Práctica N° 2 - Manipulación de Datos

Ejercicio N° 1

Lea en *Colab* el archivo `flete-aereo-vacunas-covid19-al-2021-06-28.xlsx` y realice cualquier tarea de limpieza y/o adecuación del dataset que considere necesaria.

1. Según la información que se encuentra en la base de datos, ¿cuál fue el número de vuelo que realizó una mayor cantidad de fletes? *Sugerencia:* utilice el método `value_counts()` de la librería Pandas.
2. ¿Cuántos registros no contienen información sobre el número de vuelo?
3. Calcule el promedio de lo facturado en todos los vuelos realizados **considerando únicamente los registros de viajes que se facturaron en USD** según la información contenida en la columna `factura_moneda_monto`.
4. ¿Qué porcentaje de los vuelos realizados tuvieron como origen a Rusia o a China?
5. ¿Cuál es el vuelo más reciente de los que se tiene registro en el dataset? ¿Cuántos días transcurrieron entre el primer y el último vuelo realizados?
6. Escribir el archivo en formato .parquet.

Ejercicio N°2

Lea en *Colab* el archivo `incendios-cantidad-causas-provincia_2022.csv` y realice cualquier tarea de limpieza y/o adecuación del dataset que considere necesaria.

1. Obtenga el número de incendios totales por año **para todo el país**. ¿Cuál fue el año en el que se presentó un mayor número de incendios?
2. Obtenga el número de incendios totales por año para el período 1993-2021 **en la provincia de Córdoba**.
3. Realice una tabla en la que se muestre, para cada año del periodo 1993-2021, la provincia en la que tuvo lugar el mayor número de incendios intencionales. *Sugerencia:* explore las funcionalidades del método `idxmax()` de la librería Pandas.
4. Realice un gráfico de barras para visualizar el número de incendios intencionales, por negligencia y naturales que tuvieron lugar durante el periodo 2015-2021 en la provincia de Santa Fe.
5. Obtenga el número promedio de incendios intencionales, por negligencia y naturales para la provincia de Río Negro durante el periodo 1993-2021.

Ejercicio N°3

Escriba una función para realizar una interpolación lineal por tramos para los datos de la tabla de abajo. La función deberá recibir como *input* el valor de x y producir, como *output* el valor de y correspondiente:

x	y
1	2
2	3
3	5
10	6

Ejercicio N°4

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.

Ejercicio N°5

Usando los datos de `listings_ba.csv` de Buenos Aires, 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.

Ejercicio N°6

Utilizando *regex*:

1. 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 (por ejemplo: `"htps:/pythondiario.com/"`).
2. Escriba una función que determine si una dirección de correo electrónico es un correo electrónico **de gmail** válido.
3. Escriba una función que determine si un string corresponde a una fecha válida y se encuentra en el formato YYYY-MM-DD.
4. 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_gustan = [los aviones, viajar, la mañana, el viento, soñar, la mar .... etc]`.

Ejercicio N°7

Utilizando los archivos que contienen información sobre CONICET (`conicet_personas_2020.xlsx`, `conicet_ref_sexo.xlsx` y `conicet_ref_grado_academico.xlsx`), genere una tabla en la cual se informe cuántos empleados de CONICET hay de cada sexo para cada máximo grado académico en 2020.

Ejercicio N°8

Utilizando el archivo `incendios-cantidad-causas-provincia_2022.csv` del Ejercicio N° 2, genere una tabla que muestre el número de incendios intencionales por provincia para cada año de los incluidos en dicho dataset.

Ejercicio N°9

Para cada uno de los siguientes pares de cadenas, calcule la similaridad de Jaro y Jaro-Winkler y la distancia de Levenshtein. Realice primero el cálculo en forma manual y luego verifique los resultados obtenidos utilizando herramientas de las librerías `jaro` y `levenshtein`.

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

Ejercicio N° 10

El dataset `ventas.xlsx` contiene los registros de una serie de ventas realizadas en el último tiempo en un local de productos electrónicos. Por otra parte, cuenta con el dataset `clientes_base.xlsx`, el cual contiene información sobre los clientes registrados en dicho establecimiento.

1. ¿Cuál fue el monto total de venta de productos *iPad* y *MacBook*?
2. Utilizando el método `merge()` de `pandas`, realice un **outer join** entre ambos DataFrames utilizando la columna `nombre_cliente` como *key*. ¿Qué observa en el DataFrame resultante?
3. Considerando que en `clientes_base.xlsx` los nombres de los clientes se encuentran exentos de errores ortográficos y tipográficos, ¿en qué porcentaje de los registros que conforman el dataset `ventas.xlsx` el nombre del cliente coincide con el de un cliente correctamente registrado?
4. Teniendo en cuenta lo observado en los ítems anteriores, utilice herramientas de la librería `fuzzywuzzy` para realizar un *fuzzy join* de ambos datasets por el campo `nombre_cliente`. ¿De qué ciudad es el cliente que más compras realizó en el local?

Ejercicio N° 11

1. Escriba la tabla de datos en cada uno de los siguientes formatos usando un procesador de texto.

- csv, con delimitador |
- txt
- yaml
- xml
- json
- html

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

2. Agregue una descripción al producto `jean` que diga: “skinny”.
3. Cargue en *Colab* los archivos .yaml, .json, .csv y .txt generados en el inciso 1 utilizando funciones de `pandas`. Visualice el resultado para verificar la correcta creación del archivo.