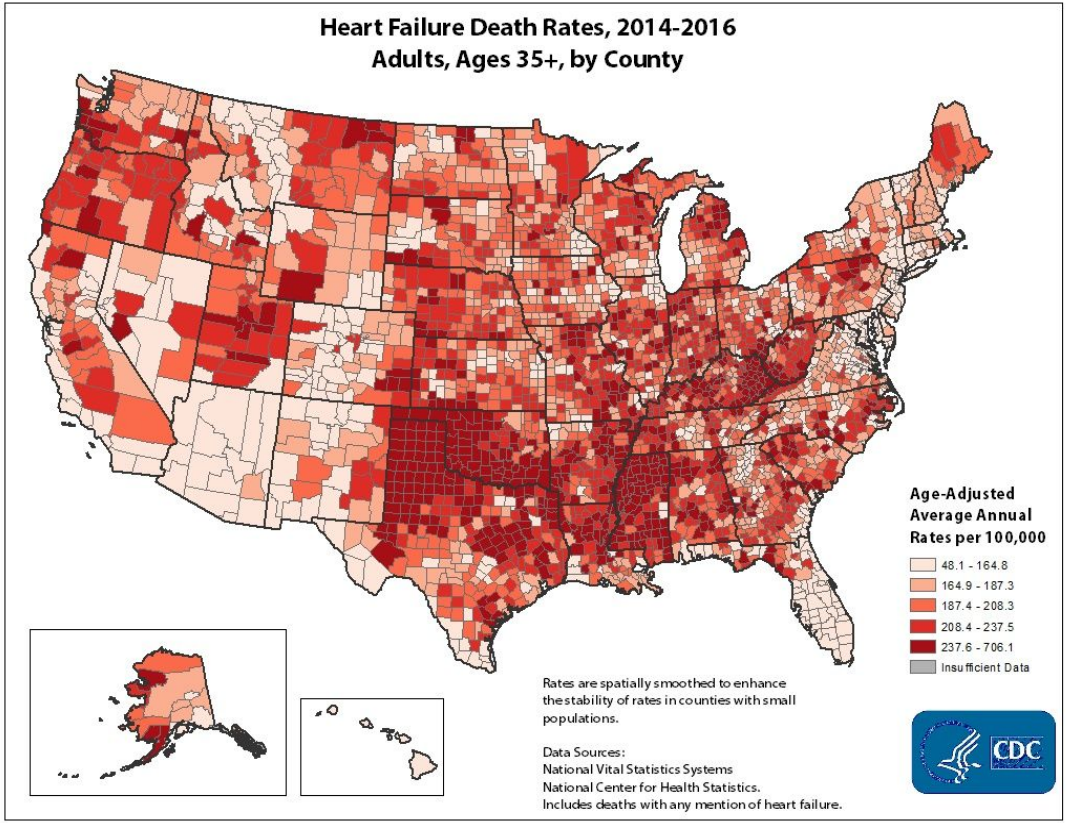
Heart Failure

Mark Perez

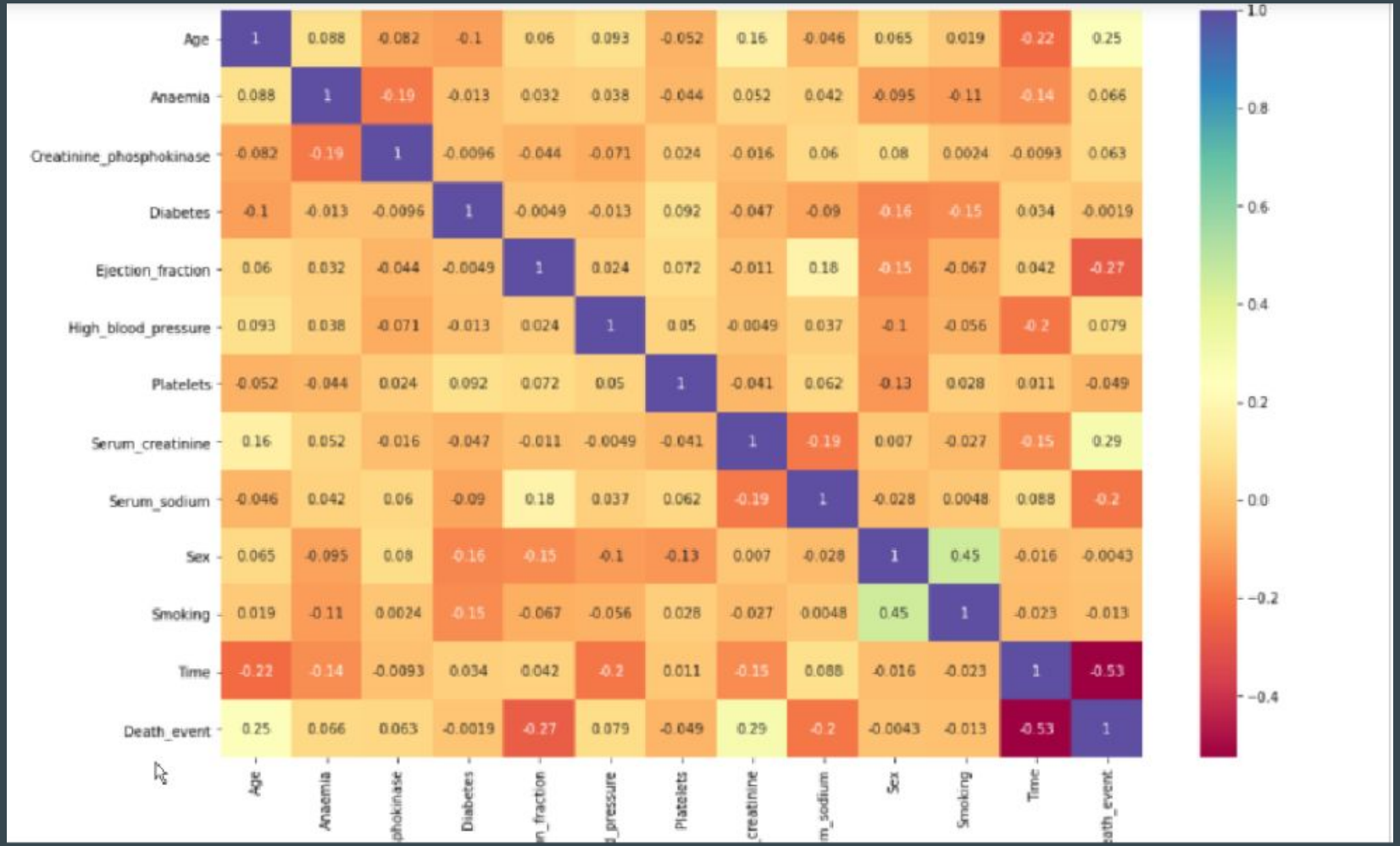
Heart Failure greatly affects us

About 6.2 million in United States.



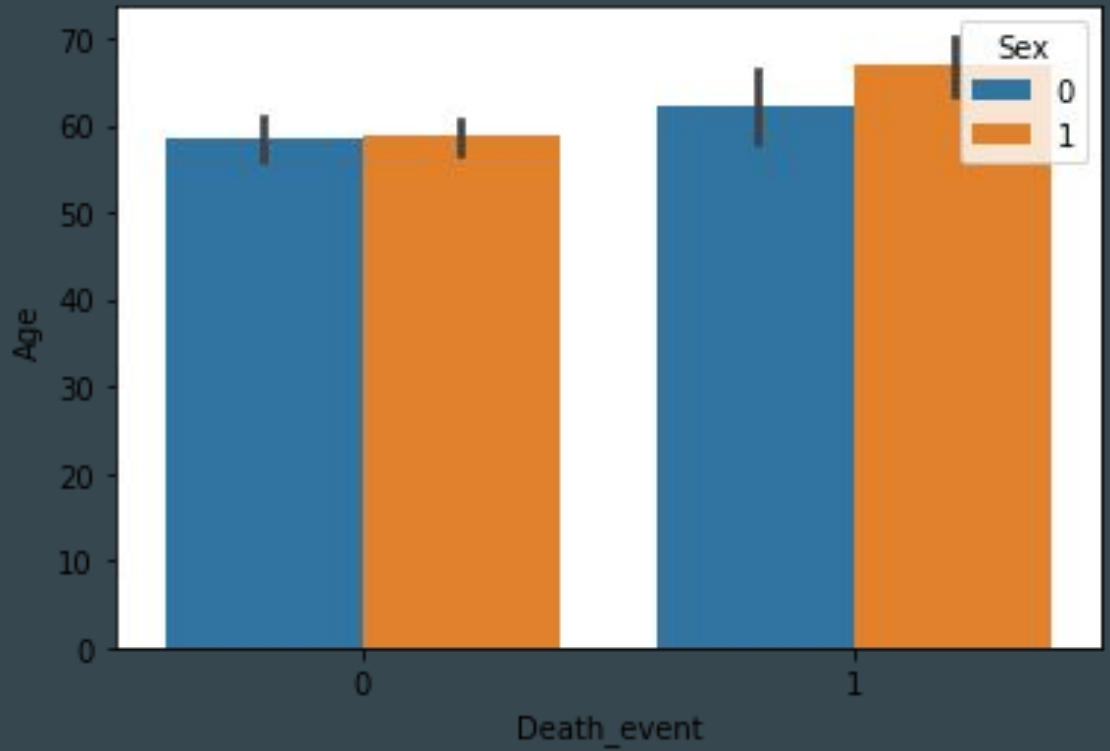
Can we predict what causes heart failure?

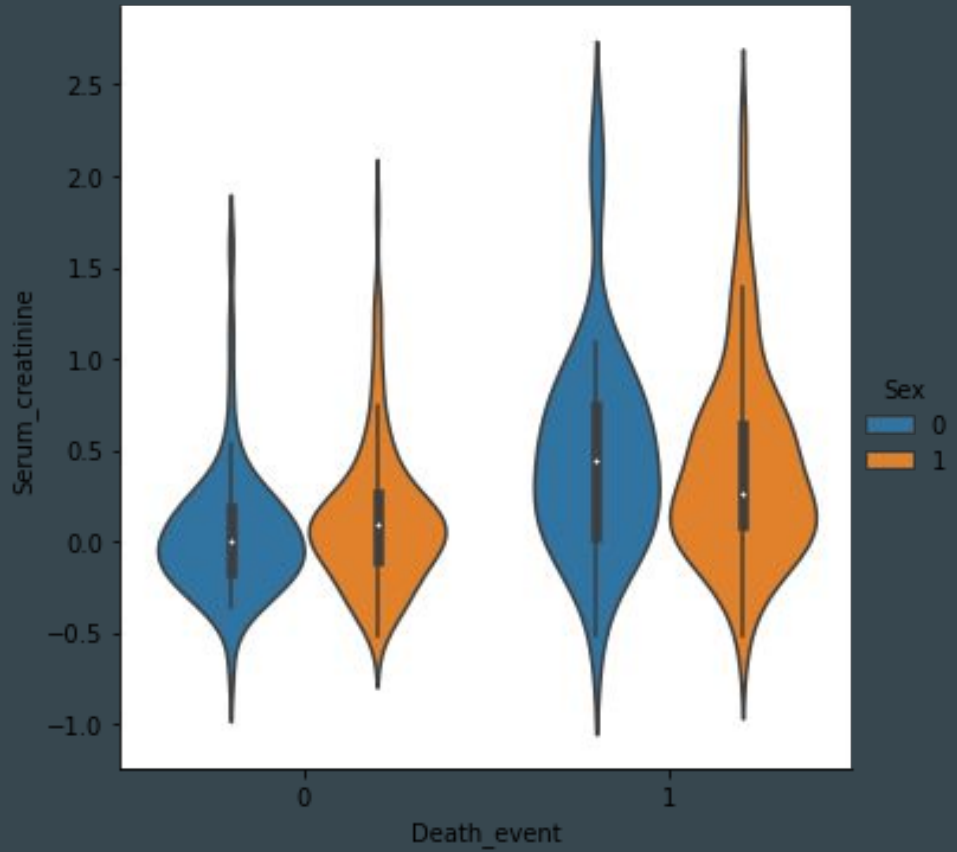
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 299 entries, 0 to 298
Data columns (total 13 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   age                       299 non-null    float64
 1   anaemia                   299 non-null    int64  
 2   creatinine_phosphokinase  299 non-null    int64  
 3   diabetes                  299 non-null    int64  
 4   ejection_fraction         299 non-null    int64  
 5   high_blood_pressure       299 non-null    int64  
 6   platelets                 299 non-null    float64
 7   serum_creatinine          299 non-null    float64
 8   serum_sodium              299 non-null    int64  
 9   sex                       299 non-null    int64  
 10  smoking                   299 non-null    int64  
 11  time                      299 non-null    int64  
 12  DEATH_EVENT               299 non-null    int64  
dtypes: float64(3), int64(10)
memory usage: 30.5 KB
```

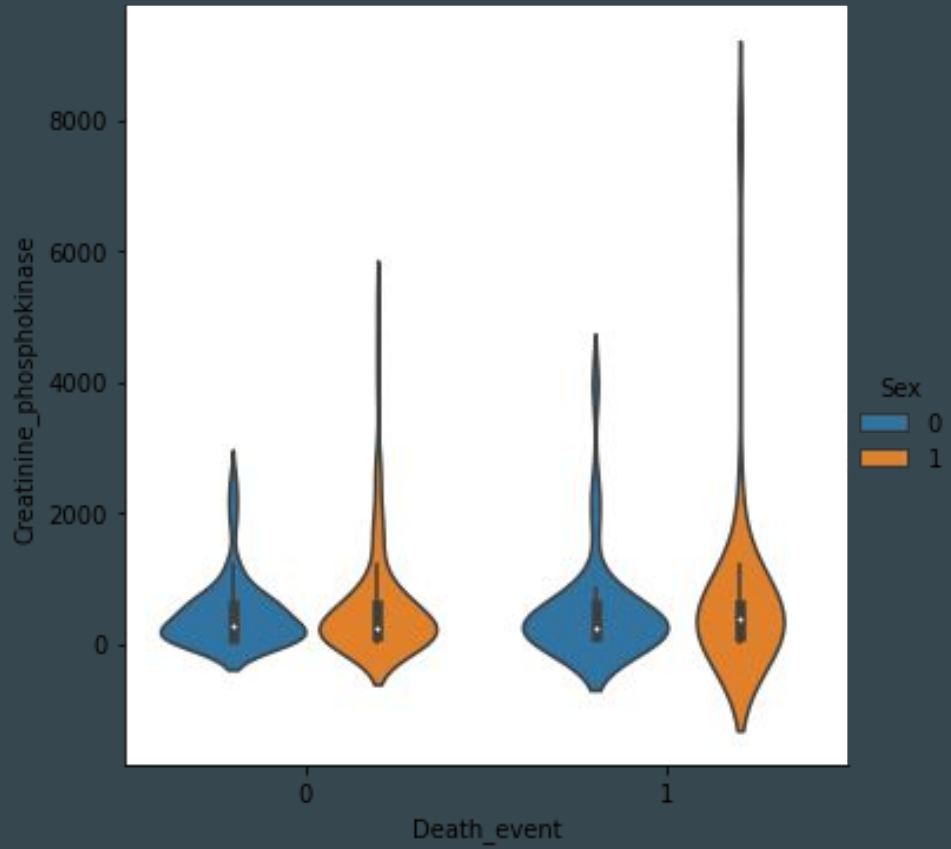

Age 1 0.088 -0.082 -0.1 0.06 0.093 -0.052 0.16 -0.046 0.065 0.019 -0.22 0.25
Anaemia 0.088 1 -0.19 -0.013 0.032 0.038 -0.044 0.052 0.042 -0.095 -0.11 -0.14 0.066
Creatinine_phosphokinase -0.082 -0.19 1 -0.0096 -0.044 -0.071 0.024 -0.016 0.06 0.08 0.0024 -0.0093 0.063
Diabetes -0.1 -0.013 -0.0096 1 -0.0049 -0.013 0.092 -0.047 -0.09 -0.16 -0.15 0.034 -0.0019
Ejection_fraction 0.06 0.032 -0.044 -0.0049 1 0.024 0.072 -0.011 0.18 -0.15 -0.067 0.042 -0.27
High_blood_pressure 0.093 0.038 -0.071 -0.013 0.024 1 0.05 -0.0049 0.037 -0.1 -0.056 -0.2 0.079
Platelets -0.052 -0.044 0.024 0.092 0.072 0.05 1 -0.041 0.062 -0.13 0.028 0.011 -0.049
Serum_creatinine 0.16 0.052 -0.016 -0.047 -0.011 -0.0049 -0.041 1 -0.19 0.007 -0.027 -0.15 0.29
Serum_sodium -0.046 0.042 0.06 -0.09 0.18 0.037 0.062 -0.19 1 -0.028 0.0048 0.088 -0.2
Sex 0.065 -0.095 0.08 -0.16 -0.15 -0.1 -0.13 0.007 -0.028 1 0.45 -0.016 -0.0043
Smoking 0.019 -0.11 0.0024 -0.15 -0.067 -0.056 0.028 -0.027 0.0048 0.45 1 -0.023 -0.013
Time -0.22 -0.14 -0.0093 0.034 0.042 -0.2 0.011 -0.15 0.088 -0.016 -0.023 1 -0.53
Death_event 0.25 0.066 0.063 -0.0019 -0.27 0.079 -0.049 0.29 -0.2 -0.0043 -0.013 -0.53 1
Age
Anaemia
Diabetes
Platelets
Sex
Smoking
Time
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Interesting correlations

- Death_event - Serum_creatinine
- Death_event - Age
- Death_event - high_blood_pressure


Sex
0
1
70
60
50
40
30
20
10
0
Age
0
1
Death_event


2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
Serum_creatinine
0
1
Death_event
Sex
0
1


8000
6000
4000
2000
0
Creatinine_phosphokinase
0
1
Death_event
Sex
0
1

How accurately can we predict?

Random Forest

```
In [31]: from sklearn.metrics import classification_report
         print(classification_report(y_pred,y_test))
```

```
              precision    recall  f1-score   support

           0       0.94      0.72      0.82        69
           1       0.49      0.86      0.62        21

    accuracy                           0.76        90
   macro avg       0.71      0.79      0.72        90
weighted avg       0.84      0.76      0.77        90
```

Linear Regression

	precision	recall	f1-score	support
0	0.94	0.76	0.84	66
1	0.57	0.88	0.69	24
accuracy			0.79	90
macro avg	0.76	0.82	0.76	90
weighted avg	0.84	0.79	0.80	90

XGBClassifier

	precision	recall	f1-score	support
0	0.89	0.77	0.82	61
1	0.62	0.79	0.70	29
accuracy			0.78	90
macro avg	0.75	0.78	0.76	90
weighted avg	0.80	0.78	0.78	90

In [46]:
```
for feature_list_index in sfm.get_support(indices=True):
    print(feat_labels[feature_list_index])
```

```
Serum_creatinine
Time
```

In [48]:
```
for feature_list_index in sfm.get_support(indices=True):
    print(feat_labels[feature_list_index])
```

```
Time
```

Further Analysis

- Accumulate more data
- Explore other models
- Explore more hyperparameters