**DATA CLEANING PROCESS**

**DATA CLEANING OBJECTIVES**: To identify data quality issues such as ***null cells****,* ***unstandardized formatting,*** ***unusual characteries*** and *inconsistencies* that could affect the data quality and integrity. To development ways to mitigate them, and ensure that the data is suitable and reliable for analysis. The tool used for this integral part of data analysis is **SQL.**

**ABOUT THE DATASET:** The financial loan dataset which consists of columns attributes relating to the loan collection parameters contains ***57 columns*** and ***466,285 rows*** was sourced from Kaggle. It contains the following columns:

|  |  |
| --- | --- |
| **Changed Column Name** | **Description** |
| [ID] | A unique LC assigned ID for the loan listing. |
| [Borrowers\_ID ] | A unique LC assigned Id for the borrower member. |
| [Loan\_Amount] | The requested loan amount by the borrower. |
| Funded\_Amount | The total amount committed to that loan. |
| [Funded\_Amnt\_Invested] | The total amount committed by investors for that loan |
| [Term] | The term of the loan usually expressed as the number of months |
| [Intrest\_Rate] | The interest rate on the loan |
| [Installment] | The monthly payment made by the borrower |
| [Grade] | The loan grade assigned by the lending institution based on the borrower's credit worthiness |
| [Sub\_Grade] | A more granular classification than the grade representing the risk level within each grade |
| [Employment\_Tittle] | The job title of the borrower |
| [emp\_length] | The length of employment of the borrower in years |
| [Home\_Ownership] | The type of home ownership status (e.g.own mortgage rent). |
| [Annual\_Income] | The annual income reported by the borrower. |
| [Verification\_Status] | Indicates whether the borrower's income was verified by the lending institution |
| [Issued\_Date] | The date when the loan was issued. |
| [Loan\_Status] | The current status of the loan (e.g. current fully paid charged off) |
| [Paymnt\_Plan] | Indicates if there is a payment plan for the loan |
| URL | The URL of the loan listing on the lending platform. |
| [Description] | A description provided by the borrower for the purpose of the loan. |
| [Purpose] | The purpose of the loan (e.g. debt consolidation home improvement). |
| [Loan\_Title] | The loan title or name provided by the borrower. |
| Zip\_code | The first three digits of the borrower's zip code. |
| [Address\_State] | The state of the borrower's address. |
| [Debt\_to\_Income\_Ratio] | Debt-to-Income ratio representing the borrower's total monthly debt payments divided by their gross monthly income |
| [Delinquency\_2yrs] | The number of 30+ days delinquencies in the borrower's credit file for the past 2years |
| [Earliest\_Credit\_Line] | The date the borrower's earliest reported credit line was opene |
| [Inquiries\_Last\_6months] | The number of inquiries by creditors in the last 6 months. |
| [months\_Since\_Last\_Delinq] | The number of months since the borrower's last delinquenc |
| [Months\_Since\_Last\_Record] | The number of months since the last public record. |
| [Open\_Account] | The number of open credit lines in the borrower's credit file. |
| [Public\_Record] | The number of derogatory public records. |
| [Revolving\_Bal] | Total credit revolving balance (i.e. total credit card balances). |
| [Revolving\_Ultilization] | Revolving line utilization rate or the amount of credit the borrower is using relative to their total credit limit |
| [Total\_Account] | The total number of credit lines currently in the borrower's credit files |
| [Initial\_List\_Status] | The initial listing status of the loan. |
| [Outstanding\_Principal] | Remaining outstanding principal for total amount funded. |
| [Outstanding\_Prncp\_Invested] | Remaining outstanding principal for portion of total amount funded investors |
| [Total\_Payment] | Payments received to date for total amount funded. |
| [Total\_Payment\_Invested] | Payments received to date for portion of total amount funded by investors |
| [Total\_Recved\_Prncp] | Total principal received to date. |
| [Total\_Recved\_Intrest] | Total interest received to date. |
| [Total\_Recived\_Late\_Fee] | Total late fees received to date. |
| [Recoveries] | Post charge-off gross recovery. |
| [Collection\_Recovery\_Fee] | Post charge-off collection fee. |
| [Last\_Pymnt\_Date] | Last month payment was received. |
| [Last\_Pymnt\_Amnt] | Last total payment amount received. |
| [Next\_Pymnt\_Date] | Next scheduled payment date. |
| [last\_Credit\_Pulled\_Date] | The most recent month LC pulled credit for this loan. |
| [Collections\_12\_Months\_Ex\_Medicals] | Number of collections in the last 12 months excluding medical collections |
| [Months\_since\_last\_Major\_Derogatory] | Months since most recent 90-day or worse rating. |
| [Policy\_Code] | publicly available policy\_code=1. |
| [Application\_Type] | Indicates whether the loan is an individual application or ajoint application with two co-borrowers |
| [Account\_now\_Delinq] | The number of accounts on which the borrower is now delinquent. |
| [Total\_Collection\_Amnt] | Total collection amounts ever owed. |
| [Total\_Current\_Bal] | Total current balance of all accounts |
| [Total\_Revolving\_High\_Credit\_limit] | Total revolving high credit/credit limit. |

**DATA QUALITY ASSESSMENT:** Thorough assessment was carried out and the following data quality issues was identified:

* **Null values:** Empty cells were observed in 18 columns of the datasets, including the following;
* earliest\_cr\_line
* delinq\_2yrs
* inq\_last\_6mths
* mths\_since\_last\_delinq
* mths\_since\_last\_record
* open\_acc
* pub\_rec
* revol\_bal
* revol\_util
* total\_acc
* last\_pymnt\_d
* next\_pymnt\_d
* last\_credit\_pull\_d
* collections\_12\_mths\_ex\_med
* mths\_since\_last\_major\_derog
* acc\_now\_delinq
* tot\_coll\_amt
* tot\_cur\_bal
* total\_rev\_hi\_lim
* **Unstandardized date format:** The following date columns were identified to be improperly formatted:
* issue\_d
* earliest\_cr\_line
* last\_pymnt\_d
* next\_pymnt\_d
* **Incomplete** **values**: The ***“Zip code”*** column contains incomplete values, having a mixture of numerical and text values which isn’t useful for the analysis.

* **Incorrect data types**: The numerical columns and the date columns which are ***36*** and ***4*** respectively, were observed to contain incorrect data types which can affect the reliability of the data for analysis.
* **Unwanted columns**: 2 columns including the ***“Zip code”*** column and the ***“URL”*** column were identified to be of no importance to the analysis.

**DATA IMPORTATION:** The dataset was confirmed to be in a ***CSV format*** before importation to ***SQL tool.*** This began by creating a schema named “***risk\_analysis\_project***” in SQL to house the dataset before the importation process was carried out

**DATA CLEANING AND PREPARATION:** Upon identifying the data quality issues as well as ways to mitigate them, the following steps were in SQL to clean the dataset, making it suitable for the analysis