To: Professor Stephen Coggeshall
From: Alok Abhishek
Date: 04/02/2018
Subject: DSO 562: Fraud Analytics – Credit Card Transaction DRQ

Data Quality Report: Credit Card Transaction data set

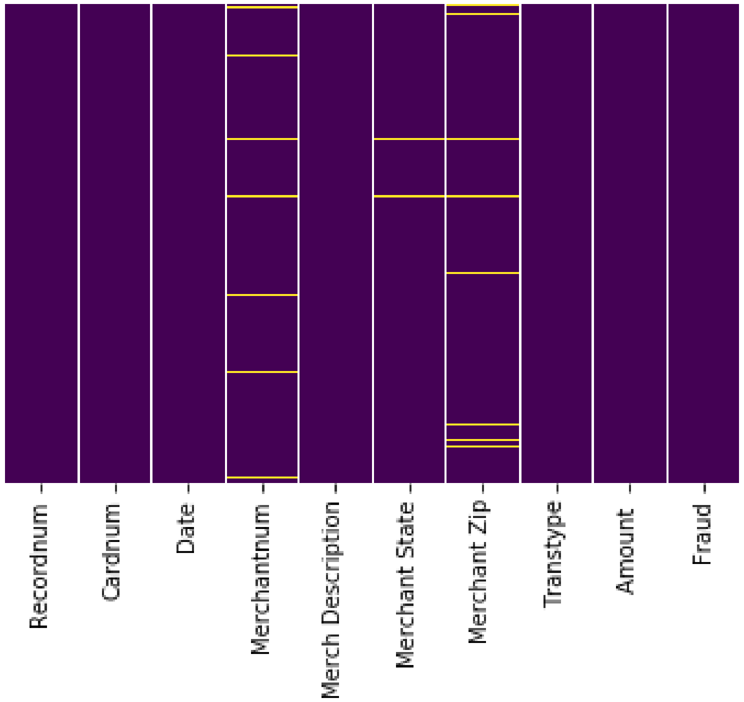
Data Description: The data is corporate credit card transaction data which covers card spending by a Government agency over the year 2010. Records are labeled with Fraud label.

There are 96,708 records and 10 columns with total data size of 6.3 MB.

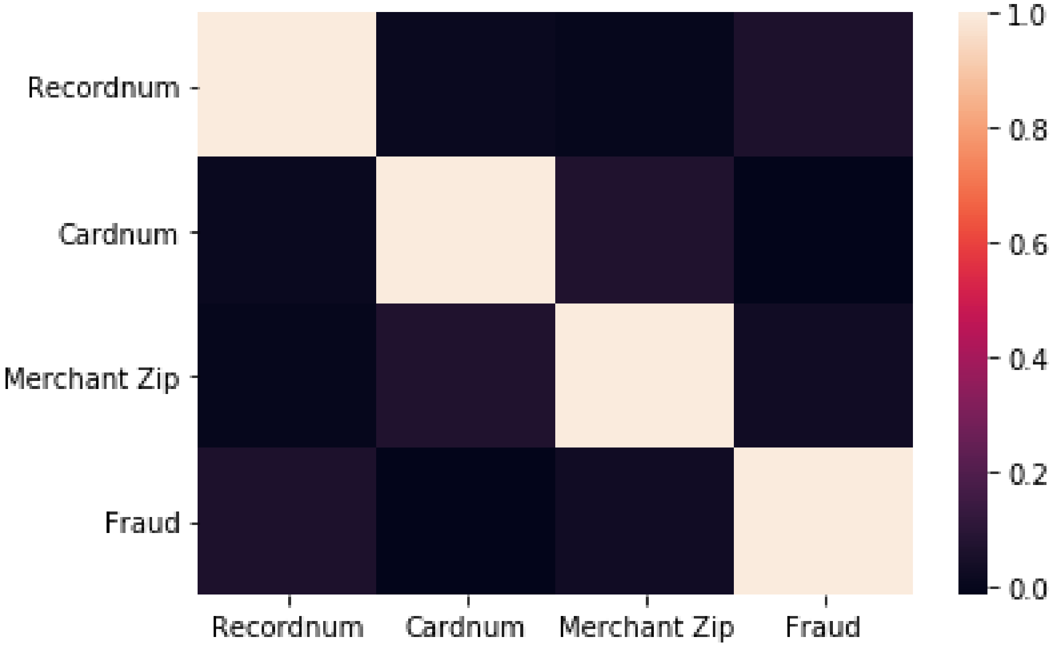
Data Summary:

Predictor	Data Type	Count	Mean	Std	Min	25%	50%	75%	Max	Percentage Populated	# of unique values
Cardnum	int64	96708	5142201000	53913	5142110000	5142152000	5142196000	5142246000	5142311000	100.00%	1644
Date	object	96708	6/26/2010		1/1/2010	4/3/2010	6/27/2010	9/13/2010	12/31/2010	100.00%	365
Merchantnum	object	96708								100.00%	13090
Merch Description	object	93333								96.51%	13124
Merchant State	object	96708								100.00%	227
Merchant Zip	float64	95513	44709.8176	28376.097	1	20855	38118	63103	99999	98.76%	4567
Transtype	object	92052								95.19%	4
Amount	object	96708	$427.87	$10,008.41	$0.01	$33.45	$137.90	$427.72	$3,102,045.53	100.00%	34875
Fraud	int64	96708	0.010485	0.101859	0	0	0	0	1	100.00%	2

As we can see from heat map of missing values the data set has some missing value for Merchant Number, Merchant State, and Merchant Zip Code.



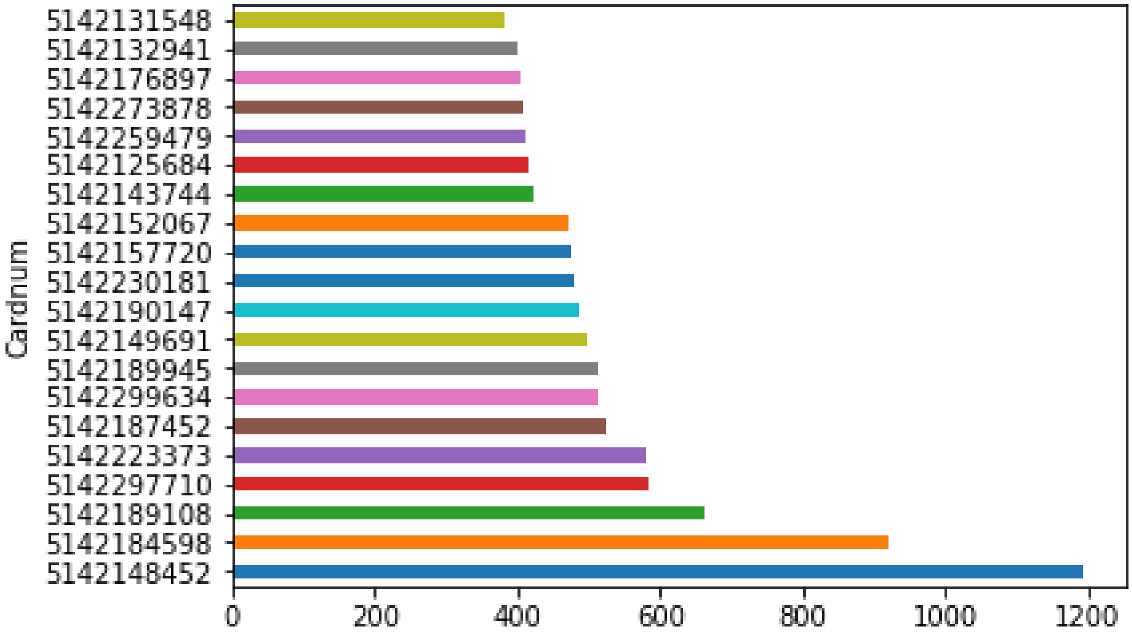
Also, from correlation plot we can see that there are no correlations in between different variables.



Card Number:

Predictor	Data Type	Count	Mean	Std	Min	25%	50%	75%	Max	Percentage Populated	# of unique values
Cardnum	int64	96708	5142201000	53913	5142110000	5142152000	5142196000	5142246000	5142311000	100.00%	1644

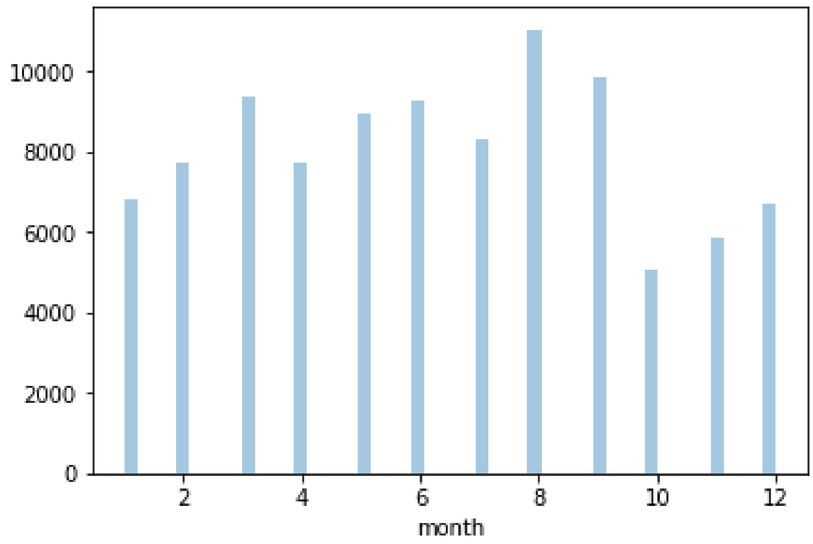
Looks like all the cards are master card as the card number stats from 51. Some of the most used card numbers are as following:



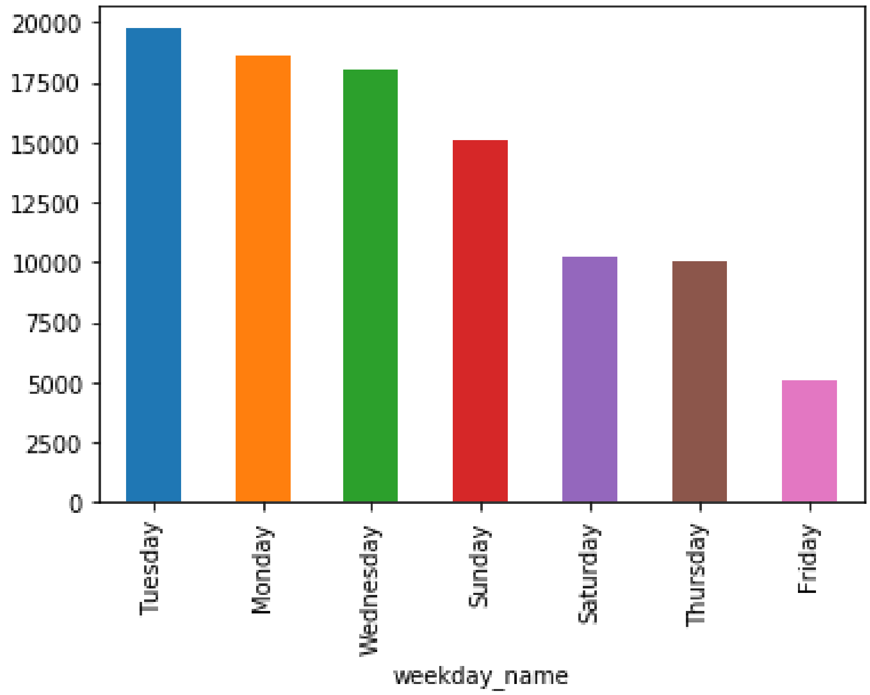
Date:

Predictor	Data Type	Count	Mean	Std	Min	25%	50%	75%	Max	Percentage Populated	# of unique values
Date	object	96708	6/26/2010		1/1/2010	4/3/2010	6/27/2010	9/13/2010	12/31/2010	100.00%	365

Looks like number of transactions on credit card gradually increases from Jan to Aug and then gradually reduces till end of the year and holiday season. October is the budget end month when new budget is planned and therefore the low purchase in October may indicate slower spending in final budget month.



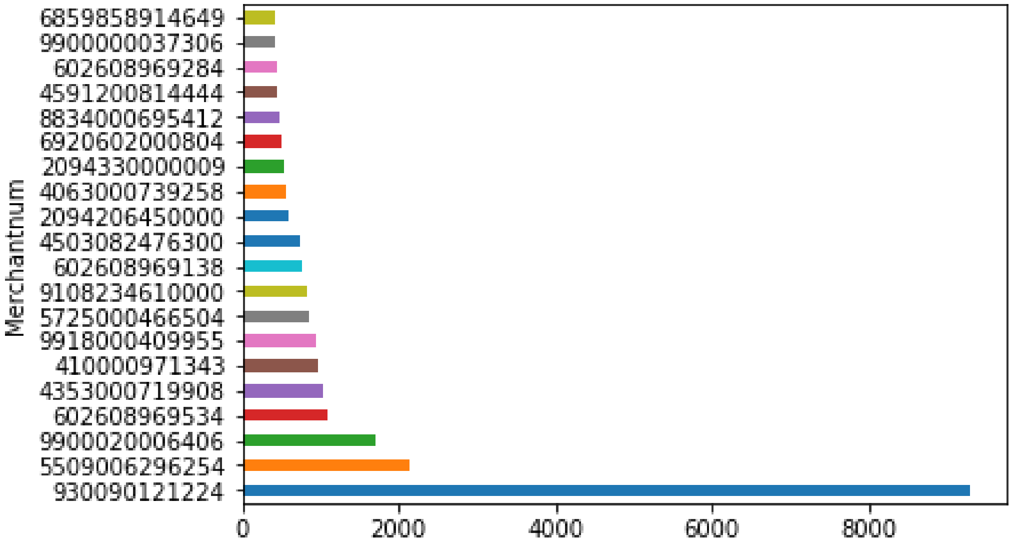
Looks like number of transactions peaks on Tuesday and we see lower number of transactions over weekends.



Merchant Number:

Predictor	Data Type	Count	Percentage Populated	# of unique values
Merchantnum	object	96708	100.00%	13090

Some of the most used merchant numbers are as following:

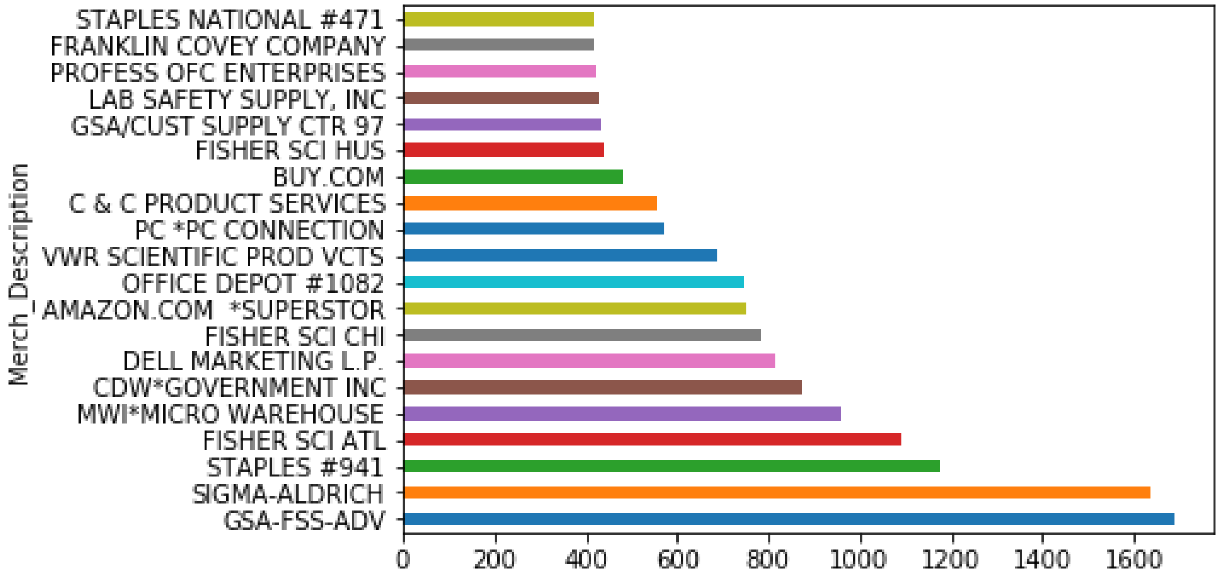


Merchant number	Count#
930090121224	9310
5509006296254	2131
9900020006406	1714
602608969534	1092
4353000719908	1020
410000971343	982
9918000409955	956
5725000466504	872
9108234610000	817
602608969138	783
4503082476300	746
2094206450000	590
4063000739258	568
2094330000009	533
6920602000804	523
8834000695412	478
4591200814444	463
602608969284	442
9900000037306	435
6859858914649	432

Merchant Description:

Predictor	Data Type	Count	Percentage Populated	# of unique values
Merch Description	object	93333	96.51%	13124

GSA (Government Services Administration) seems like the most common merchant. Following is the lost of top 20 merchants.



Some of the merchants like Staples and UPS or Amazon have small variation in their names. Fixing some of these variations. I see the following data:

Merchant	# of transactions
Amazon	1,025
Staples	2,656
FedEx	11,775
Government	3,493

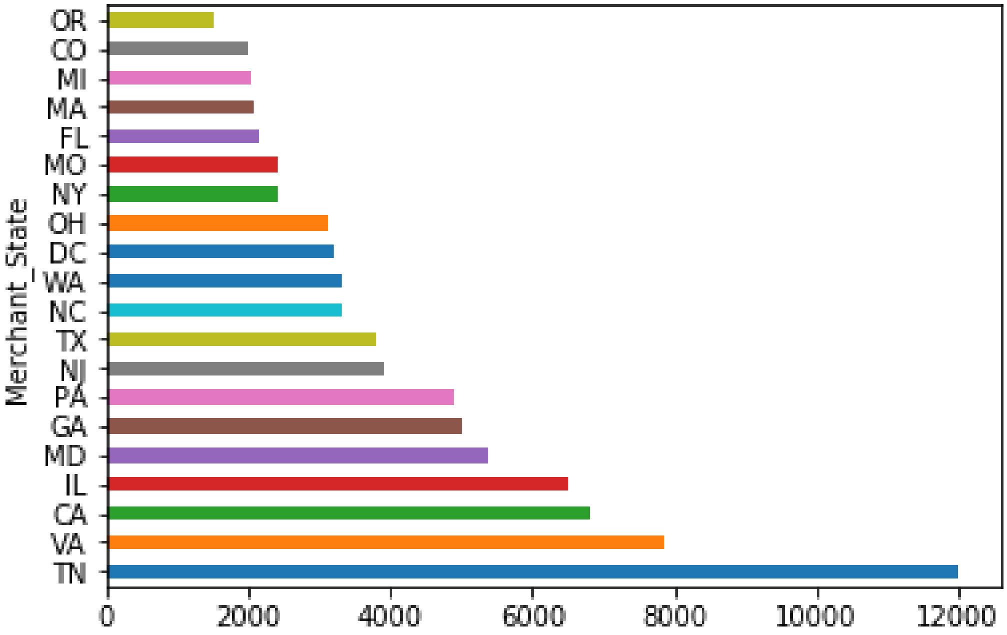
After removing the variations in merchant name, it seems like FedEx is the most frequently used merchant. This aligns well with the most frequent amount for transaction.

Merchant State:

Predictor	Data Type	Count	Percentage Populated	# of unique values
Merchant State	object	96708	100.00%	227

Number of unique values for state is 227 which shows that it usages not only the 52 states in the USA but also some other states which are numbered these could be military posts etc

Top 20 state usages:

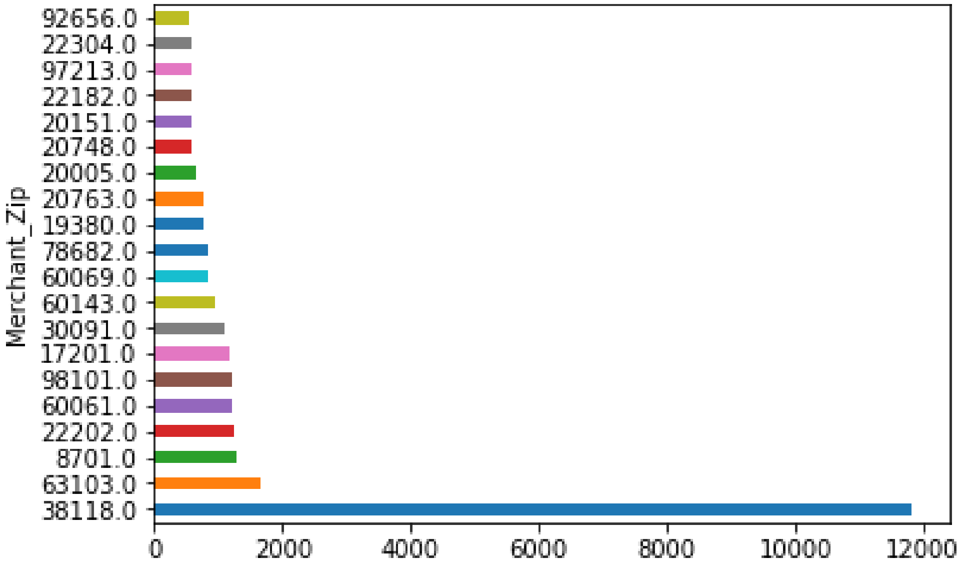


Merchant Zip:

Predictor	Data Type	Count	Mean	Std	Min	25%	50%	75%	Max	Percentage Populated	# of unique values
Merchant Zip	float64	95513	44709.8176	28376.09735	1	20855	38118	63103	99999	98.76%	4567

Zip codes are numbered from 1 to 99999 and there are 4567 unique zip codes. Some of these zip codes are not 5 digits.

Most frequent top 20 zip codes.

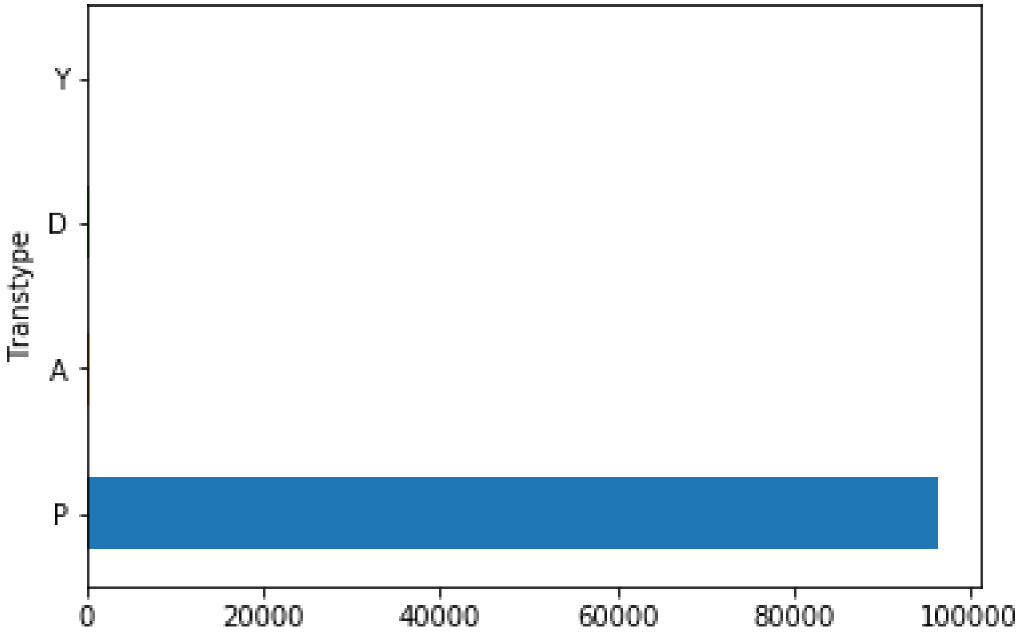


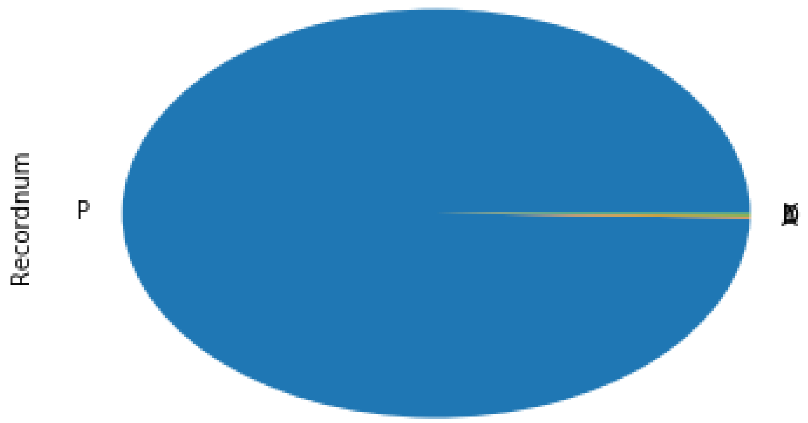
Transaction Type:

Predictor	Data Type	Count	Percentage Populated	# of unique values
Transtype	object	92052	95.19%	4

Transaction Type	# of transactions
P	**96,352**
A	**181**
D	**173**
Y	**1**

There are 4 different types of transactions, perhaps these are P – Purchase, A – Authorization, D – Delayed Capture.

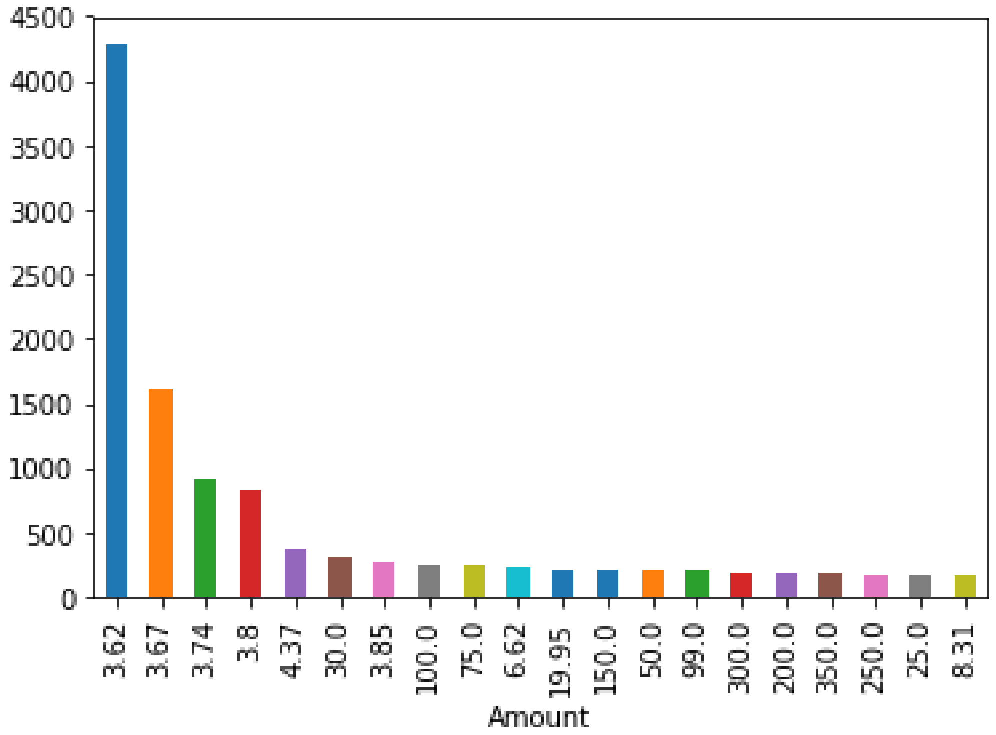




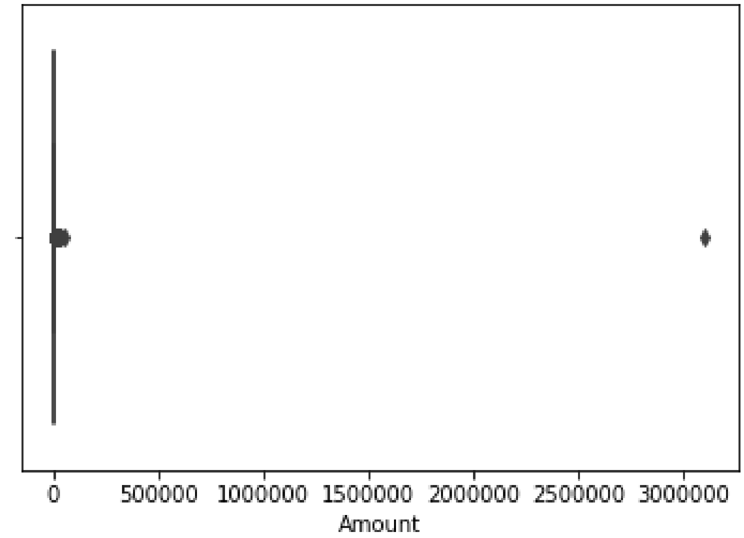
Amount:

Predictor	Data Type	Count	Mean	Std	Min	25%	50%	75%	Max	Percentage Populated	# of unique values
Amount	object	96708	$427.87	$10,008.41	$0.01	$33.45	$137.90	$427.72	$3,102,045.53	100.00%	34875

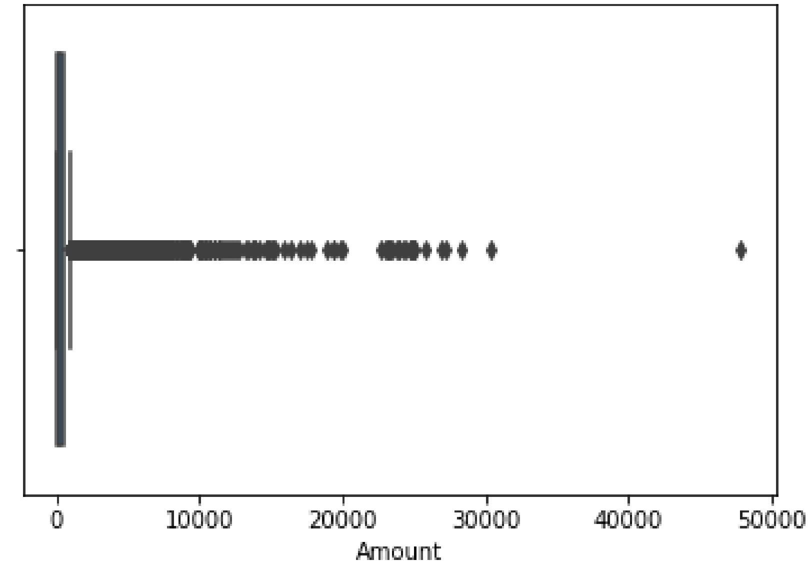
Transaction of about 4$ is the most common transaction, $3.62 being the most common. May be these transactions are cost of sending mail.



$3,102,045.53 seems to be the maximum value of transaction. With average transaction amount of of $427 and standard deviation of $10,000.

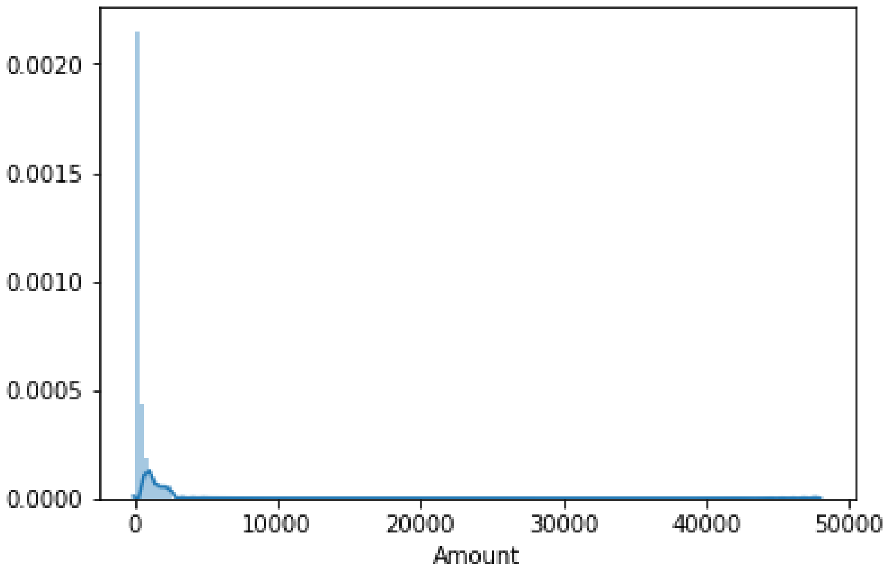


As you can see from above boxplot transaction of $3 million is an outlier. It seems like it was a transaction done in Mexican Pesos and not converted to USD. I will be excluding this transaction from further analysis. Box plot after deleting the $3M transaction.

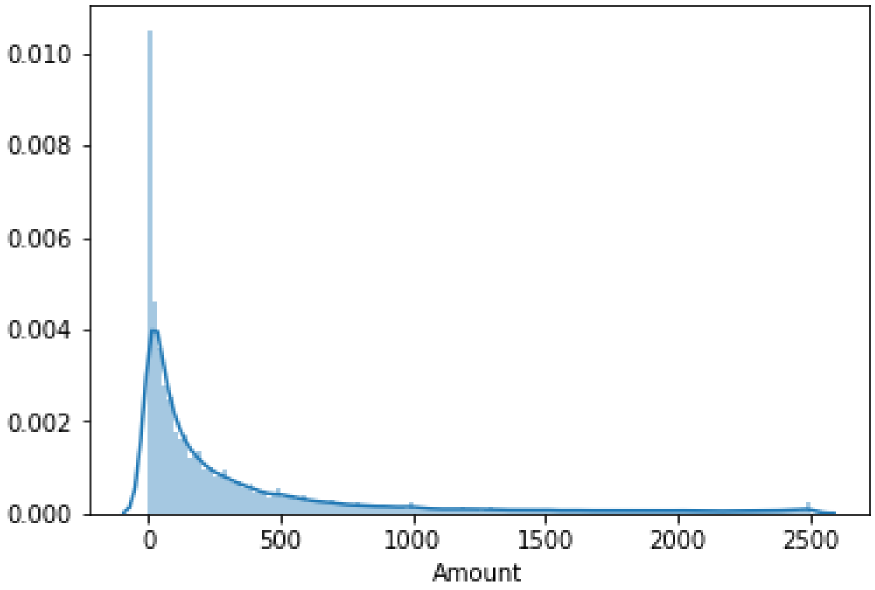


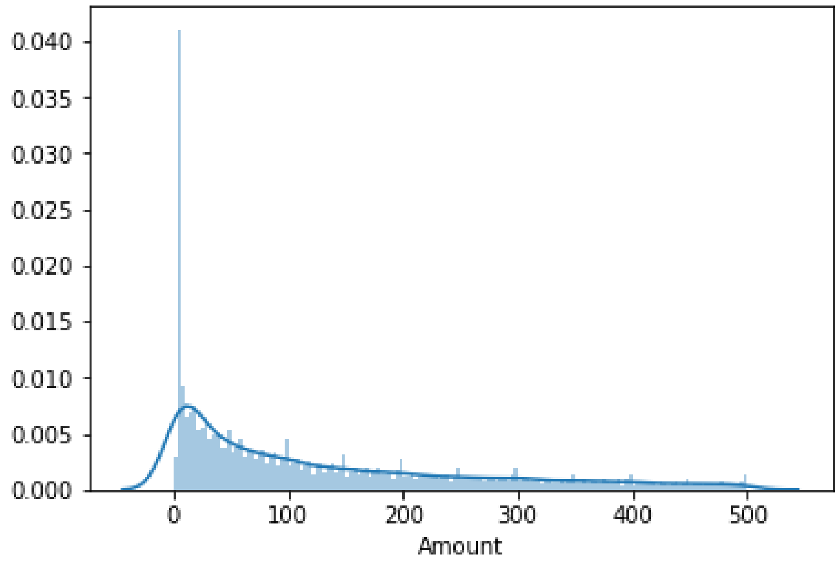
This box plot shows that most of the transactions are of small value and then some transactions are of high value shows as outliers in the box plot.

Similar trend can be seen in the distribution plot:



Looking a little closer in the spending revels similar trend.



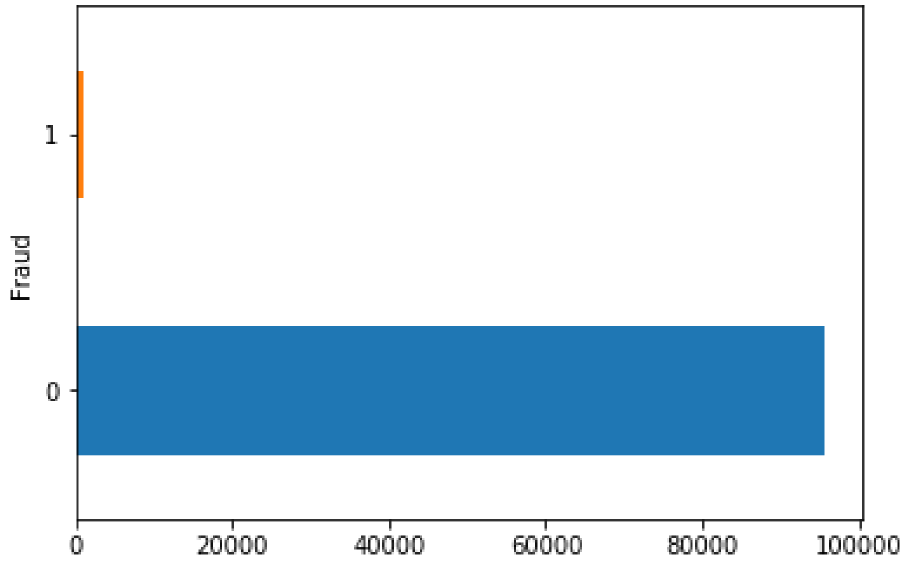


Looking even closer it seems like there is a spike in number of transaction at every $50 interval till $500.

Fraud:

Predictor	Data Type	Count	Mean	Std	Min	25%	50%	75%	Max	Percentage Populated	# of unique values
Fraud	int64	96708	0.010485	0.101859	0	0	0	0	1	100.00%	2

Most of the entries are labeled as not fraud.



Fraud Label	Count
0	**95,694**
1	**1,014**