Data & Analytics Challenge

In the area we have a theory we would like you to test and confirm or reject:

“Because the United Kingdom (UK) was the main country to colonize the United States of America (USA), and UK is on the east side of USA, there are more cities/towns with UK names alongside USA´s East Coast than West Coast”.

We would like you to work on some open datasets from https://public.opendatasoft.com as your primary source.

We wait for your answer in PDF format, which includes:

1. Your point about our theory, supported on any analysis needed.

2. A brief comment about the problems you have faced while working on the challenge.

3. Improvements you could do in the future to provide a better response.

DELIVERED TASK

1) From a historical standpoint, the theory makes sense. Nomination used to be inspired on the culture and context from its main characters, and because of that is pretty normal there are more cities with UK names alongside the USA´s East Coast, the originary settlement place for british colonists. On the other side, Wild West conquest was much later in time, and had a more complex cast.

However, the difference in the total amount of cities/towns between both coasts should be considered as a factor to relativize the final answer.

From a data standpoint, here is a copy from the Jupyter Notebook I have worked on (for the script, see attached notebook).

2) *First issue* I faced was **defining cities belonging to each group** (East Coast, West Coast, UK). Those are different administrative border levels. I resolved it using “Timezone” feature, which allowed evening all groups.

Second issue was **it could be more than one city with the same name**. I solved this creating a list with unique UK cities names and then iterate US cities column, so I could define if they match or they do not, generating a binary mark.

Finally, a *third issue* was **cities with composite names** (e.g., Newport East). Not exactly matches were excluded, so this issue can not be resolved (there are more US cities with UK names than the results, as the number of combine words for US cities with UK names increase exponentially).

3) I could improve the criteria to even cities groups, because “Timezone” feature includes cities that are not considered part of any USA´s coast and also have data source categorization mistakes, as it seems on the map. State divisions are included as a feature in the dataset, so I could be more effective defining cities belonging to coasts based on them.

Moreover, I could be more careful with formal requirements, as renaming Data Frames columns.

CONCLUSION

“The deep gap between US Coast cities counter drove us to find a better comparative indicator, based on percentages. “Percentage of Cities with UK names over Total Cities by Coast” returns 8,6% for East Coast and 4,9% for West Coast, so even though there is a difference, it is insignificant, so we could say the starting hypothesis must be rejected”.