README

Al Sabay

6/25/2018

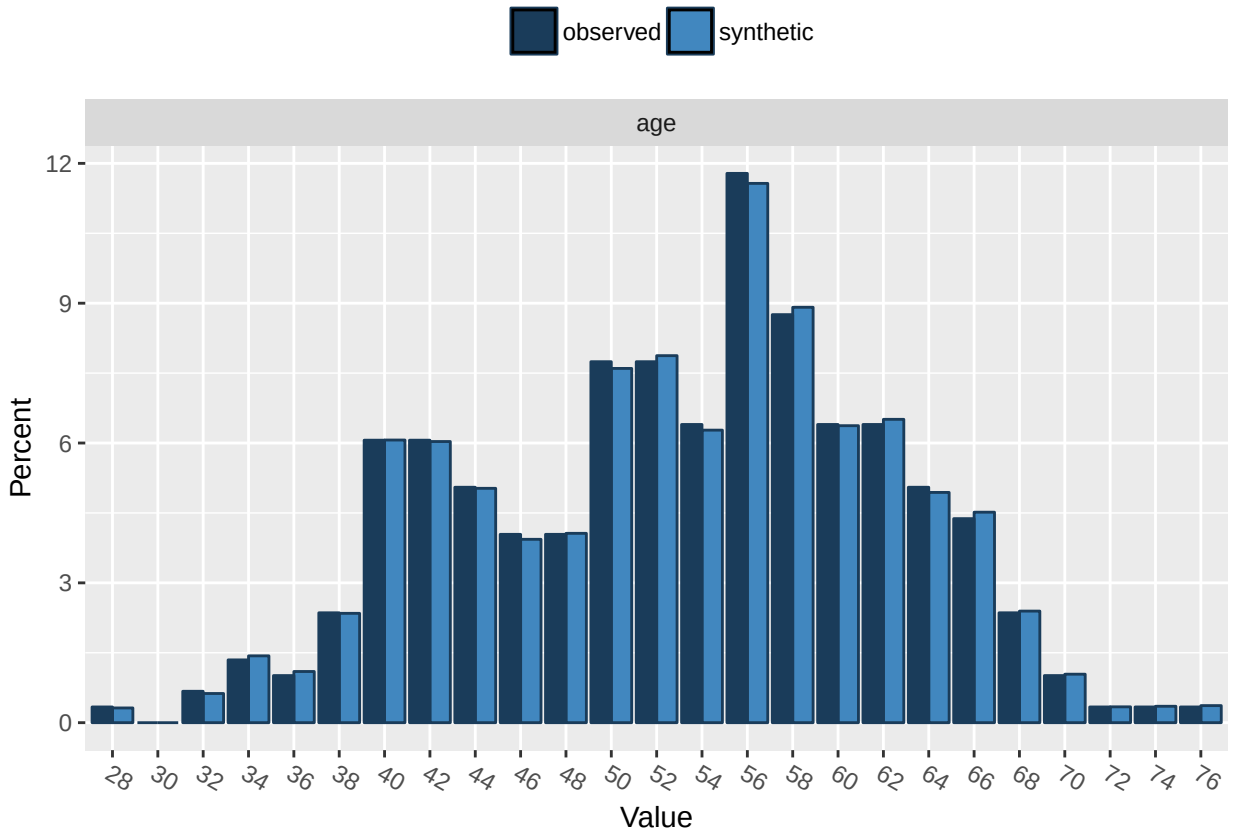
Synthesizing Heart Disease Data for Machine Learning

In many data applications especially in the medical field, data often comes in small samples and always subject the HIPPAA Privacy Laws. Small data samples often don't meet Machine Learning needs due to its lack of volume.

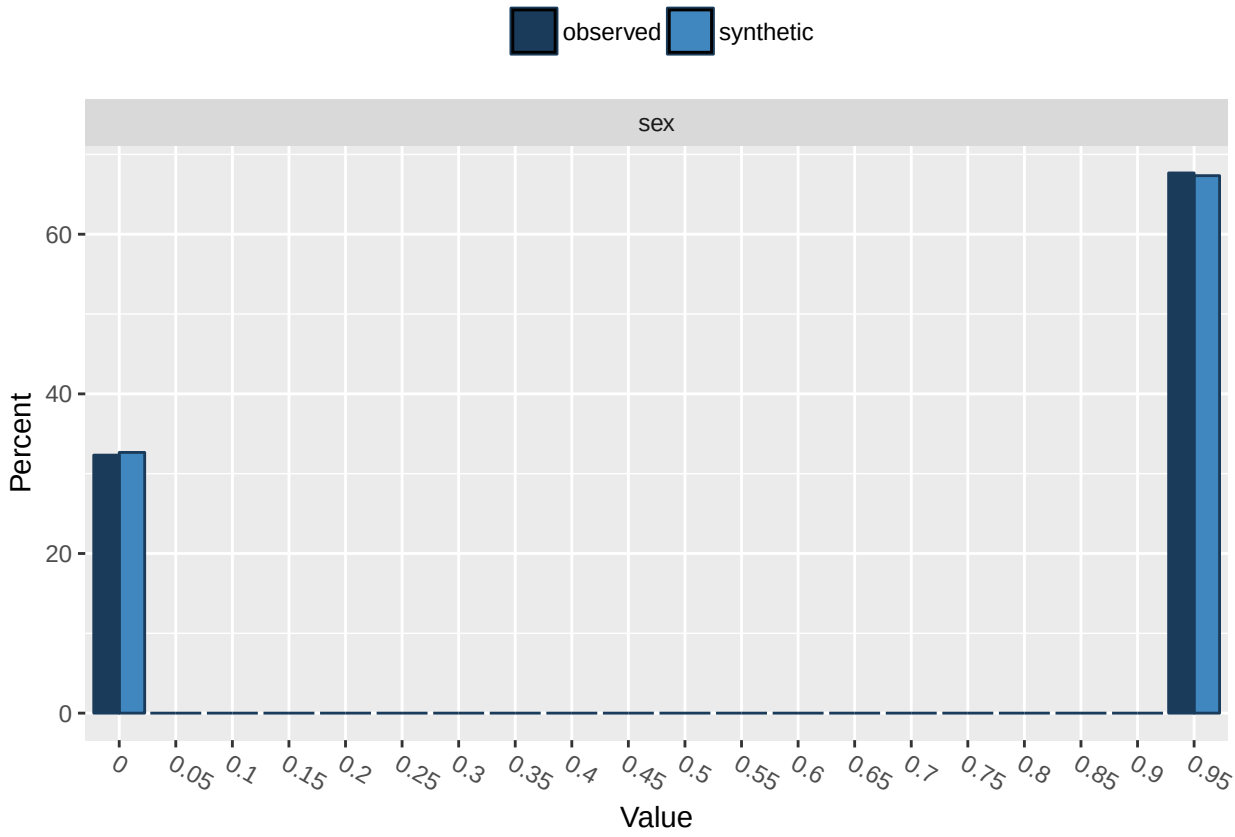
Using the Cleveland dataset (from http://archive.ics.uci.edu/ml/datasets/heart+disease), this example shows how we can use the "synthpop" R package to produce anonymized data in volume for research purposes. Although the resulting synthesized data matches the "real" data characteristics, it must be clearly understood that the resulting Synthesized Data is not "real data" and use should be limited to research studies (for ex: Machine and or Deep Learning).

Plots

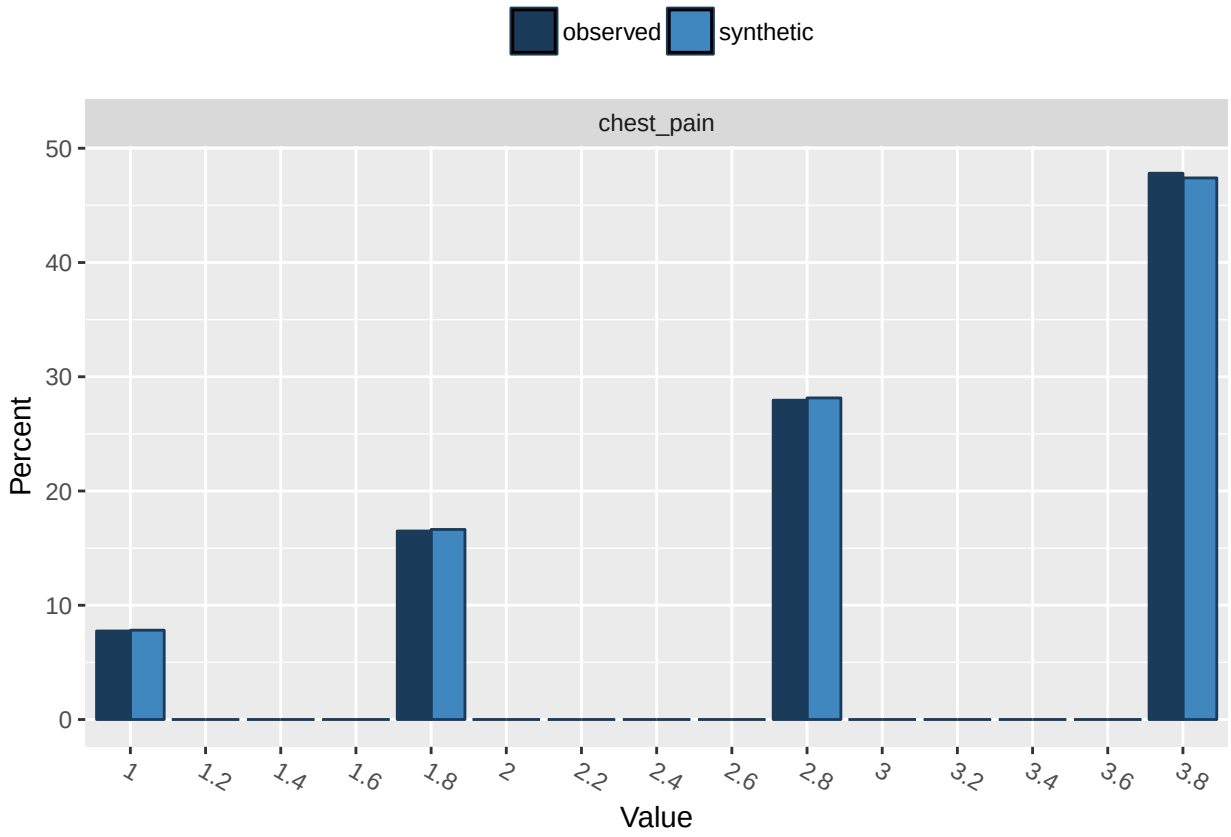
```
##
## Comparing percentages observed with synthetic
##
## $age
##                   28 30        32       34       36       38       40
## observed  0.3367003  0 0.6734007 1.346801 1.010101 2.356902 6.060606
## synthetic 0.3160000  0 0.6260000 1.434000 1.098000 2.346000 6.064000
##                   42       44       46       48       50       52       54
## observed  6.060606 5.050505 4.040404 4.040404 7.744108 7.744108 6.397306
## synthetic 6.032000 5.028000 3.934000 4.062000 7.600000 7.874000 6.276000
##                   56       58       60       62       64       66       68
## observed  11.78451 8.754209 6.397306 6.397306 5.050505 4.377104 2.356902
## synthetic 11.57000 8.912000 6.372000 6.508000 4.940000 4.516000 2.394000
##                   70        72        74        76
## observed  1.010101 0.3367003 0.3367003 0.3367003
## synthetic 1.040000 0.3400000 0.3520000 0.3660000
```



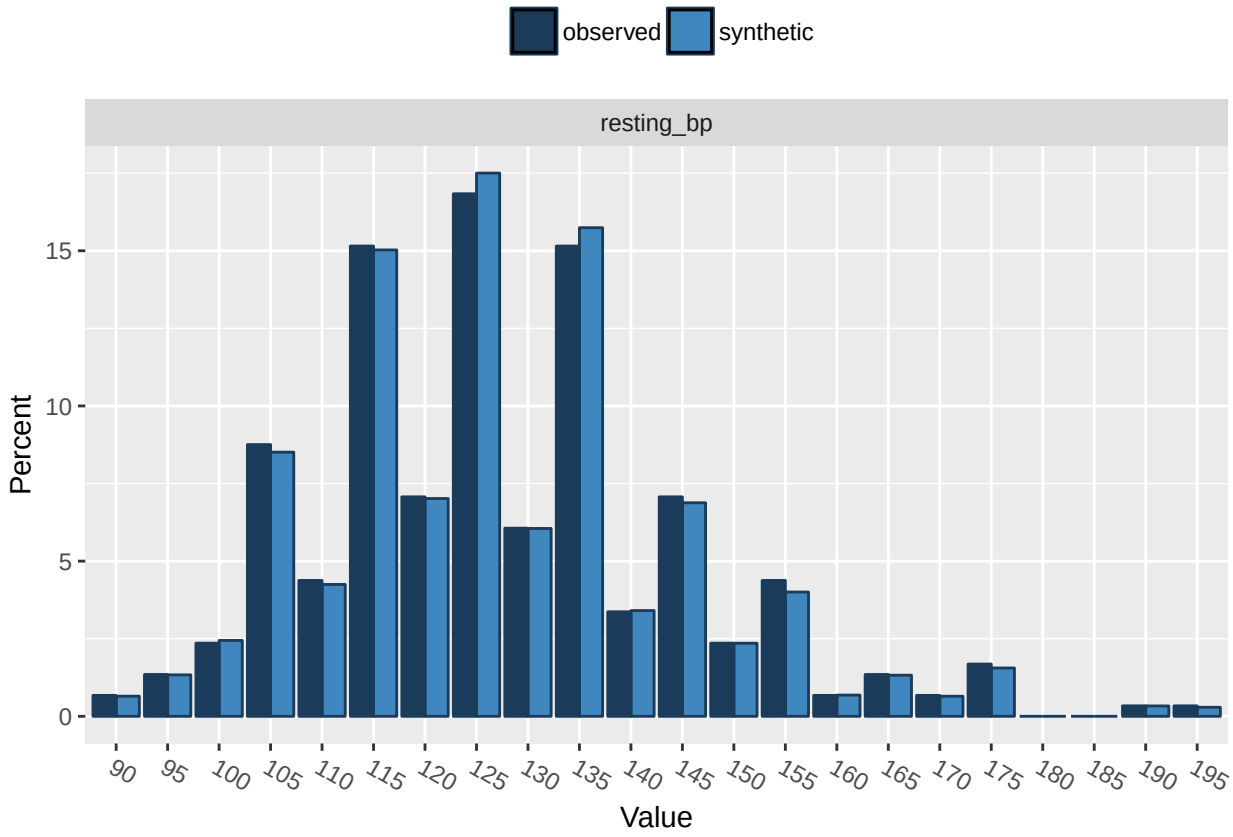

```
##
## Comparing percentages observed with synthetic
##
## $sex
##                  0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6
## observed  32.32323    0   0    0   0    0   0    0   0    0   0    0   0
## synthetic 32.66200    0   0    0   0    0   0    0   0    0   0    0   0
##           0.65 0.7 0.75 0.8 0.85 0.9     0.95
## observed     0   0    0   0    0   0 67.67677
## synthetic    0   0    0   0    0   0 67.33800
```



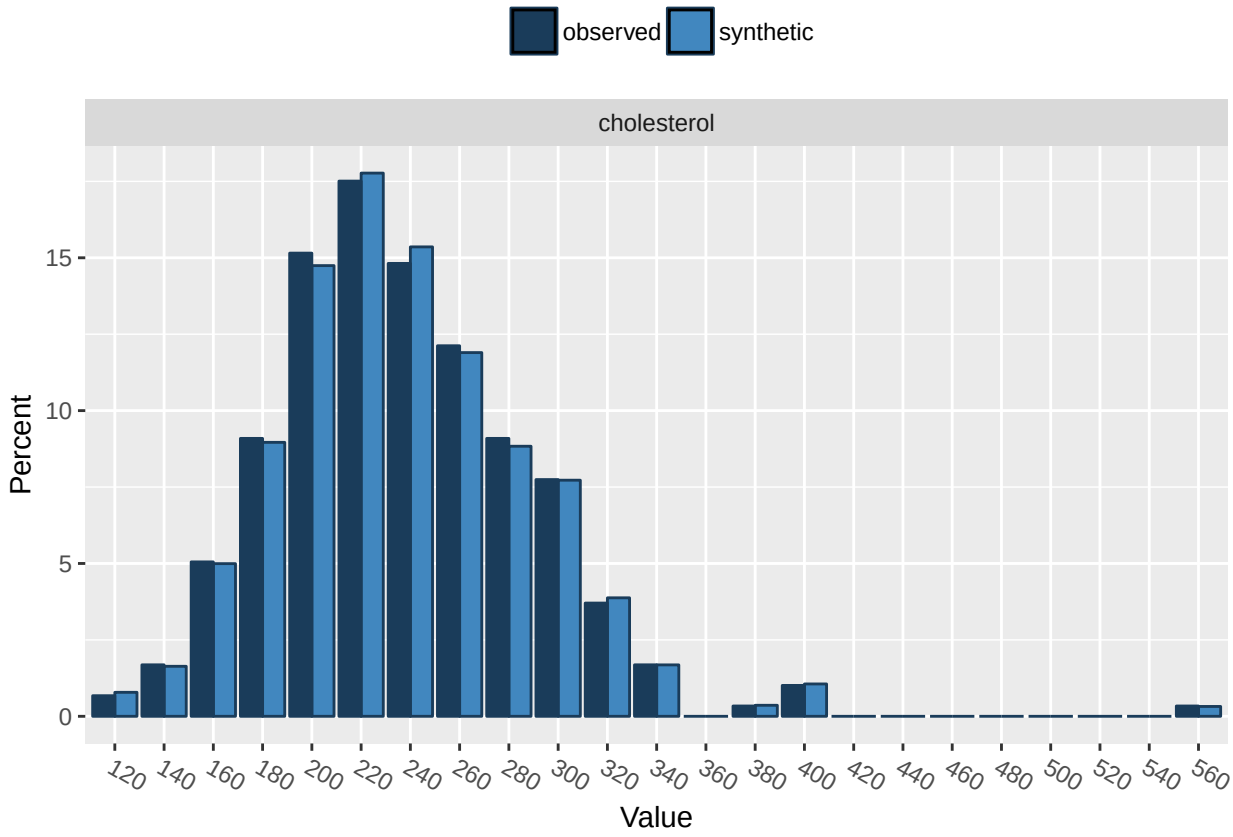

```
##
## Comparing percentages observed with synthetic
##
## $chest_pain
##                   1 1.2 1.4 1.6      1.8 2 2.2 2.4 2.6      2.8 3 3.2 3.4
## observed  7.744108   0   0   0 16.49832 0   0   0   0 27.94613 0   0   0
## synthetic 7.818000   0   0   0 16.63200 0   0   0   0 28.14600 0   0   0
##           3.6      3.8
## observed    0 47.81145
## synthetic   0 47.40400
```



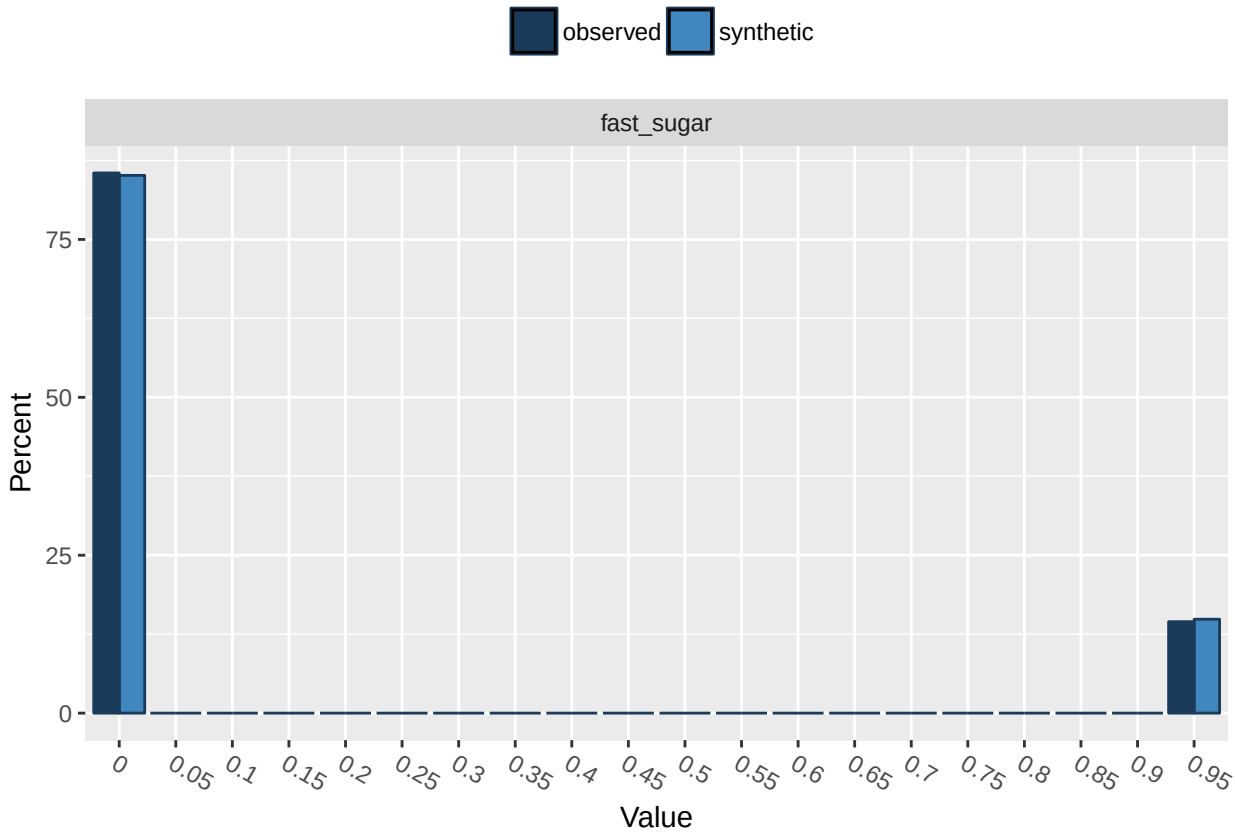

```
##
## Comparing percentages observed with synthetic
##
## $resting_bp
##                   90       95      100      105      110      115      120
## observed  0.6734007 1.346801 2.356902 8.754209 4.377104 15.15152 7.070707
## synthetic 0.6480000 1.334000 2.444000 8.510000 4.248000 15.02600 7.016000
##                  125      130      135      140      145      150      155
## observed  16.83502 6.060606 15.15152 3.367003 7.070707 2.356902 4.377104
## synthetic 17.50200 6.052000 15.74200 3.406000 6.880000 2.354000 4.004000
##                  160       165       170      175 180 185       190
## observed  0.6734007 1.346801 0.6734007 1.683502   0   0 0.3367003
## synthetic 0.6840000 1.324000 0.6460000 1.556000   0   0 0.3340000
##                  195
## observed  0.3367003
## synthetic 0.2900000
```



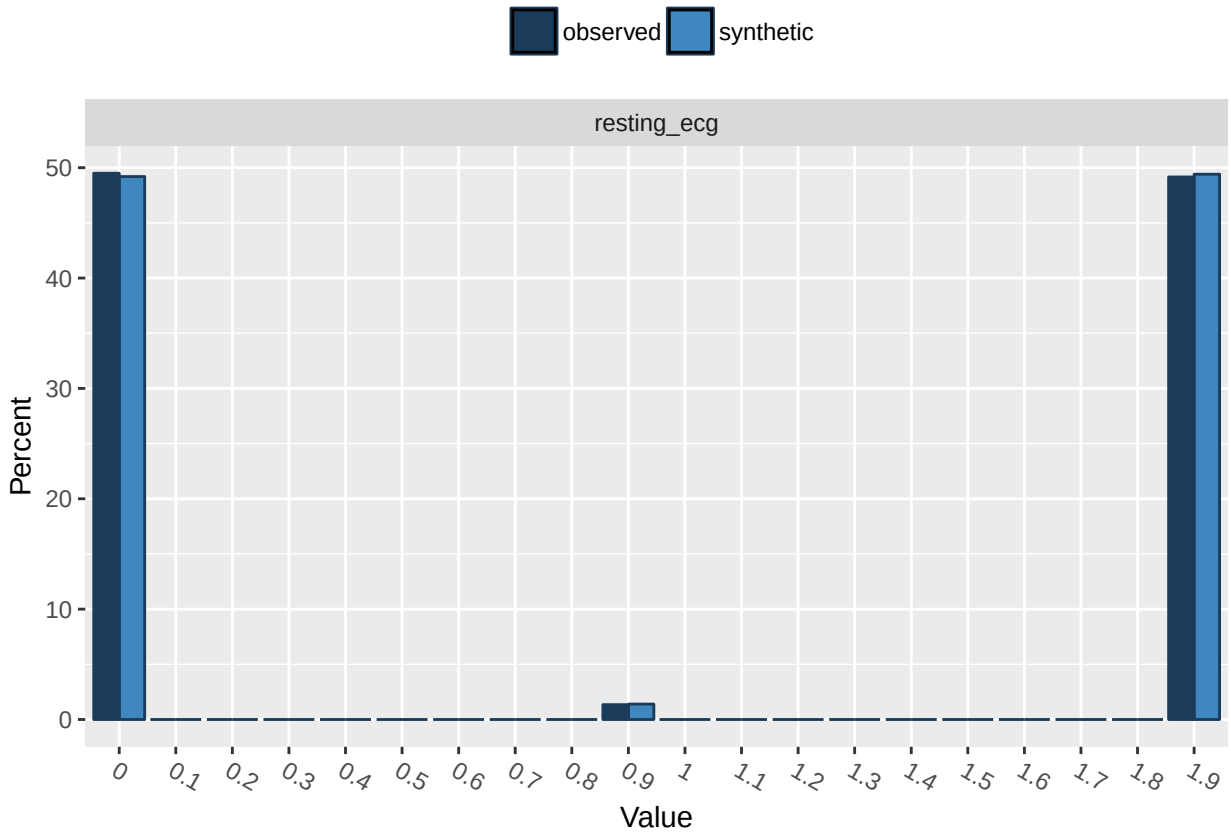

```
##
## Comparing percentages observed with synthetic
##
## $cholesterol
##                  120      140      160      180      200      220      240
## observed  0.6734007 1.683502 5.050505 9.090909 15.15152 17.50842 14.81481
## synthetic 0.7840000 1.636000 4.994000 8.962000 14.74400 17.77000 15.35600
##                  260      280      300      320      340 360       380
## observed  12.12121 9.090909 7.744108 3.703704 1.683502   0 0.3367003
## synthetic 11.89800 8.834000 7.724000 3.876000 1.682000   0 0.3600000
##                  400 420 440 460 480 500 520 540       560
## observed  1.010101   0   0   0   0   0   0   0 0.3367003
## synthetic 1.058000   0   0   0   0   0   0   0 0.3220000
```



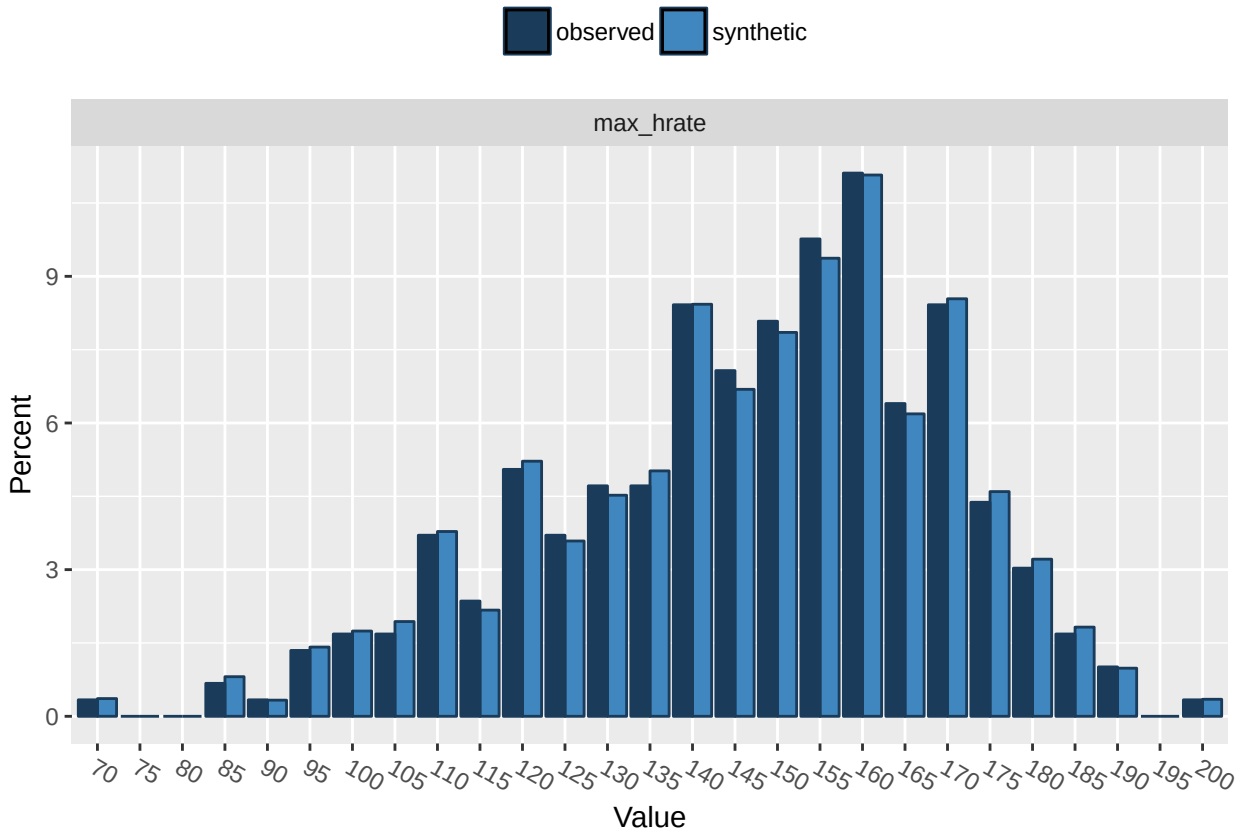

```
##
## Comparing percentages observed with synthetic
##
## $fast_sugar
##                  0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6
## observed  85.52189    0   0    0   0    0   0    0   0    0   0    0   0
## synthetic 85.13600    0   0    0   0    0   0    0   0    0   0    0   0
##           0.65 0.7 0.75 0.8 0.85 0.9     0.95
## observed     0   0    0   0    0   0 14.47811
## synthetic    0   0    0   0    0   0 14.86400
```



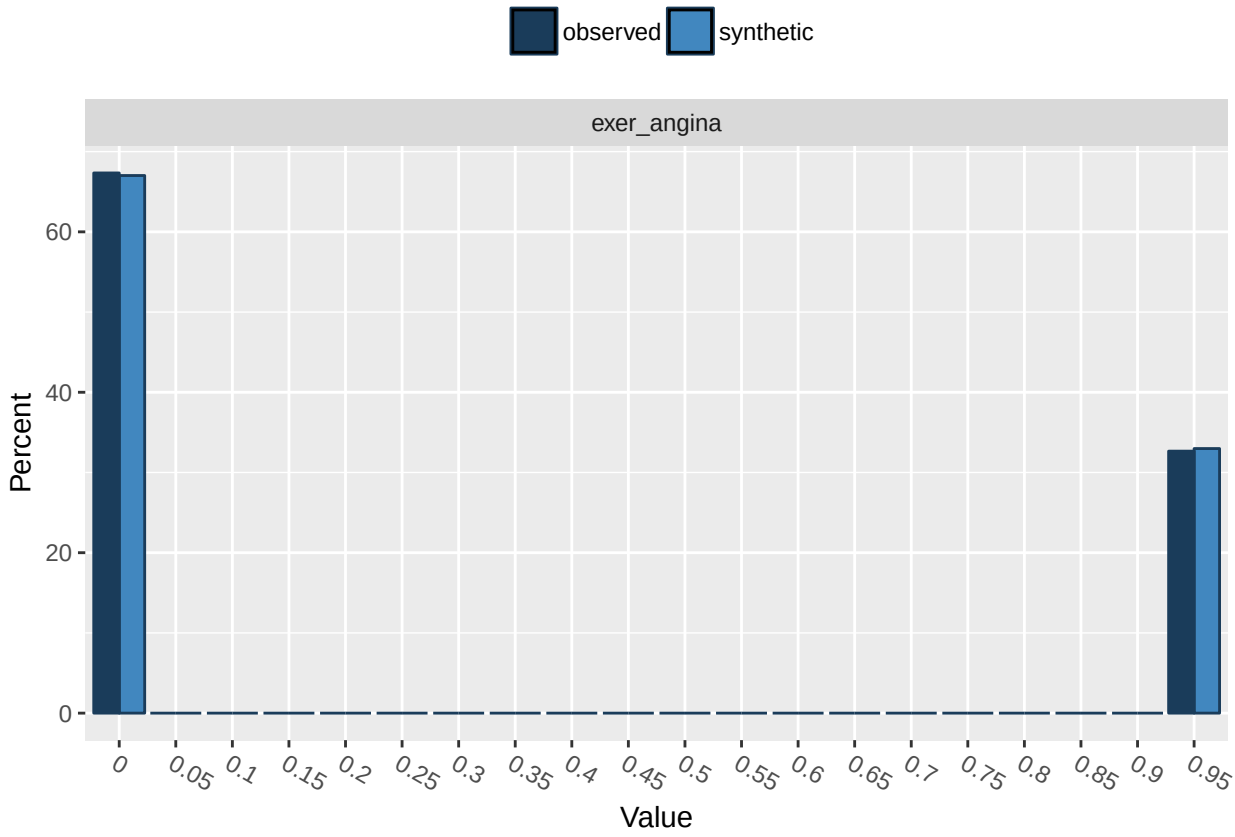

```
##
## Comparing percentages observed with synthetic
##
## $resting_ecg
##                  0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8       0.9 1 1.1 1.2 1.3
## observed  49.49495   0   0   0   0   0   0   0   0 1.346801 0   0   0   0
## synthetic 49.19600   0   0   0   0   0   0   0   0 1.396000 0   0   0   0
##           1.4 1.5 1.6 1.7 1.8      1.9
## observed    0   0   0   0   0 49.15825
## synthetic   0   0   0   0   0 49.40800
```



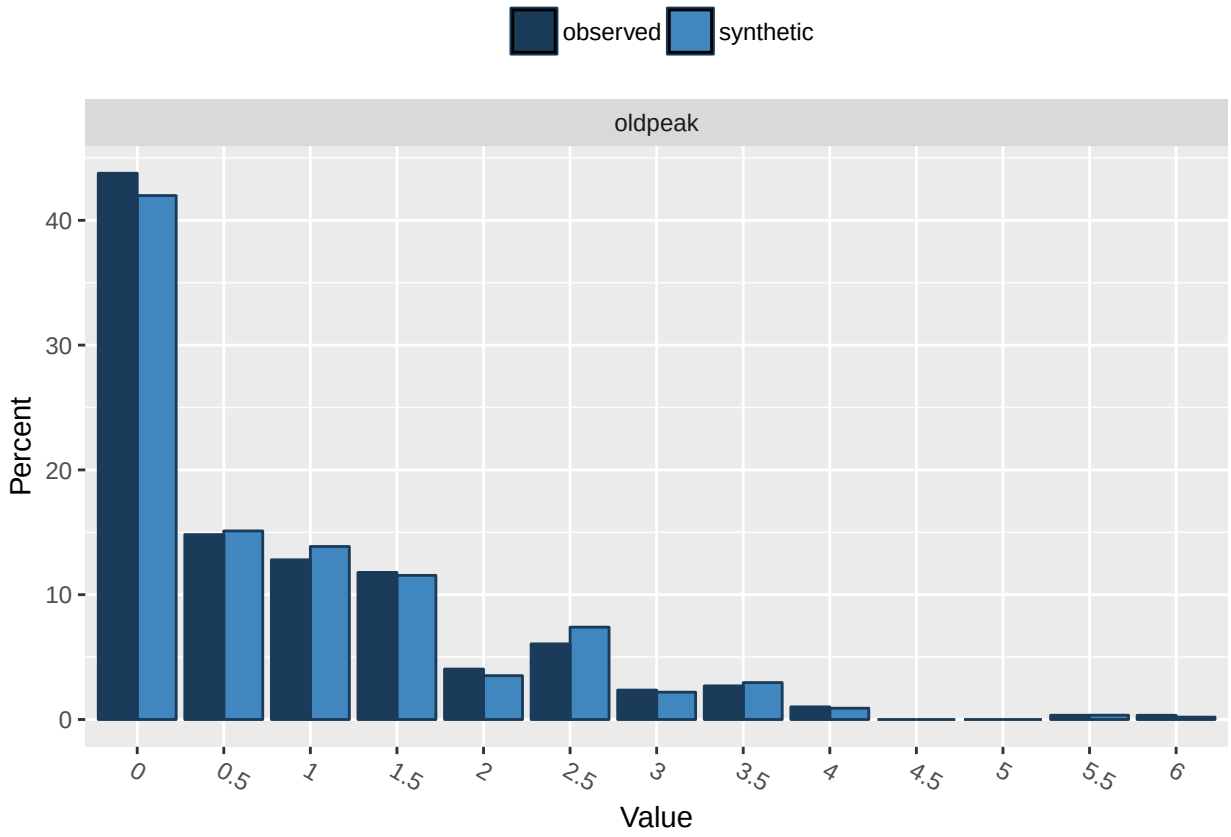

```
##
## Comparing percentages observed with synthetic
##
## $max_hrate
##                   70 75 80        85        90       95       100       105
## observed  0.3367003  0  0 0.6734007 0.3367003 1.346801 1.683502 1.683502
## synthetic 0.3620000  0  0 0.8100000 0.3300000 1.416000 1.744000 1.938000
##                110      115      120      125      130      135      140
## observed  3.703704 2.356902 5.050505 3.703704 4.713805 4.713805 8.417508
## synthetic 3.780000 2.172000 5.218000 3.586000 4.522000 5.020000 8.428000
##                145      150     155      160      165      170      175
## observed  7.070707 8.080808 9.76431 11.11111 6.397306 8.417508 4.377104
## synthetic 6.688000 7.852000 9.37000 11.07200 6.186000 8.540000 4.596000
##                180      185      190 195       200
## observed  3.030303 1.683502 1.010101   0 0.3367003
## synthetic 3.214000 1.824000 0.984000   0 0.3480000
```



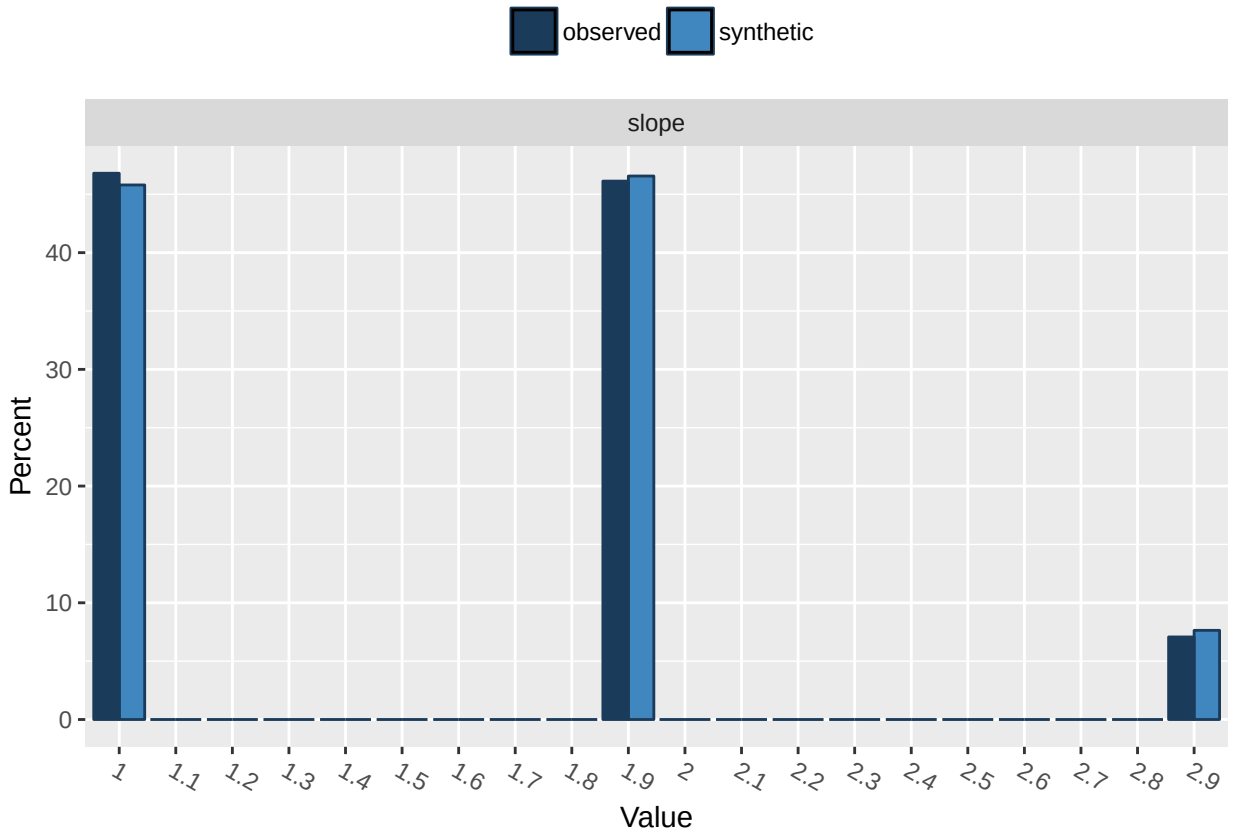

```
##
## Comparing percentages observed with synthetic
##
## $exer_angina
##                   0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6
## observed  67.34007    0   0    0   0    0   0    0   0    0   0    0   0
## synthetic 67.02000    0   0    0   0    0   0    0   0    0   0    0   0
##           0.65 0.7 0.75 0.8 0.85 0.9     0.95
## observed     0   0    0   0    0   0 32.65993
## synthetic    0   0    0   0    0   0 32.98000
```



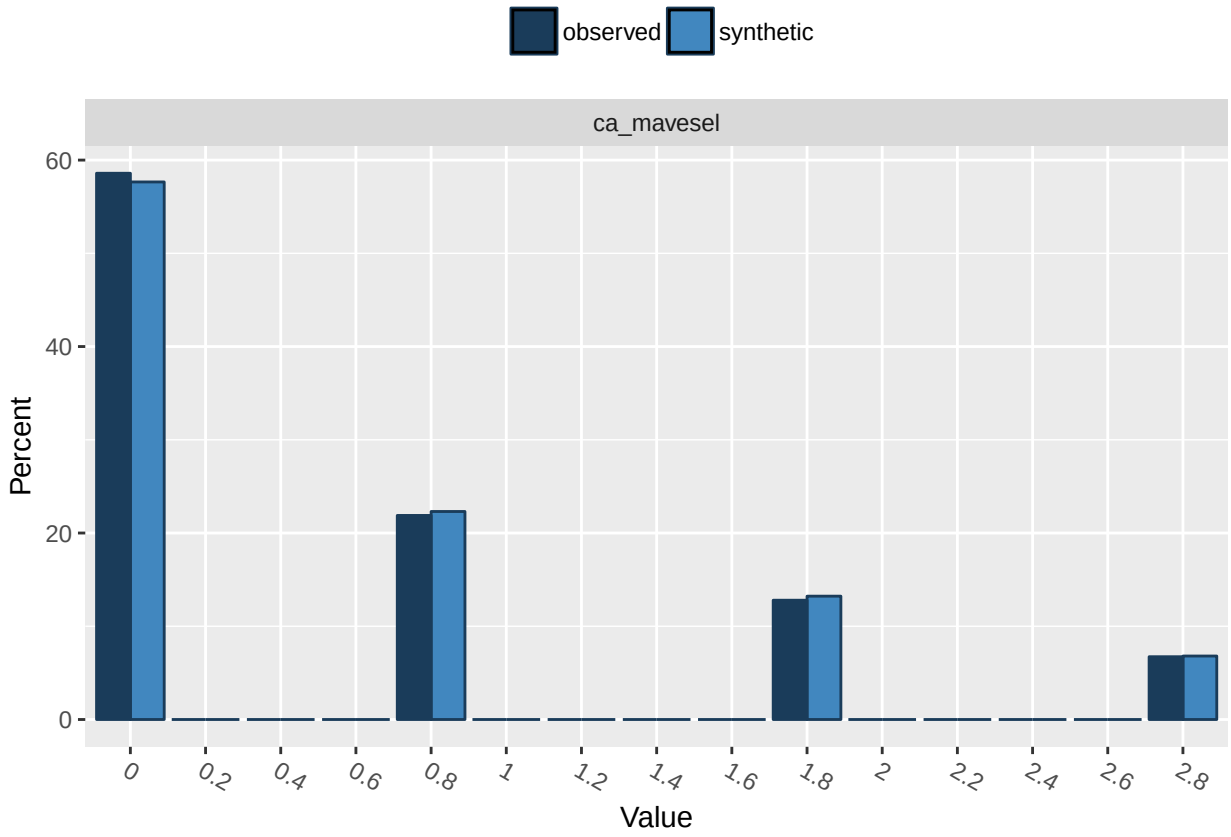

```
##
## Comparing percentages observed with synthetic
##
## $oldpeak
##                   0      0.5        1      1.5        2      2.5        3
## observed  43.77104 14.81481 12.79461 11.78451 4.040404 6.060606 2.356902
## synthetic 41.98200 15.11000 13.86400 11.54800 3.512000 7.400000 2.188000
##                 3.5        4 4.5 5       5.5         6
## observed  2.693603 1.010101   0 0 0.3367003 0.3367003
## synthetic 2.958000 0.904000   0 0 0.3440000 0.1900000
```



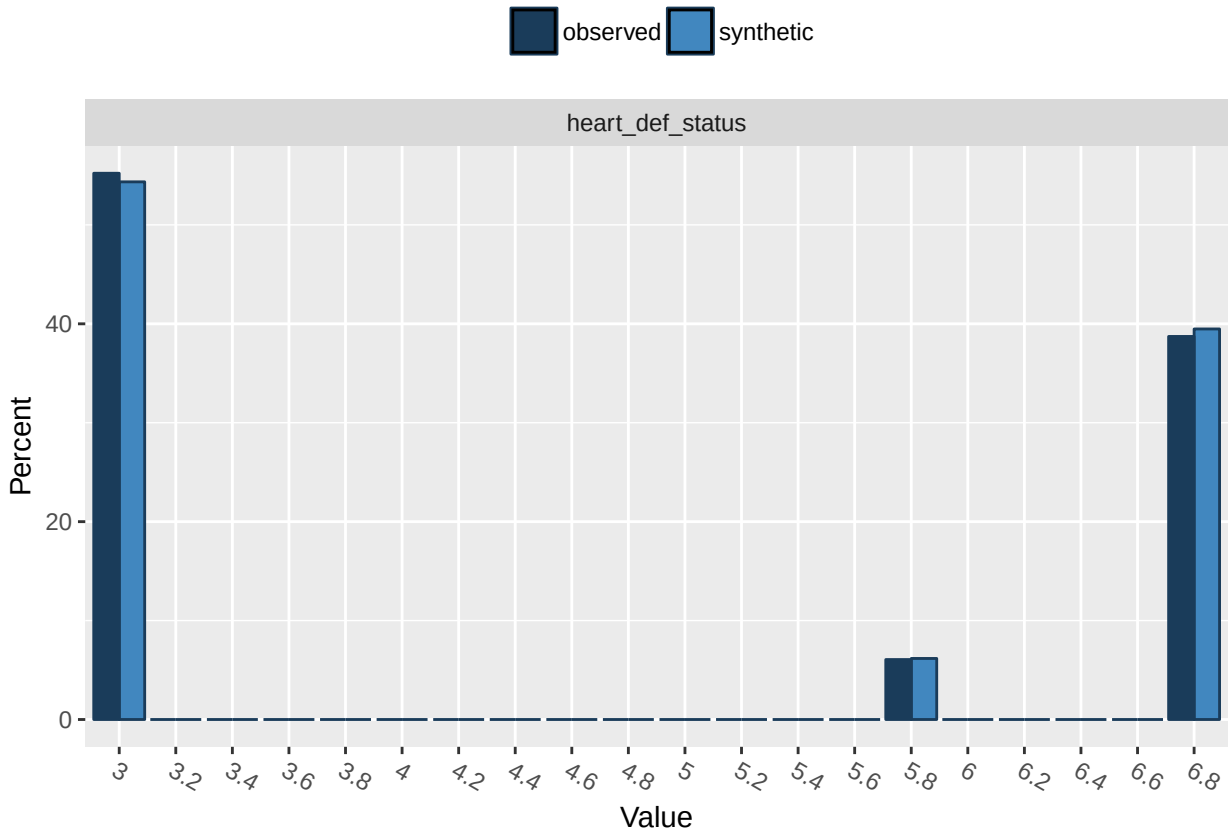

```
##
## Comparing percentages observed with synthetic
##
## $slope
##                  1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8      1.9 2 2.1 2.2 2.3
## observed  46.80135   0   0   0   0   0   0   0   0 46.12795 0   0   0   0
## synthetic 45.79800   0   0   0   0   0   0   0   0 46.56400 0   0   0   0
##           2.4 2.5 2.6 2.7 2.8      2.9
## observed    0   0   0   0   0 7.070707
## synthetic   0   0   0   0   0 7.638000
```



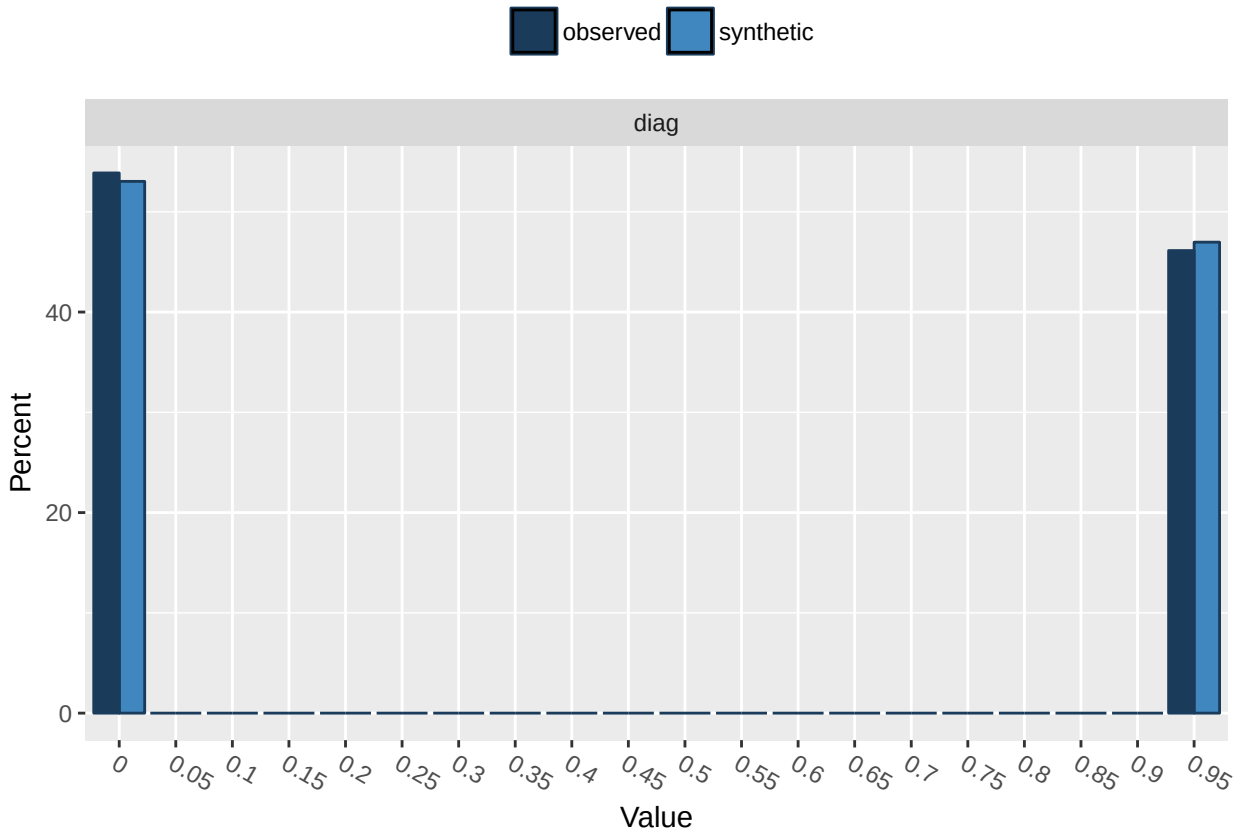

```
##
## Comparing percentages observed with synthetic
##
## $ca_mavesel
##                  0 0.2 0.4 0.6      0.8 1 1.2 1.4 1.6      1.8 2 2.2 2.4
## observed  58.58586   0   0   0 21.88552 0   0   0   0 12.79461 0   0   0
## synthetic 57.65400   0   0   0 22.31000 0   0   0   0 13.22800 0   0   0
##           2.6      2.8
## observed    0 6.734007
## synthetic   0 6.808000
```




```
##
## Comparing percentages observed with synthetic
##
## $heart_def_status
##                   3 3.2 3.4 3.6 3.8 4 4.2 4.4 4.6 4.8 5 5.2 5.4 5.6
## observed  55.21886   0   0   0   0 0   0   0   0   0 0   0   0   0
## synthetic 54.35000   0   0   0   0 0   0   0   0   0 0   0   0   0
##                 5.8 6 6.2 6.4 6.6      6.8
## observed  6.060606 0   0   0   0 38.72054
## synthetic 6.170000 0   0   0   0 39.48000
```




```
##
## Comparing percentages observed with synthetic
##
## $diag
##                  0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6
## observed  53.87205    0   0    0   0    0   0    0   0    0   0    0   0
## synthetic 53.03000    0   0    0   0    0   0    0   0    0   0    0   0
##           0.65 0.7 0.75 0.8 0.85 0.9     0.95
## observed     0   0    0   0    0   0 46.12795
## synthetic    0   0    0   0    0   0 46.97000
```


observed
synthetic
diag
Percent
40
20
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
Value