**Statistical Verification**

**Numeric Attributes Statistical Verification**

customer\_id: The mean (500000.50 vs. 499289.11, -0.14%) isslightly lower in the sample, with similar standard deviations (288675.28 vs. 288930.14, +0.09%) and slight quartile shifts (Q1: 250000.75 vs. 249060.25, Q2: 500000.50 vs. 498793.50, Q3: 750000.25 vs. 750988).

age: The mean (48.50 vs. 48.63, +0.27%) is slightly higher in the sample, with similar standard deviations (17.87 vs. 17.97, +0.56%) and no quartile shifts (Q1: 33 vs. 33, Q2: 49 vs. 49, Q3: 64 vs. 64).

membership\_years: The mean (4.50 vs. 4.48, -0.44%) is slightly lower in the sample, with similar standard deviations (2.87 vs. 2.86, -0.35%) and a slight quartile shift in the median (Q1: 2 vs. 2, Q2: 4 vs. 5, Q3: 7 vs. 7).

number\_of\_children: The mean (2.00 vs. 1.99, -0.69%) is slightly lower in the sample, with identical standard deviations (1.41 vs. 1.41, 0.00%) and no quartile shifts (Q1: 1 vs. 1, Q2: 2 vs. 2, Q3: 3 vs. 3).

transaction\_id: The mean (499891.73 vs. 500528.98, +0.13%) is slightly higher in the sample, with similar standard deviations (288706.06 vs. 288038.00, -0.23%) and slight quartile shifts (Q1: 249878.75 vs. 254204, Q2: 499559.50 vs. 500069, Q3: 750071.25 vs. 750157.75).

product\_id: The mean (4999.56 vs. 4974.41, -0.50%) is slightly lower in the sample, with similar standard deviations (2886.80 vs. 2884.48, -0.08%) and slight quartile shifts (Q1: 2498 vs. 2486.75, Q2: 4999 vs. 4941, Q3: 7498 vs. 7466).

quantity: The mean (5.00 vs. 5.00, +0.04%) is nearly identical in the sample, with similar standard deviations (2.58 vs. 2.59, +0.19%)წ

unit\_price: The mean (500.26 vs. 498.32, -0.39%) is slightly lower in the sample, with similar standard deviations (288.46 vs. 289.31, +0.29%) and slight quartile shifts (Q1: 250.31 vs. 245.23, Q2: 500.41 vs. 497.67, Q3: 750.16 vs. 751.53).

discount\_applied: The mean (0.25 vs. 0.25, -0.57%) is slightly lower in the sample, with similar standard deviations (0.14 vs. 0.14, +0.26%) and a slight quartile shift in Q1 (Q1: 0.13 vs. 0.12, Q2: 0.25 vs. 0.25, Q3: 0.37 vs. 0.37).

transaction\_hour: The mean (11.51 vs. 11.55, +0.37%) is slightly higher in the sample, with similar standard deviations (6.92 vs. 6.91, -0.21%) and a slight quartile shift in Q1 (Q1: 5 vs. 6, Q2: 12 vs. 12, Q3: 18 vs. 18).

week\_of\_year: The mean (26.50 vs. 26.28, -0.85%) is slightly lower in the sample, with similar standard deviations (15.01 vs. 15.04, +0.24%) and slight quartile shifts (Q1: 14 vs. 13, Q2: 27 vs. 26, Q3: 39 vs. 39).

month\_of\_year: The mean (6.50 vs. 6.47, -0.35%) is slightly lower in the sample, with similar standard deviations (3.46 vs. 3.47, +0.51%) and a slight quartile shift in Q2 (Q1: 3 vs. 3, Q2: 7 vs. 6, Q3: 10 vs. 10).

avg\_purchase\_value: The mean (254.89 vs. 254.92, +0.01%) is nearly identical in the sample, with similar standard deviations (141.49 vs. 142.25, +0.53%) and slight quartile shifts (Q1: 132.22 vs. 131.63, Q2: 254.93 vs. 255.11, Q3: 377.35 vs. 379.35).

avg\_discount\_used: The mean (0.25 vs. 0.25, +0.27%) is slightly higher in the sample, with similar standard deviations (0.14 vs. 0.14, +0.27%) and no quartile shifts (Q1: 0.13 vs. 0.13, Q2: 0.25 vs. 0.25, Q3: 0.38 vs. 0.38).

online\_purchases: The mean (49.45 vs. 49.54, +0.19%) is slightly higher in the sample, with similar standard deviations (28.86 vs. 29.02, +0.53%) and slight quartile shifts (Q1: 24 vs. 24, Q2: 49 vs. 50, Q3: 74 vs. 75).

in\_store\_purchases: The mean (49.48 vs. 49.39, -0.19%) is slightly lower in the sample, with similar standard deviations (28.88 vs. 28.86, -0.08%) and a slight quartile shift in Q3 (Q1: 24 vs. 24, Q2: 49 vs. 49, Q3: 75 vs. 74).

avg\_items\_per\_transaction: The mean (5.50 vs. 5.50, -0.08%) is nearly identical in the sample, with similar standard deviations (2.60 vs. 2.59, -0.11%) and slight quartile shifts (Q1: 3.26 vs. 3.23, Q2: 5.50 vs. 5.51, Q3: 7.75 vs. 7.72).

avg\_transaction\_value: The mean (255.12 vs. 256.60, +0.58%) is slightly higher in the sample, with similar standard deviations (141.43 vs. 141.34, -0.06%) and slight quartile shifts (Q1: 132.51 vs. 133.84, Q2: 255.23 vs. 257.31, Q3: 377.67 vs. 378.86).

total\_returned\_items: The mean (4.50 vs. 4.44, -1.26%) is slightly lower in the sample, with similar standard deviations (2.87 vs. 2.88, +0.18%) and no quartile shifts (Q1: 2 vs. 2, Q2: 4 vs. 4, Q3: 7 vs. 7).

total\_returned\_value: The mean (500.39 vs. 502.03, +0.33%) is slightly higher in the sample, with similar standard deviations (288.72 vs. 288.06, -0.23%) and slight quartile shifts (Q1: 250.63 vs. 252.31, Q2: 500.40 vs. 501.02, Q3: 750.39 vs. 750.32).

total\_sales: The mean (5056.06 vs. 5068.05, +0.24%) is slightly higher in the sample, with similar standard deviations (2859.10 vs. 2855.31, -0.13%) and slight quartile shifts (Q1: 2577.87 vs. 2593.34, Q2: 5059.70 vs. 5082.49, Q3: 7534.80 vs. 7531.47).

total\_transactions: The mean (49.99 vs. 49.87, -0.25%) is slightly lower in the sample, with similar standard deviations (28.57 vs. 28.66, +0.31%) and no quartile shifts (Q1: 25 vs. 25, Q2: 50 vs. 50, Q3: 75 vs. 75).

total\_items\_purchased: The mean (250.04 vs. 250.22, +0.07%) is nearly identical in the sample, with similar standard deviations (143.98 vs. 144.03, +0.03%) and slight quartile shifts (Q1: 125 vs. 127, Q2: 250 vs. 250, Q3: 375 vs. 374).

total\_discounts\_received: The mean (499.67 vs. 498.82, -0.17%) is slightly lower in the sample, with similar standard deviations (288.58 vs. 288.76, +0.06%) and slight quartile shifts (Q1: 249.76 vs. 247.29, Q2: 499.51 vs. 498.26, Q3: 749.54 vs. 747.59).

avg\_spent\_per\_category: The mean (505.18 vs. 507.82, +0.52%) is slightly higher in the sample, with similar standard deviations (286.06 vs. 287.27, +0.42%) and slight quartile shifts (Q1: 257.24 vs. 260.19, Q2: 505.14 vs. 508.84, Q3: 753.06 vs. 757.52).

max\_single\_purchase\_value: The mean (505.00 vs. 507.75, +0.54%) is slightly higher in the sample, with similar standard deviations (286.07 vs. 284.80, -0.44%) and slight quartile shifts (Q1: 256.84 vs. 261.15, Q2: 505.22 vs. 511.44, Q3: 753.21 vs. 754.49).

min\_single\_purchase\_value: The mean (5.04 vs. 5.02, -0.56%) is slightly lower in the sample, with similar standard deviations (2.86 vs. 2.87, +0.65%) and slight quartile shifts (Q1: 2.57 vs. 2.50, Q2: 5.04 vs. 5.02, Q3: 7.51 vs. 7.49).

product\_rating: The mean (3.00 vs. 3.00, +0.06%) is nearly identical in the sample, with similar standard deviations (1.15 vs. 1.15, -0.08%) and no quartile shifts (Q1: 2 vs. 2, Q2: 3 vs. 3, Q3: 4 vs. 4).

product\_review\_count: The mean (499.24 vs. 498.44, -0.16%) is slightly lower in the sample, with similar standard deviations (288.45 vs. 289.19, +0.26%) and slight quartile shifts (Q1: 250 vs. 248, Q2: 499 vs. 495, Q3: 749 vs. 751).

product\_stock: The mean (49.52 vs. 49.48, -0.08%) is nearly identical in the sample, with similar standard deviations (28.88 vs. 29.00, +0.42%) and a slight quartile shift in Q1 (Q1: 25 vs. 24, Q2: 49 vs. 49, Q3: 75 vs. 75).

product\_return\_rate: The mean (0.25 vs. 0.25, +0.04%) is nearly identical in the sample, with similar standard deviations (0.14 vs. 0.15, +0.50%) and no quartile shifts (Q1: 0.13 vs. 0.13, Q2: 0.25 vs. 0.25, Q3: 0.38 vs. 0.38).

product\_weight: The mean (5.05 vs. 5.03, -0.44%) is slightly lower in the sample, with similar standard deviations (2.86 vs. 2.85, -0.14%) and slight quartile shifts (Q1: 2.58 vs. 2.54, Q2: 5.06 vs. 5.03, Q3: 7.53 vs. 7.49).

product\_shelf\_life: The mean (181.88 vs. 182.21, +0.18%) is slightly higher in the sample, with similar standard deviations (105.23 vs. 105.20, -0.03%) and no quartile shifts (Q1: 91 vs. 91, Q2: 182 vs. 183, Q3: 273 vs. 273).

promotion\_id: The mean (499.92 vs. 505.16, +1.05%) is slightly higher in the sample, with similar standard deviations (288.45 vs. 287.70, -0.26%) and slight quartile shifts (Q1: 250 vs. 259, Q2: 500 vs. 507, Q3: 750 vs. 753).

customer\_zip\_code: The mean (54993.64 vs. 55061.54, +0.12%) is slightly higher in the sample, with similar standard deviations (25975.81 vs. 26005.75, +0.12%) and slight quartile shifts (Q1: 32477.75 vs. 32756.25, Q2: 54966 vs. 54789, Q3: 77493 vs. 77617.50).

store\_zip\_code: The mean (54972.77 vs. 54892.82, -0.15%) is slightly lower in the sample, with similar standard deviations (25981.48 vs. 25997.39, +0.06%) and slight quartile shifts (Q1: 32473 vs. 32266.75, Q2: 54961 vs. 54997, Q3: 77451 vs. 77304.25).

distance\_to\_store: The mean (49.98 vs. 49.87, -0.22%) is slightly lower in the sample, with similar standard deviations (28.86 vs. 28.86, -0.02%) and slight quartile shifts (Q1: 24.97 vs. 24.99, Q2: 49.96 vs. 49.72, Q3: 74.95 vs. 74.80).

customer\_support\_calls: The mean (9.50 vs. 9.45, -0.48%) is slightly lower in the sample, with similar standard deviations (5.76 vs. 5.76, -0.02%) and no quartile shifts (Q1: 4 vs. 4, Q2: 9 vs. 9, Q3: 14 vs. 14).

website\_visits: The mean (49.51 vs. 49.42, -0.18%) is slightly lower in the sample, with similar standard deviations (28.87 vs. 28.86, -0.02%) and a slight quartile shift in Q3 (Q1: 25 vs. 24, Q2: 50 vs. 50, Q3: 75 vs. 74).

days\_since\_last\_purchase: The mean (182.03 vs. 181.56, -0.26%) is slightly lower in the sample, with similar standard deviations (105.36 vs. 105.63, +0.26%) and a slight quartile shift in Q1 (Q1: 91 vs. 89, Q2: 182 vs. 182, Q3: 273 vs. 273).

**Categorical Attributes Statistical Verification**

gender: The top category frequency (Other: 33.27% vs. 33.60%, +0.99%) is slightly higher in the sample, with similar category distributions (Male: 33.43% vs. 33.25%, Female: 33.29% vs. 33.16%) and no missing categories.

income\_bracket: The top category frequency (High: 33.31% vs. 34.05%, +2.20%) is slightly higher in the sample, with similar category distributions (Medium: 33.39% vs. 32.61%, Low: 33.30% vs. 33.35%) and no missing categories.

loyalty\_program: The top category frequency (Yes: 50.00% vs. 50.36%, +0.72%) is slightly higher in the sample, with similar category distributions (No: 50.00% vs. 49.64%) and no missing categories.

churned: The top category frequency (Yes: 49.85% vs. 50.68%, +1.67%) is slightly higher in the sample, with similar category distributions (No: 50.15% vs. 49.32%) and no missing categories.

marital\_status: The top category frequency (Divorced: 33.42% vs. 33.63%, +0.61%) is slightly higher in the sample, with similar category distributions (Married: 33.29% vs. 33.07%, Single: 33.29% vs. 33.31%) and no missing categories.

education\_level: The top category frequency (PhD: 25.05% vs. 25.31%, +1.04%) is slightly higher in the sample, with similar category distributions (High School: 25.06% vs. 24.52%, Bachelor’s: 24.99% vs. 24.91%, Master’s: 24.89% vs. 25.27%) and no missing categories.

occupation: The top category frequency (Employed: 25.08% vs. 25.37%, +1.13%) is slightly higher in the sample, with similar category distributions (Self-Employed: 25.08% vs. 25.13%, Unemployed: 24.99% vs. 24.77%, Retired: 24.85% vs. 24.75%) and no missing categories.

product\_category: The top category frequency (Toys: 20.06% vs. 20.52%, +2.30%) is slightly higher in the sample, with similar category distributions (Clothing: 20.04% vs. 19.68%, Furniture: 20.03% vs. 19.80%, Electronics: 19.94% vs. 19.84%, Groceries: 19.93% vs. 20.17%) and no missing categories.

payment\_method: The top category frequency (Credit Card: 24.95% vs. 25.50%, +2.18%) is slightly higher in the sample, with similar category distributions (Mobile Payment: 25.04% vs. 25.30%, Debit Card: 25.01% vs. 24.93%, Cash: 24.99% vs. 24.28%) and no missing categories.

store\_location: The top category frequency (Location B: 24.97% vs. 25.42%, +1.79%) is slightly higher in the sample, with similar category distributions (Location C: 25.03% vs. 25.08%, Location D: 25.00% vs. 24.79%, Location A: 24.99% vs. 24.73%) and no missing categories.

day\_of\_week: The top category frequency (Tuesday: 14.38% vs. 14.68%, +2.05%) is slightly higher in the sample, with similar category distributions (Sunday: 14.31% vs. 14.27%, Friday: 14.31% vs. 14.40%, Thursday: 14.28% vs. 13.87%, Saturday: 14.26% vs. 14.47%, Monday: 14.23% vs. 14.36%, Wednesday: 14.22% vs. 13.97%) and no missing categories.

purchase\_frequency: The top category frequency (Daily: 24.97% vs. 25.29%, +1.29%) is slightly higher in the sample, with similar category distributions (Yearly: 25.10% vs. 24.93%, Monthly: 24.97% vs. 25.09%, Weekly: 24.95% vs. 24.69%) and no missing categories.

preferred\_store: The top category frequency (Location C: 25.07% vs. 25.55%, +1.90%) is slightly higher in the sample, with similar category distributions (Location B: 25.08% vs. 24.84%, Location A: 24.96% vs. 24.99%, Location D: 24.89% vs. 24.63%) and no missing categories.

product\_name: The top category frequency (Product C: 25.01% vs. 25.20%, +0.77%) is slightly higher in the sample, with similar category distributions (Product A: 25.04% vs. 24.76%, Product B: 25.02% vs. 24.93%, Product D: 24.92% vs. 25.12%) and no missing categories.

product\_brand: The top category frequency (Brand Z: 33.32% vs. 33.50%, +0.52%) is slightly higher in the sample, with similar category distributions (Brand Y: 33.39% vs. 33.39%, Brand X: 33.29% vs. 33.12%) and no missing categories.

product\_size: The top category frequency (Medium: 33.37% vs. 33.54%, +0.51%) is slightly higher in the sample, with similar category distributions (Large: 33.41% vs. 33.49%, Small: 33.22% vs. 32.98%) and no missing categories.

product\_color: The top category frequency (Red: 20.02% vs. 20.29%, +1.32%) is slightly higher in the sample, with similar category distributions (Green: 20.09% vs. 19.94%, Black: 20.00% vs. 20.18%, Blue: 19.98% vs. 19.56%, White: 19.91% vs. 20.05%) and no missing categories.

product\_material: The top category frequency (Glass: 24.97% vs. 25.38%, +1.62%) is slightly higher in the sample, with similar category distributions (Plastic: 25.13% vs. 24.75%, Metal: 24.97% vs. 24.98%, Wood: 24.94% vs. 24.90%) and no missing categories.

promotion\_type: The top category frequency (Flash Sale: 33.41% vs. 33.94%, +1.57%) is slightly higher in the sample, with similar category distributions (20% Off: 33.36% vs. 33.22%, Buy One Get One Free: 33.23% vs. 32.85%) and no missing categories.

promotion\_effectiveness: The top category frequency (Low: 33.28% vs. 33.65%, +1.11%) is slightly higher in the sample, with similar category distributions (Medium: 33.43% vs. 33.59%, High: 33.29% vs. 32.77%) and no missing categories.

promotion\_channel: The top category frequency (Online: 33.38% vs. 33.87%, +1.47%) is slightly higher in the sample, with similar category distributions (Social Media: 33.34% vs. 32.66%, In-store: 33.28% vs. 33.48%) and no missing categories.

promotion\_target\_audience: The top category frequency (New Customers: 50.05% vs. 50.29%, +0.48%) is slightly higher in the sample, with similar category distributions (Returning Customers: 49.95% vs. 49.71%) and no missing categories.

customer\_city: The top category frequency (City C: 25.00% vs. 25.35%, +1.40%) is slightly higher in the sample, with similar category distributions (City B: 25.07% vs. 24.89%, City A: 24.99% vs. 25.05%, City D: 24.94% vs. 24.72%) and no missing categories.

customer\_state: The top category frequency (State Z: 33.47% vs. 33.97%, +1.49%) is slightly higher in the sample, with similar category distributions (State Y: 33.30% vs. 33.05%, State X: 33.22% vs. 32.98%) and no missing categories.

store\_city: The top category frequency (City B: 25.14% vs. 25.63%, +1.93%) is slightly higher in the sample, with similar category distributions (City C: 24.98% vs. 25.13%, City D: 24.94% vs. 24.92%, City A: 24.93% vs. 24.34%) and no missing categories.

store\_state: The top category frequency (State Z: 33.27% vs. 33.67%, +1.20%) is slightly higher in the sample, with similar category distributions (State X: 33.37% vs. 33.50%, State Y: 33.36% vs. 32.83%) and no missing categories.

holiday\_season: The top category frequency (Yes: 49.97% vs. 50.19%, +0.43%) is slightly higher in the sample, with similar category distributions (No: 50.03% vs. 49.82%) and no missing categories.

season: The top category frequency (Fall: 24.97% vs. 26.00%, +4.09%) is higher in the sample, with similar category distributions (Winter: 25.10% vs. 24.65%, Summer: 25.02% vs. 24.67%, Spring: 24.92% vs. 24.70%) and no missing categories.

weekend: The top category frequency (Yes: 50.10% vs. 50.49%, +0.77%) is slightly higher in the sample, with similar category distributions (No: 49.90% vs. 49.52%) and no missing categories.

email\_subscriptions: The top category frequency (Yes: 50.11% vs. 50.65%, +1.07%) is slightly higher in the sample, with similar category distributions (No: 49.89% vs. 49.36%) and no missing categories.

app\_usage: The top category frequency (Medium: 33.42% vs. 34.09%, +1.99%) is slightly higher in the sample, with similar category distributions (High: 33.31% vs. 33.04%, Low: 33.27% vs. 32.88%) and no missing categories.

social\_media\_engagement: The top category frequency (Low: 33.26% vs. 33.63%, +1.09%) is slightly higher in the sample, with similar category distributions (Medium: 33.41% vs. 33.09%, High: 33.33% vs. 33.29%) and no missing categories.

**Summary of Similarities**

**Numerical Attributes**

* Mean Differences: Most numerical columns have mean differences of less than 0.5% (e.g., age: +0.27%, total\_sales: +0.24%), with the largest differences in total\_returned\_items (-1.26%) and promotion\_id (+1.05%). These are within acceptable sampling variation for a 5.5% sample.
* Standard Deviations: Standard deviations are highly consistent, with differences typically <0.7% (e.g., min\_single\_purchase\_value: +0.65%), indicating similar variability.
* Quartile and Range: Quartiles and ranges are nearly identical, with minor shifts (e.g., unit\_price Q1: 250.31 vs. 245.23) that do not significantly alter distributions. The sample captures the full range of values for most columns, with slight truncations in some cases (e.g., unit\_price max: 1000 vs. 999.99).
* Key Insight: The numerical attributes are highly representative, making the sampled dataset suitable for analyses involving sales, discounts, customer behavior, and other numerical metrics.

**Categorical Attributes**

* Top Category Frequency Differences: Most categorical columns show top category frequency differences of less than 2% (e.g., gender: +0.99%, loyalty\_program: +0.72%), with the largest difference in season (+4.09% for Fall). These differences are within expected sampling variation.
* Category Distributions: All categories present in the full dataset are preserved in the sampled dataset, ensuring no loss of diversity. Distributions are similar, with minor shifts (e.g., churned: Yes increases from 49.85% to 50.68%).
* Key Insight: The categorical attributes are generally representative, with slight over-representations in certain categories (e.g., season: Fall, product\_category: Toys) that are unlikely to significantly impact most analyses.

**Overall Representativeness**

* The sampled dataset closely mirrors the full dataset across both numerical and categorical attributes, with differences small enough to ensure reliable results for general analyses.
* The largest numerical difference (total\_returned\_items: -1.26%) and categorical difference (season: +4.09%) are still within acceptable bounds for a 5.5% sample, especially given the preservation of all categories and ranges.

**Outcome**

The sampled dataset is sufficiently representative of the full dataset for most analytical purposes. The numerical attributes show minimal differences in means, standard deviations, and distributions, while the categorical attributes preserve all categories with minor frequency shifts.