figure C.1 Strength-Training participation per income bracket

```
[14]: plt.figure(figsize=(8,5))
sns.barplot(
    data=df_income,
    x="Stratification1", y="Data_Value",
    hue="Stratification1", palette="viridis", legend=False
)

plt.title("Strength-Training Participation by Income Bracket")
plt.xlabel("Income Bracket")
plt.ylabel("Percent of Adults Who Strength Train ≥2 Days/Week")
plt.xticks(rotation=30, ha="right")
plt.tight_layout()
plt.show()
```



A simple bar chart shows that adults in higher income brackets reported higher participation in strength training, with a clear upward trend across income brackets.

Figure C.2 Trend in Strength Training Across Income Levels

```
[15]: plt.figure(figsize=(8,5))
sns.lineplot(
    data=df_income.sort_values("Income_Midpoint"),
    x="Income_Midpoint", y="Data_Value",
    marker="o", linewidth=3.5, color="deepskyblue"
)
plt.title("Trend in Strength Training Across Income Levels")
plt.xlabel("Income (USD Midpoint)")
plt.ylabel("Strength Training 2 or more Days/Week")
plt.tight_layout()
plt.show()
```

A simple line chart to show a steady positive relationship between income and strength frequency.

Figure C.3 Dual-Line Chart (Physical Activity Patterns Across Income Levels)

A dual-line chart to visually compare adults who strength train vs those who do not.

J. References

Centers for Disease Control and Prevention (CDC). (2024). *Behavioral Risk Factor Surveillance System (BRFSS) – Nutrition, Physical Activity, and Obesity Dataset.* Retrieved from Kaggle.

Spittman, S. (2023). *CDC Data: Nutrition, Physical Activity, and Obesity [Dataset]*. Kaggle. https://www.kaggle.com/datasets/spittman1248/cdc-data-nutrition-physical-activity-obesity