Devesh Petwal

Assignment 1 Solution

Q1. Are there any missing columns?

Ans. No missing columns.

Q2. Are there any missing column names or errors in the column names? If so, name those columns.

Ans. No missing column names or errors.

Q3. Are there any values in the columns missing?

Ans. Agencies - 2718 missing values and Companies - 2315 missing values.

Q4. How is data organized in each column? Is it properly organized?

Ans. Through the perspective of data analysis, the data is poorly organized as there are several categorical values, repetitive columns and some of the columns will not add any value in our analysis.

Q5. Is data in the proper shape for further analysis? If not, why? Explain.

Ans. Though there is a lot of information through different variables, still the data is not in proper shape for further analysis. Thus, there are lots of changes to be made before further analysis and the reasons are as follows:

1. company_name_id could have been a numeric id.

2. There are lot of missing values which needs to be imputed with suitable method.

3. The format of url is different throught the column.

4. Format of displaying full_time_employees is different throught the column.

5. In revenue_source multiple sources are given which needs to be standardized for analysis.

6. The menaing of some columns are hard to make out and thus a discription of variables is required.

7. There are a lot of unwanted characters.

8. The count in used_by_fte needs to be standardized.

Q6. How will you fix this dataset? Describe the methods you will use to fix this dataset for further analysis? It can be missing values, NAs, etc.

Ans. There are several steps we can take, such as:

1. Imputing missing values by mean for numeric values and mode for categorical values.

2. Standardizing the values in columns with a particular format.

3. Normalizing the data.

4. Removing duplicate data.

5. Removing the outliers.

6. Finding the meaning for those variables which we do not understand.

Q7. How are the two datasets linked to each other? Is there a common “primary key” to connect the two datasets?

Ans. “company_category” and “used_by_catrgory” of Companies and Agencies respectively can be used as a primary key that link the two datasets.

Exercise 2:

Q1. How many variables are in the dataset?

Ans. 23 variable in the dataset.

Q2. Name all the variables?

Ans.

```
##  [1] "sid"
##  [2] "id"
##  [3] "position"
##  [4] "created_at"
##  [5] "created_meta"
##  [6] "updated_at"
##  [7] "updated_meta"
##  [8] "meta"
```

[9] "ID "

[10] "Traffic Volume Count Location Address"

[11] "Street"

[12] "Date of Count"

[13] "Total Passing Vehicle Volume"

[14] "Vehicle Volume By Each Direction of Traffic"

[15] "Latitude"

[16] "Longitude"

[17] "Location"

[18] "Boundaries - ZIP Codes"

[19] "Community Areas"

[20] "Zip Codes"

[21] "Census Tracts"

[22] "Wards"

[23] ":@computed_region_awaf_s7ux"

Q3. What is the total traffic of vehicles on 100^{th} street to 115^{th} street?

Ans. Total Traffic of vehicles on 100th street to 115th street is: 264000.

Q4. What is the total traffic of vehicles on geolocations, (41.651861, -87.54501) and (41.66836, -87.620176)

Ans. Total Traffic between these geolocations: 13600.