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

The goal of this lab is to identify and implement techniques for mining data. In this lab you will identify patterns, extreme and subtle feature about data. You will identify basic descriptors for the data and categorize data according to the specifications defined in the Parse Worksheet you completed in Week 2. After completing this lab, you will:

1. List at least three (3) questions you feel you can answer with the data sets you have acquired (Week 1) and parsed (Week 2).
2. Your questions must incorporate ALL three (3) of the data sets you've acquired from Lab 1: Tableau Dataset, Additional Dataset #1, and Additional Dataset #2
3. List any assumptions you are making in this stage of the data visualization process.

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

Understand	***Describe*** the type of techniques to be used to better understand the data.
Apply	***Execute*** techniques and methods (statistical methods) on the data.
Evaluate	***Examine*** the resulting data and determine if it enables you to answer the question being solved.
Analysis	***Identify*** patterns, extreme and subtle features about the data.
Create	***Determine*** if the data can support the question to be answered.

In the table below list each variable in the Tableau dataset, its data type (parsing) and a basic statistical or mining technique that can be applied to better understand the variable.

Part I: Tableau Data set: Global Burden of Disease

A. Basic Descriptors

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

Variable	**Data Type**	**Basic mining procedure**
Country Code	String	String Length
Country Name	String	String Length
Year	Date/Integer	Average, Max, Min, Range of data, frequency, chronological range
Age Group (category)	String	String Length
Sex	String	String Length or Bool (true or false), Mode
Number of Deaths	Integer	Average, Max, Min, Median

Death Rate per 100,000	Integer	Average, Max, Min

Add more rows to the table above as needed.

B. Categorize

Consider what variables are similar and what variables are different. This will help you to categorize the data. Are the data normal, ordinal or ratio? Take a look at this webpage and video: https://www.graphpad.com/support/faq/what-is-the-difference-between-ordinal-interval-and-ratiovariables-why-should-i-care/

Review the different types of data and indicate the data types in your variables table: https://www.centralriversaea.org/wp-content/uploads/2017/03/F_Four-Types-of-Data-Revised-5.10.17.pdf

Answer: This data falls under four categories. The first being nominal since the three of the columns, country code, country name, and sex, are all groups that fall under categories. The second category would be an interval which would be for number of deaths and death rate per 100,000 since these columns can have a true zero. The third would be ordinal which would be for age group since there is an ordering of the age group from young to old. Finally, the last category is interval for year since there is no true zero to go off of when it comes to year.

C. Temporal

Is the data temporal (represent time, over several years, in years, days, minutes, seconds)?

Answer: Yes, this data is temporal since the data ranges from 1970 to 2010.

D. Range and Distribution

What is the distribution of the data? Few values, small size, evenly spread, sparse or dense? Explain.

Answer: This data distribution is evenly distributed but still dense. I say this because the range of it is from 1970 to 2010 with 58,906 rows of data. All of these rows can be different from one another based on year, sex, country, and age group. This data can span across multiple areas of subject and lead to many different findings based on what the user is looking for.

Part II: First (1st) additional data set: Death in the United States in 2015

A. Basic Descriptors

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

Variable	**Data Type**	**Basic mining procedure**
resident_status	Integer	Median, Mode, Average
education_1989_revision	Integer	Median, Mode
education_2003_revision	Integer	Median, Mode
education_reporting_flag	Integer	Bool
month_of_death	Integer	Mean, Median, Mode
sex	String	String length
detail_age_type	Integer	Mean, Median, Mode, Max, Min
detail_age	Integer	Mean, Median, Mode, Max, Min
age_substitution_flag	Integer	Bool
age_recode_52	Integer	Bool
age_recode_27	Integer	Bool
age_recode_12	Integer	Bool
infant_age_recode_22	Integer	Bool
place_of_death_and_decedents_status	Integer	Mean, Median, Mode
marital_status	String	String length
day_of_week_of_death	Integer	Mean, Median, Mode
current_data_year	String	String length
injury_at_work	Integer	Bool
manner_of_death	Integer	Mean, Median, Mode

method_of_disposition	String	String length
autopsy	String	String length
activity_code	Integer	Mean, Median, Mode
place_of_injury_for_causes_w00_y34_except_y06_and_y07_	Integer	Mean, Median, Mode
icd_code_10th_revision	String	String length
358_cause_recode	Integer	Mean, Median, Mode
113_cause_recode	Integer	Mean, Median, Mode
130_infant_cause_recode	Integer	Mean, Median, Mode
39_cause_recode	Integer	Mean, Median, Mode
number_of_entity_axis_conditions	Integer	Mean, Median, Mode
entity_condition_1	String	String length
entity_condition_2	String	String length
entity_condition_3	String	String length
entity_condition_4	String	String length
entity_condition_5	String	String length
entity_condition_6	String	String length
entity_condition_7	String	String length
entity_condition_8	String	String length
entity_condition_9	String	String length
entity_condition_10	String	String length
entity_condition_11	String	String length
entity_condition_12	String	String length

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

entity_condition_13	String	String length
entity_condition_14	String	String length
entity_condition_15	String	String length
entity_condition_16	String	String length
entity_condition_17	String	String length
entity_condition_18	String	String length
entity_condition_19	String	String length
entity_condition_20	String	String length
number_of_record_axis_conditions	Integer	Mean, Median, Mode
record_condition_1	String	String length
record_condition_2	String	String length
record_condition_3	String	String length
record_condition_4	String	String length
record_condition_5	String	String length
record_condition_6	String	String length
record_condition_7	String	String length
record_condition_8	String	String length
record_condition_9	String	String length
record_condition_10	String	String length
record_condition_11	String	String length
record_condition_12	String	String length
record_condition_13	String	String length
record_condition_14	String	String length
record_condition_15	String	String length
record_condition_16	String	String length
record_condition_17	String	String length

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

record_condition_18	String	String length
record_condition_19	String	String length
record_condition_20	String	String length
race	Integer	Mean, Median, Mode
bridged_race_flag	Integer	Bool
race_imputation_flag	Integer	Bool
race_recode_3	Integer	Mean, Median, Mode
race_recode_5	Integer	Mean, Median, Mode
hispanic_origin	Integer	Mean, Median, Mode
hispanic_originrace_recode	Integer	Mean, Median, Mode

Add more rows to the table above as needed.

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

Part III: Second (2nd) additional data set: Cancer Rates by US State

A. Basic Descriptors

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

Variable	**Data Type**	**Basic mining procedure**
State	String	String Length
Range	String	String length, Mode
Rate	Integer	Average, Max, Min, Median, Mode

Add more rows to the table above as needed.

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 country had the the most deaths in females in 1998?

Q2: What was Michigan's cancer rate, according to the Cancer Rate by US State?

Q3: What was the most common cause death found in 2015 within the United States?

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

1. The first assumption that I am making at this point in the data is that my data is ready to be imported into Tableau or some other data visualization software and ready to create visualizations around. Also, I have been able to get the data ready to be imported at this pint

2. The second assumption that I'm making at this stage in the data visualization process is that I know how to tell what good data is like and what bad data is like. For instance, my second dataset is not ready to be imported and is a bad set of data since there are so many columns/fields to look at and there are so many codes that I would need to know to fully understand the data. Not only that, the columns or fields are very messy with naming conventions that it is hard to understand the data.

3. The final assumption I am making at this stage in the data visualization process is that data comes in many forms and sizes. And that the data can be mined in different ways from taking the average of a column to looking if the data is true or false. There are many different ways to interpret data that we have to make sure there are no biases to what we are interpreting or assuming about the data.