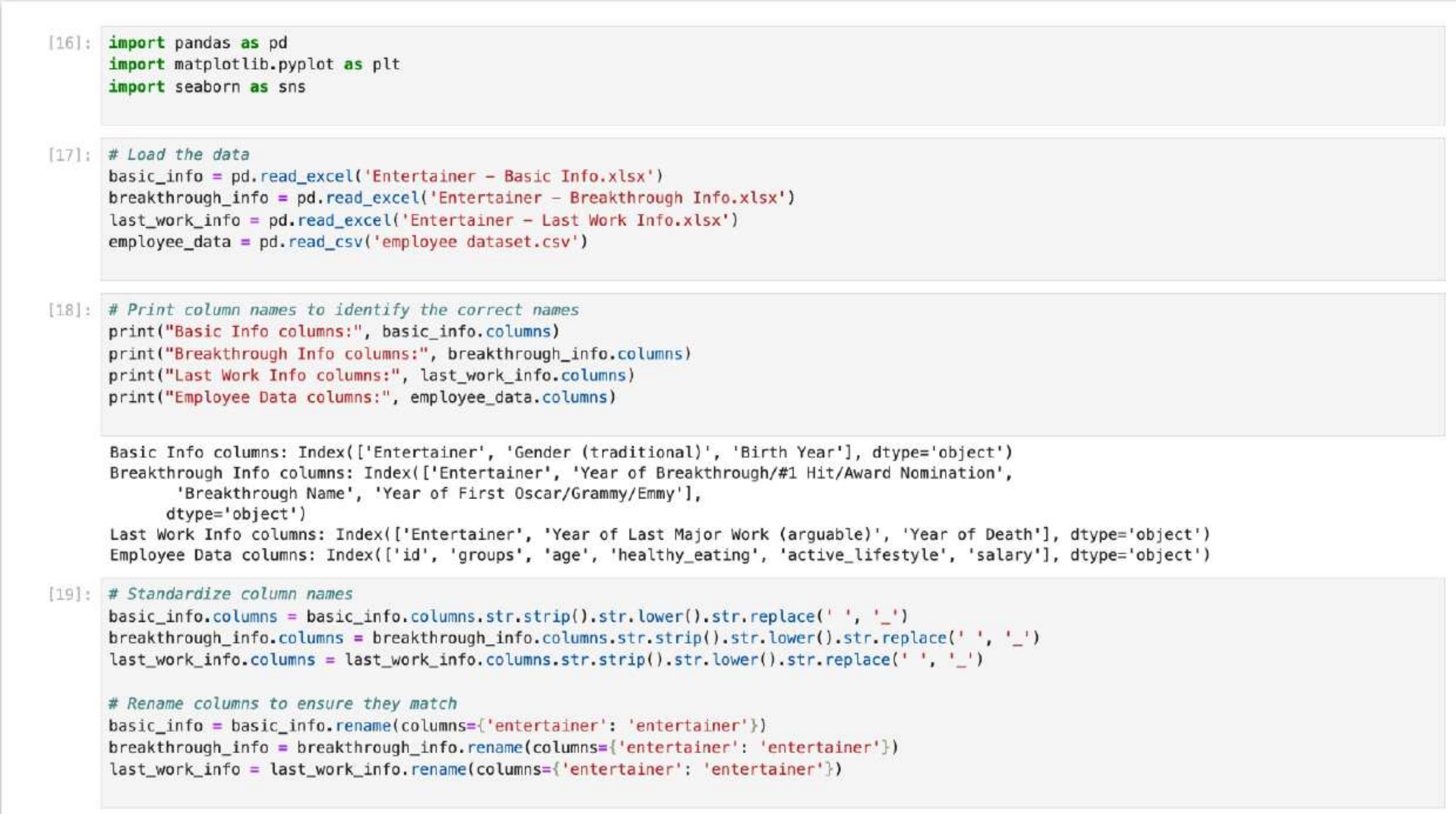


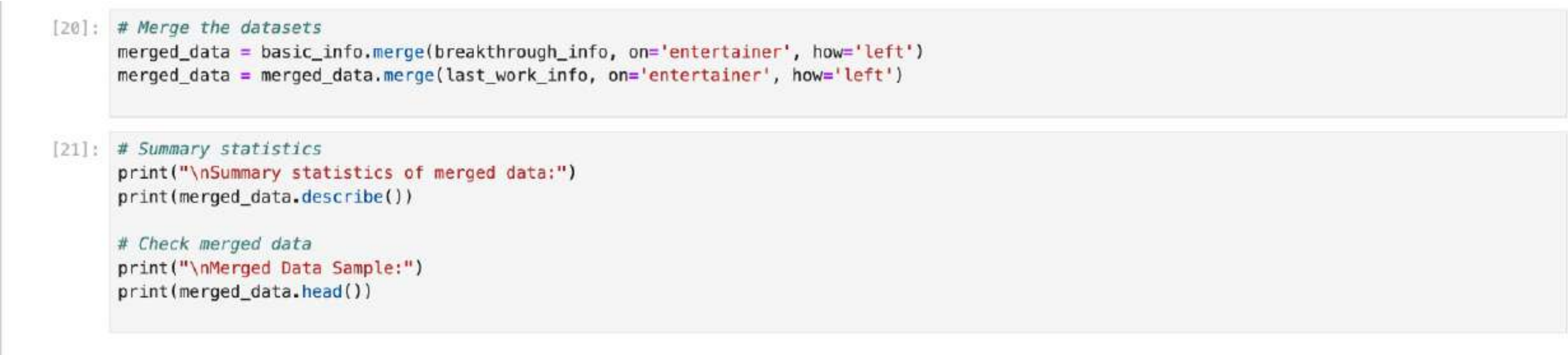
```
[16]: import pandas as pd
      import matplotlib.pyplot as plt
      import seaborn as sns
```

```
[17]: # Load the data
      basic_info = pd.read_excel('Entertainer - Basic Info.xlsx')
      breakthrough_info = pd.read_excel('Entertainer - Breakthrough Info.xlsx')
      last_work_info = pd.read_excel('Entertainer - Last Work Info.xlsx')
      employee_data = pd.read_csv('employee dataset.csv')
```

```
[18]: # Print column names to identify the correct names
      print("Basic Info columns:", basic_info.columns)
      print("Breakthrough Info columns:", breakthrough_info.columns)
      print("Last Work Info columns:", last_work_info.columns)
      print("Employee Data columns:", employee_data.columns)
```

```
Basic Info columns: Index(['Entertainer', 'Gender (traditional)', 'Birth Year'], dtype='object')
Breakthrough Info columns: Index(['Entertainer', 'Year of Breakthrough/#1 Hit/Award Nomination',
       'Breakthrough Name', 'Year of First Oscar/Grammy/Emmy'],
      dtype='object')
Last Work Info columns: Index(['Entertainer', 'Year of Last Major Work (arguable)', 'Year of Death'], dtype='object')
Employee Data columns: Index(['id', 'groups', 'age', 'healthy_eating', 'active_lifestyle', 'salary'], dtype='object')
```

```
[19]: # Standardize column names
      basic_info.columns = basic_info.columns.str.strip().str.lower().str.replace(' ', '_')
      breakthrough_info.columns = breakthrough_info.columns.str.strip().str.lower().str.replace(' ', '_')
      last_work_info.columns = last_work_info.columns.str.strip().str.lower().str.replace(' ', '_')

      # Rename columns to ensure they match
      basic_info = basic_info.rename(columns={'entertainer': 'entertainer'})
      breakthrough_info = breakthrough_info.rename(columns={'entertainer': 'entertainer'})
      last_work_info = last_work_info.rename(columns={'entertainer': 'entertainer'})
```

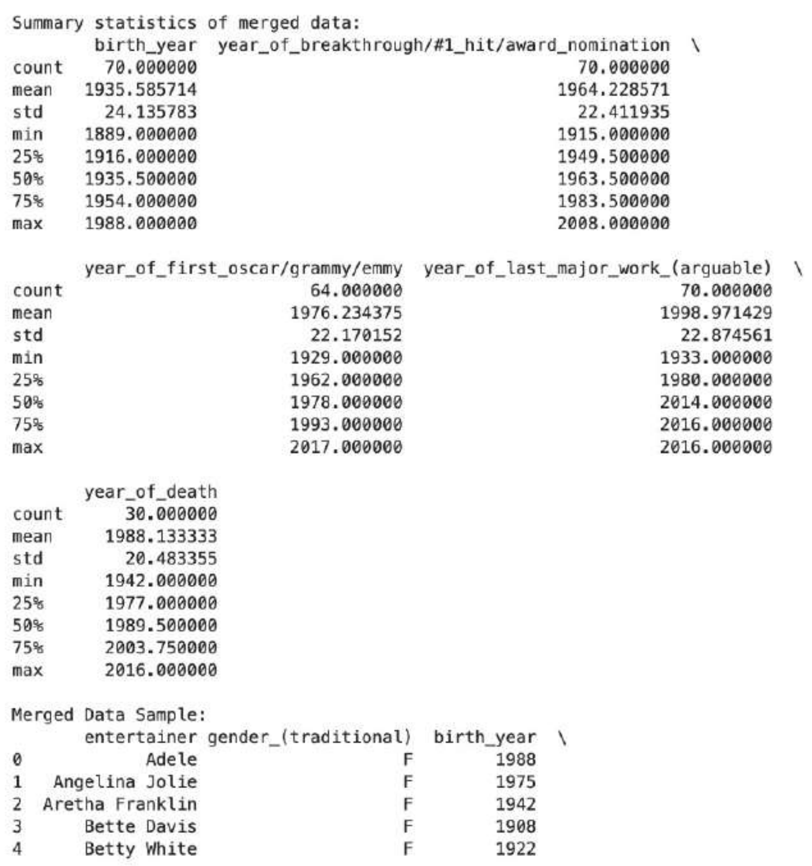


[20]:

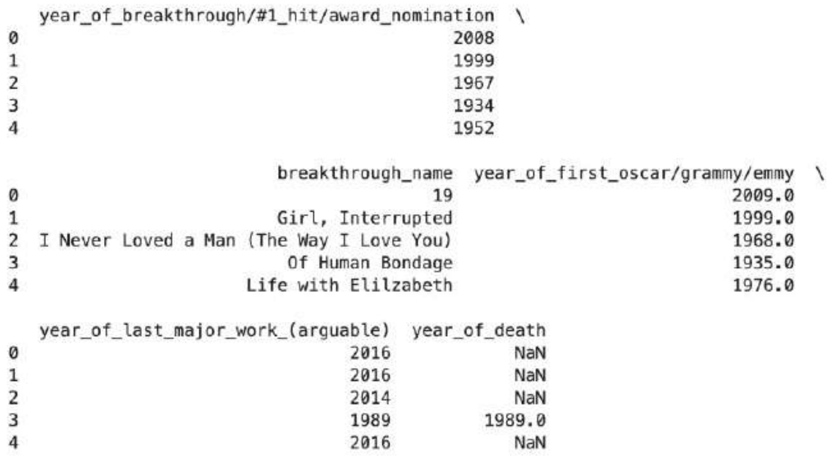
```
# Merge the datasets
merged_data = basic_info.merge(breakthrough_info, on='entertainer', how='left')
merged_data = merged_data.merge(last_work_info, on='entertainer', how='left')
```

[21]:

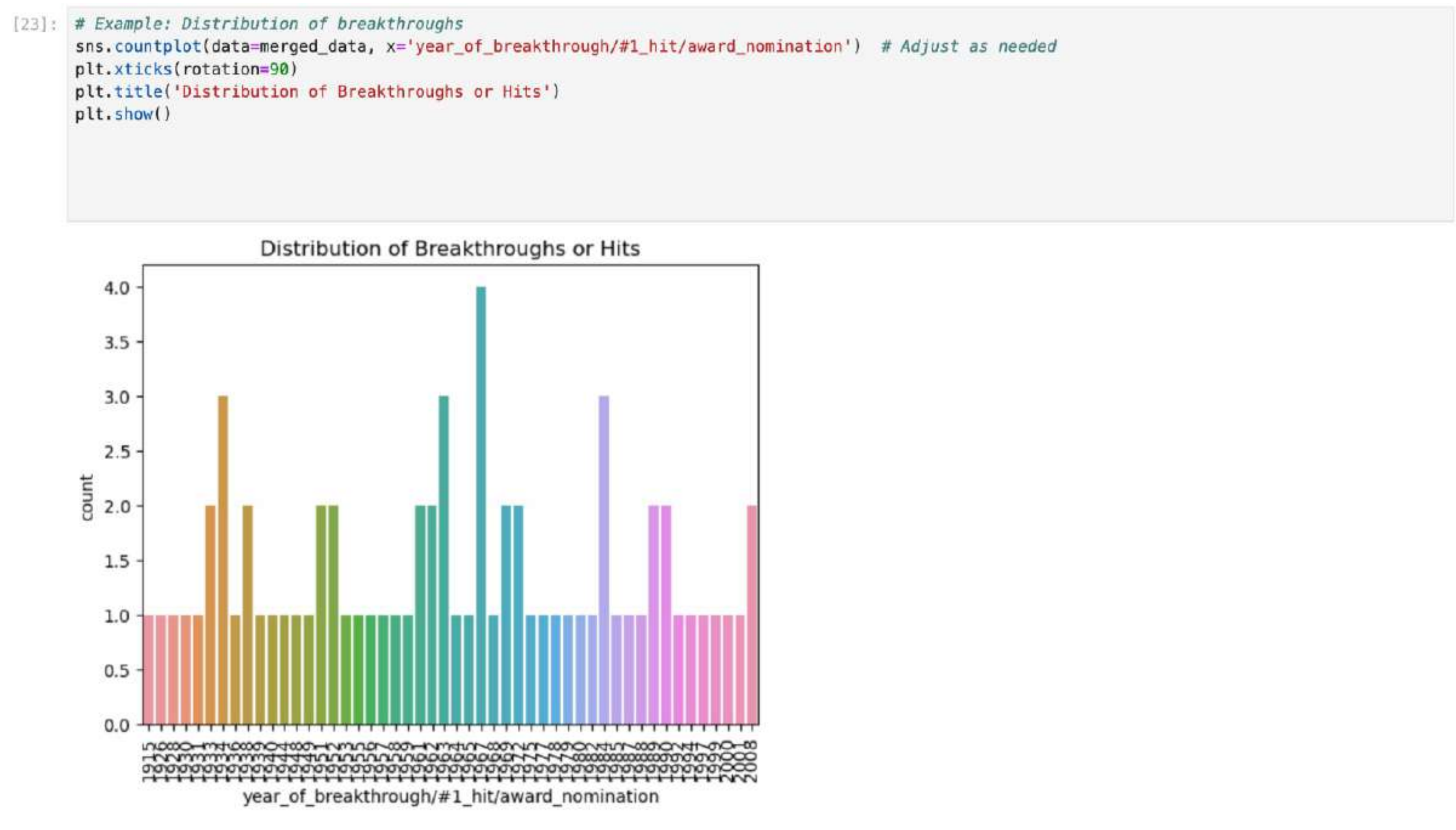
```
# Summary statistics
print("\nSummary statistics of merged data:")
print(merged_data.describe())

# Check merged data
print("\nMerged Data Sample:")
print(merged_data.head())
```



```
Summary statistics of merged data:
        birth_year  year_of_breakthrough/#1_hit/award_nomination  \
count    70.000000                                     70.000000   
mean   1935.585714                                   1964.228571   
std      24.135783                                     22.411935   
min    1889.000000                                   1915.000000   
25%    1916.000000                                   1949.500000   
50%    1935.500000                                   1963.500000   
75%    1954.000000                                   1983.500000   
max    1988.000000                                   2008.000000   

       year_of_first_oscar/grammy/emmy  year_of_last_major_work_(arguable)  \
count                        64.000000                           70.000000   
mean                       1976.234375                         1998.971429   
std                          22.170152                           22.874561   
min                        1929.000000                         1933.000000   
25%                        1962.000000                         1980.000000   
50%                        1978.000000                         2014.000000   
75%                        1993.000000                         2016.000000   
max                        2017.000000                         2016.000000   

       year_of_death  
count      30.000000  
mean     1988.133333  
std        20.483355  
min      1942.000000  
25%      1977.000000  
50%      1989.500000  
75%      2003.750000  
max      2016.000000  

Merged Data Sample:
       entertainer gender_(traditional)  birth_year  \
0            Adele                    F        1988   
1   Angelina Jolie                    F        1975   
2  Aretha Franklin                    F        1942   
3      Bette Davis                    F        1908   
4      Betty White                    F        1922   
```

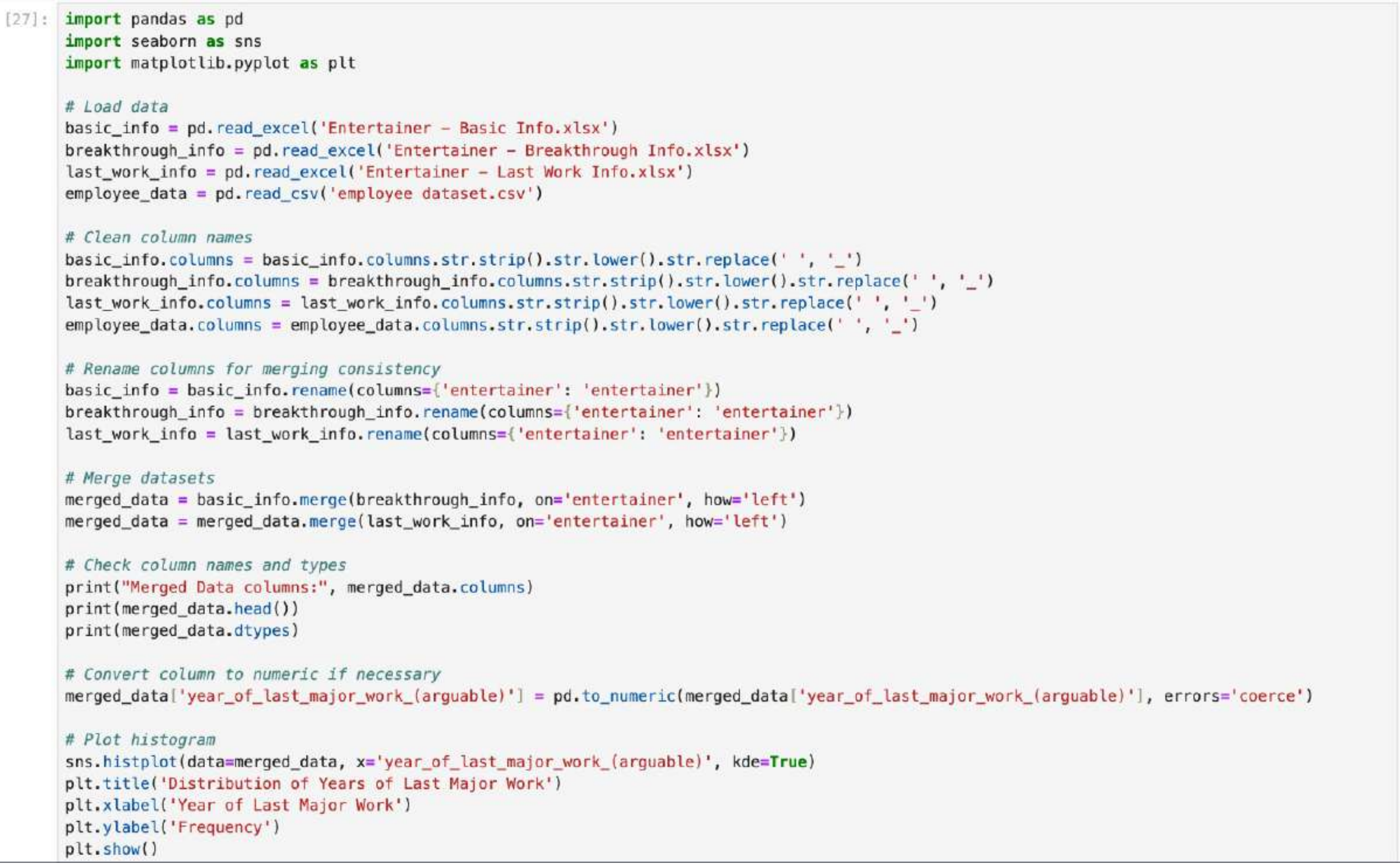


```
   year_of_breakthrough/#1_hit/award_nomination  \
0                                          2008   
1                                          1999   
2                                          1967   
3                                          1934   
4                                          1952   

                        breakthrough_name  year_of_first_oscar/grammy/emmy  \
0                                      19                           2009.0   
1                        Girl, Interrupted                           1999.0   
2  I Never Loved a Man (The Way I Love You)                          1968.0   
3                         Of Human Bondage                           1935.0   
4                     Life with Elilzabeth                           1976.0   

   year_of_last_major_work_(arguable)  year_of_death  
0                                2016            NaN  
1                                2016            NaN  
2                                2014            NaN  
3                                1989         1989.0  
4                                2016            NaN  
```

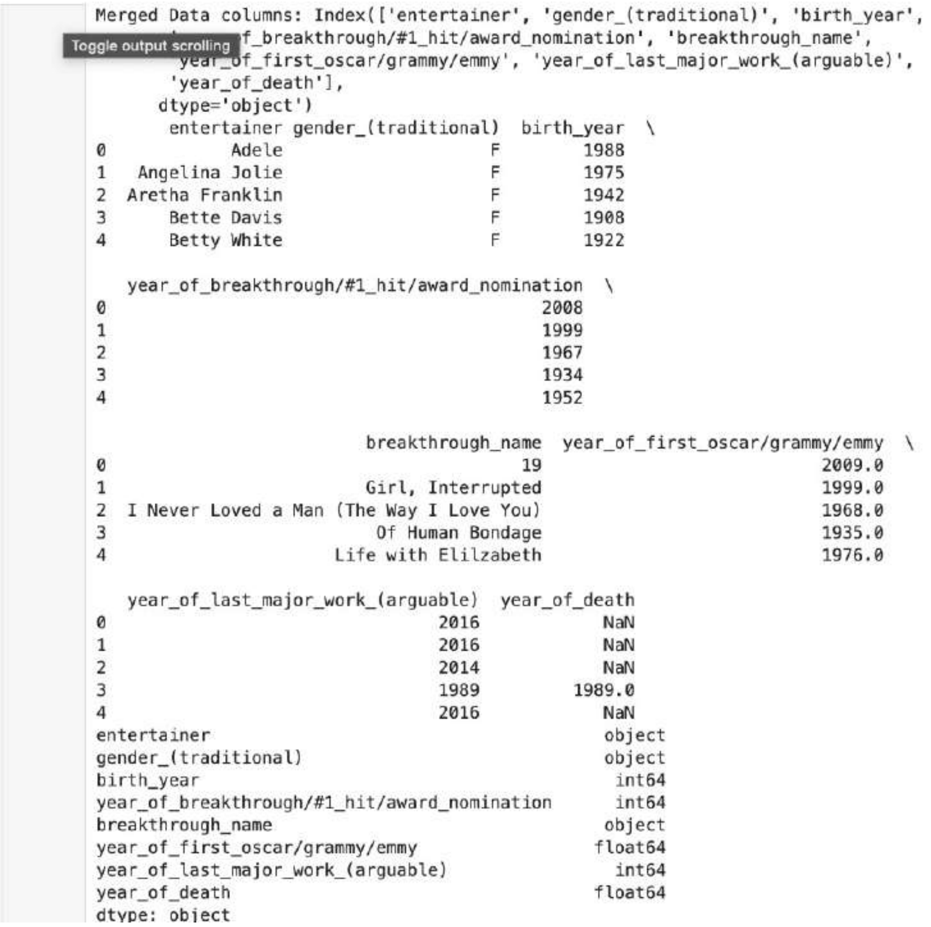


```
[23]: # Example: Distribution of breakthroughs
      sns.countplot(data=merged_data, x='year_of_breakthrough/#1_hit/award_nomination')  # Adjust as needed
      plt.xticks(rotation=90)
      plt.title('Distribution of Breakthroughs or Hits')
      plt.show()
```

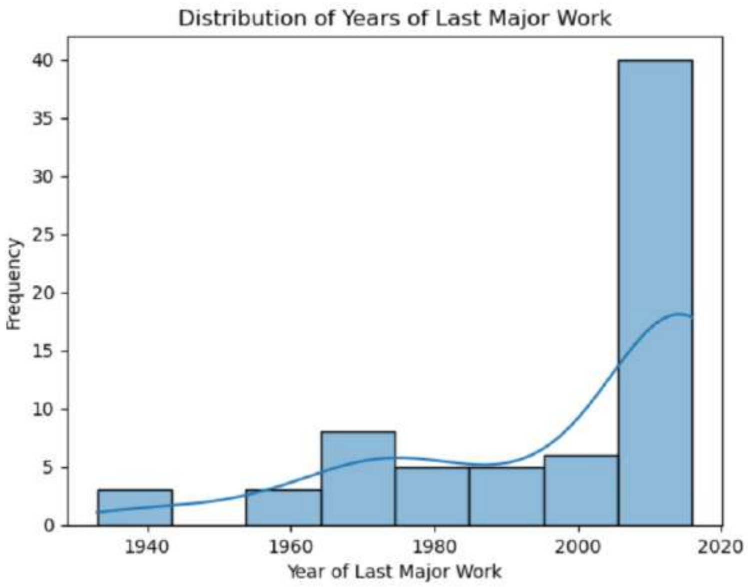
Distribution of Breakthroughs or Hits
count
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1915
2008
year_of_breakthrough/#1_hit/award_nomination



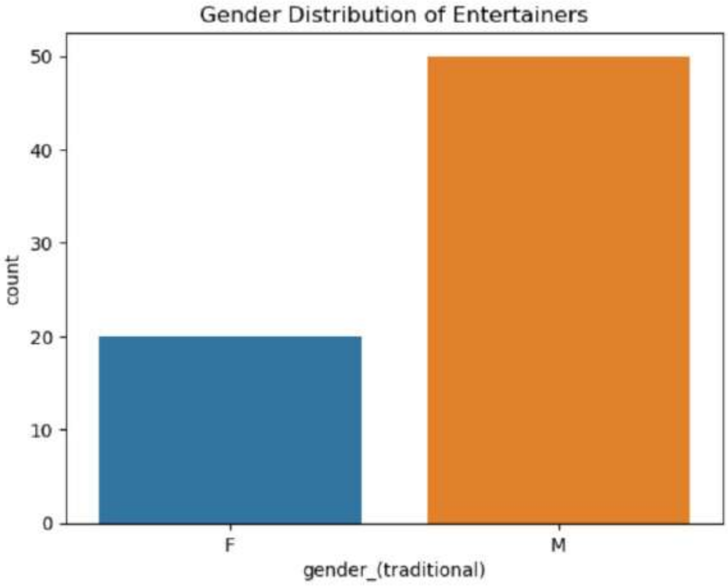
```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load data
basic_info = pd.read_excel('Entertainer - Basic Info.xlsx')
breakthrough_info = pd.read_excel('Entertainer - Breakthrough Info.xlsx')
last_work_info = pd.read_excel('Entertainer - Last Work Info.xlsx')
employee_data = pd.read_csv('employee dataset.csv')

# Clean column names
basic_info.columns = basic_info.columns.str.strip().str.lower().str.replace(' ', '_')
breakthrough_info.columns = breakthrough_info.columns.str.strip().str.lower().str.replace(' ', '_')
last_work_info.columns = last_work_info.columns.str.strip().str.lower().str.replace(' ', '_')
employee_data.columns = employee_data.columns.str.strip().str.lower().str.replace(' ', '_')

# Rename columns for merging consistency
basic_info = basic_info.rename(columns={'entertainer': 'entertainer'})
breakthrough_info = breakthrough_info.rename(columns={'entertainer': 'entertainer'})
last_work_info = last_work_info.rename(columns={'entertainer': 'entertainer'})

# Merge datasets
merged_data = basic_info.merge(breakthrough_info, on='entertainer', how='left')
merged_data = merged_data.merge(last_work_info, on='entertainer', how='left')

# Check column names and types
print("Merged Data columns:", merged_data.columns)
print(merged_data.head())
print(merged_data.dtypes)

# Convert column to numeric if necessary
merged_data['year_of_last_major_work_(arguable)'] = pd.to_numeric(merged_data['year_of_last_major_work_(arguable)'], errors='coerce')

# Plot histogram
sns.histplot(data=merged_data, x='year_of_last_major_work_(arguable)', kde=True)
plt.title('Distribution of Years of Last Major Work')
plt.xlabel('Year of Last Major Work')
plt.ylabel('Frequency')
plt.show()
```



```
Merged Data columns: Index(['entertainer', 'gender_(traditional)', 'birth_year',
                  f_breakthrough/#1_hit/award_nomination', 'breakthrough_name',
        'year_of_first_oscar/grammy/emmy', 'year_of_last_major_work_(arguable)',
       'year_of_death'],
      dtype='object')
       entertainer gender_(traditional)  birth_year  \
0            Adele                    F        1988
1   Angelina Jolie                    F        1975
2  Aretha Franklin                    F        1942
3      Bette Davis                    F        1908
4      Betty White                    F        1922

   year_of_breakthrough/#1_hit/award_nomination  \
0                                          2008
1                                          1999
2                                          1967
3                                          1934
4                                          1952

                          breakthrough_name  year_of_first_oscar/grammy/emmy  \
0                                        19                           2009.0
1                         Girl, Interrupted                           1999.0
2  I Never Loved a Man (The Way I Love You)                           1968.0
3                          Of Human Bondage                           1935.0
4                       Life with Elilzabeth                          1976.0

   year_of_last_major_work_(arguable)  year_of_death
0                                2016            NaN
1                                2016            NaN
2                                2014            NaN
3                                1989         1989.0
4                                2016            NaN
entertainer                                      object
gender_(traditional)                             object
birth_year                                        int64
year_of_breakthrough/#1_hit/award_nomination      int64
breakthrough_name                                object
year_of_first_oscar/grammy/emmy                 float64
year_of_last_major_work_(arguable)                int64
year_of_death                                   float64
dtype: object
```



```
[28]: sns.countplot(data=merged_data, x='gender_(traditional)')  # Adjust as needed
      plt.title('Gender Distribution of Entertainers')
      plt.show()
```


Gender Distribution of Entertainers
50
40
30
20
10
0
count
F
M
gender_(traditional)