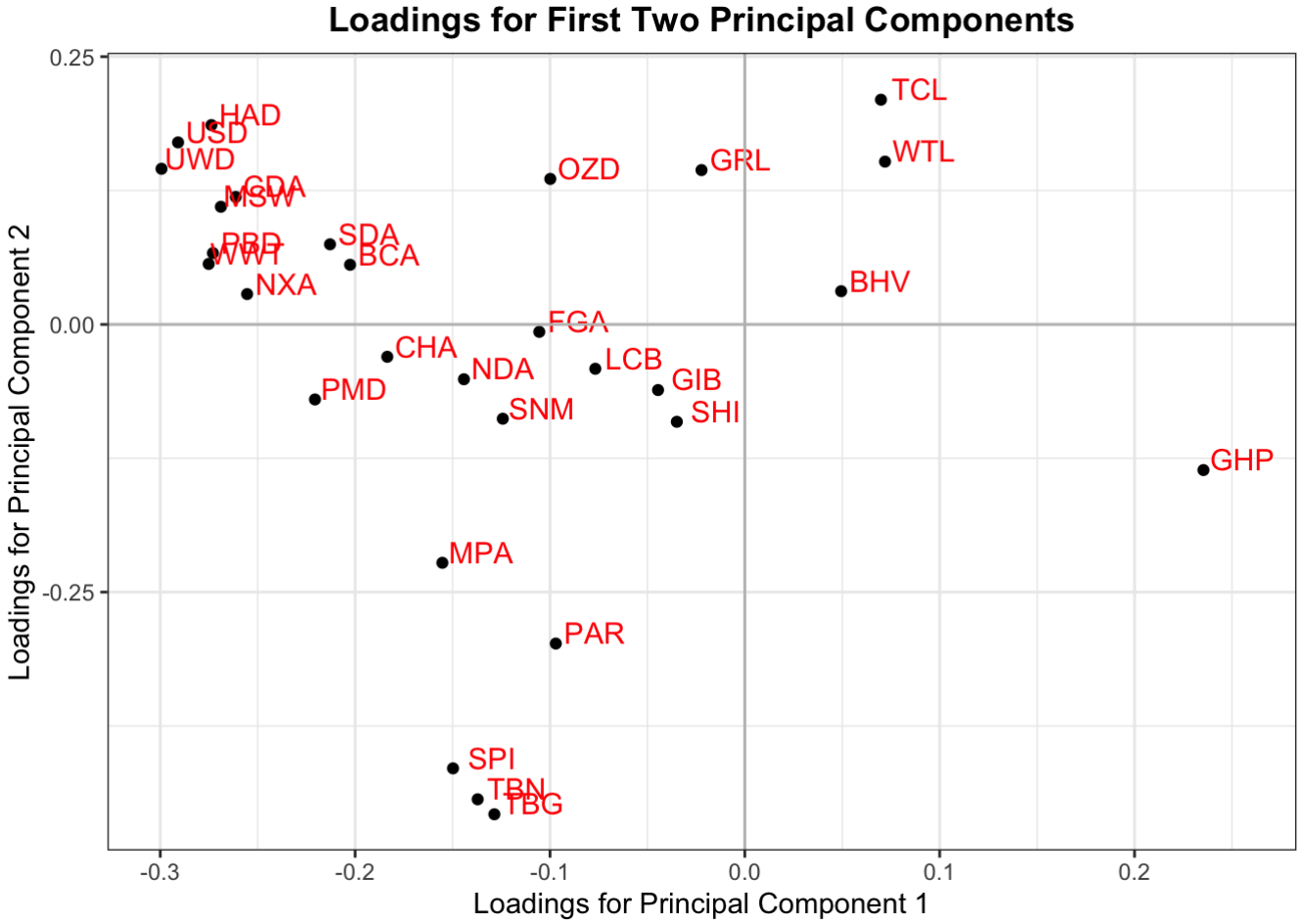
Data Analysis of Environmental Data

Huong Ngo

Introduction

The datasets used for this project are from the Environmental Performance Index project by Yale University. The first dataset contains the environmental indicators that the project has measured for each country. The environmental indicators are abbreviated in the dataset and their full names can be referred in the technical appendix. The second dataset contains information about the region of the specific country, a binary variable indicating whether the country is a developer or less developed country, another binary variable indicating whether the country has emerging markets, the GDP of a specific country and the composite indicator of environmental performance (EPI) of the country.

Loadings from First Two Principal Components



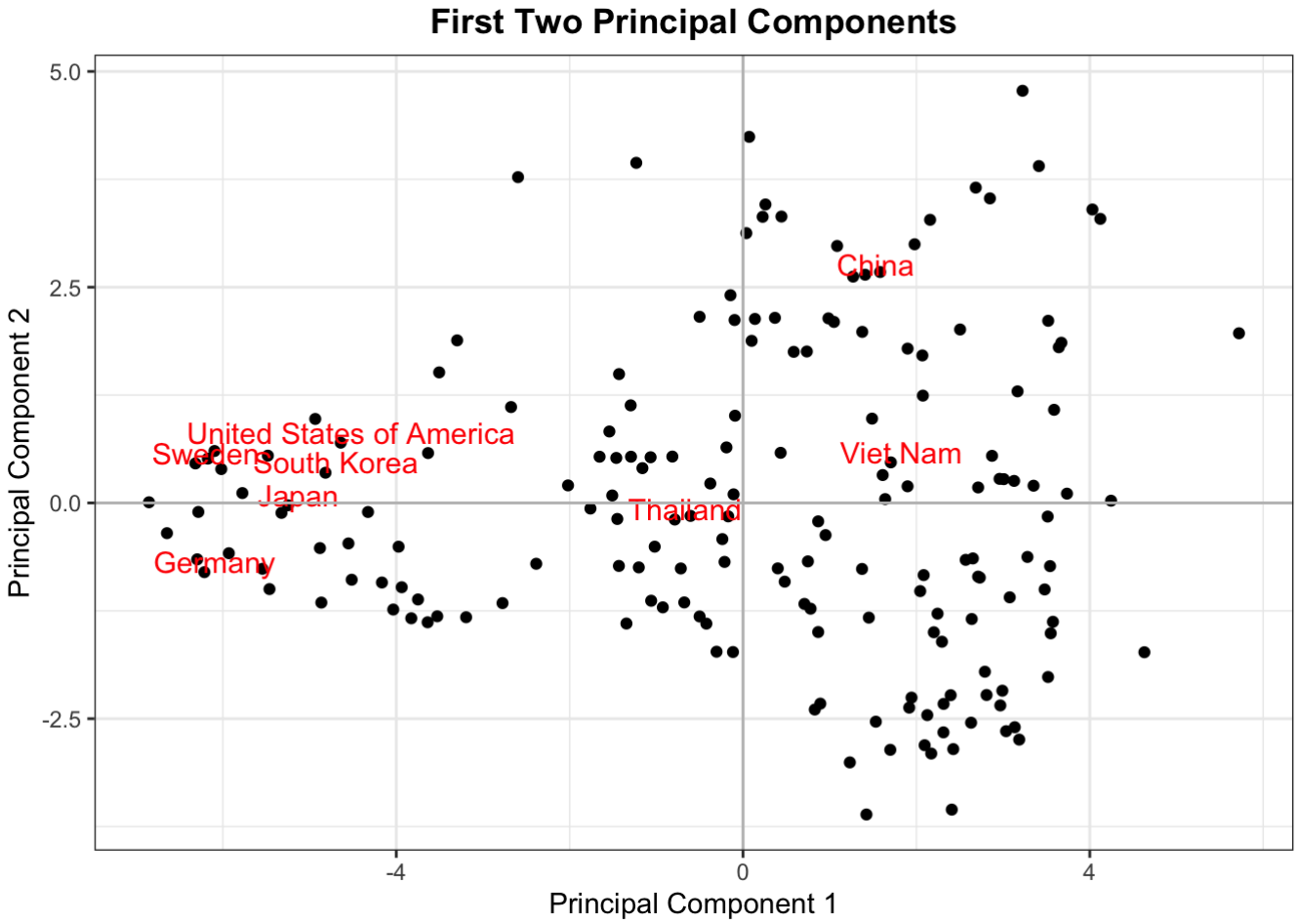
The plot above represents the loadings (coefficients) for the first two principal components. The x-axis has the loadings for principal component 1, and the y-axis has the loadings for principal component 2.

Firstly, we can note that we can't discard the second principal component or the first one because both contribute a great amount of variation captured in the dataset. The only variable that has its variation almost fully captured by the first principal component is FGA - F.gas Growth Rate.

Another thing we can note is that majority of the variables have negative coefficients in the first principal component. Along with that, GHP - total greenhouse gas (GHG) emissions per capita - is the driving factor in the principal component. We can interpret this principal component to be the effects of greenhouse gases on the other indicators of the environment as we can expect that more greenhouse gases means less of good environmental health and vitality overall (positive coefficient for GHP, negative for most of the other variables). We can be even more specific and see that the impact of greenhouse gases is very strong given that its weight (given by the dataset sourcer) is far less than the weight of other variables like PBD - Lead Exposure, MSW - Controlled Solid Waste, but it is dominant in the first principal components.

Another relationship can be interpreted from the loadings of the second component. Seeing that TCL - Tree Cover Loss and WTL - Wetland Loss are dominating coefficients, and on the other extreme, TBG - Terrestrial Biome Protection (Global) and SPI - Species Protection Index have negative coefficients, we can interpret the second principal components to reflect the harmful environmental actions to cause loss of species and damage of biomes.

First Two Principal Components

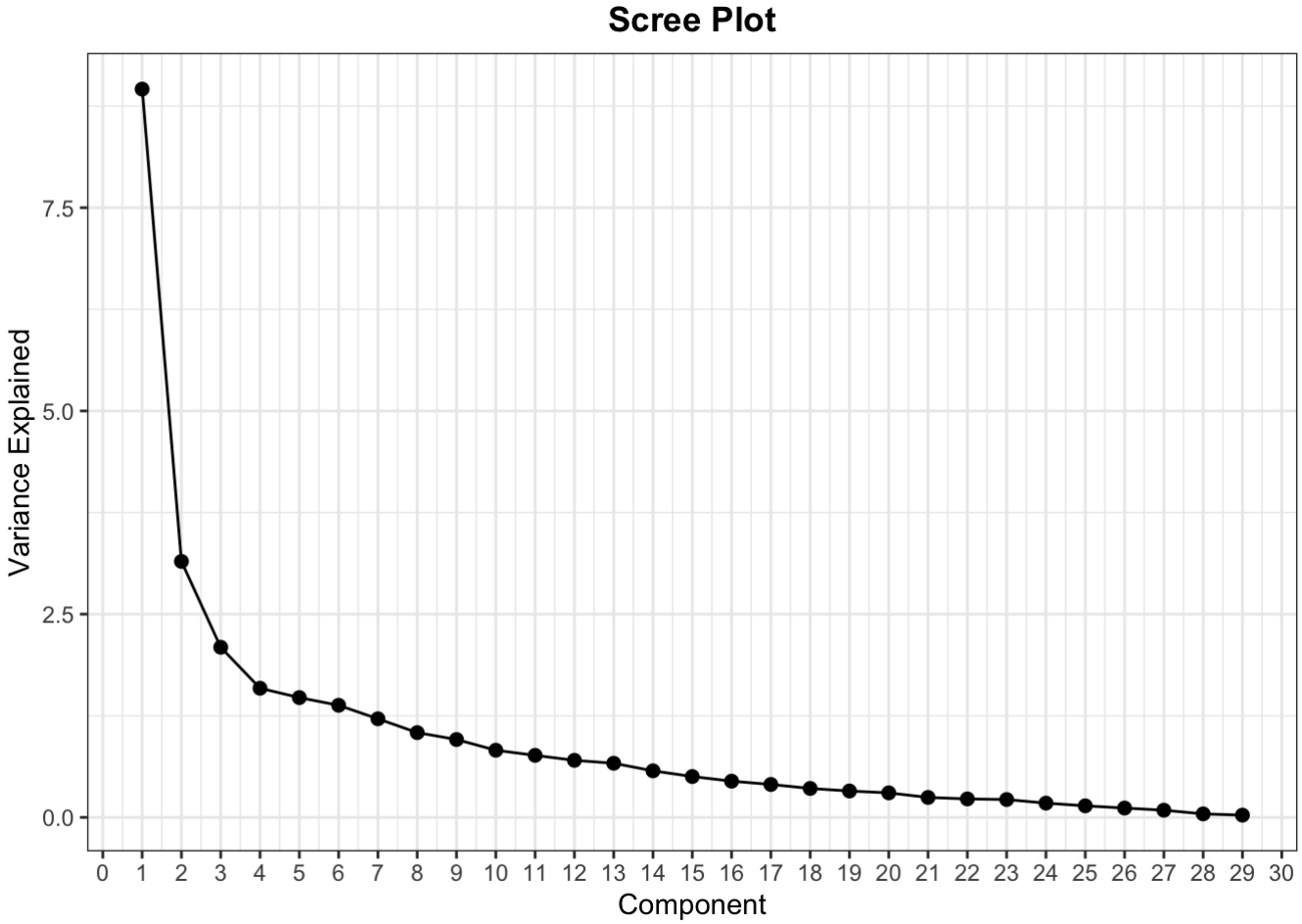


The plot above represents the first two principal components. The x-axis is the principal component 1, and the y-axis is the principal component 2.

From the labeled representative countries on the scatterplot, we can see that there is a group of countries that are quite similar in terms of environmental factors. United States of America, Sweden, South Korea and Japan can be seen quite close to each other and forming a cluster. We can interpret this as how these countries share many similarities in values of the environmental indicators, which could further mean that they are quite similar in environmental goals, policies and progress. Thailand, China and Vietnam on the other hand are quite far from

each other and from that group, with China being the furthest from the cluster. We can interpret this as how these countries share very few similarities in terms of how they deal with their local environment and that they also don't share similar environmental principles with the countries in the cluster, with China sharing the least similarities. The climate of those countries could explain why they don't share much similarties with the ones in the cluster.

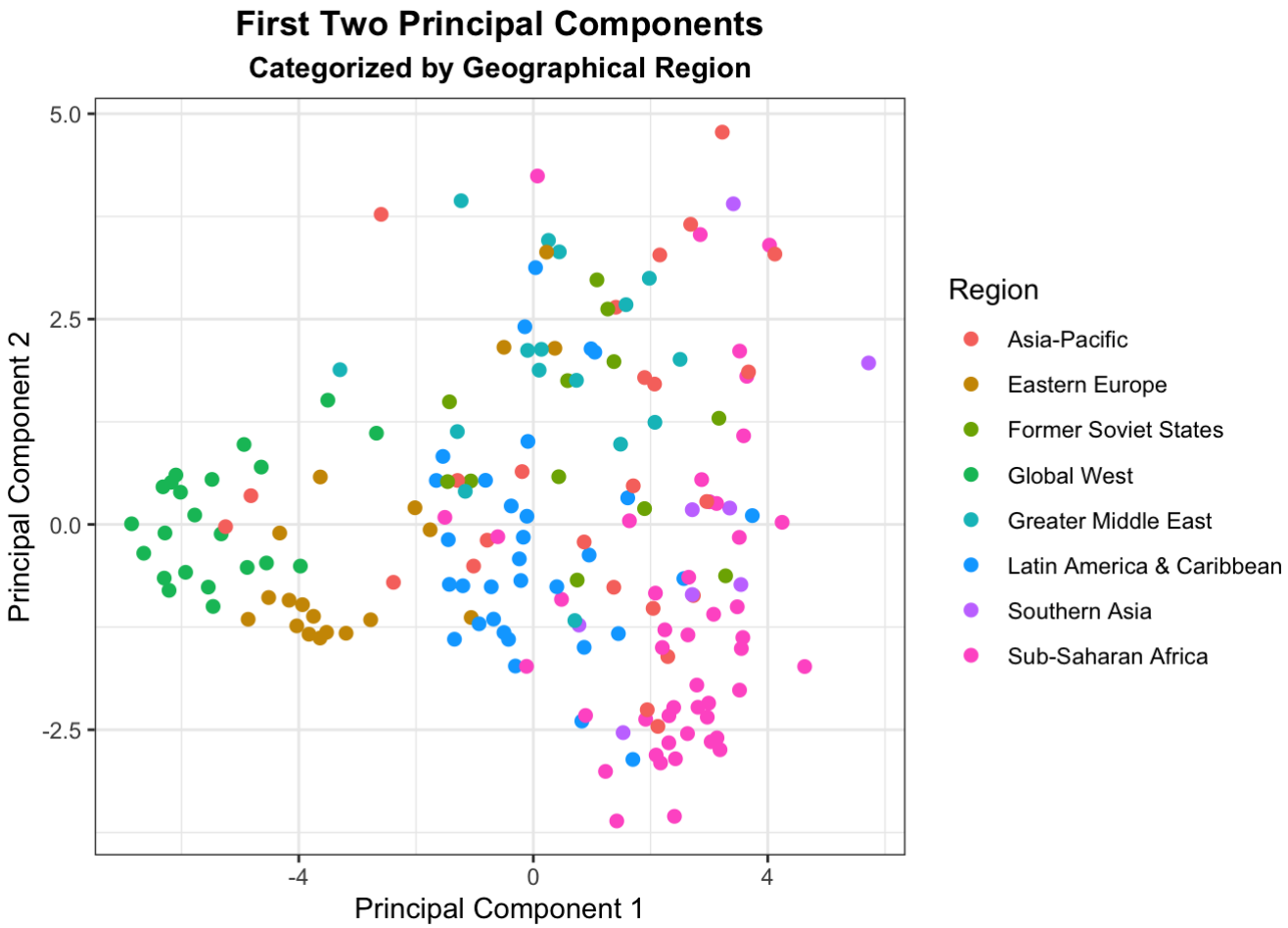
Choosing the Ideal Number of Principal Components



The plot above represents the scree plot of the principal components. The x-axis represents the variance explained by a principal component, and the y-axis has the principal component number/index.

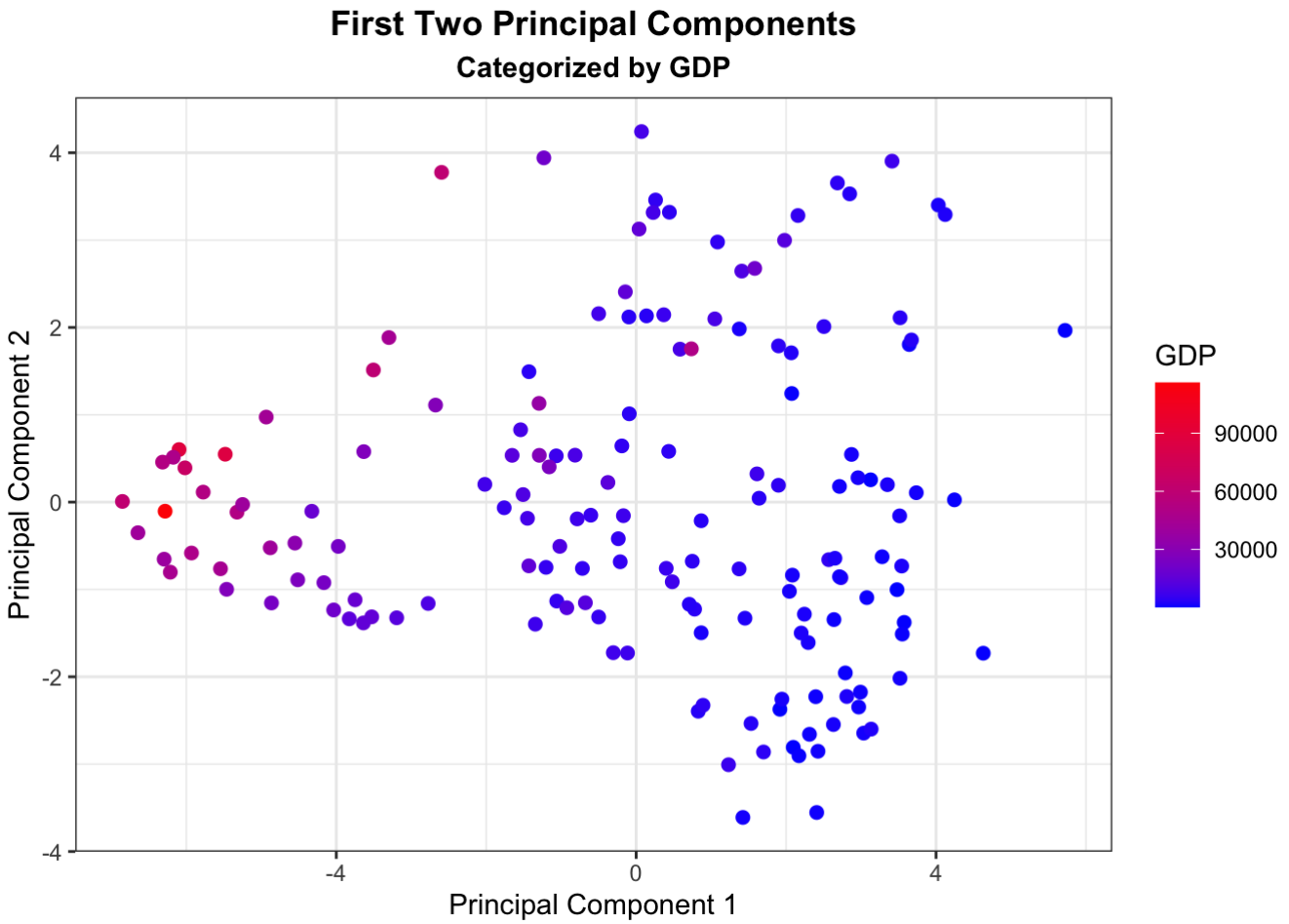
Three principal components is the ideal number to use. With a scree plot, we want to look for the "elbow" - a specific component number where after that component, the difference in variance starts looking consistent/the change in variance captured by each principal component is not drastic relative to the first few changes. Above, we can see that the change starts becoming not drastic after the third principal component. Therefore, we just want to pick the first, second and third. We also don't want to pick more given that it can be extremely difficult to create graphs in 4 dimensions/more and interpret them.

EDA with PCA



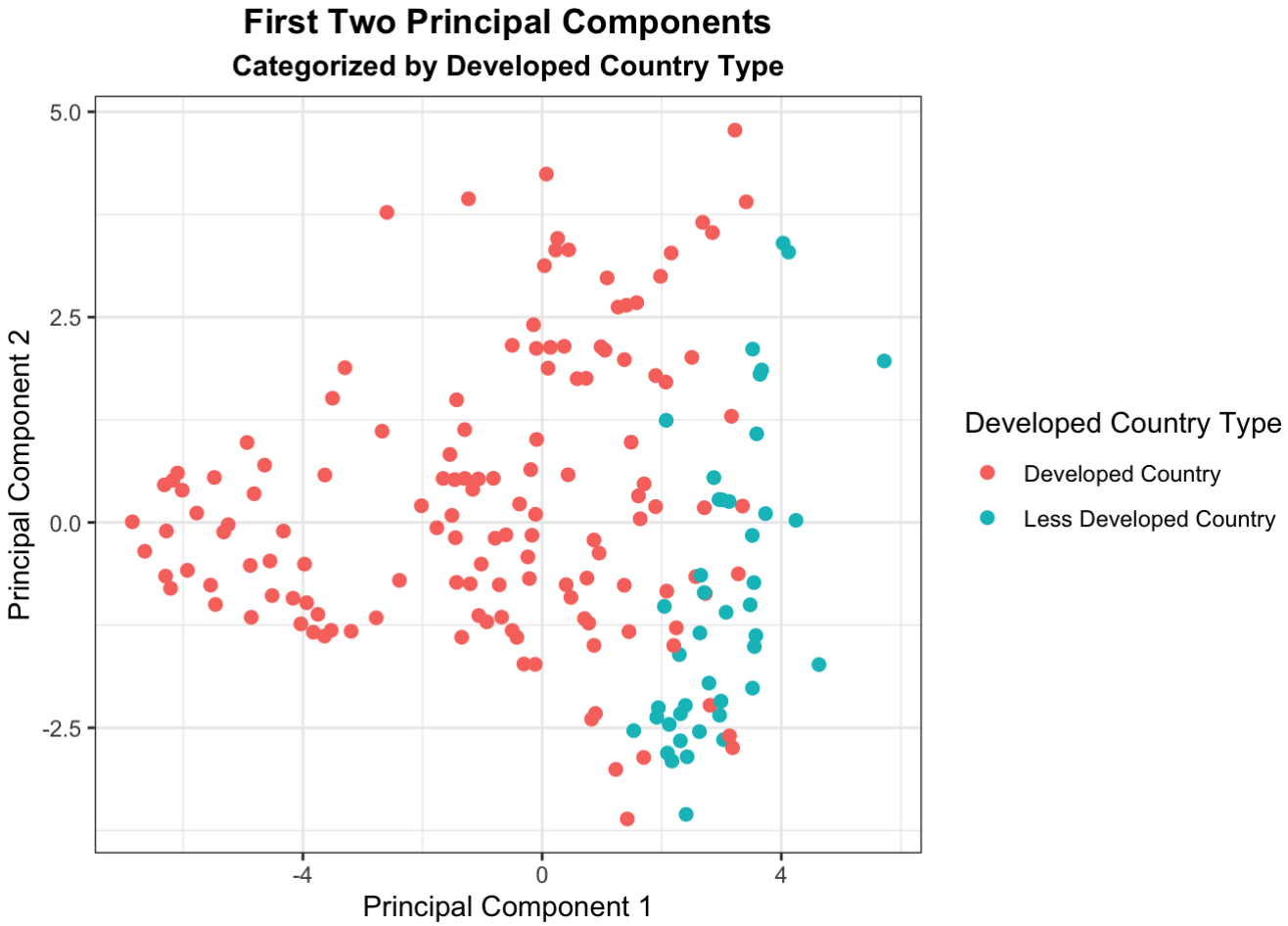
The plot above represents the first two principal components, but the country variables are categorized by their associated geographical region. The x-axis is the first principal component, and the y-axis is the second principal component.

From the scatterplot above, we can see that countries in the Global West, Latin America & Caribbean and Sub-Saharan Africa share many similarities about environmental indicator values. Essentially, the countries in their respective share many environmental aspects such as policies, goals and initiatives among each other. However, the other regions such as Asia-Pacific, Southern Asia and Greater Middle East have countries that don't share much similarities with each other compared to the former. This is quite interesting to note, and an aspect that can play a role into this phenomenon would be varying climates among the countries in the latter regions and how spread out the countries are within those regions.



The plot above represents the first two principal components, but the country variables are categorized by their associated GDP. The x-axis is the first principal component, and the y-axis is the second principal component.

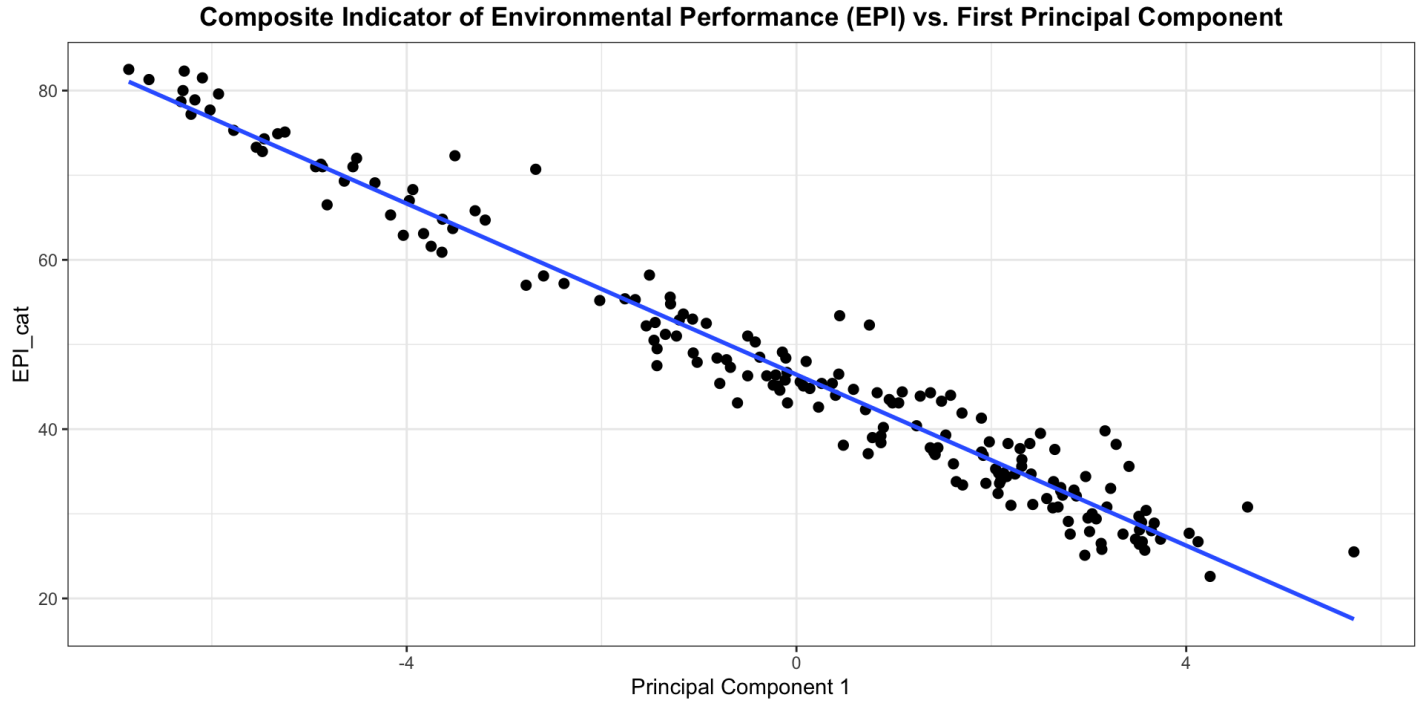
Looking at this figure, we can see that there are certain groups/clusters of countries formed by the differences in GDP. A notable thing we can take away from the figure is that countries with low GDP's are clustered on the right side of the plot, countries with the middle GDP's are clustered somewhat in the center and countries with high GDP's are clustered on the left side of the plot. Combining this with our interpretations of the first scatterplot, we can interpret GDP to be a good indicator of the environmental progress a country is making. This reflects how countries with a low GDP have less infrastructure and resources to improve their environment while other countries with a high GDP can spend a lot of money in improving and sustaining their local environment.



The plot above represents the first two principal components, but the country variables are categorized by whether they are a developed country or less developed country. The x-axis is the first principal component, and the y-axis is the second principal component.

The third scatterplot also reflects a similar sentiment as the second scatterplot, but uses a different category to emphasize the relationship between a country's economy and their environment. This can be seen where the less developed countries are clustered on the right side of the scatterplot and the more developed countries are clustered on the left. A less developed country has economic effects (lower GDP) and will have less resources and capital to work on their environmental progress.

EPI vs. First Principal Component




```
##                pc1    epi_cat
## pc1      1.0000000 -0.9783297
## epi_cat -0.9783297  1.0000000
```

The plot above is a scatterplot to showcase the relationship between the EPI variable and first principal component. The x-axis is the first principal component, and the y-axis is the EPI variable.

To compare the first principal component to the variable EPI in the second dataset, we can look at the scatterplot above and reference the correlation coefficient: `-0.978`.

Looking at the scatterplot, it is clear that their a linear and inverse relationship between the 2 variables. This is also confirmed by the negative correlation coefficient, which also means that there is a negative correlation between these 2 variables. Given that they share the same best fit line, we can see that they are similar to each other. Moreover, given that the first principal component usually captures the most variance in the dataset and EPI is a variable used to summarize all the indicators, they should be similar. We can take note that the negative slope could originate from the principal components being a product of the scaled dataset, while the EPI values could be directly from the dataset. That's also probably why the EPI values are so much larger in magnitude that the coefficients of the first principal component.

Conclusion

From my exploration, I learned a few main things. Firstly, greenhouses gases is a leading factor for the performance of other environmental indicators, and it has a negative correlation with a lot of environmental indicators. Secondly, I learned that loss of land is negatively correlated with how protected a biome can be. I also learned that certain countries share similarities in terms of environmental performance (and maybe as an extension environmental policies and goals) even if they might not be from the same geographical region. To add onto that, geographical region doesn't seem to be a factor in how "environmentally" similar countries are. In fact, there are regions with countries that don't share much similarities at all. Finally, I learned that a country's financial and development state is related to a country's environmental progress. From the visualizations above, we can make a conjecture that a country that has less capital and resources have a lower environmental performance and a more challenging time improving their local environment.