The dataset that I used for this assignment was about the ‘World Happiness Report’ for 2021. The data contains a happiness score for most countries in the world as well as data about various social statistics. Before I started working with this data, I came up with the following three questions: What is the happiest place in the world?, What social statistics have an impact on happiness?, and How does the happiest country compare with the US?. While performing my EDA I added the question, Are there huge discrepancies in happiness scores?. The first three questions would provide one with an overall sense of how happiness is measured and where the ‘happy’ countries are. The new one would look for trends in happiness scores based on the different regions in the world.

In performing the EDA, I mostly looked at various tables which showcased the happiness scores for different countries and regions and in order to rank them. Moving forward, I made some histograms like the ones towards the end to see the various distributions of data for different social statistics. This allowed me to get a sense of what the values mean. However, I ran into some issues with the generosity data since it also had negative values which made understanding the values a little difficult. I also made some heatmaps to see the relation between various social statistics and the happiness score to see what the score depended on the most but chose to not display that as it was not the most visually appealing.

In order to answer my questions, I chose to create a map which would allow me to display the happiness scores for all the countries in the world at once. Sadly, there were some countries missing and I had to rename various others to match the format used so I could plot them. Moving forward, I grouped the countries by region and made a bubble plot to visualize trends amongst the various regions in the world. There was no real need to perform any transformations here since the labelling for the regions was quite uniform. Next, I filtered the data to only the ten highest countries and changed the range of the x-axis to better capture how close each of them are. For the next visualization, I chose to use a sun-plot to directly compare the two countries. In doing so, I isolated the data for each of the countries and created the sun-plot. Finally, for the final few visualizations, I chose to use histograms with lines displaying where the highest scoring country (Finland) is and one that is middle of the pack (USA). In order to do so, I isolated data regarding different statistics and plotted it, as well as where the two countries stand. The GDP data was already logged so I did not have to log it myself.