Final Project: Exploring Factors Influencing Health and Lifestyle Choices

Project Title: **"Data-Driven Insights on Health and Lifestyle: Analyzing Correlations and Testing Hypotheses"**

Objective:

The goal of this project is to use a real-world health and lifestyle dataset to uncover relationships and test hypotheses regarding factors that may influence health outcomes, such as BMI, exercise habits, smoking, or diet. Students will apply hypothesis testing, correlation analyses (Pearson, Spearman, Kendall's Tau, Point-Biserial, and Phi), and other techniques they have learned in class to derive insights.

Project Components:

1. **Dataset**:

   Students can use a publicly available dataset such as:

   - Kaggle's Health Nutrition Dataset.
   - WHO Global Health Data.
   - CDC Behavioral Risk Factor Surveillance System (BRFSS).
   - UCI Machine Learning Repository: Lifestyle Choices and Disease Data.

2. **Project Requirements**:

   I. Exploratory Data Analysis (EDA):

   - Perform an initial analysis of the dataset. Summarize key statistics such as mean, median, and standard deviation for continuous variables (e.g., BMI, hours of exercise per week, daily calorie intake).

- Create visualizations (e.g., histograms, boxplots) to show the distribution of variables like age, weight, or hours of exercise.

II. Hypothesis Testing:

- **Propose hypotheses** based on the dataset. For example:
  - Hypothesis 1: "Smokers have a higher average BMI than non-smokers."
  - Hypothesis 2: "People who exercise more than 3 times per week have a lower average cholesterol level."
- For each hypothesis:
  - Conduct an appropriate **t-test**, **z-test**, or **ANOVA** to determine if there is statistical significance.
  - Clearly state the null and alternative hypotheses, calculate the p-value, and interpret the results.

III. Correlation Testing:

- **Identify correlations** between pairs of variables. For example:
  - Is there a **Pearson correlation** between age and cholesterol levels?
  - Is there a **Spearman correlation** between rank of exercise frequency and sleep quality?
  - **Point-Biserial correlation**: Examine whether being in a healthy weight range (binary) is correlated with daily fruit consumption (continuous).
  - **Phi coefficient**: Check if attending a yearly health checkup (yes/no) correlates with smoking habits (yes/no).
  - **Kendall's Tau**: Look for monotonic relationships between health factors, such as hours of exercise and sleep duration.
- Visualize the correlations with heatmaps and scatterplots, and interpret their strength and direction.

IV. Interpretation & Real-World Application:

- After conducting the hypothesis tests and correlation analyses, students should provide **insights** into the data:

- What factors seem to influence BMI, cholesterol, or other health outcomes?
- How might this data guide health policy recommendations or personal lifestyle changes?
- Do certain health behaviors (like smoking or exercise) have a measurable impact on health outcomes like heart rate or sleep duration?

V. Presentation & Reporting:

- Create a detailed **report** (or **presentation**) that includes:
  - Introduction to the dataset and research questions.
  - Summary of methods used (EDA, hypothesis testing, correlation testing).
  - Visualizations to support the findings.
  - Conclusion with real-world applications and recommendations.

3. **Evaluation Criteria**:
   - **Data exploration**: Quality of EDA, completeness, and visualizations.
   - **Hypothesis testing**: Correctly performed statistical tests with clear null and alternative hypotheses, proper interpretation of results.
   - **Correlation analyses**: Comprehensive application of different correlation tests, accurate interpretation, and relevance to real-world insights.
   - **Conclusions and presentation**: Clear, well-structured conclusions drawn from the analysis, including real-world implications, supported by data.

Example Real-World Scenarios for Inspiration:

1. **Healthcare**:
   - Investigate the relationship between lifestyle choices (e.g., exercise, diet, smoking) and health outcomes (e.g., BMI, cholesterol, blood pressure).
   - Hypothesis: "People who engage in regular physical activity have a lower risk of high blood pressure."

2. **Marketing**:
    - Explore how customer satisfaction (binary: satisfied/not satisfied) relates to spending habits (continuous: average spending per customer).
    - Hypothesis: "Customers who engage with loyalty programs tend to spend more on average."
3. **Education**:
    - Analyze how student study habits (hours studied, participation in study groups) correlate with exam performance (grades, pass/fail).
    - Hypothesis: "Students who attend study sessions have a higher chance of passing their exams."

Future: Extension for Machine Learning

As an optional challenge, students could explore building a **logistic regression model** to predict a binary outcome (e.g., pass/fail or healthy/unhealthy) based on continuous variables (e.g., study hours or BMI) and interpret the model's coefficients in relation to their earlier findings.
