DATA QUALITY REPORT ON CREDIT CARD TRANSACTION DATA

**Description**

**Dataset Name:** Credit Card Transaction Data

**Data Source:** DSO 562 project (Actual credit card purchases from a US government organization)

**Time Period:** Jan. 2006 – Dec. 2006

**Number of Fields:** #10

**Number of Records:** 96,753

**Description:**

This is a dataset from actual credit card purchases from a US government organization. It includes 96,753 records from January 2006 to December 2006 with 10 fields. We aim to use supervised learning algorithms on this dataset to predict transaction fraud.

**Summary Table:**

**Numeric Fields:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| FIELD  NAME | PERCENT  POPULATED | MINIMUM  VALUE | MAXIMUM  VALUE | MEAN  VALUE | STANDARD  DEVIATION | PERCENT  ZERO |
| Date | 100 | 2006-01-01 | 2006-12-31 | NaN | NaN | 0 |
| Amount | 100 | 0.01 | 3,102,045.53 | 427.89 | 10,006.14 | 0 |

**Categorical Fields:**

|  |  |  |  |
| --- | --- | --- | --- |
| FIELD  NAME | PERCENT POPULATED | NUM OF UNIQUE VALUES | MOST COMMON VALUE |
| Recnum | 100% | 96,753 | NaN |
| Cardnum | 100% | 1,645 | 5142148452 |
| Merchnum | 96.51% | 13,091 | 930090121224 |
| Merch description | 100% | 13,126 | GSA-FSS-ADV |
| Merch state | 98.76% | 227 | TN |
| Merch zip | 95.19% | 4,567 | 38118.0 |
| Transtype | 100% | 4 | P |
| Fraud | 100% | 2 | 0 |

**Distribution of Fields:**

**Filed 1:**

**Name:** Recnum

**Description:** Unique identifier for each transaction. This field has 96,753 unique integers.

**Filed 2:**

**Name:** Cardnum

**Description:** Credit card number for each transaction. This field has no missing values and 1,645 unique values. The most common value in this field is 5142148452, which occurs 1,192 times in this field. The graph below shows top 15 credit card numbers in the dataset.

Chart, bar chart

Description automatically generated

**Filed 3:**

**Name:** Date

**Description:** Date for each transaction. This field has no missing values and ranges from 2006-01-01 to 2006-12-31 (daily basis), 365 unique values. The graph below shows the distribution of date and I divide it into 52 bins for each bin represents a week in year 2006. Plus, even if the number of fraud changes frequently, we assume that the distribution has no seasonality.

Chart, histogram

Description automatically generated

We also plot the transactions on the daily and monthly basis. For the first plot, we can find a regular fluctuation in the number, but the trends do not differ a lot between different fluctuations. In the meantime, if we look at the plot on the monthly basis, we can see an obvious drop in September. However, we still assume that our data has no seasonality.

Chart

Description automatically generated

Chart, line chart

Description automatically generated

We can also plot the number of transactions grouping by fraud or not. In the graph below, the number of transactions without fraud is labeled as green and the number of transactions with fraud is labeled as red. We can easily find that on the first three months, both fraud and non-fraud transactions remain relatively stable. However, from late April to December, on the weekly basis, fraud transactions are fluctuated a lot when normal transactions are relatively stable. More importantly, for most cases, the beginning of a month and the middle of a month have higher probability to occur fraud peaks.

Chart, line chart, histogram

Description automatically generated

**Filed 4:**

**Name:** Merchnum

**Description:** Merchant number that each transaction occurs. This field has 3.49% missing values and 13,091 unique values. The most common value in this field is 930090121224, which occurs 9,310 times in this field. The graph below shows top 15 merchant numbers in the dataset on a log basis.

Chart

Description automatically generated

**Filed 5:**

**Name:** Merch description

**Description:** Merchant description that each transaction occurs. This field has no missing values and 13,126 unique values. The most common value in this field is ‘GSA-FSS-ADV’, which occurs 1,688 times in this field. The graph below shows top 15 merchant description in the dataset.

Chart, bar chart

Description automatically generated

**Filed 6:**

**Name:** Merch state

**Description:** Merchant state that each transaction occurs. This field has 1.24% missing values and 227 unique values. The most common value in this field is ‘TN’, which occurs 12,035 times in this field. The graph below shows top 15 merchant state in the dataset.

Chart, bar chart

Description automatically generated

**Filed 7:**

**Name:** Merch zip

**Description:** Merchant zip that each transaction occurs. This field has 4.81% missing values and 4,567 unique values. The most common value in this field is 38118.0, which occurs 11,868 times in this field. The graph below shows top 15 merchant zip in the dataset on a log basis. In the data cleaning process, we need to transform it into string.

Chart, bar chart

Description automatically generated

**Filed 8:**

**Name:** Transtype

**Description:** Transaction type for each transaction. This field has no missing values and 4 unique values. The most common value in this field is ‘P’, which occurs 96,398 times in this field and accounts for most of the transactions. The graph below shows transaction types in the dataset on a log basis.

Chart, bar chart

Description automatically generated

**Filed 9:**

**Name:** Amount

**Description:** Transaction amount for each transaction. This field has no missing values and ranges from 0.01 to 3,102,045.53. The average value is 427.89 and standard deviation is 10,006.14. The graph below shows the distribution of amount ranging from 0 to 50,000, which accounts for over 99.99% of the population, and I divide it into 2,000 bins.

Chart, histogram

Description automatically generated

**Filed 10:**

**Name:** Fraud

**Description:** Fraud label for each transaction. This field has no missing values and 2 unique values. 0 represents no fraud and 1 represents fraud record. The most common value in this field is 0, which occurs 95,694 times in this field and accounts for most of the transactions. The graph below shows fraud label in the dataset on a log basis.

Chart, bar chart

Description automatically generated