Phase 2 of the CRISP DM is Data Understanding Phase. It starts with initial data collection and proceeds with activities to get familiar with the data , identify data quality problems , to discover insights into the data ,and to detect interesting subsets to form hypothesis for hidden information. If We are acquiring data from multiple resources integration of Data is an additional issue .It comprises of following 4 tasks.

a.**Task 1->Collect Initial Data:** We acquire data listed in Project resources and then We load data to enhance our understanding data and this might lead to initial preparation steps.

Out put of the task:

Initial Data Collection report.(Data downloaded from Edwisor website in the form of train.csv and test.csv )

b.**Task.2->Describe Data:** Here We describe Data -examine the gross or surface properties of the acquired data and report on the results.

Output:

Data Description report

Train.csv consists of 200000 records with 202 variables( 1st variable is customer Id\_code, 2nd variable is Target (values either 0 or 1) and remaining all the variables are explanatory variables with which We need to make prediction)

Test.csv consists of 200000 records with 201 variables of them (1st variable is customer Id\_code and remaining all variables are explanatory variable . Using Train data We need to predict target in test.csv)

c.**Task.3->Explore Data:**  In this task We do Data Exploration

Output: Data Exploration report.

Check R file and ipynb file for Data Exploration

We have summary of data , type of all the explanatory variables which double in R or float in Python.

d.**Task.4->Verify Data:** In this phase We Verify Data Quality aka check for any missing values and is the data enough for our analysis.

Output: Data Quality report

There were no missing values in train.csv and test.csv and the Quality of data provided is good .