**Kickstarter Project Data Wrangling**

In this project, I use web robot’s kickstarter dataset with json format. After loading json into pandas dataframe, I noticed all the useful data was in ‘data’ column as dictionaries for each row, and other columns was only scraper robot information. After loaded ‘data’ columns’ each dictionaries into a new dataframe, the dataset have 205227 rows and 38 columns.

I used ipython notebook to inspect each columns one by one, some columns’ information is very clear and straightforward, like ‘backers\_count’, ‘country’, ‘id’, ‘pledged’, ‘currency’, ‘name’ etc. Those columns don’t need any additional wrangling and are good to use as loaded.

Some columns include heavy text for description or urls links to website, API or photos. As in this project, I don’t plan on working with Nature language processing or image processing, I simply ignore this columns for now. After all data cleaning and wrangling was completed, I saved two versions of csv, ‘cleaned\_full\_dataset’ and ‘cleaned\_useful\_dataset’. In the full version, I kept all the original columns together with some generated columns in case I’d like to analyze this information in the future. In the useful version, I saved only the columns I planned on using for this capstone project to save space. Full dataset is 1.09 GB while the useful version is only 65.4 MB, which make loading data and analyzing much easier and convenient, while I still have additional information to load if I find them useful later on.

There’s four columns named ‘created\_at’, ‘launched\_at’, ‘deadline’ and ‘state\_changed\_at’, which are all time information. However the dataset use Linux timestamp to store those information, which is very standard but not easy to interpret and analyze. So in order to convert those columns to more readable standard datetime, I define a function convert\_time, which take dataframe and column as parameters, and convert Linux timestamp to datetime64[ns] type.

For currency related columns, there are ‘currency’, ‘currency\_symbol’, ‘current\_currency’, ‘fx\_rate’, ‘static\_usd\_rate’, ‘pledged’, ‘goal’, ‘usd\_pledged’. After checking those columns, I get the information that ‘pledged’ and ‘goal’ are both using ‘currency’ as unit, while ‘static\_usd\_rate’ is the exchange rate between ‘currency’ and US dollar. If currency column is USD, this rate will be 1.0, and ‘usd\_pledged’ column is simply ‘pledged’ times ‘static\_usd\_rate’. While ‘current\_currency’ for most of the time is USD, but sometimes is other currency, and for some cases ‘currency’ is USD while ‘current currency’ is some other currency. I didn’t figure out what information ‘fx\_rate’ and ‘current\_currency’ is providing, so I put them to the drop list, still saved in the full version of the cleaned dataset, but are excluded in the useful version. For this project, I think ‘pledged’, ‘goal’, ‘currency’, ‘usd\_pledged’ and ‘static\_usd\_rate’ can provide enough information for my analysis.

There are a few columns are a series of dictionaries, like location, category and creator. I either iterate the dataframe or use list comprehension to retract useful information from those columns, and store them into a new column. From ‘category’, I used list comprehension to retract main category and sub category, and store as ‘main\_cat’, ‘sub\_cat’ columns in the dataframe. From ‘location’, I retract location name and location state. For some entries those two columns are the same, but for most cases, name is city name, while state is state name. There are some rows have location as NaN, so simply use list comprehension will provide list with wrong index, so I iterate dataframe to retract those two contents. From ‘creator’, I retract creator ID and creator name.

There are some other columns which content seems related but I can’t figure out a clear meaning, like ‘friends’, ‘is\_backing’, ‘is\_starred’, ‘permissions’. Most rows have those columns empty, but for the same 108 rows there’s some input. As I couldn’t get the reason for that, I decided to drop them from useful version of my cleaned dataset.

After cleaning all columns one by one, I picked the columns I want to use and saved as ‘cleaned\_useful\_dataset.csv’. In this cleaned dataset, there are 205227 rows and 27 columns. All rows are objects or numbers, no more dictionaries in the dataset. Only location name, location state and creator-name have NaN cells. Creator\_name columns have 205226 non-null objects, so only one cell is missing value, and the related creator\_id only have one project as well, so I couldn’t fill this information. For the missing locations, there are 204977 location state and 205044 location names. I will drop those entries for location based analysis, but can still use them for other analysis unrelated to locations. Data wrangling code can be viewed in data\_wrangling.ipynb in capstone project folder.