本文将从以下几点进行分析：

- 背景
- 提出问题
- 数据清洗
- 数据可视化分析
- 结论和建议

背景

本文是从葡萄牙的两家酒店——度假酒店和城市酒店的运营数据入手，对酒店的运营情况进行分析，最后给出建议。

提出问题

- 分析酒店订单变化的原因
- 分析顾客取消订单的原因
- 分析不同代理商订单

数据清洗

加载工具包、读取数据

In [70]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from datetime import datetime

hotel = pd.read_csv('/Users/jipeng/Downloads/hotel_bookings.csv')
```

查看数据集大小并观察前5行数据

In [71]:

```
#print(hotel.shape)
#结果为(119390, 32)

# 设置查看列不省略
pd.set_option('display.max_columns',None)
hotel.head()
```

Out[71]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_num
0	Resort Hotel	0	342	2015	July	
1	Resort Hotel	0	737	2015	July	
2	Resort Hotel	0	7	2015	July	
3	Resort Hotel	0	13	2015	July	
4	Resort Hotel	0	14	2015	July	

观察可发现某些列，如agent、company等，含有空值，但是这些列的空值含有特殊意义，不应该删除。

对数据进行描述性统计

In [72]:

```
hotel.describe()
```

Out[72]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date
count	119390.000000	119390.000000	119390.000000	119390.000000	
mean	0.370416	104.011416	2016.156554	27.165173	
std	0.482918	106.863097	0.707476	13.605138	
min	0.000000	0.000000	2015.000000	1.000000	
25%	0.000000	18.000000	2016.000000	16.000000	
50%	0.000000	69.000000	2016.000000	28.000000	
75%	1.000000	160.000000	2017.000000	38.000000	
max	1.000000	737.000000	2017.000000	53.000000	

对数据进行画图观察其分布。

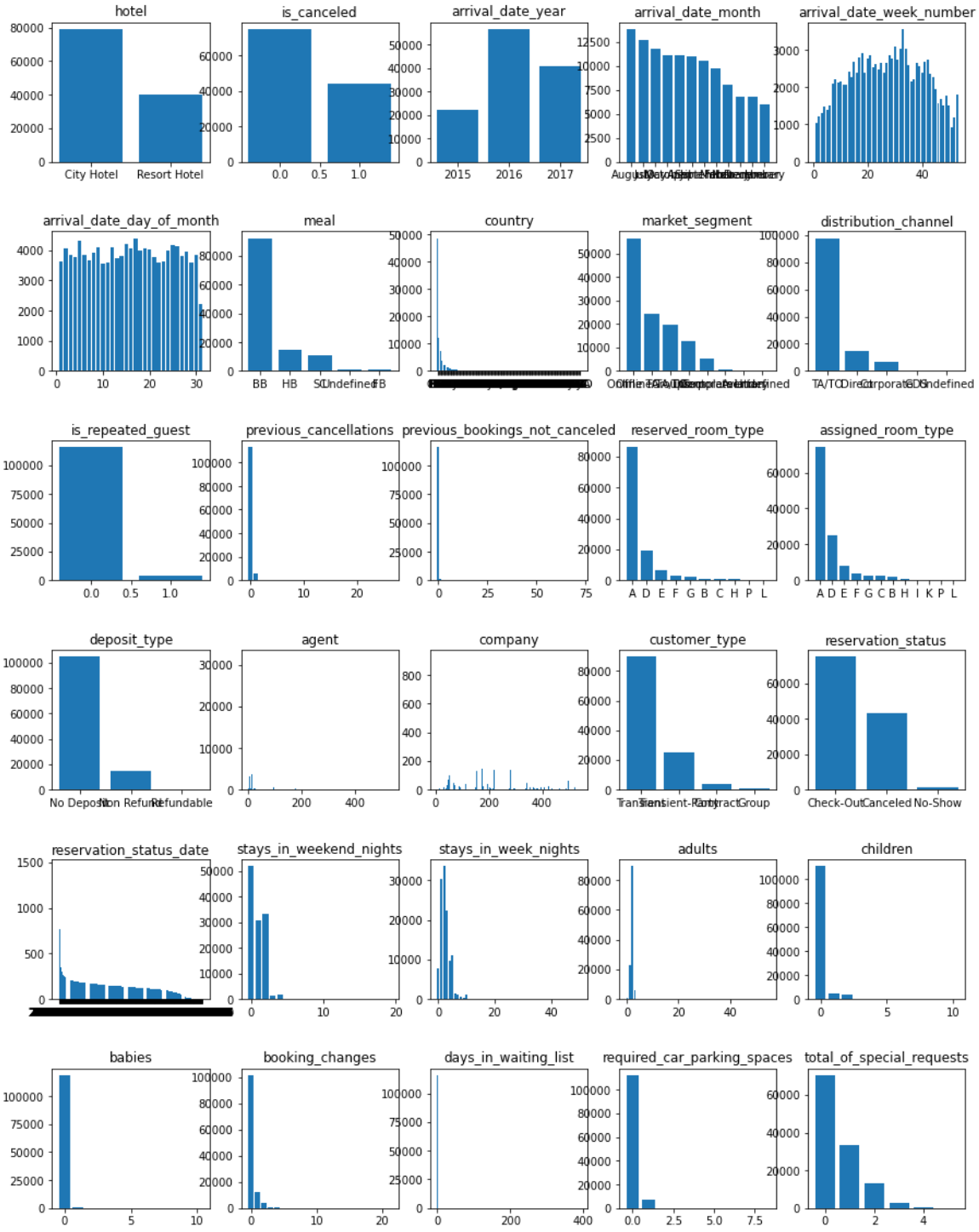
In [73]:

```
# 离散变量
bar = ["hotel", "is_canceled", "arrival_date_year", "arrival_date_month", "arriv
al_date_week_number",
       "arrival_date_day_of_month", "meal", "country", "market_segment", "distri
bution_channel",
       "is_repeated_guest", "previous_cancellations", "previous_bookings_not_can
celed",
       "reserved_room_type", "assigned_room_type", "deposit_type", "agent", "com
pany", "customer_type",
       "reservation_status", "reservation_status_date", "stays_in_weekend_night
s", "stays_in_week_nights",
       "adults", "children", "babies", "booking_changes", "days_in_waiting_list"
, "required_car_parking_spaces",
       "total_of_special_requests"]

# 连续变量画密度图
not_bar = ["lead_time", "adr"]
```

对离散或分类变量计数并画柱状图

In [74]:

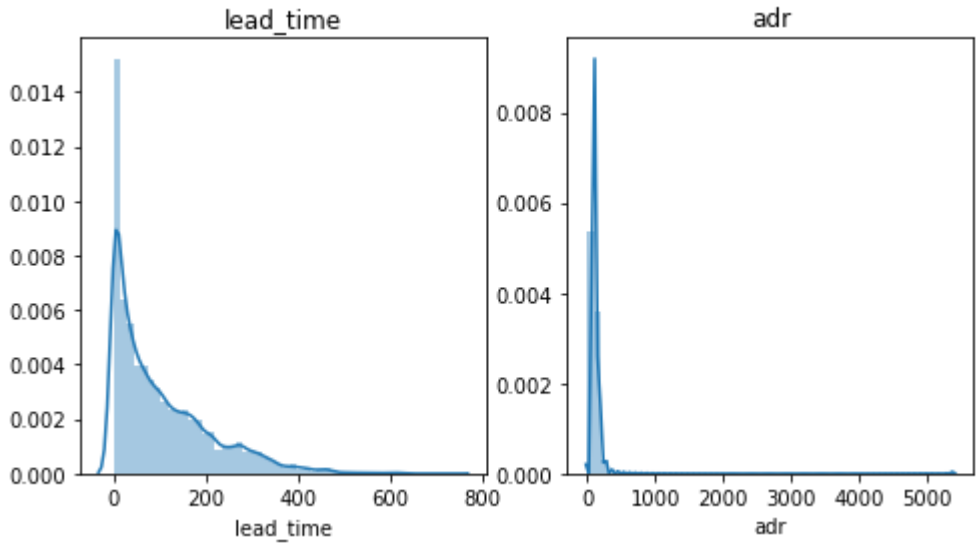
```
fig = plt.figure()
fig.set_size_inches(15, 20)
plt.subplots_adjust(left=None, bottom=None, right=None, top=None, wspace=None, h
space=0.5)
for i in range(len(bar)):
    ax = fig.add_subplot(6, 5, i+1)

    data_col = hotel.loc[:, bar[i]].value_counts()
    plt.bar(data_col.index, data_col.values)
    ax.set(#xlabel='arrival_date_week_number',
           title = bar[i])
plt.show()
```


hotel
City Hotel
Resort Hotel
is_canceled
arrival_date_year
2015
2016
2017
arrival_date_month
arrival_date_week_number
arrival_date_day_of_month
meal
BB
HB
country
market_segment
distribution_channel
TA/TO
Direct
Corporate
GDS
Undefined
is_repeated_guest
previous_cancellations
previous_bookings_not_canceled
reserved_room_type
assigned_room_type
deposit_type
No Deposit
Non Refund
Refundable
agent
company
customer_type
Transient
Transient-Party
Contract
Group
reservation_status
Check-Out
Canceled
No-Show
reservation_status_date
stays_in_weekend_nights
stays_in_week_nights
adults
children
babies
booking_changes
days_in_waiting_list
required_car_parking_spaces
total_of_special_requests

对连续变量画密度图

In [75]:

```
fig = plt.figure()
fig.set_size_inches(8, 4)
plt.subplots_adjust(left=None, bottom=None, right=None, top=None, wspace=None, h
space=1)

for i in range(len(not_bar)):
    ax = fig.add_subplot(1, 2, i+1)
    sns.distplot(hotel.loc[:, not_bar[i]])
    ax.set(title = not_bar[i])

plt.show()
```



从图和描述性统计中我们可以发现以下几种异常值：

- adr这一列存在一个样本的adr值是其他adr值的十倍以上，因此可以删除此样本。
- adults、children、babies三列之和为0。

In [76]:

```
hotel.drop(axis=0, index=48515, inplace=True)
hotel.drop(hotel.index[list(
    hotel["adults"] + hotel["children"] + hotel["babies"] == 0)], inplace=True)
hotel["agent"] = hotel["agent"].astype(str)
```

数据可视化分析

将年份和月份组成日期格式。

In [4]:

```
hotel["date"] = (hotel["arrival_date_year"].apply(lambda x: str(x))+ " "
 + hotel["arrival_date_month"].apply(lambda x: x[0:3])).apply(lambda x: datetime
.strptime(x, '%Y %b'))
```

In [13]:

```
hotel.head()
```

Out[13]:

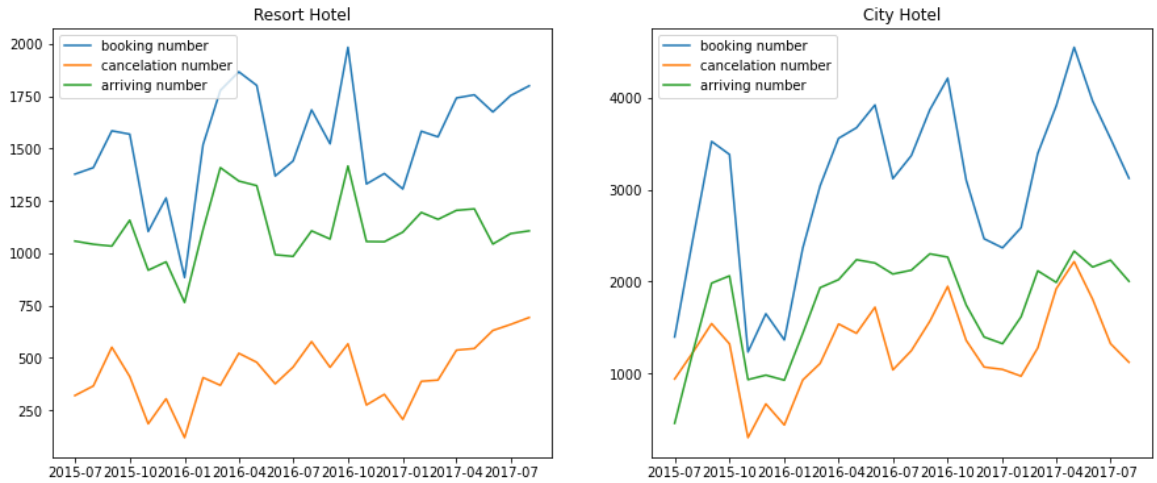
	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_num
0	Resort Hotel	0	342	2015	July	
1	Resort Hotel	0	737	2015	July	
2	Resort Hotel	0	7	2015	July	
3	Resort Hotel	0	13	2015	July	
4	Resort Hotel	0	14	2015	July	

酒店运营分析

入住数

In [14]:

```
is_canceled = hotel.groupby(["hotel", "date"])["is_canceled"]
fig = plt.figure(figsize=(15, 6))

ax = fig.add_subplot(1, 2, 1)
ax.set(title = "Resort Hotel")
sns.lineplot(data=is_canceled.count()["Resort Hotel"])
sns.lineplot(data=is_canceled.sum()["Resort Hotel"])
sns.lineplot(data=(is_canceled.count() - is_canceled.sum())["Resort Hotel"])
plt.legend(labels=['booking number','cancelation number','arriving number'])

ax = fig.add_subplot(1, 2, 2)
ax.set(title = "City Hotel")
sns.lineplot(data=is_canceled.count()["City Hotel"])
sns.lineplot(data=is_canceled.sum()["City Hotel"])
sns.lineplot(data=(is_canceled.count() - is_canceled.sum())["City Hotel"])
plt.legend(labels=['booking number','cancelation number','arriving number'])

plt.show()
```

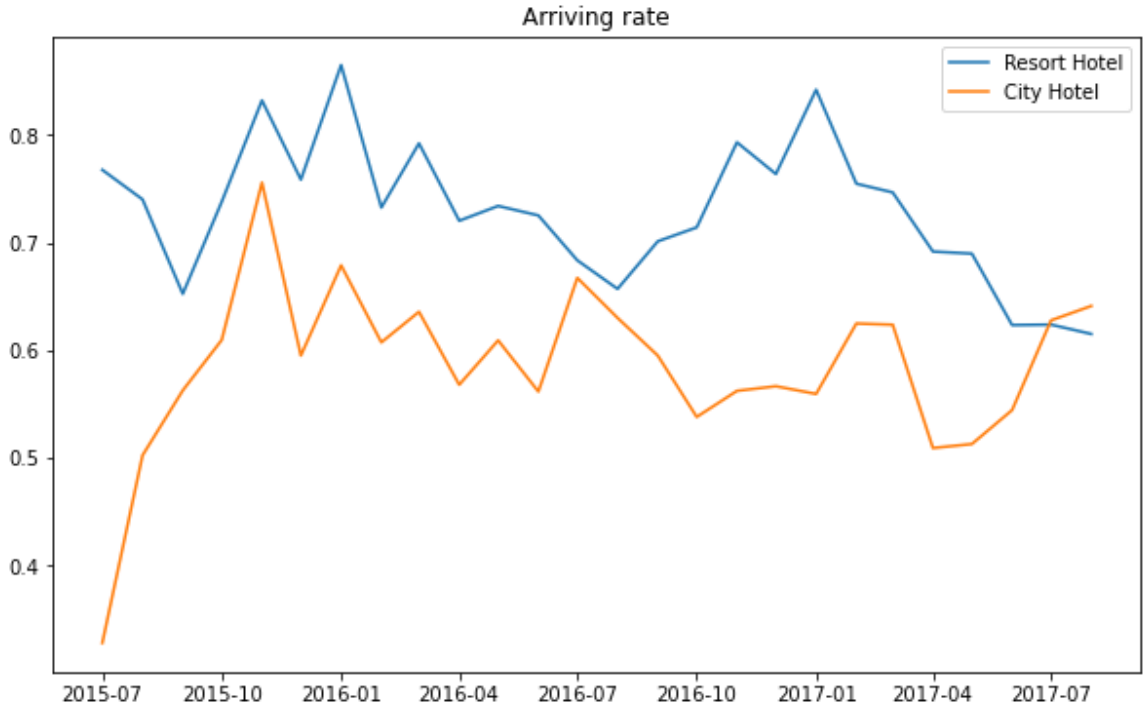


由图可知，对于度假酒店而言，每年的春秋两季入住人数较多；而对于城市酒店而言，4-9月入住人数一直较多。对于这两个酒店而言，冬季的入住人数都比较低。

入住率

In [15]:

```
fig = plt.figure(figsize=(10, 6))
sns.lineplot(data = ((is_canceled.count() - is_canceled.sum()) / is_canceled.cou
nt())["Resort Hotel"])
sns.lineplot(data = ((is_canceled.count() - is_canceled.sum()) / is_canceled.cou
nt())["City Hotel"])
plt.legend(labels=['Resort Hotel', 'City Hotel'])
plt.title('Arriving rate')
plt.show()
```



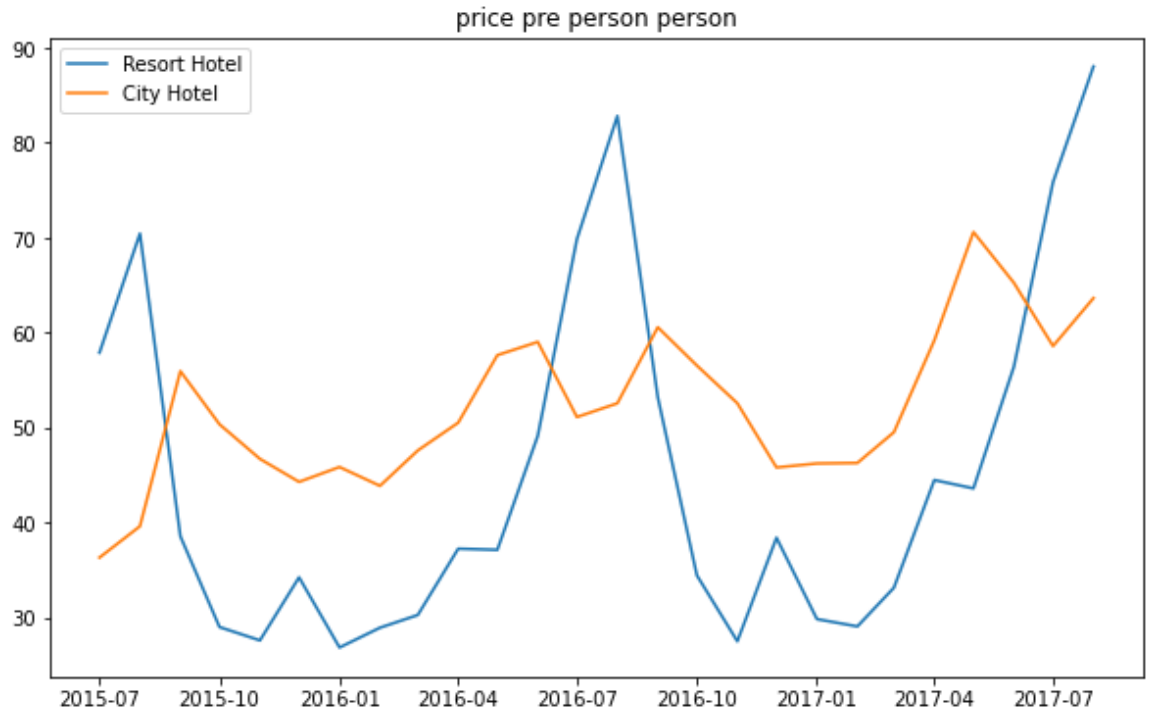
从图中可知，度假酒店的入住率平均会比城市酒店的入住率高10%左右。无论是度假酒店还是城市酒店，入住率都在冬季较高。 另外，该城市酒店在2015年7月的入住率极低（估计在30%左右），但是此数据的时间处于所给数据的开头，考虑有可能是数据不全， 需要进一步根据其他数据确定是否是这一原因。如果这一现象真实存在，需要考虑该酒店的运营管理等是否在这一段时间内有重大不足。

每晚每人价格

一般不把婴儿计入入住人数，因此可以根据成人和儿童数计算每人每晚价格。

In [16]:

```
info_hotel_date = hotel.groupby(["hotel", "date"])
adr_pp = info_hotel_date["adr"].sum() / (info_hotel_date["adults"].sum() + info_
hotel_date["children"].sum())

fig = plt.figure(figsize=(10, 6))
sns.lineplot(data = adr_pp["Resort Hotel"])
sns.lineplot(data = adr_pp["City Hotel"])
plt.legend(labels=['Resort Hotel', 'City Hotel'])
plt.title('price pre person person')
plt.show()
```



从图中可以看出，度假酒店的平均每人每晚价格在每年八月达到最高值，而在每年十二月也会有一个小的涨幅。城市酒店的平均每人每晚价格在每年的春季和秋季达到最高。无论是度假酒店还是城市酒店，平均每人每晚价格都有大而缓慢的上涨趋势。

酒店运营分析小结

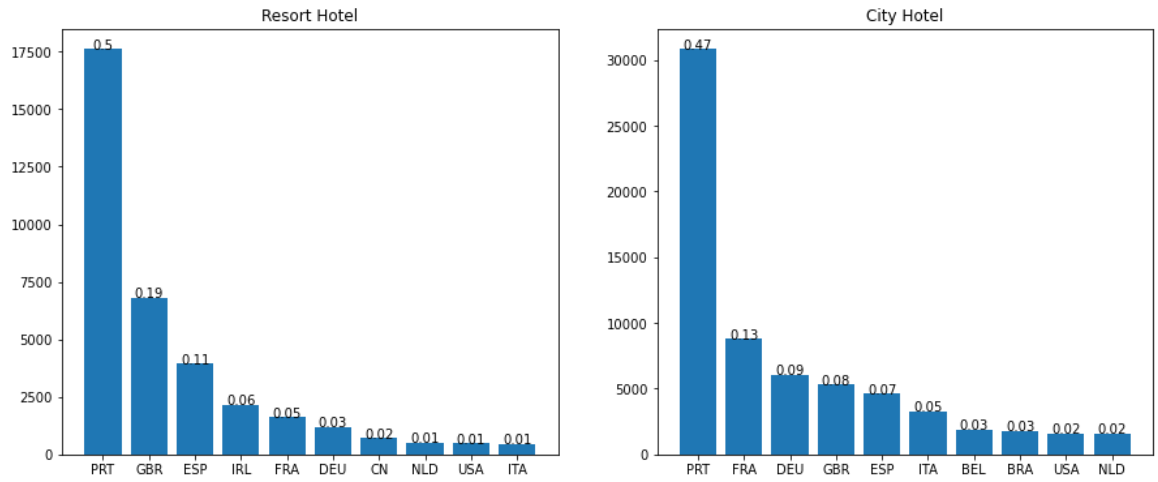
度假酒店在夏季的入住数、入住率较低，但是夏季的平均价格却较平时高很多，也许价格是导致入住情况较差的原因，可以考虑在夏季适当调低价格以谋求更好的入住情况。而城市酒店的入住情况在春秋季最好，夏季稍差，冬季最不好，其价格和入住情况成正比，春秋季最高，夏季稍低，冬季最低。因此，城市酒店的总体运营情况较为合理

用户画像分析

地理位置

In [11]:

```
N = 10

area = hotel.groupby(["hotel", "country"]).count()["is_canceled"]
area_resort = area['Resort Hotel'].sort_values(ascending = False).iloc[: N]
sum_resort = sum(area_resort)
max_resort = max(area_resort)
area_city = area['City Hotel'].sort_values(ascending = False).iloc[: N]
sum_city = sum(area_city)
max_city = max(area_city)

fig = plt.figure(figsize=(15, 6))
ax = fig.add_subplot(1, 2, 1)
ax.set(title = "Resort Hotel")
plt.bar(area_resort.index, area_resort.values)
for i in range(N):
    plt.text(i * 1, area_resort[i], str(round(area_resort[i] / sum_resort, 2)),
ha='center')

ax = fig.add_subplot(1, 2, 2)
ax.set(title = "City Hotel")
plt.bar(area_city.index, area_city.values)
for i in range(N):
    plt.text(i * 1, area_city[i], str(round(area_city[i] / sum_city, 2)), ha='ce
nter')

plt.show()
```



备注（国家简称）：

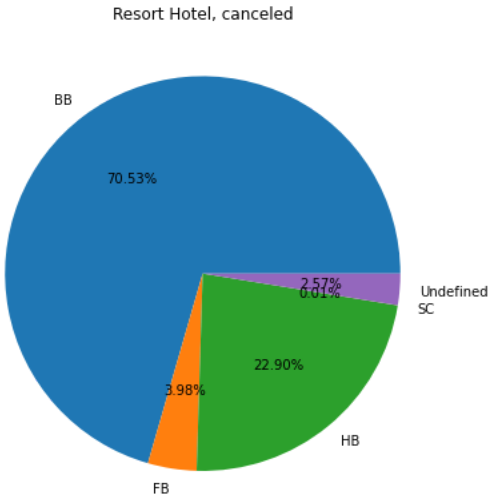
- PRT：葡萄牙，
- GBR：英国，
- ESP：西班牙，
- IRL：爱尔兰，
- FRA：法国，
- DEU：德国，
- CN：中国，
- ITA：意大利，
- NLD：荷兰，
- BEL：比利时

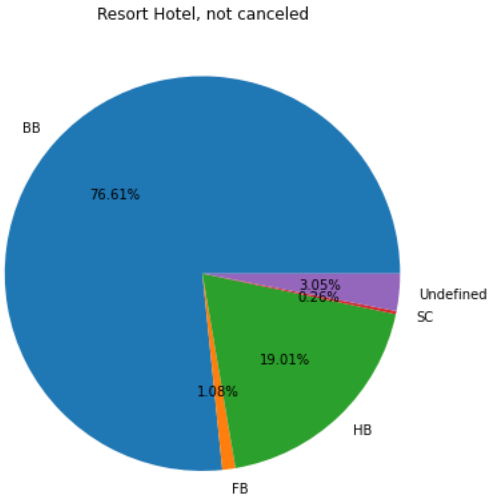
由柱状图可知，这两个酒店的95%以上的顾客都来自于欧洲国家。因此，可以针对这一点，将酒店的服务、餐食等设计的比较符合欧洲人的生活习惯也许会能获得更多的客户。

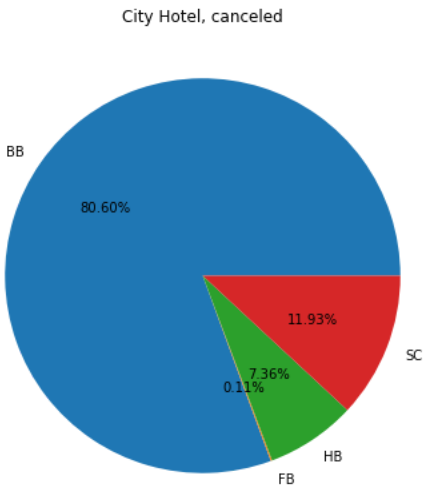
饮食偏好

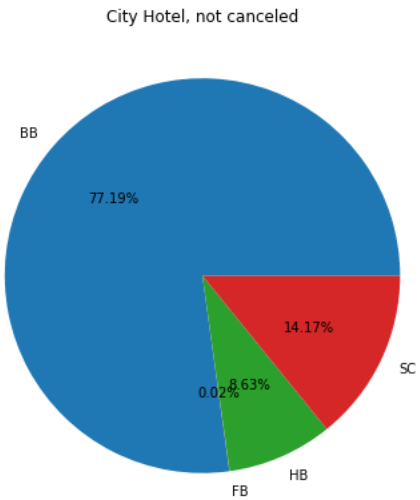
In [22]:

```
meal = hotel.groupby(["hotel", "is_canceled", "meal"]).count().iloc[:, 0]

meal_resort_cancel = meal["Resort Hotel", 1]
meal_resort_notcancel = meal["Resort Hotel", 0]

meal_city_cancel = meal["City Hotel", 1]
meal_city_notcancel = meal["City Hotel", 0]

fig = plt.figure(figsize=(15, 15))

ax = fig.add_subplot(2, 2, 1)
ax.set(title = "Resort Hotel, canceled")
plt.pie(meal_resort_cancel.values, labels=meal_resort_cancel.index, autopct='%1.
2f%%')

ax = fig.add_subplot(2, 2, 2)
ax.set(title = "Resort Hotel, not canceled")
plt.pie(meal_resort_notcancel.values, labels=meal_resort_notcancel.index, autopc
t='%1.2f%%')

ax = fig.add_subplot(2, 2, 3)
ax.set(title = "City Hotel, canceled")
plt.pie(meal_city_cancel.values, labels=meal_city_cancel.index, autopct='%1.2f%%
')

ax = fig.add_subplot(2, 2, 4)
ax.set(title = "City Hotel, not canceled")
plt.pie(meal_city_notcancel.values, labels=meal_city_notcancel.index, autopct='%
1.2f%%')

plt.show()
```







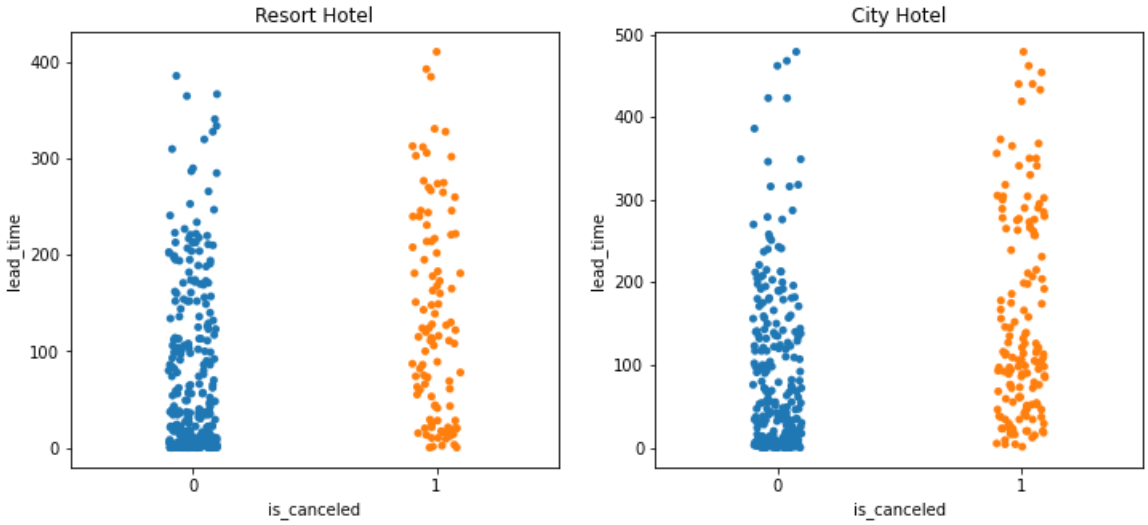


由图可知这两个酒店的取消和没有取消的订单对应的餐食比例差别不大，因此可以认为餐食对订单是否取消没有太大影响。

预订间隔

In [17]:

```
resort = hotel[hotel["hotel"] == "Resort Hotel"][["is_canceled", "lead_time"]]
city = hotel[hotel["hotel"] == "City Hotel"][["is_canceled", "lead_time"]]

fig = plt.figure(figsize=(12, 5))

ax = fig.add_subplot(1, 2, 1)
ax.set(title = "Resort Hotel")
resort = resort.sample(frac=0.01,axis=0, random_state=123)
sns.stripplot(x="is_canceled",y="lead_time",data=resort)

ax = fig.add_subplot(1, 2, 2)
ax.set(title = "City Hotel")
city = city.sample(frac=0.005,axis=0, random_state=123)
sns.stripplot(x="is_canceled",y="lead_time",data=city)

plt.show()
```



由图可知，对这两家酒店而言，都有在不取消订单、提前预订天数较短的区域聚集的特点，因此可以认为提前预订天数较长的订单有较大可能性被取消。

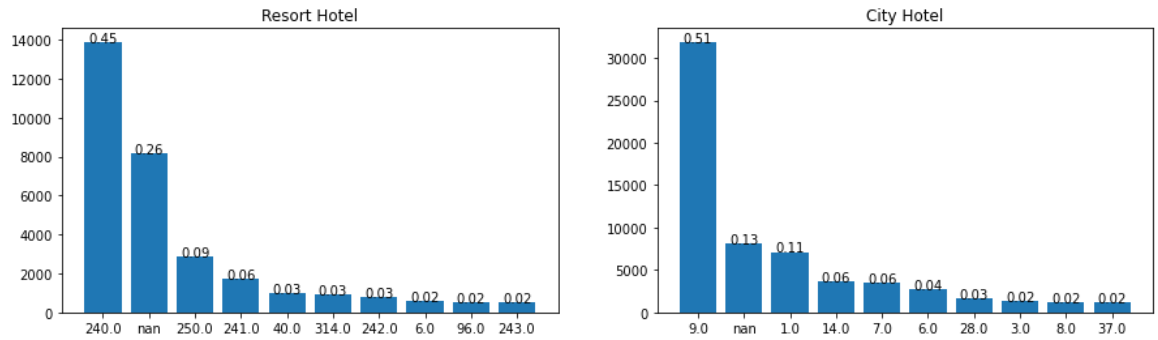
用户画像分析小结

两个酒店的用户最主要都来源于欧洲，因此酒店的设计、服务等可以向迎合欧洲用户习惯的方向管理运营；餐食对订单的取消没有太大的影响，不需要太多考虑这一方面；提前预订天数较长的订单被取消的概率更大。

代理商分析

In [67]:

```
N = 10
hotel_agent = hotel.groupby(["hotel", "agent"]).count().iloc[:, 0]
resort = hotel_agent["Resort Hotel"].sort_values(ascending = False).iloc[: N]
#resort.index = resort.index.astype(str)
sum_resort = sum(resort)
max_resort = max(resort)
city = hotel_agent["City Hotel"].sort_values(ascending = False).iloc[: N]
#city.index = city.index.astype(str)
sum_city = sum(city)
max_city = max(city)

fig = plt.figure(figsize=(15, 4))
ax = fig.add_subplot(1, 2, 1)
ax.set(title = "Resort Hotel")
plt.bar(resort.index, resort.values)
for i in range(N):
    plt.text(i * 1, resort[i], str(round(resort[i] / sum_resort, 2)), ha='cente
r')

ax = fig.add_subplot(1, 2, 2)
ax.set(title = "City Hotel")
plt.bar(city.index, city.values)
for i in range(N):
    plt.text(i * 1, city[i], str(round(city[i] / sum_city, 2)), ha='center')

plt.show()
```

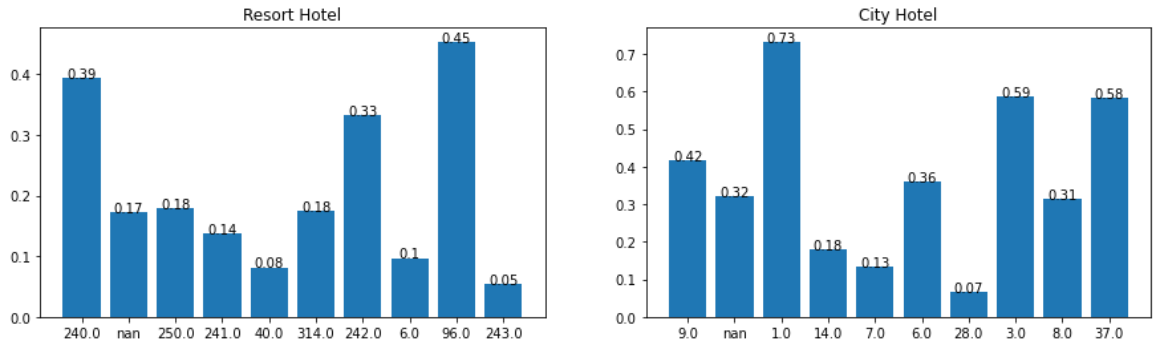


由图中可知，对度假酒店而言，ID为240、250、241的这三家代理商非常重要，通过代理商预订的订单有80%左右都是从这三家代理商完成的。而对城市酒店而言，ID为9、1、14、7、6的这几家代理商非常重要，提供了90%左右的订单。这两家公司可以重点在这些代理商进行更多的宣传。

但是不能单纯从订单量来看，还需要观察代理商订单的取消率。

In [69]:

```
N = 10
hotel_agent = hotel.groupby(["hotel", "is_canceled", "agent"]).count().iloc[:, 0
]
resort_cancel = (hotel_agent["Resort Hotel", 1] / (hotel_agent["Resort Hotel", 1
] + hotel_agent["Resort Hotel", 0]))[list(resort.index)]
city_cancel = (hotel_agent["City Hotel", 1] / (hotel_agent["City Hotel", 1] + ho
tel_agent["City Hotel", 0]))[list(city.index)]

fig = plt.figure(figsize=(15, 4))
ax = fig.add_subplot(1, 2, 1)
ax.set(title = "Resort Hotel")
plt.bar(resort_cancel.index, resort_cancel.values)
for i in range(N):
    plt.text(i * 1, resort_cancel[i], str(round(resort_cancel[i], 2)), ha='cente
r')

ax = fig.add_subplot(1, 2, 2)
ax.set(title = "City Hotel")
plt.bar(city_cancel.index, city_cancel.values)
for i in range(N):
    plt.text(i * 1, city_cancel[i], str(round(city_cancel[i], 2)), ha='center')

plt.show()
```



由图可知，对于度假酒店而言，ID为240的代理商虽然有最多的订单量，但是订单取消率较高，需要进一步考察原因，而其他ID为250、241、40的代理商订单取消率较低。对于城市酒店而言，ID为9、1、6的代理商虽然订单量很多，取消率也非常大，ID为14、7、28的代理商订单量、取消率情况都较好。

结论和建议

可以从以下几点进行建议：

- 酒店运营方面：度假酒店可以在夏季适当降低价格以获取更多的顾客
- 用户分析方面：这两家酒店都可以将餐食、服务等设计得主要符合欧洲用户的生活习惯。对提前预订天数较长的客户，可以主动联系（比如给予一些攻略上的帮助、或者分发当地著名景点的宣传片、优惠券等），以拉拢客户降低取消订单的概率。
- 代理商方面：这两家酒店的预订数最大的代理商都存在取消率较高、订单质量较差的问题，因此需要进一步探究是否存在宣传过度而服务不佳导致顾客流失严重的问题。