Analyze all input features.

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

Identify outliers for continuous features.

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.

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 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.

1)Analyze all input features

Loan_ID, Customer, Date, Gender, Married, Dependants, Age, Application_Type, Education, Self_Employed, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, Credit_History, Property_Area, Application_Number, Amount_Requested, Loan_Status

2 -3) Data Report + Outliers

Continuous Features

Age

- Mean: 37
- Median: 41
- Count: 513
- Min: -2
- Max: 180

ApplicantIncome

- Mean: 5403
- Median: 3813
- Count: 614

- Min: 150
- Max: 81000

CoapplicantIncome

- Mean: 1621
- Median: 1189
- Count: 614
- Min: 0
- Max: 41667

LoanAmount

- Mean: 146
- Median: 128
- Count: 592
- Min: 9
- Max: 700

Loan_Amount_Term

- Mean: 342
- Median: 360
- Count: 600
- Min: 12
- Max: 480

Amount_Requested

- Mean: 146
- Median: 128
- Count: 592
- Min: 9
- Max: 700

Categorical Features

LoanID
Customer
Date
Gender
Married
Dependants

Application_Type
Education
Self-Employed
Credit_History
Property_Area
Application_Number
Loan_Status

4)
Missing values: I noticed a lot of mising values in this excel sheet and removed them by highlighting everything, pressing Ctrl + G, pressing Special, pressing blanks and going to the highlighted value right clicking and then deleting

Constants: I didn't remove constants because every value is a constant

Duplicate values: I didn't remove duplicate values because a lot of the dataset has duplicate values but that doesn't mean it should be removed

6)
I removed Application_Type because every application type was loan with no other values so it didn't make sense to keep in

I was unable to figure out how to do cardinality and outliers