CGT 270 Data Visualization

Module 1 ● Week 2

**Lab 2: Parsing Data**

Name: **Megan Jacobs**

The goal of this lab is to understand the structure of data. In this lab you will change data into a format that tags each part of the data with its intended use. After completing this lab every element of the data, you selected (Tableau dataset) and the two (2) additional datasets you acquired in lab last week will be broken into its individual parts. Answer the following questions and complete the table for each dataset.

1. List the name of the Tableau Dataset you selected in the Acquire Lab:
   1. The name of the Tableau Dataset that I selected is called Global Burden of Disease which was done by the Institute of Health Metrics and Evaluations (IHME)
2. How many rows (records) are in the data set?
   1. 58,906 rows
3. How many columns (variables) are in the data set?
   1. There are 7 columns within the data set
      1. String = 4 variables, date = 1 variable, integer = 2 variables
4. What assumptions are you making about the data?
   1. An assumption that I am making about this data is that age grouping spans about every 4 years, unless they are under the age of one. Besides that there is not much room for assumption when looking at this data.

**What you should be able to do (at the end of this lab):**

|  |  |
| --- | --- |
| Remember | ***Describe*** what happens in the **parse** stage. |
| Understand | ***Describe*** the data in detail according to the parsing specifications. |
| Apply | ***Demonstrate*** the ability to change data into a useful format for future processing. |
| Evaluate | ***Categorize*** the data according to parsing specs. |
| Analysis | ***Identify*** specific features about the data. |
| Create | ***Generate*** a parsed listing of the data. |

**Tableau Data Set**

**In the table below list each variable and its data type (add more rows as needed):**

|  |  |  |
| --- | --- | --- |
|  | **Variable** | **Data type** |
| **1** | Country Code | String |
| **2** | Country Name | String |
| **3** | Year | Date/Integer |
| **4** | Age Group (category) | String |
| **5** | Sex | String |
| **6** | Number of Deaths | Integer |
| **7** | Death Rate per 100,000 | Integer |
|  |  |  |

You may add more rows and attach additional pages if needed.

Save this file as: **LastnameFirstInitial-CGT270Fall2021-Lab2Parsing.pdf**

CGT 270 Data Visualization

Module 1 ● Week 2

**Lab 2: Parsing Data**

**Additional Data Set #1**

1. List the name of the first (1st) additional data set you acquired in the Acquire Lab:
   1. Death in the United States in 2015
2. How many rows (records) are in the data set?
   1. 1,048,576
3. How many columns (variables) are in the data set?
   1. 77 Columns
4. What assumptions are you making about the data?
   1. A majority of the columns have certain codes corresponding to them. So for instance if I wanted to see the resident statues, I would need to check the json codes file included with the dataset to understand the code.

**In the table below list each variable and its data type (add more rows as needed):**

|  |  |  |
| --- | --- | --- |
|  | **Variable** | **Data Type** |
|  | resident\_status | integer |
|  | education\_1989\_revision | Integer |
|  | education\_2003\_revision | Integer |
|  | education\_reporting\_flag | Integer |
|  | month\_of\_death | Integer |
|  | sex | string |
|  | detail\_age\_type | Integer |
|  | detail\_age | integer |
|  | age\_substitution\_flag | Integer |
|  | age\_recode\_52 | Integer |
|  | age\_recode\_27 | Integer |
|  | age\_recode\_12 | Integer |
|  | infant\_age\_recode\_22 | Integer |
|  | place\_of\_death\_and\_decedents\_status | integer |
|  | marital\_status | string |
|  | day\_of\_week\_of\_death | integer |
|  | current\_data\_year | String |
|  | injury\_at\_work | Integer |
|  | manner\_of\_death | integer |
|  | method\_of\_disposition | string |
|  | autopsy | string |
|  | activity\_code | integer |
|  | place\_of\_injury\_for\_causes\_w00\_y34\_except\_y06\_and\_y07\_ | integer |
|  | icd\_code\_10th\_revision | string |
|  | 358\_cause\_recode | integer |
|  | 113\_cause\_recode | integer |
|  | 130\_infant\_cause\_recode | integer |
|  | 39\_cause\_recode | integer |
|  | number\_of\_entity\_axis\_conditions | integer |
|  | entity\_condition\_1 | string |
|  | entity\_condition\_2 | string |
|  | entity\_condition\_3 | string |
|  | entity\_condition\_4 | string |
|  | entity\_condition\_5 | string |
|  | entity\_condition\_6 | string |
|  | entity\_condition\_7 | string |
|  | entity\_condition\_8 | string |
|  | entity\_condition\_9 | string |
|  | entity\_condition\_10 | string |
|  | entity\_condition\_11 | string |
|  | entity\_condition\_12 | string |
|  | entity\_condition\_13 | string |
|  | entity\_condition\_14 | string |
|  | entity\_condition\_15 | string |
|  | entity\_condition\_16 | string |
|  | entity\_condition\_17 | string |
|  | entity\_condition\_18 | string |
|  | entity\_condition\_19 | string |
|  | entity\_condition\_20 | string |
|  | number\_of\_record\_axis\_conditions | integer |
|  | record\_condition\_1 | string |
|  | record\_condition\_2 | string |
|  | record\_condition\_3 | string |
|  | record\_condition\_4 | string |
|  | record\_condition\_5 | string |
|  | record\_condition\_6 | string |
|  | record\_condition\_7 | string |
|  | record\_condition\_8 | string |
|  | record\_condition\_9 | string |
|  | record\_condition\_10 | string |
|  | record\_condition\_11 | string |
|  | record\_condition\_12 | string |
|  | record\_condition\_13 | string |
|  | record\_condition\_14 | string |
|  | record\_condition\_15 | string |
|  | record\_condition\_16 | string |
|  | record\_condition\_17 | string |
|  | record\_condition\_18 | string |
|  | record\_condition\_19 | string |
|  | record\_condition\_20 | string |
|  | race | integer |
|  | bridged\_race\_flag | integer |
|  | race\_imputation\_flag | integer |
|  | race\_recode\_3 | integer |
|  | race\_recode\_5 | integer |
|  | hispanic\_origin | integer |
|  | hispanic\_originrace\_recode | integer |

**Additional Data Set #2**

1. List the name of the second (2nd) additional data set you acquired in the Acquire Lab:
   1. Cancer Rates by US State
2. How many rows (records) are in the data set?
   1. 52 rows
3. How many columns (variables) are in the data set?
   1. 3 Columns
4. What assumptions are you making about the data?
   1. Some assumptions being made are that you would know the state based on the initials of the state (ex: Alaska is AK). Another assumption I would make about this data is that range column is based on averages of the biggest and smallest numbers taken from a much larger data set and then put into that range.

**In the table below list each variable and its data type (add more rows as needed):**

|  |  |  |
| --- | --- | --- |
|  | **Variable** | **Data type** |
| **1** | State | String |
| **2** | Range | String |
| **3** | Rate | Integer |
| **4** |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

You may add more rows and attach additional pages if needed.

Save this file as: **LastnameFirstInitial-CGT270Fall2021-Lab2Parsing.pdf**