Assignment 2

Nick Karlman

Report: Subset of TableA's Attributes

Attributes:

Table A Schema: GDP ID, City Proper/Metropolitan Area, Country/Region, Official est. GDP ($ Billions), GDP Year, Metropolitan Population, Population Year

Table B Schema: City ID, City, Region, Country, Population, Population Year

Set S: City, Country/Region, Population, Population Year

City:

Missing Values: There are no missing values or Nulls in this attribute.

Type: Textual. Value Length: Average 18.9, Maximum 78, Minimum 3

Outliers: There are no outliers.

Value within Values: Country/Region truly has multiple values within one. Some countries may have cities but not regions, others the opposite, and some have both. This inconsistency causes some records to have extra information. For example, the records of cities in the U.S. in this attribute have the format City, State. If we are ever interested in breaking down the data further, we can separate this attribute into two.

Quality: I have discovered that there are 25 duplicate city records in the table. These duplicates have come from a secondary table from the source page. I scrapped the data from both tables and combined them, not realizing that there are now some cities represented twice. The more recent record should be kept, and the older one removed. Otherwise, if the names of the cities are correct, then there should not be any other issues in this data. When there are cities from all around the world, it is worth double checking the spelling and punctuation.

Country/Region:

Missing Values: There are no missing values or Nulls in this attribute.

Type: Textual. Value Length: Average 9.4, Maximum 23, Minimum 4

Outliers: The representation in TableA is slightly skewed. In "Country/Region Representation vs Population" (below) it is apparent that the U.S. is in more than 40% of the records in TableA.

Obviously, this is not an outlier, but worth noting that the U.S. and other countries have more (or less) representation in TableA to how their population compares.

Value within Values: The country names are very clean and concise.

Quality: I see no other data quality problems with this attribute.

Population:

Missing Values: There are 6.7% (67/999) Nulls. Since this is a population, a replacement value of -1 should satisfy.

Type: Numerical

Outliers: Of course, in the world, not every country has equal population. Instead, the representation of a country in the table is interesting when paired with the population of that country. If you have not already, view "Country/Region Representation vs Population" below.

Format: Integer Format – No commas or decimals

Quality: Assuming the sources are correct, there are no data quality issues.

Population Year:

Missing Values: There are 6.7% (67/999) Nulls. Since this is a year, a replacement value should be a numerical value. Since year can be categorical as well, a null or -1 would be sufficient to separate these records from those with recorded values. View "Population Year Representation" below.

Type: Numeric or Categorical

Outliers:1.8% (18/999) of records are 10 years or older. If decisions were to be made with this data, it should be known that there are some outdated records.

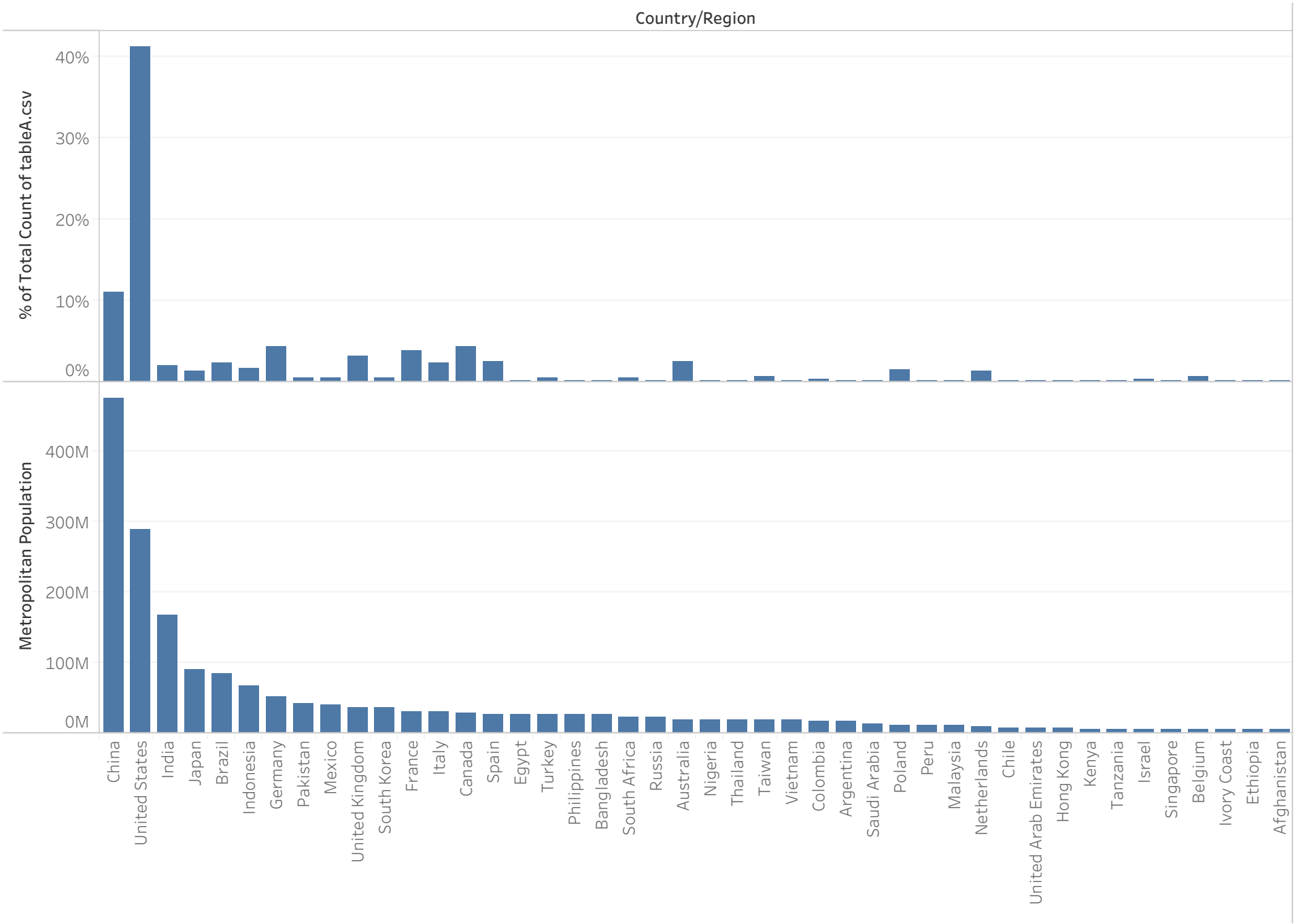
Format: The YYYY format is consistent among all real values.

Quality: If this data is factual, I see no other data quality issues with this attribute.

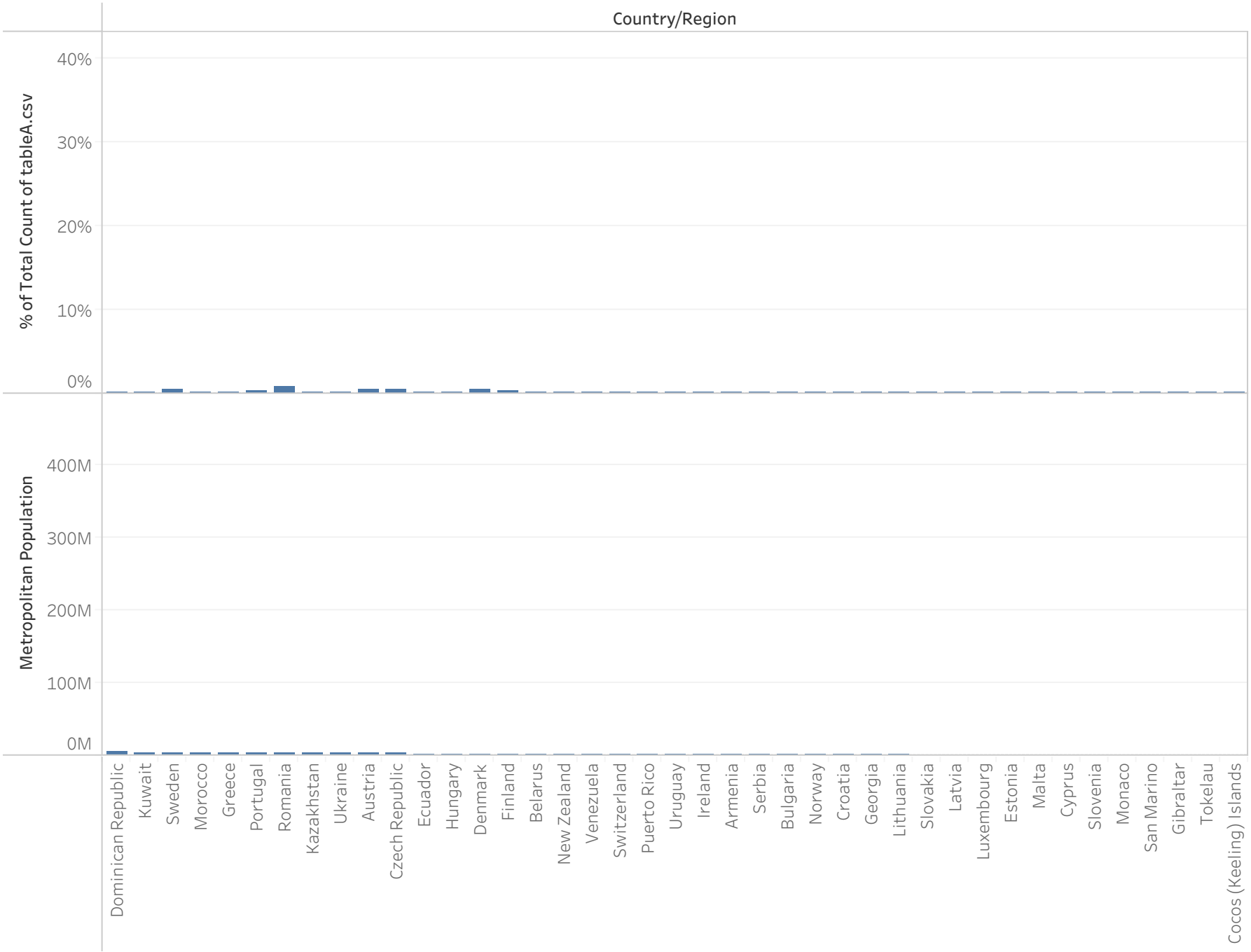
Summary

Since the source of this table is Wikipedia, there is no telling if the data within is factual without further exploration. Overall, the data in this set S seems to be factual and clean. There is room for improvement though, data can be extracted further, duplicates can be removed, and nulls can be valued to fit the attribute type.

Country/Region Representation vs Population



Country/Region Representation vs Population



Population Year Representation