In [13]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import nltk
import seaborn as sns
from sklearn.model_selection import train_test_split
from bs4 import BeautifulSoup
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize
```

In [14]:
```
hotel_reviews = pd.read_csv('hotel_reviews.csv')
hotel_reviews.head()
```

Out[14]:

	Index	Name	Area	Review_Date	Rating_attribute	Rating(Out of 10)	Review_Text
0	0	Hotel The Pearl	Paharganj, New Delhi	Jul-23	Best budget friendly hotel	9.0	Hotel the pearl is perfect place to stay in De...
1	1	Hotel The Pearl	Paharganj, New Delhi	Aug-23	Amazing place	9.0	Location of the hotel is perfect. The hotel is...
2	2	Hotel The Pearl	Paharganj, New Delhi	Aug-23	Overall good stay. Economic.	9.0	Location, Indian food.
3	3	Hotel The Pearl	Paharganj, New Delhi	Aug-23	Lovely	9.0	The location and the hotel itself is great. Ne...
4	4	Hotel The Pearl	Paharganj, New Delhi	Aug-23	Great hotel Great staff and great staying	9.0	Friendly and smiling staffs.. The reception st...

Index: An identifier for each review.

Name: The name of the hotel.

Area: The area where the hotel is located.

Review_Date: The date when the review was posted.

Rating_attribute: A short description/title for the review.

Rating(Out of 10): The rating given by the user on a scale of 1 to 10.

Review_Text: The detailed review text.

Data Cleaning and Type Conversion

In [15]:
```
hotel_reviews['Review_Date'] = pd.to_datetime(hotel_reviews['Review_Date'], errors='coer
hotel_reviews['Review_Date'].head()
```

Out[15]:
```
0   2023-07-01
1   2023-08-01
2   2023-08-01
3   2023-08-01
4   2023-08-01
Name: Review_Date, dtype: datetime64[ns]
```

In [16]:
```
# Check for any duplicated rows
```

```
duplicated_rows = hotel_reviews.duplicated()
duplicated_rows.sum()
```

Out[16]:
```
0
```

In [17]:
```
invalid_ratings = hotel_reviews['Rating(Out of 10)'][(hotel_reviews['Rating(Out of 10)']
invalid_ratings.count()
```

Out[17]:
```
0
```

In [18]:
```
missing_values= hotel_reviews.isnull().sum()
missing_values
```

Out[18]:
```
Index                0
Name                 0
Area                 0
Review_Date          0
Rating_attribute     0
Rating(Out of 10)    0
Review_Text          7
dtype: int64
```

In [11]:
```
#drop missing values
hotel_reviews.dropna(subset=['Review_Text'], inplace=True)

missing_values
```

Out[11]:
```
Index                0
Name                 0
Area                 0
Review_Date          0
Rating_attribute     0
Rating(Out of 10)    0
Review_Text          0
dtype: int64
```

In [19]:
```
data_types = hotel_reviews.dtypes
data_types
```

Out[19]:
```
Index                          int64
Name                          object
Area                          object
Review_Date          datetime64[ns]
Rating_attribute              object
Rating(Out of 10)            float64
Review_Text                   object
dtype: object
```

In [21]:
```
# Check the number of remaining rows
remaining_rows = hotel_reviews.shape[0]
remaining_rows
```

Out[21]:
```
7001
```

In [23]:
```
# Feature Transformation: Extract year and month from 'Review_Date'
hotel_reviews['Review_Year'] = hotel_reviews['Review_Date'].dt.year
hotel_reviews['Review_Month'] = hotel_reviews['Review_Date'].dt.month
hotel_reviews.head()
```

Out[23]:

	Index	Name	Area	Review_Date	Rating_attribute	Rating(Out of 10)	Review_Text	Review_Year	Review_Month
0	0	Hotel The Pearl	Paharganj, New Delhi	2023-07-01	Best budget friendly hotel	9.0	Hotel the pearl is perfect place	2023	7

							to stay in De...		
1	1	Hotel The Pearl	Paharganj, New Delhi	2023-08-01	Amazing place	9.0	Location of the hotel is perfect. The hotel is...	2023	8
2	2	Hotel The Pearl	Paharganj, New Delhi	2023-08-01	Overall good stay. Economic.	9.0	Location, Indian food.	2023	8
3	3	Hotel The Pearl	Paharganj, New Delhi	2023-08-01	Lovely	9.0	The location and the hotel itself is great. Ne...	2023	8
4	4	Hotel The Pearl	Paharganj, New Delhi	2023-08-01	Great hotel Great staff and great staying	9.0	Friendly and smiling staffs.. The reception st...	2023	8

Exploratory Data Analysis (EDA):

In [25]:
```
# Descriptive Statistics

# Summary statistics for numerical columns
summary_statistics = hotel_reviews.describe(include=[float, int])
summary_statistics
```

Out[25]:

	Index	Rating(Out of 10)	Review_Year	Review_Month
count	7001.00000	7001.000000	7001.000000	7001.000000
mean	3500.00000	7.030981	2022.760320	6.298386
std	2021.15895	2.882846	0.525876	2.505812
min	0.00000	1.000000	2020.000000	1.000000
25%	1750.00000	6.000000	2023.000000	5.000000
50%	3500.00000	8.000000	2023.000000	7.000000
75%	5250.00000	9.000000	2023.000000	8.000000
max	7000.00000	10.000000	2023.000000	12.000000

Numerical Columns: Index: Ranges from 0 to 7000, which seems to be just a unique identifier for each review.

Rating(Out of 10): Ranges from 1 to 10.

The mean rating is approximately 7.03, indicating a generally positive trend in the reviews.

The standard deviation is approximately 2.88, showing a moderate spread of ratings.

Review_Year: Ranges from 2020 to 2023, indicating that the dataset contains reviews from these years.

Review_Month: Ranges from 1 to 12, representing all months in a year.

In [26]:
```
# Distribution of ratings
rating_distribution = hotel_reviews['Rating(Out of 10)'].value_counts(normalize=True).so
```

```
rating_distribution
```

Out[26]:

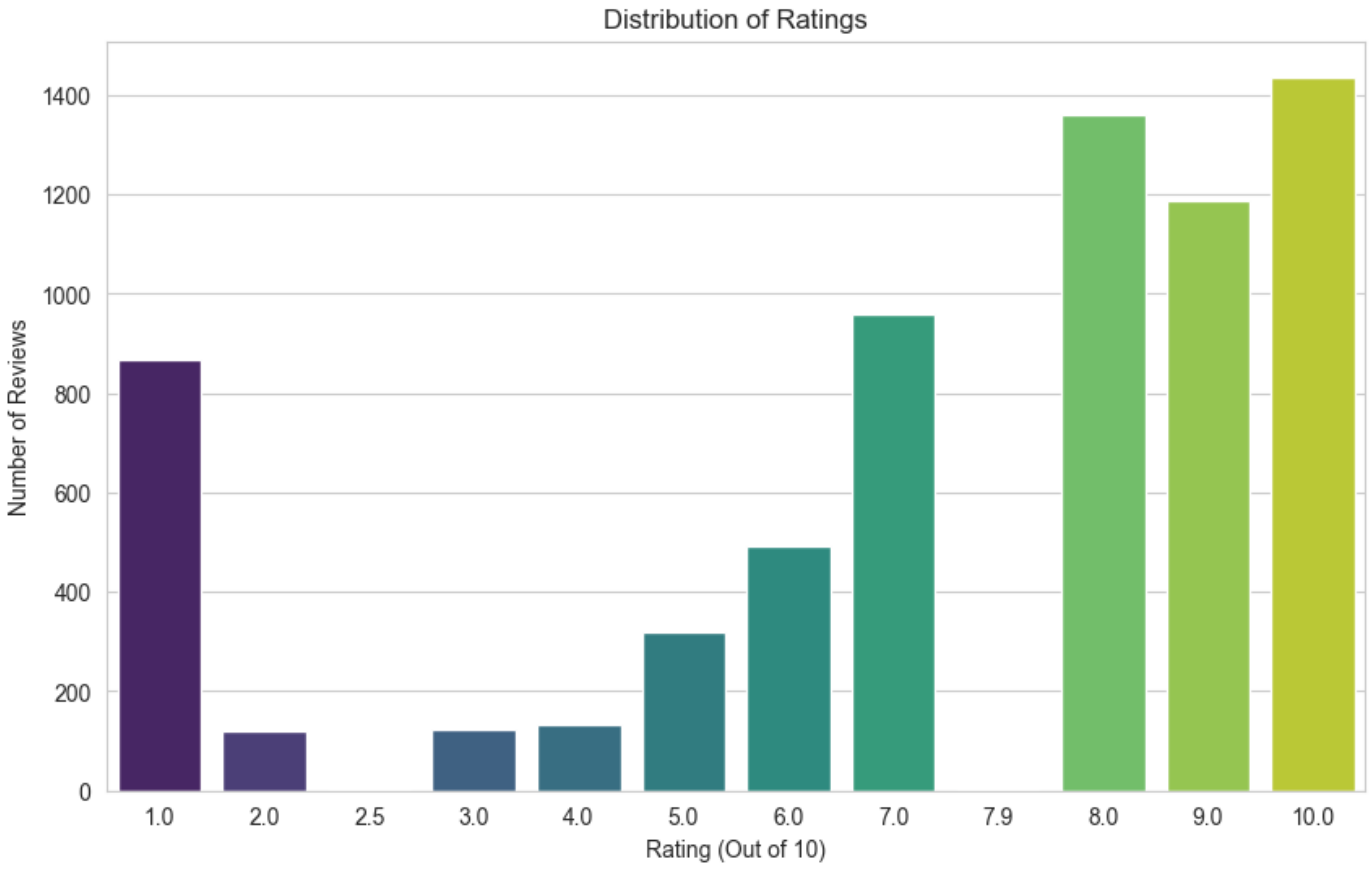
```
Rating(Out of 10)
1.0     0.123697
2.0     0.016998
2.5     0.000286
3.0     0.017712
4.0     0.018997
5.0     0.045708
6.0     0.070419
7.0     0.136980
7.9     0.000143
8.0     0.194258
9.0     0.169690
10.0    0.205114
Name: proportion, dtype: float64
```

Rating Distribution: The ratings are skewed towards the higher end, with the majority of ratings being 8, 9, or 10. The lowest ratings (1 and 2) make up a smaller portion of the dataset.

In [29]:

```
# 1. Rating Distribution
plt.figure(figsize=(10, 6))
sns.countplot(x='Rating(Out of 10)', data=hotel_reviews, palette='viridis')
plt.title('Distribution of Ratings')
plt.xlabel('Rating (Out of 10)')
plt.ylabel('Number of Reviews')
plt.show()
```

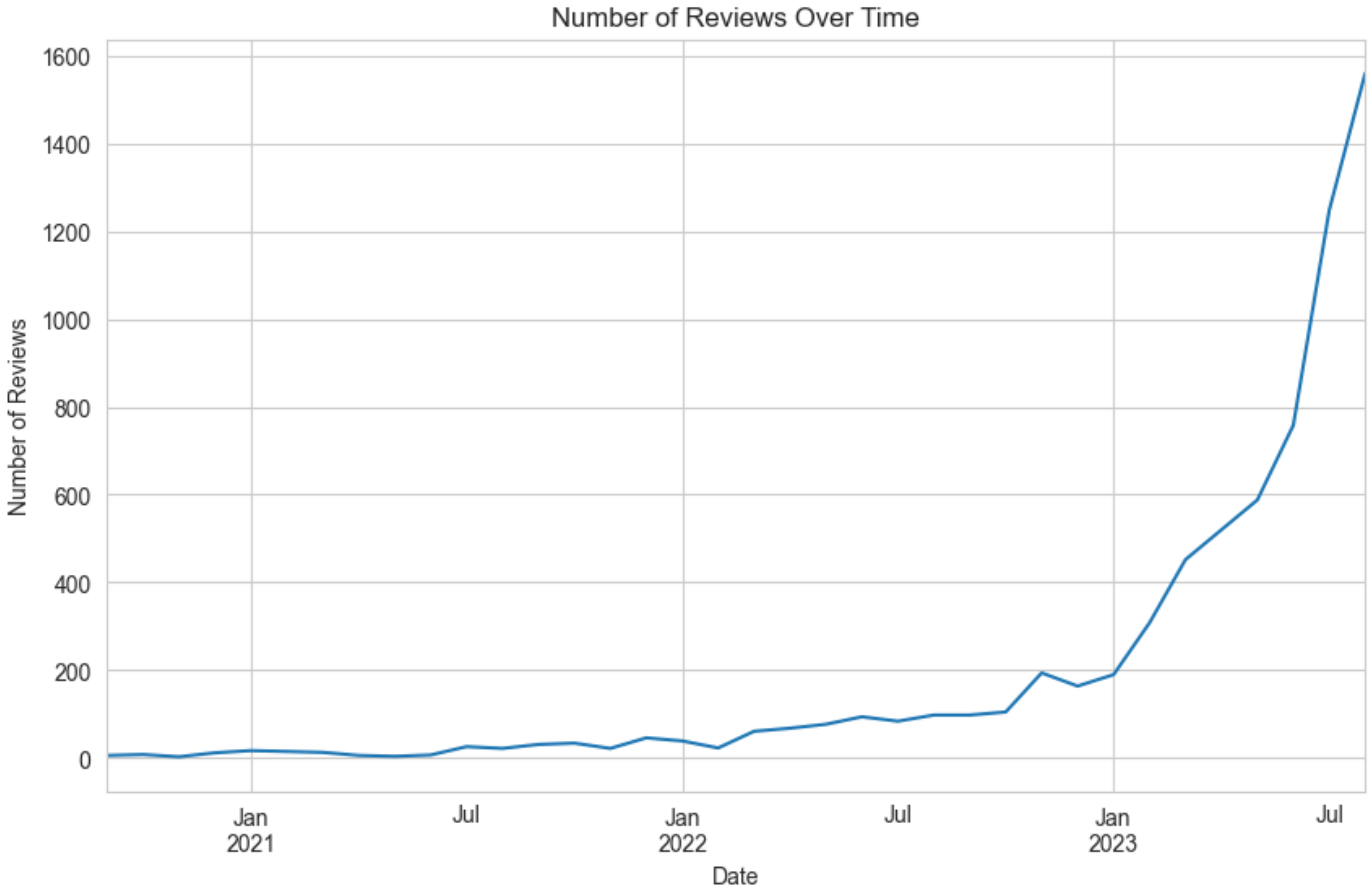
```
C:\Users\Majid\AppData\Local\Temp\ipykernel_10772\2847970913.py:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0.
Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.countplot(x='Rating(Out of 10)', data=hotel_reviews, palette='viridis')
```



The majority of reviews have high ratings, especially 8, 9, and 10, indicating overall positive experiences. There are relatively few reviews with low ratings (1 and 2). The distribution is skewed towards the higher

end, reflecting a tendency of guests to leave positive reviews.

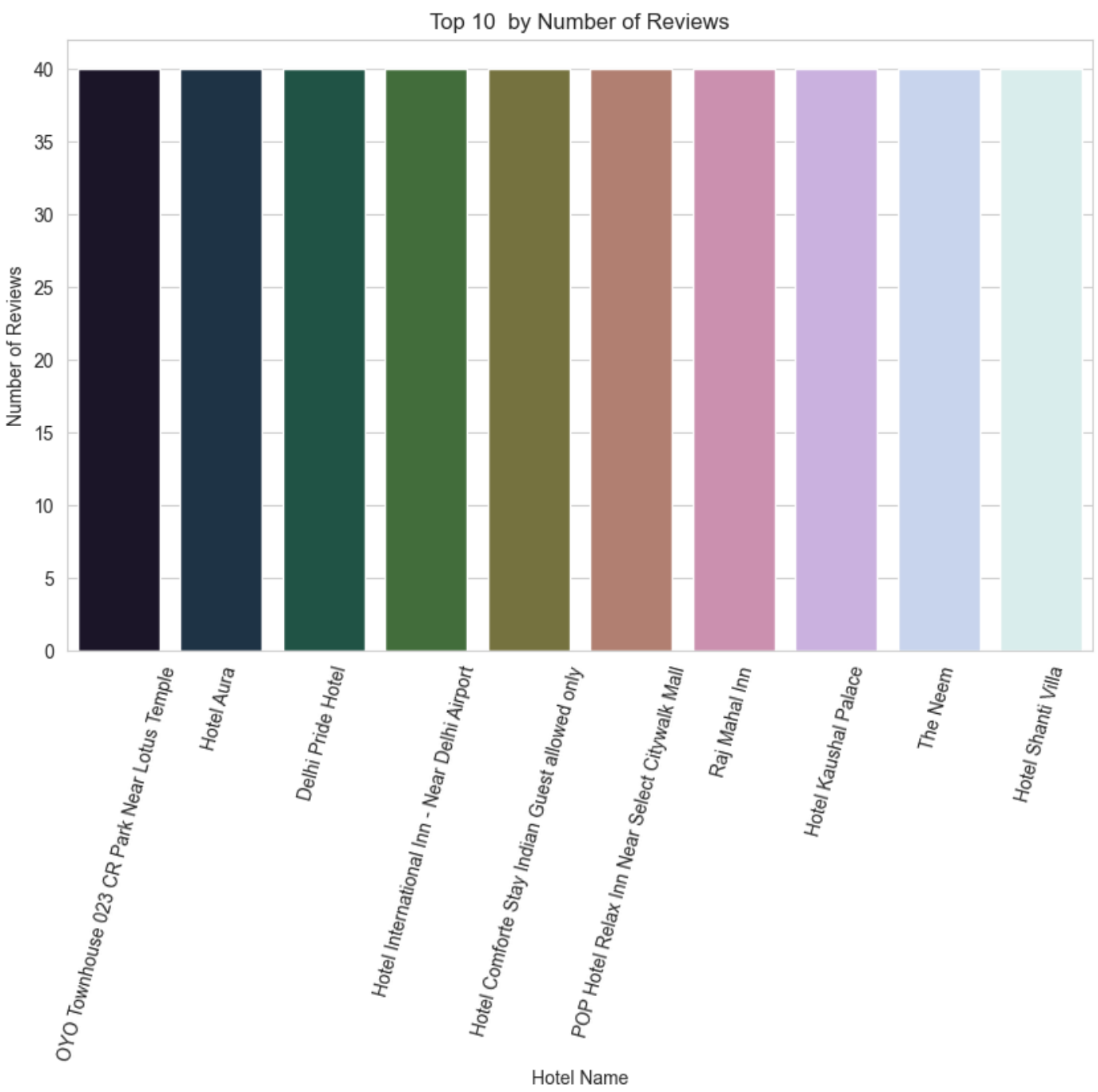
In [30]:
```
#Number of reviews over time
plt.figure(figsize=(10, 6))
hotel_reviews.set_index('Review_Date').resample('M').size().plot()
plt.title('Number of Reviews Over Time')
plt.xlabel('Date')
plt.ylabel('Number of Reviews')
plt.show()
```



There is a noticeable trend of increasing activity over time, with more reviews being posted in recent months. There are fluctuations in the number of reviews from month to month, which could be due to seasonal variations or specific events.

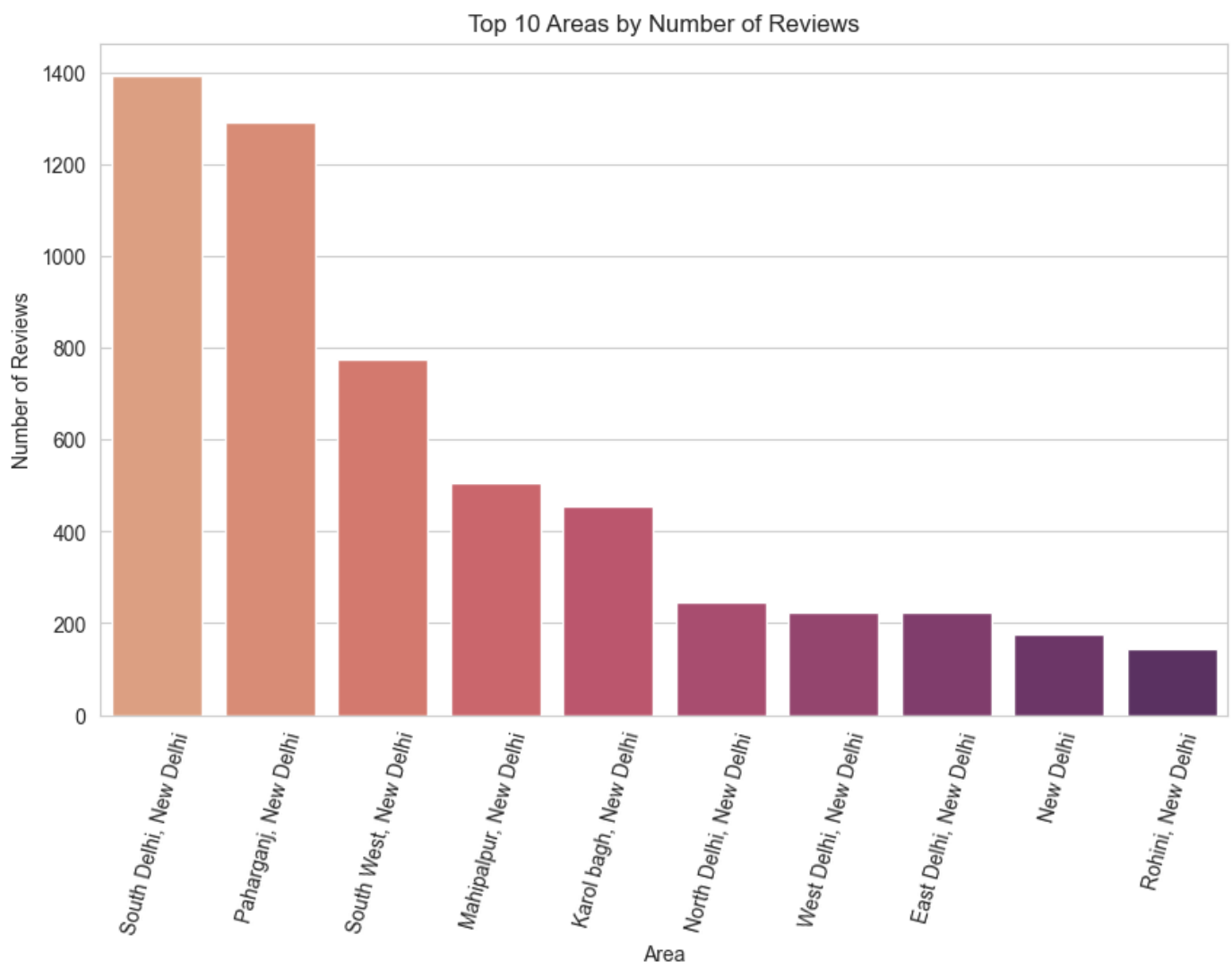
In [43]:
```
#Reviwes to hotels
top_hotels= hotel_reviews["Name"].value_counts().head(10)
plt.figure(figsize=(10, 6))
sns.barplot(x=top_hotels.index, y=top_hotels.values, palette="cubehelix")
plt.title('Top 10  by Number of Reviews')
plt.xlabel('Hotel Name')
plt.ylabel('Number of Reviews')
plt.xticks(rotation=75)
plt.show()
```

```
C:\Users\Majid\AppData\Local\Temp\ipykernel_10772\681622217.py:4: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0.
Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(x=top_hotels.index, y=top_hotels.values, palette="cubehelix")
```



In [44]:
```
#4. Reviews by Area
top_areas = hotel_reviews['Area'].value_counts().head(10)
plt.figure(figsize=(10, 6))
sns.barplot(x=top_areas.index, y=top_areas.values, palette='flare')
plt.title('Top 10 Areas by Number of Reviews')
plt.xlabel('Area')
plt.ylabel('Number of Reviews')
plt.xticks(rotation=75)
plt.show()
```

```
C:\Users\Majid\AppData\Local\Temp\ipykernel_10772\1743108804.py:4: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0.
Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(x=top_areas.index, y=top_areas.values, palette='flare')
```



There are noticeable differences in the number of reviews across different areas, indicating varying levels of guest activity or hotel concentration. Some areas have significantly more reviews than others, suggesting they might be popular tourist destinations or have a higher density of hotels.

In [2]:
```
!pip install pyppeteer
```

```
Collecting pyppeteer
  Downloading pyppeteer-1.0.2-py3-none-any.whl (83 kB)
     ---------------------------------------- 0.0/83.4 kB ? eta -:--:--
     ---- ----------------------------------- 10.2/83.4 kB ? eta -:--:--
     ---- ----------------------------------- 10.2/83.4 kB ? eta -:--:--
     ------------- ------------------------- 30.7/83.4 kB 325.1 kB/s eta 0:00:01
     ----------------- -------------------- 41.0/83.4 kB 217.9 kB/s eta 0:00:01
     ---------------------------- --------- 61.4/83.4 kB 297.7 kB/s eta 0:00:01
     ---------------------------------------- 83.4/83.4 kB 360.3 kB/s eta 0:00:00
Requirement already satisfied: appdirs<2.0.0,>=1.4.3 in c:\users\majid\anaconda3\lib\sit
e-packages (from pyppeteer) (1.4.4)
Requirement already satisfied: certifi>=2021 in c:\users\majid\anaconda3\lib\site-packag
es (from pyppeteer) (2023.7.22)
Requirement already satisfied: importlib-metadata>=1.4 in c:\users\majid\anaconda3\lib\s
ite-packages (from pyppeteer) (6.0.0)
Collecting pyee<9.0.0,>=8.1.0 (from pyppeteer)
  Downloading pyee-8.2.2-py2.py3-none-any.whl (12 kB)
Requirement already satisfied: tqdm<5.0.0,>=4.42.1 in c:\users\majid\anaconda3\lib\site-
packages (from pyppeteer) (4.65.0)
Requirement already satisfied: urllib3<2.0.0,>=1.25.8 in c:\users\majid\anaconda3\lib\si
te-packages (from pyppeteer) (1.26.16)
Collecting websockets<11.0,>=10.0 (from pyppeteer)
  Downloading websockets-10.4-cp311-cp311-win_amd64.whl (101 kB)
```

```
     ---------------------------------------- 0.0/101.4 kB ? eta -:--:--
     ---------------------------------------- 0.0/101.4 kB ? eta -:--:--
     ---------------------------------------- 0.0/101.4 kB ? eta -:--:--
     ---------------------------------------- 0.0/101.4 kB ? eta -:--:--
     ---------------------------------- --- 92.2/101.4 kB 2.6 MB/s eta 0:00:01
     ------------------------------------ 101.4/101.4 kB 2.9 MB/s eta 0:00:00
Requirement already satisfied: zipp>=0.5 in c:\users\majid\anaconda3\lib\site-packages
(from importlib-metadata>=1.4->pyppeteer) (3.11.0)
Requirement already satisfied: colorama in c:\users\majid\anaconda3\lib\site-packages (f
rom tqdm<5.0.0,>=4.42.1->pyppeteer) (0.4.6)
Installing collected packages: pyee, websockets, pyppeteer
Successfully installed pyee-8.2.2 pyppeteer-1.0.2 websockets-10.4
```