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: *U.S. Construction Spending, by value and category, 2002-16*

A. Basic Descriptors

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

Variable	**Data Type**	**Basic mining procedure**
per_idx	**Integer**	Average, max, min
per_name	**Date**	Chronological Order
cat_idx	**Integer**	Average, max, min
cat_code	**String**	String length
cat_desc	**String**	String length
cat_indent	**Boolean**	# of True vs False
dt_idx	**Integer**	Average, max, min
dt_code	**String**	String length
dt_desc	**String**	String length
dt_unit	**String**	String length
et_idx	**Integer**	Average, max, min
et_code	**String**	String length

et_desc	**String**	String length
et_unit	**String**	String length
geo_idx	**Integer**	Average, max, min
geo_code	**String**	String length
geo_desc	**String**	String length
is_adj	**Boolean**	# of True vs False
val	**Integer**	Average, max, min
serialid	**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/

- The main data featured, which is the millions of dollars spent monthly, would be ratio. Building Categories would be nominal, and date would be interval data.

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, it represents monthly data.

D. Range and Distribution

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

- The distribution is evenly spread, because several pieces of data are split into multiple rows; otherwise, it would be too dense.

Part II: First (1st) additional data set: PBNRESCONS (Public Non-Residential Construction Spending)

A. Basic Descriptors

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

Variable	Data Type	Basic mining procedure
realtime_start	**Date**	Chronological Order
value	**Integer**	Average, max, min
date	**Date**	Chronological Order
realtime_end	**Date**	Chronological Order

Add more rows to the table above as needed.

Part III: Second (2nd) additional data set: TLPRVCONS (Total Private Construction Spending)

A. Basic Descriptors

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

Variable	Data Type	Basic mining procedure
realtime_start	**Date**	Chronological Order
value	**Integer**	Average, max, min
date	**Date**	Chronological Order
realtime_end	**Date**	Chronological Order

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: At which point of the year is construction spending the highest?

Q2: Is the average monthly residential construction spending lower on average than the total construction spending and public construction spending?

Q3: Are construction categories the most expensive due to their building's function, or due to their quantity?

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

1. **Assumption #1**
    a. **I think the summer months will be the highest for monthly spending.**

2. **Assumption #2**
    a. **I think total residential construction cost would be less than public, due to public infrastructure receiving government subsidies and being funded by larger companies.**

3. **Assumption #3**
    a. **This is somewhat related to the last assumption; there are most likely more private buildings being built monthly than public ones, but the overall cost of each building project would most likely outweigh the quantity.**

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