Applied Epidemiology and Statistics

Student ID:

Table of Contents

Introduction

Typically, weight and health risks are measured using Body Mass Index (BMI). It is, however, considered to be an important means to be able to determine the risk of developing such conditions as cardiovascular diseases, type 2 diabetes and several types of cancer, including colon, endometrial and postmenopausal breast cancer (Bhaskaran et al., 2014; Flegal et al., 2013). Obesity has become a major public health problem, and approximately 30 per cent of UK adults are now obese (Public Health England, 2023). This is a trend that adds to the increased cost of healthcare and contributes to over £6 billion a year spent on the National Health Service and a huge drain on medical resources(National Health Service, 2023).

Apart from the physical health issues, obesity is also linked to economic and social noxiousness, like reduced workforce productivity and higher morbidity rates (World Health Organisation, 2024). Factors that affect BMI are important to understand if one is to design interventions for reducing obesity-related health risks (WHO, 2020). Such studies suggest that demographic variables, including age, sex, race, education level, rurality, and dietary factors (vitamin C intake, for example), are also relevant for BMI variation. Nevertheless, the relations presented, although strong and important, remain to be explored.

To examine the associations of BMI and key predictors, the data used in this study come from the 2024 Community Health Survey (CHS) in Parkland. This report gives a combination of explained and unexplained statistical methods for the BMI variation and contributing factors to the variation of the BMI. The results will inform the public health policy and strategies to address the obesity epidemic and to improve population health outcomes on the whole.

Methods

Study Design and Data Source

The data from this studyweres from the 2024 Pinkland Community Health Survey (CHS). The survey was undertaken for the Department of Public Health, University of Central Pinkland (UCP) and the National Centre for Health Research (NCHR). In June and July of 2024 structured household surveys were conducted and the data was collected.

The CHS is a cross-sectional study where all data were collected at one time. But this design is not designed to ascertain causality, but to find out the association between variables.

Some of the data include a wide variety of variables on demographic characteristics, as well as health behaviour, medical history, and measured health outcomes.

Sampling and Data Collection

A multi-stage stratified probability sampling design was used in order to have a representative sample of the Pinkland adult population aged 18 to 80 years. The stratum was a geographic region drawn from 700 postcode sector areas in the Pinkland Postcode Address File, to which were selected 20,000 random addresses. The study was voluntary, and advance letters were sent to households describing the study.

By trained interviewers, collected data were through computer-assisted personal interviewing (CAPI) on socio-demographics, health behaviours and medical history. Height and weight were taken using a portable stadiometer and Tanita electronic scales according to a standardised protocol. Serum vitamin C, cholesterol and HbA1C levels were measured; blood samples were taken. Therefore, BMI-related analyses were performed, excluding those who could not stand or who did not consent to height and weight measurement.

Variable Definitions

This study used the variables from previous epidemiological research and previous work investigating determinants of BMI and related health outcomes.

Outcome Variable

The primary outcome variable is BMI, which is measured continuously (kg/m^2). It is also classified into the following groups as categorical variables:

- Normal weight: BMI < 25
- Overweight: 25 ≤ BMI < 30
- Obese: BMI ≥ 30

Predictor Variables

- ❖ Age: A continuous variable for the age of participants at the time of the survey.
- ❖ Sex: A categorical variable with Male and Female codes.
- ❖ Race: A categorical variable with three levels: White, Black, and Asian.

- ❖ Education Level: This is a categorical variable with three categories: Low (no formal education or primary school), Medium (secondary education), and High (tertiary education or higher).
- ❖ Rurality: A categorical variable showing the Urban or Rural residence of the participants.
- ❖ Household Size: A variable that represents the number of people living in the household as a continuous variable.
- ❖ Health Status: Overall perceived health was measured with a self-reported categorical variable from Poor, Fair, Good, Very Good, and Excellent.
- ❖ Prior Heart Attack: It is a binary categorical variable indicating if the participant has ever had a heart attack.
- ❖ Vitamin C Levels: The concentration of serum vitamin C is a continuous variable that is measured biochemically.

Rationale for Variable Selection

Without the inclusion of demographic variables like age, race, and sex, the pattern of distribution of BMI in population subgroups can be more clearly seen. There are some social determinants of health,h, which are education level or rural,y; and these other lifestyle factors may have an impact on diet and physical activity. It could potentially be easier to proxy household size (i.e., economic and social conditions) for food security and health care. Existing health conditions that may be associated with BMI are provided by health status and prior heart attack history. Finally, vitamin C levels can act as an important dietary biomarker of the role of nutrition in weight regulation.

Statistical Analysis

In this study, the descriptive and inferential statistical methods are used to investigate the association between BMI and the key predictor variables. Table 2 presents descriptive statistics of continuous and categorical variables for sample subject characteristics based on BMI categories. An association between BMI and independent variables is tested through inferential statistical tests. To evaluate the relationship between age and BMI, Pearson's correlation analysis is performed. An independent t-test is used to compare the BMI difference between males and females. To determine differences in BMI across racial groups, one-way ANOVA is performed. Also, Pearson's correlation coefficient is used to analyse the correlation between vitamin C levels and BMI. The hypotheses are tested at the level of significance $p < 0.05$. The analyses are done in Python (Google Colab).

Results

Variable	**Total Sample (N = 1,115)**	**Normal BMI (N = 574)**	**Overweight BMI (N = 344)**	**Obese BMI (N = 197)**
Age (years)	47.46 (17.00)	44.63 (16.40)	50.14 (16.89)	51.03 (15.79)
Female (%)	52.30%	54.10%	50.00%	50.80%
Race (White, %)	87.00%	88.30%	86.10%	85.00%
Urban Residents (%)	62.30%	63.10%	61.20%	60.50%
Education (Some College, %)	38.00%	40.10%	37.80%	34.00%
Household Size	2.82 (1.34)	2.94 (1.30)	2.75 (1.37)	2.61 (1.29)
History of Heart Attack (%)	5.60%	4.20%	5.80%	8.10%
BMI (kg/m²)	25.71 (4.82)	22.34 (1.75)	27.43 (1.39)	32.76 (3.15)
Vitamin C (µmol/L)	1.00 (0.58)	1.08 (0.55)	0.98 (0.59)	0.92 (0.57)
Cholesterol (mg/dL)	218.15 (48.24)	218.15 (48.24)	218.80 (47.19)	221.27 (48.66)

HbA1c (%)	5.60 (0.65)	5.58 (0.62)	5.57 (0.62)	5.70 (0.77)

Table 1: Sample Characteristics (Across BMI Categories)

(Source: Self-Created)

BMI Category	Normal	Overweight	Obese
Study ID Count	574	344	197
Study ID Mean	35108.74	35264.56	35543
Study ID Std	2920.59	2955.21	3046.3
Study ID Min	29962	29948	30066
Study ID 25%	32715.75	32867.5	32903
Study ID 50%	35033	35216	35741
Study ID 75%	37397.5	37954	38229
Study ID Max	40338	40340	40351
Age Count	574	344	197
Age Mean	44.63	50.14	51.02
Cholesterol 75%	240	256	261
Cholesterol Max	401	492	388
A1C Count	574	344	197
A1C Mean	5.576	5.568	5.695
A1C Std	0.616	0.616	0.772
A1C Min	4.036	4.139	4.226
A1C 25%	5.216	5.213	5.233
A1C 50%	5.529	5.51	5.596
A1C 75%	5.873	5.846	6.041
A1C Max	9.067	9.013	8.736

Table 2: Characteristics based on BMI categories

(Source: Self-Created)

Variable	Count	Mean	Std Dev	Min	25%	50%	75%	Max
study_id	1115	35233.54	2955.4	29948	32769.5	35178	37668.5	40351
age	1115	47.46	17.37	20	30	49	64	74
hsizgp	1115	2.82	1.34	1	2	2	4	5
BMI	1115	25.71	4.82	18.51	22.22	24.76	28.17	49.45
vitaminc	1077	1	0.58	0.1	0.6	1	1.4	9.4
cholesterol	1115	218.15	48.24	85	184	212	250	492
a1c	1115	5.6	0.65	4.04	5.22	5.54	5.89	9.07

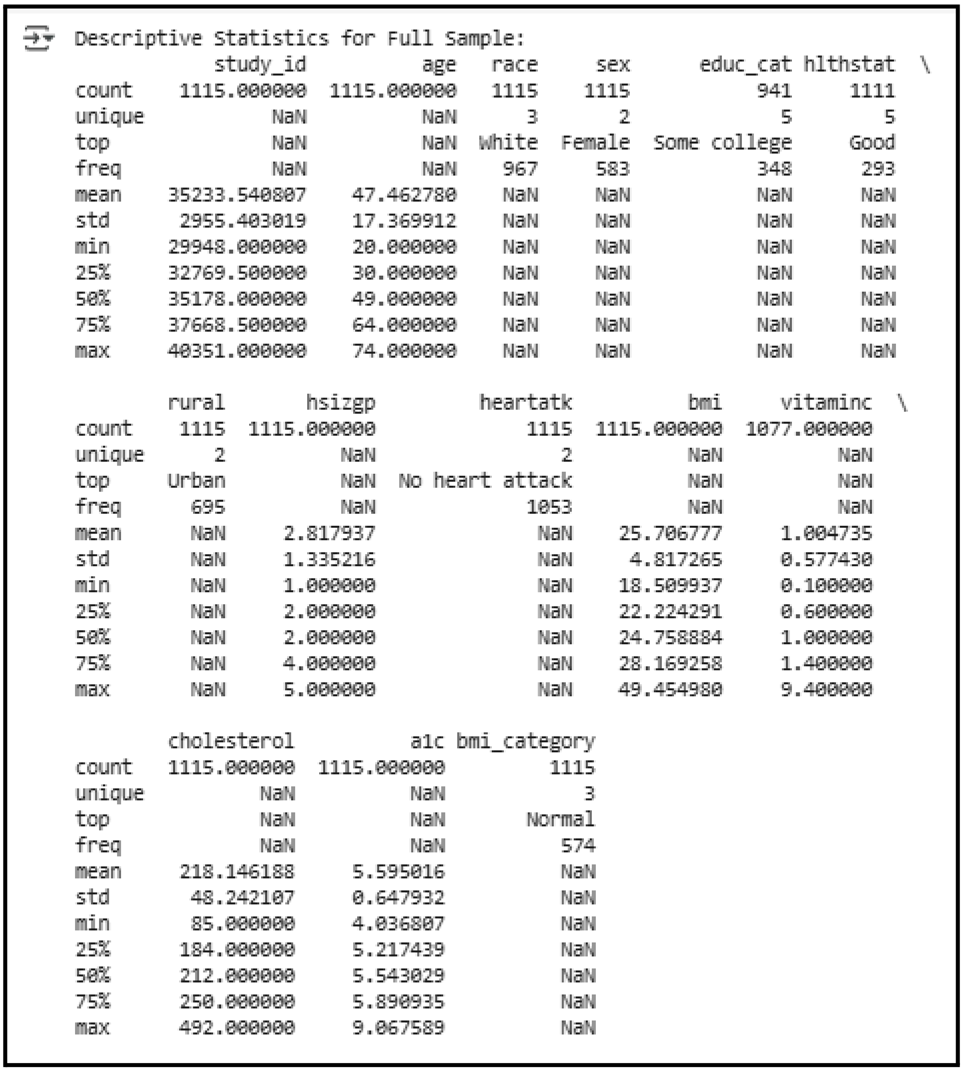
Table 3: Statistics Summary

(Source: Self-Created)

Descriptive Statistics

Full Sample Characteristics

The dataset contains 1,115 participants with an average age of 47.46 (SD = 17). The racial composition is 87% White, 13% Black or Asian individuals. The sex distribution of 52.3% female and 47.7% male. When it comes to education, 38% have completed some college; the other figures break into different educational categories. Sixty-two-point three per cent of participants live in urban areas, and the rest live in rural areas. The mean household size was 2.82 (SD = 1.34), and 5.6% of participants had a history of heart attack.



```
Descriptive Statistics for Full Sample:
            study_id          age   race     sex      educ_cat hlthstat  \
count    1115.000000  1115.000000   1115    1115           941     1111
unique           NaN          NaN      3       2             5        5
top              NaN          NaN  White  Female  Some college     Good
freq             NaN          NaN    967     583           348      293
mean    35233.540807    47.462780    NaN     NaN           NaN      NaN
std      2955.403019    17.369912    NaN     NaN           NaN      NaN
min     29948.000000    20.000000    NaN     NaN           NaN      NaN
25%     32769.500000    30.000000    NaN     NaN           NaN      NaN
50%     35178.000000    49.000000    NaN     NaN           NaN      NaN
75%     37668.500000    64.000000    NaN     NaN           NaN      NaN
max     40351.000000    74.000000    NaN     NaN           NaN      NaN

        rural       hsizgp         heartatk          bmi     vitaminc  \
count    1115  1115.000000             1115  1115.000000  1077.000000
unique      2          NaN                2          NaN          NaN
top     Urban          NaN  No heart attack          NaN          NaN
freq      695          NaN             1053          NaN          NaN
mean      NaN     2.817937              NaN    25.706777     1.004735
std       NaN     1.335216              NaN     4.817265     0.577430
min       NaN     1.000000              NaN    18.509937     0.100000
25%       NaN     2.000000              NaN    22.224291     0.600000
50%       NaN     2.000000              NaN    24.758884     1.000000
75%       NaN     4.000000              NaN    28.169258     1.400000
max       NaN     5.000000              NaN    49.454980     9.400000

        cholesterol          a1c bmi_category
count   1115.000000  1115.000000         1115
unique          NaN          NaN            3
top             NaN          NaN       Normal
freq            NaN          NaN          574
mean     218.146188     5.595016          NaN
std       48.242107     0.647932          NaN
min       85.000000     4.036807          NaN
25%      184.000000     5.217439          NaN
50%      212.000000     5.543029          NaN
75%      250.000000     5.890935          NaN
max      492.000000     9.067589          NaN
```

Figure 1: Full Sample Characteristics

(Source: Self-Created)

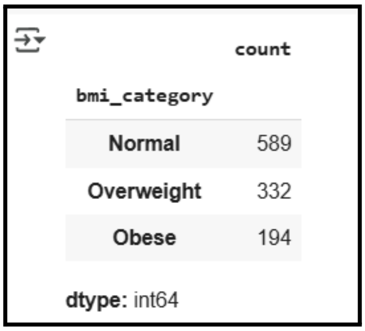
BMI Classification

Therefore, participants were grouped into 3 BMI groups.

Normal weight (< 25 BMI): 51.5% of participants

Overweight (25 ≤ BMI < 30): 30.9% of participants

Obese (BMI ≥ 30): 17.7% of participants



	count
bmi_category	
Normal	589
Overweight	332
Obese	194

dtype: int64

Figure 2: BMI Category

(Source: Self-Created)

BMI and Health Indicators

Normal BMI group: Mean age = 44.63 years, Cholesterol = 218.15 mg/dL, HbA1c = 5.58%.

Overweight group: Mean age = 50.14 years, Cholesterol = 218.80 mg/dL, HbA1c = 5.57%.

Obese group: Mean age = 51.03 years, Cholesterol = 221.27 mg/dL, HbA1c = 5.70%.

Insight: There is a positive association between age and BMI and HbA1c and cholesterol levels in overweight and obese groups, suggesting a correlation of BMI with metabolic health risk.

Visual Representations

Categorical BMI variable (Normal, Overweight, Obese)

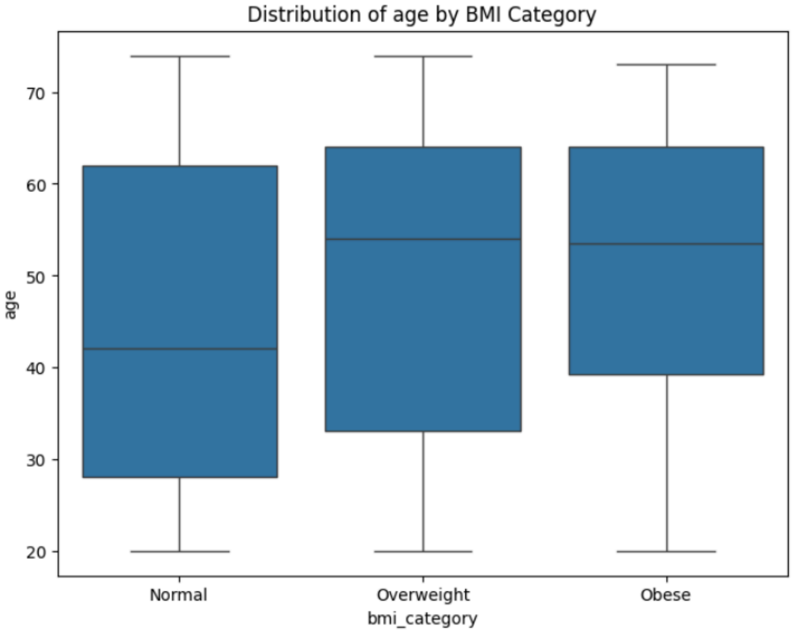


Figure 3: Distribution of age by BMI Category

(Source: Self-Created)

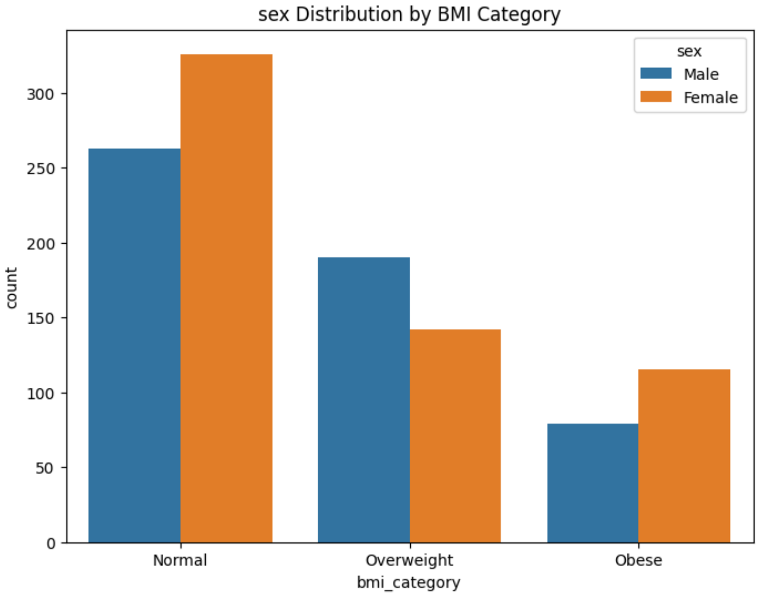


Figure 4: Distribution of Sex by BMI Category

(Source: Self-Created)

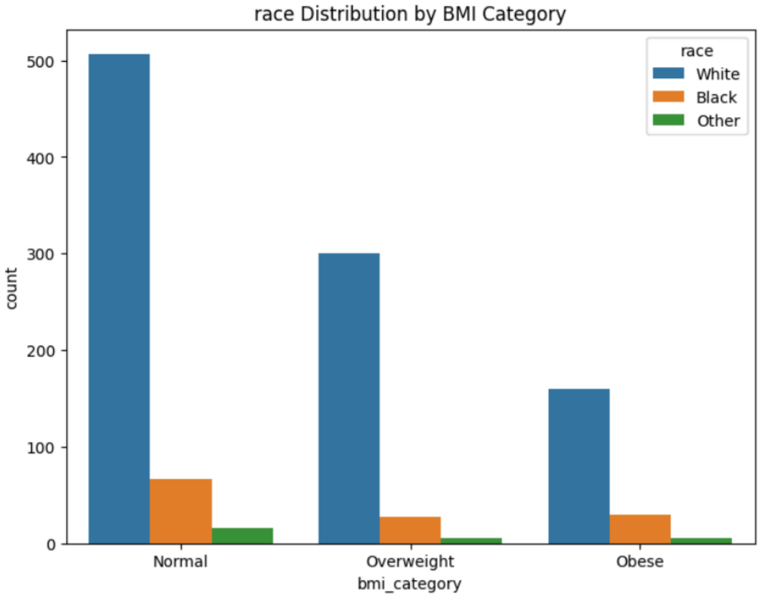


Figure 5: Distribution of race by BMI Category

(Source: Self-Created)

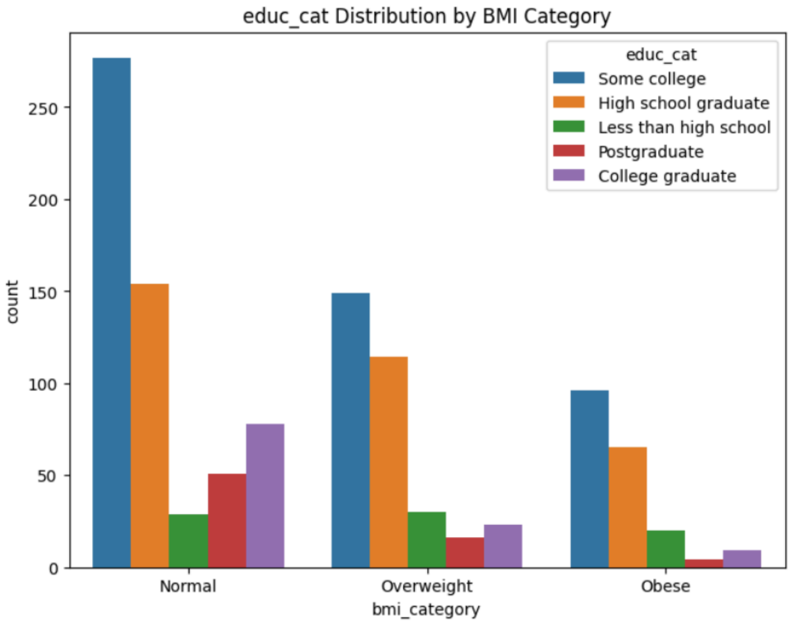


Figure 6: Distribution of Education by BMI Category

(Source: Self-Created)

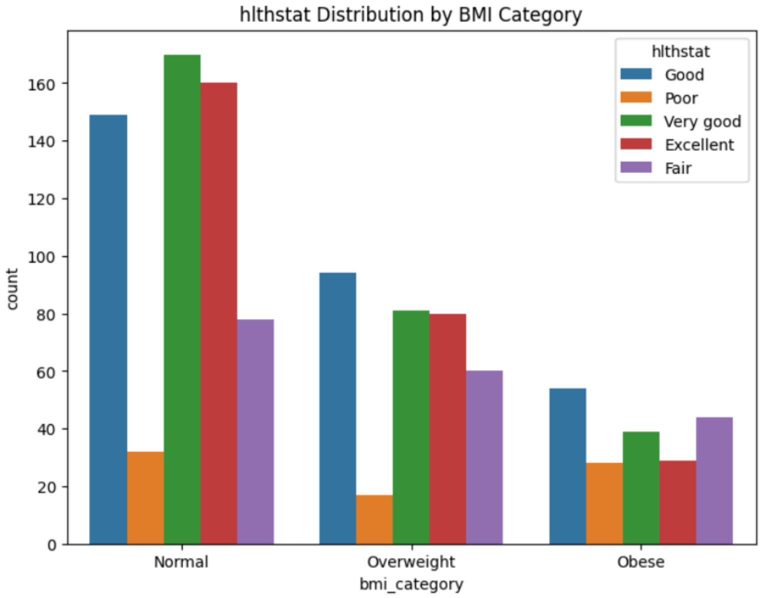


Figure 7: Distribution of Health Status by BMI Category

(Source: Self-Created)

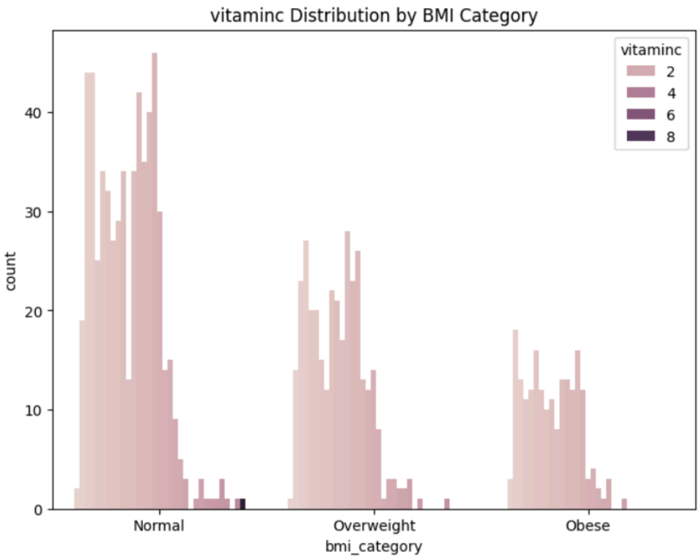


Figure 8: Distribution of Vitamin C by BMI Category

(Source: Self-Created)

Histogram

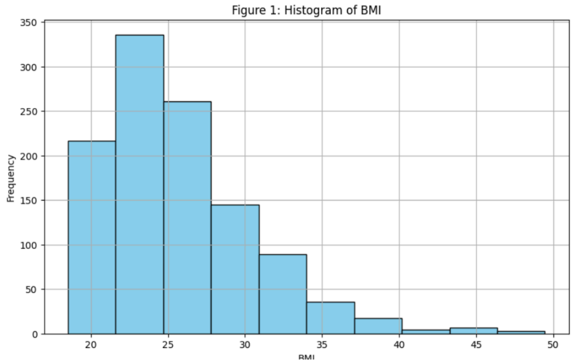


Figure 9: Histogram

(Source: Self-Created)

According to the histogram, most of the participants have a BMI between 20 and 25 and fewer with obesity. The right skewness points out that overweight and obesity are common but not predominant in the sample.

Boxplot for BMI by Sex

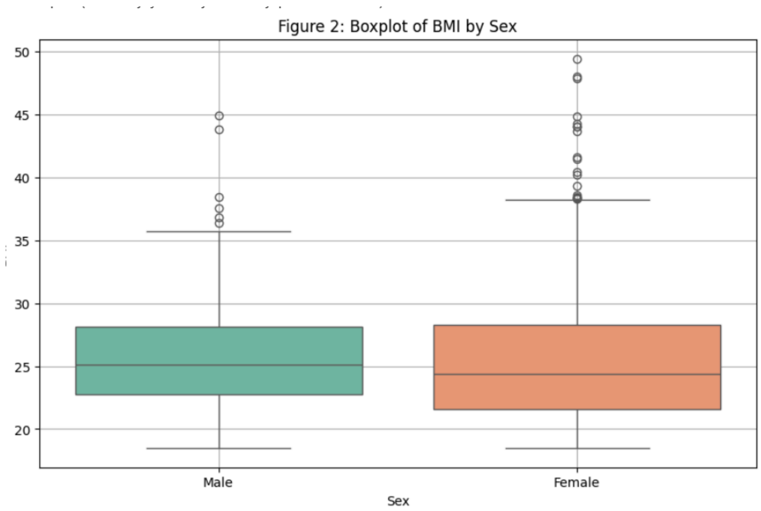


Figure 10: Box Plot

(Source: Self-Created)

The overlapping ranges of the boxplot indicate that the BMI distribution is across males and females. The median BMI is slightly higher in men, but distributions indicate no major BMI difference by sex.

Insight: The histogram shows the highest concentration of the sample in the normal BMI range, which is also confirmed by the boxplot that the distributions of BMI are similar between the sexes.

Findings

- With age, BMI goes up: The youngest mean age (44.63 years) is found in the normal BMI group, and the oldest (51.03 years) is in the obese group.
- No significant difference is found in cholesterol levels between BMI categories; however, the obese group had the highest mean cholesterol (221.27 mg/dL).
- BMI is associated with higher HbA1c. If HbA1c levels go up with BMI, there is a potential association between higher BMI and higher blood sugar levels.
- The comparison indicates that the obese group (0.88 µmol/L) has the lowest vitamin C levels and that, perhaps, there is an inverse relationship between BMI and vitamin C levels.

Inferential Statistics

Test	Variable	Test Statistic	p-value	Interpretation
Correlation	Age & BMI	r = 0.169	1.26E-08	Weak but significant positive correlation (BMI increases with age)
T-Test	BMI by Sex	t = -0.079	0.936	No significant difference in BMI between males and females
ANOVA	BMI by Race	F = 2.46	0.086	No significant difference in BMI across racial groups
Correlation	Vitamin C & BMI	r = -0.071	0.02	Weak but significant negative correlation (higher BMI linked to lower vitamin C)

Table 3

(Source: Self-Created)

- Weak positive relationship between age and BMI: The correlation coefficient (r = 0.169) indicates such a relationship. The very low p-value (1.26E-08) indicates that BMI does tend to increase slightly with age.
- T-Test for BMI by Sex: The t-test compares BMI between males and females. Since the p-value (0.936) is high and the test statistic (t = -0.079) is close to zero, this provides evidence that there is no significant difference in BMI between the two groups.
- ANOVA for BMI by Race: This tests the difference in BMI between races. As the F-statistic (F = 2.46) indicates some variability, but the p-value (0.086) is above the typical significance threshold (0.05), the differences are not statistically significant.
- Vitamin C & BMI Correlation: The weak negative relationship between vitamin C and BMI is shown by the correlation coefficient (r = -0.071). Since this trend is unlikely due to chance, the p-value (0.02) is statistically significant.

Null Hypotheses

Age and BMI: Simply state that H_0 (no correlation between age and BMI).

Sex and BMI: H_0 has no difference in BMI between males and females.

Race and BMI: H_0 is that there are no BMI differences across the racial groups.

Vitamin C and BMI: No correlation between vitamin C and BMI is stated by H_0.

Findings

- While age and BMI have a weak but statistically significant positive correlation ($r = 0.169$, $p < 0.001$), it suggests that BMI will tend to increase slightly with age.
- Non-significant t-test ($p = 0.936$) indicates that sex does not significantly influence BMI.
- Although the result is near significance, there are no significant racial differences in BMI (ANOVA: $p = 0.086$).
- There is a weak negative correlation ($r = -0.071$, $p = 0.020$) between vitamin C and BMI, such that a higher BMI is associated with a lower vitamin C concentration.

Policy Recommendations

- Nutritional Education and Public Awareness Campaigns

National campaigns that promote a balanced diet in general and especially increase fruit and vegetable consumption to achieve adequate vitamin C levels should be implemented nationwide. Promote community-based nutrition programs for high-risk populations like those with lower education levels or those in rural areas.

- School-Based Interventions

Mandatory nutritional education in the school curricula works towards healthier eating habits among children and adolescents. Use the enforcement of guidelines that aim to make the food in school meals more nutrient-rich and less processed.

- Healthcare System Integration

General practitioners and healthcare providers should encourage general practitioners and healthcare providers to screen for BMI and dietary assessments during routine check-ups. Offer subsidised dietitians and nutritionists access to those who are overweight or obese.

- Urban Planning and Physical Activity Promotion

Enabling physical activity through the expansion of public parks, walking trails, and recreational facility access. Set up policies that encourage people to pedal and walk instead of driving because of the improvement of infrastructure and safety measures.

- Further Research and Longitudinal Studies

Longitudinal studies are conducted to assess the longitudinal trends in BMI andevaluatef the effectiveness of the intervention strategies. Examine additional social and economic factors that affect obesity to create more targeted public health initiatives.

Limitations

In this study, it is difficult to prove that body mass index (BMI) directly affects the variables studied. For instance, BMI is associated with age, but it is unclear whether ageing causes BMI to rise or whether other factors are involved. The results may not apply to a larger population because the sample only includes about 100 people. This is all the more difficult due to differences in factors like race and education. The study would be strengthened and the sample would be more diverse and better represent the population with a bigger sample.

Conclusion

The demographics and health-related predictors are found to be significant in determining the BMI in this study. The results confirm that age, sex, race, as well as dietary factors, including vitamin C intake, affect BMI variation. The outcomes reinforce the significance of directing public well-being interventions to diminish the risk associated with vigilance and reduce the physical issue cost to the healthcare system. With the increasing prevalence of obesity and its adverse consequences, evidence-based strategies such as nutritional education, integration of the healthcare system, and urban planning that promotes physical activity should be put in place by policymakers. More research should be done in the future to understand additional socio-economic determinants related to BMI and the long-term effects of these interventions for healthcare applications. When a fix is taken for these issues in a multifaceted way, better population health outcomes will resul,t and in the long term, take some of the pressures off the existing healthcare system.

Reference

Bhaskaran, K., Douglas, I., Forbes, H., dos-Santos-Silva, I., Leon, D.A. and Smeeth, L. (2014). Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5·24 million UK adults. *The Lancet*, [online] 384(9945), pp.755–765. doi:https://doi.org/10.1016/s0140-6736(14)60892-8.

Centers for Disease Control and Prevention (2022). *Adult Obesity Facts*. [online] Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/obesity/data/adult.html.

Flegal, K.M., Kit, B.K., Orpana, H. and Graubard, B.I. (2013). Association of All-Cause Mortality With Overweight and Obesity Using Standard Body Mass Index Categories. *JAMA*, 309(1), p.71. doi:https://doi.org/10.1001/jama.2012.113905.

Malik, V.S., Willett, W.C. and Hu, F.B. (2012). Global obesity: trends, risk factors and policy implications. *Nature Reviews Endocrinology*, [online] 9(1), pp.13–27. doi:https://doi.org/10.1038/nrendo.2012.199.

Ng, M., Fleming, T., Robinson, M., Thomson, B., Graetz, N., Margono, C., Mullany, E.C., Biryukov, S., Abbafati, C., Abera, S.F., Abraham, J.P., Abu-Rmeileh, N.M.E., Achoki, T., AlBuhairan, F.S., Alemu, Z.A., Alfonso, R., Ali, M.K., Ali, R., Guzman, N.A. and Ammar, W. (2014). Global, regional, and National Prevalence of Overweight and Obesity in Children and Adults during 1980-2013: a Systematic Analysis for the Global Burden of Disease Study 2013. *Lancet (London, England)*, [online] 384(9945), pp.766–81. doi:https://doi.org/10.1016/S0140-6736(14)60460-8.

NHS Choices (2019). *Healthy weight*. [online] NHS. Available at: https://www.nhs.uk/live-well/healthy-weight/.

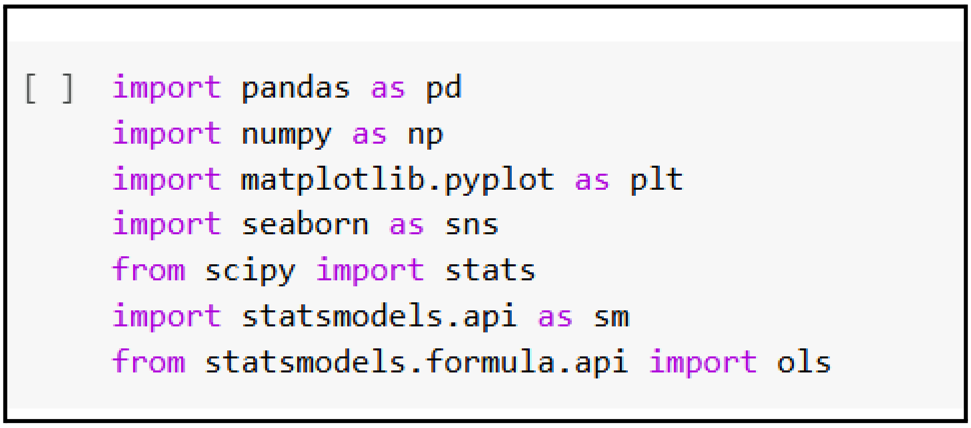
Swinburn, B.A., Sacks, G., Hall, K.D., McPherson, K., Finegood, D.T., Moodie, M.L. and Gortmaker, S.L. (2011). The Global Obesity pandemic: Shaped by Global Drivers and Local Environments. *The Lancet*, 378(9793), pp.804–814. doi:https://doi.org/10.1016/s0140-6736(11)60813-1.

The GBD 2015 Obesity Collaborators (2017). Health Effects of Overweight and Obesity in 195 Countries over 25 Years. *New England Journal of Medicine*, 377(1), pp.13–27. doi:https://doi.org/10.1056/nejmoa1614362.

WHO (2020). *Obesity*. [online] WHO. Available at: https://www.who.int/health-topics/obesity.

Appendix

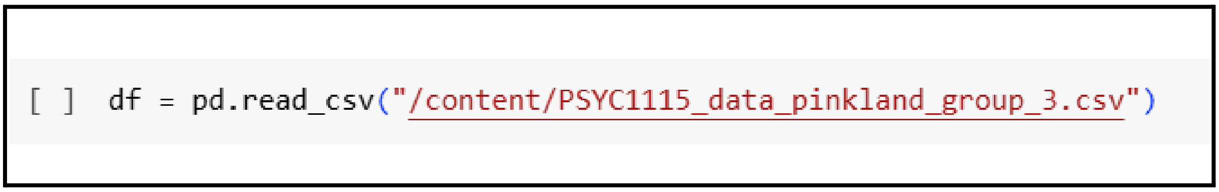
Appendix: Import Libraries & Datasets



```
[ ]  import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import seaborn as sns
     from scipy import stats
     import statsmodels.api as sm
     from statsmodels.formula.api import ols
```

Figure 11: Import Libraries

(Source: Self-Created)

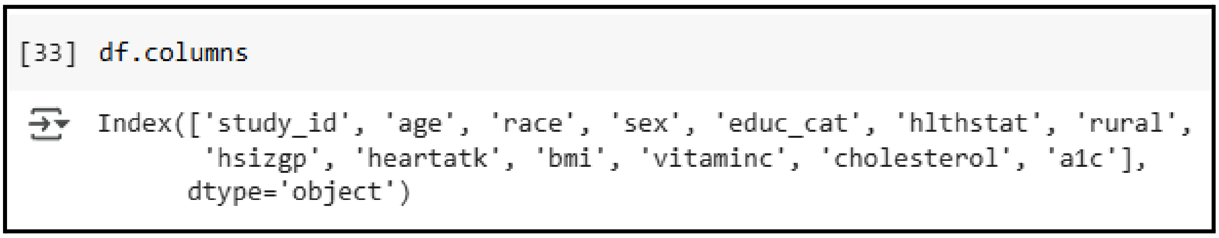


```
[ ]  df = pd.read_csv("/content/PSYC1115_data_pinkland_group_3.csv")
```

Figure 12: Import Dataset

(Source: Self-Created)

Appendix: Data Exploration

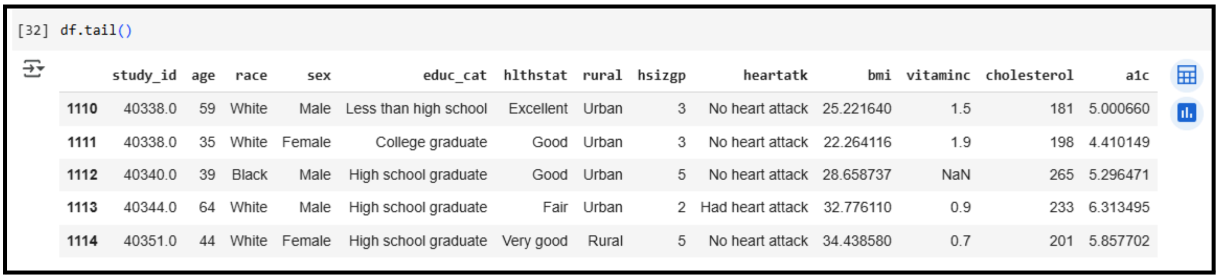


```
[33] df.columns

     Index(['study_id', 'age', 'race', 'sex', 'educ_cat', 'hlthstat', 'rural',
            'hsizgp', 'heartatk', 'bmi', 'vitaminc', 'cholesterol', 'a1c'],
           dtype='object')
```

Figure 13: All Columns

(Source: Self-Created)

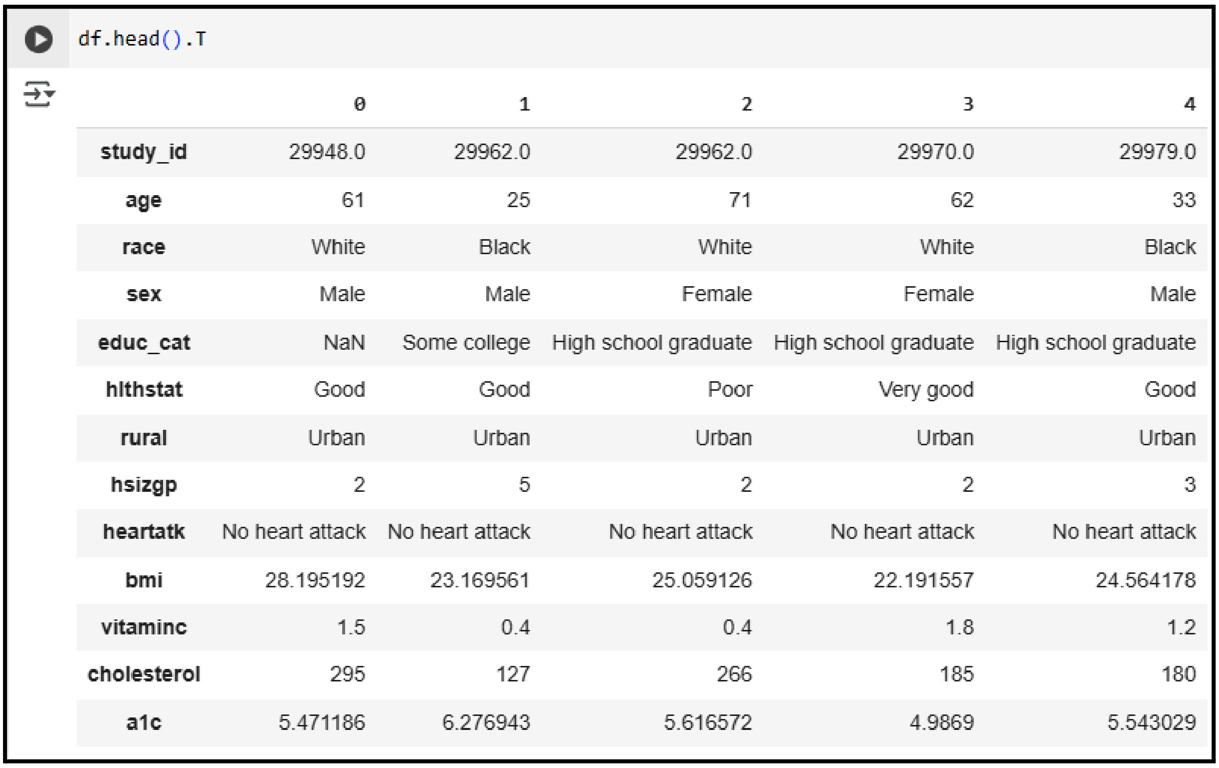


```
[32] df.tail()
```

	study_id	age	race	sex	educ_cat	hlthstat	rural	hsizgp	heartatk	bmi	vitaminc	cholesterol	a1c
1110	40338.0	59	White	Male	Less than high school	Excellent	Urban	3	No heart attack	25.221640	1.5	181	5.000660
1111	40338.0	35	White	Female	College graduate	Good	Urban	3	No heart attack	22.264116	1.9	198	4.410149
1112	40340.0	39	Black	Male	High school graduate	Good	Urban	5	No heart attack	28.658737	NaN	265	5.296471
1113	40344.0	64	White	Male	High school graduate	Fair	Urban	2	Had heart attack	32.776110	0.9	233	6.313495
1114	40351.0	44	White	Female	High school graduate	Very good	Rural	5	No heart attack	34.438580	0.7	201	5.857702

Figure 14: Last 5 Rows

(Source: Self-Created)

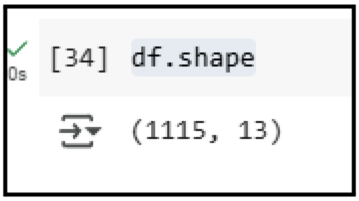


```
df.head().T
```

	0	1	2	3	4
study_id	29948.0	29962.0	29962.0	29970.0	29979.0
age	61	25	71	62	33
race	White	Black	White	White	Black
sex	Male	Male	Female	Female	Male
educ_cat	NaN	Some college	High school graduate	High school graduate	High school graduate
hlthstat	Good	Good	Poor	Very good	Good
rural	Urban	Urban	Urban	Urban	Urban
hsizgp	2	5	2	2	3
heartatk	No heart attack	No heart attack	No heart attack	No heart attack	No heart attack
bmi	28.195192	23.169561	25.059126	22.191557	24.564178
vitaminc	1.5	0.4	0.4	1.8	1.2
cholesterol	295	127	266	185	180
a1c	5.471186	6.276943	5.616572	4.9869	5.543029

Figure 15: First 5 Rows

(Source: Self-Created)



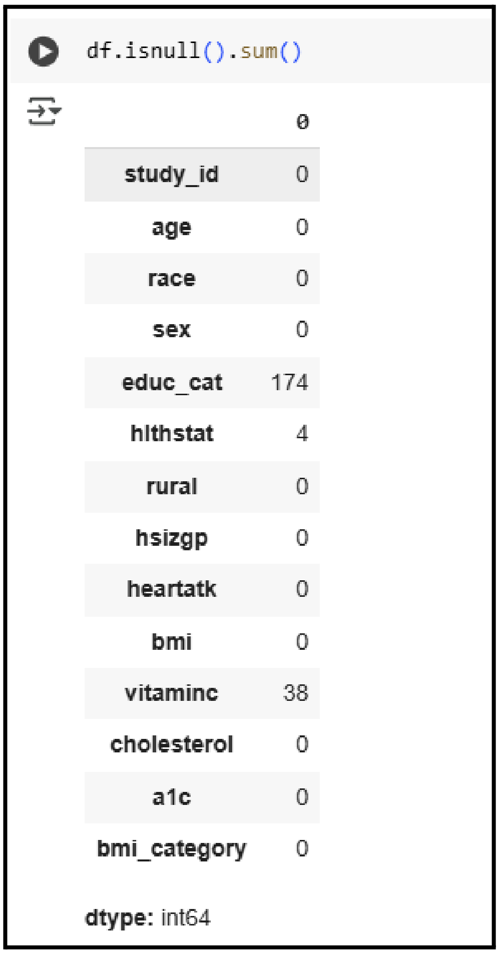
```
[34] df.shape

(1115, 13)
```

Figure 16: Shape of the Dataset

(Source: Self-Created)

Appendix: Null Values Handle



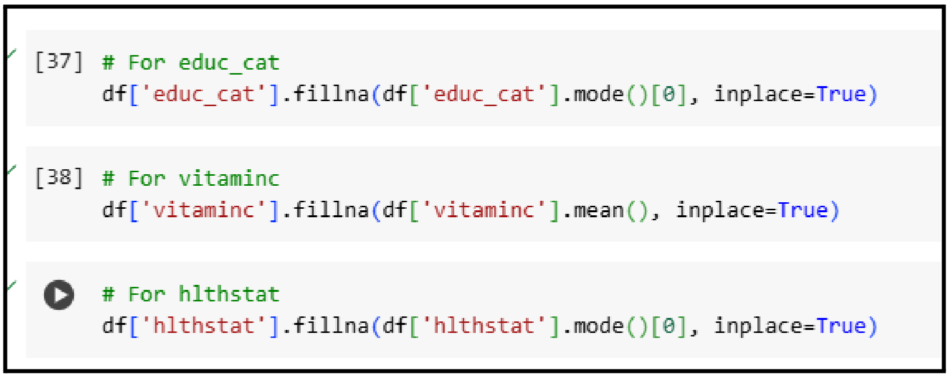
```python
df.isnull().sum()
```

```
                0
study_id        0
age             0
race            0
sex             0
educ_cat      174
hlthstat        4
rural           0
hsizgp          0
heartatk        0
bmi             0
vitaminc       38
cholesterol     0
a1c             0
bmi_category    0

dtype: int64
```

Figure 17: Checking Null Values

(Source: Self-Created)



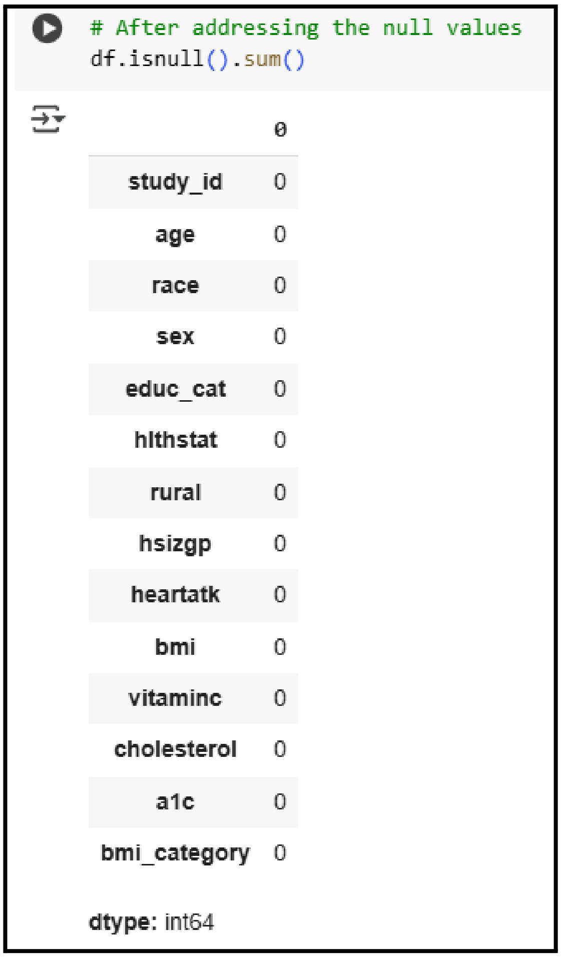
```python
[37] # For educ_cat
     df['educ_cat'].fillna(df['educ_cat'].mode()[0], inplace=True)

[38] # For vitaminc
     df['vitaminc'].fillna(df['vitaminc'].mean(), inplace=True)

     # For hlthstat
     df['hlthstat'].fillna(df['hlthstat'].mode()[0], inplace=True)
```

Figure 18: Handling the Null Values

(Source: Self-Created)



```
# After addressing the null values
df.isnull().sum()
```

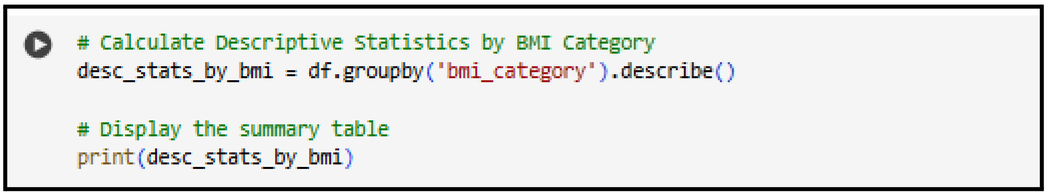
```
                 0
     study_id    0
          age    0
         race    0
          sex    0
     educ_cat    0
     hlthstat    0
        rural    0
       hsizgp    0
     heartatk    0
          bmi    0
     vitaminc    0
  cholesterol    0
          a1c    0
 bmi_category    0

dtype: int64
```

Figure 19: After Handling the Null values

(Source: Self-Created)

Appendix: Descriptive statistics

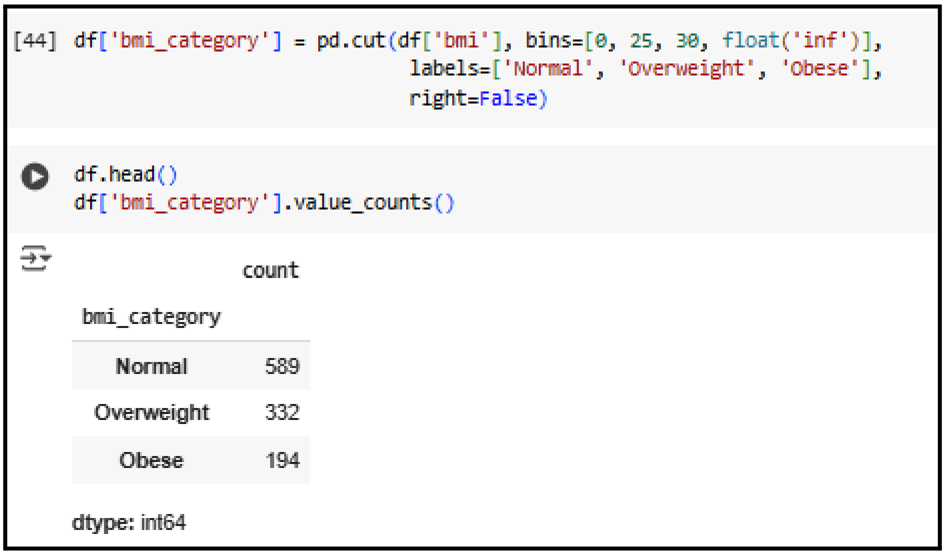


```
# Calculate Descriptive Statistics by BMI Category
desc_stats_by_bmi = df.groupby('bmi_category').describe()

# Display the summary table
print(desc_stats_by_bmi)
```

Figure 20: DS by BMI Category

(Source: Self-Created)



```
[44] df['bmi_category'] = pd.cut(df['bmi'], bins=[0, 25, 30, float('inf')],
                                  labels=['Normal', 'Overweight', 'Obese'],
                                  right=False)
```

```
df.head()
df['bmi_category'].value_counts()
```

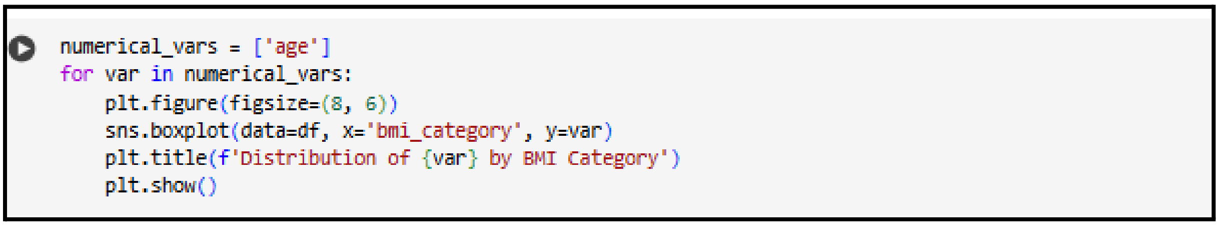
```
                    count
bmi_category

    Normal          589
    Overweight      332
    Obese           194

dtype: int64
```

Figure 21: Make Category of BMI

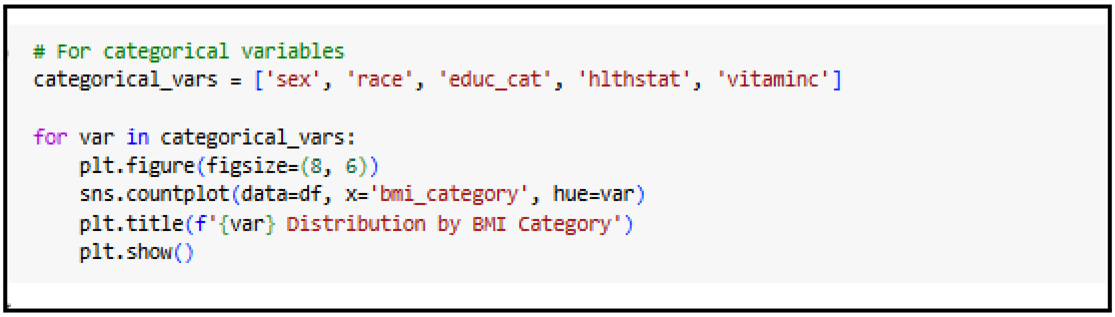
(Source: Self-Created)



```
numerical_vars = ['age']
for var in numerical_vars:
    plt.figure(figsize=(8, 6))
    sns.boxplot(data=df, x='bmi_category', y=var)
    plt.title(f'Distribution of {var} by BMI Category')
    plt.show()
```

Figure 22: Code of distribution of age by BMI

(Source: Self-Created)

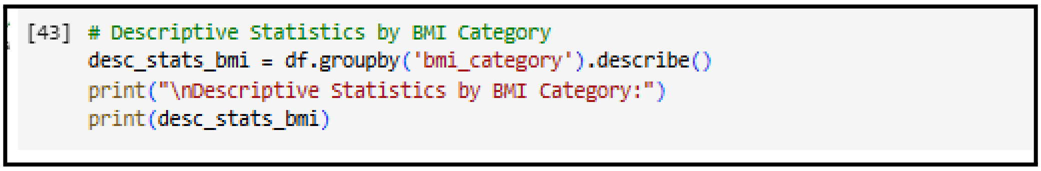


```
# For categorical variables
categorical_vars = ['sex', 'race', 'educ_cat', 'hlthstat', 'vitaminc']

for var in categorical_vars:
    plt.figure(figsize=(8, 6))
    sns.countplot(data=df, x='bmi_category', hue=var)
    plt.title(f'{var} Distribution by BMI Category')
    plt.show()
```

Figure 23: Code of Distribution of BMI categories

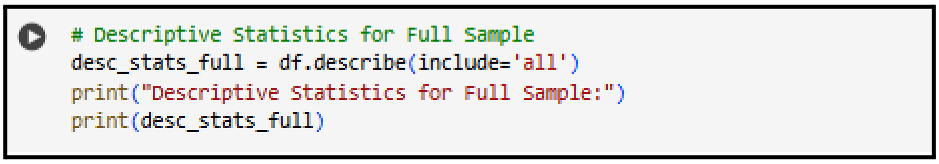
(Source: Self-Created)



```
[43] # Descriptive Statistics by BMI Category
     desc_stats_bmi = df.groupby('bmi_category').describe()
     print("\nDescriptive Statistics by BMI Category:")
     print(desc_stats_bmi)
```

Figure 24: DS statistics by BMI category

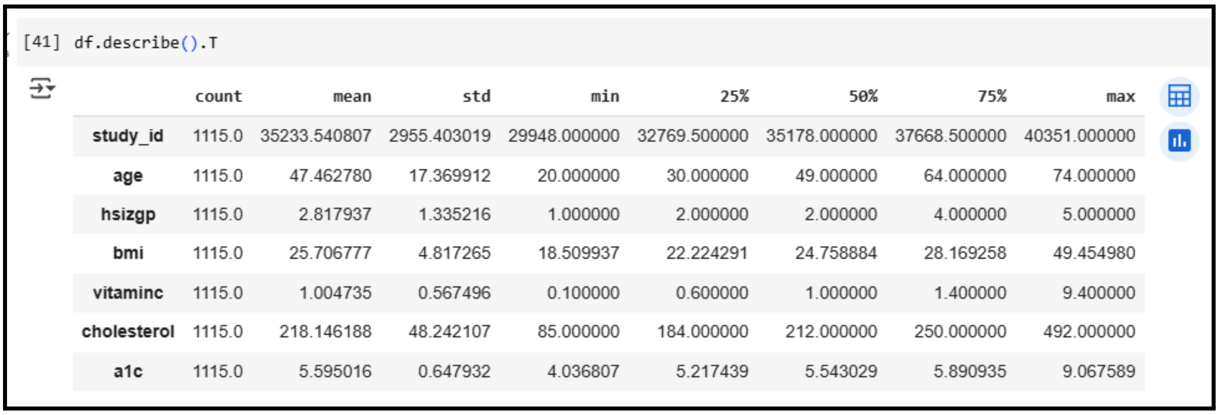
(Source: Self-Created)



```
# Descriptive Statistics for Full Sample
desc_stats_full = df.describe(include='all')
print("Descriptive Statistics for Full Sample:")
print(desc_stats_full)
```

Figure 25: DS for Full Sample

(Source: Self-Created)

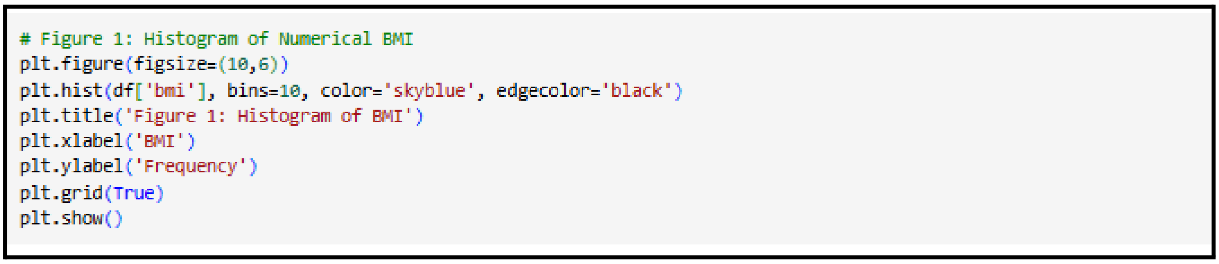


```
[41] df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
study_id	1115.0	35233.540807	2955.403019	29948.000000	32769.500000	35178.000000	37668.500000	40351.000000
age	1115.0	47.462780	17.369912	20.000000	30.000000	49.000000	64.000000	74.000000
hsizgp	1115.0	2.817937	1.335216	1.000000	2.000000	2.000000	4.000000	5.000000
bmi	1115.0	25.706777	4.817265	18.509937	22.224291	24.758884	28.169258	49.454980
vitaminc	1115.0	1.004735	0.567496	0.100000	0.600000	1.000000	1.400000	9.400000
cholesterol	1115.0	218.146188	48.242107	85.000000	184.000000	212.000000	250.000000	492.000000
a1c	1115.0	5.595016	0.647932	4.036807	5.217439	5.543029	5.890935	9.067589

Figure 26: Describe Code

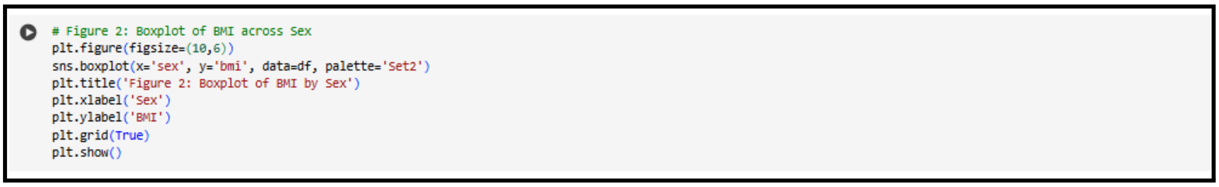
(Source: Self-Created)



```
# Figure 1: Histogram of Numerical BMI
plt.figure(figsize=(10,6))
plt.hist(df['bmi'], bins=10, color='skyblue', edgecolor='black')
plt.title('Figure 1: Histogram of BMI')
plt.xlabel('BMI')
plt.ylabel('Frequency')
plt.grid(True)
plt.show()
```

Figure 27: Histogram Code

(Source: Self-Created)

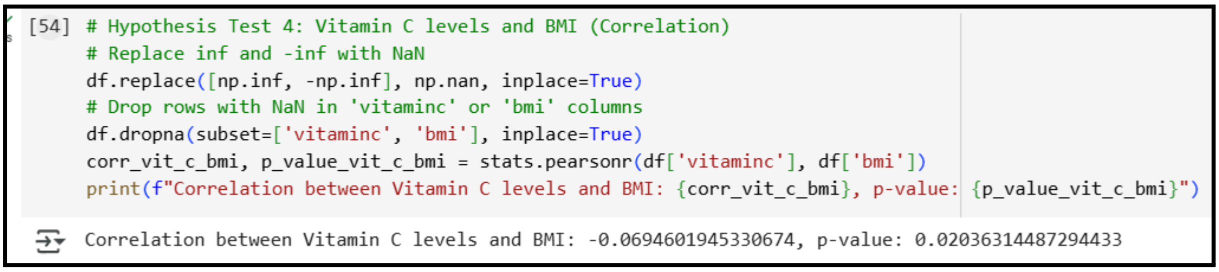


```
# Figure 2: Boxplot of BMI across Sex
plt.figure(figsize=(10,6))
sns.boxplot(x='sex', y='bmi', data=df, palette='Set2')
plt.title('Figure 2: Boxplot of BMI by Sex')
plt.xlabel('Sex')
plt.ylabel('BMI')
plt.grid(True)
plt.show()
```

Figure 28: Boxplot Code

(Source: Self-Created)

Appendix: Inferential Statistics

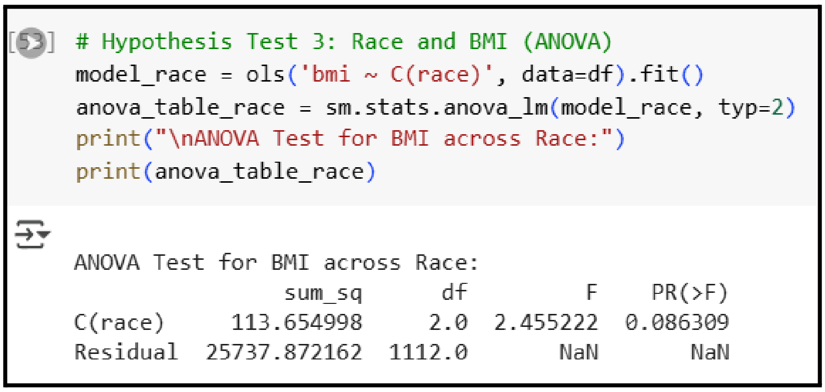


```
[54] # Hypothesis Test 4: Vitamin C levels and BMI (Correlation)
     # Replace inf and -inf with NaN
     df.replace([np.inf, -np.inf], np.nan, inplace=True)
     # Drop rows with NaN in 'vitaminc' or 'bmi' columns
     df.dropna(subset=['vitaminc', 'bmi'], inplace=True)
     corr_vit_c_bmi, p_value_vit_c_bmi = stats.pearsonr(df['vitaminc'], df['bmi'])
     print(f"Correlation between Vitamin C levels and BMI: {corr_vit_c_bmi}, p-value: {p_value_vit_c_bmi}")

     Correlation between Vitamin C levels and BMI: -0.0694601945330674, p-value: 0.02036314487294433
```

Figure 29: Code of Hypothesis Test 4

(Source: Self-Created)

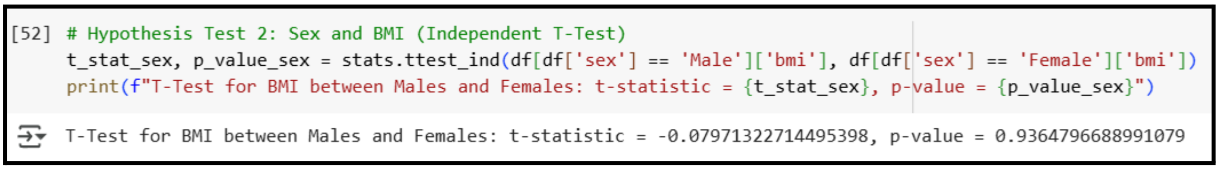


```
[53] # Hypothesis Test 3: Race and BMI (ANOVA)
     model_race = ols('bmi ~ C(race)', data=df).fit()
     anova_table_race = sm.stats.anova_lm(model_race, typ=2)
     print("\nANOVA Test for BMI across Race:")
     print(anova_table_race)


     ANOVA Test for BMI across Race:
                     sum_sq      df         F    PR(>F)
     C(race)     113.654998     2.0  2.455222  0.086309
     Residual  25737.872162  1112.0       NaN       NaN
```

Figure 30: Hypothesis Test 3

(Source: Self-Created)

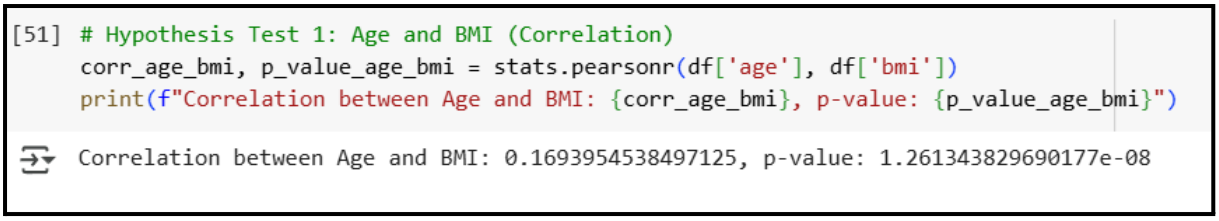


```
[52] # Hypothesis Test 2: Sex and BMI (Independent T-Test)
     t_stat_sex, p_value_sex = stats.ttest_ind(df[df['sex'] == 'Male']['bmi'], df[df['sex'] == 'Female']['bmi'])
     print(f"T-Test for BMI between Males and Females: t-statistic = {t_stat_sex}, p-value = {p_value_sex}")

     T-Test for BMI between Males and Females: t-statistic = -0.07971322714495398, p-value = 0.9364796688991079
```

Figure 31: Hypothesis Test 2

(Source: Self-Created)

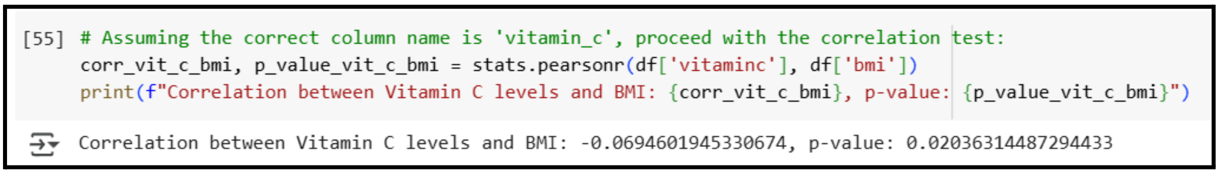


```
[51] # Hypothesis Test 1: Age and BMI (Correlation)
     corr_age_bmi, p_value_age_bmi = stats.pearsonr(df['age'], df['bmi'])
     print(f"Correlation between Age and BMI: {corr_age_bmi}, p-value: {p_value_age_bmi}")

     Correlation between Age and BMI: 0.1693954538497125, p-value: 1.261343829690177e-08
```

Figure 32: Hypothesis Test 1

(Source: Self-Created)



```
[55] # Assuming the correct column name is 'vitamin_c', proceed with the correlation test:
     corr_vit_c_bmi, p_value_vit_c_bmi = stats.pearsonr(df['vitaminc'], df['bmi'])
     print(f"Correlation between Vitamin C levels and BMI: {corr_vit_c_bmi}, p-value: {p_value_vit_c_bmi}")

     Correlation between Vitamin C levels and BMI: -0.0694601945330674, p-value: 0.02036314487294433
```

Figure 33: Code of correlation test

(Source: Self-Created)