Universidad Tecnológica Nacional
Facultad Regional San Nicolás
Aula Chivilcoy

Tema: "Un Enfoque Sencillo para Obtener Ideas Valiosas".

Grado: Tecnicatura en Programación

Materia: Metodología de la investigación

Profesor: Walter Rivas

Autor: Desia, Federico Martin

Localidad: Chivilcoy

Fecha de entrega: 03/11/2022

El presente informe se encuentra en el diario inglés "Gulf News ", el día 22/10/2022 en donde se expresa lo siguiente:

"Data science can be such a force for change - but do not fall for all of the hype".

"La ciencia de datos puede ser una gran fuerza para el cambio, pero no se deje engañar por todo el bombo publicitario" (traducción).

A la hora de buscar información local o del mundo o simplemente levantarse por la mañana y empezar a leer los diarios de la ciudad, se encuentra demasiada información. El típico dicho "Dato mata relato, dato es información" no está del todo bien, ¿por qué?, porque dato no necesariamente es información, puede haber 100 datos, 1000 datos, etc, pero puede que esos datos, sean irrelevantes, y por ende no sirva. Con un simple análisis de datos es posible ver más allá del relato y ver de dónde viene el dato.

¿Qué es la ciencia de datos?

Ciencia de datos se basa en plantear hipótesis, preguntas que luego se buscará responder a través de conclusiones basadas en el procesamiento de datos. Alguien plantea la hipótesis (ejemplo: el departamento de marketing, ventas, etc).

Cuando se empieza como analista de datos es bueno aprender programación, ya que, se hace la tarea mucho más fácil. Sea un matemático o no. Si lo hay perfecto, pero es mucho más simple.

¿La ciencia de datos tiene futuro?

Si, según el Foro Económico Mundial, el mundo va a generar más de 450 exabytes de datos al día para el 2025. Cada vez que se navega en la web dejamos un rastro digital. Esta huella informativa puede ser captada para sacar conclusiones y tomar decisiones. Además, las empresas necesitan conocer a sus clientes y sus comportamientos antes de tomar decisiones. El factor del crecimiento del mercado incluye: la adopción de soluciones basadas en la nube, la creciente aplicación de la plataforma de ciencia de datos en varias industrias y la creciente necesidad de extraer información detallada de datos voluminosos para obtener una ventaja comercial competitiva.

¿Confiabilidad de la ciencia de datos?

Cuando buscan, generan varios resultados al azar, pero esos resultados
no se sabe si son precisos o no, las personas están considerando el uso del análisis de datos como un sistema para un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados.

La ciencia de datos es relevante en todos los sectores

La ciencia de datos forma la base de decisiones importantes en varias industrias, desde predicciones en la agricultura, la toxicidad en los medicamentos y hasta la asignación de fondos en instrumentos financieros.

Fuente: https://gulfnews.com/business/analysis/data-science-can-be-such-a-force-for change---but-do-not-fall-for-all-of-the-hype-1.1666448031132

Planteamiento del problema

El presente trabajo, se establece por la siguiente razón: Como se puede hacer un simple análisis de datos de un dataset y no caer en un relato venga de donde venga, ya sea, el diario, una tesis, etc.

Estará destinado a toda persona que sea curiosa, solo se necesita de una conexión a internet, una computadora o celular. Y recomendable saber lo básico de programación.

Se va a trabajar en un dataset seleccionado, transformándolo a data frame y sacar una conclusión del dataset tomado, a través de un enfoque cuantitativo deductivo.

Se puede hacer con cualquier tema de interés ajeno.

Se hará de una manera lo más simple posible, también para que la persona sin conocimiento previo pueda llevarlo a cabo. Se recomienda tener conocimiento básico en programación.

Ideas de solución:

Planteado el problema, ahora se planteará dos soluciones para interpretar una información que se está leyendo, se tendrá dos caminos:

Idea 1: Plantear un análisis de datos simple: Se plantea llevar a cabo un simple análisis de datos, a través de la manipulación de un dataset, llegando a una conclusión deductiva. Se requiere tener una computadora, que contenga instalado Python, Spyder y librerías afines, además se requiere tener una dataset de interés descargado. La ventaja, es qué se podrá manipular los datos. Sus
desventajas, es qué puede que lleve tiempo encontrar el dataset, y que el dataset no esté por completo y esto lleve más tiempo para completarlo.

Idea 2: Confiar en lo que se ve: Se plantea que el análisis es hecho por otra persona, y se

confía ciegamente en que esa información es verídica. Se elige creer. Solo se necesita de una computadora, un celular y de conexión a internet. Su ventaja es que ahorra tiempo, pero sus desventajas es que no se sabe como sé ha llevado a cabo la manipulación de los datos, además, podría contener: sesgo, datos incompletos y datos atípicos.

Selección de la idea del proyecto:

Para continuar con este proyecto se decide obviamente ir por la **Idea 1: Se plantea un análisis de datos simple,** se elige esta opción, debido a que es menos sesgada y más sentido común tiene, además se podrá elegir los datos para tener en cuenta y llegar a una conclusión. Esto no quiere decir que sea el mejor análisis, pero se asegura de que es un análisis basado en el sentido común.

La **Idea 2: Se confía en lo que está viendo**, La idea 2 queda descartada por ser mayores las desventajas que tiene y poca ventaja. Es más rápido, pero no sé sabe cómo la persona llevó a cabo la manipulación de los datos, si estuvo sesgado o no a la hora de escribir el artículo.

Hipótesis de trabajo: La expectativa al realizar este trabajo, es que pueda ser llevado de una manera simple un análisis de datos, y de cómo resultado una conclusión que brinde más que una simple lectura y también que se verifique que los datos sean los correctos.

Desarrollo del Proyecto:

Este proyecto va a describir los pasos a implementar para llevar a cabo un análisis de datos, de la manera más simple. Debo aclarar que esta no es la única manera de hacerlo, si lo desea usted puede buscar otra manera de hacer un análisis de datos sencillo.

Esto está pensado para que cualquier persona lo pueda llevar a cabo, siguiendo el código paso a paso. Pero repito, es preferible que sepa lo básico de programación.

Puede suceder que usted esté leyendo un artículo y que este artículo contenga una información falsa, o que simplemente usted diga, voy a llevar a cabo un análisis propio para ver esta información desde otro ángulo.

Para llevarlo a cabo se utilizará un dataset a modo de ejemplo. **Spyder** que es parte del ambiente del lenguaje **Python**, para escribir el código, y además se trabajará con la librería **Pandas**, para leer un archivo **CSV.**

Marco teórico:

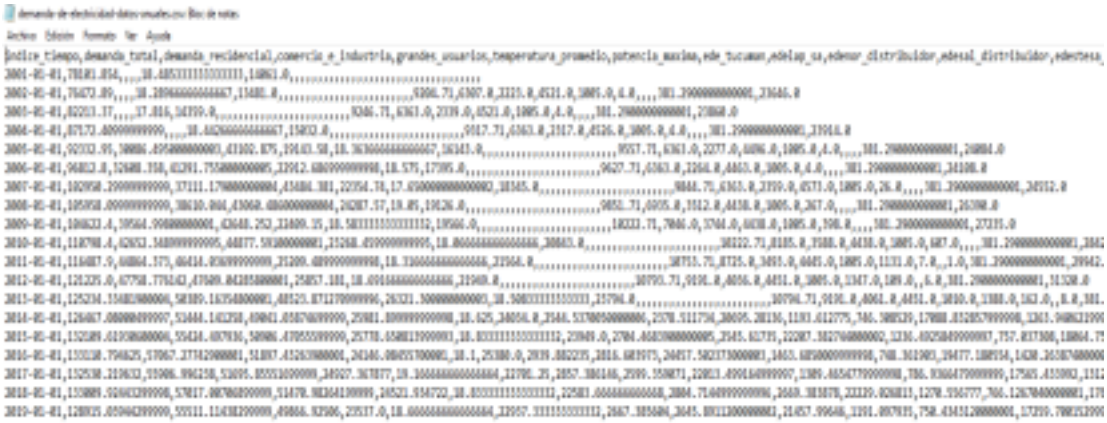
Python es un lenguaje de programación con diversas funciones. Spyder es el ambiente de Python para el análisis de datos y librerías (Pandas) que permite leer formatos de archivos tipo **xlsx(transformarlo a data frame) o CSV y muchas cosas más.**
En este caso usamos la librería llamada “Pandas”, para leer archivos tipo .CSV (dataset) y luego transformarlo a dataframe.

Para utilizar la librería Pandas, primero deben descargarla.

La diferencia de dataset y dataframe:

Dataset es un conjunto de datos y es una colección de datos habitualmente tabulados. Aquí una imagen de ejemplo:

(Aquí se está analizando la demanda de electricidad anual)



Dataframe es una" hoja de datos". Aquí una imagen (a través de Python/Spyder):

	indice_tiempo	demanda_residencial	edenor_distribuidor
13	2014-01-01	51444.141258	20695.281360
14	2015-01-01	55424.497936	22207.382744
15	2016-01-01	57067.277429	24457.502373
16	2017-01-01	55906.996238	22013.499165
17	2018-01-01	57017.007069	22229.026813
18	2019-01-01	55511.114383	21457.996460

Diferencias

Un dataset como un dataframe son conjuntos de datos organizados en forma de tabla o matriz**.** Lo que diferencia a un dataframe de un dataset es que un dataframe es un dataset que a la vez está organizado en columnas**,** de modo que en el data frame tendremos los datos estructurados y cada columna con su nombre correspondiente.

¿De dónde se extraen los datos o dataset?

Hay varias formas de hacerlo y varias plataformas, pero para este proyecto lo haré más simple, se descarga en la web de:

https://data.humdata.org/

el siguiente dataset:
https://data.humdata.org/dataset/covid-19-vaccinations/resource/146b9a31-ecc4-494c-9832-5788c4645fd8

Se analizará el dataset "Coronavirus (COVID-19) Vaccinations" que muestra la cantidad de dosis de vacunación contra el COVID-19 administradas por cada 100 personas dentro de una población determinada. Se deberá tener en cuenta que no mide el número total de personas que han sido vacunadas (que generalmente son dos dosis).

, y este dataset se comparará con el Monitoreo público de Vacunación del gobierno de Argentina:

https://www.argentina.gob.ar/coronavirus/vacuna/aplicadas

Importante: El dataset descargado, se actualizó la última vez el 20/10/2022. Y el Monitoreo público de Vacunación del gobierno de Argentina, la última actualización fue el 24/10/2022. Así que, el dataset está desactualizado, pero se podrá llegar a una estimación muy cercana

a la del 24/10/2022.

Código

#Detalla el paso a paso: #

```
#Esta es la libreria pandas
import pandas

#Cargamos el archivo csv (señalando/ubicacion)y lo guardanos en la variable llamada archivo_csv
archivo_csv=pandas.read_csv("C:\\Users\\Familia\\OneDrive\\Escritorio\\Metodologia\\data.csv")

#Convertimos a DataFrame   pandas.Dataframe(NombreVariableDondeGuardeCSV)
dataframe=pandas.DataFrame(archivo_csv)

# .info() nos dara informacion sobre el contenido de nuestro DataFrame
print(dataframe.info())

# Aqui tomamos las columnas que nos interesa de nuestro DataFrame.
columna_dataframe=["location", "total_vaccinations", "people_fully_vaccinated"]

# Aca decimos "de lo que yo te marque de columna_dataframe, del dataset dataframe, guardamelo en location"
location=dataframe[columna_dataframe]

# Aqui es como un indice, le decimos de la variable location dame lo que sea exactamente igual == a  "Argentina"
#y guardalo en arg_location
arg_location=location[location["location"]=="Argentina"]

#input
print(arg_location)
```

¿Cómo se sabe si el dataset está completo?

.info() nos muestra lo siguiente:Indice(0,1,2,etc), Columnas, filas, la cantidad de valores que contiene la fila y el tipo de valor.

```
RangeIndex: 134802 entries, 0 to 134801
Data columns (total 16 columns):
 #   Column                              Non-Null Count   Dtype
---  ------                              --------------   -----
 0   location                            134802 non-null  object
 1   iso_code                            134802 non-null  object
 2   date                                134802 non-null  object
 3   total_vaccinations                  65615 non-null   object
 4   people_vaccinated                   62822 non-null   float64
 5   people_fully_vaccinated             60126 non-null   float64
 6   total_boosters                      35807 non-null   float64
 7   daily_vaccinations_raw              54381 non-null   float64
 8   daily_vaccinations                  133871 non-null  object
 9   total_vaccinations_per_hundred      65614 non-null   float64
 10  people_vaccinated_per_hundred       62822 non-null   float64
 11  people_fully_vaccinated_per_hundred 60126 non-null   float64
 12  total_boosters_per_hundred          35807 non-null   float64
 13  daily_vaccinations_per_million      133870 non-null  float64
 14  daily_people_vaccinated             133406 non-null  float64
 15  daily_people_vaccinated_per_hundred 133406 non-null  float64
dtypes: float64(11), object(5)
memory usage: 16.5+ MB
None
```

Range Index son las filas, y 134.802 son la cantidad de filas con valor.

Ejemplo en la columna “location” --- 134.802 non-null, quiere decir que hay 134.802 valores

non-null, es decir no nulos, que contiene valor. Por ende, la columna “location” esta completa. Pero puede contener algunas que no, como por ejemplo “total vaccinations”

Se ejecuta:

	location	total_vaccinations	people_fully_vaccinated
4892	Argentina	20490	2.0
4893	Argentina	40592	7.0
4894	Argentina	43398	7.0
4895	Argentina	43525	7.0
4896	Argentina	46837	10.0
...	...	...	...
5550	Argentina	109788438	37828011.0
5551	Argentina	109801040	37829592.0
5552	Argentina	109813120	37831060.0
5553	Argentina	109825584	37832559.0
5554	Argentina	109827384	37832727.0

Monitoreo público de Vacunación

Resulta que el análisis del total de dosis aplicada, y vacunados con esquema completo, se acerca al monitoreo público de vacunación, se elaboró el análisis sabiendo que el dataset está atrasado, pero está acercándose al monitoreo de vacunación, y esto quiere decir que la información es verídica y está siendo actualizada.

<u>Lenguaje técnico a tener en cuenta:</u>

Range Index = = filas.

Non—null = = No nulo, es decir, contiene valor.

A las variables ponerle nombre relacionado al contenido.

Necesitamos primero instalar la librería y luego importarla con “import”.

info() nos permite ver las columnas y que tipo de datos son.

NAN-es el valor que se asigna a una celda cuando no tiene datos.

<u>Conclusión del proyecto</u>

En conclusión el proyecto que se presentó, demuestra que cualquier con conocimientos básicos puede llevarlo a cabo. Además de aportar una noción mayor a lo que simplemente se puede leer, aporta mayor flexibilidad y mayor capacidad para reaccionar a condiciones adversas. El proyecto puede ser mucho más amplio, y de seguir agregando temas relacionado con la programación y con el aprendizaje automático.

El resultado del análisis mostró que el Monitoreo público de Vacunación del gobierno de Argentina está correctamente actualizado y que los datos son congruentes. El proyecto puede ser mejorado y el código se puede escribir de mil maneras diferentes.