# Data Science Coding Challenge- Detailed Design and Implementation Document

# Part 1: Data Exploration and Evaluation

Create an exploratory data analysis project. Load the data and perform any necessary cleaning and aggregations to explore and better understand the dataset. Based on your exploration, please describe your high level findings in a few sentences. Please include two data visualizations and two summary statistics to support these findings.

**Steps:**

* Refer PRDXN\_code/Data Exploration folder for the code.
* Code: part1\_ipynb.ipynb (Jupyter Notebook) can refer to part1\_ipynb.py if jupyter is not installed, else read README.txt for instructions to install Jupyter.
* Approach:
  + Load the data from cv
  + Remove irrelevant fields – Which have no relation to the insights/visualizations we are going to derive in the following sections
  + Identify the columns of importance and get an idea of the data values
  + Standardize non-uniform data, convert to integer or binary values wherever possible
  + Get the concentration of nulls in the columns and discard the ones that have more than 80% of null values.
  + Load the cleansed dataframe into a csv (for tableau visualization).
* Visualization:
  + 1)Number of loan applications based on geography/state:

<https://public.tableau.com/profile/fauzia.shaikh#!/vizhome/Part1-PRDXNLCAssignment/Part1_1>

* + 2)Number of loan accounts based on employment length and status of the loans:

<https://public.tableau.com/profile/fauzia.shaikh#!/vizhome/Part1-PRDXNLCAssignment/Part1_2>

* + Tableau workbooks have not been uploaded in Github due to size constraints.

# Part 2: Engineering

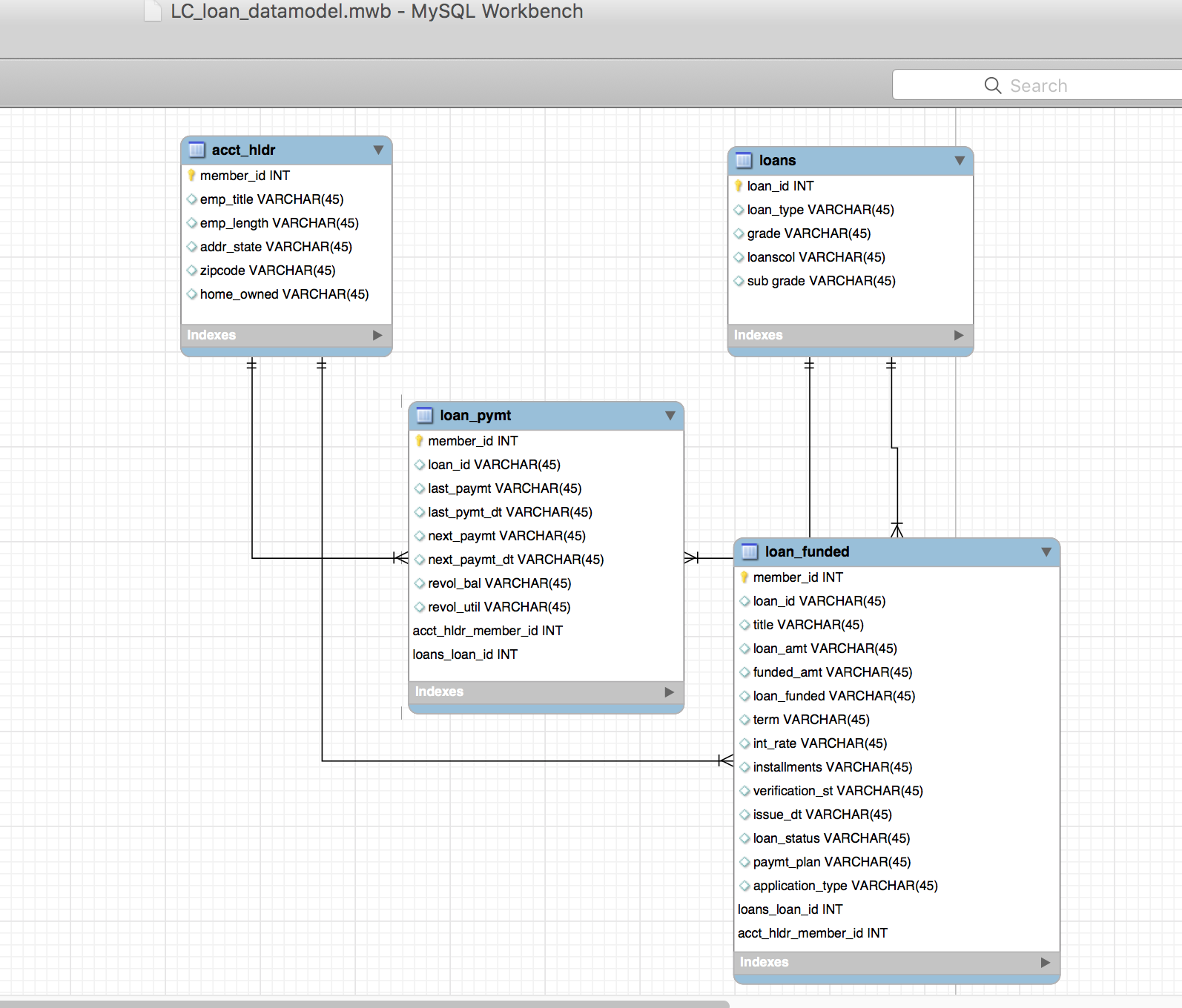
Please build a prototype of a production data pipeline that will feed an analysis system (data warehouse) based on this dataset. This system will allow data scientists and data analysts to interactively query and explore the data, and will also be used for machine learning model building and testing. You may drop fields that you consider are not important for your analysis based on your evaluation in Part1.

* Create a data model / schema in a database engine of your choice
* Develop code that will persist the dataset into this storage system
* Include any data validation routines that you think may be necessary

Prioritize simplicity in your data model and processing code. Explain your thought process and document any alternate data models you considered along the way.

**Steps:**

* Source:Assumption: The loan.csv file will land at a particular location at a particular time, based on some cadence.
* ETL:Python script to run based on the cadence of the above source file, to cleanse,standardize and aggregate data if necessary.
* Load the data in sqllite db (database.sqlite) based on the current model chosen.
* Current model: Single table (with relevant and standardized fields only) with the assumption based on the data available, that each member\_id will have only 1 loan account associated with him/her.Creating the table in sqlite for the downstream teams.
* Alternate model:
  + Separate tables containing account holder details(name,title,address,employement),loans per account holder(cars,home,personal),loan\_details(Different loans available with Lending club),loan\_payment details (per acct\_id per loan\_id,paymnt details) but there’s only 1 account per id so it is already normalized otherwise it would have had redundancy
  + Managing PK,Referential integrity,contraints,etc
  + Sample ER model:



(Note:Data types need to be updated by going through each field one by one and understand the possible values and range, have skipped it for now given the time constraint.)

* Actual ER model can be found at PRDXN\_code/Data Engineering/data model/PRDXN\_LC\_loan\_data\_model.mwb (Need to install mysql workbench to view it,installation for mac is included in the same directory)

Note: The script is quite scrappy and can be updated to have a modular approach ( different functions for different tasks alongwith exception handling, in an iterative way.

# Part 3: Business Analysis

# After getting the data in the warehouse, your business analysts are interested in getting answers to the following, please write SQL queries and share the resultant data.

# Assuming the loans with status that are ‘Current’, ‘Issued’ and ‘Fully Paid’ as “Good Loans”, what is the percentage of good loans across each the 36- and 60-month terms.

# Solution:

# Disclaimer (Would need confirmation for these values): I'm not including the loans with status as ‘Does not meet the credit policy…’ as good loans inpite of Status:Fully Paid/Charged Off.

# %Good loans across 36month term:92.65%

# %Good loans across the 60month term:90.72%

# Query:

# Select (c.total\_good\_36\*100)/a.total\_36 as pct\_good\_loan\_36, (d.total\_good\_60\*100)/b.total\_60 as pct\_good\_loan\_60

# FROM

# (Select count(\*) as total\_36 from loan where term='36') a,

# (Select count(\*) as total\_60 from loan where term='60') b,

# (Select count(\*) as total\_good\_36 from loan where loan\_status in ('Current','Issued','Fully Paid') and term='36') c,

# (Select count(\*) as total\_good\_60 from loan where loan\_status in ('Current','Issued','Fully Paid') and term='60') d

# Output File: Located in PRDXN\_code/SQLs/output/SQL1.csv

# What are the title(s) of employee(s) who took the most loans and least number of loans?

# Solution:Gives top 10 employee titles with the most and least number of loans

# Query:

# select \* from

# (Select emp\_title,count(\*) as ct from loan where emp\_title is NOT NULL group by emp\_title order by ct desc limit 10 ) a

# UNION ALL

# select \* from (Select emp\_title,count(\*) as ct from loan group by emp\_title order by ct asc limit 10 ) b

# Output File: Located in PRDXN\_code/SQLs/output/SQL2.csv

# What is the most common purpose of the loans that are considered “Bad Loans” (please use definition mentioned for “Good Loans” in #1 above).

# Solution:

# All possible values for loan\_status:

# >> df.loan\_status.unique()

# array(['Fully Paid', 'Charged Off', 'Current', 'Default',

# 'Late (31-120 days)', 'In Grace Period', 'Late (16-30 days)',

# 'Does not meet the credit policy. Status:Fully Paid',

# 'Does not meet the credit policy. Status:Charged Off', 'Issued'], dtype=object)

# As per #1, anything that is not ‘Current’,’Issued’, or ‘Fully Paid’ falls under the “Bad Loans” category.

# Disclaimer (Would need confirmation for these values): Including the loans with status as ‘Does not meet the credit policy…’ as bad loans as they were excluded in #1.

# Query:

# Select purpose from loan where loan\_status NOT IN ('Current','Issued','Fully Paid') group by purpose order by count(\*) desc limit 1

# Output File: Located in PRDXN\_code/SQLs/output/SQL3.csv

# Part 4: Modeling

# Your Product Manager wants you to suggest a good machine learning model to predict loan defaults so as to not invest in such loans. Please build models using different algorithms and explain how would you choose the one that is effective. Assume that (i) you are given the ability to invest in each loan independently; (ii) you invest immediately following loan origination and hold to maturity (36 months); and (iii) all loan fields that would be known upon origination are made available to you.

# <I’d require more time to research on this part as I’m not a data scientist(yet) and have very limited understanding about the ML concepts and algorithms.So will skip this one for now>

# Part 5: Visualization

# Please create a workbook in Tableau with the following dashboards –

# YoY Trend of loan amounts bucketed by tenure and state :

# <https://public.tableau.com/profile/fauzia.shaikh#!/vizhome/Part5-PRDXNLCAssignment/Sheet1>

# 2. Profitability\* analysis state-wise for each purpose :

# \*For simplicity calculate profitability as (total\_pymnt/funded\_amnt)^1/3 - 1

# <https://public.tableau.com/profile/fauzia.shaikh#!/vizhome/Part5-PRDXNLCAssignment/Sheet2>

# Loan amount trend forecast for each grade:

# <https://public.tableau.com/profile/fauzia.shaikh#!/vizhome/Part5-PRDXNLCAssignment/Sheet3>

# Tableau workbooks have not been uploaded on Github due to size constraints.Please download them from the above given links.

# Artifacts:

# Github link: <https://github.com/fauzia-s/PRDXN_LC_Assignment/tree/master/PRDXN_code>