DATA ANALYSIS - HOTEL BOOKING

Importing Libraries

In [1]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

Loading The Dataset

In [2]:

```
df = pd.read_csv(r'C:\Users\ADMIN\Desktop\Python Projects - Data Sci Lovers\hotel_booking.csv')
```

Exploratory Data Analysis and Data Cleaning

In [3]:

```
df.head()
```

Out[3]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_
0	Resort Hotel	0	342	2015	July	27	1	
1	Resort Hotel	0	737	2015	July	27	1	
2	Resort Hotel	0	7	2015	July	27	1	
3	Resort Hotel	0	13	2015	July	27	1	
4	Resort Hotel	0	14	2015	July	27	1	

5 rows × 36 columns

In [4]:

```
df.shape
```

Out[4]:

```
(119390, 36)
```

In [5]:

```
df.columns
```

Out[5]:

```
Index(['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', 'children', 'babies', 'meal',
       'country', 'market_segment', 'distribution_channel',
       'is_repeated_guest', 'previous_cancellations',
       'previous_bookings_not_canceled', 'reserved_room_type',
       'assigned_room_type', 'booking_changes', 'deposit_type', 'agent',
       'company', 'days_in_waiting_list', 'customer_type', 'adr',
       'required_car_parking_spaces', 'total_of_special_requests',
       'reservation_status', 'reservation_status_date', 'name', 'email',
       'phone-number', 'credit_card'],
      dtype='object')
```

In [6]:

```
df.drop(columns=['name','email','phone-number','credit_card'],axis=1,inplace=True)
```

In [7]:

```
df.shape
```

Out[7]:

```
(119390, 32)
```

In [8]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
hotel                           119390 non-null object
is_canceled                     119390 non-null int64
lead_time                       119390 non-null int64
arrival_date_year               119390 non-null int64
arrival_date_month              119390 non-null object
arrival_date_week_number        119390 non-null int64
arrival_date_day_of_month       119390 non-null int64
stays_in_weekend_nights         119390 non-null int64
stays_in_week_nights            119390 non-null int64
adults                          119390 non-null int64
children                        119386 non-null float64
babies                          119390 non-null int64
meal                            119390 non-null object
country                         118902 non-null object
market_segment                  119390 non-null object
distribution_channel            119390 non-null object
is_repeated_guest               119390 non-null int64
previous_cancellations          119390 non-null int64
previous_bookings_not_canceled  119390 non-null int64
reserved_room_type              119390 non-null object
assigned_room_type              119390 non-null object
booking_changes                 119390 non-null int64
deposit_type                    119390 non-null object
agent                           103050 non-null float64
company                         6797 non-null float64
days_in_waiting_list            119390 non-null int64
customer_type                   119390 non-null object
adr                             119390 non-null float64
required_car_parking_spaces     119390 non-null int64
total_of_special_requests       119390 non-null int64
reservation_status              119390 non-null object
reservation_status_date         119390 non-null object
dtypes: float64(4), int64(16), object(12)
memory usage: 29.1+ MB
```

In [9]:

```
df['reservation_status_date']=pd.to_datetime(df['reservation_status_date'])
```

In [10]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
hotel                           119390 non-null object
is_canceled                     119390 non-null int64
lead_time                       119390 non-null int64
arrival_date_year               119390 non-null int64
arrival_date_month              119390 non-null object
arrival_date_week_number        119390 non-null int64
arrival_date_day_of_month       119390 non-null int64
stays_in_weekend_nights         119390 non-null int64
stays_in_week_nights            119390 non-null int64
adults                          119390 non-null int64
children                        119386 non-null float64
babies                          119390 non-null int64
meal                            119390 non-null object
country                         118902 non-null object
market_segment                  119390 non-null object
distribution_channel            119390 non-null object
is_repeated_guest               119390 non-null int64
previous_cancellations          119390 non-null int64
previous_bookings_not_canceled  119390 non-null int64
reserved_room_type              119390 non-null object
assigned_room_type              119390 non-null object
booking_changes                 119390 non-null int64
deposit_type                    119390 non-null object
agent                           103050 non-null float64
company                         6797 non-null float64
days_in_waiting_list            119390 non-null int64
customer_type                   119390 non-null object
adr                             119390 non-null float64
required_car_parking_spaces     119390 non-null int64
total_of_special_requests       119390 non-null int64
reservation_status              119390 non-null object
reservation_status_date         119390 non-null datetime64[ns]
dtypes: datetime64[ns](1), float64(4), int64(16), object(11)
memory usage: 29.1+ MB
```

In [11]:

```
df.describe()
```

Out[11]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights
count	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000
mean	0.370416	104.011416	2016.156554	27.165173	15.798241	0.927599
std	0.482918	106.863097	0.707476	13.605138	8.780829	0.998613
min	0.000000	0.000000	2015.000000	1.000000	1.000000	0.000000
25%	0.000000	18.000000	2016.000000	16.000000	8.000000	0.000000
50%	0.000000	69.000000	2016.000000	28.000000	16.000000	1.000000
75%	1.000000	160.000000	2017.000000	38.000000	23.000000	2.000000
max	1.000000	737.000000	2017.000000	53.000000	31.000000	19.000000

In [12]:

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

Out[12]:

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reserved_room_type	assigned_room_type
count	119390	119390	119390	118902	119390	119390	119390	119390
unique	2	12	5	177	8	5	10	12
top	City Hotel	August	BB	PRT	Online TA	TA/TO	A	A
freq	79330	13877	92310	48590	56477	97870	85994	74053

In [13]:

```
df.describe(include='object').columns
```

Out[13]:

```
Index(['hotel', 'arrival_date_month', 'meal', 'country', 'market_segment',
       'distribution_channel', 'reserved_room_type', 'assigned_room_type',
       'deposit_type', 'customer_type', 'reservation_status'],
      dtype='object')
```

In [14]:

```
for col in df.describe(include='object').columns:
    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 [15]:

```
df.isnull().sum()
```

Out[15]:

```
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 [16]:

```
df.drop(['agent','company'],axis=1,inplace=True)
```

In [17]:

```
df.dropna(inplace=True)
```

In [18]:

```
df.isnull().sum()
```

Out[18]:

```
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
```

In [19]:

```
df.describe()
```

Out[19]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights
count	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
std	0.483168	106.903309	0.707459	13.589971	8.780324	0.996216
min	0.000000	0.000000	2015.000000	1.000000	1.000000	0.000000
25%	0.000000	18.000000	2016.000000	16.000000	8.000000	0.000000
50%	0.000000	69.000000	2016.000000	28.000000	16.000000	1.000000
75%	1.000000	161.000000	2017.000000	38.000000	23.000000	2.000000
max	1.000000	737.000000	2017.000000	53.000000	31.000000	16.000000

In [20]:

```
df = df[df['adr']<5000]
```

In [21]:

```
df.shape
```

Out[21]:

```
(118897, 30)
```

Data Analysis And Visualization

In [22]:

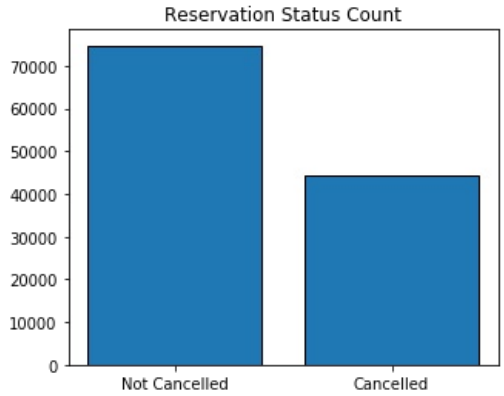
```
cancelled_perc = df['is_canceled'].value_counts(normalize = True)
cancelled_perc
```

Out[22]:

```
0    0.628653
1    0.371347
Name: is_canceled, dtype: float64
```

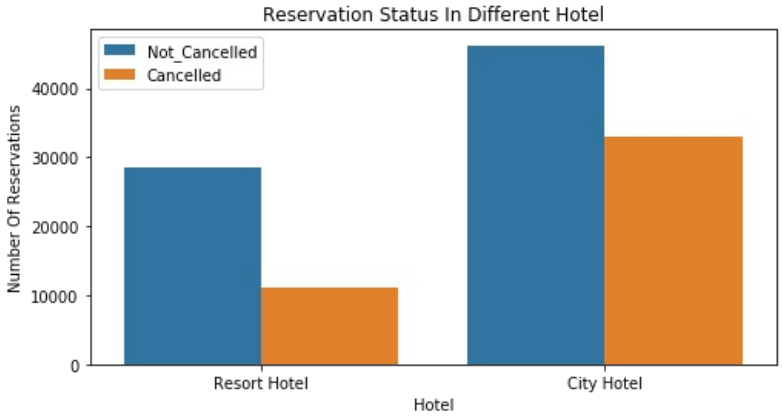
In [23]:

```
plt.figure(figsize=(5,4))
plt.bar(['Not Cancelled','Cancelled'],df['is_canceled'].value_counts(),edgecolor='k')
plt.title('Reservation Status Count')
plt.show()
```



In [24]:

```
plt.figure(figsize=(8,4))
sns.countplot(x='hotel',hue='is_canceled',data=df)
plt.legend(['Not_Cancelled','Cancelled'])
plt.title('Reservation Status In Different Hotel')
plt.xlabel('Hotel')
plt.ylabel('Number Of Reservations')
plt.show()
```



In [25]:

```
df['hotel'].value_counts()
```

Out[25]:

```
City Hotel      79301
Resort Hotel    39596
Name: hotel, dtype: int64
```

In [26]:

```
City_Hotel = df[df['hotel']=='City Hotel']
City_Hotel['is_canceled'].value_counts(normalize = True)
```

Out[26]:

```
0    0.582918
1    0.417082
Name: is_canceled, dtype: float64
```

In [27]:

```
Resort_Hotel = df[df['hotel']=='Resort Hotel']
Resort_Hotel['is_canceled'].value_counts(normalize = True)
```

Out[27]:

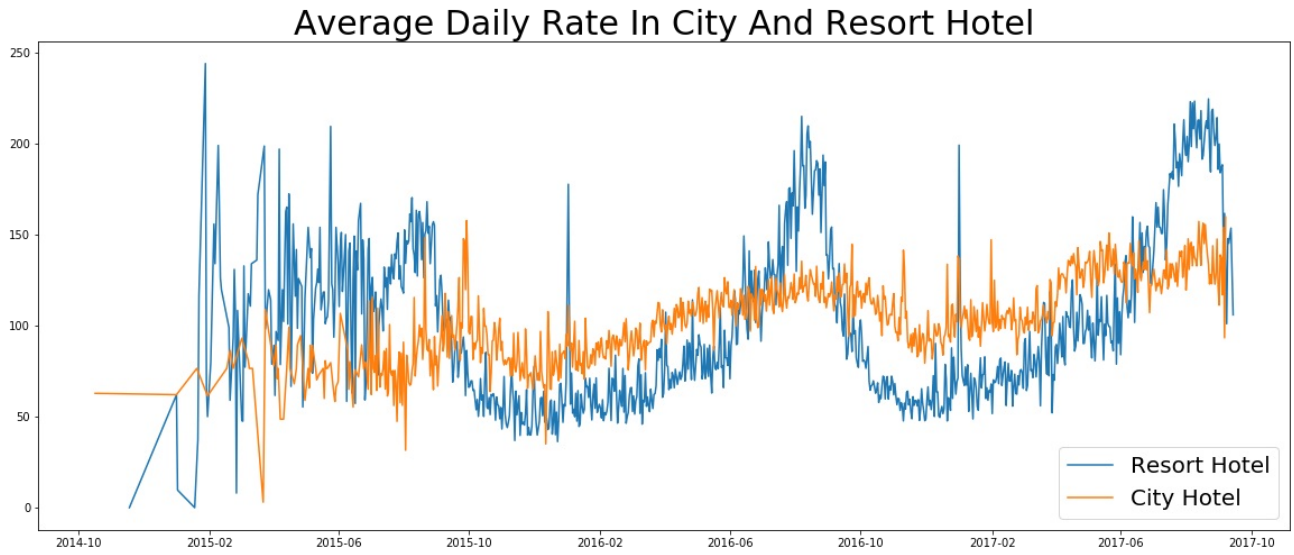
```
0    0.72025
1    0.27975
Name: is_canceled, dtype: float64
```

In [28]:

```
Resort_Hotel=Resort_Hotel.groupby('reservation_status_date')[['adr']].mean()
City_Hotel=City_Hotel.groupby('reservation_status_date')[['adr']].mean()
```

In [29]:

```
plt.figure(figsize=(20,8))
plt.title('Average Daily Rate In City And Resort Hotel',fontsize=30)
plt.plot(Resort_Hotel.index,Resort_Hotel['adr'],label='Resort Hotel')
plt.plot(City_Hotel.index,City_Hotel['adr'],label='City Hotel')
plt.legend(fontsize=20)
plt.show()
```



In [30]:

```
df.head(2)
```

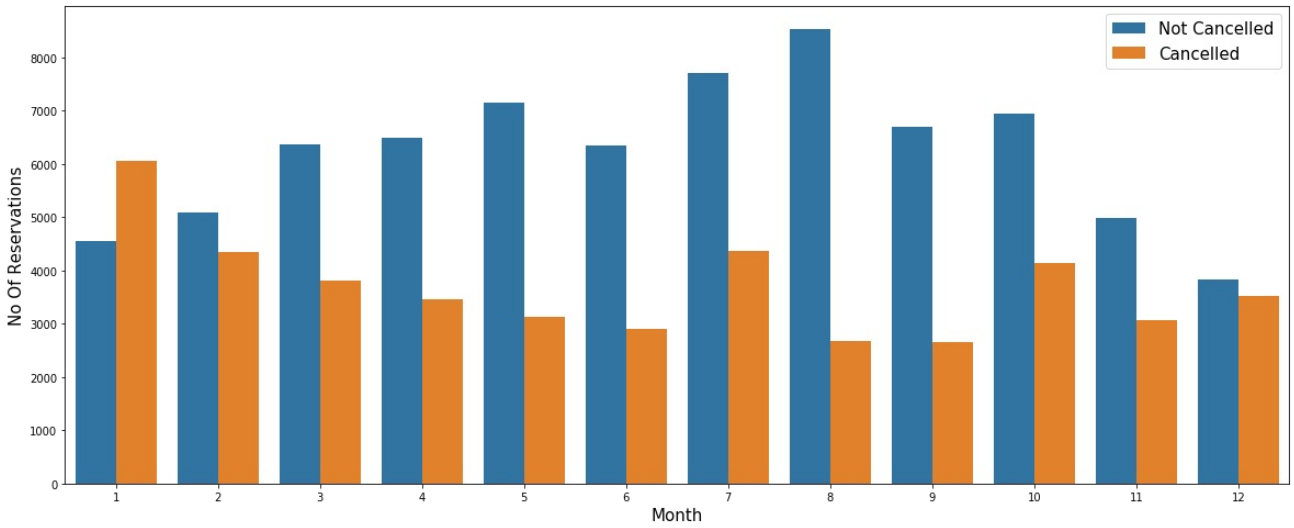
Out[30]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_
0	Resort Hotel	0	342	2015	July	27	1	
1	Resort Hotel	0	737	2015	July	27	1	

2 rows × 30 columns

In [31]:

```
df['month']=df['reservation_status_date'].dt.month
plt.figure(figsize=(20,8))
sns.countplot(x='month',hue='is_canceled',data=df)
plt.legend(['Not Cancelled','Cancelled'],fontsize=15)
plt.xlabel('Month',fontsize=15)
plt.ylabel('No Of Reservations',fontsize=15)
plt.show()
```



In [32]:

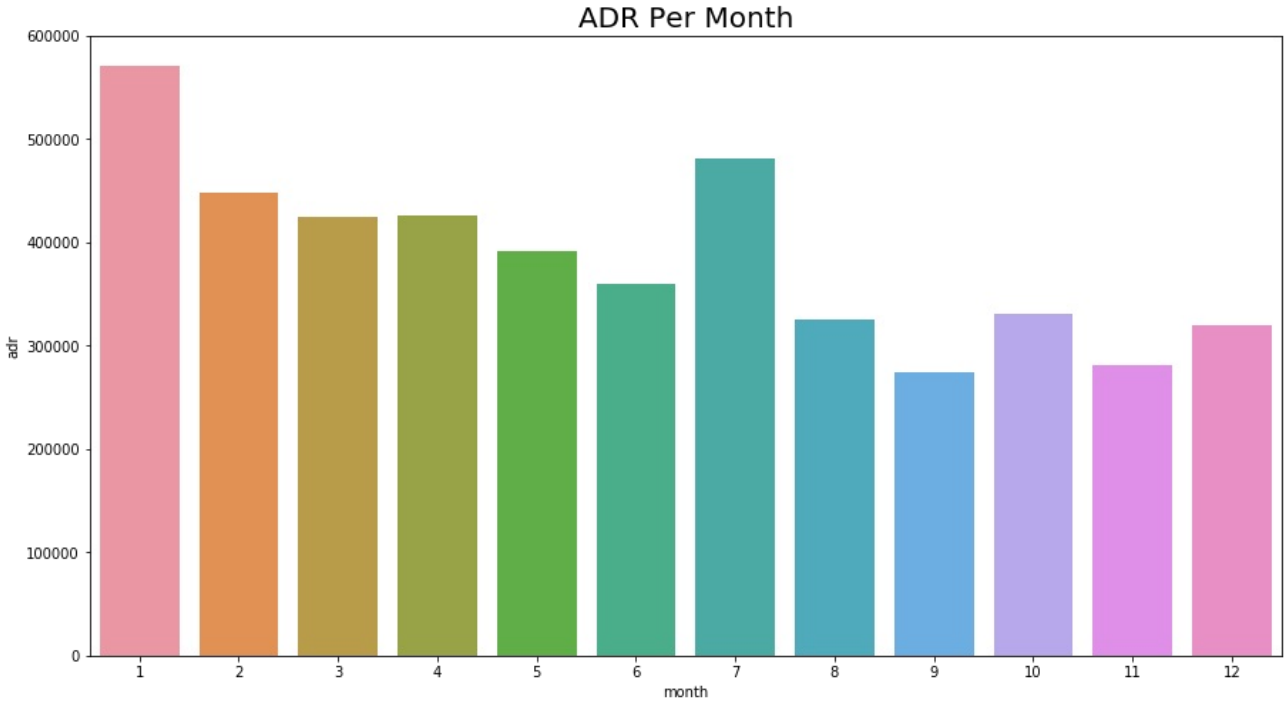
```
df[df['is_canceled']==1].groupby('month')[['adr']].sum()
```

Out[32]:

month	adr
1	571203.57
2	447675.76
3	424872.56
4	425315.23
5	391438.08
6	359003.34
7	481376.17
8	325668.43
9	274176.33
10	330968.87
11	281367.90
12	319272.65

In [33]:

```
plt.figure(figsize=(15,8))
plt.title('ADR Per Month',fontsize = 20)
sns.barplot('month','adr',data=df[df['is_canceled']==1].groupby('month')[['adr']].sum().reset_index())
plt.show()
```



In [34]:

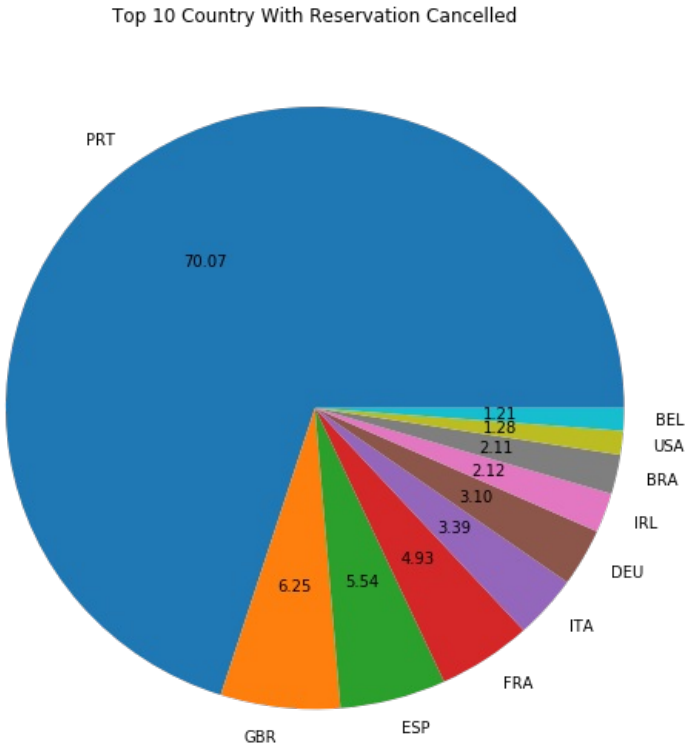
```
cancelled_data = df[df['is_canceled']==1]
top_10_country=cancelled_data['country'].value_counts()[:10]
top_10_country
```

Out[34]:

```
PRT    27514
GBR     2453
ESP     2177
FRA     1934
ITA     1333
DEU     1218
IRL      832
BRA      830
USA      501
BEL      474
Name: country, dtype: int64
```

In [35]:

```
plt.figure(figsize=(9,9))
plt.title('Top 10 Country With Reservation Cancelled')
plt.pie(top_10_country,autopct='%.2f',labels=top_10_country.index)
plt.show()
```



In [36]:

```
df['market_segment'].value_counts()
```

Out[36]:

```
Online TA        56402
Offline TA/TO    24159
Groups           19806
Direct           12448
Corporate         5111
Complementary      734
Aviation           237
Name: market_segment, dtype: int64
```

In [37]:

```
df['market_segment'].value_counts(normalize= True)
```

Out[37]:

```
Online TA        0.474377
Offline TA/TO    0.203193
Groups           0.166581
Direct           0.104696
Corporate        0.042987
Complementary    0.006173
Aviation         0.001993
Name: market_segment, dtype: float64
```

In [38]:

```
cancelled_data['market_segment'].value_counts(normalize= True)
```

Out[38]:

```
Online TA        0.469696
Groups           0.273985
Offline TA/TO    0.187466
Direct           0.043486
Corporate        0.022151
Complementary    0.002038
Aviation         0.001178
Name: market_segment, dtype: float64
```

In [39]:

```
cancelled_df_adr = cancelled_data.groupby('reservation_status_date')[['adr']].mean()
cancelled_df_adr.reset_index(inplace=True)
cancelled_df_adr.sort_values('reservation_status_date',inplace = True)
cancelled_df_adr
```

Out[39]:

	reservation_status_date	adr
0	2014-10-17	62.800000
1	2014-11-18	0.000000
2	2015-01-01	62.062779
3	2015-01-02	9.633750
4	2015-01-18	0.000000
5	2015-01-20	76.500000
6	2015-01-21	37.301209
7	2015-01-22	116.568333
8	2015-01-28	244.000000
9	2015-01-29	66.000000
10	2015-01-30	56.805970
11	2015-02-02	79.740000
12	2015-02-05	155.832500
13	2015-02-06	134.000000
14	2015-02-09	199.000000
15	2015-02-10	165.000000
16	2015-02-11	126.213333
17	2015-02-12	119.700000
18	2015-02-17	76.500000
19	2015-02-19	99.000000
20	2015-02-20	84.947500
21	2015-02-23	85.050000
22	2015-02-24	130.900000
23	2015-02-25	108.175000
24	2015-02-26	8.000000
25	2015-02-27	108.330000
26	2015-03-03	48.969474
27	2015-03-04	47.475000
28	2015-03-05	132.800000
29	2015-03-06	77.175000
...	...	...
872	2017-08-01	189.788696
873	2017-08-02	196.841667
874	2017-08-03	187.509375
875	2017-08-04	143.532647
876	2017-08-05	151.887778
877	2017-08-06	174.150000
878	2017-08-07	173.199412
879	2017-08-08	165.666667
880	2017-08-09	194.426471
881	2017-08-10	134.925278
882	2017-08-11	135.554286
883	2017-08-12	198.000000
884	2017-08-13	192.516667
885	2017-08-14	197.347333
886	2017-08-15	185.053333
887	2017-08-16	163.768667

888	2017-08-17	185.456875
889	2017-08-18	166.126667
890	2017-08-19	169.766667
891	2017-08-20	181.150000
892	2017-08-21	121.100000
893	2017-08-22	153.490000
894	2017-08-23	181.710000
895	2017-08-24	153.953333
896	2017-08-25	228.250000
897	2017-08-26	178.200000
898	2017-08-27	167.300000
899	2017-08-28	81.416667
900	2017-08-29	144.253333
901	2017-08-31	189.750000

902 rows × 2 columns

In [40]:

```
not_cancelled_data = df[df['is_canceled']==0]

not_cancelled_df_adr = not_cancelled_data.groupby('reservation_status_date')[['adr']].mean()
not_cancelled_df_adr.reset_index(inplace = True)
not_cancelled_df_adr.sort_values('reservation_status_date',inplace = True)
not_cancelled_df_adr
```

Out[40]:

	reservation_status_date	adr
0	2015-07-01	31.650000
1	2015-07-02	95.924000
2	2015-07-03	97.645205
3	2015-07-04	110.480000
4	2015-07-05	100.308485
5	2015-07-06	91.540000
6	2015-07-07	97.738824
7	2015-07-08	80.219600
8	2015-07-09	87.349444
9	2015-07-10	86.710741
10	2015-07-11	119.891563
11	2015-07-12	104.993750
12	2015-07-13	94.961471
13	2015-07-14	98.500000
14	2015-07-15	107.928077
15	2015-07-16	117.222045
16	2015-07-17	98.267722
17	2015-07-18	130.875806
18	2015-07-19	114.181959
19	2015-07-20	106.237385
20	2015-07-21	129.542188
21	2015-07-22	84.309697
22	2015-07-23	121.609655
23	2015-07-24	101.218224
24	2015-07-25	136.088889
25	2015-07-26	107.603896
26	2015-07-27	147.684848
27	2015-07-28	105.155000
28	2015-07-29	76.018033
29	2015-07-30	126.337255
...	...	...

775	2017-08-14	163.144051
776	2017-08-15	161.853611
777	2017-08-16	164.287097
778	2017-08-17	169.668257
779	2017-08-18	165.124433
780	2017-08-19	174.151091
781	2017-08-20	168.537565
782	2017-08-21	166.638750
783	2017-08-22	174.188889
784	2017-08-23	147.256023
785	2017-08-24	148.506500
786	2017-08-25	146.346028
787	2017-08-26	172.672165
788	2017-08-27	152.494744
789	2017-08-28	150.247197
790	2017-08-29	149.113494
791	2017-08-30	172.943662
792	2017-08-31	161.504861
793	2017-09-01	128.816324
794	2017-09-02	164.153182
795	2017-09-03	156.190370
796	2017-09-04	138.166216
797	2017-09-05	143.067619
798	2017-09-06	148.826875
799	2017-09-07	137.947368
800	2017-09-08	101.012500
801	2017-09-09	147.778333
802	2017-09-10	145.272500
803	2017-09-12	153.570000
804	2017-09-14	105.930000

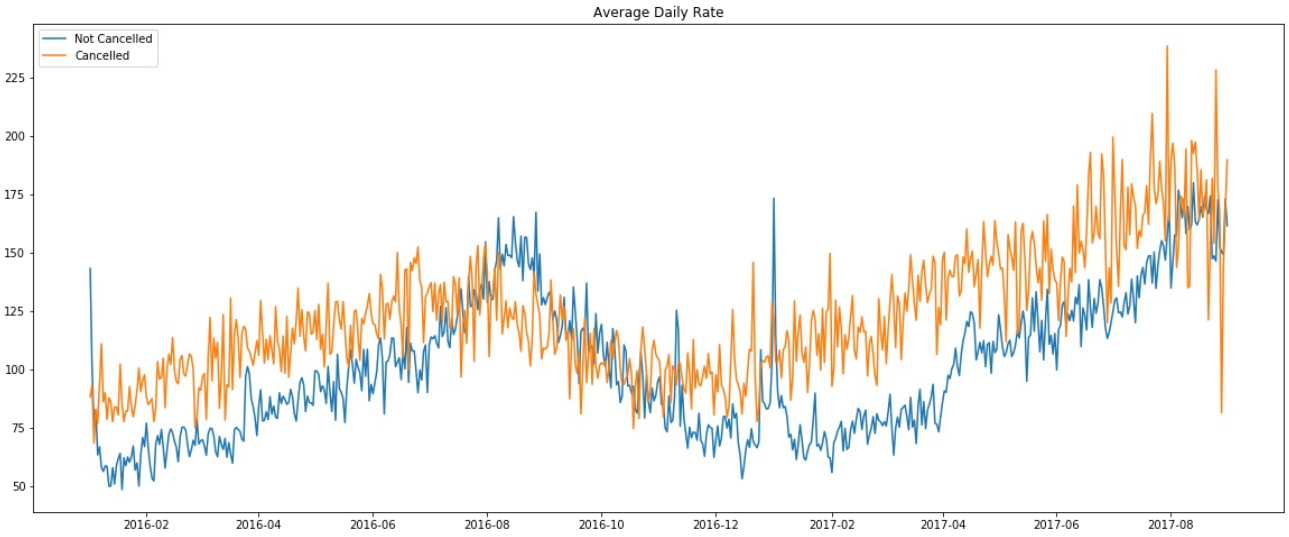
805 rows × 2 columns

In [41]:

```
cancelled_df_adr = cancelled_df_adr[(cancelled_df_adr['reservation_status_date']>'2016') & (cancelled_df_adr['res
ervation_status_date']<'2017-09')]
not_cancelled_df_adr=not_cancelled_df_adr[(not_cancelled_df_adr['reservation_status_date']>'2016') & (not_cancell
ed_df_adr['reservation_status_date']<'2017-09')]
```

In [42]:

```
plt.figure(figsize=(20,8))
plt.title('Average Daily Rate')
plt.plot(not_cancelled_df_adr['reservation_status_date'],not_cancelled_df_adr['adr'],label = 'Not Cancelled')
plt.plot(cancelled_df_adr['reservation_status_date'],cancelled_df_adr['adr'],label = 'Cancelled')
plt.legend()
plt.show()
```



In [ ]:

```

```