The dataset is called "Framingham Heart Study Dataset" and it contains information about patients from the Framingham Heart Study. It includes a variety of demographic and health-related information such as age, sex, cholesterol levels, and whether the patient has been diagnosed with a cardiovascular disease.

**Demographic:**

* Sex: male or female (Nominal)
* Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)

**Behavioural**

* Education: 0: Less than High School and High School degrees, 1: College Degree and Higher
* Current Smoker: whether or not the patient is a current smoker (Nominal)
* Cigs Per Day: the number of cigarettes that the person smoked on average in one day. (Can be considered continuous as one can have any number of cigarettes, even half a cigarette.)

**Medical (history)**

* BP Meds: whether or not the patient was on blood pressure medication (Nominal)
* Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
* Prevalent Hyp: whether or not the patient was hypertensive (Nominal)
* Diabetes: whether or not the patient had diabetes (Nominal)
* Medical(current)
* Tot Chol: total cholesterol level (Continuous)
* Sys BP: systolic blood pressure (Continuous)
* Dia BP: diastolic blood pressure (Continuous)
* BMI: Body Mass Index (Continuous)
* Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
* Glucose: glucose level (Continuous)
* Predict variable (desired target)
* 10-year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)

**Data sample:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **male** | **age** | **education** | **currentSmoker** | **cigsPerDay** | **BPMeds** | **prevalentStroke** | **prevalentHyp** | **diabetes** | **totChol** | **sysBP** | **diaBP** | **BMI** | **heartRate** | **glucose** | **TenYearCHD** |
| **0** | 1 | 39 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 195 | 106 | 70 | 26.97 | 80 | 77 | 0 |
| **1** | 0 | 46 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 250 | 121 | 81 | 28.73 | 95 | 76 | 0 |
| **2** | 1 | 48 | 1 | 1 | 20 | 0 | 0 | 0 | 0 | 245 | 127.5 | 80 | 25.34 | 75 | 70 | 0 |
| **3** | 0 | 61 | 3 | 1 | 30 | 0 | 0 | 1 | 0 | 225 | 150 | 95 | 28.58 | 65 | 103 | 1 |
| **4** | 0 | 46 | 3 | 1 | 23 | 0 | 0 | 0 | 0 | 285 | 130 | 84 | 23.1 | 85 | 85 | 0 |

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Non-Null | Count | Dtype |
|  |  |  |  |
| male | 4238 | non-null | int64 |
| age | 4238 | non-null | int64 |
| education | 4133 | non-null | float64 |
| currentSmoker | 4238 | non-null | int64 |
| cigsPerDay | 4209 | non-null | float64 |
| BPMeds | 4185 | non-null | float64 |
| prevalentStroke | 4238 | non-null | int64 |
| prevalentHyp | 4238 | non-null | int64 |
| diabetes | 4238 | non-null | int64 |
| totChol | 4188 | non-null | float64 |
| sysBP | 4238 | non-null | float64 |
| diaBP | 4238 | non-null | float64 |
| BMI | 4219 | non-null | float64 |
| heartRate | 4237 | non-null | float64 |
| glucose | 3850 | non-null | float64 |
| TenYearCHD | 4238 | non-null | int64 |

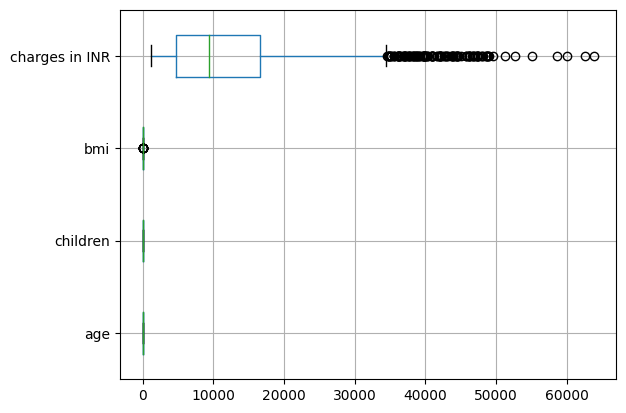
**Following features have null values:**

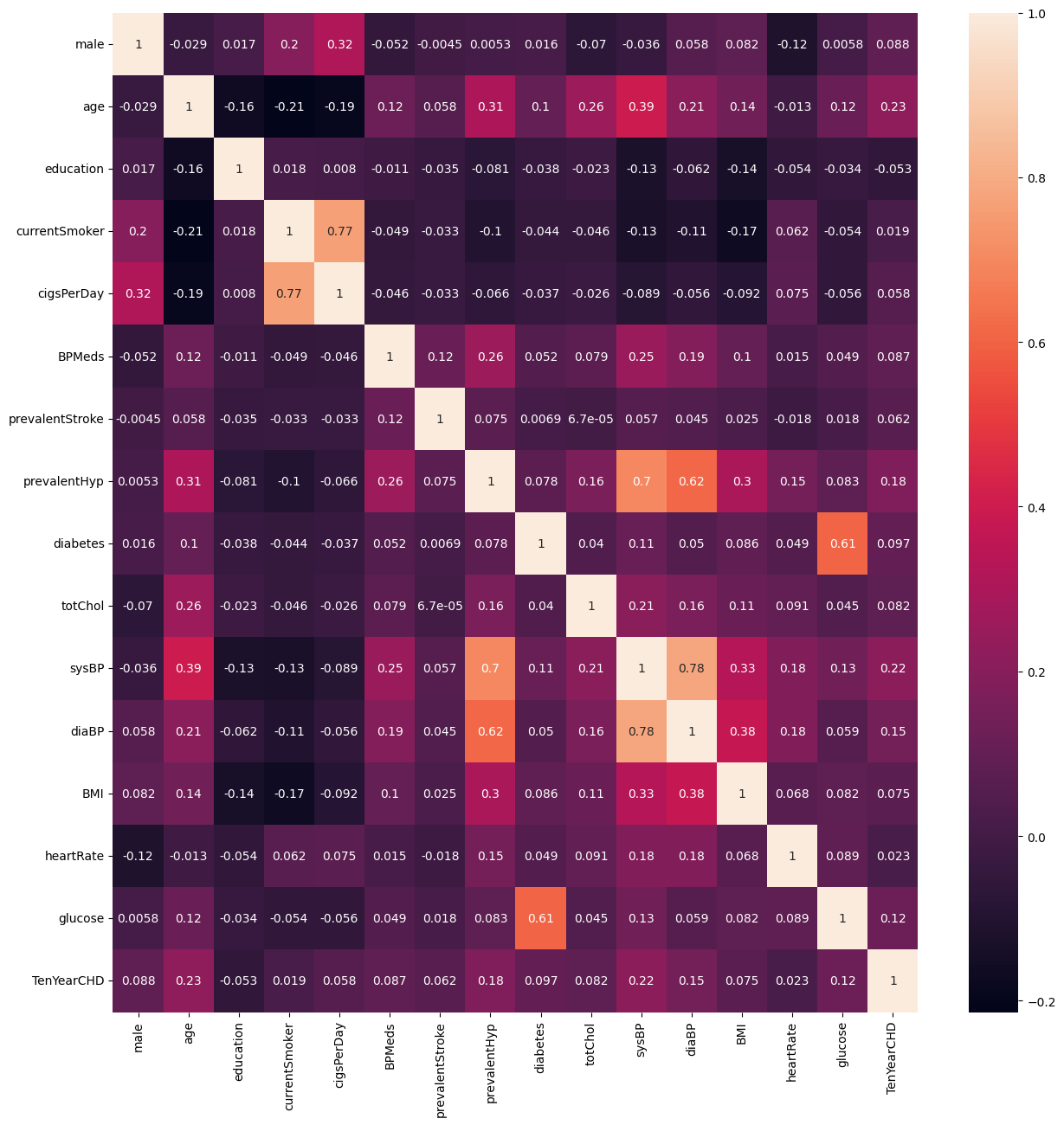
|  |  |
| --- | --- |
| Features | Count |
| education | 105 |
| cigsPerDay | 29 |
| BPMeds | 53 |
| totChol | 50 |
| BMI | 19 |
| heartRate | 1 |
| glucose | 388 |

In this case as both categorical and continuous data are present, for continuous data missing values are filled with mean of that column and for categorical data null values will be filled with most repeated data of that feature i.e. mode.

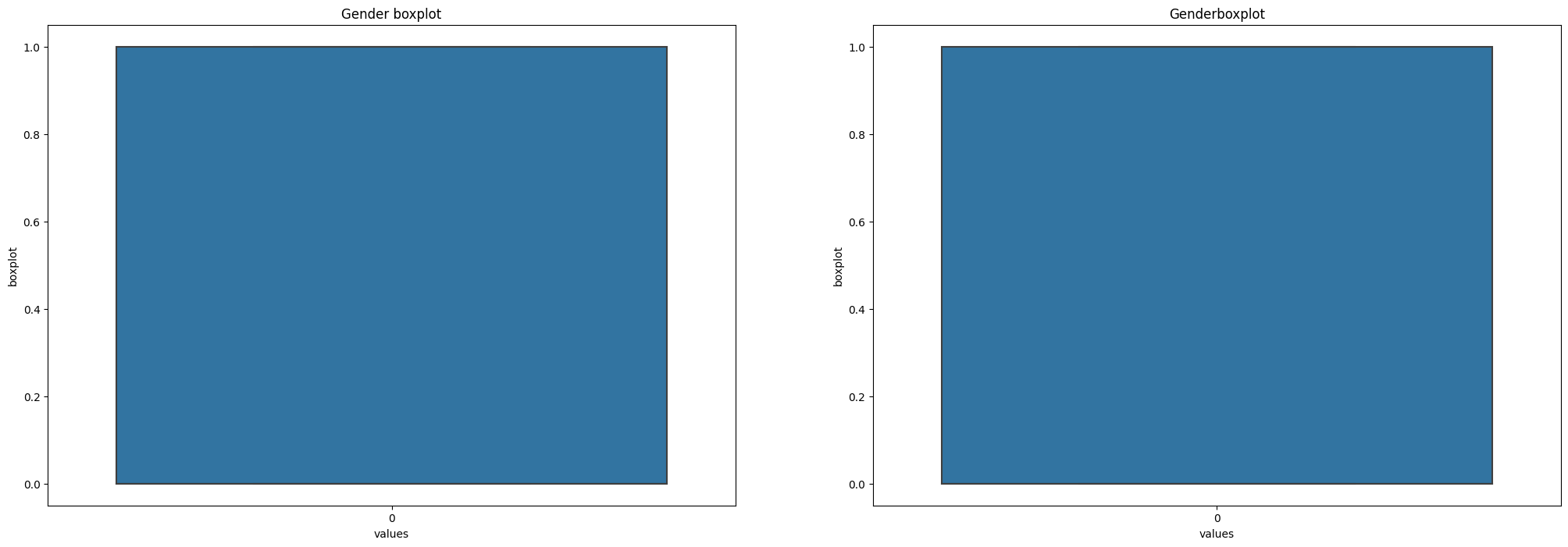
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | male | age | education | currentSmoker | cigsPerDay | BPMeds | prevalentStroke | prevalentHyp | diabetes | totChol | sysBP | diaBP | BMI | heartRate | glucose | TenYearCHD |
| count | 4238 | 4238 | 4133 | 4238 | 4209 | 4185 | 4238 | 4238 | 4238 | 4188 | 4238 | 4238 | 4219 | 4237 | 3850 | 4238 |
| mean | 0.429212 | 49.58495 | 1.97895 | 0.494101 | 9.003089 | 0.02963 | 0.005899 | 0.310524 | 0.02572 | 236.7216 | 132.3524 | 82.89346 | 25.802 | 75.87892 | 81.96675 | 0.151958 |
| std | 0.495022 | 8.57216 | 1.019791 | 0.500024 | 11.920094 | 0.169584 | 0.076587 | 0.462763 | 0.158316 | 44.59033 | 22.0381 | 11.91085 | 4.08011 | 12.0266 | 23.96 | 0.359023 |
| min | 0 | 32 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 107 | 83.5 | 48 | 15.54 | 44 | 40 | 0 |
| 25% | 0 | 42 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 206 | 117 | 75 | 23.07 | 68 | 71 | 0 |
| 50% | 0 | 49 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 234 | 128 | 82 | 25.4 | 75 | 78 | 0 |
| 75% | 1 | 56 | 3 | 1 | 20 | 0 | 0 | 1 | 0 | 263 | 144 | 89.875 | 28.04 | 83 | 87 | 0 |
| max | 1 | 70 | 4 | 1 | 70 | 1 | 1 | 1 | 1 | 696 | 295 | 142.5 | 56.8 | 143 | 394 | 1 |

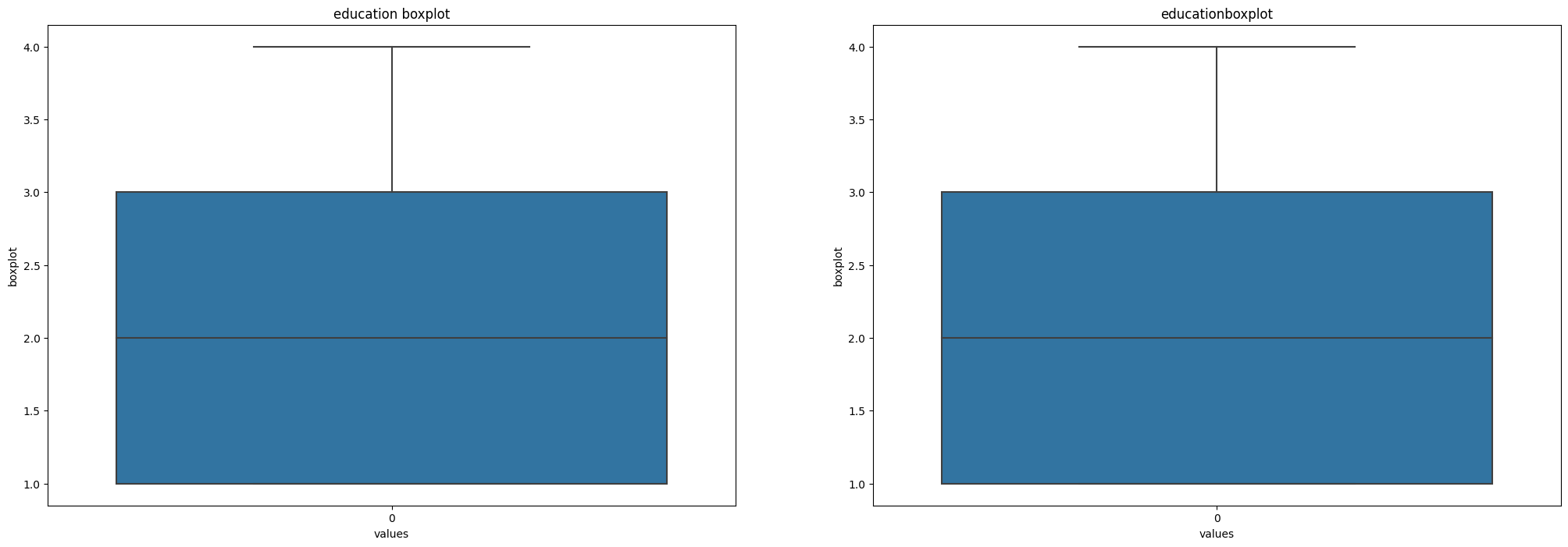
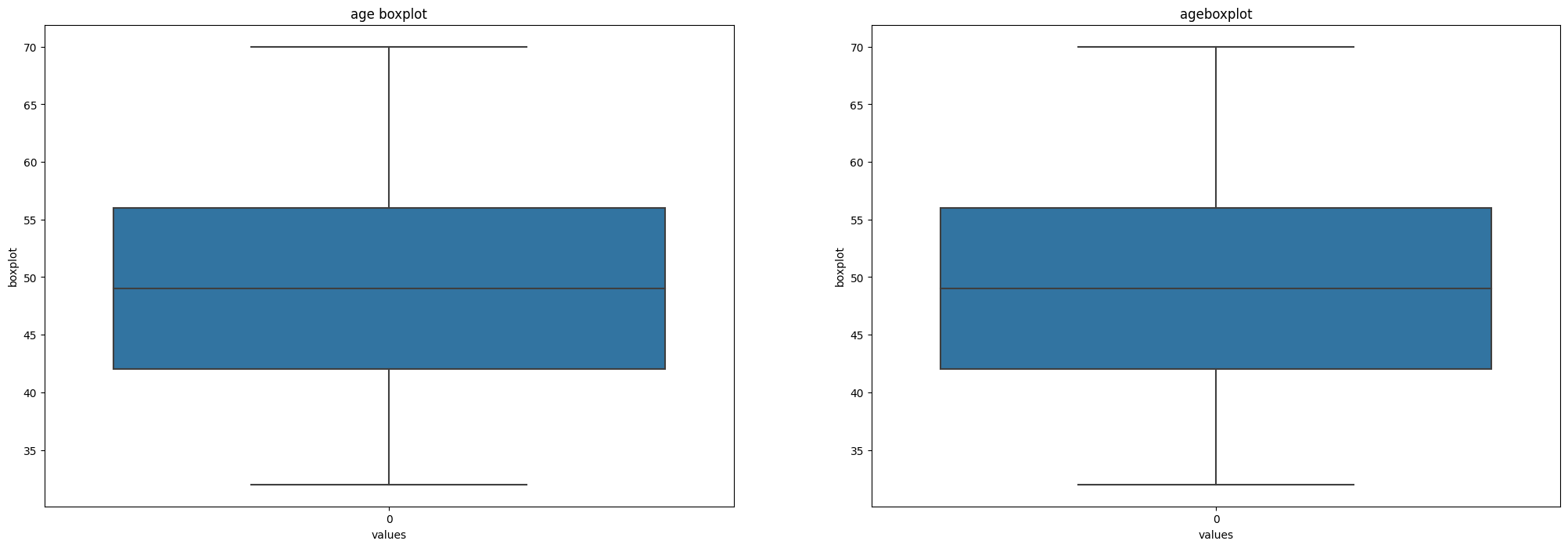
Descriptive analysis:

Box plot for outliers:

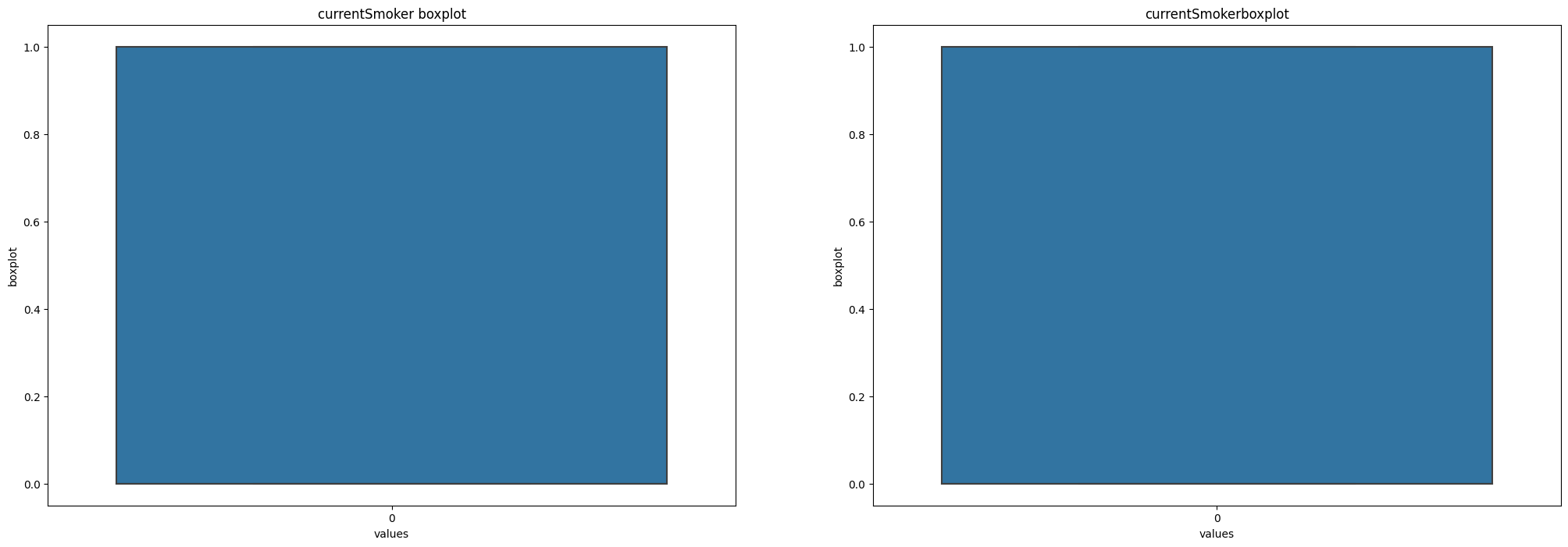
Correlation:

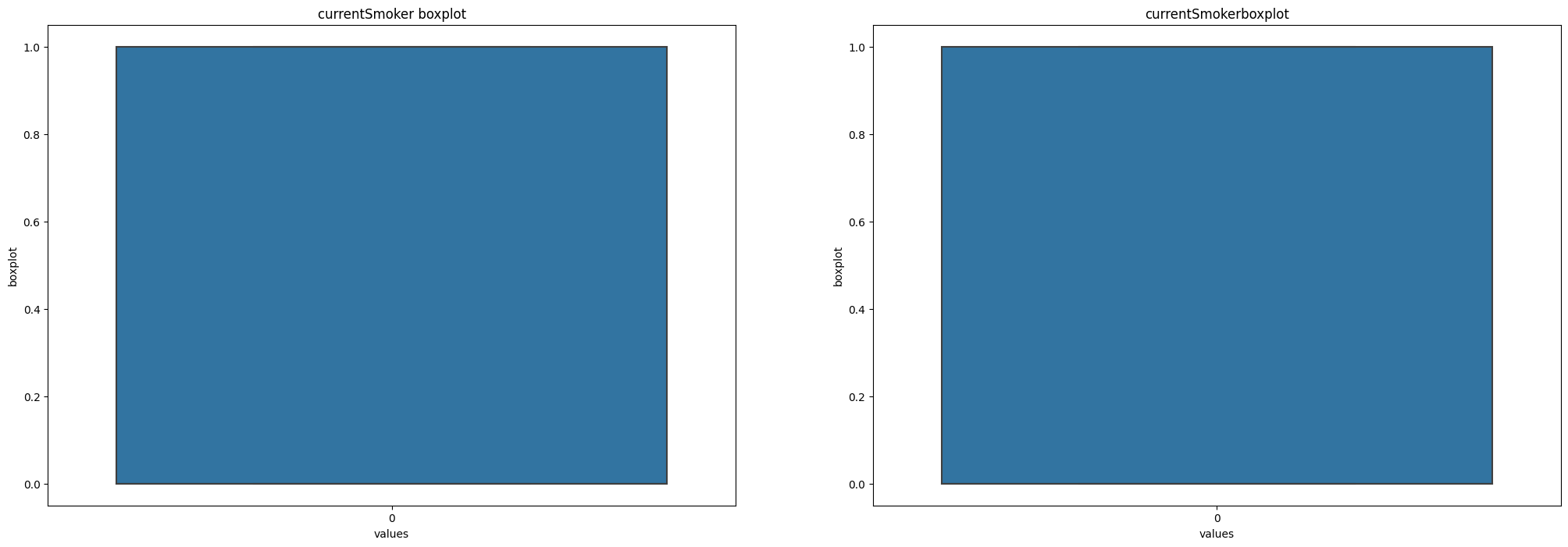
Outliers correction:

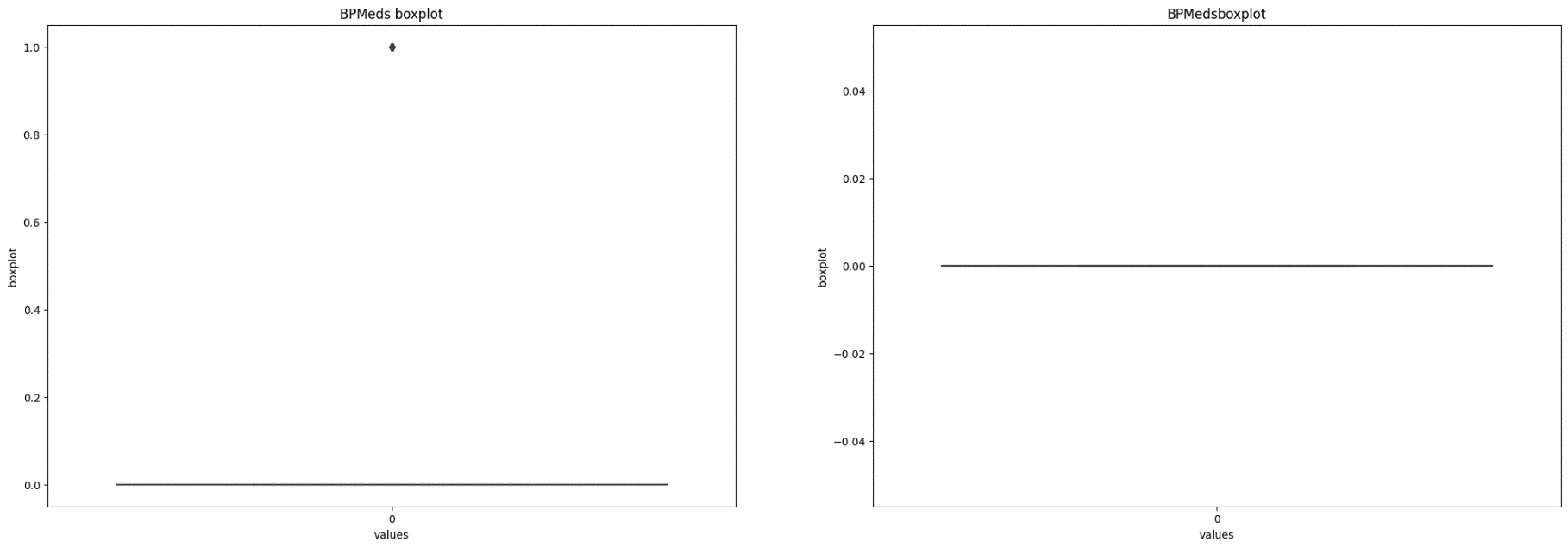
Gender: Feature Gender does not have any outliers

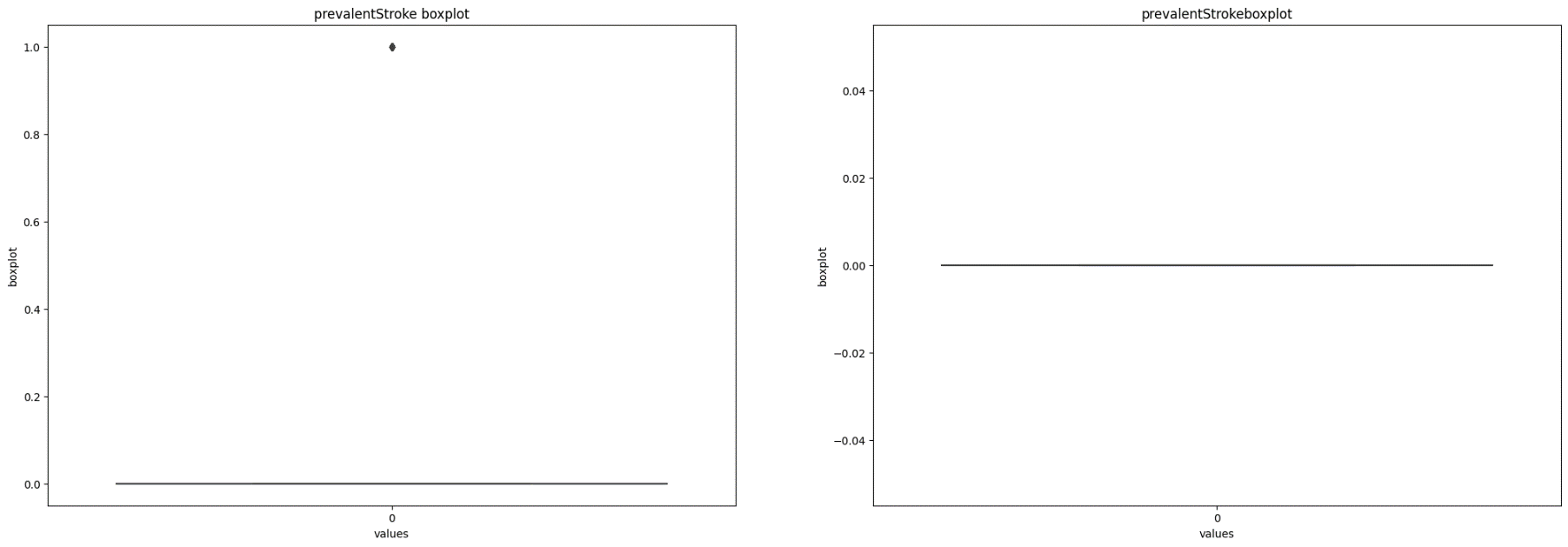
Age: Feature age does not have any outliers

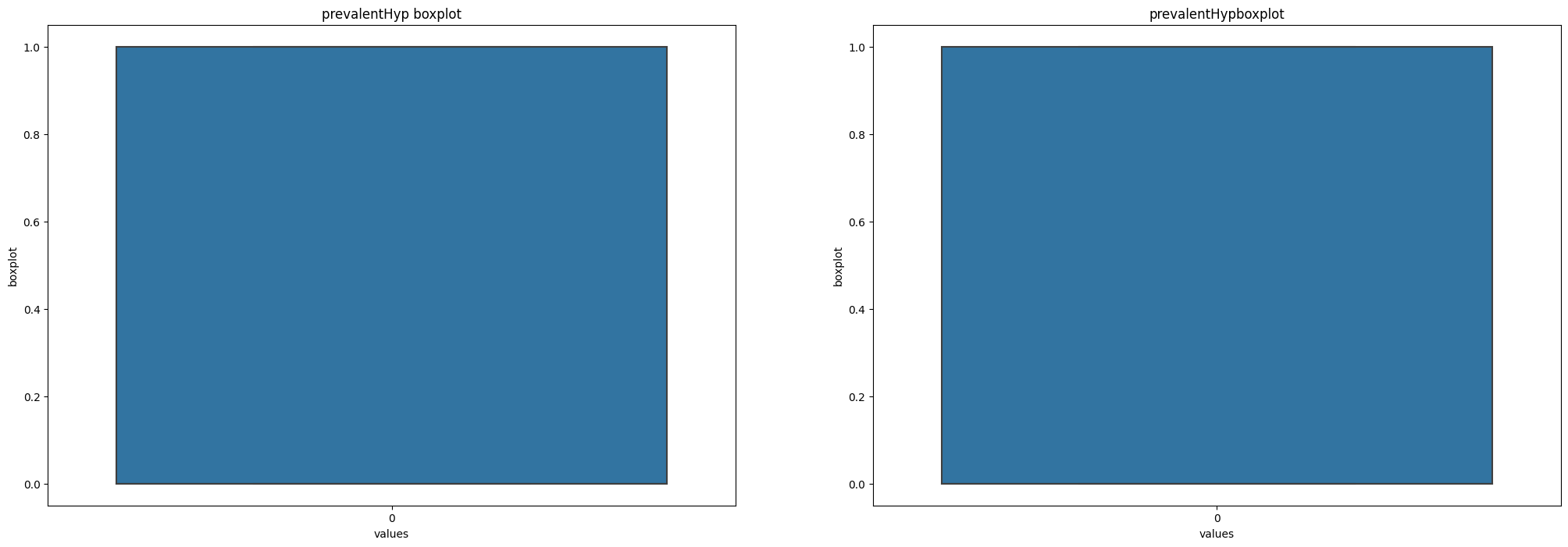
Education: Feature education does not have any outliers

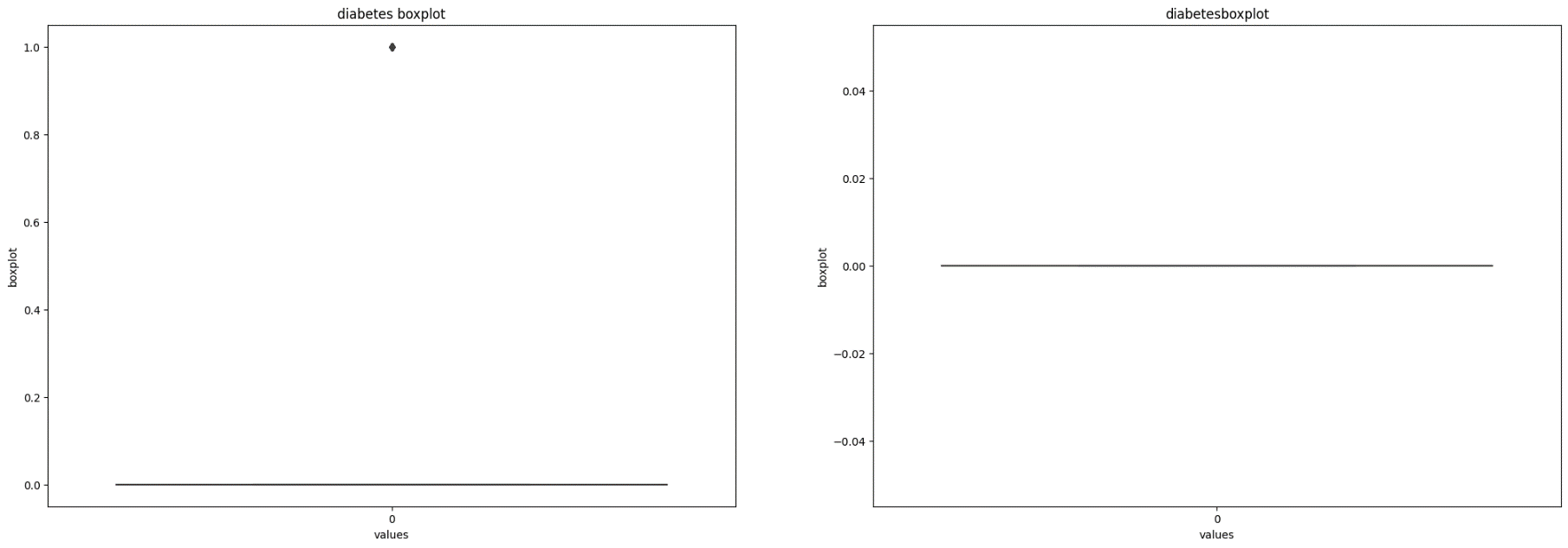
Current smoker: Feature currentSmoker does not have any outliers

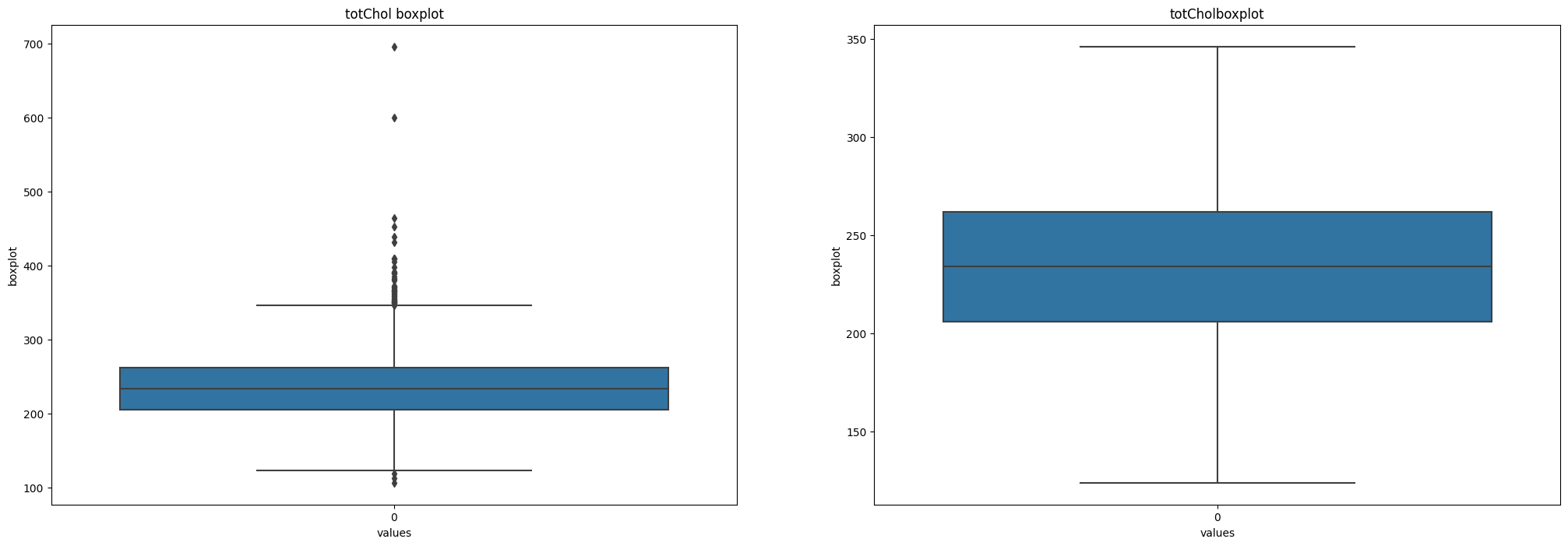
Cigs per day: Feature cigsPerDay has outliers

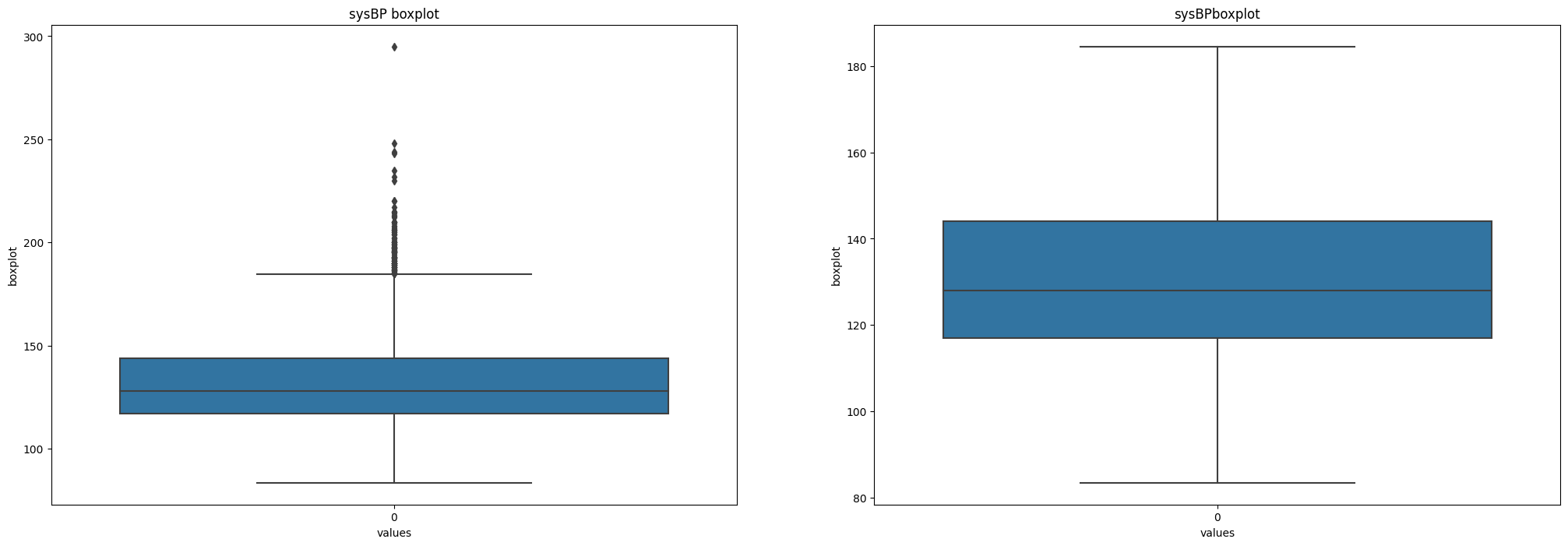
BP meds: feature BPMeds has outliers

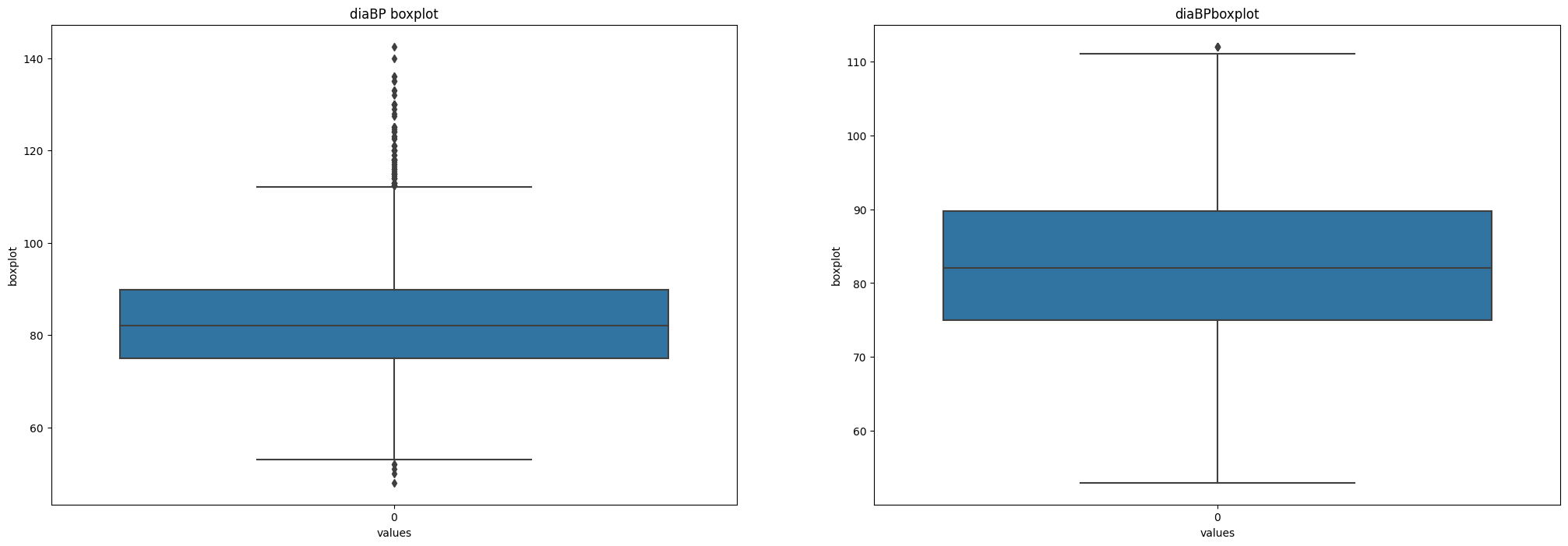
Prevalent Stroke: feature prevalentStroke has outliers

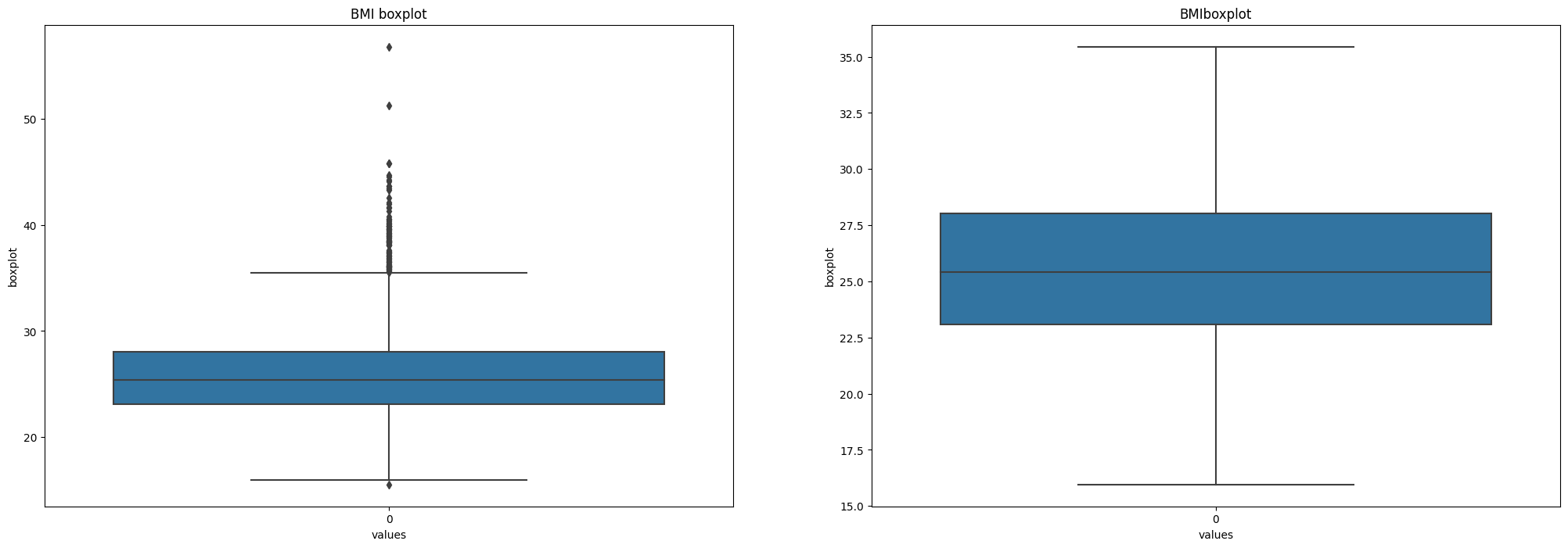
PrevalentHYP: feature prevalentHyp does not have any outliers

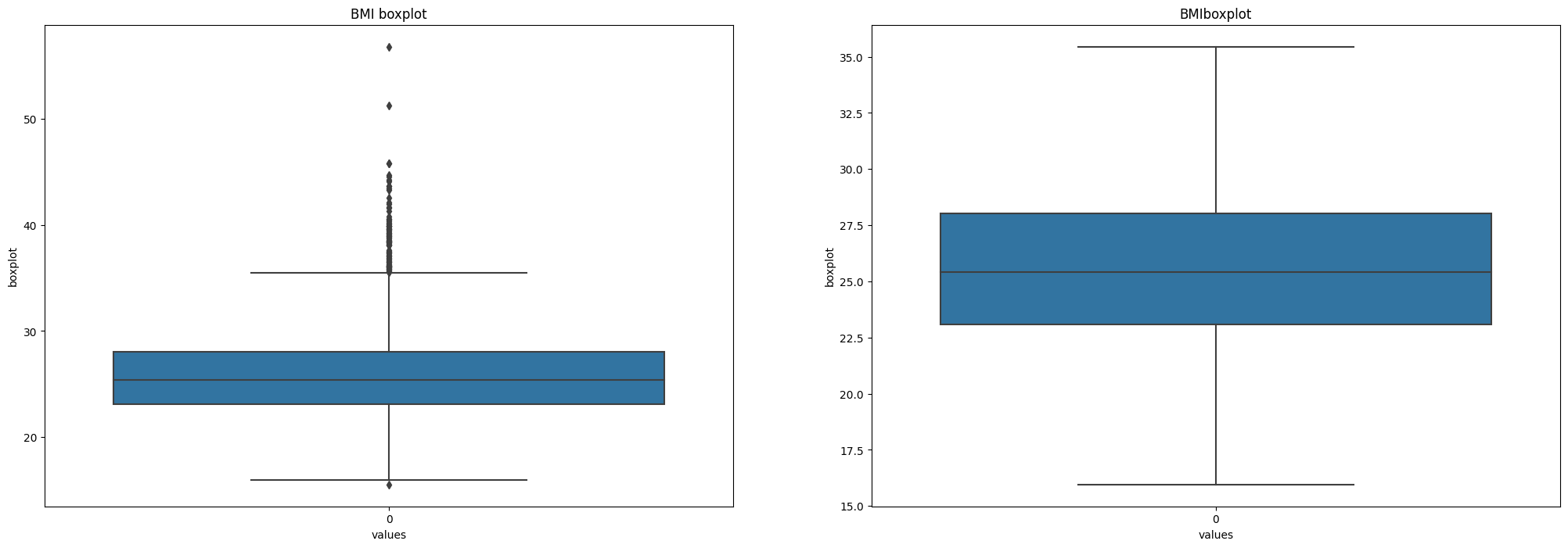
Diabetes: feature diabetes has outliers

totChol: feature totChol has outliers

sysBP: feature sysBP has outliers

diaBP: feature diaBP has outliers

BMI: feature BMI has outliers

HeartRate: feature heartRate has outliers

Applying PCA to eliminate features with higher correlation:

Explained variance ratio after PCA is: [0.24840278 0.40636 0.50189688 0.59010986 0.67416949 0.7552134 ,0.82839909 0.88517931 0.93313903]

PCA table:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PC\_1 | PC\_2 | PC\_3 | PC\_4 | PC\_5 | PC\_6 | PC\_7 | PC\_8 | PC\_9 |
| -1.84583 | -1.0584 | -0.71632 | 1.970003 | -0.563263 | 0.829554 | -1.288571 | 1.172637 | 0.199029 |
| 0.153157 | -1.18968 | 1.221128 | 0.215011 | -0.25428 | -0.48494 | -1.966735 | 0.456413 | 0.273427 |
| -0.93083 | 1.506166 | -0.689592 | -0.90883 | 0.324003 | -0.71256 | -0.373691 | 0.366742 | -0.39413 |
| 1.796789 | 1.67594 | 0.284366 | 0.020195 | -1.019271 | 2.031391 | 1.687903 | -1.701849 | 1.653977 |
| -0.71272 | 1.062352 | 1.911223 | -0.182913 | 0.526993 | 1.313516 | -0.340716 | -0.509729 | -0.14712 |

Hyperparameters:

DecisionTreeClassifier: criterion= 'log\_loss', max\_depth= 3, max\_features= 3

KNeighborsClassifier: weights ='uniform', algorithm ='auto'

RandomForestClassifier: max\_depth= 9, max\_features=4, n\_estimators= 100

SVC: gamma='scale',C= 0.1

AdaBoostClassifier: n\_estimators= 200 ,learning\_rate=0.1

GradientBoostingClassifier(n\_estimators=50 ,learning\_rate= 0.1

XGBClassifier(eta=0.5 ,max\_depth= 10,gamma=0 ,reg\_lambda = 10,alpha=10

**Model building and evaluation:**

|  |  |  |
| --- | --- | --- |
|  | **model\_name** | **accuracy\_score** |
| 0 | LogisticRegression | 0.869497 |
| 4 | SVC | 0.869497 |
| 3 | RandomForestClassifier | 0.867925 |
| 6 | GradientBoostingClassifier | 0.867138 |
| 5 | AdaBoostClassifier | 0.856918 |
| 7 | XGBClassifier | 0.847484 |
| 2 | KneighborClassifier | 0.841981 |
| 1 | DecisionTreeClassifier | 0.764937 |

**Applying Cross validation:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | model\_names | cv\_score | cv\_std |
| 0 | Logisticregression | 0.849928 | 0.003728 |
| 4 | SVC | 0.847805 | 0.001189 |
| 6 | GradientBoostingClassifier | 0.847098 | 0.003764 |
| 5 | AdaBoostClassifier | 0.844734 | 0.00823 |
| 3 | RandomForestClassifier | 0.843086 | 0.00764 |
| 7 | XGBClassifier | 0.838365 | 0.005575 |
| 2 | KNeighborsClassifier | 0.831055 | 0.009096 |
| 1 | DecisionTreeClassifier | 0.756009 | 0.016723 |

**Cross Validation post hyperparameter:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | model\_names | cv\_score | cv\_std |
| 0 | LogisticRegression | 0.849928 | 0.003728 |
| 4 | SVC | 0.848042 | 0.001105 |
| 3 | RandomForestClassifier | 0.84804 | 0.00374 |
| 5 | AdaBoostClassifier | 0.847569 | 0.003393 |
| 6 | GradientBoostingClassifier | 0.846862 | 0.003393 |
| 1 | DecisionTreeClassifier | 0.846626 | 0.002074 |
| 7 | XGBClassifier | 0.841198 | 0.007102 |
| 2 | KNeighborsClassifier | 0.831055 | 0.009096 |

Using Clusters to improve accuracy:

**Model building and evaluation:**

|  |  |  |
| --- | --- | --- |
|  | model\_name | accuracy\_score |
| 3 | RandomForestClassifier | 0.874214 |
| 0 | LogisticRegression | 0.871069 |
| 4 | SVC | 0.869497 |
| 6 | GradientBoostingClassifier | 0.867925 |
| 5 | AdaBoostClassifier | 0.860063 |
| 2 | KneighborClassifier | 0.857704 |
| 7 | XGBClassifier | 0.849843 |
| 1 | DecisionTreeClassifier | 0.761792 |

**Applying Cross validation:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | model\_names | cv\_score | cv\_std |
| 0 | Logisticregression | 0.850164 | 0.003588 |
| 3 | RandomForestClassifier | 0.848512 | 0.003372 |
| 4 | SVC | 0.848042 | 0.001105 |
| 5 | AdaBoostClassifier | 0.846863 | 0.003506 |
| 6 | GradientBoostingClassifier | 0.845918 | 0.005378 |
| 7 | XGBClassifier | 0.833176 | 0.004476 |
| 2 | KNeighborsClassifier | 0.830578 | 0.010357 |
| 1 | DecisionTreeClassifier | 0.756726 | 0.021117 |

**Cross Validation post hyperparameter:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | model\_names | cv\_score | cv\_std |
| 7 | XGBClassifier | 0.851343 | 0.004674 |
| 0 | LogisticRegression | 0.850164 | 0.003588 |
| 3 | RandomForestClassifier | 0.848984 | 0.002654 |
| 5 | AdaBoostClassifier | 0.84875 | 0.002184 |
| 1 | DecisionTreeClassifier | 0.848042 | 0.001472 |
| 4 | SVC | 0.848042 | 0.001105 |
| 6 | GradientBoostingClassifier | 0.847807 | 0.004216 |
| 2 | KNeighborsClassifier | 0.830578 | 0.010357 |

Using feature selection to remove features contributing less:

Result:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | age | cigsPerDay | totChol | sysBP | diaBP | BMI | heartRate | glucose |
| 0 | 39 | 0 | 195 | 106 | 70 | 26.97 | 80 | 77 |
| 1 | 46 | 0 | 250 | 121 | 81 | 28.73 | 95 | 76 |
| 2 | 48 | 20 | 245 | 127.5 | 80 | 25.34 | 75 | 70 |
| 3 | 61 | 30 | 225 | 150 | 95 | 28.58 | 65 | 103 |
| 4 | 46 | 23 | 285 | 130 | 84 | 23.1 | 85 | 85 |

Model evaluation on new model:

**Model building and evaluation:**

|  |  |  |
| --- | --- | --- |
|  | model\_name | accuracy\_score |
| 0 | LogisticRegression | 0.871069 |
| 4 | SVC | 0.869497 |
| 3 | RandomForestClassifier | 0.868711 |
| 5 | AdaBoostClassifier | 0.865566 |
| 6 | GradientBoostingClassifier | 0.860849 |
| 2 | KneighborClassifier | 0.858491 |
| 7 | XGBClassifier | 0.851415 |
| 1 | DecisionTreeClassifier | 0.771226 |

**Applying Cross validation:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | model\_names | cv\_score | cv\_std |
| 0 | Logisticregression | 0.851108 | 0.004668 |
| 4 | SVC | 0.848042 | 0.001105 |
| 5 | AdaBoostClassifier | 0.84639 | 0.006008 |
| 3 | RandomForestClassifier | 0.845208 | 0.00639 |
| 6 | GradientBoostingClassifier | 0.844265 | 0.003223 |
| 7 | XGBClassifier | 0.829873 | 0.006938 |
| 2 | KNeighborsClassifier | 0.829398 | 0.010434 |
| 1 | DecisionTreeClassifier | 0.755543 | 0.019462 |

**Cross Validation post hyperparameter:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | model\_names | cv\_score | cv\_std |
| 0 | LogisticRegression | 0.851108 | 0.004668 |
| 7 | XGBClassifier | 0.850872 | 0.004724 |
| 5 | AdaBoostClassifier | 0.849929 | 0.002403 |
| 3 | RandomForestClassifier | 0.849694 | 0.003452 |
| 4 | SVC | 0.848042 | 0.001105 |
| 6 | GradientBoostingClassifier | 0.847096 | 0.005399 |
| 1 | DecisionTreeClassifier | 0.846863 | 0.003368 |
| 2 | KNeighborsClassifier | 0.829398 | 0.010434 |