Credit Consumption Prediction Challenge

Team Members

-Durgesh

-Mandar

-Omkar

Problem Statement

- Predict the average spend for a different set of customers in the test set for the coming 3 months.

Potential Business Problems

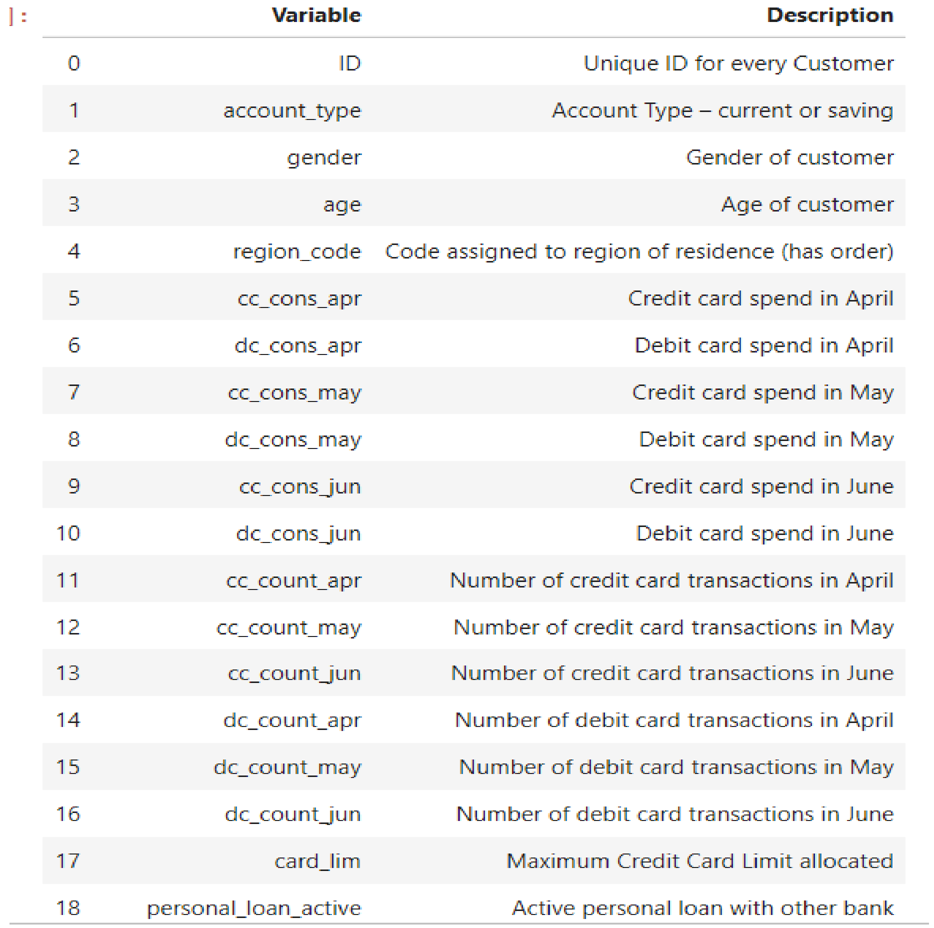
- To understand these patterns thoroughly and get insights on the customer persona and the spending patterns.

- allow banks to build strategic partnerships with vendors for discounts or other plans to reward and retain customers.

- To profile people and tailor the loans or financial products based on those.

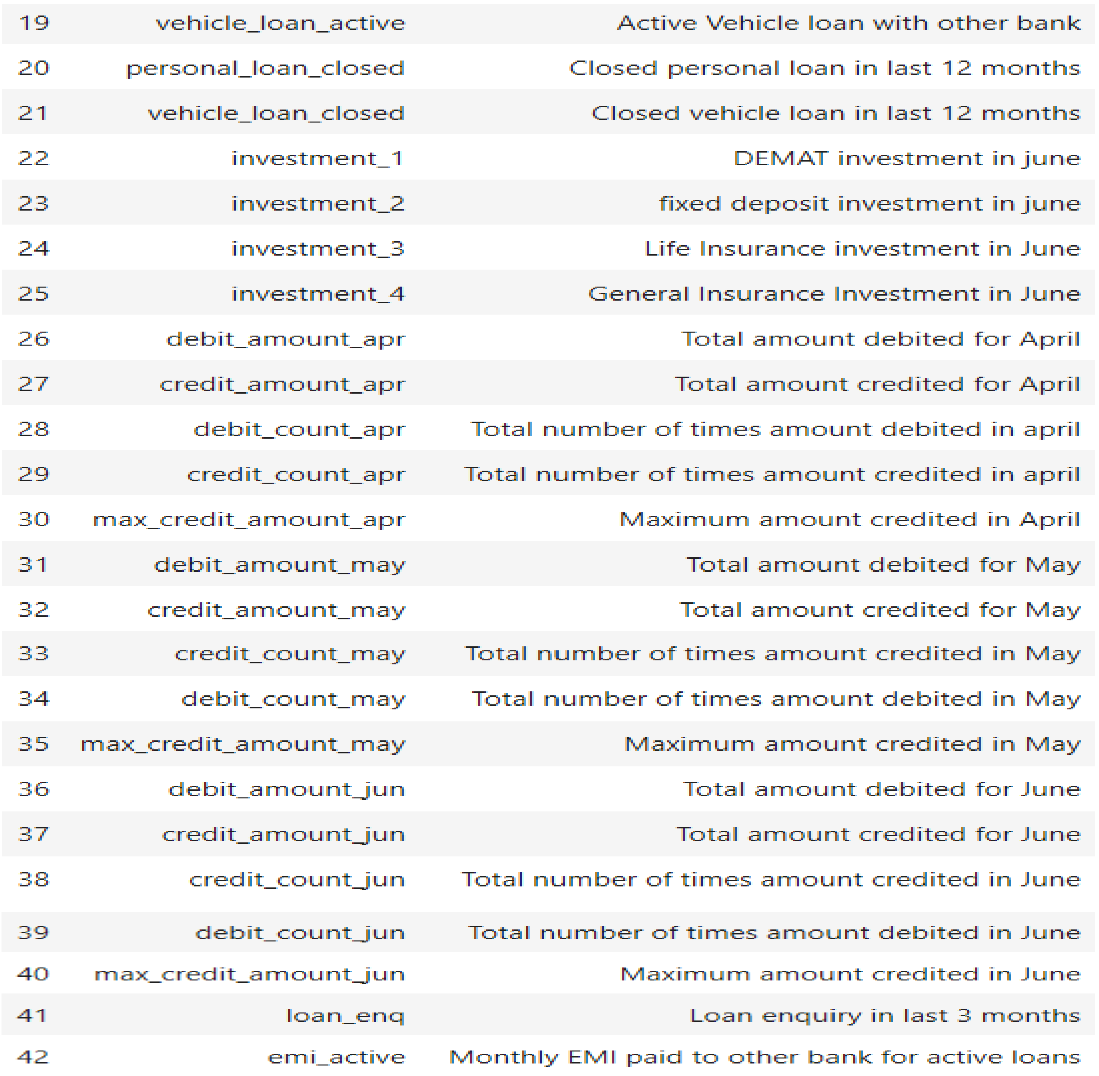
DATA

- **Dataset Information:** The data consists of records of roughly 15000 clients and 44 features. There are 43 predictors and 1 target that gives expected average spend in the coming 3 months (July, August & September).

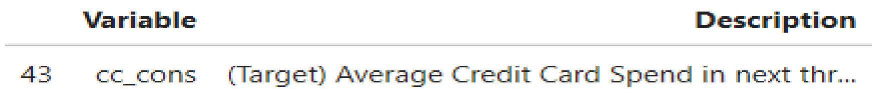


]:

	Variable	Description
0	ID	Unique ID for every Customer
1	account_type	Account Type – current or saving
2	gender	Gender of customer
3	age	Age of customer
4	region_code	Code assigned to region of residence (has order)
5	cc_cons_apr	Credit card spend in April
6	dc_cons_apr	Debit card spend in April
7	cc_cons_may	Credit card spend in May
8	dc_cons_may	Debit card spend in May
9	cc_cons_jun	Credit card spend in June
10	dc_cons_jun	Debit card spend in June
11	cc_count_apr	Number of credit card transactions in April
12	cc_count_may	Number of credit card transactions in May
13	cc_count_jun	Number of credit card transactions in June
14	dc_count_apr	Number of debit card transactions in April
15	dc_count_may	Number of debit card transactions in May
16	dc_count_jun	Number of debit card transactions in June
17	card_lim	Maximum Credit Card Limit allocated
18	personal_loan_active	Active personal loan with other bank



19	vehicle_loan_active	Active Vehicle loan with other bank
20	personal_loan_closed	Closed personal loan in last 12 months
21	vehicle_loan_closed	Closed vehicle loan in last 12 months
22	investment_1	DEMAT investment in june
23	investment_2	fixed deposit investment in june
24	investment_3	Life Insurance investment in June
25	investment_4	General Insurance Investment in June
26	debit_amount_apr	Total amount debited for April
27	credit_amount_apr	Total amount credited for April
28	debit_count_apr	Total number of times amount debited in april
29	credit_count_apr	Total number of times amount credited in april
30	max_credit_amount_apr	Maximum amount credited in April
31	debit_amount_may	Total amount debited for May
32	credit_amount_may	Total amount credited for May
33	credit_count_may	Total number of times amount credited in May
34	debit_count_may	Total number of times amount debited in May
35	max_credit_amount_may	Maximum amount credited in May
36	debit_amount_jun	Total amount debited for June
37	credit_amount_jun	Total amount credited for June
38	credit_count_jun	Total number of times amount credited in June
39	debit_count_jun	Total number of times amount debited in June
40	max_credit_amount_jun	Maximum amount credited in June
41	loan_enq	Loan enquiry in last 3 months
42	emi_active	Monthly EMI paid to other bank for active loans



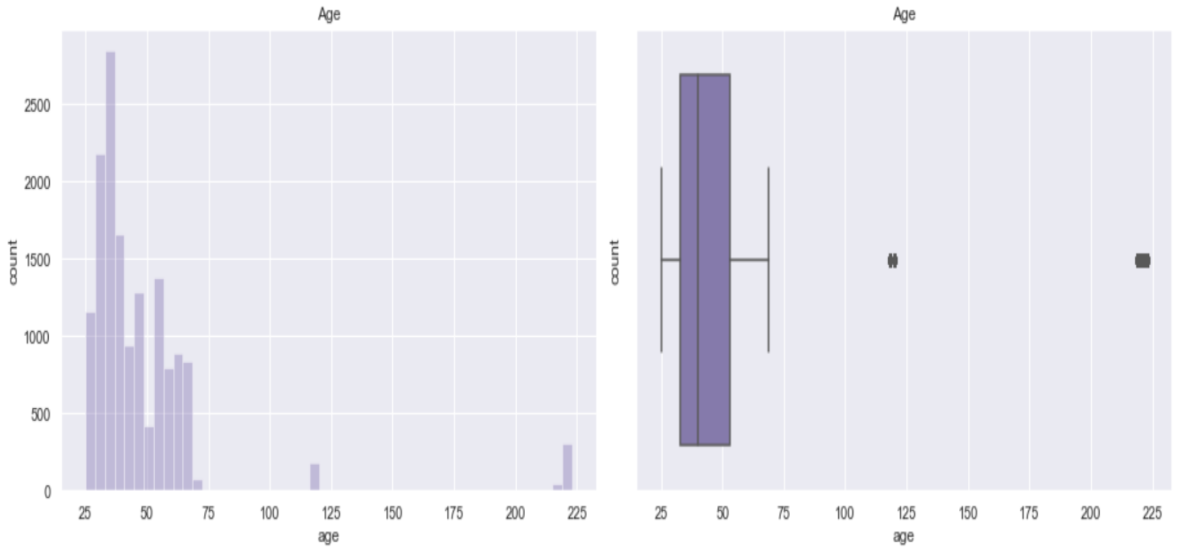
	Variable	Description
43	cc_cons	(Target) Average Credit Card Spend in next thr...

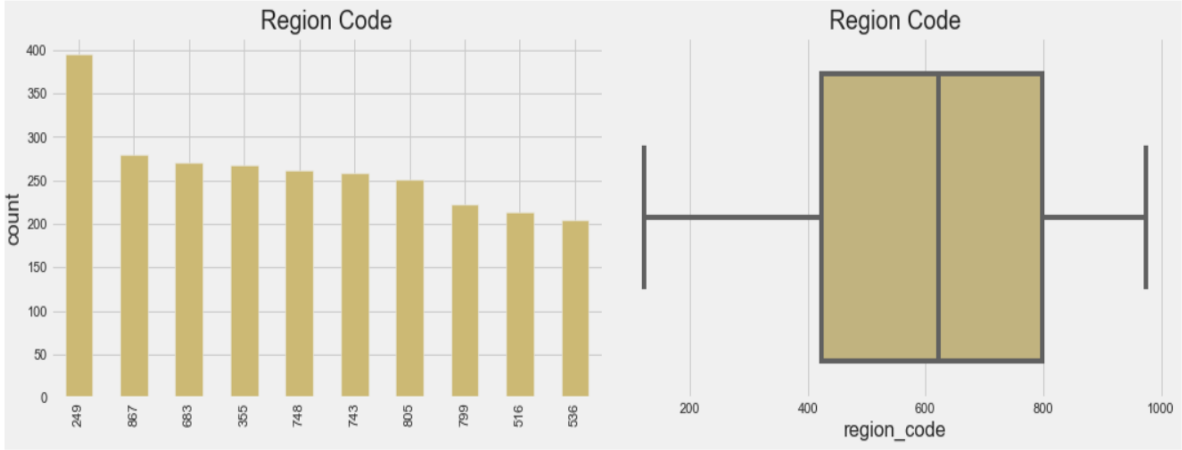
Evaluation Metric

- The average predicted spend of customers for the next three months would be evaluated using **Root of Mean Squared Logarithmic Error i.e RMSLE**.

First steps - EDA

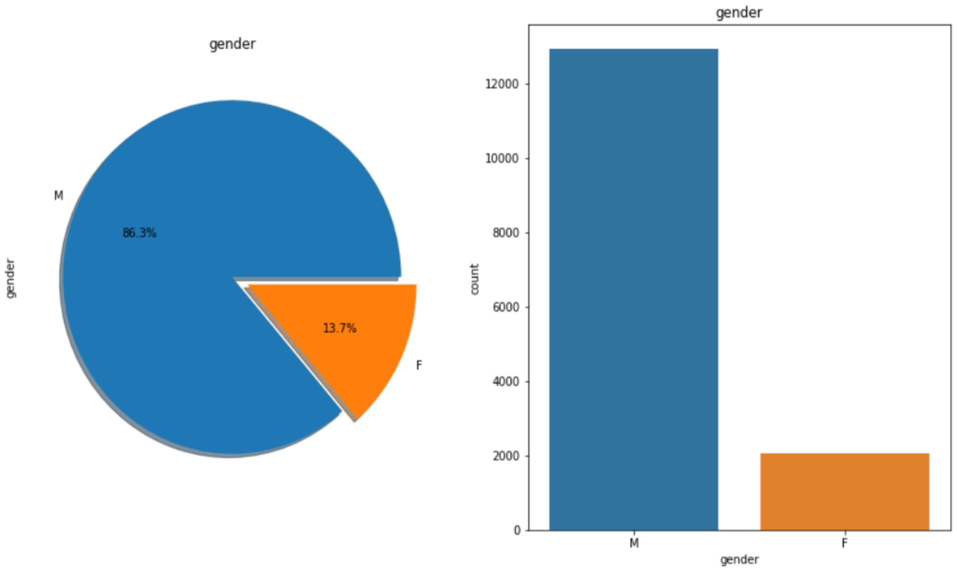
Age and Region Code

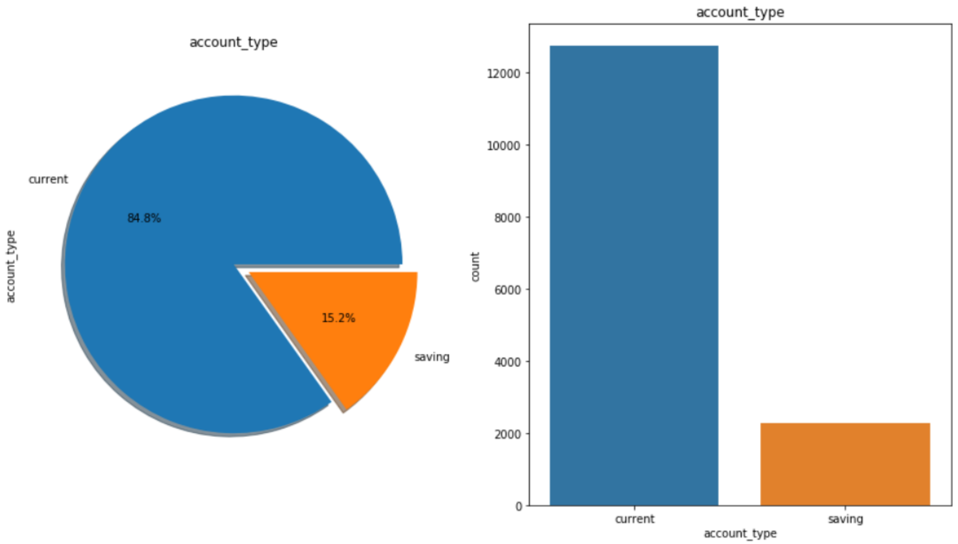




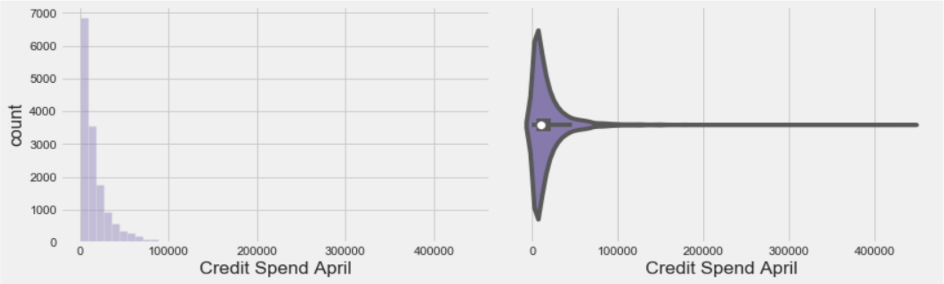
1. It can be observed 95% of the customers are from the age group of 25 to 70 and the remaining 5 % are due to incorrect entries as seen in the above graph.
2. Majority of them are in the 30-40 age bracket

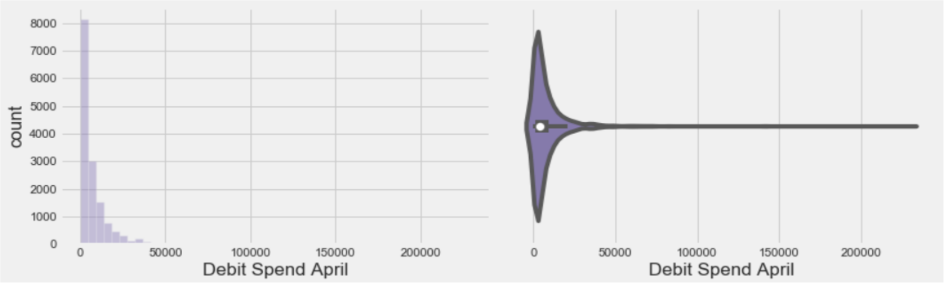
Gender and Account Type

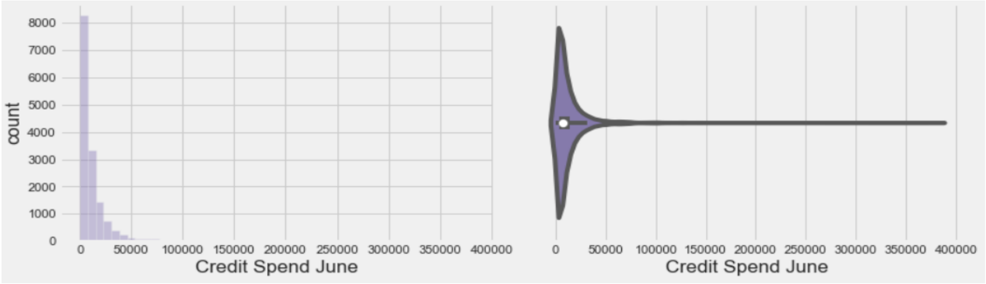


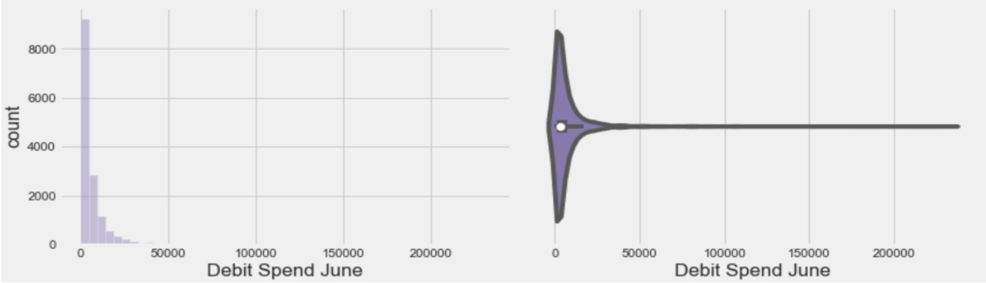


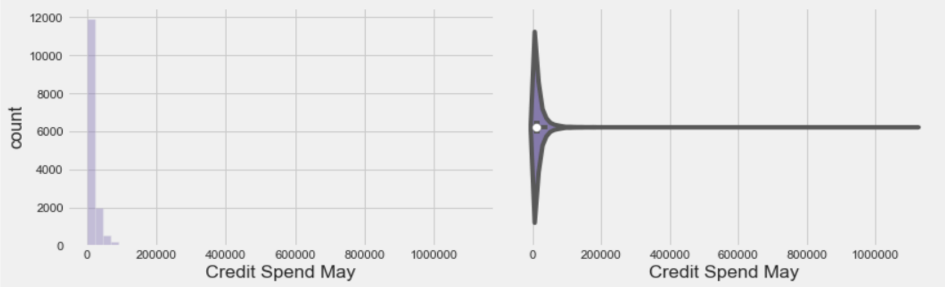
Credit and Debit Card Spends

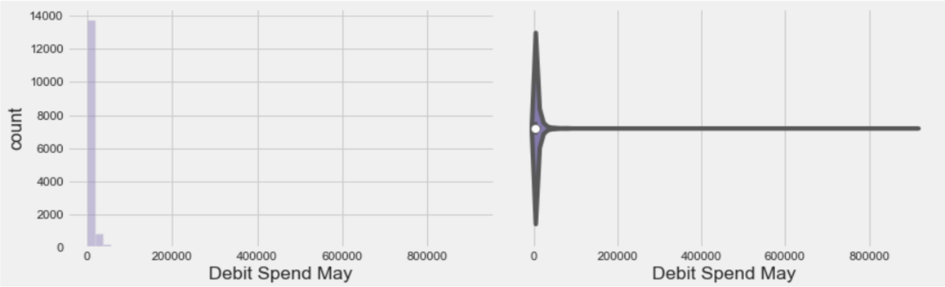








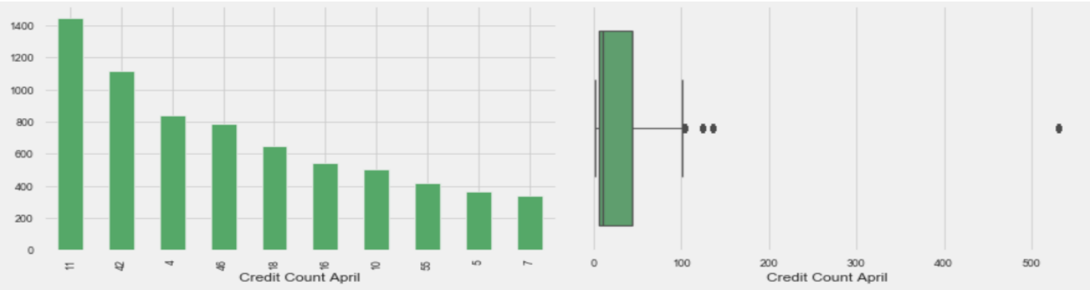


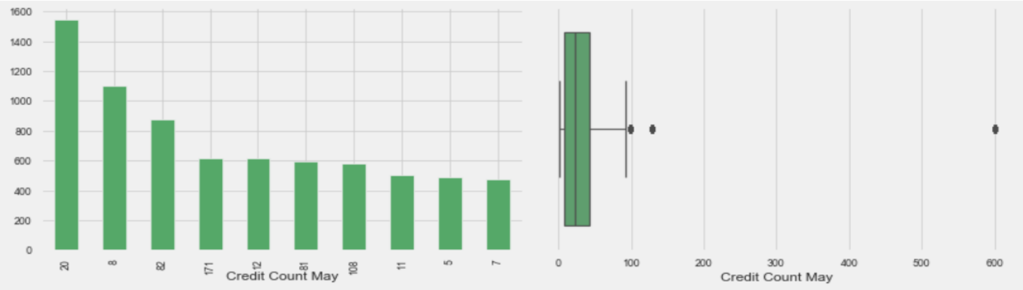


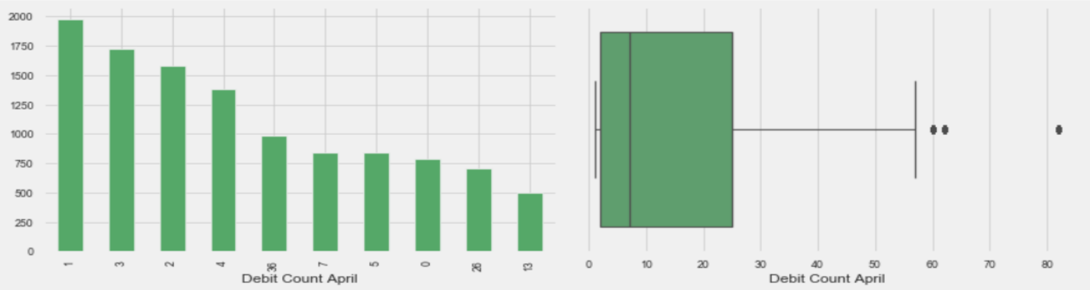
Observations from Credit and Debit Card Spend

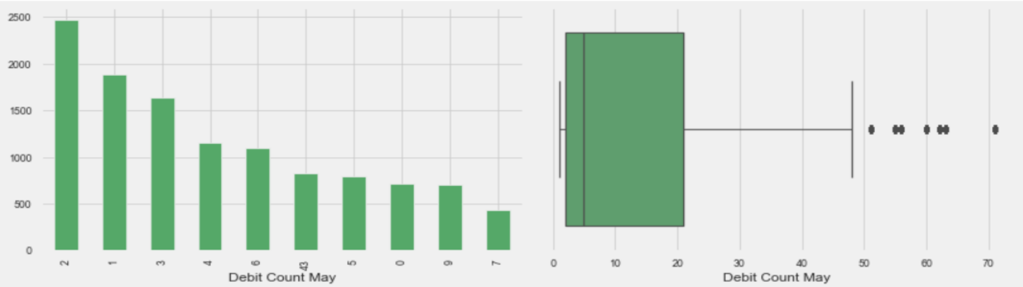
1. From the above plots we can see the that maximum number of amounts spend using credit or debit card are less than 50000.
2. The range is particlarly large for the month of May with range extendig upto 10 lakh for credit card 8 lakh for debit card spend. the month of may had siginificnatly higher spends

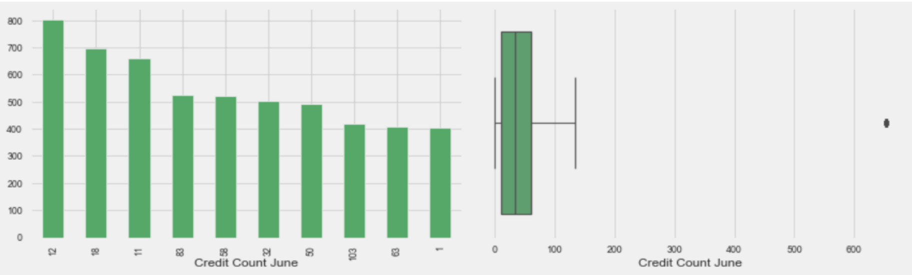
Number of Credit and Debit Card Transaction

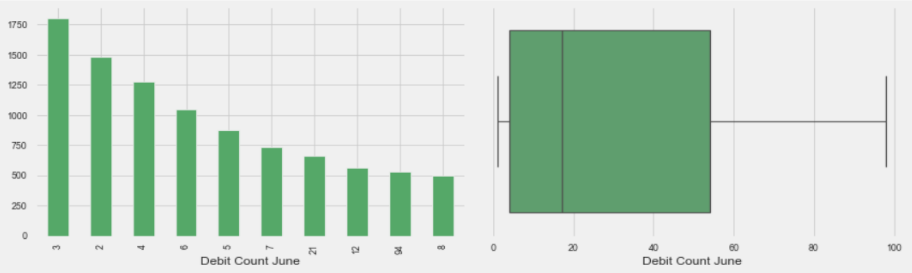




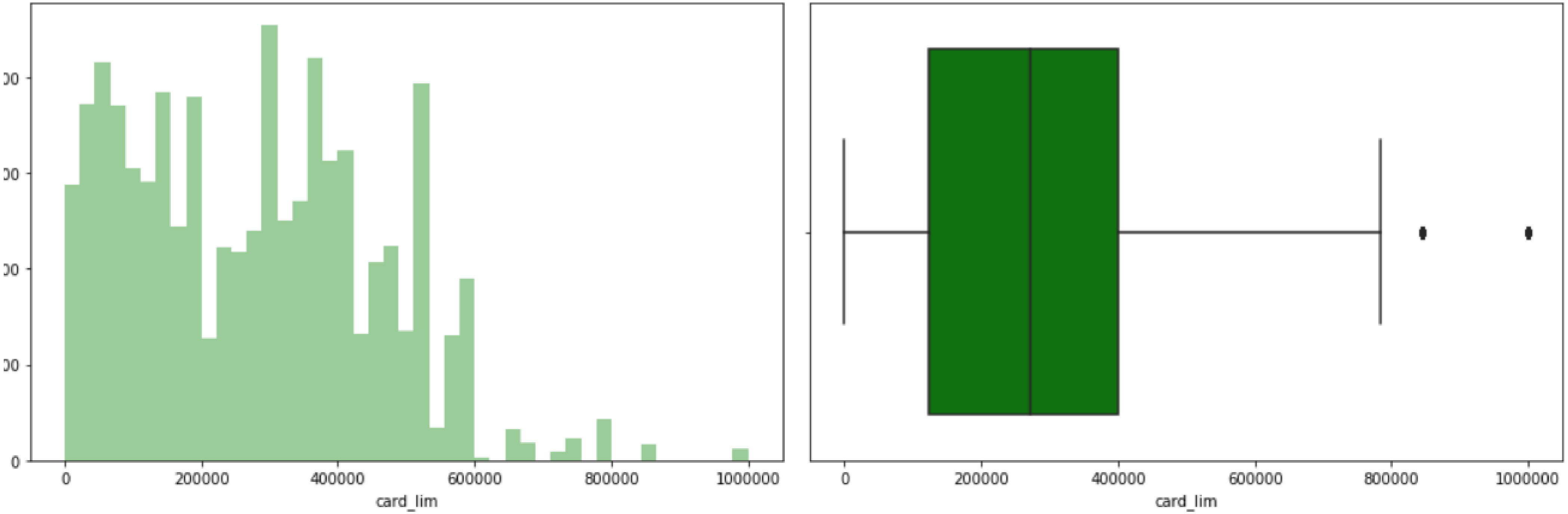






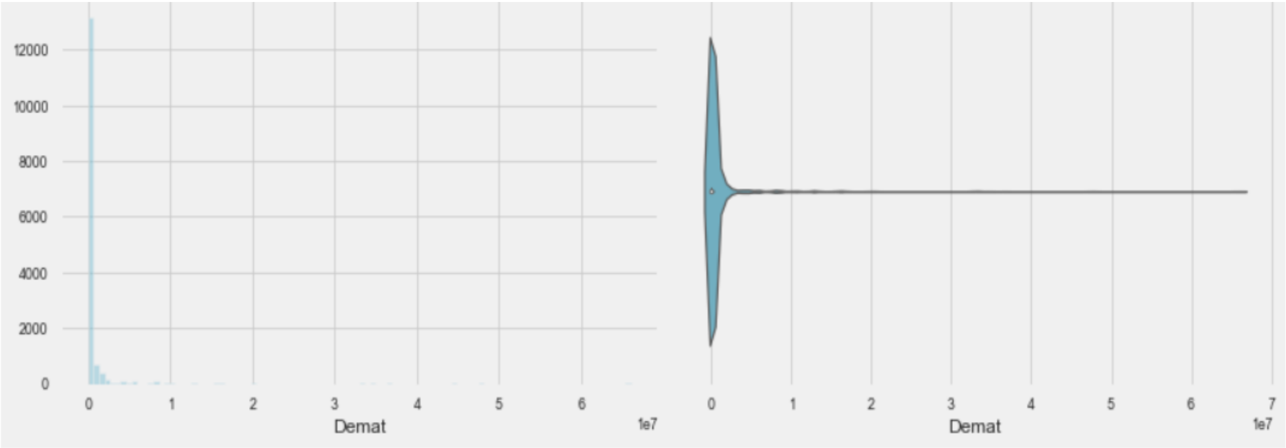


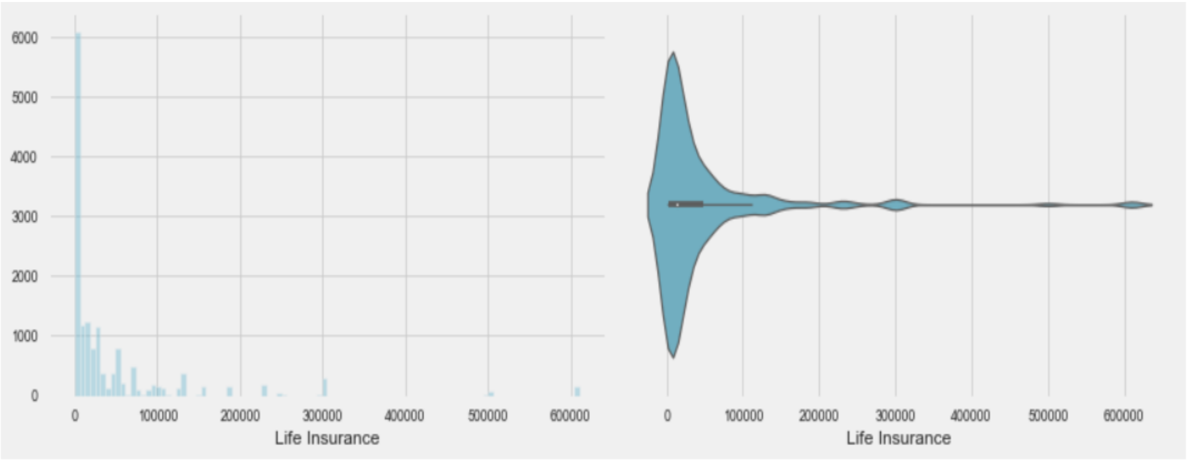
Card Limit

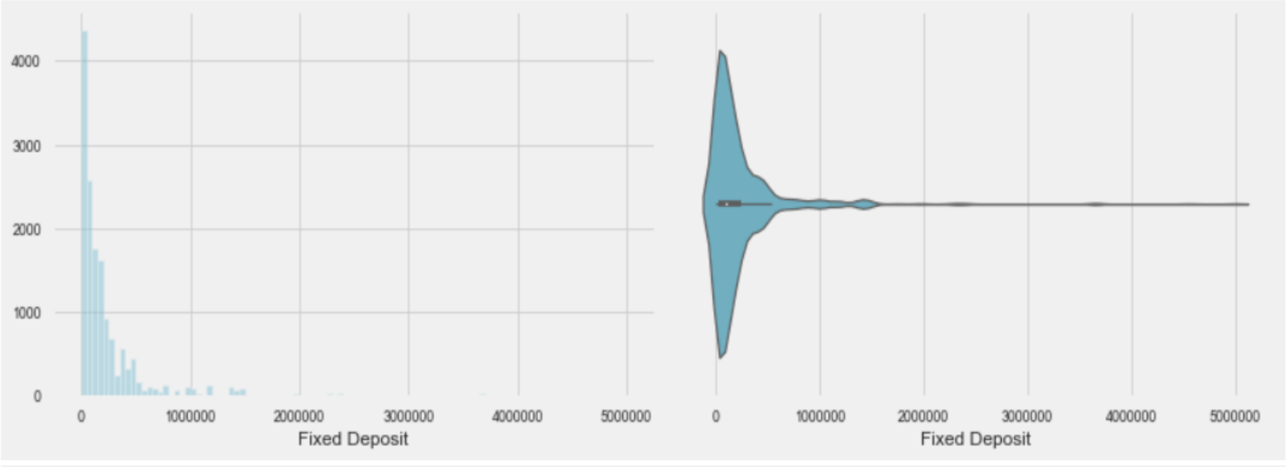


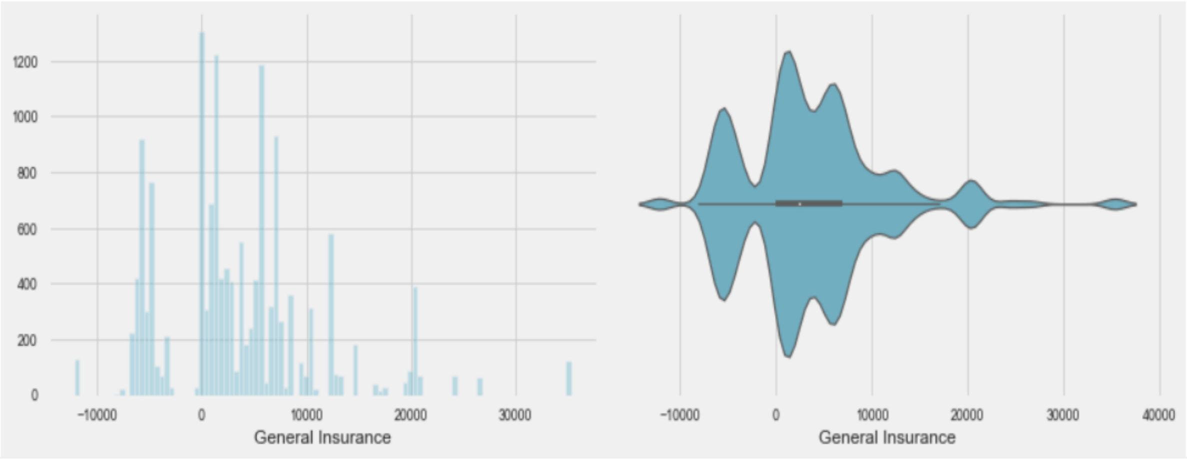
1. The card limit of majority customers is less than 6 lakhs
2. There are some customers(extreme outliers) with maximum card limit upto 10 lakhs
3. There are some customers who have card limit of 0 which is an incorrect value.

Investments

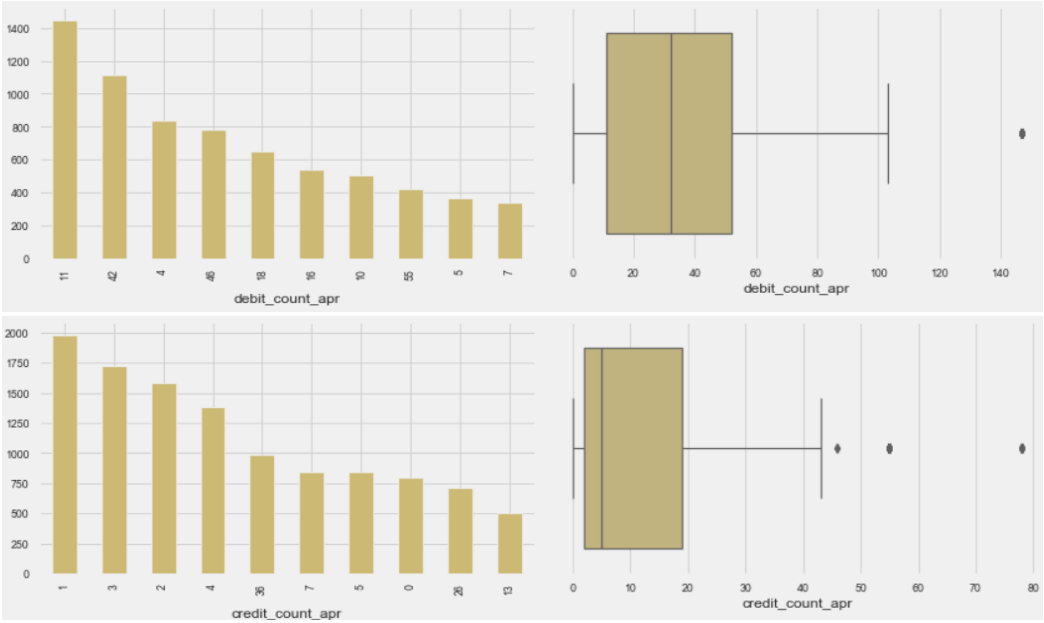


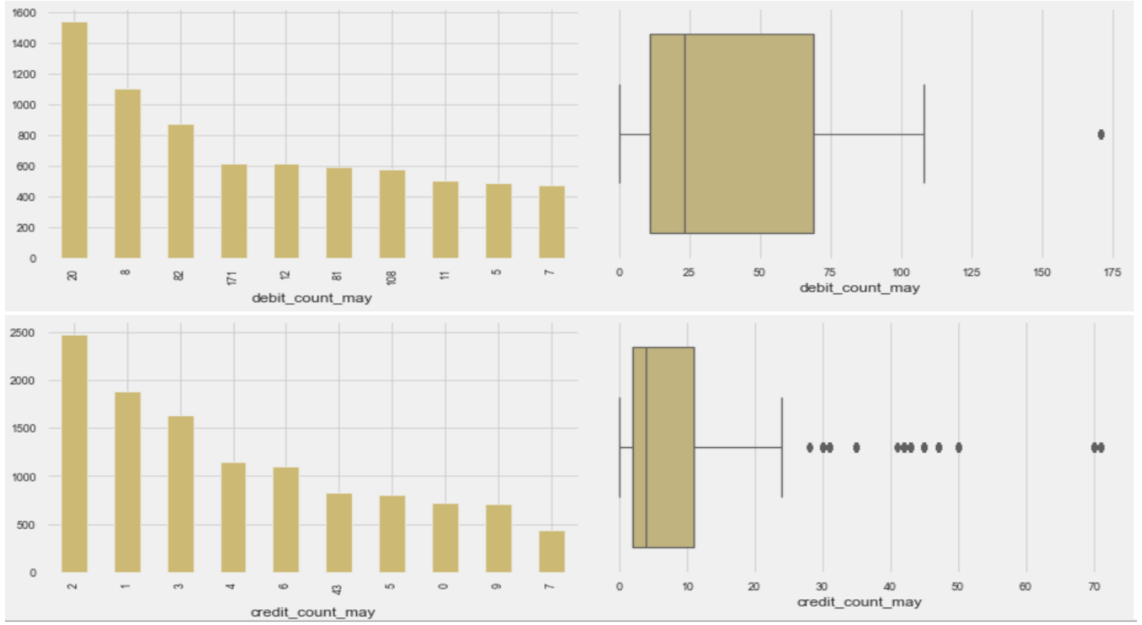


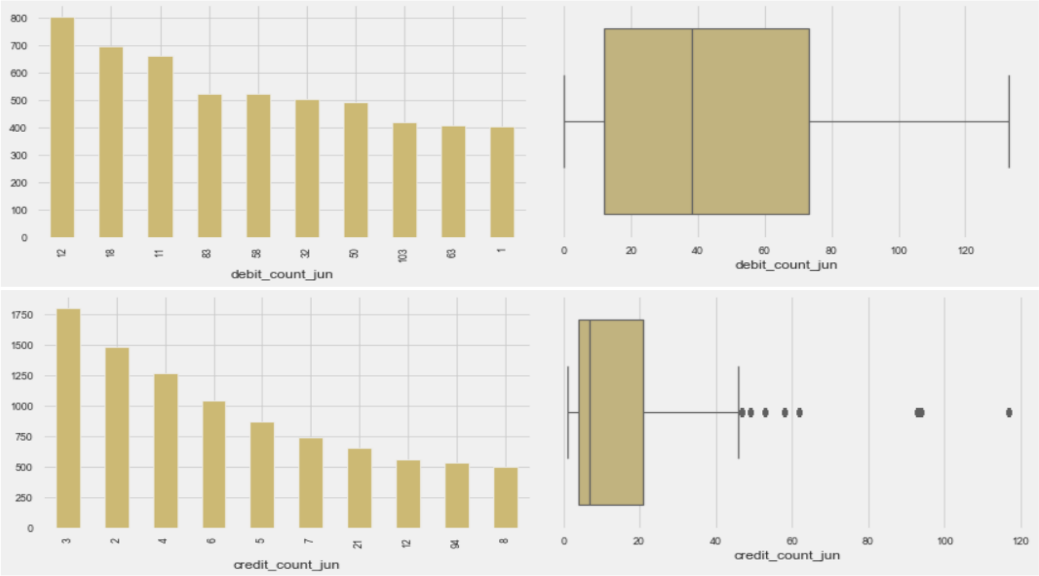




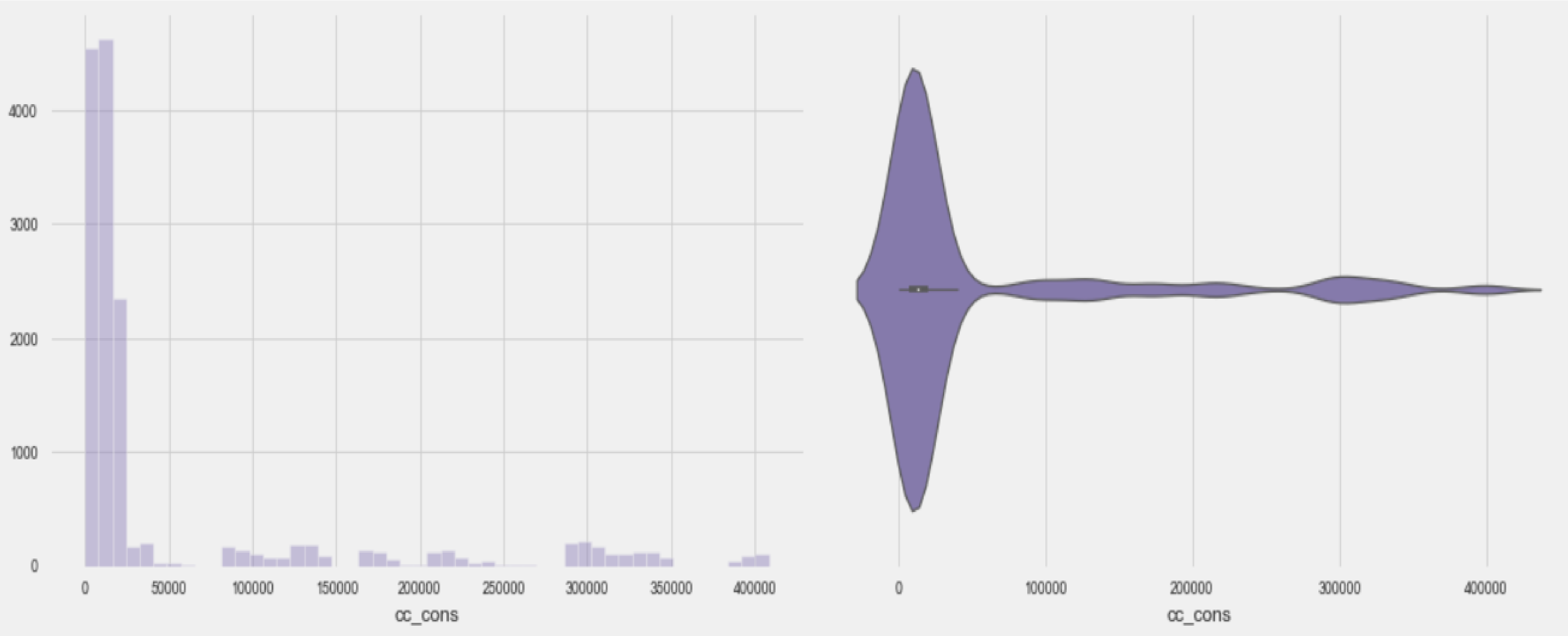
Total number of times amount debited







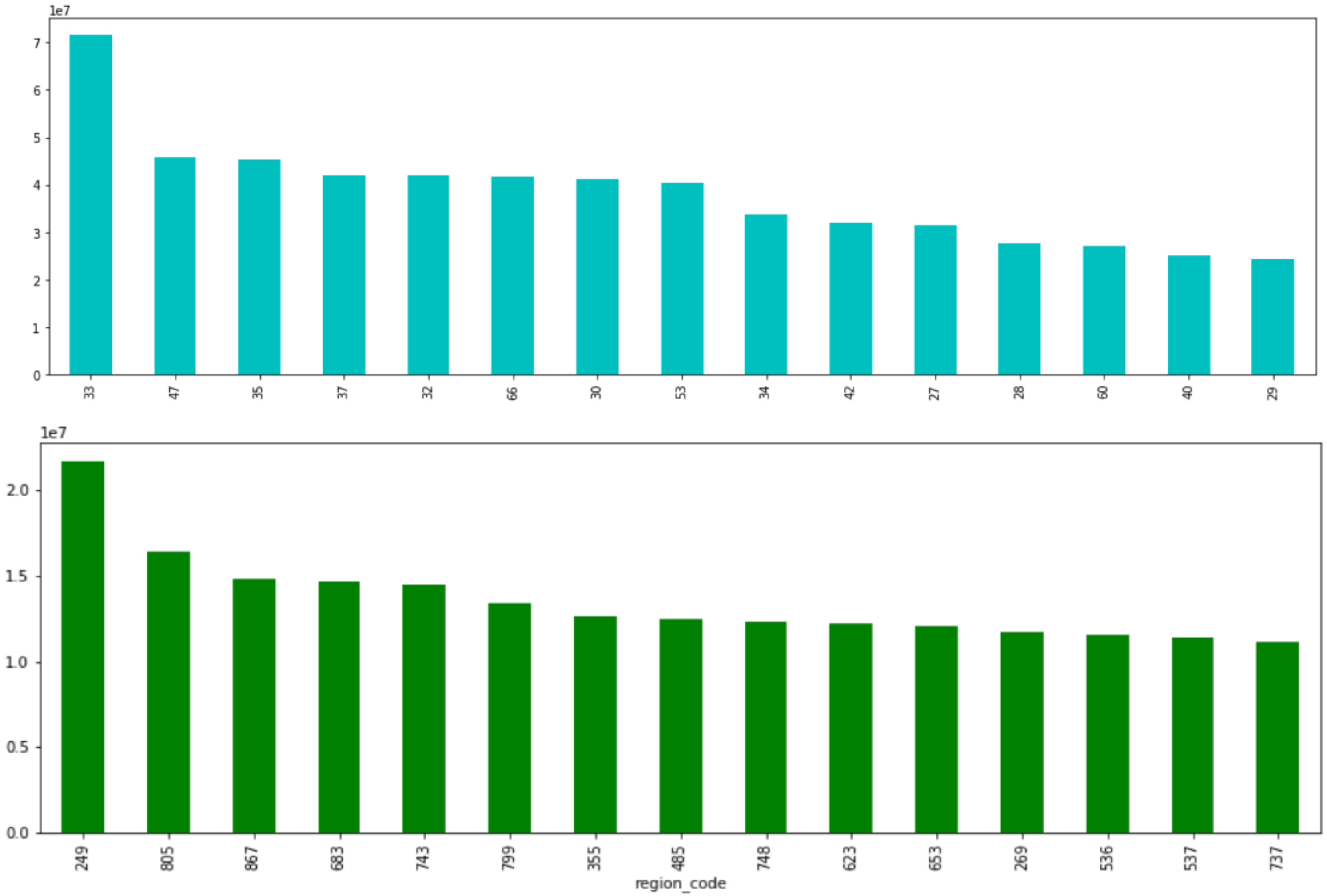
Target Distribution



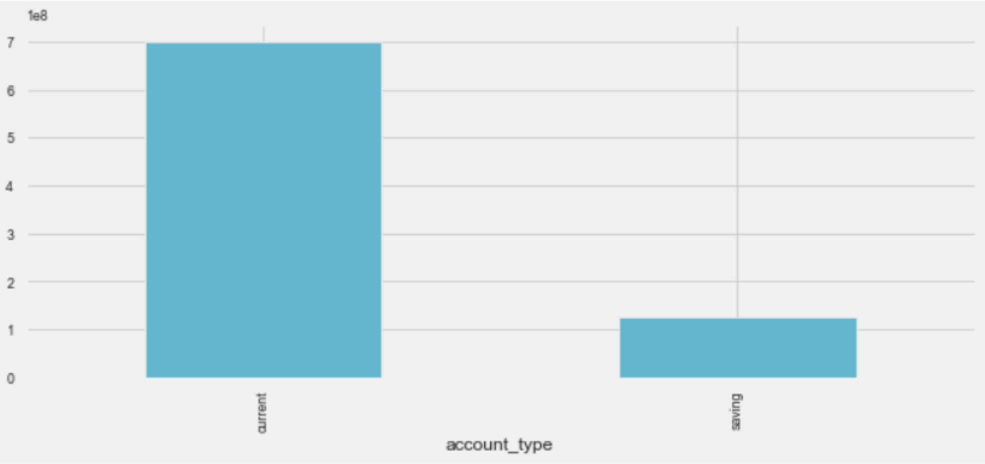
Majority of Target column (cc_cons) values lie in range between 0 to 50000 and the distribution is highly skewed to right

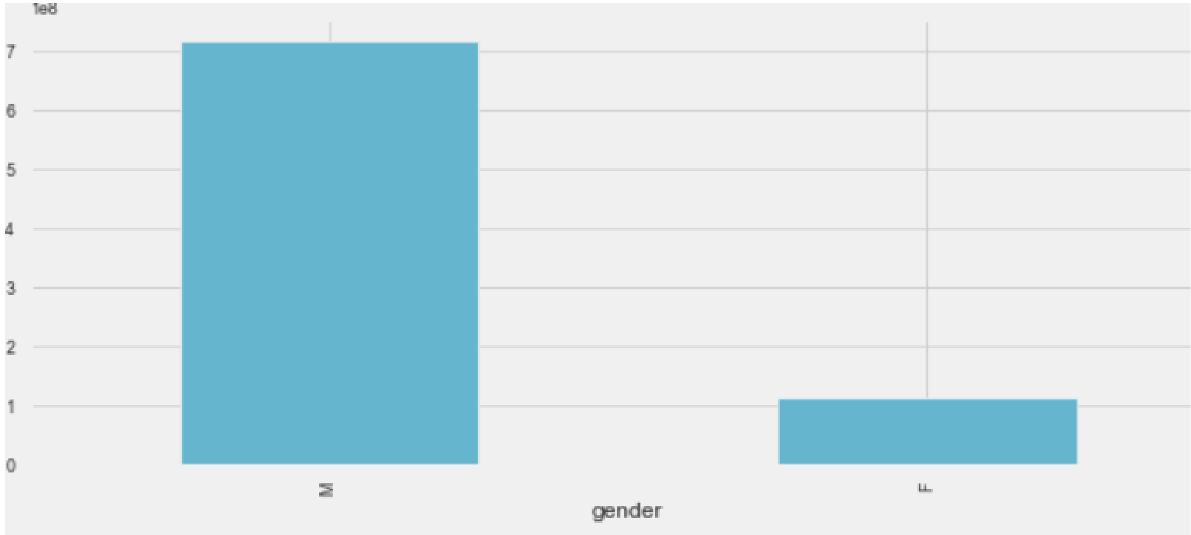
Bivariate Analysis

Age and Region code with Target



Account Type and Gender with respect to Target

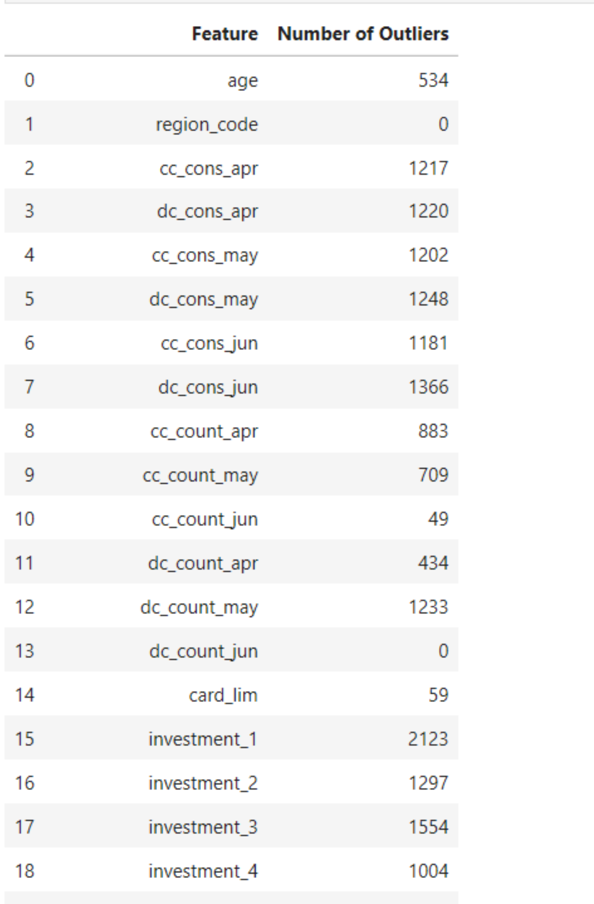


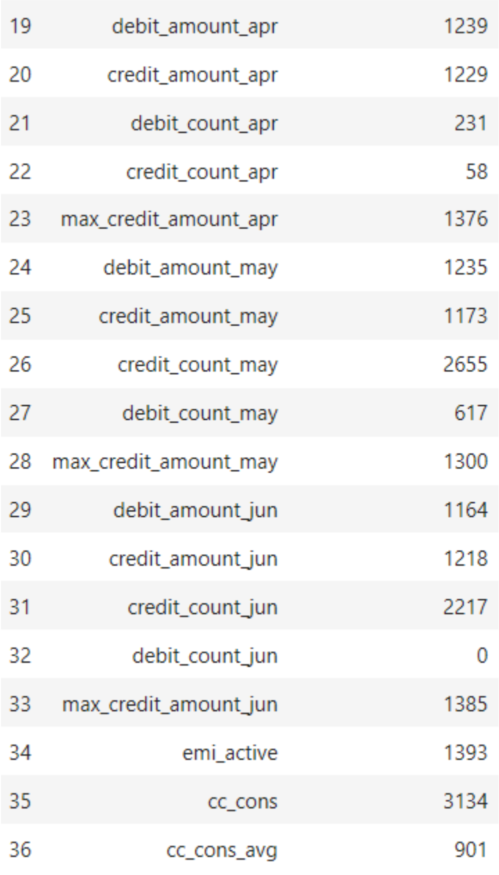


Pipeline

Outlier Treatment :

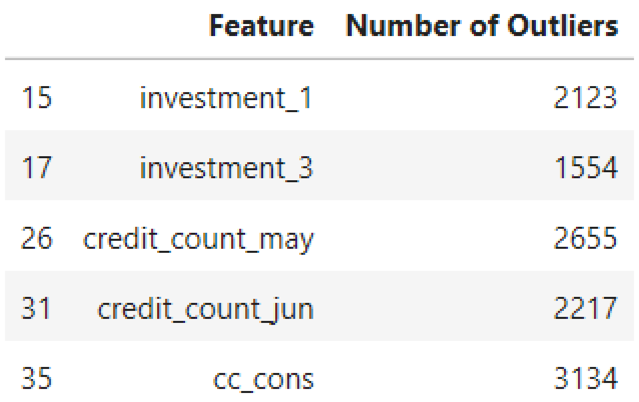
- The Outliers in the continuous features were detected and treated using a method called **Winsorization**.





	Feature	Number of Outliers
0	age	534
1	region_code	0
2	cc_cons_apr	1217
3	dc_cons_apr	1220
4	cc_cons_may	1202
5	dc_cons_may	1248
6	cc_cons_jun	1181
7	dc_cons_jun	1366
8	cc_count_apr	883
9	cc_count_may	709
10	cc_count_jun	49
11	dc_count_apr	434
12	dc_count_may	1233
13	dc_count_jun	0
14	card_lim	59
15	investment_1	2123
16	investment_2	1297
17	investment_3	1554
18	investment_4	1004
19	debit_amount_apr	1239
20	credit_amount_apr	1229
21	debit_count_apr	231
22	credit_count_apr	58
23	max_credit_amount_apr	1376
24	debit_amount_may	1235
25	credit_amount_may	1173
26	credit_count_may	2655
27	debit_count_may	617
28	max_credit_amount_may	1300
29	debit_amount_jun	1164
30	credit_amount_jun	1218
31	credit_count_jun	2217
32	debit_count_jun	0
33	max_credit_amount_jun	1385
34	emi_active	1393
35	cc_cons	3134
36	cc_cons_avg	901

- In winsorization, outliers of some columns could not be treated:



	Feature	Number of Outliers
15	investment_1	2123
17	investment_3	1554
26	credit_count_may	2655
31	credit_count_jun	2217
35	cc_cons	3134

- We treated the predictor outliers in the above table using logarithm,ic and square root transformations
- The outlier in cc_cons(target) was imputed with the mean values of credit consumption columns from the months of April, May and June

➢ There were no trends observed in Personal and Vehicle Loan columns and were dropped from the dataset. Also , loan enquiry had a single value in all rows and hence was dropped.

Feature Selection :

- Following methods were used for feature selection :
  - Correlation
  - RFE

Models and Approaches

- Three vanilla models were assessed without performing any hyperparameter tuning and without treatment
- of class imbalance of the target. The models were:
    - Linear Regression
    - Random Forest Regressor
    - Gradient Boosting Regressor

- None of the three models were not able to give a good RMSLE value.
- This called for performing hyperparameter tuning using Grid Search.

Hyper Parameter Tuned models

- After performing hyperparameter tuning using Grid Search and the following results were observed on the features selected using RFE method.
- Also we tried an Ensemble model of Linear Regression, Random Forest Regressor and Gradient Boosting Regressor.

Models used and their scores

Models	RMSLE
Linear Regression	1.633
Random Forest Regression	1.638
XGBoost Regressor	1.644

Insights & Decisions

Customers to be targeted

- Age : 30 – 40
- Region Code : 400 -800
- Account Type : Current Account holders had the maximum credit consumption and banks should try to retain current account holders. Current holders are mostly held by people in Business. Hence, bank could make efforts towards acquiring customers who own a business.
- The customer spends were highest in the month of May.

THANK YOU