Yishay Polatov: 316302140

# Part 1 preview

* The initial 10 rows of the demographics dataset both prior to and following sorting by Country.
* Before sorting

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Country | Life Expectancy Both | Life Expectancy Female | Life Expectancy Male | Urban Population Percentage | Urban Population Absolute | Population Density |
| Afghanistan | 66.5 | 68.1 | 68.1 | 26.7 | 11704638 | 67 |
| Albania | 80 | 81.7 | 81.7 | 70.3 | 1948831 | 101 |
| Algeria | 76.7 | 78.1 | 78.1 | 74.4 | 35291973 | 20 |
| Angola | 65 | 67.5 | 67.5 | 68.8 | 26847887 | 31 |
| Antigua and Barbuda | 77.9 | 80.6 | 80.6 | 28.5 | 26823 | 214 |
| Argentina | 77.7 | 80.2 | 80.2 | 96 | 44010474 | 17 |
| Armenia | 76 | 79.7 | 79.7 | 63.8 | 1882155 | 104 |
| Aruba | 76.6 | 79.1 | 79.1 | 44.7 | 48340 | 601 |
| Australia | 84.2 | 86 | 86 | 86.5 | 23335357 | 4 |
| Austria | 82.3 | 84.6 | 84.6 | 58.6 | 5337973 | 111 |

* After sorting

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Country | Life Expectancy Both | Life Expectancy Female | Life Expectancy Male | Urban Population Percentage | Urban Population Absolute | Population Density |
| Afghanistan | 66.5 | 68.1 | 68.1 | 26.7 | 11704638 | 67 |
| Albania | 80 | 81.7 | 81.7 | 70.3 | 1948831 | 101 |
| Algeria | 76.7 | 78.1 | 78.1 | 74.4 | 35291973 | 20 |
| Angola | 65 | 67.5 | 67.5 | 68.8 | 26847887 | 31 |
| Antigua and Barbuda | 77.9 | 80.6 | 80.6 | 28.5 | 26823 | 214 |
| Argentina | 77.7 | 80.2 | 80.2 | 96 | 44010474 | 17 |
| Armenia | 76 | 79.7 | 79.7 | 63.8 | 1882155 | 104 |
| Aruba | 76.6 | 79.1 | 79.1 | 44.7 | 48340 | 601 |
| Australia | 84.2 | 86 | 86 | 86.5 | 23335357 | 4 |
| Austria | 82.3 | 84.6 | 84.6 | 58.6 | 5337973 | 111 |

* The first 5 rows of the GDP and Population datasets before and after sorting
  + GDP before sorting

|  |  |
| --- | --- |
| **Country** | **GDP\_per\_capita\_PPP** |
| **Afghanistan** | 2144.167 |
| **Albania** | 16353.81 |
| **Algeria** | 14496.87 |
| **Andorra** | 59332.2 |
| **Angola** | 7408.127 |

* + GDP after sorting

|  |  |
| --- | --- |
| **Country** | **GDP\_per\_capita\_PPP** |
| **Afghanistan** | 2144.167 |
| **Albania** | 16353.81 |
| **Algeria** | 14496.87 |
| **Andorra** | 59332.2 |
| **Angola** | 7408.127 |

* + Population before sorting

|  |  |
| --- | --- |
| **Country** | **Population** |
| **Afghanistan** | 40000360 |
| **Africa** | 1.41E+09 |
| **Africa (UN)** | 1.41E+09 |
| **Albania** | 2849591 |
| **Algeria** | 44761051 |

* + Population after sorting

|  |  |
| --- | --- |
| **Country** | **Population** |
| **Afghanistan** | 40000360 |
| **Africa** | 1.41E+09 |
| **Africa (UN)** | 1.41E+09 |
| **Albania** | 2849591 |
| **Algeria** | 44761051 |

* Describe GDP

|  |  |
| --- | --- |
|  | **GDP\_per\_capita\_PPP** |
| **count** | 213 |
| **mean** | 25822.60454084507 |
| **std** | 25794.941595488068 |
| **min** | 836.6656 |
| **25%** | 6214.017 |
| **50%** | 16353.81 |
| **75%** | 38862.09 |
| **max** | 137947.3 |

* Describe Population

|  |  |
| --- | --- |
|  | **Population** |
| **count** | 260 |
| **mean** | 168775241.3653846 |
| **std** | 7.25E+08 |
| **min** | 515 |
| **25%** | 522403.8 |
| **50%** | 6827910 |
| **75%** | 34253335 |
| **max** | 7.95E+09 |

* A printed confirmation of shapes and column names for all three datasets.
  + **demographics columns**:

['LifeExpectancy\_Both', 'LifeExpectancy\_Female', 'LifeExpectancy\_Male', 'UrbanPopulation\_Percentage', 'UrbanPopulation\_Absolute', 'PopulationDensity', 'Country']

* + **demographics shape**: (200, 7)
  + **GDP columns**: ['Country', 'GDP\_per\_capita\_PPP']
  + **GDP shape**: (213, 2)
  + **Population columns**: ['Country', 'Population']
  + **Population shape**: (260, 2)

## Demographics Data Analysis

* Stats: for each numeric column in the raw data: mean, standard deviation, minimum, maximum, median, and count of missing values.

Summary statistics per field:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **mean** | **std** | **min** | **max** | **median** | **missing** |
| **Life Expectancy Both** | 73.9775 | 7.00458885841 | 54.8 | 85.8 | 74.85 | 0 |
| **Life Expectancy Female** | 76.596499999 | 7.12605367330 | 55.1 | 88.4 | 78.05 | 0 |
| **Life Expectancy Male** | 76.596499999 | 7.12605367330 | 55.1 | 88.4 | 78.05 | 0 |
| **Urban Population Percentage** | 57.276 | 25.1460342637 | 0 | 98.8 | 60.55 | 0 |
| **Urban Population Absolute** | 23592450 | 83188523 | 0 | 9.57E+08 | 4745905 | 0 |
| **Population Density** | 148.82901554404145 | 168.295891424 | 2 | 929 | 87 | 7 |

* Pearson correlation coefficient between Life Expectancy Both and Population Density.

# Cleaning Summary

* Demographics dataset
  + Non-numeric values in Population Density: 7 – replaced with Nan
  + Country name mismatches corrected: 6
  + Dropped 0 rows with invalid Life Expectancy (kept 200 of 200)
* GDP dataset
  + Country name mismatches corrected in GDP: 23
  + Missing GDP per capita entries: 0
  + GDP outliers detected (Tukey): 6 (not dropped)
  + Dropped 0 duplicate rows based on country (final count: 213)
* Population dataset
  + Country name mismatches corrected in population: 31
  + Missing population values: 0
  + Population outliers detected: 1 (not dropped)
  + Dropped 0 duplicate countries (final count: 260)

# After Feature engineering

* Table after feature engineering **columns**: ['LifeExpectancy Both', 'LifeExpectancy Female', 'LifeExpectancy Male', 'UrbanPopulation Percentage', 'UrbanPopulation Absolute', 'PopulationDensity', 'GDP\_per\_capita\_PPP', 'TotalGDP', 'LogGDPperCapita', 'Population', 'LogPopulation']
* Table after feature engineering **shape**: (173, 11)