Angela Chen & Alex Heilgeist
HCDE 411
February 17th, 2017
A3 - Exploratory Visual Analysis Report

Introduction

Visuals and imagery are frequently used to prompt immediate cognitive recognition of the information they attempt to convey. They surpass textual content when it comes to efficiency in understanding and processing information, which is why they are key components in presenting data. Information visualization is a tool used to tell a story with purpose, without overwhelming viewers with meaningless and obscure text. It can be used to answer questions, form relationships, draw comparisons, inform the public, etc. It is meant to communicate in a way that our brains are cognizant of and hard-wired to interpret by using colors, shapes, and other visual cues.

There is a plethora of data in the world that is readily available to us, waiting to be turned into a meaningful, shareable story. In this report, we formulated three questions based off of data extracted from the World Development Indicators data set. We will be discussing the visualizations that we created, which look at the correlation between GDP and adult literacy, electric power consumption over time, and prevalence of HIV worldwide. Our intent was to create visualizations that can effectively and efficiently answer these proposed questions and successfully turn abstract numbers and words into clean, relevant visual data.

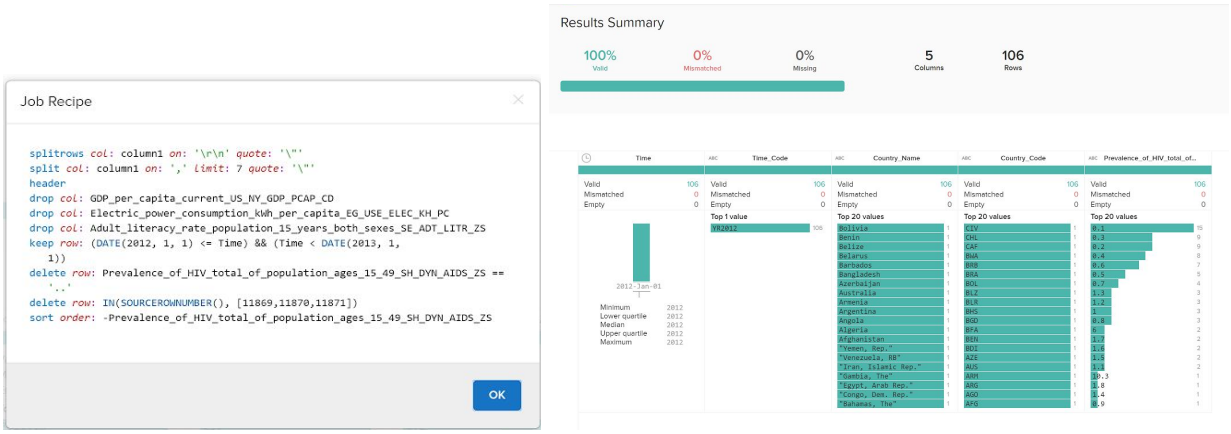
Data Profile

Where: World Development Indicators Site
Size: 264 Rows, 4 Columns, 723 kb
Types: Electricity Consumption, Adult Literacy Rate, Prevalence of HIV, GDP per capita
Quality: Large gaps in literacy rates, Smaller gaps in Electricity, GDP per capita, HIV. Particularly recent data is scarce, leading to the usage of 2015 data and earlier, such as 2012 data.

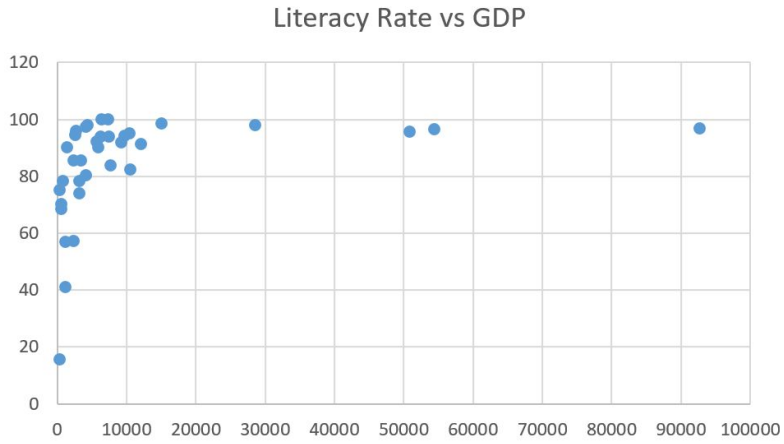
Q1. How did a country's amount of wealth (in GDP) correlate with adult literacy in 2015?

We formulated this question in hopes that we would see some correlation between the health of a country's economy and the percentage of the population that is literate. We hypothesized that the data would trend upward, showing that as a country's GDP increased, so would the literacy rate. We assumed that countries with more wealth and a healthily stimulated economy would have more access to certain educational resources, tools and infrastructure that would help current and future generations to learn to read and write.

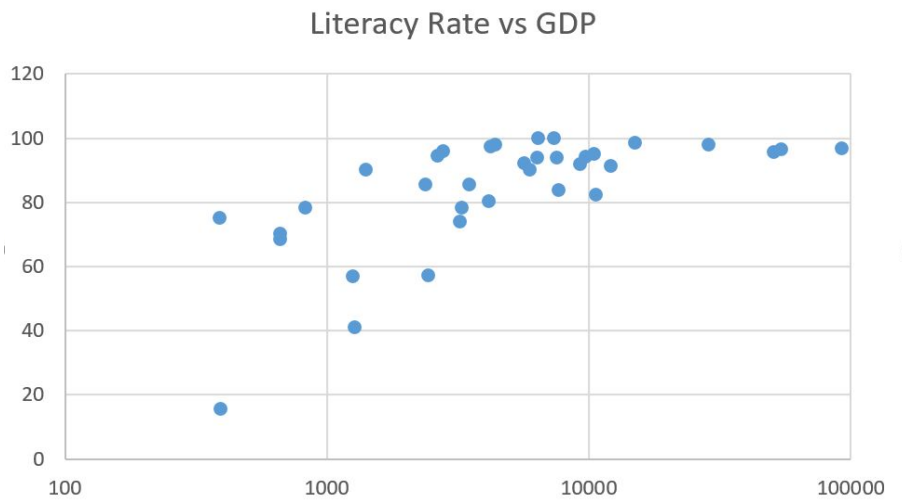
While creating our data extracts from the World Development Indicators data set, we found that a majority of countries did not or hardly had any reports on adult literacy rates in the last ~25 years (1992-2016). We were surprised that even some of the more developed countries such as Japan and the United States did not have data on their country's adult literacy rate. In Wrangler, we had to manipulate the data to get rid of all the countries that did not have information about literacy so that we could focus on the countries that did.



Because of the lack of data points, we considered looking at the literacy rates over time but felt comparisons across years would lead to an inaccurate comparison between countries. Thus, we decided to only focus on a single year to change the fewest number of variables as possible. We took an extract of the literacy rates and GDP (in current $US) from 2012 and created a simple scatter plot to see if we could identify any visible trends. 2012 was selected because many of our other visualizations had high quantities of data in the year 2012, and we wanted to remain consistent.

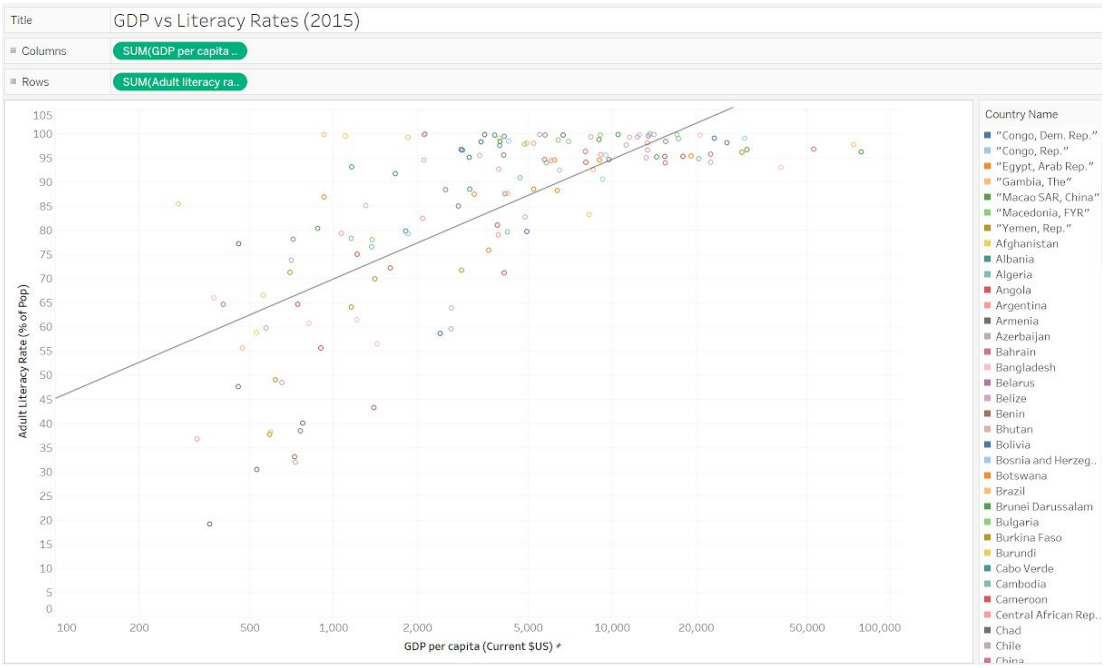


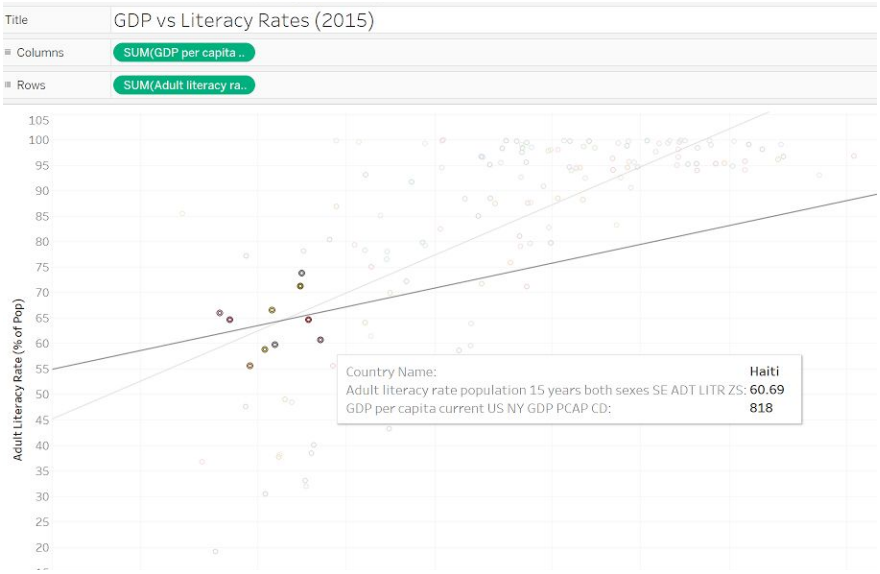
While the scatter plot showed an increasing function, the scaling on the x-axis, which represents the GDP, caused occlusion in the plotted data points, making it difficult to see a definitive pattern in the data. We played with the scaling a little bit and changed the GDP scale from a linear scale to a logarithmic scale, which distributed the data points across the plot and reduced the amount of occlusion in the set.



This visualization draft better maintains Graphical Integrity by presenting the value relationships accurately. The size of the visualization better fits the data by utilizing the available space of the scales more effectively than the linear scale, with its clustering to the left, was able to do. After analyzing the data and the relationship between the two dimensions, the data for 2012 seemed too insufficient to accurately answer the question that we had originally posed. As a result, we went back to the full data extraction and shifted our focus to the year 2015, despite the fact that it sacrifices consistency with our other visualizations, because the report for literacy rates was significantly more complete.

With more data points, we were able to see a stronger, more conclusive correlation between the two dimensions, adult literacy and GDP. Using Tableau, we created an interactive scatter plot to highlight important information about the data and present more context to amplify cognition for the viewer.





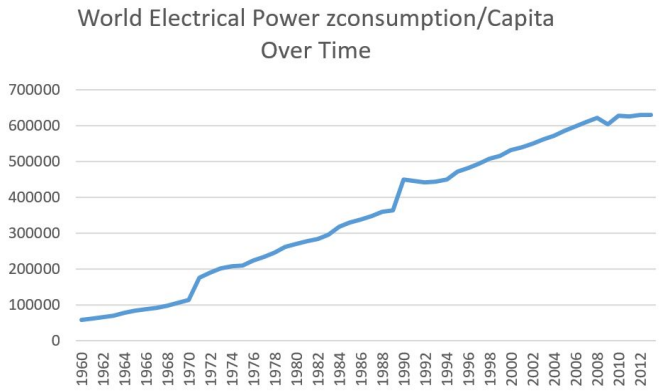
This scatter plot of data points effectively shows that there is a positive correlation between GDP and adult literacy rates. The visualization accurately presents the relationship between the two dimensions to maintain its Graphical Integrity and is carefully scaled so that it supports perceptual monitoring for the viewer. The logarithmic regression line that is plotted on the data is meant to enhance pattern detection. Though the scale begins at 100 instead of zero, we decided to make that tradeoff of potential distortion for the ability to amplify cognition. The visualization also utilizes focus and context as well as details-on-demand to give the viewer additional, specific information if they choose to take a closer look at a smaller set of data points. In all, the visualization successfully achieves its purpose in showing the correlation between GDP and adult literacy rate.

Q2. How does the world's Electric Power Consumption change over time?

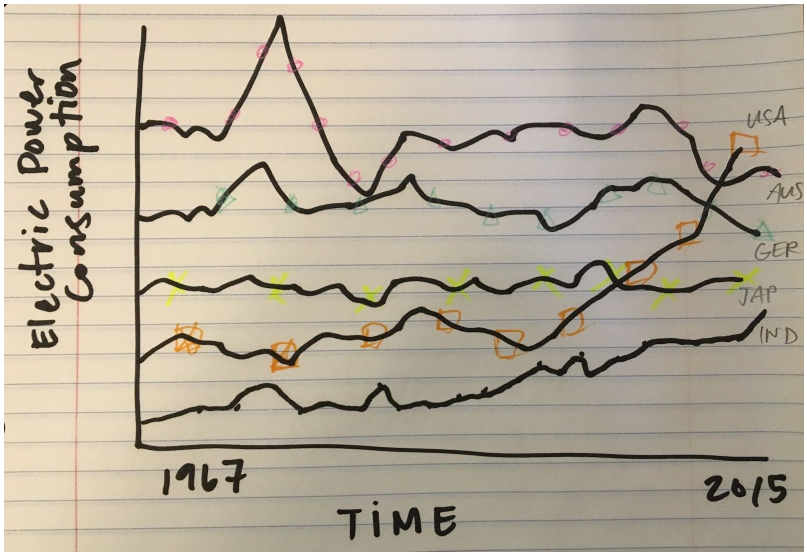
Unlike the adult literacy data, most countries have reported data about their electric power consumption (kWh per capita), which should allow for richer analysis on the change over time. Given that the burning of fossil fuels over the decades has had a dramatic, negative impact on the environment, regions of the world are looking for alternative sources of energy that are less harmful to the planet. People have turned to electric energy to reduce the emission of greenhouse gases into the atmosphere by using this renewable energy to power homes, buildings, cars, vehicles, etc. We think it is reasonable to assume that the world's electric power consumption has increased over time and most likely, increased dramatically in the past few years as people are looking for more opportunities to practice sustainability and stop climate change.

In the data extraction from the World Bank, we looked at worldwide electric power consumption over the last fifty years. In Wrangler, we eliminated the countries that did not have data in the fifty-year span and were left with 106 countries with viable data. Because of the time span and the number of countries we were looking at, there was an overwhelming amount of information

presented. We wanted to synthesize the data by aggregating each country's consumption into a sum for the year. We plotted it on a simple line graph to view the change-over-time relationship with power consumption on the y-axis and time on the x-axis.



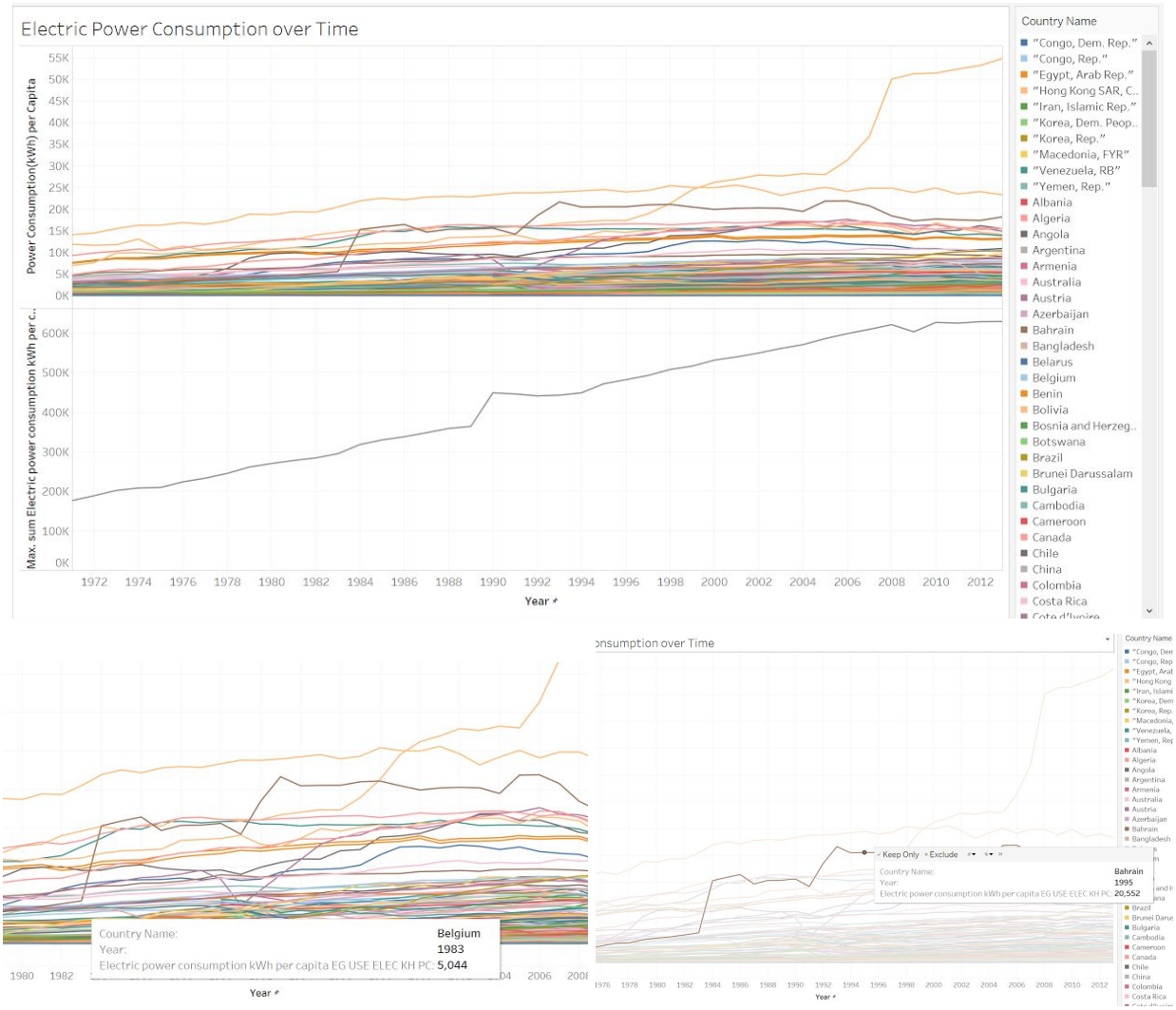
The graph showed exactly what we assumed would happen with power consumption over time – a steady increase in kWh/capita over the last fifty years. Though this is a simple, straightforward representation that can easily answer the question we proposed, we thought it would be beneficial to show how specific countries have progressed in their consumption over the last fifty years. Every country may utilize electric power differently or have limited resources to extract this type of energy and as a result, have completely different output from others. Some countries might have had a spike in energy consumption while others might have faltered around a constant value. We decided it would be interesting to see how individual countries utilized electric energy and made a quick sketch of how we wanted to visualize the data.



Similar to the previous line graph, electric power consumption would be on the y-axis and time would be on the x-axis with each line being a selected country. The sketch uses dummy data to show how we visualized the data using multivariate analysis. With this visualization, we can see

how the countries compare with one another as well as how each country has changed in its electric energy consumption over time. Each country could be encoded with a different color to make them more distinguishable. We could also include the worldwide consumption line that we plotted originally for an additional comparison opportunity.

On Tableau, we experimented with the visualization that contained the line graphs for every country with available data as well as the single line for worldwide energy consumption over time. We debated whether to keep just the visualization with all the encoded countries or to display both that and the worldwide electric consumption plot. In the end, we decided to keep both visualizations so that the viewer could compare the overall trend of worldwide consumption with the individual countries. Though they are on different y-scales, the purpose is not to look at specific numbers but to compare the orientation of the lines.



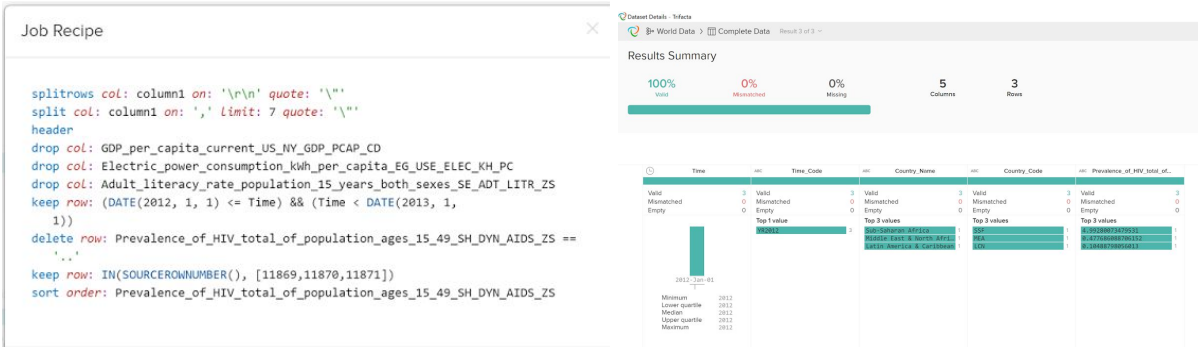
This question could have been answered by simply providing the single line plot that represented total electric energy consumption over the last fifty years. However, by simultaneously providing a breakdown of electric consumption for each country, the data

becomes more interesting and viewers can draw comparisons and form relationships with the additional information. It is easy to see from the line graph on the lower half of the visualization that worldwide electric energy consumption has increased in the last fifty years. The axes are clearly labeled and scaled properly to increase Graphical Integrity and Excellence. It helps to amplify cognition and processing because of its simplicity, which makes it incredibly easy to determine any patterns or trends. The top half of the visualization merely augments the data in a way that supports perceptual monitoring. Not only are users able to see how specific countries contribute to the total consumption over time, but it also employs focus and context interactions to view details while preserving the surrounding context for orientation. Finally, the visualization provides details-on-demand upon hovering over a specific line. With everything that is integrated into this visualization, we were able to go above and beyond in answering the proposed question.

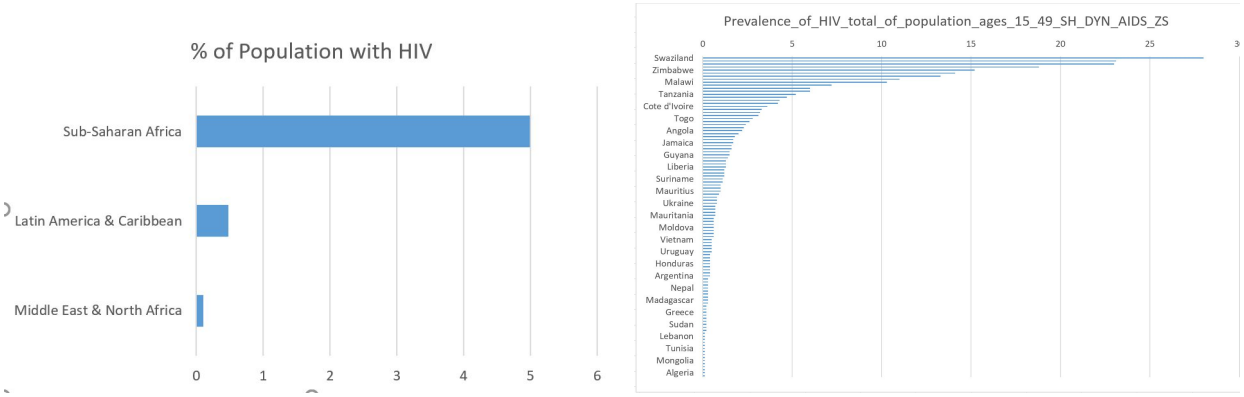
Q3. What is the prevalence of HIV in the countries of the world?

HIV is a troubling and unfortunately, widespread health problem, especially in areas around the developing world. To many of us, HIV is a distant concept; something that many of us do not deal with on a daily basis, and only concern ourselves with in passing when it comes to avoiding it such as not sharing needles and practicing safe sexual behaviors. Even if it does become an issue, modern medicine has made it more into an inconvenience than a death sentence. However, for many of the less fortunate around the globe, HIV is an everyday problem, with staggeringly high infection rates in some places, and a lack of modern medicine that would save lives. We hypothesized, due to the general reputation of HIV, that it would be predominantly found in Africa, with other developing countries being found to have high rates as well.

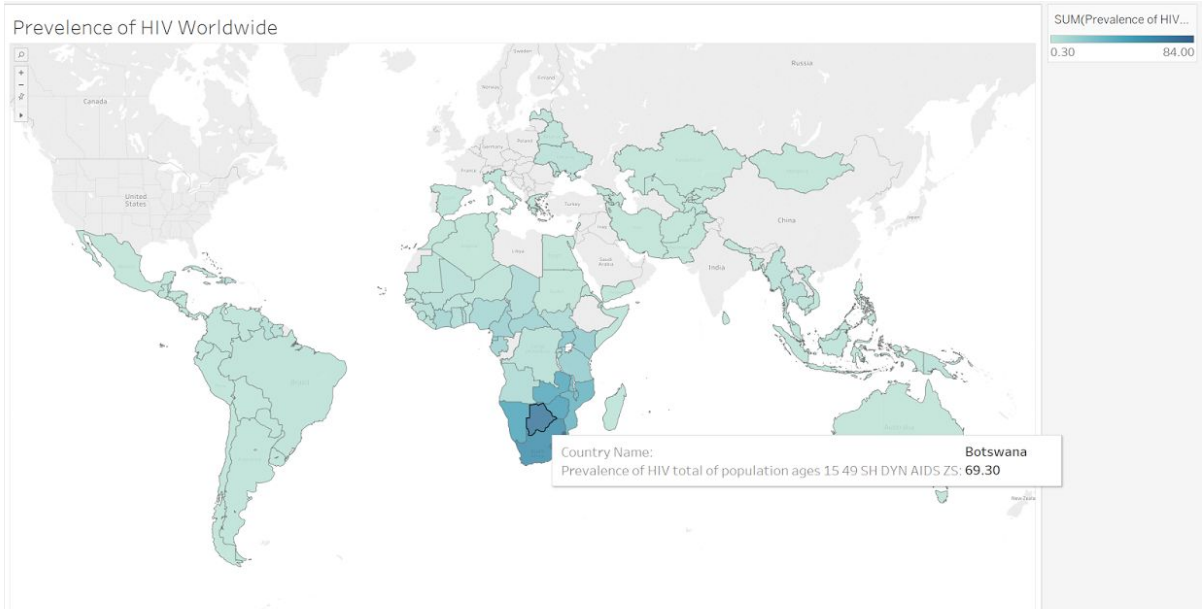
Prevalence is the number of cases of a disease that are present in a particular population at a given time. When analyzing the data in Wrangler, we found that the data for HIV prevalence appeared to be mostly confined to developing nations. Many developed nations did not have points in the dataset, such as the U.S or European countries. However, the data did contain most of Africa, south America, Oceania, and Central Asia, allowing at least a comparison of infection rates between developing nations in different locations around the globe. We eliminated blank data to ensure the integrity of our visualization.



Our end goal for our visualization was to create a geographical map showcasing HIV prevalence globally. Initially, without powerful tools such as Tableau, simple graphs were the best we were able to do. We experimented with multiple ways of representing the data; one by region in general, and one by specific countries. We eventually chose to focus on the graph of specific countries in order to provide more granular data as well as to simplify the display of the data later on in a geographical representation, which displays by country, not region.



As consistent with our hypothesis, the highest prevalence of HIV was by far in Africa, though specifically, Sub-Saharan Africa. In this preliminary representation, the horizontal bar graph has a country axis, and a % axis, displaying the percentage of the population infected with HIV. As stated, we found that these graphs, while accurate, failed to give a sense of location for the data. The simple regional graph was too basic and simplified, while the country graph found it difficult organizationally to showcase a country's location relative to another. To address this issue, we went to Tableau, in order to make a map based visualization



By utilizing a map based visualization, we can see much more clearly the clustering of high HIV infection rates in Sub Saharan Africa. The visualization uses shades of color to represent the

severity of infection rates. Missing data was displayed as a neutral grey. Different colorations tended to meld missing data with low data, and were thus found not suitable for our purposes, as the misleading nature of the visualization would sacrifice Graphical Integrity. The visualization had additional interactive functionality, such as details-on-demand that presented the name of the country as well as the exact HIV prevalence when hovered over. Countries can be group selected to isolate their data from the rest of the visualization as well. These features increase Graphical Excellence by enabling user understanding through interactivity.

Reflection

Overall, our data looks at a number of metrics that describe the difference in quality of life for the developed versus developing world. We used a number of tools to produce our final visualization. The World Development Indicators website's interface allowed us to parse our data set down to a manageable size, and focus on the specific variables we used. However, the provided tool could not properly cull the dataset as desired, and was generally clunky. Trifacta Wrangler was our tool of choice for taking the raw data from the WDIS, and refining it into a form that could be represented in a visualization. Wrangler was powerful, and allowed fine grain control over the data. It easily identified bad data, which made it useful for cutting that data out. However, the application was slow, and took a long time for actions to be completed. A slow response time is annoying to work with, especially if, like us, the data needed multiple iterations of refinement to ensure the greatest quality. As for our visualization, we used Tableau, a powerful tool that intuitively created visualizations with a few clicks. Tableau had a bit of a higher learning curve and certain ways of representing data were not immediately intuitive. But after a brief trial run, the visualizations produced were highly interactive, as well as very aesthetically pleasing, and required minimal effort. We look forward to seeing the future of visualization tools, as they are certainly powerful. But tools such as Wrangler need optimization on the code side to function at an acceptable speed.