Dataframe

In [2]:
```
from IPython.display import HTML

HTML('''<script>
code_show=true;
function code_toggle() {
 if (code_show){
 $('div.input').hide();
 } else {
 $('div.input').show();
 }
 code_show = !code_show
}
$( document ).ready(code_toggle);
</script>''')
```

Out[2]:

In [89]:
```
df.head()
```

Out[89]:

	finess	mois	annee	sexe	ghm2	GHS	age	duree	supp_rea	supp_si	...	population_region	Libell GHI
0	10007987	6	2012	1	20Z042	7268	36	9	0	0	...	5 542 094	Ethylism ave dépendanc niveau
1	10007987	10	2015	1	20Z042	7268	39	8	0	0	...	5 542 094	Ethylism ave dépendanc niveau
2	10007987	7	2015	1	20Z042	7268	50	10	0	0	...	5 542 094	Ethylism ave dépendanc niveau
3	10780054	1	2015	1	20Z042	7268	71	11	0	0	...	5 542 094	Ethylism ave dépendanc niveau
4	10780054	7	2017	1	20Z042	7282	21	14	0	0	...	5 542 094	Ethylism ave dépendanc niveau

5 rows × 45 columns

In [90]:
```
df_c = df[df['grp_cln']=='16']
```

In [12]:
```
df_c['cost']
```

```
1807673    2422.771948
```

Out[12]:
```
1808197    2422.771948
1808592    1549.361917
1808593    1549.361917
1808594    1549.361917
              ...
1863967    3263.681464
1863968    1581.862432
1863974    1581.862432
1863977    2422.771948
1863983    2422.771948
Name: cost, Length: 32525, dtype: float64
```

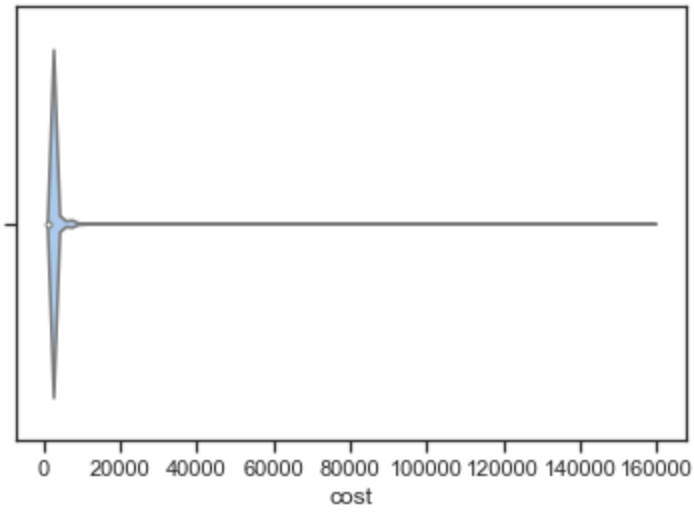
Cost

In [76]:
```
df_c['cost'].describe()
```

Out[76]:
```
count     32525.000000
mean       1618.025333
std        1059.379885
min        1253.658151
25%        1549.361917
50%        1549.361917
75%        1549.361917
max      159416.147591
Name: cost, dtype: float64
```

In [91]:
```
sns.violinplot(data=df_c,x='cost')
```

Out[91]:
```
<AxesSubplot:xlabel='cost'>
```



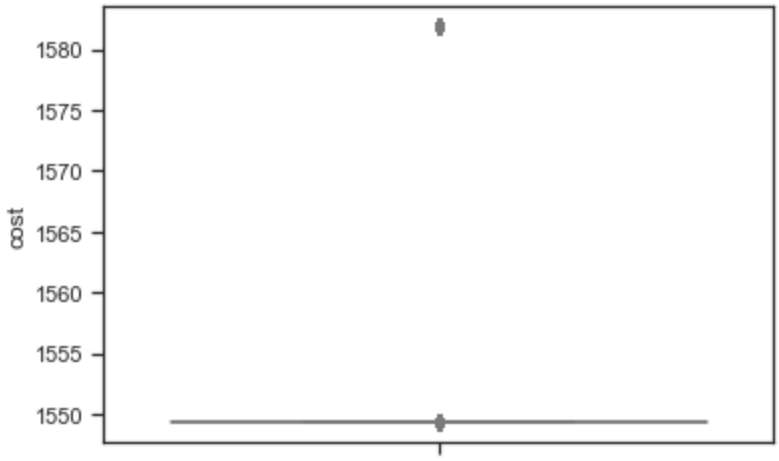
In [92]:
```
df_c[df_c['cost'] > (5000)]['cost']
```

Out[92]:
```
1808599     5707.403041
1810285    14195.505174
1814248     7468.229045
1814250     8543.357738
1814622     9677.739617
               ...
1861375    11153.879949
1861376    13909.826068
1862274     5786.410013
1862682    17559.143239
1863719     7468.229000
Name: cost, Length: 74, dtype: float64
```

In [93]:
```
df_c_trim = df_c[(df_c['cost']<1600)&(df_c['cost']>1300)]
```

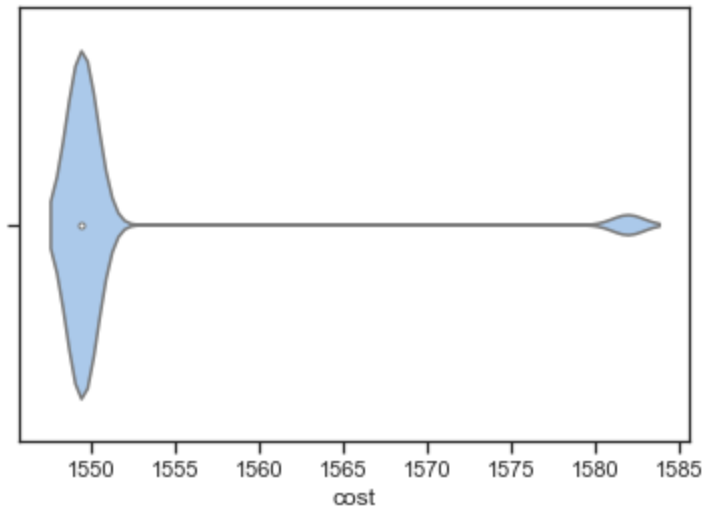
In [94]:
```
sns.boxplot(data=df_c_trim,y='cost')
```

Out[94]:
```
<AxesSubplot:ylabel='cost'>
```



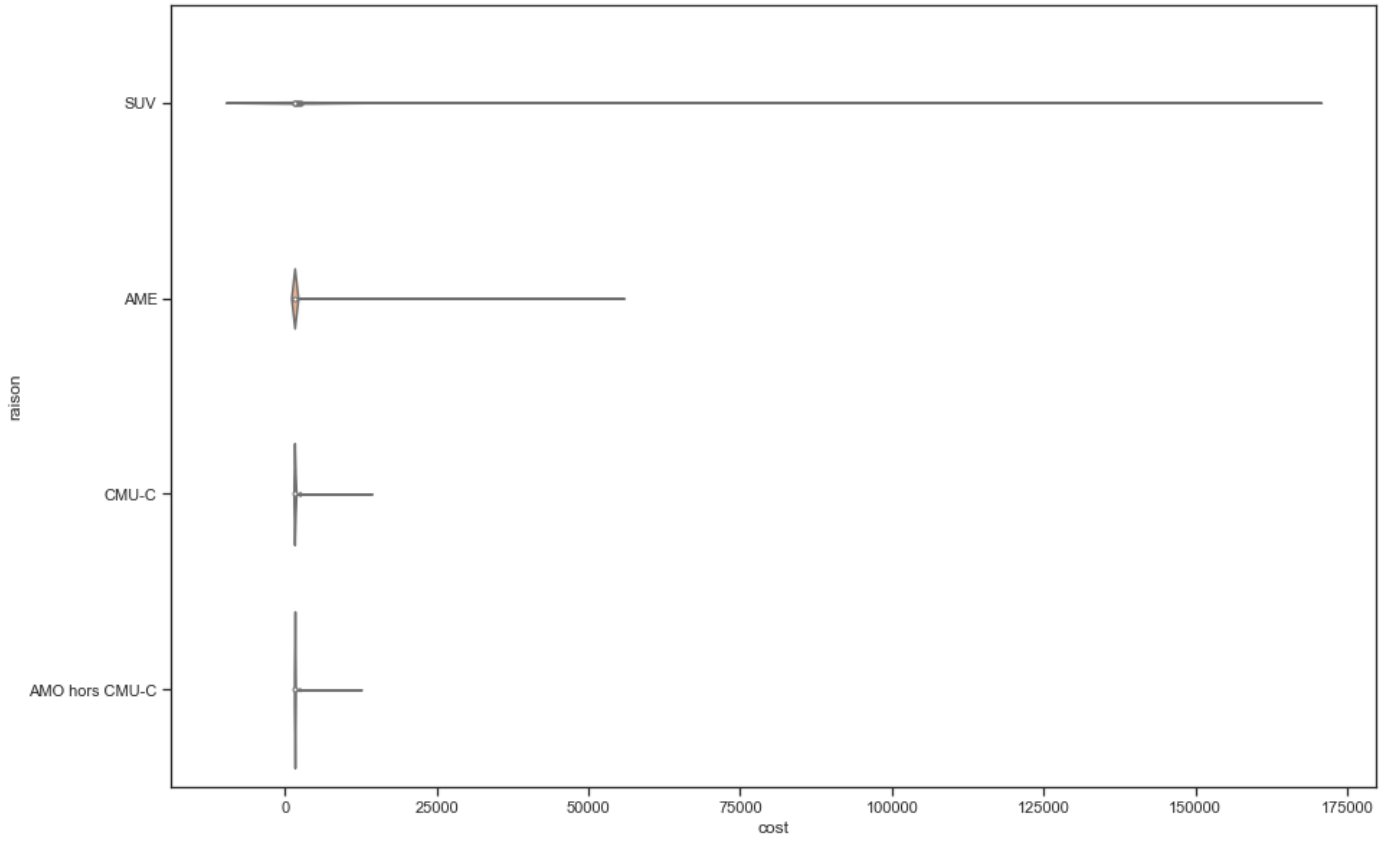
In [95]:
```
sns.violinplot(data=df_c_trim,x='cost')
```

Out[95]:
```
<AxesSubplot:xlabel='cost'>
```



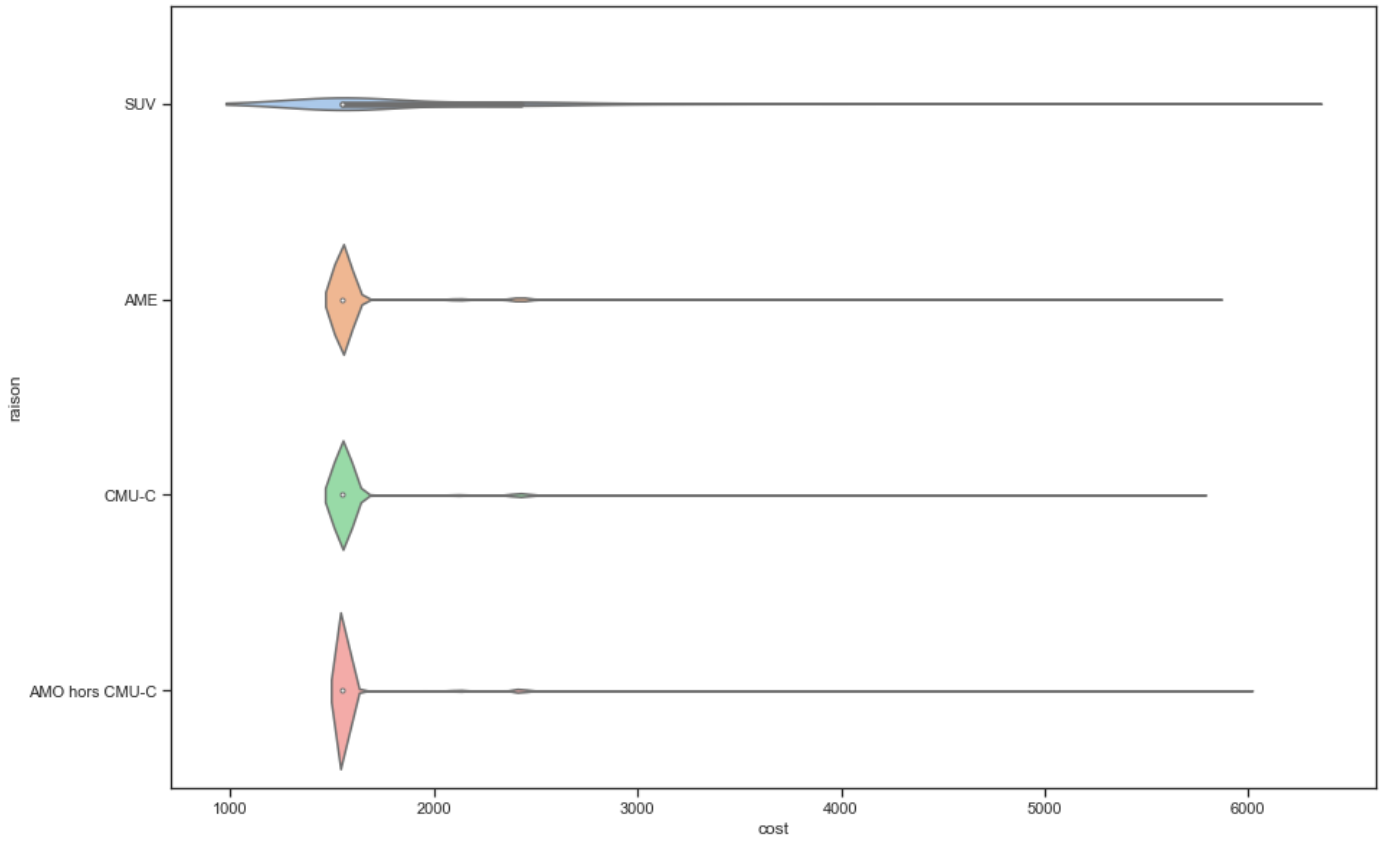
In [128…
```
plt.figure(figsize=(15,10))
sns.violinplot(data=df_c,x='cost',y='raison')
```

Out[128]:
```
<AxesSubplot:xlabel='cost', ylabel='raison'>
```



In [129…
```
plt.figure(figsize=(15,10))
sns.violinplot(data=df_c_trim,x='cost',y='raison')
```

Out[129]:
```
<AxesSubplot:xlabel='cost', ylabel='raison'>
```



In [132…
```
df_c[df_c['raison']=='SUV']['cost'].describe()
```

Out[132]:
```
count       118.000000
mean       4120.050742
std       14724.370337
min        1549.361917
```

```
25%        1549.361917
50%        1549.361917
75%        2422.771948
max      159416.147591
Name: cost, dtype: float64
```

In [133…
```
df_c[df_c['raison']=='AMO hors CMU-C']['cost'].describe()
```

Out[133]:
```
count    15305.000000
mean      1589.292423
std        284.738154
min       1549.361917
25%       1549.361917
50%       1549.361917
75%       1549.361917
max      12513.686142
Name: cost, dtype: float64
```

In [134…
```
df_c[df_c['raison']=='CMU-C']['cost'].describe()
```

Out[134]:
```
count     6327.000000
mean      1613.731142
std        436.584210
min       1549.361917
25%       1549.361917
50%       1549.361917
75%       1549.361917
max      14195.505174
Name: cost, dtype: float64
```

In [135…
```
df_c[df_c['raison']=='AME']['cost'].describe()
```

Out[135]:
```
count    10775.000000
mean      1633.959212
std        858.141820
min       1253.658151
25%       1549.361917
50%       1549.361917
75%       1549.361917
max      55620.210000
Name: cost, dtype: float64
```

In [97]:
```
def age_class(age):
    if age in list(range(0,18)):
        return '0 à 17 ans'
    elif age in list(range(18,36)):
        return '18 à 35 ans'
    elif age in list(range(36,54)):
        return '36 à 53 ans'
    elif age in list(range(54,72)):
        return '54 à 71 ans'
    elif age in list(range(72,90)):
        return '72 à 89 ans'
    else:
        return '90 ans et plus'
```

In [98]:
```
df_c['age_class'] = df_c['age'].apply(age_class)
```

```
<ipython-input-98-64996861a71c>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_
```

```
guide/indexing.html#returning-a-view-versus-a-copy
  df_c['age_class'] = df_c['age'].apply(age_class)
```

In [100…

In [102…
```
df_c.sort_values(by='age_class',inplace=True)
```

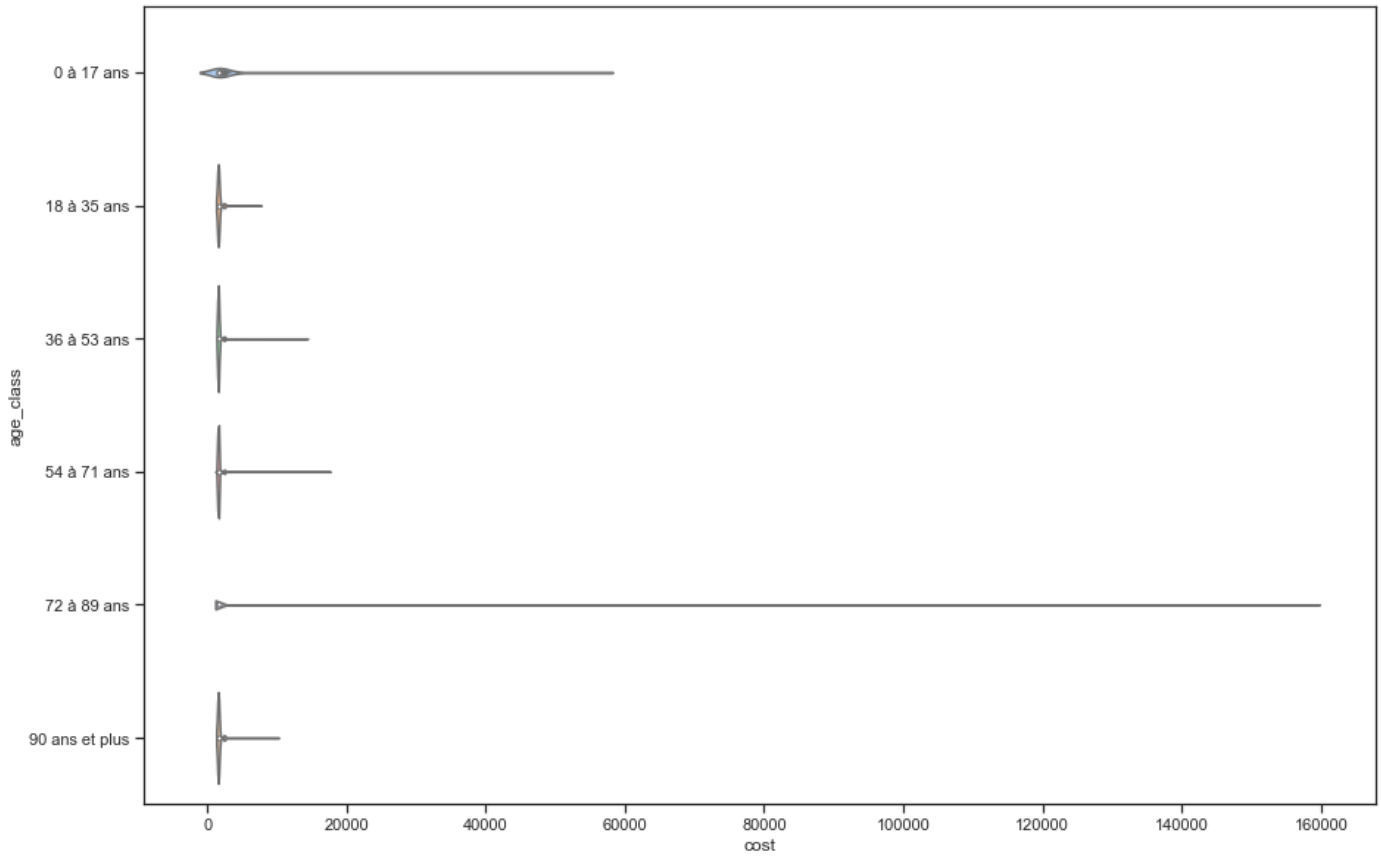
```
<ipython-input-102-7d14fcbfc09c>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_
guide/indexing.html#returning-a-view-versus-a-copy
  df_c.sort_values(by='age_class',inplace=True)
```

In [117…
```
df_c_trim = df_c[(df_c['cost']<((6000)))&(df_c['cost']>1400)]
```

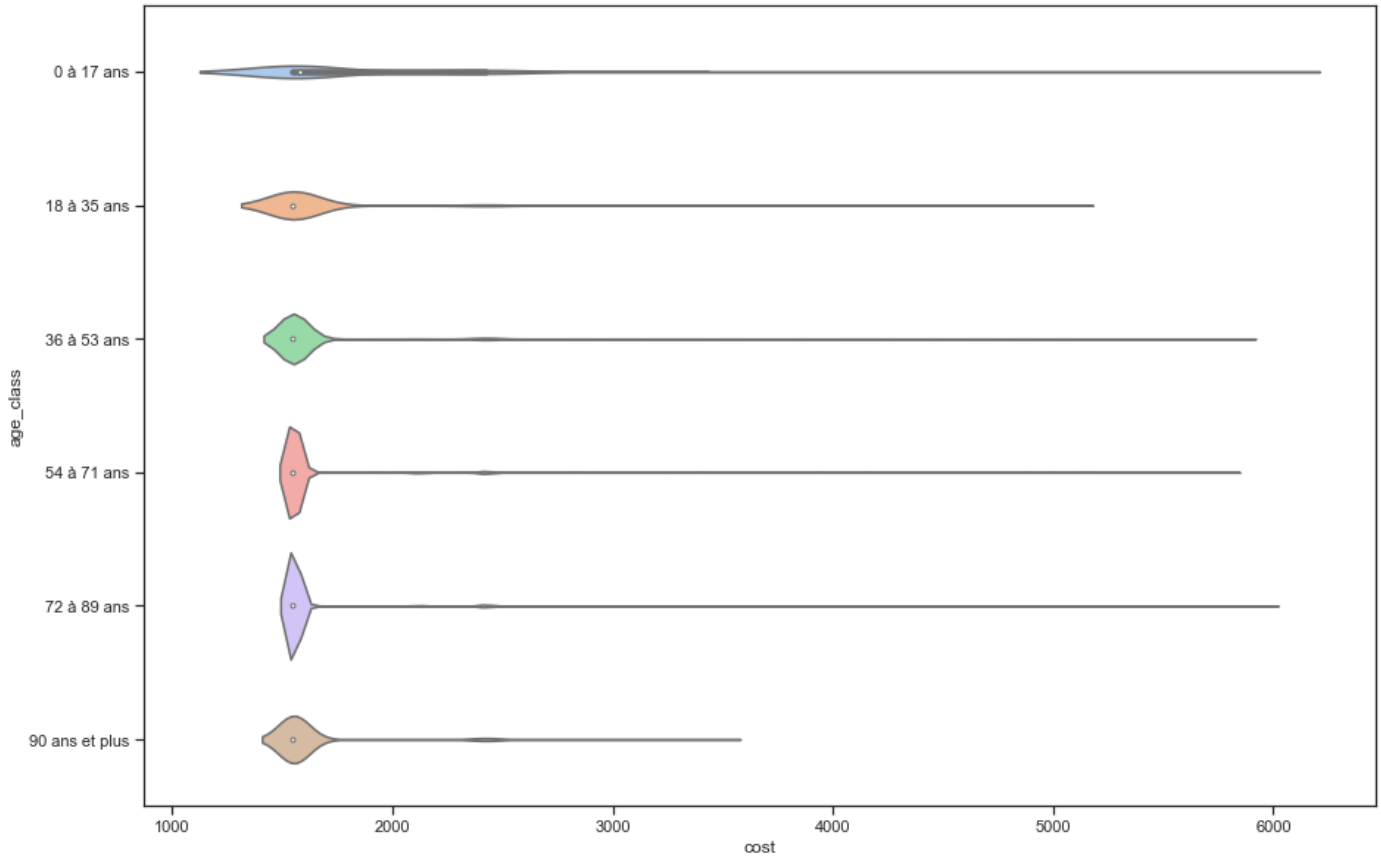
In [118…
```
plt.figure(figsize=(15,10))
sns.violinplot(data=df_c,x='cost',y='age_class')
```

Out[118]:
```
<AxesSubplot:xlabel='cost', ylabel='age_class'>
```



In [119…
```
plt.figure(figsize=(15,10))
sns.violinplot(data=df_c_trim,x='cost',y='age_class')
```

Out[119]:
```
<AxesSubplot:xlabel='cost', ylabel='age_class'>
```



In [136…
```
df_c[df_c['age_class']=='0 à 17 ans']['cost'].describe()
```

Out[136]:
```
count      363.000000
mean      2530.086301
std       4302.024686
min       1549.361917
25%       1549.361917
50%       1581.862432
75%       2422.771948
max      55620.210000
Name: cost, dtype: float64
```

In [137…
```
df_c[df_c['age_class']=='18 à 35 ans']['cost'].describe()
```

Out[137]:
```
count      662.000000
mean      1680.275694
std        536.519581
min       1549.361917
25%       1549.361917
50%       1549.361917
75%       1549.361917
max       7468.229045
Name: cost, dtype: float64
```

In [138…
```
df_c[df_c['age_class']=='36 à 53 ans']['cost'].describe()
```

Out[138]:
```
count     2569.000000
mean      1642.980365
std        550.190437
min       1549.361917
25%       1549.361917
50%       1549.361917
75%       1549.361917
max      14195.505174
Name: cost, dtype: float64
```

In [139…
```
df_c[df_c['age_class']=='54 à 71 ans']['cost'].describe()
```

Out[139]:
```
count    14303.000000
mean      1604.215935
std        424.292397
min       1253.658151
25%       1549.361917
50%       1549.361917
75%       1549.361917
max      17559.143239
Name: cost, dtype: float64
```

In [140…
```
df_c[df_c['age_class']=='72 à 89 ans']['cost'].describe()
```

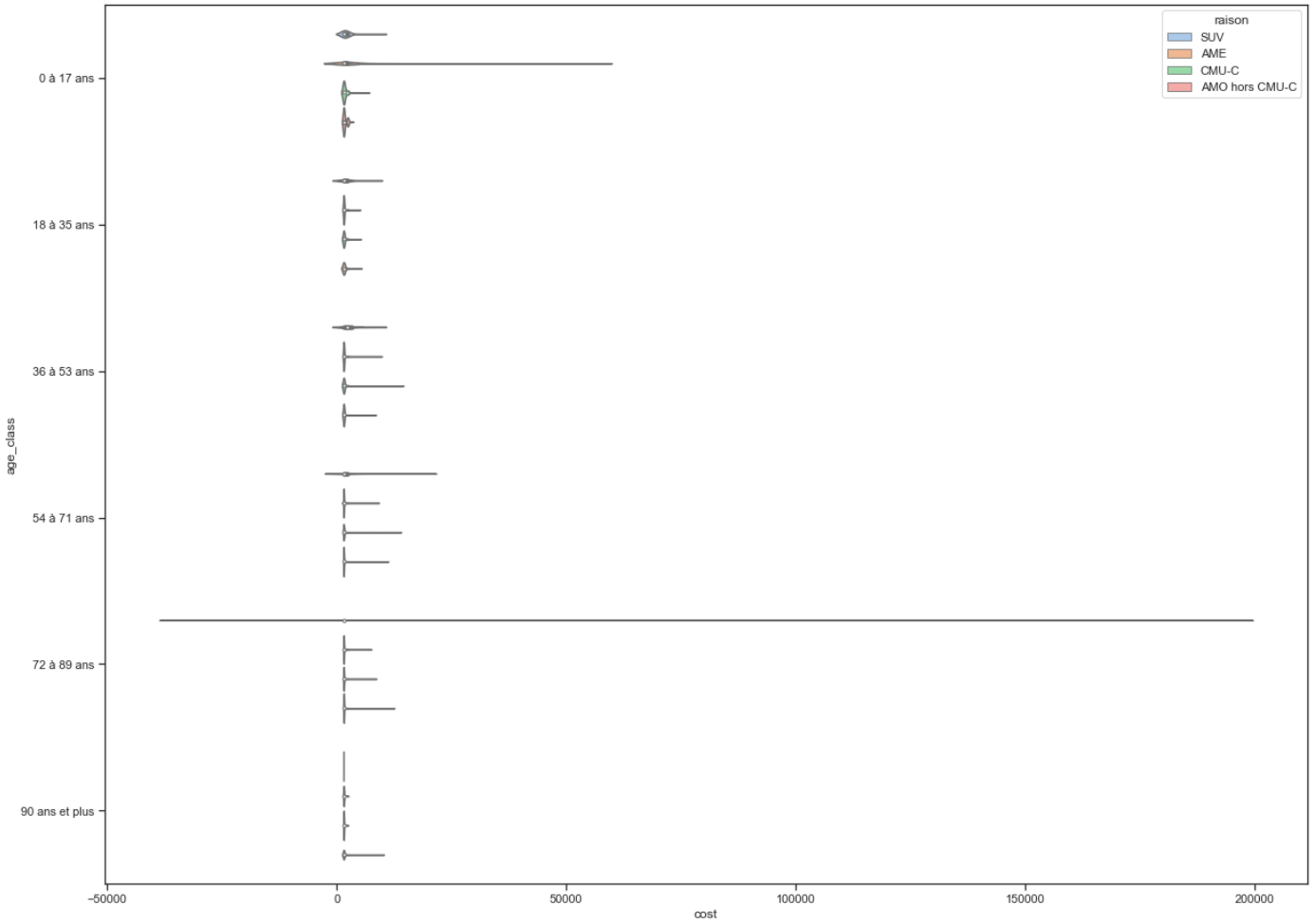
Out[140]:
```
count     14191.000000
mean       1600.763486
std        1349.738523
min        1549.361917
25%        1549.361917
50%        1549.361917
75%        1549.361917
max      159416.147591
Name: cost, dtype: float64
```

In [141…
```
df_c[df_c['age_class']=='90 ans et plus']['cost'].describe()
```

Out[141]:
```
count     437.000000
mean     1631.941464
std       466.237416
min      1549.361917
25%      1549.361917
50%      1549.361917
75%      1549.361917
max      9990.957594
Name: cost, dtype: float64
```

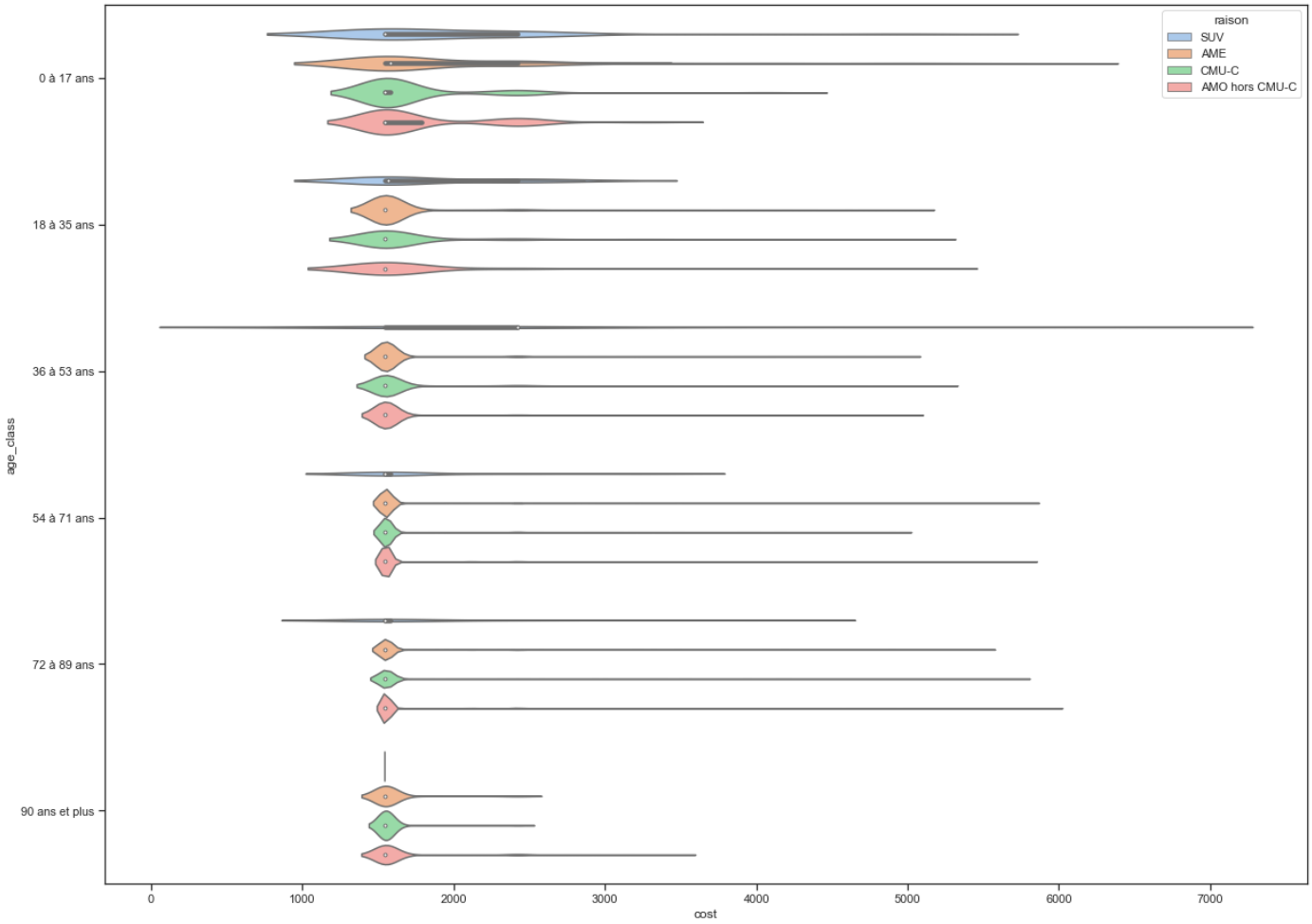
In [120…
```
plt.figure(figsize=(20,15))
sns.violinplot(data=df_c,x='cost',y='age_class',hue='raison')
```

Out[120]:
```
<AxesSubplot:xlabel='cost', ylabel='age_class'>
```



In [121…
```
plt.figure(figsize=(20,15))
sns.violinplot(data=df_c_trim,x='cost',y='age_class',hue='raison')
```

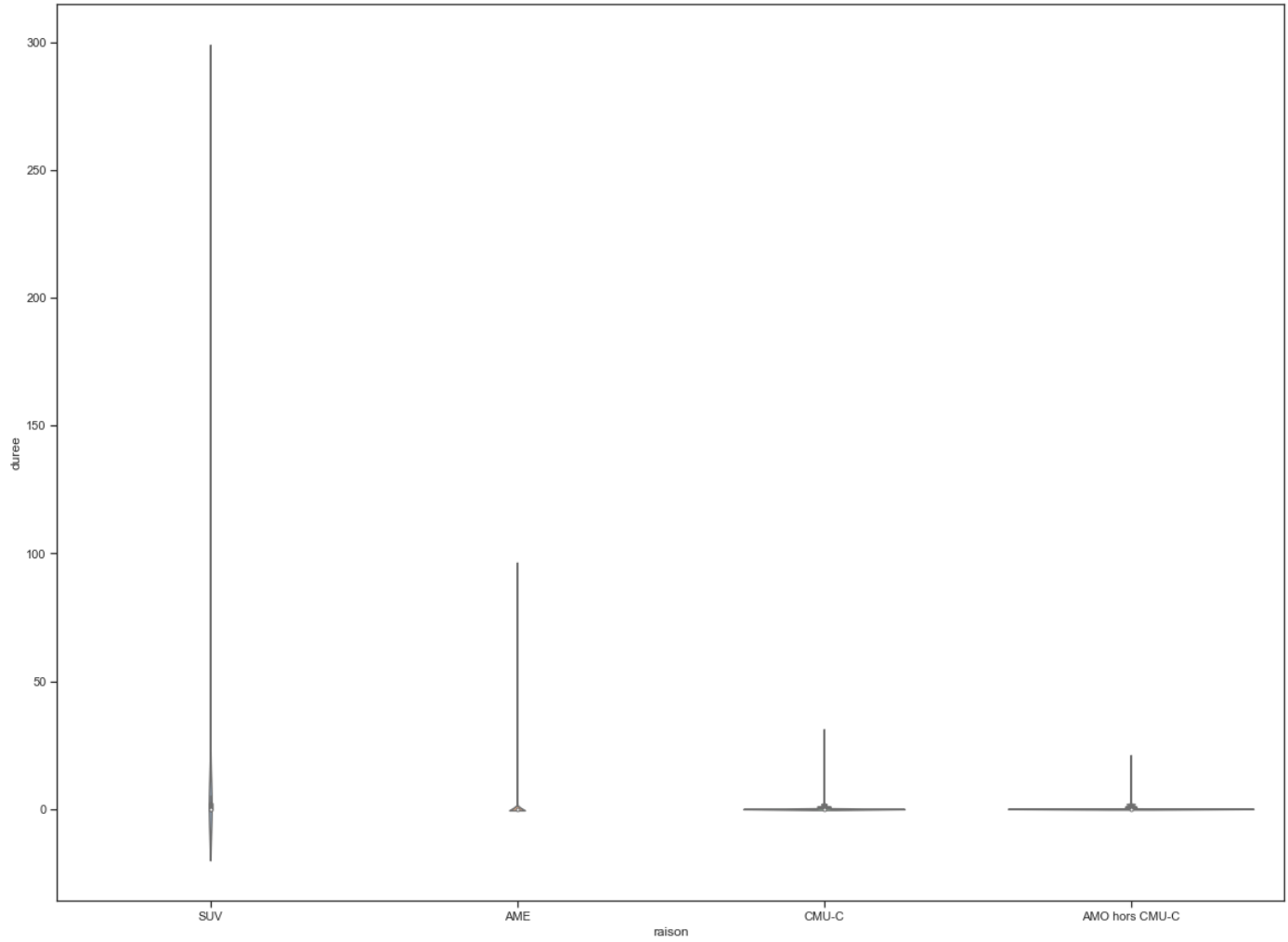
Out[121]:
```
<AxesSubplot:xlabel='cost', ylabel='age_class'>
```



Durée

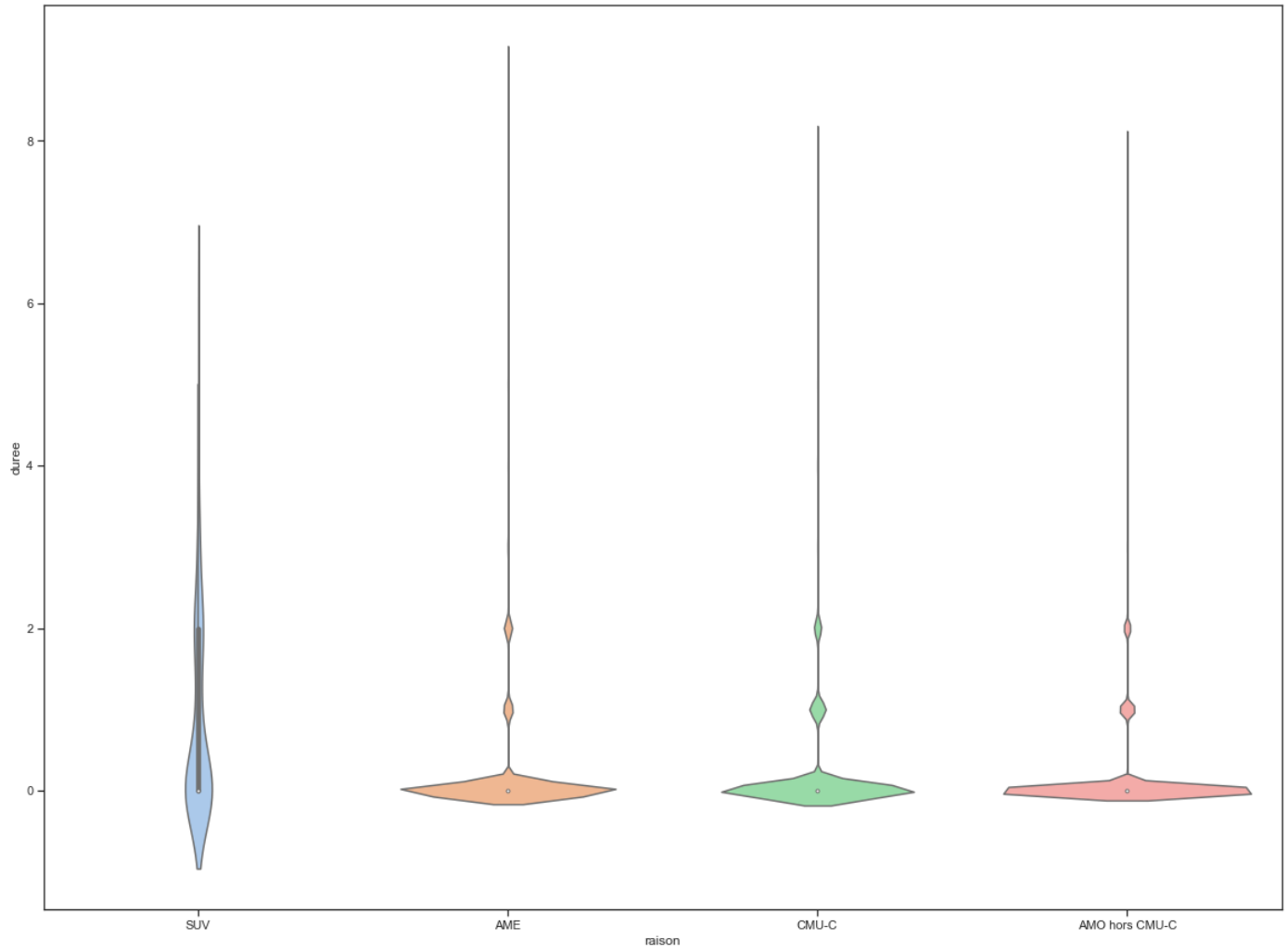
In [153…
```
plt.figure(figsize=(20,15))
sns.violinplot(data=df_c,y='duree',x='raison')
```

Out[153]:
```
<AxesSubplot:xlabel='raison', ylabel='duree'>
```



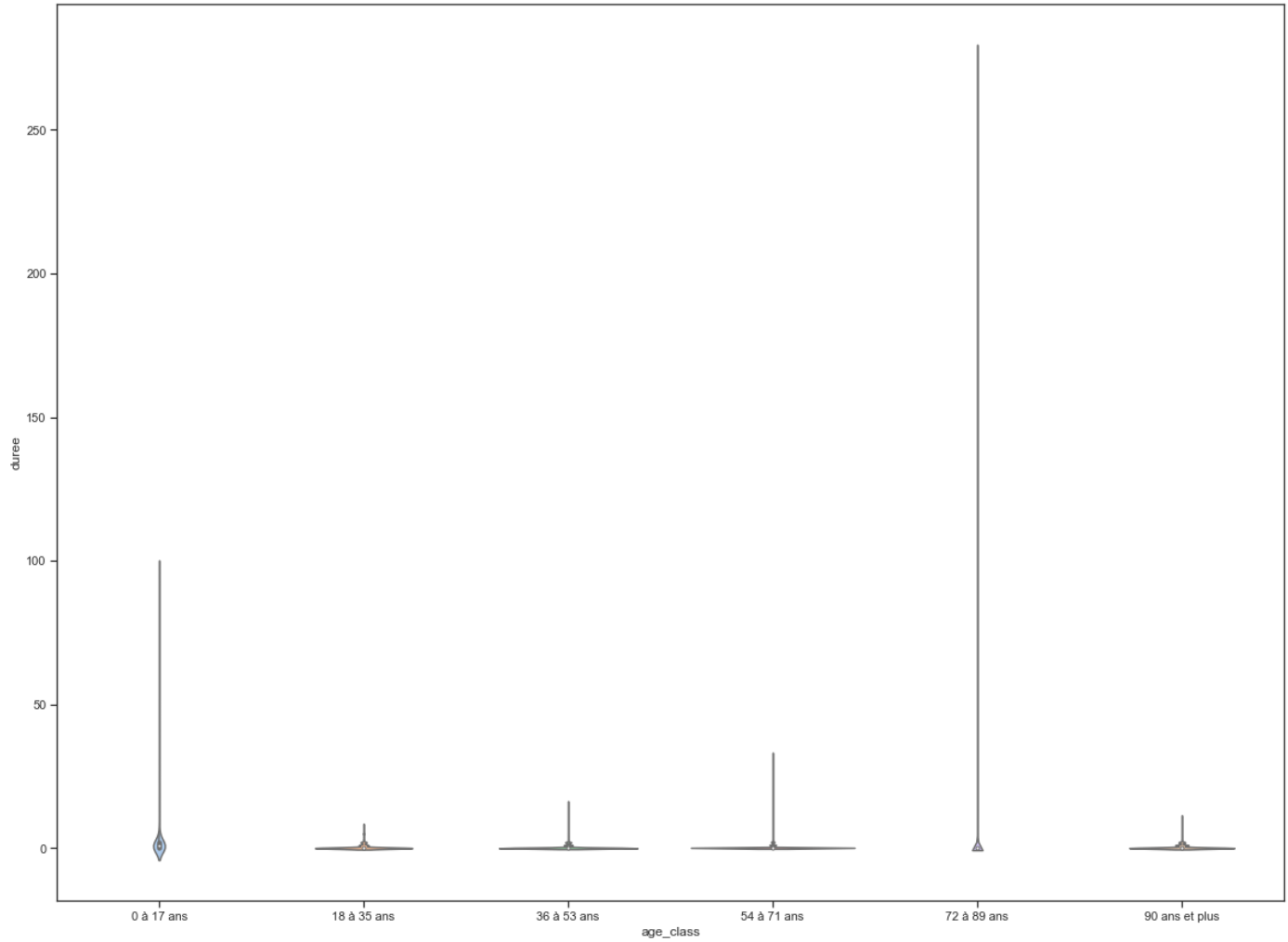
In [154…
```
plt.figure(figsize=(20,15))
sns.violinplot(data=df_c_trim,y='duree',x='raison')
```

Out[154]:
```
<AxesSubplot:xlabel='raison', ylabel='duree'>
```



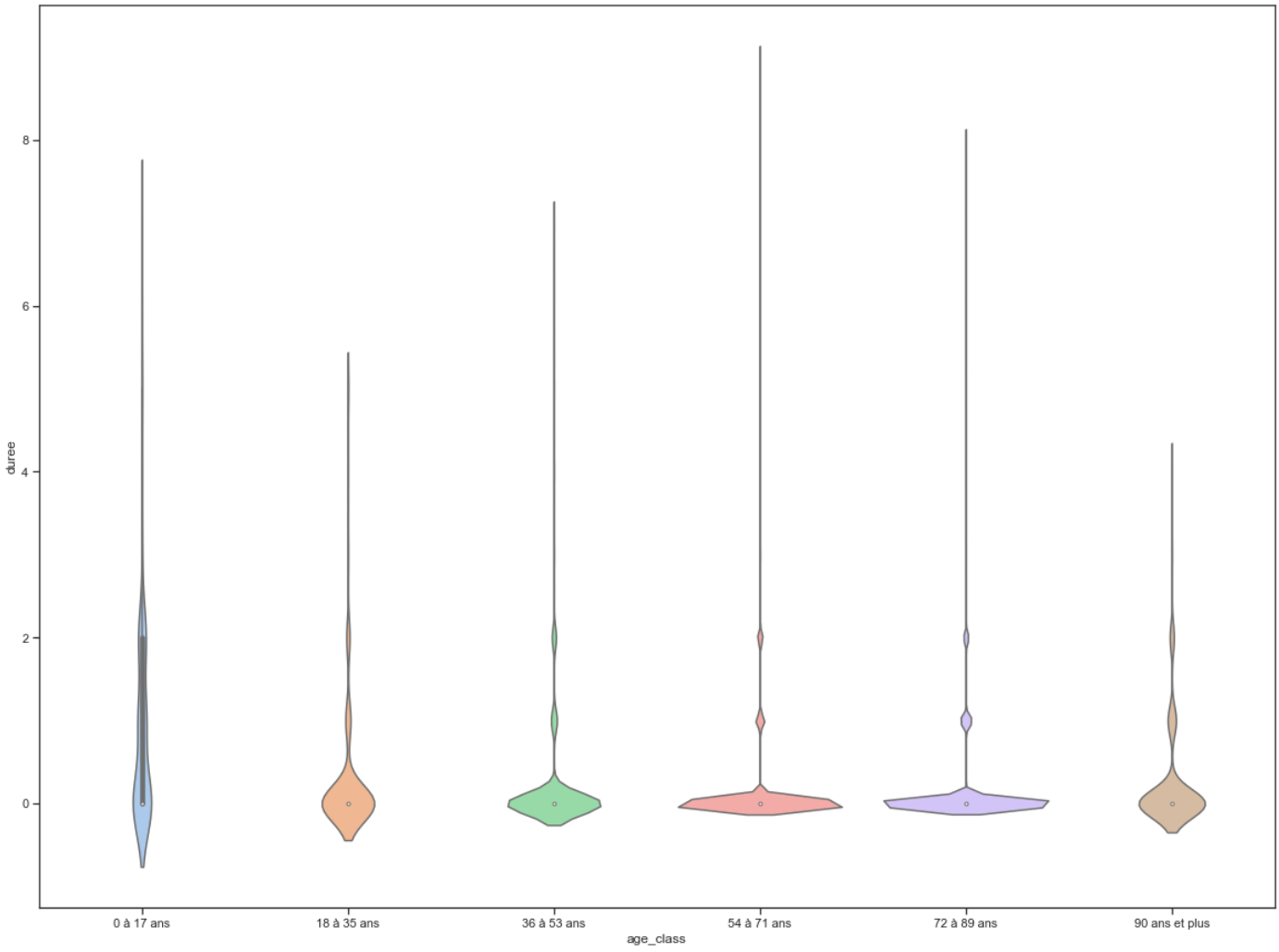
In [155…
```
plt.figure(figsize=(20,15))
sns.violinplot(data=df_c,y='duree',x='age_class')
```

Out[155]:
```
<AxesSubplot:xlabel='age_class', ylabel='duree'>
```



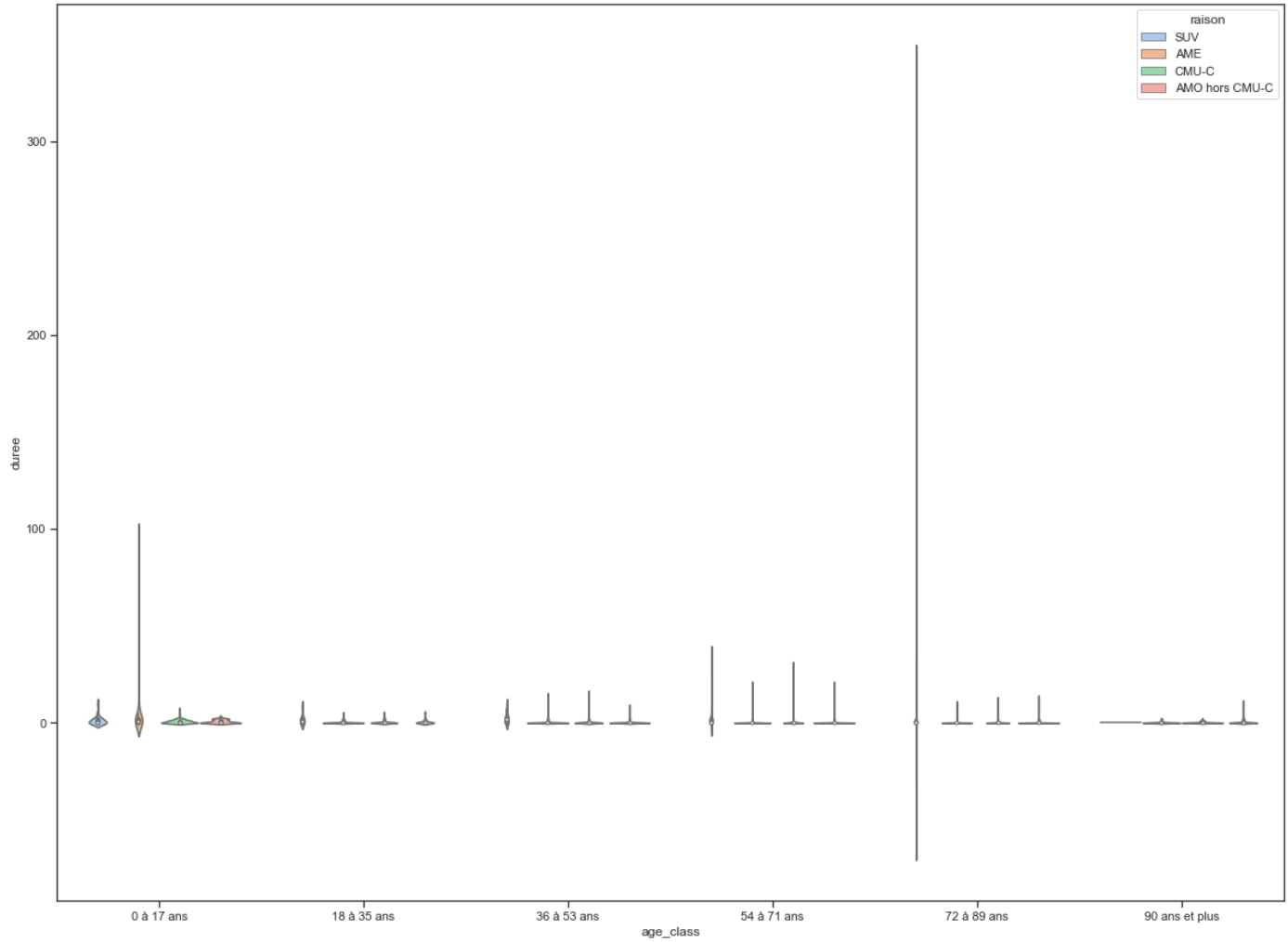
In [156…
```
plt.figure(figsize=(20,15))
sns.violinplot(data=df_c_trim,y='duree',x='age_class')
```

Out[156]:
```
<AxesSubplot:xlabel='age_class', ylabel='duree'>
```



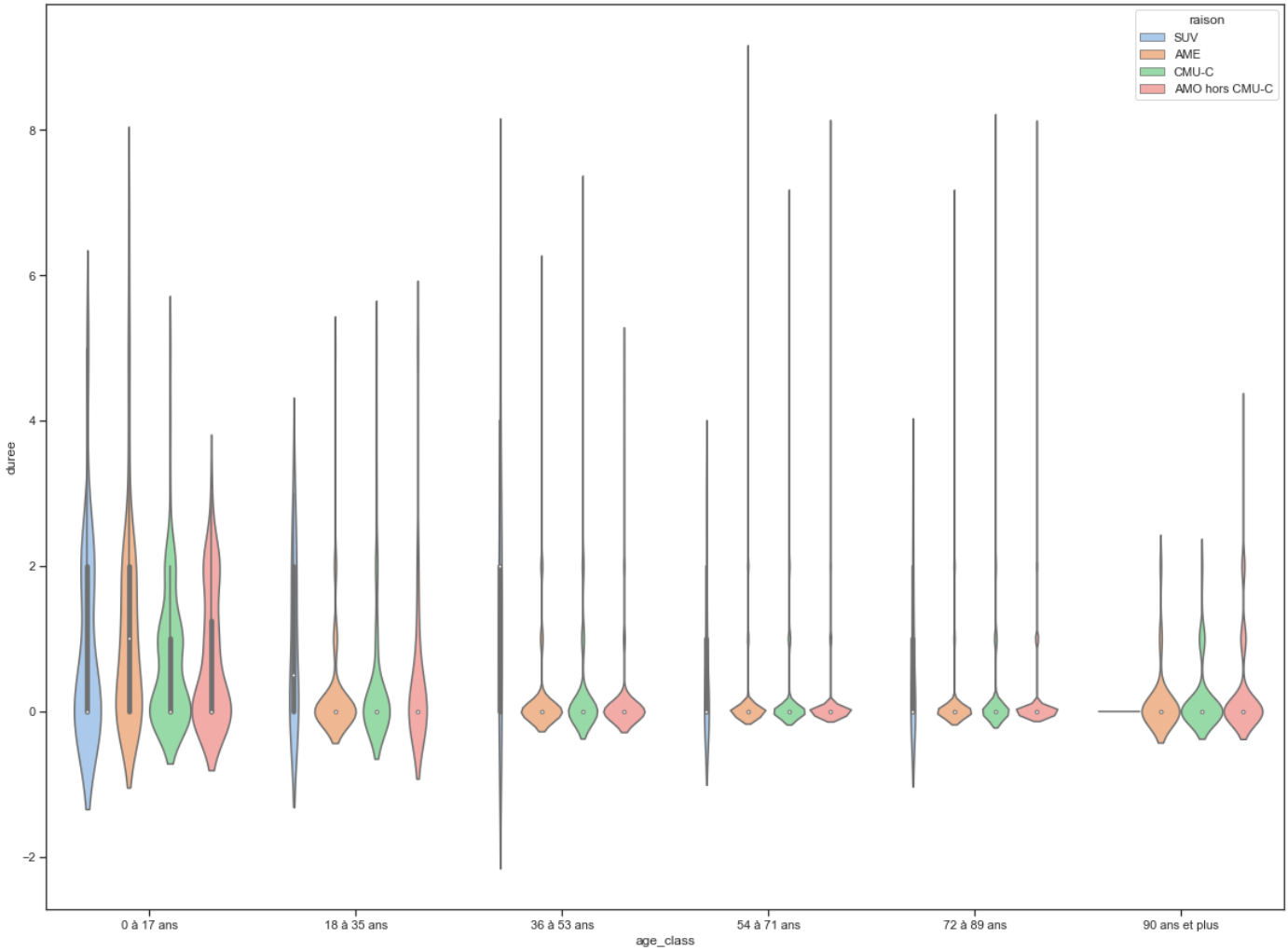
In [126…
```
plt.figure(figsize=(20,15))
sns.violinplot(data=df_c,y='duree',x='age_class',hue='raison')
```

Out[126]:
```
<AxesSubplot:xlabel='age_class', ylabel='duree'>
```



In [127…
```
plt.figure(figsize=(20,15))
sns.violinplot(data=df_c_trim,y='duree',x='age_class',hue='raison')
```

Out[127]:
```
<AxesSubplot:xlabel='age_class', ylabel='duree'>
```



In [ ]:

In [148…
```
df_c[df_c['raison']=='SUV']['duree'].describe()
```

Out[148]:
```
count    118.000000
mean       4.406780
std       25.913478
min        0.000000
25%        0.000000
50%        0.000000
75%        2.000000
max      279.000000
Name: duree, dtype: float64
```

In [149…
```
df_c[df_c['raison']=='AMO hors CMU-C']['duree'].describe()
```

Out[149]:
```
count    15305.000000
mean         0.117935
std          0.534495
min          0.000000
25%          0.000000
50%          0.000000
75%          0.000000
max         21.000000
Name: duree, dtype: float64
```

In [150…
```
df_c[df_c['raison']=='CMU-C']['duree'].describe()
```

Out[150]:
```
count    6327.000000
mean        0.186976
std         0.878468
min         0.000000
```

```
25%          0.000000
50%          0.000000
75%          0.000000
max         31.000000
Name: duree, dtype: float64
```

In [151…
```
df_c[df_c['raison']=='AME']['duree'].describe()
```

Out[151]:
```
count    10775.000000
mean         0.183759
std          1.422464
min          0.000000
25%          0.000000
50%          0.000000
75%          0.000000
max         96.000000
Name: duree, dtype: float64
```

In [152…
```
df_c[df_c['age_class']=='0 à 17 ans']['duree'].describe()
```

Out[152]:
```
count    363.000000
mean       1.884298
std        6.809796
min        0.000000
25%        0.000000
50%        1.000000
75%        2.000000
max       96.000000
Name: duree, dtype: float64
```

In [143…
```
df_c[df_c['age_class']=='18 à 35 ans']['duree'].describe()
```

Out[143]:
```
count    662.000000
mean       0.309668
std        0.910120
min        0.000000
25%        0.000000
50%        0.000000
75%        0.000000
max        8.000000
Name: duree, dtype: float64
```

In [144…
```
df_c[df_c['age_class']=='36 à 53 ans']['duree'].describe()
```

Out[144]:
```
count    2569.000000
mean        0.217205
std         0.888399
min         0.000000
25%         0.000000
50%         0.000000
75%         0.000000
max        16.000000
Name: duree, dtype: float64
```

In [145…
```
df_c[df_c['age_class']=='54 à 71 ans']['duree'].describe()
```

Out[145]:
```
count    14303.000000
mean         0.133399
std          0.823098
min          0.000000
25%          0.000000
50%          0.000000
75%          0.000000
max         33.000000
Name: duree, dtype: float64
```

In [146…
```
df_c[df_c['age_class']=='72 à 89 ans']['duree'].describe()
```

Out[146]:
```
count    14191.000000
mean         0.142344
std          2.394357
min          0.000000
25%          0.000000
50%          0.000000
75%          0.000000
max        279.000000
Name: duree, dtype: float64
```

In [147…
```
df_c[df_c['age_class']=='90 ans et plus']['duree'].describe()
```

Out[147]:
```
count    437.000000
mean       0.258581
std        0.777818
min        0.000000
25%        0.000000
50%        0.000000
75%        0.000000
max       11.000000
Name: duree, dtype: float64
```

In [ ]:
```

```