Entertainer data analysis

April 29, 2024

1 Entertainer Data Analysis

[25]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[7]:
```
df1 = pd.read_csv('Entertainer - Basic Info.csv')
df2 = pd.read_csv('Entertainer - Breakthrough Info.csv')
df3 = pd.read_csv('Entertainer - Last work Info.csv')
```

[8]:
```
mer_df = pd.merge(df1, df2, how='left')
```

[9]:
```
mer_data = pd.merge(mer_df,df3, how = 'left')
mer_data.head(20)
```

[9]:
```
          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
5         Bing Crosby                    M        1903
6            Bob Hope                    M        1903
7       Carol Burnett                    F        1933
8      Carole Lombard                    F        1908
9       Carrie Fisher                    F        1956
10         Cary Grant                    M        1904
11    Charlie Chaplin                    M        1889
12          Clara Bow                    F        1905
13        Clark Gable                    M        1901
14    David Letterman                    M        1947
15    Debbie Reynolds                    F        1932
16  Denzel Washington                    M        1954
17      Dick Van Dyke                    M        1925
18  Donald Sutherland                    M        1935
19     Dustin Hoffman                    M        1937
```

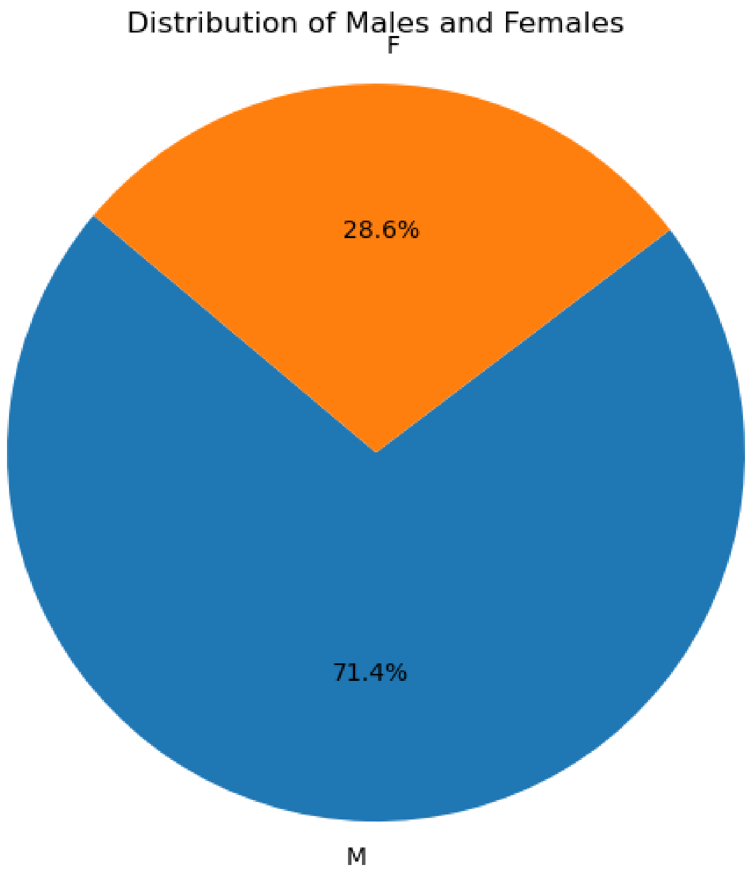
```
    Year of Breakthrough/#1 Hit/Award Nomination  \
0                                           2008   
1                                           1999   
2                                           1967   
3                                           1934   
4                                           1952   
5                                           1931   
6                                           1938   
7                                           1959   
8                                           1934   
9                                           1977   
10                                          1933   
11                                          1915   
12                                          1926   
13                                          1934   
14                                          1982   
15                                          1952   
16                                          1989   
17                                          1961   
18                                          1967   
19                                          1967   

                           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   
5                              Several Songs                           1962.0   
6                  The Big Broadcast of 1938                           1940.0   
7                       The Garry Moore Show                           1962.0   
8                          Twentieth Century                              NaN   
9                                  Star Wars                              NaN   
10          She Done Him Wrong, I'm No Angel                           1970.0   
11                                 The Tramp                           1929.0   
12                                   Mantrap                              NaN   
13                     It Happened One Night                           1934.0   
14           Late Night with David Letterman                           1981.0   
15                       Singin' in the Rain                              NaN   
16                                     Glory                           1989.0   
17    Bye Bye Birdie, The Dick Van Dyke Show                           1964.0   
18                           The Dirty Dozen                           1995.0   
19                              The Graduate                           1980.0   

    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
5                                 1974          1977.0
6                                 1972          2003.0
7                                 2016             NaN
8                                 1942          1942.0
9                                 2016          2016.0
10                                1966          1986.0
11                                1967          1977.0
12                                1933          1965.0
13                                1960          1960.0
14                                2015             NaN
15                                2006          2016.0
16                                2016             NaN
17                                2015             NaN
18                                2016             NaN
19                                2016             NaN
```

2 Explanatory Data Analysis

[56]:

```
# Vizualization of number of males and females in the field of entertainment
gender_counts = mer_data['Gender (traditional)'].value_counts()
plt.figure(figsize=(8, 6))
plt.pie(gender_counts, labels=gender_counts.index, autopct='%1.1f%%',␣
 ↪startangle=140)
plt.title('Distribution of Males and Females')
plt.axis('equal')
plt.show()
```



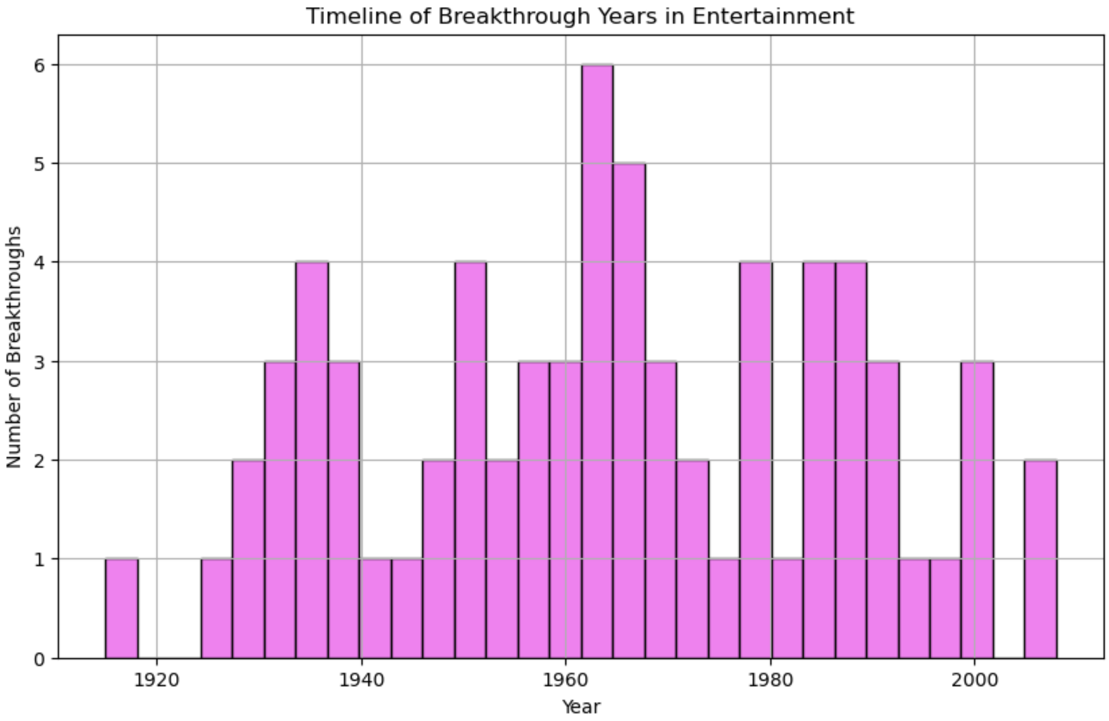
[81]:
```
numerical_columns = mer_data.select_dtypes(include=['number'])
numerical_columns.corr()
```

[81]:
```
                                              Birth Year  \
Birth Year                                      1.000000
Year of Breakthrough/#1 Hit/Award Nomination    0.941740
Year of First Oscar/Grammy/Emmy                 0.772863
Year of Last Major Work (arguable)              0.723677
Year of Death                                   0.561124

                                              Year of Breakthrough/#1 Hit/Award
Nomination  \
Birth Year
0.941740
Year of Breakthrough/#1 Hit/Award Nomination
1.000000
Year of First Oscar/Grammy/Emmy
0.842745
Year of Last Major Work (arguable)
```

```
0.748675
Year of Death
0.511714

                                              Year of First Oscar/Grammy/Emmy  \
Birth Year                                                           0.772863
Year of Breakthrough/#1 Hit/Award Nomination                         0.842745
Year of First Oscar/Grammy/Emmy                                      1.000000
Year of Last Major Work (arguable)                                   0.608658
Year of Death                                                        0.360906

                                              Year of Last Major Work (arguable)
\
Birth Year                                                              0.723677
Year of Breakthrough/#1 Hit/Award Nomination                            0.748675
Year of First Oscar/Grammy/Emmy                                         0.608658
Year of Last Major Work (arguable)                                      1.000000
Year of Death                                                           0.848694

                                              Year of Death
Birth Year                                         0.561124
Year of Breakthrough/#1 Hit/Award Nomination       0.511714
Year of First Oscar/Grammy/Emmy                    0.360906
Year of Last Major Work (arguable)                 0.848694
Year of Death                                      1.000000
```

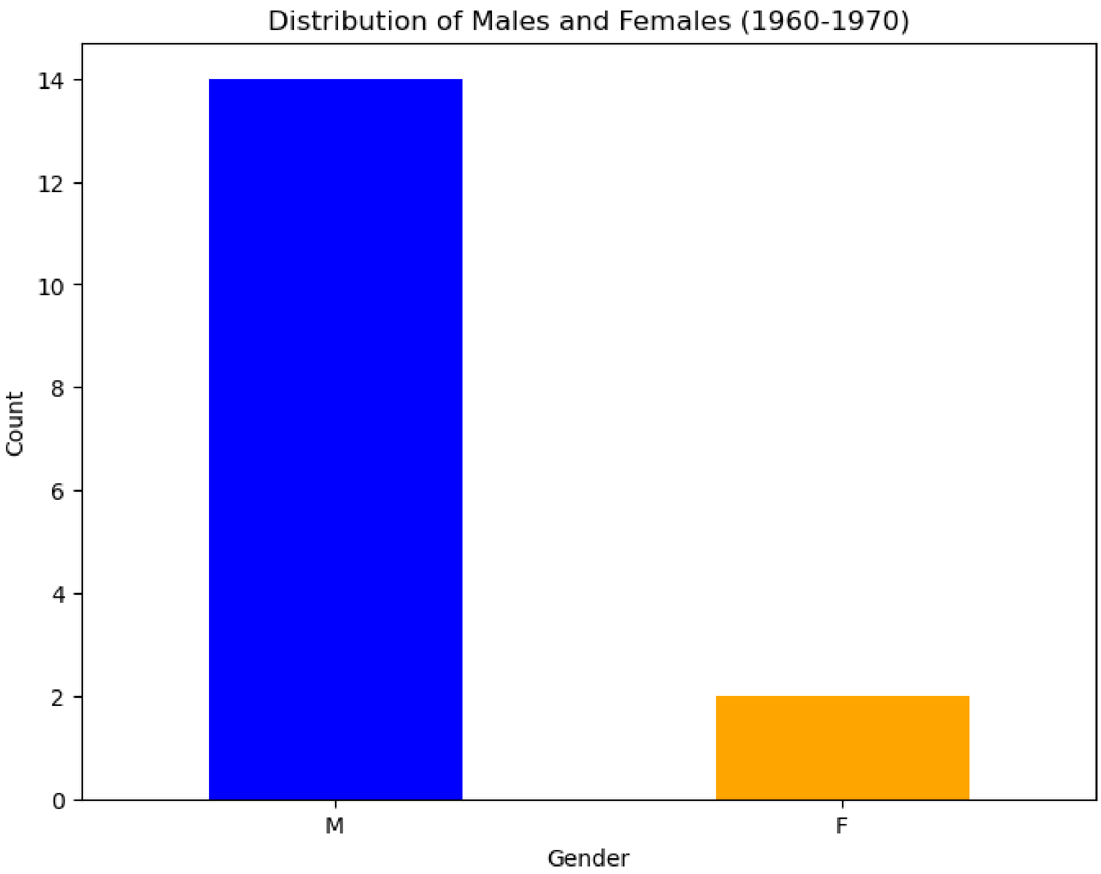
[33]:

```
# visualizing the trending period of entertainment from this dataset
plt.figure(figsize=(10, 6))
plt.hist(mer_data['Year of Breakthrough/#1 Hit/Award Nomination'], bins=30,␣
 ↪color='violet', edgecolor='black')
plt.title('Timeline of Breakthrough Years in Entertainment')
plt.xlabel('Year')
plt.ylabel('Number of Breakthroughs')
plt.grid(True)
plt.show()

# during the period of 1960-70, entertainment was the trending the most
```



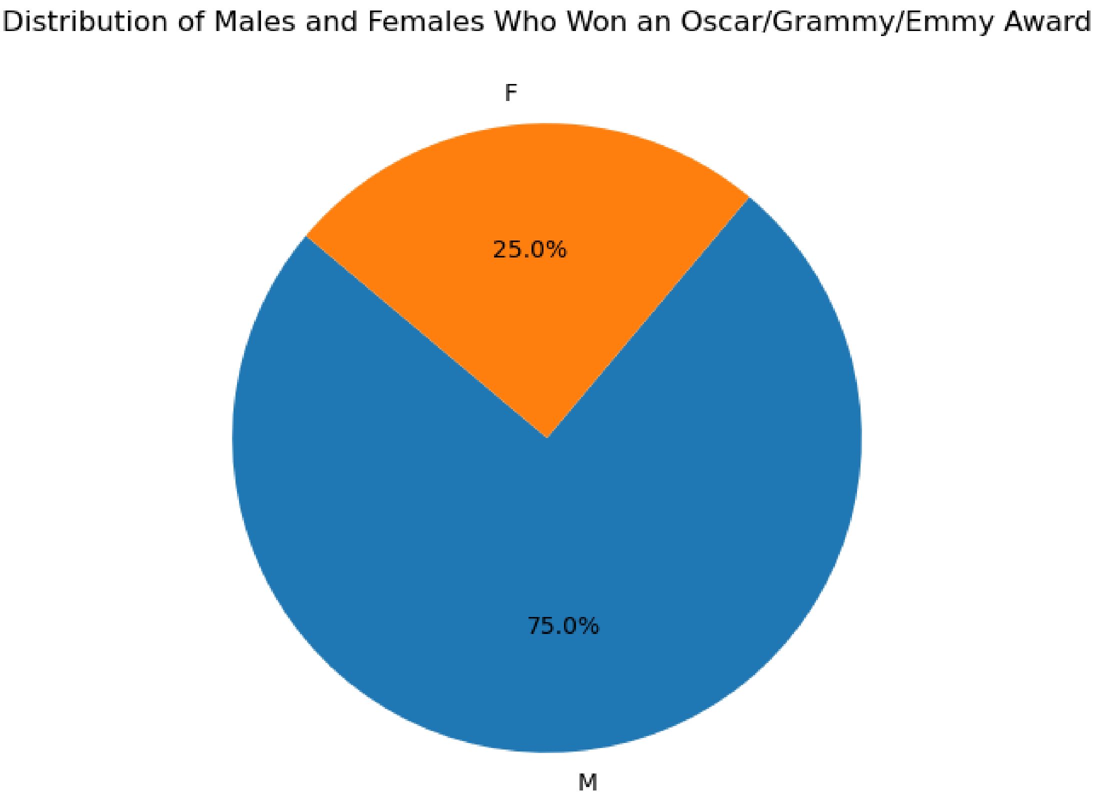
[39]:

```
filtered_df = mer_data[(mer_data['Year of Breakthrough/#1 Hit/Award␣
  ↪Nomination'] >= 1960) & (mer_data['Year of Breakthrough/#1 Hit/Award␣
  ↪Nomination'] <= 1970)]
gender_counts = filtered_df['Gender (traditional)'].value_counts()
plt.figure(figsize=(8, 6))
gender_counts.plot(kind='bar', color=['blue', 'orange'])
plt.title('Distribution of Males and Females (1960-1970)')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.xticks(rotation=0)
plt.show()

# During this trending period, the number of male entertainer is high
```

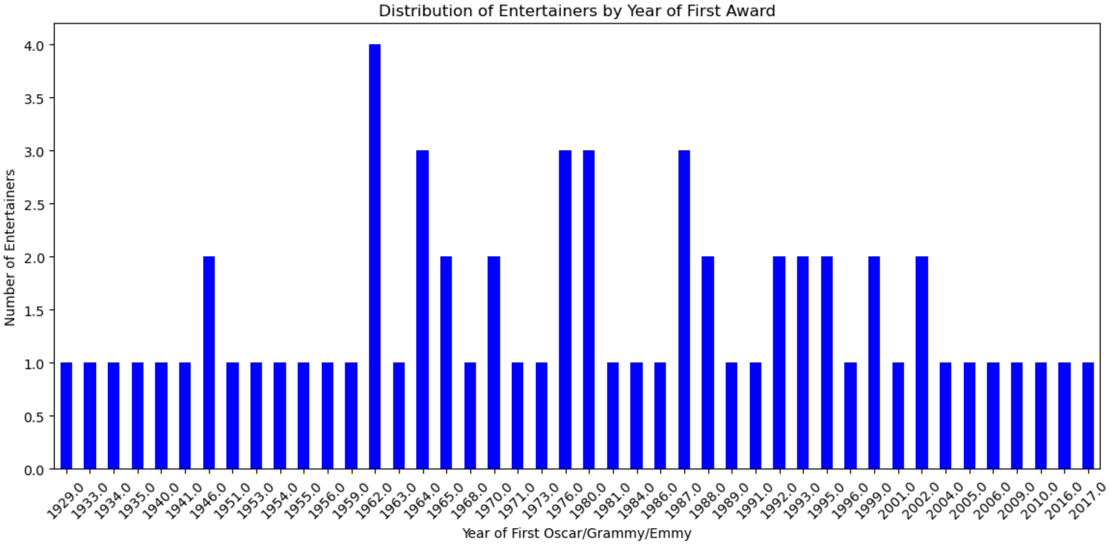


[58]:

```
modified_df = mer_data.copy()
modified_df['Year of First Oscar/Grammy/Emmy'] = modified_df['Year of First␣
  ↪Oscar/Grammy/Emmy'].notna().astype(int)
winners_df = modified_df[modified_df['Year of First Oscar/Grammy/Emmy'] == 1]
gender_counts = winners_df['Gender (traditional)'].value_counts()
plt.figure(figsize=(6, 6))
plt.pie(gender_counts, labels=gender_counts.index, autopct='%1.1f%%',␣
  ↪startangle=140)
plt.title('Distribution of Males and Females Who Won an Oscar/Grammy/Emmy␣
  ↪Award')
plt.show()
```

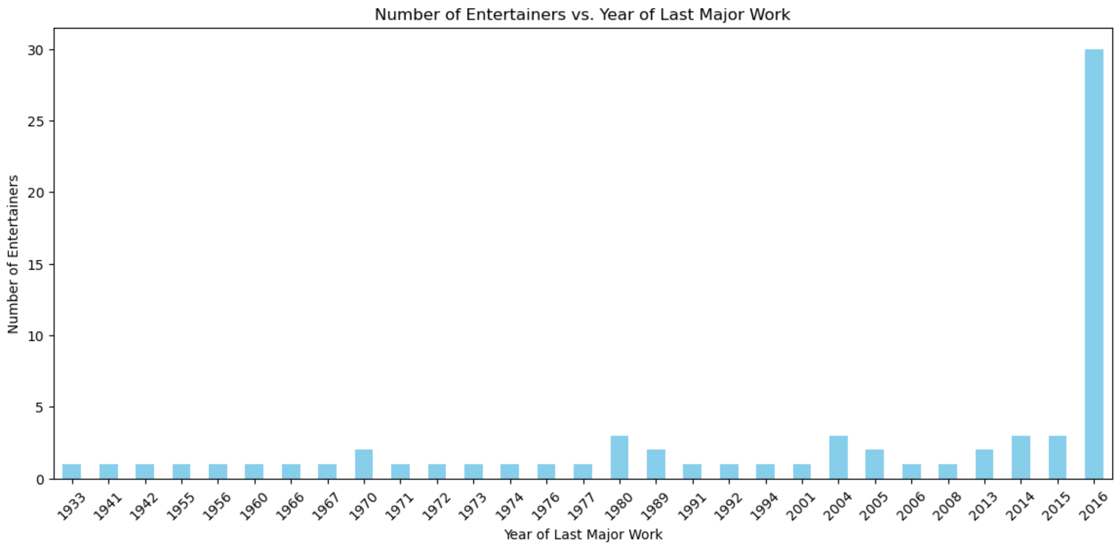


[85]:
```
# Visualize the year in which maximum number of Oscar was won by Entertainers
award_year_counts = mer_data['Year of First Oscar/Grammy/Emmy'].value_counts().
 ↪sort_index()

# Plot the bar graph
plt.figure(figsize=(14, 6))
award_year_counts.plot(kind='bar', color='blue')
plt.title('Distribution of Entertainers by Year of First Award')
plt.xlabel('Year of First Oscar/Grammy/Emmy')
plt.ylabel('Number of Entertainers')
plt.xticks(rotation=45)  # Rotate x-axis labels if necessary
plt.show()

# In 1962 the Oscar award were awarded to maximum number of entertainers
```



[72]:

```
# visualizing  the relationship between entertainers and their Year of Last Major Work (arguable).
year_counts = mer_data['Year of Last Major Work (arguable)'].value_counts().sort_index()
plt.figure(figsize=(14, 6))
year_counts.plot(kind='bar', color='skyblue')
plt.title('Number of Entertainers vs. Year of Last Major Work')
plt.xlabel('Year of Last Major Work')
plt.ylabel('Number of Entertainers')
plt.xticks(rotation=45)  # Rotate x-axis labels if necessary
plt.show()

# Most of the entertainer's last work is in the year 2016
```



[79]:

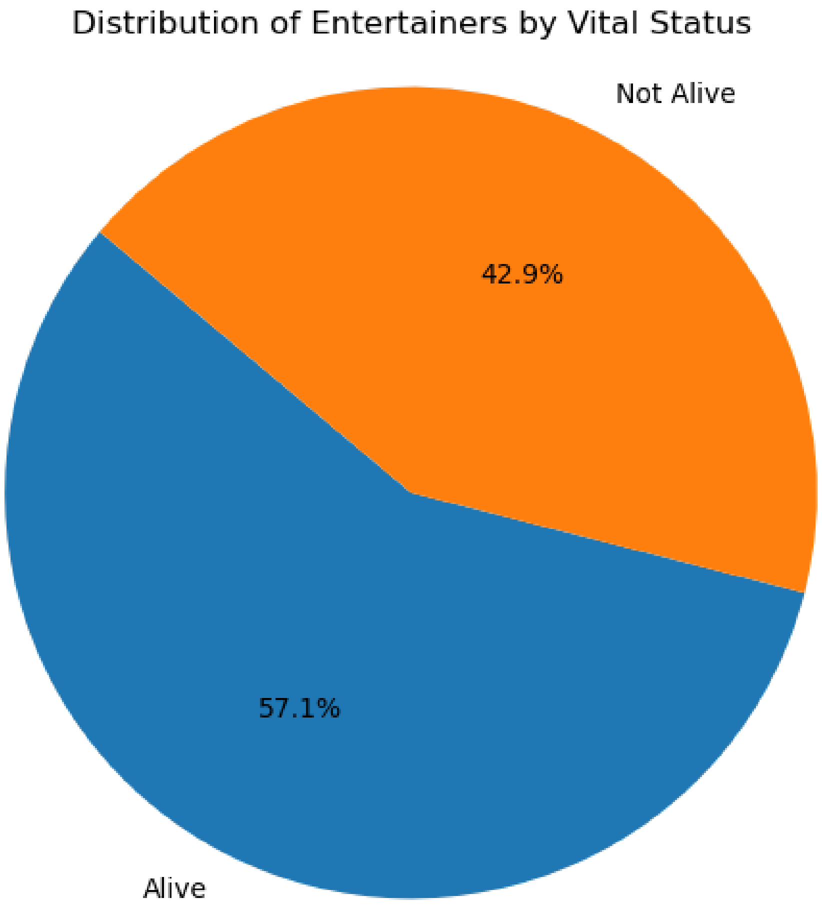
```
alive_df = mer_data.copy()
alive_df['Year of Death'].fillna('Alive', inplace=True)
num_alive = (alive_df['Year of Death'] == 'Alive').sum()
print("Number of entertainers who are still alive:", num_alive)
```

```
Number of entertainers who are still alive: 40
```

```
/tmp/ipykernel_3979/2298961232.py:2: FutureWarning: Setting an item of
incompatible dtype is deprecated and will raise in a future error of pandas.
Value 'Alive' has dtype incompatible with float64, please explicitly cast to a
compatible dtype first.
  alive_df['Year of Death'].fillna('Alive', inplace=True)
```

[80]:

```
# visualizing the graph of entertainers who are alive

alive_df = mer_data.copy()  # Make a copy to avoid modifying the original␣
  ↪DataFrame
alive_df['Vital Status'] = alive_df['Year of Death'].apply(lambda x: 'Alive' if␣
  ↪pd.isnull(x) else 'Not Alive')

# Count the number of entertainers in each vital status category
vital_status_counts = alive_df['Vital Status'].value_counts()

# Plot the pie chart
plt.figure(figsize=(6, 6))
plt.pie(vital_status_counts, labels=vital_status_counts.index, autopct='%1.
  ↪1f%%', startangle=140)
plt.title('Distribution of Entertainers by Vital Status')
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle
plt.show()
```


Distribution of Entertainers by Vital Status
Not Alive
42.9%
57.1%
Alive