CGT 270 Data Visualization
Module 1
Week 3
Lab 3: Mining Data

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

A. Basic Descriptors

Variable	Data Type	Basic mining procedure
Country or territory name	String/nominal	String length
ISO 2-character country/territory code	String/nominal	String length
ISO 3-character country/territory code	String/nominal	String length
ISO numeric country/territory code	Integer/interval	Median
Region	String/nominal	String length
Year	Integer/interval	Median
Estimated total population number	Integer/ratio	Average, median, max, min
Estimated prevalence of TB (all forms) per 100,000 population	Float/ratio	Average, median, max, min
Estimated prevalence of TB (all forms) per 100 000 population, low bound	Float/ratio	Average, median, max, min
Estimated prevalence of TB (all forms) per 100 000 population, high bound	Float/ratio	Average, median, max, min
Estimated prevalence of TB (all forms)	Float/ratio	Average, median, max, min
Estimated prevalence of TB (all forms), low bound	Float/ratio	Average, median, max, min
Estimated prevalence of TB (all forms), high bound	Float/ratio	Average, median, max, min
Method to derive prevalence estimates	String/nominal	String length
Estimated mortality of TB cases (all forms, excluding HIV) per 100 000 population	Float/ratio	Average, median, max, min

Estimated mortality of TB cases (all forms, excluding HIV), per 100 000 population, low bound	Float/ratio	Average, median, max, min
Estimated mortality of TB cases (all forms, excluding HIV), per 100 000 population, high bound	Float/ratio	Average, median, max, min
Estimated number of deaths from TB (all forms, excluding HIV)	Float/ratio	Average, median, max, min
Estimated number of deaths from TB (all forms, excluding HIV), low bound	Float/ratio	Average, median, max, min
Estimated number of deaths from TB (all forms, excluding HIV), high bound	Float/ratio	Average, median, max, min
Estimated mortality of TB cases who are HIV-positive, per 100 000 population	Float/ratio	Average, median, max, min
Estimated mortality of TB cases who are HIV-positive, per 100 000 population, low bound	Float/ratio	Average, median, max, min
Estimated mortality of TB cases who are HIV-positive, per 100 000 population, high bound	Float/ratio	Average, median, max, min
Estimated number of deaths from TB in people who are HIV-positive	Float/ratio	Average, median, max, min
Estimated number of deaths from TB in people who are HIV-positive, low bound	Float/ratio	Average, median, max, min
Estimated number of deaths from TB in people who are HIV-positive, high bound	Float/ratio	Average, median, max, min
Method to derive mortality estimates	String/nominal	String length
Estimated incidence (all forms) per 100 000 population	Float/ratio	Average, median, max, min
Estimated incidence (all forms) per 100 000 population, low bound	Float /ratio	Average, median, max, min
Estimated incidence (all forms) per 100 000 population, high bound	Float/ratio	Average, median, max, min

Save this document as: **LastnameFirstInitial-CGT270Fall21-Lab3Mine.pdf**

Estimated number of incident cases (all forms)	Float/ratio	Average, median, max, min
Estimated number of incident cases (all forms), low bound	Float/ratio	Average, median, max, min
Estimated number of incident cases (all forms), high bound	Float/ratio	Average, median, max, min
Method to derive incidence estimates	String/nominal	String length
Estimated HIV in incident TB (percent)	Float/ratio	Average, median, max, min
Estimated HIV in incident TB (percent), low bound	Float/ratio	Average, median, max, min
Estimated HIV in incident TB (percent), high bound	Float/ratio	Average, median, max, min
Estimated incidence of TB cases who are HIV-positive per 100 000 population	Float/ratio	Average, median, max, min
Estimated incidence of TB cases who are HIV-positive per 100 000 population, low bound	Float/ratio	Average, median, max, min
Estimated incidence of TB cases who are HIV-positive per 100 000 population, high bound	Float/ratio	Average, median, max, min
Estimated incidence of TB cases who are HIV-positive	Float/ratio	Average, median, max, min
Estimated incidence of TB cases who are HIV-positive, low bound	Float/ratio	Average, median, max, min
Estimated incidence of TB cases who are HIV-positive, high bound	Float/ratio	Average, median, max, min
Method to derive TBHIV estimates	String	String length
Case detection rate (all forms), percent	Float/ratio	Average, median, max, min
Case detection rate (all forms), percent, low bound	Float/ratio	Average, median, max, min
Case detection rate (all forms), percent, high bound	Float/ratio	Average, median, max, min

Save this document as: **LastnameFirstInitial-CGT270Fall21-Lab3Mine.pdf**

B. Categorize

For most parts of the data in my original dataset, the variables are numerical and, most likely, ratio. That makes sense because the purpose of this dataset is to collect information (mainly numbers) of TB cases, special case detections, and high/low bounds of those calculations. For these data, virtually all kinds of statistical calculation can be performed, including average, median, max, min, standard deviation, etc.

For region, country, and detection method, the data type is most likely string with nominal properties, because region names/ISO-codes and detection methods are all categorical and cannot be put in obvious order. For this type of data, the possibilities of further calculation are low.

For some exceptions, like years and numerical country/region codes, I classified them as interval data, for they are presented in numerical form, but cannot be calculated as ratio type of data because they do not have a meaningful zero point. For this type of data, only partial calculations can be used, like median.

C. Temporal

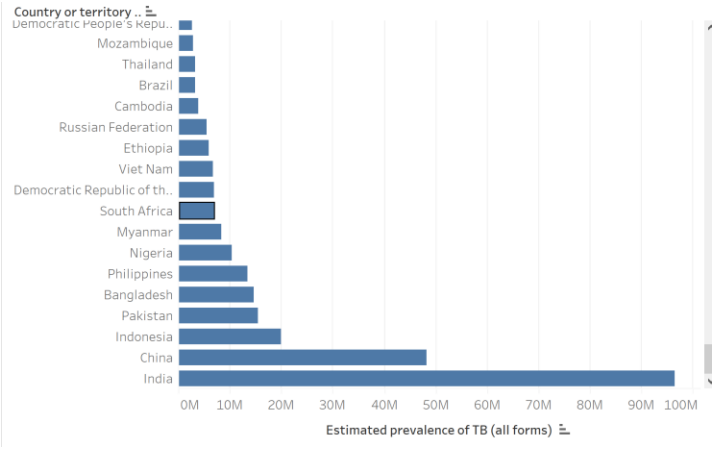
The data in this set **is** temporal, because year is one of the variables presented in the data set and could be updated as time passed. So it records values from year 1990 to year 2013.

D. Range and Distribution

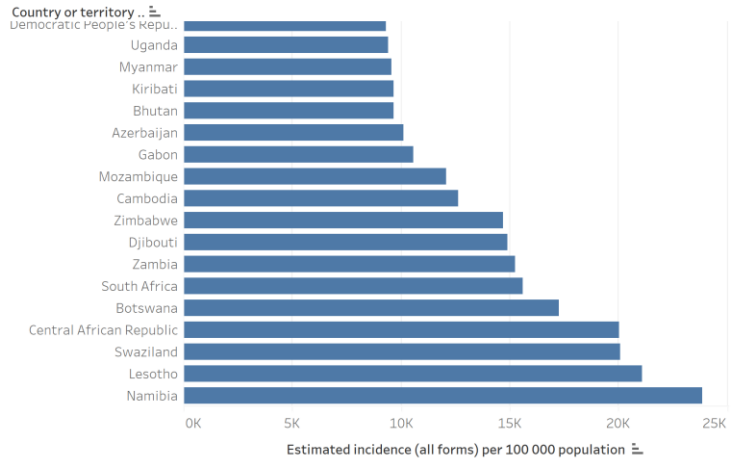
The dataset is very big with thousands of rows. Sometimes it's extremely dense and sometimes it's sparse. It depends on different type of data it's collecting.

The ratio data in this data set can be roughly divided into two types: absolute value and relative value. Examples for absolute values are TB incidence in a certain country, estimated population in a certain country, or estimated number of deaths of TB in a certain country. Examples for relative values are estimated TB mortality in 100,000 population or estimated HIV in TB cases.

For absolute values, the distribution of the data is extremely widespread. For regions that have a high population number, the absolute value of that region will be much higher than that of regions with low population. Thus, China and India are two main outliers in most of the absolute cases with higher values than every other region. It also causes the density to be way high up in regions of lower population than that of regions that have high population.



But when it comes to relative values, the frequency of TB incidence, mortality, and many other aspects all present different trends. For least developed countries like Zimbabwe, Namibia, and Central African Republic. For most of the developed countries and some of the developing countries, the value significantly lowered.



Part II: First (1st) additional data set: *Treatment coverage Data by country*, data from Global Health Observatory data repository (GHO).

A. Basic Descriptors

List the variables from Week 2's parsing lab and provide basic mining procedures.

Variable	Data Type	Basic mining procedure
Country	String/nominal	String length
Year	Integer/interval	Average, median, max, min
Tuberculosis treatment coverage	Float/ratio	Average, median, max, min
Tuberculosis treatment coverage (low bound)	Float/ratio	Average, median, max, min

Tuberculosis treatment coverage (high bound)	Float/ratio	Average, median, max, min
Number of incident tuberculosis cases	Integer/ratio	Average, median, max, min
Number of incident tuberculosis cases (low bound)	Integer/ratio	Average, median, max, min
Number of incident tuberculosis cases (high bound)	Integer/ratio	Average, median, max, min
Tuberculosis – new and relapse cases	Integer/ratio	Average, median, max, min

Part III: Second (2nd) additional data set: TB incidence by Age Sex and Risk factor, data from WHO Global Tuberculosis Programme.

A. Basic Descriptors

List the variables from Week 2's parsing lab and provide basic mining procedures.

Variable	Data Type	Basic mining procedure
Country	String/ nominal	String length
ISO2	String/ nominal	String length
ISO3	String/ nominal	String length
ISO_numeric	Integer/interval	median
Year	Integer/interval	median
Measure	String/ nominal	String length
Unit	String/ nominal	String length
Age_group	Alphanumeric/nominal	String length
Sex	Character/ nominal	String length
Risk_factor	String/ nominal	String length
Best	Integer/ratio	Average, median, max, min
Lo	Integer/ratio	Average, median, max, min
Hi	Integer/ratio	Average, median, max, min

Part IV: Questions and Assumptions

List at least three (3) questions you feel you can answer using the datasets you have acquired and mined. You MUST use complete sentences. Your questions must incorporate ALL three (3) of the data sets you've acquired.

Q1: What type of data represents, in most of the cases, TB incidence in every 100,000 population?

Q2: Is there any outliers in the sets regarding the numerical data representing TB incidence?

Q3: What is the general trending of the TB incidence according to the values of TB cases in each year?

List 3 assumptions you are making in this stage of the data visualization process:

1. **Assumption #1:** The four types of data mentioned in this part of the visualization process are going to be processed in different ways and separately in later processes in order to ensure a organized outcome.

2. **Assumption #2:** Data mining helps the process by categorizing different kinds of data in dataset and helps us build a clearer picture of what we're dealing with.

3. **Assumption #3:** Collecting information about the data and determining if the data is temporal can help us keep track of the future changes and their prospect of later analysis ana visualization.

Save this document as: **LastnameFirstInitial-CGT270Fall21-Lab3Mine.pdf**