Data Analysis

In [1]:
```
from IPython.display import Image
Image('Desktop/235365-(2)-نصائح-لمذاكرة-الدروس-قبل-الإمتحانات.jpg')
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
File E:\Anaconda\lib\site-packages\IPython\core\display.py:1032, in Image._da
ta_and_metadata(self, always_both)
   1031 try:
-> 1032     b64_data = b2a_base64(self.data).decode('ascii')
   1033 except TypeError as e:

TypeError: a bytes-like object is required, not 'str'

The above exception was the direct cause of the following exception:

FileNotFoundError                         Traceback (most recent call last)
File E:\Anaconda\lib\site-packages\IPython\core\formatters.py:973, in MimeBun
dleFormatter.__call__(self, obj, include, exclude)
    970     method = get_real_method(obj, self.print_method)
    972     if method is not None:
--> 973         return method(include=include, exclude=exclude)
    974     return None
    975 else:

File E:\Anaconda\lib\site-packages\IPython\core\display.py:1022, in Image._re
pr_mimebundle_(self, include, exclude)
   1020 if self.embed:
   1021     mimetype = self._mimetype
-> 1022     data, metadata = self._data_and_metadata(always_both=True)
   1023     if metadata:
   1024         metadata = {mimetype: metadata}

File E:\Anaconda\lib\site-packages\IPython\core\display.py:1034, in Image._da
ta_and_metadata(self, always_both)
   1032     b64_data = b2a_base64(self.data).decode('ascii')
   1033 except TypeError as e:
-> 1034     raise FileNotFoundError(
   1035         "No such file or directory: '%s'" % (self.data)) from e
   1036 md = {}
   1037 if self.metadata:

FileNotFoundError: No such file or directory: 'Desktop/235365-(2)-نصائح-لمذاكرة-الدروس-قبل-الإمتحانات.jpg'
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
File E:\Anaconda\lib\site-packages\IPython\core\display.py:1032, in Image._da
ta_and_metadata(self, always_both)
   1031 try:
-> 1032     b64_data = b2a_base64(self.data).decode('ascii')
   1033 except TypeError as e:

TypeError: a bytes-like object is required, not 'str'

The above exception was the direct cause of the following exception:

FileNotFoundError                         Traceback (most recent call last)
File E:\Anaconda\lib\site-packages\IPython\core\formatters.py:343, in BaseFor
matter.__call__(self, obj)
    341     method = get_real_method(obj, self.print_method)
    342     if method is not None:
--> 343         return method()
    344     return None
    345 else:

File E:\Anaconda\lib\site-packages\IPython\core\display.py:1054, in Image._re
pr_png_(self)
   1052 def _repr_png_(self):
   1053     if self.embed and self.format == self._FMT_PNG:
-> 1054         return self._data_and_metadata()

File E:\Anaconda\lib\site-packages\IPython\core\display.py:1034, in Image._da
ta_and_metadata(self, always_both)
   1032     b64_data = b2a_base64(self.data).decode('ascii')
   1033 except TypeError as e:
-> 1034     raise FileNotFoundError(
   1035         "No such file or directory: '%s'" % (self.data)) from e
   1036 md = {}
   1037 if self.metadata:

FileNotFoundError: No such file or directory: 'Desktop/235365--نصائح-لمذاكرة-الدروس
(2)-قبل-الإمتحانات.jpg'
```

Out[1]: <IPython.core.display.Image object>

In [2]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
from sklearn.cluster import KMeans
```

In [3]:
```
df=pd.read_excel('Desktop/3 hotels1.xlsx')
```

In [4]:
```
df.head()
```

Out[4]:

	travelCode	userCode	name	place	days	price	total	data14	date1	data12	da
0	424	3	Hotel CB	Rio de Janeiro (RJ)	1	165.99	165.99	2023	7/13/2023	1900-01-07	
1	623	4	Hotel BD	Natal (RN)	4	242.88	971.52	2023	7/13/2023	1900-01-07	
2	79050	766	Hotel A	Florianopolis (SC)	3	313.02	939.06	2023	7/13/2023	1900-01-07	
3	109969	1089	Hotel AF	Sao Paulo (SP)	4	139.10	556.40	2023	7/13/2023	1900-01-07	
4	126658	1252	Hotel AF	Sao Paulo (SP)	2	139.10	278.20	2023	7/13/2023	1900-01-07	

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40552 entries, 0 to 40551
Data columns (total 11 columns):
 #   Column      Non-Null Count  Dtype         
---  ------      --------------  -----         
 0   travelCode  40552 non-null  int64         
 1   userCode    40552 non-null  int64         
 2   name        40552 non-null  object        
 3   place       40552 non-null  object        
 4   days        40552 non-null  int64         
 5   price       40552 non-null  float64       
 6   total       40552 non-null  float64       
 7   data14      40552 non-null  int64         
 8   date1       40552 non-null  object        
 9   data12      40552 non-null  datetime64[ns]
 10  data13      40552 non-null  int64         
dtypes: datetime64[ns](1), float64(2), int64(5), object(3)
memory usage: 2.9+ MB
```

In [6]:
```
df.describe()
```

Out[6]:

	travelCode	userCode	days	price	total	data14	
count	40552.000000	40552.000000	40552.000000	40552.000000	40552.000000	40552.000000	40
mean	67911.794461	666.963726	2.499679	214.439554	536.229513	2020.518248	
std	39408.199333	391.136794	1.119326	76.742305	319.331482	0.977547	
min	0.000000	0.000000	1.000000	60.390000	60.390000	2019.000000	
25%	33696.750000	323.000000	1.000000	165.990000	247.620000	2020.000000	
50%	67831.000000	658.000000	2.000000	242.880000	495.240000	2020.000000	
75%	102211.250000	1013.000000	4.000000	263.410000	742.860000	2021.000000	
max	135942.000000	1339.000000	4.000000	313.020000	1252.080000	2023.000000	

In [7]:
```
df['travelCode'].mean()
```

Out[7]:
```
67911.79446143223
```

In [8]:
```
df['travelCode'].median()
```

Out[8]:
```
67831.0
```

In [9]:
```
df['userCode'].var()
```

Out[9]:
```
152987.99190255022
```

In [10]:
```
df['userCode'].std()
```

Out[10]:
```
391.13679436042605
```

In [11]:
```
df['userCode'].mode()
```

Out[11]:
```
0    1104
Name: userCode, dtype: int64
```

In [12]:
```
IQR = df['travelCode'].quantile(0.75) - df['travelCode'].quantile(0.25)
IQR
```

Out[12]:
```
68514.5
```

In [30]:
```
print(df.columns)
```

```
Index(['travelCode', 'userCode', 'name', 'place', 'days', 'price', 'total',
       'data14', 'date1', 'data12', 'data13'],
      dtype='object')
```

Sorting

In [35]:
```
df.sort_values(by="price", ascending=False).head()
```

Out[35]:

	travelCode	userCode	name	place	days	price	total	data14	date1	data12
28581	21693	212	Hotel A	Florianopolis (SC)	3	313.02	939.06	2020	2020-12-11 00:00:00	1900 01-1
29238	29661	283	Hotel A	Florianopolis (SC)	1	313.02	313.02	2020	2020-11-06 00:00:00	1900 01-0
22877	50567	492	Hotel A	Florianopolis (SC)	1	313.02	313.02	2020	4/23/2020	1900 01-0
7181	57990	565	Hotel A	Florianopolis (SC)	2	313.02	626.04	2021	9/30/2021	1900 01-0
22867	44656	436	Hotel A	Florianopolis (SC)	1	313.02	313.02	2020	4/23/2020	1900 01-0

In [36]:
```
df.sort_values(by="place", ascending=False).head()
```

Out[36]:

	travelCode	userCode	name	place	days	price	total	data14	date1	data12	data13
13312	57204	557	Hotel AF	Sao Paulo (SP)	1	139.1	139.1	2021	1/21/2021	1900-01-01	21
9656	41024	396	Hotel AF	Sao Paulo (SP)	3	139.1	417.3	2021	5/13/2021	1900-01-05	13
21040	121959	1210	Hotel AF	Sao Paulo (SP)	2	139.1	278.2	2020	6/25/2020	1900-01-06	25
31947	105597	1047	Hotel AF	Sao Paulo (SP)	3	139.1	417.3	2020	2020-06-08 00:00:00	1900-01-08	6
21037	120564	1196	Hotel AF	Sao Paulo (SP)	3	139.1	417.3	2020	6/25/2020	1900-01-06	25

In [45]:
```
df.sort_values(by="name", ascending=False).head()
```

Out[45]:

	travelCode	userCode	name	place	days	price	total	data14	date1	data12	da
19452	110822	1098	Hotel Z	Aracaju (SE)	2	208.04	416.08	2020	8/20/2020	1900-01-08	
3547	71812	696	Hotel Z	Aracaju (SE)	2	208.04	416.08	2022	2/17/2022	1900-01-02	
32064	16537	164	Hotel Z	Aracaju (SE)	4	208.04	832.16	2020	2020-06-02 00:00:00	1900-01-02	
20720	117372	1163	Hotel Z	Aracaju (SE)	1	208.04	208.04	2020	7/16/2020	1900-01-07	
20718	117178	1161	Hotel Z	Aracaju (SE)	1	208.04	208.04	2020	7/16/2020	1900-01-07	

In [44]:
```
df.sort_values(by="data14", ascending=False).head()
```

Out[44]:

	travelCode	userCode	name	place	days	price	total	data14	date1	data12	dat
0	424	3	Hotel CB	Rio de Janeiro (RJ)	1	165.99	165.99	2023	7/13/2023	1900-01-07	
633	26285	251	Hotel K	Salvador (BH)	2	263.41	526.82	2023	2023-09-02 00:00:00	1900-01-02	
621	2469	21	Hotel BD	Natal (RN)	2	242.88	485.76	2023	2023-09-02 00:00:00	1900-01-02	
622	8480	82	Hotel AF	Sao Paulo (SP)	3	139.10	417.30	2023	2023-09-02 00:00:00	1900-01-02	
623	9006	90	Hotel Z	Aracaju (SE)	2	208.04	416.08	2023	2023-09-02 00:00:00	1900-01-02	

In [40]:
```
df.sort_values(by="days", ascending=False).head()
```

Out[40]:

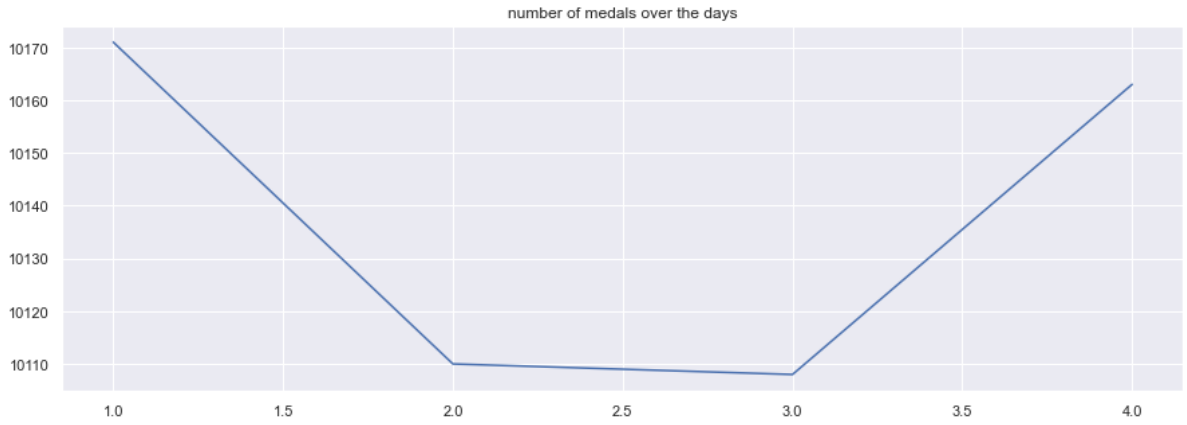
	travelCode	userCode	name	place	days	price	total	data14	date1	data1
40551	135894	1338	Hotel BP	Brasilia (DF)	4	247.62	990.48	2019	2019-03-10 00:00:00	190 01-'
31888	76065	741	Hotel AU	Recife (PE)	4	312.83	1251.32	2020	2020-06-08 00:00:00	190 01-(
12923	3726	38	Hotel A	Florianopolis (SC)	4	313.02	1252.08	2021	1/28/2021	190 01-(
12924	3829	39	Hotel A	Florianopolis (SC)	4	313.02	1252.08	2021	1/28/2021	190 01-(
12928	5422	54	Hotel BD	Natal (RN)	4	242.88	971.52	2021	1/28/2021	190 01-(

In [46]:
```
df[(df["price"] == 0) & (df["name"] == "No")]["total"].max()
```

Out[46]: nan

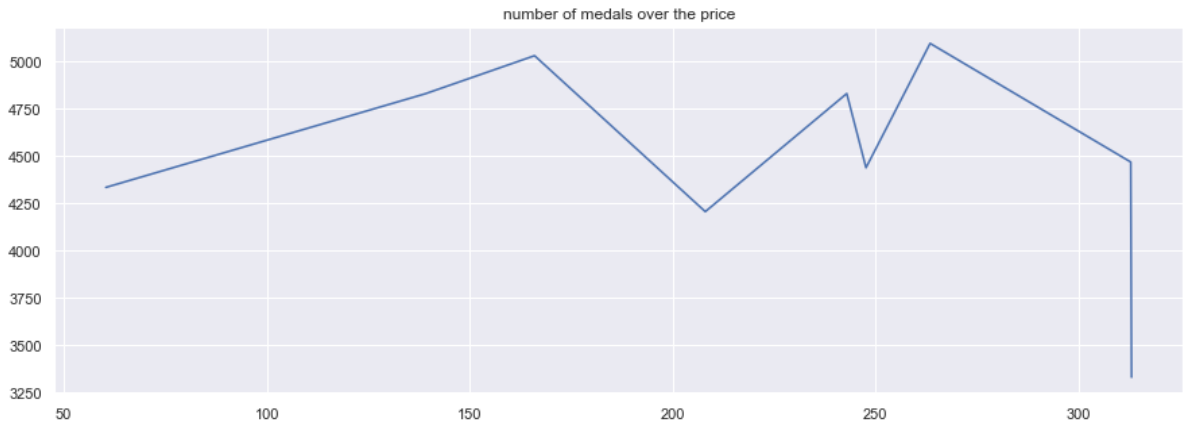
In [13]:
```
c
```

Out[13]: <AxesSubplot:title={'center':'number of medals over the days'}>

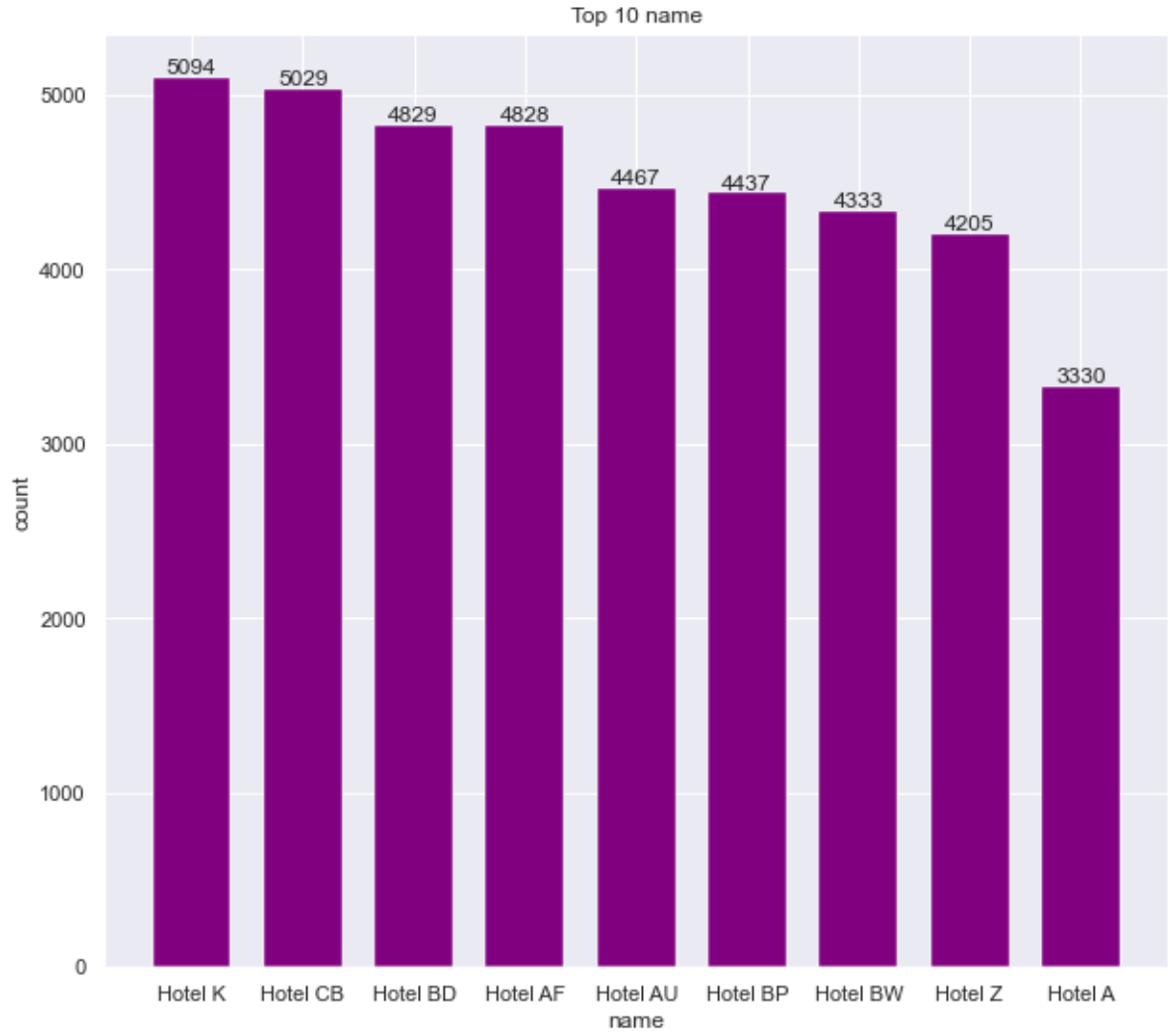


In [14]:
```
plt.figure(figsize=(15,5))
plt.title('number of medals over the price')
df.price.value_counts().sort_index().plot()
```

Out[14]: <AxesSubplot:title={'center':'number of medals over the price'}>



In [15]:
```
itemNames = df['name'].value_counts().index[:10]
itemValues = df['name'].value_counts().values[:10]
plt.figure(figsize=(10,9))
plt.ylabel('count', fontsize='medium')
plt.xlabel('name', fontsize='medium')
plt.title('Top 10 name')
plt.bar(itemNames,itemValues, width = 0.7,color='purple',linewidth=0.4)
for i in range(len(itemNames)):
    plt.text(i,itemValues[i],itemValues[i],ha='center',va='bottom')
plt.show()
```



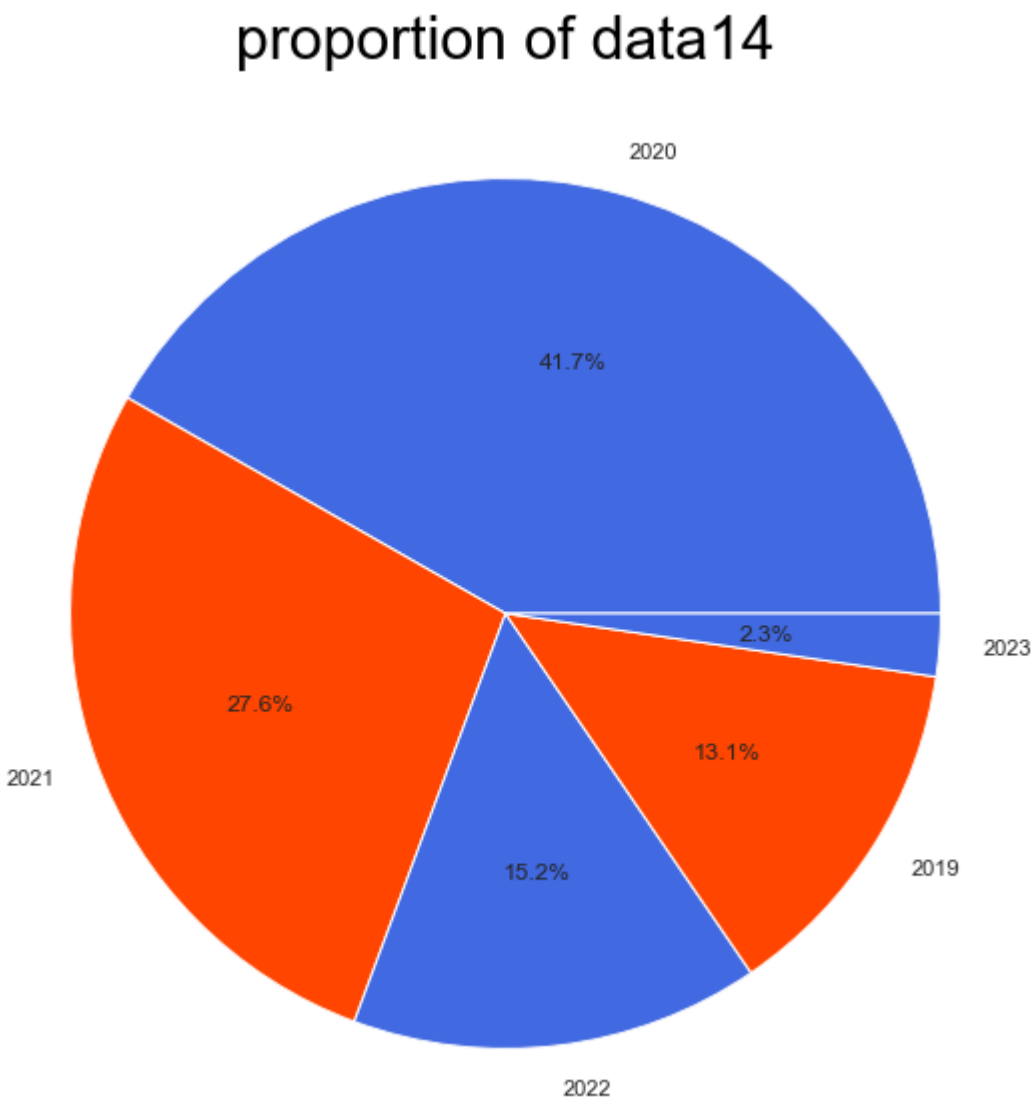
In [1]:
```
itemNames = df['place'].value_counts().index[:9]
itemValues = df['place'].value_counts().values[:9]
plt.figure(figsize=(15,9))
plt.ylabel('count', fontsize='medium')
plt.xlabel('place', fontsize='medium')
plt.title('Top 10 place')
plt.bar(itemNames,itemValues, width = 0.7,color='purple',linewidth=0.4)
for i in range(len(itemNames)):
    plt.text(i,itemValues[i],itemValues[i],ha='center',va='bottom')
plt.show()
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Input In [1], in <cell line: 1>()
----> 1 itemNames = df['place'].value_counts().index[:9]
      2 itemValues = df['place'].value_counts().values[:9]
      3 plt.figure(figsize=(15,9))

NameError: name 'df' is not defined
```

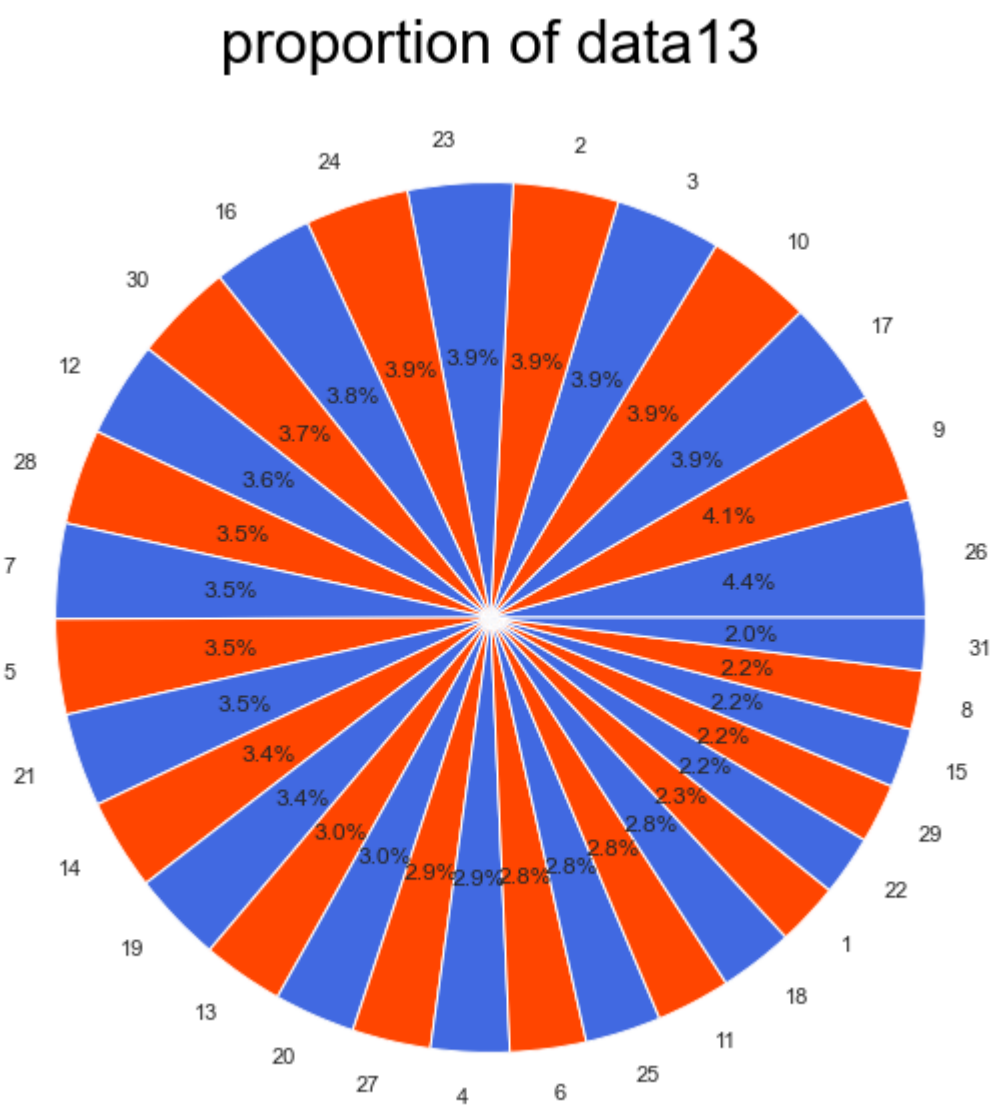
In [17]:
```
labels = df.data14.value_counts().index
colors = ['royalblue','orangered']
data14 = df.data14.value_counts().values
plt.figure(figsize = (10,10))
plt.pie(data14, labels=labels, colors=colors, autopct='%1.1f%%')
plt.title('proportion of data14',color = 'black',fontsize = 30)
```

Out[17]: Text(0.5, 1.0, 'proportion of data14')



In [18]:
```
labels = df.data13.value_counts().index
colors = ['royalblue','orangered']
data13 = df.data13.value_counts().values
plt.figure(figsize = (10,10))
plt.pie(data13, labels=labels, colors=colors, autopct='%1.1f%%')
plt.title('proportion of data13',color = 'black',fontsize = 30)
```

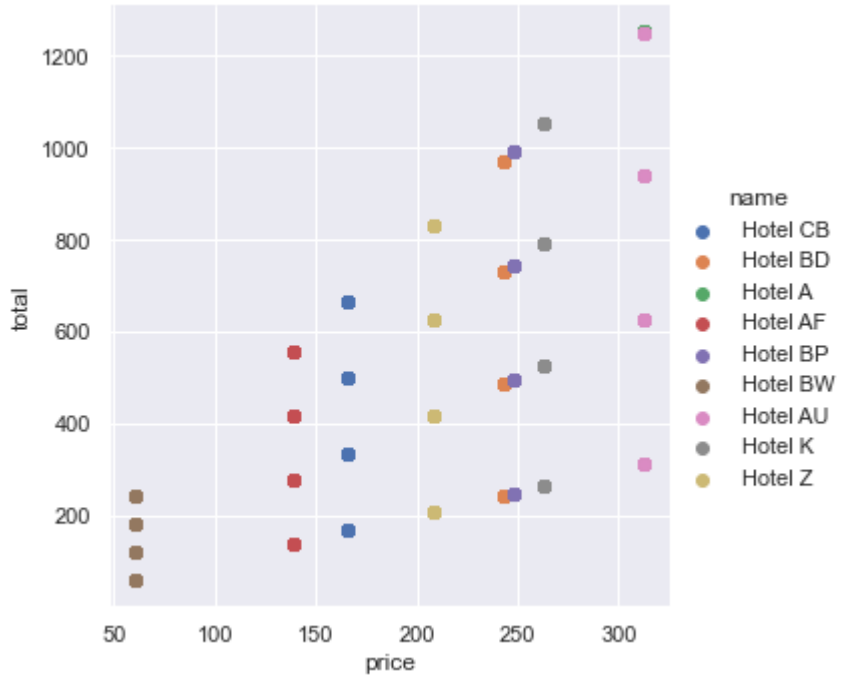
Out[18]: Text(0.5, 1.0, 'proportion of data13')



In [21]:
```
sns.FacetGrid(df, hue="name", size=5) \
   .map(plt.scatter, "price", "total") \
   .add_legend()
```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\axisgrid.py:337: UserWarni
ng: The `size` parameter has been renamed to `height`; please update your cod
e.
  warnings.warn(msg, UserWarning)
```

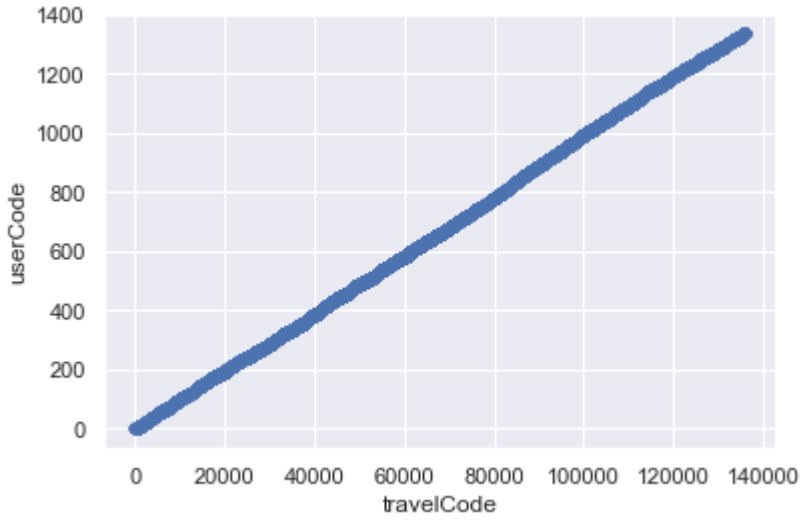
Out[21]: <seaborn.axisgrid.FacetGrid at 0xc2245c8>



In [23]:
```
df.plot(kind="scatter", x="travelCode", y="userCode")
```

```
*c* argument looks like a single numeric RGB or RGBA sequence, which should b
e avoided as value-mapping will have precedence in case its length matches wi
th *x* & *y*.  Please use the *color* keyword-argument or provide a 2D array
with a single row if you intend to specify the same RGB or RGBA value for all
points.
```

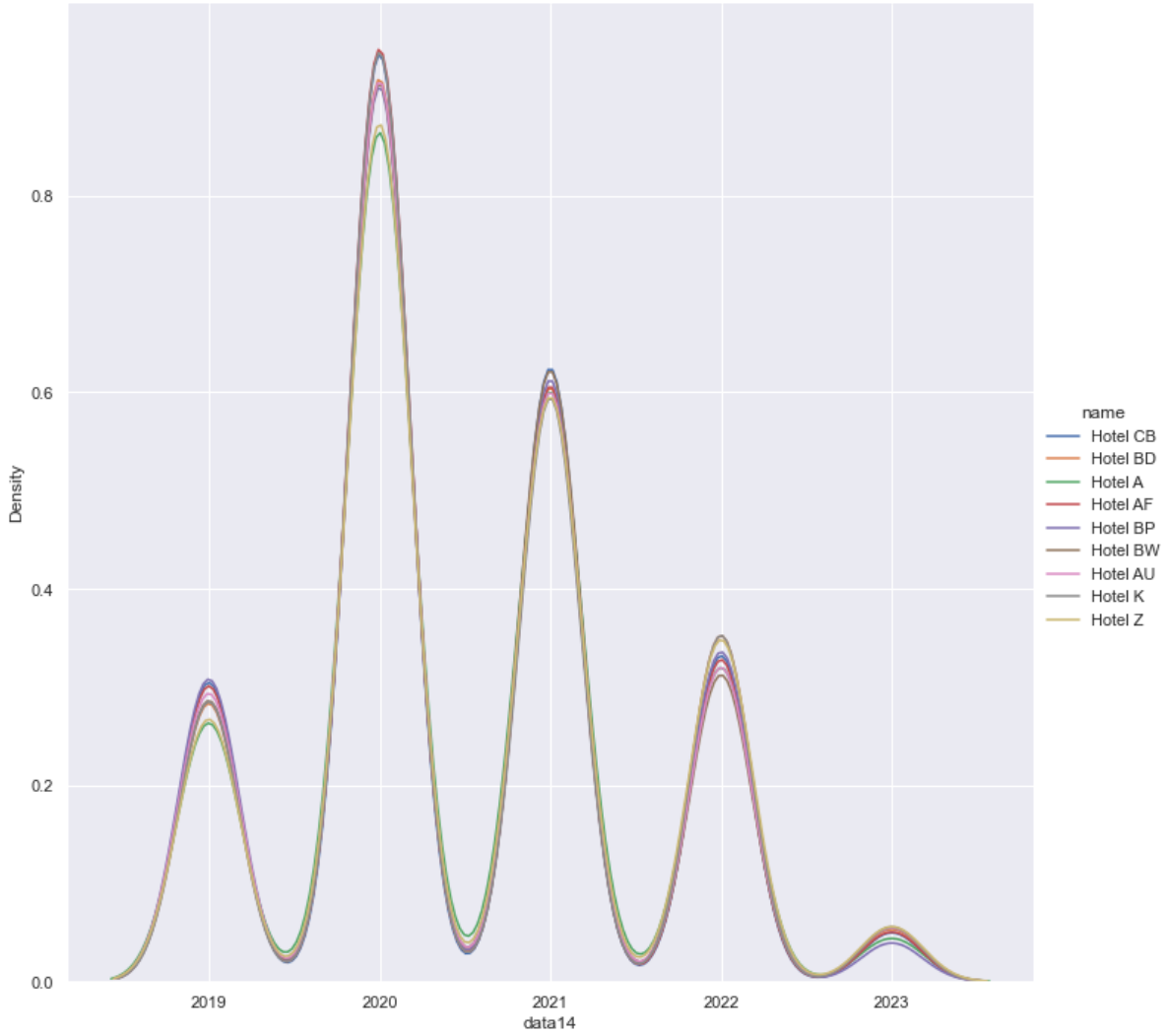
Out[23]:
```
<AxesSubplot:xlabel='travelCode', ylabel='userCode'>
```



In [26]:
```
sns.FacetGrid(df, hue="name", size=10) \
   .map(sns.kdeplot, "data14") \
   .add_legend()
```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\axisgrid.py:337: UserWarni
ng: The `size` parameter has been renamed to `height`; please update your cod
e.
  warnings.warn(msg, UserWarning)
```

Out[26]: <seaborn.axisgrid.FacetGrid at 0xa87e178>



Thank You !!