Rubén Alejandro López Reynoso

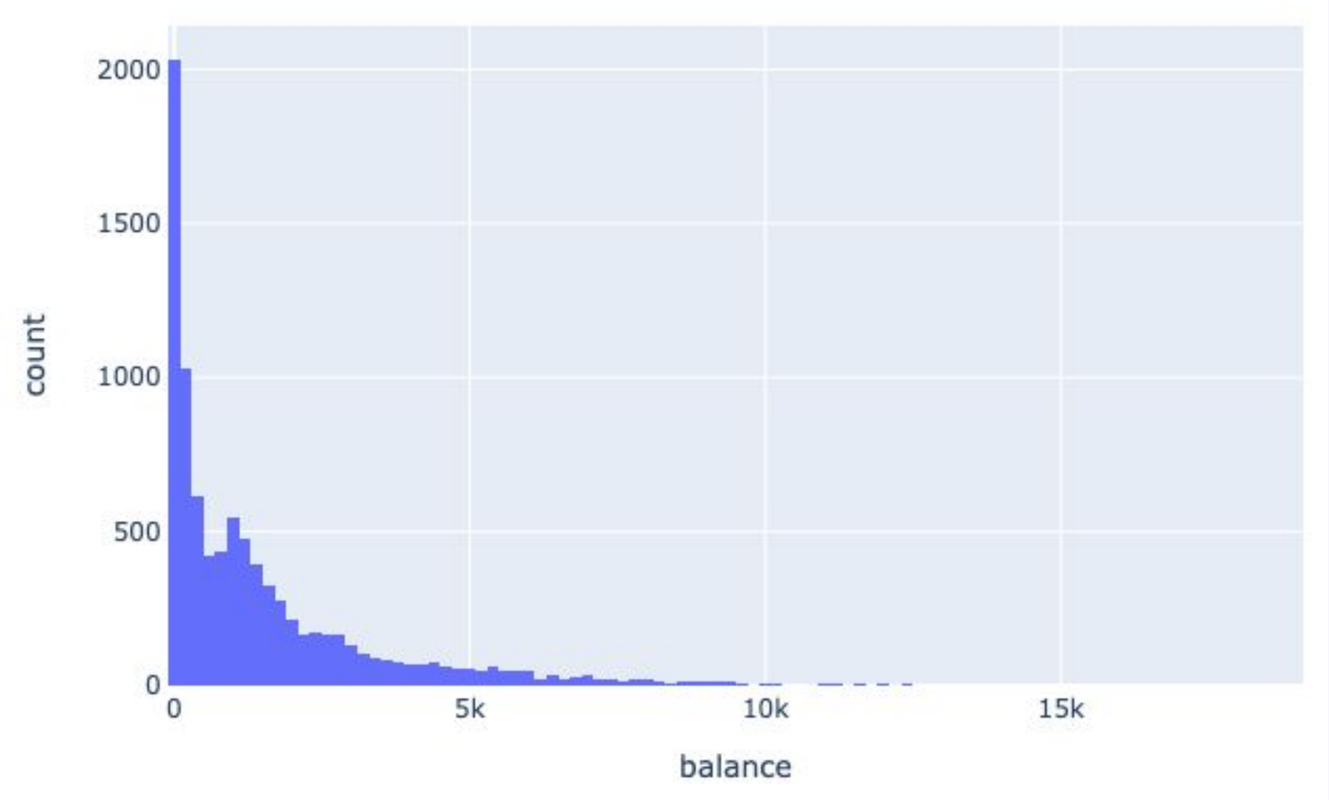
Data Analyst Jr. Stori Card

Table of contents

1. Question 1
2. Question 2
3. Question 3
4. Closing Slide

1. Question 1

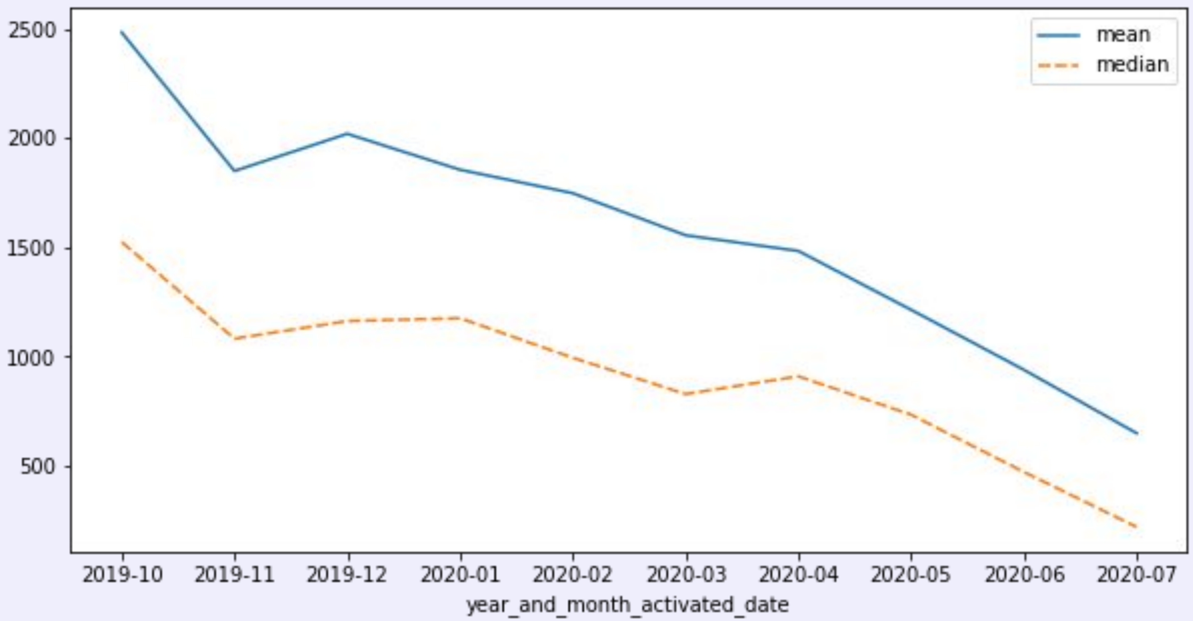
1.1 Balance Histogram

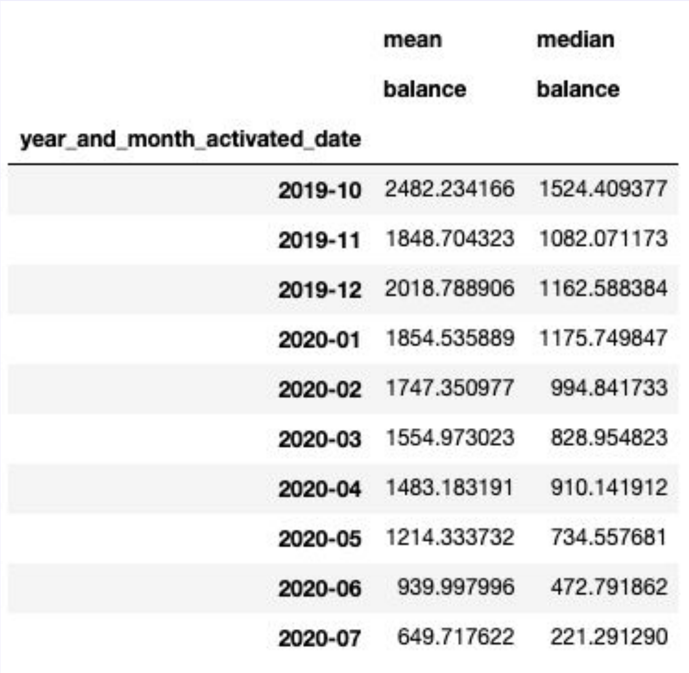


1.2 Insights;

1. Around 80% of the counts have less of 2K in balance
2. The population is skewed to the left
3. Attached you can find the plot in *histograma.html*

1.3 Mean and Median





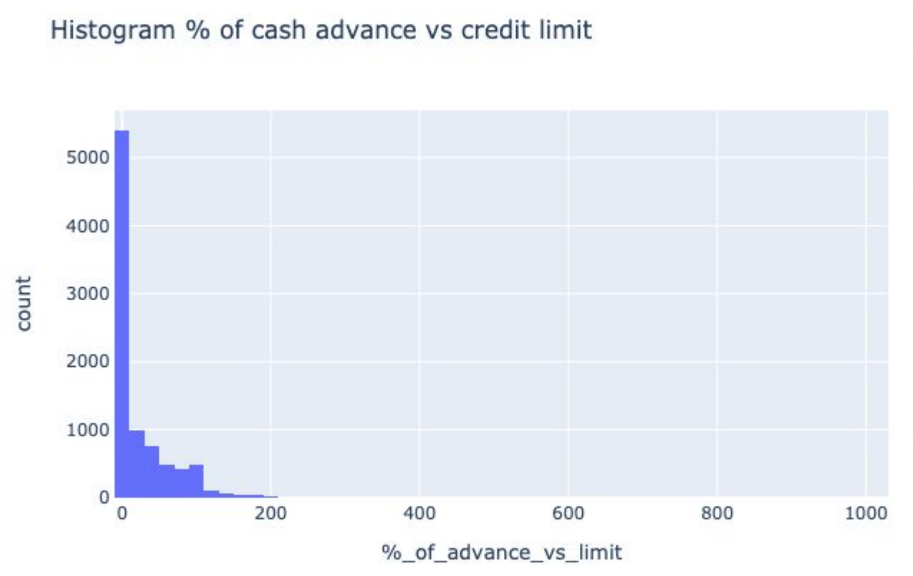
	mean balance	median balance
year_and_month_activated_date		
2019-10	2482.234166	1524.409377
2019-11	1848.704323	1082.071173
2019-12	2018.788906	1162.588384
2020-01	1854.535889	1175.749847
2020-02	1747.350977	994.841733
2020-03	1554.973023	828.954823
2020-04	1483.183191	910.141912
2020-05	1214.333732	734.557681
2020-06	939.997996	472.791862
2020-07	649.717622	221.291290

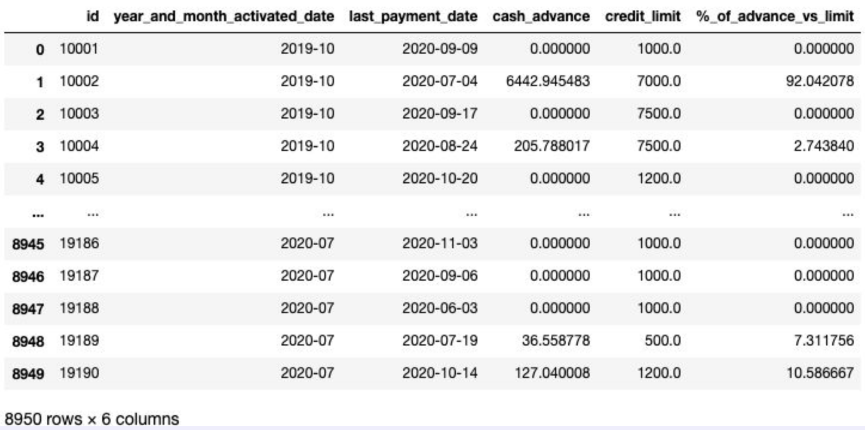
Insights;

1. Average and median balance were falling down every month

2. Question 2

2.1 Reported table





	id	year_and_month_activated_date	last_payment_date	cash_advance	credit_limit	%_of_advance_vs_limit
0	10001	2019-10	2020-09-09	0.000000	1000.0	0.000000
1	10002	2019-10	2020-07-04	6442.945483	7000.0	92.042078
2	10003	2019-10	2020-09-17	0.000000	7500.0	0.000000
3	10004	2019-10	2020-08-24	205.788017	7500.0	2.743840
4	10005	2019-10	2020-10-20	0.000000	1200.0	0.000000
...	...	...	...	...	...	...
8945	19186	2020-07	2020-11-03	0.000000	1000.0	0.000000
8946	19187	2020-07	2020-09-06	0.000000	1000.0	0.000000
8947	19188	2020-07	2020-06-03	0.000000	1000.0	0.000000
8948	19189	2020-07	2020-07-19	36.558778	500.0	7.311756
8949	19190	2020-07	2020-10-14	127.040008	1200.0	10.586667

8950 rows × 6 columns

2.1 Insights;

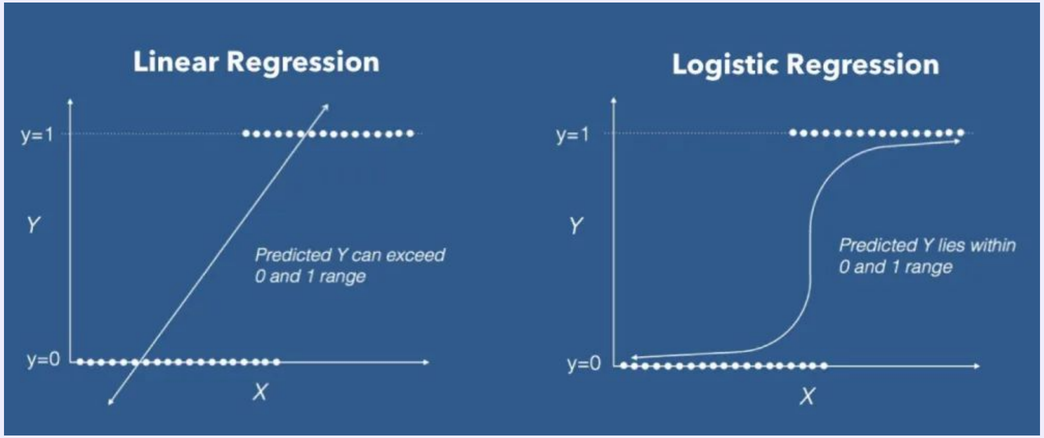
1. Around 80% of the clients have less of 80% of cash advance vs credit limit
2. The population is skewed to the left
3. Attached you can find the plot in *histograma_credito.html* and the information in *pregunta_2.csv)*

3. Question 3

3.1 Predictive Model for fraud

3.1 Insights;

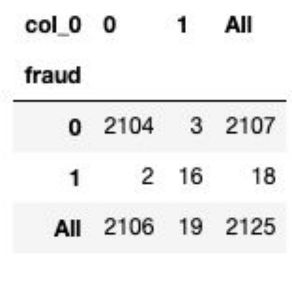
1. Fraud is a nominal variable, which means it only can take the value of True or false (1 or 0 respectively)
2. It was apply a machine learning method for prediction called Logistic regression, the details are in the *stori_card.ipynb* file.
3. The selected variables were the following: *'balance', 'balance_frequency', 'oneoff_purchases', 'installments_purchases', 'cash_advance', 'purchases_frequency', 'oneoff_purchases_frequency', 'purchases_installments_frequency', 'cash_advance_frequency', 'cash_advance_trx', 'purchases_trx', 'credit_limit', 'payments', 'prc_full_payment'*


Linear Regression
y=1
Y
y=0
X
Predicted Y can exceed 0 and 1 range
Logistic Regression
y=1
Y
y=0
X
Predicted Y lies within 0 and 1 range

3.1 Predictive Model for fraud

3.1 Evaluation of model

1. In order to Evaluate the model it was calculate a confusion matrix, in order to obtain the probability of false positives or false negatives:

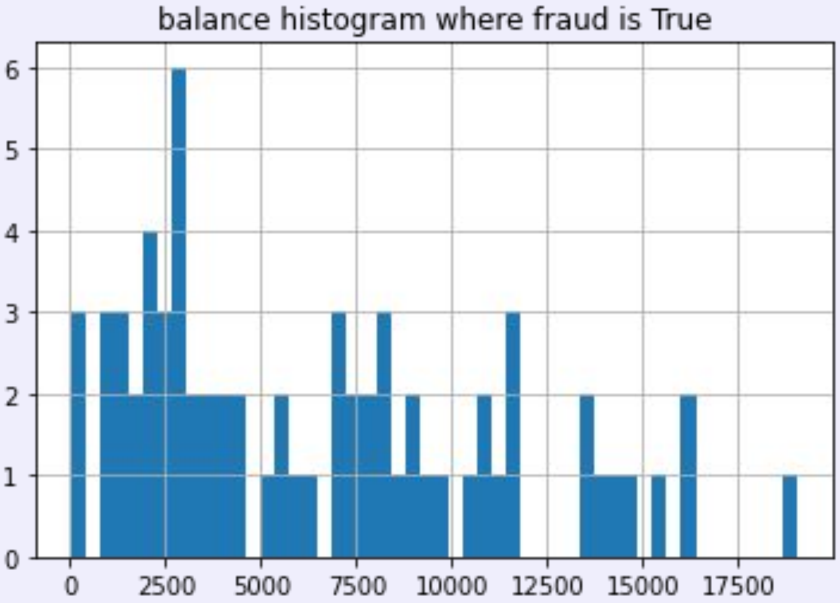


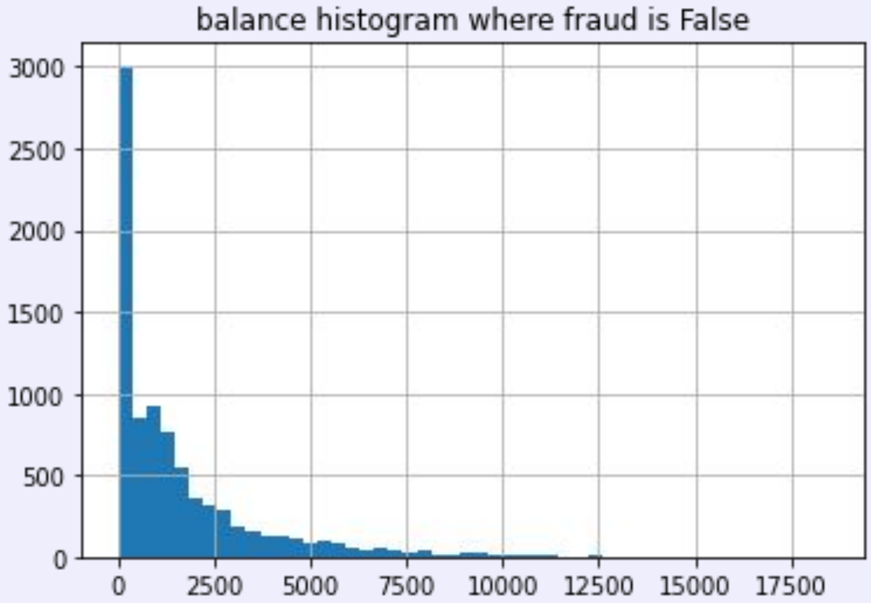
col_0	0	1	All
fraud			
0	2104	3	2107
1	2	16	18
All	2106	19	2125

3.1 Evaluation of the model

1. The probability of obtaining a false positive is 0.147% (3 / 2, 107 of the predictions)
2. The probability of obtaining a false negative es 11.11% (2/18 of the predictions)

3.2 Most predictive variable for fraud





3.2 Insights;

1. Where fraud is True most of the cases have a balance above 2500
2. Where Fraud is False the balance is under 2500

Thank you!