Assessment 5

Exploratory Data Analysis and Visualization

Ziyang Lin

Introduction

This report outlines the strategy and results of exploratory data analyses and visualizations on the 2017 World Bank Development Indicators data set. Each record in this data set corresponds to a single country where the variables are mainly development indicators of that country. The data set contains 14 columns in total with a uniquely identifiable country name, 2 categorical variables for regions and income levels, and the remaining 11 variables are numerical development indicators which include but not limit to GDP per capita, unemployment rate for both sex, life expectancy for both sex, and CO_2 emission per capita.

We primarily focus on the below topics of identifying issues of data quality, such as missing data or outliers; conducting a brief exploration of the univariate and multivariate distribution of several variables in the data; and detecting any clustering behaviour with the aid of dimension reduction techniques.

Data Quality

From an initial glimpse at the data, we observe that all columns besides educational expense have non-`NA` values for all countries. In particular, there are 12 such observations, which constitutes 1.3% of the total entries. Our next step will be to characterize the types of missingness in the data.

We plot the missing education expense values against all 12 variables. In Figure 1 we see that the missingness appears mainly for countries with high or lower middle income. In terms of geographic region, there is only one country in North America in this data set, and this country has missing education expense record. Even though the number of missing observations and the total sample size is relatively small, we can still suspect that there is dependence between the missingness and these categorical variables, so that the missingness is certainlyly not MCAR.

However, examining missingness against the remaining 10 numerical variables (codes are supplied in the RMarkdown notebook), we find the distributions of missing versus non-missing values of education expenses conditioned on each variable are similar, suggesting that no missingness dependence exists for the numerical variables. Hence we can categorize the missingness as MAR.

For this type of missingness, a regression imputation should yield unbiased results but will have a standard error that is too small. Using multiple imputation can avoid this problem, and it also produces reliable results regardless of our level of confidence in a particular imputed value. We perform multiple imputation by applying a linear model of `GDP.percap` against the others (including `Education.Expend`) and pool the results. The imputed values are then used populate the `NA` entries.

Next we move on to outlier detections. We first investigate the empirical distribution of all numerical variables using histogram (codes in RMarkdown notebook). We observe that the distributions of these variables appear to be mostly different from each other. To have a unified standard for outlier detection, we choose to use the modified Z-score statistic.

In Figure 2, for each variable, we computed the modified Z-score, and the red dash lines represent the outlier threshold of $|M_i| > 3.5$. We observe that some variables such as GDP per capita and CO_2 emission per

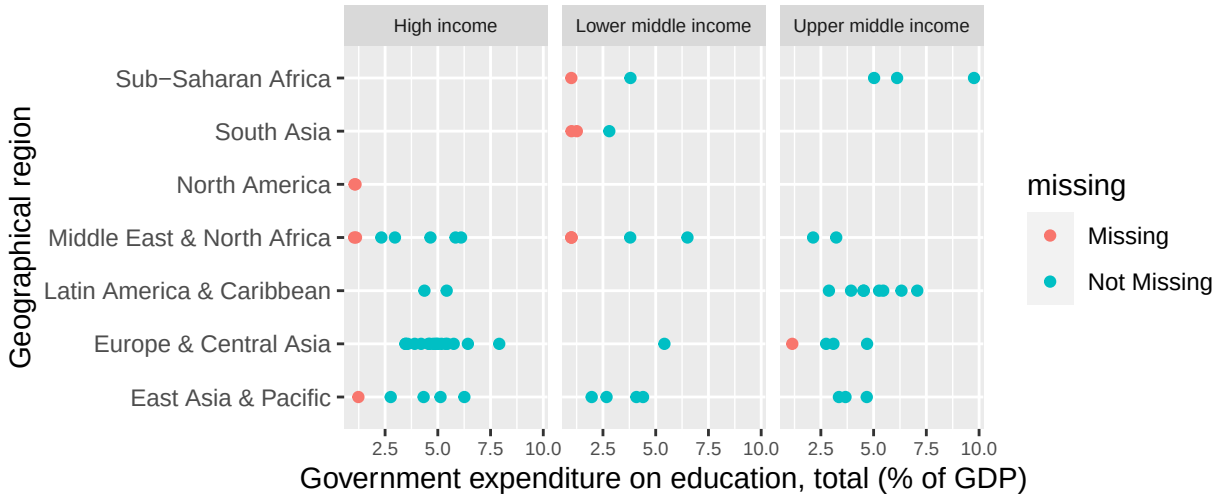

Figure 1: Missingness plot for dducaiton expense against categorical variables

capita have some outlying observations. In particular, two observations for mortality rate under-5 have very large modified Z-score values. We are unable to compute this statistic for the variable access to electricity as percentage of population since its MAD is computed at 0. We applied the normal boxplot method for this variable and discovered all values except 100 are outside of the interquartile range, thus are categorized as outliers.

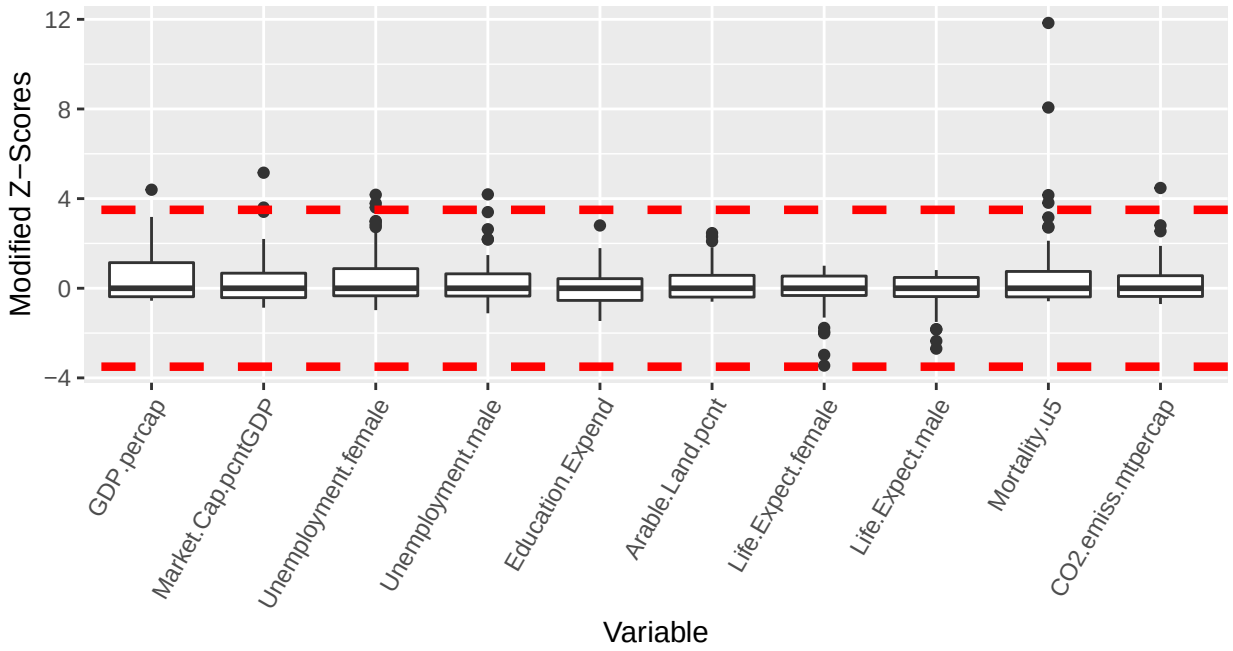

Figure 2: Side-by-side boxplots to visualize outliers for 10 numerical variables

Univariate and Multivariate Exploration

Next we move on to exploring the univariate and multivariate distribution of some variables. We want to study the distribution of GDP per capita, and its marginal and joint distribution with some other variables such as income level and region, life expectancy, and CO_2 emission as well as their correlation structure.

Figure 3 shows that GDP per capita features an asymmetric right-skewed distribution, but the Q-Q plot does not suggests over or underdispersion, so we should expect its empirical kurtosis to be less than 3.

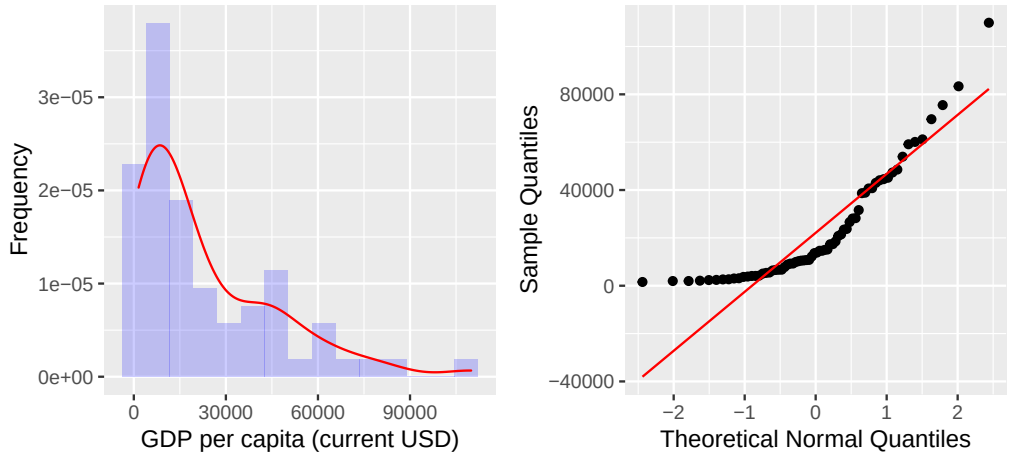

Figure 3: Empirical distribution of GDP per capita

In Figure 4 we compare the distributions of GDP per capita against income levels and geographical regions, and we discovered that these two variables both have visually significant impacts on the center and spread of GDP per capita. In particular, we found that higher income levels translate to a higher GDP per capita, and that South Asia and Sub-Saharan Africa have significantly lower GDP per capita than other regions. In fact, these two regions have no countries that are categorized as high income in this data set.

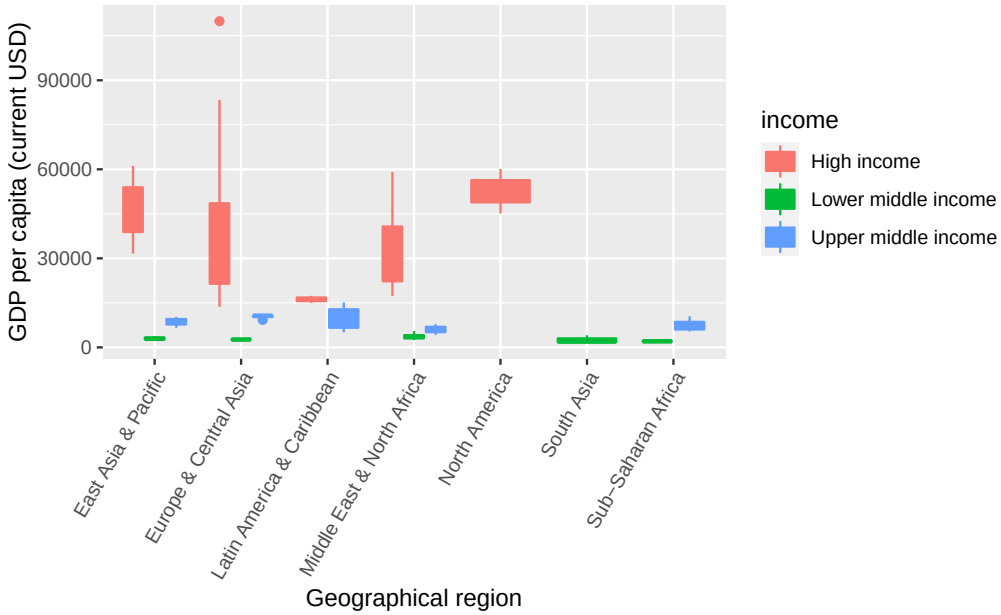

Figure 4: Distributions of GDP per capita against income levels and geographical regions

Consider Figure 5 for the joint and marginal distribution of GDP per capita against female life expectancy at birth and CO_2 emission per capita respectively. Visually we see GDP per capita varies with each of these variables, such that higher life expectancy or higher CO_2 emission corresponds to higher GDP per capita. The left panel indicates that the relationship may not be linear, and the right panel also shows that the variance outside of the linear relationship increases as GDP per capita increases.

To conclude this section, we produce a heatmap to visualize all pairwise dependence in Figure 6. We discovered

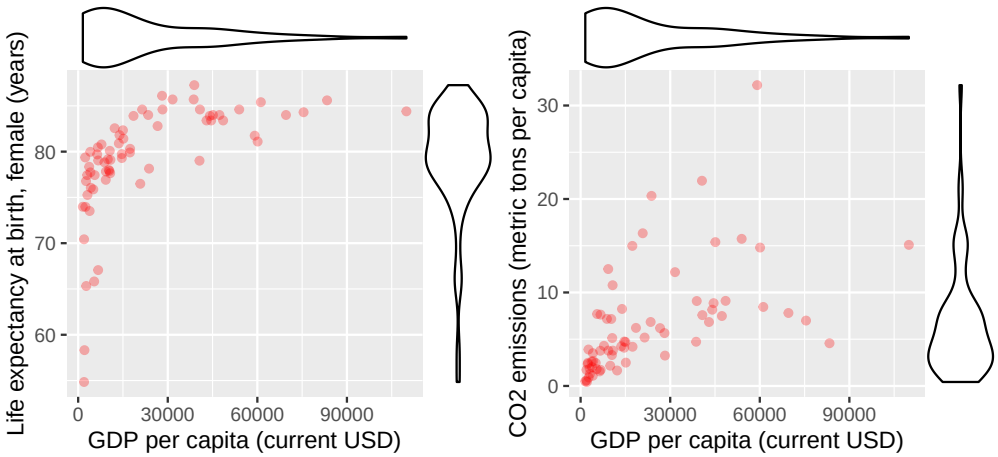

Figure 5: Joint and marginal distribution of GDP per capita against female life expectancy and CO2 emission

in the diagonal from bottom-left to top-right that GDP per capita is moderately to strongly correlated with all other variables.

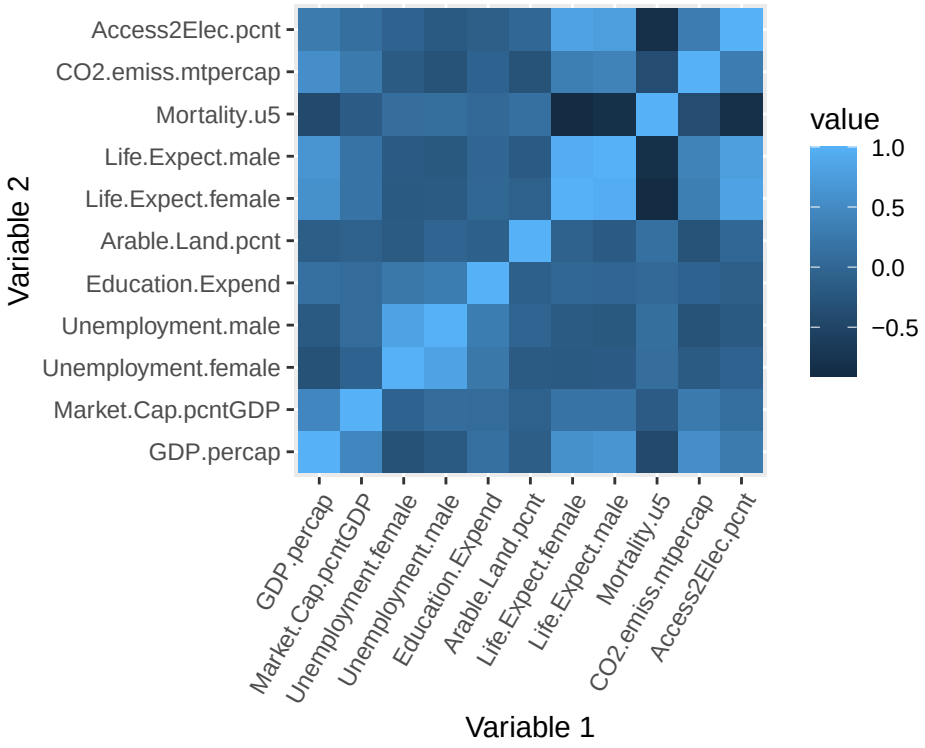

Figure 6: Pairwise dependence of all numerical variables in a heatmap

Cluster Analysis

Considering the structure of the data set, we may naturally believe that clustering behavior would exist for the other numerical variables that relate to the income or region categorical variables. We first construct a generalized pairs plot conditioned on income level, and the plot does not show any distinctive clustering behavior for any combination of the variables. The plot is too large to be contained in this report (but

provided in RMarkdown notebook), but we can generate some contour plots for some pairs of the variables as in Figure 7, and observe there is no obvious clusters.

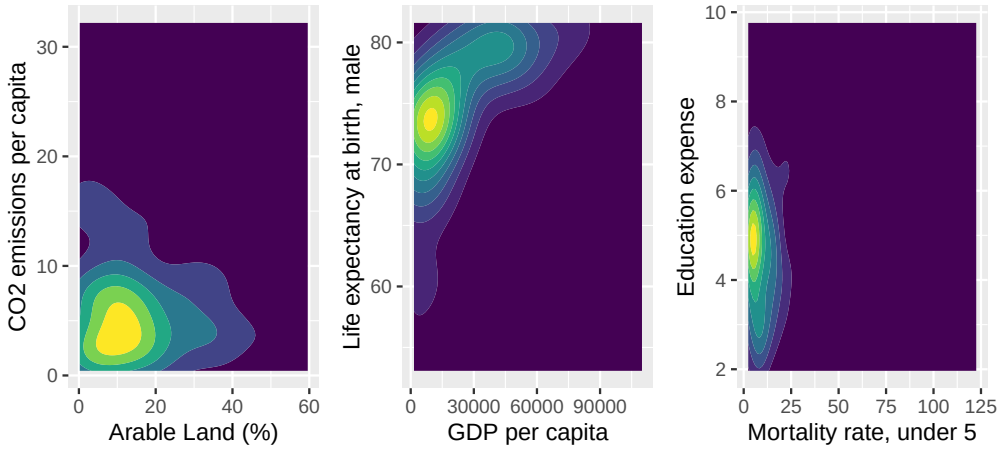

Figure 7: Pairwise contour plots for some pairs of variables

However, only from the above observations cannot let us conclude the non-existence of clusters. Clustering behaviors are sometimes difficult to visualize in high dimension, but we can project the data onto a low-dimensional embedding. We will try 3 different approaches: principal component analysis, multidimensional scaling, and t-distributed stochastic neighbour embedding. The outputting density plots are shown in Figure 8.

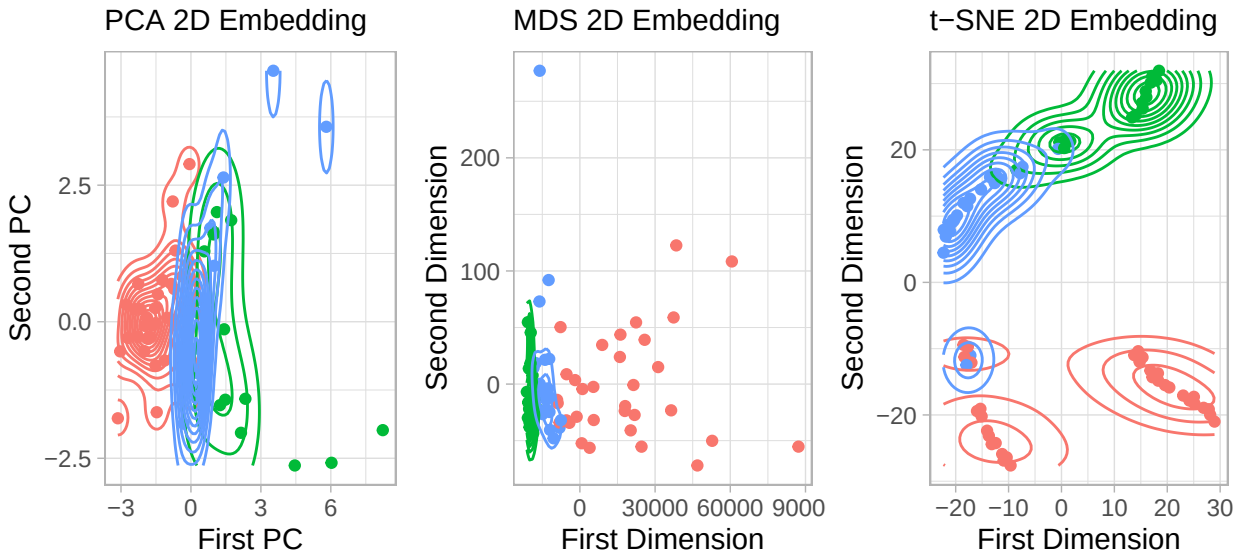

Figure 8: Cluster visualizations with the help of 3 dimension reduction approaches

We observed that the t-SNE method produces the most separable clusters, although the boundary is still not entirely clear. In fact, when contrasting with the income levels, the data appears to have more than 3 clusters, but this already allows us to conclude that there is clustering behavior in the data, and dimension reduction technique helps us extract such clusters in visualizations.

To pursue clustering analysis further, we can apply the k-Means clustering algorithm on both the full data set and the PCA-reduced data set. The codes are appended in RMarkdown notebook and we can observe that dimension reduction helps identify the clusters.