In [53]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
```

In [54]:
```
data=pd.read_csv("C:\\Users\\harsh\\Downloads\\amazon.csv",encoding='iso-8859-1',pa
data.head()
```

Out[54]:

	year	state	month	number	date
0	1998	Acre	Janeiro	0.0	1998-01-01
1	1999	Acre	Janeiro	0.0	1999-01-01
2	2000	Acre	Janeiro	0.0	2000-01-01
3	2001	Acre	Janeiro	0.0	2001-01-01
4	2002	Acre	Janeiro	0.0	2002-01-01

In [55]:
```
data.dtypes
```

Out[55]:
```
year                 int64
state               object
month               object
number             float64
date      datetime64[ns]
dtype: object
```

In [56]:
```
data['date'].dt.year
```

Out[56]:
```
0       1998
1       1999
2       2000
3       2001
4       2002
        ...
6449    2012
6450    2013
6451    2014
6452    2015
6453    2016
Name: date, Length: 6454, dtype: int64
```

In [57]:
```
data.shape
```

Out[57]:
```
(6454, 5)
```

In [58]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6454 entries, 0 to 6453
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype         
---  ------  --------------  -----         
 0   year    6454 non-null   int64         
 1   state   6454 non-null   object        
 2   month   6454 non-null   object        
 3   number  6454 non-null   float64       
 4   date    6454 non-null   datetime64[ns]
dtypes: datetime64[ns](1), float64(1), int64(1), object(2)
memory usage: 252.2+ KB
```

In [59]:
```
data.duplicated().sum()
```

Out[59]:
```
32
```

In [60]:
```
data=data.drop_duplicates()
data
```

Out[60]:

	year	state	month	number	date
0	1998	Acre	Janeiro	0.0	1998-01-01
1	1999	Acre	Janeiro	0.0	1999-01-01
2	2000	Acre	Janeiro	0.0	2000-01-01
3	2001	Acre	Janeiro	0.0	2001-01-01
4	2002	Acre	Janeiro	0.0	2002-01-01
...	...	...	...	...	...
6449	2012	Tocantins	Dezembro	128.0	2012-01-01
6450	2013	Tocantins	Dezembro	85.0	2013-01-01
6451	2014	Tocantins	Dezembro	223.0	2014-01-01
6452	2015	Tocantins	Dezembro	373.0	2015-01-01
6453	2016	Tocantins	Dezembro	119.0	2016-01-01

6422 rows × 5 columns

In [61]:
```
data.isna().sum()
```

Out[61]:
```
year      0
state     0
month     0
number    0
date      0
dtype: int64
```

In [62]:
```
data.describe(include='all',datetime_is_numeric=True)
```

Out[62]:

	year	state	month	number	date
count	6422.000000	6422	6422	6422.000000	6422
unique	NaN	23	12	NaN	NaN
top	NaN	Rio	Agosto	NaN	NaN
freq	NaN	697	540	NaN	NaN
mean	2007.490969	NaN	NaN	108.815178	2007-06-29 10:46:40.622859008
min	1998.000000	NaN	NaN	0.000000	1998-01-01 00:00:00
25%	2003.000000	NaN	NaN	3.000000	2003-01-01 00:00:00
50%	2007.000000	NaN	NaN	24.497000	2007-01-01 00:00:00
75%	2012.000000	NaN	NaN	114.000000	2012-01-01 00:00:00
max	2017.000000	NaN	NaN	998.000000	2017-01-01 00:00:00
std	5.731806	NaN	NaN	191.142482	NaN

In [63]:
```
data['month'].unique()
```

Out[63]:
```
array(['Janeiro', 'Fevereiro', 'Março', 'Abril', 'Maio', 'Junho', 'Julho',
       'Agosto', 'Setembro', 'Outubro', 'Novembro', 'Dezembro'],
      dtype=object)
```

In [64]:
```
data['month_new']=data['month'].map({'Janeiro':'January','Fevereiro':'February','Ma
```

In [65]:
```
data['month_new']
```

Out[65]:
```
0         January
1         January
2         January
3         January
4         January
           ...
6449     December
6450     December
6451     December
6452     December
6453     December
Name: month_new, Length: 6422, dtype: object
```

In [66]:
```
data['month_new'].unique()
```

Out[66]:
```
array(['January', 'February', 'March', 'April', 'May', 'June', 'July',
       'August', 'September', 'October', 'November', 'December'],
      dtype=object)
```

In [67]:
```
data['month_new'].value_counts()
```

Out[67]:
```
August       540
September    540
October      540
November     540
June         539
July         539
January      535
February     535
March        534
April        534
May          533
December     513
Name: month_new, dtype: int64
```

In [68]:
```
data
```

Out[68]:

	year	state	month	number	date	month_new
0	1998	Acre	Janeiro	0.0	1998-01-01	January
1	1999	Acre	Janeiro	0.0	1999-01-01	January
2	2000	Acre	Janeiro	0.0	2000-01-01	January
3	2001	Acre	Janeiro	0.0	2001-01-01	January
4	2002	Acre	Janeiro	0.0	2002-01-01	January
...	...	...	...	...	...	...
6449	2012	Tocantins	Dezembro	128.0	2012-01-01	December
6450	2013	Tocantins	Dezembro	85.0	2013-01-01	December
6451	2014	Tocantins	Dezembro	223.0	2014-01-01	December
6452	2015	Tocantins	Dezembro	373.0	2015-01-01	December
6453	2016	Tocantins	Dezembro	119.0	2016-01-01	December

6422 rows × 6 columns

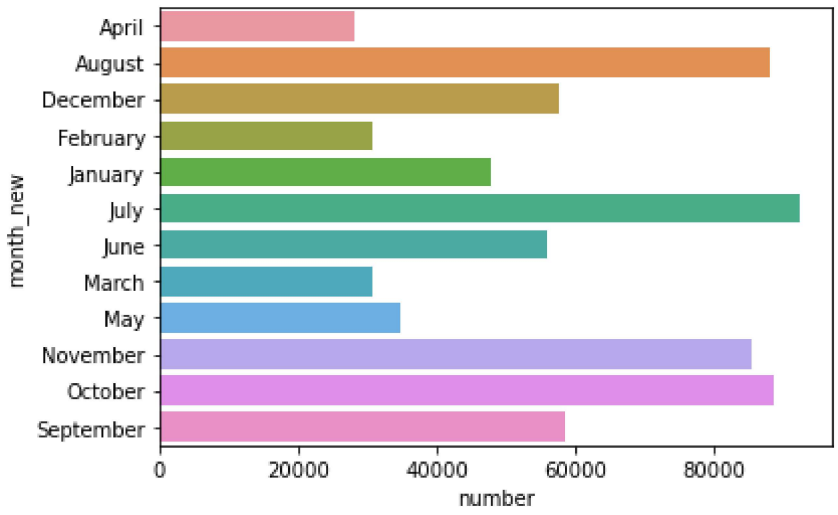
In [76]:
```
top=data.groupby('month_new').sum()[['number']]
top
```

Out[76]:

month_new	number
April	28184.770
August	88050.435
December	57535.480
February	30839.050
January	47681.844
July	92319.113
June	55997.675
March	30709.405
May	34725.363
November	85508.054
October	88681.579
September	58578.305

In [77]:
```
sns.barplot(x='number',y=top.index,data=top)
```

Out[77]:
```
<AxesSubplot:xlabel='number', ylabel='month_new'>
```



In [78]:
```
data.head(2)
```

Out[78]:

	year	state	month	number	date	month_new
0	1998	Acre	Janeiro	0.0	1998-01-01	January
1	1999	Acre	Janeiro	0.0	1999-01-01	January

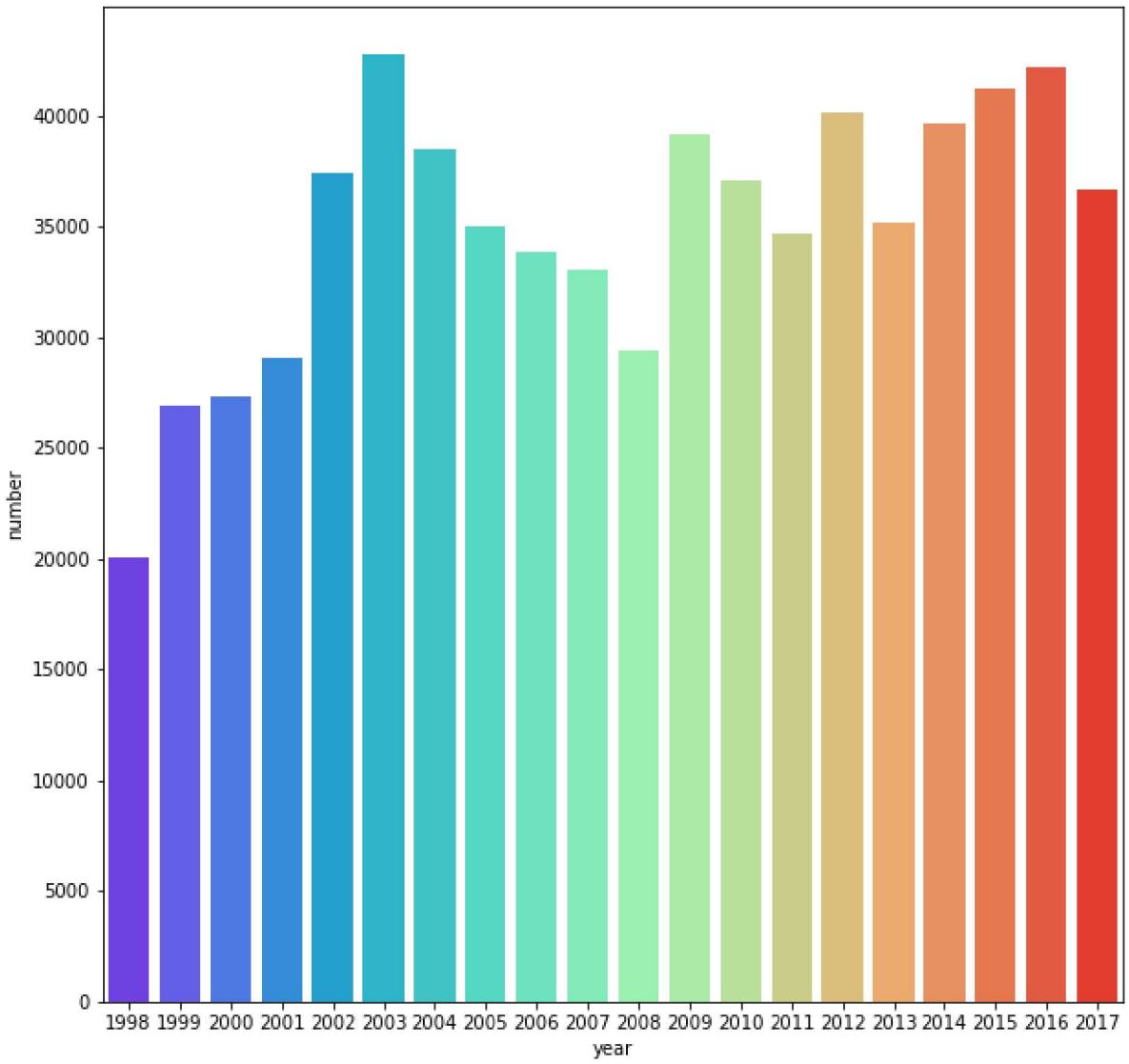
In [82]:
```
top=data.groupby('year').sum()
top
```

Out[82]:

year	number
1998	20013.971
1999	26882.821
2000	27351.251
2001	29054.612
2002	37390.600
2003	42760.674
2004	38450.163
2005	35004.965
2006	33824.161
2007	33028.413
2008	29378.964
2009	39116.178
2010	37037.449
2011	34633.545
2012	40084.860
2013	35137.118
2014	39621.183
2015	41208.292
2016	42212.229
2017	36619.624

In [95]:
```
plt.figure(figsize=(10,10))
sns.barplot(y='number',x=top.index,data=top,palette='rainbow')
```

Out[95]:
```
<AxesSubplot:xlabel='year', ylabel='number'>
```



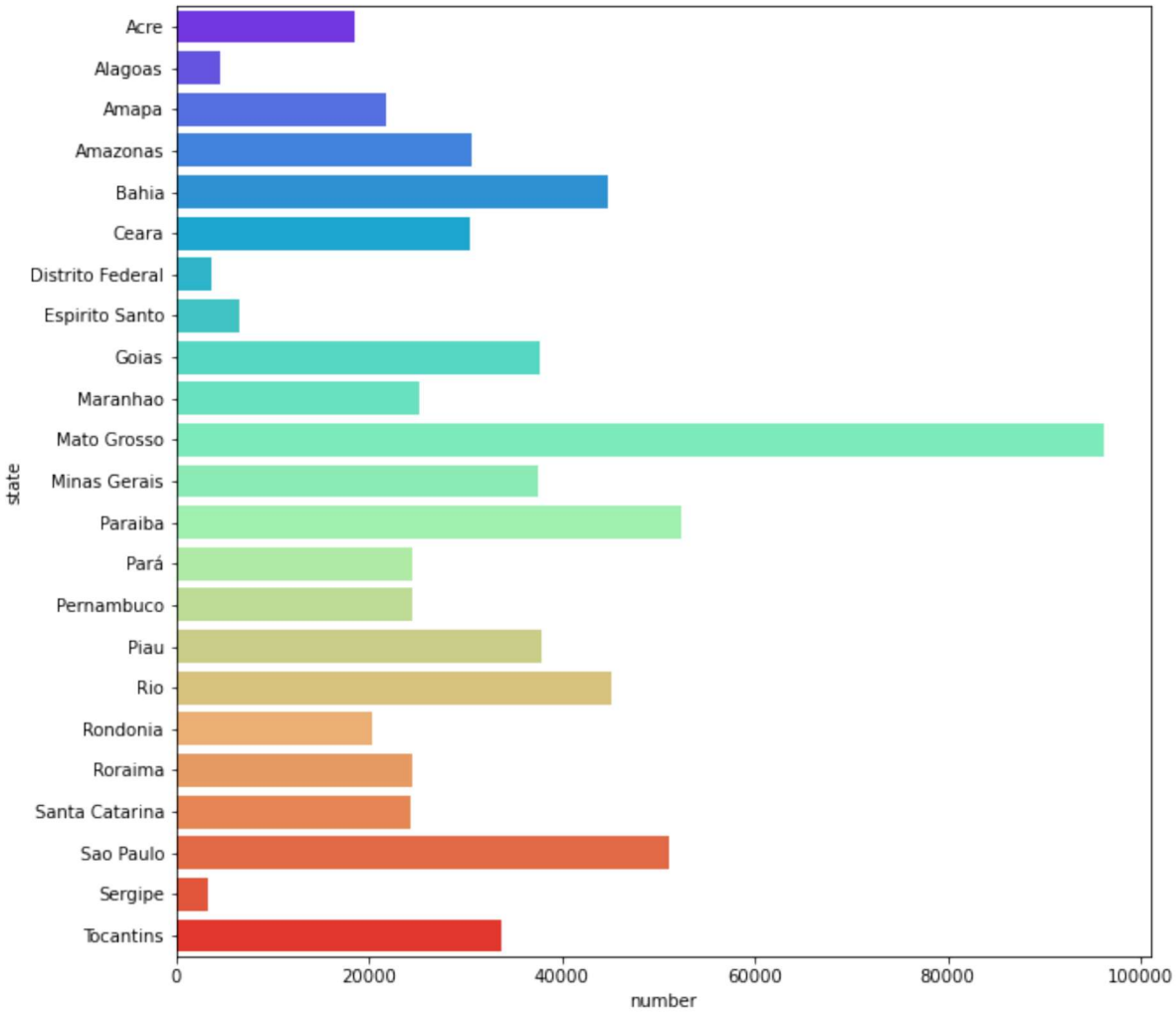
In [97]:
```
top=data.groupby('state').sum()[['number']]
top
```

Out[97]:

state	number
Acre	18464.030
Alagoas	4606.000
Amapa	21831.576
Amazonas	30650.129
Bahia	44746.226
Ceara	30428.063
Distrito Federal	3561.000
Espirito Santo	6546.000
Goias	37695.520
Maranhao	25129.131
Mato Grosso	96246.028
Minas Gerais	37475.258
Paraiba	52426.918
Pará	24512.144
Pernambuco	24498.000
Piau	37803.747
Rio	45094.865
Rondonia	20285.429
Roraima	24385.074
Santa Catarina	24359.852
Sao Paulo	51121.198
Sergipe	3237.000
Tocantins	33707.885

In [98]:
```
plt.figure(figsize=(10,10))
sns.barplot(x='number',y=top.index,data=top,palette='rainbow')
```

Out[98]:
```
<AxesSubplot:xlabel='number', ylabel='state'>
```



In [ ]:

In [102…
```
data.loc[data['state']=='Amazonas']['number'].sum()
```

Out[102]:
```
30650.129
```

In [106…
```
data.groupby(['state','year']).sum().loc['Amazonas']
```

Out[106]:

year	number
1998	946.000
1999	1061.000
2000	853.000
2001	1297.000
2002	2852.000
2003	1524.268
2004	2298.207
2005	1657.128
2006	997.640
2007	589.601
2008	2717.000
2009	1320.601
2010	2324.508
2011	1652.538
2012	1110.641
2013	905.217
2014	2385.909
2015	1189.994
2016	2060.972
2017	906.905

In [107…

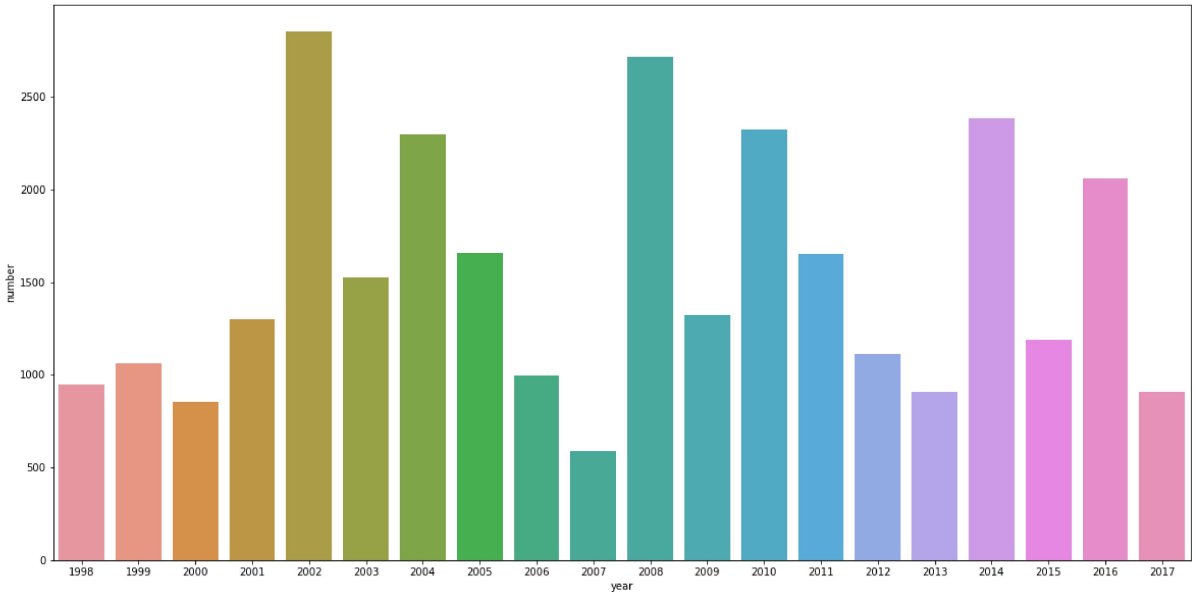
```
df=data.groupby(['state','year']).sum().loc['Amazonas']
```

In [112…

```
plt.figure(figsize=(20,10))
sns.barplot(x=df.index,y='number',data=df)
```

Out[112]:

```
<AxesSubplot:xlabel='year', ylabel='number'>
```



In [ ]:

In [132…
```
df=data.loc[data['state']=='Amazonas']
df
```

Out[132]:

	year	state	month	number	date	month_new
718	1998	Amazonas	Janeiro	0.0	1998-01-01	January
719	1999	Amazonas	Janeiro	3.0	1999-01-01	January
720	2000	Amazonas	Janeiro	7.0	2000-01-01	January
721	2001	Amazonas	Janeiro	3.0	2001-01-01	January
722	2002	Amazonas	Janeiro	17.0	2002-01-01	January
…	…	…	…	…	…	…
952	2012	Amazonas	Dezembro	80.0	2012-01-01	December
953	2013	Amazonas	Dezembro	236.0	2013-01-01	December
954	2014	Amazonas	Dezembro	293.0	2014-01-01	December
955	2015	Amazonas	Dezembro	565.0	2015-01-01	December
956	2016	Amazonas	Dezembro	133.0	2016-01-01	December

239 rows × 6 columns

In [133…
```
df['date'].dt.day_of_week
```

Out[133]:
```
718    3
719    4
720    5
721    0
722    1
      ..
952    6
953    1
954    2
955    3
956    4
Name: date, Length: 239, dtype: int64
```

In [141…
```
df2=df.groupby(df['date'].dt.day_of_week).sum()['number']
df2
```

Out[141]:
```
date
0    1886.601
1    6474.217
2    3910.177
3    5754.802
4    5446.480
5    4162.666
6    3015.186
Name: number, dtype: float64
```

In [142…
```
import calendar
```

In [143…
```
daynames=[calendar.day_name[x] for x in range(0,7)]
daynames
```

Out[143]:
```
['Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday']
```

In [146…
```
df2.index=daynames
df2
```

Out[146]:
```
Monday       1886.601
Tuesday      6474.217
Wednesday    3910.177
Thursday     5754.802
Friday       5446.480
Saturday     4162.666
Sunday       3015.186
Name: number, dtype: float64
```

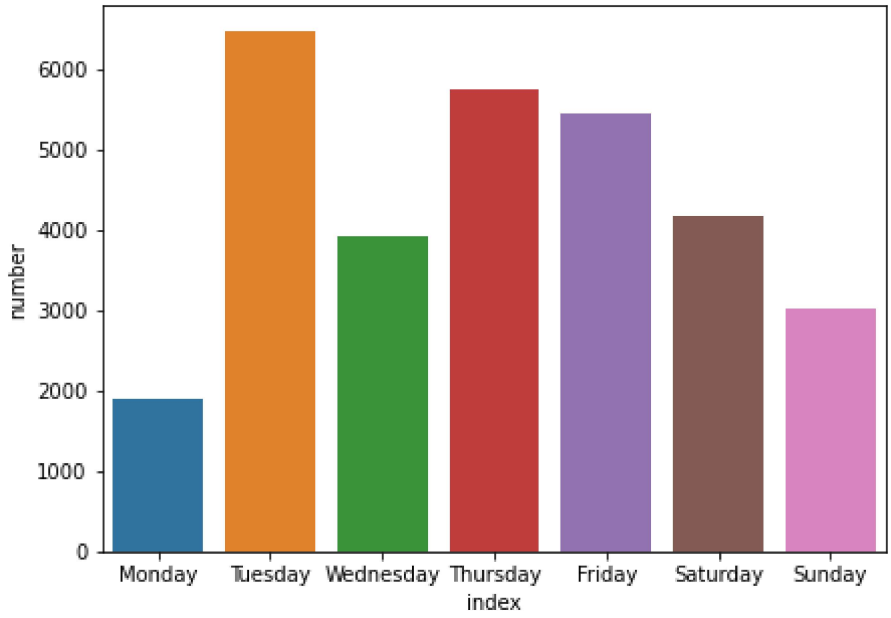
In [154…
```
df3=df2.reset_index()
df3
```

Out[154]:

	index	number
0	Monday	1886.601
1	Tuesday	6474.217
2	Wednesday	3910.177
3	Thursday	5754.802
4	Friday	5446.480
5	Saturday	4162.666
6	Sunday	3015.186

In [158…
```
plt.figure(figsize=(7,5))
sns.barplot(x='index',y='number',data=df3)
```

Out[158]:
```
<AxesSubplot:xlabel='index', ylabel='number'>
```



In [159…
```
df3
```

Out[159]:

	index	number
0	Monday	1886.601
1	Tuesday	6474.217
2	Wednesday	3910.177
3	Thursday	5754.802
4	Friday	5446.480
5	Saturday	4162.666
6	Sunday	3015.186

In [ ]:
```
```

In [173…
```
df=data.groupby(['year','month_new']).sum().loc[2015]
df
```

Out[173]:

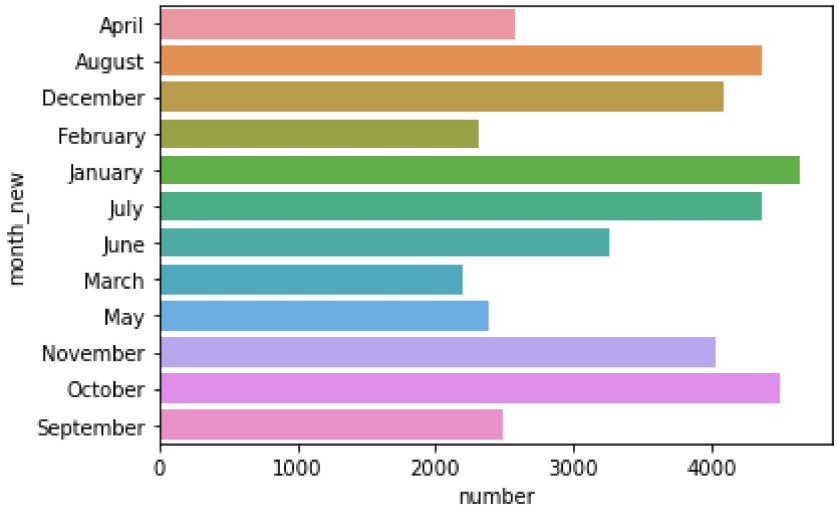
month_new	number
April	2573.000
August	4363.125
December	4088.522
February	2309.000
January	4635.000
July	4364.392
June	3260.552
March	2202.000
May	2384.000
November	4034.518
October	4499.525
September	2494.658

In [175…

```
sns.barplot(y=df.index,x='number',data=df)
```

Out[175]:

```
<AxesSubplot:xlabel='number', ylabel='month_new'>
```



In [179…

```
avg=data.groupby('state').mean()[['number']]
avg
```

Out[179]:

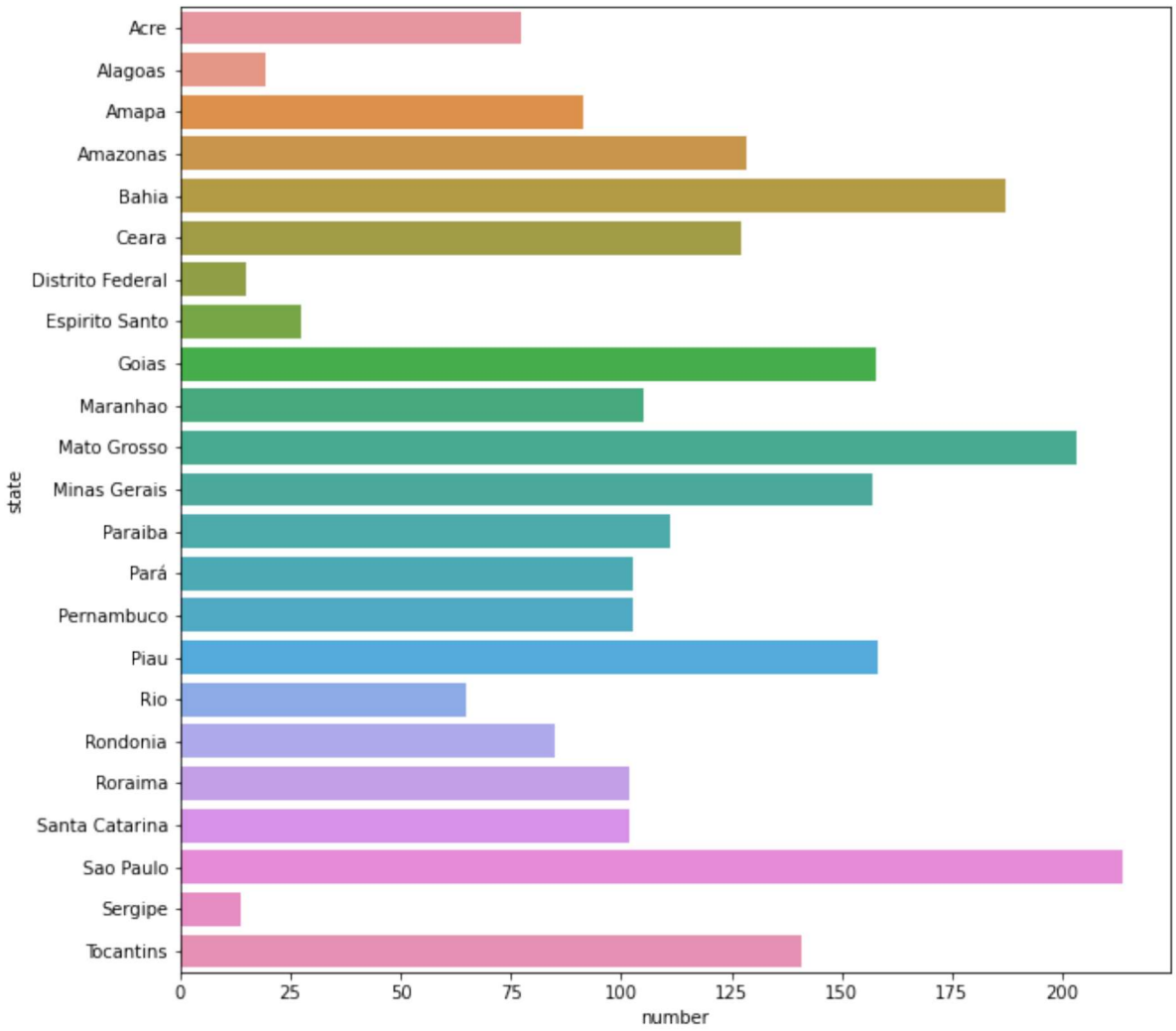
state	number
Acre	77.255356
Alagoas	19.271967
Amapa	91.345506
Amazonas	128.243218
Bahia	187.222703
Ceara	127.314071
Distrito Federal	14.899582
Espirito Santo	27.389121
Goias	157.721841
Maranhao	105.142808
Mato Grosso	203.479975
Minas Gerais	156.800243
Paraiba	111.073979
Pará	102.561272
Pernambuco	102.502092
Piau	158.174674
Rio	64.698515
Rondonia	84.876272
Roraima	102.029598
Santa Catarina	101.924067
Sao Paulo	213.896226
Sergipe	13.543933
Tocantins	141.037176

In [183…

```
plt.figure(figsize=(10,10))
sns.barplot(y=avg.index,x='number',data=avg)
```

Out[183]:

```
<AxesSubplot:xlabel='number', ylabel='state'>
```



In [190…
```
data.loc[data['month_new']=='December']['state'].unique()
```

Out[190]:
```
array(['Acre', 'Alagoas', 'Amapa', 'Amazonas', 'Bahia', 'Ceara',
       'Distrito Federal', 'Espirito Santo', 'Goias', 'Maranhao',
       'Mato Grosso', 'Minas Gerais', 'Pará', 'Paraiba', 'Pernambuco',
       'Piau', 'Rio', 'Rondonia', 'Roraima', 'Santa Catarina',
       'Sao Paulo', 'Sergipe', 'Tocantins'], dtype=object)
```

In [ ]: