Exploratory Analysis

Naive Frequentists

2021-10-08

Read in data and create factors

```
dat = read_csv("heart.csv")
dataset_wfactors = dat %>% mutate(anaemia = factor(anaemia),
                              diabetes = factor(diabetes),
                              high_blood_pressure = factor(high_blood_pressure),
                              sex = factor(sex),
                              smoking = factor(smoking),
                              DEATH_EVENT = factor(DEATH_EVENT)
                              )

# Dataset without factors
dataset = dataset_wfactors %>%
  select(-c(anaemia, diabetes, high_blood_pressure, sex, smoking, DEATH_EVENT))
```

As with any exploratory analysis, the first step is to determine which variable are categorical, and which are continuous. We see that there are 6 categorical variables in this dataset. They were converted to `factors` to ensure ease of use with $R's$ functions.

Missing Values

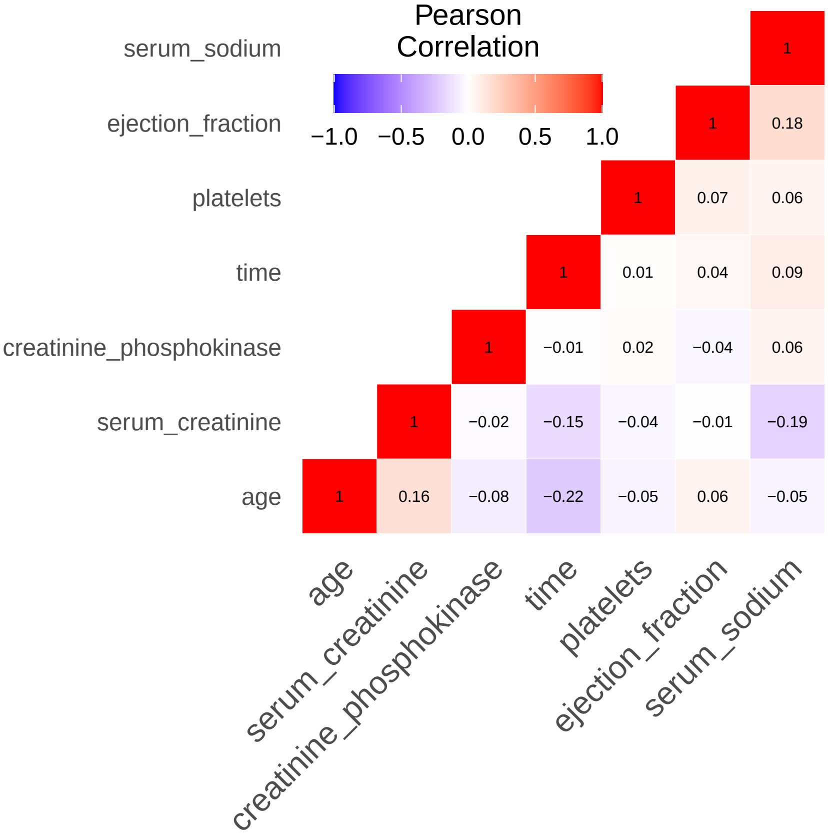
```
apply(dataset_wfactors, 2, FUN = function(x) sum(is.na(x)))
```

```
##                      age                  anaemia creatinine_phosphokinase
##                        0                        0                        0
##                 diabetes        ejection_fraction      high_blood_pressure
##                        0                        0                        0
##                platelets         serum_creatinine             serum_sodium
##                        0                        0                        0
##                      sex                  smoking                     time
##                        0                        0                        0
##              DEATH_EVENT
##                        0
```

The next step was to screen for missing values as many R functions cannot handle NAs. In addition, missing values can be due to underlying issues with the treatment and/or data collection procedure. Fortunately, the data does not have any missingness.

Summary Stats

Continuous Correlation Matrix

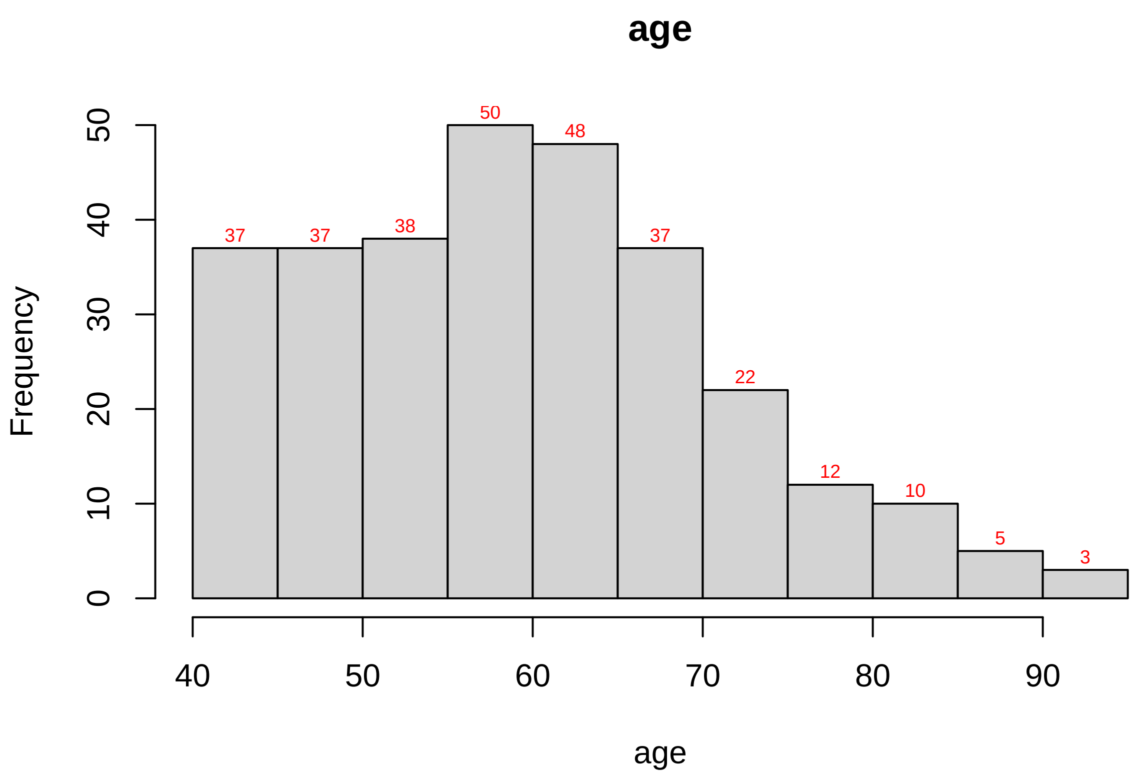


The above graphic indicates positive and negative associations between variables. There also is a clear clustering among certain variables that share a common direction of association. We notice moderate negative relationshps between time/age, time/serum__creatinine, and serum__sodium/serum__creatinine. Our next step will be to gain additional insight into the biological mechanisms that explain why certain variables associate with each other.

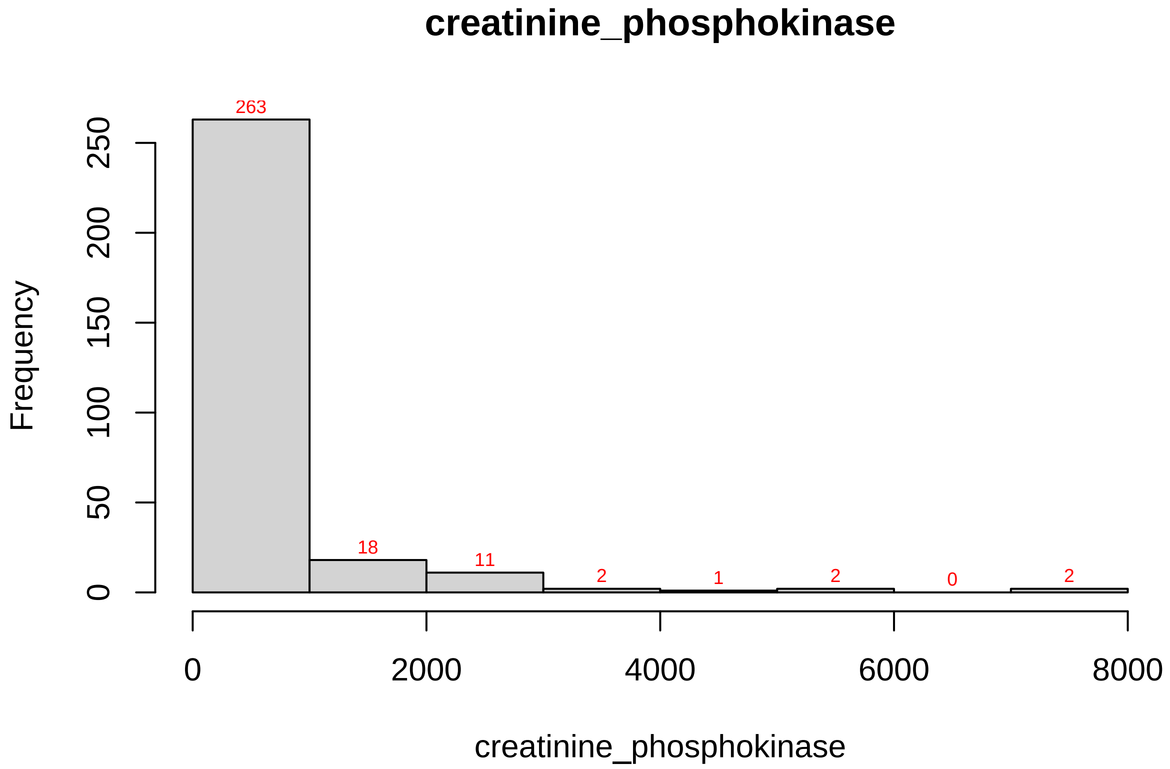
Histograms

Continuous Variables

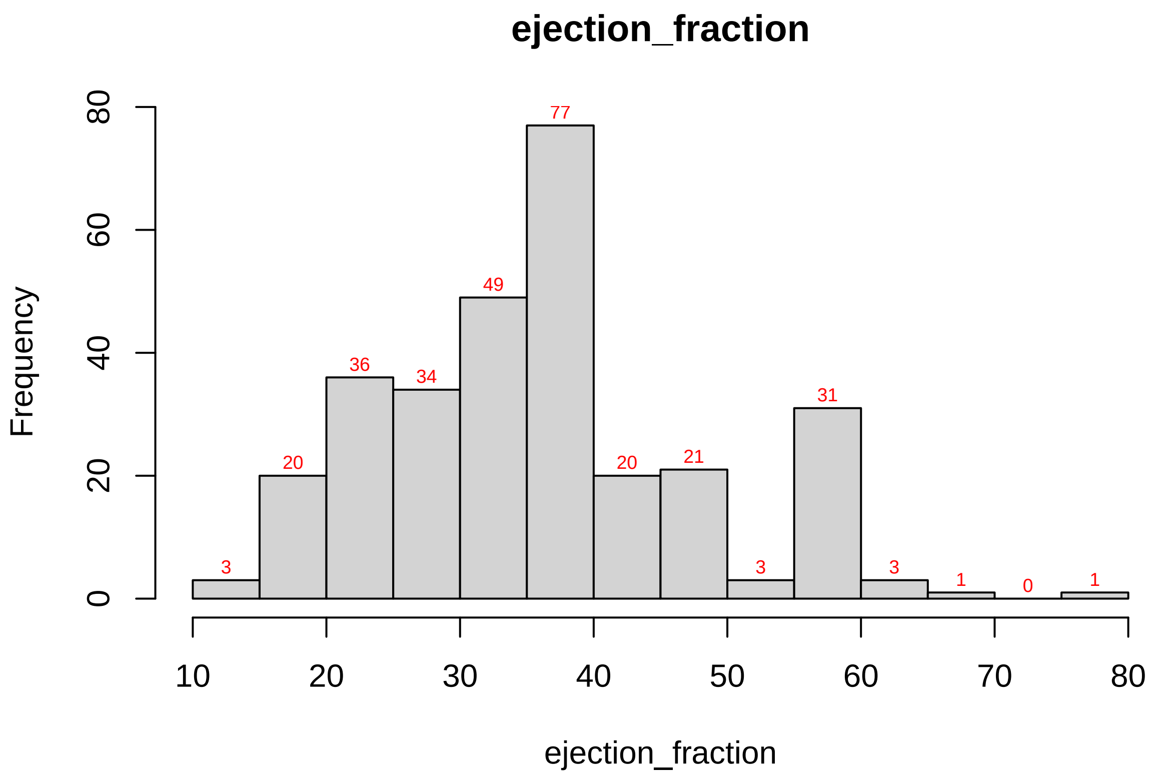
```
do = data.frame(dataset)
for(i in 1:ncol(do)) {
    h = hist(do[, i], main = names(do)[i], xlab = names(do)[i])
    text(h$mids, h$counts, labels=h$counts, adj=c(0.5, -0.5), cex = 0.6, col = "red")
    print("")
}
```


age
Frequency
0
10
20
30
40
50
37
37
38
50
48
37
22
12
10
5
3
40
50
60
70
80
90
age

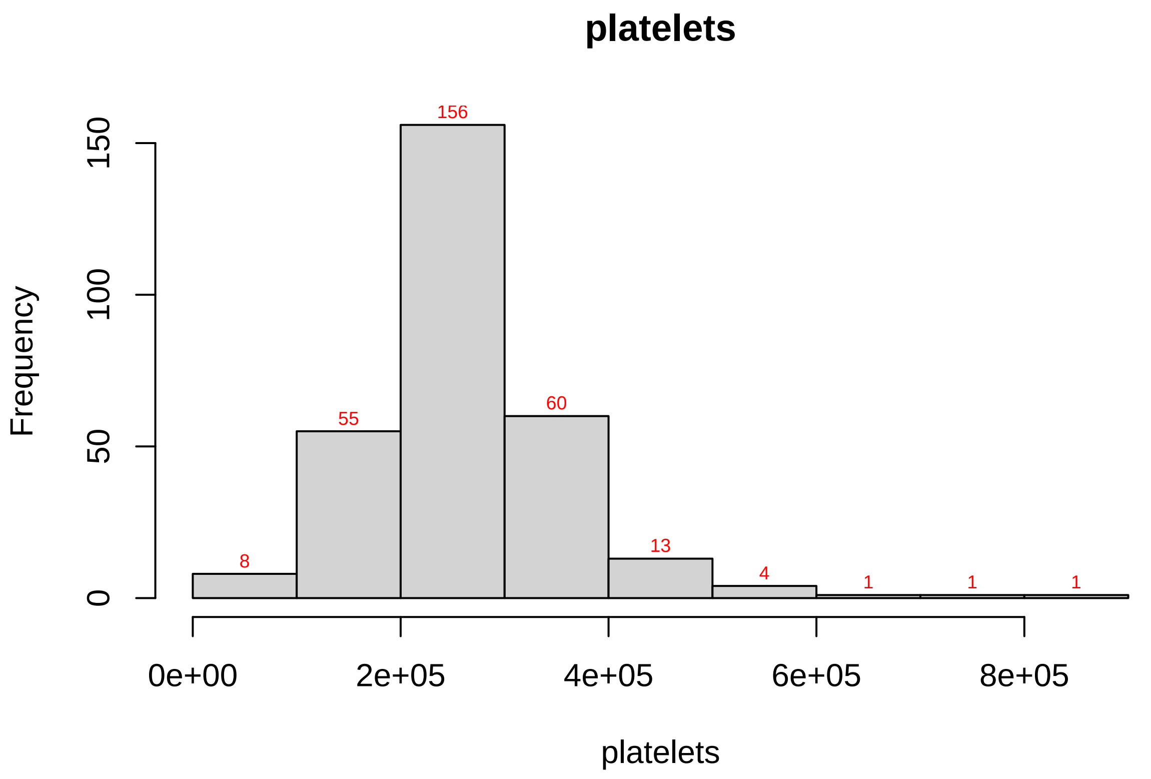
```
## [1] ""
```


creatinine_phosphokinase
Frequency
0
50
100
150
200
250
263
18
11
2
1
2
0
2
0
2000
4000
6000
8000
creatinine_phosphokinase

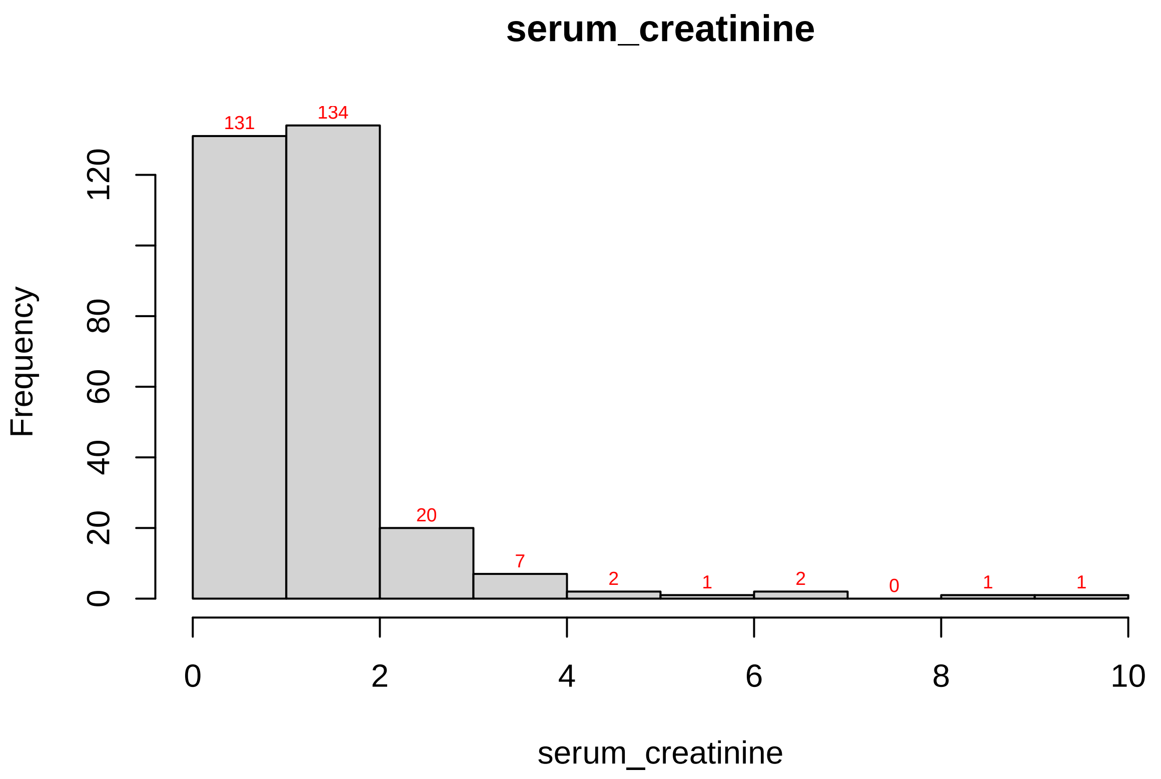
```
## [1] ""
```


ejection_fraction
Frequency
0
20
40
60
80
3
20
36
34
49
77
20
21
3
31
3
1
0
1
10
20
30
40
50
60
70
80
ejection_fraction

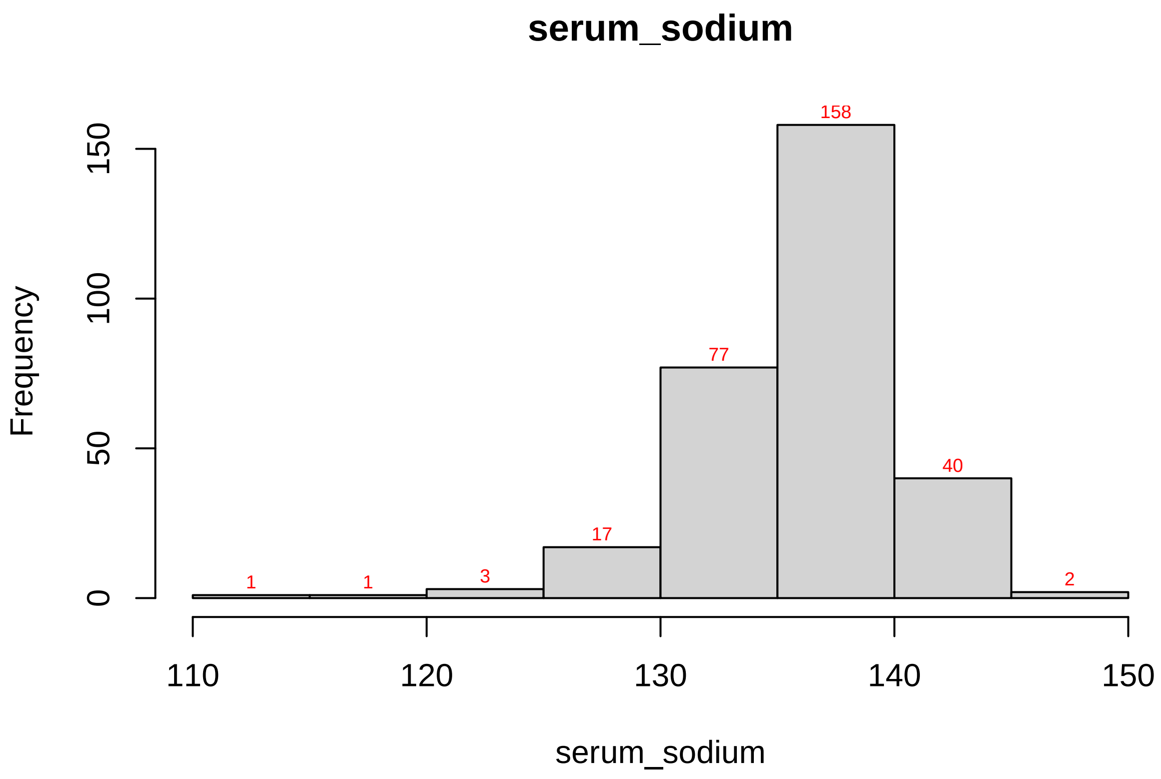
```
## [1] ""
```


platelets
Frequency
0
50
100
150
8
55
156
60
13
4
1
1
1
0e+00
2e+05
4e+05
6e+05
8e+05
platelets

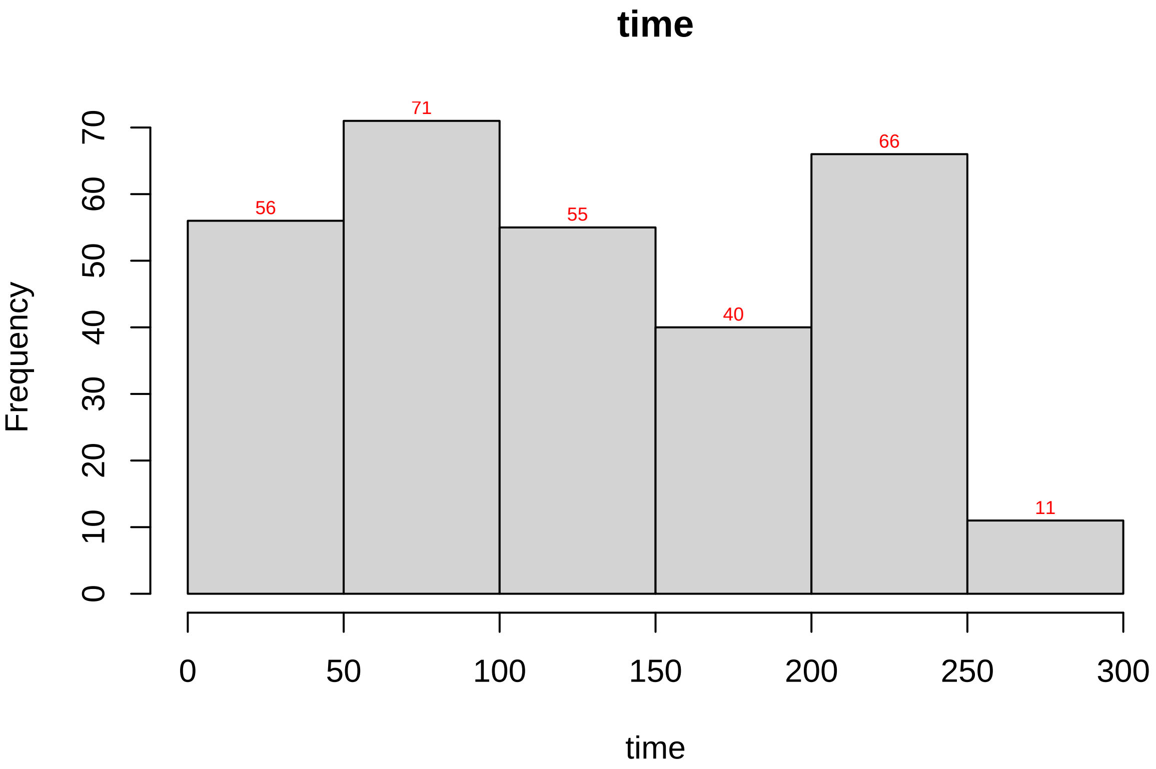
```
## [1] ""
```


serum_creatinine
Frequency
0
20
40
60
80
120
131
134
20
7
2
1
2
0
1
1
0
2
4
6
8
10
serum_creatinine

```
## [1] ""
```


serum_sodium
Frequency
0
50
100
150
1
1
3
17
77
158
40
2
110
120
130
140
150
serum_sodium

```
## [1] ""
```



```
## [1] ""
```

The histograms showcase the different distributions present in the data. The variables of serum_creatinine, creatinine_phosphokinase, and age have a right-skew, while serum_sodium, platelets, and ejection_fraction are more normally distributed. The variable time is more uniformly distributed. Again, more domain knowledge will give us a better understanding of the typical distributions that these variables take. This will be useful in helping us gauge the generalizability of our results.

Factors

```
dataset_wfactors %>%
  select(c(anaemia, diabetes, high_blood_pressure, sex, smoking, DEATH_EVENT)) %>%
  apply(2, table) %>% pander
```

	anaemia	diabetes	high_blood_pressure	sex	smoking	DEATH_EVENT
0	170	174	194	105	203	203
1	129	125	105	194	96	96

```
dataset_wfactors %>% group_by(sex) %>% count(DEATH_EVENT) %>% pander
```

sex	DEATH_EVENT	n
0	0	71
0	1	34
1	0	132

sex	DEATH_EVENT	n
1	1	62

```
dataset_wfactors %>% group_by(smoking) %>% count(DEATH_EVENT) %>% pander
```

smoking	DEATH_EVENT	n
0	0	137
0	1	66
1	0	66
1	1	30

```
dataset_wfactors %>% group_by(anaemia) %>% count(DEATH_EVENT) %>% pander
```

anaemia	DEATH_EVENT	n
0	0	120
0	1	50
1	0	83
1	1	46

The above tables show a common ration of 2 to 1 for many of the factors. Smoking/non-Smoking is roughly 2:1, as is high/not-high blood pressure. This is also true when we look at the counts for Death-event for each sex; the ratio of dying to not dying is roughly 2 to 1 for both men and women.

Regression Models

Initial Logistic Screening (all variables)

```
logistic_1 <- glm(data = dataset_wfactors,
                  DEATH_EVENT ~ ., family = "binomial")


pander(summary(logistic_1))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	10.18	5.657	1.801	0.07177
age	0.04742	0.0158	3.001	0.00269
anaemia1	-0.00747	0.3605	-0.02072	0.9835
creatinine_phosphokinase	0.0002222	0.0001779	1.249	0.2117
diabetes1	0.1451	0.3512	0.4133	0.6794
ejection_fraction	-0.07666	0.01633	-4.695	2.668e-06
high_blood_pressure1	-0.1027	0.3587	-0.2862	0.7747
platelets	-1.2e-06	1.889e-06	-0.635	0.5254
serum_creatinine	0.6661	0.1815	3.67	0.0002425
serum_sodium	-0.06698	0.03974	-1.686	0.09186
sex1	-0.5337	0.4139	-1.289	0.1973
smoking1	-0.01349	0.4126	-0.0327	0.9739

	Estimate	Std. Error	z value	Pr(>|z|)
time	-0.02104	0.003014	-6.981	2.923e-12

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	375.3 on 298 degrees of freedom
Residual deviance:	219.6 on 286 degrees of freedom

The inital screening above provides evidence of which terms might be most useful in predicting the death_event outcome. In particular, it seems that age, ejection_fraction, serum_creatinine, and time are the most significant covariates in our model. Further steps will be to use other metrics (AIC, BIC, etc.) to perform model selection and assess our model.

95% CI for Odds-Ratios

```
logistic_1 %>% confint %>% exp %>% pander # Exponentiate
```

Waiting for profiling to be done...

	2.5 %	97.5 %
(Intercept)	0.4449	2.422e+09
age	1.018	1.083
anaemia1	0.4858	2.008
creatinine_phosphokinase	0.9999	1.001
diabetes1	0.5797	2.31
ejection_fraction	0.8955	0.955
high_blood_pressure1	0.4423	1.815
platelets	1	1
serum_creatinine	1.371	2.859
serum_sodium	0.8635	1.011
sex1	0.2568	1.311
smoking1	0.4382	2.225
time	0.973	0.9846