**Data Cleanup & Fusion**

First I loaded the cs171\_survey\_4years file into Google Refine and then exported it as a csv file so it would be in the same format as the cs171\_survey\_2013 file. Both files were then loaded into Google Refine again and the following clean up was done:

1. Combined the ‘Comfortable with Programming’ field with the ‘Programming skill’ field
2. Added the year 2013 to the year column for the 2013 survey data
3. Combined the ‘Degree’ and ‘Primary Concentration’ fields
4. Combined the ‘Secondary Concentration’ and ‘Secondary Concentration (if any)’ fields
5. Combined the ‘Languages’ and ‘What other languages do you know?’ fields
6. Combined the ‘Coding skills’ and the ‘Overall, how comfortable are you with programming?’ fields
7. Combined the ‘Programming Experience’ and the ‘How long have you been programming’ fields
8. Removed the ‘How comfortable are you with design’ column as it was not present in the previous years’ surveys.
9. Replaced occurrences of ‘Division of Continuing Education’ with DCE in the type field
10. Performed the following clean up on the student’s locations
    1. Separated out the 2013 data into individual columns for City, State and Country
    2. Normalized the Country name (i.e United States to USA)
    3. Normalized the State abbreviations (i.e. Mass to MA)
    4. General cleanup of some students comments
    5. Removed the town column as it wasn’t present in the previous years’ surveys
11. Normalized the OS data so that the different OS groups were separated into:
    1. Windows
    2. Mac OS
    3. Linux/Unix
    4. Created a Total OS column that counted how many different OSes the student knows. This is a helper column used to answer question #4.

**Visualizing Data in Tableau**

1. Which (primary) programming language is preferred by male and female students, respectively?

Male and Female students both overwhelmingly prefer C for a programming language.

1. Which (primary) programming languages do students feel more comfortable with?

On average students feel most comfortable with Ruby with an overall average of 4.368. While only 19 students listed Ruby as their primary language those students rated their coding skill with Ruby very high.

1. What relationships are discernible between programming experience (in years) and (primary) programming languages of choice?

For those students with 0 to 3 years programming experience their preferred language is C (the cs50 effect?). For the students that have been programming for over three years their preferred language is Java.

1. How do OS choices relate to students’ reported comfort as programmers?

As students have more comfort in programming (coding skills) the use of Linux as an operating system goes up though it is still the least likely used out of Windows, Mac OS and Linux. Those who are the least comfortable with programming tend to use Macs slightly more than Windows.

1. What is the relationship between age, programming language, and programming experience, and how does it change over time?

The younger age groups (18 to 24) with a high level of programming skill rank Ruby as their programming language of choice with Python and Java close behind, though it should be noted that the number of students in this age group that rank C as their overall language of preference (regardless of skill) far outweighs other languages. Those in the same age group with lower coding skills choose basic and HTML/CSS as their language of choice. Ruby has remained the top choice for skilled programmers over the past five years.

Those in the middle age group (25 to 44) are very similar to the 18 to 24 age group in terms of languages chosen based on coding skill.

The students in the 45 to 64 age group have PHP and Java as their chosen language for those with high coding skill and basic and HTML/CSS as the main choice for those with lower coding skills. It should be noted that there are not a lot of data points for this age group.

1. How does the 2013 data compare to the data from years prior? Is there anything notably different about the data from 2013 relative to the past, given enrollment in the class has increased quite a bit? What trends can you pick out from the data, and what can you predict (if anything) about the future of CS171?

There are several things to note about the 2013 data:

* 1. Enrollment by students in the United States, while high in previous years, went from 114 in 2011 to 185 students in 2013. One would expect this to continue in future years.
  2. The enrollment of male students increased from 91 in 2011 to 136 in 2013 as compared to women which saw an increase from 50 in 2011 to only 67 in 2013.
  3. The top degree students are pursuing continues to be Computer Science with a large increase from 26 in 2011 to 46 in 2013. Economics was a distant second with no noticeable increase over the past 5 years.
  4. While the data is incomplete the number of DCE students has shown a slow, but steady increase over the past five years, however the number of Harvard College students still outweighs the number of DCE students (136 compared to 73).
  5. The number of students choosing laptops over desktops is skyrocketing over the past five years and can be expected to continue.
  6. The number of students who choose C as their language of choice had decreased in the past few years (though it was still the far and away leader), but in 2013 it saw a big increase from 41 to 61 students listing it as their primary language. Java also continues to grow in popularity at a slower pace with Python a close third showing a big increase from 2011 (13 to 30).