DATA-FLOW ANALYSIS

CREDIT ONE

CONTENTS

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 - ZUMEL AND MOUNT

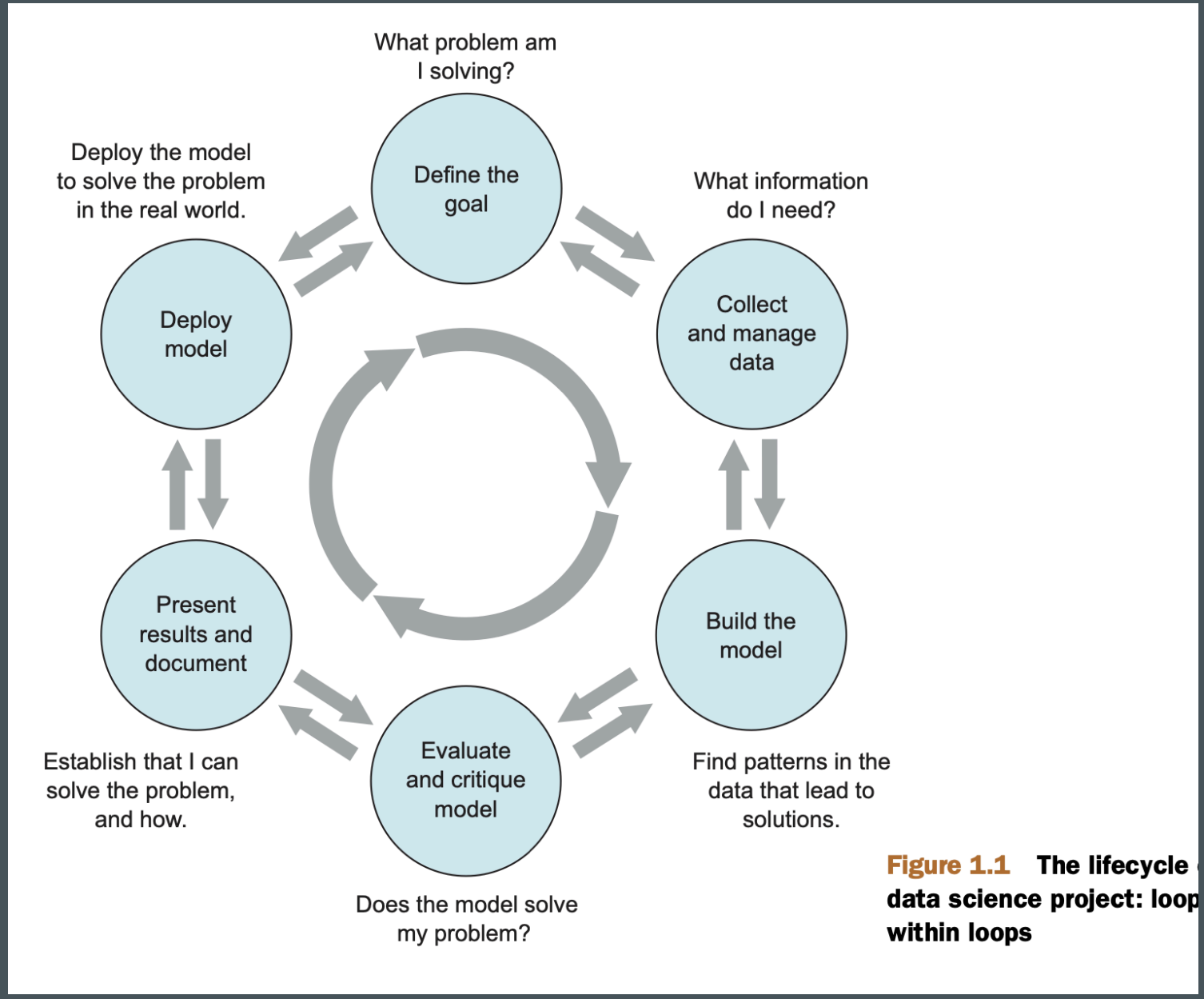


Figure 1.1 The lifecycle data science project: loop within loops

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 las Payment, Amount of bill statement, Amount of previous payment, clients behaivour.
- Credit score, income typo 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, kkn, Decision trees.

Evaluate and Critique

Evaluate and Critique Model

- Evaluation of the model. Is the goal meet?
- 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.

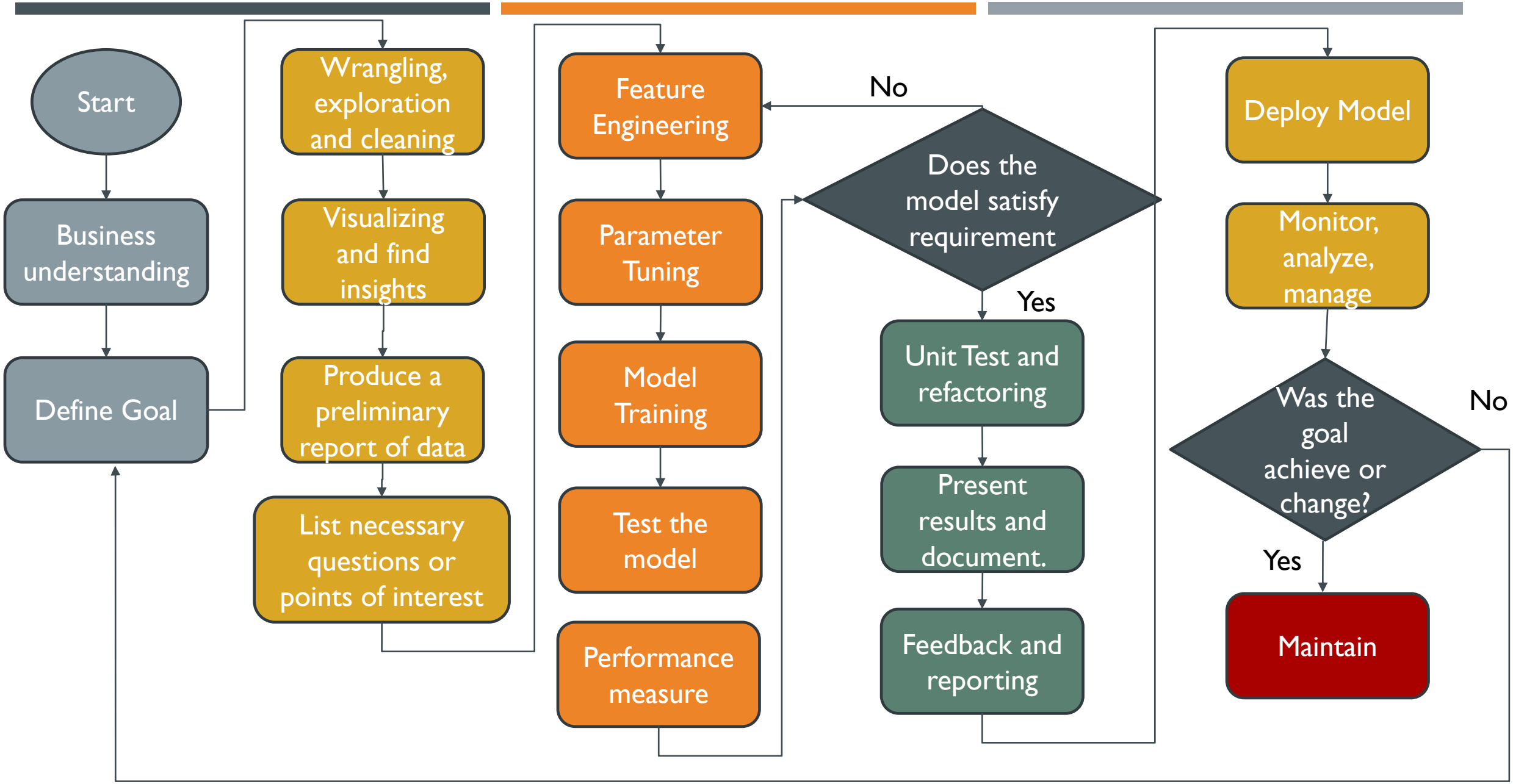
Cleaning the Data: The task of cleaning the data includes the removal of redundant, missing words, duplication, and redundant data.

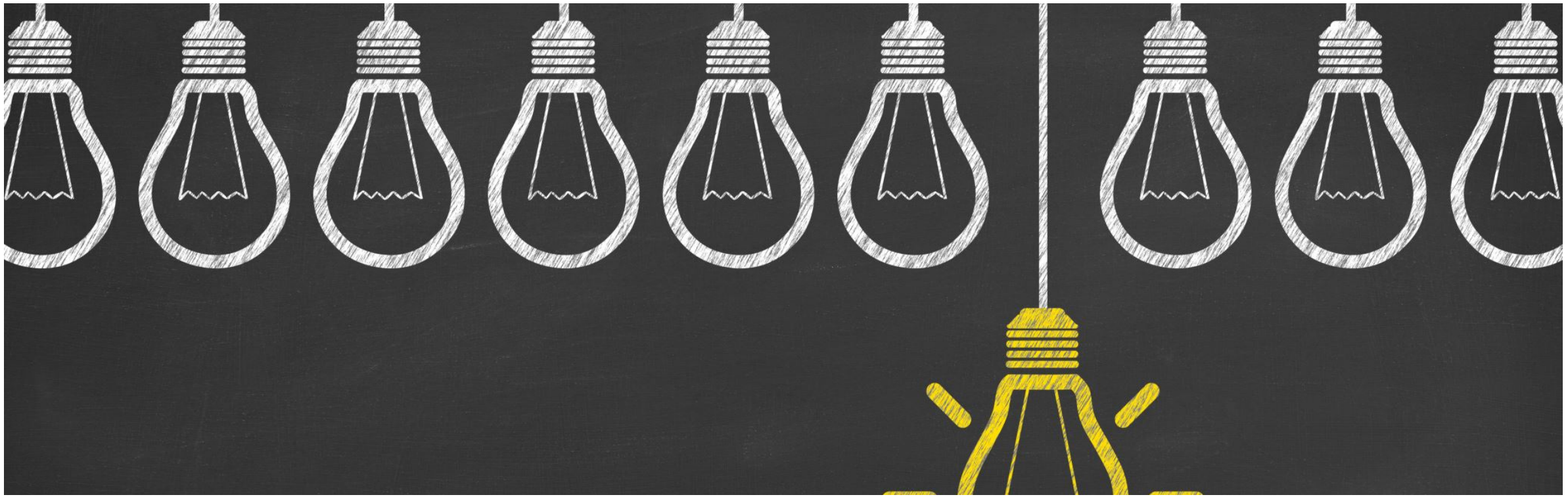
KNOWN ISSUES WITH THE DATA

- No missing data
- PAY_0 needs be renamed to PAY_1
- "default payment next month" needs to be renamed
- Double headers in the file, we can get rid of X1 …. X23 and Y
- ID is not relevant.
- We can create a group for age and

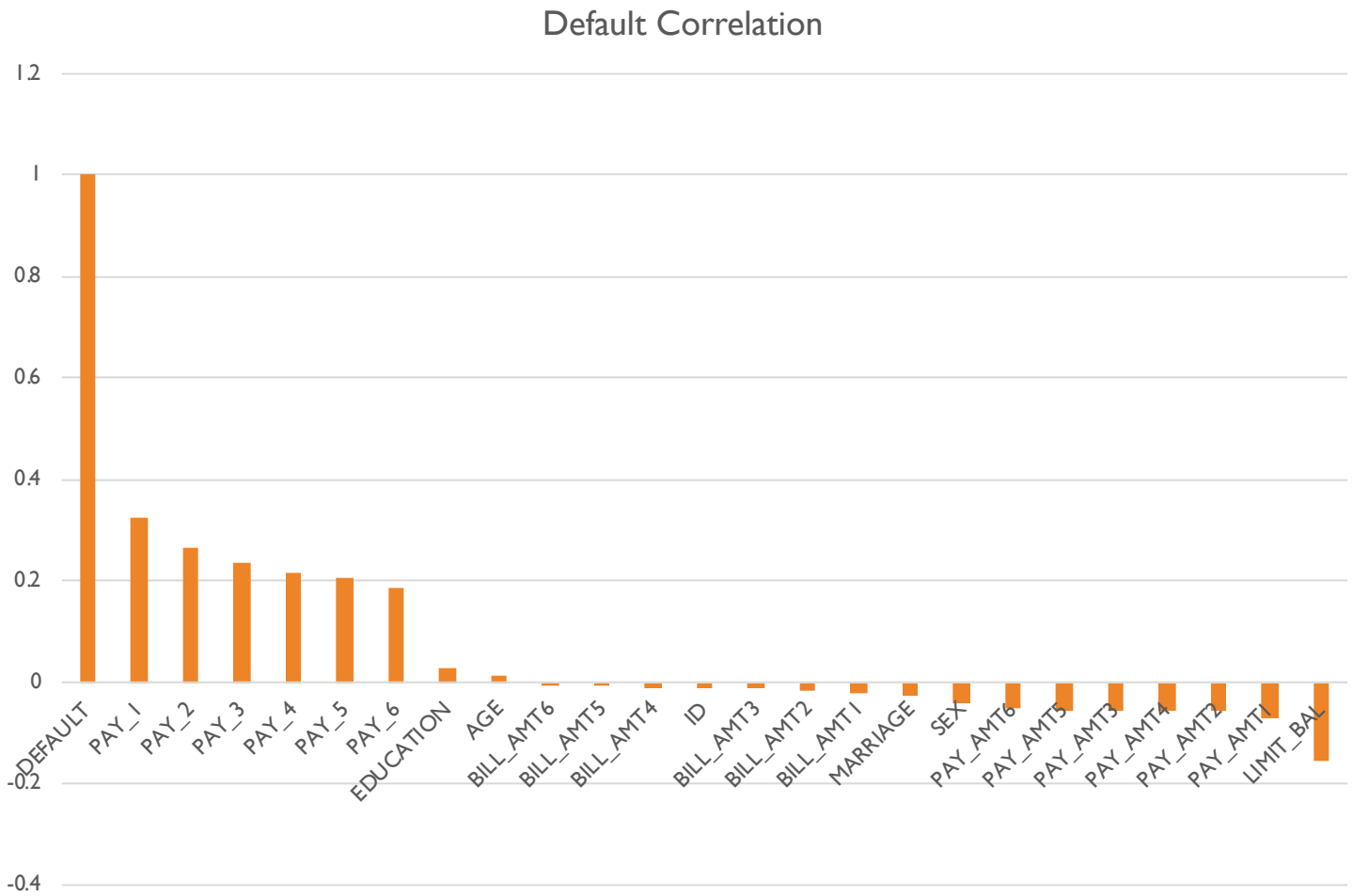
- Additional questions:
  - Why when the payment was paid in full, we have Default data under history of last payments?

DATA SCIENCE FLOWCHART VISUALIZING THE DETAILED PROCESS


Start
Business understanding
Define Goal
Wrangling, exploration and cleaning
Visualizing and find insights
Produce a preliminary report of data
List necessary questions or points of interest
Feature Engineering
Parameter Tuning
Model Training
Test the model
Performance measure
No
Does the model satisfy requirement
Yes
Unit Test and refactoring
Present results and document.
Feedback and reporting
Deploy Model
Monitor, analyze, manage
Was the goal achieve or change?
No
Yes
Maintain

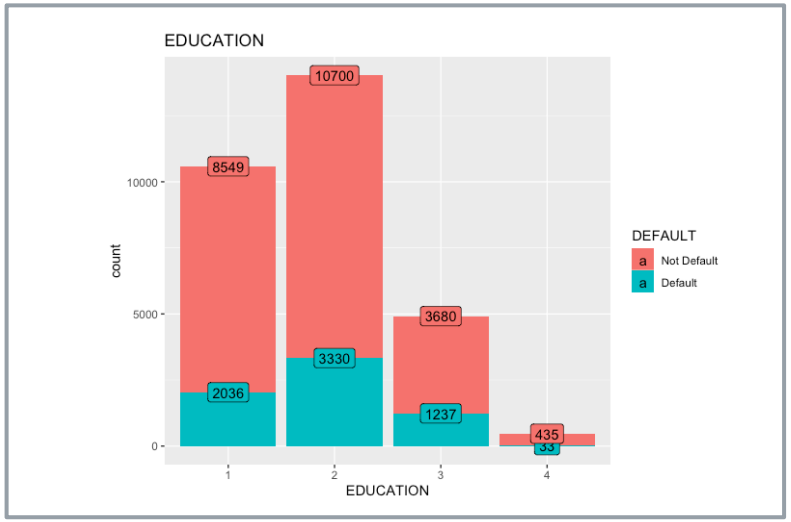


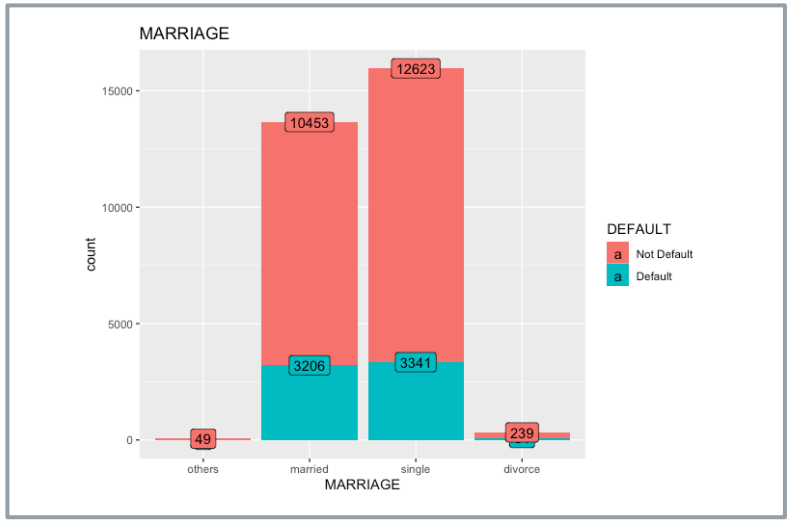
INITIAL INSIGHTS OF THE DATA



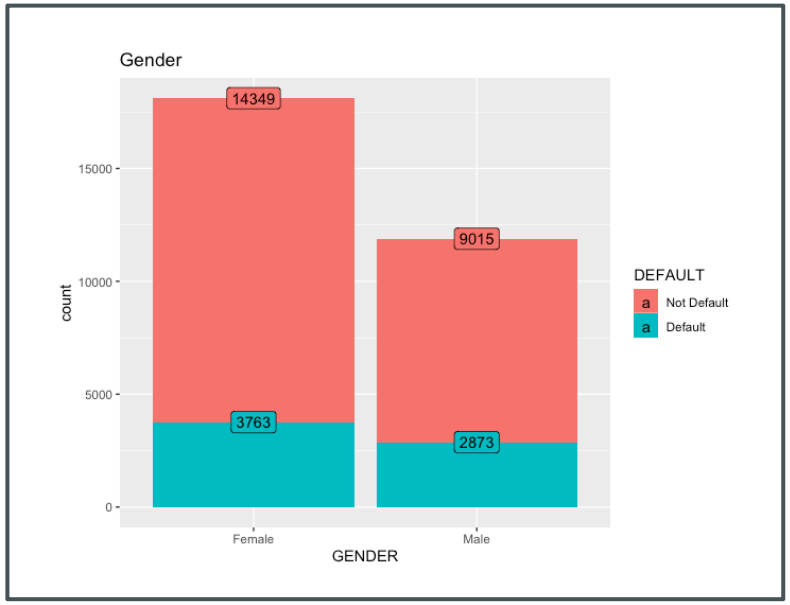
CORRELATION

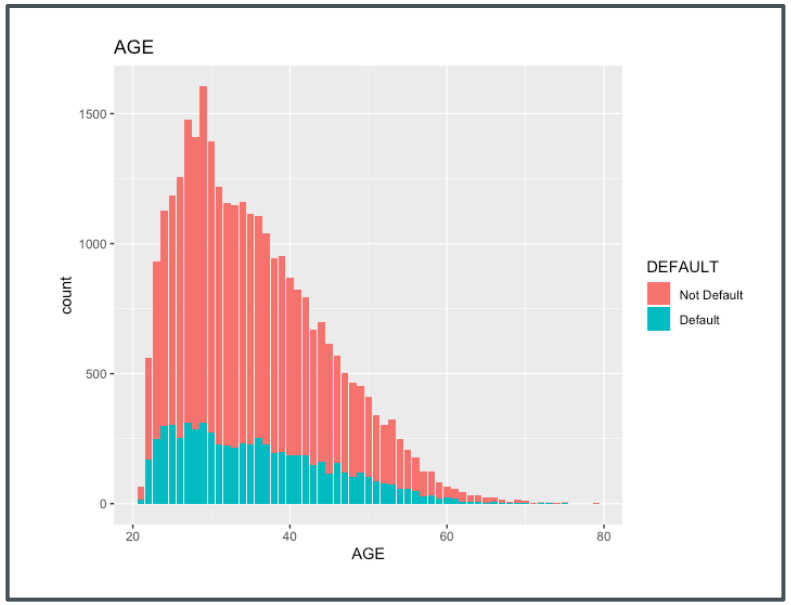
INSIGHTS OF THE DATA





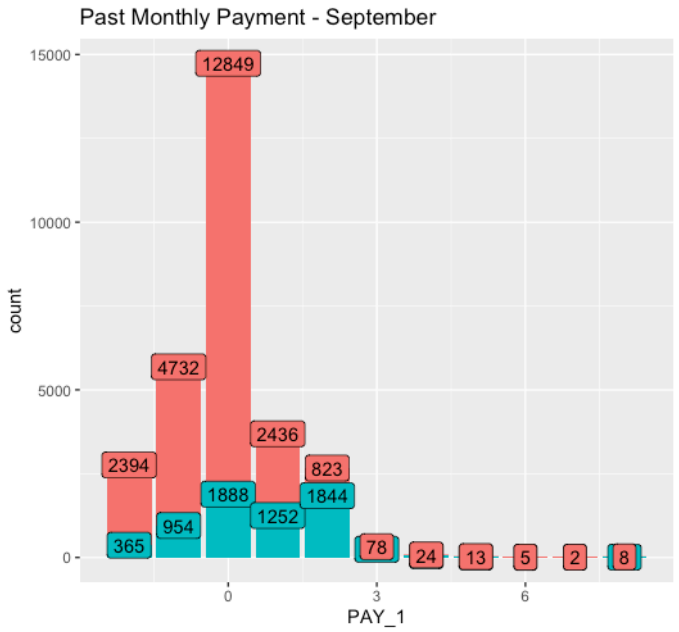
- Education
  - University group as the higher default count
- Marital Status
  - Marriage and single marital status has almost the same default count

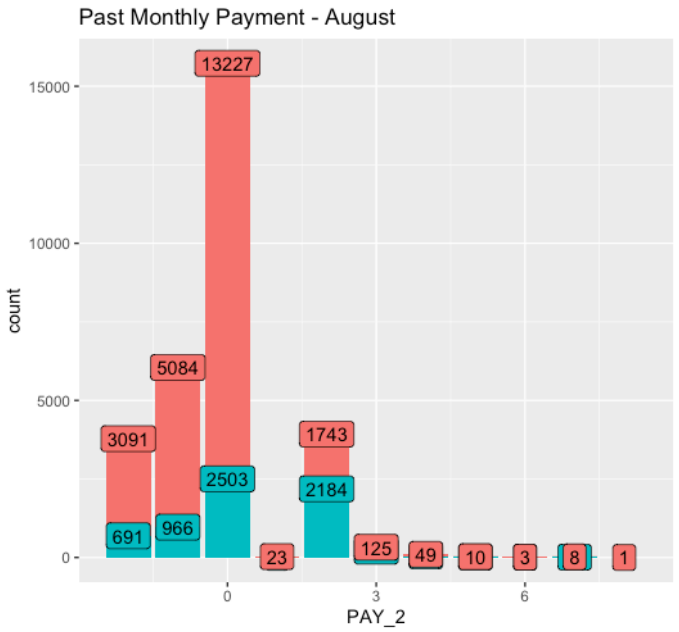


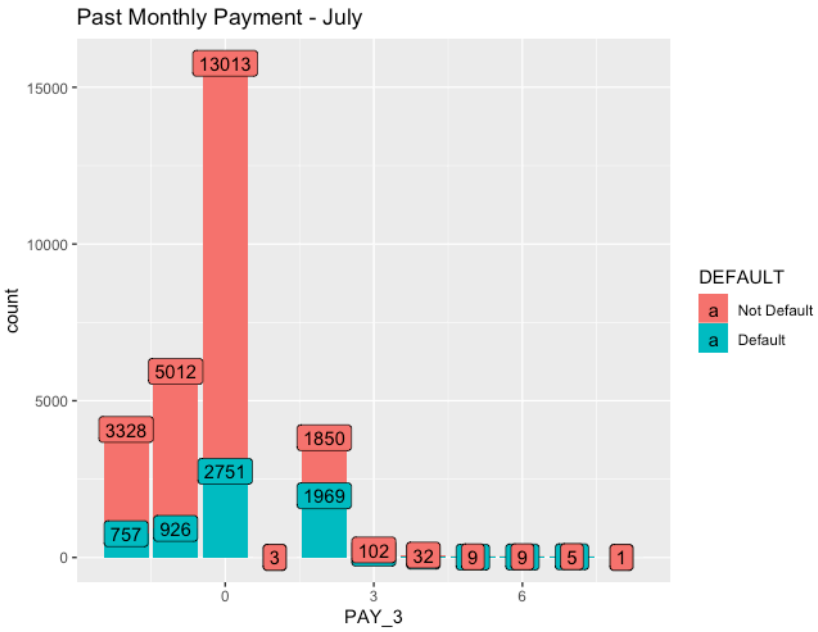


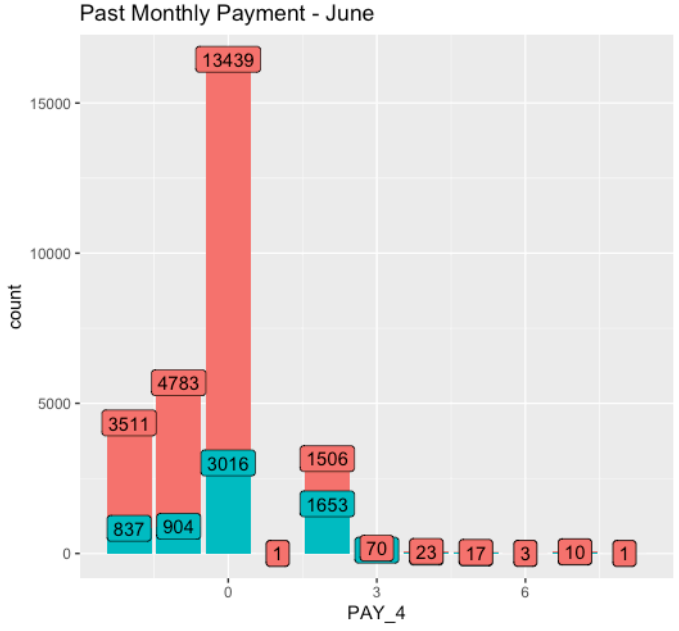
INSIGHTS OF THE DATA

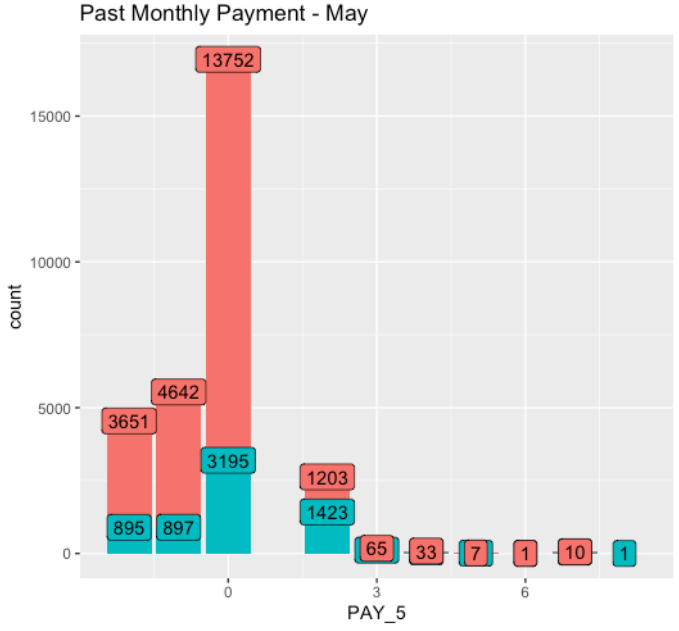
- Gender
  - Females has a higher default count
  - Females has a higher not default count
  - In general, there are more loans for females
- Age
  - Default count is higher for the group less than 30 years old.
  - Default count decreases as age increases

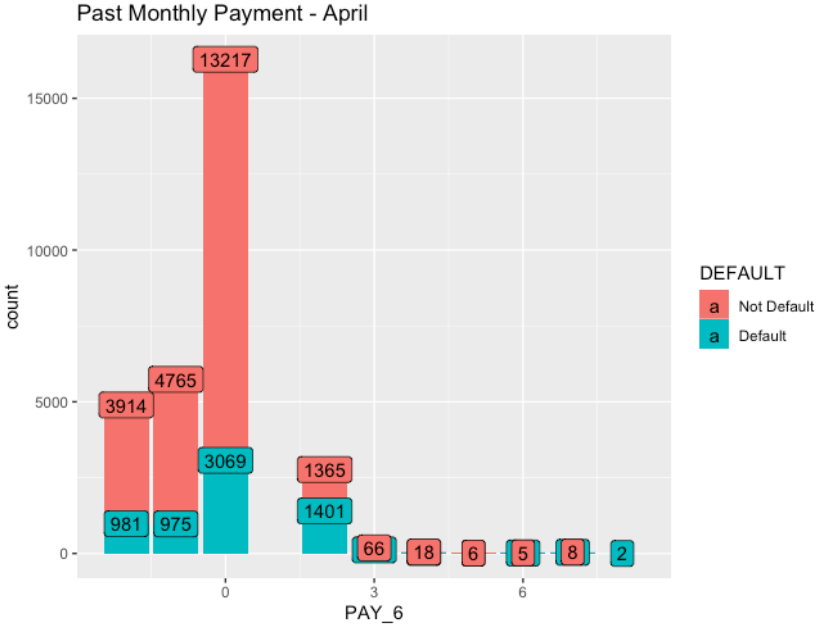

Past Monthly Payment - September
count
PAY_1
DEFAULT
a Not Default
a Default
12849
4732
2436
2394
823
1888
1844
1252
954
365
78
24
13
5
2
8


Past Monthly Payment - August
count
PAY_2
DEFAULT
a Not Default
a Default
13227
5084
3091
2503
1743
2184
691
966
23
125
49
10
3
8
1


Past Monthly Payment - July
count
PAY_3
DEFAULT
a Not Default
a Default
13013
5012
3328
2751
1850
1969
757
926
3
102
32
9
9
5
1


Past Monthly Payment - June
count
PAY_4
DEFAULT
a Not Default
a Default
13439
4783
3511
3016
1506
1653
837
904
1
70
23
17
3
10
1


Past Monthly Payment - May
count
PAY_5
DEFAULT
a Not Default
a Default
13752
4642
3651
3195
1203
1423
895
897
65
33
7
1
10
1


Past Monthly Payment - April
count
PAY_6
DEFAULT
a Not Default
a Default
13217
4765
3914
3069
1365
1401
981
975
66
18
6
5
8
2

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