5 steps included in this tutorial

1. Obtain the data:
2. The data has been obtained from Kaggle, and reference the link to it:

<https://www.kaggle.com/burak3ergun/loan-data-set/code>

1. Take a look at the data, how many rows and columns?614\*13 What are the columns and what do they represent?

Loan\_ID: SE

Gender: gender of the applicant

Married: yes/no

Dependents: # of dependents

Education: level of education

Self\_Employed: yes/no

Property\_Area: SE

Purpose: The purpose of the loan

loan amount: SE

annual income: SE

dti: The debt-to-income ratio of the borrower (amount of debt divided by annual income)

fico credit score: SE

Loan\_Status: Yes (approved)/No (not approved)

1. Scrub the data:
2. Create a duplicate row; row with loan\_id: LP001003 is duplicated
3. Create Gender/Married in the same row and then split it.
4. Create a clearly wrong data point: credit score for applicant with loan\_id: LP001020 is wrong, extra trailing zero.
5. Create a missing data point(s): credit score for applicant with loan\_id: LP001008 is missing
6. Create a datapoint in the wrong format: loan amount for applicant with loan\_id: LP001020 is incorrectly formatted.
7. Explore the data(EDA):
8. What are the datatypes of the columns? We need to know because we will treat them differently.

Loan\_ID: categorical

Gender: categorical

Married: categorical

Dependents: categorical

Education: categorical

Self\_Employed: categorical

Property\_Area: categorical

Purpose: categorical

loan amount: categorical

annual income: categorical

dti: numerical-float

fico credit score: numerical

Loan\_Status: categorical

1. What are the various patterns: explore the descriptive stats and the distributions of some of the columns.

See EDA word document

1. What is the relationship between the different variables? Through a correlation (it's how two variables move in relation to one another) matrix.

See EDA word document

1. Model data
2. Create a nominal logistic regression model to classify applicants into eligible/not eligible for the requested loan. See model results in model data word document.
3. Divide previous applicants into k clusters to target each with a customized mail marketing campaign. See model results in model data word document.
4. Interpret data

In this step, try to draw a conclusion from the classification model as to which variables contributed the most to the class outcome. Moreover, try to characterize the clusters generated by K-means to understand the way they should be dealt with commercially.

Once done, go through all the steps and explain what additions or differences might occur at each step and how to handle it.