Part 1 Preview

3.1 Web-crawling

Below are the first 10 rows of `df_demographics` after crawling the website and casting fields to appropriate types. The crawler extracted the countries in alphabetic order, so sorting by `Country` did not affect the output.

Country	Life Exp. (Both)	Life Exp. (Female)	Life Exp. (Male)	Urban Pop.	Urban Pop. %	Pop. Density
Afghanistan	66.5	68.1	64.9	11704638	26.7	67
Albania	80.0	81.7	78.1	1948831	70.3	101
Algeria	76.7	78.1	75.3	35291973	74.4	20
Angola	65.0	67.5	62.4	26847887	68.8	31
Antigua and Barbuda	77.9	80.6	74.9	26823	28.5	214
Argentina	77.7	80.2	75.1	44010474	96.0	17
Armenia	76.0	79.7	71.8	1882155	63.8	104
Aruba	76.6	79.1	74.0	48340	44.7	601
Australia	84.2	86.0	82.4	23335357	86.5	4
Austria	82.3	84.6	80.0	5337973	58.6	111

Shape & Column Names

```
df_demographics.shape:  (200, 6)

df_demographics.columns:  ['LifeExpectancy_Both', 'LifeExpectancy_Female', 'LifeExpectancy_Male',

'UrbanPopulation_Absolute', 'UrbanPopulation_Percentage', 'PopulationDensity']
```

Demographics Data Analysis

Below are the results of runnign `describe()` on each of the numeric fields in the crawled demographics dataset:

Stat.	Life Exp. (Both)	Life Exp. (Female)	Life Exp. (Male)	Urban Pop.	Urban Pop. %	Pop. Density
mean	73.98	76.60	71.38	23592449.77	57.28	156.02
std	7.00	7.13	6.98	83188523.42	25.15	176.96
min	54.80	55.10	53.50	0.00	0.00	2.00
25%	68.90	71.47	66.57	766036.75	40.22	35.50
50%	74.85	78.05	71.20	4745905.00	60.55	88.00
75%	79.12	82.03	76.30	16095833.75	77.43	222.50
max	85.80	88.40	83.10	956553854.00	98.80	929.00
missing (or zero)	0.00	0.00	0.00	10.00	10.00	0.00

The pearson correlation between `LifeExpectancy_Both` and `PopulationDensity` is 0.18, which is not considered to be a high correlation.

3.2 Loading the 2021 CSVs

GDP per Capita

First 5 rows before sorting *(see `output/gdp_before_sort.csv`)*

Country	GDP_per_capita_PPP
Afghanistan	2144.1665
Albania	16353.8090
Algeria	14496.8650
Andorra	59332.2030
Angola	7408.1265

First 5 rows after sorting *(see `output/gdp_after_sort.csv` — note data was already in A→Z order)*

Country	GDP_per_capita_PPP
Afghanistan	2144.1665
Albania	16353.8090
Algeria	14496.8650
Andorra	59332.2030
Angola	7408.1265

Descriptive statistics *(see output/gdp_describe.csv)*

Stat.	GDP_per_capita_PPP
count	213.000000
mean	25822.604541
std	25794.941595
min	836.665600
25%	6214.017000
50%	16353.809000
75%	38862.090000
max	137947.340000

Shape & Column Names

```
df_gdp.shape:  (213, 2)

df_gdp.columns:  ['Country', 'GDP_per_capita_PPP']
```

Population

First 5 rows before sorting *(see output/pop_before_sort.csv)*

Country	Population
Afghanistan	40000360
Africa	1413750475
Africa (UN)	1413753005
Albania	2849591
Algeria	44761051

First 5 rows after sorting *(see `output/pop_after_sort.csv` — sorted A→Z)*

Country	Population
Afghanistan	40000360
Africa	1413750475
Africa (UN)	1413753005
Albania	2849591
Algeria	44761051

Descriptive statistics *(see `output/pop_describe.csv` — note scientific notation for large values)*

Stat.	Population
count	2.600000e+02
mean	1.687752e+08
std	7.254974e+08
min	5.150000e+02
25%	5.224038e+05
50%	6.827910e+06
75%	3.425334e+07
max	7.954448e+09

Shape & Column Names

`df_pop.shape: (260, 2)`

`df_pop.columns: ['Country', 'Population']`

Part 2 Preview

Taken from `output/cleaning_summary.pdf`

4.1 Clean `df_demographics`

(a) Some fields may have been missing or scarped incorrectly resulting in nonsensical values or `NaN` $\rightarrow$ we made sure columns have appropriate numeric data types, and dropped rows with missing values, or values not in expected range.

(b) All countries had valid, numeric life expectancy values, in range $[40, 100]$.

(c) All countries had valid, numeric population density and urban population values.

(d) Country names are inconsistent between datasets $\rightarrow$ In order to keep as many countries as possible when merging datasets down the line, we applied a normalization scheme to country names.

(f) Overall, 7 countries were affected. For Brunei, a leading space was removed. The rest of the mismatches were caused by irregular capitalization (as Python's `str.title()` does not handle words such as "and" or "of" and abbreviations).

Old value	New value
Antigua and Barbuda	Antigua And Barbuda
Bosnia and Herzegovina	Bosnia And Herzegovina
Brunei	Brunei
Côte d'Ivoire	Côte D'Ivoire
DR Congo	Dr Congo
State of Palestine	State Of Palestine
Trinidad and Tobago	Trinidad And Tobago

- Rows before cleaning: 200
- Rows after cleaning: 200 (0 dropped)

4.2 Clean `df_gdp`

Issues encountered:

- Non-numeric characters (commas, currency symbols, etc.) in the `GDP_per_capita_PPP` column $\rightarrow$ prevented direct float conversion
- Missing values introduced when malformed strings were coerced to `NaN` $\rightarrow$ needed to be documented and removed
- Potential outliers in the GDP distribution $\rightarrow$ important to flag for downstream analysis
- Duplicate country entries $\rightarrow$ required de-duplication logic
- Inconsistent country names $\rightarrow$ needed standardization

Actions taken:

1. **Type conversion**
   Cast `GDP_per_capita_PPP` to string, stripped all non-digit/decimal characters, then converted to numeric (no invalid rows found).

2. **Missing-value handling**
   Exported rows where `GDP_per_capita_PPP` was `NaN` to `output/dropped_gdp.csv`, then dropped them from the DataFrame (no rows were dropped).

3. **Outlier detection (Tukey method)**
   Calculated Q1, Q3, and IQR; flagged values outside [Q1 – 1.5·IQR, Q3 + 1.5·IQR] as outliers (found 6 outliers).

4. **Duplicate removal**
   Removed duplicate rows by `Country` (no duplicates were found).

5. **Country-name standardization**
   Applied the same normalization as `df_demographics`.

- Rows before cleaning: 213
- Rows after cleaning: 213 (0 dropped)

4.3 Clean `df_pop`

Issues encountered:

- Non-numeric characters (spaces, commas, text) in the `Population` column $\rightarrow$ prevented direct numeric conversion

- Missing values created when malformed strings were coerced to `NaN` $\rightarrow$ needed to be documented and removed
- Potential outliers in the population distribution (on a $\log_{10}$ scale) $\rightarrow$ important to flag for review
- Duplicate country entries $\rightarrow$ required de-duplication logic
- Inconsistent country names $\rightarrow$ needed standardization

Actions taken:

1. **Type conversion**
   Cast Population to string, removed all non-digit/decimal characters, then converted to numeric (no invalid rows found).
2. **Missing-value handling**
   Rows Population was `NaN` were dropped (no such rows found).
3. **Outlier detection (Tukey on log scale)**
   Computed $\log_{10}$(`Population`), then calculated Q1, Q3, and IQR; flagged values outside [Q1 − 1.5·IQR, Q3 + 1.5·IQR] as outliers (1 outlier found).
4. **Duplicate removal**
   Removed duplicate rows by `Country` (no duplicates were found).
5. **Country-name standardization**
   Applied the same normalization as `df_demographics`.

- Rows before cleaning: 260
- Rows after cleaning: 260 (0 dropped)

Part 3 Preview

5.1 New Feature

Below is the result of running `describe()` on the new column `TotalGDP`:

mean	889352331458.71
std	3095183898278.81
min	391387777.17
25%	36903699628.76
50%	112857767960.96
75%	566276868415.05
max	29108919292117.52

5.2+5.3 Transforms

We applied $\log_{10}$ transformation and z-score normalization to columns `GDP_per_capita_PPP` and `Population`, in order to compress scale and center the data, providing a more comprehensible comparison. To the column `LifeExpectancy_Both`, which already is approximately normally distributed, we only applied z-score normalization. Here are the results of running `describe()` on the normalized columns:

Stat.	GDP	Population	Life Exp.
mean	-0.00	-0.00	-0.00
std	1.00	1.00	1.00
min	-2.44	-2.39	-2.71
25%	-0.76	-0.58	-0.74
50%	0.09	0.07	0.12
75%	0.84	0.71	0.75
max	1.94	2.71	1.68

5.4 Data Integration

After normalizing the `Country` column across dataframes, and setting it as index, we merged them. We were left with 173 countries:

`Afghanistan, Albania, Algeria, Angola, Antigua And Barbuda, Argentina, Armenia, Aruba, Australia, Austria, ...`

The countries **dropped** from the crawled demographics dataset are:

`Cabo Verde, Cuba, Curaçao, Czech Republic (Czechia), Côte D'Ivoire, Dr Congo, Eritrea, French Guiana, French Polynesia, Guadeloupe, Guam, Martinique, Mayotte, Micronesia, New Caledonia, North Korea, Réunion, Sao Tome & Principe, South Sudan, St. Vincent & Grenadines, State Of Palestine, Taiwan, Timor-Leste, U.S. Virgin Islands, Venezuela, Western Sahara, Yemen`