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: *EMSI_MillennialsvsBabyBoomers*

A. **Basic Descriptors**

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

Variable	**Data Type**	**Basic mining procedure**
Occupation	String	Length:
Generation	String	Length:
2007 Jobs	Integer	Min: **116,155** Max: **6,630,203** Average: **1,351,203**
2013 Jobs	Integer	Min: **124,627** Max: **6,389,681** Average: **1,395,889**
Job Change	Integer	Min: **-478,287** Max: **644,835** Average: **44,687**
Job Share of All Jobs 2007	Floating Point	Average: 0.21 Median: 0.19
Job Share of All Jobs 2013	Floating Point	Average: 0.21 Median: 0.19

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-ratio-variables-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

Occupation: Nominal

Generation: Ordinal

2007 Jobs: Ratio

2013 Jobs: Ratio

Job Share of All Jobs 2007: Ratio

Job Share of All Jobs 2013: Ratio

C. Temporal

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

Yes, this data is temporal. They data is collected between the year 2007 and 2013.

D. Range and Distribution

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

The distribution of this data is that is it's a small size (only 47 rows), noting that the data may seem to be sparse, with little data points before the mean and after the mean.

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

A. Basic Descriptors

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

Variable	**Data Type**	**Basic mining procedure**
Month	String	Length: 3
Year	Integer	Range: 2007 – 2013
Total Jobs Change	Integer	Median: 1,300 Mode:1,600 Average: 1,159
Private Sector Change	Integer	Median: 1,200 Mode: 1,600 Average: 967

Government Total Change	Integer	Median: 100 Mode:-100 Average: 142

Add more rows to the table above as needed.

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

A. Basic Descriptors

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

Variable	**Data Type**	**Basic mining procedure**
Month	String	Range: January - December
Year	String	Range: 2007 – 2019
Total Jobs	Integer	Mode:2,611,600 Median: 2,611,700 Average: 2,632,781
Private Sectors	Integer	Mode:2,262,000 Median: 2,129,150 Average: 2,132,714
Governments	Integer	Mode:146,100 Median: 144,350 Average: 140,729

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: What can the mode indicate in the datasets that relate to job change?

Q2: Is the data in all datasets are gathered in similar fashion. Are there gaps in gathering data such missing months, years, and etc.

Q3: Does the current calculation done from mining reveal a pattern?

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

1. **The negative values in some calculations may indicate an change in a particular variable.**
2. **An high average with a low min value may indicate the data is skewed.**
3. **An high average may indicate outliers in the data.**

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