Project: Capstone Project 1 - Data Wrangling

This document present findings from some data wrangling and EDA performed on a dataset of customers who have paid off their loans or not.

The dataset is available as a csv file here: <https://www.kaggle.com/zhijinzhai/loandata>

To do the data wrangling steps and EDA the dataset is imported into a Pandas DataFrame.

The findings are:

* The dataset as 500 rows in total.
* By removing all the NaN values from all rows we get a total of 100 rows that have no missing data.
* The NaN values are present only in ‘paid\_off\_time’ and ‘past\_due\_date’ columns.
* Some of the analysis that can be performed are correlation between factors like Gender, Education, Age, Loan Amount (Principal), Duration (term) and Loan\_Status (whether locan was paid off or not)
* It is an observation in the dataset that the males have better paid-off to not-paid-off ratio of total no. of loans males have taken, compared to same metrics in females.
* Higher Education as a positive correlation to loan being paid-off, especially in males.

The code for above analysis is in the ‘data\_wrangling\_capstone.ipynb’ file.