PREPARE AND EXPLORE THE DATA FOR DEFAULTS ON LOANS

CREDIT ONE

Prepared for **Guido Rossum**
Prepared by **Lucia Oviedo**
July14, 2020

TABLE OF CONTENTS

Overview *3*

Getting to know our data 3

Investigative Questions: 4

Prepare and Explore the data *5*

What recommendations would you give to the Guido regarding your findings? 5

Data Types 5
Statistical Analysis 6
Correlation 6
Covariance 7
GENDER 7
AGE 7
AGE Vs GENDER 7
EDUCATION 8
MARRIAGE 8
BILL_AMT 8
PAY_AMT 9
DEFAULT VS LIMIT / EDUCATION 9
DEFAULT VS LIMIT / AGE 10
DEFAULT VS LIMIT_BAL / PAY_1 10
DEFAULT VS BILL_AMT / PAY_1 11

Did you learn anything of potential business value from this analysis? 11

What are the main lessons you've learned from this experience? 11

OVERVIEW

Problem:

1. Increase in customer default rates - This is bad for Credit One since we approve the customers for loans in the first place.
2. Revenue and customer loss for clients and, eventually, loss of clients for Credit One

Getting to know our data

We have a data set with 30000 entries and 25 columns

Attribute Information:

LIMIT_BAL Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.

SEX (1 = male; 2 = female).

EDUCATION (1 = graduate school; 2 = university; 3 = high school; 0, 4, 5, 6 = others). *** unify others to one unique value

MARRIAGE (1 = married; 2 = single; 3 = divorce; 0=others).

AGE (year).

PAY_1 ... PAY_6 History of past payment (from April to September, 2005)
-2: No consumption;
-1: Paid in full;
0: The use of revolving credit;
1 = payment delay for one month;
2 = payment delay for two months; . . .;
8 = payment delay for eight months;
9 = payment delay for nine months and above.

BILL_AMT1 ... BILL_AMT6 Amount of bill statement (NT dollar) (September 2005 to April 2005)

PAY_AMT1 ... PAY_AMT6 Amount of previous payment (NT dollar) (September 2005 to April 2005)

DEFAULT Y=0 then not default, Y=1 then default"

Investigative Questions:

1. How do you ensure that customers can/will pay their loans? Can we do this?
   There is no way to ensure that customers will pay their loans. But we can make sure that they can pay their loans based on their income and limiting their credit balance.

2. As you progress through the tasks at hand begin thinking about how to solve this problem. Here are some lessons we learned form a similar problem we addressed last year:
   We cannot control customer spending habits, but we can control limit credit balance and we know our customer payment behavior, since we have their history payments. So, based on that we can predict if the customer is going to pay next month or not.

3. We must on the problem(s) we can solve: What attributes in the data can we deem to be statistically significant to the problem at hand? PAY_1, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6, EDUCATION, AGE

4. What concrete information can we derive from the data we have? We will answer this question in the report

5. What proven methods can we use to uncover more information and why? We can create models implementing Machine Learning techniques that will help us to predict which customers will be DEFAULT on their payments.

PREPARE AND EXPLORE THE DATA

What recommendations would you give to the Guido regarding your findings?

Data Types

Nominal values: SEX, EDUCATION, MARRIAGE, Default Payment next month (not standard name needs to get renamed), PAY_0 .... PAY_6

Numeric values: LIMIT_BAL, AGE, BILL_AMT1 .... BILL_AMT6, PAY_AMT1 .... PAY_AMT6

	count	mean	std	min	25%	50%	75%	max
LIMIT_BAL	30000.0	167484.322667	129747.661567	10000.0	50000.00	140000.0	240000.00	1000000.0
SEX	30000.0	1.603733	0.489129	1.0	1.00	2.0	2.00	2.0
EDUCATION	30000.0	1.842267	0.744494	1.0	1.00	2.0	2.00	4.0
MARRIAGE	30000.0	1.551867	0.521970	0.0	1.00	2.0	2.00	3.0
AGE	30000.0	35.485500	9.217904	21.0	28.00	34.0	41.00	79.0
PAY_1	30000.0	-0.016700	1.123802	-2.0	-1.00	0.0	0.00	8.0
PAY_2	30000.0	-0.133767	1.197186	-2.0	-1.00	0.0	0.00	8.0
PAY_3	30000.0	-0.166200	1.196868	-2.0	-1.00	0.0	0.00	8.0
PAY_4	30000.0	-0.220667	1.169139	-2.0	-1.00	0.0	0.00	8.0
PAY_5	30000.0	-0.266200	1.133187	-2.0	-1.00	0.0	0.00	8.0
PAY_6	30000.0	-0.291100	1.149988	-2.0	-1.00	0.0	0.00	8.0
BILL_AMT1	30000.0	51223.330900	73635.860576	-165580.0	3558.75	22381.5	67091.00	964511.0
BILL_AMT2	30000.0	49179.075167	71173.768783	-69777.0	2984.75	21200.0	64006.25	983931.0
BILL_AMT3	30000.0	47013.154800	69349.387427	-157264.0	2666.25	20088.5	60164.75	1664089.0
BILL_AMT4	30000.0	43262.948967	64332.856134	-170000.0	2326.75	19052.0	54506.00	891586.0
BILL_AMT5	30000.0	40311.400967	60797.155770	-81334.0	1763.00	18104.5	50190.50	927171.0
BILL_AMT6	30000.0	38871.760400	59554.107537	-339603.0	1256.00	17071.0	49198.25	961664.0
PAY_AMT1	30000.0	5663.580500	16563.280354	0.0	1000.00	2100.0	5006.00	873552.0
PAY_AMT2	30000.0	5921.163500	23040.870402	0.0	833.00	2009.0	5000.00	1684259.0
PAY_AMT3	30000.0	5225.681500	17606.961470	0.0	390.00	1800.0	4505.00	896040.0
PAY_AMT4	30000.0	4826.076867	15666.159744	0.0	296.00	1500.0	4013.25	621000.0
PAY_AMT5	30000.0	4799.387633	15278.305679	0.0	252.50	1500.0	4031.50	426529.0
PAY_AMT6	30000.0	5215.502567	17777.465775	0.0	117.75	1500.0	4000.00	528666.0
DEFAULT	30000.0	0.221200	0.415062	0.0	0.00	0.0	0.00	1.0

Statistical Analysis

- LIMIT_BAL mean value of credit amount is 167484.322667
- AGE mean value of years old that ask for credit is 35.485500
- BILL_AMT average goes from 38871.760400 to 51223.330900 and it is increasing every month
- PAY_AMT average of the pay amount goes from 5215.502567 to 5921 and it is increasing every month

Incongruence found:

Maximum LIMIT_BAL US 1,000,000 and max BILL_AMT3 is 1,664,089 witch is over the LIMIT_BAL.

This observation has a LIMIT_BAL of 500000 but bill_amt3 is of 1664089 (three times more than the LIMIT_BAL). PAY_AMT2 is of $1,684,259 getting a negative balance.

A deeper analysis of all the BILL_AMT that is greater than LIMIT_BAL plus 30% of loan interest was made. And 498 observations out of 30,000 has a bill amount much greater than the limit balance (1.6% of the Data Set) Need to follow and client should identify the root cause of this flaw.

Correlation

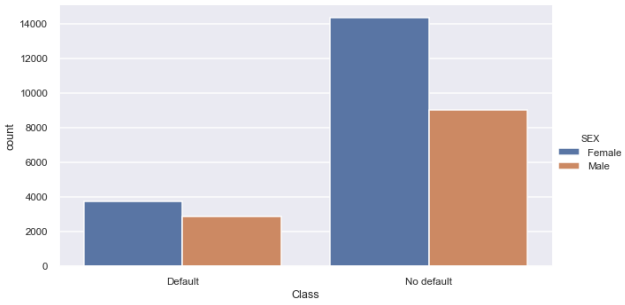
	DEFAULT
DEFAULT	1
PAY_1	0.324794
PAY_2	0.263551
PAY_3	0.235253
PAY_4	0.216614
PAY_5	0.204149
PAY_6	0.186866
EDUCATION	0.033842
AGE	0.01389
BILL_AMT6	-0.005372
BILL_AMT5	-0.00676
BILL_AMT4	-0.010156
BILL_AMT3	-0.014076
BILL_AMT2	-0.014193
BILL_AMT1	-0.019644
MARRIAGE	-0.024339
SEX	-0.039961
PAY_AMT6	-0.053183
PAY_AMT5	-0.055124
PAY_AMT3	-0.05625
PAY_AMT4	-0.056827
PAY_AMT2	-0.058579
PAY_AMT1	-0.072929
LIMIT_BAL	-0.15352

Covariance

- Positive trends of DEFAULT: EDUCATION, AGE, PAY_1 .... PAY_6
- Negative trends of DEFAULT: LIMIT_BAL, SEX, MARRIAGE, BILL_AMT1 ... BILL_AMT6, PAY_AMT1 ... PAY_AMT6

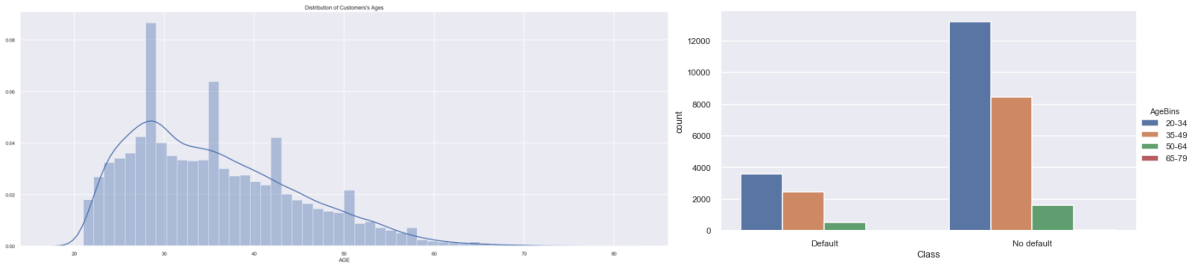
GENDER

- As we already know from Correlation, there is no relation on SEX and default behavior.
- There is a higher number of defaults for females but also there is a higher number of no defaults for females



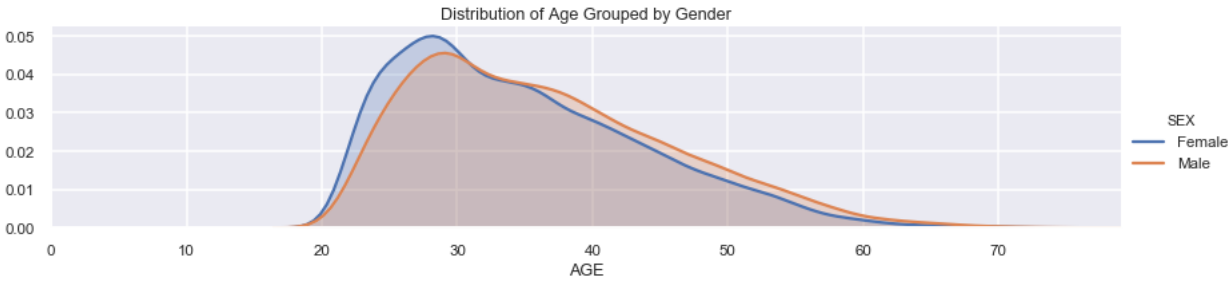
AGE

- Customers with less than 30 years old has a higher count for credit loans in the data set.
- Customers from 20-34 years old has a higher default count in the data set, followed by 35-49 and 50-64.



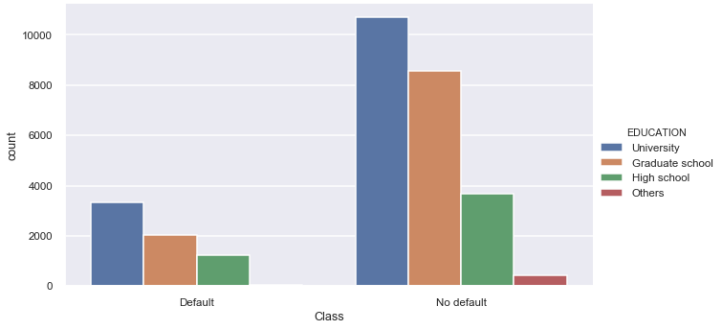
AGE Vs GENDER

- Distribution is pretty much the same, there are more females count for credit loans in the data set



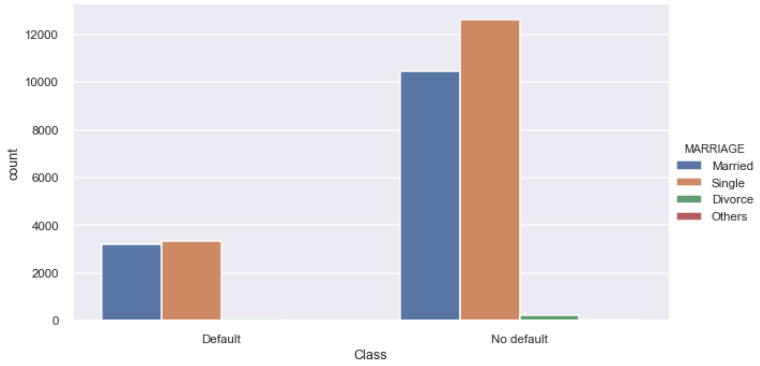
EDUCATION

- Customers with a University degree has a higher default count in the data set, followed by Graduate school and High School degree.
- Others has not Default data



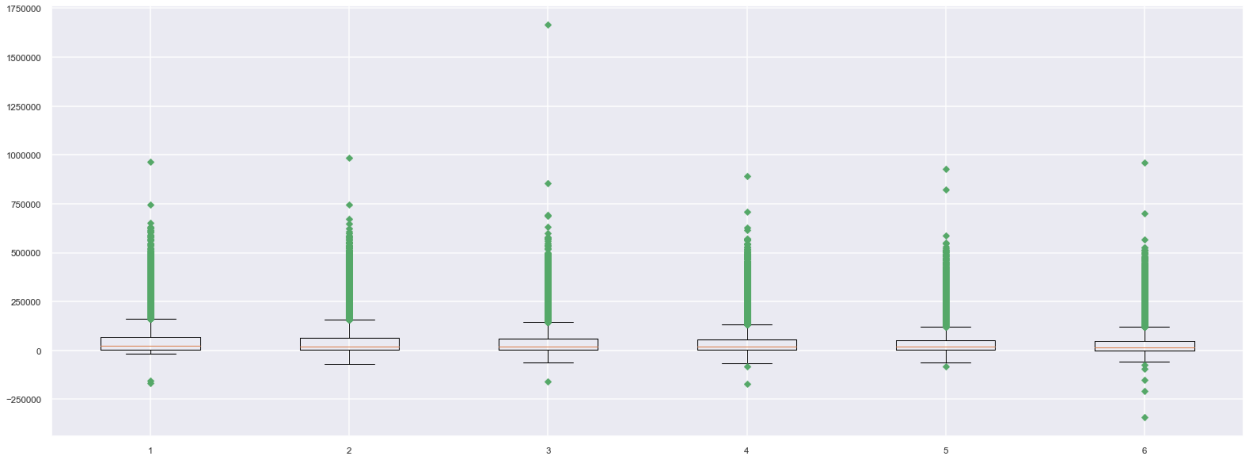
MARRIAGE

- As we already detected in correlation, Marriage doesn't have a correlation with default



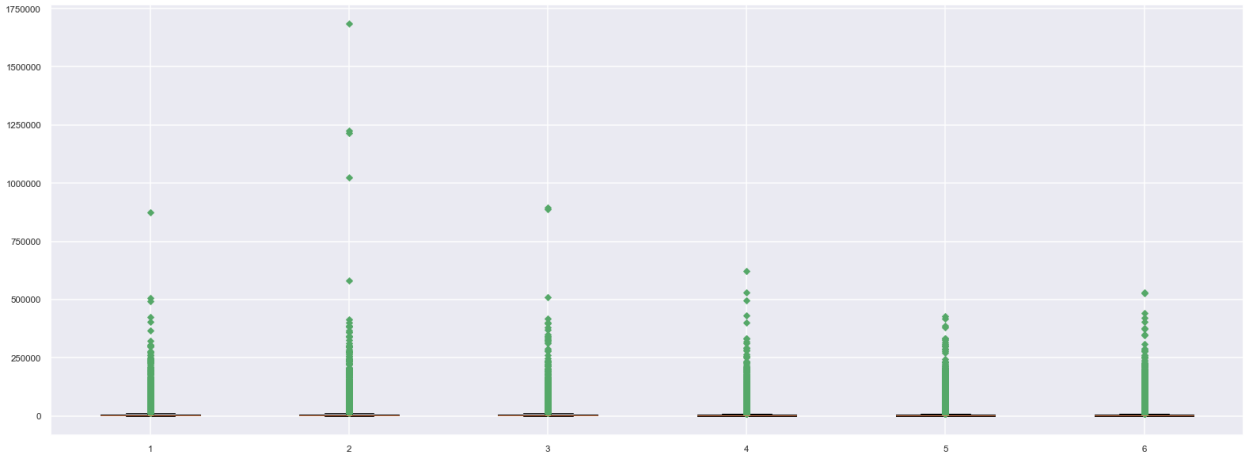
BILL_AMT

- BILL_AMT looks to be constant across all months.
- BILL_AMT1 minimum value is very close to 0, all others has a negative amount.



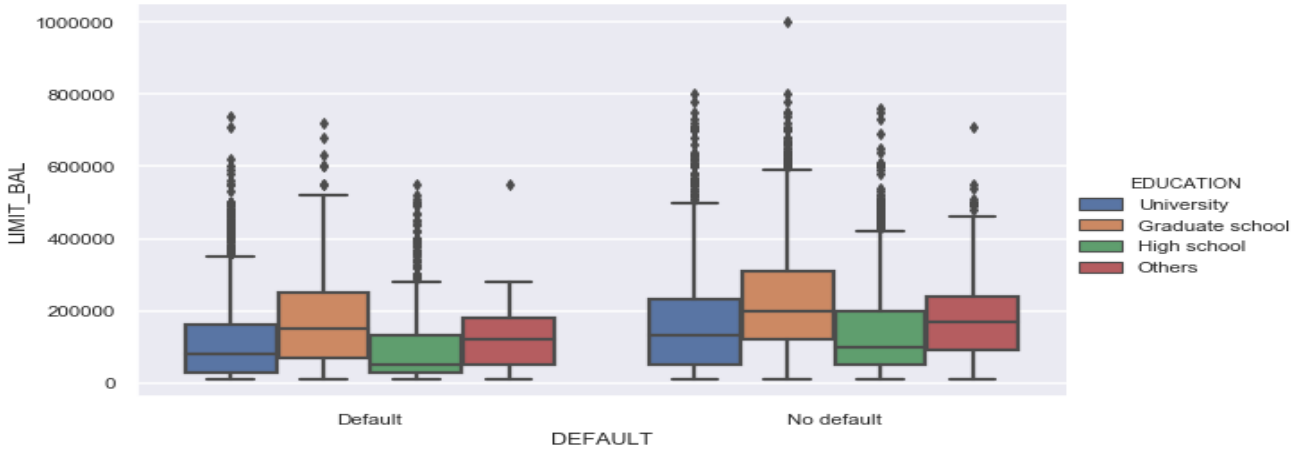
PAY_AMT

- Most PAY_AMT are low, we cannot really appreciate them in a boxplot
- Few PAY_AMT are high, more in PAY_AMT2



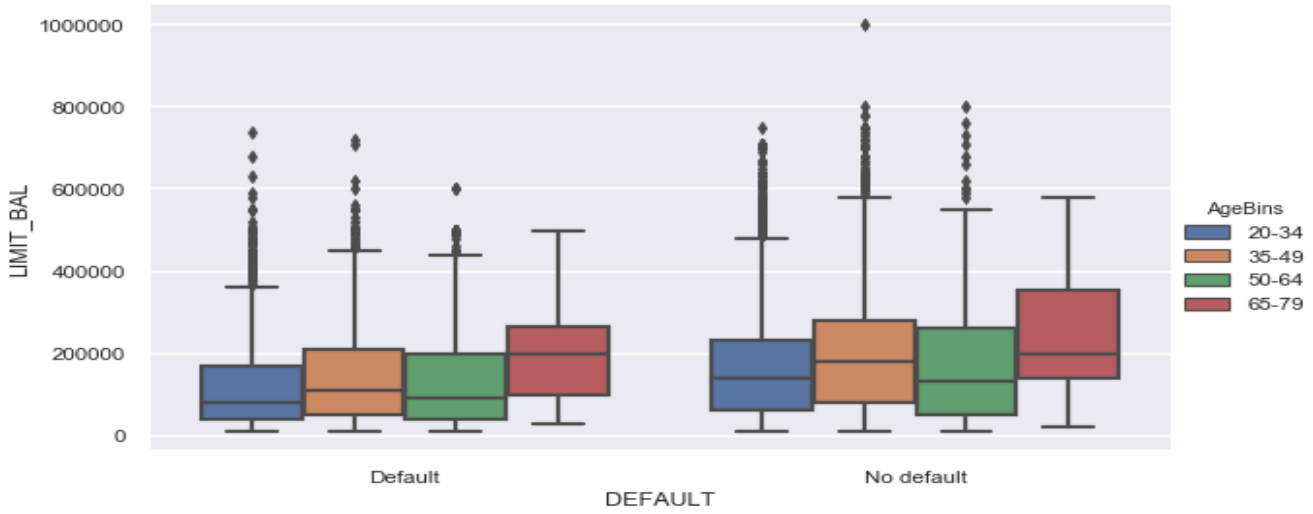
DEFAULT VS LIMIT / EDUCATION

- In average customers who has default has a LIMIT_BAL AVERAGE OF 200,000
- In average customers who are Graduate School has a higher LIMIT_BAL, followed by OTHERS
- In average customers who DEFAULD and are Graduate School has a higher LIMIT_BAL
- In average customers who DEFAULD and are High School has a lower LIMIT_BAL
-



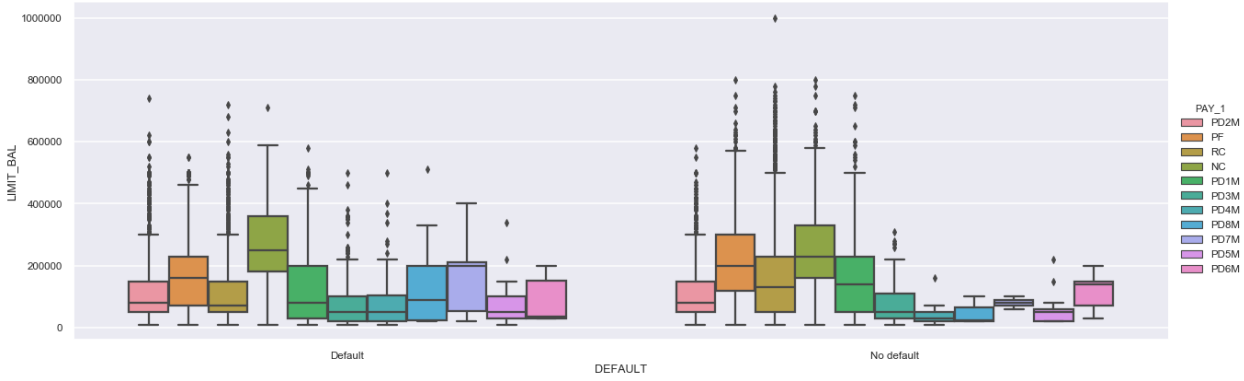
DEFAULT VS LIMIT / AGE

- In average customers from 65 -79 years old has a higher LIMIT_BAL of 200000, followed by 35-49
- In average customers in the range 35-49 and 50-64 has the lowest and almost the same LIMIT_BAL
- In average customers who DEFAULT has a lower LIMIT_BAL than people who don't DEFAULT



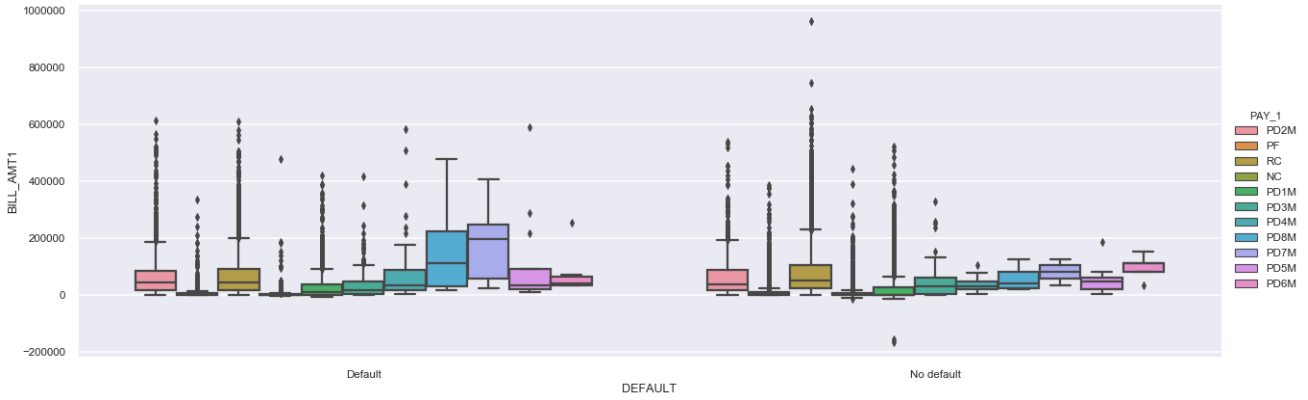
DEFAULT VS LIMIT_BAL / PAY_1

- Pay in Full is found under Default Data, it might be an inconsistency
- I would suggest lower the Limit credit amount for those customers that has a delayed payment of 5 or more months. We can compare them against those customers who doesn't default and those who actually DEFAUL has a higher LIMIT_BAL. Same goes for PAY_2 … PAY_8



DEFAULT VS BILL_AMT / PAY_1

- As expected, we can detect a bigger BILL_AMT1 for payment delay for 4 months to 8 months
- Pay in full is low for PAY_AMT1



Did you learn anything of potential business value from this analysis?

Yes, I learned how to create useful visualization to understand the data, create some insights, find errors in the data and discover some patterns.

I learn how to interpreter boxplot and create visualization for more than two variables.

What are the main lessons you've learned from this experience?

The main lessons that I learn is how to create an Exploratory Data Analysis as part of a initial stage of data analysis with the help of visualization and statistics about each variable. Also, I learn how to create visualization from 2 or more variables and be able to extract some important information