In [2]:
```python
import numpy as np
import pandas as pd
import seaborn as sea
import matplotlib.pyplot as mtn
```

In [4]:
```python
df = pd.read_csv(r'C:\Users\chira\Desktop\CSV_Files\Hotel Booking CSV\hotel booking data.csv', encoding = 'unicode_escape')
```

In [4]:
```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  int64
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_week_nights            119390 non-null  int64
 9   adults                          119390 non-null  int64
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  country                         118902 non-null  object
 14  market_segment                  119390 non-null  object
 15  distribution_channel            119390 non-null  object
 16  is_repeated_guest               119390 non-null  int64
 17  previous_cancellations          119390 non-null  int64
 18  previous_bookings_not_canceled  119390 non-null  int64
 19  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64
 26  customer_type                   119390 non-null  object
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  object
dtypes: float64(4), int64(16), object(12)
memory usage: 29.1+ MB
```

In [5]:
```python
df.head()
```

Out[5]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	...	deposit_type	agent	company	days_in_waiting_list	customer_type	adr	required_car_parking_spaces	total_of_specia
0	Resort Hotel	0	342	2015	July	27	1	0	0	2	...	No Deposit	NaN	NaN	0	Transient	0.0	0	
1	Resort Hotel	0	737	2015	July	27	1	0	0	2	...	No Deposit	NaN	NaN	0	Transient	0.0	0	
2	Resort Hotel	0	7	2015	July	27	1	0	1	1	...	No Deposit	NaN	NaN	0	Transient	75.0	0	
3	Resort Hotel	0	13	2015	July	27	1	0	1	1	...	No Deposit	304.0	NaN	0	Transient	75.0	0	
4	Resort Hotel	0	14	2015	July	27	1	0	2	2	...	No Deposit	240.0	NaN	0	Transient	98.0	0	

5 rows × 32 columns

In [7]:
```python
df["reservation_status_date"] = pd.to_datetime(df["reservation_status_date"], format = "mixed")
```

In [9]:
```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  int64
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_week_nights            119390 non-null  int64
 9   adults                          119390 non-null  int64
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  country                         118902 non-null  object
 14  market_segment                  119390 non-null  object
 15  distribution_channel            119390 non-null  object
 16  is_repeated_guest               119390 non-null  int64
 17  previous_cancellations          119390 non-null  int64
 18  previous_bookings_not_canceled  119390 non-null  int64
 19  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64
 26  customer_type                   119390 non-null  object
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  datetime64[ns]
dtypes: datetime64[ns](1), float64(4), int64(16), object(11)
memory usage: 29.1+ MB
```

1. unique values of object datatype

In [10]:
```python
df.describe(include = "object")
```

Out[10]:

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reserved_room_type	assigned_room_type	deposit_type	customer_type	reservation_status
count	119390	119390	119390	118902	119390	119390	119390	119390	119390	119390	119390
unique	2	12	5	177	8	5	10	12	3	4	3
top	City Hotel	August	BB	PRT	Online TA	TA/TO	A	A	No Deposit	Transient	Check-Out
freq	79330	13877	92310	48590	56477	97870	85994	74053	104641	89613	75166

In [11]:
```python
for col in df.describe(include = "object"):
    print(col)
    print(df[col].unique())
    print("_"*50)
```

```
hotel
['Resort Hotel' 'City Hotel']
__________________________________________________
arrival_date_month
['July' 'August' 'September' 'October' 'November' 'December' 'January'
 'February' 'March' 'April' 'May' 'June']
__________________________________________________
meal
['BB' 'FB' 'HB' 'SC' 'Undefined']
__________________________________________________
country
['PRT' 'GBR' 'USA' 'ESP' 'IRL' 'FRA' nan 'ROU' 'NOR' 'OMN' 'ARG' 'POL'
 'DEU' 'BEL' 'CHE' 'CN' 'GRC' 'ITA' 'NLD' 'DNK' 'RUS' 'SWE' 'AUS' 'EST'
 'CZE' 'BRA' 'FIN' 'MOZ' 'BWA' 'LUX' 'SVN' 'ALB' 'IND' 'CHN' 'MEX' 'MAR'
 'UKR' 'SMR' 'LVA' 'PRI' 'SRB' 'CHL' 'AUT' 'BLR' 'LTU' 'TUR' 'ZAF' 'AGO'
 'ISR' 'CYM' 'ZMB' 'CPV' 'ZWE' 'DZA' 'KOR' 'CRI' 'HUN' 'ARE' 'TUN' 'JAM'
 'HRV' 'HKG' 'IRN' 'GEO' 'AND' 'GIB' 'URY' 'JEY' 'CAF' 'CYP' 'COL' 'GGY'
 'KWT' 'NGA' 'MDV' 'VEN' 'SVK' 'FJI' 'KAZ' 'PAK' 'IDN' 'LBN' 'PHL' 'SEN'
 'SYC' 'AZE' 'BHR' 'NZL' 'THA' 'DOM' 'MKD' 'MYS' 'ARM' 'JPN' 'LKA' 'CUB'
 'CMR' 'BIH' 'MUS' 'COM' 'SUR' 'UGA' 'BGR' 'CIV' 'JOR' 'SYR' 'SGP' 'BDI'
 'SAU' 'VNM' 'PLW' 'QAT' 'EGY' 'PER' 'MLT' 'MWI' 'ECU' 'MDG' 'ISL' 'UZB'
 'NPL' 'BHS' 'MAC' 'TGO' 'TWN' 'DJI' 'STP' 'KNA' 'ETH' 'IRQ' 'HND' 'RWA'
 'KHM' 'MCO' 'BGD' 'IMN' 'TJK' 'NIC' 'BEN' 'VGB' 'TZA' 'GAB' 'GHA' 'TMP'
 'GLP' 'KEN' 'LIE' 'GNB' 'MNE' 'UMI' 'MYT' 'FRO' 'MMR' 'PAN' 'BFA' 'LBY'
 'MLI' 'NAM' 'BOL' 'PRY' 'BRB' 'ABW' 'AIA' 'SLV' 'DMA' 'PYF' 'GUY' 'LCA'
 'ATA' 'GTM' 'ASM' 'MRT' 'NCL' 'KIR' 'SDN' 'ATF' 'SLE' 'LAO']
__________________________________________________
market_segment
['Direct' 'Corporate' 'Online TA' 'Offline TA/TO' 'Complementary' 'Groups'
 'Undefined' 'Aviation']
__________________________________________________
distribution_channel
['Direct' 'Corporate' 'TA/TO' 'Undefined' 'GDS']
__________________________________________________
reserved_room_type
['C' 'A' 'D' 'E' 'G' 'F' 'H' 'L' 'P' 'B']
__________________________________________________
assigned_room_type
['C' 'A' 'D' 'E' 'G' 'F' 'I' 'B' 'H' 'P' 'L' 'K']
__________________________________________________
deposit_type
['No Deposit' 'Refundable' 'Non Refund']
__________________________________________________
customer_type
['Transient' 'Contract' 'Transient-Party' 'Group']
__________________________________________________
reservation_status
['Check-Out' 'Canceled' 'No-Show']
__________________________________________________
```

In [12]:
```python
pd.isnull(df).sum()
```

Out[12]:
```
hotel                                  0
is_canceled                            0
lead_time                              0
arrival_date_year                      0
arrival_date_month                     0
arrival_date_week_number               0
arrival_date_day_of_month              0
stays_in_weekend_nights                0
stays_in_week_nights                   0
adults                                 0
children                               4
babies                                 0
meal                                   0
country                              488
market_segment                         0
distribution_channel                   0
is_repeated_guest                      0
previous_cancellations                 0
previous_bookings_not_canceled         0
reserved_room_type                     0
assigned_room_type                     0
booking_changes                        0
deposit_type                           0
agent                              16340
company                           112593
days_in_waiting_list                   0
customer_type                          0
adr                                    0
required_car_parking_spaces            0
total_of_special_requests              0
reservation_status                     0
reservation_status_date                0
dtype: int64
```

In [13]:
```python
df.drop(columns = ["agent", "company"], axis = 1, inplace = True)
```

In [16]:
```python
df.dropna(inplace = True)
```

In [17]:
```python
pd.isnull(df).sum()
```

Out[17]:
```
hotel                             0
is_canceled                       0
lead_time                         0
arrival_date_year                 0
arrival_date_month                0
arrival_date_week_number          0
arrival_date_day_of_month         0
stays_in_weekend_nights           0
stays_in_week_nights              0
adults                            0
children                          0
babies                            0
meal                              0
country                           0
market_segment                    0
distribution_channel              0
is_repeated_guest                 0
previous_cancellations            0
previous_bookings_not_canceled    0
reserved_room_type                0
assigned_room_type                0
booking_changes                   0
deposit_type                      0
days_in_waiting_list              0
customer_type                     0
adr                               0
required_car_parking_spaces       0
total_of_special_requests         0
reservation_status                0
reservation_status_date           0
dtype: int64
```

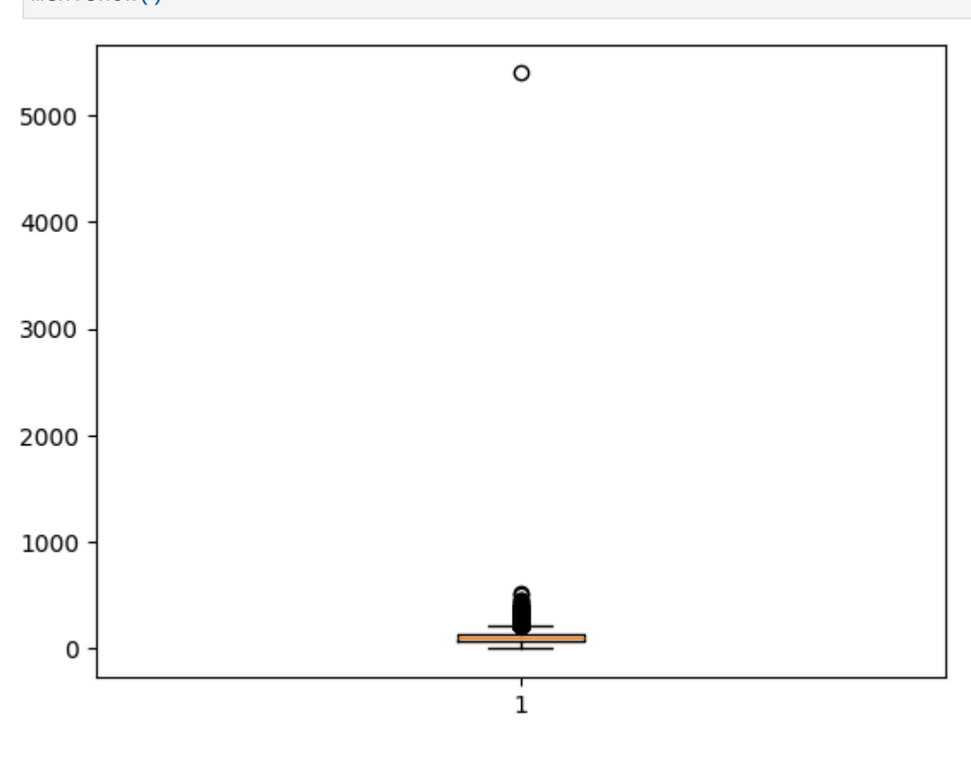
2. Create a BoxPlot for "Adr" Values

In [18]:
```python
df.describe()
```

Out[18]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	children	babies	is_repeated_guest	previous_cancellations	previous_bookings_not_canceled	booking_changes	da
count	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	118898.000000	
mean	0.371352	104.311435	2016.157656	27.166555	15.800880	0.928897	2.502145	1.858391	0.104207	0.007948	0.032011	0.087142	0.131634	0.221181	
min	0.000000	0.000000	2015.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	18.000000	2016.000000	16.000000	8.000000	0.000000	1.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
50%	0.000000	69.000000	2016.000000	28.000000	16.000000	1.000000	2.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
75%	1.000000	161.000000	2017.000000	38.000000	23.000000	2.000000	3.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
max	1.000000	737.000000	2017.000000	53.000000	31.000000	16.000000	41.000000	55.000000	10.000000	10.000000	1.000000	26.000000	72.000000	21.000000	
std	0.483168	106.903309	0.707459	13.589971	8.780324	0.996216	1.900168	0.578576	0.399172	0.097380	0.176029	0.845869	1.484672	0.652785	

In [20]:
```python
mtn.boxplot(df["adr"])
mtn.show()
```



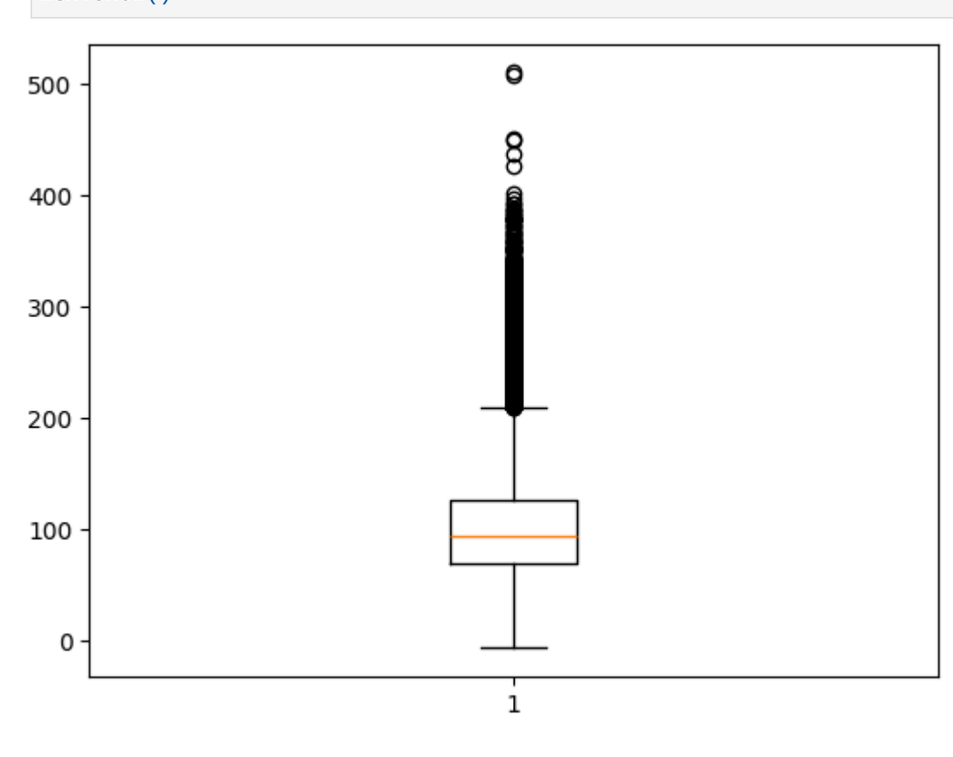
In [23]:
```python
df1 = df[(df["adr"] < 5000)]
```

In [24]:
```python
df1.describe()
```

Out[24]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	children	babies	is_repeated_guest	previous_cancellations	previous_bookings_not_canceled	booking_changes	da
count	118897.000000	118897.000000	118897.000000	118897.000000	118897.000000	118897.000000	118897.000000	118897.000000	118897.000000	118897.000000	118897.000000	118897.000000	118897.000000	118897.000000	
mean	0.371347	104.312018	2016.157657	27.166674	15.800802	0.928905	2.502157	1.858390	0.104208	0.007948	0.032011	0.087143	0.131635	0.221175	
min	0.000000	0.000000	2015.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	18.000000	2016.000000	16.000000	8.000000	0.000000	1.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
50%	0.000000	69.000000	2016.000000	28.000000	16.000000	1.000000	2.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
75%	1.000000	161.000000	2017.000000	38.000000	23.000000	2.000000	3.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
max	1.000000	737.000000	2017.000000	53.000000	31.000000	16.000000	41.000000	55.000000	10.000000	10.000000	1.000000	26.000000	72.000000	21.000000	
std	0.483167	106.903570	0.707462	13.589966	8.780321	0.996217	1.900171	0.578578	0.399174	0.097381	0.176030	0.845872	1.484678	0.652784	

In [25]:
```python
mtn.boxplot(df1["adr"])
mtn.show()
```



In [26]:
```python
df.head()
```

Out[26]:

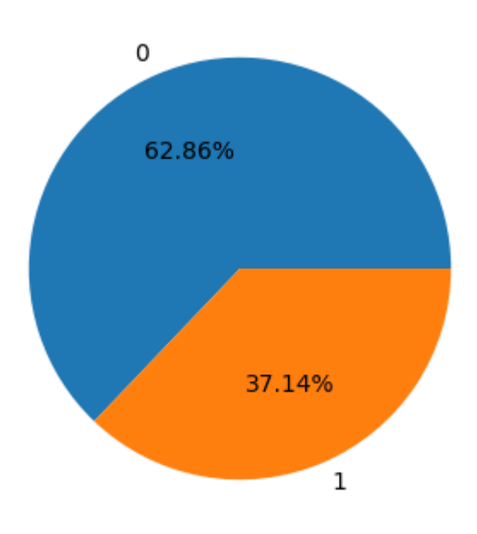
	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	...	assigned_room_type	booking_changes	deposit_type	days_in_waiting_list	customer_type	adr	required_car_parking_s
0	Resort Hotel	0	342	2015	July	27	1	0	0	2	...	C	3	No Deposit	0	Transient	0.0	
1	Resort Hotel	0	737	2015	July	27	1	0	0	2	...	C	4	No Deposit	0	Transient	0.0	
2	Resort Hotel	0	7	2015	July	27	1	0	1	1	...	C	0	No Deposit	0	Transient	75.0	
3	Resort Hotel	0	13	2015	July	27	1	0	1	1	...	A	0	No Deposit	0	Transient	75.0	
4	Resort Hotel	0	14	2015	July	27	1	0	2	2	...	A	0	No Deposit	0	Transient	98.0	

5 rows × 30 columns

3. How many reservations are cancelled or not

In [29]:
```python
canc = df["is_canceled"].value_counts(normalize = True)
```

In [33]:
```python
mtn.figure(figsize = (6,4))
mtn.pie(canc, autopct = "%1.2f%%", labels = canc.index)
mtn.show()
```



4. Which hotel have more number of cancellation rate

In [34]:
```python
df.head()
```

Out[34]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	...	assigned_room_type	booking_changes	deposit_type	days_in_waiting_list	customer_type	adr	required_car_parking_s
0	Resort Hotel	0	342	2015	July	27	1	0	0	2	...	C	3	No Deposit	0	Transient	0.0	
1	Resort Hotel	0	737	2015	July	27	1	0	0	2	...	C	4	No Deposit	0	Transient	0.0	
2	Resort Hotel	0	7	2015	July	27	1	0	1	1	...	C	0	No Deposit	0	Transient	75.0	
3	Resort Hotel	0	13	2015	July	27	1	0	1	1	...	A	0	No Deposit	0	Transient	75.0	
4	Resort Hotel	0	14	2015	July	27	1	0	2	2	...	A	0	No Deposit	0	Transient	98.0	

5 rows × 30 columns

In [12]:
```python
cancel = df[(df["is_canceled"] == 1)]
```

In [13]:
```python
cancel.head()
```

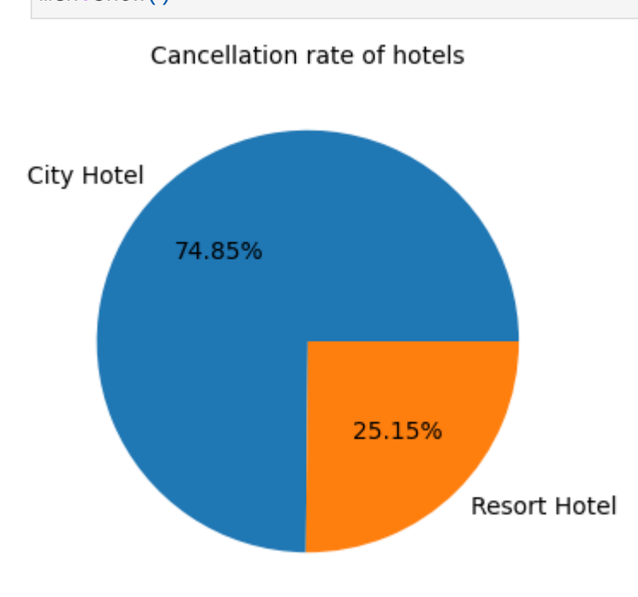
Out[13]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	...	agent	company	days_in_waiting_list	customer_type	adr	required_car_parking_spaces	total_of_special_requests
8	Resort Hotel	1	85	2015	July	27	1	0	3	2	...	240.0	NaN	0	Transient	82.0	0	1
9	Resort Hotel	1	75	2015	July	27	1	0	3	2	...	15.0	NaN	0	Transient	105.5	0	0
10	Resort Hotel	1	23	2015	July	27	1	0	4	2	...	240.0	NaN	0	Transient	123.0	0	0
27	Resort Hotel	1	60	2015	July	27	1	2	5	2	...	240.0	NaN	0	Transient	107.0	0	2
32	Resort Hotel	1	96	2015	July	27	1	2	8	2	...	NaN	NaN	0	Transient	108.3	0	2

5 rows × 33 columns

In [21]:
```python
both = cancel["hotel"].value_counts(normalize = True)
```

In [22]:
```python
mtn.figure(figsize = (5,4))
mtn.pie(both, autopct = "%1.2f%%", labels = both.index)
mtn.title("Cancellation rate of hotels", fontsize = 10)
mtn.show()
```



5. What is the cancellation rate in resort hotel?

In [15]:
```python
cancel["hotel"].value_counts(normalize = True)
```

Out[15]:
```
hotel
City Hotel      0.748508
Resort Hotel    0.251492
Name: proportion, dtype: float64
```

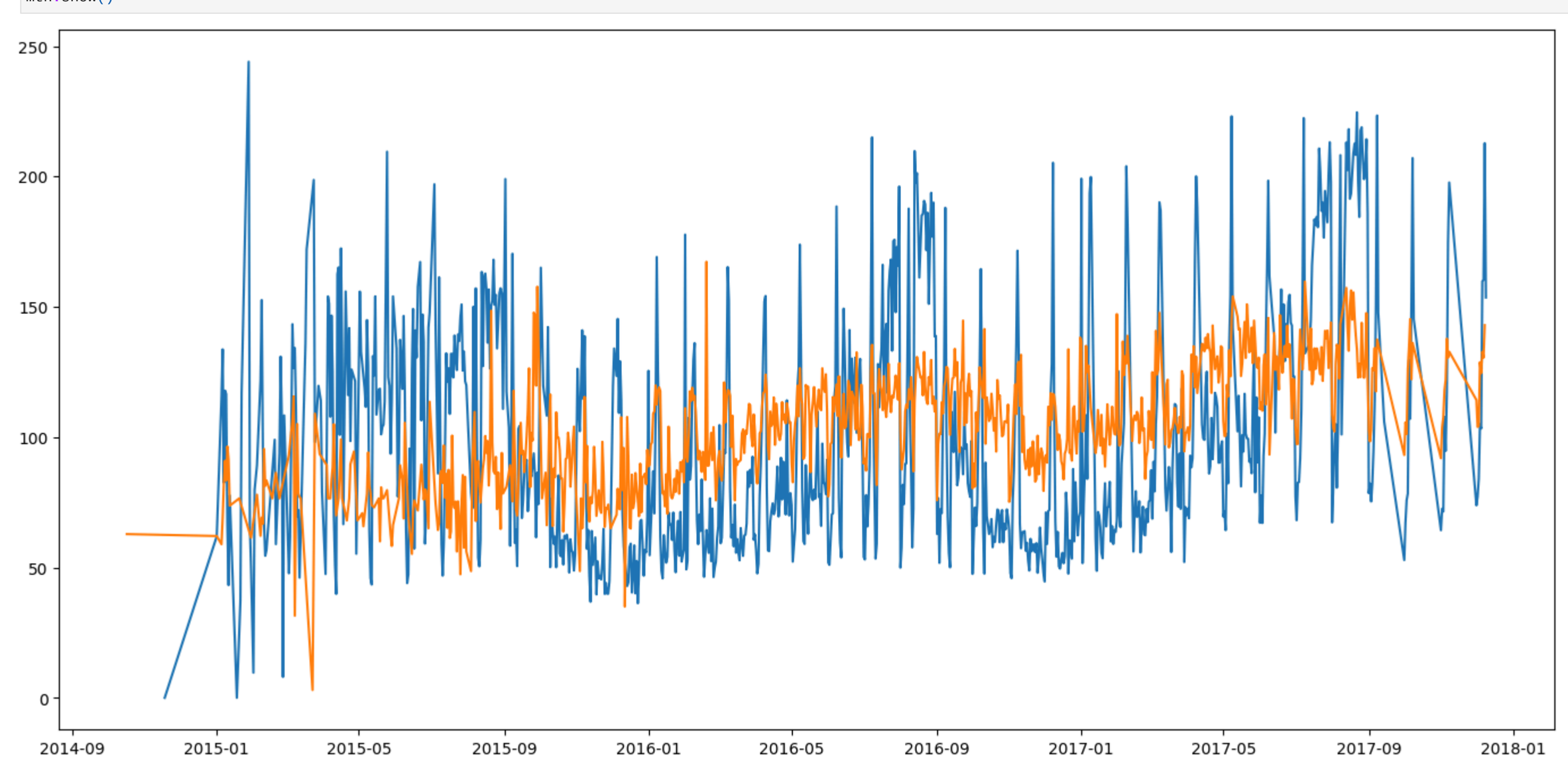
Cancellation Rate in Resort hotel is 25%

6. Average daily rate in city and resort hotels

In [67]:
```python
ree = df[(df["hotel"] == "Resort Hotel")]
cii = df[(df["hotel"] == "City Hotel")]
```

In [68]:
```python
resort = ree.groupby(["reservation_status_date"])[["adr"]].mean()
cittt = cii.groupby(["reservation_status_date"])[["adr"]].mean()
```

In [70]:
```python
mtn.figure(figsize = (17,8))
mtn.plot(resort.index, resort["adr"], label = "resort")
mtn.plot(cittt.index, cittt["adr"], label = "cittt")
mtn.show()
```



7. Reservation status per month

In [71]:
```python
df.head()
```

Out[71]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	...	assigned_room_type	booking_changes	deposit_type	days_in_waiting_list	customer_type	adr	required_car_parking_s
0	Resort Hotel	0	342	2015	July	27	1	0	0	2	...	C	3	No Deposit	0	Transient	0.0	
1	Resort Hotel	0	737	2015	July	27	1	0	0	2	...	C	4	No Deposit	0	Transient	0.0	
2	Resort Hotel	0	7	2015	July	27	1	0	1	1	...	C	0	No Deposit	0	Transient	75.0	
3	Resort Hotel	0	13	2015	July	27	1	0	1	1	...	A	0	No Deposit	0	Transient	75.0	
4	Resort Hotel	0	14	2015	July	27	1	0	2	2	...	A	0	No Deposit	0	Transient	98.0	

5 rows × 30 columns

In [8]:
```python
df["month"] = df["reservation_status_date"].dt.month
```

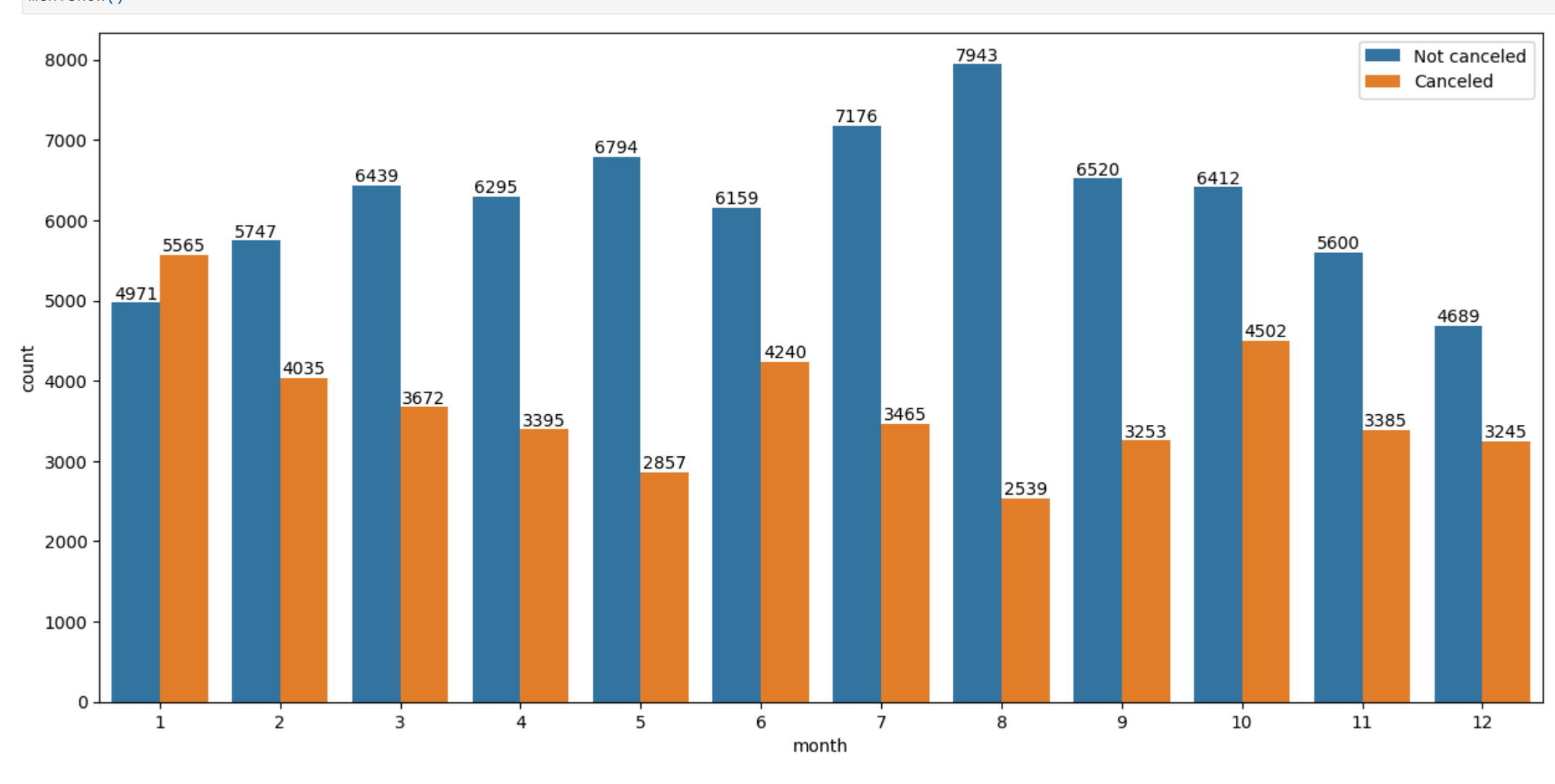
In [73]:
```python
df.head(2)
```

Out[73]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	...	booking_changes	deposit_type	days_in_waiting_list	customer_type	adr	required_car_parking_spaces	total_of_special_
0	Resort Hotel	0	342	2015	July	27	1	0	0	2	...	3	No Deposit	0	Transient	0.0	0	
1	Resort Hotel	0	737	2015	July	27	1	0	0	2	...	4	No Deposit	0	Transient	0.0	0	

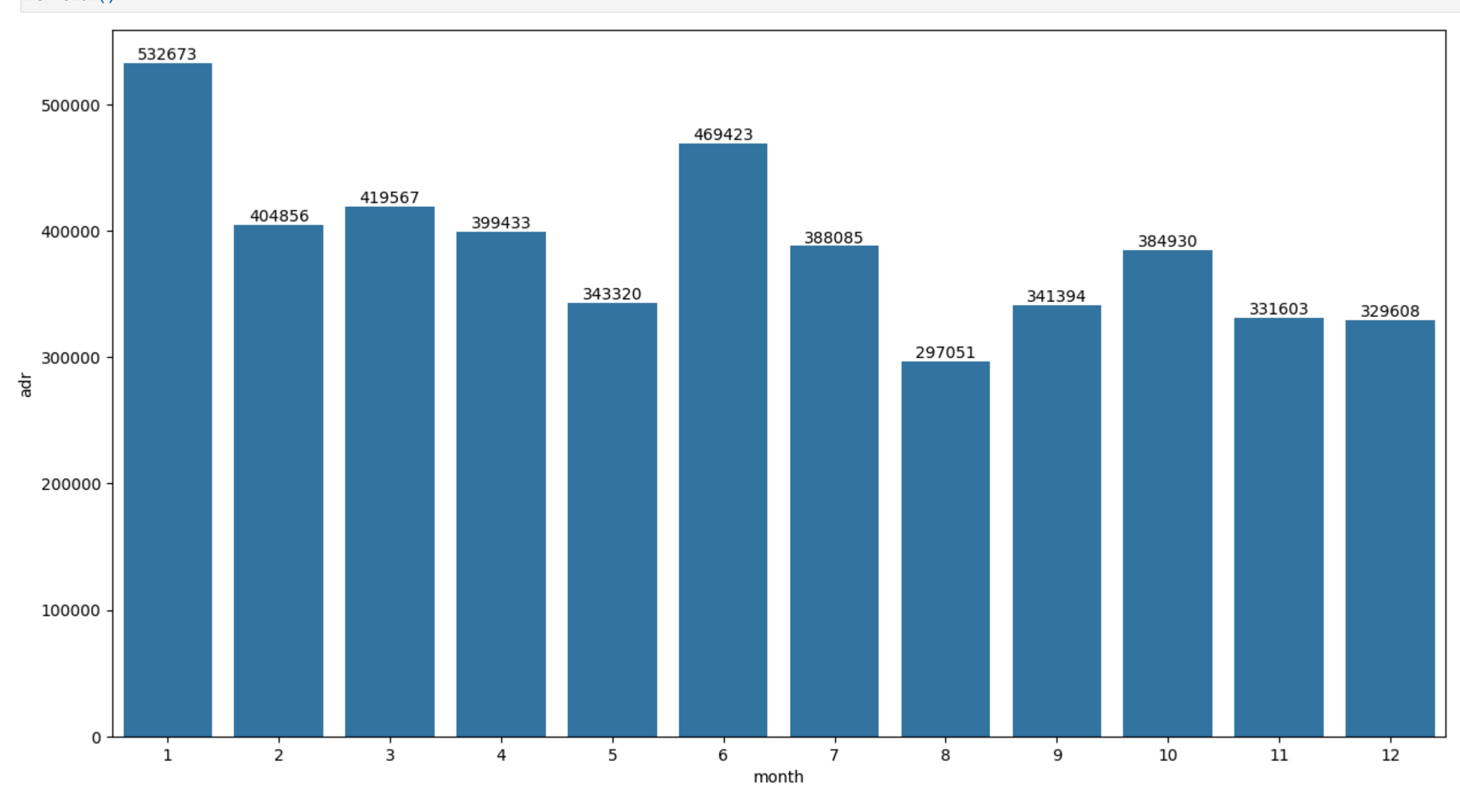
2 rows × 31 columns

In [85]:
```python
mtn.figure(figsize = (15,7))
ax = sea.countplot(x = "month", hue = "is_canceled", data = df)
for container in ax.containers:
    ax.bar_label(container)
mtn.legend(["Not canceled", "Canceled"])
mtn.show()
```



8. ADR per month in cancelled data

In [10]:
```python
mtn.figure(figsize = (15,8))
ax = sea.barplot(x = "month", y = "adr", data = df[(df["is_canceled"] == 1)].groupby("month")[["adr"]].sum().reset_index())
ax.bar_label(ax.containers[0])
mtn.show()
```



9. Top 10 countries on the behalf of cancelled data

In [94]:
```python
top = cancel["country"].value_counts(normalize = True).head(10)
```

mtn.figure(figsize = (18,7)) mtn.pie(top, autopct = "%1.2f%%", labels = top.index) mtn.show()

10. Market segment value counts

In [99]:
```python
df["market_segment"].value_counts(normalize = True)
```

Out[99]:
```
market_segment
Online TA        0.474373
Offline TA/TO    0.203199
Groups           0.166580
Direct           0.104695
Corporate        0.042986
Complementary    0.006173
Aviation         0.001993
Name: proportion, dtype: float64
```

11. Cancelled data in market segments

In [101…
```python
cancel["market_segment"].value_counts(normalize = True)
```

Out[101…
```
market_segment
Online TA        0.469685
Groups           0.273979
Offline TA/TO    0.187484
Direct           0.043485
Corporate        0.022150
Complementary    0.002038
```

```
Acceleration                    0.0011176
Model: gpt-oss-20b, 20 gen. Tokenize
```