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: *Top Baby Names by State*

A. Basic Descriptors

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

Variable	**Data Type**	**Basic mining procedure**
Year	Integer	Average, max, min
Top name	String	String length
State	String	String length
Gender	Character	Mode
Occrence	Integer	Average, max, min

Add more rows to the table above as needed.

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

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-ratio-variables-why-should-i-care/

1, Year-Interval

2, Top name-Nominal

3, Sate-Nominal

4, Gender-Ordinal

5, Occrence-Ratio

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

C. Temporal

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

Yes, the data is temporal, and it is between 1910-2012

D. Range and Distribution

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

The distribution of the data is dense data; this is because all the data points concentrate on those popular names.

Part II: First (1st) additional data set: *name_gender_dataset*

A. Basic Descriptors

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

Variable	**Data Type**	**Basic mining procedure**
Name	String	String length
Gender	Character	Mode
Count	Integer	Average, max, min
Probability	Float	Average, max, min

Add more rows to the table above as needed.

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

Part III: Second (2nd) additional data set: *StateNames*

A. Basic Descriptors

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

Variable	**Data Type**	**Basic mining procedure**
ID	Integer	Average, max, min
Name	String	String length
Year	Integer	Average, max, min
Gender	Character	Mode
State	String	String length
Count	Integer	Average, max, min

Add more rows to the table above as needed.

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

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: Which baby names appear the most?

Q2: Which babies' name has the biggest probability to be given?

Q3: which states has the most popular name?

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

1. **Assumption #1 I assume that the maximum number of occurrences is the most popular name.**
2. **Assumption #2 I assume that the name that has the biggest probability to be given is the most popular name.**
3. **Assumption #3 I assume that the state that has a bigger population will have more chance to have the most popular name**

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