**PAN Card Validation – Project Summary Report**

**Project Objective**

Process and validate 10,000 PAN numbers from a raw CSV file in PostgreSQL, ensuring compliance with official Indian PAN format rules, removing invalid and duplicate entries, and delivering a clean, compliance-ready dataset with summary metrics.

**Key Outcomes**

1. 10,000 PANs processed end-to-end in PostgreSQL
2. 3,186 valid PANs retained
3. 5,839 invalid PANs flagged
4. 975 incomplete/duplicate entries removed prior to validation
5. 100% automated validation (regex + UDFs + CTE pipeline)

**Problem & Business Impact**

Before → After

* Invalid, duplicate, and badly formatted PANs in raw data

→ Only compliant, standardized PANs flow downstream

* Manual verification was slow and error-prone

→ Fully automated SQL pipeline; verification time reduced to near-zero

* Compliance risk from improper PAN handling

→ Enforced format/sequence rules; reduced compliance and audit risk

**Approach & Tools**

1. **Validation Logic**

* Exactly 10 characters
* Pattern: AAAAA9999A
* No two adjacent characters (letters or digits) are identical
* First 5 letters must not form an alphabetic sequence (e.g., ABCDE)

- Next 4 digits must not form a numeric sequence (e.g., 1234)

2. **Data Loading & Cleaning**

Imported CSV into PostgreSQL

Removed null/empty rows and exact duplicates

Standardized formatting (TRIM, UPPER)

3. **Results Generation**

Final table with Valid/Invalid status

Roll-up metrics for quick reporting

**Value to Organization**

1. Compliance Assurance: All retained PANs conform to official structure and logic rules
2. Efficiency: Eliminates manual checks; saves analyst hours at scale
3. Risk Mitigation: Early detection of invalid/suspicious PANs avoids downstream issues
4. Reusability: SQL script is parameterized and repeatable for future datasets

**Tools Used**

PostgreSQL (SQL, UDFs, regex, CTEs)

CSV (source data)

**PAN Card Validation – Project Summary Report**

**Executive Summary**

Validated and standardized 10,000 PAN numbers from a raw CSV dataset using PostgreSQL. Implemented a fully automated SQL pipeline incorporating regex, UDFs, and CTE-based logic to ensure compliance with Indian PAN format rules. Achieved a 68% reduction in invalid/duplicate entries and delivered a compliance-ready dataset for downstream business use.

**Key Results**

**Metric Value**

Total PANs Processed 10,000

Valid PANs Retained 3,186

Invalid PANs Flagged 5,839

Incomplete/Duplicates Removed 975

Validation Automation Rate 100%

**Business Impact**

**Before After**

Multiple invalid, duplicate, and ill-formatted PANs. Only compliant, standardized

PANs flow downstream

Manual verification was slow and error-prone Fully automated SQL validation with near-zero verification time

Compliance risk from improper PAN handling Enforced format rules, reducing

Compliance & audit risks

**Approach & Tools**

**Validation Logic**

1. Exactly 10 characters
2. Format: AAAAA9999A
3. No two adjacent identical characters/digits
4. First 5 letters not in sequential order (e.g., ABCDE)
5. Next 4 digits not in sequential order (e.g., 1234)

**Data Loading & Cleaning**

1. Data Loading & Cleaning
2. Imported CSV into PostgreSQL
3. Removed null/empty rows & exact duplicates
4. Standardized formatting (TRIM, UPPER)

**Results Generation**

1. Final table with Valid / Invalid status
2. Summary metrics for quick reporting

**Value to the Organization**

1. Compliance Assurance – All retained PANs conform to official format & logical rules
2. Efficiency – Eliminates manual checks; saves analyst hours at scale
3. Risk Mitigation – Early detection of invalid/suspicious PANs avoids downstream issues
4. Reusability – Parameterized SQL script repeatable for future datasets

**Tools Used**

1. PostgreSQL
2. CSV (Source Data)