In [1]:
```
## Load Packages
```

In [2]:
```
import os
import pandas as pd
import numpy as np
import seaborn as sns
from matplotlib import pyplot as plt
```

Read and Understand the Data

In [3]:
```
hotel_data = pd.read_csv('/Users/doyin/Documents/Data Projects/hotel_bookings.csv')
hotel_data.head(5) # check first 5 rows of data
```

Out[3]:

	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	N
1	Resort Hotel	0	737	2015	July	27	1	0	0	2	...	No Deposit	NaN	N
2	Resort Hotel	0	7	2015	July	27	1	0	1	1	...	No Deposit	NaN	N
3	Resort Hotel	0	13	2015	July	27	1	0	1	1	...	No Deposit	304.0	N
4	Resort Hotel	0	14	2015	July	27	1	0	2	2	...	No Deposit	240.0	N

5 rows × 32 columns

In [4]:
```
hotel_data.dtypes # understanding the format of the data
```

Out[4]:
```
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
```

Deciding on Variables

In [10]:
```
# deleting agent and company column

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

In [5]:
```
# looking further into unsure columns to decide if data is useful
hotel_data["market_segment"].describe()
```

Out[5]:
```
count     119390
unique         8
top       Online TA
freq       56477
Name: market_segment, dtype: object
```

In [6]:
```
hotel_data["market_segment"].unique()
```

Out[6]:
```
array(['Direct', 'Corporate', 'Online TA', 'Offline TA/TO',
       'Complementary', 'Groups', 'Undefined', 'Aviation'], dtype=object)
```

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

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

Transforming Variables

In [7]:
```
# cancelled column variable can be changed

hotel_data["is_canceled"].dtype
```

Out[7]:
```
dtype('int64')
```

In [30]:
```
new_is_canceled = pd.Categorical(hotel_data["is_canceled"])
new_is_canceled = new_is_canceled.rename_categories(["Cancelled","Booked"])

new_is_canceled.describe()
```

Out[162]:

	counts	freqs
categories		
Cancelled	75166	0.629584
Booked	44224	0.370416

In [17]:
```
hotel_data["Booking Status"] = new_is_canceled
```

In [18]:
```
hotel_data.head(5)
```

Out[18]:

	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	...	agent	company	days_in_
0	Resort Hotel	0	342	2015	July	27	1	0	0	2	...	NaN	NaN	
1	Resort Hotel	0	737	2015	July	27	1	0	0	2	...	NaN	NaN	
2	Resort Hotel	0	7	2015	July	27	1	0	1	1	...	NaN	NaN	
3	Resort Hotel	0	13	2015	July	27	1	0	1	1	...	304.0	NaN	
4	Resort Hotel	0	14	2015	July	27	1	0	2	2	...	240.0	NaN	

5 rows × 33 columns

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

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[222]:

	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_waiti
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 [28]:
```
# combining children and babies

hotel_data["children"].describe()
```

Out[28]:
```
count    119386.000000
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 [21]:
```
np.sum(hotel_data.isnull())
```

Out[213]:
```
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
Booking Status                       0
Arrival Date                         0
dtype: int64
```

In [22]:
```
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[22]:
```
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 [23]:
```
Cancelled = hotel_data[hotel_data["Booking Status"] == "Cancelled"]
Cancelled_per = (Cancelled.shape[0]/hotel_data.shape[0])*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[23]:
```
150.83551119615525
```

In [24]:
```
hotel_data["arrival_date_month"].value_counts()
```

Out[24]:
```
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
```

Visualising the data

In [25]:
```
plt.figure(figsize = (12,4))
sns.countplot(x = 'arrival_date_month', hue = 'hotel',data = hotel_data, palette = 'Blues')

plt.title("Hotel Bookings per Month")

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

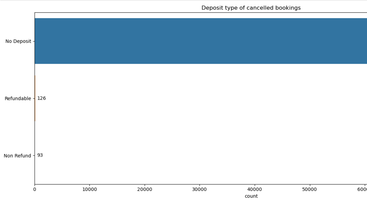
Out[25]:

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

plt.title("Hotel Bookings")

Text(0.5, 1.0, 'Hotel Bookings')
```

Out[32]:



In [33]:
```
plt.figure(figsize = (10,5))

sns.countplot(x = 'hotel', data = hotel_data, hue = 'Booking Status', palette = 'Greens')
plt.title('Cancelation vs Confirmed Bookings')

plt.tight_layout()
```


In [29]:
```
# 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%")
```

```
print above 70%
```

In [30]:
```
fig, ax = plt.subplots(figsize = (14, 6))

sns.countplot(ax = ax,
              x = "deposit_type",
              data = Cancelled,
              order = ['No Deposit', 'Refundable', 'Non Refund'])
ax.bar_label(ax.containers[0], padding = 4)
plt.title("Deposit type of cancelled bookings")

plt.tight_layout()
```


In [ ]:
```

```