**Estructura de los códigos en la consolidación y definición del marco muestral**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Lenguaje de programación:** Python

**Paquetes requeridos:**

1. Pandas
2. Numpy

El proyecto fue desarrollado en 3 scripts, los cuales se dividen en:

1. **funciones.py**

En este script se encuentran las funciones definidas dentro de los procesos repetitivos para el cruce de cada uno de los registros administrativos relacionados al sector transporte. Las funciones establecidas son:

df\_estandar: esta función permite filtrar las columnas de interés en el proceso de filtrar cruzar bases de datos.

encontrar\_posicion:

deff\_direct: esta función permite agregar la información de la procedencia de los directorios que se están cruzando.

renombrar: renombra las columnas recibiendo como parámetro el diccionario de los índices en que se relacionan las columnas con las columnas de interés.

cruce: con esta función se realiza el cruce de la base de datos, teniendo presente las funciones anteriores para consolidar la base final.

imputacion\_ingresos: esta función realiza la imputación a la variable de ingresos por medio de la distancia mínima, utilizando la variable de personal ocupado.

ocupados: la aplicación de esta función fue útil para registrar las variables de personal ocupado.

1. **Cruce.py**

Inicialmente en este script se realiza la lectura de la ruta de la carpeta en la que se encuentran los registros administrativos que se desean cruzar.

De esta manera, se realiza la lectura de todos los archivos que se encuentren en la carpeta. Como resultado, se obtiene un diccionario de la subcarpeta y del nombre del archivo, por ejemplo: {Supertransporte: Directorio\_Supertransporte.xlsx}.

Seguidamente se realiza la lectura de la estructura en que se desea mantener la base final, esta estructura se encuentra registrada de la siguiente manera: Consolidad\_directorio.xlsx

Una vez importados los archivos se procede a realizar el cruce entre todos los archivos como sigue el ejemplo a continuación:

En el caso de cruzar la DIG con el archivo del fest tenemos:

r\_DIG = os.path.join(r\_directorios, #Ruta del archivo

                        n\_directorios[0],

                        sub\_carpetas[n\_directorios[0]][3])

df\_DIG = pd.read\_excel(r\_DIG, sheet\_name= 0, #Lectura del archivo

                          decimal = ',')

indic = {0:4, 1:5, 2:7, 3:8, 4:15, 6:0, 7:2, 8:13, 9:14, 11:17, 12:16,

13:9, 14:10, 15:11, 16:12, 17:26, 18:23, 20:25} #Parametros relacionados

df\_consol = cruce(df\_fest\_f, df\_DIG, names\_obj, indic, sub\_carpetas[n\_directorios[0]][3]) #Cruce de la información

Este proceso fue iterado en todas las bases que se desean cruzar, teniendo presente las particularidades en algunas de estas.

Finalmente, se realiza una caracterización para identificar que cantidad de empresas esta abarcando el consolidado final de directorios respecto a los directorios que se están cruzando, esto se realiza para el fest y para la DIG que es el registro administrativo que más aporta información. Además, se proporciona un archivo que contine la información de la cantidad de duplicados que se encontraron, pero no se conservaron en el consolidado final, ya que este esta conformado por empresas únicas, es decir no hay duplicados. El insumo de duplicados es con fines de evaluar si se puede estar perdiendo información valiosa para futuros procesos.

1. **Muestreo.py**

En cuanto al muestreo, se realiza la lectura de la base consolidada y se actualiza respecto a la información más actualizada, posteriormente, se aplican los criterios de inclusión y exclusión para su operación para finalmente establecer el marco final.

Como insumos en este proceso se obtiene el marco final que presenta la cantidad de empresas que se contiene por CIIU principal para las actividades económicas de interés. Además, se proporciona un archivo .xlsx que contine la información desagregada por cada CIIU, con fines de evaluar y caracterizar algunas estadísticas de interés para el diseño muestral.