CGT 270 Data Visualization
Module 1
Week 3
Lab 3: Mining Data
Name: Trami Nguyen

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: Wimbledon Champions

A. Basic Descriptors

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

Variable	**Data Type**	**Basic mining procedure**
Gender	String	String length
Champion	String	String length
Minutes	Integer	Average, maximum, minimum
Runner-up Nationality	String	String length
Champion Nationality	String	String length
Runner-up	String	String length
Score	String	String length
Runner-up seed	Character	Mode
Champion seed	Integer	Average, maximum, minimum

Year	Integer	Average, maximum, minimum
Runner-up Nationality	String	String length
Runner-up	String	String length

B. Categorize

Consider what variables are similar and what variables are different. This will help you to categorize the data. Are the data nominal, 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/

Nominal

- Gender, champion, runner-up nationality, champion nationality, runner-up, runner-up nationality, runner-up

Ordinal

- Runner-up seed, champion seed

Ratio

- Minutes, score, year

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 because it ranges over several years.

D. Range and Distribution

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

Minutes – this is sparse because there are many missing values for how long the matches took, most likely because they were not recorded and put out to the public.

Runner-up seed and champion seed- they are sparse as well because there are missing values.

Part II: First (1st) additional data set: Association of Tennis Professional Matches

A. Basic Descriptors

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

Variable	Data Type	Basic mining procedure
ID of Tournament	String	String length
Name of Tournament	String	String length
Surface of Court	String	String length
# of People in Tournament	Integer	Average, maximum, minimum
Tournament Level	Character	Mode
Start Date of Tournament	Integer	Average, maximum, minimum
Match Number	Integer	Average, maximum, minimum
Winner ID	Integer	Average, maximum, minimum
Seed of Winner	Integer	Average, maximum, minimum
Winner Entry	Character	Mode

Part III: Second (2nd) additional data set: Padel Tennis World Championship

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
Host (final location)	String	String length
Nationality	String	String length
Gold Medal Game	String	String length
Bronze Medal Game	String	String length

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 year had the most champion matches?

Q2: Were there leading (>1 win) champions over the years?

Q3: In what location/city are champion matches mostly held?

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

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

1. **Assumption #1: Most numerical values (integers) are mined with mode, median, and averages.**

2. **Assumption #2: String data types cannot be distributed.**

3. **Assumption #3: The data is accurate.**

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