Capstone project

Business Report

By Ruchita Parulekar

Contents:

1) Introduction to the business problem

a) Problem Statement

This business problem is a supervised learning example for a credit card company. The objective is to predict the probability of default (whether the customer will pay the credit card bill or not) based on the variables provided. There are multiple variables on the credit card account, purchase and delinquency information which can be used in the modeling.

PD modeling problems are meant for understanding the riskiness of the customers and how much credit is at stake in case the customer defaults. This is an extremely critical part in any organization that lends money [both secured and unsecured loans].

b) Need of the study/project

Credit card companies operate in a landscape where it is critical to minimize risk while maximizing profitability. The analysis of credit card default probability plays a pivotal role in achieving this balance. By accurately assessing the likelihood of default, companies can proactively manage potential losses, optimize credit limits, and tailor interest rates for individual customers. This analysis empowers companies to make informed decisions, enhance customer relationships, and maintain a healthier bottom line. As a data analyst, delving into this realm provides an opportunity to contribute to the industry's stability and success.

c) Understanding business/social opportunity

In the domain of financial risk assessment, accurately predicting the probability of credit card default is of utmost importance. This project focuses on exploring the predictive potential of data analysis, utilizing Python as the primary tool to analyze historical data records. By developing a strong model, the goal is to precisely forecast the probability of credit card holders defaulting on payments.

2) Data Report

a) Visual inspection of data (rows, columns, descriptive details)

head()

	userid	default	acct_amt_added_12_24m	acct_days_in_dc_12_24m	acct_days_in_rem_12_24m	acct_days_in_term_12_24m	acct_incoming_debt_vs_paid_0_24m
0	4567129.0	0.0	0.0	0.0	0.0	0.0	0.0
1	2635118.0	0.0	0.0	0.0	0.0	0.0	NaN
2	4804232.0	0.0	0.0	0.0	0.0	0.0	NaN
3	1442693.0	0.0	0.0	NaN	NaN	NaN	NaN
4	4575322.0	0.0	0.0	0.0	0.0	0.0	NaN

5 rows × 36 columns

tail()

	userid	default	acct_amt_added_12_24m	acct_days_in_dc_12_24m	acct_days_in_rem_12_24m	acct_days_in_term_12_24m	acct_incoming_debt_vs_paid_0_24
99974	4648093.0	NaN	56102.0	0.0	0.0	0.0	0.0641
99975	1247657.0	NaN	0.0	0.0	0.0	0.0	Na
99976	NaN	NaN	NaN	NaN	NaN	NaN	Na
99977	NaN	NaN	NaN	NaN	NaN	NaN	Na
99978	0.0	10000.0	0.0	11836.0	11836.0	11836.0	59315.0000

5 rows × 36 columns

Shape of the dataset.

```
(99979, 36)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99979 entries, 0 to 99978
Data columns (total 36 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   userid                             99977 non-null  float64
 1   default                            89977 non-null  float64
 2   acct_amt_added_12_24m              99977 non-null  float64
 3   acct_days_in_dc_12_24m             88141 non-null  float64
 4   acct_days_in_rem_12_24m            88141 non-null  float64
 5   acct_days_in_term_12_24m           88141 non-null  float64
 6   acct_incoming_debt_vs_paid_0_24m   40662 non-null  float64
 7   acct_status                        45604 non-null  float64
 8   acct_worst_status_0_3m             45604 non-null  float64
 9   acct_worst_status_12_24m           33216 non-null  float64
 10  acct_worst_status_3_6m             42275 non-null  float64
 11  acct_worst_status_6_12m            39627 non-null  float64
 12  age                                99977 non-null  float64
 13  avg_payment_span_0_12m             76141 non-null  float64
 14  avg_payment_span_0_3m              50672 non-null  float64
 15  merchant_category                  99977 non-null  object
 16  merchant_group                     99968 non-null  object
 17  has_paid                           88943 non-null  float64
 18  max_paid_inv_0_12m                 88943 non-null  float64
 19  max_paid_inv_0_24m                 88943 non-null  float64
 20  name_in_email                      88943 non-null  object
 21  num_active_div_by_paid_inv_0_12m   70052 non-null  float64
 22  num_active_inv                     88943 non-null  float64
 23  num_arch_dc_0_12m                  88943 non-null  float64
 24  num_arch_dc_12_24m                 88943 non-null  float64
 25  num_arch_ok_0_12m                  88943 non-null  float64
 26  num_arch_ok_12_24m                 88943 non-null  float64
 27  num_arch_rem_0_12m                 88943 non-null  float64
 28  status_max_archived_0_6_months     88943 non-null  float64
 29  status_max_archived_0_12_months    88943 non-null  float64
 30  status_max_archived_0_24_months    88943 non-null  float64
 31  recovery_debt                      88943 non-null  float64
 32  sum_capital_paid_acct_0_12m        88943 non-null  float64
 33  sum_capital_paid_acct_12_24m       88943 non-null  float64
 34  sum_paid_inv_0_12m                 88943 non-null  float64
 35  time_hours                         88943 non-null  float64
dtypes: float64(33), object(3)
memory usage: 27.5+ MB
```

- The data frame has 99979 rows and 36 columns, and there are missing values in almost all of the columns.
- It has 3 (Two) Object Data Types and 33 (eight) Float Data Types

	count	mean	std	min	25%	50%	75%	max
userid	99977.0	2.998947e+06	1.154211e+06	0.0	2.000260e+06	2.998815e+06	4.000633e+06	4.999868e+06
default	89977.0	1.254543e-01	3.333776e+01	0.0	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+04
acct_amt_added_12_24m	99977.0	1.225503e+04	3.548133e+04	0.0	0.000000e+00	0.000000e+00	4.937000e+03	1.128775e+06
acct_days_in_dc_12_24m	88141.0	3.573252e-01	4.028733e+01	0.0	0.000000e+00	0.000000e+00	0.000000e+00	1.183600e+04
acct_days_in_rem_12_24m	88141.0	5.178850e+00	4.594340e+01	0.0	0.000000e+00	0.000000e+00	0.000000e+00	1.183600e+04
acct_days_in_term_12_24m	88141.0	4.211774e-01	3.997377e+01	0.0	0.000000e+00	0.000000e+00	0.000000e+00	1.183600e+04
acct_incoming_debt_vs_paid_0_24m	40662.0	2.789992e+00	2.953340e+02	0.0	0.000000e+00	1.520899e-01	6.629926e-01	5.931500e+04
acct_status	45604.0	2.234431e+00	2.546089e+02	1.0	1.000000e+00	1.000000e+00	1.000000e+00	5.437300e+04
acct_worst_status_0_3m	45604.0	2.365165e+00	2.546086e+02	1.0	1.000000e+00	1.000000e+00	1.000000e+00	5.437300e+04
acct_worst_status_12_24m	33216.0	3.347212e+00	3.663034e+02	1.0	1.000000e+00	1.000000e+00	2.000000e+00	6.676100e+04
acct_worst_status_3_6m	42275.0	2.550183e+00	2.806343e+02	1.0	1.000000e+00	1.000000e+00	1.000000e+00	5.770200e+04
acct_worst_status_6_12m	39627.0	2.776062e+00	3.031610e+02	1.0	1.000000e+00	1.000000e+00	1.000000e+00	6.035000e+04
age	99977.0	3.601592e+01	1.300174e+01	0.0	2.500000e+01	3.400000e+01	4.500000e+01	1.000000e+02
avg_payment_span_0_12m	76141.0	1.828429e+01	8.725376e+01	0.0	1.080000e+01	1.490909e+01	2.100000e+01	2.383600e+04
avg_payment_span_0_3m	50672.0	1.596251e+01	2.192071e+02	0.0	8.400000e+00	1.300000e+01	1.828571e+01	4.930500e+04
has_paid	88943.0	8.657455e-01	3.409275e-01	0.0	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
max_paid_inv_0_12m	88943.0	9.362712e+03	1.367242e+04	0.0	2.387500e+03	6.170000e+03	1.140000e+04	2.790000e+05
max_paid_inv_0_24m	88943.0	1.141961e+04	1.543171e+04	0.0	3.685000e+03	7.720000e+03	1.386500e+04	5.385000e+05
num_active_div_by_paid_inv_0_12m	70052.0	3.818166e-01	7.137492e+01	0.0	0.000000e+00	0.000000e+00	1.000000e-01	1.889100e+04
num_active_inv	88943.0	6.259852e-01	1.610529e+00	0.0	0.000000e+00	0.000000e+00	1.000000e+00	4.700000e+01
num_arch_dc_0_12m	88943.0	6.261313e-02	3.821109e-01	0.0	0.000000e+00	0.000000e+00	0.000000e+00	1.700000e+01
num_arch_dc_12_24m	88943.0	5.922894e-02	3.623208e-01	0.0	0.000000e+00	0.000000e+00	0.000000e+00	1.300000e+01
num_arch_ok_0_12m	88943.0	7.786448e+00	1.674280e+01	0.0	0.000000e+00	3.000000e+00	8.000000e+00	2.610000e+02
num_arch_ok_12_24m	88943.0	6.846576e+00	1.606794e+01	0.0	0.000000e+00	2.000000e+00	7.000000e+00	3.130000e+02
num_arch_rem_0_12m	88943.0	4.839954e-01	1.395611e+00	0.0	0.000000e+00	0.000000e+00	0.000000e+00	4.200000e+01
status_max_archived_0_6_months	88943.0	8.217398e-01	7.166438e-01	0.0	0.000000e+00	1.000000e+00	1.000000e+00	3.000000e+00
status_max_archived_0_12_months	88943.0	1.074182e+00	7.763897e-01	0.0	1.000000e+00	1.000000e+00	2.000000e+00	5.000000e+00
status_max_archived_0_24_months	88943.0	1.248193e+00	8.205178e-01	0.0	1.000000e+00	1.000000e+00	2.000000e+00	5.000000e+00
recovery_debt	88943.0	3.601925e+00	1.162108e+02	0.0	0.000000e+00	0.000000e+00	0.000000e+00	1.641100e+04
sum_capital_paid_acct_0_12m	88943.0	1.086026e+04	2.663062e+04	0.0	0.000000e+00	0.000000e+00	8.959500e+03	5.714750e+05
sum_capital_paid_acct_12_24m	88943.0	6.614946e+03	1.924381e+04	0.0	0.000000e+00	0.000000e+00	1.025000e+02	3.418590e+05
sum_paid_inv_0_12m	88943.0	4.103591e+04	9.459642e+04	0.0	3.395500e+03	1.705700e+04	4.573900e+04	2.962870e+06
time_hours	88943.0	1.534165e+01	5.030877e+00	0.0	1.163181e+01	1.580833e+01	1.955417e+01	2.399972e+01

c) Understanding of attributes (variable info, renaming if required)

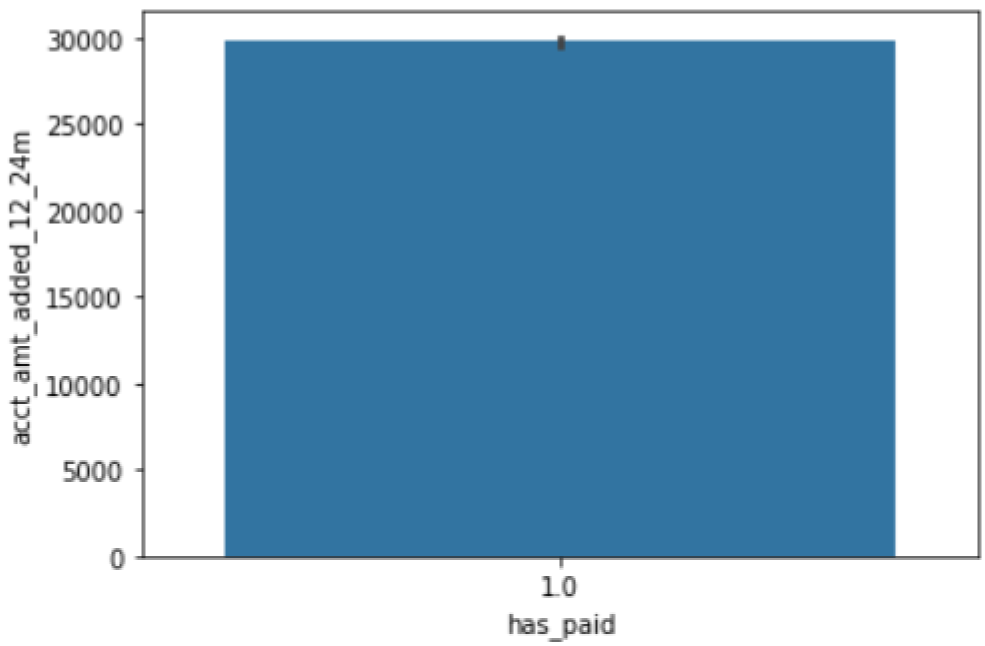
Columns	Description
userid	The unique user id of the customer who is holding the credit card.
default	Target Variable. 1 - Indicates the user has defaulted. 0 - Indicates that the person has not defaulted
acct_amt_added_12_24m	The total amount of the purchases made using the credit card between 24 months in the past to the present date to the 12 months in the past to the current date.
acct_days_in_dc_12_24m	The total number of days that the Credit Card Account has stayed in the Debt-Collection Status between 24 months in the past to the present date to the 12 months in the past to the current date. . Note: Debt-Collection Status: If a Customer has not even paid a minimum amount of the bill, then the account goes into a state called as debt-collection wherein the previous dues from the customer needs to be collected using an agency.
acct_days_in_rem_12_24m	The total number of days that the Credit Card Account has stayed in the Reminder Status between 24 months in the past to the present date to the 12 months in the past to the current date. Note: Reminder Status: If a Customer has not yet paid the Credit Card Bill even after the last due date, the bank used to send reminders for making the payment. If an account starts receiving reminder messages, then it goes to the reminder status.
acct_days_in_term_12_24m	The total number of days that the Credit Card Account has stayed in the Termination Status between 24 months in the past to the present date to the 12 months in the past to the current date. Note: Termination Status: If a Customer has paid the Credit Card Bill even after multiple reminders, then his card gets terminated and he will not be able to make any transactions using the credit card unless it gets activated again.
acct_incoming_debt_vs_paid_0_24m	The ratio of the amount collected out of the total debt in an account by an agency to the total debt amount of the account in the previous 24 months from the current date.

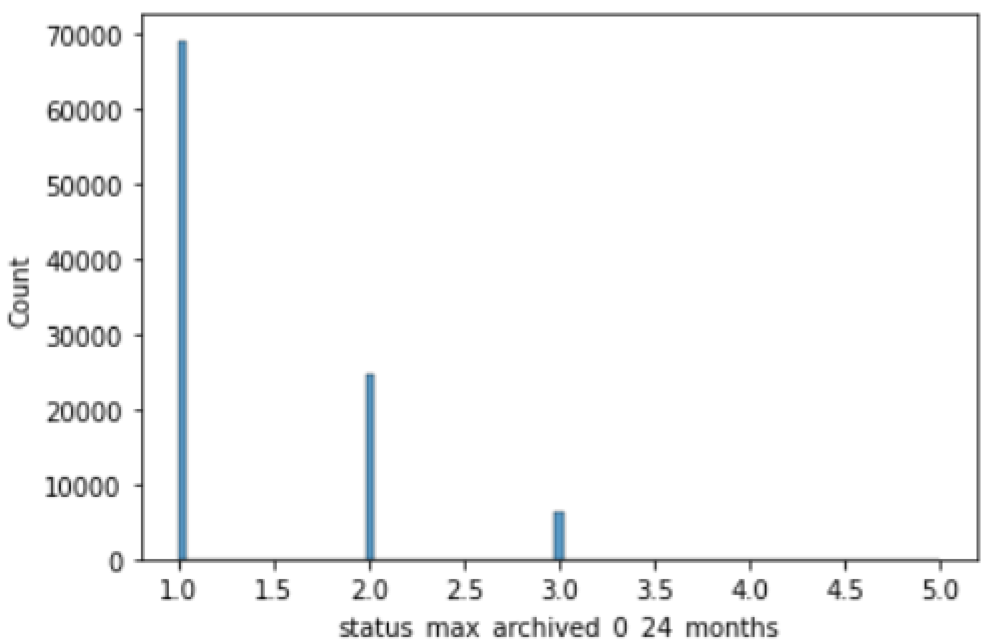
acct_status	The current status of the account. 1 represents an active account, while 0 represents an inactive account.
acct_worst_status_0_3m	The total number of days that the Credit Card Account has stayed in the Worst Status between 3 months in the past to the present date . Note: Worst Status: If a Customer has not even paid a minimum amount of the bill for more than 30 days post the due date, then the account goes into a state called as worst date.
acct_worst_status_12_24m	The total number of days that the Credit Card Account has stayed in the Worst Status between 24 months in the past to the present date and 12 months in the past to the current date .
acct_worst_status_3_6m	The total number of days that the Credit Card Account has stayed in the Worst Status between 6 months in the past to the present date and 3 months in the past to the current date .
acct_worst_status_6_12m	The total number of days that the Credit Card Account has stayed in the Worst Status between 12 months in the past to the present date and 6 months in the past to the current date .
age	The age of the customer.
avg_payment_span_0_12m	The average payment span that the customer has taken in days after the credit card bill was generated in the last one year.
avg_payment_span_0_3m	The average payment span that the customer has taken in days after the credit card bill was generated in the last three months.
merchant_category	The category of the merchant.
merchant_group	The group of the merchants.
has_paid	Whether the customer has paid the current credit card bill or not. True - Paid. False - Unpaid.
max_paid_inv_0_12m	The maximum credit card bill amount that has been paid by the customer in the last one year.
max_paid_inv_0_24m	The maximum credit card bill amount that has been paid by the customer in the last two years.
name_in_email	Name of the customer in email.
num_active_div_by_paid_inv_0_12m	Ratio of the number of unpaid bills to the paid bills in the last one year.

num_active_inv	Number of the active invoices (unpaid bills)
num_arch_dc_0_12m	number of archived purchases that were in debt collection status in the last one year
num_arch_dc_12_24m	number of archived purchases that were in debt collection status between 24 months in the past to the present date and 12 months in the past to the current date .
num_arch_ok_0_12m	number of archived purchases that were paid in the last one year.
num_arch_ok_12_24m	number of archived purchases that were paid between 24 months in the past to the present date and 12 months in the past to the current date .
num_arch_rem_0_12m	number of archived purchases that were in the reminder status in the last one year.
status_max_archived_0_6_months	maximum number of times the account was in archived status in the last 6 months.
status_max_archived_0_12_months	maximum number of times the account was in archived status in the last one year.
status_max_archived_0_24_months	maximum number of times the account was in archived status in the last two years.
recovery_debt	The total amount that has been recovered out of the entire debt amount on the account.
sum_capital_paid_acct_0_12m	sum of principal balance paid on account in the last one year.
sum_capital_paid_acct_12_24m	sum of principal balance paid on account between 24 months in the past to the present date and 12 months in the past to the current date .
sum_paid_inv_0_12m	The total amount of the paid invoices in the last one year.
time_hours	The total hours spent by the customer in purchases made using the credit card.

3. Exploratory Data Analysis

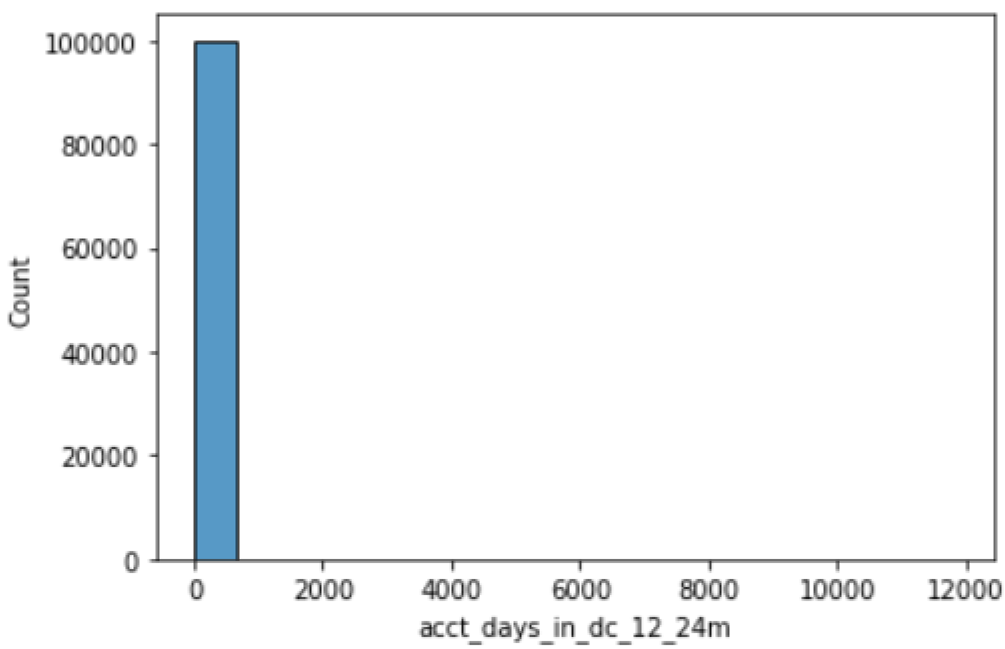
a) Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)





Status_max_archived_0_24_months:

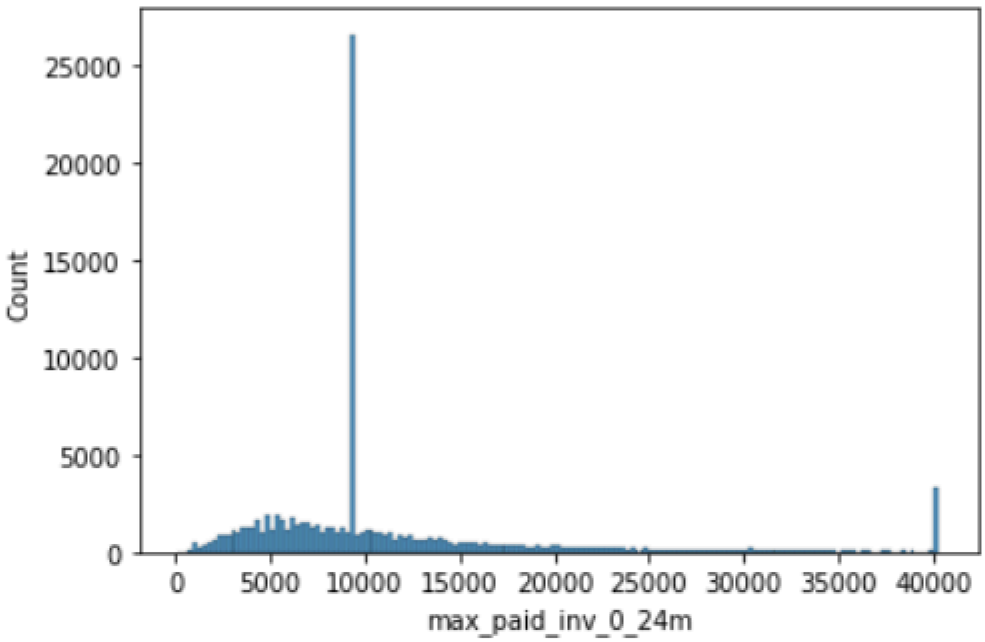
- The data point "status_max_archived_0_24_months" refers to the maximum number of times a specific credit card account was archived within the last two years (24 months).
- The graph suggests that more than 6000 accounts have been archived 1 time during the time frame of 24 months
- About 3000 have been archived 2 times in that period
- **1:** Represents an account that wasn't archived at all in the last two years.
- **3:** Represents an account that was archived three times within the last two years. This would be the maximum number of times archived for that specific account.



Acct_days_in_dc_12_24:

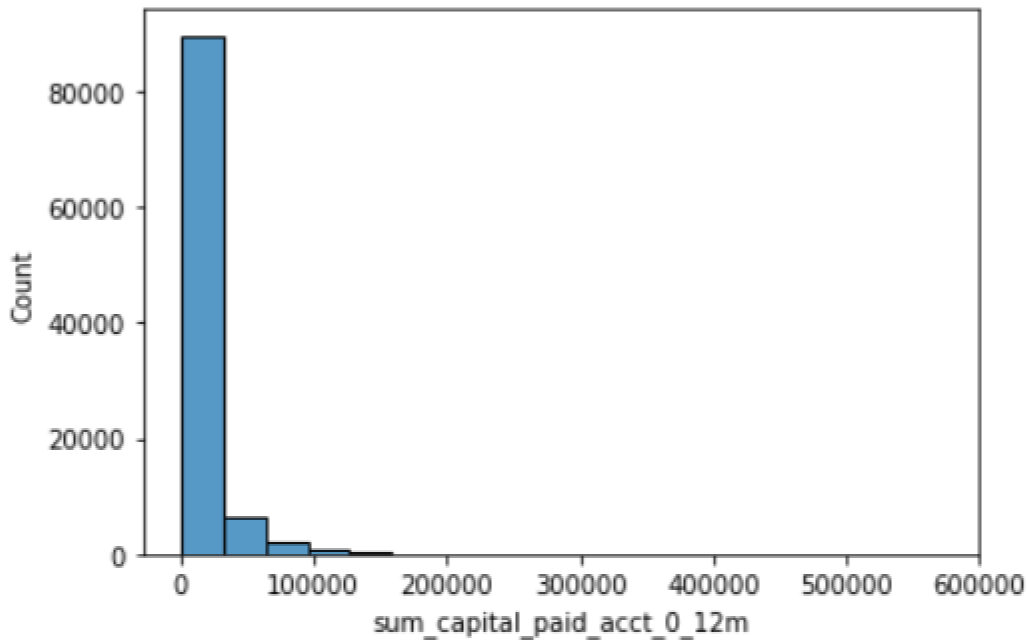
- The graph shows the total number of days that credit card accounts have been in debt collection status over the past 24 months.
- The x-axis labeled "acct_days_in_dc_12-24m" indicates the range of days in debt collection, while the y-axis labeled "Count" represents the number of accounts that fall within that range.
- **Debt Collection Trend:** It appears that there are a higher number of accounts with a lower number of debt collection days (between 0 and 20 days). This could

indicate that these accounts are being addressed relatively quickly by debt collection agencies.



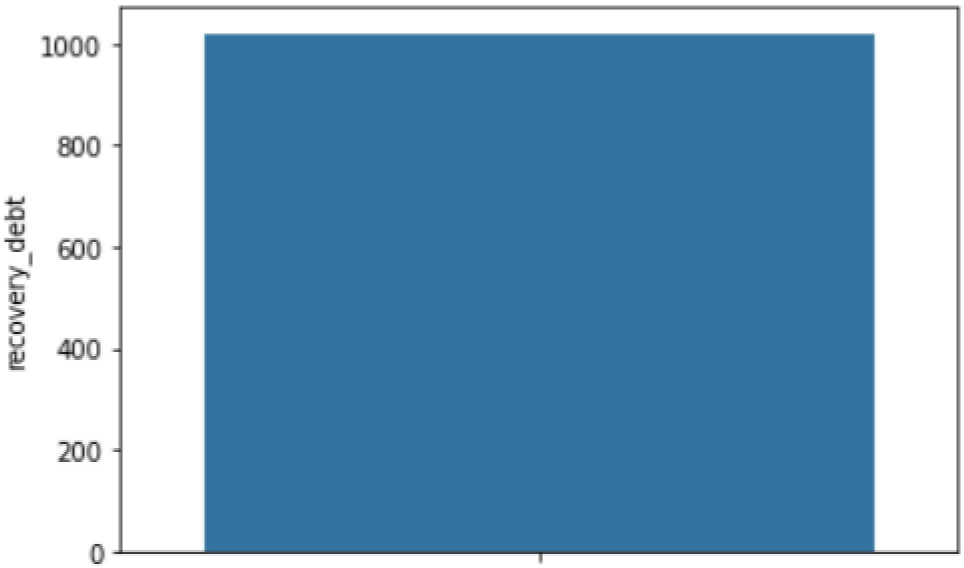
Max_paid_inv_0_24m:

- The data shows the distribution of the maximum credit card bill amounts paid by customers within the last two years (24 months).
- The x-axis represents different ranges of maximum payment amounts (e.g., \$0-\$5,000, \$5,000-\$10,000, etc.), and the y-axis represents the number of customers who paid within those ranges.
- The bar at the lower end (e.g., \$0-\$5,000) might suggest a large portion of customers who typically pay smaller amounts or have low credit limits.
- A high bar in the middle (e.g., \$10,000-\$15,000) could indicate a concentration of customers with moderate credit limits or spending habits.
- The bar at the higher end (e.g., >\$25,000) might represent customers with high credit limits or who make larger purchases occasionally.



Sum_capital_paid_acct_0_12m:

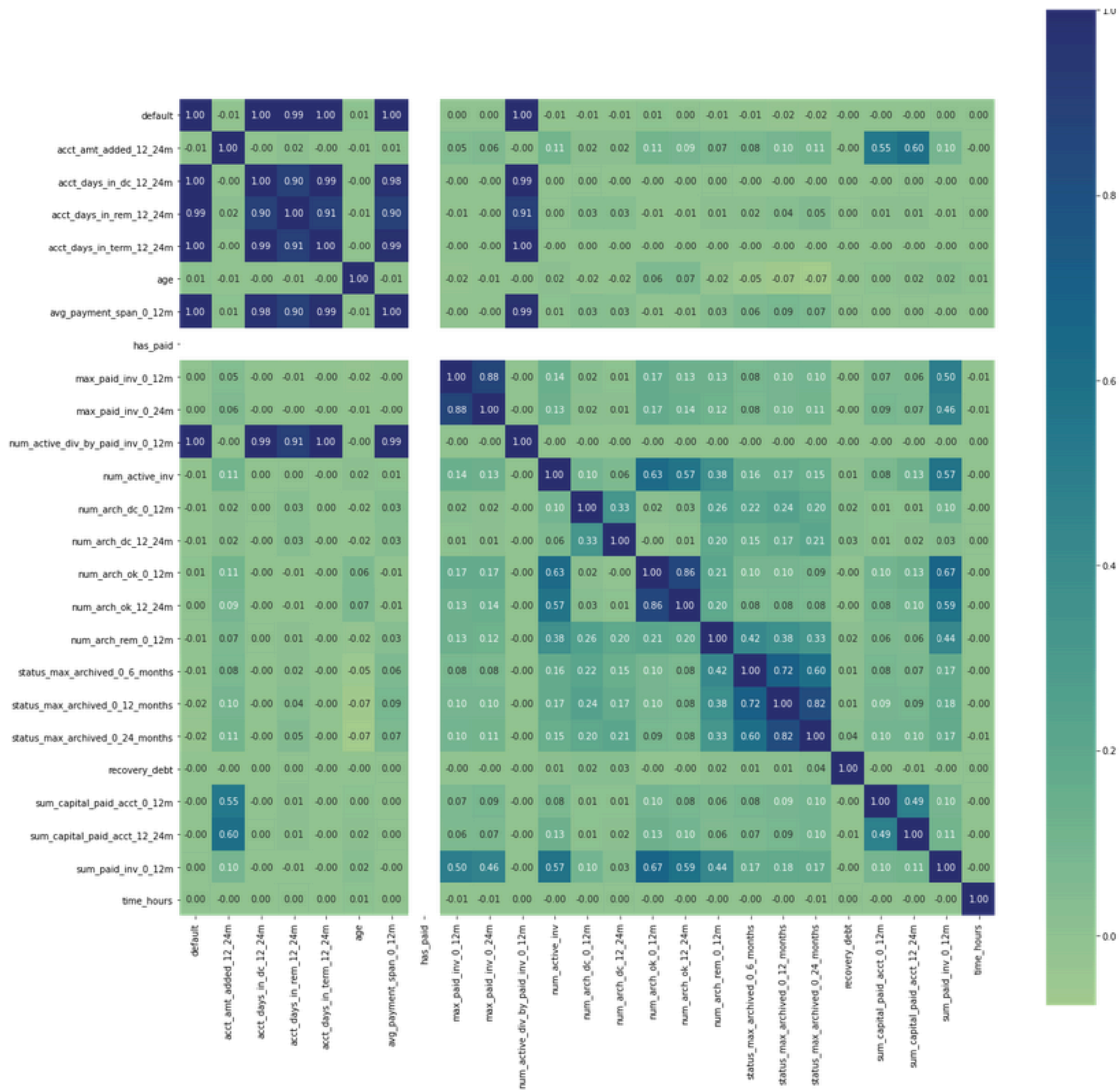
- The x-axis labeled "sum_capital_paid_acct_0_12m" represents the range of total principal paid (likely in dollar amounts), and the y-axis labeled "Count" represents the number of accounts that fall within those ranges.
- **Distribution of Principal Payments:** There appears to be a higher concentration of accounts with lower total payments made in the last year (between $10,000 and $30,000). This could indicate a customer base that carries a significant balance or makes minimum payments.
- **Higher Payment Accounts:** There are also a decent number of accounts with higher total principal payments made within the last year (between $40,000 and $60,000). This segment might represent customers who pay down their balances more aggressively or have higher credit limits.



Recovery_debt:

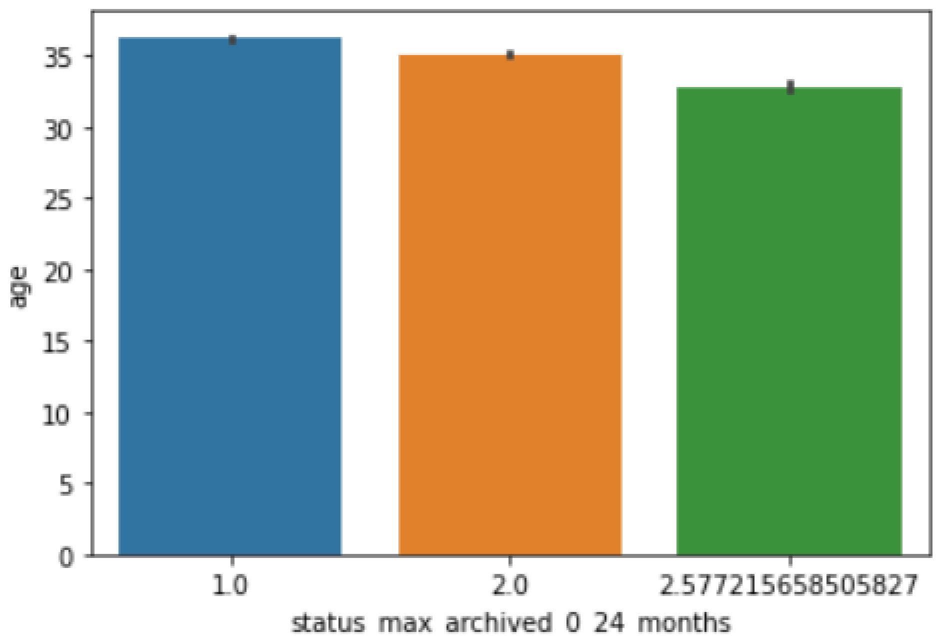
- *This graph represents the total amount recovered of the entire debt amount on the account.*
- *Almost all the accounts have recovered 1018 amount as their recovery debt.*

b) Bivariate analysis (relationship between different variables , correlations)

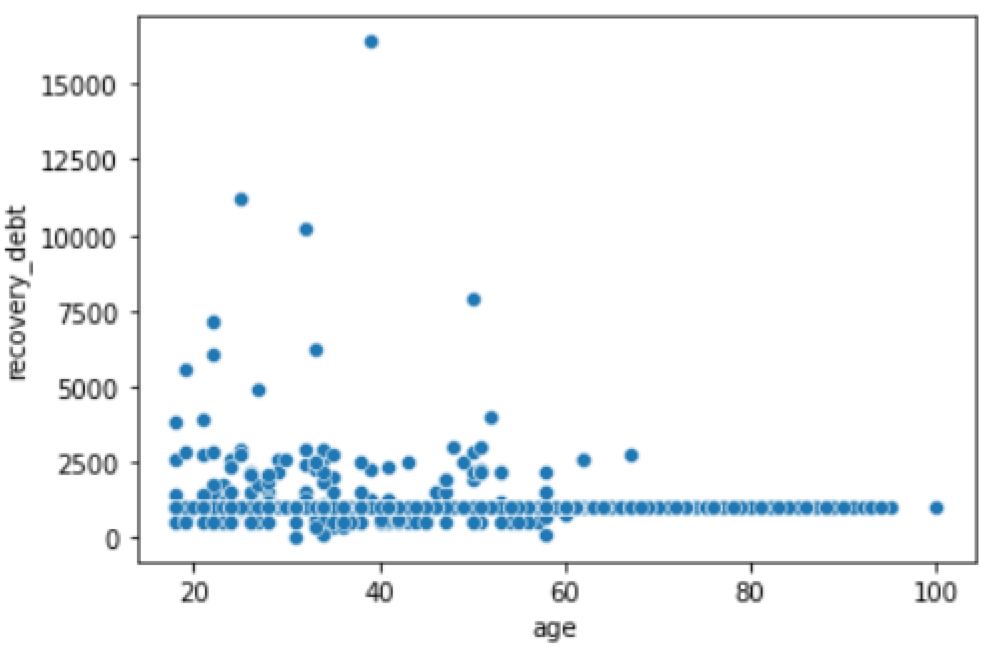


Heatmaps:

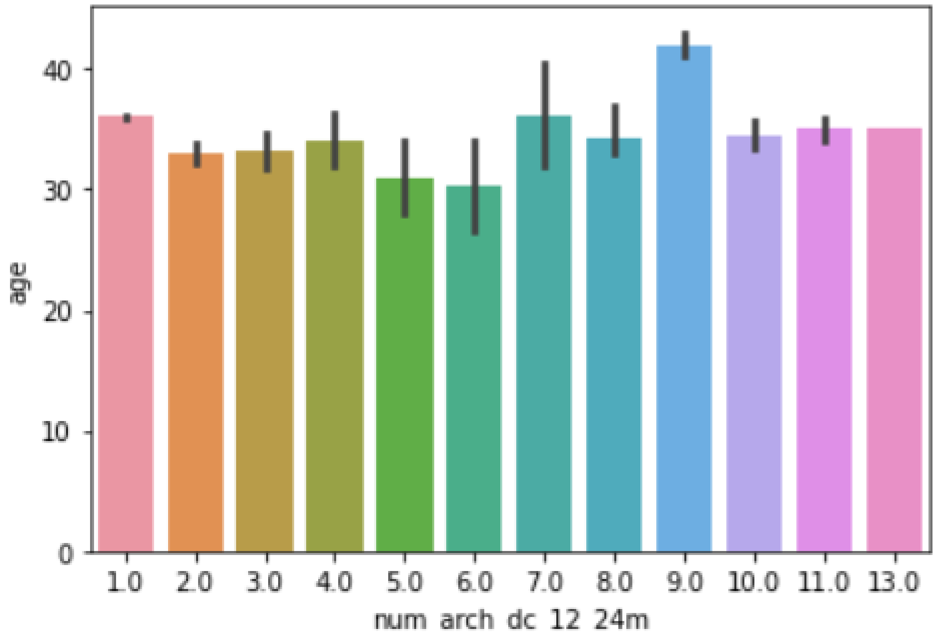
- As we can see in the heat map, there are a few variables that are highly correlated.
- This shows that there is multicollinearity present in the data and thus we will have to reduce it during model building.
- For this, we will use VIF(Variance Inflation Factor) technique.
- VIF is a statistical measure used in regression analysis to assess multicollinearity among independent variables. It measures the degree to which the accuracy of a regression model is affected when independent variables are correlated with each other.



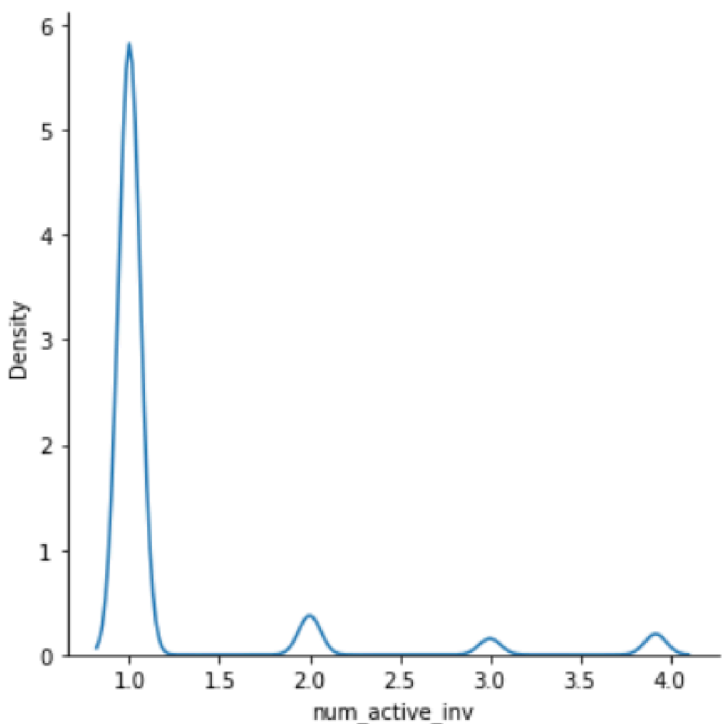
- *Status_max_archived_0_24_months and age:*



- *Age and recovery_debt*



- *Num_arch_dc_12_24m and age:*



- This represents Total number of active invoices per user, I.e. Unpaid bills.
- Looking at the proportion of this variable, 69% users have no active invoices which is a good thing.

- Here, the distribution is highly right skewed. Majority of active invoices count fall in the range of 1 to 5 but there are a lot of outliers with few users having more than 30 active invoices.

c) Removal of unwanted variables (if applicable)

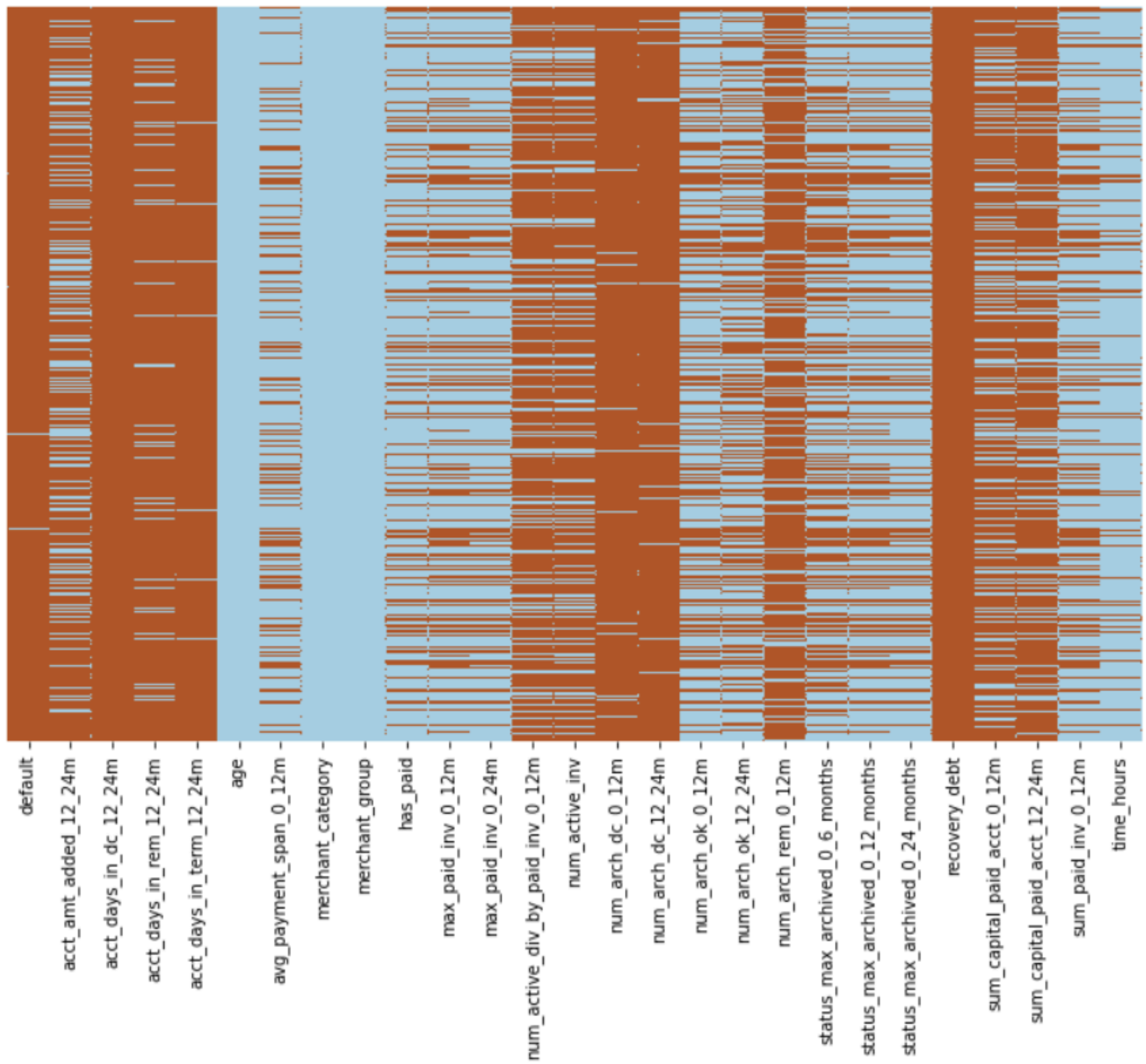
```
MISSING VALUES PER VARIABLE -->


default                              10002
acct_amt_added_12_24m                    2
acct_days_in_dc_12_24m               11838
acct_days_in_rem_12_24m              11838
acct_days_in_term_12_24m             11838
acct_incoming_debt_vs_paid_0_24m     59317
acct_status                          54375
acct_worst_status_0_3m               54375
acct_worst_status_12_24m             66763
acct_worst_status_3_6m               57704
acct_worst_status_6_12m              60352
age                                      2
avg_payment_span_0_12m               23838
avg_payment_span_0_3m                49307
merchant_category                        2
merchant_group                          11
has_paid                             11036
max_paid_inv_0_12m                   11036
max_paid_inv_0_24m                   11036
num_active_div_by_paid_inv_0_12m     29927
num_active_inv                       11036
num_arch_dc_0_12m                    11036
num_arch_dc_12_24m                   11036
num_arch_ok_0_12m                    11036
num_arch_ok_12_24m                   11036
num_arch_rem_0_12m                   11036
status_max_archived_0_6_months       11036
status_max_archived_0_12_months      11036
status_max_archived_0_24_months      11036
recovery_debt                        11036
sum_capital_paid_acct_0_12m          11036
sum_capital_paid_acct_12_24m         11036
sum_paid_inv_0_12m                   11036
time_hours                           11036
dtype: int64
```

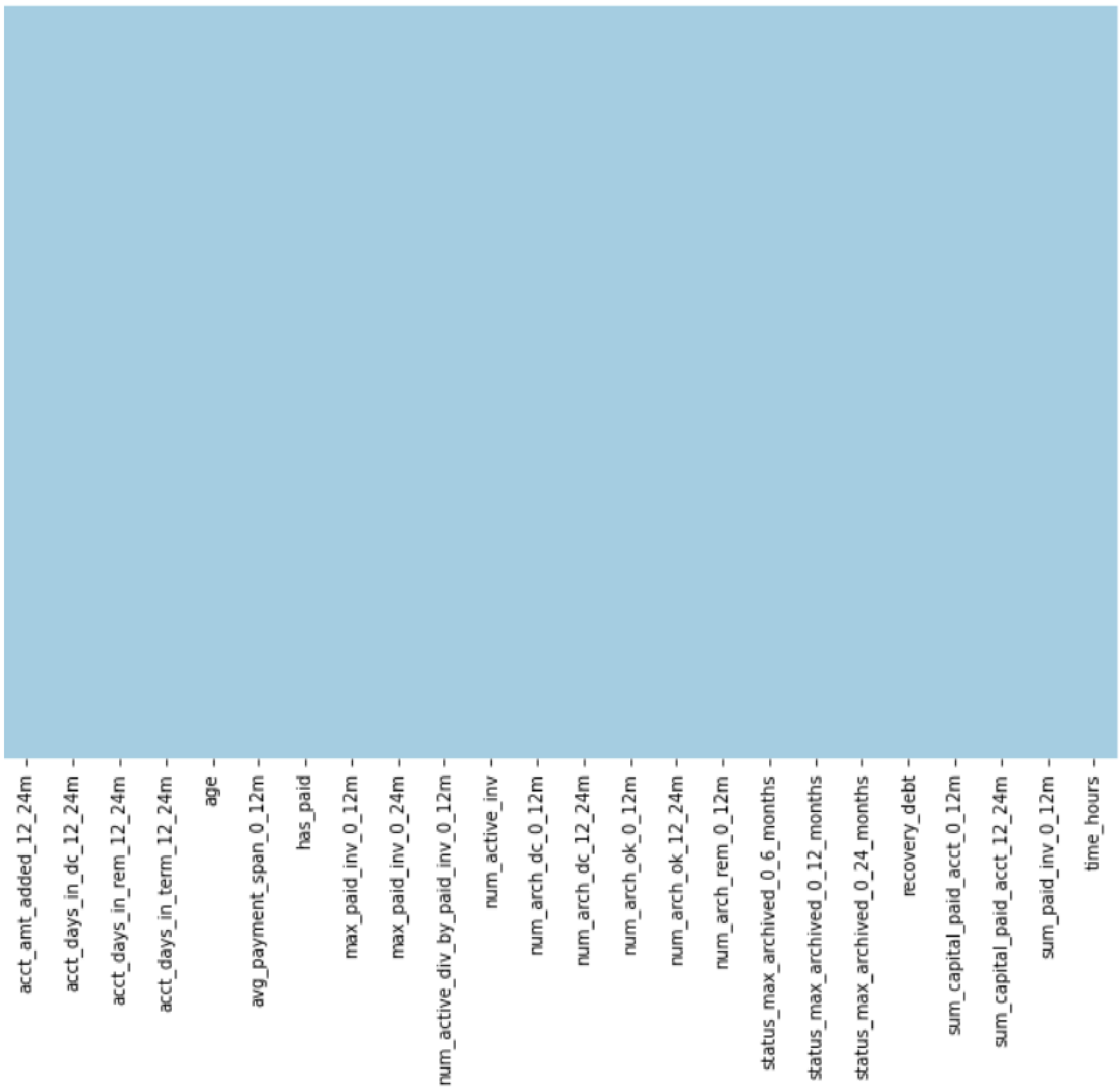
- There are a lot of missing values in the dataset.
- Total missing values is around 20% of the whole dataset. Therefore we should be quite careful as to how we are going to treat these missing values.
- Among them we should drop the variables which have at least 30% of the column as missing values, as imputing such large number of missing values will lead to what we call it an Analyst bias.
- It refers to an approach related bias, since missing values imputation differ from analyst to analyst and there is no sure shot way to do it

```
PERCENT OF MISSING VALUES TO THE WHOLE DATA --->


acct_worst_status_12_24m            66.78
acct_worst_status_6_12m             60.36
acct_incoming_debt_vs_paid_0_24m    59.33
acct_worst_status_3_6m              57.72
acct_status                         54.39
acct_worst_status_0_3m              54.39
avg_payment_span_0_3m               49.32
num_active_div_by_paid_inv_0_12m    29.93
avg_payment_span_0_12m              23.84
acct_days_in_dc_12_24m              11.84
acct_days_in_rem_12_24m             11.84
acct_days_in_term_12_24m            11.84
time_hours                          11.04
max_paid_inv_0_12m                  11.04
sum_paid_inv_0_12m                  11.04
has_paid                            11.04
max_paid_inv_0_24m                  11.04
num_active_inv                      11.04
sum_capital_paid_acct_12_24m        11.04
sum_capital_paid_acct_0_12m         11.04
recovery_debt                       11.04
status_max_archived_0_24_months     11.04
status_max_archived_0_12_months     11.04
status_max_archived_0_6_months      11.04
num_arch_rem_0_12m                  11.04
num_arch_ok_12_24m                  11.04
num_arch_ok_0_12m                   11.04
num_arch_dc_12_24m                  11.04
num_arch_dc_0_12m                   11.04
default                             10.00
merchant_group                       0.01
age                                  0.00
merchant_category                    0.00
acct_amt_added_12_24m                0.00
dtype: float64
```

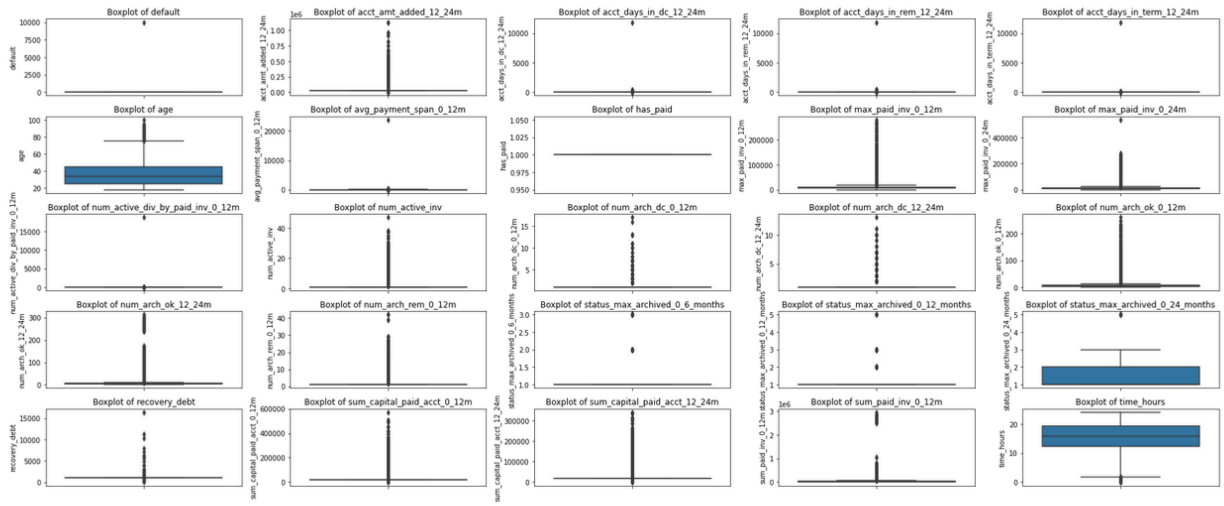


- After dropping the variables with more than 30% missing values, we are left with 24 independent variables.
- Imputing the remaining missing values:
- Since the number of missing values in the 'merchant_category' and 'merchant_group', we will simply drop those missing values.
- The target variable 'default' has around 10% values as missing, so while imputing the remaining missing values, we will impute the target variable using mode value i.e. 0 in this case and resample the data using “SMOTE” later on.
  For other continuous variables we will be imputing the missing values with median value as there are outliers present in the data.
- As we can see, missing values from all the variables have been treated:


acct_amt_added_12_24m
acct_days_in_dc_12_24m
acct_days_in_rem_12_24m
acct_days_in_term_12_24m
age
avg_payment_span_0_12m
has_paid
max_paid_inv_0_12m
max_paid_inv_0_24m
num_active_div_by_paid_inv_0_12m
num_active_inv
num_arch_dc_0_12m
num_arch_dc_12_24m
num_arch_ok_0_12m
num_arch_ok_12_24m
num_arch_rem_0_12m
status_max_archived_0_6_months
status_max_archived_0_12_months
status_max_archived_0_24_months
recovery_debt
sum_capital_paid_acct_0_12m
sum_capital_paid_acct_12_24m
sum_paid_inv_0_12m
time_hours

d) Outlier treatment (if required)

<u>Checking the outliers per column using boxplot</u>



- All the variables have outliers except 'time_hours'.
- Some variables have one or two outliers while some have a large number of outliers.
- Thus, we are going to use the Interquartile-range (IQR) method for treating the outliers.
- All the outliers got treated using the IQR method, but some of the variables were found to have only zeroes left while some had only ones left after the outlier treatment. Therefore after dropping such variables, finally we were left with 14 Variables.