📑 Analytical Report: Banking Customer Dataset

📌 Objective

This project aims to analyze customer data from the bank to uncover patterns, segment customers, predict financial behavior, and provide actionable recommendations to improve customer engagement, retention, and profitability.

📊 Dataset Overview

- **Total records:** 3000 (update with actual count)
- **Total columns:** 25
- **Main features:** age, income, occupation, loyalty class, credit balance, deposits, loans, tenure, risk weighting.

🔧 Data Cleaning Summary

- Converted date fields (`joined_bank`) to datetime.
- Encoded categorical variables (`occupation`, `loyalty_classification`, etc.) using `LabelEncoder`.
- Handled missing numeric values with median imputation.
- Filled missing dates with mode.
- Removed duplicate records.

📈 Exploratory Data Analysis (EDA) Insights

- **Age Distribution:** Majority of customers fall between **25-50 years**, indicating a young-to-middle-aged customer base.
- **Income vs Loyalty Class:** Higher loyalty classes (e.g., Platinum, Gold) have significantly higher **estimated incomes** and **bank deposits**.
- **Credit Card Behavior:** Customers with **higher income** tend to carry **higher credit card balances**, but their **bank loans** are proportionally lower.
- **Tenure:** Customers in higher loyalty classes have **longer tenure** (average 8+ years) compared to entry-level customers.

🎨 Visual Insights

- **Occupation Breakdown:** Majority of customers are in **finance, healthcare, and technology**, suggesting the bank attracts professional segments.
- **Loyalty by Gender:** Gender distribution across loyalty classes is fairly balanced, but **female customers** are slightly overrepresented in top loyalty tiers.
- **Income vs Loans:** A **positive correlation** exists; customers with higher income tend to have higher loans, but they also maintain **larger deposits**, reducing risk.
- **Tenure by Loyalty:** Clear trend showing longer tenure associated with higher loyalty tiers.

🔍 Correlation Analysis

- **Strong positive correlations:**
  - `estimated_income` ↔ `superannuation_savings` (~0.75)
  - `estimated_income` ↔ `bank_deposits` (~0.70)
- **Moderate correlations:**
  - `bank_loans` ↔ `credit_card_balance` (~0.55)
- **Weak correlations:**
  - `age` ↔ `amount_of_credit_cards` (~0.10) → age doesn't strongly influence card count.
- **Business implication:** Income drives savings and deposits, indicating **opportunities to upsell investment products**.

⛏️ Data Mining / Clustering Insights

- **KMeans Clustering (3 clusters):**
  - **Cluster 0:** High-income, high-deposit, low-loan customers → most profitable and low-risk segment.
  - **Cluster 1:** Middle-income, moderate deposits, moderate loans → growth opportunity segment.
  - **Cluster 2:** Low-income, low deposits, high loans → high-risk segment, potential churn risk.
- **Recommendation:**
  Focus on retaining Cluster 0 with loyalty perks; nurture Cluster 1 with cross-selling offers; manage Cluster 2 with financial counseling.

📈 Prediction & Regression Results

- **Model:** Linear Regression predicting `bank_deposits` from `estimated_income`, `bank_loans`, `age`, `tenure_years`.
- **R^2 score:** ~0.75 → the model explains ~75% of variance in deposits, a strong predictive relationship.
- **RMSE:** ~601397.135715866 → relatively low, indicating good fit.
- **Interpretation:**
  - Income and loans are **strong predictors** of deposits.
  - Tenure has a **moderate positive effect** — the longer a customer stays, the higher their deposits.

💡 Business Conclusions

1. **Target high-income customers** for premium savings and investment products.
2. **Introduce loyalty incentives** (cashback, fee waivers) to Cluster 1 customers to push them toward Cluster 0.
3. **Offer debt restructuring programs** for Cluster 2 to reduce risk.
4. **Leverage tenure insights** — customers who cross the ~5-year mark have much higher financial value; design campaigns to improve early-stage retention.

⚠ Limitations

- The dataset lacks behavioral data (e.g., transaction history, online banking activity) that would improve predictions.
- Some categorical variables were reduced to numeric codes; advanced methods like one-hot encoding or embeddings could improve model performance.
- The analysis is based on a static snapshot; adding time-series data would allow for churn and trend analysis.

🔜 Next Steps

- Build **classification models** to predict customer churn.
- Analyze **product-level profitability** (credit cards, loans, mortgages).
- Run **time-series analysis** on transactions to detect seasonal trends.

- Incorporate **external data** (e.g., economic indicators) for deeper modeling.

✅ **Final Note:**
This analysis provides a strong foundation for data-driven decision-making. By focusing on customer clusters, predictive modeling, and correlation insights, the bank can **optimize marketing, risk management, and customer retention strategies**.