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: *<Employment Changes in Great Britain by Industry >*

A. Basic Descriptors

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

Variable	**Data Type**	**Basic mining procedure**
City	String	String length
Country	String	String length
SIC-1	Character	Mode
SIC-1 name	String	String length
SIC-2	Integer	Max, min
Industry	String	String length
Jobs 2011	Integer	Average, max, min, mode, range

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

Jobs 2014	Integer	Average, max, min, mode, range
Change	Integer	Max, min, range, mode, average
Percent Change	Float/double	Max, min, range, mode, average

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

- Nominal
  - City
  - Country
  - SIC-1 name
  - Industry
- Ordinal
  - SIC-1
  - SIC-2
  - Change
  - Percent change
- Ratio
  - Jobs 2011
  - Jobs 2014

C. Temporal

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

I have variables, number of jobs in 2011, number of jobs in 2014, change, and percent change. The variables only record the number of jobs and not the actual time.

D. Range and Distribution

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

- Jobs in 2011
    - Max: 369,867
    - Min: 0
- Jobs in 2014
    - Max: 378,602
    - Min: 0
- Change
    - Max: 48,718
    - Min: -17,247

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

Part II: First (1st) additional data set: *<US Unemployment Rate by County 1990-2016t>*

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	Max, min
Month	String	String length
State	String	String length
County	String	String length
Rate	Double	Max, min, average, range

Add more rows to the table above as needed.

Part III: Second (2nd) additional data set: *<French employment, salaries, population per town>*

A. Basic Descriptors

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

Variable	Data Type	Basic mining procedure
CODGEO (geographic code)	Integer	Max, min
LIBGEO (town name)	String	String length
REG (region number)	Integer	Max, min
DEP (department number)	Integer	Max, min
E14TST (total number of firms)	Integer	Max, min, average
E14TS0ND (# unknown/null size firms)	Integer	Max, min, average, range
E14TS1 (# firms with 1-5 employees)	Integer	Max, min, average, range
E14TS6 (# firms with 6-9 employees)	Integer	Max, min, average, range
E14TS10 (# firms with 10-19 employees)	Integer	Max, min, average, range
E14TS20 (# firms with 20-49 employees)	Integer	Max, min, average, range
E14TS50 (# firms with 50-99 employees)	Integer	Max, min, average, range
E14TS100 (# firms with 100-199 employees)	Integer	Max, min, average, range
E14TS200 (# with 200-499 employees)	Integer	Max, min, average, range
E14TS500 (# firms with more than 500 employees)	Integer	Max, min, average, range

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 is the average employment rate by country in the UK, America, and France?

Q2: What is the change in North American and European employment rates in the last 10 years?

Q3: How does industry and location affect employment rates?

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

1. **Location has a vast effect on employment rates as the US, UK, and France have much different rates of employment**

2. **Employment rates over time can change a lot as seen in the UK employment rate changes between 2011 and 2014.**

3. Industry has a major impact on employment rates which is shown through number of firm employees in firms per town in France and Industry changes in the UK.