In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
df = pd.read_csv("C:/Users/Sonia Kaushik/Downloads/hotel_booking.csv")
df.head(5)
```

Out[2]:

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

5 rows × 36 columns

In [3]:
```
df.shape
```

Out[3]:
```
(119390, 36)
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 36 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 
 32  name                            119390 non-null  object 
 33  email                           119390 non-null  object 
 34  phone-number                    119390 non-null  object 
 35  credit_card                     119390 non-null  object 
dtypes: float64(4), int64(16), object(16)
memory usage: 32.8+ MB
```

In [5]:
```
df.drop_duplicates(inplace=True)
```

In [6]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 119390 entries, 0 to 119389
Data columns (total 36 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 
 32  name                            119390 non-null  object 
 33  email                           119390 non-null  object 
 34  phone-number                    119390 non-null  object 
 35  credit_card                     119390 non-null  object 
dtypes: float64(4), int64(16), object(16)
memory usage: 33.7+ MB
```

In [7]:
```
df.drop(['name','email','phone-number','required_car_parking_spaces','credit_card
```

In [8]:
```
df.head()
```

Out[8]:

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

5 rows × 30 columns

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

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 119390 entries, 0 to 119389
Data columns (total 30 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  total_of_special_requests       119390 non-null  int64  
 28  reservation_status              119390 non-null  object 
 29  reservation_status_date         119390 non-null  object 
dtypes: float64(3), int64(15), object(12)
memory usage: 28.2+ MB
```

In [10]:
```
df['Family']=df['adults']+df['children']+df['babies']
```

In [11]:
```
df['Family'].info()
```

```
<class 'pandas.core.series.Series'>
Int64Index: 119390 entries, 0 to 119389
Series name: Family
Non-Null Count   Dtype  
--------------   -----  
119386 non-null  float64
dtypes: float64(1)
memory usage: 1.8 MB
```

In [12]:
```
df.drop(['adults','children','babies','company','agent'],axis=1,inplace=True)
```

In [13]:
```
df.head()
```

Out[13]:

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

5 rows × 26 columns

In [14]:
```
df['hotel'].replace(['Resort Hotel','City Hotel'],[4,5],inplace=True)
```

In [15]:
```
df['arrival_date_month'].value_counts()
```

Out[15]:
```
August       13877
July         12661
May          11791
October      11160
April        11089
June         10939
September    10508
March         9794
February      8068
November      6794
December      6780
January       5929
Name: arrival_date_month, dtype: int64
```

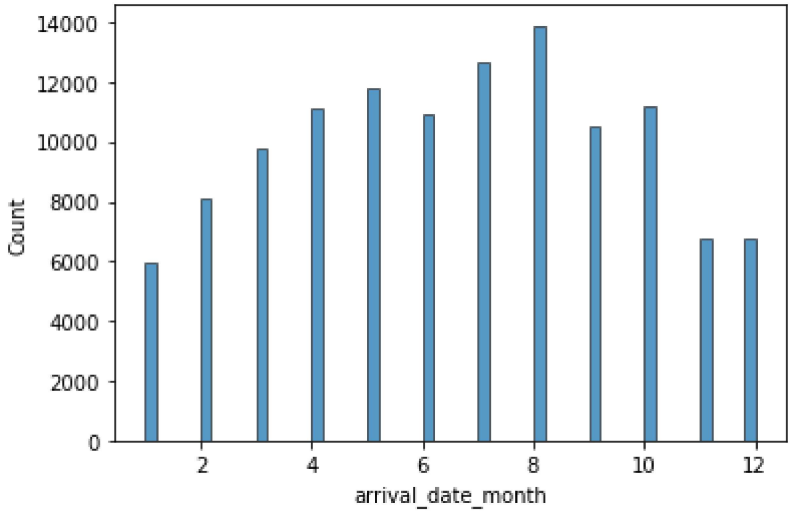
In [16]:
```
d={'January':1,'February':2,'March':3,'April':4,'May':5,'June':6,'July':7,'August
```

In [17]:
```
df.arrival_date_month=df.arrival_date_month.map(d)
```

In [18]:
```
df['arrival_date_month']
```

Out[18]:
```
0         7
1         7
2         7
3         7
4         7
         ..
119385    8
119386    8
119387    8
119388    8
119389    8
Name: arrival_date_month, Length: 119390, dtype: int64
```

In [19]:
```
import seaborn as sns
sns.histplot(x='arrival_date_month', data=df);
```



In [20]:
```
a = df.select_dtypes(object).columns
for i in a:
    print (i, df[i].nunique())
```

```
meal 5
country 177
market_segment 8
distribution_channel 5
reserved_room_type 10
assigned_room_type 12
deposit_type 3
customer_type 4
reservation_status 3
reservation_status_date 926
```

In [21]:
```
def categories(row):
    if ((row['deposit_type'] == 'No Deposit') | (row['deposit_type'] == 'Refundab
        return 0
    elif (row['deposit_type'] == 'Non Refund'):
        return 1
```

In [22]:
```
df['Deposit_type']=df.apply(lambda row:categories(row),axis=1)
```

In [23]:
```
def categories(row):

    if row['Family']>5:

        return "0"
    elif row['Family']<=5:
        return "1"
```

In [24]:
```
df['family']=df.apply(lambda row:categories(row),axis=1)
```

In [25]:
```
df.drop(['Family','deposit_type'],axis=1,inplace=True)
```

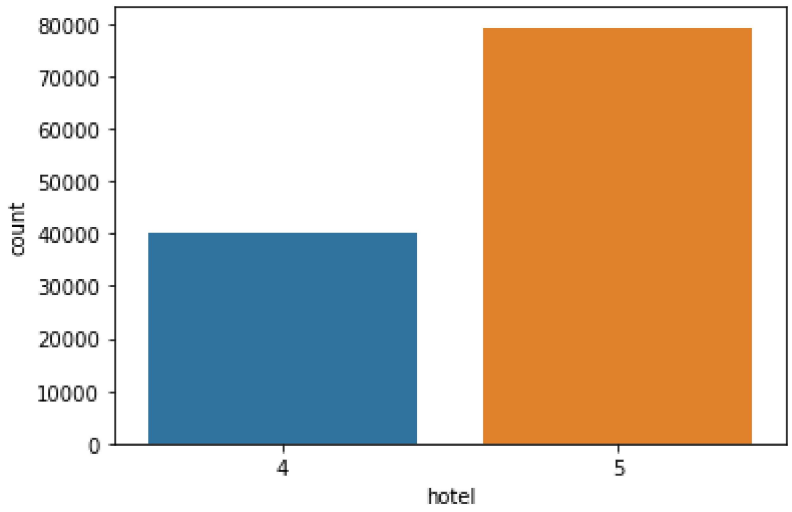
In [26]:
```
df['family']=df['family'].fillna(df['family'].mean)
```

In [27]:
```
df['country']=df['country'].fillna(df['country'].mode)
```

In [28]:
```
df['hotel'].value_counts()
```

Out[28]:
```
5    79330
4    40060
Name: hotel, dtype: int64
```

In [29]:
```
sns.countplot(x='hotel', data=df);
```

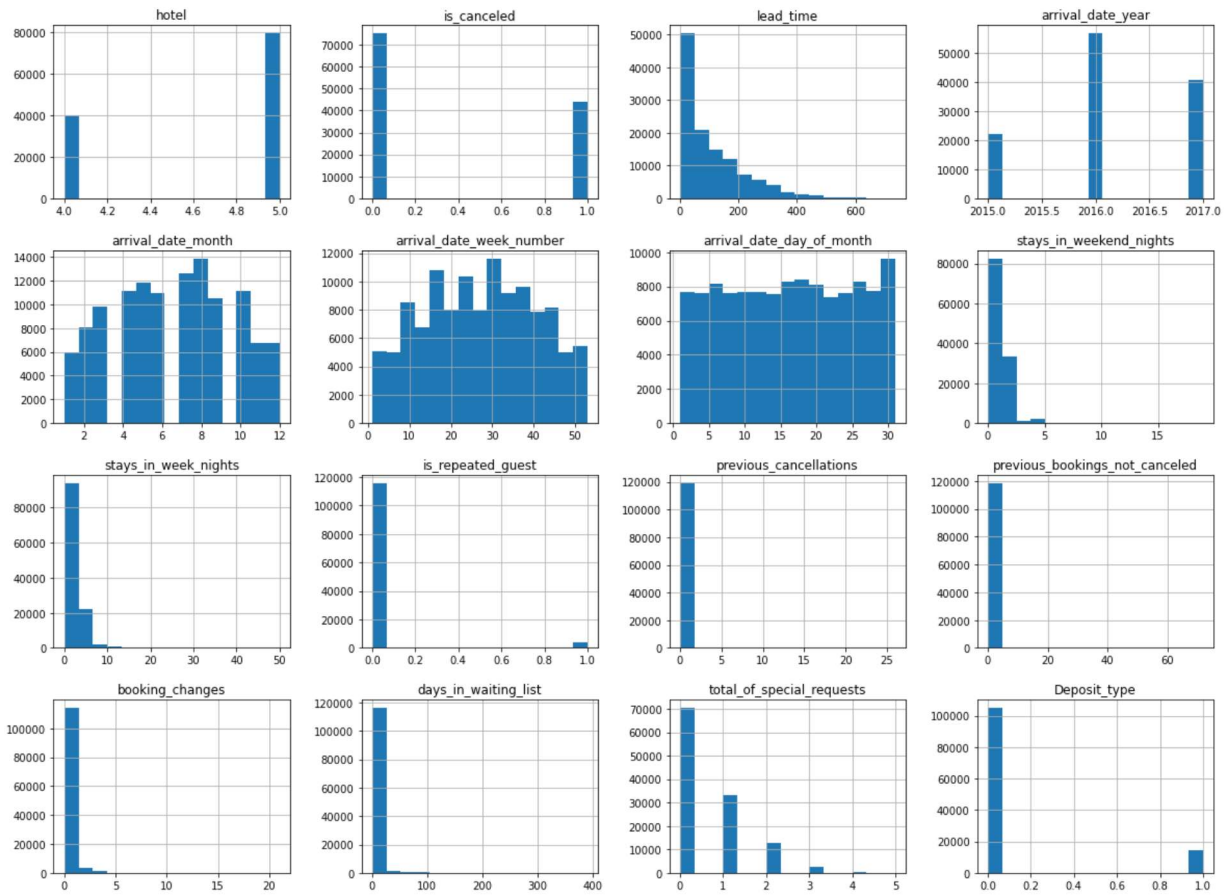


In [30]:
```
df['country']=df['country'].fillna(df.mode().iloc[0])
```

In [31]:
```
import matplotlib.pyplot as plt
%matplotlib inline
```

In [32]:
```
df.hist(bins=15,figsize=(20,15))
```

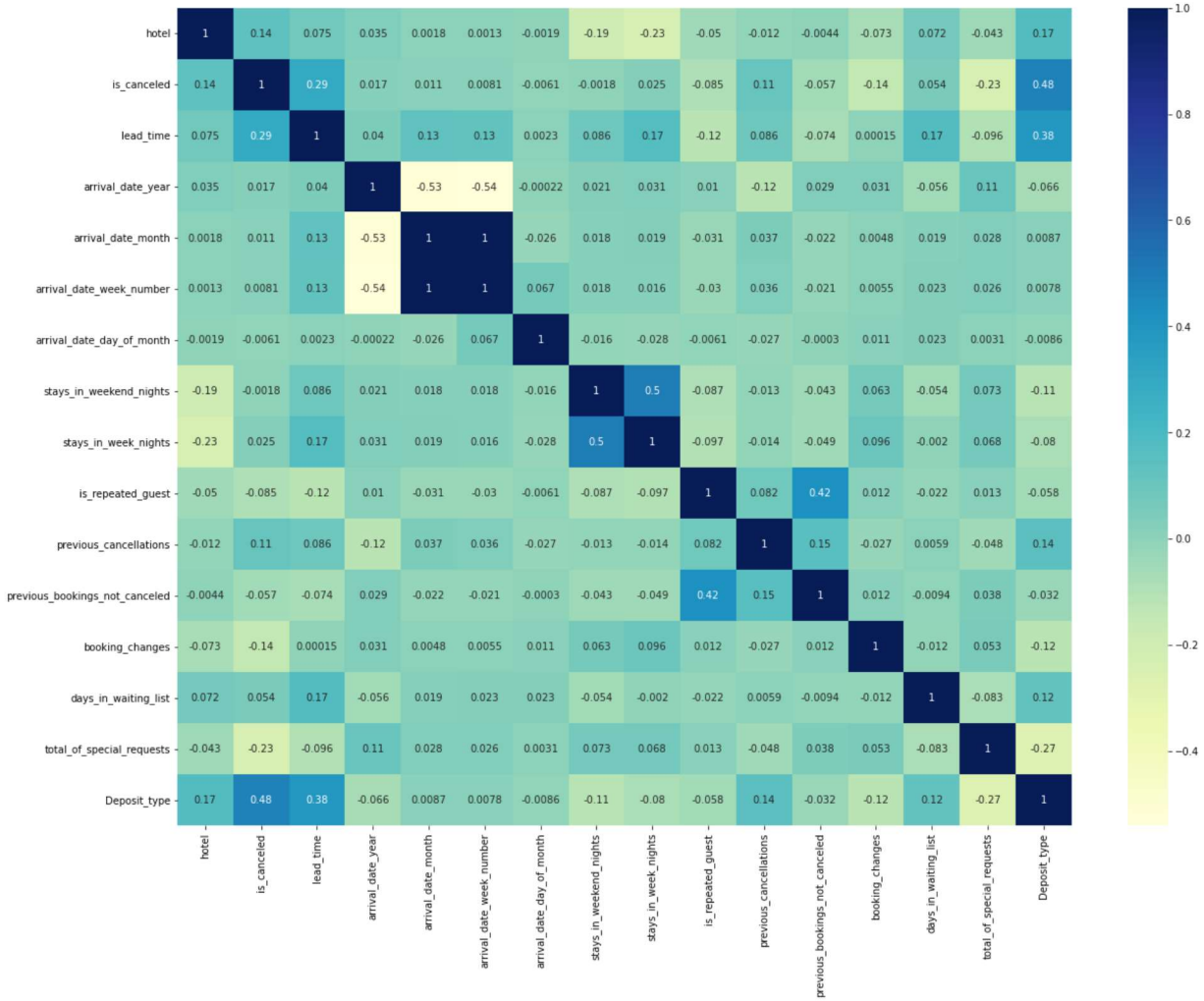
Out[32]:
```
array([[<AxesSubplot:title={'center':'hotel'}>,
        <AxesSubplot:title={'center':'is_canceled'}>,
        <AxesSubplot:title={'center':'lead_time'}>,
        <AxesSubplot:title={'center':'arrival_date_year'}>],
       [<AxesSubplot:title={'center':'arrival_date_month'}>,
        <AxesSubplot:title={'center':'arrival_date_week_number'}>,
        <AxesSubplot:title={'center':'arrival_date_day_of_month'}>,
        <AxesSubplot:title={'center':'stays_in_weekend_nights'}>],
       [<AxesSubplot:title={'center':'stays_in_week_nights'}>,
        <AxesSubplot:title={'center':'is_repeated_guest'}>,
        <AxesSubplot:title={'center':'previous_cancellations'}>,
        <AxesSubplot:title={'center':'previous_bookings_not_canceled'}>],
       [<AxesSubplot:title={'center':'booking_changes'}>,
        <AxesSubplot:title={'center':'days_in_waiting_list'}>,
        <AxesSubplot:title={'center':'total_of_special_requests'}>,
        <AxesSubplot:title={'center':'Deposit_type'}>]], dtype=object)
```



In [33]:
```
df.info()
```

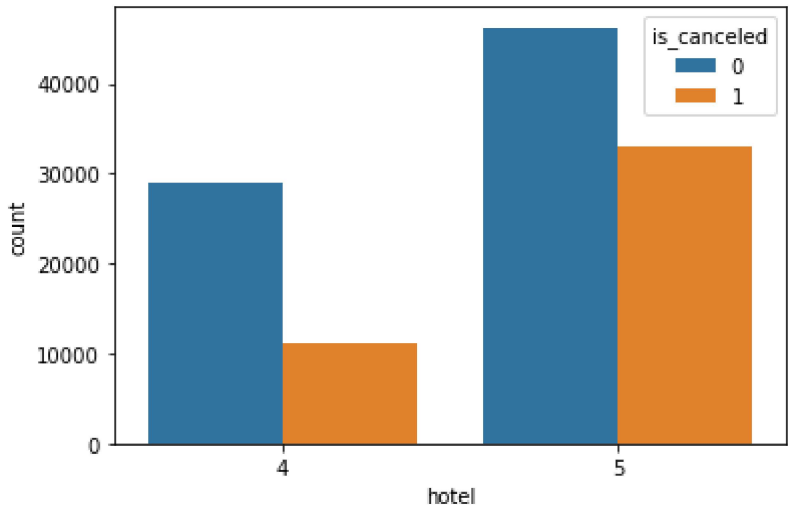
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 119390 entries, 0 to 119389
Data columns (total 26 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  int64
 1   is_canceled                     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  int64
 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   meal                            119390 non-null  object
 10  country                         119390 non-null  object
 11  market_segment                  119390 non-null  object
 12  distribution_channel            119390 non-null  object
 13  is_repeated_guest               119390 non-null  int64
 14  previous_cancellations          119390 non-null  int64
 15  previous_bookings_not_canceled  119390 non-null  int64
 16  reserved_room_type              119390 non-null  object
 17  assigned_room_type              119390 non-null  object
 18  booking_changes                 119390 non-null  int64
 19  days_in_waiting_list            119390 non-null  int64
 20  customer_type                   119390 non-null  object
 21  total_of_special_requests       119390 non-null  int64
 22  reservation_status              119390 non-null  object
 23  reservation_status_date         119390 non-null  object
 24  Deposit_type                    119390 non-null  int64
 25  family                          119390 non-null  object
dtypes: int64(16), object(10)
memory usage: 28.6+ MB
```

In [34]:
```
plt.figure(figsize=(20,15))
ax = sns.heatmap(df.corr(),cmap='YlGnBu',annot=True)
plt.show()
```



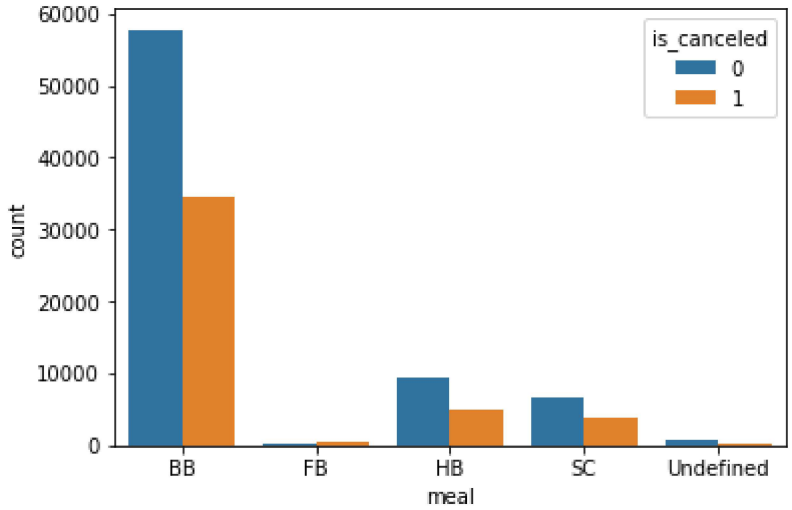
In [35]:
```
sns.countplot(x='hotel',hue='is_canceled',data=df)
```

Out[35]: <AxesSubplot:xlabel='hotel', ylabel='count'>



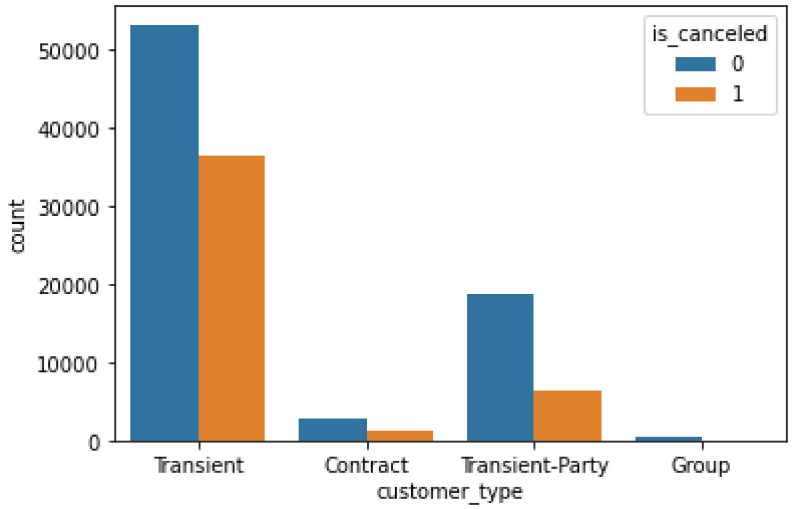
In [36]:
```
sns.countplot(x='meal',hue='is_canceled',data=df)
```

Out[36]: <AxesSubplot:xlabel='meal', ylabel='count'>

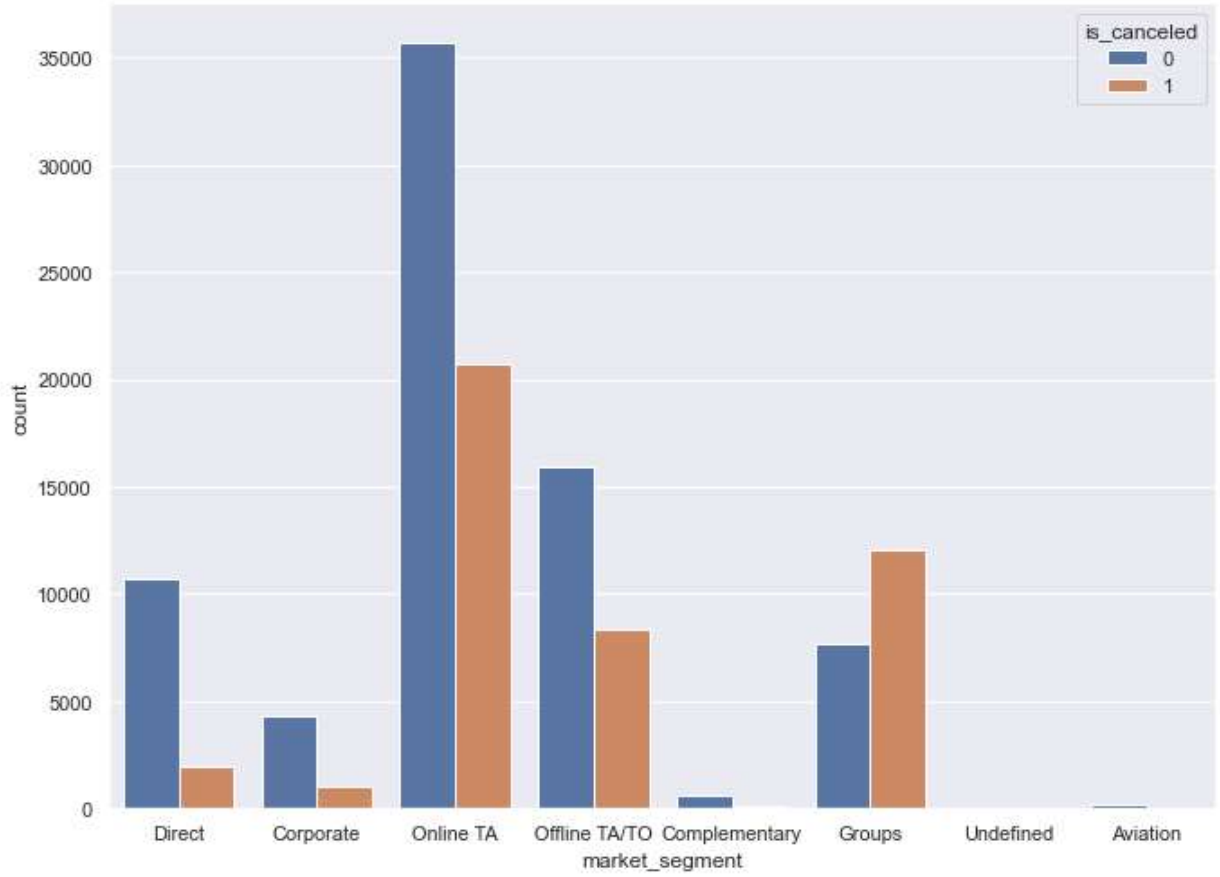


In [37]:
```
sns.countplot(x='customer_type',hue='is_canceled',data=df)
```

Out[37]:
```
<AxesSubplot:xlabel='customer_type', ylabel='count'>
```

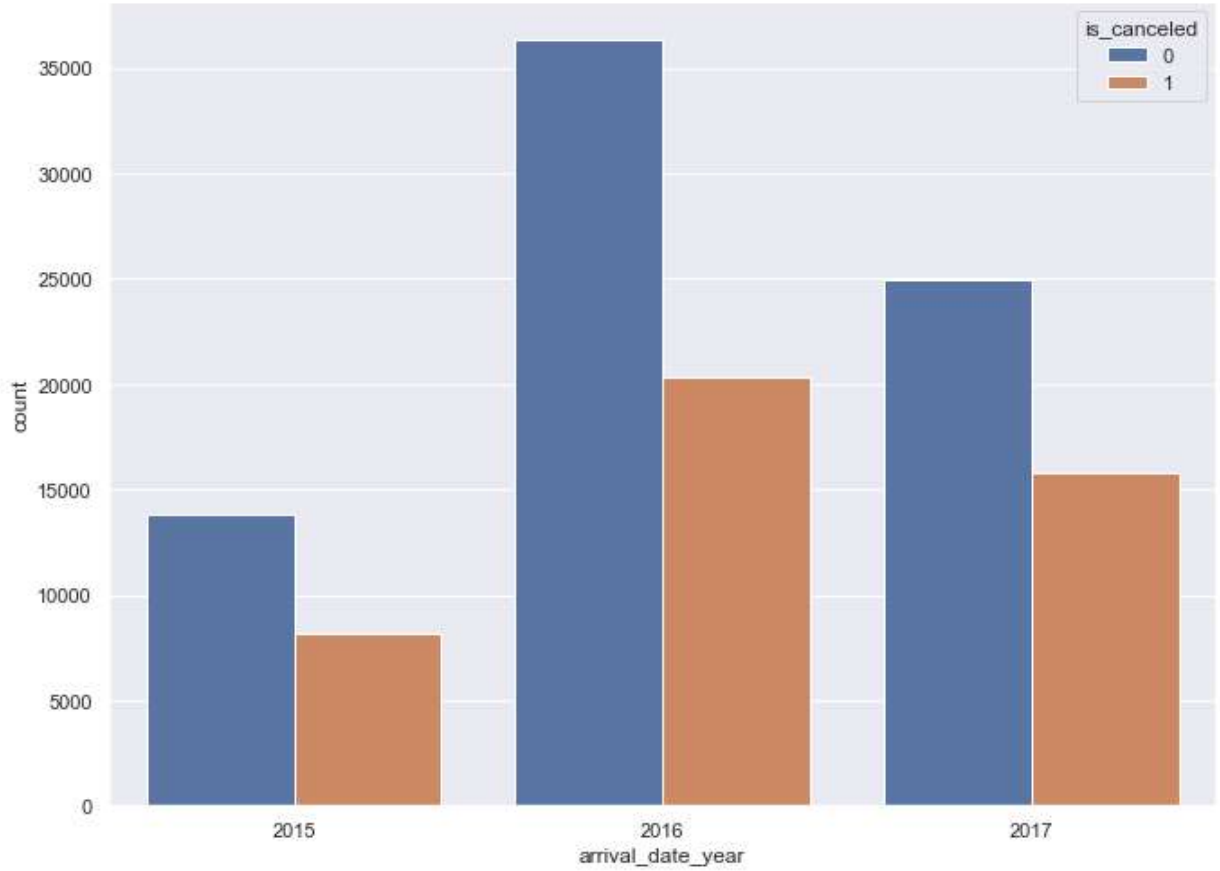


In [64]:
```
sns.countplot(x='market_segment',hue='is_canceled',data=df)
sns.set(rc={"figure.figsize":(11,8.2)})
```



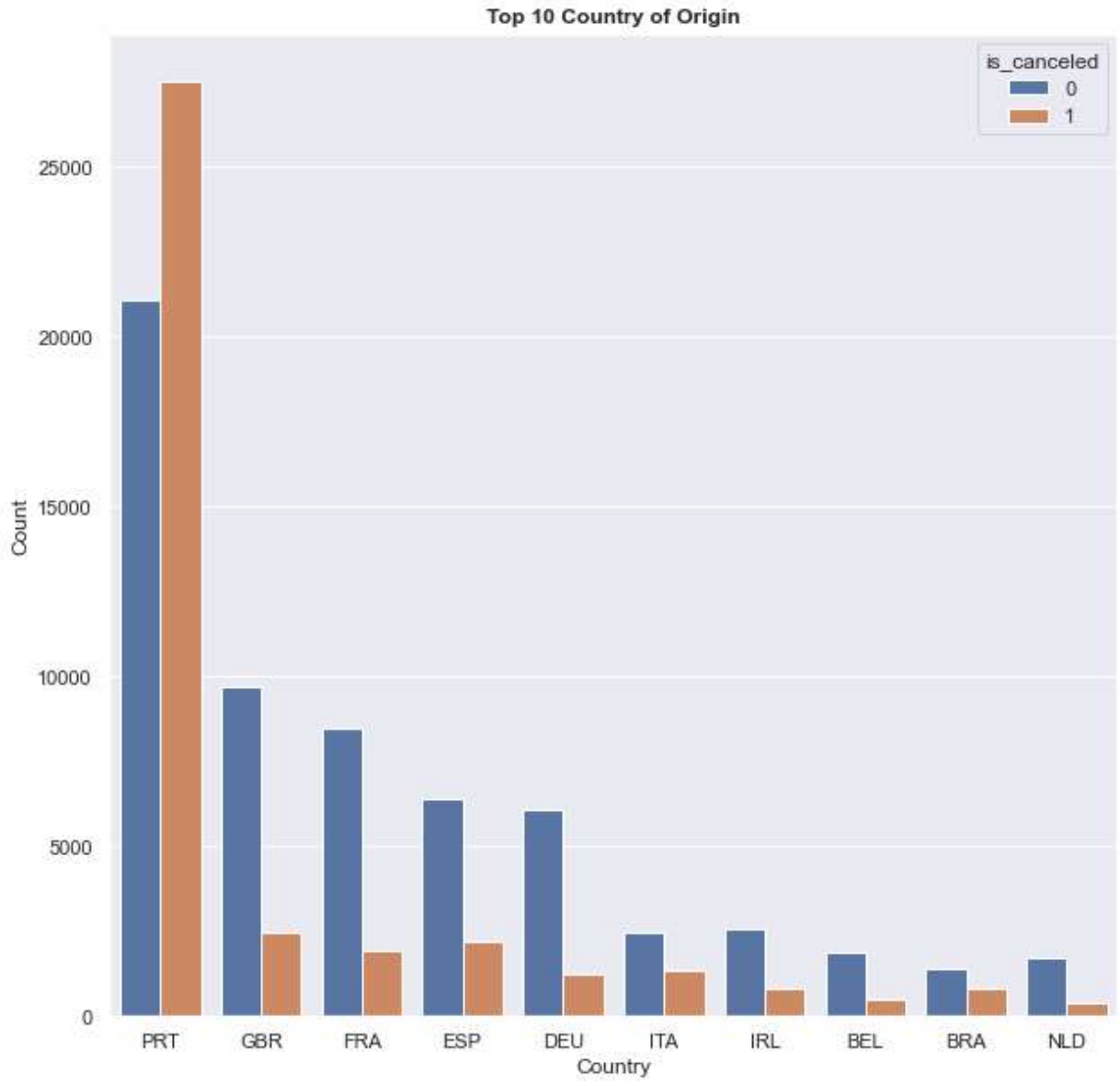
In [39]:
```
sns.countplot(x='arrival_date_year',hue='is_canceled',data=df)
```

Out[39]:
```
<AxesSubplot:xlabel='arrival_date_year', ylabel='count'>
```



In [40]:
```
plt.figure(figsize=(10,10))
sns.countplot(x='country', data=df,order=pd.value_counts(df['country']).iloc[:10]
plt.title('Top 10 Country of Origin', weight='bold')
plt.xlabel('Country', fontsize=12)
plt.ylabel('Count', fontsize=12)
```

Out[40]: Text(0, 0.5, 'Count')



In [45]:
```
df1=pd.get_dummies(df,columns=['meal','reservation_status_date','distribution_cha
```

In [46]:
```
from sklearn.model_selection import train_test_split
X = df1.drop('is_canceled', axis = 1)
y = df1['is_canceled']

X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.2,random_stat
```

In [47]:
```
from sklearn import preprocessing

scaler = preprocessing.StandardScaler()
X_train_sc = scaler.fit_transform(X_train)
X_test_sc = scaler.transform(X_test)
```

In [48]:
```
from sklearn.ensemble import RandomForestClassifier
model = RandomForestClassifier(n_estimators=400)
```

In [49]:
```
model.fit(X_train_sc,y_train)
```

Out[49]:
```
RandomForestClassifier(n_estimators=400)
```

In [51]:
```
y_pred = model.predict(X_test_sc)
```

In [52]:
```
from sklearn.metrics import accuracy_score,confusion_matrix
accuracy_score(y_test,y_pred)
```

Out[52]:
```
1.0
```

In [53]:
```
confusion_matrix(y_test,y_pred)
```

Out[53]:
```
array([[15057,     0],
       [    0,  8821]], dtype=int64)
```

In [ ]: