# Notes

# Data Patterns:

**UN Data(without “UN” in the filename):**

Stuff to remove :

* 1st row title - MANUAL
* The rows with missing HDI ranks (last 18 rows)
* 1st column : HDI Rank

Column names :

* Country(names)
* Years (Consider year 2018)

Merge on: “Country”

Preprocessing:

* Country names to change:

|  |  |
| --- | --- |
| Bolivia (Plurinational State of) | Bolivia |
| Congo (Democratic Republic of the) | Congo, Dem. Rep. |
| Czechia | Czech Republic |
| Eswatini (Kingdom of) | Eswatini |
| Micronesia (Federated States of) | Micronesia, Fed. Sts. |
| Tanzania (United Republic of) | Tanzania |
| CÃ´te d'Ivoire | Cote d'Ivoire |
| Congo | Congo, Rep. |
| Hong Kong, China (SAR) | Hong Kong SAR, China |
| Iran (Islamic Republic of) | Iran, Islamic Rep. |
| Saint Kitts and Nevis | St. Kitts and Nevis |
| Korea (Republic of) | Korea, Rep. |
| Saint Lucia | St. Lucia |
| Moldova (Republic of) | Moldova |
| Saint Vincent and the Grenadines | St. Vincent and the Grenadines |
| Venezuela (Bolivarian Republic of) | Venezuela |
| Viet Nam | Vietnam |

**WB Data:**

Stuff to remove:

* Remove first four rows - MANUAL

Column Names :

* Country Name,
* Country Code
* Indicator Name
* Indicator code
* years(1960 - 2020)

Merge on : “Country Name”

Preprocessing:

* Country names to change:

|  |  |
| --- | --- |
| Bahamas, The | Bahamas |
| Egypt, Arab Rep. | Egypt |
| Gambia, The | Gambia |
| Kyrgyz Republic | Kyrgyzstan |
| Lao PDR | Lao People's Democratic Republic |
| Slovak Republic | Slovakia |
| Venezuela, RB | Venezuela |
| Yemen, Rep. | Yemen |

# Open Questions:

1. What is the base for getting all country names? (IDs)
   1. This will help avoid spelling mistakes
   2. World Bank seems to have more countries than UN, so that can be used
   3. Having a list of countries we need to work on would surely help
   4. \*\*\*\*\*\*Two different country lists:
      1. <https://history.state.gov/countries/all>
      2. <https://data.worldbank.org/country>
2. Do we have any use for the continent groups like (South Asia, East Asia, Sub-Saharan Africa, etc.)? Is the world data (row) helping us in any way?
   1. No ideas on this as of yet but it could be interesting to see how the rankings break down by region
3. Missing Countries:
   1. We have a total of 190 countries common in both data sources
   2. World Bank
      1. Palestine (would West Bank and Gaza be the same?)
   3. UN Data
      1. South Korea (Korea, Dem. People’s Rep.)
      2. Somalia
      3. Greenland
   4. Analysis of missing countries available here: <https://docs.google.com/spreadsheets/d/1gZ5f-S8L9BpqjwyxK5JpHHnE8IVuo-6M/edit#gid=59518916>
      1. Countries with #N/A are not found
      2. Each should contain a comment next to it to help us understand if we plan to use these countries

Next Steps:

1. Preprocess on the data source
2. Look for Central Tendency Measures
   1. Work with **median** right now
   2. Modularize the code for ebay edit later
   3. If it changes by variable, add this to the control file
   4. Check if using county groups would be a good idea
3. Come up ranking
   1. Look into few variables
      1. Finalize their data source
   2. Formulate a scoring functions
      1. How to create meaning index
      2. Happiness Report maybe a good starting point
      3. Look into More papers if needed
4. Go through the progress report
5. Come up with indices
   1. Create the indices, same as WHR (using our own scoring functions)
   2. Get all categories
   3. Get 1-2 category scoring functions finalized (modular code)
6. We can try to show how indices
7. Doing a correlation analysis on this would be helpful (?)

------------------------------------------------------------

Next Steps for the report:

* Datasets
  + Describe more about the datasets being used
  + Why did we finalize on these
  + Maybe some info on the Europe data (?)
    - We need to check if we have all metrics from this one as well
* Data Cleaning
  + Describe the current methods
  + Country selection
  + Cleaning by Data Source
* Ranking Function
  + This may change as per data type
  + Initial index ready for a category
  + Develop and explain how this can be changed easily by some robust python implementation
* Visualizations
  + Compare one of the calculated index with the WHR and show on a world map
  + Derive observations and next steps on this

-----------------------------------------------------------------

Notes from WHR20:

* They have given more weight to interpersonal and institutional confidence over some negative components: Can we use any of these to show that they actually affect happiness in a country?
  + Ill-health
  + **Unemployment**
  + **Low income -- Poverty % (check correlation with unemployment)**
  + *Discrimination (how to operationalize this)*
  + Family Breakdown
  + *Safety in Streets (how to operationalize this)*
* Variables used by WHR20:
  + GDP per Capita
  + Generosity
  + Social Support
  + Freedom to make decisions
  + Life expectancy
  + Freedom from corruption
* Life Evaluation as a metric:
  + How they calculate this:
  + Can we make a scoring function for this as well?
  + Can be used to check correlation with various variables, which can help us select our final variables
  + Previous years may need to be considered
* Data used:
  + They use data from 2017-2019, as not all countries take the survey each year
  + We can also take the latest available data in a series of months
  + We can make a linear regression for each country to predict their value for a given year and use that instead (Interpolating missing data)
* Cantril Ladder is used for these variables: