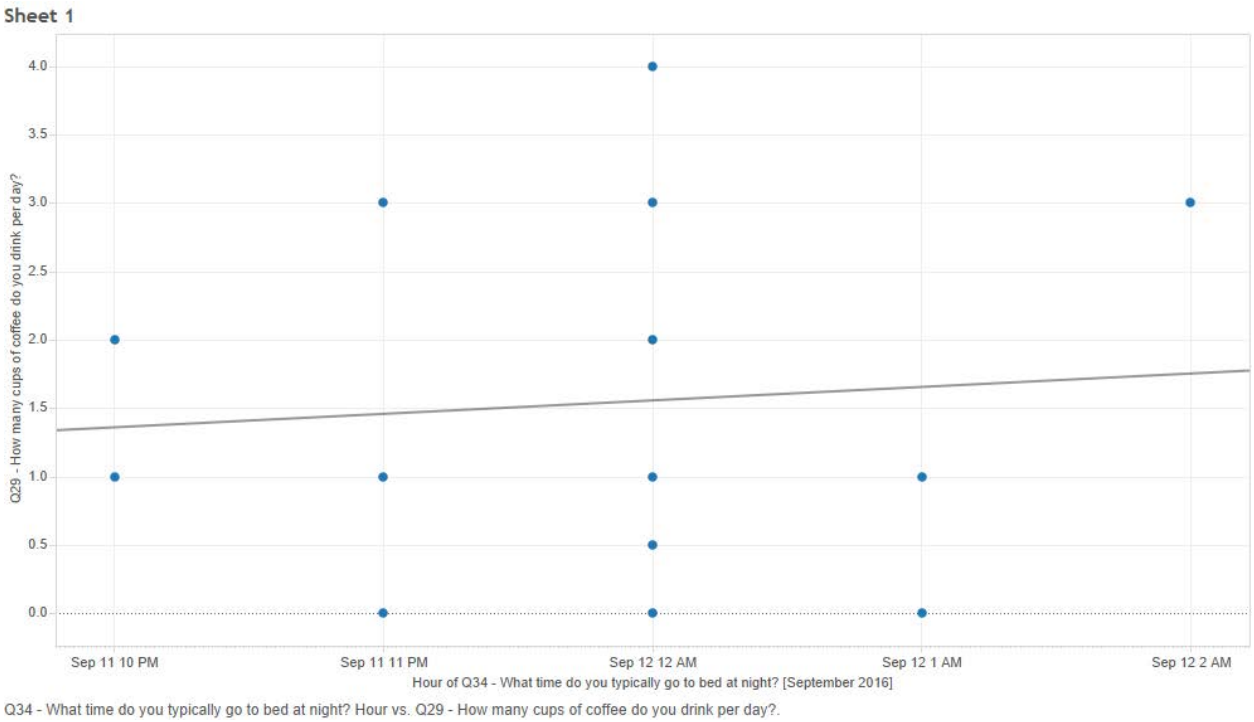
5/6)

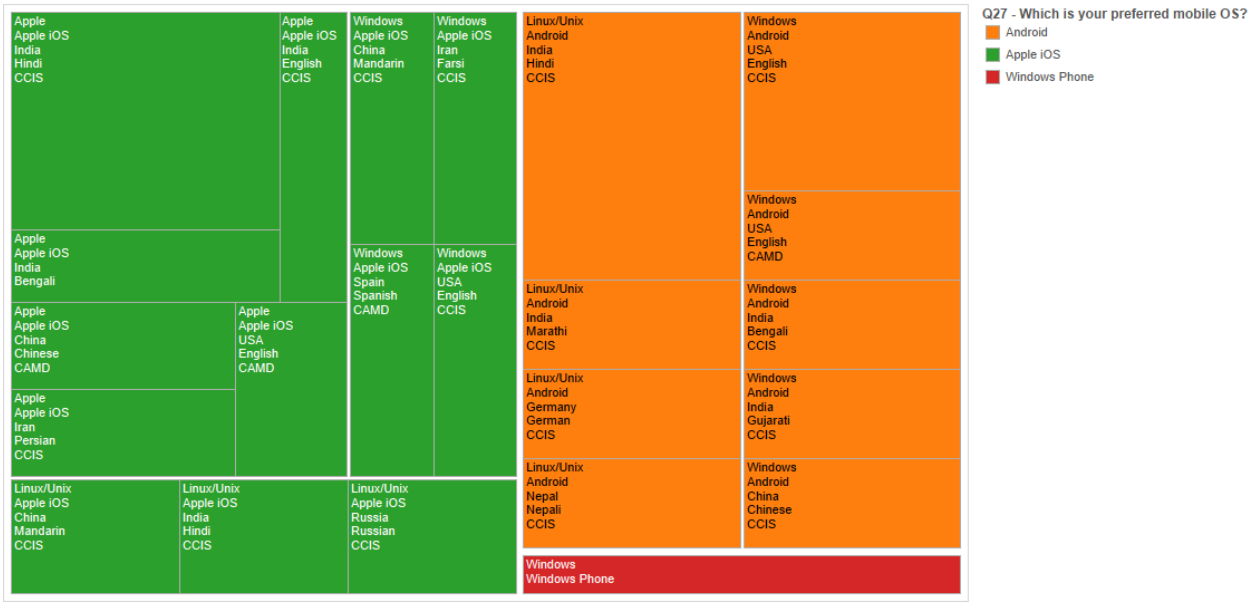
i. Correlation between bedtime and coffee consumption



No, there is no correlation between bedtime and coffee consumption. I chose to represent the data as a scatterplot in order to look for linear correlations between the two variables shown. (I approximated a scatterplot using a line graph, since the scatterplot function requires two measures as input, and reducing my axes to measures just returned a simple count of entries). Each piece of data is represented using the x and y position (channels) of a single point (mark) on the plot. Unfortunately, I was unable to find a way to make Tableau use half hour increments rather than hour increments, so the data is binned to the hour rather than the 30 minute period. It is easy enough to see that there is no linear relationship in the data from a simple plot, but I also added a trendline to get a quantitative value. The R^2 value is 0.007 (I could not get the trendline equation or metrics to plot except in an interactive tooltip, so it is not included in the graph).

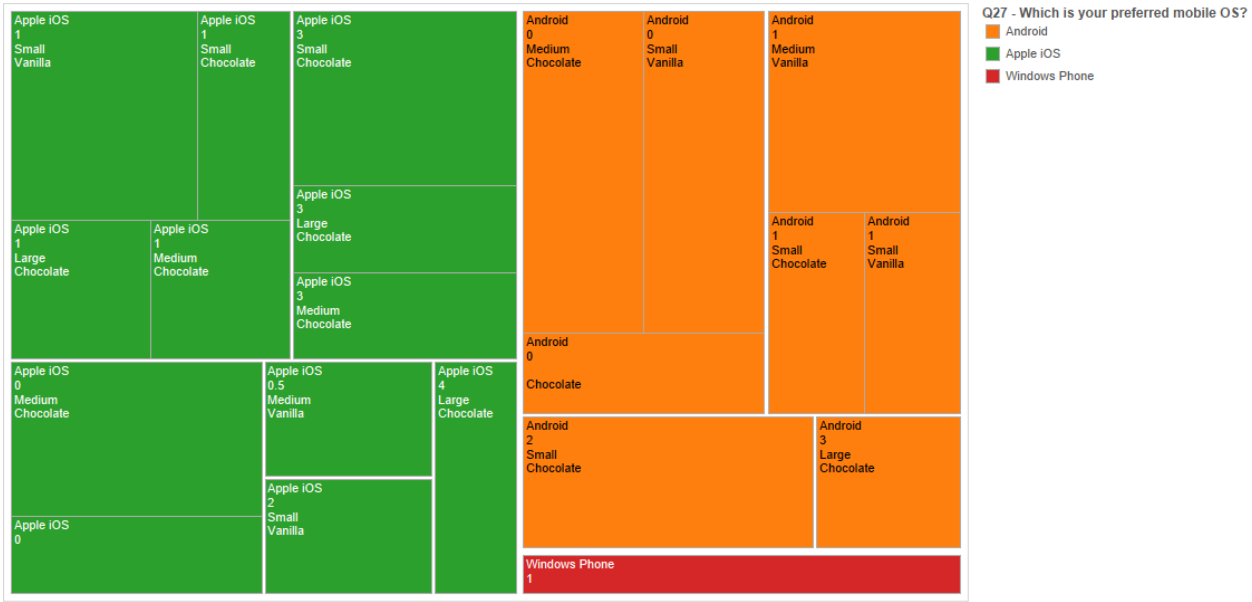
ii. Typical profile for mobile phone with Android vs. Apple OS

Sheet 9



Q26 - Which is your preferred OS?, Q27 - Which is your preferred mobile OS?, Q22 3 - Country:, Q23 - What is your native (first) language? and Q35 - In which college are you enrolled for your degree program?. Color shows details about Q27 - Which is your preferred mobile OS?. Size shows count of Q27 - Which is your preferred mobile OS?. The marks are labeled by Q26 - Which is your preferred OS?, Q27 - Which is your preferred mobile OS?, Q22 3 - Country:, Q23 - What is your native (first) language? and Q35 - In which college are you enrolled for your degree program?. The view is filtered on Q27 - Which is your preferred mobile OS?, which keeps Android, Apple iOS and Windows Phone.

Sheet 10



Q27 - Which is your preferred mobile OS?, Q29 - How many cups of coffee do you drink per day?, Q32 - What size cup of coffee (or coffee/expresso-based beverage) do you typicall... and Q30 - Chocolate or vanilla?. Color shows details about Q27 - Which is your preferred mobile OS?. Size shows count of Q27 - Which is your preferred mobile OS?. The marks are labeled by Q27 - Which is your preferred mobile OS?, Q29 - How many cups of coffee do you drink per day?, Q32 - What size cup of coffee (or coffee/expresso-based beverage) do you typicall... and Q30 - Chocolate or vanilla?. The view is filtered on Q27 - Which is your preferred mobile OS?, which keeps Android, Apple iOS and Windows Phone.

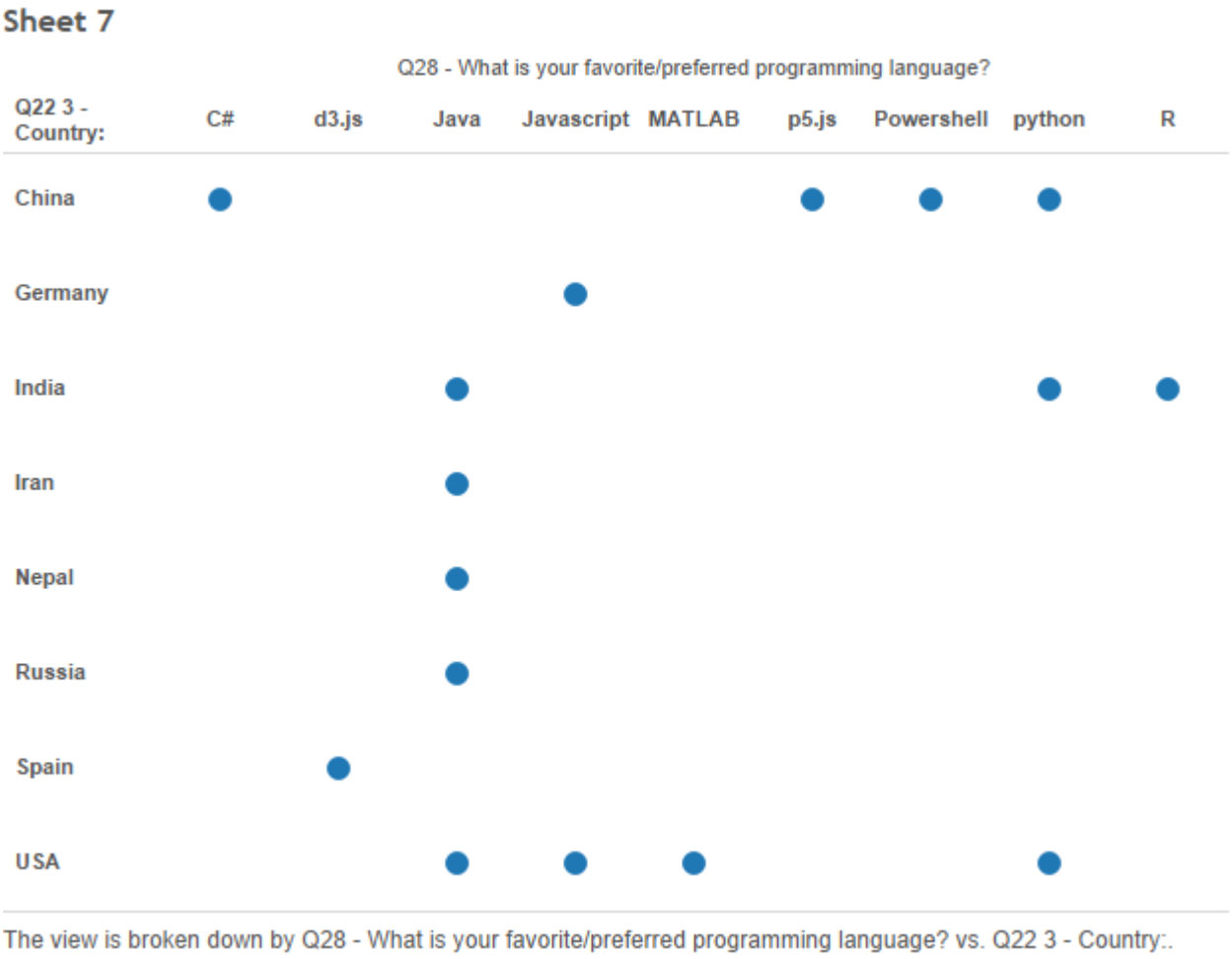
No, I do not see a particular profile for users of different mobile phone systems. I found it extremely hard to visualize answers to this question in any meaningful way in Tableau. There are so many categorical variables (and the segmentation between variables is so fine) that it is very difficult to show more than a couple at a time in a single plot. It might be possible to do a side-by-side comparison for all of the variables using small multiples on the same x axis, but that seems even more difficult to aggregate mentally.

The treemaps above encode the categorical variables as areas. The largest divisions are the mobile phone systems, and I reinforced this division using color to help differentiate between the major zones quickly. I chose to split the group into two graphics, with related variables plotted in each one. I skipped variables such as US State where a majority of respondents did not provide a value. I chose to use text labels for all of the other variables encoded in a graphic, to simplify interpretation.

In the end, I included operating system, programming, country, first language and college in the first graphic, and variables related to taste preferences or consumption patterns (choc/van, coffee, etc) in the second. The class is pretty evenly split between Apple iOS and Android, with a slight preference for Apple. There does not seem to be a strong relationship between OS and any of the other variables plotted.

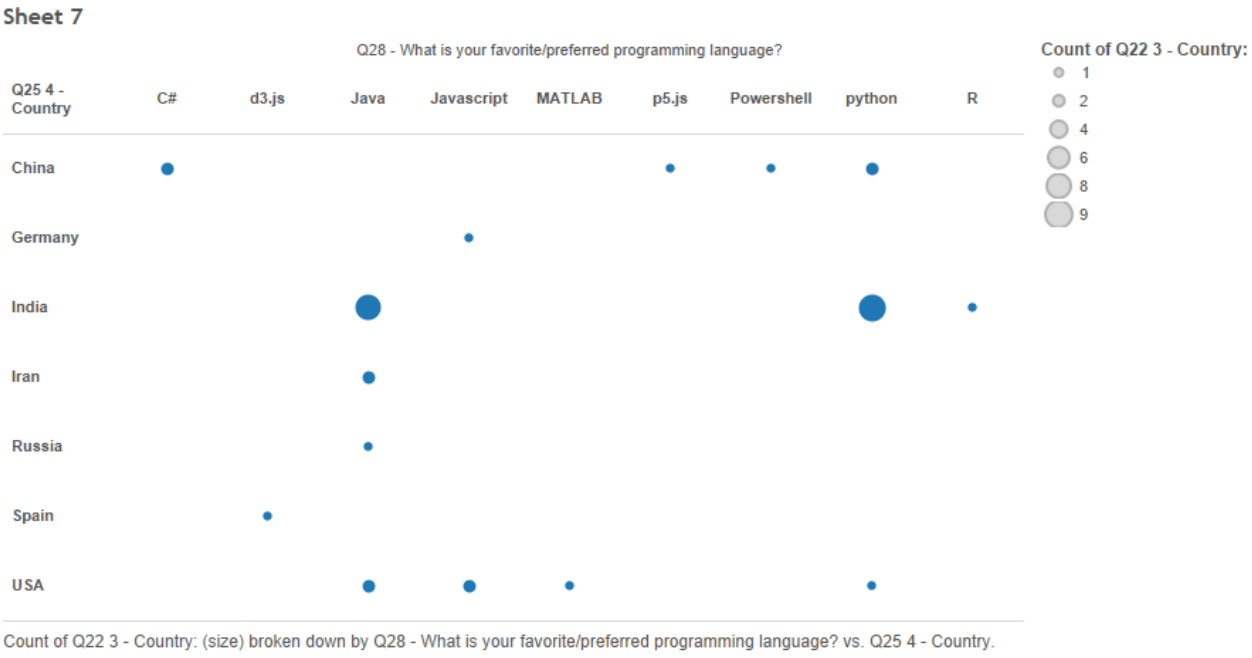
Honestly, I think that for this data, a table would be clearer and easier to read. I attempted to construct a table in Tableau, but it insisted on interleaving the variables rather than presenting them in series in a structured table layout. If both axes were quantitative, then a small multiples series of correlation diagrams could also be helpful.

iii. Do people from the same country prefer the same programming languages?

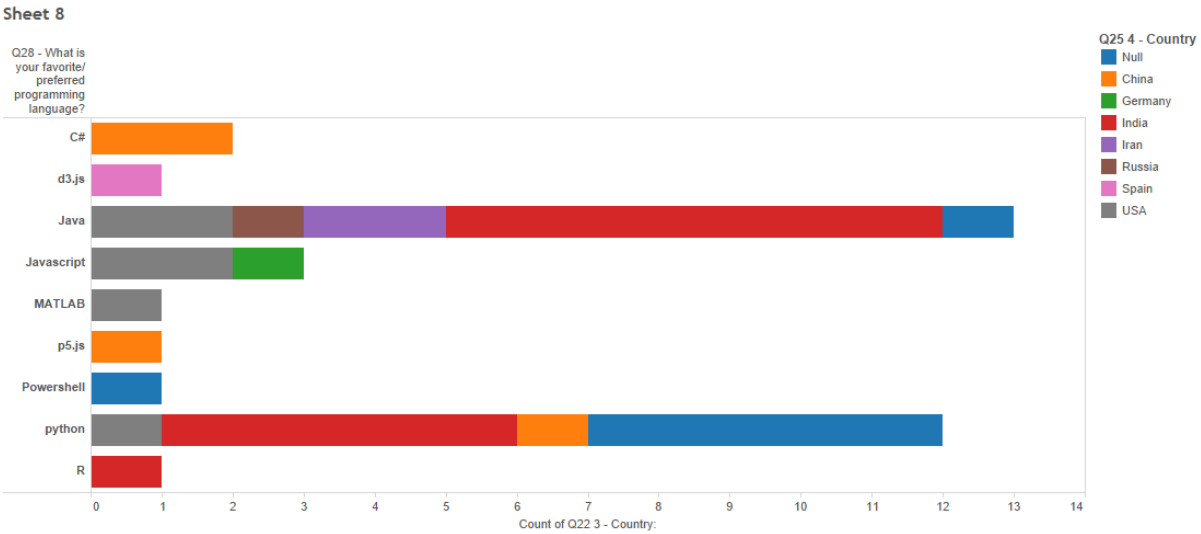


No, there is no connection between country of origin and preferred programming language. I felt that this data was better represented by a tabular plot, because it has two categorical axes that are not readily converted to ordinal values required for a quantitative/continuous scatterplot axis. The dataset is small enough that it is possible to see trends fairly quickly in a table, so I felt that this was sufficient.

I made two versions of this plot; the second is shown below. I am ambivalent about this one. I chose to scale the size of the marks using the number of people from each country that preferred a particular language. I did this because I didn't think that the first graphic did a good job of accounting for differences in sampling size for the different countries – we have only one Spaniard but many Indians in the dataset, and so it is somewhat misleading to represent those different numbers of respondents as a single point on the graph. The second version was meant to ameliorate this problem by showing the number of people from each country who liked a particular piece of software. It is much clearer to see the weighting in the second graph than in the first. However, this graph also requires the user to keep track of how many total people are from each country in order to have a sense of whether a big spot means a larger population or a strong preference (looking just at the size of a single dot, it would be easy to say that Indians prefer Java).

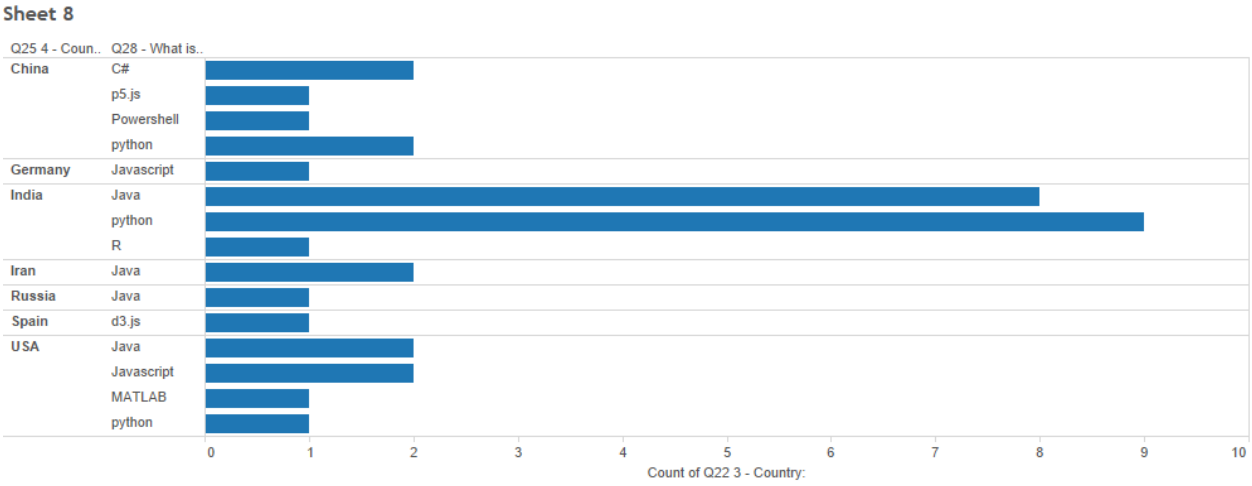


I also tried representing the data as a bar chart, with colors to represent the different countries. This isn't ideal, since there are so many categories to encode, but both axes have a large number of categories to represent. The good thing about this version is that it shows the total population of users for a particular language, encoded as the position of the end of the bar chart. It also shows the national origins of the users for each language, which is helpful. While it is possible to compare across bars using colors, it is difficult to do so using this graphic. It is easy to see that a large number of Java programmers are Chinese, but harder to see how many Chinese users prefer Java.



Count of Q22 3 - Country: for each Q28 - What is your favorite/preferred programming language?. Color shows details about Q25 4 - Country.

In the end, I think I prefer a simpler side-by-side bar chart by country for the different programming languages to answer this question. Each country is represented by a series of bars, which show the programming languages used by people from that country. It appears that Indians tend to use Java or python and everything else is reasonably spread out among countries. This graphic does not make it easy to compare how many users of a particular language come from each country; that is better encoded in the previous examples.



Count of Q22 3 - Country: for each Q28 - What is your favorite/preferred programming language? broken down by Q25 4 - Country.

For a first glance overview, I prefer the tablular view in the first example shown. It readily shows the distribution of languages and countries, and it's possible to get a sense of incidence level. Since most of the data is in counts of 1 or 2, this seems like a pretty reasonable reflection of the data. To get into the specific details of preference within a country, I think the final bar chart is best.

iv. Is there a relationship between a student's college of origin and their preferred programming languages?

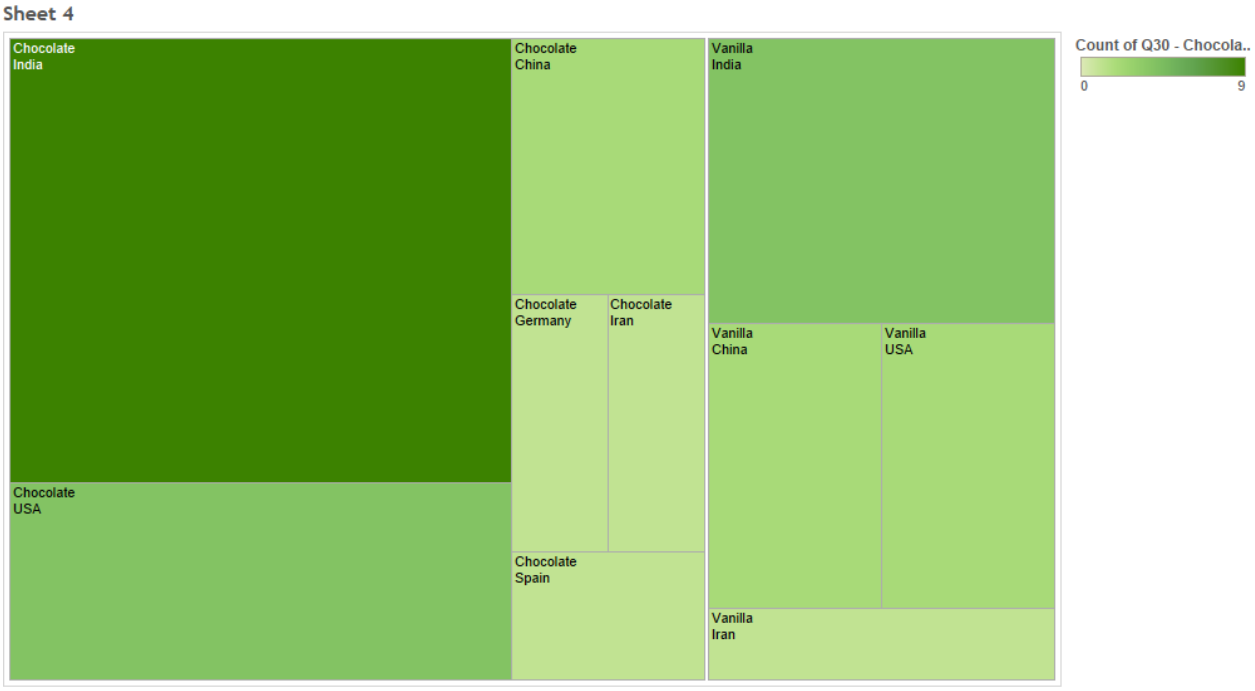
Sheet 3

Q28 - What is your favorite/preferred programming language?	Q35 - In which college are you enrolled for your degree program? CAMD	CCIS
C#		●
d3.js	●	
Java	●	●
Javascript	●	●
MATLAB		●
p5.js	●	
Powershell		●
python		●
R		●

The view is broken down by Q35 - In which college are you enrolled for your degree program? vs. Q28 - What is your favorite/preferred programming language?.

Yes. Students from CAMD tend to focus on Javascript and web-based languages, with a little Java thrown in (likely due to Processing). Students from computer science do not have contact with d3 or p5.js, but have a wide range of preferences more traditional programming languages. This simple graph has the same problem of not representing count as the previous problem, but I think this question is adequately answered with a binary view.

v. What is the relative preference for chocolate and vanilla by country?



Q30 - Chocolate or vanilla? and Q25 4 - Country. Color shows count of Q30 - Chocolate or vanilla?. Size shows count of Q30 - Chocolate or vanilla?. The marks are labeled by Q30 - Chocolate or vanilla? and Q25 4 - Country.

This graphic uses rectangular area to encode the number of people from each country that prefer chocolate or vanilla. There is also some encoding by position, since chocolate and vanilla are grouped in a structured way to make the differences easier to see. I used color to reinforce the count values, but I think it could probably be omitted, as it is redundant. In general, it is possible to see that there are more chocolate lovers than vanilla lovers (about a 2:1 ratio). More Indians and Americans prefer chocolate over vanilla, but the Chinese and Iranian students are equally balanced between the two flavors. I think that this partition layout does a decent job of representing the data, and allowing simple comparison between the two categories. Even though the flavor labels are redundant, I found them helpful for reinforcing the distinctions between the different classes.

vi. Do citizens of some countries drink more coffee than others?

Yes. I thought this was an interesting and straightforward visualization, representing two categorical axes in a table format, again with points as marks and using both the x/y position and size channels. Chinese and American students tend not to drink as much coffee as Indian students, and the other countries are scattered over the range (given only a single data point, I'm not sure it's worth looking for "trends"). Since the number of cups of coffee is a numerical axis, using the size to represent the count of students from a particular country leads to fewer conflicts/confusions than it did in example iii. Here, I am also not attempting to compare across a row of the table, so it is less important to be able to judge the total number of people in order to make informed inferences.

7) Recreate in c3.js