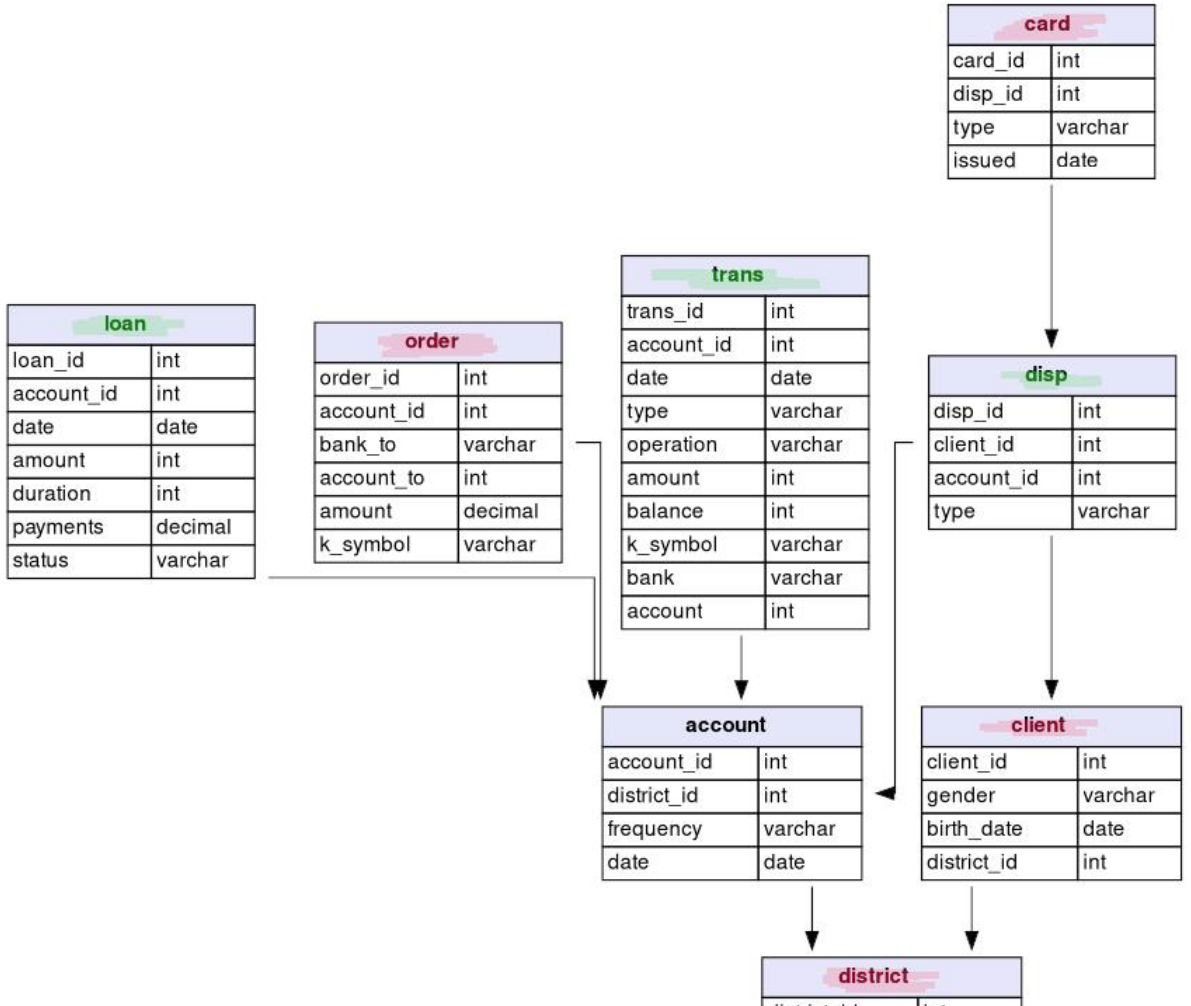
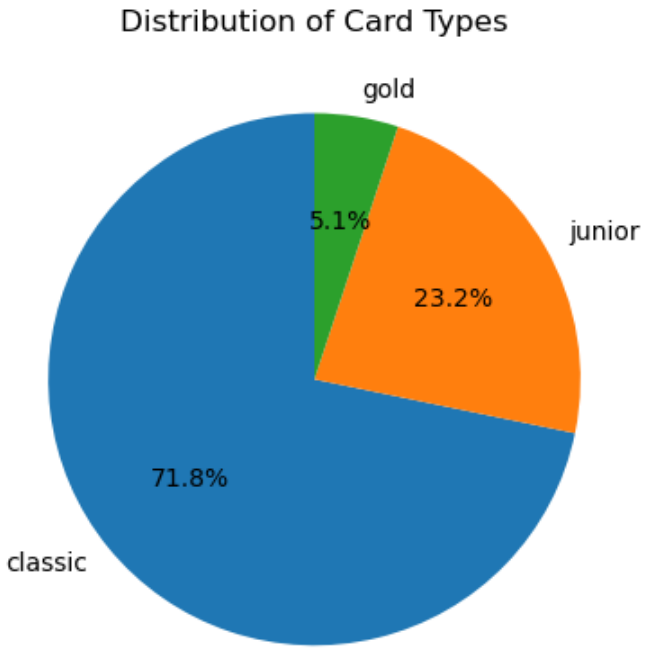
Should we loan?

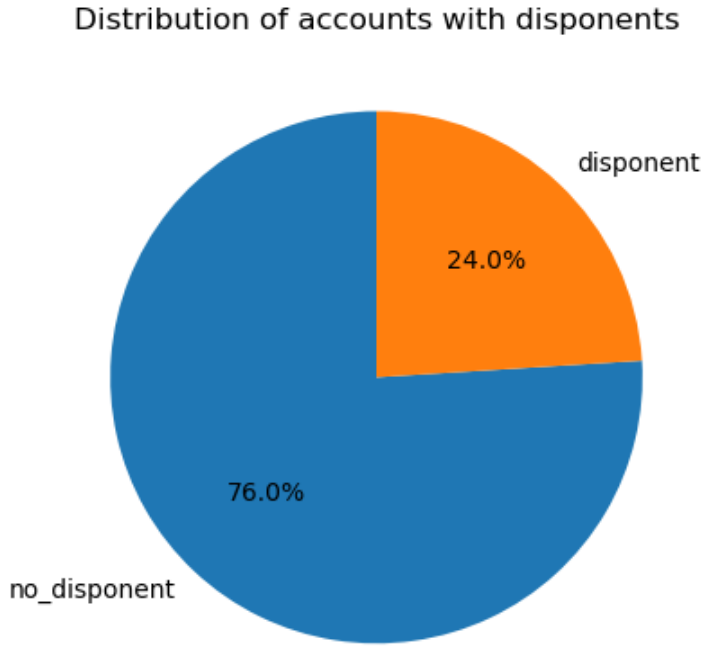
Project by Pedro Leite and Stefan Samfirescu

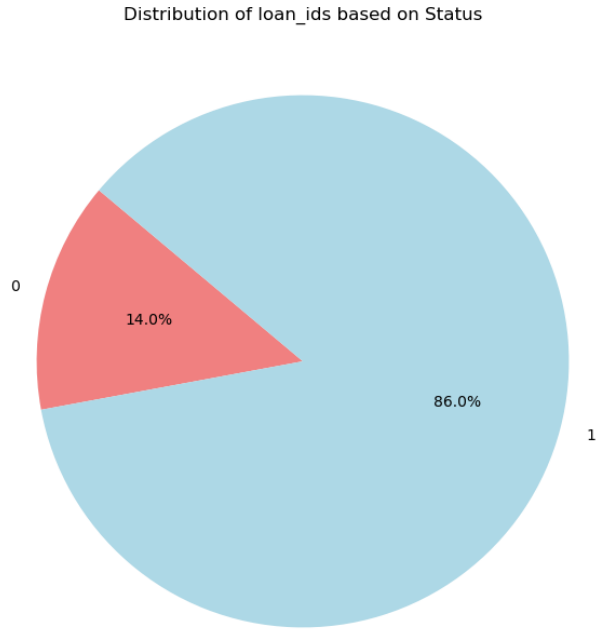
Data Understanding and Preparation



Data Understanding and Preparation

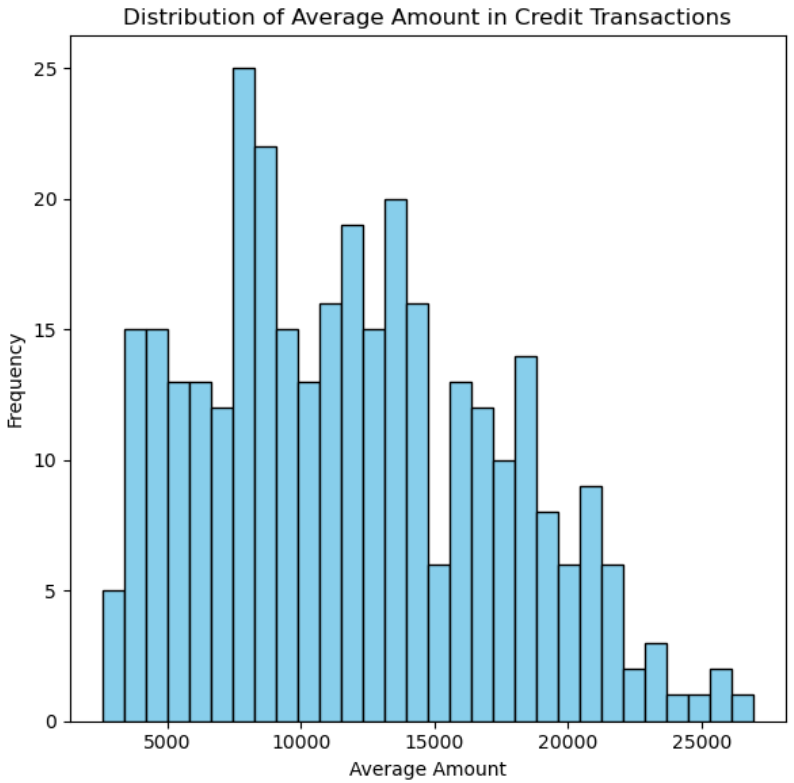


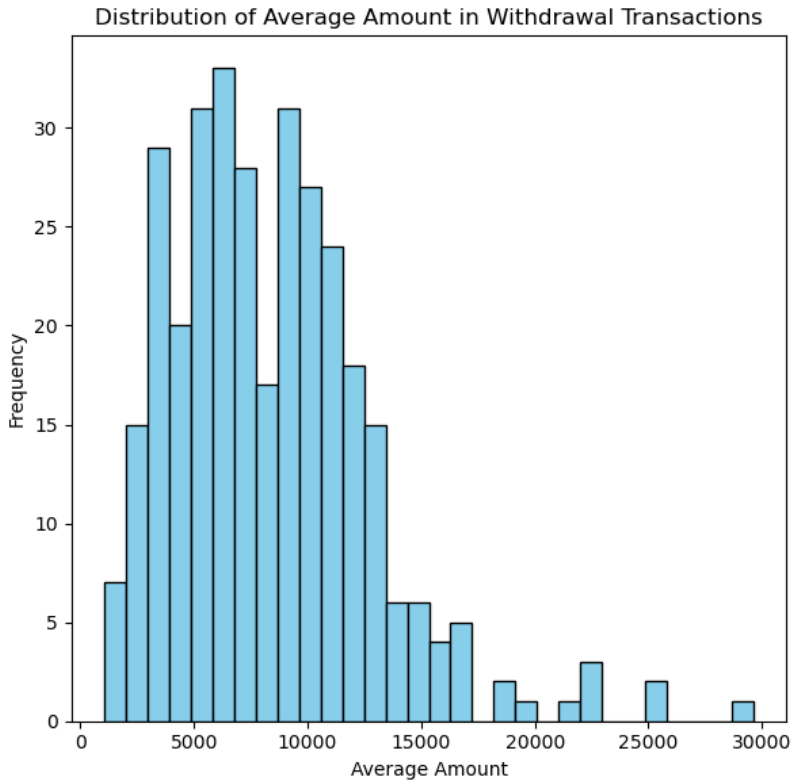


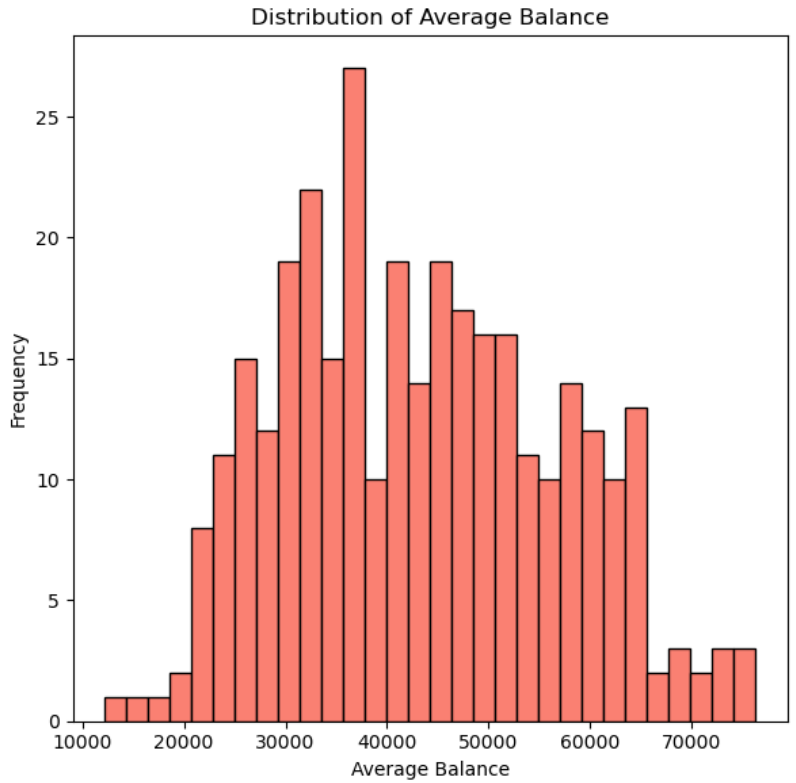


Only 328 unique loan_id's, data aggregation will result in only 328 samples to train on
We didn't include card types because we only found 11 accounts with card

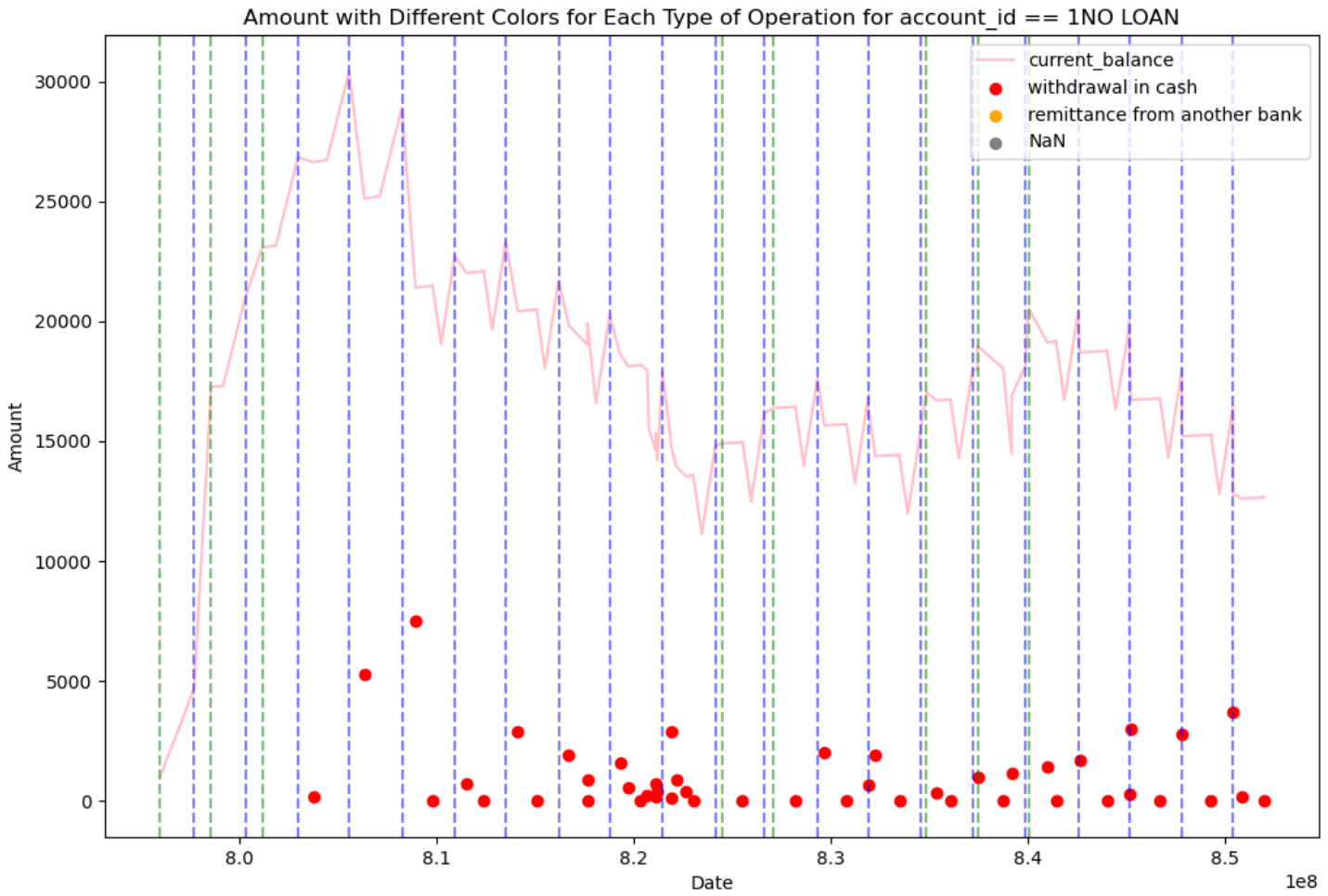
Data Understanding and Preparation



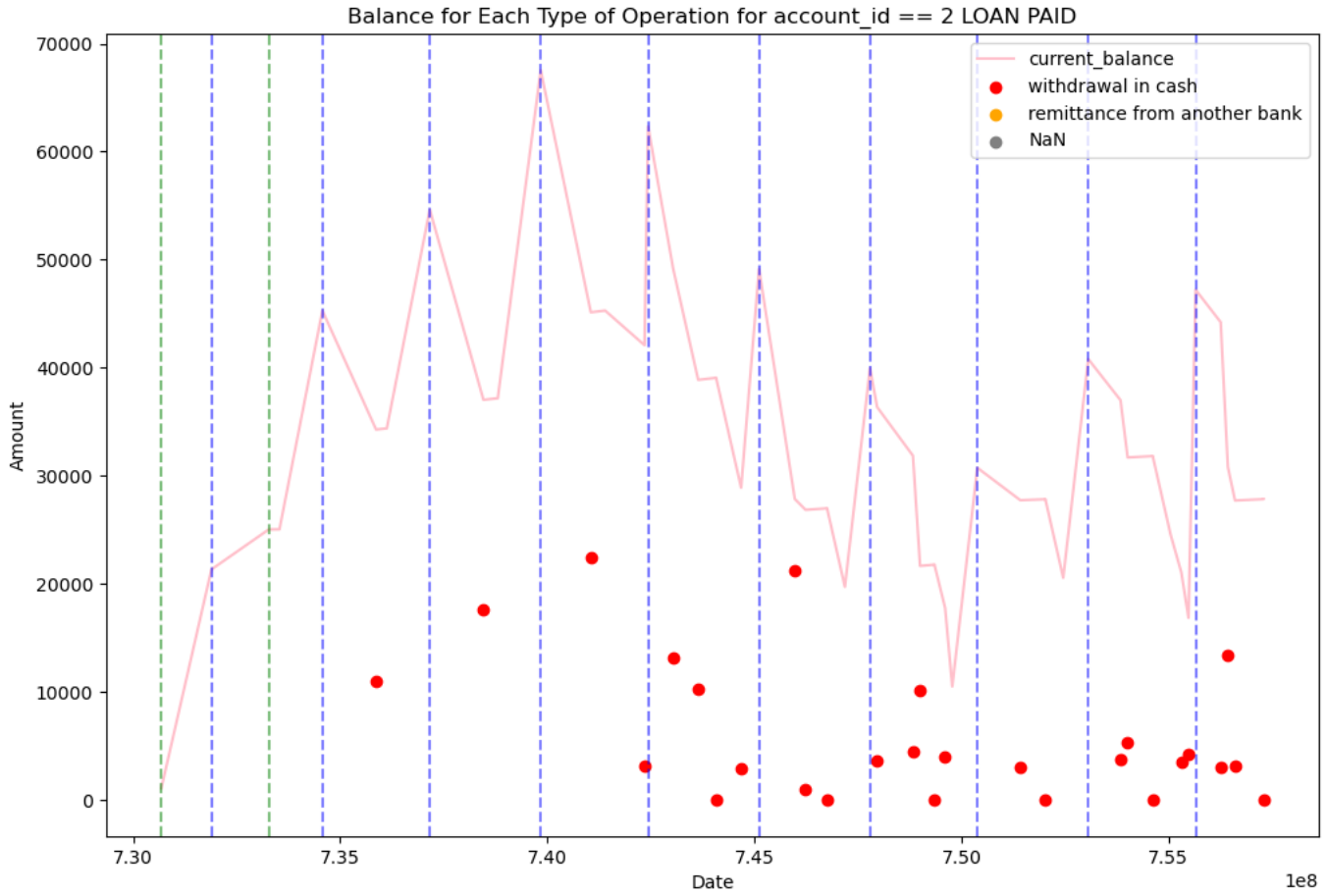




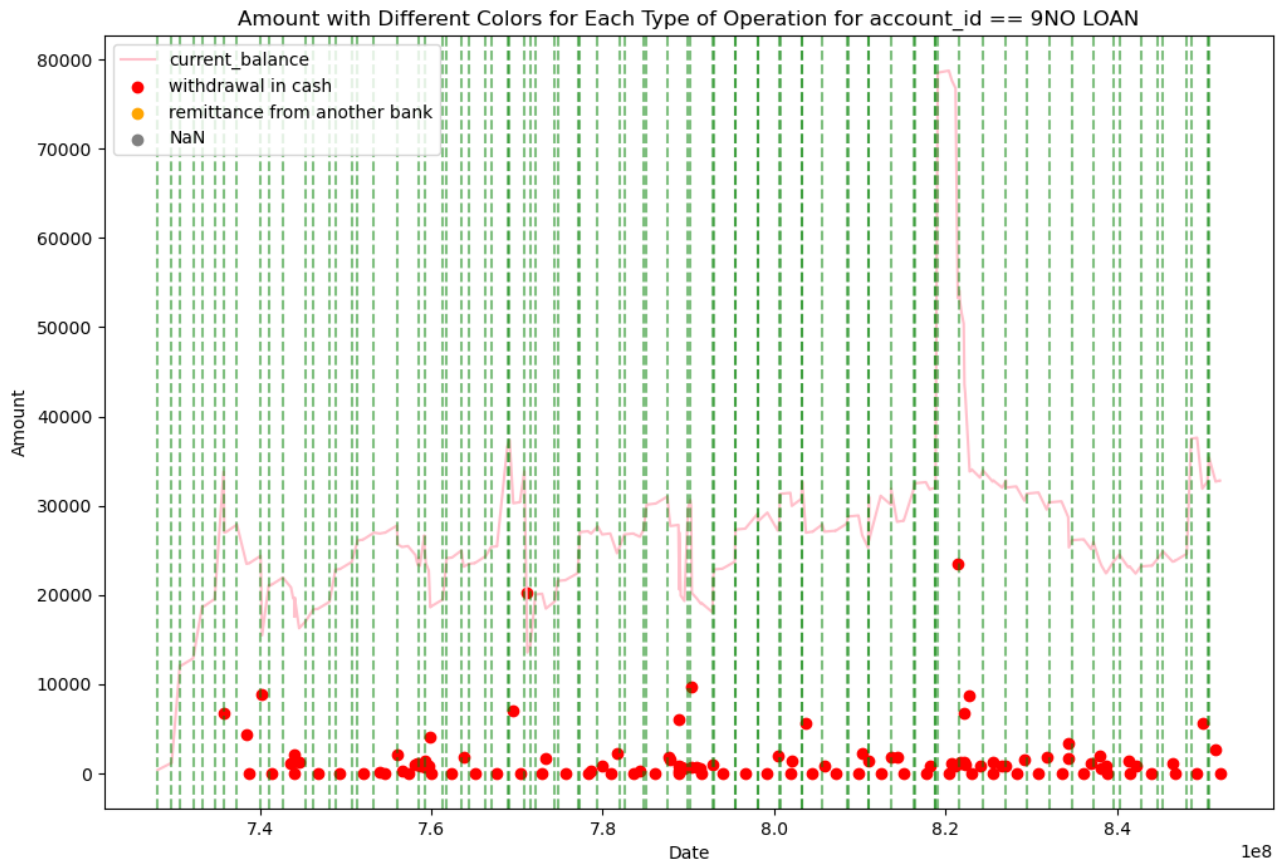
Data Understanding and Preparation



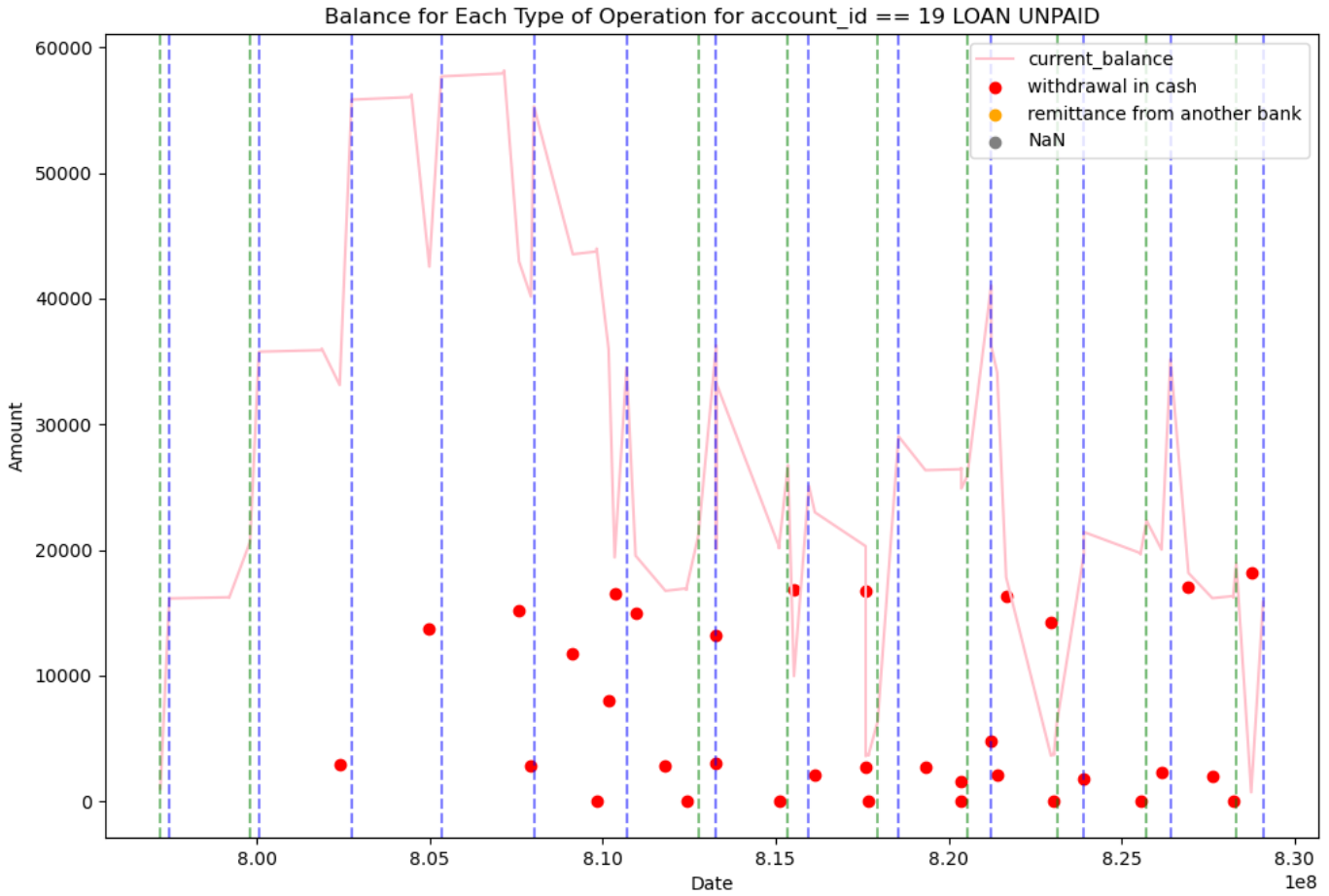
Data Understanding and Preparation



Data Understanding and Preparation



Data Understanding and Preparation



Data Understanding and Preparation

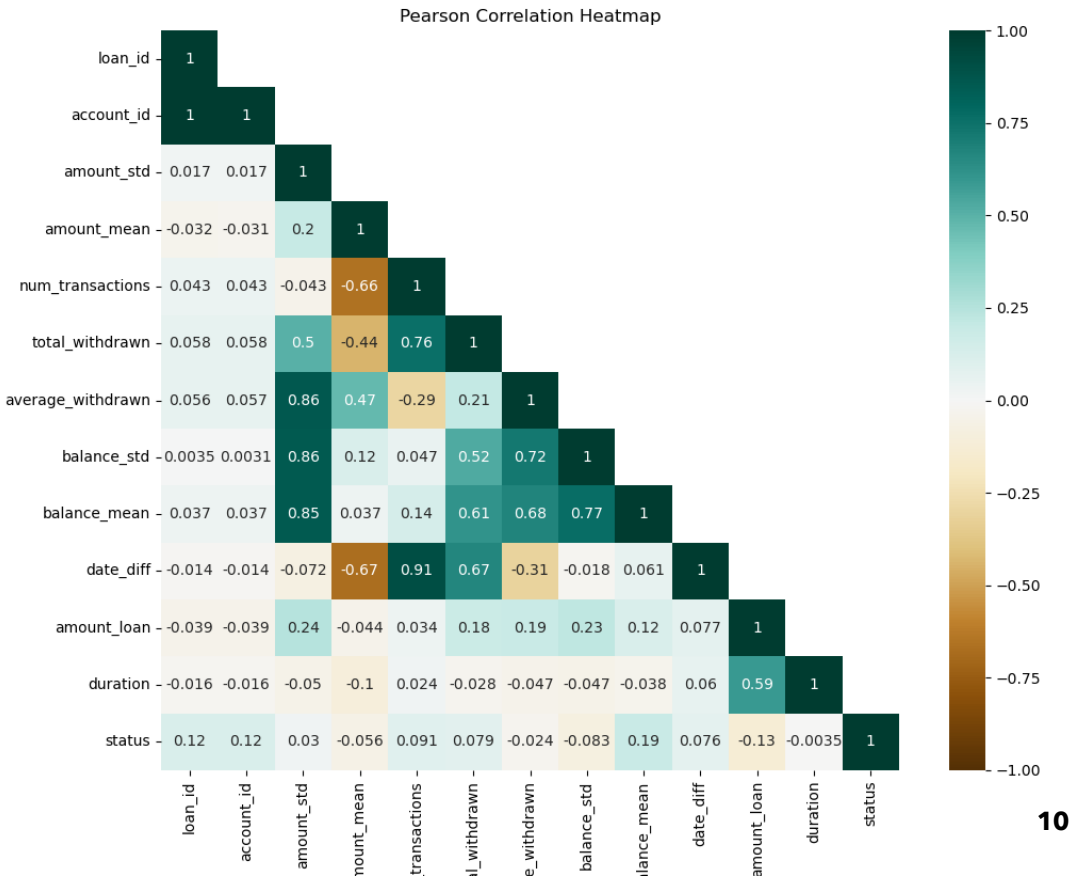
- Aggregated data snapshot

```
loan_id,account_id,amount_std,amount_mean,num_transactions,total_withdrawn,average_withdrawn,balance_std,balance_mean,date_diff,amount_loan,duration,status,disp
4959,2,11895.998464916345,515.5685185185185,54,200844.59999999998,6276.393750000001,12061.705681876094,32590.624074074072,10,80952,24,1,disponent
4961,19,9205.689080326661,198.18,80,226324.80000000002,6656.611764705884,15039.248404932323,25197.0925,12,30276,12,0,no_disponent
4973,67,20882.029392885444,189.51520000000002,125,789552.4,8972.18636363636,20955.646998372446,52523.2448,18,165960,24,1,no_disponent
4996,132,21020.897593232345,2545.4129032258065,31,190733.0,12715.533333333333,21638.258869671175,62778.09032258065,5,88440,12,1,disponent
5002,173,10052.730130005115,933.8466666666667,30,85630.6,4757.2555555555555,11517.175248212317,38709.829999999994,6,104808,12,1,disponent
5032,290,15349.418522400405,412.90491803278684,61,260479.60000000003,6678.964102564103,20572.457529737083,49084.381967213114,11,123696,48,1,no_disponent
5044,344,8168.072604859274,673.3971428571429,70,160567.8,3734.134883720931,9576.327716908689,35614.74428571429,12,100980,60,1,no_disponent
5045,347,17258.630286253032,577.6386363636364,88,451311.39999999997,9026.227999999997,22038.48158908669,42084.31136363636,16,187224,24,0,no_disponent
5060,426,16589.841841132715,1149.122950819672,61,319518.0,9129.085714285715,21418.913855883657,57340.50819672131,7,252060,60,0,no_disponent
5082,501,13169.319978136018,1358.5964285714285,56,166646.0,5207.687500000001,16330.173222315432,41729.61607142857,12,262980,60,1,no_disponent
5088,544,14032.198318519318,511.4116883116883,77,344249.0,7483.673913043478,13400.064096640634,44829.54415584415,15,91152,24,1,no_disponent
5125,789,3462.2725634371154,422.46470588235286,51,51582.99999999999,1842.2499999999998,6550.595429955129,23255.149019607845,9,73056,48,1,disponent
5126,790,15857.719330202128,1273.040625,64,261411.80000000002,7468.908571428573,22812.186440376034,36369.975,14,208128,48,0,no_disponent
```

Data Understanding and Preparation

- Pearson correlation with loan status:

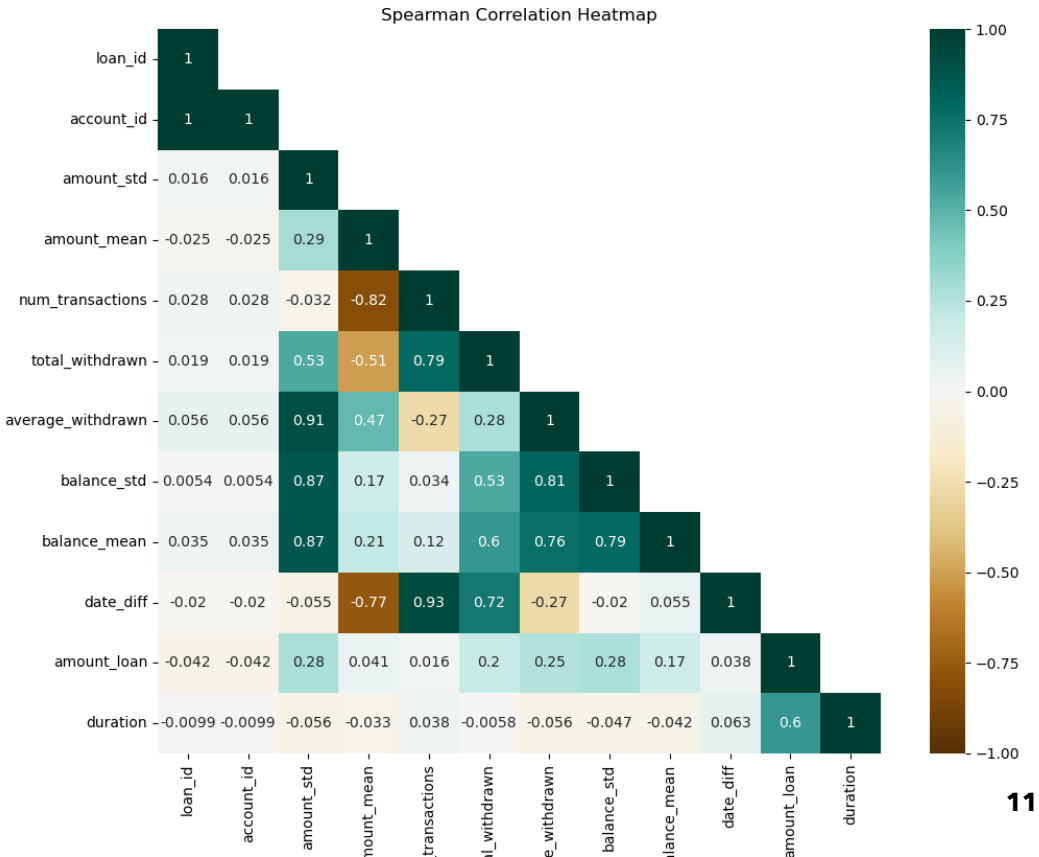
	X	Y	r
0	status	loan_id	0.121917
1	status	account_id	0.122633
2	status	amount_std	0.029760
3	status	amount_mean	-0.056448
4	status	num_transactions	0.091208
5	status	total_withdrawn	0.078689
6	status	average_withdrawn	-0.024413
7	status	balance_std	-0.083467
8	status	balance_mean	0.193985
9	status	date_diff	0.076181
10	status	amount_loan	-0.128237
11	status	duration	-0.003537



Data Understanding and Preparation

- Spearman correlation:
- Correlations between features:

	X	Y	r
0	loan_id	account_id	1.000000
1	num_transactions	date_diff	0.929124
2	amount_std	average_withdrawn	0.906459
3	amount_std	balance_std	0.874023
4	amount_std	balance_mean	0.873036
5	average_withdrawn	date_diff	-0.267229
6	num_transactions	average_withdrawn	-0.269604
7	amount_mean	total_withdrawn	-0.512973
8	amount_mean	date_diff	-0.767940
9	amount_mean	num_transactions	-0.819169



Data Understanding and Preparation

- Feature selection based on p-values:

P values for numerical features:

	Pearson Corr.	p-value
loan_id	0.1219	0.0273
account_id	0.1226	0.0264
amount_std	0.0298	0.5912
amount_mean	-0.0564	0.3081
num_transactions	0.0912	0.0992
total_withdrawn	0.0787	0.1551
average_withdrawn	-0.0244	0.6596
balance_std	-0.0835	0.1314
balance_mean	0.1940	0.0004
date_diff	0.0762	0.1687
amount_loan	-0.1282	0.0202
duration	-0.0035	0.9491

P values for categorical features:
(Chi-square test)

target	0	1
disp		
disponent	0	75
no_disponent	46	207

p-value: 0.00014869537509726651891 56

Dropped features: loan_id, account_id, num_transactions, total_withdrawn, amount_mean, amount_std

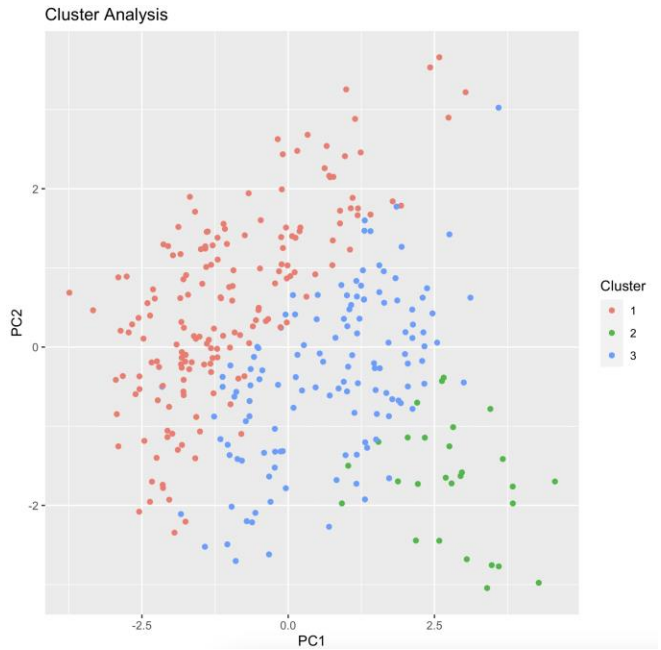
Data Understanding and Preparation

- Encoded and filtered data snapshot

```
average_withdrawn,balance_std,balance_mean,date_diff,amount_loan,duration,status,disp
6276.393750000001,12061.705681876094,32590.624074074072,10,80952,24,1,1
6656.611764705884,15039.248404932323,25197.0925,12,30276,12,0,0
8972.18636363636,20955.646998372446,52523.2448,18,165960,24,1,0
12715.533333333333,21638.258869671175,62778.09032258065,5,88440,12,1,1
4757.2555555555555,11517.175248212317,38709.829999999994,6,104808,12,1,1
6678.964102564103,20572.457529737083,49084.381967213114,11,123696,48,1,0
3734.134883720931,9576.327716908689,35614.74428571429,12,100980,60,1,0
9026.227999999997,22038.48158908669,42084.31136363636,16,187224,24,0,0
9129.085714285715,21418.913855883657,57340.50819672131,7,252060,60,0,0
5207.687500000001,16330.173222315432,41729.61607142857,12,262980,60,1,0
7483.673913043478,13400.064096640634,44829.54415584415,15,91152,24,1,0
1842.2499999999998,6550.595429955129,23255.149019607845,9,73056,48,1,1
7468.908571428573,22812.186440376034,36369.975,14,208128,48,0,0
5049.245161290322,13056.987306542755,21447.624590163938,10,215616,48,0,0
4360.941463414635,7695.862800862465,29380.839743589742,15,24312,12,1,0
10126.164705882351,25825.858181660515,63857.236752136756,17,48624,24,1,0
6523.961111111112,13408.02036465337,33687.85862068966,11,538500,60,1,0
22425.466666666667,34581.40134837078,66292.372,4,187104,24,1,0
1922.3738317757015,6554.760512834303,30847.14932432432,20,87216,48,1,0
```

Descriptive Modelling

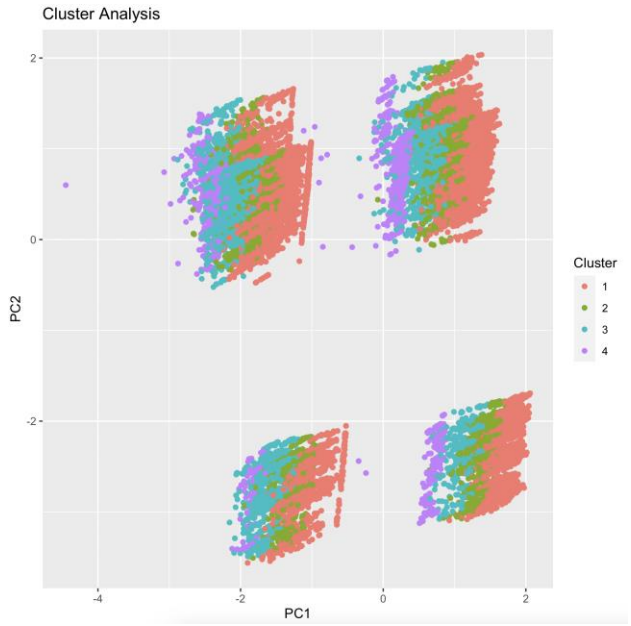
loan_id	num_transactions	total_withdrawn	average_withdrawn	balance_std	balance_mean	amount_loan	duration	status	disp
4959	54	200844.59999999998	6276.393750000001	12061.705681876094	32590.624074074072	80952	24	1	1
4961	80	226324.80000000002	6656.611764705884	15039.248404932323	25197.0925	30276	12	0	0
4973	125	789552.4	8972.18636363636	20955.646998372446	52523.2448	165960	24	1	0
4996	31	190733.0	12715.533333333333	21638.258869671175	62778.09032258065	88440	12	1	1
5002	30	85630.6	4757.2555555555555	11517.175248212317	38709.829999999994	104808	12	1	1



Silhouette Coefficient 0.37

Descriptive Modelling

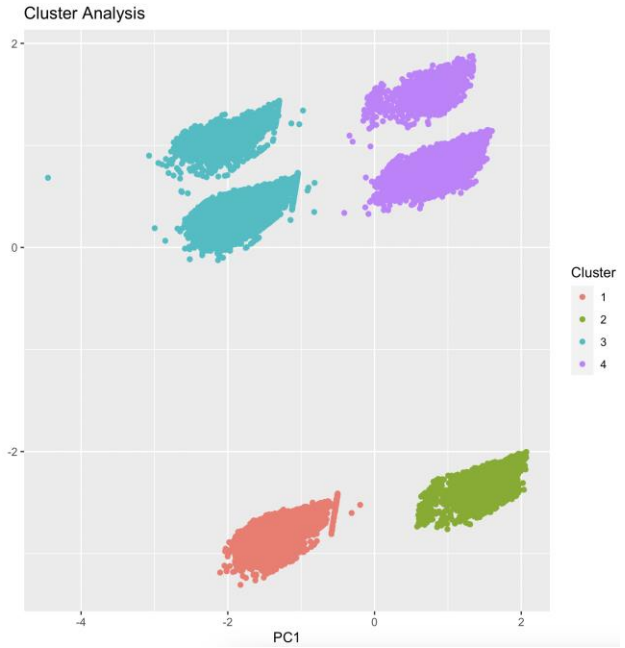
amount_loan	duration	payments	status	balance	date_diff	type_credit	type_withdrawal	disp_disponent	disp_no_disponent
80952	24	3373	1	1100.0	0.0	1	0	1	0
80952	24	3373	1	21336.0	1209600.0	1	0	1	0
80952	24	3373	1	25036.0	1382400.0	1	0	1	0
80952	24	3373	1	25049.5	259200.0	1	0	1	0
80952	24	3373	1	45285.5	1036800.0	1	0	1	0



Silhouette Coefficient 0.51

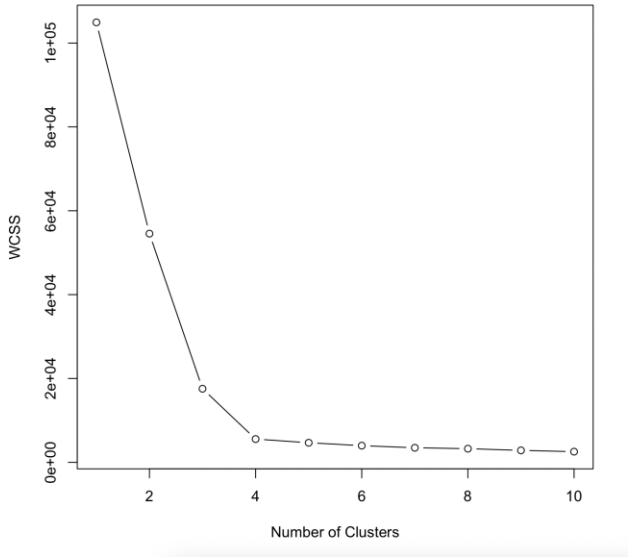
Descriptive Modelling

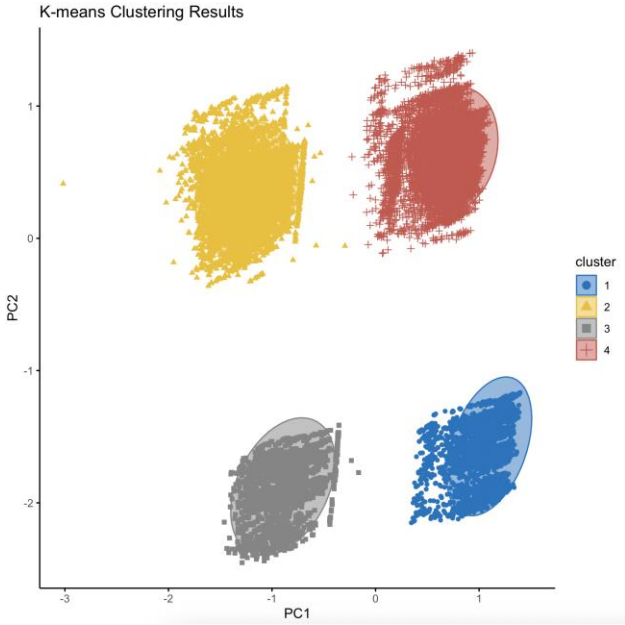
amount_loan	duration	payments	status	balance	date_diff	type_credit	type_withdrawal	disp_disponent	disp_no_disponent
80952	24	3373	1	1100.0	0.0	1	0	1	0
80952	24	3373	1	21336.0	1209600.0	1	0	1	0
80952	24	3373	1	25036.0	1382400.0	1	0	1	0
80952	24	3373	1	25049.5	259200.0	1	0	1	0
80952	24	3373	1	45285.5	1036800.0	1	0	1	0



Silhouette Coefficient 0.81

Descriptive Modelling



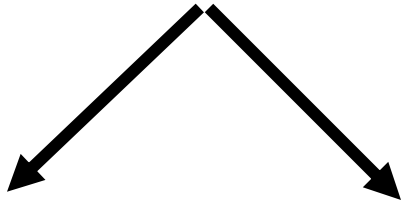


Silhouette Coefficient 0.79

Predictive Modelling

loan_id	num_transactions	total_withdrawn	average_withdrawn	balance_std	balance_mean	amount_loan	duration	status	disp
4959	54	200844.59999999998	6276.393750000001	12061.705681876094	32590.624074074072	80952	24	1	1
4961	80	226324.80000000002	6656.611764705884	15039.248404932323	25197.0925	30276	12	0	0
4973	125	789552.4	8972.18636363636	20955.646998372446	52523.2448	165960	24	1	0
4996	31	190733.0	12715.533333333333	21638.258869671175	62778.09032258065	88440	12	1	1
5002	30	85630.6	4757.2555555555555	11517.175248212317	38709.829999999994	104808	12	1	1

- Original dataframe
- Dataframe with only the PCs (that hold 80% of the information)
- Dataframe with the original attributes and the PCs



80% training 20% testing

Predictive Modelling

Naive Bayes

- **Original dataframe:**
- Accuracy – 39% / Error Rate – 61% / AUC – 64%

- **Dataframe with only the PCs:**
- Accuracy – 98% / Error Rate – 2% / AUC – 95%

- **Dataframe with the original attributes and the PCs:**
- Accuracy – 79% / Error Rate – 21% / AUC – 88%

Predictive Modelling

Decision Trees

- **Original dataframe:**
- Accuracy – 91% / Error Rate – 9% / AUC – 85%

- **Dataframe with only the PCs:**
- Accuracy – 99% / Error Rate – 1% / AUC – 98%

- **Dataframe with the original attributes and the PCs:**
- Accuracy – 99% / Error Rate – 1% / AUC – 99%

Predictive Modelling

KNN (K=5)

- **Original dataframe:**
- Accuracy – 83% / Error Rate – 17% / AUC – 49%

- **Dataframe with only the PCs:**
- Accuracy – 82% / Error Rate – 18% / AUC – 48%

- **Dataframe with the original attributes and the PCs:**
- Accuracy – 83% / Error Rate – 17% / AUC – 49%

Predictive Modelling

Logistic Regression

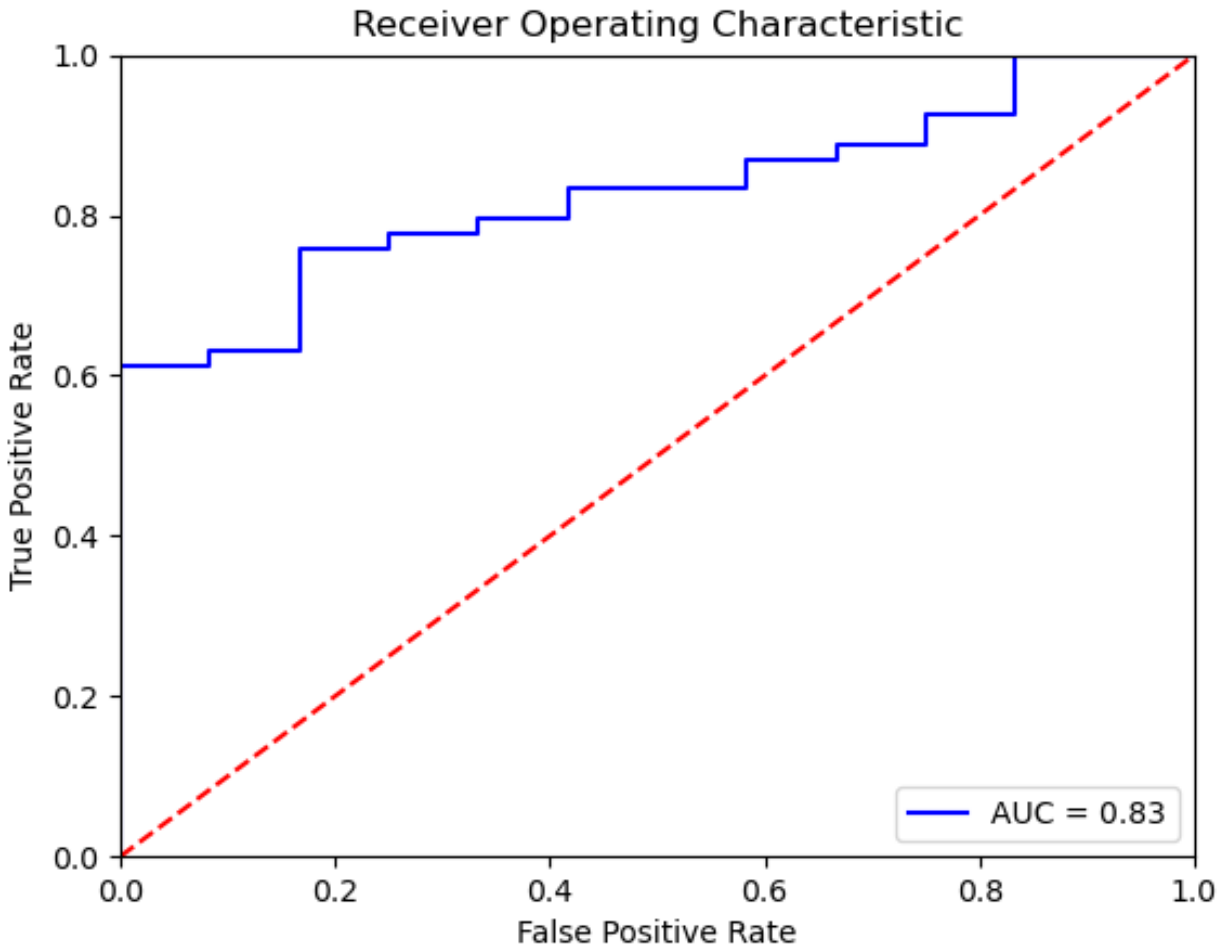
- **Original dataframe:**
- Accuracy – 81% / Error Rate – 0.19% / AUC – 81%

- **Dataframe with only the PCs:**
- Accuracy – 100% / Error Rate – 0% / AUC – 100%

- **Dataframe with the original attributes and the PCs:**
- Accuracy – 100% / Error Rate – 0% / AUC – 100%

Predictive Modelling

Logistical Regression initial results (Python version)

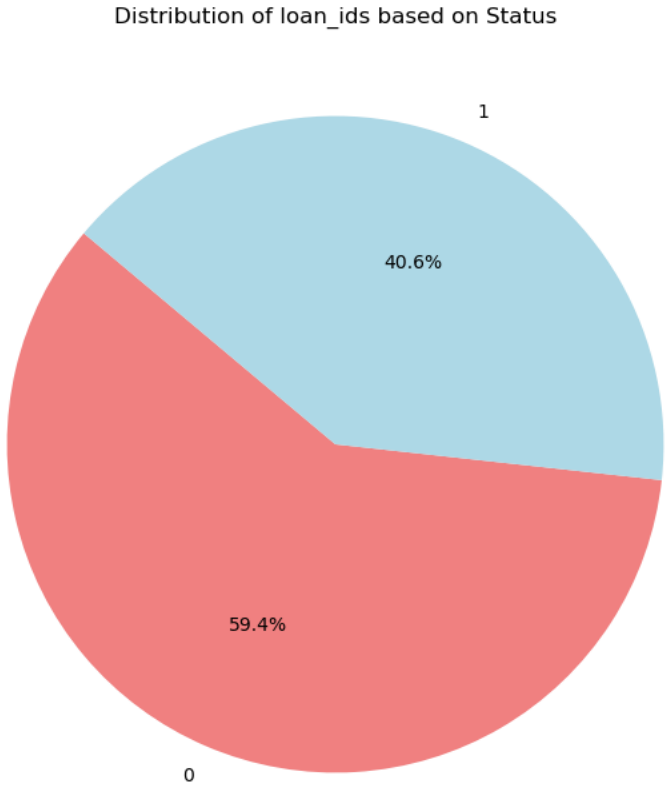


Predictive Modelling

Improving the training dataset:

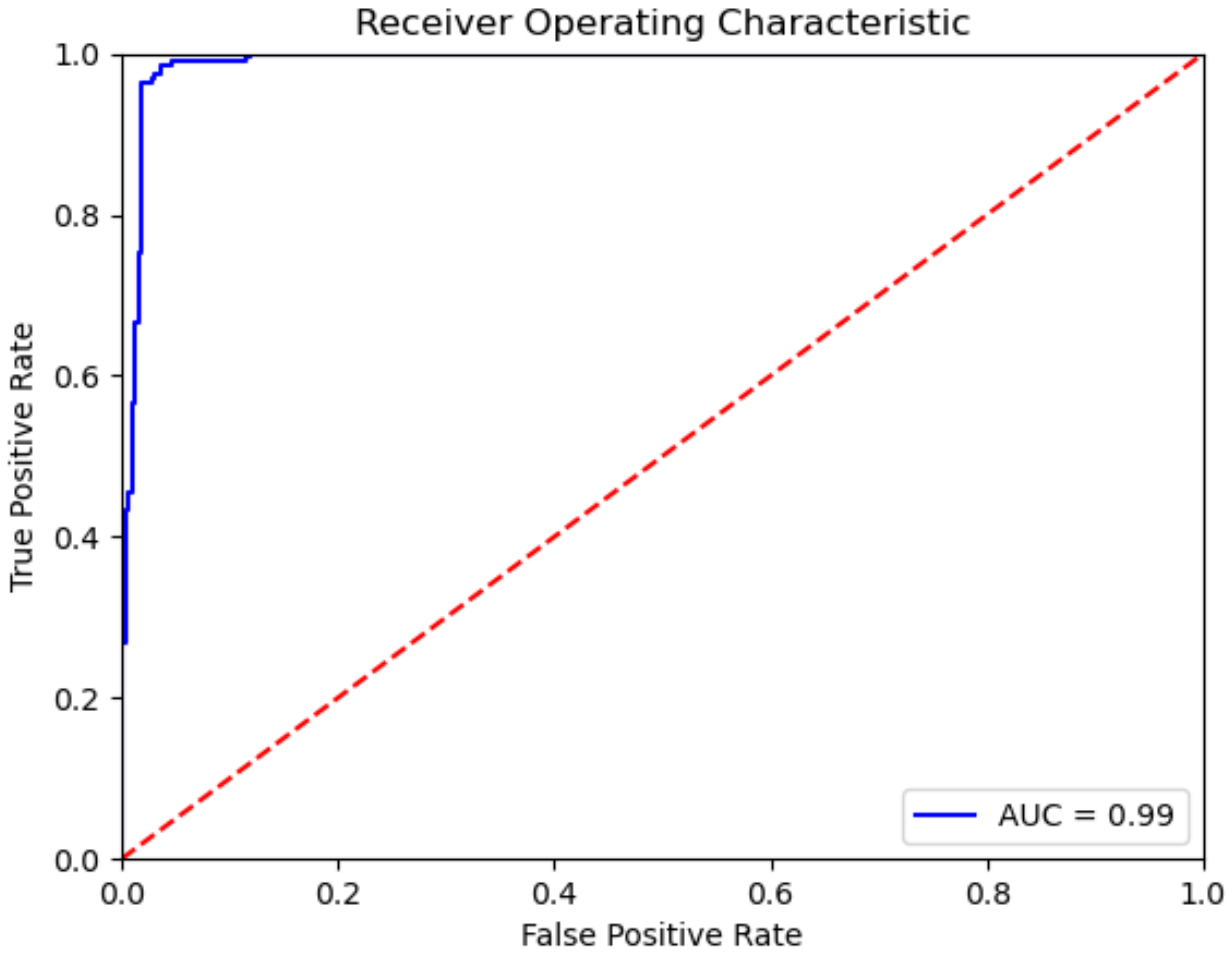
- "rule of thumb" formula for **amount_loan**:
    - (balance_mean - 2 * balance_std) * date_diff
- approximation for loan **duration:**
    - find_closest_value(date_diff, [12, 24, 48, 60])
- apply the logistic regression trained model => obtain **status** for accounts without loans
- merge with the current dataset and retrain

Predictive Modelling



Predictive Modelling

Logistical Regression improved results



Conclusions and Future Considerations

- Given the exploratory nature of the project, a lot of trial and error was encountered
- Data understanding and feature engineering was the most essential stage, its execution strongly influences training and prediction results
- If the deadline would've been extended, perhaps more interesting approaches and results would be observed. Ex: try another generative approach for new accounts with generated loan data
- Regarding collaboration, team communication, synchronization was essential to scheduling and fulfilling tasks, overall very good atmosphere