Dataset Overview and Issues

The dataset contains 97 records with four columns: `age`, `income`, `city`, and `signup_time`.

Initial inspection revealed missing values, invalid negative ages, extreme income values, inconsistent city names, and mixed date formats.

Cleaning Steps and Rationale

- **Type conversion:** Converted numeric and datetime columns using safe parsing to handle invalid values.
- **Missing values:** Imputed age with the median, income with the mean, city with the mode, and signup time with the most frequent value.
- **Domain correction:** Removed negative ages based on domain knowledge and re-imputed them.
- **Outlier handling:** Applied 1st–99th percentile capping to age and income to reduce the influence of extreme values while keeping all records.
- **String cleaning:** Normalized city names by lowercasing and removing punctuation and extra spaces.
- **Canonical mapping:** Mapped city abbreviations (e.g., “ny”, “sf”, “la”) to standard names.
- **Validation:** Confirmed no missing values, no negative ages, and correct data types.

Trade-offs and Limitations

Imputation and outlier capping may reduce data variability and obscure rare extreme cases. City mapping depends on predefined rules and may not capture unseen categories.

Final Outcome

The final dataset is clean, consistent, and ready for analysis or modeling.