

```
In [3]: ## Load Packages

In [5]: import os
import pandas as pd
import numpy as np
import seaborn as sns
from matplotlib import pyplot as plt

In [6]: ## Read and Understand the Data

In [7]: hotel_data = pd.read_csv('/Users/doyin/documents/Data Projects/hotel_bookings.csv')
hotel_data.head(5) # check first 5 rows of data

Out[7]:
```

|   | 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 | ... | deposit_type | agent | comp |
|---|--------------|-------------|-----------|-------------------|--------------------|--------------------------|---------------------------|-------------------------|----------------------|--------|-----|--------------|-------|------|
| 0 | Resort Hotel | 0           | 342       | 2015              | July               | 27                       | 1                         | 0                       | 0                    | 2      | ... | No Deposit   | NaN   | ...  |
| 1 | Resort Hotel | 0           | 737       | 2015              | July               | 27                       | 1                         | 0                       | 0                    | 2      | ... | No Deposit   | NaN   | ...  |
| 2 | Resort Hotel | 0           | 7         | 2015              | July               | 27                       | 1                         | 0                       | 1                    | 1      | ... | No Deposit   | NaN   | ...  |
| 3 | Resort Hotel | 0           | 13        | 2015              | July               | 27                       | 1                         | 0                       | 1                    | 1      | ... | No Deposit   | 304.0 | ...  |
| 4 | Resort Hotel | 0           | 14        | 2015              | July               | 27                       | 1                         | 0                       | 2                    | 2      | ... | No Deposit   | 240.0 | ...  |

5 rows × 32 columns

```
In [8]: hotel_data.dtypes # understanding the format of the data

Out[8]:
```

|                                |         |
|--------------------------------|---------|
| hotel                          | object  |
| is_canceled                    | int64   |
| lead_time                      | int64   |
| arrival_date_year              | int64   |
| arrival_date_month             | object  |
| arrival_date_week_number       | int64   |
| arrival_date_day_of_month      | int64   |
| stays_in_weekend_nights        | int64   |
| stays_in_week_nights           | int64   |
| adults                         | int64   |
| children                       | float64 |
| babies                         | int64   |
| meal                           | object  |
| country                        | object  |
| market_segment                 | object  |
| distribution_channel           | object  |
| is_repeated_guest              | int64   |
| previous_cancellations         | int64   |
| previous_bookings_not_canceled | int64   |
| reserved_room_type             | object  |
| assigned_room_type             | object  |
| booking_changes                | int64   |
| deposit_type                   | object  |
| agent                          | float64 |
| company                        | float64 |
| days_in_waiting_list           | int64   |
| customer_type                  | object  |
| adr                            | float64 |
| required_car_parking_spaces    | int64   |
| total_of_special_requests      | int64   |
| reservation_status             | object  |
| reservation_status_date        | object  |
| dtype:                         | object  |

```
In [9]: ## Deciding on Variables

In [10]: # deleting agent and company column

del hotel_data["agent"]
del hotel_data["company"]

In [11]: # looking further into unsure columns to decide if data is useful
hotel_data["market_segment"].describe()

Out[11]:
```

|        |                               |
|--------|-------------------------------|
| count  | 119390                        |
| unique | 8                             |
| top    | Online TA                     |
| freq   | 56477                         |
| Name:  | market_segment, dtype: object |

```
In [12]: hotel_data["market_segment"].unique()

Out[12]:
```

```
array(['Direct', 'Corporate', 'Online TA', 'Offline TA/T0',
      'Complementary', 'Groups', 'Undefined', 'Aviation'], dtype=object)
```

```
In [13]: del hotel_data["market_segment"]

In [14]: del hotel_data["distribution_channel"]

In [15]: ## Transforming Variables

In [16]: # cancelled column variable can be changed

hotel_data["is_canceled"].dtype

Out[16]:
```

```
dtype('int64')
```

```
In [17]: new_is_canceled = pd.Categorical(hotel_data["is_canceled"])
new_is_canceled = new_is_canceled.rename_categories(["Cancelled", "Booked"])
new_is_canceled.describe()

Out[17]:
```

|           | counts | freqs    |
|-----------|--------|----------|
| Cancelled | 75166  | 0.629584 |
| Booked    | 44224  | 0.370416 |

```
In [18]: hotel_data["Booking Status"] = new_is_canceled

In [19]: hotel_data.head(5)

Out[19]:
```

|   | 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 | ... | booking_changes | deposit_type |
|---|--------------|-------------|-----------|-------------------|--------------------|--------------------------|---------------------------|-------------------------|----------------------|--------|-----|-----------------|--------------|
| 0 | Resort Hotel | 0           | 342       | 2015              | July               | 27                       | 1                         | 0                       | 0                    | 2      | ... | 3               | No De        |
| 1 | Resort Hotel | 0           | 737       | 2015              | July               | 27                       | 1                         | 0                       | 0                    | 2      | ... | 4               | No De        |
| 2 | Resort Hotel | 0           | 7         | 2015              | July               | 27                       | 1                         | 0                       | 1                    | 1      | ... | 0               | No De        |
| 3 | Resort Hotel | 0           | 13        | 2015              | July               | 27                       | 1                         | 0                       | 1                    | 1      | ... | 0               | No De        |
| 4 | Resort Hotel | 0           | 14        | 2015              | July               | 27                       | 1                         | 0                       | 2                    | 2      | ... | 0               | No De        |

5 rows × 29 columns

```
In [20]: # del hotel_data[is_canceled]

In [21]: # Combining Dates Columns

hotel_data["Arrival Date"] = pd.to_datetime(hotel_data["arrival_date_year"].astype(str) + '/' + hotel_data["arrival_date_month"].astype(str) + '/' + hotel_data["arrival_date_day_of_month"].astype(str))

In [ ]:

In [22]: hotel_data.head(5)

Out[22]:
```

|   | 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 | ... | deposit_type | days_in_wait |
|---|--------------|-------------|-----------|-------------------|--------------------|--------------------------|---------------------------|-------------------------|----------------------|--------|-----|--------------|--------------|
| 0 | Resort Hotel | 0           | 342       | 2015              | July               | 27                       | 1                         | 0                       | 0                    | 2      | ... | No Deposit   |              |
| 1 | Resort Hotel | 0           | 737       | 2015              | July               | 27                       | 1                         | 0                       | 0                    | 2      | ... | No Deposit   |              |
| 2 | Resort Hotel | 0           | 7         | 2015              | July               | 27                       | 1                         | 0                       | 1                    | 1      | ... | No Deposit   |              |
| 3 | Resort Hotel | 0           | 13        | 2015              | July               | 27                       | 1                         | 0                       | 1                    | 1      | ... | No Deposit   |              |
| 4 | Resort Hotel | 0           | 14        | 2015              | July               | 27                       | 1                         | 0                       | 2                    | 2      | ... | No Deposit   |              |

5 rows × 30 columns

```
In [23]: # combining children and babies

hotel_data["children"].describe()

Out[23]:
```

|       |                          |          |
|-------|--------------------------|----------|
| count | 119386                   | 0.000000 |
| mean  | 0.103890                 |          |
| std   | 0.398561                 |          |
| min   | 0.000000                 |          |
| 25%   | 0.000000                 |          |
| 50%   | 0.000000                 |          |
| 75%   | 0.000000                 |          |
| max   | 10.000000                |          |
| Name: | children, dtype: float64 |          |

```
In [24]: np.sum(hotel_data.isnull())

Out[24]:
```

|                                |       |
|--------------------------------|-------|
| 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   |
| 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     |
| Booking Status                 | 0     |
| Arrival Date                   | 0     |
| dtype:                         | int64 |

```
In [25]: new_children = np.where(hotel_data["children"].isnull(),
                                np.mean(hotel_data["children"]),
                                hotel_data["children"])

hotel_data["children"] = new_children

np.sum(hotel_data["children"].isnull())

Out[25]:
```

```
0
```

```
In [26]: hotel_data["Children"] = hotel_data["children"].astype(int) + hotel_data["babies"].astype(int)

In [27]: # hotel cancelled bookings
hotel_data["Booking Status"].value_counts()

Out[27]:
```

|           |                              |
|-----------|------------------------------|
| Cancelled | 75166                        |
| Booked    | 44224                        |
| Name:     | Booking Status, dtype: int64 |

```
In [28]: Cancelled = hotel_data[hotel_data["Booking Status"] == "Cancelled"]
Cancelled_per = (hotel_data["Booking Status"].shape[0]/len(Cancelled))*100

#here I have shown how to use shape[0] to get the total number of rows also use len() to get the total number of rows cancelled

Cancelled_per

Out[28]:
```

```
158.83511161961525
```

```
In [29]: hotel_data["arrival_date_month"].value_counts()

Out[29]:
```

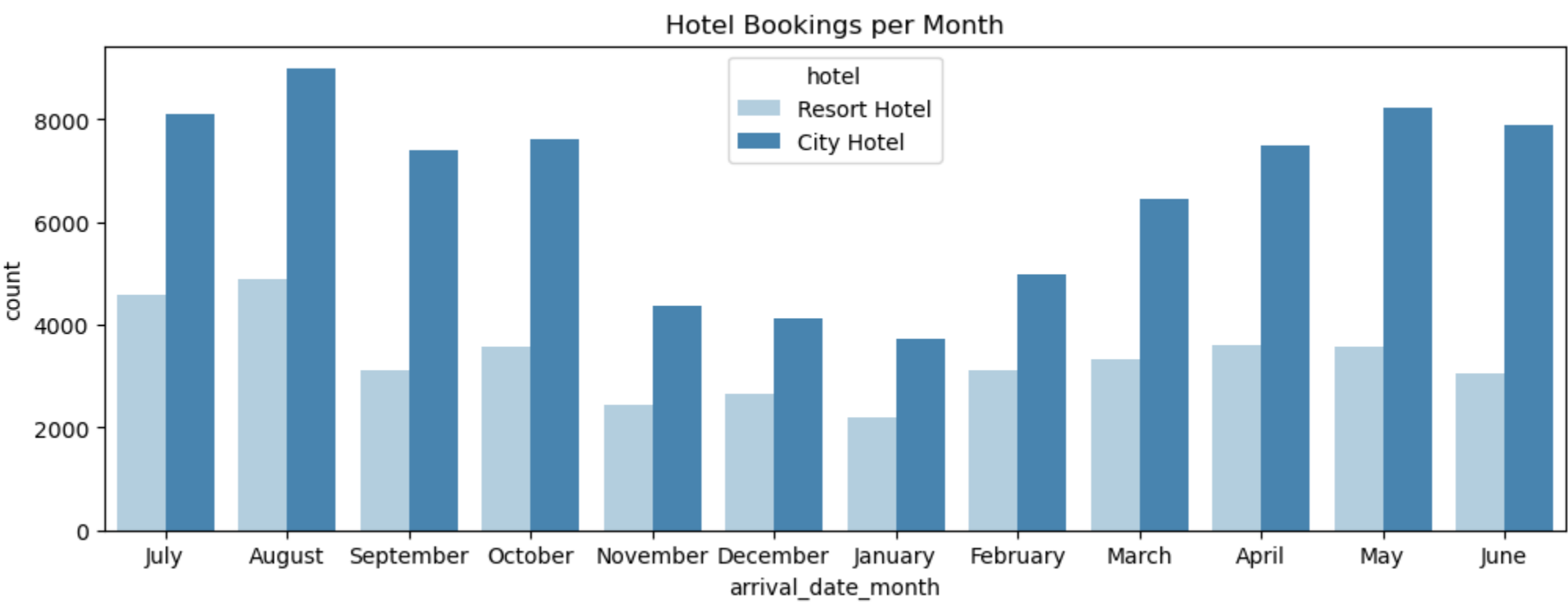
|           |                                  |
|-----------|----------------------------------|
| August    | 13877                            |
| July      | 12661                            |
| May       | 11791                            |
| October   | 11160                            |
| April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

```
In [31]: plt.figure(figsize = (12,4))
sns.countplot(x = 'arrival_date_month', hue = 'hotel', data = hotel_data, palette = 'Blues')

plt.title("Hotel Bookings per Month")

Out[31]:
```

Text(0.5, 1.0, 'Hotel Bookings per Month')




```
In [32]: plt.figure(figsize = (10,5))
sns.countplot(x = 'hotel', data = hotel_data)

plt.title("Hotel Bookings")

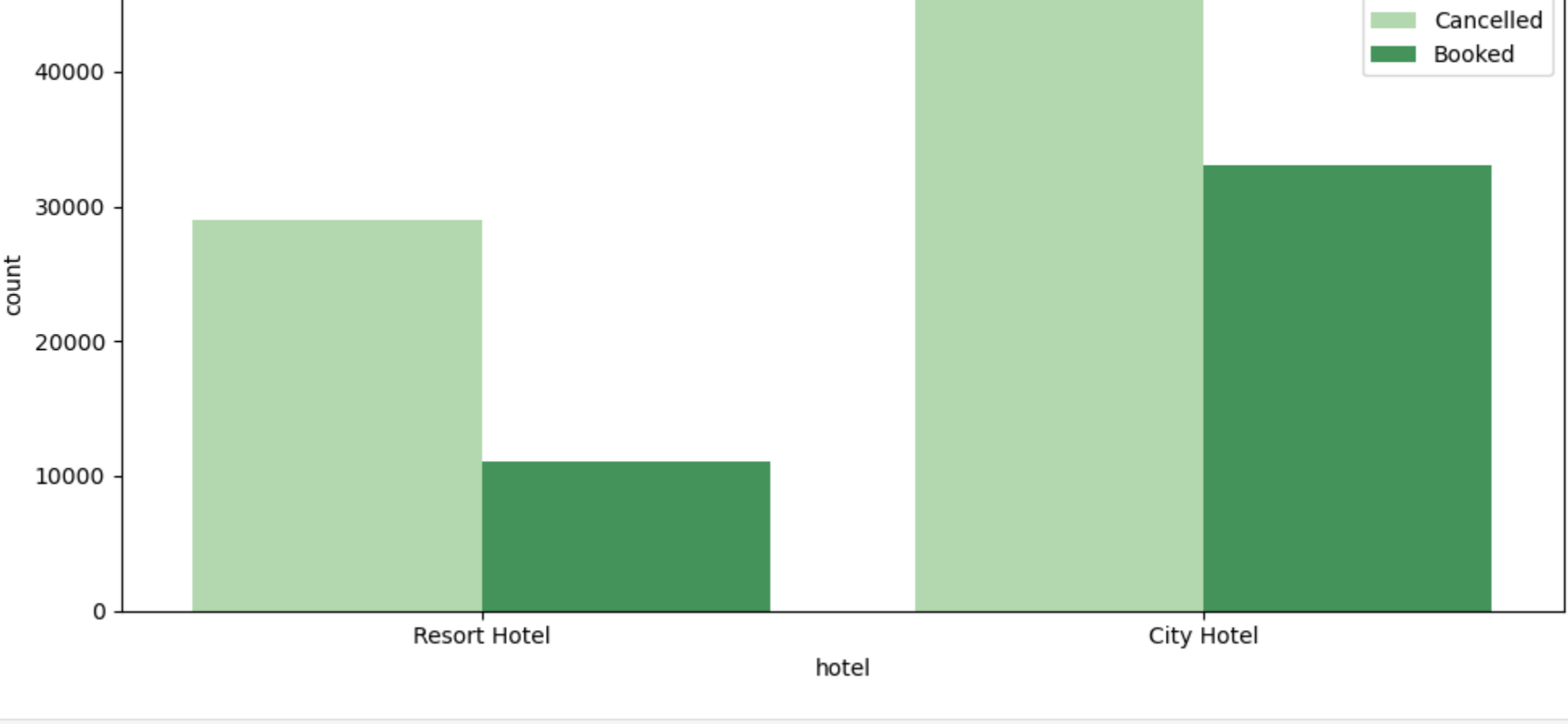
Out[32]:
```

Text(0.5, 1.0, 'Hotel Bookings')



```
In [33]: plt.figure(figsize = (10,5))
sns.countplot(x = 'hotel', data = hotel_data, hue = 'Booking Status', palette = 'Greens')
plt.title("Cancellation vs Confirmed Bookings")
plt.tight_layout()

Out[33]:
```



```
In [42]: # if null values in column is more than 70% of overall rows

for col in hotel_data.columns:
    if np.sum(hotel_data[col].isnull()) > (hotel_data.shape[0]*0.70):
        print("print above 70%")

In [ ]:
```