# Project 2 Milestone 3 - White Paper

## Topic

The risks associated with high blood pressure and high cholesterol levels have been a prevalent topic of discussion in America for many years. This issue is personally relevant to me, as my parents and many older relatives are on medication for high blood pressure, and my father must carefully monitor his diet to prevent a rise in his cholesterol levels. I aim to analyze and explore potential correlations between blood pressure, cholesterol levels, and cardiovascular disease.

## Background History

Cardiovascular diseases are the leading cause of death worldwide, with 17.9 million people dying from CVDs in 2019, according to the World Health Organization (World, n.d.). The WHO also highlights other factors that increase the risk of cardiovascular disease, including an unhealthy diet, smoking, alcohol abuse, and physical inactivity (World, n.d.). In addition, the CDC states that high cholesterol significantly raises the risk of heart disease and stroke (Centers, n.d.). Thankfully, smoking and alcohol are also included in the dataset I have chosen for this project.

## Research Questions

The goal of this project is to conduct research and data analysis on historical cardiovascular disease data to determine if individuals with high blood pressure and/or high cholesterol are at an increased risk for cardiovascular disease. Specifically, I aim to answer the following research questions:

1. Is there an increased risk of cardiovascular disease for individuals with high blood pressure?

2. Is there an increased risk of cardiovascular disease for individuals with high cholesterol levels?
3. Are there other contributing factors, such as alcohol consumption, smoking, and glucose levels, that could affect the risk?

**Data Explanation**

I found a dataset on Kaggle.com containing 70,000 records of patient data, collected during medical examinations. The dataset includes 11 variables and a target variable, making it ideal for model development. The binary target variable indicates the presence or absence of cardiovascular disease. There was no need for any data cleanup. I checked for null values in the dataset, and since there were none, I began my analysis right away.

**Methods**

In this analysis, I used various methods to explore and model the relationship between features and the target variable, cardiovascular disease (cardio). Exploratory Data Analysis (EDA) provided initial insights into the dataset, utilizing descriptive statistics and visualizations such as histograms and box plots. I conducted correlation analysis using Pearson correlations for numeric features, Point-Biserial correlation for binary features, and the Chi-Square test for categorical features to identify significant relationships. I used Logistic Regression as a baseline model, with coefficients revealing key predictors like age, blood pressure, and weight, achieving an accuracy of 69.81%. Additionally, I employed the Random Forest classifier, an ensemble model, which slightly improved accuracy to 71.43% and provided insights into feature importance, highlighting age, blood pressure, and weight as the strongest predictors.

**Analysis**

***Box Plot Analysis***

The box plots reveal some interesting insights about the data, particularly the presence of outliers and differences in variability. Individuals with cardiovascular disease (cardio = 1) exhibit more variability in their data compared to those without, but overall, the two groups share similar ranges. This suggests that while there may be some extreme values or outliers, the general trends in both groups don't appear drastically different at a first glance.

***Pearson Correlations***

The Pearson correlation analysis highlights age and cholesterol as the strongest positive predictors of cardiovascular disease. Both variables exhibit a moderate correlation with the target variable, indicating their potential as key predictors. On the other hand, weight, glucose, and blood pressure show weaker correlations, suggesting they may not be as predictive in their current form. Features such as gender, alcohol intake, height, smoking, and physical activity demonstrate almost negligible correlations, implying little to no linear relationship with the likelihood of cardiovascular disease.

***Chi-Squared Test***

The results from the chi-squared test identify several significant predictors of cardiovascular disease. Gender, cholesterol, glucose, smoking, and physical activity all have p-values less than 0.05, indicating they are statistically significant predictors. Alcohol intake is marginally non-significant, with a p-value close to 0.05, suggesting it might not have as strong a relationship with cardiovascular disease as some of the other variables. This reinforces the notion that lifestyle factors like smoking and physical activity are more strongly associated with the disease, though alcohol intake remains a potential, albeit weaker, factor.

***Point-Biserial Correlation***

The point-biserial correlation analysis reveals that smoking and physical activity have weak but statistically significant negative correlations with cardiovascular disease. Specifically, more physical activity appears to be slightly associated with a reduced risk of cardiovascular disease, while smoking has a weak negative relationship. Alcohol intake, on the other hand, shows a very weak and marginally non-significant correlation. Despite these weaker correlations, physical activity stands out as a notable predictor, and the weaker impact of smoking and alcohol consumption suggests that other factors like age and cholesterol might be more critical.

***Logistic Regression Model***

The Logistic Regression model shows a decent level of accuracy, though it doesn't perform exceptionally well. The strongest predictors in the model are systolic blood pressure (ap_hi), weight, and cholesterol, with large coefficients pointing to their importance in predicting cardiovascular disease. Conversely, glucose and diastolic blood pressure (ap_lo) have minimal effects on the model, implying that these features might not be as important in predicting the disease. Features like gender, smoking, alcohol intake, and height contribute very little, as evidenced by their small coefficients.

***Random Forest Model***

The Random Forest model yields slightly better performance than the Logistic Regression model but still isn’t performing at an exceptional level. The most important predictors identified by the model are age, systolic blood pressure (ap_hi), and weight—consistent with established risk factors for cardiovascular disease. Cholesterol, diastolic blood pressure (ap_lo), and height also play notable roles but are less important than the top three predictors. Features such as glucose, gender, physical activity, smoking, and alcohol intake contribute very little to the predictions, suggesting that their relationship with cardiovascular

disease may be weaker or more complex, possibly requiring non-linear models for accurate prediction.

**Conclusion**

Across all analyses, it's clear that age, cholesterol, and systolic blood pressure are consistently identified as important predictors of cardiovascular disease. However, lifestyle factors like smoking and physical activity, despite being statistically significant in some tests, seem to have a weaker overall influence compared to these more established factors. The models tested (Logistic Regression, Random Forest) show that while some predictors align well with known risk factors, further improvements in model performance might be possible by refining the selection and treatment of features.

**Assumptions**

The dataset's author, Svetlana Ulianova, mentions that there are three types of input features: factual information, medical exam results, and patient-provided data. I assume all information was ethically sourced and that patients provided truthful answers.

**Limitations**

One limitation is that the dataset does not include specific cholesterol values, such as the levels of "good" (HDL) and "bad" (LDL) cholesterol. Instead, cholesterol levels are categorized as "normal," "above normal," or "well above normal." This limitation is unfortunate because having precise cholesterol values would have enhanced the depth of my analysis. Additionally, I would have liked the dataset to include a variable measuring sodium intake, which I believe would be beneficial. It would also be interesting to include a variable that tracks the amount of sunlight a patient receives daily, as I think that could have added valuable insights.

**Future Uses/Additional Applications**

While I am not a medical expert, I believe this dataset could serve as the foundation for a larger study. More variables could be added, and additional patient data could be incorporated. Over time, these datasets and analyses could be updated and improved. It's a fascinating topic, and I'm surprised I hadn't considered it earlier.

**Recommendations**

If I were to make recommendations to anyone continuing this analysis, I would suggest including the variables I mentioned earlier: sodium intake, specific cholesterol levels, vitamin D intake, and so on. Alternatively, starting a new survey with a fresh group of patients and incorporating these variables could be a valuable approach.

**Ethical Assessment**

There were no significant ethical considerations. I did not need to modify the data in any way. I do not claim ownership of the dataset, and it was ethically sourced on my end. I assume that the dataset's author ethically sourced the data when compiling and publishing it on Kaggle.com. No personal data is included in the dataset, so there are no security concerns.

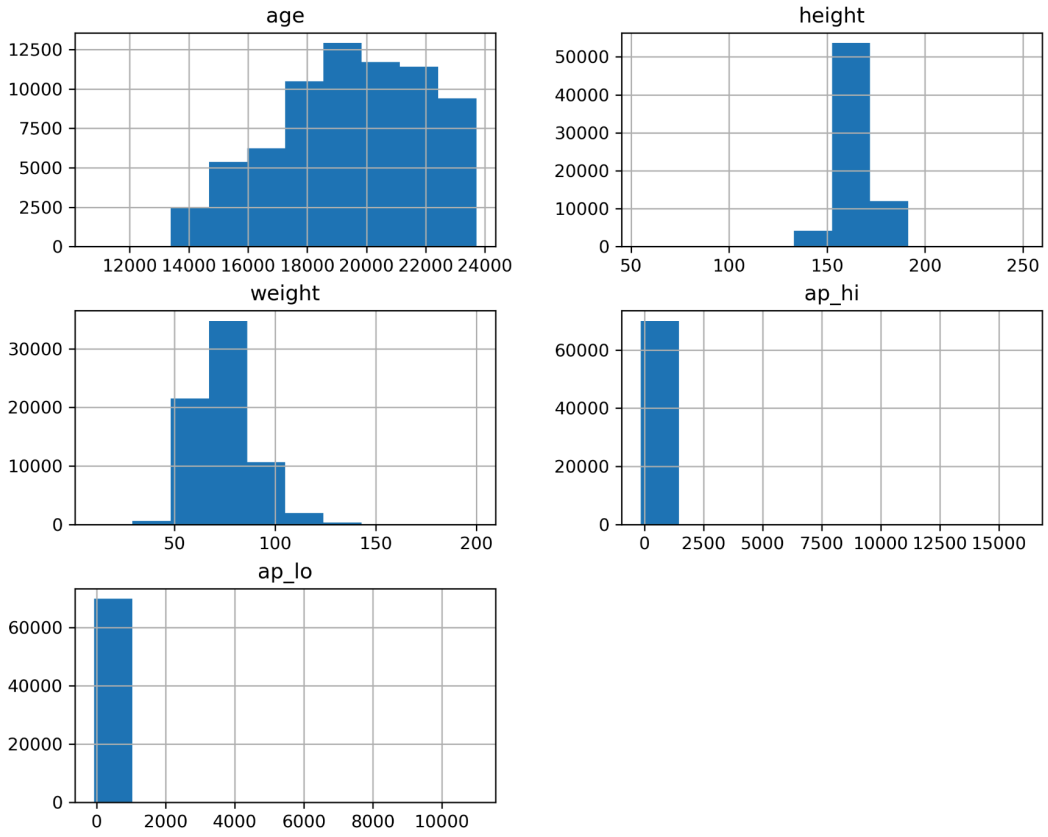
## References

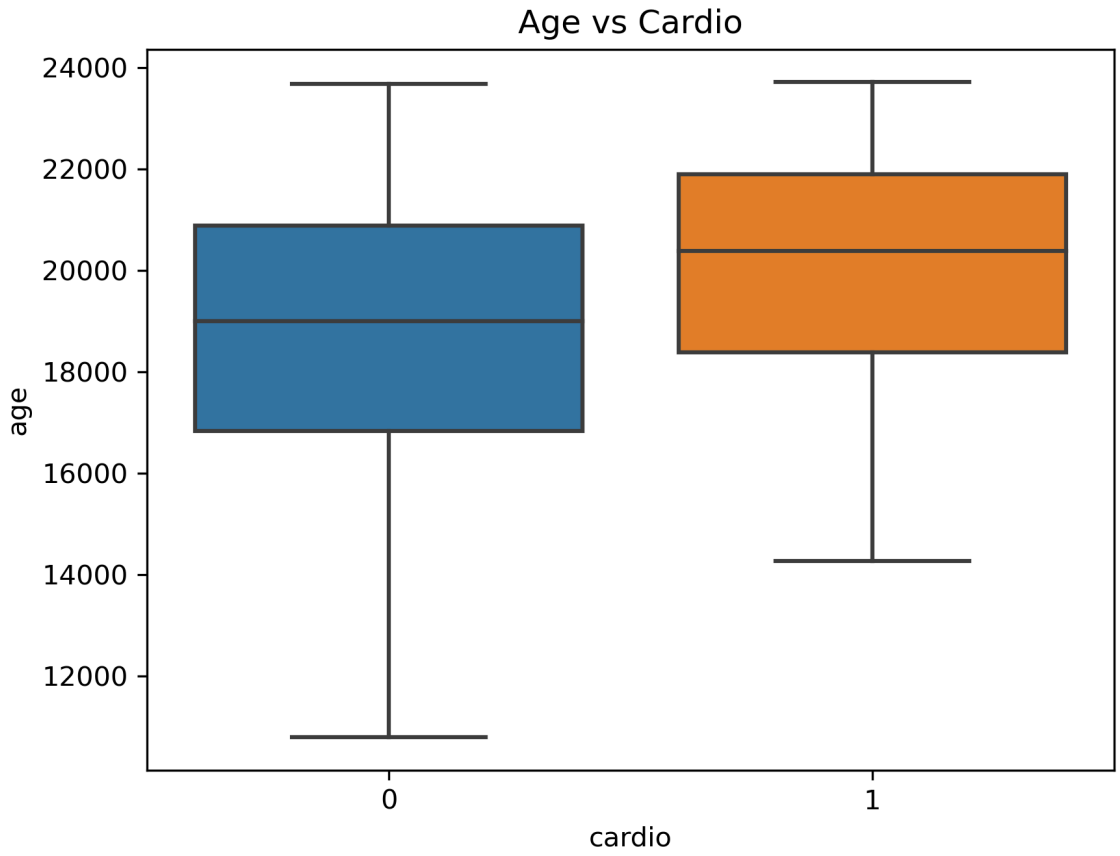
Ulianove, S. (2019, January 20). *Cardiovascular disease dataset.* Kaggle. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

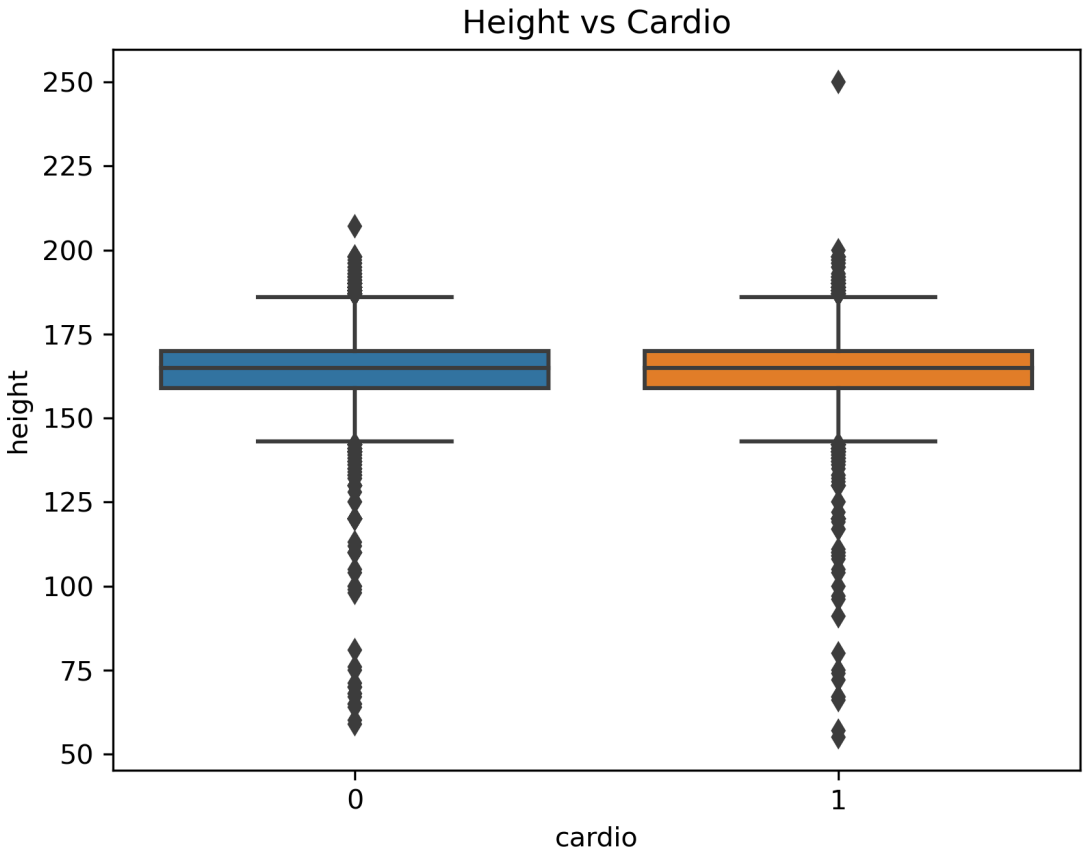
World Health Organization. (n.d.). *Cardiovascular diseases (cvds).* World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

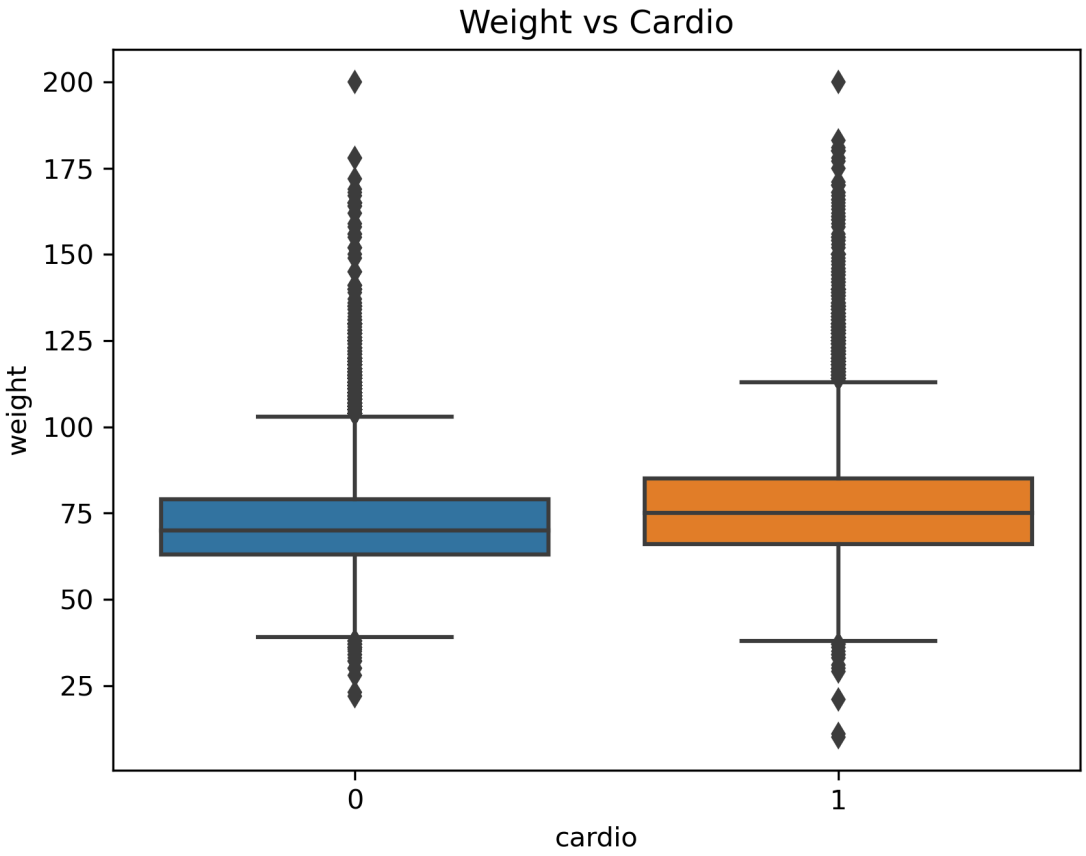
Centers for Disease Control and Prevention. (n.d.). *About cholesterol.* Centers for Disease Control and Prevention. https://www.cdc.gov/cholesterol/about/index.html

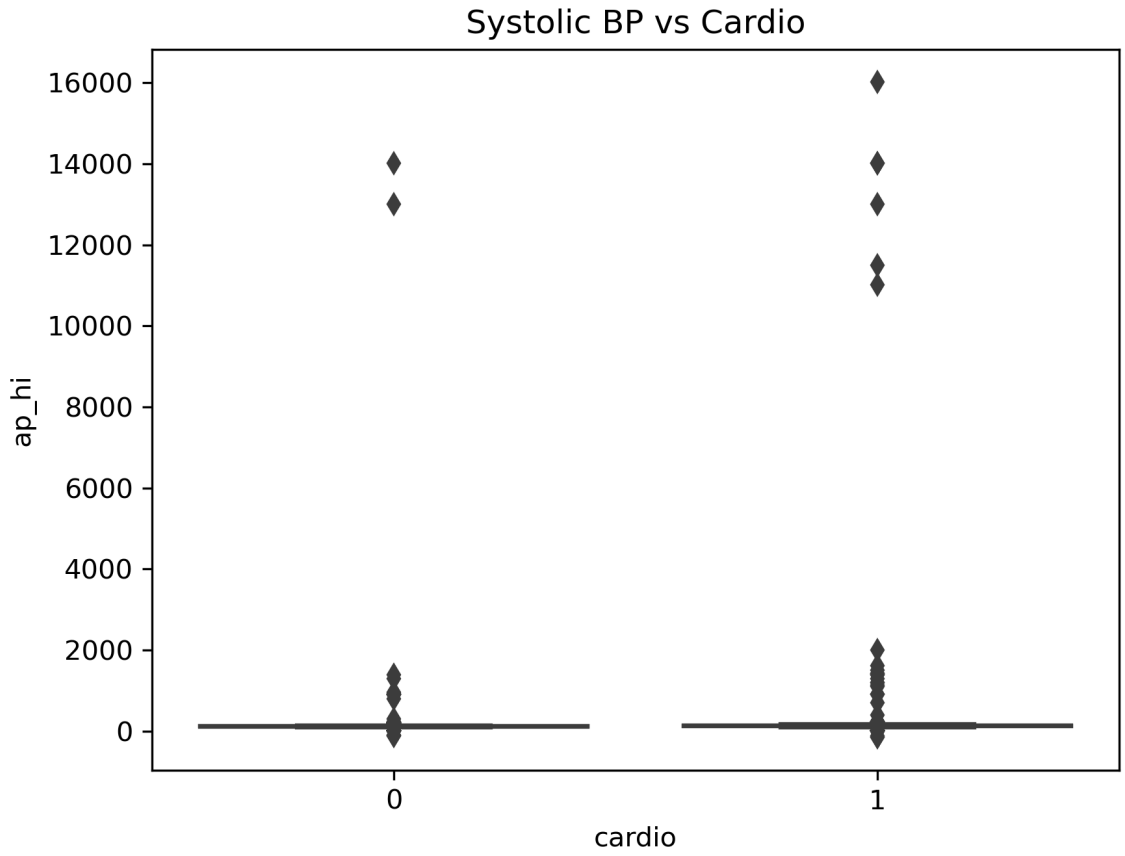
# Appendix A: Data Visualizations

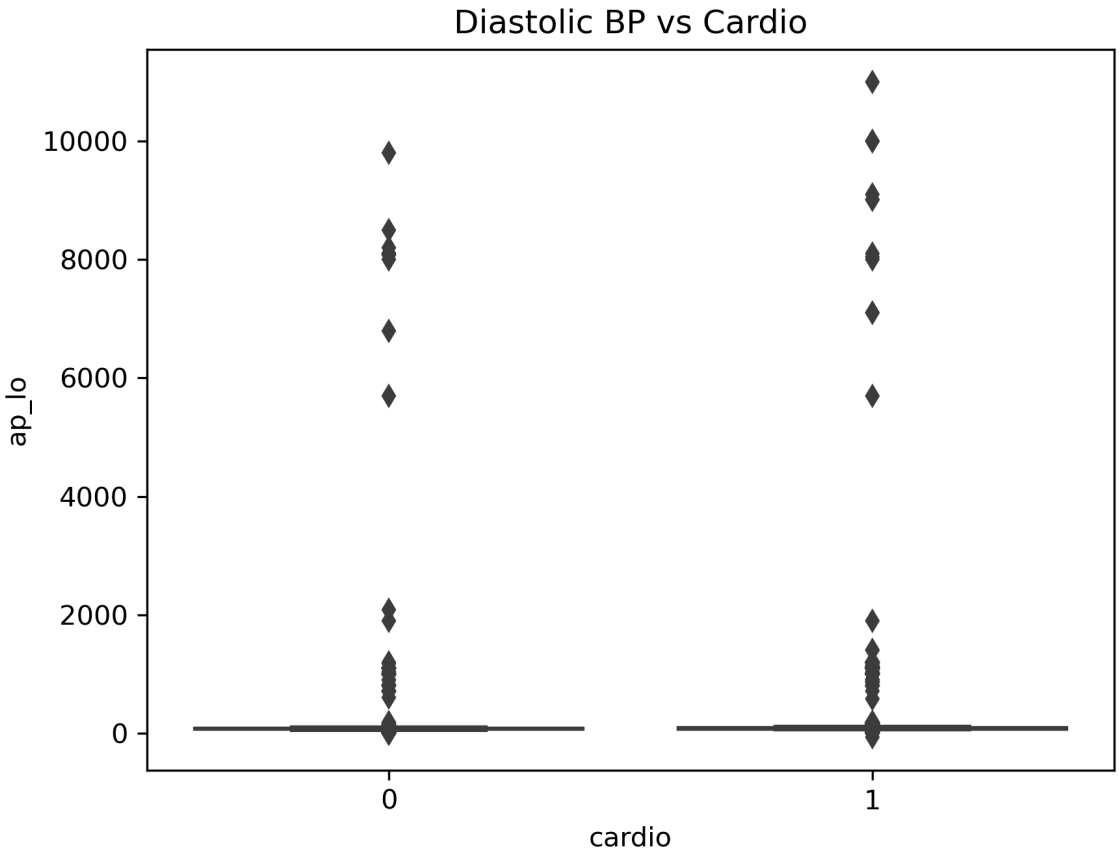


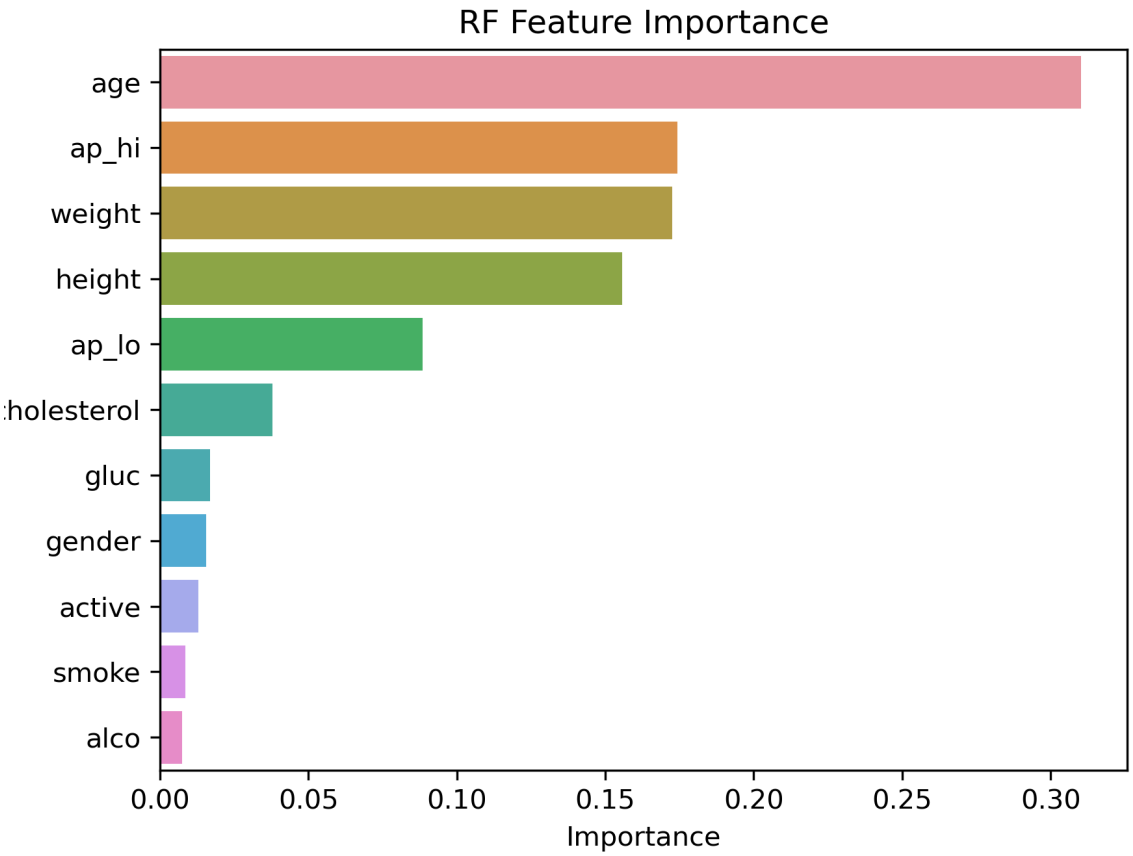

Age vs Cardio
24000
22000
20000
18000
16000
14000
12000
age
0
1
cardio


Height vs Cardio
250
225
200
175
150
125
100
75
50
height
0
1
cardio


Weight vs Cardio
200
175
150
125
100
75
50
25
weight
0
1
cardio


Systolic BP vs Cardio
16000
14000
12000
10000
8000
6000
4000
2000
0
ap_hi
0
1
cardio


Diastolic BP vs Cardio
ap_lo
10000
8000
6000
4000
2000
0
0
1
cardio


RF Feature Importance
age
ap_hi
weight
height
ap_lo
cholesterol
gluc
gender
active
smoke
alco
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Importance

## Appendix B: Potential Audience Questions

1. How do you ensure the quality and reliability of the dataset you used for your analysis?
    a. The author did not provide a source for the data, but there were no errors in the dataset, so I'm confident in the data's quality.
2. Why did you choose to use box plots for your initial analysis? Are there other visualization methods you considered?
    a. I chose box plots because they effectively show the distributions between variables and the target variable. I also considered using pair plots but had difficulty organizing the visualization, so I ultimately decided to use box plots instead.
3. What specific impact does the lack of detailed cholesterol data (HDL/LDL levels) have on your conclusions?
    a. HDL is considered the "good" cholesterol and LDL the "bad" cholesterol. Having both as separate variables would have added valuable detail to the analysis and potentially refined the conclusions.
4. Can you elaborate on how the p-value for cholesterol (0) influences your interpretation of its significance?
    a. A p-value less than 0.05 suggests strong evidence against the null hypothesis, meaning that cholesterol is statistically significant in the analysis of cardiovascular disease. However, I was surprised to see a p-value of 0.0, which suggests an extremely low likelihood that the result occurred by chance.
5. You mentioned that age was the most important feature in your random forest model—why do you think age has such a strong correlation with cardiovascular disease?

   a. I think age has a strong correlation with cardiovascular disease because, as we age, the effects of poor habits accumulate, and the body's ability to recover diminishes. While age is certainly not the only factor, it is a key component when considering cardiovascular risk.

6. You noted that alcohol consumption and smoking didn't show a significant correlation with cardiovascular disease risk in your analysis. Why do you think this might be the case, given the well-known risks associated with these factors?
   a. The analysis used binary variables for alcohol and smoking, which may not capture the full range of consumption patterns. For example, someone who drinks alcohol occasionally might be classified the same as someone who drinks heavily. A more detailed measure of these variables could have provided a clearer picture of their relationship with cardiovascular disease.
7. What challenges do you anticipate if you expand this study to include more variables, such as sodium intake or vitamin D levels?
   a. A major challenge would be going back to fill in these variables for each patient in the current dataset. It might be more effective to start a new study with expanded variables and minor adjustments to the original design.
8. Since you're working with historical data, how confident are you that these findings would apply to current populations, especially considering changes in diet and lifestyle over time?
   a. I'm not sure about the exact age of the data. Knowing the specific time frame would help clarify this. However, I do know that society has become less healthy

over time, which could affect how relevant the findings are to current populations. It may be worthwhile to conduct a new study to see if these trends hold today.

9. What are the potential practical applications of your findings? Could this analysis help in predicting cardiovascular risk in real-world settings?
    a. If some of the variables I mentioned were added, and adjustments were made to existing ones, it could improve the model's accuracy. A well-calibrated model could be useful for predicting cardiovascular risk in real-world settings.
10. Could you explain more about your recommendation to include sunlight exposure as a variable? How do you think it would impact the analysis of cardiovascular disease risk?
    a. Vitamin D, which is synthesized through sunlight exposure, is essential for health. It’s not often discussed in relation to cardiovascular disease, so I think including it as a variable could provide new insights. People tend to focus on medications, but sunlight exposure might play a more significant role than we realize, and incorporating this variable could offer surprising findings.