In [ ]:
```
## Inspect ##
```

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

In [ ]:
```
df = pd.read_csv('/content/hotel_bookings.csv')
```

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

Out[ ]:
```
(119390, 32)
```

In [ ]:
```
df.sample()
```

Out[ ]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arriva
99305	City Hotel	0	99	2016	October	

1 rows × 32 columns

In [ ]:
```
df.columns
```

Out[ ]:
```
Index(['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', 'children', 'babies', 'meal',
       'country', 'market_segment', 'distribution_channel',
       'is_repeated_guest', 'previous_cancellations',
       'previous_bookings_not_canceled', 'reserved_room_type',
       'assigned_room_type', 'booking_changes', 'deposit_type', 'agent',
       'company', 'days_in_waiting_list', 'customer_type', 'adr',
       'required_car_parking_spaces', 'total_of_special_requests',
       'reservation_status', 'reservation_status_date'],
      dtype='object')
```

In [ ]:
```
df.describe()
```

Out[ ]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number
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 [ ]:
```
df.dtypes
```

Out[ ]:

	0
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 [ ]: ##Cleaning ##
```

```
In [ ]: df[['hotel', 'arrival_date_month', 'meal', 'country', 'market_segment', '
           'reserved_room_type', 'assigned_room_type', 'deposit_type', 'customer_
           'reservation_status', 'reservation_status_date']] = df[['hotel', 'arri
                                                                  'market_segment'
           'reserved_room_type', 'assigned_room_type', 'deposit_type', 'customer_
           'reservation_status', 'reservation_status_date']].astype('string')
```

```
In [ ]: df.dtypes
```

Out[ ]:

	0
hotel	string[python]
is_canceled	int64
lead_time	int64
arrival_date_year	int64
arrival_date_month	string[python]
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	string[python]
country	string[python]
market_segment	string[python]
distribution_channel	string[python]
is_repeated_guest	int64
previous_cancellations	int64
previous_bookings_not_canceled	int64
reserved_room_type	string[python]
assigned_room_type	string[python]
booking_changes	int64
deposit_type	string[python]
agent	float64
company	float64
days_in_waiting_list	int64
customer_type	string[python]
adr	float64
required_car_parking_spaces	int64
total_of_special_requests	int64
reservation_status	string[python]
reservation_status_date	string[python]

dtype: object

In [ ]:

```
df.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

```
In [ ]: df.duplicated().sum()
```

```
Out[ ]: np.int64(31994)
```

```
In [ ]: df = df.drop_duplicates()
        display(df.shape)
```

```
(87396, 32)
```

```
In [ ]: df['company'].isnull().sum()
```

```
Out[ ]: np.int64(82137)
```

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

```
Out[ ]: (87396, 32)
```

```
In [ ]: df = df.dropna(subset=['children', 'country'])
```

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

```
Out[ ]: (86940, 32)
```

```
In [ ]: df = df.fillna('Not_Available')
```

```
In [ ]: df.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
                          agent       0
                        company       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 [ ]: df.dtypes
```

Out[ ]:

	0
hotel	string[python]
is_canceled	int64
lead_time	int64
arrival_date_year	int64
arrival_date_month	string[python]
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	string[python]
country	string[python]
market_segment	string[python]
distribution_channel	string[python]
is_repeated_guest	int64
previous_cancellations	int64
previous_bookings_not_canceled	int64
reserved_room_type	string[python]
assigned_room_type	string[python]
booking_changes	int64
deposit_type	string[python]
agent	object
company	object
days_in_waiting_list	int64
customer_type	string[python]
adr	float64
required_car_parking_spaces	int64
total_of_special_requests	int64
reservation_status	string[python]
reservation_status_date	string[python]

dtype: object

In [ ]:
```
## Visualisation ##
```

In [ ]:
```
df.columns
```

Out[ ]:
```
Index(['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', 'children', 'babies', 'meal',
       'country', 'market_segment', 'distribution_channel',
       'is_repeated_guest', 'previous_cancellations',
       'previous_bookings_not_canceled', 'reserved_room_type',
       'assigned_room_type', 'booking_changes', 'deposit_type', 'agent',
       'company', 'days_in_waiting_list', 'customer_type', 'adr',
       'required_car_parking_spaces', 'total_of_special_requests',
       'reservation_status', 'reservation_status_date'],
      dtype='object')
```

In [ ]:
```
df['adults'].unique()
```

Out[ ]:
```
array([ 2,  1,  3,  4, 40, 26, 50, 27, 55,  0, 20,  6,  5, 10])
```

In [ ]:
```
df['country'].nunique()
```

Out[ ]:
```
177
```

In [ ]:
```
df['hotel'].unique()
```

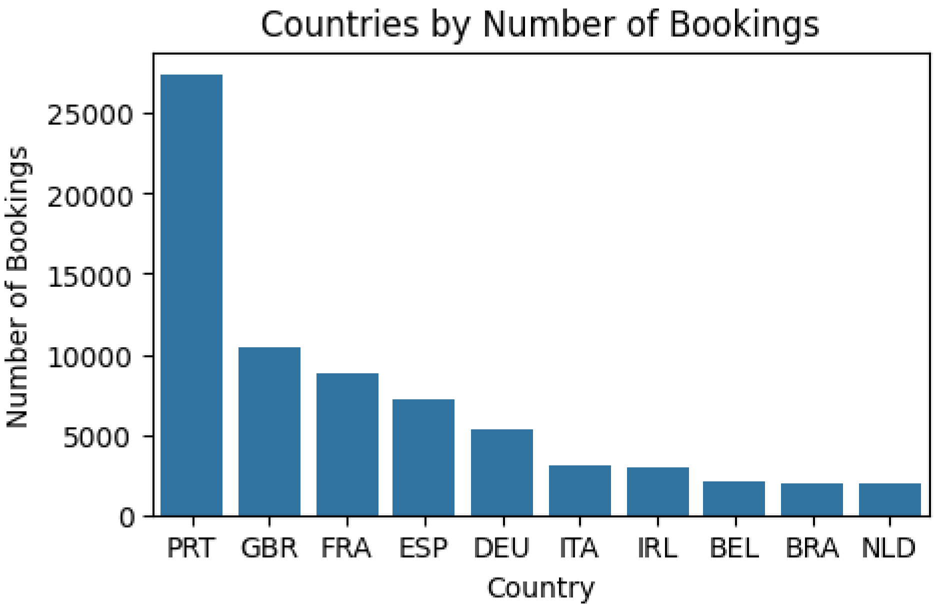
Out[ ]:
```
<StringArray>
['Resort Hotel', 'City Hotel']
Length: 2, dtype: string
```

In [ ]:
```
df['meal'].unique()
```

Out[ ]:
```
<StringArray>
['BB', 'FB', 'HB', 'SC', 'Undefined']
Length: 5, dtype: string
```

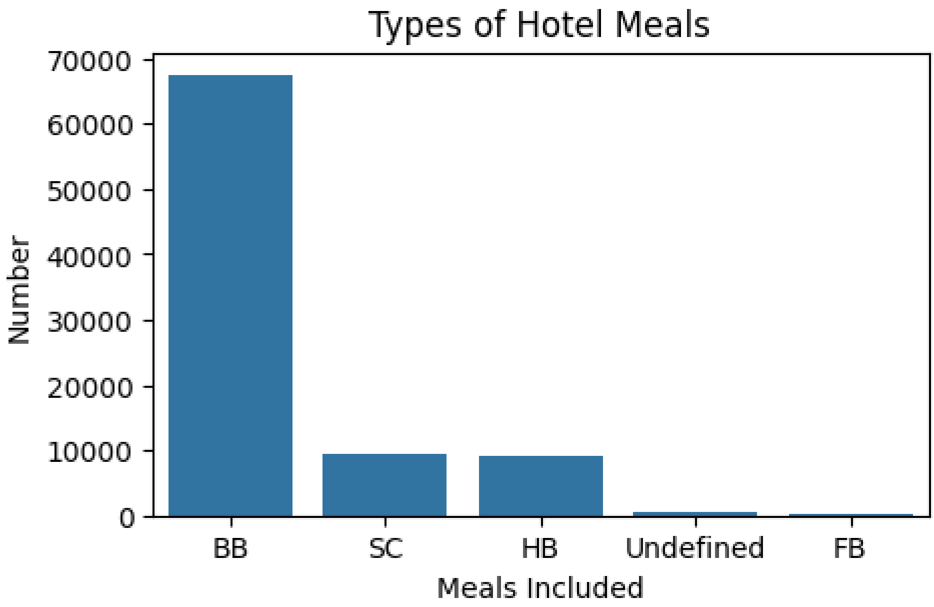
In [ ]:
```
df['hotel'].unique()
```

In [ ]:
```
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(5, 3))
country_counts = df['country'].value_counts().head(10)
sns.barplot(x=country_counts.index, y=country_counts.values)
plt.title('Countries by Number of Bookings')
plt.xlabel('Country')
plt.ylabel('Number of Bookings')
plt.xticks(rotation=0)
plt.show()
```



In [ ]:

```
plt.figure(figsize=(5, 3))
country_counts = df['meal'].value_counts()
sns.barplot(x=country_counts.index, y=country_counts.values)
plt.title('Types of Hotel Meals')
plt.xlabel('Meals Included')
plt.ylabel('Number')
plt.xticks(rotation=0)
plt.show()
```



In [ ]:

```
display(numerical_columns)
```

```
['is_canceled',
 'lead_time',
 'arrival_date_year',
 'arrival_date_week_number',
 'arrival_date_day_of_month',
 'stays_in_weekend_nights',
 'stays_in_week_nights',
 'adults',
 'children',
 'babies',
 'is_repeated_guest',
 'previous_cancellations',
 'previous_bookings_not_canceled',
 'booking_changes',
 'days_in_waiting_list',
 'adr',
 'required_car_parking_spaces',
 'total_of_special_requests']
```

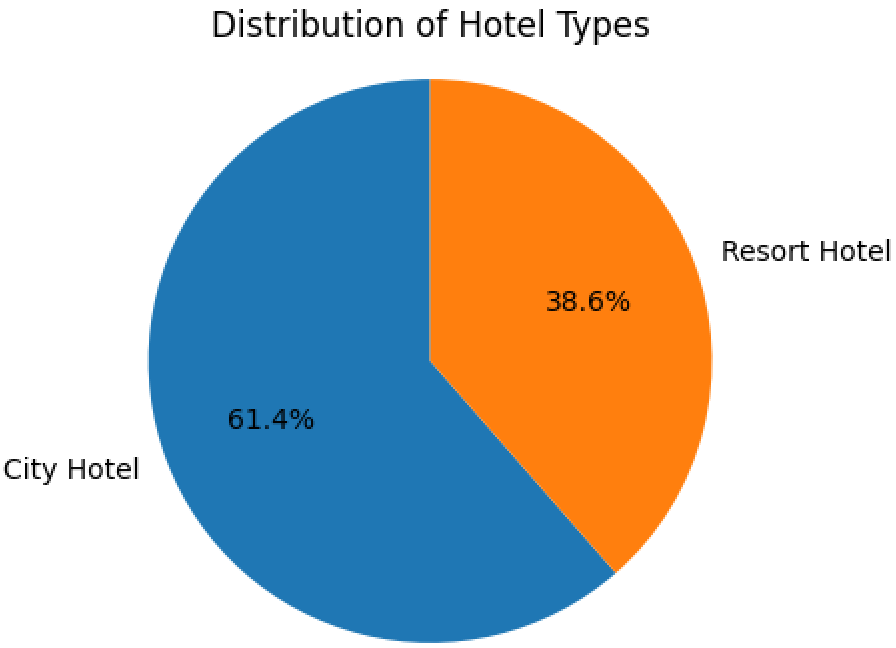
In [ ]:
```
df['is_repeated_guest'].unique()
```

Out[ ]:
```
array([0, 1])
```

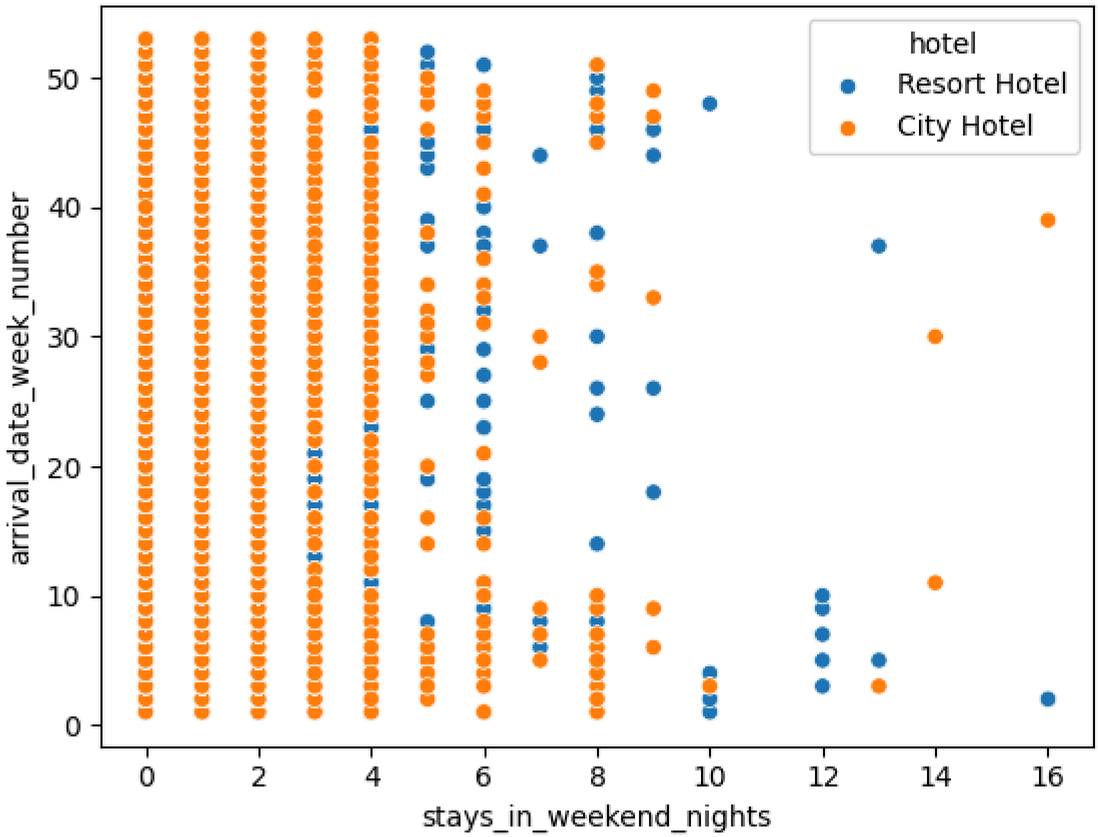
In [ ]:
```
binary_columns = df.columns[df.nunique() == 2]
print(binary_columns)
```

```
Index(['hotel', 'is_canceled', 'is_repeated_guest'], dtype='object')
```

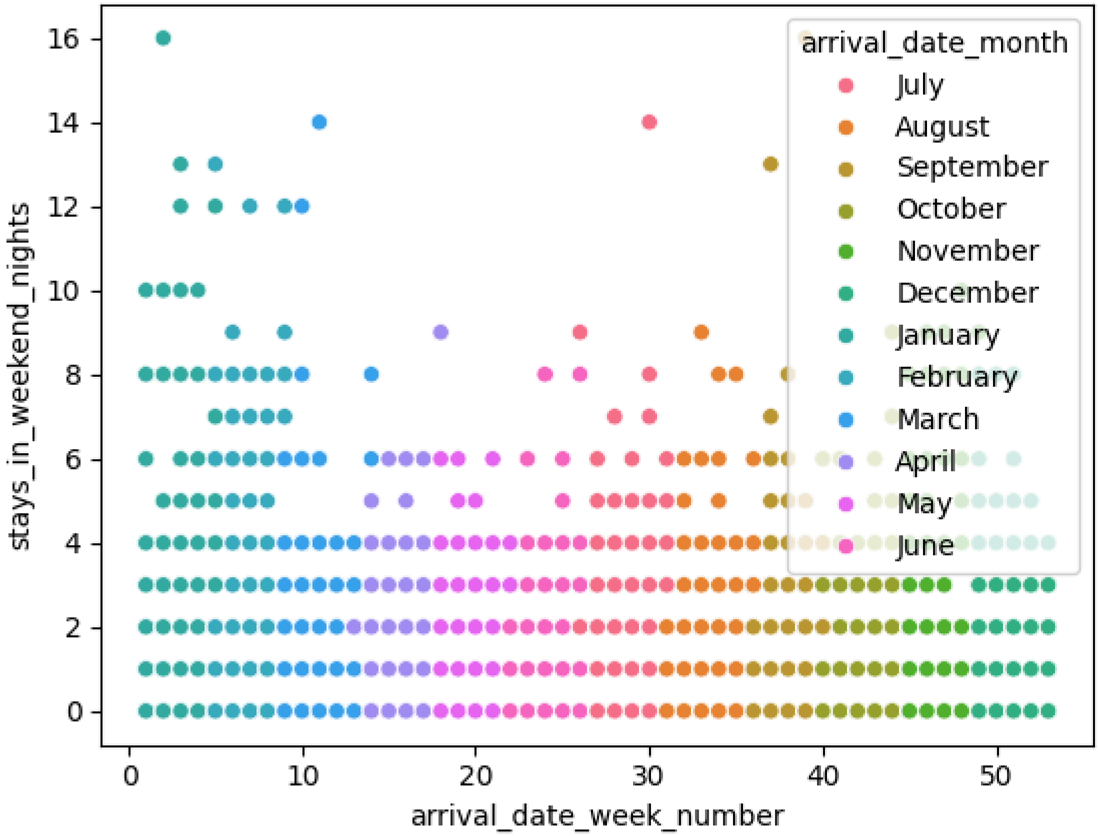
In [ ]:
```
import matplotlib.pyplot as plt
hotel_counts = df['hotel'].value_counts()

plt.figure(figsize=(4, 4))
plt.pie(hotel_counts, labels=hotel_counts.index, autopct='%1.1f%%', start
plt.title('Distribution of Hotel Types')
plt.axis('equal')
plt.show()
```



In [ ]:
```
sns.scatterplot(data=df, x="stays_in_weekend_nights", y="arrival_date_wee
plt.show()
```



```
In [ ]: sns.scatterplot(data=df, x="arrival_date_week_number", y="stays_in_weeken
        plt.show()
```


arrival_date_month
July
August
September
October
November
December
January
February
March
April
May
June
stays_in_weekend_nights
arrival_date_week_number