

# Online Data Preparation

Luca Di Luccio and M.Tanveer Jan

Università degli Studi di Salerno, Italia {l.diluccio1, m.jan}@studenti.unisa.it

**Abstract.** The advent of IoT devices is enabling automated data collection from every area of interest in the modern society. Little sensors can be placed around a city or any other point of interest in order to create smart networks that can read the data and use it properly. The measurement of such devices are important artefacts that helps companies from all around the world to better develop their business strategies. For this very reason there has been an outstanding growth in this research field. One of the most important aspects about data management is the data cleaning and preparation. This step of the data collection process is often neglected because of the effort involved in such operation, effort that can be easily be seen as superfluous. Unfortunately, the *Garbage in - Garbage out* principle applies to our models too. Raw data must be cleaned, augmented and structured to better convey its meaning. Before the advent of the Internet of Things the data needed could be easily processed as the whole dataset was known *a priori*, but the sheer size of the modern day data needs can't ensure it. Right now we must think about doing preparation processes without any knowledge of the next set of data, making the cleaning process harder and less reliable. Our research will explore the most common ways to deal with such problems, the tools to make a better data preparation and what results we have found during the implementation of our toy project.

**Keywords:** Online data · Data preparation · Internet of Things.

## 1 Introduction

Data collection has become one of the most important activities of any business as data itself can be obtained more easily than ever and can help in an increasing number of contexts. Even if the asset is vital for the company, it is not uncommon to find data collection strategies that are not performing as well as needed because of little-to-none data manipulation. We find important to specify that Internet of Things and/or sensors networks are only good at data extraction, but most of the times the output of such networks does not represent useful information for the company. Different techniques for data preparation are used in such environments in order to refine the data acquired; the usual work flow may vary based on the company needs and their policies for data management.

Data preparation is an umbrella term that describes the manipulation of raw data in order to make it easily manageable and accurately analysed as needed.

It is the first step in data analytics projects and can include different operations such as:

- Data Ingestion
- Data Fusion
- Data Cleaning
- Data Imputation
- Data Editing
- Data Delivery
- many others

The data acquired must undergo those procedures in order to become useful and structured information e.g. measurement from a gyroscope are collected as tuple of raw data "-5, 6.8, 9.5" that we must modify in order to reach a useful representation to it "the device is upside down".

### 1.1 Application contexts for data preparation

A research from [6] indicate that the poor data quality is a primary reason for about 40% of all business initiatives failing to achieve their targeted goals. Data Profiling and Data Cleansing are two essential building blocks of these Information Management initiatives and before we get closer into the details of the use cases for these components here are some basic definitions:

- **Data Profiling** is the process of examining the data collected in some existing data source and obtaining statistics and more insightful information about the data stored.
- **Data Validation** is the process of ensuring that the data collected is clean, useful and correct as unbalances and/or mistakes in the dataset can off-balance the whole computation.
- **Data Quality Assessment** is the process of finding and exposing problems with the dataset in order to plan data cleaning and data enrichment strategies.
- **Data Cleansing** (also named: *Data Cleaning* or *Data Scrubbing*) is the process of detecting and correcting (according to the selected strategies) corrupted or inaccurate records from a record set.

After this high-level definition, let's take a look into specific use cases where especially the Data Profiling capabilities are supporting the end users in their day-by-day work and enabling them to get better understanding or insight into the data they are using.

Let's start with a classical Data Warehouse use case, where a Data Architect or Data Warehouse designer is working on the data provisioning side of the data warehouse and is going to integrate data from multiple data sources. It is essential that the end user can not only get the information about tables, views and technical objects are available at the databases or data sources, but also

to get an understanding what kind of content is stored within the tables with specific labels or headlines e.g. a column named "Color" is only an indicator of what might be stored within the column, nothing else. Data Profiling enables users to get a quick insight e.g. by using the frequency distribution we can visualise how many different attributes exist and what are the most frequent ones. Based on data profiling results the Data Architect can then also identify if different, inconsistent representations of same information exists in the whole dataset. Based on that finding he/she can already set up his mapping tables within the data flow to standardise and unify the new content.

Another use case that we studied is [7] data preparation framework for multi-database language by Kai & Eike. They described data preparation as a crucial part as good quality data can have huge impact on the results. They proposed certain steps for data to be prepared for modeling. The steps could be carried out in any order depends on that data and the result you wish to see and sometime could be iterated multiple time to get best possible results. Steps that are described by them are data selection, data transformation, data cleaning and data reduction. Selection of the data has to be done at the start as the identifying a clear domain of the data is really important, whereas the other steps can be overlapped and be done as you wish to see fit.

**Data Transformation** - Collection of such a huge amount of data can really be a headache for sorting it out. First of all as data is being collected from different sources so in order to transform our data into similar form, we should have transparent access to these sources. Moreover all conflicts must be solved especially structural and semantic conflicts. Other than that data discrepancies should also be addressed using reconciliation functions and user-defined ones.

**Data Cleaning** - Cleaning the data can affect the results of the product quite significantly. Due to collection of data from different sources there can be a lot of inconsistencies which could result in poor results. There are a lot of sub-problem that may exist in the data collected, we will be discussing only a few of them. The first problem that exists in most of the data collected is the duplication of same values. We must identify the duplicates and try to adjust the data according to it, only those values should be kept where it has most effect on the results. The second is missing values. Same as data is duplicated, there may also be some data that has missing values, so we must fill those values with specified values i.e. in numerical context the missing values must be labelled as 0 or NA. Third problem is detecting outliers in the data that belongs to specified domain, if the values doesn't belong to the domain from which data is collected or it belongs to domain but we are not interested in that parameter to which the data belongs to so it must be trashed out so that we can have results. The fourth problem that may have a huge impact on the results is the noise that data contains although its number is not as large as the others but the impact of even less number of noise is huge. Noise is described as key factor in data mining as described by R.Y. Wang, V.C. Storey, C.P. Firth, in their journal "A Framework for Analysis of Data Quality Research" [8].

**Data Reduction** - In today's technological era and deployment of IoT devices in every field of life had made a huge difference on the data is being collected for different purposes. Huge number of data is being collected which need to be reduced in size as it not possible to handle such large amount. There are several ways to approach this. The most common are GroupBy and Aggregation, projection, sampling and discretization.

## 1.2 Evolution of data

Information technology have come a long way since the early 90s. Many advancements are made since then in all fields of information technology. Data is the a key pillar of today's technological era and there are now a days a huge amount of data available to almost every aspect of life to be analysed and used for further research. Decade ago, data was only used by experts team of IT and if a person needs access to it for their use they would have to request it to the team upon providing them with a report. Now a days it has changed the course in opposite direction, almost all the people have access to all kind of data for their intelligence purposes. With the advancements and usage of Artificial and Machine learning in daily routines, data has become more important. Even though the ML has become common in consumer world, it has only showed the true advantages in enterprises. Many financial corporations are still in the process of figuring out what can be stocked from the amount of data they have and how much will be sufficient to start a project and measure its returns in investment by using ML as the tool.

In corporate world, the true value of ML can only be understood if you understand the evolution of data and how it has transformed. In the early days of internet, most of the focus was on the Online transaction processing systems (OLTP). OLTP was the backbone for powering website and that's why every effort was being made for improving the reliability, and its availability. With the growth of OLTPs and implementation of E-Commerce site much more things came into existence like who is using and from where are they using so that new customers must be brought into it. This gave birth to new applications that can be used for business intelligence and analytics engines. After all these, social media came into existence which in terms gave a sense into user-generated content. Now the analytics engines that were designed for corporate world were not suitable for these kind of contents, new sorts of analytics engines were needed deal with content like this. These kind of advancements have changed the way now the people interact with the systems and also changed their way of decision making.

Take an example of commercial travel. Many of the websites early were used to point out lowest prices one place to your destination. Then were introduced the websites that can compare prices across all the sites and gave you results with different parameters of your choice. Now we have sites that let us choose a budget, based on that budget it will suggest us destination and days to travel

in. Our decision has been transformed from decision based on the data available to actions being recommended to us and then to decide whether to pursue it or not. It will only be possible if our goals are clearly defined such as our budget. This kind of intelligence is still not being developed in corporate world. From what we have seen the reason is not because of insufficient data but its because the lack of defining the goals clearly.

## 2 Online data

In computer science, an *online* algorithm is one that can process its input piece-by-piece in a serial fashion, i.e., in the order that the input is fed to the algorithm, without having the entire input available from the start. When we refer to Online Data we have to think about a dataset that must be used while the data is still being produced, usually in real-time. It is possible to consider the *Online Data* research field as the one that concerns the problems of computing with incomplete information.

### 2.1 Challenges in Automation of data preparation

Data preparation is being considered as crucial step in data analysis, previously it was mostly referred to as "janitorial work". With the passage of time and involvement of data into artificial intelligence, it has now gained priority over every other tasks that are involved in the process. There are number of challenges due to which the automation of data preparation has struck a road block in today's time, but hopefully will be soon be possible as most of the giant techs are have started working on this as they have realised the importance of data for artificial intelligence and machine learning.

Many problems may arise while working with Online Data, this is due to the unpredictability of the measurements. It is not hard to imagine a sensor network that sends critical information to a common node (e.g. a collector node) in order to do basic data preparation before sending it to the company's business critical software. If the collector node of the network has no knowledge about the specification of the data that receives from the sensors it will be impossible to correctly identify what procedures to apply to such data.

**Time consuming** One of the main challenge that is faced in preparation of data is time. More that 80% of overall time is spent of cleaning and preparing of data for further processing. Now this percentage of time is spent when there is a human interaction with the process without a human interaction and to guide the process it will take ages for data to be prepared due to which automation application a not possible to design at this stage while human is still interacting with the process of preparation of the data.

**Integration from multiple sources** The other most common challenge that is hurdle in the way for automation of data preparation is integration of data that is collected from multiple sources. Difference variety of data is being collected which is hard to synchronise together if they do not have similar forms and attributes which in term create a huge hurdle for designing the automation tool. Even after developing an application with this problem, it will need human interaction for sorting out the varying form of data.

**Missing Features** Machine learning is based on the principle of "Garbage in - Garbage out", that's why data cleaning and analysis are the key steps for data preparation. In real-time scenario data comes in many forms and it contains alot of noise i.e null values, missing values, extra data and missing attributes or features. Missing features like these are one of the most common problem that occurs in data pre-processing. These missing features can be the result of malfunctioning sensors, duplication of values or wrong pixels(image processing case). Such data need to be studied and have complete information about the whole scenario only then can decisions be made for solving missing features problem.

### 3 Tool for data preparation

In the last few years the data preparation approaches have increased with the rise of the public attention to this new field. Right now there are many tools that can help the knowledgeable user to have a better understanding about the data collected; while many of this tools have shown new powerful approaches to the the different problems of the field we still couldn't find anything helpful to make it work with online data. This is due to the fact that online data is not finite nor has a definite definition, making modification to the data in such is an impossible task as we need to make choices based on different aspects of the code e.g. a network sensor may send a measurement that, based on the sensor may be a floating point number "12.26685" or an integer multiplied by a certain amount "1226685", and as we should be able to select the best strategy to do as few computation as possible without knowing if in the future we will add new floating point sensor or integer ones.

Because of those limitations due to the very nature of the online data, our research of the tools has been more focused on the ones that can help us with the right amount of flexibility to carry on our project.

### 3.1 Trifacta Wrangler

[3] Trifacta wrangler [3] is one of the most advance and high rated software in today's market. The compant that has developed this software has been working on data cleaning and preparation tool for over 20 years. It transform data for downstream analytics and visualization. The Giant techs like google uses these softwares as the based for running their own data preparation software



Fig. 1: Trifacta Approach to data preparation

### 3.2 OpenRefine

[1] OpenRefine, formerly *Google Refine*, is a standalone open source desktop application for data cleanup and transformation to other formats. It behaves like a spreadsheet but allows for powerful transformations on thousands of tuples with a simple and complete interface.

It helps the knowledgeable user to have a better understanding on the different classes of information. One of the most important operation in OpenRefine is clustering. It refers to the operation of "finding groups of different values that might be alternative representations of the same thing" e.g. the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences. OpenRefine offers different approaches to the clustering problem and the end user can select the one that relates more to his problem:

- Key Collision Methods
  - Fingerprint
  - N-Gram Fingerprint
  - Phonetic Fingerprint
- Nearest Neighbor Methods
  - Levenshtein Distance

- PPM

OpenRefine also support powerful tool for search and replace, making bulk modification easier and a simple interface to have a clear understanding of all the data rows in real time i.e. it is possible to visualise the data as graph in order to show the correlation of the different fields and/or the outliers of the column/row.

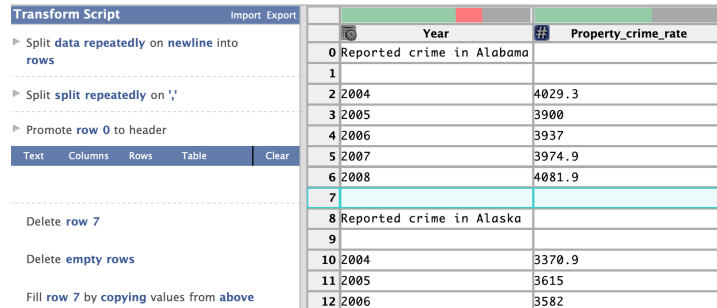
The use of this approach, even if it is really powerful, is limited by the dataset as it must be known a priori, making it not useful for the problems about online data preparation. Implementing a support for such operation is to be considered because of the increase in productivity that it will carry on e.g. the dataset will be automatically modified as the new tuple will be added in a programmatic way.

### 3.3 Tabula

[2] Tabula is java based software developed by Knight Foundation and Shuttleworth Foundation for both Windows and MAC. It is designed to extract tables from portable documents into comma separated value format by a simple web interface or can also be installed as a desktop application. Its not one of the best too in the market a quite useful and handy one, although there are some drawbacks to it, first and the most is that it cannot extract tables from a scanned file. The files that will be used to extract tables must be text-based only then will this work.

### 3.4 Data Wrangler

[5] Data Wrangler is another tool based on the solution of Trifacta. Its was developed in 2011 by a group of researchers working at Stanford University and University of California, Berkeley. This system provides a mixed-initiative interface that maps user interaction to suggest data transform and presents NL description and visual transform preview to help assess each suggestion.



The screenshot shows the Data Wrangler interface. On the left, a 'Transform Script' panel lists three steps: 'Split data repeatedly on newline into rows', 'Split split repeatedly on \',', and 'Promote row 0 to header'. Below the script are buttons for 'Text', 'Columns', 'Rows', 'Table', and 'Clear'. On the right, a data table is displayed with two columns: 'Year' and 'Property\_crime\_rate'. The table contains data for 'Reported crime in Alabama' and 'Reported crime in Alaska' for the years 2004, 2005, and 2006. Row 7 is highlighted in blue.

	Year	Property_crime_rate
0	Reported crime in Alabama	
1		
2	2004	4029.3
3	2005	3900
4	2006	3937
5	2007	3974.9
6	2008	4081.9
7		
8	Reported crime in Alaska	
9		
10	2004	3370.9
11	2005	3615
12	2006	3582

Fig. 2: Data Wrangler Software



### 3.5 Python

Python is a programming language and the most popular one since its was introduced into machine learning and data science. It is considered one of the most useful skill for a developer as it can be used in almost every field of technology.

The main reason behind the popularity and success of python is its usage into data science and machine learning field. Many libraries have been developed for data analysis that can be used with python and can do alot of stuff that may be not possible to do in the interactive softwares that are available. some of the most popular libraries are used for data preparation are:

- NUMPY
- Matplotlib
- Pandas
- Plotly
- Theano

The drawback that python have over other interactive softwares available in the market is that its a programming language and as such the end user must be quite proficient with developing in order to master the tools that are available for it.

In our project we have chosen to work with python to build a script that can help us doing data preparation as it is the most versatile tool in the field. The software we wrote is not the perfect solution for such problems, but is a good showcase of how to approach online data preparation problems.

## 4 Real-World Scenario

A wide range of applications have been discussed in the previous sections; in order to make this research work a bit more grounded into reality we decided to develop a python script that can help the data cleaning process. The script we developed is responsible for performing data preparation across multiple fields for a limited type of noise due to the limitation presented in the last chapters. The implementation of the cleaning algorithm is composed of two distinct phases:

- Specification Phase
- Prediction Phase

**Specification Phase** The Specification phase is based on building a robust system that knows how to deal with the expected data formatting, providing a few data preparation routines that can help to argument the information from the data itself. All those modification are based on the information the end user give us; we must have a clear knowledge about each column of data, i.e datatype, outliers etc. Basically it is a clear image of the data that we're going to expect from the system and we build a script to deal with it accordingly.

**Prediction Phase** The Prediction phase is the weak spot of the system as it is impossible to correctly implement and more research is needed. In this toy project we have written some basic rules to work with null values and other incomplete data.

Basing on our current knowledge of the technologies it is impossible to write a script that can do basic decision on the data itself without any input, the script we wrote do decisions made upon the data by using the information that were specified in the first phase, like applying data imputation technique to replace the null values with median of the upper and lower limit.

#### 4.1 Dataset

To better develop this toy example we have looked at different kind of datasets, spanning different fields and research studies. At the end of our studies we have found the *Titanic Disaster Dataset*. The reasoning behind such peculiar choice resides in the simplicity of the dataset both in terms of size (i.e is not a large dataset) and complexity (i.e. has less than 15 columns and is easily contained in a single file), but at the same time is best suitable for performing data preparation tasks as it contains multiple noise fields and missing values that need to be cleaned before proceeding further.

#### 4.2 Test environment

**Server** Flask is a lightweight web application framework. Its a simple wrapper around Werkzeug and Jinja, and has become of the most popular python web application framework.

**Implementation** The dataset is hosted on a flask server and, in order to use the records that are available, we request a single entry from the it at each http request. Several cleaning rules have been set up for the different records. To amount of noise that exist in today's data is huge and it span different field of studies, we were able to implement only algorithms to clean it, and they are specifically designed for this dataset.

The process starts by sending a single request to the hosted server which replies with a single record from the dataset. The record that is sent as a response is in JSON format. The response is appended to an empty dataframe, that is initialised before the request with the titles of the different columns. The first procedure of data preparation is making all the values are into numerical ones, so the model is able to make a sense out of it based on applying different algorithms.

The data received as response consists of many details among which one is gender, as we said that the first order of data cleaning is that we map the string values into numerical, so we map gender with its corresponding numerical value.

Another task is that dataset consists of details about the passenger's siblings in two columns named "SibSP" and "Parch". These columns are combined into a single column named "FamilyMember" and the values is rounded off to nearest floor value. The graph can be viewed for combined family size related survival rate of the passenger with Siblings and with Parents.

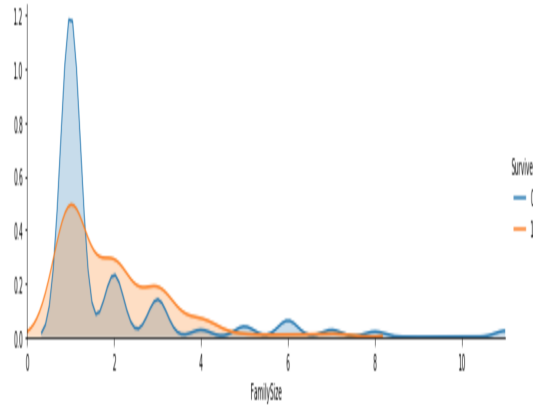


Fig. 3: Family member consolidated

Similarly the "Name" field is characterised by a lot of inconsistencies because of the multiple values for one title, so first we extract the nobility titles out of the that field and we place them into new column called "Title". Now the next step is to remove the inconsistencies from the column we just created i.e "Mr" and "Mister", both are similar title but with different forms, all the inconsistencies like these are replace with the single values, and then transformed them into numerical values. Here is the redundant data of all the titles.

```
print(train_df['Title'].value_counts(ascending=True, dropna=False))
```

Mme	1
Sir	1
Capt	1
the Countess	1
Ms	1
Lady	1
Jonkheer	1
Don	1
Major	2
Mlle	2
Col	2
Rev	6
Dr	7
Master	40
Mrs	125
Miss	182
Mr	517

Name: Title, dtype: int64

Fig. 4: Redundant Titles

Lets have a look into the survived people based on their gender which has been calculated from the title columns, after removing the redundancy.

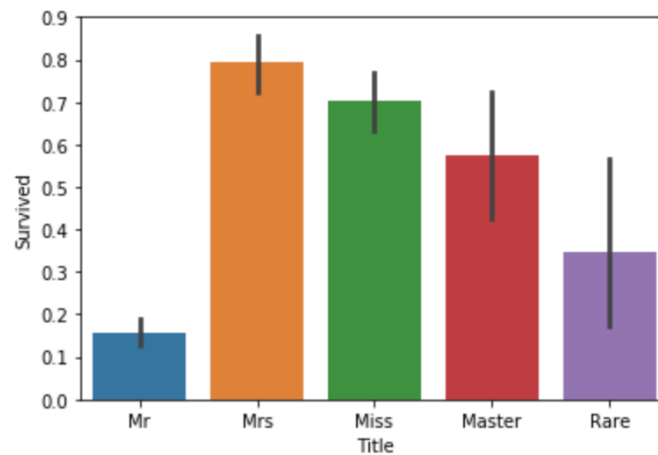


Fig. 5: Survival Chart Based of Titles extracted

The dataset has an "Embarked" field that has very few null values. Because more than 95% of the values are the same it is safe to assume that we can fill the nulls with the most used value of the other field without unbalancing the dataset. Our tool can detect such possibility and automatically fill it.

The dataset also expose a "Fare" field that needs to be adjusted before storing the data. If we have a look at the information in the "Fare" columns we will see that there are some extreme values which need to be addressed before passing it to model to running algorithms, unfortunatly doing this kind of modification requires a good knowledge of the dataset that we don't have because of the

online nature of the dataset i.e. we can't predict if the whole dataset will be unbalanced or not and we can't convert the data accordingly e.g log forms in order to get rid of the outliers.

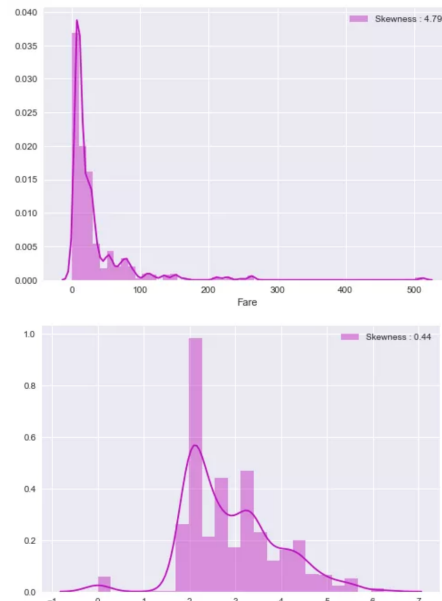


Fig. 6: Fare After and Before Logarithmic form skewness

Above figure shows the skewness of the fares before and after the conversion of fare into log, which have decreased from 4.79 into 0.44. Now also let's have a look at the correlation among all the fields of the dataset to get a clear description of which field is related to which one and up to what extent.

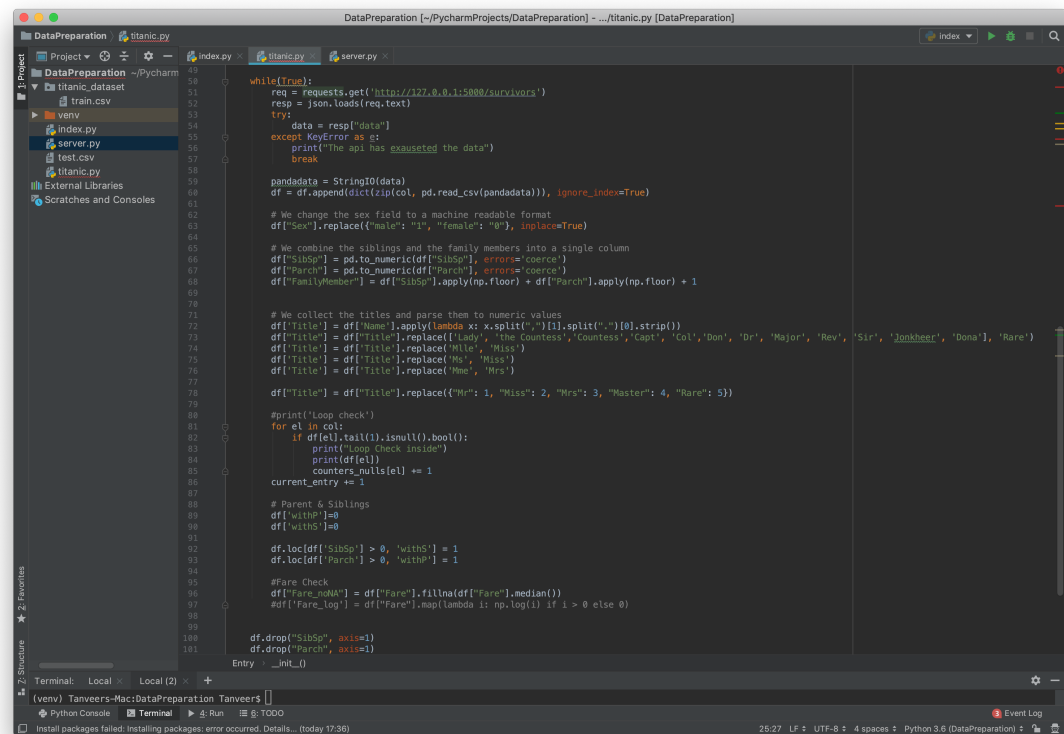


### 4.3 The Tool we develop

There are a number of tools used to develop this script for data preparation. The data preparation tasks are performed on a publicly available dataset named as "Titanic Survivors" [4].

Several python libraries are imported for performing of data cleaning tasks which are as follows:

- Pandas
- Numpy
- StringIO
- Math
- requests
- json
- CSV



```

40 while(True):
41     req = requests.get('http://127.0.0.1:5000/survivors')
42     resp = json.loads(req.text)
43     try:
44         data = resp["data"]
45     except KeyError as e:
46         print("The api has exhausted the data")
47         break
48     pandadata = StringIO(data)
49     df = df.append(dict(zip(col, pd.read_csv(pandadata))), ignore_index=True)
50
51     # We change the sex field to a machine readable format
52     df["Sex"].replace("male": "1", "female": "0", inplace=True)
53
54     # We combine the siblings and the family members into a single column
55     df["SibSp"] = pd.to_numeric(df["SibSp"], errors='coerce')
56     df["Parch"] = pd.to_numeric(df["Parch"], errors='coerce')
57     df["FamilyMember"] = df["SibSp"].apply(np.floor) + df["Parch"].apply(np.floor) + 1
58
59     # We collect the titles and parse them to numeric values
60     df["Title"] = df["Name"].apply(lambda x: x.split(",")[1].split(" ")[0].strip())
61     df["Title"] = df["Title"].replace(["Lady", "the Countess", "Countess", "Capt", "Col", "Don", "Dr", "Major", "Rev", "Sir", "Jonkheer", "Dona", "Rare"])
62     df["Title"] = df["Title"].replace("Mlle", "Miss")
63     df["Title"] = df["Title"].replace("Mme", "Miss")
64     df["Title"] = df["Title"].replace("Mr": 1, "Miss": 2, "Mrs": 3, "Master": 4, "Rare": 5))
65
66     #print('Loop check')
67     for el in col:
68         if df[el].tail(1).isnull().bool():
69             print("Loop Check inside")
70             print(df[el])
71             counters_nulls[el] += 1
72             current_entry += 1
73
74     # Parent & Siblings
75     df["withP"] = 0
76     df["withS"] = 0
77     df.loc[df["SibSp"] > 0, "withS"] = 1
78     df.loc[df["Parch"] > 0, "withP"] = 1
79
80     #Fare Check
81     df["Fare_mdn"] = df["Fare"].fillna(df["Fare"].median())
82     df["Fare_log"] = df["Fare"].map(lambda i: np.log(i) if i > 0 else 0)
83
84     df.drop("SibSp", axis=1)
85     df.drop("Parch", axis=1)
86
87     Entry = _init_()

```

Apart from these libraries, Flask Restful API is used for hosting a local server. The datasets is placed into this server and responds to the requests that

are made from the script we develop in order to fake the continuous stream of data.

The response from the server with data cleaned up to some extent can be seen in the results by running the script

[12] train\_df

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...)	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S

Fig. 7: Raw Data

[30] train\_df

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Title	FamilySize	FamilyGroup	withP	withS	Fare_log	FareGroup
0	1	0	3	1	22.000000	1	0	7.2500	0	1	2	2	0	1	1.981001	1
1	2	1	1	0	38.000000	1	0	71.2833	1	3	2	2	0	1	4.266662	4
2	3	1	3	0	26.000000	0	0	7.9250	0	2	1	1	0	0	2.070022	2
3	4	1	1	0	35.000000	1	0	53.1000	0	3	2	2	0	1	3.972177	4
4	5	0	3	1	35.000000	0	0	8.0500	0	1	1	1	0	0	2.085672	2
5	6	0	3	1	28.235556	0	0	8.4583	2	1	1	1	0	0	2.135148	2
6	7	0	1	1	54.000000	0	0	51.8625	0	1	1	1	0	0	3.948596	4
7	8	0	3	1	2.000000	3	1	21.0750	0	4	5	4	1	1	3.048088	3
8	9	1	3	0	27.000000	0	2	11.1333	0	3	3	3	1	0	2.409941	2
9	10	1	2	0	14.000000	1	0	30.0708	1	3	2	2	0	1	3.403555	3
10	11	1	3	0	4.000000	1	1	16.7000	0	2	3	3	1	1	2.815409	3
11	12	1	1	0	58.000000	0	0	26.5500	0	2	1	1	0	0	3.279030	3
12	13	0	3	1	20.000000	0	0	8.0500	0	1	1	1	0	0	2.085672	2
13	14	0	3	1	39.000000	1	5	31.2750	0	1	7	4	1	1	3.442819	4
14	15	0	3	0	14.000000	0	0	7.8542	0	2	1	1	0	0	2.061048	2
15	16	1	2	0	55.000000	0	0	16.0000	0	3	1	1	0	0	2.772589	3
16	17	0	3	1	2.000000	4	1	29.1250	2	4	6	4	1	1	3.371597	3
17	18	1	2	1	33.736559	0	0	13.0000	0	1	1	1	0	0	2.564949	2
18	19	0	3	0	31.000000	1	0	18.0000	0	3	2	2	0	1	2.890372	3
19	20	1	3	0	28.235556	0	0	7.2250	1	3	1	1	0	0	1.977547	1
20	21	0	2	1	35.000000	0	0	26.0000	0	1	1	1	0	0	3.258097	3
21	22	1	2	1	34.000000	0	0	13.0000	0	1	1	1	0	0	2.564949	2
22	23	1	3	0	15.000000	0	0	8.0292	2	2	1	1	0	0	2.083085	2
23	24	1	1	1	28.000000	0	0	35.5000	0	1	1	1	0	0	3.569533	4
24	25	0	3	0	8.000000	3	1	21.0750	0	2	5	4	1	1	3.048088	3
25	26	1	3	0	38.000000	1	5	31.3875	0	3	7	4	1	1	3.446410	4
26	27	0	3	1	28.235556	0	0	7.2250	1	1	1	1	0	0	1.977547	1
27	28	0	1	1	19.000000	3	2	263.0000	0	1	6	4	1	1	5.572154	4
28	29	1	3	0	28.235556	0	0	7.8792	2	2	1	1	0	0	2.064226	2
29	30	0	3	1	28.235556	0	0	7.8958	0	1	1	1	0	0	2.066331	2

Fig. 8: Cleaned Data

#### 4.4 Results

As we discussed before, this field of studies is very limited because of the unpredictability of the data itself. During the development phase, the team has tried to implement as many features into the online data preparation script as possible but the results haven't been up to the standards dictated by other data

preparation tools.

Not all the transformation needed to have a good data preparation tool have been performed because of the limitation with the online data, but with the help of the specifications and some really simple null values detection techniques we can approximate the results.

## 5 Conclusion

Based of the papers and materials that are studied for the sake of completeness of this paper, we see that data preparation has a huge impact on the results, but unfortunately there are not complete tools that do these tasks automatically due to the complexity of the problems that exists in the data. Large number of researchers and corporation are now more focus on finding solutions for solving this problem and we may see some good outcomes out of these research, but at this point it's difficult to have a complete automation tool. Many data analyst uses different tools for different problems that may arise



## References

1. OpenRefine. <http://openrefine.org/>
2. Tabula. <https://tabula.technology/>
3. Trifacta Wrangler. <https://www.trifacta.com/products/>
4. Kaggle: Titanic Dataset. <https://www.kaggle.com/c/titanic/data>
5. Kandel, S., Paepcke, A., Hellerstein, J., Heer, J.: Wrangler: Interactive visual specification of data transformation scripts. In: ACM Human Factors in Computing Systems (CHI) (2011), <http://vis.stanford.edu/papers/wrangler>
6. Research, G.: Gartner. <https://www.gartner.com/en/documents/1819214/measuring-the-business-value-of-data-quality>
7. Sattler, K.U., Schallehn, E.: A data preparation framework based on a multidatabase language (08 2002)
8. Wang, R.Y., Storey, V.C., Firth, C.P.: A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering* **7**, 623–640 (1995)