Tafara Makayi

Jeff Smolinski

4447 Final Project

11/22/2020

**Final Project Details and Rubric**

In Tools 1, we are concerned with data cleaning and exploratory analysis. Please select a project that has enough scope for the following activities.

For the final project, you will create a Github repository for your project and tag it with the label **TOOL1\_FINAL\_PROJECT** by the due date. The github repository must have a .ipynb notebook file with output and associated code. Having output in the notebook cell is very important if your dataset is big or we won’t be able to run the notebook in a reasonable time. Also, we should be able to run your project with a Binder link. The binder link should be in the **README.md** file. Please check https://mybinder.org/ to see how to create a Binder link. If this service is down, this step is not required.

Your final report should read like a data-driven story/scientific study (data science). This is really important if you want to publish your story as a blog on the web or share with stakeholders.

Scientific publications have their own style and content requirements.

Use code cells and markdown cells to carry out your analysis. Please write the report using the following section format guidelines. You can create more sections if it is more natural to do so, depending on your project. Please write each section like a report and address the points mentioned in the following rubric. Try to make your report more enjoyable to read.

• Proper tagging of Github repository for final report as per deadlines (0.5 = 0.25 + 0.25 points)

• Dataset and motivation slide (1 points)

**The EFSA Comprehensive European Food Consumption Database**

<https://data.europa.eu/euodp/en/data/dataset/the-efsa-comprehensive-european-food-consumption-database/resource/18cce1f2-1586-4bf9-9a28-b5103aa413f6>

**World Development Indicators**

<https://datacatalog.worldbank.org/dataset/world-development-indicators>

**Health Nutrition And Population Statistics**

<https://datacatalog.worldbank.org/dataset/health-nutrition-and-population-statistics>

MAYBE - **Plotly Express build in population and GDP data by Country from 2007**

df = px.data.gapminder().query("year==2007")

* used to help feed the choropleth

How/why the dataset was collected and a description of the metadata of your dataset.

• Actual task definition/research question (2 points)

**Motivation**: Does national development impact child mortality or is child mortality a function of nutrition alone? Alternatively, are there countries which have low development index but high new child and youth mortality because of the consumption of some nutritional undesirables?

The data files were downloaded from the above links. Two of the files required conversion from XLS to CSV and CSV to CSV.

The data files were downloaded from the above links. Two of the files required conversion from XLS to CSV and CSV to CSV.

**EFSA Nutritional Data**

avg\_chronic : chronic consumption of alcohol

avg\_chronic\_bw : chronic consumption of alcohol by body weight

avg\_acute : acute consumption of alcohol

**Chronic vs Acute consumption:**

Summary statistics from the Comprehensive Database have been published for both chronic and acute consumption.

**Chronic Consumption:**

Intake statistics have been calculated based on individual average consumption over the total survey period. The average value for each subject was then considered only once when calculating the chronic average consumption and other statistics related to chronic consumption at population level.

**Acute Consumption:**

Statistics have been calculated based on every single reporting day. For example, if subjects in a population had recorded their consumption by means of a 7 day food record, the average intake of each individual over the 7 days was calculated. Acute consumption figures were calculated using each reporting day independently, and in summing eating occasions for a considered food. All days from each subject (7 days in the above reported example) were used to calculate the acute average consumption and the other statistics related to acute consumption at population level.

**Gram/Day and Gram/Kg of body weight per day:**

Indicates if the data on consumption is relative to a gram per day assumption or to a gram per kg of body weight per day.

**Mortality:**

Mortality rate, or death rate, is a measure of the number of deaths (in general, or due to a specific cause) in a particular population, scaled to the size of that population, per unit of time.

**Ploty Express World data**

Per capita gross domestic product (GDP) is a metric that breaks down a country's economic output per person and is calculated by dividing the GDP of a country by its population.

**Wikipedia GDP per Capita**

https://en.wikipedia.org/wiki/List\_of\_countries\_by\_GDP\_(nominal)\_per\_capita

What real-world problem are you trying to solve? What are the input and output of your analysis?

• Literature review (2 points)

**Establish relationship between Mortality Rates, Chronic Nutritional Indicators and National Economic Performance**

We have managed to aggregate data from multiple sources in order to create our visualizations. Unfortunately, our nutritional dataset, despite having a large amount of granular data was restricted to European countries. This has caused us to drop the remainder of the world in the short term. We have pulled together data for mortality rates, gross domestic product per person, and average chronic consumption of alcohol as an indicator of poor nutrition.

What are the input and output of your analysis?

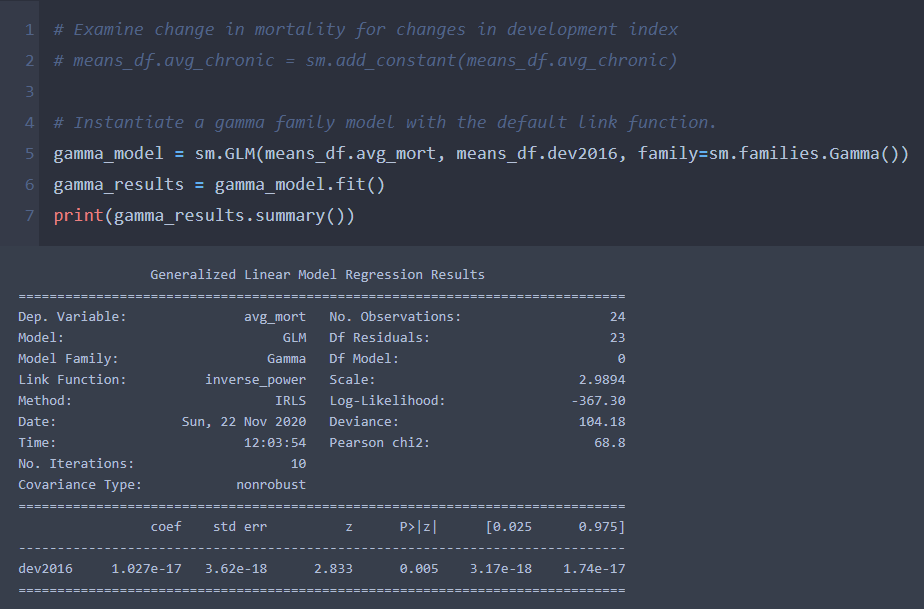
dev\_index ~ avg\_chronic + avg\_chronic\_bw + avg\_acute + diet + mortality

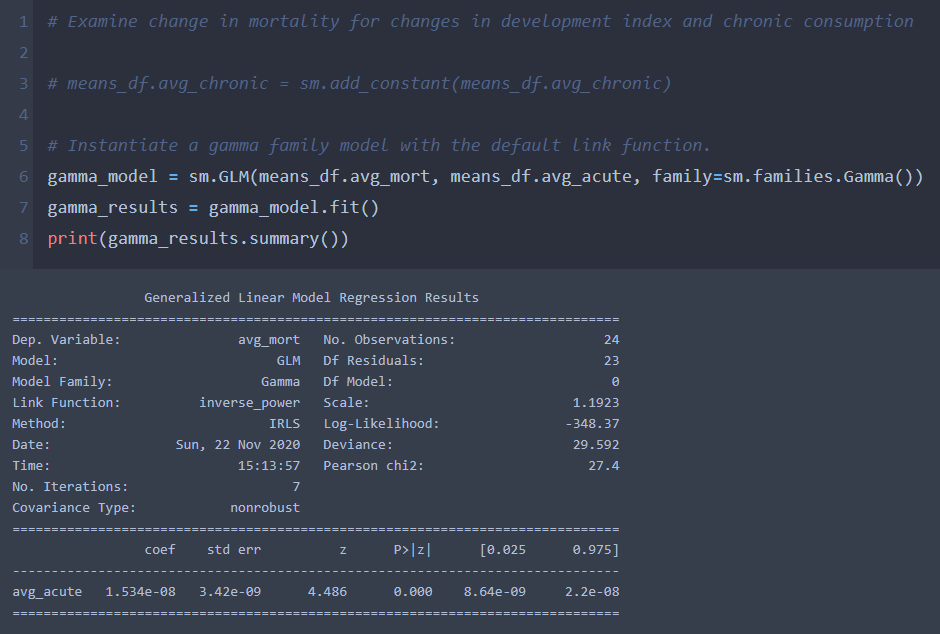
It is assumed that child and youth mortality is the output of development index, nutrition and alcohol consumption.

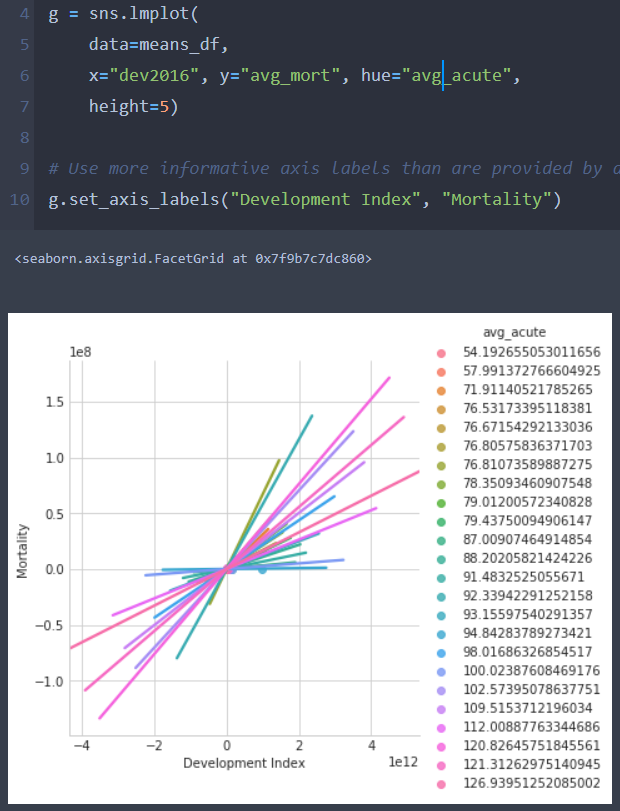
GDP\_US\_2018 ~ mortality GDP\_US\_2018 ~ avg\_chronic

|  |  |
| --- | --- |
|  |  |

The least mean square line for GDPperCap ~ Average Chronic Nutrition values virtually matches that of the Mortality numbers.







What other work has been done in this area, and how is your work novel compared to others?

• Quality of cleaning (6 points, 2 points each)

The WHO’s Nutrition and Food Safety (NFS) Department regularly addresses the burden of disease from physical, chemical and microbial hazards in food and unhealthy diets, maternal and child malnutrition, overweight and obesity. The UN also has a keen interest in the topic through UNICEF.

Both the WHO and the UNICEF’s 2020 report, like most research on the topic, established that there is a clear link between infant and young child feeding practices and household characteristics. This is now a well-accepted fact.

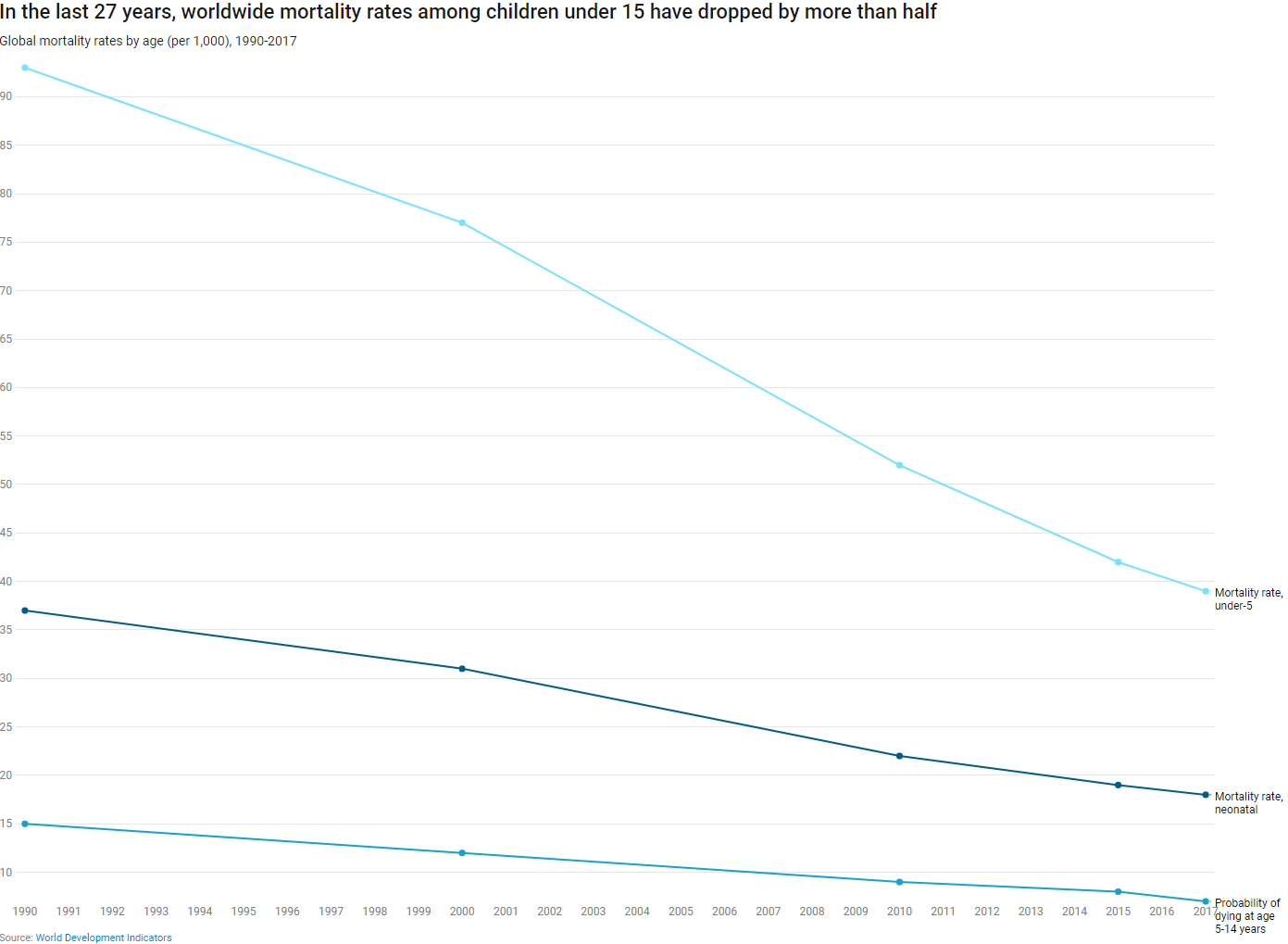
Our research tried to connect child mortality to national development index, in the face of national chronic and acute consumption of alcoholic drinks. We had to merge three (3) data sets from different sources, because the issue is often looked at in isolation. We established that child mortality cannot be explained by national development and chronic and acute consumption alone. Our findings show that these factors account for only 12% of child mortality. This implies that there are other significant factors that cause child mortality.

[**Global Nutrition Report 2020 - UNICEF DATA**](https://data.unicef.org/resources/global-nutrition-report-2020/)

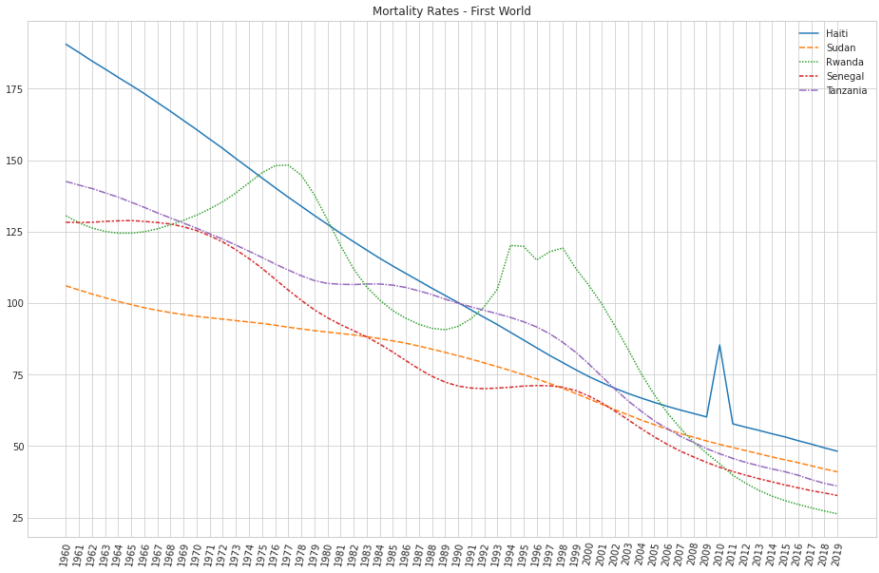
After browsing through peer work, we can see that there are similar mortality visualizations but they are asking different questions leading to different conclusions from their plotting. We wanted to pin it down to the link between nutrition and fatality and GDPperCapita index.

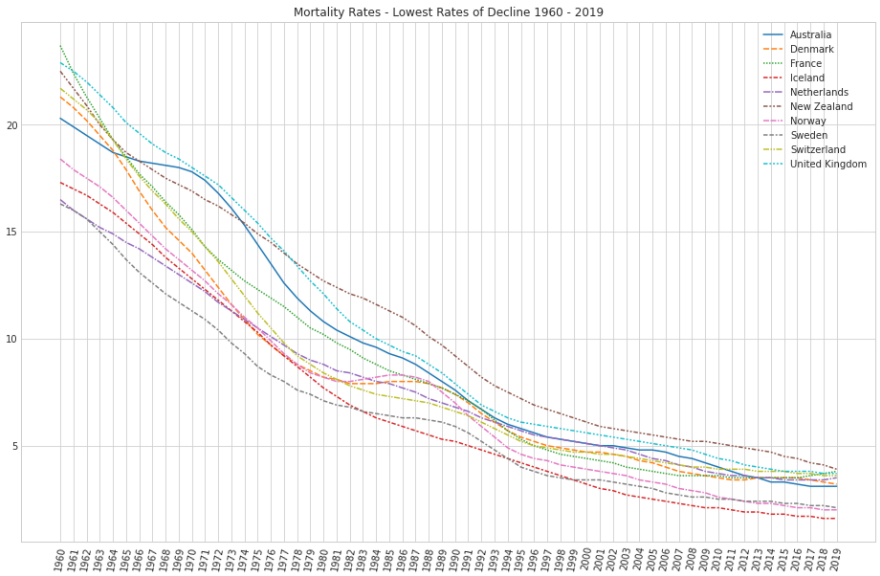
<https://datawrapper.dwcdn.net/Gczfa/2/>

“In the last 27 years, worldwide mortality rates among children under 15 have dropped by more than half”

****

**VS OUR DATA EXPLORATION**

****



- Data cleaning and type conversion activity. Please share anything unusual you faced during this activity.

We were fortunate in that we managed to avoid any categorical factor designation issues. Our practice with mapping and applying and pivoting is paying off. We didn't get too entirely sidelined with any special character or value scrubbing issues. The SQL table merging that we did was relatively straightforward. The fact that we are juggling so many different plotting libraries all with their own syntax is a bit frustrating.

Had to pivot table for time series.

Had to merge multiple DataFrames according to Country Code for choropleth

Had to groupby PopulationGroupL2DESC for violinplot

- What did you do about missing values and why? Handling missing values properly is very important.

The goal was to try to massage the data into the most complete form possible. Large blocks of missing values did cause us to reduce the trajectory of our ambition to just the European countries. We did have to remove Slovakia due to missing data. We did use the ability to aggregate data and fill missing cells by way of averages. We can show that any deviation in those cases would result in a near zero difference in our conclusions.

Mortality Data Set

2020 had many missing values so it was omitted.

na rows where entirely omitted versus filling with mean.

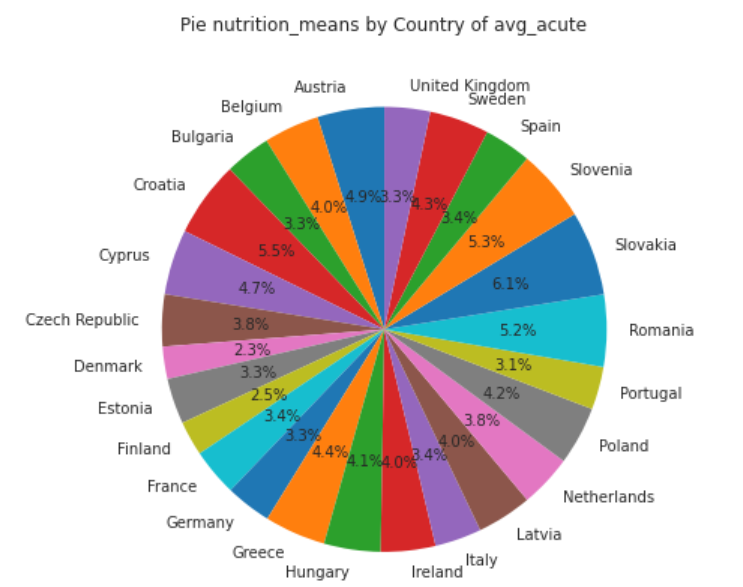
- New feature/attribute creation and data summary statistics and interpretation.

• Visualization (8 points, 2 points each)

We aggregated means across countries in order to compare nutrition and mortality to GDPperCap.  
  
Created new feature/attribute called Diet to represent the .variable ExposurehierarchyL1DESC in nutrition\_data as shown below:

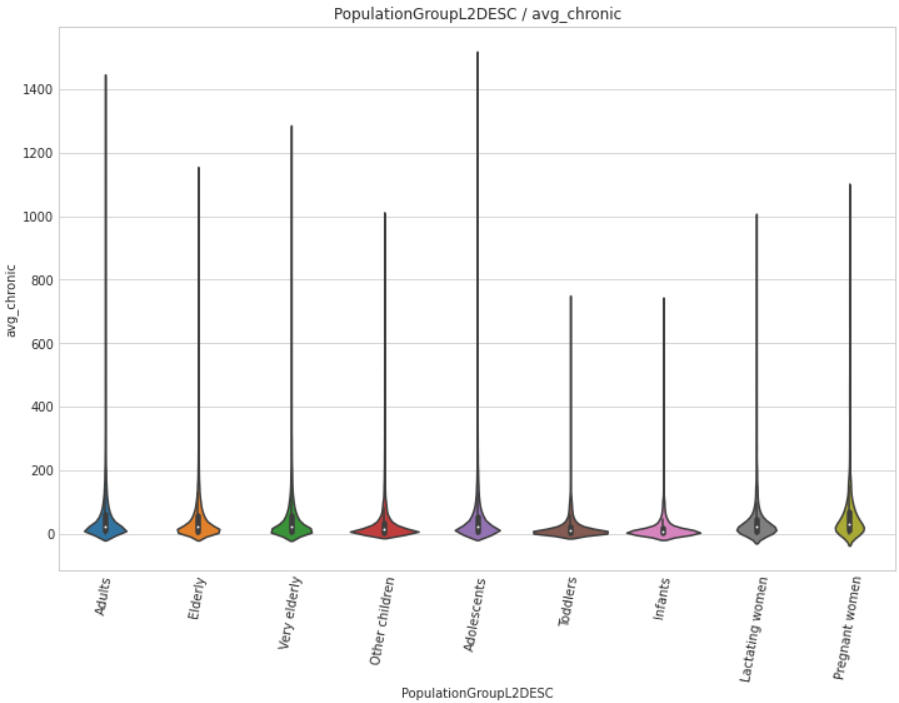
Example:

* Edible\_Oil : 'Animal and vegetable fats and oils and primary derivatives thereof',
* Cof\_Coc\_Tea : Coffee, cocoa, tea and infusions',
* Composite : Composite dishes',
* Eggs: Eggs and egg products',
* FishnSeaFood: Fish, seafood, amphibians, reptiles and invertebrates',
* Fruit: Fruit and fruit products',
* FruitnVeg\_Juices: Fruit and vegetable juices and nectars (including concentrates)',
* Grains: Grains and grain-based products',
* Legumes: Legumes, nuts, oilseeds and spices'

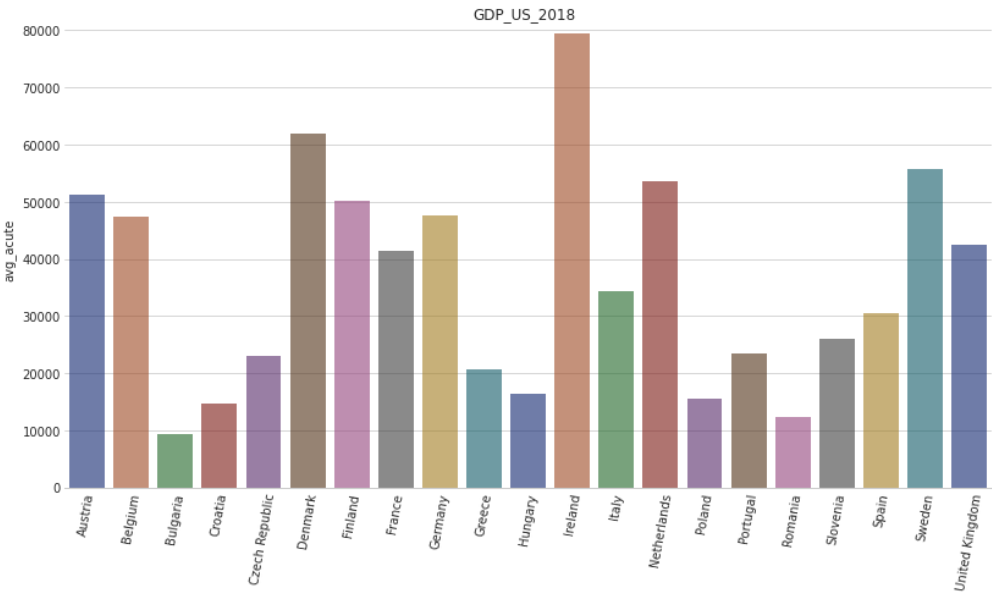


Averages tend to be unexpectedly similar. Slovakia, Croatia, and Romania are expected in comparison to their GDP values. Austria is a bit of a surprise.

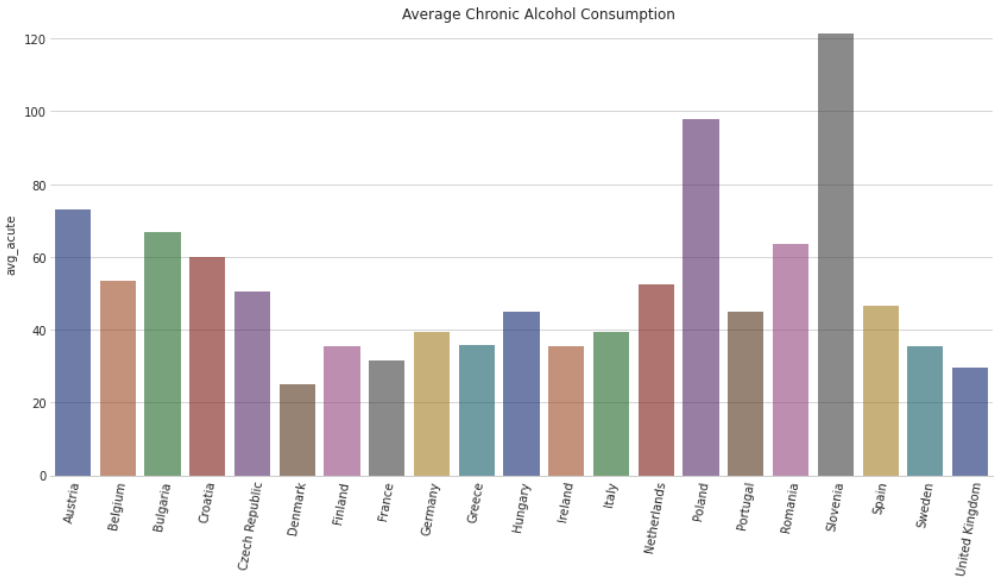
- Data visualization activity (box plot, bar plot, violin plot, and pairplot to see relationships and distribution, etc.).



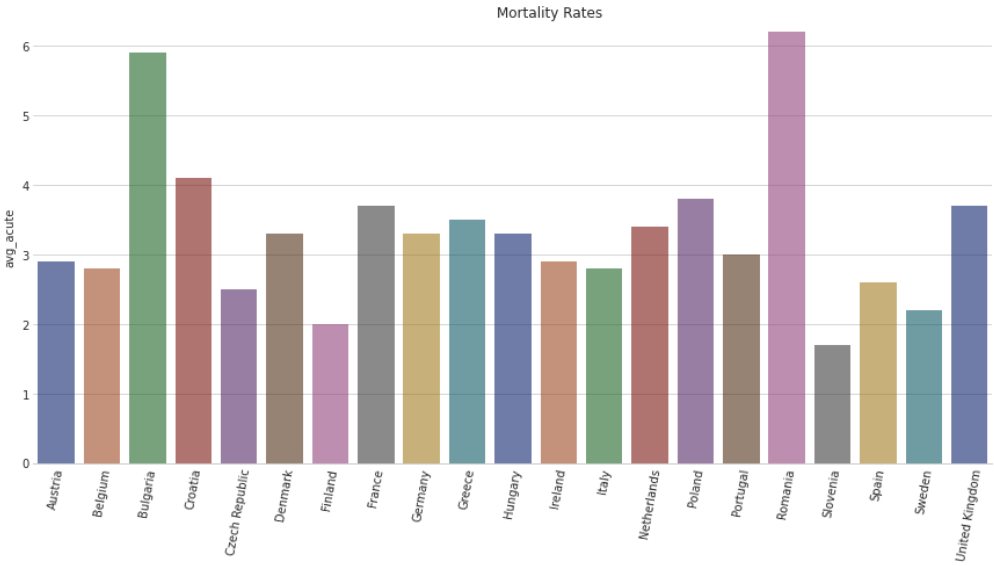
The above violin plot shows that there exists Chronic Alcohol Consumption in toddlers and infants, albeit a bit flatter in all the groups of young children.



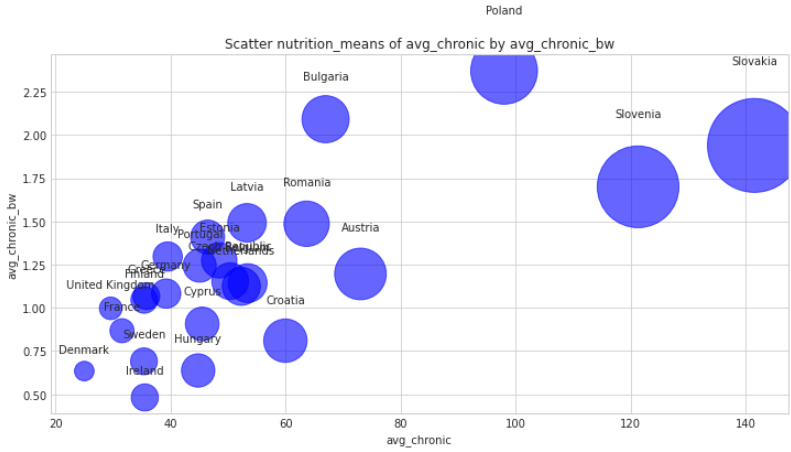
Ireland has an impressive GDP. Bulgaria, Croatia, Hungary, Poland, Romania are quite poor in comparison with other European Countries.

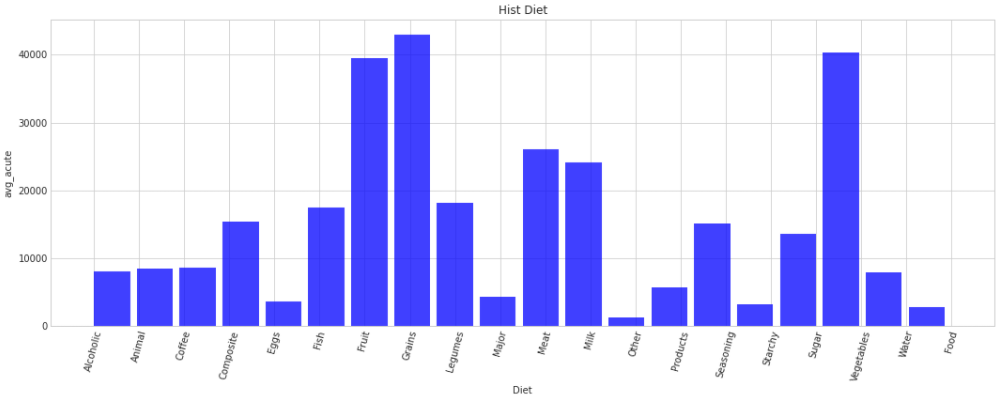


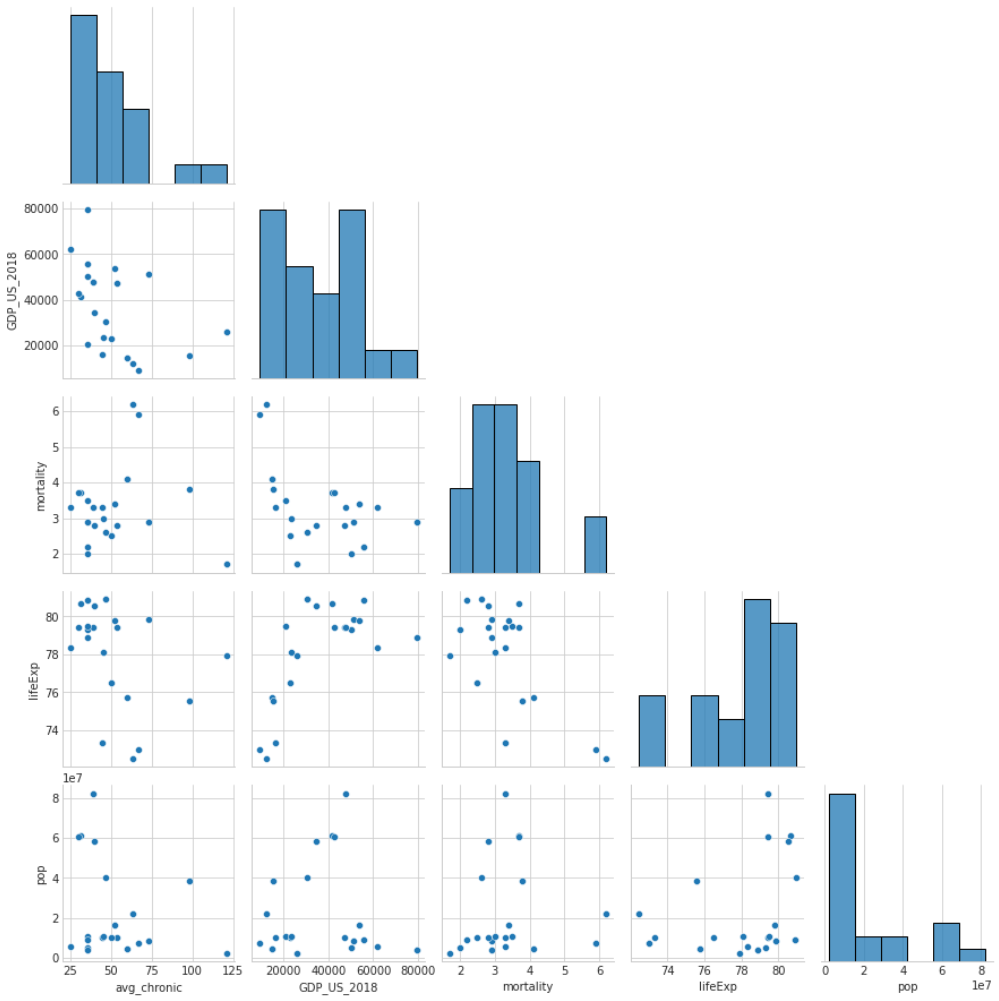
Ireland has a surprisingly low rating of Chronic Alcohol Consumption. While Slovenia and Poland are abnormally high.



Bulgaria and Romania seem to be dangerous places in congruence with their GDP but in opposition to their national alcohol consumption averages.

Slovenia and Bulgaria are outliers on this plot.





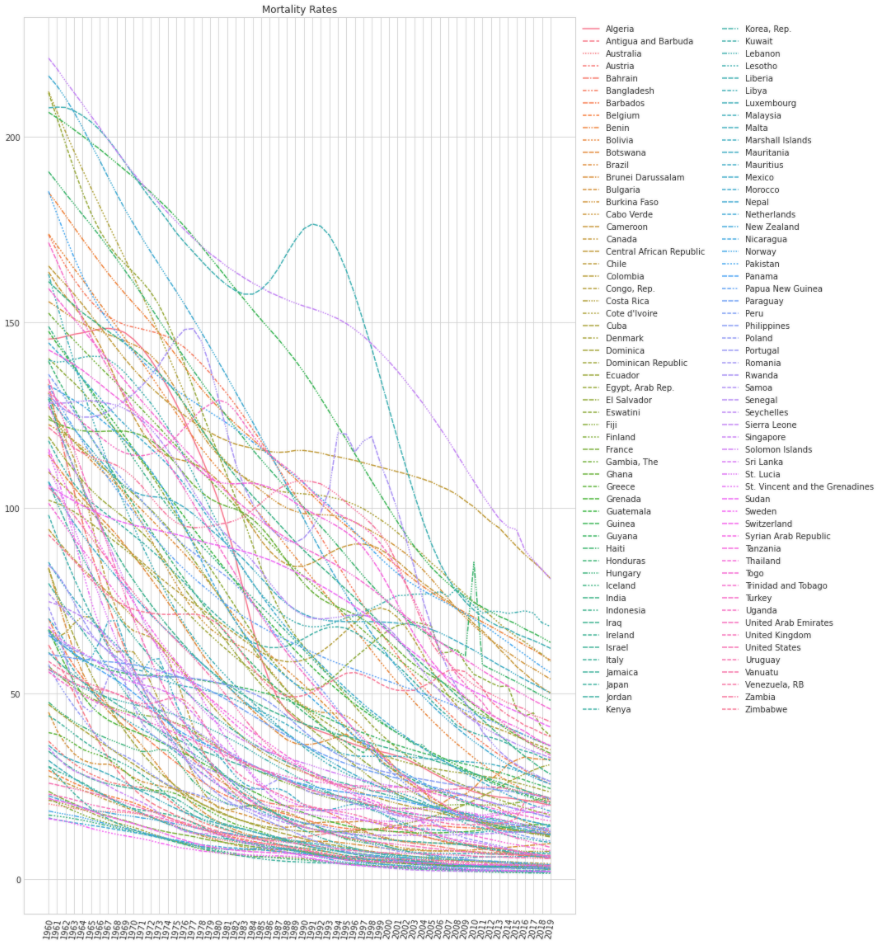
* Life expectancy values in this visualization are from 2007.
* I'm not getting much from this plotting... added it to fulfill requirements.

- Describe anything you find in the data after each visualization.

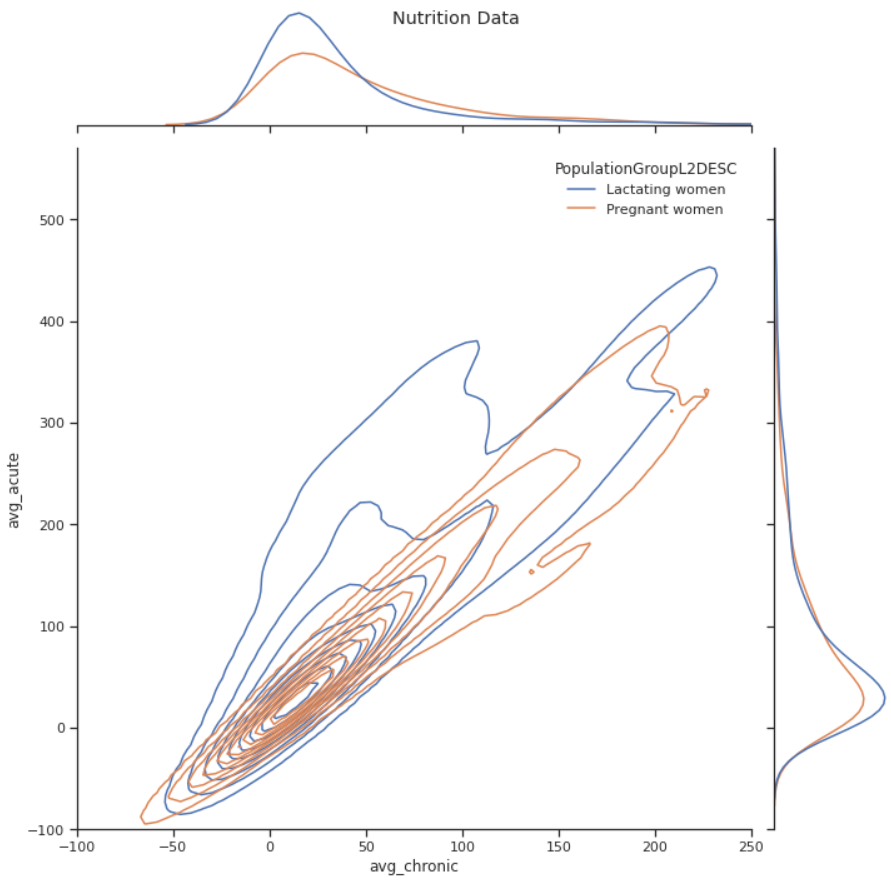
See above.

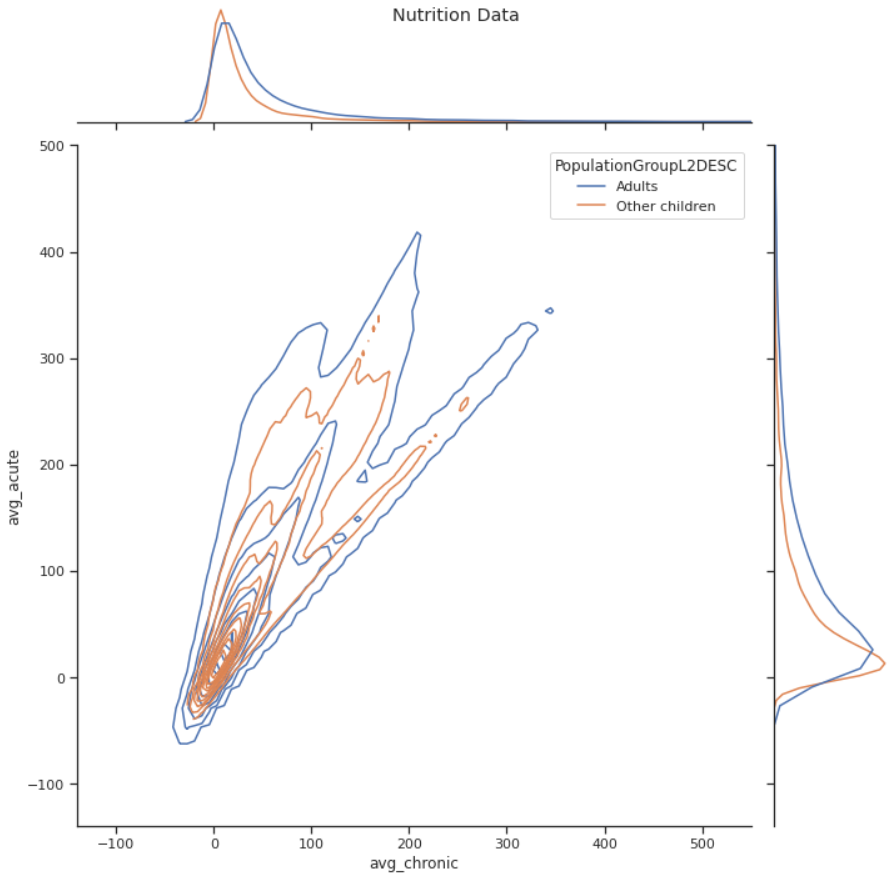
- What data visualization helped you understand about data distribution.

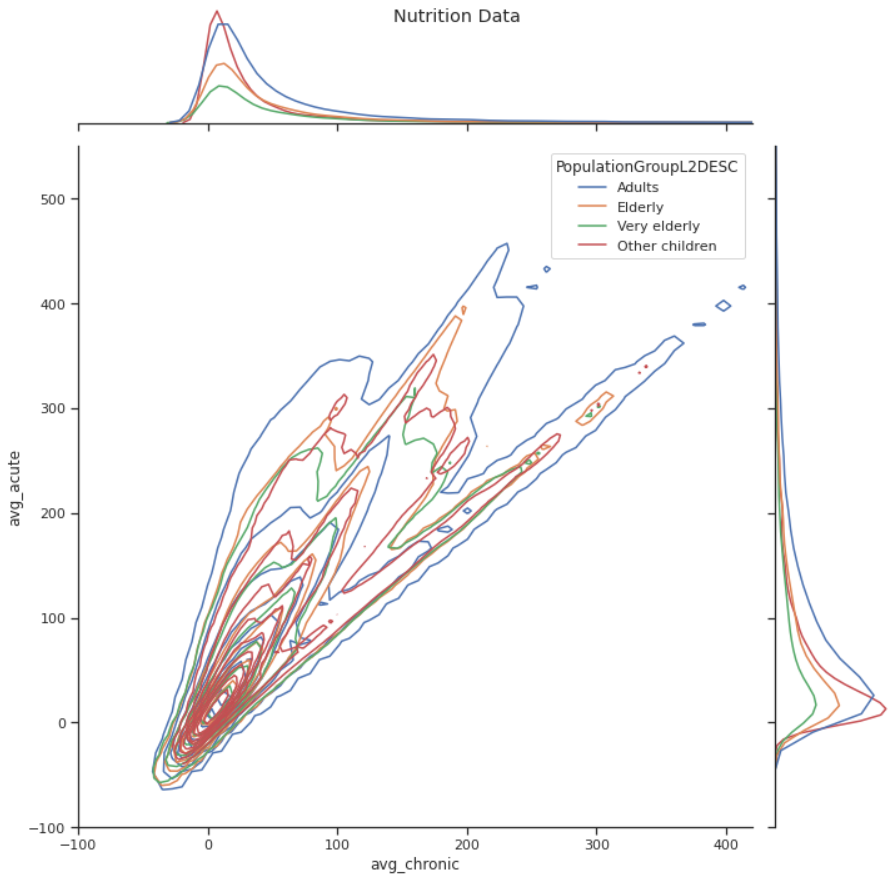
Histograms and barplots are a great place to start. They tend to show large and small conglomerations.



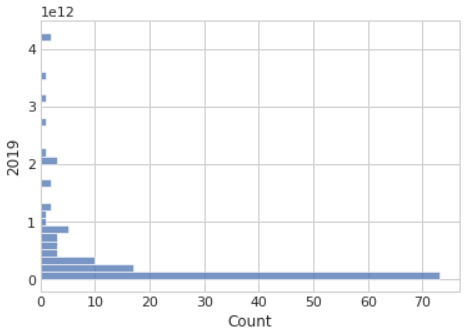
One of the most enlightening plots was a simple line plot of all the country mortality numbers over time. We could clearly identify the tragedy in Haiti and the conflict in Rwanda. I would have liked to normalize those numbers according to population density but we needed to focus on the task at hand.

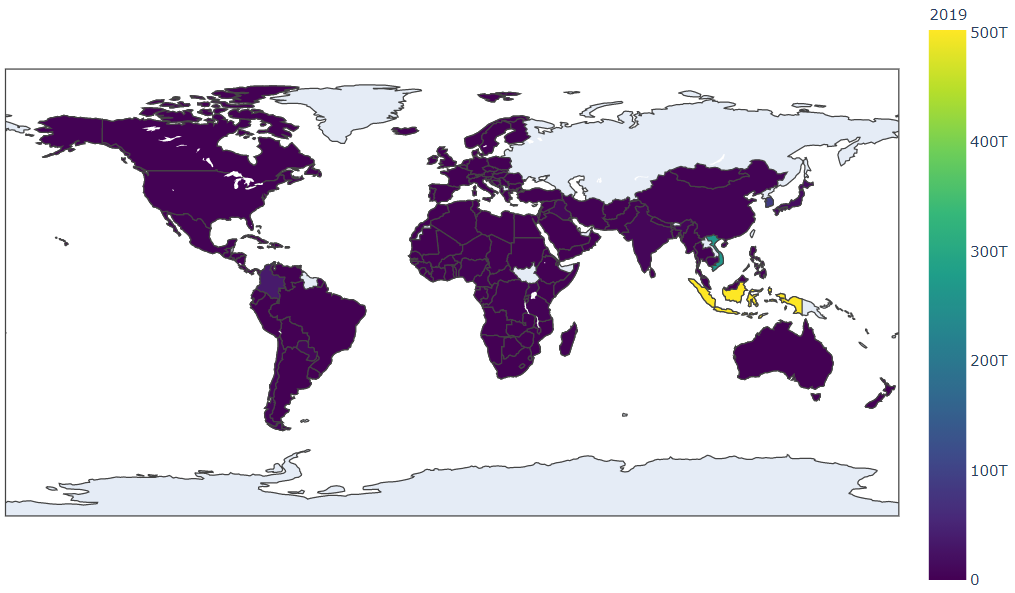






- What you did about possible outlier as per data distribution visualization. (Did you confirm with your client whether it is actually an outlier or put a disclosure statement in your notebook if you decided to remove it?)





We did a high level aggregation of the vast amounts of numbers into means by country in order to answer our research question. This aggregation eliminated most egregious outlier situations.  
  
Although in our exploration, we did run into situations where we were trying to plot color gradients on choropleth maps. The choropleth world map plot was showing high concentrations in South East Asia. In order to increase the gradient contrast across countries throughout the remaining world, we needed to set the range of the color ramp to essentially push/omit the greatest concentrations out of the range.