I Have a Dataset! Now What?

Catherine R. Barber

February 22, 2022

Session Goal

In this session, you will learn to use four questions to evaluate and understand a dataset:

- What is the integrity of the dataset?
- What type of data does the dataset contain?
- What can the data tell me?
- What techniques can I use to analyze the data?

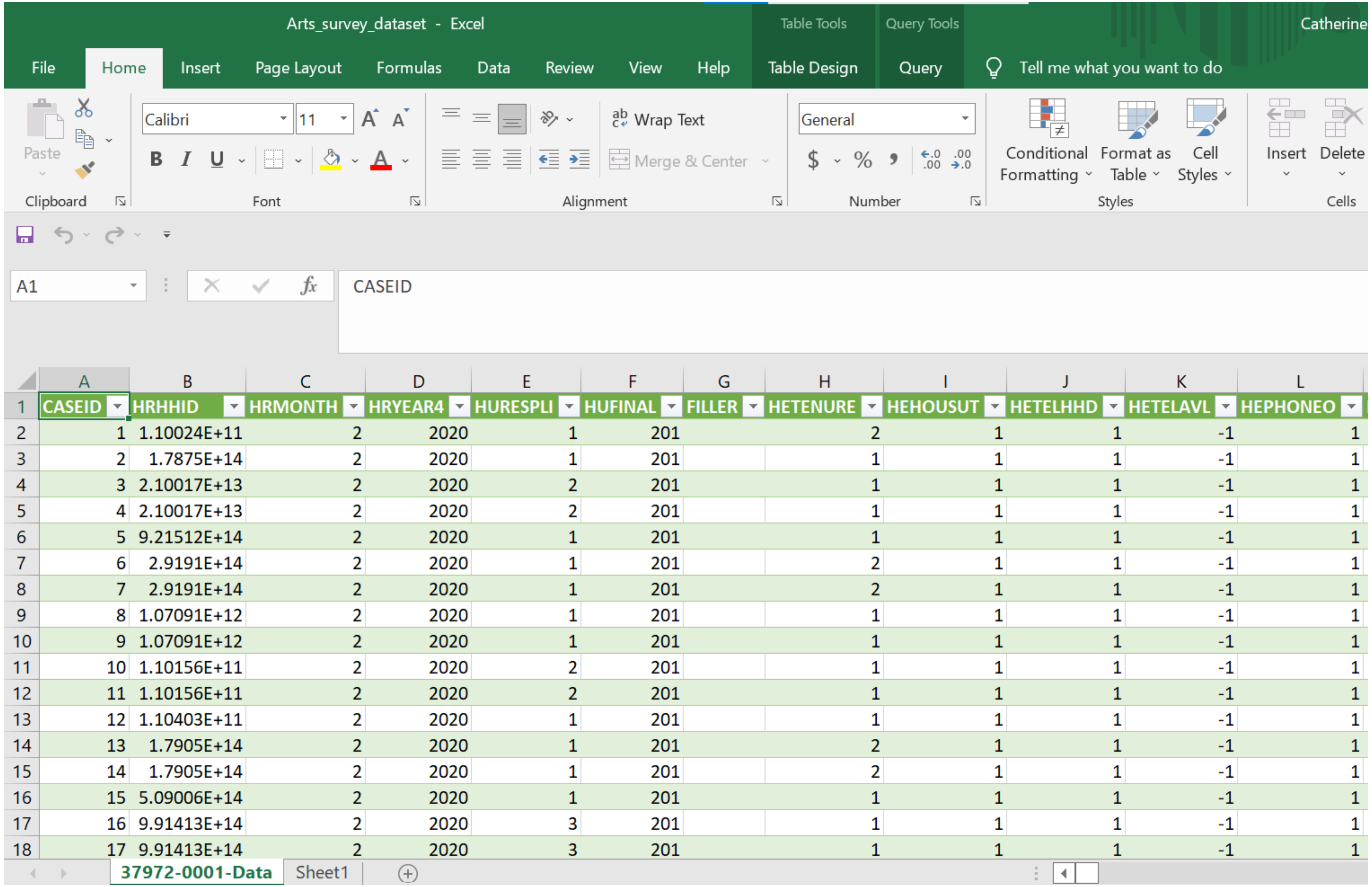
To Begin: A Few Terms

- Data: pieces of information
- Variable: factor of interest that can vary
    - Examples: age; temperature; test score; extent of agreement
- Value: score, code, or other content that represents position on the variable
    - Examples: 46 years old; 50 degrees Fahrenheit; 85%; Strongly disagree

Dataset Example

- 2020 Arts Basic Survey – U.S. Bureau of the Census for the National Endowment for the Arts
  - Source: National Endowment for the Arts, United States. Bureau of the Census, and United States. Bureau of Labor Statistics. Arts Basic Survey, United States, 2020. Inter-university Consortium for Political and Social Research [distributor], 2021-05-03. https://doi.org/10.3886/ICPSR37972.v1
- Conducted in February, 2020
- Approximately 35,000 survey respondents; half responded to the arts questions
- Exporting and importing data – see video link in chat!

Sample dataset

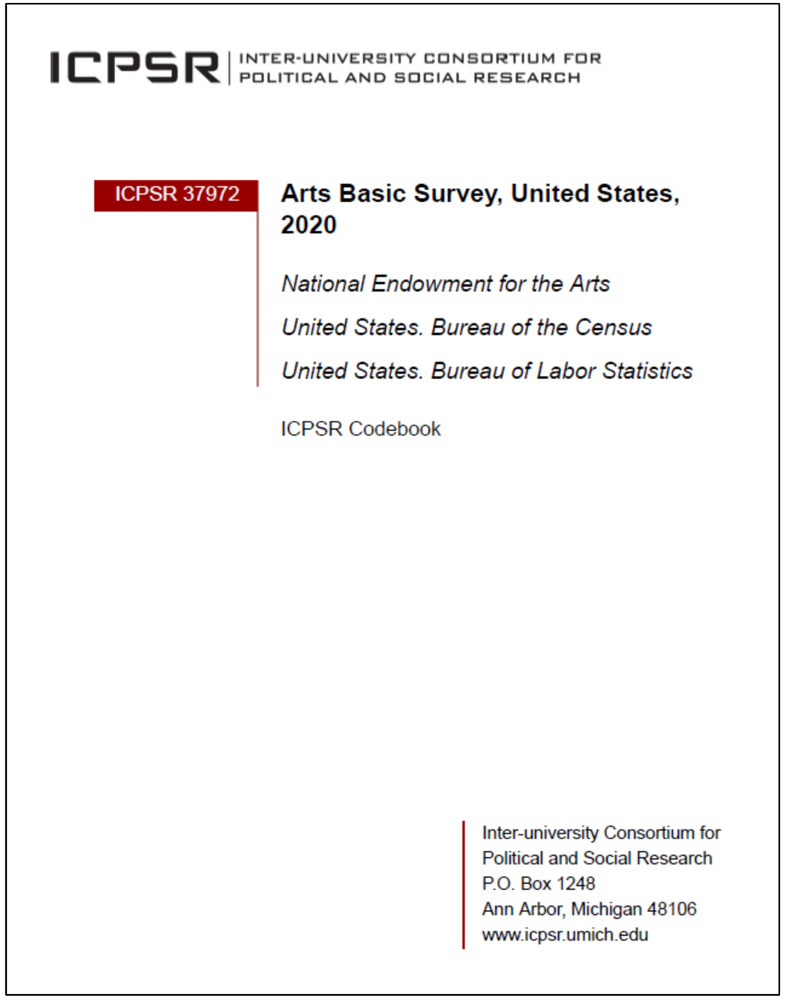


A1 CASEID

CASEID	HRHHID	HRMONTH	HRYEAR4	HURESPLI	HUFINAL	FILLER	HETENURE	HEHOUSUT	HETELHHD	HETELAVL	HEPHONEO
1	1.10024E+11	2	2020	1	201		2	1	1	-1	1
2	1.7875E+14	2	2020	1	201		1	1	1	-1	1
3	2.10017E+13	2	2020	2	201		1	1	1	-1	1
4	2.10017E+13	2	2020	2	201		1	1	1	-1	1
5	9.21512E+14	2	2020	1	201		1	1	1	-1	1
6	2.9191E+14	2	2020	1	201		2	1	1	-1	1
7	2.9191E+14	2	2020	1	201		2	1	1	-1	1
8	1.07091E+12	2	2020	1	201		1	1	1	-1	1
9	1.07091E+12	2	2020	1	201		1	1	1	-1	1
10	1.10156E+11	2	2020	2	201		1	1	1	-1	1
11	1.10156E+11	2	2020	2	201		1	1	1	-1	1
12	1.10403E+11	2	2020	1	201		1	1	1	-1	1
13	1.7905E+14	2	2020	1	201		2	1	1	-1	1
14	1.7905E+14	2	2020	1	201		2	1	1	-1	1
15	5.09006E+14	2	2020	1	201		1	1	1	-1	1
16	9.91413E+14	2	2020	3	201		1	1	1	-1	1
17	9.91413E+14	2	2020	3	201		1	1	1	-1	1

Question 1: What is the integrity of the dataset?

- Source
- Sample
- Method of data collection
- Reliability and validity of measures



ICPSR | INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

ICPSR 37972

Arts Basic Survey, United States, 2020

National Endowment for the Arts

United States. Bureau of the Census

United States. Bureau of Labor Statistics

ICPSR Codebook

Inter-university Consortium for
Political and Social Research
P.O. Box 1248
Ann Arbor, Michigan 48106
www.icpsr.umich.edu

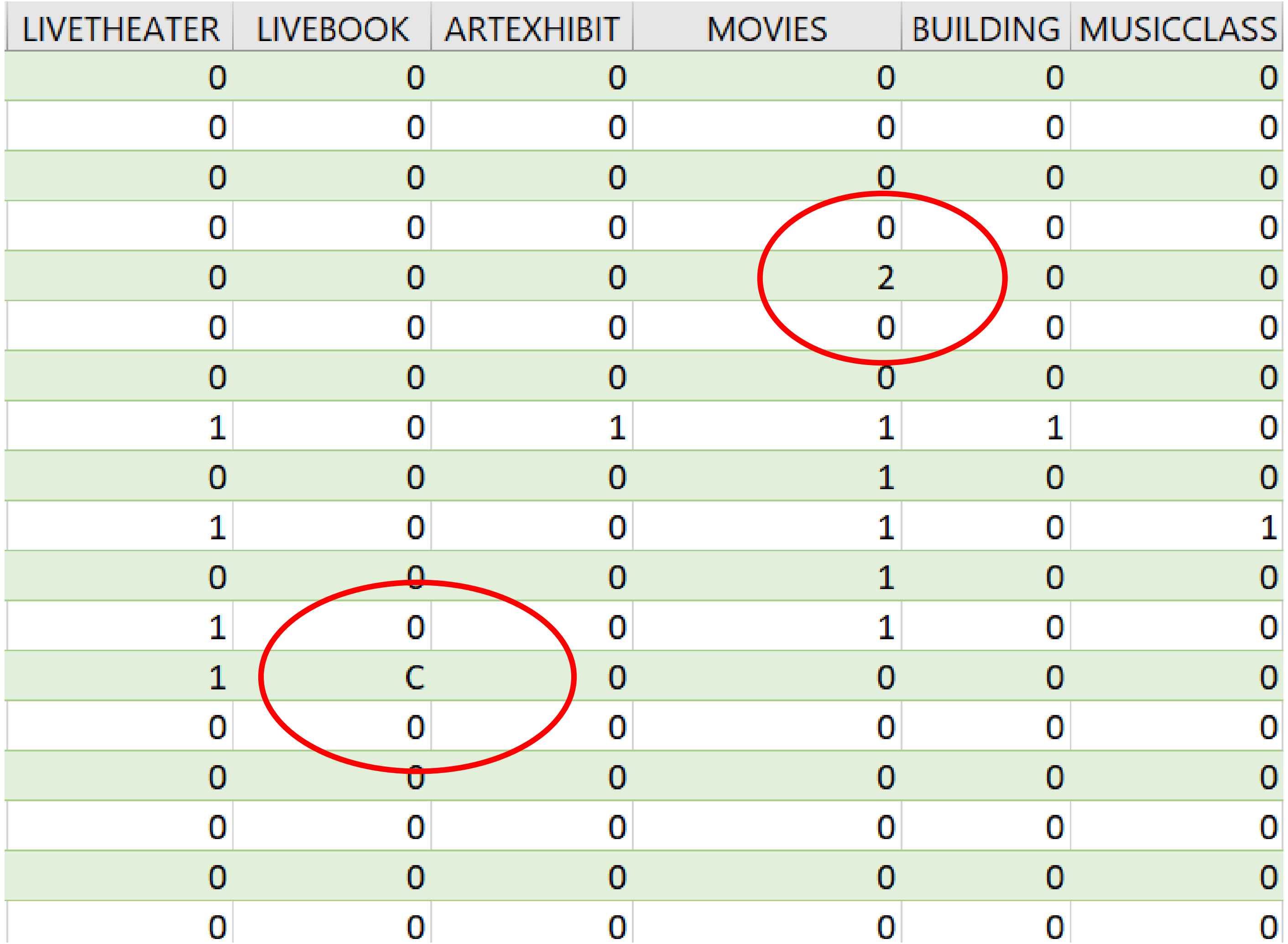
Missing Data and Cleaning the Dataset

Why are data missing? Is there a pattern?

How much cleaning is needed?

- Errors, blank rows, duplicates, outliers
- Formatting

Cleaning the Dataset

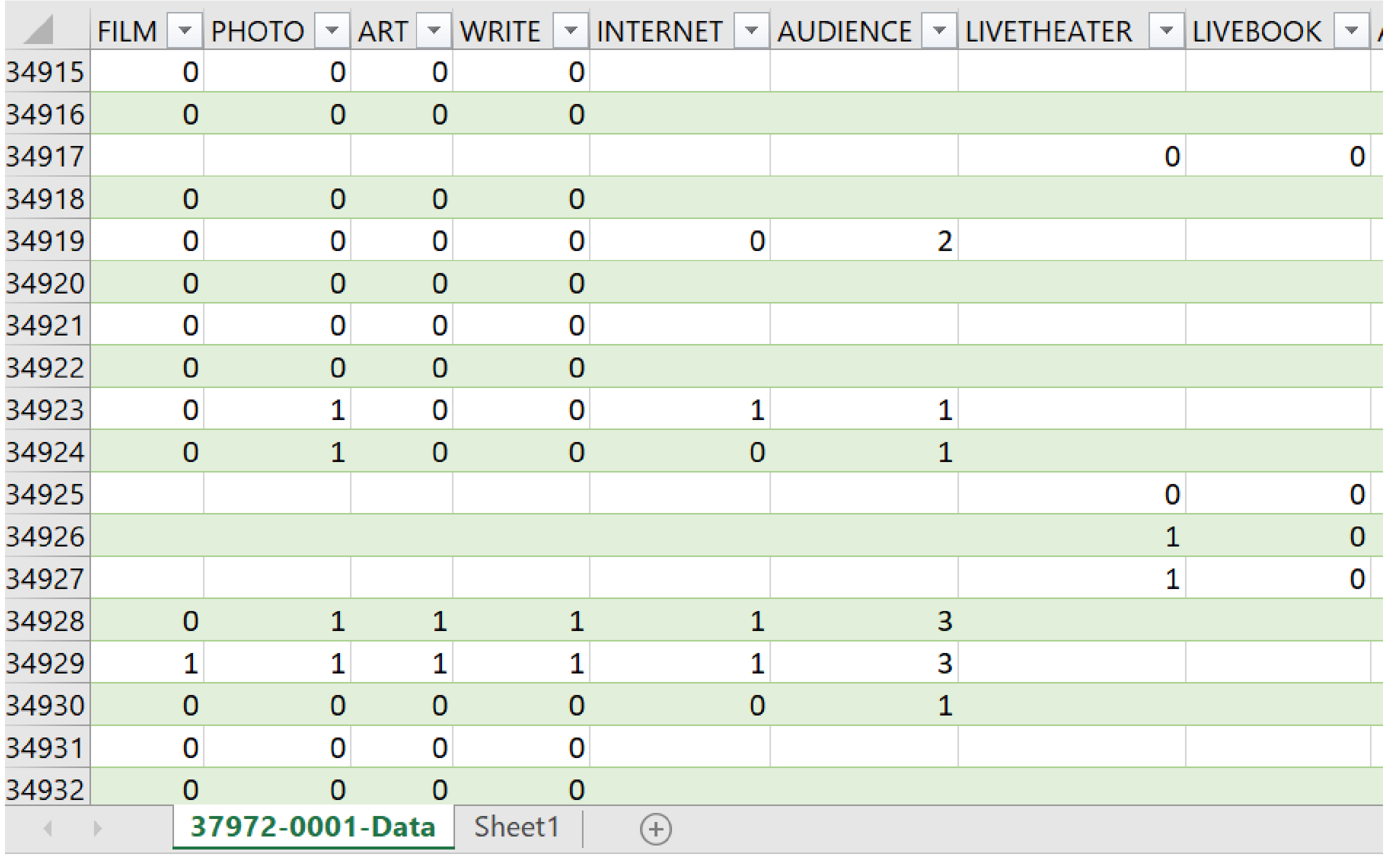


LIVETHEATER	LIVEBOOK	ARTEXHIBIT	MOVIES	BUILDING	MUSICCLASS
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	2	0	0
0	0	0	0	0	0
0	0	0	0	0	0
1	0	1	1	1	0
0	0	0	1	0	0
1	0	0	1	0	1
0	0	0	1	0	0
1	0	0	1	0	0
1	C	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

Question 2: What type of data does the dataset contain?

- Qualitative – text/visual
- Quantitative – numeric
- Mixed – both

Quantitative Data



	FILM	PHOTO	ART	WRITE	INTERNET	AUDIENCE	LIVETHEATER	LIVEBOOK
34915	0	0	0	0				
34916	0	0	0	0				
34917							0	0
34918	0	0	0	0				
34919	0	0	0	0	0	2		
34920	0	0	0	0				
34921	0	0	0	0				
34922	0	0	0	0				
34923	0	1	0	0	1	1		
34924	0	1	0	0	0	1		
34925							0	0
34926							1	0
34927							1	0
34928	0	1	1	1	1	3		
34929	1	1	1	1	1	3		
34930	0	0	0	0	0	1		
34931	0	0	0	0				
34932	0	0	0	0				

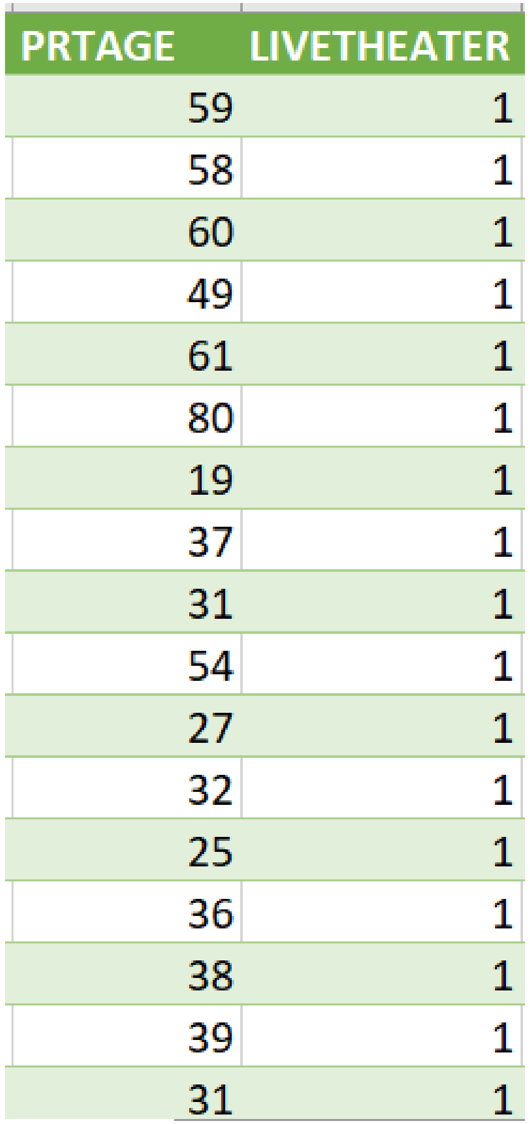
- Categorical variables: dichotomies or discrete groups
- Continuous variables: measurable

Question 3: What can the data tell me?

Qualitative: experiences, stories, themes

Quantitative:

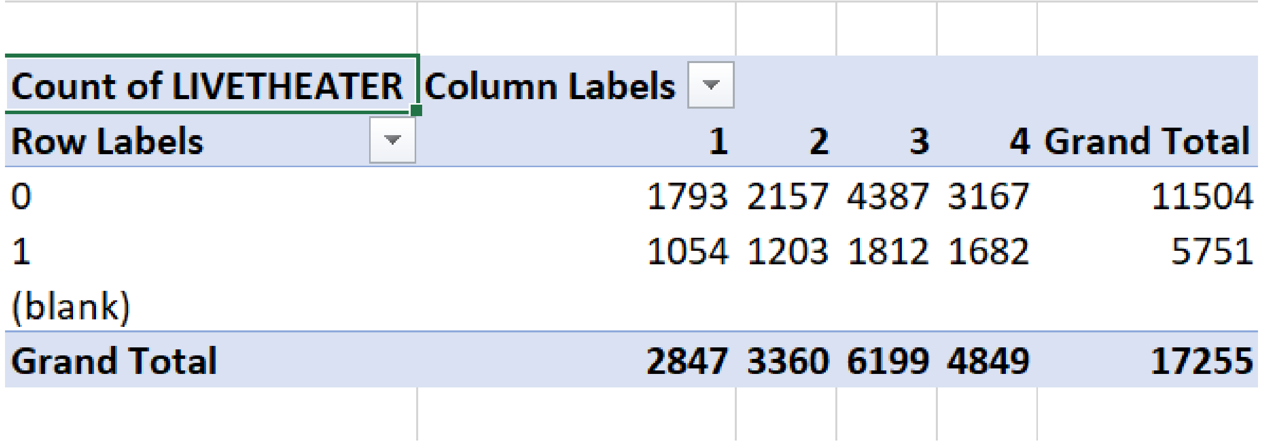
- Central tendency and variability (see image)
- Frequency/percentage
- Relationship



PRTAGE	LIVETHEATER
59	1
58	1
60	1
49	1
61	1
80	1
19	1
37	1
31	1
54	1
27	1
32	1
25	1
36	1
38	1
39	1
31	1

Question 4: What techniques can I use to analyze the data?

- Qualitative: coding, thematic analysis
- Quantitative:
  - Data visualization
  - Descriptive statistics
  - Inferential statistics



Count of LIVETHEATER	Column Labels				
Row Labels	1	2	3	4	Grand Total
0	1793	2157	4387	3167	11504
1	1054	1203	1812	1682	5751
(blank)					
Grand Total	**2847**	**3360**	**6199**	**4849**	**17255**

Quiz!

- Type your responses in the chat, but wait to post your response!
  - Evaluating a dataset begins with examining what aspect?
  - What are two main types of quantitative variables?
  - **True or False:** With qualitative data, research questions focus on frequency, central tendency, and comparison and relationship.
  - What is typically the first step in analyzing quantitative data?

Quiz Answers

- Evaluating a dataset begins with examining what aspect?
  - The integrity of the dataset
- What are two main types of quantitative variables?
  - Categorical and continuous
- True or False: With qualitative data, research questions focus on frequency, central tendency, and relationship.
  - False: This is true of quantitative data.
- What is typically the first step in analyzing quantitative data?
  - Data visualization

Wrap-up and Q&A

Four questions for understanding and evaluating a dataset:

- What is the integrity of the dataset?
- What type of data does the dataset contain?
- What can the data tell me?
- What techniques can I use to analyze the data?

Questions?

Thank you!