**Comprehensive Code Description: Customer Churn Analysis**

**Data Segmentation**

The analysis begins by dividing the customer dataset into two distinct groups using the 'Exited' column as the primary segmentation criterion. The first group (df\_0) comprises customers who have maintained their relationship with the bank, while the second group (df\_1) consists of customers who have churned. This fundamental separation enables comparative analysis between retained and lost customers across multiple demographic and financial dimensions.

* Customers are split using the Exited column.
  + Code:
    - df[df['Exited']==0] → df\_0 (stayed)
    - df[df['Exited']==1] → df\_1 (left)

**Hypothesis 1: Age**

* Visualize distributions with histograms + KDE.
  + Code: sns.histplot()
* Compute mean & std for each group.
  + Code: .mean(), .std()
* Run statistical test.
  + Code: stats.ttest\_ind(df\_0['Age'], df\_1['Age'], equal\_var=False)
* Validate with bootstrapping.
  + Code: bs\_choice() → resamples Age for bootstrap p-values.
* **Result:** Age shows **strong, statistically significant differences** (older customers churn more).

**Hypothesis 2: Credit Score**

* Compare distributions with histograms.
  + Code: sns.histplot(df\_0['CreditScore']), sns.histplot(df\_1['CreditScore'])
* Calculate descriptive stats.
  + Code: .mean(), .std()
* Perform statistical test.
  + Code: stats.ttest\_ind(df\_0['CreditScore'], df\_1['CreditScore'], equal\_var=False)
* **Result:** Credit score differences are **weak and not consistently significant**.

**Hypothesis 3: Balance**

* Analyze balances with all accounts.
  + Code: sns.histplot(df\_0['Balance']), sns.histplot(df\_1['Balance'])
  + stats.ttest\_ind(df\_0['Balance'], df\_1['Balance'], equal\_var=False)
* Repeat excluding zero balances.
  + Code: df\_0[df\_0['Balance']>0], df\_1[df\_1['Balance']>0] + t-test.
* **Result:** Balance is a **very strong churn predictor** — customers with higher balances tend to leave more.

**Hypothesis 4: Estimated Salary**

* Visualize distributions.
  + Code: sns.histplot(df\_0['EstimatedSalary']), sns.histplot(df\_1['EstimatedSalary'])
* Perform Welch’s t-test.
  + Code: stats.ttest\_ind(df\_0['EstimatedSalary'], df\_1['EstimatedSalary'], equal\_var=False)
* Apply bootstrap for robust validation.
  + Code: bs\_choice() again.
* **Result:** Salary shows **no meaningful differences** between groups.

**Statistical Framework**

* Threshold: α = 0.05 for rejecting null hypotheses.
* Both **p-values** and **effect sizes (Cohen’s d)** guide interpretation.
  + Code: cohens\_d(x, y)

**Comprehensive Conclusion**

* Ranking features by predictive strength (based on effect size & p-values):
  1. **Balance** (strongest predictor, high effect size, significant p-value)
  2. **Age** (moderately strong predictor, significant p-value)
  3. Credit Score (weak, inconsistent significance)
  4. Salary (not significant)

➡️ **Final Highlight:** The **Balance** variable is the **strongest churn predictor**, followed by **Age**. Salary and Credit Score contribute little to churn prediction.

**Methodological Robustness**

* Layers of analysis:
  + **Visualization** → sns.histplot()
  + **Parametric testing** → stats.ttest\_ind()
  + **Non-parametric validation** → bs\_choice() bootstrap resampling
* Ensures reliable, multi-angle conclusions.

**Data Segmentation Phase**

The analysis begins by dividing the customer dataset into two distinct groups using the 'Exited' column as the primary segmentation criterion. The first group (df\_0) comprises customers who have maintained their relationship with the bank, while the second group (df\_1) consists of customers who have churned. This fundamental separation enables comparative analysis between retained and lost customers across multiple demographic and financial dimensions.

**Hypothesis 1: Age Analysis**

The age hypothesis investigation employs a versatile statistical approach to determine if significant age differences exist between retained and churned customers. The analysis initiates with visual exploration using overlapping histogram distributions with kernel density estimates, providing immediate visual insights into age pattern differences. Descriptive statistics including mean and standard deviation calculations quantify central tendency and variability within each group.

The analytical rigor extends to inferential statistics through independent samples t-testing, assessing whether observed age differences reach statistical significance. To complement parametric testing, a bootstrapping methodology is implemented, which involves resampling the age data with replacement to create empirical sampling distributions. This technique calculates shifted distributions aligned to the overall mean age, enabling computation of bootstrap-derived p-values that validate the t-test results through robust resampling techniques.

**Hypothesis 2: Credit Score Evaluation**

Credit score analysis follows a similar methodological framework, beginning with comparative distribution visualization. The investigation tests whether creditworthiness, as measured by credit scores, significantly influences customer retention. Statistical testing focuses on determining if systematic credit score differences exist between the two customer segments, employing hypothesis testing to ascertain whether credit score serves as a meaningful churn predictor.

**Hypothesis 3: Balance Assessment**

The balance analysis incorporates additional analytical sophistication by addressing the common banking scenario of zero-balance accounts. The investigation proceeds through multiple phases: initial comprehensive balance comparison including all accounts, followed by focused analysis excluding zero-balance accounts. This stratified approach distinguishes between overall balance patterns and active-account balance behaviors, providing nuanced insights into how account funding levels correlate with churn probability.

**Hypothesis 4: Salary Investigation**

Estimated salary analysis employs the most comprehensive methodological approach, integrating both traditional parametric testing and advanced resampling techniques. The investigation visualizes salary distribution patterns across customer segments, followed by standard t-testing. The analytical depth increases with bootstrap implementation specifically tailored for salary data, creating empirical distributions of mean differences through extensive resampling to derive robust statistical significance measures.

**Statistical Interpretation Framework**

Throughout all hypotheses, a consistent statistical significance threshold (typically α=0.05) determines whether null hypotheses are rejected. Each conclusion explicitly states whether sufficient evidence exists to claim significant differences between retained and churned customers for each examined feature. The interpretation considers both practical significance (magnitude of differences) and statistical significance (probability observations occurred by chance).

**Comprehensive Conclusion Synthesis**

The final analysis phase synthesizes findings across all examined features through systematic comparison of statistical significance levels. The framework identifies the most potent churn predictor by ranking features according to their p-values, with the most significant feature demonstrating the strongest statistical evidence for differentiation between customer segments. Additional contextual insights include directional analysis (whether churned customers show higher or lower values) and magnitude assessment of observed differences.

**Methodological Robustness**

The analytical approach ensures robustness through multiple verification layers: visual distribution analysis, parametric statistical testing, and bootstrap validation. This triangulation of methods provides confidence in conclusions by addressing various statistical assumptions and potential limitations. The implementation systematically progresses from data preparation through hypothesis testing to actionable insights, delivering a comprehensive understanding of factors driving customer churn in the banking context.

The entire analytical process transforms raw customer data into evidence-based insights, enabling data-driven decision-making for customer retention strategies and identifying the most critical variables for predictive churn modeling.