DS ASSIGNMENT
ENVEL AI

VRUTIK HALANI

* First Glance of Data:

Number of Columns : 30

Number of Categorical Columns : 14

Categorical columns :

1. Account_balance → Can be used as a target Column
2. Check
3. CS_FICO_str
4. CS_internal
5. feeCode
6. feeDescription
7. isCredit
8. returnCode
9. Status
10. Student → Can be used as a target Column
11. Subtype
12. subtypeCode
13. Type
14. institutionName

Columns that can be dropped and are of No use for prediction :

1. Transaction Count (since all are distinct values)

Target Columns:

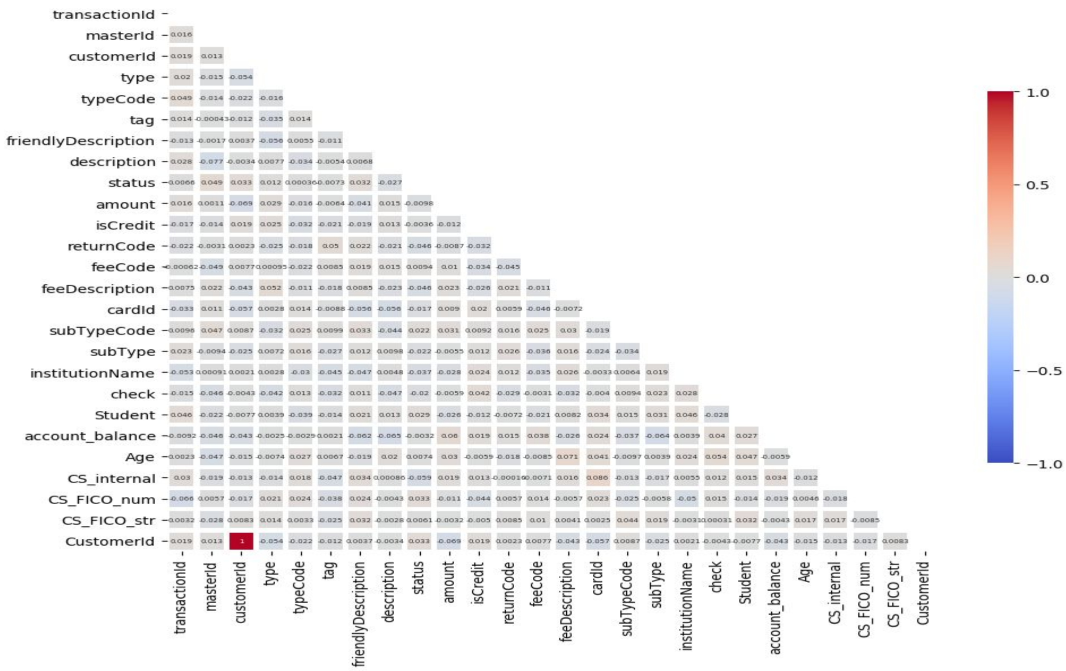
1. Account_Balance : Given data about the customer and all his transactions we can use the data to predict classification of Account_Balance of given customer at a point of Transaction
2. Student : Given all the data about the customer and his transactions the model can learn to classify if given customer is a student and depending on his situation suggestions for Institutions and other expenses can be given to him by searching for other similar users in the database
3. Amount : Given other data we can create a regressor model to predict the amount of given transaction for particular customer based on similar other customers for the transaction.
4. Age : Use the data to predict Customer's age which then can be used to suggest different things to the user.

*Statistical Analysis of Data at First :

Steps involved:

1. Check all the column names and depict the behaviour of each column.
2. Listing all column names of the data and number of unique values involved in that column.
3. Listing number of Unique values the column possesses and then analysing the data based on that to determine which columns to drop.
4. Label Encoding done to remove strings and all converted to numbers.
5. Drop the columns leading to no data.
6. Create Correlation Matrix - Heat Map for the data.
7. Create Correlation for Target Variable - using online API
8. Create different Graphs for checking the statistics of data.

* Heat Map and Correlation with Target :



Column name	Data type	Nullability	Missing% (Count)	Invalid values	Distinct values	Correlation with Target
account_balance Target	Categorical	Nullable	0% (0)	0% (0)	4	---
check	Categorical	Nullable	0% (0)	0% (0)	2	0.122
CS_FICO_str	Categorical	Nullable	0% (0)	0% (0)	6	0.122
CS_internal	Categorical	Nullable	0% (0)	0% (0)	3	0.122
feeCode	Categorical	Nullable	0% (0)	0% (0)	3	0.122
feeDescription	Categorical	Nullable	0% (0)	0% (0)	4	0.122
institutionName	Categorical	Nullable	0% (0)	0% (0)	18	0.123
isCredit	Categorical	Nullable	0% (0)	0% (0)	2	0.122
returnCode	Categorical	Nullable	0% (0)	0% (0)	4	0.122
status	Categorical	Nullable	0% (0)	0% (0)	4	0.122
Student	Categorical	Nullable	0% (0)	0% (0)	2	0.122
subType	Categorical	Nullable	0% (0)	0% (0)	2	0.122
subTypeCode	Categorical	Nullable	0% (0)	0% (0)	2	0.122
type	Categorical	Nullable	0% (0)	0% (0)	7	0.122
Age	Numeric	Nullable	0% (0)	0% (0)	18	0.494
amount	Numeric	Nullable	0% (0)	0% (0)	1,475	0.495
cardId	Numeric	Nullable	0% (0)	0% (0)	571	0.495
CS_FICO_num	Numeric	Nullable	0% (0)	0% (0)	403	0.496
customerId	Numeric	Nullable	0% (0)	0% (0)	567	0.495
description	Numeric	Nullable	0% (0)	0% (0)	573	0.495
friendlyDescription	Numeric	Nullable	0% (0)	0% (0)	569	0.494
masterId	Numeric	Nullable	0% (0)	0% (0)	570	0.494
tag	Numeric	Nullable	0% (0)	0% (0)	572	0.496
transactionCount	Numeric	Nullable	0% (0)	0% (0)	1,475	0.494
transactionId	Numeric	Nullable	0% (0)	0% (0)	555	0.495
typeCode	Numeric	Nullable	0% (0)	0% (0)	575	0.494
availableDate	Timestamp	Nullable	0% (0)	0% (0)	1,475	0.351
createdDate	Timestamp	Nullable	0% (0)	0% (0)	1,475	0.351
settledDate	Timestamp	Nullable	0% (0)	0% (0)	1,475	0.351
voidedDate	Timestamp	Nullable	0% (0)	0% (0)	1,475	0.351

Column name	Data type	Nullability	Missing% (Count)	Invalid values	Distinct values	Correlation with Target
account_balance	Categorical	Nullable	0% (0)	0% (0)	4	0.494
check	Categorical	Nullable	0% (0)	0% (0)	2	0.494
CS_FICO_str	Categorical	Nullable	0% (0)	0% (0)	6	0.497
CS_internal	Categorical	Nullable	0% (0)	0% (0)	3	0.494
feeCode	Categorical	Nullable	0% (0)	0% (0)	3	0.493
feeDescription	Categorical	Nullable	0% (0)	0% (0)	4	0.494
institutionName	Categorical	Nullable	0% (0)	0% (0)	18	0.51
isCredit	Categorical	Nullable	0% (0)	0% (0)	2	0.493
returnCode	Categorical	Nullable	0% (0)	0% (0)	4	0.495
status	Categorical	Nullable	0% (0)	0% (0)	4	0.493
Student	Categorical	Nullable	0% (0)	0% (0)	2	0.494
subType	Categorical	Nullable	0% (0)	0% (0)	2	0.492
subTypeCode	Categorical	Nullable	0% (0)	0% (0)	2	0.493
type	Categorical	Nullable	0% (0)	0% (0)	7	0.498
Age Target	Numeric	Nullable	0% (0)	0% (0)	18	—
amount	Numeric	Nullable	0% (0)	0% (0)	1,475	0.496
cardId	Numeric	Nullable	0% (0)	0% (0)	571	0.496
CS_FICO_num	Numeric	Nullable	0% (0)	0% (0)	403	0.496
customerId	Numeric	Nullable	0% (0)	0% (0)	567	0.496
description	Numeric	Nullable	0% (0)	0% (0)	573	0.497
friendlyDescription	Numeric	Nullable	0% (0)	0% (0)	569	0.494
masterId	Numeric	Nullable	0% (0)	0% (0)	570	0.495
tag	Numeric	Nullable	0% (0)	0% (0)	572	0.496
transactionCount	Numeric	Nullable	0% (0)	0% (0)	1,475	0.495
transactionId	Numeric	Nullable	0% (0)	0% (0)	555	0.495
typeCode	Numeric	Nullable	0% (0)	0% (0)	575	0.495
availableDate	Timestamp	Nullable	0% (0)	0% (0)	1,475	0.521
createdDate	Timestamp	Nullable	0% (0)	0% (0)	1,475	0.521
settledDate	Timestamp	Nullable	0% (0)	0% (0)	1,475	0.521
voidedDate	Timestamp	Nullable	0% (0)	0% (0)	1,475	0.521

Column name	Data type	Nullability	Missing% (Count)	Invalid values	Distinct values	Correlation with Target
account_balance	Categorical	Nullable	0% (0)	0% (0)	4	0.495
check	Categorical	Nullable	0% (0)	0% (0)	2	0.493
CS_FICO_str	Categorical	Nullable	0% (0)	0% (0)	6	0.498
CS_internal	Categorical	Nullable	0% (0)	0% (0)	3	0.495
feeCode	Categorical	Nullable	0% (0)	0% (0)	3	0.495
feeDescription	Categorical	Nullable	0% (0)	0% (0)	4	0.495
institutionName	Categorical	Nullable	0% (0)	0% (0)	18	0.502
isCredit	Categorical	Nullable	0% (0)	0% (0)	2	0.493
returnCode	Categorical	Nullable	0% (0)	0% (0)	4	0.494
status	Categorical	Nullable	0% (0)	0% (0)	4	0.494
Student	Categorical	Nullable	0% (0)	0% (0)	2	0.493
subType	Categorical	Nullable	0% (0)	0% (0)	2	0.494
subTypeCode	Categorical	Nullable	0% (0)	0% (0)	2	0.493
type	Categorical	Nullable	0% (0)	0% (0)	7	0.496
Age	Numeric	Nullable	0% (0)	0% (0)	18	0.496
amount Target	Numeric	Nullable	0% (0)	0% (0)	1,475	—
cardId	Numeric	Nullable	0% (0)	0% (0)	571	0.496
CS_FICO_num	Numeric	Nullable	0% (0)	0% (0)	403	0.495
customerId	Numeric	Nullable	0% (0)	0% (0)	567	0.499
description	Numeric	Nullable	0% (0)	0% (0)	573	0.496
friendlyDescription	Numeric	Nullable	0% (0)	0% (0)	569	0.496
masterId	Numeric	Nullable	0% (0)	0% (0)	570	0.496
tag	Numeric	Nullable	0% (0)	0% (0)	572	0.495
transactionCount	Numeric	Nullable	0% (0)	0% (0)	1,475	0.496
transactionId	Numeric	Nullable	0% (0)	0% (0)	555	0.495
typeCode	Numeric	Nullable	0% (0)	0% (0)	575	0.496
availableDate	Timestamp	Nullable	0% (0)	0% (0)	1,475	0.52
createdDate	Timestamp	Nullable	0% (0)	0% (0)	1,475	0.52
settledDate	Timestamp	Nullable	0% (0)	0% (0)	1,475	0.52
voidedDate	Timestamp	Nullable	0% (0)	0% (0)	1,475	0.52

	Column name	Data type	Nullability	Missing% (Count)	Invalid values	Distinct values	Correlation with Target
	account_balance	Categorical	Nullable	0% (0)	0% (0)	4	0.122
	check	Categorical	Nullable	0% (0)	0% (0)	2	0.122
	CS_FICO_str	Categorical	Nullable	0% (0)	0% (0)	6	0.122
	CS_internal	Categorical	Nullable	0% (0)	0% (0)	3	0.122
	feeCode	Categorical	Nullable	0% (0)	0% (0)	3	0.122
	feeDescription	Categorical	Nullable	0% (0)	0% (0)	4	0.122
	institutionName	Categorical	Nullable	0% (0)	0% (0)	18	0.123
	isCredit	Categorical	Nullable	0% (0)	0% (0)	2	0.122
	returnCode	Categorical	Nullable	0% (0)	0% (0)	4	0.122
	status	Categorical	Nullable	0% (0)	0% (0)	4	0.122
✓	Student Target	Categorical	Nullable	0% (0)	0% (0)	2	---
	subType	Categorical	Nullable	0% (0)	0% (0)	2	0.122
	subTypeCode	Categorical	Nullable	0% (0)	0% (0)	2	0.122
	type	Categorical	Nullable	0% (0)	0% (0)	7	0.122
	Age	Numeric	Nullable	0% (0)	0% (0)	18	0.494
	amount	Numeric	Nullable	0% (0)	0% (0)	1,475	0.493
	cardId	Numeric	Nullable	0% (0)	0% (0)	571	0.493
	CS_FICO_num	Numeric	Nullable	0% (0)	0% (0)	403	0.493
	customerId	Numeric	Nullable	0% (0)	0% (0)	567	0.493
	description	Numeric	Nullable	0% (0)	0% (0)	573	0.493
	friendlyDescription	Numeric	Nullable	0% (0)	0% (0)	569	0.494
	masterId	Numeric	Nullable	0% (0)	0% (0)	570	0.493
	tag	Numeric	Nullable	0% (0)	0% (0)	572	0.493
	transactionCount	Numeric	Nullable	0% (0)	0% (0)	1,475	0.494
	transactionId	Numeric	Nullable	0% (0)	0% (0)	555	0.494
	typeCode	Numeric	Nullable	0% (0)	0% (0)	575	0.493
	availableDate	Timestamp	Nullable	0% (0)	0% (0)	1,475	0.349
	createdDate	Timestamp	Nullable	0% (0)	0% (0)	1,475	0.349
	settledDate	Timestamp	Nullable	0% (0)	0% (0)	1,475	0.349
	voidedDate	Timestamp	Nullable	0% (0)	0% (0)	1,475	0.349

*Label Encoding :

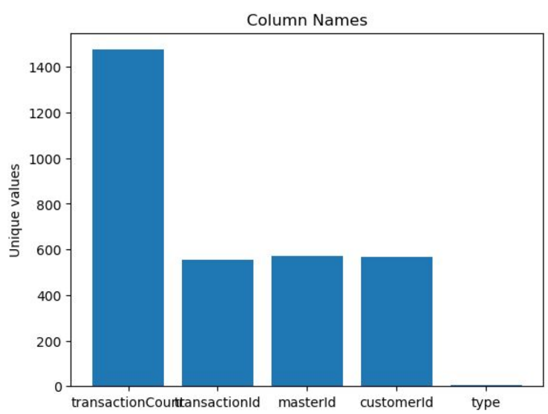
- * type:
  - CorePro Deposit → 1
  - CorePro Recurring Withdrawal → 2
  - CorePro Withdrawal → 3
  - Interest Adjustment → 4
  - Interest Paid → 5
  - Internal CorePro Transfer → 6
  - Manual Adjustment → 7
- * isCredit , subTypeCode , subType , check:
  - Y → 1
  - N → 0

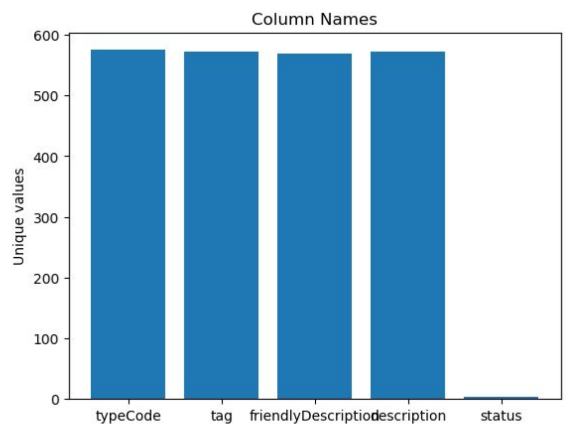
- * Fee Description :
  - Abeus Papam → 1
  - Dominus Vobiscum → 2
  - Gallia est omnis divisa in partes → 3
  - Veni, Vidi, Vici → 4
- * Instituion Name :
  - Bank of America → 1
  - Barclays → 2
  - Budapest Bank → 3
  - Capital One → 4
  - CHASE Bank → 5
  - CIT Group → 6
  - Citigroup → 7
  - Citizens Bank → 8
  - Citizens_Bank → 9
  - First Rand Bank → 10
  - HSBC Bank USA → 11
  - MFUG Union Bank → 12
  - New York Community Bank → 13

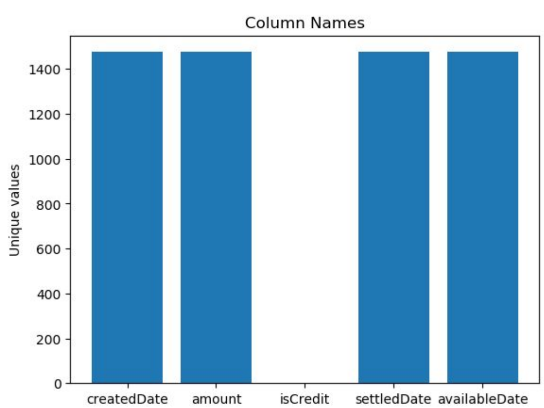
Royal Bank of Scotland→ 14
Santander Bank → 15
State Street Corporation → 16
Toronto Dominion Bank→ 17
Webster Bank → 18

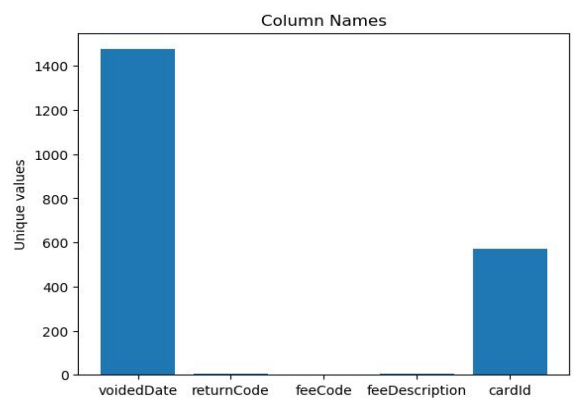
*Graphs made:

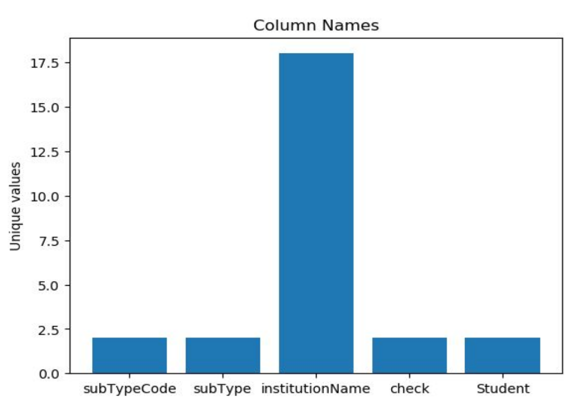
1. Unique values in each column

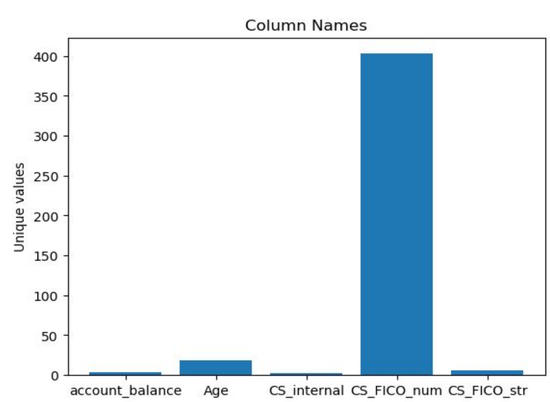




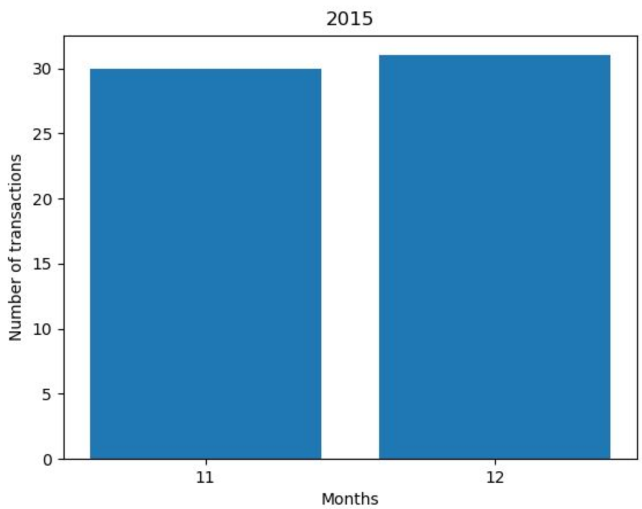


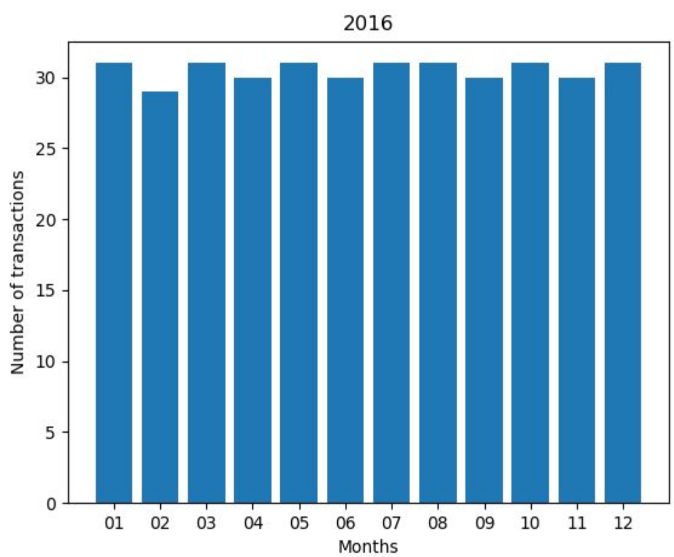


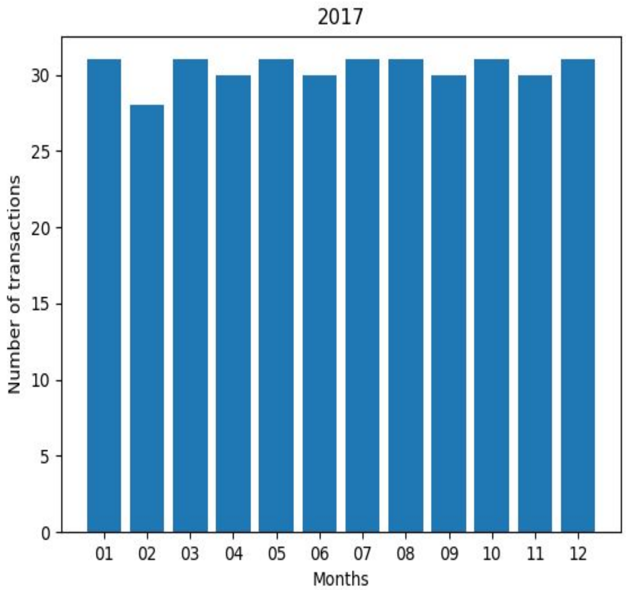


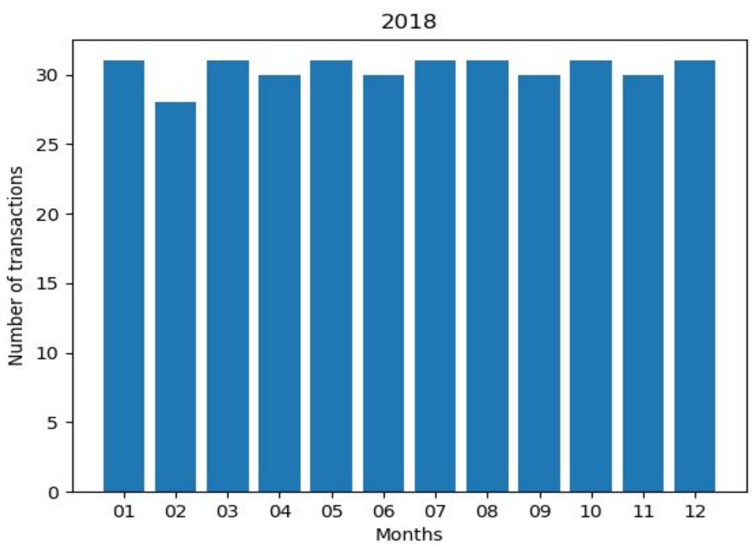


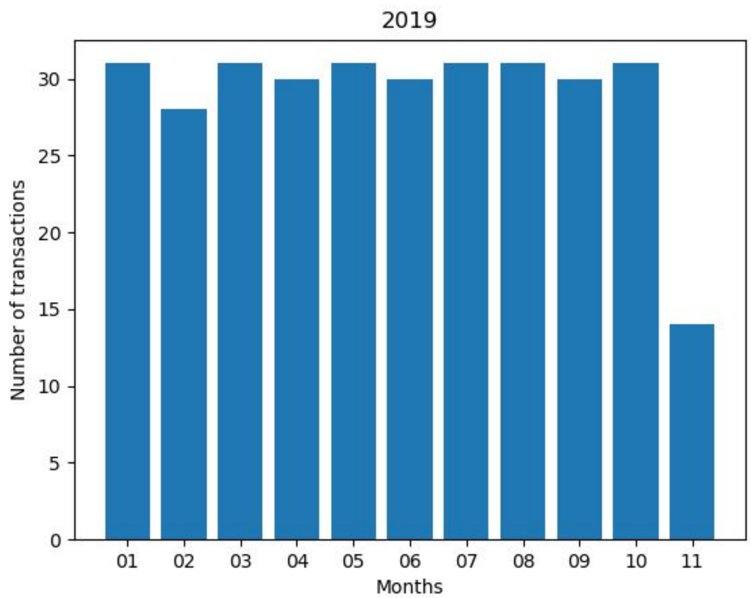
2. Every year graph - month-wise for number of transactions in every month in every year



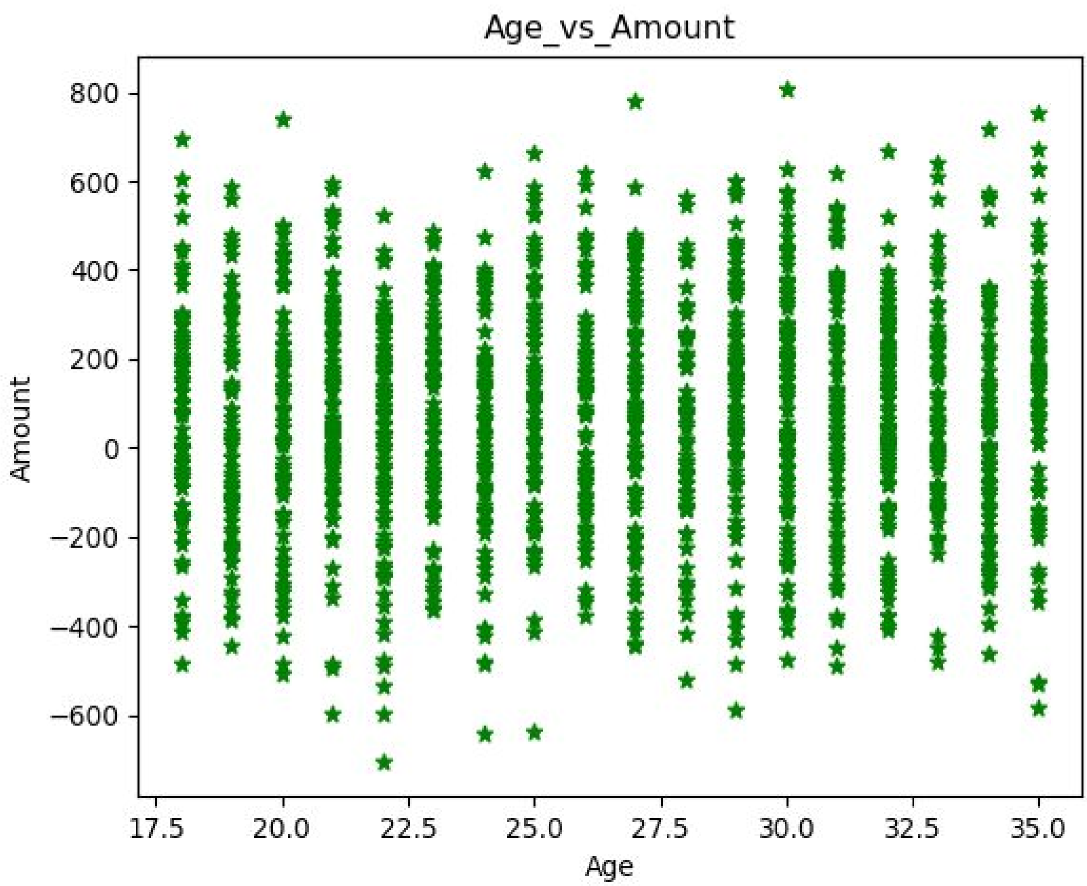



2017
Number of transactions
0
5
10
15
20
25
30
01 02 03 04 05 06 07 08 09 10 11 12
Months

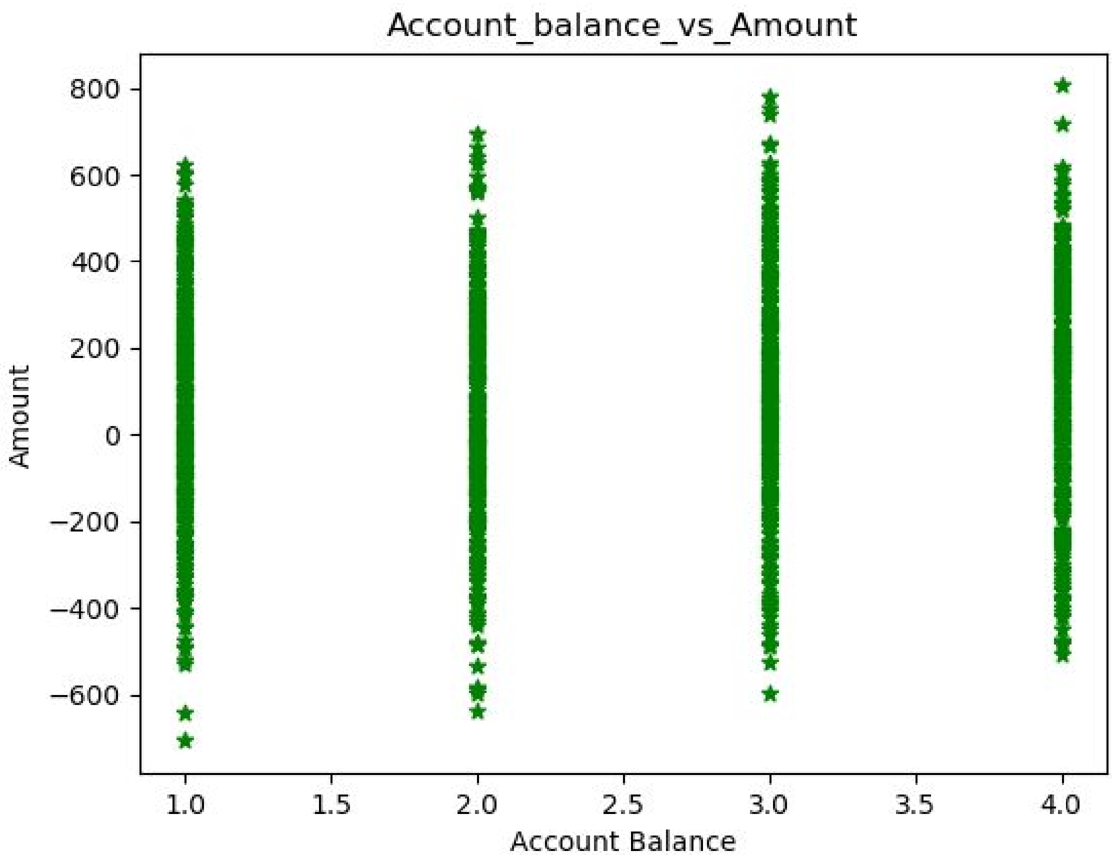

2018
Number of transactions
0
5
10
15
20
25
30
01 02 03 04 05 06 07 08 09 10 11 12
Months



3. Age vs Amount of transaction graph

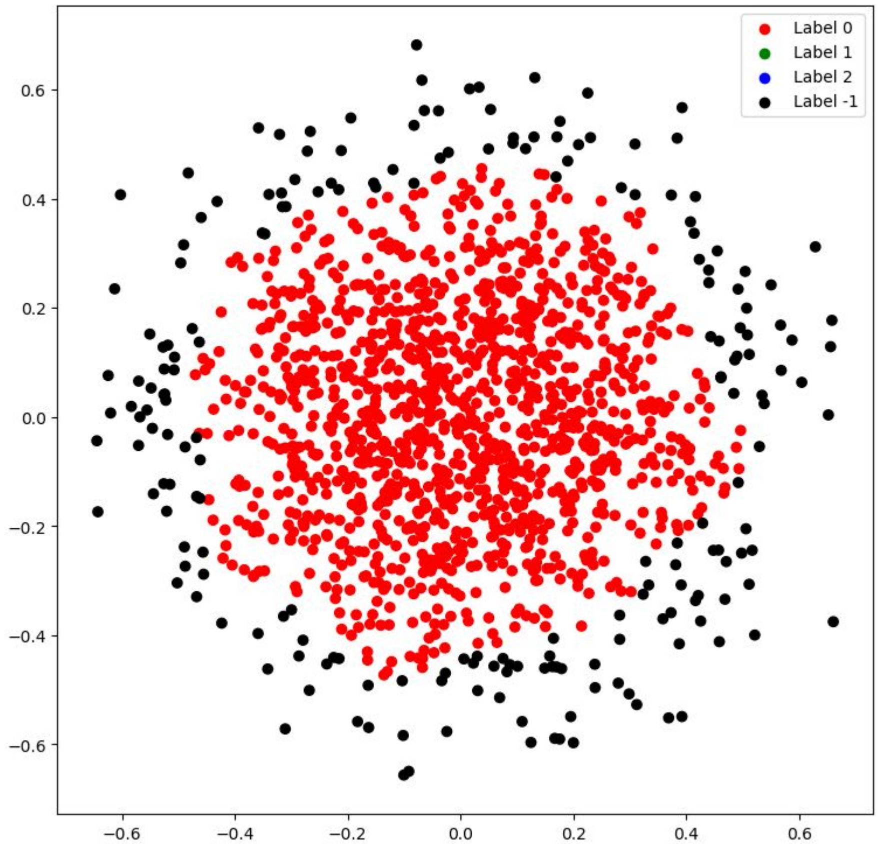


4. Account Balance vs Amount graph



These graphs help us understand the data and seeing the graphs we can see that if we want to predict some of the data through Machine Learning we need to remove the anomalies which can be clearly seen through the graph.

To remove the anomalies , we can use DBSCAN Clustering algorithm and using appropriate parameters we can detect anomalies in the data which can then be removed for better predictions.



DBSCAN

* General Proposition for Anomaly Detection :

We can use Generative Adversarial Networks (GANs) that are basically constructed for images and fake image detection on Tabular data like the given data.

What we can do is , using the Encoder function on train data we can create encoded data by regression and classification neural networks and then taking this data as input for the generator function we create fake data corresponding to every train data and then measuring the distance between train data and corresponding fake data we can give anomaly scores. The higher the distance the higher the anomaly score.

Note : Due to lack of time I was not able to apply this algorithm on the given data but I have applied it to another dataset as a part of my research and this algorithm gives about 70% accuracy. We can tune the parameters for further corrections.

* After Removing Anomalies :

After removing the anomalies from the data we can use different methods for prediction of target variable.
In this assignment I did not remove the anomalies but I tried predicting the target variable "**Student**" using different Classifiers.

*** Classifiers used :**

1. Random Forest Classifier with parameter tuning
2. XG Boost Classifier along with Random Forest
3. KNN (K Nearest Neighbours)
4. Gradient Boosting Classifier Algorithm with parameter tuning

*** Results Achieved :**

1. For RandomForestClassifier : Max Accuracy = 0.5799457994579946
2. For KNN : Max Accuracy = 0.5094850948509485
3. For XGBoost : Max Accuracy = 0.5745257452574526
4. For Gradient Boosting : Max Accuracy = 0.564

From this we can conclude that XGBoost or Gradient Boosting Algorithms should be used for the dataset since it took almost 2000 iterations to make Random Forest Classifier Best Model but XGBoost and Gradient Descent gave almost similar accuracy in less time. If we have to create a Rest API then Random Forest can be a good idea but since we need to keep training the model on live data time is important so using Boosting Algorithms is the best way to trade off between time since there is not much difference in accuracy.

The lower accuracy scores are due to less data available for training the models. Similar and other techniques can also be used for predictions of other Target Variables such as **“Amount” , “Account_Balance” and “Age”.**

“Amount” and “Account_Balance” being numeric data and not categorical data we can use Neural Networks to predict the data or Random Forest Regressor.

* Final Observation :

Since the data given was too random (proved by Correlation HeatMap) , to make predictions is very difficult. Also when taken into consideration all the features for neural networks it takes too much time.

What we can conclude from this exercise , we can separate anomalies using DBSCAN and also plotting various graphs. There were points in graphs above which can clearly be interpreted as anomalies without any algorithms. Those points were seen in Account Balance vs Amount graph and Age vs Amount graph.

Also Gradient Boosting Algorithm had very less change in accuracy over many iterations.