INFOECO COOKBOOK | "DO" EXERCISE

DATASET SUMMARIES

DESIGNED BY BOB GRADECK AND LIZ MONK, 2022

OVERVIEW

This exercise will ask you to determine which Pittsburgh neighborhood has the largest number of food-producing trees. You will answer this question by downloading a dataset of street trees in Pittsburgh and using a spreadsheet to generate a pivot table. In the exercise below you will navigate to the Western Pennsylvania Regional Data Center's open data portal, find the dataset of street trees, download this dataset and open it in Microsoft Excel, and use a pivot table to produce summary statistics on which neighborhood has the largest number of trees.

To complete this exercise you will need a computer with Microsoft Excel, an internet connection, and a sense of the types of trees that fruit squirrels find delicious. If you've never made a pivot table before, this exercise should take you between 10 and 15 minutes to complete.

DIRECTIONS

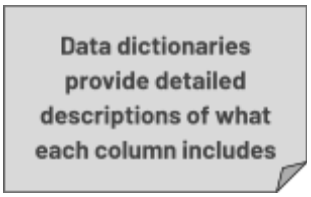
Part 1: Download a Dataset

1. Navigate to the Western Pennsylvania Regional Data Center's website, at www.wprdc.org
2. In the upper right-hand corner of the screen, type the word "tree" in the search box where you see the phrase "Search for open data." Hit enter or click the magnifying glass to see your results.
3. You'll now see a listing of datasets that contain information about trees. Click the link that says "City of Pittsburgh Trees" to open the dataset record. Here you will see links to the dataset and metadata about the dataset which includes a description of the data, when the data was last updated, and who created the data.

Metadata is a structured record containing information about datasets

4. Click the "Download' link next to the "City of Pittsburgh Trees" dataset with a yellow icon. The yellow icon designates a .CSV (comma-separated value) file format that can be opened in a spreadsheet.
5. Find the downloaded file and open it using Microsoft Excel.
6. Now that the file is open, familiarize yourself with the data.
    a. Each row in the spreadsheet is a tree somewhere on the streets of Pittsburgh.
    b. Scroll up and down and left and right. Take note of the column names in the first row of the table, and look to see which columns contain a non-scientific tree name and the neighborhood name. Also look at the ID column–we'll be using this later.

c. Check out the data dictionary if you're interested in learning about what is in the fields. You can find it on the same WPRDC page where you found the dataset.



Part 2: Make a Pivot Table

1. Click in any of the cells in the table to start the process of making a pivot table.
2. In the menu at the top of the screen, select the “Insert” menu, and then click on the PivotTable icon.
3. A “Create Pivot Table” dialog box will open on your screen asking you where your data is located, and where you want to place the pivot table. Excel should recognize the range of the data, so under “choose the data that you want to analyze” keep the “select a table or range” selected.
4. Select “New Worksheet” under the “Choose where to…” prompt and select “OK”.
5. When pivot table worksheet opens, you’ll see a navigation pane open titled “Pivotable Fields” on the right side of the screen.
6. Drag the “neighborhood” field into the “Rows” box.
7. Drag the “common_name” field into the “Filters” box.
8. Drag the “id” field into the “Values” box.
9. You’ll notice that data is now filling in the table to the left. Neighborhood names are showing up under the “Row Labels” column header, and a long number is showing up under the “sum of id” column. The long number is actually the sum of all IDs for trees in this neighborhood. To change the sum of all IDs for trees in the neighborhood to a count of all trees in the neighborhood do the following: in the “Values” box in the bottom-right hand corner of the screen, click on the arrow next to “sum of id” and select the “Value Field Settings” option.
    a. If you are completing this assignment on a mac, you will click the “i” information icon next the “sum of id”
10. In the dialog box that opens, choose to summarize values by count. Other options available include calculating an average, and generating the highest and lowest values in a range. Click “ok” to close the dialog box. In the table, you can now see the number of trees in each neighborhood.
11. Now we need to produce a count of the trees that would provide a squirrel with food. Just above the table in cell B1you’ll see text that says “(All).” Click on the drop-down arrow. You now see a full listing of tree species included in the database, listed by common name. We’ll make use of the “filter” feature in the pivot table to select only the information we want to include in our summary table, i.e., to filter out any non-delicious varieties of tree. Click the box that says “select multiple items” since more than one type of tree produces fruit.

    a. If you are completing this assignment on Mac, you will not need to click the box that says “select multiple items” as you will be able to select whatever you want without this setting checked.

12. You can then select the option at the top of the list marked “All” to clear all selections, then select each tree that looks like it would be important to a squirrel’s decision about where to live. Oak trees produce acorns, and Pawpaw trees are also a local delicacy. Select any others you’d like to include.

13. Click “ok” to apply the filter, and you have your table with counts of the most-delicious trees available. You can sort this list by number to determine which neighborhood has the largest number of food-producing trees. There’s a good chance that the Squirrel Hill neighborhood will be near if not at the top of the list.

    a. If you are completing this assignment on a Mac, there will not be an “ok” option to apply the filter--you will just click out of the filter box.