Data Preprocessing and extracting insights

Load the dataset

```
import pandas as pd
hotel_bookings = pd.read_csv("hotel_bookings.csv")
```

Handling Missing Values

```
hotel_bookings['children'].fillna(0, inplace=True)
hotel_bookings['country'].fillna(hotel_bookings['country'].mode()[0], inplace=True)
hotel_bookings['agent'].fillna(0, inplace=True)
hotel_bookings['company'].fillna(0, inplace=True)
hotel_bookings
```

	index	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_
0	0	Resort Hotel	0	342	2015	July	27	1	
1	1	Resort Hotel	0	737	2015	July	27	1	
2	2	Resort Hotel	0	7	2015	July	27	1	
3	3	Resort Hotel	0	13	2015	July	27	1	
4	4	Resort Hotel	0	14	2015	July	27	1	
...	...	...	...	...	...	...	...	...	
36518	36518	Resort Hotel	0	219	2017	May	19	13	
36519	36519	Resort Hotel	0	195	2017	May	20	16	
36520	36520	Resort Hotel	0	154	2017	May	20	16	
36521	36521	Resort Hotel	0	0	2017	May	20	20	
36522	36522	Resort Hotel	0	118	2017	May	20	16	

36523 rows × 33 columns

Convert reservation_status_date to datetime

```
hotel_bookings['reservation_status_date'] = pd.to_datetime(hotel_bookings['reservation_status_date'])
hotel_bookings
```

	index	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_
0	0	Resort Hotel	0	342	2015	July	27	1	
1	1	Resort Hotel	0	737	2015	July	27	1	
2	2	Resort Hotel	0	7	2015	July	27	1	
3	3	Resort Hotel	0	13	2015	July	27	1	
4	4	Resort Hotel	0	14	2015	July	27	1	
...	...	...	...	...	...	...	...	...	
36518	36518	Resort Hotel	0	219	2017	May	19	13	
36519	36519	Resort Hotel	0	195	2017	May	20	16	
36520	36520	Resort Hotel	0	154	2017	May	20	16	
36521	36521	Resort Hotel	0	0	2017	May	20	20	
36522	36522	Resort Hotel	0	118	2017	May	20	16	

36523 rows × 33 columns

Create total_stay_nights feature

```
hotel_bookings['total_stay_nights'] = hotel_bookings['stays_in_weekend_nights'] + hotel_bookings['stays_in_week_nights']
hotel_bookings
```

	index	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_
0	0	Resort Hotel	0	342	2015	July	27	1	
1	1	Resort Hotel	0	737	2015	July	27	1	
2	2	Resort Hotel	0	7	2015	July	27	1	
3	3	Resort Hotel	0	13	2015	July	27	1	
4	4	Resort Hotel	0	14	2015	July	27	1	
...	...	...	...	...	...	...	...	...	
36518	36518	Resort Hotel	0	219	2017	May	19	13	
36519	36519	Resort Hotel	0	195	2017	May	20	16	
36520	36520	Resort Hotel	0	154	2017	May	20	16	
36521	36521	Resort Hotel	0	0	2017	May	20	20	
36522	36522	Resort Hotel	0	118	2017	May	20	16	

36523 rows × 34 columns

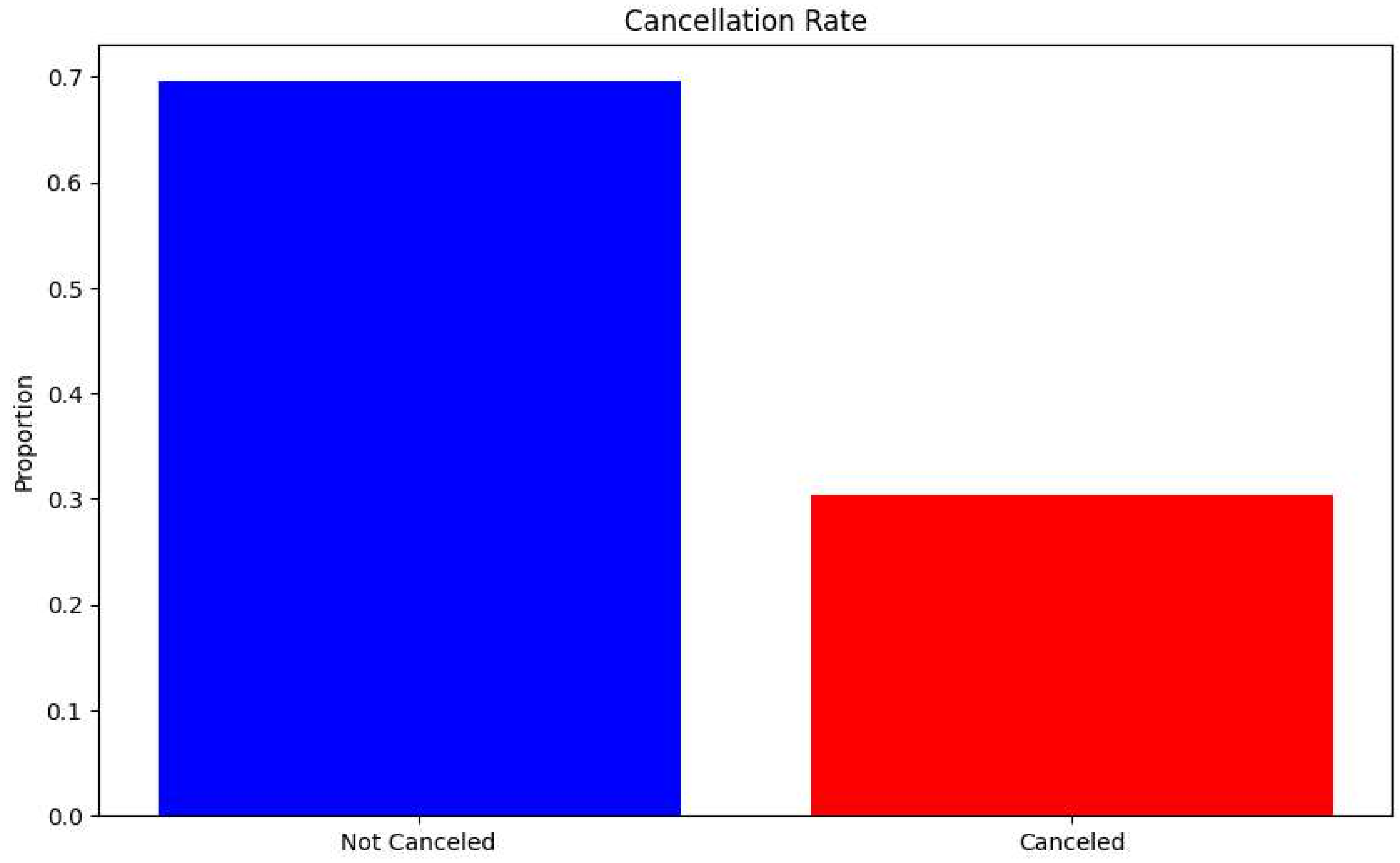
Insights Extraction and Visualization

Analyze cancellation rate

```
cancellation_rate = hotel_bookings['is_canceled'].mean()
cancellation_rate
```

```
0.30449305916819536
```

```
import matplotlib.pyplot as plt
plt.figure(figsize=(10, 6))
plt.bar(['Not Canceled', 'Canceled'], [1 - cancellation_rate, cancellation_rate], color=['blue', 'red'])
plt.title('Cancellation Rate')
plt.ylabel('Proportion')
plt.show()
```

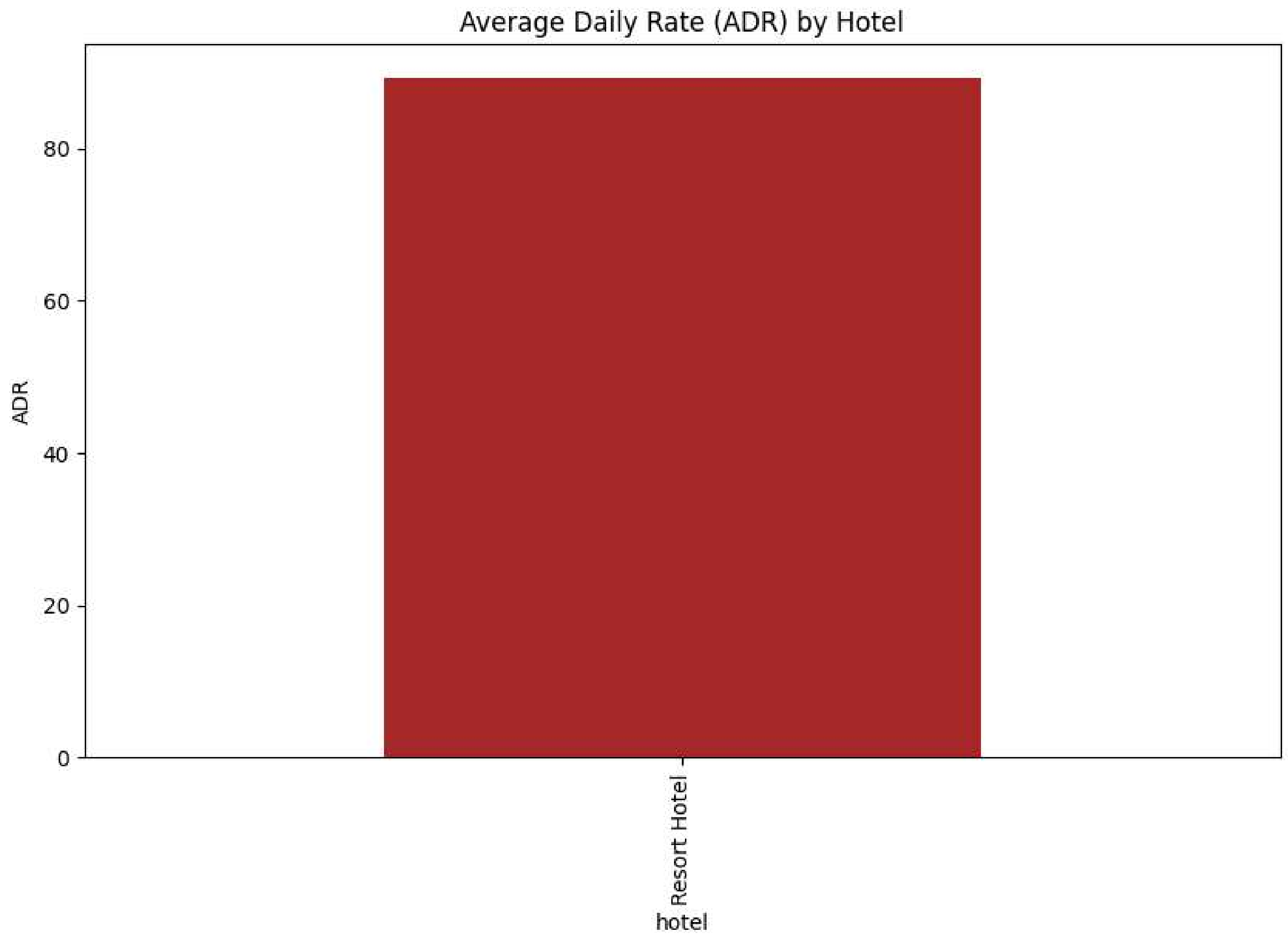


Average daily rate (ADR) for different hotels

```
adr_by_hotel = hotel_bookings.groupby('hotel')['adr'].mean()
adr_by_hotel
```

```
hotel
Resort Hotel    89.234286
Name: adr, dtype: float64
```

```
plt.figure(figsize=(10, 6))
adr_by_hotel.plot(kind='bar', color=['brown', 'orange'])
plt.title('Average Daily Rate (ADR) by Hotel')
plt.ylabel('ADR')
plt.show()
```

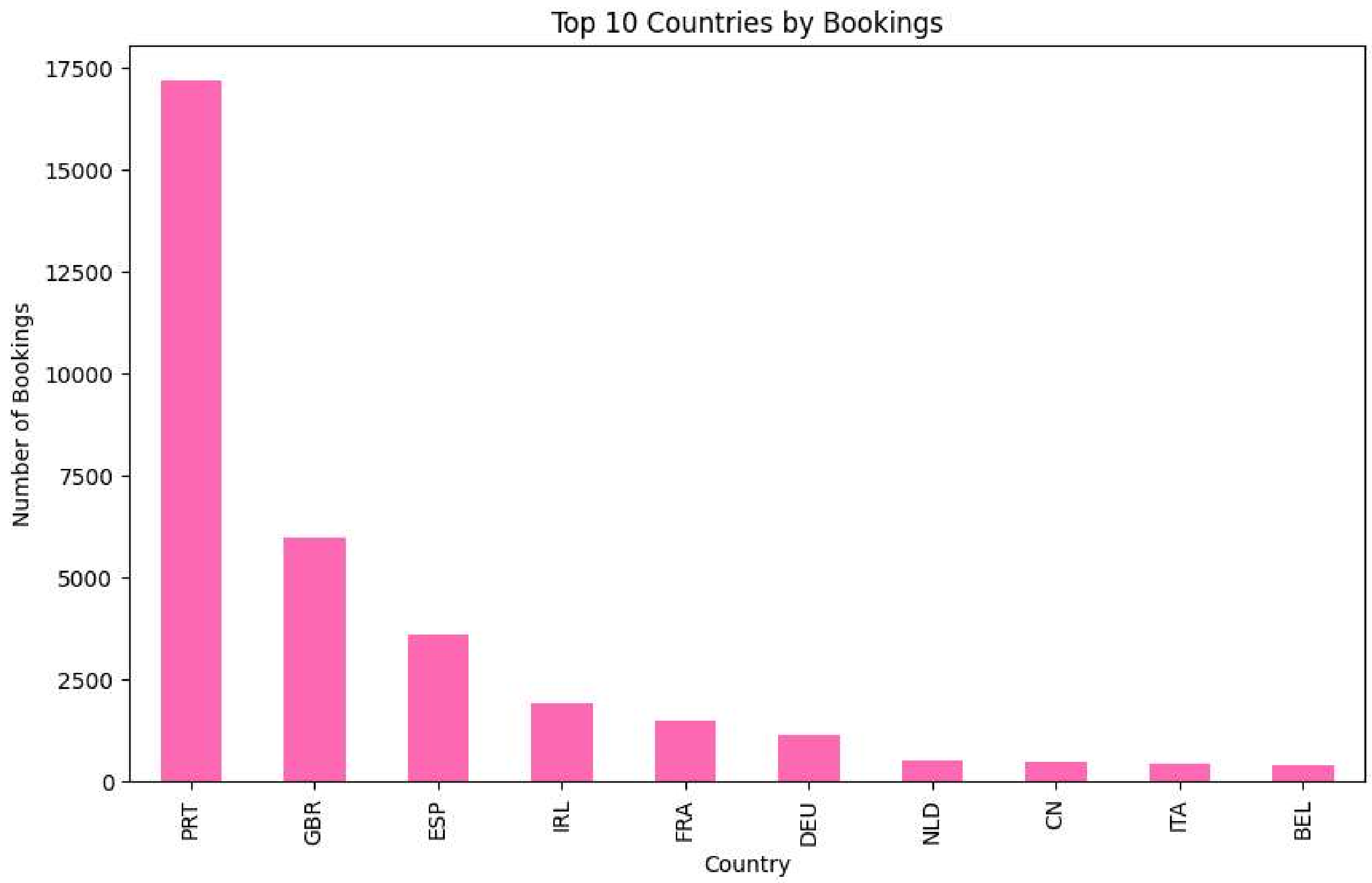


Distribution of bookings by country

```
bookings_by_country = hotel_bookings['country'].value_counts().head(10)
bookings_by_country
```

```
country
PRT    17194
GBR     5972
ESP     3577
IRL     1917
FRA     1489
DEU     1125
NLD      487
CN       473
ITA      413
BEL      376
Name: count, dtype: int64
```

```
plt.figure(figsize=(10, 6))
bookings_by_country.plot(kind='bar', color='hotpink')
plt.title('Top 10 Countries by Bookings')
plt.ylabel('Number of Bookings')
plt.xlabel('Country')
plt.show()
```

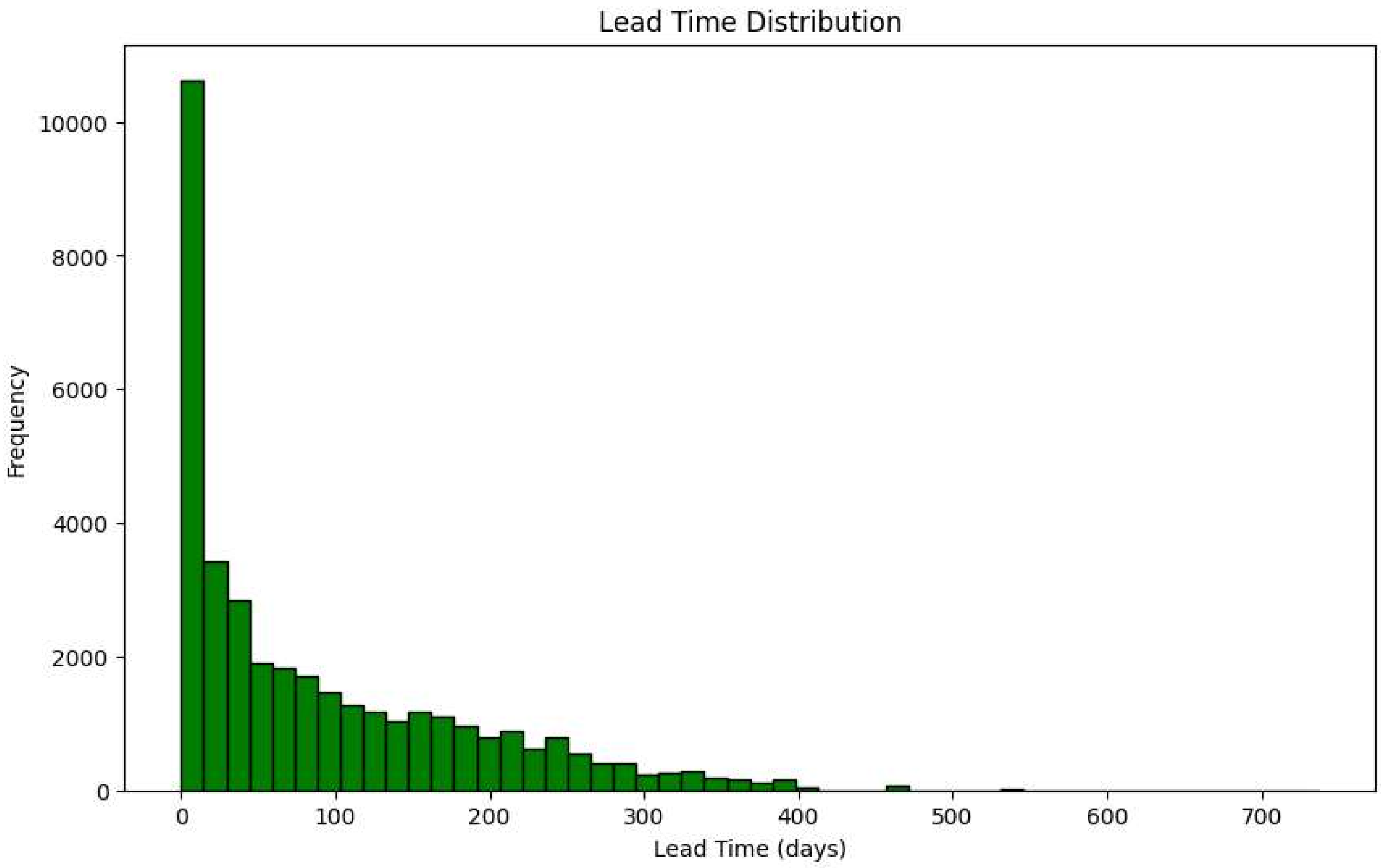


Lead time distribution

```
lead_time_distribution = hotel_bookings['lead_time'].describe()
lead_time_distribution
```

```
count    36523.000000
mean        89.889357
std         96.240146
min          0.000000
25%         10.000000
50%         54.000000
75%        149.000000
max        737.000000
Name: lead_time, dtype: float64
```

```
plt.figure(figsize=(10, 6))
plt.hist(hotel_bookings['lead_time'], bins=50, color='green', edgecolor='black')
plt.title('Lead Time Distribution')
plt.xlabel('Lead Time (days)')
plt.ylabel('Frequency')
plt.show()
```


Lead Time Distribution
Frequency
Lead Time (days)