Writeup: Reflections and Implications

1. What biases did you expect to find in the data (before you started working with it), and why?
2. What (potential) sources of bias did you discover in the course of your data processing and analysis?
3. What might your results suggest about (English) Wikipedia as a data source?
4. Can you think of a realistic data science research situation where using these data (to train a model, perform a hypothesis-driven research, or make business decisions) might create biased or misleading results, due to the inherent gaps and limitations of the data?

I will try to answer these 4 questions in my writeup as below:

Before I started working on the project, the first thing I thought is why (English) Wikipedia? How about non-English speaking country, since I'm from China, English is not our native language we use every day. I will suggest if we can get Wikipedia data from all different languages, not just (English) from Politicians by Country from the English-language Wikipedia (figshare.com), then it will avoid biases happening from the data sources.

Then during the data processing and analysis, we are doing comparation between countries and regions by 'articles_per_population', the second thing which impacts a lot is population of the county. As we know, China and India have the most population of the world. Therefore, there's no surprising when we saw them among the list of "Bottom 10 countries by coverage: 10 lowest-ranked countries in terms of number of politician articles as a proportion of country population", and the smallest population countries are listed on the "Top 10 countries by coverage: 10 highest-ranked countries in terms of number of politician articles as a proportion of country population". From the data analysis result, we also see something interesting that North Korean and Saudi Arabia among the list of "Top 10 countries by relative quality: 10 highest-ranked countries in terms of the relative proportion of politician articles that are of GA and FA-quality". Why? I guess because they get the most attention from the whole world about their polity and their government decisions.

If we use these data to do a realistic data science research and to train a model, as we learned from Samuel, S. AI's Islamophobia problem https://www.vox.com/future-perfect/22672414/ai-artificial-intelligence-gpt-3-bias-muslim and Duarte, N., Llanso, E., & Loup, A. "Mixed Messages? The Limits of Automated Social Media Content Analysis". https://cdt.org/wp-content/uploads/2017/12/FAT-conference-draft-2018.pdf. I think it will indeed bring bias and unfairness from the training data source, since non-English language is one of NPL's limitations. What's more, we also need to bring the data from difference data sources, not only by Western sources, which can help us to reduce the biases happening in the data source.