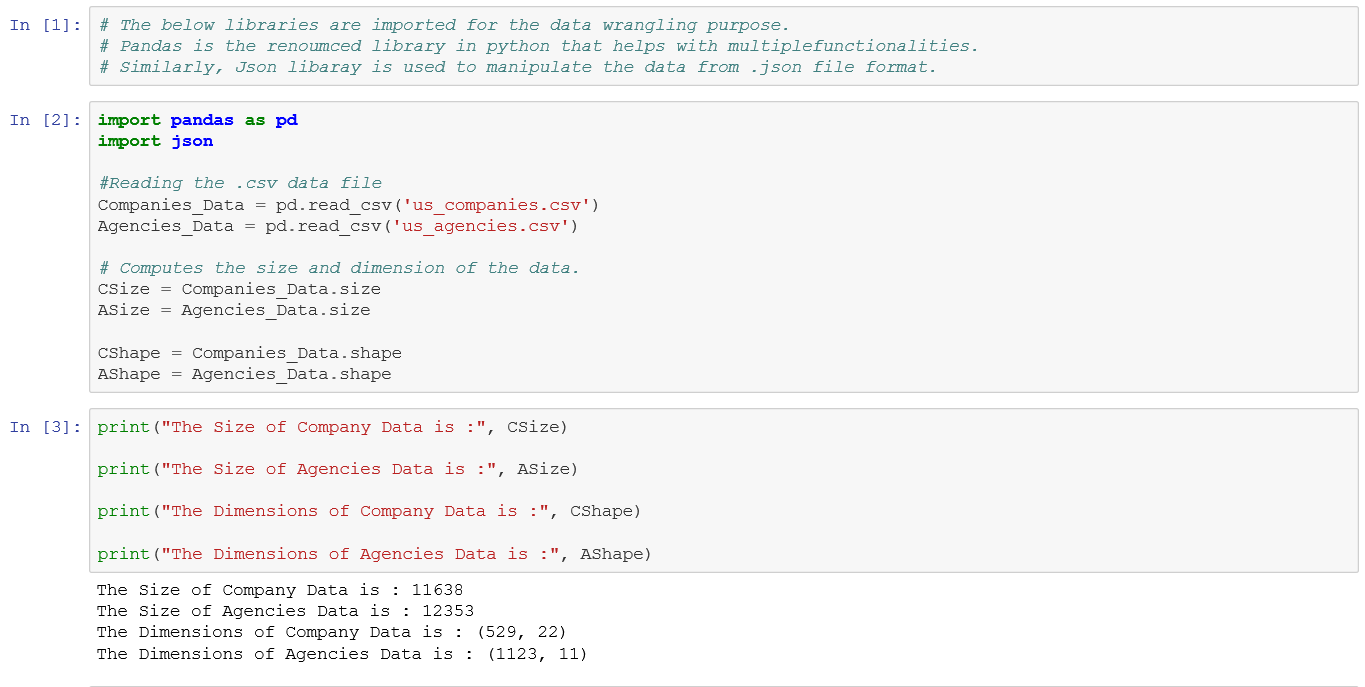
**Analytics in Practice – Data Wrangling**

**Summary**

The libraries such as Pandas, Json, Pandas\_profiling in Python plays an important role in data wrangling process of preprocessing. We also have libraries like dplyr, ISLR, etc in R language. Any real time data have various constraints, such as the data inconsistency, missing values, duplicity, or major issue with reading the data due to its format. Thus, just loading the data does not necessarily help with descriptive statistics of the data, but some time it takes more extraction by manipulation, as we have done it with the json file. Cleaning of data is must to gain the information it store. Therefore, applying the statistics measures we can gain an insight of the data to build a model and further help with Machine learning.

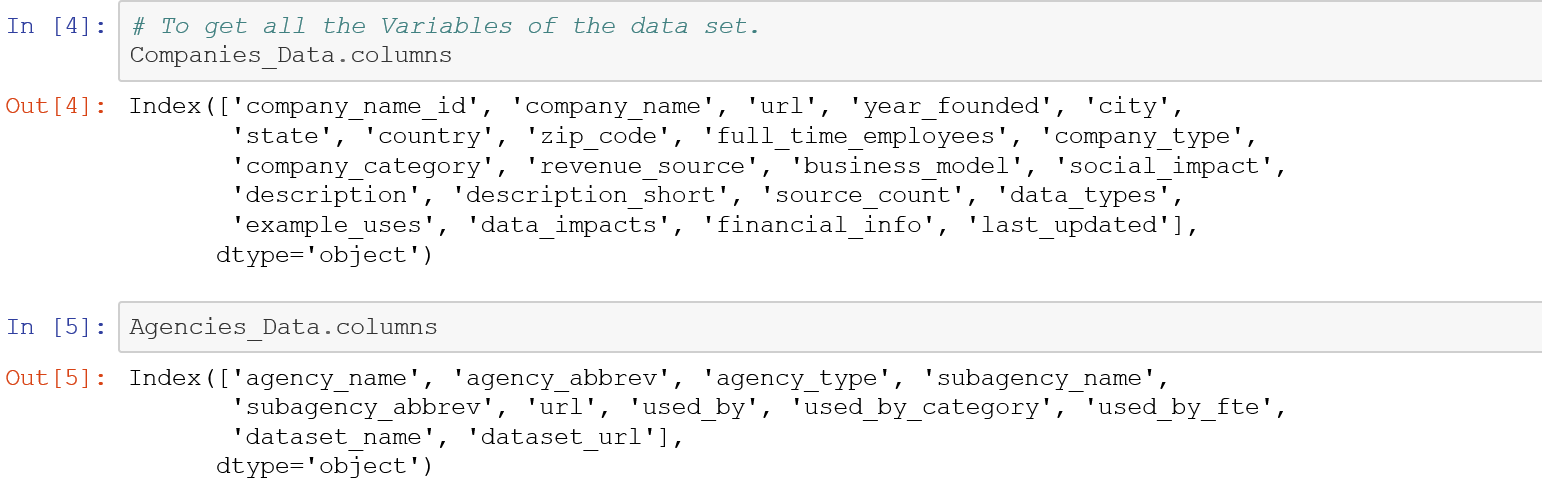
Here while dealing with the 3 different data set of varied format we have learnt the techniques to read these files and summaries the data they have. The data can be of different data type which can create a difficulty in manipulating them, but there are methods to be used to change and manipulate the data. We have also learnt the drawback of an raw data and also the cleaning techniques of them. Extracting the possible information also could be gathered for better data visualization.



**Questioner I**

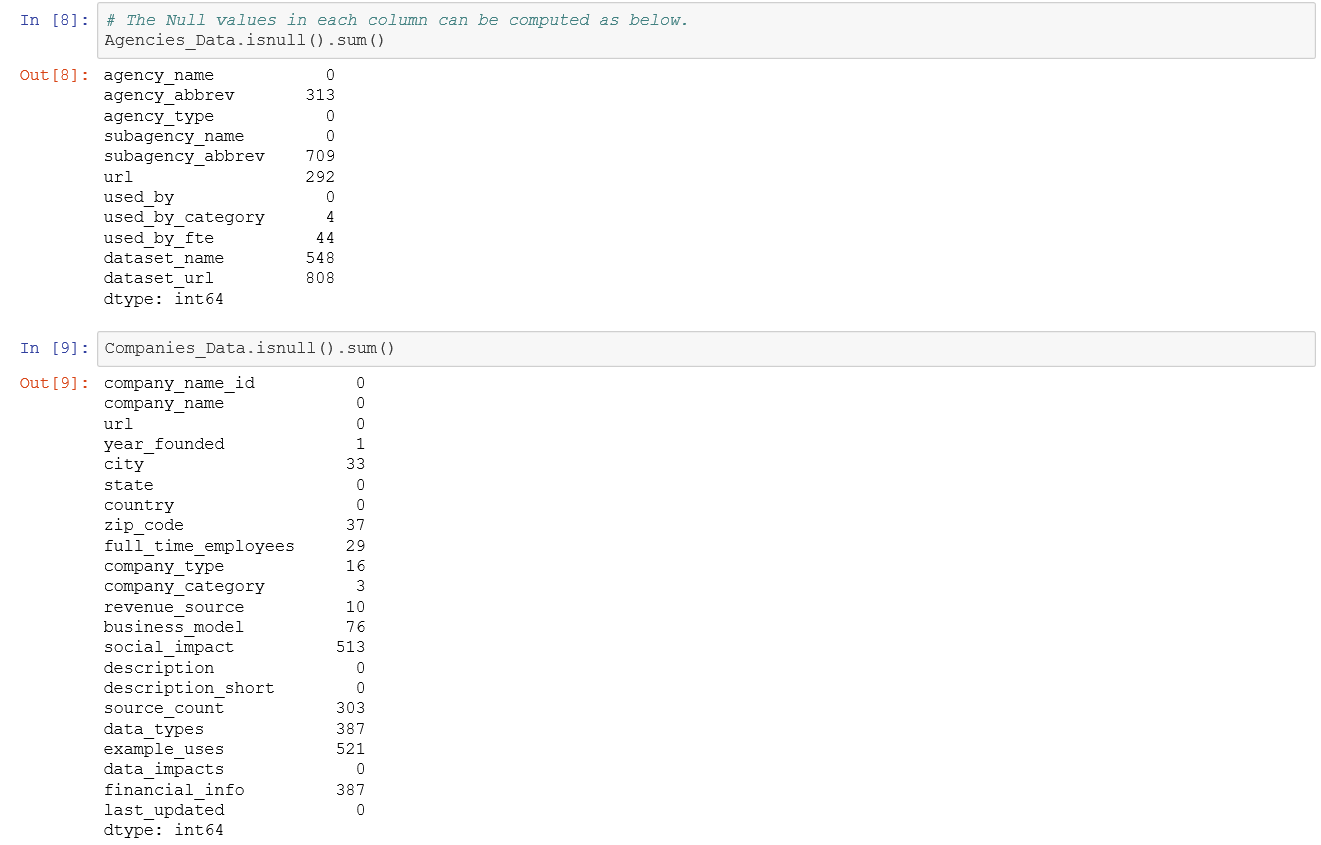
# **Question 1:** Are there any missing columns?

By looking at the Data columns we cannot determine the missing columns unless we have DOMAIN knowledge about the data. But simply looking the columns there do not seem to be any missing columns.



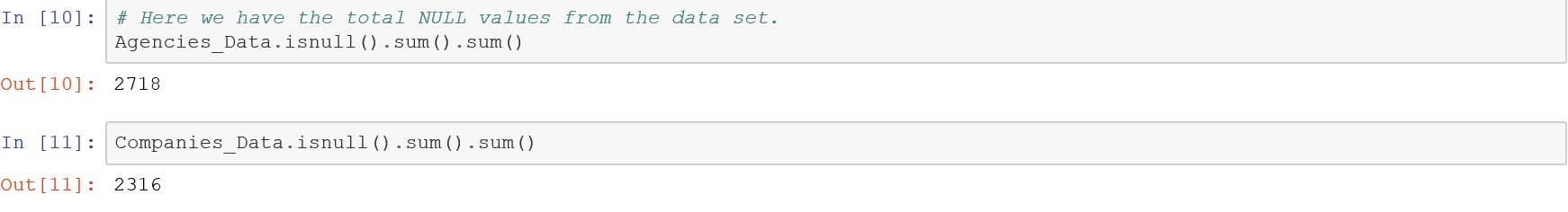
# **Question 2:** Are there any missing column names or errors in the column names? If so, name those columns.

All the Columns are named in both the dataset. The column used\_by\_fte does look like a column that could have a better name.



# **Question 3:** Are there any values in the columns missing?

There are many missing values in the dataset. Above is the detail view of all the missing values column wise and total missing values for Agencies data is 2718, whereas for companies data is 2316.

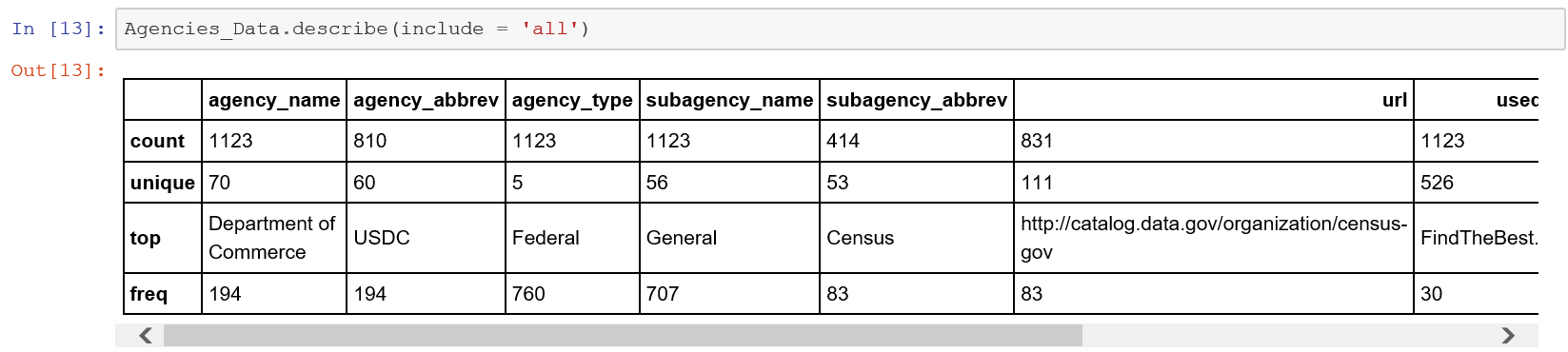


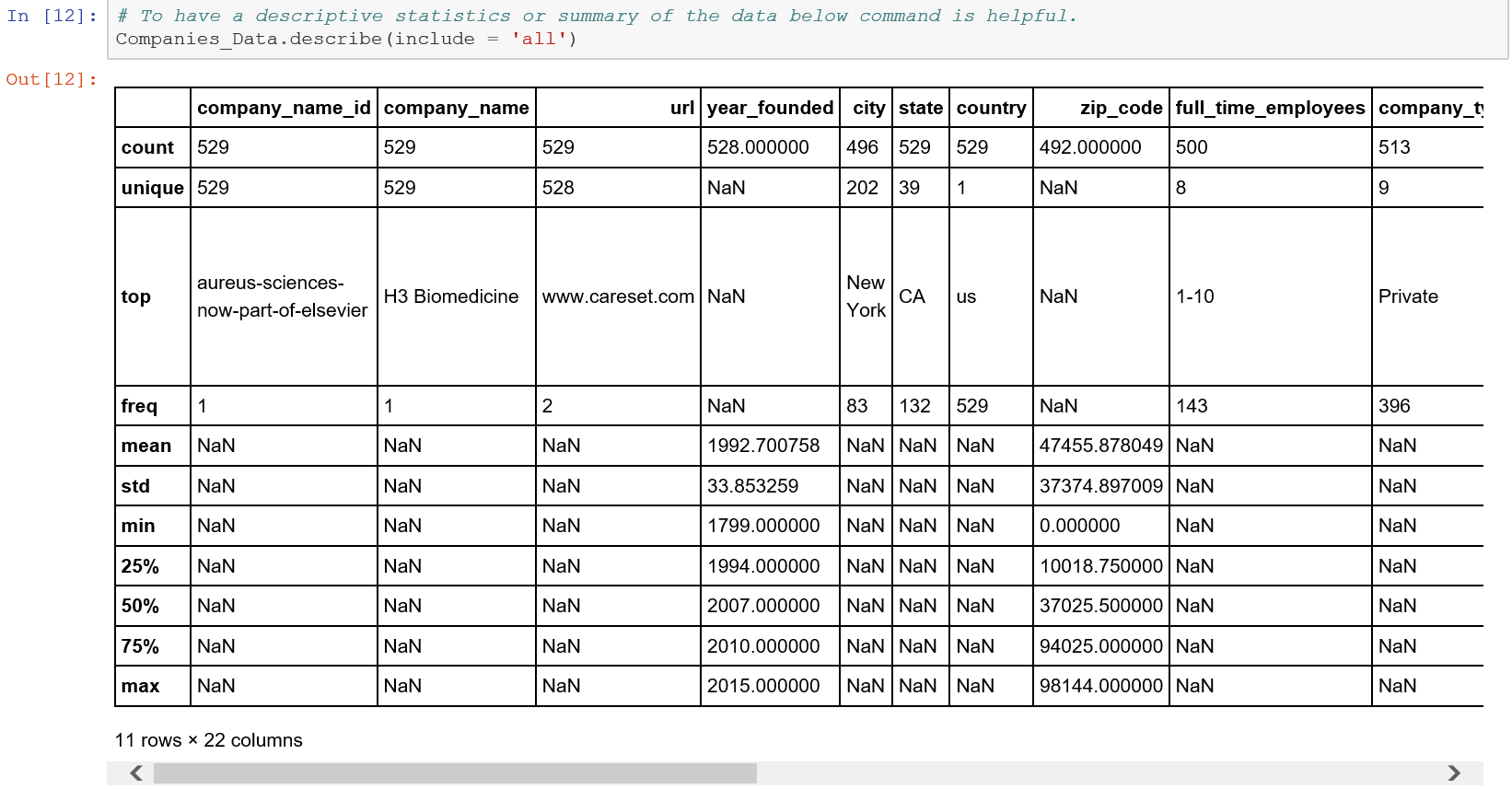
# **Question 4:** How is data organized in each column? Is it properly organized?

The summary statistics of the data have a good picture how the data is organized.

For Agencies the data is a categorical value and hence the distribution cannot be obtained. However, for Companies data the distribution is available.

The data organized is well manner in the columns but due to missing values it is difficult to extract the distribution of the data. Also the data\_impact column is not organized as it contains an empty list in it.





# **Question 5:** Is data in the proper shape for further analysis? If not, why? Explain.

The data are not in the proper shape for further analysis.

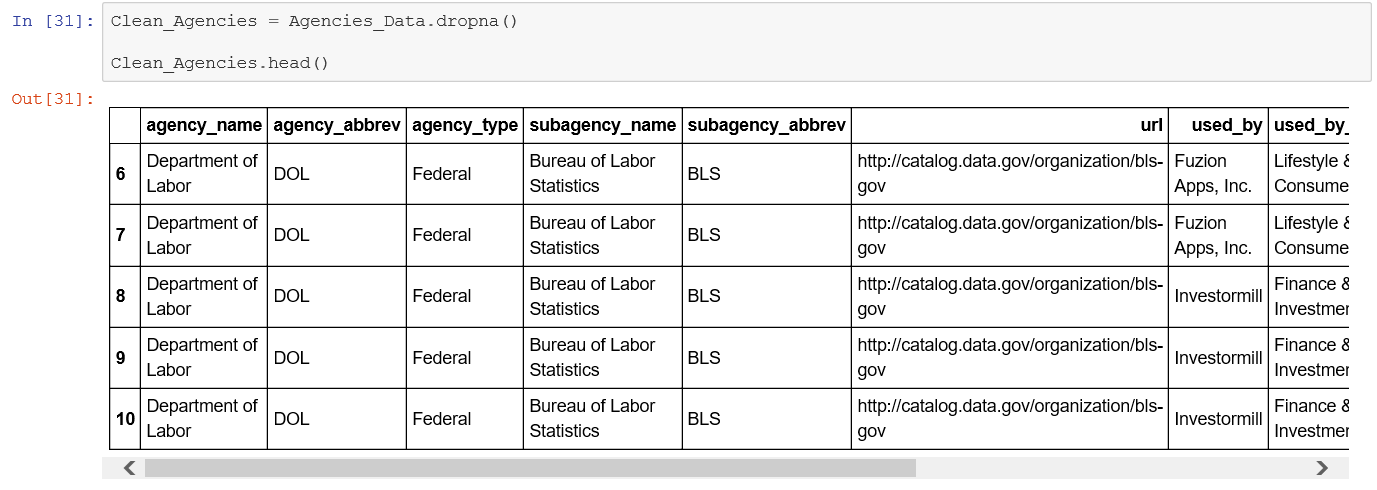
* There are lots of missing values in the dataset (73%).
* There is also redundancy in the data.
* The data is made complex by direct storage of the list values in column like data\_impact.
* Few columns are having highly correlated data which might affect the training later.

**Question 6:** How will you fix this dataset? Describe the methods you will use to fix this dataset for further analysis?

The Data set could be cleaned using below measures:

* Remove the missing values from the column.
* The columns could be dropped with very few data entries.
* The redundancy can be eliminated be removing correlated columns. (Remove duplicated)
* The columns with the list values can be extracted to generate new columns for cleaner visibility.





**Question 7:** How are the two datasets linked to each other? Is there a common “primary key” to connect the two datasets?

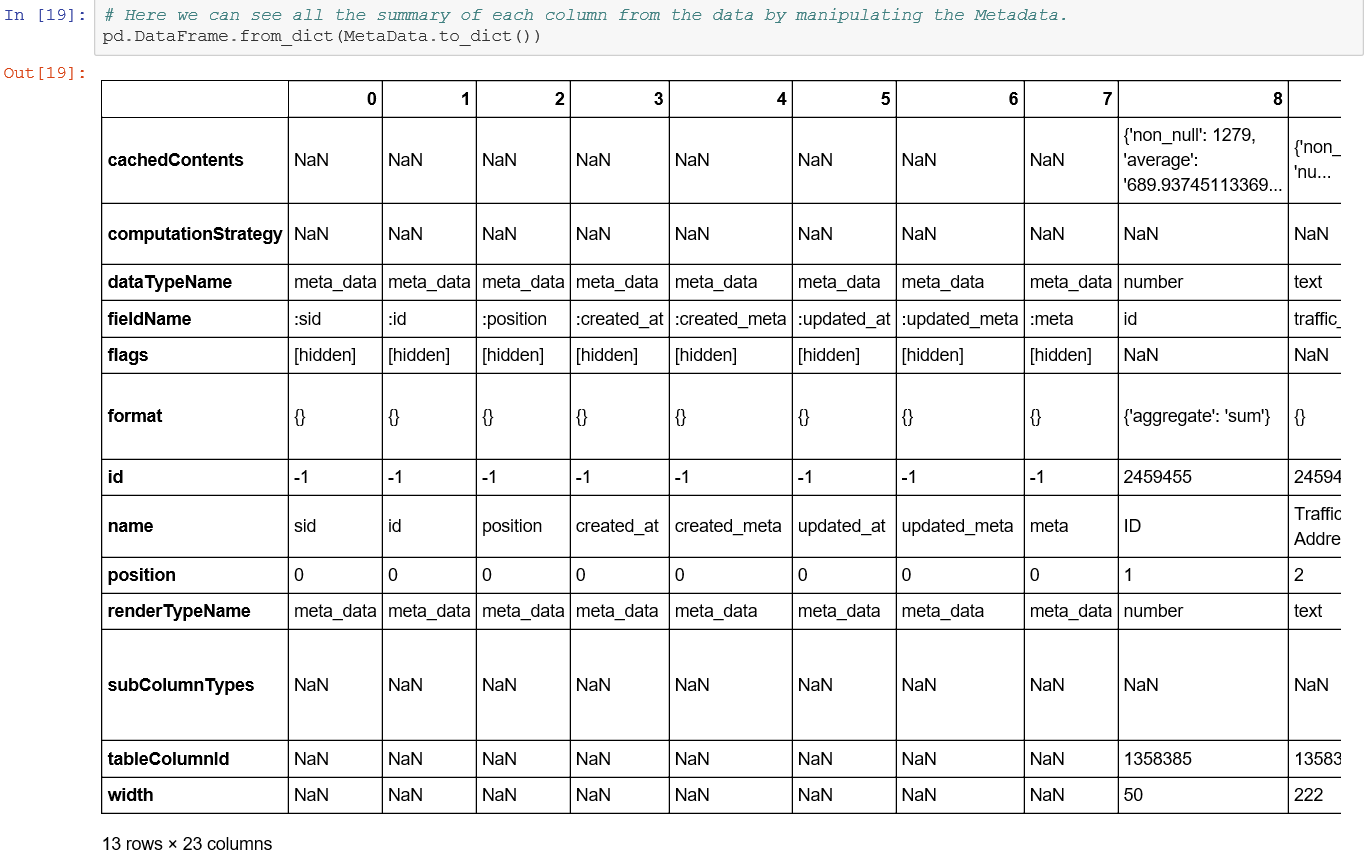
The 2 dataset are linked to each other using the company name: Comapnies\_Data:company\_name = Agencies\_Data:used\_by.

# **Questioner II**

# **Question 1:** How many variables are in the dataset?

There are 23 variables in the data set.





# **Question 2:** Name all the variables?

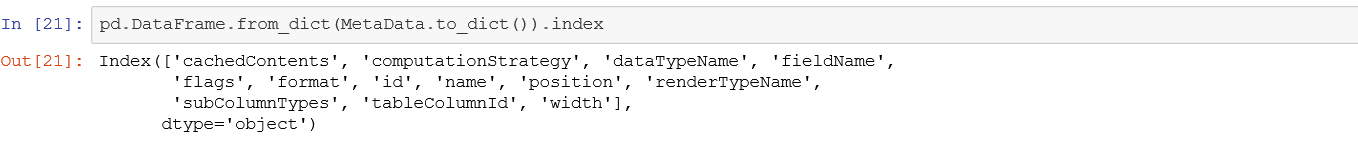
Index(['cachedContents', 'computationStrategy', 'dataTypeName', 'fieldName',

'flags', 'format', 'id', 'name', 'position', 'renderTypeName',

'subColumnTypes', 'tableColumnId', 'width'],

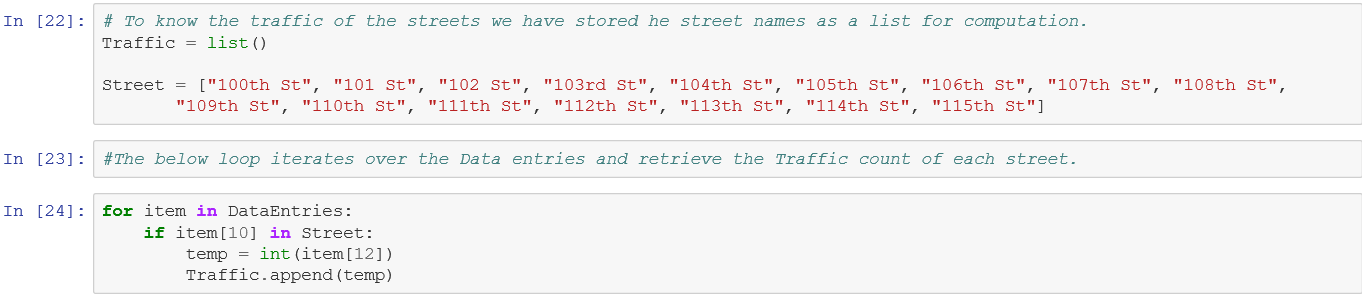
dtype='object')

Above are the names of all the variables from the dataset.



# **Question 3:** What is the total traffic of vehicles on 100th street to 115th street?

The traffic count is computed below.



The Total traffic between street 100th to 115 is 443200

# **Question 4:** What is the total traffic of vehicles on geo locations, (41.651861, -87.54501) and (41.66836, -87.620176)

The total traffic for the given geographic location is computed below.

Total traffic with at Latitude = 41.66836 and longitude = -87.620176 is 8900

Total traffic with at Latitude = 41.651861 and longitude = -87.54501 is 4700

Total traffic at the given coordinates will be 13600

