PLTW COMPUTER SCIENCE

Activity 3.1.1

Time Series and Trends

Introduction

Image ©Microsoft

Think of something you have observed changing. It could be a plume of smoke from a car or a campfire, winding upward; it could be the sounds of the city or the birds waking up in the morning. As things change over time, patterns emerge. Creating visualizations of data can help reveal those trends. A data visualization is any graphical representation of a data set.

Time can be represented in several ways, including:

- Plotting time as the x-coordinate in a scatter plot or line graph.
- Animating any type of data visualization, with each frame visualizing data from one point in time.

Have you ever seen a representation of your height as you got older? Did it fit in one of these two categories?

Materials

- Computer with Enthought Canopy distribution of *Python*® programming language, Flash, and access to Internet

Resources

3.1.1 Source Files

3.1.1 Optional Source Files

Note: When uncompressed, 3.1.1 Source Files contain 22Mb of data and the *optional* 3.1.1 Optional Source Files contain 107Mb of data

Procedure

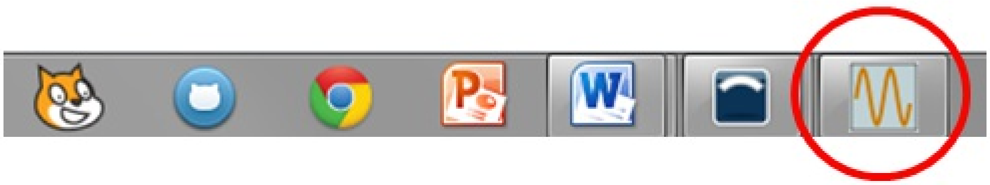
Part I: Name Frequencies

Copyright © 2017 Project Lead The Way, Inc. All rights reserved.

In this activity you will explore two data sets that represent values that change over time.

1. Form pairs as directed by your teacher. Meet or greet each other to practice professional skills. Set team norms.
2. Using Notepad++, open the data file corresponding to your year of birth in the `SourceFiles` folder. Select **Search > Find** and find your name.
    - How many babies of each gender in your year of birth were given the same name as you?
    - Look up another name as similar to yours as possible – maybe a variation in spelling or a common nickname. How many babies of each gender in your year of birth were given that name?
3. If directed by your teacher, use Notepad++ to open the data file corresponding to the state in which you were born, and find the line containing your name and your year of birth. Describe the algorithm you used for finding that line.
4. You are going to construct a graph showing the U.S.-wide popularity of your name, with popularity changing over time. How could you do this by hand? Describe the algorithm you would follow to gather the necessary data from all of the files if you were to do this without programming.
5. Computing changes how we deal with data because people with programming skills can automate time-consuming tasks.

    Launch Canopy. Open an editor window and open the `visualize_names.py` program in the code editor. Execute the program. A plot should be produced. You might have to select the plot's icon in the taskbar.



    The plot shows the frequency of these names vs. time. The frequency of something tells how many times it occurred. Estimate how long it would have taken you to produce this time series plot by hand.

6. Read and discuss the code with your partner. In the code shown below, note the first use of each variable, and identify the variable role (fixed value, most recent, stepper, walker, accumulator, aggregator, one-way flag, or best-so-far) for each variable. If a variable is used for different roles at different lines of code, note that too.

```
import os.path

# Get the directory name for data files
directory = os.path.dirname(os.path.abspath(__file__))

name1 = 'Juan'
name2 = 'Juanita'

#initialize the aggregators
years1=[]
number_of_people1=[]
```

Copyright © 2017 Project Lead The Way, Inc. All rights reserved.

```
years2=[]
number_of_people2=[]

# Scan one year's file at a time
for year in range(1880,2013):
    # Open the file
    filename = os.path.join(directory, 'yob' + str(year) + '.txt')
    datafile = open(filename, 'r')
    # Go through all the names that year
    for line in datafile:
        name, gender, number = line.split(',')
        # Aggregate based on name1
        if name == name1 and gender == 'M':
            years1.append(year)
            number_of_people1.append(number)
        #Aggregate based on name2
        if name == name2 and gender == 'F':
            years2.append(year)
            number_of_people2.append(number)
    # Close that year's file
    datafile.close()

# Plot on one set of axes.
import matplotlib.pyplot as plt
fig, ax = plt.subplots(1, 1)
ax.plot(years1, number_of_people1, '#0000FF')
ax.plot(years2, number_of_people2, '#FF00FF')

ax.set_title('U.S. Babies Named ' +
             name1 + ' (blue) or ' + name2 + ' (pink)')
fig.show()
```

7. Reflect on how well the comments in the code assisted you in understanding why variables were being used and what they represented. Where are more comments needed?
8. Creating code to automate a task often takes as long as just doing the task by hand, especially if it is the first time you are using a library. Estimate how long it would take you as a pair to create this program, including the research you would have to do to find and learn from documentation. Describe how you arrived at your estimate.
9. One reason to automate tasks with a program is to make the task easy to do again after modifying it slightly. Estimate how much time you will need to create another data visualization from this data set, describing the pattern of baby name frequency vs. time for a different name.
10. Writing code to automate a task is worth the work if you will need to repeat the task a certain number of times or more. Based on your answers to Steps 5, 8, and 9, how many visualizations of the names data would make the effort to write this program worthwhile?
11. Modify the code to visualize the pattern of data for your name and a variation of your name as described in Step 2. Repeat with your partner's name and a variation on that name. Paste screenshots of the data visualizations here.
12. Describe the trends you observe and the knowledge you gained about the data you visualized. How do the names compare? When was the peak frequency? Were there periods of time with especially high rates of increasing or decreasing frequency for these names?
13. Using computational power to analyze data is most powerful if the analysis can be generalized. A solution is generalized if it can be applied to solve related problems or answer

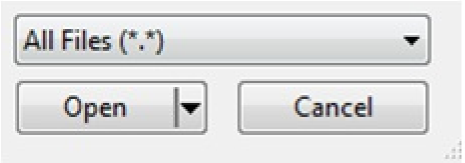
Copyright © 2017 Project Lead The Way, Inc. All rights reserved.

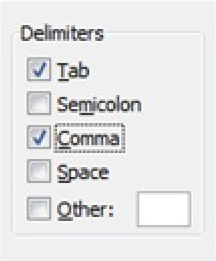
related questions. One way to generalize a solution is to avoid hard-coding values in your code, or at least to use fixed-value variables when you hard code a value. Hard coding arguments in your program makes your code less flexible.

- It is best if hard-coded data is kept in a single place in the program. Line 39 of the code has hard-coded data. What are the disadvantages of this?
- Change line 39 so that it relies on the fixed-value variables that were assigned at the beginning of the program.

14. The program could be made more flexible if other lines in the program used fixed-value variables. Which variables could be created to improve this program? List them. For each one identify:
    - The line number of code where data is currently hard coded.
    - The name of the variable you would use instead.
    - The code you would place at the top of the program to initialize this variable.

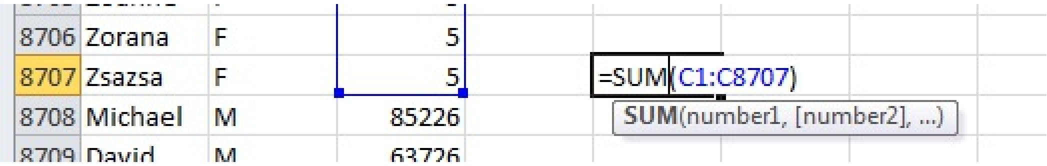
15. What are some questions you could answer from these data sets using computational power that you would not be able to answer in a reasonable amount of time without writing code? Brainstorm as a class.
16. These data files contain the frequency of each name. The frequency is the number of times the name occurred. You might be more interested in what percentage of the people born each year were given a particular name. These percentages would be an example of relative frequencies. Relative frequency is the frequency of something (like a name) as a portion of (divided by) the total. The relative frequencies add up to 100% for each year.
    - Describe an algorithm that could be used to create a data file of relative frequencies from a data file of frequencies.
    - Algorithms can be described using a natural (human) language, pseudocode, flow diagrams, or a computer language. Which type of description did you use in Step 16a, and why?
    - In Microsoft Excel®, open the file containing data for your birth year. Since it is a CSV file instead of an XLSX file, you will have to select **All Files** in the Open File dialog box as shown on the left below. In the Import Wizard, include **Comma** in the **delimiters** as shown on the right below.




Delimiters
☑ Tab
☐ Semicolon
☑ Comma
☐ Space
☐ Other:

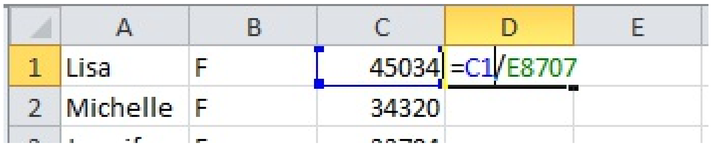
    - Calculate the relative frequency of the names with the following steps.
        - Use the scrollbar to find the first and last names of your gender. In an empty cell in column E, create a formula for the total of all people of your gender as shown below. Use the formula `=sum(first_cell:last_cell)`. To enter this colon-separated range, you can select the range with your mouse while editing the formula. For a large range like this, select a cell at one end of the range, use the scroll bar to move to the other end of the range, and then use shift-click to select the cell at the other end of the range. Once you have selected the range, you can press enter to accept the formula.

Copyright © 2017 Project Lead The Way, Inc. All rights reserved.



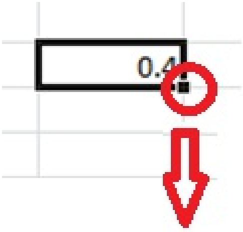
8706	Zorana	F	5	
8707	Zszasza	F	5	=SUM(C1:C8707)
8708	Michael	M	85226	SUM(number1, [number2], ...)
8709	David	M	63726	

- Take note of the cell containing the sum you just created. Create a formula for the relative frequency of the first name, `=cellWithAbsoluteFrequency/cellWithSum`. An example is shown below.

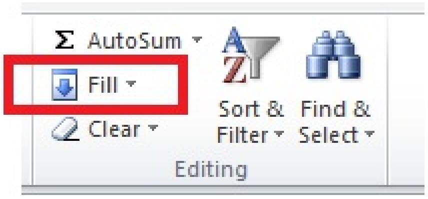


	A	B	C	D	E
1	Lisa	F	45034	=C1/E8707	
2	Michelle	F	34320		

- Change the formula you just created so that the reference to the sum has an absolute reference for the row. An absolute reference to a row or column will continue to refer to the same row or column when copied to another cell. You can create an absolute reference in Excel by inserting a `$` before the row or column number. For example, `=C1/E$8707` will continue to refer to row 8707 when copied.
- Use the fill tool to copy the formula to all the other rows for your gender. For a small spreadsheet, you can drag the fill anchor as shown below.



For a larger spreadsheet like this one, select a range to fill, including the cell from which you want to copy. You can use the select-scrollbar-shiftclick method from step i above. Then select the fill tool from the ribbon, as shown below.



- Use `Ctrl-f` to find your name in the spreadsheet. What percent of males or females born in your birth year were given your name?

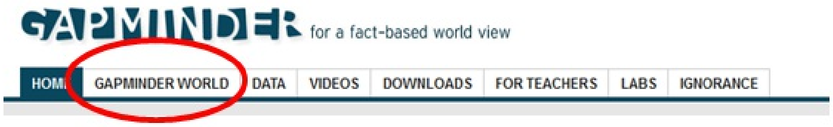
17. The baby naming data are from the Social Security Administration. They provide interactive web pages that let you query a single name against a limited database, and they also provide static web pages offering download of the full data set in the compressed files you received, provided at http://www.ssa.gov/OACT/babynames/limits.html. The page with the zipped data files says, “To maintain an acceptable performance level on our servers, we provide only the top 1000 names through our forms.” Explain – in terms of CPU operations – why using

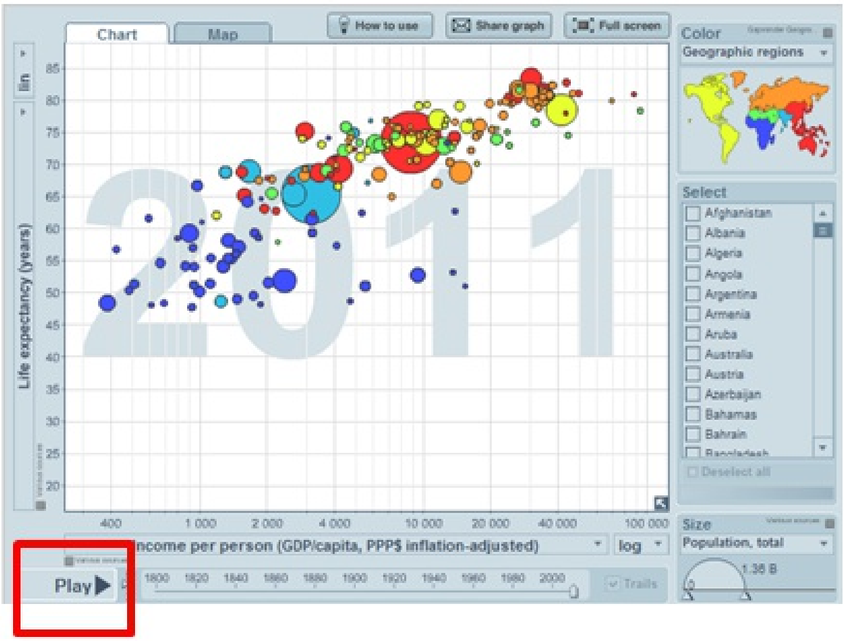
Copyright © 2017 Project Lead The Way, Inc. All rights reserved.

only the most common 1000 names makes their web pages load more quickly.

Part II: Country Data

18. Navigate to http://gapminder.org and load the Gapminder tool that resembles the image below, noting that the year 2011 may be more current. Click the play button and observe.





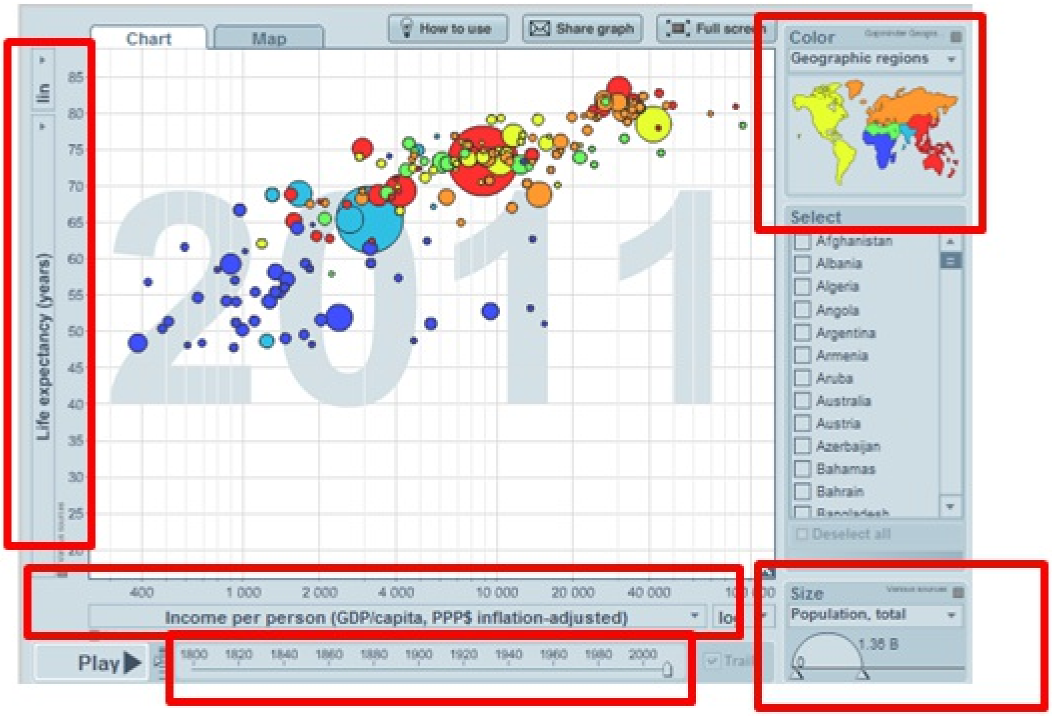
Income per person (GDP/capita, PPP$ inflation-adjusted)

Color: Geographic regions

Size: Population, total

19. Notice that the visualization is displaying five variables. What does each of the following represent?

Copyright © 2017 Project Lead The Way, Inc. All rights reserved.



- x-coordinate of a dot:
- y-coordinate of a dot:
- radius of a dot:
- color of a dot:
- motion of a dot while the animation is playing:

20. Describe one general pattern that you observe.
21. Describe one anomaly that you observe. An anomaly is a data point or single pattern in a data set that does not follow the general pattern of the data set.
22. Near the bottom right of the tool, select the **OPTIONS** icon. Select new variables for the x-axis and y-axis from the drop-down menus. Click play again, and identify a few countries that do interesting things when the graph is animated. You may also choose to identify countries of special interest to you. Not all countries report data for all variables in all years, so depending on what variables you select, you might not see much data. (If you need to reset the tool, go back to the gapminder web page and reload it.)
23. Create a slide show: Pause the animation to capture screenshots at various years. Annotate and comment on the screen shots to explain data trends you see and assumptions you make from the animation.
24. The baby names data contained five variables: name, gender, geographic state, time, and frequency. The line graph represented time on the x-axis and frequency on the y-axis. Each visualization represented one or two single values of name and gender and a single sum across all states.

    To extract knowledge from data with a lot of variables, we have to be creative in how we represent data. Consider a visualization of the data set in Part II that would use a time series line graph like those created in Part I. Sketch an example of how that might look.

25. Compare the visualizations in Part I and Part II. Which type of visualization was better? For what types of data sets would you use a plot with time on the x-axis? For what types of data sets would you use an animation to represent time?

Copyright © 2017 Project Lead The Way, Inc. All rights reserved.

Conclusion

1. Computers already are better than humans at some aspects of data analysis. Humans continue to be better than computers at other aspects of data analysis. In both Parts I and II, you used a computer to create a visualization but used your brain to identify patterns in the data. Data visualization combines the strengths of humans and computers. What part of the discovery of knowledge in data are computers better at? What part of the discovery of knowledge in data are humans better at? Why?
2. Why are algorithms helpful when dealing with large data sets?
3. Look for a visualization of data that represents time as part of the visualization. Paste a copy of the visualization here. Describe what you think are strengths and weaknesses of the visualization. What were some good creative decisions made by the authors of the visualization, and what could they have done better? Are there any indications that the visualization does not correctly represent the data?

Copyright © 2017 Project Lead The Way, Inc. All rights reserved.