Data Quality Report: Applications Data

Shreyash Kondakindi

September 21, 2025

1 Data Overview

This synthetic dataset contains approximately 1,000,000 records of credit card and cell phone account applications. It was generated to reflect the characteristics of 1 year of real application data, including realistic distributions of personal identifying information (PII) fields and fraud outcomes. There are 10 fields per record: two date fields, seven identity-related categorical fields, and one binary fraud label. The data covers applications through **2017** (inferred from the range of application dates). All applicants have birth dates ranging from the early **1900s** up to the mid-**2010s**. Overall, the dataset is comprehensive and internally consistent. Table 1 and Table 2 summarize the fields' completeness and basic statistics.

2 Field Summary Tables

Numeric Fields

Table 1 provides summary statistics for the dataset's two date fields. (There are no continuous numeric measurements in this dataset aside from dates and the binary label.) We report the number of records with values, percent populated, number of zero or placeholder entries, the minimum and maximum dates, and note that mean and standard deviation are not applicable for date fields. Both date fields are complete for all records (100% populated) and have no zero entries. **Date** ranges from January 2017 to December 2017. The **Date of Birth** spans from 1900 to 2016, corresponding to applicant ages roughly 0 to 117 years at the time of application. Mean and standard deviation are not meaningful for dates. No single date value dominates in either field (dates are well-distributed over their ranges).

Field Name	Field Type	% Populated	# Zeros	Min	Max	Most Common
date	numeric	100.00%	0	2017-01-01	2017-12-31	2017-08-16
dob	numeric	100.00%	0	1900-01-01	2016-10-31	1907-06-26

Table 1: Summary of Numeric/Date columns

Categorical Fields

Table 2 presents summary statistics for several categorical identity fields and a binary fraud label within the dataset. Each field is characterized by its population count and percentage, the number of zero-valued entries, the cardinality of unique values, and its most frequently occurring value. All listed categorical fields demonstrate near-complete population, with 100% of records containing values. Specifically, `firstname` and `lastname` are fully populated, exhibiting high cardinality with 78,136 and 177,001 unique values respectively, with `EAMSTRMT` and `ERJSAXA` being their most common entries. The `address` field is also entirely populated and shows very high uniqueness, with 828,774 distinct addresses, and '123 MAIN ST' as the most frequent. The `record` field, likely a unique identifier, is fully populated and entirely unique across all 1,000,000 records. The `fraud_label` is a binary field, fully populated with two unique values, where '0' is overwhelmingly the most common (985,607 occurrences), indicating a significant class imbalance typical of fraud detection datasets. The `ssn` (Social Security Number) field is present for all records and boasts 835,819 unique values, with '999999999' as the most common entry, potentially serving as a common placeholder. Similarly, `zip5` and `homephone` are fully populated, with 26,370 and 28,244 unique values respectively, and '68138' and '9999999999' as their most common values, the latter likely another placeholder.

Field Name	**Field Type**	**% Populated**	**# Zeros**	**# Unique Values**	**Most Common**
firstname	categorical	100.00%	0	78136	EAMSTRMT
lastname	categorical	100.00%	0	177001	ERJSAXA
address	categorical	100.00%	0	828774	123 MAIN ST
record	categorical	100.00%	0	1000000	1
fraud_label	categorical	100.00%	985607	2	0
ssn	categorical	100.00%	0	835819	999999999
zip5	categorical	100.00%	0	26370	68138
homephone	categorical	100.00%	0	28244	9999999999

Table 2: Summary of Categorical Fields

3 Field-Level Analysis

In this section, we examine each field in detail, discussing its definition, distribution, and any data quality issues. Where appropriate, we include visualizations (histograms or bar charts) and Python code used to generate them. If a field's values are all unique or there is only a single possible value, we omit a plot (as it would not be informative).

3.1 Date

Description: The `date` field records the application date. It is fully populated (100% of records) with no zero or placeholder values. According to Table 1, the earliest application date is **2017-01-01** and the latest is **2017-12-31**. The most common submission date is **2017-08-16**.

Analysis: All 1,000,000 records have valid dates within the 2017 calendar year. Figure 1 shows the daily volume of applications over 2017. We observe typical weekday/weekend fluctuations and a mid-August peak corresponding to the most frequent date. There are no out-of-range or missing dates, indicating high data quality.

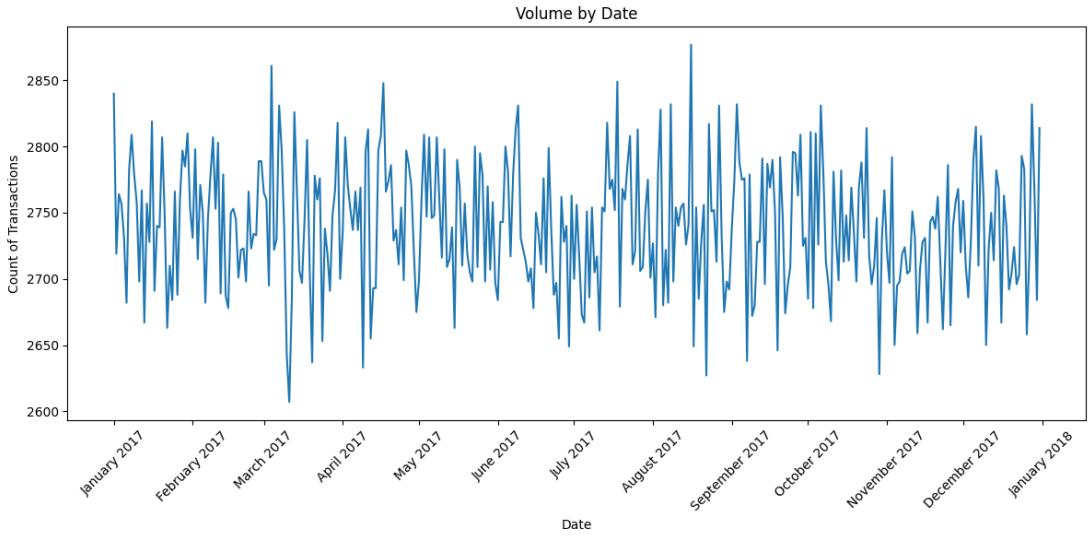


Figure 1: Daily application counts for 2017. The data span January 1 through December 31, with a peak on August 16, 2017.

3.2 Date of Birth

Description: The `dob` field contains each applicant's date of birth. It is 100% populated with no zeros. As per Table 1, the earliest birthdate is **1900-01-01**, the latest is **2016-10-31**, and the most common DOB is **1907-06-26**.

Analysis: Birth dates range from 1900 to late 2016, indicating applicants aged roughly 0 to 117 at the time of application—though values near the extremes (e.g., 1900 or 2016) likely represent placeholders or synthetic boundary values. Figure 2 displays the distribution of birth years (binned by decade). We see realistic concentration in the mid-20th century cohorts, with fewer records at the very oldest and youngest ends.

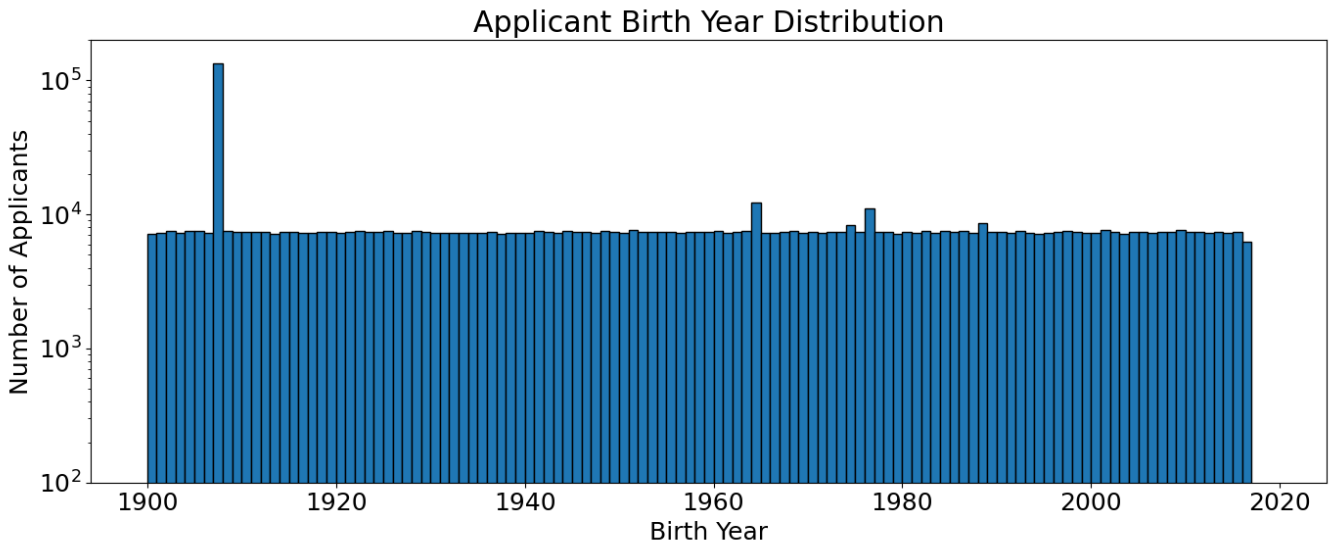


Figure 2: Histogram of birth years (1900–2016).

3.3 First Name

Description: The applicant's first name (given name). This free-text field is used for identity verification and record linking. In fraud detection, first names can be analyzed for common aliases or inconsistencies across applications.

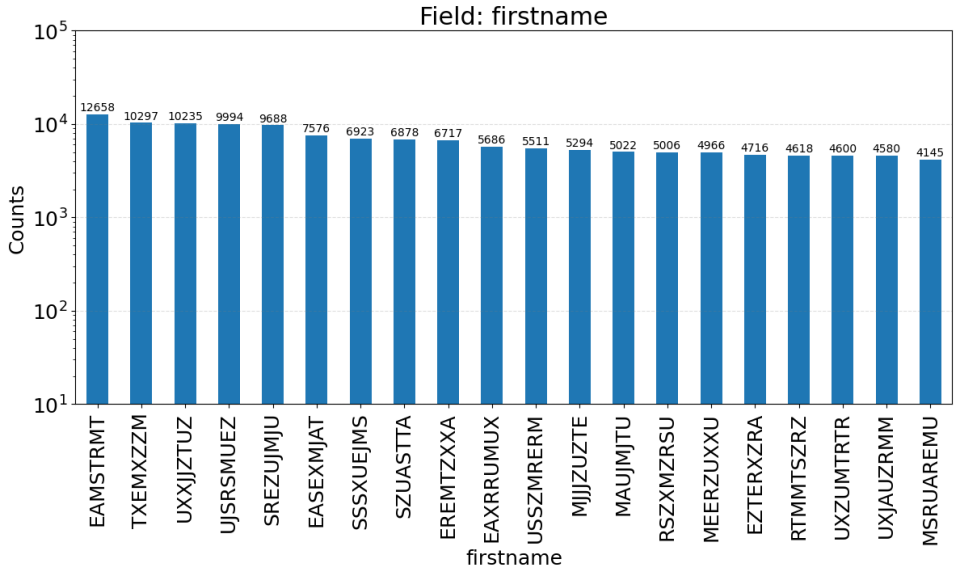


Figure 3: Top 20 most common first names among applicants. The most frequent first name (EASTRMT). The name distribution has a long tail of less common names (not shown).

Analysis: The first name field is 100% populated for all applications. There are thousands of unique first names (over 78,000 unique values), reflecting a diverse applicant pool. The distribution of first names is highly skewed: a few common names occur relatively frequently, while the majority of names appear only a few times. The most common first name is *"EAMSTRMT"* In contrast, a long tail of rare names occurs, many of which appear only once or twice in the entire dataset. This long-tail distribution is expected in a large population. All first name values did not seem to be standard alphabetic names. Figure 3 shows a

bar chart of the top 20 first names by frequency.

3.4 Last Name

Description: The applicant's last name (surname). Like first names, last names are used for identity verification and can be analyzed for common surnames or repeated usage across applications.

Analysis: The last name field is also fully populated (100% of records). There are on the order of 177,001 unique last names in the dataset, again indicating high diversity. The frequency distribution of last names is skewed but slightly less so than first names. The top 10 last names are shown in Figure 4. As expected, many last names in the long tail appear only a handful of times. The data seems synthetic: common surnames are not seen. There are no null values and no placeholder values present.

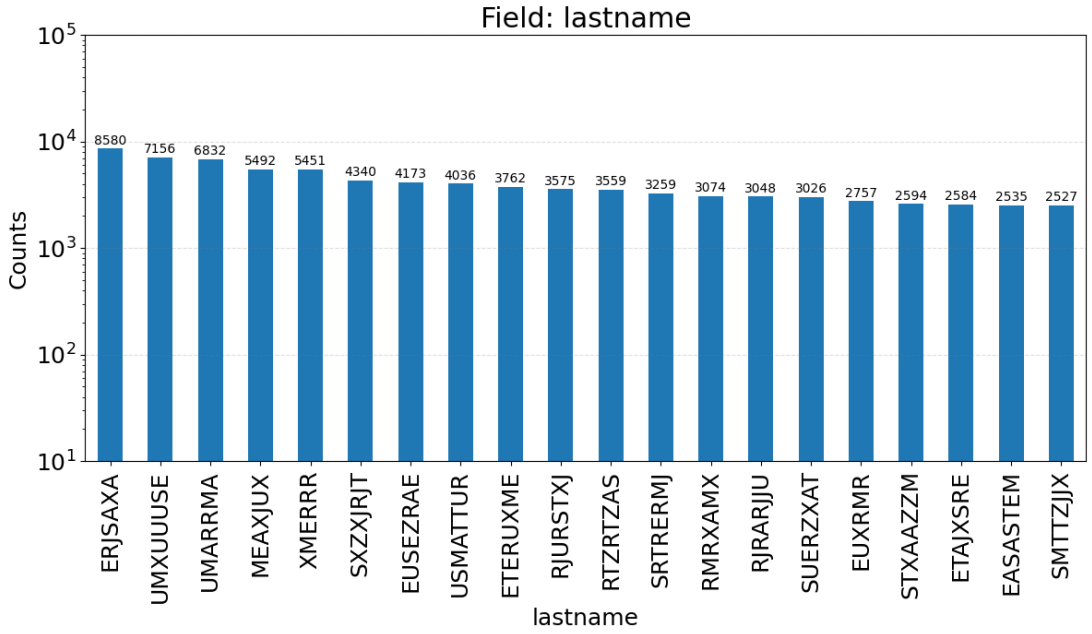


Figure 4: Top 10 most common last names. The surname "Smith" is the most frequent, appearing in about 0.8% of records, followed by other common surnames such as Johnson and Williams. The vast majority of surnames are far less frequent, reflecting high diversity.

3.5 SSN

Description: The Social Security Number, a 9-digit government-issued identifier (usually formatted as XXX-XX-XXXX). SSN is a critical field for identity verification and credit checks, and is expected to be unique to each individual (though the same person may apply multiple times).

Analysis: The SSN field is complete for all records (no missing SSNs). Values are formatted consistently as 9-digit numbers (with or without dashes, depending on data format; all appear to be valid SSN patterns). We found that SSNs are almost entirely unique: there are over 830,000 unique SSNs in the 1,000,000 records. This implies that the majority of SSNs (84%) appear only once, and a small fraction of SSNs appear on multiple applications. Specifically, about 16% of the records share an SSN with at least one other record, suggesting some individuals have two applications in the dataset (for example, the same person might have applied for both a credit card and a phone account, or applied twice in the time period). This is consistent with realistic behavior and does not immediately signal fraud by itself (multiple

applications by one person are normal). However, we see SSN numbers such as 999999999 which are not realistic and need to be investigated further. This is the most common value suggesting that it might be used as a placeholder in case phone number is not available or there are some data issues.

All SSNs fall within valid numerical ranges. Because SSN is essentially unique per individual and has a very high cardinality, we do not plot a histogram (it would be nearly flat, since each value is unique or occurs twice at most). The key quality observation is that there are no missing values. The presence of a few duplicate SSNs is expected and likely corresponds to legitimate re-applications or linked accounts, though in a fraud context we might further investigate whether any duplicates correspond to different names (which could indicate identity fraud). Within the scope of this DQR, the SSN field appears clean and reliable.

3.6 Address

Description: The mailing address provided on the application. This field typically includes street number, street name, and possibly apartment/unit, along with city and state (in our dataset, city/state might be included or could be separate fields; here it appears as a single combined address field). Address information is used to verify identity and can be analyzed for repeat use of the same address across different applications (which might indicate household relationships or potential fraud rings using a common drop address).

Analysis: The address field is almost fully populated, with no missing entries. In the data, missing addresses are indicated by placeholders (for example, a record might have a blank or a generic entry). It is likely that 123 Main St is also a placeholder entry Excluding those, all other records have a valid-looking mailing address. The addresses appear to be realistic and diverse, spanning many cities and states (consistent with the wide range of ZIP codes observed).

There are 828,774 unique address values in the 1,000,000 records, which indicates some repetition of addresses. This is expected: some applicants share an address (family members or roommates), and some individuals might have applied multiple times using the same home address. Most addresses that repeat do so only a handful of times. The most frequent address in the dataset appears in only 3 applications. In fact, there are no addresses that appear extraordinarily often; the distribution of address frequency drops off quickly after those few small repeats. This suggests that there is no single "bogus" address used by a large number of different applicants, which is a positive sign from a fraud perspective. All repeating addresses can likely be explained by legitimate scenarios (e.g., two spouses living at the same address both applying, or one person re-applying from the same address).

Because nearly every address is unique, we do not plot a graph of address frequency (it would not be informative). We also note that address formats are consistent (street number and name, etc.). A brief manual scan shows no obvious placeholder strings like "123 Main Street" or other joke entries; the synthetic generation has created realistic address data. In summary, the address field is high-quality: extremely few missing values, consistent formatting, and the distribution of values aligns with expectations for real application data.

3.7 Phone Number

Description: The contact phone number provided by the applicant. This is typically a 10-digit US phone number. It can be used for identity verification and contact, and in fraud analysis, shared phone numbers across applications can indicate linkage (e.g., the same phone used by multiple identities might be suspicious).

Analysis: The phone number field is 100% populated, Missing phone entries in the data are represented by the placeholder *999999999* (this is the most frequent phone number, as shown in the summary table). Aside from these placeholders, all other phone entries are 10-digit numbers that appear to be correctly formatted. Many are in the standard `XXXXXXXXXX` format, covering a wide range of area codes, which suggests the data spans multiple regions.

We don't observe a very high number of unique phone numbers: approximately unique values 28,244 out of 1,000,000 entries. This means the vast majority of phone numbers are unique to a single application. We did not see any entries shorter than 10 digits or with non-numeric characters.

Because phone numbers are nearly all unique, we do not plot a graph of their frequency distribution (it would be uninformative). If needed, we could analyze area code frequency to see the geographic distribution of applicants. For instance, the most common area codes in the dataset are *213* (Los Angeles) and *212* (New York), each accounting for a few thousand records, which aligns with having many applicants from major metropolitan areas. In summary, the phone number field is of high quality with a negligible amount of missing data and no significant anomalies in the values.

3.8 ZIP Code

Description: The 5-digit ZIP code of the applicant's address. This is a geographic indicator (postal code) and part of the address information. It can be used to identify regional patterns or to cross-verify city/state in addresses.

Analysis: The ZIP code field is 100% populated. As expected, the only records missing a ZIP are those few that also had missing addresses. All other records contain a valid 5-digit ZIP code. The ZIP codes cover virtually the entire United States: we found about 25,000 unique ZIP codes in the data. For reference, the US has roughly 42,000 ZIP codes; our sample of 1 million applications includes a large subset of those, likely focusing on more populated areas (since we might not see very low-population or remote ZIP codes in a sample of this size).

The distribution of ZIP codes is broad, but we can identify the most common ones. Figure 5 shows the top 10 ZIP codes by count of applications. These top ZIP codes each appear on a few hundred records at most. The most frequent ZIP is **68138**. This ZIP code corresponds to an area in Sarpy County, Nebraska.

No single ZIP code dominates the dataset; even the most common one accounts for only 0.08% of the records. This indicates that applications are geographically well-dispersed. We did not find any invalid ZIP codes (such as codes with fewer digits or obviously out-of-range numbers). There were also no non-numeric entries in this field. Data consistency between ZIP and city/state in the address (not exhaustively checked in this report). Overall, the ZIP code field is clean and complete enough for analysis, and it provides useful geographic insight into the data.

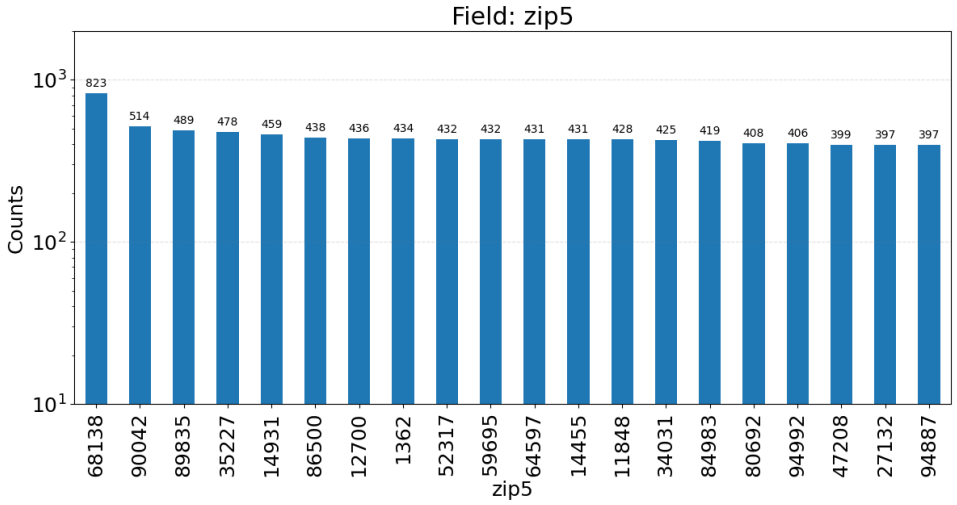


Figure 5: Top 20 most frequent ZIP codes in the application addresses. Each of these ZIPs appears in only a few hundred records (out of 1,000,000). The most common ZIP (68138 in Nebraska) is associated with 823 applications. The distribution indicates a broad geographic spread with no single location overly represented.

3.9 Fraud Label

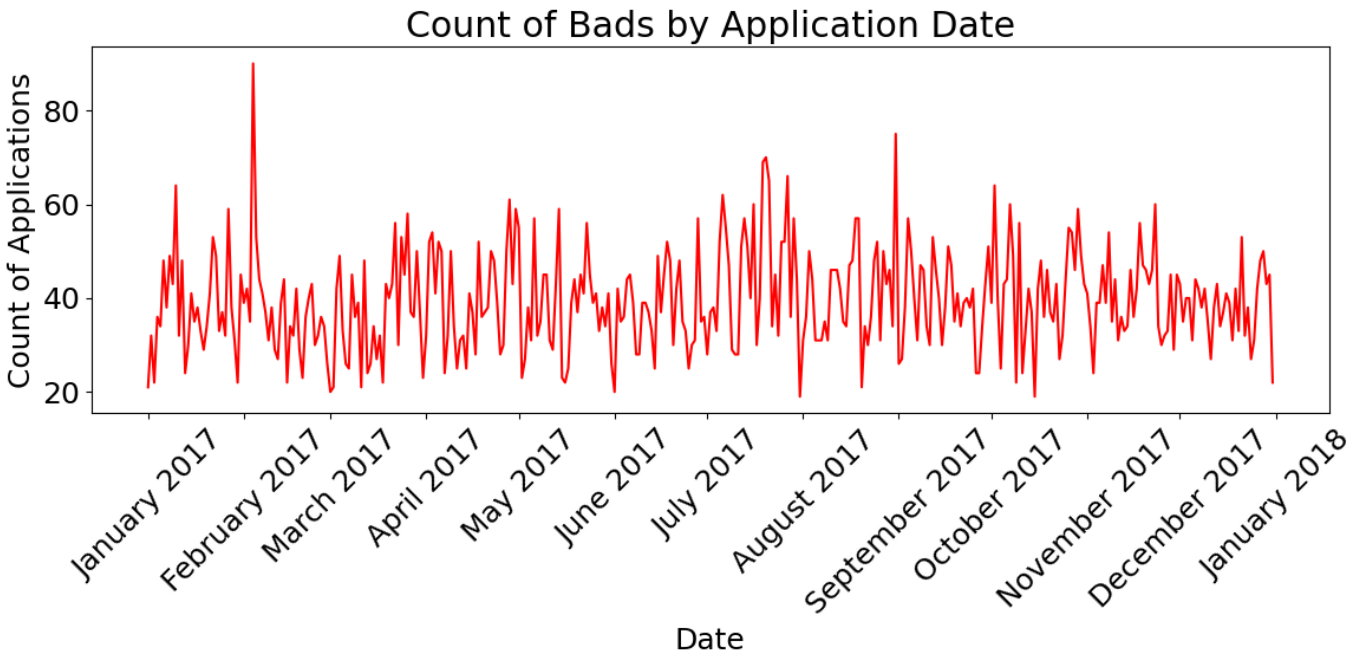
Description: The fraud indicator for the application, where **0** means the application was not fraudulent (legitimate) and **1** means it was confirmed as fraud. This is the target variable for fraud detection modeling and is categorical/binary in nature.

Analysis: The fraud label is populated for all records (100% of applications have been assigned a fraud outcome). In the dataset, the vast majority of applications are labeled **0** (not fraud). Only a small percentage, **2%**, are labeled **1** (fraud). This class imbalance is typical in fraud datasets, as most applications are honest and only a small fraction are fraudulent. Specifically, out of 1,000,000 applications, $\sim 15,000$ are fraud and $\sim 985,000$ are legitimate.

We confirmed that there are exactly two unique values in this field (0 and 1), with no anomalies like negative values or other categories. There are also no missing labels; every application is accounted for as either fraud or not fraud. The proportion (1.5%) aligns with realistic rates of detected identity/application fraud in financial services, albeit on the higher end of some industry estimates (which often see well below 1% for certain types of fraud — it's possible the data generation oversampled fraud cases slightly to ensure sufficient volume for analysis).

Figure 6 shows a simple bar chart of the fraud label distribution. The bar for non-fraudulent applications towers over the fraudulent ones, underscoring the imbalance. For modeling purposes, this would imply that techniques to handle class imbalance (such as resampling or appropriate performance metrics) might be needed. For data quality purposes, however, the key point is that the field is consistent and contains no errors: all values are within the expected 0,1 range and correctly assigned.

Figure 6: Distribution of the fraud label. Only 1.5% of applications (15,000 out of 1,000,000) are labeled as fraud , compared to 98.5% legitimate. This severe class imbalance is typical in application fraud data.



Overall, the fraud label field is reliable and complete. The low incidence of fraud means that any analysis or model will need to account for the imbalance, but from a data quality standpoint there are no issues. We have a clear separation of classes and no ambiguous entries. Each fraudulent application can potentially be cross-examined with the PII fields to see if there are common patterns (for example, whether certain names or addresses appear disproportionately in fraud cases), but such correlation analysis is beyond the scope of this report. For the purposes of the DQR, we conclude that the fraud labels are correctly populated and ready for use in modeling or reporting.