Predictive Modeling of Cardiometabolic Risk Using Lifestyle and Socioeconomic Factors

Ellie Landoch

elandoch@bellarmine.edu

DS450 Data Science Senior Capstone

10 February 2026

Introduction

Cardiovascular disease is one of the leading causes of death in the United States, though many cases are preventable through early identification of risk factors and lifestyle intervention. Cardiovascular risk is influenced by demographic characteristics, clinical measurements, and lifestyle behaviors. Predictive analytics can be used to identify individuals at higher risk and support preventive strategies.

The dataset used in this project is a cardiometabolic-focused subset of the National Health and Nutrition Examination Survey (NHANES), a health survey conducted by the Centers for Disease Control and Prevention (CDC). NHANES data can be accessed directly through the CDC NHANES website or public data repositories such as Kaggle. This dataset was selected because it contains demographic, behavioral, clinical, and laboratory variables related to cardiometabolic health, making it a good dataset for predictive modeling of cardiometabolic risk.

The purpose of this exploratory data analysis (EDA) is to understand the structure, quality, and characteristics of the data before building prediction models. This EDA evaluates data completeness, variable distributions, relationships between predictors, and potential data quality issues, such as missing data and outliers. The results of this analysis will guide future data preprocessing and model development.

Dataset Description

This dataset comes from NHANES and includes approximately 10,000 observations with demographic, behavioral, clinical, and laboratory variables related to cardiometabolic health.

The variables in this dataset represent several major categories of health information. Demographic variables include age, sex, race/ethnicity, education level, income ratio, and marital status. Clinical measurements include blood pressure readings, body measurements, cholesterol levels, triglycerides, and glycohemoglobin. Behavioral and lifestyle variables include smoking history, alcohol consumption, and physical activity. Laboratory variables include measures related to kidney function, blood chemistry, and metabolic health.

The dataset contains both continuous and categorical variables. Continuous variables include measurements such as blood pressure, body mass index, cholesterol levels, and laboratory test values. Categorical variables include survey responses, diagnostic indicators, and behavioral indicators. Variables are classified as nominal, ordinal, interval, or ratio based on the measurement scale. Most laboratory and clinical measurement variables fall into the ratio scale, while most survey response variables fall into nominal or ordinal categories.

The table below provides details about each variable, including variable name, data type, value range, and percentage of data missing. A full data dictionary with detailed variable definitions is available in the project GitHub repository.

Variable Name	Data Type	% Data Missing	Min value	Max value
SEQN	Nominal	0	73557	83731
BPXSY1	Ratio	29.51	66	228
BPXSY2	Ratio	27.18	66	230
BPXSY3	Ratio	27.19	62	228
BPXSY4	Ratio	94.94	80	212
BPXDI1	Ratio	29.51	0	122
BPXDI2	Ratio	27.18	0	116
BPXDI3	Ratio	27.19	0	118
BPXDI4	Ratio	94.94	0	128
BPXPULS	Nominal	6.53	1	2
BPXSY_AVG	Ratio	25.99	64	229
BPXDI_AVG	Ratio	25.99	0	128
BP_CATEGORY	Ordinal	25.99		
HTN_STAGE	Ordinal	25.99		
HYPERTENSION_YN	Nominal	0	0	1
BPQ020	Nominal	36.47	1	9
BPQ030	Nominal	78.63	1	9
BPQ050A	Nominal	82.16	1	9
RIDAGEYR	Ratio	0	0	80
RIAGENDR	Nominal	0	1	2
RIDRETH1	Nominal	0	1	5
DMDEDUC2	Ordinal	43.3	1	9
INDFMPIR	Ratio	7.71	0	5
DMDMARTL	Nominal	43.3	1	99
BMXWT	Ratio	4.44	3.1	222.6
BMXHT	Ratio	10.89	79.7	202.6
BMXBMI	Ratio	11.01	12.1	82.9
BMXWAIST	Ratio	14.88	40.2	177.9
BMXARML	Ratio	8.59	9.9	47.9
BMXARMC	Ratio	8.59	10.4	59.4
BMXLEG	Ratio	27.25	24.4	51.9
LBXGH	Ratio	34.71	3.5	17.5
LBXIN	Ratio	69.6	0.14	682.48
DIQ010	Nominal	3.99	1	9
DIQ050	Nominal	4	1	9
DIQ070	Nominal	88.21	1	9
LBXTC	Ratio	25.07	69	813

LBDHDD	Ratio	25.07	10	173
LBDLDL	Ratio	69.48	14	375
LBXTR	Ratio	69.08	13	4233
RATIO_TC_HDL	Ratio	25.07	1.311828	25.1
RATIO_TG_HDL	Ratio	69.08	0.184971	103.2439
LBXSCR	Ratio	35.6	0.29	17.41
URXUMA	Ratio	20.86	0.21	9600
URXUCR	Ratio	73.56	8	659
eGFR_CKD_EPI_2021	Ratio	35.6	2.015401	172.3696
ACR_MG_PER_G	Ratio	73.56	0.21164	9000
LBXSAL	Ratio	35.6	2.4	5.6
LBXWBCSI	Ratio	16.03	2.3	55.7
LBXPLTSI	Ratio	16.03	18	723
SMQ020	Nominal	39.92	1	9
SMQ040	Nominal	74.65	1	3
ALQ120Q	Ordinal	55.98	0	999
ALQ120U	Nominal	64.69	1	3
PAQ605	Nominal	29.75	1	7
PAQ620	Nominal	29.75	1	9
PAD615	Ratio	88.52	10	9999
PAQ650	Nominal	29.76	1	9
DBQ010	Nominal	81.67	1	9
DBQ700	Ordinal	36.47	1	9
SLQ050	Nominal	36.47	1	9
SLQ060	Nominal	36.47	1	9
HSD010	Ordinal	36.44	1	9
BPQ080	Nominal	36.47	1	9

Dataset Structure and Key Measures

The dataset contains multiple blood pressure measurements per individual, as well as derived average blood pressure variables. Average systolic blood pressure, BPXSY_AVG, and average diastolic blood pressure, BPXDI_AVG, represent overall blood pressure status and reduce measurement variability compared to individual readings. Blood pressure classification variables such as BP_CATEGORY and HTN_STAGE provide clinically defined groupings of blood pressure levels.

Lifestyle and behavioral variables are based on standardized survey questions. Smoking variables measure both lifetime smoking exposure and current smoking status. Physical activity variables measure participation in moderate and vigorous activity. Alcohol consumption variables measure both frequency and amount of alcohol use.

The dataset also includes important laboratory indicators of cardiometabolic health, including cholesterol measurements, triglycerides, glycohemoglobin, and kidney function indicators such as estimated glomerular filtration rate and albumin-to-creatinine ratio. These measures are commonly used in clinical practice to assess cardiovascular and metabolic risk.

Missing Data and Data Completeness

Exploratory analysis showed that several variables have missing data, particularly in some laboratory and specialized survey measurements. Some variables have low levels of missing data and are suitable for predictive modeling with minimal preprocessing. Other variables contain moderate to high levels of missing data and may require imputation or exclusion during modeling. Missing data is expected in NHANES datasets because not all participants complete every survey section or laboratory test.

Variables such as LBDLDL, LBXTR, URXUCR, and ACR_MG_PER_G contain high levels of missing data and will require careful evaluation before modeling. Variables with high missing data may reduce model stability or reduce usable sample size.

The presence of missing data does not reduce the usefulness of the dataset but must be addressed during preprocessing and considered during model building. The next steps of this project will include determining acceptable levels of missing data, applying imputation methods, and evaluating how missing data influences model performance.

Dataset Relevance to Project

This dataset is well-suited for predicting cardiometabolic risk because it contains variables strongly associated with cardiovascular and metabolic disease, including blood pressure, cholesterol levels, BMI, diabetes indicators, and lifestyle behaviors. The dataset also includes demographic and socioeconomic variables that allow analysis of potential differences in cardiometabolic risk across demographic groups.

By combining clinical, behavioral, and demographic variables, this dataset supports the development of predictive models that analyze all major contributors to cardiometabolic risk. This supports the project goal of predicting cardiometabolic risk using clinical, lifestyle, and socioeconomic factors.

Dataset Summary Statistics

Dataset summary statistics were calculated for all numeric variables to understand typical value ranges and variability across the patients. Clinical measures such as blood pressure, BMI, cholesterol, and glycohemoglobin show wide variation across the population, which is expected in a nationally representative health dataset. BMI and several laboratory variables show right-skewed distributions, meaning most participants fall in lower ranges while a smaller group shows elevated values. Blood pressure values are generally within expected clinical ranges but include some extreme values that may represent severe hypertension or measurement variation.

The table below shows the distribution of the numeric variables, including measures of center and spread.

	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
SEQN	10175	78644	2937.414	73557	76100.5	78644	81187.5	83731
BPXSY1	7172	118.1235	18.07815	66	106	116	128	228
BPXSY2	7409	118.2305	18.1812	66	106	116	128	230
BPXSY3	7408	117.9995	18.07985	62	106	114	128	228
BPXSY4	515	125.666	22.60809	80	108	126	140	212
BPXDI1	7172	65.76994	14.96011	0	58	66	76	122
BPXDI2	7409	65.23795	15.70024	0	58	66	74	116
BPXDI3	7408	65.03564	16.23317	0	58	68	74	118
BPXDI4	515	69.01359	15.8064	0	60	70	78	128
BPXPULS	9511	1.013668	0.116116	1	1	1	1	2
BPXSY_AVG	7531	118.1635	18.07021	64	106	115	128	229
BPXDI_AVG	7531	65.19081	15.44211	0	58	67	75	128
HYPERTENSION_YN	10175	0.300737	0.458601	0	0	0	1	1
BPQ020	6464	1.66909	0.514554	1	1	2	2	9
BPQ030	2174	1.226311	0.617529	1	1	1	1	9
BPQ050A	1815	1.130028	0.379544	1	1	1	1	9
RIDAGEYR	10175	31.48413	24.42165	0	10	26	52	80
RIAGENDR	10175	1.508305	0.499956	1	1	2	2	2
RIDRETH1	10175	3.091892	1.263305	1	2	3	4	5
DMDEDUC2	5769	3.518807	1.236032	1	3	4	5	9
INDFMPIR	9390	2.252153	1.634907	0	0.87	1.705	3.6075	5
DMDMARTL	5769	2.57185	2.626299	1	1	1	5	99
BMXWT	9723	62.59905	32.33162	3.1	37.95	65.3	83.5	222.6
BMXHT	9067	155.8838	23.17627	79.7	149.5	162	171.05	202.6
BMXBMI	9055	25.67824	7.955137	12.1	19.7	24.7	30.2	82.9
BMXWAIST	8661	87.27205	22.5426	40.2	71.2	87.8	102.8	177.9
BMXARML	9301	33.14101	7.40942	9.9	30.5	35.5	38.1	47.9
BMXARMC	9301	28.48576	7.961971	10.4	22.6	29.3	34	59.4
BMXLEG	7402	38.57771	4.04782	24.4	36	38.6	41.3	51.9

LBXGH	6643	5.642556	1.00485	3.5	5.2	5.4	5.8	17.5
LBXIN	3093	13.52681	18.63839	0.14	6.08	9.47	15.35	682.48
DIQ010	9769	1.947589	0.345375	1	2	2	2	9
DIQ050	9768	1.979423	0.186265	1	2	2	2	9
DIQ070	1200	1.555833	0.583524	1	1	2	2	9
LBXTC	7624	179.5341	40.954	69	151	175	204	813
LBDHDD	7624	53.10519	15.23084	10	42	51	61	173
LBDLDL	3105	106.2213	34.98866	14	81	103	127	375
LBXTR	3146	112.3067	115.6071	13	60	88	133	4233
RATIO_TC_HDL	7624	3.628951	1.336934	1.311828	2.717949	3.333333	4.236842	25.1
RATIO_TG_HDL	3146	2.521973	3.599036	0.184971	1.017648	1.68573	2.821795	103.2439
LBXSCR	6553	0.880172	0.487262	0.29	0.69	0.82	0.98	17.41
URXUMA	8052	41.21885	238.9102	0.21	4.5	8.4	17.625	9600
URXUCR	2690	127.5784	81.98228	8	65	112	171	659
eGFR_CKD_EPI_2021	6553	100.7588	26.02435	2.015401	84.0138	102.2523	119.7087	172.3696
ACR_MG_PER_G	2690	42.99992	294.298	0.21164	5.042355	7.961165	15.38675	9000
LBXSAL	6553	4.282085	0.343649	2.4	4.1	4.3	4.5	5.6
LBXWBCSI	8544	7.379506	2.302574	2.3	5.8	7.1	8.6	55.7
LBXPLTSI	8544	251.1951	66.05402	18	206	244	288	723
SMQ020	6113	1.580402	0.511762	1	1	2	2	9
SMQ040	2579	2.13765	0.942517	1	1	3	3	3
ALQ120Q	4479	4.70931	34.42836	0	1	2	4	999
ALQ120U	3593	1.921514	0.853701	1	1	2	3	3
PAQ605	7148	1.836738	0.375266	1	2	2	2	7
PAQ620	7148	1.679771	0.487423	1	1	2	2	9
PAD615	1168	187.5086	433.8161	10	60	120	240	9999
PAQ650	7147	1.712887	0.461026	1	1	2	2	9
DBQ010	1865	1.28311	0.541476	1	1	1	2	9
DBQ700	6464	2.95823	0.995792	1	2	3	4	9
SLQ050	6464	1.757426	0.44843	1	2	2	2	9
SLQ060	6464	1.924505	0.427139	1	2	2	2	9
HSD010	6467	2.768053	0.970974	1	2	3	3	9
BPQ080	6464	1.723855	0.715436	1	1	2	2	9

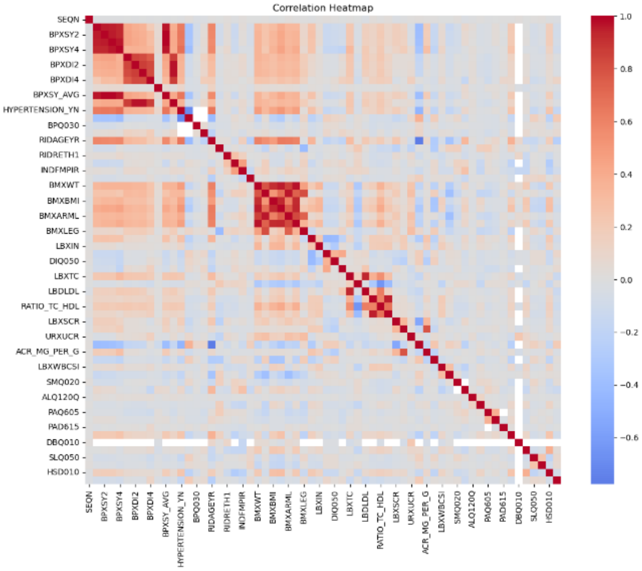
Frequency Distributions

Frequency distributions were calculated for categorical variables. For example, the hypertension indicator shows the distribution of individuals classified as hypertensive versus non-hypertensive. This provides context for variable balance and helps inform future modeling decisions. The hypertension distribution shows approximately 30% positive cases, indicating moderate imbalance that may influence model evaluation metrics.

HYPERTENSION_YN	proportion
0	69.9263
1	30.0737

Correlation Analysis

Correlation analysis showed strong relationships between several variables. Blood pressure readings strongly correlate with their calculated averages, and body composition measures such as weight, waist circumference, and BMI are also highly correlated. Lipid-related laboratory measures and ratio variables also show expected relationships because ratio variables are calculated from base cholesterol values. These correlated variable groups will be considered during feature selection to avoid redundant predictors in modeling. Demographic and socioeconomic variables showed weaker direct correlations with clinical measures but may still improve prediction when combined with other variables.



Data Anomalies

Missing data is present in several laboratory and survey variables. Some variables have little missing data and can be used with minimal preprocessing, while others with more missing data may require imputation or adding a missing data flag. Outliers were observed in BMI, blood pressure, and some laboratory variables. These will be evaluated during preprocessing to determine whether transformation, scaling, or outlier capping is appropriate. Extreme values can disproportionately influence model training, especially for distance-based models like k-Nearest Neighbors, which will be used in the project.

Implications for Predictive Modeling

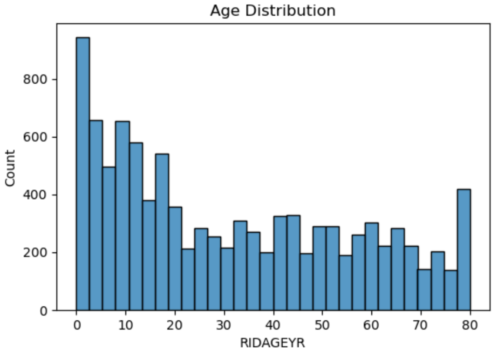
This dataset should work well for predictive modeling because it includes important clinical and lifestyle health measures. Some preprocessing will still be needed, especially for correlated variables, skewed data, and missing values.

Dataset Graphical Exploration

Graphical analysis was conducted to better understand the distribution of key demographic and clinical variables and to identify patterns, relationships, and potential outliers. Visualizations used included distribution plots (histograms), boxplots, bar charts, scatterplots, and correlation heat maps. These visualizations help show how individual health variables relate to cardiometabolic risk.

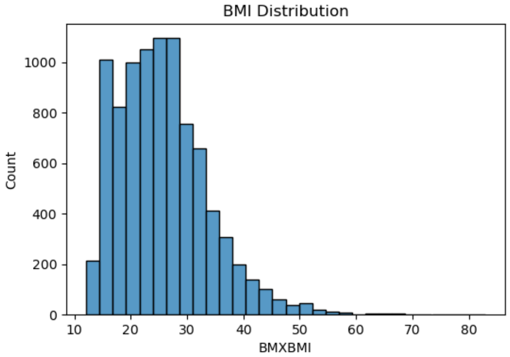
Age Distribution

The age distribution histogram shows representation of all ages, with higher counts among younger and middle-aged individuals and fewer observations at older ages, but has one additional spike of data values at the very highest point of the data range. The distribution gradually declines with increasing age, which is expected in population-based health surveys. This supports the use of age as an important predictor in cardiometabolic risk modeling.



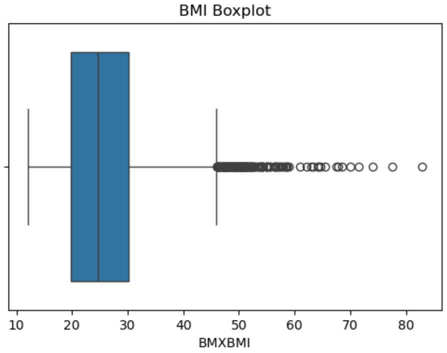
Body Mass Index (BMI) Distribution

The BMI distribution histogram shows a right-skewed distribution, with most individuals falling between 20-30 BMI. A smaller portion of individuals show much higher BMI values. This pattern is consistent with population health trends, where most individuals fall within normal to overweight ranges and fewer individuals fall into obesity or severe obesity categories.



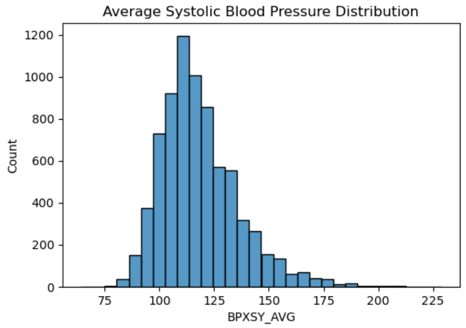
BMI Outlier Analysis

The BMI boxplot confirms the presence of high-value outliers, with several observations extending well above typical BMI ranges. These may represent severe obesity, medical conditions, or potential measurement variation. Outlier handling will be evaluated during preprocessing.



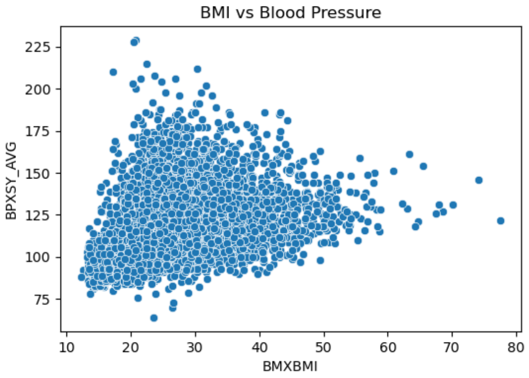
Blood Pressure Distribution

The average systolic blood pressure distribution shows an approximately bell-shaped distribution centered around typical clinical ranges. Most individuals fall within normal or pre-hypertensive ranges, while fewer individuals show extremely high blood pressure values. These extreme values may represent severe hypertension or measurement variation.



Relationship Between BMI and Blood Pressure

The scatterplot comparing BMI and average systolic blood pressure shows a weak to moderate positive relationship between the variables. This means that higher BMI values are generally associated with higher blood pressure values. However, variability suggests blood pressure is influenced by multiple factors. This shows that both variables are important for prediction because they reflect different aspects of health risk.



Implications for Predictive Modeling

The graphical analysis confirms that key predictors such as age, BMI, and blood pressure show meaningful variation across the dataset. Skewed distributions and outliers suggest that normalization or transformation may improve model performance. The relationship found between BMI and blood pressure supports including both variables in cardiometabolic risk prediction models.

Data Quality

Several data quality issues were identified during the EDA that will need to be addressed before modeling. Missing data is present in some laboratory and survey variables, which is expected in NHANES because not all participants complete every test or survey section. Some variables have small amounts of missing data, while others have larger gaps that may require imputation or additional handling.

Outliers were seen in BMI, blood pressure, and some laboratory variables. These may represent true clinical extremes or measurement variation and will need to be reviewed during preprocessing.

Strong relationships were also observed among some groups of variables, particularly blood pressure measures, body measurements, and lipid laboratory values. These relationships between variables may require selecting only certain variables to include in the model to avoid redundancy.

Overall, the dataset is usable for predictive modeling, but missing data, outliers, and correlated variables will need to be addressed during preprocessing.

Summary of Findings

This analysis shows that the NHANES cardiometabolic subset is well-suited for predictive modeling of cardiometabolic risk using demographic, lifestyle, clinical, and laboratory variables. The dataset includes predictors commonly associated with cardiovascular and metabolic risk, such as blood pressure, body mass index, and cholesterol measures. Lifestyle variables, including smoking, physical activity, and alcohol use, provide behavioral context that supports prevention-focused modeling.

The dataset also contains demographic and socioeconomic variables such as age, sex, race/ethnicity, education, and income ratio. Even though these variables don't show strong direct correlations with clinical measures, they are still important because they can help explain differences in health across populations and may influence risk. While direct healthcare access variables are limited in this dataset, demographic and socioeconomic indicators provide useful context for understanding variation in cardiometabolic risk across populations.

Several issues with the data appeared during the exploratory analysis that will need to be addressed before building the prediction models.

Some variables have a large amount of missing data, especially certain lab and survey measures. Variables with particularly high missing data include LBDLDL, LBXTR, URXUCR, and ACR_MG_PER_G, which may require special preprocessing consideration. Instead of automatically removing these variables, the amount of missing data will be evaluated first. Some variables contain substantial missing data and will need to be reviewed before modeling. For variables with moderate missing data, imputation may be used if appropriate.

Extreme values were also found in variables such as BMI, average systolic blood pressure, and some lab results. These values may represent real clinical extremes or measurement variation. These values will be reviewed and adjusted if needed so they do not overly influence model results.

Some variables are highly correlated, particularly repeated blood pressure measurements, body size measures, and cholesterol-related measures. This will be addressed during variable selection.

Despite its limitations, the dataset includes key clinical and lifestyle variables and enough data to support model development. Next steps include preprocessing, variable selection, and model development.