**Histogram Analysis**

This document provides a summary interpretation of the histograms generated for each feature in the dataset, along with recommendations for machine learning modeling.

**Key Observations and Recommendations:**

**Numerical Features:**

| **Column** | **Distribution & Observations** | **Recommendation for ML** |
| --- | --- | --- |
| **age\_skewed** | Right-skewed; older customers are fewer but show higher churn risk. | **Keep**. Consider **binning** (e.g., young, middle-aged, old) or using **as-is** for tree models. |
| **creditscorerange** | Fairly uniform; no strong peaks or concentration. | **Optional**. Low correlation; keep for tree-based models or **drop** if model performance doesn't improve. |
| **tenurerange** | Bimodal distribution; peaks at low and high tenures. | **Keep**. Useful indicator of customer loyalty; try **no transformation** first. |
| **numofproducts** | Skewed; most customers have 1–2 products. | **Keep**. Important feature; especially useful for tree-based models. |
| **balancerange** | Highly right-skewed; most customers have low balances. | **Keep**. For linear models, apply **log transformation**; for trees, **use as-is**. |
| **estimatedsalaryrange** | Fairly uniform with some minor gaps. | **Optional**. Weak direct signal; keep for tree models. Might be dropped after feature importance analysis. |

**Categorical Features (Encoded):**

| **Column** | **Distribution & Observations** | **Recommendation for ML** |
| --- | --- | --- |
| **gender\_label** | Roughly balanced; slight difference between male and female. | **Optional**. Weak predictor; keep for completeness but could be dropped if feature selection is needed. |
| **isactivemember** | Slight imbalance; more active members than inactive. | **Keep**. Strong relationship with churn; important feature. |
| **hascrcard** | Slightly more customers have a credit card. | **Optional**. Very low correlation; can **keep** for testing but **drop** if model doesn't improve. |
| **geography\_germany** | Smaller group but more churn-prone. | **Keep**. Important geographical signal for churn. |
| **geography\_spain** | Smaller group; less churn-prone. | **Keep**. Use alongside Germany for geographical separation. |
| **geography\_france** | Majority group; relatively neutral churn behavior. | **Drop**. To avoid dummy variable trap (keep only 2 of 3 geography features). |
| **exited** (Target Variable) | Imbalanced (more non-churners). | **Target variable**. Handle imbalance using techniques like **SMOTE**, **class weighting**, or **stratified sampling**. |

**General Recommendations:**

* Apply scaling or transformations (like log or binning) **only if needed** depending on model type (linear vs. tree-based).
* Monitor feature importance after first modeling run; consider removing low-impact features.
* Address target class imbalance during training to avoid bias towards majority class.
* Retain geographic and activity features — they show clear separations valuable for predictive modeling.

