# **Tasks**

**Initial Tasks**:

1. Data Cleaning: Handling missing values, correcting data types, and removing duplicates.
2. Descriptive Statistics: Generating summary statistics for numerical and categorical variables.
3. Data Visualization: Creating visualizations to understand the distribution of variables and relationships between them.
4. Time Series Analysis: Analyzing trends and patterns over time.
5. Segmentation Analysis: Identifying and analyzing different customer segments.
6. Booking and Cancellation Analysis: Examining booking attributes and their impact on cancellations and ADR.
7. Actionable Insights: Providing recommendations based on the analysis to improve hotel operations and customer satisfaction.

**Objective Questions**:

1. What is the average lead time for bookings that are canceled versus those that are not canceled?

Solution:

Code:

1. cancelled\_lead\_time = booking[booking['is\_canceled']==1]['lead\_time'].mean()
2. not\_cancelled\_lead\_time = booking[booking['is\_canceled']==0]['lead\_time'].mean()
3. print("The average Lead Time for Canceled Bookings were: " ,cancelled\_lead\_time, " and for Non Cancelled Bookings are ", not\_cancelled\_lead\_time)
4. alt = booking.groupby('is\_canceled')['lead\_time'].mean()
5. print(alt)
6. sns.barplot(data=alt)

Output:

The average Lead Time for Canceled Bookings were: 105.71925078043705 and for Non

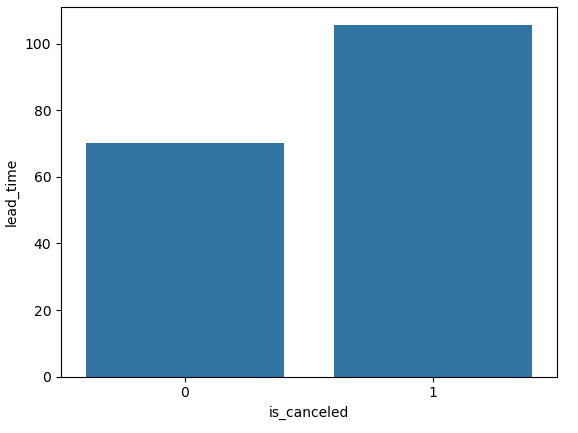
Cancelled Bookings are 70.0995881396853

is\_canceled

0 70.099588

1 105.719251

Name: lead\_time, dtype: float64



Explanation & Insights:

The average number of Cancelled Number of Booking are 105 while the Non Cancelled average number of Bookings are 70

Therefore we can conclude there are a higher number of booking cancellations on average as compared to those which are not

1. Which month has the highest number of bookings?

Solution:

Code:

1. hnb=booking['arrival\_date\_month'].value\_counts()
2. hnb.idxmax(axis=0)

Output:

'August'

Explanation & Insights:

To get the Month wise highest number of bookings we need to find count month wise, then take the 1st entry

Month August seems to have the most number of total bookings over the years.

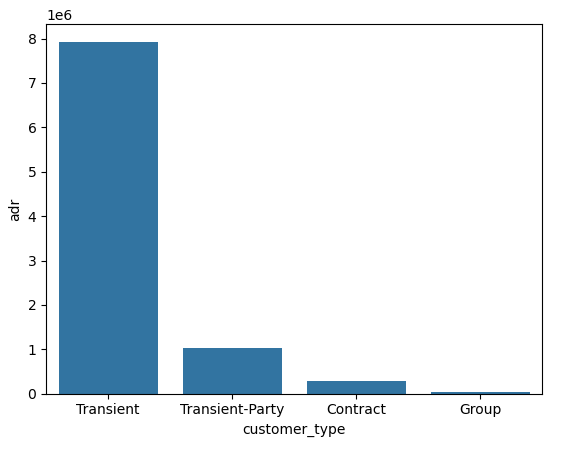
1. What is the distribution of ADR (Average Daily Rate) across different customer types?

Solution:

Code:

1. adrc=booking[['customer\_type','adr']]
2. adrc1=pd.DataFrame(adrc.groupby('customer\_type')['adr'].sum().reset\_index())
3. adrc1 =adrc1.sort\_values(by='adr',ascending=False)
4. sns.barplot(data=adrc1, x='customer\_type', y='adr')
5. plt.show()

Output:



Explanation & Insights:

To get the above output, We need to find the sum of ‘adr’ w.r.t ‘customer\_type’.

Transient has the highest ADR among all ‘customer\_types’

1. How many bookings had special requests, and what is the average number of special requests per booking?

Solution:

Code:

1. sr=booking[booking['total\_of\_special\_requests']>0]['total\_of\_special\_requests'].value\_counts()
2. sr.sum()
3. sr2=booking[booking['total\_of\_special\_requests']>0]['total\_of\_special\_requests'].mean()
4. sr2.round(1)

Output:

1. np.int64(43502)
2. np.float64(1.4)

Explanation & Insights:

* + - 1. We need to find the count of bookings where special request is not ‘0’
      2. We need to find the mean of special request where special request is not ‘0’

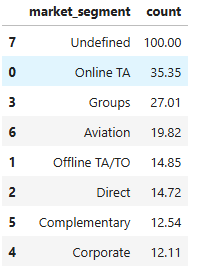
1. What is the cancellation rate for each market segment?

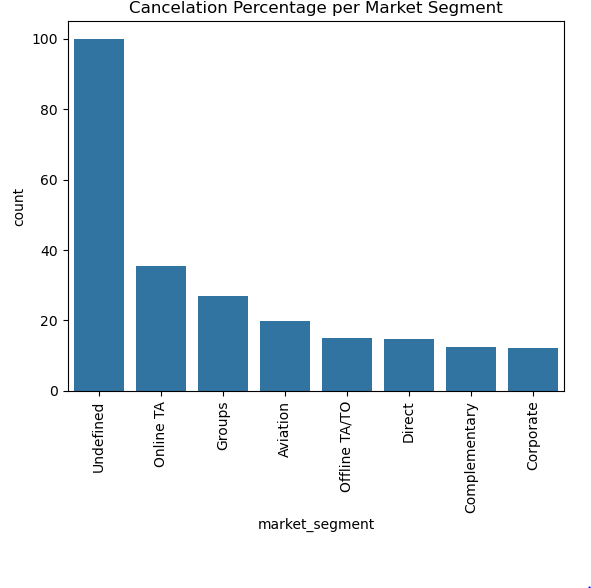
Solution:

Code:

* 1. cancellation = booking[booking['is\_canceled']==1]['market\_segment'].value\_counts()
  2. total = booking['market\_segment'].value\_counts()
  3. non\_cancellation = booking[booking['is\_canceled']==0]['market\_segment'].value\_counts()
  4. cancellation\_rate=pd.DataFrame((cancellation/total)\*100)
  5. cancellation\_rate['count']=cancellation\_rate['count'].round(2)
  6. cancellation\_rate=cancellation\_rate.reset\_index().sort\_values(by='count',ascending=False)
  7. cancellation\_rate
  8. sns.barplot(data=cancellation\_rate, x= 'market\_segment', y='count')
  9. plt.title('Cancelation Percentage per Market Segment')
  10. plt.xticks(rotation=90)
  11. plt.show()

Output:





Explanation & Insights:

We will first find the number of bookings that were cancelled wrt market segment, second find the total number of bookings. Next we need to find the mean the cancelled/total wrt market segment

Undefined market section has the highest cancellation rate in market segment, while corporate and complementary have the lowest mean.

1. Which distribution channel has the highest average ADR?

Solution:

Code:

* 1. dcadr= booking[['distribution\_channel','adr']]
  2. ADR = pd.DataFrame(dcadr.groupby('distribution\_channel')['adr'].mean())
  3. avg\_ADR = ADR.sort\_values(by='adr',ascending=False)
  4. highest\_avg\_ADR = avg\_ADR.head(1)
  5. highest\_avg\_ADR

Output:



Explanation & Insights:

We will find the average of ‘adr’ w.r.t ‘distribution channel’ and consider the highest entry.

GDS has the highest average adr.

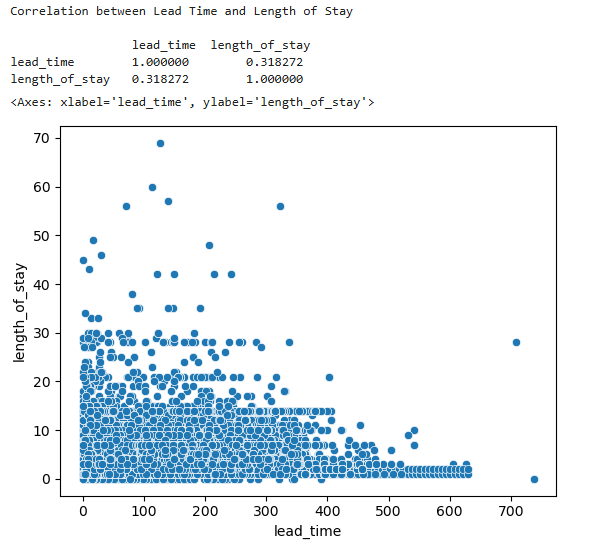
1. What is the correlation between lead time and the length of stay?

Solution:

Code:

1. booking['length\_of\_stay']=booking['stays\_in\_weekend\_nights']+booking['stays\_in\_week\_nights']
2. corre = booking[['lead\_time','length\_of\_stay']]
3. print("Correlation between Lead Time and Length of Stay\n\n",corre.corr(method='pearson'))
4. data=pd.DataFrame(corre.groupby('length\_of\_stay')['lead\_time'].mean().reset\_index())
5. data=data.sort\_values(by='length\_of\_stay')
6. # Plot to see the relationship between Lead Time and Length of Stay
7. sns.scatterplot(data=corre, x= 'lead\_time', y= 'length\_of\_stay')

Output:



Explanation & Insights:

The average lead time of the length of stay between 0-30 lie mainly between 0-400

However it can be seen that

* The length of stay between 0-10 lies between the avg lead time 0-600
* The length of stay between 10-20 lies between the avg lead time 0-450
* The length of stay between 20-30 lies between the avg lead time 0-300
* The length of stay >30 lies between the avg lead time 0-200

1. How many bookings include children and/or babies?

Solution:

Code:

1. next = booking[(booking['children']>0) & (booking['babies']>=0)]
2. child\_baby\_bookings = next[['children','babies']].value\_counts()
3. child\_baby\_bookings
4. child\_baby\_bookings.sum()

Output:

children babies

1 0 4552

2 0 3561

1 1 140

3 0 75

2 1 32

1 2 3

10 0 1

Name: count, dtype: int64

np.int64(8364)

Explanation & Insights:

We calculate the count of children and/or babies, then find the total sum of the count.

The total number of bookings having children and/or babies is 8,364

Also note the no. of bookings having only children are relatively higher as compared to those with babies included as well.

1. What is the average lead time for repeated guests compared to new guests?

Solution:

Code:

* 1. repeat\_vs\_new=booking.groupby('is\_repeated\_guest')['lead\_time'].mean()
  2. repeat\_vs\_new
  3. sns.barplot(data=repeat\_vs\_new,color='green')
  4. repeat\_vs\_new=pd.DataFrame(repeat\_vs\_new.reset\_index())
  5. repeat\_vs\_new['is\_repeated\_guest']=repeat\_vs\_new['is\_repeated\_guest'].replace(0,'new\_guests').replace(1,'repeated\_guests')
  6. print(repeat\_vs\_new)
  7. plt.pie(data=repeat\_vs\_new, x='lead\_time', labels='is\_repeated\_guest',autopct='%.1f%%')
  8. plt.show()

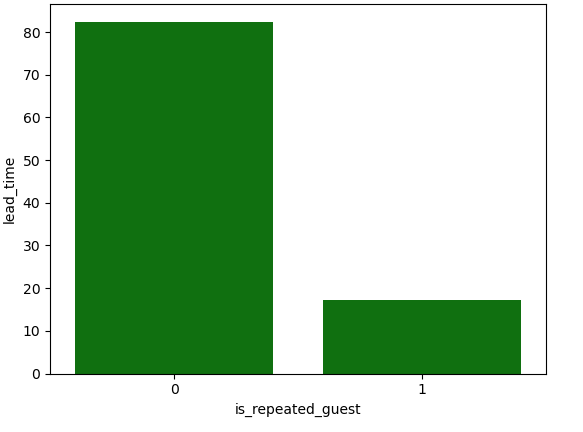
Output:

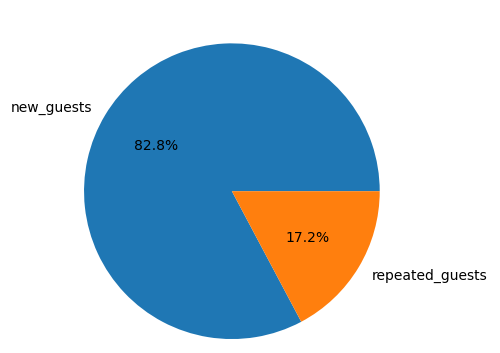
is\_repeated\_guest

0 82.442255

1 17.160469

Name: lead\_time, dtype: float64





Explanation & Insights:

We will calculate the avg. ‘lead time’ w.r.t ‘is repeated guest’ column

The average lead time for new guests (80 days) is higher than the average lead time for repeated guests (20 days)

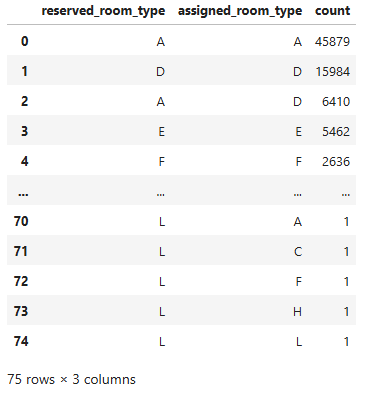
1. What is the most common combination of reserved room type and assigned room type?

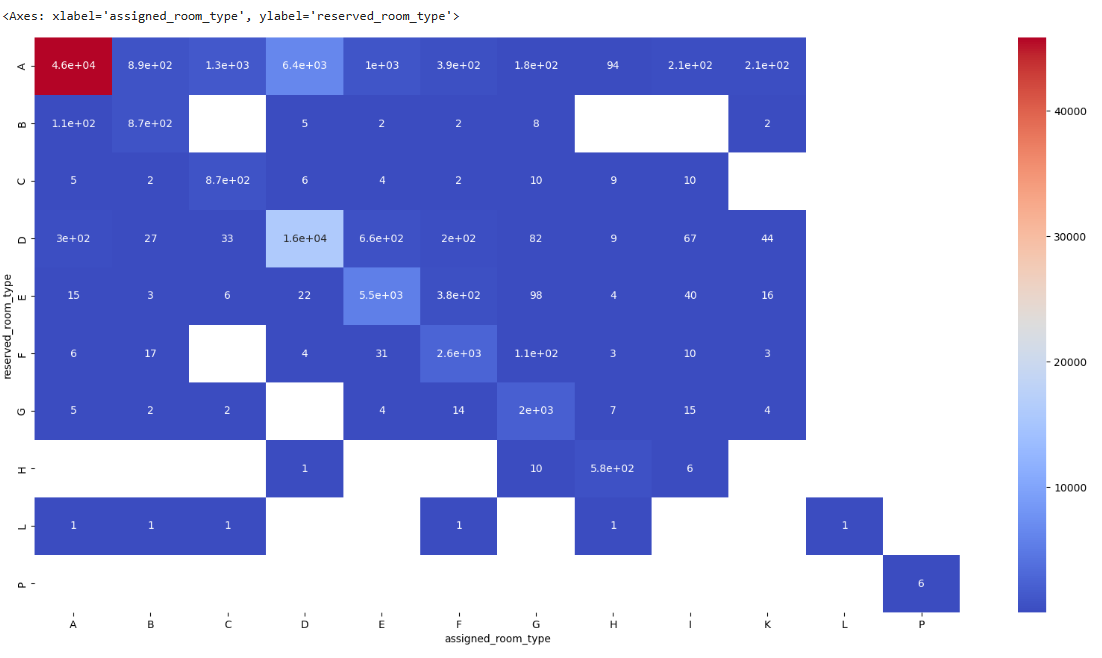
Solution:

Code:

* 1. mcc=pd.DataFrame(booking[['reserved\_room\_type','assigned\_room\_type']].value\_counts().reset\_index())
  2. mcc
  3. data = mcc.pivot(index='reserved\_room\_type',columns='assigned\_room\_type',values='count')
  4. plt.figure(figsize=(20, 10))
  5. sns.heatmap(data,annot=True, cmap='coolwarm')

Output:





Explanation & Insights:

The combination of ‘assigned\_room\_type’ ‘A’ and ‘assigned\_Room\_type’ ‘A’ is the highest, followed by combination of ‘assigned\_room\_type’ ‘D’ and ‘assigned\_Room\_type’ ‘D’.

**Subjective Questions:**

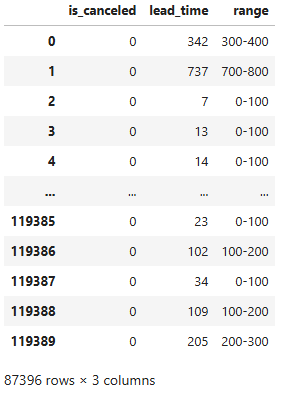
1. How does the lead time affect the likelihood of booking cancellations?

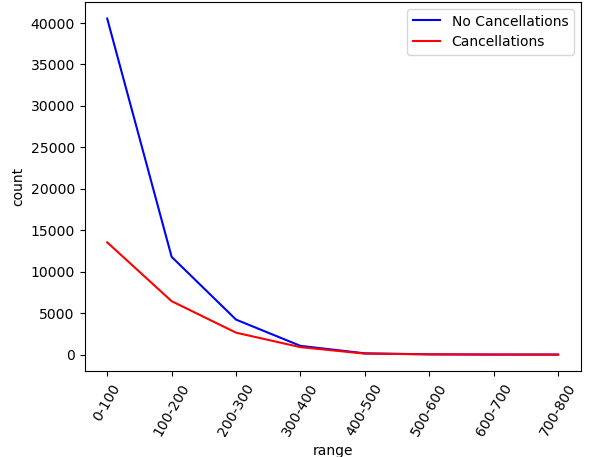
Solution:

Code:

* 1. likelihood=booking[['is\_canceled','lead\_time']]
  2. likelihood.sort\_values(by='lead\_time', ascending=False)
  3. new=[]
  4. likelihood=pd.DataFrame(likelihood)
  5. bin = [0,100,200,300,400,500,600,700,800]
  6. lables=['0-100','100-200','200-300','300-400','400-500','500-600','600-700','700-800']
  7. likelihood['range']=pd.cut(likelihood['lead\_time'],bins=bin,labels=lables)
  8. likelihood
  9. likelihood\_range=likelihood.groupby(['is\_canceled','range'])['lead\_time'].value\_counts()
  10. likely=likelihood\_range.dropna().round(2).reset\_index()
  11. likely=likely.groupby(['is\_canceled','range'])['count'].sum().reset\_index()
  12. print(likely)
  13. sns.lineplot(data=likely[likely['is\_canceled']==0], x='range', y='count', label ='No Cancellations',color='blue')
  14. sns.lineplot(data=likely[likely['is\_canceled']==1], x='range', y='count', label ='Cancellations',color='red')
  15. plt.xticks(rotation=60)
  16. plt.show()

Output:





Explanation & Insights:

The Total Number of Bookings (cancelled & not cancelled) having average Lead Time between 0-100 are highest (approx. 40000, 13000 respectively)

Both No. of Bookings Exponentially reduce till around lead time between 0-400, However the Total number of bookings for lead time range between 400-800 remains same for both.

1. What impact do special requests have on booking cancellations?

Solution:

Code:

* 1. src = booking.groupby('is\_canceled')['total\_of\_special\_requests'].sum()
  2. src
  3. sns.barplot(data=src)

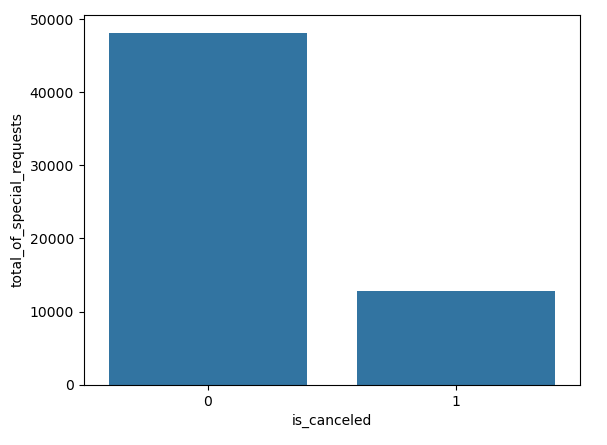
Output:

is\_canceled

0 48182

1 12870

Name: total\_of\_special\_requests, dtype: int64



Explanation & Insights:

We will find the sum of the total number of special request w.r.t. cancelled.

The total number of special requests for non cancelled bookings is higher than the cancelled bookings.

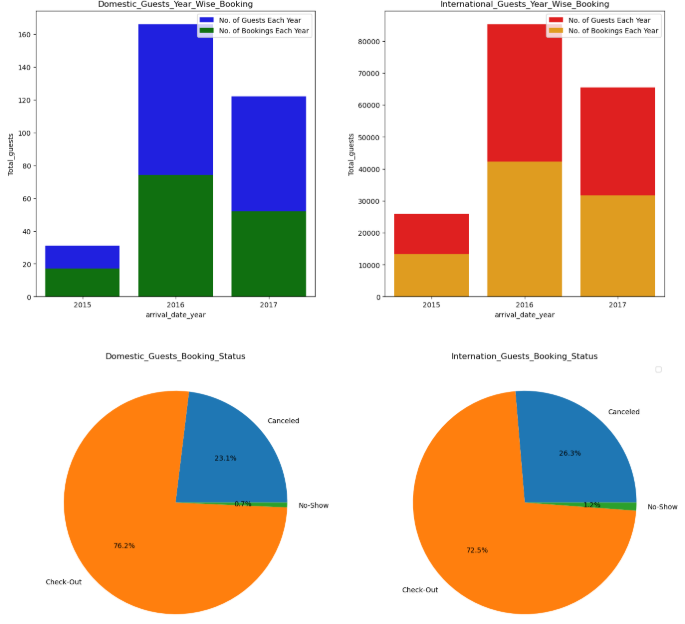
1. What trends can be observed in the booking behavior of international guests compared to domestic guests?

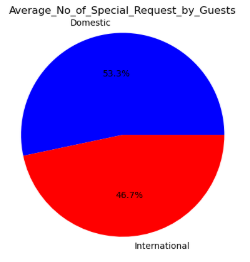
Solution:

Code:

* 1. data=booking
  2. data['Guest\_type']=np.where(data['country']=='IND',"Domestic","International")
  3. year\_wise\_booking=pd.DataFrame(data[['arrival\_date\_year','Guest\_type']].value\_counts().reset\_index())
  4. guest\_reservation=pd.DataFrame(data[['Guest\_type','reservation\_status']].value\_counts().reset\_index())
  5. guest\_numbers=data.groupby(['Guest\_type','arrival\_date\_year'])['Total\_guests'].sum().reset\_index()
  6. guest\_sp\_request=data.groupby('Guest\_type')['total\_of\_special\_requests'].mean().reset\_index()
  7. guest\_numbers
  8. fig,axes=plt.subplots(2,2,figsize=(15,15))
  9. fig.tight\_layout(pad=7.0)
  10. #Guests booking over the years
  11. sns.barplot(x='arrival\_date\_year',y='Total\_guests',data=guest\_numbers[guest\_numbers['Guest\_type']=='Domestic'].sort\_values(by='arrival\_date\_year'),ax=axes[0,0],color='blue',label='No. of Guests Each Year').set\_title('No\_of\_Domestic\_Guests\_each\_year')
  12. sns.countplot(x='arrival\_date\_year',data=booking[booking['Guest\_type']=='Domestic'].sort\_values(by='arrival\_date\_year'),ax=axes[0,0],color="green",label='No. of Bookings Each Year').set\_title('Domestic\_Guests\_Year\_Wise\_Booking')
  13. plt.legend()
  14. sns.barplot(x='arrival\_date\_year',y='Total\_guests',data=guest\_numbers[guest\_numbers['Guest\_type']=='International'].sort\_values(by='arrival\_date\_year'),ax=axes[0,1],color='red',label='No. of Guests Each Year').set\_title('No\_of\_International\_Guests\_each\_year')
  15. sns.countplot(x='arrival\_date\_year',data=booking[booking['Guest\_type']=='International'].sort\_values(by='arrival\_date\_year'),ax=axes[0,1],color="orange",label='No. of Bookings Each Year').set\_title('International\_Guests\_Year\_Wise\_Booking')
  16. plt.legend()
  17. #Booking Status of guests
  18. '''sns.countplot(x='reservation\_status',data=booking[booking['Guest\_type']=='Domestic'].sort\_values(by='reservation\_status'),ax=axes[1,0],color='blue').set\_title('Domestic\_Guests\_Booking\_Status')
  19. sns.countplot(x='reservation\_status',data=booking[booking['Guest\_type']=='International'].sort\_values(by='reservation\_status'),ax=axes[1,1],color='red').set\_title('International\_Guests\_Booking\_Status')
  20. '''
  21. axes[1,0].pie(data=guest\_reservation[guest\_reservation['Guest\_type']=='Domestic'].sort\_values(by='reservation\_status'),x='count',labels='reservation\_status',autopct='%.1f%%',radius=1)
  22. axes[1,0].set\_title('Domestic\_Guests\_Booking\_Status')
  23. axes[1,1].pie(data=guest\_reservation[guest\_reservation['Guest\_type']=='International'].sort\_values(by='reservation\_status'),x='count',labels='reservation\_status',autopct='%.1f%%',radius=1)
  24. axes[1,1].set\_title('Internation\_Guests\_Booking\_Status')
  25. plt.show()
  26. plt.pie(labels='Guest\_type',x='total\_of\_special\_requests',data=guest\_sp\_request,colors=['blue','red'],autopct='%.1f%%',radius=1.1)
  27. plt.title('Average\_No\_of\_Special\_Request\_by\_Guests')
  28. plt.show()

Output:





Explanation & Insights:

We will find the difference between Domestic (Indian National) and International guests on various fronts namely Total No. of Bookings and Total No. of Guests each year, Booking Status and Average No of Special requests.

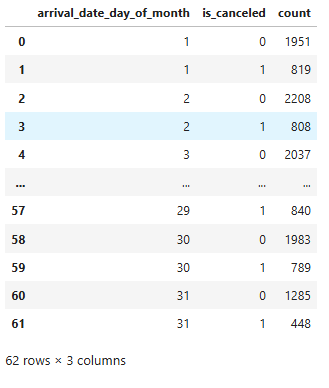
1. What is the impact of the day of the week on booking cancellations?

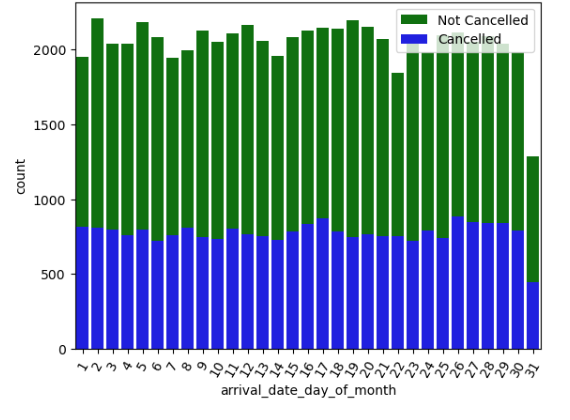
Solution:

Code:

* 1. impact = booking[['arrival\_date\_day\_of\_month','is\_canceled']]
  2. cancellation\_impact=pd.DataFrame(impact.groupby('arrival\_date\_day\_of\_month')['is\_canceled'].value\_counts().reset\_index())
  3. cancellation\_impact
  4. #pd.concat([pd.DataFrame(cancellation\_impact[cancellation\_impact['is\_canceled']==0]),pd.DataFrame(cancellation\_impact[cancellation\_impact['is\_canceled']==1])],axis=1)
  5. sns.barplot(data=cancellation\_impact[cancellation\_impact['is\_canceled']==0],x='arrival\_date\_day\_of\_month',y='count',color='green',label='Not Cancelled')
  6. sns.barplot(data=cancellation\_impact[cancellation\_impact['is\_canceled']==1],x='arrival\_date\_day\_of\_month',y='count',color='blue',label='Cancelled')
  7. plt.xticks(rotation=60)
  8. plt.legend()
  9. plt.show()

Output:





Explanation & Insights:

We will calculate the no of bookings that were cancelled and not cancelled each day

We see that on average the cancelled are at 700 bookings per day while Not cancelled bookings are on average 2000.

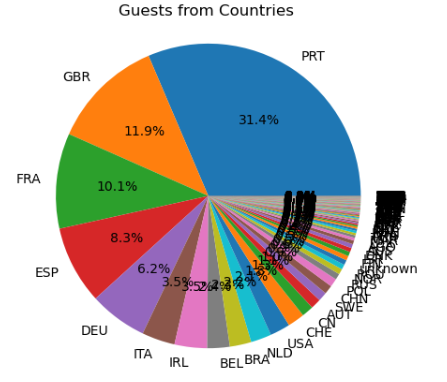
1. What are the most common countries for international guests, and how do their booking patterns compare to domestic guests?

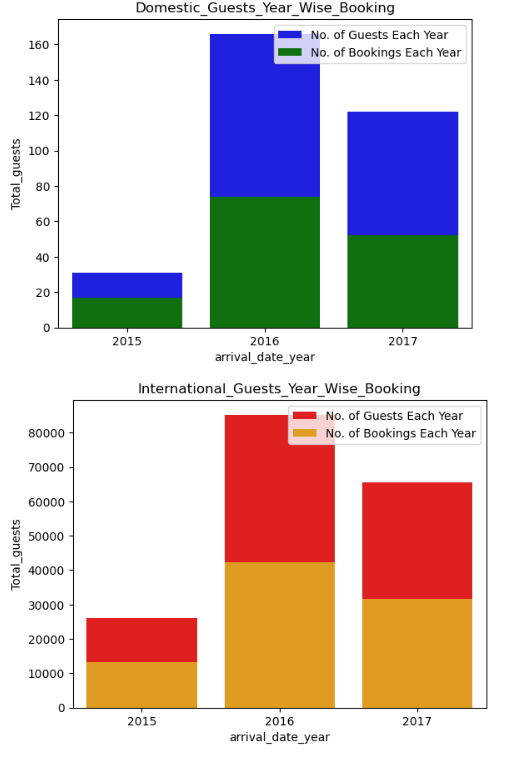
Solution:

Code:

* 1. most\_common\_countries=booking[booking['country']!='IND']['country'].value\_counts().reset\_index()
  2. plt.pie(data=most\_common\_countries, x='count', labels='country',autopct='%.1f%%',radius=1.1)
  3. plt.title('Guests from Countries ')
  4. plt.show()
  5. guest\_numbers=booking.groupby(['Guest\_type','arrival\_date\_year'])['Total\_guests'].sum().reset\_index()
  6. sns.barplot(x='arrival\_date\_year',y='Total\_guests',data=guest\_numbers[guest\_numbers['Guest\_type']=='Domestic'].sort\_values(by='arrival\_date\_year'),color='blue',label='No. of Guests Each Year').set\_title('No\_of\_Domestic\_Guests\_each\_year')
  7. sns.countplot(x='arrival\_date\_year',data=booking[booking['Guest\_type']=='Domestic'].sort\_values(by='arrival\_date\_year'),color="green",label='No. of Bookings Each Year').set\_title('Domestic\_Guests\_Year\_Wise\_Booking')
  8. plt.legend()
  9. plt.show()
  10. sns.barplot(x='arrival\_date\_year',y='Total\_guests',data=guest\_numbers[guest\_numbers['Guest\_type']=='International'].sort\_values(by='arrival\_date\_year'),color='red',label='No. of Guests Each Year').set\_title('No\_of\_International\_Guests\_each\_year')
  11. sns.countplot(x='arrival\_date\_year',data=booking[booking['Guest\_type']=='International'].sort\_values(by='arrival\_date\_year'),color="orange",label='No. of Bookings Each Year').set\_title('International\_Guests\_Year\_Wise\_Booking')
  12. plt.legend()
  13. plt.show()

Output:





Explanation & Insights:

To get the most common counties for international guests, we can find the total number of bookings w.r.t countries

To get the booking patterns of international guest vs domestic, we can find the count of bookings and count of guests each year for both domestic and international guests.

2016 and 2017 are seen to have the most number of bookings but number of guests is only 50% of number of bookings. However in 2015 the number of guests is 60% of number of bookings