CS/INFO 3300
Project 1 Description
Participants: Alexis Vinzons, Bernardo Gampel, Olusola Adegunloye

Part A. Description of the data

For our project, we used two different datasets: the Global Innovation Index (GII) 2017 (https://www.globalinnovationindex.org/analysis-indicator) and a dataset classifying countries by its political regime on a scale from fully autocratic (-10) to fully democratic (+10) over the past 200+ years (https://ourworldindata.org/democracy#note-8). In short, we were interested in better understanding the impact of a country's political regime to the different components of innovation in that country.

Because the way the Innovation Index is calculated changes year by year-as well as the availability of the data- we decided to base our analysis solely on year 2017. Additionally, the political regime data was only available until year 2015, so an underlying assumption for our analysis is that worldwide political regimes in 2017 have stayed relatively the same as in the year 2015.

The only variable we cared about from the first political regime dataset is the political regime score (from -10 to 10) whereas in the GII dataset we are interested in extracting a couple of different variables. In particular, we are interested in extracting the GII values as well as the Sub-Index values for each of the 127 countries listed in the dataset. The total Global Innovation Index in simply a weighted average of the seven different Sub-Index values.

Both data files were .csv formatted. The political regime dataset was formatted in a "parsing-friendly" way when using d3.csv- so that we did not have to reformat the file in any way. On the other hand, the GII dataset was formatted weirdly (lots of blank spaces) so we changed it around to make it easily readable by d3.csv. More specifically, we deleted all reported metrics that were not of interest and transposed the data so that the countries would appear in the rows and the metrics of interest in the columns.

Finally, we had to aggregate the two datasets and only select the data that would be useful to our analysis. The first function "parseInnov" was used to create a list of objects- each containing the name of the country and the fields that we were interested in extracting for our analysis. The second function "formatPr" was used to create a hashmap with the countries' political regime score in 2015 only. Lastly, the function "joininnov_x_pr" merges the data from the previous to functions into a list of objects in O(innovation_data.length) time complexity. Each object has the name of the country with its corresponding political regime score and innovation data.

It is important to note that even though our individual datasets had 168 (political regime) and 127 (innovation) countries listed, the aggregated data only had 119 countries. This happened because some countries' names did not match and were very hard to cross reference.

Part B. Mapping from data to visual elements

We plotted Global Innovation Index (GII) value by political regime, where each country is a svg circle of the same size. We used a linear scale to position points along the x and y axes. Points further to the right of the plot represent democracies while those on the far left represent autocracies. We used a color scale to further distinguish the political regime- where blue represents democratic and red represents autocratic, a standard color symbolism for political regimes. The points are also slightly opaque, to render clusters more salient. Jittering was also added for data points greater than 6 because that is where most of the clustering happened. While we considered sizing the circles by population we decided to size each country the same. In doing so, we draw more attention to the quantity of countries at a given index or political regime value, rather than distracting the reader's attention to more populous countries.

The decision to highlight certain countries were based on noticeable outliers, maximum and minimum GII, as well as familiarity to the audience (e.g. those in United States).

The political regime data set describes countries with a score [-10,-6] as autocracy, [-5,0] as closed anocracy, [1,5] in open anocracy, and [5,10] as democracy. We used opaque rectangles to highlight the differences in these categorical distinctions in the second plot. We also chose to specifically highlight those with GII values above the mean to propose that, while more democratic countries score higher on GII, there still exists some autocratic countries with high GII scores.

The last plot shows the differences in the averages of sub-indexes between democratic and autocratic nations. We determined that stacked bar charts would allow for easy comparison across political regime, as well as across sub-indexes. A linear scale was used to align the width of the bars to the svg while band scales were used to do the height of the bars and the y positions. Finally, we maintained the blue-red color scheme for political regime.

Part C. The story

The first plot scatter plot shows a U-shaped pattern when comparing the total Innovation Index for each country against their political regimes. Interesting enough, this pattern contradicts our initial hypothesis that the higher on the political spectrum (from autocratic to democratic) a country is, the higher its innovation index would be in comparison to countries that are lower on the spectrum. What our data instead shows is that countries that are well-defined as either an autocracy or a democracy are likely to perform better that countries that are mixing democratic and autocratic features (aka. Anocracy).

We then decided to partition the data by ignoring the anocratic countries and instead focusing on the comparison between high performing autocratic and high performing democratic nations. Our second plot shows that a greater percentage of democratic countries fall above the average innovation index when compared against autocratic countries. Furthermore, the fact that countries like Qatar and the United Arab Emirates are

so high on the index can likely be attributed to the fact that they are two of the largest oil producing economies in the world, rather than to the fact that they are ruled by autocratic regime

We then decided to partition the data by ignoring the anocratic countries and instead focusing on the comparison between high performing autocratic and high performing democratic nations. Our second plot shows that a greater percentage of democratic countries fall above the average innovation index when compared against autocratic countries. Furthermore, the fact that countries like Qatar and the United Arab Emirates are so high on the index can likely be attributed to the fact that they are two of the largest oil producing economies in the world, rather than to the fact that they are ruled by autocratic regimes.

Taking it a step further, our final piece of analysis (third plot) aims to answer the question: what are the main factors contributing to the observed differences in the innovation index between autocratic and democratic nations? While looking at the results in the third plot, it immediatly stands out how some sub-indexes' average scores differ by more than 20% whereas other sub-index differences are well below 10%. One might hypothesize, for instance, that the reason why creative outputs differs by so much is because autocratic regimes would rather stay under the radar, and isolated from the foreign world. On the other hand, they still need to be fairly developed on infrastructure in order to efficiently run their country, which is why we see less deviation from the average infrastructure score of the democratic countries.

Visuals that we try but resulted in little to no insight:

1. Bar plots displayed in sub-index ranking order: we expected a pattern resembling a cosine function but instead found no particular interesting trend.

2. Individual countries' sub-index composition comparisons: we believed that the sub-index composition of high performing autocratic nations would be very different from high performing democratic ones; however, they turned out to be pretty similar.