GENERAL ASSEMBLY

FUNDAMENTALS OF DATA IN EXCEL

LEARNING OBJECTIVES

In today’s lesson, we will:

1. Discuss data cleaning best practices.
2. Review strategies to prepare and clean a data set.
3. Practice asking the “right” questions of our data.

DATA CLEANING

"We're ready to scrub the data." #betterdata

FUNDAMENTALS OF DATA IN EXCEL

BEST PRACTICES

1. KEEP A COPY OF ALL UNTOUCHED, RAW DATA.

2. Document ALL of the steps you take in your analysis.

3. Always create a working summary sheet for yourself, that includes the following:
    a. A directory of other sheets.
    b. An explanation of analysis.
    c. A short summary of your results.

HOW TO CLEAN DATA IN EXCEL

FUNCTIONS + METHODS

- Duplicates
- Auto filter
- Concatenate()
- Trim blanks/spaces
- Boolean operators
- Strategy for nulls.
- Cleaning non-printing characters
- Uniform appearance (upper, lower, proper)
- Date field selections
- Numeric column settings
- Text to columns
- Aggregation (%, avg, sum)
- Substitute data if needed

DATA REFERENCING IN EXCEL: NULLS

A `NULL` is any missing value in your data. There are four primary strategies for handling `NULL` values:

1. Delete them (only with caution).
2. Ignore them (some may have meaning).
3. Impute values (e.g. median or zeros).
4. Find missing values (using reference resources).

Guideline: If over 15% of a dataset is filled with `NULL` values, find new data!

BUSINESS SCENARIO

BUSINESS SCENARIO

In this scenario, we'll be working for the Washington State Governor's Office as a **policy analyst**.

Policy analysts often use many different data sources to evaluate policy decisions and make recommendations for how to allocate resources, hold entities accountable, and more.

We'll be using data from the American Community Survey (ACS), which is a ***random survey*** given to U.S. residents each year.

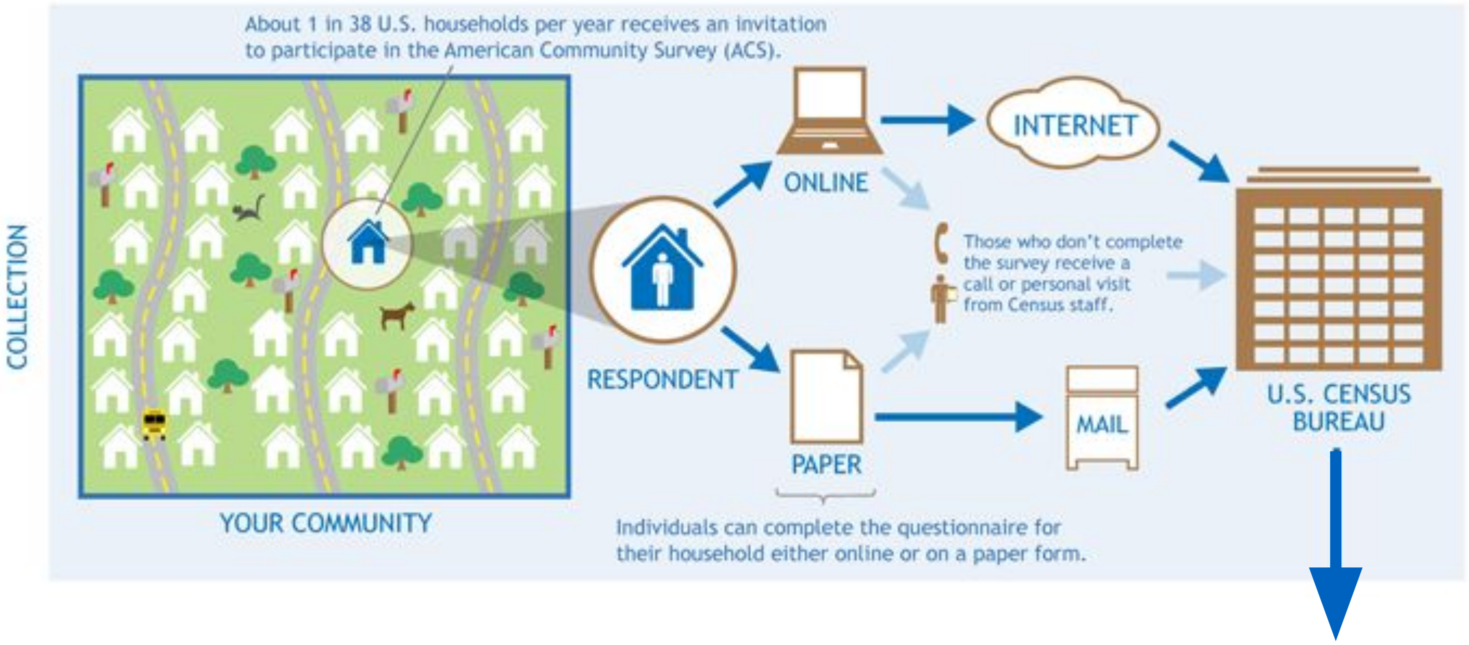
BUSINESS SCENARIO

The ACS survey is a bit different from the standard US census:

- During a census, everyone is asked to respond.
- Sampling is used in off years to provide *estimated* data about the population.
- For the ACS, approximately 1/38 households are asked to respond.
- Because it's a sample, each variable is an estimate that has an associated degree of error, based on number of respondents, sampling strategy, and more.

BUSINESS SCENARIO

Here is a diagram from the U.S. Census Bureau depicting how the ACS works:



About 1 in 38 U.S. households per year receives an invitation to participate in the American Community Survey (ACS).

Those who don't complete the survey receive a call or personal visit from Census staff.

Individuals can complete the questionnaire for their household either online or on a paper form.

Source: http://www.census.gov/programs-surveys/acs/about/how-the-acs-works.html

INDEPENDENT PRACTICE: CLEANING OUR DATA SET

INDEPENDENT PRACTICE: CLEANING OUR DATA SET

A significant amount of any project involves cleaning data (usually between 50–80%) and performing basic exploratory analysis in order to develop a working knowledge of the dataset.

Only after you understand the nuances of a data set and have normalized its structure (by dealing with any `NULL`s, for instance) can you advance and begin analyzing it.

***Source**: https://goo.gl/UivUn4

INDEPENDENT PRACTICE: CLEANING OUR DATA SET

Here’s important information you need to know about our data set:

- Many of the ACS tables have data aggregated by census tracts.
- Census tracts are small areas — sometimes as small as a few blocks in a densely populated area — that the ACS uses for tabulation.
- Each census tract has an ID, which is the “id” field in our data set.
- Because we’re working for the **state of Washington**, our data set only includes census tracts in Washington.

INDEPENDENT PRACTICE: CLEANING OUR DATA SET

Here’s more important information you need to know about our data set:

- Sometimes data are reported as a **total of those counted**.
  - For example, the data set contains the total female population of a census tract, but it doesn’t have the percentage of people who are female.
- Other times, data are **reported as percentages**.
  - For example, the unemployment rate is provided in the data set.
- In order to be able to perform analysis using the ACS data set, we’ll need to make some changes and do some exploration.

INDEPENDENT PRACTICE: CLEANING OUR DATA SET

Let’s go over each column header and see what the data looks like.

a. Let’s freeze the top row so we always see the column header: `View > Freeze Panes > Freeze Top Row`.

a. Let’s Wrap Text on the first row to let us read all of the column headers: `Home > clicking row number to highlight >Wrap Text`.

ACTIVITY: CLEANING OUR DATA SET (BREAK OUT GROUPS)

EXERCISE

DIRECTIONS

1. Open “**PART_1_ACS_Dataset**” tab in *Day 1 - Data Analytics with Excel*
2. Based on your experience, choose either the BASE or STRETCH tab to complete (30 min).

You may work with a partner, checking in with each other after answering each question.

DELIVERABLE

Complete the BASE or STRETCH tab in 2014_acs_select_WA.xlsx.
Be prepared to share an observation or analysis from question No. 4.

GUIDED PRACTICE: REVIEWING OUR SOLUTIONS

GUIDED PRACTICE: REVIEW SOLUTIONS

Goals:

1. Combine values to create meaningful relationships.
2. Create percentage columns for each column that’s not already a percent or rate.
3. Create a ratio for total population (E) and population of males (M).
4. Create a new column with the header “% Male of Population.”
5. Enter the formula “=F2/E2” in the first cell under the header (row 2).
6. Double-click the square (bottom-right corner of the cell) to copy down all rows.

GUIDED PRACTICE: REVIEW SOLUTIONS

Goal: Convert all percentage columns to values between 0–1 (inclusive) with a format of “00.00%.”

Two methods we can use:

- Create a new column equal to the old column but divided by 100.
- Use Paste Special’s “Divide” feature.

GUIDED PRACTICE: REVIEW SOLUTIONS

Goal: Remove all rows with no data.

- What do you do with rows that are completely empty?
- Always document what you delete and make a note of the reason.

GUIDED PRACTICE: REVIEW SOLUTIONS

Goal: Create a common code for cells with no data.

- It’s important to have empty or null values coded in a consistent way.
- This dataset has them represented with both blank and “**-**” cells.
    - If there are textual data amid numeric data, Excel will be unable to plot that data correctly.
- It’s best practice is to recode all empty cells as blank.

GUIDED PRACTICE: REVIEW SOLUTIONS

Are there any data that could be erroneous? If so, what are our options?

- Click through each column’s filter dropdown and take a look at the values. Do you see anything out of order?
- For now, there’s not much we can do except document our findings. Make a note about the age field and others that might be questionable.
- To help us determine if the median age is unreasonable, you may need to look up information about the questionable tract using external reference data to confirm.

GUIDED PRACTICE: SOLUTIONS & FINDINGS TO SHARE

What were some interesting findings?

- Exploratory data analysis is always helpful for finding potentially erroneous data, as well as understanding what data you have in your data set.
- Scatterplots are great tools for looking at relationships between two variables.

CONCLUSION

CONCLUSION

To recap, in this lesson we:

1. Discussed data cleaning best practices.
2. Reviewed strategies to prepare and clean a data set.
3. Practiced asking the “right” questions of our data.

Q&A

RESOURCES

RESOURCES

- For additional resources, check out the student resource directory.
- A thorough guide to the steps of data cleaning:

  https://www.siop.org/tip/backissues/Jan05/PDF/423_089to096.pdf.

- To find these census tracts on a map, you can use this website:

  https://www.huduser.gov/qct/qctmap.html. To search, enter the portion of the ID after "US."

CITATIONS

- The data sets used were compiled from the American Community Survey (ACS): https://www.census.gov/programs-surveys/acs/.
- The data sets were downloaded directly from the American FactFinder site: http://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml.
- All data sets are from the 2014 5-Year Estimate ACS.
- Summary of the ACS Data Collection: http://www.census.gov/programs-surveys/acs/about/how-the-acs-works.html.