GANs

Jonathan P. Latner, PhD

October 16, 2023

The question

When are GANs efficient/effective in creating synthetic data sets relative to Synthpop (Datasynthesizer)?

What do we know?

What do we not know?

Data

- 2 rows/observations = [1.000, 5.000]
- 3 cols/variables = [10, 15, 20]
- 4 values per variable = [5, 10, 15, 20]
- Datasets = 24 ($2 \times 3 \times 4$)
- Synthpop, Datasynthesizer, CTGAN set to default values
- Note: 1 synthetic copy per dataset because focus is on duration

Figure 1: Duration: CTGAN =< Synthpop if (cols > 20 | vals > 20)

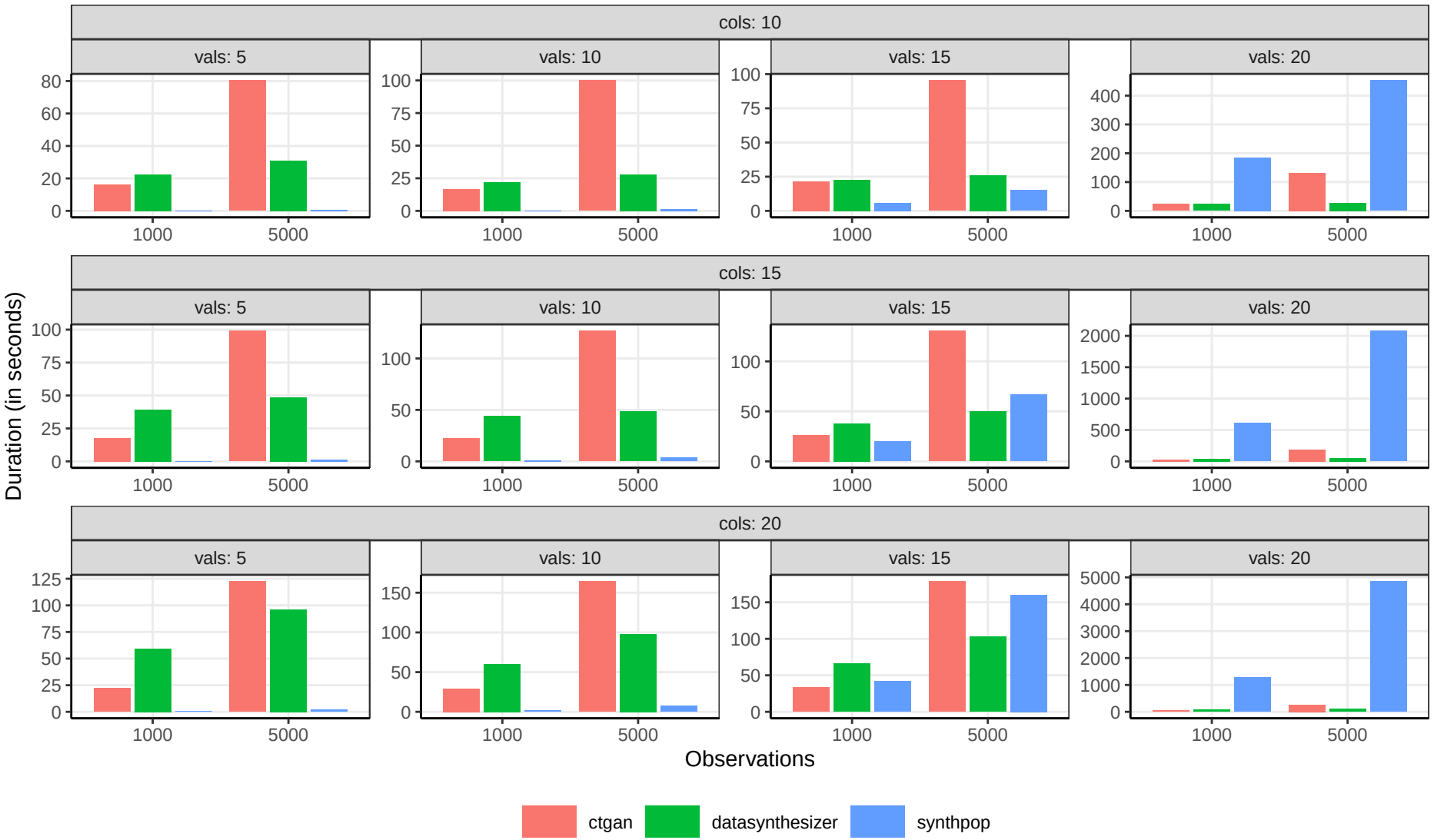


Figure 2: Kolmogorov-Smirnov utility measure

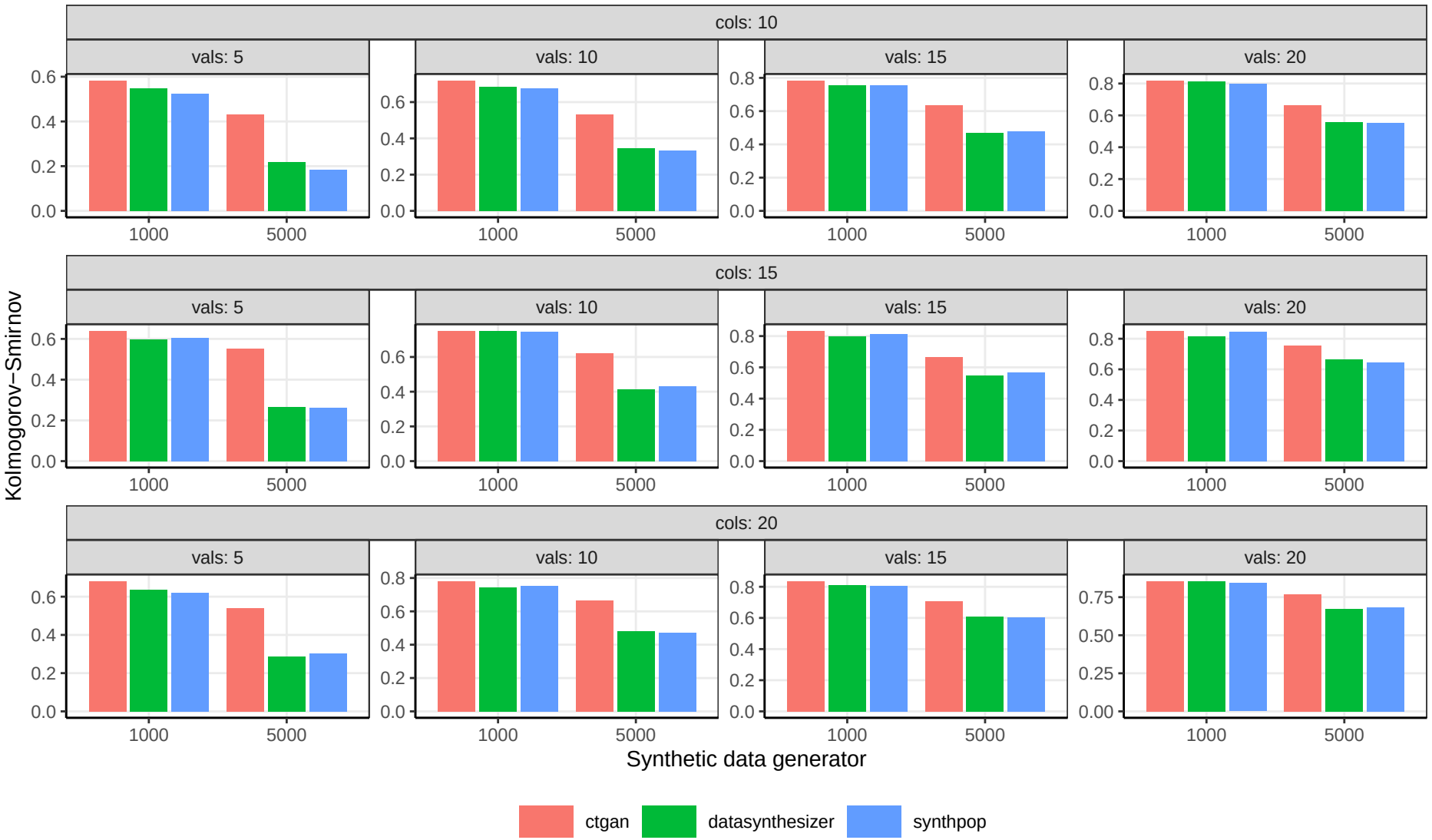
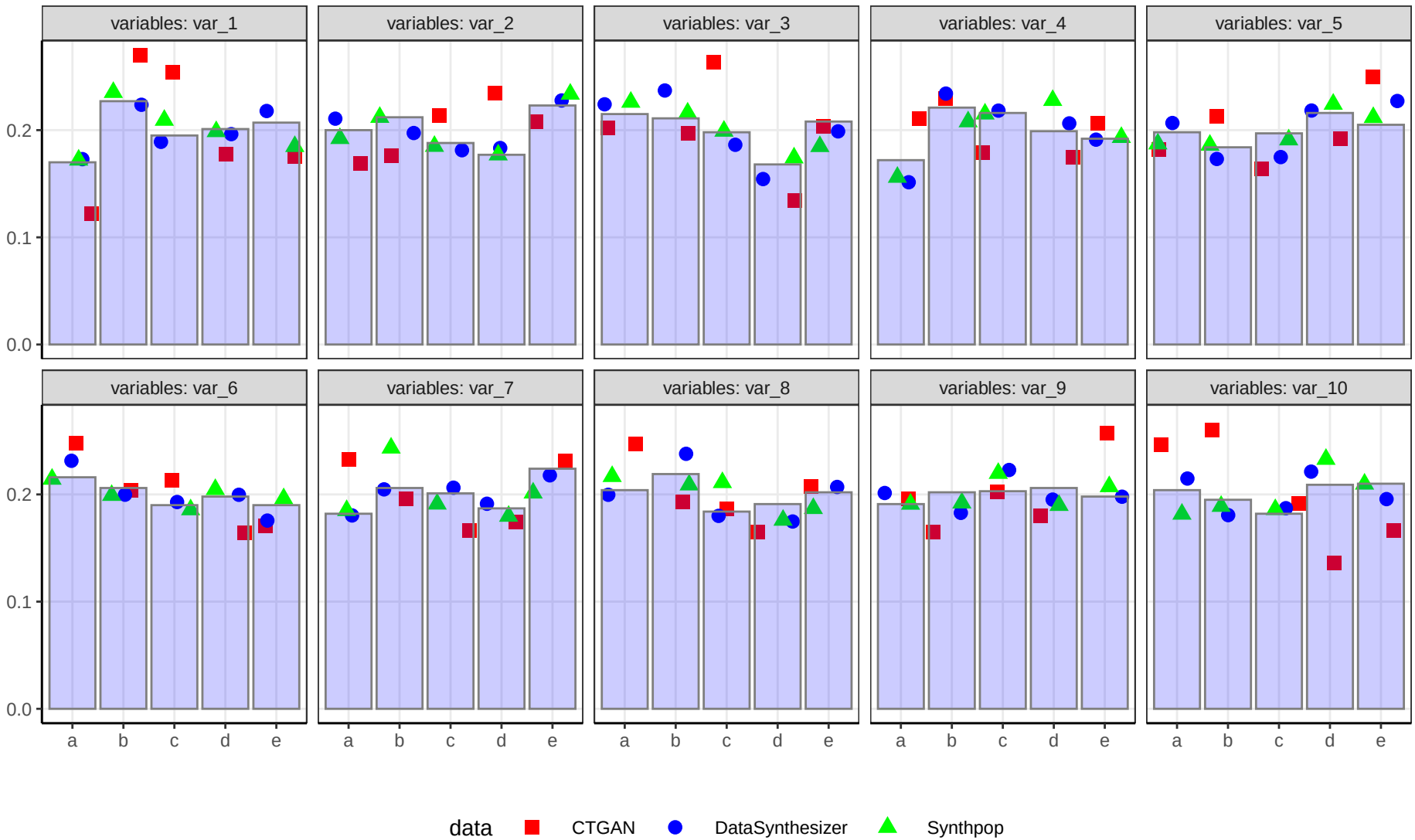


Figure 3: Example frequency for cols = 10, rows = 1.000, vals = 5



Data

- 3 rows/observations = [50.000, 100.000, 200.000]
- 3 cols/variables = [10, 15, 20]
- Datasets = 9
  - Different distributions, depending on # variables
- Note: 1 synthetic copies per dataset
- Focus on tuning CTGAN
  - epochs = [10, 20, 30, 40, 50, 75, 100]
  - batch size = [500, 1.000, 5.000, 10.000]

Figure 4: Frequency for cols = 10, rows = 100.000

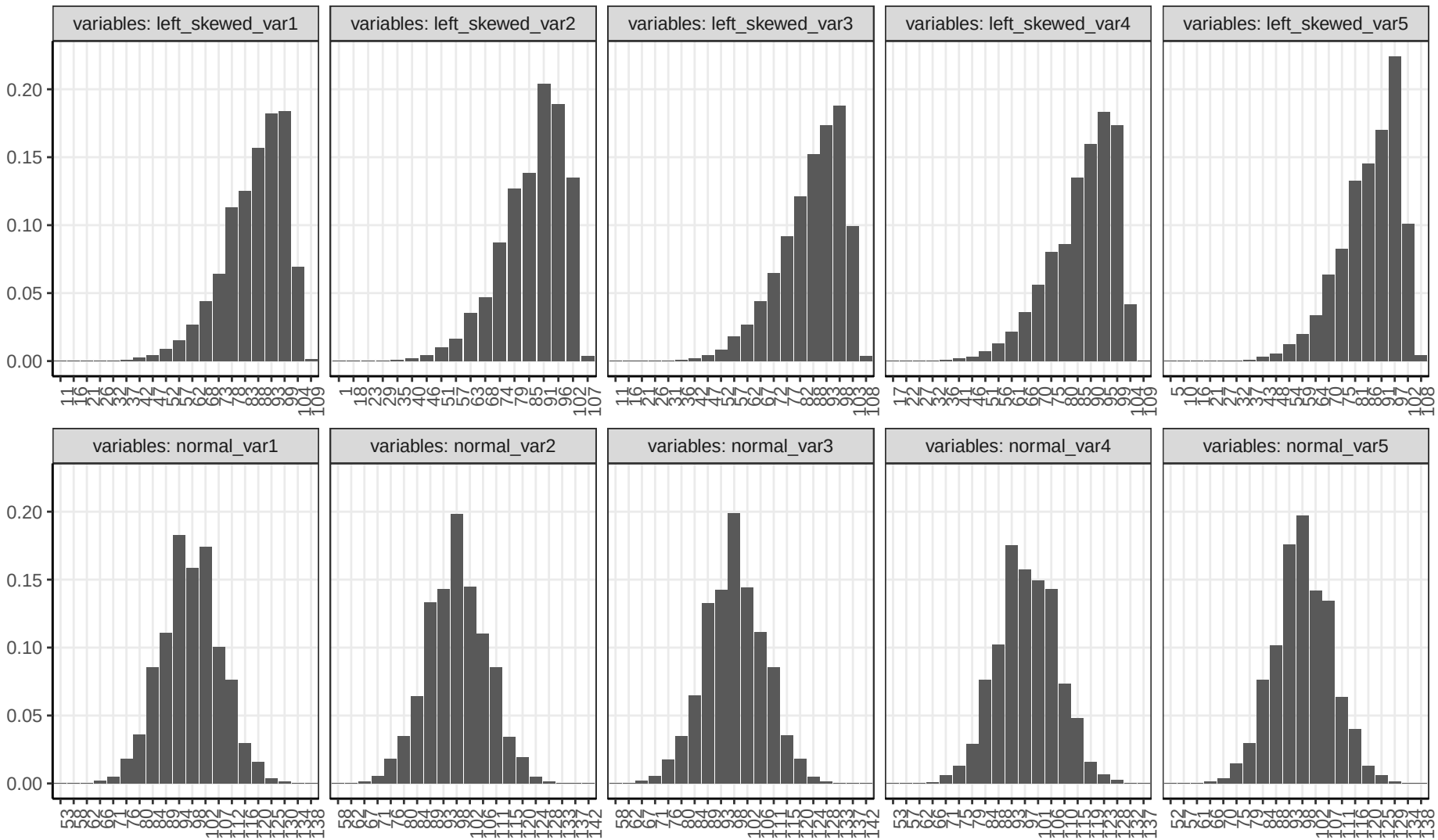


Figure 5: Frequency for cols = 15, rows = 100.000

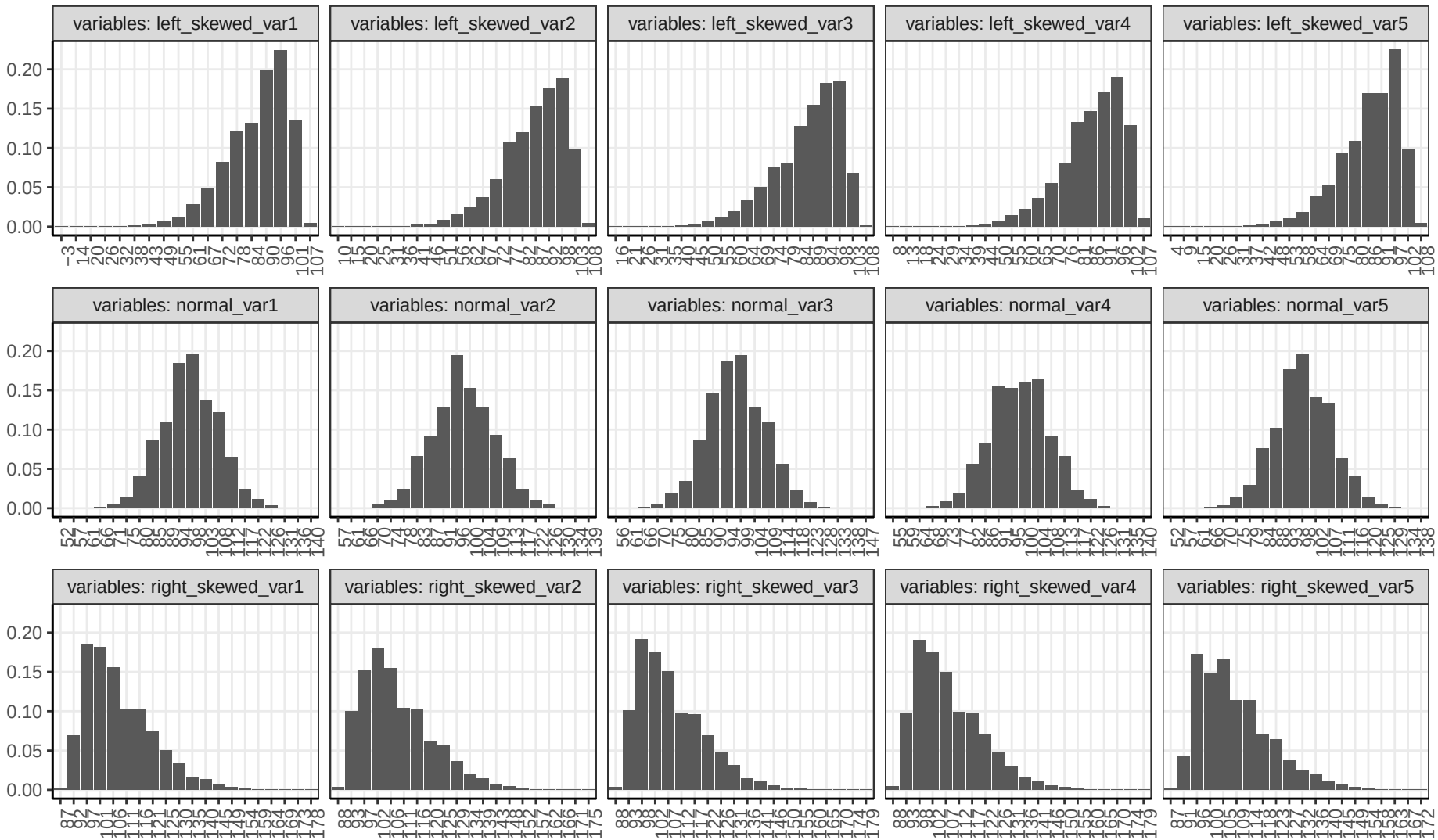


Figure 6: Frequency for cols = 20, rows = 100.000

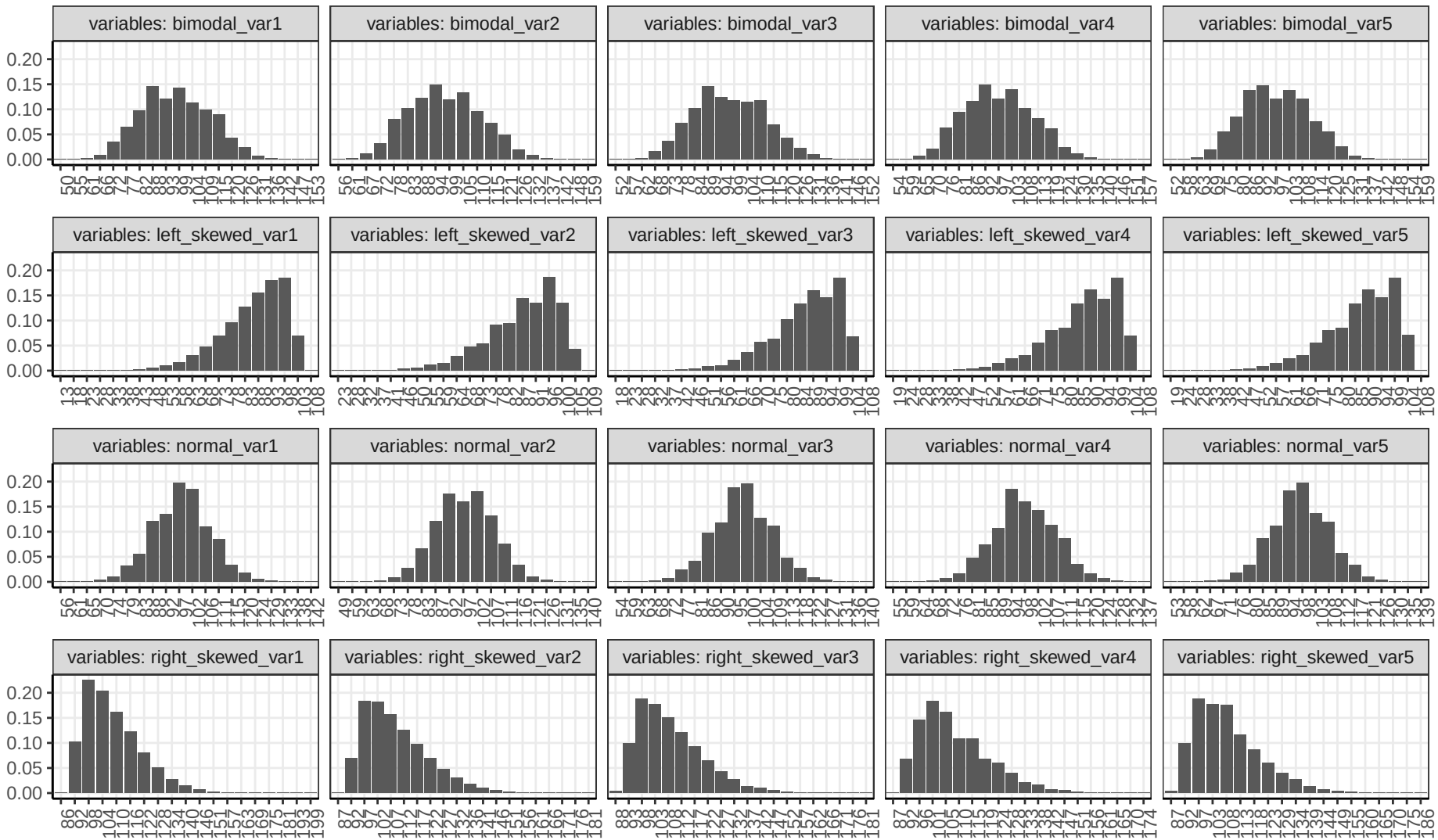


Figure 7: Duration (CTGAN)

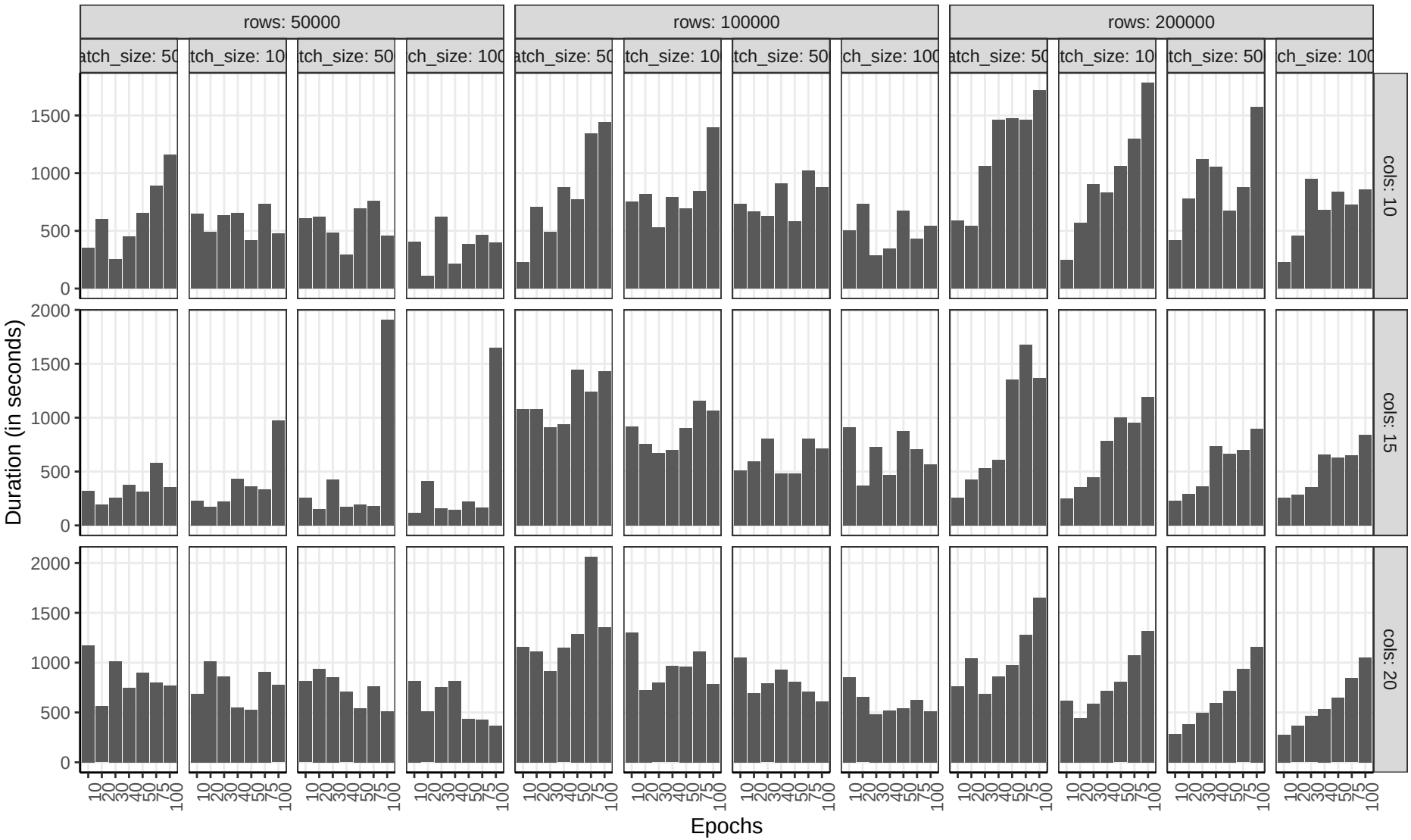


Figure 8: Kolmogorov-Smirnov utility measure (CTGAN)

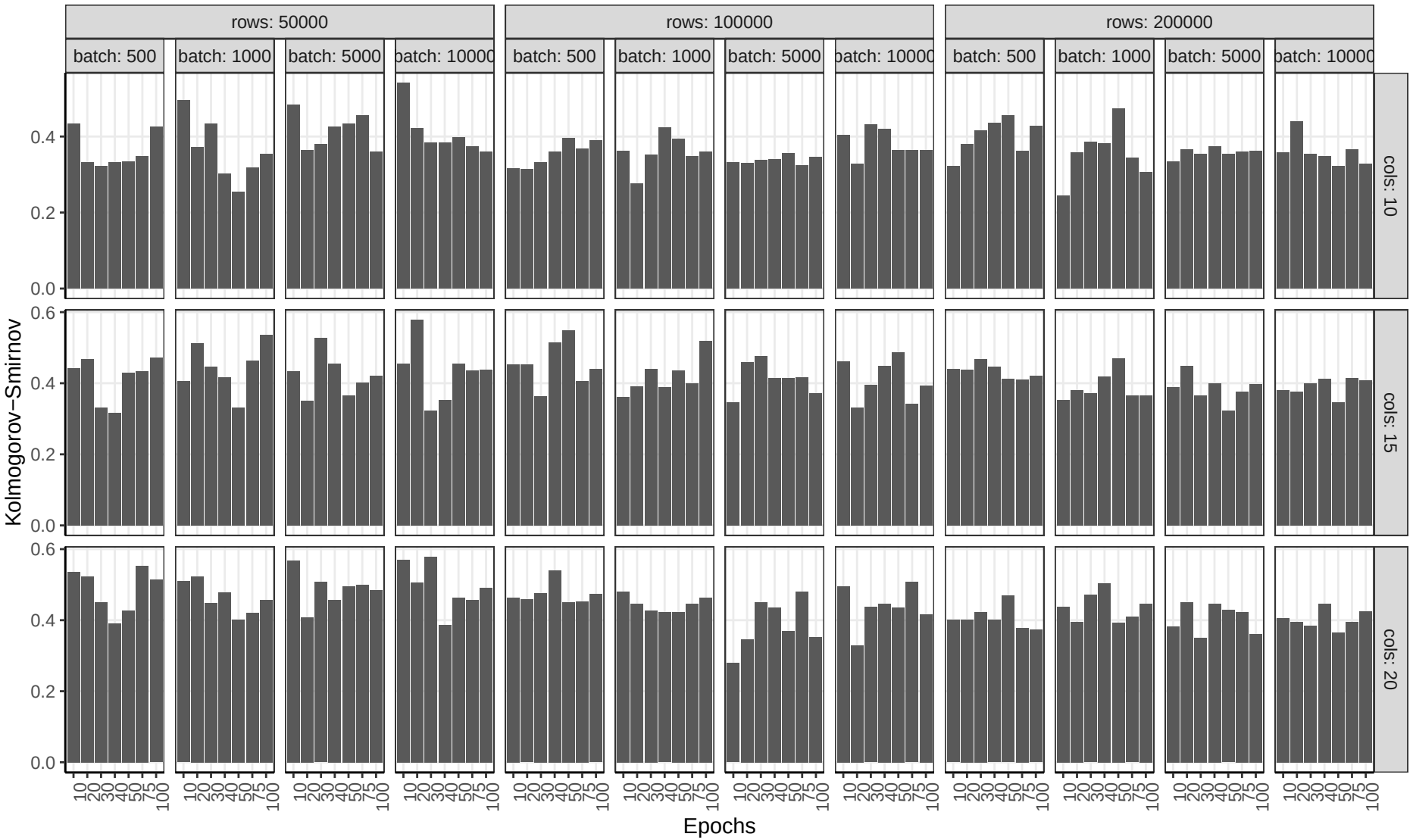
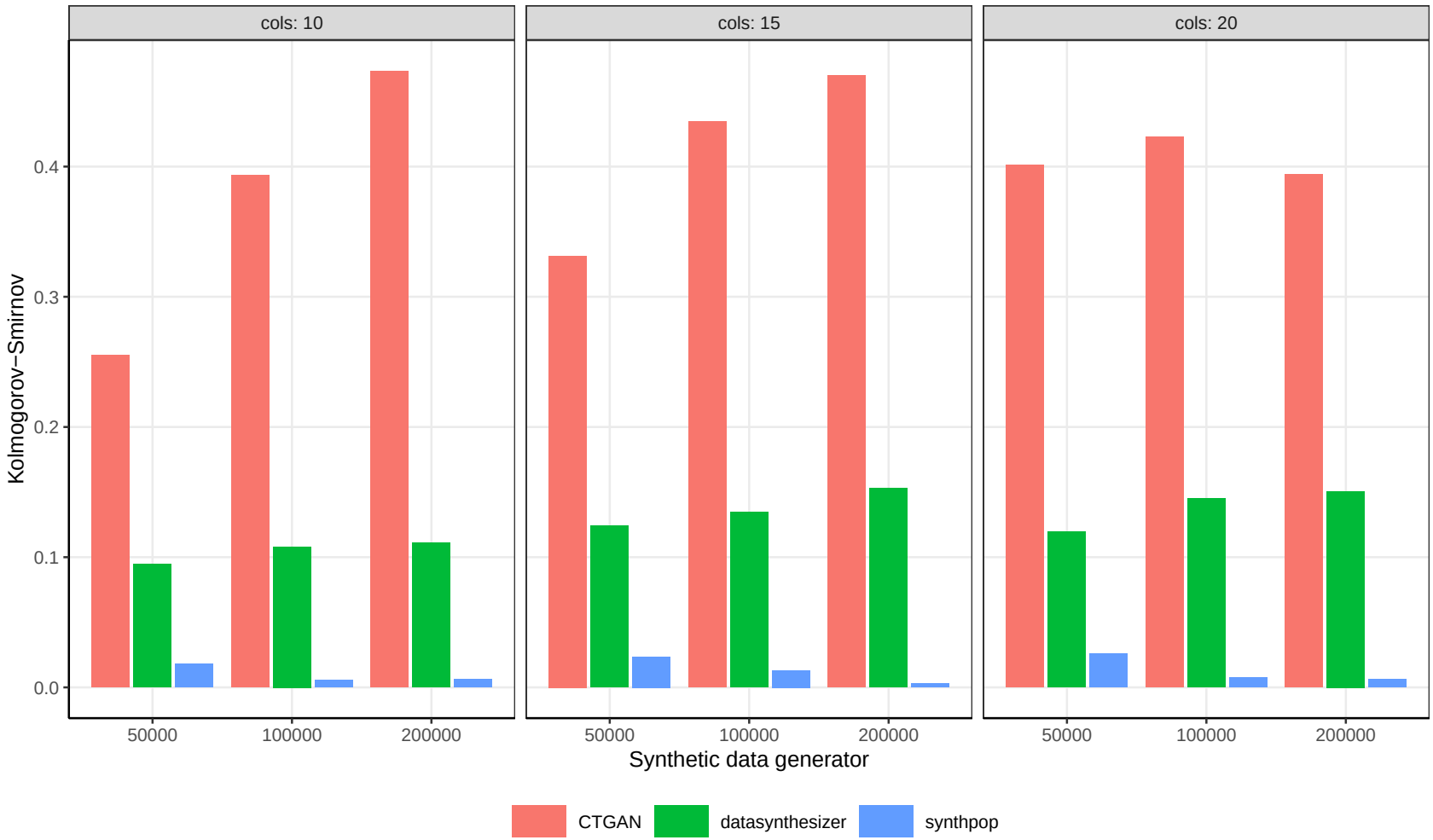
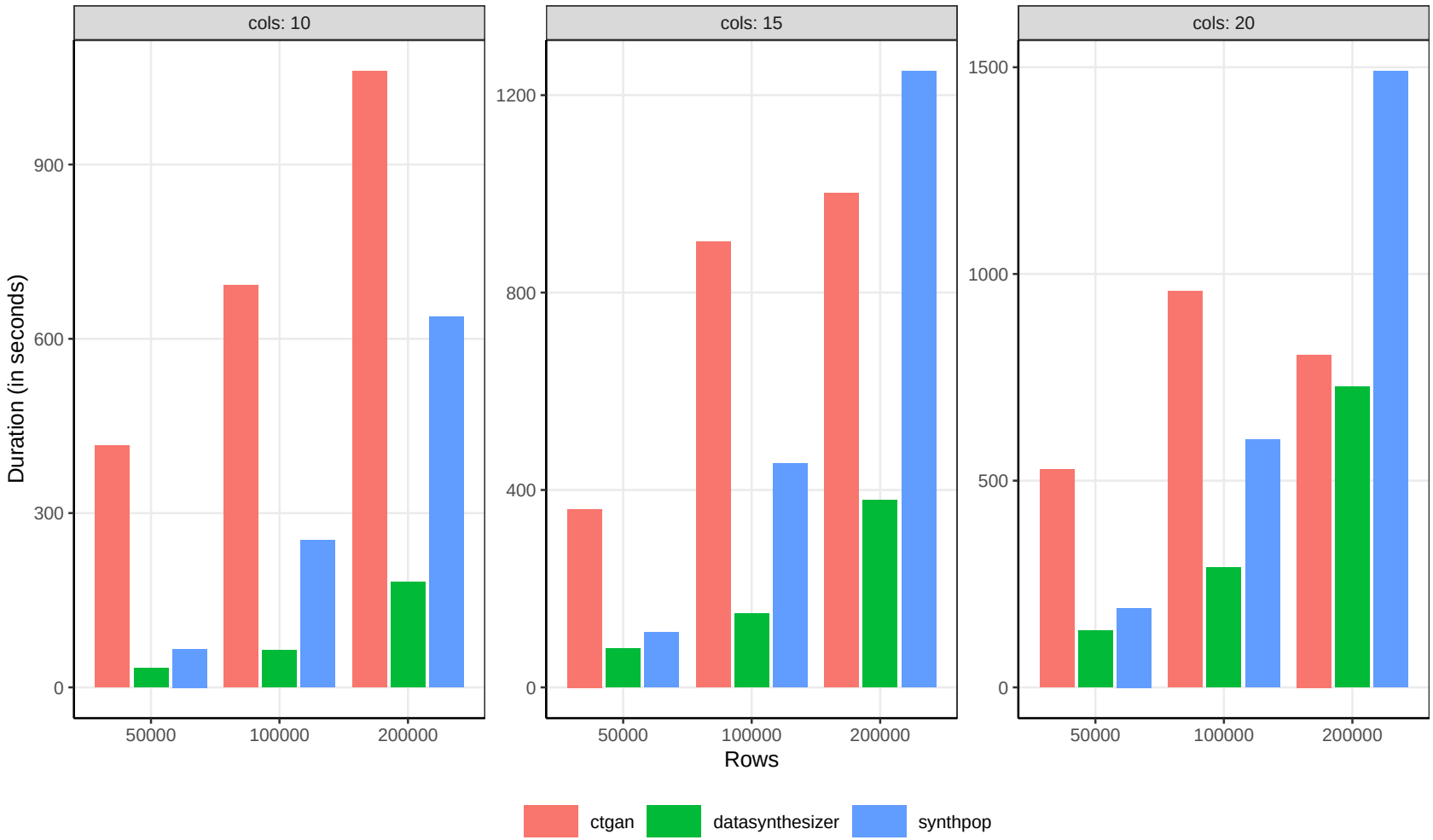


Figure 9: Kolmogorov-Smirnov utility measure



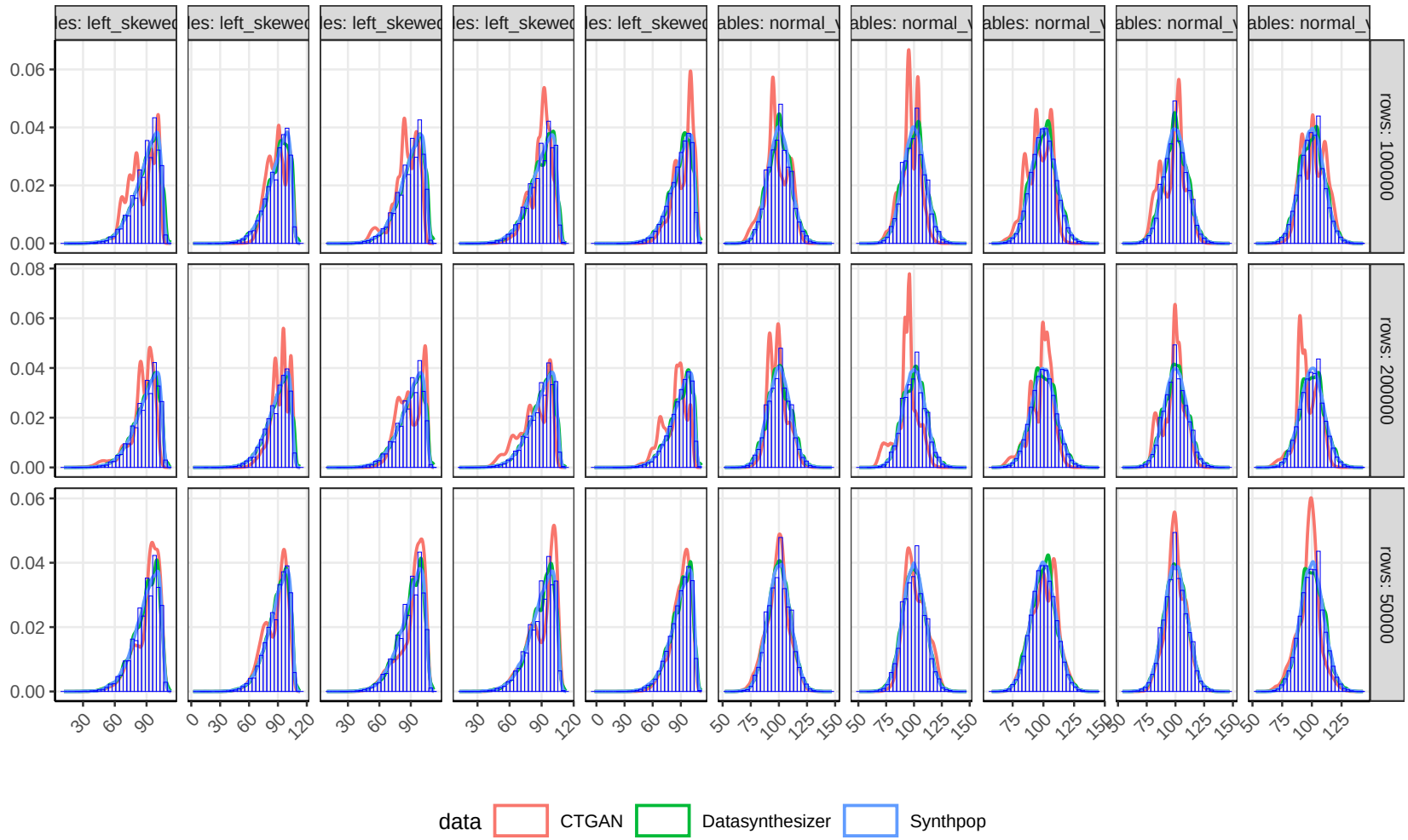
Note: CTGAN hyperparameter: batch size = 1000 and epochs = 50 (see figure 8)

Figure 10: Duration



Note: CTGAN hyperparameter: batch size = 1000 and epochs = 50 (see figure 8)

Figure 11: Example frequency for cols = 10



Note: CTGAN hyperparameter: batch size = 1000 and epochs = 50 (see figure 8)