Codebook

En este codebook se indica el proceso realizado, la información utilizada así como aquellos cambios realizados a las variables con el objetivo de que se pueda realizar una investigación reproducible en un futuro.

- ## Datos utilizados

Para la realización del proyecto se han obtenido los datos del Instituto Nacional de Estadística (INE), el enlace de obtención es el siguiente:
https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176952&menu=resultados&secc=1254736195203&idp=1254735572981#!tabs-1254736195203

En concreto los datos los tenemos en formato .csv, en la carpeta data tenemos los datos de los conjuntos, las direcciones son las siguientes:

data/datos_hogares_2020/ECHHogares2020.csv
data/datos_personas_2020/ECHPersonas2020.csv

Luego tenemos los diccionarios donde se detalla la información sobre el nombre de las variables, que significan y la forma de codificación que tienen. Son archivos de tipo .xlsx que también se encuentran en la carpeta data del proyecto con direcciones:

data/datos_hogares_2020/dr_ECHHogares_2020.xlsx
data/datos_personas_2020/dr_ECHPersonas_2020.xlsx

- ## Unión de datasets

Hemos trabajado con un dataset que constituía la unión de el dataset de hogares y personas, el método utilizado es mediante un full join con las keys {ID_VIV, PERIODO, TAMANO, IDQ_PV, CA, FACCAL}

- ## Cambio de nombres de variables

Para mayor claridad, cambiamos el nombre de variables una vez están unidos los conjuntos de datos.

El formato de cambio es el siguiente:
EJEMPLO
{NUEVONOMBRE ; VIEJONOMBRE}

NUEVOS NOMBRES

{ PROVINCIA ; IDQ_PV,
F_CONS_EDI ; FEDI,
PERS_HOGAR ; TAMTOHO,
NAC_HO ; NACHO,
PAIS_NACIM ; PNACIMT,
EDAD_LLEG_ESP ; EDADFLLEG,
TPO_LLEG_ESP ; TPOFLLEG,
PAIS_NAC ; PNACT,
PAIS_NAC_NACIM ; PNACNACIMT,
TPO_NAC_ESP ; TPOFNACESP,
PAIS_NACIM_PADRE ; PNACIMPADRET,
PAIS_NACIM_MADRE ; PNACIMMADRET,
RELAC_MIEMBROS ; P01,
ACTIVIDAD ; RELACT}

- Variables eliminadas y creadas

Las variables que hemos decidido omitir a la hora de implementar el proceso de limpieza del dataset son P2,P3….P19, el motivo es el alto porcentaje de NA's que presentan (En el apartado gráfico se ha realizado una gráfica donde si se utiliza desde P1, P2…P13).

También se eliminan las variables ANONEDI y ANEDI por que apenas aportan información respecto a la variable F_CONS_EDI.

Para el dataframe que utilizamos como ejemplo de conjunto de datos que se podría utilizar para modelizar hemos creado variables dummy para aquellas variables con un formato de codificación erróneo, hemos creado las variables dummy de las siguientes variables: (REGVI, NAC_HO, SEXO, EC, NACIM, COCINA, NAC, NACNACIMESP, NACIMPADRE, NACIMMADRE, OCUPA, PAREJA, SEXOPAR, NACPAR, ECPAR, HIJOSDEAMBOS).

- Tipos de variables

Hemos utilizado dos dataframes distintos durante la realización del proyecto, uno donde se transformaban muchas variables que aparecen como tipo integer a character y luego de character a factor, en el otro también convertimos algunas variables de integer a character y de character a factor pero la mayoría de las variables de tipo integer, permanecen así.

En los diccionarios se indica en la columna **Tipo**, el formato inicial que tienen las variables, **A** corresponden a columnas de tipo integer en r mientras que **N** son las variables de tipo numérico.

Para convertir las variables de integer a character hemos utilizado los diccionarios a los que hacemos referencia anteriormente, nosotros hemos hecho este proceso en los chunks *variables de hogar a carácter* y *df personas* en el código recogido en el archivo .rmd del proyecto, se podría copiar y pegar directamente.

Al final de los dos dataframes con los que trabajamos realizan las siguientes conversiones:

df ← ***Primero a caracter y posteriormente a factor*** (CA, PROVINCIA , TAMANO, COCINA, REGVI, TIPOVIV, FEDI, TIPOHO, NAC_HO, NUCLEOFAM, SEXO, EC, NACIM, NAC, NACNACIMESP, RELAC_MIEMBROS, ESTUDIOS, ACTIVIDAD , OCUPA, SITUHO, SITUHO_D, PAREJA, SEXOPAR, NACPAR, ECPAR, HIJOSDEAMBOS, SITUNUCLEOFAM, NACIMPADRE, NACIMMADRE, PAIS_NACIM, PAIS_NAC, PAIS_NAC_NACIM , PAIS_NACIM_PADRE, PAIS_NACIM_MADRE)

Tipo fecha (PERIODO)

Tipo numérico (METROSVI)

df_num ← ***Primero a caracter y posteriormente a factor*** (CA, PROVINCIA , TIPOHO, RELAC_MIEMBROS, SITUHO, SITUHO_D, PAIS_NACIM, PAIS_NAC, PAIS_NAC_NACIM , PAIS_NACIM_PADRE, PAIS_NACIM_MADRE)

Tipo fecha (PERIODO)

Conversión a dummy (REGVI, NAC_HO, SEXO, EC, NACIM, COCINA, NAC, NACNACIMESP, NACIMPADRE, NACIMMADRE, OCUPA, PAREJA, SEXOPAR, NACPAR, ECPAR, HIJOSDEAMBOS).

- Imputación de valores NA

Se han creado nuevas categorías para las variables con valores ausentes para el dataframe df, la imputación ha sido mediante categorías tipo texto que luego se convierten a factores al hacer la conversión de character a factor.

Un ejemplo de cómo se van a establecer las condiciones és el siguiente:

VARIABLE: ***si condición/es,*** “Valor Imputado”

Aquí se enlista como se ha realizado la imputación en cada caso:

RELAC_MIEMBROS: ***si NA***, "Miembro encuestado"

NHIJOME_NUCLEO: ***si PAREJA == "No convive en pareja" y NA***, "Sin hijos/Sin pareja

NHIJO_NUCLEO: ***si PAREJA == "No convive en pareja" y NA***, "Sin hijos/Sin pareja

SITUNUCLEOFAM: ***si PAREJA == "No convive en pareja" y NA***, "Sin hijos/Sin pareja

NHIJOMENOR: ***si PAREJA == "No convive en pareja" y NA***, "Sin hijos/Sin pareja

NHIJOMENOR:***si PAREJA == "No convive en pareja" y NA***, "Sin hijos/Sin pareja

NUCLEOFAM: ***si PAREJA == "No convive en pareja" y NA***, "Sin pareja”

NUCLEOFAM: ***si PAREJA == "Convive con cónyuge de distinto sexo"y NA***, "Pareja casada con o sin hijos, con o sin otras personas"

NUCLEOFAM: ***si PAREJA == "Convive con pareja de hecho de distinto sexo" y NA***, "Pareja de hecho con o sin hijos, con o sin otras personas"
NUCLEOFAM: ***si PAREJA == "Convive con pareja de hecho del mismo sexo" y NA***, "Pareja de hecho con o sin hijos, con o sin otras personas"

NUCLEOFAM: ***si PAREJA == "Convive con cónyuge del mismo sexo" y NA***, "Pareja casada con o sin hijos, con o sin otras personas"

ECPAR: ***si NA***, "Sin pareja"

NACPAR: ***si NA***, "Sin pareja"

HIJOSDEAMBOS: ***si NA***, "Sin pareja"

SEXOPAR: ***si NA,*** "Sin pareja"

NHIJOPAR: ***si HIJOSDEAMBOS == "Conviviendo sin hijos" y NA***, 0

NHIJOPAR: ***si HIJOSDEAMBOS == "Conviviendo con hijos todos comunes" y NA***, NHIJO_NUCLEO

NHIJOPAR: ***si HIJOSDEAMBOS == "Conviviendo con hijos no comunes" y NA***, NHIJO_NUCLEO

NHIJOPAR: ***si HIJOSDEAMBOS ==*"Sin pareja" y NA**, "Sin pareja"

NHIJO***: si NHIJO_NUCLEO == "Sin hijos/Sin pareja" y NA***, "Sin hijos/Sin pareja"

NHIJO: ***si NA,*** NHIJO_NUCLEO

ACTIVIDAD: ***si NA***, "No procede"

ESTUDIOS: ***si NA,*** "Menor de 16 años"

OCUPA: ***si NA,*** "No trabaja"

EDAD_LLEG_ESP = ifelse(is.na(EDADFLLEG), "Nacido en España", EDADFLLEG),

TPO_LLEG_ESP = ifelse(is.na(TPOFLLEG), "Nacido en España", TPOFLLEG),

TPO_NAC_ESP : ***si NAC == "Española" y NA***, "Nacionalidad española de nacimiento"

TPO_NAC_ESP : ***si NAC == "No tiene nacionalidad española" y NA***, "No tiene nacionalidad española"

TPO_NAC_ESP : ***si NAC == "Española y otras" y NA***, "Nacionalidad española de nacimiento"

NACNACIMESP: ***si NA***, "No tiene nacionalidad española"

- Valores eliminados

Aparecen en las variables PAIS_NACIM_PADRE, PAIS_NACIM_MADRE y PAIS_NACIM unos valores que el INE tuvo que codificar incorrectamente como países, aquellos con un valor de “966” y “555”, se han eliminado.

- Anomalías

En el conjunto de datos hemos detectado una serie de datos anómalos, para eliminarlos hemos aplicado el siguiente filtro:

DORMITORIOS <= 8
ASEOS <= 4
TRASTEROS <= 4
COMEDORES <= 3
HABVI <= 10
OTRASHAB <= 2

- Outliers

Hemos buscado outliers en las variables METROSVI y DENSIDADVI. Para METROSVI hemos considerado sólo aquellos valores que cumplen la regla 3sigma mientras que para DENSIDADVI sólo hemos considerado aquellos valores que entran dentro del intervalo del método Boxplot.