

DATA-FLOW ANALYSIS

CREDIT ONE



CONTENTS

Goal

Data science process framework

Descriptions and location of related data sources

Manage the data for the project

Known issues with the data

Flowchart visualizing the detailed process

Initial insights of the data



Over the past year or so Credit One has seen an increase in the number of customers who have defaulted on loans they have secured from various partners, and Credit One, as their credit scoring service, could risk losing business if the problem is not solved right away.



The objective of our Data Science team is to design and implement a creative, empirical solution that assist Credit One improve their credit scoring service.



Primary Goal is to reduce the number of customers that defaulted on loans by at least 10%, creating and implementing a model that predicts which loan applicants are likely to default.

GOAL

DATA SCIENCE PROCESS FRAMEWORK

FRAMEWORK - ZUMELAND
MOUNT

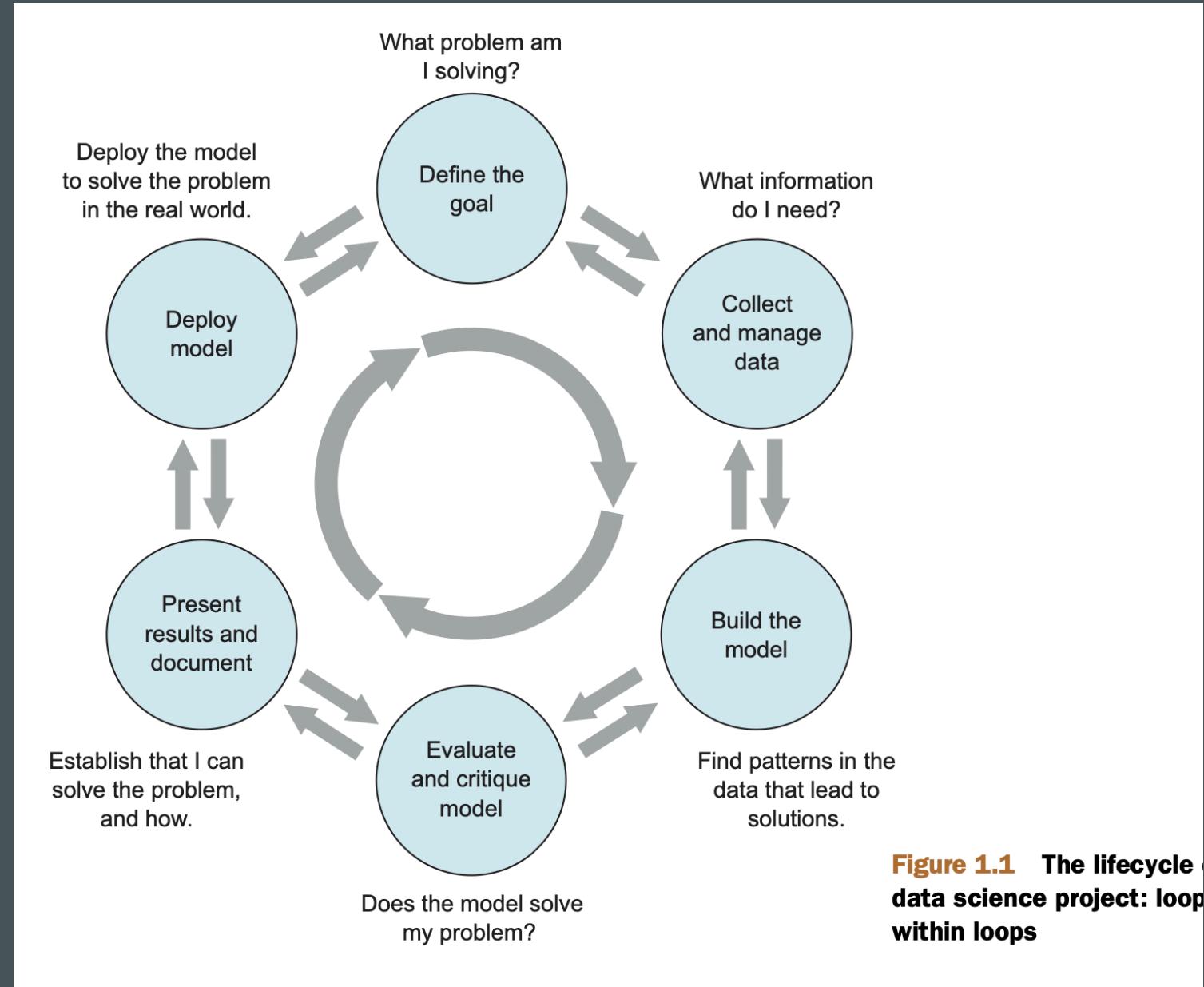


Figure 1.1 The lifecycle of a data science project: loops within loops

I am proposing to follow Zumel and Mount Framework because it separate the data science process within 6 important stages that guides us through the whole process step by step.

It is important to keep in mind that the boundaries between the stages are fluid, and the activities of one stage will often overlap those of other stages. And always we will go back and forth before moving into a new stage.

Other important reason Zumel and Mount was selected is because at the end of one project may lead into a follow-up project and since we already have the stages defined the whole process becomes more practical.

FRAMEWORK - ZUMEL AND MOUNT

FRAMEWORK - ZUMEL AND MOUNT



Define

Define the Goal
•Primary Goal is to reduce the number of customers that defaulted on loans by at least 10%, creating and implementing a model that predicts which loan applicants are likely to default.



Collect and Manage

Collect and Manage Data
•General, History of last Payment, Amount of bill statement, Amount of previous payment, clients behaviour.
•Credit score, income type of loan, Credit score, APR, Available credit, Credit utilization, collections are not included in the dataset..



Build

Build the Model
•Training set of 75% and Test set 25%
•Classification model will be evaluated: C5.0, Random Forest, SVM, kNN, Decision trees.



Evaluate and Critique

Evaluate and Critique Model
•Evaluation of the model. Is the goal met?
•Is the model accurate enough to meet the stakeholders' needs?
•Does it perform better than "the obvious guess" and any techniques being used currently?
•Do the results of the model make sense in the context of the real-world problem domain?



Present

Present Results and Document
•How should stakeholders interpret the model?
•How confident should they be in its predictions?
•When should they potentially overrule the model's predictions?



Deploy

Deploy Model
•How is the model to be handed off to "production"?
•How often, and under which circumstances, should the model be revised?

DESCRIPTIONS AND LOCATION OF RELATED DATA SOURCES

Amount
of the
Given
Credit

General

- Gender
- Education
- Marital Status
- Age

History of
last
Payment

- April to September

Amount
of Bill
Statement

- April to September

Amount
of
previous
payment

- April to September

Client's
Behaviour

- Default
- No default

IS THE DATA QUALITY GOOD ENOUGH?



Type of Loan



Credit Score



APR



Available credit



Credit Utilization



Number of
Collections

- Some additional factors that will help us determine if a loan should be granted

MANAGE THE DATA FOR THE PROJECT



Compliant with Data Privacy Regulation: Personal data privacy and protection is our priority in this organization.

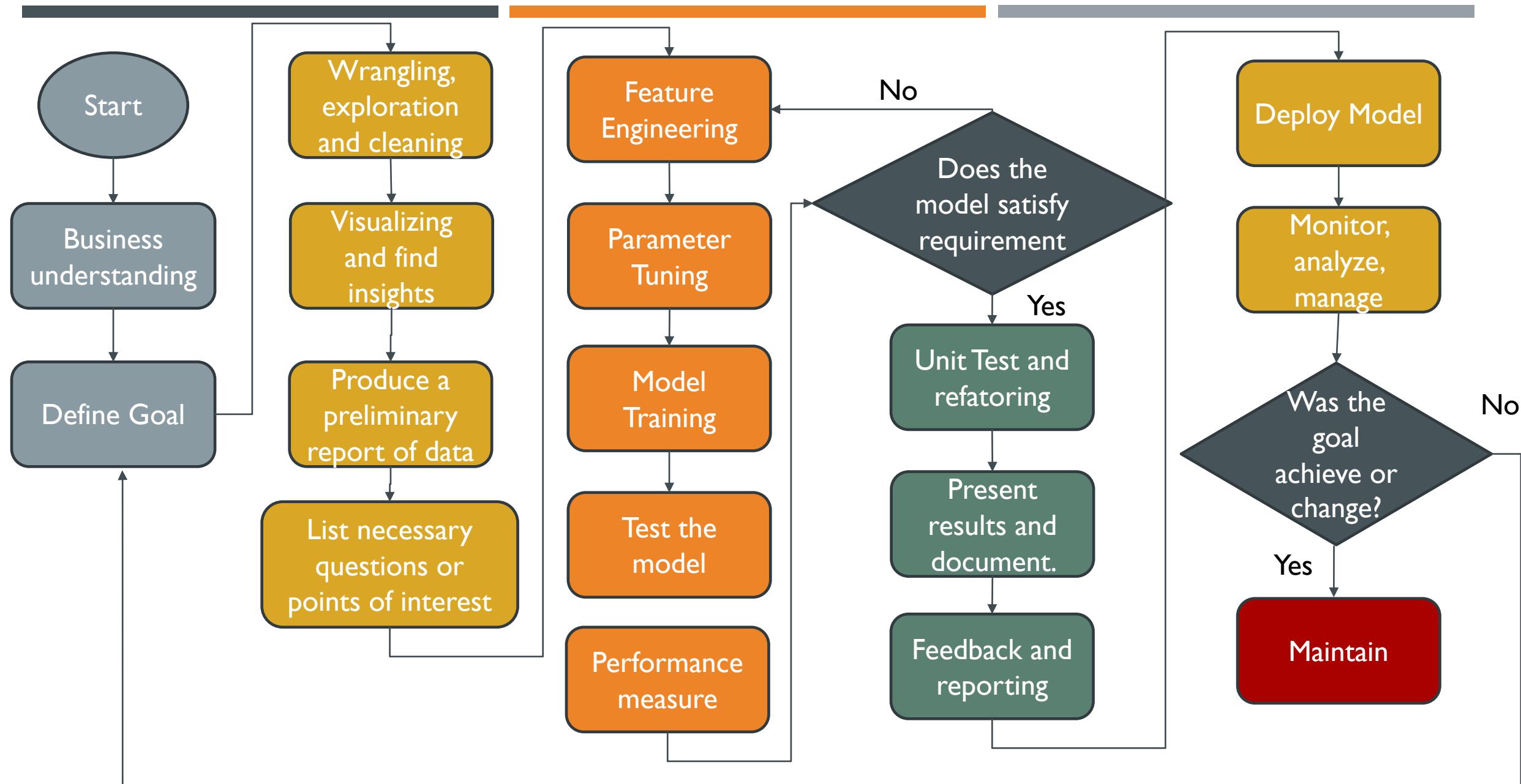


Cleaning the Data: The task of cleaning the data includes the removal of redundant, missing words, duplication, and redundant data.

KNOWN ISSUES WITH THE DATA

- No missing data
- PAY_0 needs be renamed to PAY_I
- “default payment next month” needs to be renamed
- Additional questions:
 - Why when the payment was paid in full, we have Default data under history of last payments?

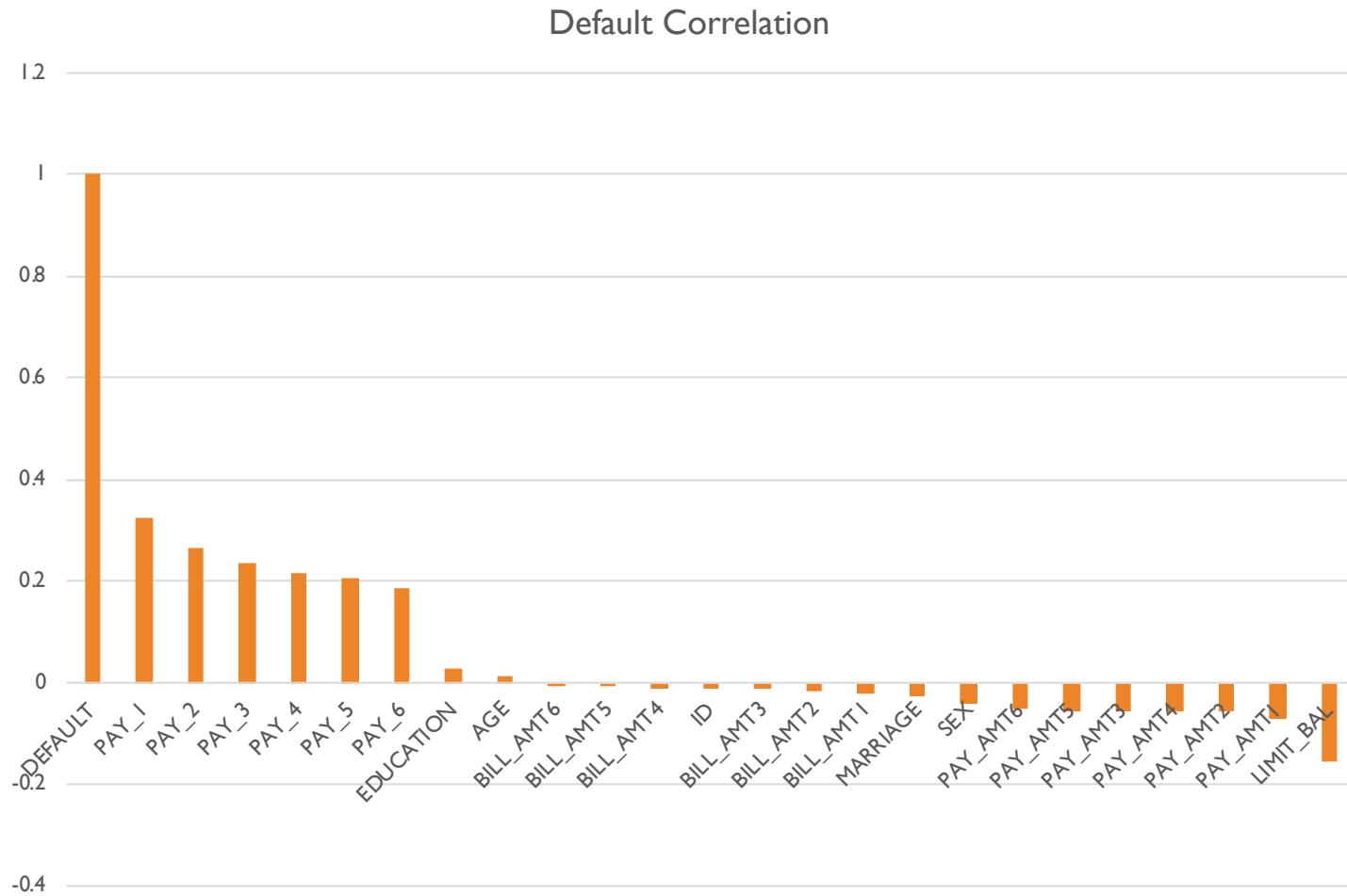
DATA SCIENCE FLOWCHART VISUALIZING THE DETAILED PROCESS





INITIAL INSIGHTS OF THE DATA

CORRELATION



INSIGHTS OF THE DATA



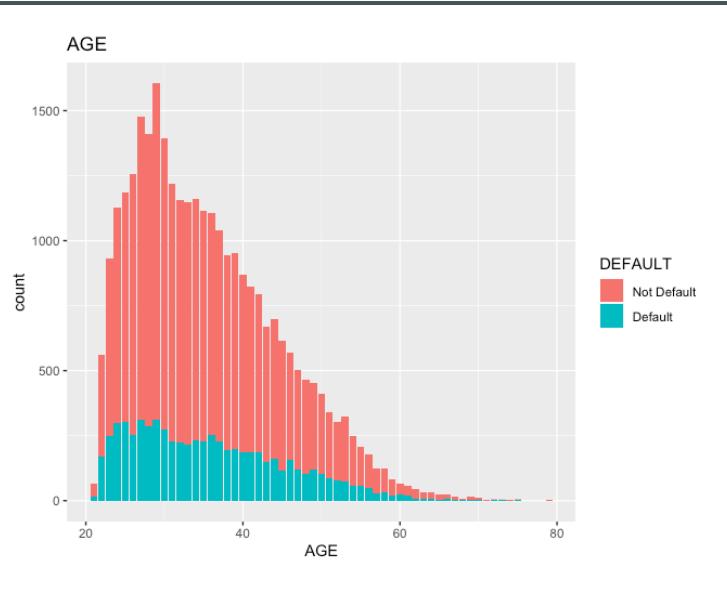
- Education
 - University group as the higher default count
- Marital Status
 - Marriage and single marital status has almost the same default count

INSIGHTS OF THE DATA



■ Gender

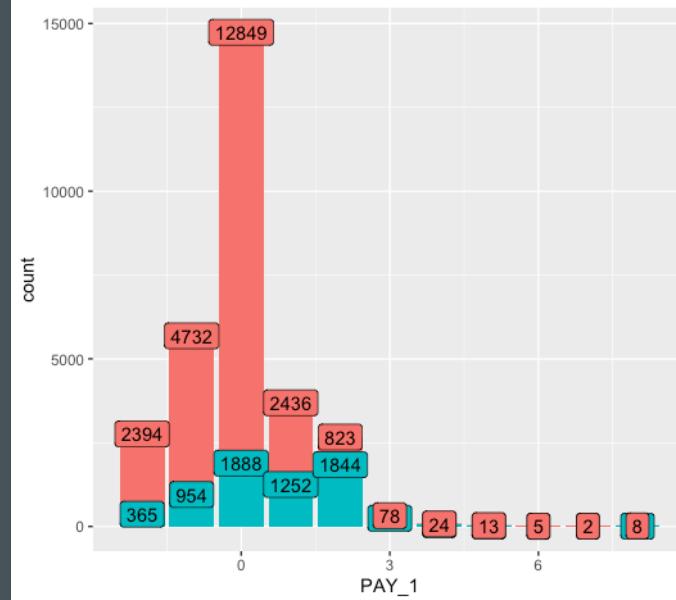
- Females has a higher default count
- Females has a higher not default count
- In general, there are more loans for females



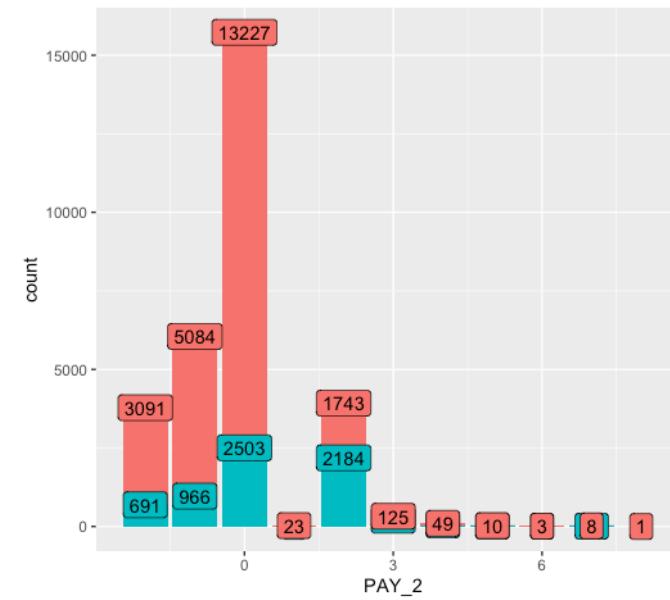
■ Age

- Default count is higher for the group less than 30 years old.
- Default count decreases as age increases

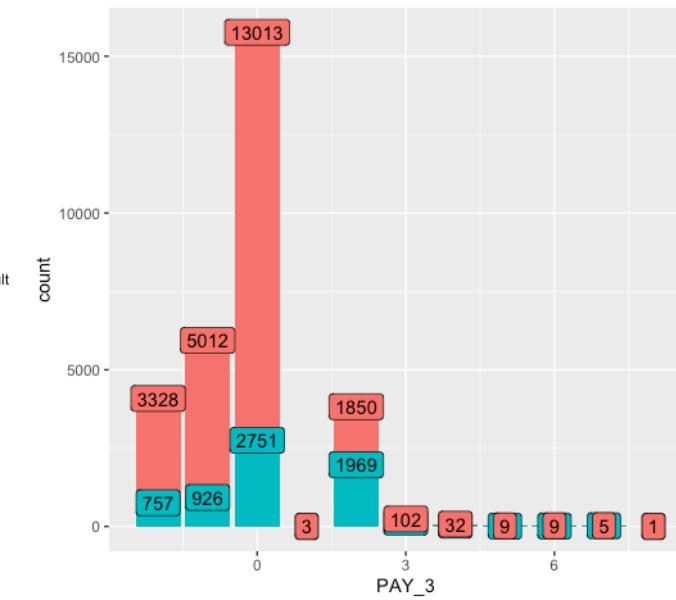
Past Monthly Payment - September



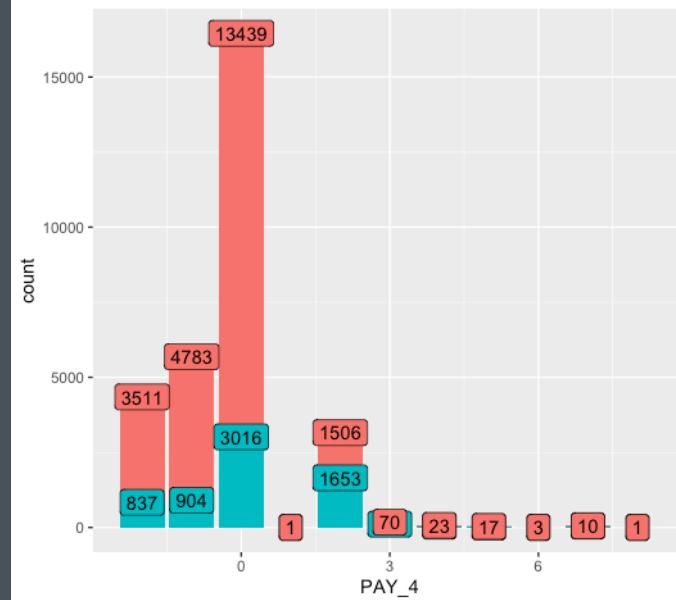
Past Monthly Payment - August



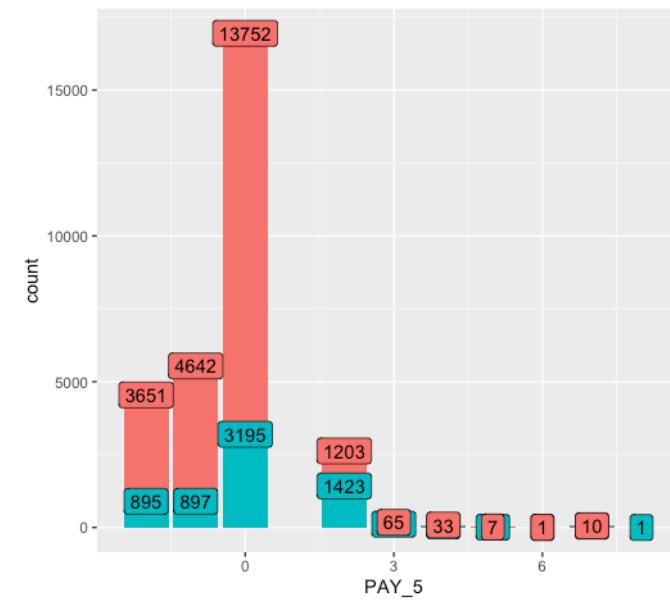
Past Monthly Payment - July



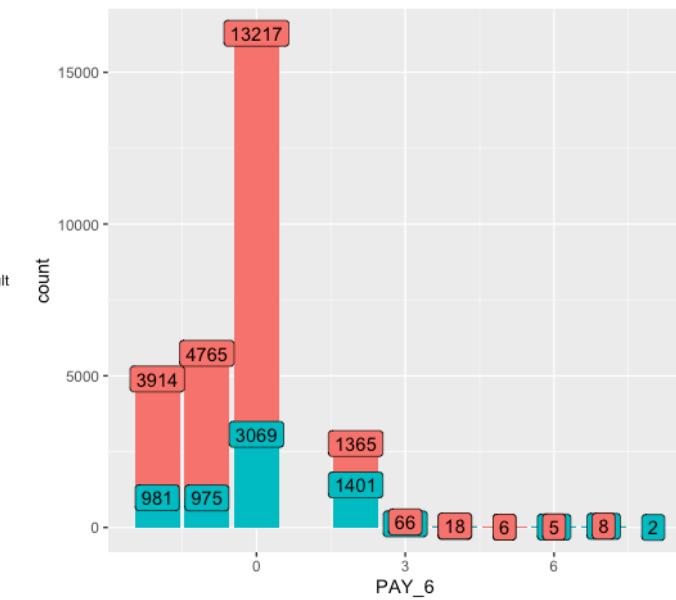
Past Monthly Payment - June



Past Monthly Payment - May



Past Monthly Payment - April



DEFAULT
a Not Default
a Default

DEFAULT
a Not Default
a Default

INSIGHTS OF THE DATA

- -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;
- Questions
 - Why if the payment was paid in full, we have Default data?
 - Insights
 - Payment delay for one month is higher in September
 - Payment delay for two months is higher in July
 - Payment delay for more than two months is in average less than 100

QUESTIONS

