

## Appendix A: Dataset Preparation

Prior to analysis, anomalies queries were run to identify rate outliers and inconsistent (within provider) rates likely reflecting data entry errors. This included a manual review of provider records with:

- Weekly rates below \$25
- Weekly rates above \$400
- Full day rates where the difference between age group rates (Under 1 year vs 1, 1 vs 2, 2 vs 3, 3 vs 4, 4 vs 5 years/kindergarten) was greater than |\$50| (absolute value)
- Day camp rates where the difference between high rates and low rates was greater than |\$25|
- Daily drop-in rates below \$15
- Daily drop-in rates above \$100
- Daily drop-in rates with differences between age group rates (2 vs 3, 3 vs 4, 4 vs 5 years/kindergarten, 5years/kindergarten vs 5 years and older) was greater than |\$15|

Flagged records were cleaned as follows:

- Set \$0 rates to missing
- Set weekly full-day rates below \$25 to missing (including some because they were inconsistent across the age groups; all rates for four providers)
- Set weekly day camp rates below \$20 to missing (all rates for five providers)
- Set daily drop-in rates below \$5 to missing (all rates for one provider)
- For rates clearly having an extra digit based on other reported rates (typically, but not always, a zero), removed the extra digits
  - For one provider, a prevalent pattern was not identified (rates of \$424, \$424, \$3,690, \$3690, \$3,690), so removed zeros based on website-published tuition rates.
- For rates clearly missing a digit at the beginning or end of the number based on other reported rates, added the appropriate digit (e.g., for one provider with rates of \$397, \$376, \$365, \$347, \$30, changed the last rate to \$330)
- In other cases, where the appropriate update could not be determined, the single inconsistent rate was deleted (e.g., for one provider with rates of \$993, \$898, \$804, \$704, \$160, the \$160 rate was removed)

There were several providers (approximately 30) that provided seemingly discrepant lower weekly full-day rates for 4-year-olds, but as this seemed intentional, those rates were left as they were. (e.g., one provider with rates of \$185, \$150, \$150, \$140, \$80). It may be these providers were reporting before and after care/meals for publicly funded Georgia's Pre-K classes.

Additionally, despite the above-mentioned data cleaning, there are some maximum rates that may seem unlikely for a specific Zone, and there are some drop-in rates that seem more likely to be weekly rather than daily rates (more than \$200/day for four providers). The latter may be due to providing two sets of weekly rates before the drop-in rates and/or the grid data entry format.