KYC & Client Onboarding Data Quality Audit

Amal S

Agenda

1 Introduction

2 Primary Goals

3 KYC Data Risk Overview

4 Strategic Roadmap for KYC Data Integrity

5 Summary

Introduction

Know Your Customer(KYC) compliance is a core requirement in banking and financial services. Ensuring accurate and complete client data helps institutions prevent fraud, meet regulatory obligations, and improve operational efficiency.

This project simulates a real-world scenario of client onboarding, where incomplete or inconsistent data can trigger compliance risks.

Using Python, I created synthetic client records and implemented logic to detect missing values, data entry errors, and sanctioned country flags.

The project also includes fuzzy matching techniques to correct misspelled country names — a common issue in data entry.

Visual summaries of flagged issues provide actionable insights for improving KYC data quality and onboarding workflows.

Primary goals

Ensure completeness, accuracy, and compliance of client onboarding data using Python-based validation and flagging logic.

Flagged KYC Issues by Type

KYC Data Risk Overview

Issue Type	Number of Records	Risk Level	Suggested Action
Incomplete KYC	241	🔴 High	Improve follow-up and automate document reminders
Missing Full Name	13	🟡 Moderate	Make full name mandatory at entry point
Missing Full Name, Incomplete KYC	6	🔴 High	Flag for manual review before account activation
Incomplete KYC, High-Risk Country	5	🔴 High	Escalate for compliance officer review
High-Risk Country	3	🔴 High	Block or require enhanced due diligence (EDD)
Missing Full Name, High-Risk Country	2	🟠 Medium	Cross-check with source documents
Missing Full Name, Incomplete KYC, High-Risk Country	2	🔴 Critical	Multi-risk: escalate immediately
Incomplete KYC, Missing Aadhaar Number	1	🔴 High	Indian client: Aadhaar should be required
OK (No issues)	227	🟢 Low	No action needed

> Without data, you're just another person with an opinion
>
> W. Edwards Deming

Tools & Technologies Used

Tool/Tech	Purpose
Python (pandas)	Data cleaning, simulation
Faker	Generate synthetic client data
RapidFuzz	Correct misspelled country names
Matplotlib/seaborn	Visualizations
PowerPoint	Final reporting & presentation

Key Takeaways / Lessons Learned

- Real-world KYC processes can have **hidden data quality issues**
- Simple Python tools can help detect **compliance risks at scale**
- Fuzzy matching and rule logic are powerful for **automated anomaly detection**
- This project strengthened my confidence in **data-driven decision-making**

Strategic Roadmap for KYC Data Integrity

Planning
Define Key Fields & Flags

Validation
Build Validation & Simulation

Cleaning
Data Correction & Enrichment

Analysis
Detect Patterns & Risks

Reporting
Deliver Insightful Outcomes

Strategic Priorities in KYC Data Quality

KYC Data Integrity & Automation

- Develop logic to detect incomplete or non-compliant KYC entries
- Automate anomaly detection using synthetic datasets
- Clean and standardize messy inputs (e.g., country misspellings)
- Visualize risk concentrations across account types

Scalable Risk Flagging Framework

- Extend rule-based detection to real-time onboarding systems
- Integrate fuzzy matching with live data validation pipelines
- Enable proactive compliance alerts to avoid future fines
- Build dashboards (Power BI or internal tools) for compliance teams

How we got there

Data Generation & Preparation

"Laying the foundation with synthetic data"

Generated realistic client onboarding data using Python & Faker

Simulated anomalies (e.g. missing fields, high-risk countries)

Structured dataset to reflect real-world KYC scenarios

Validation & Flagging Logic

"Applying rules to find risk and gaps"

Created rule-based logic to flag incomplete or suspicious data

Used fuzzy matching (RapidFuzz) to fix country spelling errors

Identified patterns in risk by account type (Retail, HNI, Corporate)

Insight Delivery & Reporting

"Turning data into actionable insights"

Visualized issues using bar charts & summary tables

Presented findings in a professional PowerPoint report

Suggested real-world actions (e.g. Aadhaar enforcement, alerts)

Summary

In this project, I simulated and audited KYC onboarding data to identify compliance risks and data quality issues.
Using Python and rule-based logic, I flagged incomplete records, corrected country misspellings with fuzzy matching, and visualized risk patterns across account types.
This hands-on approach helped me understand the real-world importance of data accuracy in financial onboarding systems — and how even small errors can have regulatory consequences.
I believe clean, validated data is the foundation of trust in any client-facing organization.

Thank you

Amal S

amal17ek@gmail.com

linkedin.com/in/amal-s-9a5b86310