In [0]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [0]:

```
data = pd.read_csv('https://raw.githubusercontent.com/philurnezhys/visualization_lab/master/lab1/hotel_bookings.csv?token=ANZN4CBGEDHID5IVMP673RC6NACI2')
```

In [0]:

```
data.describe()
```

Out[0]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_dat
count	119390.000000	119390.000000	119390.000000	119390.000000	
mean	0.370416	104.011416	2016.156554	27.165173	
std	0.482918	106.863097	0.707476	13.605138	
min	0.000000	0.000000	2015.000000	1.000000	
25%	0.000000	18.000000	2016.000000	16.000000	
50%	0.000000	69.000000	2016.000000	28.000000	
75%	1.000000	160.000000	2017.000000	38.000000	
max	1.000000	737.000000	2017.000000	53.000000	

In [0]:

```
data.head()
```

Out[0]:

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

In [0]:

```
data.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
```

Подготовка данных

In [0]:

```
data.isnull().sum()
```

Out[0]:

```
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
```

4 пропуска в детях

488 пропуска в странах

Пропущенные ячейки заполняем средним возрастом всех детей

In [0]:

```
average_children = round(data["children"].mean())
data["children"] = data["children"].fillna(value=average_children)
```

In [0]:

```
data["country"].value_counts().head(5)
```

Out[0]:

```
PRT    48590
GBR    12129
FRA    10415
ESP     8568
DEU     7287
Name: country, dtype: int64
```

Пропущенные ячейки заполняем Португалией, потому что это самая популярная страна в датасете

In [0]:

```
data["country"] = data["country"].fillna(value="PRT")
```

Удаляем ячейки с Агентами и Компаниями из-за ненадобности

In [0]:

```
data.drop(["agent"], axis = 1, inplace = True)
data.drop(["company"], axis = 1, inplace = True)
```

In [0]:

```
data.isnull().sum()
```

Out[0]:

```
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 [0]:

```
sns.countplot(data.hotel)
```

Out[0]:

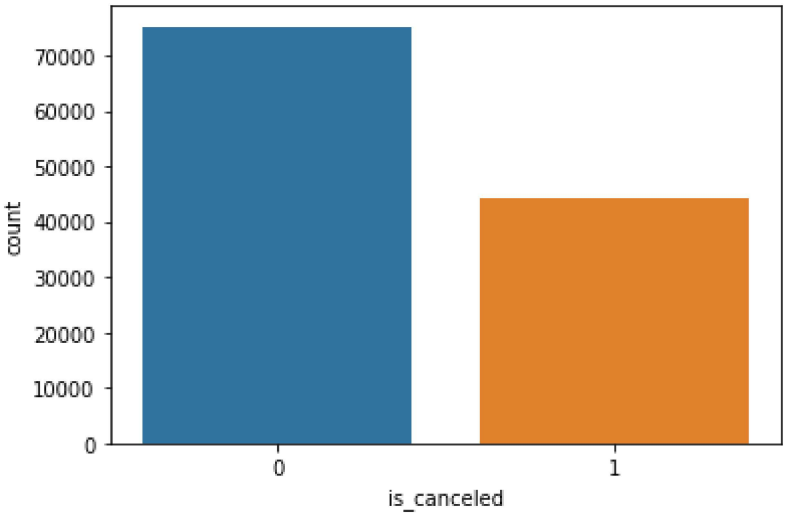
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe4d5779eb8>
```

In [0]:

```
sns.countplot(data.is_canceled)
```

Out[0]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe4d3e91128>
```

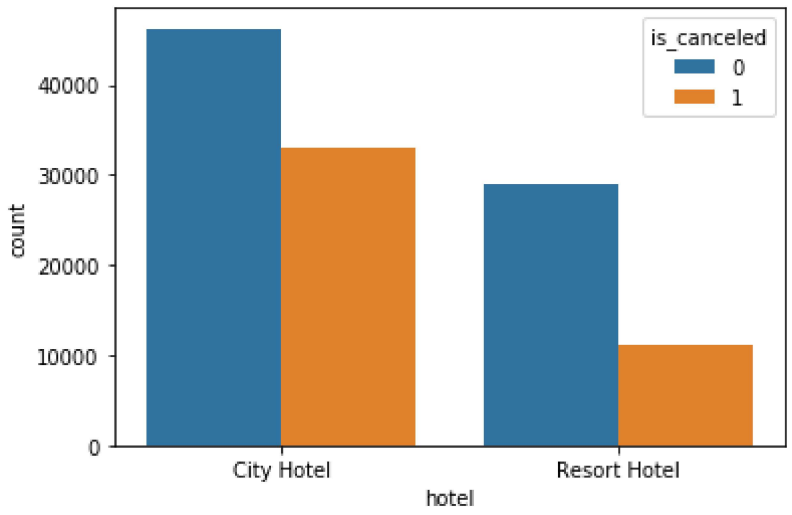


In [0]:

```
cancel=data.groupby(["hotel","is_canceled"]).lead_time.count().reset_index()
cancel.columns=["hotel","is_canceled","count"]
sns.barplot(x="hotel", y="count", hue="is_canceled", data=cancel)
```

Out[0]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe4d3e363c8>
```



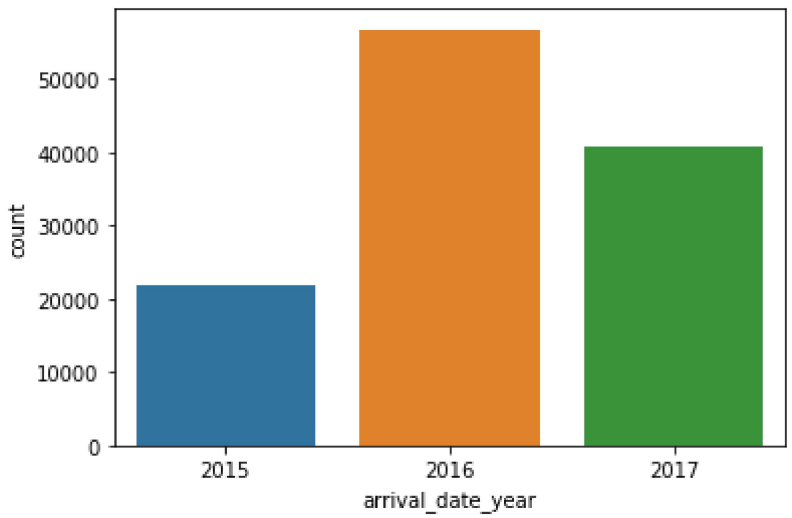
Резервацию отелей отменяют больше в обычных городских отелях

In [0]:

```
sns.countplot(data.arrival_date_year)
```

Out[0]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe4d35f31d0>
```

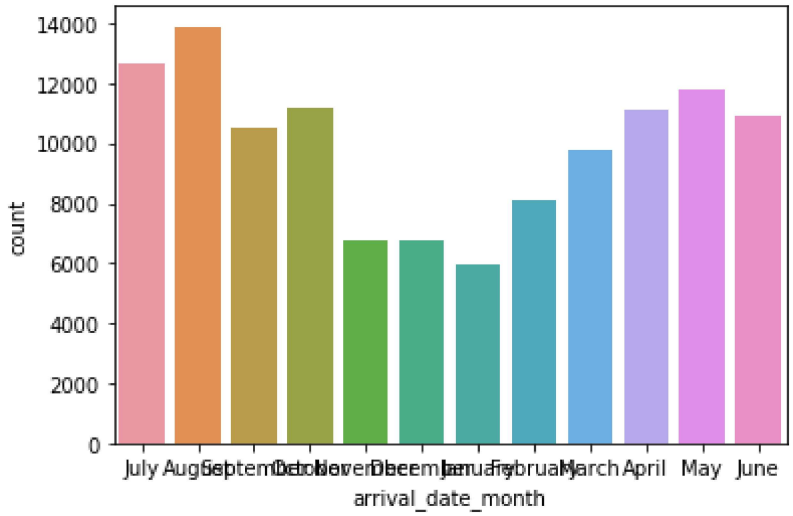


In [0]:

```
sns.countplot(data.arrival_date_month)
```

Out[0]:

```
<Figure size 1800x432 with 0 Axes>
```




```
<Figure size 1800x432 with 0 Axes>
```

По данному графику видим, что больше всего людей заселяются в отель в августе

In [0]:

```
data_august = data.loc[data.arrival_date_month=='August']
plt.figure(figsize=(20,6))
sns.countplot(data_august.arrival_date_day_of_month, hue=data_august.arrival_date_year)
```

Out[0]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe4d2cd0860>
```