# **Data Wrangling Steps – Capstone Project 1**

## **Data Source**

* The data set was obtained from Kaggle website.

## **Cleaning**

* The data was fairly clean as it was coming from a reliable source.
* I first tried to understand the data in the file:
  + Using head() function I tried to see first 5 rows of my data. By using pandas set\_option(‘max\_columns’, None) function I made sure I see all columns that are available in the dataset.
  + Using info() function I learned that I have 74 columns and 887379 entries in my dataset.
  + I decided to not drop any columns as there aren’t too many in the file to begin with even though a few columns like ‘sub-grade’ and ‘emp\_title’ seemed less important.
  + The describe() function helped in understanding the distribution of each numerical column, their min, max and mean values, the standard deviation etc.
* Loans had around 10 different statuses which I bucketed into 5 namely, ‘Fully Paid’, ‘Default’, ‘Current’,’ Late’ and finally the last catch-all bucket named ‘Unknown’ to catch any exception which might arise when the same code is applied to new records in the company. The column does not have any null values currently but in future it can have if the dataset size is increased and thus the last category will help us catch those records.
* When understanding borrowers, a column which recorded employment length of each borrower had around 12 unique values which I bucketed into 5 categories namely ‘10+ years’, ‘1 to 5 years’, ‘6 to 9 years’, ‘Less than 1 year’ and finally the last catch-all bucket named ‘n/a’ where we do not have the data.
* The employment length column also had null values and thus I made sure I knew how many null values exist and bucket them appropriately when cleaning up employment\_length column.
* There were four records for which annual income was not available, and were filled using median of annual income filled these.
* I tried to check top 10 states for fully paid and default loans.
* I drew a regression line for annual income and loan amount for the fully paid and default loans to see if they have any correlation.