

## **1. Project Title:** Customer Segmentation for Christmas Campaign

Name	Email	Country	College	Specialization
------	-------	---------	---------	----------------

Diyora Radhika Radhikadiyora2023@gmail.com Germany IU Internationale Data Science

**Submission Type:** PDF + Jupyter Notebook

---

## **2. Problem Description**

XYZ Bank wants to roll out **personalized Christmas offers** to its customers.

- Business Objective: Avoid sending the same offer to all customers, target campaigns efficiently.
- Requirement: Maximum 5 customer segments for campaign efficiency.

Challenges:

- Manual segmentation is inefficient.
  - Need to uncover hidden patterns in customer behavior.
- 

## **3. Dataset Overview**

**Source:** cust\_seg.csv.zip

**Number of Records:** [Insert Count]

**Key Features:**

Feature	Description
age	Customer age
renta	Gross income
antiguedad	Seniority (months)
ind_empleado	Employee type
ind_*_ult1	Product ownership (Savings, Current Accounts, Loans, etc.)
sexo, pais_residencia	Demographics
fecha_dato, ncodpers	IDs & Dates

---

## **4. Data Cleaning & Transformation**

### **4.1 Missing Values Handling**

### **Techniques Used by Team Members:**

- **Member 1:** Mean imputation for numeric columns (age, renta)
- **Member 2:** Median imputation for numeric columns segmented by sexo
- **Member 3:** Model-based imputation for antiguedad using regression

### **Before Cleaning:**

- Null counts: age: 1200, renta: 4500, antiguedad: 0

### **After Cleaning:**

- Null counts: All 0
- 

### **4.2 Outlier Handling**

- **Technique 1:** IQR-based capping for renta and age
- **Technique 2:** Z-score filtering for extreme total\_products values

**Before:** Max renta: 5,000,000, Min age: -2

**After:** Capped at 1st and 99th percentile

---

### **4.3 Categorical Variable Transformation**

- One-hot encoding applied for: sexo, ind\_empleado, pais\_residencia, tiprel\_1mes
  - Drop first category to avoid dummy variable trap
- 

### **4.4 Feature Engineering**

- **Total Products:** Sum of all product ownership columns (ind\_\*\_ult1)
  - **Cluster Label:** Assigned after K-Means (5 clusters)
- 

### **4.5 Summary Table of Cleaned Data**

Feature	Mean	Median	Min	Max	Nulls
age	45.6	44	18	90	0
renta	128,000	120,000	4,000	500,000	0
total_products	2.1	1	0	15	0

*(Add screenshots from your notebook here)*

---

### **5. Cluster Analysis (Optional)**

## **Clusters: 5**

- Cluster 0: Older, high-income, multi-product → premium offers
- Cluster 1: Young, low-product, medium-income → entry-level offers
- Cluster 2: Tiny outlier cluster → ignore for campaign
- Cluster 3 & 4: Mid-age, varying products → standard or loyalty offers

## **Cluster Summary Table:**

Cluster	Customers	Avg Age	Avg Income	Avg Total Products
---------	-----------	---------	------------	--------------------

0	88,629	47	150k	5.88
1	10,784	43	106k	1.0
2	2	57	106k	2.0
3	448,529	39.6	123k	1.02
4	452,056	46	142k	2.02

(Insert PCA plot screenshot here)

---

## **6. Recommendations**

- **Cluster-based Targeting:** Different offers based on engagement and income
  - **High-value clusters (0 & 4):** Premium products and loyalty rewards
  - **Low-product clusters (1 & 3):** Simple or promotional offers
  - **Ignore outliers (2)** for large-scale campaigns
- 

## **7. GitHub Collaboration**

- Code contributions from all team members present in the repo
  - Peer reviews documented as comments
  - Merge workflow applied for final notebook
- 

## **8. References / Links**

- GitHub Repository: <https://github.com/RadhikaRanchhodhaiDiyora/VC/tree/week9>
- Dataset: cust\_seg.csv.zip