In [11]:

```
import pandas as pd
import tabula
import numpy as np
import glob
import re
pd.set_option('display.max_columns', 50)
pd.set_option('display.max_rows', 550)
dane = pd.read_csv('C:/Users/Laptop/Desktop/Grypa/Dane/dane.csv')
dane.drop(dane.columns[dane.columns.str.contains('unnamed', case = False)], axis=1, in
dane['Rok_str'] = dane['Rok_str'].astype(str)
dane
```

53	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2011/02B.pdf	2011	
54	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2011/02C.pdf	2011	
55	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2011/02D.pdf	2011	
56	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2011/03A.pdf	2011	
57	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2011/03B.pdf	2011	
58	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2011/03C.pdf	2011	
59	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2011/03D.pdf	2011	
60	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2011/04A.pdf	2011	
61	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2011/04B.pdf	2011	
62	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2011/04C.pdf	2011	
63	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2011/04D.pdf	2011	
64	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2011/05A.pdf	2011	

In [10]:

```
suma = dane['0 to 4'].sum()
suma
```

Out[10]:

```
8018443
```

In [15]:

```
#przygotowanie do zaciągnięcia danych
dane['Rok_str'] = dane['Rok'].astype(str)
dane['Sezon'] = ""
dane['0 to 4'] = ""
dane['5 to 14'] = ""
dane['15 to 64'] = ""
dane['65+'] = ""
```

In [18]:

```
#zaciągnięcie danych
for x in range(507):
    ad = dane.loc[x, "Plik"]

    liczba = tabula.read_pdf(ad, area = (230, 240, 240, 294), pages=1)
    liczba_str = str(liczba)
    O_to_4 = re.sub('[^0-9]', '', liczba_str)
    dane.loc[x, "0 to 4"] = O_to_4
    print(O_to_4)

    liczba2 = tabula.read_pdf(ad, area = (230, 330, 240, 360), pages=1)
    liczba2_str = str(liczba2)
    piec_to_14 = re.sub('[^0-9]', '', liczba2_str)
    dane.loc[x, "5 to 14"] = piec_to_14
    print(piec_to_14)

    liczba3 = tabula.read_pdf(ad, area = (230, 400, 240, 440), pages=1)
    liczba3_str = str(liczba3)
    pietnascie_to_64 = re.sub('[^0-9]', '', liczba3_str)
    dane.loc[x, "15 to 64"] = pietnascie_to_64
    print(pietnascie_to_64)

    liczba4 = tabula.read_pdf(ad, area = (230, 485, 240, 530), pages=1)
    liczba4_str = str(liczba4)
    szescpiec = re.sub('[^0-9]', '', liczba4_str)
    dane.loc[x, '65+'] = szescpiec
    print(szescpiec)

dane.to_csv('C:/Users/Laptop/Desktop/Grypa/Dane/dane.csv')
```

```
2568
2710
9754
1582
2296
2841
9316
1474
2718
3364
9180
1345
2524
3301
7883
1590
2760
3341
8395
```

In [12]:

```
#przygotowanie do wprawadzenia sezonów
lista_lat = dane['Rok'].unique()
lista_lat_str = dane['Rok_str'].unique()
slow1 = dict(zip(lista_lat_str, lista_lat))
print(slow1)
```

```
{'2010': 2010, '2011': 2011, '2012': 2012, '2013': 2013, '2014': 2014, '201
5': 2015, '2016': 2016, '2017': 2017, '2018': 2018, '2019': 2019, '2020': 20
20}
```

In [14]:

```
lista_sezonow = []
for x in lista_lat_str:
    nazwa_sezonu = 'sezon_' + x
    lista_sezonow.append(nazwa_sezonu)
    #print(nazwa_sezonu)
lista_sezonow
```

Out[14]:

```
['sezon_2010',
 'sezon_2011',
 'sezon_2012',
 'sezon_2013',
 'sezon_2014',
 'sezon_2015',
 'sezon_2016',
 'sezon_2017',
 'sezon_2018',
 'sezon_2019',
 'sezon_2020']
```

In [16]:

```
#podział danych na pliki z sezonami
for x in lista_lat_str:
    sez1 = dane.loc[(dane['Rok'] == slow1[x]) & (dane['Miesiac'] > 8)]
    sez2 = dane.loc[(dane['Rok'] == slow1[x] + 1) & (dane['Miesiac'] < 9)]
    sezon = sez1.append(sez2)
    #sezon.rename({"Unnamed: 0":"a"}, axis="columns", inplace=True)
    #sezon.drop(["a"], axis=1, inplace=True)
    sezon1 = sezon.reset_index(drop=True)
    sezon1['Sezon'] = x
    sezon1.to_csv('C:/Users/Laptop/Desktop/Grypa/Dane/Sezony/Sezon_' + x + '.csv')
    #print(sezon1)
```

In [22]:

```
#scalanie sezonów w jeden plik
path = r'C:/Users/Laptop/Desktop/Grypa/Dane/Sezony'
all_files = glob.glob(path + "/*.csv")
li = []
for x in all_files:
    plik = pd.read_csv(x, index_col=None, header=0)
    li.append(plik)
total = pd.concat(li, axis=0, ignore_index=True)
total.drop(total.columns[total.columns.str.contains('unnamed', case = False)], axis=1,
total['Rok_str'] = total['Rok_str'].astype(str)
total.to_csv('C:/Users/Laptop/Desktop/Grypa/Dane/' + 'total' +'.csv')
```

In [23]:

```
total["Unique_ID"] = total['Rok_str'] + '_' + total['Tydzien']
```

In [24]:

```
total
```

Out[24]:

	URL	Plik	Rok	Miesia
0	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2010/09A.pdf	2010	9
1	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2010/09B.pdf	2010	9
2	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2010/09C.pdf	2010	9
3	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2010/09D.pdf	2010	9
4	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2010/10A.pdf	2010	10
5	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2010/10B.pdf	2010	10
6	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2010/10C.pdf	2010	10
7	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2010/10D.pdf	2010	10
8	http://wwwold.pzh.gov.pl/oldpage/epimeld/grypa...	C:/Users/Laptop/Desktop/Grypa/Dane/2010/11A.pdf	2010	1

In [45]:

```
basic = pd.DataFrame(total, columns = ['Unique_ID', 'Total'])
basic['nr_tyg'] = basic.reset_index().index
basic
```

Out[45]:

	Unique_ID	Total	nr_tyg
0	2010_09A	3604	0
1	2010_09B	6892	1
2	2010_09C	10398	2
3	2010_09D	12631	3
4	2010_10A	15750	4
5	2010_10B	19872	5
6	2010_10C	18634	6
7	2010_10D	18753	7
8	2010_11A	16372	8
9	2010_11B	18387	9

In [52]:

```
import matplotlib
from matplotlib import pyplot as plt
```

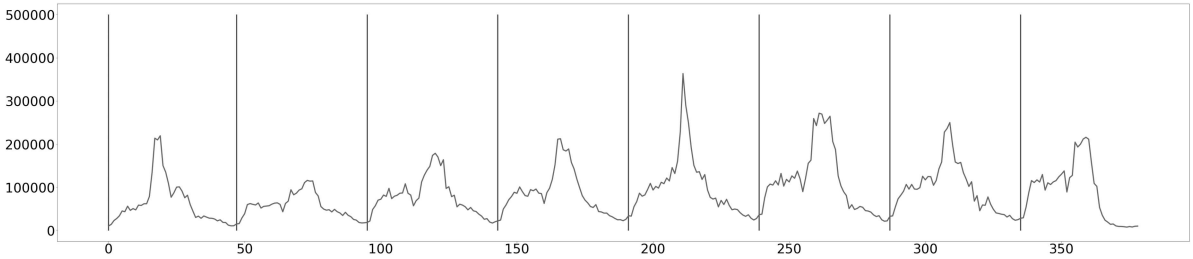
In [101]:

```
basic2['Total'].plot(figsize=(70, 15), linewidth=4)
plt.vlines(x=[0, 47, 95, 143, 191, 239, 287, 335], ymin=0, ymax=500000, lw=4, color =
plt.xticks(fontsize = 45)
plt.yticks(fontsize = 45)

```

Out[101]:

```
(array([-100000.,       0.,  100000.,  200000.,  300000.,  400000.,
         500000.,  600000.]), <a list of 8 Text yticklabel objects>)
```



In [98]:

```
basic2 = basic.iloc[96:475, 0:]
basic2 = basic2.reset_index(drop=True)
```

In [95]:

```
1 basic2
```

Out[95]:

	Unique_ID	Total	nr_tyg
95	2012_08D	9932	95
96	2012_09A	10530	96
97	2012_09B	14375	97
98	2012_09C	23150	98
99	2012_09D	27864	99
100	2012_10A	34232	100
101	2012_10B	45357	101
102	2012_10C	43668	102
103	2012_10D	56433	103
104	2012_11A	46712	104

In [57]:

```
1 plt.figure()
2 x = basic['nr_tyg']
3 y = basic['Total']
4 plt.show()
```

```
<Figure size 432x288 with 0 Axes>
```

In [46]:

```
1 basic.plot(figsize=(40, 15))
2
3 pyplot.show()
```