In [35]:
```
import pandas as pd
df=pd.read_csv(r"D:\Documents\python\projects\Census.csv")
df
```

Out[35]:

	District_code	State_name	District_name	Population	Male	Female	Literate	Wo
0	1	JAMMU AND KASHMIR	Kupwara	870354	474190	396164	439654	22
1	2	JAMMU AND KASHMIR	Badgam	753745	398041	355704	335649	21
2	3	JAMMU AND KASHMIR	Leh(Ladakh)	133487	78971	54516	93770	1
3	4	JAMMU AND KASHMIR	Kargil	140802	77785	63017	86236	5
4	5	JAMMU AND KASHMIR	Punch	476835	251899	224936	261724	16
...	...	...	...	...	...	...	...	...
635	636	PONDICHERRY	Mahe	41816	19143	22673	36470	1
636	637	PONDICHERRY	Karaikal	200222	97809	102413	154916	6
637	638	ANDAMAN AND NICOBAR ISLANDS	Nicobars	36842	20727	16115	25332	1
638	639	ANDAMAN AND NICOBAR ISLANDS	North AND Middle Andaman	105597	54861	50736	78683	3
639	640	ANDAMAN AND NICOBAR ISLANDS	South Andaman	238142	127283	110859	190266	9

640 rows × 25 columns

In [36]:
```
print("The variables of the Census data are \n",df.columns)
print("The shape of the data is\n ",df.shape)
df.fillna('0')
```

```
The variables of the Census data are
 Index(['District_code', 'State_name', 'District_name', 'Population', 'Male',
       'Female', 'Literate', 'Workers', 'Male_Workers', 'Female_Workers',
       'Cultivator_Workers', 'Agricultural_Workers', 'Household_Workers',
       'Hindus', 'Muslims', 'Christians', 'Sikhs', 'Buddhists', 'Jains',
       'Secondary_Education', 'Higher_Education', 'Graduate_Education',
       'Age_Group_0_29', 'Age_Group_30_49', 'Age_Group_50'],
      dtype='object')
The shape of the data is
  (640, 25)
```

Out[36]:

	District_code	State_name	District_name	Population	Male	Female	Literate	Wo
0	1	JAMMU AND KASHMIR	Kupwara	870354	474190	396164	439654	22
1	2	JAMMU AND KASHMIR	Badgam	753745	398041	355704	335649	21
2	3	JAMMU AND KASHMIR	Leh(Ladakh)	133487	78971	54516	93770	7
3	4	JAMMU AND KASHMIR	Kargil	140802	77785	63017	86236	5
4	5	JAMMU AND KASHMIR	Punch	476835	251899	224936	261724	16
...	...	...	...	...	...	...	...	...
635	636	PONDICHERRY	Mahe	41816	19143	22673	36470	1
636	637	PONDICHERRY	Karaikal	200222	97809	102413	154916	6
637	638	ANDAMAN AND NICOBAR ISLANDS	Nicobars	36842	20727	16115	25332	1
638	639	ANDAMAN AND NICOBAR ISLANDS	North AND Middle Andaman	105597	54861	50736	78683	3
639	640	ANDAMAN AND NICOBAR ISLANDS	South Andaman	238142	127283	110859	190266	9

640 rows × 25 columns

In [37]:
```
print('Hide the indexes of the dataframe \n ')
df=pd.read_csv(r"D:\Documents\python\projects\Census.csv",index_col=0)
```

```
Hide the indexes of the dataframe
```

In [38]:
```
print("How do we set the heading of the dataframe \n")
df.head()
```

```
How do we set the heading of the dataframe
```

Out[38]:

District_code	State_name	District_name	Population	Male	Female	Literate	Workers	N
1	JAMMU AND KASHMIR	Kupwara	870354	474190	396164	439654	229064	
2	JAMMU AND KASHMIR	Badgam	753745	398041	355704	335649	214866	
3	JAMMU AND KASHMIR	Leh(Ladakh)	133487	78971	54516	93770	75079	
4	JAMMU AND KASHMIR	Kargil	140802	77785	63017	86236	51873	
5	JAMMU AND KASHMIR	Punch	476835	251899	224936	261724	161393	

5 rows × 24 columns

In [39]:
```
print("Show data related to the districts 'New Delhi, Lucknow and Jaipur'")
df[df['District_name'].isin(['New Delhi', 'Lucknow', 'Jaipur'])]
```

```
Show data related to the districts 'New Delhi, Lucknow and Jaipur'
```

Out[39]:

District_code	State_name	District_name	Population	Male	Female	Literate	Workers
94	NCT OF DELHI	New Delhi	142004	77942	64062	114179	59541
110	RAJASTHAN	Jaipur	6626178	3468507	3157671	4300965	2464893
157	UTTAR PRADESH	Lucknow	4589838	2394476	2195362	3127260	1542806

3 rows × 24 columns

In [40]:
```
print(f"""Calculate State-wise: \n
    A. The total number of the populaton is \n {df.groupby('State_name').Population
    B. The total number of the population with different religions \n {df.groupby('
    """)
```

```
Calculate State-wise:

    A. The total number of the populaton is
 State_name
UTTAR PRADESH     199812341
MAHARASHTRA       112374333
BIHAR             104099452
WEST BENGAL        91276115
ANDHRA PRADESH     84580777
Name: Population, dtype: int64
    B. The total number of the population with different religions
               Muslims  Christians   Sikhs  Buddhists    Jains
State_name
UTTAR PRADESH  38483967      356448  643500     206285   213267
WEST BENGAL    24654825      658618   63523     282898    60141
BIHAR          17557809      129247   23779      25453    18914
MAHARASHTRA    12971152     1080073  223247    6531200  1400349
ASSAM          10679345     1165867   20672      54993    25949
```

In [41]:
```
print(f"""
    How many male workers are there in Maharashtra state ? \n  {df[df.State_name=='
```

```
   How many male workers are there in Maharashtra state ?
 32616875
```

In [42]:
```
print("What statistical measures would you use to summarize the population distribu
population_summary = df.groupby('District_name')['Population'].sum().sort_values(as
print(population_summary)
```

```
What statistical measures would you use to summarize the population distribution acr
oss different districts?
District_name
Thane                         11060148
North Twenty Four Parganas    10009781
Bangalore                      9621551
Pune                           9429408
Mumbai Suburban                9356962
                                ...
Nicobars                         36842
Upper Siang                      35320
Lahul AND Spiti                  31564
Anjaw                            21167
Dibang Valley                     8004
Name: Population, Length: 634, dtype: int64
```

In [43]:
```
print("How would you calculate the literacy rate for each district based on the pro
# Calculate literacy rate for each district
df['Literacy_Rate'] = (df['Literate'] / df['Population']) * 100
print(df[['District_name', 'Literacy_Rate']])
```

```
How would you calculate the literacy rate for each district based on the provided da
ta?
                          District_name  Literacy_Rate
District_code
1                                Kupwara      50.514388
2                                 Badgam      44.530843
3                           Leh(Ladakh)       70.246541
4                                Kargil       61.246289
5                                 Punch       54.887749
...                                 ...            ...
636                               Mahe        87.215420
637                           Karaikal        77.372117
638                           Nicobars        68.758482
639           North  AND Middle Andaman       74.512534
640                       South Andaman       79.896028

[640 rows x 2 columns]
```

In [44]:
```
print("Can you identify any outliers in the population distribution? How would you
# Identify outliers using z-score
from scipy.stats import zscore

outliers = df[(zscore(df['Population']) > 3) | (zscore(df['Population']) < -3)]
print(outliers.head())
```

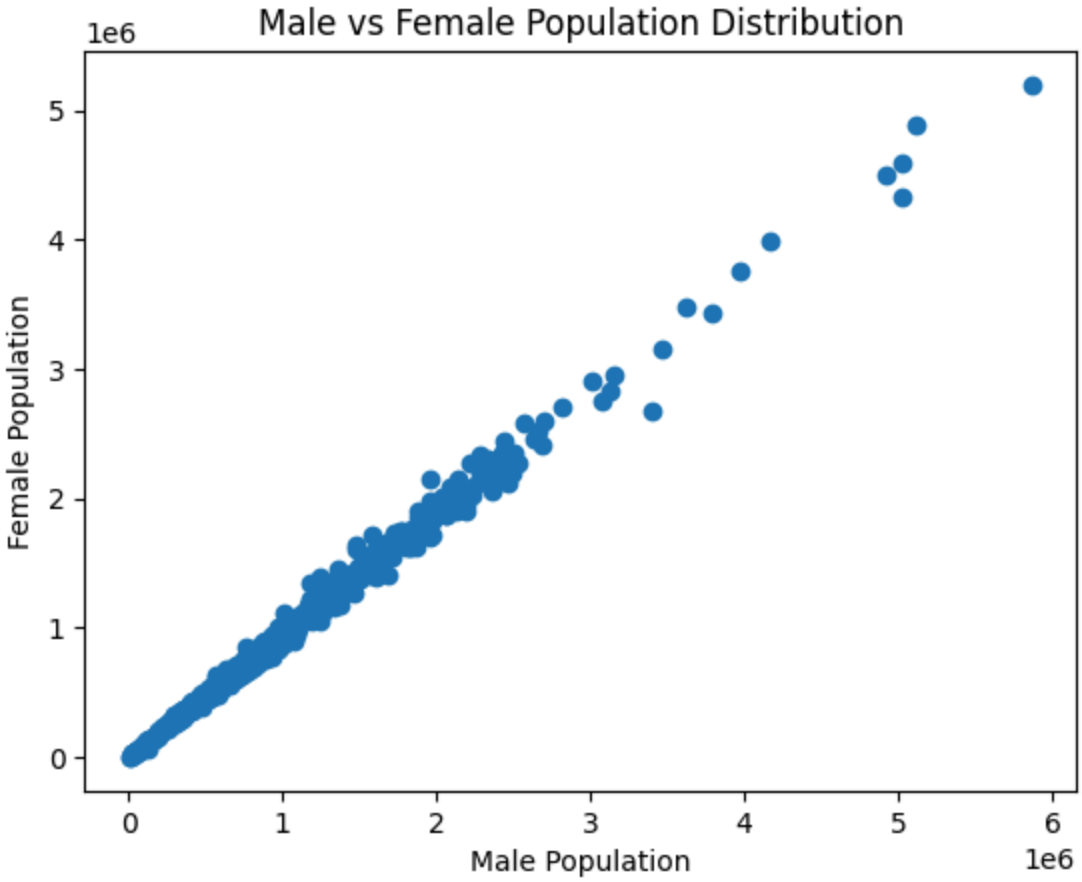
Can you identify any outliers in the population distribution? How would you handle t
hem?

District_code	State_name	District_name	Population	Male	Female	Literate	Workers	Male_Workers	Female_Workers	Cultivator_Workers	...	Sikhs	Buddhists	Jains	Secondary_Education	Higher_Education	Graduate_Education	Age_Group_0_29	Age_Group_30_49	Age_Group_50	Literacy_Rate
110	RAJASTHAN	Jaipur	6626178	3468507	3157671	4300965	2464893	1714947	749946	744374	...	18782	1020	81079	659389	455527	703673	4091148	1646480	884057	64.908685
333	WEST BENGAL	Murshidabad	7103807	3627564	3476243	4055834	2589907	1985667	604240	381076	...	766	348	3037	443254	230242	170051	4390564	1717594	991413	57.093809
335	WEST BENGAL	Barddhaman	7717563	3966889	3750674	5247208	2911251	2293083	618168	342166	...	16675	1602	1674	698251	405935	439823	4149420	2265543	1292913	67.990478
337	WEST BENGAL	North Twenty Four Parganas	10009781	5119389	4890392	7608693	3571624	2945189	626435	288058	...	9394	5818	4452	989053	670977	852923	5104876	2968996	1921143	76.012582
343	WEST BENGAL	South Twenty Four Parganas	8161961	4173778	3988183	5531657	2964494	2356571	607923	355350	...	2783	2494	972	564417	316571	322157	4763943	2151474	1235558	67.773627

[5 rows x 25 columns]

In [45]:
```
print("What insights can you gain from comparing the male and female population dis
import matplotlib.pyplot as plt

# Plotting male and female population distributions
plt.scatter(df['Male'], df['Female'])
plt.xlabel('Male Population')
plt.ylabel('Female Population')
plt.title('Male vs Female Population Distribution')
plt.show()
```

What insights can you gain from comparing the male and female population distributio
ns across districts?



In [46]:
```
print("Calculate the average population density for the entire dataset. How might y
# Calculate average population density
average_density = df['Population'].mean()
print("Average Population Density:", average_density)
```

```
Calculate the average population density for the entire dataset. How might you inter
pret this value in the context of urbanization or rural areas?
Average Population Density: 1891960.9015625
```

In [51]:
```
print("Using statistical tests, determine if there is a significant difference in l
from scipy.stats import ttest_ind

# Perform t-test to compare literacy rates between states
state1 = df[df['State_name'] == 'JAMMU AND KASHMIR']['Literacy_Rate']
state2 = df[df['State_name'] == 'RAJASTHAN']['Literacy_Rate']
t_stat, p_value = ttest_ind(state1, state2)
print("T-Statistic:", t_stat)
print("P-Value:", p_value)
```

```
Using statistical tests, determine if there is a significant difference in literacy
rates between districts in different states.
T-Statistic: 0.2608527695786823
P-Value: 0.7952169556706871
```

In [54]:
```
print("If the p-value is less than the chosen significance level (e.g., 0.05), we r
```

```
If the p-value is less than the chosen significance level (e.g., 0.05), we reject th
e null hypothesis.
 Inference: There is no significant difference in literacy rates between districts i
n JAMMU AND KASHMIR and those in RAJASTAN state.
```

In [58]:
```
print('Investigate the correlation between population size and literacy rates. What
correlation = df['Population'].corr(df['Literacy_Rate'])
print("Correlation between Population and Literacy Rate:", correlation)

print(f"If the correlation coefficient is close to 1, it indicates a strong positiv
```

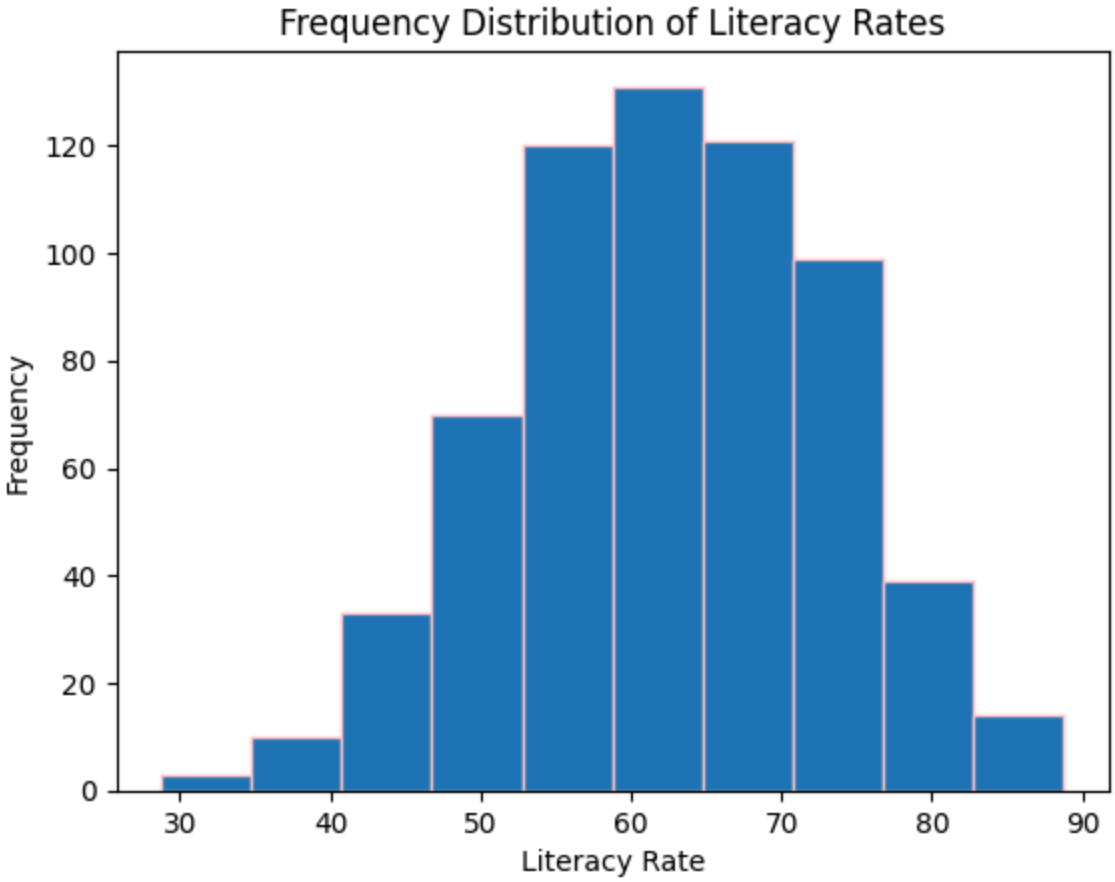
```
Investigate the correlation between population size and literacy rates. What does th
is correlation, if any, suggest about education access in different areas?
Correlation between Population and Literacy Rate: 0.0702645677502296
If the correlation coefficient is close to 1, it indicates a strong positive correla
tion, close to -1 indicates a strong negative correlation, and close to 0 indicates
no correlation.
 Inference: There is no correlation between population size and literacy rate eviden
ced by a correlation coefficient of 0.0702645677502296
```

In [60]:
```
print("Perform a hypothesis test to determine if there is a significant difference
#Hypothesis Testing for Population and Literacy Rate:
# Perform t-test to compare literacy rates of districts above and below average pop
above_avg = df[df['Population'] > average_density]['Literacy_Rate']
below_avg = df[df['Population'] < average_density]['Literacy_Rate']
t_stat, p_value = ttest_ind(above_avg, below_avg)
print("T-Statistic:", t_stat)
print("P-Value:", p_value)
print(f"{p_value} is greater than 0.05. We therefore reject the assumption that  th
```

```
Perform a hypothesis test to determine if there is a significant difference in the l
iteracy rates of districts with populations above and below the average population s
ize.
T-Statistic: 0.4427352448423842
P-Value: 0.6581073099427797
0.6581073099427797 is greater than 0.05. We therefore reject the assumption that  th
ere is a significant difference in the literacy rates of districts with populations
above and below the average population size
```

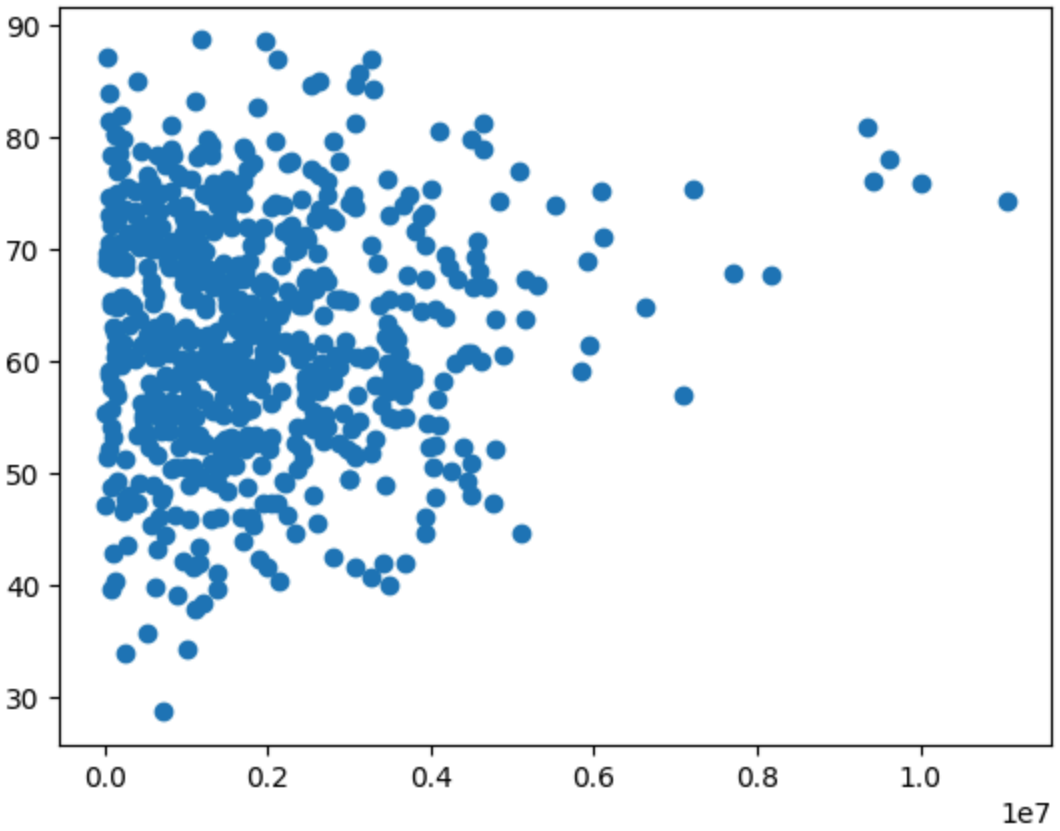
In [61]:
```
print("Create a frequency distribution of literacy rates across all districts. What
#Frequency Distribution of Literacy Rates:
# Plot histogram of literacy rates
plt.hist(df['Literacy_Rate'], bins=10, edgecolor='pink')
plt.xlabel('Literacy Rate')
plt.ylabel('Frequency')
plt.title('Frequency Distribution of Literacy Rates')
plt.show()
```

```
Create a frequency distribution of literacy rates across all districts. What does th
e distribution tell us about the overall literacy levels in the region?
```



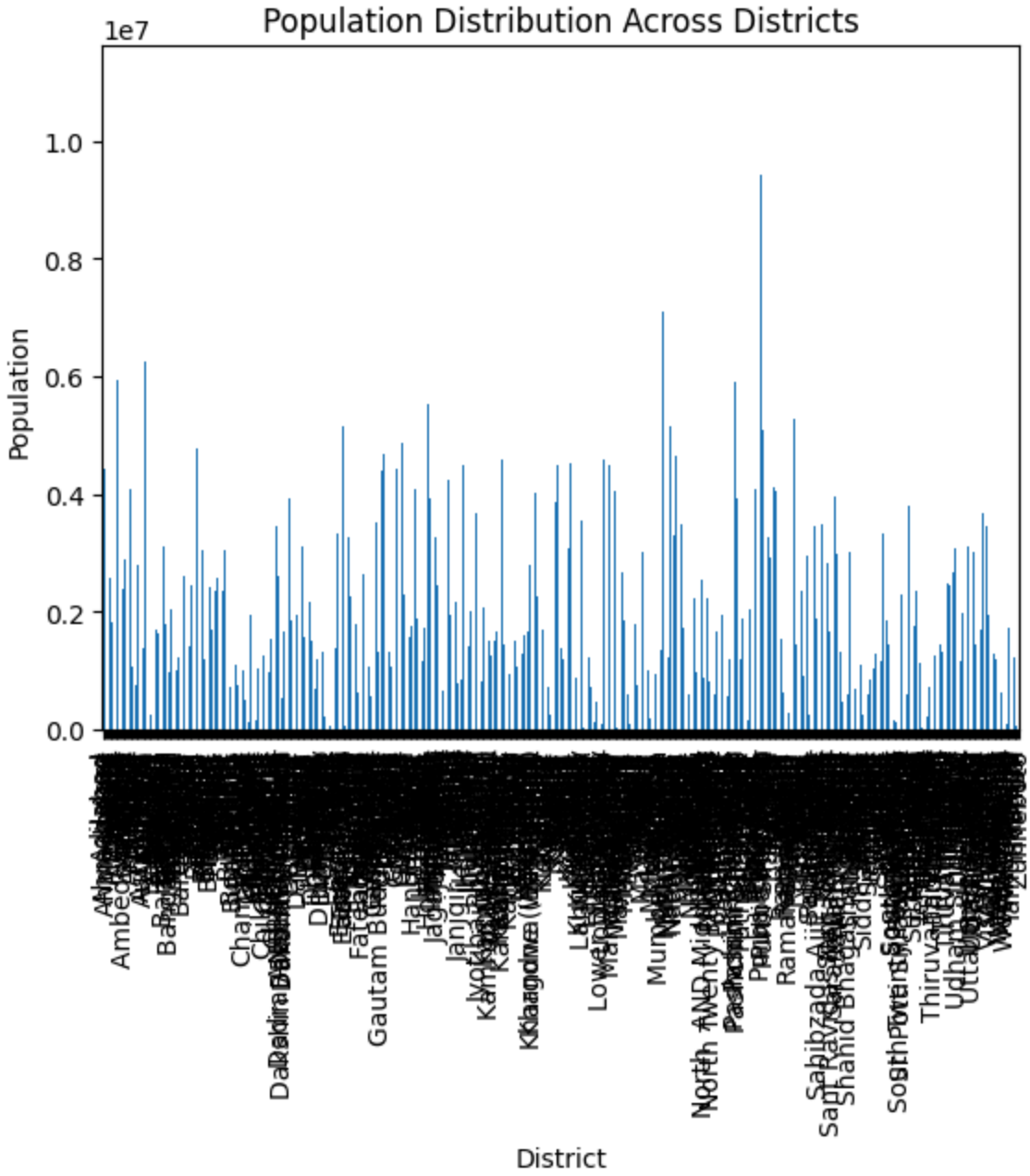
In [77]:
```
print("Using regression analysis, predict the literacy rate for a district based on
#Regression Analysis for Predicting Literacy Rates:
from sklearn.linear_model import LinearRegression

# Perform linear regression to predict literacy rate based on population
X = df[['Population']]
y = df['Literacy_Rate']
model = LinearRegression().fit(X, y)
print("Coefficient:", model.coef_)
print("Intercept:", model.intercept_)
plt.scatter(X,y)
print(f"{model.intercept_}% is the most likely literacy rate")
```

```
Using regression analysis, predict the literacy rate for a district based on its pop
ulation size.
Coefficient: [4.78978927e-07]
Intercept: 61.55699036434614
61.55699036434614% is the most likely literacy rate
```



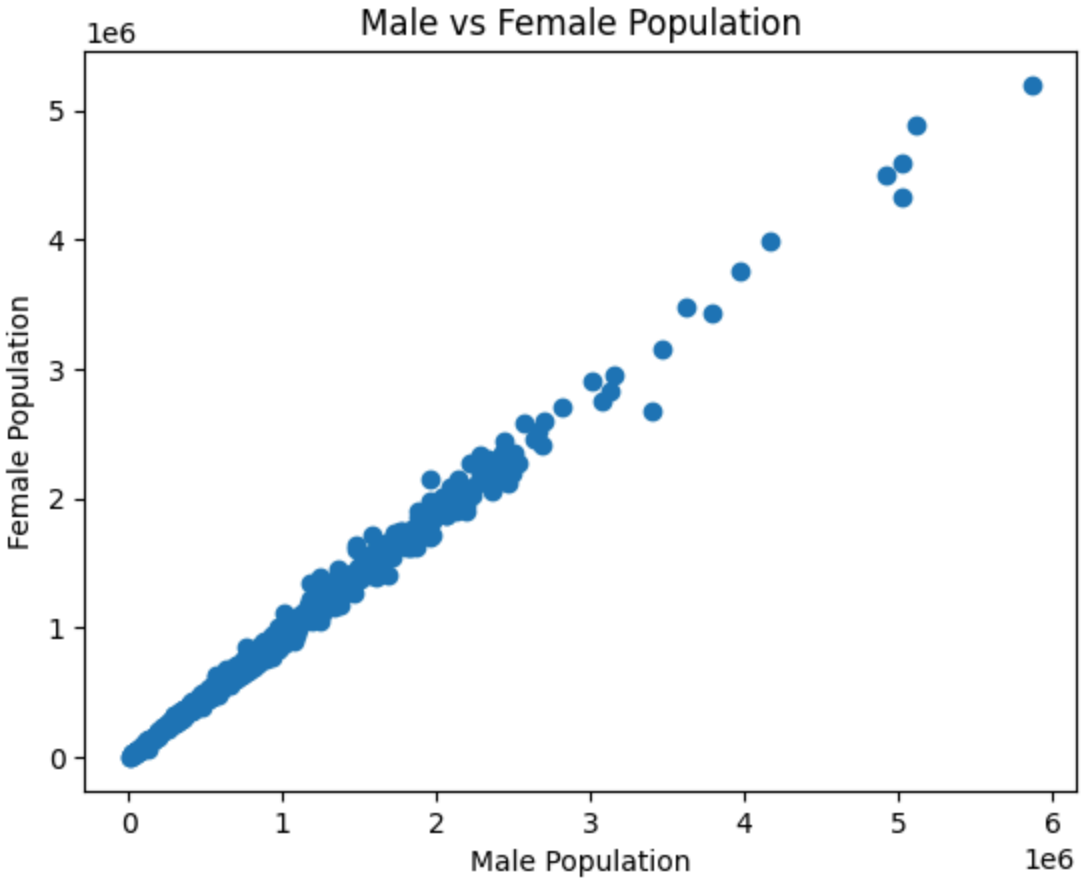
In [80]:
```
print("Create a bar plot showing the population distribution across different distr
#Bar Plot of Population Distribution:
# Plot bar chart of population distribution across districts
df.groupby('District_name')['Population'].sum().plot(kind='bar')
plt.xlabel('District')
plt.ylabel('Population')
plt.title('Population Distribution Across Districts')
plt.show()
```

```
Create a bar plot showing the population distribution across different districts.
```



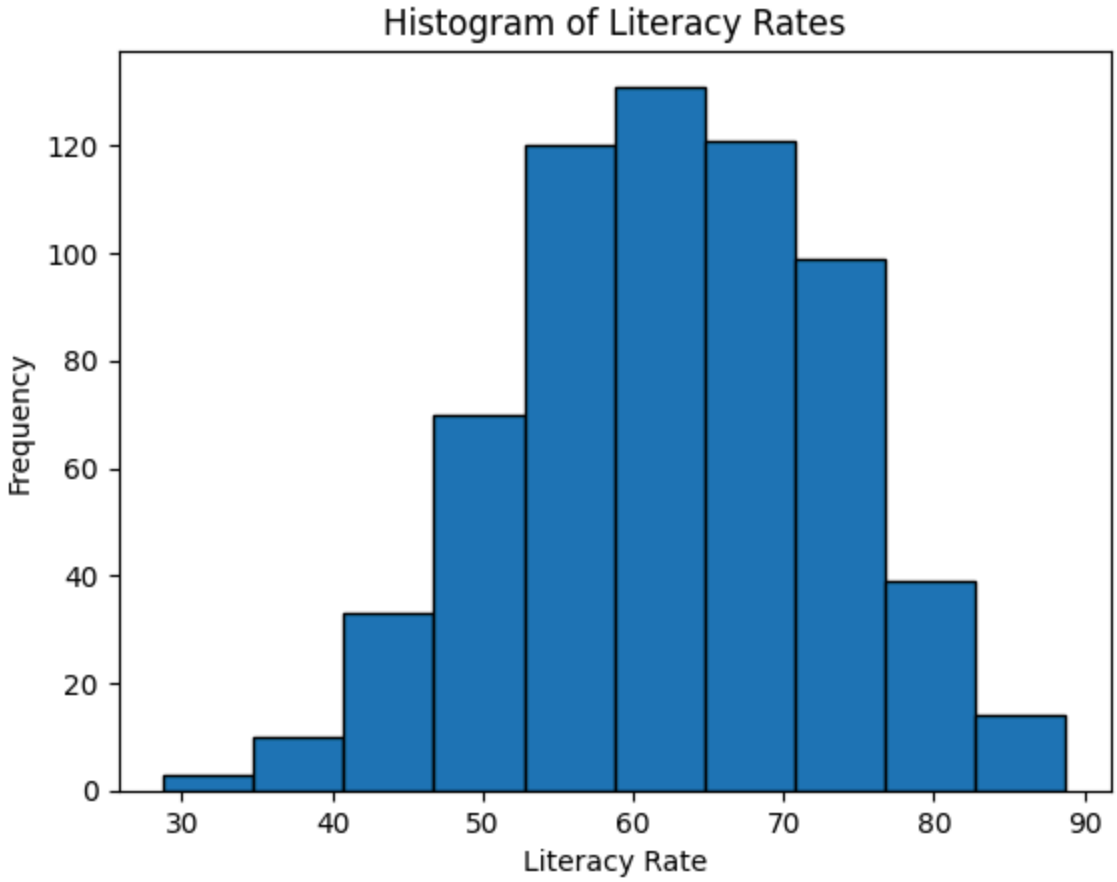
In [81]:
```
print("Generate a scatter plot showing the relationship between male and female pop
#Scatter Plot of Male vs Female Population:
plt.scatter(df['Male'], df['Female'])
plt.xlabel('Male Population')
plt.ylabel('Female Population')
plt.title('Male vs Female Population')
plt.show()
```

Generate a scatter plot showing the relationship between male and female populations in each district.



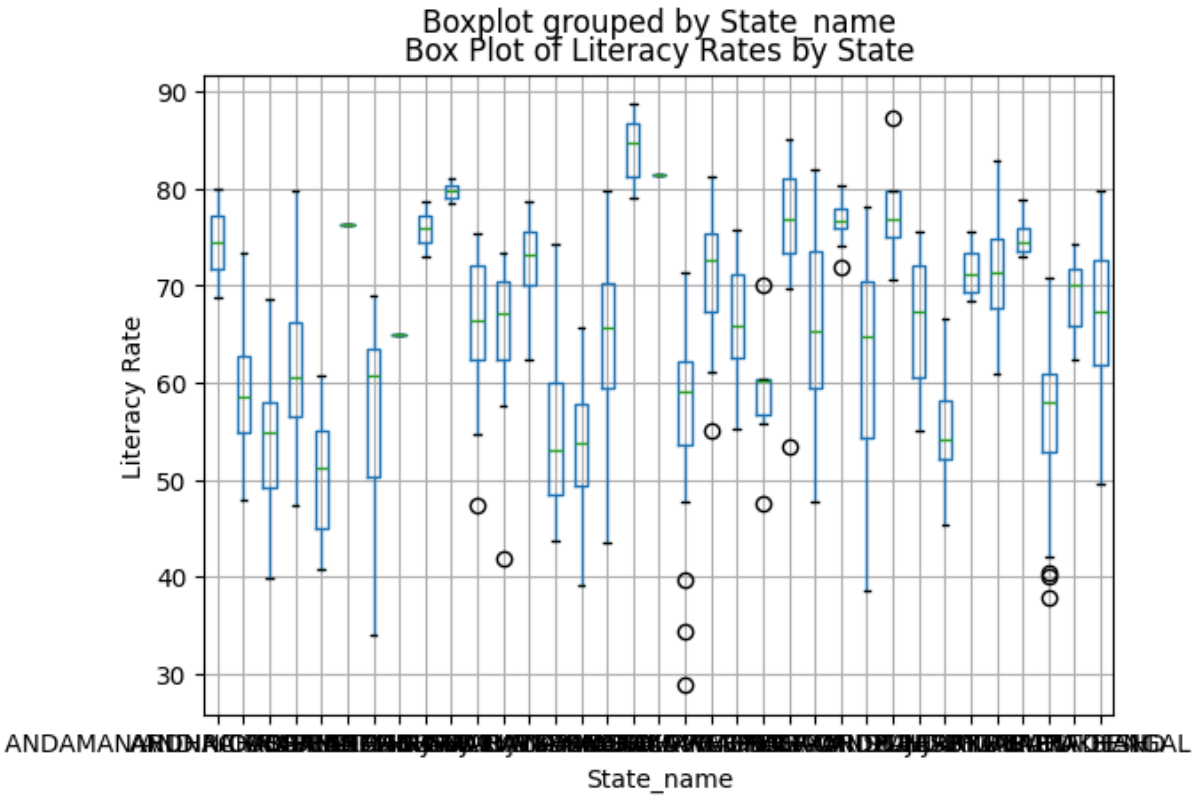
In [82]:
```
print("17.Plot a histogram of literacy rates across all districts.")
# Histogram of literacy rates
plt.hist(df['Literacy_Rate'], bins=10, edgecolor='black')
plt.xlabel('Literacy Rate')
plt.ylabel('Frequency')
plt.title('Histogram of Literacy Rates')
plt.show()
```

```
17.Plot a histogram of literacy rates across all districts.
```



In [83]:
```
print("Create a box plot comparing the literacy rates between different states")
# Box plot comparing literacy rates between states
df.boxplot(column='Literacy_Rate', by='State_name')
plt.ylabel('Literacy Rate')
plt.title('Box Plot of Literacy Rates by State')
plt.show()
```

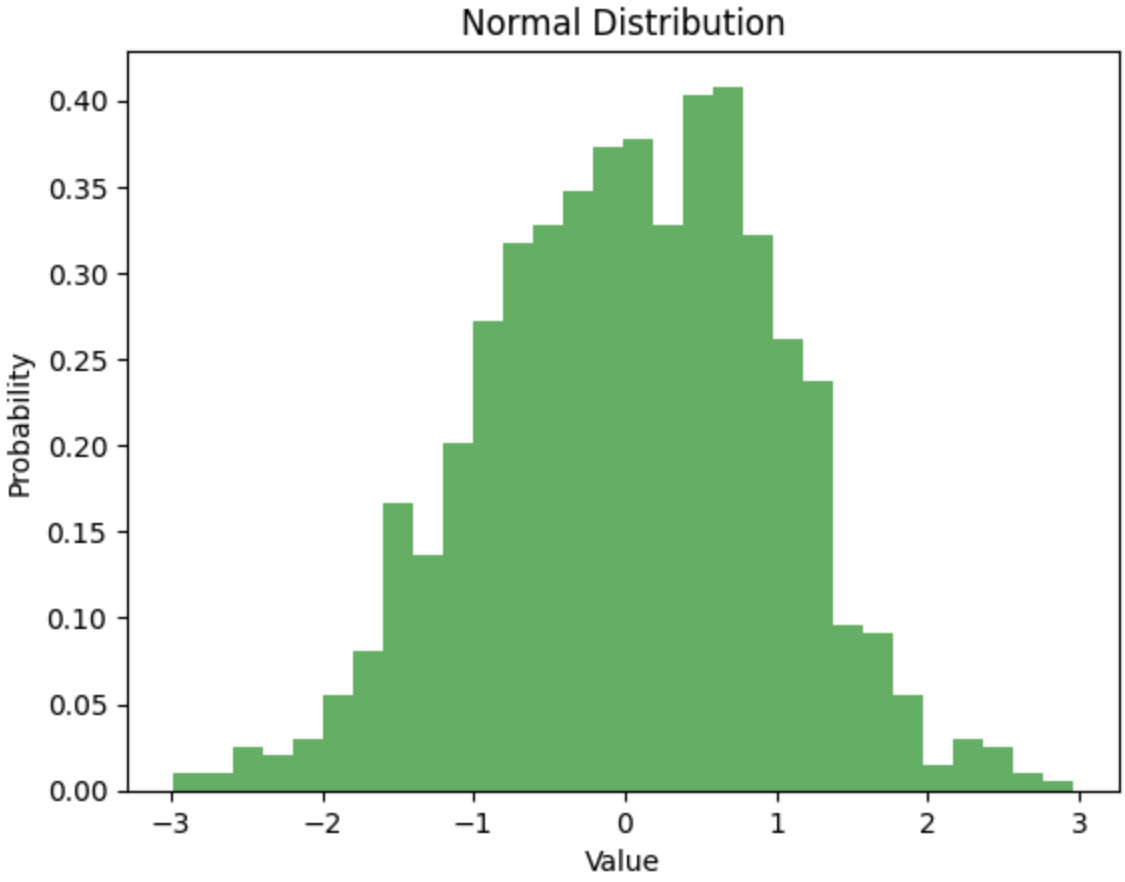
```
Create a box plot comparing the literacy rates between different states
```



In [86]:
```
from scipy.stats import f_oneway

# Perform ANOVA to analyze variance in literacy rates between different states
state1 = df[df['State_name'] == 'JAMMU AND KASHMIR']['Literacy_Rate']
state2 = df[df['State_name'] == 'PONDICHERRY']['Literacy_Rate']
state3 = df[df['State_name'] == 'ANDAMAN AND NICOBAR ISLANDS']['Literacy_Rate']
f_stat, p_value = f_oneway(state1, state2, state3)
print("F-Statistic:", f_stat)
print("P-Value:", p_value)
```

```
F-Statistic: 17.772371670716574
P-Value: 1.365120473088497e-05
```

In [87]:
```
# Example of generating and plotting a normal distribution
import numpy as np
import matplotlib.pyplot as plt

mean = 0
std_dev = 1
samples = np.random.normal(mean, std_dev, 1000)
plt.hist(samples, bins=30, density=True, alpha=0.6, color='g')
plt.title('Normal Distribution')
plt.xlabel('Value')
plt.ylabel('Probability')
plt.show()
```



In [88]:
```
#Calculating Confidence Interval:

from scipy.stats import t

# Calculate confidence interval for literacy rate with 95% confidence
n = len(df['Literacy_Rate'])
mean = df['Literacy_Rate'].mean()
std_dev = df['Literacy_Rate'].std()
t_critical = t.ppf(0.975, df=n-1)
margin_of_error = t_critical * (std_dev / (n ** 0.5))
confidence_interval = (mean - margin_of_error, mean + margin_of_error)
print("Confidence Interval:", confidence_interval)
```

```
Confidence Interval: (61.64602348097321, 63.28037605431719)
```

In [ ]: