Mapas de calor y boxplots de base de datos

Integrantes del equipo:

Carlos Sánchez Mejorada Raynal A01702188

María de los Angeles Arista Huerta A01369984

Ariann Fernando Arriaga Alcántara A01703556

In [1]:
```
import pandas as pd
import seaborn as sb
import numpy as np; np.random.seed(0)
import matplotlib.pyplot as plt
data=pd.read_csv('heart_failure_clinical_records_dataset.csv')
data.shape
```

Out[1]:
```
(299, 13)
```

In [3]:
```
data.head
```

Out[3]:
```
<bound method NDFrame.head of       age  anaemia  creatinine_phosphokinase  diabetes  ej
ection_fraction  \
0    75.0        0                       582         0                 20
1    55.0        0                      7861         0                 38
2    65.0        0                       146         0                 20
3    50.0        1                       111         0                 20
4    65.0        1                       160         1                 20
..    ...      ...                       ...       ...                ...
294  62.0        0                        61         1                 38
295  55.0        0                      1820         0                 38
296  45.0        0                      2060         1                 60
297  45.0        0                      2413         0                 38
298  50.0        0                       196         0                 45

     high_blood_pressure  platelets  serum_creatinine  serum_sodium  sex  \
0                      1  265000.00               1.9           130    1
1                      0  263358.03               1.1           136    1
2                      0  162000.00               1.3           129    1
3                      0  210000.00               1.9           137    1
4                      0  327000.00               2.7           116    0
..                   ...        ...               ...           ...  ...
294                    1  155000.00               1.1           143    1
295                    0  270000.00               1.2           139    0
296                    0  742000.00               0.8           138    0
297                    0  140000.00               1.4           140    1
298                    0  395000.00               1.6           136    1

     smoking  time  DEATH_EVENT
0          0     4            1
1          0     6            1
2          1     7            1
3          0     7            1
4          0     8            1
..       ...   ...          ...
294        1   270            0
295        0   271            0
```

```
296        0   278           0
297        1   280           0
298        1   285           0

[299 rows x 13 columns]>
```

Descripción estadística de los datos

In [4]:
```
data.describe()
```

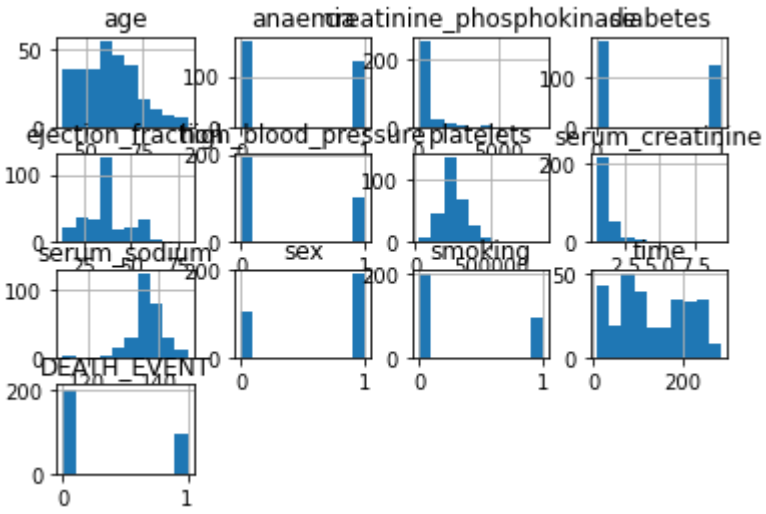
Out[4]:

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressu
count	299.000000	299.000000	299.000000	299.000000	299.000000	299.0000
mean	60.833893	0.431438	581.839465	0.418060	38.083612	0.351
std	11.894809	0.496107	970.287881	0.494067	11.834841	0.478
min	40.000000	0.000000	23.000000	0.000000	14.000000	0.0000
25%	51.000000	0.000000	116.500000	0.000000	30.000000	0.0000
50%	60.000000	0.000000	250.000000	0.000000	38.000000	0.0000
75%	70.000000	1.000000	582.000000	1.000000	45.000000	1.0000
max	95.000000	1.000000	7861.000000	1.000000	80.000000	1.0000

Histograma de los campos

In [5]:
```
data.drop([0,1]).hist()
```

Out[5]:
```
array([[<AxesSubplot:title={'center':'age'}>,
        <AxesSubplot:title={'center':'anaemia'}>,
        <AxesSubplot:title={'center':'creatinine_phosphokinase'}>,
        <AxesSubplot:title={'center':'diabetes'}>],
       [<AxesSubplot:title={'center':'ejection_fraction'}>,
        <AxesSubplot:title={'center':'high_blood_pressure'}>,
        <AxesSubplot:title={'center':'platelets'}>,
        <AxesSubplot:title={'center':'serum_creatinine'}>],
       [<AxesSubplot:title={'center':'serum_sodium'}>,
        <AxesSubplot:title={'center':'sex'}>,
        <AxesSubplot:title={'center':'smoking'}>,
        <AxesSubplot:title={'center':'time'}>],
       [<AxesSubplot:title={'center':'DEATH_EVENT'}>, <AxesSubplot:>,
        <AxesSubplot:>, <AxesSubplot:>]], dtype=object)
```



Correlación de los datos por método de pearson

In [6]:
```
data.corr(method='pearson')
```

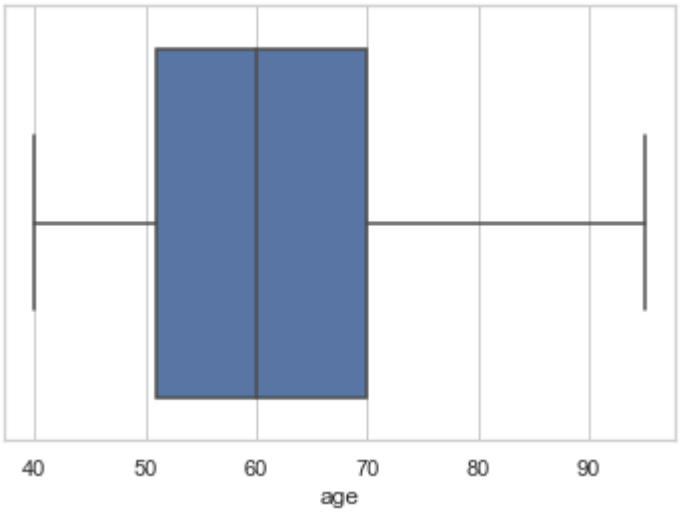
Out[6]:

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	hi
age	1.000000	0.088006	-0.081584	-0.101012	0.060098	
anaemia	0.088006	1.000000	-0.190741	-0.012729	0.031557	
creatinine_phosphokinase	-0.081584	-0.190741	1.000000	-0.009639	-0.044080	
diabetes	-0.101012	-0.012729	-0.009639	1.000000	-0.004850	
ejection_fraction	0.060098	0.031557	-0.044080	-0.004850	1.000000	
high_blood_pressure	0.093289	0.038182	-0.070590	-0.012732	0.024445	
platelets	-0.052354	-0.043786	0.024463	0.092193	0.072177	
serum_creatinine	0.159187	0.052174	-0.016408	-0.046975	-0.011302	
serum_sodium	-0.045966	0.041882	0.059550	-0.089551	0.175902	
sex	0.065430	-0.094769	0.079791	-0.157730	-0.148386	
smoking	0.018668	-0.107290	0.002421	-0.147173	-0.067315	
time	-0.224068	-0.141414	-0.009346	0.033726	0.041729	
DEATH_EVENT	0.253729	0.066270	0.062728	-0.001943	-0.268603	

Gráfica de caja y bigotes

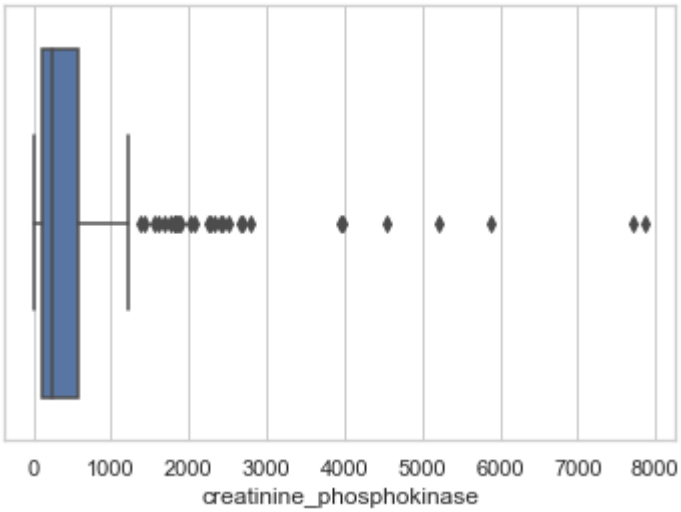
Edad

In [8]:
```
sb.set_theme(style="whitegrid")
ax=sb.boxplot(x=data["age"])
```



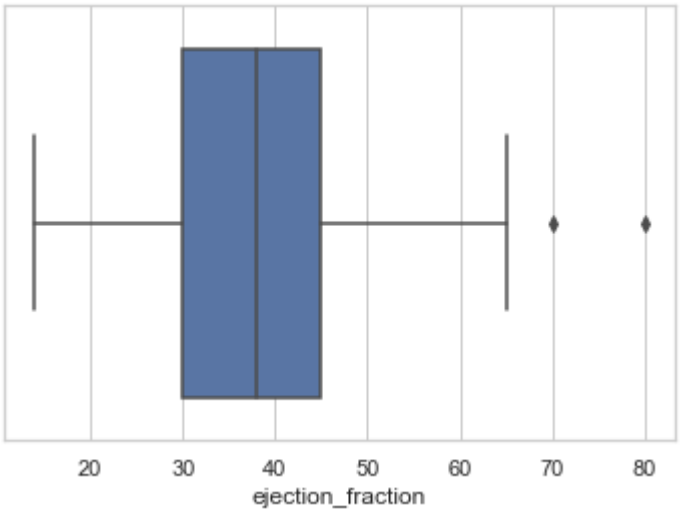
creatinine phosphokinase

In [9]:
```
sb.set_theme(style="whitegrid")
ax=sb.boxplot(x=data["creatinine_phosphokinase"])
```



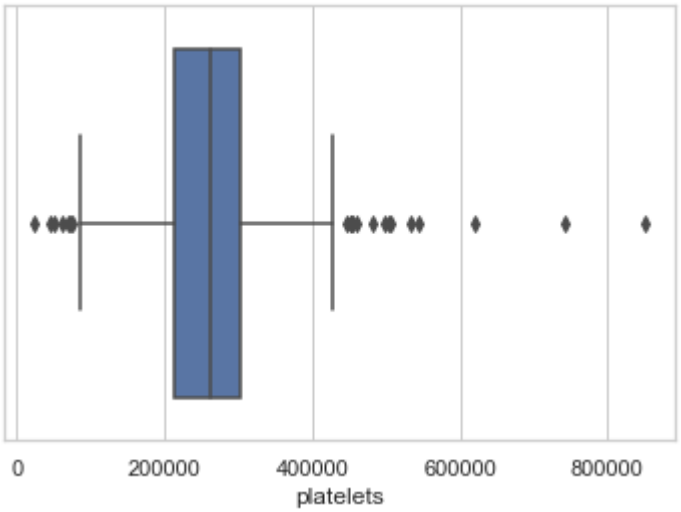
Fracción de eyección

In [12]:
```
sb.set_theme(style="whitegrid")
ax=sb.boxplot(x=data["ejection_fraction"])
```



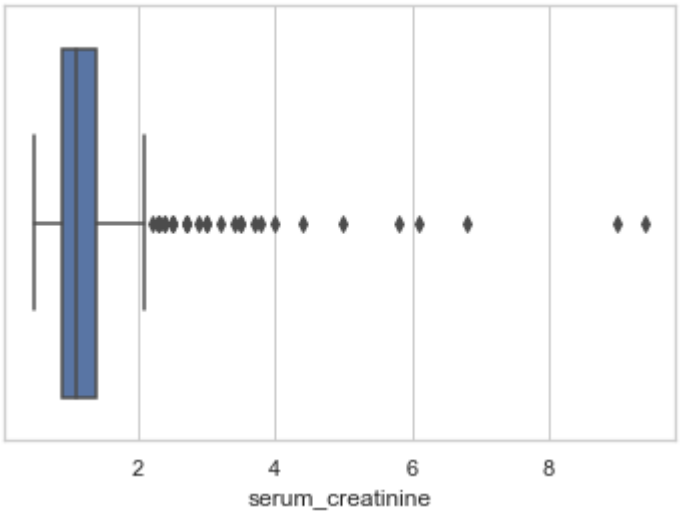
Plaquetas

In [13]:
```
sb.set_theme(style="whitegrid")
ax=sb.boxplot(x=data["platelets"])
```



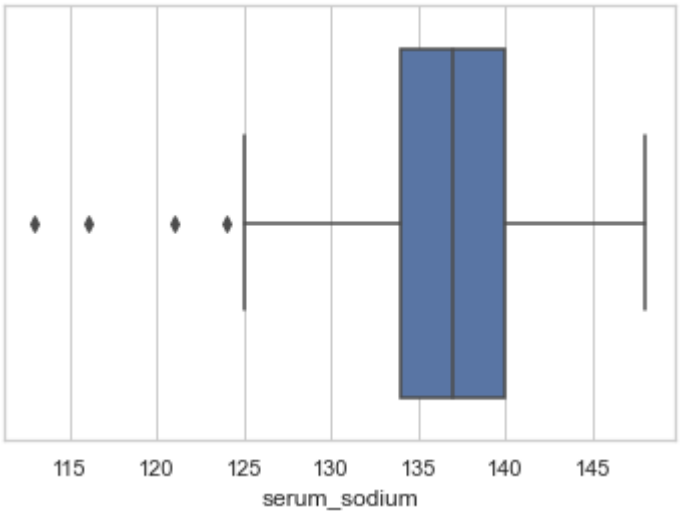
Suero de creatina

In [14]:
```
sb.set_theme(style="whitegrid")
ax=sb.boxplot(x=data["serum_creatinine"])
```



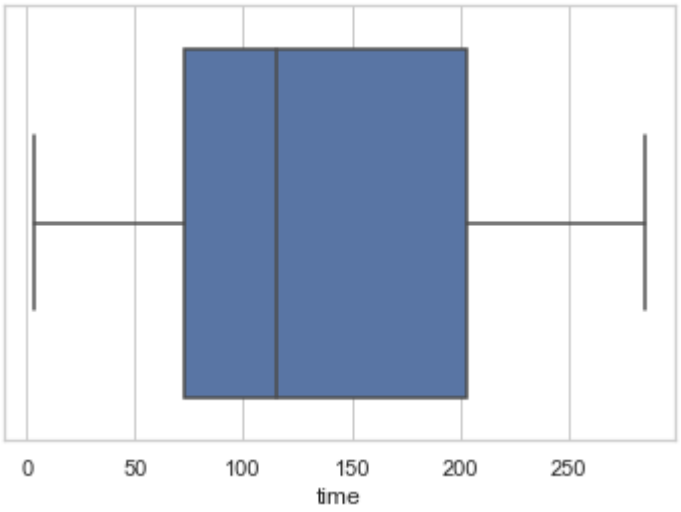
Suero de sodio

In [18]:
```
sb.set_theme(style="whitegrid")
ax=sb.boxplot(x=data["serum_sodium"])
```



Tiempo de seguimiento del tratamiento

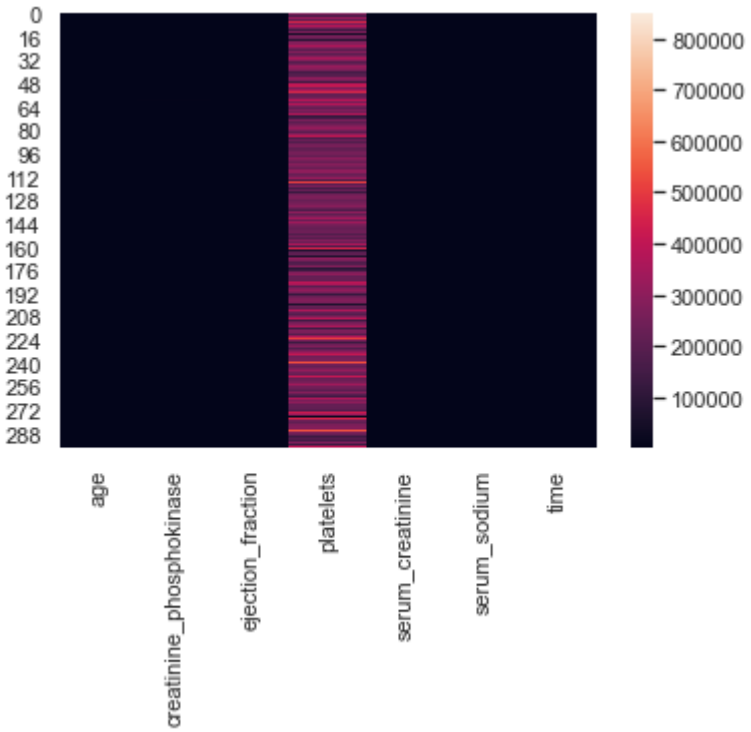
In [17]:
```
sb.set_theme(style="whitegrid")
ax=sb.boxplot(x=data["time"])
```



Mapas de calor

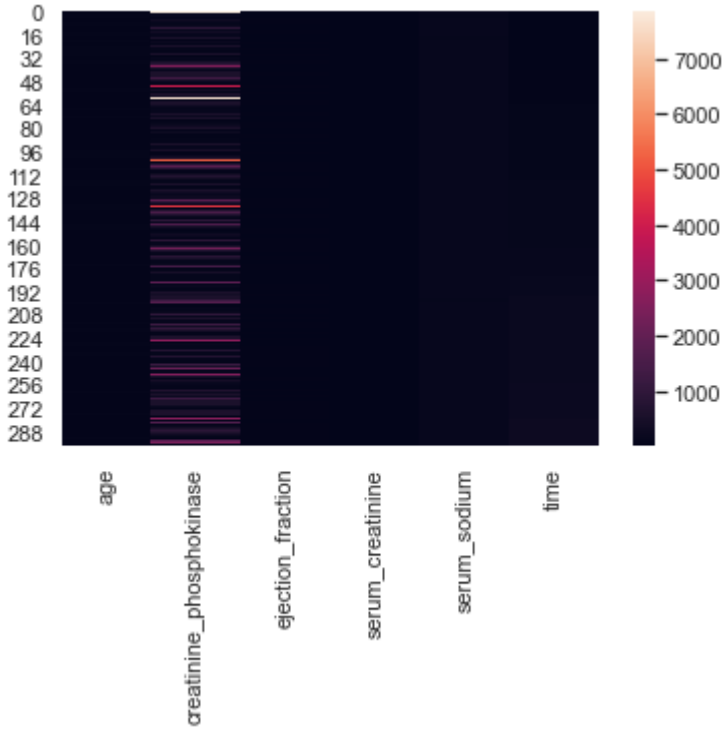
Con plaquetas

In [55]:
```
Heart_health = pd.read_csv('heart_failure_clinical_records_dataset_new.csv')
ax=sb.heatmap(Heart_health)
```



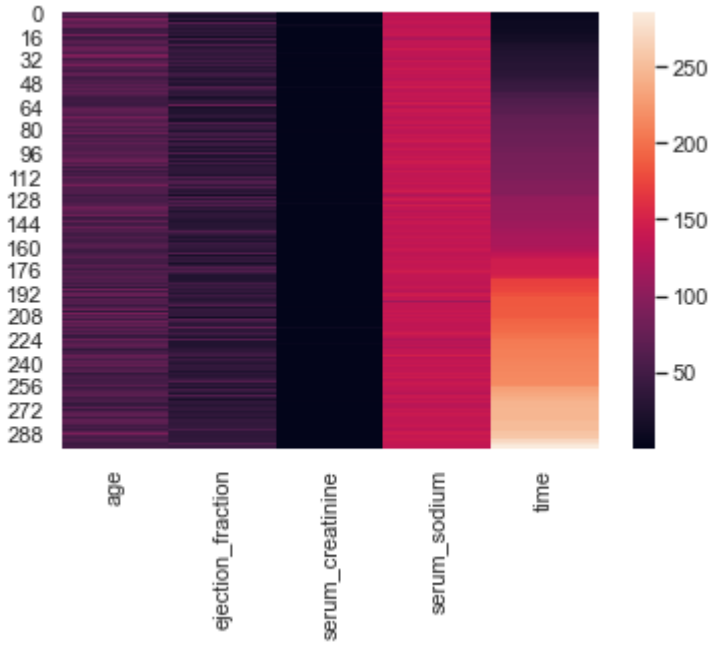
Sin plaquetas

In [21]:
```
Heart_health2 = pd.read_csv('heart_failure_clinical_records_dataset_new_2.csv')
ax=sb.heatmap(Heart_health2)
```



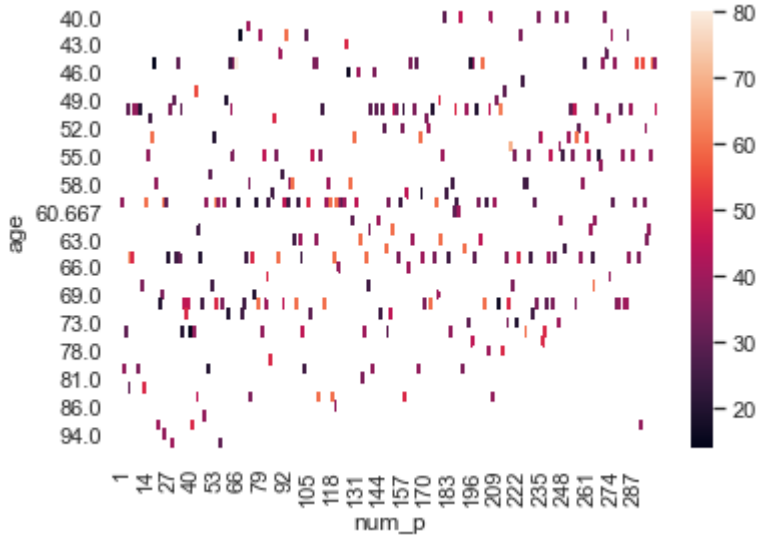
Sin creatina

In [23]:
```
Heart_health3 = pd.read_csv('heart_failure_clinical_records_dataset_new_3.csv')
ax=sb.heatmap(Heart_health3)
```



Comparación de la edad de los pacientes con su número de fracción de eyección

In [63]:
```
Heart_H=pd.read_csv('heart_failure_clinical_records_dataset_new_4.csv')
Heart_H= Heart_H.pivot('age','num_p','ejection_fraction')
ax=sb.heatmap(Heart_H)
```



Preguntas detonadoras de análisis

¿Hay alguna variable que no aporta información? Todas las variables booleanas su aportación es relativamente nula ya que tienen una clasificación bastante generalizada la cual se categoriza en dos secciones 1 u 0 y pues en nuestro analisis su valor informativo es realmente bajo. Estas siendo:anaemia, high_blood_pressure,sex,DEATH_EVENT Cabe aclarar que apesar de que si hay un pequeño aporte de información por parte de estas variable, su utilidad para un análisis de datos es casi nulo.

Si tuvieras que eliminar variables, ¿cuáles quitarías y por qué? Ceatinine Phosphokinase: dispersión de datos y valores atípicos, poca correlación con los datos Plaquetas: dispersión de datos y valores atípicos Sexo: valores atípicos Todas estas variables fueron seleccionadas para poder ser una opción para eliminar debido a que generan o tienen un alto impacto en la fiabilidad de nuestro análisis. Si bien es importante agregar que al tener cada una de estas variables tiene una alta dispersión y varios valores atípicos podemos decir que nuestro proceso es altamente afectado

¿Existen variables que tengan datos extraños? las variables anaemia, high_blood_pressure,sex,DEATH_EVENT,diabetes,smoking presentan valores booleanos, lo cual significa que no pueden ser usadas para nuestra evaluación de los datos

Si comparas las variables, ¿todas están en rangos similares? ¿Crees que esto afecte? No, si existe una variabilidad en los rangos de las variables y si consideramos que esto puede afectar nuestro proceso porque si bien dicho proceso tiene un rango fijo y al cada variable tener su propio rango esto genera cierta inestabiliedad en este mismo. Aunque si dividimos las variables por dos clasificaciones entonces si comparten rangos comunes, las dos clasificaciones son datos mayores a 2000 y los datos menores a 1000.

¿Puedes encontrar grupos qué se parezcan? ¿Qué grupos son estos? Si, existen dos tipos de grupos en como se clasifican los datos. Tenemos los datos que se clasifican entre 0 y 1 y los datos que se clasifican en función a como transcurre el proceso. Y dentre de estos datos están las clasificaciones previamente mencionadas donde sus datos superaban o no las 2000 unidades.