IE0005 Introduction To Data Science & Artificial Intelligence

PREDICTORS OF CARDIOVASCULAR DISEASE

EL06 Group 7
Phua Wei An
Pagdanganan Robert Martin Gosioco
Tan Chuan Bing
Nguyen Hoang Minh

CONTENT

Dataset:

- Cardiovascular Disease Prediction

Objective:

- To build a prediction model to determine the variable that best indicates the likelihood of cardiovascular disease
- Suggest to low-income countries what equipment and testing methodology to channel limited hospital resources into, for early detection and prevention of cardiovascular disease in these lower-income places

1. Initial Data Preparation
2. Exploratory Analysis & Further Prep
3. Machine Learning Techniques
4. Findings

1: INITIAL DATA PREPARATION

DATA PREPARATION

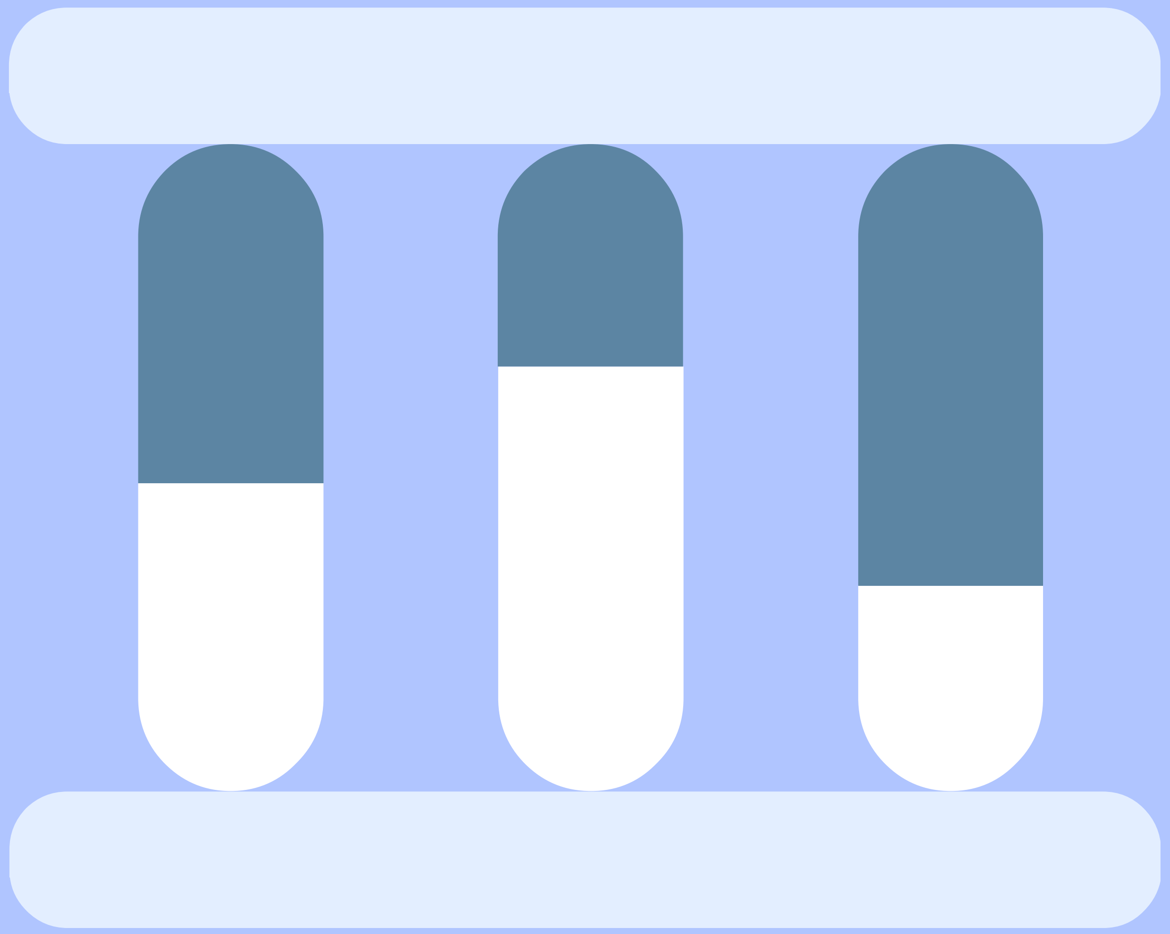
01 SIMPLIFY CATEGORICAL DATA FOR EASIER UNDERSTANDING

02 ADDED IN NEW POSSIBLE VARIABLE: BMI

03 CONVERT AGE FROM DAYS TO YEARS FOR EASIER UNDERSTANDING

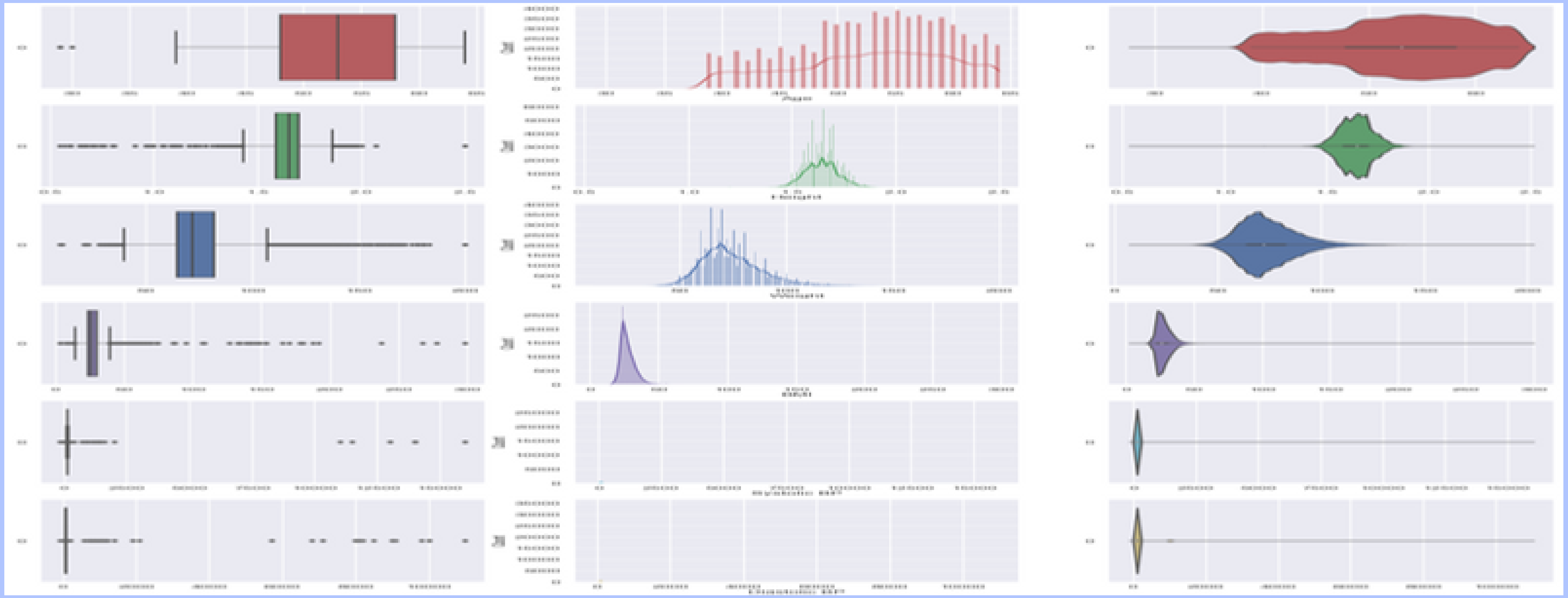
```python
#Replaced binary int with strings

cardioData['Gender'].replace([1,2],['Female','Male'],inplace=True)
cardioData['Smoke'].replace([0,1],['No','Yes'],inplace=True)
cardioData['Cholesterol'].replace([1,2,3],['Normal','Above Normal','Well Above Normal'],inplace=True)
cardioData['Glucose'].replace([1,2,3],['Normal','Above Normal', 'Well Above Normal'],inplace=True)
cardioData['Physical Activity'].replace([0,1],['No','Yes'],inplace=True)
cardioData['Cardiovascular Disease'].replace([0,1],['No','Yes'],inplace=True)
cardioData['Alcohol Intake'].replace([0,1],['No','Yes'],inplace=True)
cardioData['Age'] = (cardioData['Age']/365).astype(int) #Change age from days to years
cardioData['Height'] = (cardioData['Height']/100).astype(float)
cardioData['BMI'] = (cardioData['Weight']/(cardioData['Height']*cardioData['Height'])).round(2) #Add BMI
del cardioData['id'] #Removing "id" column from the dataset
```

2: EXPLORATORY DATA ANALYSIS AND OBSERVATION



EXPLORATORY ANALYSIS AND OBSERVATIONS

we then visualize the statistical distributions



Boxplot **Densityplot** **Violinplot**

EXPLORATORY ANALYSIS AND OBSERVATIONS

we then visualize the relations between pairs of variables using Seaborn pairplot

Pairplot

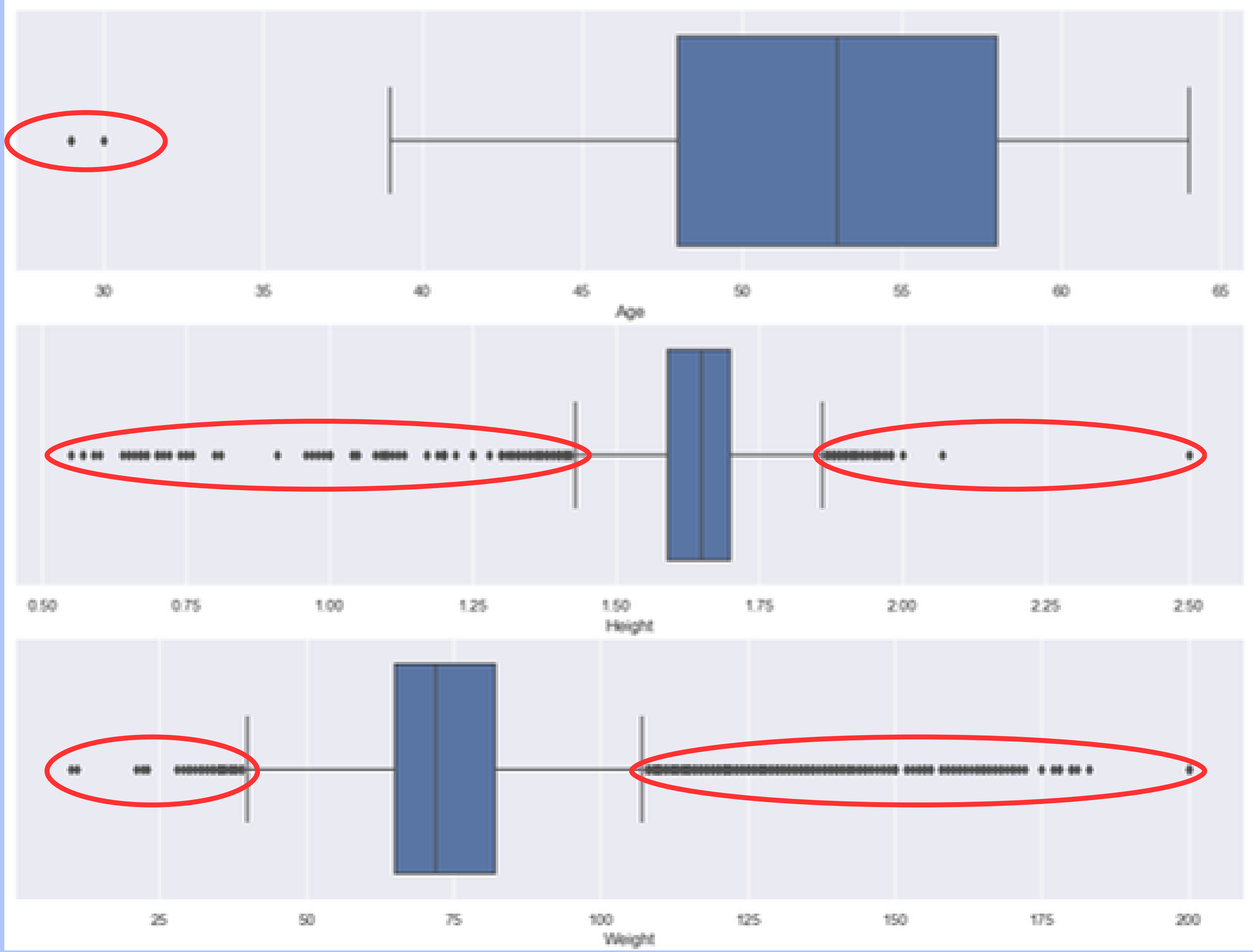
EXPLORATORY ANALYSIS AND OBSERVATIONS

visualizing the anomalies in the pairplot

Pairplot

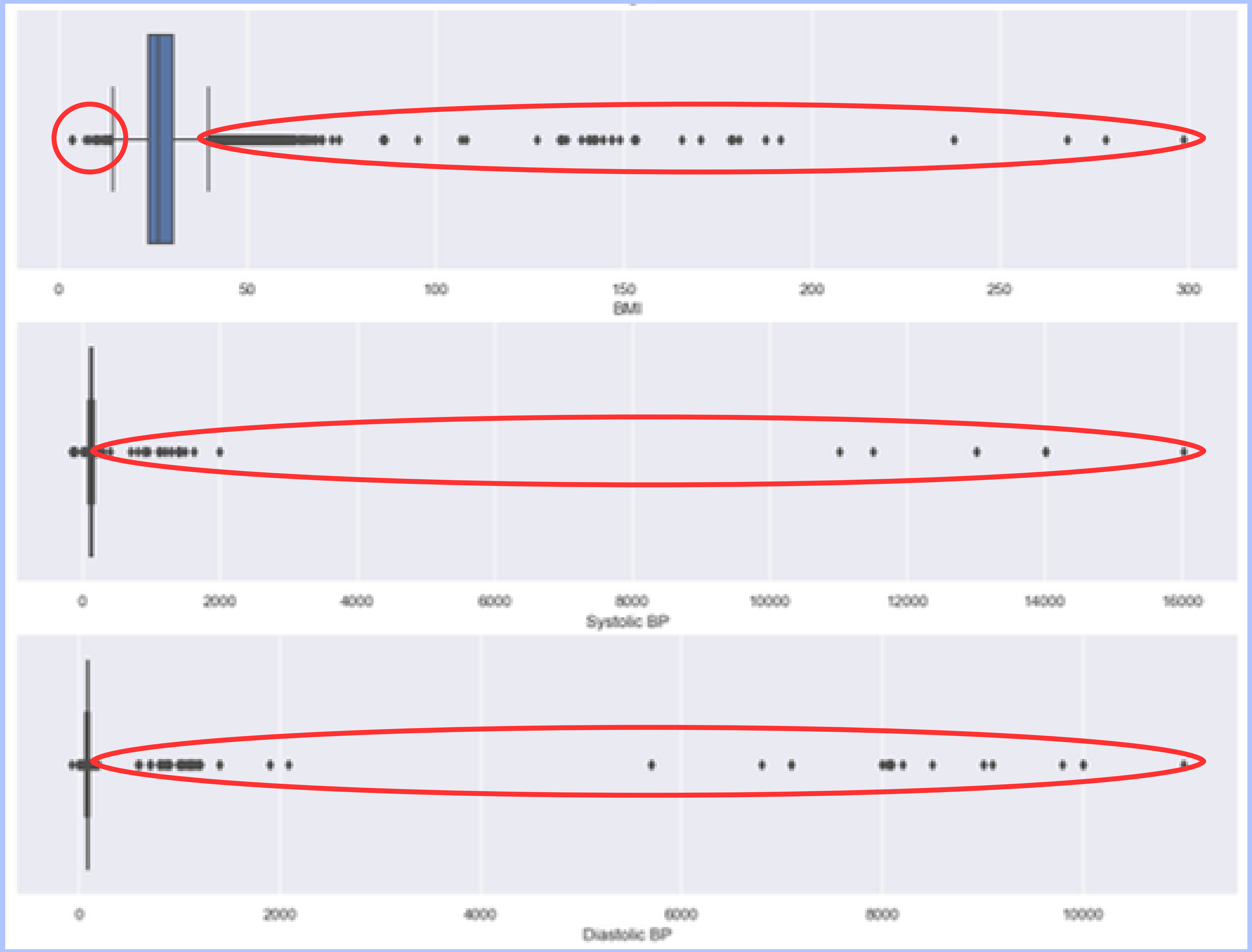
EXPLORATORY ANALYSIS AND OBSERVATIONS

After importing the dataset, we did boxplots for all 6 variables **before** cleaning



EXPLORATORY ANALYSIS AND OBSERVATIONS

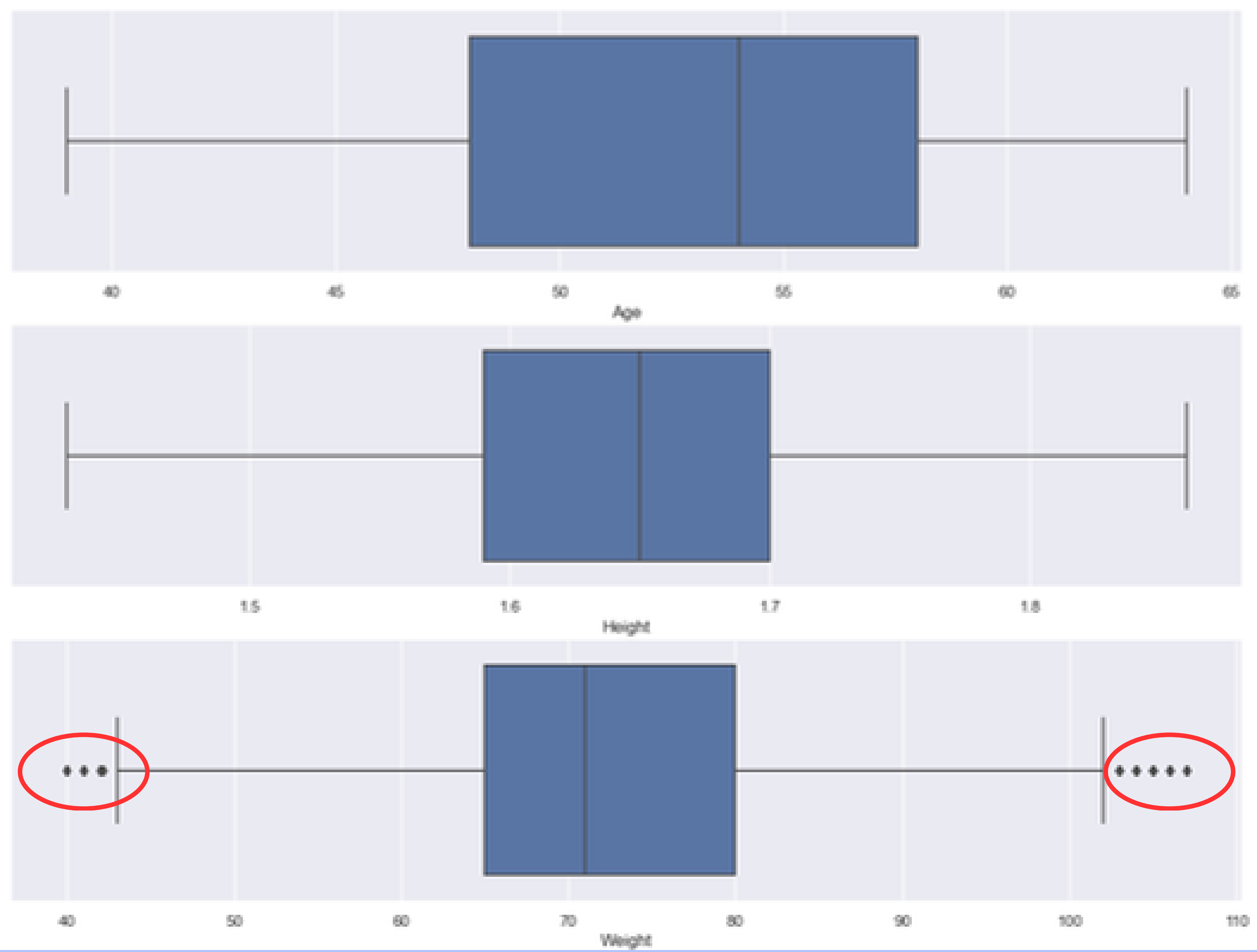
After importing the dataset, we did boxplots for all 6 variables **before** cleaning



EXPLORATORY ANALYSIS AND OBSERVATIONS

visualizing the variables using boxplot for the **cleaned** data

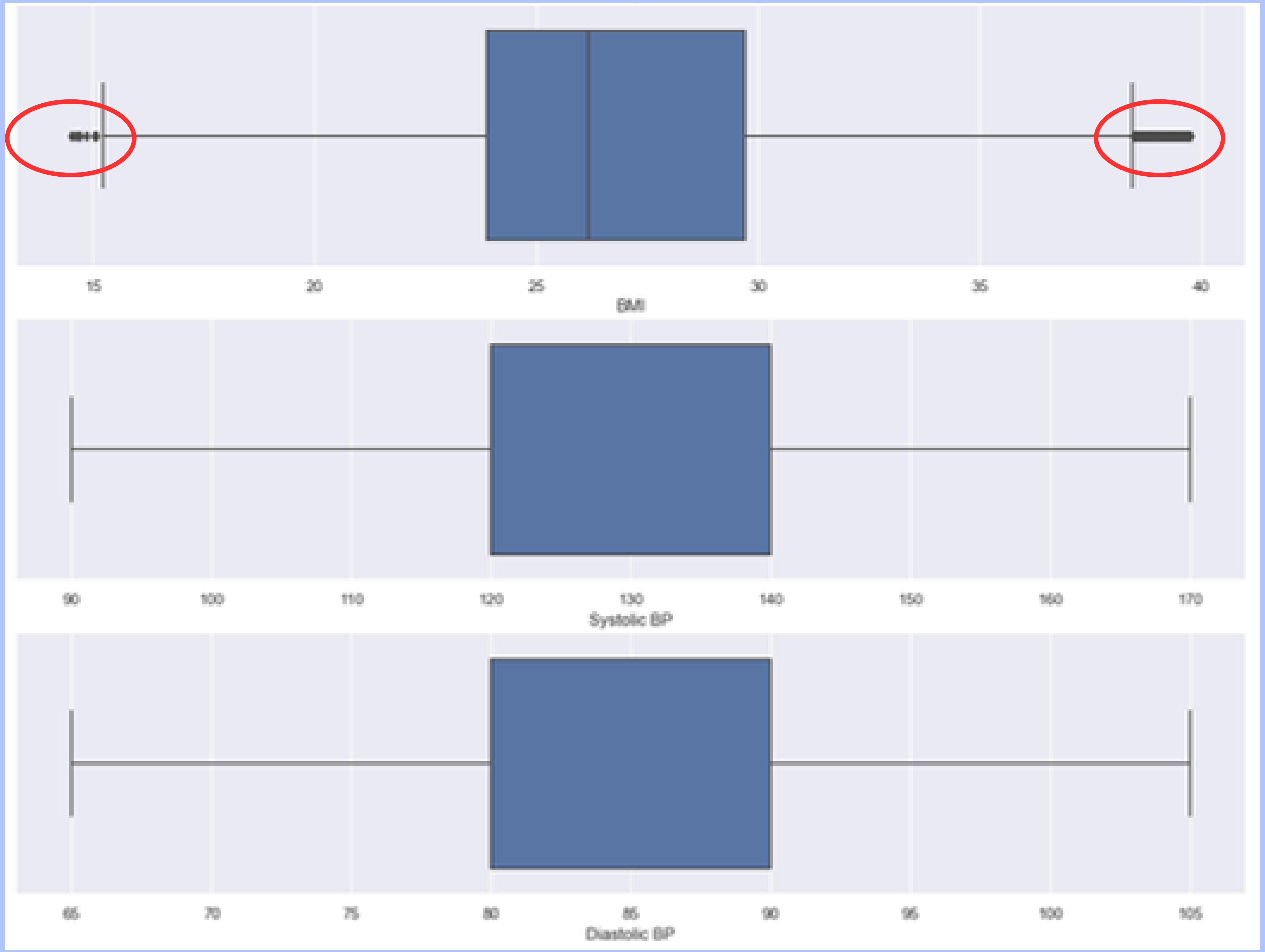
Boxplot



EXPLORATORY ANALYSIS AND OBSERVATIONS

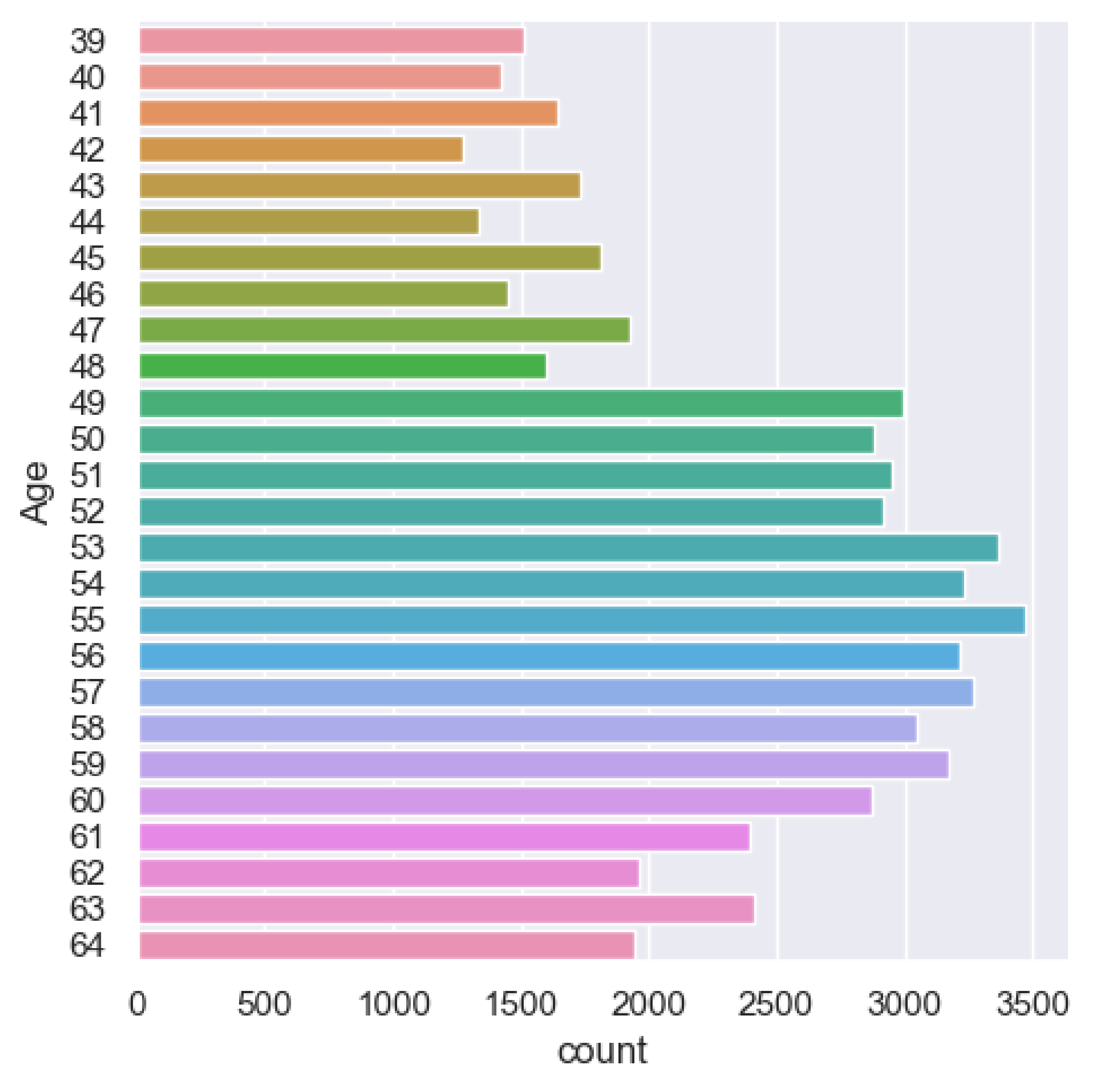
visualizing the variables using boxplot for the **cleaned** data

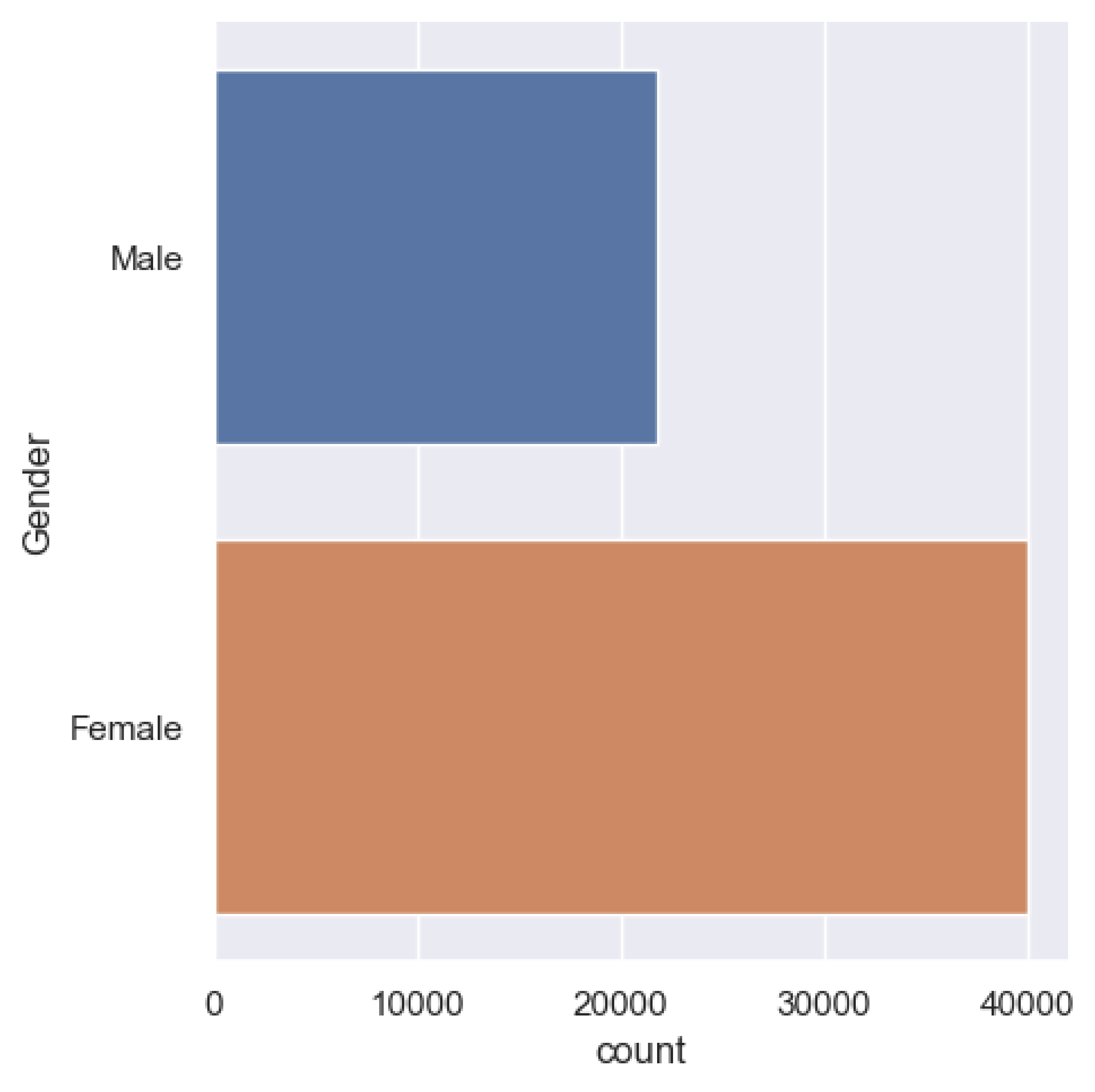
Boxplot

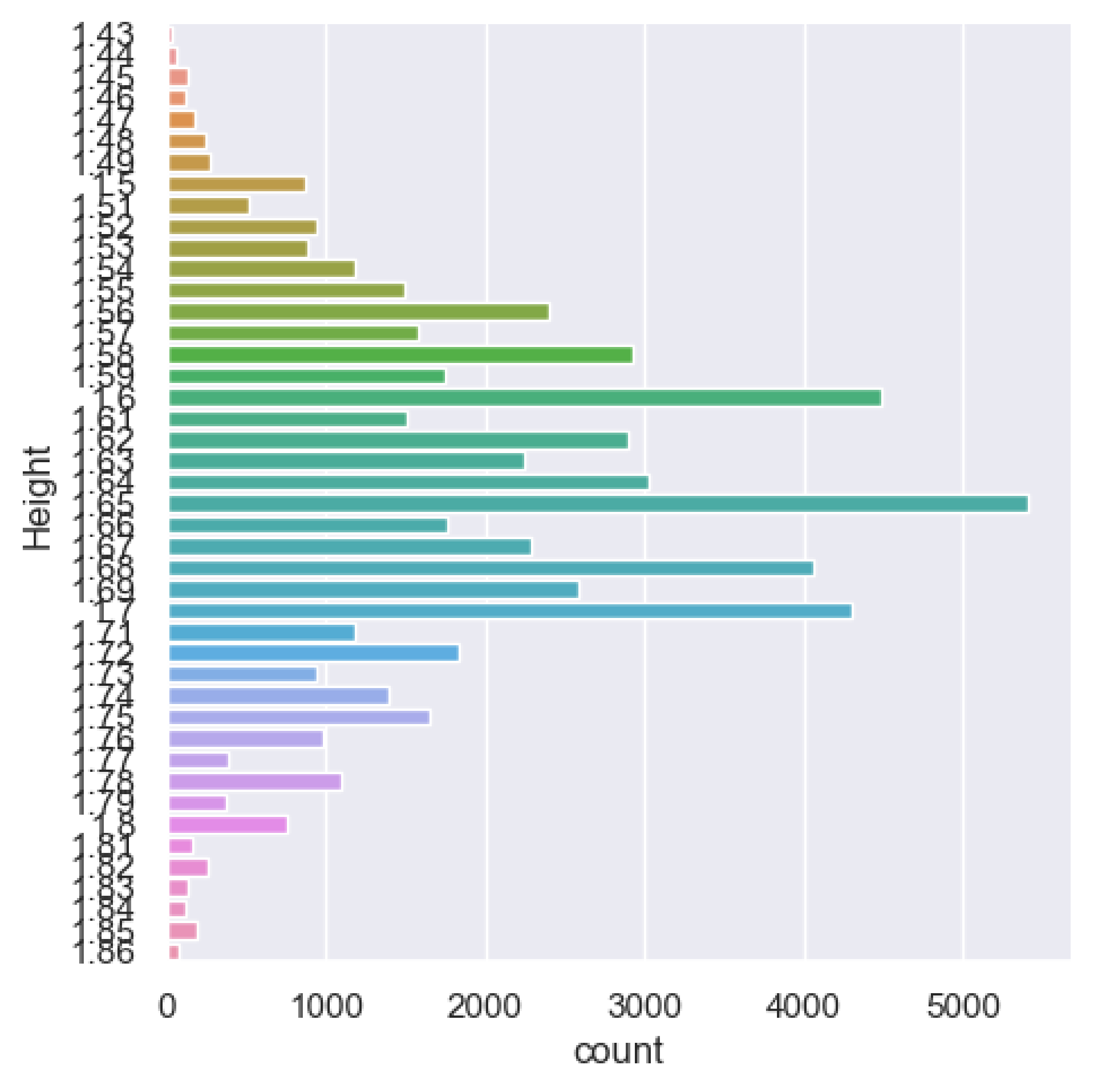


EXPLORATORY ANALYSIS AND OBSERVATIONS

we then visualize the distributions for the variables using catplot

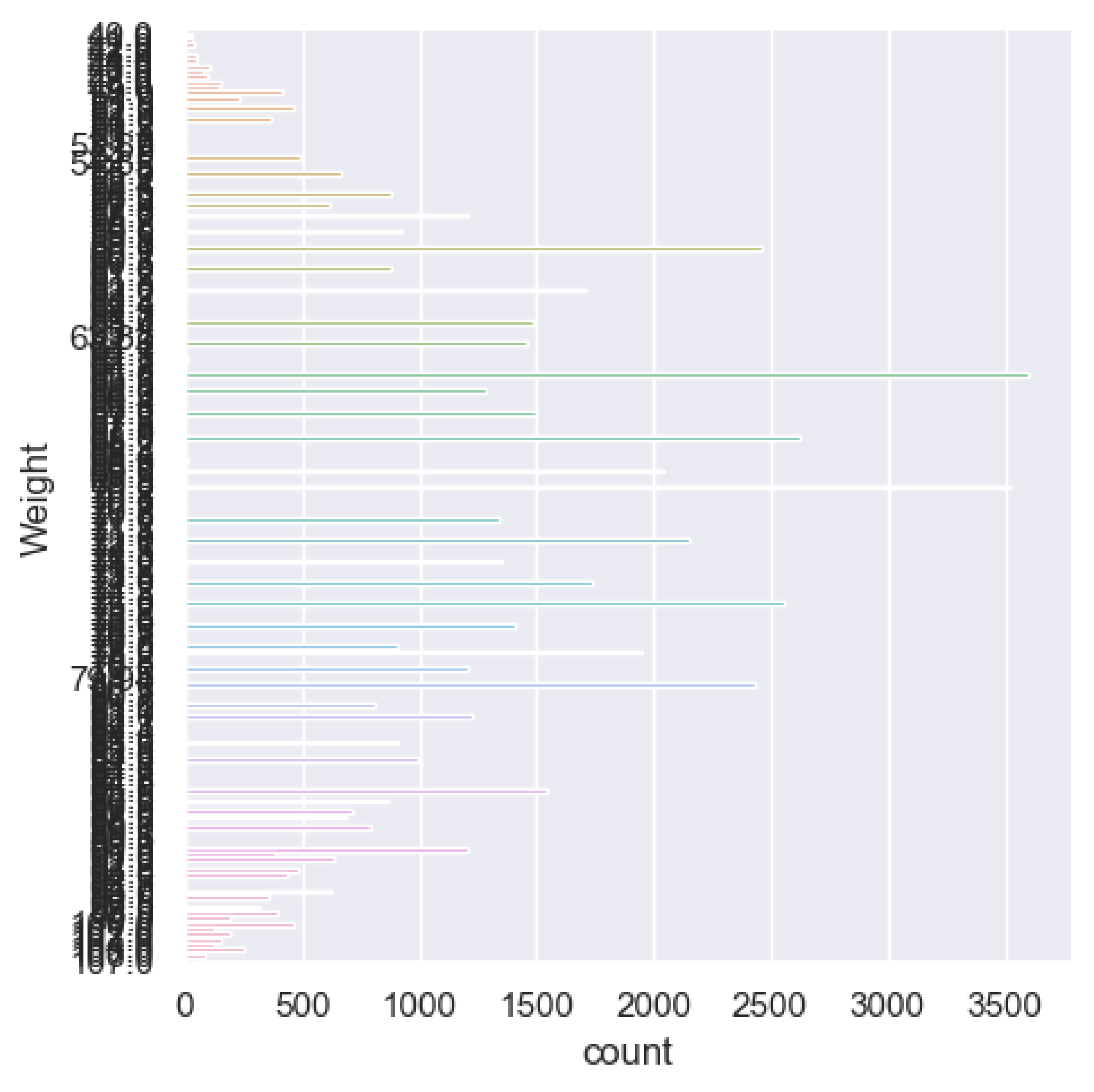


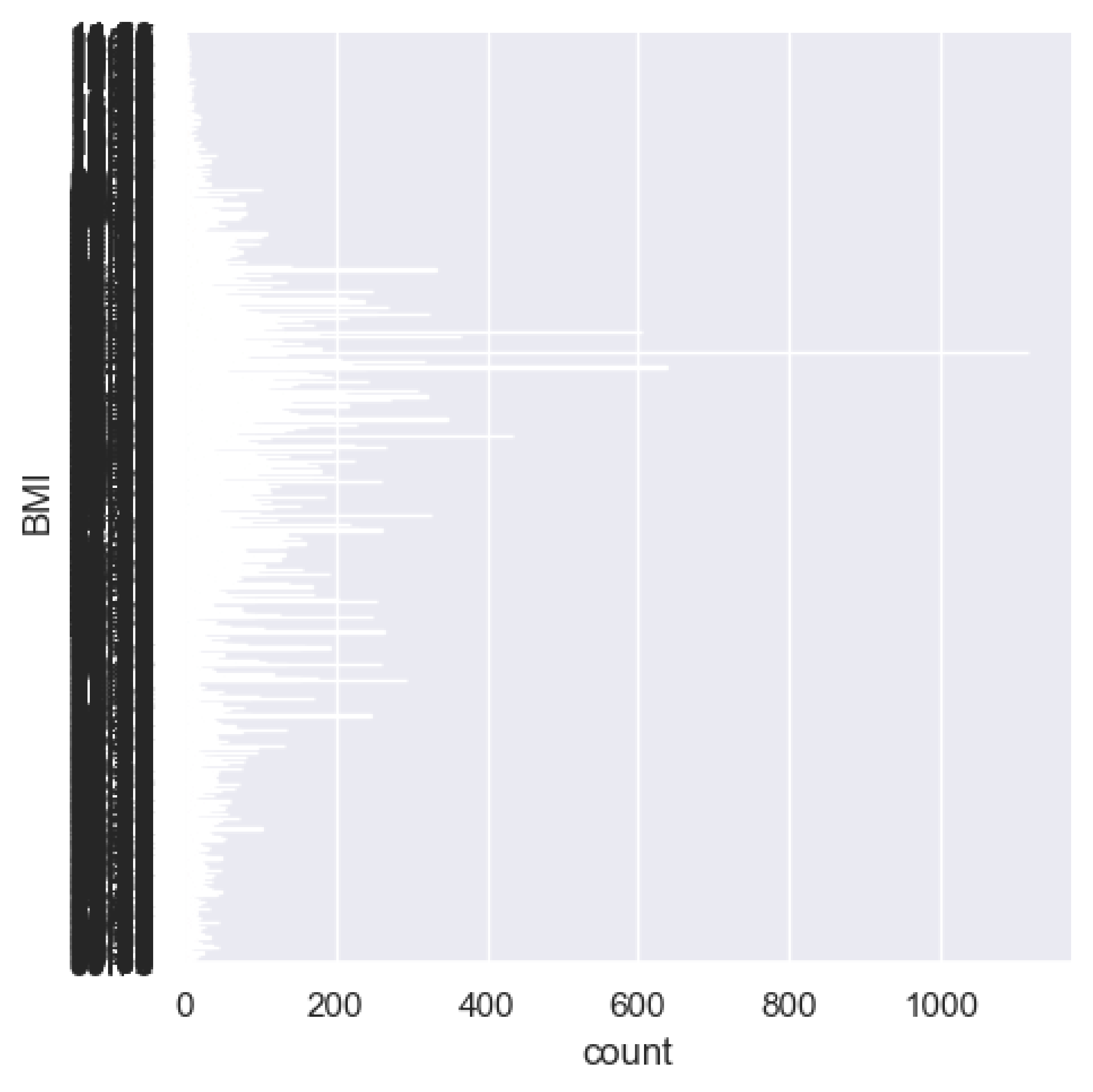


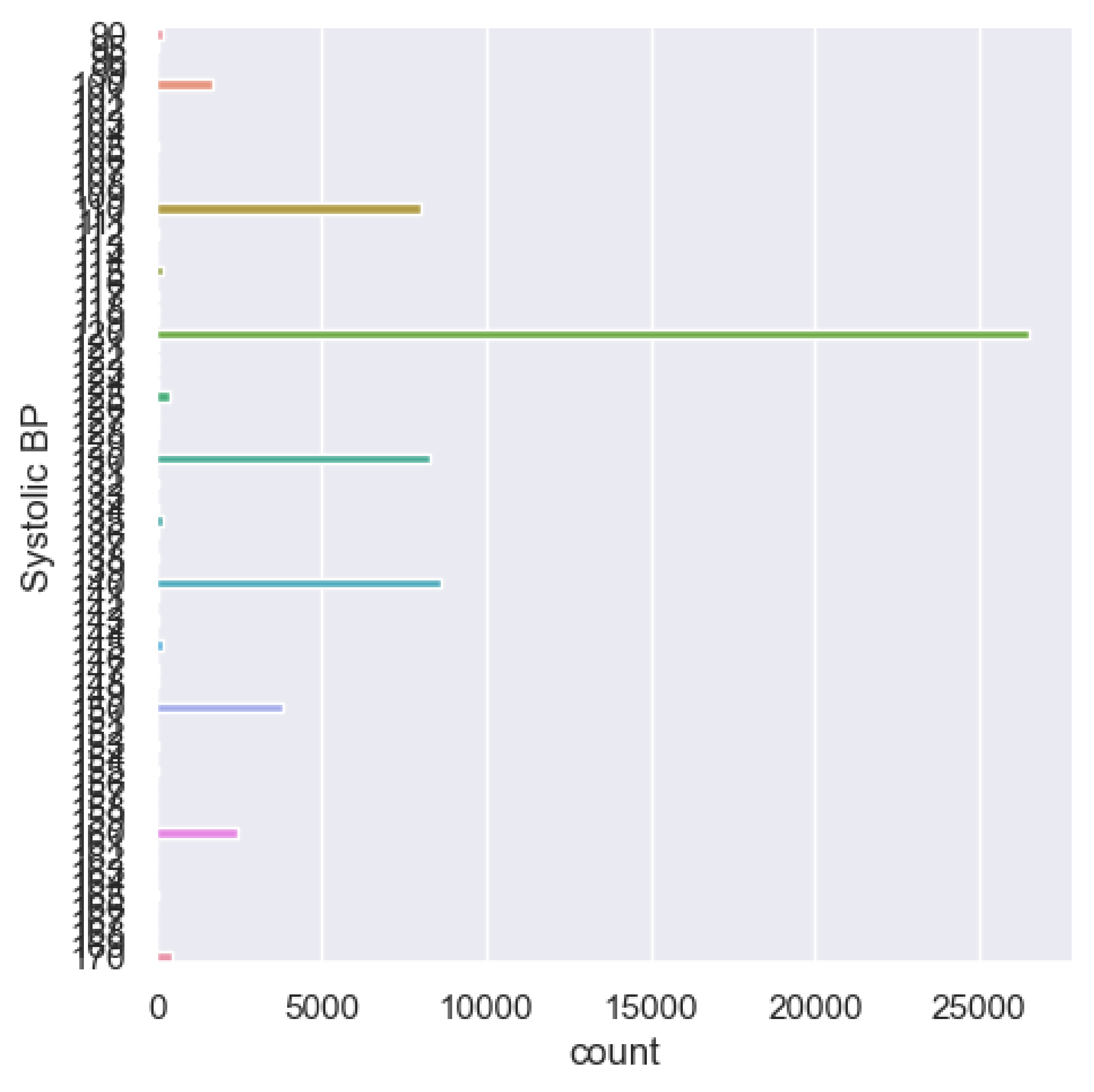


EXPLORATORY ANALYSIS AND OBSERVATIONS

we then visualize the distributions for the variables using catplot

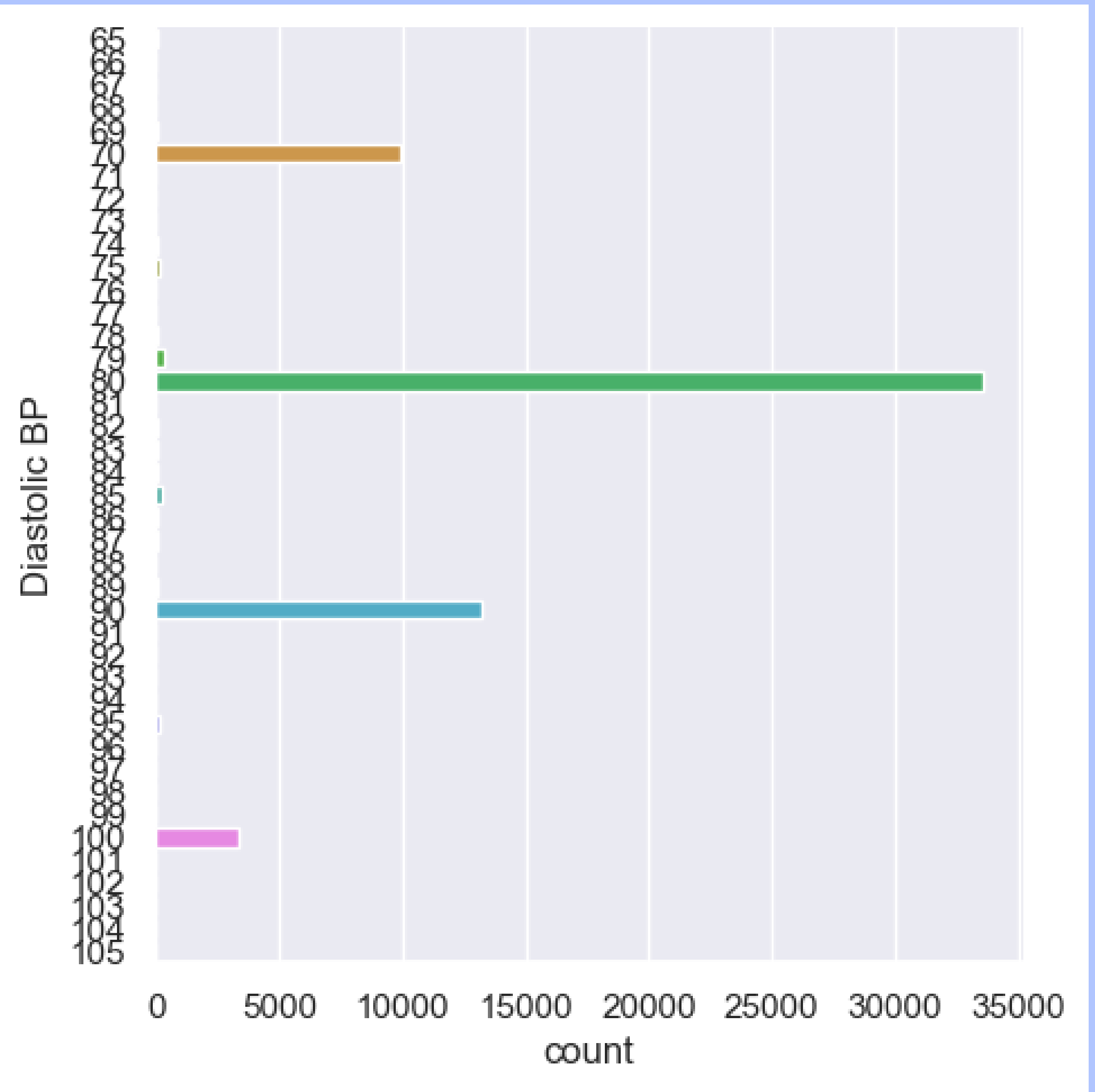


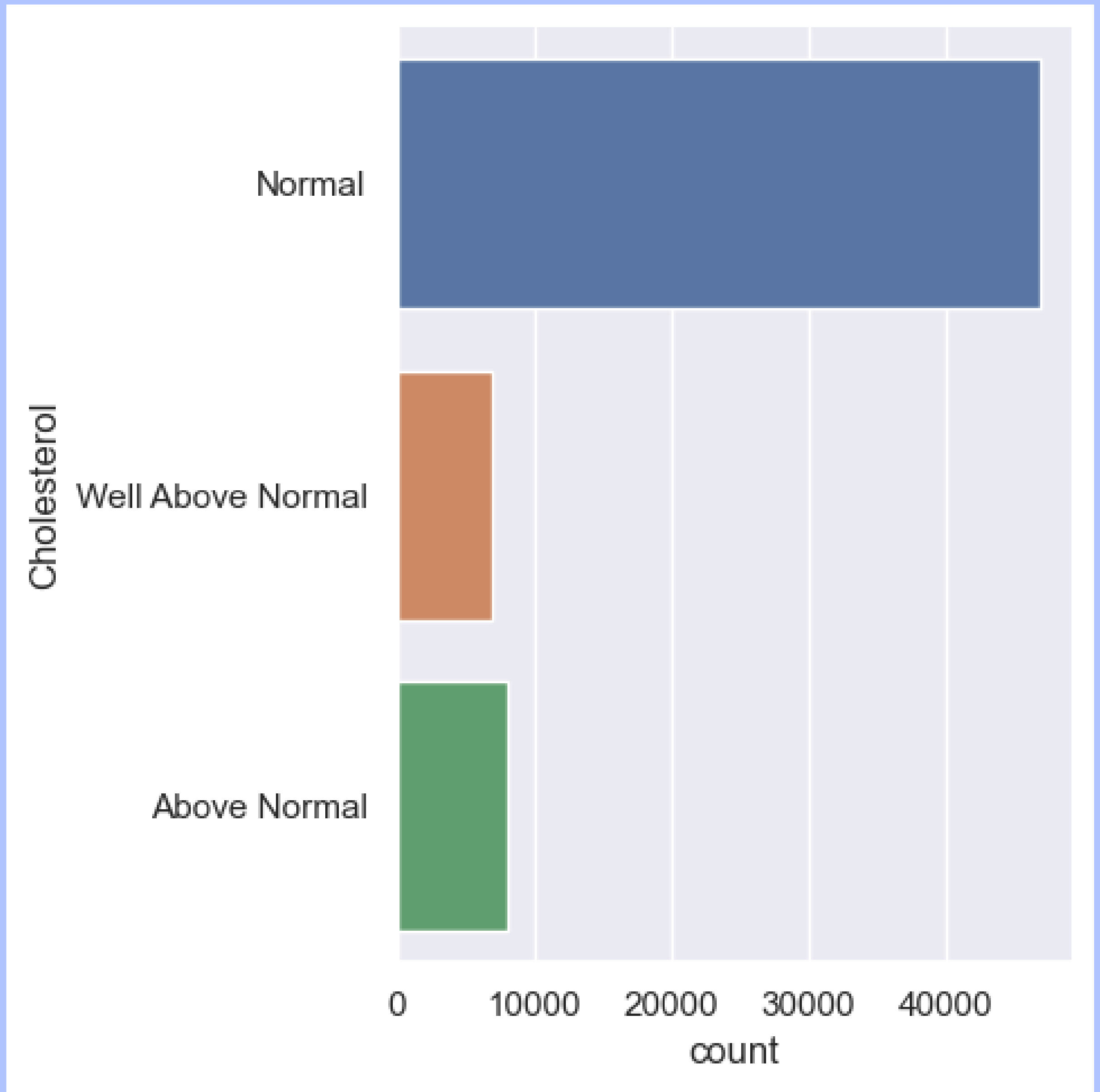


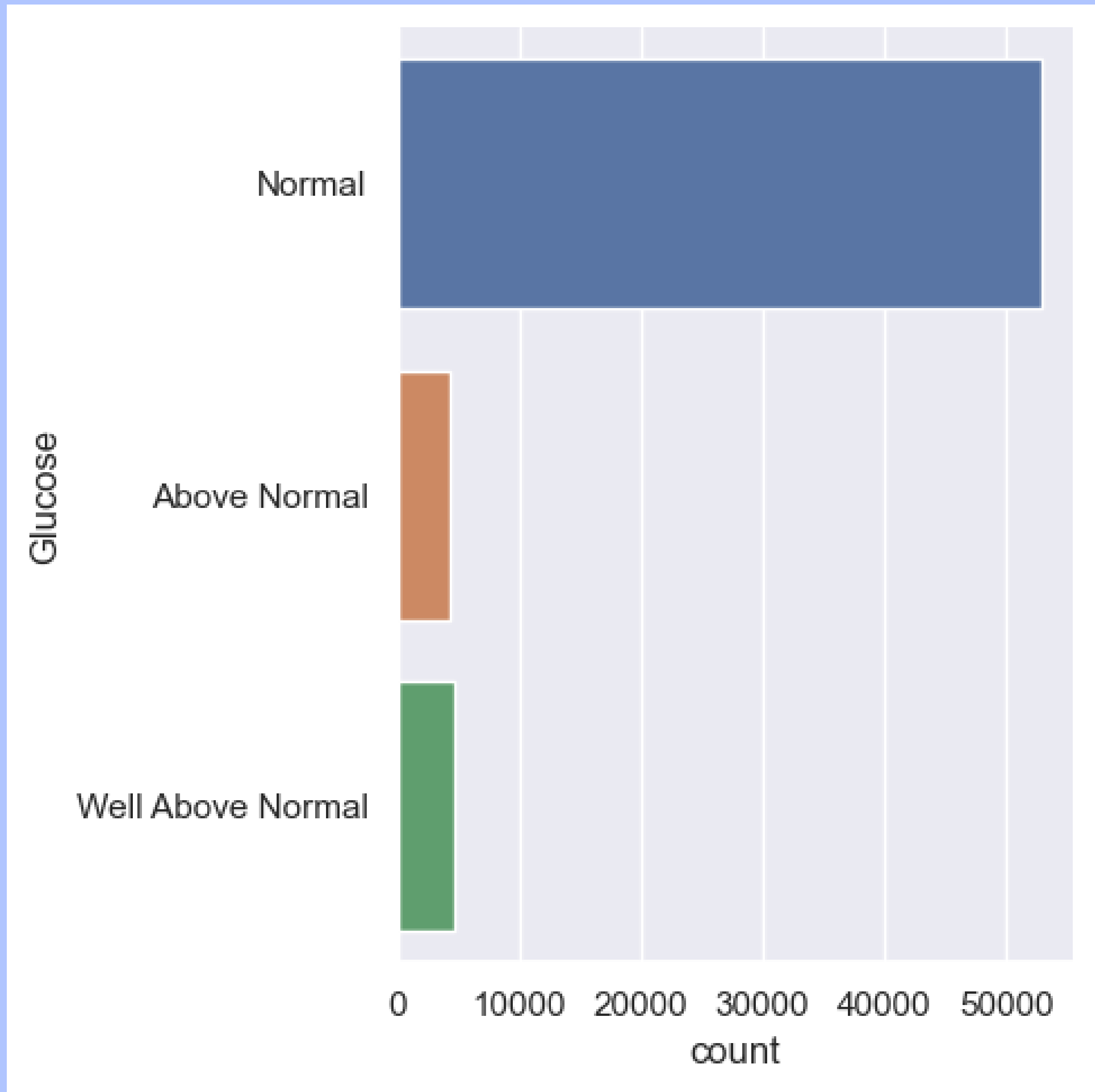


EXPLORATORY ANALYSIS AND OBSERVATIONS

we then visualize the distributions for the variables using catplot

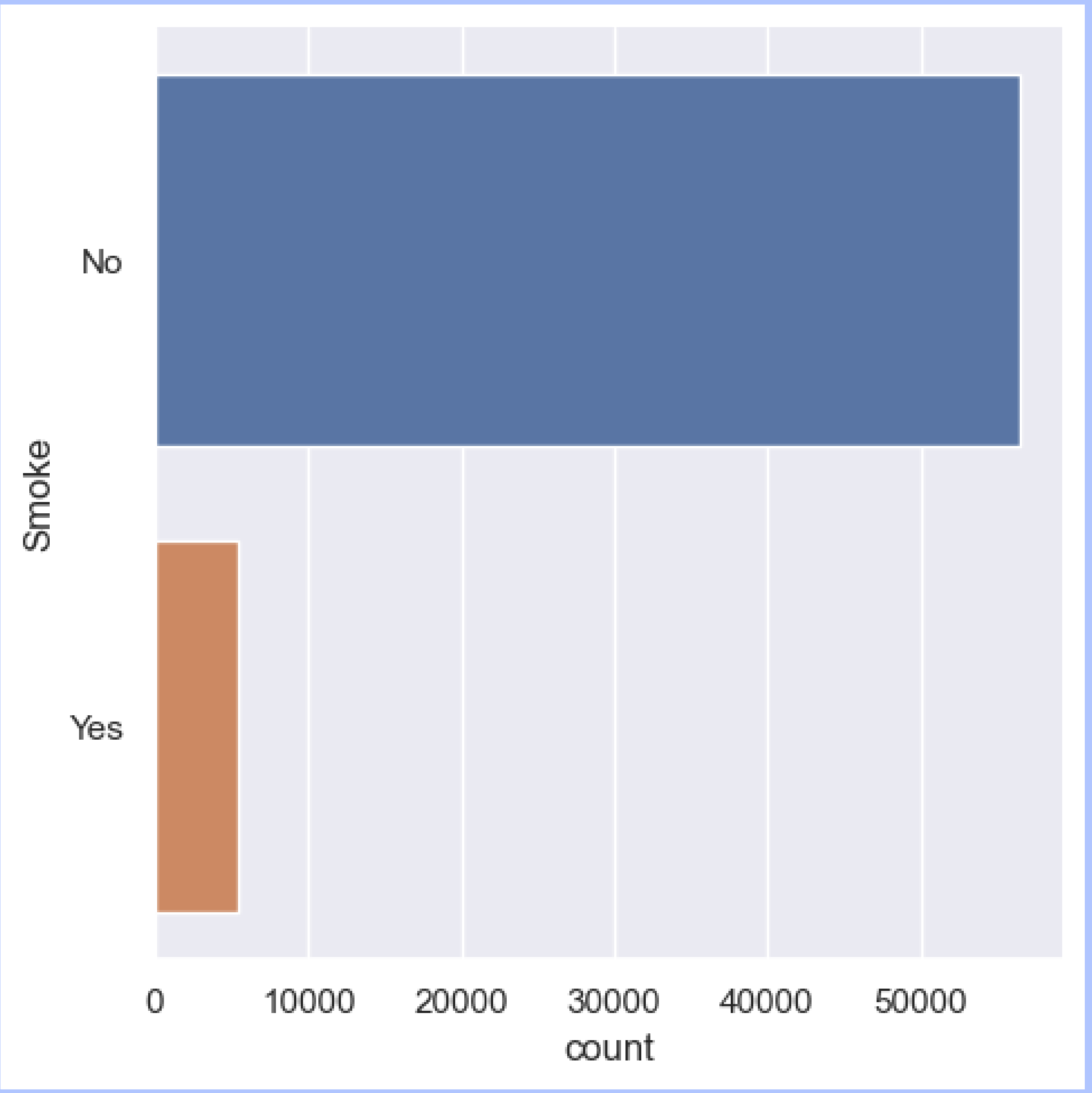


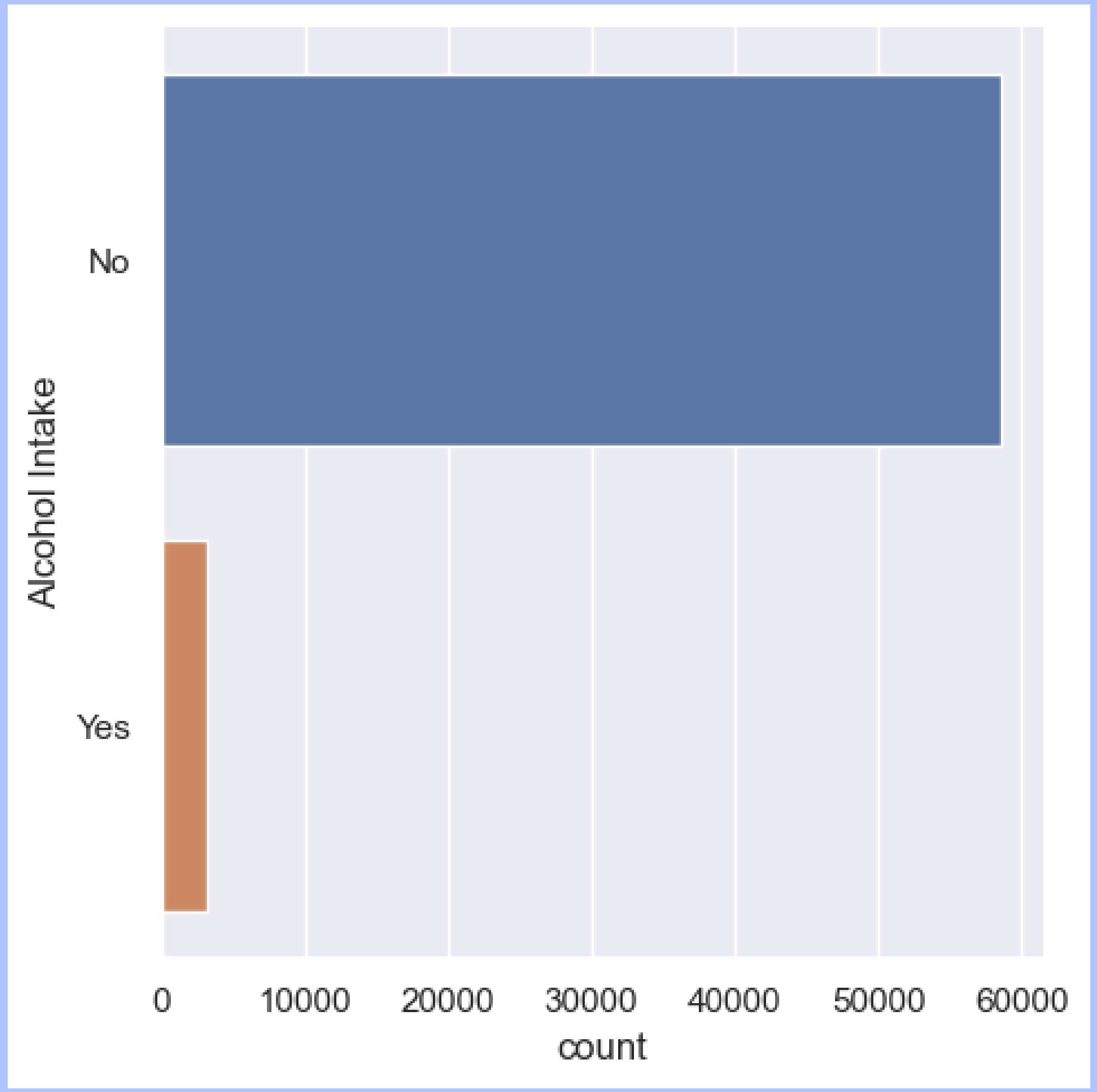


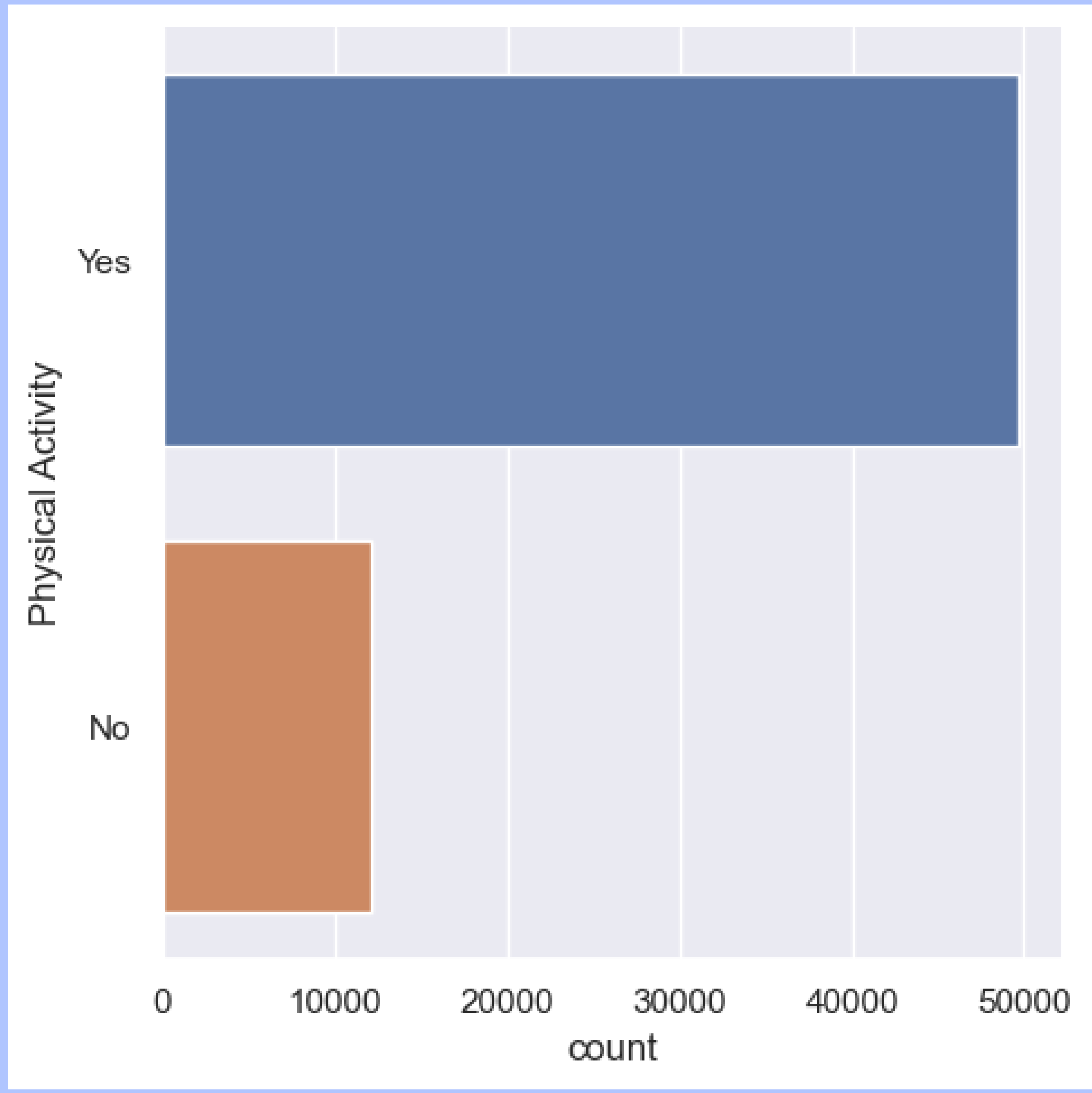


EXPLORATORY ANALYSIS AND OBSERVATIONS

we then visualize the distributions for the variables using catplot

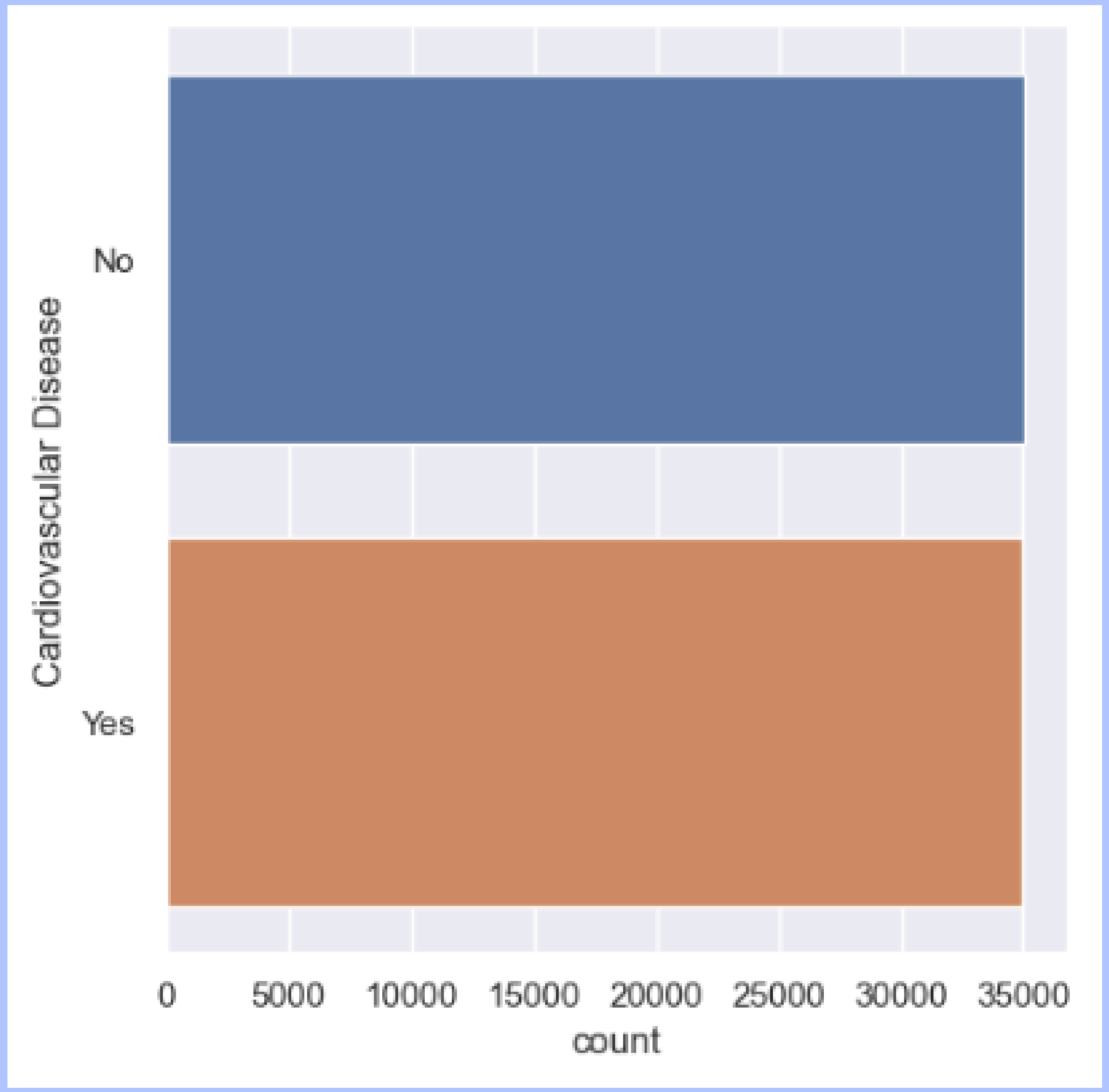




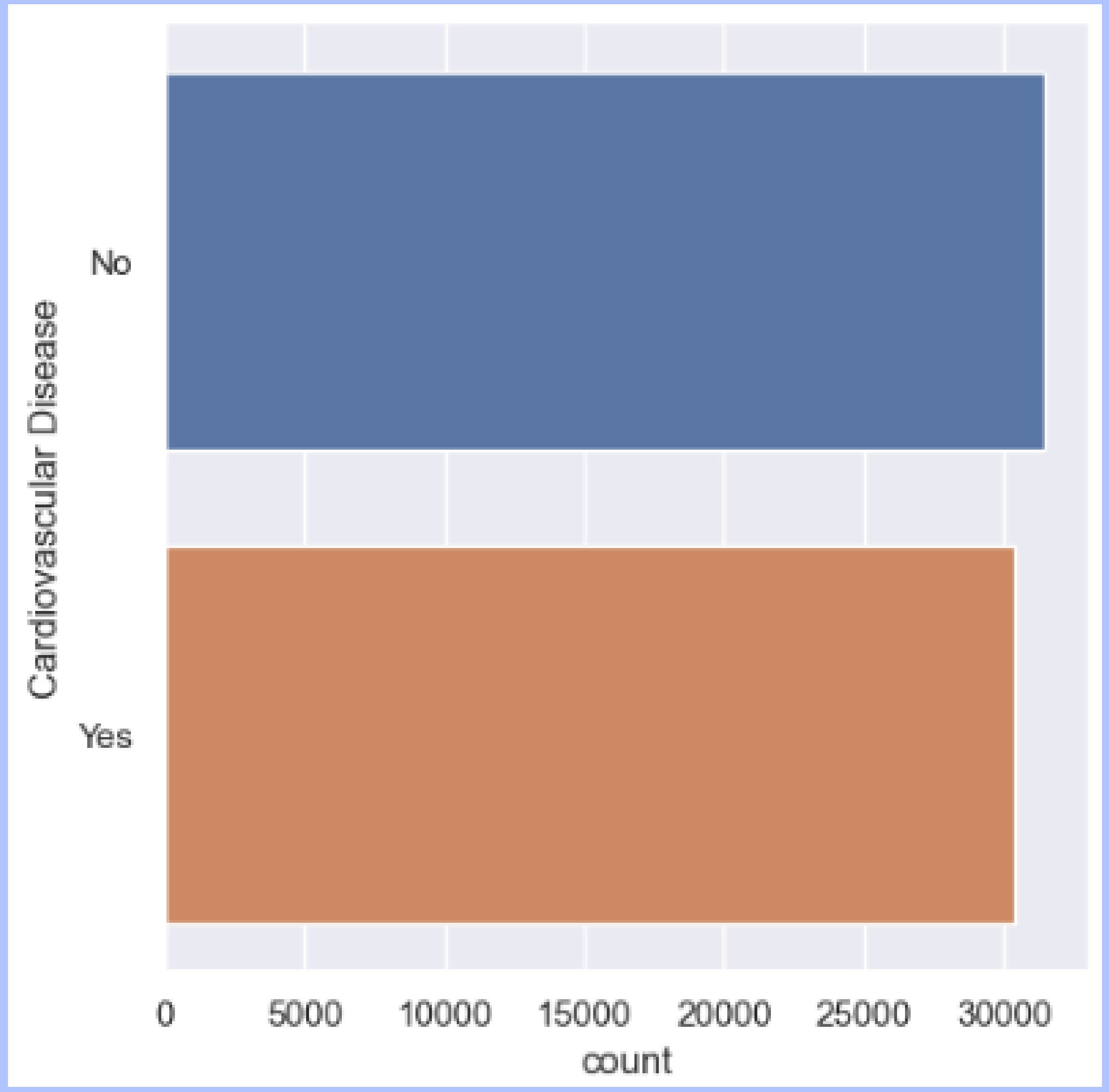


EXPLORATORY ANALYSIS AND OBSERVATIONS

we then visualize the distributions for the variables using catplot



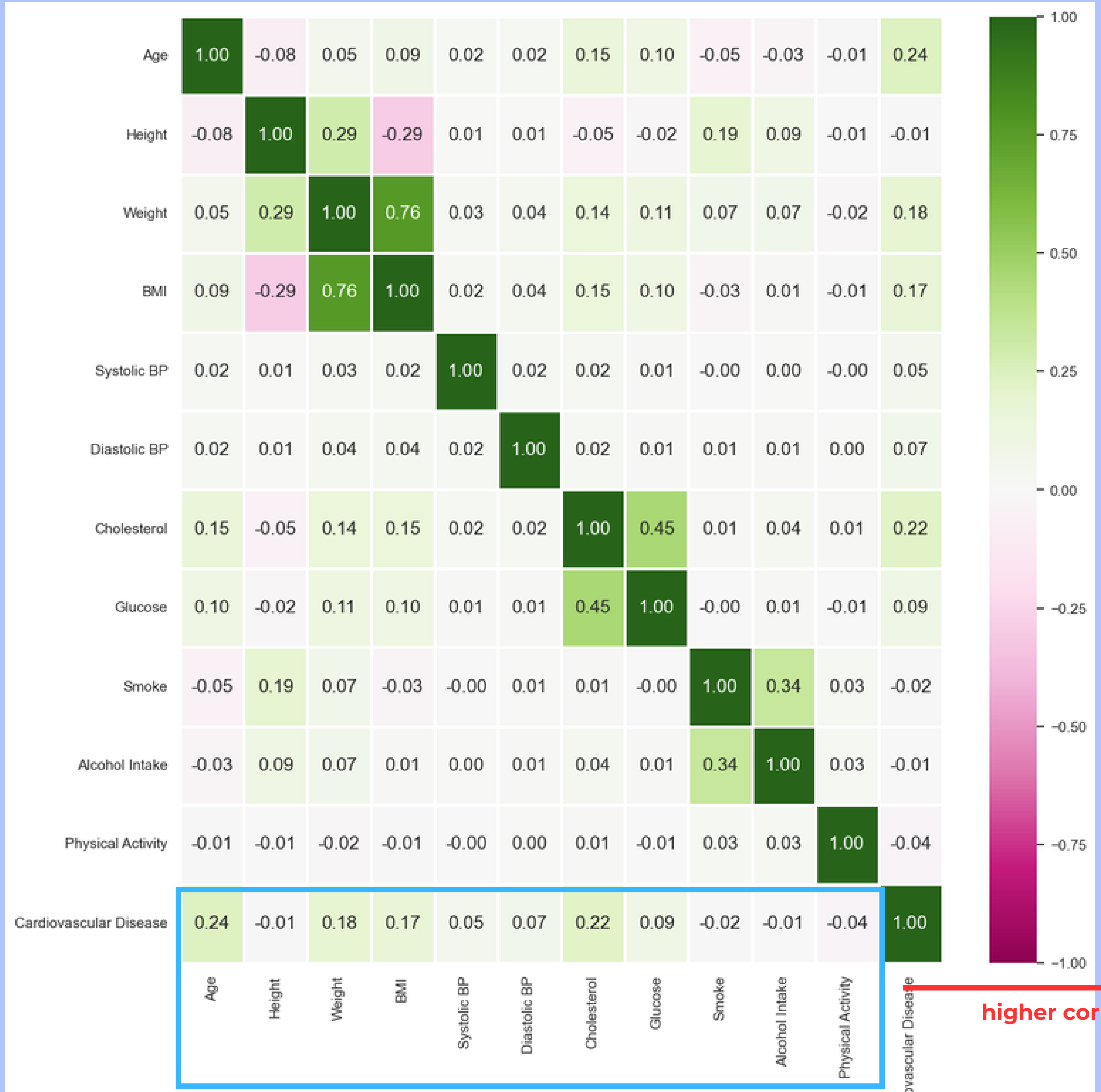
before cleaning



after cleaning

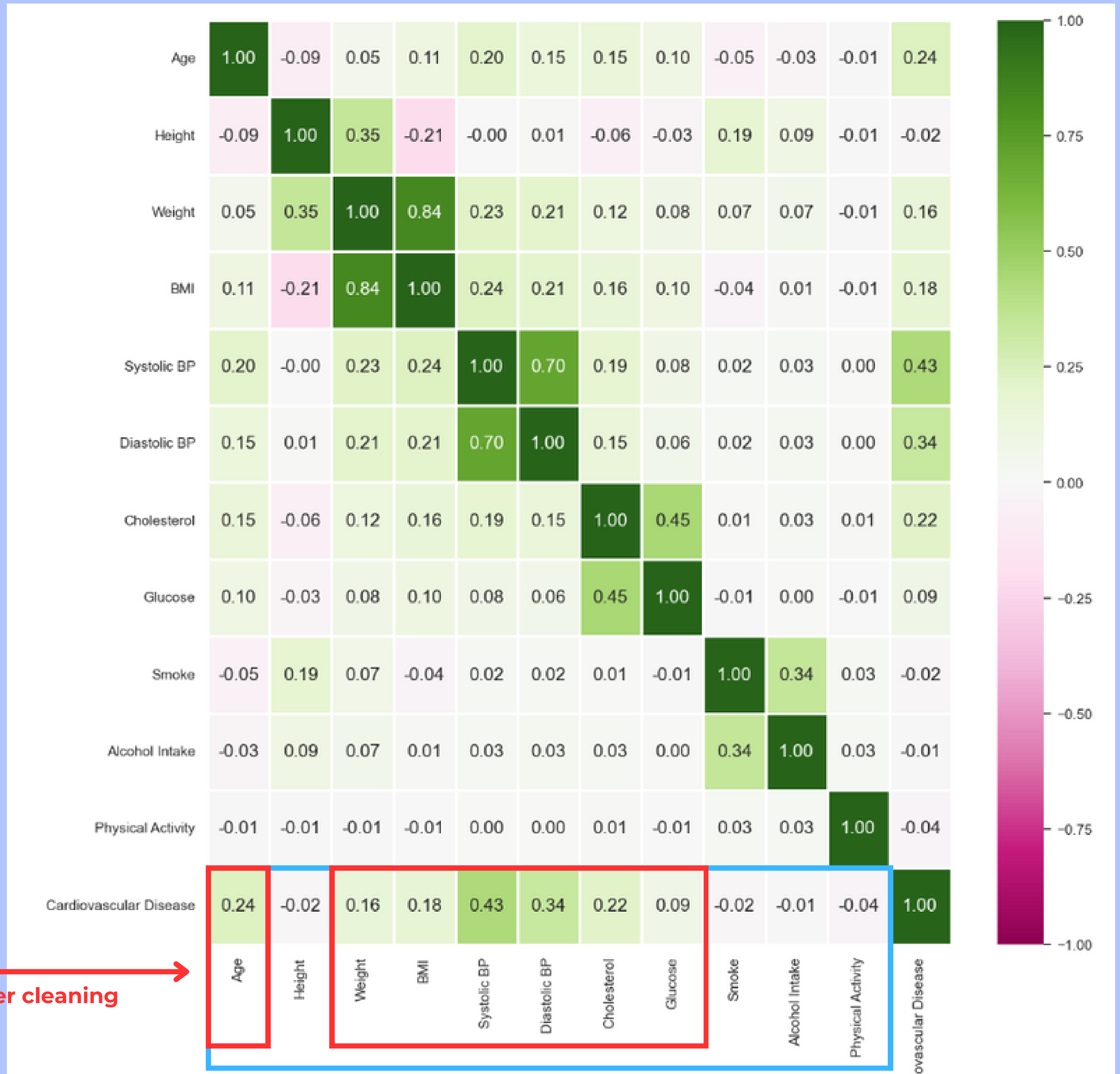
EXPLORATORY ANALYSIS AND OBSERVATIONS

Heatmap



before cleaning

higher correlations after cleaning



after cleaning

3: MACHINE LEARNING MODELS

TYPES OF ML MODELS

1. Logistic Regression
2. K-Means Clustering
3. Decision Tree
4. Random Forest

LOGISTIC REGRESSION

Require categorical variables

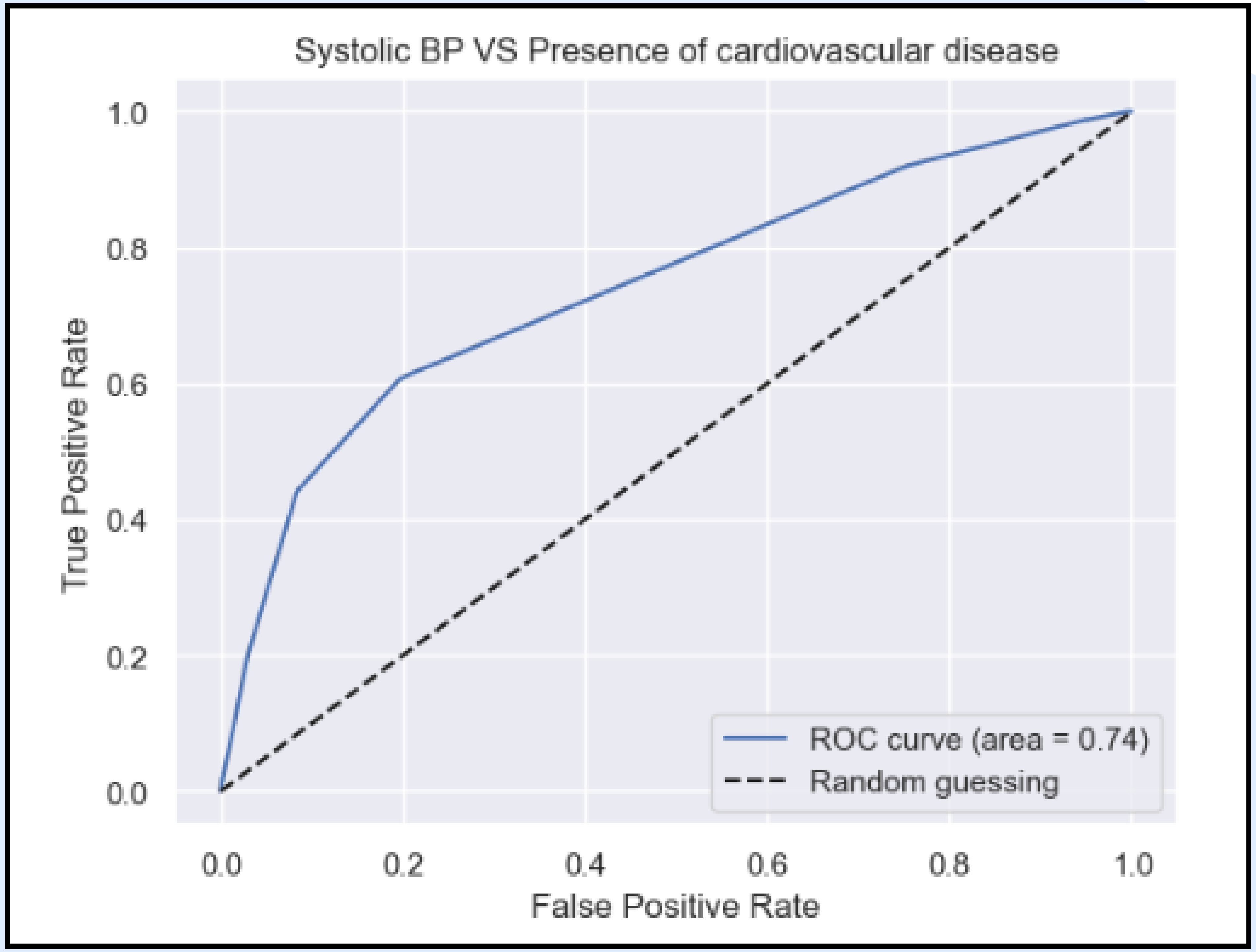
Our dependent variable, Presence of cardiovascular disease, is binary [0, 1]

	Age	Gender	Height	Weight	BMI	Systolic BP	Diastolic BP	Cholesterol	Glucose	Smoke	Alcohol Intake	Physical Activity	Cardiovascular Disease
0	50	Male	1.68	62.0	21.97	110	80	1	1	0	0	1	0
1	55	Female	1.56	85.0	34.93	140	90	3	1	0	0	1	1
2	51	Female	1.65	64.0	23.51	130	70	3	1	0	0	0	1
3	48	Male	1.69	82.0	28.71	150	100	1	1	0	0	1	1
4	60	Female	1.51	67.0	29.38	120	80	2	2	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
61779	53	Female	1.72	70.0	23.66	130	90	1	1	0	0	1	1
61780	57	Female	1.65	80.0	29.38	150	80	1	1	0	0	1	1
61781	52	Male	1.68	76.0	26.93	120	80	1	1	1	0	1	0
61782	61	Female	1.63	72.0	27.10	135	80	1	2	0	0	0	1
61783	56	Female	1.70	72.0	24.91	120	80	2	1	0	0	1	0

SYSTOLIC BP VS CARDIOVASCULAR DISEASE

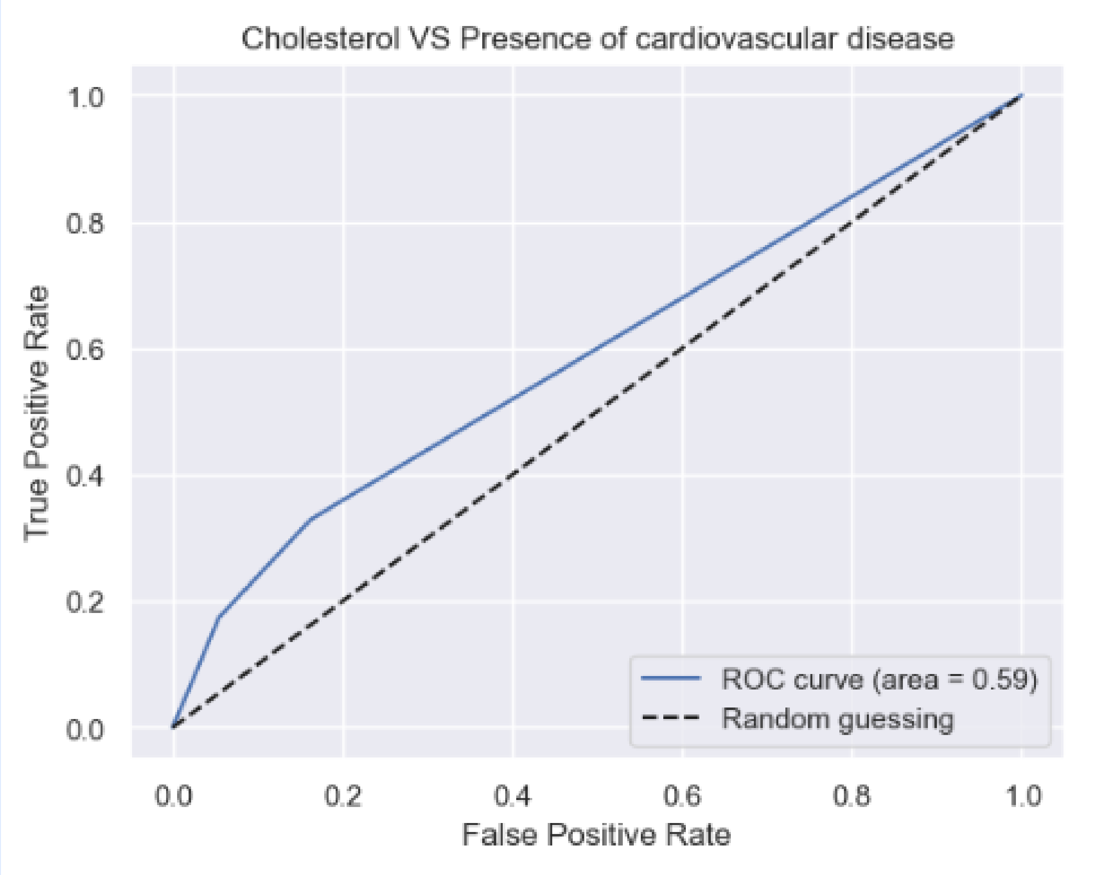
```
Intercept        : b =  [-9.70092741]
Coefficients     : a =  [[0.07696742]]

              precision    recall  f1-score   support

           0       0.68      0.80      0.74      9503
           1       0.74      0.61      0.67      9033

    accuracy                           0.71     18536
   macro avg       0.71      0.70      0.70     18536
weighted avg       0.71      0.71      0.70     18536

AUC-ROC: 0.7420967146326647
Accuracy: 0.706894691411307
```

- Precision: Prediction of Classes
- Recall: Correctly Identify of Classes
- F1-Score: Weighted Average of Precision and Recall
- Support: Number of Instances
- Accuracy: Correctly Classified of Classes



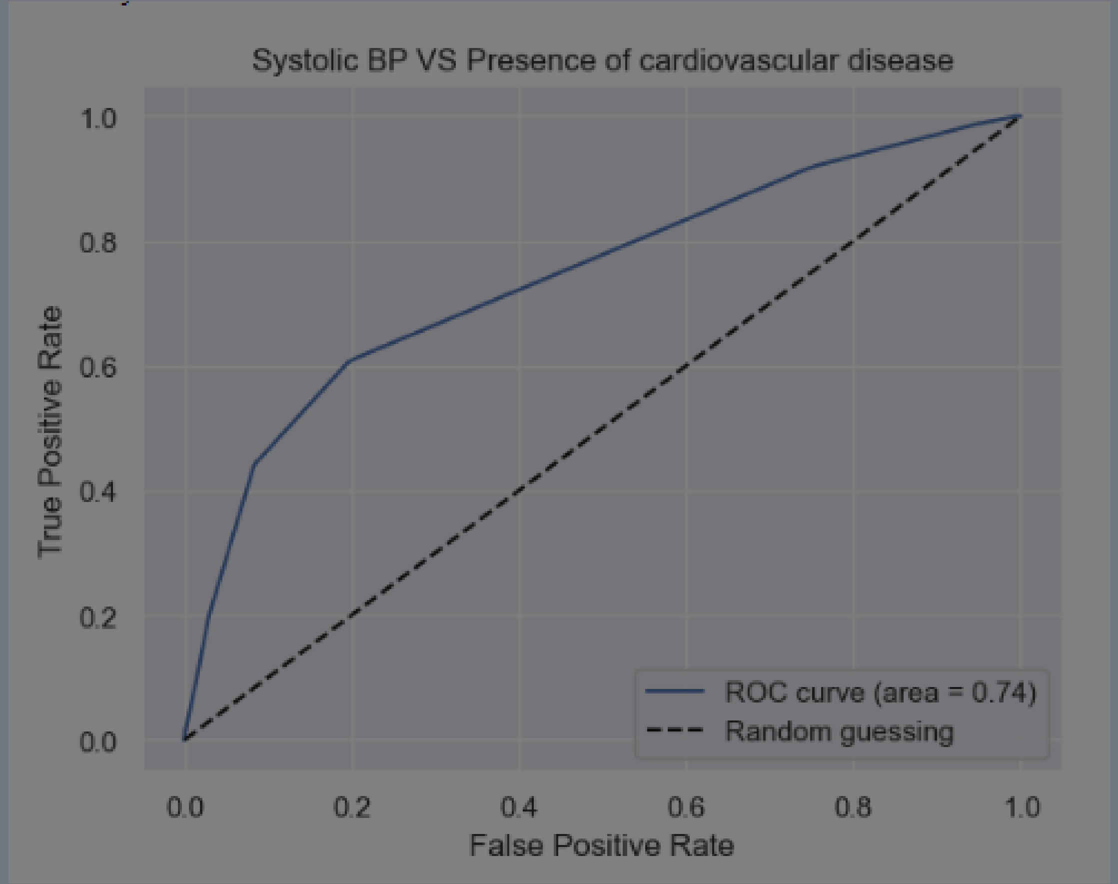
Cholesterol vs Cardiovascular Disease

```
Intercept        : b =  [-0.9628198]
Coefficients     : a =  [[0.70070073]]

              precision    recall  f1-score   support

           0       0.57      0.84      0.68      9503
           1       0.66      0.33      0.44      9033

    accuracy                           0.59     18536
   macro avg       0.61      0.58      0.56     18536
weighted avg       0.61      0.59      0.56     18536

AUC-ROC: 0.5885370452738803
Accuracy: 0.5897712559343979
```



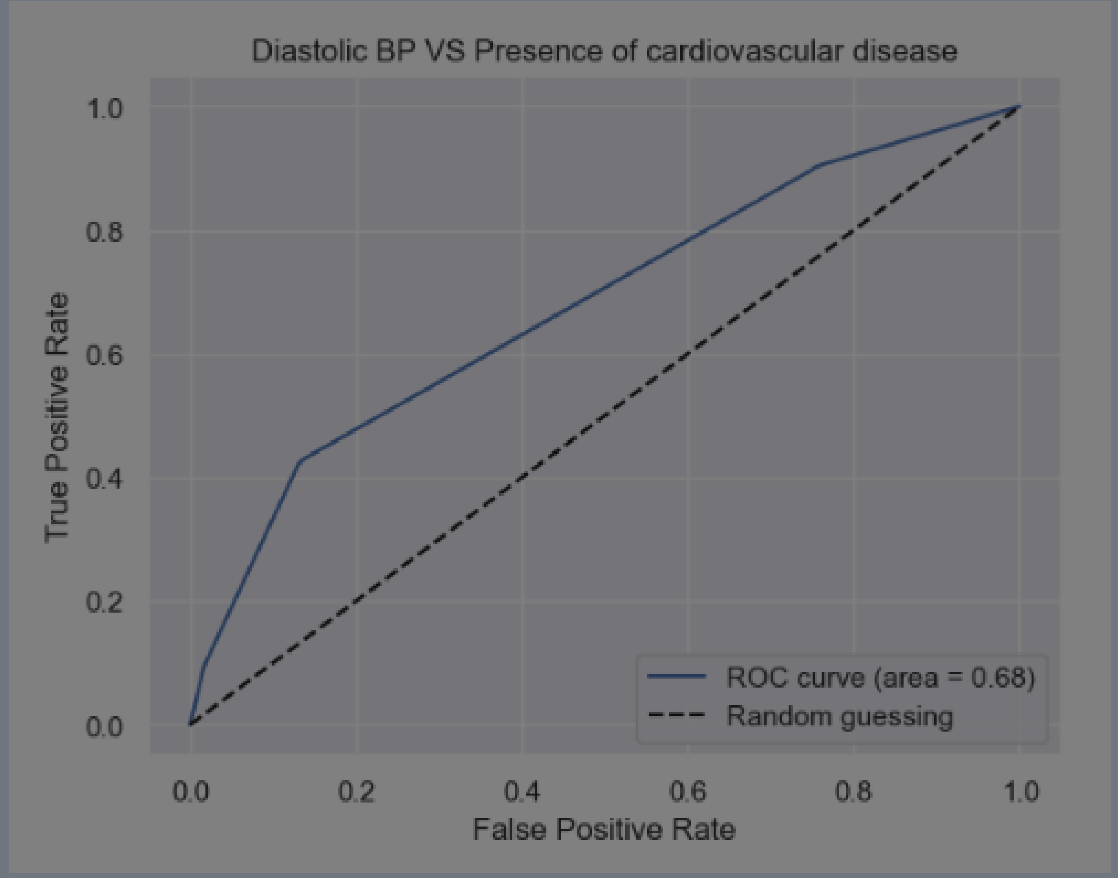
Systolic BP vs Cardiovascular Disease

```
Intercept        : b =  [-9.70092741]
Coefficients     : a =  [[0.07696742]]

              precision    recall  f1-score   support

           0       0.68      0.80      0.74      9503
           1       0.74      0.61      0.67      9033

    accuracy                           0.71     18536
   macro avg       0.71      0.70      0.70     18536
weighted avg       0.71      0.71      0.70     18536

AUC-ROC: 0.7420967146326647
Accuracy: 0.7068946914113077
```



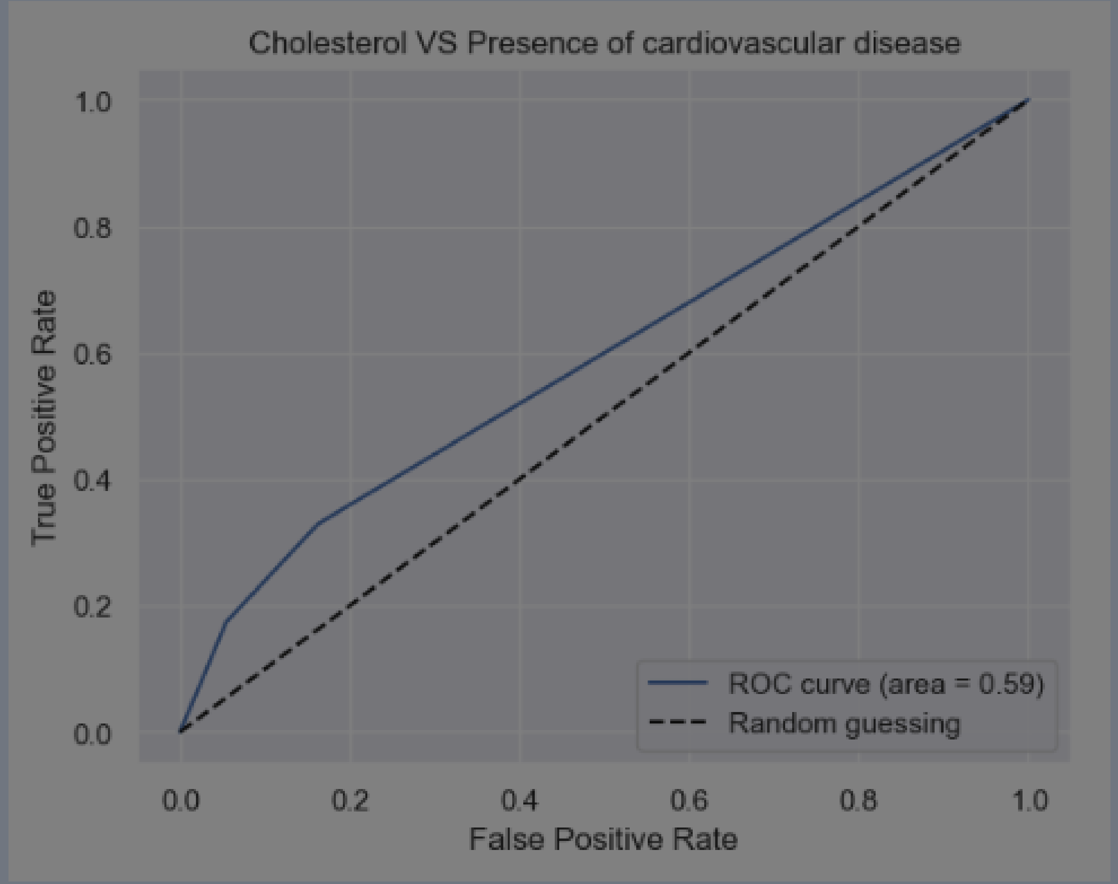
Diastolic BP vs Cardiovascular Disease

```
Intercept        : b =  [-8.15131333]
Coefficients     : a =  [[0.09960165]]

              precision    recall  f1-score   support

           0       0.61      0.86      0.72      9503
           1       0.75      0.43      0.55      9033

    accuracy                           0.65     18536
   macro avg       0.68      0.65      0.63     18536
weighted avg       0.68      0.65      0.63     18536

AUC-ROC: 0.6773027935184843
Accuracy: 0.6515968925334484
```



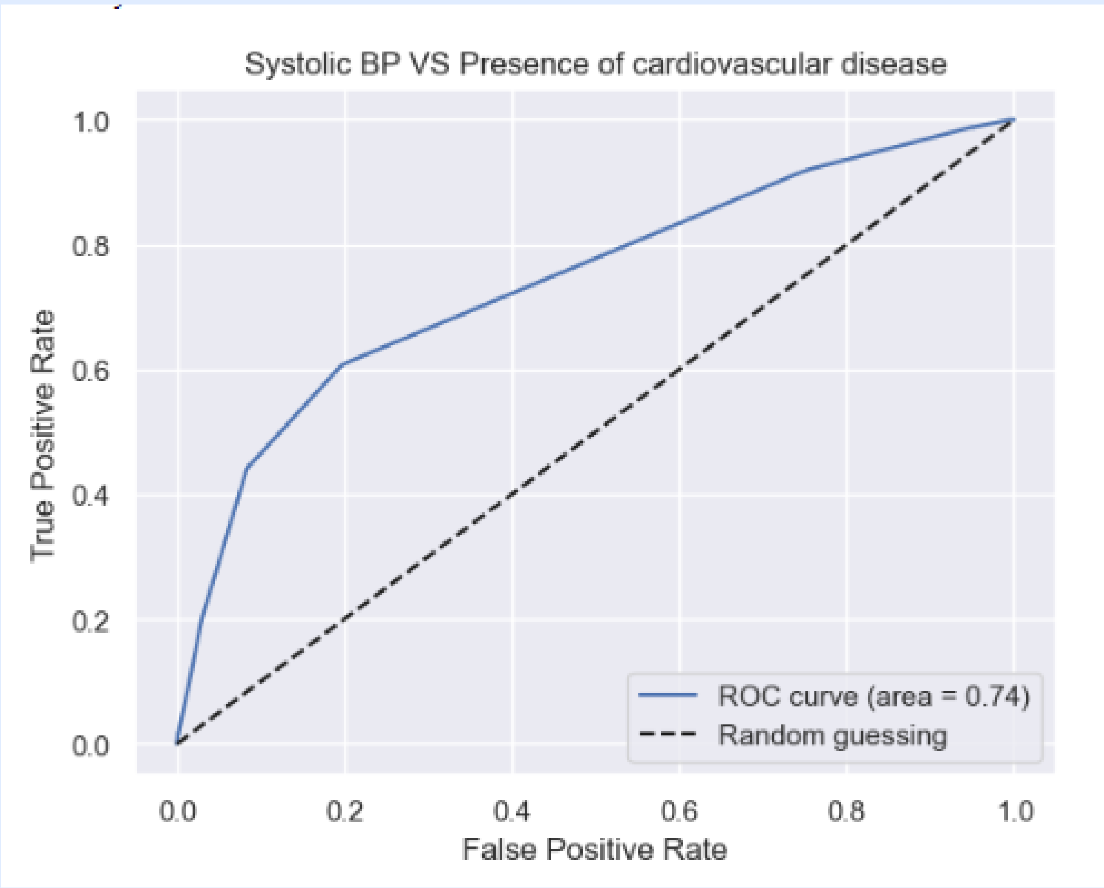
Cholesterol vs Cardiovascular Disease

```
Intercept       : b =  [-0.9628198]
Coefficients    : a =  [[0.70070073]]

              precision    recall  f1-score   support

           0       0.57      0.84      0.68      9503
           1       0.66      0.33      0.44      9033

    accuracy                           0.59     18536
   macro avg       0.61      0.58      0.56     18536
weighted avg       0.61      0.59      0.56     18536

AUC-ROC: 0.5885370452738803
Accuracy: 0.5897712559343979
```



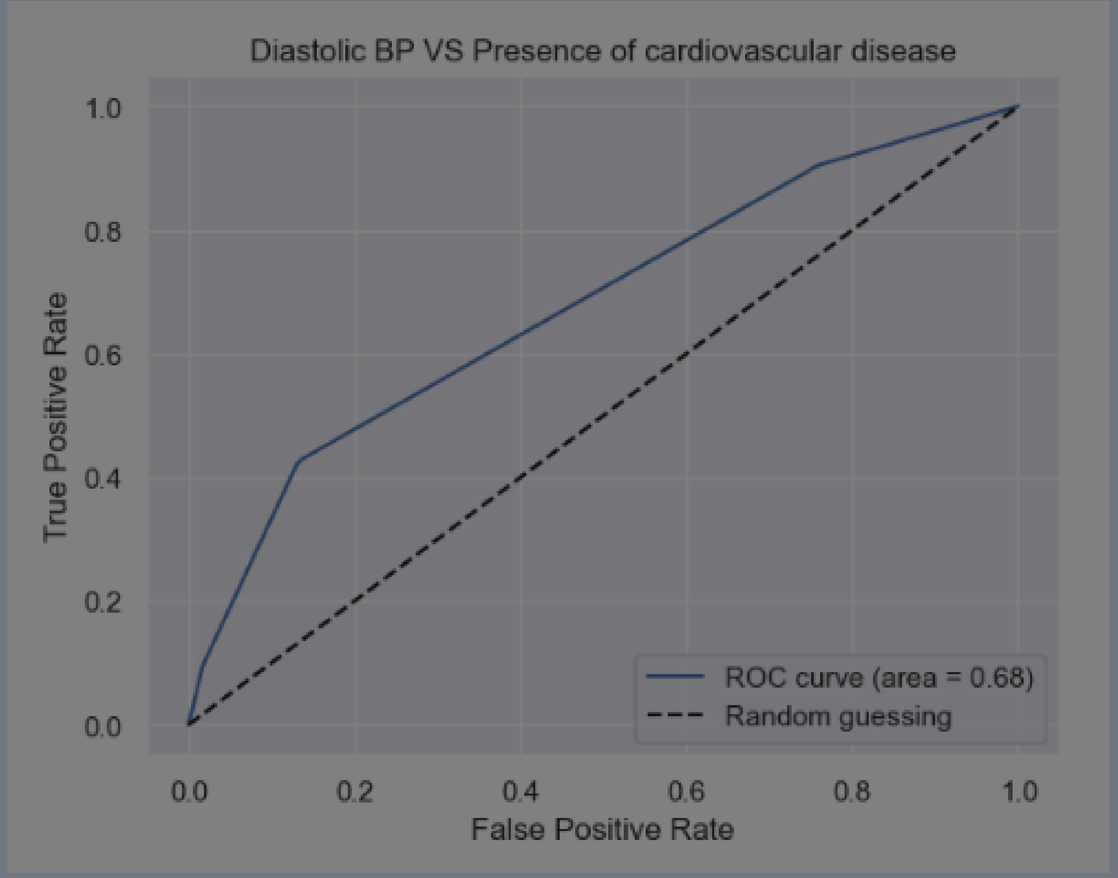
Systolic BP vs Cardiovascular Disease

```
Intercept       : b =  [-9.70092741]
Coefficients    : a =  [[0.07696742]]

              precision    recall  f1-score   support

           0       0.68      0.80      0.74      9503
           1       0.74      0.61      0.67      9033

    accuracy                           0.71     18536
   macro avg       0.71      0.70      0.70     18536
weighted avg       0.71      0.71      0.70     18536

AUC-ROC: 0.7420967146326647
Accuracy: 0.7068946914113077
```



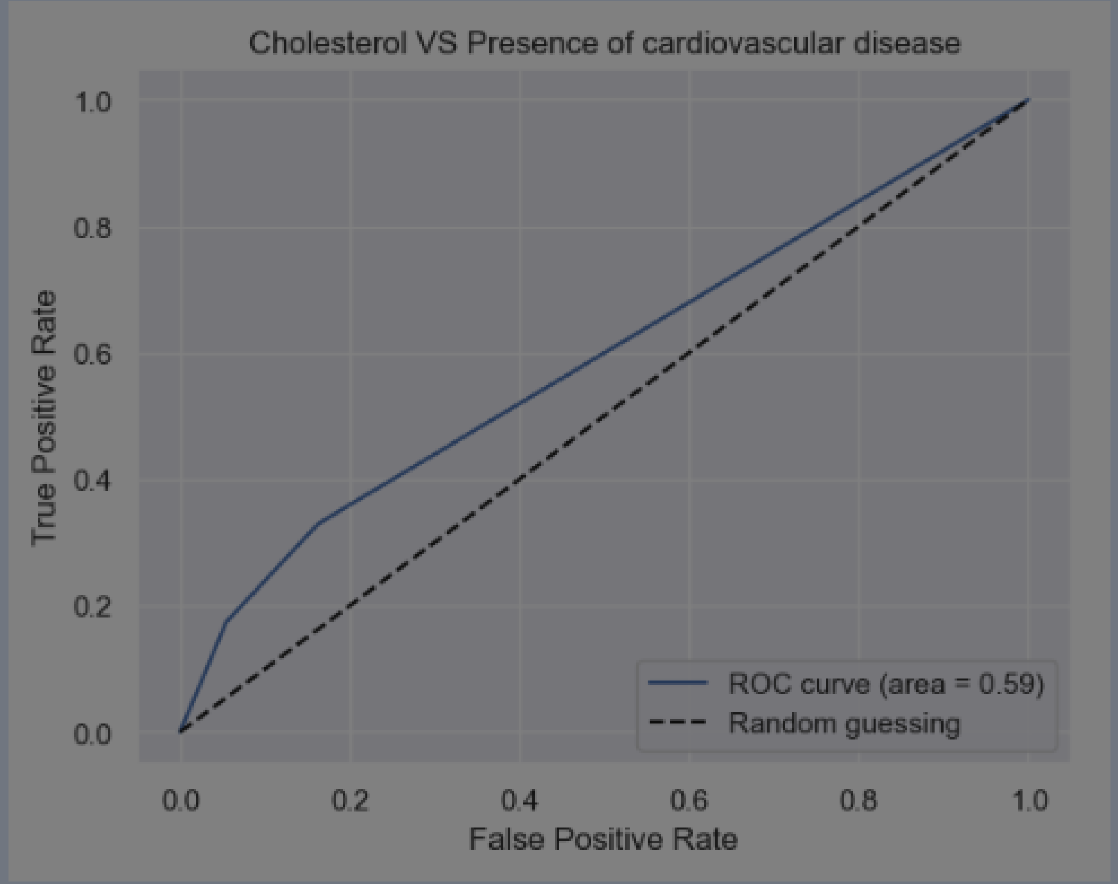
Diastolic BP vs Cardiovascular Disease

```
Intercept       : b =  [-8.15131333]
Coefficients    : a =  [[0.09960165]]

              precision    recall  f1-score   support

           0       0.61      0.86      0.72      9503
           1       0.75      0.43      0.55      9033

    accuracy                           0.65     18536
   macro avg       0.68      0.65      0.63     18536
weighted avg       0.68      0.65      0.63     18536

AUC-ROC: 0.6773027935184843
Accuracy: 0.6515968925334484
```



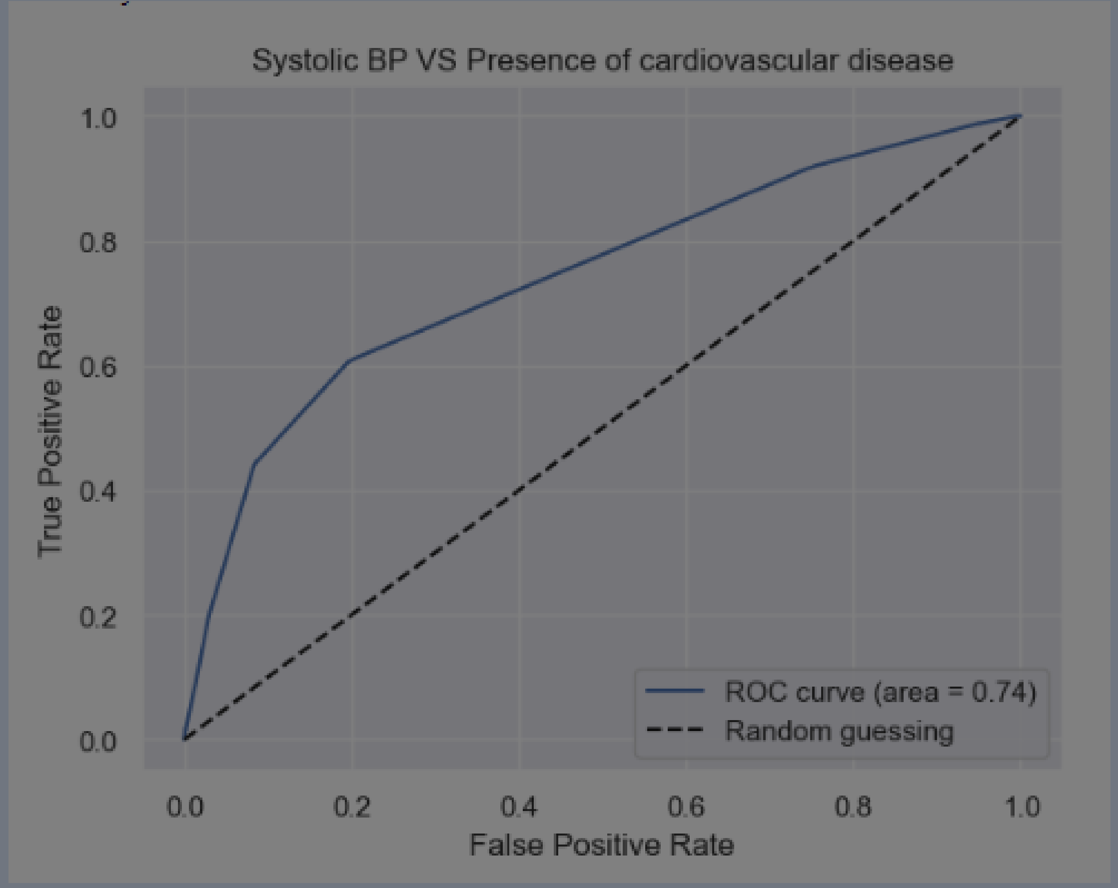
Cholesterol vs Cardiovascular Disease

```
Intercept        : b =  [-0.9628198]
Coefficients     : a =  [[0.70070073]]

              precision    recall  f1-score   support

           0       0.57      0.84      0.68      9503
           1       0.66      0.33      0.44      9033

    accuracy                           0.59     18536
   macro avg       0.61      0.58      0.56     18536
weighted avg       0.61      0.59      0.56     18536

AUC-ROC: 0.5885370452738803
Accuracy: 0.5897712559343979
```



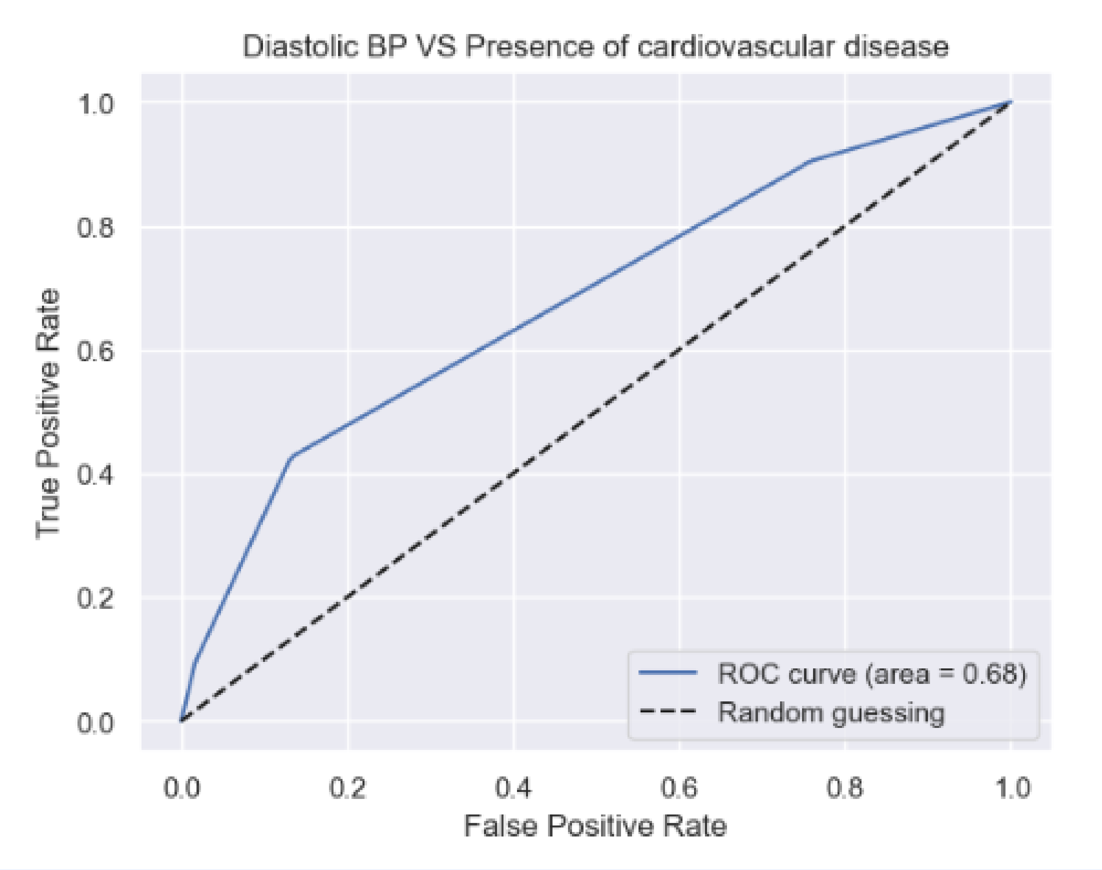
Systolic BP vs Cardiovascular Disease

```
Intercept        : b =  [-9.70092741]
Coefficients     : a =  [[0.07696742]]

              precision    recall  f1-score   support

           0       0.68      0.80      0.74      9503
           1       0.74      0.61      0.67      9033

    accuracy                           0.71     18536
   macro avg       0.71      0.70      0.70     18536
weighted avg       0.71      0.71      0.70     18536

AUC-ROC: 0.7420967146326647
Accuracy: 0.7068946914113077
```



Diastolic BP vs Cardiovascular Disease

```
Intercept        : b =  [-8.15131333]
Coefficients     : a =  [[0.09960165]]

              precision    recall  f1-score   support

           0       0.61      0.86      0.72      9503
           1       0.75      0.43      0.55      9033

    accuracy                           0.65     18536
   macro avg       0.68      0.65      0.63     18536
weighted avg       0.68      0.65      0.63     18536

AUC-ROC: 0.6773027935184843
Accuracy: 0.6515968925334484
```



WHICH MODEL IS BEST BASED ON ITS METRICS?

Cholesterol vs Presence of Cardiovascular Disease (A)	Systolic BP vs Presence of Cardiovascular Disease (B)	Diastolic BP vs Presence of Cardiovascular Disease (C)	Best
Intercept : b = [-0.9628198] Coefficients : a = [[0.70070073]] precision recall f1-score support 0 0.57 0.84 0.68 9503 1 0.66 0.33 0.44 9033 accuracy 0.59 18536 macro avg 0.61 0.58 0.56 18536 weighted avg 0.61 0.59 0.56 18536 AUC-ROC: 0.5885370452738803 Accuracy: 0.5897712559343979	Intercept : b = [-9.70092741] Coefficients : a = [[0.07696742]] precision recall f1-score support 0 0.68 0.80 0.74 9503 1 0.74 0.61 0.67 9033 accuracy 0.71 18536 macro avg 0.71 0.70 0.70 18536 weighted avg 0.71 0.71 0.70 18536 AUC-ROC: 0.7420967146326647 Accuracy: 0.7068946914113077	Intercept : b = [-8.15131333] Coefficients : a = [[0.09960165]] precision recall f1-score support 0 0.61 0.86 0.72 9503 1 0.75 0.43 0.55 9033 accuracy 0.65 18536 macro avg 0.68 0.65 0.63 18536 weighted avg 0.68 0.65 0.63 18536 AUC-ROC: 0.6773027935184843 Accuracy: 0.6515968925334484	B

<u>Higher</u> Precision, recall, F1-score, AUC

COMPARISON TO PRE-CLEANING

Original Dataset	Cleaned Dataset

Original Dataset

```
Intercept        : b =  [-5.6937604]
Coefficients     : a =  [[0.0450365]]

              precision    recall  f1-score   support

           0       0.68      0.80      0.74     10461
           1       0.76      0.63      0.69     10539

    accuracy                           0.72     21000
   macro avg       0.72      0.72      0.72     21000
weighted avg       0.72      0.72      0.72     21000

AUC-ROC: 0.7551067439216099
Accuracy: 0.7177619047619047
```

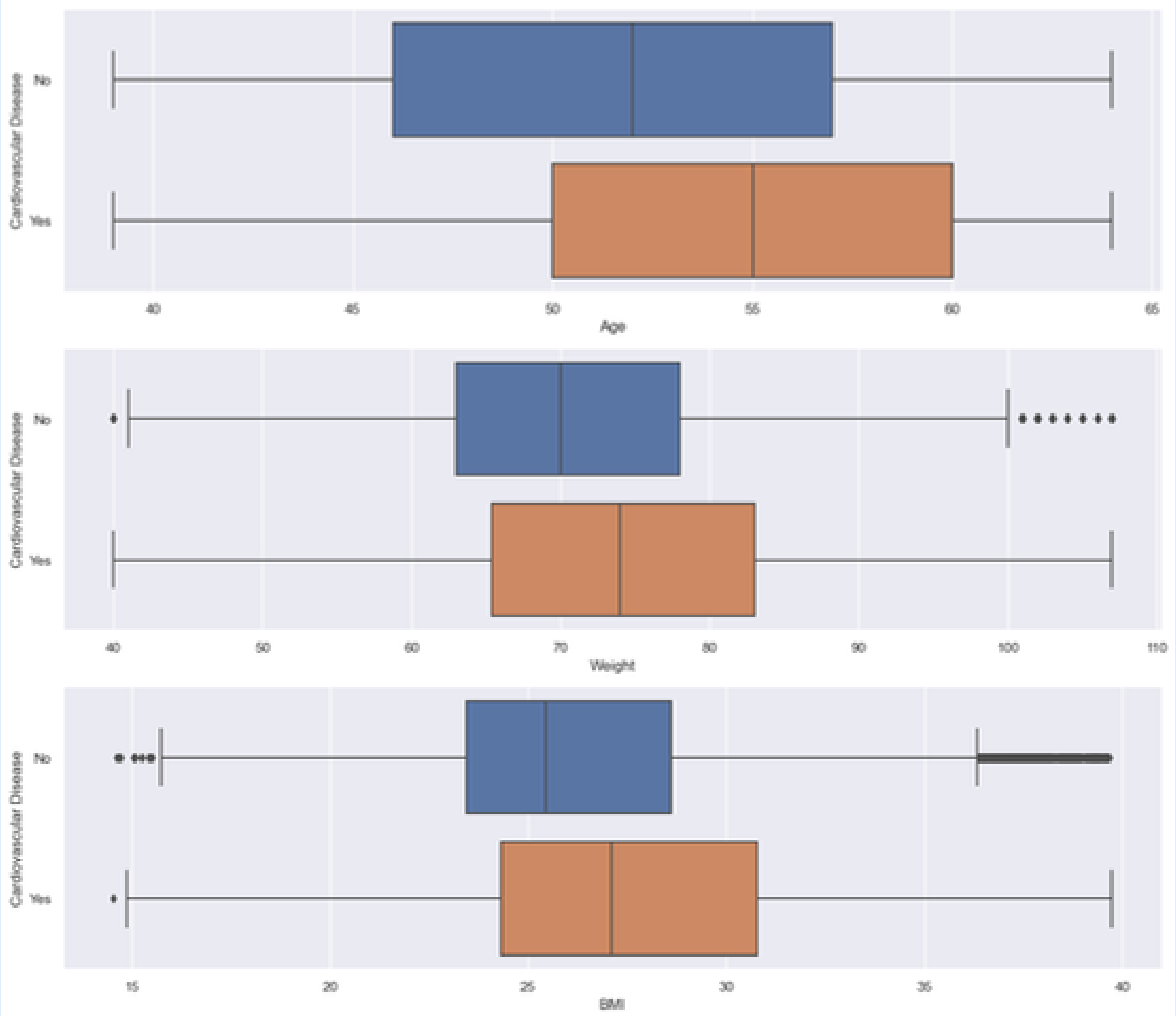
Cleaned Dataset

```
Intercept        : b =  [-9.70092741]
Coefficients     : a =  [[0.07696742]]

              precision    recall  f1-score   support

           0       0.68      0.80      0.74      9503
           1       0.74      0.61      0.67      9033

    accuracy                           0.71     18536
   macro avg       0.71      0.70      0.70     18536
weighted avg       0.71      0.71      0.70     18536

AUC-ROC: 0.7420967146326647
Accuracy: 0.706894691411307
```

Very similar AUC and Accuracy

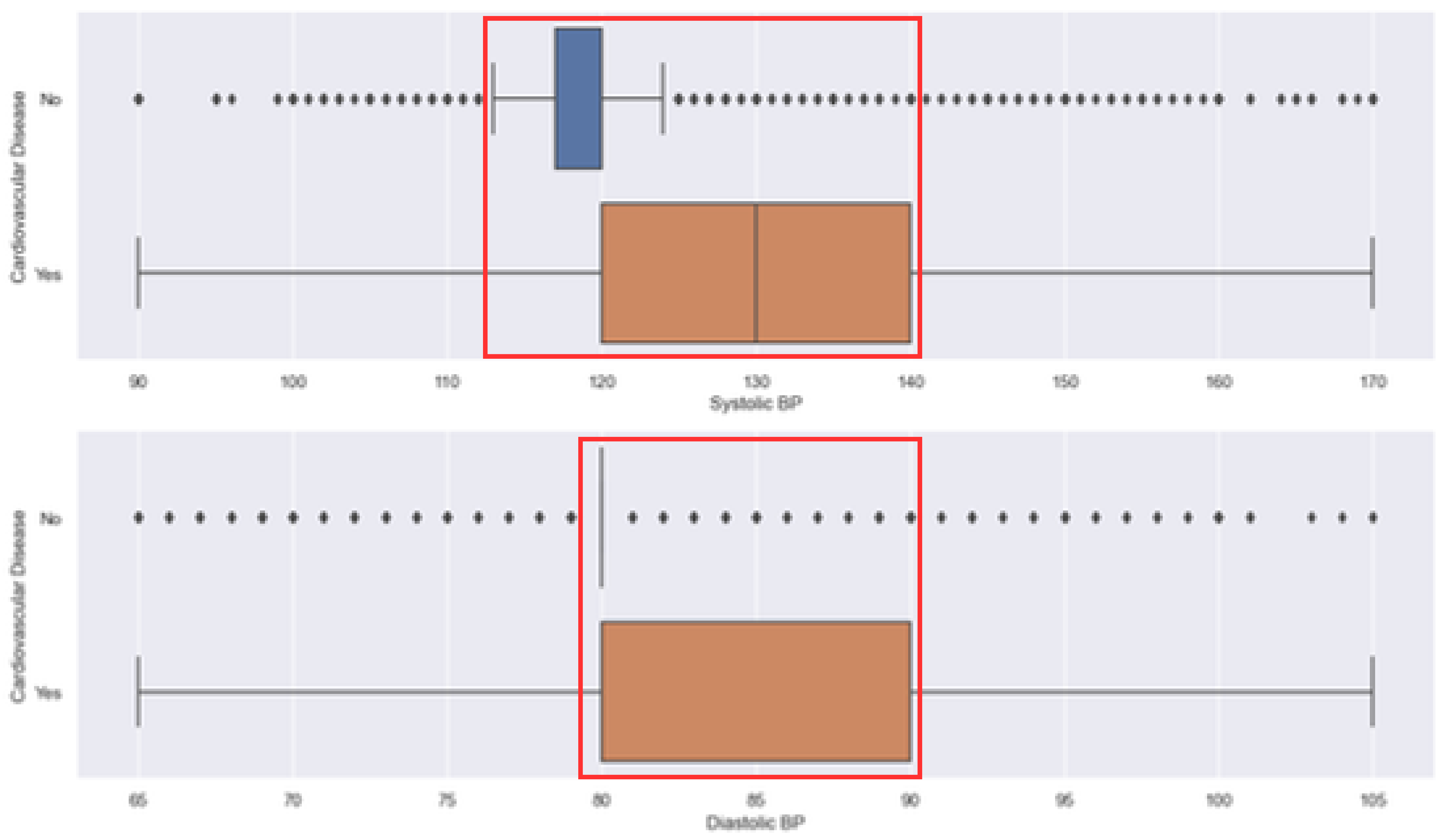
K-MEANS CLUSTERING

To group similar data points together and discover underlying patterns

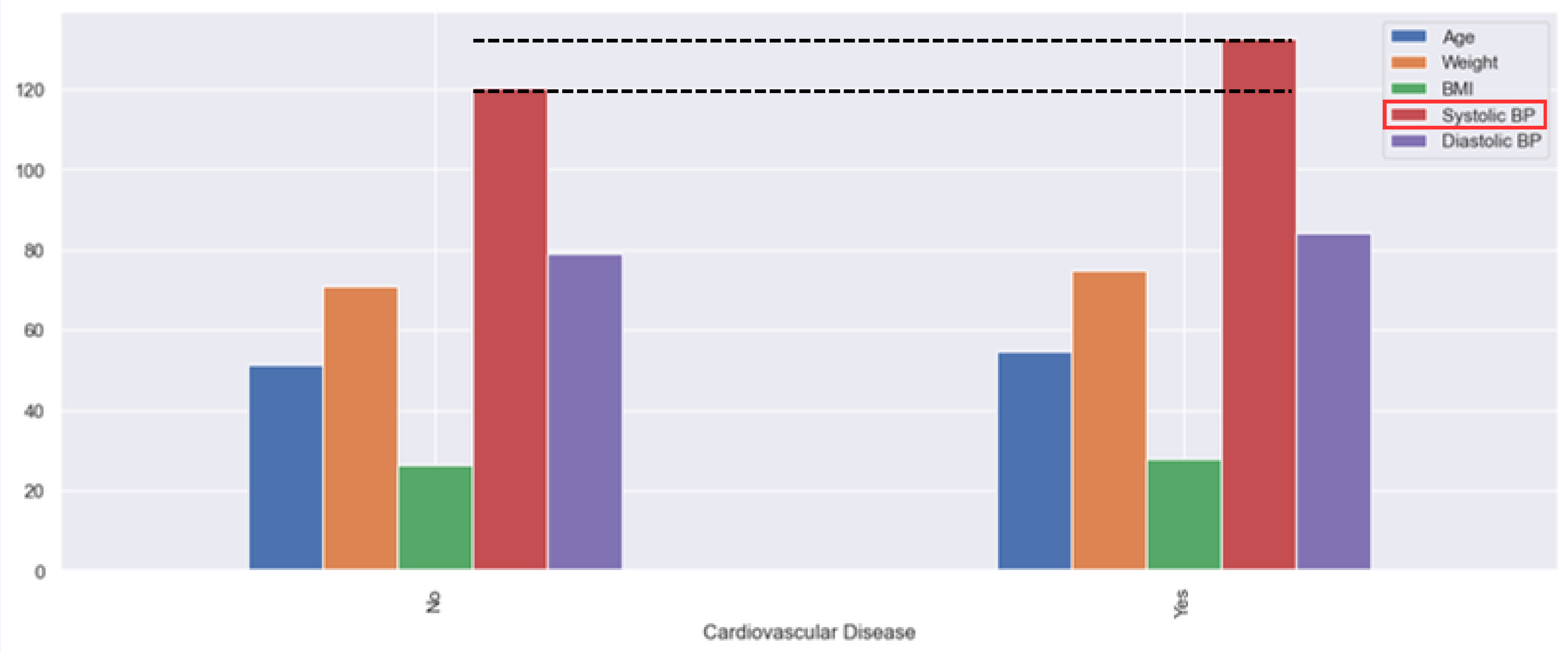


K-MEANS CLUSTERING

BP variables for people with Cardiovascular Disease are **generally higher** than those without Cardiovascular Disease



K-MEANS CLUSTERING



biggest difference

DECISION TREE

Description: Tree-like models are useful for classification tasks and uses categorical data.

Purpose of Decision Tree in the context of our Project:
Classifies the factors used, measures the effectiveness for predicting the likelihood of the country being happy or unhappy.

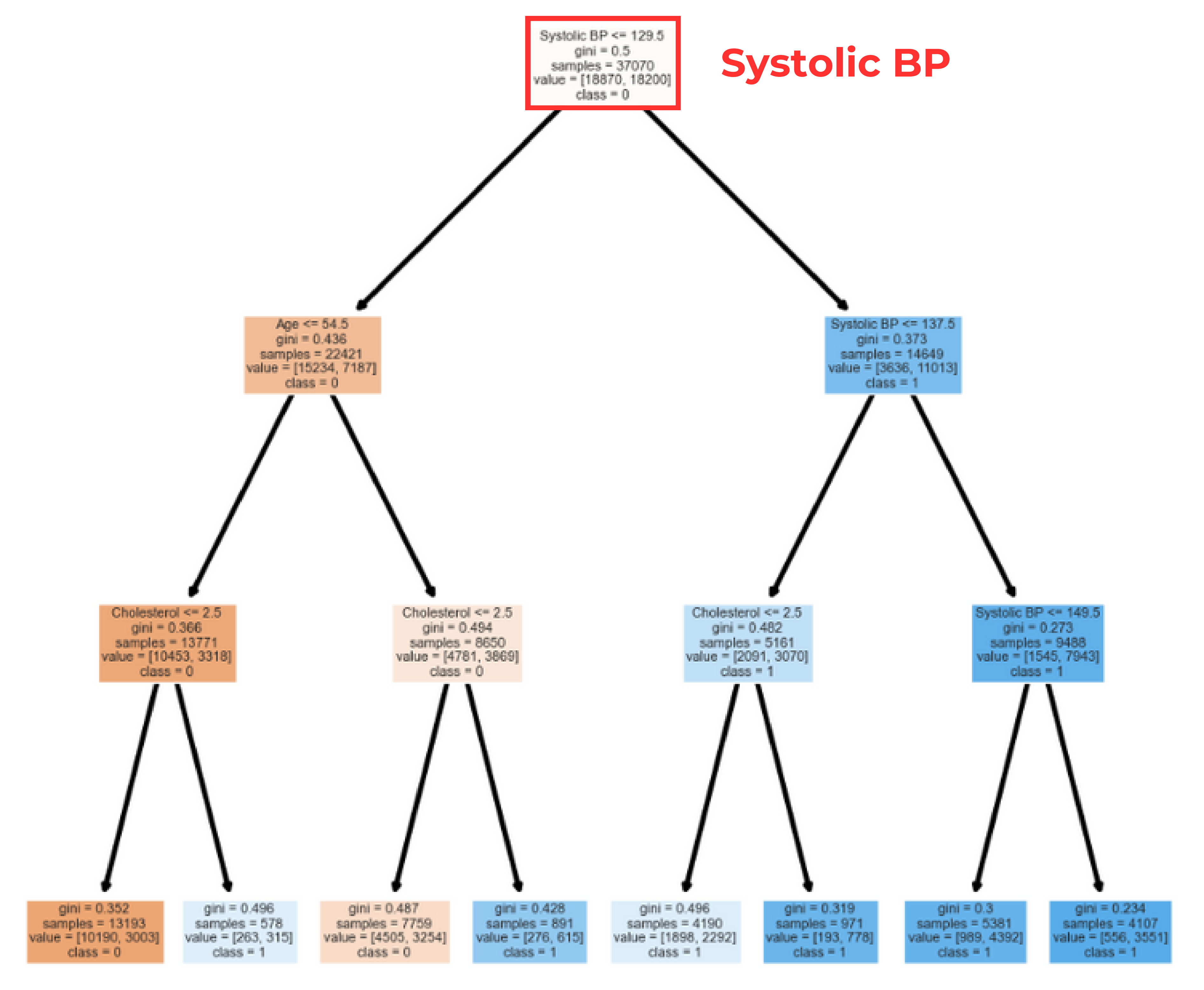
- Response variable: Presence of Cardiovascular Disease

- Predictor factors:

1. Age
2. Weight
3. BMI
4. Systolic BP
5. Diastolic BP
6. Cholesterol

DECISION TREE

```
Goodness of Fit of Model        Train Dataset
Classification Accuracy         : 0.7214459131373078

Goodness of Fit of Model        Test Dataset
Classification Accuracy         : 0.7191470421623372
```



Observations:
Gini Index of the decision tree is relatively low (0.0 - 0.5), denoting high purity or low impurity.

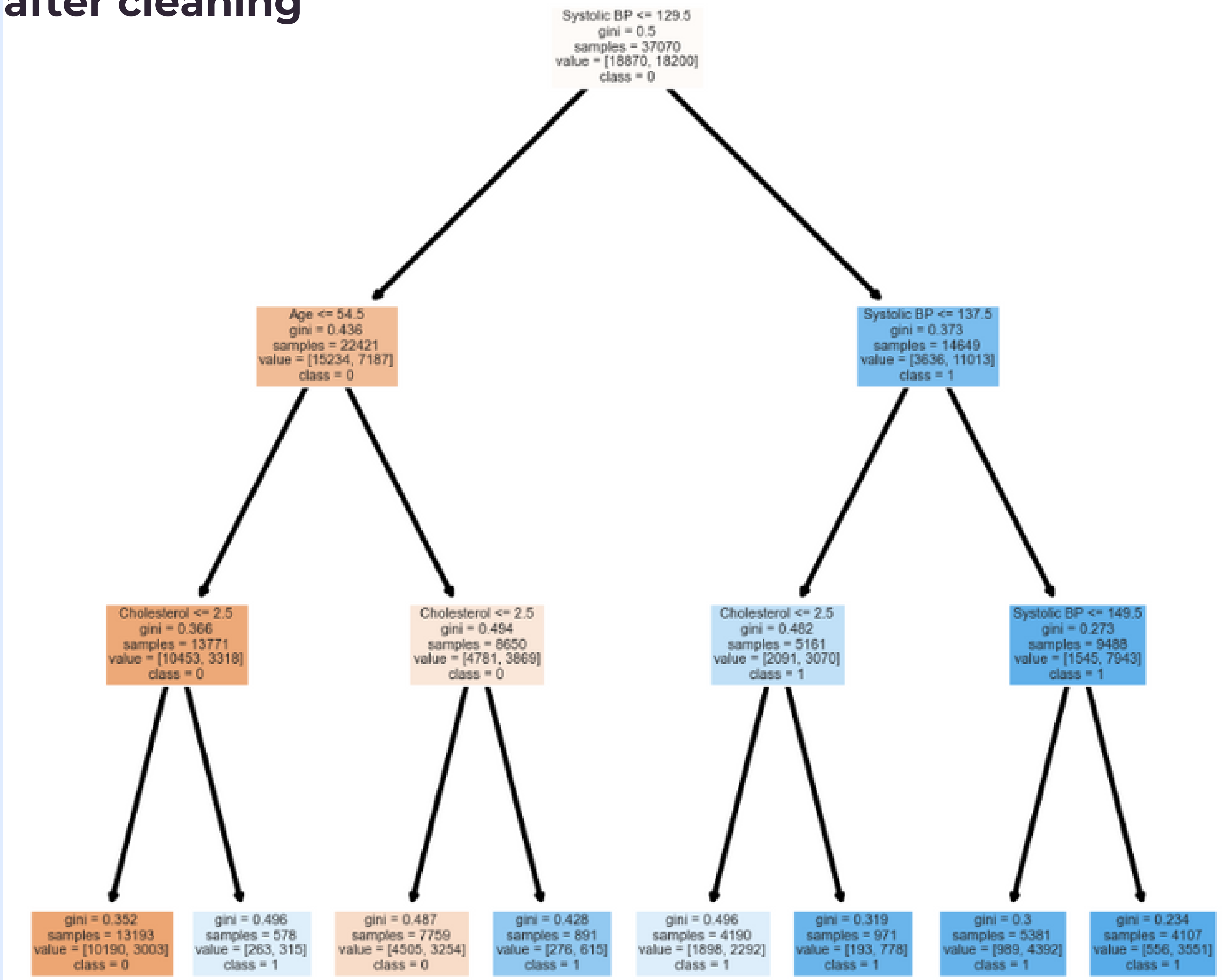
In a multivariate decision tree, **overfitting** may occur.

Preventative Measures:
Limiting the number of variables and setting max_depth to 3.

DECISION TREE

before cleaning

after cleaning



RANDOM FOREST

Description: Multiple decision trees are used to give a prediction based on the factors in relation to the presence of Cardiovascular Disease

Purpose of Decision Tree in the context of our Project:
Classifies the factors used, measures the effectiveness for predicting the likelihood of cardiovascular disease.

How Random Forest works:

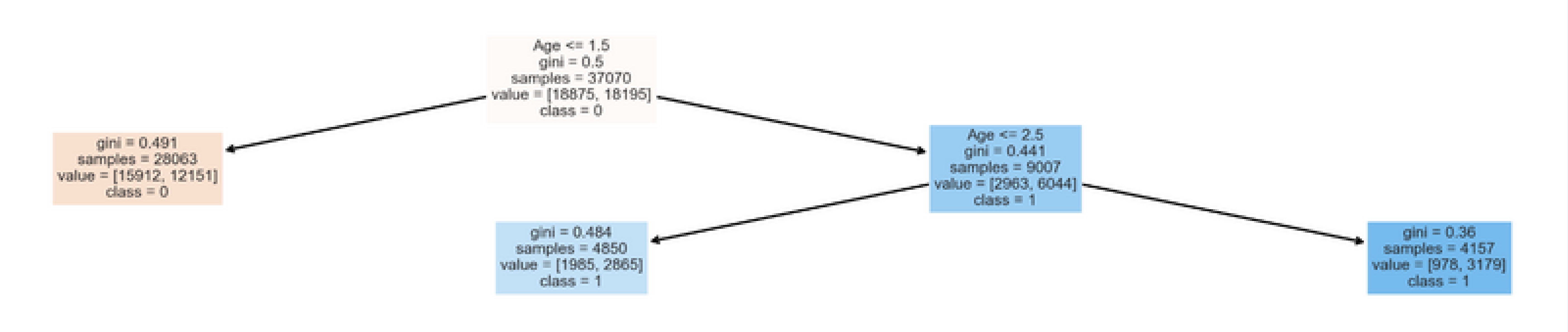
1. Select random samples from a given dataset and split into Train and Test sets
2. Construct a decision tree for each sample and get a prediction result from each decision tree.
3. Perform a vote for each predicted result.
4. Select the prediction result with the most votes as the final prediction.

RANDOM FOREST

Observations:
Gini Index is also relatively low (0.0-0.5), denoting the high purity and low impurity

To improve the accuracy of our random forest, we used a validation test set.

```
Accuracy on validation set: 0.6883952415634863
```



4: FINDINGS AND OUTCOMES

	Logistic Regression	**Decision Tree**	**Random Forest**
Advantages	Strong at **Data Analysis**, especially in **binary outcomes.** **Great** at measuring **relationships** between **predictors** and **target variables**.	Great at capturing **non-linear relationship** between predictors. Great at **Predicting**.	Strong at providing **accurate prediction** than other models. Can **capture non-linear relationships** between predictors. **Less prone to overfitting.**
Disadvantages **(Limitations)**	**May not perform well** if the relationships are **non-linear**. **Not suitable for smaller dataset.**	May be prone to risk of **overfitting** due to complex tree, especially a **smaller dataset**.	More **difficult to interpret** than a single Decision Tree.

Outcome

Through this project, we analyzed the data set, trained a classification model with classification, clustering and anomaly predication and evaluated the data.

We have built a relatively effective model of >70% accuracy with an AUC >0.74 to predict the likelihood of cardiovascular disease.

Outcome

Each of our Machine Learning models generally support that

<u>Systolic Blood Pressure</u>

is the most important variable in predicting the presence of cardiovascular disease.

CONCLUSION

SYSTOLIC BP AND DIASTOLIC BP
are the most important indicators of cardiovascular disease

- Blood pressure measurement is a relatively simple and non-invasive procedure.
- Lower income countries should focus on testing for BP since BP monitoring devices are cheap and widely available and can be easily purchased and maintained by healthcare facilities.

JOB DISTRIBUTION

Name:	Phua Wei An	Pagdanganan Robert Martin Gosioco	Tan Chuan Bing	Nguyen Hoang Minh
Initial Data Preparation	✓	✓	✓	support
Exploratory Analysis + Further Data Cleaning	✓	support	✓	✓
Logistic Regression	✓	support	support	support
K-Means Clustering	✓	support	support	support
Decision Tree	✓	support	support	support
Random Forest	✓	support	support	support
Findings + Conclusion	✓	support	support	✓