Project 1 : **Analysing Hotel Booking Cancellation problem**

Steps involved in this project:

1. Understanding the problem statement
2. Gathering the Data sets
3. Identifying the Dataset you require to analyse
4. Explore and clean the data
5. Analysing the cleaned data to get useful insights based on:
6. Problem statement
7. Hypothetical & Research questions
8. Preparing Report
9. Presenting with Dashboards

Step 1: **Understanding the problem statement:**

**Business Problem Statement: In recent years, City Hotels and Resort Hotel have seen high cancellation rates. Each Hotel is now dealing with number of issues as a result, including fewer revenues and less than ideal hotel room use. Consequently, lowering cancellation rates is both Hotels primary goal in order to increase their efficiency in generating revenue, and for us to offer thorough business advice to address this problem.**

**The Analysis of Hotel booking cancellations as well as other factors that have no bearing on their business and yearly revenue generation are the main topics of this report.**

**Assumptions:**

* No unusual occurrences(No outliers) between 2015 and 2017 will have a substantial impact on the data used
* The information is still current and can be used to analyse a hotel’s possible plans in an efficient manner
* There are no unanticipated negatives to the hotel employing any advised technique
* The hotels are not currently using any of the suggested solutions
* The biggest factor affecting the effectiveness of earning income is booking cancellation
* Cancellations result in vacant rooms for the booked length of time
* Clients make hotel reservations the same year they make cancellations

**Understandings:**

1. **Understanding the problem from given statement:** Hotel Businessis facing huge loss due to high cancellation rate from customers, so we need to identify what is the main reason for customer cancelling the bookings?
2. **Generating Research questions:**
3. What are the variables that affects cancellation of hotel reservations?
4. How can we reduce the booking cancellation rate?
5. How will hotels be assisted in making pricing and promotional decisions?
6. **Generating Hypothesis:**
7. More cancellation occurs when price is high
8. When there is a longer waiting list customer tend to cancel the ticket
9. Poor quality service may lead to customer vacate the rooms earlier than reserved days

Step 2**: Gathering the Data sets:**

Collected data from Kagle website : <https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand>

Step 3: **Identifying the Dataset you require to analyse:** Using google colab note book to perform data analysis process.

* Loading all the Libraries
* Importing data

Note: We have imported data after removing customer personal details.

Step 4: **Explore and clean the data :** Using google colab note book to perform data analysis process.

* Performing EDA
* Cleaning the data

Note: In Jupyter note book we are using code df= pd.read\_file extension(csv,xlsx,xlsm,etc) then inside the bracket give the file name including extension of file. If you saved the file within Jupyter note book in the same folder of your notebook then no need to put the file path you can directly mention the file name.

Note: It’s not mandatory to mention df this is the name we are assigning to load the data. So we can give any name we want.

Note: Object data type means categorical columns like Yes/No, true/False, Male/Female etc.

Note: Describe formula applies on numerical data types but if we add include formula inside that and select the Object data type then we can get details about object data type as well

Note: When we apply isnull function to check the missing values and we came to know that there are some columns having very high number of missing values it’s almost 60-90% of total rows(It’s not possible for anyone to handle) , then we can decide will this column required for us to anlyse or is this column not much important for our analysis. If we found that the particular column is of no use then we can remove those columns from our data set.

Note: When we describe the data and we are able to find that there is a huge gap between min and Max value that not going to make any sense it will not gives the accurate report. This is called Outliers. We can either remove them from the data set or keep it like that. But we can’t use them in better way so we can remove the access value rows from that column.

EDA Concepts applied:

1. Df.head() to check the top 5 rows
2. Df.tail() to check the bottom 5 rows
3. Df.shape to check total how many rows and columns exists
4. Df.columns to check the column headers
5. Df.info() used to check the data type and count of null values in particular column (here if count shows less that total number of rows that means there are null values you can calculate by deducting total minus current column count)
6. Df[‘column name’]= Pd.to\_datetime(df[‘column name’]) is used to convert data type of Date column from object to date format
7. Df.describe(include =’object’) it will give some statistics such as count of object data type columns count, unique values in a particular column, top values, frequency,etc. if we didn’t mention include function then describe function gives statistics details for only numeric data type columns there we will get the count, min, max, 25%, 50%,75% data.
8. For col in df.describe (include = ‘object’).columns:

Print(col)

Print(df[col].unique())

Print(‘-‘\*50) this code is for adding line after each paragraph

This formula is used to get the names of unique values in each object columns

1. Df.isnull().sum() to check the number of missing values in each column. If there are more than 0.01% to 1% of data is missing out of 1Lakh rows then we can remove those empty rows but if more than 1% of data is missing in particular column then it’s difficult to clean them with values so we can remove those columns from data set
2. Df.drop([‘column name’], axis=1, inplace=True) to delete the columns from data set. Here axis=1 means selecting columns and inplace=True is to apply changes in data frame.
3. Df.dropna(inplace=True) this function is used to remove the null values in columns if particular column having upto 1% null rows of 1lakh total rows then we can use this to delete the null value rows
4. Df= df[df[‘column name’] mention the value till how much you want to show in your data set] this function is used to maintain the data set within the range where we can use them efficiently to plot the values.

Step 5: **Analysing the cleaned data to get useful insights:**

* Analysing the data in python(google colab)

1. Calculating the percentage of cancellation because this is the major issue which hotel is facing currently.

Formula to check: variable name(cancelled\_Per) = data frame name(df) [“column name”].value\_counts(normalize=True)

**Note: Value\_counts** gives the category name and count of how many times it repeated in the column. If we add normalize= True inside the value\_counts it will give the percentage of that categorise.

* Building Visualization

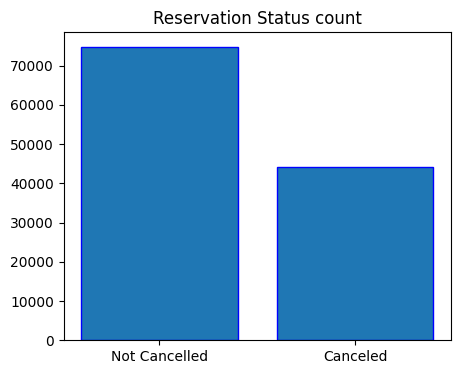
Using the above count value we are applying matplot lib to present our data visually. Here is the formula

plt.figure(figsize=(5,4))

plt.title('Reservation Status count')

plt.bar(['Not Cancelled','Canceled'],df['is\_canceled'].value\_counts(), edgecolor='b', width = 0.8) #here we used the above formula to get the count. One changes we did is we are not using normalize=True because we want the exact count not the percentage.

plt.show()



1. **Calculating the total number of Reservation by Hotel wise, to identify which hotel got more reservation: We are representing this via visualization using Matplotlib and seaborn. It will show how much reservation happened in each hotel and in that reservation how much cancelled.**

**Code:**

plt.figure(figsize=(7,4))

ax1= sns.countplot(x ='hotel', hue ='is\_canceled', data = df, palette = 'Blues')

legend\_lables,\_=ax1. get\_legend\_handles\_labels()

ax1.legend(bbox\_to\_anchor=(1,1))

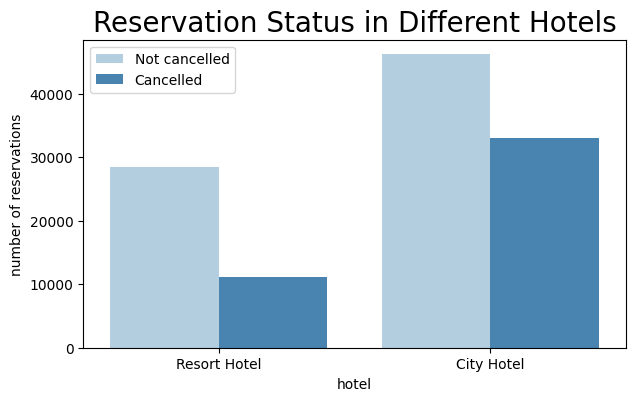
plt.title('Reservation Status in Different Hotels', size=20)

plt.xlabel('hotel')

plt.ylabel('number of reservations')

plt.legend(['Not cancelled', 'Cancelled'])

plt.show()

****

Above graph shows us that City hotels getting more Bookings compare to Resort Hotels, but at the same time City hotel bookings are getting cancelled more than Resort Hotels. So we need to identify the exact reason why city Hotel Bookings are getting cancelled and also need to identify why Resort Hotels are getting low reservation bookings.

1. **We are checking the Percentage of Hotel boking and cancellation by different Hotels:**

**Code:**

resort\_hotel = df[df['hotel'] == 'Resort Hotel']

resort\_hotel['is\_canceled'].value\_counts(normalize=True)\*100 which gives us

Not cancelled =72.024952,

Cancelled = 27.975048

City\_hotel = df[df['hotel'] == 'City Hotel']

City\_hotel['is\_canceled'].value\_counts(normalize= True)\*100 which gives us

Not cancelled = 58.291825,

Cancelled = 41.708175

Note: From the Above Counts we can clearly understand that City Hotels Bookings are only 58% not getting cancelled comparing to City Hotel's not cancelling rate of 72%

1. **Now we are checking resort wise daily Avg**

**booking price: This will helps us to get to know the Avg price charging by each hotels on a daily basis. So that we can identify whether variation in price will effect the cancellation?**

**Code:**

resort\_hotel = resort\_hotel.groupby('reservation\_status\_date')[['adr']].mean()

City\_hotel = City\_hotel.groupby('reservation\_status\_date')[['adr']].mean()

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

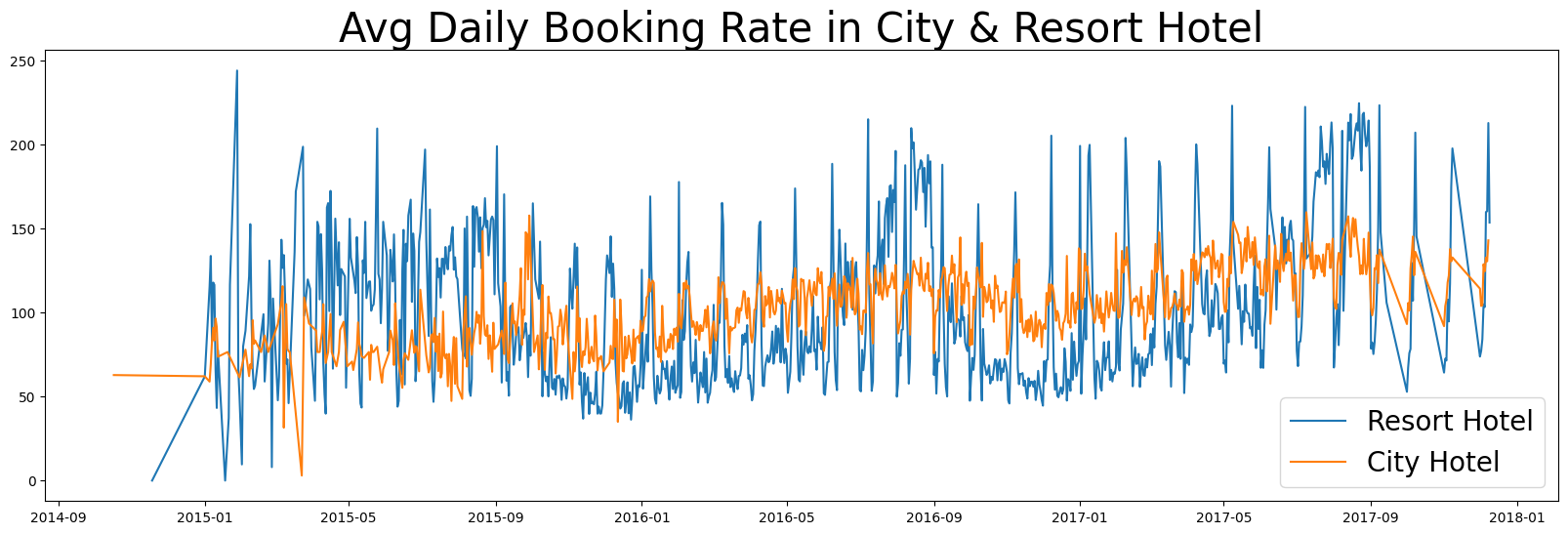
plt.title('Avg Daily Booking Rate in City & Resort Hotel', size= 30)

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()

****

**Note:** Above graph proves that avg price for booking in City hotel is much lesser than resort Hotel for most of the days. May be during Weekends and festival season price may gone up as we assuming

1. **We are now checking in which month highest reservation happens and which month Highest cancelation happens.**

**Code:**

df['month']= df['reservation\_status\_date'].dt.month

plt.figure(figsize=(16,8))

ax1 = sns.countplot(x = 'month', hue = 'is\_canceled', data = df, palette = 'bright')

legend\_lables,\_ =  ax1. get\_legend\_handles\_labels()

ax1.legend(bbox\_to\_anchor=(1,1))

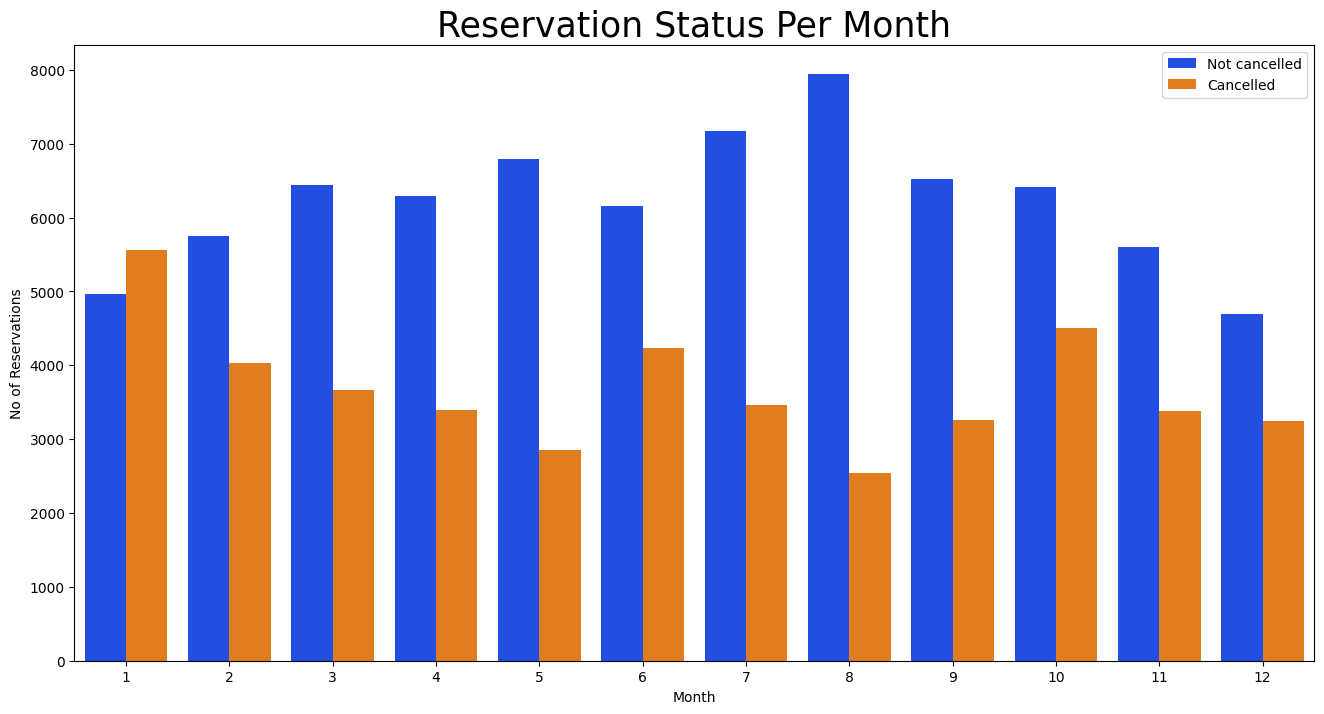
plt.title('Reservation Status Per Month', size=25)

plt.xlabel('Month')

plt.ylabel('No of Reservations')

plt.legend(['Not cancelled', 'Cancelled'])

plt.show()

****

**Note: In the above graph we can understand that in the month of August there is a high reservation happened and in the month of Jan and Dec Lowest reservation happened. Same way In January highest number of cancellation happened and in August lowest cancellation happened. It shows that During high peak time cancellation is less but during Low peak reservation is less and cancellation is more.**

1. **To understand the above graph even better we are now calculating the avg price per month so that we can understand is that price is the reason to increase and decrease the reservation and cancellation?**

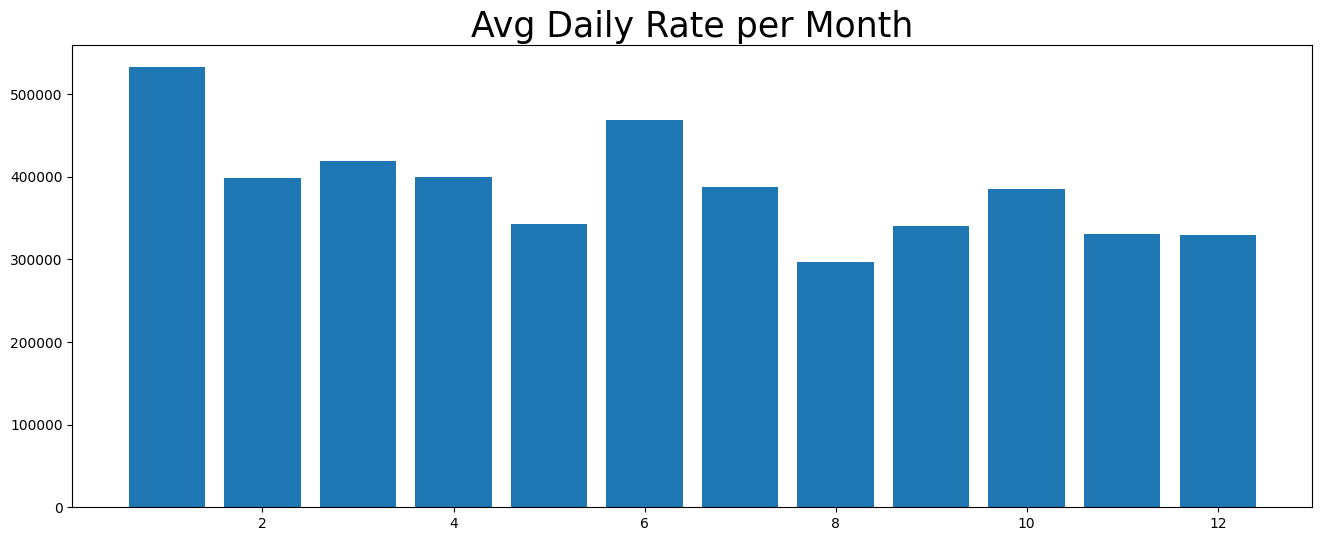
**Code:**

plt.figure(figsize=(16,6))

plt.title('Avg Daily Rate per Month', fontsize=25)

plt.bar('month','adr', data = df[df['is\_canceled'] == 1].groupby('month')[['adr']].sum().reset\_index())

plt.show()

****

**Note: If we check the above graph it’s clearly shows that During the month of Jan and June price of reservation is High so that most of the cancellation happened in these month and also less reservations happened in those months. It’s clearly states that when hike in the booking price reservation goes down and cancellation goes up and vice versa.**

1. **Now we are checking country wise cancellation rate to understand which country making highest booking and cancellation, so that we can understand which country we need to focus more to attract theme even better.**

**Note: There are around 177 countries,we can't analyse on each and every country. So we are going with Top 10 countries on the basis of Highest bookings cancelled.**

**Code:**

cancelled\_data = df[df['is\_canceled']==1]

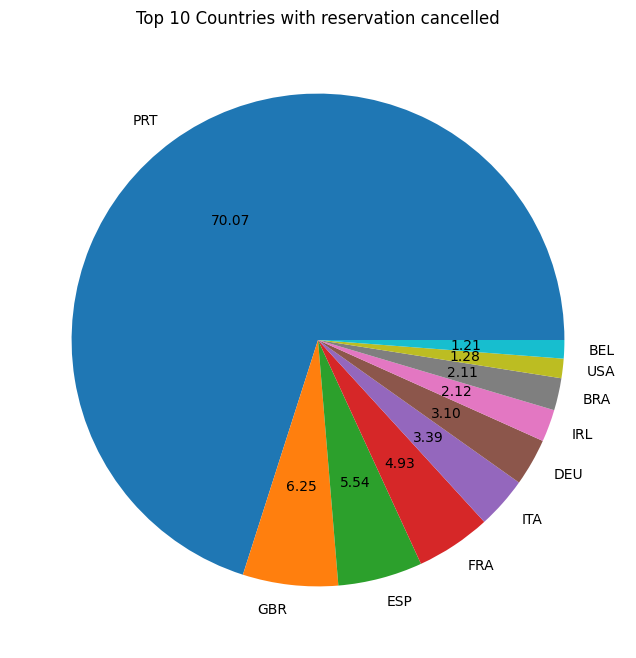
top\_10\_country = cancelled\_data['country'].value\_counts()[:10]

plt.figure(figsize=(8,8))

plt.title('Top 10 Countries with reservation cancelled')

plt.pie(top\_10\_country, autopct = '%.2f', labels= top\_10\_country.index)

plt.show()

****

**Note: Above pie chart shows that Portugal country is Highest among other countries on Cancellation of Bookings. So we can suggest the Hotels to focus more on Portugal country hotels on improvising the quality of service, properly fixing the pricing, adding different dishes as per countries requirement, arranging some events where public can enjoy with fun. Running more advertises, promotional discount offers etc.**

1. **Now we need to identify from where customers coming either through online portal or through offline travel agencies. So we can understand is the online booking takes lot of time or offline booking takes lot of time.**

**Code:**

df['market\_segment'].value\_counts()

**It gives the result as**

Online TA - 56402

Offline TA/TO - 24159

Groups - 19806

Direct - 12448

Corporate - 5111

Complementary - 734

Aviation - 237

Note: Our Hypothesis was that people coming highest from Offline travel agencies but it’s went wrong. While calculating percentage of booking segment we got that 47% of booking is happening through Online Travel Agents.

Online TA 47.437698

Offline TA/TO 20.319268

Groups 16.658116

Direct 10.469566

Corporate 4.298679

Complementary 0.617341

Aviation 0.199332

If we check the Cancelled rate from booking segment that count is as below:

Code: cancelled\_data['market\_segment'].value\_counts(normalize=True)\*100

Result:

Online TA 46.969560

Groups 27.398532

Offline TA/TO 18.746603

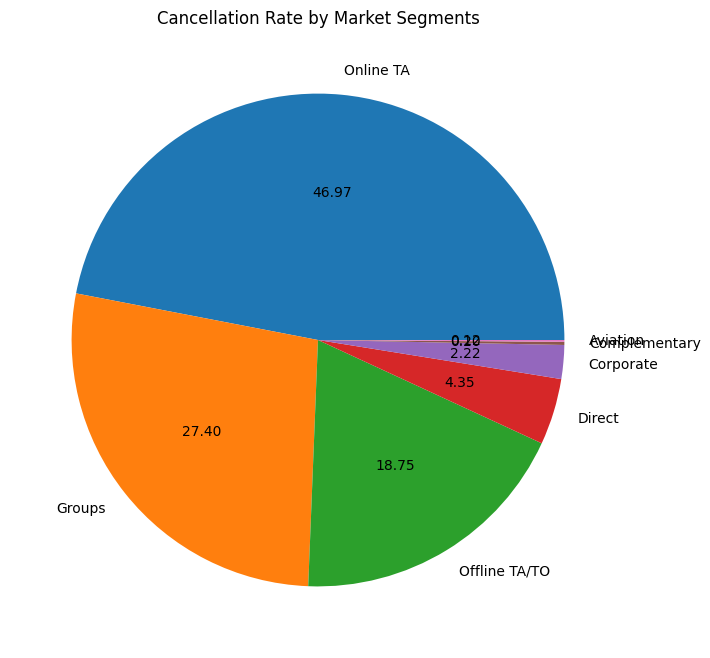
Direct 4.348614

Corporate 2.215075

Complementary 0.203841

Aviation 0.117775

Note: In cancellation also Online bookings are getting cancelled more than other segments.



1. **Now we are checking the Avg Daily rate price is high in cancelled tickets or non cancelled tickets?**

**Code:**

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)

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)

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

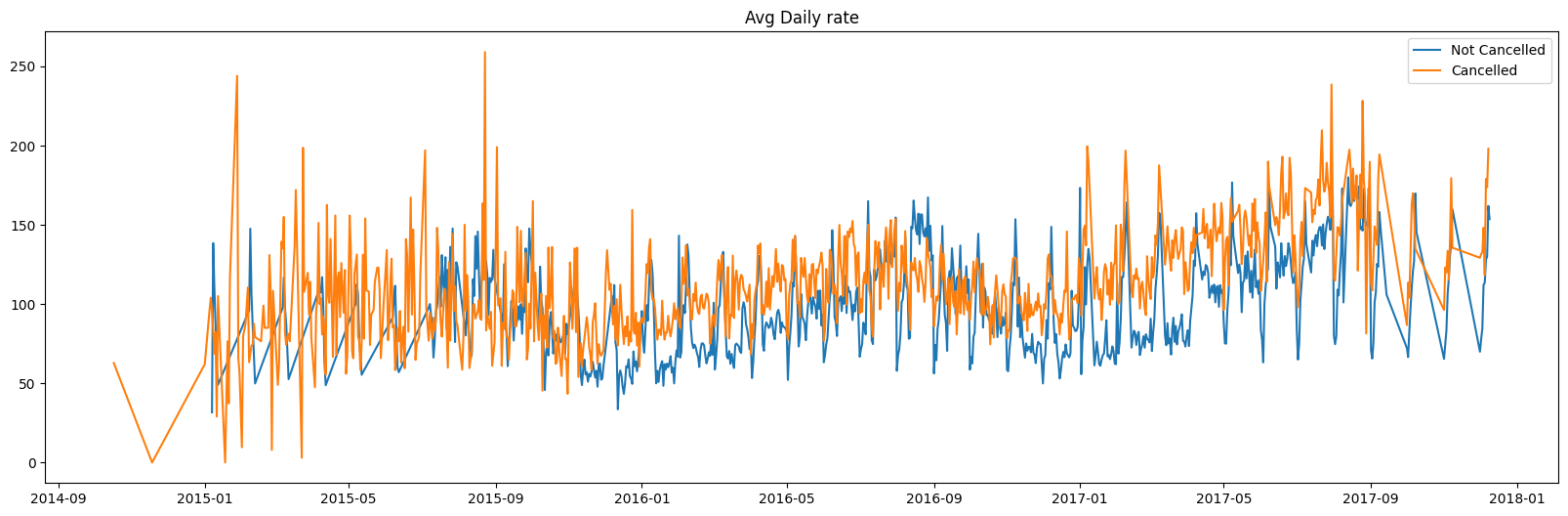
plt.title('Avg Daily rate')

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()

****

**Note: Here we got the messy data and in 2015 & 2017 there are less data so we are filtering only 2016 data using below code.**

cancelled\_df\_adr = cancelled\_df\_adr[(cancelled\_df\_adr['reservation\_status\_date']>'2016') & (cancelled\_df\_adr['reservation\_status\_date']<'2017-09')]

not\_cancelled\_df\_adr = not\_cancelled\_df\_adr[(not\_cancelled\_df\_adr['reservation\_status\_date']>'2016') & (not\_cancelled\_df\_adr['reservation\_status\_date']<'2017-09')]

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

plt.title('Avg 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()

**Chart, line chart

Description automatically generated**

**Note: If we see the above line graph we can clearly consider that whenever price gone up tickets got cancelled and vice versa. So main reason for cancellation is pricing of bookings.**

**Step 6: Preparing Report :**

**We already prepared a report with detailed info about each and every problems and hypothesis we performed. I will present it in Different doc with proper structure.**

**Final Suggestions to the Hotel Owners:**

1. Cancellation rates rises when the price goes up. In order to prevent cancellation of reservations, hotels could work on their pricing strategies and try to lower the price for specific hotels in specific Countries where cancellation rate is high(refer Top 10 countries). Adding Some discounts will helps to attract many customers.
2. As the Reservation rate is High and Cancellation rate is also high in City Hotel compare to Resort Hotel, so they needs to provide some discounts and Special offers to Couples on Weekend or Holidays that can help them to get more not cancelled customers.
3. In the Month of January Hotels getting highest number of Cancellation, so they need to perform some strategic marketing campaign with some discounts and offers. This will help them to hold their customers.
4. We saw that in Portugal highest number of cancellation is happening, so to reduce that Hotels should focus on providing high quality services and adding new items in their food menu according to customer preferences and countries culture. They need to conduct some events to attract more customers.