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 the Machine Learning needs.

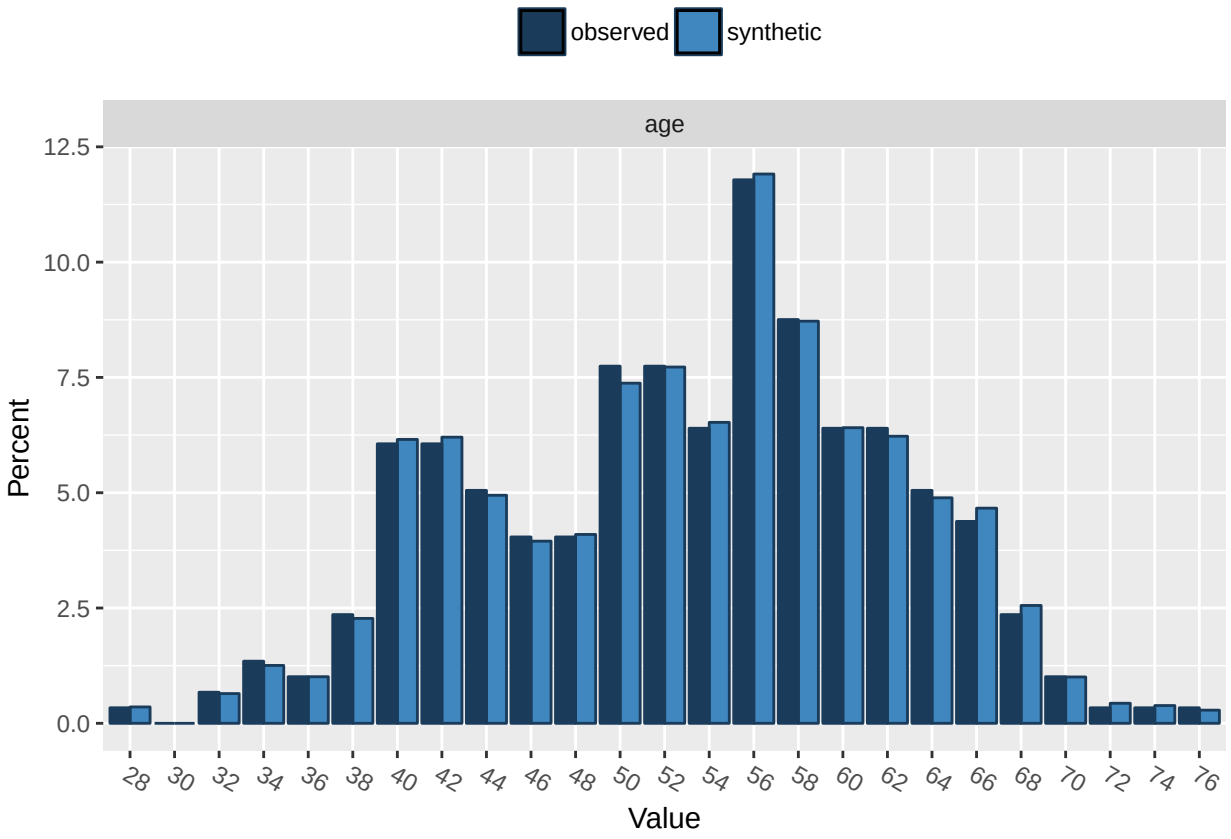
This example shows how we can use the "synthpop" R package to produce anonymized data in volume for research purposes. Although the resulting synthesized data matches the "real" data characteristics, it must be clearly understood that it is not "real data" and use should only be limited to research studies using Machine and or Deep Learning.

Including Plots

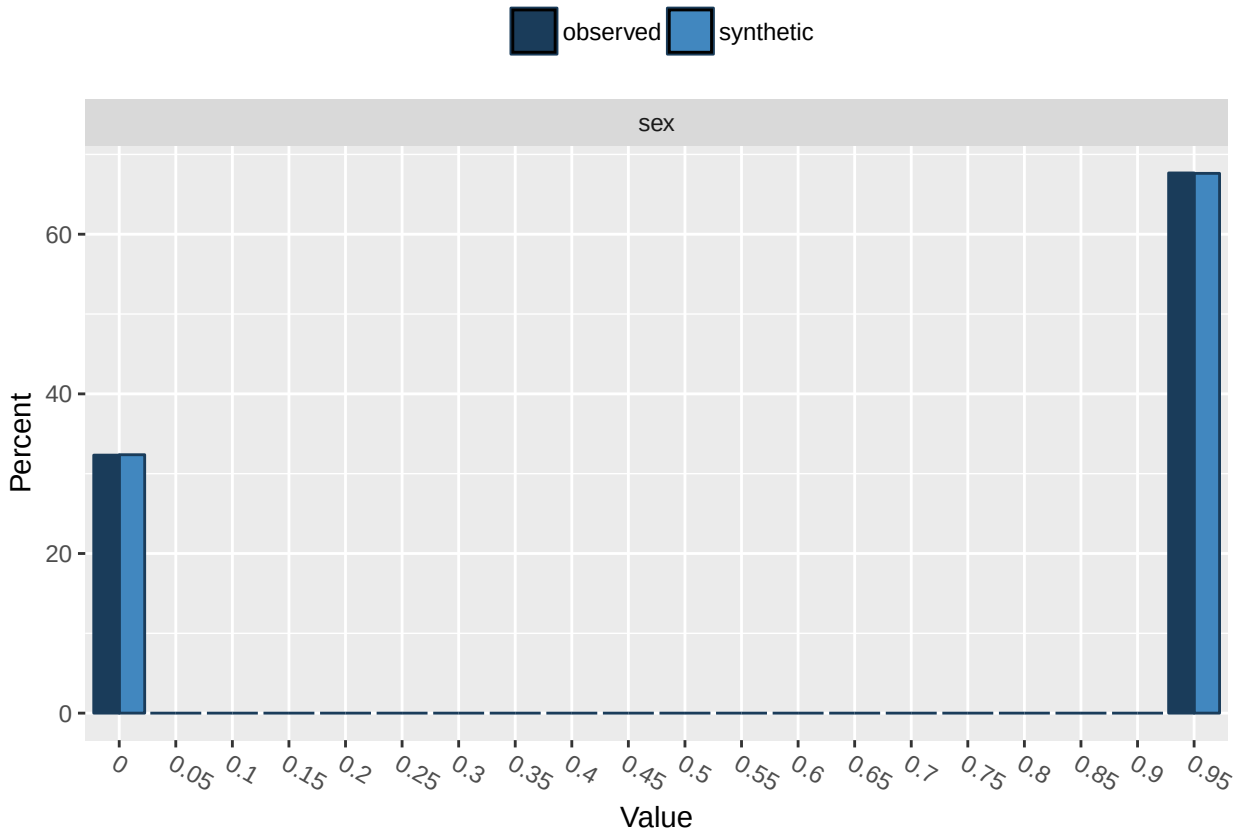
```
## Loading required package: lattice

## Loading required package: MASS

## Loading required package: nnet

## Loading required package: ggplot2

## Sample(s) of size 20000 will be generated from original data of size 297.
##
## syn  variables
## 1    age sex chest_pain resting_bp cholesterol fast_sugar resting_ecg max_hrate exer_angina oldpeak
##      slope ca_mavesel heart_def_status diag

##
## Call:
## glm(formula = diag ~ age + sex + chest_pain + resting_bp + cholesterol +
##     fast_sugar + resting_ecg + max_hrate + exer_angina + oldpeak +
##     slope + ca_mavesel + heart_def_status, family = "binomial",
##     data = df_syn)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0928  -0.6699  -0.2810   0.6575   2.7497
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -8.2635117  0.3087704 -26.763  < 2e-16 ***
## age              0.0282341  0.0024588  11.483  < 2e-16 ***
## sex              0.3694125  0.0458329   8.060 7.63e-16 ***
## chest_pain       0.6689440  0.0226617  29.519  < 2e-16 ***
## resting_bp       0.0089361  0.0011399   7.840 4.52e-15 ***
## cholesterol      0.0003772  0.0003817   0.988 0.323121
## fast_sugar      -0.0815213  0.0543808  -1.499 0.133852
## resting_ecg      0.0244078  0.0196284   1.243 0.213686
## max_hrate       -0.0012778  0.0010336  -1.236 0.216354
## exer_angina      0.1689618  0.0457476   3.693 0.000221 ***
```

```
## oldpeak           0.4270630  0.0221591  19.273  < 2e-16 ***
## slope             0.1027812  0.0379058   2.711 0.006698 **
## ca_mavesel        0.5604028  0.0228024  24.576  < 2e-16 ***
## heart_def_status  0.4388169  0.0104871  41.843  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27688  on 19999  degrees of freedom
## Residual deviance: 17842  on 19986  degrees of freedom
## AIC: 17870
##
## Number of Fisher Scoring iterations: 5

## Warning: Note that all these results depend on the synthesis model being correct.
##
## Fit to synthetic data set with a single synthesis.
## Inference to coefficients and standard errors that
## would be obtained from the observed data.
##
## Call:
## glm.synds(formula = diag ~ age + sex + chest_pain + resting_bp +
##     cholesterol + fast_sugar + resting_ecg + max_hrate + exer_angina +
##     oldpeak + slope + ca_mavesel + heart_def_status, family = "binomial",
##     data = syn_cleveland)
##
## Combined estimates:
##                   xpct(Beta) xpct(se.Beta) xpct(z) Pr(>|xpct(z)|)
## (Intercept)      -8.26351170    2.53380020 -3.2613      0.0011090 **
## age               0.02823405    0.02017738  1.3993      0.1617253
## sex               0.36941252    0.37610954  0.9822      0.3260043
## chest_pain        0.66894402    0.18596389  3.5972      0.0003217 ***
## resting_bp        0.00893609    0.00935386  0.9553      0.3394071
## cholesterol       0.00037716    0.00313241  0.1204      0.9041614
## fast_sugar       -0.08152135    0.44625422 -0.1827      0.8550498
## resting_ecg       0.02440780    0.16107261  0.1515      0.8795553
## max_hrate        -0.00127782    0.00848180 -0.1507      0.8802487
## exer_angina       0.16896180    0.37540960  0.4501      0.6526577
## oldpeak           0.42706300    0.18183980  2.3486      0.0188458 *
## slope             0.10278122    0.31105904  0.3304      0.7410800
## ca_mavesel        0.56040279    0.18711910  2.9949      0.0027454 **
## heart_def_status  0.43881695    0.08605831  5.0991      3.413e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Sample(s) of size 20000 will be generated from original data of size 297.
##
## syn  variables
## 1    age sex chest_pain resting_bp cholesterol fast_sugar resting_ecg max_hrate exer_angina oldpeak
##      slope ca_mavesel heart_def_status diag

##
## Comparing percentages observed with synthetic
##
```

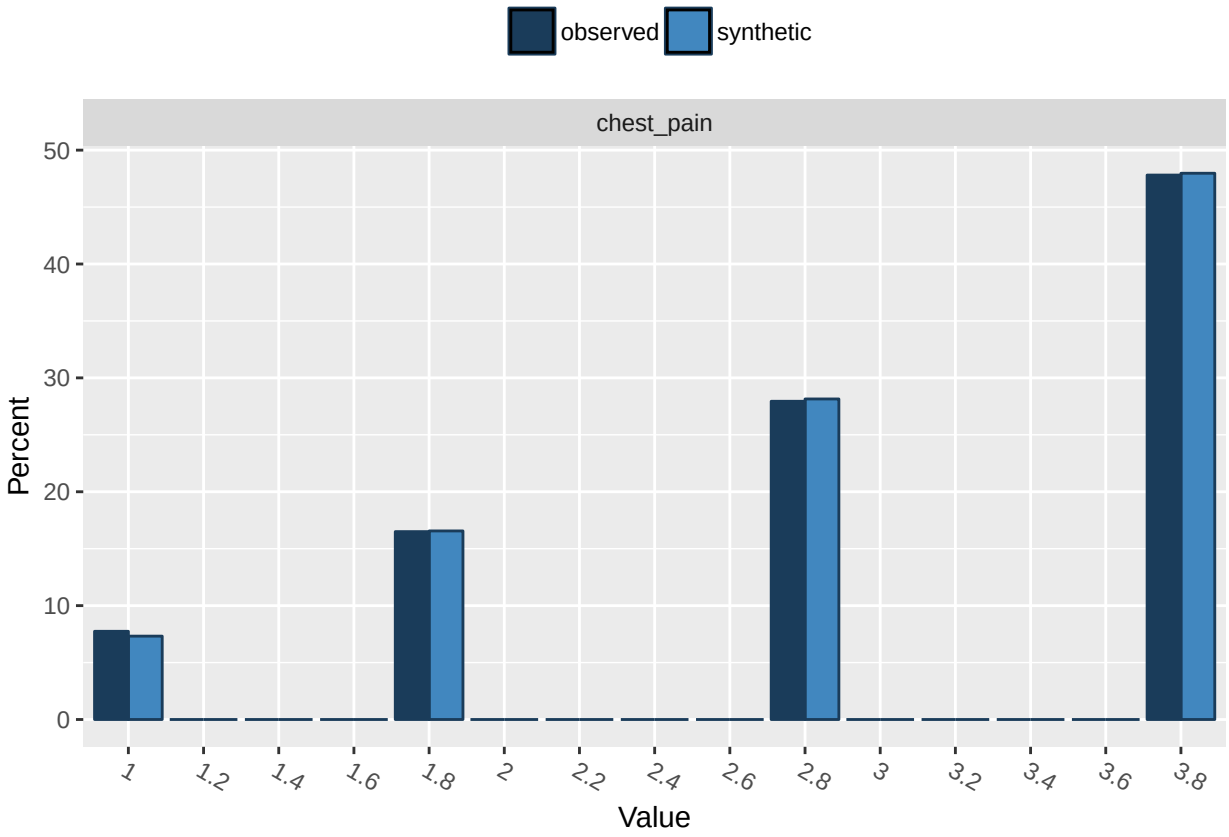
```
## $age
##                   28 30        32       34       36       38       40
## observed  0.3367003  0 0.6734007 1.346801 1.010101 2.356902 6.060606
## synthetic 0.3550000  0 0.6450000 1.255000 1.010000 2.275000 6.155000
##                 42       44       46       48       50       52       54
## observed  6.060606 5.050505 4.040404 4.040404 7.744108 7.744108 6.397306
## synthetic 6.205000 4.945000 3.950000 4.095000 7.375000 7.725000 6.525000
##                 56       58       60       62       64       66       68
## observed  11.78451 8.754209 6.397306 6.397306 5.050505 4.377104 2.356902
## synthetic 11.91000 8.720000 6.410000 6.225000 4.890000 4.665000 2.555000
##                 70        72        74        76
## observed  1.010101 0.3367003 0.3367003 0.3367003
## synthetic 1.005000 0.4350000 0.3850000 0.2850000
```



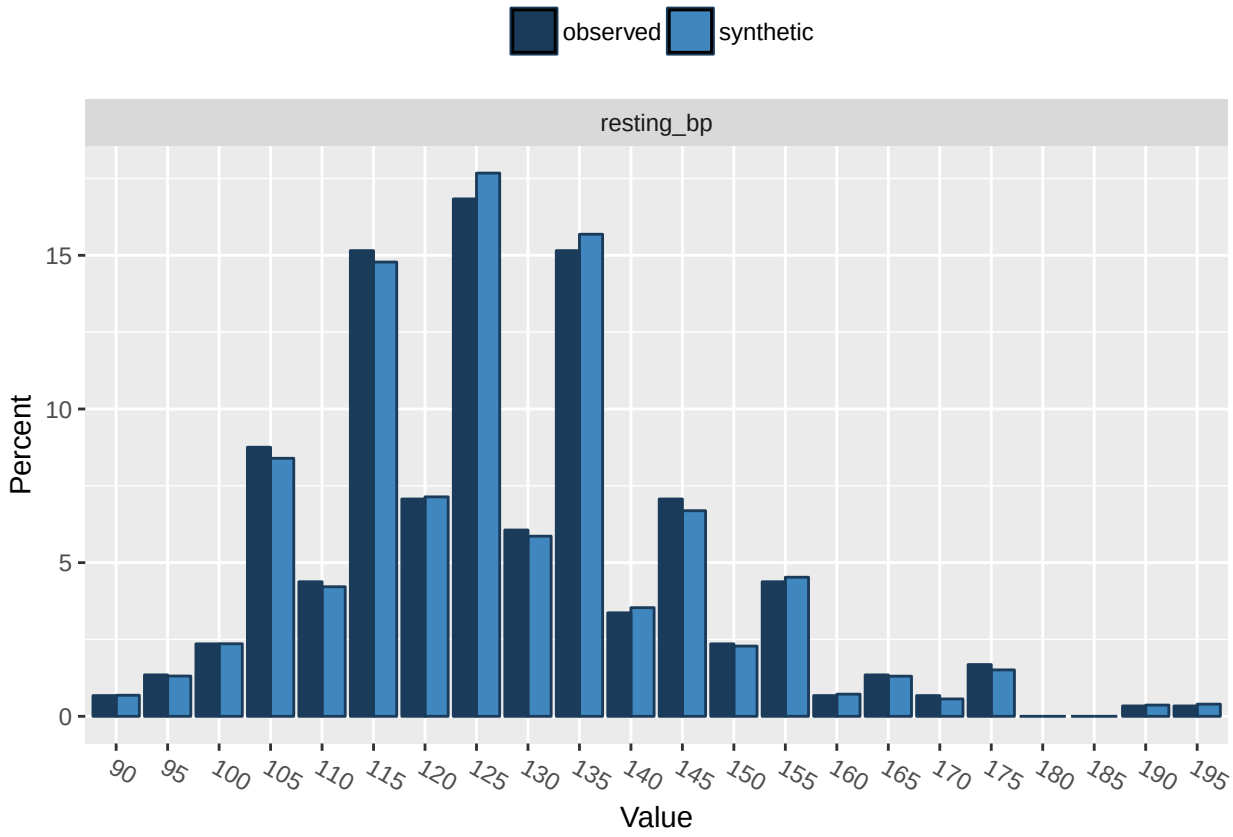

```
##
## Comparing percentages observed with synthetic
##
## $sex
##                   0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6
## observed  32.32323    0   0    0   0    0   0    0   0    0   0    0   0
## synthetic 32.37500    0   0    0   0    0   0    0   0    0   0    0   0
##           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 67.62500
```



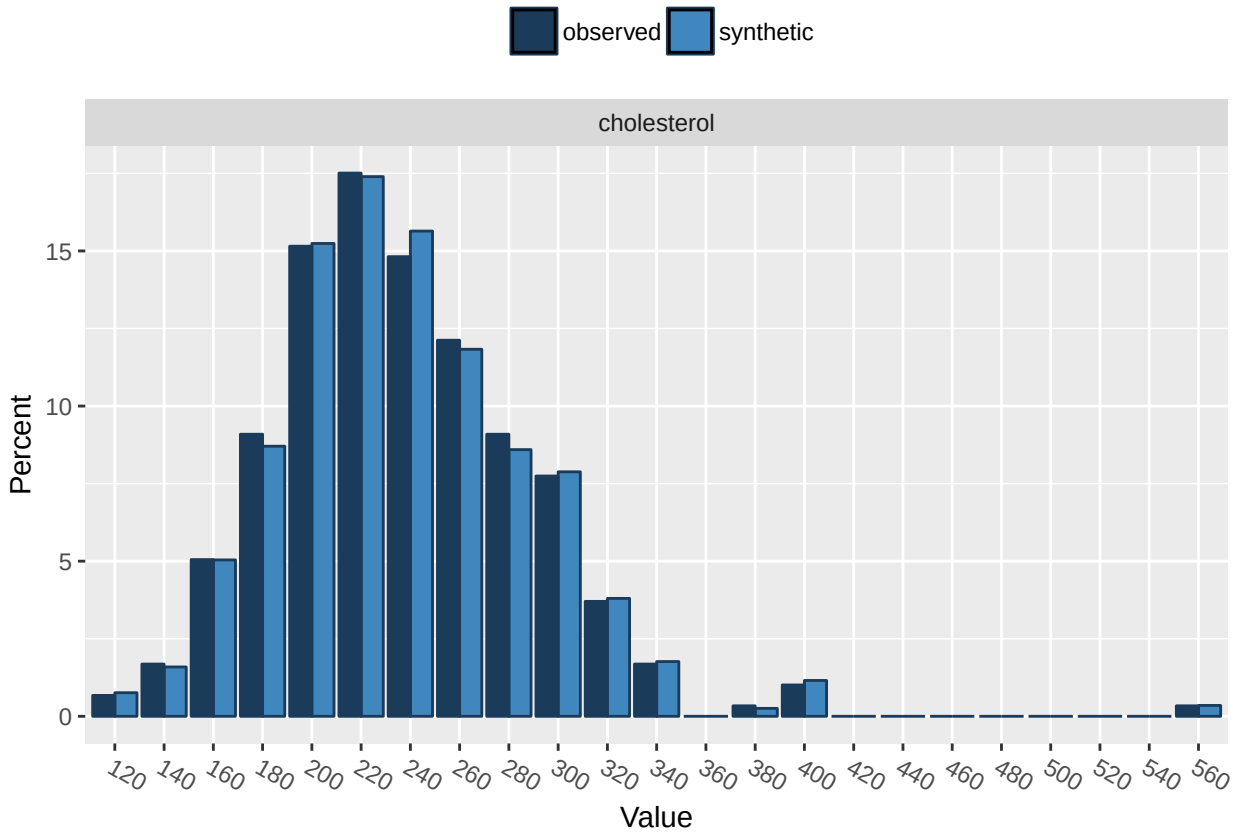

```
##
## Comparing percentages observed with synthetic
##
## $chest_pain
##                   1 1.2 1.4 1.6      1.8 2 2.2 2.4 2.6      2.8 3 3.2 3.4
## observed  7.744108   0   0   0 16.49832 0   0   0   0 27.94613 0   0   0
## synthetic 7.320000   0   0   0 16.56000 0   0   0   0 28.15000 0   0   0
##           3.6      3.8
## observed    0 47.81145
## synthetic   0 47.97000
```



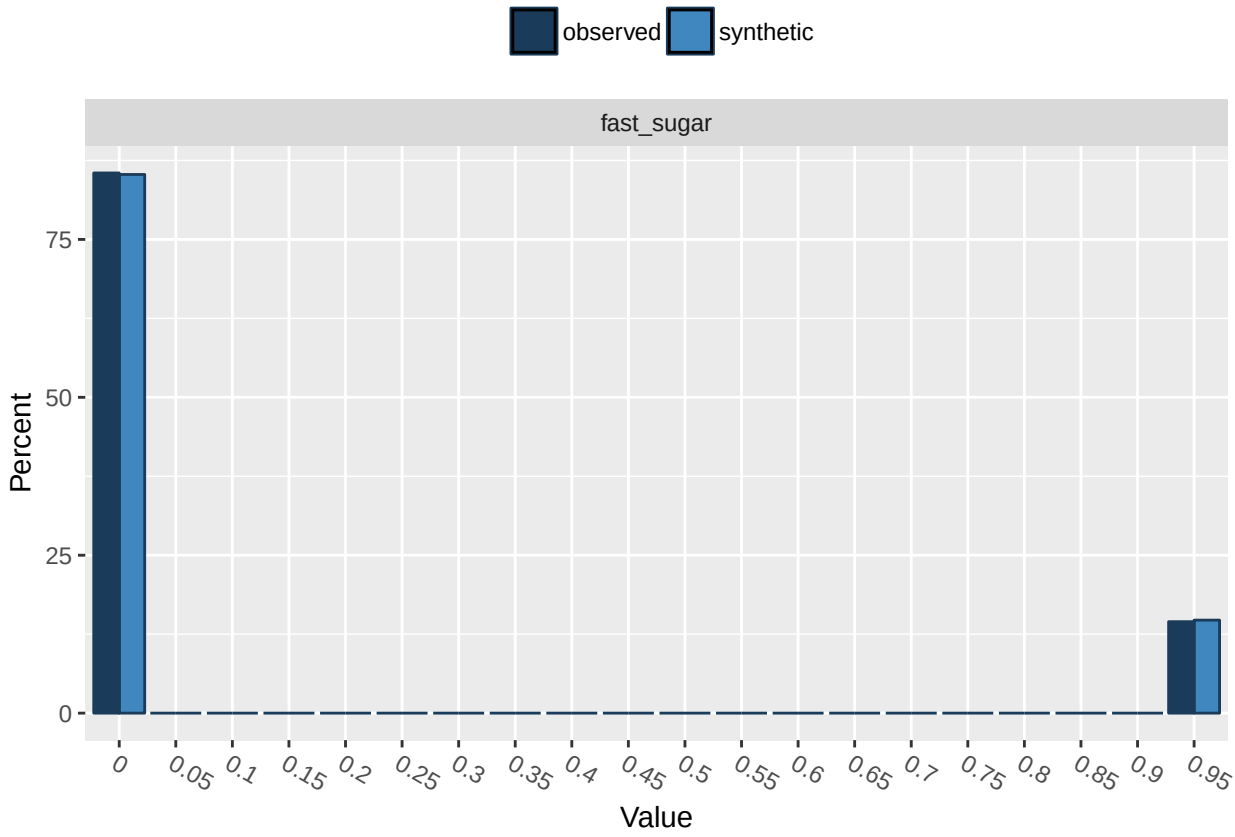

```
##
## Comparing percentages observed with synthetic
##
## $resting_bp
##                   90       95      100      105      110      115      120
## observed  0.6734007 1.346801 2.356902 8.754209 4.377104 15.15152 7.070707
## synthetic 0.6850000 1.310000 2.360000 8.395000 4.215000 14.78000 7.140000
##                  125      130      135      140      145      150      155
## observed  16.83502 6.060606 15.15152 3.367003 7.070707 2.356902 4.377104
## synthetic 17.67500 5.860000 15.68500 3.535000 6.690000 2.285000 4.525000
##                  160       165       170      175 180 185       190
## observed  0.6734007 1.346801 0.6734007 1.683502   0   0 0.3367003
## synthetic 0.7200000 1.305000 0.5650000 1.510000   0   0 0.3650000
##                  195
## observed  0.3367003
## synthetic 0.3950000
```



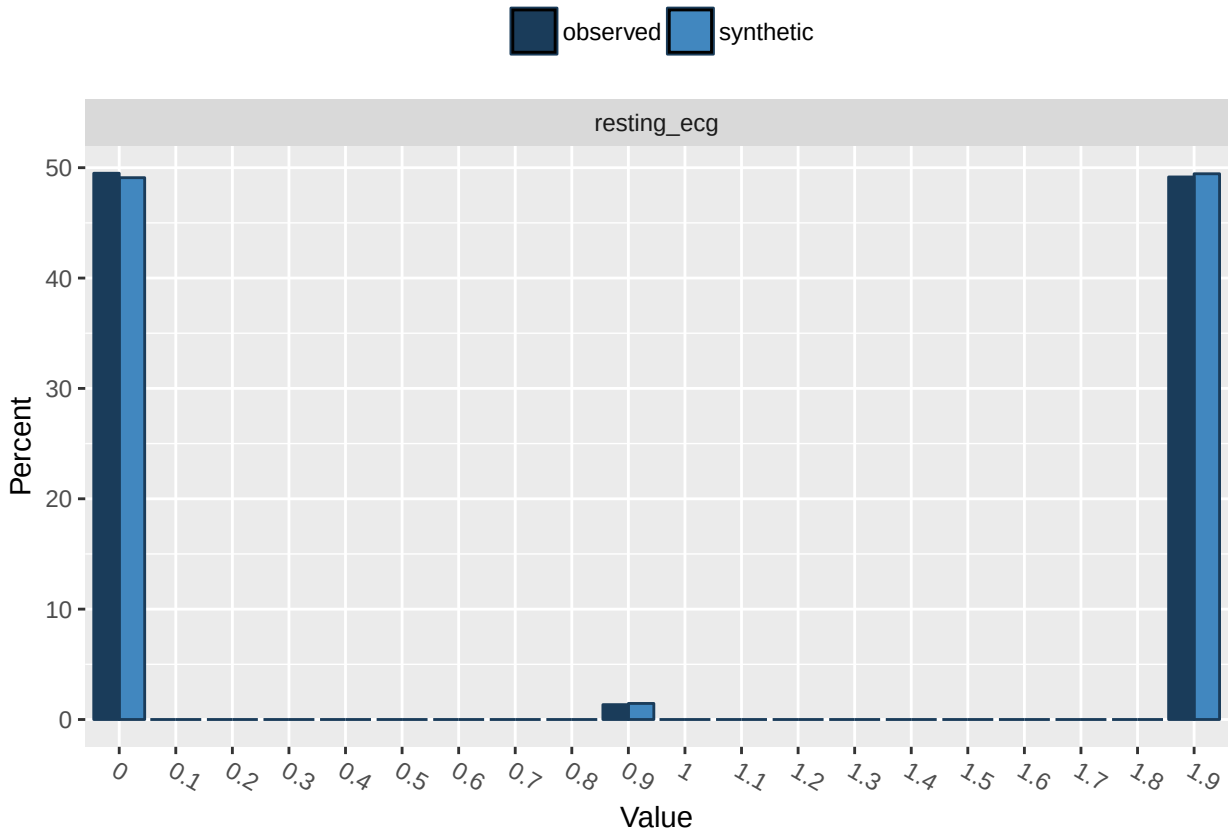

```
##
## Comparing percentages observed with synthetic
##
## $cholesterol
##                  120      140      160      180      200      220      240
## observed  0.6734007 1.683502 5.050505 9.090909 15.15152 17.50842 14.81481
## synthetic 0.7600000 1.590000 5.040000 8.705000 15.24000 17.39500 15.64000
##                  260      280      300      320      340 360       380
## observed  12.12121 9.090909 7.744108 3.703704 1.683502   0 0.3367003
## synthetic 11.83000 8.595000 7.880000 3.800000 1.765000   0 0.2550000
##                  400 420 440 460 480 500 520 540       560
## observed  1.010101   0   0   0   0   0   0   0 0.3367003
## synthetic 1.155000   0   0   0   0   0   0   0 0.3500000
```



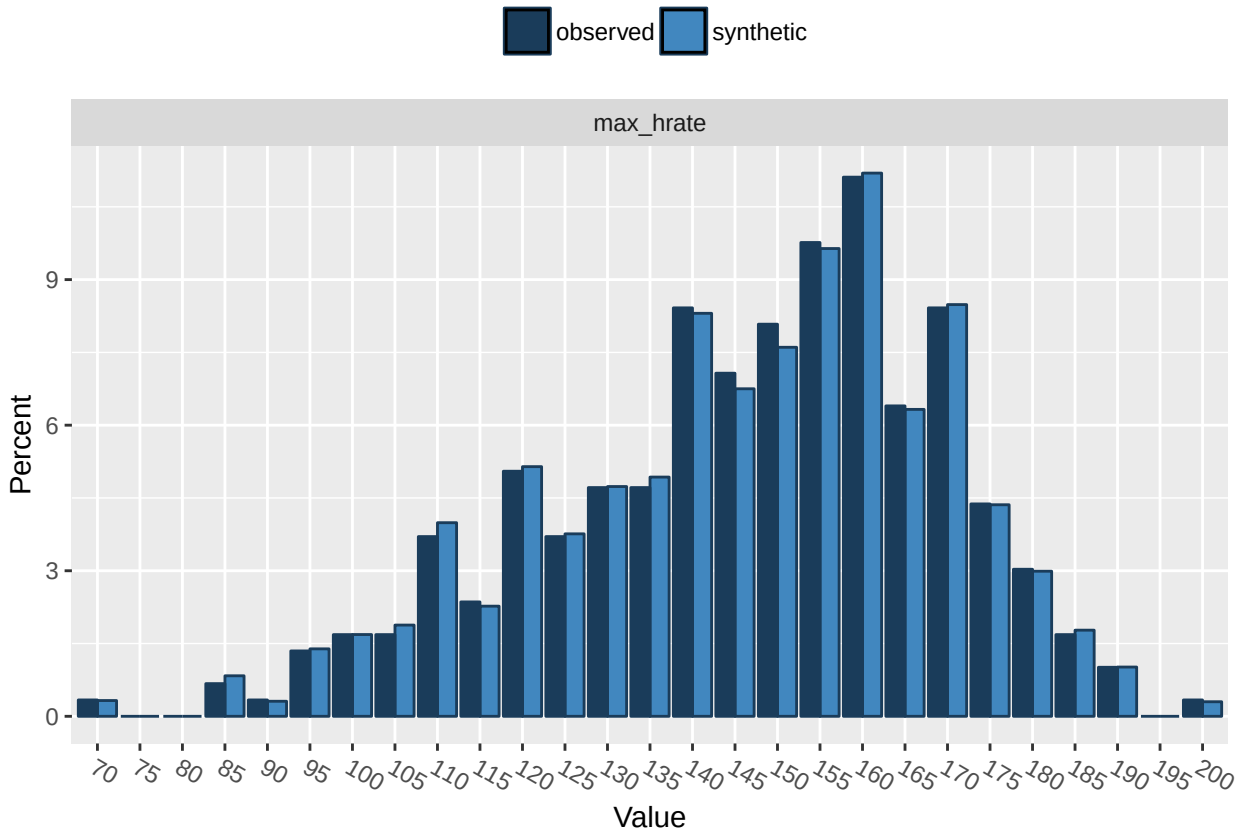

```
##
## Comparing percentages observed with synthetic
##
## $fast_sugar
##                   0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6
## observed  85.52189    0   0    0   0    0   0    0   0    0   0    0   0
## synthetic 85.28000    0   0    0   0    0   0    0   0    0   0    0   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 14.72000
```



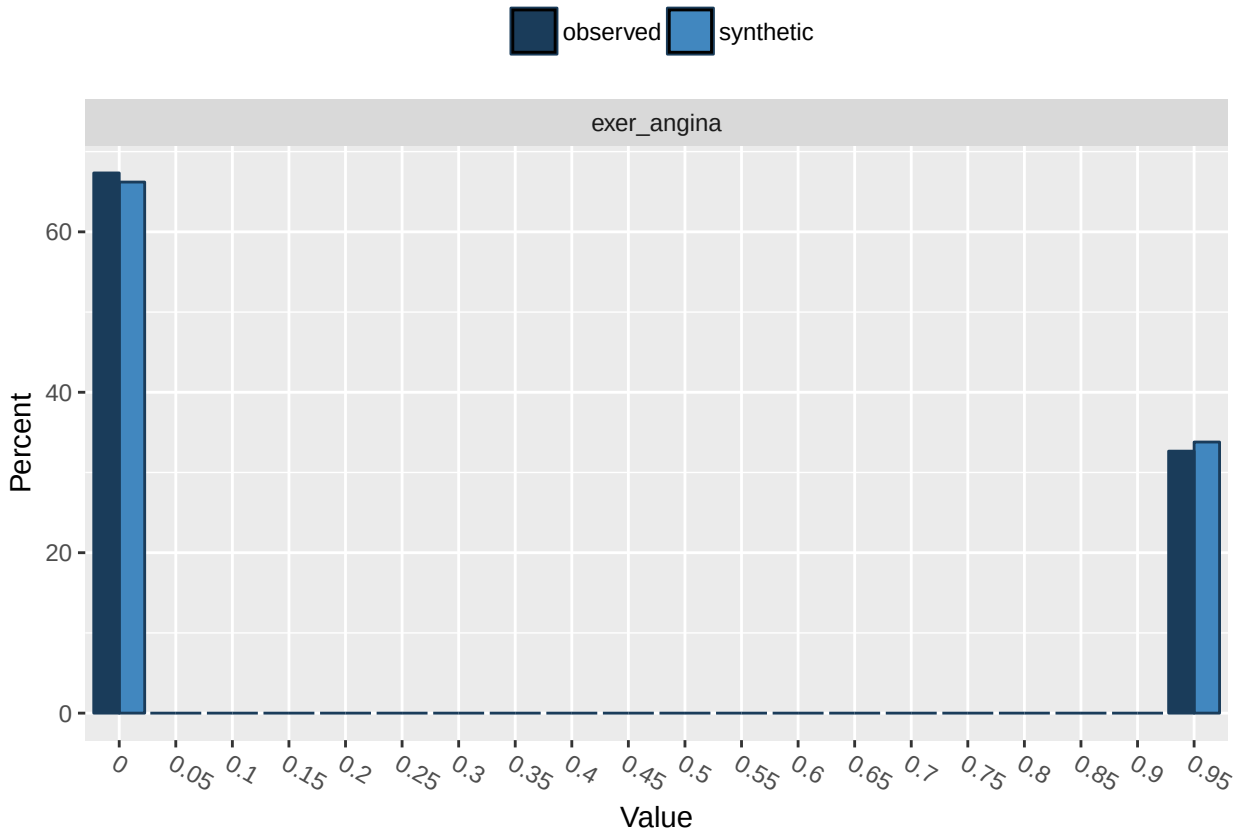

```
##
## Comparing percentages observed with synthetic
##
## $resting_ecg
##                   0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8       0.9 1 1.1 1.2 1.3
## observed  49.49495   0   0   0   0   0   0   0   0 1.346801 0   0   0   0
## synthetic 49.09500   0   0   0   0   0   0   0   0 1.455000 0   0   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 49.45000
```



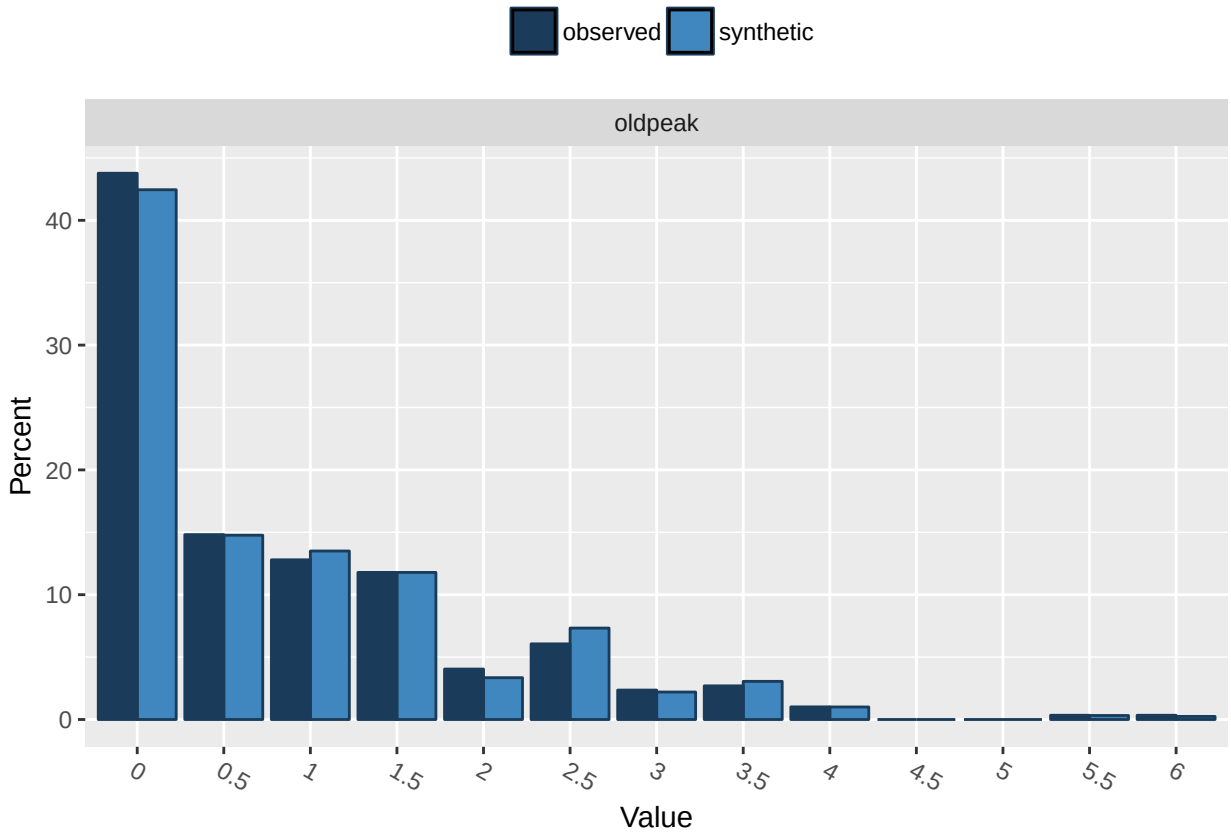

```
##
## Comparing percentages observed with synthetic
##
## $max_hrate
##                   70 75 80        85        90       95       100       105
## observed  0.3367003  0  0 0.6734007 0.3367003 1.346801 1.683502 1.683502
## synthetic 0.3250000  0  0 0.8350000 0.3100000 1.390000 1.685000 1.880000
##                110      115      120      125      130      135      140
## observed  3.703704 2.356902 5.050505 3.703704 4.713805 4.713805 8.417508
## synthetic 3.990000 2.270000 5.145000 3.760000 4.735000 4.930000 8.305000
##                145      150     155      160      165      170      175
## observed  7.070707 8.080808 9.76431 11.11111 6.397306 8.417508 4.377104
## synthetic 6.750000 7.605000 9.64000 11.19500 6.325000 8.485000 4.360000
##                180      185      190 195       200
## observed  3.030303 1.683502 1.010101   0 0.3367003
## synthetic 2.990000 1.775000 1.015000   0 0.3000000
```



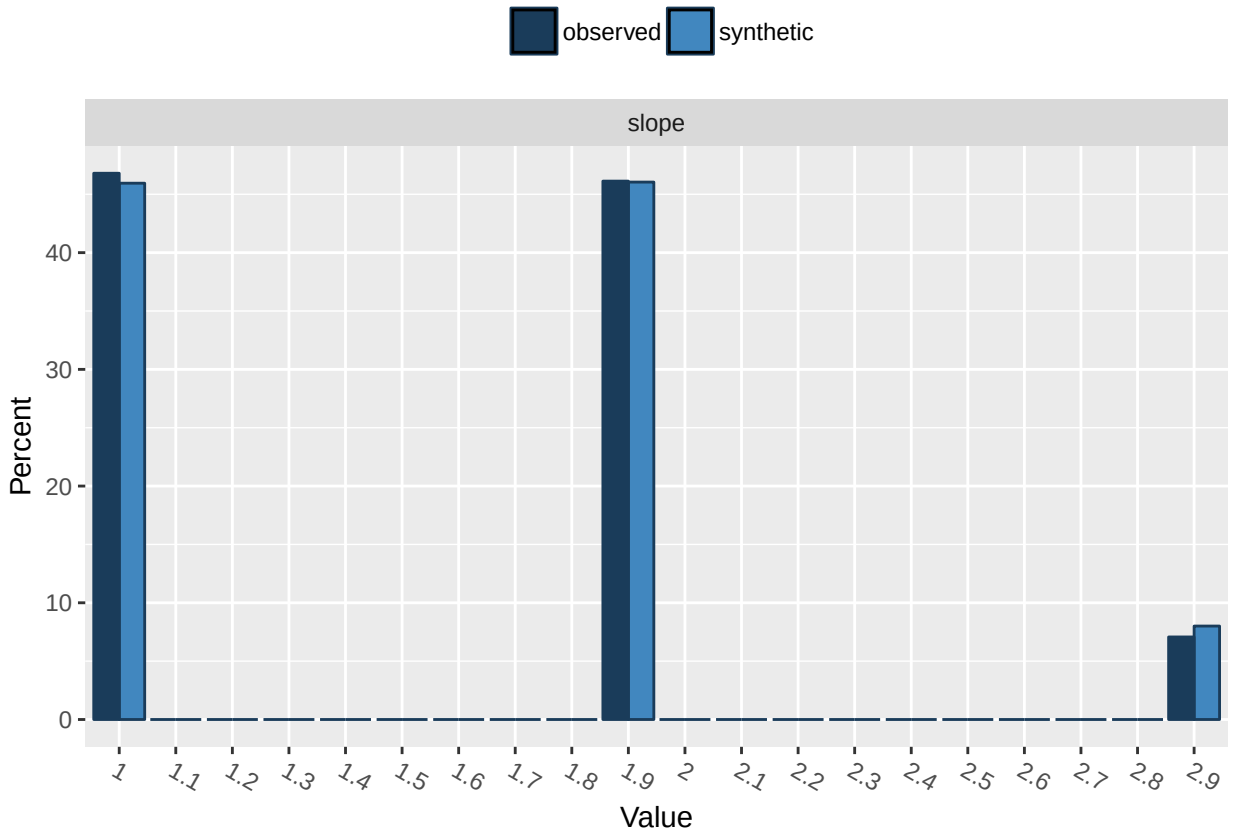

```
##
## Comparing percentages observed with synthetic
##
## $exer_angina
##                   0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6
## observed  67.34007    0   0    0   0    0   0    0   0    0   0    0   0
## synthetic 66.20500    0   0    0   0    0   0    0   0    0   0    0   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 33.79500
```



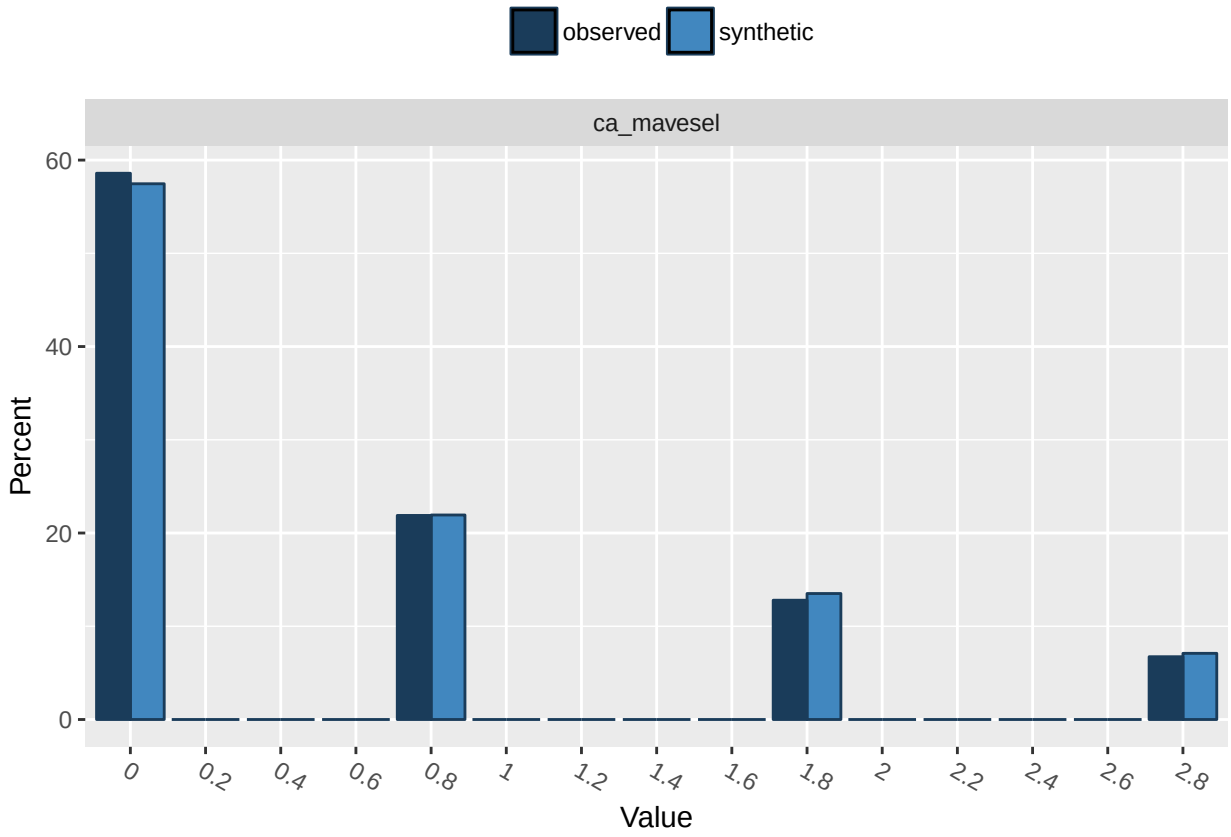

```
##
## Comparing percentages observed with synthetic
##
## $oldpeak
##                   0      0.5        1      1.5        2      2.5        3
## observed  43.77104 14.81481 12.79461 11.78451 4.040404 6.060606 2.356902
## synthetic 42.45000 14.76500 13.50000 11.78500 3.350000 7.325000 2.200000
##                 3.5        4 4.5 5       5.5         6
## observed  2.693603 1.010101   0 0 0.3367003 0.3367003
## synthetic 3.055000 1.005000   0 0 0.3200000 0.2450000
```



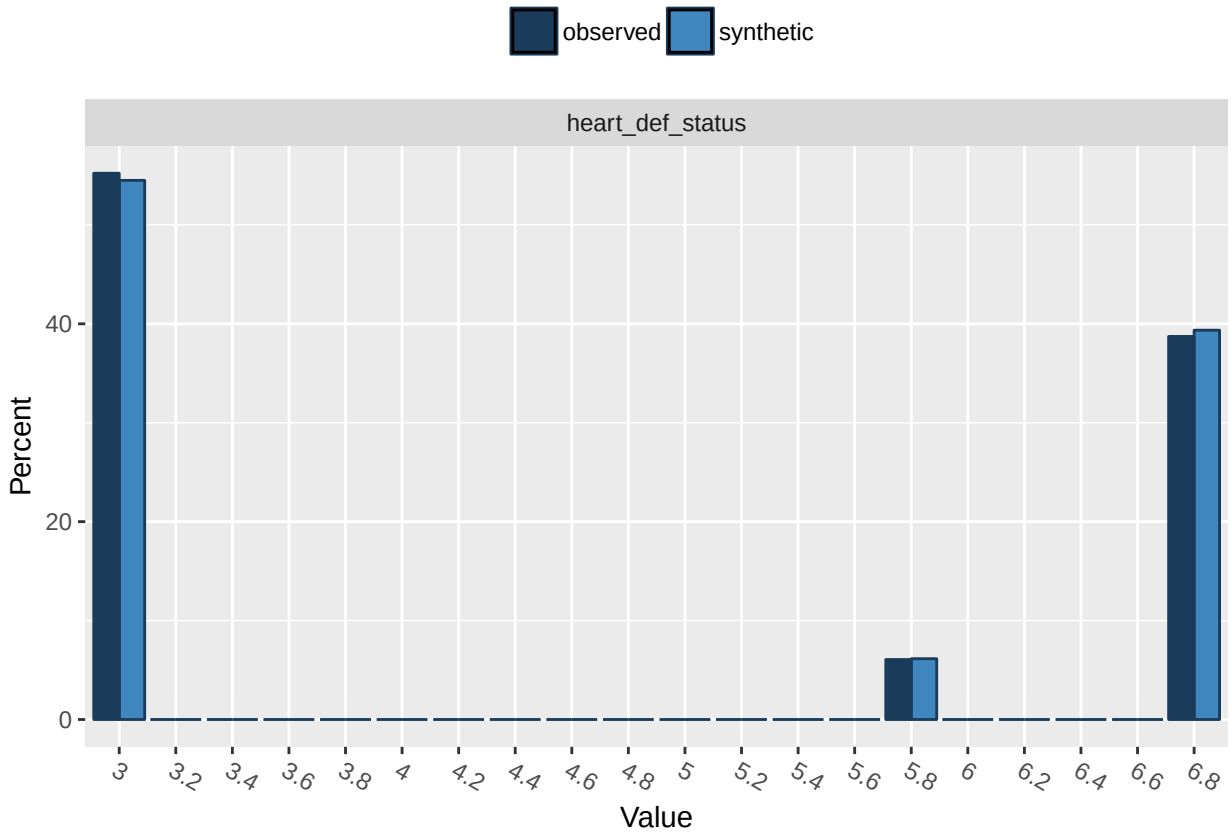

```
##
## Comparing percentages observed with synthetic
##
## $slope
##                  1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8      1.9 2 2.1 2.2 2.3
## observed  46.80135   0   0   0   0   0   0   0   0 46.12795 0   0   0   0
## synthetic 45.95000   0   0   0   0   0   0   0   0 46.04500 0   0   0   0
##           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 8.005000
```



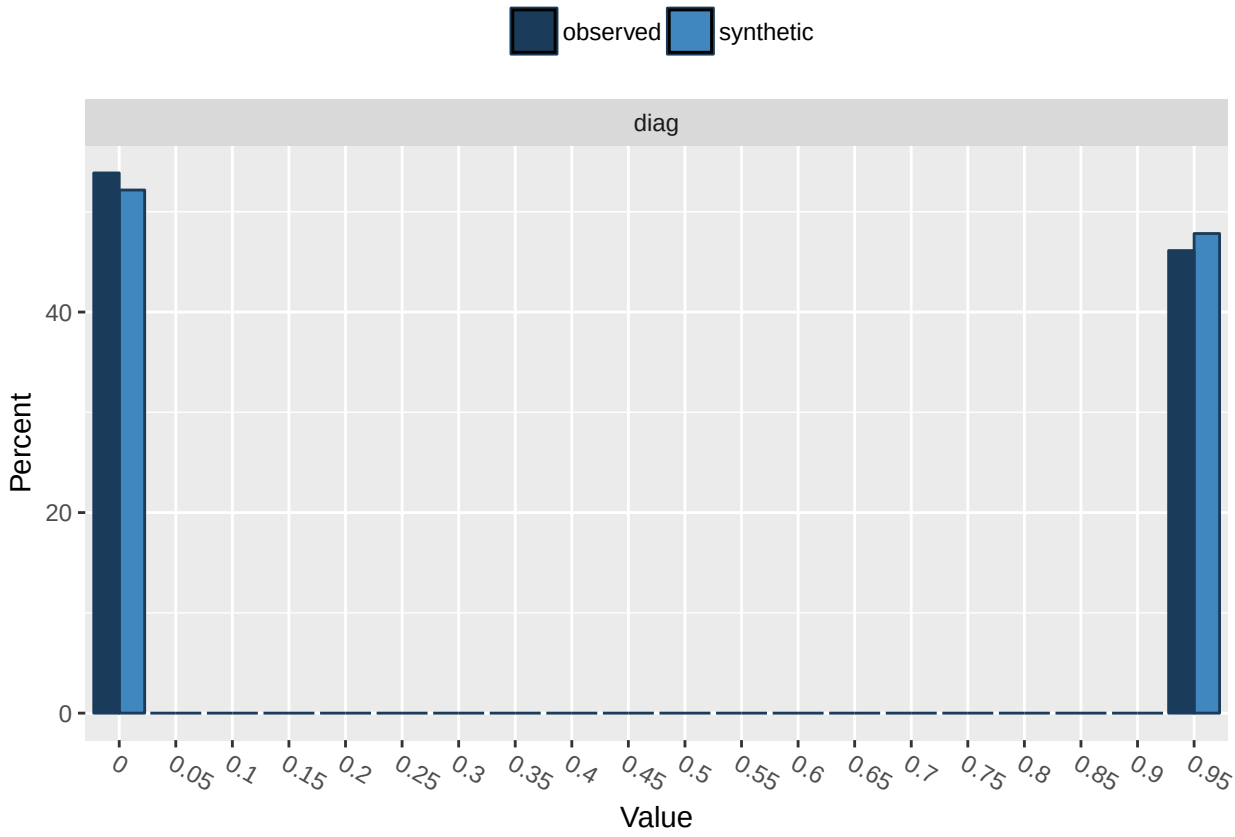

```
##
## Comparing percentages observed with synthetic
##
## $ca_mavesel
##                  0 0.2 0.4 0.6      0.8 1 1.2 1.4 1.6      1.8 2 2.2 2.4
## observed  58.58586   0   0   0 21.88552 0   0   0   0 12.79461 0   0   0
## synthetic 57.46000   0   0   0 21.93500 0   0   0   0 13.51000 0   0   0
##           2.6      2.8
## observed    0 6.734007
## synthetic   0 7.095000
```




```
##
## 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 54.50000   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 6.145000 0   0   0   0 39.35500
```




```
##
## 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 52.17000    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 47.83000
```


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