Pontificia Universidad Católica de Chile
IMT2000 - Ciencia de datos: Qué, cómo y por qué
Primer Semestre 2023

Proyecto C - Entrega 1

Fecha de entrega: 29 de mayo, 2023

En el proyecto C, trabajarán con datos extraídos del mundo de Game of Thrones (libros de la saga Canción de Hielo y Fuego).

Contexto

Canción de Hielo y Fuego es una multipremiada serie de novelas y novelas cortas de fantasía épica escritas por el novelista y guionista estadounidense George R. R. Martin. La historia de Canción de Hielo y Fuego se sitúa en un mundo ficticio medieval, principalmente en un continente llamado Poniente pero también en un vasto continente oriental, conocido como Essos. La mayor parte de los personajes son humanos pero a medida que la serie avanza aparecen otras razas. Hay tres líneas argumentales en la serie: la crónica de la guerra civil dinástica por el control de Poniente entre varias familias nobles; la creciente amenaza de los Otros, apenas contenida por un inmenso muro de hielo que protege el norte de Poniente; y el viaje de Daenerys Targaryen, la hija exiliada del rey que fue asesinado en otra guerra civil hace quince años, quien busca regresar a Poniente a reclamar sus derechos. Estas tres historias interactúan entre sí y son extremadamente co-dependientes.

Datos

Para esta entrega, deberán usar los archivos:

- battles.csv: Describe batallas que han ocurrido en la serie.
- deaths.csv: Describe a los personajes y sus muertes.

Ambos datasets fueron obtenidos de Kaggle. Estos los encontrarán en Canvas, en la carpeta de Proyecto C junto a un .txt que explica cada columna.

Ejercicios

Cargar y describir los datos (10 puntos)

1. Cargar cada uno de los archivos en un DataFrame utilizando la librería pandas.
2. Seleccionar y mantener sólo las columnas que utilizará. Justifique para cada columna que conserve. Esto lo debe hacer para cada dataset.
3. Indique cuántas filas y columnas tiene cada dataset.
4. Elija 2 funciones de pandas que ayuden a entender los datos y explique el resultado. Cada función se debe aplicar y explicar por cada dataset.

Limpieza y preprocesamiento de datos (5 puntos)

- En caso de haber datos faltantes, duplicados o sucios, identificarlos y luego decida que hacer con ellos. Justifique y ejecute.

Nuevos datos (15 puntos)

1. Crear una columna en el dataset battles.csv que muestre en cuántos libros aparece cada personaje.
2. Crear un dataset de las Casas Nobles según lealtades (Allegiances de deaths.csv), que incluya la información de cuántos nobles hay por cada casa, el porcentaje de mujeres, de hombres y cuántos de los personajes están vivos y cuántos muertos.
3. Hay personajes que sabemos que están muertos, pues nos aparece el año de su muerte, pero no sabemos como murieron. Suponiendo que murieron en batalla, indique en qué batalla pudieron haber muerto. En caso de haber más de una batalla posible, quédese con aquella donde combatió la mayor cantidad de gente. Justifique bien su razonamiento para resolver el ejercicio.

Preguntas (20 puntos)

1. ¿Cuál es el comandante que más batallas ha ganado?
2. Cuántas mujeres que sean nobles, que estén vivas y de casas aliadas que han ganado al menos una batalla aparecen por cada libro?
3. Basándose en las muertes de los personajes, ¿Cuál es la casa que más muertes de aliados tuvo por cada año?
4. En algún punto de la serie, se propone que los ejércitos más numerosos son los que ganan las batalllas, ¿Esto es cierto? Justifique.
5. ¿Cuál casa tiene más personajes no nobles importantes? Considérese importante como aquel personaje que haya estado vivo al menos por 10 capítulos desde que lo presentaron en la serie.

Orden y formato (10 puntos)

Este puntaje se asigna según criterios de orden, código legible, formato correcto, ortografía y calidad de explicaciones. Todos parten con los 10 puntos y se irá descontando según aplique.

Bonus

Pueden acceder hasta a +10 décimas bonus aquellos que realicen algo interesante con el dataset. Tienen que justificar por qué eso nos entrega más información y explicar bien. La asignación queda a criterio del corrector.

Entrega

Deberán entregar un archivo en formato .ipynb que contenga su trabajo. Se habilitará en Canvas un cuestionario para subir la entrega. Sólo un integrante por grupo deberá subirlo.