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.

1. **Part I: Tableau Data set:** The 2014 Inc. 5000

A. Basic Descriptors

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

Variable	**Data Type**	**Basic mining procedure**
_input	Void	
_num	Integer	Average = 2500.5 max = 5000 min = 1
_widgetName	String	String length
_source	String	String length
_resultNumber	Integer	Average = 2500.5 max = 5000 min = 1
_PageUrl	String	String length
_Hostname (eg .com)	String	String length
_Path name (eg .javacript, .json	string	String length
_id	Integer	Average =20036.575

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

		max =26620 min =4
_rank	Integer	Average = 2500.5 max = 5000 min = 1
_workers	Integer	Average =208.9698 max =34219 min = 0
_Company	Character	Character length
_url	character	Character length
_state_l	character	Character length
_state_s	character	Character length
_city	String	String length
_metro	character	Character length
_growth	float	Average =516.43989 max =158956.91 min =42.447
_revenue	Integer	Average =43058182.4 max =5528202691 min =1953000
_industry	character	Character length
_yrs_on_list	integer	Average =2.744 max =12 min =1

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/

Texutal

- Company, state_l, state_s, city, metro, industry

Ordinal

- _widgetName, _source, _pageUrl, Url

Interval

- _num
- _resultNumber
- _id
- _rank
- _workers
- _growth

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

- _revenue
- _yrs_on_list

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 only for the year 2014 from January through December of 2014.

D. Range and Distribution

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

- The dataset is large with 5000 rows.
- The data looks like it is evenly distributed but there are a couple outliers.
    - the range is between 1-5000 and the average of 2500.5

Part II: First (1st) additional data set: Business and Industry Reports

A. Basic Descriptors

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

Variable	Data Type	Basic mining procedure
Time_series_code	**String**	String length
Date	**Date**	Range length
Value	**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-ratio-variables-why-should-i-care/

Texutal

- Time_series_code

Interval

- Value, Date

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)?

This data is over serval years from 1992-2015.

D. Range and Distribution

What is the distribution of the data? Few values, small size, evenly spread, sparse or dense? Explain
This is a large set of datasets and is evenly spread from over time.

Part III: Second (2nd) additional data set: Stock Exchange data

A. Basic Descriptors

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

Variable	Data Type	Basic mining procedure
Index	**Character**	Average, max, min
date	**Date**	Range
open	**Float**	Average, max, min
high	**Float**	Average, max, min
low	**Float**	Average, max, min
close	**Float**	Average, max, min
adjclose	**Float**	Average, max, min
volume	**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-ratio-variables-why-should-i-care/

Texutal

- index

Interval

- Date
- Open
- High
- Low
- Close
- Adjclose
- Volume

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)?

It seems like the data is over several years.

D. Range and Distribution

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

The values are in a large scale of 112458 rows, and it is evenly spread because the large dataset made it possible for it to be even over the years.

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: Has the revenue increase over the year of 2014?

Q2: Did the value for the business report increase over several years?

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

Q3: what is the average open and close rate for the stocks?

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

1. **Assumption #1 m**
    a. **identifying outliers is important because in some of the data so it might affect some of the results for max, min, and average.**

2. **Assumption #2**
    a. **Missing dates might affect visualization process.**

3. **Assumption #3**
    a. **Some dataset's values are null, so it affects the how it is distributed.**