REPORT – PROJECT 02

1. Executive Summary

The dataset comprises of Credit Card Transaction data. Data cleaning was performed on the raw dataset and subsequently important predictors to indicate fraud were created and selected. Post which, various models were explored, and their performance was analyzed, which then resulted in finalizing the model that performed the best on this data. Using our model, we can catch 56.2 % of the fraud in the top 3% of records, and we anticipate savings of $21,000,000/year.

Description of data

The dataset contains credit card transaction details from a US government organization. The information in the dataset pertains to the year of 2010. The dataset contains 96,753 records and 10 fields. It contains information about the transaction type and amount, merchant description and location, card number, and the record number. The dependent variable in this dataset is the fraud column which represents if that transaction is fraudulent or not.

Summary Tables:

i) **Numeric Table**

Field Name	% Populated	Min	Max	Mean	Std Dev	% Zero
Date	100	2010-01-01	2010-12-31	N/A	N/A	0
Amount	100	0.01	3,102,045.53	427.88	10,006.14	0

ii) **Categorical Table**

Field Name	% Populated	# Blank	# Zeros	# Unique Values	Most Common Values
Recnum	100.00	0	0	96,753	N/A
Cardnum	100.00	0	0	1,645	5142148452
Merchnum	96.51	3,375	231	13,091	930090121224
Merch Description	100.00	0	0	13,126	GSA-FSS-ADV
Merch state	98.76	1,195	0	227	TN
Merch zip	95.19	4,656	0	4,567	38118
Transtype	100.00	0	0	4	P
Fraud	100.00	0	0	2	0

2. Data Cleaning

Outlier Removal

We removed the transaction which showed an abnormally high amount of $3102045.53. In addition to this, we only kept the transactions labeled as “P”.

Imputation Logic

We removed the null values for Merch state, Merch num and Merch zip columns. Below is the imputation logic.

i) **Merch state** –

- If zip values for null-merch state values are within the range 00600 – 00799, 00900 – 00999, the merch state is hardcoded as PR (Puerto Rico).
- 3 dictionaries with "Merch state" as keys are created with its respective ‘merch zip’, ‘merch num’ and ‘merch description’ values (non-null only), and the mode ‘merch state’ value within these key-value pairs are picked to be imputed.
- For the remaining null values, merch state is labelled as “unknown” for the two merch description categories – RETAIL DEBIT ADJUSTMENT and RETAIL CREDIT ADJUSTMENT
- Finally, the remaining are labelled as “unknown”.

ii) **Merch num** –

- All zeros are replaced with null
- A dictionary with "Merch num" as keys is created with its respective ‘merch description’ values (non-null only), and the mode merch num value within these key-value pairs are picked to be imputed.
- For the remaining null values, merch num is labelled as “unknown” for the two merch description categories – RETAIL DEBIT ADJUSTMENT and RETAIL CREDIT ADJUSTMENT
- For the remaining nulls, for which there is no available merch description, ‘merch num’ values are synthesized by increments of the max ‘merch num ‘values available in the dataset.

iii) **Merch zip** –

- A dictionary with "merch zip" as keys is created with its respective ‘merch num’ and ‘merch zip’ values (non-null only), and the mode ‘merch zip’ value within these key-value pairs are picked to be imputed.
- For the remaining null values, ‘merch zip’ is labelled as “unknown” for the two merch description categories – RETAIL DEBIT ADJUSTMENT and RETAIL CREDIT ADJUSTMENT
- Finally, the remaining are labelled as “unknown”.

3. Variable Creation

This step mainly involves feature engineering and coming up with new variables. The aim of this step is to come up with a variety of variables that we think affect the predictions of fraudulent transactions. For instance, we created Benford’s law - variables (measures the unusualness of the first digit distribution relative to Benford’s law) for both Cardnum and Merchnum and we created the Risk variable which gives us the likelihood of fraud for that day of the week. We additionally created variables with different entities and time periods (e.g., Velocity Change, Day since, Frequency and Amount variables). The below table gives us the details of these variables created.

Summary of variables

Description of Variables	# Variables Created	
Original variables excluding 'Recnum' and 'Fraud'	8	
Risk Variable	1	
New entities combining/concatenating original fields	13	
Days - Since Variables - # Days since a variable with an entity has been seen	18	
Benfords Law Variables - Measures the unusualness of the first digit distribution relative to Benford's Law	2	
Velocity Change Variables - Fraction of the number of applications with the same application - attributes over the past 0/1 day out of the total number of applications with the same application - attributes over {3,7,14,30} days	144	
Frequency Variables - # Applications unique per entity in the last {0,1,3,7,14,30} days	972	
Amount Variables - a) Includes maximum frequency across fields in the last {0,1,3,7,14,30} days b) Also includes maximum, minimum and mean age, median per field		
Additional Columns - a) Month b) DOW	2	
Amount Parameters - Amount - statistics per 'card num' to flag unusual transactions	4	New
Days_Since+Merchnum+Frequency+Amount	102	New
Total	**1266**	

New Variables:

1. Amount – statistics per 'card num' to flag unusual transactions (4 variables):
   Max_cardnum: The maximum amount for that card number
   Min_cardnum: The minimum amount for that card number
   Median_difference_Cardnum: The percentage difference between the amount and the median amount for that card number
   Mean_difference_cardnumber: The percentage difference between the amount and the mean amount for that card number.

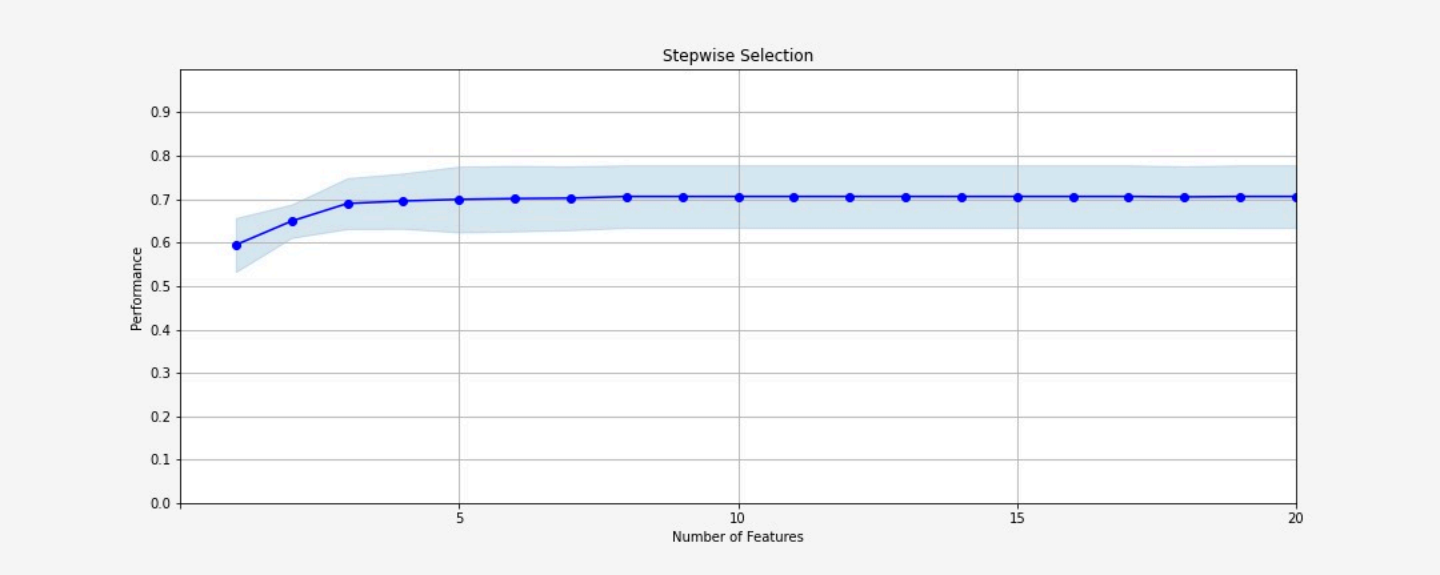
2. Days_Since+Merchnum+Frequency+Amount(102 variables):
   This is a combination of entities to check the frequency of the card numbers with the days since the card number has been used, along with the frequency and the merchant number.
   This essentially tells us if there have been a lot of transactions happening at a particular merchant in a short span of time.

4. Feature Selection

In this step we first ran a filter with around 300 columns, including 2 months of OOT data and then explored different models to run the wrapper with both forward and backward selection. The final model selected was the forward LGBM with num_filters = 300 and num_wrappers = 20. This model was selected because of the following reasons:

- This model gave the highest performance of almost over 0.7 when compared to the other models.

- LGBM forward selection model was quicker to run (around 1 hour)
- This model run selected a good mix of short-term, long-term variables and different entities.



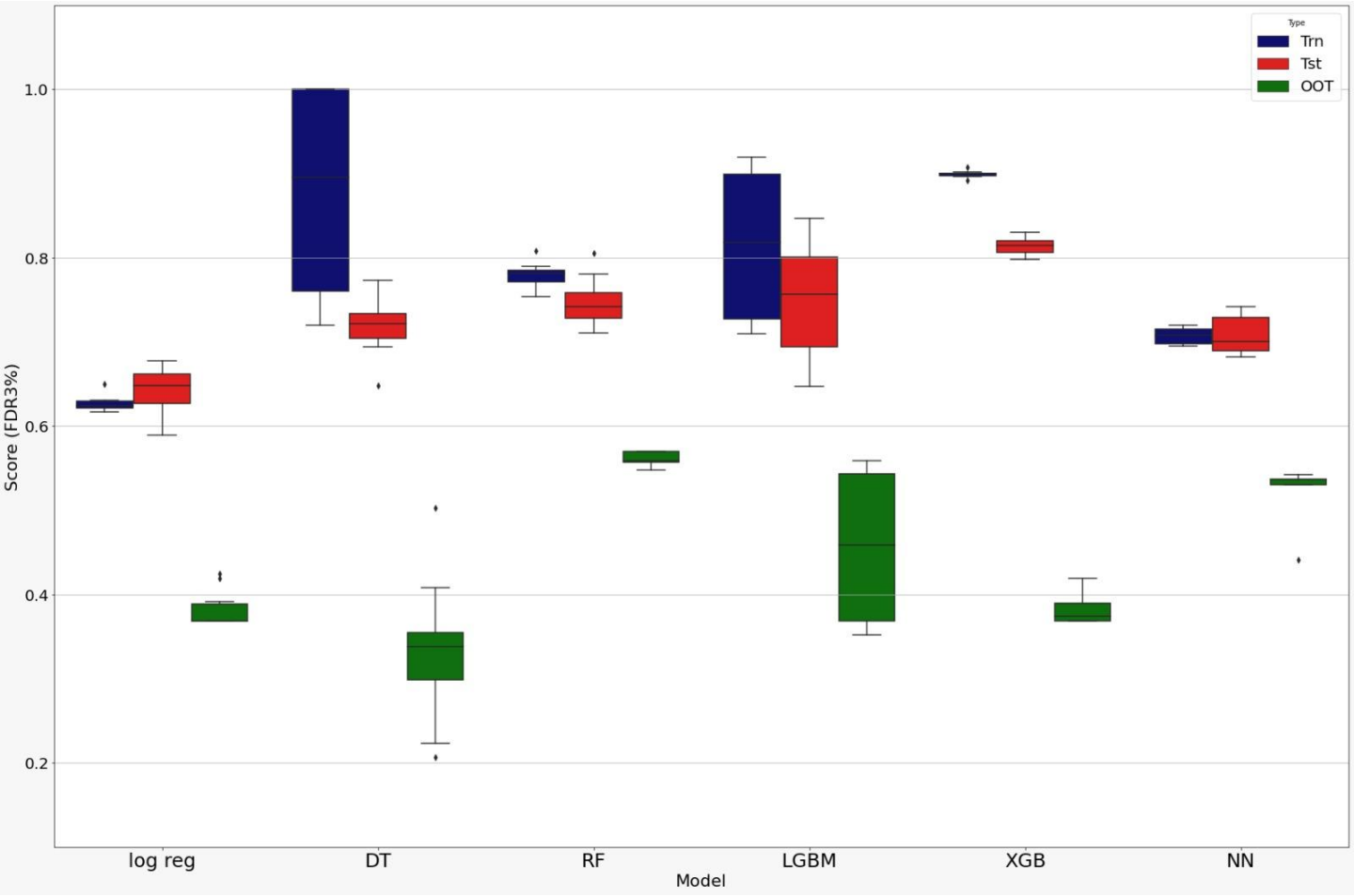
List of Final Variables

wrapper order	variable	filter score
1	card_zip_total_14	0.650480728
2	card_zip3_max_30	0.612028897
3	card_merch_total_0	0.544203337
4	Card_Merchdesc_avg_60	0.516781821
5	merch_zip_total_14	0.439451061
6	Card_Merchnum_desc_avg_3	0.517344017
7	Merchnum_desc_avg_0	0.508087779
8	Card_Merchnum_desc_avg_0	0.506822721
9	Card_Merchnum_desc_avg_7	0.517248507
10	Card_Merchdesc_avg_14	0.516417757
11	Card_Merchdesc_avg_7	0.515927775
12	Card_Merchnum_desc_avg_60	0.513812325
13	Card_Merchdesc_avg_1	0.512517464
14	Card_Merchnum_desc_avg_1	0.508992002
15	Card_Merchdesc_avg_0	0.506801743
16	Card_Merchdesc_avg_3	0.516342115
17	Merchnum_desc_avg_1	0.504887803
18	Merchnum_avg_1	0.505884535
19	card_merch_avg_30	0.520234365
20	card_merch_avg_3	0.522054963

5. Preliminary Model Exploration

After having selected our features, we run the data through different models. Here, we explored 6 models - Logistic Regression, Decision Tree, Random Forest, XGB, Neural Network and LGBM. We tuned each of these models with different combinations of hyperparameters and observed a range of performances for each model, as seen below. We ran each model 10 times and averaged the results.

Model		Parameter						Average FDR @ 3%		
Logistic Regression	Number of variables	Penalty	c	solver		l1_ratio		Train	Test	OTT
1	10	L1	1	saga		None		0.629	0.626	0.374
2	10	L1	0.9	saga		None		0.626	0.636	0.376
3	10	L1	0.7	saga		None		0.626	0.63	0.379
4	10	L2	1	liblinear		None		0.618	0.625	0.439
6	10	L2	1	saga		1		0.629	0.632	0.38
7	10	Elasticnet	1	liblinear		1		0.632	0.606	0.453
8	10	Elasticnet	1	liblinear		0.7		0.63	0.62	0.447
Decision Tree	Number of variables	criterion	splitter	max_depth	min_samples_split	min_samples_leaf		Train	Test	OTT
1	10	gini	random	2	1000	500		0.198	0.206	0.151
2	10	gini	best	10	50	50		0.791	0.738	0.426
3	10	gini	best	15	40	35		0.858	0.747	0.343
4	10	gini	best	20	20	30		0.88	0.7482	0.349
5	10	entropy	random	25	15	25		0.771	0.69	0.403
6	10	entropy	best	30	10	10		1	0.722	0.315
7	10	gini	best	30	8	8		0.999	0.71	0.327
8	10	gini	randomt	25	5	5		0.9761	0.685	0.341899
Random Forest	Number of variables	n_estimators	criterion	max_depth	min_samples_split	min_samples_leaf		Train	Test	OTT
1	10	3	gini	2	50	30		0.569	0.578	0.336
2	10	20	entropy	5	45	25		0.766	0.732	0.544
3	10	30	gini	8	40	20		0.783	0.759	0.546
4	10	40	entropy	10	35	17		0.911	0.808	0.512
5	10	50	gini	15	30	15		0.949	0.794	0.474
6	10	70	entropy	20	20	10		0.999	0.810	0.379
7	10	100	entropy	30	5	3		1.000	0.818	0.347
XGB	Number of variables	n_estimators	max_depth	Booster	eta	subsample	min_child_weight	Train	Test	OTT
1	10	1	2	gbtree	0.3	1	1	0.529	0.524	0.259
2	10	3	6	gbtree	0.3	1	1	0.714	0.704	0.531
3	10	50	100	gbtree	0.3	1	1	1.000	0.799	0.346
4	10	30	100	gbtree	0.3	0.8	5	0.928	0.801	0.365
5	10	30	100	gbtree	0.01	1	1	0.726	0.705	0.529
6	10	100	50	gblinear	0.3	1	1	0.622	0.629	0.440
7	10	100	50	gbtree	0.3	0.4	10	0.891	0.791	0.399
Neural Network	hidden_layer_sizes	activation	alpha	solver	learning_rate	learning_rate_iter		Train	test	OOT
1	2	tanh	0.01	sgd	constant	0.1		0.615994	0.616643	0.408939
2	5	tanh	0.1	adam	constant	0.1		0.589421	0.568421	0.394972
3	20	tanh	0.001	sgd	constant	0.1		0.728767	0.711636	0.450838
4	1	identity	0.01	sgd	constant	0.1		0.62238	0.622127	0.439665
6	30	identity	0.01	adam	constant	0.01		0.507853	0.522132	0.415084
7	50	tanh	0.01	sgd	constant	0.1		0.709589	0.718479	0.518994
8	100	tanh	0.01	sgd	constant	0.1		0.704539	0.71827	0.534637
LGBM	Number of variables	max_depth	N_estimations	num_leaves	subsample	boosting_type	learning_rate	Train	Test	OTT
1	10	2	20	2	1	gbdt	0.1	0.635	0.635	0.391
2	10	2	50	2	1	gbdt	0.1	0.642	0.624	0.417
3	10	4	100	4	1	gbdt	0.03	0.743	0.731	0.541
4	10	4	200	5	1	gbdt	0.03	0.797	0.75	0.549
5	10	5	300	7	0.8	GOSS	0.01	0.798	0.765	0.486
6	10	5	500	8	0.8	GOSS	0.01	0.834	0.779	0.457
7	10	6	800	8	0.8	GOSS	0.01	0.858	0.811	0.414
8	10	6	1000	10	0.8	GOSS	0.01	0.892	0.818	0.393



6. Final Model Performance

During our model exploration we noticed that the random forest model performed the best in terms of oot and time taken to run.

TRAIN

TRAIN	# Records	# Goods	# Bads	Fraud Rate								
	10116	9584	532	0.053								
	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Good	Cumulative Bads	% Goods	% Bads FDR	KS	FPR
1	506	173	333	34.1897233	65.8103	506	173	333	0.345744	61.32597	60.98022	0.51952
2	506	428	78	84.5849802	15.415	1012	601	411	1.201111	75.69061	74.4895	1.462287
3	505	475	30	94.0594059	5.94059	1517	1076	441	2.150409	81.21547	79.06506	2.439909
4	506	475	31	93.8735178	6.12648	2023	1551	472	3.099706	86.92449	83.82479	3.286017
5	506	486	20	96.0474308	3.95257	2529	2037	492	4.070987	90.60773	86.53675	4.140244
6	506	500	6	98.8142292	1.18577	3035	2537	498	5.070248	91.71271	86.64246	5.094378
7	506	499	7	98.6166008	1.3834	3541	3036	505	6.06751	93.00184	86.93433	6.011881
8	505	502	3	99.4059406	0.59406	4046	3538	508	7.070768	93.55433	86.48356	6.964567
9	506	505	1	99.8023715	0.19763	4552	4043	509	8.080021	93.73849	85.65847	7.943026
10	506	504	2	99.6047431	0.39526	5058	4547	511	9.087275	94.10681	85.01954	8.898239
11	506	502	4	99.2094862	0.79051	5564	5049	515	10.09053	94.84346	84.75293	9.803883
12	506	504	2	99.6047431	0.39526	6070	5553	517	11.09779	95.21179	84.114	10.74081
13	505	500	5	99.009901	0.9901	6575	6053	522	12.09705	96.1326	84.03555	11.59579
14	506	505	1	99.8023715	0.19763	7081	6558	523	13.1063	96.31676	83.21046	12.5392
15	506	504	2	99.6047431	0.39526	7587	7062	525	14.11356	96.68508	82.57153	13.45143
16	506	505	1	99.8023715	0.19763	8093	7567	526	15.12281	96.86924	81.74644	14.38593
17	506	503	3	99.4071146	0.59289	8599	8070	529	16.12807	97.42173	81.29367	15.2552
18	505	504	1	99.8019802	0.19802	9104	8574	530	17.13532	97.60589	80.47057	16.17736
19	506	506	0	100	0	9610	9080	530	18.14657	97.60589	79.45932	17.13208
20	506	504	2	99.6047431	0.39526	10116	9584	532	19.15383	97.97422	78.82039	18.01504

TEST

TEST	# Records	# Goods	# Bads	Fraud Rate								
	6744	6437	307	0.046								
	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Good	Cumulative Bads	% Goods	% Bads FDR	KS	FPR
1	337	143	194	42.4332344	57.5668	337	143	194	0.428362	57.56677	57.1384	0.737113
2	337	298	39	88.4272997	11.5727	674	441	233	1.321032	69.13947	67.81843	1.892704
3	338	320	18	94.6745562	5.32544	1012	761	251	2.279603	74.48071	72.20111	3.031873
4	337	320	17	94.9554896	5.04451	1349	1081	268	3.238175	79.52522	76.28705	4.033582
5	337	329	8	97.6261128	2.37389	1686	1410	276	4.223707	81.89911	77.6754	5.108696
6	337	336	1	99.7032641	0.29674	2023	1746	277	5.230207	82.19585	76.96564	6.303249
7	337	330	7	97.9228487	2.07715	2360	2076	284	6.218734	84.273	78.05426	7.309859
8	338	336	2	99.408284	0.59172	2698	2412	286	7.225234	84.86647	77.64123	8.433566
9	337	336	1	99.7032641	0.29674	3035	2748	287	8.231735	85.1632	76.93147	9.574913
10	337	336	1	99.7032641	0.29674	3372	3084	288	9.238235	85.45994	76.22171	10.70833
11	337	334	3	99.1097923	0.89021	3709	3418	291	10.23874	86.35015	76.1114	11.7457
12	337	334	3	99.1097923	0.89021	4046	3752	294	11.23925	87.24036	76.0011	12.7619
13	338	335	3	99.112426	0.88757	4384	4087	297	12.24276	88.13056	75.88781	13.76094
14	337	333	4	98.8130564	1.18694	4721	4420	301	13.24027	89.31751	76.07724	14.68439
15	337	336	1	99.7032641	0.29674	5058	4756	302	14.24677	89.61424	75.36747	15.74834
16	337	337	0	100	0	5395	5093	302	15.25627	89.61424	74.35798	16.86424
17	337	336	1	99.7032641	0.29674	5732	5429	303	16.26277	89.91098	73.64821	17.91749
18	338	336	2	99.408284	0.59172	6070	5765	305	17.26927	90.50445	73.23518	18.90164
19	337	335	2	99.4065282	0.59347	6407	6100	307	18.27277	91.09792	72.82515	19.86971
20	337	337	0	100	0	6744	6437	307	19.28227	91.09792	71.81565	20.96743

OOT

OOT	# Records	# Goods	# Bads	Fraud Rate								
	2419	2291	128	0.053								
	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Good	Cumulative Bads	% Goods	% Bads FDR	KS	FPR
1	121	71	50	58.677686	41.3223	121	71	50	0.595738	27.93296	27.33722	1.420
2	121	76	45	62.8099174	37.1901	242	147	95	1.233428	53.07263	51.8392	1.547
3	121	114	7	94.214876	5.78512	363	261	102	2.189965	56.98324	54.79328	2.559
4	121	120	1	99.1735537	0.82645	484	381	103	3.196845	57.5419	54.34505	3.699
5	121	119	2	98.3471074	1.65289	605	500	105	4.195335	58.65922	54.46388	4.762
6	121	121	0	100	0	726	621	105	5.210606	58.65922	53.44861	5.914
7	121	120	1	99.1735537	0.82645	847	741	106	6.217486	59.21788	53.00039	6.991
8	121	120	1	99.1735537	0.82645	968	861	107	7.224367	59.77654	52.55217	8.047
9	121	114	7	94.214876	5.78512	1089	975	114	8.180903	63.68715	55.50625	8.553
10	121	121	0	100	0	1210	1096	114	9.196174	63.68715	54.49098	9.614
11	121	120	1	99.1735537	0.82645	1331	1216	115	10.20305	64.24581	54.04276	10.574
12	121	119	2	98.3471074	1.65289	1452	1335	117	11.20154	65.36313	54.16158	11.410
13	121	120	1	99.1735537	0.82645	1573	1455	118	12.20842	65.92179	53.71336	12.331
14	121	120	1	99.1735537	0.82645	1694	1575	119	13.2153	66.48045	53.26514	13.235
15	121	121	0	100	0	1815	1696	119	14.23058	66.48045	52.24987	14.252
16	121	120	1	99.1735537	0.82645	1936	1816	120	15.23746	67.03911	51.80165	15.133
17	120	115	5	95.8333333	4.16667	2056	1931	125	16.20238	69.8324	53.63002	15.448
18	121	120	1	99.1735537	0.82645	2177	2051	126	17.20926	70.39106	53.1818	16.278
19	121	119	2	98.3471074	1.65289	2298	2170	128	18.20775	71.50838	53.30063	16.953
20	121	121	0	100	0	2419	2291	128	19.22302	71.50838	52.28536	17.898

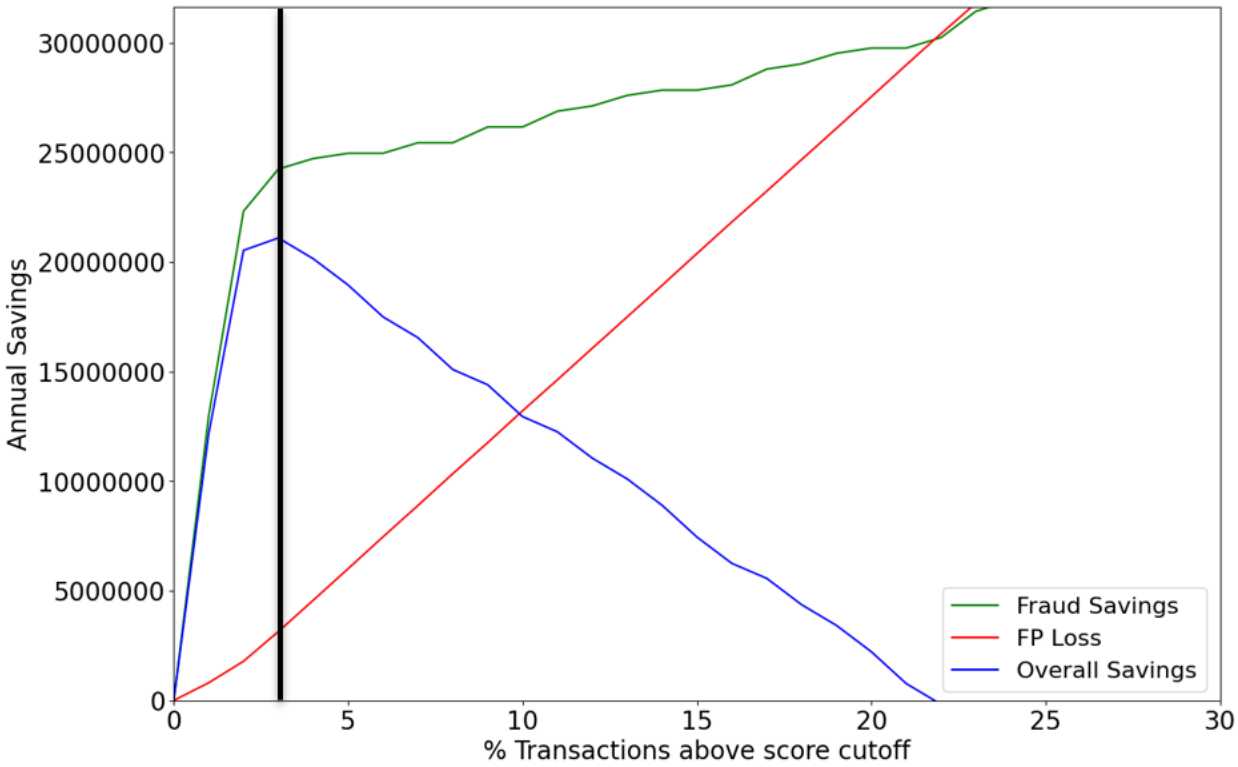
Summary of the model

The final model we choose from exploration is Random Forest with the parameters n_estimators=35, max_depth=10, min_samples_split=45,min_samples_leaf=30,max_features=5. With this we get the results as -

Train – 0.820
Test – 0.760
OOT – 0.562

We see that we can achieve fraud detection rate at 3% with 56.20%. This means that our model is able to capture 56.20% of all fraud at 3%. Our model rejects 3% of application to get 56.20% accuracy. Model rejects 3% of the application and capture 56.20% of the fraudulent transaction.

7. Financial Curves



The recommended score cutoff is around 3%. The estimated savings using the model is around $21,000,000. In the above graph, the blue line represents the overall savings, the green line represents the Fraud amount caught, and the red line represents the lost revenue.

From the business fraud manager, we need 2 key numbers:
1. Estimated $ lost for every fraud transaction missed ($400).
2. Estimated lost sales from every false positive, a good transaction that we denied ($20).

The Overall Savings will be the difference between these two values. We use the out-of-time model performance to plot these three numbers at different score penetration percentiles.
We then look at the Overall Savings at all the possible score cutoff thresholds and select a conservative maximal location where the savings saturates and does not increase beyond a point.

In this case, the recommended cutoff is 3%, which gives us the estimated savings of $21,000,000.
This value comes from the graph and computation: Estimated $ lost ($24,000,000) - Estimated lost sales ($3,000,000)

8. Summary

The dataset contains credit card transaction details from a US government organization in 2010, with 96,753 records and 10 fields. The independent variables are transaction type, amount, merchant description, location, card number, and the record number. The dependent variable is the fraud column that represents whether a transaction is fraudulent or not.

Data cleaning was performed, which included - removing outliers (Amount > 3000000) and keeping only desired values (Transtype = 'P'). For fields with missing values, we preformed imputation by hardcoding values or leveraging other fields within the dataset, using specific logic. Predominantly, the field Merch state was paired with 'merch zip', 'merch num' and 'merch description' values using dictionaries and imputed with its mode value within each pair. Similarly, Merch num was paired with 'merch description', 'Merch zip' was paired with 'merch num' and 'merch zip' values to impute mode-values of these fields in missing cells. Values that could no longer be imputed were labelled as "unknown".

For feature engineering, the aim is to generate a broad range of variables that can potentially influence the predictions of fraudulent transactions. For instance, we produced Benford's law variables to assess the abnormality of the first digit distribution concerning 'Cardnum' and 'Merchnum', as well as the Risk variable, which determines the probability of fraud for each day of the week. We also created various variables for different entities and time periods, such as Velocity Change, Day since, Frequency, and Amount variables.
We also created 2 other categories of variables on our own - statistics per 'card num' to flag unusual transactions to flag transactions that are not peculiar to a particular card and Days_Since+Merchnum+Frequency+Amount variables to tell us if there have been a lot of transactions happening at a particular merchant in a short span of time.

Feature selection was performed using a filter and wrapper method. We explored feature selection with LGBM Forward Selection, Random Forest Forward Selection, LGBM Backward Selection, Cat Boost Forward Selection. The final model selected was the forward LGBM with num_filters = 300 and num_wrappers = 20. This model gave a performance of 70% and the complexity was not very high, due to which execution time was only about 1 hour. This model run selected a good mix of short-term, long-term variables and different entities with filter scores ranging from 0.447 to 0.654 with jumps in between, as expected.

Finally, we explored 6 models with different hyperparameter combinations as shown in the model exploration section of this document. The test and train performances were high, with most around 70% while the OOT performances were around 40-50%. We also tuned 4 models to attain overfitting by varying 1 hyperparameter in each model in uninform increments to understand their nature some more.

The final model selected by us that gave the best performance statistics was Random Forest with the below parameters: n_estimators=35, max_depth=10, min_samples_split=45, min_samples_leaf=30, max_features=5. With this, we got about 80% performance for test and train data and 55% for OOT data. Thus, we see that we can achieve fraud detection rate at 3% with 56.20%.

9. Appendix

1. Data Description

The dataset comprises of Credit Card Transaction data.
It has 96753 records and 10 fields – 'Recnum', 'Cardnum', 'Date', 'Merchnum', 'Merch description', 'Merch state', 'Merch zip', 'Transtype', 'Amount', 'Fraud'.
The data spans across the year 2010.

2. Summary Tables

Numerical Table

Field Name	% Populated	Min	Max	Mean	Stdev	% Zero
Date	100	2010-01-01	2010-12-31	N/A	N/A	0
Amount	100	0.01	3,102,045.53	427.88	10,006.14	0

Categorical Table

Field Name	% Populated	# Unique Values	Most Common Value
Recnum	100	96,573	N/A
Cardnum	100	1,645	5142148452
Merchnum	96.512	13,091	930090121224
Merch Description	100	13,126	GSA-FSS-ADV
Merch state	98.765	227	TN
Merch zip	95.188	4,567	38118
Transtype	100	4	P
Fraud	100	2	0

3. Visualization of Each Field

a. Field Name: Recnum

This is an ordinal unique positive integer which is incremental. This is assigned to each record entry from 1 to 96,753.

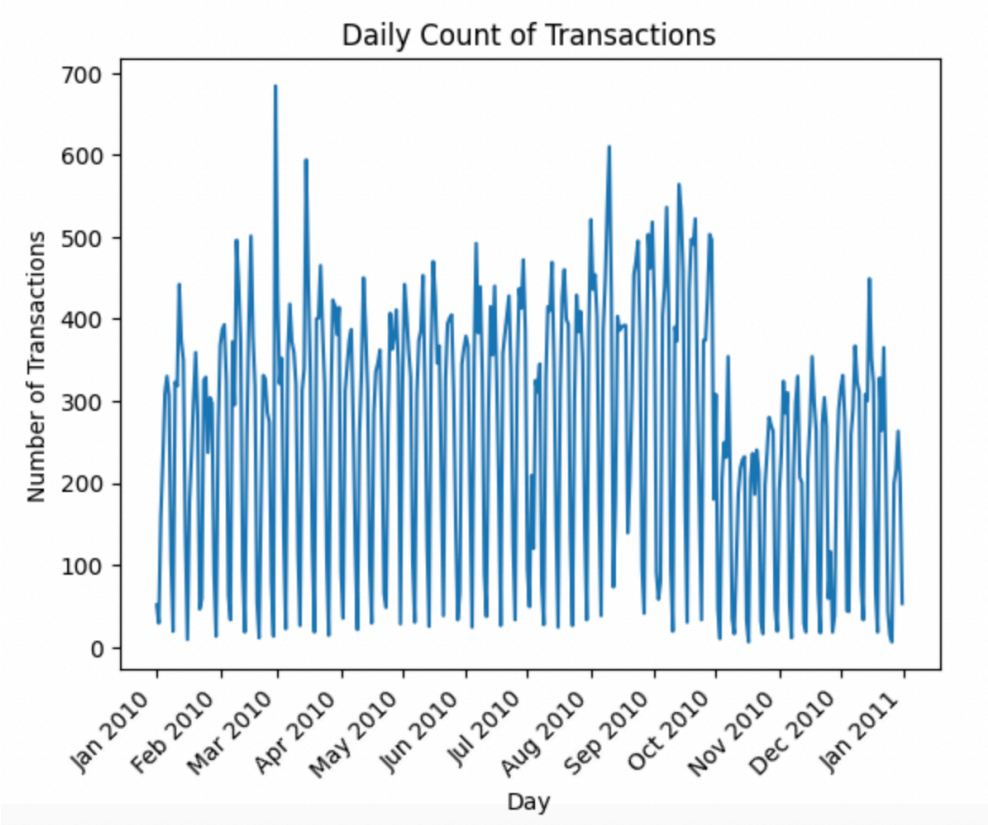
b. Field Name: Date

This is a numerical field which captures the date on which the transactions occurred. The transactions are captured between 2010-01-01 and

2010-12-31. The count of transactions has been captured on a daily, weekly, and monthly basis:

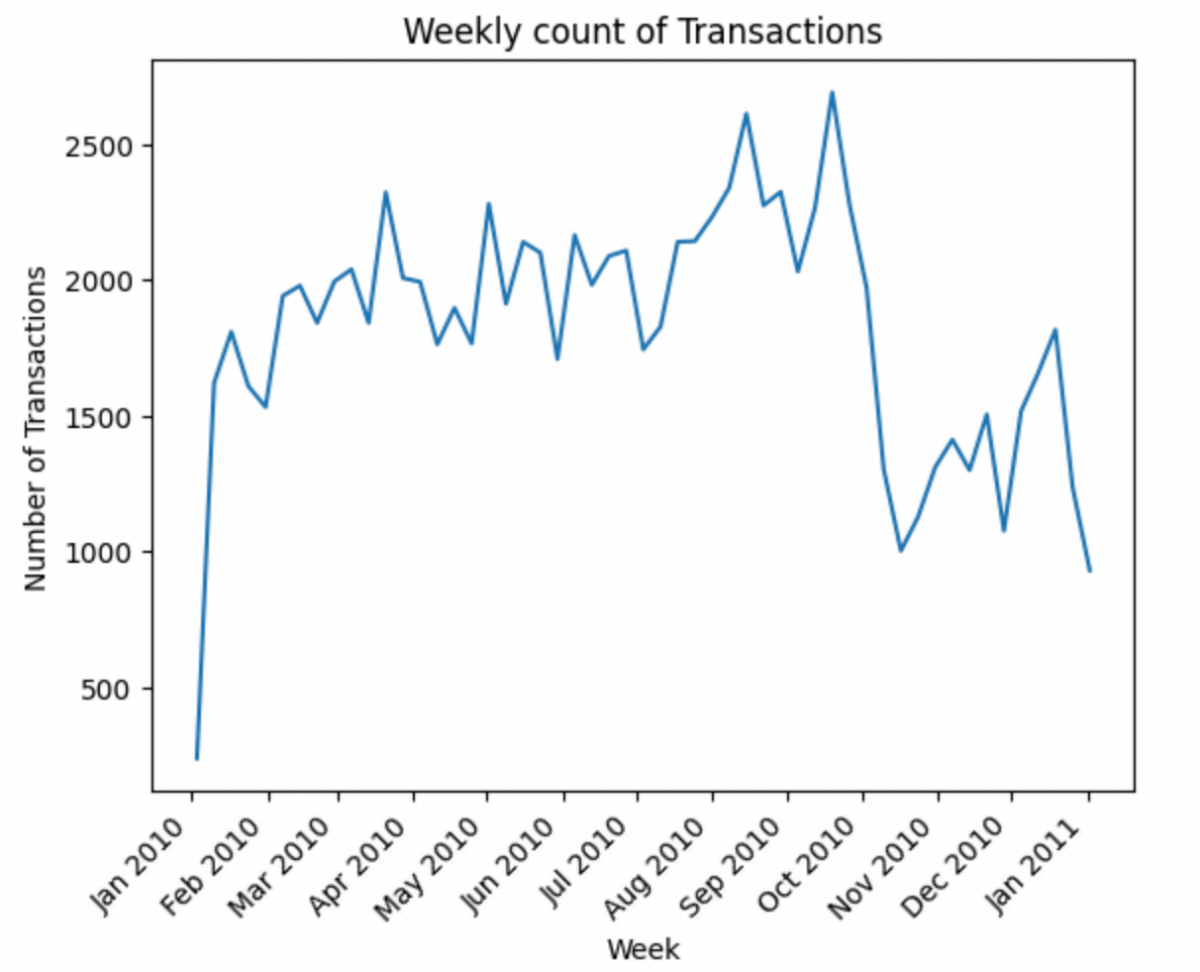
Daily Count of Transactions –

It is seen that on January 1st, 2010, there were 51 transactions and on December 31st, 2010, there were 53 transactions.
The highest number of transactions occurred on 28th February 2010, capturing 684 transactions on that day.



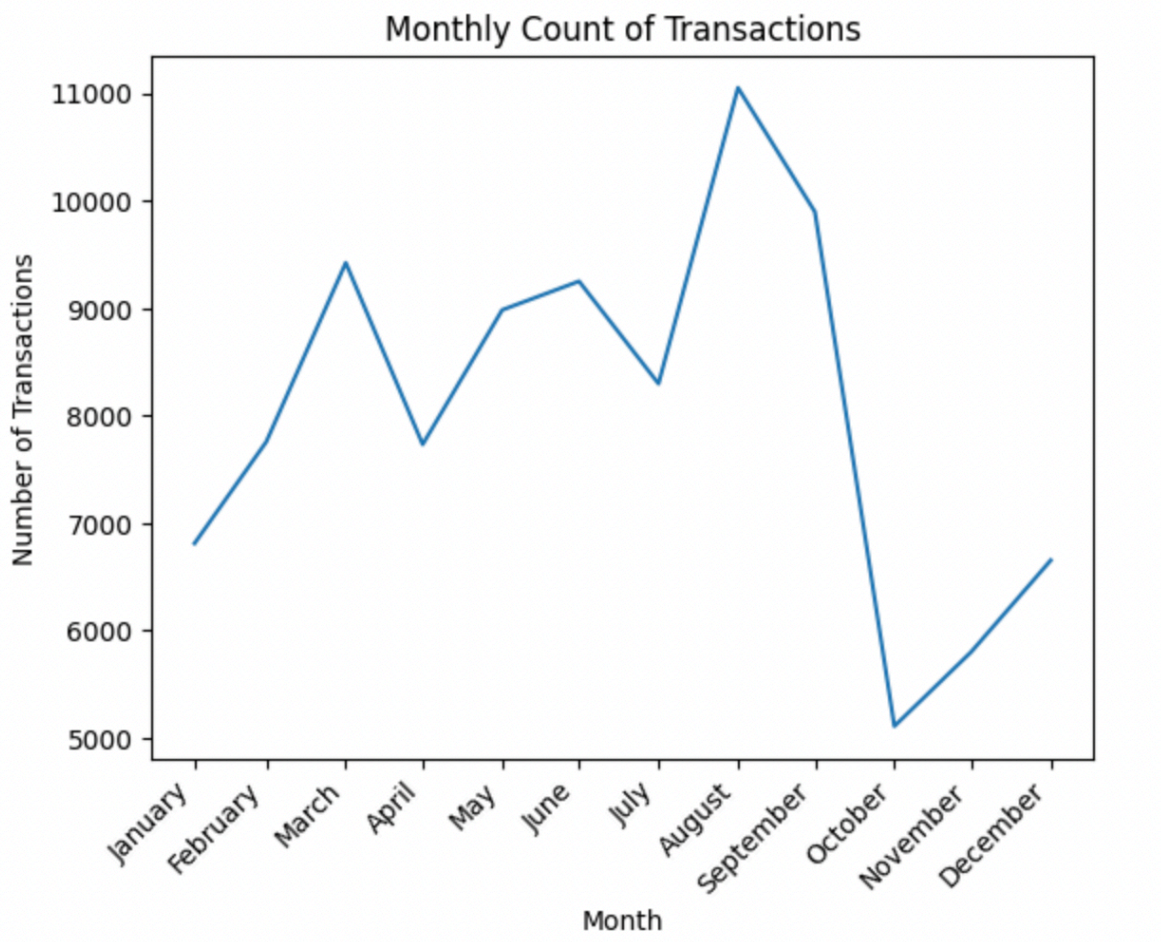
Weekly Count of Transactions –

Week 01(week ending January 3rd, 2010) captures 239 transactions and the last week (week ending February 2nd 2011) captures 931 transactions as seen below.



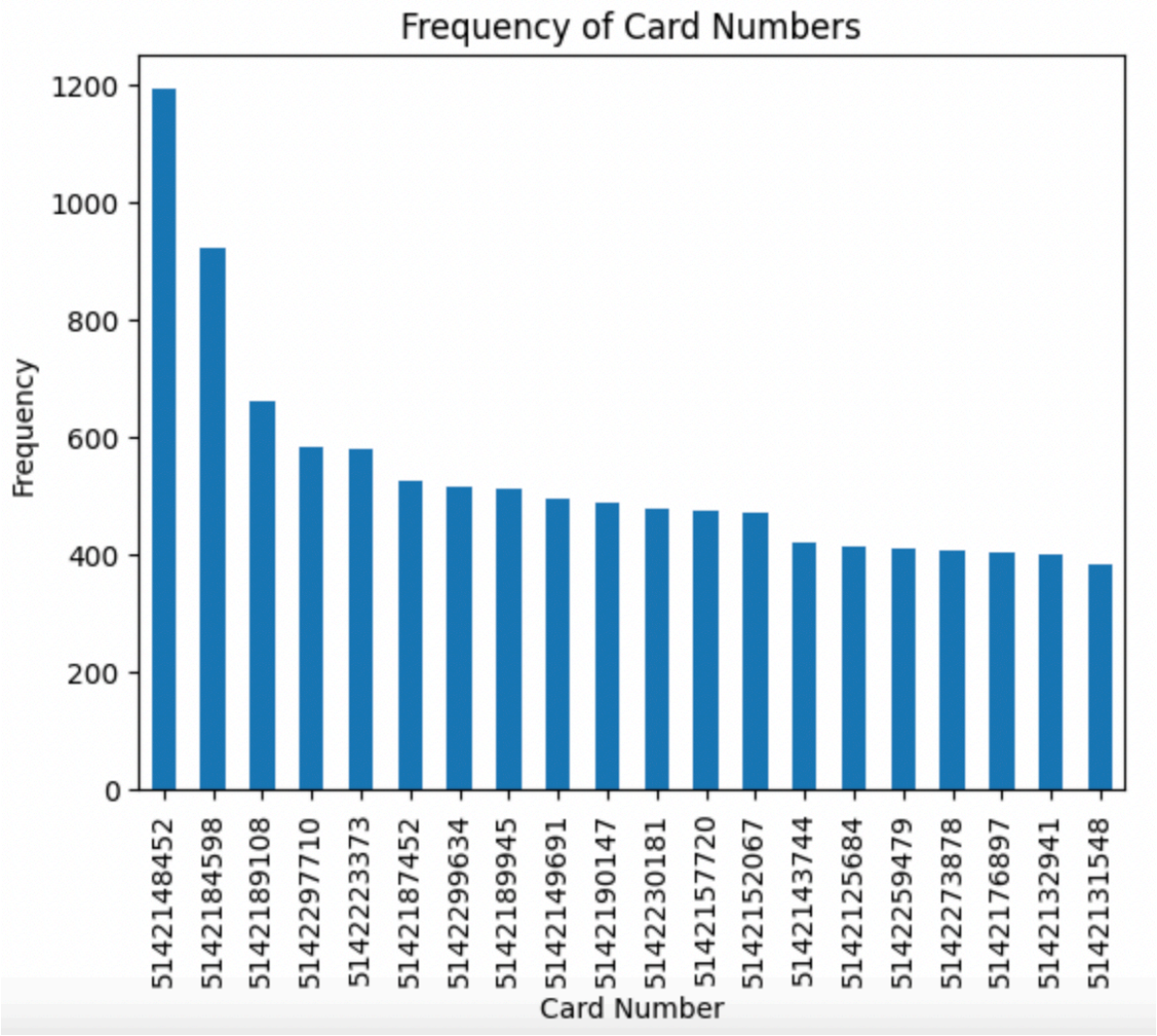
Monthly Count of Transactions –

August is the month where the highest number of transactions (11,050) have been captured. October is the month with the least number of transactions (5,109).



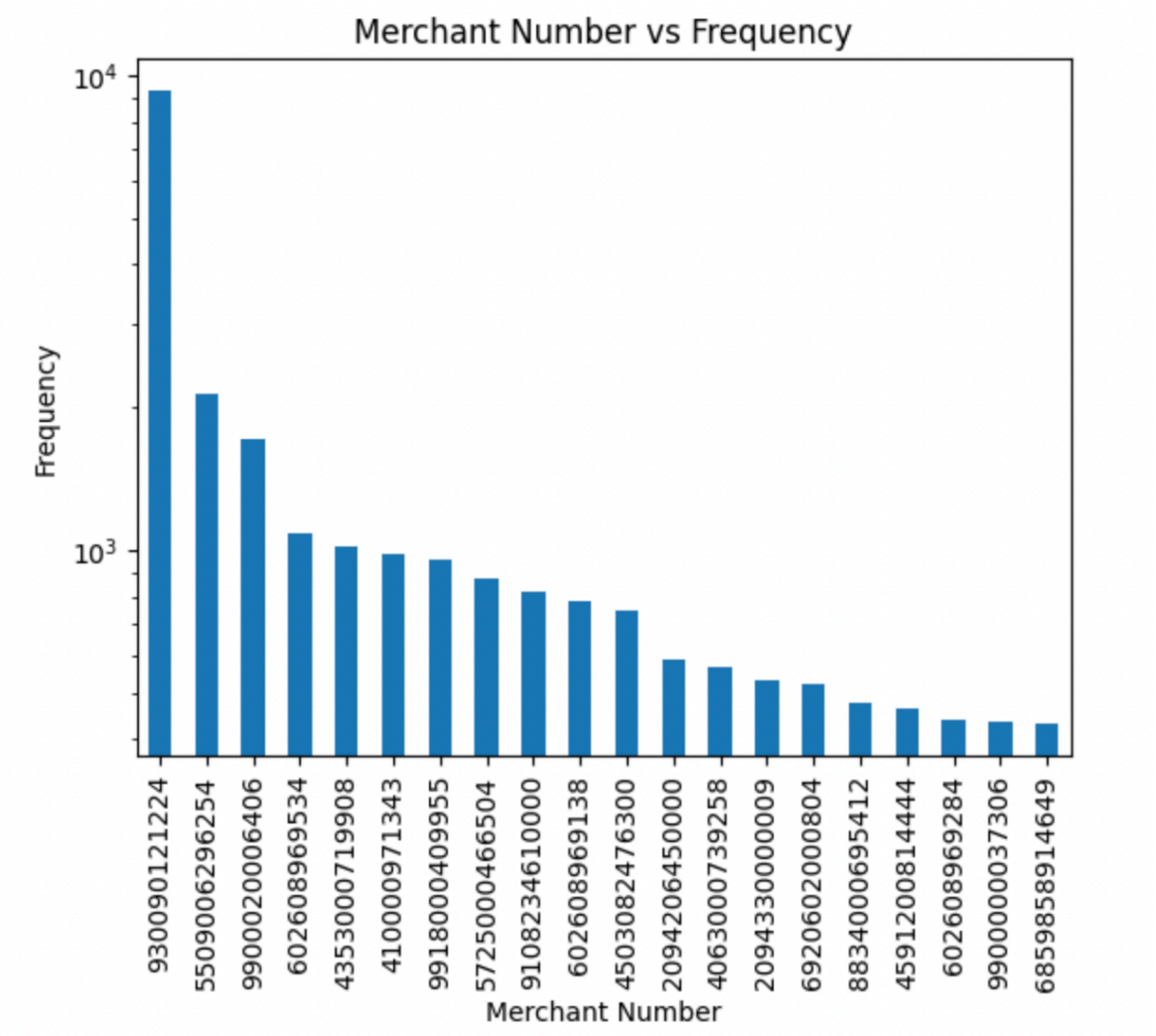
c. Field Name: Cardnum

This is a categorical field which displays the card number used to make each transaction. The card number 5142148452 was used 1192 number of times, making this card number the most used, in the dataset.



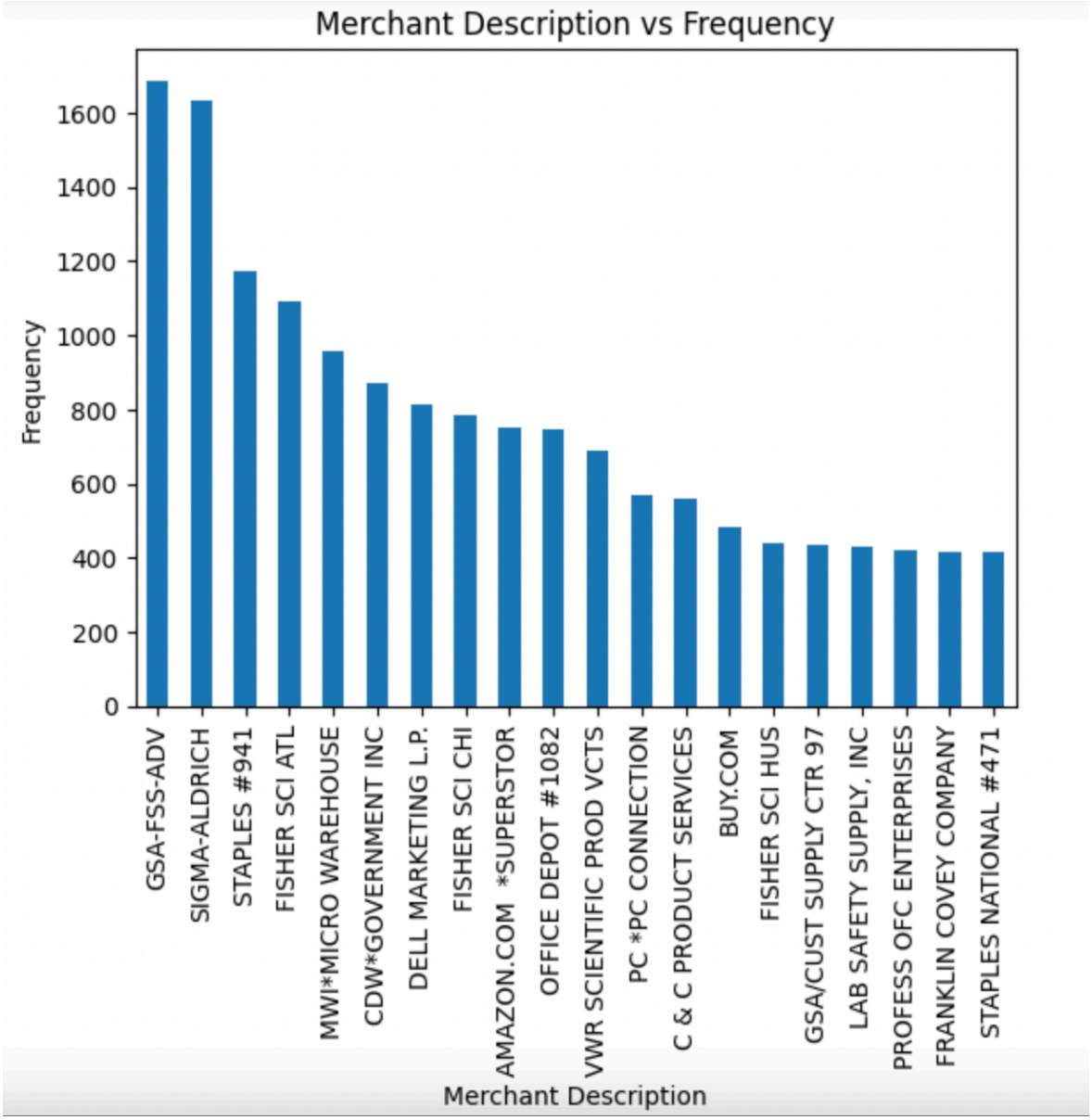
d. Field Name: Merchnum

A merchant number is a code that is issued by credit card processors that tells the banks and credit card processors where the money is going. 930090121224 is the merchnum with the highest number of transactions (9310).



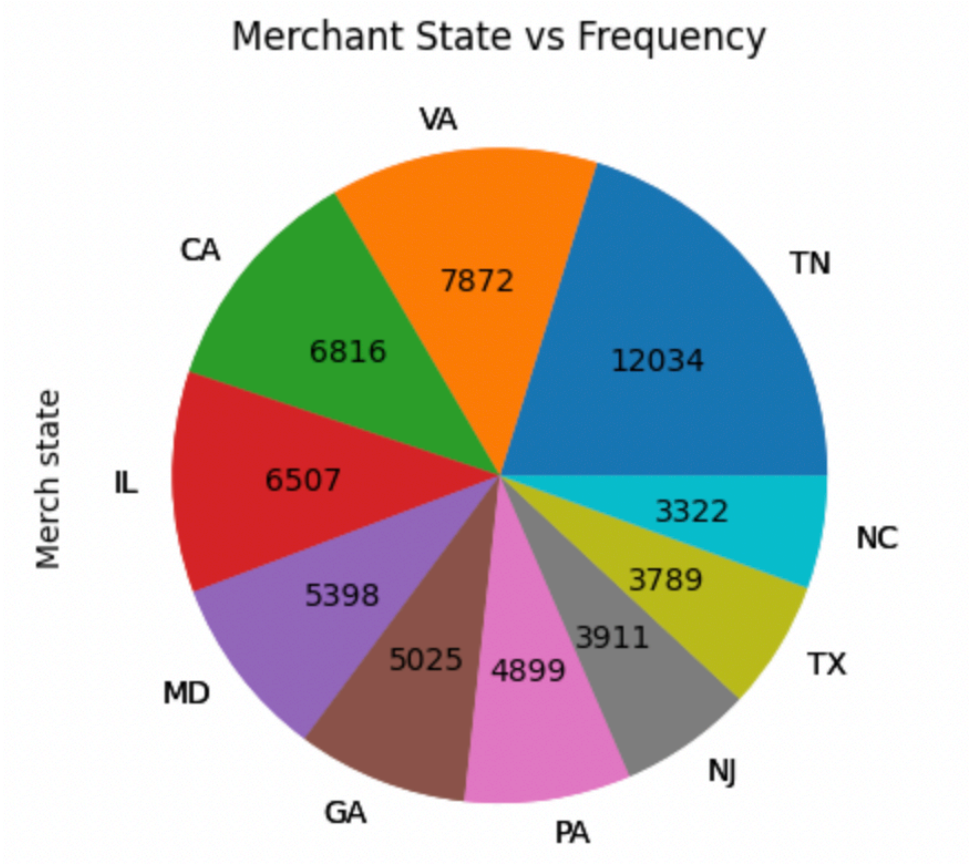
e. Field name: Merchant description

This field describes the merchants, with GSA-FSS-ADV being the merchant description with the highest occurrence (1688).



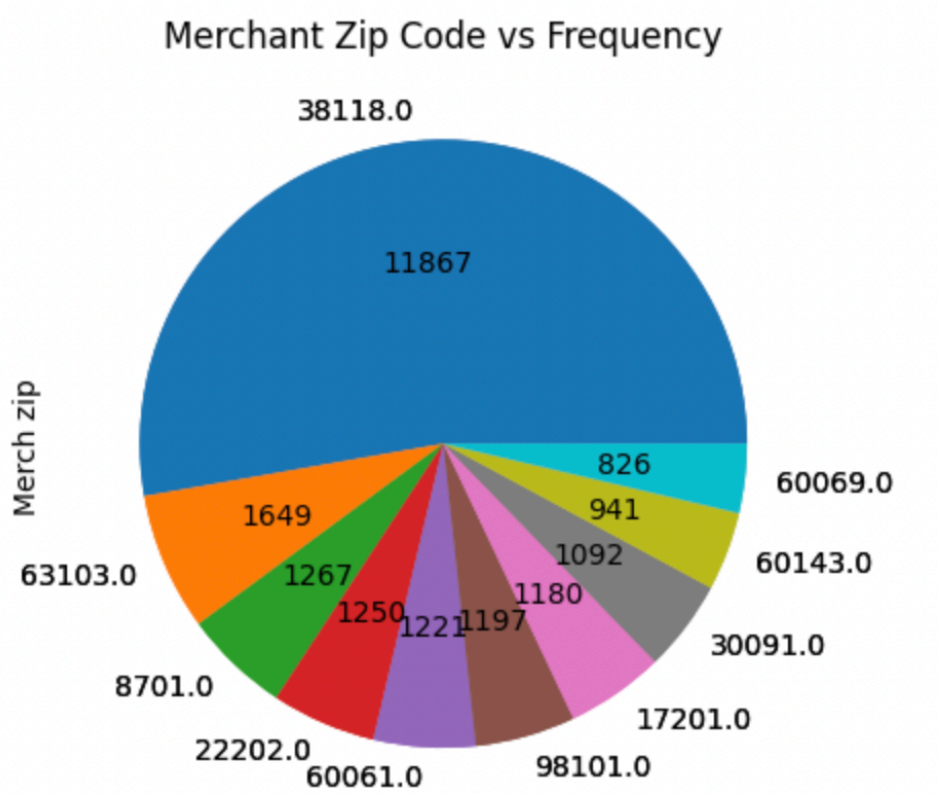
f. Field name: Merch state

Merchant state describes the state where the transactions take place. As seen below, most of the transactions originate in Tennessee (12,034).



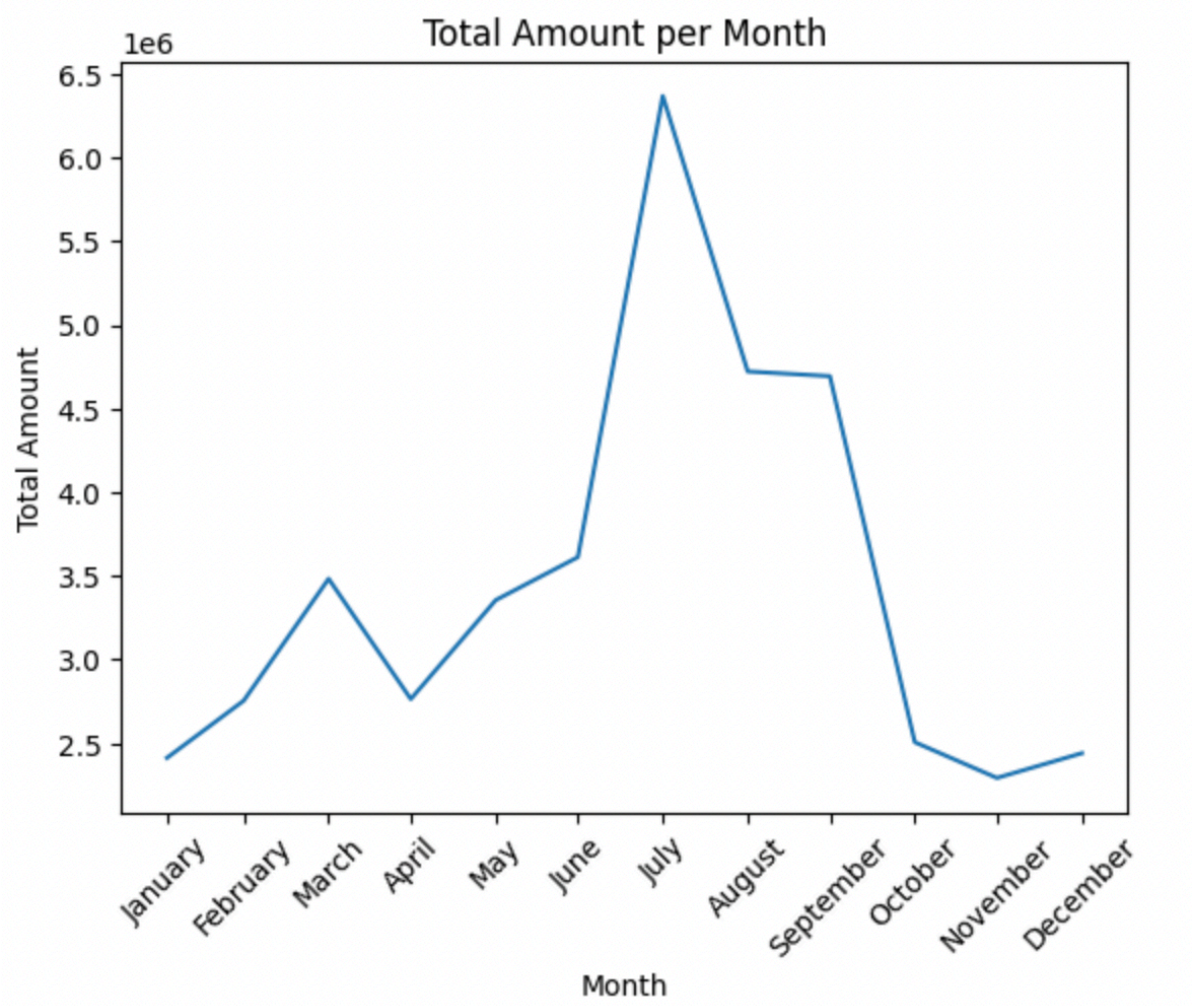
g. Field name: Merch zip

Captures the merchant zip codes. The zip code 38118 has the highest number of transactions (11867).

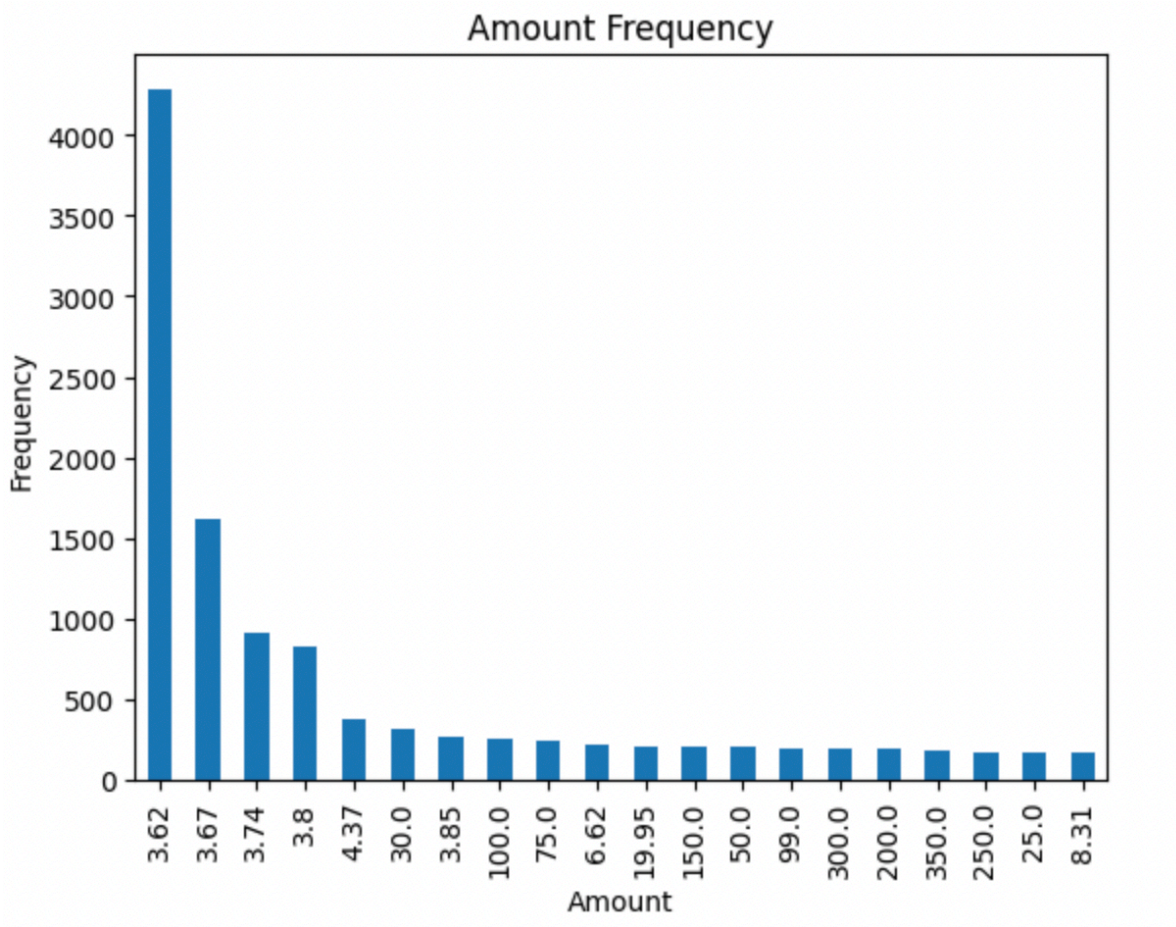


h. Field Name: Amount

This field displays the amount transacted in each of the transactions.
The below plot shows the total amount transacted in each month. July is the month where the largest amount has been transacted (63,67,486).

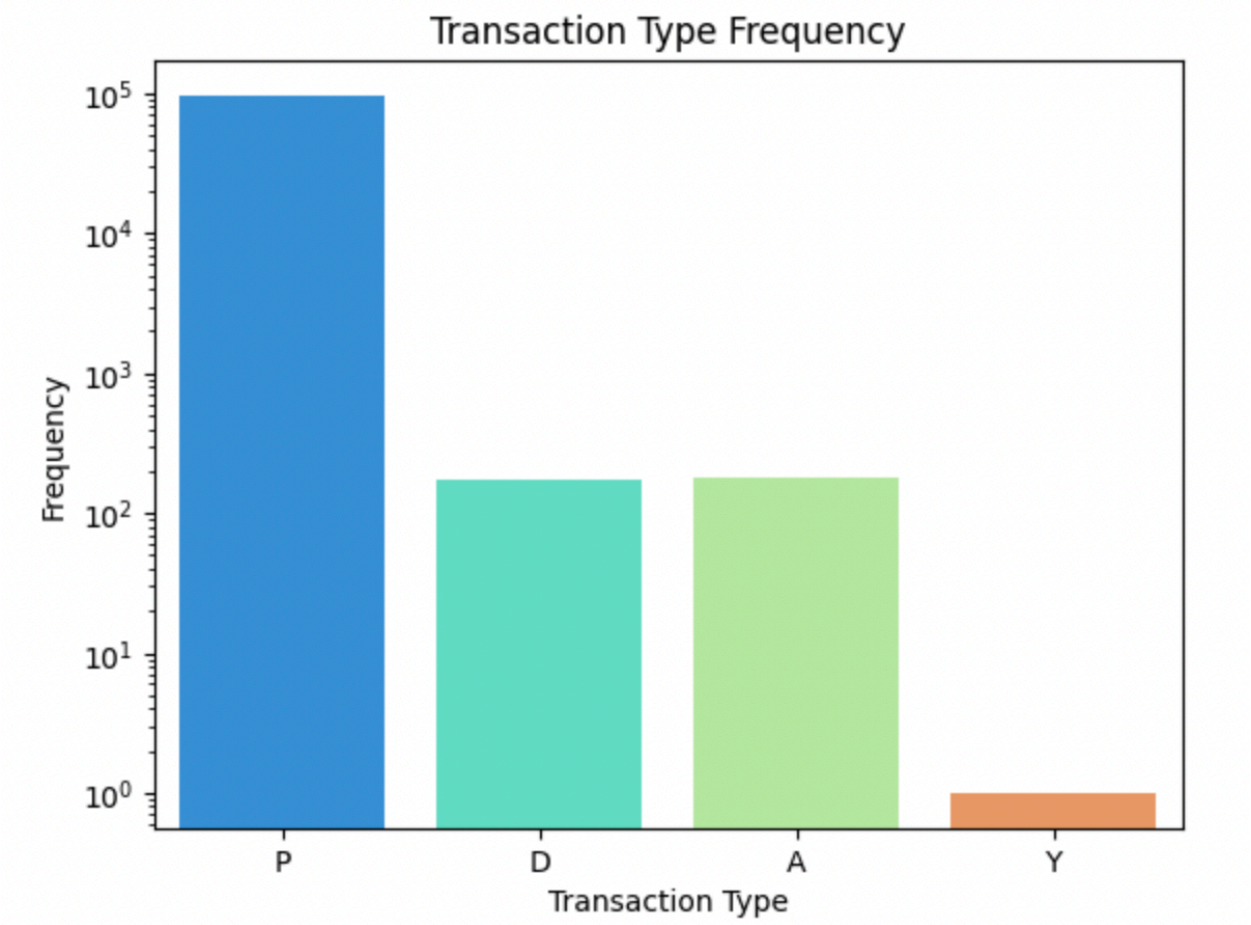


The below plot shows the most common dollar amounts that were transacted. The frequency is depicted in the log scale. The amount which is transacted the most is 3.62, which has occurred 4,283 times.



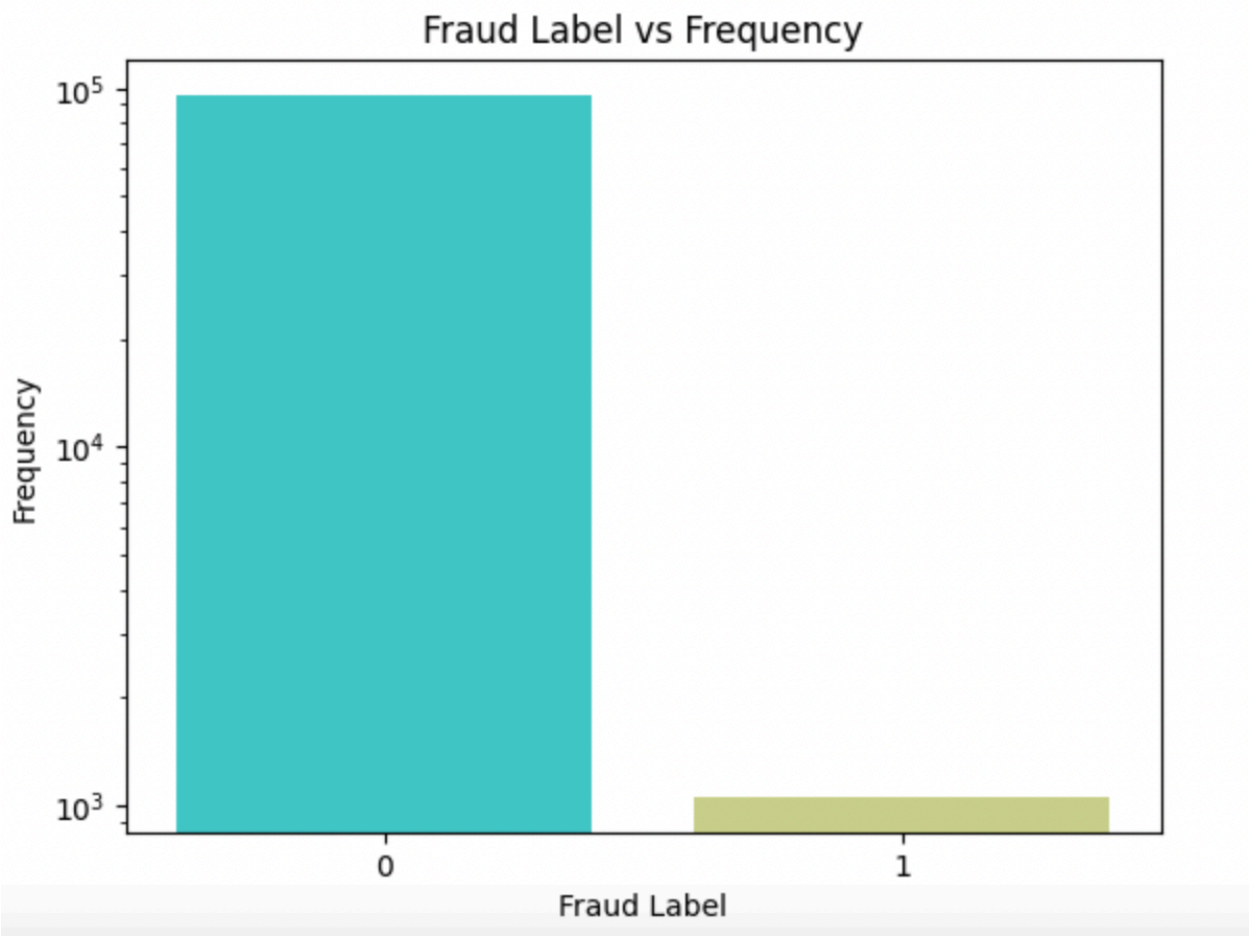
i. Transaction Type

Categorizes the transactions into 4 categories – P, D, A, Y. Most of the transactions fall under the "P" category.



j. Field Name: Fraud

This field depicts whether the transaction is categorized as “Fraud” or not. All the transactions with Fraud = 1 are Fraud transactions and the ones with Fraud = 0 are not fraud.



Fraud Label	Count	Percentage %
0	95,694	99%
1	1059	1%