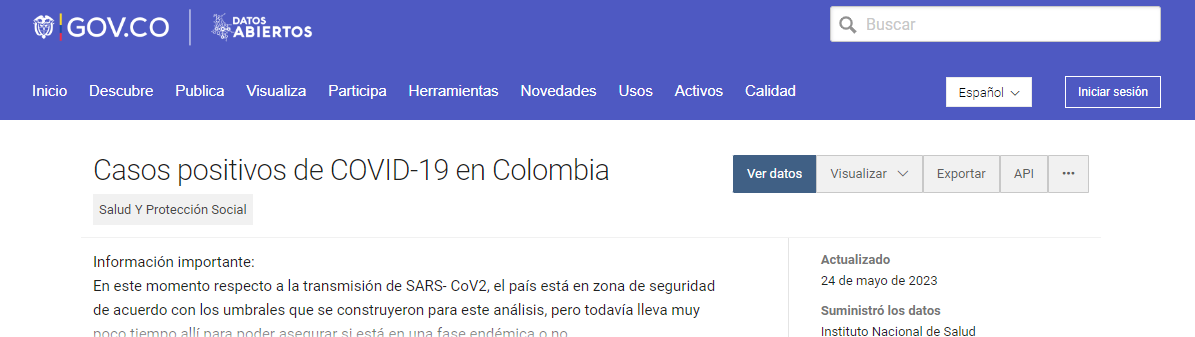
Sustentación de prueba técnica por Cesar Augusto Bedoya Escobar.

1. **Alcance del proyecto y captura de datos.**

Para esa etapa se plantea como alcance el medir el impacto social de los afectados por la pandemia generada por el covid 19, por medio de las personas más afectadas clínicamente y el poder generar propuestas de recuperación social y económica por aquellas zonas más afectadas. La idea no es aprovecharse de las muertes o dificultades generadas por la epidemia, por el contrario es poder generar valor y resiliencia a lo sucedido y qué mejor forma que fortalecer aquellos sectores más impactados.

Para esta tarea se da como tarea inicial el identificar a nivel Colombia, por ahora cuales son estas zonas y poder identificar las oportunidades que se puedan aprovechar desde el conjunto poblacional más afectado qué por factores de las restricciones de salud pública pudieran haber perdido su trabajo o forma de vida. Con la ayuda de la base de datos suministrada por el gobierno [gov.co]: https://www.datos.gov.co, fuente de datos que nos puede suministrar tanto la información de las fechas de los contagios en las fechas cuando se aplican más restricciones a la movilidad y la interacción social además de también mirar cuales son las áreas qué todavía presenta contagios y que se pueden aprovechar para mejorar o abordar planes de mejoramiento desde los medios virtuales y no presenciales.

Para descargar y visualizar los datos desde la página tenemos las opciones en la parte superior derecha y en preferencia para obtener los datos lo mejor es exportar como csv para excel, de esta forma podremos tener un formato más cómodo para su manipulación.



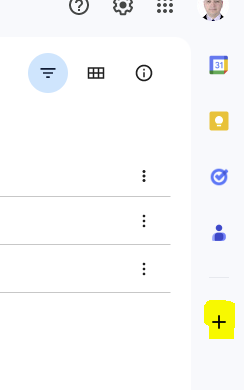
Opción Exportar, copiar url de enlace del botón [CSV para Excel]

https://www.datos.gov.co/api/views/gt2j-8ykr/rows.csv?accessType=DOWNLOAD&bom=true&format=true

1. **Explorar y evaluar los datos.**

Ya con la url de los datos, podemos trabajar de varias formas pero para nuestro ejercicio me enfocare en 2.

* La primera de ella es poder explorar la data desde python con la librería de pandas, aunque no siempre es la mejor alternativa por lo masivo de los datos, esta nos permite realizar manipulaciones pequeñas con los datos, entre estas manipulaciones podemos explorar la cantidad de datos, la estructura y composición de los datos para validar si contienen vacíos o variaciones y lo más importante el poder dar respuesta al alcance que nos planteamos, para lograr esta tarea nos apoyamos con de la herramienta de google colaboratory, extensión que es gratuita y nos brinda un interfaz de trabajo similar a Júpiter. Para obtener esta, desde una cuenta de correo de gmail buscamos el símbolo de + a la derecha de las herramientas, como se muestra en la ilustración.



y esta opción nos dará acceso a una nueva ventana donde podemos buscar extensiones donde con la ayuda de la barra de búsqueda podemos escribir colaboratory y nos instalará dicha extensión y poder crear y trabajar con un entorno de python sin la necesidad de instalarlo en nuestra máquina.

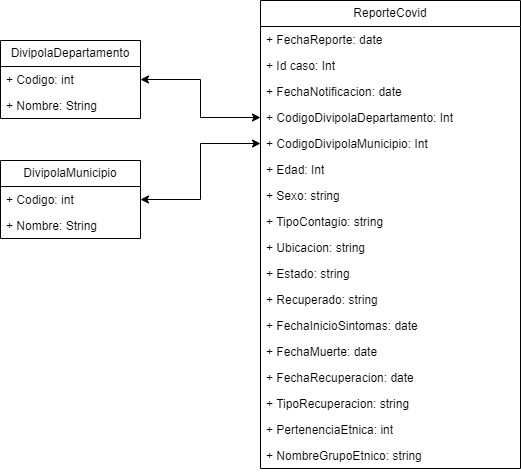
Con el área de trabajo ya disponible, se puede utilizar la url de los datos y con la instancia de pandas, poder cargar los datos y comenzar a trabajar con ellos. Entre los entregables dispuse de las consultas y algunas de las transformaciones que se realizaron con los datos para poder trabajar con la data.

Uno de los problemas que resaltan es el formato que tiene los números en especial para el Id\_Caso y el Codigo\_DIVIPOLA\_departamento, que presentan separadores de mil, sin embargo esto se puede corregir fácilmente

1. **Definir el modelo de datos.**

Ya con el acceso a los datos y una lectura de que podemos encontrar en ellos podemos abordar la idea de cómo organizarlos y de la construcción de un modelo que nos permita agilizar las consultas o futuras consultas de los mismos.

de esta información se creó el siguiente modelo de datos relacional que podemos crear.



De este se puede determinar que son muy pocos los campos que se pueden normalizar, sin embargo nos ayuda para caracterizar las zonas afectadas lo cual es útil para el alcance del estudio que planeamos realizar.

Con esto claro, se plantea crear una base de datos relacional y pensando que no se espera un volumen de información mayor a 10 millones de registros es ideal para almacenar la información y no tener que extraerla constantemente y para efectos de estar actualizando constantemente se puede planear incrementos delta de la data donde se consulte con intervalos de fechas de 1 meses e ir actualizando la información con el id\_caso.

pensando en esta tarea se plantea una tarea de ETL que periódicamente consulte este delta y los ingente en la base de datos.

Para dar respuesta a estas necesidades se abordó la idea de trabajar con una cuenta azure que me da la posibilidad de realizar ambas actividades, adicional que me permite almacenar la información para futuros reportes que se puedan generar.

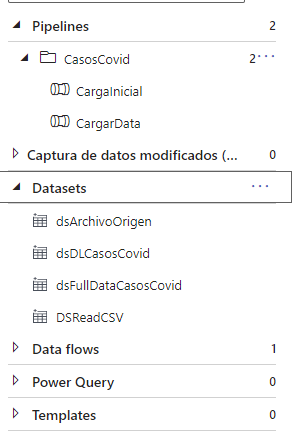
1. Ejecutar la ETL.

Ya definida la herramienta que es azure, utilizaremos la cuenta de azure de:

* usuario: [cesar7895123@hotmail.com](mailto:cesar7895123@hotmail.com) y contraseña: c3s4r7895123\*

Estas mismas claves aplican para la cuenta de github.

Donde se crearon 2 pipelines



Donde se creó 1 para la carga inicial y el otro para las cargas delta.

Todos los cambios se realizaron en la rama: dev\_azdfCasosCovid cuyos cambios no se lograron pasar a la main hasta no terminar con los ajustes.