

# Heart Failure Prediction

## PREDICTIVE MODELING REPORT

*Submitted by*

**UJJVAL PATEL    JINAY SHAH    PURAV SHAH**

**[20162121021]    [20162121025]    [20162121026]**

*Guided by*

**PROF. SHEETAL MAKHIJA**

*In partial fulfillment for the award of the degree of*

**BACHELOR OF TECHNOLOGY**

*In*

**Computer Science Engineering Department**

**Institute of Computer Technology**



**Ganpat University, Mehsana**

**[FEB 2023]**

## Business Understanding

- Heart failure prediction models are a type of healthcare analytics that can be used to identify patients who are at high risk of developing heart failure. This type of model can be used to improve patient outcomes and reduce healthcare costs by enabling healthcare providers to intervene early and provide proactive treatment.
- Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide.
- Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality by heart failure.
- Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies.
- People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

## Data Description Report

### 1. Data Quantity

- **What is the format of data?**

Dataset given is in CSV format which is known as comma separated values. It is a text file in which there are records and in between the records there is a comma which separates the values column wise. Each line of the file is a data record. Each record consists of one or more fields, separated by commas.

- **Identify the method used to capture the data?**

Not Mentioned

- **How large is the dataset?**

Number of Fields = 13

Number of Records = 299

## 2. Data Quality

- Does the data include the characteristics relevant to the business question?

Yes, the data include the characteristics relevant to the business question which is related to the prediction of heart failure

- What data types are present?

The Measurement levels are continuous and the data types present are Double and long.

**Type**

Settings

Read values to instantiate data ⓘ

[Read values](#) [Clear all values](#) [Clear values](#)











Find in column Field

<input type="checkbox"/>	Field	Measure ⓘ	Role ⓘ	Value mode ⓘ	Values
<input type="checkbox"/>	# age	Continuous	Input	Instantiated	40.0, 95.0
<input type="checkbox"/>	# anaemia	Continuous	Input	Instantiated	0, 1
<input type="checkbox"/>	# creatinine_phosph	Continuous	Input	Instantiated	23, 7861
<input type="checkbox"/>	# diabetes	Continuous	Input	Instantiated	0, 1
<input type="checkbox"/>	# ejection_fraction	Continuous	Input	Instantiated	14, 80
<input type="checkbox"/>	# high_blood_press	Continuous	Input	Instantiated	0, 1
<input type="checkbox"/>	# platelets	Continuous	Input	Instantiated	25100.0, 850000.0
<input type="checkbox"/>	# serum_creatinine	Continuous	Input	Instantiated	0.5, 9.4

Default mode ⓘ

[Cancel](#) [Save](#)









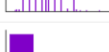



## View Output: Data Audit

Audit		Quality		Search Columns	
Name	Graph	Measurement	Type		
age		 Continuous	Double		
anaemia		 Continuous	Long		
creatinine_phosphokinase		 Continuous	Long		
diabetes		 Continuous	Long		
ejection_fraction		 Continuous	Long		

- Did you compute basic statistics for the key attributes? What insight did this provide into the business question?



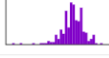



View Output: Data Audit

Compare

Audit		Quality		Search Columns						
Name	Graph	Measurement	Type	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
age		 Continuous	Double	40.000	95.000	60.834	11.895	0.419	0	299
anaemia		 Continuous	Long	0	1	0.431	0.496	0.275	0	299
creatinine_phosphokinase		 Continuous	Long	23	7,861	581.839	970.288	4.418	0	299
diabetes		 Continuous	Long	0	1	0.418	0.494	0.331	0	299
ejection_fraction		 Continuous	Long	14	80	38.084	11.835	0.550	0	299
high_blood_pressure		 Continuous	Long	0	1	0.351	0.478	0.620	0	299







View Output: Data Audit

Compare

Audit	Quality	Search Columns								
Name	Graph	Measurement	Type	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
platelets		Continuous	Double	25,100.000	850,000.000	263358.029	97804.237	1.448	0	299
serum_creatinine		Continuous	Double	0.500	9.400	1.394	1.035	4.411	0	299
serum_sodium		Continuous	Long	113	148	136.625	4.412	-1.038	0	299
sex		Continuous	Long	0	1	0.649	0.478	-0.620	0	299
smoking		Continuous	Long	0	1	0.321	0.468	0.763	0	299
time		Continuous	Long	4	285	130.261	77.614	0.127	0	299

View Output: Data Audit

Compare

Audit	Quality	Search Columns								
Name	Graph	Measurement	Type	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
serum_creatinine		Continuous	Double	0.500	9.400	1.394	1.035	4.411	0	299
serum_sodium		Continuous	Long	113	148	136.625	4.412	-1.038	0	299
sex		Continuous	Long	0	1	0.649	0.478	-0.620	0	299
smoking		Continuous	Long	0	1	0.321	0.468	0.763	0	299
time		Continuous	Long	4	285	130.261	77.614	0.127	0	299
DEATH_EVENT		Continuous	Long	0	1	0.321	0.468	0.763	0	299

1. Age: age of the patient [years]
2. Anaemia: [1: anaemia, 0:Normal]
3. Creatinine\_phosphokinase: Creatinine phosphokinase of the patient
4. Diabetes: [1: diabetes, 0:Normal]
5. Ejection\_fraction: Ejection fraction of the patient in percentage
6. High\_blood\_pressure: [1:high\_blood\_pressure, 0:Normal]
7. Platelets: Platelets of the patient
8. Serum\_creatinine: Serum creatinine of the patient
9. Serum\_sodium: Serum sodium of the patient
- 10.Sex: sex of the patient [1: Male, 0: Female]
- 11.Smoking: [1: Smoking, 0:Normal]
- 12.time:
- 13.DEATH\_EVENT: Output Class or Target [1:True, 0:False]

- **Are you able to prioritise relevant attributes? If not, are business analysts available to provide further insight?**

1. Relevant Attributes

- Age
- Anaemia
- Creatinine\_Phosphokinase
- Diabetes
- Ejection\_fraction
- High\_blood\_pressure
- Platelets
- Serum\_creatinine
- Serum\_sodium
- Smoking

2. Non Relevant Attributes

- Sex
- Time

## Data Exploration Report

### 1. What sort of hypotheses have you formed about the data?

A relational hypothesis is one that suggests variables are related in some way. So, in our case the death\_event is related to the values from various fields like high blood pressure, diabetes, smoking, anaemia, creatinine phosphokinase, ejection fraction, high platelets count, serum creatinine, low serum sodium.

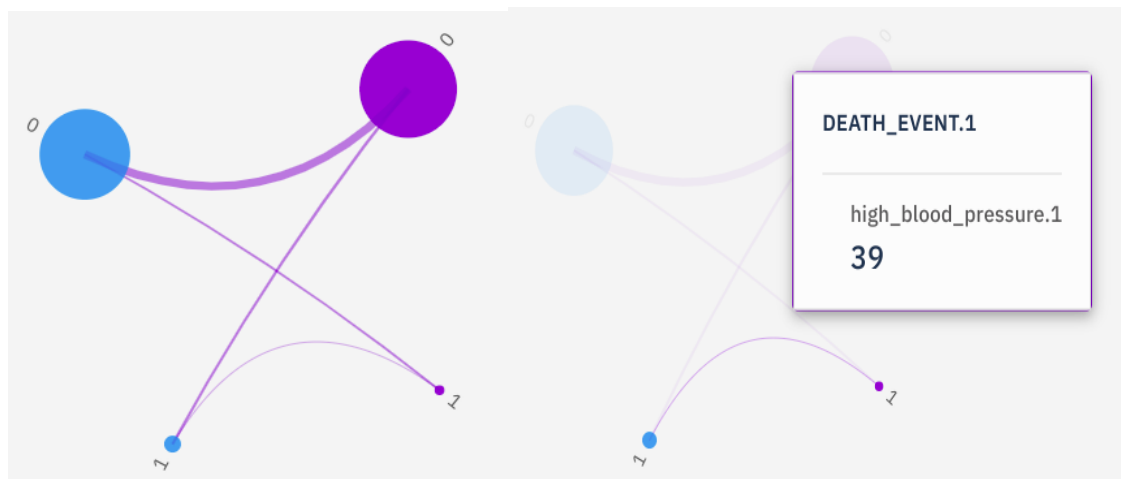
If the patient is having problems in the particular field or the combination of two or more fields than the patient might face heart failure.

View Output: DEATH\_EVENT x high\_blood\_pressure Compare ×

Field value: high\_blood\_pressure

DEATH_EVENT	0	1
0	137	66
1	57	39

Cells contain: cross-tabulation of fields (including missing values)  
Chi-square = 1.883, df = 1, probability = 0.17



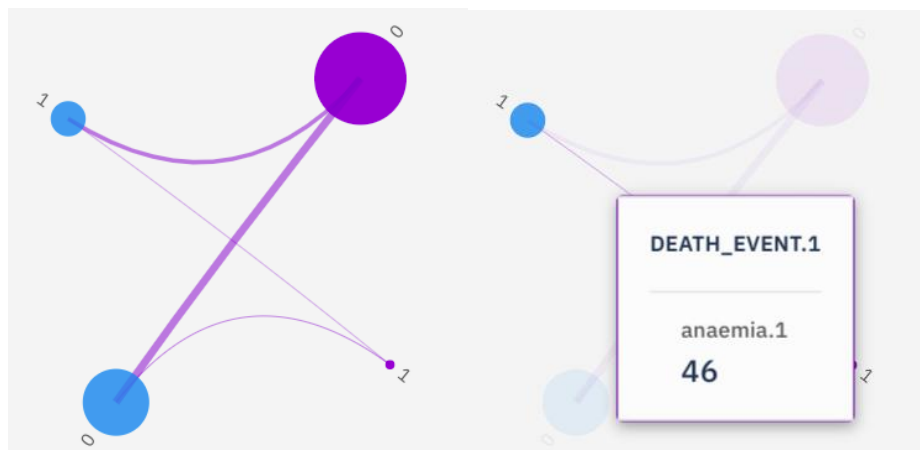
- 39 patients were having high blood pressure and they died.
- Even though 66 patients were having high blood pressure, they didn't die.
- Even though 57 patients were not having high blood pressure, they died.

View Output: DEATH\_EVENT x anaemia Compare ×

Field value: anaemia

DEATH_EVENT	0	1
0	120	83
1	50	46

Cells contain: cross-tabulation of fields (including missing values)  
Chi-square = 1.313, df = 1, probability = 0.252



- 46 patients were having anaemia and they died.
- Even though 83 patients were smoking, they didn't die.
- Even though 50 patients were not smoking, they died.



View Output: DEATH\_EVENT x diabetes

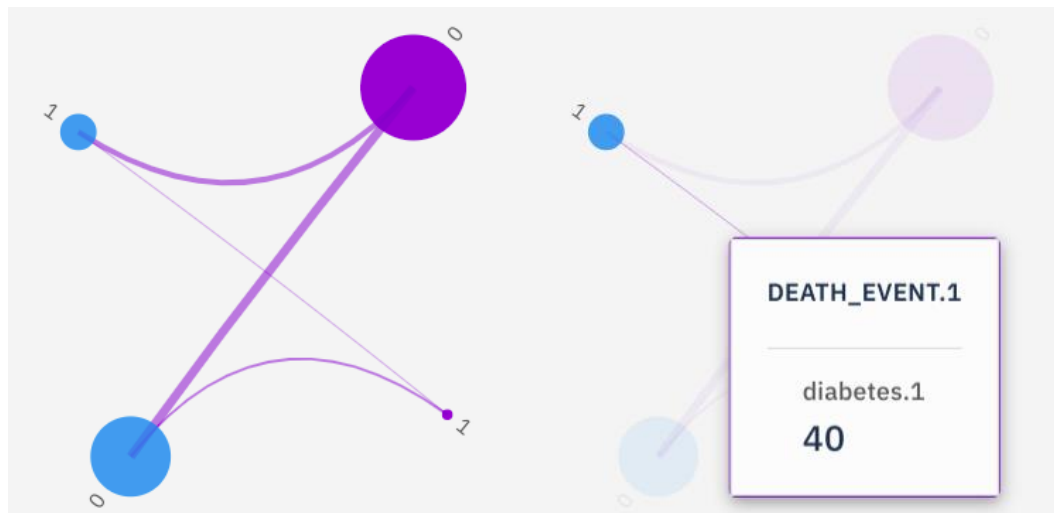
[Compare](#)

x

Field value: diabetes

DEATH_EVENT	0	1
0	118	85
1	56	40

Cells contain: cross-tabulation of fields (including missing values)  
Chi-square = 0.001, df = 1, probability = 0.973



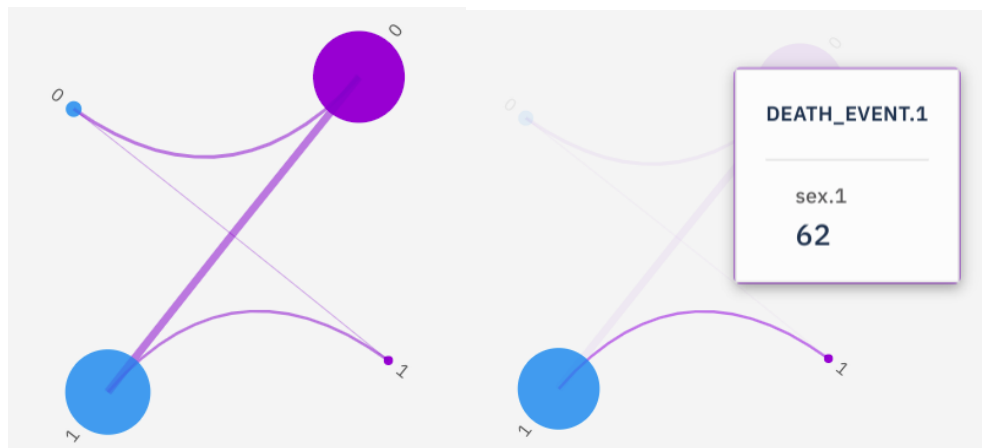
- 40 patients were having diabetes and they died.
- Even though 85 patients were having diabetes, they didn't die.
- Even though 56 patients were not having diabetes, they died.

View Output: DEATH\_EVENT x sex Compare ×

Field value: sex

DEATH_EVENT	0	1
0	71	132
1	34	62

Cells contain: cross-tabulation of fields (including missing values)  
Chi-square = 0.006, df = 1, probability = 0.941



- 71 patients who were female but did not die.
- 34 patients who were female, but they did die.
- 132 patients who were male but did not die.
- 62 patients who were male but did die.

View Output: DEATH\_EVENT x smoking

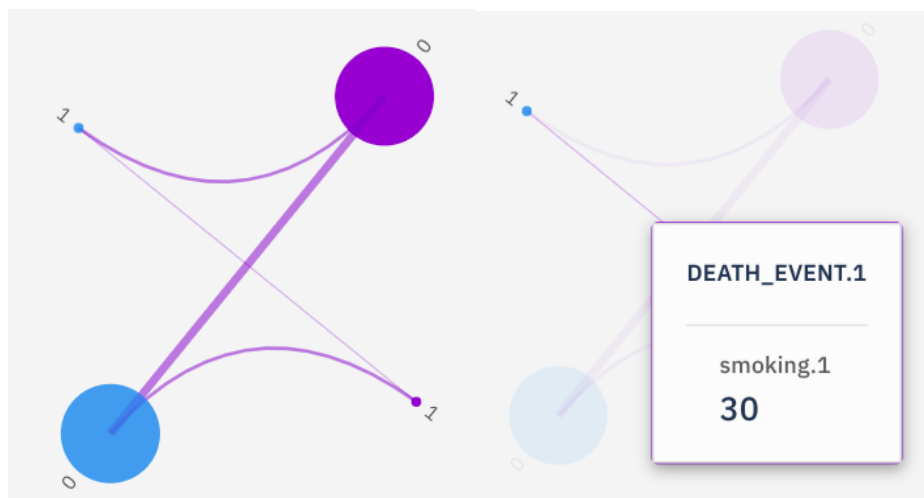
[Compare](#)

✕

Field value: smoking

DEATH_EVENT	0	1
0	137	66
1	66	30

Cells contain: cross-tabulation of fields (including missing values)  
Chi-square = 0.048, df = 1, probability = 0.827



- 30 patients were smoking, and they died.
- Even though 66 patients were smoking, they didn't die.
- Even though 66 patients were not smoking, they also died.

## 2. Which attributes seem promising for further analysis?

Fields like creatinine phosphokinase, ejection fraction, platelets, serum creatinine and serum sodium were continuous but we were not able to find their relationship with death event field so we changed the measurement level of these fields by reclassifying them into 3 classes

1. If the value is in ideal range then the value assigned is 1
2. If the value is less then the ideal range then the value assigned is 0
3. If the value is greater than ideal range, then the value assigned is 2

### → Creatinine\_phosphokinase:

Value between 10 and 120 => 1 (ideal)

Value less than 10 => 0 (low)

Value greater than 120 => 2 (high)

### → Ejection\_fraction:

Value between 50% and 70% => 1 (ideal)

Value less than 50% => 0 (low)

Value greater than 70% => 2 (high)

### → platelets:

Value between 150000 and 450000 => 1 (ideal)

Value less than 150000 => 0 (low)

Value greater than 450000 => 2 (high)

### → serum\_creatinine:

For adult men, 0.74 to 1.35 => 1 (ideal)

For adult men, less than 0.74 => 0 (low)

For adult men, greater than 1.35 => 2 (high)

For adult women, 0.59 to 1.04 => 1 (ideal)

For adult women, less than 0.59 => 0 (low)

For adult women, greater than 1.04 => 2 (high)

→ **serum\_sodium:**

Value between 135 to 145 => 1 (ideal)

Value less than 135 => 0 (low)

Value greater than 145 => 2 (high)

### 3. Have your explorations revealed new characteristics about the data?

Yes, while exploring the matrix node and finding the relationship of each column with DEATH\_EVENT we came to the conclusion that the DEATH\_EVENT can't be predicted by single attributes. We will need the combination of attributes to predict the death event correctly.

View Output: DEATH\_EVENT x creatinine\_phosphokinase\_reclassified

[Compare](#)

×

Field value: creatinine\_phosphokinase\_reclassified

DEATH_EVENT	1	2
0	58	145
1	19	77

Cells contain: cross-tabulation of fields (including missing values)  
Chi-square = 2.628, df = 1, probability = 0.105

View Output: DEATH\_EVENT x ejection\_fraction\_reclassified

[Compare](#)

×

Field value: ejection\_fraction\_reclassified

DEATH_EVENT	0	1	2
0	157	45	1
1	82	14	0

Cells contain: cross-tabulation of fields (including missing values)  
Chi-square = 2.905, df = 2, probability = 0.234

View Output: DEATH\_EVENT x platelets\_reclassified

[Compare](#)

×

Field value: platelets\_reclassified

DEATH_EVENT	0	1	2
0	16	179	8
1	11	80	5

Cells contain: cross-tabulation of fields (including missing values)  
Chi-square = 1.341, df = 2, probability = 0.512

View Output: DEATH\_EVENT x platelets\_reclassified

[Compare](#)

✕

Field value: serum\_creatinine\_reclassified

DEATH_EVENT	0	1	2
0	16	134	53
1	2	41	53

Cells contain: cross-tabulation of fields (including missing values)  
Chi-square = 25.255, df = 2, probability = 0

View Output: DEATH\_EVENT x serum\_sodium\_reclassified

[Compare](#)

✕

Field value: serum\_sodium\_reclassified

DEATH_EVENT	0	1	2
0	41	161	1
1	42	53	1

Cells contain: cross-tabulation of fields (including missing values)  
Chi-square = 18.609, df = 2, probability = 0

View Output: DEATH\_EVENT x serum\_creatinine\_reclassified

[Compare](#)

✕

Field value: serum\_creatinine\_reclassified

DEATH_EVENT	0	1	2
0	16	134	53
1	2	41	53

Cells contain: cross-tabulation of fields (including missing values)  
Chi-square = 25.255, df = 2, probability = 0

#### 4. How have these explorations changed your initial hypothesis?

No, the given fields were related to the prediction of heart failure only so there were no changes observed during exploration.

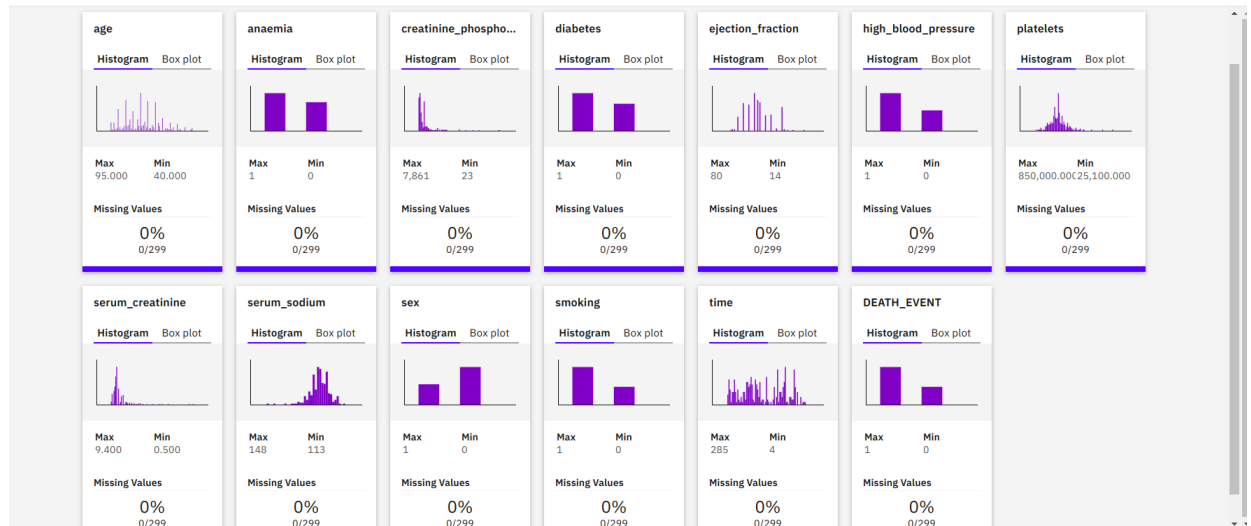
## Data Quality Report

### 1. Have you identified missing attributes and blank fields? If so, is there meaning behind such missing values?

There are no missing attributes and blank fields in this file.

View Output: Data Audit

Compare  



### 2. Are there spelling inconsistencies that may cause problems in later merges or transformations?

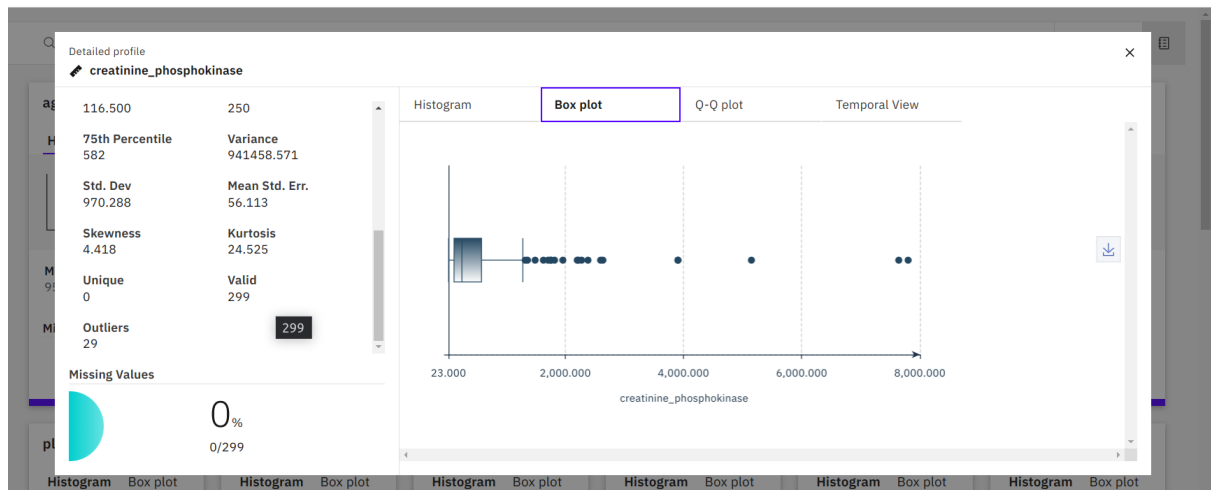
There is no spelling inconsistency yet identified in any field.

### 3. Have you explored deviations to determine whether they are "noise" or phenomena worth analysing further?

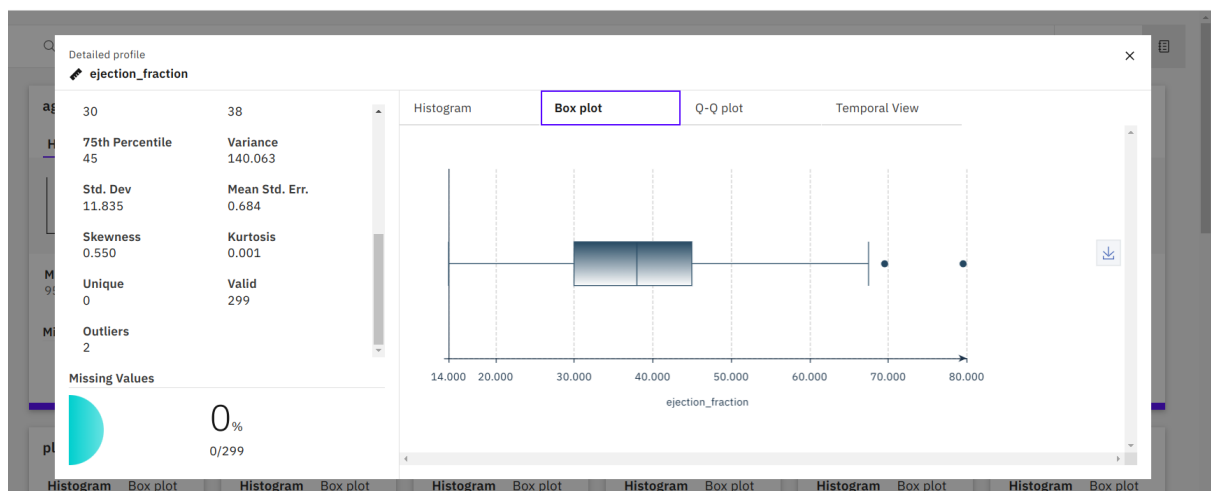
There are outliers present in creatinine\_phosphokinase, ejection\_fraction, platelets, serum\_creatinine, serum\_sodium and all are valid outliers and no invalid outliers are present.

- There are 29 valid outliers in creatinine\_phosphokinase
- There are 2 valid outliers in ejection\_fraction
- There are 21 valid outliers in platelets
- There are 29 valid outliers in serum\_creatinine
- There are 4 valid outliers in serum\_sodium

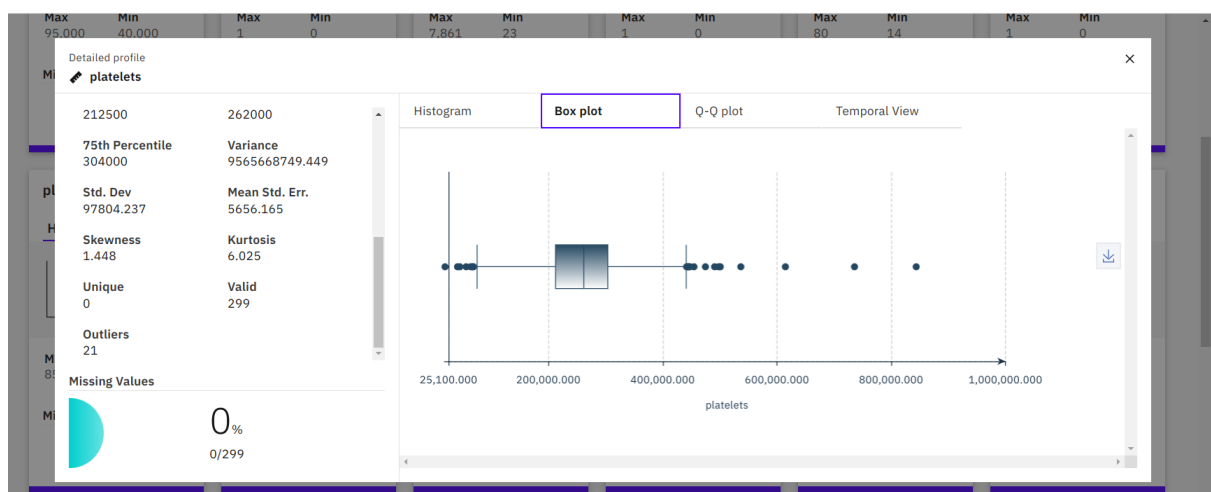
View Output: Data Audit

[Compare](#)

View Output: Data Audit

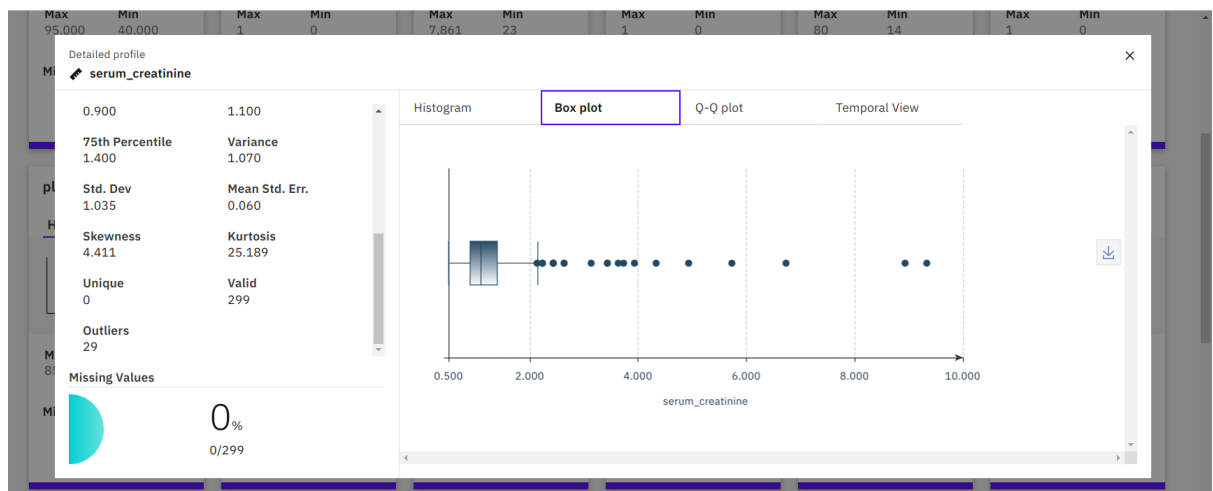
[Compare](#)

View Output: Data Audit

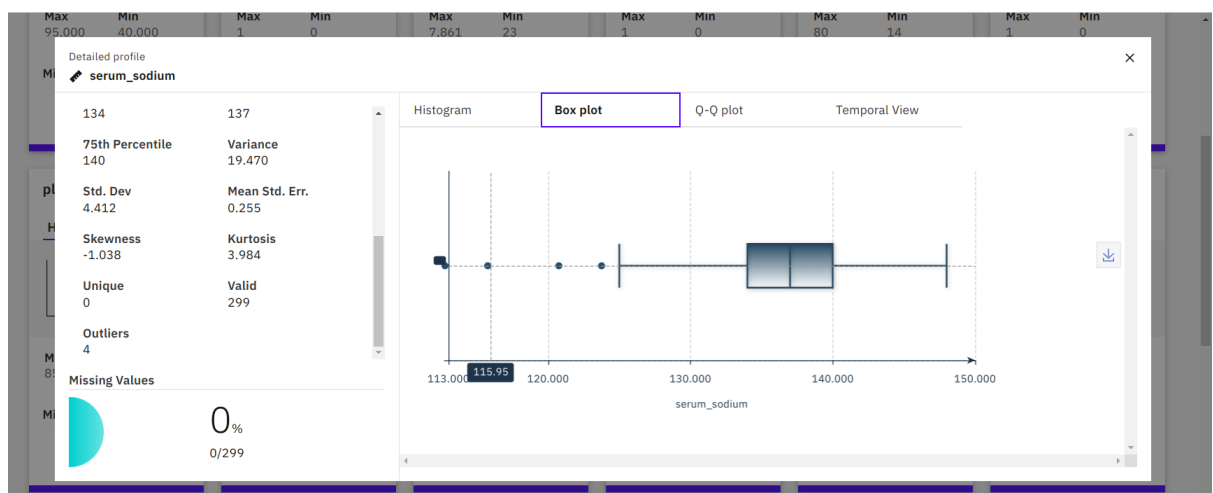
[Compare](#)



View Output: Data Audit



View Output: Data Audit



4. Have you conducted a plausibility check for values? Take notes on any apparent conflicts (such as teenagers with high income levels).

There is no plausibility.

5. Have you considered excluding data that has no impact on your hypotheses?

We can remove columns sex and time.

6. Is the data stored in flat files? If so, are the delimiters consistent among files? Does each record contain the same number of fields?

It's a CSV file so it's delimited by comma.