|  |
| --- |
| MINOR PROJECT  20 |
| street view with city buildings, market and street sings  Exploratory Data Analysis  DOMAIN: BANKING |
| July 12  INTERMILES (DATA SCIENCE)  Authored by: NANDINI PANJWANI |

# Problem Statement:

# Happy Bank provides various credit cards to customers. The manager of Happy Bank

# is disturbed by more and more customers leaving their credit card services. The team

# did a customer survey to check customer attrition. Various customer attributes like

# Customer Age, Credit Limit, Dependent Count. The team would really appreciate it if

# one could predict for them who is going to get churned so they can proactively go to the

# customer to provide them better services and turn customers& decisions in the

# opposite direction.

# STEPS with DISSCUSION

# Using Python for EDA

# Import required Libraries in python

# Loading Data

# Data Summary i.e

.The data\_overview() function, returns an overview of a given database, informing the user of important factors, such as the quantity of null records, the primitive type and the number of categorical entries for each column. Observing the result generated for the database in question, it is possible to state:

* There are no null records for the database available;
* Of the 23 columns available, 6 are categorical and 17 are numeric;
* The categorical column with the most registered entries is \_Education*Level* with 7 different entries;

# \*the shape of the data -> (10127,21) It shows rows and columns in Data

# \*the percentage of missing values in each column of the data- No MISSING VALUES

# \*Check for any duplicate rows.- NO DUPLICATE ROWS

# 5. The distribution of the Customer Age column. Check the basic statistics like mean, median, and standard deviation of the age column.

# We can observe in below boxplot that the distribution of customer ages in our dataset follows a fairly normal distribution; thus, further use of the age feature can be done with the normality assumption.

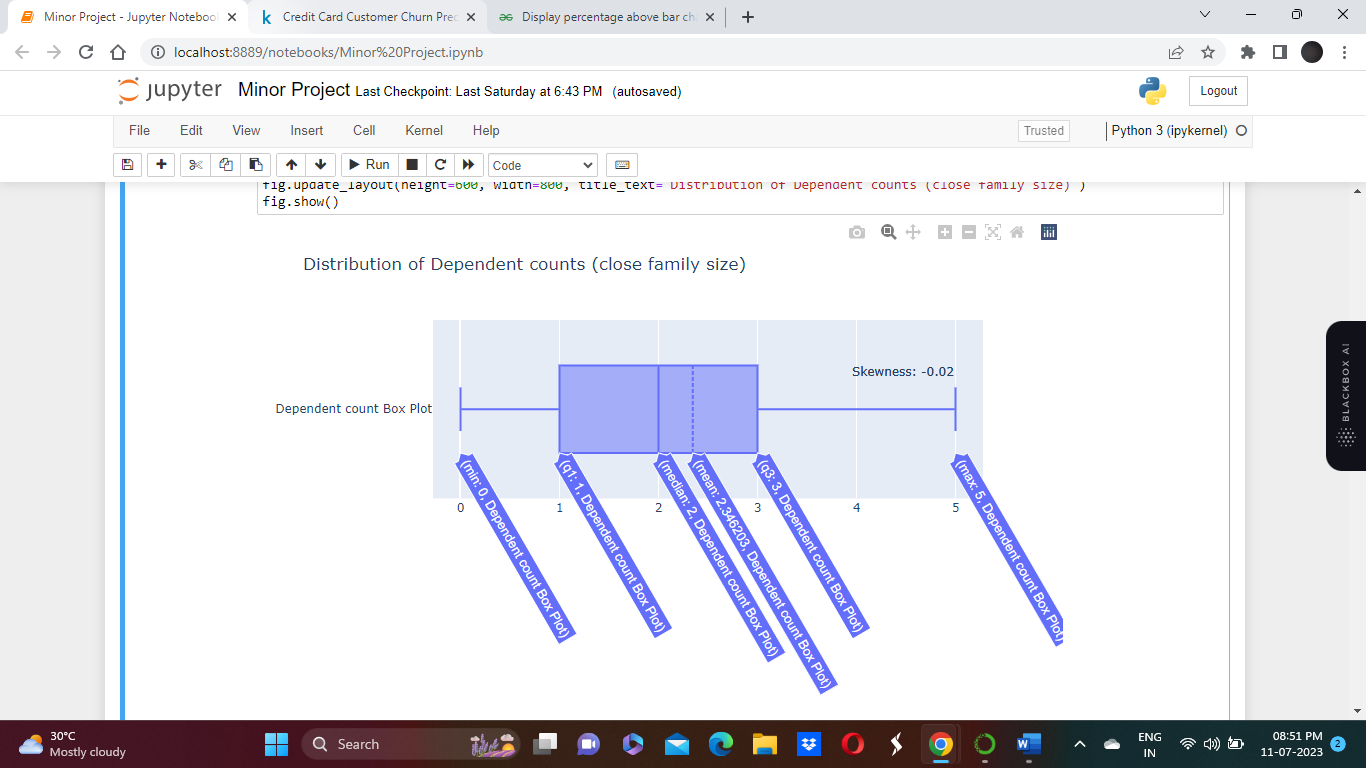
# 

# 

Mean value for the Customer Age column is approx 46 and the median is also 46. This shows that majority of the customers are under 46 years of age.

6.BOXPLOT of DEPENDENT counts

As we can observe that the distribution of Dependent counts is fairly normally distributed with a slight right skew (The mean is greater than the median)



# 7. BOXPLOT of Total products held by customers

# 

# The distribution of the total number of products held by the customer seems closer to a uniform distribution and may appear useless as a predictor for churn status.

# 8. PIE CHART of Education level and Different income levels

# 

# Most of the customers with unknown education status lack any education, we can state that more than 70% of the customers have a formal education level. About 35% have a higher level of education.

# 

# Majority of the customers’ income level is less than 40k dollars and only 7% of the customers have income of 120k+ dollars.

# 8. Box plot of total revolving amount

# Total revolving balance of 0 would mean the customer never uses the credit card

# 

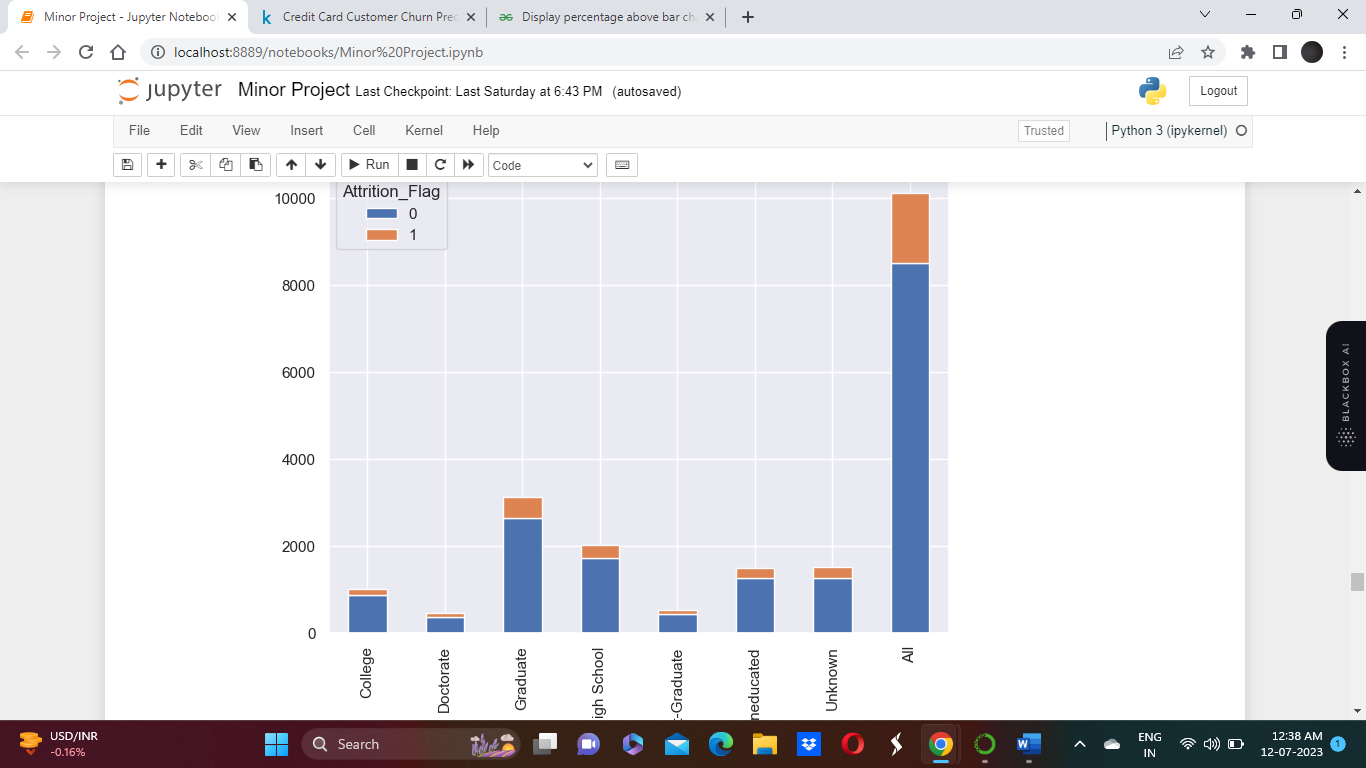
# Box Plot of Card Category

# 

The chart above show important factors related to how customers can be categorized in terms of consumption variables at this financial institution. Regarding customer relationship time, it is possible to perceive a slight positive correlation between "card level" and "long relationship time" with the bank. The breakdown by attrition flag indicates that there is no direct relationship between these two variables with clients who migrated to another institution.

# 9. Percentage segment bar graph between Education\_Level and

# Attrition\_Flag of the customers.

****

The graph above is important to have a better understanding of the audience present at the base in relation to the level of education of customers.

Major Existing Customers are Graduated.

# 10. Plot a percentage segment bar graph between Income\_Category andAttrition\_Flag of the customers.

The above graphs allow us to infer that:

#The vast majority of this institution's clients fall into the portion with annual earnings of less than $40K;

#There is a subtle relationship between age and salary range, indicating that customers with high annual earnings are usually part of an older audience.

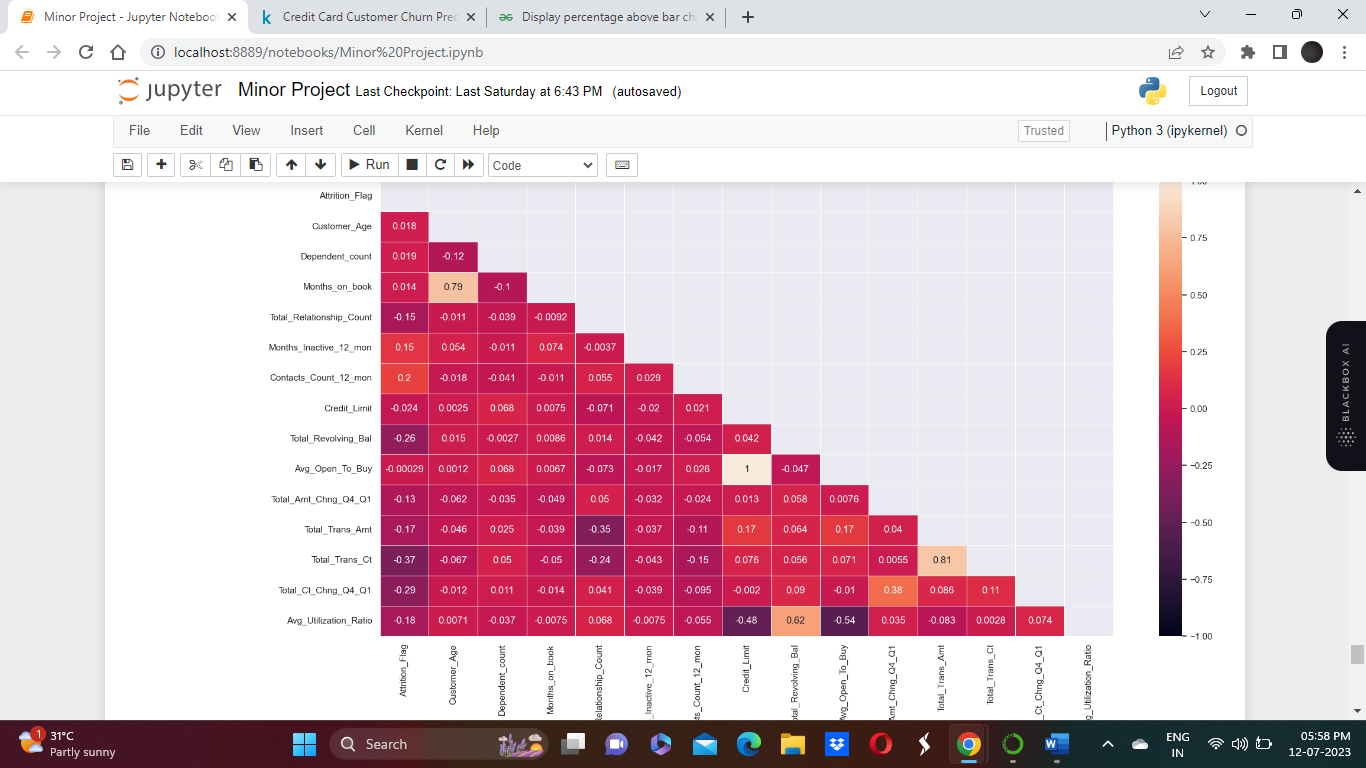
11. A sub data frame which consists of all the

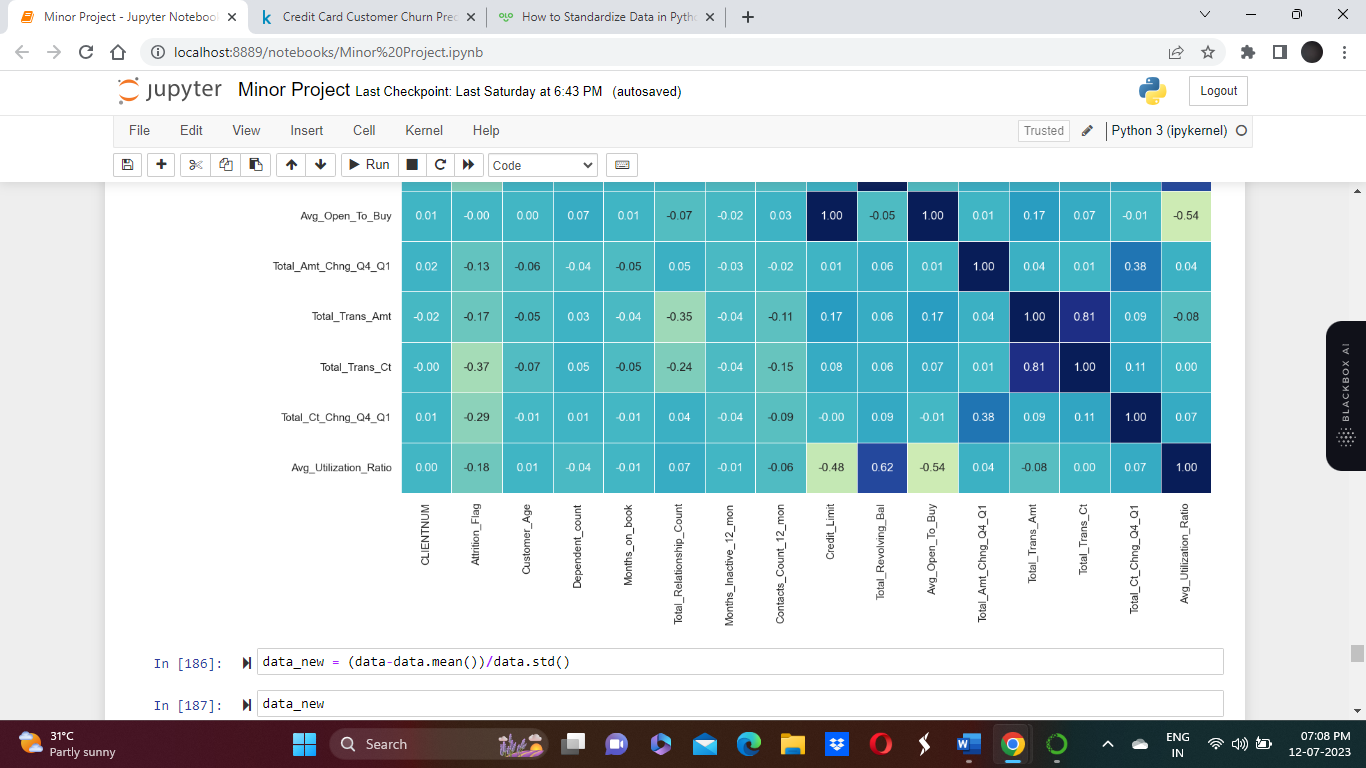
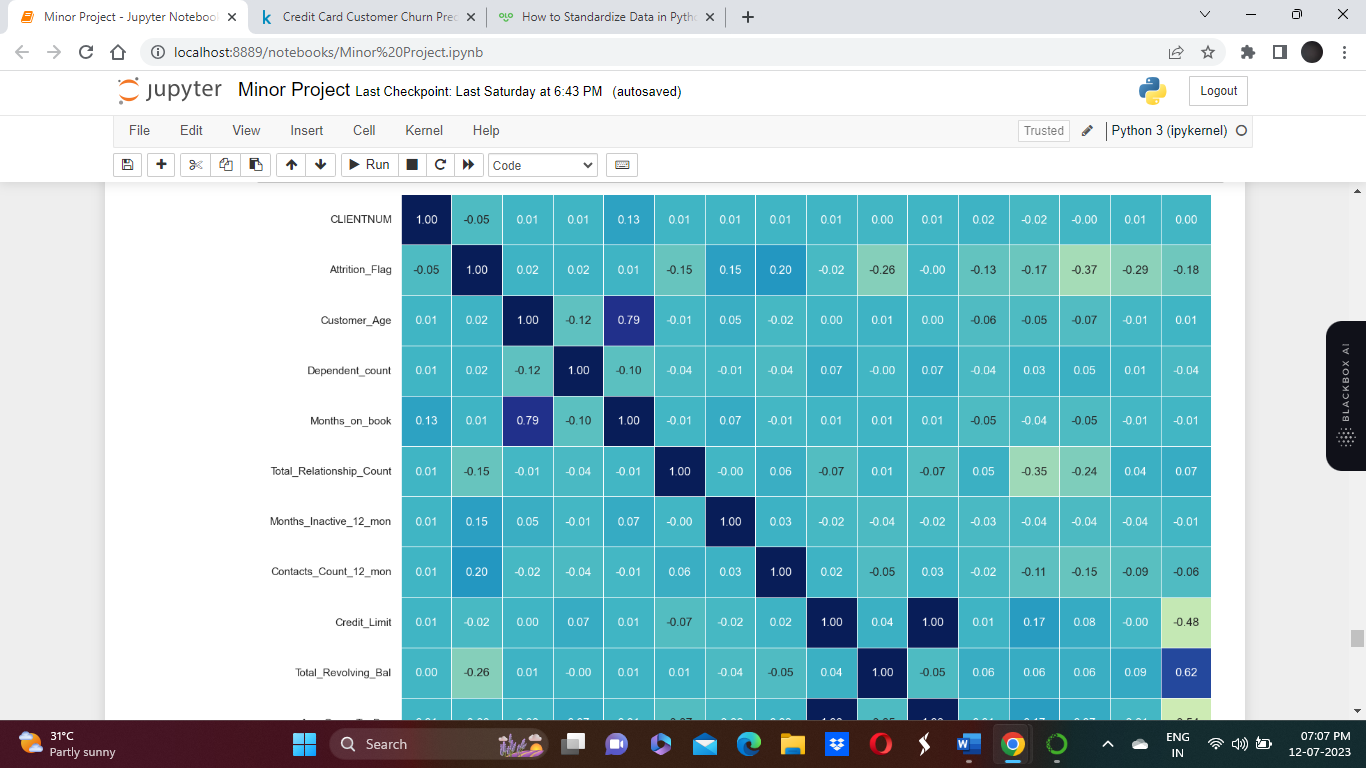
numerical columns(i.e.int64,float64).

Plot a clear heatmap to view the correlation using seaborn.

A sub data frame built after dropping “Clientium” column and a heatmap was generated using seaborn library

After an initial approach to important variables that describe the profile of the financial institution's customers, it is possible to analyze the numerical attributes at once in terms of correlation. This approach is important to give an overview, in a single view, of how the variables are correlated with each other and with a target variable (Attrition flag)

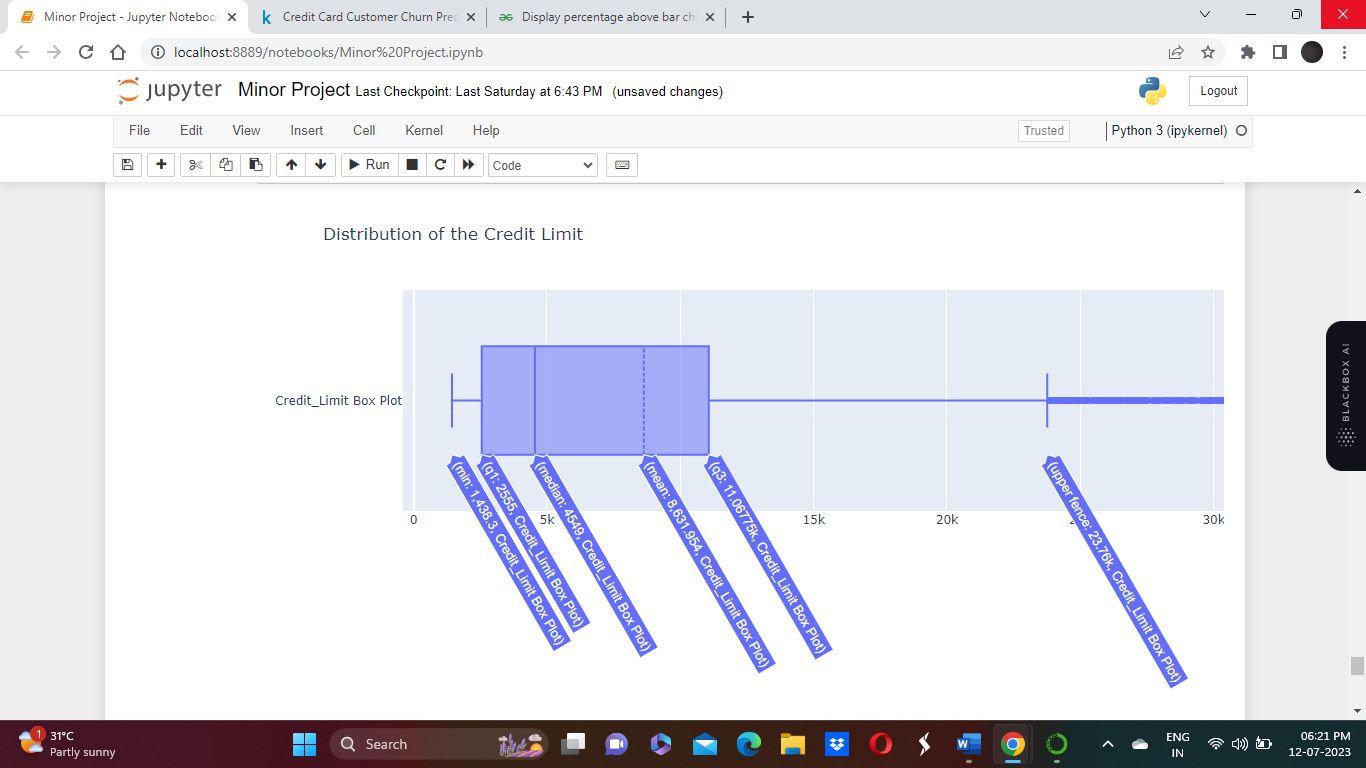
* +  There are clusters formed with respect to attrition for the variables total revolving amount, total amount change Q4 to Q1, total transaction amount, total transaction count, total transaction count change Q4 to Q1
  + There are strong correlation between a few columns as well, which we'll check in below correlation heatmap.
* Credit Limit and Average Open to Buy have 100% collinearity
* Months on book and Customer Age have quite strong correlation
* Average Utilization Ration and Total Revolving Balance are also a bit correlated it appears
* Attrition Flag does not have highly strong correlation with any of the numeric variables
* Customer Churn appears to be uncorrelated with Customer Age, Dependent Count, Months on Book, Open to Buy, Credit Limit, we'll remove these from dataset.

 Credit Limit and Average Open to Buy have 100% collinearity

* Months on book and Customer Age have quite strong correlation
* Average Utilization Ration and Total Revolving Balance are also a bit correlated it appears
* Attrition Flag does not have highly strong correlation with any of the numeric variables
* Customer Churn appears to be uncorrelated with Customer Age, Dependent Count, Months on Book, Open to Buy, Credit Limit, we'll remove these from dataset

12.Box Plot of Credit\_limit

Credit limit has outliers



13. Standarzing the columns by Removing the spaces from column names, and standardizing the column names to lower case.

We can verify that the mean and standard deviation of each column is equal to 0 and 1, respectively.