In [160…
```
df['adr'].describe()
```

Out[160…
```
count    119389.000000
mean        101.786744
std          48.153554
min          -6.380000
25%          69.290000
50%          94.560000
75%         126.000000
max         510.000000
Name: adr, dtype: float64
```

# DATA ANALYSIS AND VISUALIZATION

In [165…
```
df['is_canceled'].unique()

# 0 is not cancelled and 1 is cancelled (No and Yes)
```

Out[165…
```
array([0, 1])
```

In [322…
```
total_booking = df.groupby('hotel')['is_canceled'].count().reset_index()
total_booking
```

Out[322…

| | hotel | is_canceled |
|---|---|---|
| **0** | City Hotel | 79329 |
| **1** | Resort Hotel | 40060 |

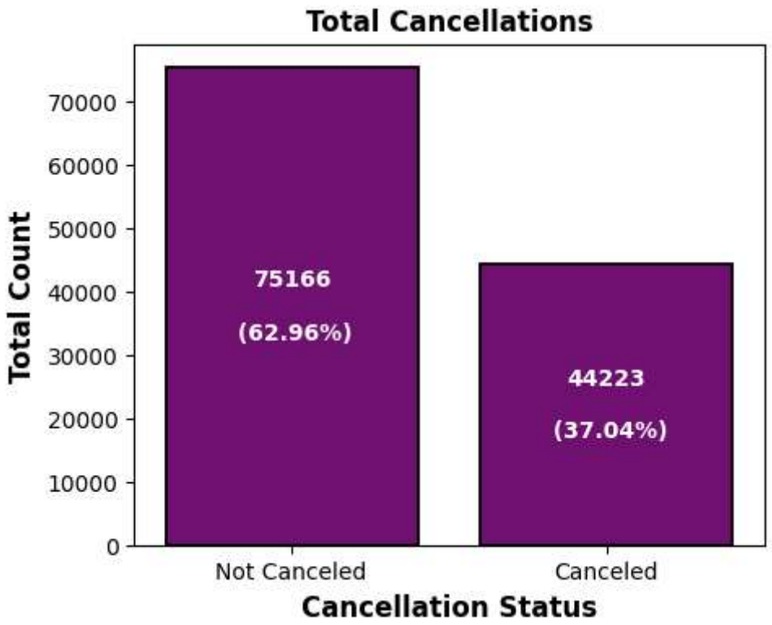
In [178…
```
overall_canc = df['is_canceled'].value_counts().reset_index()
overall_canc
```

Out[178…

| | is_canceled | count |
|---|---|---|
| **0** | 0 | 75166 |
| **1** | 1 | 44223 |

In [302…
```
plt.figure(figsize=(5,4))

ax = sns.barplot(x = 'is_canceled', y = 'count', data=overall_canc, color='purple', linewidth = 1.3, edgecolor='black')
plt.xlabel('Cancellation Status', fontsize=12, fontweight='bold')
plt.ylabel('Total Count', fontsize=12, fontweight='bold')
plt.title("Total Cancellations", fontsize=12, fontweight='bold')

total = overall_canc['count'].sum()

for x in ax.patches:
    height = x.get_height()
    percentage = (height / total) * 100  # Convert to percentage
    plt.text(
        x.get_x() + x.get_width() / 2,  # X-position
        height / 2,  # Y-position
        f'{int(height)}\n\n ({percentage:.2f}%)',  # Display as percentage
        ha='center',
        va='center',
```

```
        color='white',
        fontweight='bold'
    )


plt.xticks(ticks=[0, 1], labels=['Not Canceled', 'Canceled'])

plt.show()
```




**Observation** - Overall, 37.04% of bookings are getting cancelled

In [327…

```
df_grouped = df.groupby('hotel')['is_canceled'].value_counts().to_frame(name='count')
df_grouped['percentage'] = df.groupby('hotel')['is_canceled'].value_counts(normalize=True) * 100
df_grouped
```

Out[327…

| | | count | percentage |
|---|---|---|---|
| **hotel** | **is_canceled** | | |
| **City Hotel** | **0** | 46228 | 58.273771 |
| | **1** | 33101 | 41.726229 |
| **Resort Hotel** | **0** | 28938 | 72.236645 |
| | **1** | 11122 | 27.763355 |

In [311…

```
plt.figure(figsize=(6,5))

ax = sns.countplot(x = 'hotel', hue='is_canceled', data=df, palette='Blues', edgecolor = 'black', linewidth = 1.3)

for x in ax.patches:
    height = x.get_height()
    if height > 0:
        plt.text(
```

```
            x.get_x() + x.get_width() / 2,
            height / 2,
            f'{height:.0f}',
            ha = 'center',
            va = 'center',
            color = 'black',
            fontweight = 'bold')

plt.xlabel('Hotel', fontsize = 12, fontweight = 'bold')
plt.ylabel('Cancellations', fontsize = 12, fontweight = 'bold')
plt.title('Hotel wise Cancellation Rate', fontsize = 12, fontweight = 'bold')

plt.legend(['Not Cancelled', 'Cancelled'])
plt.show()
```


**Observation** - 41.72% of bookings are cancelled from City Hotel and 27.7% are cancelled from Resort Hotel

In [341…

```
df.columns
```

Out[341…
```
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',
       'days_in_waiting_list', 'customer_type', 'adr',
       'required_car_parking_spaces', 'total_of_special_requests',
       'reservation_status', 'reservation_status_date'],
      dtype='object')
```

In [335…
```
df.groupby(['hotel'])['adr'].mean()
```

Out[335…
```
hotel
City Hotel      105.237722
Resort Hotel     94.952930
Name: adr, dtype: float64
```

In [339…
```
resort_hotel = df[df['hotel'] == 'Resort Hotel']
city_hotel = df[df['hotel'] == 'City Hotel']
```

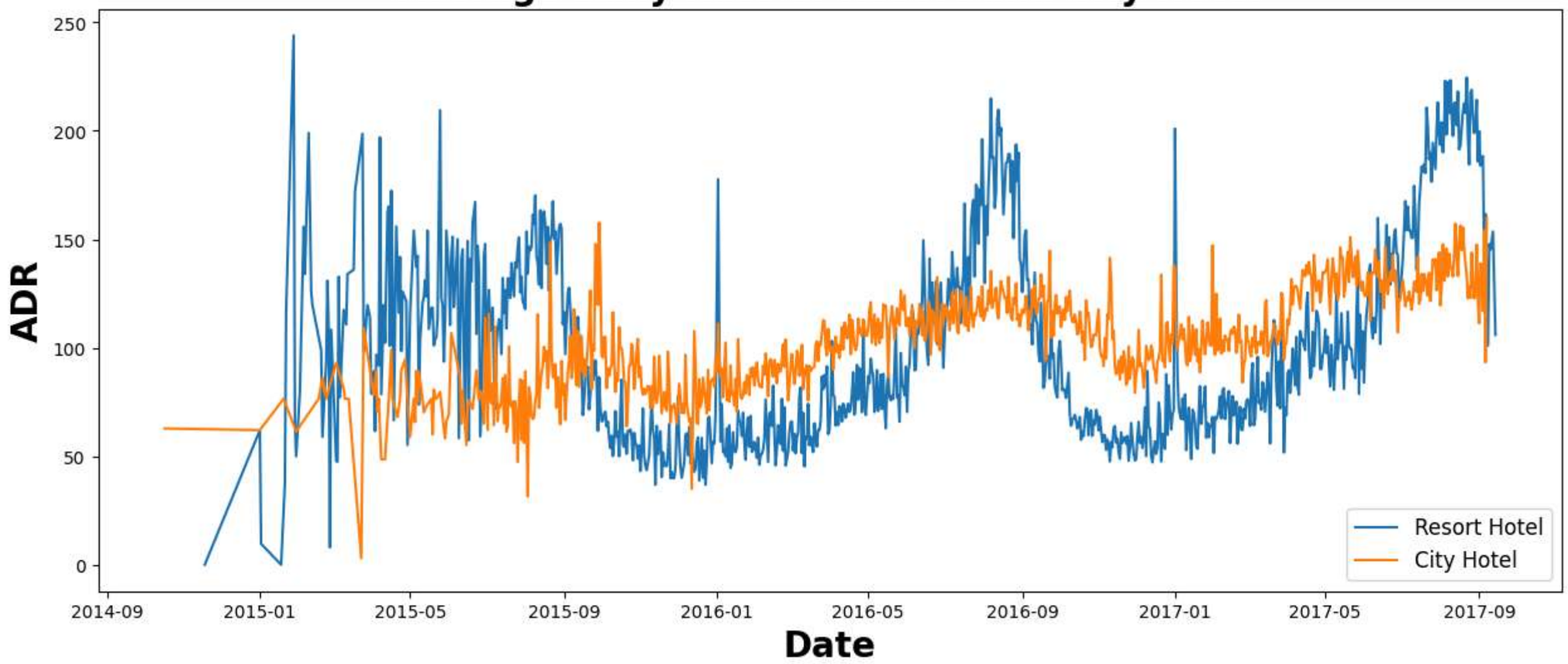
In [347…
```
resort_avg_daily_adr = resort_hotel.groupby('reservation_status_date')['adr'].mean().reset_index()
city_avg_daily_adr = city_hotel.groupby('reservation_status_date')['adr'].mean().reset_index()
```

In [353…
```
resort_avg_daily_adr.head(2)
```

Out[353…

| | reservation_status_date | adr |
|---|---|---|
| **0** | 2014-11-18 | 0.000000 |
| **1** | 2015-01-01 | 61.966667 |

In [371…
```
plt.figure(figsize=(15,6))

plt.plot('reservation_status_date', 'adr', data = resort_avg_daily_adr)
plt.plot('reservation_status_date', 'adr', data = city_avg_daily_adr)

plt.xlabel("Date", fontsize=20, fontweight='bold')
plt.ylabel("ADR", fontsize=20, fontweight='bold')
plt.title("Average Daily Rate of Resort and City Hotel", fontsize=20, fontweight='bold')
plt.legend(['Resort Hotel', 'City Hotel'], fontsize = 12)

plt.show()
```




**Observation** - City Hotel usually has lesser Average Daily Rate compared to Resort Hotel.

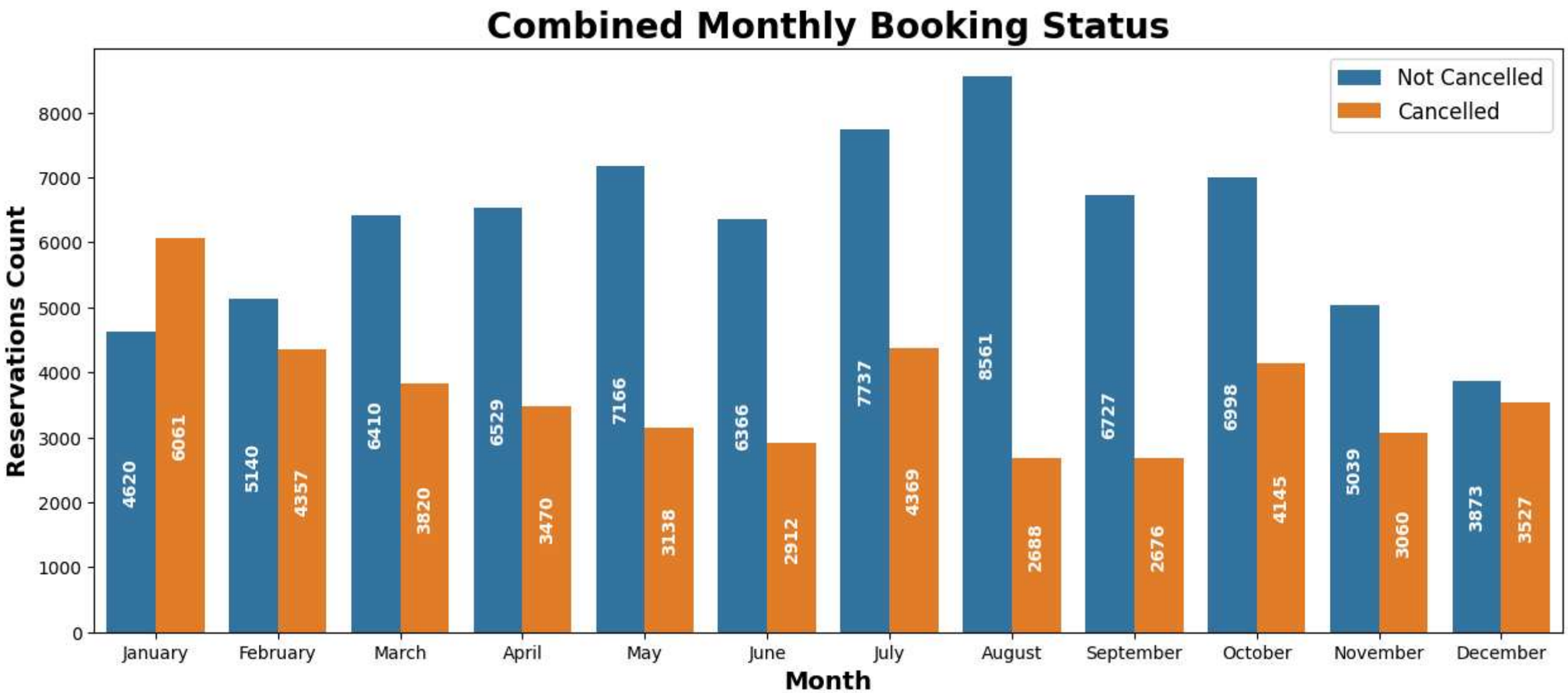
In [375...
```
df['month'] = df['reservation_status_date'].dt.month_name()
df['month'].unique()
```

Out[375...
```
array(['July', 'May', 'April', 'June', 'March', 'August', 'September',
       'January', 'November', 'October', 'December', 'February'],
      dtype=object)
```

In [382...
```
month_order = [
    "January", "February", "March", "April", "May", "June",
    "July", "August", "September", "October", "November", "December"
]

df['month'] = pd.Categorical(df['month'], categories=month_order, ordered=True)
```

In [388...
```
plt.figure(figsize=(15,6))

bx = sns.countplot(x = 'month', hue = 'is_canceled', data=df)

for x in bx.patches:
    height = x.get_height()
    if height > 0:
        plt.text(
```

```
        x.get_x() + x.get_width() / 2,
            height / 2,
            f'{height:.0f}',
            ha = 'center',
            va = 'center',
            color = 'white',
            fontweight = 'bold',
            rotation = 90)

plt.xlabel("Month", fontsize=14, fontweight='bold')
plt.ylabel("Reservations Count", fontsize=14, fontweight='bold')
plt.title("Combined Monthly Booking Status", fontsize=20, fontweight='bold')
plt.legend(['Not Cancelled', 'Cancelled'], fontsize = 12)

plt.show()
```




**Observation** - In January, the cancellations were most whereas August had most bookings with less cancellations.

In [405…

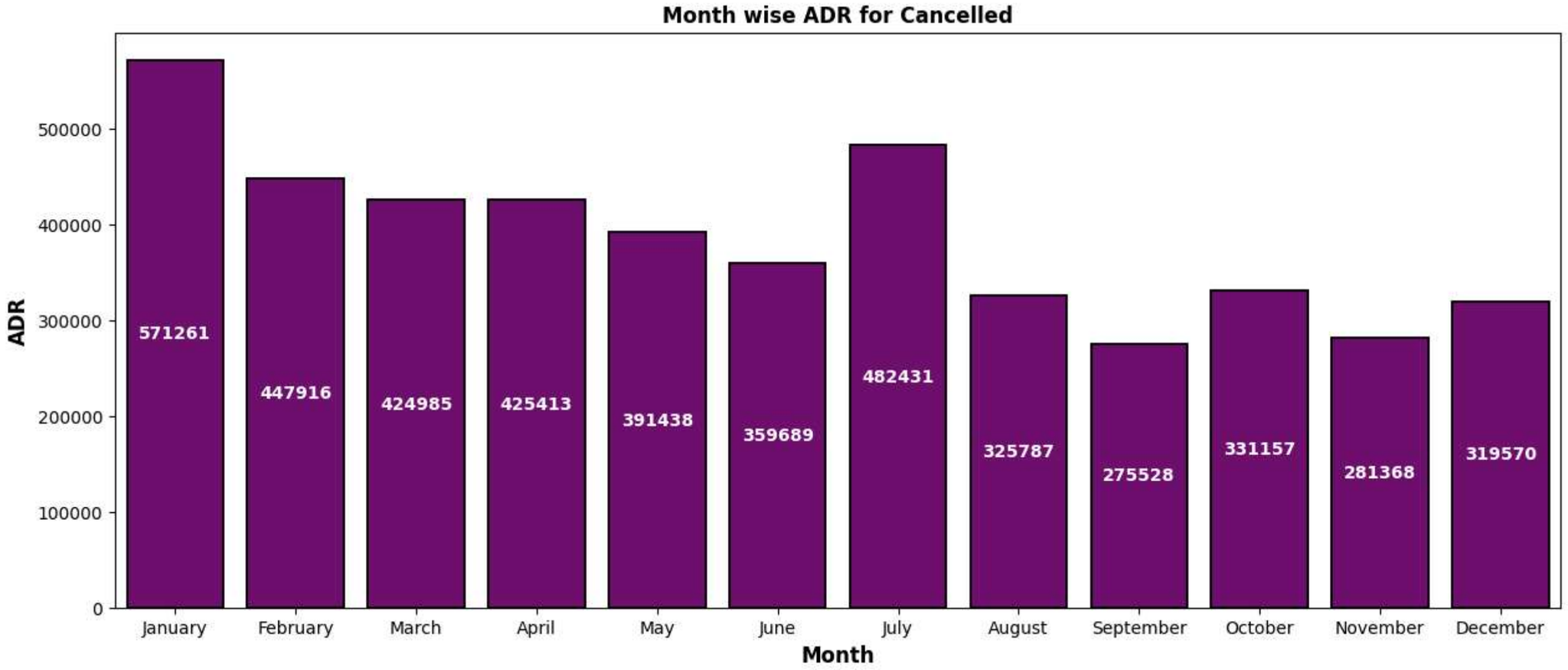
```
monthly_adr_canc = df[df['is_canceled'] == 1].groupby('month')['adr'].sum().reset_index()
monthly_adr_canc
```

Out[405…

| | month | adr |
|---|---|---|
| **0** | January | 571261.17 |
| **1** | February | 447915.76 |
| **2** | March | 424985.36 |
| **3** | April | 425413.23 |
| **4** | May | 391438.08 |
| **5** | June | 359688.89 |
| **6** | July | 482430.62 |
| **7** | August | 325786.93 |
| **8** | September | 275528.33 |
| **9** | October | 331156.87 |
| **10** | November | 281367.90 |
| **11** | December | 319569.53 |

In [409…

```
plt.figure(figsize=(15,6))

bx = sns.barplot(x = 'month', y = 'adr', data=monthly_adr_canc, color='purple', linewidth = 1.3, edgecolor='black')
plt.xlabel('Month', fontsize=12, fontweight='bold')
plt.ylabel('ADR', fontsize=12, fontweight='bold')
plt.title("Month wise ADR for Cancelled", fontsize=12, fontweight='bold')

for x in bx.patches:
    height = x.get_height()
    if height > 0:
        plt.text(
        x.get_x() + x.get_width() / 2,
        height / 2,
        f'{height:.0f}',
        ha='center',
        va='center',
        color='white',
        fontweight='bold'
    )

plt.show()
```




**Observation** - August has very low ADR hence, very few cancellations whereas, January has high ADR hence many cancellations.

In [458…
```
# Top 10 Country with cancellation

cancelled_bookings = df[df['is_canceled'] == 1]
noncancelled_bookings = df[df['is_canceled'] == 0]
```

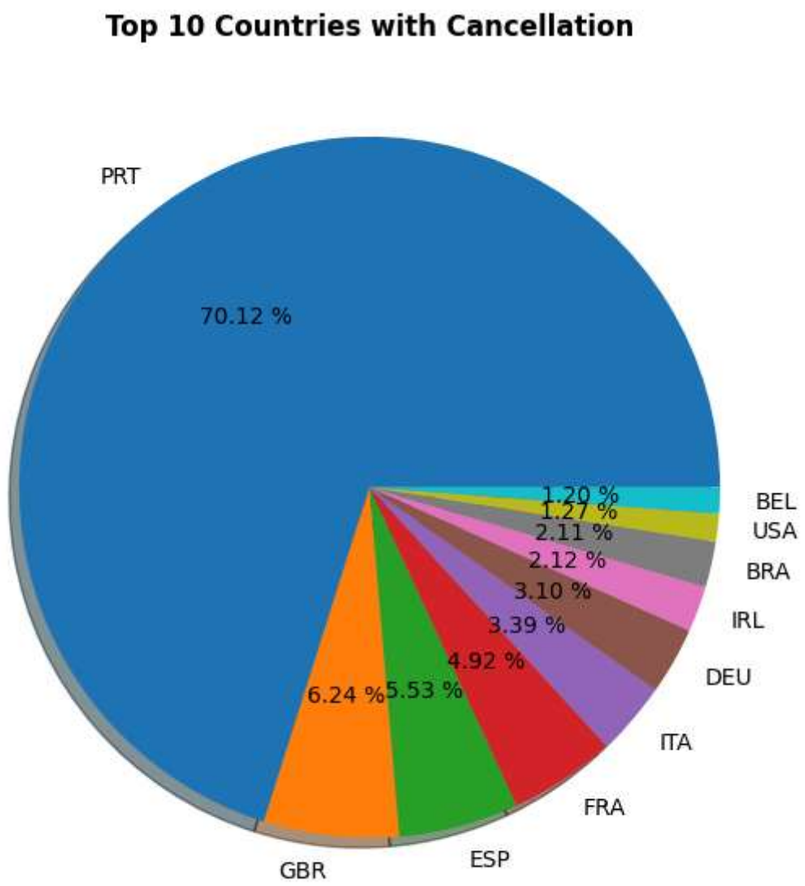
In [432…
```
top_10_country = cancelled_bookings['country'].value_counts()[:10].reset_index()
top_10_country

# [:10] and nlargest(10) is same
```

Out[432…

| | country | count |
|---|---|---|
| **0** | PRT | 27585 |
| **1** | GBR | 2453 |
| **2** | ESP | 2177 |
| **3** | FRA | 1934 |
| **4** | ITA | 1333 |
| **5** | DEU | 1218 |
| **6** | IRL | 832 |
| **7** | BRA | 830 |
| **8** | USA | 501 |
| **9** | BEL | 474 |

In [450…

```
plt.figure(figsize=(10,7))

plt.pie('count', labels = 'country', data = top_10_country, autopct='%1.2f %%', shadow = True)
plt.title("Top 10 Countries with Cancellation", fontsize=12, fontweight='bold')
plt.show()
```




**Observation** - Portugal has the highest cancellation rate

In [461…
```
df['market_segment'].value_counts()
```

Out[461…
```
market_segment
Online TA        56477
Offline TA/TO    24218
Groups           19811
Direct           12606
Corporate         5295
Complementary      743
Aviation           237
Undefined            2
Name: count, dtype: int64
```

In [462…
```
cancelled_bookings['market_segment'].value_counts()
```

Out[462…
```
market_segment
Online TA        20739
Groups           12097
Offline TA/TO     8310
Direct            1934
Corporate          992
Complementary       97
Aviation            52
Undefined            2
Name: count, dtype: int64
```

In [463…
```
noncancelled_bookings['market_segment'].value_counts()
```

Out[463…
```
market_segment
Online TA        35738
Offline TA/TO    15908
Direct           10672
Groups            7714
Corporate         4303
Complementary      646
Aviation           185
Name: count, dtype: int64
```

In [476…
```
canc_plot = cancelled_bookings.groupby('reservation_status_date')['adr'].mean().reset_index().sort_values('reservation_status_date')
noncanc_plot = noncancelled_bookings.groupby('reservation_status_date')['adr'].mean().reset_index().sort_values('reservation_status_date')
```

In [475…
```
canc_plot
```

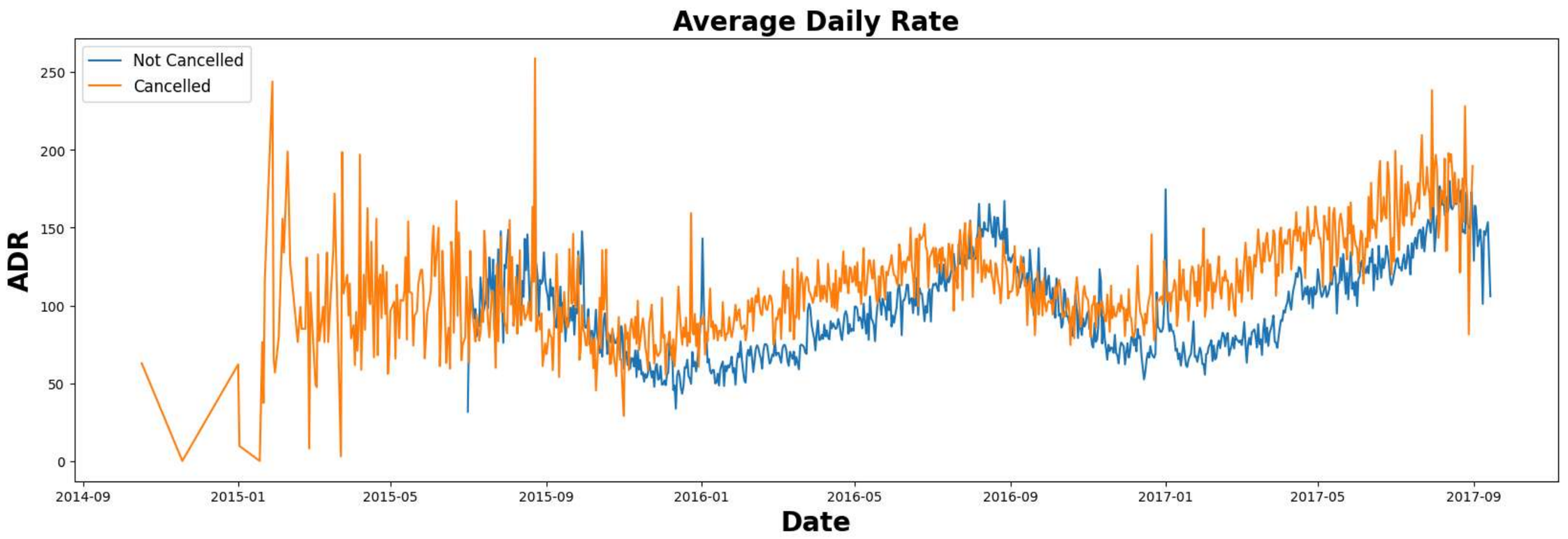
Out[475…

| | 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 |
| **...** | ... | ... |
| **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 [479…
```
plt.figure(figsize=(20,6))

plt.plot('reservation_status_date', 'adr', data = noncanc_plot)
plt.plot('reservation_status_date', 'adr', data = canc_plot)

plt.xlabel("Date", fontsize=20, fontweight='bold')
plt.ylabel("ADR", fontsize=20, fontweight='bold')
```

```
plt.title("Average Daily Rate", fontsize=20, fontweight='bold')
plt.legend(['Not Cancelled', 'Cancelled'], fontsize = 12)

plt.show()
```




In [504…

```
# Since date data is very inconsistent, we filter the date data and choose only from Jan 2016 upto Sept 2017

cancelled_bookings = cancelled_bookings[(cancelled_bookings['reservation_status_date'] > '2016') & (cancelled_bookings['reservation_status_date'] < '2017-09')].sort_values('reservation_status
pd.concat([cancelled_bookings.head(2), cancelled_bookings.tail(2)])
```

Out[504…

| | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **48511** | City Hotel | 1 | 83 | 2016 | March | 13 | 24 | 1 | 3 | 2 | 0.0 | 0 | BB |
| **82578** | City Hotel | 1 | 228 | 2016 | July | 31 | 27 | 0 | 2 | 2 | 0.0 | 0 | BB |
| **13792** | Resort Hotel | 1 | 24 | 2017 | August | 35 | 31 | 0 | 1 | 2 | 0.0 | 0 | HB |
| **13789** | Resort Hotel | 1 | 24 | 2017 | August | 35 | 31 | 0 | 1 | 1 | 0.0 | 0 | HB |

In [505…

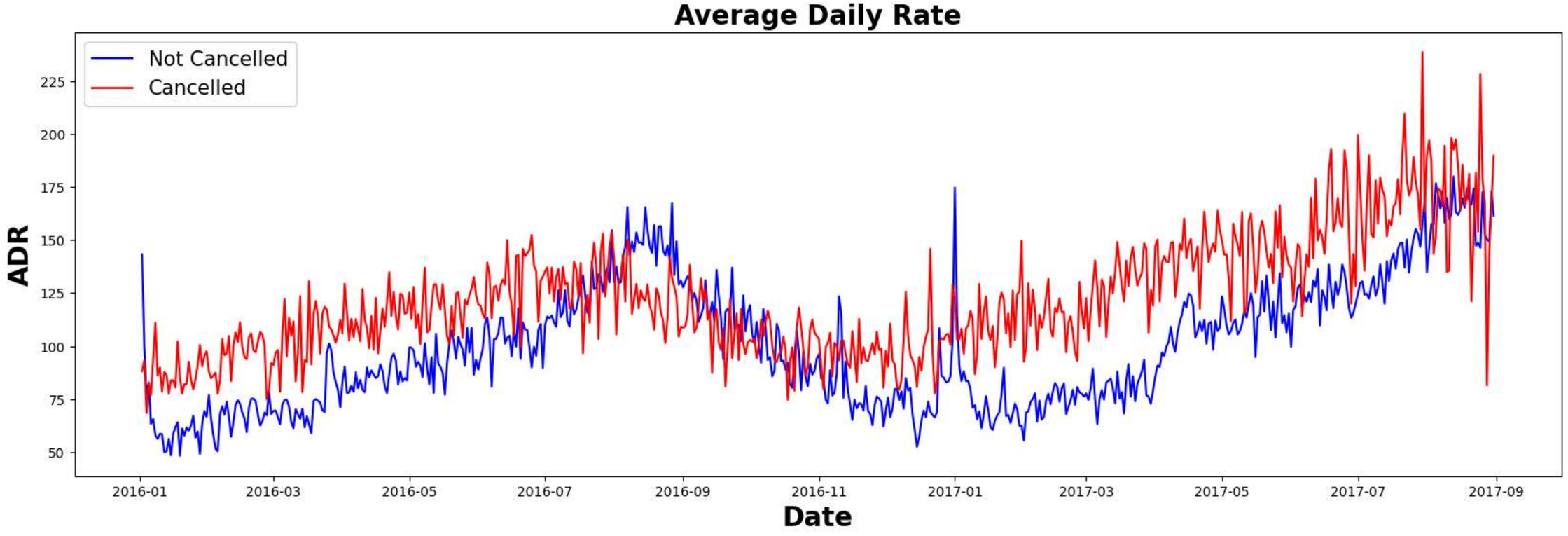
```
noncancelled_bookings = noncancelled_bookings[(noncancelled_bookings['reservation_status_date'] > '2016') & (noncancelled_bookings['reservation_status_date'] < '2017-09')].sort_values('reserv

pd.concat([noncancelled_bookings.head(2), noncancelled_bookings.tail(2)])
```

Out[505…

| | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **46357** | City Hotel | 0 | 160 | 2015 | December | 53 | 30 | 0 | 3 | 2 | 0.0 | 0 | BB |
| **82558** | City Hotel | 0 | 2 | 2015 | December | 53 | 30 | 0 | 3 | 2 | 1.0 | 1 | HB |
| **39865** | Resort Hotel | 0 | 8 | 2017 | August | 35 | 30 | 0 | 1 | 1 | 0.0 | 0 | BB |
| **119177** | City Hotel | 0 | 3 | 2017 | August | 35 | 28 | 1 | 2 | 1 | 0.0 | 0 | SC |

In [517…

```
plt.figure(figsize=(20,6))

sns.lineplot(x = 'reservation_status_date', y = 'adr', data = noncancelled_bookings, color = 'blue', label = 'Not Cancelled', ci=None)
sns.lineplot(x = 'reservation_status_date', y = 'adr', data = cancelled_bookings, color = 'red', label = 'Cancelled', ci = None)

plt.xlabel("Date", fontsize=20, fontweight='bold')
plt.ylabel("ADR", fontsize=20, fontweight='bold')
plt.title("Average Daily Rate", fontsize=20, fontweight='bold')
plt.legend(fontsize = 15)

plt.show()
```




**Observation -** ADR has impact on Cancellation Rates