

# Data Glacier – Virtual Internship

## Week 10: Exploratory Data Analysis (EDA) Report

---

**Team Name:** DataVision

**Team Members:**

- **Raj Pawar**  
Email: [rajpawar32646@gmail.com](mailto:rajpawar32646@gmail.com)  
Country: UK  
College: University of Liverpool  
Specialization: Data Analyst
  - **Naga Pavithra Jajala**  
Email: [pavithrajajala8naga@gmail.com](mailto:pavithrajajala8naga@gmail.com)  
Country: UK  
College: Birmingham City University  
Specialization: Data Analyst
- 

### Problem Description

XYZ Credit Union is experiencing low product adoption among its customer base. The organization seeks to use historical customer data to identify behavioural and demographic patterns that can inform better marketing and customer engagement strategies. Our task is to perform a thorough exploratory data analysis (EDA) on the cleaned dataset to generate business insights and inform future predictive modelling.

---

### Dataset Used

- **Source:** Cleaned version of the original train.csv file.
  - **Final Cleaned File:** Final\_cleaned.csv
  - **Size:** Approximately 3GB
  - **Format:** Structured CSV with 48 columns and over 13 million rows
- 

### EDA Summary

#### 1. Customer Profile Analysis

- **Gender Distribution:**
  - Male: ~6.2 million
  - Female: ~7.4 million
  - Observation: **Slightly higher representation of female customers.**

- **Age Distribution:**
  - Mean Age: ~40.2 years
  - Age Range: 18 to 100
  - Observation: **Broad range with concentration around the 30–60 year range. Outliers and incorrect entries were handled during cleansing.**
- **Customer Segment Distribution:**
  - ‘Particulares’ dominate (~60%), followed by ‘Universitario’ and ‘Top’ customers.
  - Segment 'Universitario' could represent younger clients or students.
- **Country of Residence:**
  - **Overwhelmingly Spanish** (ES) customers (~99%)
  - Other countries (FR, AR, GB, etc.) are negligible in comparison.
- **Customer Seniority:**
  - Ranges from 0 to over 250 months
  - Mean: ~79.3 months
  - **Outliers and erroneous values handled** (e.g., negative seniority)

## 2. Product Ownership Analysis

- **Top Products Owned:**
  - Current Account (~8.9M), Particular Account (~1.7M), Direct Debit, Payroll Account
  - Observation: **Bread-and-butter products (accounts and direct debits) dominate the landscape**
- **Product Count per Customer:**
  - Mean: 1.46
  - Median: 1
  - **52.5% of customers have only one product**
  - Implication: High potential for cross-selling initiatives

## 3. Time-Based Analysis

- **Customer Count by Month:**
  - **Monthly customer base is stable**, ranging from ~600k to 930k
- **Monthly Product Holding Trends:**
  - Product ownership was tracked for each product across months
  - Some products (e.g., current\_account, direct\_debit) show stable or increasing trends

- Others (e.g., long\_term\_deposit, short\_term\_deposit) show decline, possibly due to policy or interest rate changes

#### 4. Churn and Loyalty Insights

- **High Churn Products:**
  - Direct Debit: ~136K churned
  - Current Account, Credit Card, Salary, Pension also show significant churn
- **High Adoption Products:**
  - Direct Debit (~153K added), Pension (~84K), Salary (~73K)
- **Insight:**
  - These products are transactional and flexible, possibly tied to monthly salaries or benefit payments
  - Targeted retention strategies can focus on these categories
- **Rates Computed:**
  - Churn and Addition Rates calculated relative to total users per product
  - Most products show churn rates below 10%, with notable exceptions like short\_term\_deposit (26%) and credit\_card (11.6%)

#### 5. Correlation & Feature Interaction

- **Age & Product Ownership:**
  - Older customers tend to hold more products; age positively correlates with mortgage, pension, and funds
- **Segment vs Product Ownership:**
  - 'Top' segment customers tend to own more premium products like securities, mortgages
  - 'Universitario' segment mostly owns basic accounts like e\_account or payroll
- **Seniority vs Products:**
  - Customers with longer seniority show diversified portfolios
- **Gender Trends:**
  - Minor variances; however, female customers show slightly higher adoption in savings-related products

---

### Final Recommendations

1. **Cross-Selling Opportunities:**

- 52% of customers own only one product. Focused campaigns can be deployed targeting this group

## **2. Segment-Specific Offers:**

- ‘Top’ segment can be pitched high-value financial products
- ‘Universitario’ segment can be introduced to savings or investment options post-graduation

## **3. Retention Strategy:**

- High churn in direct\_debit and credit\_card segments need closer evaluation. Consider engagement strategies or bundling products

## **4. Age-Based Targeting:**

- Young adults (20–30) underrepresented in long-term investment products, tailor content/ads for this cohort

## **5. Product Lifecycle Monitoring:**

- Trends in deposits (short/medium/long-term) suggest customer preferences are shifting, worth exploring external economic factors

## **6. Model Readiness:**

- Data is now structured, clean, and temporally sorted, ready for modelling tasks such as churn prediction or customer segmentation

---

## **Conclusion**

This EDA process provided us with rich customer insights and product usage patterns across time. It set the stage for Week 11’s hypothesis testing and deeper analytics. The collaborative and structured EDA ensured completeness, and the findings are directly tied to potential business actions.

---