Entendimiento de los datos:

Como son tantas variables, solo se van a analizar aquellas que seleccionamos para la pregunta analítica. Se analizará completitud, validez y unicidad.

Para la unicidad, se observa que no hay ninguna fila repetida, es decir, hay unicidad:

```
[8]  duplicated_rows = df.loc[df.duplicated(keep=False)].shape[0]
```

```
duplicated_rows
```

```
0
```

Validez y completitud de las variables seleccionadas:

NPCFP14C: enfermedades mentales o de comportamiento

```
[6]  print(df['NPCFP14C'].value_counts())
```

```
NPCFP14C
2.0    11593
1.0      201
Name: count, dtype: int64
```

```
print(df['NPCFP14C'].isnull().sum())
```

```
1
```

Se observan 11794 registros, de los cuales apenas 1 es nulo (es decir, prácticamente completitud del 100%). Por su parte, según el diccionario, los únicos valores posibles para esta variable son 1 y 2, por lo que también hay validez.

NPCFP14J: enfermedades nutricionales o del metabolismo

```
[11]  print(df['NPCFP14J'].value_counts())
```

```
NPCFP14J
2.0    10840
1.0      954
Name: count, dtype: int64
```

```
print(df['NPCFP14J'].isnull().sum())
```

```
1
```

Se observan 11794 registros, de los cuales apenas 1 es nulo (es decir, prácticamente completitud del 100%). Por su parte, según el diccionario, los únicos valores posibles para esta variable son 1 y 2, por lo que también hay validez.

NPCFP13A: consultas generales por año

```
[14] print(df['NPCFP13A'].value_counts())

NPCFP13A
1.0    7230
2.0    4564
Name: count, dtype: int64

[15] print(df['NPCFP13A'].isnull().sum())

1
```

Se observan 11794 registros, de los cuales apenas 1 es nulo (es decir, prácticamente completitud del 100%). Por su parte, según el diccionario, los únicos valores posibles para esta variable son 1 y 2, por lo que también hay validez.

NPCFP13F: ir a consulta psicológica en el último año

```
[17] print(df['NPCFP13F'].value_counts())

NPCFP13F
2.0    11243
1.0      551
Name: count, dtype: int64

print(df['NPCFP13F'].isnull().sum())

1
```

Se observan 11794 registros, de los cuales apenas 1 es nulo (es decir, prácticamente completitud del 100%). Por su parte, según el diccionario, los únicos valores posibles para esta variable son 1 y 2, por lo que también hay validez.

NPCFP36: Actividad física semanal

```
[20] print(df['NPCFP36'].value_counts())

NPCFP36
4.0    5157
1.0    2090
2.0    2039
3.0    1209
Name: count, dtype: int64

print(df['NPCFP36'].isnull().sum())

1300
```

Se observan 10495 registros, es decir, acá sí hay incompletitud, decidimos reemplazar esos valores nulos por el promedio de los registros ya que esta variable hace referencia a la actividad física. Por su parte, según el diccionario, los únicos valores posibles para esta variable son enteros de 1 a 4, por lo que también hay validez.

NPCFP38: Si ha fumado en los últimos 30 días

```
[23] print(df['NPCFP38'].value_counts())
```

```
NPCFP38
3.0    9564
1.0     536
2.0     395
Name: count, dtype: int64
```

```
print(df['NPCFP38'].isnull().sum())
```

```
1300
```

La incompletitud es idéntica a la variable anterior, pero acá se decide reemplazar los valores nulos con 3 debido a que hace referencia a que la gente no hay fumado en los últimos 30 días. Es la suposición más segura de las 3 opciones.

NOMBRE_ESTRATO:

```
[37] print(df['NOMBRE_ESTRATO'].value_counts())
```

```
NOMBRE_ESTRATO
Patio Bonito                    681
Arborizadora                    506
Las Margaritas                  321
San Francisco                   265
Tintal Sur                      256
                               ...
Ciudad Salitre Oriental          10
Localidad Usaquen resto           9
Localidad San Cristobal resto     9
Localidad Ciudad Bolivar resto    5
Granjas de Techo                  1
Name: count, Length: 99, dtype: int64
```

```
print(df['NOMBRE_ESTRATO'].isnull().sum())
```

```
0
```

```
[38] print(df['NOMBRE_ESTRATO'].describe())
```

```
count            11795
unique              99
top       Patio Bonito
freq               681
Name: NOMBRE_ESTRATO, dtype: object
```

No existen valores nulos, o sea, hay completitud, el estrato más popular es “Patio bonito”.

NPCHP34: nivel educativo

```
[41] print(df['NPCHP34'].value_counts())
```

```
NPCHP34
4.0     49
2.0     31
1.0     22
3.0     11
6.0     10
99.0     7
5.0      4
7.0      3
8.0      3
10.0     2
9.0      1
Name: count, dtype: int64
```

```
print(df['NPCHP34'].isnull().sum())
```

```
11652
```

Para esta variable hay apenas 143 registros (de casi 12.000). Lo que hicimos fue otorgarles el valor correspondiente a la moda (4), que significa “Toda la secundaria”.

NPCFP1: afiliado a EPS

```
[44] print(df['NPCFP1'].value_counts())
```

```
NPCFP1
1.0    10698
2.0     1023
9.0       73
Name: count, dtype: int64
```

```
print(df['NPCFP1'].isnull().sum())
```

```
1
```

En este caso se observa un error en la validez, ya que los posibles valores de esta variable son 1, 2 y 3. Lo que hicimos fue reemplazar los valores de 9 por 3, ya que asumimos que hay un typo por no haber ningún registron con valor 3. Apenas hay un valor nulo, le dimos el valor de la moda.

NHCCPCTRL2: cuántas personas componen el hogar

```
[48] print(df['NHCCPCTRL2'].describe())

count    11795.000000
mean         3.488512
std          1.486278
min          1.000000
25%          2.000000
50%          3.000000
75%          4.000000
max         11.000000
Name: NHCCPCTRL2, dtype: float64
```

```
print(df['NHCCPCTRL2'].value_counts())

NHCCPCTRL2
4     3267
3     3014
2     2291
5     1484
1      816
6      544
7      208
8      106
10      28
11      19
9       18
Name: count, dtype: int64
```

En promedio, un hogar está conformado por 3.48 personas. Hay completitud debido a que no hay ningún valor nulo. Tiene sentido también que el hogar esté conformado por al menos una persona.