set up the envirement:

In [ ]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
plt.style.use("seaborn-whitegrid")
plt.rc("figure", autolayout=True)
plt.rc(
    "axes",
    labelweight="bold",
    labelsize="large",
    titleweight="bold",
    titlesize=14,
    titlepad=10,
)
```

In [ ]:

```
table_DS = pd.read_excel('./Adherent_AYD_ALD_PEC_DS.xlsx',sheet_name='Adherent_DS')
table_PEC = pd.read_excel('./Adherent_AYD_ALD_PEC_DS.xlsx',sheet_name='Adherent_PEC')
```

In [ ]:

```
table_DS.head()
```

Out[ ]:

	Année	Mois	Matricule	Date_naissance	Genre	ville	CJT	Enf	DS	Statut_ad
0	2017	1	10016	1955-12-26	Masculin	AGADIR	1	2	1	Vali
1	2017	5	10018	1959-05-23	Masculin	CASABLANCA	1	2	3	Vali
2	2017	6	10018	1959-05-23	Masculin	CASABLANCA	1	2	1	Vali
3	2017	1	1004	1954-06-30	Masculin	CASABLANCA	1	2	1	Vali
4	2017	2	1004	1954-06-30	Masculin	CASABLANCA	1	2	1	Vali

In [ ]:

```
table_PEC.head()
```

Out[ ]:

	Année	Mois	Matricule	CJT	Enf	ALD	PEC	MT_Devis	MT_PEC	Max_date_prest	ex
0	2017	3	10016	1	2	1	1	57696.92	28848.46	2017-03-22	100
1	2017	6	10016	1	2	1	1	1072170.00	56430.00	2017-06-12	100
2	2017	7	10016	1	2	1	1	38019.76	19009.88	2017-07-21	100
3	2017	11	10016	1	2	1	2	92983.72	74991.86	2017-11-30	100
4	2017	12	10016	1	2	1	1	1379400.00	62700.00	2017-12-15	100

In [ ]:

```
table_DS.drop(columns='existe',inplace=True)
table_PEC.drop(columns='existe',inplace=True)
```

In [ ]:

```
table_DS.rename(columns={'Année':'Annee'},inplace=True)
table_PEC.rename(columns={'Année':'Annee'},inplace=True)
```

In [ ]:

```
table_DS.ville.fillna('UNKOWN',inplace=True)
```

In [ ]:

```
table_DS.shape
```

Out[ ]:

```
(32746, 13)
```

the progration over time of:

- charges (Mt_rem).
- Mt_remb
- number of request.

In [ ]:

```
def func_by_month_every_year_feature(df,feature,func):
    Mt_eng_ds_over_time = pd.DataFrame()
    Mt_eng_ds_over_time['2017'] = df[table_DS.Annee == 2017].groupby(['Mois'])[feature]
.agg(func)
    Mt_eng_ds_over_time['2018'] = df[table_DS.Annee == 2018].groupby(['Mois'])[feature]
.agg(func)
    Mt_eng_ds_over_time['2019'] = df[table_DS.Annee == 2019].groupby(['Mois'])[feature]
.agg(func)
    Mt_eng_ds_over_time['2020'] = df[table_DS.Annee == 2020].groupby(['Mois'])[feature]
.agg(func)
    Mt_eng_ds_over_time['2021'] = df[table_DS.Annee == 2021].groupby(['Mois'])[feature]
.agg(func)
    return Mt_eng_ds_over_time
```

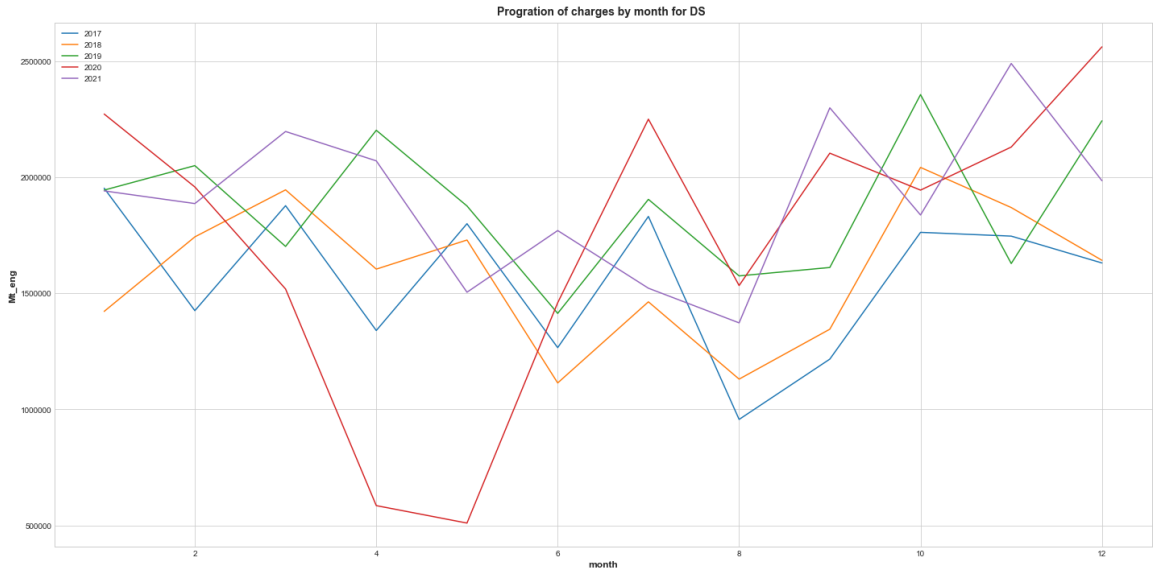
In [ ]:

```
Mt_eng_ds_over_time = func_by_month_every_year_feature(table_DS,'Mt_remb',sum)
```

In [ ]:

```
plt.figure(figsize=(20,10))
plt.title('Progration of charges by month for DS ')
for i in range(2017,2022):
    sns.lineplot(data=Mt_eng_ds_over_time[str(i)],label=str(i))

plt.ylabel('Mt_eng')
plt.xlabel('month')
plt.show()
```



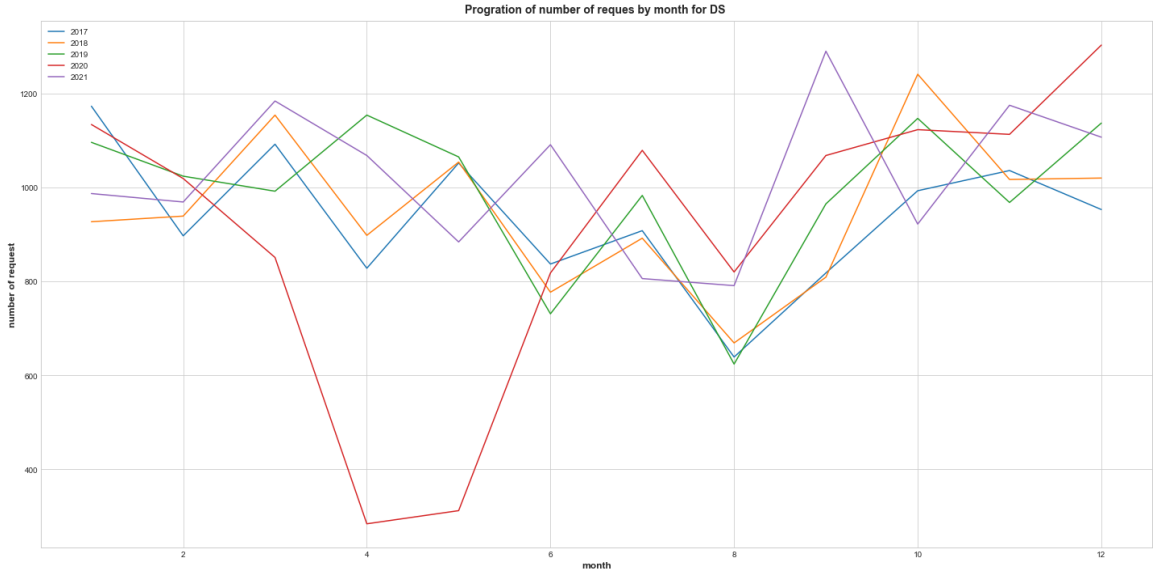
In [ ]:

```
number_of_ds_overtime = func_by_month_every_year_feature(table_DS,'DS',sum)
```

In [ ]:

```
plt.figure(figsize=(20,10))
plt.title('Progration of number of reques by month for DS ')
for i in range(2017,2022):
    sns.lineplot(data=number_of_ds_overtime[str(i)],label=str(i))

plt.ylabel('number of request')
plt.xlabel('month')
plt.show()
```



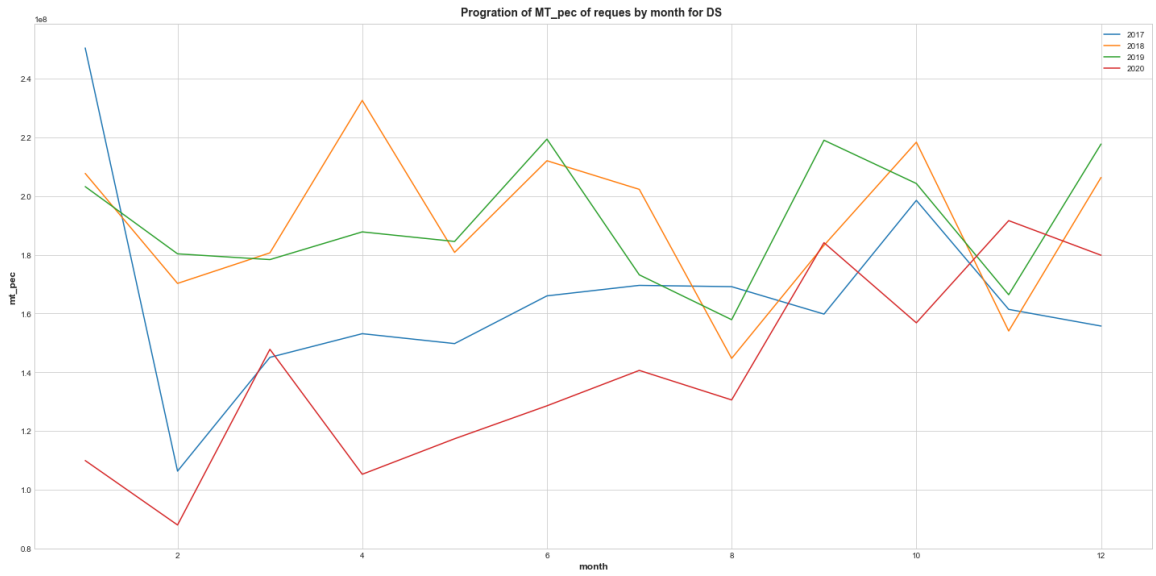
In [ ]:

```
Mt_eng_pec_over_time = func_by_month_every_year_feature(table_PEC,'MT_PEC',sum)
```

In [ ]:

```
plt.figure(figsize=(20,10))
plt.title('Progration of MT_pec of reques by month for DS ')
for i in range(2017,2022):
    sns.lineplot(data=Mt_eng_pec_over_time[str(i)],label=str(i))

plt.ylabel('mt_pec')
plt.xlabel('month')
plt.show()
```



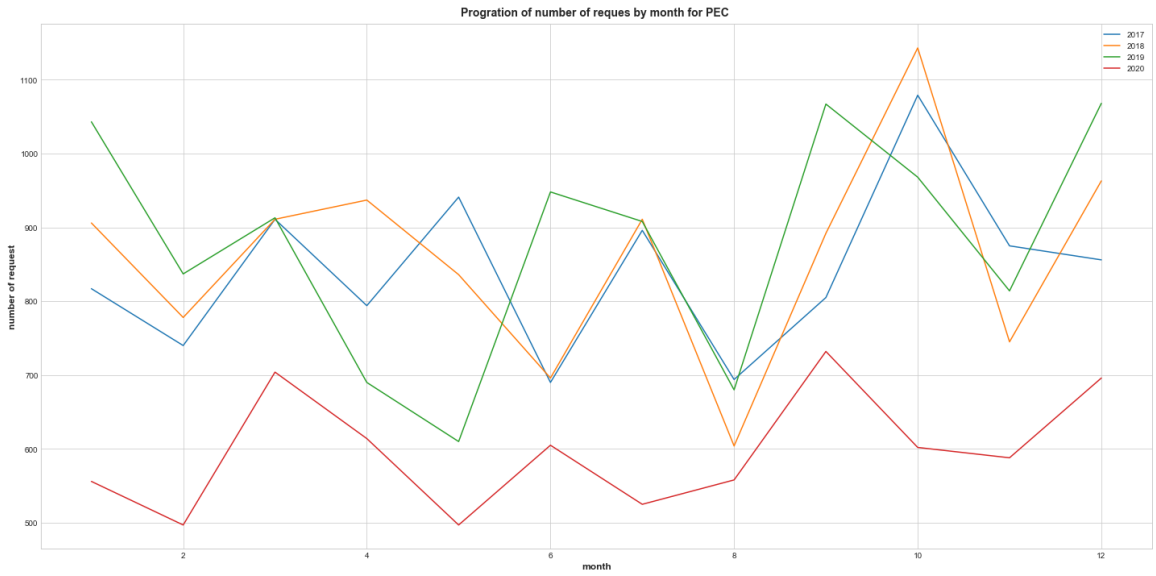
In [ ]:

```
number_of_pec_over_time = func_by_month_every_year_feature(table_PEC,'PEC',sum)
```

In [ ]:

```
plt.figure(figsize=(20,10))
plt.title('Progration of number of reques by month for PEC ')
for i in range(2017,2022):
    sns.lineplot(data=number_of_pec_over_time[str(i)],label=str(i))

plt.ylabel('number of request')
plt.xlabel('month')
plt.show()
```



In [ ]:

```
table_DS.head()
```

Out[ ]:

	Annee	Mois	Matricule	Date_naissance	Genre	ville	CJT	Enf	DS	Statut_ac
0	2017	1	10016	1955-12-26	Masculin	AGADIR	1	2	1	Vali
1	2017	5	10018	1959-05-23	Masculin	CASABLANCA	1	2	3	Vali
2	2017	6	10018	1959-05-23	Masculin	CASABLANCA	1	2	1	Vali
3	2017	1	1004	1954-06-30	Masculin	CASABLANCA	1	2	1	Vali
4	2017	2	1004	1954-06-30	Masculin	CASABLANCA	1	2	1	Vali

In [ ]:

```
table_DS.dtypes
```

Out[ ]:

```
Annee                      int64
Mois                       int64
Matricule                  int64
Date_naissance    datetime64[ns]
Genre                     object
ville                     object
CJT                        int64
Enf                        int64
DS                         int64
Statut_adh                object
Mt_eng                   float64
Mt_remb                  float64
Max_date_depot    datetime64[ns]
dtype: object
```

In [ ]:

```
table_DS['Age'] = (table_DS.Max_date_depot - table_DS.Date_naissance)/np.timedelta64(1,
'Y')
table_DS.drop(columns=['Date_naissance','Max_date_depot'],inplace=True)
```

In [ ]:

```
table_DS.head()
```

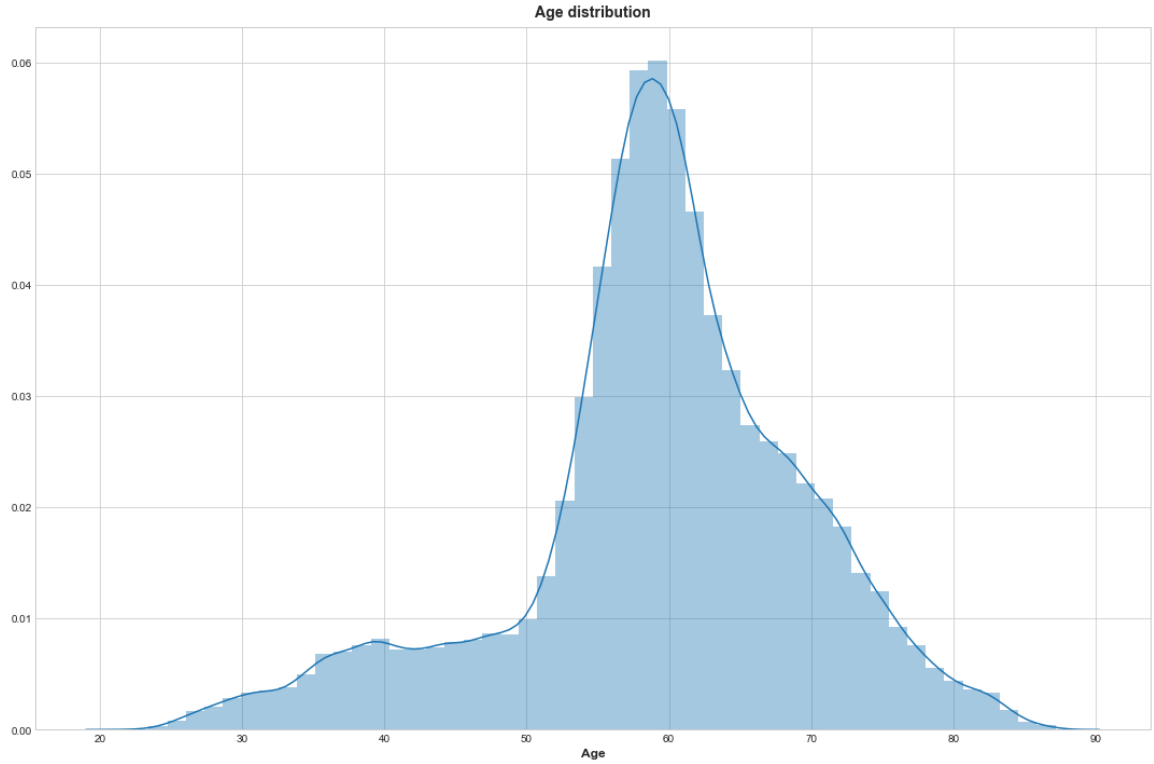
Out[ ]:

	Annee	Mois	Matricule	Genre	ville	CJT	Enf	DS	Statut_adh	Mt_eng	Mt_
0	2017	1	10016	Masculin	AGADIR	1	2	1	Validé	16000.0	40
1	2017	5	10018	Masculin	CASABLANCA	1	2	3	Validé	4116.4	29
2	2017	6	10018	Masculin	CASABLANCA	1	2	1	Validé	640.0	1
3	2017	1	1004	Masculin	CASABLANCA	1	2	1	Validé	5834.2	23
4	2017	2	1004	Masculin	CASABLANCA	1	2	1	Validé	3358.8	22

age distribution:

In [ ]:

```
plt.figure(figsize=(15,10))
plt.title('Age distribution')
sns.distplot(table_DS['Age'])
plt.show()
```



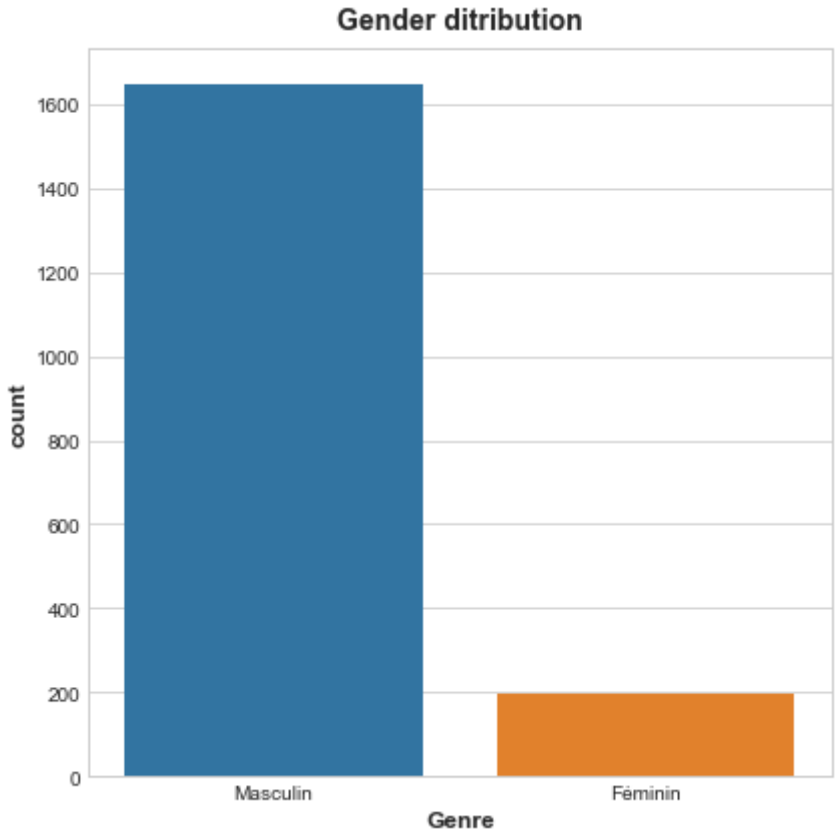
gender distribution

In [ ]:

```
filter_by_gender = table_DS[['Matricule','Genre']].drop_duplicates()
```

In [ ]:

```
plt.figure(figsize=(6,6))
plt.title('Gender ditribution')
sns.countplot(x='Genre',data=filter_by_gender)
plt.show()
```



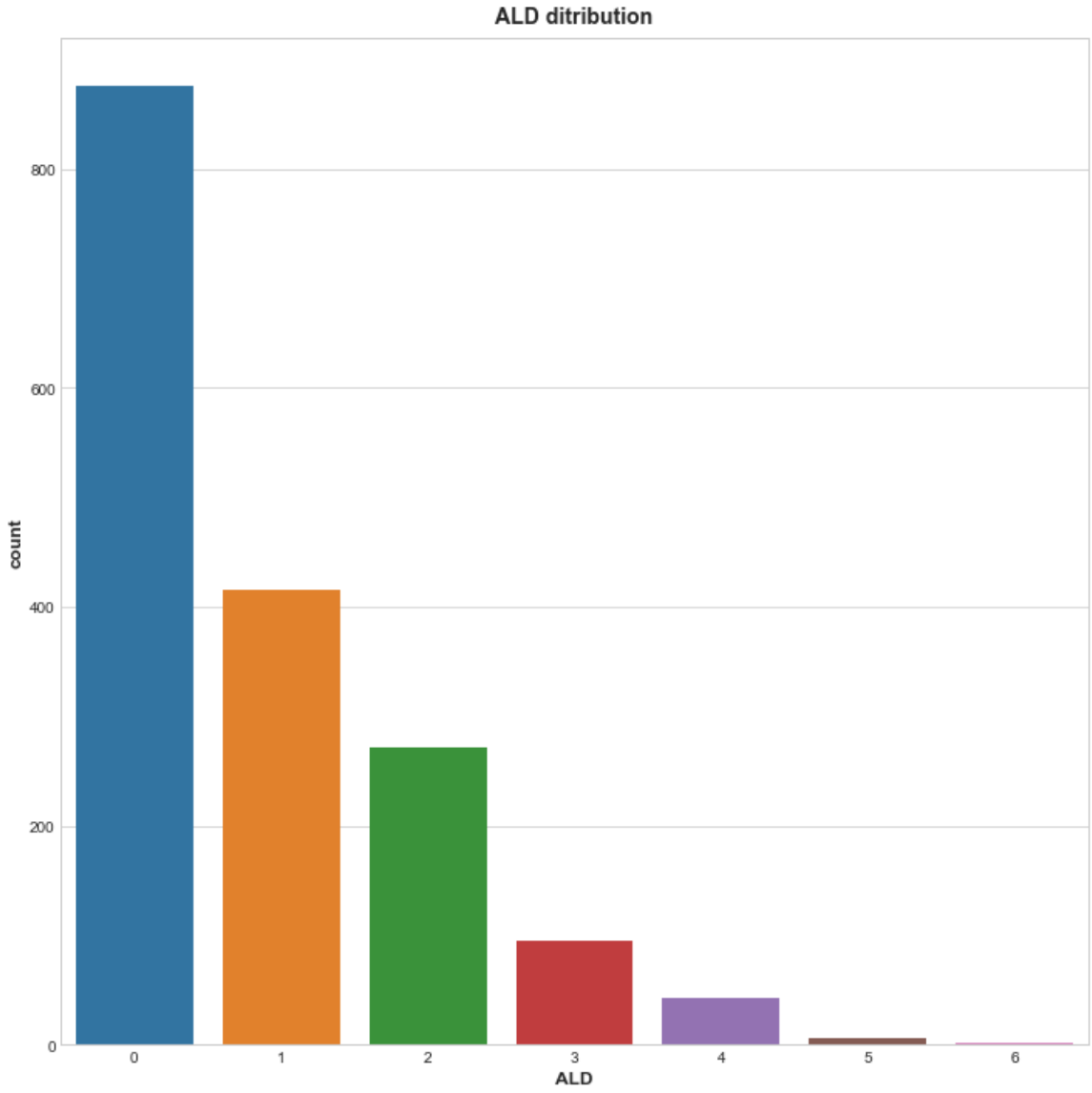
ALD distribution:

In [ ]:

```
filter_by_ALD = table_PEC[['Matricule','ALD']].drop_duplicates()
```

In [ ]:

```
plt.figure(figsize=(10,10))
plt.title('ALD ditribution')
sns.countplot(x='ALD',data=filter_by_ALD)
plt.show()
```



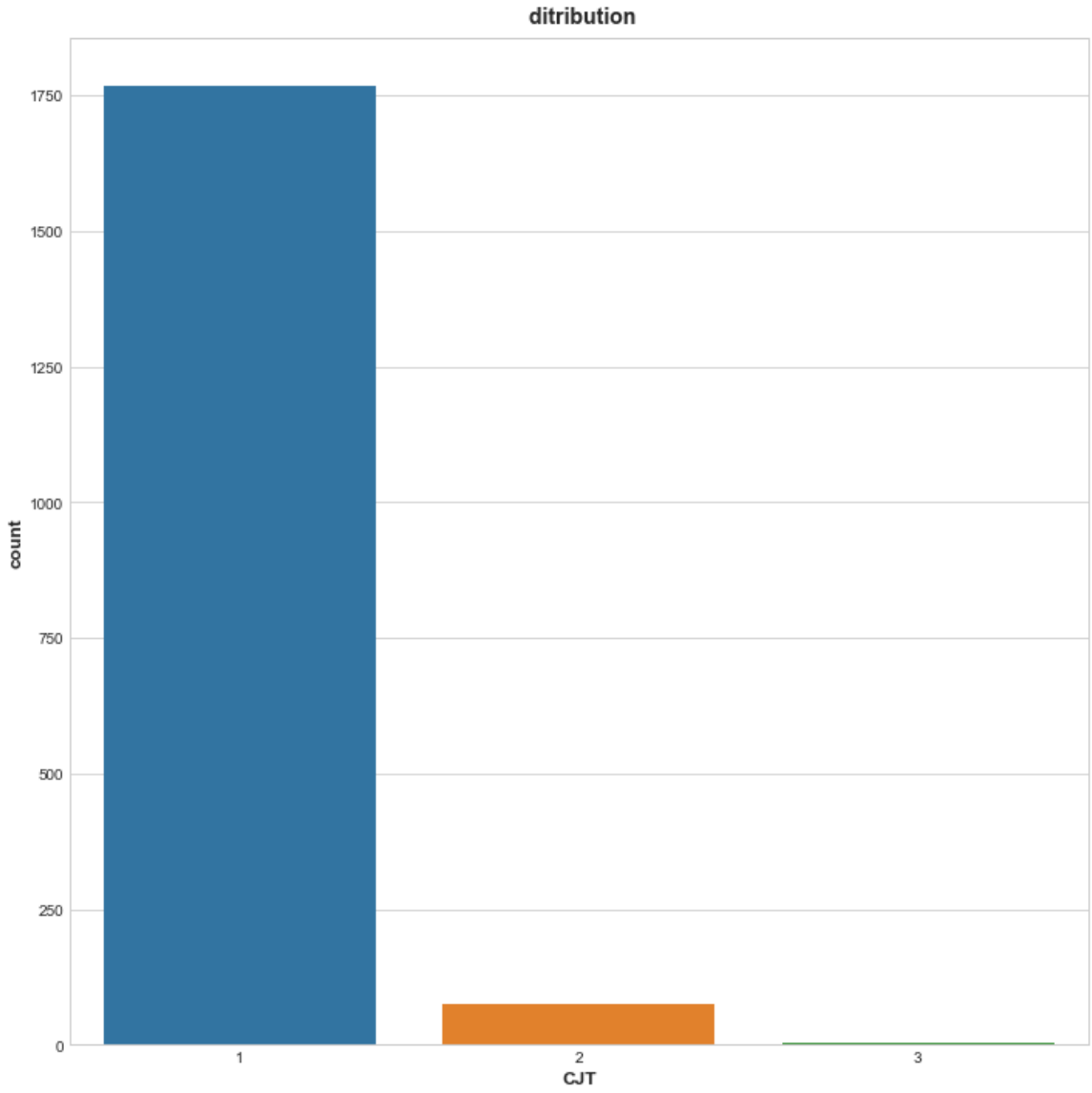
Maried /not Maried

In [ ]:

```
filter_by_CJT = table_DS[['Matricule','CJT']].drop_duplicates()
```

In [ ]:

```
plt.figure(figsize=(10,10))
plt.title(' ditribution')
sns.countplot(x='CJT',data=filter_by_CJT)
plt.show()
```



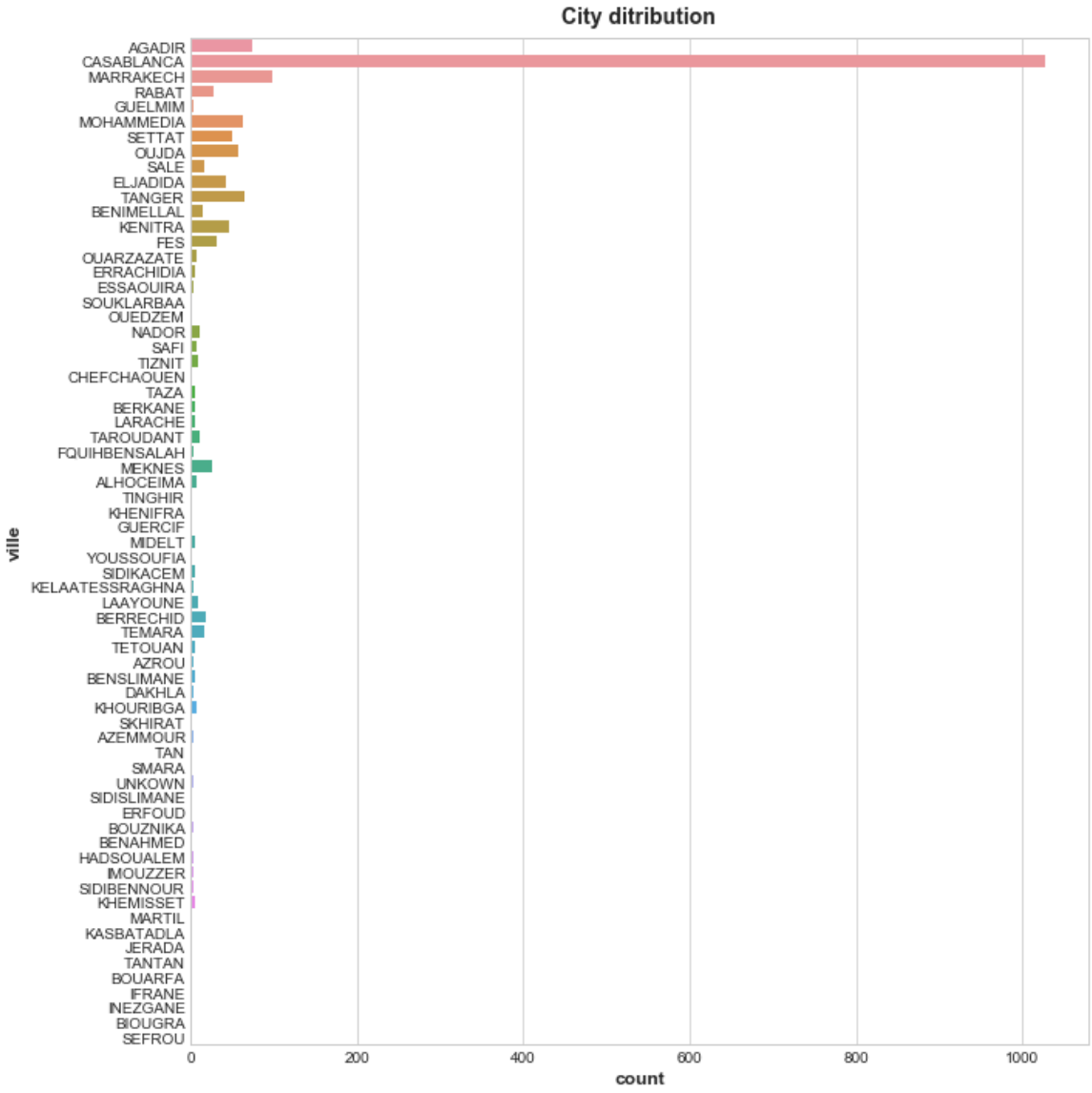
cities distribution:

In [ ]:

```
filter_by_ville = table_DS[['Matricule','ville']].drop_duplicates()
```

In [ ]:

```
plt.figure(figsize=(10,10))
plt.title('City ditribution')
sns.countplot(y='ville',data=filter_by_ville)
plt.show()
```



distribution of charges of DS:

In [ ]:

```
filter_by_charges_DS = table_DS.groupby('Matricule').Mt_remb.agg(sum)
```

In [ ]:

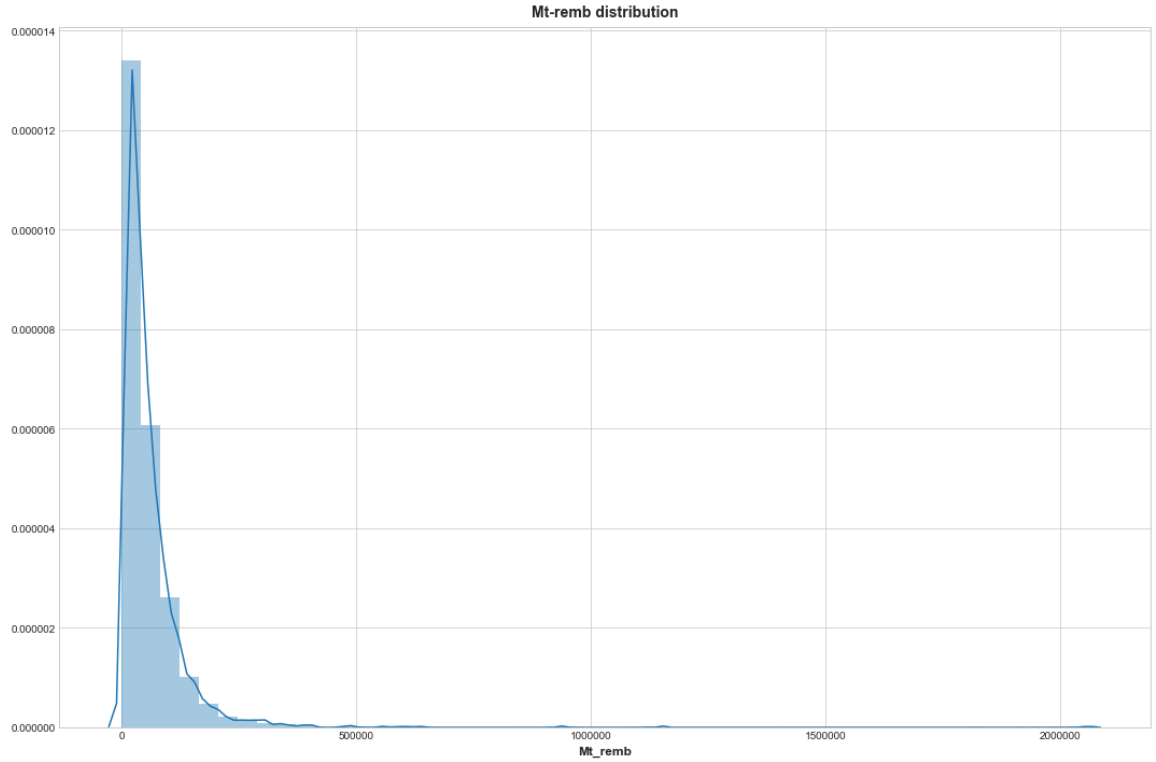
```
filter_by_charges_DS.head()
```

Out[ ]:

```
Matricule
30     13447.82
36    137832.43
40     61230.19
54     44069.55
75     19424.00
Name: Mt_remb, dtype: float64
```

In [ ]:

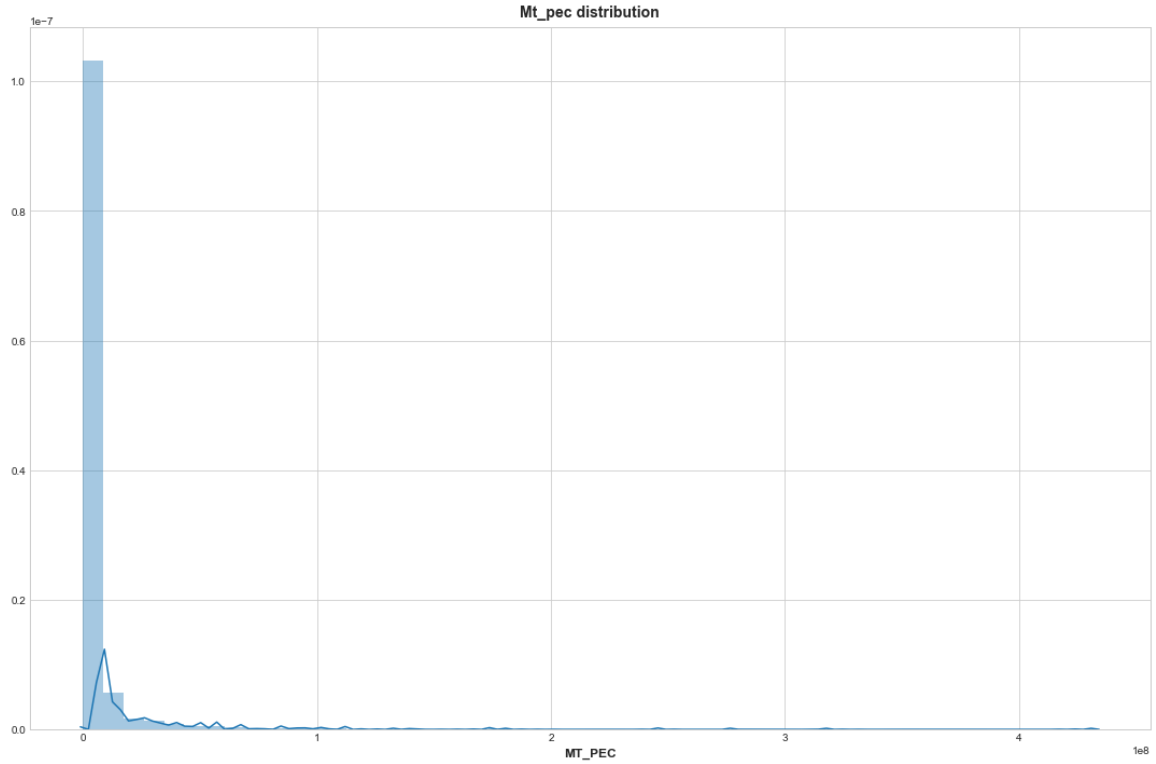
```
plt.figure(figsize=(15,10))
plt.title('Mt-remb distribution')
sns.distplot(filter_by_charges_DS)
plt.show()
```



distribution of charges of PEC:

In [ ]:

```
filter_by_charges_PEC = table_PEC.groupby('Matricule').MT_PEC.agg(sum)
plt.figure(figsize=(15,10))
plt.title('Mt_pec distribution')
sns.distplot(filter_by_charges_PEC)
plt.show()
```



data cleaning for table DS:

In [ ]:

```
table_DS.head()
```

Out[ ]:

	Annee	Mois	Matricule	Genre	ville	CJT	Enf	DS	Statut_adh	Mt_eng	Mt_
0	2017	1	10016	Masculin	AGADIR	1	2	1	Validé	16000.0	40
1	2017	5	10018	Masculin	CASABLANCA	1	2	3	Validé	4116.4	29
2	2017	6	10018	Masculin	CASABLANCA	1	2	1	Validé	640.0	1
3	2017	1	1004	Masculin	CASABLANCA	1	2	1	Validé	5834.2	23
4	2017	2	1004	Masculin	CASABLANCA	1	2	1	Validé	3358.8	22

In [ ]:

```
table_DS.shape
```

Out[ ]:

```
(32746, 12)
```

In [ ]:

```
data_DS = table_DS.drop(columns=['Mois','Mt_eng'])
```

In [ ]:

```
data_DS['Mt_remb'] = data_DS.groupby(['Annee','Matricule']).Mt_remb.transform(sum)
```

In [ ]:

```
data_DS['Age'] = data_DS.Age.astype(int)
```

In [ ]:

```
data_DS.shape
```

Out[ ]:

```
(32746, 10)
```

In [ ]:

```
data_DS = data_DS.drop_duplicates()
```

In [ ]:

```
data_DS.shape
```

Out[ ]:

```
(21164, 10)
```

In [ ]:

```
data_ALD = table_PEC[['Matricule','ALD']].drop_duplicates()
data_DS = data_DS.merge(data_ALD,on='Matricule',how='left')
```

In [ ]:

```
data_DS = data_DS.drop(columns=['Annee','Matricule','Statut_adh','DS'])
```

In [ ]:

```
data_DS.head()
```

Out[ ]:

	Genre	ville	CJT	Enf	Mt_remb	Age	ALD
0	Masculin	AGADIR	1	2	4040.00	61	1.0
1	Masculin	CASABLANCA	1	2	3150.56	57	1.0
2	Masculin	CASABLANCA	1	2	3150.56	58	1.0
3	Masculin	CASABLANCA	1	2	31191.20	62	2.0
4	Masculin	CASABLANCA	1	2	31191.20	63	2.0

In [ ]:

```
data_DS.to_csv('./data_DS.csv')
```

data cleaning for table PEC:

In [ ]:

```
data_PEC = table_PEC.drop(columns=['Max_date_prest','MT_Devis','PEC','Mois'])
```

In [ ]:

```
data_PEC['MT_PEC'] = data_PEC.groupby(['Annee','Matricule']).MT_PEC.transform(sum)
```

In [ ]:

```
data_age = table_DS[['Matricule','Age','Genre']]
data_age.Age = data_age.Age.astype(int)
data_age.Age = data_age.groupby('Matricule').transform('mean')
data_age = data_age.drop_duplicates()
data_age.Age = data_age.Age.astype(int)
```

In [ ]:

```
data_age.head()
```

Out[ ]:

	Matricule	Age	Genre
0	10016	63	Masculin
1	10018	60	Masculin
3	1004	63	Masculin
14	10047	66	Masculin
16	10057	65	Masculin

In [ ]:

```
data_PEC.head()
```

Out[ ]:

	Annee	Matricule	CJT	Enf	ALD	MT_PEC
0	2017	10016	1	2	1	241980.2
1	2017	10016	1	2	1	241980.2
2	2017	10016	1	2	1	241980.2
3	2017	10016	1	2	1	241980.2
4	2017	10016	1	2	1	241980.2

In [ ]:

```
data_PEC = data_PEC.merge(data_age,on='Matricule',how='left')
```

In [ ]:

```
data_PEC.head()
```

Out[ ]:

	Annee	Matricule	CJT	Enf	ALD	MT_PEC	Age	Genre
0	2017	10016	1	2	1	241980.2	63.0	Masculin
1	2017	10016	1	2	1	241980.2	63.0	Masculin
2	2017	10016	1	2	1	241980.2	63.0	Masculin
3	2017	10016	1	2	1	241980.2	63.0	Masculin
4	2017	10016	1	2	1	241980.2	63.0	Masculin

In [ ]:

```
data_PEC = data_PEC.drop_duplicates()
data_PEC = data_PEC.drop(columns=['Annee'])
```

In [ ]:

```
data_PEC.MT_PEC = data_PEC.groupby('Matricule').MT_PEC.transform('mean')
```

In [ ]:

```
data_PEC = data_PEC.drop_duplicates()
```

In [ ]:

```
data_PEC.head()
```

Out[ ]:

	Matricule	CJT	Enf	ALD	MT_PEC	Age	Genre
0	10016	1	2	1	141141.196	63.0	Masculin
5	10018	1	2	1	891492.240	60.0	Masculin
11	1004	1	2	2	115206.916	63.0	Masculin
17	10046	1	3	1	5859.675	NaN	NaN
18	10047	1	3	0	31254.400	66.0	Masculin

In [ ]:

```
data_PEC.to_csv('./data_PEC.csv')
```

the end