Nombre: Andres Felipe Bello Zapata.

1. **Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.**

Desde el inicio de la maestría he sentido gran interés por implementar los conocimientos en el día a día, y este trabajo no fue la excepción. Esto hace parte de un proyecto en el que venia trabajando que consiste en tomar información de unas actividades que se deben desarrollar, con unas fechas específicas.

Anteriormente tenia que ingresar a la plataforma del trabajo, descargar las tareas, verificar cuales se había efectuado, verificar las fechas de cumplimiento, y notificar a las personas responsables de realizar algunas de ellas. Inicialmente realice el proceso de automatización de tal forma que solo tuviese que descargar la información, la cual puede ser exportada desde la pagina web a un Excel, y a través de un algoritmo, comparo todas las actividades con las que tengo ya almacenadas, de tal forma que el automáticamente identifique cuales están pendientes, cuantos días faltan y basado en la cantidad de días pendientes notifique a las personas responsables de las mismas, y a mi como administrador de las tareas.

Gracias a las lecturas de esta materia y a investigaciones independientes, en este momento el acceso a internet y la captura de la información se realiza de manera automática, siendo únicamente mi responsabilidad, el ejecutar un código, y magia!!!, se captura la información directamente de la web, se analizan las actividades pendientes y se reportan a los responsables con solo click.

2. **Definir un título para el dataset. Elegir un título que sea descriptivo**

En titulo escogido es **REFERENCIA_PLAN_ACCION**, en el siguiente punto realizo una explicación del motivo por el cual se escogió.

3. **Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).**

El dataset que obtengo a partir del web scraping, es un archivo de tipo csv, que sirve a mi algoritmo como referencia para identificar tareas nuevas, tareas eliminadas, estado de tareas, etc. Por ese motivo decidí dar el nombre al archivo de **REFERENCIA_PLAN_ACCION**. El Plan de Acción, básicamente hace referencia a un conjunto de actividades que se implementan en el transcurso del año, con un objetivo específico, consistente en mejorar los niveles de HSE (Healthy, Safety and Enviroment) en la organización en la que laboro. Tuve la intención de generar un data set usando PostgreSQL, sin embargo, el programa de revisión y notificación automática de la información ya lo tenia contemplado desde un archivo de tipo xls. Por tal motivo importé la biblioteca de pandas e implemente uno de sus métodos para exportar la información.

4. **Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente**

La representación gráfica del dataset la realizo en 3 tomas, considerando que la base de datos es bastante amplia. Agregué colores y demás con el fin de hacer un poco más atractiva la información, ya que el formato csv, no es muy agradable a la vista. Igualmente, la columna inicial y final, se crean únicamente porque es información que pasa por la web, sin embargo deben ser eliminados

posteriormente como parte del proceso de limpieza de datos. Algunos nombres también se identifican con una simbología extraña, debido principalmente al no reconocimiento de tildes y demás.

Toma 1

	X	ID	SEQUENCE_NUM	ACT_NAME	INICIO_PLAN	FIN_PLAN	TYPE_NAME	UNIDAD	PRIORIDAD	NSABLE_ACTI
0		98825	C-0421-2020	Verificar el im	9/21/2020	12/15/2020	Plan Accion	CACOM-5	Alta	Abdon Estiber
1		98826	H-0422-2020	Verificar el im	1/1/2020	6/20/2020	Plan Accion	CACOM-5	Alta	Abdon Estiber
2		98827	H-0423-2020	Verificar el im	6/21/2020	12/15/2020	Plan Accion	CACOM-5	Alta	Abdon Estiber
3		98828	L-0424-2020	Realizar una re	1/1/2020	6/20/2020	Plan Accion	CACOM-5	Alta	Ricardo Augus
4		98833	L-0429-2020	Realizar un bal	6/21/2020	12/15/2020	Plan Accion	CACOM-5	Alta	Ricardo Augus
5		98831	S-0427-2020	Realizar una au	1/1/2020	6/20/2020	Plan Accion	CACOM-5	Alta	Andres Felipe
6		98832	S-0428-2020	Verificar el cur	6/21/2020	12/15/2020	Plan Accion	CACOM-5	Alta	Andres Felipe
7		98834	M-0430-2020	A travÃ©s de l	1/1/2020	12/15/2020	Plan Accion	CACOM-5	Alta	Ricardo Augus
8		98835	C-0431-2020	Realizar una in	1/1/2020	3/20/2020	Plan Accion	CACOM-5	Alta	Abdon Estiber
9		98836	C-0432-2020	Verificar el efe	3/21/2020	6/20/2020	Plan Accion	CACOM-5	Alta	Abdon Estiber
10		98837	C-0433-2020	Realizar una in	6/21/2020	9/20/2020	Plan Accion	CACOM-5	Alta	Abdon Estiber
11		101213	A-0718-2020	EN REFERENCIA	1/7/2020	6/30/2020	Plan Accion	CACOM-5	Alta	Victor Alfonso
12		98864	A-0460-2020	Realizar un anÁ	1/1/2020	6/20/2020	Plan Accion	CACOM-5	Alta	Juan Guillerm
13		98865	A-0461-2020	Realizar un anÁ	6/21/2020	12/15/2020	Plan Accion	CACOM-5	Alta	Juan Guillerm
14		98866	G-0462-2020	Dictar una capa	1/1/2020	12/15/2020	Plan Accion	CACOM-5	Alta	Juan Guillerm
15		98867	N-0463-2020	Realizar reuniÁ	1/1/2020	12/15/2020	Plan Accion	CACOM-5	Alta	Juan Guillerm
16		98868	N-0464-2020	CapacitaciÃ³n	1/1/2020	12/15/2020	Plan Accion	CACOM-5	Alta	Juan Guillerm
17		98869	O-0465-2020	Realizar una au	1/1/2020	6/20/2020	Plan Accion	CACOM-5	Alta	Juan Guillerm
18		98870	O-0466-2020	Verificar el cur	6/21/2020	12/15/2020	Plan Accion	CACOM-5	Alta	Juan Guillerm

Toma 2

ESTADO_ACTIVIDAD	OR_ACTIVI	PROGRESO_ACTIVIDAD	PORCENTAJE_ACTIVIDAD	PROGRAMA_PREVENCION	PLAN_ID	PLAN_NAME	PLAN_YEAR
No Iniciada	0	0	0	C. Programa de Confiabilidad	48	Plan de AcciÃ³	2020
No Iniciada	0	0	0	H. Programa de PrevenciÃ³n d	48	Plan de AcciÃ³	2020
No Iniciada	0	0	0	H. Programa de PrevenciÃ³n d	48	Plan de AcciÃ³	2020
No Iniciada	0	0	0	L. Programa de PrevenciÃ³n d	48	Plan de AcciÃ³	2020
No Iniciada	0	0	0	L. Programa de PrevenciÃ³n d	48	Plan de AcciÃ³	2020
No Iniciada	0	0	0	S. Programa de AuditorÃ-as e	48	Plan de AcciÃ³	2020
No Iniciada	0	0	0	S. Programa de AuditorÃ-as e	48	Plan de AcciÃ³	2020
No Iniciada	0	0	0	M. Programa de PrevenciÃ³n d	48	Plan de AcciÃ³	2020
Cumplida	100	100	100	C. Programa de Confiabilidad	48	Plan de AcciÃ³	2020
No Iniciada	0	0	0	C. Programa de Confiabilidad	48	Plan de AcciÃ³	2020
No Iniciada	0	0	0	C. Programa de Confiabilidad	48	Plan de AcciÃ³	2020
No Iniciada	0	0	0	A. Programa de PrevenciÃ³n e	48	Plan de AcciÃ³	2020
No Iniciada	0	0	0	A. Programa de PrevenciÃ³n e	48	Plan de AcciÃ³	2020
No Iniciada	0	0	0	A. Programa de PrevenciÃ³n e	48	Plan de AcciÃ³	2020
No Iniciada	0	0	0	G. Programa de PrevenciÃ³n d	48	Plan de AcciÃ³	2020
No Iniciada	0	0	0	N. Programa de PrevenciÃ³n d	48	Plan de AcciÃ³	2020
No Iniciada	0	0	0	N. Programa de PrevenciÃ³n d	48	Plan de AcciÃ³	2020
No Iniciada	0	0	0	O. Programa de PrevenciÃ³n d	48	Plan de AcciÃ³	2020
No Iniciada	0	0	0	O. Programa de PrevenciÃ³n d	48	Plan de AcciÃ³	2020

Toma 3

PLAN_TYPE_NAM	PLAN_STATUS	PROGRAM_NAME	PRIORIDAD_ACTIVIDAD	ASOCIADO_A	PROGRAMA	Y
DESAES	No Legalizado	C. Programa de Conf	Alta	Plan de acciÃ³n	PREVENCI	
DESAES	No Legalizado	H. Programa de Prev	Alta	Plan de acciÃ³n	PREVENCI	
DESAES	No Legalizado	H. Programa de Prev	Alta	Plan de acciÃ³n	PREVENCI	
DESAES	No Legalizado	L. Programa de Preve	Alta	Plan de acciÃ³n	PREVENCI	
DESAES	No Legalizado	L. Programa de Preve	Alta	Plan de acciÃ³n	PREVENCI	
DESAES	No Legalizado	S. Programa de Audi	Alta	Plan de acciÃ³n	PREVENCI	
DESAES	No Legalizado	S. Programa de Audi	Alta	Plan de acciÃ³n	PREVENCI	
DESAES	No Legalizado	M. Programa de Prev	Alta	Plan de acciÃ³n	PREVENCI	
DESAES	No Legalizado	C. Programa de Conf	Alta	Plan de acciÃ³n	PREVENCI	
DESAES	No Legalizado	C. Programa de Conf	Alta	Plan de acciÃ³n	PREVENCI	
DESAES	No Legalizado	C. Programa de Conf	Alta	Plan de acciÃ³n	PREVENCI	
DESAES	No Legalizado	A. Programa de Prev	Alta	SRV	PREVENCI	
DESAES	No Legalizado	A. Programa de Prev	Alta	Plan de acciÃ³n	PREVENCI	
DESAES	No Legalizado	A. Programa de Prev	Alta	Plan de acciÃ³n	PREVENCI	
DESAES	No Legalizado	G. Programa de Prev	Alta	Plan de acciÃ³n	PREVENCI	
DESAES	No Legalizado	N. Programa de Prev	Alta	Plan de acciÃ³n	PREVENCI	
DESAES	No Legalizado	N. Programa de Prev	Alta	Plan de acciÃ³n	PREVENCI	
DESAES	No Legalizado	O. Programa de Prev	Alta	Plan de acciÃ³n	PREVENCI	
DESAES	No Legalizado	O. Programa de Prev	Alta	Plan de acciÃ³n	PREVENCI	

NOTA: Solo se muestran unas filas de la base de datos, ya que la descarga contiene alrededor de 50 filas.

5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

A continuación se describe los campos que contiene el dataset:

- **X:** No tiene significado alguno, se genera debido a que al hacer el web-scraping, la tabla tomada retorna dicho valor, y si no coinciden el numero de columnas con la información que se va a ingresar, la biblioteca pandas retorna un error.
- **ID:** Representa un número de referencia para la actividad.
- **SEQUENCE_NUM:** Número de referencia de la actividad. Este número suele usarse por el personal responsable de las actividades como parámetro de búsqueda en el sistema.
- **ACT_NAME:** Realiza una descripción de la tarea a desarrollar.
- **INICIO_PLAN:** Fecha a partir de la cual se tiene la tarea disponible en el sistema. No todas las fechas de las actividades son iguales ya que algunas son cargadas durante el año por razones de identificación de planes de mejora.
- **FIN_PLAN:** Fecha límite para ejecutar la actividad. Esta fecha la usa como referencia el algoritmo para saber en que momento debe enviar los correos.
- **TYPE_NAME:** Hace referencia al origen de la actividad. Puede ser parte del programa, o ser generada como adicional
- **UNIDAD:** Hace referencia a la Unidad responsable de la actividad, para este caso es la misma en todo el dataset, ya que el filtro se realizo en el momento en el cual se genero le dataset.
- **PRIORIDAD:** Hace referencia a la prioridad, para la ejecución de la actividad. Este campo es generado por la gerencia.
- **RESPONSABLE_ACTIVIDAD:** Este campo nombra a la persona responsable de la actividad. Este campo sirve al algoritmo para determinar a quien se le debe enviar el correo.

- **ESTADO_ACTIVIDAD:** El estado de la actividad informa el estado actual de la misma. El algoritmo utiliza ese campo para no tener en cuenta las cumplidas.
- **VALOR_ACTIVIDAD:** Esta columna no tengo conocimiento de su objetivo, ya que es usada por la gerencia para estadísticas internas.
- **PROGRESO_ACTIVIDAD:** Este campo es diligenciado por el responsable, en donde indica el estado de avance. Para efectos de calificación, solo se tienen en cuenta valores absolutos, o se hizo completo o no se ha efectuado.
- **PORCENTAJE_ACTIVIDAD:** Esta columna no tengo conocimiento de su objetivo, ya que es usada por la gerencia para estadísticas internas.
- **PROGRAMA_PREVENCION:** Este campo indica el programa que se esta atacando con la actividad. Es de gran utilidad para realizar análisis estadísticos de la información.
- **PLAN_ID:** Esta columna no tengo conocimiento de su objetivo, ya que es usada por la gerencia para estadísticas internas.
- **PLAN_NAME:** Esta columna no tengo conocimiento de su objetivo, ya que es usada por la gerencia para estadísticas internas.
- **PLAN_YEAR:** Determina el año del plan. Esta casilla también es filtrada dentro del proceso de web scraping, motivo por el cual es igual en todos los valores.
- **PLAN_TYPE_NAME:** Esta columna no tengo conocimiento de su objetivo, ya que es usada por la gerencia para estadísticas internas.
- **PLAN_STATUS:** Esta columna no tengo conocimiento de su objetivo, ya que es usada por la gerencia para estadísticas internas.
- **PROGRAM_NAME:** Se repite esta casilla, hace referencia a la misma información suministrada en la casilla **PROGRAMA_PREVENCION**.
- **PRIORIDAD_ACTIVIDAD:** Esta columna no tengo conocimiento de su objetivo, ya que es usada por la gerencia para estadísticas internas.
- **ASOCIADO_A:** Esta columna no tengo conocimiento de su objetivo, ya que es usada por la gerencia para estadísticas internas.
- **PROGRAMA:** Esta columna no tengo conocimiento de su objetivo, ya que es usada por la gerencia para estadísticas internas.
- **Y:** No tiene significado alguno, se genera debido a que al hacer el web-scraping, la tabla tomada retorna dicho valor, y si no coinciden el numero de columnas con la información que se va a ingresar, la biblioteca pandas retorna un error.

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

Agradecimientos a las siguientes fuentes de información:

Web:

- Stackoverflow
- Udemy: “Curso Moden Web Scraping With Python Using Scrapy…”

Textos:

- Texto Web Scraping del modulo (Laia Subirats Maté/ Mireia Calvo González)
- Web Scraping (Richard Lawson)

Resumen:

- Resumen personal de información que he encontrado en la web

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder

El conjunto de datos es interesante, particularmente para mi trabajo, ya que dicha información integra actividades que promueven el desarrollo de HSE en la organización. La aplicación del web scraping en este contexto permite realizar el proceso de recolección del la información de manera automática, reduciendo carga de trabajo en el personal de la sección, y garantizando que se realice de manera correcta, considerando que se han presentado oportunidades en las que se ingresan filtros alterados.

Mediante este proceso no pretendo responder una pregunta en especial, pero si pretendo hacer ver a la gerencia, la importancia de tener una cultura orientada al dato en la organización. Tanto para optimizar procesos, como para analizar información y, en fin; un gran sin número de herramientas que trae consigo este campo. De la mano de esa finalidad, esta la de mejorar el proceso de revisión, control y notificación de actividades relacionadas con la sección.

8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección

Para esta base de datos se implementó la biblioteca de pandas, la cual permite realizar diferentes tipos de operaciones con los datos. Además de permitir integrar la información con bibliotecas para generar gráficas (matplotlib) y de aprendizaje (sklearn).