**Agenda:**

* Country Specific PCA
  + Data interpolation and extrapolation
* All Countries PCA
  + Additional data interpolation and extrapolation

**Country Specific PCA**

My goal was to compute the PCA for a specific country, where the rows of the input dataset were years and the columns were specific features. I computed a country-specific PCA for a country from each of the four World Bank defined income levels (low income, lower middle income, upper middle income, and high income). To compute these PCAs, I needed to perform data interpolation and extrapolation, as countries reported different indicators on different years, with some countries only reporting certain indicators after a specific year.

Data Interpolation:

I selected data between 2000 and 2019, as it appeared that the highest proportion of indicators had non-NAN values during this time period.

1. I removed all feature columns that had less than three non-NAN values, as I could not confidently determine whether this feature followed a linear trend from only 2 data points.
2. I determined whether the indicator’s data was linear or more random by finding the difference between each successive non-NAN value. If these differences were either all positive or all negative, I considered the feature to follow a linear trend, where the indicator consistently changed in one direction over time.
3. If the data did not follow this strictly linear trend, I considered the values to have the potential to change in either direction.
   1. I calculated the average of the feature’s non-NAN datapoints and replaced the NAN values with this average.
   2. This average was used for both interpolation and extrapolation purposes.
4. If the data followed a strictly linear trend, I used linear interpolation to approximate missing data points that were upper and lower bounded by existing non-NAN values.
   1. To do this, I used Panda’s interpolate method.

Linear Data Extrapolation:

Linear extrapolation formula:

I used this formula to perform both forward and backward linear extrapolation.

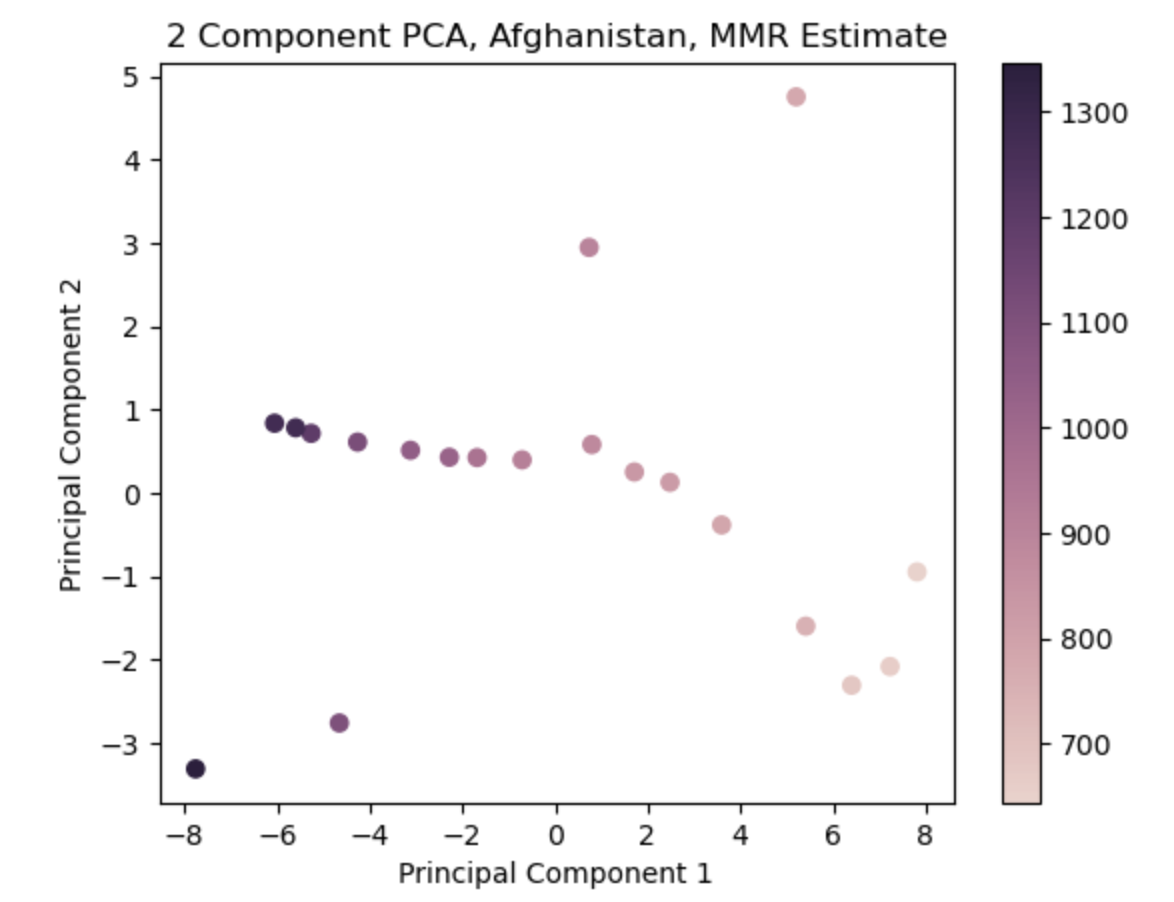
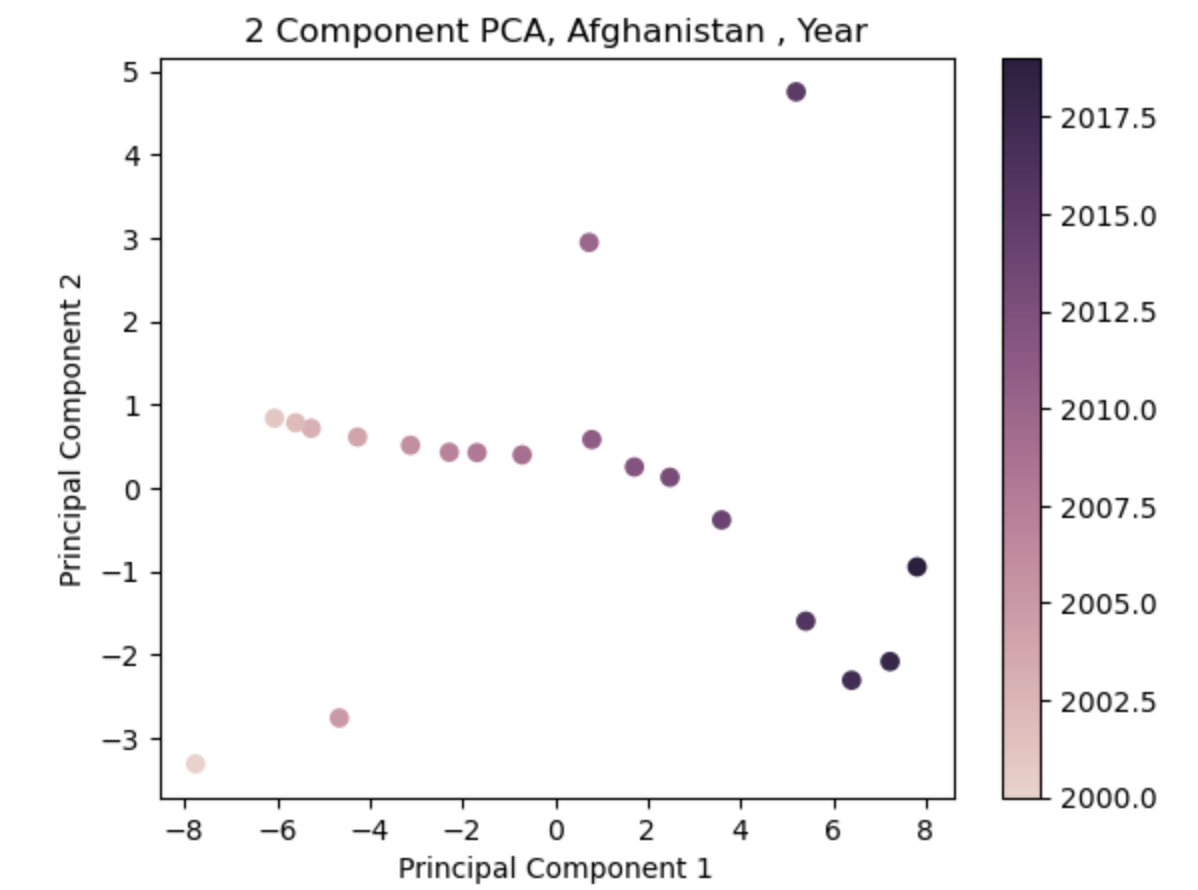
PCA:

After interpolating and extrapolating the data, I standardised the data so it could be used for PCA analysis.

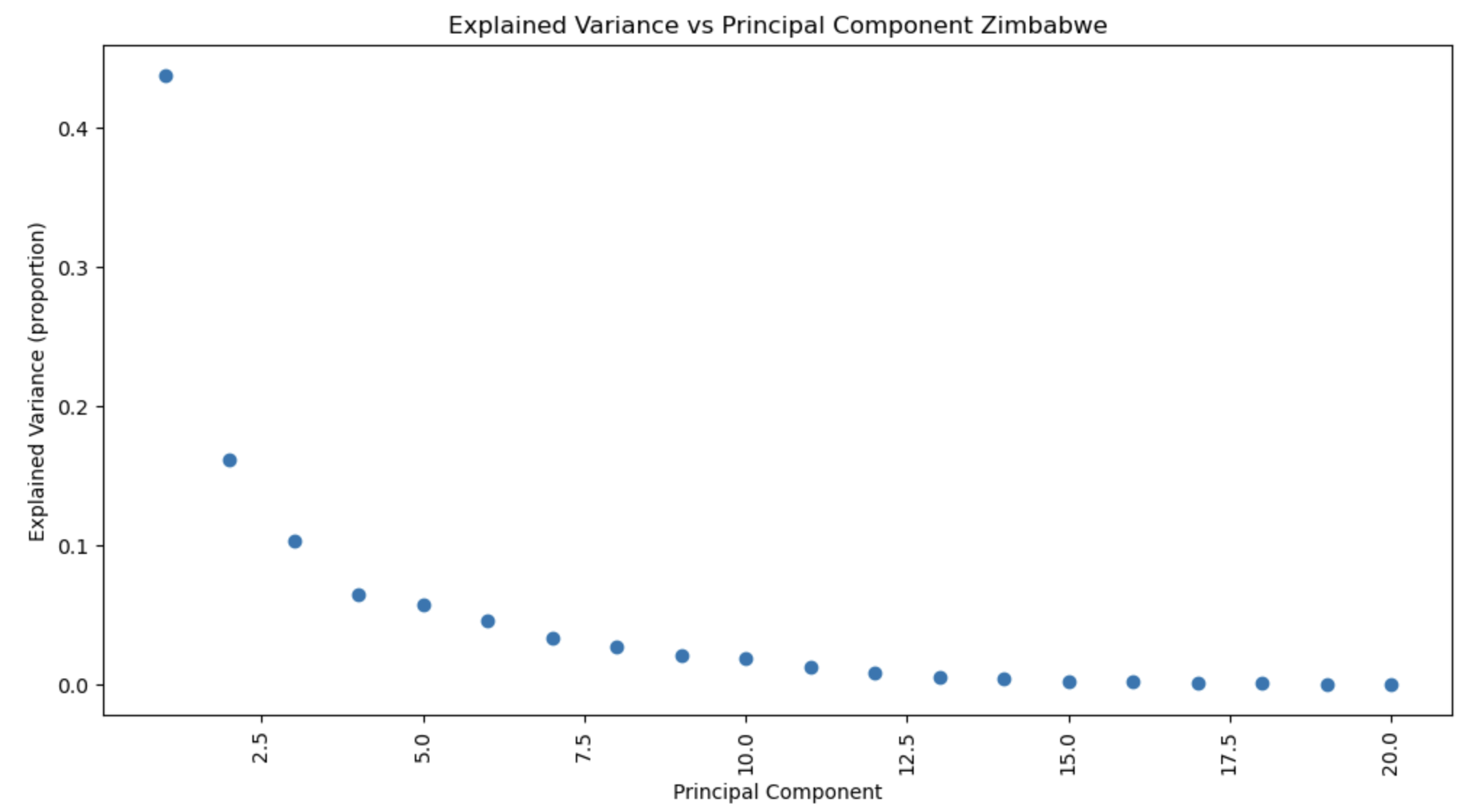
*Afghanistan (low income):*

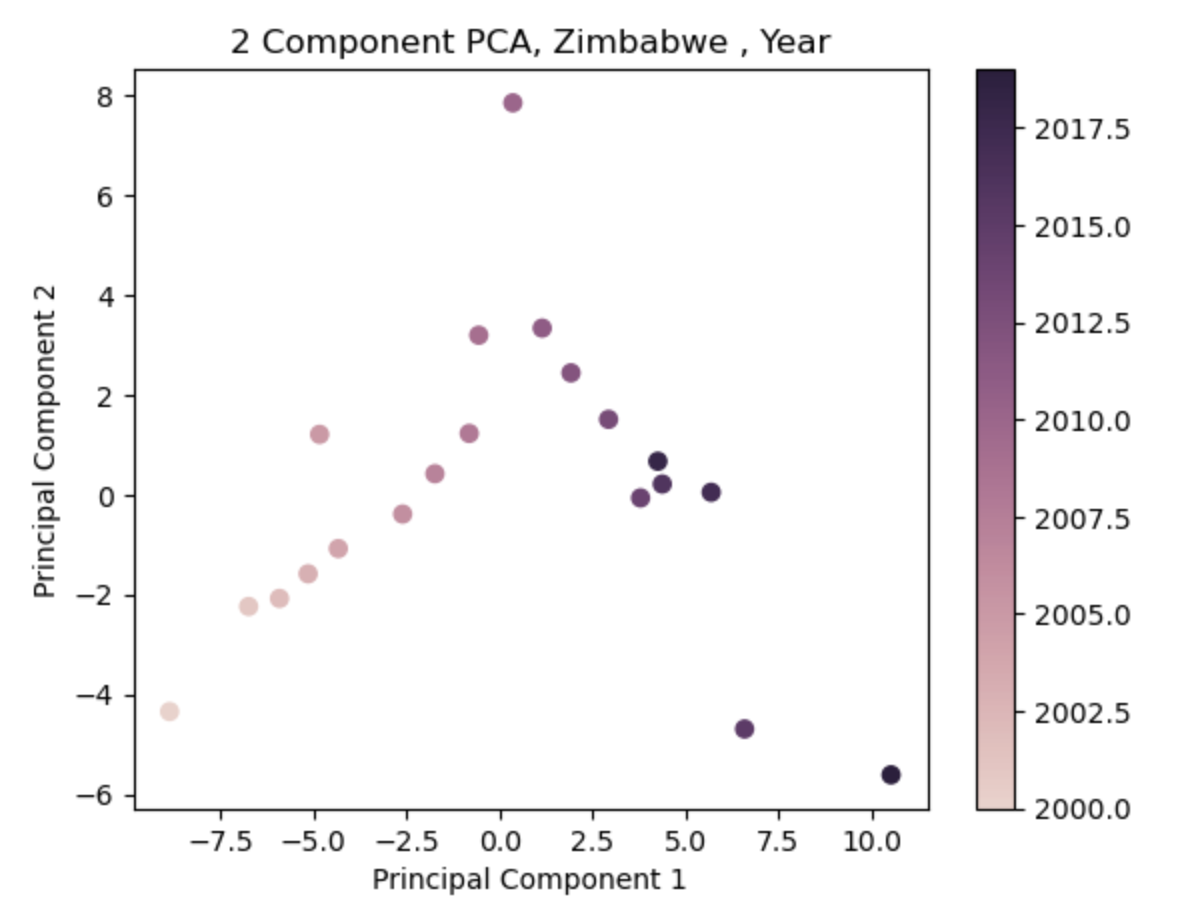
*A graph with blue dots

AI-generated content may be incorrect.*

**

*Zimbabwe (lower middle income):*

**

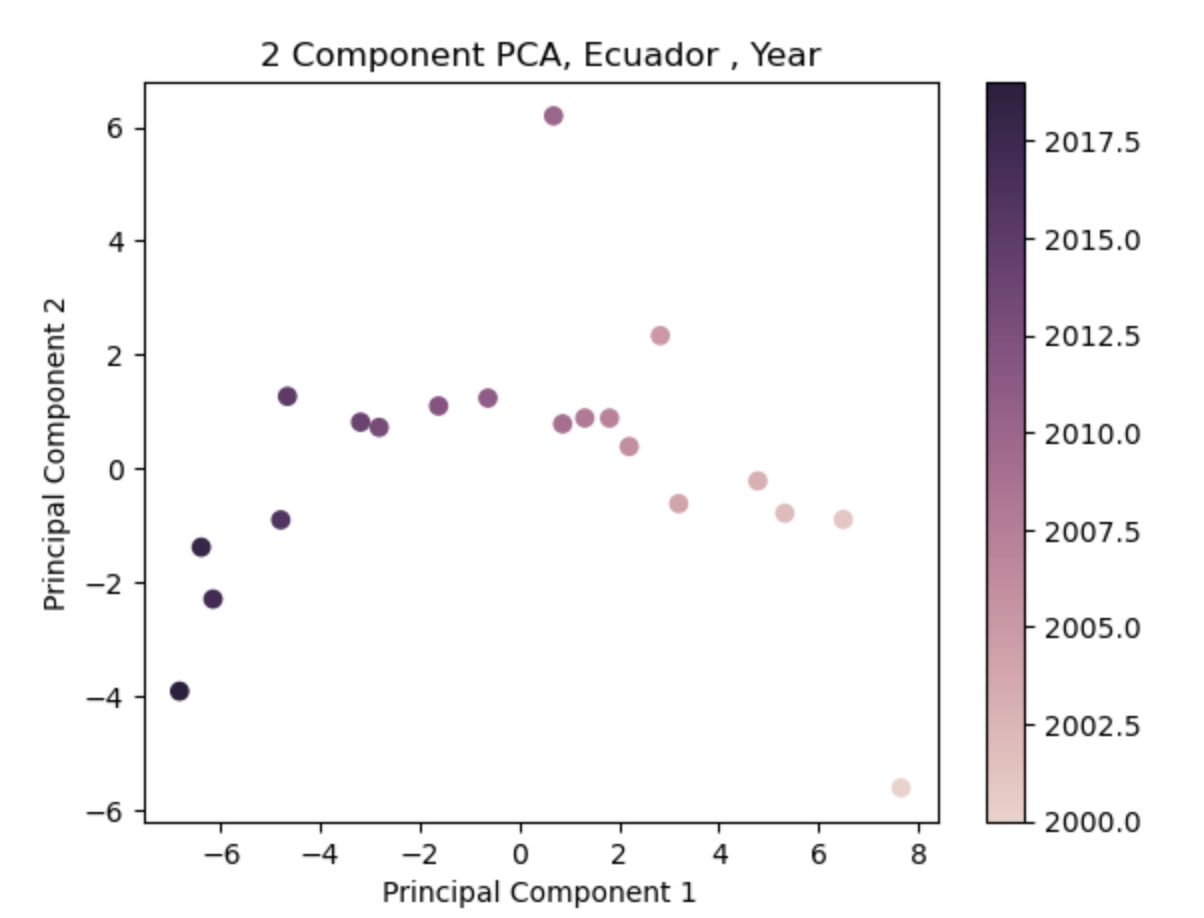
** *A graph of a graph with purple dots

AI-generated content may be incorrect.*

*Ecuador (upper middle income):*

*A graph of a function

AI-generated content may be incorrect.*

*A graph of a component

AI-generated content may be incorrect.*

*Denmark (high income):*

*A graph of a function

AI-generated content may be incorrect.*

*A graph of a diagram

AI-generated content may be incorrect.A graph of a component

AI-generated content may be incorrect.*

PCA Observations:

* The first principal component explained the highest proportion of variation in the lowest income country (>70%)’s data, and between 40 and 60% for the others.
* While there did not appear to be large clusters, the year and MMR estimate appeared to change moving along the first principal axis.
  + However, there were small clusters present, indicating feature sharing between small groups of years.

**All Countries PCA (for a specific year):**

Additional Data Cleaning:

A column from a country’s dataset was dropped if it had less than three non-NAN values. Unfortunately, different countries had different columns that needed to be dropped.

1. I calculated the average of each feature for different income level countries. I replaced missing values with the average for their income level.
2. If the entire income level was missing a value, I used the average from the lower down income level.

PCA:

*2010:*

*A graph of a graph with blue dots

AI-generated content may be incorrect.*

*A graph with purple and blue dots

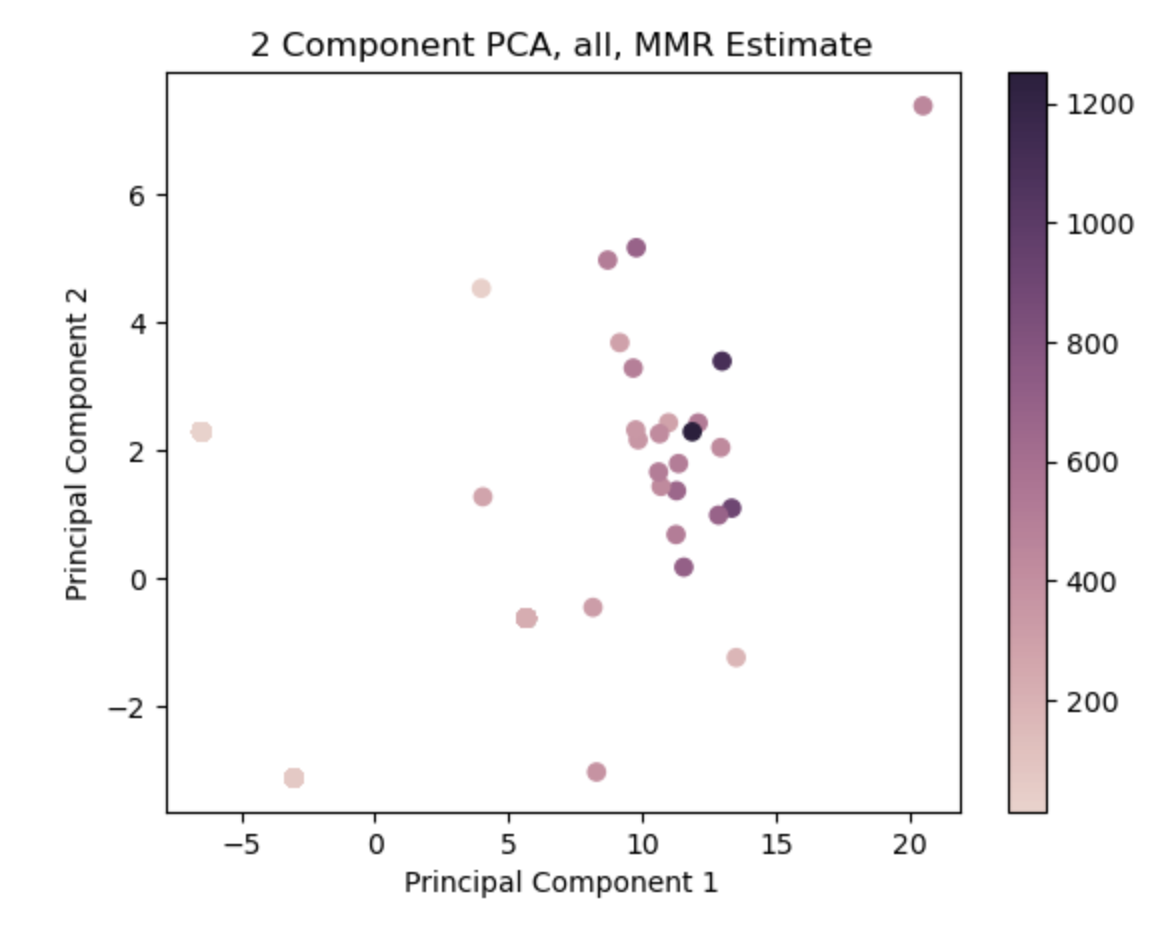
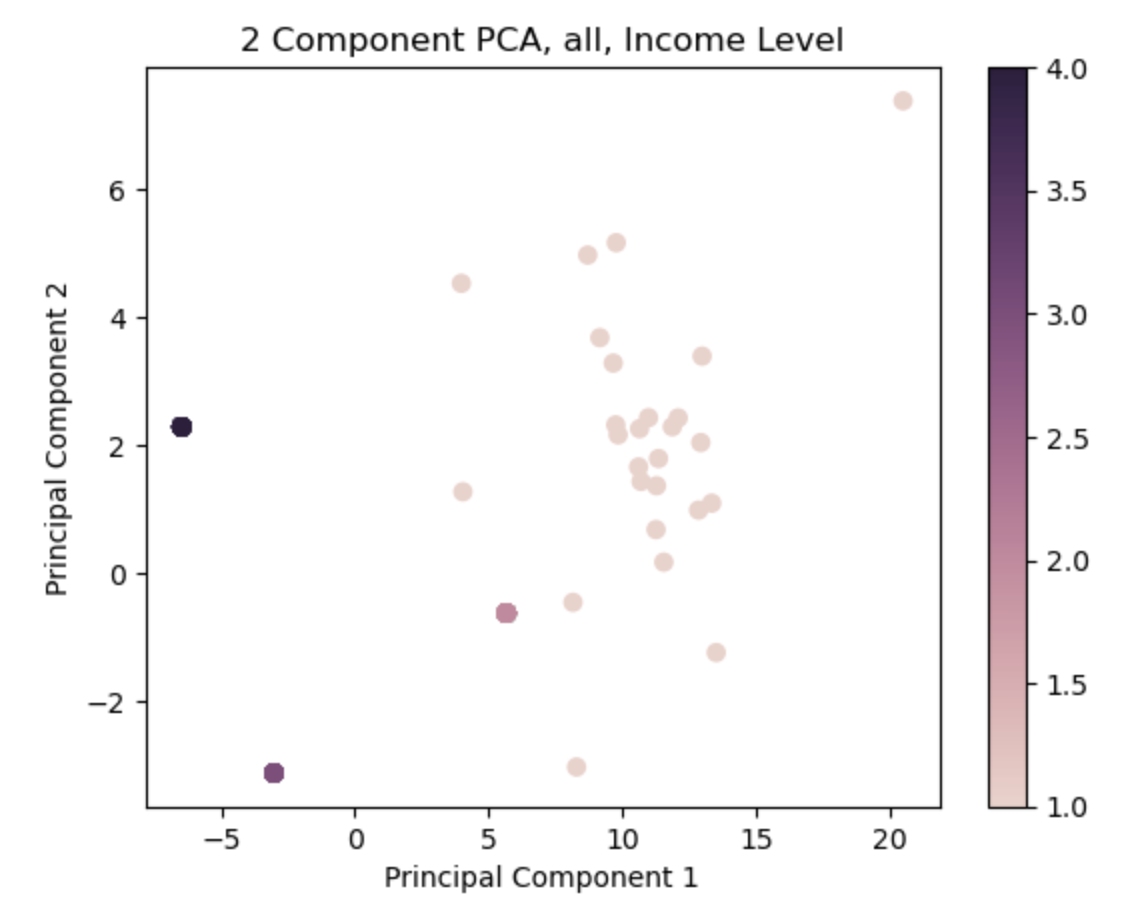
AI-generated content may be incorrect.A graph of a component

AI-generated content may be incorrect.*

*2017:*

*A graph with blue dots

AI-generated content may be incorrect.*

**

PCA Observations:

* Regardless of year, the first principal component explained approximately 70% of variation.
* There was a cluster of lower income countries and high MMR estimates, with the clusters becoming slightly closer together between 2010 and 2017.
  + Indicating potential feature similarity.

**To-Dos and Questions**

* Discussion of interpolation/extrapolation methods, especially for when I collated all the countries.
  + Would you suggest a more robust method for approximating missing values than using income level averages?
* Look at more imputation methods
  + More than linear regression
  + Visualisation of the different imputation methods
    - See distribution of the data
  + K-nearest neighbours
* Separate plots based on income level?
* Look at distribution of data for each income level
* Look at methods for seeing which distribution data best fits
  + **Kolmogorov-smirnov test** 
    - Looks at many distributions and tells you which distribution your data fits best
* Distribution of missing data
  + Missing data over the years across all countries
  + And maybe also by income level
  + Similar thing for distribution of missing data