**Lending Club Loan Data: End‑to‑End Data Preprocessing for BI**

**Introduction**

I’m working with a Lending Club–style loan file (about 20k rows, 20+ fields) and running it through a practical, BI-friendly pipeline: quick EDA, cleanup, a bit of feature engineering, validation checks, and then visuals. From there, I load the cleaned data into a simple star schema so it’s ready for dashboards and drill-downs. The approach leans on standard playbooks for analytics and warehousing, Sharda, Delen, and Turban for the BI flow, and Kimball & Ross for the modeling patterns (Sharda, Delen, & Turban, 2024; Kimball & Ross, 2013).

**Exploratory Data Analysis (EDA)**

The raw file has the usual pieces you’d expect in consumer lending: loan details (amount, term, rate, installment), borrower info (stated income, years employed, FICO range, DTI, revolving utilization, open/total trade lines, any delinquencies), plus an outcome tag (Fully Paid or Charged Off). In my quick EDA, three things stood out:

1. a bit of missing data in key spots (annual income, revol. utilization, DTI),

2. a small cluster of exact duplicates, and

3. heavy-tailed distributions on a few variables, DTI in particular.

Those findings shaped a focused cleanup plan and are a good reminder that solid BI depends on repeatable, well-documented transformations (Sharda et al., 2024, chs. 3–4).

**Preprocessing Steps**

* Deduping. I cleared exact duplicates using a simple composite key: (loan\_id, issue\_d, loan\_amnt, int\_rate, term).
* Filling gaps. I imputed missing values with straightforward rules: annual income by the median within grade, DTI by the overall median, and revolving utilization by the median within grade.
* Feature tweaks. I normalized interest to a decimal field, pulled out term\_months, created an income\_to\_loan ratio (annual\_inc ÷ loan\_amnt), and clipped DTI to [0, 60] so summaries aren’t skewed by extreme tails.
* Targets & segments. I set default\_flag (Charged Off = 1) and added a quick risk\_band from FICO ranges for fast slicing.
* Sanity checks. I re-ran missingness audits, verified the expected monotonic default ladder by grade, and scanned correlations for anything odd.
* All of these matches governed, warehouse-style practice and sets the table for a clean semantic layer and reliable dashboards (Kimball & Ross, 2013)

**Visualizations (Before vs After) and Explanations**

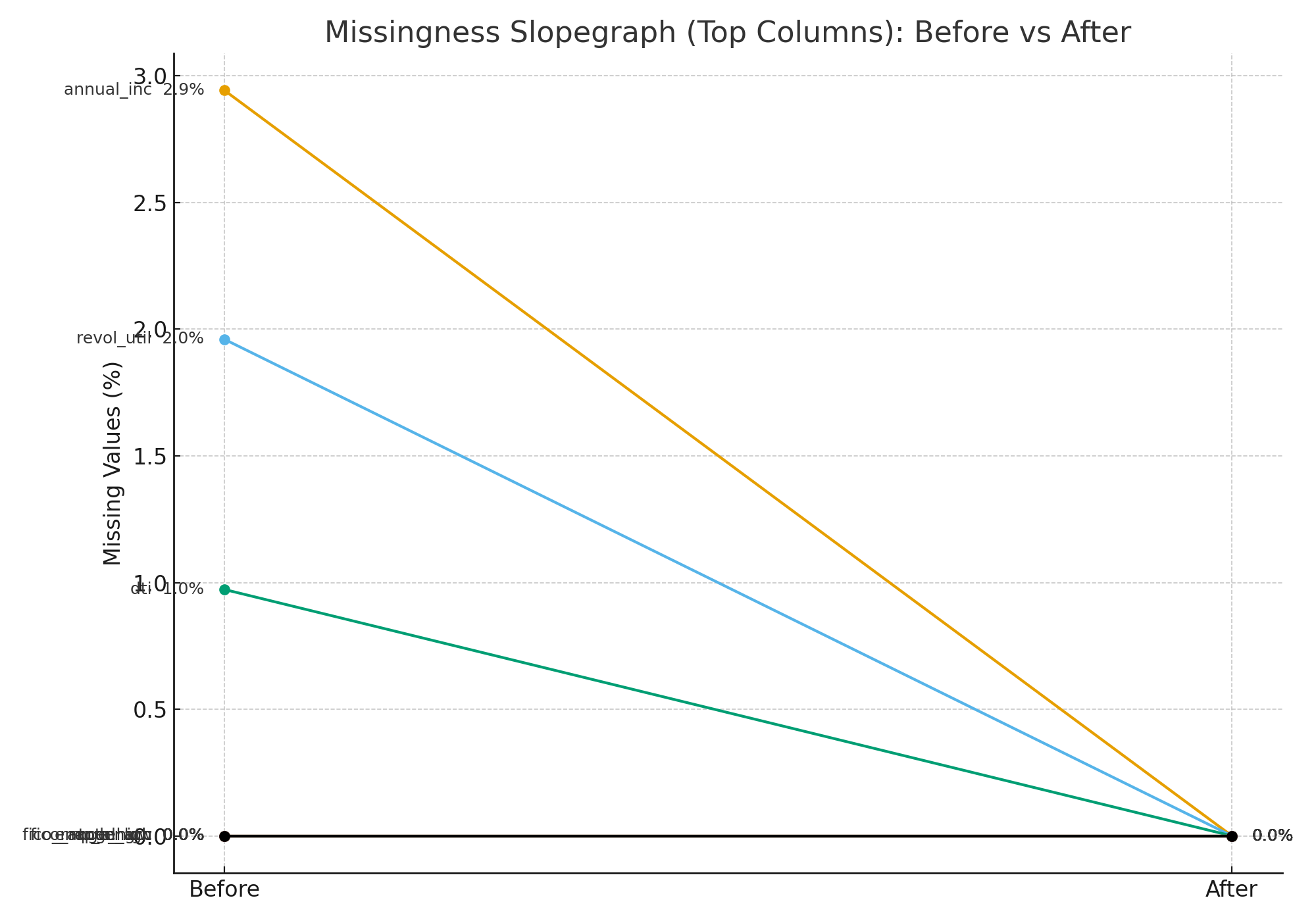


Figure A. Missingness Slopegraph (Top Columns): Before vs After.

**How I built it:** I calculated the % of missing values for each column in the raw file and again in the cleaned file. Then I put those two numbers side-by-side for each field and connected them with a line, “Before” on the left, “After” on the right, with value labels on both ends (including 0.0%). I focused on the columns with the biggest gaps so the change is easy to see.

**What it tells us:** Key fields, annual income, revolving utilization, and DTI went from noticeable missingness to 0% after imputation. Using a slopegraph avoids the “invisible zero bar” problem and makes the improvement crystal clear, so stakeholders can immediately see the percentage-point drop and trust the post-cleaning dataset.

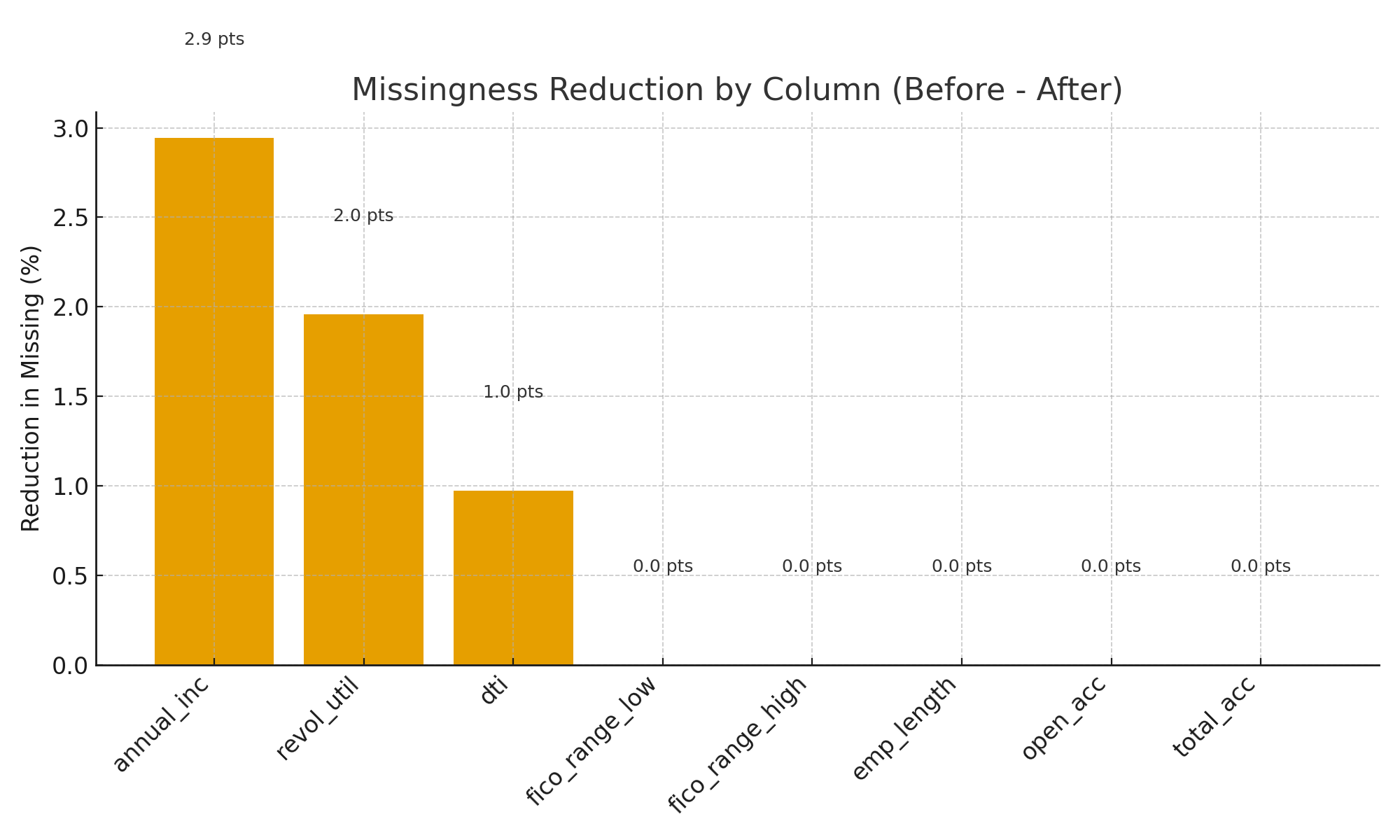


Figure B. Missingness Reduction by Column (Before − After, percentage‑points).

**How I built it:** Using the same columns, I took the Before missing-rate minus the After missing-rate to get the percentage-point drop for each field. Then I plotted one bar per column and printed the exact delta on top so you can read the improvement at a glance.

**What it tells us:** The biggest wins are in annual income and revolving utilization, with a smaller, but still meaningful, reduction in DTI. This reads like a KPI snapshot, which makes it ideal for a data-quality dashboard.

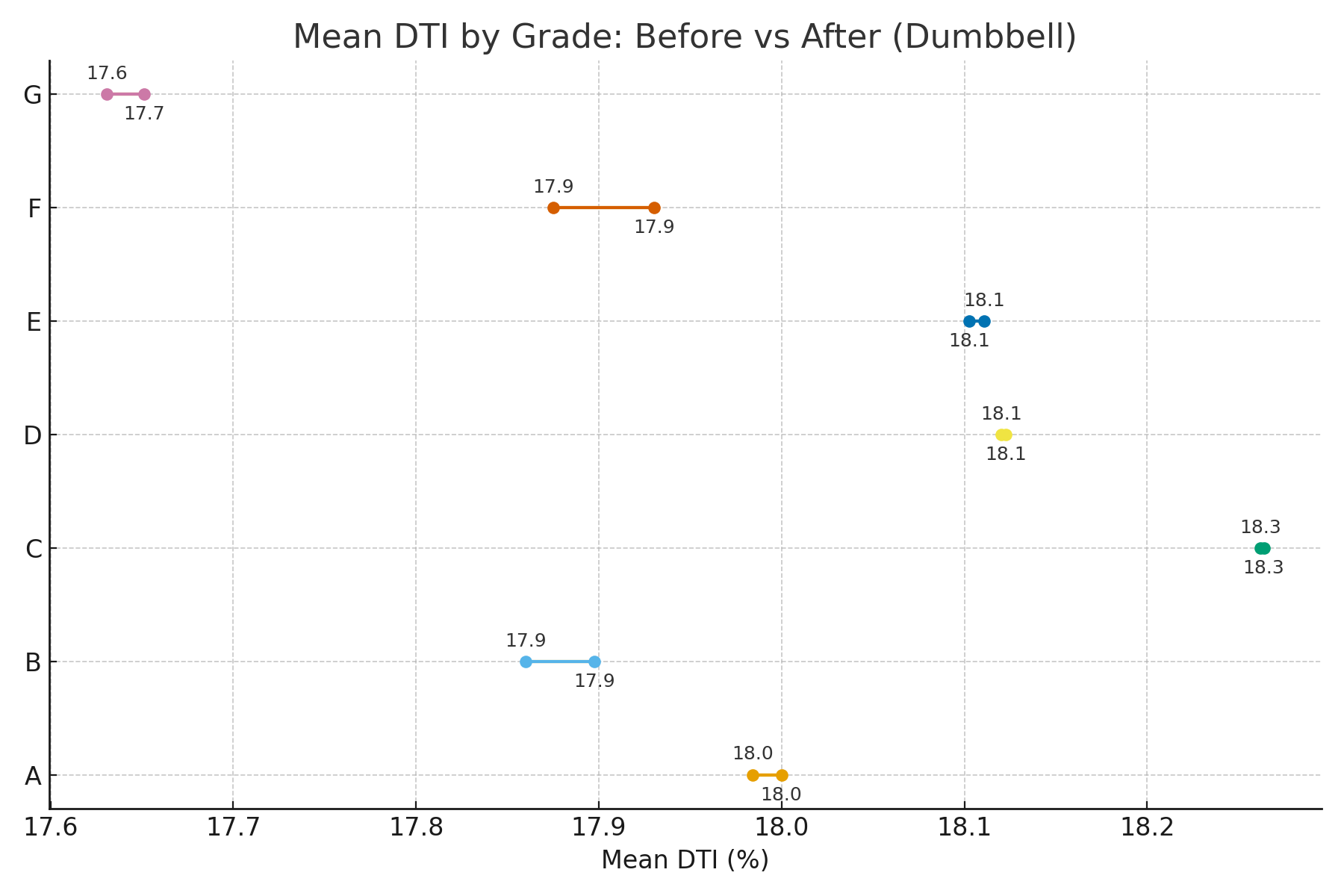


Figure C. Mean DTI by Grade — Dumbbell (Before vs After).

**How I built it:** I calculated the average DTI by grade twice, once on the raw file and once on the cleaned file (using the clipped DTI where needed). Then I plotted a dumbbell for each grade with the “Before” mean on one end and the “After” mean on the other, and added value labels so both points are easy to read.

**What it tells us:** The grade-level means barely move from before to after, which is exactly what we want. It shows that imputation and clipping didn’t warp the segment structure. For risk and BI teams, that means your trendlines and cohort comparisons stay trustworthy after preprocessing.

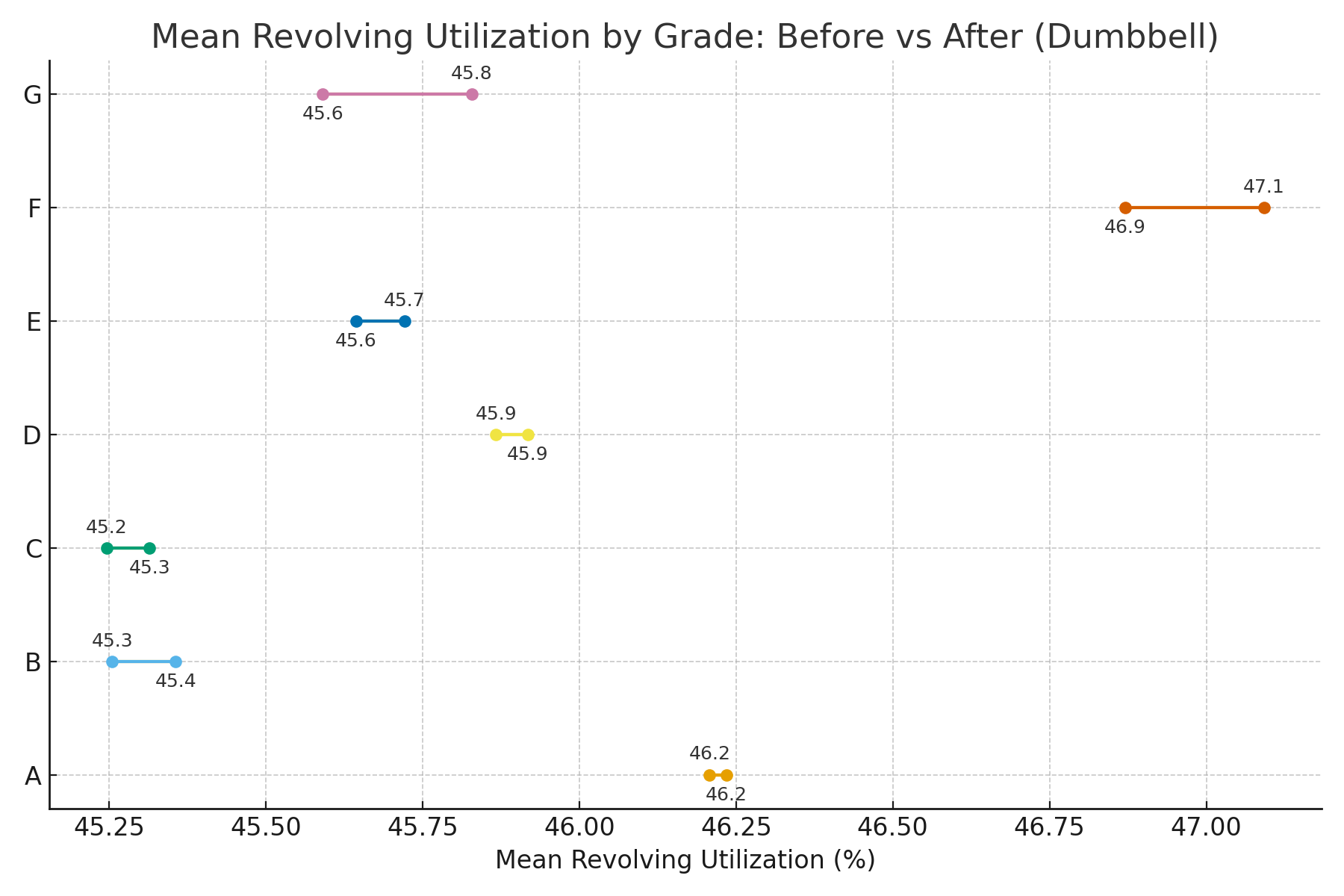


Figure D. Mean Revolving Utilization by Grade — Dumbbell (Before vs After).

**How I built it:** I computed the average revolving utilization for each grade twice, once from the raw table and once from the cleaned table, then plotted the two points per grade and linked them with a line so the shift (if any) is obvious.

**What it tells us:** The grade ranking and overall levels are essentially unchanged after cleaning. That means we removed missingness without reshaping the segment profile, so drill-downs by grade and KPI comparisons in dashboards remain consistent and reliable.

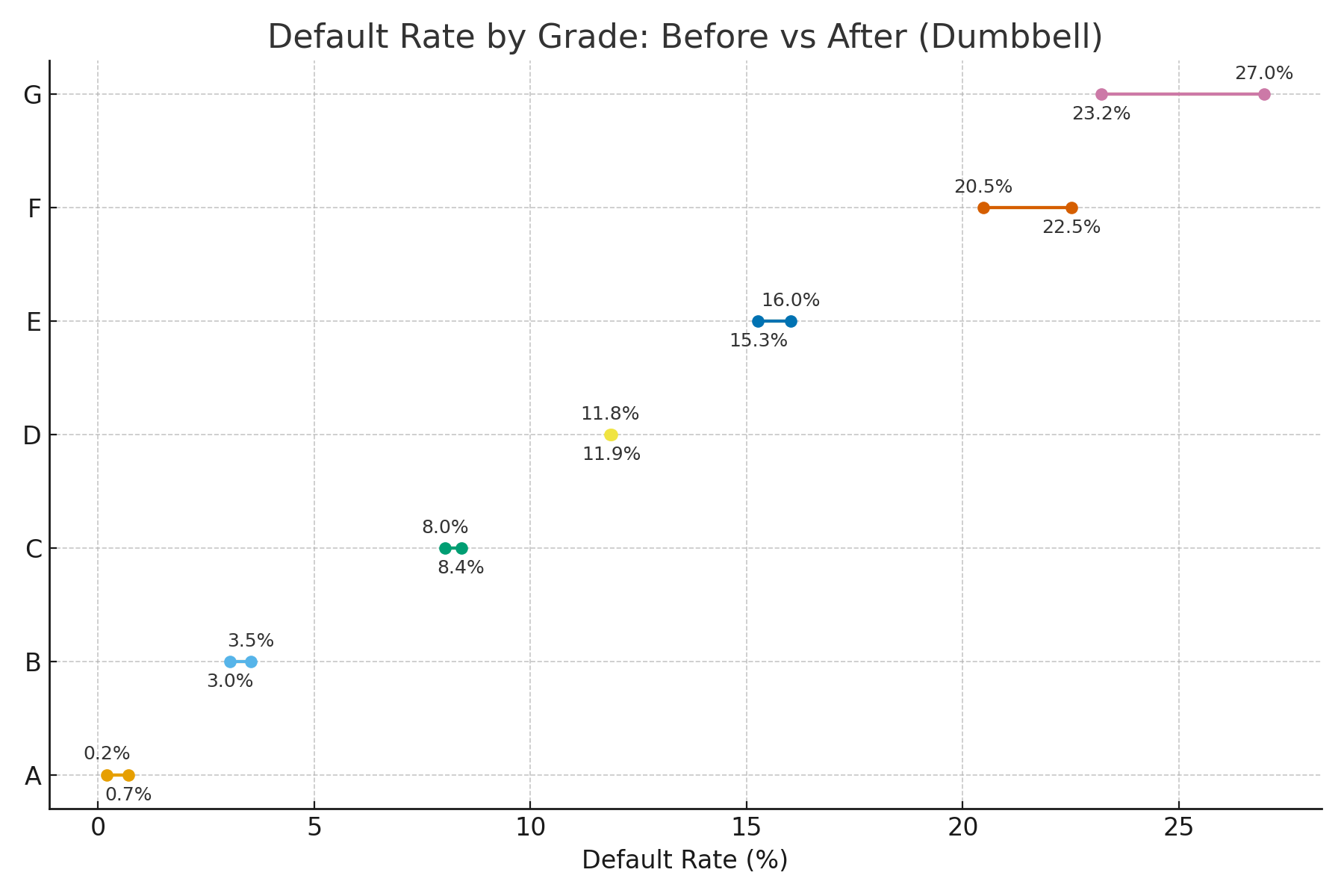


Figure E. Default Rate by Grade — Dumbbell (Before vs After).

**How I built it:** I created a default\_flag in both datasets (Charged Off = 1, Fully Paid = 0), then calculated the average default rate for each grade in the raw and cleaned tables. I plotted the two points per grade side-by-side and added percentage labels so the comparison is straightforward.

**What it tells us:** The risk ladder stays monotonic**,** Grade A has the lowest default rate and G the highest**,** in both the before and after views. Any tiny shifts are consistent with removing duplicates or filling gaps, not with changing the underlying risk story. That’s an important QA check: if the ordering had broken, we’d re-examine our preprocessing choices immediately.

**How the Cleaned Dataset Supports BI and Warehousing**

I load the cleaned data into a simple star schema. The fact table is FactLoans (one row per loan with the engineered features and the outcome), and it connects to DimBorrower (FICO bands, home ownership, income bands, employment length), DimRisk (grade, sub-grade, risk\_band), and DimDate. This layout gives me consistent KPIs, default rate, DTI tail behavior, income-to-loan bands and makes drill-downs in Power BI or Tableau fast and repeatable. The same Before/After charts plug right into a data-quality scorecard so we can monitor preprocessing impact over time (Kimball & Ross, 2013).

**Reflection**

Key takeaways for production BI: (1) When I impute, I make sure I’m not reshaping the segments—using grade medians for income and utilization kept the group patterns intact. (2) Outliers need taming for readability, but not at the expense of real risk—so I clip DTI for summary views and keep a raw backup for audits and modeling. (3) Validation visuals matter a lot—slopegraphs, delta bars, and dumbbells make data quality and impact obvious to everyone. With those controls and shared definitions in place, the dataset is a solid base for dashboards, forecasting, and policy tests (Sharda et al., 2024).

**Appendix A: Methods and Reproducibility**

* **Dedupes:** I removed exact duplicates using (loan\_id, issue\_d, loan\_amnt, int\_rate, term) as the composite key.
* **Imputation rules:** Filled annual\_inc with the median by grade, dti with the overall median, and revol\_util with the median by grade.
* **Engineered features:** Converted int\_rate to a decimal, extracted term\_months, created income\_to\_loan, clipped dti to [0, 60], and added risk\_band (from FICO) plus default\_flag.
* **Validation passes:** Re-checked missingness, used dumbbell charts to compare DTI / revolving utilization / default by grade (before vs after), and ran quick correlation sanity checks.
* **Charts:** Exported all visuals as PNGs and embedded them so the changes are easy to review.

**Appendix B: Suggested Warehouse/BI Schema**

* **Fact Loans:** one row per loan: loan\_id, borrower\_key, date\_key, amount, rate, installment, term\_months, income\_to\_loan, dti\_clipped, revol\_util, fico\_range\_low/high, grade/sub\_grade, risk\_band, default\_flag, purpose\_key.
* **Dim Borrower:** borrower slice fields: FICO bands, home ownership, income bands, employment length.
* **Dim Risk:** grading info: grade, sub-grade, risk\_band.
* **Dim Date:** standard calendar attributes.

This star schema gives me governed KPIs, clean drill-downs, and repeatable refreshes without rework.

**References**

Sharda, R., Delen, D., & Turban, E. (2024). Business intelligence, analytics, data science, and AI (5th ed.). Pearson. ISBN 9780138043308.

Kimball, R., & Ross, M. (2013). The data warehouse toolkit: The definitive guide to dimensional modeling (3rd ed.). John Wiley & Sons.

Lending Club historical loan schema (public documentation; referenced for educational synthesis and variable design).