Data Wrangling: An Essential Journey for Data Science

Darshit Jagetiya

MS in Data Analytics, San Jose State University

Dr. Glen Mules

May 9, 2020

**Abstract**

Data Analytics has been adopted by every major company in the world. All these companies have their own approach to obtain information from the data according to their desired output. However, poor quality data can impact their analysis and lead to wrong interpretations which can cause a huge cost to pay. This makes all these companies to perform Data Wrangling. This article focuses on all the steps of Data wrangling that a Data Scientist should apply on a dataset to make it usable. The paper also covers a case study on which various Data Wrangling tasks can be performed to get to the stage of the final stage of the analysis. The case study is based on a real-life data of funding received by Indian Startups. It shows how initial data wrangling tasks will eliminate much of a data analyst’s future work in the final analysis part. At last, the paper discusses many outcomes that can be achieved by applying programming and statistical concept on the dataset.

*Keywords:*  Data Wrangling, Data Munging, Deduplication, Industry vertical, Sub-verticals.

**Table of Contents**

**Section Page No.**

1. Overview……………………………………………………………………....4
2. Introduction…………………………………………………………................5
3. Sources of the Dataset…………………………………………………………5
4. Description and structure of the dataset………………………………….........6
5. Data Wrangling………………………………………………………………..7
6. Data set columns……………………………………………………....7
7. Data type………………………………………………………………7
8. Deduplication…………………………………………………….........8
9. Eliminating multiple values from a column...…………………………9
10. Filling the null values and dropping unwanted columns……………...10
11. Determining the outliers………………………………………………11
12. Achievements possible………………………………………………………..12
13. Summary………………………………………………………………...……14
14. References…………………………………………………………………… 15

**Overview**

Born and brought up in India, I have always been interested in the emerging tech startups around the country. Being a tech enthusiast, I used to find new means to follow these startups and how they are receiving necessary funding from the major investors across the country. When it was to select a dataset for a project, I already knew what to find and I found this dataset for Indian startup funding. The dataset I have chosen describes various information related to the Indian startups and their investors for various type of industries.

Outcomes from the analysis on this data set can be beneficial to millions of aspiring startup founders who are hesitant to initiate due to lack of knowledge about funding situation in the country and perfect locations for the type of industry of their startup. Apart from that, time series analysis will show how is the funding situation in the country with the time, what is the best time throughout the year to receive funding and which industries are preferred by the investors. This dataset can be very helpful to such ambitious future founders including me.

**Introduction**

Data Analysis can be defined as a method of inspecting, cleansing, transforming raw data, and then applying statistical techniques on that data to determine the useful information. This process of extracting key information from the data is just like distinguishing the signal from the noise (Konold &amp; Pollatsek, 2002). Among all these steps, the first three steps of examining and transforming data consumes around 50-80% of the time of a data analyst (Kandel et al., 2011). These steps collectively can be known as *Data Wrangling* or *Data Munging* (Endel & Piringer, 2015).

Data Wrangling is often considered as the tedious part of the Data Analytics process. Again, there are no other software can fully replace this task as it will vary for each dataset. This makes it equally important as applying the statistical and programming skills in which the data quality is improved by removing the errors and the irregularities. The main reasons for a poor quality of data is due to errors made during data entries and the differences in data representation among different sources. These errors can make huge variations in statistical calculations and thus lead to wrong interpretations. This makes data wrangling a necessary task to get insights from a pile of data either to make a machine learning model or for analysis. As mentioned in Godsey 2017, spending a little extra time on data wrangling can save a lot of pain later.

**Source of the dataset**

The dataset that I have chosen for this project is based on Indian startups. It mentions the various investors, their funding, startup names, and locations. It is opensource and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. The dataset was formed and collected by a site Trak.in which is a blog based on Indian Business, Technology, Mobile & Startups which features trending news, views, and analysis on the trending topics.

**Description and structure of the data:**

The dataset consists of 10 columns and 3044 rows. The column names in the dataset are Sr No., Date, Startup Name, Industry Vertical, Sub-vertical, City Location, Investors Name, Investment Type, Amount in USD, and Remarks.

The meaning of each of these columns and what they represent is explained below:

* Sr. No: The unique id for reference for each of the startups in our dataset.
* Startup Name: The official name of the startup which received the funding.
* Date: It represents the date of funding.
* Industry vertical: It is the category of the startup. For example, Logistics, E-Commerce, Hospitality etc. are few of the industry verticals.
* Subvertical: It is the subcategory of the Industry vertical. For example, online seafood and meat ordering startup subvertical in E-Commerce industry vertical.
* City Location: It is the city in which the headquarters of the startup is based.
* Investor name: It is the name of the investors.
* Amount in USD: It is the funding given by the investors.
* Remarks: The additional comments if required related to the corresponding startup.

The datatype and the structural changes in these columns are explained in the next section of this article which shows all the wrangling measures that are required. These columns mentioned here are the new names are the column names assigned after changing in the original dataset as required for further wrangling. The next section shows the various data wrangling steps that can be performed on this dataset to get the accurate inferences

**Data Wrangling**

As discussed before, data wrangling is the process made up of certain steps from taking data in raw and unstructured form to make it usable for a statistical and programming software. For different datasets, the steps of wrangling will be different to get good results during analysis. This section explains such measures that should be applied on the Indian Startup’s Dataset. Many of these steps can be found in the book *Think Like a Data Scientist* by Brian Godsey

A typical data wrangling tasks include column renaming, datatype changing, deduplication and record matching, eliminating multiple values, filling the null values, dropping unwanted columns, and determining the outliers. Each of these tasks are explained below.

1. **Data set columns**

The first step of the data wrangling process would be check and rename each of the columns to something short as the column name is used multiple times in the analysis, and to something related to the data in that column which will be informative at the time of analysis.

1. **Datatype**

The second step in the process would be to check for the datatype for all the columns and change it as per the information filled in each of the column. Sometimes, data cannot be changed into the required datatype. For example, a data might be tagged along with the symbol currency after the amount. In such a scenario, string manipulation is to be used to extract the amount from the given values. In the funding amount column of our dataset, the numbers are separated with commas after every three places. These commas would throw an error if tried to convert to float or integer before removing them. To tackle such issues, the funding amount column can be parsed with regular expression to allow only integer values. Similarly, on converting date to timeseries, it was found that few of the values have dots and commas instead of a slash to separate month and the year. Such errors can also be removed using regular expressions which will allow only a specific format for dates and time.

In this dataset, the datatype of ID and Funding amount would change to integer, funding date would change to time series and the rest would change to string. Changing the funding amount to integer will help in determining the mean, median and other statistical analysis whereas ID will help in fetching the data and identifying each of the startups as ID is unique. Changing the datatype of funding date will help in determining the timeline about when various funding was released and thus, help in future analysis. The remaining columns have words and alphabets for which string is the suitable data type.

1. **Deduplication**

One of the key steps to eliminate data duplication is to convert all the string characters into lower case before starting operations. A computer algorithm will take same alphabets with different cases as two different alphabets and thus will result in data duplication. Sometimes, word duplication takes place due to a space between two words. For example, data set and dataset mean the same, but will be considered differently by the algorithm. Similar cases are noticed in the investment type and startup name columns.

Most of the investment types are divided among Private Equity, Seed Funding, and Crowd Funding. On listing all the unique investment types, it was found that the dataset has duplicated its values. Crowd Funding & CROWD funding, Private Equity & PrivateEquity, Seed Funding & SeedFunding are few of such other examples. Such duplications are possible in startup names as well. It is noticed that the dataset has different names for same startups. For example, Oyorooms, Oyo Rooms and Oyo are all same but will be all considered as different names by the algorithm. Some top startups are listed as per the website name. For example, Flipkart and flipkart.com are one and the same.

There are ways in which these duplicate values can be removed. To remove duplicate values from this dataset, these are the measures that can be performed.

1. Uncapitalizing the columns where duplication is possible due to lower and upper cases of capitalization. For this dataset, it is best to uncapitalize startup names, investment type, industry vertical, sub-vertical, city, and investors name.
2. Using Regular Expressions to remove special characters from strings, keep optional spacing, and removing website extensions (.com, .in).
3. By creating a copy of data frame and checking subset for each value in one data frame and comparing with all values in the other data frame. Then a new list of subsets will be small, and all the duplicate values can be identified.
4. Drawing word clouds are one of the best ways to find the duplicate values.

The above tasks will remove most of the duplicate values for this dataset.

1. **Eliminating multiple values from a column:**

Sometimes there can be more than one value for a few columns in a dataset. There are two columns identified having multiple values in this dataset as well. Some startups are listed with multiple locations and multiple investors for a single entry. The first step to perform with multiple values is to separate both the values using split function and defining the delimiter. For different data columns, the criteria for selecting the values will be different.

For startups with multiple investors, there are two options that can be considered. First is to take the first value and remove the other and second is to make different row entries for each investor and divide the amount equally among all the investors. In this dataset, taking only one investor might highly impact the dataset when finding top investors in the company and few more queries. Thus, creating new entry for each investor is the best option. Similarly, for multiple locations, the amount can be divided between locations or select the first location which will generally be the headquarters of the startup.

1. **Filling the null values and dropping unwanted columns**

The dataset has 945 rows which does not have funding amount. There are certain ways in which these values can be approximately determined by performing appropriate calculations. The dataset can be grouped on Industry vertical, Subvertical, and Location of the startup and then calculate median values for each of the group. These median values will be used to fill the null values in the funding amount column. There are 25 startups which does not have funding amount along with these 3 columns. Determining funding for these rows will not give accurate values as any other data column is not suitable to calculate these values. It is best to delete these rows from the data frame as the null values will be affecting the overall calculations and will impact the analysis.

We have 2625 rows in the remarks column which are null out of the 3044 rows which constitutes 86.23% of the total rows. The remarks include the type of funding received by the startup such as Series A, B or C funding, Strategic Funding, Additional Funding, or government backed Venture Capital funding. The remarks column does not add sufficient input for any of the aims. Thus, it is best to drop the remarks column. There are no missing values in the funding date and the startup name which will help to determine if the number of startups are increasing or decreasing with the time. There are 4 missing values in the investment type. These values can be determined by finding their corresponding investors in the investors name column. The type of investments these investors have been doing most of the time will determine the null values. For the subvertical column, there are 956 values which are empty. Also, the column subvertical has hundreds of different unique subvertical type which will not be sufficient for our analysis to determine funding for each type. Thus, after determining the null values for other columns, it is best to drop the subvertical column from our data frame.

For startups, which does not have location, can be considered as based at multiple locations or not a specific location. For such startups, location can be filled by ‘Not Specific’ in the city column. There are some startups which does not specify any Industry Vertical. Such startups can be the one who does not lie towards any specific industry and are unique towards its own. Instead of deleting such startups, they can be assigned to a new category -- ‘Other’. These startups will help in determining good locations for startups.

Now that we have filled all the null values in our data set, we will move forward on the next data wrangling steps for data wrangling.

1. **Determining the outliers**

The best way to determine the outliers is to plot histogram or some other visual plots (Godsey 2017). Outliers are the extreme values in a dataset which are not realistic. In the selected dataset, there is a funding column which can have extreme values. For example, a negative value, or a value less than few thousand dollars may be an outlier for the funding. These values will easily catch one’s eye and the outlier can be eliminated. In some other scenarios, a negative value may be a debt for a company where it should not be eliminated. But for this dataset, it seems to be an outlier as funding from an investor cannot be negative.

These are all the wrangling and munging steps that are required on this dataset to make it usable for analysis.

**Achievements possible**

After discussing the data wrangling steps, this section discusses the achievements and the inferences that can be made by imposing statistical and programming techniques on the clean data. There are many important questions that are possible to answer by performing a good analysis on this dataset. Answers to such questions will help many aspiring startup founders in India and others across the world who plan to invest in India. Below are the questions and the way to find the answer to those questions:

1. **Which industries are being favored by the investors for funding?**

* The answer to this question will help in determining the rising industry domain where most of the investors are interested. The number of startups in which the investors invested can be calculated by grouping the industries based on industry vertical and the total count of startups for an industry vertical can be calculated.

1. **How is the funding ecosystem changing with time in India?**

* For each year, the total funding for each industry domain will determine the shift in the interests of the investor. This can be done for each location as well which will determine which locations were preferred by the investors before which are not good anymore and which can be the new hub for startups.

1. **Does location play a role to get funding from the investors? If yes, what cities are preferred by the investors for each type of industry?**

* One can find the number of investors for each location by grouping on locations and the count of investors will give total investors for each location. Similarly, total funding for each location can be calculated by grouping on location and summing the funds, total funding per location per industry vertical can be calculated by grouping on location and industry vertical then summing the funding. For both of these total funding values, average can be calculated by dividing the total startups for each group made.

1. **Who are the major investors in the Indian Startup ecosystem?**

* For each of the year, total investments by an investor can be calculated by grouping on startups and calculating the total funding amount for each investor. Top investors for each location and/or industry vertical can be calculated by grouping them along with the investors.

1. **How much funding does startups generally get in India for various industries?**

* Total funding for each industry vertical can be analyzed by grouping the startups on industry verticals and sum the funding values for each industry vertical. These total values can be averaged out by the number of startups in each industry vertical. This will give the average funding for each industry vertical.

1. **What is the best time throughout the year for a startup to get funding?**

* For each year, the funding invested in different months can be calculated by grouping in months or quarters. This will give the general funding for each time unit. Total funding at a time for each industry can be found by grouping with time as well as industry type.

One can perform the operations mentioned above by applying many programming languages using a tool like Jupyter Notebook or in Hadoop Ecosystem. However, it can be interpreted that the data wrangling steps seems to be the challenging task for a data scientist rather than actual analysis. This is the reason, why it consumes 80% of the time of a data scientist. If the data is wrangled, then only answers to these questions are possible and useful information can be extracted from this dataset.

**Summary**

Data analysis can provide answers to many questions in today’s world. Although, to reach the stage of analysis, performing the steps of Data wrangling effectively is very important. With the dataset for Indian Startups, many important outcomes can be made which can turn out to be beneficial for millions of aspiring Indian startup founders. However, to deal with so many irregularities, such as missing values, multiple values, and determining outliers, a well set of measures becomes a necessity for a data wrangler. The article has showed such steps by considering the case study based on the data of Indian Startups. Although, a standard set of measures can never be possible as neither the datasets are same, nor are the irregularities. It can be inferred that high level of coding is required even to perform data wrangling along with the analysis on the dataset.

Therefore, future research should be conducted in more realistic settings by uploading this dataset on Hadoop ecosystem and perform the remaining tasks of data wrangling and analysis that have been discussed in this article. These remaining tasks requires considerable coding experience which would be something I would work on to get the best output out of this wonderful dataset. Collectively, it can be concluded that a key task of a data scientist includes data wrangling along with the analysis which might impact the output significantly.

**References**

Endel, F., & Piringer, H. (2015). Data Wrangling: Making data useful again. Retrieved from <https://doi.org/10.1016/j.ifacol.2015.05.197>

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., Ham, F. V., Riche, N. H., . . . Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. Retrieved from doi.org/10.1177/1473871611415994

(Kandel et al., 2011)

Konold, C., &amp; Pollatsek, A. (2002). Data Analysis as the Search for Signals in Noisy Processes. Journal for Research in Mathematics Education, 33(4), 259. Retrieved from doi:10.2307/749741

Frankenfield, J. (2019, April 27). How Data Analytics Work. Retrieved from https://www.investopedia.com/terms/d/data-analytics.asp

Godsey, B. (2017). *Think Like a Data Scientist tackle the data science process step-by-step*. Shelter Island, New York: Manning.