In [53]:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
import pickle as pckl
import seaborn
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
```

In [26]:

```python
#creating dataframe

df=pd.read_csv("Capstone project phase 1.csv")
df
```

Out[26]:

	State	Year	Population of Each state	Litracy rate	Area in Sq Km	Total Crimes	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unnamed: 12	Unnamed: 13
0	Andhra Pradesh	2001.0	75728400.0	66.40	1,62,975	130089.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Arunachal Pradesh	2001.0	1098328.0	66.95	83,743	2342.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	Assam	2001.0	26638600.0	73.18	78,438	36877.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	Bihar	2001.0	82879910.0	69.82	94,163	88432.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	Chhattisgarh	2001.0	20834530.0	71.04	1,35,192	38460.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
380	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
381	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
382	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
383	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
384	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

385 rows × 14 columns

In [27]:

```python
df.drop(columns=['Unnamed: 6','Unnamed: 7','Unnamed: 8','Unnamed: 9','Unnamed: 10','Unnamed: 11','Unnamed: 12','Unnamed: 13'],inplace=True)
df
```

Out[27]:

	State	Year	Population of Each state	Litracy rate	Area in Sq Km	Total Crimes
0	Andhra Pradesh	2001.0	75728400.0	66.40	1,62,975	130089.0
1	Arunachal Pradesh	2001.0	1098328.0	66.95	83,743	2342.0
2	Assam	2001.0	26638600.0	73.18	78,438	36877.0
3	Bihar	2001.0	82879910.0	69.82	94,163	88432.0
4	Chhattisgarh	2001.0	20834530.0	71.04	1,35,192	38460.0
...	...	...	...	...	...	...
380	NaN	NaN	NaN	NaN	NaN	NaN
381	NaN	NaN	NaN	NaN	NaN	NaN
382	NaN	NaN	NaN	NaN	NaN	NaN
383	NaN	NaN	NaN	NaN	NaN	NaN
384	NaN	NaN	NaN	NaN	NaN	NaN

385 rows × 6 columns

In [28]:

```python
df
```

Out[28]:

	State	Year	Population of Each state	Litracy rate	Area in Sq Km	Total Crimes
0	Andhra Pradesh	2001.0	75728400.0	66.40	1,62,975	130089.0
1	Arunachal Pradesh	2001.0	1098328.0	66.95	83,743	2342.0
2	Assam	2001.0	26638600.0	73.18	78,438	36877.0
3	Bihar	2001.0	82879910.0	69.82	94,163	88432.0
4	Chhattisgarh	2001.0	20834530.0	71.04	1,35,192	38460.0
...	...	...	...	...	...	...
380	NaN	NaN	NaN	NaN	NaN	NaN
381	NaN	NaN	NaN	NaN	NaN	NaN
382	NaN	NaN	NaN	NaN	NaN	NaN
383	NaN	NaN	NaN	NaN	NaN	NaN
384	NaN	NaN	NaN	NaN	NaN	NaN

385 rows × 6 columns

In [68]:

```python
#remove Nan valiue


df.dropna(subset=['State','Year','Population of Each state','Litracy rate','Total Crimes'], inplace=True)
```

In [58]:

```python
df.
```

Out[58]:

	State	Year	Population of Each state	Litracy rate	Area in Sq Km	Total Crimes
0	Andhra Pradesh	2001.0	75728400.0	66.40	1,62,975	130089.0
1	Arunachal Pradesh	2001.0	1098328.0	66.95	83,743	2342.0
2	Assam	2001.0	26638600.0	73.18	78,438	36877.0
3	Bihar	2001.0	82879910.0	69.82	94,163	88432.0
4	Chhattisgarh	2001.0	20834530.0	71.04	1,35,192	38460.0
...	...	...	...	...	...	...
319	Tamil Nadu	2012.0	635963102.0	81.33	1,30,058	200474.0
320	Tripura	2012.0	32659810.0	88.75	1,12,077	6264.0
321	Uttar Pradesh	2012.0	179673604.0	69.78	2,40,928	198093.0
322	Uttarakhand	2012.0	89449107.0	79.64	53,483	8882.0
323	West Bengal	2012.0	86571309.0	78.08	88,752	161427.0

324 rows × 6 columns

In [59]:

```python
df.head(15)
```

Out[59]:

	State	Year	Population of Each state	Litracy rate	Area in Sq Km	Total Crimes
0	Andhra Pradesh	2001.0	75728400.0	66.40	1,62,975	130089.0
1	Arunachal Pradesh	2001.0	1098328.0	66.95	83,743	2342.0
2	Assam	2001.0	26638600.0	73.18	78,438	36877.0
3	Bihar	2001.0	82879910.0	69.82	94,163	88432.0
4	Chhattisgarh	2001.0	20834530.0	71.04	1,35,192	38460.0
5	Goa	2001.0	1348900.0	87.40	3,702	2341.0
6	Gujarat	2001.0	50597200.0	79.31	1,96,024	103419.0
7	Haryana	2001.0	50597200.0	76.64	44,212	38759.0
8	Himachal Pradesh	2001.0	6077453.0	83.78	55,673	11499.0
9	Jharkhand	2001.0	26946070.0	66.40	79,716	25447.0
10	Karnataka	2001.0	52734986.0	75.60	1,91,791	109098.0
11	Kerala	2001.0	31839000.0	93.91	38,863	103847.0
12	Madhya Pradesh	2001.0	60385090.0	70.63	3,08,252	181741.0
13	Maharashtra	2001.0	96752500.0	82.91	3,07,713	171233.0
14	Manipur	2001.0	2294480.0	79.85	22,327	2489.0

In [69]:

```python
#check the mean
df.describe()
```

Out[69]:

	Year	Population of Each state	Litracy rate	Total Crimes
count	324.000000	3.240000e+02	324.000000	324.000000
mean	2006.500000	4.886474e+08	77.405864	70789.993827
std	3.457392	1.220932e+09	8.484150	68989.911324
min	2001.000000	5.419020e+05	8.450000	443.000000
25%	2003.750000	3.189886e+07	71.120000	3284.500000
50%	2006.500000	8.124579e+07	76.780000	45563.500000
75%	2009.250000	5.174197e+08	82.210000	129734.250000
max	2012.000000	8.298676e+09	94.910000	220335.000000

In [64]:

```python
#EDA(Exploratory data analisys)
#to checking the dimension
df.shape
```

Out[64]:

```
(324, 6)
```

In [65]:

```python
#if we ant to fetchout the column names
#to checking the output
df.columns
```

Out[65]:

```
Index(['State', 'Year', 'Population of Each state', 'Litracy rate',
       'Area in Sq Km ', 'Total Crimes'],
      dtype='object')
```

In [66]:

```python
df.isnull().sum().sum()
```

Out[66]:

```
0
```

In [67]:

```python
#checking the value counts of each column
for i in df.columns:
    print(df[i].value_counts())
    print("\n")
```

```
Andhra Pradesh       12
Manipur              12
Uttarakhand          12
Uttar Pradesh        12
Tripura              12
Tamil Nadu           12
Sikkim               12
Rajasthan            12
Punjab               12
Odisha               12
Nagaland             12
Mizoram              12
Meghalaya            12
Maharashtra          12
Arunachal Pradesh    12
Madhya Pradesh       12
Kerala               12
Karnataka            12
Jharkhand            12
Himachal Pradesh     12
Haryana              12
Gujarat              12
Goa                  12
Chhattisgarh         12
Bihar                12
Assam                12
West Bengal          12
Name: State, dtype: int64


2001.0    27
2002.0    27
2003.0    27
2004.0    27
2005.0    27
2006.0    27
2007.0    27
2008.0    27
2009.0    27
2010.0    27
2011.0    27
2012.0    27
Name: Year, dtype: int64


5.079720e+07    3
5.759120e+05    2
5.283599e+07    2
5.299686e+09    2
5.059720e+07    2
               ..
1.351096e+08    1
2.099453e+07    1
8.298676e+09    1
2.873960e+08    1
8.657131e+07    1
Name: Population of Each state, Length: 313, dtype: int64


67.40    11
69.92    10
94.91     8
67.95     8
76.78     7
         ..
93.93     1
76.64     1
87.50     1
83.79     1
69.78     1
Name: Litracy rate, Length: 120, dtype: int64


1,62,975    12
22,327      12
53,483      12
2,40,928    12
1,12,077    12
1,30,058    12
7,096       12
3,42,239    12
50,362      12
1,55,707    12
16,579      12
21,081      12
22,429      12
3,07,713    12
83,743      12
3,08,252    12
38,863      12
1,91,791    12
79,716      12
55,673      12
44,212      12
1,96,024    12
3,702       12
1,35,192    12
94,163      12
78,438      12
88,752      12
Name: Area in Sq Km , dtype: int64


2286.0      2
31439.0     2
1059.0      2
552.0       2
176833.0    1
           ..
69350.0     1
8634.0      1
130181.0    1
3081.0      1
161427.0    1
Name: Total Crimes, Length: 320, dtype: int64
```

In [62]:

```python
correlation_coefficient = df['Litracy rate'].corr(df['Total Crimes'])
print(f"Correlation Coefficient: {correlation_coefficient}")
```

```
Correlation Coefficient: -0.22692306371930482
```

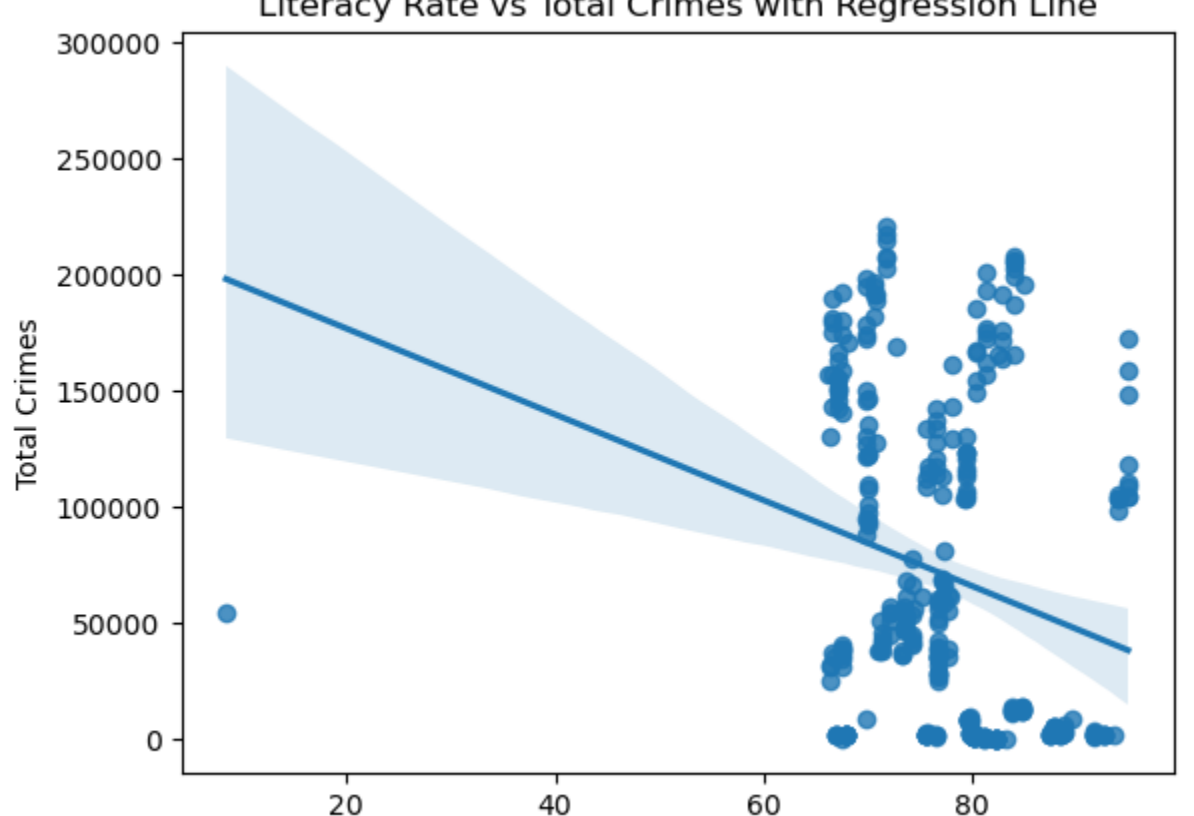
In [63]:

```python
X = df[['Litracy rate']]
y = df['Total Crimes']

model = LinearRegression()
model.fit(X, y)

# Print the coefficients
print(f"Coefficient (slope): {model.coef_[0]}")
print(f"Intercept: {model.intercept_}")
```

```
Coefficient (slope): -1845.2528326417053
Intercept: 213623.3840007335
```

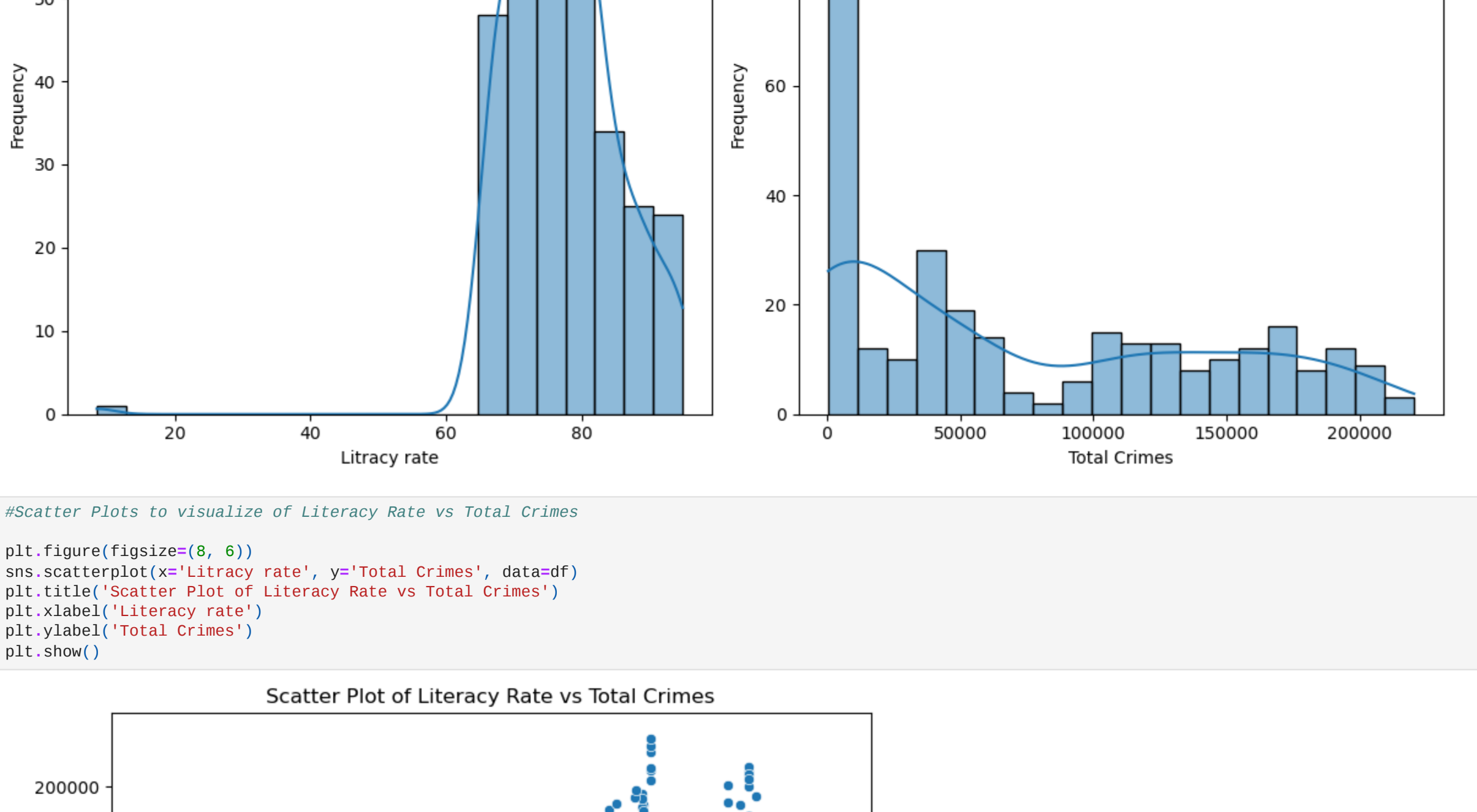
In [70]:

```python
sns.regplot(x='Litracy rate', y='Total Crimes', data=df)
plt.title('Literacy Rate vs Total Crimes with Regression Line')
plt.xlabel('Litracy rate')
plt.ylabel('Total Crimes')
plt.show()
```



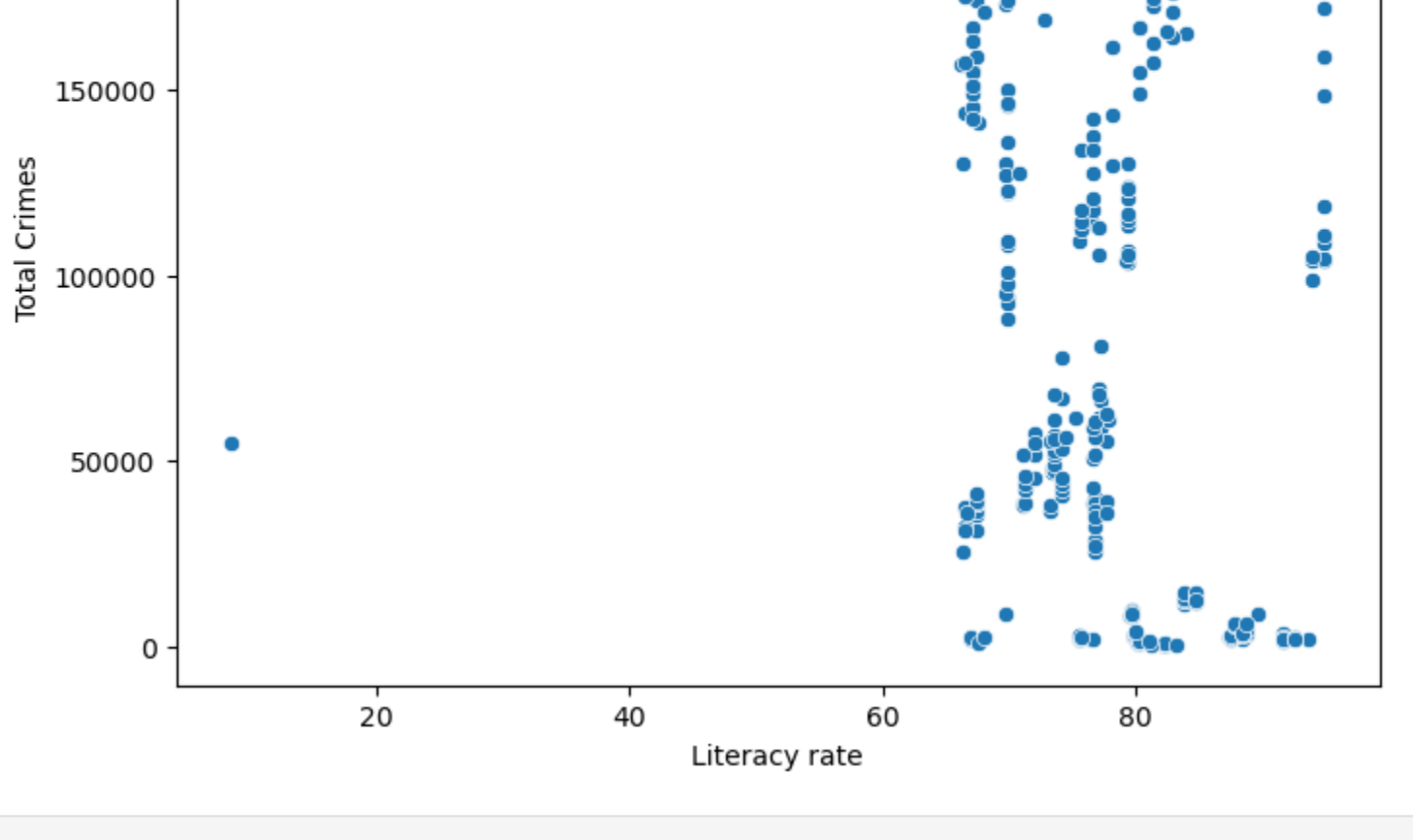
In [86]:

```python
import matplotlib.pyplot as plt
import seaborn as sns

# Assuming 'df' is your DataFrame with columns 'LiteracyRate' and 'TotalCrimes'
fig, axes = plt.subplots(nrows=1, ncols=2, figsize=(12, 6))

# Histogram for Literacy Rate
sns.histplot(df['Litracy rate'], bins=20, kde=True, ax=axes[0])
axes[0].set_title('Histogram of Literacy Rate')
axes[0].set_xlabel('Litracy rate')
axes[0].set_ylabel('Frequency')

# Histogram for Total Crimes
sns.histplot(df['Total Crimes'], bins=20, kde=True, ax=axes[1])
axes[1].set_title('Histogram of Total Crimes')
axes[1].set_xlabel('Total Crimes')
axes[1].set_ylabel('Frequency')

plt.tight_layout()
plt.show()
```



In [89]:

```python
#Scatter Plots to visualize of Literacy Rate vs Total Crimes

plt.figure(figsize=(8, 6))
sns.scatterplot(x='Litracy rate', y='Total Crimes', data=df)
plt.title('Scatter Plot of Literacy Rate vs Total Crimes')
plt.xlabel('Literacy rate')
plt.ylabel('Total Crimes')
plt.show()
```



In [ ]: