

Dataset Loaded Successfully

	record_id	age_group	age	weight	gender	region	smoking_status	...	bmi	blood_pressure	diabetes	hypertension	cholesterol_level	glucose_level
0	REC_0001	18-25	21	71	Male	South	Smoker	...	30.6	133.2	False	False	162.8	81.4
1	REC_0002	36-45	45	46	Male	South	Non-Smoker	...	21.5	131.2	True	False	224.7	126.5
2	REC_0003	26-35	29	60	Female	West	Non-Smoker	...	24.6	149.7	False	True	198.3	99.2
3	REC_0004	36-45	39	95	Female	North	Smoker	...	18.8	135.1	False	False	267.6	110.1
4	REC_0005	46-60	59	56	Other	North	Smoker	...	24.8	153.9	False	True	218.4	95.2

[5 rows x 15 columns]

Dataset Info:  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 500 entries, 0 to 499  
Data columns (total 15 columns):

#	Column	Non-Null Count	Dtype
0	record_id	500 non-null	object
1	age_group	500 non-null	object
2	age	500 non-null	int64
3	weight	500 non-null	int64
4	gender	500 non-null	object
5	region	500 non-null	object
6	smoking_status	500 non-null	object
7	exercise_frequency	500 non-null	object
8	bmi	500 non-null	float64
9	blood_pressure	500 non-null	float64
10	diabetes	500 non-null	bool
11	hypertension	500 non-null	bool
12	cholesterol_level	500 non-null	float64
13	glucose_level	500 non-null	float64
14	visit_date	500 non-null	object

Missing Values:

record_id	0
age_group	0
age	0
weight	0
gender	0
region	0
smoking_status	0
exercise_frequency	0
bmi	0
blood_pressure	0
diabetes	0
hypertension	0
cholesterol_level	0
glucose_level	0
visit_date	0

dtype: int64

Data Cleaning Done

Descriptive Statistics:

	age	weight	bmi	blood_pressure	diabetes	hypertension	cholesterol_level	glucose_level
count	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000
mean	46.558000	70.088000	25.253800	134.787600	0.422000	0.390000	216.763400	120.08060
std	17.056274	14.226773	4.003471	14.436734	0.494373	0.488238	36.429397	23.76824
min	18.000000	45.000000	18.500000	110.000000	0.000000	0.000000	150.300000	80.30000
25%	32.000000	58.000000	21.700000	122.400000	0.000000	0.000000	186.925000	99.20000
50%	47.000000	70.500000	25.100000	134.150000	0.000000	0.000000	219.250000	119.80000
75%	62.000000	82.000000	28.825000	147.775000	1.000000	1.000000	248.350000	140.90000
max	75.000000	95.000000	32.000000	159.700000	1.000000	1.000000	280.000000	159.70000

CONFIDENCE INTERVAL FOR AGE (95%)  
Mean Age: 46.56  
Confidence Interval: (45.06, 48.05)

CONFIDENCE INTERVAL FOR AGE (95%)  
Mean Age: 46.56  
Confidence Interval: (45.06, 48.05)

CHI-SQUARE TEST (Smoking vs Diabetes)  
Chi-Square Value: 0.7211478412815252  
P-value: 0.6972760301069733  
Decision: Fail to Reject  $H_0$

T-TEST (BMI vs Diabetes)  
T-statistic: -1.3185258161724853  
P-value: 0.18793374675356622  
Decision: Fail to Reject  $H_0$

ANOVA TEST (Exercise Frequency vs BMI)  
F-statistic: 1.5729928918027194  
P-value: 0.19501424739757406  
Decision: Fail to Reject  $H_0$

COVARIANCE & CORRELATION  
Covariance (Age, BMI): -1.221463326653303  
Correlation (Age, BMI): -0.01788790900196501

FINAL CONCLUSION:

- Lifestyle and health indicators show significant relationships.
- Smoking and exercise patterns impact disease prevalence.
- BMI is significantly different for diabetic individuals.
- Age and BMI show measurable correlation.