Desafío analítico Data & Analytics

En el área tenemos una hipótesis de pasillo que nos gustaría que tú la revisaras para confirmar o rechazar:

"Dado que el Reino Unido (UK) fue uno de los principales países que colonizaron los Estados Unidos (USA), y UK se encuentra en el lado este de USA, entonces hay más ciudades/poblados con nombres de ciudades de UK en la costa este de USA en comparación a la costa oeste".

Nos gustaría que ocupes alguno de los datasets abiertos de la página https://public.opendatasoft.com como base para tu análisis.

Esperamos que tu respuesta vuelva en un documento en formato PDF e incluya:

1. Tu juicio sobre nuestra hipótesis, con cualquier análisis de apoyo que generes.
2. Una sección breve sobre cuales fueron tus principales dificultades que te enfrentaste en el análisis.
3. Que mejoras o futuro análisis podrían hacerse para mejorar tu respuesta.
4. Una copia de cualquier tipo de código que hayas producido y no hayas incluido en el punto 1.

ENTREGABLE

1) A priori, desde un enfoque histórico, el juicio tiene sentido. La nominación suele inspirarse en la cultura y el contexto de los protagonistas de cada historia, con lo cual es factible que existan más ciudades con nombres del Reino Unido en la costa este de los Estados Unidos, lugar originario de asentamiento de los colonos ingleses. Por otra parte, la conquista del lejano oeste fue muy posterior en el tiempo y tuvo como reparto a una mayor complejidad de actores.

Un factor que puede relativizar la respuesta al problema planteado es la cantidad total de poblados/ciudades existentes en cada costa.

Desde una aproximación de los datos, adjunto una copia del Jupyter Notebook en el que estuve trabajando para resolver el desafío (para el código, ver Notebook aparte).

2) La *primera dificultad* que tuve fue **determinar la pertenencia de las ciudades a cada grupo** (Costa Este, Costa Oeste, Reino Unido), ya que estos implican divisiones administrativas de distinto nivel. Ello pudo ser resuelto utilizando la variable "Timezone", referencia con la cual pude igualar los tres grupos.

La *segunda dificultad* fue **el descubrimiento de que puede existir más de una ciudad o poblado con el mismo nombre.** En ese sentido resolví crear una lista con nombres únicos de las ciudades del Reino unido para luego iterar la columna con los nombres de las ciudades estadounidenses, y así determinar si coinciden o no en la nominación, generando una marca binaria

Por último, una *tercera dificultad* radicó en que existen **ciudades con nombres compuestos** (más de una palabra) y la operación de iteración de nombres deja afuera las coincidencias no exactas (por ejemplo, Newport East). De esta forma existen más ciudades estadounidenses con nombres del Reino Unido, tal y como lo muestra evaluar el número combinado de palabras de las ciudades estadounidenses con nombres del Reino Unido en cada costa.

3) Podría mejorar el criterio de igualación de los grupos de ciudades, ya que la variable "Timezone" puede incluir ciudades que no están consideradas parte de cada costa de los Estados Unidos, así como también errores de la fuente de datos, como se aprecia en el mapa. El dataset dispone de los datos de los Estados a los cuales pertenece cada ciudad, con lo cual podría afinarse la determinación de la pertenencia a una u otra costa.

Por otra parte, mejorar cuestiones formales, como renombrar las columnas de los Data Frames.

CONCLUSIÓN

"Teniendo en cuenta la enorme disparidad en la cantidad de ciudades de cada costa, la comparación pertinente debe realizarse en porcentajes. El indicador "Porcentaje de Ciudades con nombres del Reino Unido sobre el Total de Ciudades de cada Costa" arroja 8,6 % para la Costa Este y 4,9 % para la Costa Oeste, con lo cual si bien existe una diferencia, es mínima, y la hipótesis inicial debe ser rechazada".