Assignment:

Data Preparation and Descriptive Analytics

This assignment should be completed individually.

You will complete this assignment using Tableau Prep and Excel. Tableau Prep is a data cleaning and preparation tool, whereas Tableau Desktop, which we will start using in a couple of weeks, is a data visualization tool. Please follow the directions below for downloading and installing the two Tableau products. Note that the product key provided below is exclusively for our class, so DO NOT share it with anyone as there are only enough licenses for our class.

1. Download the latest version of Tableau Desktop and Tableau Prep Builder here
2. Click on the link above and select "Download Tableau Desktop" and "Download Tableau Prep Builder". On the form, enter your school email address for Business E-mail and enter the name of your school for Organization.
3. Activate with your product key: TC33-8D5E-62E0-EAB0-517B
4. Already have a copy of Tableau Desktop installed? Update your license in the application: Help menu → Manage Product Keys

Once you have installed Tableau Prep, you can watch the training videos prepared and published by Tableau to learn how to use the software. Here is the list of the videos:

https://www.tableau.com/learn/training/20204

Here are the videos that you will watch for this assignment:

- Tableau Prep
  - Getting Started with Tableau Prep
  - The Tableau Prep Builder Interface
  - The Input Step
  - The Cleaning Step
  - Group and Replace
  - The Aggregate Step
  - The Join Step
  - The Output Step

Also, Chapter 10 of the recommended textbook (Visual Analytics with Tableau) provides information on how to use Tableau Prep for data preparation. While reading that chapter is helpful, it is not required.

Use the datasets named "beers.xlsx" and "breweries.xlsx" uploaded on Canvas to perform each step and answer the following questions. Your goal is to clean and consolidate the two datasets to be used to analyze different characteristics of beers such as ABV (Alcohol by Volume), IBU (international Bitterness Units) and drink size. Datasets for this assignment were modified from sets available on Kaggle.

1) Open and review the two datasets. What are the quality issues with the beer and brewery data? Explain how you identified the issues.
2) How many attributes/fields does each table include? Is there any attribute common between the two tables?

3) Do you think you will need to change the structure of the data (e.g., split any fields or combine fields) to enable future analysis?

4) Using Tableau Prep Builder clean and aggregate the datasets as instructed below and generate a .xlsx file as an output. Explain each data cleaning and aggregation step and provide a final screenshot of the Tableau Prep Builder interface with the flow chart and profile pane visible on it to enable me to understand how you have done it. Here is the data preparation process:

   a) Connect to both datasets and add a clean step to each. Explore the data using the profile pane. Identify duplicate field names between the two sets and rename to avoid confusion once both sets are joined.

   b) Split the column *location* into two columns (*City* and *State*) from "breweries". Perform any grouping as required to correct data entry errors within the state field.

   c) Check each field to ensure correct data roles and types (*City* and *State* should be geographic, any IDs should be string).

   d) Review the "beers" dataset to identify and correct errors in data entry (Hint: Look for outliers in the *ABV* field and use manual grouping to correct these values).

   e) Join the two datasets using the most appropriate fields for the join operation. Which field did you use to join the tables? How may performing the join operation be useful?

5) During join or in data entry, data is mistakenly duplicated. Address the duplicate beer issue in the new dataset using the Aggregate function to ensure each beer is only reported once. (Hint: Check the "Aggregated Fields" section of this step - the default action for "number" types is "SUM." Is this right? What is a better measure when removing duplicates?) To learn more about ways to remove duplicates see here and here. Also, to learn more about the aggregate function see here.

6) Were there any duplicates in the dataset? What beers and how many times were they duplicated? What measure did you use to maintain data integrity during this step?

7) Clean and organize the combined and unduplicated dataset and save the output to a .xlsx file.

8) How many beers, beer styles, and breweries are listed in the final dataset?

9) What is the average IBU across all the beers? Using Excel charts and/or tables show how the average ABV and average IBU are different among the following beer styles: American IPA, Hefeweizen, Oatmeal Stout, and Wheat Ale. Briefly explain the results.

10) Is there any correlation between ABV and IBU? Explain in a sentence or two.

11) Did you perform any *listwise or pairwise deletion* during the data preparation process? How would performing each of these techniques impact your answers to questions 8 and 9? (You do not need to perform both; whichever you have done, explain how using the other one would change the results.)

What to submit: You will turn in three files:

1. A Word document containing answers to the questions with charts and tables, as needed. A screenshot should also be provided in response to question 4.
2. A copy of the Tableau packaged flow file (.tflx) that shows your work in Tableau Prep. Do not submit the .tfl file.
3. A .xlsx file that contains the cleaned dataset (output of Tableau Prep).