**Final Report**

**1. EXECUTIVE SUMMARY: REGRESSION**

1. Question of Interest

What is the relationship between a patient’s level of the CPK enzyme (creatinine phosphokinase) and their personal demographics and health history?

1. Motivation

This question is worth exploring because high levels of the CPK enzyme can indicate that there are internal damages to a patient’s brain or muscle tissue (Chen). This enzyme is particularly important to understand when it comes to heart disease, as abnormal levels of CPK are often associated with conditions such as myocarditis, cardiomyopathy, and heart attacks (Chen). Therefore, in addition to understanding what high levels of this enzyme mean for a patient’s health, it is equally as important to investigate behavioral and health characteristics that can lead to its increase in the body. Thus, we believe it is worth analyzing the ways in which the variables in this dataset are associated with heightened CPK levels. Overall, if our analysis shows that there are strong correlations between the demographic and health characteristics of a patient and the level of CPK in the blood, we can potentially glean insights into ways to prevent the enzyme from spiking as a result of damaged muscle tissue.

Relevant stakeholders for this question could include a number of people. Firstly, we believe that doctors, specifically cardiologists, would be very interested in the answers to this question. Given that the CPK enzyme levels are associated with damaging heart conditions, cardiologists would likely be interested in obtaining any insight into relationships between a patient’s demographics and lifestyle and their CPK levels, as the prevention of high CPK levels could potentially mitigate the risk of heart failure. Additionally, patients who are at high-risk for heart disease are important stakeholders when it comes to this question. If there are tangible actions that can be taken to prevent high CPK levels and therefore potential heart damage, these patients would likely be very interested in the answers to this question.

1. Analyses

The analyses undertaken in our group's previous milestones significantly contribute to a further understanding of the connection between a patient's CPK enzyme levels and their personal health history and demographics. Through a series of regression analyses, we systematically addressed this question, yielding insights into the influence of specific predictors on CPK enzyme levels. Notably, our findings highlight the role played by the presence of anemia and age as the most influential predictors of CPK enzyme levels. Furthermore, we discovered this relationship by identifying specific conditions that shape the expected CPK enzyme levels. The selected predictors, including anemia, serum creatinine, and platelet counts, consistently demonstrate a robust impact on the accurate prediction of CPK enzyme levels.

1. Recommendations

The analyses conducted by our group contribute to a deeper understanding of the intricate relationship between a patient's CPK enzyme levels and their personal health history and demographics. These insightful findings provide actionable recommendations for various stakeholders. Firstly, healthcare professionals must prioritize the education of patients with elevated CPK enzyme levels, making clear to them the associated risks to heart health. Additionally, individuals at risk of heart disease should be motivated to embrace lifestyle changes as a proactive measure. Secondly, public health authorities can implement preventive measures through educational campaigns and advocacy for policies fostering healthier lifestyles. Lastly, there is a call for medical professionals to integrate CPK education into training programs, ensuring broader awareness within the healthcare community. Overall, these recommendations aim to empower stakeholders for the proactive management of potential heart-related risks and outcomes.

**2. DATA AND VARIABLE DESCRIPTION: REGRESSION**

1. Data Set

The data set used for this analysis is made up of information from the medical records of nearly 300 heart failure patients at the Faisalabad Institute of Cardiology in Pakistan, all of which was collected during the year 2015. The dataset includes information about the lifestyle habits and clinical health of each patient, describing the potential risk factors and predictors of death from heart failure.

1. Source

We found this free dataset, called “Heart Failure Prediction,” on the website, Kaggle. The user obtained the original source material from BioMed Central (also known as BMC), which is owned by the medical research publisher, Springer Nature.

1. Variable Description

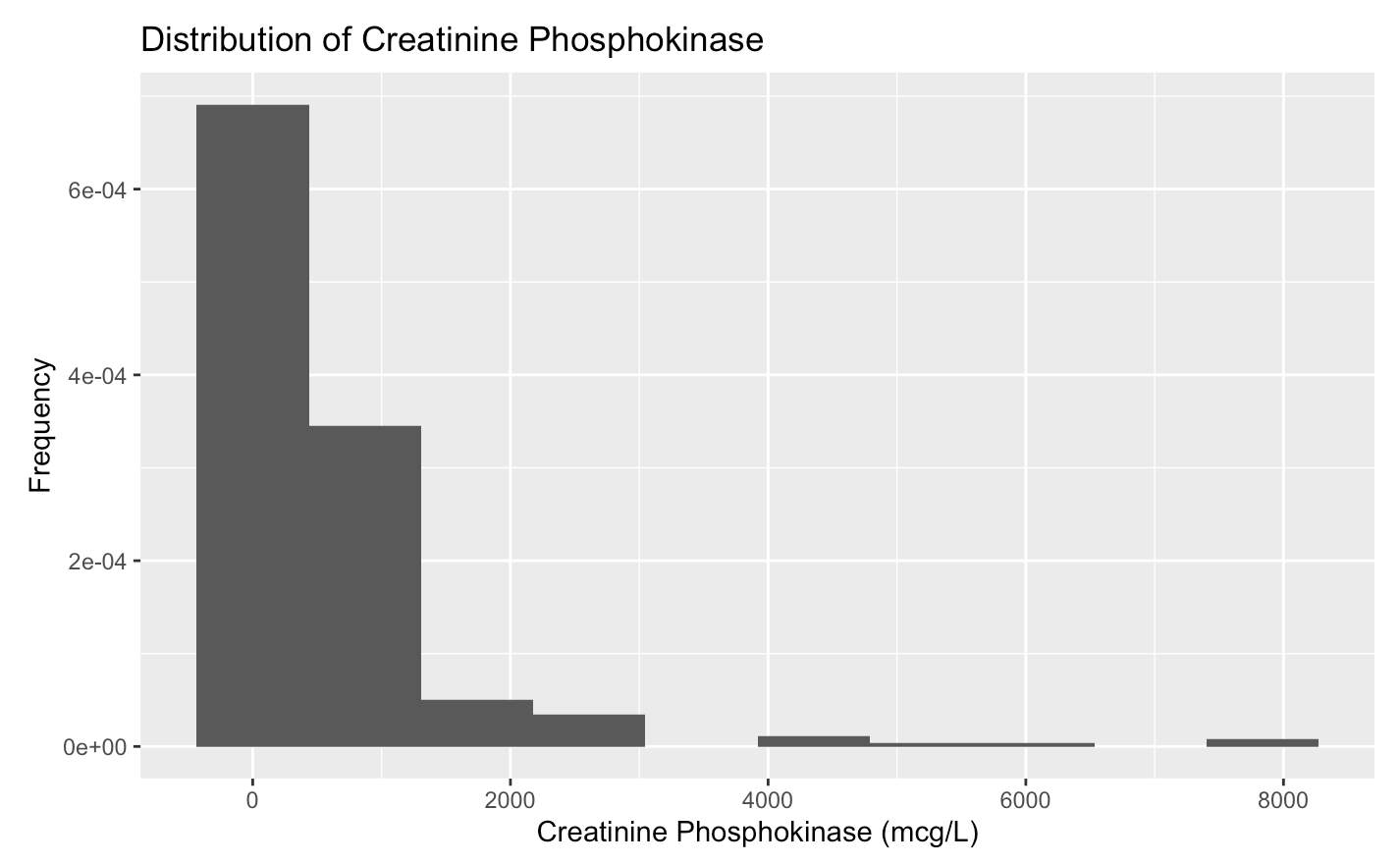
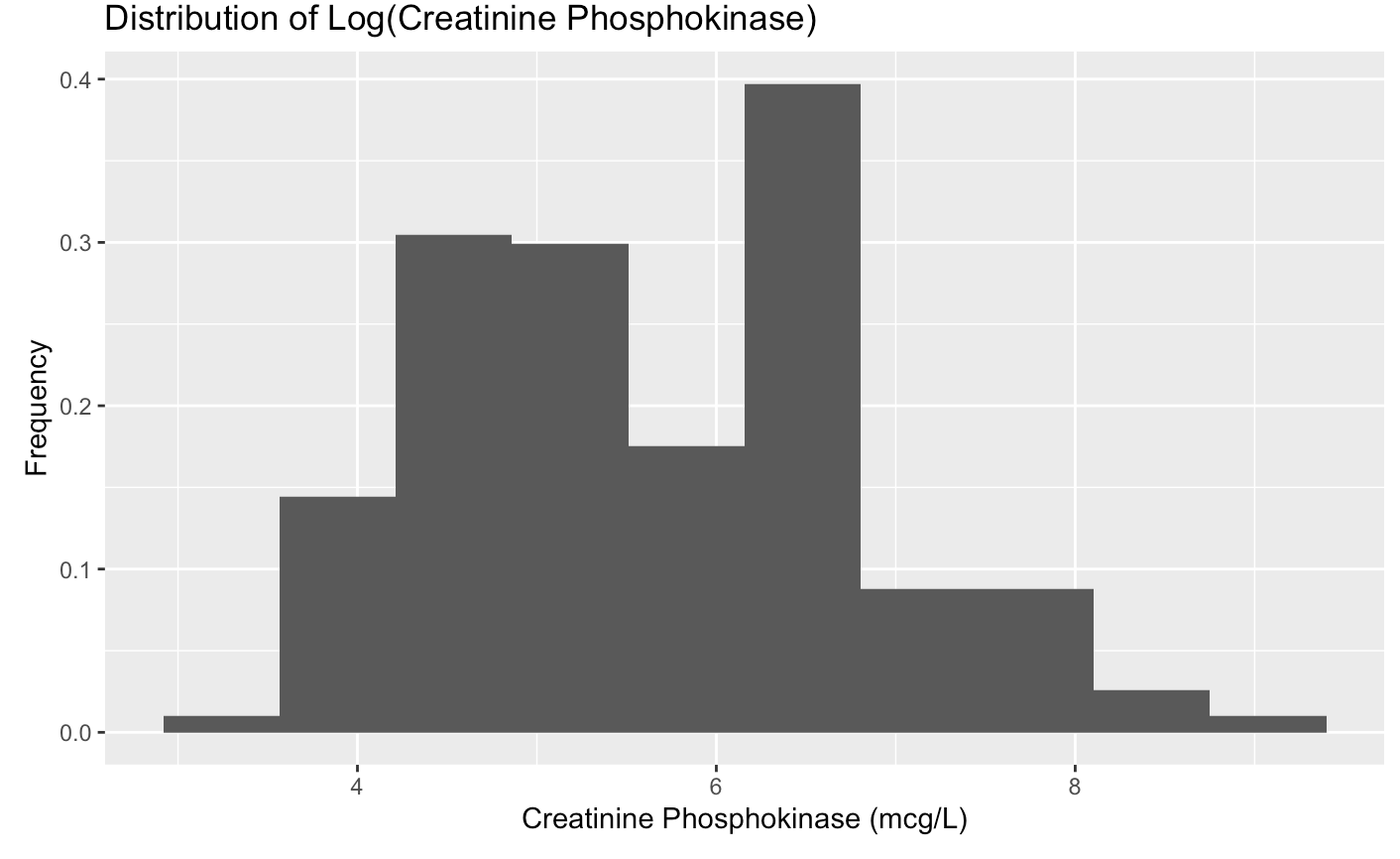
|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Description** | **Type** |
| Age | Age of the patient in years | Quantitative, continuous |
| Anaemia | Patient’s anemia status - whether or not the patient has anemia (decreased hemoglobin) | Categorical, binary  *- levels: Anemic, Non-Anemic* |
| Creatinine\_phosphokinase\*\* | Level of the CPK enzyme in the blood (mcg/L) | Quantitative, continuous |
| Diabetes | Patient’s diabetes status - whether or not the patient has diabetes | Categorical, binary  *- levels: Diabetic, Non-Diabetic* |
| Ejection\_Fraction | Percentage of blood leaving the heart at each contraction | Quantitative, continuous |
| High\_blood\_pressure | Patient’s hypertension status - whether or not the patient has hypertension | Categorical, binary  *- levels: Hypertensive, Non-Hypertensive* |
| Platelets | Level of platelets in the blood (kiloplatelets/mL) | Quantitative, continuous |
| Serum\_creatinine | Level of serum creatinine in the blood (mg/dL) | Quantitative, continuous |
| Serum\_sodium | Level of serum sodium in the blood (mEq/L) | Quantitative, continuous |
| Sex | Patient’s gender - whether the patient is male or female | Categorical, binary  *- levels: Male, Female* |
| Smoking | Patient’s smoking status - whether or not the patient smokes | Categorical, binary  *- levels: Smoker, Non-Smoker* |

\*\* indicates the response variable

**3. REGRESSION QUESTIONS**

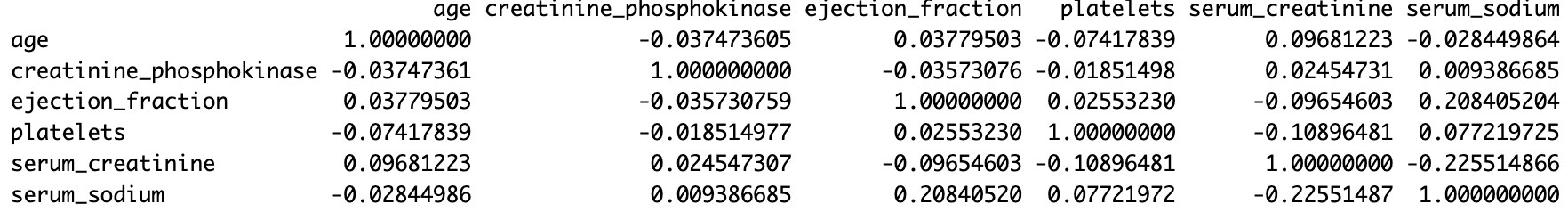
**3.1) Exploratory Data Analysis**

***Figure 1A: Figure 1B:***

*** ***

From 1A, we can see that the original response variable, ‘creatinine\_phosphokinase,’ is heavily skewed to the right. In order to correct for this, we transformed the variable using the log function, producing a much more symmetric, albeit not perfectly even distribution (1B).

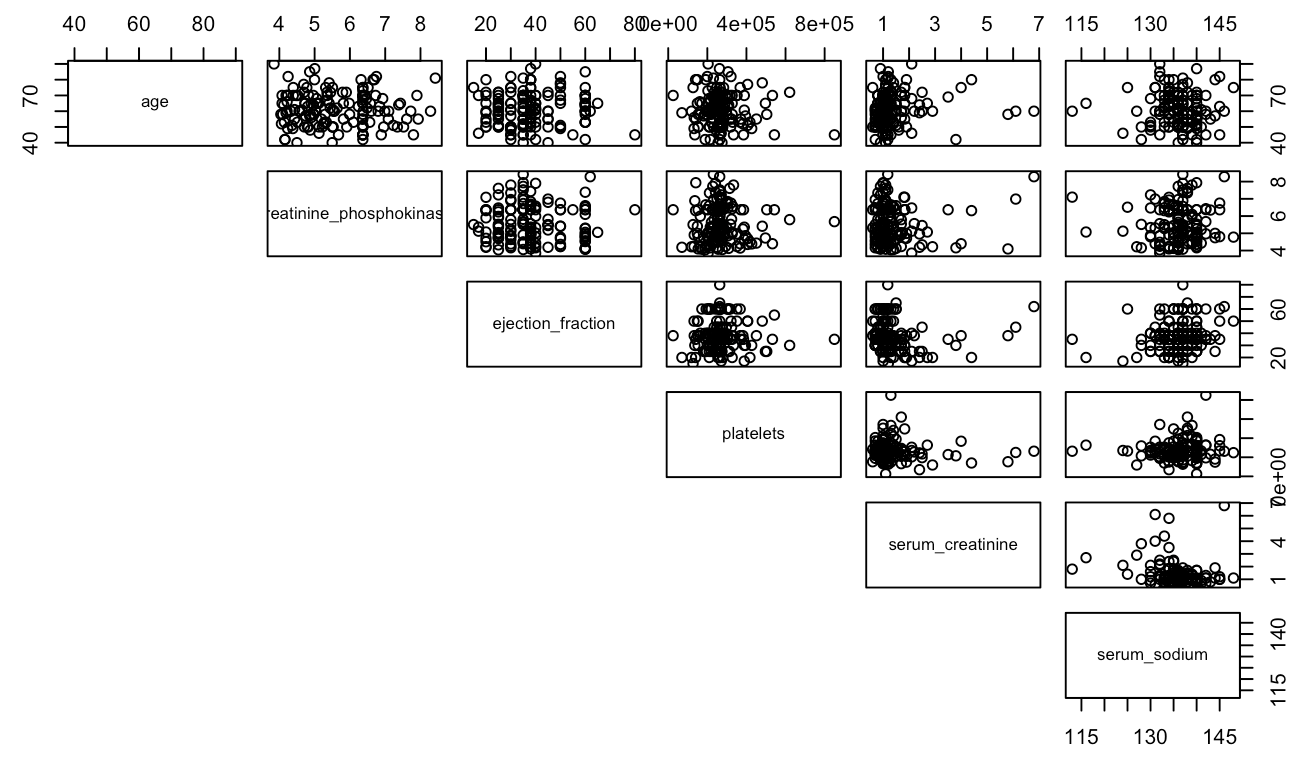
***Figure 2***

******

In the correlation matrix presented above, there are several weak linear relationships between the variables. To begin, age exhibits a weak negative correlation with creatinine\_phosphokinase (-0.0375) and platelets (-0.0742), indicating almost no correlation with creatinine\_phosphokinase and a slight tendency for platelet levels to decrease with age. Moreover, in addition to these two negative associations, age is positively correlated with ejection\_fraction and serum\_creatinine, suggesting that these variables tend to increase with age.

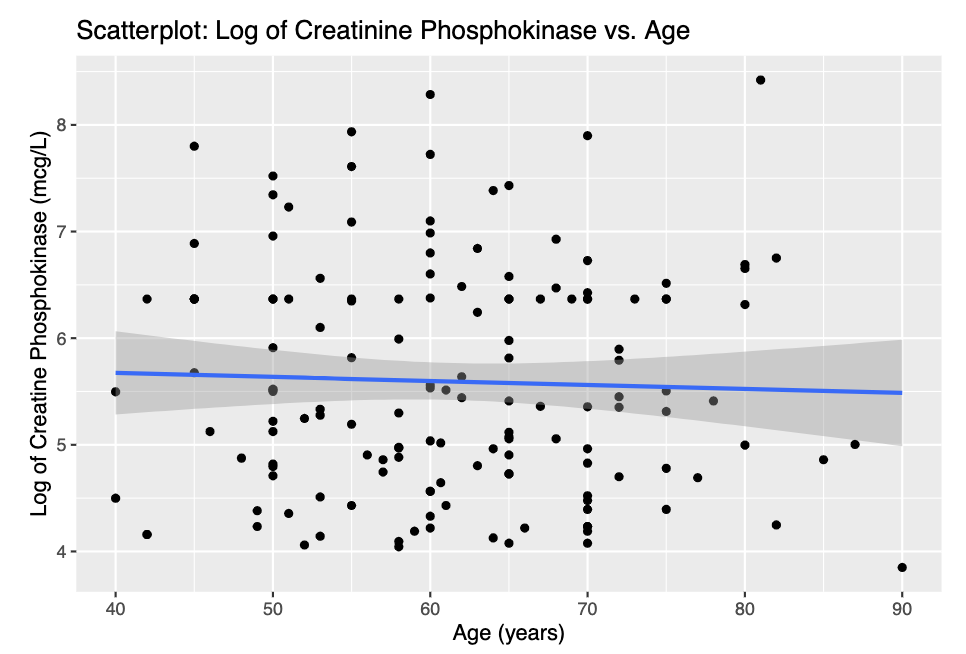
Creatinine\_phosphokinase, on the other hand, has a weak negative correlation (-0.0185) with platelets, implying a slight decrease in platelet levels as creatinine\_phosphokinase levels increase. Lastly, there is a weak negative correlation (-0.2255) between serum\_creatinine and serum\_sodium, signifying a tendency for serum sodium levels to decrease as serum\_creatinine levels increase. Furthermore, serum\_sodium and ejection\_fraction exhibit a weak positive correlation (0.2084), suggesting that serum\_sodium levels tend to rise as ejection fraction increases. As none of the variables are strongly correlated with one another, we can conclude that multicollinearity is likely not a significant issue for this dataset.

***Figure 3***



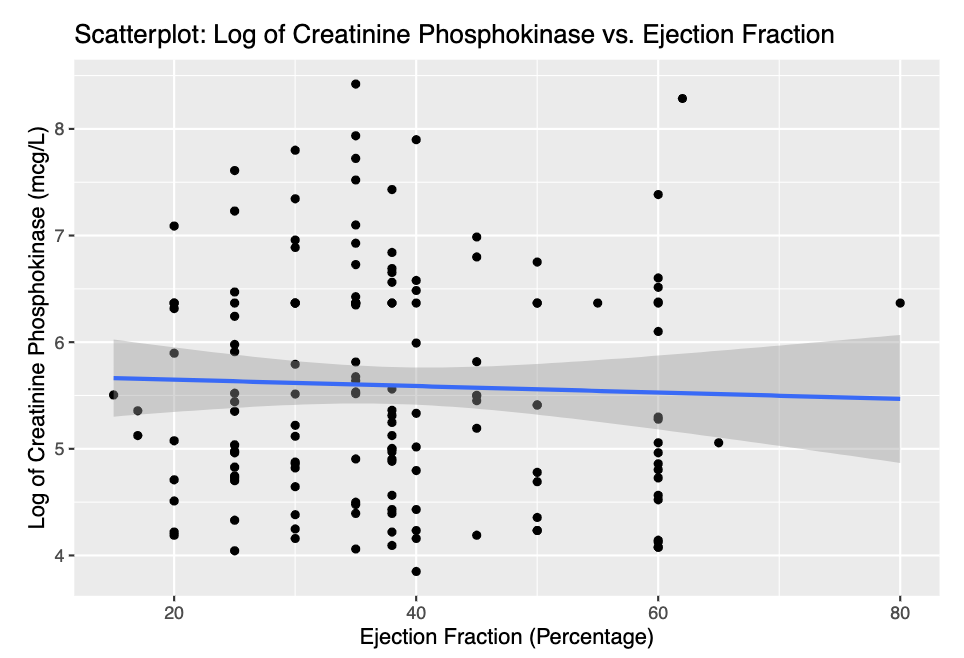
Examining the provided scatterplot matrix, we can see that the majority of relationships between the quantitative variables are very weak. Serum\_sodium exhibits two positive associations, one with platelets and the other with ejection\_fraction. This implies that as serum\_sodium increases, both platelets and ejection\_fraction tend to increase. Additionally, one of the most prominent negative relationships in the matrix is observed between serum\_creatinine and platelets, signifying that as serum creatinine levels rise, platelet counts tend to decrease. In a broader context, the scatterplot matrix reveals that the majority of these predictors appear to have no significant relationships and are largely independent of each other.

***Figure 4***

******

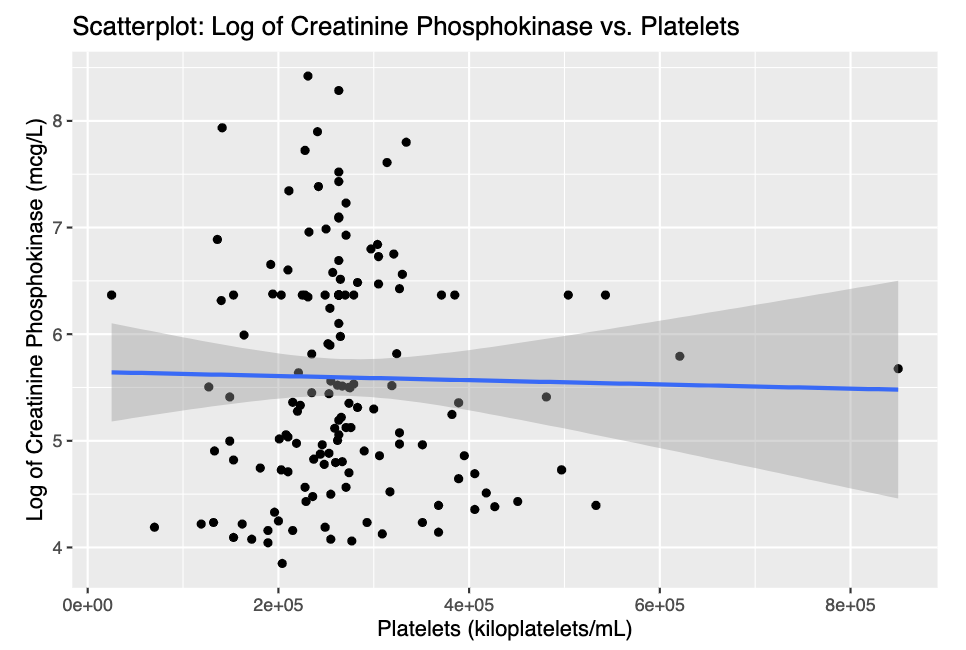
Looking at the scatterplot above, we can see that the relationship between age and creatinine\_phosphokinase is very weak. With a correlation of -0.0375 (Figure 2), we know that a very weak negative correlation exists, however the points are extremely scattered and do not display a clear trend.

***Figure 5***

******

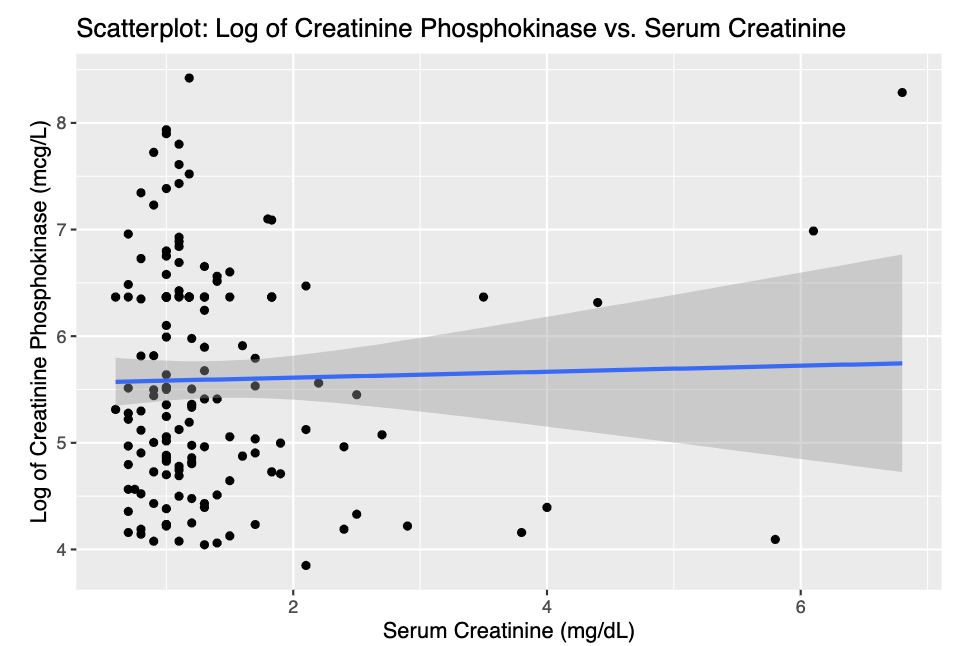
Examining the scatterplot above, we can see that there is not a clear, identifiable relationship present between ejection\_fraction and creatinine\_phosphokinase. The scatterplot does illustrate a handful of outliers that lie at the intersection of low ejection fraction and high CPK levels. These outliers may be a result of the low ejection fraction combining with other personal demographics and health characteristics to produce high CPK enzyme levels, but they could also be a result of natural variation within the population of individuals sampled. Given the weak correlation of -0.0357 (Figure 2), we can conclude that the relationship between these variables is likely negligible.

***Figure 6***



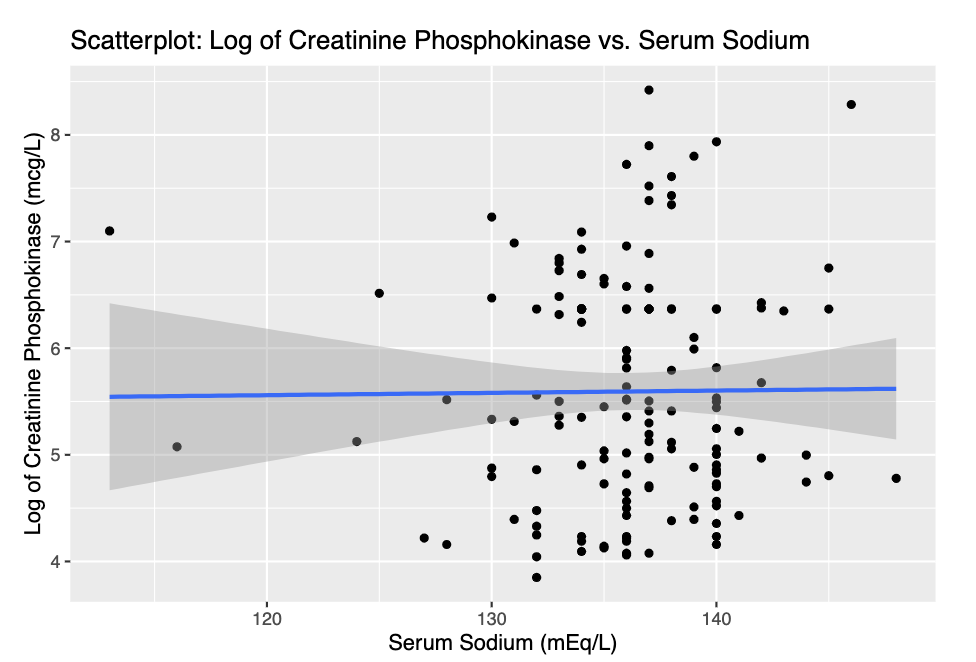
In the above scatterplot, we analyze the relationship between platelets and creatinine\_phosphokinase. A slightly negative correlation is evident, as a majority of the points display a downward trend. This trend indicates that as platelet levels increase, there is a tendency for creatinine phosphokinase levels to decrease. This association suggests that low platelet levels may pose an increased risk for higher creatinine phosphokinase levels, potentially linked to eventual heart failure.

***Figure 7***

******

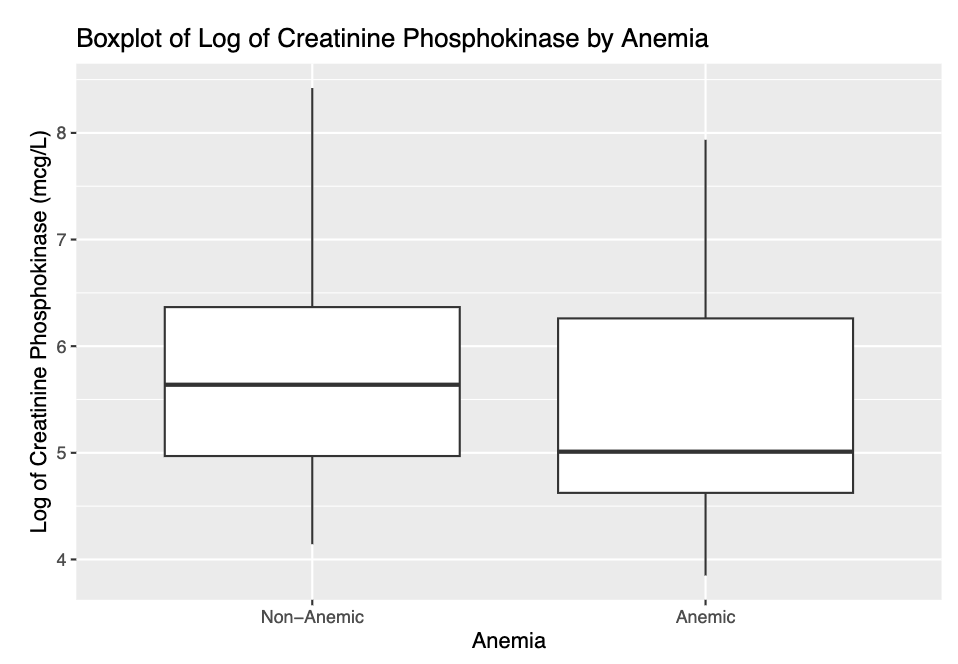
In this scatterplot, it is apparent that there is an absence of a significant relationship between creatinine\_phosphokinase and serum\_creatinine. The majority of points are densely clustered in the lower left corner, suggesting that creatinine\_phosphokinase tends to remain relatively constant across varying serum\_creatinine levels. Notably, the plot showcases outliers where creatinine\_phosphokinase remains steady in tandem with serum\_creatinine, while in other cases, creatinine\_phosphokinase increases while certain serum creatinine levels remain unchanged. This disparity confirms the lack of a strong correlation between the two variables.

***Figure 8***



Observing the presented scatterplot, a slight positive relationship between serum sodium levels and creatinine phosphokinase becomes evident. This relationship suggests that elevations in serum sodium levels correspond, to a limited extent, with increases in creatinine phosphokinase. Such a correlation may signify potential heart damage, raising concerns about the likelihood of heart failure, as heightened sodium levels could contribute to increased cardiac damage.

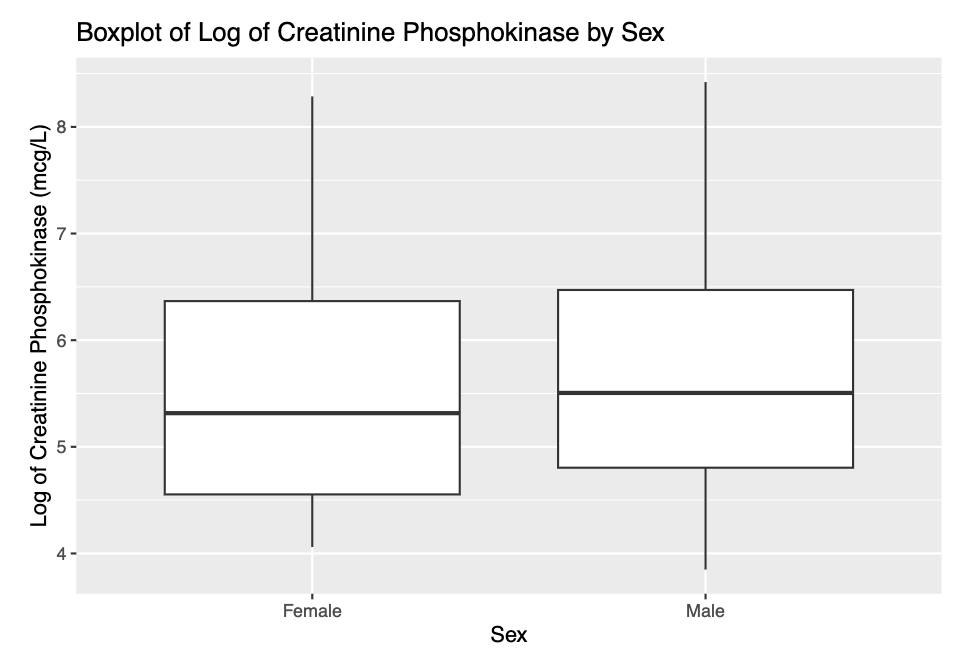
***Figure 9***

******

This boxplot displays the difference in CPK levels between anemic and non-anemic patients. The median value is higher for the first group than the second. This indicates that non-anemic people have higher median CPK levels than anemic people.

The non-anemic group also has more outliers, showing more extreme cases of high CPK concentration. This is surprising because it is typically understood that those that are anemic are likely to have higher levels of CPK, as high creatinine levels indicate issues with kidney function, potentially leading to anemia.

***Figure 10***

******

This graph indicates that men and women are very similar regarding median CPK levels, with men having slightly higher levels on average. This was not surprising, as men on average have a higher percentage of muscle mass than women. The box plot also shows that men have a slightly higher interquartile range (IQR) than women. This indicates that there is more variability in CPK levels among men.

**3.2) Shrinkage Methods**

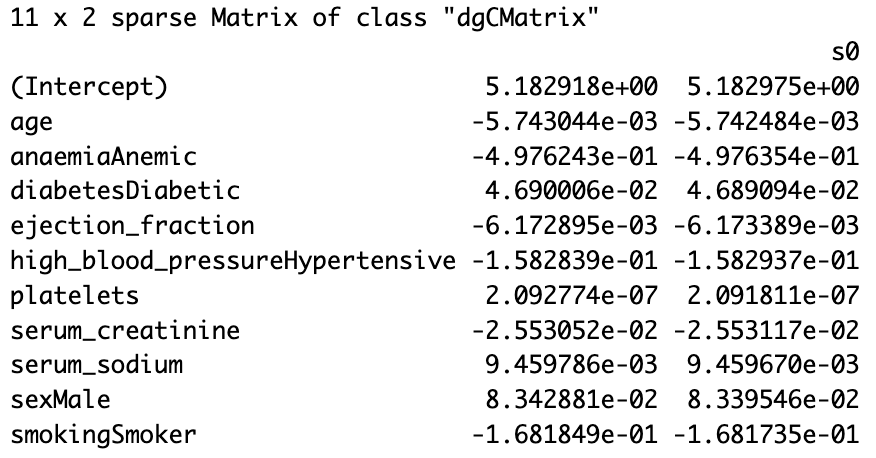
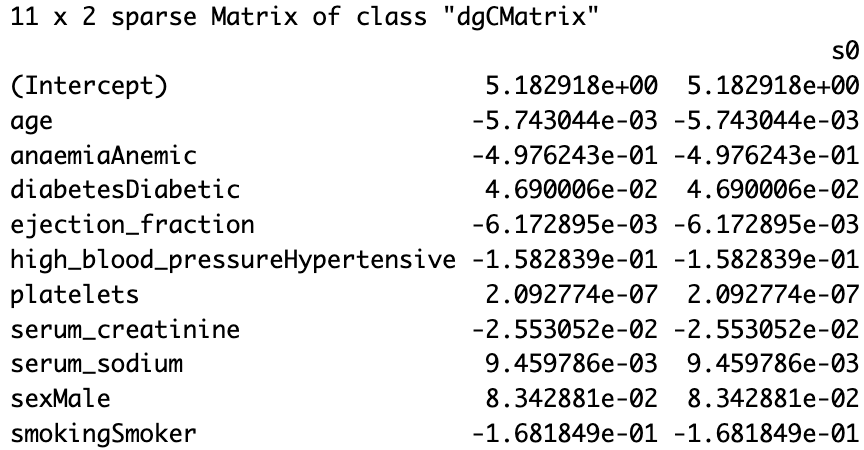
1. Included Predictors

As highlighted in the exploratory data analysis, none of the variables are strongly or directly related to one another, so there was no need to exclude variables as a result of multicollinearity. That being said, one of the variables, DEATH\_EVENT, was excluded, as it does not make sense in the context of the question. We do not know whether someone will die until after the fact, making it a useless predictor in the context of this regression analysis.

1. Threshold Value

When choosing the appropriate threshold value for the minimization procedure, it is important that the estimated coefficients for ridge regression are very close to, or ideally equal to, the estimated coefficients from ordinary least squares. This is because when the tuning parameter (λ) is equal to 0, the coefficients should be the same. In order to ensure equality in these coefficients, we first used the default threshold, finding that the coefficients were not equal. Thus, we decided to reduce the threshold to 10-23, which in turn gave us estimated ridge regression coefficients that are equal to the estimated ordinary least squares coefficients.

**Threshold: 10-7 Threshold: 10-23**

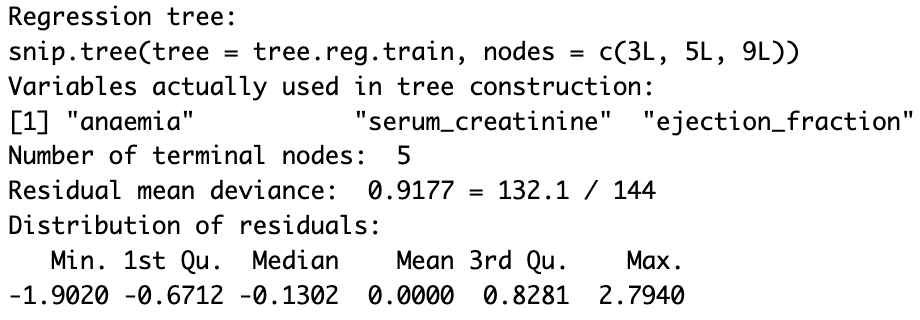
 

**3.3) Regression Trees**

1. Recursive Binary Splitting vs. Pruned

We decided to present the pruned regression tree, rather than the original recursive binary splitting tree, as it is simpler and easier to interpret. The original tree has 21 terminal nodes, while the pruned tree has only 5. Additionally, the test MSE for the pruned tree is lower, at about 1.488, than the test MSE of the original tree, at about 1.860.

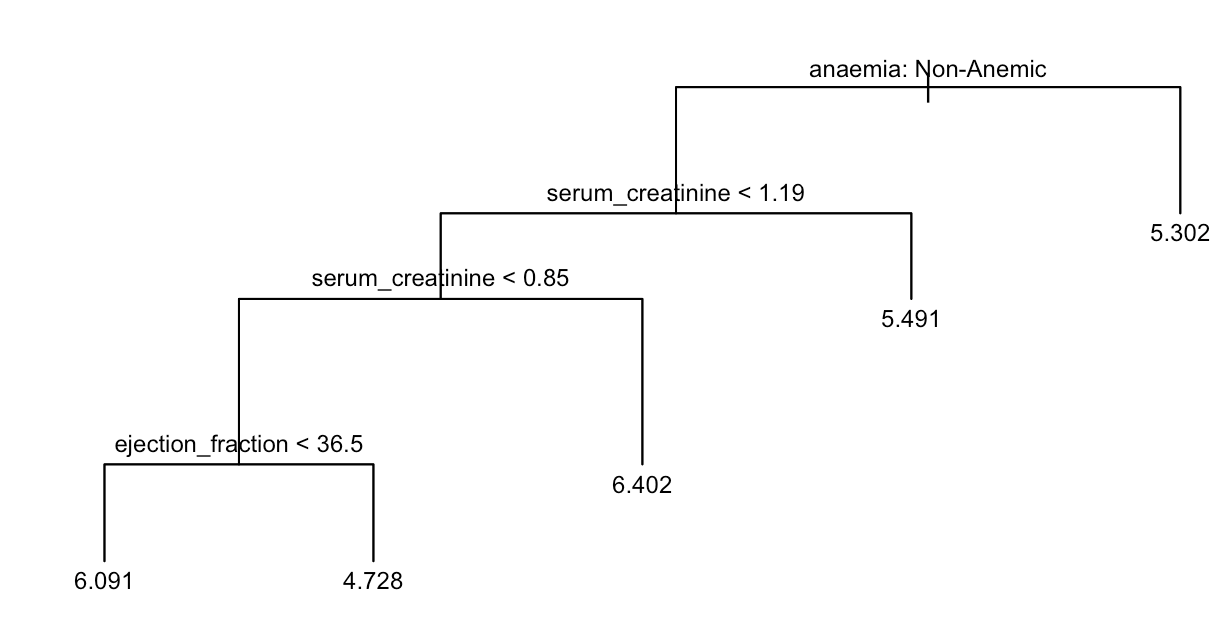
1. Summary Output



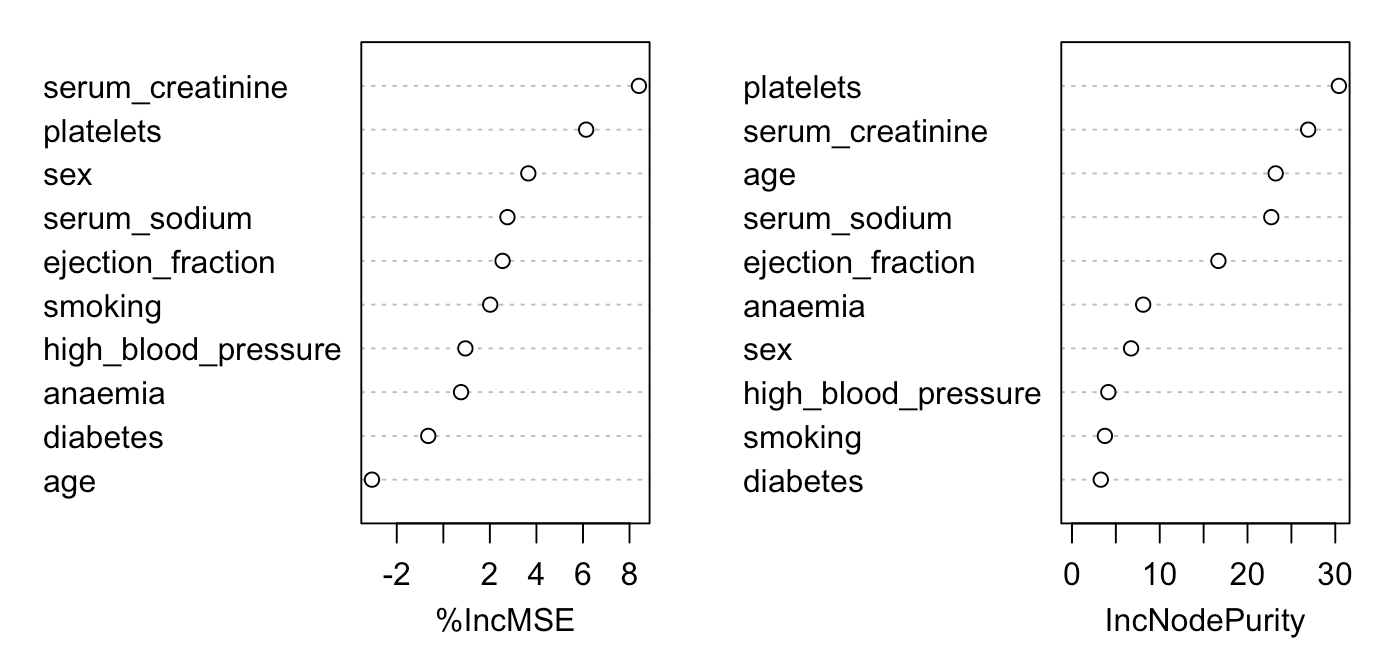
1. Terminal Nodes

The pruned tree has 5 terminal nodes.

1. Graphical Output



1. Important Predictors

****

**3.4) Summary of Findings**

1. Table of MSEs

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Least Squares** | **Ridge** | **Lasso** | **Pruned Tree** | **Random Forests** |
| 1.438754 | 1.444791 | 1.402563 | 1.487567 | 1.398362 |

1. **MSE Commentary**

For this analysis, a test MSE of zero would indicate that our model is performing perfectly. Thus, as it is extremely unlikely that this would occur in actuality, we can evaluate the performance of each of the methods based on the size of the test MSEs in relation to the response variable, with smaller values indicating better performance. While each of the test MSEs is relatively close to zero (none being larger than 1.5), we must consider the log transformation that we performed on the response variable prior to completing the analysis. This transformation to the log of creatinine\_phosphokinase produced a variable that ranges from around 3 to 8 mcg/L. Thus, while the values of the MSEs are small, they represent a moderate error size when considered in relation to the response variable. As a result, we can conclude from the test MSEs that our model performed moderately well and could likely be further improved.

1. **Discussion**

Each of the subsections above provides valuable insight into our question of interest. Firstly, through the exploratory data analysis, we can see that there are very few significant relationships between the quantitative predictors and the response variable, with ejection\_fraction and age exhibiting the most significant correlations of -0.0378 and -0.0374, respectively. From this, we can determine that there is a weak, negative relationship between these two predictor variables and the response, meaning that when ejection\_fraction and age decrease, the CPK enzyme levels tend to increase slightly. Through the boxplots, we can see that the median CPK enzyme level is higher for non-anemic patients and lower for female patients. This is indicative of a potential correlation between these variables and the response.

From the shrinkage methods section, each of the three methods are useful in analyzing our question of interest. First, ridge regression can be used to identify which patient attributes are most significant in predicting CPK enzyme levels. The regression will shrink the less important predictors but still retain them with reduced impact through regularization. Secondly, the lasso regression can be used for variable selection, helping to identify which variables are most related to CPK enzyme levels. In the case of our analysis, lasso determined these predictors to be “anaemia” and “age.” Finally, OLS regression can be used to determine the relationship between CPK enzyme levels and the variables without any form of regularization by offering direct estimates of each coefficient.

Lastly, from the regression tree section, we can even further answer the question of interest. It begins with an initial node based on the presence of anemia in the patient, leading to further divisions into decision nodes. These nodes, which branch further into additional decision nodes or terminal nodes, offer pivotal insights into our area of interest. Through the values of variables within each node, we determine if an individual conforms to certain conditions, thereby determining the anticipated CPK enzyme level. For instance, a patient without anemia, exhibiting serum creatinine levels below 1.19 and 0.85, alongside platelet counts under 245,000, is predicted to have a log CPK enzyme level of 7.239, which when exponentiated, is 1,392.7 mcg/L. This predictive value is based on the average observation within the training set, characterizing that specific segment of the data. These predictors were chosen for this decision tree because they all could have a strong impact on the response variable.

1. **Best Method**

Upon evaluating our five models, it becomes evident that the random forest model outperforms in accurately predicting CPK enzyme levels based on our specific predictors. Comparing the Test MSE values, the Random Forest exhibited the lowest score (1.398362). The reason for the random forest model achieving heightened accuracy could potentially be attributed to its methodology of aggregating multiple trees trained on varied data subsets, resulting in superior predictive performance.

**3.5 Previous Comments**

In evaluating the performance of the various regression methods, it is important to note that the response variable was log-transformed. This, in effect, changes the original distribution, which was heavily right-skewed, to a more normal and symmetric distribution. This transformation remained present throughout the regression exploration, only reversing this log transformation (done by exponentiating the variable) for the upcoming classification analysis.

**4. EXECUTIVE SUMMARY: CLASSIFICATION**

1. Question of Interest

Are there different health characteristics for patients who die from heart failure than for those who do not die from heart failure?

1. Motivation

This question is worth exploring because heart failure is a common cause of death, with almost 380,000 mentions on death certificates in 2018 in the United States alone (CDC). Heart failure also has a large financial impact, costing $30.7 billion in 2012 in medicine and healthcare costs (CDC). Further research on risk factors can lead to earlier treatment, which would decrease both the death rate and the financial cost of heart failure. An analysis that pinpoints the main predictors of heart failure will help physicians estimate the direness of a patient’s condition. This will allow them to allocate resources to the most serious cases and save lives of people with high risk factors. This knowledge could also help public servants pursue beneficial health policy. If a predictor is especially correlated with death by heart failure, policymakers can redirect funding to research the root cause of that predictor and how to alleviate the issue.

Relevant stakeholders for this question could include a number of people. Firstly, as was true with our regression question, we believe cardiologists would be very interested in the answers to this question. By understanding the risk factors that lead to heart failure, cardiologists can implement treatment plans at an earlier and more effective stage. This would potentially prevent a heart condition from developing, worsening, or causing a death. Additionally, patients who are at a high risk for heart failure are stakeholders in this question, the answers to which could point to a course of prevention and mitigation that might not otherwise be clear. Finally, we believe public servants are stakeholders. As mentioned above, health policy could benefit significantly from further knowledge into heart failure, and public servants could channel funding into the appropriate areas. This could not only reduce the number of deaths due to heart failure, but also limit the financial impact of heart failure in the United States.

1. Analyses

Our analyses have shed light on some of the greatest risk factors associated with a death event from heart failure. In multiple analyses, we found that the three main predictors of a death event were serum creatinine, ejection fraction, and age. A serum creatinine level of above 1.815 mg/dL was found to be a very strong indicator of death, with most patients passing away at this level. Ejection fraction and age were also critical, though their effects depended on other variables. A high ejection fraction (above 32.5%) was associated with death *if* the patient was old enough. To be more specific, patients above 79 years old were the most at risk. Similarly, a low ejection fraction had a higher likelihood of death *if* the serum creatinine levels were moderate to high, above 0.85 mg/dL. In summary, an elderly patient with a high ejection fraction, or a high-serum-creatinine patient with a low ejection fraction, faces an increased chance of death.

1. Recommendations

These findings can help doctors, citizens, and public servants make more informed decisions about this widespread health issue. Firstly, doctors should utilize serum creatinine tests more frequently for patients with a family history of heart failure. Currently, such tests are predominantly used to test for kidney issues like kidney disease (Mayo). Greater prescription of these tests will catch predicted cases of death from heart failure before they happen, and will save lives. A lesser, but still important, action of medical professionals is to also be more willing to prescribe ejection fraction tests. Echocardiogram, cardiac catheterization, magnetic resonance imaging (MRI), computerized tomography (CT), and nuclear medicine scan are examples of such tests (Mayo). Because these are more expensive tests and ejection fraction is a less crucial predictor than serum creatinine, medical practitioners should allocate resources efficiently and prioritize serum creatinine tests.

Policymakers also have a hand in making these tests more available. They can do so through changes in Medicaid and Medicare policies. Firstly, state and federal leaders should ensure that insurance covers these necessary tests. This will prevent at-risk people from declining these tests to save money. Secondly, lawmakers should make a priority of increasing the number of serum creatinine tests. A shortage of tests could lead to rationing, which may prevent a person with high serum creatinine from identifying their risk and receiving proper treatment. Increased funds to the development of such tests would ensure greater prevention of death.

The most important stakeholder is the person experiencing a death event. As regular citizens have less medical knowledge and power than a cardiologist or policymaker, they are limited in their options to prevent death events. Individuals and families should recognize the correlation between age and death from heart failure, and seek out greater medical attention with greater age. Additionally, as their doctors are more informed on serum creatinine and ejection fraction levels, patients with a high risk of heart failure should follow the advice of their doctors.

**5. VARIABLE DESCRIPTION**

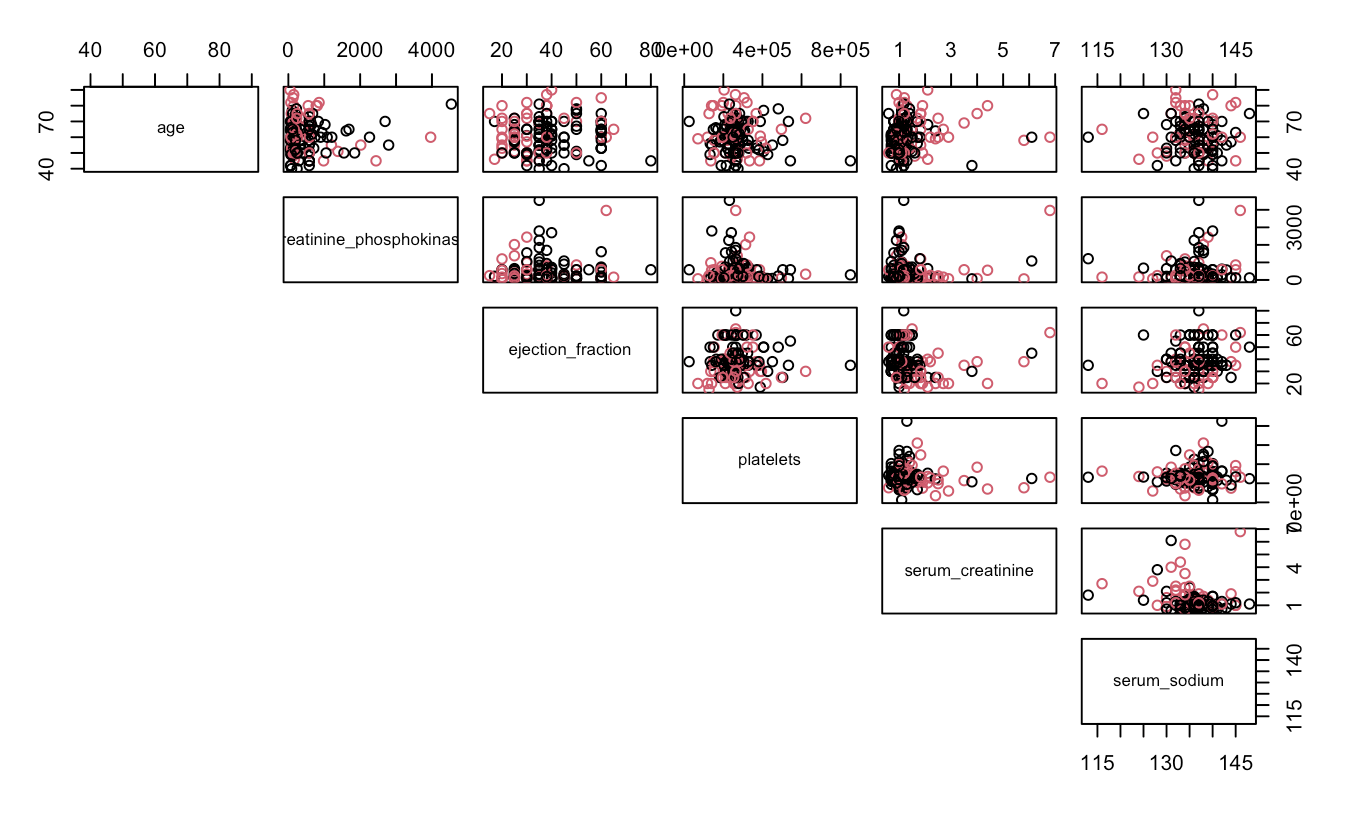
|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Description** | **Type** |
| DEATH\_EVENT \*\* | Whether or not the patient died from heart failure | Categorical, binary  *- Levels: Death, No Death* |
| Age | Age of the patient in years | Quantitative, continuous |
| Anaemia | Patient’s anemia status - whether or not the patient has anemia (decreased hemoglobin) | Categorical, binary  *- levels: Anemic, Non-Anemic* |
| Creatinine\_phosphokinase | Level of the CPK enzyme in the blood (mcg/L) | Quantitative, continuous |
| Diabetes | Patient’s diabetes status - whether or not the patient has diabetes | Categorical, binary  *- levels: Diabetic, Non-Diabetic* |
| Ejection\_Fraction | Percentage of blood leaving the heart at each contraction | Quantitative, continuous |
| High\_blood\_pressure | Patient’s hypertension status - whether or not the patient has hypertension | Categorical, binary  *- levels: Hypertensive, Non-Hypertensive* |
| Platelets | Level of platelets in the blood (kiloplatelets/mL) | Quantitative, continuous |
| Serum\_creatinine | Level of serum creatinine in the blood (mg/dL) | Quantitative, continuous |
| Serum\_sodium | Level of serum sodium in the blood (mEq/L) | Quantitative, continuous |
| Sex | Patient’s gender - whether the patient is male or female | Categorical, binary  *- levels: Male, Female* |
| Smoking | Patient’s smoking status - whether or not the patient smokes | Categorical, binary  *- levels: Smoker, Non-Smoker* |

\*\* indicates the response variable

**6. CLASSIFICATION QUESTION**

**6.1 Exploratory Data Analysis**

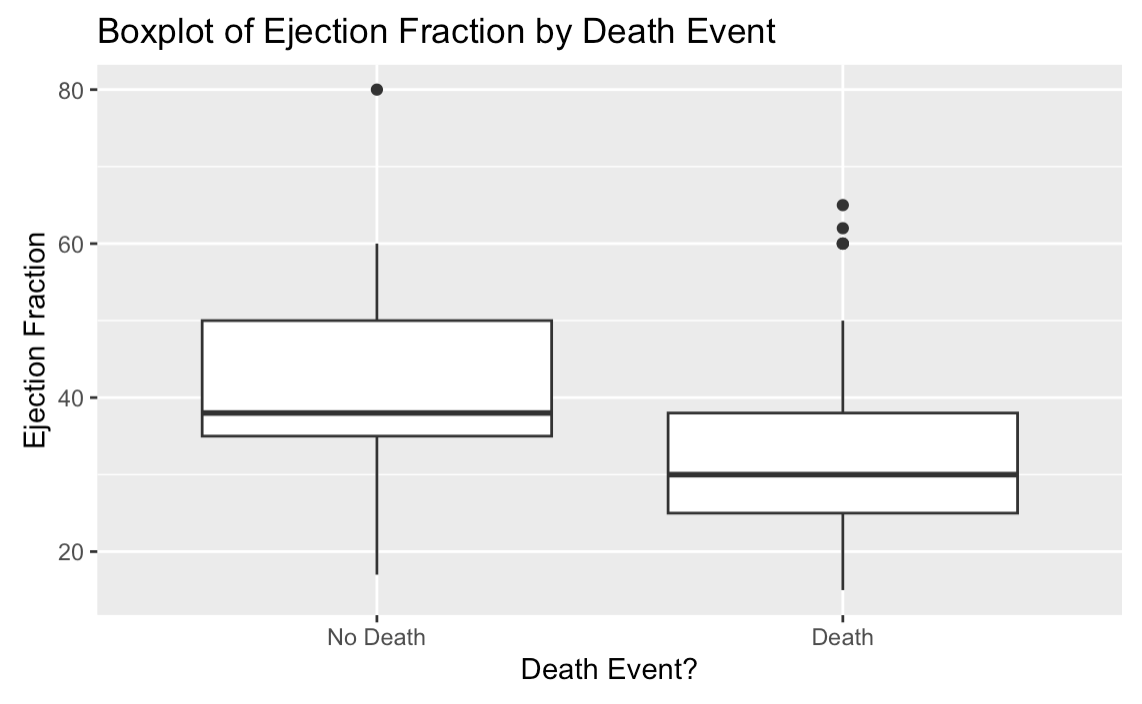
***Figure 1***

******

Examining the provided scatterplot matrix, we can identify several noteworthy relationships. Serum\_sodium exhibits two positive associations, one with platelets and the other with ejection\_fraction. This implies that as serum\_sodium increases, both platelets and ejection\_fraction tend to increase. Additionally, one of the most prominent negative relationships in the matrix is observed between serum\_creatinine and platelets, signifying that as serum creatinine levels rise, platelet counts tend to decrease. In a broader context, the scatterplot matrix reveals that the majority of these predictors appear to have no significant relationships. This suggests that these variables are largely independent of each other, and their behaviors do not exhibit strong correlations.

It is also worth noting that there is significant overlap between the patients who died from heart failure and those who did not die, and the red (death) and black (no death) points are somewhat separated for only a few of the variables. We would prefer to see complete separation between these data points, so this scatterplot indicates that our classification will likely perform only moderately well.

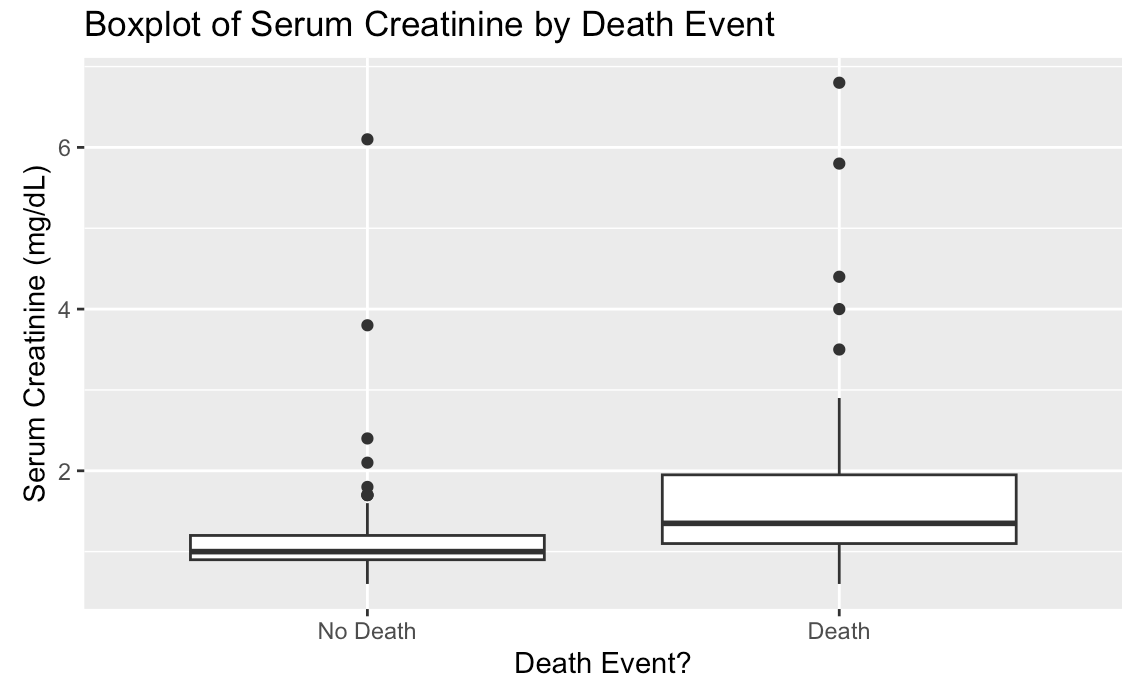
***Figure 2***

******

In this boxplot, we see that the median ejection fraction level is noticeably higher for patients who survived heart-failure. This suggests that individuals with higher ejection fraction levels are more likely to survive from heart failure than those with lower ejection fraction levels. This is unsurprising to us, as high ejection fraction levels are usually indicative of good cardiac health, whereas lower levels indicate that the heart is not effectively pumping blood around the body in a strong enough manner (Penn Heart and Vascular Blog).

In addition to the differences in the medians, we can see that the boxplot for the patients who survived heart failure exhibits a larger interquartile range (IQR) for ejection fraction level as compared to the boxplot for those who died from heart failure. This indicates that there is greater variability in ejection fraction levels for patients who survived heart failure.

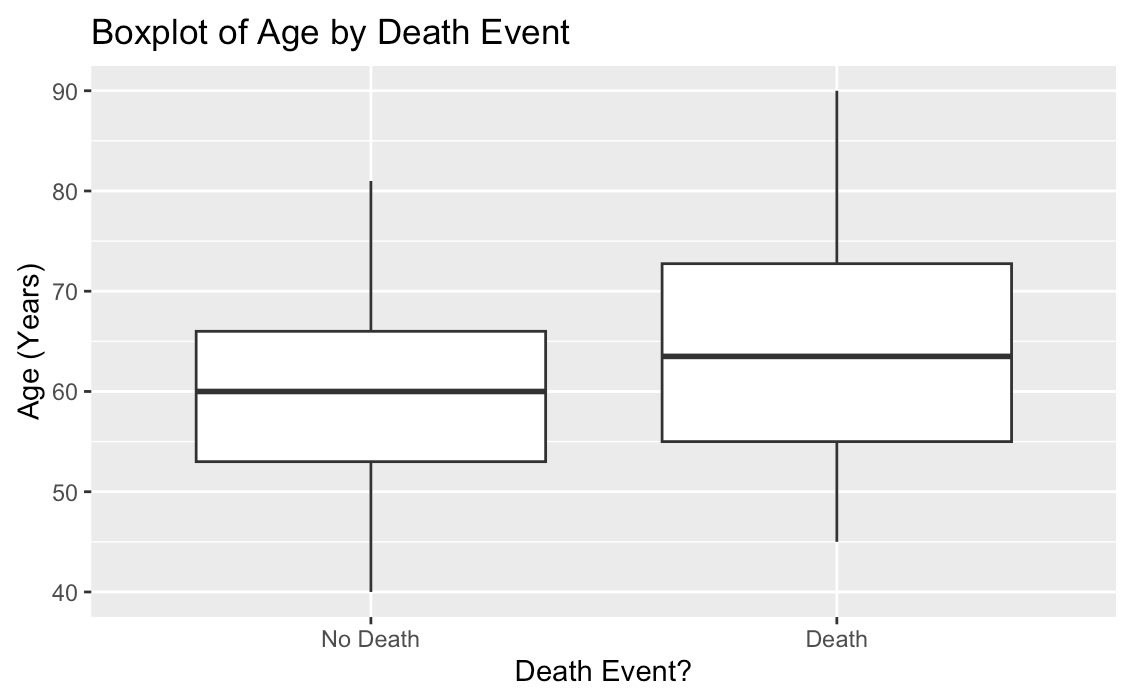
***Figure 3***

******

In this boxplot, we see that the median level of serum creatinine is higher for patients who died from heart failure. This suggests that individuals with a higher level of serum creatinine in their blood are more likely to die from heart failure as compared to those with lower levels of serum creatinine. This finding is not surprising, as higher levels of serum creatinine can indicate that a patient is suffering from poor renal function, and this is commonly seen in patients with poor cardiac health (Mullens, et. al.).

Additionally, we can see that the interquartile range (IQR) is greater on the boxplot for patients who died from heart failure as compared to the boxplot for those who survived. This indicates that there is greater variability in the level of serum creatinine for patients who died from heart failure.

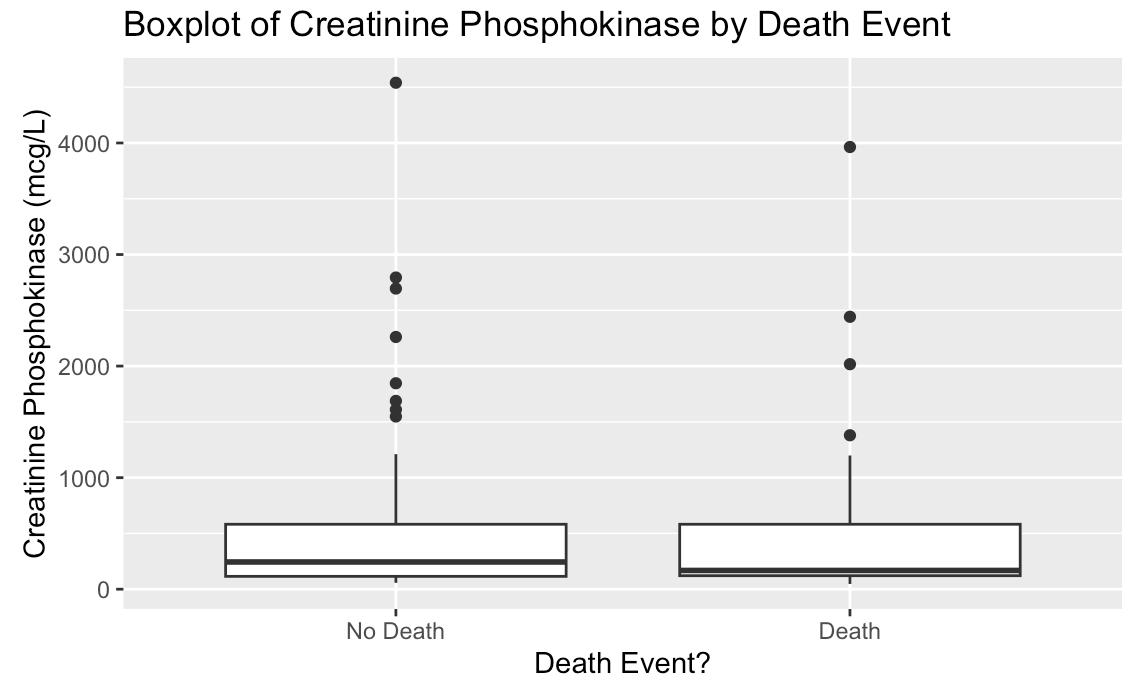
***Figure 4***



In this boxplot, we can see that patients who died from heart failure exhibit a higher median age as compared to patients who survived heart failure. This suggests that, on average, individuals who die from heart failure are older than those who do not. This is not surprising, as heart failure is the most common cause of hospitalization for people over 65 years old, as aging can weaken the heart muscles and disrupt their ability to function correctly (Medline Plus).

When we examine the IQR, we observe that the boxplot for patients who died also displays a larger IQR. This indicates more variability in ages for individuals who die from heart failure. On the other hand, the boxplot for patients who survived shows a narrower IQR, which implies that there is less age variability within this group. These insights from the boxplot provide valuable information about the distribution and variability of age in the context of death events, emphasizing differences in median age and IQR between the two groups.

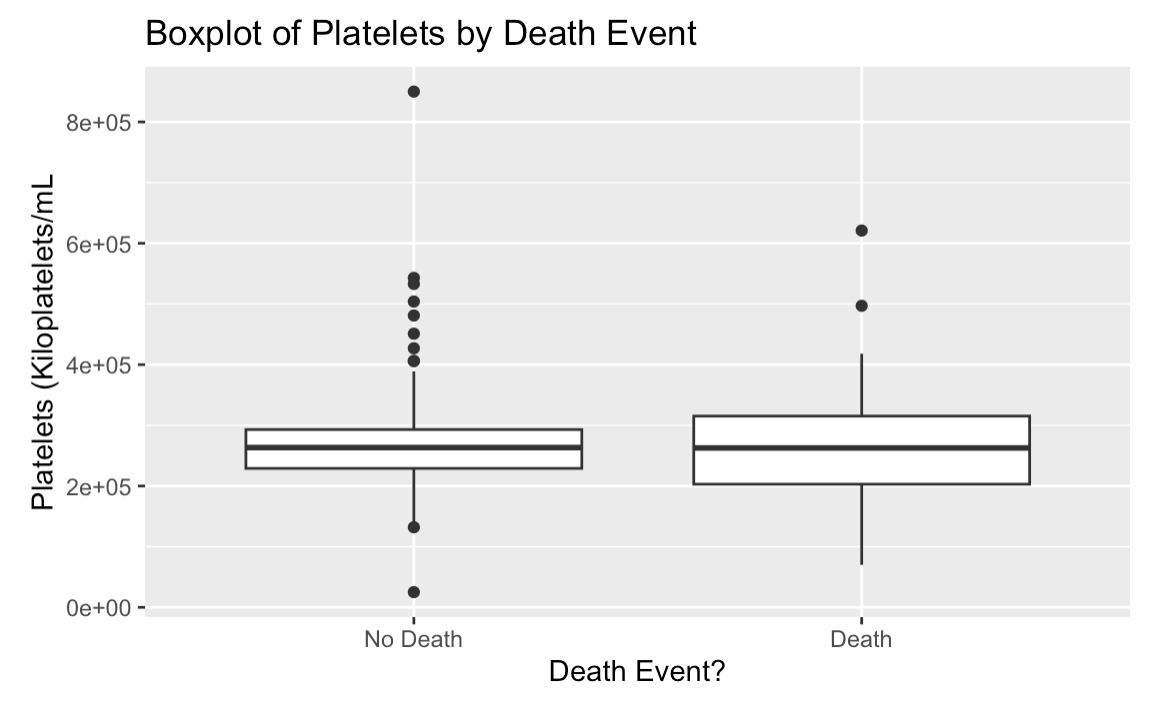
***Figure 4***



In this boxplot, we see that the median levels of creatinine phosphokinase are fairly similar, but the median is slightly higher for patients who survived heart failure. This indicates that it may be more likely for patients with lower levels of creatinine phosphokinase to survive, although the difference is so slight that we cannot draw any clear conclusions from this graph. Additionally, the interquartile ranges (IQR) are also very similar, meaning that there is not a large difference in the variability in the level of creatinine phosphokinase for patients who died as compared to those who survived.

This finding is surprising to use, as research shows that3(Chen). Thus, we would have expected there to see higher levels of this enzyme in patients who died from heart failure. We are hopeful that further investigation will shed light on why our data does not reflect this, although we recognize that it could be due to the random training/test split we conducted.

***Figure 6***

******

This boxplot demonstrates that the median level of platelets for patients is relatively similar, regardless of whether or not a death event occurred. This was surprising, as higher platelet levels are associated with thrombocytosis and thrombocythemia, which can cause blood clots and heart failure (National Heart, Lung, and Blood Institute). Therefore, we expected death events to have a higher median platelet level. Additionally, the IQR is much larger for those experiencing a death event. This indicates that there is more variability in platelet levels for those who died from heart failure. On the other hand, the cases with no death event have a smaller IQR, indicating less variability.

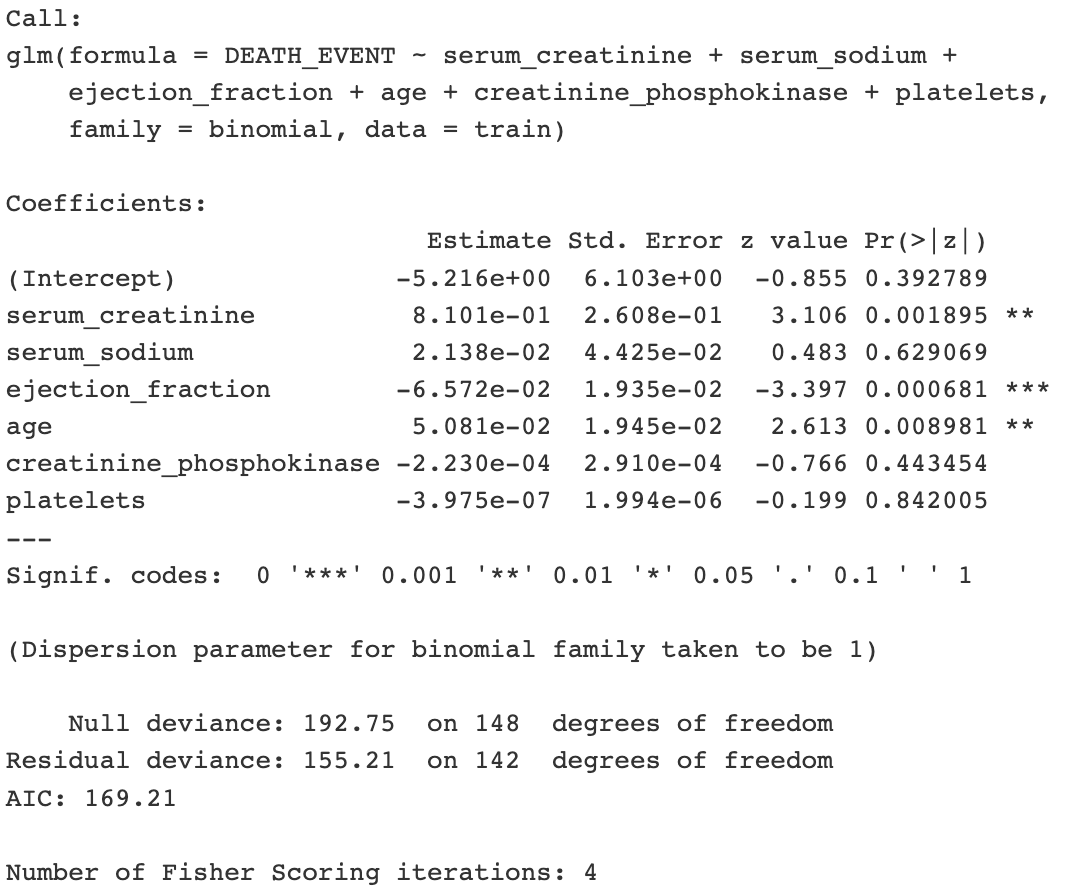
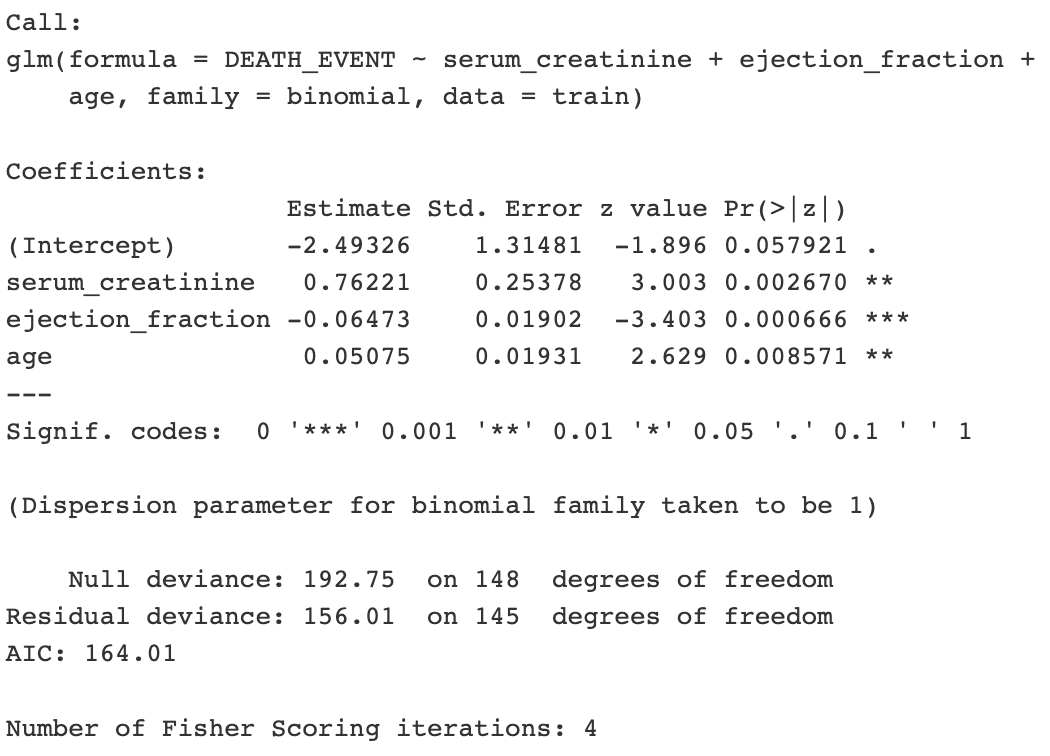
**6.2) Logistic Regression**

1. Included Predictors

|  |  |
| --- | --- |
| Age | The risk of heart disease increases with age, as the heart muscle can weaken and arteries can stiffen (U.S. Department). This may increase the likelihood of a death event. |
| Anaemia | Anaemia affects blood volume which disrupts the proper heart function and may lead to death. (Anemia) |
| Diabetes | Diabetes, and the high blood sugar that results, can damage the heart and blood vessels, which could lead to death. (Centers) |
| Ejection\_fraction | A lower ejection fraction can indicate that the heart is not effectively pumping blood and is potentially damaged, indicating a health issue ending in death. (Mayo) |
| High\_blood\_pressure | Can add stress to the heart and potentially lead to death or injury. (Mayo) |
| Platelets | Abnormal levels of platelets can indicate heart disease, which many people pass away from. (2022 AHA). |
| Serum\_creatinine | Can indicate impaired kidney function and lead to heart disease, which many people pass away from. (Chen) |
| Serum\_sodium | High serum sodium levels can cause an excess retention of water and put more pressure on the heart as it pumps blood. (Mayo) |
| Sex | Men tend to have a higher risk of heart disease, and thus muscle damage than women, which could lead to health issues. (Bots) |
| Smoking | Smoking can increase damage to the heart’s walls, increasing the chance of death. (US Department) |
| Creatinine\_Phosphokinase | High levels of the CPK enzyme can indicate that there are internal damages to a patient’s brain or muscle tissue, which could cause death from heart failure. (Chen) |

1. Summary Output

**Original model: Reduced/Improved model:**

**** 

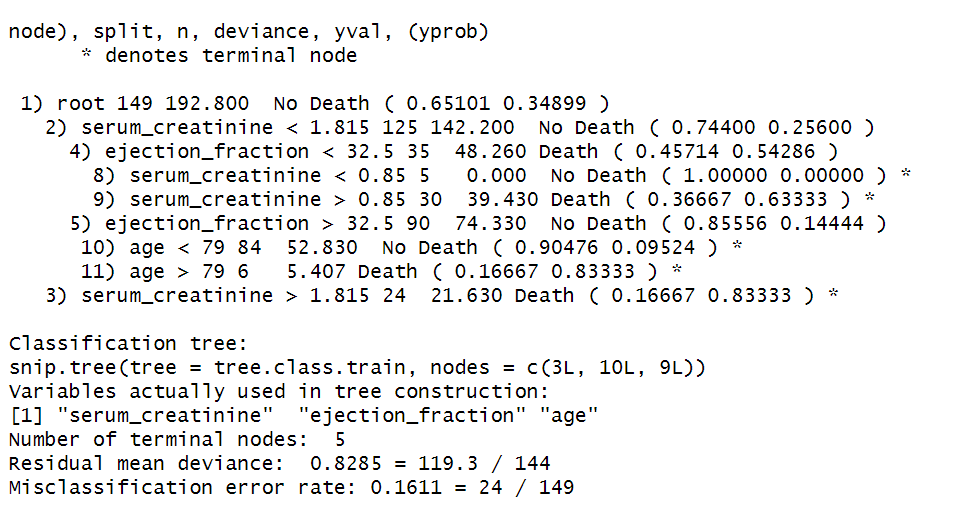
In both the original and the improved model, serum\_creatine, ejection\_fraction, and age were all found to be significant. In the original model, serum\_sodium, creatinine\_phosphokinase, and platelets were found to be insignificant, and after conducting a likelihood ratio test, we decided our model was better without them and removed them. We were surprised that neither gender nor smoking were considered significant when we had tried adding them.

**6.3) Classification Trees**

1. Motivation Behind Pruned Tree

The motivation behind the pruning of our original classification tree was to reduce the size of the decision tree. Pruning the tree removed the parts of the tree that did not provide significant classifying power. Pruning helped to reduce the presence of overfitting in our tree, improving the prediction accuracy.

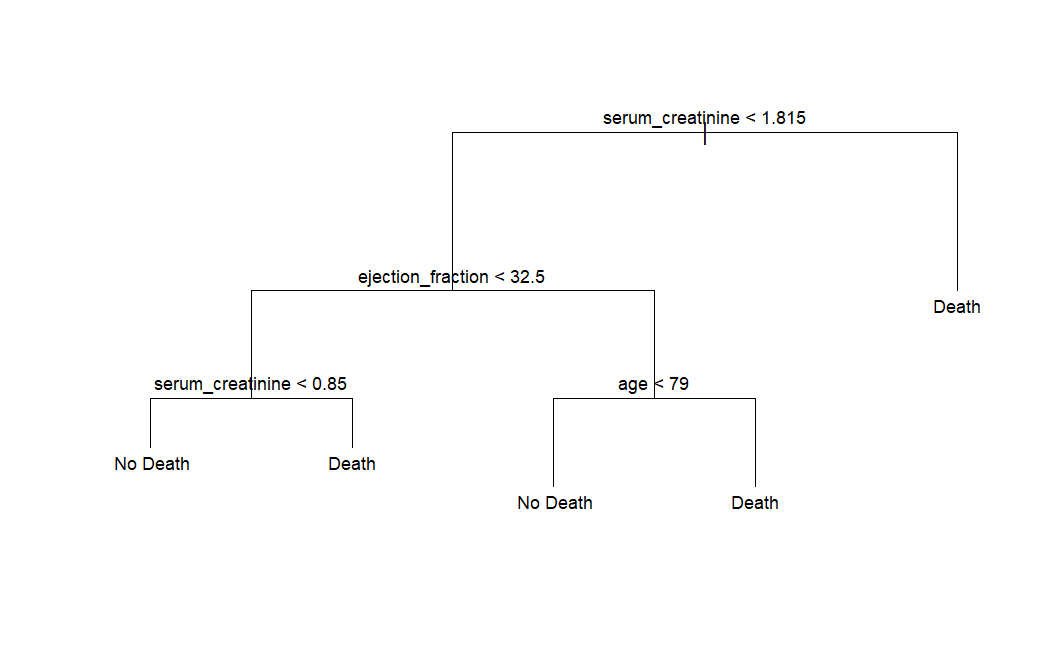
1. Summary Output

****

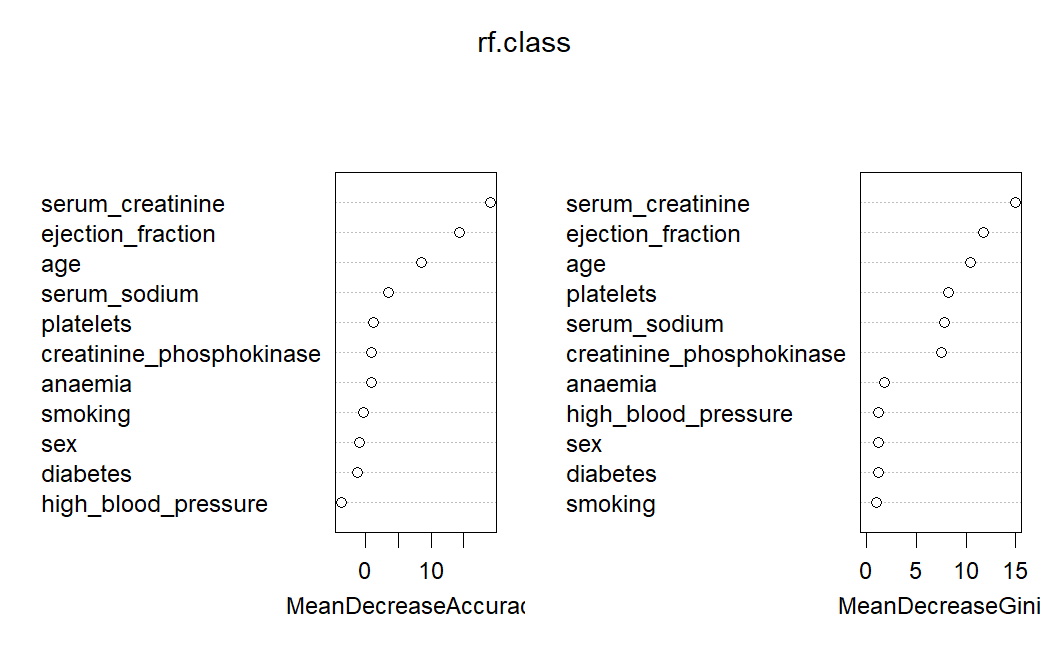
1. Terminal Nodes

This tree has 5 terminal nodes.

1. Graphical Output

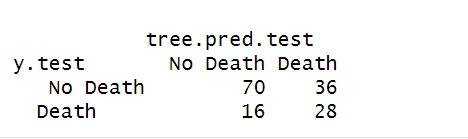


1. Random Forest Important Predictors



**6.4) Summary of Findings**

1. Confusion Matrices
   1. **Logistic Regression**

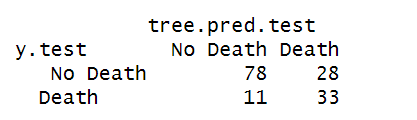
****

Overall Test Error Rate (Threshold=0.5): 0.3466667

False Positive Rate (Threshold=0.5): 0.3396226

False Negative Rate (Threshold=0.5): 0.3636364

* 1. **Pruned Classification Tree**

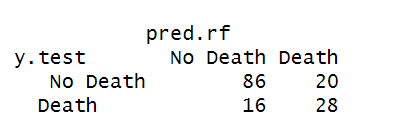


Overall Test Error Rate (Threshold=0.5): 0.26

False Positive Rate (Threshold=0.5): 0.2641509

False Negative Rate (Threshold=0.5): 0.25

* 1. **Random Forests**

****

Overall Test Error Rate (Threshold=0.5): 0.24

False Positive Rate (Threshold=0.5): 0.1886792

False Negative Rate (Threshold=0.5): 0.3636364

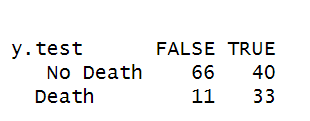
1. Table

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Recursive Binary Splitting** | **Pruned Classification Tree** | **Random Forest** |
| **Test Error Rate** | 0.3466667  *Threshold = 0.5* | 0.26  *Threshold = 0.5* | 0.24  *Threshold = 0.5* |
| **FPR** | 0.3396226  *Threshold = 0.5* | 0.2641509  *Threshold = 0.5* | 0.1886792  *Threshold = 0.5* |
| **FNR** | 0.3636364  *Threshold = 0.5* | 0.25  *Threshold = 0.5* | 0.3636364  *Threshold = 0.5* |

1. Threshold Discussion

After reviewing the confusion matrix, we conclude that our false negative rate is far too high. A false negative in a medical context is far more dangerous than a false positive. If a patient is incorrectly predicted to not experience a death event, they will likely miss out on life-saving care. On the other hand, if a patient is incorrectly predicted to die (false positive), there will be monetary loss from wasted medical treatments, but they will not lose their lives. Therefore, to improve our model, it is crucial to reduce the false negative rate (FNR).

1. Adjusted Threshold Confusion Matrices
   1. **Logistic Regression**

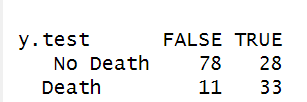
****

Overall Test Error Rate (Threshold=0.2): 0.34

False Positive Rate (Threshold=0.2): 0.37736

False Negative Rate (Threshold=0.2): 0.25

* 1. **Pruned Classification Tree**

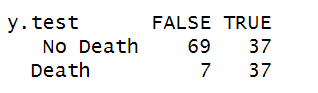
****

Overall Test Error Rate (Threshold=0.1): 0.26

False Positive Rate (Threshold=0.1): 0.2641509

False Negative Rate (Threshold=0.1): 0.25

* 1. **Random Forests (Threshold = 0.3)**

****

Overall Test Error Rate (Threshold=0.3): 0.29333

False Positive Rate (Threshold=0.3): 0.34906

False Negative Rate (Threshold=0.3): 0.15909

1. Table

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Recursive Binary Splitting** | **Pruned Classification Tree** | **Random Forest** |
| **Test Error Rate** | 0.34  *Threshold = 0.2* | 0.26  *Threshold = 0.1* | 0.29333  *Threshold = 0.3* |
| **FPR** | 0.37736  *Threshold = 0.2* | 0.2641509  *Threshold = 0.1* | 0.34906  *Threshold = 0.3* |
| **FNR** | 0.25  *Threshold = 0.2* | 0.25  *Threshold = 0.1* | 0.15909  *Threshold = 0.3* |

1. Question of Interest

The exploratory data analysis (EDA), logistic regression, and classification trees in subsections 6.1 to 6.3 aims to address the question of accurately predicting heart failure, with a specific emphasis on minimizing false positives and false negatives. The EDA explored significant relationships among predictors, such as serum sodium with platelets and ejection fraction, providing valuable insights into potential predictive power. Logistic regression identified age, serum creatinine, and ejection fraction as significant predictors, aligning with the known associations between these factors and heart failure outcomes. Pruning the classification tree aimed at reducing overfitting and enhancing predictive accuracy. The relation of predictors like age, serum creatinine, and ejection fraction to the response variable underscores their importance in predicting heart failure outcomes. Adjusting thresholds for Logistic Regression and Random Forests addressed the critical concern of minimizing false negatives in a medical context, ensuring timely interventions for individuals at the highest risk. This approach facilitates optimized resource allocation and improved patient outcomes, aligning with the goal of enhancing heart failure prediction models.

1. Discussion of Methods

The model comparisons reveal noteworthy insights, particularly regarding the trade-off between interpretability and predictive performance. The pruned classification tree, characterized by its simplicity and interpretability, stands out with the lowest test error rate (0.26), making it a valuable tool for understanding the underlying patterns in predicting heart failure. Its False Positive Rate (FPR) of 0.2641509 and False Negative Rate (FNR) of 0.25 further emphasize its effectiveness in achieving a fair prediction. On the other hand, the Random Forest model exhibits a distinct advantage in lowering the False Negative Rate (FNR), recording a notable value of 0.15909. This suggests that the Random Forest model excels in identifying true cases of heart failure, minimizing instances where it fails to recognize positive cases. This attribute makes Random Forest a compelling choice in scenarios where reducing the risk of missing actual cases holds significant importance, such as in healthcare applications.

However, this improvement in predictive performance comes at the expense of interpretability. Random Forest models, being ensembles of numerous decision trees, are inherently more complex and challenging to interpret compared to pruned classification trees. The aggregated decision-making process across multiple trees makes it difficult to pinpoint the exact features or rules responsible for a specific prediction.

**7. FURTHER WORK**

In order to improve the accuracy and usefulness of our models, there are a few additional actions we could take. Firstly, it would be helpful to rerun our analyses on more comprehensive datasets. The dataset we used considered only 299 cases of heart failure documented by the Faisalabad Institute of Cardiology in Pakistan in 2015. A more in-depth analysis would consider a larger sample size to provide more accurate results. Additionally, a more geographically diverse sample size would allow us to extrapolate our results to other regions of the world. For example, Russia, Hungary, and Romania suffer from the highest heart failure rates in the world due to factors like dietary and smoking habits (Heart). Therefore, the factors predicting heart failure in Pakistan are likely very different from those in these countries. As death rates often rely on national health regulations, this more holistic analysis could provide policymakers with better insight regarding which countries’ policies are most effective and should be mimicked.

Additionally, a further application would be the use of “panel” data – a dataset that follows the same individuals’ health characteristics over time. Our current dataset simply shows their characteristics at a single point in time. However, it is possible that identifying trends over time may provide better insight on the likelihood of a death event from heart failure. For example, perhaps a sharp increase in ejection fraction is a better predictor of death than a simple high measure at a point in time. Identifying such trends can allow doctors to track their patients’ health more effectively and identify problematic shifts. It would also decrease medical costs by identifying those most at risk of imminent heart failure, and allocating resources to those patients.

**References:**

*Anemia and heart failure: Guidance for clinicians and Trialists*. American College of Cardiology. (n.d.). https://www.acc.org/Latest-in-Cardiology/Articles/2021/07/06/12/18/Anemia-and-Heart-Failure#:~:text=Anemia%2C%20the%20reduction%20in%20the,key%20function%20of%20the%20circulation

Bots, S. H., Peters, S. A. E., & Woodward, M. (2017, March 1). *Sex differences in coronary heart disease and stroke mortality: A global assessment of the effect of ageing between 1980 and 2010*. BMJ Global Health. https://gh.bmj.com/content/2/2/e000298

Centers for Disease Control and Prevention. (2022, June 20). *Diabetes and Your Heart*. Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/library/features/diabetes-and-heart.html#:~:text=Over%20time%2C%20high%20blood%20sugar,and%20can%20damage%20artery%20walls

Chen, X., Jin, H., Wang, D., Liu, J., Qin, Y., Zhang, Y., Zhang, Y., & Xiang, Q. (2023, February 28). *Serum creatinine levels, traditional cardiovascular risk factors and 10-year cardiovascular risk in Chinese patients with hypertension*. Frontiers. https://www.frontiersin.org/articles/10.3389/fendo.2023.1140093/full#:~:text=The%20relationship%20between%20creatinine%20and,risk%20factors%20in%20hypertensive%20patients

Heart disease rates by country 2023. (n.d.). https://worldpopulationreview.com/country-rankings/heart-disease-rates-by-country

U.S. Department of Health and Human Services. (n.d.). *Heart Health and Aging*. National Institute on Aging. https://www.nia.nih.gov/health/heart-health-and-aging

2022 AHA/ACC/HFSA guideline for the management of heart failure ... (n.d.). https://www.ahajournals.org/doi/10.1161/CIR.0000000000001062