correlation-vs-causation

January 29, 2024

0.0.1 Table of Contents

        4. Mechanistic Causation
        5. Interventionist Causation
        6. Counterfactual Causation
        7. Existential Causation
        8. Transcendental Causation
        9. Social Causation
        10. Psychological Causation
    - Examples Illustrating Causation
    - Challenges in Establishing Causation
4. **Correlation vs. Causation**
    - Key Differences
        1. Definition
        2. Nature of Relationship
        3. Direction
        4. Measurement
        5. Temporal Order
        6. Inference
    - Example using BMI dataset
        - Correlation Analysis
        - Regression Analysis
        - Interpreting Results
        - Impact of Sex on BMI
5. **Conclusion**
    - Summary of Correlation and Causation
    - Practical Insights from Data Analysis
    - Tips for Distinguishing Correlation from Causation

[ ]:

0.1 Introduction

Correlation and causation are two concepts that are often confused in statistics and data analysis. Correlation means that there is a statistical association between two variables, such that when one variable changes, the other variable tends to change in a certain direction. Causation means that a change in one variable causes a change in another variable, implying a cause-and-effect relationship between them.

[ ]:

0.1.1 Correlation

Definition: Correlation refers to a statistical relationship between two or more variables where changes in one variable are associated with changes in another. In simpler terms, it measures the degree to which two variables move in relation to each other.

Key Points: 1. **No Causation:** Correlation does not imply causation. Just because two variables are correlated does not mean that one causes the other. 2. **Strength and Direction:** Correlation can be positive, negative, or zero. A positive correlation means that as one variable increases, the other tends to increase as well. A negative correlation indicates that as one variable increases, the

other tends to decrease. 3. **Correlation Coefficient:** The strength and direction of correlation are quantified by the correlation coefficient (r). It ranges from -1 to 1, with -1 indicating a perfect negative correlation, 1 indicating a perfect positive correlation, and 0 indicating no correlation.

Example: *Positive Correlation:* There is a positive correlation between the amount of time spent studying and exam scores. As study time increases, exam scores tend to increase.

[ ]:

[78]:
```
from IPython.display import Image

# Specify the path to your image file
image_path = 'corr.PNG'

# Display the image
Image(filename=image_path)
```

[78]:

The formula for correlation is:

$$\rho = \frac{\sum_{i=1}^{N}(X_i - X)(Y_i - Y)}{\sqrt{\sum_{i=1}^{N}(X_i - X)^2}\sqrt{\sum_{i=1}^{N}(Y_i - Y)^2}}$$

Where:

- ρ is the correlation coefficient
- X_i and Y_i are the individual values of variables X and Y
- X and Y are the means of variables X and Y
- N is the number of observations

0.1.2 Types of correlation

There are several types of correlation coefficients used to measure the strength and direction of relationships between variables. The most common ones include:

1. **Pearson Correlation Coefficient (r):**
   - Measures the linear relationship between two continuous variables.
   - Values range from -1 to 1, where -1 indicates a perfect negative linear correlation, 1 indicates a perfect positive linear correlation, and 0 indicates no linear correlation.
2. **Spearman Rank Correlation Coefficient (or rs):**
   - Measures the strength and direction of monotonic relationships (not necessarily linear).
   - It is based on the ranks of the data rather than the actual values.
   - Useful when dealing with ordinal or non-normally distributed data.

3. **Kendall Tau Rank Correlation Coefficient (or Kendall's):**
   - Similar to Spearman's rank correlation but focuses on concordant and discordant pairs of data points.
   - It is also suitable for ordinal data and is less sensitive to outliers.
4. **Point-Biserial Correlation Coefficient (r_pb):**
   - Measures the strength and direction of the relationship between a dichotomous (binary) variable and a continuous variable.
   - Essentially a special case of the Pearson correlation coefficient.
5. **Phi Coefficient ( ):**
   - Measures the association between two dichotomous variables.
   - Often used in the context of 2x2 contingency tables.
6. **Cramér's V:**
   - A measure of association between two nominal (categorical) variables.
   - Similar to the Phi coefficient but applicable to larger contingency tables.
7. **Biserial Correlation Coefficient (r_b):**
   - Measures the strength and direction of the relationship between a dichotomous variable and a continuous variable.
   - Similar to the point-biserial correlation but used when the dichotomous variable is thought of as being continuous.
8. **Distance Correlation:**
   - Measures the association between two sets of points in a high-dimensional space.
   - It is a measure of dependence that is not restricted to linear relationships.

It's essential to choose the appropriate correlation coefficient based on the nature of your data and the type of variables involved. For example, Pearson correlation is suitable for linear relationships between continuous variables, while Spearman or Kendall correlation may be more appropriate for non-linear relationships or ordinal data.

[ ]:

0.1.3 Directions of correlation

When we refer to the "types" of correlation, we often mean the direction of the relationship between two variables. The direction of correlation can be positive, negative, or zero, indicating how the variables change in relation to each other. Let's explore these terms:

1. **Positive Correlation:**
   - A positive correlation exists when an increase in one variable is associated with an increase in the other.
   - The correlation coefficient (r) is positive, and the points on a scatter plot tend to form an upward-sloping line.
2. **Negative Correlation:**
   - A negative correlation occurs when an increase in one variable is associated with a decrease in the other.
   - The correlation coefficient (r) is negative, and the points on a scatter plot tend to form a downward-sloping line.
3. **Zero Correlation:**
   - Zero correlation indicates no linear relationship between two variables.

- The correlation coefficient (r) is close to 0, and the points on a scatter plot appear scattered without a clear trend.

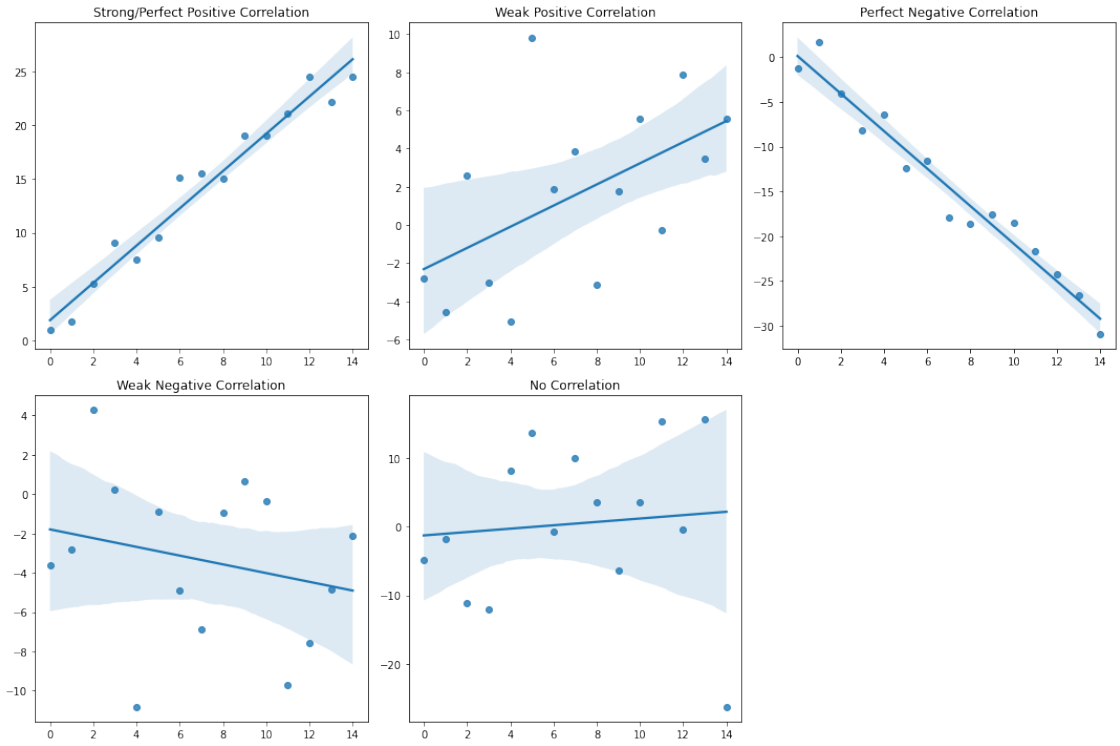
4. **Perfect Positive Correlation:**
   - A perfect positive correlation occurs when all data points fall exactly on a straight line with a positive slope.
   - The correlation coefficient (r) is +1.
5. **Perfect Negative Correlation:**
   - A perfect negative correlation occurs when all data points fall exactly on a straight line with a negative slope.
   - The correlation coefficient (r) is -1.
6. **No Correlation (Uncorrelated):**
   - The absence of a linear relationship between two variables.
   - The correlation coefficient (r) is exactly 0.

Understanding the direction and strength of correlation is crucial in interpreting relationships between variables. Additionally, correlation coefficients are sensitive to outliers, and non-linear relationships may not be adequately captured by linear correlation measures.

[ ]:

[79]:
```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# Set a seed for reproducibility
np.random.seed(42)

# Generate synthetic data
size = 15

# Strong/Perfect Positive Correlation
strong_positive_x = np.arange(size)
strong_positive_y = 2 * strong_positive_x + np.random.normal(scale=2, size=size)

# Weak Positive Correlation
weak_positive_x = np.arange(size)
weak_positive_y = 0.5 * weak_positive_x + np.random.normal(scale=5, size=size)

# Perfect Negative Correlation
perfect_negative_x = np.arange(size)
perfect_negative_y = -2 * perfect_negative_x + np.random.normal(scale=2,␣
↪size=size)

# Weak Negative Correlation
weak_negative_x = np.arange(size)
weak_negative_y = -0.5 * weak_negative_x + np.random.normal(scale=5, size=size)

# No Correlation
```

```
no_correlation_x = np.arange(size)
no_correlation_y = np.random.normal(scale=10, size=size)

# Plotting with Seaborn
fig, axes = plt.subplots(2, 3, figsize=(15, 10))

# Strong/Perfect Positive Correlation
sns.regplot(x=strong_positive_x, y=strong_positive_y, ax=axes[0, 0],␣
 ↪label='Strong/Perfect Positive')
axes[0, 0].set_title('Strong/Perfect Positive Correlation')

# Weak Positive Correlation
sns.regplot(x=weak_positive_x, y=weak_positive_y, ax=axes[0, 1], label='Weak␣
 ↪Positive')
axes[0, 1].set_title('Weak Positive Correlation')

# Perfect Negative Correlation
sns.regplot(x=perfect_negative_x, y=perfect_negative_y, ax=axes[0, 2],␣
 ↪label='Perfect Negative')
axes[0, 2].set_title('Perfect Negative Correlation')

# Weak Negative Correlation
sns.regplot(x=weak_negative_x, y=weak_negative_y, ax=axes[1, 0], label='Weak␣
 ↪Negative')
axes[1, 0].set_title('Weak Negative Correlation')

# No Correlation
sns.regplot(x=no_correlation_x, y=no_correlation_y, ax=axes[1, 1], label='No␣
 ↪Correlation')
axes[1, 1].set_title('No Correlation')

# Remove empty subplot
fig.delaxes(axes[1, 2])

# Adjust layout
plt.tight_layout()
plt.show()
```



[ ]:

[ ]:

[ ]:

[ ]:

0.1.4 Causation

Definition: Causation implies a cause-and-effect relationship between two variables, where changes in one variable directly result in changes in another. Establishing causation is more complex than observing correlation.

Key Points: 1. **Cause and Effect:** In a causal relationship, changes in the independent variable directly cause changes in the dependent variable. 2. **Experimental Design:** Establishing causation often requires experimental designs where one variable is manipulated while keeping other factors constant. This helps rule out alternative explanations for the observed relationship. 3. **Temporal Order:** Causation requires a temporal order, meaning the cause must precede the effect in time.

Example: *Causation:* Administering a new drug to patients and observing a subsequent improvement in their symptoms establishes a causal relationship between the drug and symptom relief.

0.1.5 Some types of causation

Below are some types of causation 1. **Factual Causation:** - **Explanation:** Factual causation, often assessed using the counterfactual test, asks whether the cause was a necessary condition for the effect to occur. - **Example:** "But for the icy road conditions, the car would not have skidded and collided with the barrier."

2. **Legal Causation (Proximate Cause):**
   - **Explanation:** Legal causation, or proximate cause, evaluates whether the cause was closely related to the effect, considering factors like foreseeability and proximity.
   - **Example:** "The shop owner's negligence in not fixing the broken step was the proximate cause of the customer's fall and subsequent injury."
3. **Probabilistic Causation:**
   - **Explanation:** Probabilistic causation assesses whether the cause increases the probability of the effect occurring.
   - **Example:** "Smoking probabilistically causes lung cancer, as smokers have a higher probability of developing lung cancer than non-smokers."
4. **Mechanistic Causation:**
   - **Explanation:** Mechanistic causation explores how the cause produced the effect, focusing on the underlying mechanisms and processes.
   - **Example:** "The release of histamine mechanistically causes allergy symptoms, as it triggers inflammation and other immune responses."
5. **Interventionist Causation:**
   - **Explanation:** Interventionist causation assesses whether manipulating the cause can change the effect.
   - **Example:** "Administering a painkiller intervenes to cause relief, as it alters the perception of pain in the nervous system."
6. **Counterfactual Causation:**
   - **Explanation:** Counterfactual causation considers what would have happened if the cause had not occurred.
   - **Example:** "If the safety measures had not been implemented, the counterfactual causation suggests that more accidents might have occurred in the workplace."
7. **Existential Causation:**
   - **Explanation:** Existential causation focuses on the actual existence of the cause and its role in bringing about the effect.
   - **Example:** "The existence of a genetic mutation is an existential cause of the inherited disorder."
8. **Transcendental Causation:**
   - **Explanation:** Transcendental causation considers metaphysical or philosophical aspects of causation that go beyond observable phenomena.
   - **Example:** "Some philosophers argue that consciousness is the transcendental cause underlying all human experiences."
9. **Social Causation:**
   - **Explanation:** Social causation examines how social factors contribute causally to certain outcomes or phenomena.
   - **Example:** "Social causation theory posits that socioeconomic factors, such as poverty and discrimination, causally impact mental health disparities in a population."
10. **Psychological Causation:**

- **Explanation:** Psychological causation explores how mental processes, thoughts, and behaviors act as causes for various psychological outcomes.
- **Example:** "Psychological causation suggests that chronic stress can cause changes in brain function, contributing to the development of anxiety disorders."

These explanations and examples provide a deeper understanding of the different types of causation and how they are applied in various contexts.

[ ]:

0.1.6 Correlation VS Causation

0.1.7 Examples Illustrating Correlation and Causation:

1. Ice Cream Sales and Drowning Incidents: - *Correlation:* There is a positive correlation between ice cream sales and the number of drowning incidents. During the summer months, both increase. - *Causation:* No direct causation. The increase in ice cream sales and drowning incidents is due to a common factor—warm weather.

2. Sunscreen Use and Skin Cancer: - *Correlation:* There is a negative correlation between sunscreen use and the incidence of skin cancer. As sunscreen use increases, skin cancer rates tend to decrease. - *Causation:* It's plausible that sunscreen use may prevent skin cancer, but establishing causation requires controlled experiments.

3. Education Level and Income: - *Correlation:* Positive correlation between education level and income. Generally, higher education is associated with higher income. - *Causation:* While education can contribute to higher income, other factors (like job market conditions) also play a role. Causation is complex in this context.

4. Exercise and Weight Loss: - *Correlation:* Positive correlation between regular exercise and weight loss. - *Causation:* While exercise contributes to weight loss, it's not the only factor. Diet, genetics, and metabolism also play roles.

5. Example: **-** Correlation:** There is a strong positive correlation between ice cream sales and drowning incidents. However, buying ice cream does not cause drownings; both are influenced by a common factor, warmer weather. - **Causation:** Administering a vaccine causes an increase in immunity. The act of vaccination directly leads to a change in the immune status of an individual. ### Tips for Distinguishing Correlation from Causation:

1. **Consider Alternative Explanations:**
   - Correlation may be coincidental. Consider other factors that could explain the observed relationship.
2. **Temporal Sequence:**
   - If establishing causation, ensure that the cause precedes the effect in time.
3. **Controlled Experiments:**
   - Conducting controlled experiments helps isolate variables and establish causation.
4. **Consistency and Replication:**
   - A causal relationship should be consistent across different studies and populations.
5. **Biological Plausibility:**
   - A causal relationship is more plausible if it aligns with known biological mechanisms.
6. **Critical Thinking:**

- Approach data with skepticism and critically evaluate the study design and methodology.

0.1.8 Common Mistakes to Avoid:

1. **Assuming Correlation Implies Causation:**
   - Correlation is not sufficient evidence for causation. Look for additional evidence.
2. **Ignoring Confounding Variables:**
   - Failure to account for confounding variables can lead to inaccurate conclusions about causation.
3. **Post Hoc Fallacy:**
   - Assuming that because one event follows another, the first event caused the second.
4. **Overlooking Bidirectional Causation:**
   - Recognize that causation can sometimes be bidirectional, where variables influence each other.

By understanding these nuances, researchers and analysts can make more informed decisions and contribute to sound scientific inquiry.

[ ]:

[ ]:

0.1.9 Differences:

Correlation and causation are two concepts often used in statistics and research, but they represent different aspects of relationships between variables. Here are the key differences between correlation and causation:

1. **Definition:**
   - **Correlation:** Correlation refers to a statistical measure that describes the extent to which two variables change in relation to each other. It quantifies the strength and direction of a linear relationship between variables.
   - **Causation:** Causation, on the other hand, denotes a cause-and-effect relationship between two variables, where a change in one variable (the cause) leads to a change in another variable (the effect).
2. **Nature of Relationship:**
   - **Correlation:** Correlation indicates the degree to which two variables are associated or co-vary. It does not imply a direct cause-and-effect relationship; it merely shows that as one variable changes, the other tends to change as well.
   - **Causation:** Causation implies a direct influence, where changes in one variable result in changes in another. It suggests that one variable is responsible for the observed effect in the other.
3. **Direction:**
   - **Correlation:** Correlation can be positive, negative, or zero. A positive correlation indicates that both variables increase or decrease together, a negative correlation implies that as one variable increases, the other decreases, and zero correlation means no linear relationship.
   - **Causation:** Causation involves a specific direction from the cause to the effect. Changes in the cause lead to changes in the effect, and not vice versa.

4. **Measurement:**
   - **Correlation:** Measured using correlation coefficients such as Pearson's r, Spearman's , or Kendall's . The values range from -1 to 1, indicating the strength and direction of the correlation.
   - **Causation:** There is no single statistical measure for causation. Establishing causation often involves a combination of study design, statistical analysis, and consideration of other factors.
5. **Temporal Order:**
   - **Correlation:** Does not require a specific temporal order. Correlation can exist at a single point in time or over a period.
   - **Causation:** Requires a clear temporal order, where the cause precedes the effect in time.
6. **Inference:**
   - **Correlation:** Correlation does not imply causation. Even a strong correlation does not prove that changes in one variable cause changes in the other; it could be coincidental or influenced by confounding variables.
   - **Causation:** Establishing causation requires additional evidence, often from experimental studies, to support a direct cause-and-effect relationship.

[ ]:

[ ]:

0.1.10 Example using BMI dataset

[80]:
```
import pandas as pd
```

[81]:
```
#Import data
df = pd.read_csv("bmi_data.csv")
df.head()
```

[81]:

	Sex	Age	Height(Inches)	Weight(Pounds)	BMI
0	Female	21	65.78331	112.9925	18.357646
1	Female	35	71.51521	136.4873	18.762652
2	Female	27	69.39874	153.0269	22.338985
3	Male	24	68.21660	142.3354	21.504612
4	Female	18	67.78781	144.2971	22.077669

[82]:
```
df.shape
```

[82]:
```
(25000, 5)
```

[83]:
```
# Data types
df.dtypes
```

[83]:
```
Sex               object
Age                int64
```

```
Height(Inches)    float64
Weight(Pounds)    float64
BMI               float64
dtype: object
```

[84]:
```
#check for missing values
df.isna().sum()
```

[84]:
```
Sex                0
Age                0
Height(Inches)    19
Weight(Pounds)    16
BMI               50
dtype: int64
```

[85]:
```
#Drop Missing values
df = df.dropna()
```

[86]:
```
df.isna().sum()
```

[86]:
```
Sex               0
Age               0
Height(Inches)    0
Weight(Pounds)    0
BMI               0
dtype: int64
```

[87]:
```
# Statistics
df.describe()
```

[87]:

	Age	Height(Inches)	Weight(Pounds)	BMI
count	24950.000000	24950.000000	24950.000000	24950.000000
mean	26.497836	67.992821	127.077390	19.321368
std	5.190667	1.901551	11.663509	1.552091
min	18.000000	60.278360	78.014760	13.070879
25%	22.000000	66.704955	119.307525	18.278339
50%	27.000000	67.995700	127.152500	19.302160
75%	31.000000	69.271823	134.893550	20.357547
max	35.000000	75.152800	170.924000	26.023756

[88]:
```
categorical_variables = df.select_dtypes(include=['object'])
numerical_variables = df.select_dtypes(exclude=['object'])
print (categorical_variables.columns)
print (numerical_variables.columns)
```

```
Index(['Sex'], dtype='object')
Index(['Age', 'Height(Inches)', 'Weight(Pounds)', 'BMI'], dtype='object')
```

[89]:
```
# BMIClasses
# Count of unique values in the 'BMIClass' column
bmi_class_counts = df['Sex'].value_counts(normalize = True)

# Display the counts
print(bmi_class_counts)
```

```
Sex
Male      0.503527
Female    0.496473
Name: proportion, dtype: float64
```

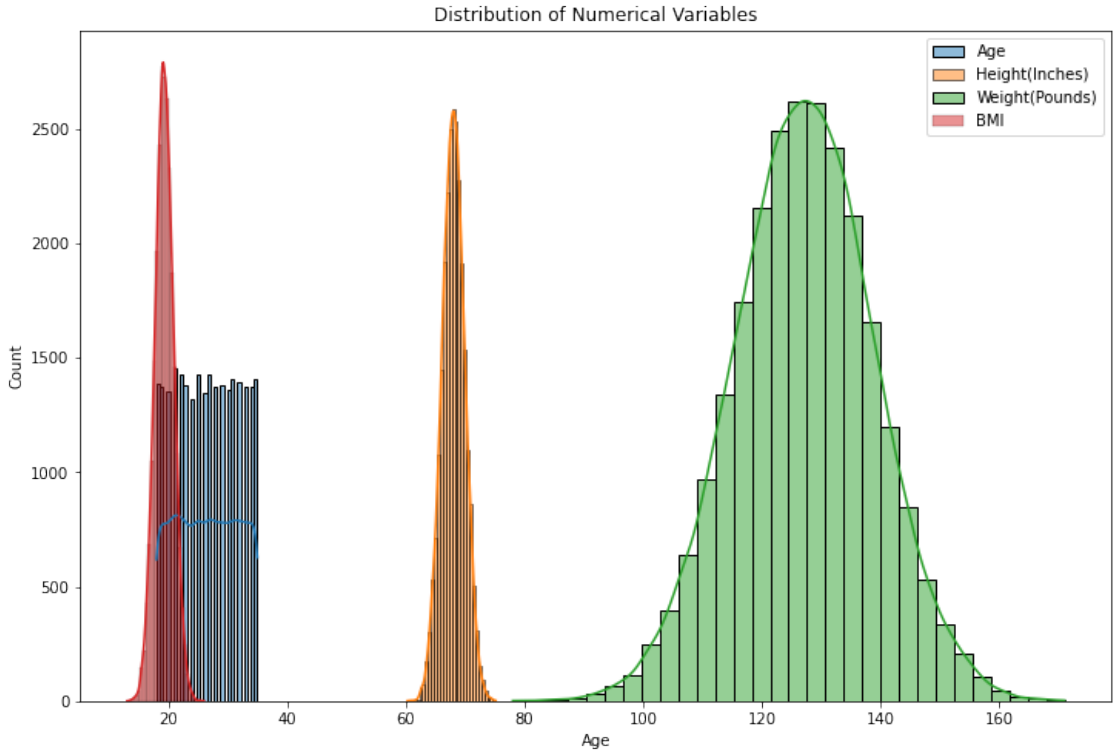
[ ]:
```

```

0.1.11 Correlation

Different ways of visualizing correlation

[90]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

[91]:
```
# Distribution plots
plt.figure(figsize=(12, 8))

# Plotting histograms for numerical variables
for column in numerical_variables.columns:
    sns.histplot(df[column], kde=True, label=column, bins=30)

plt.title('Distribution of Numerical Variables')
plt.legend()
plt.show()
```



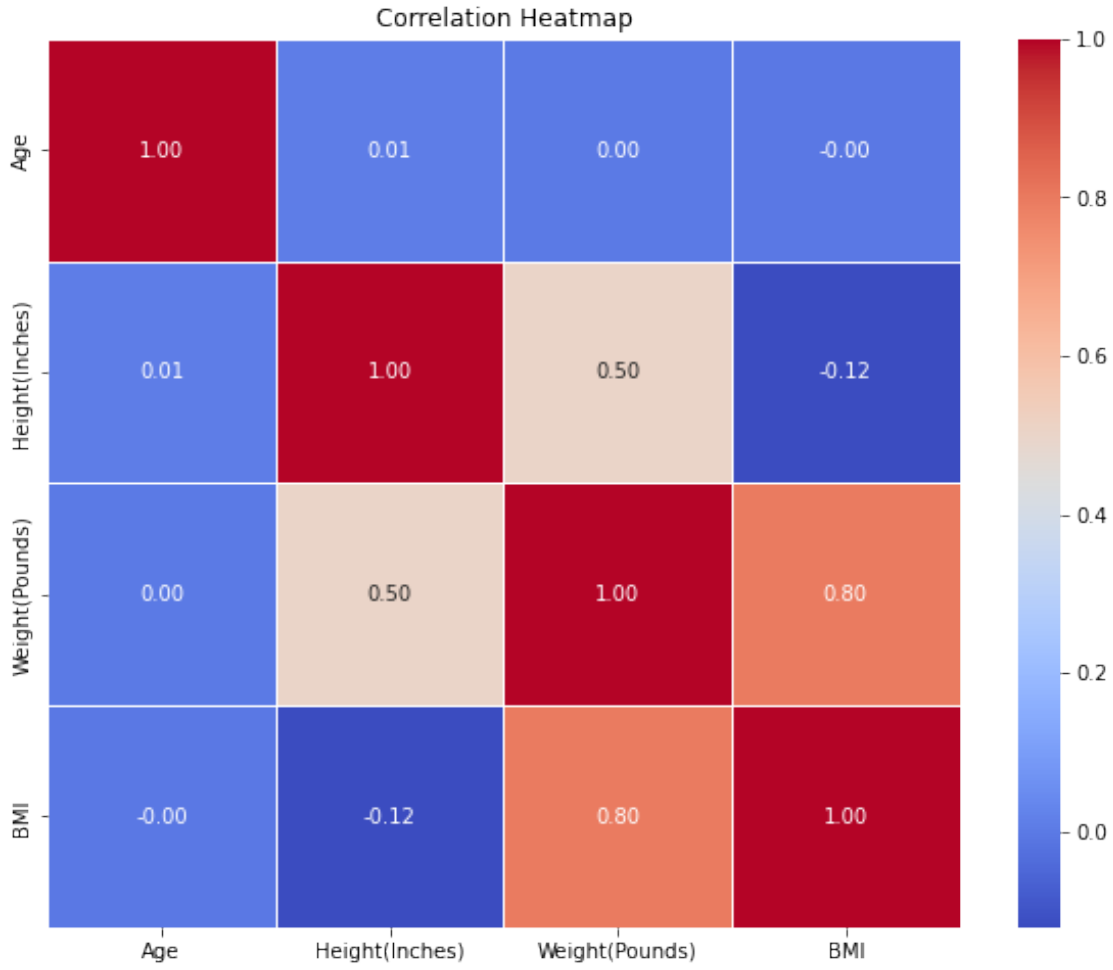
1. Correlation matrix

[92]:
```
correlation_matrix = numerical_variables.corr()
print(correlation_matrix)
```

```
                     Age  Height(Inches)  Weight(Pounds)       BMI
Age             1.000000        0.006268        0.001151 -0.003323
Height(Inches)  0.006268        1.000000        0.502925 -0.121230
Weight(Pounds)  0.001151        0.502925        1.000000  0.795607
BMI            -0.003323       -0.121230        0.795607  1.000000
```

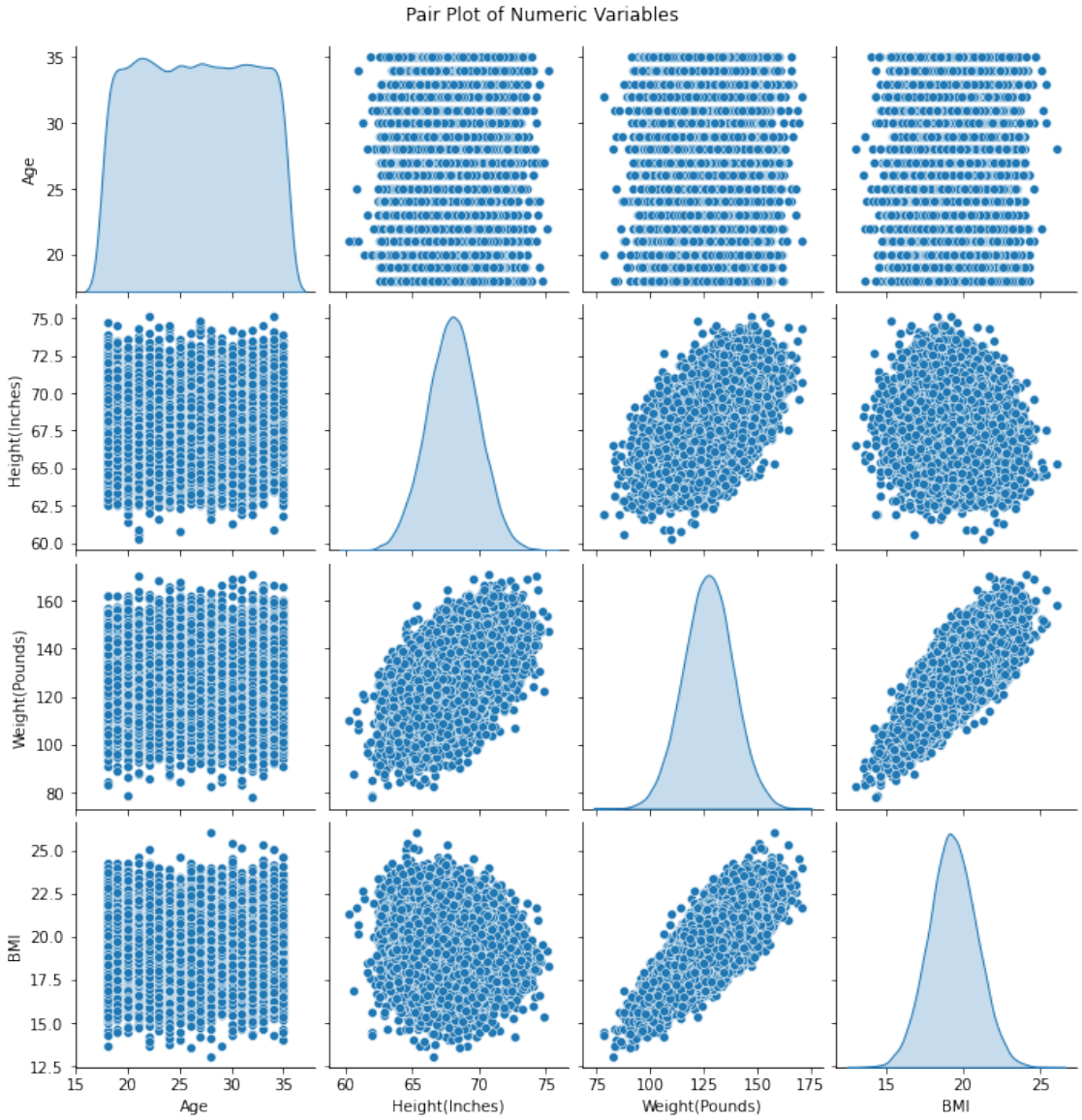
2. Heatmap

[93]:
```
correlation_matrix = numerical_variables.corr()
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f',␣
 ↪linewidths=.5)
plt.title('Correlation Heatmap')
plt.show()
```



3. Pairplots

[94]:
```
sns.pairplot(numerical_variables, diag_kind='kde')
plt.suptitle('Pair Plot of Numeric Variables', y=1.02)
plt.show()
```

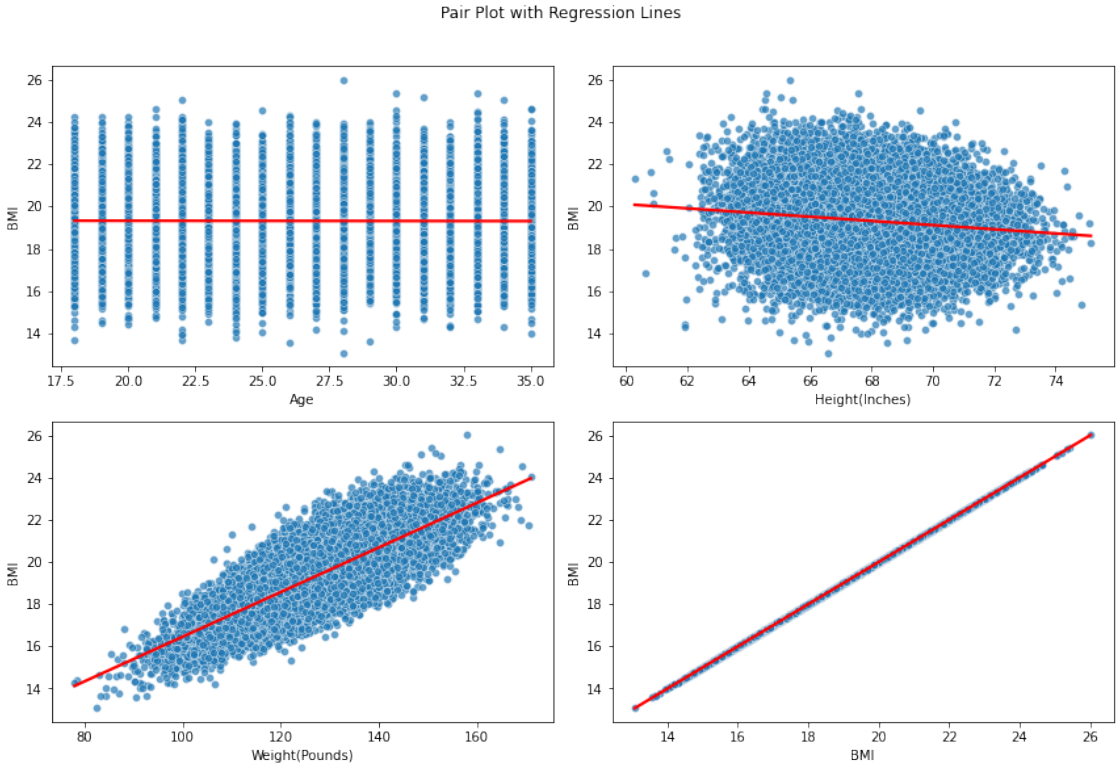


[ ]:

4. Regression Line plot

[95]:
```
# Pair plot with regression lines using matplotlib
plt.figure(figsize=(12, 8))

# Plotting scatter plots and regression lines
for i, column in enumerate(numerical_variables.columns):
    plt.subplot(2, 2, i+1)
    sns.scatterplot(x=column, y='BMI', data=numerical_variables, alpha=0.7)
```

```
    sns.regplot(x=column, y='BMI', data=df, scatter=False, ci=None,␣
 ↪line_kws={'color':'red'})

plt.suptitle('Pair Plot with Regression Lines', y=1.02)
plt.tight_layout()
plt.show()
```



The correlation matrix shows the Pearson correlation coefficients between pairs of numerical variables (Age, Height(Inches), Weight(Pounds), BMI) in your dataset. Here's how to interpret the results:

1. **Age and BMI:**
   - The correlation coefficient between Age and BMI is approximately -0.0033. This value is close to zero, indicating a very weak correlation between Age and BMI. The negative sign suggests a slight tendency for BMI to decrease with increasing age, but the correlation is negligible.
2. **Height(Inches) and BMI:**
   - The correlation coefficient between Height(Inches) and BMI is approximately -0.1212. This value indicates a weak negative correlation. Taller individuals tend to have a slightly lower BMI on average. However, the correlation is not very strong.
3. **Weight(Pounds) and BMI:**
   - The correlation coefficient between Weight(Pounds) and BMI is approximately 0.7956. This value indicates a strong positive correlation. As expected, there is a significant

positive relationship between weight and BMI—individuals with higher weight tend to have a higher BMI. The positive sign suggests a direct proportional relationship.

It's important to note that correlation does not imply causation. While there is a correlation between weight and BMI, for example, it doesn't mean that an increase in weight causes an increase in BMI or vice versa. Correlation only measures the strength and direction of a linear relationship between variables.

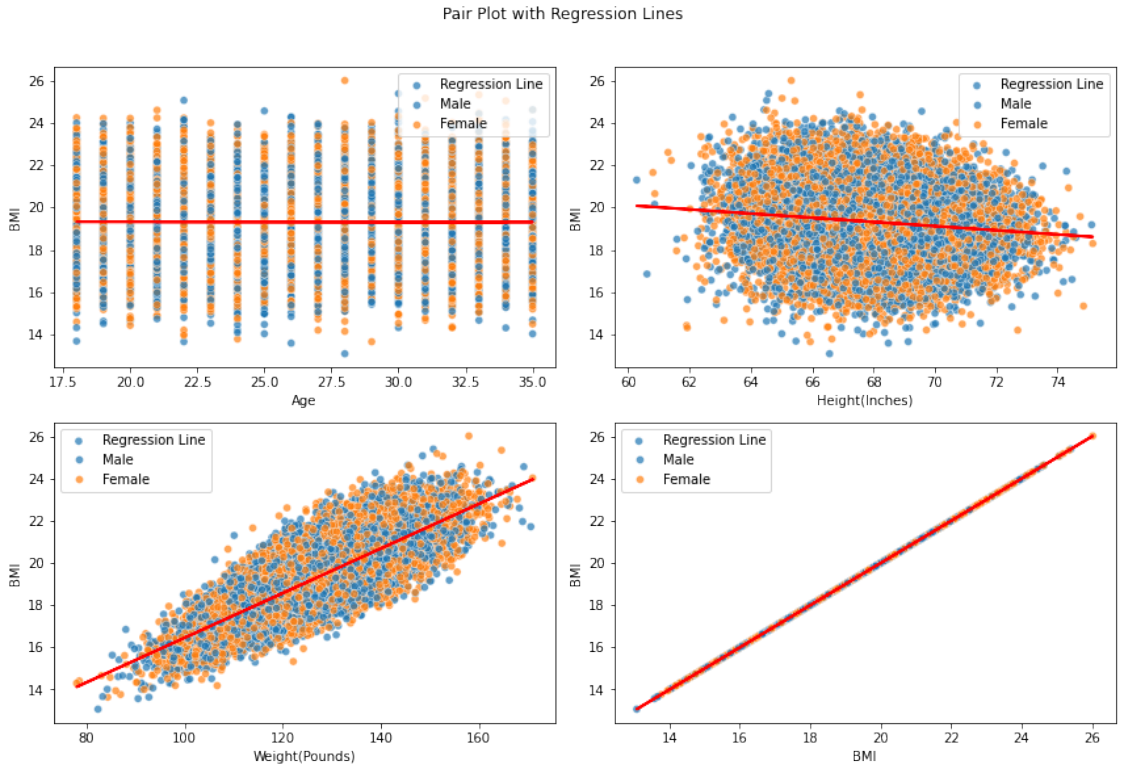
To explore causation, you would need to conduct further analyses, such as experimental studies or controlled experiments. The correlation results provide insights into associations between variables but do not establish a causal relationship. Always be cautious when inferring causation based solely on correlation.

[ ]:

[ ]:

[96]:
```
from scipy.stats import linregress

# Pair plot with regression lines using matplotlib
plt.figure(figsize=(12, 8))

# Store regression coefficients
coefficients = []

# Plotting scatter plots and regression lines
for i, column in enumerate(numerical_variables.columns):
    plt.subplot(2, 2, i+1)

    # Scatter plot
    sns.scatterplot(x=column, y='BMI', data=df, hue='Sex', alpha=0.7)

    # Regression line
    result = linregress(df[column], df['BMI'])
    coefficients.append((column, result.slope, result.intercept))

    # Plotting regression line
    x_values = df[column]
    y_values = result.intercept + result.slope * x_values
    plt.plot(x_values, y_values, color='red', linewidth=2)

    # Adding legend for the regression line
    plt.legend(['Regression Line', 'Male', 'Female'])

plt.suptitle('Pair Plot with Regression Lines', y=1.02)
plt.tight_layout()
plt.show()

# Display regression coefficients
```

```
for column, slope, intercept in coefficients:
    print(f"{column}: Slope = {slope:.4f}, Intercept = {intercept:.4f}")
```




```
Age: Slope = -0.0010, Intercept = 19.3477
Height(Inches): Slope = -0.0990, Intercept = 26.0493
Weight(Pounds): Slope = 0.1059, Intercept = 5.8673
BMI: Slope = 1.0000, Intercept = 0.0000
```

The regression coefficients you provided indicate how changes in each variable are associated with changes in BMI. Let's interpret each set of coefficients:

1. **Age: Slope = -0.0010, Intercept = 19.3477**
   - For each additional year of age, the BMI decreases by approximately 0.0010. The intercept of 19.3477 represents the estimated BMI when age is zero. However, interpreting the intercept in the context of age may not be meaningful, as age is not a variable that can be zero in a realistic scenario.
2. **Height(Inches): Slope = -0.0990, Intercept = 26.0493**
   - For each additional inch of height, the BMI decreases by approximately 0.0990. The intercept of 26.0493 represents the estimated BMI when height is zero, which is not practically meaningful.
3. **Weight(Pounds): Slope = 0.1059, Intercept = 5.8673**
   - For each additional pound of weight, the BMI increases by approximately 0.1059. The intercept of 5.8673 represents the estimated BMI when weight is zero, which is not a realistic scenario.

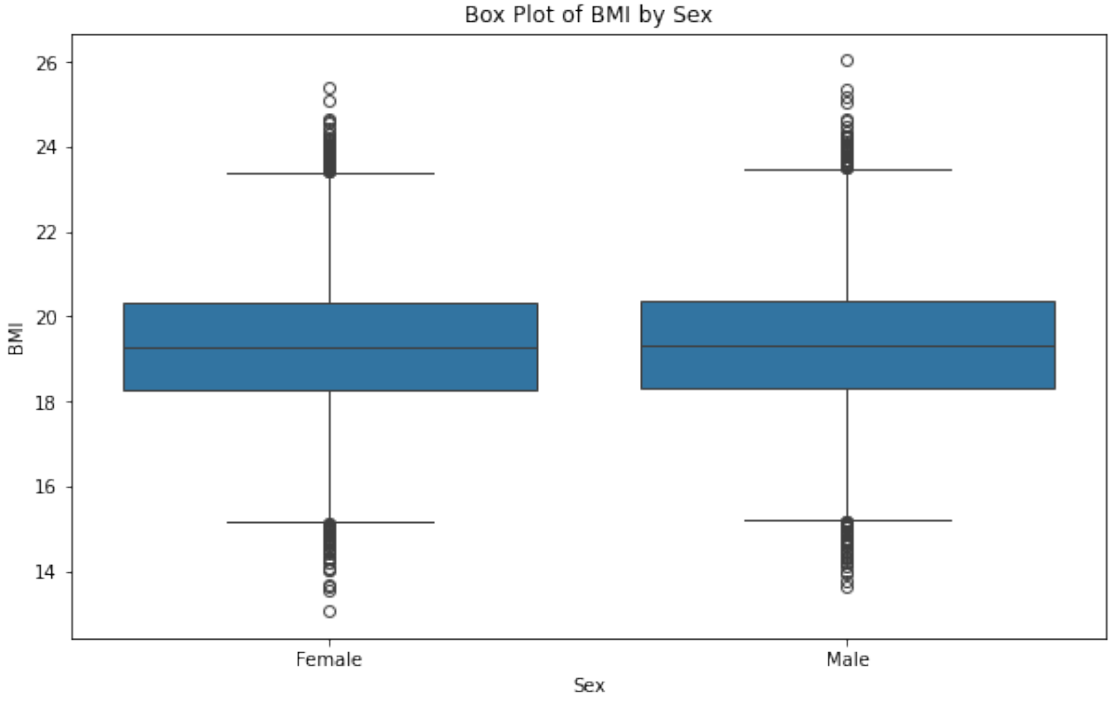
4. **BMI: Slope = 1.0000, Intercept = 0.0000**
    - The coefficients for BMI itself are trivial. The slope of 1.0000 means that a one-unit increase in BMI results in a one-unit increase in BMI. The intercept of 0.0000 is also trivial and is expected, as BMI is zero when BMI is zero.

It's important to note that interpreting intercepts in the context of these variables may not be meaningful, especially when the variable cannot practically be zero (e.g., age, height, weight). The slope indicates the estimated change in BMI for a one-unit change in the corresponding variable.

Additionally, keep in mind that correlation and regression do not establish causation. These coefficients represent associations, and the direction and strength of the associations may not imply causation. Further investigation and consideration of confounding factors are necessary for a more nuanced understanding.

Correlation between BMI and Sex

[97]:
```
# Box plot for BMI by sex
plt.figure(figsize=(10, 6))
sns.boxplot(x='Sex', y='BMI', data=df)
plt.title('Box Plot of BMI by Sex')
plt.show()
```



[98]:
```
# Summary statistics of BMI by sex
bmi_summary_by_sex = df.groupby('Sex')['BMI'].describe()

# Display the summary statistics
```

```
print(bmi_summary_by_sex)
```

```
          count       mean       std        min        25%        50%  \
Sex
Female  12387.0  19.300909  1.545866  13.070879  18.261913  19.278544
Male    12563.0  19.341540  1.558003  13.641129  18.303884  19.323838

              75%        max
Sex
Female  20.327472  25.402883
Male    20.380405  26.023756
```

Analyzing the summary statistics of BMI by sex, we can observe the following:

1. **Mean BMI:**
   - The mean BMI for females is approximately 19.30, while the mean BMI for males is approximately 19.34. This suggests a very slight difference in the average BMI between the two groups.
2. **Variability:**
   - The standard deviation (std) for BMI is similar for both females (approximately 1.55) and males (approximately 1.56). This indicates that the variability in BMI within each group is comparable.
3. **Distribution:**
   - Looking at the quartiles (Q1, Q2, Q3), the median (Q2) BMI for both females (around 19.28) and males (around 19.32) is quite close. This suggests that the middle 50% of BMI values in each group are similar.
4. **Range:**
   - The range of BMI values (difference between the maximum and minimum) is slightly wider for males (from around 13.64 to 26.02) compared to females (from around 13.07 to 25.40).

The overall impact of sex on BMI appears to be minimal. The mean BMI values are very close, and the variability within each group is similar.

[ ]:

[ ]:

0.1.12 Conclusion

The results from the correlation analysis provide insights into the statistical relationships between different variables and BMI. Here are some key observations:

1. **Age and BMI:**
   - There is a very weak negative correlation between age and BMI (correlation coefficient -0.0033). However, this correlation is practically negligible, indicating that age has minimal impact on BMI.
2. **Height and BMI:**
   - There is a weak negative correlation between height and BMI (correlation coefficient -0.1212). Taller individuals tend to have a slightly lower BMI on average, but the

correlation is not very strong.

3. **Weight and BMI:**
   - There is a strong positive correlation between weight and BMI (correlation coefficient 0.7956). As expected, individuals with higher weight tend to have a higher BMI. This correlation is significant and indicates a direct proportional relationship.
4. **Sex and BMI:**
   - The correlation analysis did not directly include the variable "Sex" because it's categorical. However, based on subsequent analyses, the mean BMI values for females and males were found to be very close, suggesting that sex has a minimal impact on BMI.

Concerning the Difference Between Correlation and Causation:

- **Correlation does not imply causation:**
  - Correlation measures the strength and direction of a linear relationship between two variables but does not establish a causal connection. For example, the correlation between weight and BMI is strong, but it doesn't imply that an increase in weight causes an increase in BMI or vice versa.
- **Causation requires additional evidence:**
  - Establishing causation involves more than just observing a correlation. It requires further investigation, controlled experiments, and consideration of potential confounding variables. The observed correlations can guide hypotheses about causation, but additional research is needed to draw causal conclusions.

In summary, while correlations provide valuable insights into associations between variables, it's essential to be cautious when inferring causation. Additional analyses, experiments, and consideration of potential confounding factors are necessary to establish causal relationships definitively. Always interpret correlation results within the appropriate context and consider the limitations of observational studies.

[ ]: