

1. Analyze all input features.

Loan ID, Customer, Date, Gender, Married, Dependents, Age, Application Type, Education, Self Employed, Applicant Income, Co Applicant Income, Loan Amount, Loan Amount Term, Credit History, Property Area, Application Number, Amount Requested, Loan Status

2. Create a categorical and continuous data quality report highlighting statistical metrics including count, min, max, mean, mode, median, cardinality etc.

Categorical:

- Loan ID
- Customer
- Date
- Gender
- Married
- Application Type
- Education
- Self Employed
- Property Area
- Application Number
- Loan Status

Continuous:

- Dependents
 - Min:0
 - Max:100
 - Mean:1.1
 - Mode:0
 - Median:0
 - Cardinality:
- Age
 - Min: -2
 - Max: 180
 - Mean: 37.4
 - Mode: 55
 - Median: 41
 - Cardinality:
- Applicant Income
 - Min: 150
 - Max: 81000
 - Mean: 5403.5
 - Mode: 2500
 - Median: 3812.5
 - Cardinality:

- Co Applicant Income
 - Min: 0
 - Max: 41667
 - Mean: 1621.2
 - Mode: 0
 - Median: 1188.5
 - Cardinality:
- Loan Amount
 - Min: 9
 - Max: 700
 - Mean: 146.4
 - Mode: 120
 - Median: 128
 - Cardinality:
- Loan Amount Term
 - Min: 12
 - Max: 480
 - Mean: 342
 - Mode: 360
 - Median: 360
 - Cardinality:
- Credit History
 - Min: 0
 - Max: 1
 - Mean: 0.8
 - Mode: 1
 - Median: 1
 - Cardinality:
- Amount Requested
 - Min: 9
 - Max: 700
 - Mean: 146.4
 - Mode: 120
 - Median: 128
 - Cardinality:

Identify outliers for continuous features.

- Dependents: 2,4
- Age:
- Applicant Income: 16667, 16525, 19730, 14683, 39147, 11417, 12500,
- Co Applicant Income: 33837, 6666, 20000, 7873, 6250, 7210
- Loan Amount: 324, 480, 311, 296, 570, 308,
- Loan amount term
- Credit history:
- Amount requested: 324, 480, 311, 296, 380, 570, 304, 279, 320, 315

Apply feature elimination techniques like missing values, constants, duplicates etc. and identify the list of input features which can be eliminated/dropped, and explain why each feature was dropped.

- Application type can be eliminated because the only answer is loan

Last step will be to upload your final dataset (cleaned) with the list of input and output features after eliminating bad quality features. You can also consider eliminating rows or observations which does not make sense to you. Please add your comments why rows were dropped (if any). Please use excel functions to find statistical metrics like mean, median, count, min, max etc. Please use Git to check-in the final files. I'm looking for 2 files 1) Final dataset, file naming convention - Loan_'First Name'.csv (Loan_Noble.csv) 2) Summary with explanations, file naming convention - 'Summary_First Name'.txt (Summary_Noble.tx, feel free to use ppt or xls or pdf if that's easy)

Please refer to the session recordings and slides first and if you still have questions, please reach out to me.