CGT 270 Data Visualization
Module 1 ● Week 2
Lab 2: Parsing Data

Name: **Jiaxiang Li**

Tableau Data Set: *Tuberculosis Burden by Country - The World Health Organization estimates the prevalence and mortality of Tuberculosis by country*, Tableau public.

	Variable	Data type
1	Country or territory name	String
2	ISO 2-character country/territory code	String
3	ISO 3-character country/territory code	String
4	ISO numeric country/territory code	Integer
5	Region	String
6	Year	Integer
7	Estimated total population number	Integer
8	Estimated prevalence of TB (all forms) per 100,000 population	Float
9	Estimated prevalence of TB (all forms) per 100 000 population, low bound	Float
10	Estimated prevalence of TB (all forms) per 100 000 population, high bound	Float
11	Estimated prevalence of TB (all forms)	Float
12	Estimated prevalence of TB (all forms), low bound	Float
13	Estimated prevalence of TB (all forms), high bound	Float
14	Method to derive prevalence estimates	String
15	Estimated mortality of TB cases (all forms, excluding HIV) per 100 000 population	Float
16	Estimated mortality of TB cases (all forms, excluding HIV), per 100 000 population, low bound	Float

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

17	**Estimated mortality of TB cases (all forms, excluding HIV), per 100 000 population, high bound**	**Float**
18	**Estimated number of deaths from TB (all forms, excluding HIV)**	**Float**
19	**Estimated number of deaths from TB (all forms, excluding HIV), low bound**	**Float**
20	**Estimated number of deaths from TB (all forms, excluding HIV), high bound**	**Float**
21	**Estimated mortality of TB cases who are HIV-positive, per 100 000 population**	**Float**
22	**Estimated mortality of TB cases who are HIV-positive, per 100 000 population, low bound**	**Float**
23	**Estimated mortality of TB cases who are HIV-positive, per 100 000 population, high bound**	**Float**
24	**Estimated number of deaths from TB in people who are HIV-positive**	**Float**
25	**Estimated number of deaths from TB in people who are HIV-positive, low bound**	**Float**
26	**Estimated number of deaths from TB in people who are HIV-positive, high bound**	**Float**
27	**Method to derive mortality estimates**	**String**

28	**Estimated incidence (all forms) per 100 000 population**	**Float**
29	**Estimated incidence (all forms) per 100 000 population, low bound**	**Float**
30	**Estimated incidence (all forms) per 100 000 population, high bound**	**Float**
31	**Estimated number of incident cases (all forms)**	**Float**
32	**Estimated number of incident cases (all forms), low bound**	**Float**
33	**Estimated number of incident cases (all forms), high bound**	**Float**
34	**Method to derive incidence estimates**	**String**
35	**Estimated HIV in incident TB (percent)**	**Float**
36	**Estimated HIV in incident TB (percent), low bound**	**Float**
37	**Estimated HIV in incident TB (percent), high bound**	**Float**
38	**Estimated incidence of TB cases who are HIV-positive per 100 000 population**	**Float**
39	**Estimated incidence of TB cases who are HIV-positive per 100 000 population, low bound**	**Float**
40	**Estimated incidence of TB cases who are HIV-positive per 100 000 population, high bound**	**Float**
41	**Estimated incidence of TB cases who are HIV-positive**	**Float**

42	**Estimated incidence of TB cases who are HIV-positive, low bound**	**Float**
43	**Estimated incidence of TB cases who are HIV-positive, high bound**	**Float**
44	**Method to derive TBHIV estimates**	**String**
45	**Case detection rate (all forms), percent**	**Float**
46	**Case detection rate (all forms), percent, low bound**	**Float**
47	**Case detection rate (all forms), percent, high bound**	**Float**

A total of 47 columns are presented in the chart, containing data types of string, float, and integer. Most of the records are focused on the estimated number of different group's population, TB detections, with many of them with both a high and a low bound number.

A total of 5120 rows of records are presented in the data set, showing all regions in all countries in the globe, with every column filled with data (with minor exceptions of special regions/circumstances, in that case, the block is left blank with *null* data).

According to the data, we can see how the survey was structured and conducted. With the general estimation of population and TB incidence calculated, the data showed how different confounding variables were found and calculated (that is, the "risk factor" mentioned in additional dataset #2, such as HIV).

Additional Data Set #1: *Treatment coverage Data by country*, data from Global Health Observatory data repository (GHO).

	Variable	**Data type**
1	**Country**	**String**
2	**Year**	**Integer**
3	**Tuberculosis treatment coverage**	**Float**
4	**Tuberculosis treatment coverage (low bound)**	**Float**
5	**Tuberculosis treatment coverage (high bound)**	**Float**

6	**Number of incident tuberculosis cases**	**Integer**
7	**Number of incident tuberculosis cases (low bound)**	**Integer**
8	**Number of incident tuberculosis cases (high bound)**	**Integer**
9	**Tuberculosis – new and relapse cases**	**Integer**

Additional data set #1 originally had 5 columns. But two of the columns (*tuberculosis treatment coverage* and *number of incident tuberculosis cases*) were consist of normal estimation, low bound, and high bound. Data parsing was performed to ensure consistency of categories for columns in this set and that of the original data set. Thus, the parsed data set has 9 columns and 3857 records.

According to the set, approximate proportions of TB cases and treatment coverage can be deduced. Combined with the information of nation/region population of presented in the original data set, along with the new/relapse cases category in this one, further connection of each country's TB treatment coverage, population, and TB cases can be made.

Additional Data Set #2: TB incidence by Age Sex and Risk factor, data from WHO Global Tuberculosis Programme.

	Variable	**Data type**
1	**Country**	**String**
2	**ISO2**	**String**
3	**ISO3**	**String**
4	**ISO_numeric**	**Integer**
5	**Year**	**Integer**
6	**Measure**	**String**
7	**Unit**	**String**
8	**Age_group**	**Alphanumeric**
9	**Sex**	**Character**
10	**Risk_factor**	**String**
11	**Best**	**Integer**
12	**Lo**	**Integer**
13	**Hi**	**Integer**

1. List the name of the second (2nd) additional data set you acquired in the Acquire Lab:
2. How many rows (records) are in the data set?
3. How many columns (variables) are in the data set?
4. What assumptions are you making about the data?

The second additional data set is like the original data set, with variables of country, ISO, year, and TB cases of normal, low bound, and high bound. But it includes variables of age-group, gender, and risk factor. With these added, it has a total of 13 columns and 7299 rows (countries/regions).

With the data presented in this set, we can estimate the influence of gender, age, and risk factors have toward the incidence of TB.