Report on The Internal Lab Assignment AAPS on identifying anomalous patients in the cardiac ICU dataset

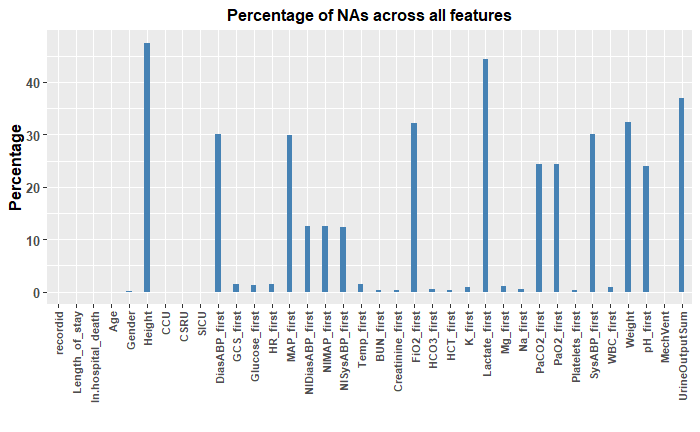
Load ICU dataset the ICU\_filtered.csv

the csv file had 36 columns with 7886 rows

1. 'data.frame': 7886 obs. of 36 variables:
2. $ recordid : int 132539 132540 132541 132543 132545 132547 132548 132551 132554 132555 ...
3. $ Length\_of\_stay : int 5 8 19 9 4 6 9 6 17 8 ...
4. $ In.hospital\_death: int 0 0 0 0 0 0 0 1 0 0 ...
5. $ Age : int 54 76 44 68 88 64 68 78 64 74 ...
6. $ Gender : int 0 1 0 1 0 1 0 0 0 1 ...
7. $ Height : num NA 175 NA 180 NA ...
8. $ CCU : int 0 0 0 0 0 1 0 0 0 0 ...
9. $ CSRU : int 0 1 0 0 0 0 0 0 0 1 ...
10. $ SICU : int 1 0 0 0 0 0 0 0 0 0 ...
11. $ DiasABP\_first : int NA 67 81 NA NA 79 98 50 NA 51 ...
12. $ GCS\_first : int 15 3 7 15 15 7 15 15 15 10 ...
13. $ Glucose\_first : int 205 105 141 129 113 264 94 132 113 106 ...
14. $ HR\_first : num 73 88 100 79 93 78 73 111 127 67 ...
15. $ MAP\_first : num NA 79 107 NA NA 101 136 83 NA 67 ...
16. $ NIDiasABP\_first : int 65 38 84 63 41 89 88 51 71 57 ...
17. $ NIMAP\_first : num 92.3 49.3 100.3 86.7 75.3 ...
18. $ NISysABP\_first : int 147 72 133 134 144 129 187 100 138 134 ...
19. $ Temp\_first : num 35.1 35.2 37.8 36.3 37.8 35.8 36.3 38 37.3 34.8 ...
20. $ BUN\_first : int 13 16 8 23 45 15 32 81 21 19 ...
21. $ Creatinine\_first : num 0.8 0.8 0.4 0.9 1 1.4 3.4 0.9 0.7 1.1 ...
22. $ FiO2\_first : num NA 1 1 NA NA 0.5 NA 1 NA 1 ...
23. $ HCO3\_first : int 26 21 24 28 18 19 25 18 21 23 ...
24. $ HCT\_first : num 33.7 24.7 28.5 41.3 22.6 41.6 31.9 32.6 28.3 31.5 ...
25. $ K\_first : num 4.4 4.3 3.3 4 6 5.1 3.7 4.2 3.9 4.6 ...
26. $ Lactate\_first : num NA NA 1.3 NA NA NA NA 1.4 NA NA ...
27. $ Mg\_first : num 1.5 3.1 1.9 2.1 1.5 1.7 1.9 2.2 1.6 1.8 ...
28. $ Na\_first : int 137 139 137 140 140 141 140 141 139 141 ...
29. $ PaCO2\_first : int NA 34 37 NA NA 45 NA 30 NA 42 ...
30. $ PaO2\_first : int NA 344 65 NA NA 78 NA 128 NA 486 ...
31. $ Platelets\_first : int 221 164 72 391 109 276 325 91 696 141 ...
32. $ SysABP\_first : int NA 105 148 NA NA 150 205 103 NA 98 ...
33. $ WBC\_first : num 11.2 7.4 4.2 11.5 3.8 24 6.2 16.1 15.2 9 ...
34. $ Weight : num NA 80.6 56.7 84.6 NA NA 87 48.4 60.7 66.1 ...
35. $ pH\_first : num NA 7.45 7.51 NA NA 7.29 NA 7.4 NA 7.39 ...
36. $ MechVent : int 0 1 1 0 0 1 0 1 0 1 ...
37. $ UrineOutputSum : int NA 5 14 NA NA 12 NA 12 NA 4 ...

Plot fraction of missing values (NAs) in each column of the data frame

this will show the count of the Null values present in each columns



Drop columns with more than 20% missing values

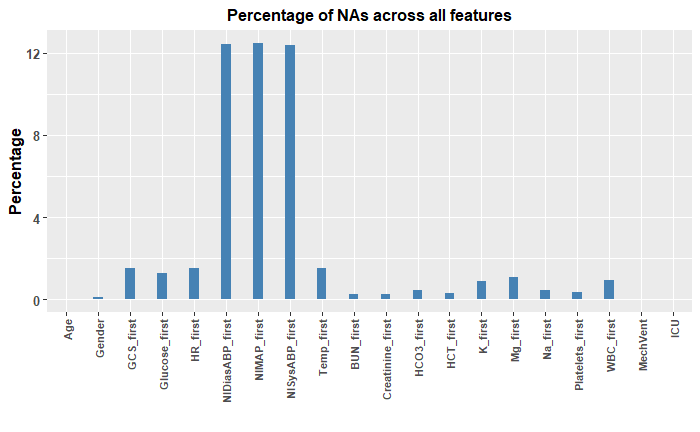
This will help us to remove the columns with more than 20% missing values, we do this because it will be difficult Impute the values for a column that as more than 20% missing values.

Combining 3 different CCU, CSRU, SICU categorical columns into one categorical column named ICU.

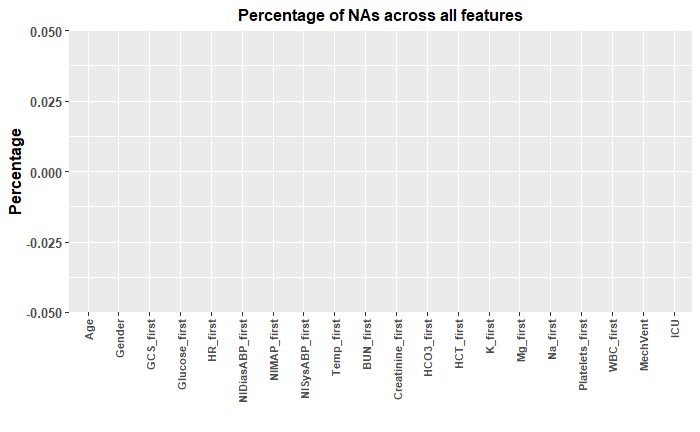
Collate 4 different ICU types (CCU, CSRU, SICU, None) into one column

called 'ICU', remove separate ICU columns and the following columns:

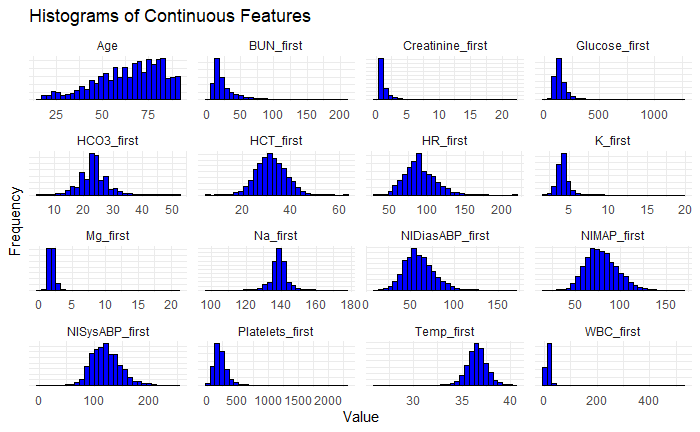
After Dropping and collating the ICU columns, seeing the missing values

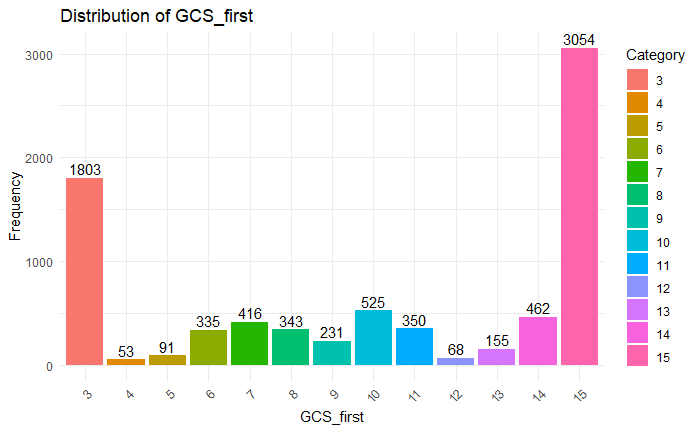
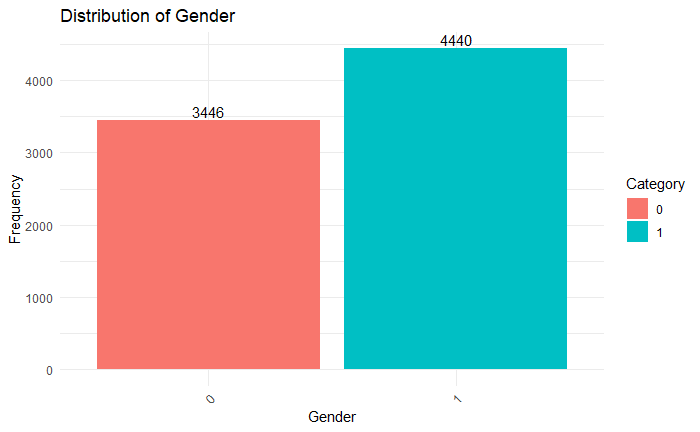
Impute missing values using the MICE package, this will fill the missing values based on each colums

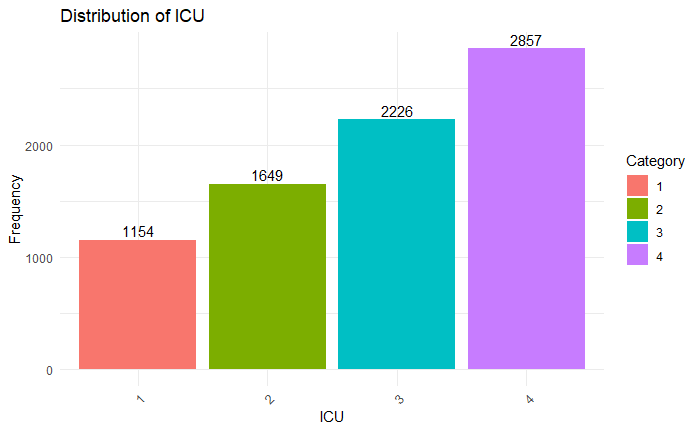
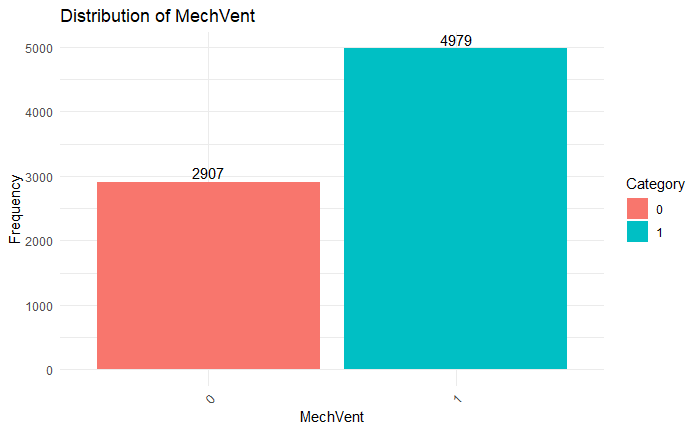
After that no missing values



Split the columns into categorical and continuous based on features and plotting them for better visualization and get an Idea on the data

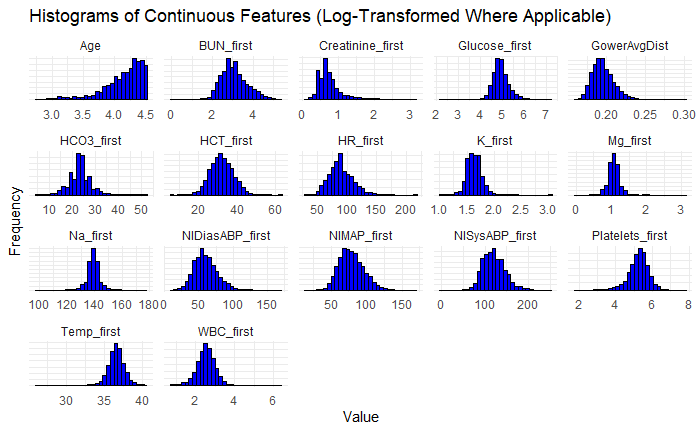






Make the categorical into continuous using Gowers Distance get one continuous feature for the four categorical features Gender, GCS\_first, MechVent, ICU.

Transformed features for Mahalanobis distance applied log transformation on features like "Age", "BUN\_first", "categorical\_avg\_distance", "categorical\_dist\_to\_ref", "Creatinine\_first", "Glucose\_first", "K\_first", "Mg\_first", "Platelets\_first", "WBC\_first"



Apply PCA to all the 17 continuous features find the most variance explained aprox(92%)

|  |  |  |
| --- | --- | --- |
| **PC** | **Variance Explained** | **Cumulative Variance** |
| **PC1** | 0.162570757 | 0.1625708 |
| **PC2** | 0.133007319 | 0.2955781 |
| **PC3** | 0.093050056 | 0.3886281 |
| **PC4** | 0.082768551 | 0.4713967 |
| **PC5** | 0.068800754 | 0.5401974 |
| **PC6** | 0.06480727 | 0.6050047 |
| **PC7** | 0.058375365 | 0.6633801 |
| **PC8** | 0.051632032 | 0.7150121 |
| **PC9** | 0.050391883 | 0.765404 |
| **PC10** | 0.046679712 | 0.8120837 |
| **PC11** | 0.040262674 | 0.8523464 |
| **PC12** | 0.037811435 | 0.8901578 |
| **PC13** | 0.034738974 | 0.9248968 |
| **PC14** | 0.030174315 | 0.9550711 |
| **PC15** | 0.023677299 | 0.9787484 |
| **PC16** | 0.013501665 | 0.9922501 |
| **PC17** | 0.007749939 | 1 |

Considered only the top 13 PCA directions as they are combined explaining about 92% variance in the dataset

Upon these 13 directions applied the Mahalanobis distance got a continuous feature for the 13 PCA directions

Marked the outliers as red Number of outlier points (99% threshold): 460 got these many points as outliers

plotted graphs for the first 5 directions

