*\*This document is included in the final capstone. This document just contains data wrangling information*

Pundir, Saurabh

saurabhpundir.data@gmail.com

Table of Contents

[Data Wrangling 1](#_Toc490295596)

[I: Getting File 1](#_Toc490295597)

[II: File Content: 1](#_Toc490295598)

[III: Choosing Extraction Interval: 1](#_Toc490295599)

[IV: Reading Data: 1](#_Toc490295600)

[V: Cleaning Data: 2](#_Toc490295601)

[V: Storing Data: 2](#_Toc490295602)

[V: Merging Datasets: 3](#_Toc490295603)

[I. Data Dictionary: 3](#_Toc490295604)

# **Data Wrangling**

## **I: Getting File**

Data is available as CSV file on webrobots.com. The data is available based on a monthly basis as a zip file. Each month file size is 140 MB.

## **II: File Content:**

Each months zip file contains nearly 40 CSV files of approximately 19 MB size each. Each CSV file contains around 4000 project information. There are 32 distinct fields for each project. The fields detail mentioned in data dictionary section.

## **III: Choosing Processing Interval:**

The approach is to start from latest month file available(July-2017) and keep adding previous months data. Reading and consolidation of multiple month data showed that there are not many new projects added progressively in each month files. The latest file contains most projects from previous months. So data from Jan 2017 till current (July-2017) is considered for processing.

**IV. 3 Step Approach**

Data wrangling performed using three stages process.

1. Extracting data for each month.
2. Data Cleaning and consolidating data.
3. Storing data for further processing.

## **IV: Extracting Data:**

There are multiple files in each month’s folder. For load sharing purposes, each month folder processed separately. Reading files inside the folder is performed using “*os.path”* module. Each data extracted from individual file gets appended to a data frame for a month.

For an initial understanding of data the information on rows, columns, and data type of the data frame are explored. Each month has 170K distinct projects. After initial analysis, the fields which need further extraction are identified. In this stage, the following fields are identified for further consideration.

1. Id: This is unique project id
2. Location: JSON format. It required further extraction.
3. Creator: JSON format. It needed further extraction.
4. Category: JSON format. It needed further extraction.
5. URL: JSON format. It required further extraction.
6. Deadline: Unix date format
7. Launched at: Unix date format
8. State change at: Unix date format

*Note: This doesn’t mean other fields are insignificant or not processed.*

The code related to this part is in Capstone\_DataWrangling.ipynb file.

## **V: Cleaning Data:**

After reading monthly data file into a data frame. The fields mentioned above were extracted further by looping each data frame row. The JSON data is further extracted using JSON normalize. The date fields are converted using date time conversion from Unix time to UTC date (“%Y-%m-%d %H:%M:%S”) format.

The new temporary data frame is built in process for above extracted and converted data. This temporary data frame is merged using join and project id as a common key into main data frame after all sub fields value is extracted.

The list of sub fields extracted from JSON columns

* **creator**
  + name
  + is\_registered
  + user id
* **location**
  + city
  + state
  + type (district, town, city)
* **Category**
  + Project category type

**Missing Data 1(Location):**

Location information is not available for all rows in the data frame. So, this was handled comparing location with the null & empty value before extracting. Every data frame around 2K records will not have any location value.

**Missing Data 1(user info):**

The field name is\_registered under user info is not available from and before May 2017. This field is not relevant and hence filled with empty if not available.

Code related to this part is in Capstone\_DataWrangling.ipynb file.

## **V: Storing Data:**

**Challenge:**

The challenge in above process of cleaning and extracting data for each month in the data frame is the processing time. Processing each month files take 8-10 hours on medium configuration machine available. Processing 6 months data took almost three days to complete.

After cleaning data and creating final data frame for the monthly file. The file is stored in the hard disk to avoid re-running the time-consuming process. The pickle object sterilization is utilized to store each month data frame. After all monthly files are available, as individual files, a single pickle file is created with all data appended.

This file is unpickled and processed in further steps. The process of unpickling and making file available is under 5 minutes.

This approach provided the benefit of processing the data once and then utilizing it further at a faster pace.

**Further Work:**

The Hadoop and another related map -reduce technology can be utilized to make this process better. But currently, this was not the scope of the project.

The code related to this part is in Capstone\_DataWrangling.ipynb file.

## **V: Merging Datasets:**

The pickle files created for each month mentioned in above process are utilized to read and create a final data frame. The two data frames are created from this data.

**All unique projects:**

Considering the latest months data frame (July) as the base data frame, the records from all previous months are appended to create a data frame considering **unique** projects. At the end of the process, we have unique **216828** projects available. This data frame will be utilized further stages of the project.

**Historical information:**

Considering the latest months data frame (July) as the base data frame, the records from all previous months are appended to create a data frame with projects where stage change from launched state to another state like successful, failed or canceled. This data frame is currently being just created with the concept of getting some insight of project historical time line (not sure). —Need mentor discussion.

The code related to this part is in Capstone\_Read.ipynb file.

## Data Dictionary:

1. Kickstarter.com:

Data: <https://webrobots.io/kickstarter-datasets/>

The dataset is available for March 2016 for every month. This data is collected from web crawler created by the company. The latest dataset contains following fields.

1. Id: the unique id
2. Photo: Info for all photo associated with the project.
3. Name: name of the project
4. Blurb: intro & detail
5. Goal: Amount needed to be raised
6. Pledged: Actual amount raised
7. State: Current state of project (canceled, failed, successful)
8. Slug: brief description
9. Disable\_communication: communication allowed to creator
10. country: country of campaign origin
11. currency: the currency of campaign origin
12. currency\_symbol: currency symbol of campaign origin
13. currency\_trailing\_code: TRUE, if conversion needs to happen in user currency
14. deadline: UNIX time stamp for a project deadline
15. state\_changed\_at: Unix timestamp
16. created\_at: UNIX time stamp for project created
17. launched\_at: UNIX time stamp for project started
18. staff\_pick: TRUE, if staff picked
19. backers\_count: Total user backed the project
20. static\_usd\_rate: USD conversion rate from original currency
21. Usd\_pledged: pledge amount in USD after conversion:
22. creator: details like user name for the creator
23. location: location of the project
24. category: category of the project
25. profile:
26. spotlight: feature spotlight available or not
27. URLs: Url info for a project
28. source\_url: seems like URL for category

More data available on scrapping:

1. First 10 Backers Come from Top Cities
2. First 10 Backers Come from Top countries
3. Number of new backers
4. Number of returning backer