In [1]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt


#Importacion de los datasets
#Dataset import

vehiculos = pd.read_csv('E:/Análisis de datos/DATASETS/BCN accidentes/Accidentes vehiculos

personas = pd.read_csv('E:/Análisis de datos/DATASETS/BCN accidentes/Accidents personas TO

tipos = pd.read_csv('E:/Análisis de datos/DATASETS/BCN accidentes/Accidents tipos TOTAL.cs

cantidades = pd.read_csv('E:/Análisis de datos/DATASETS/BCN accidentes/Accidents cantidad

causas= pd.read_csv('E:/Análisis de datos/DATASETS/BCN accidentes/Accidentes causas TOTAL.


#Revisamos uno de los datasets
#We check one of the datasets
vehiculos.head()
```

```
E:\Anaconda\lib\site-packages\IPython\core\interactiveshell.py:3444: DtypeWarning: Columns
(1,3,5,10,11,25,26,27,28,29) have mixed types.Specify dtype option on import or set low_me
mory=False.
  exec(code_obj, self.user_global_ns, self.user_ns)
E:\Anaconda\lib\site-packages\IPython\core\interactiveshell.py:3444: DtypeWarning: Columns
(10) have mixed types.Specify dtype option on import or set low_memory=False.
  exec(code_obj, self.user_global_ns, self.user_ns)
E:\Anaconda\lib\site-packages\IPython\core\interactiveshell.py:3444: DtypeWarning: Columns
(22,23) have mixed types.Specify dtype option on import or set low_memory=False.
  exec(code_obj, self.user_global_ns, self.user_ns)
```

Out[1]:

	Codi expedient	Codi districte	Nom districte	Codi barri	Nom barri	Codi carrer	Nom Carrer	Nom Carrer 2	Num postal caption	Descripció dia setmana	...
0	2017S007932,1.0,Ciutat Vella,4.0,"Sant Pere, S...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
1	2017S007932,1.0,Ciutat Vella,4.0,"Sant Pere, S...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
2	2017S007933	1.0	Ciutat Vella	3.0	La Barceloneta	224904.0	Joan Borbó Comte Barcelona	NaN	0044 0045	Dilluns	...
3	2017S007933	1.0	Ciutat Vella	3.0	La Barceloneta	224904.0	Joan Borbó Comte Barcelona	NaN	0044 0045	Dilluns	...
4	2017S007933	1.0	Ciutat Vella	3.0	La Barceloneta	224904.0	Joan Borbó Comte Barcelona	NaN	0044 0045	Dilluns	...

5 rows × 29 columns

In [2]:

```
#Revisamos las columnas de un dataset
```

In [ ]:
```
#We check its columns
vehiculos.columns
```

Out[2]:
```
Index(['Codi expedient', 'Codi districte', 'Nom districte', 'Codi barri',
       'Nom barri', 'Codi carrer', 'Nom Carrer', 'Nom Carrer 2',
       'Num postal caption', 'Descripció dia setmana', 'Dia setmana',
       'Descripció tipus dia', 'NK Any', 'Mes de any', 'Nom mes', 'Dia de mes',
       'Hora de dia', 'Data Completa', 'Descripció causa vianant',
       'Descripció tipus de vehicle', 'Descripció model',
       'Descripció model MAYUS', 'Descripció model simplificat',
       'Descripció marca', 'Descripció color', 'Descripció carnet',
       'Antiguitat carnet', 'Coordenada UTM (Y)', 'Coordenada UTM (X)'],
      dtype='object')
```

In [3]:
```
#Revisamos las columnas de otro de los datasets.
#We also check the columns of the next dataset.
personas.columns
```

Out[3]:
```
Index(['Codi expedient', 'Codi districte', 'Nom districte', 'Codi barri',
       'Nom barri', 'Codi carrer', 'Nom carrer', 'Nom Carrer 2',
       'Num postal caption', 'Descripció dia setmana', 'Dia setmana',
       'Descripció tipus dia', 'NK Any', 'Mes de any', 'Nom mes', 'Dia de mes',
       'Descripció torn', 'Hora de dia', 'Data completa',
       'Descripció causa vianant', 'Desc. Tipus vehicle implicat',
       'Descripció sexe', 'Descripció tipus persona', 'Edat',
       'Descripció victimització', 'Altre informació', 'Coordenada UTM (Y)',
       'Coordenada UTM (X)', 'Longitud', 'Latitud'],
      dtype='object')
```

In [4]:
```
#Eliminamos columnas que coinciden con el primer dataset.
#We will be deleting the columns that match this dataset with the previous one.
personas = personas.drop(columns=['Codi districte', 'Nom districte', 'Codi barri',
       'Nom barri', 'Codi carrer', 'Nom carrer', 'Nom Carrer 2',
       'Num postal caption', 'Descripció dia setmana', 'Dia setmana',
       'Descripció tipus dia', 'NK Any', 'Mes de any', 'Nom mes', 'Dia de mes',
       'Descripció torn', 'Hora de dia', 'Data completa',
       'Descripció causa vianant', 'Coordenada UTM (Y)',
       'Coordenada UTM (X)', 'Longitud', 'Latitud'])
```

In [5]:
```
#Volvemos a revisar:
#We check the columns again:
personas.columns
```

Out[5]:
```
Index(['Codi expedient', 'Desc. Tipus vehicle implicat', 'Descripció sexe',
       'Descripció tipus persona', 'Edat', 'Descripció victimització',
       'Altre informació'],
      dtype='object')
```

In [6]:
```
#Haremos lo mismo con el resto de datasets.
#We will be doing the same with the rest of the datasets:
tipos.columns
```

Out[6]:
```
Index(['Codi expedient', 'Codi districte', 'Nom districte', 'Codi barri',
       'Nom barri', 'Codi carrer', 'Nom carrer', 'Nom Carrer 2',
       'Num postal caption', 'Descripció dia setmana', 'Dia setmana',
       'Descripció tipus dia', 'NK Any', 'Mes de any', 'Nom mes', 'Dia de mes',
       'Data completa', 'Hora de dia', 'Descripció torn',
       'Descripció tipus accident', 'Coordenada UTM (Y)', 'Coordenada UTM (X)',
       'Longitud', 'Latitud'],
      dtype='object')
```

In [7]:

```
tipos = tipos.drop(columns=['Codi districte', 'Nom districte', 'Codi barri',
       'Nom barri', 'Codi carrer', 'Nom carrer', 'Nom Carrer 2',
       'Num postal caption', 'Descripció dia setmana', 'Dia setmana',
       'Descripció tipus dia', 'NK Any', 'Mes de any', 'Nom mes', 'Dia de mes',
       'Data completa', 'Hora de dia', 'Descripció torn','Coordenada UTM (Y)', 'Coordenada
       'Longitud', 'Latitud'])
```

In [8]:
```
tipos.columns
```

Out[8]:
```
Index(['Codi expedient', 'Descripció tipus accident'], dtype='object')
```

In [9]:
```
cantidades.columns
```

Out[9]:
```
Index(['Codi expedient', 'Codi districte', 'Nom districte', 'NK barri',
       'Nom barri', 'Codi carrer', 'Nom carrer', 'Nom carrer 2',
       'Num postal caption', 'Descripció dia setmana', 'Dia de setmana',
       'Descripció tipus dia', 'NK Any', 'Mes de any', 'Nom mes', 'Dia de mes',
       'Data completa', 'Hora de dia', 'Descripció torn',
       'Descripció causa vianant', 'Número de morts',
       'Número de lesionats lleus', 'Número de lesionats greus',
       'Número de víctimes', 'Número de vehicles implicats',
       'Coordenada UTM (Y)', 'Coordenada UTM (X)', 'Longitud', 'Latitud'],
      dtype='object')
```

In [10]:
```
cantidades = cantidades.drop(columns=['Codi districte', 'Nom districte', 'NK barri',
       'Nom barri', 'Codi carrer', 'Nom carrer', 'Nom carrer 2',
       'Num postal caption', 'Descripció dia setmana', 'Dia de setmana',
       'Descripció tipus dia', 'NK Any', 'Mes de any', 'Nom mes', 'Dia de mes',
       'Data completa', 'Hora de dia', 'Descripció torn',
       'Descripció causa vianant','Coordenada UTM (Y)', 'Coordenada UTM (X)', 'Longitud',
```

In [11]:
```
cantidades.columns
```

Out[11]:
```
Index(['Codi expedient', 'Número de morts', 'Número de lesionats lleus',
       'Número de lesionats greus', 'Número de víctimes',
       'Número de vehicles implicats'],
      dtype='object')
```

In [12]:
```
causas.columns
```

Out[12]:
```
Index(['Codi expedient', 'Codi districte', 'Nom districte', 'Codi barri',
       'Nom barri', 'Codi carrer', 'Nom carrer', 'Nom carrer 2',
       'Num postal caption', 'Descripció dia setmana', 'Dia setmana',
       'Descripció tipus dia', 'NK Any', 'Mes de any', 'Nom mes', 'Dia de mes',
       'Data Completa', 'Hora de dia', 'Descripció torn',
       'Descripció causa mediata', 'Coordenada UTM (Y)', 'Coordenada UTM (X)',
       'Longitud', 'Latitud'],
      dtype='object')
```

In [13]:
```
causas = causas.drop(columns=['Codi districte', 'Nom districte', 'Codi barri',
       'Nom barri', 'Codi carrer', 'Nom carrer', 'Nom carrer 2',
       'Num postal caption', 'Descripció dia setmana', 'Dia setmana',
       'Descripció tipus dia', 'NK Any', 'Mes de any', 'Nom mes', 'Dia de mes',
       'Data Completa', 'Hora de dia', 'Descripció torn',
      'Coordenada UTM (Y)', 'Coordenada UTM (X)',
       'Longitud', 'Latitud'])
```

In [14]:
```
causas.columns
```

Out[14]:
```
Index(['Codi expedient', 'Descripció causa mediata'], dtype='object')
```

In [15]:
```
#Unimos los datasets
#We join the datasets.
accidents = pd.merge(personas,vehiculos, how='inner', on=["Codi expedient"])
```

In [16]:
```
accidents = pd.merge(accidents,tipos, how='inner', on=["Codi expedient"])
```

In [17]:
```
accidents = pd.merge(accidents,cantidades, how='inner', on=["Codi expedient"])
```

In [18]:
```
accidents = pd.merge(accidents,causas, how='inner', on=["Codi expedient"])
```

In [19]:
```
#Revisamos el nuevo dataset creado a partir de los joins anteriores
#We check the new dataset create by joining the previous ones.
accidents.columns
```

Out[19]:
```
Index(['Codi expedient', 'Desc. Tipus vehicle implicat', 'Descripció sexe',
       'Descripció tipus persona', 'Edat', 'Descripció victimització',
       'Altre informació', 'Codi districte', 'Nom districte', 'Codi barri',
       'Nom barri', 'Codi carrer', 'Nom Carrer', 'Nom Carrer 2',
       'Num postal caption', 'Descripció dia setmana', 'Dia setmana',
       'Descripció tipus dia', 'NK Any', 'Mes de any', 'Nom mes', 'Dia de mes',
       'Hora de dia', 'Data Completa', 'Descripció causa vianant',
       'Descripció tipus de vehicle', 'Descripció model',
       'Descripció model MAYUS', 'Descripció model simplificat',
       'Descripció marca', 'Descripció color', 'Descripció carnet',
       'Antiguitat carnet', 'Coordenada UTM (Y)', 'Coordenada UTM (X)',
       'Descripció tipus accident', 'Número de morts',
       'Número de lesionats lleus', 'Número de lesionats greus',
       'Número de víctimes', 'Número de vehicles implicats',
       'Descripció causa mediata'],
      dtype='object')
```

In [20]:
```
accidents.head()
```

Out[20]:

	Codi expedient	Desc. Tipus vehicle implicat	Descripció sexe	Descripció tipus persona	Edat	Descripció victimització	Altre informació	Codi districte	Nom districte	Codi barri	...	A
0	2010S000001	Motocicleta	Home	Conductor	30.0	Ferit Lleu	NaN	2.0	Eixample	7.0	...	
1	2010S000001	Motocicleta	Home	Conductor	30.0	Ferit Lleu	NaN	2.0	Eixample	7.0	...	
2	2010S000001	Motocicleta	Home	Conductor	30.0	Ferit Lleu	NaN	2.0	Eixample	7.0	...	
3	2010S000001	Motocicleta	Home	Conductor	30.0	Ferit Lleu	NaN	2.0	Eixample	7.0	...	

	Codi expedient	Desc. Tipus vehicle implicat	Descripció sexe	Descripció tipus persona	Edat	Descripció victimització	Altre informació	Codi districte	Nom districte	Codi barri	...	A
4	2010S000001	Motocicleta	Home	Conductor	30.0	Ferit Lleu	NaN	2.0	Eixample	7.0	...	

5 rows × 42 columns

In [21]:
```
accidents.dtypes
```

Out[21]:
```
Codi expedient                   object
Desc. Tipus vehicle implicat     object
Descripció sexe                  object
Descripció tipus persona         object
Edat                            float64
Descripció victimització         object
Altre informació                 object
Codi districte                  float64
Nom districte                    object
Codi barri                      float64
Nom barri                        object
Codi carrer                     float64
Nom Carrer                       object
Nom Carrer 2                     object
Num postal caption               object
Descripció dia setmana           object
Dia setmana                      object
Descripció tipus dia             object
NK Any                          float64
Mes de any                      float64
Nom mes                          object
Dia de mes                      float64
Hora de dia                      object
Data Completa                    object
Descripció causa vianant         object
Descripció tipus de vehicle      object
Descripció model                 object
Descripció model MAYUS           object
Descripció model simplificat     object
Descripció marca                 object
Descripció color                 object
Descripció carnet                object
Antiguitat carnet               float64
Coordenada UTM (Y)              float64
Coordenada UTM (X)              float64
Descripció tipus accident        object
Número de morts                 float64
Número de lesionats lleus       float64
Número de lesionats greus       float64
Número de víctimes              float64
Número de vehicles implicats    float64
Descripció causa mediata         object
dtype: object
```

In [22]:
```
#Cambiamos los nombres de las columnas añadiendo _ paras que sea más sencillo manipularlas
#Column name changes to include _

accidents.columns = accidents.columns.str.replace(' ','_')
accidents.dtypes
```

Out[22]:
```
Codi_expedient                  object
Desc._Tipus_vehicle_implicat    object
Descripció_sexe                 object
Descripció_tipus_persona        object
Edat                           float64
Descripció_victimització        object
Altre_informació                object
Codi_districte                 float64
Nom_districte                   object
Codi_barri                     float64
Nom_barri                       object
Codi_carrer                    float64
Nom_Carrer                      object
Nom_Carrer_2                    object
Num_postal_caption              object
Descripció_dia_setmana          object
Dia_setmana                     object
Descripció_tipus_dia            object
NK_Any                         float64
Mes_de_any                     float64
Nom_mes                         object
Dia_de_mes                     float64
Hora_de_dia                     object
Data_Completa                   object
Descripció_causa_vianant        object
Descripció_tipus_de_vehicle     object
Descripció_model                object
Descripció_model_MAYUS          object
Descripció_model_simplificat    object
Descripció_marca                object
Descripció_color                object
Descripció_carnet               object
Antiguitat_carnet              float64
Coordenada_UTM_(Y)             float64
Coordenada_UTM_(X)             float64
Descripció_tipus_accident       object
Número_de_morts                float64
Número_de_lesionats_lleus      float64
Número_de_lesionats_greus      float64
Número_de_víctimes             float64
Número_de_vehicles_implicats   float64
Descripció_causa_mediata        object
dtype: object
```

In [23]:
```
#Cambiamos NA por 0 para poder cambiar el tipo de datos de las columnas después,
#ya que necesitamos que algunos de ellos sean int para hacer calculos posteriores

#We change NA's to 0, to be able to change the data type for some of the columns,
#since we need to make some calculations with them later:

accidents = accidents.fillna(0)
```

In [24]:
```
accidents.Edat= accidents.Edat.astype('int')
```

In [25]:
```
accidents.NK_Any= accidents.NK_Any.astype('int')
```

In [26]:
```
accidents.Mes_de_any= accidents.Mes_de_any.astype('int')
```

In [27]:
```
accidents.Dia_de_mes= accidents.Dia_de_mes.astype('int')
```

In [28]:
```
accidents.Codi_districte= accidents.Codi_districte.astype('int')
```

In [29]:
```
accidents.Codi_barri= accidents.Codi_barri.astype('int')
```

In [30]:
```
accidents.Codi_carrer= accidents.Codi_carrer.astype('int')
```

In [31]:
```
accidents.dtypes
```

Out[31]:
```
Codi_expedient                    object
Desc._Tipus_vehicle_implicat      object
Descripció_sexe                   object
Descripció_tipus_persona          object
Edat                               int32
Descripció_victimització          object
Altre_informació                  object
Codi_districte                     int32
Nom_districte                     object
Codi_barri                         int32
Nom_barri                         object
Codi_carrer                        int32
Nom_Carrer                        object
Nom_Carrer_2                      object
Num_postal_caption                object
Descripció_dia_setmana            object
Dia_setmana                       object
Descripció_tipus_dia              object
NK_Any                             int32
Mes_de_any                         int32
Nom_mes                           object
Dia_de_mes                         int32
Hora_de_dia                       object
Data_Completa                     object
Descripció_causa_vianant          object
Descripció_tipus_de_vehicle       object
Descripció_model                  object
Descripció_model_MAYUS            object
Descripció_model_simplificat      object
Descripció_marca                  object
Descripció_color                  object
Descripció_carnet                 object
Antiguitat_carnet                float64
Coordenada_UTM_(Y)               float64
Coordenada_UTM_(X)               float64
Descripció_tipus_accident         object
Número_de_morts                  float64
Número_de_lesionats_lleus        float64
Número_de_lesionats_greus        float64
Número_de_víctimes               float64
Número_de_vehicles_implicats     float64
Descripció_causa_mediata          object
dtype: object
```

In [32]:
```
#En caso de que queramos eliminar los valores nulls según un umbral determinado:
#In case we wanted to delete some of the null values

total = accidents.isnull().sum().sort_values(ascending=False)
percent = (accidents.isnull().sum()/accidents.isnull().count()).sort_values(ascending=Fals
missing_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
missing_data[missing_data['Total'] > 0]
```

Out[32]:

	Total	Percent

In [33]:

```
#Sin embargo en este caso no funcionará, ya que ya hemos cambiado los valores,
#y porque los valores no están dentro de los márgenes establecidos en el siguiente código.

#In this case it won't work, we have already changed the values to 0 and also,
#there would not be values inside the margins set below:

accidents = accidents[missing_data[missing_data['Percent'] < 0.15].index]

accidents
```

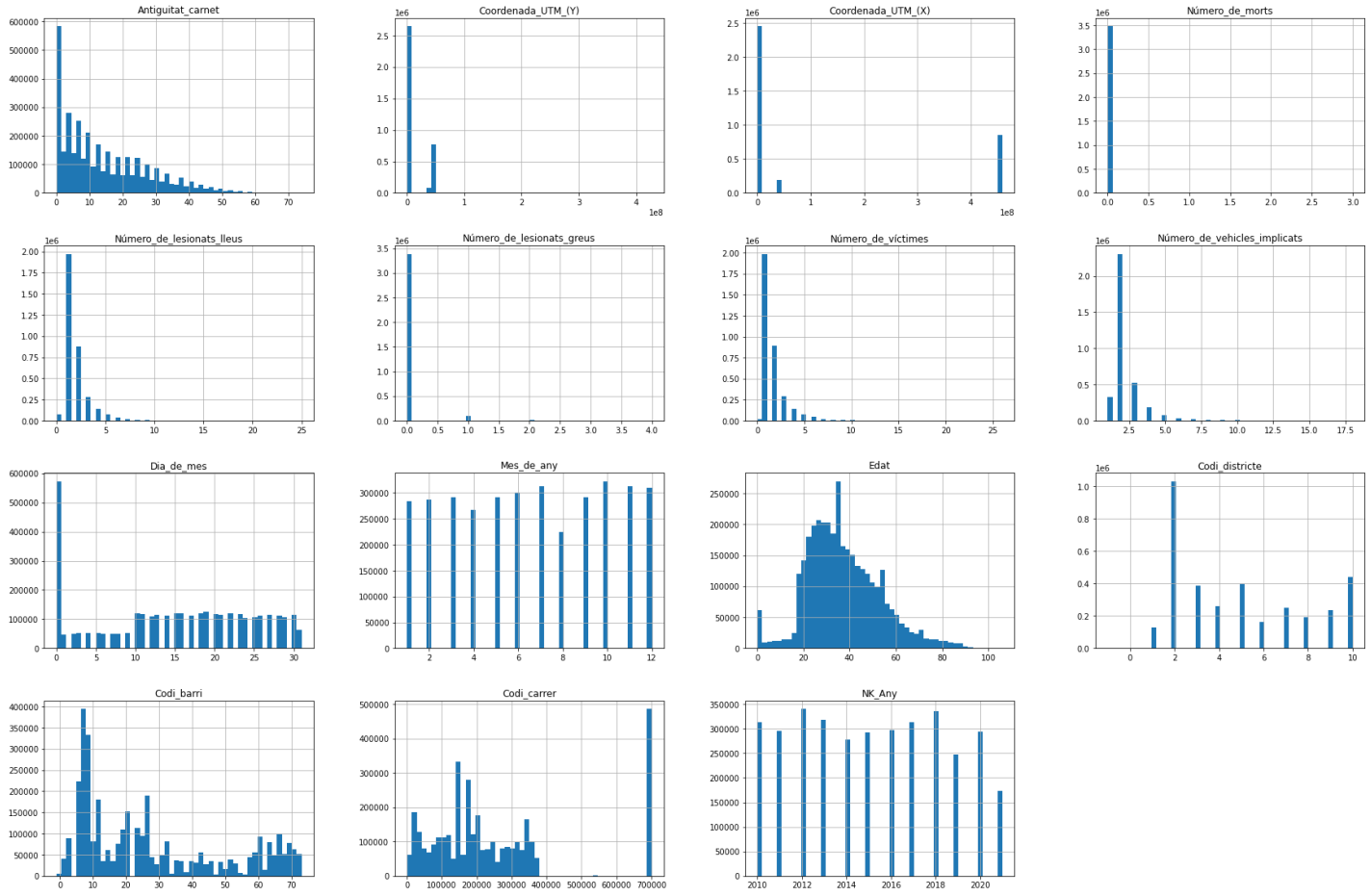
Out[33]:

	Codi_expedient	Descripció_carnet	Data_Completa	Descripció_causa_vianant	Descripció_tipus_de_vehicle	De
0	2010S000001	A	01/01/2010	No és causa del vianant	Motocicleta	
1	2010S000001	A	01/01/2010	No és causa del vianant	Motocicleta	
2	2010S000001	A	01/01/2010	No és causa del vianant	Motocicleta	
3	2010S000001	A	01/01/2010	No és causa del vianant	Motocicleta	
4	2010S000001	A	01/01/2010	No és causa del vianant	Motocicleta	
...	...	...	...	...	...	
3497425	2021S000615	B	06/02/2021	No és causa del vianant	Turisme	
3497426	2021S000615	B	06/02/2021	No és causa del vianant	Turisme	
3497427	2021S000615	B	06/02/2021	No és causa del vianant	Turisme	
3497428	2021S000615	B	06/02/2021	No és causa del vianant	Turisme	
3497429	2021S000615	B	06/02/2021	No és causa del vianant	Turisme	

3497430 rows × 42 columns

In [34]:

```
#Realizamos una visualización rápida de los valores numéricos:
#Quick viz of all numeric values:

accidents.hist(bins=50, figsize=(30,20));
```

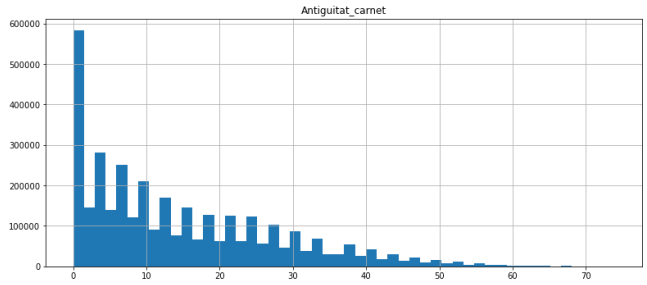


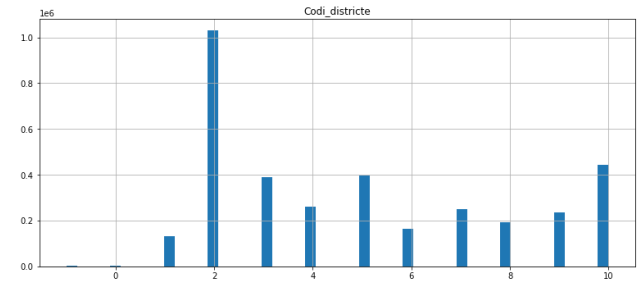
In [35]:

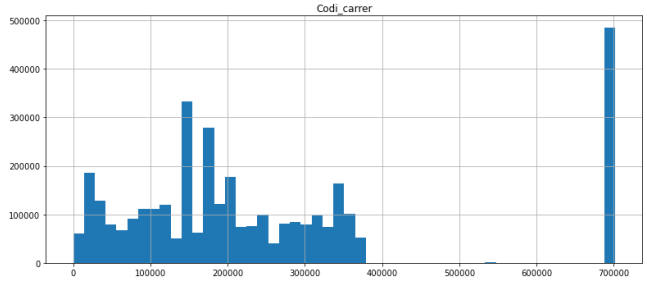
```
#Podemos visualizar solo algunas de las columnas:
#Viz for some of the columns:
accidents.hist(bins=50, figsize=(30,20), column=["Antiguitat_carnet", "Codi_districte", "C
```

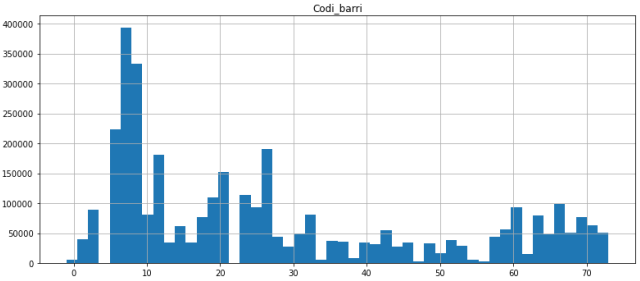
Out[35]:

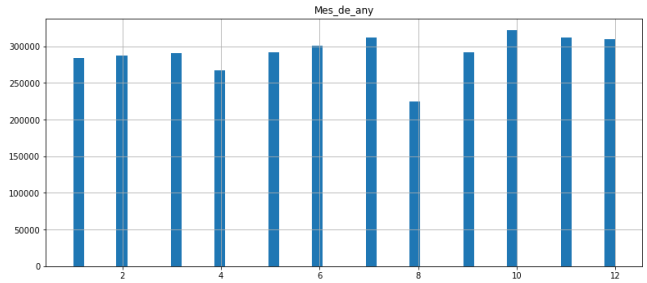
```
array([[<AxesSubplot:title={'center':'Antiguitat_carnet'}>,
        <AxesSubplot:title={'center':'Codi_districte'}>],
       [<AxesSubplot:title={'center':'Codi_carrer'}>,
        <AxesSubplot:title={'center':'Codi_barri'}>],
       [<AxesSubplot:title={'center':'Mes_de_any'}>,
        <AxesSubplot:title={'center':'NK_Any'}>]], dtype=object)
```

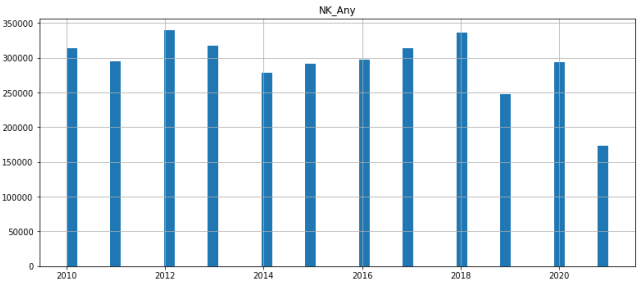












In [36]:

```
#Revisamos medias, ds, min, max, quantiles:
#Average, median, min, max and quantiles:
accidents.describe().T
```

Out[36]:

	count	mean	std	min	25%	50%	75%	
Antiguitat_carnet	3497430.0	1.440352e+01	1.308552e+01	0.0	4.00	11.00	23.00	
Coordenada_UTM_(Y)	3497430.0	1.302998e+07	1.709693e+07	-1.0	4579483.36	4583438.47	4589063.89	4
Coordenada_UTM_(X)	3497430.0	1.157936e+08	1.955084e+08	-1.0	429973.96	433495.24	45853754.00	4
Número_de_morts	3497430.0	4.689157e-03	7.275931e-02	0.0	0.00	0.00	0.00	
Número_de_lesionats_lleus	3497430.0	1.767213e+00	1.438525e+00	0.0	1.00	1.00	2.00	
Número_de_lesionats_greus	3497430.0	3.569021e-02	2.138575e-01	0.0	0.00	0.00	0.00	
Número_de_víctimes	3497430.0	1.807592e+00	1.459981e+00	0.0	1.00	1.00	2.00	
Número_de_vehicles_implicats	3497430.0	2.352155e+00	1.121876e+00	1.0	2.00	2.00	2.00	
Dia_de_mes	3497430.0	1.490283e+01	9.829759e+00	0.0	6.00	16.00	23.00	
Mes_de_any	3497430.0	6.592169e+00	3.482338e+00	1.0	4.00	7.00	10.00	
Edat	3497430.0	3.722514e+01	1.570051e+01	0.0	26.00	35.00	47.00	
Codi_districte	3497430.0	4.916233e+00	2.978809e+00	-1.0	2.00	4.00	7.00	
Codi_barri	3497430.0	2.757680e+01	2.206358e+01	-1.0	8.00	21.00	43.00	
Codi_carrer	3497430.0	2.523259e+05	2.044247e+05	-1.0	115603.00	191204.00	323203.00	
NK_Any	3497430.0	2.015216e+03	3.344676e+00	2010.0	2012.00	2015.00	2018.00	

In [37]:

```
accidents.describe()
```

Out[37]:

	Antiguitat_carnet	Coordenada_UTM_(Y)	Coordenada_UTM_(X)	Número_de_morts	Número_de_lesionats_lleus
count	3.497430e+06	3.497430e+06	3.497430e+06	3.497430e+06	3.497430e+06
mean	1.440352e+01	1.302998e+07	1.157936e+08	4.689157e-03	1.767213e+00
std	1.308552e+01	1.709693e+07	1.955084e+08	7.275931e-02	1.438525e+00
min	0.000000e+00	-1.000000e+00	-1.000000e+00	0.000000e+00	0.000000e+00
25%	4.000000e+00	4.579483e+06	4.299740e+05	0.000000e+00	1.000000e+00
50%	1.100000e+01	4.583438e+06	4.334952e+05	0.000000e+00	1.000000e+00
75%	2.300000e+01	4.589064e+06	4.585375e+07	0.000000e+00	2.000000e+00
max	7.400000e+01	4.275191e+08	4.591069e+08	3.000000e+00	2.500000e+01

In [38]:

```
#También podemos hacerlo para algun quantil determinado, etc:
#Showing only a quantile:
accidents.quantile(0.25)
```

Out[38]:

```
Antiguitat_carnet                    4.00
Coordenada_UTM_(Y)             4579483.36
Coordenada_UTM_(X)              429973.96
Número_de_morts                      0.00
Número_de_lesionats_lleus            1.00
Número_de_lesionats_greus            0.00
Número_de_víctimes                   1.00
Número_de_vehicles_implicats         2.00
Dia_de_mes                           6.00
Mes_de_any                           4.00
Edat                                26.00
Codi_districte                       2.00
Codi_barri                           8.00
Codi_carrer                     115603.00
NK_Any                            2012.00
Name: 0.25, dtype: float64
```

In [39]:

```
#O podemos revisar solo una de las columnas:
#Checking only one of the columns:

accidents["Antiguitat_carnet"].describe()
```

Out[39]:

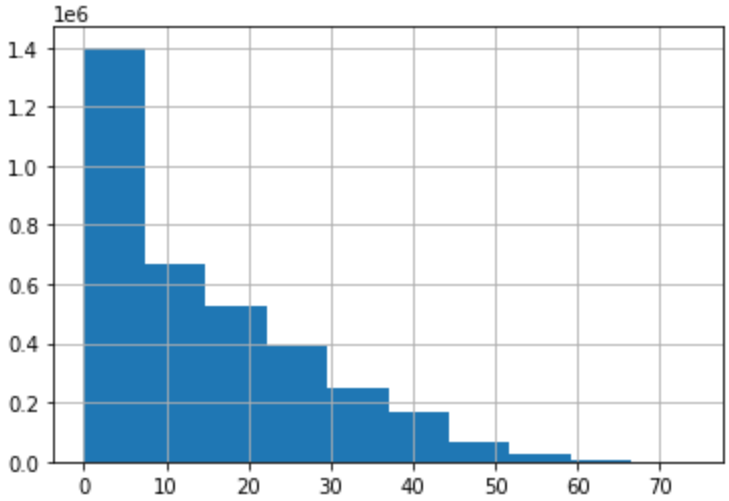
```
count    3.497430e+06
mean     1.440352e+01
std      1.308552e+01
min      0.000000e+00
25%      4.000000e+00
50%      1.100000e+01
75%      2.300000e+01
max      7.400000e+01
Name: Antiguitat_carnet, dtype: float64
```

In [40]:

```
accidents["Antiguitat_carnet"].hist()
```

Out[40]:

```
<AxesSubplot:>
```



In [42]:

```
#Cantidad de expedientes (únicos) con víctimas:
#Total amount of expedients (unique) with victims:
Victimas = accidents.groupby("Número_de_víctimes").Codi_expedient.nunique()
print(Victimas)
```

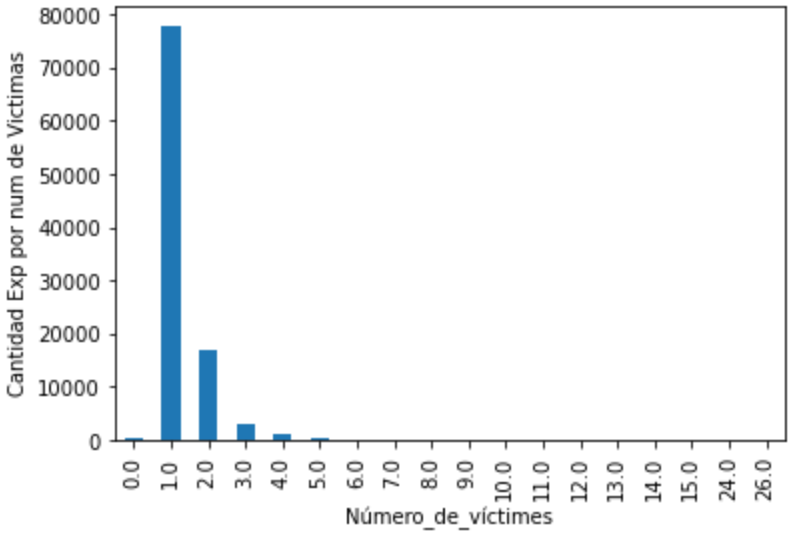
```
Número_de_víctimes
0.0       529
1.0     77865
2.0     16962
3.0      3025
4.0      1016
5.0       374
6.0       175
7.0        68
8.0        37
9.0        20
10.0       11
11.0        7
12.0        7
13.0        2
14.0        1
15.0        1
24.0        1
26.0        1
Name: Codi_expedient, dtype: int64
```

In [43]:

```
#Plot de las victimas calculadas antes
#Plot for victims/count of unique exp

ax = Victimas.plot(kind = 'bar')
ax.set_ylabel('Cantidad Exp por num de Victimas')
```

Out[43]:

```
Text(0, 0.5, 'Cantidad Exp por num de Victimas')
```



In [44]:

```
#Cantidad de victimas por año:
#Totl amount of victims per year:

accidents.groupby('NK_Any')['Número_de_víctimes'].agg(['sum'])
```

Out[44]:

NK_Any	sum
2010	571489.0
2011	530892.0
2012	680791.0
2013	602196.0
2014	502690.0
2015	525448.0
2016	547150.0
2017	542569.0
2018	641865.0
2019	468870.0
2020	411580.0
2021	296388.0

In [45]:

```
#Subset, filtro con muertos y graves menores de 30 años
#Subset and filter with dead and sever injured under 30 years of age:

Morts_Greus_Edat = accidents[(accidents.Descripció_victimització == "Mort") |
                             (accidents.Descripció_victimització == "Ferit Greu") & (accid
Morts_Greus_Edat.head()
```

Out[45]:

	Codi_expedient	Descripció_carnet	Data_Completa	Descripció_causa_vianant	Descripció_tipus_de_vehicle	Descri
1542	2010S000047	A	04/01/2010	No és causa del vianant	Motocicleta	
1543	2010S000047	A	04/01/2010	No és causa del vianant	Motocicleta	

	Codi_expedient	Descripció_carnet	Data_Completa	Descripció_causa_vianant	Descripció_tipus_de_vehicle	Descri
1544	2010S000047	A	04/01/2010	No és causa del vianant	Motocicleta	
1545	2010S000047	A	04/01/2010	No és causa del vianant	Motocicleta	
1546	2010S000047	A	04/01/2010	No és causa del vianant	Motocicleta	

5 rows × 43 columns

In [46]:
```
#Media de víctimas según Edat:
#Average victimes per age:

accidents[['Edat', 'Número_de_víctimes']].groupby(['Edat'], as_index=False).mean()
```

Out[46]:

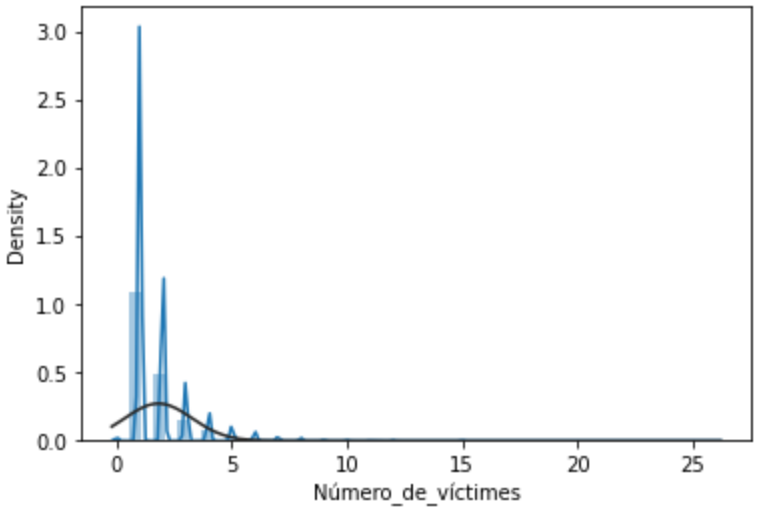
	Edat	Número_de_víctimes
0	0	2.084782
1	1	3.372169
2	2	3.443284
3	3	2.636014
4	4	2.479566
...	...	...
96	96	1.476190
97	97	1.478261
98	98	1.000000
99	101	1.000000
100	106	1.000000

101 rows × 2 columns

In [47]:
```
#Aplicamos el test de normalidad al número de víctimas.
#Normality test applied to victims.

from scipy.stats import norm
sns.distplot(accidents['Número_de_víctimes'], fit = norm)
```

```
E:\Anaconda\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot` is
a deprecated function and will be removed in a future version. Please adapt your code to u
se either `displot` (a figure-level function with similar flexibility) or `histplot` (an a
xes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[47]:
```
<AxesSubplot:xlabel='Número_de_víctimes', ylabel='Density'>
```



In [48]:
```
#Calculamos el coeficiente de asimetria. Al ser mayor de 1, los valores son más densos hac
#En una distribución normal, el valor debería acercarse a 0.

#Asymetry coefficient for victims. Since it's larger than 1, values are more dense to the
#On a normal distribution, the value should be closer to 0.

accidents['Número_de_víctimes'].skew()
```

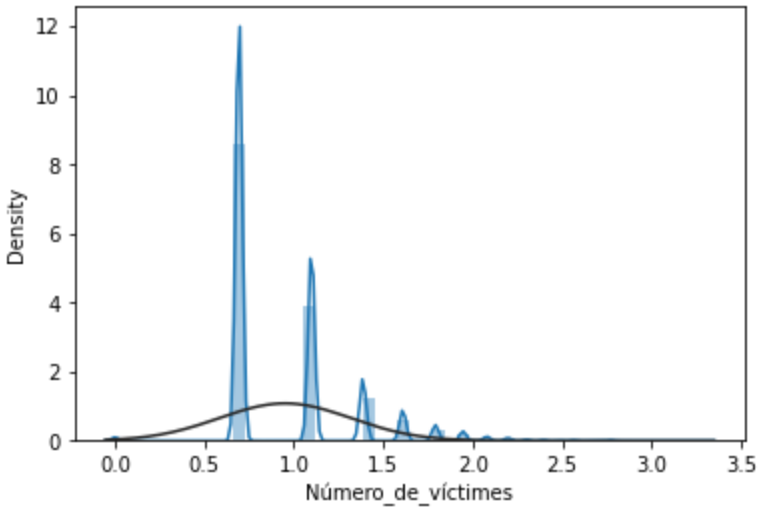
Out[48]:
```
4.257247071747584
```

In [49]:
```
#Calculamos el valor kurtosis para averiguar la relación del pico central con los extremos
#El valor debería ser cercano a 1 para ser coherente con la normalidad de la variable.

#Kurtosis value calculation to check the relationship between the central peak with the fa
#The value should be colse to 1 to be consistent with the normality of the variable.

accidents['Número_de_víctimes'].kurt()
```

Out[49]:
```
36.51316577644762
```

In [50]:
```
#Si tenemos una variable que no es normal, debemos aplicar el logaritmo a la variable:
#We apply the log to make "Numero de victimes" reach normality.

accidents['Número_de_víctimes'] = np.log1p(accidents['Número_de_víctimes'])
sns.distplot(accidents['Número_de_víctimes'], fit = norm)
```

Out[50]:
```
E:\Anaconda\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot` is
a deprecated function and will be removed in a future version. Please adapt your code to u
se either `displot` (a figure-level function with similar flexibility) or `histplot` (an a
xes-level function for histograms).
  warnings.warn(msg, FutureWarning)
<AxesSubplot:xlabel='Número_de_víctimes', ylabel='Density'>
```



In [51]:
```
#Comprobamos si la variable ha sido normalizada:
#We check the previous values again:

accidents['Número_de_víctimes'].skew()
```

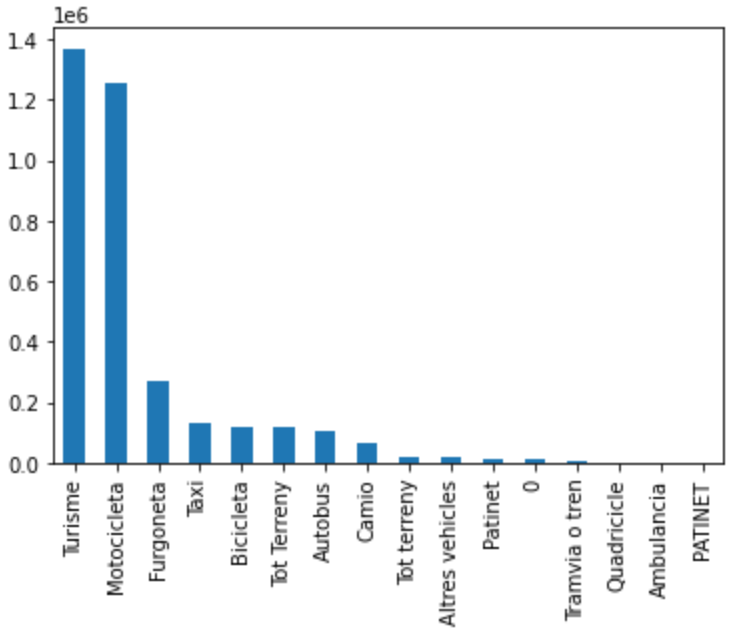
Out[51]:
```
1.4101229756005436
```

In [52]:
```
accidents['Número_de_víctimes'].kurt()
```

Out[52]:
```
2.465432491985602
```

In [53]:
```
#Vamos a revisar a continuación variables categóricas:
#Let's check now categorical values:
accidents["Descripció_tipus_de_vehicle"].value_counts().plot(kind='bar')
```

Out[53]:
```
<AxesSubplot:>
```

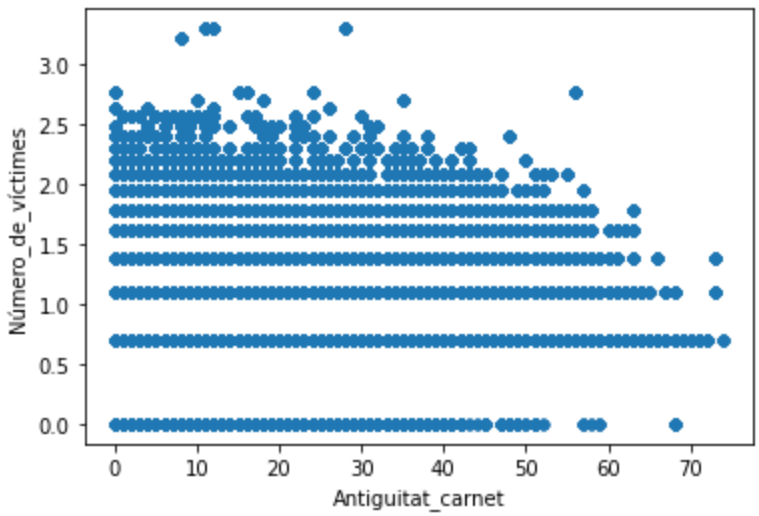


In [54]:
```
#Realizaremos ahora un análisis bivariante entre "Antiguitat Carnet" y "Número de víctimes
#Now a bivariant check between "Antiguitat Carnet" and "Número de víctimes".

var = 'Antiguitat_carnet'
```

```
data = pd.concat([accidents['Número_de_víctimes'], accidents['Antiguitat_carnet']], axis=1
data.plot.scatter(x='Antiguitat_carnet', y='Número_de_víctimes')
```

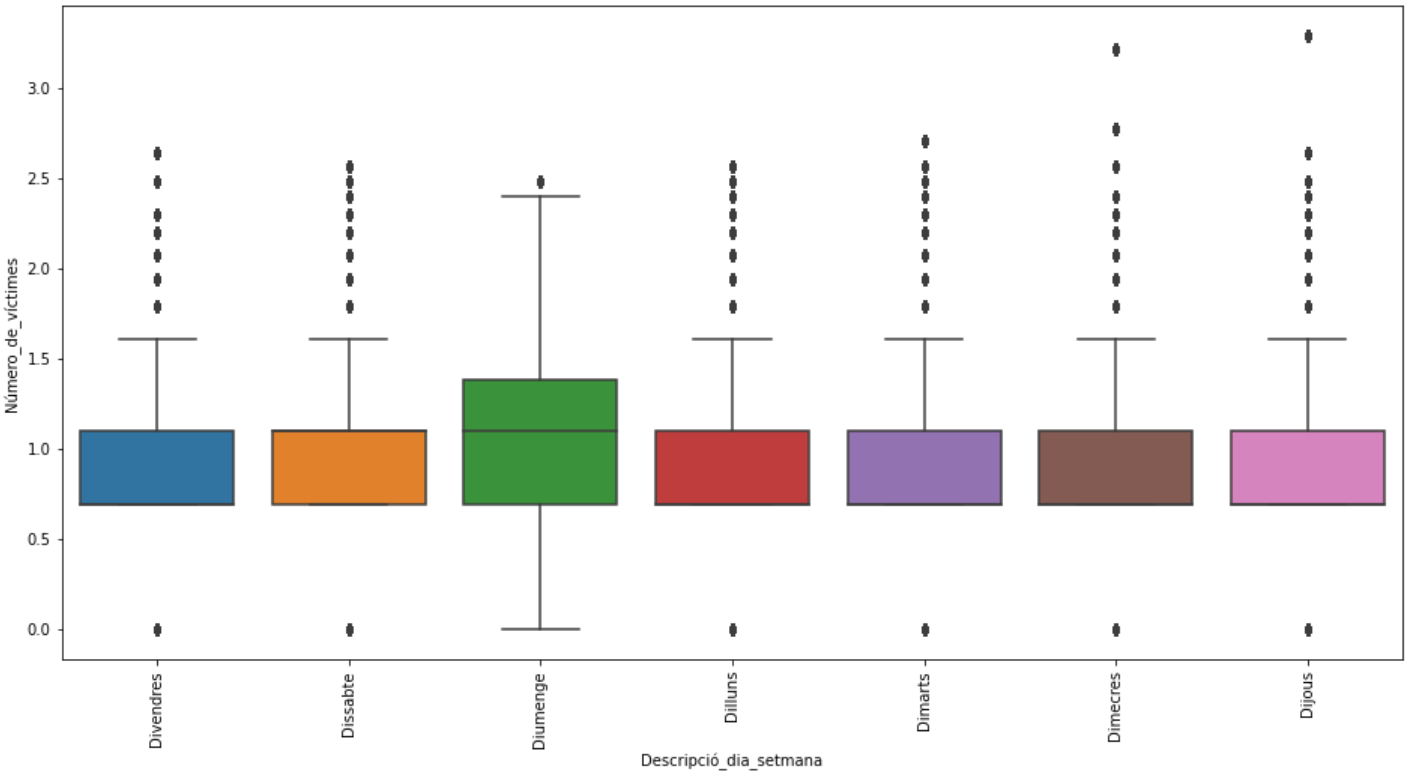
Out[54]:

```
<AxesSubplot:xlabel='Antiguitat_carnet', ylabel='Número_de_víctimes'>
```



In [55]:

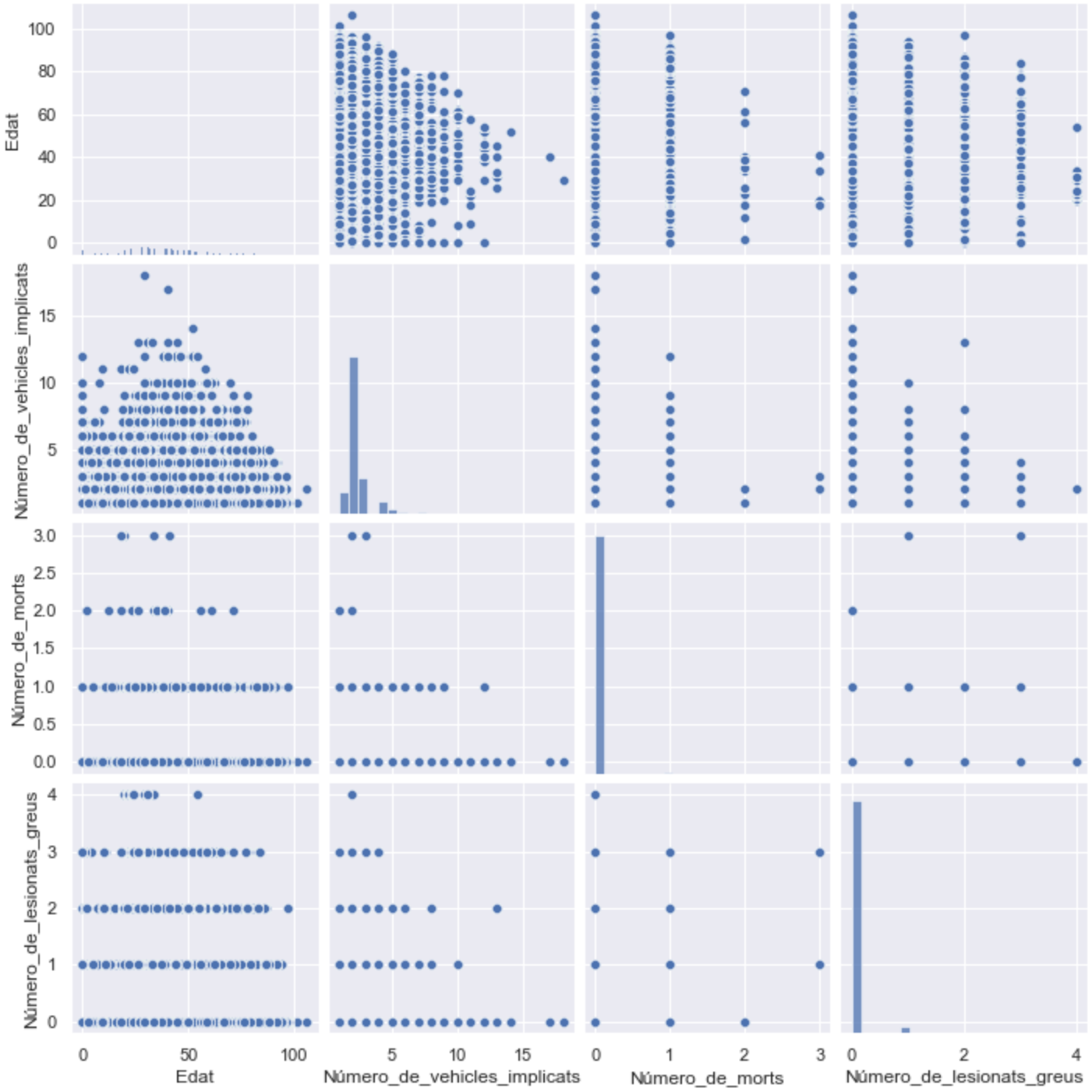
```
#Analizamos variables categóricas
#Another check of categorical values:

var = 'Descripció_dia_setmana'
data = pd.concat([accidents['Número_de_víctimes'], accidents[var]], axis=1)
f, ax = plt.subplots(figsize=(16, 8))
fig = sns.boxplot(x=var, y='Número_de_víctimes', data=data)
plt.xticks(rotation=90);
```



In [56]:

```
#Revisamos relaciones cruzadas entre varias variables a la vez:
#We check crossed relationships between some variables:

sns.set()
cols = ['Edat', 'Número_de_vehicles_implicats','Número_de_morts',
                    'Número_de_lesionats_greus']
```

```
sns.pairplot(accidents[cols], height = 2.5)
plt.show();
```



In [57]:

```
#Vamos a crear un dataframe para hacer una correlacion pero solo de las columnas que nos i
#We create a new dataframe to correlate some of the columns we are interested in:

acc_corr= accidents[['Edat', 'Número_de_víctimes','Dia_de_mes', 'Antiguitat_carnet',
                     'Número_de_vehicles_implicats','Número_de_morts',
                     'Número_de_lesionats_lleus','Número_de_lesionats_greus']]
acc_corr.head()
```

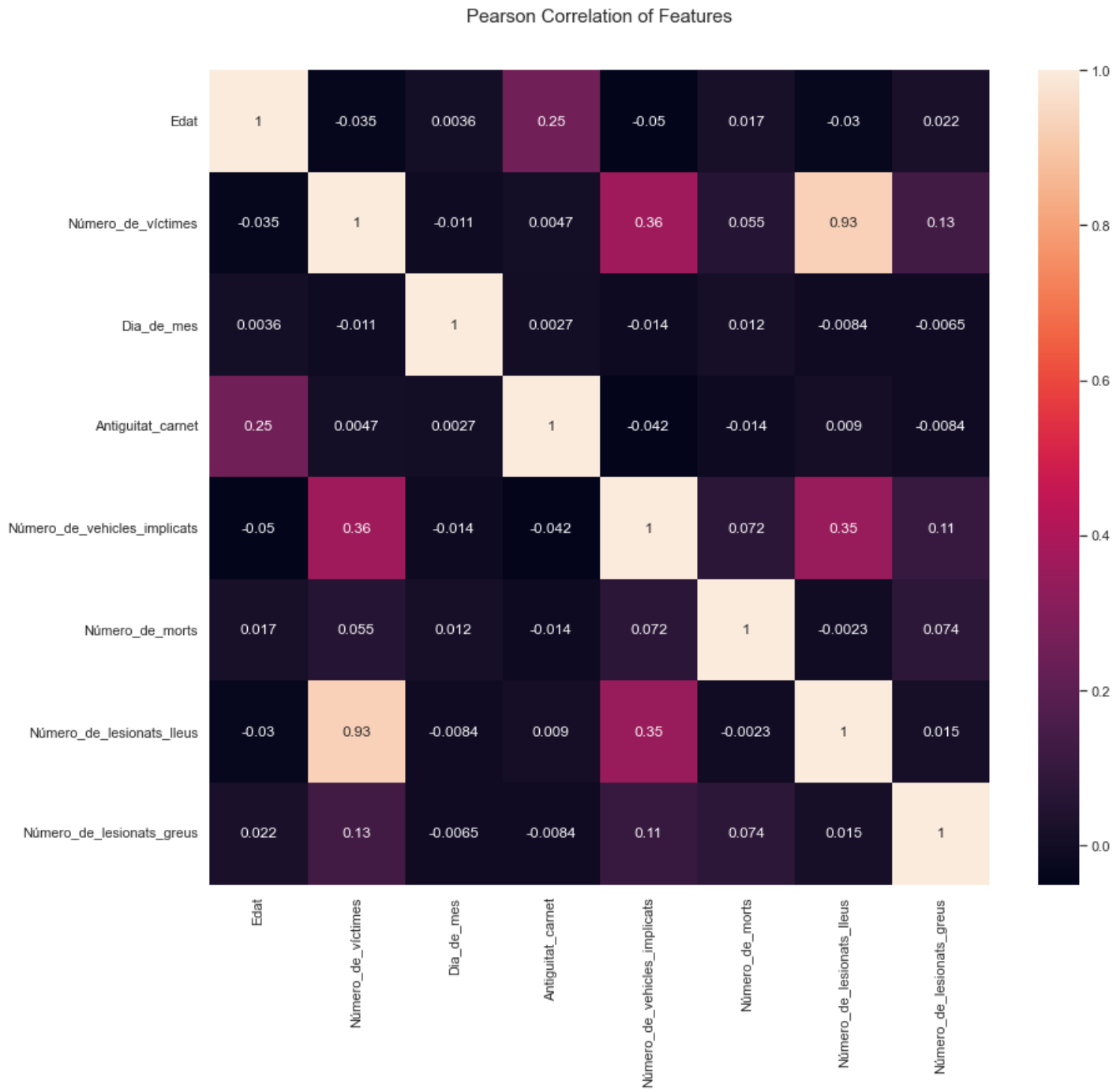
Out[57]:

	Edat	Número_de_víctimes	Dia_de_mes	Antiguitat_carnet	Número_de_vehicles_implicats	Número_de_morts	Núme
0	30	1.098612	1	3.0	2.0	0.0	
1	30	1.098612	1	3.0	2.0	0.0	
2	30	1.098612	1	3.0	2.0	0.0	

	Edat	Número_de_víctimes	Dia_de_mes	Antiguitat_carnet	Número_de_vehicles_implicats	Número_de_morts	Núme
3	30	1.098612	1	3.0	2.0	0.0	
4	30	1.098612	1	3.0	2.0	0.0	

In [58]:

```
#Creamos la representación de la correlación.
#We create the visual representation of the correlation.

correlation_matrix = acc_corr.corr()
correlation_matrix

plt.figure(figsize=(14,12))
plt.title('Pearson Correlation of Features', y=1.05, size=15)
sns.heatmap(correlation_matrix, annot=True);

#A parte de la correlación entre distrito-barrio, lesionados leves-victimas, vemos que las
#correlaciones son más altas para Edad-Antiguitat_Carnet,Num_de_Victimes-Num_vehicles_impl
#y lesionats lleus-vehicles implicats

#We can see some obviuos correlations between distrito-barrio, lesionados leves-victimas,
# Edad-Antiguitat_Carnet,Num_de_Victimes-Num_vehicles_implicats and y lesionats lleus-vehi
```

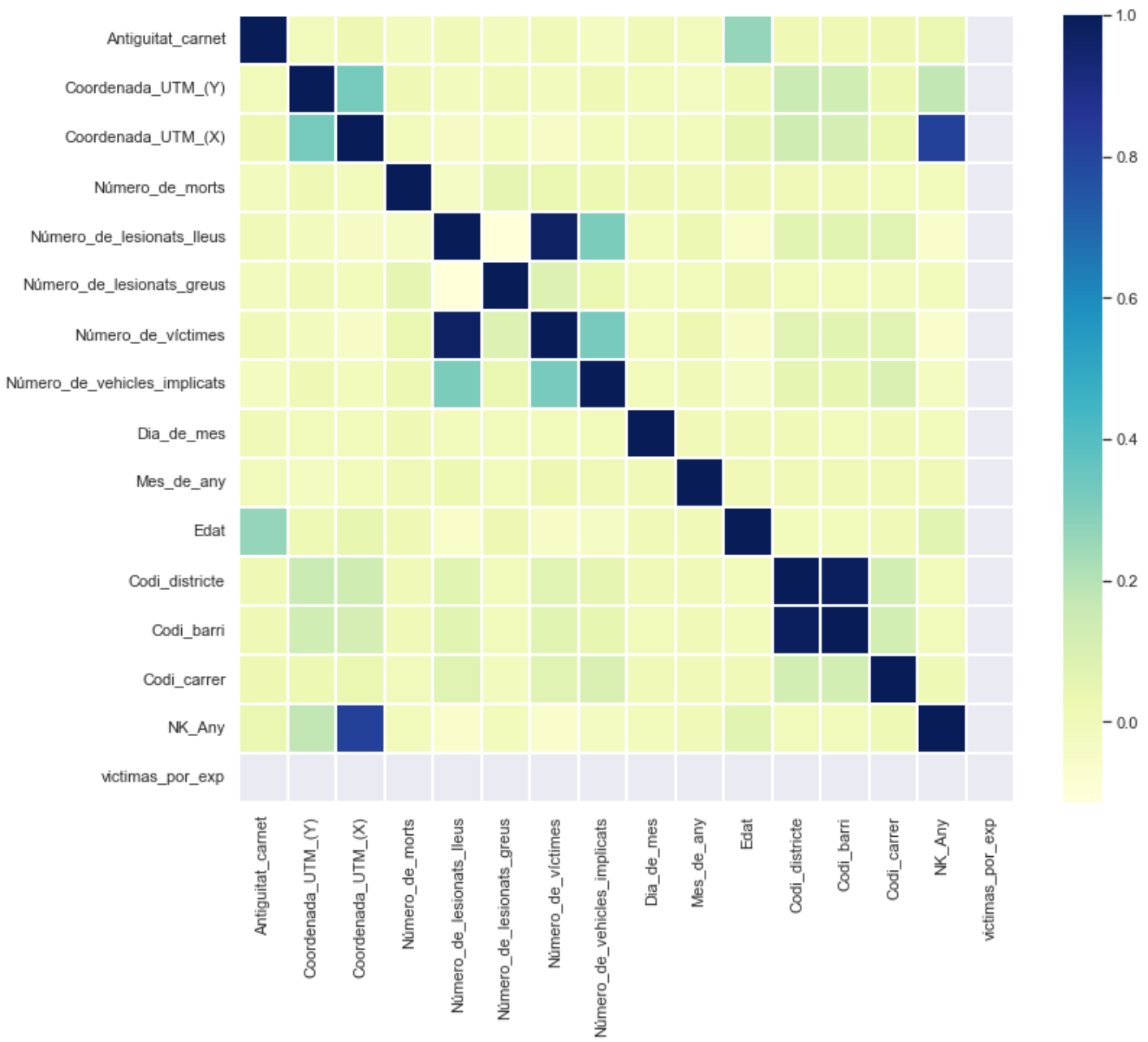


In [59]:

```
#Realizaremos el mismo estudio en otro formato:
#Another way of creating a similar study:

corrmat = accidents.corr(method='spearman')
f, ax = plt.subplots(figsize=(12, 10))
sns.heatmap(corrmat, ax=ax, cmap="YlGnBu", linewidths=0.1)
```

Out[59]:

```
<AxesSubplot:>
```

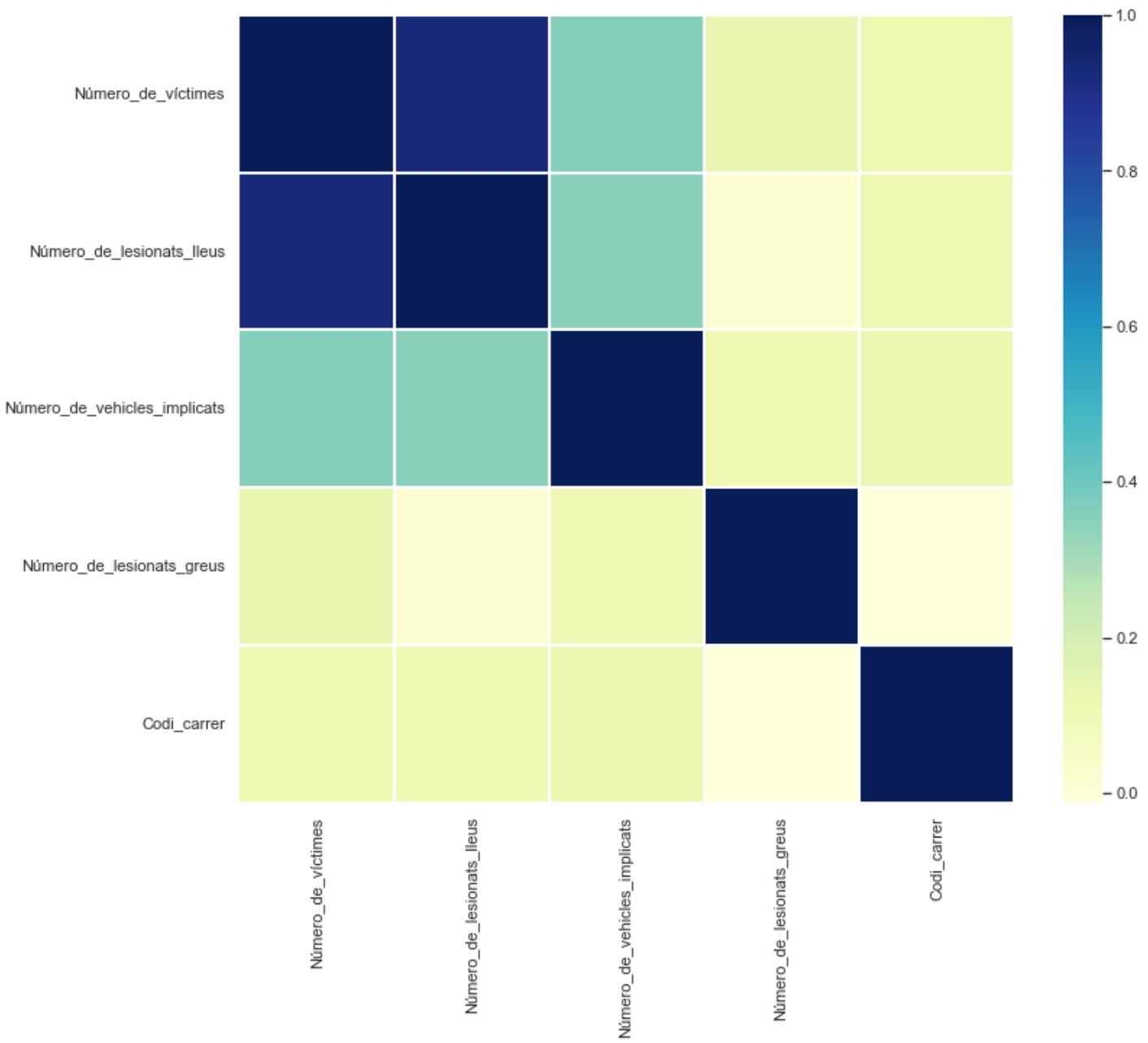


In [62]:

```
#Revisamos solo las 5 var más relacionadas con "Numero de victimes".
#We can also check the 5 variables most related with "Numero de victimes".

k = 5
cols = corrmat.nlargest(k, 'Número_de_víctimes')['Número_de_víctimes'].index
cm = np.corrcoef(accidents[cols].values.T)
f, ax = plt.subplots(figsize=(12, 10))
sns.heatmap(cm, ax=ax, cmap="YlGnBu", linewidths=0.1, yticklabels=cols.values, xticklabels
```

Out[62]:

```
<AxesSubplot:>
```



In [66]:
```
nbconvert --to html notebook.ipynb
```

```
  File "C:\Users\quique\AppData\Local\Temp/ipykernel_7700/479693054.py", line 1
    nbconvert --to html notebook.ipynb
                       ^
SyntaxError: invalid syntax
```

In [ ]: