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: Airbnb listings in New York

A. Basic Descriptors

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

Variable	**Data Type**	**Basic mining procedure**
Host ID	Integer	Mode
Host Since Month	String	String length
Host Since Day	Integer	Max, min, mode
Host Since Year	Integer	Max, min, mode
Name	String	String length
Neighborhood	String	String length
Property Type	String	String length
Review Scores Rating (bin)	Integer	Max, min, avg
Room Type	String	String length
Zip code	Integer	Mode
Beds	Integer	Max, min, mode, avg
Number of Records	Integer	Max, min, avg
Number of Reviews	Integer	Max, min, avg
Price	Integer	Max, min, avg

Review Score Rating	Integer	Max, min, avg

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 data types that are similar are name, neighborhood, property type, room type, and host since month. These data are all nominal. Host ID, host since day, host since year, and zip code are all similar because they are ordinal. On the other hand, review scores rating, beds, number of records, number of reviews, review score rating, and price are similar because they are all ratio.

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)?

Host Since month, day, and year represent time over several years, months, and days, so the data is temporal.

D. Range and Distribution

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

The distribution of Host ID is evenly spread. The data seems to be spread evenly on both sides of the mean. The distribution of Host Since Day is dense because there's a lot of different number of dates listed. The distribution of Host Since Year is evenly spread because the spread of a few different years is even. The distribution of the Review Scores Rating is evenly spread because the data seems to be evenly spread on both sides of the mean. The distribution of zip code is dense because there is a lot of different zip codes. The distribution of beds is evenly spread because the spread of 3 different values is even. The distribution of number of reviews is evenly spread because there are a bunch of different values that seem to be even on both the high side and low side. The distribution of price seems to be evenly spread because the data is spread evenly on both sides of the mean.

Part II: First (1st) additional data set: Airbnb listings in Los Angeles

A. Basic Descriptors

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

Variable	Data Type	Basic mining procedure
ID	Integer	Mode
Name	String	String length
Host ID	Integer	Mode
Host Name	String	String length
Neighborhood	Integer	Mode
Latitude	Floating point	Max, min
Longitude	Floating point	Max, min
Room Type	String	String length
Price	Integer	Max, min, avg
Minimum nights	Integer	Max, min, avg
Number of Reviews	Integer	Max, min, avg.
Last Review Month	String	String length
Last Review Day	Integer	Mode
Last Review Year	Integer	Mode
Reviews per Month	Floating point	Max, min
Calculating host listings count	Integer	Max, min, avg
Availability	Integer	Max, min, avg
City	String	String length

Add more rows to the table above as needed.

Part III: Second (2nd) additional data set: Hotel Booking Demand

A. Basic Descriptors

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

Variable	Data Type	Basic mining procedure
Hotel	String	String length
Is Cancelled	Integer	Max, min, avg
Lead Time	Integer	Max, min, avg
Arrival Date Year	Integer	Mode
Arrival Date Month	String	String length
Stays in weekend Nights	Integer	Max, min, avg
Stays in Weeks Nights	Integer	Max, min, avg
Adults	Integer	Max, min, avg
Children	Integer	Max, min, avg
Babies	Integer	Max, min, avg
Meal	String	String length
Country	String	String length
Market Segment	String	String length
Distribution Channel	String	String length
Is repeated guest	Integer	Max, min, avg
Previous Cancellations	Integer	Max, min, avg
Previous bookings not cancelled	Integer	Max, min, avg
Booking Changes	Integer	Max, min, avg
Deposit Types	String	String length
Agent	Integer	Max, min, avg
Company	Integer	Max, min, avg
Days in waiting list	Integer	Max, min, avg
Customer Type	String	String length

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: Do all the datasets have temporal data?

Q2: For the integers, are most of the data evenly spread?

Q3: Is the data mostly ordinal or ratio?

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

1. **Assumption #1**
   I assume that the airbnbs with more rooms are higher priced
2. **Assumption #2**
   I assume that location, especially downtown areas, has an influence on price
3. **Assumption #3**
   I assume that the number of stays over weekend nights are more than week days

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