Alex Lopes

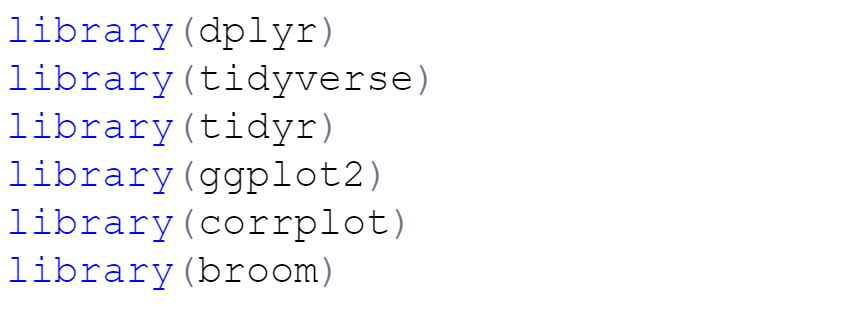
Statistical Analysis Final

Introduction

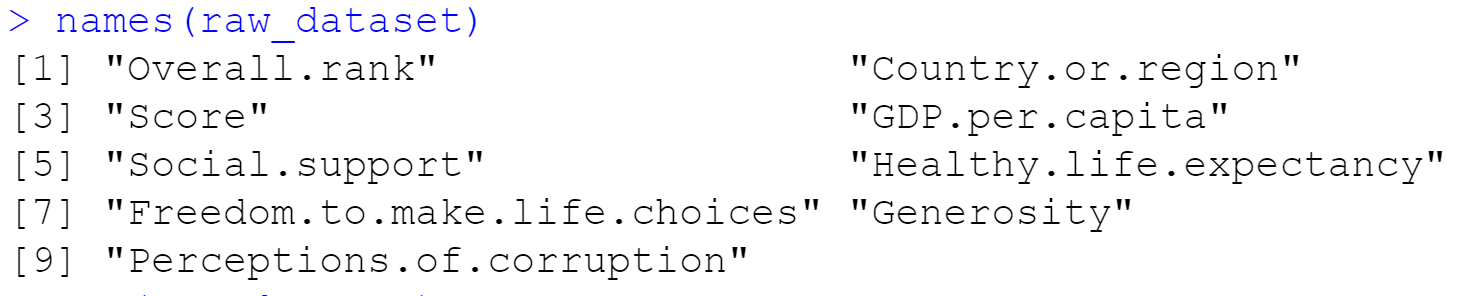
Introduce the dataset

Methods of statistical analysis on this dataset to extract information from it. We will try to contextualize this information for further usage. The dataset we will be using today is in reference to the Happiness of many different countries. There are a few different variables all which are correlated but have different weights on the happiness. We will look into this data to se how the correlating factors play out with each other.

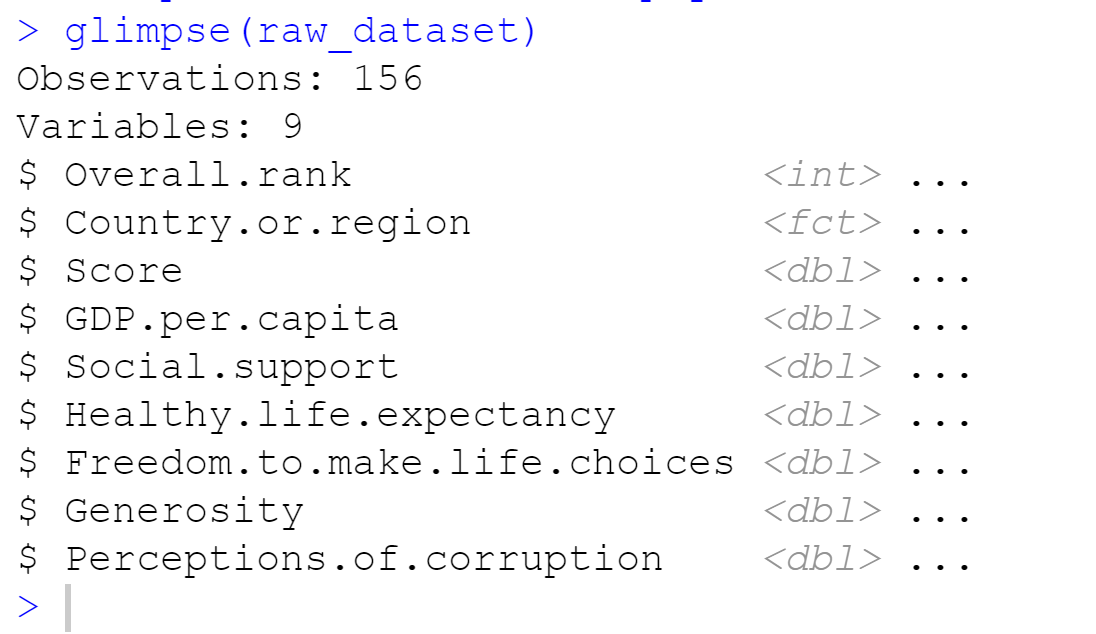
First in the process we will need to select and import our libraries. We will import dplyr,tidyverse, tidyr and broom for our main data cleaning. Ggplot2 will be our main tool for visualization with corrplot for correlation visualizations.



Once we have loaded the datasets we will load our data into R as a data frame and begin examining the data. The data we will be specifically looking refers to the year of 2019 even though there are other years in our dataset we will be focusing on the most recent. Now we begin reviewing our data at a arge level to get a general idea of what we are working with. We start with reviewing the names within the dataset to see how the columns are organized and the data they contain.



We can continue to view more information on the columns themselves by using the glimpse function on that dataset. Here we can see more of what the data types are within each column, dbl are really numeric, as well as notice there are 9 variables with 156 observations.



Cleaning the dataset is a big part of the process as it not only usually takes the longest but requires attention to fine details. The first thing we are going to do with our dataset is change the column Country.or.Region to just Country. When reviewing the dataset itself there we no regional values in that column so the region in the name was redundant. However, in the 2015 dataset there is region included as a separate column. The 2015 dataset is uploaded and then the column is added and organized by country to in 2019 dataset, our main dataset, using the dplyr functions.

After the dataset is review for missing values to correct if needed and then review all of the values possible within the column.



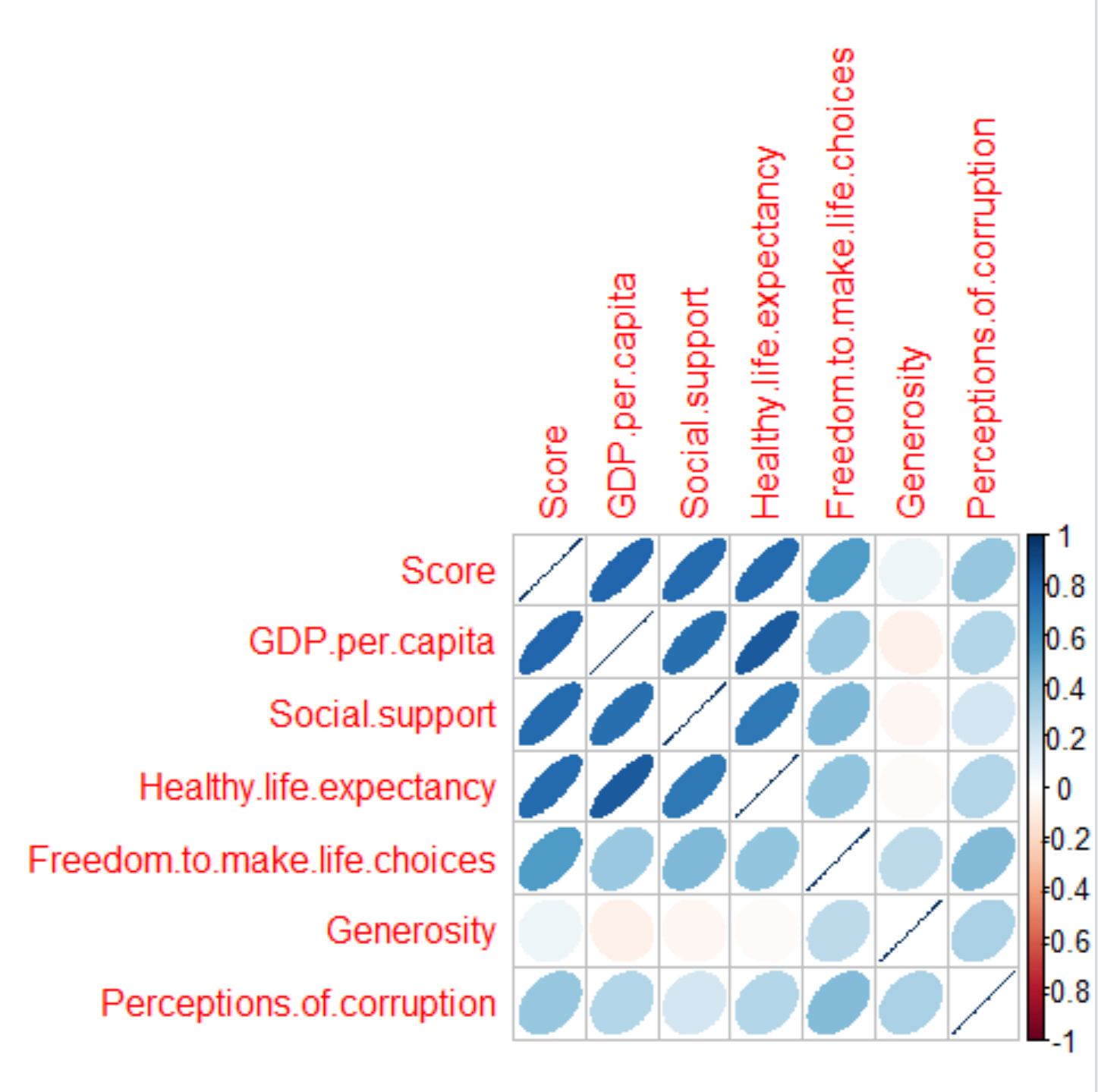
And then we count the number of entries per region



It appears the Regions with the most entries are “Sub-Saharan Africa”, and “Central and Eastern Europe” with “Latin America and Caribbean”, “Western Europe” following “Middle East and Northern Africa” following. “North America” and “Australia and New Zealand” have the lowest both with a count of 2.

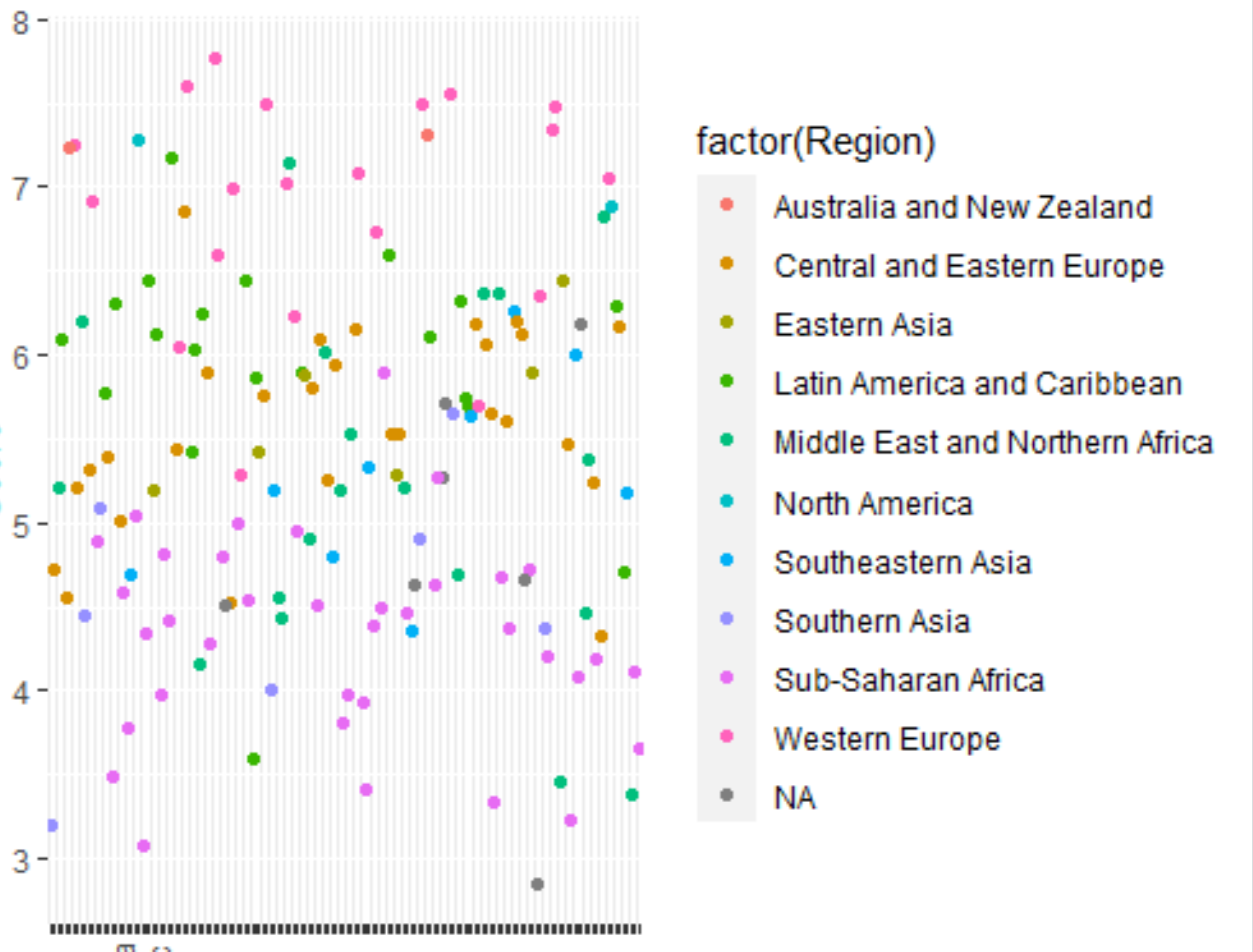
We can now create new data frames based on the categorized groups if we need them for further exploration.

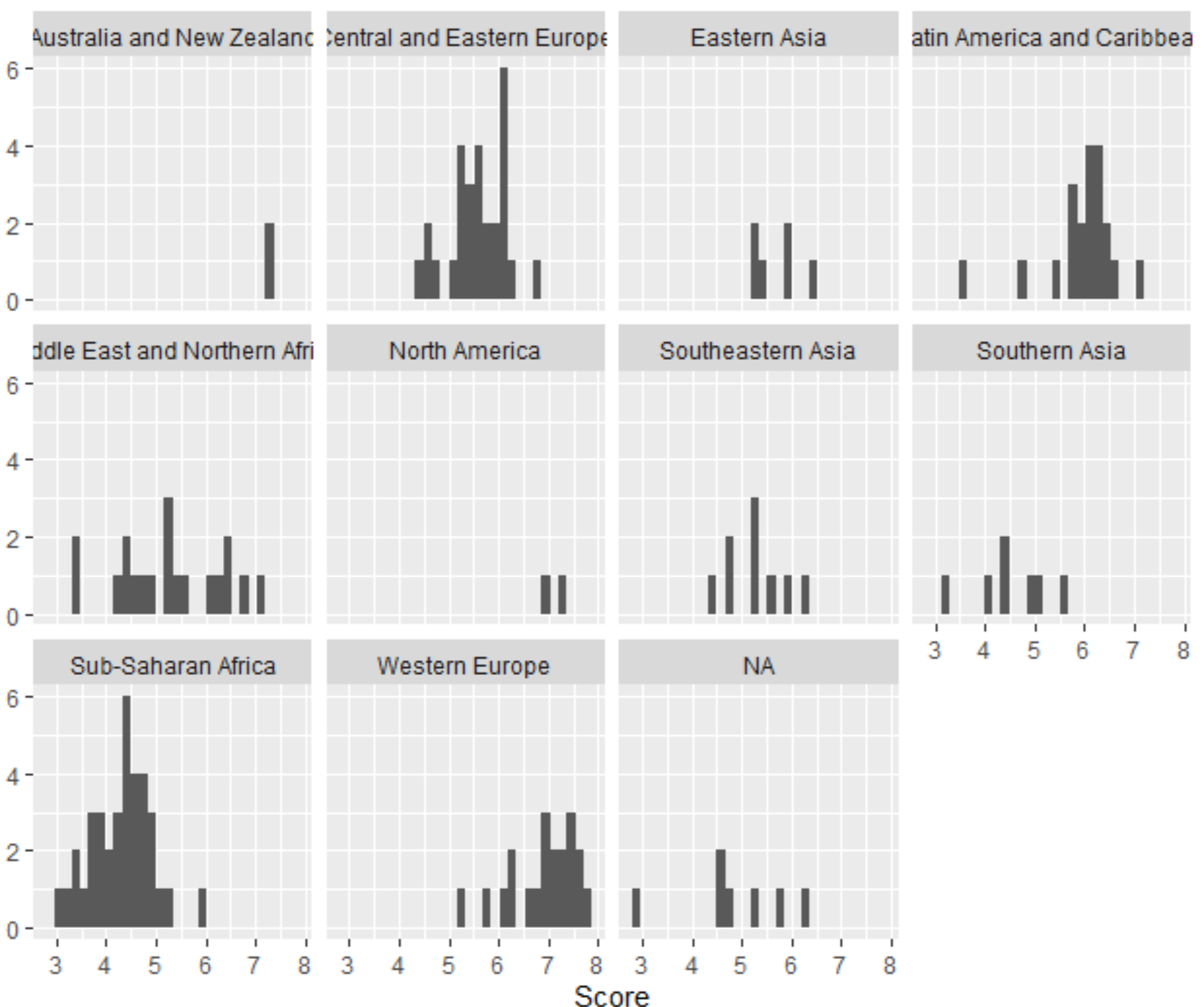
Next we will use data visualizations such as scatter plots, box plots, and histograms to analyze the numerical variables within our dataset. To prepare our data and get a sense of what factors to pay attention to we will run a correlation coefficient on the numerical data within the dataset. This will show us how strong each variable is correlated. We are really looking for the strongest correlations with the Score which references the happiness score of each country although we can also see how each variable can have an impact on another.

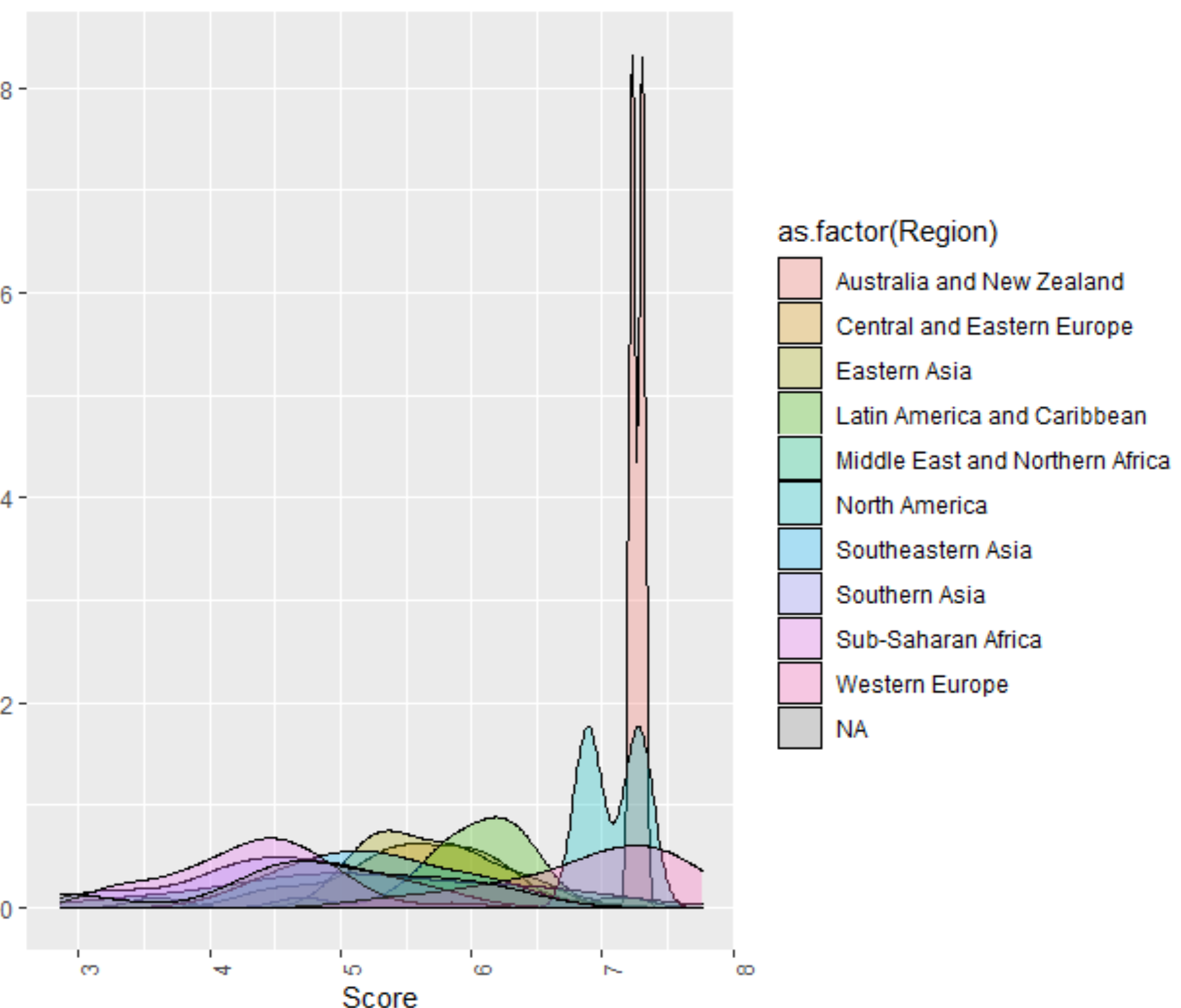


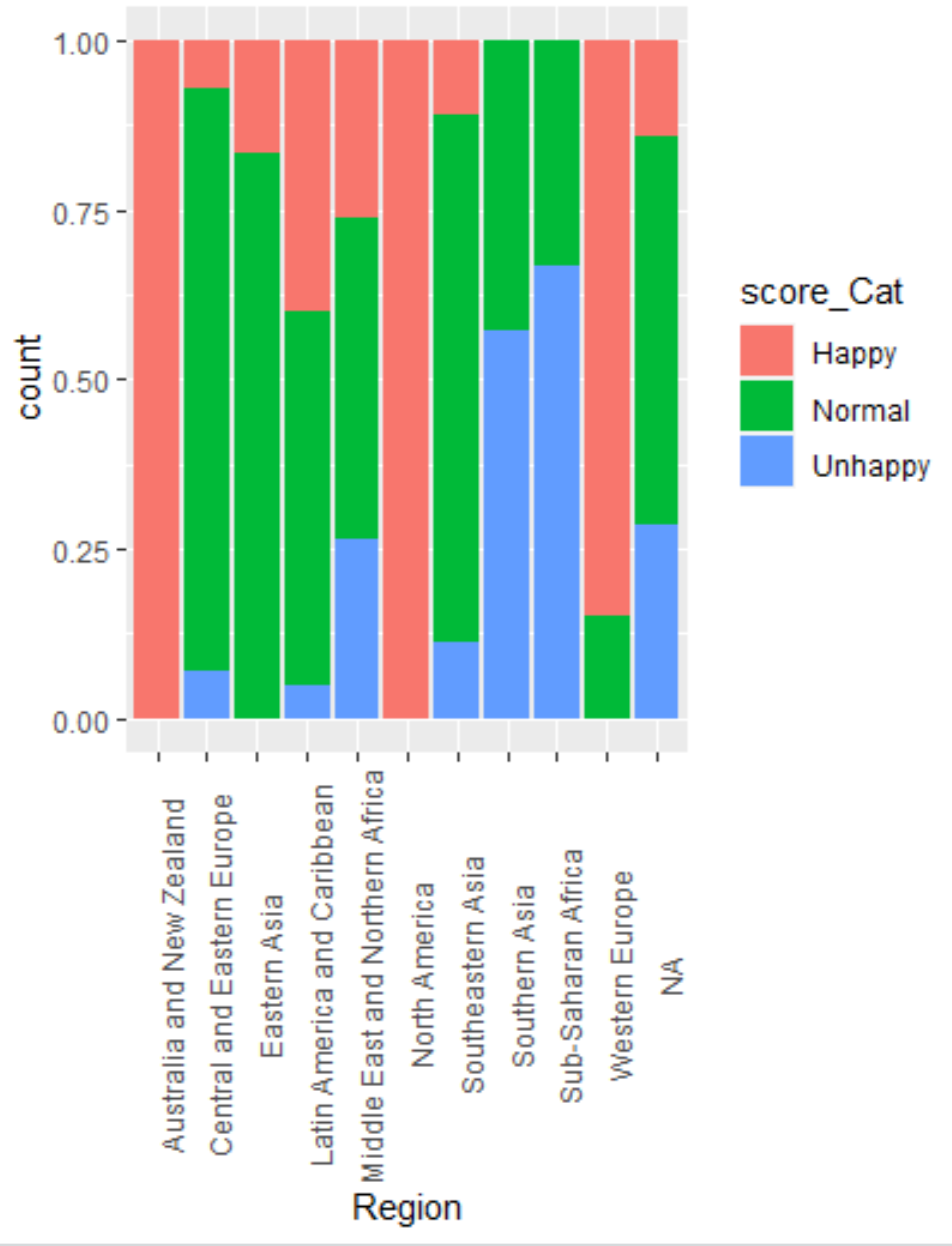
An interesting observation was how less Freedom.to.make.choices is correlated to the score in refence to GDP.per.capita and Social.support. Another interesting observation is GDP.per.capita and Healthy.life.expectancy. Next for further cleaning we created some categorical columns to add to the dataset. For each of the top 4 variables: “GDP.per.capita”, “Social.support”, “Healthy.life.expectancy”, and “Freedom.to.make.choices”.

Additionally, constructing a scatterplot and a histogram and help reveal more information within the data. First a scatterplot based on the country, and score with the color factored on region to see if there are any obvious point to view when looking at all of the point.



Next we did the same with a histogram to try to review the distributions within the regions themselves. As well as an overlaid density plot to review the data side by side and directly in relation to each other. 

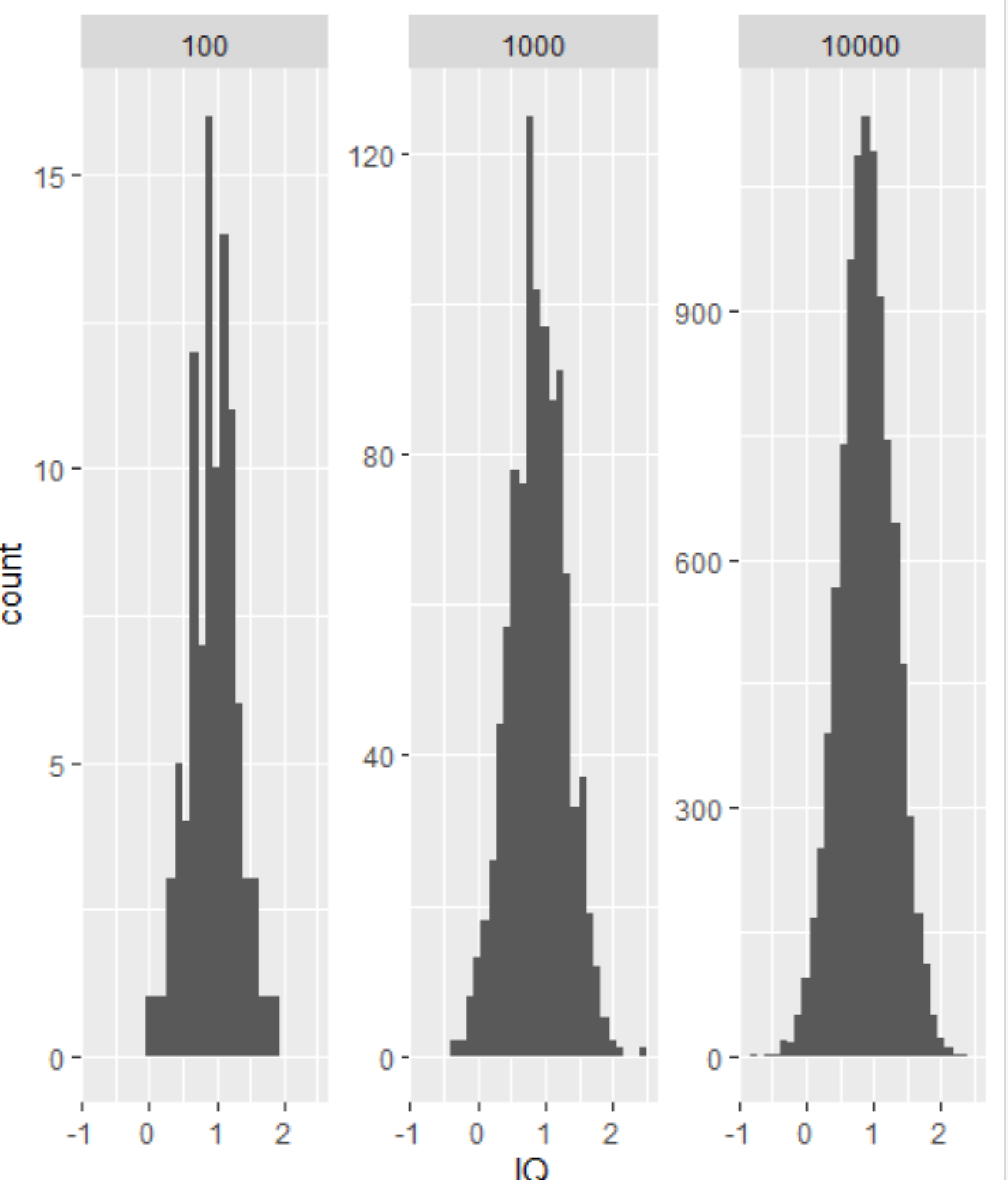




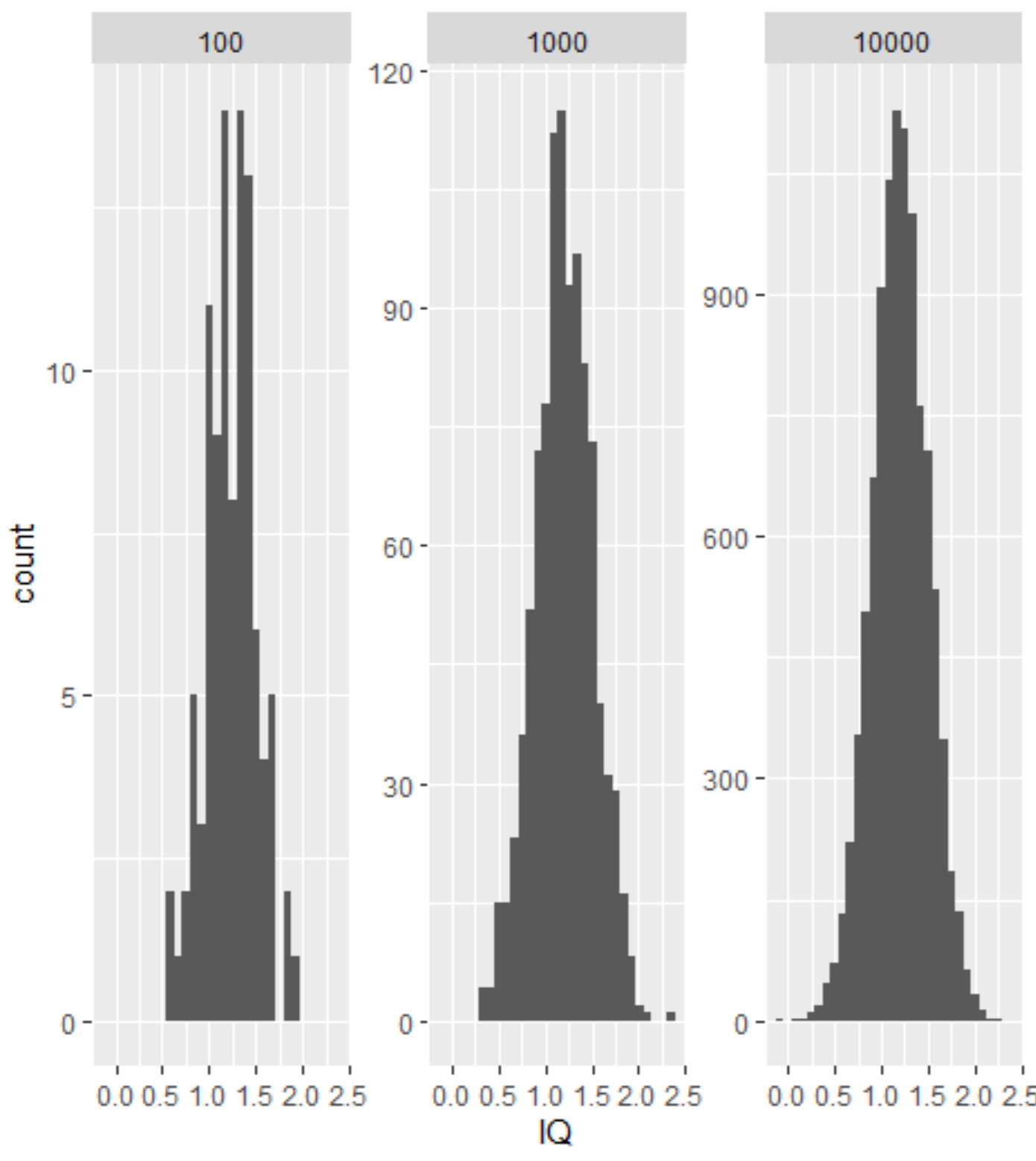
Here we can view these proportions in relation to each. Seeing Australia and New Zealand are the only 2 in their respective category it is no surprise that they have a complete bar. Other regions containing more countries such Latin America and the Caribbean have a god mix in relativity.

Now we can construct discrete probability distributions for some of our numerical data. We will use GDP.per.capita, Social.support, and Healthy.life.expectany as our main variables.Using the Law of Large Numbers to help show the type of distributions we have in our various categories.

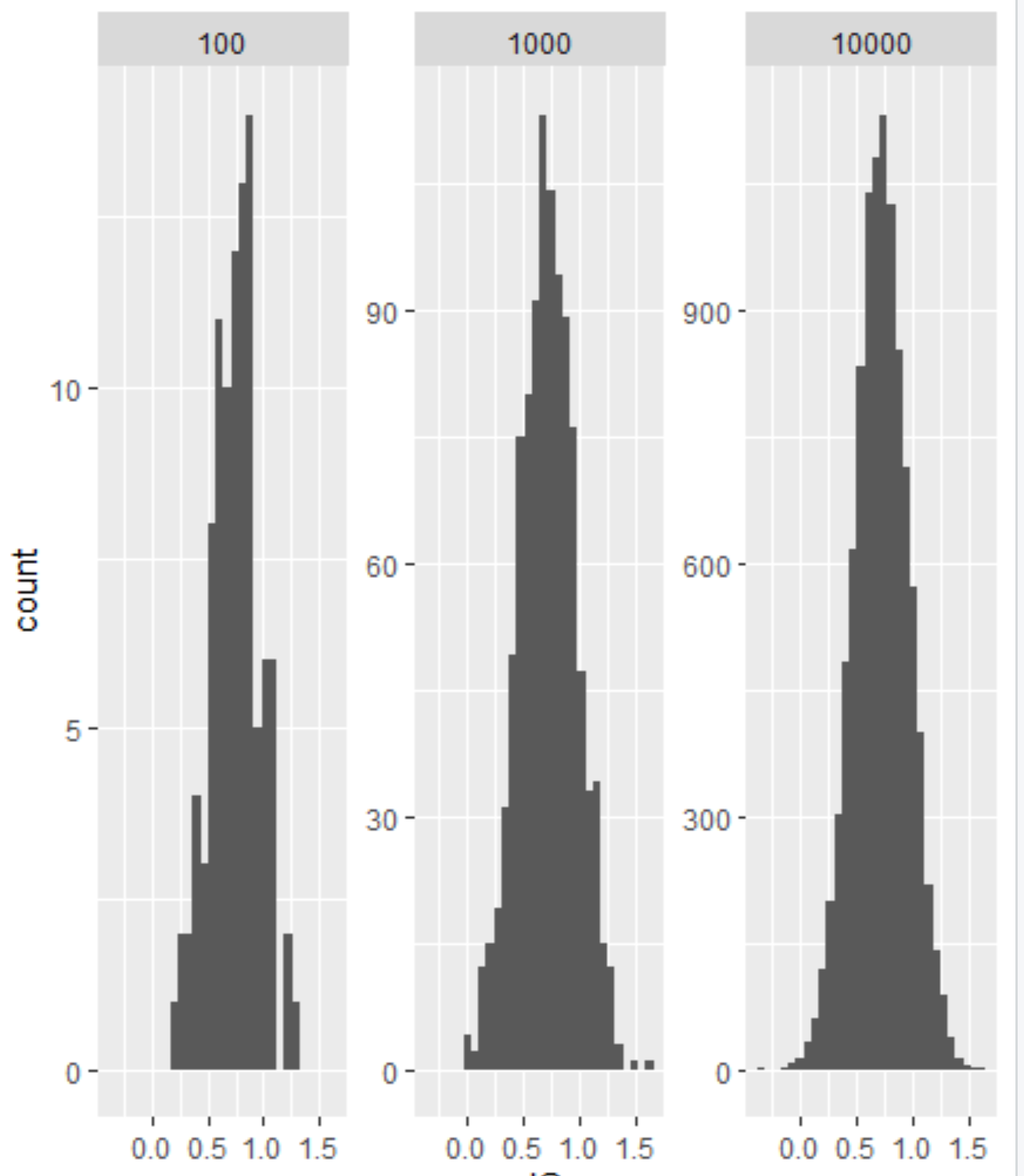
GDP Per Capita



Social Support

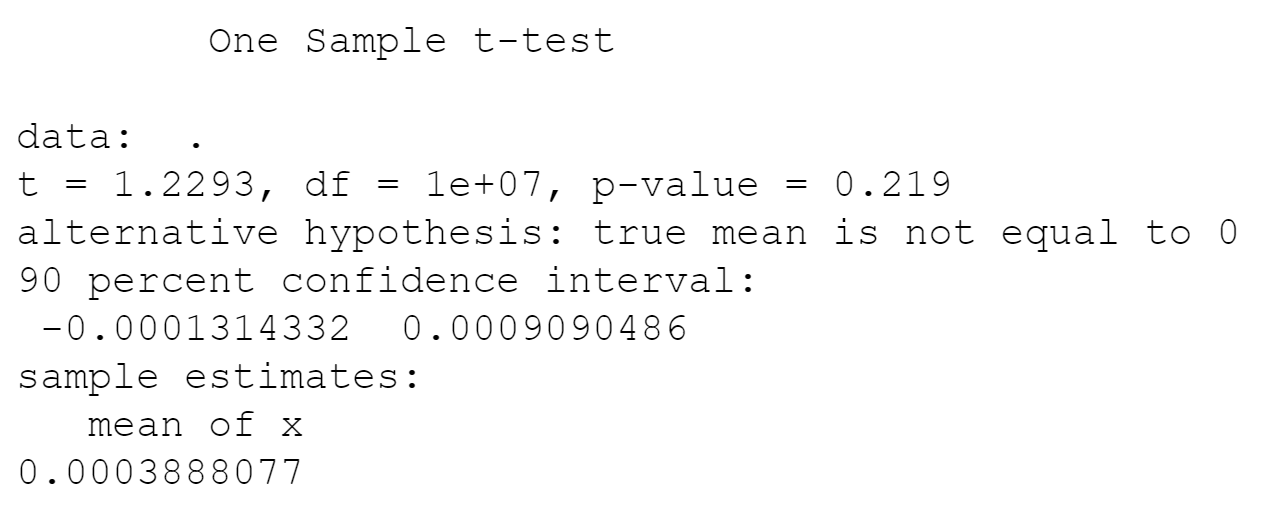


Healthy Life Expectancy

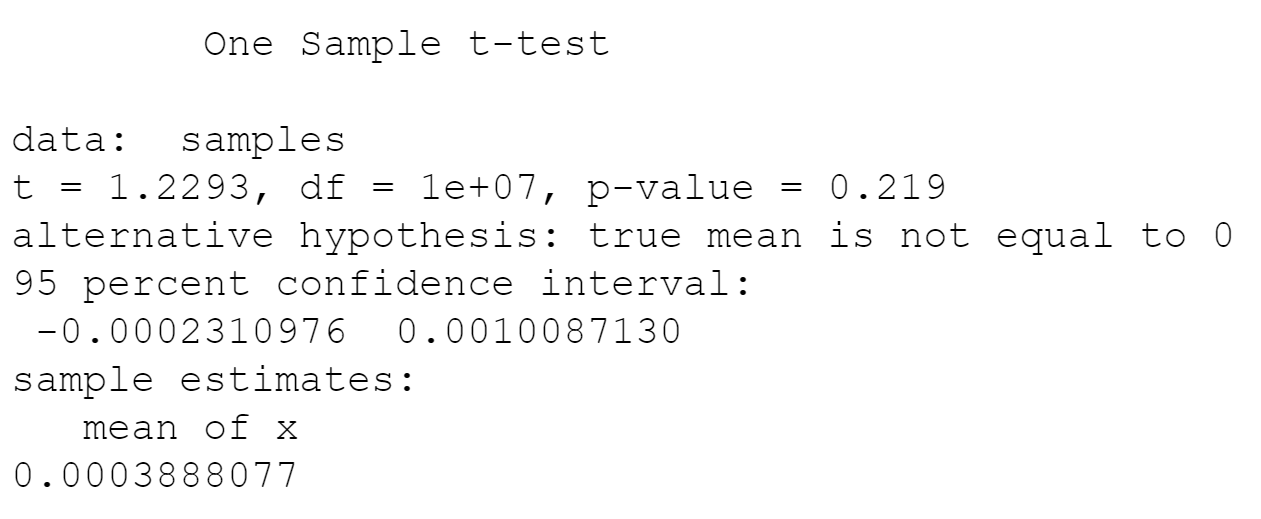


With these samples we can see the normal distribution become more and more apparent as the sample size increase, due to the weight of each individual point becoming less, and we can construct confidence intervals for sample means.

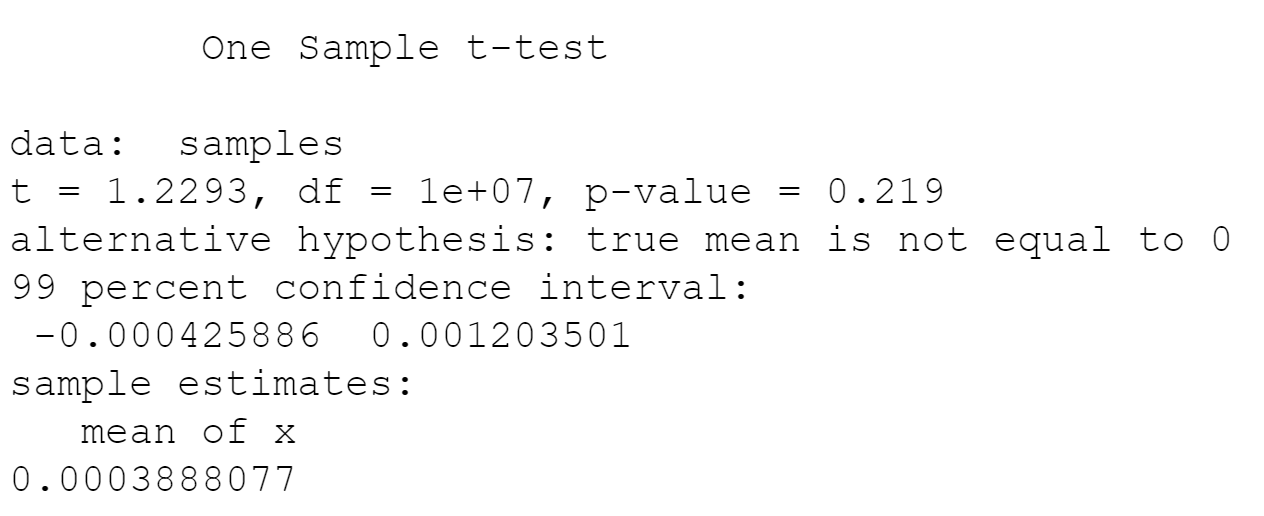
90% confidence:



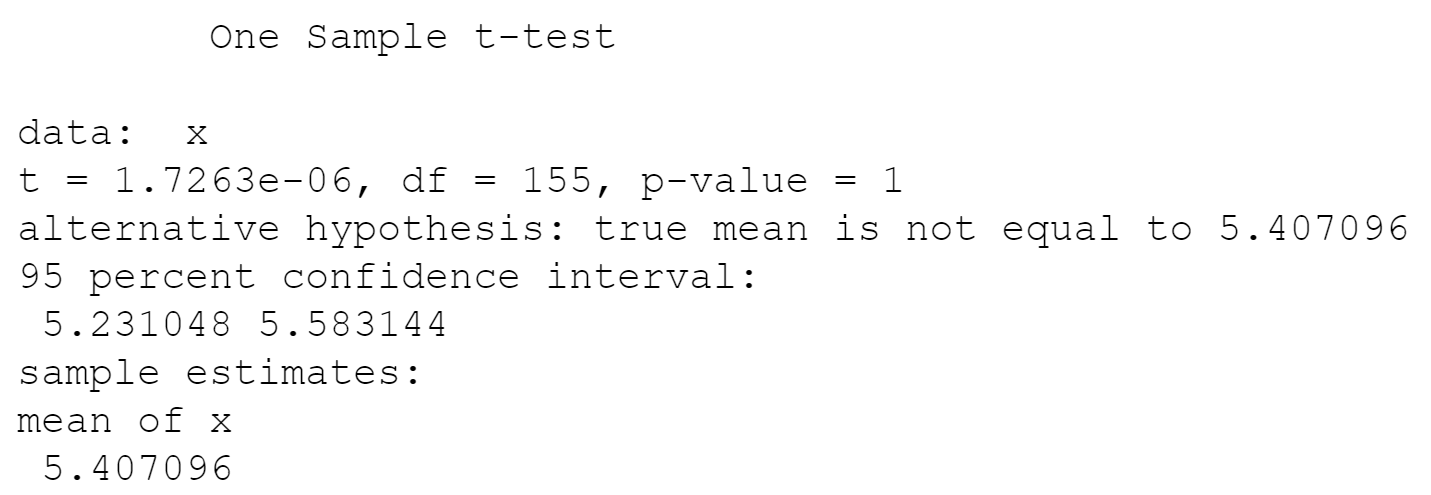
95% confidence:



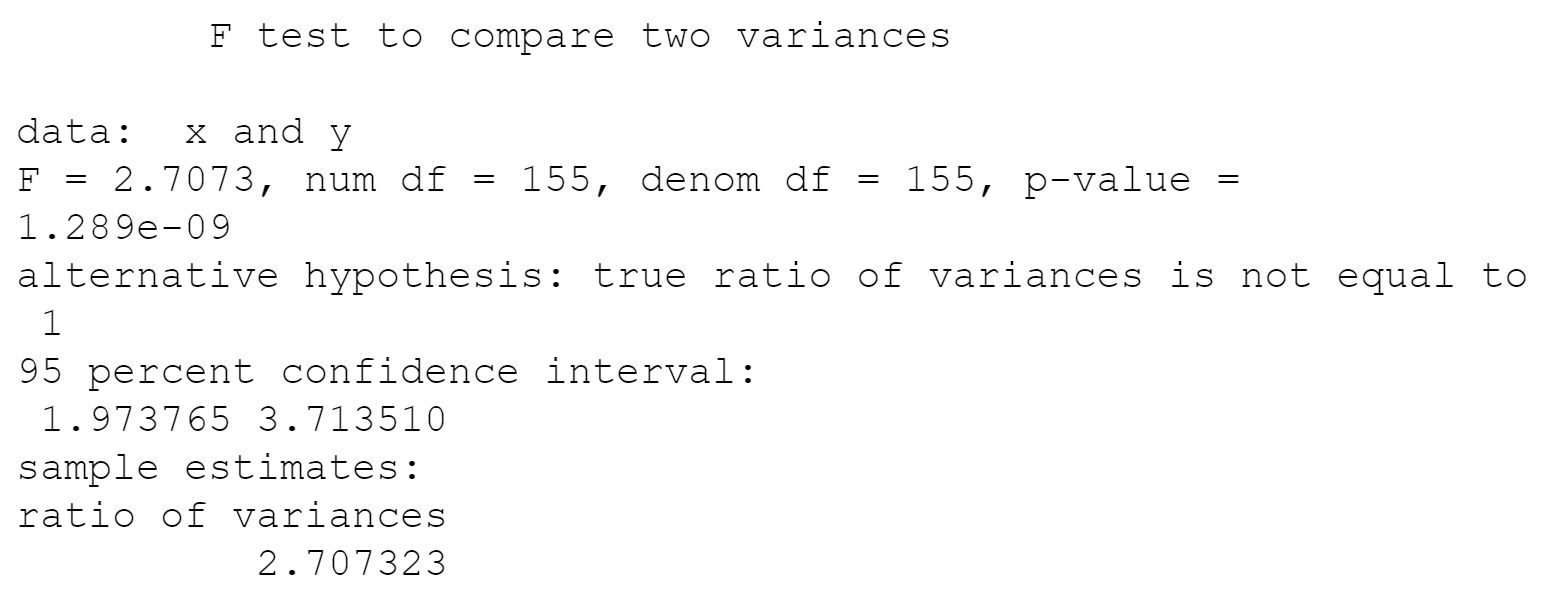
99% confidence:

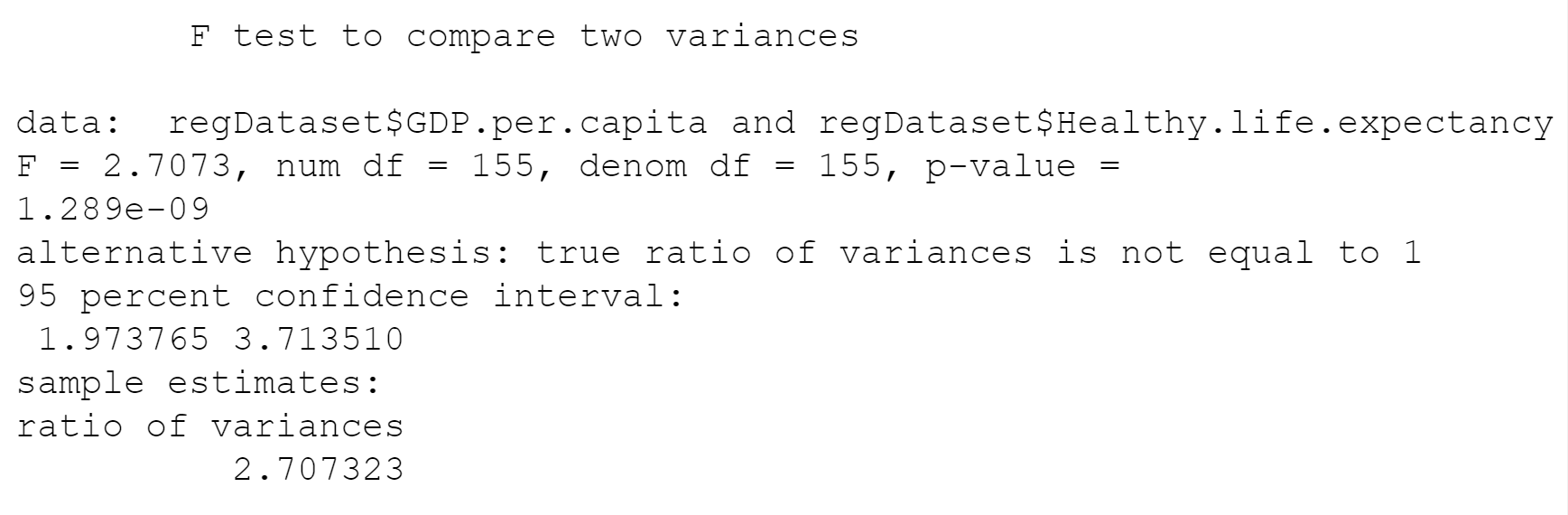


And the t-test on a single variable

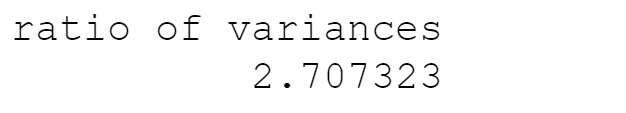


Next we will examine and compare the variances in the GDP.per.capita and Healthy.life.expectancy columns within our dataset though f-testing, or variance testing and examine the ratio of the variances as well as the pvalue.

# F-test



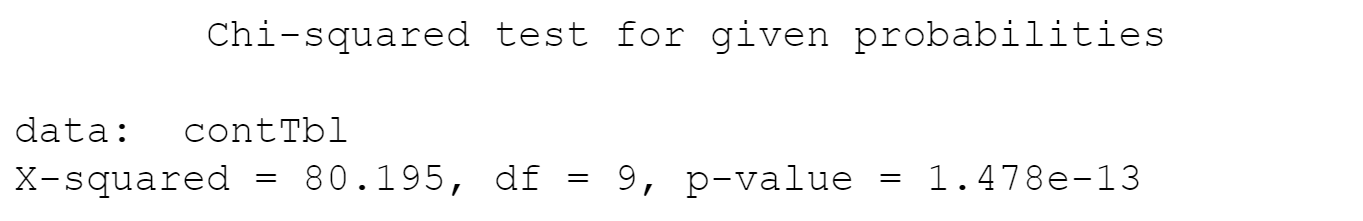
# ratio of variances



res.ftest$p.value

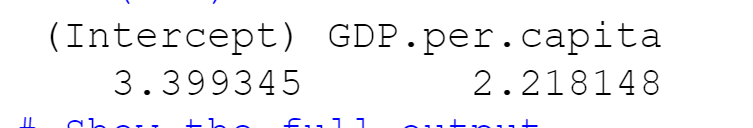


All of this does not answer how dependent the score was on any of our top three factors. Created a contingency table to prepare for the chi squared function. Inside the table we took the Score categories which determine if a country is listed as “Happy” and the region to try to view the independence between them through testing their variences.

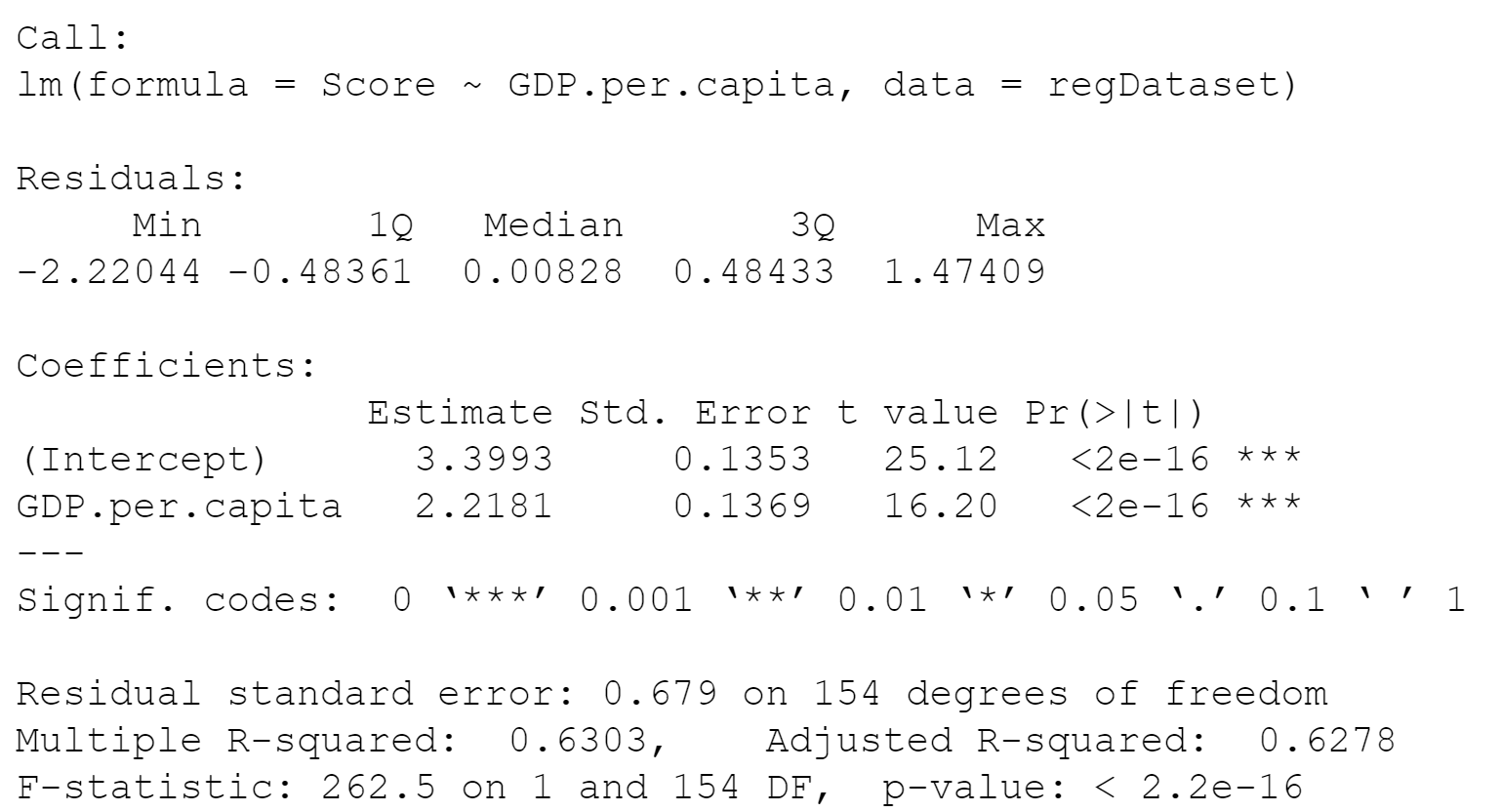


We will finish off with preparing out data for linear regression models which we will plot at with a scatterplot and add a regression line.

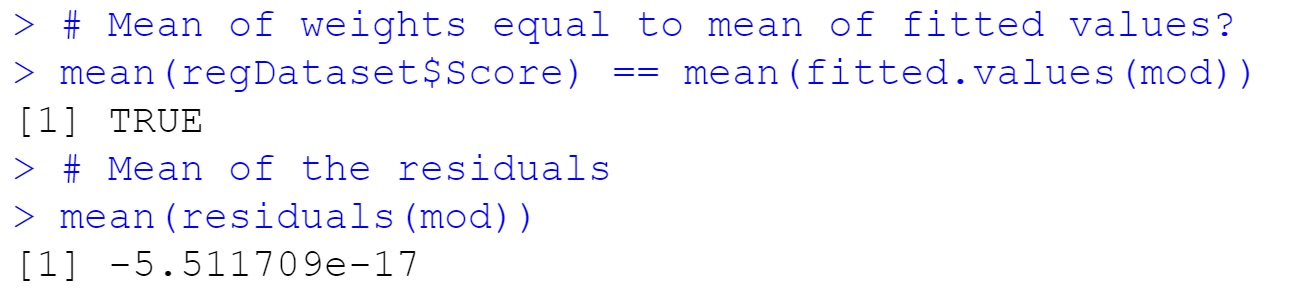
review fitted values of regression model and review their coefficients

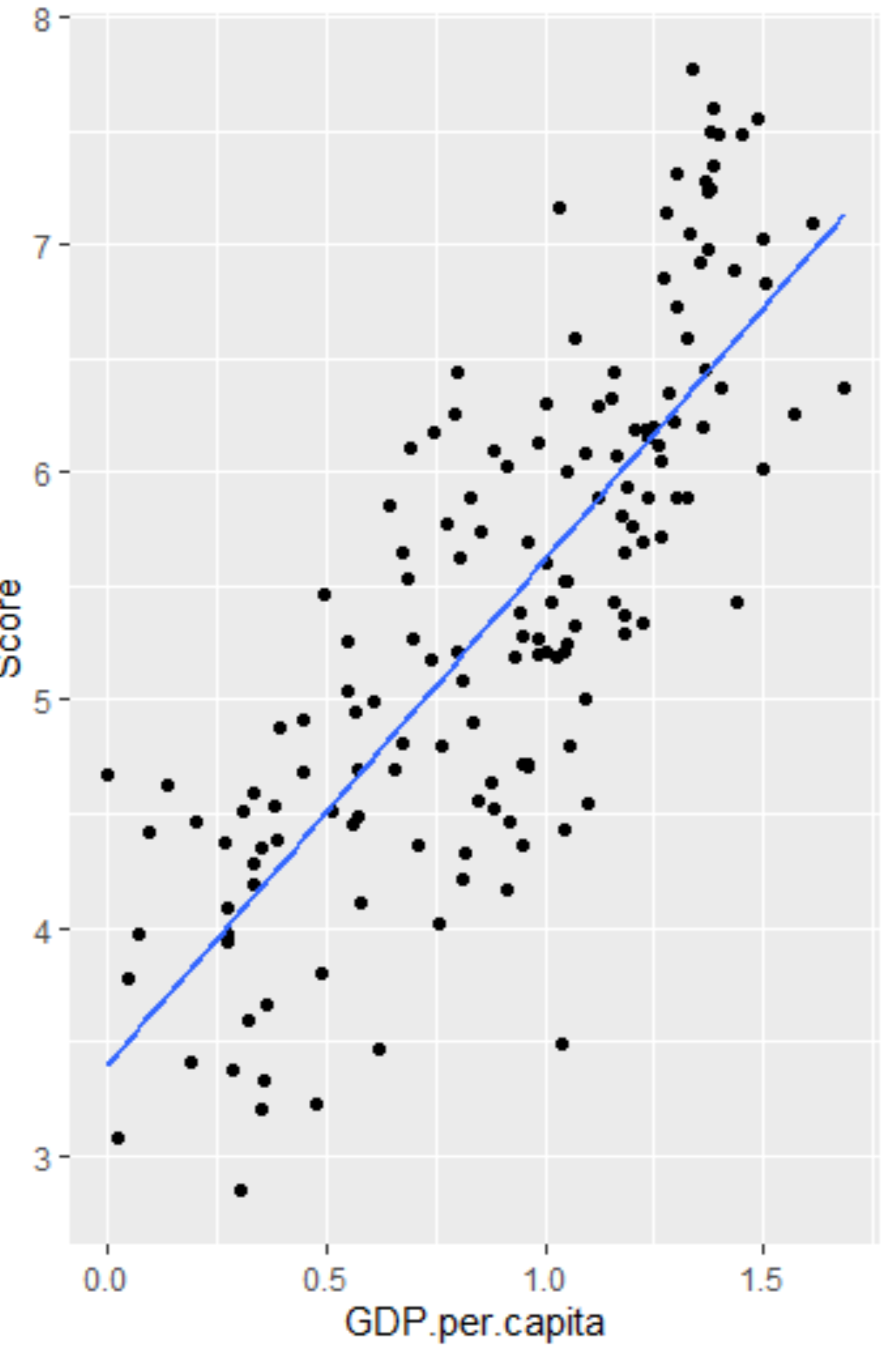


Full summary Output.



Review if the Mean of weights equal to mean of fitted values?





After viewing the scatterplots and linear models and correlating all of this data back with the corrplots earlier. From the observations GDP.per.capita, Healthy.life.expectancy, and Social.Support all have a positive correlation with a higher happiness score but the Score does not depend on any of the three more than the other. Freedom.to.make.life.choices was high on the correlation last but not as strong as the other three were. Also based on our data we can infer of a positive relationship between GDP.per.capita and Healthy.life.expectancy.