Kickstarter Project: Data Cleaning Steps

1.1 General Data Cleaning Steps:

- Convert launched_date and deadline_date into month and year
- Convert state (whether project failed or succeeded) into a numeric variable with failed as zero and successful as 1
- Calculate the difference between start launch_date and deadline_date
- Create new variable goal_per_day or divide the usd_goal by number of days between the deadline_date and launch_date

1.2 Imputation using KNN

- For the countries variable, around 3K observations are missing
    - Replace 'N,0' variable with 'nan'
    - Convert all the countries into label encoding (replace category name with number; this is the only way KNN will process the data)
    - Apply KNN algorithm

1.3 Part of Speech Tagging

- In order to extract information about part of each of each word, we would need to use nltk's part of speech tagger (pos_tag)
- Extract the part of speech tag from the output of pos_tag
- Then , apply set to get only unique post_tags and join them
- Remove any extraneous symbols (:,#)
- Create dummies for each unique post_tag