Color backgroundTitle and subtitle with crop mark graphicAuthor and company name with crop mark graphic

**Overall Description:**

There are altogether 96,753 observations and 10 variables in this dataset. It includes the

transaction information of the credit card payment of the workers in federal government from

January 1st to December 31st in 2010. Each transaction is labeled as fraud (represented by 1)

or non-fraud (represented by 0). In total, there are 1059 fraudulent transactions in this dataset, and the overall fraud rate is 1.099%.

**Statistic Summary:**

**Table 1** shows the statistical summary for numeric fields in the dataset.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **# of records that have a value** | **% populated** | **# of unique values** | **# of records with value zero** | **Mean** | **std** | **min** | **max** |
| **Amount** | 96753 | 100 | 34909 | 0 | 427.89 | 10006.14 | 0.01 | 3102045.53 |

**Table 2** shows the information for categorical fields in the dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **# of records that have a value** | **% populated** | **# of unique values** | **# of records with value zero** | **Most Common Field Value** |
| Recordnum | 96753 | 1 | 96753 | 0 | NA |
| Cardnum | 96753 | 1.00 | 1645 | 0 | 5142148452 |
| Date | 96753 | 1.00 | 365 | 0 | 2/28/10 |
| Merchnum | 93378 | 0.97 | 13092 | 0 | 930090121224 |
| Merch description | 96753 | 1.00 | 13126 | 0 | GSA-FSS-ADV |
| Merch state | 95558 | 0.99 | 228 | 0 | TN |
| Merch zip | 92097 | 0.95 | 4568 | 0 | 38118 |
| Transtype | 96753 | 1.00 | 4 | 0 | P |
| Fraud | 96753 | 1.00 | 2 | 95694 | 0 |

**Field Analysis:**

1. **Recordnum:** a categorical variable with 96753 unique values. It is the unique ordinal reference number for each transaction record.
2. **Cardnum:** a categorical variable with 1645 unique values. It is the number of the credit card used in each transaction.

A picture containing fence

Description automatically generated

1. **Date:** a date variable which takes the value ranging from 2010-01-01 to 2010-12-31. It is the date when the credit card transaction actually happened. We can tell from the graphs that there is a strong seasonality in the transaction payment trend throughout the year 2010.

A picture containing computer

Description automatically generated

A close up of a mans face

Description automatically generatedA close up of text on a white background

Description automatically generated

A screenshot of a cell phone

Description automatically generated

1. **Merchantnum:** a categorical variable with 13092 unique values. It is the unique identifier for each merchant.

**A screenshot of a cell phone

Description automatically generated**

1. **Merch Description:** a categorical variable with 13126 unique values. It contains some detailed information about the merchant in each credit card transaction, such as the name of the merchant, etc. Top 10 variables show in the chart.

|  |  |
| --- | --- |
| GSA-FSS-ADV | 1688 |
| SIGMA-ALDRICH | 1635 |
| STAPLES #941 | 1174 |
| FISHER SCI ATL | 1093 |
| MWI\*MICRO WAREHOUSE | 958 |
| CDW\*GOVERNMENT INC | 872 |
| DELL MARKETING L.P. | 816 |
| FISHER SCI CHI | 783 |
| AMAZON.COM \*SUPERSTOR | 750 |
| OFFICE DEPOT #1082 | 748 |

1. **Merchant State:** a categorical variable with 228 unique values. It consists of not only the twoletter abbreviation of American states, but also some digit combinations that indicate some specific geographical locations all around the globe. Top 10 variables show in the chart.

|  |  |
| --- | --- |
| TN | 12035 |
| VA | 7872 |
| CA | 6817 |
| IL | 6508 |
| MD | 5398 |
| GA | 5025 |
| PA | 4899 |
| NJ | 3912 |
| TX | 3790 |
| NC | 3322 |

1. **Merchant Zip:** a categorical variable with 4568 unique values. It indicates the zip code of the merchant. Top 10 variables show in the chart.

|  |
| --- |
| 38118.0 11868 |
| 63103.0 1650 |
| 8701.0 1267 |
| 22202.0 1250 |
| 60061.0 1221 |
| 98101.0 1197 |
| 17201.0 1180 |
| 30091.0 1092 |
| 60143.0 942 |
| 60069.0 826 |
| 78682.0 817 |

1. **Transtype:** a categorical variable with 4 unique values: P, A, D and Y, among which P takes up the majority.

|  |  |
| --- | --- |
| P | 96398 |
| A | 181 |
| D | 173 |
| Y | 1 |

1. **Amount:** the only numeric variable in this dataset and it indicates the amount of each transaction. A screenshot of a cell phone

   Description automatically generated
2. **Fraud:** a categorical variable with two unique values: 0 and 1, with 0 representing non-fraudulent records and 1 representing fraudulent record. It is a label that is manually added to the dataset for educational purpose.

|  |  |  |
| --- | --- | --- |
| Fraud | Count | Percentage |
| 1 | 1059 | 1.1% |
| 0 | 95338 | 98.9% |