Ciencia de Datos

Práctico N°1: Nivelación con Pandas y Seaborn

Introducción: Kaggle, es una plataforma online de acceso gratuito, lanzada en 2010, para dar soporte a la comunidad de data scientists, permitiendo explorar y publicar tanto bases de datos como modelos y desarrollar competencias (sólo con costo para empresas) para resolver desafíos de la Ciencia de Datos. Desde 2017 es una subsidiaria de Google LLC. Para poder acceder a los recursos de la plataforma es necesario registrarse previamente como usuario en caso de no tener cuenta. Entre los recursos, Kaggle proporciona cursos cortos, gratuitos y autocorregidos para adquirir destrezas básicas en el uso de las herramientas usuales en la industria.
Por otro lado, Google proporciona Colaboratory, o Colab para abreviar, un producto que permite escribir y ejecutar código en Python desde un navegador de internet, usando recursos de los propios servidores de Google, si bien en la versión gratuita están limitados.
A los fines de nivelar nuestros conocimientos, las primeras dos tareas consisten en hacer dos cursos y conseguir los correspondientes certificados y luego resolver un desafío creando una notebook en Colab.

Tarea 1: Completar el curso sobre **Pandas** de Kaggle.

Tarea 2: Completar el curso sobre **Data Visualization** de Kaggle.

Tarea 3: Para aplicar lo aprendido se proponen las siguientes actividades:

a Descargar el dataset Airplane Crashes and Fatalities upto 2023 disponible en Kaggle. Estudiar el diccionario de variables. Luego almacenar el archivo `csv` en el drive personal de la cuenta institucional de la UNC.

b Crear una instancia en Colab y averiguar cómo acceder al sistema de archivos de drive para cargar el `csv` usando Pandas, con la variable `Date` como índice de fecha. Usando lo aprendido con Pandas se obtiene un error porque el encoding del archivo no está en `UTF-8`, que es lo que se supone por default. Recordar: en la vida *los datos nunca están en el formato deseado.* Para cambiar el encoding en la lectura usar: `encoding='latin-1'`.

c Observar la información del dataset que proporcionan las funciones `info()` y `describe()`.

d Mostrar un `lineplot` usando Seaborn con el número *total* de personas embarcadas muertas (`Fatalities`) en cada uno de los días. Para esto, agrupar previamente por fecha y sumar los datos de esa columna.

e Reordenar el dataset usando la columna (`Fatalities`) en forma descendente para mostrar los dos accidentes con mayor número de muertes entre las personas embarcadas.

f Calcular el número de personas embarcadas sobrevivientes en esos dos vuelos y la correspondiente proporción sobre el número total de personas embarcadas.

g Crear un `DataFrame` con una variable que sume el número de (`Fatalities`) en cada uno de los meses del año –de enero a diciembre– (usando el índice de la base) y otra columna

con los correspondientes porcentajes sobre el total de `Fatalities`. ¿Cuáles son los meses con menores porcentajes?

h Mostrar un `barplot` usando Seaborn con los porcentajes de sobrevivientes en cada uno de los meses del año.

Para completar esta tarea, el colab debe contener al menos una celda para cada ítem con el código Pandas y el resultado de ejecutarlo.

P. Pury 2024