Capstone 2

Data Analysis of Suicides from 1990-2022 Worldwide

Introduction to the Data

The data I chose was from Kaggle. I wanted a dataset that had missing data to learn better how to handle it.

After working with the data for a while and encountering some issues, I learned that the names of the two files were switched around in my download. I did not change them for this project, but if I encountered this in the workforce, I would address the issue immediately upon finding it.

I really enjoyed this dataset. It was not my first choice, or even my second. The other two datasets I worked with were missing too much data for anything to work. With this dataset, I was able to explore missing data, charts I had only seen briefly in classes, and I really learned a lot from it.

This dataset consisted of two files: suicide rates and suicide ages.

My Process

The first thing after loading the data and using glimpse(), summary(), and names() to get familiar with the data was to clean the data.

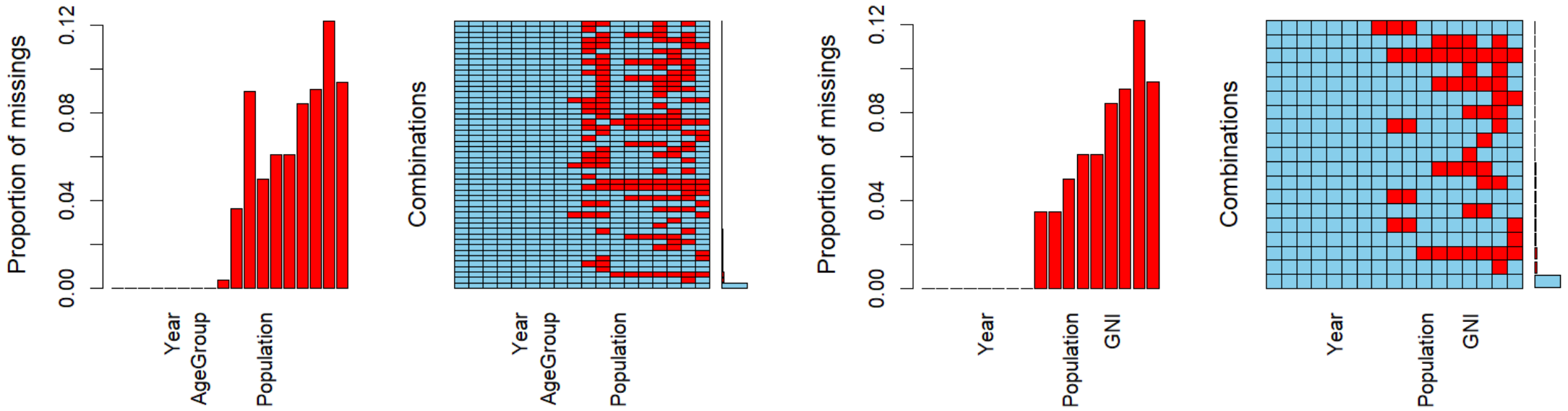
First I checked for complete cases. The suicide rates data was 79% complete and the suicide ages data was 75% complete. My next step was to visualize the missing data. I discovered a great tool in R called VIM. It helps visualize the missing data. With the charts on the next screen, you can see the data that is missing.

Next I checked for NAs with summary(). Both files had numerous NAs. I removed them and then checked for duplicated data. The suicide rates had multiple duplicated lines. I removed them, compared the data from both files, changed a name of one column in suicide rates to match the suicide ages and then compared the datasets again to make sure they could easily be joined. I used a left join to join the data.

Process continued on slide 5

Missing Data Analysis Using VIM

The left chart is the missing data from the suicide rates dataset and the right is the suicide ages dataset.



My Process With Clean Data

After cleaning the data, I stated my initial hypotheses.

Null Hypothesis: The rates of suicide do not change based on age or country.

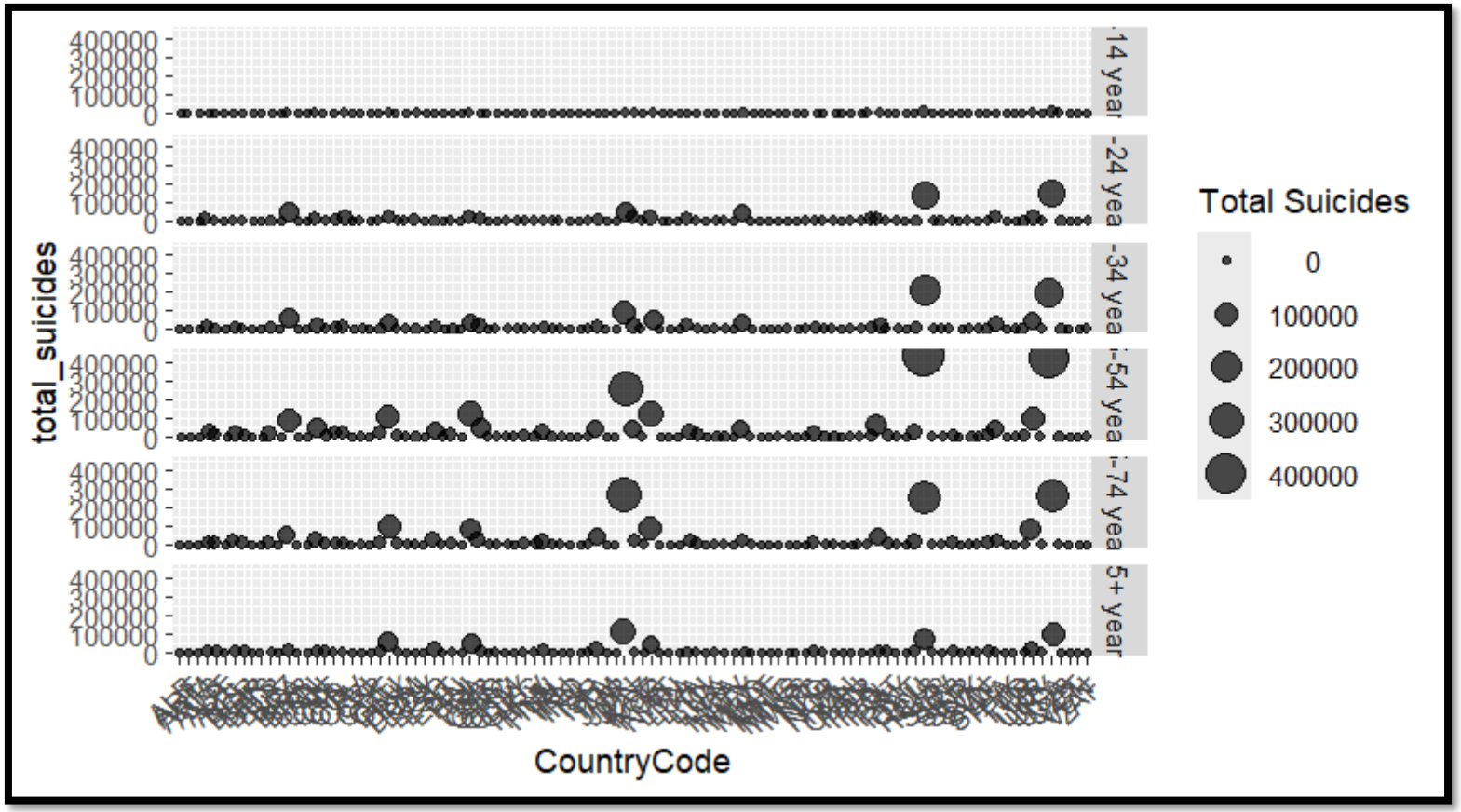
Alternative Hypothesis: The rates of suicide are lower with the younger age groups in lower income countries.

I started to explore the data. I used group_by() and found the data spanned 32 years, 101 countries, 6 age groups, and 2 genders. I wanted to test my hypothesis, so I grouped the data by country and age group with the mean suicide counts and the GNI for each country. The data was massive and hard to read on the charts. I switched focus and broke the data down a bit differently.

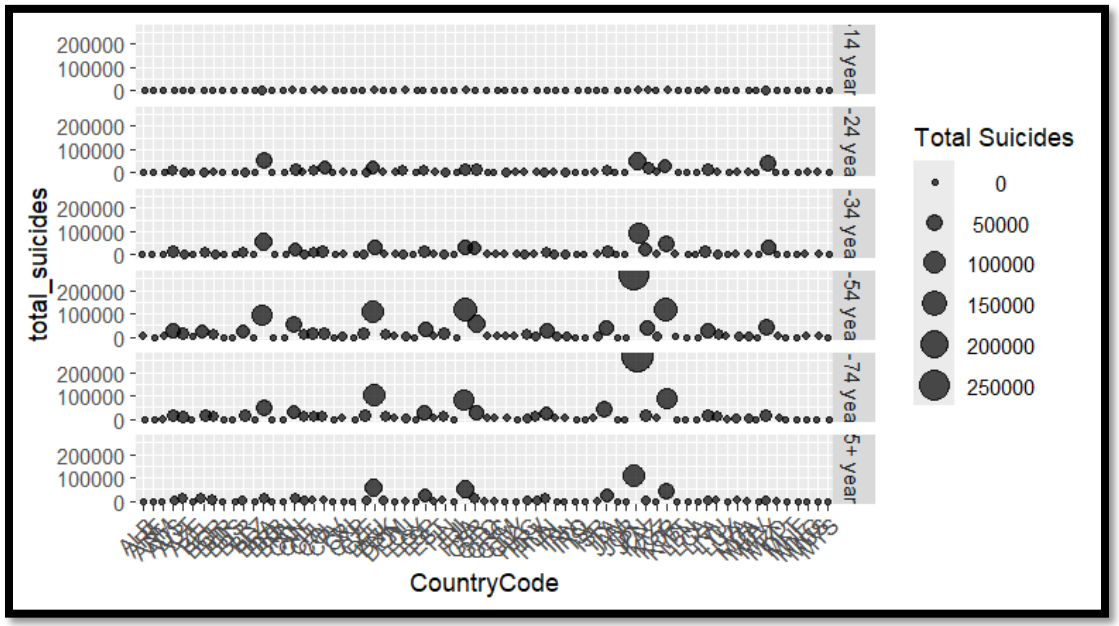
Process continued on slide 5.

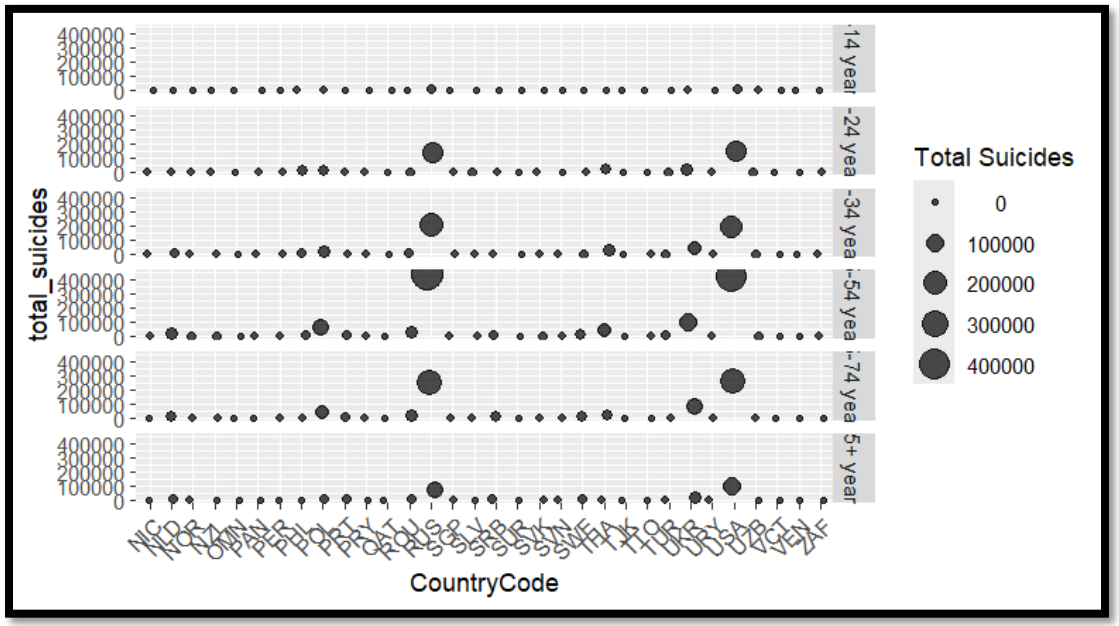
Data Broken Down by Country, Age Group and GNI

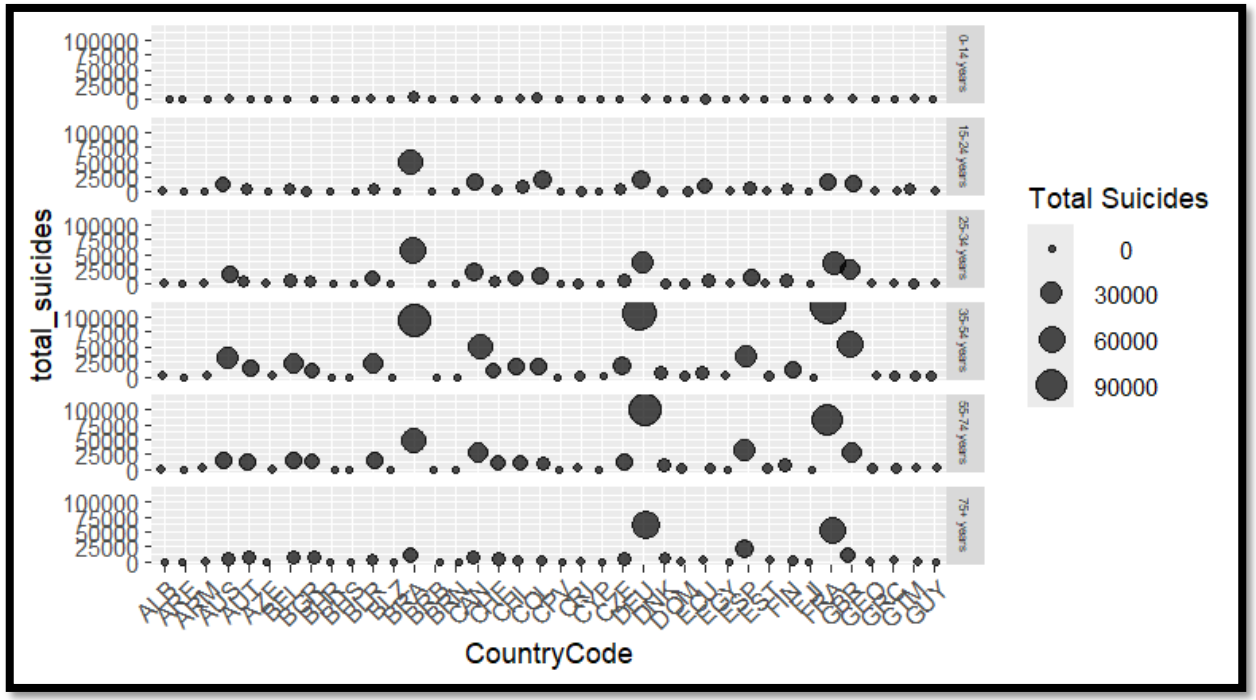
These first 2 are the are the charts I initially plotted. I used the facet grid variable to groups them nicely. As you can see, the data is massive. I could not use enough aesthetics to make it easily readable. I next moved to breaking down this chart into two based on the alphabet.

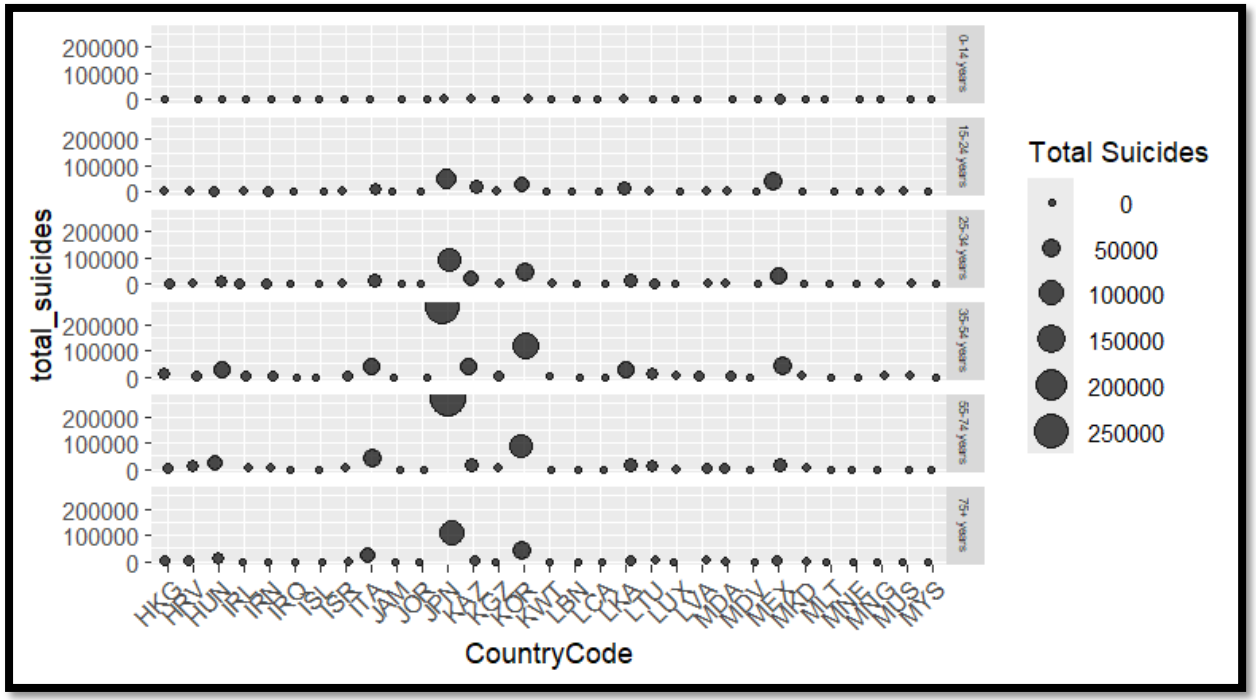


These are the charts using A to M and N to Z. The data is still pretty unreadable so the next step was to break it down into 3 charts. I used A to G, H to M, and N to Z. This worked much better.

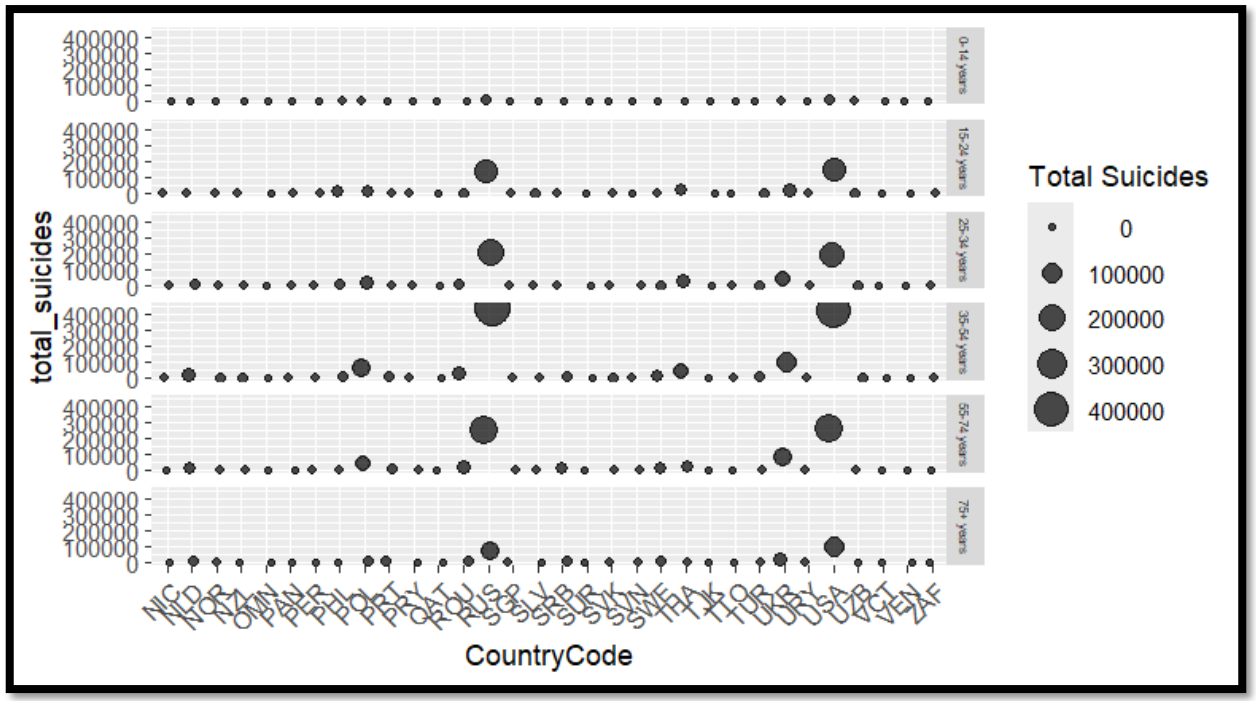






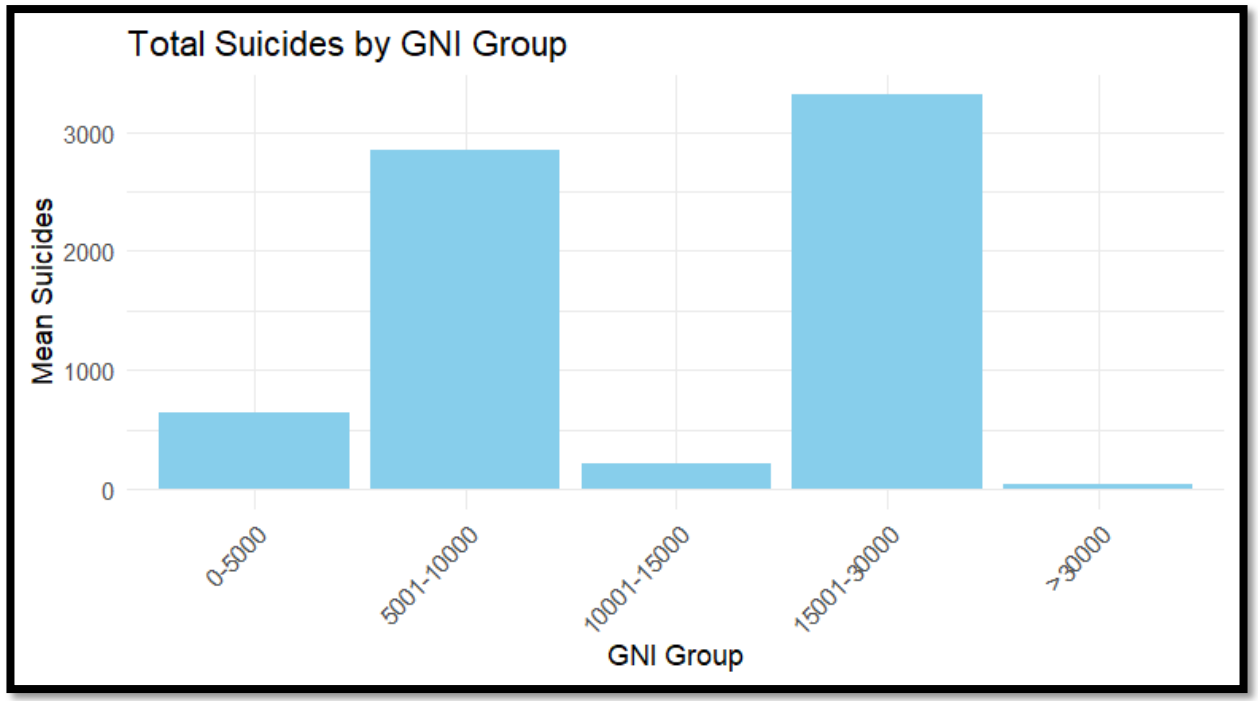


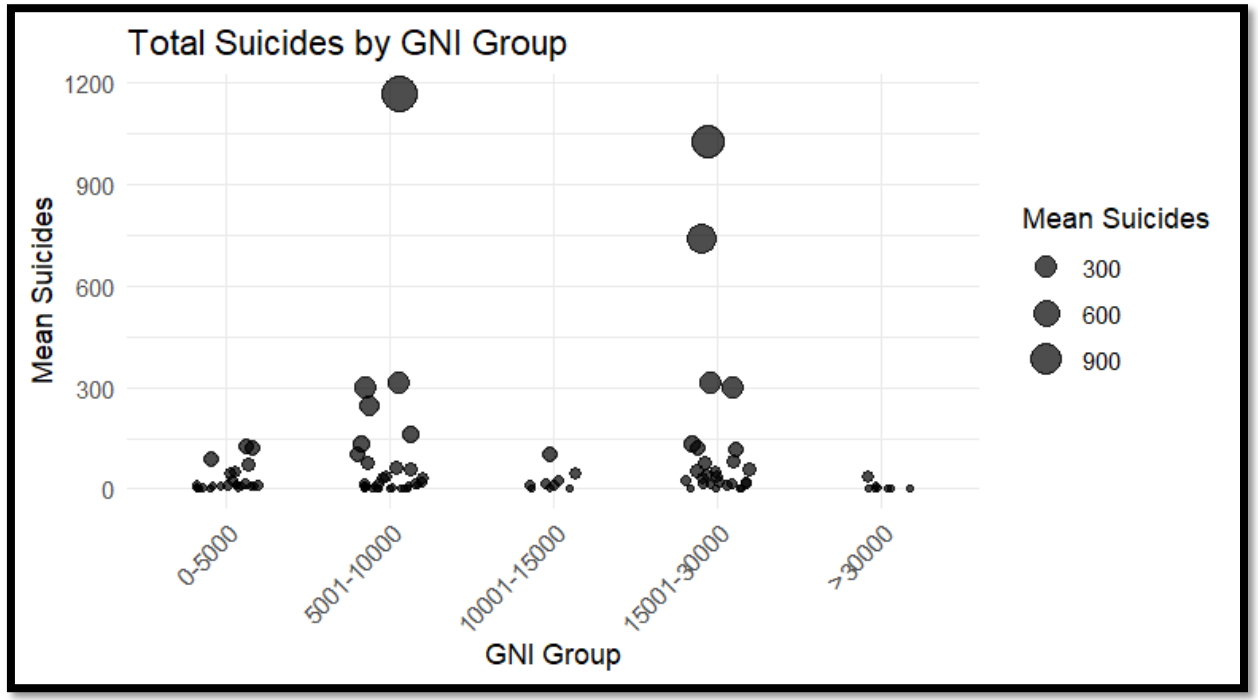
As you can see here, the data is much easier to view with 3 charts.
This led me to discover that the mean suicides rates are not higher in lower age groups. They are actually higher in the older age groups. The age group with the highest occurrence of suicide is the ages between 35 and 74.

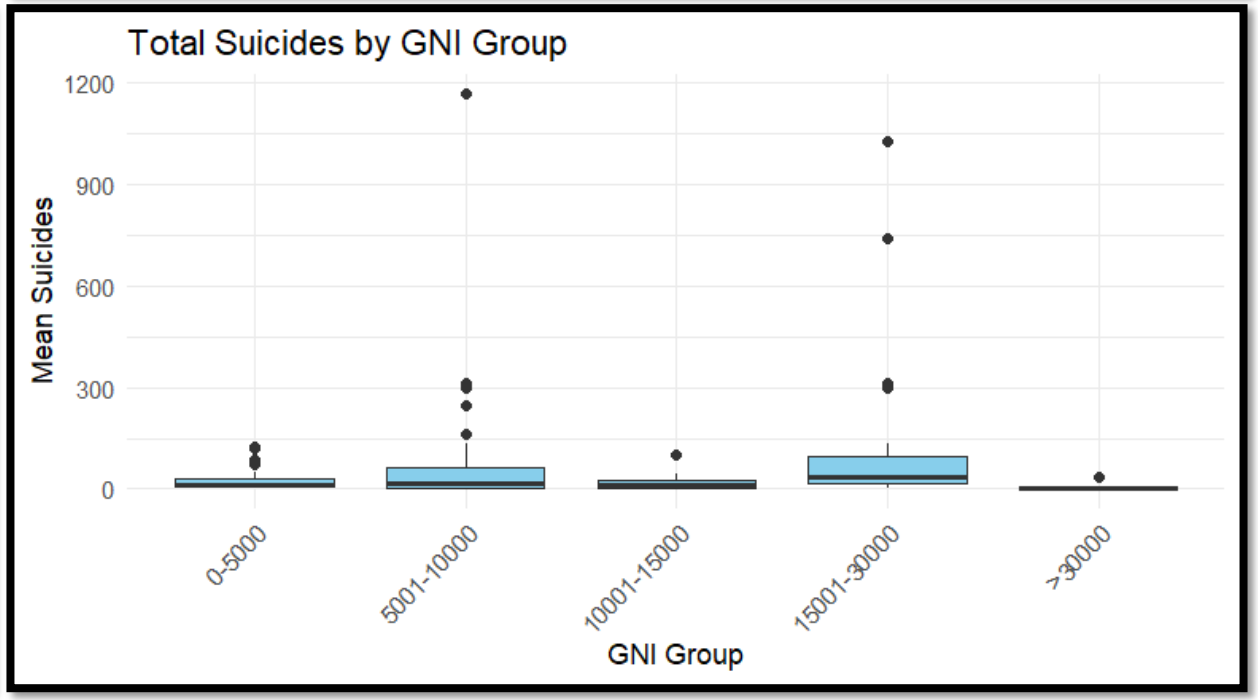


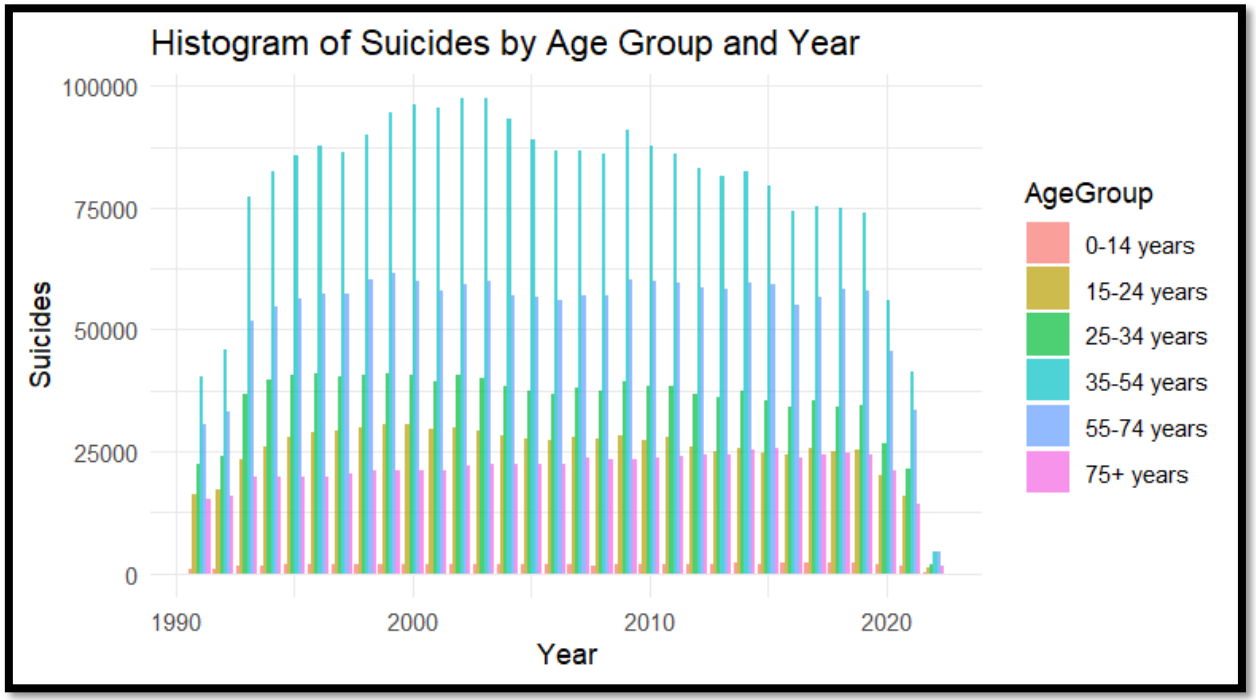
So in my initial hypothesis, I wanted to see if suicides were higher in lower income countries and in lower age groups. This shows that part of my hypothesis was incorrect.

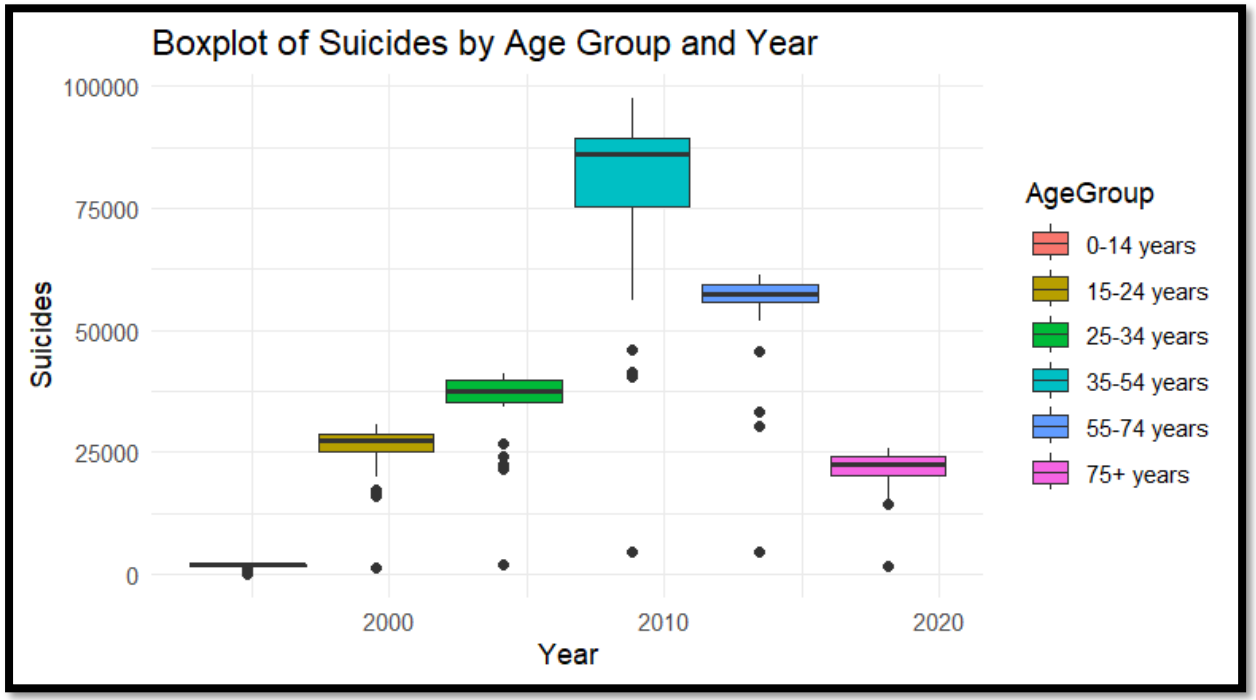
To see if my second part of my hypothesis is correct, I grouped the data by country, and by GNI based on GNI Per Capita. It seems to be fairly evenly distributed here for the most part. The bar chart displays the best here.

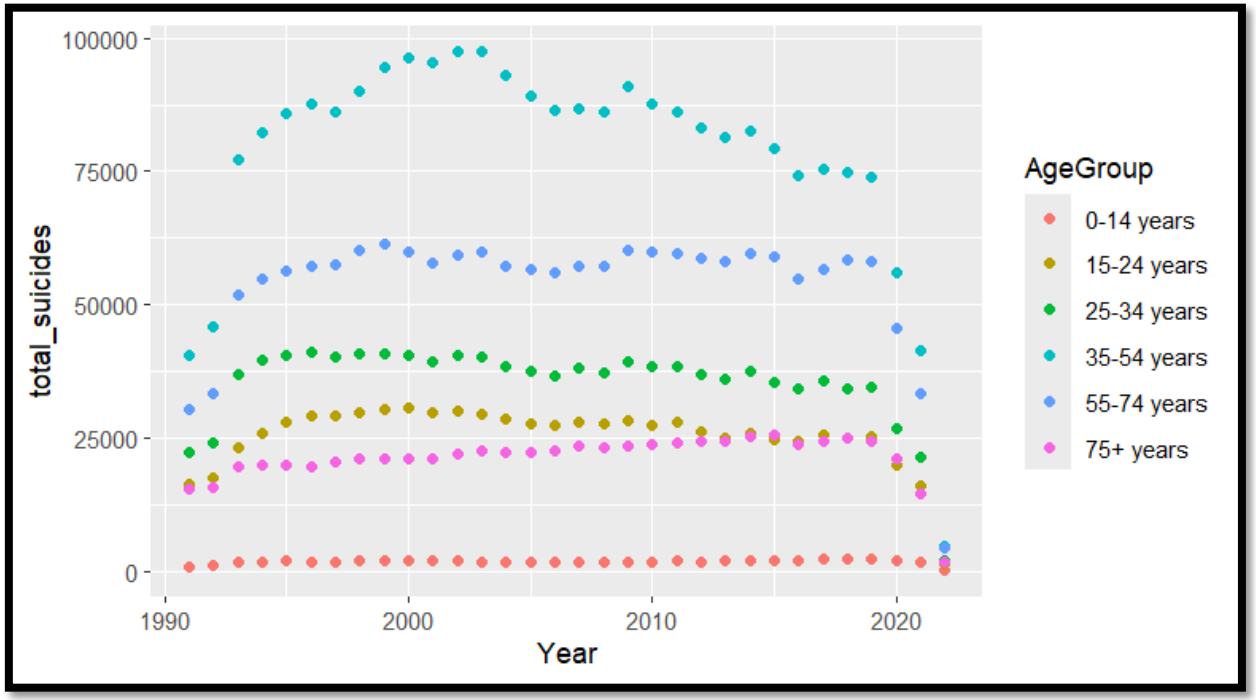






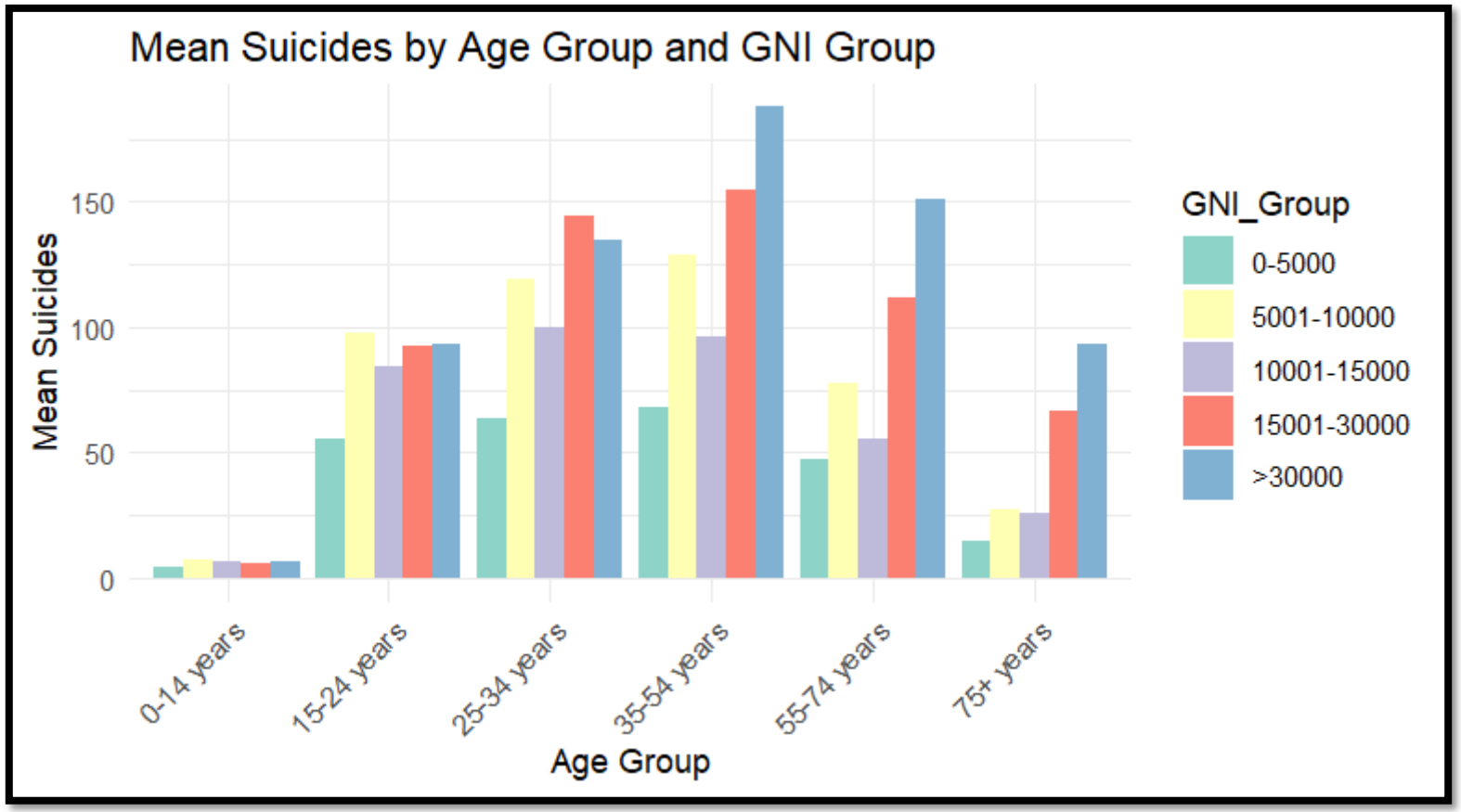






Finally, I wanted to see the data by age group and year. I wanted to see the data from multiple angles again. The histogram, while very nice, is just too congested. The boxplot and scatter are wonderful. You can easily see that my hypothesis on age was very off here.

Finally, I put the age groups and GNI groups together. As you can see from the bar chart, the highest suicides are in the 35-54 year range in most GNI groups.



Conclusions

After reviewing the data, charting it, and understanding it, I have learned that the most suicides are in the 35-54 year age group. This seems to be standard in most within most of the GNI groups I assembled. There were two exceptions: In the 5001-10000 GNI group, the highest age groups for suicide was the 15-24 year group and in the 15001-30000 GNI group, it was the 25-34 year group. The United States falls into the 15001-30000 GNI group with 24100 GNI based on GNI Per Capita. We had a mean of 1027.68 suicides between 1990 and 2022 and a total of 1,133,528.