Unidad 2 - Manipulación de datos

Fundamentos de ciencia de datos

1. Escriba la siguiente tabla en cada uno de los formatos aprendidos en clase usando un procesador de texto.
    a. csv. Con delimitador |
    b. txt
    c. yaml
    d. xml
    e. json
    f. html

id	desc_prod	precio	proveedor
0049570	camisa	2000	fashionistas
0769298	jean	6000	tu moda
8458909	polera	3000	el ropero

2. Agregar una descripción al producto `jean` que diga: “skinny”

2. Cargar el archivo *.yaml, *.json, *.csv y *.txt generados en el ejercicio 1 con el paquete correspondiente. Visualizar el resultado para verificar la correcta creación del archivo
3. Leer el archivo `flete-aereo-vacunas-covid19-al-2021-06-28.xlsx`
    a. Calcular el porcentaje de cada vuelo y verificar que la suma de los mismos sea 1
    b. Calcular el promedio de lo facturado usando la columna `factura_moneda_monto`. Realizar un boxplot para con estos datos para entender la distribución de los mismos.
    c. Informar cuándo fue el último vuelo
    d. Calcular la cantidad de días que pasaron entre el primer y el último vuelo
    e. Escribir el archivo en formato parquet
4. Leer el archivo `incendios-cantidad-causas-provincia_2022.csv`
    a. Obtener el número de incendios totales por año para todo el país
    b. Realizar un gráfico de barras con el número de incendios totales para cada año del período 1993-2021 para la provincia de Córdoba
    c. Realizar un gráfico de barras que compare el número de incendios intencionales, por negligencia y naturales para el período 2015-2021 en la provincia de Santa Fe.
    d. Obtener el promedio para todo el período del número de incendios intencionales, por negligencia y naturales para la provincia de Río Negro.
5. Escribir una función para realizar una interpolación lineal por tramos para los datos de la tabla de abajo. La función recibe como input el valor de x y como output el valor de y correspondiente:

x	**y**
1	2
2	3
3	5
10	6

7. La siguiente tabla resume la evolución de la población total argentina desde 1960 a la actualidad según los censos nacionales de población (fuente: INDEC):

Año	Población total
1960	20013793
1970	23364431
1978	
1980	27949780
1986	
1991	32615528
2001	36260130
2010	40117096
2014	
2022	46044703

Utilizando una interpolación lineal, complete la información sobre **Población total** para aquellos años en los que no se cuenta con datos de censos nacionales.

8. Utilizando regex:

   a. Escriba una función que determine si una url es válida e imprima 'URL válida' (por ejemplo para `"https://pythondiario.com/"`) si la url dada como input es válida y 'URL no válida' en caso de que no sea válida (p.ej: `"https:/pythondiario.com/"`).

   b. Escriba una función que determine si una dirección de correo electrónico es un correo electrónico **de gmail** válido.

   c. Escriba una función que determine si un string corresponde a una fecha válida y se encuentra en el formato YYYY-MM-DD.

   d. Utilizando el archivo `Me_gustas_tu-Manu_Chao.txt` , que contiene la letra de la canción 'Me gustas tu' de Manu Chao:

   - indique cuántas veces en la canción se hace referencia al verbo gustar
   - ¿cuántos verbos en infinitivo tiene la letra de la canción ?
   - Realice una lista de todas las cosas que le gustan a Manu Chao. Por ejemplo: cosas_que_le_gusta=[los aviones, viajar, la mañana, el viento, soñar, la mar …. etc]

9. Usando los datos de `listing` `s_ba.csv` de Buenos Aires que se encuentran cargados en el campus imputar los precios de alquiler faltantes empleando:

   a. La media y moda de los datos no faltantes

   b. Una medida de resumen de su elección por barrio y tipo de habitación

c. Los 10 puntos más cercanos geográficamente a cada dato faltante usando las coordenadas.

En cada uno de los casos indicar la cantidad de datos que se usaron para la imputación

10. Utilizando los archivos `conicet_personas_2020.xlsx`, `conicet_ref_sexo.xlsx` y `conicet_ref_grado_academico.xlsx` disponibles en el dataset de la unidad 2, genere una tabla en la cual se informe cuántos empleados de CONICET hay de cada sexo para cada máximo grado académico en 2020.
11. Utilizando el archivo `incendios-cantidad-causas-provincia_2022.csv` del punto 5 generar una tabla que muestre el número de incendios intencionales por provincia para cada año (por ej.: las provincias como filas y los años en las columnas).
12. Calcular la similaridad de Jaro, Jaro-Winkler y Levenshtein manualmente y luego verificarlas usando los paquetes de `jaro` en los primeros dos casos y de `enchant` en el último.

cadena 1	**cadena 2**
Mariana	Merianna
Della Ceca	Dellacecca
Córdoba 2568	Cordoba 2478