Data Cleaning

1.0 Introduction

1.1 Dataset:

The World Bank, World Development Report 2014 (World Development Report. 2016. Wdr2014-annex-tables.csv [Data set CSV File]. Retrieved from http://data.worldbank.org/data-catalog/world-development-report-2014. September 22nd, 2016.)

1.2 Description:

The dataset consists of 10 tables which includes 75 indicators relevant to the management of risk in the context of development. Each of these indicators help summarize the risk management capacity of countries based on different social, economic, financial and environmental dimensions. The dataset is a compilation of data collected by various reliable sources like the World Bank, United Nations, International Monetary Fund, research institutes, etc. It mainly contains data of 133 out of the 195 countries in the world.

1.3 Details of the license or terms of use:

This particular dataset has been made available by The World Bank and can be distributed, adapted, displayed for commercial and noncommercial purposes. Complete details of all the terms of use can be found http://data.worldbank.org/summary-terms-of-use. There are also specifications mentioned about using third-party data that may have been used to build the dataset on http://data.worldbank.org/restricted-data. However, none of the third-party sources used in this particular dataset are mentioned. Hence you can assume that none of the data included is restricted. It can be concluded that there aren't any usage constraints.

1.4 Metadata Availability:

The dataset came with a file that contained all the technical notes and metadata related to the file. Detailed descriptions of fields including the collection source, symbol notations and definitions were all mentioned. Information was well titled and could be easily correlated with the data set, wherever necessary references to other datasets were provided and the metadata was represented in logical order, making it extremely easy to understand.

2.0 Data Cleaning

2.1 Rationale:

- Overall: Countries like Argentina, Iran, Islamic Rep., Libya, Myanmar, Somalia, South Sudan, West Bank and Gaza were completely omitted because there was negligible data available for these countries. The unavailability of data is mostly because of political or economic turmoil and hence these countries are deemed to be very risky.
- Key Indicators: Within the first sheet the column of "Adult literacy rate" was deleted because the definition of literacy covered a broad spectrum of people who do not necessarily paint a complete picture of the country's human capital quality. Additionally, this information was deemed irrelevant for our analysis.
- Selected Risk: Columns like "Homicide Rate", "Poverty headcount ratio" , "Volatility of household consumption growth per capita", "Volatility of GDP growth per capita" were all deemed to be irrelevant to analysis. The column containing "Volatility of GDP growth per capita" had information for the years 1990 and 2000 and could not be combined with the GDP per capita data from sheet one.
- Human Capital: All data containing "Education quality" and "Under-5 mortality rate" were deleted because they were irrelevant and mostly inconsistent.
- Enterprise: This sheet contained information with respect to indicators related to risks involved in the enterprise sector and hence was deemed irrelevant for our analysis. Labor, goods and production are indicators that would be analyzed only if HIP dealt primarily in the commodities sector.
- Financial Sector: Few of our most weighted indicators were retrieved from this sheet. We decided to only keep indicators that depict what the saving and investment trends of the countries were. We chose to delete all factors that focused on whether savings and loans were formal or informal because they did not seem to add any valuable inferences to our analysis.
- Macroeconomics: Macroeconomics essentially deals with performance, structure and economic stability of a country and although a lot of the indicators were relevant to analyse the financial stability of the country we felt that data for "international reserves" and "worldwide governance indicators" were adequate to capture this.
- Disasters: Data related to total deaths and damages were deleted because we already have average figures which were a better indicator for our analysis.
- Other Economies: This sheet was omitted because it contained key indicators for those countries which were not included in the first sheet. There was no other data which would lead to difficulties in comparison.

- Global Temp Anomalies: This sheet contained data with respect to global temperature averages and hence was deemed irrelevant.
- Aid Commitments: This sheet was regarded as irrelevant as it contained data regarding aid commitments for Emergency Response, Reconstruction Relief, and Prevention and Preparedness.

2.2 Issues and Limitations

1. Missing some values in the table and it is unlikely to retrieve those values from somewhere else.
2. The time scales are inconsistent across different sectors in the table. For example, the selected risk indicator sheet contains data which has three distinctive time scales. They are 1993-2010, 1993-2012 and 2007-2011. In the access to social insurance sector, countries were also surveyed in different years.
3. As an insurance company, which is looking for foreign investment opportunities, the company still has issues with how to rank each indicator (from the most important indicator to the least important) after the data is cleaned and how to weight each indicator exactly (the percentage each indicator is granted).
4. There is no significant data measures the private insurance industry in each country.
5. Difficulties in running analysis because of the column titles

2.3 Steps:

1. Open the file “wdr-2014-annex-tables” in MS Excel, after extracting the xlsx
2. File from the downloaded zip folder. Make a draft copy for yourself by selecting the Save As option from the File menu.
3. Delete sheets titled “4-Enterprise”, “9-GlobalTempAnomalies ” and “10-AidCommitments”.
4. From the sheet titled “1-KeyIndicators” delete the column “Adult literacy rate” by right clicking on column X and then clicking on Delete.
5. From the sheet titled “2-SelectedRisk”, first select multiple columns and then delete them by right clicking the column and then clicking on Delete. To select multiple columns, hold the Ctrl key or the command key on a mac and then click on the columns P,Q,R,S,T,U,V,W,X,Y,Z,AA,AB,AC. Alternatively, since these are consecutive columns you can select them by clicking on column P and then holding down the shift key and the right arrow key till you reach column AC.
6. Now that we have familiarized ourselves with selecting and deleting columns in Excel, Delete the following columns from the remaining sheets:

Sheet Name	Columns to be Deleted
3-Human Capital	H,I,J,K
5-FinancialSector	B,C,D,J,K,L,M,N,O,P,Q,R,S,T,X,AL,AM,AN,AO
6-Macroeconomic	B,C,D,E,F,I,J,K
7-Disasters	B,C,D,J,K,L,R

7. Make sure that all the column have headings. If any of your column headings are missing you can always re-type them by referring to the original file.
8. Delete rows that contain the data for the following countries Argentina, Iran, Islamic Rep., Libya, Myanmar, Somalia, South Sudan, West Bank and Gaza. To do this you must first filter out column A. Click on row 2 and then under the data menu click on filter. From the drop down menu that appears in cell A2 select only the above mentioned countries and then hit on "Apply Filters" (If you are using Excel 360 or a version above, you may have to select only one country a time and click on the exit button to apply the filter.) After the sheet has been filtered, just as you would for a column, select all the rows and then delete it.
9. Unfilter your sheet by once again clicking on the filter button.
10. Combine all the sheet in the workbook into one masterfile using VLookUp. Set the Country name as the primary key and the pull data from all sheets into the columns.
11. Rename the column Titles into codes for ease in running analysis. Refer Technical Notes to understand what each column represents.
12. Rename the Sheet as "WDR" and delete all other sheets.
13. You should end up with 38 columns and 129 rows.
14. Save all your changes from File >> Save or Ctrl (command on a mac) + s.

3.0 Analytics

The analytics for initial plotting of variables was carried out using R. This document can be found in the git repository on this link and is named INFM600_0201_Hip_DraftRScript_Final. We followed this up with draft plots in R to find out top countries based on each indicator. The document for the same can be found on this link and is named INFM600_0201_TeamHIP_DraftRPlot.nb. Additionally, we have added the same versions of the document to the current folder in the repository.

References:

Macroeconomics retrieved from https://en.wikipedia.org/wiki/Macroeconomics on November 1st, 2016.