As I have selected loan default capstone project the problem statement for the project is below

Financial institutions incur significant losses due to the default of vehicle loans. This has led to the tightening up of vehicle loan underwriting and increased vehicle loan rejection rates.  
The need for a better credit risk scoring model is also raised by these institutions. This warrants a study to estimate the determinants of vehicle loan default.

There is 1 dataset data that have 41 attributes.  
You are required to determine and examine factors that affected the ratio of vehicle loan defaulters. Also, use the findings to create a model to predict the potential defaulters.

For the above problem the analysis I did is as below

* First I have loaded the dataset with the help of pandas in Jupiter notebook and analyse the datatypes
* Next in order to find the missing values uses numpy na and seen there is any duplicate values
* The variable name in data is not according with identifier so changed according ,and analyse data size, data describe
* The target variable is loan default with the count plot analysed how it is distributed
* Next with help of box plot analysed how target variable is distributed across branchId,manufacturedId,supplierId,stateID and employment type
* In order find age variable subtracted date of birth and disbursal date and analyse how it distributed with target variable
* With the help of count values find out which prof the customer has submitted the most for loan
* Analysed credit bureau score distribution with loan default by count plot
* Explore the primary and secondary account details by T test
* Analysed difference between the sanctioned and disbursed amount of primary and secondary loans and correlation of both primary and secondary loan
* There are 1344 customer who make more enquiries are high risk candidates
* From the correlation heatmap, Primary and secondary accounts,credit history, that is new loans in last six months, loans defaulted in last six months, time since first loan, are not a significant factor in estimating probability of loan defaulters
* For employment type used dummy variable
* For analysed date used logistic regression from sk learn model and divided varible to x and y
* Next split the variable in the ratio of 70 train data and 30 test data
* By logistic regression trained the data find the intercept and coefficient
* At last analyse ypred by predicating XTest data and analysed classification report with Accuracy.