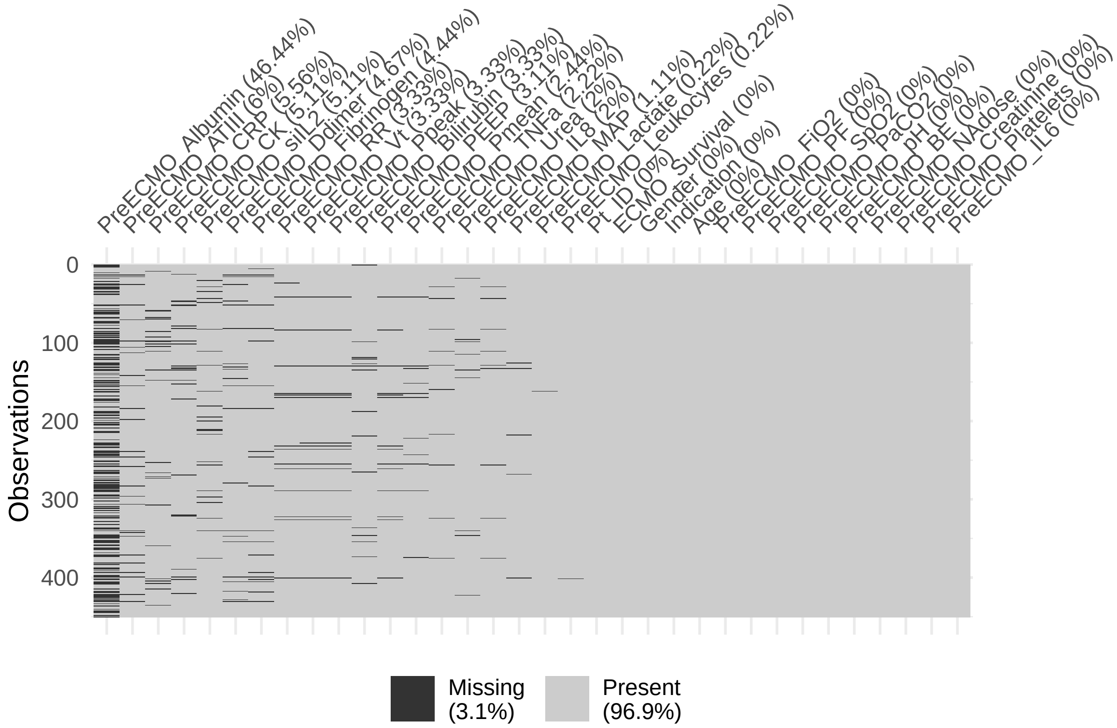
Classification of Acute Respiratory Distress Syndrome

Robert Edwards

Removing features that are not relevant to this analysis

Selecting only rows with data.



We see that `PreECMO_Albumin` is missing in 46% of observations. In listwise deletion, this will cause it to drop the entire row if a single observation has missing data. So we drop this feature.

An upset plot from the `UpSetR` package can be used to visualise the patterns of missingness, or rather the combinations of missingness across cases. To see combinations of missingness and intersections of missingness amongst variables, use the `gg_miss_upset` function:

If there are 40 intersections, there will be up to 40 combinations of variables explored. The number of sets and intersections can be changed by passing arguments `nsets = 10` to look at 10 sets of variables, and nintersects = 50 to look at 50 intersections.

This plot shows the number of missing values in each variable in a dataset. It is powered by the `miss_var_summary()` function.

This plot shows the number of missings in each column, broken down by a categorical variable from the dataset. It is powered by a `dplyr::group_by` statement followed by `miss_var_summary()`.

This plot shows the cumulative sum of missing values, reading the rows of the dataset from the top to bottom. It is powered by the `miss_case_cumsum()` function.

This plot shows the cumulative sum of missing values, reading columns from the left to the right of your dataframe. It is powered by the `miss_var_cumsum()` function.

Exploratory Data Analysis

We first visually explore the data to get a sense of the features and distributions of the data. Then we will conduct hypothesis tests for each feature based on `ECMO_Survival` as a rough idea how relevant each feature will be.

To get an idea of the distribution of the data, the following summary statistics were obtained for the categorical variable `ECMO_Survival` (Table 1) and for the continuous variables (Table 2).

Table 1: Numbers of survivors and nonsurvivors of ECMO treatment.

ECMO_Survival	n	Percent %
N	109	24.22
Y	341	75.78

Table 1 shows that out of the 450 individuals, only 75.78% of the individuals in the study sample survived ECMO treatment (341 survived vs 109 did not survive).

Table 2: Number of males and females.

Gender	n	Percent %
m	305	67.78
w	145	32.22

Table 2 shows that out of the 450 individuals, only 67.78% of the individuals in the study sample are male (305 male vs 145 female).

Table 3: Number of each disease type indication.

Indication	n	Percent %
1	66	14.67
2	181	40.22
3	31	6.89
4	28	6.22
5	71	15.78
6	12	2.67
7	61	13.56

Table 3 shows the distribution of each disease type indication:

- ALF - Acute Lung Failure - 0%
- 1 - Viral Pneumonia - 14.67%
- 2 - Bacterial Pneumonia - 40.22%
- 3 - Aspiration Pneumonitis - 6.89%

- 4 - ARDS Trauma - 6.22%
- 5 - ARDS Surgery - 15.78%
- 6 - Chemo - 2.67%
- 7 - Other - 13.56%

Table 4: Summary statistics on ARDS data continuous variables.

Variable	n	Mean	SD	Min	1st quartile	Median	3rd quartile	Max
Age	450	51.66	14.45	18	42	53	62.75	83
Indication	450	3.3	2	1	2	2	5	7
PreECMO_Albumin	450	22.17	6.57	6	18	22	27	41
PreECMO_ATIII	450	65.07	39.18	10	47	63	80	650
PreECMO_BE	450	-1.4	7.27	-39	-6	-2	3	32
PreECMO_Bilirubin	450	1.8	3.08	0.1	0.5	0.8	1.8	29.6
PreECMO_CK	450	1159.54	3492.62	9	74	200	683.5	36102
PreECMO_Creatinine	450	1.67	1.35	0.1	0.8	1.25	2.08	11.6
PreECMO_CRP	450	167.44	126.8	1	50	152	262	569
PreECMO_Ddimer	450	10.08	10.2	1	3	6	13	36
PreECMO_Fibrinogen	450	525.81	236.34	40	356	510.5	650.5	1236
PreECMO_FiO2	450	0.92	0.16	0.21	0.9	1	1	1
PreECMO_IL6	450	13807.71	47296.01	4	93.5	461.5	4426	6e+05
PreECMO_IL8	450	5658.68	31013.4	6	40	113	421	376513
PreECMO_Lactate	450	32.2	37.07	3	11	17	36	336
PreECMO_Leukocytes	450	14.64	10.06	0	8	13.1	19.6	91.5
PreECMO_MAP	450	69.46	12.22	34	61	68	76	109
PreECMO_NAdose	450	0.46	0.67	0	0.11	0.28	0.59	6.94
PreECMO_PaCO2	450	67.3	26.01	30	50.25	62	76	237
PreECMO_PEEP	450	15.06	4.23	2	12	15	17	35
PreECMO_PF	450	84.1	49.52	28	56	69	92	410
PreECMO_pH	450	7.22	0.13	6.39	7.16	7.23	7.31	7.57
PreECMO_Platelets	450	199.61	129.39	2	106.5	182	269	808
PreECMO_Pmean	450	22.39	5.1	5	20	22	25	40
PreECMO_Ppeak	450	33.8	5.79	15	30	34	37	50
PreECMO_RR	450	23.48	6.29	7	20	23	26	60
PreECMO_siIL2	450	4163.2	7893.67	27	1255	2183	4277	121123
PreECMO_SpO2	450	87.5	10.28	29	85	91	94	100
PreECMO_TNFa	450	49.61	107.72	4	14	25	46.25	1468
PreECMO_Urea	450	73.98	59.37	2	38	58	97	703
PreECMO_Vt	450	482.15	126.54	6	407.5	477	560	941

Looking at Table 4

Violin Plot

All continuous features are normally scaled to be comparable.

The boxplot shows many highly skewed variables indicating that these likely need to be transformed in some fashion (likely log transformation). For **logistic regression**, first a model will be fit without transformations and then the not significant variables will be transformed and model fit will be evaluated.

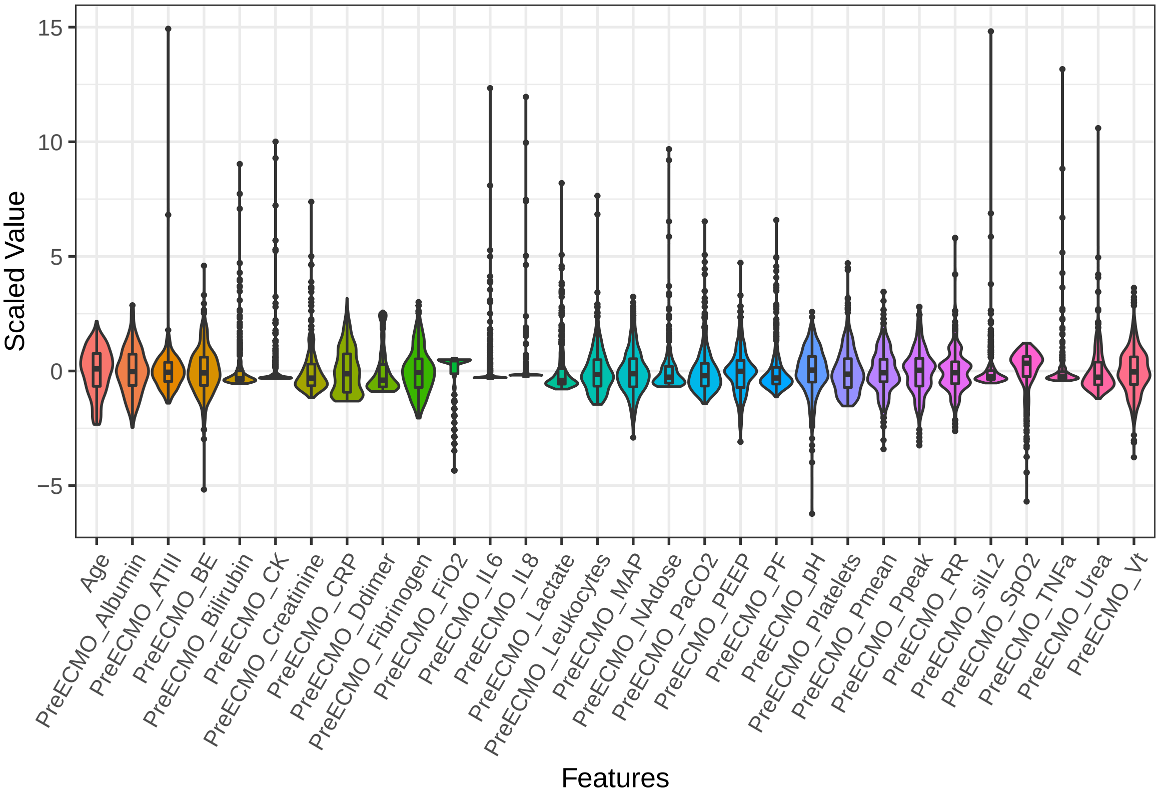


Figure 1: Violin plot of continuous variables.

```
pdf
  2
```

Violinplot After Yeo Johnson Transformation

- center
- scale
- YeoJohnson

```
pdf
  2
```

Correlation Plot

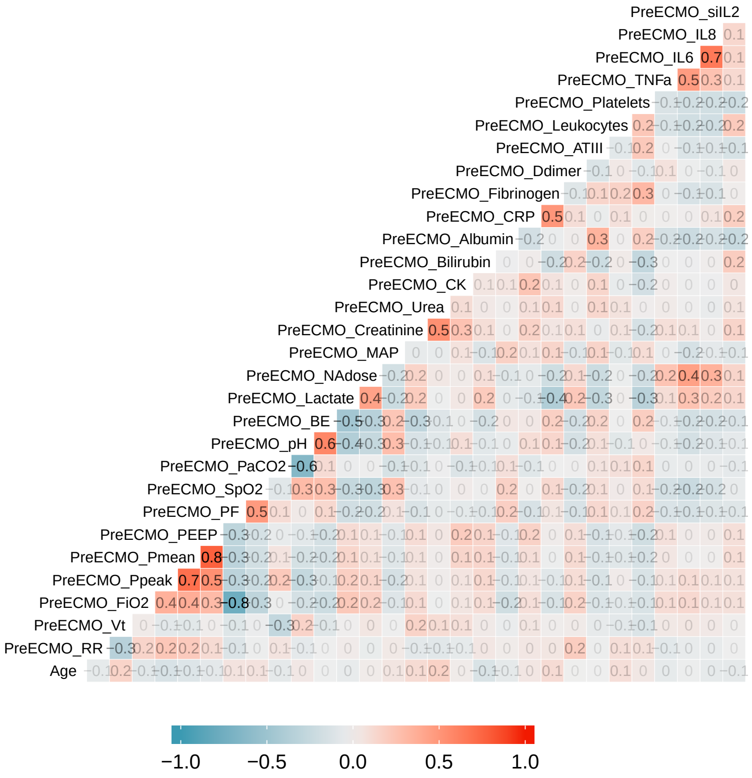


Figure 2: Correlation heatmap of contnuous variables.

Yeo Johnson Correlation Heatmap

```
pdf
  2
```

Pairs Plot

Hypothesis Testing

From the boxplot we see many variables are not normally distributed, or at the least highly skewed. To run a hypothesis test we should take note that a t-test assumes the data is normally distributed. For the skewed or non-normal variables we should use an appropriate hypothesis test such as Wilcoxon test that tests for differences in the medians of two samples.

Student's t-Tests

Table 5: Hypothesis tests for variables.

Variable	Test	df	p.value	Lower	Upper
Age	Welch Two Sample t-test	191.893	0.220	6.282	0.036
PreECMO_RR	Welch Two Sample t-test	165.424	0.692	3.569	0.004
PreECMO_Vt	Welch Two Sample t-test	171.378	-36.112	21.313	0.612
PreECMO_FiO2	Welch Two Sample t-test	208.702	-0.024	0.042	0.591
PreECMO_Ppeak	Welch Two Sample t-test	204.104	0.071	2.437	0.038
PreECMO_Pmean	Welch Two Sample t-test	181.332	-0.293	1.916	0.149
PreECMO_PEEP	Welch Two Sample t-test	156.898	-0.546	1.490	0.361
PreECMO_PF	Welch Two Sample t-test	222.155	-17.514	1.742	0.108
PreECMO_SpO2	Welch Two Sample t-test	192.720	-2.994	1.324	0.446
PreECMO_PaCO2	Welch Two Sample t-test	201.674	-3.667	6.985	0.540
PreECMO_pH	Welch Two Sample t-test	181.503	-0.065	-0.007	0.016
PreECMO_BE	Welch Two Sample t-test	162.959	-2.729	0.672	0.234
PreECMO_Lactate	Welch Two Sample t-test	129.424	4.752	25.748	0.005
PreECMO_NAdose	Welch Two Sample t-test	158.329	0.013	0.332	0.035
PreECMO_MAP	Welch Two Sample t-test	172.411	-4.503	1.017	0.214
PreECMO_Creatinine	Welch Two Sample t-test	161.874	-0.109	0.523	0.197
PreECMO_Urea	Welch Two Sample t-test	243.267	-5.381	16.925	0.309
PreECMO_CK	Welch Two Sample t-test	225.928	-804.650	524.023	0.678
PreECMO_Bilirubin	Welch Two Sample t-test	162.959	-0.281	1.154	0.231
PreECMO_Albumin	Welch Two Sample t-test	132.145	-1.242	2.402	0.530
PreECMO_CRP	Welch Two Sample t-test	189.928	-32.360	21.781	0.700
PreECMO_Fibrinogen	Welch Two Sample t-test	168.582	-60.406	48.590	0.831
PreECMO_Ddimer	Welch Two Sample t-test	155.397	0.554	5.411	0.016
PreECMO_ATIII	Welch Two Sample t-test	286.705	-10.614	3.324	0.304
PreECMO_Leukocytes	Welch Two Sample t-test	203.637	-2.790	1.313	0.479
PreECMO_Platelets	Welch Two Sample t-test	181.332	-62.969	-6.945	0.015
PreECMO_TNFa	Welch Two Sample t-test	126.035	-19.162	43.430	0.444
PreECMO_IL6	Welch Two Sample t-test	125.066	-2265.306	26171.608	0.099
PreECMO_IL8	Welch Two Sample t-test	130.139	-2600.980	14861.031	0.167
PreECMO_siIL2	Welch Two Sample t-test	164.829	-554.996	2994.096	0.177

Wilcoxon Signed-Rank Tests

Table 6: Hypothesis tests for variables.

Variable	Test	p.value
Age	Wilcoxon rank sum test with continuity correction	0.046
PreECMO_RR	Wilcoxon rank sum test with continuity correction	0.005
PreECMO_Vt	Wilcoxon rank sum test with continuity correction	0.588
PreECMO_FiO2	Wilcoxon rank sum test with continuity correction	0.976
PreECMO_Ppeak	Wilcoxon rank sum test with continuity correction	0.013
PreECMO_Pmean	Wilcoxon rank sum test with continuity correction	0.241
PreECMO_PEEP	Wilcoxon rank sum test with continuity correction	0.516
PreECMO_PF	Wilcoxon rank sum test with continuity correction	0.166
PreECMO_SpO2	Wilcoxon rank sum test with continuity correction	0.143
PreECMO_PaCO2	Wilcoxon rank sum test with continuity correction	0.303
PreECMO_pH	Wilcoxon rank sum test with continuity correction	0.014
PreECMO_BE	Wilcoxon rank sum test with continuity correction	0.225
PreECMO_Lactate	Wilcoxon rank sum test with continuity correction	0.024
PreECMO_NAdose	Wilcoxon rank sum test with continuity correction	0.011
PreECMO_MAP	Wilcoxon rank sum test with continuity correction	0.127
PreECMO_Creatinine	Wilcoxon rank sum test with continuity correction	0.136
PreECMO_Urea	Wilcoxon rank sum test with continuity correction	0.030
PreECMO_CK	Wilcoxon rank sum test with continuity correction	0.584
PreECMO_Bilirubin	Wilcoxon rank sum test with continuity correction	0.159
PreECMO_Albumin	Wilcoxon rank sum test with continuity correction	0.479
PreECMO_CRP	Wilcoxon rank sum test with continuity correction	0.884
PreECMO_Fibrinogen	Wilcoxon rank sum test with continuity correction	0.753
PreECMO_Ddimer	Wilcoxon rank sum test with continuity correction	0.004
PreECMO_ATIII	Wilcoxon rank sum test with continuity correction	0.521
PreECMO_Leukocytes	Wilcoxon rank sum test with continuity correction	0.515
PreECMO_Platelets	Wilcoxon rank sum test with continuity correction	0.004
PreECMO_TNFa	Wilcoxon rank sum test with continuity correction	0.820
PreECMO_IL6	Wilcoxon rank sum test with continuity correction	0.082
PreECMO_IL8	Wilcoxon rank sum test with continuity correction	0.000
PreECMO_siIL2	Wilcoxon rank sum test with continuity correction	0.128

χ^2 Tests

Need to convert `chars` to numeric factors in the categorical variables

Two random variables x and y are called independent if the probability distribution of one variable is not affected by the presence of another.

Assume f_{ij} is the observed frequency count of events belonging to both i-th category of x and j-th category of y. Also assume e_{ij} to be the corresponding expected count if x and y are independent. The null hypothesis of the independence assumption is to be rejected if the p-value of the following Chi-squared test statistics is less than a given significance level α.

$$\chi^2 = \sum_{ij} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Hypothesis 1

Test the hypothesis whether the `ECMO_Survival` is independent of `Gender` at .05 significance level.

	m	w
N	77	32
Y	228	113

Hypothesis 2

Test the hypothesis whether the `ECMO_Survival` is independent of disease `Indication` at .05 significance level.

	1	2	3	4	5	6	7
N	12	41	5	5	20	7	19
Y	54	140	26	23	51	5	42

Hypothesis 3

Test the hypothesis whether the `Gender` is independent of disease `Indication` at .05 significance level.

	1	2	3	4	5	6	7
m	35	123	24	25	54	9	35
w	31	58	7	3	17	3	26

Table 7: Hypothesis tests for variables.

Variables	Test	df	p.value
ECMO_Survival / Gender	Pearson's Chi-squared test with Yates' continuity correction	1	0.537
ECMO_Survival / Indication	Pearson's Chi-squared test	6	0.042
Gender / Indication	Pearson's Chi-squared test	6	0.004