Data Science Career Track Capstone 01

Data Wrangling Checkpoint

Cleaning Steps

Cleaning for the provided dataset was accomplished in three steps:

1. Removing rows (countries) with missing values, since the missing data could not be found elsewhere
2. Removing extraneous symbols (e.g., `$` and `,`)
3. Altering numeric `str` values to the `float` datatype

This process was straightforward due to the fact that the dataset was taken from Kaggle, which tends to have datasets that are partially cleaned.

Missing Values

For this dataset, missing values were removed because a key metric that would be used in the future model was missing and those values could not be replaced with external datasets. `NaN` values in the following columns were removed:

- **GDP per Capita** – Every nation does have a GDP in the real world, but since the dataset does not provide any other measure of financial prosperousness, predictions cannot be made based on finances for a country with a missing `GDP per Capita` value. This decision removed 15 countries (8% of total data) from the dataset.

- **Footprints and Resources** – There were several countries missing all footprint data and resource data except for the totals (`Total Biocapacity`, `Biocapacity Deficit or Reserve`). Since the model will use components of the footprint (e.g., `Carbon Footprint`) and the biocapacity (e.g., `Forest Land`), rows which contained `NaN` values were omitted. Every column that was missing any one of the component footprints or resources was missing all of them. `Cropland` was arbitrarily selected as the column with `NaN` to delete each row. This decision removed 10 countries (another 5% of total data) from the dataset, This included the following variables:

    - `Cropland Footprint`
    - `Grazing Footprint`
    - `Forest Footprint`
    - `Carbon Footprint`
    - `Fish Footprint`

    - `Cropland`
    - `Grazing Land`
    - `Forest Land`
    - `Fishing Water`
    - `Urban Land`

A total of 25 countries were removed from the original dataset, which is a total of 13%.

Outliers

Components of biocapacity (e.g., `Cropland`) and footprints (e.g., `Carbon Footprint`) were not evaluated for outliers. This is because different regions of the world are expected to have various strengths and weaknesses in ecological productivity. In other words, variance in these metrics is expected and is encouraged for the integrity of the model. Thus, the focus for outlier analysis was placed on aggregate metrics such as `GDP per Capita`, `Population`, `Total Biocapacity`, and `Biocapacity Deficit or Reserve`.

- `GDP per Capita`
    - **Bulk of data:** 0 to 20,000 (128 countries, 78.5%)
    - **Mid range**: 20,000 to 70,000 (32 countries, 19.6%)
    - **Major outliers:** >80,000 (3 countries, 1.8%)

- `Population`
    - **Bulk of data:** 0 to 400 (161 countries, 98.8%)
    - **No data:** 400 to 1,200 (0 countries, 0%)
    - **Major outliers:** >1,200 (2 countries, 1.2%)

- `Total Biocapacity`
    - **Bulk of data:** 0 to 30 (161 countries, 98.8%)
    - **Major outliers:** >60 (2 countries, 1.2%)

- `Biocapacity Deficit or Reserve`
    - **Countries with deficit (bulk of data):** <0 (116 countries, 71.2%)
    - **Countries with small reserve (mid-range):** 0 to 30 (45 countries, 27.6%)
    - **Major outliers:** >60 (2 countries, 1.2%)

Histograms for visualization of these distributions can be found in the Data Cleaning notebook.

Unknown Data

The `Data Quality` column is an important factor in weighing the data from each country used in the model so that more reliable data has a higher influence in the model and vice versa. This would ensure better integrity in the results of the final model. However, at this point in time, the values for `Data Quality` are unclear based on the documentation provided by the data source and outside research. There is an ongoing discussion on the Kaggle forums to try and resolve this. However, if this cannot be resolved before creating the final model, it will be dropped.