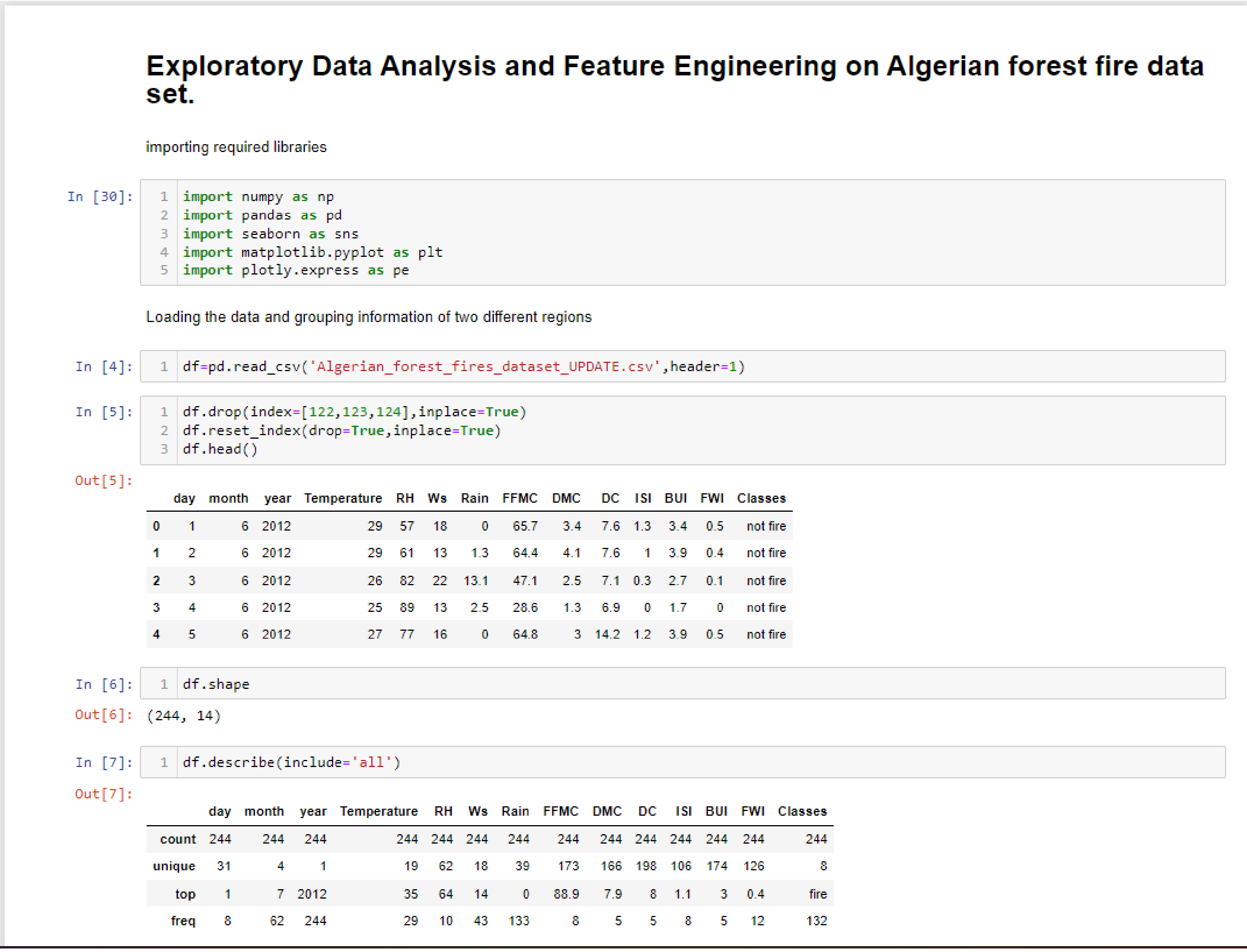


Exploratory Data Analysis and Feature Engineering on Algerian forest fire data set.

importing required libraries

In [30]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as pe
```

Loading the data and grouping information of two different regions

In [4]:

```
df=pd.read_csv('Algerian_forest_fires_dataset_UPDATE.csv',header=1)
```

In [5]:

```
df.drop(index=[122,123,124],inplace=True)
df.reset_index(drop=True,inplace=True)
df.head()
```

Out[5]:

	day	month	year	Temperature	RH	Ws	Rain	FFMC	DMC	DC	ISI	BUI	FWI	Classes
0	1	6	2012	29	57	18	0	65.7	3.4	7.6	1.3	3.4	0.5	not fire
1	2	6	2012	29	61	13	1.3	64.4	4.1	7.6	1	3.9	0.4	not fire
2	3	6	2012	26	82	22	13.1	47.1	2.5	7.1	0.3	2.7	0.1	not fire
3	4	6	2012	25	89	13	2.5	28.6	1.3	6.9	0	1.7	0	not fire
4	5	6	2012	27	77	16	0	64.8	3	14.2	1.2	3.9	0.5	not fire

In [6]:

```
df.shape
```

Out[6]:

```
(244, 14)
```

In [7]:

```
df.describe(include='all')
```

Out[7]:

	day	month	year	Temperature	RH	Ws	Rain	FFMC	DMC	DC	ISI	BUI	FWI	Classes
count	244	244	244	244	244	244	244	244	244	244	244	244	244	244
unique	31	4	1	19	62	18	39	173	166	198	106	174	126	8
top	1	7	2012	35	64	14	0	88.9	7.9	8	1.1	3	0.4	fire
freq	8	62	244	29	10	43	133	8	5	5	8	5	12	132

In [8]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 14 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   day          244 non-null    object
 1   month        244 non-null    object
 2   year         244 non-null    object
 3   Temperature  244 non-null    object
 4    RH          244 non-null    object
 5    Ws          244 non-null    object
 6   Rain         244 non-null    object
 7   FFMC         244 non-null    object
 8   DMC          244 non-null    object
 9   DC           244 non-null    object
 10  ISI          244 non-null    object
 11  BUI          244 non-null    object
 12  FWI          244 non-null    object
 13  Classes      244 non-null    object
dtypes: object(14)
memory usage: 26.8+ KB
```

In [9]:

```
df
```

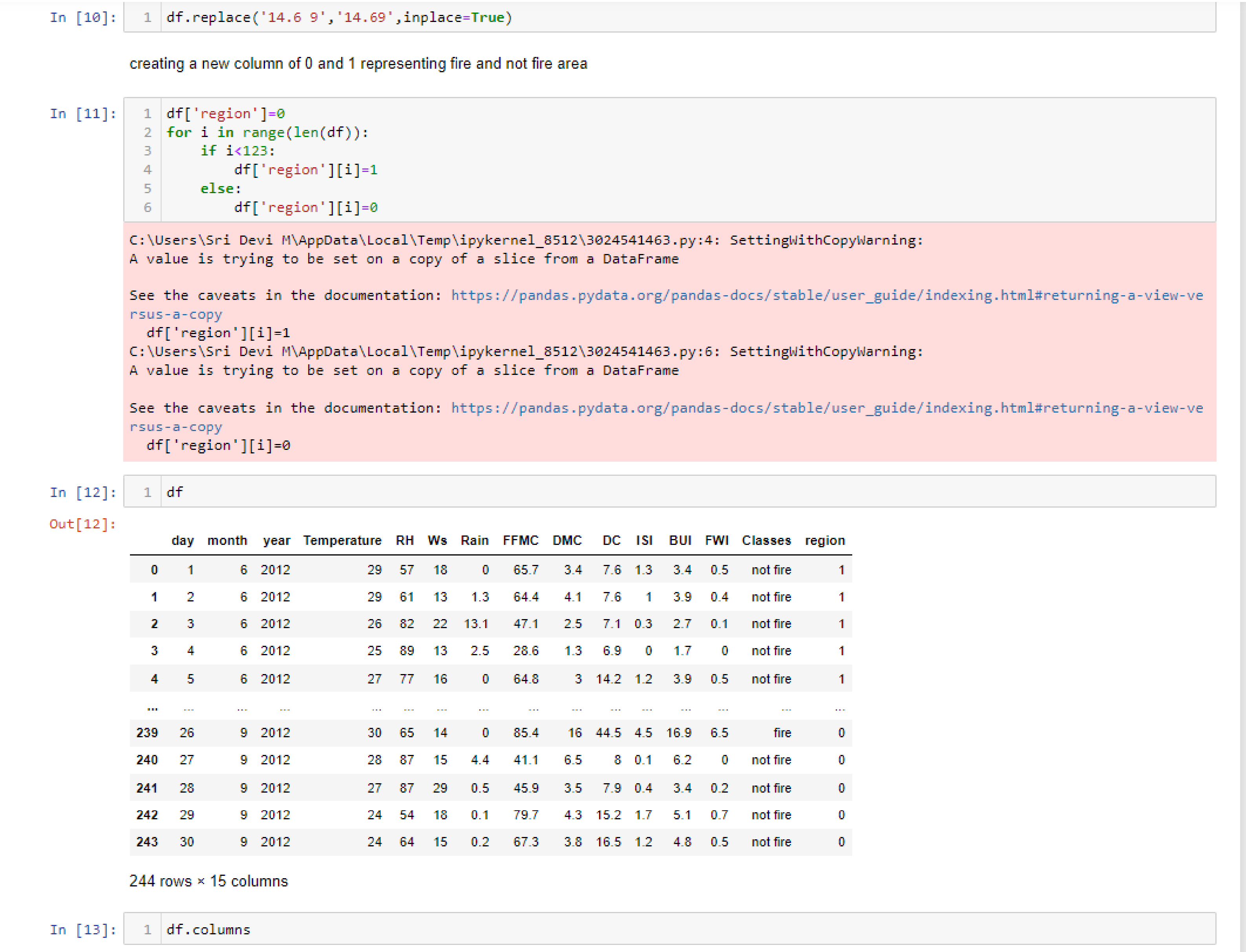
Out[9]:

	day	month	year	Temperature	RH	Ws	Rain	FFMC	DMC	DC	ISI	BUI	FWI	Classes
0	1	6	2012	29	57	18	0	65.7	3.4	7.6	1.3	3.4	0.5	not fire
1	2	6	2012	29	61	13	1.3	64.4	4.1	7.6	1	3.9	0.4	not fire
2	3	6	2012	26	82	22	13.1	47.1	2.5	7.1	0.3	2.7	0.1	not fire
3	4	6	2012	25	89	13	2.5	28.6	1.3	6.9	0	1.7	0	not fire
4	5	6	2012	27	77	16	0	64.8	3	14.2	1.2	3.9	0.5	not fire
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
239	26	9	2012	30	65	14	0	85.4	16	44.5	4.5	16.9	6.5	fire
240	27	9	2012	28	87	15	4.4	41.1	6.5	8	0.1	6.2	0	not fire
241	28	9	2012	27	87	29	0.5	45.9	3.5	7.9	0.4	3.4	0.2	not fire
242	29	9	2012	24	54	18	0.1	79.7	4.3	15.2	1.7	5.1	0.7	not fire
243	30	9	2012	24	64	15	0.2	67.3	3.8	16.5	1.2	4.8	0.5	not fire

244 rows × 14 columns

In [10]:

```
df.replace('14.6 9','14.69',inplace=True)
```



In [10]:

```
df.replace('14.6 9','14.69',inplace=True)
```

creating a new column of 0 and 1 representing fire and not fire area

In [11]:

```
df['region']=0
for i in range(len(df)):
    if i<123:
        df['region'][i]=1
    else:
        df['region'][i]=0
```

```
C:\Users\Sri Devi M\AppData\Local\Temp\ipykernel_8512\3024541463.py:4: 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-ve
rsus-a-copy
  df['region'][i]=1
C:\Users\Sri Devi M\AppData\Local\Temp\ipykernel_8512\3024541463.py:6: 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-ve
rsus-a-copy
  df['region'][i]=0
```

In [12]:

```
df
```

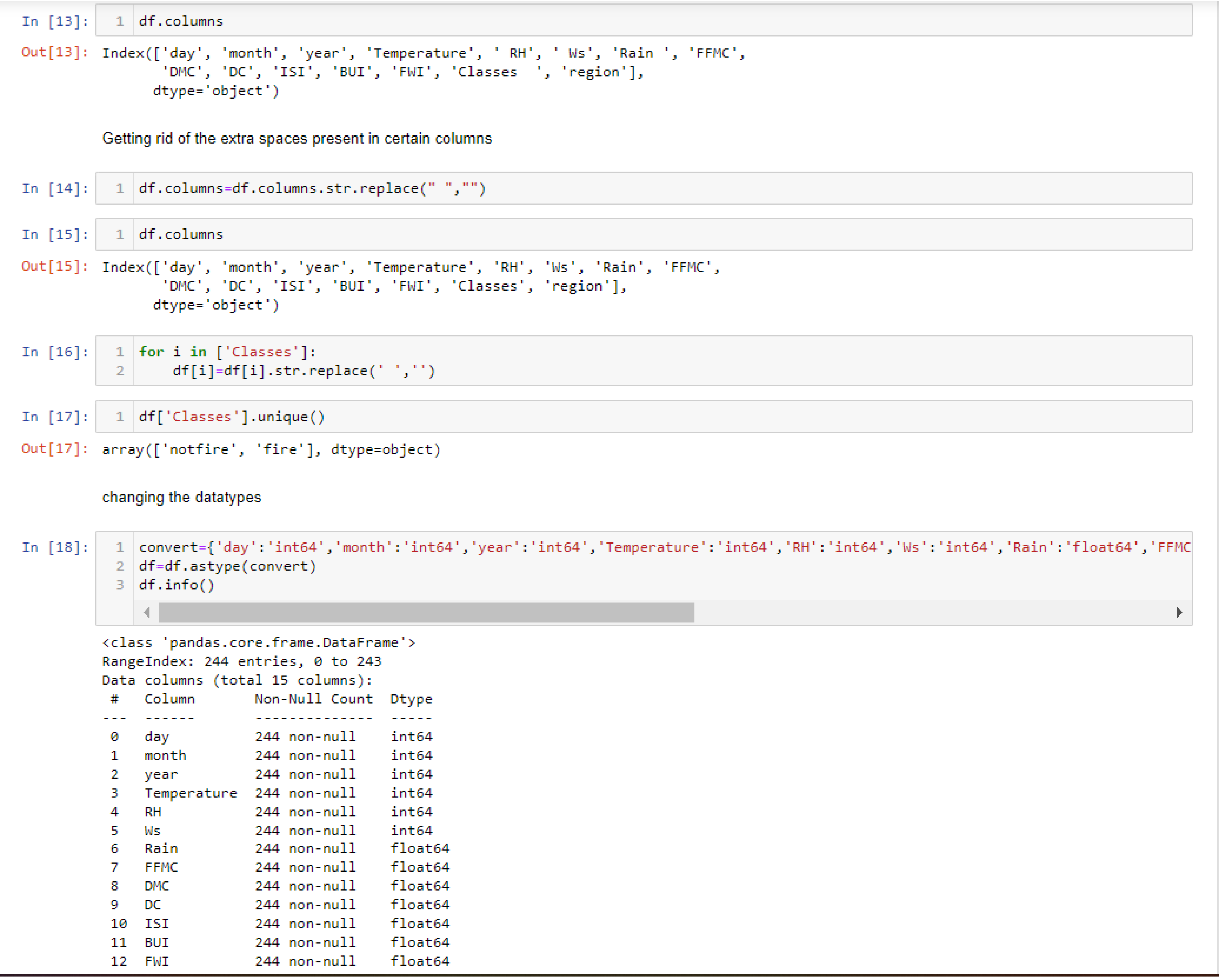
Out[12]:

	day	month	year	Temperature	RH	Ws	Rain	FFMC	DMC	DC	ISI	BUI	FWI	Classes	region
0	1	6	2012	29	57	18	0	65.7	3.4	7.6	1.3	3.4	0.5	not fire	1
1	2	6	2012	29	61	13	1.3	64.4	4.1	7.6	1	3.9	0.4	not fire	1
2	3	6	2012	26	82	22	13.1	47.1	2.5	7.1	0.3	2.7	0.1	not fire	1
3	4	6	2012	25	89	13	2.5	28.6	1.3	6.9	0	1.7	0	not fire	1
4	5	6	2012	27	77	16	0	64.8	3	14.2	1.2	3.9	0.5	not fire	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
239	26	9	2012	30	65	14	0	85.4	16	44.5	4.5	16.9	6.5	fire	0
240	27	9	2012	28	87	15	4.4	41.1	6.5	8	0.1	6.2	0	not fire	0
241	28	9	2012	27	87	29	0.5	45.9	3.5	7.9	0.4	3.4	0.2	not fire	0
242	29	9	2012	24	54	18	0.1	79.7	4.3	15.2	1.7	5.1	0.7	not fire	0
243	30	9	2012	24	64	15	0.2	67.3	3.8	16.5	1.2	4.8	0.5	not fire	0

244 rows × 15 columns

In [13]:

```
df.columns
```



```
In [13]: df.columns
```

```
Out[13]: Index(['day', 'month', 'year', 'Temperature', ' RH', ' Ws', 'Rain ', 'FFMC',
                'DMC', 'DC', 'ISI', 'BUI', 'FWI', 'Classes  ', 'region'],
               dtype='object')
```

Getting rid of the extra spaces present in certain columns

```
In [14]: df.columns=df.columns.str.replace(" ","")
```

```
In [15]: df.columns
```

```
Out[15]: Index(['day', 'month', 'year', 'Temperature', 'RH', 'Ws', 'Rain', 'FFMC',
                'DMC', 'DC', 'ISI', 'BUI', 'FWI', 'Classes', 'region'],
               dtype='object')
```

```
In [16]: for i in ['Classes']:
             df[i]=df[i].str.replace(' ','')
```

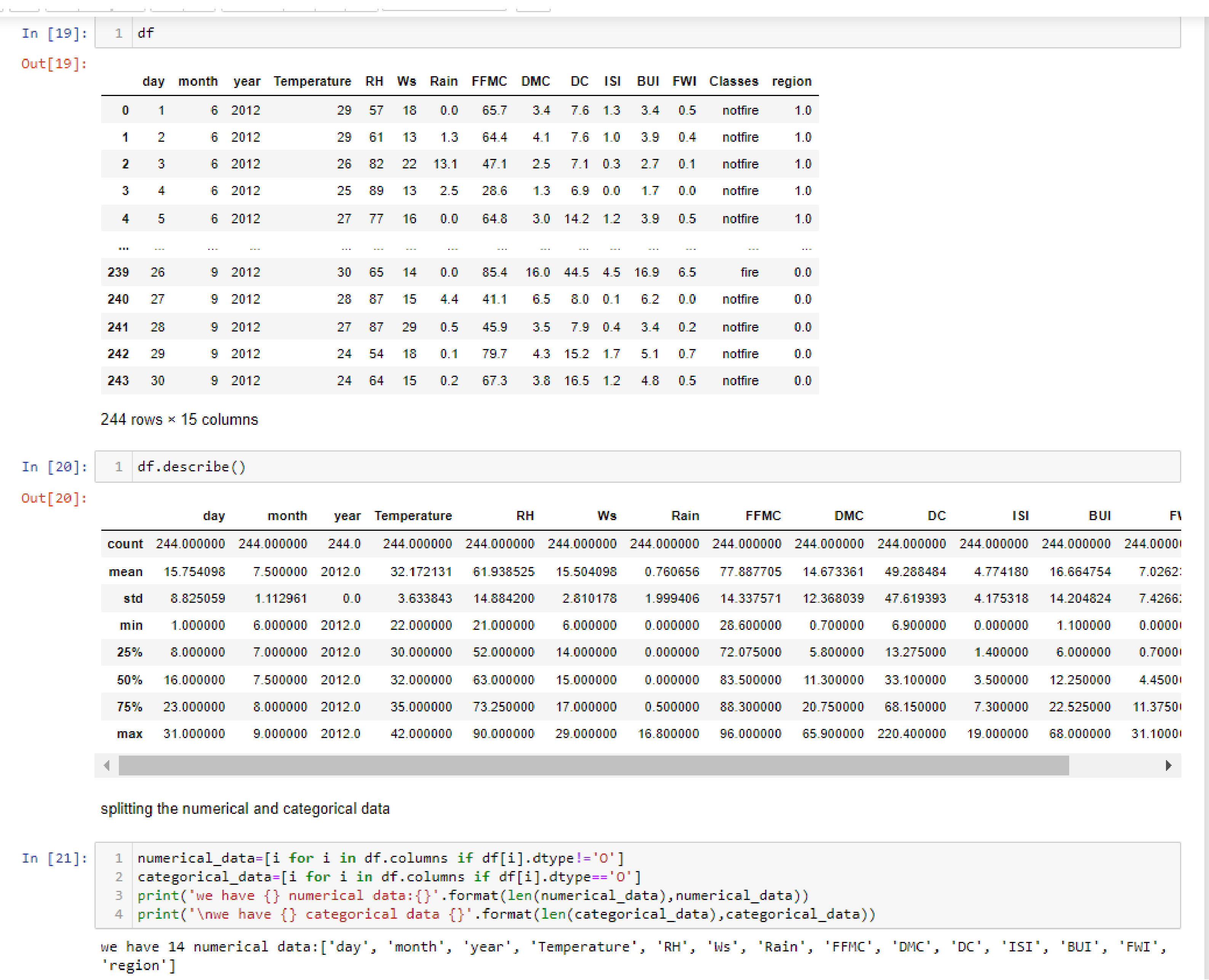
```
In [17]: df['Classes'].unique()
```

```
Out[17]: array(['notfire', 'fire'], dtype=object)
```

changing the datatypes

```
In [18]: convert={'day':'int64','month':'int64','year':'int64','Temperature':'int64','RH':'int64','Ws':'int64','Rain':'float64','FFMC
         df=df.astype(convert)
         df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 15 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   day          244 non-null    int64  
 1   month        244 non-null    int64  
 2   year         244 non-null    int64  
 3   Temperature  244 non-null    int64  
 4   RH           244 non-null    int64  
 5   Ws           244 non-null    int64  
 6   Rain         244 non-null    float64
 7   FFMC         244 non-null    float64
 8   DMC          244 non-null    float64
 9   DC           244 non-null    float64
 10  ISI          244 non-null    float64
 11  BUI          244 non-null    float64
 12  FWI          244 non-null    float64
```



In [19]:

```
df
```

Out[19]:

	day	month	year	Temperature	RH	Ws	Rain	FFMC	DMC	DC	ISI	BUI	FWI	Classes	region
0	1	6	2012	29	57	18	0.0	65.7	3.4	7.6	1.3	3.4	0.5	notfire	1.0
1	2	6	2012	29	61	13	1.3	64.4	4.1	7.6	1.0	3.9	0.4	notfire	1.0
2	3	6	2012	26	82	22	13.1	47.1	2.5	7.1	0.3	2.7	0.1	notfire	1.0
3	4	6	2012	25	89	13	2.5	28.6	1.3	6.9	0.0	1.7	0.0	notfire	1.0
4	5	6	2012	27	77	16	0.0	64.8	3.0	14.2	1.2	3.9	0.5	notfire	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
239	26	9	2012	30	65	14	0.0	85.4	16.0	44.5	4.5	16.9	6.5	fire	0.0
240	27	9	2012	28	87	15	4.4	41.1	6.5	8.0	0.1	6.2	0.0	notfire	0.0
241	28	9	2012	27	87	29	0.5	45.9	3.5	7.9	0.4	3.4	0.2	notfire	0.0
242	29	9	2012	24	54	18	0.1	79.7	4.3	15.2	1.7	5.1	0.7	notfire	0.0
243	30	9	2012	24	64	15	0.2	67.3	3.8	16.5	1.2	4.8	0.5	notfire	0.0

244 rows × 15 columns

In [20]:

```
df.describe()
```

Out[20]:

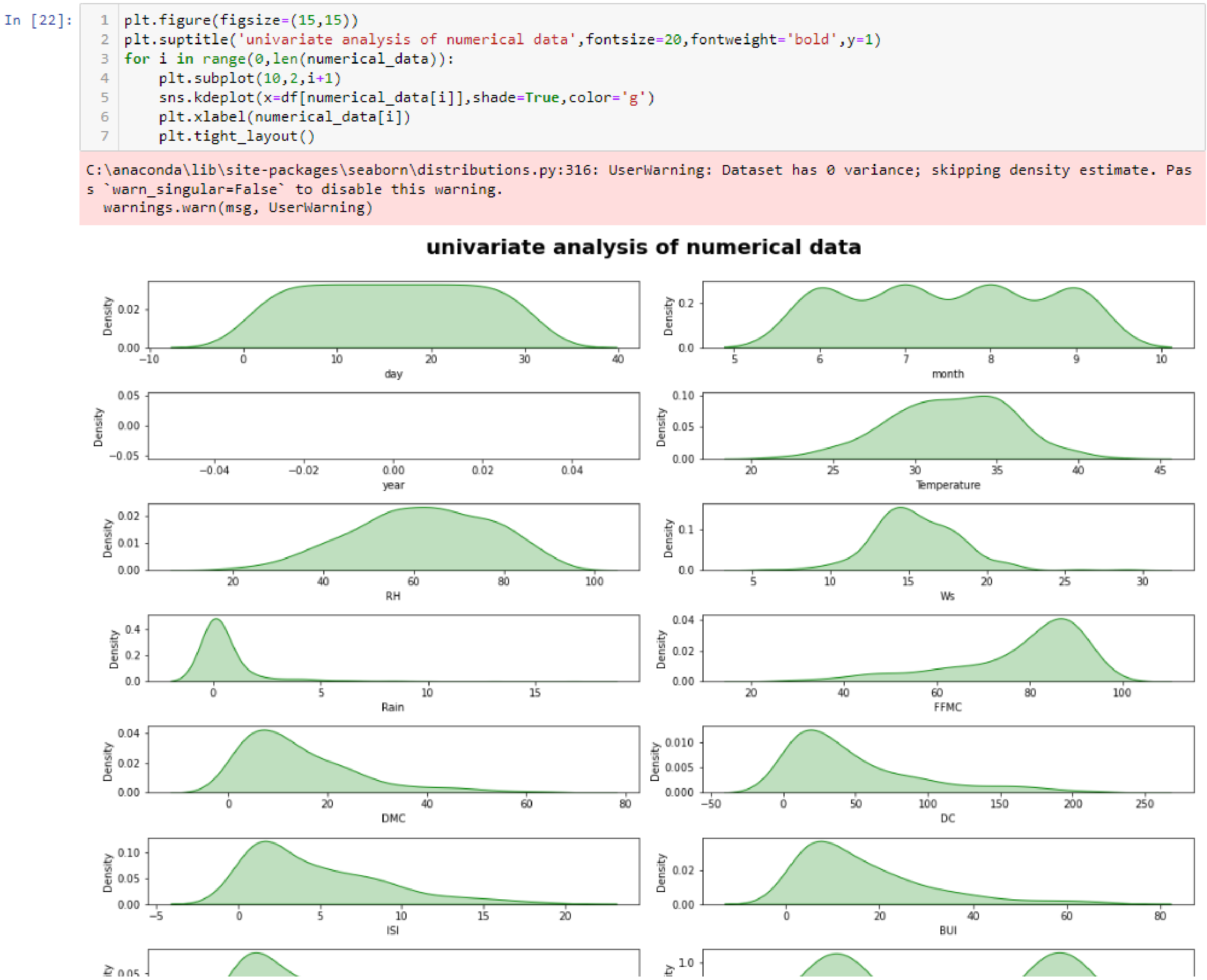
	day	month	year	Temperature	RH	Ws	Rain	FFMC	DMC	DC	ISI	BUI	F
count	244.000000	244.000000	244.0	244.000000	244.000000	244.000000	244.000000	244.000000	244.000000	244.000000	244.000000	244.000000	244.00000
mean	15.754098	7.500000	2012.0	32.172131	61.938525	15.504098	0.760656	77.887705	14.673361	49.288484	4.774180	16.664754	7.0262
std	8.825059	1.112961	0.0	3.633843	14.884200	2.810178	1.999406	14.337571	12.368039	47.619393	4.175318	14.204824	7.4266
min	1.000000	6.000000	2012.0	22.000000	21.000000	6.000000	0.000000	28.600000	0.700000	6.900000	0.000000	1.100000	0.0000
25%	8.000000	7.000000	2012.0	30.000000	52.000000	14.000000	0.000000	72.075000	5.800000	13.275000	1.400000	6.000000	0.7000
50%	16.000000	7.500000	2012.0	32.000000	63.000000	15.000000	0.000000	83.500000	11.300000	33.100000	3.500000	12.250000	4.4500
75%	23.000000	8.000000	2012.0	35.000000	73.250000	17.000000	0.500000	88.300000	20.750000	68.150000	7.300000	22.525000	11.3750
max	31.000000	9.000000	2012.0	42.000000	90.000000	29.000000	16.800000	96.000000	65.900000	220.400000	19.000000	68.000000	31.1000

splitting the numerical and categorical data

In [21]:

```
numerical_data=[i for i in df.columns if df[i].dtype!='O']
categorical_data=[i for i in df.columns if df[i].dtype=='O']
print('we have {} numerical data:{}'.format(len(numerical_data),numerical_data))
print('\nwe have {} categorical data {}'.format(len(categorical_data),categorical_data))
```

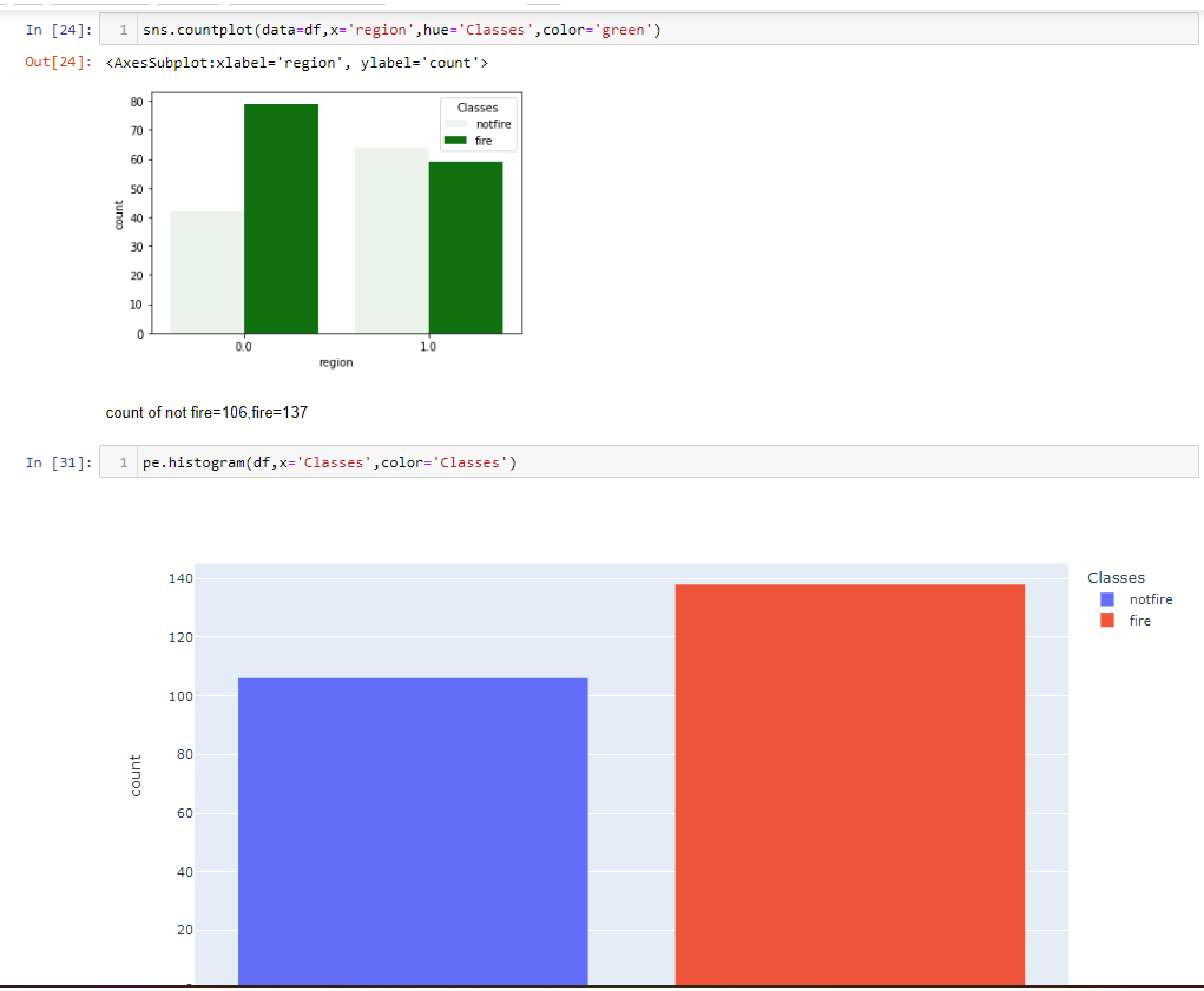
```
we have 14 numerical data:['day', 'month', 'year', 'Temperature', 'RH', 'Ws', 'Rain', 'FFMC', 'DMC', 'DC', 'ISI', 'BUI', 'FWI',
'region']
```



In [22]:

```
plt.figure(figsize=(15,15))
plt.suptitle('univariate analysis of numerical data',fontsize=20,fontweight='bold',y=1)
for i in range(0,len(numerical_data)):
    plt.subplot(10,2,i+1)
    sns.kdeplot(x=df[numerical_data[i]],shade=True,color='g')
    plt.xlabel(numerical_data[i])
    plt.tight_layout()
```

```
C:\anaconda\lib\site-packages\seaborn\distributions.py:316: UserWarning: Dataset has 0 variance; skipping density estimate. Pas
s `warn_singular=False` to disable this warning.
  warnings.warn(msg, UserWarning)
```



```
In [24]:  sns.countplot(data=df,x='region',hue='Classes',color='green')
```

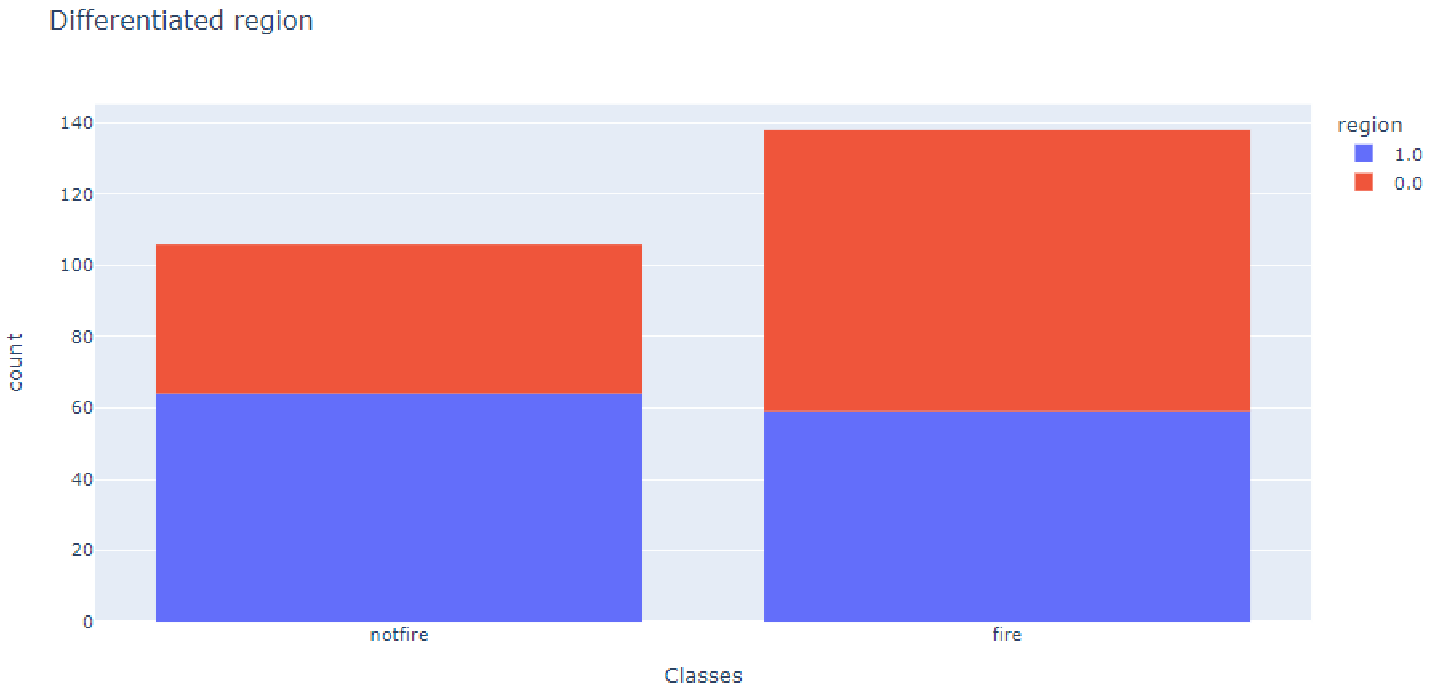
```
Out[24]: <AxesSubplot:xlabel='region', ylabel='count'>
```

count of not fire=106,fire=137

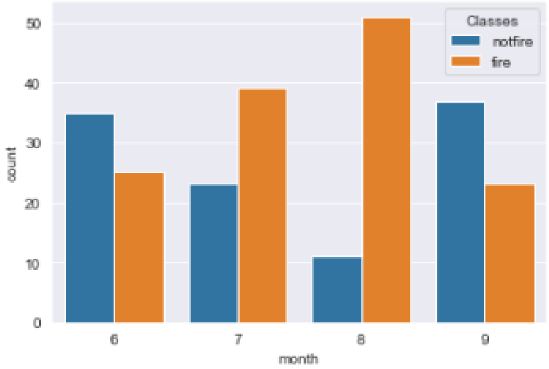
```
In [31]:  pe.histogram(df,x='Classes',color='Classes')
```

The below graph gives a clear view on both regions

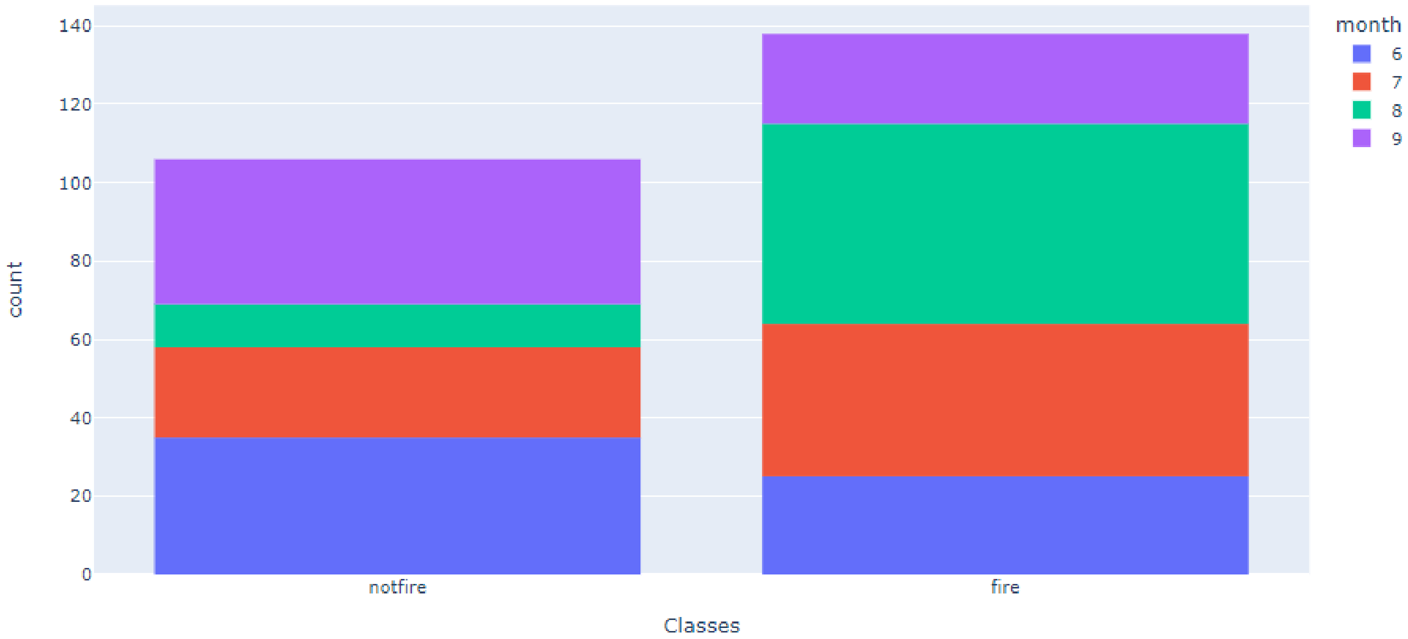
```
In [32]: pe.histogram(df,x='Classes',color='region',title='Differentiated region')
```

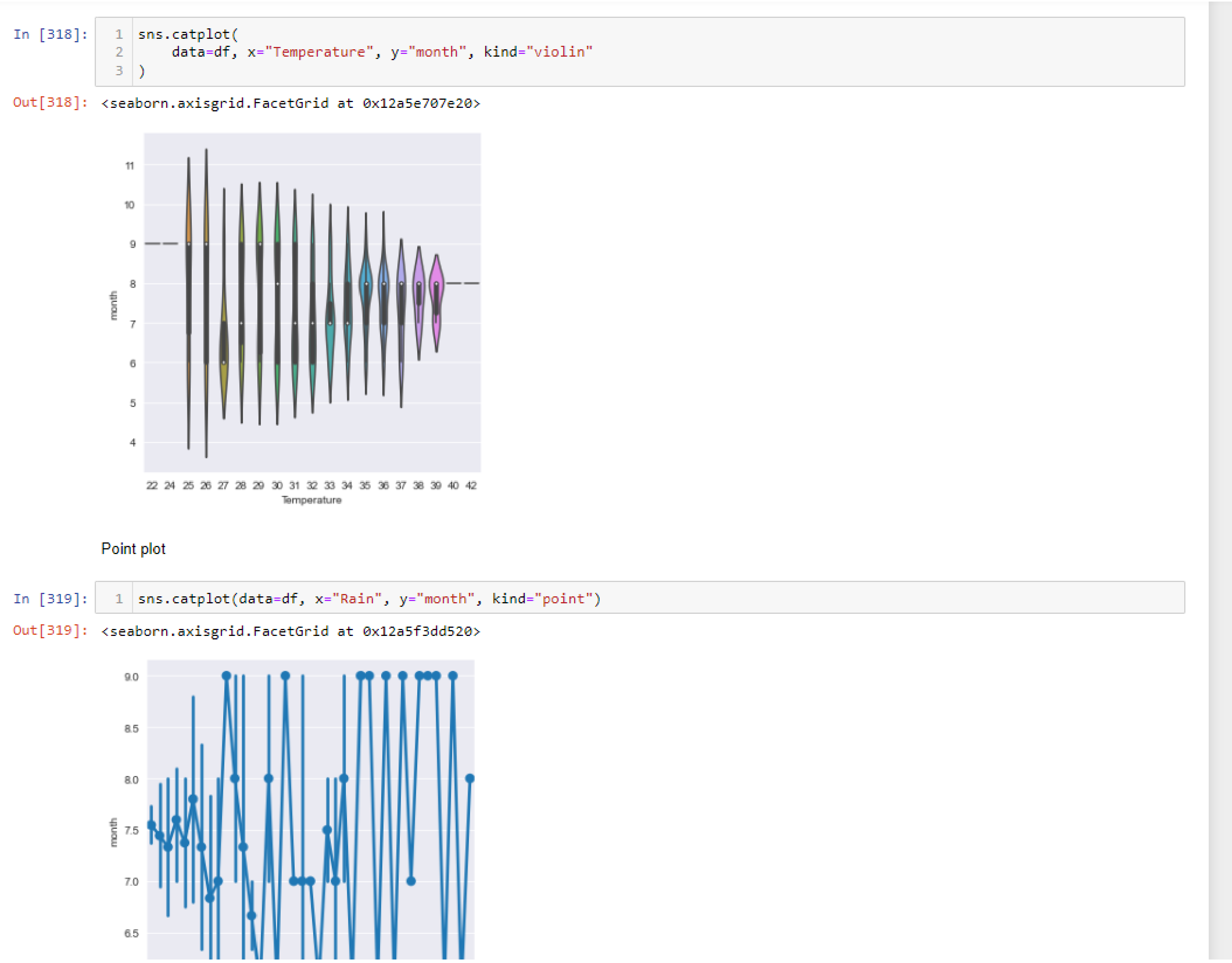


```
In [315]:  sns.countplot(data=df,x='month',hue='Classes')
Out[315]: <AxesSubplot:xlabel='month', ylabel='count'>
```



```
In [34]:  pe.histogram(df,x='Classes',color='month')
```





In [318]:

```
sns.catplot(
    data=df, x="Temperature", y="month", kind="violin"
)
```

Out[318]:

```
<seaborn.axisgrid.FacetGrid at 0x12a5e707e20>
```

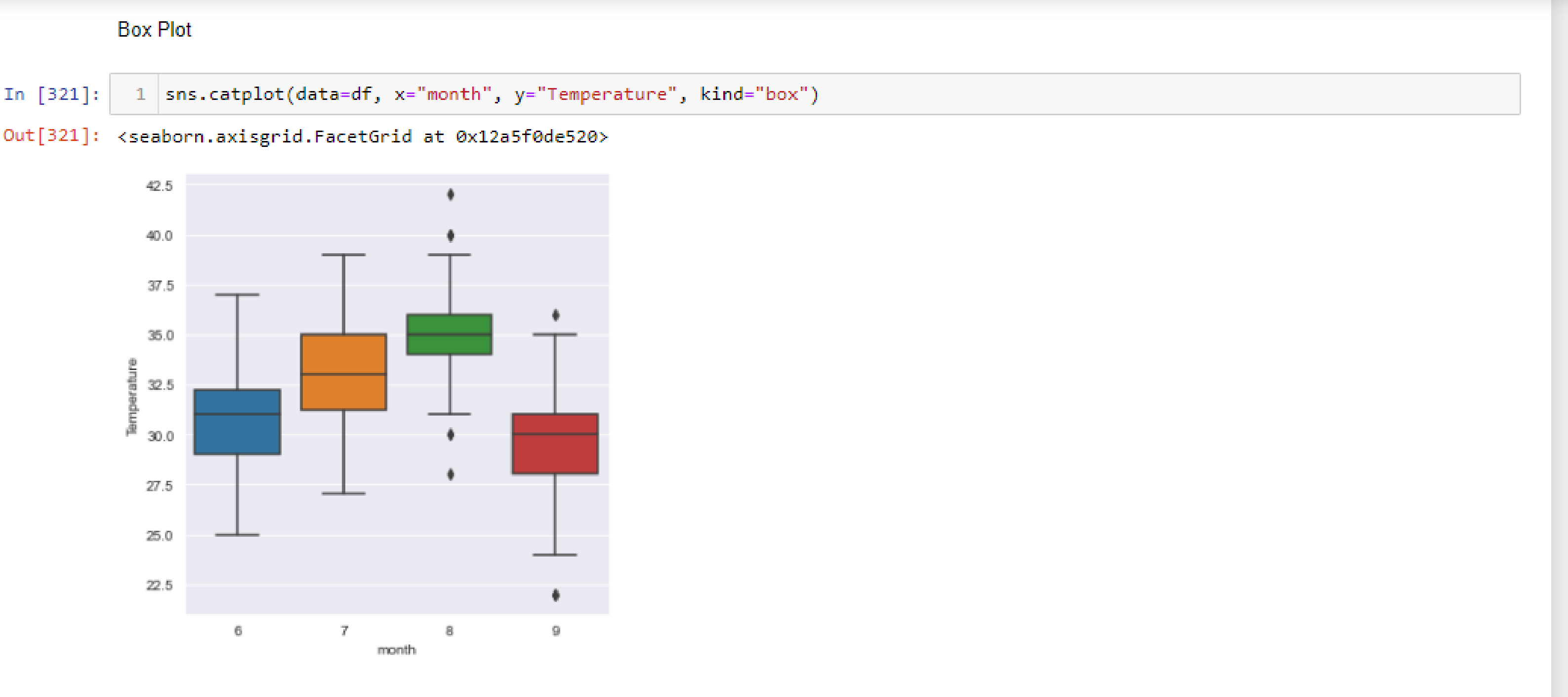
Point plot

In [319]:

```
sns.catplot(data=df, x="Rain", y="month", kind="point")
```

Out[319]:

```
<seaborn.axisgrid.FacetGrid at 0x12a5f3dd520>
```

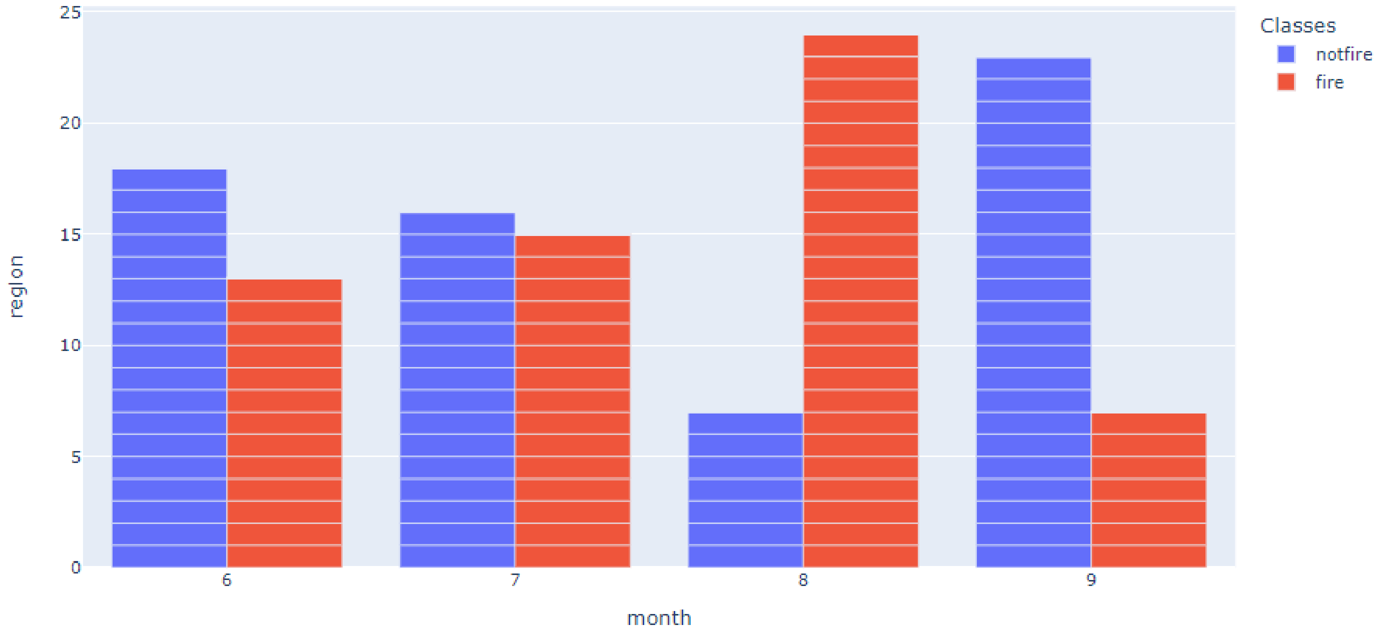


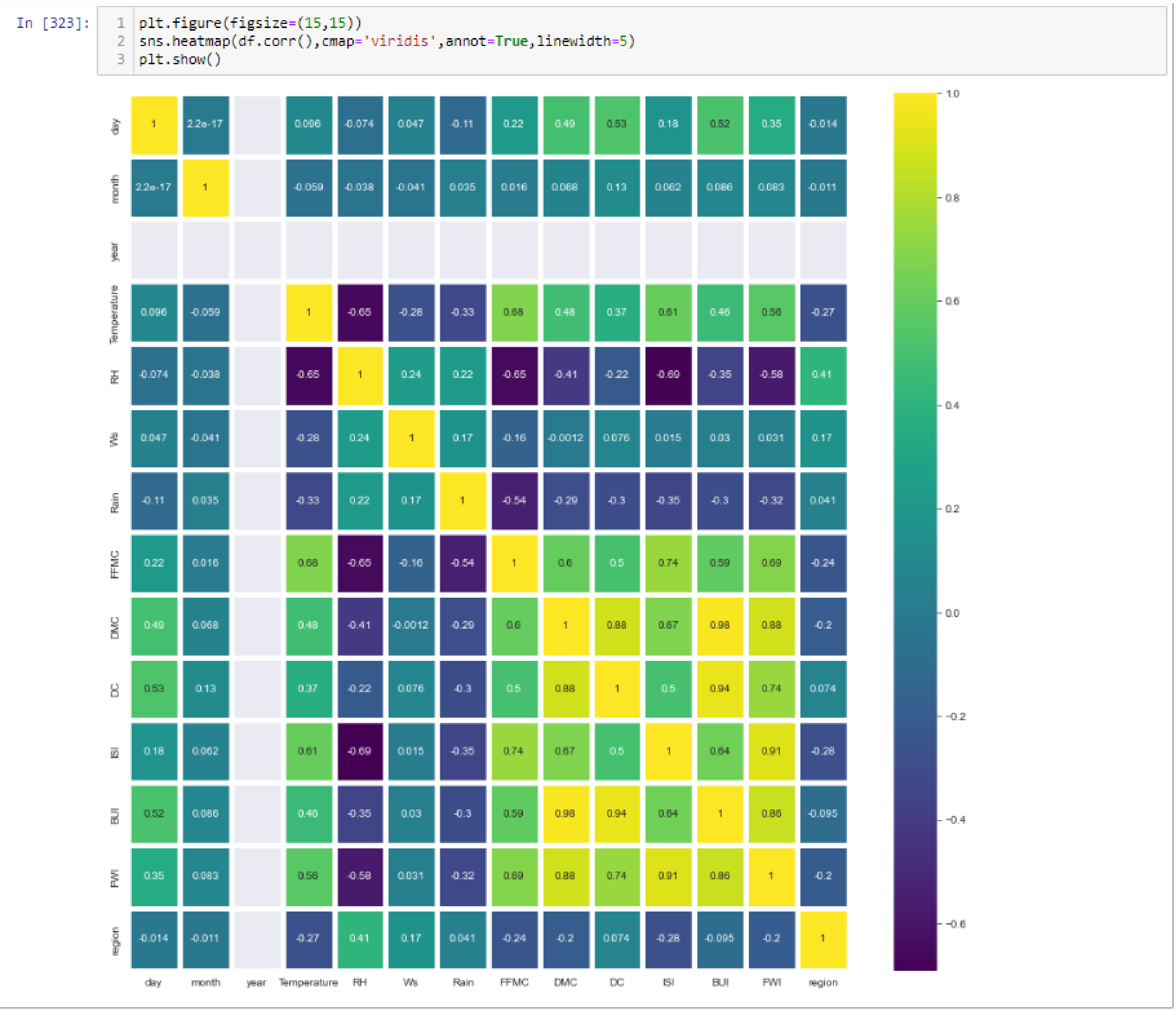
Box Plot

```
In [321]: sns.catplot(data=df, x="month", y="Temperature", kind="box")
Out[321]: <seaborn.axisgrid.FacetGrid at 0x12a5f0de520>
```

Grouped Bar

```
In [35]: pe.bar(df,x='month',y='region',color='Classes',barmode='group')
```





In [323]:

```
plt.figure(figsize=(15,15))
sns.heatmap(df.corr(),cmap='viridis',annot=True,linewidth=5)
plt.show()
```

	day	month	year	Temperature	RH	Ws	Rain	FFMC	DMC	DC	ISI	BUI	FWI	region
day	1	2.2e-17		0.096	-0.074	0.047	-0.11	0.22	0.49	0.53	0.18	0.52	0.35	-0.014
month	2.2e-17	1		-0.059	-0.038	-0.041	0.035	0.016	0.068	0.13	0.062	0.086	0.083	-0.011
year														
Temperature	0.096	-0.059		1	-0.65	-0.28	-0.33	0.68	0.48	0.37	0.61	0.46	0.56	-0.27
RH	-0.074	-0.038		-0.65	1	0.24	0.22	-0.65	-0.41	-0.22	-0.69	-0.35	-0.58	0.41
Ws	0.047	-0.041		-0.28	0.24	1	0.17	-0.16	-0.0012	0.076	0.015	0.03	0.031	0.17
Rain	-0.11	0.035		-0.33	0.22	0.17	1	-0.54	-0.29	-0.3	-0.35	-0.3	-0.32	0.041
FFMC	0.22	0.016		0.68	-0.65	-0.16	-0.54	1	0.6	0.5	0.74	0.59	0.69	-0.24
DMC	0.49	0.068		0.48	-0.41	-0.0012	-0.29	0.6	1	0.88	0.67	0.98	0.88	-0.2
DC	0.53	0.13		0.37	-0.22	0.076	-0.3	0.5	0.88	1	0.5	0.94	0.74	0.074
ISI	0.18	0.062		0.61	-0.69	0.015	-0.35	0.74	0.67	0.5	1	0.64	0.91	-0.28
BUI	0.52	0.086		0.46	-0.35	0.03	-0.3	0.59	0.98	0.94	0.64	1	0.86	-0.095
FWI	0.35	0.083		0.56	-0.58	0.031	-0.32	0.69	0.88	0.74	0.91	0.86	1	-0.2
region	-0.014	-0.011		-0.27	0.41	0.17	0.041	-0.24	-0.2	0.074	-0.28	-0.095	-0.2	1

END