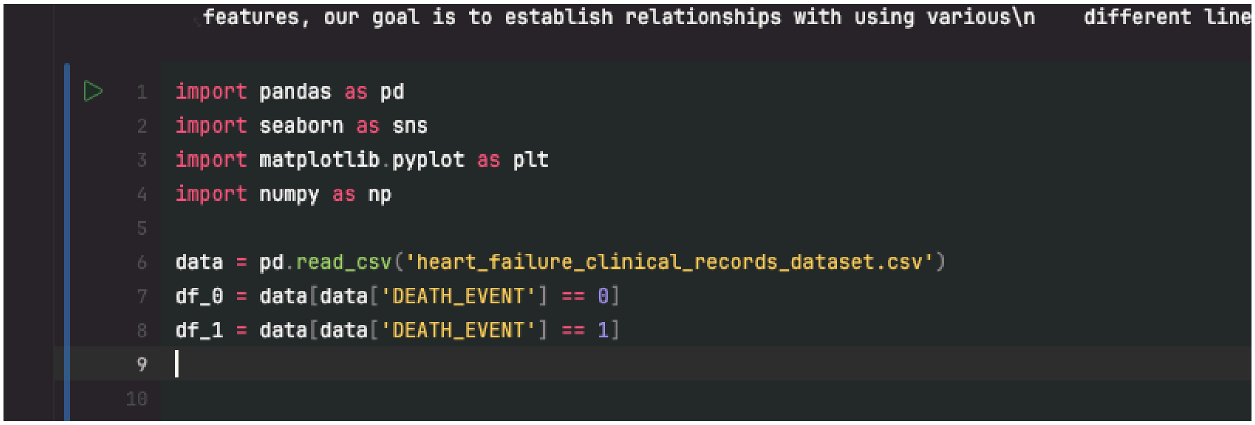
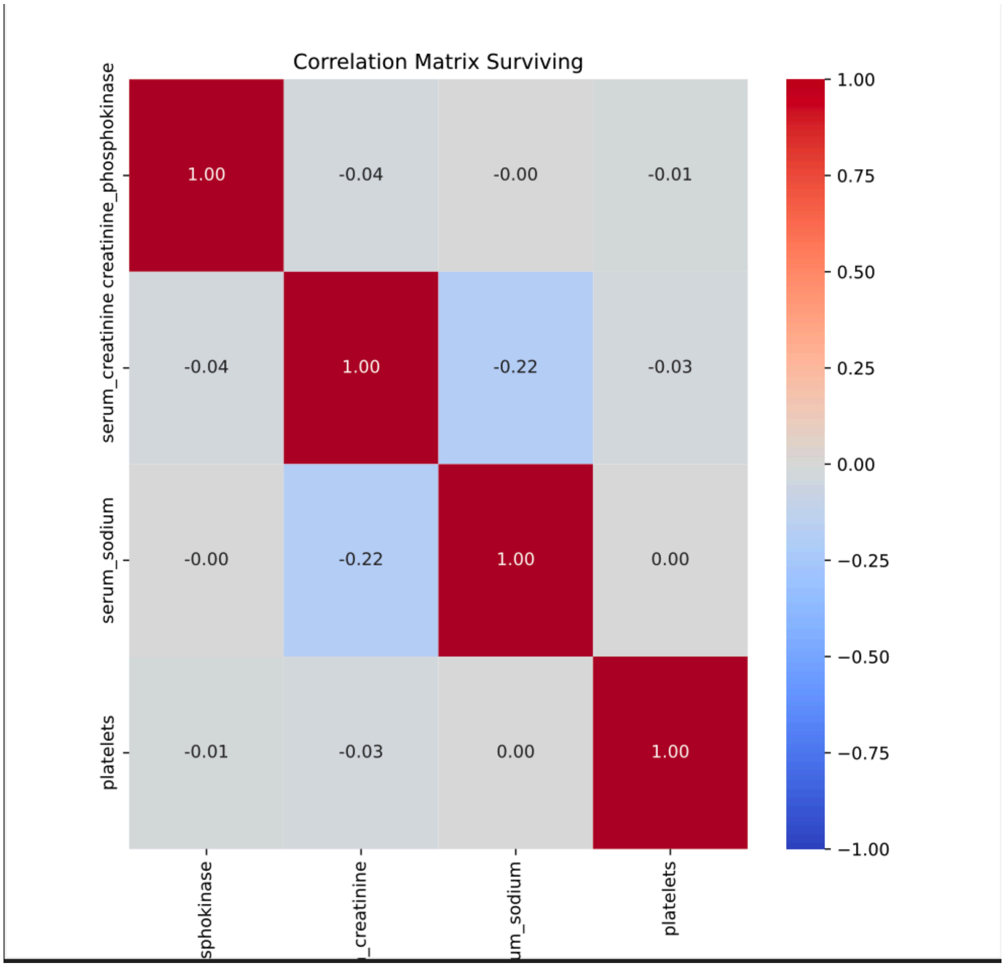
Ananya Singh
Assignment #4

Question #1 Part 1

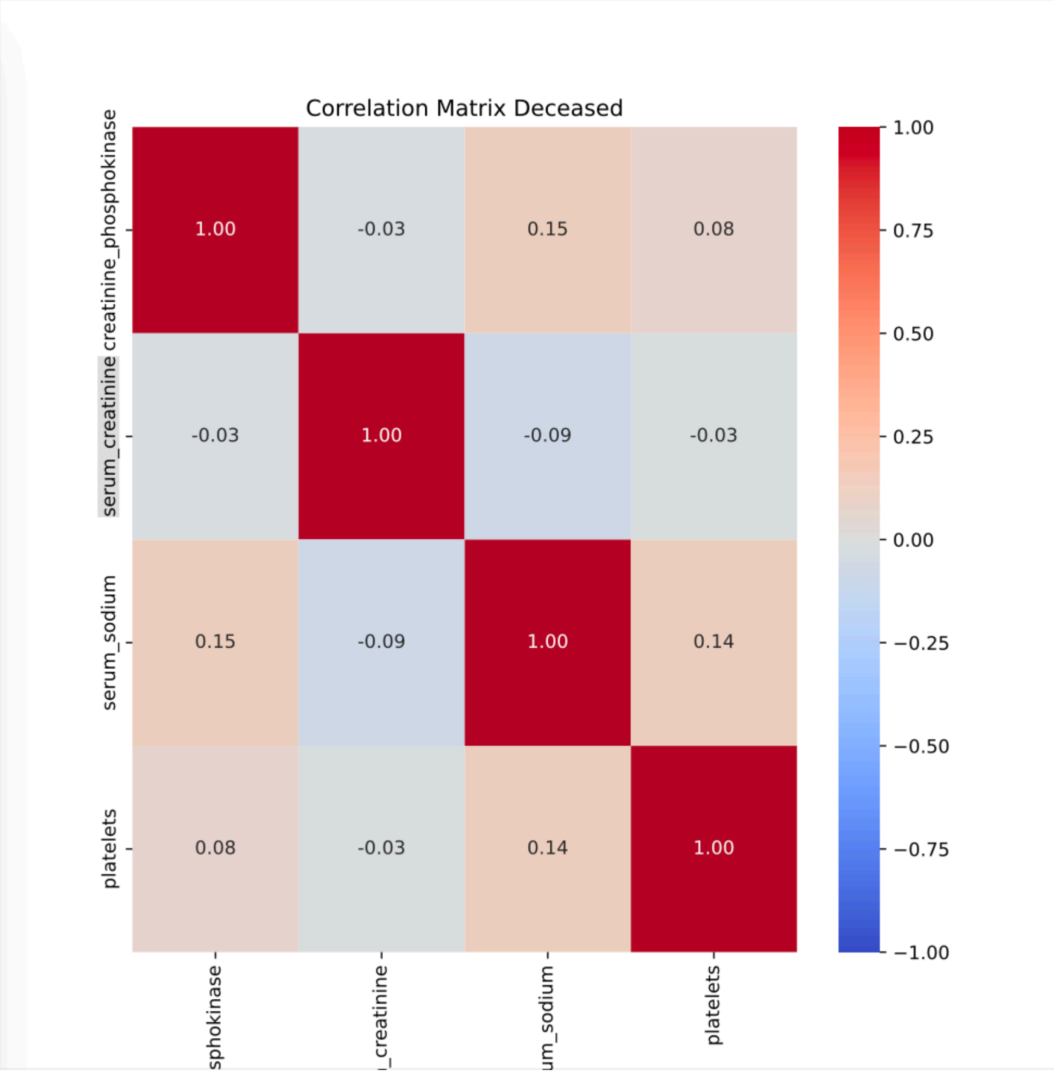


```
features, our goal is to establish relationships with using various\n    different line

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

data = pd.read_csv('heart_failure_clinical_records_dataset.csv')
df_0 = data[data['DEATH_EVENT'] == 0]
df_1 = data[data['DEATH_EVENT'] == 1]
```

Question #1 Part 2



0.75

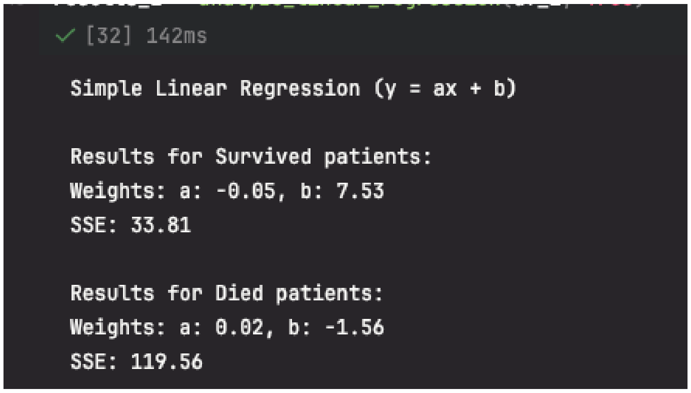


Question #1 Part 3

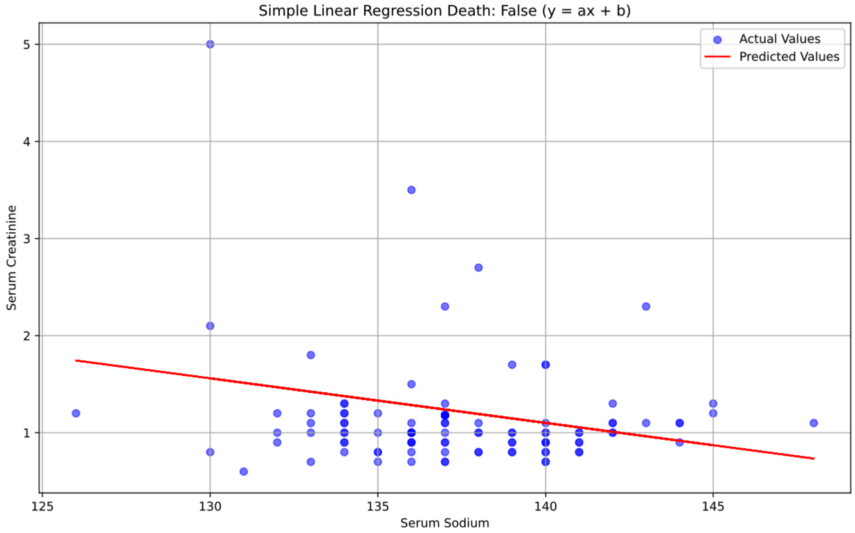
a. For surviving patients, we can see that all feature correlations are near zero therefore there is no high correlation.
b. The lowest correlation is between serum creatinine and serum sodium.
c. The highest correlation for deceased patients are between serum sodium/creatine phosphokinase and platelets/serum sodium
d. The lowest correlation is between serum creatinine and serum sodium.
e. No, the results are very different. Deceased patients correlations show stronger relationship between features whereas surviving patients correlations all have a weak linear relationship.

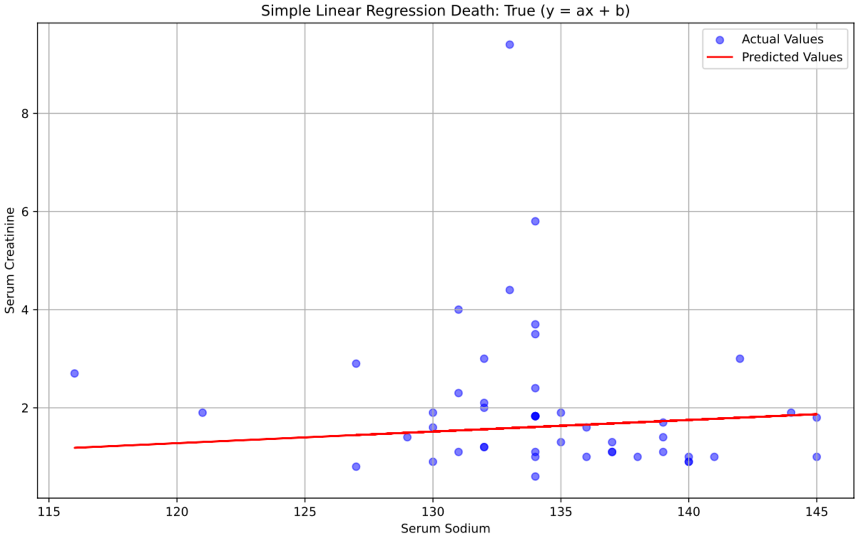
Question #2 I am part of Group #3 (X: serum sodium, Y : serum creatinine)

1. y= ax + b (simple linear regression)

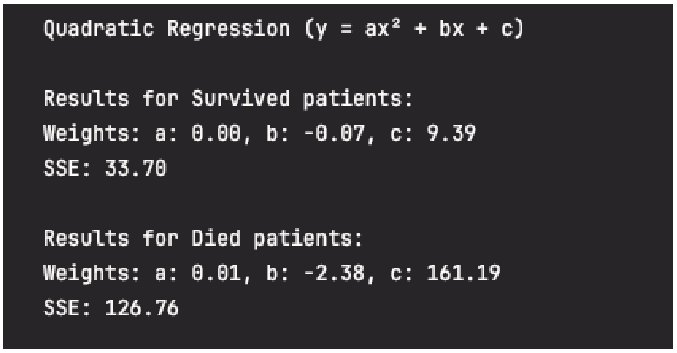


```
✓ [32] 142ms

Simple Linear Regression (y = ax + b)

Results for Survived patients:
Weights: a: -0.05, b: 7.53
SSE: 33.81

Results for Died patients:
Weights: a: 0.02, b: -1.56
SSE: 119.56
```

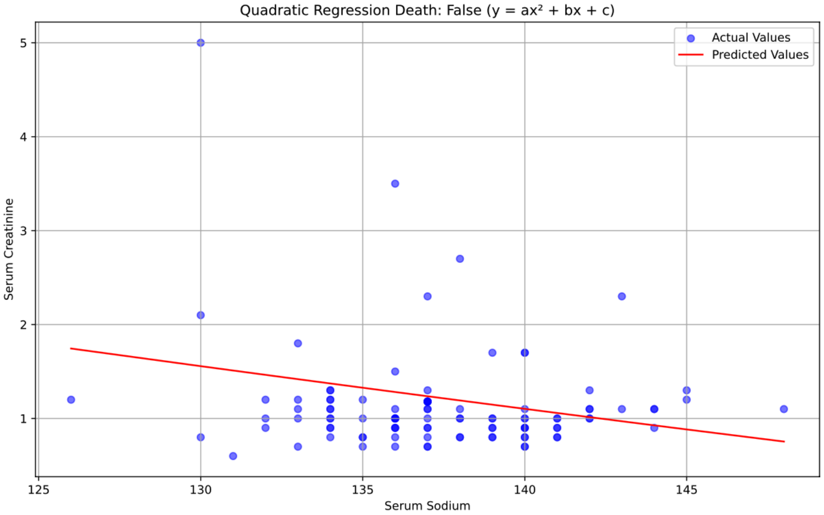


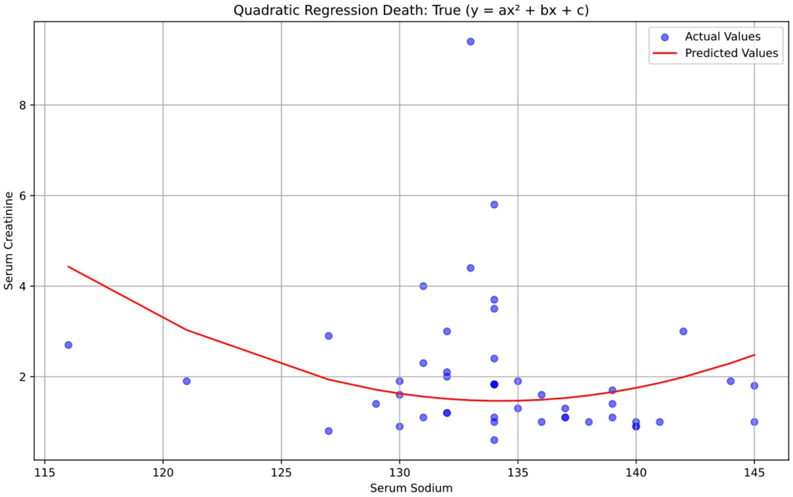


2. y= ax^2 + bx + c (quadratic)

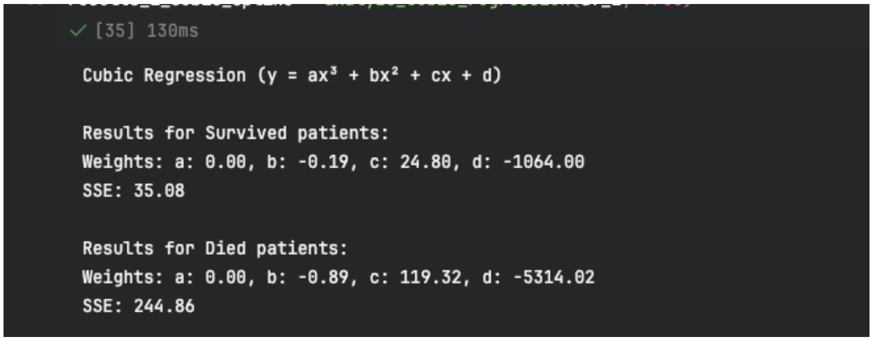


```
Quadratic Regression (y = ax² + bx + c)

Results for Survived patients:
Weights: a: 0.00, b: -0.07, c: 9.39
SSE: 33.70

Results for Died patients:
Weights: a: 0.01, b: -2.38, c: 161.19
SSE: 126.76
```

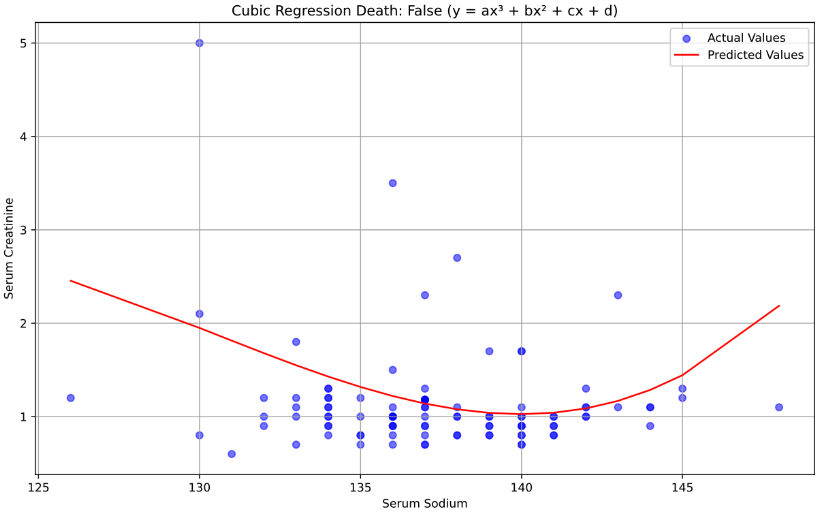


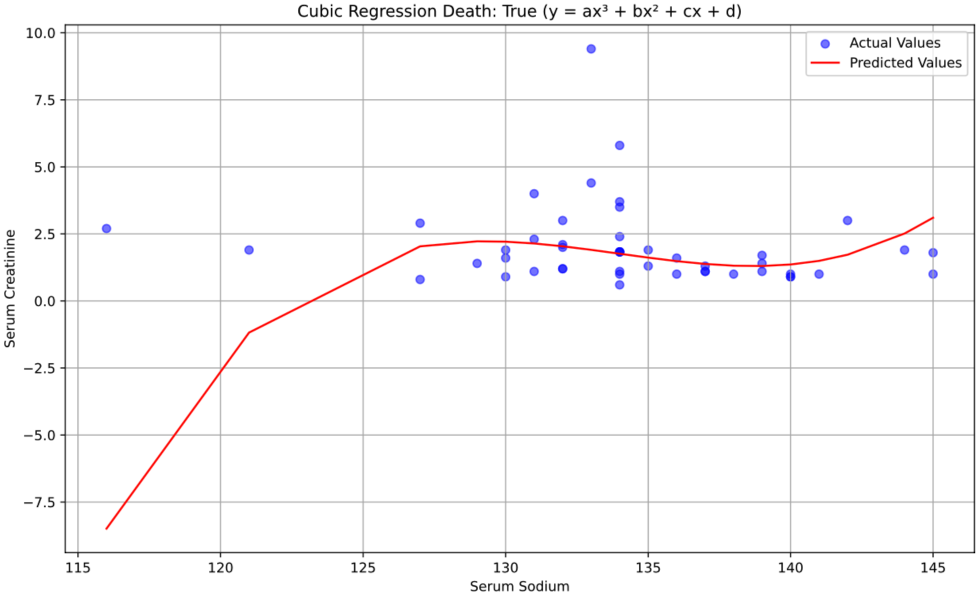


3. y= ax^3 + bx^2 + cx + d (cubic spline)

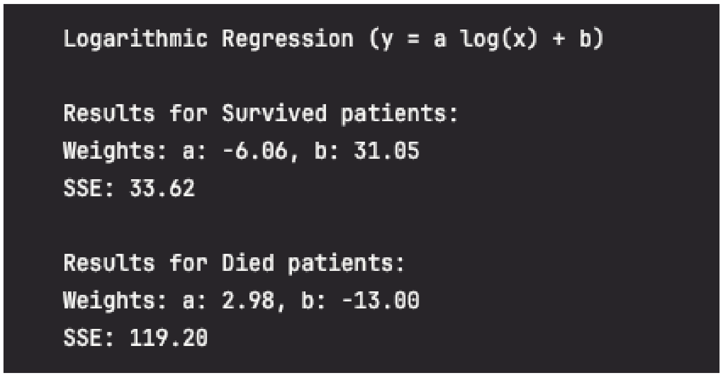


```
✓ [35] 130ms

Cubic Regression (y = ax³ + bx² + cx + d)

Results for Survived patients:
Weights: a: 0.00, b: -0.19, c: 24.80, d: -1064.00
SSE: 35.08

Results for Died patients:
Weights: a: 0.00, b: -0.89, c: 119.32, d: -5314.02
SSE: 244.86
```

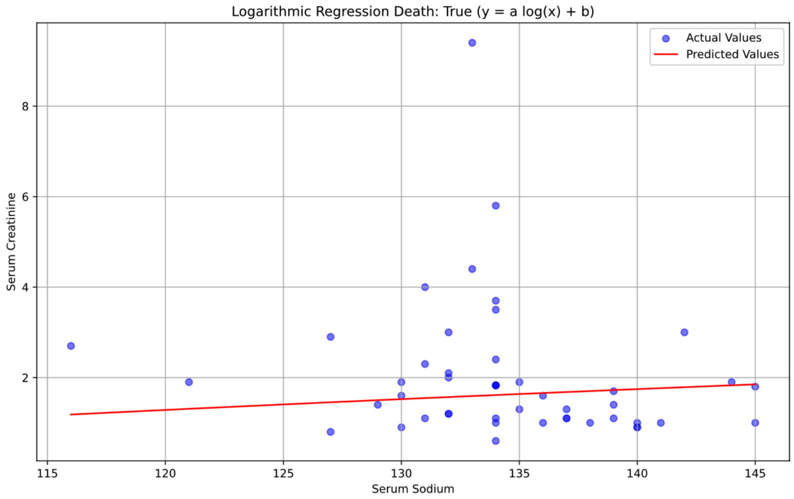




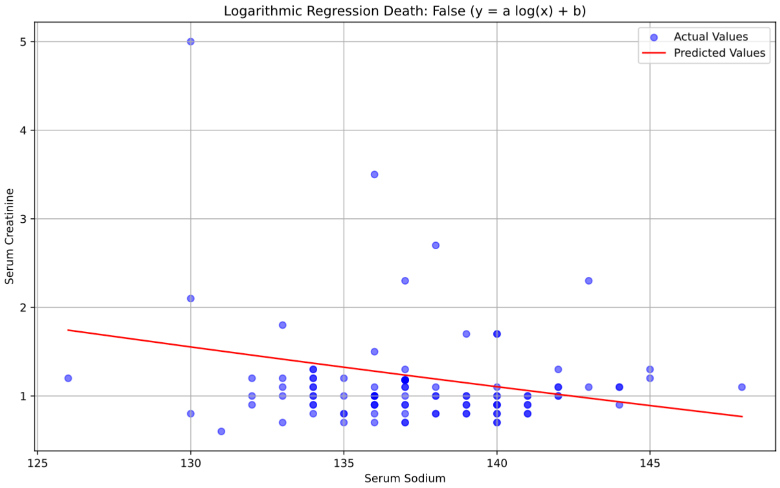
4. y= a log x + b (GLM - generalized linear model)



```
Logarithmic Regression (y = a log(x) + b)

Results for Survived patients:
Weights: a: -6.06, b: 31.05
SSE: 33.62

Results for Died patients:
Weights: a: 2.98, b: -13.00
SSE: 119.20
```

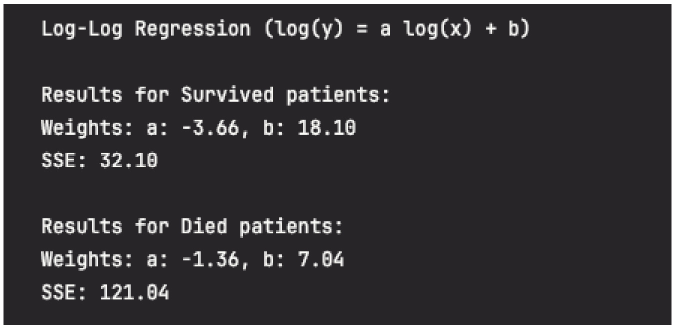


Logarithmic Regression Death: True (y = a log(x) + b)

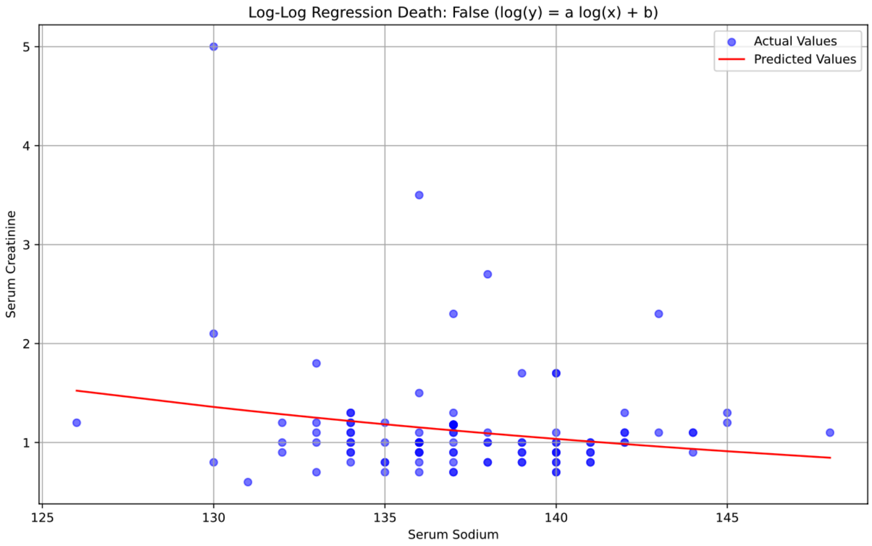


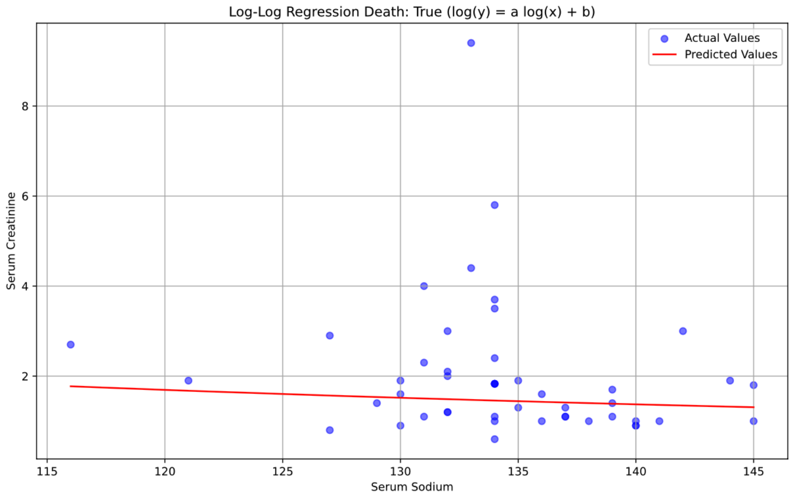
Logarithmic Regression Death: False (y = a log(x) + b)

5. log y= a log x + b (GLM - generalized linear model)

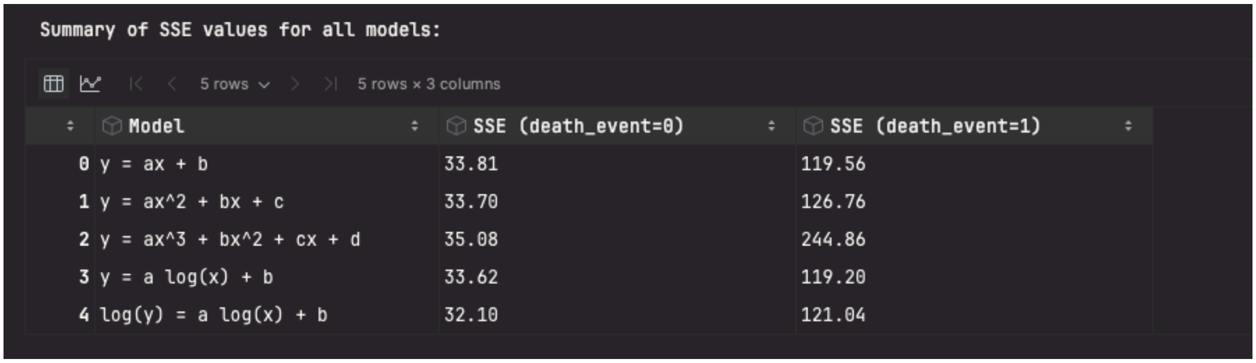


```
Log-Log Regression (log(y) = a log(x) + b)

Results for Survived patients:
Weights: a: -3.66, b: 18.10
SSE: 32.10

Results for Died patients:
Weights: a: -1.36, b: 7.04
SSE: 121.04
```





Question #5 Part 1



Summary of SSE values for all models:

	Model	SSE (death_event=0)	SSE (death_event=1)
0	y = ax + b	33.81	119.56
1	y = ax^2 + bx + c	33.70	126.76
2	y = ax^3 + bx^2 + cx + d	35.08	244.86
3	y = a log(x) + b	33.62	119.20
4	log(y) = a log(x) + b	32.10	121.04

Question #5 Part 2
surviving patients: the best model is log y= a log x + b (GLM - generalized linear model) with SSE around 32.10.
dead patients: best model is y= a log x + b (GLM - generalized linear model) with SSE around 119.20

Question #5 Part 3
surviving patients: the worst model is y= ax^3 + bx^2 + cx + d (cublc spline) with SSE around 35.08.
dead patients: worst model is y= ax^3 + bx^2 + cx + d (cublc spline) with SSE around 244.86.