CCT COLLEGE

Statistical Tchniques

Assessment

Dublin

2021

Nuno Alfredo Ribeiro Teixeira de Almeida - 2021310

CREDIT CARD SAMPLE ANALYSIS

Work developed in the analysis of a Dataset and use of the models to obtain the second general grade of the Statistical Techniques of tHDip Data Analytics - FT - Sept 2021 at CCT course.

Lectures: Aldana

Dublin

2021

SUMMARY

1. INTRODUCTION

The financial market grows each year, and with them, the customers. This statement comes from my intention to develop this work, seeking to understand some variables and meet the proposed requirements for a second-semester grade in statistical techniques.

To develop the work, I researched the Kaggle platform, a datacamp that has a vast database of Datasets. All content is available for consultation and debate among users.

Using the CA2 assessment tools, our subject is related to customer data from a credit card company, which brings us relevant information, such as relationships and essential numbers, to adhere to our study and the applicability of our theory in a possible practice.

The work will be developed in stages, namely, Introductory, where we are currently highlighting this information. We will have the Methodology part, Results, Discussion, and Conclusion.

2. METHODOLOGY

According Kuria (2020) Data collection is the process of gathering and measuring information for variables of interest in an established and systematic fashion enabling one to answer queries, state research questions, test hypotheses and evaluate the outcomes.

The project aims to present an exploratory analysis of Credit Card Customers, using Data Preparation and Statistical techniques to develop hypothesis testing based on research and work development. Using Jupyter Notebook so that mathematical operations could be conducted through specific functions and thus used. Separated by sections, it will create analyses on samples and their variables in written form.

External research referenced sites with subjects related to Credit Card and their relationship with the variables available in the Dataset. They are credit limit, gender, among others. We also used external web research and libraries available in Python language to have illustrations and perform functions in the Data Set used.

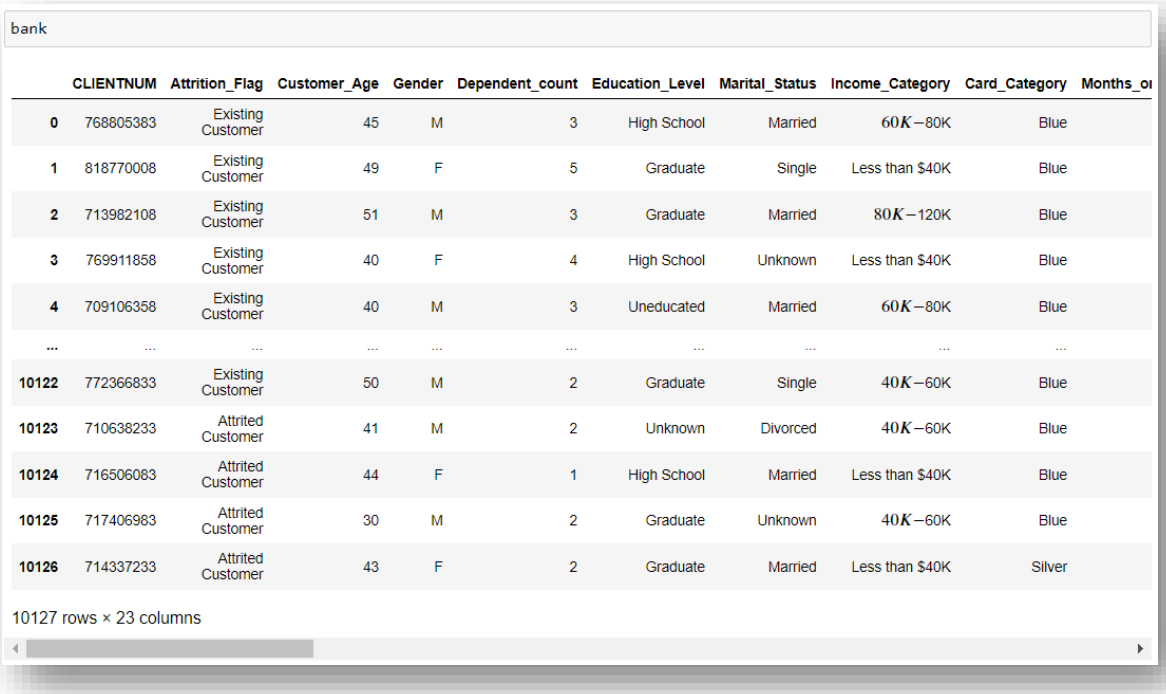
The methodology we used in this dataset includes different approaches, including an EDA (Exploratory Data Analysis) where combining different code functions (explained below) we determined the dataset, dealing with missing values, or zero values, etc. Hypotheses test using the sample to developed a prediction, Correlations techniques using between variables and linear regression mode separately.

3. RESULTS

3.1 DATA UNDERSTANDING

According to the descriptive analysis of the dataset in general, it contains 10.127 rows and 23 columns. The columns and data print are in the picture below.

Figure 1 – Dataset;

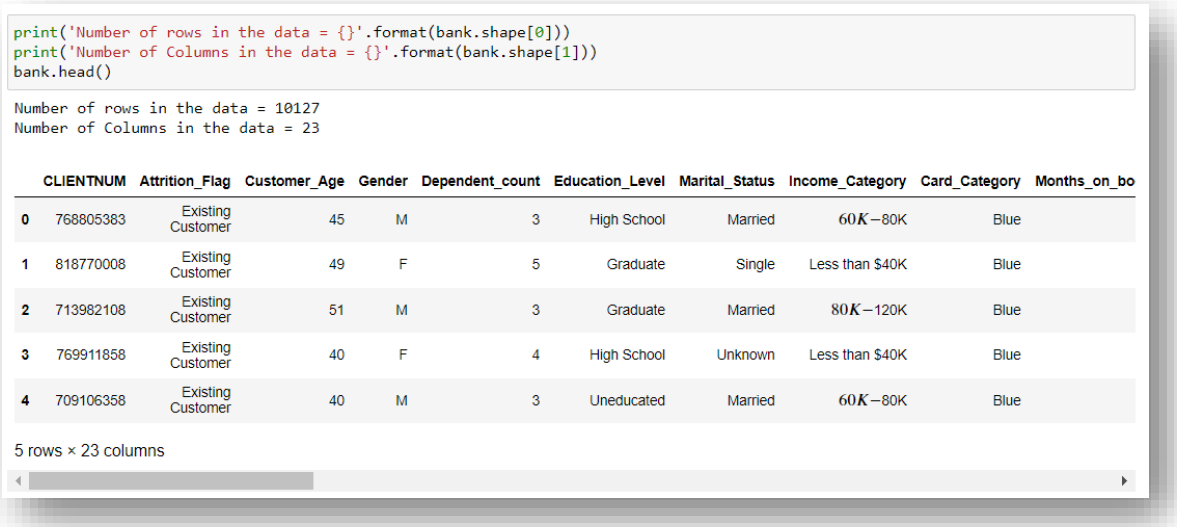


bank

	CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_o
0	768805383	Existing Customer	45	M	3	High School	Married	$60K – 80K	Blue	
1	818770008	Existing Customer	49	F	5	Graduate	Single	Less than $40K	Blue	
2	713982108	Existing Customer	51	M	3	Graduate	Married	$80K – 120K	Blue	
3	769911858	Existing Customer	40	F	4	High School	Unknown	Less than $40K	Blue	
4	709106358	Existing Customer	40	M	3	Uneducated	Married	$60K – 80K	Blue	
...	...	...	...	...	...	...	...	...	...	
10122	772366833	Existing Customer	50	M	2	Graduate	Single	$40K – 60K	Blue	
10123	710638233	Attrited Customer	41	M	2	Unknown	Divorced	$40K – 60K	Blue	
10124	716506083	Attrited Customer	44	F	1	High School	Married	Less than $40K	Blue	
10125	717406983	Attrited Customer	30	M	2	Graduate	Unknown	$40K – 60K	Blue	
10126	714337233	Attrited Customer	43	F	2	Graduate	Married	Less than $40K	Silver	

10127 rows × 23 columns

Figure 2 – Size and details about Data Set;



```
print('Number of rows in the data = {}'.format(bank.shape[0]))
print('Number of Columns in the data = {}'.format(bank.shape[1]))
bank.head()
```

```
Number of rows in the data = 10127
Number of Columns in the data = 23
```

	CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_bo
0	768805383	Existing Customer	45	M	3	High School	Married	$60K – 80K	Blue	
1	818770008	Existing Customer	49	F	5	Graduate	Single	Less than $40K	Blue	
2	713982108	Existing Customer	51	M	3	Graduate	Married	$80K – 120K	Blue	
3	769911858	Existing Customer	40	F	4	High School	Unknown	Less than $40K	Blue	
4	709106358	Existing Customer	40	M	3	Uneducated	Married	$60K – 80K	Blue	

5 rows × 23 columns

Figure 3 – Information about Data Set;

```
bank.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7081 entries, 0 to 10126
Data columns (total 20 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Cliente ID                7081 non-null   int64
 1   Customer                  7081 non-null   object
 2   Age                       7081 non-null   int64
 3   Gender                    7081 non-null   object
 4   Dependents                7081 non-null   int64
 5   Education                 7081 non-null   object
 6   Status                    7081 non-null   object
 7   Income_Category           7081 non-null   object
 8   Card_Category             7081 non-null   object
 9   Months_on_book            7081 non-null   int64
 10  Total_Relationship_Count  7081 non-null   int64
 11  Months_Inactive_12_mon    7081 non-null   int64
 12  Contacts_Count_12_mon     7081 non-null   int64
 13  Credit_Limit              7081 non-null   float64
 14  Total_Revolving_Bal       7081 non-null   int64
 15  Avg_Open_To_Buy           7081 non-null   float64
 16  Total_Amt_Chng_Q4_Q1      7081 non-null   float64
 17  Total_Trans_Amt           7081 non-null   int64
 18  Total_Trans_Ct            7081 non-null   int64
 19  Avg_Utilization_Ratio     7081 non-null   float64
dtypes: float64(4), int64(10), object(6)
memory usage: 1.1+ MB
```

This dataset consists on different types of data, such as categorical (objects/text), numbers (integer/int) and float (decimal numbers). Further, the objects will need to developing in Data preparation and Statistical Techniques.

Figure 4 – Dataset Columns names;

```
bank.columns
```

```
Index(['CLIENTNUM', 'Attrition_Flag', 'Customer_Age', 'Gender',
       'Dependent_count', 'Education_Level', 'Marital_Status',
       'Income_Category', 'Card_Category', 'Months_on_book',
       'Total_Relationship_Count', 'Months_Inactive_12_mon',
       'Contacts_Count_12_mon', 'Credit_Limit', 'Total_Revolving_Bal',
       'Avg_Open_To_Buy', 'Total_Amt_Chng_Q4_Q1', 'Total_Trans_Amt',
       'Total_Trans_Ct', 'Total_Ct_Chng_Q4_Q1', 'Avg_Utilization_Ratio',
       'Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Ina
ctive_12_mon_1',
       'Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Ina
ctive_12_mon_2'],
      dtype='object')
```

Column Description according the author from Kaggle:

<u>About Customer:</u>

- CLIENTNUM: unique client number value each customer

- Attrition_Flag: Internal event (customer activity) variable ("Existing Customer","Attrited Customer")

- Customer_Age: Demographic variable - Customer's Age in Years

- Gender: Demographic variable - M=Male, F=Female

- Dependent_count: Demographic variable - Number of dependents

- Education_Level: Demographic variable - Educational Qualification of the account holder (example: high school, college graduate, etc.)

- Marital_Status: Demographic variable - Married, Single, Divorced, Unknown

- Income_Category: Demographic variable - Annual Income Category of the account holder (< 40K, 40K - 60K, 60K− 80K, 80K− 120K, >

About Credit Card:

- Card_Category: Product Variable - Type of Card (Blue, Silver, Gold, Platinum)

- Months_on_book: Period of relationship with bank

- Total_Relationship_Count:Total no. of products held by the customer.

- Months_Inactive_12_mon:No. of Months in the last 12 months.

- Contacts_Count_12_mon:No. of Contacts in the last 12 months.

- Credit_Limit:Credit Limit on the Credit Card

- Total_Revolving_Bal:Total Revolving Balance on the Credit Card.

- Avg_Open_To_Buy:Open to Buy Credit Line (Average of last 12 months

- Total_Amt_Chng_Q4_Q1:Change in Transaction Amount (Q4 over Q1).

- Total_Trans_Amt:Total Transaction Amount (Last 12 months).

- Total_Trans_Ct:Total Transaction Count (Last 12 months).

- Total_Ct_Chng_Q4_Q1:Change in Transaction Count (Q4 over Q1).

- Avg_Utilization_Ratio:Average Card Utilization Ratio.

Figure 5 – The 5 number summary

```
bank.describe()
```

	CLIENTNUM	Customer_Age	Dependent_count	Months_on_book	Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Count_12_mon	Cre
count	1.012700e+04	10127.000000	10127.000000	10127.000000	10127.000000	10127.000000	10127.000000	1012
mean	7.391776e+08	46.325960	2.346203	35.928409	3.812580	2.341167	2.455317	863
std	3.690378e+07	8.016814	1.298908	7.986416	1.554408	1.010622	1.106225	908
min	7.080821e+08	26.000000	0.000000	13.000000	1.000000	0.000000	0.000000	143
25%	7.130368e+08	41.000000	1.000000	31.000000	3.000000	2.000000	2.000000	255
50%	7.179264e+08	46.000000	2.000000	36.000000	4.000000	2.000000	2.000000	454
75%	7.731435e+08	52.000000	3.000000	40.000000	5.000000	3.000000	3.000000	1106
max	8.283431e+08	73.000000	5.000000	56.000000	6.000000	6.000000	6.000000	3451

```
bank.describe(include=object)
```

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

Figure 6 – Dropping Columns;

```
bank.drop(['Total_Ct_Chng_Q4_Q1',
           'Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_
           'Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_
          axis = 1, inplace = True)

bank.head()
```

Figure 7 – Information and Missing Data Variables;

```
print('Number of missing values by column:')
bank.isna().sum()
```

Figure 8 – Rename Columns;

```
bank.rename(columns = {'CLIENTNUM': 'Cliente ID','Attrition_Flag': 'Customer', 'Customer_Age': 'Age','Dependent_count':'Deper
```

Figure 9 – Understanding Unknown Values;

```
bank['Status'].unique()
```

```
array(['Married', 'Single', 'Unknown', 'Divorced'], dtype=object)
```

Figure 10 – – Numbers of Unknown Values;

```
Age                       0.000000
Gender                    0.000000
Dependents                0.000000
Education                 0.149995
Status                    0.073961
Income_Category           0.109805
Card_Category             0.000000
Months_on_book            0.000000
Total_Relationship_Count  0.000000
Months_Inactive_12_mon    0.000000
Contacts_Count_12_mon     0.000000
```

Figure 11 – Duplicated Values;

```
bank.duplicated().any()
```

```
False
```

Figure 12 – Featuring Engineering with column;

```
bank['Income_Category'].unique()
```

```
array(['$60K - $80K', 'Less than $40K', '$80K - $120K', '$40K - $60K',
       '$120K +'], dtype=object)
```

lace the categories with labels which describes the level of income and Display the counts of the the new income labels

```
income_categories_labels = {'Less than $40K' : 'low', '$40K - $60K' : 'medium',
                            '$60K - $80K': 'above medium', '$80K - $120K':'high', '$120K +' : 'very high'}
bank['Income_Category'] = bank['Income_Category'].replace(income_categories_labels)

bank['Income_Category'].value_counts()
```

```
low             2792
medium          1412
high            1202
above medium    1103
very high        572
Name: Income_Category, dtype: int64
```

Figure 13 – Describe function with all variables;

```
In [21]: vg.describe()
```

Out[21]:

	Rank	Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
count	16540.000000	16540.000000	16540.000000	16540.000000	16540.000000	16540.000000	16540.000000
mean	8294.197642	2006.414510	0.265079	0.146883	0.077998	0.048191	0.538426
std	4790.703200	5.788794	0.817929	0.506129	0.309800	0.188879	1.557424
min	1.000000	1980.000000	0.000000	0.000000	0.000000	0.000000	0.010000
25%	4143.750000	2003.000000	0.000000	0.000000	0.000000	0.000000	0.060000
50%	8292.500000	2007.000000	0.080000	0.020000	0.000000	0.010000	0.170000
75%	12440.250000	2010.000000	0.240000	0.110000	0.040000	0.040000	0.480000
max	16600.000000	2020.000000	41.490000	29.020000	10.220000	10.570000	82.740000

```
In [22]: vg.describe(include=object)
```

Out[22]:

	Name	Platform	Genre	Publisher
count	16540	16540	16540	16540
unique	11442	31	12	578
top	Need for Speed: Most Wanted	PS2	Action	Electronic Arts
freq	12	2159	3309	1351

Figure 14 – Analyzing Outliers with quantile function;

```
Q1 = bank.quantile(0.25)
Q3 = bank.quantile(0.75)
IQR = Q3 - Q1
print(IQR)
```

```
Cliente ID                6.022650e+07
Age                       1.100000e+01
Dependents                2.000000e+00
Months_on_book            9.000000e+00
Total_Relationship_Count  2.000000e+00
Months_Inactive_12_mon    1.000000e+00
Contacts_Count_12_mon     1.000000e+00
Credit_Limit              8.231000e+03
Total_Revolving_Bal       1.318000e+03
Avg_Open_To_Buy           8.243000e+03
Total_Amt_Chng_Q4_Q1      2.290000e-01
Total_Trans_Amt           2.651000e+03
Total_Trans_Ct            3.600000e+01
Avg_Utilization_Ratio     4.890000e-01
dtype: float64
```

3.2 VISUALISATION

3.2.1 CATEGORICAL DATA

Figure 15 – Analyzing Outliers with quantile function;

```
Categorical Columns are : Index(['Customer', 'Gender', 'Education', 'Status', 'Income_Category',
       'Card_Category'],
      dtype='object')
```

Categorical Data

- Customer
- Gender
- Education
- Status
- Income_Category
- Card_Category

3.2.2 NUMERICAL DATA

Figure 16 – Analyzing Outliers with quantile function;

```
Numerical Columns are : Index(['Cliente ID', 'Age', 'Dependents', 'Months_on_book',
       'Total_Relationship_Count', 'Months_Inactive_12_mon',
       'Contacts_Count_12_mon', 'Credit_Limit', 'Total_Revolving_Bal',
       'Avg_Open_To_Buy', 'Total_Amt_Chng_Q4_Q1', 'Total_Trans_Amt',
       'Total_Trans_Ct', 'Avg_Utilization_Ratio'],
      dtype='object')
```

Numerical Data

- Age
- Dependents
- Months_on_book
- Total_Relationship_Count
- Months_Inactive_12_mon
- Contacts_Count_12_mon
- Credit_Limit
- Total_Revolving_Bal
- Avg_Open_To_Buy
- Total_Amt_Chng_Q4_Q1
- Total_Trans_Amt
- Total_Trans_Ct
- Total_Ct_Chng_Q4_Q1
- Avg_Utilization_Ratio

3.2.3 OUTLIERS

Figure 17 – Illustrate Outliers before Cleaning;

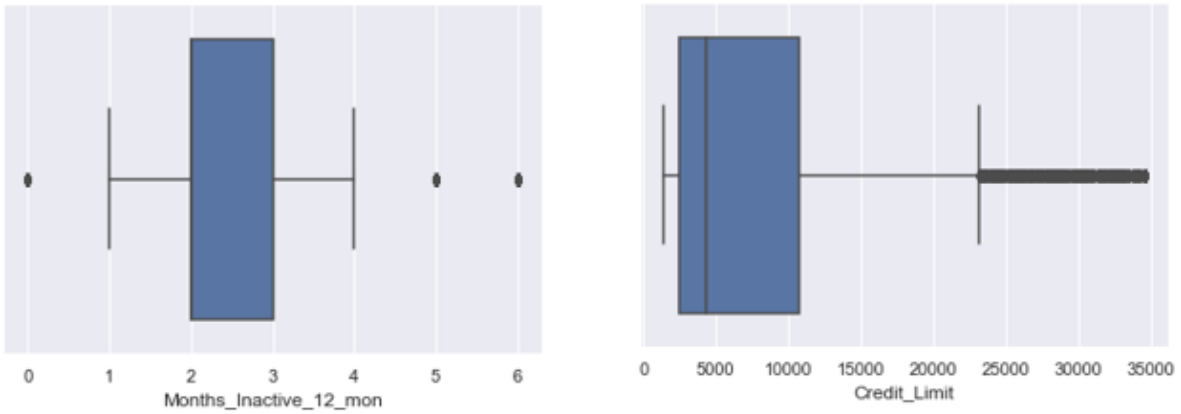


Figure 18– Illustrate Outliers results after Cleaning;

```
print('After Z-score cleaning {} rows for anlysis'.format(bank.shape[0]))
print('After Quantile cleaning {} rows for anlysis'.format(bankinter.shape[0]))
print('After cleaning {} columns for anlysis'.format(bank.shape[1]))
```

```
After Z-score cleaning 7081 rows for anlysis
After Quantile cleaning 4885 rows for anlysis
After cleaning 20 columns for anlysis
```

Figure 19 – Correlation Variables;

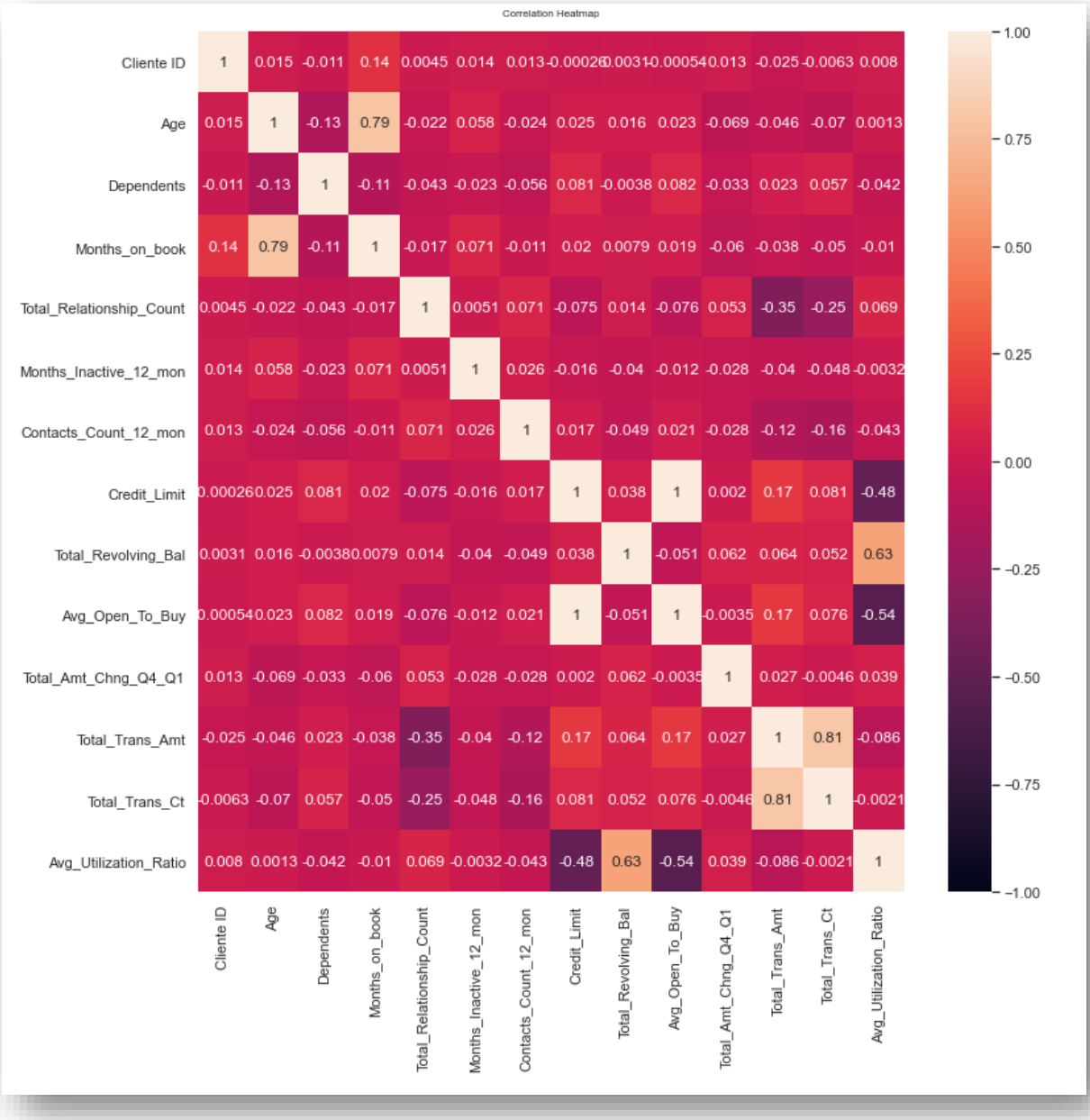


Figure 20 – Using Pairplot with Numerical Variables;

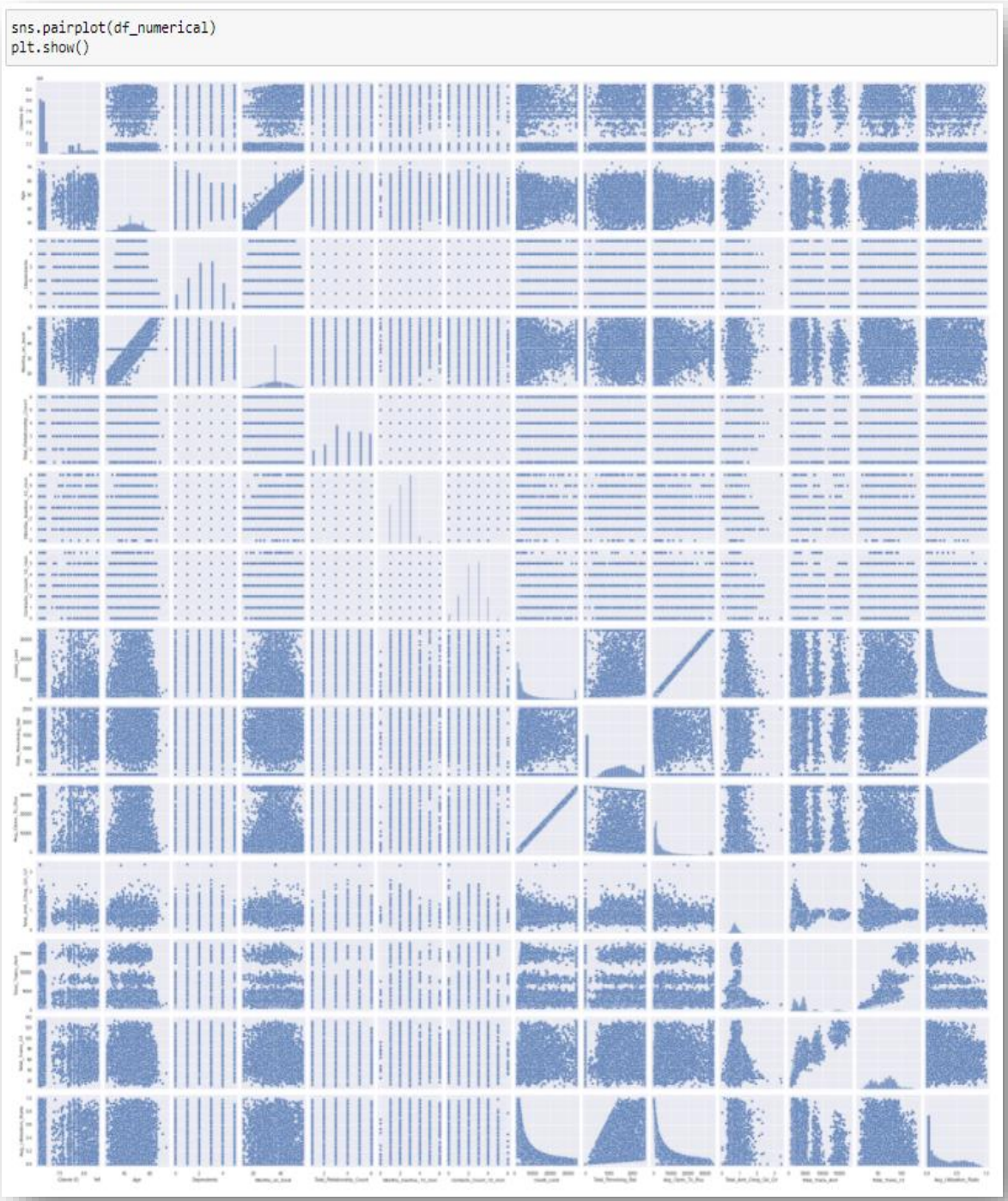


Figure 21– Using Scatterplot with Numerical Variables;

```
sns.scatterplot(data=bank, x="Credit_Limit", y="Income_Category")
```

```
<AxesSubplot:xlabel='Credit_Limit', ylabel='Income_Category'>
```

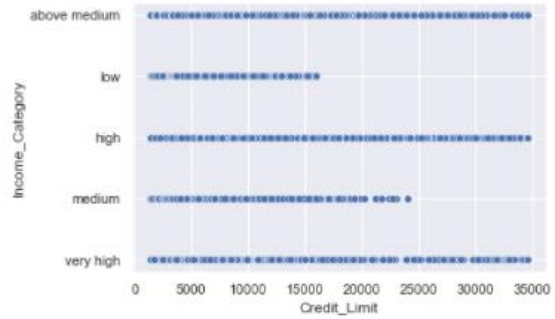
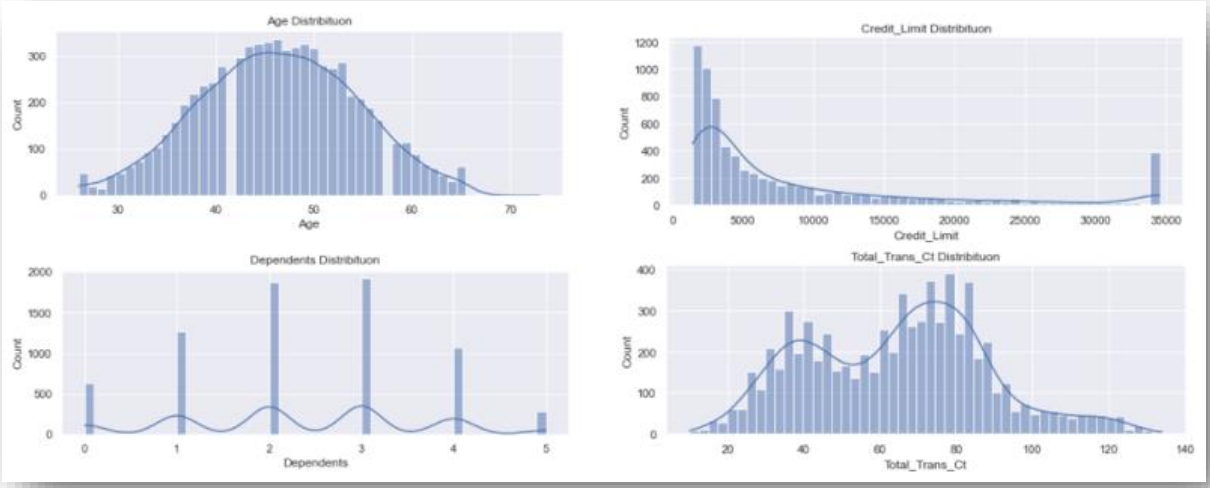


Figure 22– Illustrate Distributions according the variables values;



4. FIRST SECTION

4.1 DESCRIPTION

According to the objective of the project choose one variable, perform a Hypothesis Test based on the data, and interpret your results providing a conclusion and your own analysis. You will need to conduct research to find the parameters of the population you want to analyse. Some examples:

✓ It is believed that the average age to finish High School is 17 years old, then you perform a HT to reject or not this hypothesis.

✓ It is believed that due to COVID and remote working workers save 8.5hours on average, then you perform a HT to reject or not this hypothesis.

4.2 RESEARCH

The first moment with the Card company details, it was thought to study the credit limit and its relationship with other variables. However, it was found that gender influences a lot on some situations directly. In the Dataset defined for this study, the correlation of variables was analyzed. However, the intention with an external research was to develop the real analyzes that can be done and brought to Integrated CA.

The salary disparity between men and women means that women earn a fraction of what men do. According to updated information, there are little variances in total bank card limitations between men and women, and men tended to have access to more credit. (Konish, 2021) .

Your credit score is crucial, but it is far from the only variable in determining. Creditors want assurance that you'll be able to make good on your payment when you open an account, which is where your profits come in. Having a regular source of income can help lenders see you as a suitable borrower.

Developing our analysis around the results found, we will carry out the hypothesis test, resulting from the analysis in the first section. According Christina Majaski (2021):

Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data.

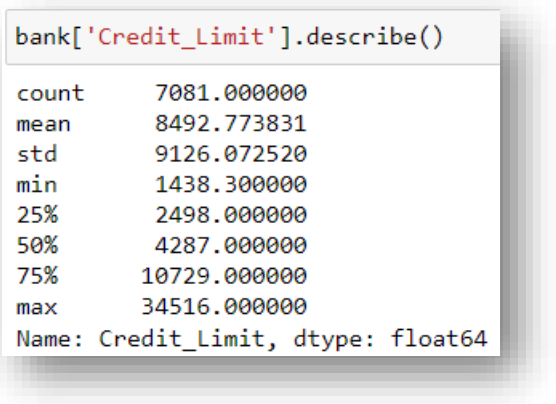
- The test provides evidence concerning the probability of the idea, given the data.
- Statistical analysts test a hypothesis by measuring and examining a random sample of the analyzed population.

The hypotheses test will work with the variables that involve Credit Limit and the results by other variables, Income Category and Gender.

Understand the variable that will be developing in that section:

Analysing that the feminine gender has little influence, let's build a hypothesis according to the 5 number summary already shown above. Based on the figure below, we will work with:

Figure 23– Describe function in Credit Limit Column;



Using se two types of Hypotheses test to highlight the results of the following statement.

4.3 ALTERNATIVE HYPOTHESIS

- H0 : The average Credit Limit on female customer is 4287 and the gender has relationship with the Income Category;
- H1 : The average Credit Limit on female customer is not 4287 and the gender has no relationship with the Income Category;

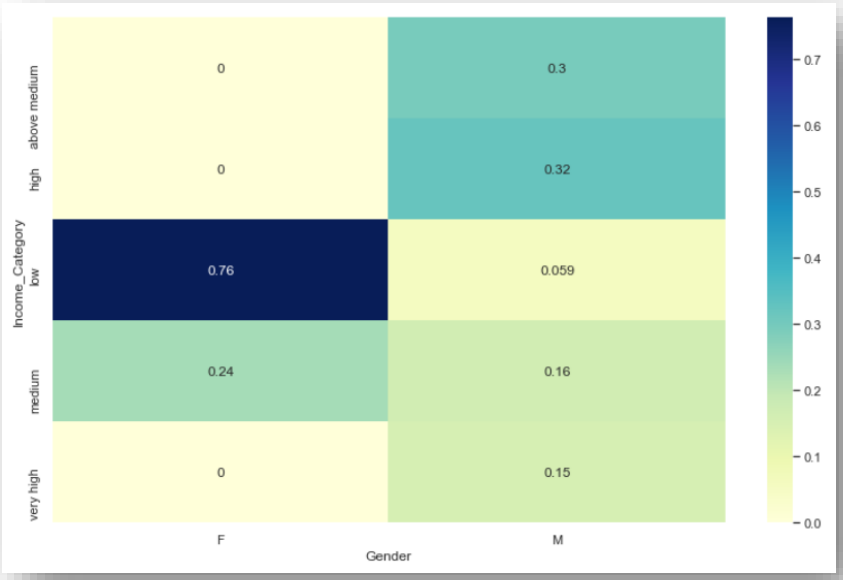
From the first step used was the A chi-square statistic WHERE is one way to show a relationship between two categorical variables. In that Analysis of the hypothese are to understand with the two variables (Gender and Income Category) has some relations between them.

Figure 24– Gender and Income Category

Gender	F	M
Income_Category		
above medium	0.00000	0.297625
high	0.00000	0.324339
low	0.76237	0.059093
medium	0.23763	0.164598
very high	0.00000	0.154344

```
contingency_table :-
 Gender                F     M
Income_Category
above medium          0  1103
high                  0  1202
low                2573   219
medium              802   610
very high             0   572
```

Figure 25– Using a heatmap to show the correlation



4.4 USING CHI-SQUARE TEST

Applying the formula Chi-square test between these variables, observed and expected values were perceived in common, when using the function (chi2_contingency) we obtained p-value results equal to 0. And thus developing the formula, to proceed with the study we obtained the following results.

Figure 26– show the statistics Values

```
p-value: 0.0
Significance level:  0.05
Degree of Freedom:  1
chi-square statistic: 4883.016920248163
critical_value: 3.841458820694124
p-value: 0.0
```

Having as the final result of the sample to verify the relationship between the 2 variables. With P-value of value 0, having the critical_value less than the statistc value, thus showing relationship between the two variables, and p value having the value less than Alpha value of 0.05.

Figure 27– Results about Hypotheses test between variables.

```
Reject H0,There is a relationship between 2 categorical variables
Reject H0,There is a relationship between 2 categorical variables
```

4.5 USING Z-TEST TO UNDERSTAND THE VALUE ABSOLUT ABOUT THE NULL HYPOTHESYS.

According Chen (2021) the z-test is also a hypothesis test in which the z-statistic follows a normal distribution. The z-test is best used for greater-than-30 samples because, under the central limit theorem, as the number of samples gets larger, the samples are considered to be approximately normally distributed.

The null and alternative hypotheses, as well as the alpha and z-score, should all be provided when doing a z-test. The statistical test should next be calculated, followed by the results and conclusion. A z-statistic, commonly known as a z-score, is a measure that describes how many standard deviations a score produced from a z-test is up or down mean population.

To develop this section, the separation of the data set into a sample was used, where the male and female groups were separated. Through the function

(groupby(data.columnname)) where we define what according to the result we can separate the data into Categories.

The picture shown the difference between the standard deviations from gender.

Figure 28– Checking values between male and female standart.

```
stdf = f_df['Credit_Limit'].std()

print('The standart deviation is {:.2f}'.format(stdf))
```

```
The standart deviation is 3163.39
```

```
stdm = m_df['Credit_Limit'].std()

print('The standart deviation is {:.2f}'.format(stdm))
```

```
The standart deviation is 10672.76
```

Regarding this the result about the mean values between female (first picture) and male (picture)has different analysis and is possible understand that there are variance between these values and the sample female has a result less than male gender.

Figure 29– Illustrate mean and distribution between credit limit by gender groups

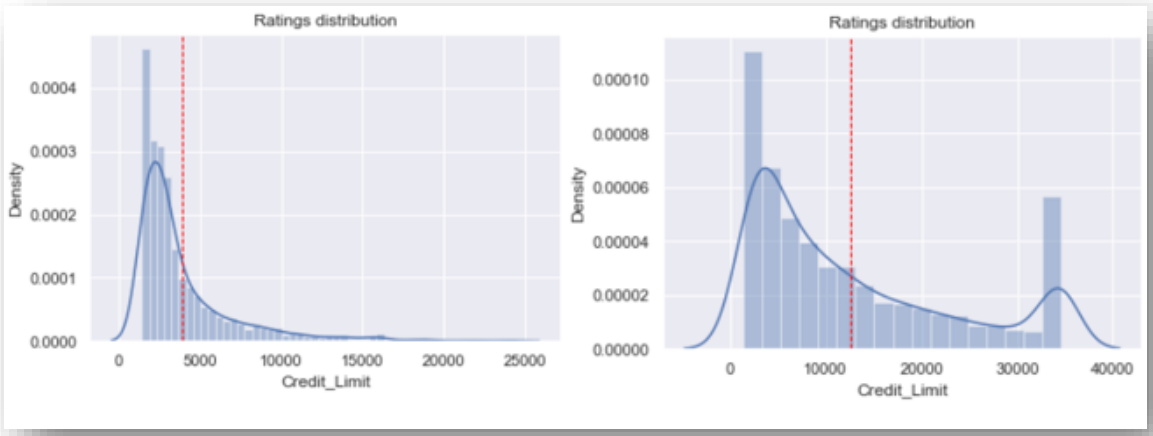


Figure 30–Getting result from the Scipy library and using stats models

```
import pandas as pd
from scipy import stats
from statsmodels.stats import weightstats as stests
ztest ,pval = stests.ztest(f_df['Credit_Limit'], x2=None, value=4000)
print(float(pval))
if pval<0.05:
    print("reject null hypothesis")
else:
    print("accept null hypothesis")
```

```
0.24386291566745577
accept null hypothesis
```

The hypotheses about the female analysis about the average Credit Limit is acceptable in terms of our development, getting the p-value 0.24.

5. SECOND SECTION

5.1 DESCRIPTION

Carry out a correlation analysis between 2 variables. Interpret your results and check if the correlation implies causation. Provide a short explanation and conclusion based on your findings.

5.2 CORRELATION AND CAUSATION

According Gregerson (2020) because both deal with relationships between variables, correlation and regression analysis are connected. The correlation coefficient is a measure of how closely two variables are related linearly. Value of correlation coefficient are always between 1 and 1. A correlation value of +1 shows a positive linear connections between two variables, a correlation coefficient of 1 indicates a negative linear relationship between two variables, and a correlation coefficient of 0 indicates no straight correlation between the two variables.

One of the variables visible for purchase through the Credit Limit directly is Open to Buy Credit Line, but analysing the influence with another variable, would be the relationship time with the bank. Variables as well as Age and Months on Book are temporal variables that, over time, build solid bases with the bank, and can influence positively or negatively. Analysing the entire sample audience and the data set showing the audience without gender distinction.

Figure 31–Using the pairplot (Kind= kde)

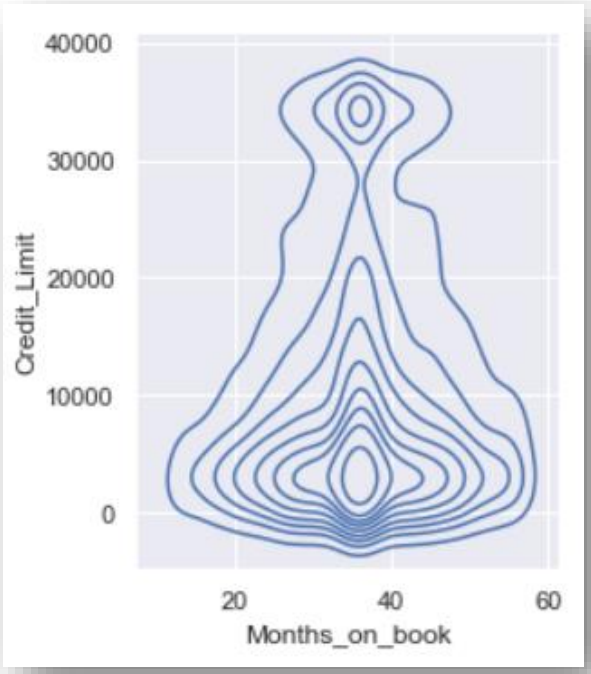


Figure 32 –Using the Correlation Table

Another way to understand correlation between variables is using correlation function, that will get the value between variables in a Pearson format, in that case the

results about our correlation is 0.02 between the variables Months_on Book and Credit Limit.

Figure 33 – Using another type of correlation

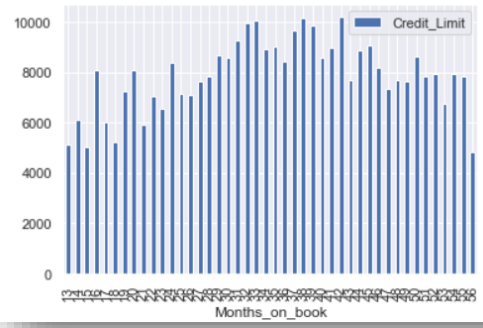
```
Months_on_book = bank["Months_on_book"]
Credit_Limit = bank["Credit_Limit"]
correlation = Months_on_book.corr(Credit_Limit)
```

```
print('The correlation between variables is {:.2f}'.format(correlation))
```

```
The correlation between variables is 0.02
```

```
bank.groupby('Months_on_book').mean().plot(y='Credit_Limit', kind='bar')
```

```
<AxesSubplot:xlabel='Months_on_book'>
```



5.3 CONCLUSION

According Sharma (2021) Correlation is also known as an association. It refers to a relation between two different entities or data points. When one thing goes up another comes down and vice-versa which means that they change together.

The most difficult element is verifying structural assumptions in the data. It require observational causal graph inference and testing around whether the variables being controlled render the intended causal impact detectable to do it properly without an expert

In the analysis of Months on book and Credit Limit variables to understand if the relationship time and their stay in the bank have Correlation and are also influenced by Causation. We can conclude that it has an approximate existence correlation in the value of 0.02, not being changed in a directly proportional factor. Therefore, the Cause relationship cannot be considered, as other relevant factors, that is, other variables may be factors that directly interfere with our objects of study in this second section.

6. THIRD SECTION

6. 1 RESEARCH

Pick two variables (different to the ones in question 1) and build a linear regression model that allows you to predict information about those variables. Interpret your results, and provide a short explanation and conclusion based on your findings.

The models used in the research were **Linear Regression**

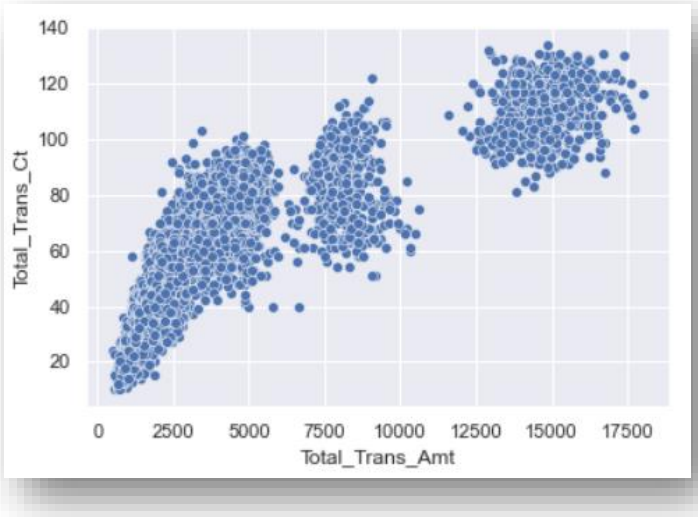
- **Linear Regression**

Linear regression has examined a set of predictor variables that perform well in predicting a dependent variable. That is, regression estimates are used to explain the relationship between a dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent variable and one independent variable is defined by the formula y = c + b * x, where y = estimated score of the dependent variable, c = constant, b = regression coefficient and x = score in the independent variable. (Müller, Guido, 2017)

6.2 LINEAR REGRESSION AND PREDICTION

According to the pairplot shown in the EDA section, we can see that some variables have linear projections such as "average_to_buy" and "Credit_Limit" but seeking to analyze other variables, and we can highlight that according to a quick survey, credit card spending has increased in recent years. With that, I decided to investigate the linear relation of the variable that brings us the total spent and the amount of spent operations of each client, and they are: "Total_Trans_Ct" and "Total Transaction Count." Remember that these variables' data are accounted for in the last 12 months).

Figure 34 – Using Scatter between variables "Total_Trans_Ct" and "Total Transaction Count."



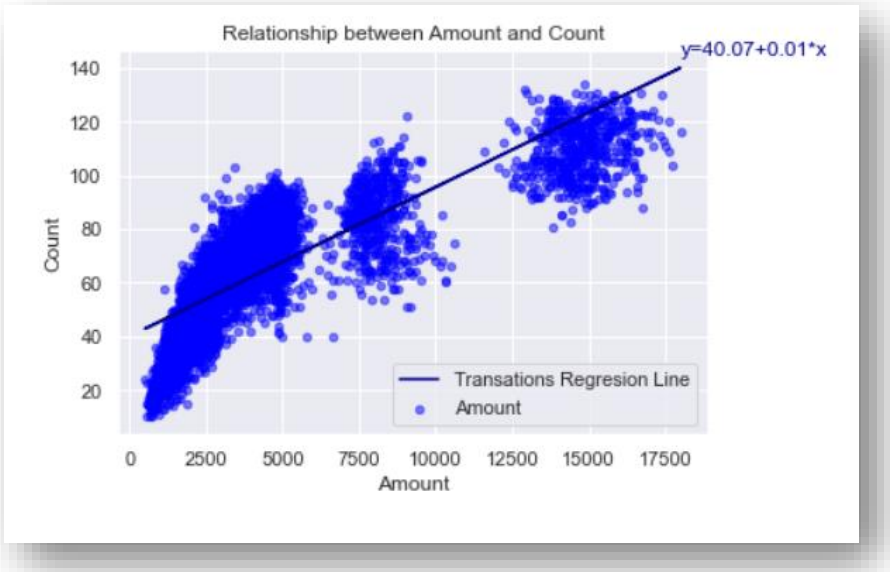

```
array([5.56076667e-03, 4.00676428e+01])
```

- Amount: A
- Count: C

Lineal Regression Model: **A = 5.56*C - 4.00**

A linear regression with a value of 5.56 for X and 4.00 for Y, thus showing the representation in the graph below starting from 40 Count having equation value of y= 40.07+0.01x represented through its regression line. It is noticeable that the values do not have a linear and clear distribution, but they are shown in the figure below.

Figure 35– Using Scatter line between variables "Total_Trans_Ct" and "Total Transaction Count."



The following example shows a prediction with a highlighted value of 230(chosen value) being changed and used according to the required prediction. The formula was highlighted through the Librarie Numpy.

Figure 36 – Using prediction mode by Numpy

7. CONCLUSION/RECOMMENDATIONS

Regarding that it developed the work to verify Hypotheses, Correlations, use a linear regression model and prepare it for prediction. The data set variables to have continuous relationships. However, it is necessary to verify which other models perform better in each situation and variable to be analyzed for a more detailed study.

In the first section, Z-test and Chi-square test models were used to define relationships between Categorical Variables and Numerical Variables separately. Thus, we can observe in Section II that although we have a credit limit as an independent factor, it is partially influenced by the time the customer is at the bank. However, through research, we can be clear that other factors can influence, more or less, as we saw highlighted in the Correlation table.

Its last section uses the function to perform the Linear Regression Illustration between the Expenses variables and being it the total amount spent and its count per operation. So hoping to find future predictions, results that can bring if they can be confirmed according to the data set used.

REFERENCES

Chen, James (2021) Z-Test. Avaiable at: https://www.investopedia.com/terms/z/z-test.asp (Accessed: 29 December 2021).

Gregerson (2020) Correlation Vs Causation, Avaiable at: https://towardsdatascience.com/correlation-vs-causation-3e3481c71fef (Accessed: 29 December 2021).

Konish, Lorie (2021) *Men tend to have higher total credit limits than female borrowers*, research finds. Avaiable at https://www.cnbc.com/2021/11/09/men-tend-to-have-higher-credit-limits-than-female-borrowers.html (Accessed: 26 December 2021).

Kuria, Derrick (2021) *DATA AND DATA COLLECTION***.** Avaiable at https://medium.com/@derrickkuria44/data-and-data-collection-7c929495d09a. (Accessed: 23 December 2021).

Majaski ,Christina (2021) Hypothesis Testing. Avaiable at: https://www.investopedia.com/terms/h/hypothesistesting.asp. (Accessed: 29 December 2021).

Müller, A.C. Guido, S. (2017) Introduction to Machine Learning with Python, A Guide for Data Scientists. USA. O'Reilly.

Sharma, Himanshu (2021) Statistics, Avaiable at: https://towardsdatascience.com/correlation-vs-causation-3e3481c71fef (Accessed: 22 December 2021).