DATA ACCIDENTS GUATEMALA

July 30, 2018

```
In [1]: import pandas as pd
        import numpy as np
```

```
/home/ebon1/Development/datascience/accidentesguatemala/venv/lib/python3.6/importlib/_bootstra
  return f(*args, **kwds)
/home/ebon1/Development/datascience/accidentesguatemala/venv/lib/python3.6/importlib/_bootstra
  return f(*args, **kwds)
```

1 Fallecidos Lecionado

1.1 Import DATA

```
In [2]: injured_2009 = pd.read_csv("L2009.csv")
        dead_2009 = pd.read_csv("F2009.csv")
        dead_injured_2010 = pd.read_csv('FYL2010.csv')
        dead_injured_2011 = pd.read_csv('FYL2011.csv')
        dead_injured_2012 = pd.read_csv('FYL2012.csv')
        dead_injured_2013 = pd.read_csv('FYL2013.csv')
        dead_injured_2014 = pd.read_csv('FYL2014.csv')
        dead_injured_2015 = pd.read_csv('FYL2015.csv')
        dead_injured_2016 = pd.read_csv('FYL2016.csv')
        dead_injured_2017 = pd.read_csv('FYL2017.csv')
        dead_2009.head()
```

Out[2]:

	num_hecho	dia_ocu	mes_ocu	dia_sem_ocu	hora_ocu	depto_ocu	areag_ocu	\
0	864	25	12	5	999	1	1	
1	656	25	8	2	0	16	2	
2	116	19	2	4	0	14	2	
3	170	7	3	6	0	5	2	
4	291	19	4	7	0	19	1	

	sexo_pil	edad_fall	g_edad	edad_m1	tipo_vehi	color_vehi	modelo_vehi	\
0	1	22	5	1	4	1	9999	
1	1	43	7	1	3	1	1986	
2	2	1	1	2	3	5	1987	
3	1	28	6	1	4	1	1988	
4	1	17	4	2	4	5	1991	

```
   causa_acc
0          1
1          4
2          3
3          2
4          1
```

1.2 Clean data for 2009

- Standarized names like sexo_les to sexo_fall_les to be able to concat table injured and dead
- concated dead_2009 with injured_2009
- created new table dead_injured_2009
- added the column zona_ocu with 99 default valeu

```
In [3]: injured_2009.rename(columns={'sexo_les': 'sexo_fall_les', 'edad_les': 'edad_fall_les'}
        dead_2009.rename(columns={'edad_fall': 'edad_fall_les', 'sexo_pil': 'sexo_fall_les'}, i
        injured_2009['fall_les'] = 2
        dead_2009['fall_les'] = 1


        dead_injured_2009 = pd.concat([injured_2009, dead_2009])
        dead_injured_2009['zona_ocu'] = 99
        dead_injured_2009['zona_ocu'] = 99
```

Standarizing the variable fall_less: - 2010: Changed index name lesio_fall -> fall_less - 2011: Changed index name condicion_pil -> fall_less - 2012: Changed index name estado_implicado -> fall_less - 2013: Changed index name estado_pil -> fall_less

```
In [4]: dead_injured_2010.rename(columns={'lesio_fall': 'fall_les'}, inplace=True)
        dead_injured_2011.rename(columns={'condicion_pil': 'fall_les'}, inplace=True)
        dead_injured_2012.rename(columns={'estado_implicado': 'fall_les'}, inplace=True)
        dead_injured_2013.rename(columns={'estado_pil': 'fall_les'}, inplace=True)

In [5]: dead_injured_2010.head()

Out[5]:    dia_ocu  mes_ocu  dia_sem_ocu  hora_ocu  depto_ocu  areag_ocu  zona_ocu  \
        0       23        9            4         7          1          1         5
        1       23        9            4         7          1          1         5
        2       28        3            7         4          1          1         7
        3        7        8            6         8          1          1         7
        4        5        7            1        20          1          1         7

           sexo_fall_les  edad_fall_les  g_edad  fall_les  tipo_vehi  Causa_acc
        0              2            999      11         1          1          1
        1              2            999      11         1          1          1
        2              2            999      11         2          1          1
        3              1            999      11         1          1          1
        4              1            999      11         1          1          1
```

```
In [6]: ## Added year to all tables

In [7]: dead_injured_2009['year'] = 2009
        dead_injured_2010['year'] = 2010
        dead_injured_2011['year'] = 2011
        dead_injured_2012['year'] = 2012
        dead_injured_2013['year'] = 2013
        dead_injured_2014['year'] = 2014
        dead_injured_2015['year'] = 2015
        dead_injured_2016['year'] = 2016
        dead_injured_2017['year'] = 2017
```

1.3 Clean data for 2010

- Standarized names Causa_acc to causa_acc
- Added columns num_hech0, edad_m1, color_vehi, modelo_vehi
- concated dead_injured_2009, dead_injured_2010
- created new table dead_injured_2009_2010

```
In [8]: dead_injured_2010.rename(columns={'Causa_acc': 'causa_acc'}, inplace=True)

        dead_injured_2010['num_hecho'] = -1
        dead_injured_2010['edad_m1'] = 9
        dead_injured_2010['color_vehi'] = -1
        dead_injured_2010['modelo_vehi'] = -1

        dead_injured_2009_2010 = pd.concat([dead_injured_2009, dead_injured_2010], sort=False)
```

1.4 Clean data for 2011

- Standarized names sexo_pil to sexo_pil and edad_pil to edad_fall_les
- Added columns on 2009_20010 marca_vehi, muni_ocu with exception values
- concated dead_injured_2009_2010, dead_injured_2011
- created new table dead_injured_2009_2011

```
In [9]: dead_injured_2011.rename(columns={'sexo_pil': 'sexo_fall_les'}, inplace=True)
        dead_injured_2011.rename(columns={'edad_pil': 'edad_fall_les'}, inplace=True)

        # 2009_2010
        dead_injured_2009_2010['marca_vehi'] = 99
        dead_injured_2009_2010['muni_ocu'] = -1

        dead_injured_2009_2011 = pd.concat([dead_injured_2009_2010, dead_injured_2011], sort=Fa
        dead_injured_2009_2011.to_csv('dead_injured_2009_2011.csv', sep=',')
```

1.5 Clean data for 2012

- Standarized names mupio_ocu to muni_ocu , g_edad_fall_les to g_edad, casusa_acc to causa_acc

- Added columns on 20012 marca_vehi, modelo_vehi with exception values
- Checked if there were any ocurrances where g_edad was larger than 11
- concated dead_injured_2009_2011, dead_injured_2012
- created new table dead_injured_2009_2012

```
In [10]: dead_injured_2012.rename(columns={'mupio_ocu': 'muni_ocu'}, inplace=True)
         mask = dead_injured_2009_2011.g_edad > 11
         dead_injured_2009_2011.loc[mask, 'g_edad'] = -1
         # Check for ocurrance where g_edd_less is more than 11
         print(dead_injured_2012[dead_injured_2012 > 11].sum())
         # We can conclude that 2012  g_edad_fall_les is the same as g_edad
         dead_injured_2012.rename(columns={'g_edad_fall_les': 'g_edad'}, inplace=True)
         dead_injured_2012.rename(columns={'casusa_acc': 'causa_acc'}, inplace=True)
         dead_injured_2012['marca_vehi'] = 99
         dead_injured_2012['modelo_vehi'] = -1
         dead_injured_2009_2012 = pd.concat([dead_injured_2009_2011, dead_injured_2012], sort=
         dead_injured_2009_2012.to_csv('dead_injured_2009_2012.csv', sep=',')
```

```
num_hecho          20069214.0
dia_ocu               79990.0
mes_ocu                7464.0
dia_sem_ocu               0.0
hora_ocu              64227.0
depto_ocu             38880.0
muni_ocu            5632049.0
areag_ocu                 0.0
zona_ocu             488846.0
sexo_fall_les             0.0
edad_fall_les        346131.0
edad_m1                   0.0
g_edad_fall_les           0.0
fall_les                  0.0
tipo_vehi              8569.0
color_vehi           101398.0
casusa_acc              693.0
year               12746020.0
dtype: float64
```

1.6 Clean data for 2013

- Standarized names
- Added columns on 2013 modelo_vehi, grupo_mode_veh(2009-2012) with exception values
- Checked if there were any ocurrances where g_edad was larger than 11
- concated dead_injured_2009_2012, dead_injured_2013
- created new table dead_injured_2009_2013

```
In [11]: dead_injured_2013.rename(columns={'color_veh': 'color_vehi', 'modelo_veh': 'modelo_ve
         dead_injured_2013.rename(columns={'mupio_ocu': 'muni_ocu'}, inplace=True)
```

```
dead_injured_2013.rename(columns={'sexo_pil': 'sexo_fall_les', 'edad_pil': 'edad_fall

# The reason for this change is that modelo_vehi is a year and the data does not have
# it seems to be duplicated with group_mode_veh so we made modelo_vehi the exception
# created on the consolidated table a nue column gropo_mode_veh with exception value
dead_injured_2013['modelo_vehi'] = 9999
dead_injured_2009_2012['grupo_mode_veh'] = -1

dead_injured_2009_2013 = pd.concat([dead_injured_2009_2012, dead_injured_2013], sort=
dead_injured_2009_2013.to_csv('dead_injured_2009_2013.csv', sep=',')
```

1.7 Clean data for 2014

- Standarized names
- Added columns on 2014 causa_acc, grupo_mode_veh exception values
- Also added to the dead_injured_2009_2013 table missing columns with exception values
- Generated edad_quinquenales based on age
- concated dead_injured_2009_2013, dead_injured_2014
- created new table dead_injured_2009_2014

In [12]:
```
dead_injured_2014.rename(columns={'marca_veh': 'marca_vehi', 'modelo_veh': 'modelo_veh
dead_injured_2014.rename(columns={'área_geo_ocu': 'area_geo_ocu'}, inplace=True)
dead_injured_2014.rename(columns={'mayor_menor': 'edad_m1'}, inplace=True)
dead_injured_2014.rename(columns={'mupio_ocu': 'muni_ocu'}, inplace=True)
dead_injured_2014.rename(columns={'color_veh': 'color_vehi'}, inplace=True)

dead_injured_2014.rename(columns={'num_correlativo': 'num_hecho'}, inplace=True)
dead_injured_2014.rename(columns={'edad_víc': 'edad_fall_les'}, inplace=True)
dead_injured_2014.rename(columns={'día_ocu': 'dia_ocu'}, inplace=True)
dead_injured_2014.rename(columns={'día_sem_ocu': 'dia_sem_ocu'}, inplace=True)
dead_injured_2014.rename(columns={'sexo_víc': 'sexo_fall_les'}, inplace=True)

dead_injured_2014['causa_acc'] = 99
dead_injured_2014['grupo_mode_veh'] = -1


dead_injured_2009_2013.rename(columns={'areag_ocu': 'area_geo_ocu'}, inplace=True)
dead_injured_2009_2013['num_corre'] = -1
dead_injured_2009_2013['corre_base'] = -1

dead_injured_2009_2013['tipo_eve'] = 99



dead_injured_2009_2013['g_hora'] = 4

# Change Exception value to -1 because it keeps growing
dead_injured_2009_2013.loc[dead_injured_2009_2013.g_edad == 11, 'g_edad'] = -1
```

```
dead_injured_2014.loc[dead_injured_2014.g_edad == 12, 'g_edad'] = -1
```

```
import math
def get_age_range(x):
    if x == 999:
        return 18
    elif x >= 80:
        return 17
    return math.floor((x/5)+1)

dead_injured_2009_2013['edad_quinquenales'] = dead_injured_2009_2013['edad_fall_les']

dead_injured_2009_2014 = pd.concat([dead_injured_2009_2013, dead_injured_2014], sort=
dead_injured_2009_2014.to_csv('dead_injured_2009_2014.csv', sep=',')

dead_injured_2009_2014.head()
```

Out[12]:

	num_hecho	dia_ocu	mes_ocu	dia_sem_ocu	hora_ocu	depto_ocu	\
0	6133	29	11	7	6	1	
1	1969	11	4	6	6	22	
2	3927	4	8	2	7	9	
3	4284	4	8	2	7	9	
4	4420	4	8	2	7	9	

	area_geo_ocu	sexo_fall_les	edad_fall_les	g_edad	...	\
0	1	1	1	1	...	
1	2	1	1	1	...	
2	2	1	1	1	...	
3	2	1	1	1	...	
4	2	1	1	1	...	

	zona_ocu	year	marca_vehi	muni_ocu	grupo_mode_veh	num_corre	\
0	99	2009	99	-1	-1	-1	
1	99	2009	99	-1	-1	-1	
2	99	2009	99	-1	-1	-1	
3	99	2009	99	-1	-1	-1	
4	99	2009	99	-1	-1	-1	

	corre_base	tipo_eve	g_hora	edad_quinquenales
0	-1	99	4	1
1	-1	99	4	1
2	-1	99	4	1
3	-1	99	4	1
4	-1	99	4	1

[5 rows x 26 columns]

1.8 Clean data for 2015

- Standarized names
- Added columns on 2014 causa_acc, grupo_mode_veh exception values
- Also added to the dead_injured_2009_2013 table missing columns with exception values
- Generated edad_quinquenales based on age
- concated dead_injured_2009_2013, dead_injured_2014
- created new table dead_injured_2009_2014

In [13]:
```
dead_injured_2015.rename(columns={'área_geo_ocu': 'area_geo_ocu'}, inplace=True)
dead_injured_2015.rename(columns={'día_ocu': 'dia_ocu'}, inplace=True)
dead_injured_2015.rename(columns={'día_sem_ocu': 'dia_sem_ocu'}, inplace=True)
dead_injured_2015.rename(columns={'núm_corre': 'num_corre'}, inplace=True)
dead_injured_2015.rename(columns={'g_modelo_veh': 'grupo_mode_veh'}, inplace=True)
dead_injured_2015.rename(columns={'marca_veh': 'marca_vehi'}, inplace=True)
dead_injured_2015.rename(columns={'modelo_veh': 'modelo_vehi'}, inplace=True)
dead_injured_2015.rename(columns={'núm_corre': 'num_hecho'}, inplace=True)
dead_injured_2015.rename(columns={'mupio_ocu': 'muni_ocu'}, inplace=True)
dead_injured_2015.rename(columns={'tipo_veh': 'tipo_vehi'}, inplace=True)
dead_injured_2015.rename(columns={'mayor_menor': 'edad_m1'}, inplace=True)
dead_injured_2015.rename(columns={'edad_per': 'edad_fall_les'}, inplace=True)
dead_injured_2015.rename(columns={'sexo_per': 'sexo_fall_les'}, inplace=True)
dead_injured_2015.rename(columns={'color_veh': 'color_vehi'}, inplace=True)
dead_injured_2015['causa_acc'] = 99
dead_injured_2015['num_corre'] = -1
dead_injured_2015['corre_base'] = -1


dead_injured_2015.rename(columns={'g_edad_60ymás': 'g_edad'}, inplace=True)

dead_injured_2015.loc[dead_injured_2015.g_edad == 12, 'g_edad'] = -1

del dead_injured_2015['año_ocu']

dead_injured_2009_2014['int_o_noint'] = 9

def g_hora_5(horaExacta):
    if (horaExacta < 11):
        return 1
    elif (horaExacta < 20):
        return 2
    elif (horaExacta < 24):
        return 3
    return 4

dead_injured_2009_2014['g_hora_5'] = dead_injured_2009_2014['hora_ocu'].apply(g_hora_5
```

```
dead_injured_2015.head()
dead_injured_2009_2015 = pd.concat([dead_injured_2009_2014, dead_injured_2015], sort=
dead_injured_2009_2015.to_csv('dead_injured_2009_2015.csv', sep=',')
```

In [14]:
```
dead_injured_2016.rename(columns={'área_geo_ocu': 'area_geo_ocu'}, inplace=True)
dead_injured_2016.rename(columns={'día_ocu': 'dia_ocu'}, inplace=True)
dead_injured_2016.rename(columns={'día_sem_ocu': 'dia_sem_ocu'}, inplace=True)
dead_injured_2016.rename(columns={'núm_corre': 'num_corre'}, inplace=True)
dead_injured_2016.rename(columns={'g_modelo_veh': 'grupo_mode_veh'}, inplace=True)
dead_injured_2016.rename(columns={'marca_veh': 'marca_vehi'}, inplace=True)
dead_injured_2016.rename(columns={'modelo_veh': 'modelo_vehi'}, inplace=True)
dead_injured_2016.rename(columns={'núm_corre': 'num_hecho'}, inplace=True)
dead_injured_2016.rename(columns={'mupio_ocu': 'muni_ocu'}, inplace=True)
dead_injured_2016.rename(columns={'tipo_veh': 'tipo_vehi'}, inplace=True)
dead_injured_2016.rename(columns={'mayor_menor': 'edad_m1'}, inplace=True)
dead_injured_2016.rename(columns={'edad_per': 'edad_fall_les'}, inplace=True)
dead_injured_2016.rename(columns={'sexo_per': 'sexo_fall_les'}, inplace=True)
dead_injured_2016.rename(columns={'color_veh': 'color_vehi'}, inplace=True)
dead_injured_2016['causa_acc'] = 99
dead_injured_2016['num_corre'] = -1
dead_injured_2016['corre_base'] = -1


dead_injured_2016.rename(columns={'g_edad_60ymás': 'g_edad'}, inplace=True)

dead_injured_2016.loc[dead_injured_2016.g_edad == 12, 'g_edad'] = -1

del dead_injured_2016['año_ocu']
# print(dead_injured_2016['año_ocu'])


dead_injured_2009_2016 = pd.concat([dead_injured_2009_2015, dead_injured_2016], sort=
dead_injured_2009_2016.to_csv('dead_injured_2009_2016.csv', sep=',')
dead_injured_2009_2016.head()
```

Out[14]:

	num_hecho	dia_ocu	mes_ocu	dia_sem_ocu	hora_ocu	depto_ocu	\
0	6133.0	29	11	7	6	1	
1	1969.0	11	4	6	6	22	
2	3927.0	4	8	2	7	9	
3	4284.0	4	8	2	7	9	
4	4420.0	4	8	2	7	9	

	area_geo_ocu	sexo_fall_les	edad_fall_les	g_edad	...	\
0	1	1	1	1	...	
1	2	1	1	1	...	
2	2	1	1	1	...	
3	2	1	1	1	...	

```
4             2              1              1       1      ...

   muni_ocu  grupo_mode_veh  num_corre  corre_base  tipo_eve  g_hora  \
0        -1              -1         -1          -1        99       4
1        -1              -1         -1          -1        99       4
2        -1              -1         -1          -1        99       4
3        -1              -1         -1          -1        99       4
4        -1              -1         -1          -1        99       4

   edad_quinquenales  int_o_noint  g_hora_5  g_edad_80ymás
0                  1            9         1            NaN
1                  1            9         1            NaN
2                  1            9         1            NaN
3                  1            9         1            NaN
4                  1            9         1            NaN

[5 rows x 29 columns]
```

In [15]:
```
dead_injured_2017.rename(columns={'área_geo_ocu': 'area_geo_ocu'}, inplace=True)
dead_injured_2017.rename(columns={'día_ocu': 'dia_ocu'}, inplace=True)
dead_injured_2017.rename(columns={'día_sem_ocu': 'dia_sem_ocu'}, inplace=True)
dead_injured_2017.rename(columns={'núm_corre': 'num_corre'}, inplace=True)
dead_injured_2017.rename(columns={'g_modelo_veh': 'grupo_mode_veh'}, inplace=True)
dead_injured_2017.rename(columns={'marca_veh': 'marca_vehi'}, inplace=True)
dead_injured_2017.rename(columns={'modelo_veh': 'modelo_vehi'}, inplace=True)
dead_injured_2017.rename(columns={'núm_corre': 'num_hecho'}, inplace=True)
dead_injured_2017.rename(columns={'mupio_ocu': 'muni_ocu'}, inplace=True)
dead_injured_2017.rename(columns={'tipo_veh': 'tipo_vehi'}, inplace=True)
dead_injured_2017.rename(columns={'mayor_menor': 'edad_m1'}, inplace=True)
dead_injured_2017.rename(columns={'edad_per': 'edad_fall_les'}, inplace=True)
dead_injured_2017.rename(columns={'sexo_per': 'sexo_fall_les'}, inplace=True)
dead_injured_2017.rename(columns={'color_veh': 'color_vehi'}, inplace=True)
dead_injured_2017['causa_acc'] = 99
dead_injured_2017['num_corre'] = -1
dead_injured_2017['corre_base'] = -1


dead_injured_2017.rename(columns={'g_edad_60ymás': 'g_edad'}, inplace=True)

dead_injured_2017.loc[dead_injured_2017.g_edad == 12, 'g_edad'] = -1

del dead_injured_2017['año_ocu']
dead_injured_2017['area_geo_ocu'] = -1

dead_injured_2009_2017 = pd.concat([dead_injured_2009_2016, dead_injured_2017], sort=
```

In [16]:
```
def g_edad_80(edadExacta):
    if (edadExacta == 999):
```

```
            return 16
        elif (edadExacta < 15):
            return 1
        elif (edadExacta >= 80):
            return 15
        edadExacta = int(((edadExacta) / 5) - 1)
        return edadExacta

del dead_injured_2009_2017['g_edad_80ymás']
del dead_injured_2009_2017['g_edad']

dead_injured_2009_2017['g_edad'] = dead_injured_2009_2017['edad_fall_les'].apply(g_eda

dead_injured_2009_2017.to_csv('dead_injured_2009_2017.csv', sep=',')
dead_injured_2009_2017.head()
```

Out[16]:

	num_hecho	dia_ocu	mes_ocu	dia_sem_ocu	hora_ocu	depto_ocu	\
0	6133.0	29	11	7	6	1	
1	1969.0	11	4	6	6	22	
2	3927.0	4	8	2	7	9	
3	4284.0	4	8	2	7	9	
4	4420.0	4	8	2	7	9	

	area_geo_ocu	sexo_fall_les	edad_fall_les	edad_m1	...	muni_ocu	\
0	1	1	1	2	...	-1	
1	2	1	1	2	...	-1	
2	2	1	1	2	...	-1	
3	2	1	1	2	...	-1	
4	2	1	1	2	...	-1	

	grupo_mode_veh	num_corre	corre_base	tipo_eve	g_hora	edad_quinquenales	\
0	-1	-1	-1	99	4	1	
1	-1	-1	-1	99	4	1	
2	-1	-1	-1	99	4	1	
3	-1	-1	-1	99	4	1	
4	-1	-1	-1	99	4	1	

	int_o_noint	g_hora_5	g_edad
0	9	1	1
1	9	1	1
2	9	1	1
3	9	1	1
4	9	1	1

[5 rows x 28 columns]

2 Hechos de Trancito

2.1 Import Data

```
In [17]: car_accidents_2009 = pd.read_csv('HDT2009.csv')
         car_accidents_2010 = pd.read_csv('HDT2010.csv')
         car_accidents_2011 = pd.read_csv('HDT2011.csv')
         car_accidents_2012 = pd.read_csv('HDT2012.csv')
         car_accidents_2013 = pd.read_csv('HDT2013.csv')
         car_accidents_2014 = pd.read_csv('HTD2014.csv')
         car_accidents_2015 = pd.read_csv('EDT2015.csv')
         car_accidents_2016 = pd.read_csv('HDT2016.csv')
         car_accidents_2017 = pd.read_csv('ADT2017.csv')

         car_accidents_2009['year'] = 2009
         car_accidents_2010['year'] = 2010
         car_accidents_2011['year'] = 2011
         car_accidents_2012['year'] = 2012
         car_accidents_2013['year'] = 2013
         car_accidents_2014['year'] = 2014
         car_accidents_2015['year'] = 2015
         car_accidents_2016['year'] = 2016
         car_accidents_2017['year'] = 2017
```

2.2 Clean Data Car Accidents 2009

```
In [18]: car_accidents_2009['zona_ocu'] = 99

         del car_accidents_2009['año_ocu']

         car_accidents_2009.loc[car_accidents_2009.g_edad_pil == 11, 'g_edad_pil'] = -1
         car_accidents_2009.head()
```

Out[18]:

	num_hecho	dia_ocu	mes_ocu	dia_sem_ocu	hora_ocu	depto_ocu	areag_ocu	\
0	1449	13	5	3	1	14	1	
1	1587	1	7	1	15	22	2	
2	679	1	3	7	17	17	2	
3	633	26	2	4	18	21	2	
4	673	1	3	7	15	10	1	

	sexo_pil	edad_pil	g_edad_pil	estado_pil	tipo_vehi	color_vehi	\
0	9	999	99	2	3	2	
1	2	13	2	1	4	4	
2	1	14	2	1	4	99	
3	1	15	2	9	1	1	
4	1	15	2	9	14	1	

	modelo_vehi	causa_acc	year	zona_ocu
0	1988	2	2009	99

```
1          9999          2  2009        99
2          9999          4  2009        99
3          9999          4  2009        99
4          2007          2  2009        99
```

2.3 Clean Data Car Accidents 2010

```
In [19]: car_accidents_2010.rename(columns={'causa_ac': 'causa_acc'}, inplace=True)
         car_accidents_2010.rename(columns={'color_v': 'color_vehi'}, inplace=True)
         car_accidents_2010.rename(columns={'modelo_v': 'modelo_vehi'}, inplace=True)
         car_accidents_2010.rename(columns={'tipo_v': 'tipo_vehi'}, inplace=True)


         car_accidents_2010['num_hecho'] = -1
         car_accidents_2010['g_edad_pil'] = -1

         car_accidents_2009_2010 = pd.concat([car_accidents_2009, car_accidents_2010], sort=Fa
         car_accidents_2009_2010.to_csv('car_accidents_2009_2010.csv', sep=',')
         car_accidents_2009_2010.head()
```

```
Out[19]:    num_hecho  dia_ocu  mes_ocu  dia_sem_ocu hora_ocu  depto_ocu  areag_ocu  \
         0       1449       13        5            3        1         14          1
         1       1587        1        7            1       15         22          2
         2        679        1        3            7       17         17          2
         3        633       26        2            4       18         21          2
         4        673        1        3            7       15         10          1

            sexo_pil  edad_pil  g_edad_pil  estado_pil  tipo_vehi  color_vehi  \
         0         9       999          99           2          3           2
         1         2        13           2           1          4           4
         2         1        14           2           1          4          99
         3         1        15           2           9          1           1
         4         1        15           2           9         14           1

            modelo_vehi  causa_acc  year  zona_ocu
         0         1988          2  2009        99
         1         9999          2  2009        99
         2         9999          4  2009        99
         3         9999          4  2009        99
         4         2007          2  2009        99
```

2.4 Clean Data Car Accidents 2011

```
In [20]: car_accidents_2011.rename(columns={'tipo_vehiculo': 'tipo_vehi'}, inplace=True)

         car_accidents_2009_2010['muni_ocu'] = -1
         car_accidents_2009_2010['marca_vehi'] = -1
```

```
def edad_m1(edad_pil):
    if (edad_pil == 999):
        return 9
    elif (edad_pil >= 18):
        return 1
    return 2

car_accidents_2009_2010['edad_m1'] = car_accidents_2009_2010['edad_pil'].apply(edad_m

car_accidents_2009_2011 = pd.concat([car_accidents_2009_2010, car_accidents_2011], so
car_accidents_2009_2011.to_csv('car_accidents_2009_2011.csv', sep=',')
car_accidents_2009_2011.head()
```

Out[20]:

	num_hecho	dia_ocu	mes_ocu	dia_sem_ocu	hora_ocu	depto_ocu	areag_ocu	\
0	1449	13	5	3	1	14	1	
1	1587	1	7	1	15	22	2	
2	679	1	3	7	17	17	2	
3	633	26	2	4	18	21	2	
4	673	1	3	7	15	10	1	

	sexo_pil	edad_pil	g_edad_pil	estado_pil	tipo_vehi	color_vehi	\
0	9	999	99	2	3	2	
1	2	13	2	1	4	4	
2	1	14	2	1	4	99	
3	1	15	2	9	1	1	
4	1	15	2	9	14	1	

	modelo_vehi	causa_acc	year	zona_ocu	muni_ocu	marca_vehi	edad_m1
0	1988	2	2009	99	-1	-1	9
1	9999	2	2009	99	-1	-1	2
2	9999	4	2009	99	-1	-1	2
3	9999	4	2009	99	-1	-1	2
4	2007	2	2009	99	-1	-1	2

In [21]:
```
car_accidents_2012.rename(columns={'condicion_pil': 'estado_pil'}, inplace=True)
car_accidents_2012.rename(columns={'g_edad': 'g_edad_pil'}, inplace=True)
car_accidents_2012.rename(columns={'mupio_ocu': 'muni_ocu'}, inplace=True)


car_accidents_2012['modelo_vehi'] = 9999
car_accidents_2012['marca_vehi'] = -1

car_accidents_2009_2012 = pd.concat([car_accidents_2009_2011, car_accidents_2012], so
car_accidents_2009_2012.to_csv('car_accidents_2009_2012.csv', sep=',')
car_accidents_2009_2012.head()
```

Out[21]:

	num_hecho	dia_ocu	mes_ocu	dia_sem_ocu	hora_ocu	depto_ocu	areag_ocu	\
0	1449	13	5	3	1	14	1	

```
1       1587        1        7            1       15         22          2
2        679        1        3            7       17         17          2
3        633       26        2            4       18         21          2
4        673        1        3            7       15         10          1

   sexo_pil  edad_pil  g_edad_pil  estado_pil  tipo_vehi  color_vehi  \
0         9       999          99           2          3           2
1         2        13           2           1          4           4
2         1        14           2           1          4          99
3         1        15           2           9          1           1
4         1        15           2           9         14           1

   modelo_vehi  causa_acc  year  zona_ocu muni_ocu  marca_vehi  edad_m1
0         1988          2  2009        99       -1          -1        9
1         9999          2  2009        99       -1          -1        2
2         9999          4  2009        99       -1          -1        2
3         9999          4  2009        99       -1          -1        2
4         2007          2  2009        99       -1          -1        2
```

In [22]:
```
car_accidents_2013.rename(columns={'color_veh': 'color_vehi'}, inplace=True)
car_accidents_2013.rename(columns={'modelo_veh': 'modelo_vehi'}, inplace=True)
car_accidents_2013.rename(columns={'marca_veh': 'marca_vehi'}, inplace=True)
car_accidents_2013.rename(columns={'tipo_veh': 'tipo_vehi'}, inplace=True)
car_accidents_2013.rename(columns={'mupio_ocu': 'muni_ocu'}, inplace=True)
car_accidents_2013.rename(columns={'mayor_menor': 'edad_m1'}, inplace=True)


car_accidents_2013['g_edad_pil'] = -1

car_accidents_2009_2012['g_edad_2'] = -1
car_accidents_2009_2012.loc[car_accidents_2009_2012.g_edad_pil == 11, 'g_edad_pil'] =

def g_hora(hora_exacta):
    try :
        hora_exacta = int(hora_exacta)
        if (hora_exacta < 11):
            return 1
        elif (hora_exacta < 20):
            return 2
        elif (hora_exacta < 24):
            return 3
    except:
        return 99
    return 4


car_accidents_2009_2012['g_hora'] = car_accidents_2009_2012['hora_ocu'].apply(g_hora)
```

```
car_accidents_2009_2013 = pd.concat([car_accidents_2009_2012, car_accidents_2013], so
car_accidents_2009_2013.to_csv('car_accidents_2009_2013.csv', sep=',')
car_accidents_2009_2013.head()
```

Out[22]:

	num_hecho	dia_ocu	mes_ocu	dia_sem_ocu	hora_ocu	depto_ocu	areag_ocu	\
0	1449	13	5	3	1	14	1	
1	1587	1	7	1	15	22	2	
2	679	1	3	7	17	17	2	
3	633	26	2	4	18	21	2	
4	673	1	3	7	15	10	1	

	sexo_pil	edad_pil	g_edad_pil	...	color_vehi	modelo_vehi	causa_acc	\
0	9	999	99	...	2	1988	2	
1	2	13	2	...	4	9999	2	
2	1	14	2	...	99	9999	4	
3	1	15	2	...	1	9999	4	
4	1	15	2	...	1	2007	2	

	year	zona_ocu	muni_ocu	marca_vehi	edad_m1	g_edad_2	g_hora
0	2009	99	-1	-1	9	-1	1
1	2009	99	-1	-1	2	-1	2
2	2009	99	-1	-1	2	-1	2
3	2009	99	-1	-1	2	-1	2
4	2009	99	-1	-1	2	-1	2

[5 rows x 22 columns]

In [23]:
```
car_accidents_2014.rename(columns={'num_correlativo': 'num_hecho'}, inplace=True)
car_accidents_2014.rename(columns={'día_ocu': 'dia_ocu'}, inplace=True)
car_accidents_2014.rename(columns={'día_sem_ocu': 'dia_sem_ocu'}, inplace=True)
car_accidents_2014.rename(columns={'mupio_ocu': 'muni_ocu'}, inplace=True)
car_accidents_2014.rename(columns={'área_geo_ocu': 'areag_ocu'}, inplace=True)
car_accidents_2014.rename(columns={'sexo_con': 'sexo_pil'}, inplace=True)
car_accidents_2014.rename(columns={'edad_con': 'edad_pil'}, inplace=True)
car_accidents_2014.rename(columns={'g_edad': 'g_edad_2'}, inplace=True)
car_accidents_2014.rename(columns={'mayor_menor': 'edad_m1'}, inplace=True)
car_accidents_2014.rename(columns={'estado_con': 'estado_pil'}, inplace=True)
car_accidents_2014.rename(columns={'tipo_veh': 'tipo_vehi'}, inplace=True)
car_accidents_2014.rename(columns={'modelo_veh': 'modelo_vehi'}, inplace=True)
car_accidents_2014.rename(columns={'color_veh': 'color_vehi'}, inplace=True)
car_accidents_2014.rename(columns={'marca_veh': 'marca_vehi'}, inplace=True)


car_accidents_2014['causa_acc'] = 99


def g_edad_nuevo(edadExacta):
```

```
        if (edadExacta == 999):
            return 16
        elif (edadExacta < 15):
            return 1
        elif (edadExacta >= 80):
            return 15

        edadExacta = int(((edadExacta) / 5) - 1)
        return edadExacta

    car_accidents_2014['g_edad_pil'] = car_accidents_2014['edad_pil'].apply(g_edad_nuevo)
```

```
    car_accidents_2009_2013['corre_base'] = -1
    car_accidents_2009_2013['tipo_eve'] = 99
```

```
    car_accidents_2009_2014 = pd.concat([car_accidents_2009_2013, car_accidents_2014], so
    car_accidents_2009_2014.to_csv('car_accidents_2009_2014.csv', sep=',')
    car_accidents_2009_2014.head()
```

Out[23]:

	num_hecho	dia_ocu	mes_ocu	dia_sem_ocu	hora_ocu	depto_ocu	areag_ocu	\
0	1449	13	5	3	1	14	1	
1	1587	1	7	1	15	22	2	
2	679	1	3	7	17	17	2	
3	633	26	2	4	18	21	2	
4	673	1	3	7	15	10	1	

	sexo_pil	edad_pil	g_edad_pil	...	causa_acc	year	zona_ocu	\
0	9	999	99	...	2	2009	99	
1	2	13	2	...	2	2009	99	
2	1	14	2	...	4	2009	99	
3	1	15	2	...	4	2009	99	
4	1	15	2	...	2	2009	99	

	muni_ocu	marca_vehi	edad_m1	g_edad_2	g_hora	corre_base	tipo_eve
0	-1	-1	9	-1	1	-1	99
1	-1	-1	2	-1	2	-1	99
2	-1	-1	2	-1	2	-1	99
3	-1	-1	2	-1	2	-1	99
4	-1	-1	2	-1	2	-1	99

[5 rows x 24 columns]

In [24]:
```
car_accidents_2015.head()
```

Out[24]:

	núm_corre	año_ocu	mes_ocu	día_ocu	día_sem_ocu	hora_ocu	g_hora	\

```
0          1     2015        1        1            4        16       3
1          2     2015        1        1            4        22       4
2          3     2015        1        1            4         2       1
3          4     2015        1        1            4         9       2
4          5     2015        1        1            4         1       1

   g_hora_5  depto_ocu  mupio_ocu  ...   g_edad_60ymás  edad_quinquenales  \
0         2          1        101  ...               2                  4
1         3          1        101  ...               5                  7
2         1          1        101  ...              12                 18
3         1          1        101  ...               3                  5
4         1          1        115  ...               3                  5

   estado_con  tipo_veh  marca_veh  color_veh  modelo_veh  g_modelo_veh  \
0           9         4         21          5        2011             5
1           1         4         21          5        9999             6
2           9         3         44          6        9999             6
3           9         4         40          5        9999             6
4           9         1         34          4        9999             6

   tipo_eve  year
0         1  2015
1         1  2015
2         1  2015
3         2  2015
4         2  2015

[5 rows x 26 columns]
```

In [25]:
```
car_accidents_2015.rename(columns={'núm_corre': 'num_hecho'}, inplace=True)
car_accidents_2015.rename(columns={'día_ocu': 'dia_ocu'}, inplace=True)
car_accidents_2015.rename(columns={'día_sem_ocu': 'dia_sem_ocu'}, inplace=True)
car_accidents_2015.rename(columns={'área_geo_ocu': 'areag_ocu'}, inplace=True)
car_accidents_2015.rename(columns={'tipo_veh': 'tipo_vehi'}, inplace=True)
car_accidents_2015.rename(columns={'modelo_veh': 'modelo_vehi'}, inplace=True)
car_accidents_2015.rename(columns={'color_veh': 'color_vehi'}, inplace=True)
car_accidents_2015.rename(columns={'marca_veh': 'marca_vehi'}, inplace=True)
car_accidents_2015.rename(columns={'mayor_menor': 'edad_m1'}, inplace=True)
car_accidents_2015.rename(columns={'edad_per': 'edad_pil'}, inplace=True)
car_accidents_2015.rename(columns={'estado_con': 'estado_pil'}, inplace=True)
car_accidents_2015.rename(columns={'g_edad_60ymás': 'g_edad_2'}, inplace=True)
car_accidents_2015.rename(columns={'mupio_ocu': 'muni_ocu'}, inplace=True)
car_accidents_2015.rename(columns={'sexo_per': 'sexo_pil'}, inplace=True)
car_accidents_2015.rename(columns={'sexo_per': 'sexo_pil'}, inplace=True)


car_accidents_2015['causa_acc'] = 99
car_accidents_2015['corre_base'] = -1
```

```
car_accidents_2015['g_edad_pil'] = -1

del car_accidents_2015['año_ocu']

car_accidents_2009_2014['g_hora_5'] = 5
car_accidents_2009_2014['g_modelo_veh'] = 6

car_accidents_2009_2015 = pd.concat([car_accidents_2009_2014, car_accidents_2015], so
car_accidents_2009_2015.to_csv('car_accidents_2009_2015.csv', sep=',')
car_accidents_2009_2015.head()
```

Out[25]:

	num_hecho	dia_ocu	mes_ocu	dia_sem_ocu	hora_ocu	depto_ocu	areag_ocu	\
0	1449	13	5	3	1	14	1	
1	1587	1	7	1	15	22	2	
2	679	1	3	7	17	17	2	
3	633	26	2	4	18	21	2	
4	673	1	3	7	15	10	1	

	sexo_pil	edad_pil	g_edad_pil	...	marca_vehi	edad_m1	\
0	9	999	99	...	-1	9	
1	2	13	2	...	-1	2	
2	1	14	2	...	-1	2	
3	1	15	2	...	-1	2	
4	1	15	2	...	-1	2	

	g_edad_2	g_hora	corre_base	tipo_eve	g_hora_5	g_modelo_veh	\
0	-1	1	-1	99	5	6	
1	-1	2	-1	99	5	6	
2	-1	2	-1	99	5	6	
3	-1	2	-1	99	5	6	
4	-1	2	-1	99	5	6	

	g_edad_80ymás	edad_quinquenales
0	NaN	NaN
1	NaN	NaN
2	NaN	NaN
3	NaN	NaN
4	NaN	NaN

[5 rows x 28 columns]

In [26]:
```
car_accidents_2016.rename(columns={'núm_corre': 'num_hecho'}, inplace=True)
car_accidents_2016.rename(columns={'día_ocu': 'dia_ocu'}, inplace=True)
car_accidents_2016.rename(columns={'día_sem_ocu': 'dia_sem_ocu'}, inplace=True)
car_accidents_2016.rename(columns={'área_geo_ocu': 'areag_ocu'}, inplace=True)
car_accidents_2016.rename(columns={'tipo_veh': 'tipo_vehi'}, inplace=True)
car_accidents_2016.rename(columns={'modelo_veh': 'modelo_vehi'}, inplace=True)
car_accidents_2016.rename(columns={'color_veh': 'color_vehi'}, inplace=True)
```

```
car_accidents_2016.rename(columns={'marca_veh': 'marca_vehi'}, inplace=True)
car_accidents_2016.rename(columns={'mayor_menor': 'edad_m1'}, inplace=True)
car_accidents_2016.rename(columns={'edad_per': 'edad_pil'}, inplace=True)
car_accidents_2016.rename(columns={'estado_con': 'estado_pil'}, inplace=True)
car_accidents_2016.rename(columns={'g_edad_60ymás': 'g_edad_2'}, inplace=True)
car_accidents_2016.rename(columns={'mupio_ocu': 'muni_ocu'}, inplace=True)
car_accidents_2016.rename(columns={'sexo_per': 'sexo_pil'}, inplace=True)
car_accidents_2016.rename(columns={'sexo_per': 'sexo_pil'}, inplace=True)

car_accidents_2016['causa_acc'] = 99
car_accidents_2016['corre_base'] = -1
car_accidents_2016['edad_m1'] = 9
car_accidents_2016['edad_pil'] = 999
car_accidents_2016['estado_pil'] = 9
car_accidents_2016['g_edad_2'] = 12
car_accidents_2016['g_edad_pil'] = -1
car_accidents_2016['sexo_pil'] = 9

del car_accidents_2016['año_ocu']


car_accidents_2009_2016 = pd.concat([car_accidents_2009_2015, car_accidents_2016], so
car_accidents_2009_2016.to_csv('car_accidents_2009_2016.csv', sep=',')
car_accidents_2009_2016.head()
```

Out[26]:

	num_hecho	dia_ocu	mes_ocu	dia_sem_ocu	hora_ocu	depto_ocu	areag_ocu	\
0	1449	13	5	3	1	14	1	
1	1587	1	7	1	15	22	2	
2	679	1	3	7	17	17	2	
3	633	26	2	4	18	21	2	
4	673	1	3	7	15	10	1	

	sexo_pil	edad_pil	g_edad_pil	...	marca_vehi	edad_m1	\
0	9	999	99	...	-1	9	
1	2	13	2	...	-1	2	
2	1	14	2	...	-1	2	
3	1	15	2	...	-1	2	
4	1	15	2	...	-1	2	

	g_edad_2	g_hora	corre_base	tipo_eve	g_hora_5	g_modelo_veh	\
0	-1	1	-1	99	5	6	
1	-1	2	-1	99	5	6	
2	-1	2	-1	99	5	6	
3	-1	2	-1	99	5	6	
4	-1	2	-1	99	5	6	

	g_edad_80ymás	edad_quinquenales
0	NaN	NaN

```
1            NaN                NaN
2            NaN                NaN
3            NaN                NaN
4            NaN                NaN

[5 rows x 28 columns]
```

In [27]:
```
car_accidents_2017.rename(columns={'núm_corre': 'num_hecho'}, inplace=True)
car_accidents_2017.rename(columns={'día_ocu': 'dia_ocu'}, inplace=True)
car_accidents_2017.rename(columns={'día_sem_ocu': 'dia_sem_ocu'}, inplace=True)
car_accidents_2017.rename(columns={'área_geo_ocu': 'areag_ocu'}, inplace=True)
car_accidents_2017.rename(columns={'tipo_veh': 'tipo_vehi'}, inplace=True)
car_accidents_2017.rename(columns={'modelo_veh': 'modelo_vehi'}, inplace=True)
car_accidents_2017.rename(columns={'color_veh': 'color_vehi'}, inplace=True)
car_accidents_2017.rename(columns={'marca_veh': 'marca_vehi'}, inplace=True)
car_accidents_2017.rename(columns={'mayor_menor': 'edad_m1'}, inplace=True)
car_accidents_2017.rename(columns={'edad_per': 'edad_pil'}, inplace=True)
car_accidents_2017.rename(columns={'estado_con': 'estado_pil'}, inplace=True)
car_accidents_2017.rename(columns={'g_edad_60ymás': 'g_edad_2'}, inplace=True)
car_accidents_2017.rename(columns={'mupio_ocu': 'muni_ocu'}, inplace=True)
car_accidents_2017.rename(columns={'sexo_per': 'sexo_pil'}, inplace=True)
car_accidents_2017.rename(columns={'sexo_per': 'sexo_pil'}, inplace=True)



car_accidents_2017['areag_ocu'] = -1
car_accidents_2017['causa_acc'] = 99
car_accidents_2017['corre_base'] = -1
car_accidents_2017['edad_m1'] = 9
car_accidents_2017['edad_pil'] = 999
car_accidents_2017['estado_pil'] = 9
car_accidents_2017['g_edad_2'] = 12
car_accidents_2017['g_edad_pil'] = -1
car_accidents_2017['sexo_pil'] = 9

del car_accidents_2017['año_ocu']


car_accidents_2009_2017 = pd.concat([car_accidents_2009_2016, car_accidents_2017], so
car_accidents_2009_2017.to_csv('car_accidents_2009_2017.csv', sep=',')
car_accidents_2009_2017.head()
```

Out[27]:

	num_hecho	dia_ocu	mes_ocu	dia_sem_ocu	hora_ocu	depto_ocu	areag_ocu	\
0	1449	13	5	3	1	14	1	
1	1587	1	7	1	15	22	2	
2	679	1	3	7	17	17	2	
3	633	26	2	4	18	21	2	
4	673	1	3	7	15	10	1	

```
   sexo_pil  edad_pil  g_edad_pil        ...          marca_vehi  edad_m1  \
0         9       999          99        ...                  -1        9
1         2        13           2        ...                  -1        2
2         1        14           2        ...                  -1        2
3         1        15           2        ...                  -1        2
4         1        15           2        ...                  -1        2

   g_edad_2  g_hora  corre_base  tipo_eve  g_hora_5 g_modelo_veh  \
0        -1       1          -1        99         5             6
1        -1       2          -1        99         5             6
2        -1       2          -1        99         5             6
3        -1       2          -1        99         5             6
4        -1       2          -1        99         5             6

   g_edad_80ymás  edad_quinquenales
0            NaN                NaN
1            NaN                NaN
2            NaN                NaN
3            NaN                NaN
4            NaN                NaN

[5 rows x 28 columns]
```

In [28]:
```
import math
def get_edad_quinquenal(x):
    if x == 999:
        return 18
    elif x >= 80:
        return 17
    return math.floor((x/5)+1)

def g_edad_nuevo(edadExacta):
    if (edadExacta < 15):
        return 1
    if (edadExacta >= 80):
        return 15
    if (edadExacta == 999):
        return 16
    edadExacta = int(((edadExacta) / 5) - 1)
    return edadExacta

del car_accidents_2009_2017['g_edad_80ymás']

car_accidents_2009_2017['edad_quinquenales'] = car_accidents_2009_2017['edad_pil'].ap
car_accidents_2009_2017['g_edad'] = car_accidents_2009_2017['edad_pil'].apply(g_edad_
car_accidents_2009_2017.to_csv('car_accidents_2009_2017.csv', sep=',')
car_accidents_2009_2017.head()
```

Out[28]:

	num_hecho	dia_ocu	mes_ocu	dia_sem_ocu	hora_ocu	depto_ocu	areag_ocu	\
0	1449	13	5	3	1	14	1	
1	1587	1	7	1	15	22	2	
2	679	1	3	7	17	17	2	
3	633	26	2	4	18	21	2	
4	673	1	3	7	15	10	1	

	sexo_pil	edad_pil	g_edad_pil	...	marca_vehi	edad_m1	g_edad_2	\
0	9	999	99	...	-1	9	-1	
1	2	13	2	...	-1	2	-1	
2	1	14	2	...	-1	2	-1	
3	1	15	2	...	-1	2	-1	
4	1	15	2	...	-1	2	-1	

	g_hora	corre_base	tipo_eve	g_hora_5	g_modelo_veh	edad_quinquenales	\
0	1	-1	99	5	6	18	
1	2	-1	99	5	6	3	
2	2	-1	99	5	6	3	
3	2	-1	99	5	6	4	
4	2	-1	99	5	6	4	

	g_edad
0	15
1	1
2	1
3	2
4	2

[5 rows x 28 columns]

3 Vehiculos

In [29]:
```
vehicles_2010 = pd.read_csv('VI2010.csv')
vehicles_2012 = pd.read_csv('VI2012.csv')
vehicles_2013 = pd.read_csv('VI2013.csv')
vehicles_2014 = pd.read_csv('VI2014.csv')
vehicles_2015 = pd.read_csv('VI2015.csv')
vehicles_2016 = pd.read_csv('VI2016.csv')
vehicles_2017 = pd.read_csv('VI2017.csv')
```

In [30]:
```
vehicles_2010['year'] = 2010
vehicles_2012['year'] = 2012
vehicles_2013['year'] = 2013
vehicles_2014['year'] = 2014
vehicles_2015['year'] = 2015
vehicles_2016['year'] = 2016
vehicles_2017['year'] = 2017
```

3.1 Vehicles 2010

```
In [31]: vehicles_2010.rename(columns={'causa_ac': 'causa_acc'}, inplace=True)
         vehicles_2010.rename(columns={'color_v': 'color_vehi'}, inplace=True)
         vehicles_2010.rename(columns={'modelo_v': 'modelo_vehi'}, inplace=True)
         vehicles_2010.rename(columns={'tipo_v': 'tipo_vehi'}, inplace=True)
         vehicles_2010.rename(columns={'condic_pil': 'estado_pil'}, inplace=True)
         vehicles_2010.rename(columns={'mupio_ocu': 'muni_ocu'}, inplace=True)

         vehicles_2010['muni_ocu'] = -1
         vehicles_2010['num_hecho'] = -1
         vehicles_2010['g_edad_pil'] = -1

         def edad_m1(edad_pil):
             if (edad_pil == 999):
                 return 9
             elif (edad_pil >= 18):
                 return 1
             return 2

         vehicles_2010['edad_m1'] = vehicles_2010['edad_pil'].apply(edad_m1)
         vehicles_2010.to_csv('vehicles_2010.csv', sep=',')
         vehicles_2010.head()
```

Out[31]:

	dia_ocu	mes_ocu	dia_sem_ocu	hora_ocu	depto_ocu	areag_ocu	zona_ocu	\
0	31	1	7	15	2	2	99	
1	29	11	1	22	9	2	99	
2	6	12	1	7	13	2	99	
3	8	8	7	17	16	2	99	
4	9	1	6	20	2	2	99	

	sexo_pil	edad_pil	estado_pil	tipo_vehi	color_vehi	modelo_vehi	\
0	9	999	9	5	2	9999	
1	1	37	1	5	7	9999	
2	1	55	1	5	99	9999	
3	1	35	1	3	3	9999	
4	1	24	9	4	4	9999	

	causa_acc	year	muni_ocu	num_hecho	g_edad_pil	edad_m1
0	3	2010	-1	-1	-1	9
1	3	2010	-1	-1	-1	1
2	4	2010	-1	-1	-1	1
3	1	2010	-1	-1	-1	1
4	1	2010	-1	-1	-1	1

```
In [32]: vehicles_2012.head()
```

Out[32]:

	num_vehí	dia_ocu	mes_ocu	dia_sem_ocu	hora_ocu	depto_ocu	mupio_ocu	\
0	1	1	1	7	12	1	101	

```
1         2        1        1            7        12          1        101
2         3        1        1            7        17          1        101
3         4        1        1            7        17          1        101
4         5        1        1            7        10          1        102

   areag_ocu  zona_ocu  sexo_pil  edad_pil  g_edad_pil  edad_m1  condic_pil  \
0          1         3         1        40           7        1           1
1          1         3         1        65          10        1           1
2          1        18         1        44           7        1           2
3          1        18         1       999          11        1           1
4          2        99         1        52           9        1           1

   tipo_vehi  color_vehi  modelo_vehi  causa_acc  year
0          1           6         1999          1  2012
1          1           1         1997          1  2012
2          1           2         9999          1  2012
3          3           5         9999          1  2012
4          3           4         9999          2  2012
```

In [33]:
```
vehicles_2012.rename(columns={'num_vehí': 'num_hecho'}, inplace=True)
vehicles_2012.rename(columns={'causa_ac': 'causa_acc'}, inplace=True)
vehicles_2012.rename(columns={'condic_pil': 'estado_pil'}, inplace=True)
vehicles_2012.rename(columns={'mupio_ocu': 'muni_ocu'}, inplace=True)



vehicles_2010_2012 = pd.concat([vehicles_2010, vehicles_2012], sort=False)
vehicles_2012.to_csv('vehicles_2010_2012.csv', sep=',')
vehicles_2010_2012.head()
```

Out[33]:
```
   dia_ocu  mes_ocu  dia_sem_ocu  hora_ocu  depto_ocu  areag_ocu  zona_ocu  \
0       31        1            7        15          2          2        99
1       29       11            1        22          9          2        99
2        6       12            1         7         13          2        99
3        8        8            7        17         16          2        99
4        9        1            6        20          2          2        99

   sexo_pil  edad_pil  estado_pil  tipo_vehi  color_vehi  modelo_vehi  \
0         9       999           9          5           2         9999
1         1        37           1          5           7         9999
2         1        55           1          5          99         9999
3         1        35           1          3           3         9999
4         1        24           9          4           4         9999

   causa_acc  year  muni_ocu  num_hecho  g_edad_pil  edad_m1
0          3  2010        -1         -1          -1        9
1          3  2010        -1         -1          -1        1
2          4  2010        -1         -1          -1        1
```

```
3          1  2010        -1         -1          -1        1
4          1  2010        -1         -1          -1        1
```

In [34]:
```
vehicles_2013.rename(columns={'num_vehí': 'num_hecho'}, inplace=True)
vehicles_2013.rename(columns={'causa_ac': 'causa_acc'}, inplace=True)
vehicles_2013.rename(columns={'color_veh': 'color_vehi'}, inplace=True)
vehicles_2013.rename(columns={'modelo_veh': 'modelo_vehi'}, inplace=True)
vehicles_2013.rename(columns={'tipo_veh': 'tipo_vehi'}, inplace=True)
vehicles_2013.rename(columns={'condic_pil': 'estado_pil'}, inplace=True)
vehicles_2013.rename(columns={'mupio_ocu': 'muni_ocu'}, inplace=True)
vehicles_2013.rename(columns={'mayor_menor': 'edad_m1'}, inplace=True)
vehicles_2013.rename(columns={'g_edad': 'g_edad_pil'}, inplace=True)

vehicles_2010_2012['marca_veh'] = 99
vehicles_2010_2012['grupo_mode_veh'] = 6

vehicles_2010_2013 = pd.concat([vehicles_2010_2012, vehicles_2013], sort=False)
vehicles_2010_2013.to_csv('vehicles_2010_2013.csv', sep=',')
vehicles_2010_2013.head()
```

Out[34]:

	dia_ocu	mes_ocu	dia_sem_ocu	hora_ocu	depto_ocu	areag_ocu	zona_ocu	\
0	31	1	7	15	2	2	99	
1	29	11	1	22	9	2	99	
2	6	12	1	7	13	2	99	
3	8	8	7	17	16	2	99	
4	9	1	6	20	2	2	99	

	sexo_pil	edad_pil	estado_pil	...	color_vehi	modelo_vehi	\
0	9	999	9	...	2	9999	
1	1	37	1	...	7	9999	
2	1	55	1	...	99	9999	
3	1	35	1	...	3	9999	
4	1	24	9	...	4	9999	

	causa_acc	year	muni_ocu	num_hecho	g_edad_pil	edad_m1	marca_veh	\
0	3	2010	-1	-1	-1	9	99	
1	3	2010	-1	-1	-1	1	99	
2	4	2010	-1	-1	-1	1	99	
3	1	2010	-1	-1	-1	1	99	
4	1	2010	-1	-1	-1	1	99	

	grupo_mode_veh
0	6
1	6
2	6
3	6
4	6

[5 rows x 21 columns]

In [35]:
```
vehicles_2014.rename(columns={'num_correlativo': 'num_hecho'}, inplace=True)
vehicles_2014.rename(columns={'causa_ac': 'causa_acc'}, inplace=True)
vehicles_2014.rename(columns={'color_veh': 'color_vehi'}, inplace=True)
vehicles_2014.rename(columns={'modelo_veh': 'modelo_vehi'}, inplace=True)
vehicles_2014.rename(columns={'tipo_veh': 'tipo_vehi'}, inplace=True)
vehicles_2014.rename(columns={'estado_con': 'estado_pil'}, inplace=True)
vehicles_2014.rename(columns={'mupio_ocu': 'muni_ocu'}, inplace=True)
vehicles_2014.rename(columns={'mayor_menor': 'edad_m1'}, inplace=True)
vehicles_2014.rename(columns={'g_edad': 'g_edad_pil'}, inplace=True)
vehicles_2014.rename(columns={'área_geo_ocu': 'areag_ocu'}, inplace=True)
vehicles_2014.rename(columns={'día_sem_ocu': 'dia_sem_ocu'}, inplace=True)
vehicles_2014.rename(columns={'edad_con': 'edad_pil'}, inplace=True)
vehicles_2014.rename(columns={'día_ocu': 'dia_ocu'}, inplace=True)
vehicles_2014.rename(columns={'sexo_con': 'sexo_pil'}, inplace=True)

vehicles_2014['grupo_mode_veh'] = 6

vehicles_2010_2013['g_hora'] = -1
vehicles_2010_2013['corre_base'] = -1
vehicles_2010_2013['tipo_eve'] = 99


vehicles_2010_2014 = pd.concat([vehicles_2010_2013, vehicles_2014], sort=False)
vehicles_2010_2014.to_csv('vehicles_2010_2014.csv', sep=',')
vehicles_2010_2014.head()
```

Out[35]:

	dia_ocu	mes_ocu	dia_sem_ocu	hora_ocu	depto_ocu	areag_ocu	zona_ocu	\
0	31	1	7	15	2	2	99	
1	29	11	1	22	9	2	99	
2	6	12	1	7	13	2	99	
3	8	8	7	17	16	2	99	
4	9	1	6	20	2	2	99	

	sexo_pil	edad_pil	estado_pil	...	year	muni_ocu	num_hecho	\
0	9	999	9	...	2010	-1	-1	
1	1	37	1	...	2010	-1	-1	
2	1	55	1	...	2010	-1	-1	
3	1	35	1	...	2010	-1	-1	
4	1	24	9	...	2010	-1	-1	

	g_edad_pil	edad_m1	marca_veh	grupo_mode_veh	g_hora	corre_base	\
0	-1	9	99	6	-1	-1	
1	-1	1	99	6	-1	-1	
2	-1	1	99	6	-1	-1	
3	-1	1	99	6	-1	-1	
4	-1	1	99	6	-1	-1	

	tipo_eve

```
0        99
1        99
2        99
3        99
4        99

[5 rows x 24 columns]
```

In [36]:
```
vehicles_2015.rename(columns={'núm_corre': 'num_hecho'}, inplace=True)
vehicles_2015.rename(columns={'color_veh': 'color_vehi'}, inplace=True)
vehicles_2015.rename(columns={'modelo_veh': 'modelo_vehi'}, inplace=True)
vehicles_2015.rename(columns={'tipo_veh': 'tipo_vehi'}, inplace=True)
vehicles_2015.rename(columns={'estado_con': 'estado_pil'}, inplace=True)
vehicles_2015.rename(columns={'mupio_ocu': 'muni_ocu'}, inplace=True)
vehicles_2015.rename(columns={'mayor_menor': 'edad_m1'}, inplace=True)
vehicles_2015.rename(columns={'g_edad': 'g_edad_pil'}, inplace=True)
vehicles_2015.rename(columns={'área_geo_ocu': 'areag_ocu'}, inplace=True)
vehicles_2015.rename(columns={'día_sem_ocu': 'dia_sem_ocu'}, inplace=True)
vehicles_2015.rename(columns={'edad_per': 'edad_pil'}, inplace=True)
vehicles_2015.rename(columns={'día_ocu': 'dia_ocu'}, inplace=True)
vehicles_2015.rename(columns={'sexo_per': 'sexo_pil'}, inplace=True)
vehicles_2015.rename(columns={'g_modelo_veh': 'grupo_mode_veh'}, inplace=True)


vehicles_2015['causa_acc'] = 99
vehicles_2015['corre_base'] = -1

vehicles_2010_2014['g_hora_5'] = -1



vehicles_2010_2015 = pd.concat([vehicles_2010_2014, vehicles_2015], sort=False)
vehicles_2010_2015.to_csv('vehicles_2010_2015.csv', sep=',')
vehicles_2010_2015.head()
```

Out[36]:

	dia_ocu	mes_ocu	dia_sem_ocu	hora_ocu	depto_ocu	areag_ocu	zona_ocu	\
0	31	1	7	15	2	2	99	
1	29	11	1	22	9	2	99	
2	6	12	1	7	13	2	99	
3	8	8	7	17	16	2	99	
4	9	1	6	20	2	2	99	

	sexo_pil	edad_pil	estado_pil	...	edad_m1	marca_veh	\
0	9	999	9	...	9	99	
1	1	37	1	...	1	99	
2	1	55	1	...	1	99	
3	1	35	1	...	1	99	
4	1	24	9	...	1	99	

```
   grupo_mode_veh  g_hora  corre_base  tipo_eve  g_hora_5  g_edad_80ymás  \
0               6      -1          -1        99        -1            NaN
1               6      -1          -1        99        -1            NaN
2               6      -1          -1        99        -1            NaN
3               6      -1          -1        99        -1            NaN
4               6      -1          -1        99        -1            NaN

   g_edad_60ymás  edad_quinquenales
0            NaN                NaN
1            NaN                NaN
2            NaN                NaN
3            NaN                NaN
4            NaN                NaN

[5 rows x 28 columns]
```

In [37]:
```
vehicles_2016.rename(columns={'num_corre': 'num_hecho'}, inplace=True)
vehicles_2016.rename(columns={'color_veh': 'color_vehi'}, inplace=True)
vehicles_2016.rename(columns={'modelo_veh': 'modelo_vehi'}, inplace=True)
vehicles_2016.rename(columns={'tipo_veh': 'tipo_vehi'}, inplace=True)
vehicles_2016.rename(columns={'estado_con': 'estado_pil'}, inplace=True)
vehicles_2016.rename(columns={'mupio_ocu': 'muni_ocu'}, inplace=True)
vehicles_2016.rename(columns={'mayor_menor': 'edad_m1'}, inplace=True)
vehicles_2016.rename(columns={'g_edad': 'g_edad_pil'}, inplace=True)
vehicles_2016.rename(columns={'área_geo_ocu': 'areag_ocu'}, inplace=True)
vehicles_2016.rename(columns={'día_sem_ocu': 'dia_sem_ocu'}, inplace=True)
vehicles_2016.rename(columns={'día_ocu': 'dia_ocu'}, inplace=True)
vehicles_2016.rename(columns={'sexo_per': 'sexo_pil'}, inplace=True)
vehicles_2016.rename(columns={'edad_per': 'edad_pil'}, inplace=True)
vehicles_2016.rename(columns={'g_modelo_veh': 'grupo_mode_veh'}, inplace=True)


vehicles_2016['causa_acc'] = 99
vehicles_2016['corre_base'] = -1



del vehicles_2016['año_ocu']

vehicles_2010_2016 = pd.concat([vehicles_2010_2015, vehicles_2016], sort=False)
vehicles_2010_2016.to_csv('vehicles_2010_2016.csv', sep=',')
vehicles_2010_2016.head()
```

Out[37]:
```
   dia_ocu  mes_ocu  dia_sem_ocu  hora_ocu  depto_ocu  areag_ocu  zona_ocu  \
0       31        1            7        15          2          2        99
1       29       11            1        22          9          2        99
2        6       12            1         7         13          2        99
```

```
3        8        8            7        17         16          2        99
4        9        1            6        20          2          2        99

   sexo_pil  edad_pil  estado_pil        ...          edad_m1  marca_veh  \
0         9       999           9        ...                9         99
1         1        37           1        ...                1         99
2         1        55           1        ...                1         99
3         1        35           1        ...                1         99
4         1        24           9        ...                1         99

   grupo_mode_veh  g_hora  corre_base  tipo_eve  g_hora_5  g_edad_80ymás  \
0               6      -1          -1        99        -1            NaN
1               6      -1          -1        99        -1            NaN
2               6      -1          -1        99        -1            NaN
3               6      -1          -1        99        -1            NaN
4               6      -1          -1        99        -1            NaN

   g_edad_60ymás  edad_quinquenales
0            NaN                NaN
1            NaN                NaN
2            NaN                NaN
3            NaN                NaN
4            NaN                NaN

[5 rows x 28 columns]
```

In [38]:
```
vehicles_2017.rename(columns={'núm_corre': 'num_hecho'}, inplace=True)
vehicles_2017.rename(columns={'color_veh': 'color_vehi'}, inplace=True)
vehicles_2017.rename(columns={'modelo_veh': 'modelo_vehi'}, inplace=True)
vehicles_2017.rename(columns={'tipo_veh': 'tipo_vehi'}, inplace=True)
vehicles_2017.rename(columns={'estado_con': 'estado_pil'}, inplace=True)
vehicles_2017.rename(columns={'mupio_ocu': 'muni_ocu'}, inplace=True)
vehicles_2017.rename(columns={'mayor_menor': 'edad_m1'}, inplace=True)
vehicles_2017.rename(columns={'g_edad': 'g_edad_pil'}, inplace=True)
vehicles_2017.rename(columns={'área_geo_ocu': 'areag_ocu'}, inplace=True)
vehicles_2017.rename(columns={'día_sem_ocu': 'dia_sem_ocu'}, inplace=True)
vehicles_2017.rename(columns={'día_ocu': 'dia_ocu'}, inplace=True)
vehicles_2017.rename(columns={'sexo_per': 'sexo_pil'}, inplace=True)
vehicles_2017.rename(columns={'edad_per': 'edad_pil'}, inplace=True)
vehicles_2017.rename(columns={'g_modelo_veh': 'grupo_mode_veh'}, inplace=True)


vehicles_2017['causa_acc'] = 99
vehicles_2017['corre_base'] = -1



del vehicles_2017['año_ocu']
```

```
vehicles_2010_2017 = pd.concat([vehicles_2010_2016, vehicles_2017], sort=False)
vehicles_2010_2017.to_csv('vehicles_2010_2017.csv', sep=',')
vehicles_2010_2017.head()
```

Out[38]:

	dia_ocu	mes_ocu	dia_sem_ocu	hora_ocu	depto_ocu	areag_ocu	zona_ocu	\
0	31	1	7	15	2	2.0	99	
1	29	11	1	22	9	2.0	99	
2	6	12	1	7	13	2.0	99	
3	8	8	7	17	16	2.0	99	
4	9	1	6	20	2	2.0	99	

	sexo_pil	edad_pil	estado_pil	...	edad_m1	marca_veh	\
0	9	999	9	...	9	99	
1	1	37	1	...	1	99	
2	1	55	1	...	1	99	
3	1	35	1	...	1	99	
4	1	24	9	...	1	99	

	grupo_mode_veh	g_hora	corre_base	tipo_eve	g_hora_5	g_edad_80ymás	\
0	6	-1	-1	99	-1	NaN	
1	6	-1	-1	99	-1	NaN	
2	6	-1	-1	99	-1	NaN	
3	6	-1	-1	99	-1	NaN	
4	6	-1	-1	99	-1	NaN	

	g_edad_60ymás	edad_quinquenales
0	NaN	NaN
1	NaN	NaN
2	NaN	NaN
3	NaN	NaN
4	NaN	NaN

[5 rows x 28 columns]

In [39]:
```
del vehicles_2010_2017['g_edad_80ymás']
del vehicles_2010_2017['g_edad_60ymás']
del vehicles_2010_2017['edad_quinquenales']


def g_edad_nuevo(edadExacta):
    if (edadExacta < 15):
        return 1
    if (edadExacta >= 80):
        return 15
    if (edadExacta == 999):
        return 16
    edadExacta = int(((edadExacta) / 5) - 1)
```

```
    return edadExacta


vehicles_2010_2017['g_edad_n'] = vehicles_2010_2017['edad_pil'].apply(g_edad_nuevo)
vehicles_2010_2017.to_csv('vehicles_2010_2017.csv', sep=',')
vehicles_2010_2017.head()
```

Out[39]:

	dia_ocu	mes_ocu	dia_sem_ocu	hora_ocu	depto_ocu	areag_ocu	zona_ocu	\
0	31	1	7	15	2	2.0	99	
1	29	11	1	22	9	2.0	99	
2	6	12	1	7	13	2.0	99	
3	8	8	7	17	16	2.0	99	
4	9	1	6	20	2	2.0	99	

	sexo_pil	edad_pil	estado_pil	...	num_hecho	g_edad_pil	edad_m1	\
0	9	999	9	...	-1	-1.0	9	
1	1	37	1	...	-1	-1.0	1	
2	1	55	1	...	-1	-1.0	1	
3	1	35	1	...	-1	-1.0	1	
4	1	24	9	...	-1	-1.0	1	

	marca_veh	grupo_mode_veh	g_hora	corre_base	tipo_eve	g_hora_5	g_edad_n
0	99	6	-1	-1	99	-1	15
1	99	6	-1	-1	99	-1	6
2	99	6	-1	-1	99	-1	10
3	99	6	-1	-1	99	-1	6
4	99	6	-1	-1	99	-1	3

[5 rows x 26 columns]