NovaWireless Synthetic Customer Generator:

A Reproducible Framework for Telecom Population Simulation

Gina Aulabaugh

Applied Data Science Coursework

February 15, 2026

Abstract

Modern telecommunications research, call-center analytics, and churn modeling experiments require realistic customer data. However, access to real customer records is legally restricted and ethically sensitive. This paper presents the NovaWireless Synthetic Customer Generator, a reproducible system for generating statistically grounded yet fully synthetic telecom customer populations. The generator derives empirical distributions from publicly available datasets (e.g., telecom churn benchmarks and telecommunications indicators) and uses probabilistic sampling to construct realistic customer records. Service contracts are replaced with device payment plans to reflect contemporary industry structures. Churn is modeled as a probability (risk score) rather than a deterministic outcome. The system produces structured outputs suitable for simulation, experimentation, and portfolio demonstration without redistributing proprietary datasets. This paper documents the architecture, statistical grounding, business logic, reproducibility controls, and ethical safeguards of the generator.

1. Introduction

Telecommunications analytics systems depend on customer-level datasets to evaluate churn behavior, revenue patterns, and service adoption. Access to real customer data is constrained by privacy laws, corporate governance, and regulatory compliance. Synthetic data provides a practical alternative, allowing researchers and engineers to simulate realistic populations without exposing sensitive information.

The NovaWireless Synthetic Customer Generator was designed to create a statistically plausible telecom customer database grounded in empirical artifacts while ensuring that no original row-level data is redistributed. The project balances three goals: statistical realism, reproducibility, and legal/ethical safety.

2. Data Grounding and Statistical Derivation

2.1 Source Datasets

Publicly available datasets were used solely to derive aggregate distributions and statistical parameters (e.g., probability mass functions for categorical attributes, histogram bins for numeric variables, and conditional churn rates). No row-level third-party data is redistributed by this project.

Representative sources include a telecom churn benchmark dataset (**ibm_telco_churn**), public call-center performance data (**mta_call_center**), and telecommunications indicators published through research-friendly portals (**owid_mobile_subs**).

2.2 Derivation Approach

External datasets inform the generator through *derived distributions* rather than record reuse. Practically, this means the baseline configuration stores only:

- Category probabilities (e.g., payment methods, internet service types)

- Numeric histogram bins and their probabilities (e.g., tenure, monthly charges)
- Aggregate churn rates and conditional churn lookups when available

This design supports realism while reducing privacy and licensing risk.

3. System Architecture

The system is implemented as two scripts:

- `generate_customers.py`: emits a synthetic customer population
- `clean_customers_v1.py`: normalizes fields and writes a stable v1 dataset and profile

Primary output artifacts include:

- `customers.csv`: raw generated customers
- `customers_v1.csv`: cleaned, stable v1 schema
- `customers_v1_profile.csv`: summary metrics (QA profile)
- `customer_generation_receipt.json`: reproducibility receipt (seed, schema, summaries)

4. Customer Generation Process

4.1 Numeric Variable Sampling

Numeric attributes such as tenure and monthly charges are sampled using histogram-based distributions:

1. Choose a bin according to its empirical probability.
2. Sample uniformly within the bin interval.
3. Clamp values to observed bounds.

This preserves realistic variance without reproducing original records.

4.2 Categorical Sampling

Categorical fields (e.g., internet service and payment method) are sampled from normalized empirical probability distributions. Sampling is controlled by a fixed random seed to support reproducibility.

5. Device Payment Plan Logic

Service contracts are not emitted in the final dataset. Instead, the generator models modern wireless commitment structures through device payment plans:

- A customer either has a device payment plan or does not.
- If present, the plan term is always 24 months.
- Months remaining vary and are sampled with a tenure-informed tendency while retaining variability.
- Device monthly payment is bounded as a plausible share of monthly charges.

If a source dataset includes contract-like categories, they may be used internally as a statistical proxy to approximate commitment patterns. The proxy is not emitted as a customer-facing field.

6. Churn Risk Modeling

The system does not label customers as churned. Instead, it emits `churn_risk_score`, interpreted as a churn probability. When available, conditional churn rates are estimated from aggregate group statistics (e.g., by internet service and tenure bin). If conditional tables are not available, a global churn rate is used.

This design supports probabilistic simulation and avoids deterministic outcome leakage.

7. Latent Behavioral Variables

Two latent behavioral traits are derived from churn risk:

- `trust_baseline`: approximately $1 -$ `churn_risk_score` with bounded noise
- `patience`: a bounded function of churn risk with noise

These are simulation control parameters rather than measured attributes.

8. Cleaning and Normalization

The cleaning stage:

- Removes unused or deprecated fields (e.g., gender, contracts, paid support fields)
- Combines streaming fields into `streaming_services` (Yes/No)
- Normalizes add-on flags to a consistent Yes/No representation
- Writes a stable v1 schema and a dataset profile summary

9. Reproducibility

Reproducibility is ensured through seed control. The generator is deterministic given the same seed and configuration. The receipt artifact records key parameters (seed, counts, schema, and summary statistics) to support auditability.

10. Ethical and Licensing Considerations

All generated records are synthetic. The project is designed to avoid redistribution of third-party datasets and does not represent any real individuals. NovaWireless is a fictional entity and is not affiliated with any real telecommunications provider or data publisher. Users who independently download external datasets should follow those datasets' terms for their own local use.

11. Conclusion

The NovaWireless Synthetic Customer Generator provides a reproducible and ethically safer framework for creating telecom-like customer populations for research and simulation. By grounding sampling in derived distributions rather than record reuse, the system supports realistic experimentation while reducing privacy and licensing risk. The resulting dataset and profile artifacts enable downstream work in churn analysis, call-center simulation, and metric-governance research.

References

Ahmadzade Jahromi, M. (n.d.). *Call center data (Call center daily performance)* [Data set]. Kaggle. Retrieved February 15, 2026, from https://www.kaggle.com/datasets/satvicoder/call-center-data

IBM. (n.d.). *Telco Customer Churn* [Data set]. Retrieved February 15, 2026, from https://raw.githubusercontent.com/IBM/telco-customer-churn-on-icp4d/master/data/Telco-Customer-Churn.csv

IBM. (n.d.). *Telco Customer Churn* [Data set]. Kaggle. Retrieved February 15, 2026, from https://www.kaggle.com/datasets/blastchar/telco-customer-churn

International Telecommunication Union (ITU). (2026). *Mobile phone subscriptions per 100 people* [Data set]. Processed by Our World in Data; original data from World Bank, *World Development Indicators*. Retrieved February 15, 2026, from https://archive.ourworldindata.org/20260202-181317/grapher/mobile-cellular-subscriptions-per-100-people.html

Metropolitan Transportation Authority. (2026). *MTA NYCT Paratransit Call Center Performance: Beginning 2016* [Data set]. Retrieved February 15, 2026, from https://data.ny.gov

World Bank. (2026). *Mobile cellular subscriptions (per 100 people)* [Data set]. World Development Indicators. Retrieved February 15, 2026, from https://data.worldbank.org/indicator/IT.CEL.SETS