Lab 2: Parsing Data

Name: Trami Nguyen

The goal of this lab is to understand the structure of data. In this lab you will change data into a format that tags each part of the data with its intended use. After completing this lab every element of the data, you selected (Tableau dataset) and the two (2) additional datasets you acquired in lab last week will be broken into its individual parts. Answer the following questions and complete the table for each dataset.

1. List the name of the Tableau Dataset you selected in the Acquire Lab: Wimbledon Champions
2. How many rows (records) are in the data set? 253 rows
3. How many columns (variables) are in the data set? 11 columns
4. What assumptions are you making about the data?
    a. There were no runner-ups for nearly half of the tournament matches
    b. The number of sets played per match ranged depending on the type of match/what tournament it was in
    c. Length of the matches ranged from short and long
    d. The length in minutes were from the start of the match to the end with no stops
    e. Not all runner-ups and champions had a seed number
    f. Many names/numbers have missing values in the chart, most likely because the information is not fully accessible to the public

What you should be able to do (at the end of this lab):

Remember	***Describe*** what happens in the **parse** stage.
Understand	***Describe*** the data in detail according to the parsing specifications.
Apply	***Demonstrate*** the ability to change data into a useful format for future processing.
Evaluate	***Categorize*** the data according to parsing specs.
Analysis	***Identify*** specific features about the data.
Create	***Generate*** a parsed listing of the data.

Tableau Data Set

In the table below list each variable and its data type (add more rows as needed):

	Variable	Data type
1	**Gender**	**String**
2	**Champion (full)**	**String**
3	**First Initial (champion)**	**Character**
4	**Middle Initial (champion)**	**Character**
5	**Last Name (champion)**	**String**
6	**Minutes**	**Integer**

7	**Runner-up Nationality**	**String**
8	**Champion Nationality**	**String**
9	**Runner-up**	**String**
10	**First Initial (runner-up)**	**Character**
11	**Middle Initial (runner-up)**	**Character**
12	**Last Name (runner-up)**	**String**
13	**Score**	**String**
14	**Set 1 Score**	**String**
15	**Set 2 Score**	**String**
16	**Set 3 Score**	**String**
17	**Set 4 Score**	**String**
18	**Set 5 Score**	**String**
19	**Runner-up Seed**	**Character**
20	**Champion Seed**	**Integer**
21	**Year**	**Integer**
22	**Runner-up Nationality (Men's)**	**String**
23	**Runner-up**	**String**
24	**First Initial (runner-up)**	**Character**
25	**Middle Initial (runner-up)**	**Character**
26	**Last Name (runner-up)**	**String**

You may add more rows and attach additional pages if needed.

Save this file as: **LastnameFirstInitial-CGT270Fall2021-Lab2Parsing.pdf**

Additional Data Set #1

1. List the name of the first (1st) additional data set you acquired in the Acquire Lab: Association of Tennis Professional Matches (Kaggle)
2. How many rows (records) are in the data set? 75 rows
3. How many columns (variables) are in the data set? 10 columns
4. What assumptions are you making about the data?
   a. The date of the tournament is in order of year, month, and day, but there are no slashes
   b. Not all winners had seed numbers, therefore were not listed in
   c. The reason for entrance of the winners to the tournament are unclear/not accessible

In the table below list each variable and its data type (add more rows as needed):

	Variable	Data type
1	**ID of Tournament**	**String**
2	**Name of the Tournament**	**String**
3	**Surface of the Court**	**String**
4	**Number of People in the Tournament**	**Integer**
5	**Tournament Level**	**Character**
6	**Start Date of Tournament**	**Integer**
7	**Year (date)**	**Integer**
8	**Month (date)**	**Integer**
9	**Day (date)**	**Integer**
10	**Match Number**	**Integer**
11	**Winner ID**	**Integer**
12	**Seed of Winner**	**Integer**
13	**Winner Entry**	**Character**
14		

You may add more rows and attach additional pages if needed.

Additional Data Set #2

1. List the name of the second (2nd) additional data set you acquired in the Acquire Lab: Padel Tennis World Championship
2. How many rows (records) are in the data set? 14 rows
3. How many columns (variables) are in the data set? 5 columns
4. What assumptions are you making about the data?
   a. The name of the variable for column 3 is "nationality" since there is no label or name in that slot

        b.   The tennis world championship occurs only every two years
        c.   There are no exact host locations, but most are in Latin America
        d.   Not all gold medal and bronze medal games were played/recorded

In the table below list each variable and its data type (add more rows as needed):

	Variable	Data type
1	**Year**	**Integer**
2	**Host (final location)**	**String**
3	**Host (location)**	**String**
4	**City (location)**	**String**
5	**Nationality**	**String**
6	**Gold Medal Game**	**String**
7	**Bronze Medal Game**	**String**
8		

You may add more rows and attach additional pages if needed.