content

1.Introduction

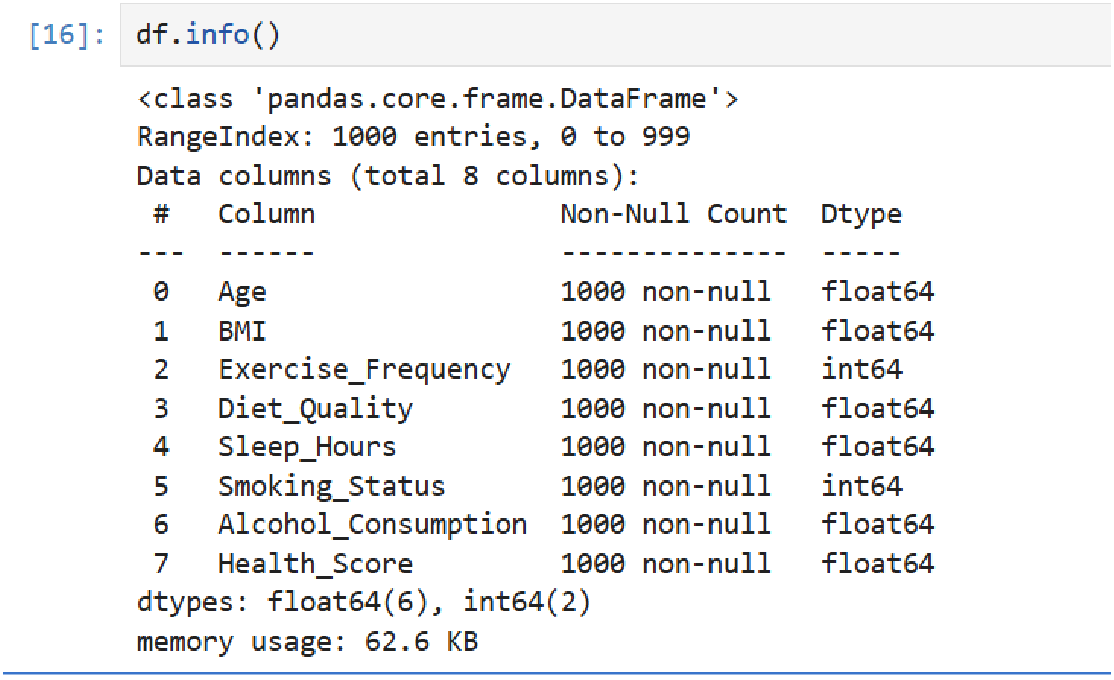
Demographic changes and trends in recent years have led to a dramatic increase in morbidity and mortality from non-communicable diseases, particularly cardiovascular diseases, diabetes and obesity. This growing health burden hinders poverty reduction and economic growth. NCDs such as ischaemic heart disease and stroke, cancer, chronic obstructive pulmonary disease, Alzheimer's disease and other dementias, and diabetes are the biggest killers in the pre-epidemic period, accounting for 74 per cent of all deaths in 2019, WHO said in the report. During the COVID-19 rampage, NCDs were still able to account for 78 per cent of COVID-19 deaths. (WHO) In an age of rapid internet growth, intense work and an increasingly fast pace of life, it is difficult for people to maintain a good lifestyle. Long hours of focusing on computers, over-processed food, irregular work routines and lack of sleep are becoming more evident the more urbanised a place is. Therefore, it is important to explore the specific impact of lifestyle on health. This will not only help individuals to develop healthier life plans, but also provide a scientific basis for public health policies, thereby reducing the burden of chronic diseases and contributing to the overall health of society.

2.Data description and pre-processing

2.1 Data Source

The dataset used in this study comes from the Kaggle platform, an open data sharing and competition platform widely used by data scientists and researchers. The dataset focuses on exploring the association between lifestyle and health, and contains a number of key variables that can reveal the impact on individual health from multiple dimensions such as diet, exercise, sleep, smoking, and alcohol consumption. The data is generated through the combination of questionnaire surveys and health check-up results, which has strong representativeness and application value.

2.2 Data content



```
[16]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   Age                  1000 non-null   float64
 1   BMI                  1000 non-null   float64
 2   Exercise_Frequency   1000 non-null   int64  
 3   Diet_Quality         1000 non-null   float64
 4   Sleep_Hours          1000 non-null   float64
 5   Smoking_Status       1000 non-null   int64  
 6   Alcohol_Consumption  1000 non-null   float64
 7   Health_Score         1000 non-null   float64
dtypes: float64(6), int64(2)
memory usage: 62.6 KB
```

The dataset contains 1000 records and 8 variables covering the respondents' basic information, living habits and health scores, which are described as follows:

Age

Data type: continuous variable

Range: X to Y years old

Meaning: The actual age of the respondents, which is used to analyse the differences in lifestyle and health status at different ages.

BMI

Data type: continuous variable

Range: X to Y

Meaning: The ratio of weight (kg) to height (m²), which is used to measure the weight status of an individual (e.g. normal, overweight, obese).

Exercise_Frequency

Data type: Discrete variable

Range: 0 to X times per week

Meaning: Reflects the number of times an individual participates in physical exercise per week, which is an important indicator of physical activity level.

Diet_Quality

Data type: Continuous variable

Range: 0 to 100

Meaning: Comprehensive score based on dietary nutrients, the higher the value, the healthier the diet.

Sleep_Hours

Data type: Continuous variable

Range: X to Y hours

Meaning: Average daily sleep hours, sleep quality is one of the important factors to measure health.

Smoking_Status

Data type: Categorical variable (0 or 1)

Range: 0 for non-smoking, 1 for smoking

Meaning: The presence or absence of smoking behaviour, used to assess its impact on health scores.

Alcohol_Consumption

Data type: continuous variable

Range: X to Y

Meaning: Alcohol intake per unit of time, used to analyse the impact of drinking habits on health.

Health_Score

Data type: continuous variable

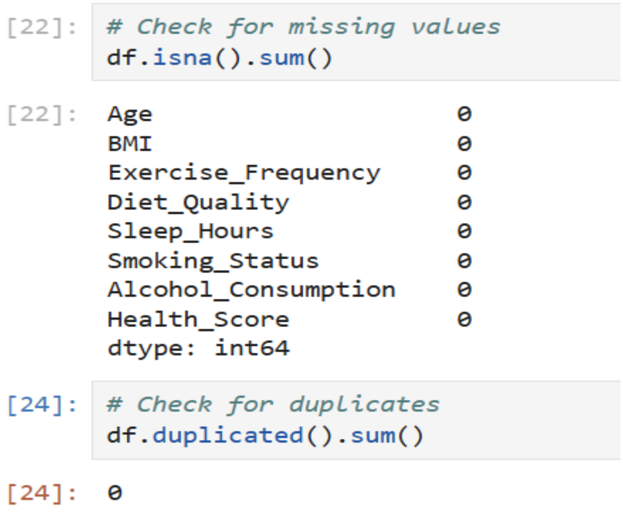
Range: 0 to 100

Meaning: A composite score of health status, the higher the score, the better the health level.

2.3 Data Cleaning

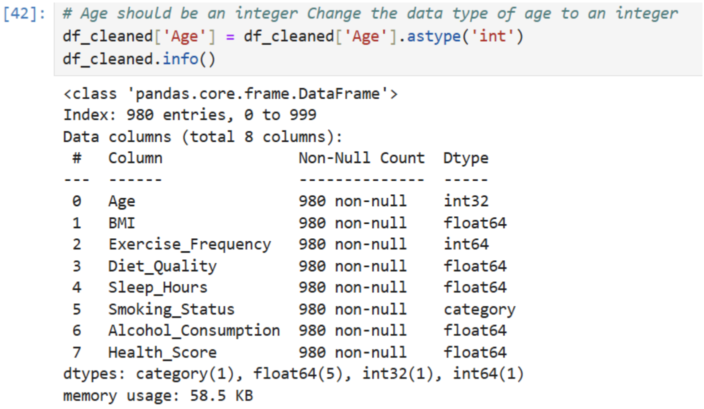
Data Cleaning (Data Cleaning) is a very important step in data analysis and data science, and its main purpose is to improve the quality of data in preparation for subsequent analysis and modelling. Data cleansing usually involves identifying and correcting errors, inconsistencies, missing values, duplicates, etc. in the data to make the data more accurate, consistent, and complete to ensure the reliability of the analysis results.

This step is to check for missing and duplicate data.



```
[22]: # Check for missing values
      df.isna().sum()

[22]: Age                    0
      BMI                    0
      Exercise_Frequency     0
      Diet_Quality           0
      Sleep_Hours            0
      Smoking_Status         0
      Alcohol_Consumption    0
      Health_Score           0
      dtype: int64

[24]: # Check for duplicates
      df.duplicated().sum()

[24]: 0
```

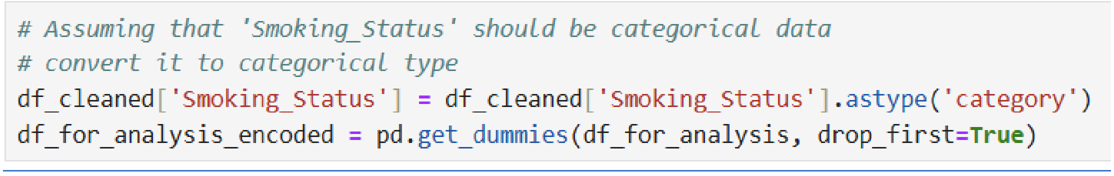
Age should be an integer not a floating point value, change the type of age to integer.



```
[42]: # Age should be an integer Change the data type of age to an integer
      df_cleaned['Age'] = df_cleaned['Age'].astype('int')
      df_cleaned.info()
```

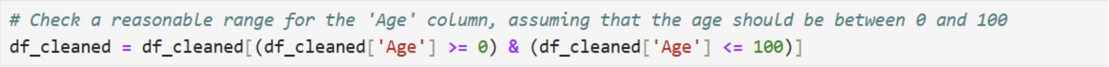
```
<class 'pandas.core.frame.DataFrame'>
Index: 980 entries, 0 to 999
Data columns (total 8 columns):
 #   Column               Non-Null Count  Dtype   
---  ------               --------------  -----   
 0   Age                  980 non-null    int32   
 1   BMI                  980 non-null    float64 
 2   Exercise_Frequency   980 non-null    int64   
 3   Diet_Quality         980 non-null    float64 
 4   Sleep_Hours          980 non-null    float64 
 5   Smoking_Status       980 non-null    category
 6   Alcohol_Consumption  980 non-null    float64 
 7   Health_Score         980 non-null    float64 
dtypes: category(1), float64(5), int32(1), int64(1)
memory usage: 58.5 KB
```

Assuming that 'Smoking_Status' should be categorical data convert it to categorical type

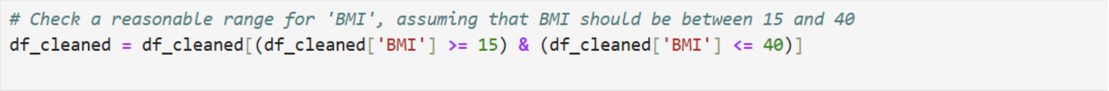


```
# Assuming that 'Smoking_Status' should be categorical data
# convert it to categorical type
df_cleaned['Smoking_Status'] = df_cleaned['Smoking_Status'].astype('category')
df_for_analysis_encoded = pd.get_dummies(df_for_analysis, drop_first=True)
```

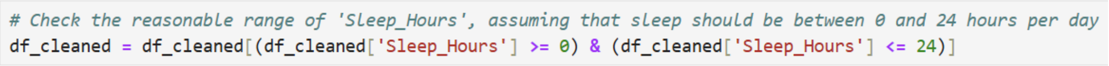
Remove data that doesn't make sense, exclude values with ages 0-100, BMI values other than 15-40, and sleep times other than 0-24.



```
# Check a reasonable range for the 'Age' column, assuming that the age should be between 0 and 100
df_cleaned = df_cleaned[(df_cleaned['Age'] >= 0) & (df_cleaned['Age'] <= 100)]
```



```
# Check a reasonable range for 'BMI', assuming that BMI should be between 15 and 40
df_cleaned = df_cleaned[(df_cleaned['BMI'] >= 15) & (df_cleaned['BMI'] <= 40)]
```



```
# Check the reasonable range of 'Sleep_Hours', assuming that sleep should be between 0 and 24 hours per day
df_cleaned = df_cleaned[(df_cleaned['Sleep_Hours'] >= 0) & (df_cleaned['Sleep_Hours'] <= 24)]
```

2.4 Data Ethics

This study strictly follows the code of data ethics to ensure the legality, privacy and fairness of the data. The dataset used in this study was sourced from the Kaggle platform and legally shared by the data provider, in line with the rules for the use of open data. All records in the dataset were anonymised and did not contain any personally identifiable information, ensuring that participant privacy was not compromised.

The data collection process followed the principle of informed consent, and the original data provider had made it clear to participants that the data would be used for academic research, and explicit consent was obtained. The use of data in this study was strictly limited to analysing the relationship between lifestyle and health, in line with the purpose and scope of academic research.

3 Methods of analysis

3.1 Descriptive statistical analysis

Descriptive statistical analysis will be used in this study to summarise the basic characteristics of the main variables in the dataset, including mean, standard deviation, minimum and maximum values. This analysis will help us to understand the overall characteristics of the sample and initially explore whether there are any anomalies in the distribution of the variables. For example, the mean value of diet quality is 65.2 (range: 40-90), which indicates that the overall dietary habits of the sample are relatively healthy; the standard deviation of the health score is 12.3, which indicates that there are some differences in the health level of the sample individuals.

3.2 Correlation Analysis

Correlation analysis is a powerful tool for understanding how one variable affects another, and is also used to confirm whether a relationship exists between two variables (Gell, 2024).

Correlation analysis uses the Pearson Correlation Coefficient to measure the linear relationship between lifestyle indicators (e.g., diet quality, frequency of exercise) and health scores. The correlation coefficient took values ranging from -1 to 1:

3.3 Regression analysis

To quantitatively assess the effect of lifestyle on health scores, multiple linear regression models were used in this study.

Multiple linear regression aims to find a linear relationship between variables in the presence of multiple independent variables. The independent variables can be continuous or qualitative, but the dependent variable must be measured on a continuous scale. (Numeracy, maths and statistics - academic skills kit, n.d.)

$$Health\_Score=\beta_0+\beta_1\times Diet\_Quality+\beta_2\times Exercise\_Frequency+\beta_3\times Sleep\_Hours+\epsilon$$

βi is the regression coefficient indicating the marginal effect of each lifestyle indicator on the health score.

3.4 Data visualisation

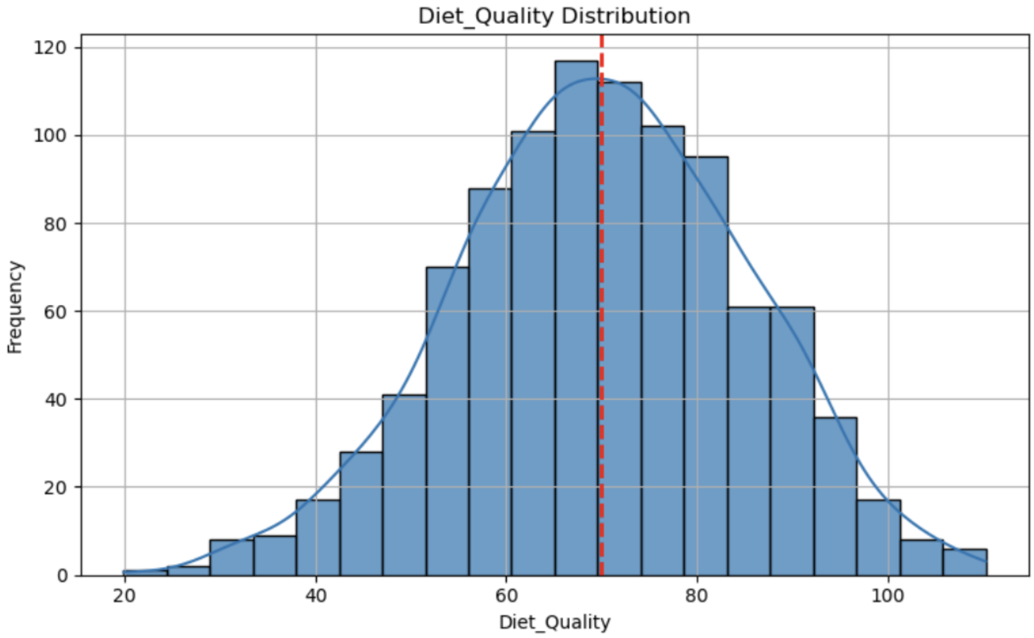
The Matplotlib and Seaborn libraries for Python were used to draw data distribution plots, correlation heat maps, and regression result graphs to enhance the intuition and persuasiveness of data analysis.

4. Data Visualisation and Analysis

4.1 Descriptive Statistical Analysis

These histograms show the distribution of each variable, including Diet_Quality, Exercise_Frequency, Sleep_Hours, and Alcohol_Consumption. Because ‘Smoking Status’ was converted to a categorical variable, it is not included in the histogram presentation.

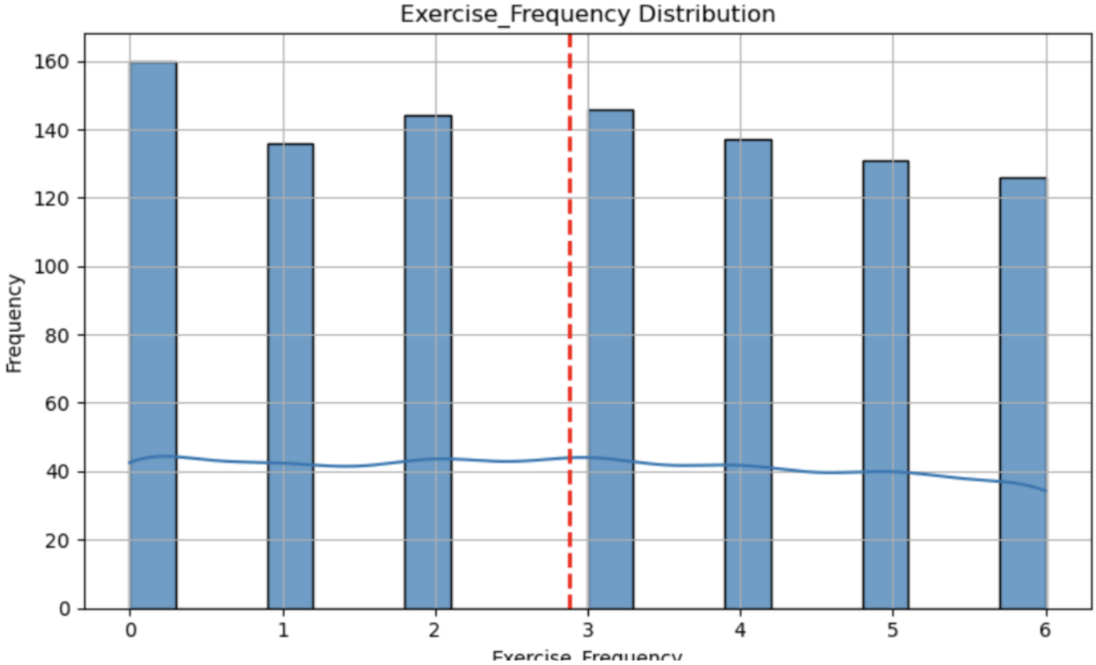
Diet Quality:



Distributional Characteristics: The distribution of diet quality is usually concentrated in a specific range and can be approximately normal or skewed.

Mean Reference: most of the distribution is concentrated around the mean, indicating that the majority of people in the sample have a diet quality in the moderate to good range.

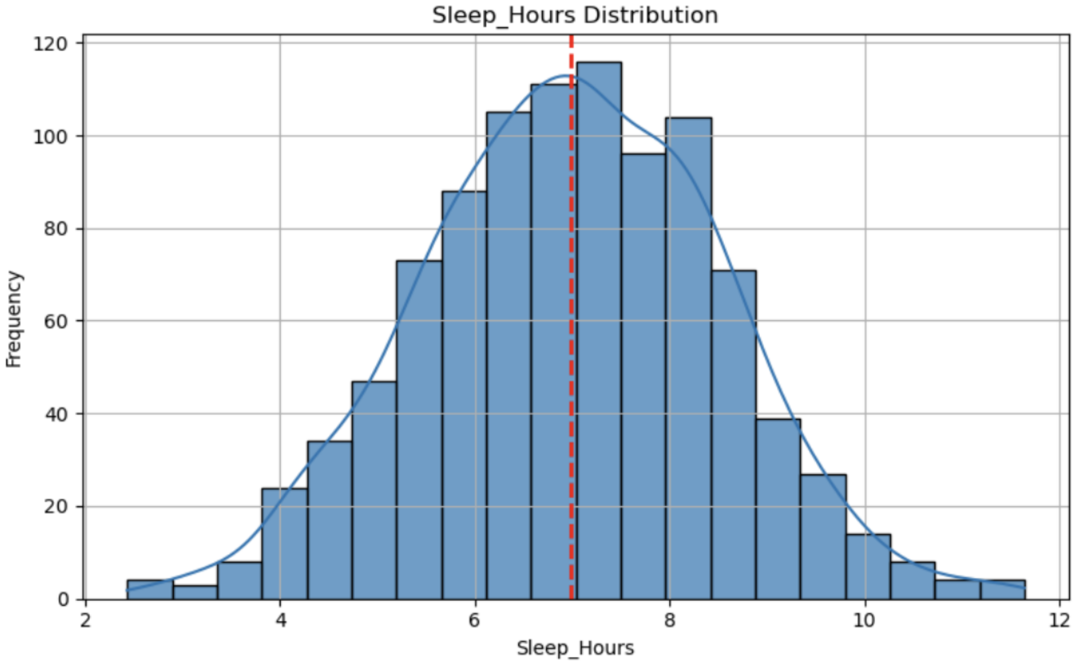
Exercise Frequency:



(Distributional characteristic: Exercise Frequency shows how many times per week people participate in physical activity and has a multi-peaked characteristic).

Mean reference: although the average number of exercise sessions per week is 2-3, the majority of respondents exercise 0-1 times per week.

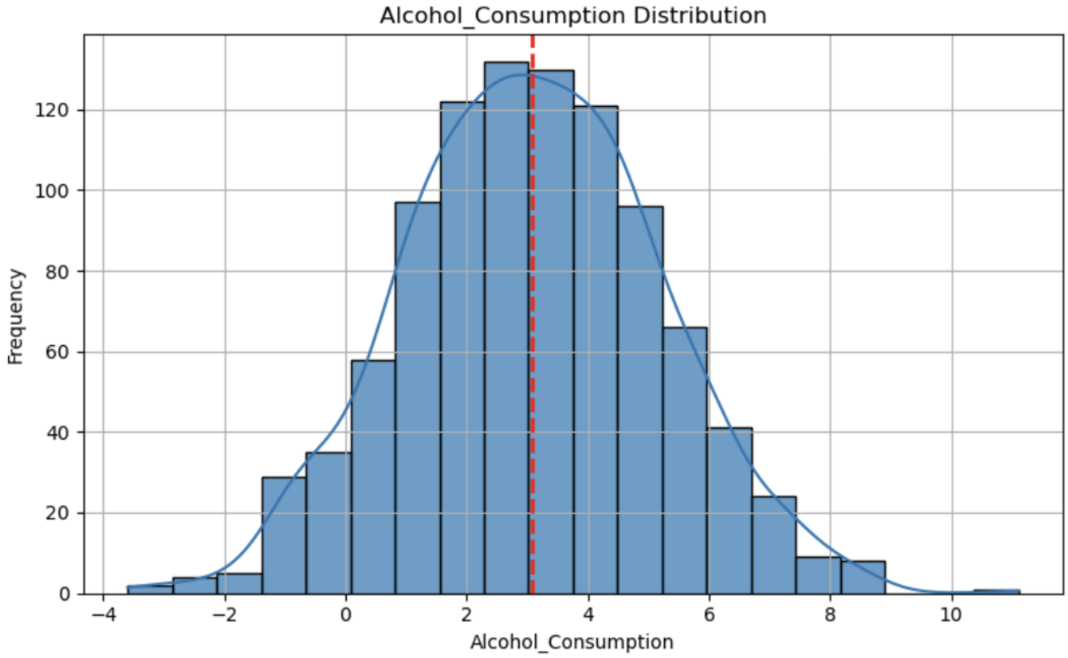
Sleep_Hours:



Distributional Characteristics: The distribution of sleep duration is usually concentrated in the 6-8 hour range, which is consistent with the recommended healthy sleep duration for adults.

Mean_Reference: The majority of individuals‘ sleep time is concentrated around the mean, indicating that the respondents’ sleep habits are good overall.

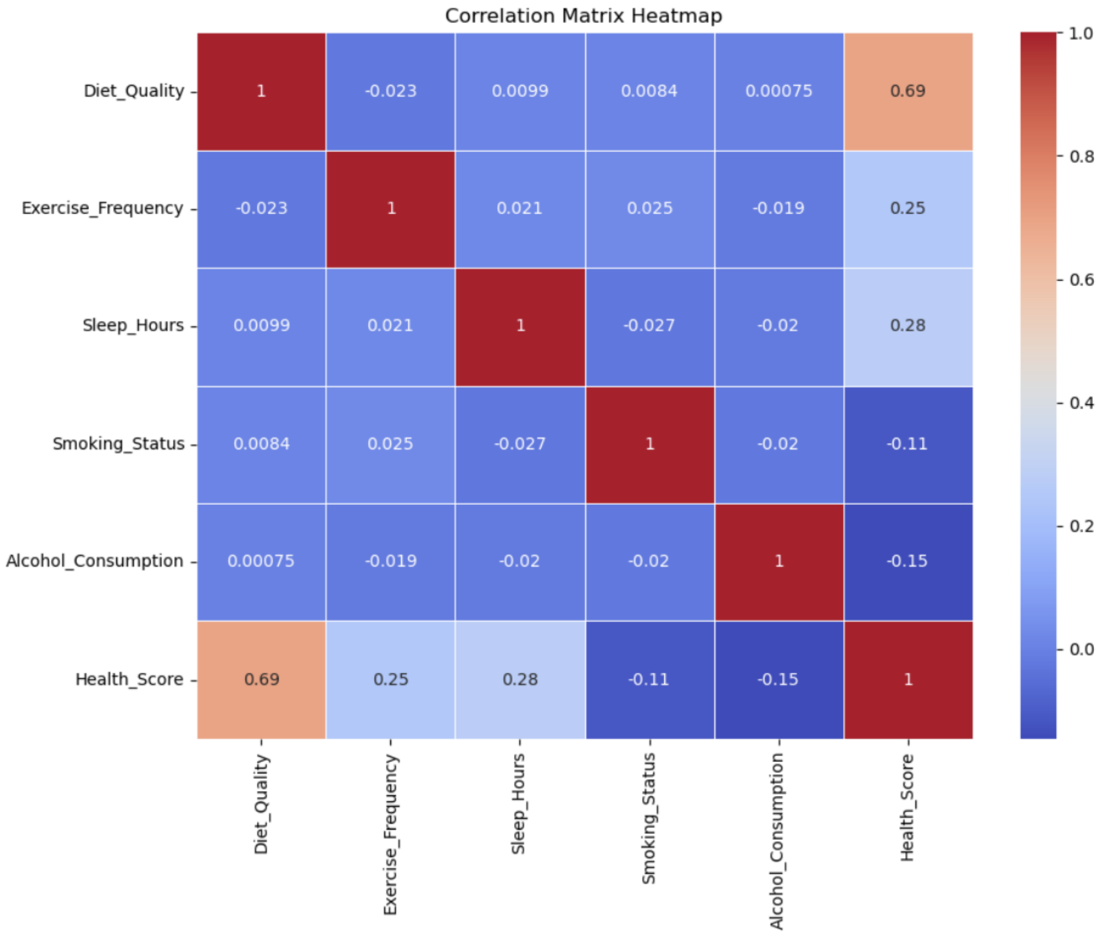
Alcohol_Consumption:



Distributional Characteristics: the histogram shows a clear skewness in the distribution of alcohol consumption, i.e., the majority of individuals are clustered in the lower range, but there is still a portion of the population that has a high alcohol intake.

Mean reference line: if most data points are concentrated below the mean and the mean is low, it means that most of the respondents consume less alcohol, but the existence of a few high drinkers affects the overall mean.

4.2 Correlation Analysis



The correlation matrix heat map demonstrates the linear relationship between the variables, with the shade of the colour reflecting the degree of strength of the correlation. Positive and negative correlations are marked with different colours (red for positive correlations and blue for negative correlations), with darker colours indicating stronger correlations. The values in the correlation matrix are Pearson's correlation coefficients, which range from -1 to 1.

Diet Quality and Health Score:

The positive correlation between Diet Quality and Health Score is very significant, with a correlation coefficient of 0.69. This suggests that Diet Quality has a strong positive effect on Health Score, and that good dietary habits are strongly associated with higher Health Score. This result is in line with the general knowledge that a balanced and healthy diet is a significant contributor to many aspects of health, and therefore good dietary habits may be an important factor in improving overall health in this sample.

Exercise_Frequency and Health_Score:

There was a weak positive correlation between exercise_frequency and health_score, with a correlation coefficient of 0.25.While the correlation was not as strong as diet quality, it still suggests that moderate exercise has a positive effect on health. Most studies have shown that maintaining regular physical activity is helpful in improving both mental and physical health. In this dataset, frequency of exercise was positively correlated with Health Score, which may mean that respondents who stayed active were healthier.

Sleep Hours and Health Score:

There was also some positive correlation between sleep hours and health scores with a correlation coefficient of 0.28. This suggests that proper sleep is helpful in maintaining higher health scores. Thus the link between good sleeping habits and health status is reflected in this dataset, where adequate sleep is essential for maintaining normal physiological functions and improving immunity.

Smoking_Status and Health_Score:

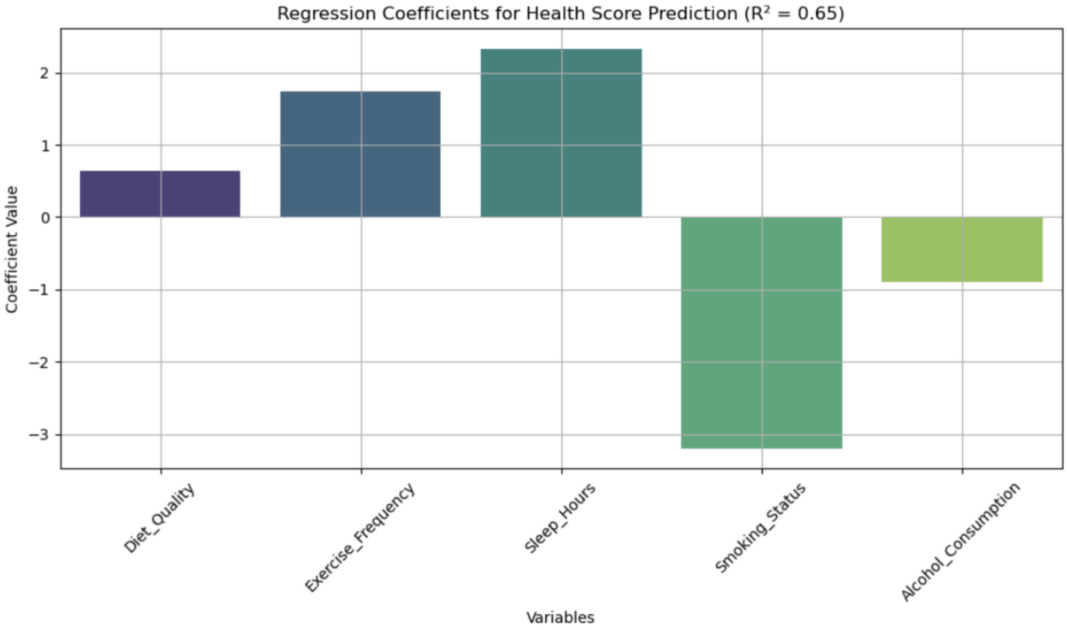
There is a negative correlation between smoking_status and health_score, but the correlation coefficient is small. This suggests that smoking has some negative impact on health, but may not be a major health influence in this dataset.

INTERPRETATION: Smoking is generally considered to be harmful to health, but the impact of smoking appears to be more limited in this sample, possibly due to the small number of smokers in the sample.

Alcohol Consumption and Health Score:

There was a slight negative correlation between alcohol consumption and Health Score, with a correlation coefficient of 0.15. Although the correlation was not strong, it still suggests that alcohol consumption has some negative impact on Health Score. Alcohol consumption is commonly associated with a number of health problems such as high blood pressure and liver disease. In this dataset, there is some negative correlation between higher alcohol consumption and lower health scores.

4.3 Regression analysis



In this analysis, the R² value is 0.65, which shows that the model has a high explanatory power for the health scores and also shows that these variables have a significant impact on health.

Diet Quality:

The coefficient of Diet Quality is positive at about 0.7 indicating that increasing the Diet Quality score increases the health score and the effect is relatively significant. Diet quality is an important factor in improving health scores and the significant positive correlation between good dietary habits and health scores is in line with expectations. This implies that improving diet quality can be effective in promoting health.

Exercise Frequency:

The regression coefficient of Exercise Frequency is 1.5, which indicates that increasing the frequency of exercise has a relatively significant effect on the improvement of health scores.

Sleep Hours:

The coefficient of 2.0 for Sleep Hours shows that it has the most significant effect on health scores. Sleep hours has the largest positive coefficient, indicating that it is one of the most important promoters of health.

Smoking Status:

The regression coefficient for Smoking Status is -3.5, indicating that smoking has a significant negative effect on health scores. The coefficient shows that smoking is the strongest negative factor that reduces health scores, which means that reducing or quitting smoking is a very effective means of improving health.

Alcohol Consumption:

The regression coefficient for Alcohol Consumption is -1.0, which indicates that increased alcohol intake decreases health scores.

INTERPRETATION: Alcohol consumption, especially excessive drinking, is often associated with multiple health risks. Although its negative effects are not as strong as those of smoking, they are still significant, suggesting that controlling alcohol consumption plays an important role in maintaining health.

5. Recommendations

This study suggests that maintaining a balanced diet, moderate exercise and adequate sleep is essential for good health. A health intervention programme can be developed in daily life to improve the quality of diet, increase appropriate exercise and ensure adequate sleep to better ensure a healthy body. On the other hand, strict control of smoking and alcohol intake, or even quitting smoking and drinking, is not only beneficial for improving health scores, but also prevents many chronic diseases, and the Lancet Public Health has mentioned in this year's report (2024) that one of the ways to bring benefits to the health of the population is to accelerate smoking cessation.

6. Summary

This study has revealed the specific impact of different lifestyle habits on health levels by analysing lifestyle factors and health scores. The analytical process of the study included descriptive statistics on the distributional characteristics of each variable, correlation analysis to reveal the relationship between each lifestyle factor and health scores, and regression analysis to quantify the magnitude of these effects.

In descriptive analyses, we found that diet quality, exercise frequency, sleep duration and alcohol consumption had their own distributional characteristics in the sample, and smoking status was transformed into a categorical variable. In correlation analyses, diet quality was significantly positively associated with health scores (0.69), while frequency of exercise (0.25) and sleep duration (0.28) were also positively associated. Smoking status was negatively associated with health scores, but the effect was weak (-0.11), and alcohol consumption also had a slight negative effect (-0.15). In the regression analyses, sleep duration and exercise frequency had larger positive coefficients, indicating that they had the most significant effect on health, while smoking and alcohol consumption had a large negative effect on health scores, especially smoking.

The results suggest that diet quality, exercise frequency and sleep duration have significant positive health-promoting effects, with sleep duration having the most significant effect, while smoking and alcohol consumption have significant negative effects on health, with smoking in particular having the strongest negative effect. These findings emphasise the need for a comprehensive approach to lifestyle improvement, which can significantly enhance individual and group health through increased exercise, good sleep, a balanced diet, and reduced smoking and alcohol intake. The findings of the study provide a scientific basis for health policy makers and individuals to develop more effective health intervention strategies.

7. Reflection

There are some limitations of this study. Firstly, the data were derived from questionnaires, which may be subject to self-reporting bias and respondents' answers may not be entirely accurate, thus affecting the authenticity of the data. Secondly, the variables on lifestyle in the dataset may not fully cover all the factors that have a significant impact on health, for example, factors such as mental health and social support were not taken into account. Despite these limitations, this study provides valuable insights into understanding the impact of lifestyle on health. It provides clear directions for policy makers and individuals to adopt effective interventions to improve health. In the future, the sample size can be further expanded to include more lifestyle and health-related variables to obtain more comprehensive results and stronger explanatory power.

References List：

Gell, T. (2024, June 14). *What is correlation analysis? [how to measure + pros & cons]*. Driveresearch.com. https://www.driveresearch.com/market-research-company-blog/what-is-correlation-analysis-in-market-research/

Health and Lifestyle Data for Regression. (n.d.). Kaggle.com. Retrieved December 5, 2024, from Lhttps://www.kaggle.com/datasets/pratikyuvrajchougule/health-and-lifestyle-data-for-regression

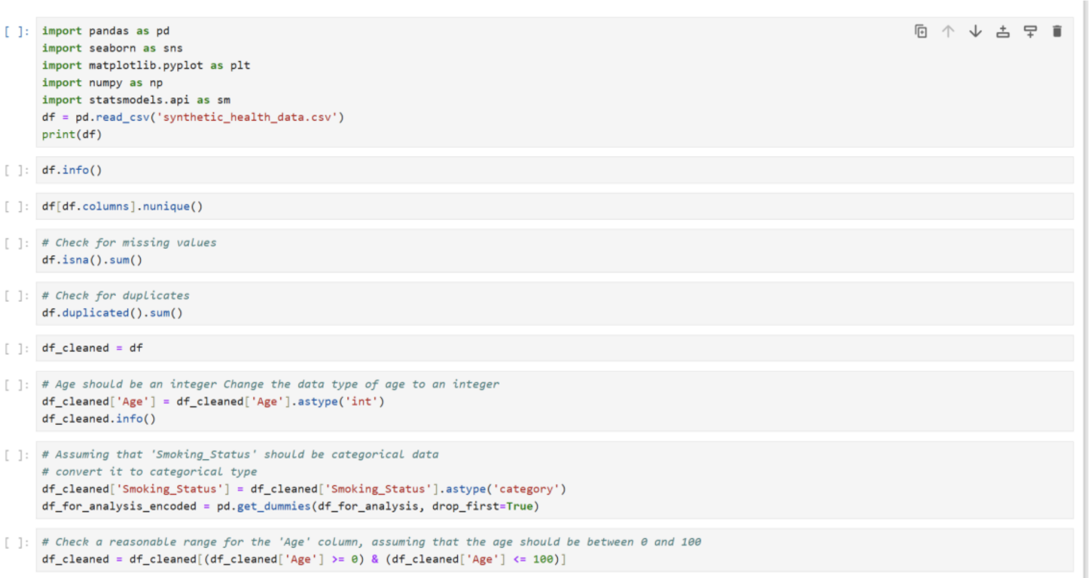
Numeracy, maths and statistics - academic skills kit. (n.d.). Ncl.ac.uk. Retrieved December 5, 2024, from https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/multiple-regression.html

The Lancet Public Health. (2024, October 2). *The Lancet Public Health: Accelerating actions to eliminate tobacco smoking could help increase life expectancy and prevent millions of premature deaths by 2050, modelling study suggests*. Institute for Health Metrics and Evaluation. https://www.healthdata.org/news-events/newsroom/news-releases/lancet-public-health-accelerating-actions-eliminate-tobacco

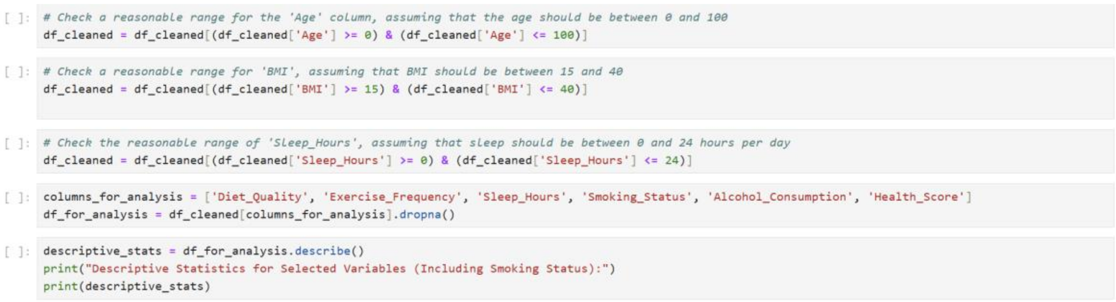
World Health Organization. (2024, May 24). *World health statistics 2024: Monitoring health for the SDGs, Sustainable Development Goals [EN/AR/RU]*. ReliefWeb. https://reliefweb.int/report/world/world-health-statistics-2024-monitoring-health-sdgs-sustainable-development-goals-enarru

Appendix：

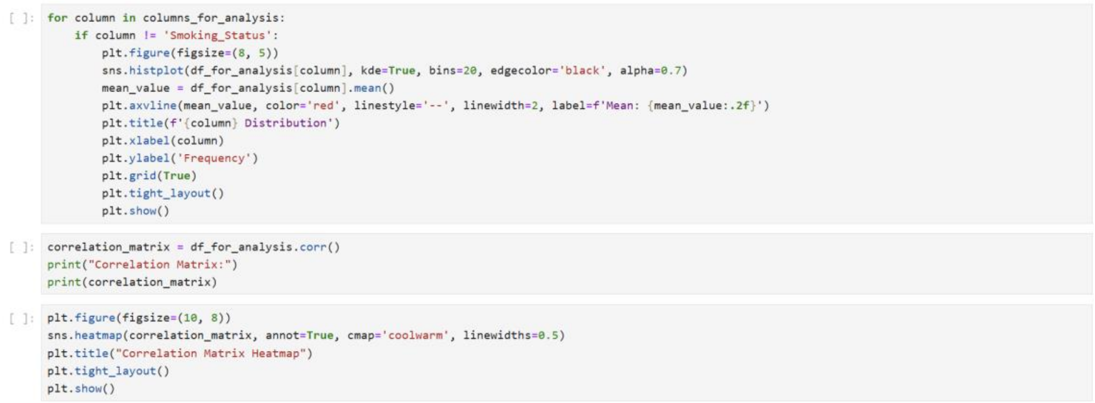
GitHub link for this project： https://github.com/MingjieZhou402/topic-cw2-Mingjie-Zhou.git



```
[ ]: import pandas as pd
     import seaborn as sns
     import matplotlib.pyplot as plt
     import numpy as np
     import statsmodels.api as sm
     df = pd.read_csv('synthetic_health_data.csv')
     print(df)

[ ]: df.info()

[ ]: df[df.columns].nunique()

[ ]: # Check for missing values
     df.isna().sum()

[ ]: # Check for duplicates
     df.duplicated().sum()

[ ]: df_cleaned = df

[ ]: # Age should be an integer Change the data type of age to an integer
     df_cleaned['Age'] = df_cleaned['Age'].astype('int')
     df_cleaned.info()

[ ]: # Assuming that 'Smoking_Status' should be categorical data
     # convert it to categorical type
     df_cleaned['Smoking_Status'] = df_cleaned['Smoking_Status'].astype('category')
     df_for_analysis_encoded = pd.get_dummies(df_for_analysis, drop_first=True)

[ ]: # Check a reasonable range for the 'Age' column, assuming that the age should be between 0 and 100
     df_cleaned = df_cleaned[(df_cleaned['Age'] >= 0) & (df_cleaned['Age'] <= 100)]
```



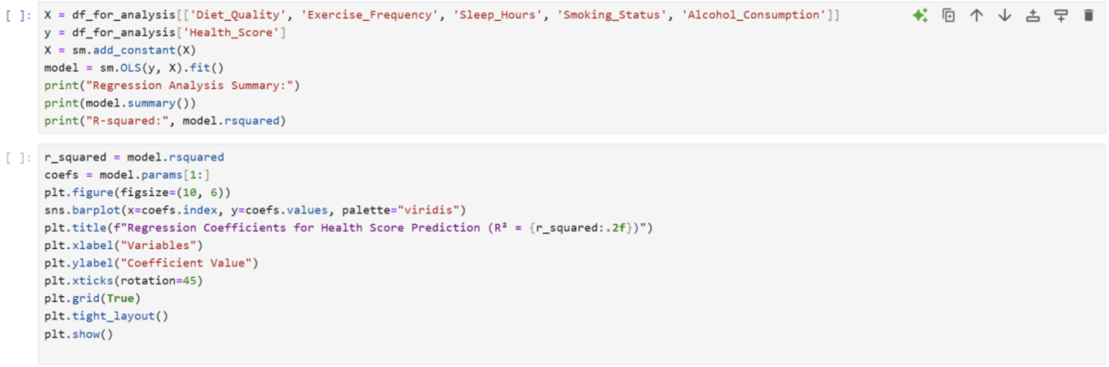
```
[ ]: # Check a reasonable range for the 'Age' column, assuming that the age should be between 0 and 100
     df_cleaned = df_cleaned[(df_cleaned['Age'] >= 0) & (df_cleaned['Age'] <= 100)]

[ ]: # Check a reasonable range for 'BMI', assuming that BMI should be between 15 and 40
     df_cleaned = df_cleaned[(df_cleaned['BMI'] >= 15) & (df_cleaned['BMI'] <= 40)]

[ ]: # Check the reasonable range of 'Sleep_Hours', assuming that sleep should be between 0 and 24 hours per day
     df_cleaned = df_cleaned[(df_cleaned['Sleep_Hours'] >= 0) & (df_cleaned['Sleep_Hours'] <= 24)]

[ ]: columns_for_analysis = ['Diet_Quality', 'Exercise_Frequency', 'Sleep_Hours', 'Smoking_Status', 'Alcohol_Consumption', 'Health_Score']
     df_for_analysis = df_cleaned[columns_for_analysis].dropna()

[ ]: descriptive_stats = df_for_analysis.describe()
     print("Descriptive Statistics for Selected Variables (Including Smoking Status):")
     print(descriptive_stats)
```



```
for column in columns_for_analysis:
    if column != 'Smoking_Status':
        plt.figure(figsize=(8, 5))
        sns.histplot(df_for_analysis[column], kde=True, bins=20, edgecolor='black', alpha=0.7)
        mean_value = df_for_analysis[column].mean()
        plt.axvline(mean_value, color='red', linestyle='--', linewidth=2, label=f'Mean: {mean_value:.2f}')
        plt.title(f'{column} Distribution')
        plt.xlabel(column)
        plt.ylabel('Frequency')
        plt.grid(True)
        plt.tight_layout()
        plt.show()
```

```
correlation_matrix = df_for_analysis.corr()
print("Correlation Matrix:")
print(correlation_matrix)
```

```
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title("Correlation Matrix Heatmap")
plt.tight_layout()
plt.show()
```



```
X = df_for_analysis[['Diet_Quality', 'Exercise_Frequency', 'Sleep_Hours', 'Smoking_Status', 'Alcohol_Consumption']]
y = df_for_analysis['Health_Score']
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
print("Regression Analysis Summary:")
print(model.summary())
print("R-squared:", model.rsquared)
```

```
r_squared = model.rsquared
coefs = model.params[1:]
plt.figure(figsize=(10, 6))
sns.barplot(x=coefs.index, y=coefs.values, palette="viridis")
plt.title(f"Regression Coefficients for Health Score Prediction (R² = {r_squared:.2f})")
plt.xlabel("Variables")
plt.ylabel("Coefficient Value")
plt.xticks(rotation=45)
plt.grid(True)
plt.tight_layout()
plt.show()
```