Cleaning & Organizing Data

This project is made possible by a grant from the U.S. Institute of Museum and Library Services.

Making Data Useful

Necessity of Data Cleanup

Getting to Know Your Data

Splitting & Concatenating Data

Filtering & Removing Duplicates

Creating & Using Lookup Tables

Data Cleaning & Organizing

Cleaning and organizing your dataset is a critical step that should be conducted before analyzing and visualizing.

Data cleaning involves …

- Fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

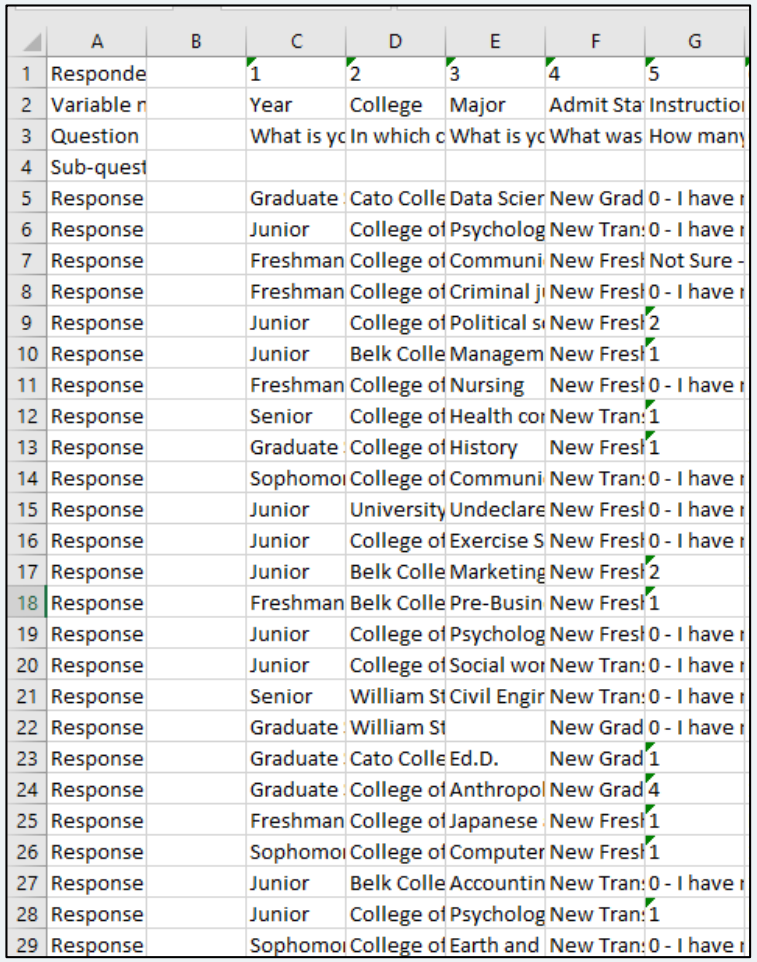
Data organizing often involves …

- Connecting two or more datasets
- Creating new categories to combine like items
- Turning words into numbers
- Modifying the date and time format
- Combining data from two columns
- Splicing data from one column into 2+

"Clean data will get you from numbers to insights much more quickly."
Source: https://www.grantbook.org/blog/how-to-clean-your-data-hygiene-best-practices

From Messy Data …

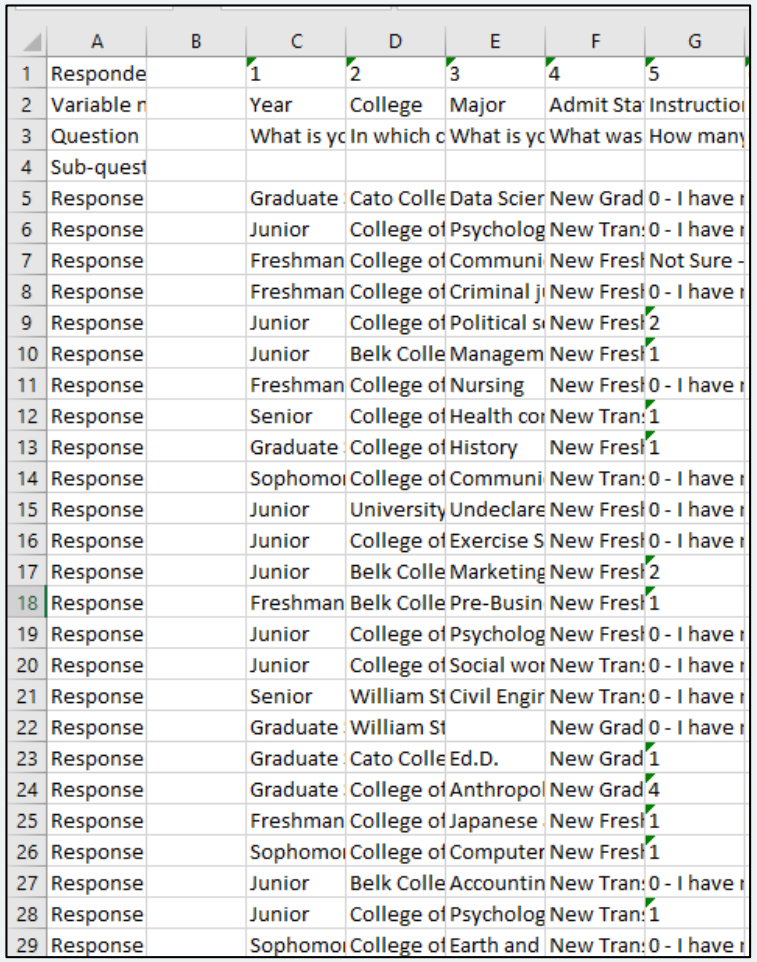
Raw Data



	A	B	C	D	E	F	G
1	Responde		1	2	3	4	5
2	Variable r		Year	College	Major	Admit Sta	Instructio
3	Question		What is yc	In which c	What is yc	What was	How many
4	Sub-quest						
5	Response		Graduate	Cato Colle	Data Scier	New Grad	0 - I have r
6	Response		Junior	College of	Psycholog	New Tran	0 - I have r
7	Response		Freshman	College of	Communi	New Fresl	Not Sure -
8	Response		Freshman	College of	Criminal j	New Fresl	0 - I have r
9	Response		Junior	College of	Political s	New Fresl	2
10	Response		Junior	Belk Colle	Managem	New Fresl	1
11	Response		Freshman	College of	Nursing	New Fresl	0 - I have r
12	Response		Senior	College of	Health co	New Tran	1
13	Response		Graduate	College of	History	New Fresl	1
14	Response		Sophomo	College of	Communi	New Tran	0 - I have r
15	Response		Junior	University	Undeclare	New Fresl	0 - I have r
16	Response		Junior	College of	Exercise S	New Fresl	0 - I have r
17	Response		Junior	Belk Colle	Marketing	New Fresl	2
18	Response		Freshman	Belk Colle	Pre-Busin	New Fresl	1
19	Response		Junior	College of	Psycholog	New Fresl	0 - I have r
20	Response		Junior	College of	Social wo	New Tran	0 - I have r
21	Response		Senior	William St	Civil Engi	New Tran	0 - I have r
22	Response		Graduate	William St		New Grad	0 - I have r
23	Response		Graduate	Cato Colle	Ed.D.	New Grad	1
24	Response		Graduate	College of	Anthropol	New Grad	4
25	Response		Freshman	College of	Japanese	New Fresl	1
26	Response		Sophomo	College of	Computer	New Fresl	1
27	Response		Junior	Belk Colle	Accountin	New Tran	0 - I have r
28	Response		Junior	College of	Psycholog	New Tran	1
29	Response		Sophomo	College of	Earth and	New Tran	0 - I have r

From Messy Data … to Clean

Raw Data

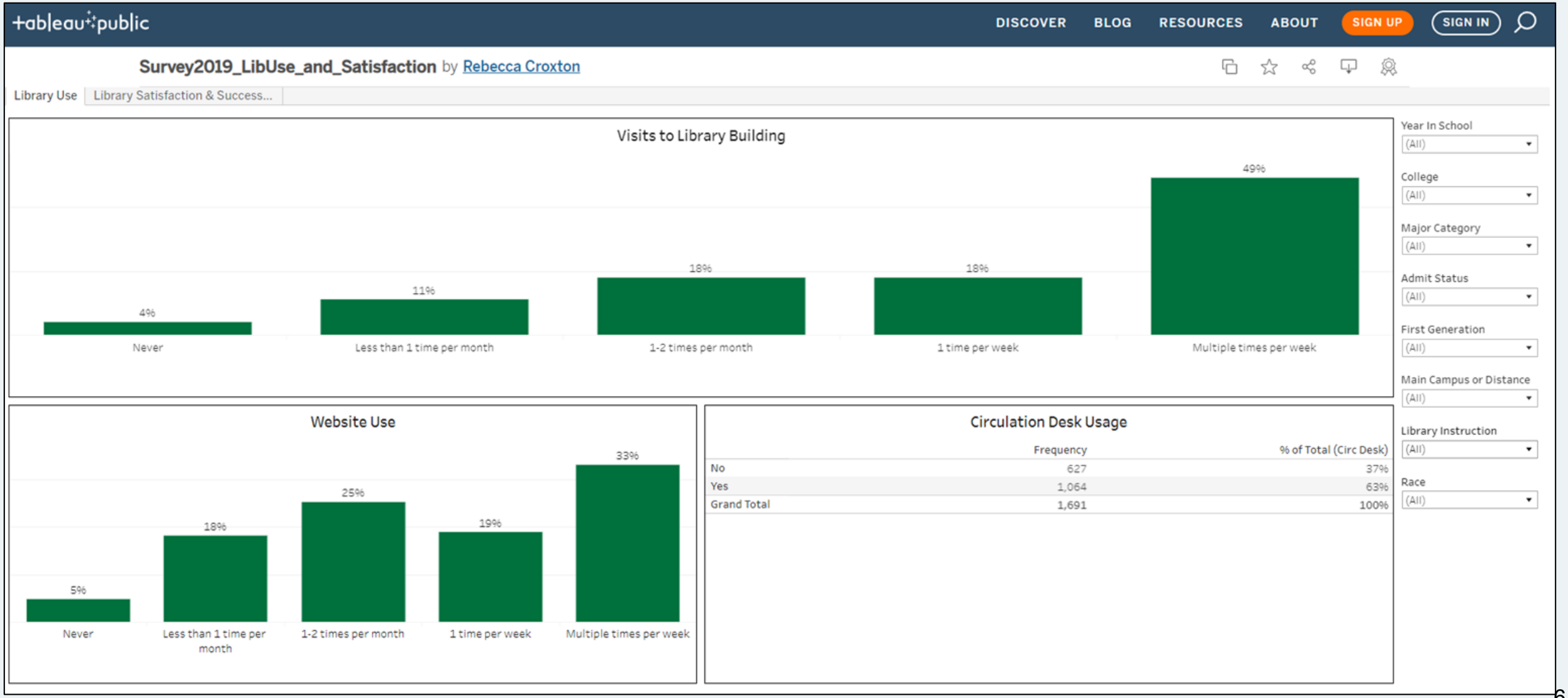


	A	B	C	D	E	F	G
1	Responde		1	2	3	4	5
2	Variable r		Year	College	Major	Admit Sta	Instructio
3	Question		What is yc	In which c	What is yc	What was	How many
4	Sub-quest						
5	Response		Graduate	Cato Colle	Data Scier	New Grad	0 - I have r
6	Response		Junior	College of	Psycholog	New Tran:	0 - I have r
7	Response		Freshman	College of	Communi	New Fresl	Not Sure -
8	Response		Freshman	College of	Criminal j	New Fresl	0 - I have r
9	Response		Junior	College of	Political s	New Fresl	2
10	Response		Junior	Belk Colle	Managem	New Fresl	1
11	Response		Freshman	College of	Nursing	New Fresl	0 - I have r
12	Response		Senior	College of	Health co	New Tran:	1
13	Response		Graduate	College of	History	New Fresl	1
14	Response		Sophomo	College of	Communi	New Tran:	0 - I have r
15	Response		Junior	University	Undeclare	New Fresl	0 - I have r
16	Response		Junior	College of	Exercise S	New Fresl	0 - I have r
17	Response		Junior	Belk Colle	Marketing	New Fresl	2
18	Response		Freshman	Belk Colle	Pre-Busin	New Fresl	1
19	Response		Junior	College of	Psycholog	New Fresl	0 - I have r
20	Response		Junior	College of	Social wor	New Tran:	0 - I have r
21	Response		Senior	William St	Civil Engir	New Tran:	0 - I have r
22	Response		Graduate	William St		New Grad	0 - I have r
23	Response		Graduate	Cato Colle	Ed.D.	New Grad	1
24	Response		Graduate	College of	Anthropol	New Grad	4
25	Response		Freshman	College of	Japanese	New Fresl	1
26	Response		Sophomo	College of	Computer	New Fresl	1
27	Response		Junior	Belk Colle	Accountin	New Tran:	0 - I have r
28	Response		Junior	College of	Psycholog	New Tran:	1
29	Response		Sophomo	College of	Earth and	New Tran:	0 - I have r

Cleaned Data

	A	B	C	D	E	F
1	ID	YearInSchool	College	MajorCategory	4_AdmitStatus_Adjuste	5_Instruction
2	1	Graduate Student	Cato College of Education	ITCS	New Graduate Student	0
3	2	Junior	College of Liberal Arts and Sciences	PSYC	New Transfer Student	0
4	3	Freshman	College of Liberal Arts and Sciences	COMS	New Freshman	1
5	4	Freshman	College of Liberal Arts and Sciences	CJUS	New Freshman	0
6	5	Junior	College of Liberal Arts and Sciences	POLS	New Freshman	2
7	6	Junior	Belk College of Business	MGMT	New Freshman	1
8	7	Freshman	College of Health and Human Services	NURS	New Freshman	0
9	8	Senior	College of Liberal Arts and Sciences	COMS	New Transfer Student	1
10	9	Graduate Student	College of Liberal Arts and Sciences	HIST	New Freshman	1
11	10	Sophomore	College of Liberal Arts and Sciences	COMS	New Transfer Student	0
12	11	Junior	University College	UCOL	New Freshman	0
13	12	Junior	College of Health and Human Services	KINE	New Freshman	0
14	13	Junior	Belk College of Business	MKTG	New Freshman	2
15	14	Freshman	Belk College of Business	BUSN	New Freshman	1
16	15	Junior	College of Liberal Arts and Sciences	PSYC	New Freshman	0
17	16	Junior	College of Health and Human Services	SOWK	New Transfer Student	0
18	17	Senior	William States Lee College of Engineering	ENGR	New Transfer Student	0
19	18	Graduate Student	William States Lee College of Engineering	ENGR	New Graduate Student	0
20	19	Graduate Student	Cato College of Education	EDUC	New Graduate Student	1
21	20	Graduate Student	College of Liberal Arts and Sciences	ANTH	New Graduate Student	4
22	21	Freshman	College of Liberal Arts and Sciences	JAPN	New Freshman	1
23	22	Sophomore	College of Computing and Informatics	ITCS	New Freshman	1
24	23	Junior	Belk College of Business	ACCT	New Transfer Student	0
25	24	Junior	College of Liberal Arts and Sciences	PSYC	New Transfer Student	1
26	25	Sophomore	College of Liberal Arts and Sciences	ESCI	New Transfer Student	0
27	26	Junior	William States Lee College of Engineering	ENGR	New Freshman	0
28	27	Freshman	University College	MATH	New Transfer Student	0
29	28	Graduate Student	College of Computing and Informatics	ITCS		0
30	29	Graduate Student	College of Computing and Informatics	ITCS	New Graduate Student	0

From Messy Data … to Clean … to Interactive Visualizations



Guided Practice Scenario

Your colleagues in Library Research & Instruction want some help analyzing 4 years worth of student engagement data.

Necessity of Data Cleanup

Getting to Know Your Data

Splitting & Concatenating Data

Filtering & Removing Duplicates

Creating & Using Lookup Tables

Before Digging in … Ask Some Questions

Are there specific questions?

- How many sessions x session type in each academic year?
- Who are they serving?
  - Undergrads, Master's, Doctoral
  - College of Enrollment
- Which is busier time of year - Fall or Spring?
- How many "repeat customers" have they had?
- Are they reaching more students over time? (Data begins in 2016)

How will the data be used?

- Interactive Dashboard to aid internal decision making (ex: staffing, marketing)
- Create a slide presentation with data visualizations to share with the Provost to advocate for more financial resources.

Two "Messy" Excel Datasets

Download the two Excel datasets to follow along.

File Name: Student Engagement.xlxs

Fields:

1. Email Address
2. Semester (Fall & Spring)
3. Year (2016, 2017, 2018, 2019)
4. Engagement Type
    a. Information Literacy Instruction
    b. Data Literacy Instruction
    c. EndNote Workshops
    d. Patent Workshops
    e. Library Tours
    f. Makerspace Training
    g. Reference Consultation

File Name: Student Information.xlxs

Fields:

1. Username
2. Student ID
3. College
    a. Arts & Architecture
    b. Business
    c. Computing & Informatics
    d. Education
    e. Engineering
    f. Health & Human Services
    g. Liberal Arts & Sciences
    h. University College
4. Major
5. Classification Faculty
    a. Graduate Student
    b. Post Bac
    c. Post Doc
    d. Undergraduate

Explore the Datasets

- How is the dataset organized?
- How many records?
- What types of data does it contain?
  - Is there anything that can be grouped or categorized?
  - What types of questions could you possibly answer with this data?
  - Is there anything that you don’t understand?

To tie two datasets together, you need a key identifier. Sometimes you have to create the "key" by Splitting and/or Concatenating

Necessity of Data Cleanup

Getting to Know Your Data

Splitting & Concatenating Data

Filtering & Removing Duplicates

Creating & Using Lookup Tables

Identify the Key (or Future Key) Identifier

They need to be exactly the same.

File Name: Student Engagement.xlxs

Fields:

1. Email Address (racroxto@uncc.edu)
2. Semester (Fall & Spring)
3. Year (2016, 2017, 2018, 2019)
4. Engagement Type
    a. Information Literacy Instruction
    b. Data Literacy Instruction
    c. EndNote Workshops
    d. Patent Workshops
    e. Library Tours
    f. Makerspace Training
    g. Reference Consultation

File Name: Student Information.xlxs

Fields:

1. Username (racroxto)
2. Student ID
3. College
    a. Arts & Architecture
    b. Business
    c. Computing & Informatics
    d. Education
    e. Engineering
    f. Health & Human Services
    g. Liberal Arts & Sciences
    h. University College
4. Major
5. Classification Faculty
    a. Graduate Student
    b. Post Bac
    c. Post Doc
    d. Undergraduate

Splitting Text to Columns & Concatenating

- **Text to Columns**
  - Parses the text in one cell (column) into two or more columns using a delimiter
    - EX: comma, semicolon, or Other such as @
  - Add a few empty columns to the right of the column you want to split.
  - Copy/paste the data from the original column into a second column
  - Highlight > Data > Text to Columns > Delimited > "Select delimiter" such as tab, comma, other (@)) > Next > Finish
  - Example: racroxto@uncc.edu → racroxto (in one column) and uncc.edu (in the next column)

- **Concatenate**
  - Joins two or more text strings into one string
  - Example: =concatenate(b2," ",c2)

1	Username (Col A)	Term (Col B)	Year (Col C)	Semester (Col D)
2	racroxto	Fall	2017	Formula: =concatenate(b2," ",c2) Returns: Fall 2017

Sometimes …

We have to create categories in order to make our data meaningful. Filtering & Removing Duplicates are important first steps in this process.

Necessity of Data Cleanup

Getting to Know Your Data

Splitting & Concatenating Data

Filtering & Removing Duplicates

Creating & Using Lookup Tables

This then allows you to …

tie data from different sources or worksheets into a single location .. which will enable your analysis and visualization!

Necessity of Data Cleanup

Getting to Know Your Data

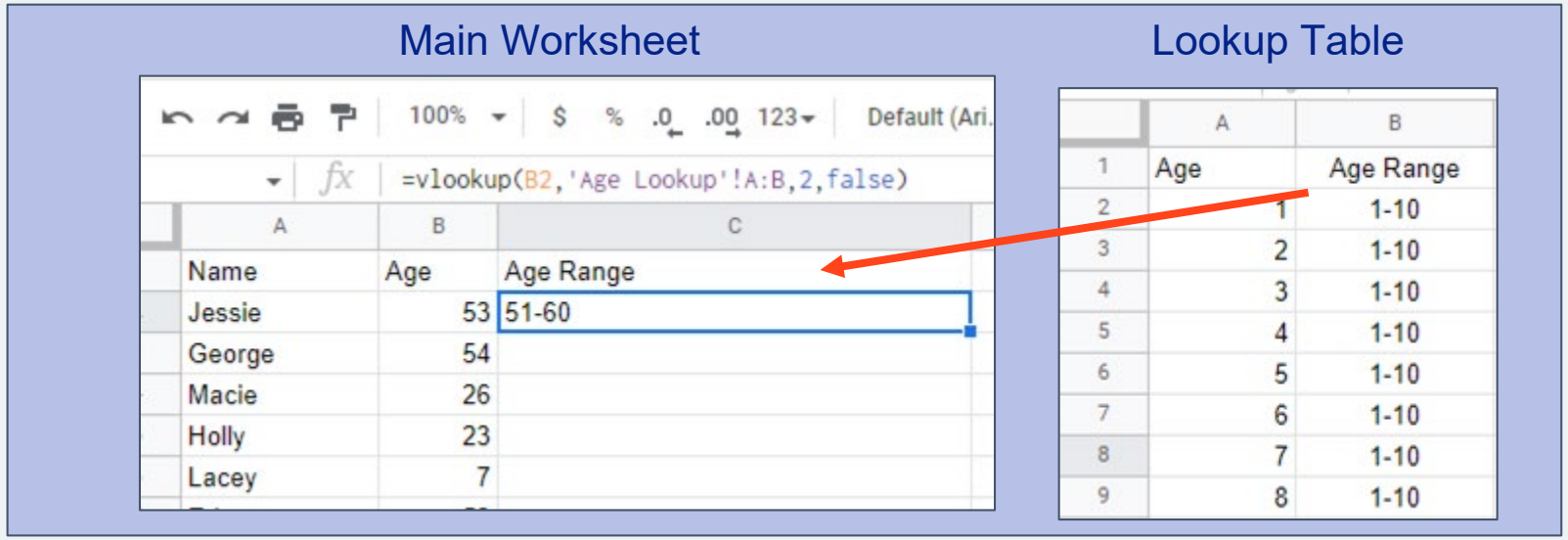
Splitting & Concatenating Data

Filtering & Removing Duplicates

Creating & Using Lookup Tables

VLookup Formulas & Lookup Tables

- This step usually follows filtering and removing duplicates.
- One worksheet will be your main or destination worksheet … where you want everything to go … and the other worksheets with the supplemental information will serve as the lookup files.
- If you're creating a new categorical field, you'll need to create a lookup table with the "definitions" … your categories.
- In the main worksheet, you use a Vlookup formula to tie the new categories to your existing data.
  - In its simplest form, the VLOOKUP function says: =VLOOKUP(What you want to look up, where you want to look for it, the column number in the range containing the value to return, return an Approximate or Exact match – indicated as 1/TRUE, or 0/FALSE).
  - =VLOOKUP(lookup_value, table_array, col_index_num, [range_lookup]



Let's Recap

Sometimes …

…understanding what Excel (or Google Sheets) can do for you

and

… knowing what to Google (search) are the most important things you need to get started in data analysis and visualization!

- ✓ Necessity of Data Cleanup
- ✓ Getting to Know Your Data
- ✓ Splitting & Concatenating Data
- ✓ Filtering & Removing Duplicates
- ✓ Creating & Using Lookup Tables

Cleaning & Organizing Data

Was this content useful?
Please provide your feedback at: https://forms.gle/mwhWAn91jtNFHv8z5

ASSOCIATION OF RESEARCH LIBRARIES