MGTA – 463: Fraud Analytics

HW5: Data Quality Report – Credit Card Transactions Data

Srikar Gunisetty

A53102026

SECTION-1: INTRODUCTION:

The data analysis contained in this report is primarily from an unidentified government organization in Tennessee. The dataset is of real government credit card transactions that was made public by the government organization for accounting purposes. However, the original dataset was not labeled and so Professor Stephen Coggeshall meticulously analyzed the data and created a binary label for each record in such a manner where the dataset is representative of realistic credit card transaction labels. Additionally, the records labeled "1" were carefully crafted by Professor Coggeshall based on his extensive subject matter expertise in fraudulent signals and activity. With each record containing a binary label, the dataset lends itself well to binary classification analysis.

File Name and Purpose: *Card Transactions.csv*

Data Collection/Source: A government organization in Tennessee and Prof. Stephen Coggeshall

Time period covered: 01/01/2006 to 12/31/2007

Number of records: 96,753

Fields: 10

Field Variables: RECNUM, CARDNUM, DATE, MERCHNUM, MERCH DESCRIPTION, MERCHANT STATE, MERCHANT ZIP, TANSTYPE, AMOUNT, FRAUD

SECTION-2: SUMMARY TABLES:

Table:1 – Field Names with Description

Field Name	Data Type	Type	Description
RECNUM	*int64*	*Categorical*	Record Number
CARDNUM	*Int64*	*Categorical*	Card Number
DATE	*Datetime64*	*Date-Time*	Transaction Date
MERCHNUM	*object*	*Categorical*	Merchant Number
MERCH DESCRIPTION	*object*	*Categorical*	Merchant Description
MERCH STATE	*object*	*Categorical*	Merchant State
MERCH ZIP	*float64*	*Categorical*	Merchant Zip code
TRANSTYPE	*object*	*Categorical*	Transaction Type
AMOUNT	*float64*	*Numerical*	Transaction Amount in [$]
FRAUD	*int64*	*Categorical*	Fraud Label

Table:2 – Summary Statistics for Numeric Fields

Field Name	Count	[%] full	Zeros	# Unique Values	mean [μ]	std [σ]	min	25%	50%	75%	max
AMOUNT	96,753	100.0	0	34,909	$427.89	$10,006.14	$0.01	$33.48	$137.98	$428.20	$3,102,046

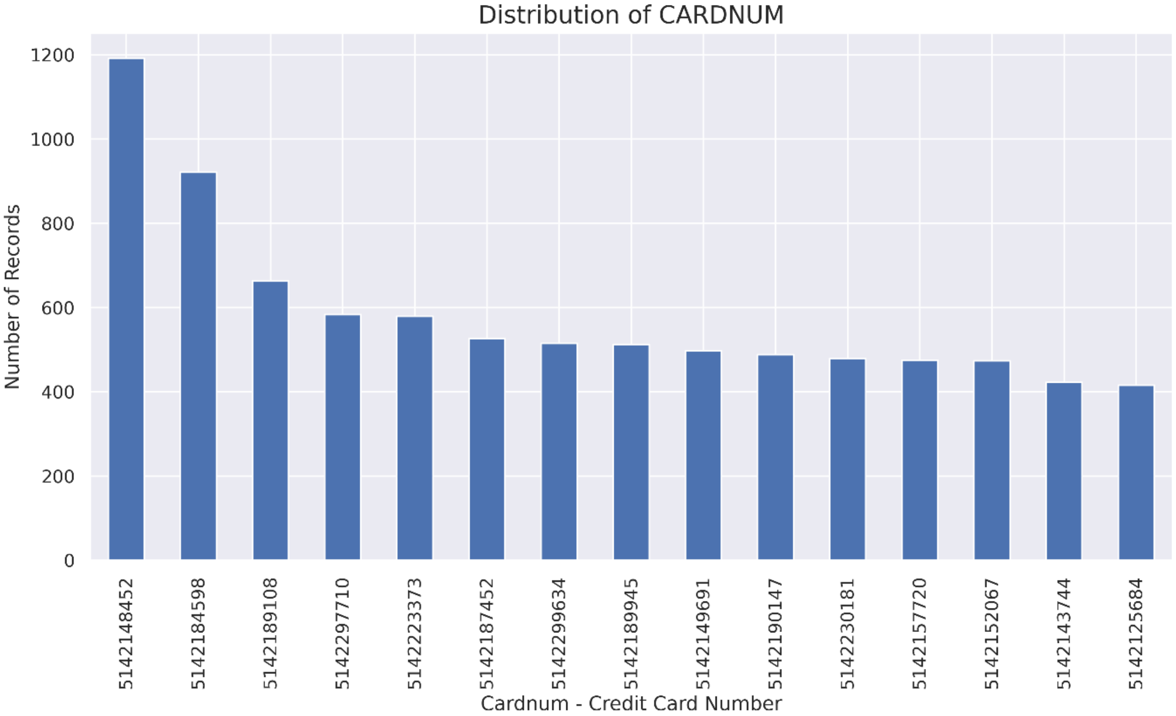
Table:3 – Summary Statistics for Categorical Fields

Field Name	Count	[%] full	# Unique	Most Frequent Value
RECNUM	96,753	100.0	96,753	N/A
CARDNUM	96,753	100.0	1,645	5142148452
DATE	96,753	100.0	365	2010-02-28
MERCHNUM	93,378	96.5	13,092	930090121224
MERCH DESCRIPTION	96,753	100.0	13,126	GSA-FSS-ADV
MERCH STATE	95,558	98.8	228	TN
MERCH ZIP	92,097	95.2	4,568	38118
TRANSTYPE	96,753	100.0	4	P
FRAUD	96,753	100.0	2	0

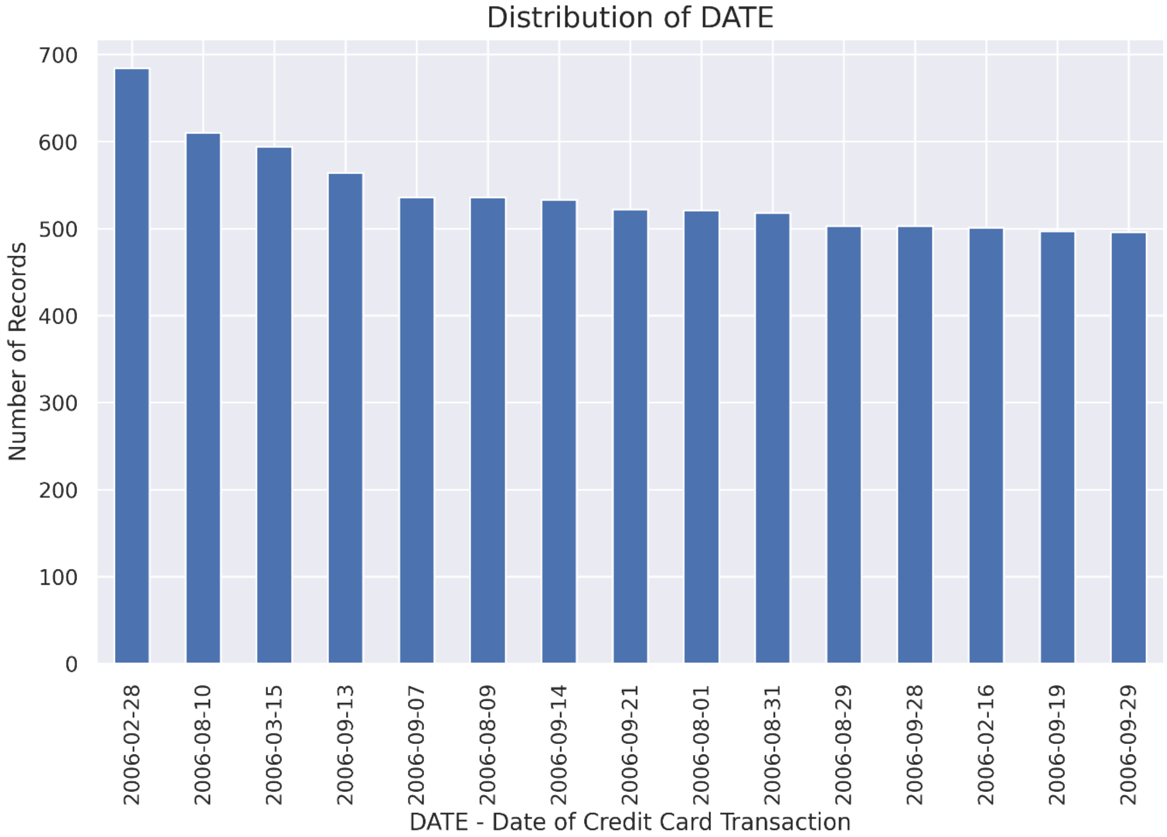
SECTION-3: FIELD EXPLORATIONS:

The subsections below provide additional detailed information about each data field within the dataset. The data fields are described in the order in which they appear

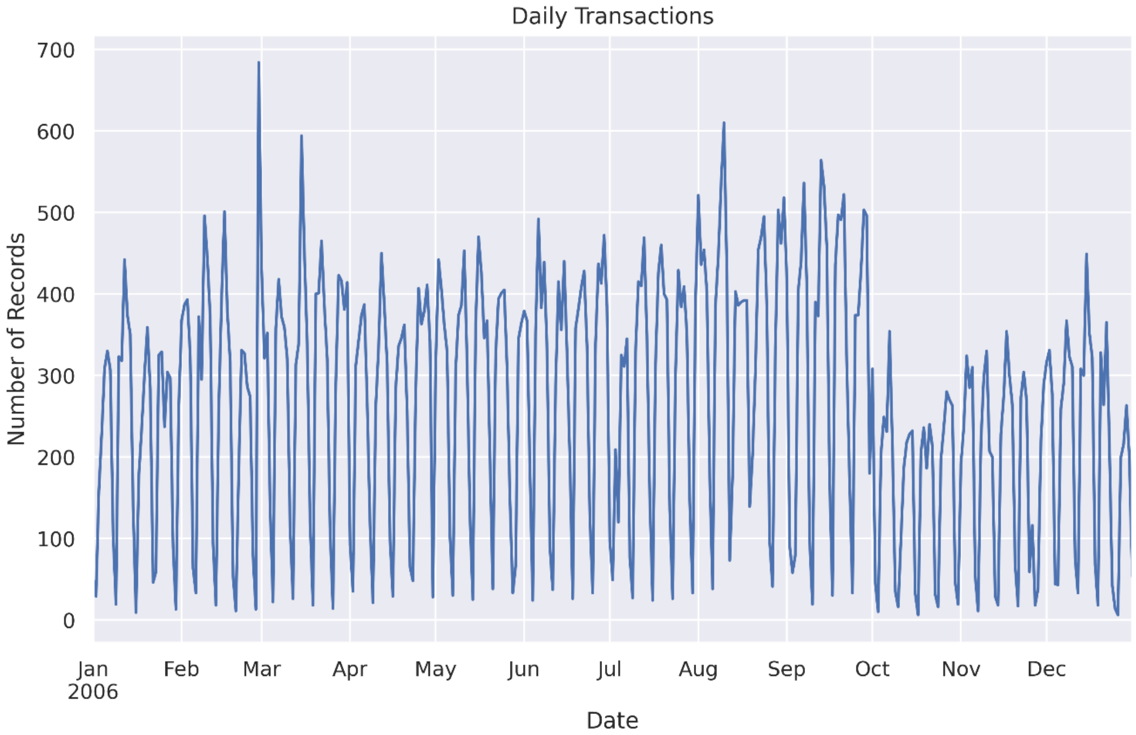
1. **RECNUM:**
   a. **Description:** A categorical data field containing an integer representing the unique card transaction record number identifier from 1 to 96,753. All records in the dataset contain a record number
2. **CARDNUM:**
   a. **Description:** A 10-digit categorical data field containing a shortened version of the credit card number for the card transaction record. The value is missing six internal digits from the true credit card number. The data field is 100% populated with 1,645 unique values. The bar chart below shows the top 15 values for the "Cardnum" data field.



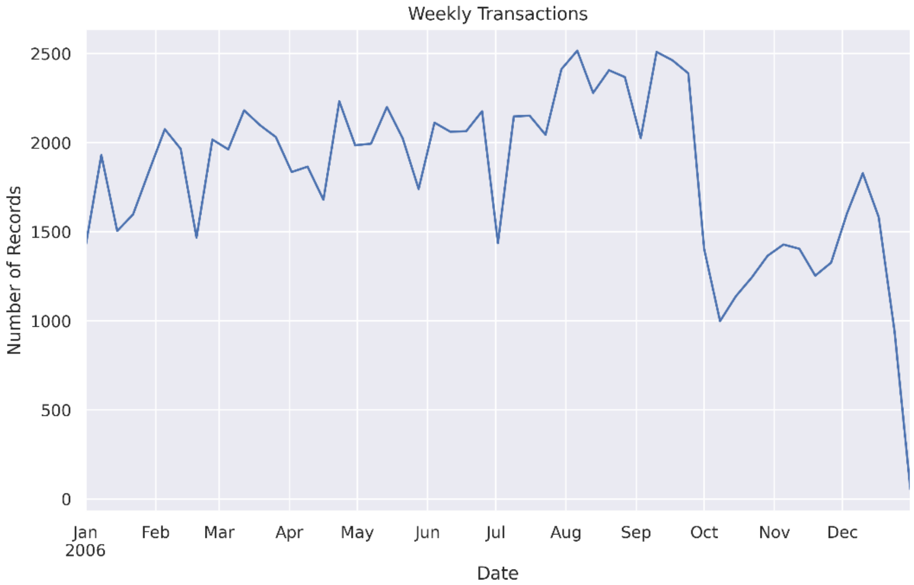
3. **DATE:**
   a) **Description:** A data field containing the date of the card transaction with a raw data format of MM/DD/YYYY. Upon importing the dataset, this data field has a converted format of YYYY-MM-DD. This data field is 100% populated and there are 365 unique values representing the 2006 calendar year.
   b) It is important to note that the dataset contains both weekly and annual seasonality. For the weekly seasonality, the data shows natural seasonality with regards to weekday and weekend activity.
   c) For the annual seasonality, the data shows activity related to government fiscal years. Government fiscal years start in October of the previous calendar year and end in September of the following calendar year. Thus, the 2006 government fiscal year started on 10/01/2006 and ended on 30/09/2006. The 2007 government fiscal year started on 10/01/2006 and ended on 30/09/2007. The dataset therefore contains records for both the 2006 and 2007 government fiscal years.
   d) The bar chart below shows the top 15 values for the "Date" data field. Given that the 2006 government fiscal year ended on 30/09/2006, we can see that a majority of the top 15 values are towards the end of the 2006 fiscal year in the months of August and September.



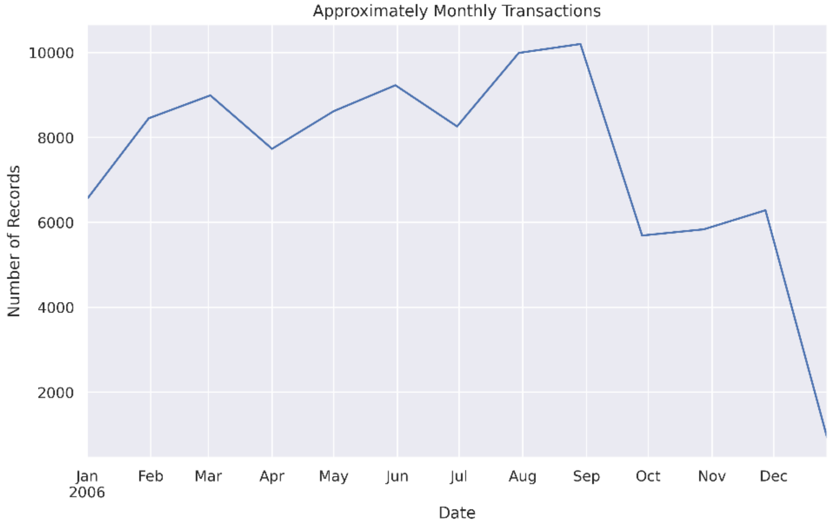
To view the weekly seasonality trends, the Daily Transactions graph below shows 52 peaks and troughs to represent the 52 weeks in a calendar year with the weekend activity depicted as the troughs.



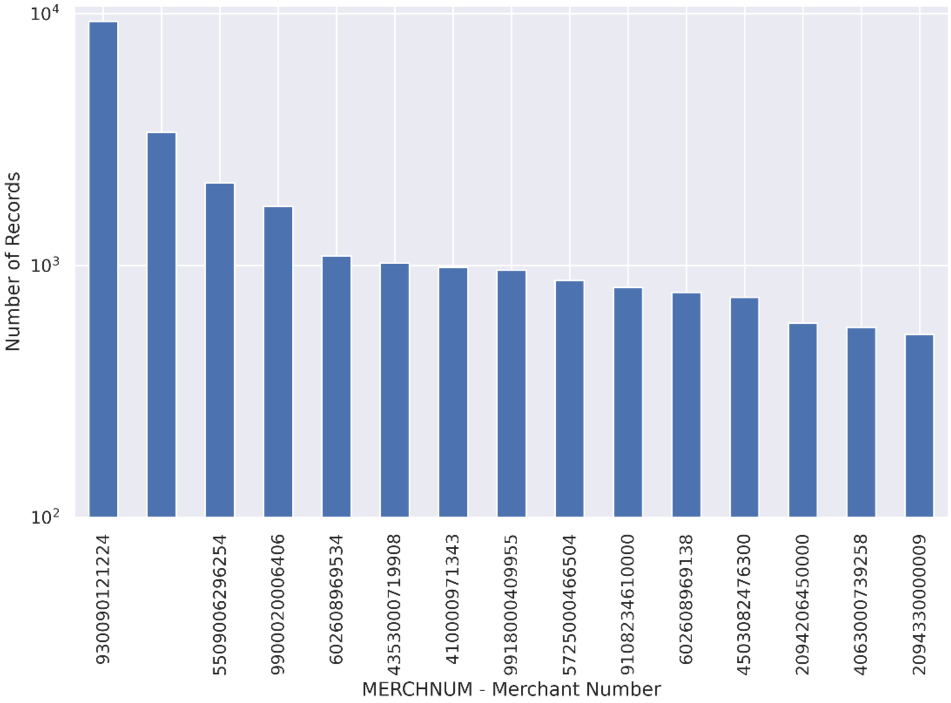
To view the annual seasonality, the Weekly Transactions graph below shows increased activity towards the end of the government fiscal year in September and then a sharp drop-in activity at the beginning of the next government fiscal year in October.



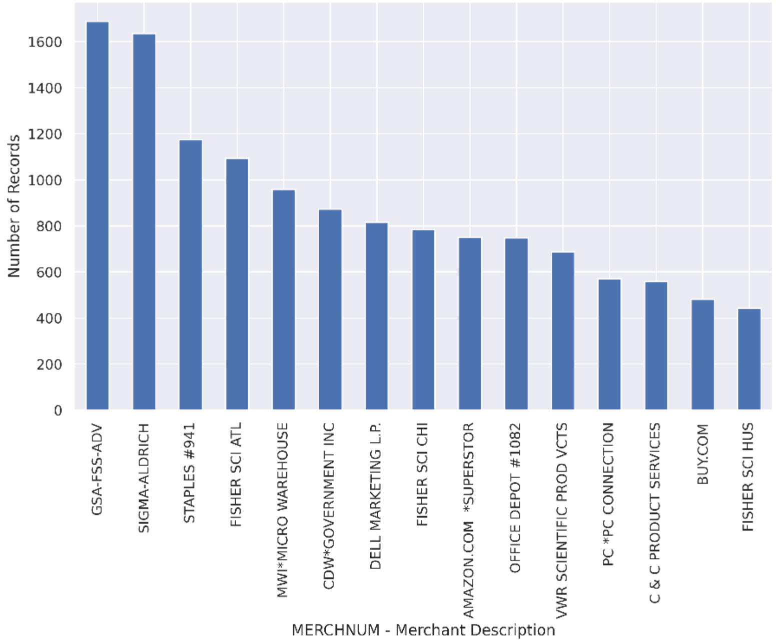
Additionally, the Monthly Transactions graph below provides a slightly different view of the annual seasonality in the data as well as the quarterly seasonality. With regards to the quarterly seasonality, there is increased activity towards the end of each quarter (March, June, September), and then a drop-in activity at the beginning of each quarter (January, April, July, October).



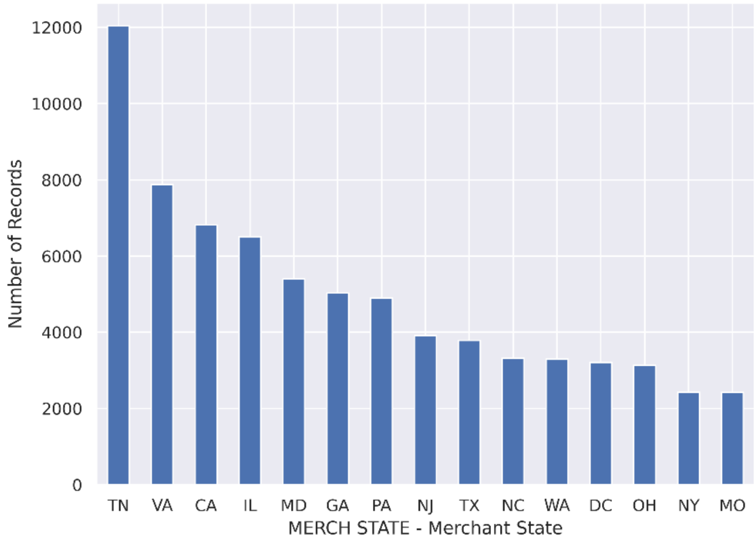
4. **MERCHNUM:**
   a) **Description:** A categorical data field containing a number identifier for the merchants in the dataset. There are 13,092 unique merchant numbers. This data field is 96.5% populated with only 3,375 records missing a value. The bar chart below provides the top 15 merchant numbers in the dataset.



5. **MERCH DESCRIPTION:**
   a) **Description:** A categorical data field containing a short text description of the merchant. This data field is 100% populated and there are 13,126 unique values for the merchant description. The bar chart below provides the top 15 merchant descriptions used for the records in the dataset.

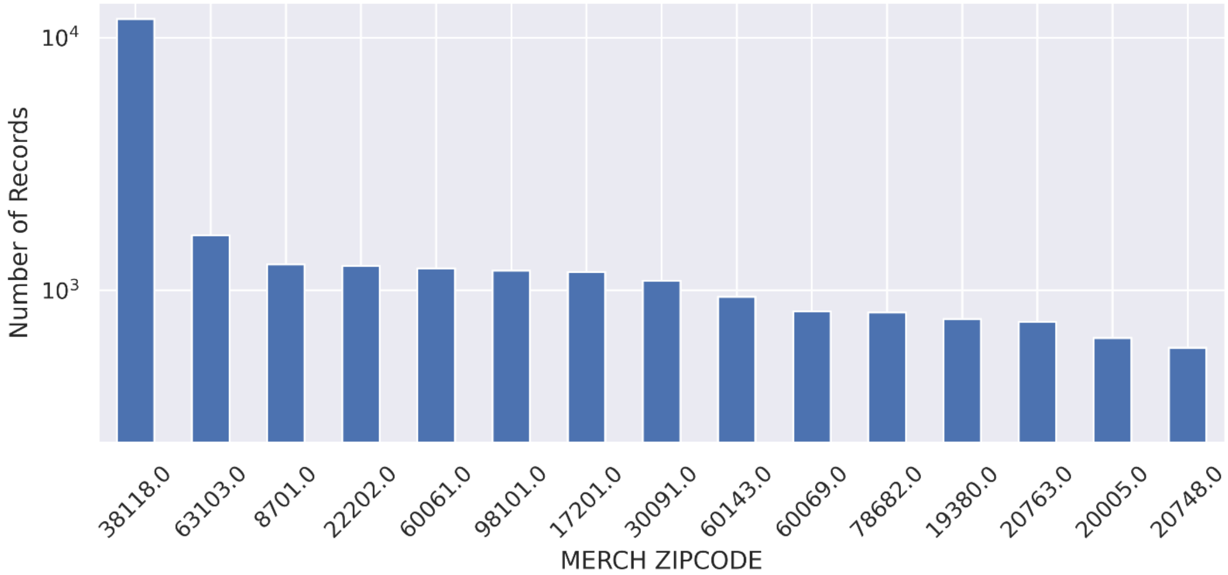


6. **MERCH STATE:**
   a) **Description:** A categorical data field containing either a 2-character or 3-digit value to represent the state associated with the merchant's location. There are 228 unique values for this data field in which 59 of the unique values use 2-character values and 168 of the unique values use 3-digit values. This data field is 98.8% populated with only 1,195 records missing a value. The bar charts below show the top 15 values for the "Merch state" data field.



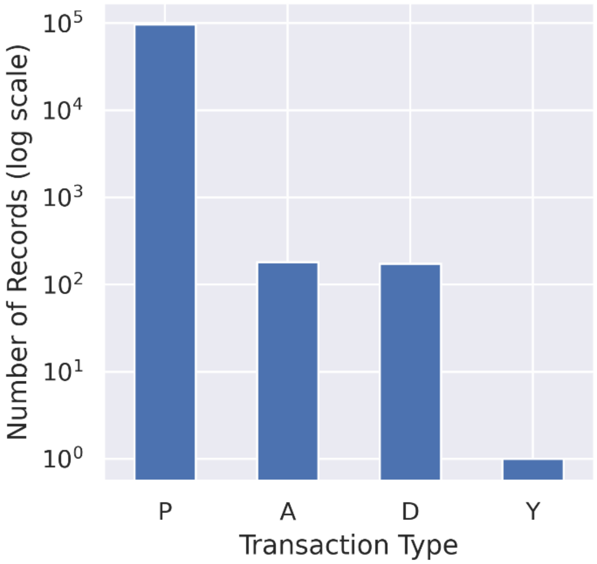
7. MERCH ZIP:

a) **Description:** A 5-digit categorical data field containing the zip code associated with the merchant's location. It is important to note that if a value has less than 5 digits, then the value has a leading zero(s). There are 4,568 unique values for this data field, and it is 95.2% populated with only 4,656 records missing a value. The bar chart below shows the top 15 values for the "Merch zip" data.



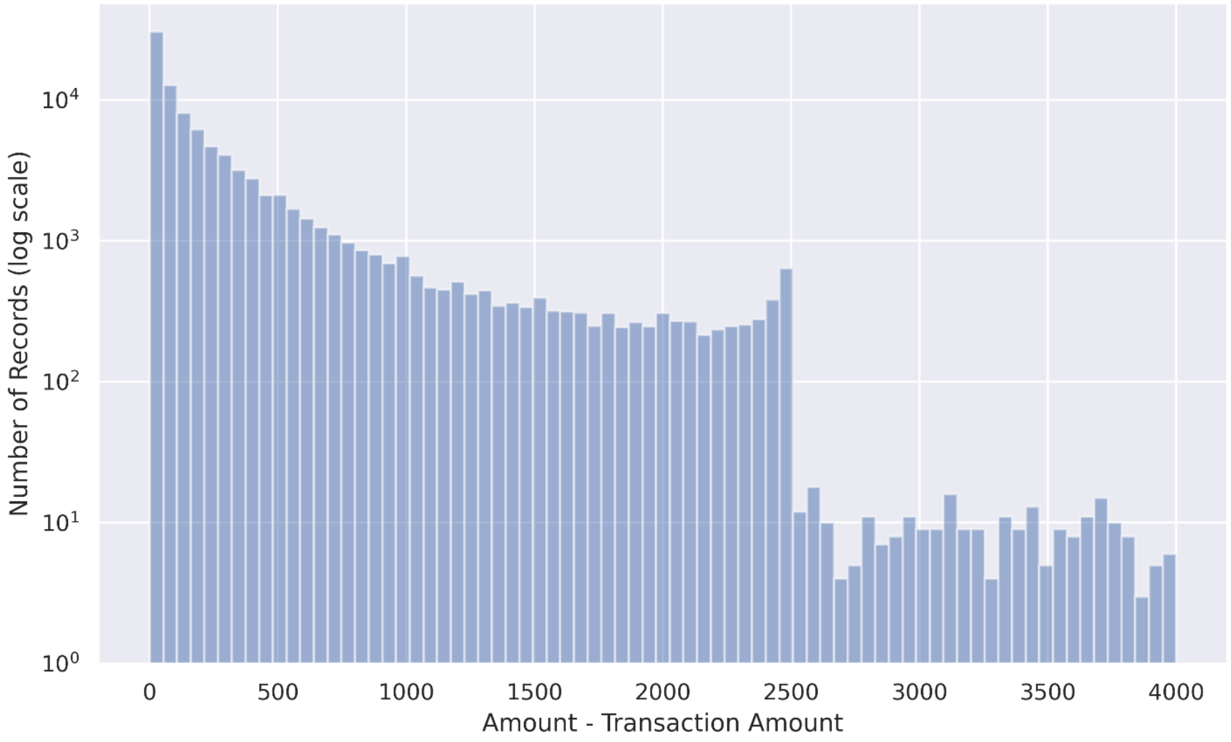
8. TRANSTYPE:

a) **Description:** A 1-character categorical data field containing the transaction type for the record. This data field is 100% populated with only four different values (P, A, D, or Y). There are 99.6% or 96,398 records containing a value of "P" for the transaction type, where "P" stands for purchase. The bar chart below shows all values for the "Transtype" data field.

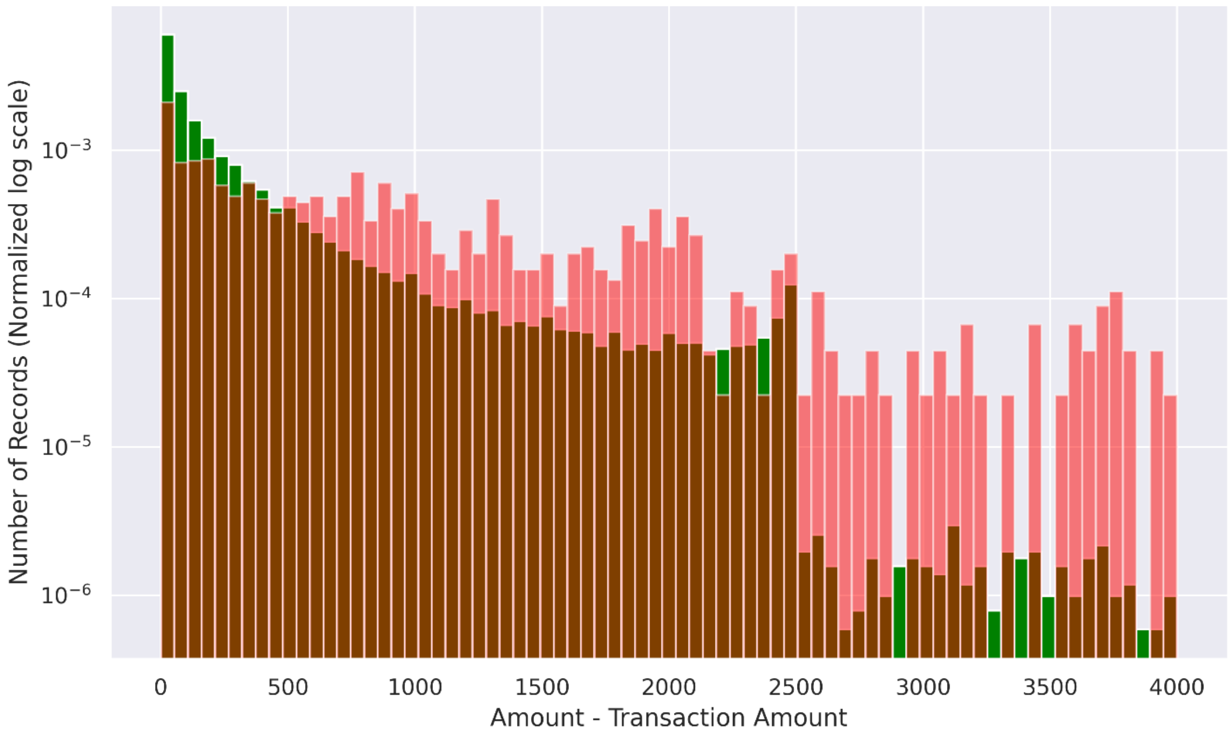


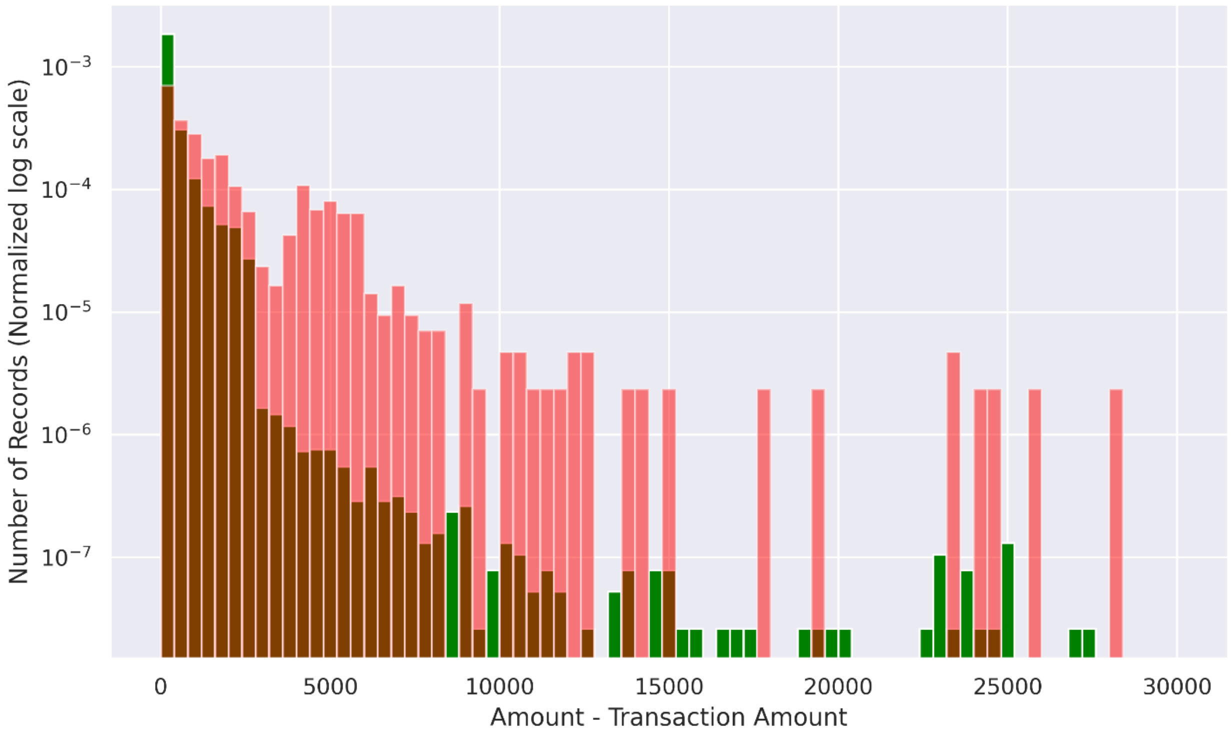
9. AMOUNT:

a) **Description:** A numerical data field containing the dollar amount charged to the credit card for that particular record. This data field is 100% populated and the values range from $0.01 to $3,102,045.53. However, it is important to note that the bulk of the purchases are below approximately $2,500 due to government purchase cards having a single purchase limit of $2,500 in 2010. The distribution plot below shows the dollar amounts for the credit card transactions in the dataset up to a value of $4,000 with the notable drop off depicted at/near the single purchase limit of $2,500.



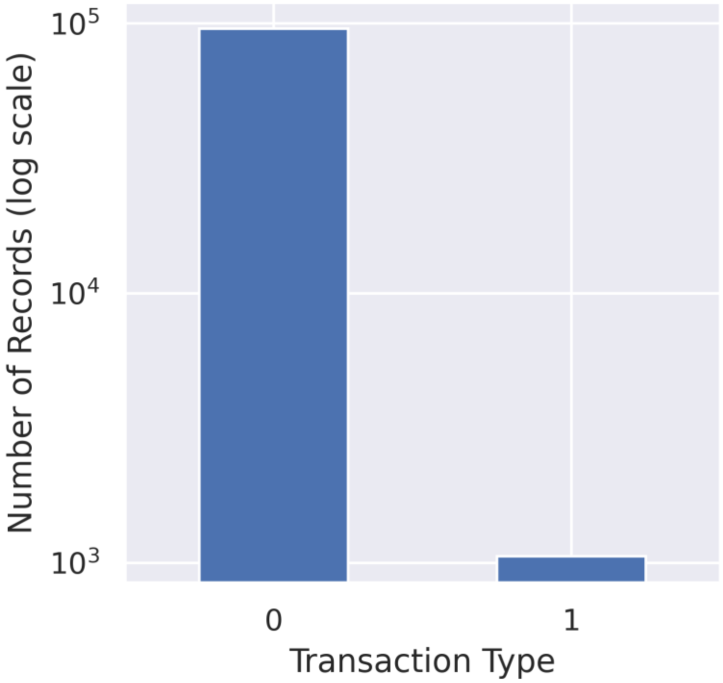
In addition, the distribution plots below show the differences in the amounts charged to the credit cards between the records labeled with a “0” (green) and “1” (red) in the “Fraud” data field. The first distribution plot shows the amounts charged up to $4,000, while the second distribution plot shows the amounts charged up to $30,000.


Number of Records (Normalized log scale)
10^{-3}
10^{-4}
10^{-5}
10^{-6}
0
500
1000
1500
2000
2500
3000
3500
4000
Amount - Transaction Amount

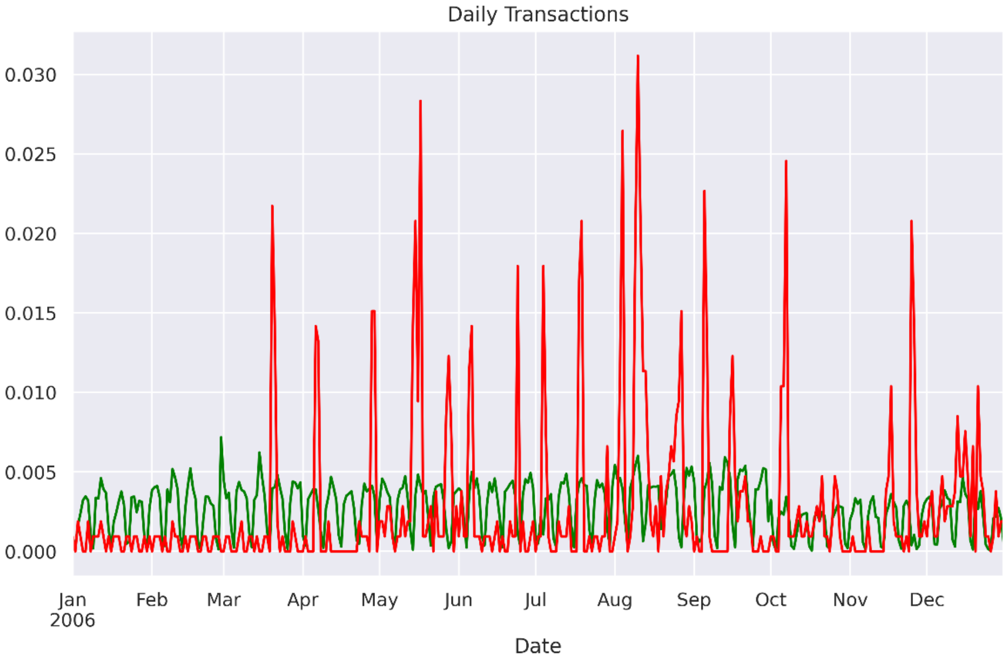

Number of Records (Normalized log scale)
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
0
5000
10000
15000
20000
25000
30000
Amount - Transaction Amount

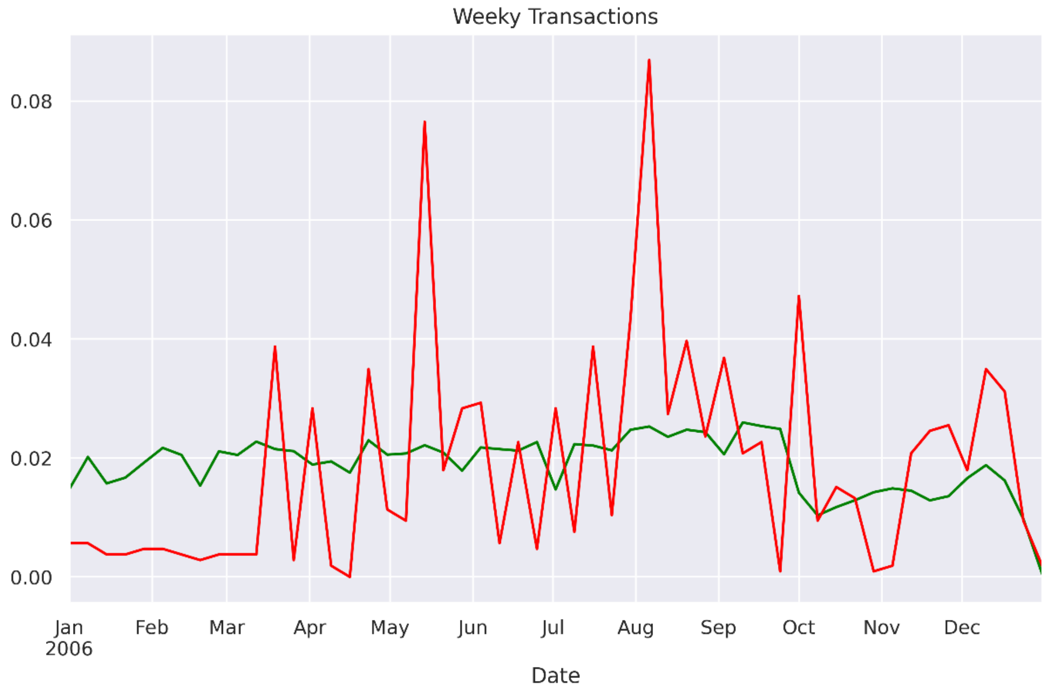
10. FRAUD:

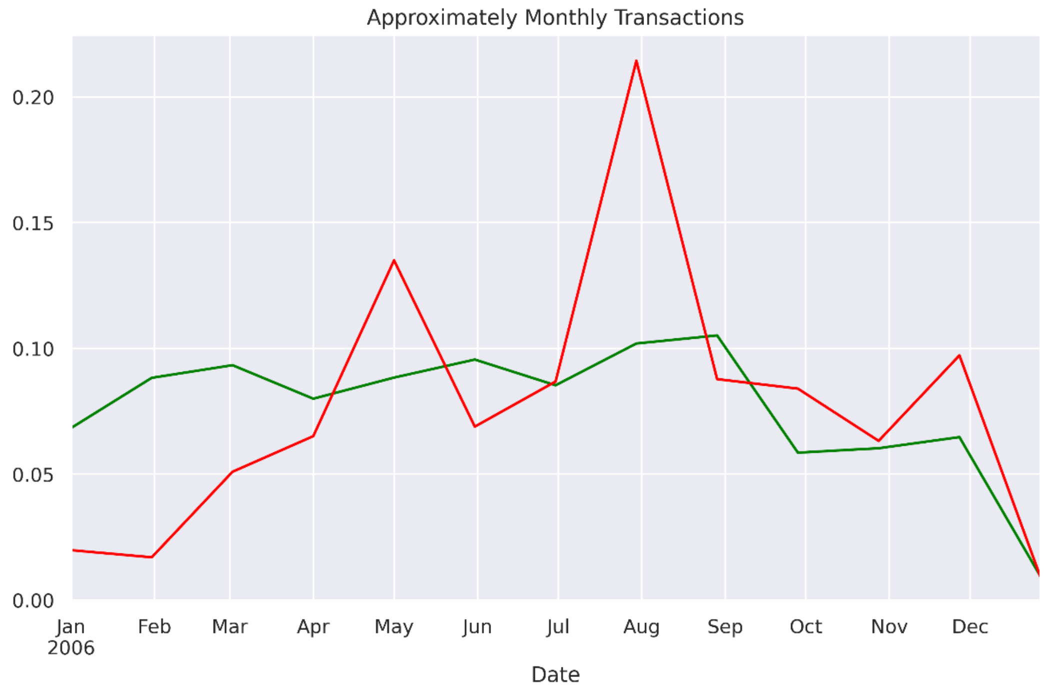
a) **Description:** A binary data field used to label the card transaction record as either a zero or one. There are 95,694 records labeled as "0" and 1,059 records labeled as "1" in the dataset. This data field is 100% populated.



MISC. DISTRIBUTIONS OF DAILY – WEEKLY – MONTHLY WITH FRAUD CLASSIFIER:




Weeky Transactions
0.08
0.06
0.04
0.02
0.00
Jan
2006
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Date


Approximately Monthly Transactions
0.20
0.15
0.10
0.05
0.00
Jan
2006
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Date