

```
# set up

In [732]:
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from pyecharts import options as opts
from pyecharts.charts import Page, File, Geo, Bar

In [734]:
# load data

In [735]:
data = pd.read_excel('travel.xlsx')

In [736]:
# examine data

In [774]:
data.head()

Out[774]:
   地点  短评  出发时间  天数  人均费用  人物  玩法  浏览量
0  西安  一场只属于我们两父子的旅行——35天9省区畅游祖国壮山河之陕西-宁夏篇  /2020/08/12  6  2400  亲子  第一次?穷游?美食?暑假  300.0
1  开封  开封第1次——11开封府  /2021/01/31  1  400  家庭  深度游?徒步?美食?冬季  16000.0
2  三亚  三亚小众玩法春天刚来，我便开始想念海南边的漫时光  /2021/01/31  4  2600  三五好友  自驾?美食?摄影  2729.0
3  拉萨  延吉十八日啊，独食“美食家”的不提独厨步  /2020/12/25  2  500  独自一人  短途周末?美食?摄影?冬季  9521.0
4  漳州  六天五夜，探秘闽南人间烟火秘境  /2021/02/19  6  1200  三五好友  自驾?春节?踏青  4901.0

In [738]:
data.shape

Out[738]:
(1601, 8)

In [739]:
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1601 entries, 0 to 1600
Data columns (total 8 columns):
#  Column  Non-Null Count  Dtype
---  --
0  地点  1601 non-null  object
1  短评  1601 non-null  object
2  出发时间  1597 non-null  object
3  天数  1601 non-null  int64
4  人均费用  1601 non-null  int64
5  人物  1601 non-null  object
6  玩法  1601 non-null  object
7  浏览量  1601 non-null  float64
dtypes: float64(1), int64(2), object(5)
memory usage: 100.2+ KB

In [740]:
# summary statistics

In [741]:
data.describe()

Out[741]:
   天数  人均费用  浏览量
count  1601.000000  1601.000000  1.601000e+03
mean    4.059863    1507.38663  2.867840e+05
std     4.862243    727.216306  5.337490e+06
min     1.000000    1.000000    3.300000e+00
25%     3.000000    1000.000000  5.416000e+03
50%     3.000000    1500.000000  1.200000e+04
75%     5.000000    2000.000000  3.200000e+04
max    99.000000   2999.000000  1.800000e+08

In [742]:
# data visualization

In [743]:
topdays_15 = data['天数'].value_counts()[1:15]
topdays_15.plot(kind='bar', figsize=(14,10), color='steelblue')
plt.style.use('ggplot')
plt.xlabel('Days')
plt.ylabel('Count')
plt.title(label="Top 15 Traveling Periods of the data",
          fontsize=30,
          color='steelblue')
# 大部分旅行天数在一个星期以下，最多频率的是5天以下，10天以上属于罕见

Out[743]:
Text(0.5, 1.0, 'Top 15 Traveling Periods of the data')

Top 15 Traveling Periods of the data


In [744]:
data['地点'].value_counts()[1:15].index.tolist()

Out[744]:
['成都',
 '重庆',
 '厦门',
 '三亚',
 '西安',
 '敦煌',
 '平遥',
 '北京',
 '南京',
 '桂林',
 '阳朔',
 '日照',
 '长沙',
 '上海']

In [745]:
data['地点'].value_counts()[1:15].tolist()

Out[745]:
[113, 60, 52, 49, 38, 37, 34, 34, 34, 30, 24, 21, 20, 19, 19]

In [746]:
bar = Bar()
bar.add_xaxis(['成都',
 '重庆',
 '厦门',
 '三亚',
 '西安',
 '敦煌',
 '平遥',
 '北京',
 '南京',
 '桂林',
 '阳朔',
 '日照',
 '长沙',
 '上海'])
bar.add_yaxis("", (113, 60, 52, 49, 38, 37, 34, 34, 34, 30, 24, 21, 20, 19, 19), color='steelblue')
bar.set_global_opts(title_opts=opts.TitleOpts(title="前15名热度城市旅游次数"))
bar.render_notebook()

Out[746]:
前15名热度城市旅游次数


In [747]:
topfees = data['人均费用'],
topfees.plot.hist(grid=True, bins=20, rwidth=0.9,
                  color='#60C8E8',figsize=(14,10))
plt.style.use('ggplot')
plt.grid(axis='y', alpha=0.75)
plt.xlabel('Fees')
plt.ylabel('Count')
plt.title(label="Histogram of Traveling Fees",
          fontsize=30,
          color='steelblue')

Out[747]:
Text(0.5, 1.0, 'Histogram of Traveling Fees')

Histogram of Traveling Fees


In [776]:
value = ['成都',
 '重庆',
 '厦门',
 '三亚',
 '西安',
 '敦煌',
 '平遥',
 '杭州',
 '北京',
 '南京',
 '桂林',
 '阳朔',
 '日照',
 '长沙',
 '上海']
data_new = data[data['地点'].isin(value)]
data_new.head()

Out[776]:
   地点  短评  出发时间  天数  人均费用  人物  玩法  浏览量
0  西安  一场只属于我们两父子的旅行——35天9省区畅游祖国壮山河之陕西-宁夏篇  /2020/08/12  6  2400  亲子  第一次?穷游?美食?暑假  300.0
2  三亚  三亚小众玩法春天刚来，我便开始想念海南边的漫时光  /2021/01/31  4  2600  三五好友  自驾?美食?摄影  2729.0
5  三亚  【三亚旅游】短妹一起超High! 四天三夜海边狂欢，住宿美食、景点交通全攻略  /2020/12/22  4  2000  闺蜜  美食?夏季?寒假  10000.0
6  三亚  在北纬18°的“东方夏威夷”三亚，赴一场微风、海韵与美食的约会  /2021/02/05  4  1000  三五好友  短途周末?自驾?美食?冬季  9148.0
7  成都  一场只属于我们两父子的旅行——35天9省区畅游祖国壮山河之四川篇  /2020/07/28  6  2100  亲子  第一次?穷游?美食?暑假  379.0

In [749]:
type(data_new)

Out[749]:
pandas.core.frame.DataFrame

In [750]:
topbrowsing = data['浏览时间'],
topbrowsing.plot.hist(grid=True, bins = 20, rwidth=0.9, color='#0504aa', alpha=0.7, figsize=(14,10))
plt.grid(axis='y', alpha=0.75)
plt.style.use('ggplot')
plt.grid(axis='y', alpha=0.75)
plt.xlabel('Browsing Times')
plt.ylabel('Count')
plt.title(label="Histogram of Browsing Times",
          fontsize=30,
          color='#0504aa')

Out[750]:
Text(0.5, 1.0, 'Histogram of Browsing Times')

Histogram of Browsing Times


In [751]:
data_new['人均费用'].groupby(data_new['地点']).mean().sort_values(ascending = False)

Out[751]:
地点
三亚    2041.081633
西安    1977.236842
厦门    1892.076923
阳朔    1790.476190
桂林    1708.958333
上海    1557.894737
重庆    1544.416667
成都    1542.796460
敦煌    1532.432432
长沙    1498.631719
杭州    1445.000000
北京    1417.852941
南京    1294.100000
平遥    1155.882353
日照    1035.000000
Name: 人均费用, dtype: float64

In [752]:
data_new['人均费用'].groupby(data_new['地点']).mean().sort_values(ascending = False).index.tolist()

Out[752]:
['三亚',
 '西安',
 '厦门',
 '阳朔',
 '桂林',
 '重庆',
 '上海',
 '成都',
 '敦煌',
 '长沙',
 '杭州',
 '北京',
 '南京',
 '平遥',
 '日照']

In [753]:
data_new['人均费用'].groupby(data_new['地点']).mean().sort_values(ascending = False).tolist()

Out[753]:
[2041.0816326530612,
 1977.2368421052631,
 1892.076923076923,
 1790.4761904761904,
 1708.9583333333333,
 1557.8947368421052,
 1544.4166666666667,
 1542.796460176991,
 1532.4324324324325,
 1498.6315789473683,
 1445.0,
 1417.8529411764705,
 1294.7,
 1155.88235411766,
 1035.0]

In [754]:
bar = Bar()
bar.add_xaxis(data_new['人均费用'].groupby(data_new['地点']).mean().sort_values(ascending = False).index.tolist())
bar.add_yaxis("", data_new['人均费用'].groupby(data_new['地点']).mean().sort_values(ascending = False).tolist(), color='#682288')
bar.set_global_opts(title_opts=opts.TitleOpts(title="前15名热度城市平均消费"))
bar.set_series_opts(label_opts=opts.LabelOpts(is_show=False))
bar.render_notebook()

Out[754]:
前15名热度城市平均消费


In [755]:
topbrowsing_selected = data_selected['浏览量']
sns.boxplot(y=topbrowsing_selected)

Out[755]:
<AxesSubplot:ylabel='浏览量'>


In [756]:
data_selected = data.loc[data['浏览量'] < 20000]
sns.boxplot(y=topbrowsing_selected)

Out[756]:
Text(0, 0.5, 'Count')


In [757]:
topbrowsing_selected.plot.hist(grid=True, bins = 20, rwidth=0.9, color='#682288', alpha=0.7, figsize=(14,10))
plt.grid(axis='y', alpha=0.75)
plt.style.use('ggplot')
plt.grid(axis='y', alpha=0.75)
plt.xlabel('Browsing Times')
plt.ylabel('Count')
plt.title(label="Histogram of Selected Browsing Times",
          fontsize=30,
          color='#682288')

Out[757]:
Text(0.5, 1.0, 'Histogram of Selected Browsing Times')

Histogram of Selected Browsing Times


In [758]:
data_new['浏览量'].groupby(data_new['地点']).mean().sort_values(ascending = False)

Out[758]:
地点
西安    1.248666e+06
成都    1.107754e+06
长沙    1.805917e+05
杭州    4.620623e+04
三亚    3.810435e+04
重庆    3.530627e+04
平遥    3.626461e+04
杭州    2.889109e+04
日照    2.815070e+04
桂林    2.814083e+04
敦煌    2.762359e+04
北京    2.344635e+04
南京    2.143987e+04
阳朔    1.950435e+04
上海    1.515084e+04
Name: 浏览量, dtype: float64

In [759]:
data_new['浏览量'].groupby(data_new['地点']).mean().sort_values(ascending = False).index.tolist()

Out[759]:
['西安',
 '成都',
 '长沙',
 '厦门',
 '三亚',
 '重庆',
 '平遥',
 '杭州',
 '日照',
 '桂林',
 '敦煌',
 '北京',
 '南京',
 '阳朔',
 '上海']

In [760]:
bar = Bar()
bar.add_xaxis(data_new['浏览量'].groupby(data_new['地点']).mean().sort_values(ascending = False).index.tolist())
bar.add_yaxis("", data_new['浏览量'].groupby(data_new['地点']).mean().sort_values(ascending = False).tolist(), color='#682288')
bar.set_global_opts(title_opts=opts.TitleOpts(title="前15名热度城市浏览量"))
bar.set_series_opts(label_opts=opts.LabelOpts(is_show=False))
bar.render_notebook()

Out[760]:
前15名热度城市浏览量


In [761]:
# pie chart

In [762]:
pd.set_option("display.max_rows", None)
data['地点'].value_counts()
```


[illegible]

[illegible]

克什克腾旗	2
金川	2
句容	2
泰山	2
祁连	2
海宁	2
六安	2
文昌	2
丽水	2
通化	2
金坛	2
会东	1
丰都	1

