CS 171/CSCI E-64: Visualization

Homework 3, Problem 3: Tableau

James Goodspeed – jgoodsp@fas.harvard.edu

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 cleanup was done:

1. Added the year 2013 to the year column for the 2013 survey data
2. Combined the 'Degree' and 'Primary Concentration' fields
3. Combined the 'Secondary Concentration' and 'Secondary Concentration (if any)' fields
4. Combined the 'Languages' and 'What other languages do you know?' fields
5. Combined the 'Coding skills' and the 'Overall, how comfortable are you with programming?' fields
6. Combined the 'Programming Experience' and the 'How long have you been programming' fields
7. Removed the 'How comfortable are you with design' column as it was not present in the previous years' surveys.
8. Replaced occurrences of 'Division of Continuing Education' with DCE in the type field
9. Performed the following clean up on the student's locations
    a. Separated out the 2013 data into individual columns for City, State and Country
    b. Normalized the Country name (i.e United States to USA)
    c. Normalized the State abbreviations (i.e. Mass to MA)
    d. General cleanup of some students comments
    e. Removed the town column as it wasn't present in the previous years' surveys
10. Normalized the OS data so that the different OS groups were separated into:
    a. Windows
    b. Mac OS
    c. Linux/Unix
11. Changed the year field from just a year (i.e. 2009) to a full date so that Tableau could parse it (i.e. 1/1/2009).
12. Changed the case on the Primary and secondary degree fields to title case (i.e. Change computer science to Computer Science) so that changes in case did not result in different categories.

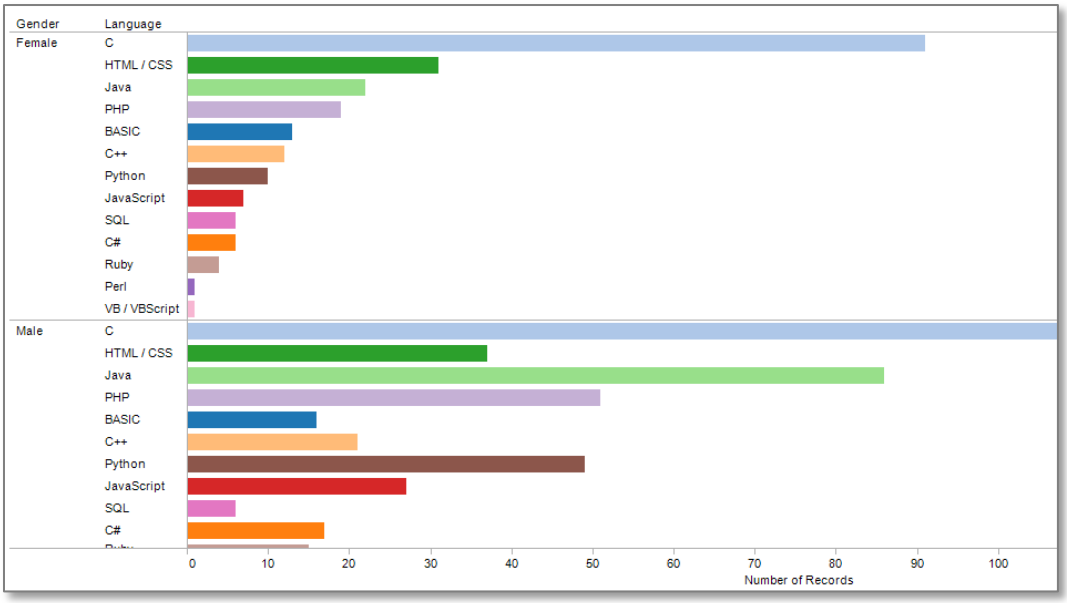
Visualizing Data in Tableau

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. The following screen shot of two bar graphs clearly shows the preference for C as a programming language for both genders. A bar chart was chosen for the ease of comparison between the data variables. Color was added to make the bars distinct.

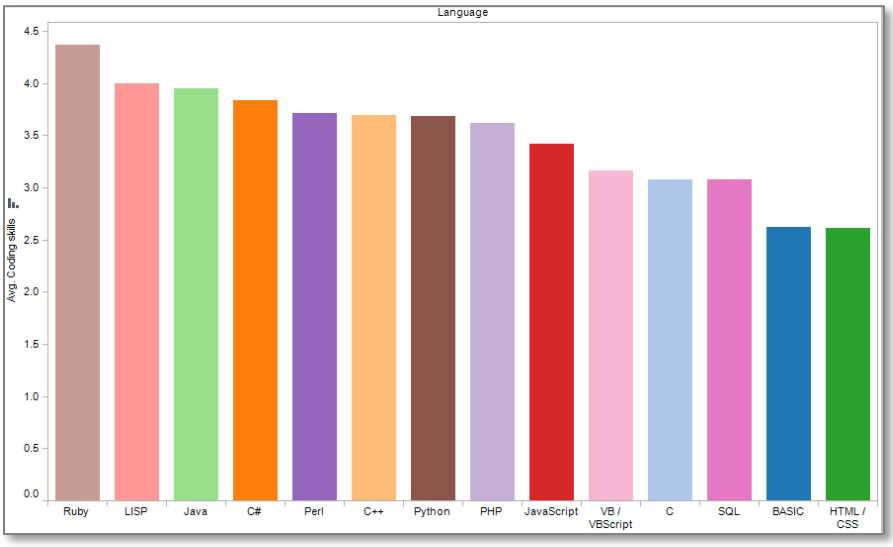
CS 171/CSCI E-64: Visualization
Homework 3, Problem 3: Tableau

James Goodspeed – jgoodsp@fas.harvard.edu



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

   On average students feel most comfortable with Ruby with an overall coding skill average of 4.368. While only 19 students listed Ruby as their primary language those students rated their coding skill with Ruby very high. The following graph shows the coding skill (comfort level) on the Y axis and the language on the X axis with Ruby as the leader with Lisp and Java a close second. Note that this measures coding skill, or comfort level, not the number of students that have chosen a particular language. A bar chart was chosen to easily order the data.



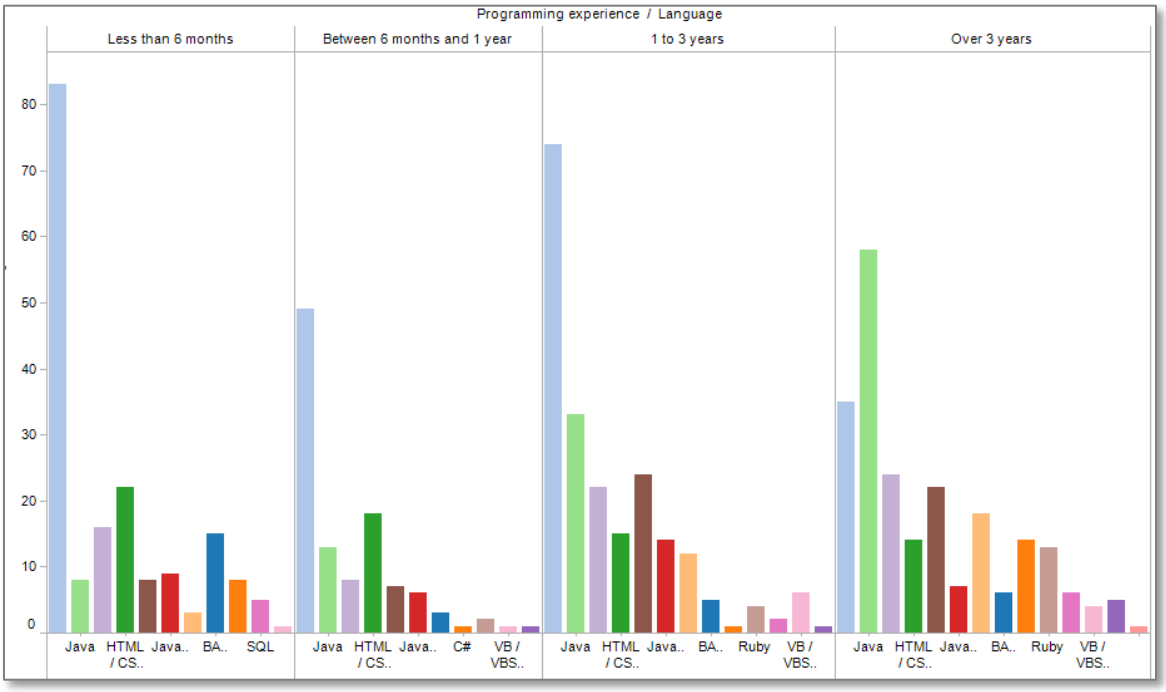
CS 171/CSCI E-64: Visualization

Homework 3, Problem 3: Tableau

James Goodspeed – jgoodsp@fas.harvard.edu

3. 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. The small multiples of bar graphs below show the results of this data in a format that is easily ordered and comparable. The light blue bar for C is strongest in the first three categories until the final, over three years of experience category.

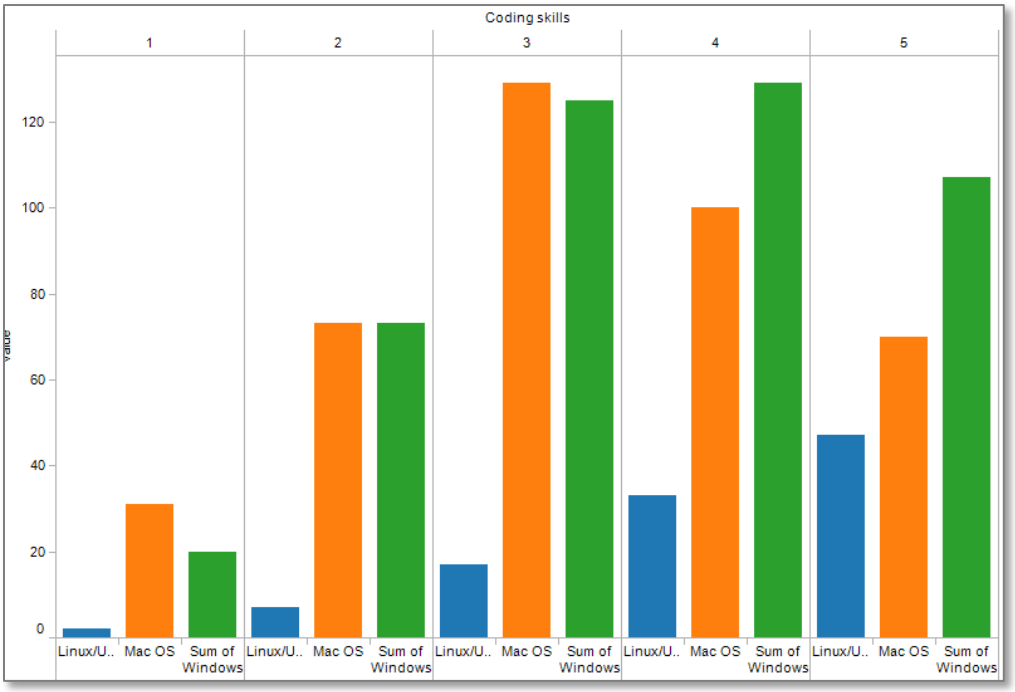


4. How do OS choices relate to students' reported comfort as programmers?

   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. Again a bar chart was chosen for the ease of comparison and color was added to aid in discerning the difference between the categories. The bar graph below shows this upward trend of Linux as users gain programming proficiency.

CS 171/CSCI E-64: Visualization
Homework 3, Problem 3: Tableau

James Goodspeed – jgoodsp@fas.harvard.edu



5. 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. Surprisingly the use of Python for skilled programmers has dropped off in recent years.

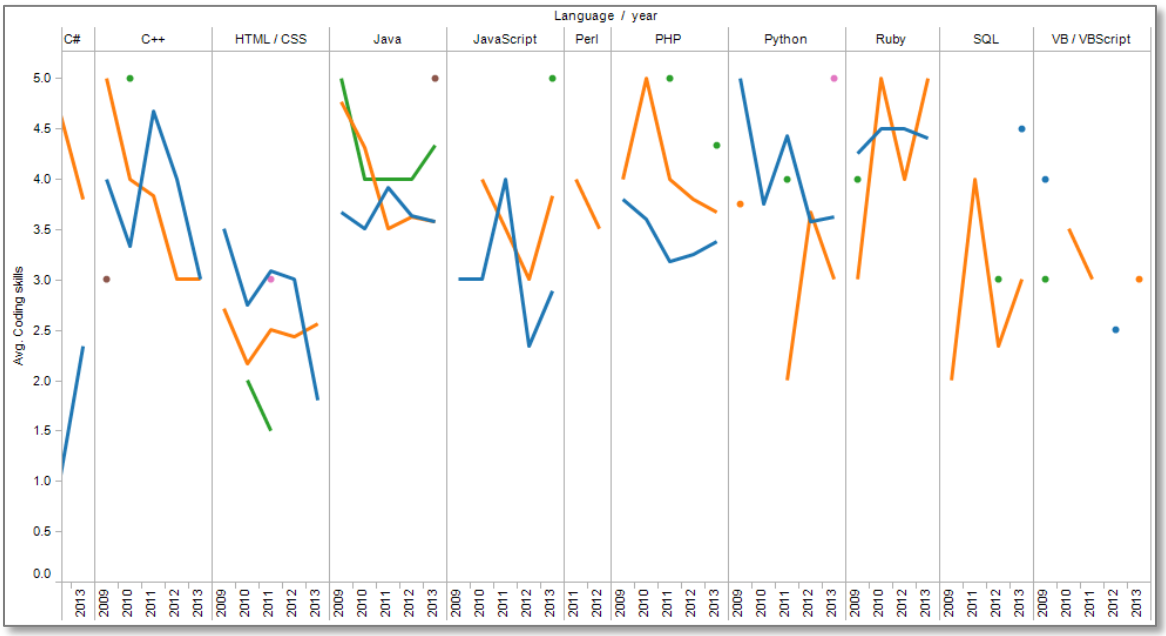
   Those in the middle age group (25 to 44) are very similar to the 18 to 24 age group in terms of languages chosen based on coding skill with higher skilled programmers choosing Ruby. Languages such as C++ that were the choice of highly skilled programmers has dropped off in recent years. Similarly PHP for the 18 to 24 and 25 to 44 groups has also dropped off for skilled programmers.

   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. Similarly there is little data for the 65+ age group though skilled programmers in

James Goodspeed – jgoodsp@fas.harvard.edu

this group prefer Java. The 17 and under age group also has little data, but this age group has a preference for Python for the skilled programmers in this group.

The visualization I chose for this is a series of small multiple line graphs to show the change in data over time. I really struggled with what type of visualization to choose for this data set and originally had a line graph that would 'page' through each age category, but ultimately decided that the choice below presented the data all at once even though it is busy. I also considered a stacked area chart, but this choice distorted the Y axis coding skill measurement because it effectively summed the different age groups so instead of having a range of 0 to 5 the range was 0 to 22.



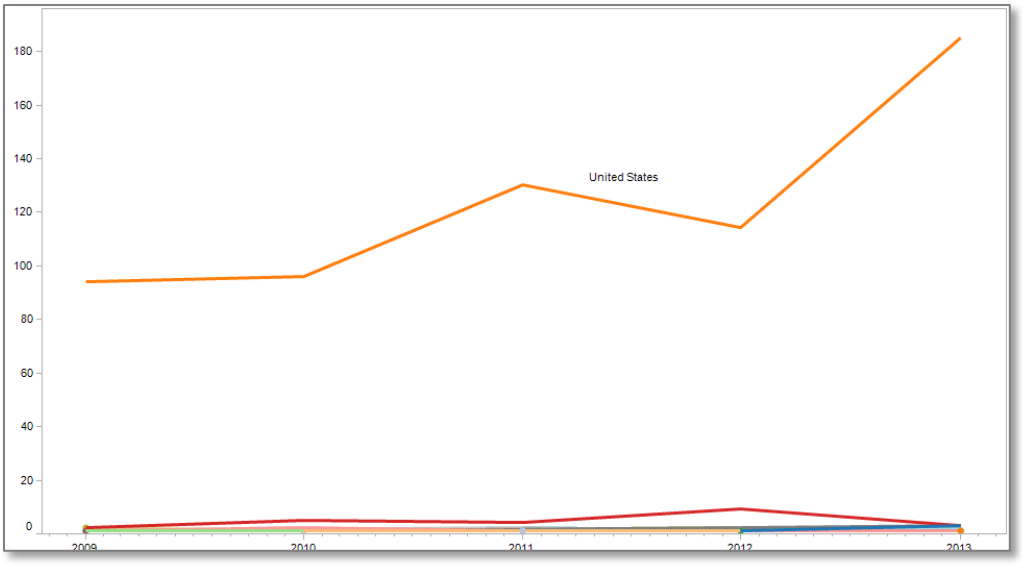
6. How does the 2013 data compare to the data from years prior? Is there anything notably different about the data from 2013 relative to the past, given enrollment in the class has increased quite a bit? What trends can you pick out from the data, and what can you predict (if anything) about the future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

   All of the following visualizations were created with line graphs to show the progression and change over time. There are several things to note about the 2013 data:

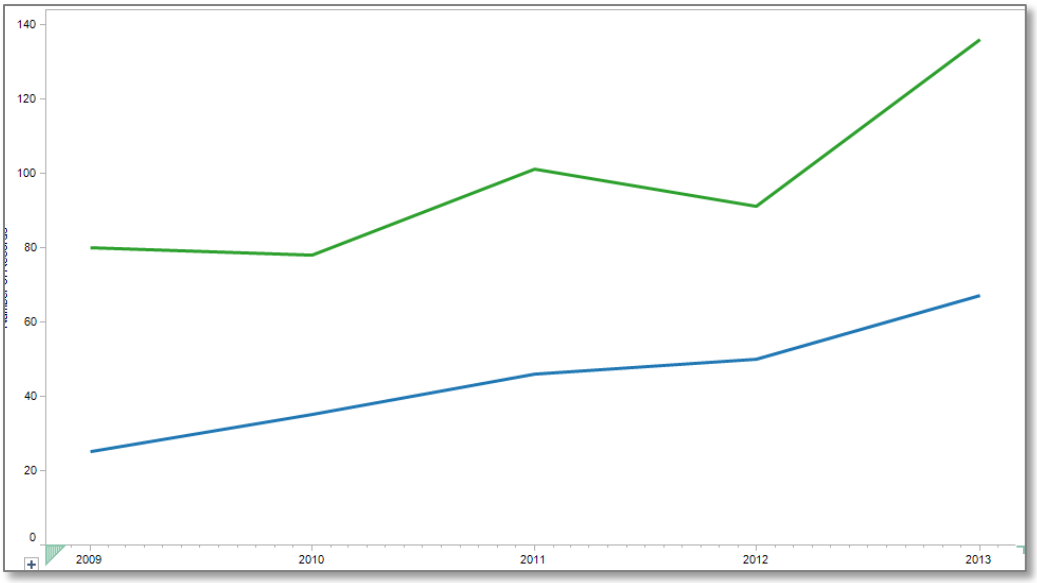
   a. Enrollment by students in the United States, while high in previous years, went from 114 in 2011 to 185 students in 2013. One would expect this to continue in future years.

CS 171/CSCI E-64: Visualization
Homework 3, Problem 3: Tableau

James Goodspeed – jgoodsp@fas.harvard.edu



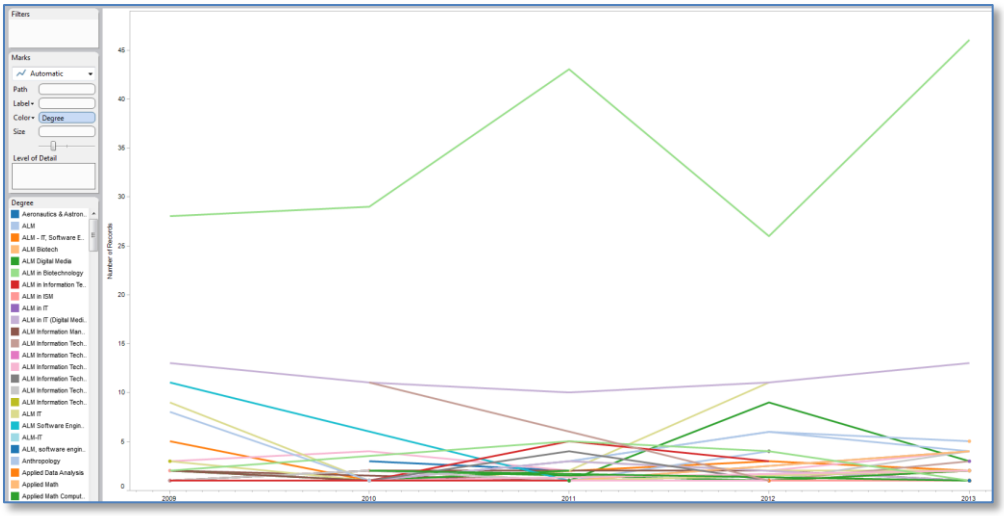
b. The enrollment of male students increased from 91 in 2011 to 136 in 2013 as compared to women which saw an increase from 50 in 2011 to only 67 in 2013. From the data it appears that this trend will continue.



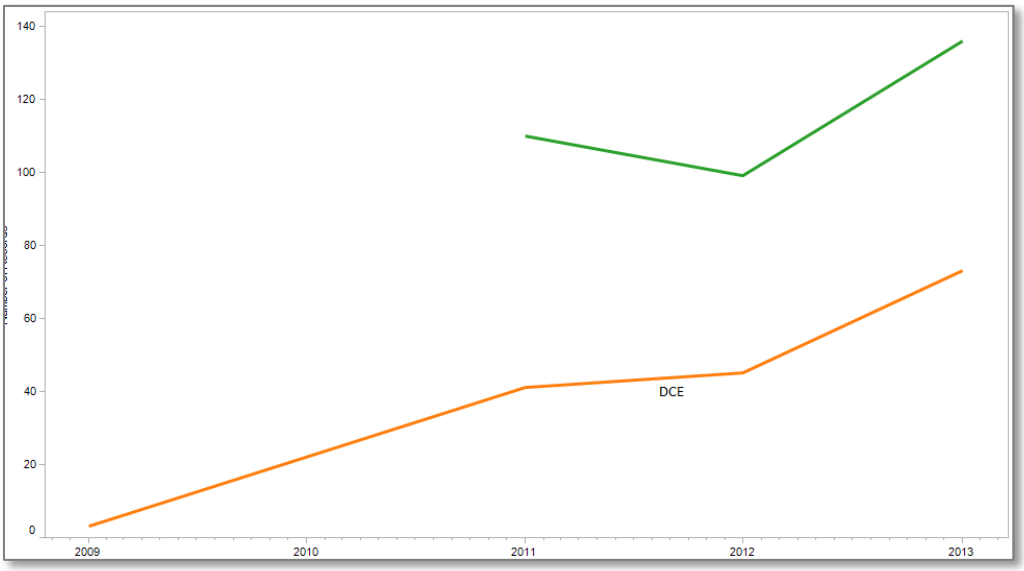
c. The top degree students are pursuing continues to be Computer Science with a large increase from 26 in 2011 to 46 in 2013. Economics was a distant second with no noticeable increase over the past 5 years. I would also expect this trend to continue as well.

CS 171/CSCI E-64: Visualization
Homework 3, Problem 3: Tableau

James Goodspeed – jgoodsp@fas.harvard.edu



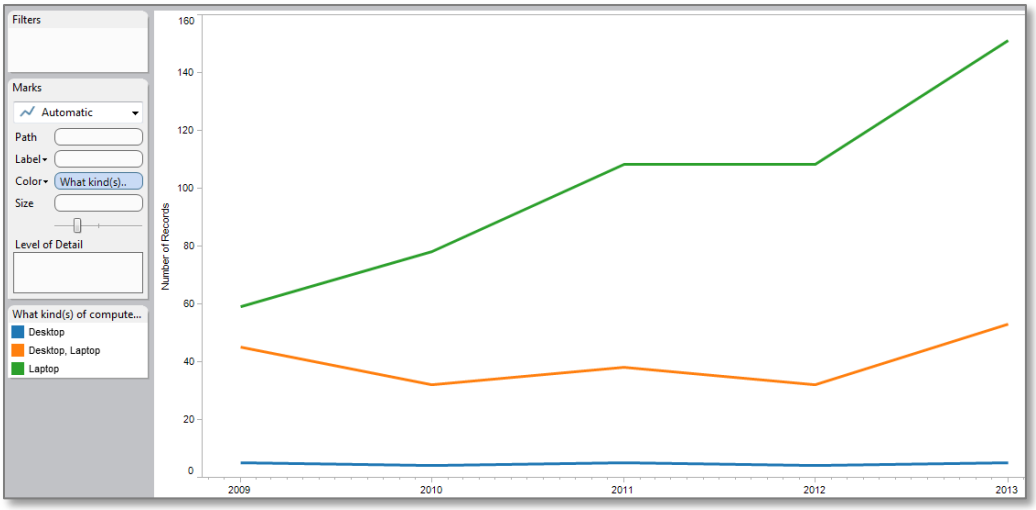
d. While the data is incomplete the number of DCE students has shown a slow, but steady increase over the past five years, however the number of Harvard College students still outweighs the number of DCE students (136 compared to 73).



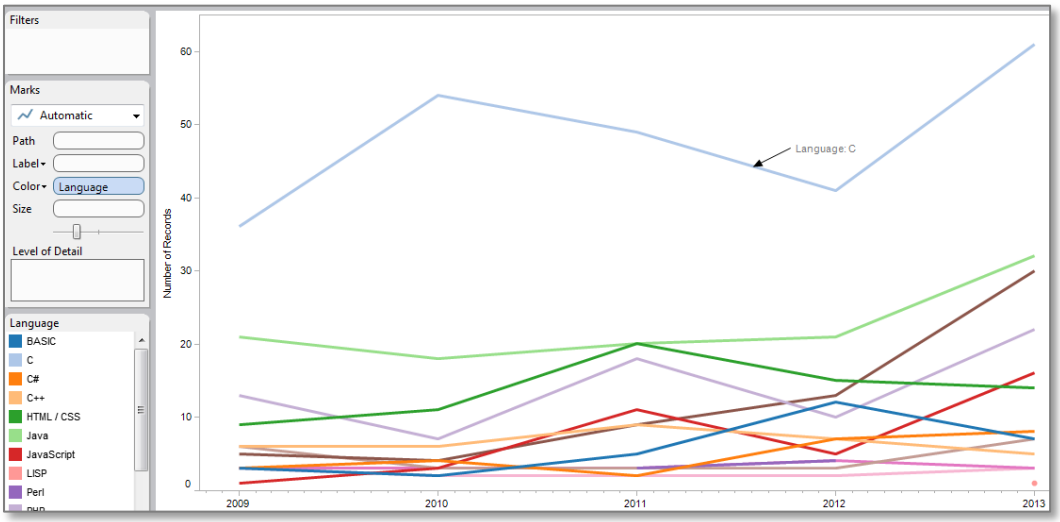
e. The number of students choosing laptops over desktops is skyrocketing over the past five years and can be expected to continue reflecting the current trends in the computer device market.

CS 171/CSCI E-64: Visualization
Homework 3, Problem 3: Tableau

James Goodspeed – jgoodsp@fas.harvard.edu



f. 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).



Dashboards in Tableau

In the following visualization all of the data points are linked together. Clicking on a country will filter that data for the other visualizations on the dashboard. Similarly clicking on any of the other graphic elements such as a bar in a bar graph will filter the data for all the other visualizations. The dashboard has variables for country, language, gender, computer device, coding skill and OS choice.

CS 171/CSCI E-64: Visualization
Homework 3, Problem 3: Tableau

James Goodspeed – jgoodsp@fas.harvard.edu