**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.

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 new child/infant mortality, nutrition AND national economic performance**

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

1. Quantitative analysis: determine if there is interaction between mortality and input variables

Null Hypothesis: There is NO interaction between mortality and input variable

Alternative Hypothesis: There is interaction between mortality and input variable

1. Is there interaction between mortality and acute consumption?

Text

Description automatically generated

By GLM(), p-value = 0.00 Reject Null Hypothesis, Conclude there is interaction between mortality and acute consumption

1. Is there interaction between mortality and development index?

Text, table

Description automatically generated

By GLM(), p-value = 0.005, Reject Null Hypothesis, Conclude there is interaction between mortality and development index

1. Are there any other causes of mortality besides the given input variables?

Multiple Linear Regression

Text, letter

Description automatically generated

From the value of the coefficient of determination above, the proportion of the change in mortality explained by the input variables is 12.1%. Conclude, there are other more significant causes of mortality besides the input variables. Alternatively, the multiple linear regression predictions DO NOT perfectly fit the data provided.

Answer to Research Question

1. Plots show that there is a linear relationship between new child and youth mortality and development index, average and acute consumption of alcohol
2. However, quantitative analyses showed that the changes in development index and chronic and acute consumption accounted for only 12.1% of the changes in mortality

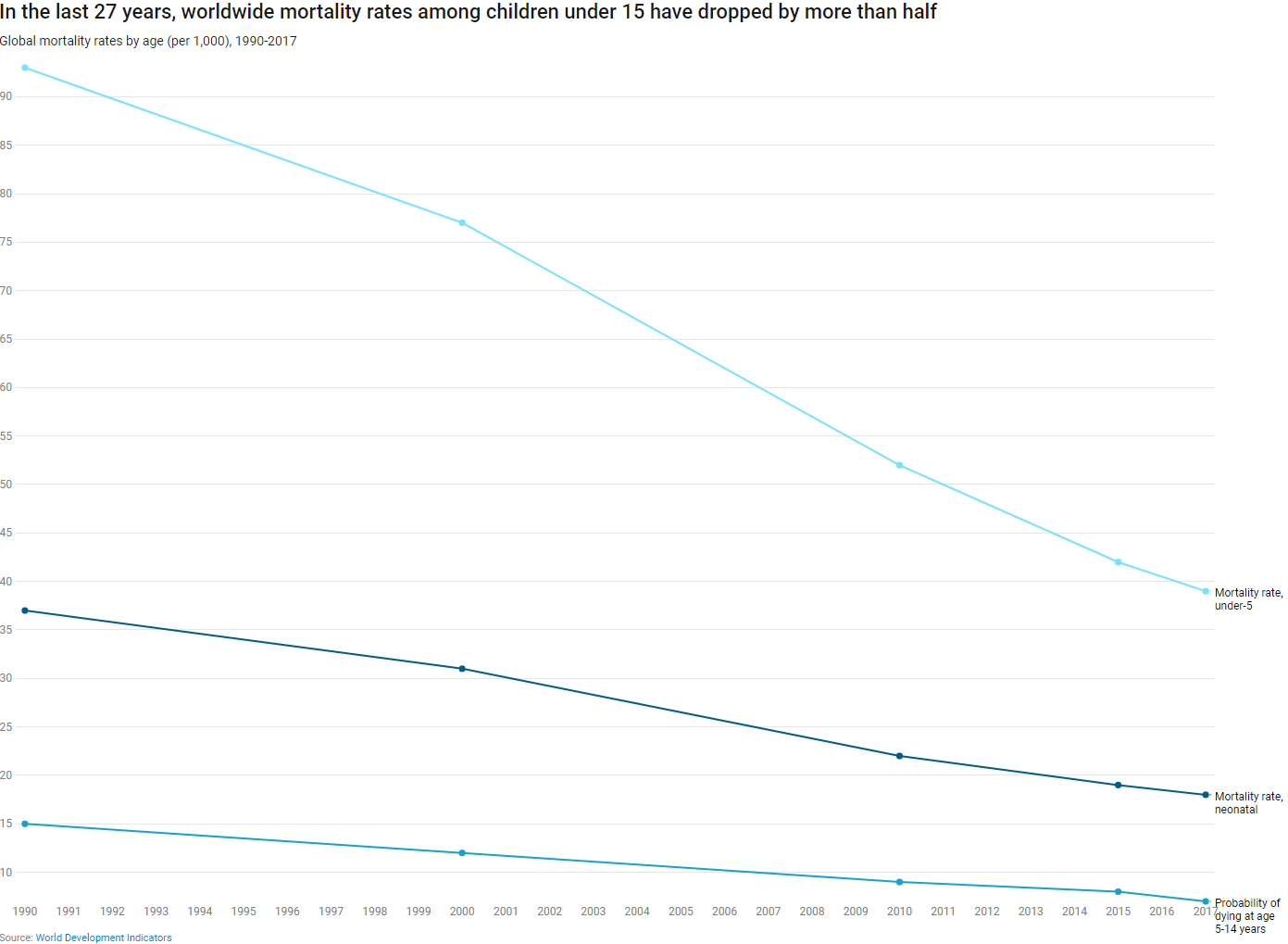
What other work has been done in this area, and how is your work novel compared to others?

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

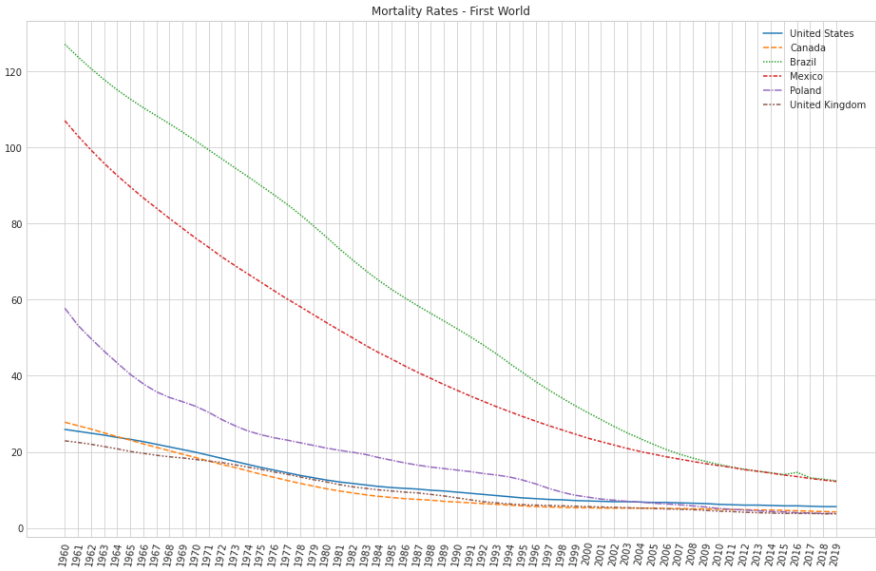
After browsing through peer work, we can see that there are similar plots but they are asking different questions leading to different conclusions from their plots. We wanted to pin it down to the link between nutrition and fatality and dev index

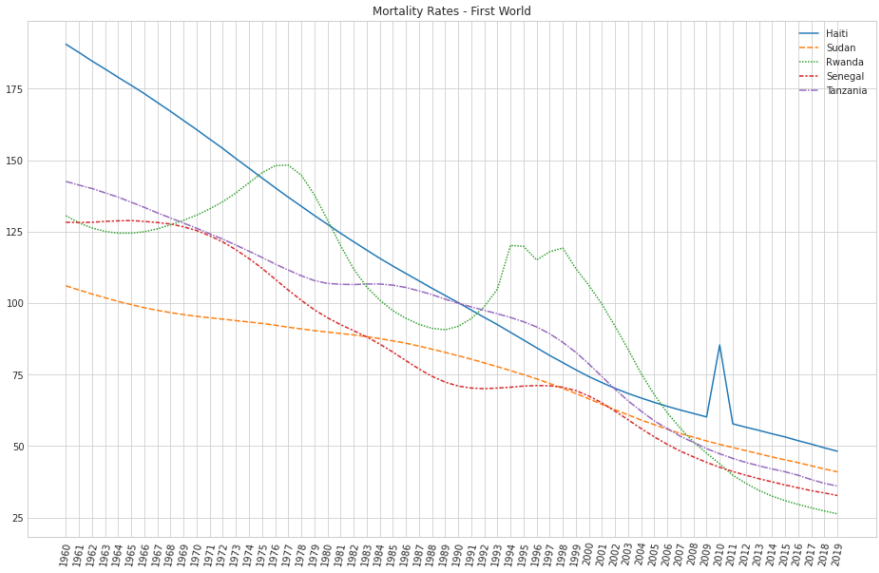
<https://datawrapper.dwcdn.net/Gczfa/2/>

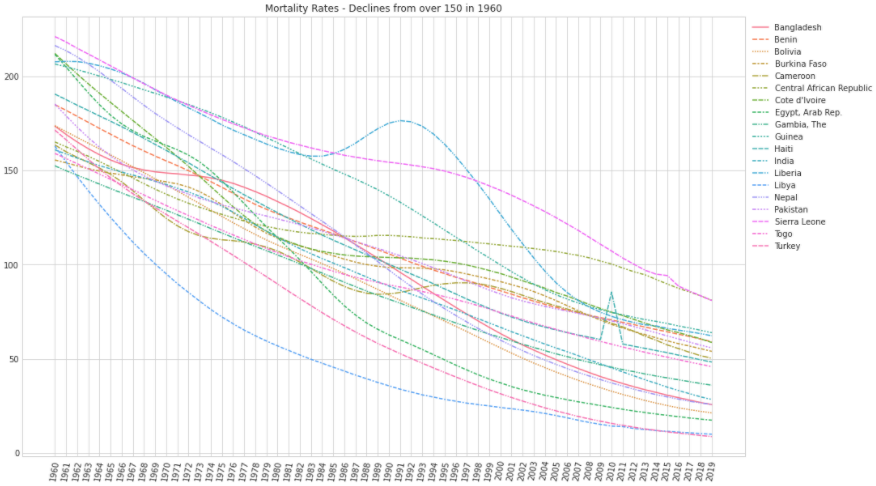
“In the last 27 years, worldwide mortality rates among children under 15 have dropped by more than half”

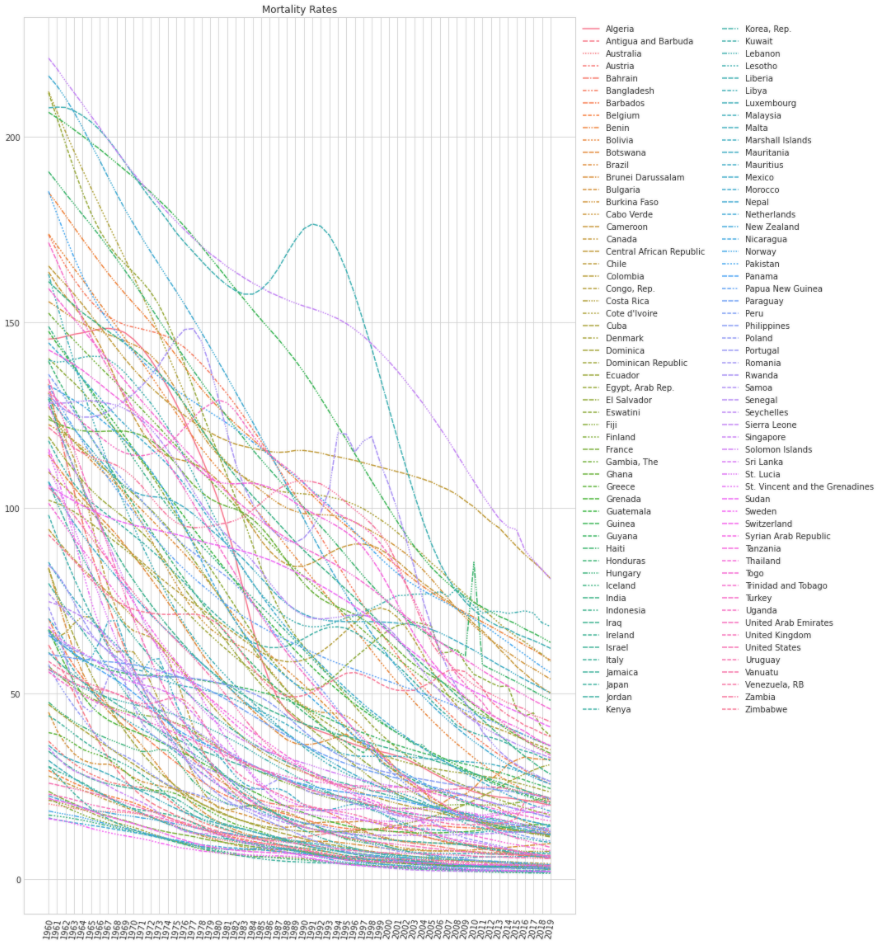
****

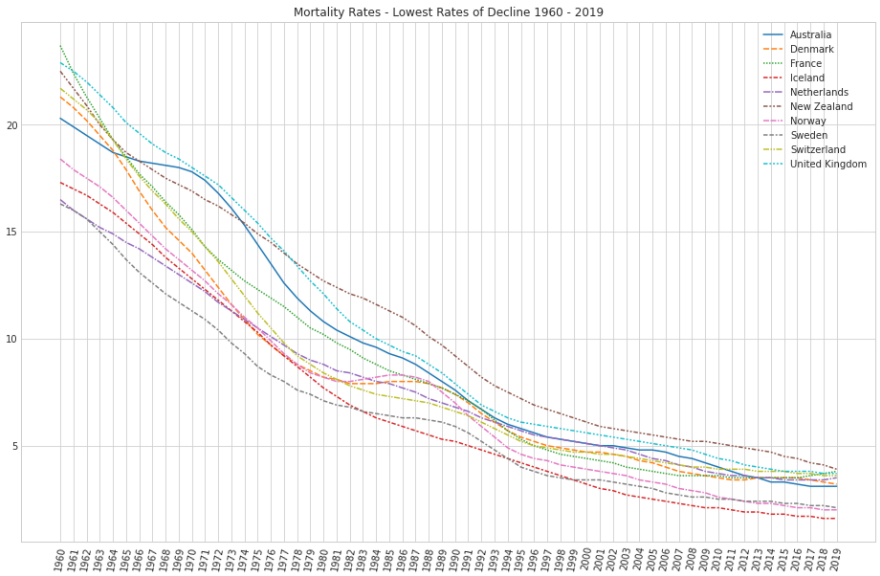
**VS OUR EXPLORATION**

****

****







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

Had to pivot table for time series.

Had to merge to 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.

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.

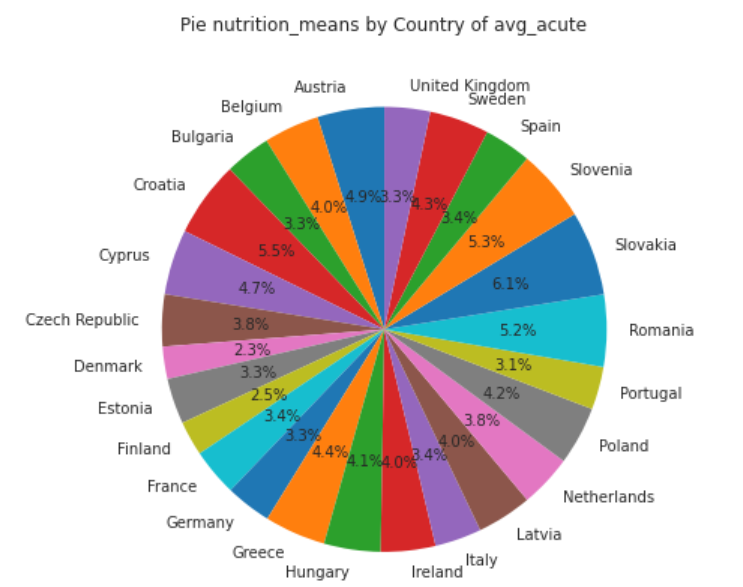
• Visualization (8 points, 2 points each)

Created means for variables by Country.

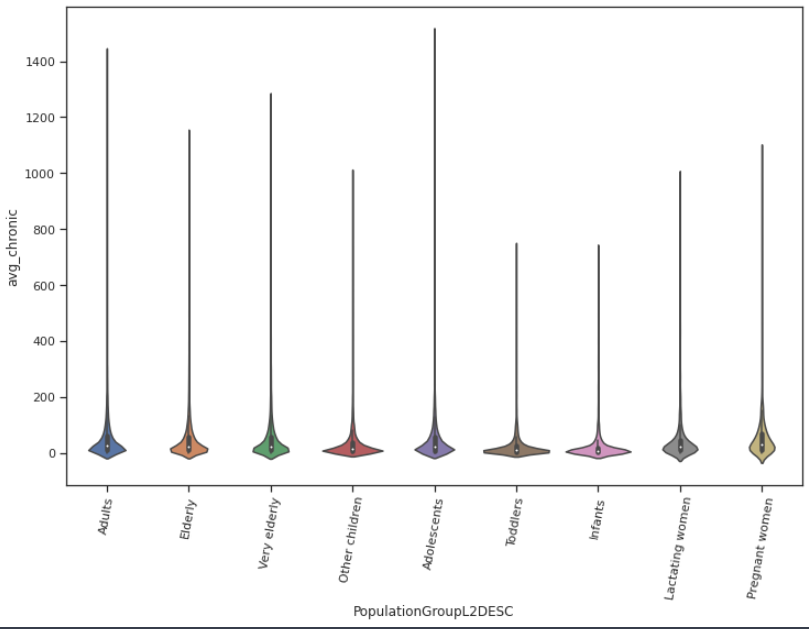
Created new feature/attribute called Diet to represent the .variable ExposurehierarchyL1DESC in nutrition\_data as shown below:

Table

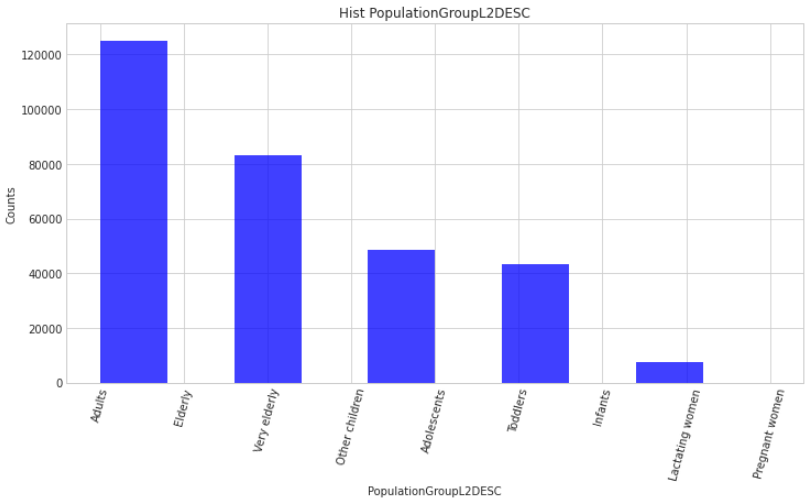
Description automatically generated

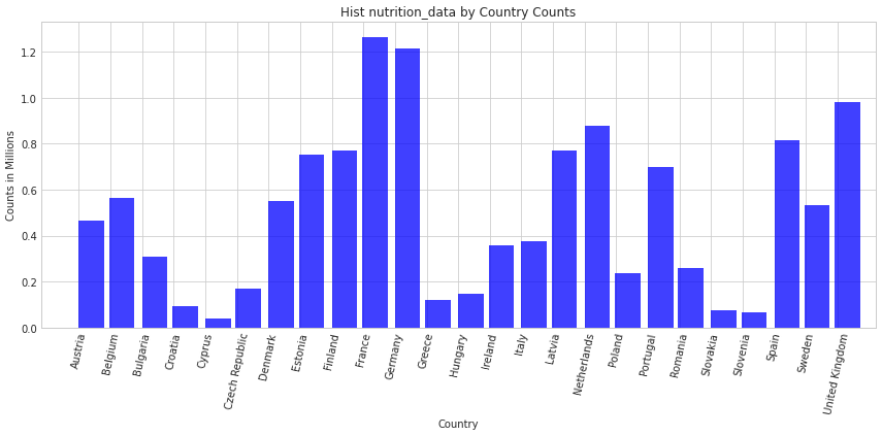


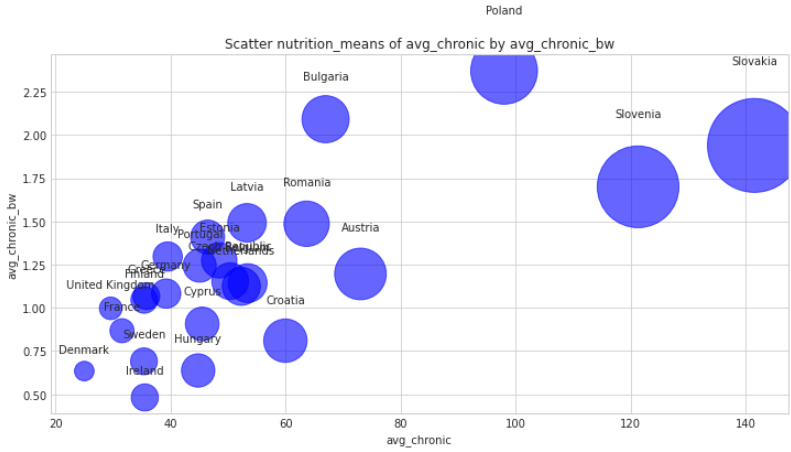
- Data visualization activity (box plot, bar plot, violin plot, and pairplot to see relationships and distribution, etc.).



Pregnant Women avg\_chronic is very high



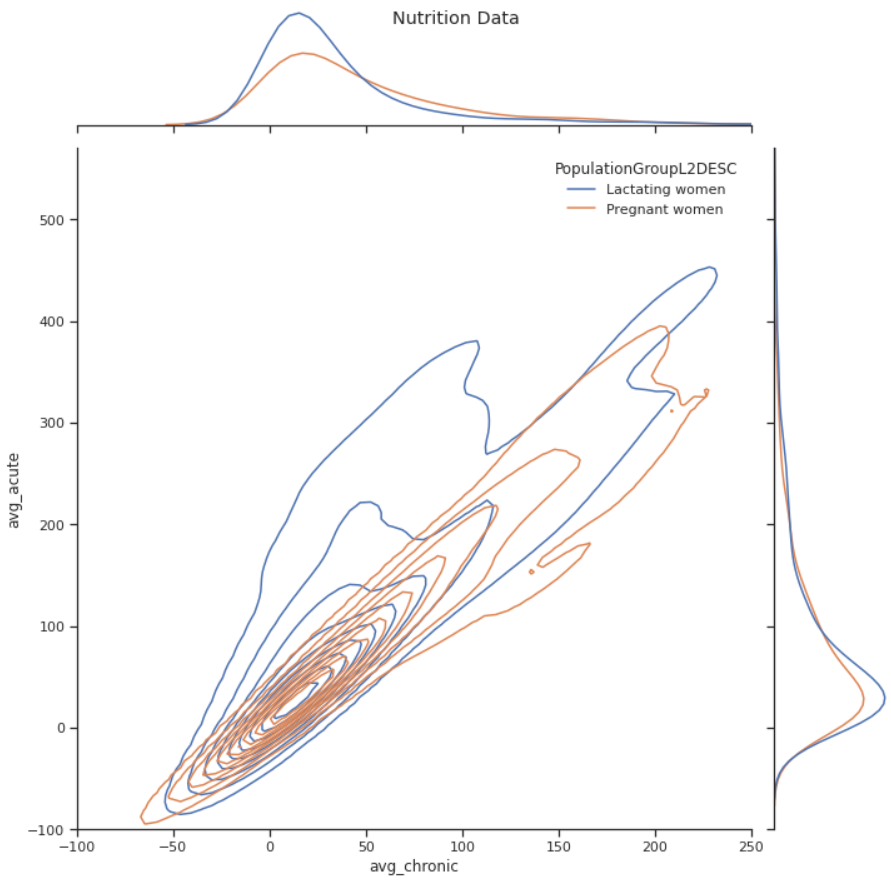


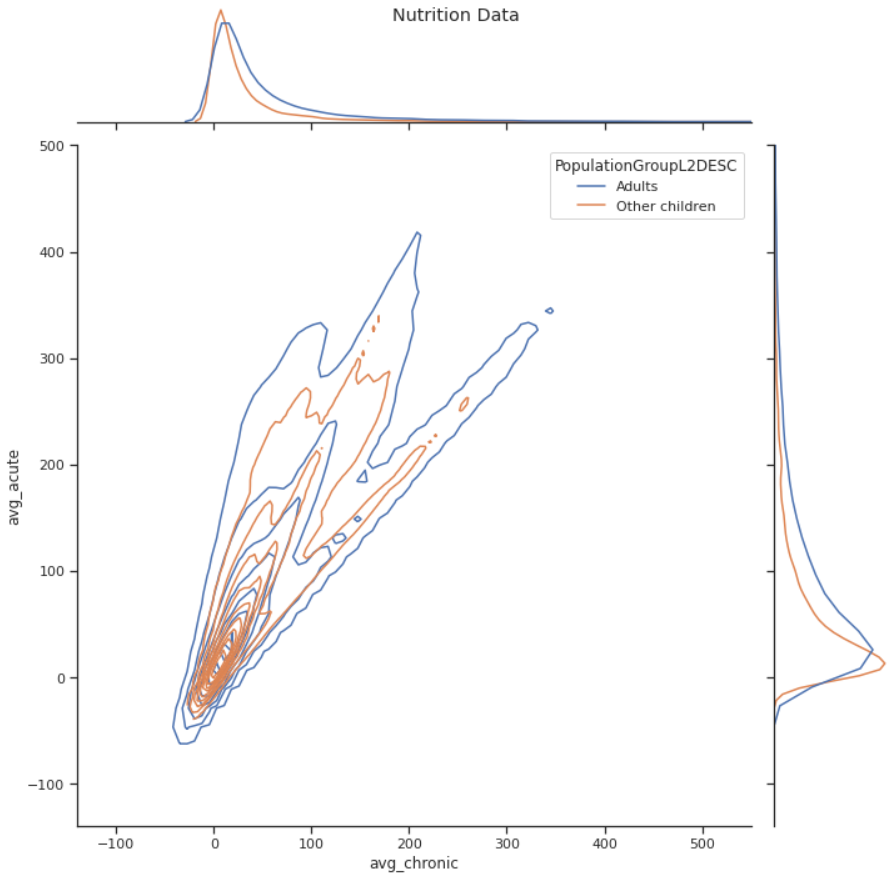


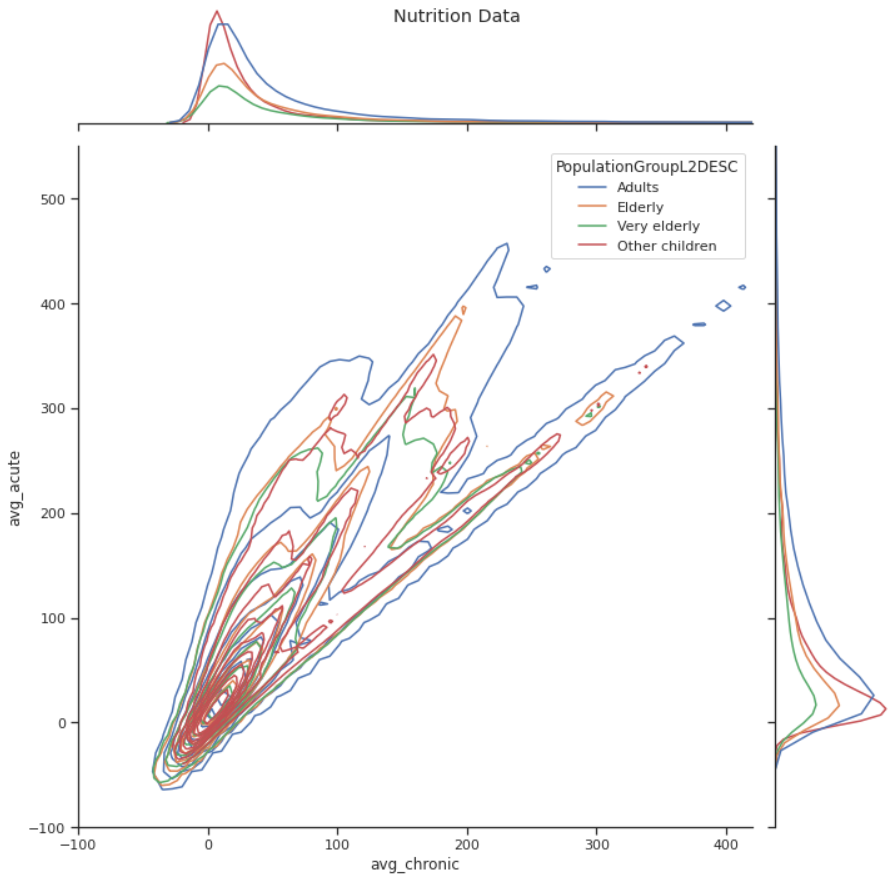
- Describe anything you find in the data after each visualization.

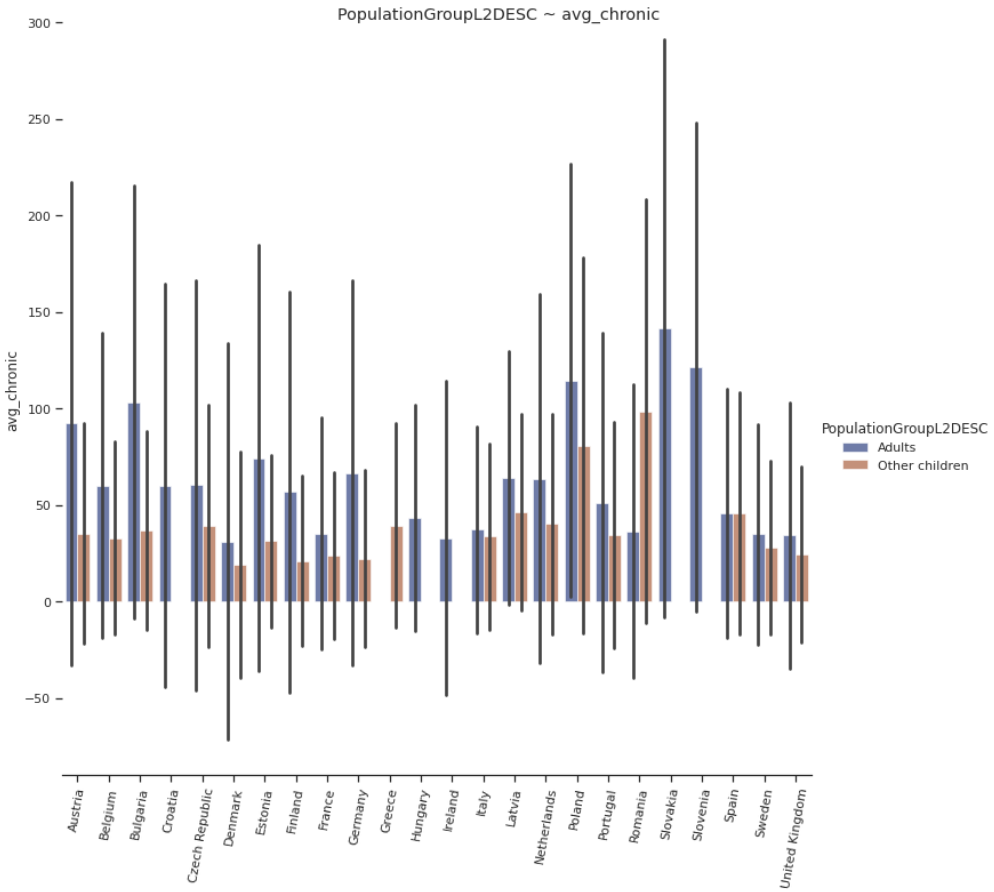
Annotate these

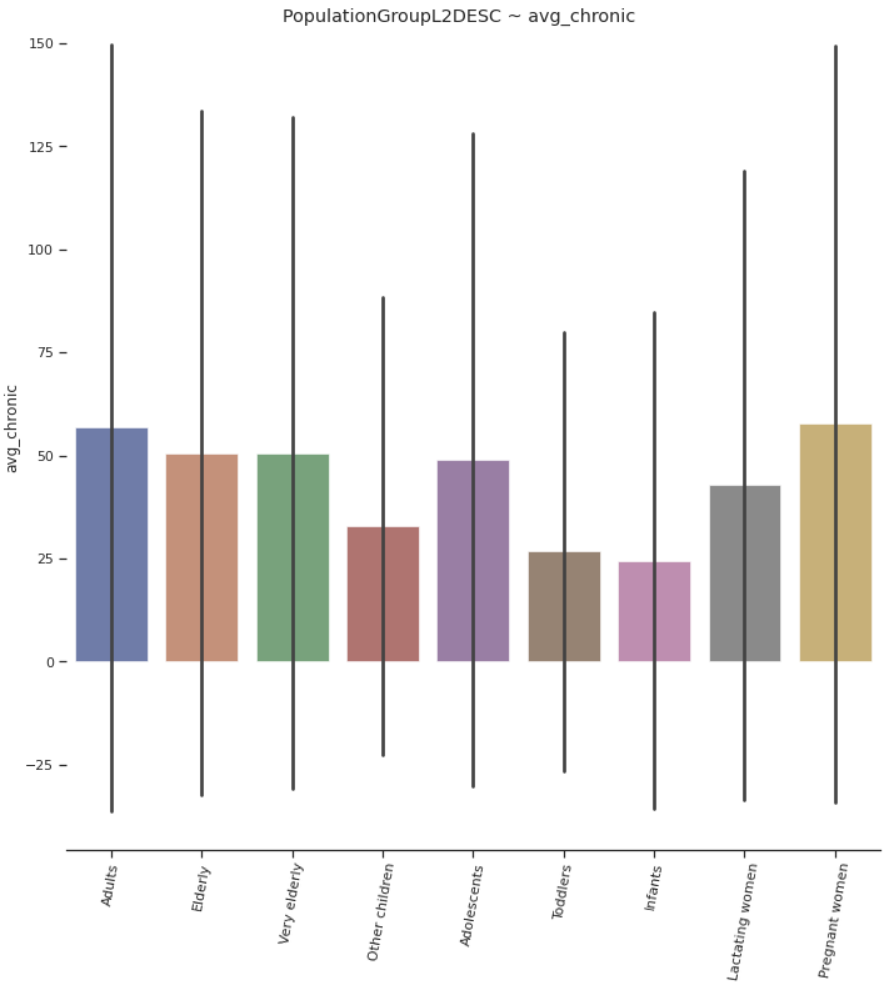
- What data visualization helped you understand about data distribution.



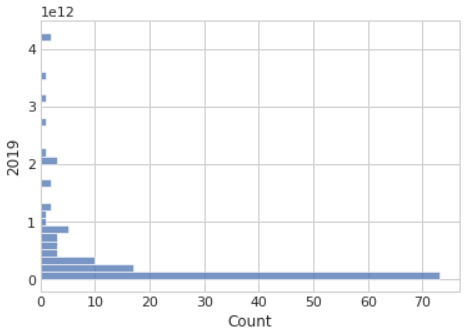


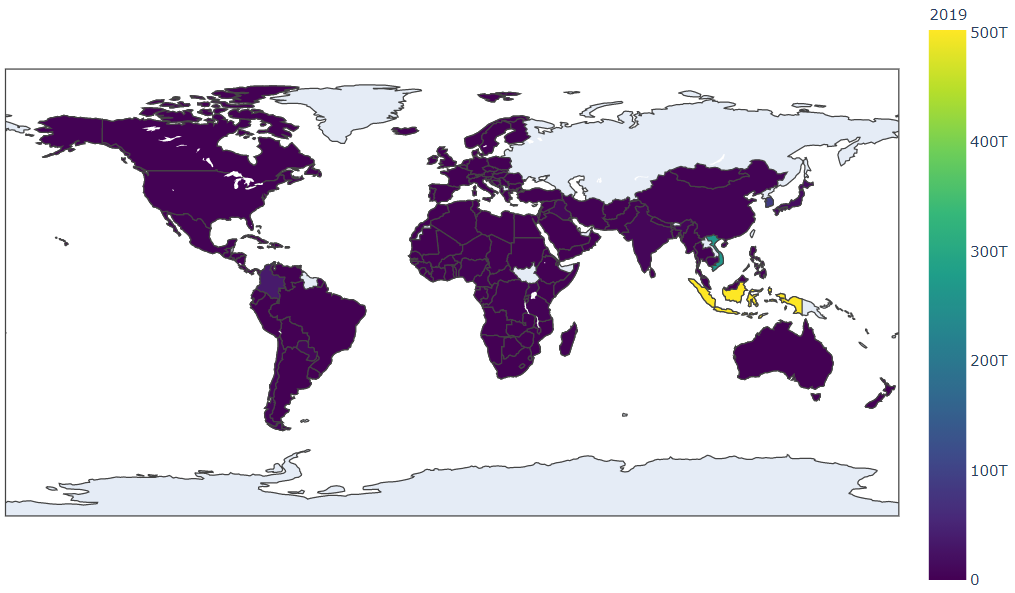






- What you did about possible outlier as per data distribution visualization. (Did you confirm with your client whether it is actually an outlier or put a disclosure statement in your notebook if you decided to remove it?)



This

This plot shows high concentration in South East Asia. In order to visualize a gradient range on the broader world we need to set the range of the color ramp to essentially push/omit the greatest concentrations.