Data Exploration Insights

After exploring the data using pandas I was able to come up with a number of insights that have helped me better understand the World Indicators dataset and how to work with it.

1. The countries table is a 31 column x 247 row table. This table contains all info about country names and other country specific data. There are 247 entries in the country name columns, however some of these entries are categorical groupings of certain countries(eg. Least developed countries: UN classification). These separators might need to be cleaned or removed for the most appropriate reporting as there are 247 entries in this column but only 191 technical countries in the world. Its numerical columns are LatestAgriculturalCensus, LatestIndustrialData, LatestTradeData, and LatestWaterWithdrawlData. The numericals provided in these columns are all dates staging the last time they received an update on these topics for each country. The unique identifier for each row is the CountryCode column.
2. The indicators table is 6 columns x 5656458 rows. This table consists of all the different indicators for each country and their value. There are 1345 indicators for each country but the indicators having the highest number of values belong to population data followed by import and export data. The years column shows a range of values from 1960-2015 with the majority of data coming from the year 2010. The values column gives a value with different parameters because each indicator follows different requirements (some might be a percentage of the whole, while others could be a max or min). The unique identifier column in this table is the Indicator code. While the foreighn key is the countryCode.
3. The series table is 20 columns x 1345 rows and reports more in depth data on each indicator. It has all 1345 indicator names and in depth description of what each indicator means and what topic it belongs to. There are 91 topics but a lot of the topic names overlap. The periodicity column shows that all values are measured annually. However I noticed that the Unit of measure column only yields 7 results. Therefore it might need to be removed. The AggrationMethod column shows how each indicator value was collected(6 possible ways). The series long definitions are very in depth and can provide all the insights needed for each individual indicator. The unique identifier is the series code and the foreign key is the indicator name.
4. The country notes table provides a source for country data country data collected. It’s uniquely identified by the countrycode and the foreign key is the series code.
5. The series notes table provides a source for indicator data country data collected. Foreign key is the series code.
6. The footnotes table provides a source for each indicator for each country for each year. Some sources are not stated

Storage

I plan on keeping the data relatively the same, The country table, Indicators table, and Series table are all very important and should remain the same in order to achieve the optimal query speed. Some slight changes I will make will be making sure all 247 rows of the country table are valid countries and have accompanying data. Secondly for the series table, I can consolidate the topic for each indicator so as to reduce the 91 closely related topic categories, down to a couple sectors. I will remove any columns that provide 90% null values like the unit of measure column. Lastly I will see if the footnotes table provides any relevant information that isn't already stated in the other nots tables. If it serves no purpose, it will be removed.