Exploratory Data Analysis for Machine Learning

Course Project

Evgeny Zorin

12. September 2021

Brief Description of the Data

The bank manager is concerned that more and more customers are dropping out of credit card services. He would really appreciate it if someone could analyze the data to find out the reason for the churn. Thus, the bank could use this analysis to predict customers who are likely to be churned out so they could proactively reach out to these customers and provide them with better services, which would likely change their minds.

The dataset was taken from Kaggle and contains **10127** different rows, each one representing a unique client.

The author recommends to drop the last two columns (Naive Bayes Classifier) since they don't provide any useful information for the study.

The remaining 21 columns are as follows:

Column	Description
CLIENTNUM	Client identification number
Attrition_Flag	Weather the customer account has been closed
Customer_Age	Age of a customer
Gender	Customer gender
Dependent_Count	Number of dependents
Education_Level	Education Qualification of account holder
Marital_Status	Married, single, divorced, or unknown
Income_Category	Annual income category of a customer
Card_Category	Type of card
Month_on_book	Period of relationship with the bank
Total_Relationship_Count	Total number of products hold by the customer
Month_Inactive_12_mon	Number of months inactive in the last 12 months
Contacts_Count_12_mon	Number of contacts in the last 12 months
Credit_Limit	Credit limit on the credit card
Total_Revolving_Bal	Total revolving balance on the card
Avg_Open_To_Buy	Last 12 months average of open to buy credit line
Total_Amt_Chng_Q4_Q1	Change in transaction amount (Q4 over Q1)
Total_Trans_Amt	Total transaction amount in last 12 months
Total_Trans_Ct	Total transaction count in last 12 months
Total_Ct_Chng_Q4_Q1	Change in transaction count (Q4 over Q1)
Avg_Utilization_Ratio	Average card utilization ratio

I also decided to drop the first column (CLIENTNUM) since it doesn't provide any useful information either.

Data Summary

- The data contains 6 categorical features and 14 numerical features.
- There are no missing or duplicated values in the data.
- Only 16% of the customers are labeled as ”Attrited Customer”.

Initial Plan for Data Exploration

The goal is to create a model that is able to predict when a customer is going to leave a service. Logistic regression classifier is most commonly used in such cases. Successful completion of the task assumes that the used dataset will not generate erroneous predictions. It is necessary to assure that the dataset consists of numerical values and that there are no outliers.

The plan is as follows:

1. Examining summary statistics of the data.
2. Examining distribution of numerical features.
3. Outliers Handling using 1.5IQR rule (if needed).
4. Examining distribution of categorical features.
5. Feature Engineering, such as binary, ordinal and one-hot encoding.
6. Log Transformation of skewed features (if needed).
7. Examining Pair Plots and Correlation Maps of the features.
8. Dropping some correlated features from the dataset (if needed).

Actions taken for data cleaning and feature engineering

Examining summary statistics of the data

- Summary statistics of numerical features:

	count	mean	std	min	25%	50%	75%	max
Customer_Age	10127.0	46.325960	8.016814	26.0	41.000	46.000	52.000	73.000
Dependent_count	10127.0	2.346203	1.298908	0.0	1.000	2.000	3.000	5.000
Months_on_book	10127.0	35.928409	7.986416	13.0	31.000	36.000	40.000	56.000
Total_Relationship_Count	10127.0	3.812580	1.554408	1.0	3.000	4.000	5.000	6.000
Months_Inactive_12_mon	10127.0	2.341167	1.010622	0.0	2.000	2.000	3.000	6.000
Contacts_Count_12_mon	10127.0	2.455317	1.106225	0.0	2.000	2.000	3.000	6.000
Credit_Limit	10127.0	8631.953698	9088.776650	1438.3	2555.000	4549.000	11067.500	34516.000
Total_Revolving_Bal	10127.0	1162.814061	814.987335	0.0	359.000	1276.000	1784.000	2517.000
Avg_Open_To_Buy	10127.0	7469.139637	9090.685324	3.0	1324.500	3474.000	9859.000	34516.000
Total_Amt_Chng_Q4_Q1	10127.0	0.759941	0.219207	0.0	0.631	0.736	0.859	3.397
Total_Trans_Amt	10127.0	4404.086304	3397.129254	510.0	2155.500	3899.000	4741.000	18484.000
Total_Trans_Ct	10127.0	64.858695	23.472570	10.0	45.000	67.000	81.000	139.000
Total_Ct_Chng_Q4_Q1	10127.0	0.712222	0.238086	0.0	0.582	0.702	0.818	3.714
Avg_Utilization_Ratio	10127.0	0.274894	0.275691	0.0	0.023	0.176	0.503	0.999

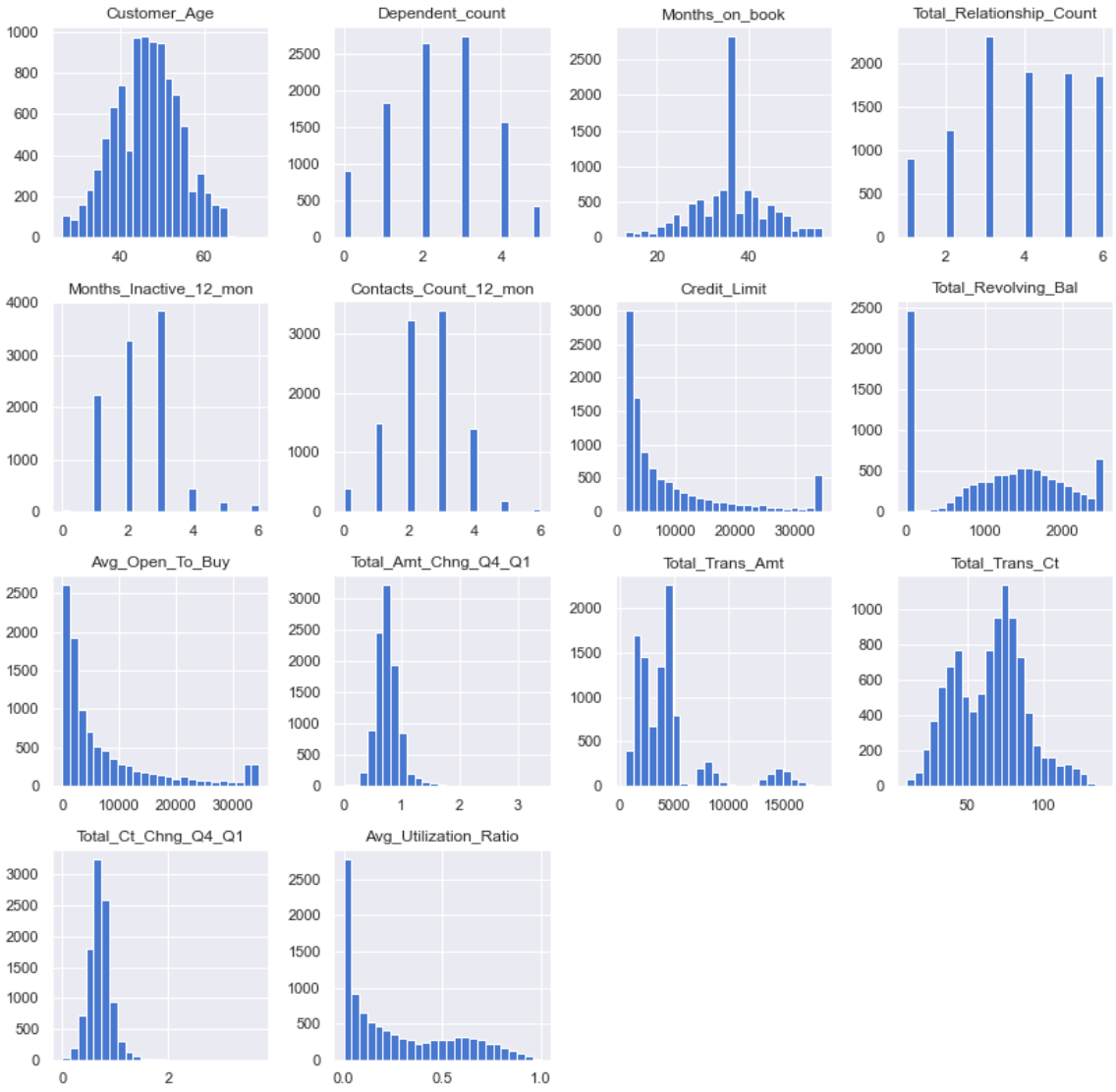
The summary statistics of numerical features indicates that most of the features have a wide range and large standard deviation. Some features, such as Total_Amt_Chng_Q4_Q1 and Credit_Limit, have significantly different magnitudes.

- Summary statistics of categorical features:

	count	unique	top	freq
Attrition_Flag	10127	2	Existing Customer	8500
Gender	10127	2	F	5358
Education_Level	10127	7	Graduate	3128
Marital_Status	10127	4	Married	4687
Income_Category	10127	6	Less than $40K	3561
Card_Category	10127	4	Blue	9436

Examining distribution of numerical features

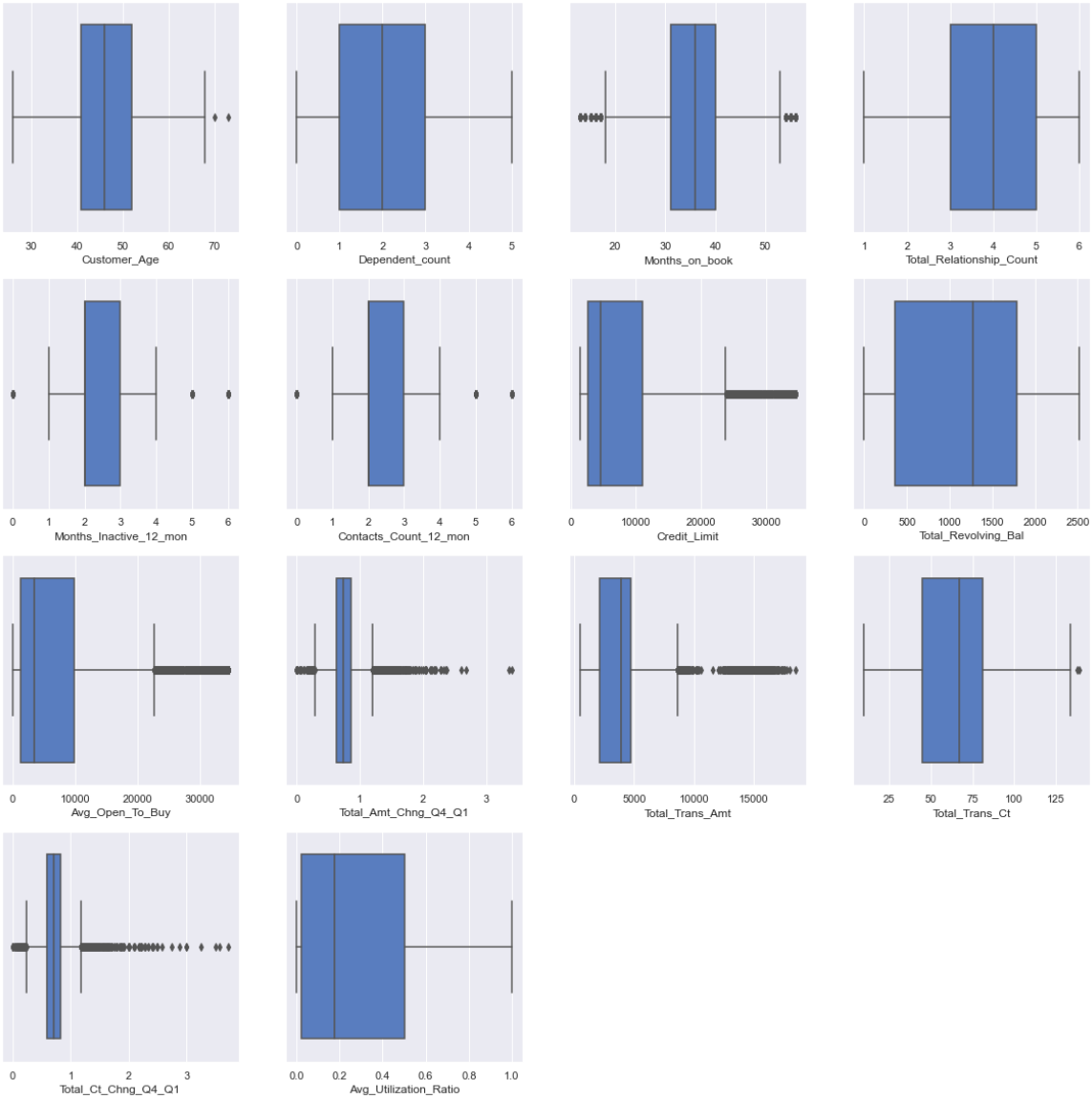
- Distributions of Numerical Features:



It can be assumed that there are potential outliers in the features.

A quick check has confirmed the assumptions. There are outliers in almost all numerical features.

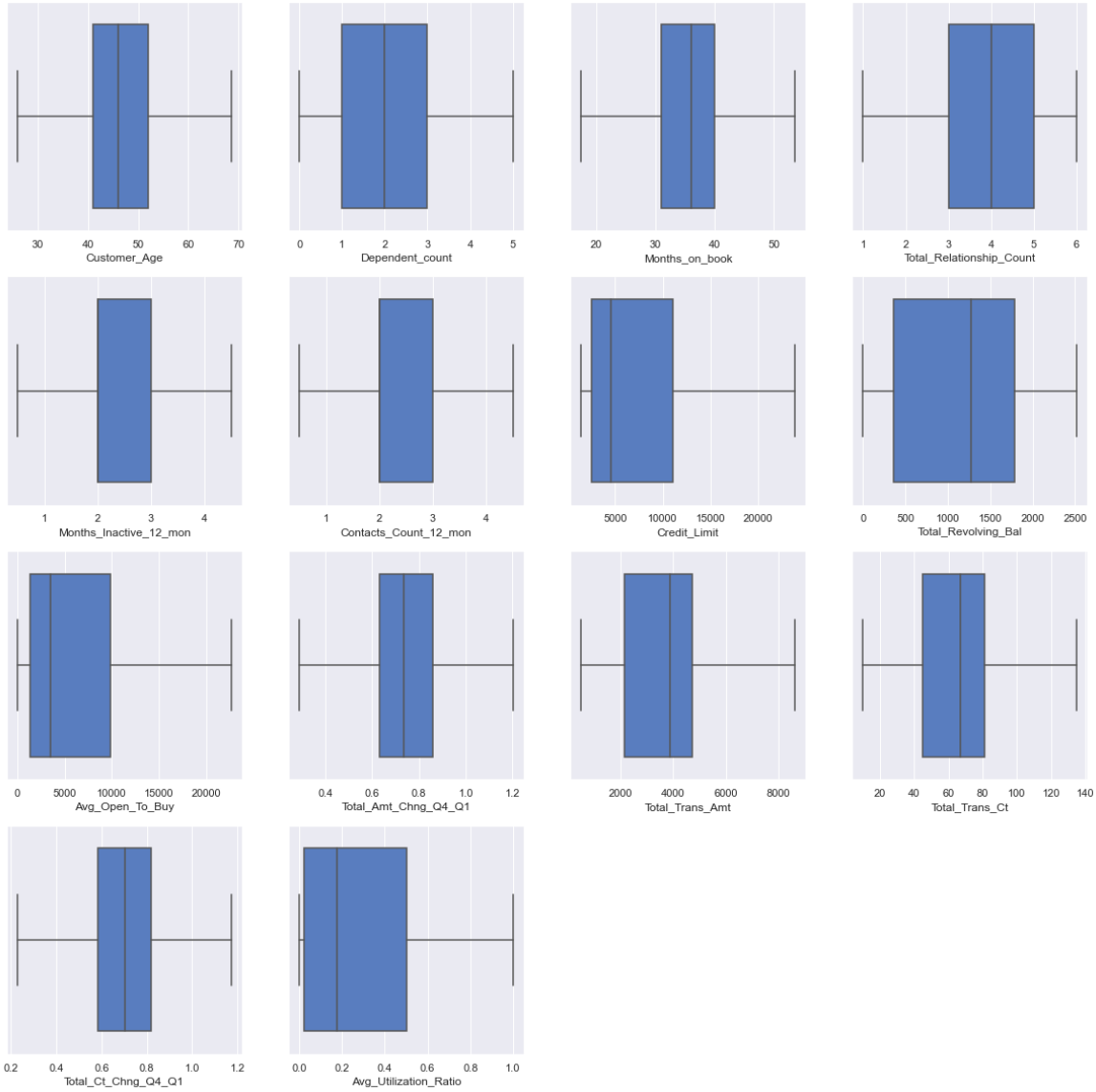
- Distribution of outliers in numerical features:



Outliers Handling using 1.5IQR rule

The 1.5IQR rule will be used to solve the problem of outliers in the data. This rule finds the lower and upper bounds for outliers in an array.
After processing the outliers, it is necessary to perform an additional check to make sure that the numerical features no longer contains outliers.

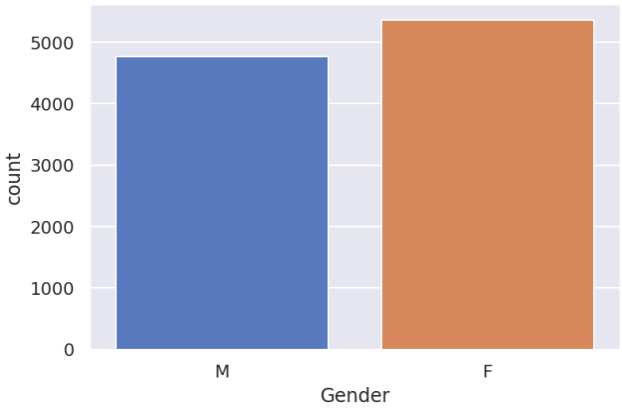
- Confirmation of data purity in numerical features:



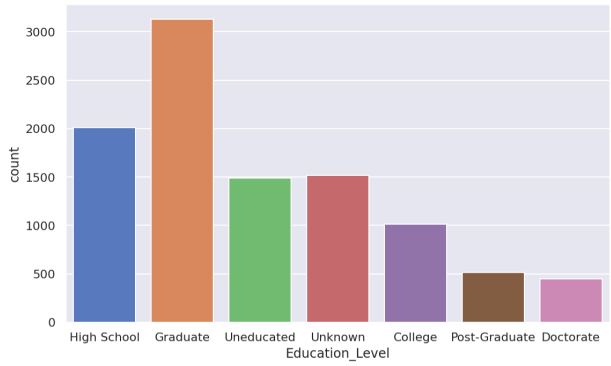
After examining the data in the numerical features, it is time to examine the categorical features.

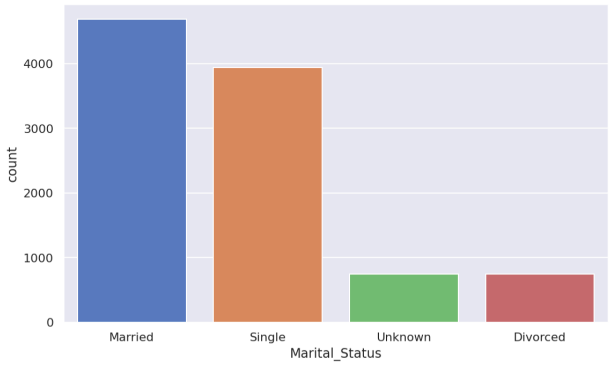
Examining distribution of categorical features

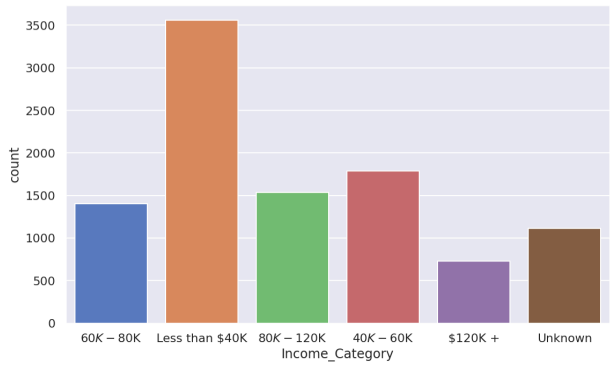
- Distribution of Attrition Flag and Gender:

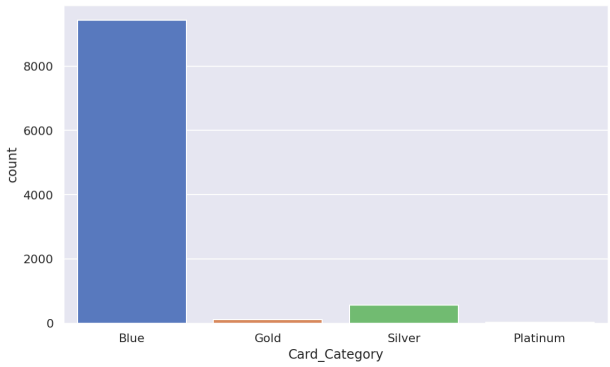


- Distribution of other categorical features:









The distribution of the label indicates that the dataset is unbalanced, whereas the gender distribution is almost balanced. Most of the customers hold the Blue card, are Married or Single, have a Graduate degree, and their annual income is less than $40K.

Feature Engineering

The dataset contains 6 categorical features. The features, such as Attrition_Flag and Gender, have only two possible values. It will be performed a binary encoding on them. The features, such as Marital_Status and Card_Category, have 4 categories. It will be performed an ordinal encoding on them. The features, such as Income_Category and Educational_Level, have 6 and 7 categories accordingly. It will be performed a one-hot encoding on them.

- Binary Encoding:

  Attrition_Flag - Existing Customer = 0, Attrited Customer = 1

  Gender - Male = 0, Female = 1

- Ordinal Encoding:

  Marital_Status - Married = 0, Single = 1, Divorced = 2, Unknown = 3

  Card_Category - Blue = 0, Silver = 1, Gold = 2, Platinum = 3

- One-Hot Encoding:

  The **Income_Category** and **Educational_Level** features have been separated into individual columns for each category. While the original feature columns have been removed.

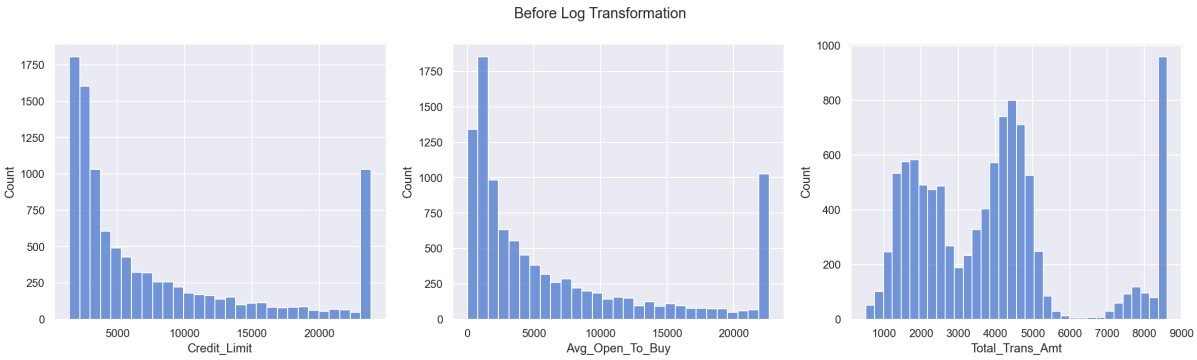
Log Transformation of skewed features

Since some of the distributions of numerical features appear to be right-skewed, Log Transformation will be performed on the features that have skewness greater than 0.75.

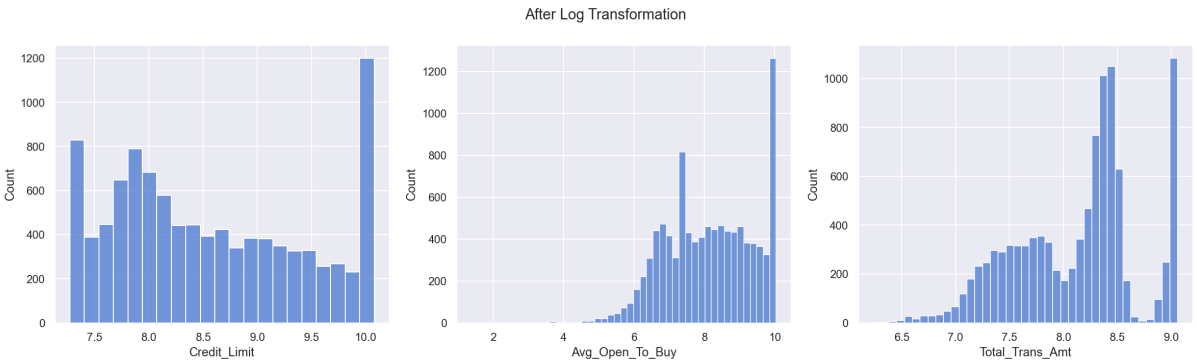
- Distribution of skewed features:

	Skew
Credit_Limit	1.197249
Avg_Open_To_Buy	1.190498
Total_Trans_Amt	0.837030

- Visualization of skewed features before Log Transformation:



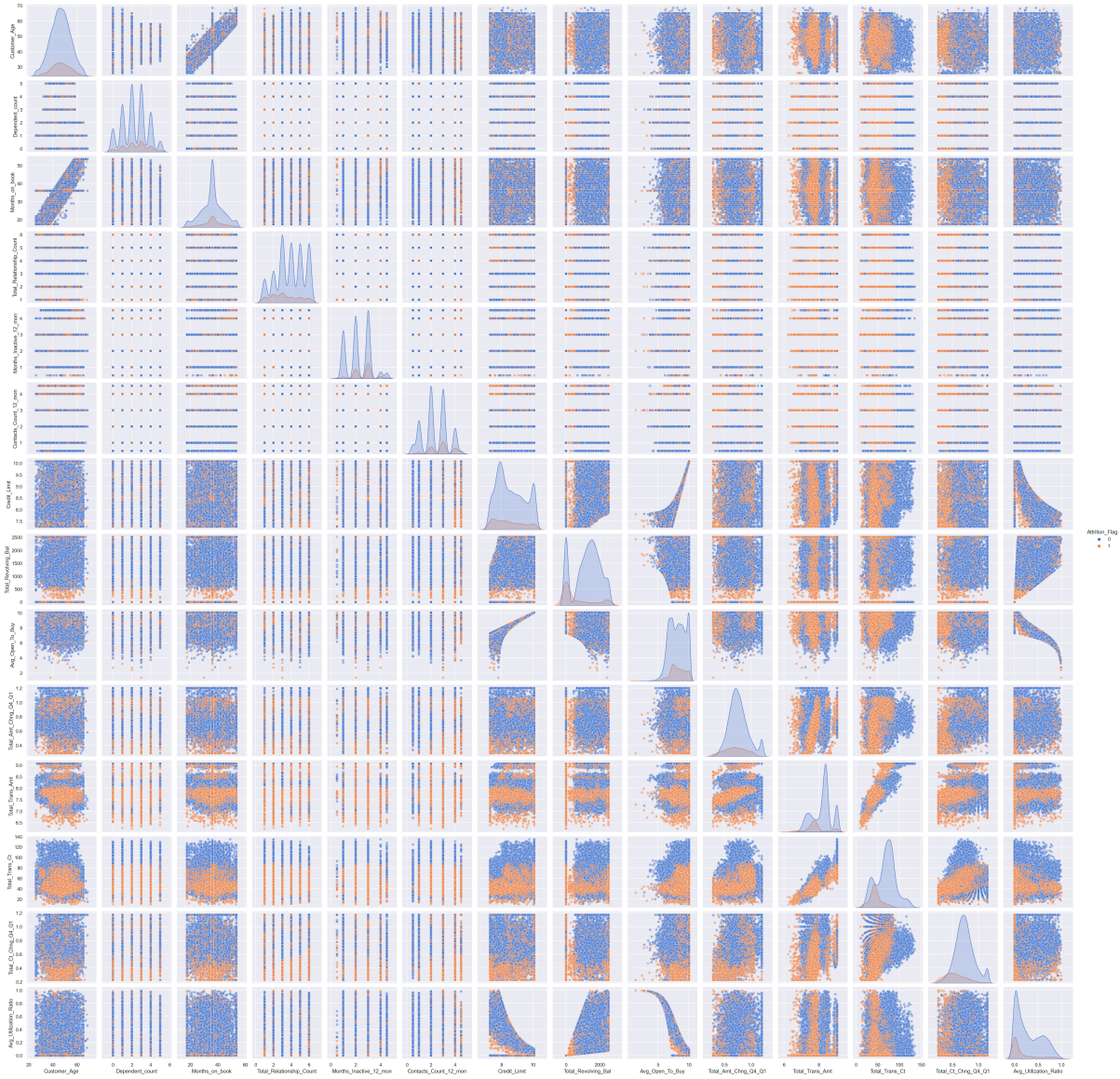
- Visualization of skewed features after Log Transformation:



It seems like the Log Transformation didn't solved the skewness of the skewed features. An examination of the paired plots and correlation maps should show whether it is reasonable to continue further log transformations for these features.

Examining Pair Plots and Correlation Maps of the features

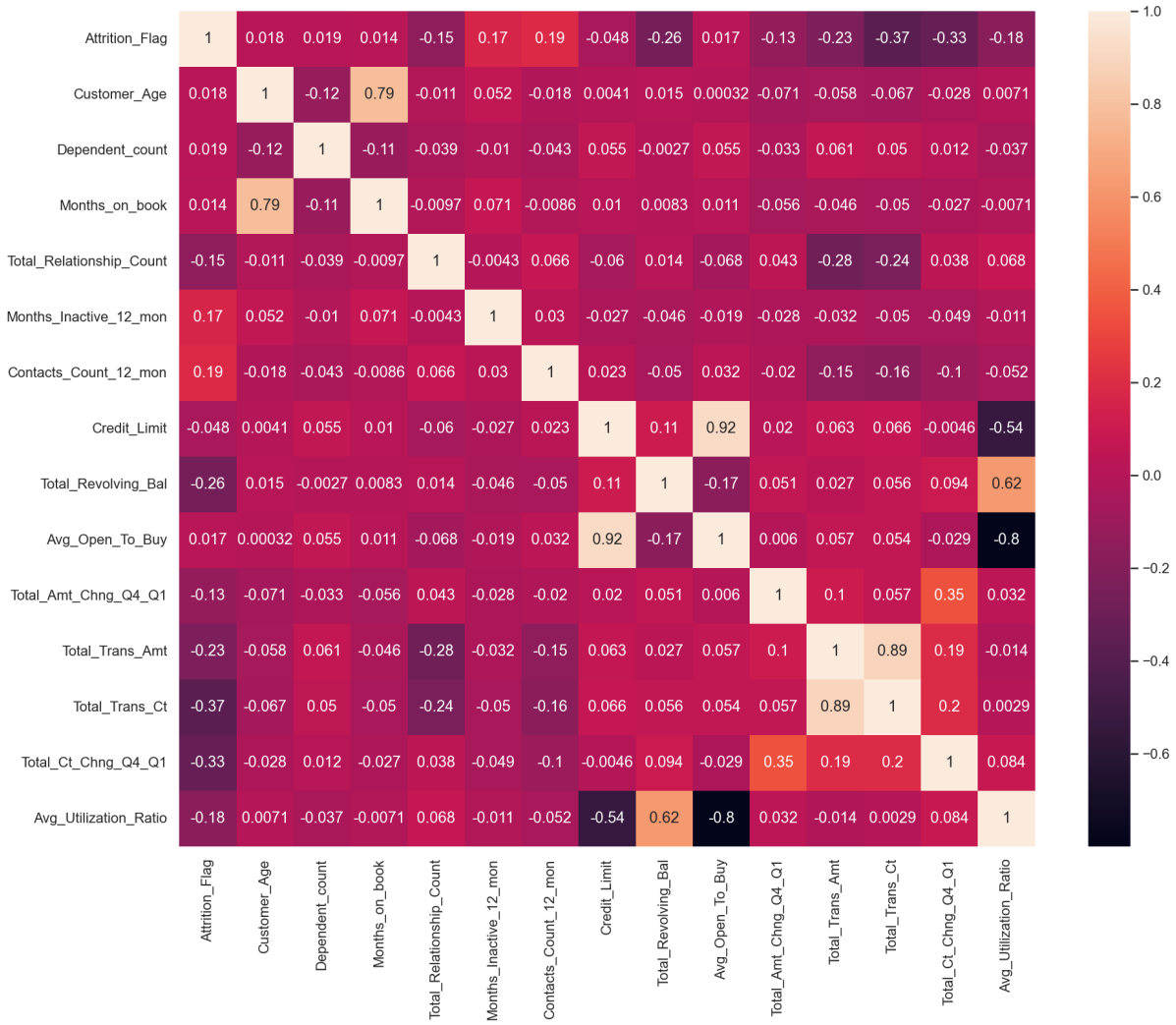
- Visualization of the relationship between numerical features and the label:



There are some correlations between the features:

- Customer_Age is correlated with Months_on_book.
- Credit_Limit is correlated with Avg_Open_To_Buy and both are correlated with Avg_Utilization_Ratio.
- Avg_Utilization_Ratio is correlated with Total_Revolving_Bal.
- Total_Amt_Chng_Q4_Q1 and Total_Trans_Amt are correlated with Total_Ct_Chng_Q4_Q1 and Total_Trans_Ct.

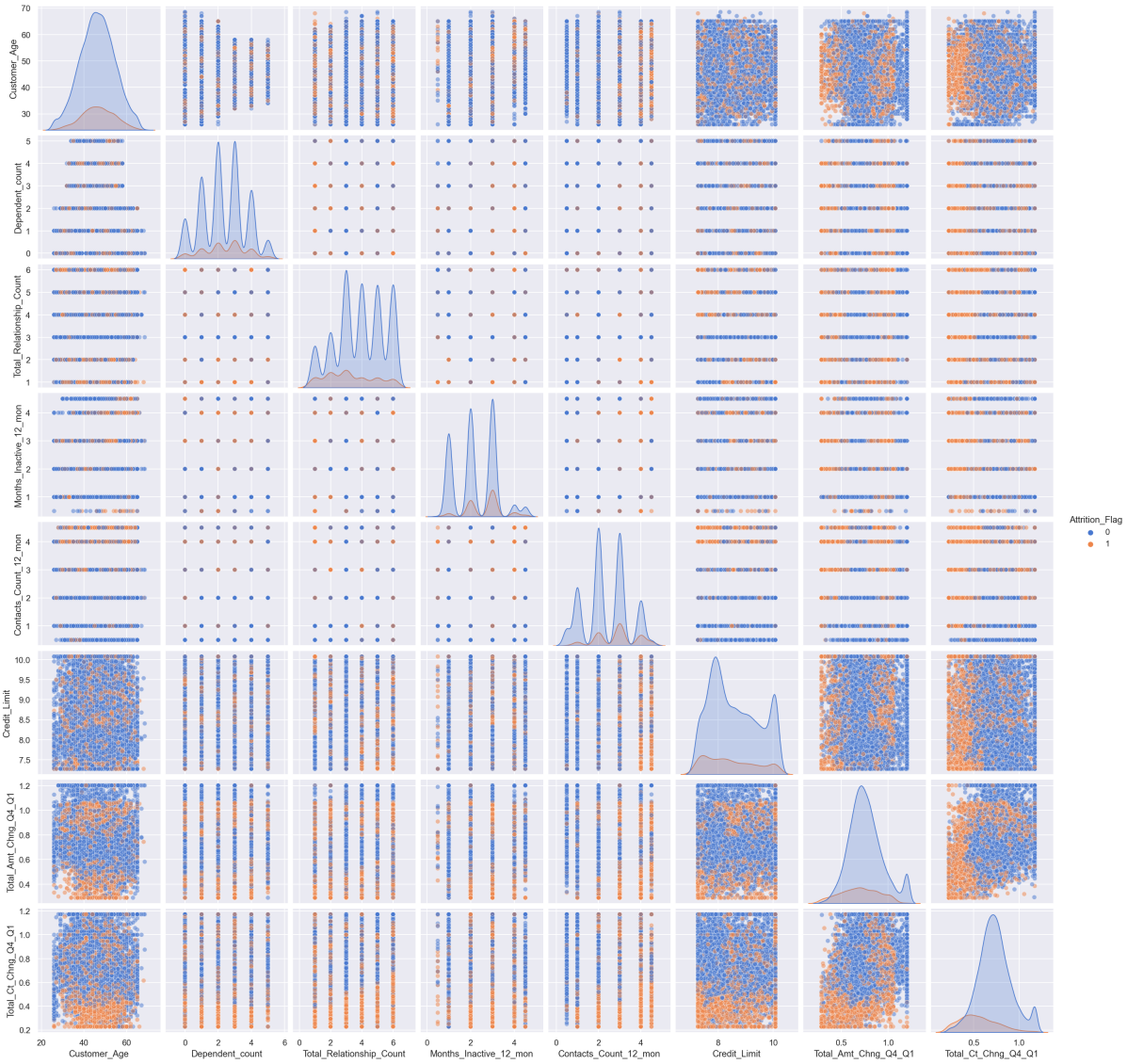
- Visualization of the correlation map:



After examining both pair plot and correlation map was decided to drop the following columns:

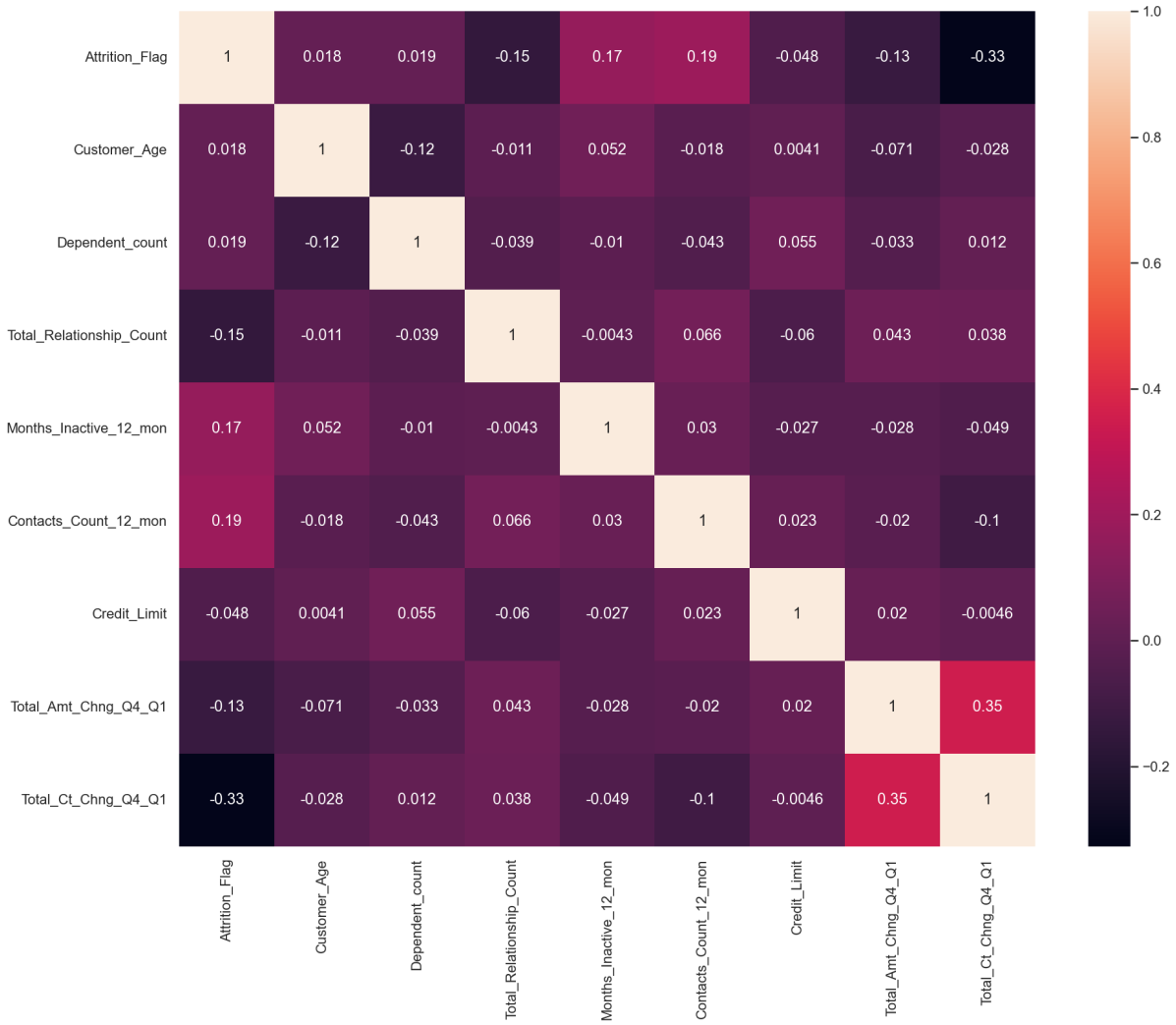
- Months_on_book
- Avg_Open_To_Buy
- Total_Trans_Amt
- Total_Trans_Ct
- Total_Revolving_Bal
- Avg_Utilization_Ratio

- Visualization of the adjusted pair plot:



The data looks much better after removing the correlated features.

- Visualization of the adjusted correlation map:



There is still a weak correlation between Total_Amt_Chng_Q4_Q1 and Total_Ct_Chng_Q4_Q1. But it is weak enough not to distort the results of further calculations.

Key Findings and Insights

The pair plot show that for some features, whether the customer is an existing customer does not affect their distribution. For example, the distribution of the age of customers are almost the same for existing and attrited customers. It suggests that most of the customers are around age 45, but it does not provide much information for predicting customer churn.

Other features have different distributions among existing customers and attrited customers. For example, if a customer has a high total revolving balance or a large amount of total transaction, this customer are more likely to be an existing customer rather than an attrited customer.

Some features have a strong and even very strong correlation. For example, the period relationship of a customer with the bank appears to be linearly correlated with the age of a customer. Although a relationship of 37 months with the bank seems to be very common and exists in all age groups. Another example is the polynomial relationship between the average open to buy credit line and the average utilization ratio. An higher average open to buy credit line is associated with a lower average utilization ratio. However, these features do not appear to be very useful in the analysis.

Hypothesis Testing

Hypothesis 1

There is a relationship between the number of dependents and the customer attrition. The more the number of dependents a client has, the lower the chance that he will churn.

OLS Regression Results

Dep. Variable:	Attrition_Flag	**R-squared:**	0.000
Model:	OLS	**Adj. R-squared:**	0.000
Method:	Least Squares	**F-statistic:**	3.653
Date:	Sun, 12 Sep 2021	**Prob (F-statistic):**	0.0560
Time:	00:34:10	**Log-Likelihood:**	-4222.5
No. Observations:	10127	**AIC:**	8449.
Df Residuals:	10125	**BIC:**	8463.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.1481	0.008	19.654	0.000	0.133	0.163
Dependent_count	0.0054	0.003	1.911	0.056	-0.000	0.011

Omnibus:	3040.106	**Durbin-Watson:**	1.941
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	6603.334
Skew:	1.847	**Prob(JB):**	0.00
Kurtosis:	4.415	**Cond. No.**	6.14

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The test results show that the P-value is 0.056. Therefore, with a confidence of 95% this hypothesis could be accepted.

Hypothesis 2

There is a relationship between the number of products hold by the customer and a customer attrition. The more the number of products a customer holds, the lower the chance that he will churn.

OLS Regression Results

Dep. Variable:	Attrition_Flag	**R-squared:**	0.023
Model:	OLS	**Adj. R-squared:**	0.022
Method:	Least Squares	**F-statistic:**	233.1
Date:	Sun, 12 Sep 2021	**Prob (F-statistic):**	4.83e-52
Time:	00:34:10	**Log-Likelihood:**	-4109.1
No. Observations:	10127	**AIC:**	8222.
Df Residuals:	10125	**BIC:**	8237.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.2958	0.010	30.947	0.000	0.277	0.315
Total_Relationship_Count	-0.0354	0.002	-15.267	0.000	-0.040	-0.031

Omnibus:	2920.450	**Durbin-Watson:**	1.939
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	6172.146
Skew:	1.787	**Prob(JB):**	0.00
Kurtosis:	4.361	**Cond. No.**	11.5

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The test results show that the P-value is extremely low and it is equal to 4.83e-52. Therefore, with a confidence of 95% this hypothesis could be rejected.

Hypothesis 3

Customer age comes from a normal distribution.
The test results show that the P-value is extremely low and it is equal to 3.85e-12. Therefore, with a confidence of 95% this hypothesis could be also rejected.

Next Steps in analyzing the data

- Scaling the features to avoid unbalanced impacts of the magnitudes on results.
- Splitting the dataset into training data and testing data.
- Performing cross-validation to avoid overfitting.
- Considering the models which could be applied to learn the patterns and make predictions.

Data Quality Summary

The quality of the data is good enough. The information appears to be accurate and mostly relevant. There is no missing and duplicated values in the dataset. However, it could be helpful to get some additional information, such as the residential location of customers.

Appendix

GitHub URL:
https://github.com/evgenyzorin/IBM-Machine-Learning/tree/main/Exploratory-Data-Analysis