Tarea 1 – Introducción a Data Science
Data Scraping y Visualización de Datos

Profesor: Brian Keith Norambuena, brian.keith@ucn.cl.
Ayudante: Matías Salas Villanueva, matias.salas@alumnos.ucn.cl.

Objetivos

1. Aplicar técnicas de Web Scraping y preprocesamiento de datos para la obtención de un conjunto de datos.
2. Aplicar técnicas de exploración de datos y visualización sobre el conjunto de datos.
3. Analizar las visualizaciones generadas aportando insights interesantes sobre los datos.

Entregables

- Un cuaderno interactivo *Jupyter Notebook* con todo el código fuente, resultados, y análisis.
- El cuaderno interactivo debe utilizar apropiadamente celdas de *markdown* y de código.

Fecha de Entrega

5 de septiembre a las 23.59 por Campus Virtual con el formato indicado en la sección de Consideraciones de Evaluación y Administrativas.

Requerimientos Mínimos

Se describen a continuación los requerimientos mínimos para esta tarea.

1. Web Scraping [3 puntos]:
   - Generalizar la obtención (scraping) de datos para **todas las páginas** de un sitio web y obtener: Nombre del cohete, Compañía lanzadora, Fecha de lanzamiento (AAAA/MM/DD), Locación del lanzamiento, Exitoso (1 o 0), Valor total del lanzamiento, Peso del Cargamento en kg (LEO + GTO en una única columna), Altura del Cohete, Altura y Diametro del carenado. **(+1 punto)**
   - **Debe generar** una columna de volumen del cohete (volumen cilindro del cohete + volumen cono del carenado). Puede agregar **opcionalmente** columnas que contengan la imagen del lanzamiento, fuente (link del lanzamiento) y carga del lanzamiento si desea extender su conjunto de datos.
   - Para cada columna del conjunto de datos debe existir una entrada limpia y normalizada, evitando el ruido o basura en las entradas. **(+1 punto)**
     Si no existe información para el valor/altura/peso cargamento, puede calcular el promedio del resto de cohetes y rellenar estos vacíos.
     Ejemplo: Si su conjunto de datos tiene una columna de RUTs, todas las entradas de esta columna deben tener el mismo formato (ya sea 12.345.678-9, 12345678-9 u otro).
   - Transformar los datos obtenidos de un dataframe a un archivo (.xlsx, .xls o .csv) y una descripción **breve** de cada columna en una celda Markdown. **(+1 punto)**
   - Se solicita la utilización del sitio https://nextspaceflight.com/launches?tab=past.
2. Visualización de los Datos [2 puntos]:
   - Debe generar visualizaciones que puedan responder a las siguientes preguntas: **(+0.5 puntos c/u)**
     ¿Qué empresas han enviado más cargamento al espacio exitosamente?
     ¿Existe alguna relación entre el valor total del lanzamiento y el peso del cargamento?
     ¿Qué empresas lanzan los cohetes más grandes en volumen (m^3)?
     ¿Cómo se distribuyen los volúmenes de los cohetes?, diferencie aquellos que tuvieron un lanzamiento exitoso y aquellos que no.
3. Análisis de Resultados [2 puntos]:
   - Explique cómo ha evolucionado cada país anualmente desde su primer lanzamiento de cohete y muestre la tasa de éxito anual para cada uno utilizando un histograma. **(+2 puntos)**
     La tasa de éxito anual se define como: $\frac{Total\ lanzamientos\ exitosos}{Total\ de\ lanzamientos} \cdot 100\%$ para cada año.

Consideraciones de Evaluación y Administrativas

- La entrega de esta tarea es de carácter individual y cualquier indicio de plagio será motivo de aplicación de nota mínima.
- El **formato de entrega** será el siguiente "NombreApellido1_NombreApellido2.ipynb" (sin tildes) mediante plataforma de campus virtual.
  Ejemplo: **PedroRojas_JosePerez.ipynb**
- A partir de las 00:00 hrs del día siguiente del día de entrega comenzará un descuento de 10 décimas **por cada hora** de retraso, Ejemplo: Si el límite de entrega es el 29 de agosto a las 23:59 hrs y la tarea es entregada el 30 de agosto a las 01:00 hrs, tiene derecho a nota máxima 6.0.
- La escala de evaluación será al 60%.
- Para recibir el puntaje completo debe ser asegurarse de que el cuaderno pueda ser ejecutado de corrido, de lo contrario se aplicará la **mitad del puntaje**.
- **Si utiliza Inteligencia Artificial generativa** (e.g., ChatGPT) durante su trabajo, por favor incluya todas sus *prompts* al final del cuaderno interactivo como documentación o entregue un anexo con esta información. No habrá penalización por el uso de estas herramientas como apoyo, pero no deberían ser utilizadas para resolver la tarea por ustedes.
- Los Créditos Extras serán puntos extras que se sumarán al final del semestre a tareas que no tengan el puntaje completo, comenzando por la tarea con menor puntaje. Los créditos extra sobrantes se anularán si el estudiante tiene puntaje máximo en todas las tareas.
- Los análisis de resultados y correspondientes conclusiones serán evaluados según la claridad de sus ideas y su presentación (escritura y formato).